



Меню



Вверх

Назад

Вперёд

Пред.

След.

Указатель

Помощь

Экран

Министерство образования Республики Беларусь
Белорусский государственный университет



ЭЛЕКТРОННЫЙ УЧЕБНО-МЕТОДИЧЕСКИЙ КОМПЛЕКС
ЧИСЛЕННЫЕ МЕТОДЫ

Настоящий электронный учебно-методический комплекс разработан по заданию
Министерства образования Республики Беларусь коллективом авторов в составе
Мандрик П. А. (научный руководитель), Репников В. И., Фалейчик Б. В.

-
- [Часть I. Руководство пользователя](#)
 - [Часть II. Учебные программы](#)
 - [Часть III. Теоретические материалы](#)
 - [Часть IV. Задачи](#)
 - [Часть V. Тесты](#)
 - [Рекомендуемая литература](#)
-



Меню



Вверх

Назад

Вперёд

Пред.

След.

Указатель

Помощь

Экран

Часть I

Руководство пользователя

Глава 1. Состав комплекса

Глава 2. Интерфейс



Глава 1

Состав комплекса

- 1.1. Учебные программы
- 1.2. Теоретические материалы
- 1.3. Задачи
- 1.4. Тесты
- 1.5. Демонстрации

В состав настоящего электронного учебно-методического комплекса входят:

- Электронный документ **Численные методы.pdf** — пособие по курсу, включающее:
 - руководство пользователя;
 - типовые учебные программы курсов;
 - теоретические материалы;
 - комплекс задач и упражнений;
 - тесты;
 - список рекомендуемой литературы по курсу.
- Интерактивные демонстрации для просмотра в Wolfram CDF Player — файлы с расширением .cdf, находящиеся в каталоге **Демонстрации**.



1.1. Учебные программы

В этой части ЭУМК содержатся типовые учебные программы курсов по дисциплинам численных методов в высших учебных заведениях Республики Беларусь. Темы из типовых программ связаны гиперссылками с соответствующими разделами теоретических материалов ЭУМК.



1.2. Теоретические материалы

Данная часть комплекса является наиболее объемной. Помимо самих теоретических материалов, разделенных на главы, эта часть документа содержит разделы «Основные понятия», «Предметный указатель» и «Доказательства теорем».

- В разделе «Основные понятия» собраны все определения, выделенные в основной части теоретических материалов.
- Раздел «Предметный указатель» содержит алфавитный список терминов, определенных в теоретической части. По щелчку на соответствующий термин пользователь попадает на соответствующее определение в разделе «Основные понятия», либо на страницу в «Теоретических материалах», на которой был определен соответствующий термин. Подавляющее большинство терминов, попавших в предметный указатель, в основном тексте выделены *наклонным шрифтом и зеленовато-голубым цветом*.
- В разделе «Доказательства теорем» собраны доказательства теорем и других утверждений, которые были вынесены из теоретических материалов с целью сделать текст менее громоздким.

1.3. Задачи

Задачи по курсу «Численные методы», которые находятся в соответствующей части электронного документа ЭУМК, могут быть снабжены ответами, решениями или указаниями к решению. Для их просмотра следует воспользоваться соответствующими гиперссылками.

Некоторые задачи в настоящем ЭУМК заимствованы из сборников [3], [5], [6].



1.4. Тесты

Каждый из тестов, входящих в состав комплекса, состоит из десяти вопросов. Тестовые вопросы могут быть двух типов.

Первый тип тестовых вопросов — простой вопрос типа «множественный выбор», в котором необходимо выбрать один правильный вариант ответа из нескольких. Заметим, что результат ответа показывается непосредственно сразу после щелчка мышью по полю выбора. Правильный ответ отмечается зеленой галочкой, неправильный — красным крестиком.

Вопросы второго типа требуют указания текстового или числового значения в специальном поле ввода. После осуществления ввода для получения результата ответа необходимо нажать на клавишу [Enter]. В случае правильного ответа вокруг поля ввода появится зеленая рамка, в противном случае рамка будет красной. Для получения правильного ответа можно нажать на кнопку [?], которой снабжено каждое поле ввода.

Начало каждого теста обозначено соответствующей надписью, в конце теста находится кнопка [Clear], нажатие на которую очищает все заполненные поля ответов.

Следует особо отметить, что тесты в настоящем ЭУМК предназначены лишь для самостоятельного контроля знаний при подготовке к зачету или экзамену.



1.5. Демонстрации

Интерактивные демонстрации представляют собой документы в формате Computable Document Format (CDF), которые реализуют демонстрационные приложения с полноценным графическим интерфейсом. Демонстрации находятся в отдельной папке с соответствующим названием. Для запуска выбранной демонстрации необходимо кликнуть правой кнопкой по иконке файла и в контекстном меню выбрать пункт «Открыть с помощью CDF Player». Если на компьютере пользователя не установлена программа Wolfram CDF Player, ее можно бесплатно скачать по адресу <http://www.wolfram.com/cdf-player/>.

При использовании демонстраций в качестве презентационного материала для лекций можно воспользоваться полноэкранным режимом, для включения/выключения которого в Wolfram CDF Player используется клавиша [F12].



Глава 2

Интерфейс

Для навигации по большому PDF-документу, которым является данный ЭУМК, используются два основных средства: панель закладок (bookmark panel) и элементы управления, размещенные в верхнем колонтитуле.

Панель закладок как правило находится в левой части окна. На ней иерархически организованные материалы ЭУМК представлены в виде дерева. Все разделы, помеченные знаком [+] могут быть раскрыты для чтения подразделов. Используя панель закладок, можно легко увидеть структуру учебника и перемещаться по его разделам. По умолчанию данная панель должна отображаться сразу при открытии документа ЭУМК. Если же она отсутствует, то либо ваша программа просмотра документов PDF не поддерживает закладки, либо отображение панели отключено. В последнем случае отображение панели закладок следует включить в меню настройки программы просмотра.

Верхний колонтитул каждой страницы ЭУМК снабжен дополнительными элементами навигации. Опишем их предназначение слева направо.

- Кнопка «Меню» служит для перехода на титульный лист, содержащий главное меню учебника.
- Навигационное меню состояния показывает в каком разделе документа находится пользователь и позволяет осуществлять быстрый переход на разделы высших уровней путём нажатия на соответствующие гиперссылки.
- Кнопка «Вверх» служит для быстрого перехода к разделу уровня на единицу выше.
- Кнопки «Назад» и «Вперед» служат для перехода по истории просмотра.



- Стрелки «Пред.» и «След.» позволяют переходить на следующую/предыдущую страницу документа. Отметим, что при использовании программы Acrobat Reader эти функции можно выполнить клавишами курсоров, клавишами [Page Up] и [Page Down], а также используя полосу прокрутки или колесико мыши.
- Кнопка «Указатель» служит для перехода к алфавитному предметному указателю по теоретическим материалам.
- По нажатии на кнопку «Помощь» пользователь переходит к разделу «Руководство пользователя».
- Кнопка «Экран» служит для быстрого включения и выключения полноэкранного режима просмотра.

Учебные материалы комплекса связаны между собой системой гиперссылок. Все гиперссылки набраны в тексте прямым шрифтом и выделены **синим цветом**. Ссылки, оформленные в виде отдельных кнопок, кроме того, заключены в квадратные скобки.



Меню



Часть II

Учебные программы

- [Глава 1. Указатель по специальностям](#)
- [Глава 2. Список программ](#)



Глава 1

Указатель по специальностям

- 1-31 03 03 Прикладная математика (по направлениям):
[Вычислительные методы алгебры](#)
[Методы численного анализа](#)
[Численные методы математической физики](#)
- 1-31 03 04 Информатика:
[Вычислительные методы алгебры](#)
[Методы численного анализа](#)
- 1-31 03 05 Актуарная математика:
[Вычислительные методы алгебры](#)
[Методы численного анализа](#)
- 1-31 03 06-01 Экономическая кибернетика (математические методы и компьютерное моделирование в экономике):
[Вычислительные методы алгебры](#)
[Методы численного анализа](#)
- 1-31 04 02 Радиофизика: [Численные методы](#)
- 1-31 04 03 Физическая электроника: [Численные методы](#)



- 1-98 01 01-01 Компьютерная безопасность (математические методы и программные системы):
[Вычислительные методы алгебры](#)
[Методы численного анализа](#)
- 1-98 01 01 Компьютерная безопасность (направление 1-98 01 01-02 радиофизические методы и программно-технические средства): [Численные методы](#)



Глава 2

Список программ

- 2.1. Численные методы
- 2.2. Вычислительные методы алгебры
- 2.3. Методы численного анализа - I
- 2.4. Методы численного анализа - II
- 2.5. Численные методы математической физики



2.1. Численные методы

Типовая учебная программа для высших учебных заведений по специальностям:

1-31 04 02 Радиофизика, 1-31 04 03 Физическая электроника,

1-98 01 01 Компьютерная безопасность (по направлениям) (направление 1-98 01 01-02 радиофизические методы и программно-технические средства)

Тема 1. Введение

Схема постановки вычислительного эксперимента. Требования, предъявляемые к вычислительным алгоритмам: устойчивость, точность, эффективность, экономичность. Основные источники погрешностей.

Тема 2. Численное решение систем линейных алгебраических уравнений

Характеристика прямых и итерационных методов. Прямые методы Гаусса, LU и LL- факторизации. Точность и устойчивость решения. Выбор главных элементов. Прямые методы для больших систем с разреженной матрицей. Схема итерационных методов первого порядка. Условие сходимости. Методы простой итерации, Гаусса–Зейделя, последовательной релаксации, неполной факторизации. Метод сопряженных градиентов. Приближенные методы вычисления собственных значений и собственных векторов.

Тема 3. Численное решение нелинейных уравнений и систем

Отделение корней, уточнение корней. Итерационные методы для уравнений и систем: итераций, Ньютона. Выбор начального приближения.

Тема 4. Приближение функций

Интерполирование функций. Интерполяционные формулы Лагранжа и Ньютона. Погрешности и условия применимости интерполяционных формул. Интерполяция локально определенными функциями: кусочно-линейная и сплайн-интерполяция. Аппроксимация функций. Метод наименьших квадратов.

Тема 5. Численное интегрирование

Методы вычисления определенных интегралов: трапеций, Симпсона. Формула Симпсона для двумерных интегралов. Вычисление многократных интегралов методом Монте – Карло.

Тема 6. Решение систем обыкновенных дифференциальных уравнений

Характеристика явных и неявных, одношаговых и многошаговых методов. Явный метод Эйлера, его



точность и устойчивость. Жесткие системы. Неявный метод Эйлера. Методы Рунге–Кутта. Явные и неявные многошаговые методы.



2.2. Вычислительные методы алгебры

Типовая учебная программа для высших учебных заведений по специальностям:

1-31 03 03 Прикладная математика (по направлениям); 1-31 03 04 Информатика;

1-31 03 05 Актуарная математика; 1-31 03 06-01 Экономическая кибернетика (математические методы и компьютерное моделирование в экономике);

1-98 01 01-01 Компьютерная безопасность (математические методы и программные системы)

1. Введение

Предмет «Вычислительные методы алгебры» и основные задачи, излагаемые в указанном курсе.

Раздел I. Методы решения систем линейных алгебраических уравнений

2. Обусловленность

Общая характеристика проблем решения систем линейных алгебраических уравнений (СЛАУ), решения задач на собственные значения, понятий корректности и устойчивости СЛАУ. Устойчивость решения СЛАУ по правой части и коэффициентная устойчивость. Число обусловленности матрицы и его свойства. Хорошо обусловленные и плохо обусловленные СЛАУ. Геометрическая интерпретация понятия обусловленности.

[Метод регуляризации](#).

3. Прямые методы решения СЛАУ

Общая характеристика прямых методов решения СЛАУ. Теорема об LU-разложении. Схема единственного деления и ее связь с теоремой об LU-разложении. Методы Гаусса с выбором главного элемента. Вычисление определителей и обращение матриц с помощью метода Гаусса. Метод квадратного корня. Метод Жордана обращения матриц. Диагонально доминирующие матрицы. Ортогональные преобразования. Методы отражений, вращений и ортогонализации. Метод прогонки решения СЛАУ с трехдиагональной матрицей. Связь метода прогонки с методом Гаусса. Теорема о корректности метода прогонки. Методы правой, встречной и циклической прогонки. Теорема о корректности метода циклической прогонки

4. Итерационные методы решения СЛАУ

Общая характеристика итерационных методов решения СЛАУ. Сходимость матричной геометрической прогрессии. Градиент функционала. Методы простой итерации и Зейделя. Теоремы сходимости. Элементы



[теории двухслойных итерационных методов](#). Основная теорема сходимости. [Методы Якоби, Гаусса-Зейделя и релаксации](#). Оптимизация сходимости итерационных процессов. Итерационные методы вариационного типа и теоремы их сходимости.

Раздел II. Методы решения задач на собственные значения

5. Полная проблема собственных значений

[Общая постановка задачи на собственные значения](#). Устойчивость задачи на собственные значения. Методы [Данилевского](#), Крылова, Леверье и видоизменение Фаддеева. Прямые методы отражений и вращений. [Итерационный метод вращений](#). QR-алгоритм. Метод бисекций решения полной проблемы собственных значений.

6. Частичная проблема собственных значений

[Степенной метод вычисления наибольшего по модулю собственного значения и его модификации](#). Метод обратных итераций. Метод λ -разности. Ускорение сходимости степенного метода.



2.3. Методы численного анализа - I

Типовая учебная программа для высших учебных заведений по специальностям:
1-31 03 03 Прикладная математика (по направлениям);

1. Введение

Предмет «Методы численного анализа» и основные задачи, излагаемые в указанном курсе.

Раздел I. Решение нелинейных уравнений и систем

2. Итерационные методы решения нелинейных уравнений и систем

Решение нелинейных уравнений. [Метод простых итераций](#). Теорема о сходимости. Ускорение сходимости метода итерации. Метод Стеффенсена. Методы типа Чебышева. Метод Ньютона для одного уравнения. Видоизменения метода Ньютона. Поиск всех корней алгебраических уравнений. Метод Лобачевского. Метод Лина. Решение систем нелинейных уравнений. Метод простых итераций. Методы Зейделя и Гаусса-Зейделя. Метод Ньютона и его видоизменения.

3. Вариационный подход к решению нелинейных систем

Сведение решения системы нелинейных уравнений к решению вариационной задачи. Метод покоординатного спуска. Метод градиентного спуска. Проблема выбора начального приближения. Метод продолжения по параметру.

Раздел II. Приближение функций

4. Наилучшие приближения

Задача о наилучшем приближении в линейных нормированных пространствах. Наилучшее приближение в гильбертовом пространстве. Среднеквадратичное приближение функций алгебраическими многочленами. Метод наименьших квадратов. Задача построения ортонормированного базиса. Наилучшее равномерное приближение. Теорема о чебышевском альтернансе. Примеры построения многочленов наилучшего равномерного приближения.

5. Интерполярование



Интерполяция в линейных нормированных пространствах. Алгебраическое интерполяция. Интерполяционный многочлен в форме Лагранжа. Интерполяционный многочлен в форме Ньютона для неравномерной сетки. Интерполяционные формулы Ньютона для равномерной сетки. Интерполяционная формула Стирлинга. Многочлены Чебышева. Минимизация остатка интерполяции. Интерполяция с кратными узлами. Многочлен Эрмита. Сходимость интерполяционного процесса. Применение интерполяции к вычислению производных. Формулы численного дифференцирования и их погрешности. Интерполяционные методы решения нелинейных уравнений.

6. Сплайн-приближения

Сплайн-интерполяция. Интерполяционный кубический сплайн. Экстремальное свойство интерполяционного кубического сплайна. Сплайн-сглаживание. Многомерная алгебраическая интерполяция. Бикубический сплайн. Приближение кривых и поверхностей. Интерполяционный параметрический сплайн.

Раздел III. Численное интегрирование

7. Интерполяционные квадратурные формулы

Квадратурные формулы и связанные с ними задачи. Интерполяционные квадратурные формулы. Квадратурные формулы Ньютона-Котеса. Простейшие квадратурные формулы (прямоугольников, трапеций, Симпсона). Правило Рунге оценки точности квадратурных формул и автоматический выбор шага интегрирования.

8. Квадратурные формулы типа Гаусса

Квадратурные формулы наивысшей алгебраической степени точности (НАСТ). Теоремы существования и единственности, о свойствах узлов квадратурных формул НАСТ. Частные случаи квадратурных формул НАСТ. Квадратурные формулы с заранее предписанными узлами и равными коэффициентами. Ослабление особенностей интегрируемой функции.

9. Приближенное вычисление кратных интегралов

Понятие о кубатурных формулах. Кубатурная формула трапеций на прямоугольной сетке. Кубатурная формула средних на прямоугольной сетке. Кубатурная формула Симпсона. Кубатурная формула средних на треугольной сетке. Кубатурная формула повышенного порядка точности на треугольной сетке.

Раздел IV. Численное решение интегральных уравнений

10. Методы решения интегральных уравнений Фредгольма и Вольтерра второго рода



Метод механических квадратур, метод замены ядра на вырожденное и метод последовательных приближений решения интегрального уравнения Фредгольма второго рода. Метод квадратур и метод последовательных приближений решения интегрального уравнения Вольтерра второго рода. Метод Галеркина решения интегральных уравнений Фредгольма и Вольтерра второго рода.

11. Методы решения некорректных задач

Понятие устойчивости и корректности задач. Уравнение Фредгольма первого рода как некорректная задача. Метод регуляризации решения некорректных задач.

Раздел V. Методы численного решения задачи Коши

12. Методы решения нежестких задач

Классификация методов. Построение одношаговых методов способом разложения решения в ряд Тейлора. Методы типа Рунге–Кутта. Построение вычислительных правил на основе принципа последовательного повышения порядка точности. Правило Рунге оценки погрешности приближенного решения. Вложенные методы типа Рунге–Кутта. Сходимость одношаговых методов. Многошаговые методы. Экстраполяционный и интерполяционный методы Адамса. Общие линейные многошаговые методы.

13. Методы решения жестких систем

Устойчивость численных методов решения задачи Коши. Жесткие задачи и методы их решения. Неявные методы Рунге–Кутта. Формулы дифференцирования назад.

Раздел VI. Решение граничных задач для обыкновенных дифференциальных уравнений

14. Методы, основанные на сведении к задаче Коши

Метод стрельбы. Метод редукции. Метод дифференциальной прогонки.

15. Вариационные методы решения граничных задач

Вариационно-проекционные методы решения граничных задач: методы моментов, Галеркина, Ритца, наименьших квадратов.

16. Сеточные методы решения граничных задач

Сеточные методы решения граничных задач. Разностная аппроксимация простейших дифференциальных операторов. Постановка разностной задачи. Погрешность аппроксимации разностных схем. Повышение порядка аппроксимации. Корректность и устойчивость разностных схем. Теорема о сходимости. Математический аппарат теории разностных схем: формулы разностного дифференцирования произведения,



суммирования почастям, разностные аналоги теорем вложения. Требования, предъявляемые к разностным схемам. Свойства консервативности и однородности разностных схем. Основные способы построения разностных схем: интегро-интерполяционный, метод Ритца, метод Галеркина, методы аппроксимации квадратичного функционала и сумматорного тождества. Принцип максимума. Монотонные разностные схемы.



2.4. Методы численного анализа - II

Типовая учебная программа для высших учебных заведений по специальностям:

1-31 03 04 Информатика; 1-31 03 05 Актуарная математика; 1-31 03 06-01 Экономическая кибернетика
(математические методы и компьютерное моделирование в экономике); 1-98 01 01-01 Компьютерная
безопасность (математические методы и программные системы)

1. Введение Предмет дисциплины «Методы численного анализа» и основные задачи, излагаемые в указанном курсе.

Раздел I. Методы решения нелинейных уравнений

2. Итерационные методы решения нелинейных уравнений и систем

Метод простых итераций решения нелинейных уравнений и систем. Теорема сходимости. Аналог метода Зейделя. Метод Ньютона для одного уравнения. Видоизменения метода Ньютона. Метод Ньютона для систем нелинейных уравнений.

3. Вариационный подход к решению нелинейных систем

Сведение решения системы нелинейных уравнений к решению вариационных задач. Метод покоординатного спуска. Метод градиентного спуска.

Раздел II. Приближение функций

4. Интерполирование

Постановка задачи интерполирования и ее разрешимость. Алгебраическое интерполирование. Интерполяционный многочлен в форме Лагранжа. Остаток интерполирования в форме Лагранжа. Разделенные разности и их свойства. Интерполяционный многочлен в форме Ньютона для неравномерной сетки. Конечные разности и их свойства. Интерполяционные формулы Ньютона для равномерной сетки. Интерполяционная формула Стирлинга. Многочлены Чебышева. Минимизация остатка интерполирования. Интерполирование с кратными узлами. Многочлен Эрмита. Остатки интерполирования с кратными узлами.

5. Сплайн-приближения



Понятие сплайн-функции. Сплайн-интерполяция. Построение кубического сплайна. Вариационная и физическая интерпретация кубического сплайна.

6. Наилучшие приближения

Задача о наилучшем приближении в линейных нормированных пространствах. Метод наименьших квадратов. Среднеквадратичные приближения. Применение интерполяции к вычислению производных. Погрешность формул приближенного дифференцирования.

Раздел III. Численное интегрирование

7. Интерполяционные квадратурные формулы

Квадратурные формулы и связанные с ними задачи. Интерполяционные квадратурные формулы. Простейшие квадратурные формулы Ньютона-Котеса. Квадратурные формулы прямоугольников, трапеций, Симпсона. Оценки точности квадратурных формул. Правило Рунге и автоматический выбор шага интегрирования.

8. Квадратурные формулы типа Гаусса

Квадратурные формулы наивысшей алгебраической степени точности (НАСТ). Критерий и свойства квадратурных формул НАСТ. Теоремы существования, единственности и о свойствах узлов квадратурных формул НАСТ. Частные случаи квадратурных формул НАСТ. Выделение особенностей интегрируемых функций.

Раздел IV. Численное решение интегральных уравнений

9. Методы решения интегральных уравнений Фредгольма и Вольтерра второго рода

Метод механических квадратур решения интегрального уравнения Фредгольма второго рода. Метод замены ядра на вырожденное. Метод последовательных приближений решения интегрального уравнения Фредгольма второго рода. Метод квадратур и метод последовательных приближений решения интегрального уравнения Вольтерра второго рода. Метод Галеркина решения интегральных уравнений Фредгольма и Вольтерра второго рода.

10. Методы решения некорректных задач

Понятие устойчивости и корректности задач. Уравнение Фредгольма первого рода как некорректная задача. Метод регуляризации решения некорректных задач.

Раздел V. Методы численного решения обыкновенных дифференциальных уравнений



11. Методы решения задачи Коши

Методы решения задачи Коши. Построение одношаговых методов способом разложения решения в ряд Тейлора. Методы типа Рунге–Кутта. Построение вычислительных правил на основе принципа последовательного повышения порядка точности. Главный член погрешности. Правило Рунге. Методы решения жестких систем. Многошаговые методы. Экстраполяционный и интерполяционный методы Адамса.

12. Методы решения краевых задач

Многоточечные и граничные задачи. Решение линейных граничных задач. Метод дифференциальной прогонки. Метод стрельбы. Метод редукции. Методы решения нелинейных задач. Метод сеток решения граничных задач. Разрешимость системы разностных уравнений. Метод разностной прогонки. Методы моментов, Галеркина, Ритца, наименьших квадратов.

Раздел VI. Методы численного решения дифференциальных уравнений с частными производными

13. Элементы теории разностных схем

Основные понятия теории разностных схем. Аппроксимация простейших дифференциальных операторов. Постановка разностной задачи. Сходимость и устойчивость разностных схем. Математический аппарат теории разностных схем.

14. Разностные схемы для основных уравнений математической физики

Разностные схемы для уравнения теплопроводности, переноса, колебания струны. Устойчивость и методы реализации. Разностная задача Дирихле для уравнения Пуассона и методы ее реализации. Метод конечных элементов. Экономичные разностные схемы для многомерного уравнения теплопроводности. Нелинейная задача теплопроводности и разностные схемы ее решения.



2.5. Численные методы математической физики

Типовая учебная программа для высших учебных заведений по специальностям:

1-31 03 03 Прикладная математика (по направлениям)

1. Введение

Математическое моделирование и вычислительный эксперимент. Роль вычислительного эксперимента в исследовании физических процессов. Типичные задачи математической физики.

Раздел I. Введение в теорию разностных схем

2. Основные понятия теории разностных схем

Сетки и сеточные функции. Сетка в криволинейной системе координат. Треугольная сетка. Разностная аппроксимация простейших дифференциальных операторов в частных производных. Самосопряженность разностного оператора Лапласа. Постановка разностных задач математической физики. Погрешность аппроксимации разностных схем. Повышение порядка аппроксимации. Аппроксимация краевых и начальных условий. Устойчивость и сходимость разностной схемы.

3. Методы исследования устойчивости

Принцип максимума. Метод гармоник. Примеры применения принципа максимума и метода гармоник. Монотонные схемы для уравнений конвективного типа. Разностные схемы для уравнения переноса.

Раздел II. Разностные схемы для стационарных уравнений

4. Схемы и способы их построения

Консервативные схемы. Интегро-интерполяционный метод построения консервативных схем. Применение интегро-интерполяционного метода в случае двухмерной стационарной задачи теплопроводности. Разностная задача Дирихле для уравнения Пуассона в прямоугольнике. Устойчивость и сходимость. Схема повышенного порядка точности.

5. Методы решения сеточных уравнений

Методы Якоби, Зейделя, верхней релаксации. Итерационный метод переменных направлений. Метод редукции.



6. Численные методы решения задач математической физики в областях сложной формы

Краевые задачи для уравнения Пуассона в области сложной формы. [Разностный метод](#). [Метод замены переменных](#). Метод контрольного объема. [Метод конечных элементов](#). [Метод граничных элементов](#).

Раздел III. Разностные схемы для нестационарных уравнений

7. Разностные схемы для одномерных уравнений

[Одномерное уравнение теплопроводности](#). Семейство двухслойных схем для уравнения с постоянными коэффициентами. Погрешность аппроксимации и устойчивость. Краевые условия третьего рода. Уравнение с переменными коэффициентами и нелинейные уравнения. [Уравнение колебаний струны](#). Семейство трехслойных схем. Погрешность аппроксимации и устойчивость.

8. Многомерные задачи

Экономичные разностные схемы для многомерных задач математической физики. Схема переменных направлений. Устойчивость. Погрешность аппроксимации.



Часть III

Теоретические материалы

[Глава 1. Основы машинных вычислений](#)

[Глава 2. Методы решения СЛАУ](#)

[Глава 3. Методы решения проблемы собственных значений](#)

[Глава 4. Решение численных уравнений](#)

[Глава 5. Приближение функций](#)

[Глава 6. Приближенное вычисление интегралов](#)

[Глава 7. Численное решение интегральных уравнений](#)

[Глава 8. Методы решения задачи Коши для обыкновенных дифференциальных уравнений](#)

[Глава 9. Методы решения граничных задач для обыкновенных дифференциальных уравнений](#)

[Глава 10. Численные методы математической физики](#)

[Предметный указатель](#)

[Определения](#)

[Доказательства теорем](#)



Глава 1

Основы машинных вычислений

- 1.1. Машинная арифметика
- 1.2. Обусловленность задачи



1.1. Машинная арифметика

- 1.1.1. Числа с плавающей точкой
- 1.1.2. Двоичные числа с плавающей точкой
- 1.1.3. Способы округления
- 1.1.4. Расширение множества чисел с плавающей точкой
- 1.1.5. Машинный эпсилон
- 1.1.6. Стандарт IEEE 754
- 1.1.7. Проблемы машинных вычислений



1.1.1. Числа с плавающей точкой

Для того, чтобы использовать вещественные числа в машинных вычислениях, необходимо решить следующую общую проблему: каким образом сохранить произвольное $x \in \mathbb{R}$ в ограниченном количестве ячеек памяти? Существует несколько способов решения этой проблемы, и наиболее распространённым является представление x в виде числа с плавающей точкой.

Определение. Пусть $\beta \in \mathbb{N}$ — основание системы счисления, $p \in \mathbb{N}$ — число значащих разрядов, d_i — цифры. Вещественное число вида

$$\pm \underbrace{d_0.d_1d_2\dots d_{p-1}}_m \times \beta^e, \quad 0 \leq d_i < \beta, \quad (1.1)$$

называется **числом с плавающей точкой** (ЧПТ). Число $m \in \mathbb{R}$ называют **мантиссой** или **значащей частью**. Число $e \in \mathbb{Z}$ называют **показателем**, или **экспонентой** (не путать с числом e).

Представление (1.1) для любого x , очевидно, не является единственным — оно зависит от «положения точки»:

$$0.0001234 = 0.0012340 \times 10^{-1} = 1.2340000 \times 10^{-4}.$$

Поэтому по умолчанию используется так называемая нормализованная форма записи ЧПТ, в которой точка ставится после первой значащей цифры.

Определение. Число с плавающей точкой с ненулевым первым разрядом ($d_0 \neq 0$) называется **нормализованным**. Множество всех нормализованных ЧПТ с основанием β , p -разрядной мантиссой, и $e_{\min} \leq e \leq e_{\max}$ условимся обозначать $\mathbb{F}_1(\beta, p, e_{\min}, e_{\max})$ или просто \mathbb{F}_1 .



1.1.2. Двоичные числа с плавающей точкой

Рассмотрим подробно случай $\beta = 2$, так как именно такая арифметика используется в большинстве современных компьютеров. Возьмём $p = 3$, $e_{\min} = -1$, $e_{\max} = 2$. Все соответствующие положительные нормализованные ЧПТ приведены в таблице, они же изображены на рис. 1.1. Степени двойки, соответствующие единичному значению мантиссы, обведены окружностями.

Таблица 1.1 Все положительные элементы множества $\mathbb{F}_1(2, 3, -1, 2)$.

$m \setminus e$	-1	0	1	10 (2)
1.00	0.1 (0.5)	1 (1)	10 (2)	100 (4)
1.01	0.101 (0.625)	1.01 (1.25)	10.1 (2.5)	101 (5)
1.10	0.110 (0.75)	1.1 (1.5)	11 (3)	110 (6)
1.11	0.111 (0.875)	1.11 (1.75)	11.1 (3.5)	111 (7)

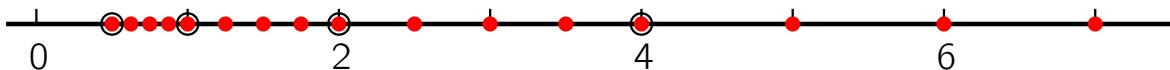


Рисунок 1.1

Приведённые данные наглядно демонстрируют следующие важные свойства, общие для всех нормализованных ЧПТ.

Свойства множества нормализованных чисел с плавающей точкой

- Отсутствие нуля: $0 \notin \mathbb{F}_1$.
- ЧПТ расположены на числовой прямой неравномерно.



3. Чем больше модуль $\xi \in \mathbb{F}_1$, тем больше и расстояние между ξ и соседними элементами \mathbb{F}_1 .

4. Между нулём и минимальным положительным $\xi_{\min} \in \mathbb{F}_1$ существует «зазор», ширина которого больше расстояния от ξ_{\min} до следующего ЧПТ в 2^{p-1} раз.

Двоичные ЧПТ выгодно отличаются от остальных тем, что в их нормализованной записи первый разряд d_0 всегда равен 1, поэтому его в памяти можно не хранить. Таким образом для хранения p -разрядной двоичной мантиссы достаточно $(p - 1)$ битов.



1.1.3. Способы округления

Понятно, что если представление x в β -ичной системе счисления содержит больше p значащих цифр, мы не можем точно записать его в виде (1.1). В этом случае можно лишь приблизить x некоторым ЧПТ, которое в дальнейшем будем обозначать $R(x)$.

Определение. *Правилом округления* для данного множества чисел с плавающей точкой $\mathbb{F} \subset \mathbb{R}$ будем называть отображение

$$R : \mathbb{R} \rightarrow \mathbb{F}$$

такое, что $R(x) = x$, если $x \in \mathbb{F}$, и $R(x) \approx x$ в противном случае.

Рассмотрим несколько способов задания R , считая $\beta = 10$, $p = 3$.

- 1) Отбрасывание «лишних» знаков ($R = R_d$): $R_d(0.12345) = 1.23 \times 10^{-1}$.
- 2) Округление вверх («школьное округление», $R = R_u$). $R_u(543.21) = 5.43 \times 10^2$, $R_u(5678) = 5.68 \times 10^3$.
В случае, когда запись x заканчивается на 5, его округляют до большего ЧПТ (вверх) : $R_u(23.45) = 2.35 \times 10^1$.
- 3) Округление до чётного ($R = R_e$). Этот способ отличается от предыдущего только трактовкой «спорного» случая, когда x находится ровно между двумя ЧПТ \underline{x} и \bar{x} . Оба эти приближения на самом деле равноправны, поэтому вместо того, чтобы всегда выбирать \bar{x} , мы будем с 50% вероятностью брать $R_e(x) = \underline{x}$ либо $R_e(x) = \bar{x}$.

Реализовать это можно, всегда выбирая из \underline{x} и \bar{x} то, мантисса которого заканчивается на чётную цифру. Таким образом, получаем: $R_e(23.45) = 2.34 \times 10^1$, но $R_e(23.55) = 2.36 \times 10^1$.

Возникает вопрос: какой из описанных способов округления лучше? Ответ на него даёт следующая теорема.

Теорема 1.1. Пусть x и y — два числа с плавающей точкой. Рассмотрим последовательность $\{x_i\}$, определённую по правилу

$$x_0 = x, \quad x_{i+1} = R(R(x_i + y) - y).$$

Если $R = R_e$, то либо $x_i = x \forall i \geq 0$, либо $x_i = x_1 \forall i \geq 1$.

Заметим, что по стандарту IEEE 754 в современных ЭВМ используется $R = R_e$.



1.1.4. Расширение множества чисел с плавающей точкой

Денормализованные числа

Специальные величины

Определение машинной арифметики

Глядя на рис. 1.1, мы видим, что для практического использования машинной арифметики нам не достаточно множества нормализованных ЧПТ \mathbb{F}_1 . Как минимум к этому множеству нужно добавить нуль (см. свойство 1). Кроме этого, современные машинные арифметики включают также специальные значения для обозначения бесконечностей, результатов некорректных операций и т. п.

Денормализованные числа

Наличие «зазора» между нулём и минимальным положительным нормализованным ЧПТ (свойство 4) может привести к серьёзным проблемам на практике. Рассмотрим, например, ЧПТ $x = 0.75$ и $y = 0.625$ из рассмотренного модельного множества $\mathbb{F}_1(2, 3, -1, 2)$. Так как $x - y = 0.125$, при любом разумном способе округления мы имеем $R(x - y) = 0$. То есть, например, выполнение обычного кода типа

```
if (x != y) then z = 1/(x - y)
```

в нашем случае приведёт к плачевному результату.

Эта проблема в современных машинных арифметиках решается дополнением множества нормализованных ЧПТ так называемыми денормализованными числами.

Определение. Вещественные числа вида

$$0.d_1d_2\dots d_{p-1} \times \beta^{e_{\min}},$$

где d_i — произвольные β -ичные цифры, называются денормализованными числами с плавающей точкой (ДЧПТ). Множество всех ДЧПТ с параметрами β , p , e_{\min} будем обозначать $\mathbb{F}_0(\beta, p, e_{\min})$ либо кратко \mathbb{F}_0 .

Введением денормализации мы сразу решаем две проблемы: получаем свойство

$$x = y \Leftrightarrow R(x - y) = 0 \quad \text{для любых ЧПТ } x, y,$$



Рисунок 1.2

а также добавляем нуль ко множеству машинных чисел. На рисунке 1.2 синим цветом обозначены денормализованные числа, соответствующие [ранее рассмотренному](#) множеству ЧПТ $\mathbb{F}_1(2, 3, -1, 2)$.

Заметим, что для хранения денормализованных ЧПТ необходимо одно дополнительное значение для экспоненты (как правило это $e_{\min} - 1$).

Специальные величины

Стандарт IEEE 754, которому соответствуют практически все современные ЭВМ, предусматривает наличие специальных значений для машинных чисел, которым соответствуют не ЧПТ, а другие объекты. Простейшие объекты такого типа — это $+\infty$ и $-\infty$ (присутствуют также $+0$ и -0). Результаты вычислений с бесконечностями являются вполне определёнными: например, если x — положительное число, то по стандарту $x/\pm\infty = \pm 0$, $x/\pm 0 = \pm\infty$ и т. д. Кроме этого, стандартом определяются так называемые не-числа (NaN-ы, от «not a number»), которые обозначают результаты некорректных арифметических операций, таких как, например, извлечение корня из отрицательного числа.

Определение машинной арифметики

Определение. *Машинными числами* будем называть элементы множества

$$M = \mathbb{F}_0 \cup \mathbb{F}_1.$$

Определение. *Машинной арифметикой с плавающей точкой* (МАПТ) будем называть множество *машинных чисел* M в совокупности с [правилом округления](#) R .

При вычислениях в МАПТ будем считать, что результаты операций сложения, вычитания, умножения и деления являются *точно округляемыми*. Это означает, что результат указанных операций всегда вычисляется точно, после чего округляется до ЧПТ по правилу R .



Меню

1.1.5. Машинный эпсилон

Параметры β , p , e_{\min} , e_{\max} и R полностью определяют свойства МАПТ, однако их знание не даёт прямой информации о том, насколько хороша или плоха соответствующая арифметика. С практической точки зрения пользователю нужны критерии, по которым можно определить качество МАПТ. Основным показателем качества будем считать точность, с которой арифметика приближает вещественные числа.

Определение. *Абсолютной погрешностью округления* для числа $x \in \mathbb{R}$ в данной МАПТ называется число

$$\Delta(x) = |x - R(x)|, \quad (1.2)$$

а *относительной погрешностью округления* — число

$$\delta(x) = \frac{|x - R(x)|}{|x|} = \frac{\Delta(x)}{|x|}. \quad (1.3)$$

Иногда относительную погрешность измеряют в процентах, умножая её на 100. Важно понимать, что при работе с машинной арифметикой уместнее всего оперировать относительными погрешностями, так как чем больше модуль x , тем больше $\Delta(x)$ (см. свойство 3).

Определение. *Машинным эпсилон* ε_M для МАПТ называется наименьшее положительное число ε , удовлетворяющее условию

$$R(1 + \varepsilon) > 1.$$

Свойства машинного эпсилон

1 (главное свойство). Для всех вещественных x таких, что $\xi_{\min} \leq |x| \leq \xi_{\max}$, где ξ_{\min} и ξ_{\max} — минимальное и максимальное положительное *нормализованное ЧПТ* соответственно, справедливо

$$\delta(x) \leq \varepsilon_M.$$

2. В диапазоне *денормализованных* чисел свойство 1, вообще говоря, не выполняется.



3. Значение машинного эпсилон зависит от [правила округления](#) и от количества бит в [мантиске](#).

4. Ни в коем случае не следует путать машинный эпсилон с машинным нулем.

Таким образом, чем меньше величина ϵ_M , тем точнее вещественные числа приближаются в машинной арифметике.



1.1.6. Стандарт IEEE 754

Международный стандарт *IEEE 754 floating point standard* определяет правила организации [машинной арифметики с плавающей точкой](#). В настоящее время ему соответствует большинство вычислительных машин. В частности, наиболее распространённый тип данных, известный как double precision floating point (тип double в C/C++) по стандарту имеет следующие параметры:

β	p	e_{\min}	e_{\max}	R
2	53	-1022	1023	R_e



1.1.7. Проблемы машинных вычислений

Потеря значимости. Эта проблема возникает при вычитании двух близких чисел, которые не являются точно представимыми в виде ЧПТ.

Приведём пример на модельной арифметике с $\mathbb{F}_1(2, 3, -1, 2)$ и $R = R_u$: пусть $x = 4.51$, $y = 4.49$. Имеем $R(x) = 5$, $R(y) = 4$, и $R(R(x) - R(y)) = 1$, тогда как $x - y = 0.02$. Таким образом мы имеем относительную погрешность вычисления равную 5000%, несмотря на то, что **относительная погрешность округления** для x и y составляет менее 12.5%. Отметим, что сложение этих двух чисел выполняется в данной арифметике точно.

Неассоциативность арифметических операций. При работе с машинными числами всегда следует помнить о том, что порядок операций существенно влияет на результат. Простейший случай — нарушение привычного свойства ассоциативности: если a , b и c — машинные числа, \circ — бинарная операция, то в общем случае $R((a \circ b) \circ c) \neq R(a \circ (b \circ c))$.



Меню

Резюме

Итак, при использовании машинной арифметики вычислитель всегда должен помнить о том, что как только он записывает в память ЭВМ число x , оно автоматически превращается в число $\tilde{x} = R(x)$, которое почти всегда не будет равно x . Кроме того, чем больше модуль x , тем больше может быть разница между x и \tilde{x} (абсолютная погрешность $\Delta(x)$). Относительная же погрешность округления, согласно свойству 1, почти всегда ограничена величиной ε_M .



1.2. Обусловленность задачи

- [1.2.1. Корректные задачи](#)
- [1.2.2. Векторные нормы](#)
- [1.2.3. Число обусловленности](#)
- [1.2.4. Резюме](#)



1.2.1. Корректные задачи

Постоянное присутствие [ошибок округления](#) при работе с [машиной арифметикой](#) предъявляет особые требования к вычислительным алгоритмам и требует дополнительного анализа решаемой задачи. Так как практически все числа представляются в ЭВМ с погрешностью, необходимо знать насколько решение чувствительно к изменениям параметров задачи.

Определение. Задача называется [корректно поставленной](#), или просто [корректной](#), если её решение (а) существует, (б) единственно и (в) непрерывно зависит от начальных данных. Если нарушено хотя бы одно из этих условий, задачу называют [некорректной](#).

Решение некорректных задач на ЭВМ — весьма серьёзная проблема. Если задача некорректна, то наличие малейшей погрешности в начальных данных (которая практически неминуемо произойдёт как только вы запишете эти данные в память) может кардинальным образом исказить решение.

Существует ещё один класс задач, формально являющихся корректными, но решения которых, тем не менее, тоже весьма плохо ведут себя при наличии погрешностей в начальных данных — это так называемые плохо обусловленные задачи. В общих чертах, плохо обусловленной называется задача, которая при маленькой *относительной* погрешности в начальных данных даёт большую *относительную* погрешность в решении. Упор на относительную погрешность делается потому, что, как мы знаем из предыдущего раздела, при округлении вещественного числа x до машинного $R(x)$ [абсолютная погрешность](#) $\Delta(x)$ зависит от величины $|x|$, в то время как относительная погрешность $\delta(x)$ постоянна для данной МАПТ.



Меню

1.2.2. Векторные нормы

В дальнейшем как параметры, так и решения рассматриваемых задач будут векторами пространства \mathbb{R}^n . Для исследования обусловленности задач нужно измерять «величины» этих векторов, для чего используются **векторные нормы**. Мы будем активно пользоваться двумя векторными нормами: **максимум-нормой**

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|, \quad (1.4)$$

и **евклидовой нормой**

$$\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}. \quad (1.5)$$

Обе эти нормы являются частными случаями **p -нормы**, определяемой формулой

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p}.$$

Введём также обозначение

$$\delta(u, v) = \frac{\|u - v\|}{\|u\|},$$

которое можно назвать «относительной нормой разности» или «относительной погрешностью» векторов u и v .



Меню

1.2.3. Число обусловленности

Рассмотрим некоторую функцию

$$f : X \rightarrow Y,$$

положив для простоты $X \subset \mathbb{R}^m$, $Y \subset \mathbb{R}^n$ (хотя в общем случае X и Y могут быть подмножествами любого линейного векторного пространства). Возьмём произвольный вектор $x \in X$ и рассмотрим задачу вычисления $y = f(x)$ в предположении, что данная задача корректна. Пусть $\tilde{x} \in X$ — «возмущённый» вектор начальных параметров, и $\tilde{y} = f(\tilde{x})$ — соответствующее «возмущённое» решение. Наша задача оценить, во сколько раз $\delta(y, \tilde{y})$ может быть больше $\delta(x, \tilde{x})$.

Определение. *Числом обусловленности* задачи вычисления $f(x)$ назовём число

$$\alpha(x) = \sup_{\tilde{x} \in M} \frac{\delta(y, \tilde{y})}{\delta(x, \tilde{x})} = \sup_{\tilde{x} \in M} \left(\frac{\|f(x) - f(\tilde{x})\|}{\|f(x)\|} \cdot \frac{\|x\|}{\|x - \tilde{x}\|} \right),$$

где $M \subset X$ — некоторая проколотая окрестность точки x . Если $\alpha(x)$ велико, задачу называют *плохо обусловленной*.

Итак, по определению имеем

$$\delta(f(x), f(\tilde{x})) \leq \alpha(x) \delta(x, \tilde{x}), \quad \forall \tilde{x} \in M.$$

Это означает, что чем больше $\alpha(x)$, тем больше чувствительность решения задачи к относительной погрешности в начальных условиях.

Замечание 1.1. Естественно, понятие «большое число обусловленности» относительно. Судить о величине обусловленности можно лишь в контексте той [машинной арифметики](#), которая используется для вычислений, а ещё точнее — от величины [машинного эпсилон](#) ϵ_M , так как эта величина ограничивает относительную погрешность округления.

Пример 1.1 (вычисление значения многочлена). Исследовать обусловленность задачи вычисления значения многочлена относительно погрешности, вносимой в коэффициенты.



Решение. Пусть $P(a_0, \dots, a_n, x) = \sum_{i=0}^n a_i x^i$. Рассмотрим задачу вычисления данного многочлена, считая фиксированными x и все коэффициенты a_i кроме какого-то одного — a_k , который и будем считать параметром. Вычислим относительную погрешность решения:

$$\delta(f(a_k), f(\tilde{a}_k)) = \frac{|P(a_0, \dots, a_n, x) - P(a_0, \dots, \tilde{a}_k, \dots, a_n, x)|}{|P(a_0, \dots, a_n, x)|} = \frac{|(a_k - \tilde{a}_k)x^k|}{|\sum_{i=0}^n a_i x^i|} = \alpha(a_k) \delta(a_k, \tilde{a}_k),$$

$$\text{где } \alpha(a_k) = \frac{|a_k x^k|}{|\sum_{i=0}^n a_i x^i|}.$$

Видим, что рассматриваемая задача может быть плохо обусловлена в двух случаях: когда x близко к одному из корней многочлена P , или когда $x \gg 1$ и k достаточно велико. Численный пример:

$$\begin{aligned} p(x) = (x - 2)^{10} &= x^{10} - 20x^9 + 180x^8 - 960x^7 + 3360x^6 - 8064x^5 + \\ &+ 13440x^4 - 15360x^3 + 11520x^2 - 5120x + 1024. \end{aligned}$$

Пусть $x = 3$, $k = 9$. Тогда $f(a_9) = f(-20) = 1$. Изменим коэффициент $a_9 = -20$ на 0.01 (что составляет 0.05%): $\tilde{a}_9 = -19.99$. Тогда получим $f(\tilde{a}_9) = 197.83$, что на 19683% больше $f(a_9)$. \square



1.2.4. Резюме

- Число обусловленности задачи показывает, во сколько раз *относительная погрешность решения* может превышать *относительную погрешность в начальных данных*.
- Величина числа обусловленности, при которой задачу можно считать плохо обусловленной, зависит от параметров используемой машинной арифметики.
- При решении плохо обусловленной задачи обычными методами *нельзя рассчитывать на получение адекватного решения*.



Глава 2

Методы решения СЛАУ

- 2.1. Обусловленность СЛАУ
- 2.2. Метод Гаусса
- 2.3. LU-разложение
- 2.4. Метод квадратного корня
- 2.5. Метод отражений
- 2.6. Метод вращений
- 2.7. Методы спуска
- 2.8. Простейшие итерационные методы
- 2.9. Форматы хранения разреженных матриц



2.1. Обусловленность СЛАУ

[2.1.1. Матричные нормы](#)

[2.1.2. Число обусловленности матрицы](#)



2.1.1. Матричные нормы

Мы приступаем к рассмотрению задачи решения системы линейных алгебраических уравнений (СЛАУ) вида

$$Ax = b, \quad (2.1)$$

где A — невырожденная квадратная матрица размерности n , $x, b \in \mathbb{R}^n$. Прежде, чем приступить к алгоритмам численного решения этой задачи, исследуем её обусловленность.

Как и в случае векторных параметров, нам нужно будет как-то измерять «величину» матрицы A . Делать это мы будем с использованием операторных матричных норм. При работе с матрицами (по крайней мере в контексте линейной алгебры) всегда важно помнить, что любая матрица определяет *линейный оператор*, то есть отображение $A : \mathbb{R}^m \rightarrow \mathbb{R}^n$, которое обладает свойством линейности

$$A(\alpha x + \beta y) = \alpha Ax + \beta Ay, \quad \forall x, y \in \mathbb{R}^n, \alpha, \beta \in \mathbb{R}.$$

Важность такой точки зрения следует хотя бы из того, что так называемое «правило умножения матриц», которые многие считают аксиомой, есть ни что иное, как алгоритм вычисления композиции линейных операторов. В таком контексте вопрос об определении «величины» матрицы сводится к определению нормы соответствующего линейного оператора.

Определение. Нормой линейного оператора A называют число

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\|.$$

Норма оператора полностью определяется векторной нормой. То есть каждая векторная норма порождает (*индуцирует*) соответствующую ей операторную матричную форму (в этом случае говорят также, что матричная норма *подчинена* векторной).

В дальнейшем мы без оговорок будем предполагать, что используемая матричная норма подчинена векторной.



Норма оператора равна максимальному «коэффициенту растяжения»: она показывает, во сколько раз под его действием может увеличиться норма вектора. Поэтому по определению для любой операторной нормы имеем важное свойство

$$\|Ax\| \leq \|A\| \|x\|. \quad (2.2)$$

Напомним как вычисляются матричные нормы, индуцированные векторными нормами $\|\cdot\|_\infty$ и $\|\cdot\|_2$.

- Векторной максимум-нормой $\|\cdot\|_\infty$ индуцируется матричная норма, вычисляемая по правилу

$$\|A\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|. \quad (2.3)$$

Данную норму будем называть (строчной) *максимум-нормой*, или *кубической* матричной нормой.

- Евклидовой векторной нормой $\|\cdot\|_2$ индуцируется матричная норма, вычисляемая по правилу

$$\|A\|_2 = \max_{1 \leq i \leq n} \sqrt{\lambda_i}, \quad (2.4)$$

где $\{\lambda_i\}_{i=1}^n$ — собственные значения матрицы A^*A . Эту норму называют *спектральной* матричной нормой.



Меню

2.1.2. Число обусловленности матрицы

Рассмотрим СЛАУ (2.1). Решение этой системы, очевидно, сводится к вычислению

$$x = A^{-1}b.$$

Исследуем обусловленность этой задачи, считая параметром вектор правой части b . Действуем по общей схеме, описанной в пункте 1.2.3:

$$f(b) = A^{-1}b,$$

$$\begin{aligned} \frac{\delta(f(b), f(\tilde{b}))}{\delta(b, \tilde{b})} &= \frac{\|A^{-1}(b - \tilde{b})\|}{\|A^{-1}b\|} \cdot \frac{\|b\|}{\|b - \tilde{b}\|} \leqslant \frac{\|A^{-1}\| \|b\|}{\|A^{-1}b\|} \leqslant \\ &\leqslant \left[\begin{array}{l} \|b\| = \|AA^{-1}b\| \leqslant \|A\| \|A^{-1}b\| \Rightarrow \\ \Rightarrow \|A^{-1}b\| \geqslant \|A\|^{-1} \|b\| \end{array} \right] \leqslant \|A^{-1}\| \|A\|. \end{aligned} \quad (2.5)$$

Исследование обусловленности относительно погрешностей в матрице A требует более тонкого подхода. Итак, пусть $f(A) = A^{-1}b$, вектор b считаем неизменным. Предположим, что возмущённая матрица \tilde{A} может быть представлена в виде $\tilde{A} = A + \varepsilon H$, где H — некоторая матрица, определяющая «направление» изменения, $0 \leqslant \varepsilon$ — вещественный параметр, контролирующий величину этого изменения. Таким образом имеем

$$\delta(A, \tilde{A}) = \varepsilon \frac{\|H\|}{\|A\|}.$$

Так как \tilde{A} является функцией ε , то решение $\tilde{x} = f(\tilde{A})$ тоже будет зависеть от параметра:

$$(A + \varepsilon H)\tilde{x}(\varepsilon) = b. \quad (2.6)$$

Дифференцируя (2.6) по ε , с учётом $\tilde{x}(0) = x = A^{-1}b$, получаем

$$\tilde{x}'(0) = -A^{-1}Hx.$$



Теперь разложим $\tilde{x}(\varepsilon)$ по формуле Тейлора в точке $\varepsilon = 0$:

$$\tilde{x}(\varepsilon) = x + \varepsilon \tilde{x}'(0) + O(\varepsilon^2) = x - \varepsilon A^{-1} H x + O(\varepsilon^2),$$

откуда

$$\begin{aligned} \delta(f(A), f(\tilde{A})) &= f(x, \tilde{x}(\varepsilon)) \leq \varepsilon \|A^{-1}\| \|H\| + O(\varepsilon^2) = \\ &= \|A\| \|A^{-1}\| \delta(A, \tilde{A}) + O(\varepsilon^2). \end{aligned} \quad (2.7)$$

Сопоставляя результаты (2.5), (2.7) получаем следующее определение.

Определение. Числом обусловленности невырожденной матрицы A называется число

$$\alpha(A) = \|A\| \|A^{-1}\|. \quad (2.8)$$

Если матрица A вырождена, её число обусловленности полагается равным бесконечности.

Замечание 2.1. Несмотря на то, что это определение ассоциировано с матрицей A , необходимо чётко понимать, что речь идёт об обусловленности задачи решения СЛАУ.

Замечание 2.2. Число обусловленности по определению зависит от нормы. В случаях, когда это необходимо, мы будем употреблять говорящие обозначения $\alpha_2(A)$ и $\alpha_\infty(A)$.

Свойства числа обусловленности матрицы

1. $\alpha(A) \geq 1 : 1 = \|A^{-1}A\| \leq \|A\| \|A^{-1}\|.$
2. $\alpha(AB) \leq \alpha(A)\alpha(B) : \|AB\| \|(AB)^{-1}\| \leq \|A\| \|A^{-1}\| \|B\| \|B^{-1}\|.$
3. Если $A = A^*$, то $\alpha_2(A) = \frac{|\lambda_{\max}|}{|\lambda_{\min}|}$, где λ_{\min} и λ_{\max} — минимальное и максимальное по модулю собственные значения матрицы A соответственно.



Замечание 2.3. Отметим, что в общем случае отсутствует прямая связь между величинами собственных значений и числом обусловленности. Например, собственные значения матрицы $A = \begin{pmatrix} 1 & \alpha \\ 0 & 1 \end{pmatrix}$ равны 1, в то время как $\alpha_2(A) \rightarrow \infty$ при $|\alpha| \rightarrow \infty$.

Замечание 2.4. Укажем на одно часто встречающееся заблуждение. Так как $\alpha(A)$ является своеобразным индикатором близости матрицы A к вырожденной, может возникнуть впечатление, что чем меньше определитель, тем больше число обусловленности. На самом же деле такой связи нет: достаточно заметить, что у A и A^{-1} взаимно обратные определители, но одинаковые числа обусловленности.

2.2. Метод Гаусса

- [2.2.1. Базовый метод Гаусса](#)
- [2.2.2. Связь метода Гаусса и LU-разложения](#)
- [2.2.3. Метод Гаусса с выбором главного элемента](#)
- [2.2.4. Матричные уравнения](#)
- [2.2.5. Обращение матрицы и вычисление определителя](#)
- [2.2.6. Метод прогонки](#)



2.2.1. Базовый метод Гаусса

Рассмотрим СЛАУ

$$Ax = b, \quad \det A \neq 0. \quad (2.9)$$

Один из способов решения этой системы заключается в переходе к эквивалентной системе (то есть, к системе с тем же решением) вида

$$Vx = g, \quad (2.10)$$

решение которой легко находится, если, например, V — верхне- или нижнетреугольная матрица. Решение системы (2.10), очевидно, может быть легко получено с помощью так называемой процедуры *обратной подстановки*, или *обратного хода*. Например, для верхнетреугольной матрицы V эта процедура выглядит так:

$$x_i = \frac{1}{v_{ii}} \left(g_i - \sum_{j=i+1}^n v_{ij}x_j \right), \quad i = n, n-1, \dots, 2, 1. \quad (2.11)$$

Переход от (2.9) к (2.10) осуществляется путём последовательного применения к обеим частям (2.9) некоторых линейных преобразований T_k :

$$T_N \dots T_2 T_1 A x = T_N \dots T_2 T_1 b,$$

то есть

$$V = \underbrace{T_N \dots T_2 T_1}_T A, \quad g = Tb.$$

Если в качестве T_k использовать элементарные преобразования, то получим *метод Гаусса*.

Введём следующие обозначения:

\underline{a}_i — i -я строчка матрицы A ,

$[A]_k$ — матрица, составленная из k первых строк и k первых столбцов матрицы A .



Базовый алгоритм метода Гаусса

```

1: for  $k = \overline{1, n-1}$  do      // Прямой ход метода
2:   for  $i = \overline{k+1, n}$  do
3:      $\underline{a}_i \leftarrow \underline{a}_i - \frac{a_{ik}}{a_{kk}} \underline{a}_k$ 
4:      $b_i \leftarrow b_i - \frac{a_{ik}}{a_{kk}} b_k$ 
5:   end for
6: end for
7: for  $k = \overline{n, 1}$  do      // Обратный ход
8:    $x_k \leftarrow \frac{1}{a_{kk}} (b_k - \sum_{j=k+1}^n a_{kj} x_j)$ 
9: end for

```

Этап алгоритма, определяемый строками 2-5, будем называть k -м шагом метода Гаусса. На этом этапе с помощью элементарных преобразований обнуляются элементы k -го столбца, находящиеся ниже главной диагонали. Матрицу системы перед выполнением k -го шага будем обозначать $A^{(k)}$ ($A^{(1)} = A$). Переход от матрицы $A^{(k)}$ к $A^{(k+1)}$ можно представить в виде $A^{(k+1)} = G_k A^{(k)}$, где

$$G_k = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & \alpha_{k+1}^{(k)} & 1 & & \\ & & & \alpha_{k+2}^{(k)} & & \ddots & \\ & & & \vdots & & \ddots & \\ & & & \alpha_n^{(k)} & & & 1 \end{bmatrix}. \quad (2.12)$$

Если M — произвольная квадратная матрица размерности n , то для вычисления $G_k M$ нужно для всех i от



$k + 1$ до n к i -й строке матрицы M прибавить k -ю, умноженную на $\alpha_i^{(k)}$. Согласно алгоритму метода Гаусса (см. строку 3) имеем

$$\alpha_i^{(k)} = -a_{ik}^{(k)} / a_{kk}^{(k)}, \quad (2.13)$$

где $a_{ij}^{(k)}$ — элементы матрицы $A^{(k)}$.

Очевидно, что если хотя бы один из элементов

$$\theta_k = a_{kk}^{(k)}$$

равен нулю, то прямой ход в базовом МГ неосуществим. В дальнейшем элементы θ_k будем называть *главными* или *ведущими*. Если же все главные элементы отличны от нуля, то приведённый алгоритм выполнится успешно.

Теорема 2.1. *Базовый алгоритм метода Гаусса осуществим тогда и только тогда, когда все главные угловые миноры матрицы A не равны нулю: $|[A]_k| \neq 0 \quad \forall k = \overline{1, n}$.*

[[Доказательство](#)]



2.2.2. Связь метода Гаусса и LU-разложения

Определение. *LU-разложением* невырожденной матрицы A называется её представление в виде

$$A = LU,$$

где L — нижнетреугольная матрица с единицами на главной диагонали, U — верхнетреугольная матрица.

Теорема 2.2 (связь метода Гаусса и LU-разложения). *Базовый алгоритм метода Гаусса для СЛАУ (2.9) выполним тогда и только тогда, когда существует LU-разложение матрицы A .*

[[Доказательство](#)]



2.2.3. Метод Гаусса с выбором главного элемента

Для того, чтобы выполнение алгоритма метода Гаусса не обрывалось при $\theta_k = 0$ (и не только), перед каждым шагом метода применяется процедура, называемая *выбором главного элемента*. Суть процедуры: путём перестановки строк или столбцов матрицы $A^{(k)}$ поставить на позицию (k, k) ненулевой элемент. При этом, чтобы не «испортить» структуру матрицы, можно использовать лишь последние $n - k$ строк и $n - k$ столбцов. Существует несколько способов выбора главного элемента.

По столбцу: среди элементов $a_{ik}^{(k)}$ для i от k до n выбирается ведущий элемент $a_{i=k}^{(k)}$, после чего переставляются местами строки k и i^* .

По строке: среди элементов $a_{kj}^{(k)}$ для j от k до n выбирается ведущий элемент $a_{k=j^*}^{(k)}$, после чего переставляются местами столбцы k и j^* .

По матрице: среди элементов $a_{ij}^{(k)}$ для i, j от k до n выбирается ведущий элемент $a_{i=j^*}^{(k)}$, после чего переставляются местами строки k и i^* и столбцы k и j^* .

Рассмотрим следующие вопросы.

- 1) Из каких соображений выбирать главный элемент?
- 2) Какой способ выбора главного элемента лучше?

Для ответа рассмотрим ещё раз матрицу G_k (2.12). Имеем

$$\alpha_\infty(G_k) = (1 + \max_i |\alpha_i^{(k)}|)^2, \quad (2.14)$$

откуда с учётом свойства 2 числа обусловленности имеем

$$\alpha_\infty(A^{(n)}) = \alpha_\infty(\tilde{G}_{n-1} A) \leq \alpha_\infty(A) \prod_{k=1}^{n-1} (1 + \max_i |\alpha_i^{(k)}|)^2.$$

Таким образом, даже если $\alpha(A)$ невелико, матрица $A^{(n)}$ может стать плохо обусловленной в случае больших значений $|\alpha_i^{(k)}|$. То есть, сам процесс метода Гаусса может «испортить» исходную систему.

Для исправления ситуации мы должны минимизировать величины (2.14). С учётом (2.13), получаем следующие ответы.



- 1) Главный элемент должен быть максимальным по модулю среди всех рассматриваемых.
- 2) Выбор главного элемента по столбцу оптimalен по соотношению «качество/скорость».



2.2.4. Матричные уравнения

Метод Гаусса естественным образом обобщается на случай матричных уравнений вида

$$AX = B, \quad (2.15)$$

где A , как и ранее, — квадратная матрица порядка n , B — матрица размеров $n \times m$, X — неизвестная матрица тех же размеров, что и B . Возможно два подхода к решению таких уравнений.

- 1) Система (2.15) эквивалентна набору из m СЛАУ вида

$$Ax_j = b_j, \quad j = \overline{1, m},$$

где x_j и b_j — столбцы матриц X и B . Применять метод Гаусса к каждой такой системе нерационально, поэтому для их решения используется [метод LU-разложения](#).

- 2) Матричный метод Гаусса. Для того, чтобы адаптировать построенный выше алгоритм к решению матричных уравнений, достаточно строку 4 заменить на

$$\underline{b}_i \leftarrow \underline{b}_i - \frac{a_{ik}}{a_{kk}} \underline{b}_k,$$

а также модифицировать алгоритм обратного хода.



2.2.5. Обращение матрицы и вычисление определителя

Обращение матрицы эквивалентно решению матричного уравнения

$$AX = I,$$

где I — единичная матрица. Для решения этого уравнения могут использоваться оба [описанных выше](#) способа.

Определитель матрицы также вычисляется с помощью метода Гаусса:

$$\tilde{G}_{n-1} A = A^{(n)} \Rightarrow 1 \cdot |A| = |A^{(n)}| = a_{11}^{(n)} a_{22}^{(n)} \dots a_{nn}^{(n)}.$$

Однако, нужно помнить про важный нюанс: если в ходе метода переставлялись строки и столбцы, то каждая такая операция меняла знак определителя на противоположный. Поэтому окончательная формула такова:

$$|A| = (-1)^p a_{11}^{(n)} a_{22}^{(n)} \dots a_{nn}^{(n)}, \quad (2.16)$$

где p — количество перестановок строк и столбцов в ходе метода.



2.2.6. Метод прогонки

Рассмотрим СЛАУ

$$\begin{bmatrix} d_1 & e_1 & & & \\ c_2 & d_2 & e_2 & & \\ c_3 & d_3 & e_3 & \ddots & \\ \ddots & \ddots & \ddots & \ddots & \\ c_{n-1} & d_{n-1} & e_{n-1} & & \\ c_n & d_n & & & \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_{n-1} \\ b_n \end{bmatrix}. \quad (2.17)$$

Матрицы такой структуры называются *трёхдиагональными*. В приложениях достаточно часто встречаются такие системы. Особая структура этой матрицы позволяет найти решение системы методом Гаусса за $O(n)$ операций. Получаемый метод называется [методом прогонки](#).

Алгоритм метода прогонки

```

1: for  $k = \overline{2, n}$  do      // Прямой ход
2:    $d_k \leftarrow d_k - e_{k-1}c_k/d_{k-1}$ 
3:    $b_k \leftarrow b_k - b_{k-1}c_k/d_{k-1}$ 
4: end for
5:  $x_n = b_n/d_n$ 
6: for  $k = \overline{n-1, 1}$  do      // Обратный ход
7:    $x_k \leftarrow (b_k - e_kx_{k+1})/d_k$ 
8: end for

```

При выполнении алгоритма прогонки мы лишены возможности выбора главного элемента, так как при этом нарушилась бы трёхдиагональная структура матрицы A . Следовательно, метод прогонки осуществим тогда и только когда, когда все главные миноры матрицы отличны от нуля. Существует также более простое для проверки достаточное условие осуществимости метода прогонки. Для его доказательства нам понадобятся следующие предварительные сведения.



Определение. *Кругом Гершгорина* для квадратной матрицы A называется замкнутый круг D_i на комплексной плоскости с центром в точке a_{ii} и радиусом

$$\rho_i = \sum_{j \neq i} |a_{ij}|.$$

Теорема 2.3 (Гершгорина). *Каждое собственное значение матрицы A лежит в одном из кругов Гершгорина.*

Определение. Если элементы матрицы A удовлетворяют условиям

$$|a_{ii}| \geq \sum_{j \neq i} |a_{ij}| \quad \forall i = \overline{1, n}, \tag{2.18}$$

то говорят, что такая матрица обладает свойством *диагонального преобладания*. Если неравенство в (2.18) строгое, говорят о *строгом диагональном преобладании*.

Теорема 2.4. *Если матрица обладает свойством строгого диагонального преобладания, то все её главные миноры отличны от нуля.* [\[Доказательство\]](#)

Следствие 2.1. *Если матрица системы (2.17) удовлетворяет условиям*

$$|d_1| > |e_1|, \quad |d_i| > |c_i| + |e_i| \quad \forall i = \overline{2, n-1}, \quad \text{и} \quad |d_n| > |c_n|,$$

то алгоритм прогонки выполним.



2.3. LU-разложение

[2.3.1. Базовый алгоритм LU-разложения](#)

[2.3.2. Выбор главного элемента](#)



2.3.1. Базовый алгоритм LU-разложения

Рассмотрим последовательность СЛАУ

$$Ax = b^{(i)}, \quad i = \overline{1, N}. \quad (2.19)$$

и предположим, что векторы $b^{(i)}$ неизвестны заранее и поступают *по одному*, то есть мы не можем свести (2.19) к матричному уравнению. При решении каждой такой СЛАУ методом Гаусса будет тратиться $O(n^3)$ операций, причём к матрице A будут применяться *одни и те же* преобразования G_k .

Поэтому разумнее, однажды проделав прямой ход (или его аналог), построить LU-разложение $A = LU$, и в дальнейшем вычислять x путём решения двух СЛАУ с треугольными матрицами:

$$LUx = b \Leftrightarrow \begin{cases} Ly = b, \\ Ux = y. \end{cases} \quad (2.20)$$

Рассмотрим алгоритм построения LU-разложения в предположении, что $|[A]_k| \neq 0 \forall k = \overline{1, n}$. По определению имеем

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \ell_{21} & 1 & 0 & \cdots & 0 \\ \ell_{31} & \ell_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \cdots & 1 \end{bmatrix}}_L \underbrace{\begin{bmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ 0 & u_{22} & u_{23} & \cdots & u_{2n} \\ 0 & 0 & u_{33} & \cdots & u_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & u_{nn} \end{bmatrix}}_U = \underbrace{\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{bmatrix}}_A. \quad (2.21)$$



При машинной реализации алгоритма матрицы L и U будем хранить на месте матрицы A :

$$A \leftarrow \tilde{A} = \underbrace{\begin{bmatrix} u_{11} & u_{12} & u_{13} & \dots & u_{1n} \\ \ell_{21} & u_{22} & u_{23} & \dots & u_{2n} \\ \ell_{31} & \ell_{32} & u_{33} & \dots & u_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \dots & u_{nn} \end{bmatrix}}_{L-I+U}, \quad \text{т. е. } \tilde{a}_{ij} = \begin{cases} u_{ij}, & i \leq j, \\ \ell_{ij}, & i > j. \end{cases}$$

Из (2.21) имеем

$$a_{ij} = \sum_{k=1}^n \ell_{ik} u_{kj} = \sum_{k=1}^{\min(i,j)} \ell_{ik} u_{kj}. \quad (2.22)$$

Выделяя последние слагаемые в суммах (2.22), получаем

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} \ell_{ik} u_{kj} \quad \text{при } i \leq j; \quad (2.23)$$

$$\ell_{ij} = \frac{1}{u_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} \ell_{ik} u_{kj} \right) \quad \text{при } i > j, \quad (2.24)$$

или

$$\tilde{a}_{ij} = \begin{cases} a_{ij} - \sum_{k=1}^{i-1} \tilde{a}_{ik} \tilde{a}_{kj} & \text{при } i \leq j; \\ \frac{1}{\tilde{a}_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} \tilde{a}_{ik} \tilde{a}_{kj} \right) & \text{при } i > j. \end{cases} \quad (2.25)$$

Таким образом, неизвестные элементы матриц L и U последовательно выражаются через a_{ij} и уже найденные ℓ_{ik} и u_{kj} .



Базовый алгоритм LU-разложения

```
1: for  $j = \overline{1, n - 1}$  do
2:   for  $i = \overline{1, n}$  do
3:      $a_{ij} \leftarrow a_{ij} - \sum_{k=1}^{\min(i,j)-1} a_{ik} a_{kj}$ 
4:     if  $i > j$  then
5:        $a_{ij} \leftarrow a_{ij} / a_{jj}$ 
6:     end if
7:   end for
8: end for
```



2.3.2. Выбор главного элемента

По аналогии с методом Гаусса, этап алгоритма *LU*-разложения, определяемый циклом в строках 2-7 будем называть j -м шагом *LU*-разложения. Для того, чтобы алгоритм был универсальным, необходимо реализовать выбор главного элемента $\tilde{a}_{jj} = u_{jj}$, на который происходит деление в строке 5.

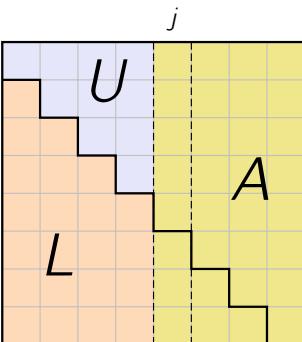


Рисунок 2.1 Вид матрицы системы перед j -м шагом алгоритма *LU*-разложения.

Рассмотрим матрицу $A^{(j)}$, которая получается из A после $(j - 1)$ -го шага разложения (рис. 2.1). К этому моменту столбцы с 1-го по $(j - 1)$ -й уже содержат часть матриц L и U , а оставшиеся столбцы являются столбцами исходной матрицы A . Имеем ли мы право переставлять в этой «составной» матрице строки и если да, то какие? Строки с 1 по $(j - 1)$ -ю переставлять нельзя, иначе нарушится структура матрицы \tilde{A} .

Перестановка же строк с j по n эквивалентна преобразованию

$$LU = A \quad \mapsto \quad PLU = PA,$$

где P — матрица перестановки.

Итак, на j -м шаге алгоритма мы имеем право переставлять строки с номерами от j до n . Поэтому элементы u_{ij} для i от 1 до $j - 1$, вычисляемые по формуле (2.23), можно найти сразу. После этого нужно



осуществить перестановку строк, но проблема в том, что ведущий элемент u_{jj} *ещё неизвестен*, как неизвестны и возможные кандидаты на его место.

Поэтому перестановка должна быть выполнена таким образом, чтобы элемент $a_{jj}^{(j)} = u_{jj}$, вычисляемый по формуле

$$u_{jj} = a_{jj} - \sum_{k=1}^{j-1} \ell_{kj} u_{kj} \quad (2.26)$$

был максимальным по модулю. Заметим, что элементы ℓ_{ij} вычисляются по формуле (2.24), которая при $i = j$ отличается от (2.26) только множителем $1/u_{jj}$. Поэтому выбор главного элемента на j -ом шаге LU-разложения осуществляется следующим образом.

- 1) Вычисляем кандидатов на роль ведущего элемента: для всех i от j до n вычисляем $a_{ij}^{(j)} = \tilde{a}_{ij}$ по второй формуле из (2.25), только без деления на \tilde{a}_{jj} ;
- 2) Среди полученных значений $a_{ij}^{(j)}$ для $i \geq j$ выбираем максимальный по модулю $a_{i^*j}^{(j)}$;
- 3) Меняем местами j -ю и i^* -ю строки матрицы $A^{(j)}$;
- 4) Для всех i от $j + 1$ до n делим $a_{ij}^{(j)}$ на $a_{jj}^{(j)}$.

Замечание 2.5. Для того, чтобы после получения LU-разложения корректно решить СЛАУ (2.20), необходимо предварительно переставить элементы вектора b в соответствии с перестановками, которые происходили в ходе разложения. Поэтому стандартная процедура должна возвращать не только матрицу \tilde{A} , но и вектор перестановок p . Кроме этого, для корректного вычисления определителя необходимо возвращать $s = \pm 1$ — значение чётности числа перестановок.



2.4. Метод квадратного корня

[2.4.1. Разложение Холецкого](#)

[2.4.2. Алгоритм метода](#)



Меню

2.4.1. Разложение Холецкого

Теорема 2.5 (разложение Холецкого). Пусть A — самосопряжённая матрица над полем \mathbb{C} : $A = A^*$. Если все главные миноры $|[A_k]|$ отличны от нуля, то существует разложение

$$A = R^*DR, \quad (2.27)$$

где R — верхнетреугольная матрица, $D = \text{diag}(d_1, d_2, \dots, d_n)$, $|d_k| = 1 \forall k = \overline{1, n}$. Формула (2.27) называется разложением Холецкого (Cholesky).

Доказательство. Так как $|[A]_k| \neq 0$, по теореме 2.2 существует LU -разложение

$$A = LU = U^*L^* = A^*,$$

откуда $L = U^*(L^*U^{-1}) = U^*H$, и

$$A = LU = U^*HU. \quad (2.28)$$

Рассмотрим матрицу $H = L^*U^{-1}$. С одной стороны H — верхнетреугольная, так как является произведением верхнетреугольных матриц L^* и U^{-1} . С другой стороны $H = (U^*)^{-1}L$, то есть H является ещё и нижнетреугольной. Следовательно,

$$H = \text{diag}(h_1, h_2, \dots, h_n).$$

Положим $d_k = h_k/|h_k|$, $\tilde{H} = \text{diag}(\sqrt{|h_1|}, \sqrt{|h_2|}, \dots, \sqrt{|h_n|})$. Тогда (2.28) даёт

$$A = U^*HU = U^*(\tilde{H}^*D\tilde{H})U = (\tilde{H}U)^*D(\tilde{H}U) = R^*DR,$$

что и требовалось доказать. □

Определение. Квадратная матрица A над полем \mathbb{R} (\mathbb{C}) называется *положительно определённой* ($A > 0$), если

$$(Ax, x) > 0 \quad \forall x \in \mathbb{R}^n (\mathbb{C}^n), x \neq 0.$$

В комплексном случае мы подразумеваем, что все (Ax, x) вещественны.



Свойства положительно определённых матриц

1. Если $A > 0$ то $|(A)_k| \neq 0 \forall k = \overline{1, n}$.
2. Если $A^* = A$, то $A > 0 \Leftrightarrow$ все собственные значения A вещественны и положительны.

Теорема 2.6. Если $A = A^*$ и $A > 0$, то существует разложение

$$A = R^* R,$$

где R — верхнетреугольная матрица.

Доказательство. Согласно свойству 1 для матрицы A существует разложение (2.27). Значит, нам достаточно показать, что матрица D — единичная. По условию имеем

$$(Ax, x) = (R^* DRx, x) = (DRx, Rx) > 0 \quad \forall x \neq 0.$$

Так как R невырождена, $\forall y \in \mathbb{C}^n \exists x : y = Rx$, то есть

$$(Dy, y) = \sum_{i=1}^n d_i y_i \bar{y}_i > 0 \quad \forall y \neq 0. \tag{2.29}$$

Возьмём в качестве y k -й единичный орт: $y_i = \delta_{ik}$. Тогда с учётом того, что $|d_k| = 1$, из (2.29) получаем $d_k = 1 \forall k = \overline{1, n}$. \square

Таким образом, в случае вещественной матрицы A теоремы 2.5 и 2.6 влекут следующие утверждения: если A симметрична ($A = A^T$) и все $|(A)_k| \neq 0$, то существует разложение вида

$$A = R^T DR, \tag{2.30}$$

где R — вещественная верхнетреугольная, D — диагональная матрица с элементами ± 1 на диагонали. Если к тому же $A > 0$, то $D = I$.



2.4.2. Алгоритм метода

Методом квадратного корня называется метод решения вещественной СЛАУ $Ax = b$ с симметричной матрицей A путём построения разложения Холецкого

$$A = R^T DR.$$

Обозначим $L = R^T$, $U = DR$. Тогда аналогично методу LU -разложения имеем

$$a_{ij} = \sum_{k=1}^{\min(i,j)} \ell_{ik} u_{kj} = \begin{bmatrix} \ell_{ik} = r_{ki}, \\ u_{kj} = d_k r_{kj} \end{bmatrix} = \sum_{k=1}^{\min(i,j)} d_k r_{ki} r_{kj}. \quad (2.31)$$

В силу симметрии достаточно рассмотреть (2.31) только для верхнего треугольника матрицы A ($i \leq j$):

$$i = j : \quad d_i r_{ii}^2 = a_{ii} - \sum_{k=1}^{i-1} d_k r_{ki}^2 = \omega_i \Rightarrow \begin{cases} d_i = \text{sign } \omega_i, \\ r_{ii} = \sqrt{|\omega_i|}; \end{cases} \quad (2.32a)$$

$$i < j : \quad r_{ij} = \frac{1}{d_i r_{ii}} \left(a_{ij} - \sum_{k=1}^{i-1} d_k r_{ki} r_{kj} \right). \quad (2.32b)$$

Детали программной реализации

- 1) Так как матрица A симметрична, достаточно хранить в памяти только её верхний треугольник.
- 2) Аналогично случаю LU -разложения расчётные формулы (2.32) позволяют последовательно (построчно) находить элементы матрицы R и хранить их на месте исходной матрицы: $A \leftarrow R$.
- 3) Решение получаемой в итоге СЛАУ $R^T DRx = b$ осуществляется путём применения двух обратных подстановок: $R^T y = b$, затем $Rx = Dy$ (так как $D = D^{-1}$).



Резюме

- Метод LU -разложения фактически представляет собой «законсервированный» метод Гаусса.
- Особенно удобен этот метод при решении большого количества СЛАУ с одной и той же матрицей A .
- При реализации LU -разложения кроме матриц L и U необходимо возвращать ещё и вектор перестановок, чтобы впоследствии корректно решить СЛАУ.
- Метод квадратного корня — частный случай LU -разложения для симметричных матриц.



2.5. Метод отражений

[2.5.1. Система из двух уравнений](#)

[2.5.2. Общая схема метода отражений](#)

[2.5.3. QR-разложение](#)

Недостатком метода Гаусса и его модификаций является то, что элементарные преобразования G_k в общем случае ухудшают обусловленность исходной системы ($\varpi_\infty(G_k) > 1$). Поэтому разработаны альтернативные методы приведения матрицы A к треугольному виду, основанные на ортогональных преобразованиях.



2.5.1. Система из двух уравнений

Рассмотрим СЛАУ

$$Ax = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}. \quad (2.33)$$

Для её решения применим к обеим частям ортогональное линейное преобразование T , которое «обнуляет» элемент a_{21} :

$$TAx = A'x = \begin{bmatrix} a'_{11} & a'_{12} \\ 0 & a'_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b'_1 \\ b'_2 \end{bmatrix} = Tb. \quad (5.1')$$

Пусть a_j — j -й столбец матрицы A , тогда $a'_j = Ta_j$. Так как ортогональные преобразования сохраняют евклидову норму векторов, имеем $\|a_j\| = \|a'_j\|$ (в дальнейшем $\|\cdot\| = \|\cdot\|_2$), поэтому $a'_{11} = \pm\|a_1\|$.

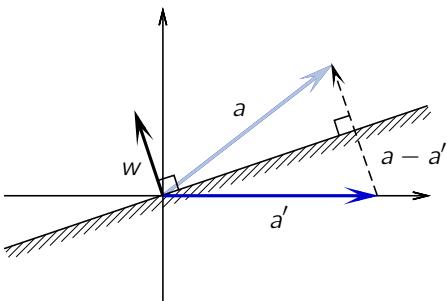


Рисунок 2.2 Преобразование отражения на плоскости.

Рассмотрим вектор $a = a_1$, под действием T переходящий в $a' = a'_1$. Наша задача — определить T как преобразование отражения относительно какой-то гиперплоскости, что равносильно нахождению вектора нормали w к этой гиперплоскости, $\|w\| = 1$. Нетрудно заметить, что векторы $a - a'$ и w коллинеарны, следовательно



$$w = \pm \frac{a - a'}{\|a - a'\|}. \quad (2.34)$$

Теперь рассмотрим обратную задачу: по данным векторам a и w найти a' . Из (2.34) имеем $\pm\|a - a'\|w = a - a' \Rightarrow a' = a - \gamma w$, $\gamma \in \mathbb{R}$. Так как $(a + a') \perp w$, $(a + a', w) = (2a - \gamma w, w) = 0 \Rightarrow \gamma = 2(a, w)$, откуда в итоге получаем

$$a' = a - 2(a, w)w. \quad (2.35)$$

Таким образом, в случае системы (2.33) прямой ход метода отражений состоит в следующем:

- 1) По формуле (2.34) найти вектор нормали w , определяющий гиперплоскость, при отражении относительно которой вектор a_1 переходит в $a'_1 = (\pm\|a_1\|, 0)^T$.
- 2) По формуле (2.35) применить найденное преобразование отражения к векторам a_2 и b .



2.5.2. Общая схема метода отражений

В случае СЛАУ произвольной размерности, [метод отражений](#), как и в [двумерном случае](#), определяется формулами (2.34), (2.35). Алгоритм метода имеет следующий вид (напомним, что a_j обозначает j -й вектор-столбец матрицы A).

- 1) Вычислить w по формуле (2.34), где $a = a_1$, $a' = a'_1 = (\pm \|a_1\|, 0, \dots, 0)^T$.
- 2) $a_1 \leftarrow a'_1$; вычислить $a_j \leftarrow a'_j \quad \forall j \neq 1$, $b \leftarrow b'$ по формуле (2.35).
- 3) Повторить шаги 1-2 для нижней правой подматрицы A и соответствующего подвектора b размерности $n - 1$, и так далее до тех пор, пока матрица A не станет верхнетреугольной.

Замечание 2.6. При реализации метода отражений существует свобода выбора знака для вектора a'_1 . Чтобы избежать вычитания близких чисел при вычислении w по формуле (2.34), этот знак выбирают таким образом, чтобы он был [противоположен](#) знаку a_{11} (a_{kk} в общем случае). Тогда при вычислении $a_1 - a'_1$ фактически будут складываться два одинаковых по модулю числа.



2.5.3. QR-разложение

Найдём в явном виде матрицу преобразования отражения, задаваемого формулой (2.35).

$$a' = a - 2(a, w)w = Ia - 2w(w^T a) = Ia - 2(ww^T)a = (I - 2ww^T)a.$$

Определение. Пусть $w \in \mathbb{R}^n(\mathbb{C}^n)$, $\|w\|_2 = 1$. Матрица

$$H = H(w) = I - 2ww^T$$

называется *матрицей отражения*. Она задаёт преобразование отражения относительно гиперплоскости с нормалью w .

Свойства матрицы отражения

1. $H(w) = H(w)^{-1}$.
2. Матрица отражения является симметричной (самосопряжённой).
3. Матрица отражения является ортогональной (унитарной).
4. Все собственные значения матрицы отражения равны ± 1 .

Доказательство. Так как матрица H унитарна, имеем $\|H\| = 1$. Пусть λ — произвольное собственное значение H , x — соответствующий собственный вектор: $Hx = \lambda x$. Тогда $\|Hx\| = \|x\| = \|\lambda x\|$, откуда $|\lambda| = 1$.

Докажем теперь, что собственные числа самосопряжённой матрицы вещественны:

$$H = H^* \Rightarrow (Hx, x) = (x, H^*x) = (x, Hx) = \overline{(Hx, x)} \Rightarrow (Hx, x) \in \mathbb{R} \quad \forall x \in \mathbb{C}^n.$$

Далее пусть x — собственный вектор. Тогда

$$(Hx, x) = \lambda(x, x) = \lambda\|x\|^2 \Rightarrow \lambda = \frac{(Hx, x)}{\|x\|^2} \in \mathbb{R}.$$

Таким образом, $|\lambda| = 1$ и $\lambda \in \mathbb{R}$, то есть $\lambda = \pm 1$. □



Меню

Замечание 2.7. Заметим, что вычисление преобразования отражения по формуле (2.35) требует $O(n)$ операций умножения и сложения, в то время как умножение на матрицу $H(w)$ требует $O(n^2)$ операций. Поэтому в явном виде $H(w)$ практически никогда не используют, а хранят только w .

Процесс преобразования A к верхнетреугольному виду по методу отражений можно представить в виде

$$A^{(k+1)} = Q_k A^{(k)},$$

где Q_k — блочная матрица вида

$$Q_k = \begin{bmatrix} I_{k-1} & 0 \\ 0 & H(w_k) \end{bmatrix}, \quad (2.36)$$

I_{k-1} — единичная матрица размерности $k - 1$, w_k — вектор нормали размерности $n - k + 1$.

Таким образом мы имеем

$$Q_{n-1} \dots Q_2 Q_1 A = \tilde{Q} A = A^{(n)},$$

откуда

$$A = QR, \quad \text{где } Q = \tilde{Q}^{-1}, R = A^{(n)}. \quad (2.37)$$

Определение. *QR-разложением* матрицы A называется её представление в виде

$$A = QR,$$

где Q — ортогональная ($Q^{-1} = Q^T$), а R — верхнетреугольная матрица.

Теорема 2.7 (о QR-разложении). Для любой вещественной квадратной матрицы A существует QR-разложение: $A = QR$, где Q — ортогональная, R — верхнетреугольная матрица с неотрицательными диагональными элементами. Если $\det A \neq 0$, то все диагональные элементы R положительны. [Доказательство]

Рассмотрим подробно алгоритм построения QR-разложения методом отражений. Из (2.37) имеем

$$Q = \tilde{Q}^{-1} = (Q_{n-1} \dots Q_1)^{-1} = Q_1 \dots Q_{n-1},$$

где Q_k определяются формулой (2.36) (здесь мы использовали тот факт, что $Q_k^{-1} = Q_k$).



Согласно замечанию 2.7 вместо того, чтобы хранить отдельно матрицу Q , мы можем хранить лишь векторы w_k , которые однозначно определяют Q . Кроме того, так как $w_k \in \mathbb{R}^{n-k+1}$, можно их хранить на месте нижнего треугольника матрицы A (для этого необходимо завести отдельный вектор для хранения диагональных элементов a_{kk}).

Предположим, что нам известно QR-разложение матрицы A . Тогда решение СЛАУ $Ax = b$ осуществляется за $O(n^2)$ операций:

$$QRx = b \Leftrightarrow \begin{cases} Qy = b, \\ Rx = y. \end{cases}$$

Таким образом, сначала вычисляется $y = Q^{-1}b = Q^T b$, а затем обратной подстановкой находится x .

Если матрица Q хранится в виде набора нормалей w_k , то вычисление вектора

$$y = Q^{-1}b = Q_{n-1} \dots Q_1 b$$

эквивалентно последовательному применению к вектору b всех преобразований отражения, использованных при построении QR-разложения.



2.6. Метод вращений

- [2.6.1. Система из двух уравнений](#)
- [2.6.2. Общий случай](#)



2.6.1. Система из двух уравнений

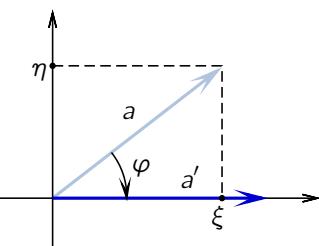


Рисунок 2.3 Преобразование вращения на плоскости.

Рассмотрим снова систему (2.33) и вектор $a = a_1$. Переход к системе (5.1') будем теперь осуществлять с помощью преобразования вращения. Угол φ необходимо выбрать так, чтобы при вращении у вектора a обнулилась вторая координата. Координаты вектора a обозначим ξ и η . С помощью перехода к полярным координатам нетрудно показать, что **матрица вращения** на угол α против часовой стрелки имеет вид

$$V = V(\alpha) = \begin{bmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{bmatrix}. \quad (2.38)$$

В нашем случае матрица преобразования $T : A \mapsto A'$ равна $V(-\varphi)$, где φ — угол между вектором a и осью x_1 :

$$T = \begin{pmatrix} c & s \\ -s & c \end{pmatrix}, \quad (2.39)$$

$$c = \frac{\xi}{\sqrt{\xi^2 + \eta^2}}, \quad s = \frac{\eta}{\sqrt{\xi^2 + \eta^2}}. \quad (2.40)$$

Итак, **метод вращений** для системы (2.33) состоит в следующем:



1) Найти $c = \cos \varphi$ и $s = \sin \varphi$ по формуле (2.40), положив $\xi = a_{11}$, $\eta = a_{21}$.

2) $a_1 \leftarrow a'_1 = (\|a_1\|, 0)^T$.

3) $a_2 \leftarrow a'_2 = Ta$, $b \leftarrow b' = Tb$, где T — матрица (2.39).



2.6.2. Общий случай

В общем случае СЛАУ из n уравнений основное отличие метода вращений от [метода отражений](#) заключается в том, что для выполнения k -го шага метода нам необходимо выполнить $n - k$ операций вращения (а не одно преобразование, как в методе отражений), по одной на каждую обнуляемую координату.

Рассмотрим матрицу

$$V_{ki} = \begin{bmatrix} & & k & & i \\ 1 & & & & \\ \dots & & & & \\ & 1 & & & \\ \hline & & c_{ki} & & s_{ki} \\ \hline & & \dots & & \\ & -s_{ki} & & c_{ki} & \\ \hline & & & & 1 \\ & & & & \dots \\ & & & & 1 \end{bmatrix} \quad \begin{array}{l} k \\ i \end{array} \quad \begin{array}{l} c_{ki} = \cos \varphi_{ki}, \\ s_{ki} = \sin \varphi_{ki}. \end{array} \quad (2.41)$$

Это [матрица элементарного вращения](#) в координатной плоскости (x_k, x_i) на угол φ_{ki} по часовой стрелке.

Рассмотрим k -й шаг метода вращений (обнуление всех элементов a_{ik} для $i = \overline{k+1, n}$). Он состоит из $n - k$ частей, каждая из которых соответствует умножению на матрицы вида (2.41):

$$Q_k = V_{kn} V_{k,n-1} \dots V_{k,k+1}. \quad (2.42)$$



Например, для системы из трёх уравнений мы будем иметь

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & c_{23} & s_{23} \\ 0 & -s_{23} & c_{23} \end{bmatrix}}_{V_{23}} \underbrace{\begin{bmatrix} c_{13} & 0 & s_{13} \\ 0 & 1 & 0 \\ -s_{13} & 0 & c_{13} \end{bmatrix}}_{V_{13}} \underbrace{\begin{bmatrix} c_{12} & s_{12} & 0 \\ -s_{12} & c_{12} & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{V_{12}} \underbrace{\begin{bmatrix} \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \end{bmatrix}}_A = \underbrace{\begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \end{bmatrix}}_{A^{(3)}}.$$

Элементы c_{ki} и s_{ki} вычисляются по формулам (2.40), где $\xi = a_{kk}$, $\eta = a_{ik}$ — текущие (а не исходные, конечно) элементы матрицы A .

Заметим, что умножение на матрицу V_{ki} изменяет в произвольном $x \in \mathbb{R}^n$ только k -й и i -й элементы:

$$V_{ki} \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ \vdots \\ x_i \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ c x_k + s x_i \\ \vdots \\ -s x_k + c x_i \\ \vdots \\ x_n \end{bmatrix} \quad (c = c_{ki}, s = s_{ki}).$$

Таким образом, i -я ($(i-k)$ -я по счёту) часть k -го шага метода вращений описывается следующим алгоритмом.

Основная часть алгоритма метода вращений

- 1: $d \leftarrow \sqrt{a_{kk}^2 + a_{ik}^2}; \quad c \leftarrow a_{kk}/d; \quad s \leftarrow a_{ik}/d;$
- 2: $a_{kk} \leftarrow d; \quad a_{ik} \leftarrow 0;$
- 3: **for** $j = k+1, n$ **do**
- 4: $tmp \leftarrow c a_{kj} + s a_{ij};$
- 5: $a_{ij} \leftarrow -s a_{kj} + c a_{ij};$



6: $a_{kj} \leftarrow tmp;$

7: end for

С помощью метода вращений, очевидно, можно строить QR-разложение матрицы A : k -й шаг метода соответствует умножению на ортогональную матрицу (2.42), поэтому аналогично методу отражений имеем

$$A = (Q_{n-1} \dots Q_1)^{-1} A^{(n)} = QR.$$



2.7. Методы спуска

- [2.7.1. Метод спуска общего вида](#)
- [2.7.2. Метод градиентного спуска](#)
- [2.7.3. Метод сопряжённых градиентов](#)



2.7.1. Метод спуска общего вида

Рассмотрим СЛАУ $Ax = b$ с симметричной и [положительно определённой](#) матрицей A . Семейство итерационных методов, называемых [методами спуска](#), строится путём перехода от задачи решения СЛАУ к задаче минимизации функционала

$$\Phi(x) = \frac{1}{2}(Ax, x) - (x, b) \rightarrow \min. \quad (2.43)$$

Минимальное значение $\Phi(x)$ достигается при

$$x = x^* = A^{-1}b$$

и равно

$$\Phi(x^*) = \varphi_0 = -\frac{1}{2}(A^{-1}b, b). \quad (2.44)$$

Следовательно, решение СЛАУ $Ax = b$ можно найти, решив эквивалентную задачу нахождения $x^* \in \mathbb{R}^n$ такого, что

$$\Phi(x^*) = \min_{x \in \mathbb{R}^n} \Phi(x) = \varphi_0.$$

Пусть x — текущее приближение к решению задачи (2.43). Построим правило (отображение) $x \rightarrow \hat{x}$, которое позволит от x перейти к уточнённому значению \hat{x} .

Выберем какой-то вектор $p \in \mathbb{R}^n$, который будем называть [направлением спуска](#). Наша задача — продвинуться из точки x в направлении p таким образом, чтобы как можно сильнее уменьшить значение функционала Φ . То есть, необходимо найти $\alpha \in \mathbb{R}$ такое, что

$$\Phi(\hat{x}) = \Phi(x + \alpha p) = \min_{\beta \in \mathbb{R}} \Phi(x + \beta p).$$

Из (2.43) с учётом того, что $(x, Ap) = (Ax, p)$ в силу симметричности A , получаем

$$\Phi(x + \alpha p) = \Phi(x) + \frac{1}{2}\alpha^2(p, Ap) - \alpha(b - Ax, p).$$



Обозначим невязку

$$b - Ax = r. \quad (2.45)$$

Тогда из условия $\frac{d}{d\alpha}\Phi(x + \alpha p) = 0$ находим

$$\alpha = \frac{(p, r)}{(Ap, p)}. \quad (2.46)$$

Таким образом, один шаг метода спуска имеет следующий общий вид.

- 1) Выбрать направление спуска p .
- 2) Вычислить $\hat{x} = x + \alpha p$, где α определяется по формуле (2.46).

Все рассматриваемые далее методы относятся к методам спуска и отличаются лишь способом выбора направления p .



2.7.2. Метод градиентного спуска

Определение. Напомним, что градиентом функции n переменных $f : \mathbb{R}^n \rightarrow \mathbb{R}$ называется вектор-функция

$$\operatorname{grad} f = \nabla f : \mathbb{R}^n \rightarrow \mathbb{R}^n,$$

определенная формулой

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n} \right).$$

Как известно, основное свойство вектора $\nabla f(x)$ состоит в том, что он показывает направление и величину наибольшей скорости возрастания f в точке x . Поэтому в *методе градиентного спуска* (МГС) в качестве *направления спуска* выбирается вектор

$$p = -\nabla \Phi(x).$$

Нетрудно показать, что

$$\nabla \Phi(x) = Ax - b = -r.$$

Таким образом, МГС это метод спуска с выбором $p = -r$.

Алгоритм метода градиентного спуска

```
1:  $x \leftarrow x^0$ 
2: while не сошлось do
3:    $r \leftarrow b - Ax$ 
4:    $\alpha \leftarrow \frac{(r, r)}{(r, Ar)}$ 
5:    $x \leftarrow x + \alpha r$ 
6: end while
```

Для доказательства сходимости метода мы будем использовать следующие



Свойства симметричных и положительно определённых матриц

1. Если $A = A^T$ и $A > 0$, то $A^{-1} = (A^{-1})^T$ и $A^{-1} > 0$.
2. Если $A = A^T$ и $A > 0$, то

$$\lambda_{\min} \leq \frac{(Ax, x)}{(x, x)} \leq \lambda_{\max} \quad (2.47)$$

$\forall x \in \mathbb{R}^n$, где λ_{\min} и λ_{\max} — минимальное и максимальное собственные значения A .

Теорема 2.8. Пусть \hat{x} — следующее за x приближение к x^* , построенное по методу градиентного спуска. Рассмотрим величину $\Delta\Phi(x) = \Phi(x) - \varphi_0$. Тогда

$$\Delta\Phi(\hat{x}) \leq \left(1 - \frac{1}{\alpha_2(A)}\right) \Delta\Phi(x). \quad (2.48)$$

Доказательство. Из (2.45) имеем $x = A^{-1}(b - r)$. Подставляя это выражение в

$$\Delta\Phi(x) = \frac{1}{2}(Ax, x) - (b, x) + \frac{1}{2}(A^{-1}b, b)$$

и приводя подобные, получаем

$$\Delta\Phi(x) = \frac{1}{2}(A^{-1}r, r). \quad (2.49)$$

Далее, согласно методу градиентного спуска имеем $\hat{x} = x + \alpha r$, $\alpha = \frac{(r, r)}{(Ar, r)}$.

Отсюда получаем

$$\hat{r} = b - A\hat{x} = b - Ax - \alpha Ar = (I - \alpha A)r.$$

Следовательно, согласно (2.49),

$$\begin{aligned} \Delta\Phi(\hat{x}) &= \frac{1}{2}(A^{-1}\hat{r}, \hat{r}) = \frac{1}{2}(A^{-1}r - \alpha r, r - \alpha Ar) = \\ &= \frac{1}{2}\left((A^{-1}r, r) - 2\alpha(r, r) + \alpha^2(r, Ar)\right) = \frac{1}{2}\left((A^{-1}r, r) - \frac{(r, r)^2}{(Ar, r)}\right) = \left(1 - \frac{(r, r)^2}{(Ar, r)(A^{-1}r, r)}\right) \Delta\Phi(x). \end{aligned}$$



Используя (2.47) и тот факт, что собственные значения взаимообратных матриц взаимообратны, имеем

$$\frac{(r, r)}{(Ar, r)} \frac{(r, r)}{(A^{-1}r, r)} \geq \frac{\lambda_{\min}}{\lambda_{\max}}.$$

Для завершения доказательства достаточно вспомнить свойство 3 числа обусловленности. □

Следствие 2.2. *Метод градиентного спуска всегда сходится (при условии $A = A^T$ и $A > 0$).*

Замечание 2.8. Сходимость метода может быть очень медленной в случае, когда $\alpha_2(A)$ велико.



2.7.3. Метод сопряжённых градиентов

Пусть v^1, v^2, \dots, v^m — линейно независимые векторы в \mathbb{R}^n . Линейную оболочку этих векторов будем обозначать

$$\text{span}(v^1, v^2, \dots, v^m) = \left\{ \sum_{k=1}^m \alpha_k v^k \mid \alpha_k \in \mathbb{R} \right\}.$$

Рассмотрим $x^0 \in \mathbb{R}^n$ и последовательность векторов x^k , $k = \overline{1, n}$, построенную по [методу спуска](#):

$$\begin{aligned} x^1 &= x^0 + \alpha_1 p^1, \\ x^2 &= x^1 + \alpha_2 p^2 = x^0 + \alpha_1 p^1 + \alpha_2 p^2, \\ &\dots \\ x^k &= x^{k-1} + \alpha_k p^k = x^0 + \sum_{j=1}^k \alpha_j p^j. \end{aligned} \tag{2.50}$$

Предположим, что векторы p^k , $k = \overline{1, n}$, линейно независимы. Рассмотрим подпространство

$$\Pi_k = \text{span}(p^1, \dots, p^k) \subset \mathbb{R}^n.$$

Матрицу размера $n \times k$, составленную из направлений спуска p^j , $j = \overline{1, k}$, обозначим P_k . Эта матрица имеет, очевидно, ранг k .

Согласно (2.50), имеем

$$x^k = x^0 + P_k y,$$

где $y = (\alpha_1, \dots, \alpha_k)^T$, то есть

$$x^k \in X_k = \{x^0 + \Pi_k\}.$$

Если мы предположим, что *каждый* вектор x_k минимизирует Φ по всему множеству X_k ,

$$\Phi(x^k) = \min_{x \in X_k} \Phi(x), \tag{2.51}$$



то это будет означать, что $x^n = x^*$, то есть для нахождения точного решения достаточно сделать не более n итераций метода спуска.

Чтобы построить метод, обладающий такими свойствами, предположим, что мы каким-то образом уже получили первые k приближений, удовлетворяющих условию (2.51). Наша цель теперь — построить следующее приближение $x^{k+1} = \hat{x}$ таким образом, чтобы выполнилось свойство (2.51) для $k + 1$. Это равносильно выбору вектора $p^{k+1} = p \notin \Pi_k$ и $\alpha^{k+1} = \alpha \in \mathbb{R}$ такого, что

$$\Phi(\hat{x}) = \Phi(x^k + \alpha p) = \min_{\beta \in \mathbb{R}} \Phi(x^k + \beta p).$$

$$\begin{aligned} \Phi(\hat{x}) &= \Phi(x^k) + \alpha(p, Ax^k - b) + \frac{\alpha^2}{2}(p, Ap) = \left[x^k = x^0 + P_k y \right] = \\ &= \Phi(x^k) + \underbrace{\frac{\alpha^2}{2}(p, Ap)}_{\varphi(\alpha)} - \underbrace{\alpha(p, b - Ax^0)}_{\psi(p)} + \underbrace{\alpha(p, AP_k y)}_{\psi(p)}. \end{aligned} \quad (2.52)$$

Последнее слагаемое портит картину. Если бы его не было, то задача минимизации $\Phi(\hat{x})$ свелась бы к минимизации выражения $\varphi(\alpha)$, т. к. $\Phi(x^k)$ не зависит от p и α .

Рассмотрим $\psi(p)$ подробнее:

$$(p, AP_k y) = \left(p, A \sum_{j=1}^k \alpha_j p^j \right) = \sum_{j=1}^k \alpha_j (p, Ap^j).$$

Мы видим, что если выбрать

$$p \perp Ap^j \quad \forall j = \overline{1, k}, \quad (2.53)$$

то $\psi(p) = 0$, и для минимизации (2.52) достаточно взять

$$\alpha = \frac{(p, b - Ax^0)}{(Ap, p)}. \quad (2.54)$$



Условия (2.53) называются *условиями А-сопряжённости*. Их выполнение влечёт за собой

$$(p, r^k) = (p, b - Ax^k) = (p, b - A(x_0 + P_k y)) = (p, b - Ax^0),$$

так что формула (2.54) в точности совпадёт с (2.46). Понятно, что условиям (2.53) в общем случае удовлетворяет бесконечно много векторов p . Для того, чтобы выбрать из них «лучший», накладывается дополнительное условие

$$\|p - r^k\|_2 \rightarrow \min. \quad (2.55)$$

Можно показать, что если p^k выбирать исходя из условий (2.53) и (2.55), то $p^{k+1} \in \text{span}(r^k, p^k)$. То есть, не нарушая общности можно положить

$$p^{k+1} = p = r + \beta p^k, \quad (2.56)$$

где неизвестный скаляр β можно найти из условия *А-сопряжённости*:

$$(Ap, p_k) = 0 \Rightarrow \beta = -\frac{(Ap_k, r)}{(Ap_k, p_k)}. \quad (2.57)$$

Собирая все вместе, получаем общий вид итерации *метода сопряжённых градиентов* (МСГ)

- 1) Если $k = 1$, положить $p^1 = r^0 = b - Ax^0$,
иначе найти β по (2.57), затем $p = p^k$ по (2.56).
- 2) Вычислить α по (2.46) с $r = r^{k-1}$.
- 3) Вычислить следующее приближение $x^k = x^{k-1} + \alpha p$.

Замечание 2.9. Для вычисления невязки r^{k+1} удобно использовать соотношение

$$r^{k+1} = b - A(x^k + \alpha p^k) = r^k - \alpha Ap^k.$$

Замечание 2.10. Несмотря на то, что по построению МСГ даёт точное решение не более чем за n итераций, на практике может оказаться, что делать все n итераций слишком дорого. С другой стороны, *ошибки округления* не гарантируют получения решения с нужной точностью за n итераций. Поэтому как правило критерием остановки итераций является (относительная) малость нормы *невязки*.



Алгоритм метода сопряжённых градиентов

```

1:  $x \leftarrow x^0; r \leftarrow b - Ax; k \leftarrow 1$ 
2: while ( $\|r\| > \epsilon \|b\|$  и  $k < k_{\max}$ ) do
3:   if  $k = 1$  then
4:      $p \leftarrow r$ 
5:   else
6:      $\beta \leftarrow -(r, q)/\gamma; p \leftarrow r + \beta p$ 
7:   end if
8:    $q \leftarrow Ap; \gamma \leftarrow (q, p)$ 
9:    $\alpha \leftarrow (p, r)/\gamma; x \leftarrow x + \alpha p$ 
10:   $r \leftarrow r - \alpha q; k \leftarrow k + 1$ 
11: end while

```

Замечание 2.11. Одна итерация МСГ требует одной операции умножения вектора на матрицу A . Особенно эффективно это реализуется в случае, когда A содержит много нулей.

Теорема 2.9. Если матрицу A , $A^T = A > 0$, можно представить в виде $A = I + B$, где матрица B имеет ранг m , тогда метод сопряжённых градиентов сходится не более чем за $m + 1$ шагов.

2.8. Простейшие итерационные методы

- [2.8.1. Принцип сжимающих отображений](#)
- [2.8.2. Общий вид стандартных итерационных методов](#)
- [2.8.3. Критерий сходимости](#)
- [2.8.4. Метод простой итерации](#)
- [2.8.5. Метод Якоби](#)
- [2.8.6. Методы Гаусса–Зейделя и релаксации](#)



Меню

2.8.1. Принцип сжимающих отображений

Расстояние между двумя точками $x, y \in \mathbb{R}^n$ условимся обозначать

$$\rho(x, y) = \|x - y\|. \quad (2.58)$$

Определение. Рассмотрим отображение $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$. Если существует константа $0 \leq \alpha < 1$, такая, что для любых $x, y \in \mathbb{R}^n$

$$\rho(f(x), f(y)) \leq \alpha \rho(x, y),$$

то отображение f называют *сжимающим*.

Определение. Последовательность точек (x^k) , $x^k \in \mathbb{R}^n$, называется *фундаментальной*, или *последовательностью Коши*, если

$$\rho(x^k, x^m) \rightarrow 0 \quad \text{при } k, m \rightarrow \infty.$$

Согласно критерию Коши, последовательность (x^k) сходится тогда и только тогда, когда она фундаментальна.

Теорема 2.10 (принцип сжимающих отображений). *Пусть $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ — сжимающее отображение. Тогда уравнение*

$$f(x) = x \quad (2.59)$$

имеет единственное решение $x = x^*$ и для любого $x_0 \in \mathbb{R}^n$ последовательность

$$x^{k+1} = f(x^k), \quad k = 0, 1, 2, \dots \quad (2.60)$$

сходится к x^* :

$$\rho(x^k, x^*) \xrightarrow{k \rightarrow \infty} 0.$$

Доказательство. Рассмотрим последовательность

$$x^1 = f(x^0), \quad x^2 = f(x^1), \quad \dots, \quad x^{k+1} = f(x^k), \quad \dots$$



и докажем, что она является [фундаментальной](#). Для этого оценим

$$\rho(x^k, x^{k+1}) = \rho(f(x^{k-1}), f(x^k)) \leq \alpha \rho(x^{k-1}, x^k) \leq \dots \leq \alpha^k \rho(x^0, x^1).$$

Пусть $m > k$. Тогда по неравенству треугольника

$$\begin{aligned} \rho(x^k, x^m) &\leq \rho(x^k, x^{k+1}) + \rho(x^{k+1}, x^{k+2}) + \dots + \rho(x^{m-1}, x^m) \leq \\ &\leq (\alpha^k + \alpha^{k+1} + \dots + \alpha^{m-1}) \rho(x^0, x^1), \end{aligned}$$

откуда получаем

$$\rho(x^k, x^m) \leq \frac{\alpha^k}{1-\alpha} \rho(x^0, x^1), \quad (2.61)$$

то есть (x^k) фундаментальна ($\alpha < 1$ по условию), и, следовательно, сходится к какому-то вектору $x^* \in \mathbb{R}^n$. Переходя к пределу в (2.60) получаем, что x^* удовлетворяет уравнению (2.59).

Для доказательства единственности предположим, что $\exists x^{**} \neq x^* : f(x^{**}) = x^{**}$. Это приводит к противоречию:

$$\rho(x^*, x^{**}) = \rho(f(x^*), f(x^{**})) \leq \alpha \rho(x^*, x^{**}) < \rho(x^*, x^{**}).$$

□

Следствие 2.3. Устремляя $m \rightarrow \infty$ в формуле (2.61), получаем априорную оценку погрешности:

$$\rho(x^k, x^*) \leq \frac{\alpha^k}{1-\alpha} \rho(x^0, x^1). \quad (2.62)$$

Кроме оценки (2.62) можно получить и апостериорную оценку погрешности, которая как правило более точна, но для получения которой необходимо проделать k итераций. Положив в (2.62) $x^0 = x^{k-1}$ получаем

$$\rho(x^k, x^*) \leq \frac{\alpha}{1-\alpha} \rho(x^{k-1}, x^k). \quad (2.63)$$

Замечание 2.12. Функция ρ (метрика), которая измеряет «расстояние» между двумя точками, не обязательно должна иметь вид (2.58). Достаточно лишь, чтобы она удовлетворяла аксиомам метрики:



- 1) $\rho(x, y) = 0 \Leftrightarrow x = y;$
- 2) $\rho(x, y) = \rho(y, x);$
- 3) $\rho(x, y) \leq \rho(x, z) + \rho(z, y).$

Замечание 2.13. Принцип сжимающих отображений играет важную роль в функциональном анализе и других разделах математики. Он может быть сформулирован не только для \mathbb{R}^n , но и для произвольных метрических пространств, обладающих свойством *полноты*. В частности, с помощью него легко доказывается существование решений для обыкновенных дифференциальных уравнений, интегральных уравнений и т. д.



Меню

2.8.2. Общий вид стандартных итерационных методов

Рассмотрим СЛАУ $Ax = b$. Составим эквивалентную ей систему (систему с тем же решением) вида

$$x = Bx + g, \quad (2.64)$$

(B — матрица, g — вектор) и соответствующий итерационный процесс

$$x^{k+1} = Bx^k + g. \quad (2.65)$$

Согласно [принципу сжимающих отображений](#) для сходимости этого процесса достаточно, чтобы отображение

$$f : x \mapsto Bx + g$$

было [сжимающим](#). Имеем

$$\rho(f(x), f(y)) = \|Bx - By\| \leq \|B\| \|x - y\| = \|B\| \rho(x, y).$$

Следовательно, имеет место следующая теорема.

Теорема 2.11 (достаточное условие сходимости итерационного процесса общего вида). *Если $\|B\| < 1$, то итерационный процесс (2.65) сходится к решению СЛАУ (2.64).*

Замечание 2.14. Если матрица B удовлетворяет условию теоремы (2.11), то для оценки погрешности метода вида (2.65) можно использовать формулы (2.62), (2.63), где $\alpha = \|B\|$.

Все рассматриваемые далее итерационные методы будут иметь вид (2.65). Различие будет заключаться лишь в способе перехода от исходной СЛАУ $Ax = b$ к системе вида (2.64).



Меню

2.8.3. Критерий сходимости

Определение. *Спектральным радиусом* $\rho(A)$ называется величина наибольшего по модулю собственного значения матрицы A .

Теорема 2.12. *Последовательность матриц $I, A, A^2, \dots, A^k, \dots$ сходится к нулевой матрице тогда и только тогда, когда $\rho(A) < 1$.*

Доказательство. \Rightarrow Пусть $A^k \xrightarrow[k \rightarrow \infty]{} 0$, и x_1 — собственный вектор единичной длины, соответствующий максимальному по модулю СЗ λ_1 . Тогда $A^k x_1 = \lambda_1^k x_1$, откуда

$$\|A^k\|_2 = |\lambda_1|^k \rightarrow 0 \Rightarrow |\lambda_1| = \rho(A) < 1.$$

\Leftarrow Пусть $\rho(A) < 1$. Рассмотрим J — жорданову каноническую форму матрицы A :

$$A = S^{-1}JS.$$

Так как $A^k = S^{-1}J^kS$ нам достаточно показать, что $J^k \rightarrow 0$.

Матрица J состоит из блоков вида $J_i = \lambda_i I + E_i$, где λ_i — собственные значения A , а E_i имеют вид

$$E_i = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}.$$

Матрица E_i нильпотентна, то есть $E_i^k = 0$ для всех k , больших некоторого числа $m_i < \infty$. Нетрудно заметить, что J^k сохраняет блочную структуру и состоит из блоков

$$J_i^k = (\lambda_i I + E_i)^k = \sum_{j=0}^k \frac{k!}{j!(k-j)!} \lambda_i^{k-j} E_i^j = \sum_{j=0}^{m_i} \frac{k!}{j!(k-j)!} \lambda_i^{k-j} E_i^j,$$



Вверх

Назад

Вперёд

Пред.

След.

Указатель

Помощь

Экран

откуда

$$\|J_i^k\|_2 \leq \sum_{j=0}^{m_i} \frac{k!}{j!(k-j)!} \|E_i^j\|_2 |\lambda_i|^{k-j} \xrightarrow{k \rightarrow \infty} 0.$$

□

Теорема 2.13 (критерий сходимости итерационного процесса общего вида). *Итерационный процесс вида $x^{k+1} = Bx^k + g$ сходится тогда и только тогда, когда $\rho(B) < 1$.*

Доказательство. Данная теорема является прямым следствием теоремы 2.12.

⇒ Рассмотрим

$$x^{k+1} - x^k = B(x^k - x^{k-1}) = \dots = B^k(x^1 - x^0) = B^k((B - I)x_0 + g).$$

Если последовательность (x^k) сходится, то $\|x^{k+1} - x^k\| \rightarrow 0 \ \forall x_0 \in \mathbb{R}^n$, откуда $\|B^k v\| \rightarrow 0 \ \forall v \in \mathbb{R}^n$, то есть $B^k \rightarrow 0 \Rightarrow \rho(B) < 1$.

⇐ Пусть $\rho(B) < 1$. Тогда из

$$x^k - x^* = B(x^{k-1} - x^*) = \dots = B^k(x^0 - x^*)$$

сразу следует, что $\|x^k - x^*\| \rightarrow 0$.

□



2.8.4. Метод простой итерации

Самый простой способ приведения СЛАУ $Ax = b$ к виду (2.64) состоит в добавлении к обеим частям вектора x :

$$x = (I - A)x + b.$$

Соответствующий итерационный метод

$$x^{k+1} = (I - A)x^k + b \quad (2.66)$$

будем называть *методом простой итерации* (МПИ).



2.8.5. Метод Якоби

Рассмотрим i -е уравнение СЛАУ $Ax = b$:

$$a_{i1}x_1 + \dots + a_{ii}x_i + \dots + a_{in}x_n = b_i.$$

Выражая из него x_i , получаем

$$x_i = \frac{1}{a_{ii}} \left(b_i - \sum_{j \neq i} a_{ij}x_j \right), \quad i = \overline{1, n}. \quad (2.67)$$

Для того, чтобы записать эту систему в векторном виде, рассмотрим разбиение матрицы A на слагаемые согласно рис. 2.4:

$$A = L + D + R. \quad (2.68)$$

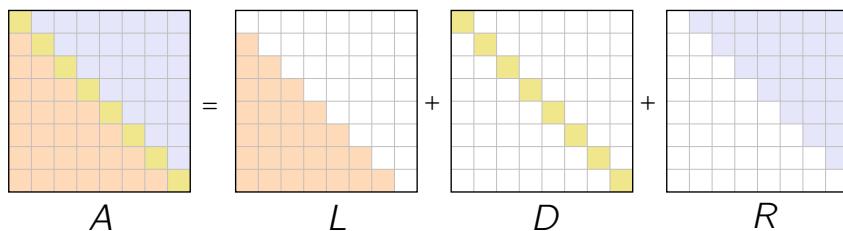


Рисунок 2.4

Тогда (2.67) примет вид

$$x = D^{-1}(b - (L + R)x) = B_J x + g_J,$$

где

$$B_J = -D^{-1}(L + R), \quad g_J = D^{-1}b. \quad (2.69)$$



Соответствующий итерационный метод

$$x_i^{k+1} = \frac{1}{a_{ii}} \left(b_i - \sum_{j \neq i} a_{ij} x_j^k \right), \quad i = \overline{1, n}, \quad k = 0, 1, 2, \dots \quad (2.70a)$$

называется *методом Якоби*. Его векторная форма имеет вид

$$x^{k+1} = B_J x^k + g_J, \quad (2.70b)$$

где B_J и g_J определяются по формуле (2.69).

Заметим, что $L + R = A - D$, поэтому для матрицы B_J существует альтернативная форма записи

$$B_J = I - D^{-1}A.$$



2.8.6. Методы Гаусса–Зейделя и релаксации

Рассмотрим i -й шаг k -ой итерации [метода Якоби](#):

$$x_i^{k+1} = \frac{1}{a_{ii}} \left(b_i - \sum_{j \neq i} a_{ij} x_j^k \right).$$

К этому моменту нам уже известны компоненты вектора x^{k+1} с номерами от 1 до $i-1$. Эти компоненты могут быть более точны, чем соответствующие компоненты текущего приближения x^k , поэтому их можно использовать в сумме (2.70a):

$$x_i^{k+1} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i+1}^n a_{ij} x_j^k \right). \quad (2.71a)$$

Векторный вариант (2.71a) имеет вид

$$x^{k+1} = D^{-1}(b - Lx^{k+1} - Rx^k), \quad \text{откуда}$$

$$x^{k+1} = B_S x^k + g_S, \quad (2.71b)$$

$$B_S = -(D + L)^{-1}R, \quad g_S = (D + L)^{-1}b.$$

Формулы (2.71a), (2.71b) определяют [метод Гаусса–Зейделя](#).

[Метод релаксации](#) получается путём взвешенного осреднения текущего приближения и приближения, построенного по методу Гаусса–Зейделя:

$$x_i^{k+1} = (1 - \omega)x_i^k + \frac{\omega}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i+1}^n a_{ij} x_j^k \right), \quad (2.72a)$$

где ω — весовой коэффициент, обычно $\omega \in (0, 2)$. Формула (2.72a) в векторной форме имеет вид

$$x^{k+1} = (1 - \omega)x^k + \omega D^{-1}(b - Lx^{k+1} - Rx^k).$$



Умножая обе части на D и группируя слагаемые, получаем

$$\begin{aligned}x^{k+1} &= B_\omega x^k + g_\omega, \\B_\omega &= (D + \omega L)^{-1}((1 - \omega)D - \omega R), \quad g_\omega = (D + \omega L)^{-1}b.\end{aligned}\tag{2.72b}$$

Замечание 2.15. При $\omega = 1$ метод релаксации, очевидно, превращается в [метод Зейделя](#).

Внимание! Для программной реализации стандартных итерационных методов используется исключительно их скалярные формы (2.70a), (2.71a), (2.72a). Соответствующие векторные записи (2.70b), (2.71b), (2.72b) используются для анализа сходимости методов.

Основные результаты о сходимости

- 1) Если матрица A обладает свойством [строгого диагонального преобладания](#), то Методы Якоби и Гаусса–Зейделя сходятся.
- 2) Если матрица A — симметричная и [положительно определённая](#), то метод релаксации сходится для всех $\omega \in (0, 2)$.

В остальных случаях сходимость методов нужно отдельно исследовать, например, с помощью теорем 2.11, 2.13.



2.9. Форматы хранения разреженных матриц

2.9.1. Координатный формат

2.9.2. Форматы CSR и CSC

2.9.3. Формат MSR



2.9.1. Координатный формат

Определение. *Разреженными* называют матрицы, содержащие большой процент нулевых элементов.

Разреженные матрицы очень часто возникают в приложениях, в частности, при численном моделировании различных физических процессов. Количество ненулевых элементов матрицы в дальнейшем будем обозначать n_z .

Понятно, что нерационально хранить в памяти ЭВМ все нулевые элементы разреженных матриц. Они, во-первых, зря занимают память, и, во-вторых, замедляют операции с матрицами. Поэтому существует ряд общепринятых способов хранения разреженных матриц.

Самый простой — так называемый *координатный формат*. В этом формате для хранения вещественной матрицы A используется три массива:

- AA — массив вещественных чисел для хранения ненулевых элементов A .
- IA — массив целых чисел для хранения номеров строк соответствующих элементов массива AA.
- JA — аналогичный массив для номеров столбцов.

Длина всех трёх массивов равна n_z .

Пример 2.1. Записать в координатном формате матрицу

2.				
		5.	7.	
		9.		
1.	4.		3.	
	8.			6.



Решение. Представление матрицы в координатном формате задаётся с точностью до перестановки элементов.

AA	1.	2.	3.	4.	5.	6.	7.	8.	9.
IA	4	1	4	4	2	5	2	5	3
JA	1	1	4	2	3	5	4	2	3





2.9.2. Форматы CSR и CSC

Если в примере [2.1](#) упорядочить элементы матрицы построчно, то информация в массиве IA окажется избыточной:

AA	2.	5.	7.	9.	1.	4.	3.	8.	6.
IA	1	2	2	3	4	4	4	5	5
JA	1	3	4	3	1	2	4	2	5

В этом случае достаточно хранить лишь указатель на элемент массива AA, с которого начинается хранение i -й строки:

IA	1	2	4	5	8	10
----	---	---	---	---	---	----

Здесь последний элемент вектора IA служит для того, чтобы можно было определить, где заканчивается n -я строка. Описанный способ хранения матриц называется *форматом CSR* (Compressed Sparse Row).

Если ненулевые элементы в массиве AA упорядочить не по строкам, а по столбцам, по аналогии получится *формат CSC* (Compressed Sparse Column).

Описание формата CSR (CSC)

- AA — массив вещественных чисел длины n_z , в котором хранятся упорядоченные по строкам (по столбцам) ненулевые элементы A .
- JA — массив целых чисел длины n_z для хранения номеров столбцов (строк) соответствующих элементов массива AA.
- IA — массив целых чисел длины $n + 1$. На i -й позиции массива хранится *номер позиции* в массиве AA, с которой начинается хранение элементов i -й строки (столбца). Более конкретно: $IA[1]=1$, $IA[i+1]=IA[i]+n_i$, где n_i — число ненулевых элементов в i -ой строке (столбце).

Формат CSR является одним из наиболее популярных.



Алгоритм умножения матрицы в формате CSR на вектор Вход: n, AA, JA, IA, x.

Выход: y.

```
for i=1 to n do
    y[i]=0
    for j=IA[i] to IA[i+1]-1 do
        y[i]=y[i]+AA[j]*x[JA[j]]
    end for
end for
```

Алгоритм умножения матрицы в формате CSC на вектор

```
y=0
for j=1 to n do
    for i=IA[j] to IA[j+1]-1 do
        y[JA[i]]=y[JA[i]]+AA[i]*x[j]
    end for
end for
```



2.9.3. Формат MSR

Во многих итерационных методах решения СЛАУ диагональные элементы матрицы A играют особую роль: доступ к ним нужно осуществлять чаще, чем к другим элементам матрицы. Рассмотренные ранее форматы, очевидно, не позволяют быстро найти диагональные элементы. Решить эту проблему помогает модификация формата CSR — *формат MSR* (Modified Sparse Row).

Основные отличия этого формата от формата CSR состоят в следующем.

- 1) Изменён формат массива AA: сначала в него *полностью* записывается главная диагональ матрицы A , а затем — недиагональные ненулевые элементы.
- 2) Массивы IA и JA объединены в один (назовём его IJ).

Рассмотрим матрицу из примера 2.1.

2.				
		5.	7.	
			9.	
1.	4.		3.	
	8.			6.

Запишем ее представление в формате MSR.

	1	2	3	4	5	6	7	8	9	10	11
AA	2.	0.	9.	3.	6.	×	5.	7.	1.	4.	8.
IJ	7	7	9	9	11	12	3	4	1	2	2



Массив AA: первые n элементов занимает главная диагональ A . Элемент на позиции $n + 1$ не используется. Начиная с $(n + 2)$ -го элемента построчно хранятся недиагональные ненулевые элементы A .

Массив IJ: $IJ[1]=n+2$, $IJ[i+1]=IJ[i]+n_i$ для $i = \overline{1, n}$, где n_i — количество недиагональных ненулевых элементов в i -й строке. Далее — номера столбцов для соответствующих элементов массива AA.

Замечание 2.16. Массивы AA и IJ полностью описывают матрицу A . Размерность матрицы легко найти по значению $IJ[1]$:

$$n=IJ[1]-2.$$

Кроме этого, длина обоих массивов равна $IJ[n+1]-1$.

Алгоритм умножения матрицы в формате MSR на вектор Вход: AA, IJ, x.

Выход: y.

```
n=IJ[1]-2
for i=1 to n do
    y[i]=AA[i]*x[i]
    for j=IJ[i] to IJ[i+1]-1 do
        y[i]=y[i]+AA[j]*x[IJ[j]]
    end for
end for
```



Глава 3

Методы решения проблемы собственных значений

- 3.1. Проблема собственных значений: общая характеристика
- 3.2. Степенной метод
- 3.3. Метод вращений Якоби
- 3.4. Метод Данилевского



3.1. Проблема собственных значений: общая характеристика

3.1.1. Сведения из линейной алгебры

3.1.2. Общая характеристика проблемы собственных значений

3.1.3. Обусловленность проблемы собственных значений



3.1.1. Сведения из линейной алгебры

Определение. Пусть A — квадратная матрица над полем \mathbb{C} . Вектор $x \neq 0 \in \mathbb{C}^n$ называется *собственным вектором* матрицы A , если существует $\lambda \in \mathbb{C}$ такое, что

$$Ax = \lambda x.$$

Число λ называется *собственным значением*, соответствующим x . Множество всех собственных значений A называется *спектром* и обозначается $\sigma(A)$.

Замечание 3.1. Собственный вектор определён с точностью до постоянного множителя:

$$Ax = \lambda x \Rightarrow A(\alpha x) = \lambda(\alpha x).$$

Поэтому все собственные векторы, соответствующие собственному значению λ образуют так называемое *собственное подпространство*. Размерность собственного подпространства (число линейно независимых собственных векторов с собственным значением λ) называют *геометрической кратностью* собственного значения.

Как известно, все собственные значения являются корнями *характеристического многочлена*

$$\varphi(\lambda) = \det(A - \lambda I).$$

Если λ — корень многочлена φ кратности k , то говорят, что *алгебраическая кратность* λ равна k .

Пример 3.1. Рассмотрим матрицу $A = I$. Она, очевидно, имеет одно собственное значение, равное 1, а собственным вектором является любой вектор $x \in \mathbb{C}^n$. Следовательно, геометрическая кратность собственного значения равна n . Характеристическое уравнение имеет вид $(1 - \lambda)^n = 0$, то есть геометрическая кратность равна алгебраической.

Пример 3.2. Рассмотрим матрицу

$$A = \begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix}.$$



Алгебраическая кратность собственного значения 1 равна 2. Собственным вектором является любой вектор вида $(\xi, 0)^T$, то есть геометрическая кратность равна 1.

Определение. Если квадратная матрица размерности n имеет n линейно независимых [собственных векторов](#), то она называется [диагонализируемой](#), а соответствующий ей линейный оператор — [оператором простой структуры](#).

Свойства диагонализируемых матриц

1. Диагонализируемая матрица может быть приведена к диагональному виду преобразованием подобия:

$$A = S^{-1}DS, \quad \det S \neq 0, \quad D = \text{diag}(\lambda_1, \dots, \lambda_n).$$

2. Матрица диагонализируема тогда и только тогда, когда алгебраическая и геометрическая кратности каждого собственного значения совпадают.



3.1.2. Общая характеристика проблемы собственных значений

Задачи на нахождение собственных значений условно делятся на два класса. Если нужно найти одно или несколько собственных значений и соответствующие им собственные векторы, то проблема собственных значений (ПСЗ) называется *частичной*. Если же необходимо найти все собственные значения и векторы, проблема называется *полной*.



3.1.3. Обусловленность проблемы собственных значений

В отличие от [задачи решения СЛАУ](#), обусловленность проблемы собственных значений в литературе исследуется посредством сравнения абсолютных, а не относительных, погрешностей.

Лемма 3.1. *Если $\det(I + A) = 0$, то $\|A\| \geq 1$ для любой подчинённой матричной нормы.* [\[Доказательство\]](#)

Теорема 3.1 (Бауэра–Файка). *Пусть $A = S^{-1}DS$ — диагонализируемая матрица с собственными значениями $\lambda_1, \dots, \lambda_n$. Рассмотрим матрицу $B = A + E$ и её собственное значение μ . Тогда*

$$\min_i |\lambda_i - \mu| \leq \alpha_p(S) \|E\|_p \quad (3.1)$$

для любой p -нормы.

Доказательство. Достаточно рассмотреть случай $\mu \notin \sigma(A)$, так как в противном случае (3.1) выполняется всегда. По условию матрица $B - \mu I$ вырождена, значит

$$\begin{aligned} 0 &= |S(B - \mu I)S^{-1}| = |S(A + E - \mu I)S^{-1}| = |D - \mu I + SES^{-1}| = \\ &= [|\tilde{D}| = |D - \mu I| \neq 0] = |\tilde{D}(I + \tilde{D}^{-1}SES^{-1})| \Rightarrow |I + \tilde{D}^{-1}SES^{-1}| = 0. \end{aligned}$$

Тогда по лемме 3.1 получаем

$$1 \leq \|\tilde{D}^{-1}SES^{-1}\|_p \leq \|\tilde{D}^{-1}\|_p \alpha_p(S) \|E\|_p.$$

Матрица \tilde{D}^{-1} — диагональная с элементами $(\lambda_i - \mu)^{-1}$. Поэтому по определению p -нормы имеем

$$\|\tilde{D}^{-1}\|_p = (\min_i |\lambda_i - \mu|)^{-1}.$$

Последние две формулы дают (3.1). □

Таким образом, число $\alpha_p(S)$ показывает чувствительность собственных значений матрицы A к возмущениям элементов этой матрицы. Необходимо отметить, что матрица S , приводящая A к диагональному виду, не единственная.



Определение. Матрица A , которая может быть диагонализирована унитарным преобразованием подобия, называется *нормальной*.

Из теоремы Бауэра–Файка сразу следует, что задача вычисления собственных значений всегда хорошо обусловлена для нормальных матриц. Следует также понимать, что одни собственные значения матрицы могут быть гораздо более чувствительны к изменениям коэффициентов, чем другие. Кроме этого, хорошая обусловленность задачи нахождения собственных значений матрицы не означает хорошей обусловленности задачи нахождения ее собственных векторов.



3.2. Степенной метод

[3.2.1. Случай 1](#)

[3.2.2. Случай 2](#)

[3.2.3. Случай 3](#)

[3.2.4. Случай 4](#)

[3.2.5. Общий случай](#)

[3.2.6. Степенной метод со сдвигом](#)

Степенной метод позволяет найти максимальное по модулю [собственное значение](#) и соответствующий ему собственный вектор вещественной [диагонализируемой матрицы](#).

Пусть A — диагонализируемая матрица, $\lambda_1, \dots, \lambda_n$ — упорядоченные по убыванию модуля собственные значения, x^1, \dots, x^n — соответствующий базис из собственных векторов. Рассмотрим несколько возможных случаев.



3.2.1. Случай 1

Пусть $\lambda_1 \in \mathbb{R}$, $|\lambda_1| > |\lambda_2| \geq \dots \geq |\lambda_n|$.

Разложим базису $\{x^i\}$ произвольный вектор $y^0 \in \mathbb{C}^n$: $y^0 = \sum_1^n \alpha_i x^i$. Предположим, что $\alpha_1 \neq 0$ и рассмотрим последовательность (y^k) :

$$y^{k+1} = Ay^k.$$

Тогда

$$y^k = A^k y^0 = \sum_1^n \alpha_i \lambda_i^k x^k = \lambda_1^k \left(\alpha_1 x^1 + \sum_2^n \alpha_i \left(\frac{\lambda_i}{\lambda_1} \right)^k x^i \right), \quad (3.2)$$

Значит, при $k \rightarrow \infty$ имеем

$$y^k \sim \lambda_1^k \alpha_1 x^1,$$

то есть вектор y^k всё сильнее приближается к **собственному подпространству**, соответствующему x^1 . Для практического применения, естественно, необходимо на каждом шаге нормировать y^k . Существует несколько вариантов нормировки, но наиболее эффективный из них следующий.

Обозначим $\max(x)$ максимальную по модулю компоненту вектора x . Тогда процесс степенного метода примет вид

$$v^{k+1} = Au^k, \quad u^{k+1} = v^{k+1} / \max(v^{k+1}), \quad u^0 = y^0.$$

При таком способе нормировки получаем $\max(u^k) = 1$ при всех $k \geq 1$, а также $u^k = y^k / \max(y^k)$. Вместо (3.2) мы теперь имеем

$$u^k = c_k \lambda_1^k \left(\alpha_1 x^1 + \sum_2^n \alpha_i \gamma_i^k x^i \right),$$

где для краткости $\gamma_i = \lambda_i / \lambda_1$, c_k — число порядка λ_1^{-k} . Тогда

$$v^{k+1} = Au^k = c_k \lambda_1^{k+1} \left(\alpha_1 x^1 + \sum_2^n \alpha_i \gamma_i^{k+1} x^i \right) = \lambda_1 u^k + c_k \lambda_1^{k+1} \sum_2^n \alpha_i (\gamma_i^{k+1} - \gamma_i^k) x^i.$$



Отсюда

$$v^{k+1} - \lambda_1 u^k \sim \lambda_1 \sum_2^n \alpha_i \gamma_i^k (\gamma_i - 1) x^i = O((\lambda_2/\lambda_1)^k).$$

Значит,

$$\max(v^k) \rightarrow \lambda_1.$$

Степенной метод, как видим, сходится со скоростью геометрической прогрессии со знаменателем $|\gamma_2|$.

Базовый алгоритм степенного метода

```
u = y0
while не сошлось do
    v ← Au
    λ ← max(v)
    u ← v/λ
end while
```



3.2.2. Случай 2

Пусть $\lambda_1 = \lambda_2 = \dots = \lambda_m \in \mathbb{R}$, $|\lambda_1| > |\lambda_{m+1}| \geq \dots \geq |\lambda_n|$.

Так как матрица A по условию [диагонализируема](#), [геометрическая кратность](#) λ_1 равна m , то есть собственные векторы x_1, \dots, x_m линейно независимы и образуют собственное подпространство X_m размерности m . Тогда формула (3.2) примет вид

$$y^k = \lambda_1^k \left(\sum_1^m \alpha_i x^i + \sum_{m+1}^n \alpha_i \left(\frac{\lambda_i}{\lambda_1} \right)^k x^i \right) \sim \lambda_1^k \sum_1^m \alpha_i x^i.$$

В этом случае [алгоритм](#) остаётся без изменений. Последовательность (u^k) сходится к какому-то вектору $u \in X_m$ — он ничем не хуже x_1, \dots, x_m .



3.2.3. Случай 3

Пусть $\lambda_1 = -\lambda_2 \in \mathbb{R}$, $|\lambda_1| > |\lambda_3| \geq \dots \geq |\lambda_n|$.

В этом случае при $k \rightarrow \infty$ получаем

$$y^{2k} \sim \lambda_1^{2k}(\alpha_1 x^1 + \alpha_2 x^2),$$

$$y^{2k+1} \sim \lambda_1^{2k+1}(\alpha_1 x^1 - \alpha_2 x^2),$$

то есть последовательность (u^k) , построенная по [базовому алгоритму](#), не сходится, а распадается на две сходящиеся подпоследовательности (u^{2k}) и (u^{2k+1}) . Эти последовательности сходятся к двум различным векторам из $\text{span}(x^1, x^2)$, причём, в отличие от предыдущего случая, ни один из них не является, вообще говоря, собственным.

Но выход есть. Имеем $\max(v^{2k}) \rightarrow \lambda_1^2$, то есть $|\lambda_1|$ мы можем найти. Положим для определённости $\lambda_1 > 0$. Тогда для достаточно больших значений k

$$u^k \approx c(\alpha_1 \lambda_1^k x^1 + \alpha_2 (-\lambda_1)^k x^2),$$

$$v^{k+1} = Au^k \approx c(\alpha_1 \lambda_1^{k+1} x^1 + \alpha_2 (-\lambda_1)^{k+1} x^2)$$

и

$$\tilde{v} = \lambda_1 u^k + v^{k+1} = 2c_k \alpha_1 \lambda_1^{k+1} x^1.$$

Таким образом, $\tilde{v} \in \text{span}(x^1)$. По желанию его можно пронормировать. Аналогично можно найти x^2 .



3.2.4. Случай 4

Пусть $\lambda_2 = \bar{\lambda}_1 \in \mathbb{C}$, $|\lambda_1| > |\lambda_3| \geq \dots \geq |\lambda_n|$.

В этом случае разложение вектора y^0 по базису x^i имеет вид

$$y^0 = \alpha_1 x_1 + \bar{\alpha}_1 \bar{x}_1 + \sum_3^n \alpha_i x^i.$$

Обозначим $\lambda_1 = \rho e^{i\varphi}$, $\alpha_1 x_j^1 = r_j e^{i\theta_j}$ тогда

$$y^k \sim \lambda_1^k \alpha_1 x_j^1 + \bar{\lambda}_1^k \bar{\alpha}_1 \bar{x}_j^1 = \rho^k r_j (e^{\varphi k + \theta_j} + e^{-(\varphi k + \theta_j)}) = 2\rho^k r_j \cos(\varphi k + \theta_j).$$

Компоненты вектора u^k осциллируют и не стремятся ни к какому пределу. В этом случае всё равно можно найти λ_1 и x_1 , см. [Дж. Уилкинсон. Алгебраическая проблема собственных значений (Наука, 1970.)].



3.2.5. Общий случай

Вообще говоря, степенной метод можно применять не только для [диагонализируемых матриц](#). Если геометрическая и алгебраическая кратности λ_1 совпадают, этот метод будет работать. В противном случае метод тоже может сходиться, но очень медленно.



Вверх

Назад

Вперёд

Пред.

След.

Указатель

Помощь

Экран

3.2.6. Степенной метод со сдвигом

Степенной метод теоретически можно применять для отыскания произвольного собственного значения λ_j и x^j , пользуясь соотношением

$$\sigma(\alpha A + \beta I) = \alpha\sigma(A) + \beta.$$

За счёт выбора α и β можно сместить спектр таким образом, что $\mu_j = \alpha\lambda_j + \beta$ станет максимальным собственным значением матрицы $\alpha A + \beta I$. Найдя μ_j степенным методом, соответственно, найдём и λ_j .

Опишем схему нахождения минимального собственного значения в случае вещественного спектра.

- 1) Пусть $0 \leq \lambda_i < m$ (когда m неизвестно, можно взять $m = \|A\|$). Заменим A на $B = mI - A$. Максимальное собственное значение B равно $m - \lambda_n$. Аналогично поступаем когда все λ_i отрицательны.
- 2) Если λ_i имеют различные знаки, рассмотрим матрицу A^2 . У неё те же собственные векторы, а собственные значения равны λ_i^2 . Таким образом задача свелась к первому случаю.



3.3. Метод вращений Якоби

3.3.1. Симметричная проблема собственных значений

3.3.2. Общая схема вращений Якоби

3.3.3. Расчётные формулы метода



3.3.1. Симметричная проблема собственных значений

Пусть матрица A — вещественная и симметричная: $A^T = A$. Проблема собственных значений для такой матрицы является более простой, чем в общем случае, так как

- A [диагонализируема](#):

$$X^{-1}AX = D = \text{diag}(\lambda_1, \dots, \lambda_n); \quad (3.3)$$

- все λ_i вещественны.

Также из (3.3) имеем $AX = XD$, то есть столбцы матрицы X являются собственными векторами A :

$$Ax_i = \lambda_i x_i.$$



3.3.2. Общая схема вращений Якоби

Суть метода состоит в построении последовательности матриц

$$A^{(0)} = A, \quad A^{(m)} = X_m^T A^{(m-1)} X_m, \quad m = 1, 2, \dots,$$

где X_m — матрицы элементарного вращения (см. (2.41)), которые строятся таким образом, что

$$\|A^{(m)} - \tilde{D}\| \xrightarrow[m \rightarrow \infty]{} 0, \quad (3.4)$$

где \tilde{D} — некоторая диагональная матрица. Таким образом, для достаточно большого M мы будем иметь

$$A^{(M)} \approx \text{diag}(\lambda_1, \dots, \lambda_n), \quad X \approx X_1 X_2 \dots X_M.$$

Главный вопрос: каким образом выбирать X_m , чтобы достичь сходимости?

Определение. *Нормой Фробениуса* называется матричная норма $\|\cdot\|_F$, определяемая как

$$\|A\|_F = \sqrt{\sum_{i,j}^n |a_{ij}|^2}.$$

Рассмотрим величины

$$\text{off}(A) = \sum_{i \neq j}^n a_{ij}^2, \quad \text{on}(A) = \sum_{i=1}^n a_{ii}^2.$$

По определению имеем

$$\text{on}(A) + \text{off}(A) = \|A\|_F^2.$$

Условие сходимости (3.4) можно теперь записать в эквивалентной форме

$$\text{off}(A^{(m)}) \xrightarrow[m \rightarrow \infty]{} 0.$$



Матрица X_m , определяющая переход от $A^{(m-1)}$ к $A^{(m)}$, является матрицей элементарного вращения и имеет вид

$$V_{pq}(\theta) = \begin{bmatrix} 1 & & & p & q \\ & \ddots & & & \\ & & 1 & & \\ & & & c & s \\ & & & -s & c \\ & & & & 1 \\ & & & & & \ddots \\ & & & & & & 1 \end{bmatrix} \quad \begin{array}{l} p \\ c = \cos \theta, \\ s = \sin \theta. \\ q \end{array} \quad (3.5)$$

Таким образом, X_m определяется тремя параметрами: p , q и θ . Чтобы понять, каким образом следует выбирать эти параметры, рассмотрим как преобразование подобия

$$A \mapsto A' = V_{pq}^T A V_{pq} \quad (3.6)$$

изменяет матрицу A (см. рисунок).

$$\begin{array}{c} \begin{array}{|c|c|c|c|c|} \hline & 1 & & & \\ \hline p & & c & & -s \\ \hline & 1 & & 1 & \\ \hline q & & s & & c \\ \hline & & & & 1 \\ \hline \end{array} & A & \begin{array}{|c|c|c|c|c|} \hline & 1 & & & \\ \hline & & 1 & & s \\ \hline & & & 1 & \\ \hline & & -s & & c \\ \hline & & & & 1 \\ \hline \end{array} & = & \begin{array}{|c|c|c|c|c|} \hline & \text{---} & \text{---} & \text{---} & \text{---} \\ \hline & \text{---} & \text{---} & \text{---} & \text{---} \\ \hline & \text{---} & \text{---} & \text{---} & \text{---} \\ \hline & \text{---} & \text{---} & \text{---} & \text{---} \\ \hline & \text{---} & \text{---} & \text{---} & \text{---} \\ \hline \end{array} \end{array}$$

V_{pq}^T



Видно, что преобразование (3.6) изменяет в матрице A только строки и столбцы с индексами p и q . В частности, диагонали матриц A и A' отличаются лишь элементами на позициях (p, p) и (q, q) . Введём следующие обозначения: $a_{pp} = a$, $a_{qq} = b$, $a'_{pp} = \alpha$, $a'_{qq} = \beta$. Тогда

$$\operatorname{on}(A') = \operatorname{on}(A) - a^2 - b^2 + \alpha^2 + \beta^2. \quad (3.7)$$

Обозначим также $a_{pq} = a_{qp} = e$, $a'_{pq} = a'_{qp} = \varepsilon$. Из (3.6) имеем соотношение

$$\begin{bmatrix} c & -s \\ s & c \end{bmatrix} \begin{bmatrix} a & e \\ e & b \end{bmatrix} \begin{bmatrix} c & s \\ -s & c \end{bmatrix} = \begin{bmatrix} \alpha & \varepsilon \\ \varepsilon & \beta \end{bmatrix}. \quad (3.8)$$

Так как ортогональное преобразование подобия сохраняет [норму Фробениуса](#), получаем

$$\alpha^2 + \beta^2 + 2\varepsilon^2 = a^2 + b^2 + 2e^2.$$

Используем это тождество в (3.7):

$$\operatorname{on}(A') = \operatorname{on}(A) + 2(e^2 - \varepsilon^2),$$

что с учётом

$$\|A\|_F = \|A'\|_F$$

равносильно

$$\operatorname{off}(A') = \operatorname{off}(A) - 2(e^2 - \varepsilon^2). \quad (3.9)$$

Из равенства (3.9) видно, что наименьшее значение $\operatorname{off}(A')$ достигается, если величины s и c (то есть угол θ) выбраны таким образом, что $\varepsilon = a'_{pq} = 0$. Для того, чтобы величина $\operatorname{off}(A)$ уменьшилась как можно сильнее нужно выбирать в качестве обнуляемого элемента a_{pq} максимальный по модулю недиагональный элемент матрицы A .

Базовый алгоритм метода Якоби. Пока $\operatorname{off}(A)$ не достаточно мало:

- 1) Среди a_{ij} для $i < j$ выбрать максимальный по модулю a_{pq} .



2) Выбрать c и s таким образом, чтобы после преобразования (3.6) получилось $a'_{pq} = 0$.

$$3) A \leftarrow V_{pq}^T A V_{pq}.$$

Согласно (3.9) для этого алгоритма имеем

$$\text{off}(A') = \text{off}(A) - 2e^2. \quad (3.10)$$

Здесь $A' = A^{(m+1)}$, $A = A^{(m)}$, $e = a_{pq}^{(m)}$. Важно понимать, что каждый шаг алгоритма в общем случае «портит» нули, сделанные на предыдущем шаге.

Оценим скорость сходимости алгоритма. Если $a_{pq} = e$ — максимальный по модулю недиагональный элемент матрицы A , то справедлива оценка

$$\text{off}(A) \leq 2Ne^2,$$

где $N = n(n - 1)/2$. Тогда из (3.10) получаем

$$\text{off}(A') \leq \left(1 - \frac{1}{N}\right) \text{off}(A),$$

что означает

$$\text{off}(A^{(m)}) \leq \left(1 - \frac{1}{N}\right)^m \text{off}(A). \quad (3.11)$$

В частности, при $n = 2$ очевидно имеем сходимость за одну итерацию.

Согласно (3.11) метод вращений Якоби сходится со скоростью геометрической прогрессии. Однако эта оценка слишком груба, и на практике метод сходится быстрее.



3.3.3. Расчётные формулы метода

Для окончательного определения метода осталось вывести формулы, по которым вычисляются элементы матрицы V_{pq} во [втором пункте](#) алгоритма. Из (3.8) получаем

$$\alpha = ac^2 + bs^2 - 2es\cos c, \quad (3.12a)$$

$$\beta = as^2 + bc^2 + 2es\cos c, \quad (3.12b)$$

$$\varepsilon = e(c^2 - s^2) + (a - b)sc = 0. \quad (3.12c)$$

Для решения уравнения (3.12c) введём переменные

$$t = \operatorname{tg} \theta = \frac{s}{c}, \quad z = \frac{b - a}{2e}.$$

После простых преобразований из (3.12c) получаем

$$t^2 + 2zt - 1 = 0,$$

и

$$t = -z \pm \sqrt{z^2 + 1}. \quad (3.13)$$

Использование этой формулы на практике приводит к большим ошибкам округления, поэтому нужно ее переписать в более подходящем для машинных вычислений виде. Домножив (3.13) на $z \pm \sqrt{z^2 + 1}$ получим

$$t = \frac{1}{z \pm \sqrt{z^2 + 1}}.$$

Важно выбрать из этих двух корней наименьший по модулю. Он равен

$$t = \frac{\operatorname{sign} z}{|z| + \sqrt{z^2 + 1}}. \quad (3.14)$$

После этого вычисляем

$$c = \frac{1}{\sqrt{1 + t^2}}, \quad s = tc. \quad (3.15)$$



Теоретически на этом этапе нам остаётся с помощью найденных значений s и c вычислить A' по формуле (3.6). Однако для того, чтобы метод был эффективен, следует организовать вычисления следующим образом.

Как мы видели, операция $A \leftarrow A'$ заключается в изменении только строк и столбцов с индексами p и q в матрице A . Для диагональных элементов $a'_{pp} = \alpha$ и $a'_{qq} = \beta$ имеют место формулы (3.12a), (3.12b), $a'_{pq} = a'_{qp} = 0$ по построению, а для $j \neq p, j \neq q$ из (3.6) имеем

$$\begin{aligned} a'_{jp} &= a'_{pj} = ca_{pj} - sa_{qj}, \\ a'_{jq} &= a'_{qj} = sa_{pj} + ca_{qj}. \end{aligned} \tag{3.16}$$

Для вычислительной устойчивости нужно представить вышеперечисленные формулы в виде

$$a'_{ij} = a_{ij} + [\text{какая-то поправка}].$$

Так, из (3.12a), (3.12b) с учётом (3.12c) получаем

$$\begin{aligned} a'_{pp} &= a_{pp} - ta_{pq}, \\ a'_{qq} &= a_{qq} + ta_{pq}, \end{aligned} \tag{3.17}$$

а вместо (3.16) имеем

$$\begin{aligned} a'_{pj} &= a_{pj} - s(a_{qj} + \tau a_{pj}), & \text{где } \tau = \frac{s}{1+c}. \\ a'_{qj} &= a_{qj} + s(a_{pj} - \tau a_{qj}), \end{aligned} \tag{3.18}$$

Замечания по практической реализации

- 1) В памяти следует хранить только верхний треугольник матрицы A .



- 2) Матрицы вращения V_{pq} , очевидно, в память не записываются, используются только величины c и s .
- 3) Преобразования (3.6) осуществляются по формулам (3.17), (3.18), при этом нужно грамотно учитывать симметрию матрицы.
- 4) При больших n выбор максимального по модулю элемента требует слишком много времени, поэтому как правило элементы a_{pq} для обнуления выбирают циклически: $a_{12}, \dots, a_{1n}, a_{23}, \dots, a_{2n}, \dots, a_{n-1,n}$. Обычно для получения решения в пределах машинной точности достаточно 5-6 таких проходов. При этом на первых двух-трёх проходах, если модуль a_{pq} достаточно мал, его пропускают.



3.4. Метод Данилевского

Метод Данилевского позволяет вычислить **характеристический многочлен** произвольной матрицы A .

Рассмотрим матрицу вида

$$\begin{bmatrix} p_1 & p_2 & \cdots & p_{n-1} & p_n \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}. \quad (3.19)$$

Такой вид матрицы называется *формой Фробениуса*. Её характеристический многочлен легко вычисляется путём рекурсивного разложения определителя по столбцу:

$$\begin{vmatrix} p_1 - \lambda & p_2 & \cdots & p_{n-1} & p_n \\ 1 & -\lambda & 0 & \cdots & 0 \\ 0 & 1 & -\lambda & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & -\lambda \end{vmatrix} = (p_1 - \lambda)(-\lambda)^{n-1} \begin{vmatrix} p_2 & \cdots & p_{n-1} & p_n \\ 1 & -\lambda & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & -\lambda \end{vmatrix} = \dots = (-1)^n(\lambda^n - p_1\lambda^{n-1} - \dots - p_{n-1}\lambda - p_n). \quad (3.20)$$

Идея метода Данилевского проста: с помощью элементарных преобразований подобия привести данную матрицу A к форме Фробениуса A' . Так как преобразования подобия сохраняют спектр матрицы, характер-



ристические многочлены матриц A и A' будут совпадать, то есть искомый характеристический многочлен может быть вычислен по формуле (3.20).

Рассмотрим алгоритм на примере матрицы A размерности 4. Будем последовательно приводить строки матрицы к нужному виду, начиная с последней.

1. Для начала «делаем единицу» на позиции (4,3): делим третий столбец на $\alpha = a_{43}$. Это эквивалентно умножению матрицы A справа на матрицу

$$T_{43} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \alpha^{-1} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Чтобы сохранить спектр, нужно дополнить это преобразование до преобразования подобия, то есть умножить результат слева на T_{43}^{-1} . Эта операция осуществляется путем умножения третьей строки на α . В результате указанных операций получаем матрицу

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

2. «Делаем нули» в последней строке: для $j \neq 3$ вычитаем из j -го столбца 3-й, умноженный на $\alpha = a_{4j}$.



Каждая такая операция должна быть дополнена до преобразования подобия, для чего необходимо к 3-й строке добавлять j -ю, умноженную на α . Получаем в итоге

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

3. Аналогично поступаем с третьей и второй строками. При этом сделанные ранее нули и единицы не пропадают:

$$\begin{bmatrix} \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \\ 0 & 0 & 1 & 0 \end{bmatrix} \mapsto \begin{bmatrix} \times & \times & \times & \times \\ \times & \times & \times & \times \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \mapsto \begin{bmatrix} \times & \times & \times & \times \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} p_1 & p_2 & p_3 & p_4 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

Общий алгоритм метода Данилевского

```

1: for  $i = \overline{n, 2}$  do
2:    $\alpha \leftarrow a_{i,i-1}$ 
3:    $a_{i-1} \leftarrow a_{i-1}/\alpha$ ;  $\underline{a_{i-1}} \leftarrow \alpha \underline{a_{i-1}}$ 
4:   for  $j \neq i$  do
5:      $\alpha \leftarrow a_{ij}$ 
6:      $a_j \leftarrow a_j - \alpha a_{i-1}$ 
7:      $\underline{a_{i-1}} \leftarrow \underline{a_{i-1}} + \alpha \underline{a_j}$ 
8:   end for

```



9: end for

Напомним, что здесь a_j — j -й вектор-столбец матрицы A , a_i — i -я строка матрицы A , n — размерность матрицы A .

После выполнения приведенного алгоритма (в невырожденном случае) на месте матрицы A будет находиться ее [форма Фробениуса](#).

Выбор главного элемента. Для минимизации погрешностей округления при машинной реализации необходимо выбирать главный элемент. Для этого перед началом i -го шага нужно выбрать максимальный по модулю элемент a_{ij^*} среди a_{ij} для $j = \overline{1, i-1}$. После этого меняем местами столбцы $i-1$ и j^* , а также соответствующие строки (для сохранения спектра).

Вырожденный случай. Предположим, на i -ом этапе не удаётся выбрать главный элемент. Это означает, что матрица имеет вид

$$A = \left[\begin{array}{c|c} A_1 & \boxtimes \\ \hline 0 & A_2 \end{array} \right],$$

где A_1, A_2 — квадратные блоки размерности $n-i$ и i соответственно, причём A_2 имеет форму Фробениуса. Тогда имеем

$$|A - \lambda I| = |A_{n-i} - \lambda I_{n-1}| \cdot |A_2 - \lambda I_2|.$$

Второй множитель мы можем вычислить сразу, поэтому остаётся лишь привести к форме Фробениуса матрицу A_1 .



Глава 4

Решение численных уравнений

- 4.1. Приближенное решение одного численного уравнения
- 4.2. Приближенное решение систем численных уравнений



4.1. Приближенное решение одного численного уравнения

4.1.1. Введение

4.1.2. Метод бисекции

4.1.3. Метод простой итерации решения численных уравнений

4.1.4. О задаче улучшения метода итерации

4.1.5. Метод Ньютона

4.1.6. Видоизменения метода Ньютона

4.1.7. Методы отыскания корней алгебраических уравнений



4.1.1. Введение

Рассмотрим уравнение вида

$$f(x) = 0, \quad (4.1)$$

где f – некоторая заданная функция, а x – неизвестная численная величина. При решении таких уравнений приходится решать две задачи:

- 1) *Отделение корней*, т.е. отыскание таких достаточно малых (смысл этого термина станет ясным по ходу изложения материала) интервалов, в которых находится один и только один корень уравнения (4.1);
- 2) Вычисление корней с требуемой точностью.

Простейший аналитический признак отделенности корня дает следующая теорема из анализа.

Теорема 4.1. Если функция $f(x)$ определена и непрерывна на некотором конечном отрезке $[a, b]$ и на концах этого отрезка принимает значения противоположных знаков, а в любой внутренней точке промежутка (a, b) функция $f'(x)$ имеет производную $f'(x)$, которая сохраняет знак, то внутри отрезка $[a, b]$ существует корень уравнения (4.1) и этот корень единственный.

Естественно, речь в данной теореме идет об отделении вещественных корней уравнения (4.1).

Часто при практическом решении задачи отделения корня пользуются графическими методами. При этом каким-либо способом строят график функции $y = f(x)$ (либо с привлечением средств анализа, либо просто по точкам на достаточно густой сетке) и визуально определяют точки его пересечения с осью абсцисс. Конечно, необходимо помнить, что так называемые «реальные» прикладные задачи чаще всего бывают «плохими» или «очень плохими», нежели «хорошими» или «очень хорошими». К числу первых, например, следует отнести случай, когда $f(x)$ имеет два очень близко ($\approx 10^{-10}$) расположенных корня. В этом плане прекрасный пример построил Уилкинсон. Это трехдиагональная матрица W_{21} , определяемая соотношениями $w_{ii} = 11 - i$, $i = \overline{1, 21}$; $w_{i+1i} = w_{ii+1} = 1$; $i = \overline{1, 20}$. Наибольшее собственное значение $\lambda_{21} \approx 10.74\dots$ совпадает с наибольшим собственным значением главной подматрицы порядка 20 в первых 15 десятичных разрядах. График $P_{21}(\lambda)$ близок к -20 на всем интервале $(10, 11)$, за исключением подинтервала с центром в точке λ_{21} с шириной, меньшей, чем 10^{-13} (!).



Иногда при графическом решении задачи об отделении корней оказывается более удобным представить исходное уравнение (4.1) в виде $\varphi(x) = \psi(x)$, а затем построить графики функций $y = \varphi(x)$ и $y = \psi(x)$ (конечно, предполагается, что это сделать проще, чем построить график исходной функции $y = f(x)$) и найти визуально точки их пересечения (естественно, найденные промежутки следует проверить, например, с помощью сформулированной выше теоремы).

Приведем пример решения задачи об [отделении корня](#). Пусть исходное уравнение имеет вид

$$x \sin x = 1. \quad (4.2)$$

Здесь $f(x) = x \sin x - 1$. Так как строить график такой функции относительно сложно, то представим уравнение (4.2) в виде (заметим, $x = 0$ не является корнем)

$$\sin x = \frac{1}{x}$$

и построим графики функций $y = \sin x$ и $y = \frac{1}{x}$.

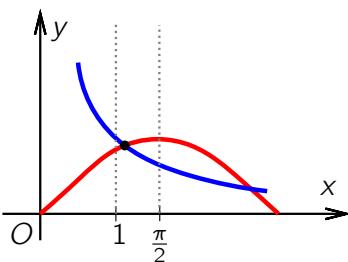


Рисунок 4.1

Отсюда видим, что ближайшая к нулю положительная точка пересечения графиков $x^* \in [1; \frac{\pi}{2}]$. Покажем, что на отрезке $\Delta = [1; \frac{\pi}{2}]$ содержится единственный корень рассматриваемого уравнения, подтвердив «графические» соображения аналитическими выкладками. Имеем: $a = 1$; $b = \frac{\pi}{2}$. Тогда

$$f(a) = \sin 1 - 1 < 0, \quad f(b) = \sin \frac{\pi}{2} \cdot \frac{\pi}{2} - 1 = \frac{\pi}{2} - 1 > 0,$$



т.е.

$$f(a) \cdot f(b) < 0, f'(x) = \sin x + x \cos x > 0 \forall x \in [a, b],$$

поэтому в силу приведенной выше теоремы уравнение (4.2) на найденном отрезке имеет единственный корень.

При решении задачи об отделении корней, конечно же, нужно стремиться использовать конкретные свойства конкретной функции, задающей данное уравнение. Так, например, для алгебраических уравнений существуют аналитические методы, позволяющие установить количество вещественных корней того или иного знака, а также их границы (теоремы Штурма, Бюдана-Фурье, Декарта и т.п.).

Простым способом отделения корней, не связанным напрямую с построением графиков, является вычисление таблицы значений функции $f(x)$ на заданной сетке точек $x_k \in [a, b]$, $k = 0, 1, \dots$, и, таким образом, вычислительный поиск точки перемены знака.



4.1.2. Метод бисекции

После того как точка перемены знака функции, задающей уравнение, найдена, дальнейшее отделение корня, т.е. уменьшение длины отрезка Δ , на котором находится корень, может быть осуществлено с помощью **метода бисекции (дихотомии, половинного деления)**, который, по сути, является и простейшей итерационной процедурой решения задачи о вычислении корней с требуемой точностью. Опишем его.

Итак, пусть мы нашли такие точки x_0 и x_1 , что $f(x_0) \cdot f(x_1) < 0$. Найдем середину отрезка $x_2 = \frac{x_0+x_1}{2}$ и вычислим $f(x_2)$. Из двух половин отрезка выберем ту, для которой $f(x_2) \cdot f(x_{\text{гран}}) < 0$. Затем новый отрезок опять делим пополам и выбираем ту половину, на концах которой функция принимает значения разных знаков, и т.д.

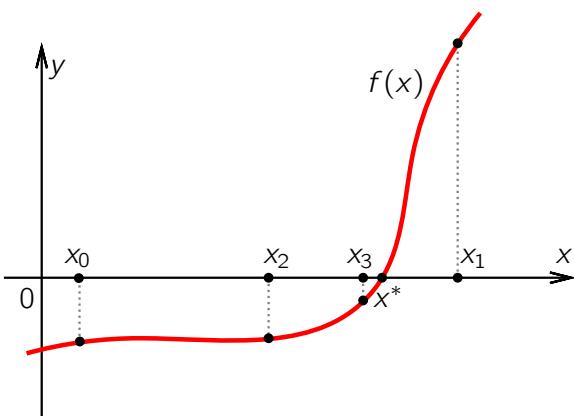


Рисунок 4.2

Если требуется найти корень с точностью ε , то продолжаем деление пополам до тех пор, пока длина отрезка не станет меньше 2ε . Тогда середина последнего отрезка даст значение корня с требуемой точностью.



Дихотомия проста и очень надежна: к простому корню она сходится для любых непрерывных функций $f(x)$, в том числе недифференцируемых; при этом она устойчива к ошибкам округления. Скорость сходимости невелика: за одну итерацию точность увеличивается примерно вдвое. Зато точность ответа гарантируется. Учитывая сказанное, можно сделать вывод: дихотомию следует применять тогда, когда требуется высокая надежность счета, а скорость сходимости малосущественна (либо как «прелюдию» к другим итерационным алгоритмам, к рассмотрению которых мы сейчас и переходим).



4.1.3. Метод простой итерации решения численных уравнений

Итак, вновь рассмотрим уравнение (4.2). Для применения метода (простой) итерации уравнение обычно приводят к виду

$$x = \varphi(x) \quad (4.3)$$

(будем говорить: к виду, удобному для итерации). Тогда решению этого уравнения можно придать следующий смысл: это неподвижная точка преобразования $y = \varphi(x)$. Способов приведения к виду (4.3) существует много. Часто для этих целей используют линейное преобразование вида $\varphi(x) = x + \psi(x)f(x)$, где $\psi(x)$ – некоторая известная непрерывная функция (по крайней мере, в окрестности отыскиваемого корня уравнения (4.1)) и не обращающаяся там в нуль. Можно также пытаться каким-либо образом выразить x из исходного уравнения; например, для уравнения

$$x^7 - 3x^2 + 1 = 0$$

это можно сделать так:

$$x = \sqrt[7]{3x^2 - 1},$$

т.е.

$$\varphi(x) = \sqrt[7]{3x^2 - 1}.$$

Будем считать, что корень x^* отделен и, таким образом, указано некоторое начальное приближение x_0 (вообще говоря, произвольное значение из отрезка отделенности). Тогда уточнение этого значения производят по [методу простой итерации](#)

$$x_{n+1} = \varphi(x_n), \quad n = 0, 1, \dots \quad (4.4)$$

Такой алгоритм обладает свойством самоисправляемости; действительно, здесь любое приближение номера n можно рассматривать как исходное приближение, поэтому отдельные сбои, допускаемые при вычислениях (естественно, это касается «ручного» счета), если они не выводят за пределы интервала сходимости, не отражаются на результатах, а влияют лишь на объем работы.



Очевидно, что успех применения алгоритма зависит не только от выбора x_0 , но и от выбора преобразования $\varphi(x)$. Исследуем грубое поведение ошибки приближенного решения в итерационном процессе (4.4). Это позволит нам предугадать ответы на два основных вопроса:

- 1) когда последовательность может быть построена;
- 2) сходимость последовательности приближений (условия и скорость сходимости).

Пусть мы построили приближение x_n к корню и его погрешность $\varepsilon_n = x^* - x_n$ достаточно мала по модулю. Выясним, как будет связана погрешность ε_n с погрешностью $\varepsilon_{n+1} = x^* - x_{n+1}$. Подставим x_n и x_{n+1} в (4.4):

$$x^* - \varepsilon_{n+1} = \varphi(x^* - \varepsilon_n).$$

Предполагая, что $\varphi(x)$ имеет непрерывную производную в окрестности точек x_n и x_{n+1} , разложим правую часть последнего равенства в ряд Тейлора, ограничившись двумя членами и взяв остаток в форме Пеано:

$$x^* - \varepsilon_{n+1} = \varphi(x^*) - \varphi'(x^*)\varepsilon_n + o(\varepsilon_n).$$

Так как $x^* = \varphi(x^*)$, то отсюда получим:

$$\varepsilon_{n+1} = \varphi'(x^*)\varepsilon_n + o(\varepsilon_n).$$

Если ε_n мало, то величиной $o(\varepsilon_n)$ в пределах принятой точности можно пренебречь. Тогда

$$\varepsilon_{n+1} \approx \varphi'(x^*)\varepsilon_n. \tag{4.5}$$

Пусть $\varphi'(x^*) \neq 0$. Тогда очевидно, что если $|\varphi'(x^*)| > 1$, то $|\varepsilon_{n+1}| > |\varepsilon_n|$. В таком случае рассчитывать на успех итераций не приходится, т.е. точка x^* будет точкой отталкивания для данного процесса. Если же $|\varphi'(x^*)| < 1$, то $|\varepsilon_{n+1}| < |\varepsilon_n|$ и можно ожидать сходимости процесса со скоростью геометрической прогрессии со знаменателем $|\varphi'(x^*)|$. Если $|\varphi'(x^*)| \ll 1$, то сходимость будет быстрой. Случай $|\varphi'(x^*)| = 1$ рассматривать не стоит, так как равенство (4.5) носит приближенный характер.



Если $0 < \varphi'(x^*) < 1$, то $\varphi'(x) > 0$ в некоторой окрестности корня и процесс будет носить монотонный характер. Если же $\varphi'(x^*) < 0$, то значения ε_n на соседних шагах итераций будут иметь противоположные знаки. Такой характер поведения ε_n особенно благоприятен для вычислений, так как позволяет высказать апостериорные суждения о точности результата (соседние приближения находятся по разные стороны от корня).

Пусть теперь $\varphi'(x^*) = 0$. Тогда $\varepsilon_{n+1} = o(\varepsilon_n)$. В этом случае можно рассчитывать на очень быструю сходимость (более быструю, чем сходимость геометрической прогрессии со сколь угодно малым знаменателем). Этот факт можно использовать для улучшения метода итерации.

Выясним, как будут связаны между собой ε_{n+1} и ε_n , если $\varphi'(x^*) = 0$. Эта связь существенно зависит от кратности корня x^* , с которой он удовлетворяет уравнению $\varphi'(x^*) = 0$.

Пусть эта кратность равна $k - 1$, т. е.

$$\varphi'(x^*) = \varphi''(x^*) = \dots = \varphi^{(k-1)}(x^*) = 0, \quad \varphi^{(k)}(x^*) \neq 0.$$

Итерации такого вида будем называть *итерациями порядка k*. В этом случае имеем:

$$x_{n+1} = x^* - \varepsilon_{n+1} = \varphi(x_n) = \varphi(x^* - \varepsilon_n) = \varphi(x^*) + (-1)^k \frac{\varphi^{(k)}(x^*)}{k!} \varepsilon_n^k + o(\varepsilon_n^k),$$

т.е.

$$\varepsilon_{n+1} \approx (-1)^{k+1} \frac{\varphi^{(k)}(x^*)}{k!} \varepsilon_n^k.$$

Отсюда

$$\frac{\varepsilon_{n+1}}{\varepsilon_n^k} \approx \frac{\varepsilon_n}{\varepsilon_{n-1}^k},$$

или

$$\frac{\varepsilon_{n+1}}{\varepsilon_n} \approx \left(\frac{\varepsilon_n}{\varepsilon_{n-1}} \right)^k.$$

Таким образом, вблизи корня количество верных значащих цифр результата на каждой итерации возрастает в k раз.

Проведенные рассуждения наглядны, но, вообще говоря, не являются абсолютно строгими. Кроме того, они ничего не говорят о том, каким должен быть выбор x_0 . Приведем сейчас утверждение, свободное от указанных недостатков.



Теорема 4.2 (о сходимости метода итерации). *Если:*

- 1) $\varphi(x)$ определена и непрерывна в области $\Delta = \{x : |x - x_0| \leq \delta\}$;
- 2) в этой области $\varphi(x)$ удовлетворяет условию Липшица $|\varphi(x') - \varphi(x'')| \leq q|x' - x''|$ для любых $x', x'' \in \Delta$ с константой $q < 1$;
- 3) справедливо неравенство $\frac{m}{1-q} \leq \delta$, где $|x_0 - \varphi(x_0)| \leq m$,

то:

- 1) последовательность приближений $\{x_n\}$ по [методу итераций](#) с исходным приближением x_0 может быть построена;
- 2) в указанной области уравнение $x = \varphi(x)$ имеет решение и притом единственное;
- 3) построенная последовательность приближений $\{x_n\}$ с ростом n сходится к этому решению;
- 4) скорость сходимости характеризуется неравенством

$$|x^* - x_n| \leq \frac{m}{1-q} q^n \quad (4.6)$$

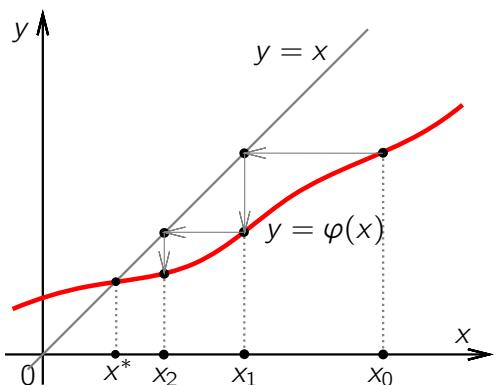
[\[Доказательство\]](#)

Замечание 4.1. При доказательстве единственности не использовалось третье условие теоремы; Второе условие теоремы можно заменить более строгим, но практически легче проверяемым: в области Δ функция $\varphi(x)$ имеет непрерывную первую производную $\varphi'(x)$ такую, что $|\varphi'(x)| < 1$ для любого $x \in \Delta$.

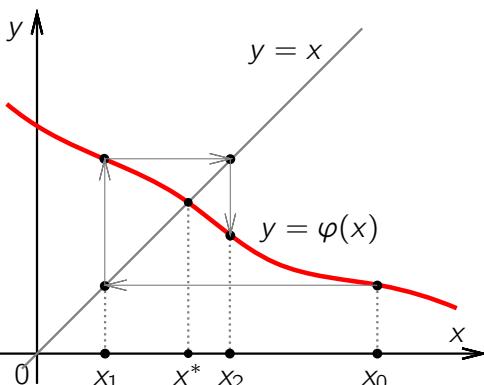
Рассмотрим геометрическую интерпретацию метода итерации. Считаем, что x и $\varphi(x)$ действительны. Пусть также $|\varphi'(x)| \leq q < 1$. Тогда возможны два случая:



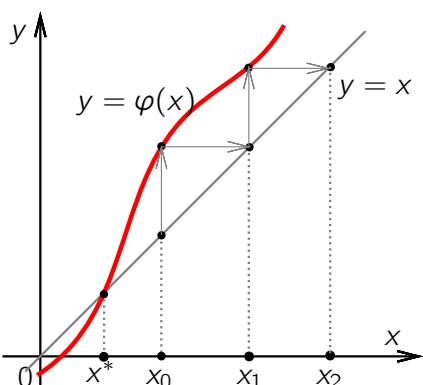
a) $0 \leq \varphi'(x) < 1$;



б) $-1 < \varphi'(x) \leq 0$.



Условие $|\varphi'(x)| < 1$ существенно. В противном случае (например, при $\varphi'(x) > 1$) картинка будет выглядеть следующим образом:





Будем считать, что процесс итерации является сходящимся. Тогда важен вопрос, когда его можно прервать, т.е. когда достигается заданная точность приближенного решения.

Конечно, зная все параметры из теоремы о сходимости и пользуясь оценкой (4.6), легко найти априорную (гарантированную (!)) оценку числа итераций:

$$n \geq \frac{\ln \frac{\epsilon(1-q)}{m}}{\ln q}. \quad (4.7)$$

С другой стороны, для тех же целей чаще пользуются результатами вычислений: процесс прерывается, если в пределах принятой точности два соседних итерационных приближения совпадают, т.е. если выполняется неравенство $|x_{n+1} - x_n| \leq \epsilon$ (часто рассматривается также аналог относительной погрешности $\frac{|x_{n+1} - x_n|}{|x_{n+1}|} \leq \epsilon$).

Однако этим правилом с достоверностью можно пользоваться лишь в случае отрицательной производной. Если же $0 \leq \varphi'(x^*) < 1$, то правилом с уверенностью пользоваться можно лишь при $q \leq \frac{1}{2}$. В случае же $\frac{1}{2} < q < 1$ оно будет давать сбои.

Обоснемуем указанные наблюдения.

Лемма 4.1.

$$|x^* - x_n| \leq \frac{q}{1-q} |x_n - x_{n-1}|. \quad (4.8)$$

[\[Доказательство\]](#)

Из леммы следует, что для достоверности предложенного выше апостериорного способа оценки погрешности (как модуля разности двух соседних приближений) необходимо требовать выполнения неравенства

$$\left| \frac{q}{1-q} |x_n - x_{n-1}| \right| \leq |x_n - x_{n-1}|,$$

откуда непосредственно получаем: оценка достоверна, если $q \leq \frac{1}{2}$.

Заметим также, что оценка (4.8) неулучшаема, поскольку она достигается, например, при $\varphi_1(x) = q(x - x^*) + \varphi(x^*)$.



4.1.4. О задаче улучшения метода итерации

Метод Стеффенсена

Улучшение итерационного процесса при помощи преобразования заданного уравнения

Как мы уже отмечали ранее, скорость сходимости **метода итерации** — это скорость сходимости геометрической прогрессии со знаменателем q (или $\varphi'(x^*)$). Таким образом, при приведении уравнения к каноническому виду желательно действовать таким образом, чтобы $\varphi'(x^*)$ имело по возможности меньшее по абсолютной величине значение. Чуть позже мы остановимся более подробно на некоторых способах, позволяющих достичь определенных успехов в данном направлении.

Сейчас же заметим пока, что если в общей схеме приведения к каноническому виду положить $\psi(x) = C$, то получим:

$$\varphi(x) = x + Cf(x),$$

откуда

$$\varphi'(x) = 1 + Cf'(x).$$

Если x^* не является кратным корнем исходного уравнения, то $f'(x^*) \neq 0$ и поэтому в окрестности точки x^* производная $f'(x)$ сохраняет знак. Тогда для выбора константы C получаем условие

$$-1 < 1 + Cf'(x) < 1$$

или

$$-2 < Cf'(x) < 0.$$

Отсюда следует, что если при всех $x \in \Delta$ (Δ — отрезок **отделенности корня**) выполняется неравенство $0 < f'(x) < M$, то при любом $C \in (-\frac{2}{M}; 0)$ метод простой итерации будет сходиться. Аналогичный результат будет иметь место и в случае $M < f'(x) < 0$ с той лишь разницей, что $C \in (0; -\frac{2}{M})$.



Метод Стеффенсена

Тот факт, что последовательность приближений метода простой итерации близка к геометрической прогрессии, позволяет применить для ускорения ее сходимости преобразование Эйткена. Опишем эту процедуру подробнее. Пользуясь правилом (4.4), по приближению x_0 построим приближения $x_1 = \varphi(x_0)$ и $x_2 = \varphi(x_1) = \varphi(\varphi(x_0))$. Применив к трем числам x_0 , x_1 , x_2 преобразование Эйткена, получим:

$$x'_1 = \frac{x_2 x_0 - x_1^2}{x_2 - 2x_1 + x_0} = \frac{x_0 \varphi(\varphi(x_0)) - \varphi^2(x_0)}{\varphi(\varphi(x_0)) - 2\varphi(x_0) + x_0}.$$

Заменив соответствующим образом индексы, получим итерационный процесс

$$x_{n+1} = \frac{x_n \varphi(\varphi(x_n)) - \varphi^2(x_n)}{\varphi(\varphi(x_n)) - 2\varphi(x_n) + x_n}, \quad (4.9)$$

который в литературе носит название *метода Стеффенсена*.

С формальной точки зрения метод Стеффенсена эквивалентен методу простой итерации

$$x_{n+1} = \Phi(x_n),$$

где

$$\Phi(x) = \frac{x \varphi(\varphi(x)) - \varphi^2(x)}{\varphi(\varphi(x)) - 2\varphi(x) + x}.$$

При этом условия сходимости такого итерационного процесса оказываются значительно более благоприятными, нежели условия сходимости процесса, на основе которого он построен. Справедлива

Лемма 4.2. *Если для функции $\varphi(x)$ вблизи точки x^* справедливо представление*

$$\varphi(x) = x^* + \alpha(x - x^*) + o(x - x^*), \quad \alpha \neq 0, \quad \alpha \neq 1, \quad (4.10)$$

то

$$\Phi(x) = x^* + o(x - x^*). \quad (4.11)$$

[[Доказательство](#)]



Из доказанной леммы следует, что функция $\Phi(x)$ непрерывна в окрестности точки x^* , причем

$$\lim_{x \rightarrow x^*} \Phi(x) = x^*.$$

Если теперь доопределить ее, положив $\Phi(x^*) = x^*$, то $\Phi(x)$ будет непрерывна в точке x^* и, кроме того,

$$\Phi'(x^*) = \lim_{x \rightarrow x^*} \frac{\Phi(x) - \Phi(x^*)}{x - x^*} = 0.$$

Полученный результат позволяет надеяться на то, что сходимость метода Стеффенсена в условиях леммы 4.2 будет квадратичной (это будет иметь место в случае, если существует непрерывная производная $\Phi'(x)$ в окрестности точки x^*).

Упражнения.

- 1) Доказать, что в условиях леммы 1 в окрестности точки x^* существует непрерывная производная $\Phi'(x)$.
- 2) Доказать, что в условиях леммы 1 метод Стеффенсена будет сходиться и в случае $\alpha = 1$.
- 3) Доказать, что если исходный итерационный процесс с функцией $\varphi(x)$ имеет порядок $k > 1$, то метод Стеффенсена будет иметь порядок не менее $2k - 1$.

Улучшение итерационного процесса при помощи преобразования заданного уравнения

Рассмотрим сейчас вопрос о том, как практически можно реализовать идею построения итерационного процесса (выбора $\varphi(x)$) таким образом, чтобы обеспечить обращение в 0 производных от функции $\varphi(x)$ в точке x^* . Вновь вернемся к описанному ранее способу приведения к каноническому виду:

$$\varphi(x) = x + \psi(x) f(x). \quad (4.12)$$

Подберем сейчас функцию S_n таким образом, чтобы удовлетворить условию

$$\varphi'(x^*) = \varphi'(x)|_{f(x)=0} = 0.$$

Имеем:

$$\varphi'(x)|_{f(x)=0} = 1 + \psi'(x) f(x) + \psi(x) f'(x)|_{f(x)=0} = 1 + \psi(x) f'(x) = 0.$$



Отсюда следует (при условии $f'(x^*) \neq 0$), что $\psi(x) = -\frac{1}{f'(x)}$.

Соответствующий итерационный процесс имеет вид

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad n = 0, 1, \dots \quad (4.13)$$

Он имеет второй порядок и носит название *метода Ньютона* (более подробно этот метод разобран в разделе 4.1.5).

Чтобы обеспечить более чем двукратное увеличение порядка итераций, линейным преобразованием уже не обойтись. Естественным обобщением (4.12) служит формула

$$\varphi(x) = x + \psi_1(x)f(x) + \psi_2(x)f^2(x) + \dots + \psi_{k-1}(x)f^{k-1}(x) \quad (4.14)$$

Здесь коэффициенты $\psi_i(x)$ можно трактовать как коэффициенты разложения функции $\Psi(x) = \Phi(x, f(x))$ в ряд Маклорена по второму аргументу при ограничении $\psi_0(x) = x$. Задача состоит в выборе этих коэффициентов. Будем требовать, чтобы выполнялись условия

$$\varphi^{(i)}(x) \Big|_{f(x)=0} = 0, \quad i = 1, \dots, k-1. \quad (4.15)$$

Эти условия приводят к системе равенств, из которых коэффициенты $2|h_n|$ могут быть найдены последовательно:

$$\varphi'(x)|_{f(x)=0} = 1 + \psi'_1(x)f(x) + \psi_1(x)f'(x) + \psi'_2(x)f^2(x) + 2\psi_2(x)f(x)f'(x) + \dots |_{f(x)=0} =$$

$$= 1 + \psi_1(x)f'(x) = 0,$$

$$\varphi''(x)|_{f(x)=0} = \psi''_1(x)f(x) + 2\psi'_1(x)f'(x) + \psi_1(x)f''(x) + 2\psi_2(x)(f'(x))^2 + \dots |_{f(x)=0} =$$

$$= 2\psi'_1(x)f'(x) + \psi_1(x)f''(x) + 2\psi_2(x)(f'(x))^2 = 0,$$



.....

Многоточиями в каждой из формул заменены те слагаемые, которые содержат в качестве сомножителя $f'(x)$ либо какую-то из степеней последней. Из данной системы непосредственно находим:

$$\psi_1(x) = -\frac{1}{f'(x)};$$

$$\psi_2(x) = -\frac{\psi'_1(x)f'(x)}{(f'(x))^2} - \frac{\psi_1(x)f''(x)}{2(f'(x))^2} = \left[\psi'_1(x) = \frac{f''(x)}{(f'(x))^2} \right] = -\frac{f''(x)f'(x)}{(f'(x))^4} +$$

$$+\frac{f''(x)}{2(f'(x))^3} = -\frac{f''(x)}{2(f'(x))^3}.$$

Таким образом, итерационный процесс третьего порядка будет иметь вид

$$\varphi^{(i)}(x) \Big|_{f(x)=0} = 0, \quad i = 1, \dots, k-1, \quad x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} - \frac{f^2(x_n)f''(x_n)}{2(f'(x_n))^3}. \quad (4.16)$$

В литературе описанный способ построения итерационных процессов высших порядков носит название [метода Чебышева](#). Существуют и другие способы решения рассмотренной задачи.

Упражнение. Построить примеры итерационных процессов чебышевского типа выше третьего порядка.



4.1.5. Метод Ньютона

Вновь возвращаемся к уравнению в исходной форме (4.1). Предположим, что каким-либо способом получено приближение x_n к корню x^* ($n \geq 0$). Погрешность данного приближения, как и ранее, обозначим через ε_n : $\varepsilon_n = x^* - x_n$. Очевидно, при известном x_n отыскание корня равносильно отысканию погрешности. Имеем:

$$f(x_n + \varepsilon_n) = 0.$$

ε_n обычно в сравнении с x^* есть величина малая по модулю. Это делает последнее уравнение более благоприятным с точки зрения вычислений. Разложив левую часть в ряд Тейлора, получим:

$$0 = f(x_n) + \frac{\varepsilon_n}{1!} f'(x_n) + \dots + \frac{\varepsilon_n^k}{k!} f^{(k)}(x_n) + O(\varepsilon_n^{k+1}).$$

Если теперь отбросить остаточный член $O(\varepsilon_n^{k+1})$, получим приближенное алгебраическое уравнение

$$f(x_n) + \frac{\varepsilon_n}{1!} f'(x_n) + \dots + \frac{\varepsilon_n^k}{k!} f^{(k)}(x_n) \approx 0 \quad (4.17)$$

(здесь мы вновь четко видим ситуацию: замена исходной задачи другой, в каком-то смысле более простой).

Решив это уравнение (если мы умеем это делать), мы, вообще говоря, не найдем само значение ε_n , а получим лишь некоторое приближенное значение (поправку) Δx_n . Прибавив эту поправку к x_n , получим новое приближение:

$$x_{n+1} = x_n + \Delta x_n.$$

Можно ожидать, что оно будет более близким к $x^* \in S_n$, чем x_n . Далее описанную процедуру можно повторить. При удачном выборе x_0 и k можно рассчитывать на достаточно быструю сходимость.

Недостаток этого подхода состоит, вообще говоря, в необходимости на каждом шаге решать алгебраическое уравнение степени k , что само по себе представляет вовсе не простую задачу. Поэтому чаще всего данный подход используют при $k = 1$. В этом случае из (4.17) имеем:

$$f(x_n) + f'(x_n) \Delta x_n = 0,$$



откуда $\Delta x_n = -\frac{f(x_n)}{f'(x_n)}$ или

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad n = 0, 1, \dots \quad (4.18)$$

Мы получили [метод Ньютона](#) (в разделе 4.1.4 он был получен из других соображений). Так как поправка Δx_n находится из линейного уравнения, то метод Ньютона называют также [методом линеаризации](#).

Рассмотрим геометрическую интерпретацию метода. Для этого к кривой $y = f(x)$ в точке с абсциссой x_n проведем касательную. Ее уравнение имеет вид

$$y - f(x_n) = f'(x_n)(x - x_n). \quad (4.19)$$

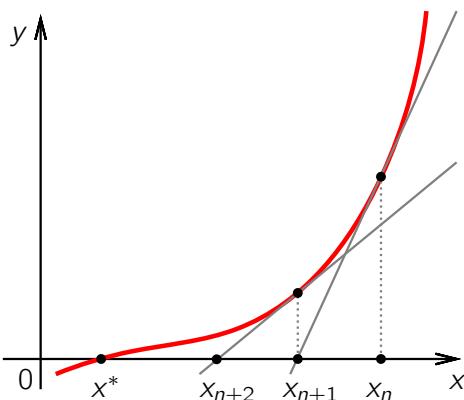


Рисунок 4.3

Найдем абсциссу точки пересечения касательной с осью Ox . Полагая в уравнении (4.19) $y = 0$, получим:

$$x = x_n - \frac{f(x_n)}{f'(x_n)}.$$



Таким образом, в методе Ньютона мы вместо приближения к корню по кривой приближаемся к нему по последовательности касательных прямых (отсюда – еще одно название: метод касательных).

Проведем исследование метода Ньютона по той же схеме, которой мы пользовались для метода (простой) итерации.

Вначале обсудим вопрос о выборе x_0 , а именно: если корень отделен, то что следует брать в качестве начального приближения? Ответ на этот вопрос может быть получен из геометрических соображений: начальное приближение x_0 целесообразно выбирать так, чтобы выполнялось неравенство (условие Фурье)

$$f(x_0) \cdot f''(x_0) > 0. \quad (4.20)$$

В противном случае, как это видно из [рисунка 4.4](#), может возникнуть ситуация, при которой уже первое приближение «вылетает» за пределы отрезка отделенности.

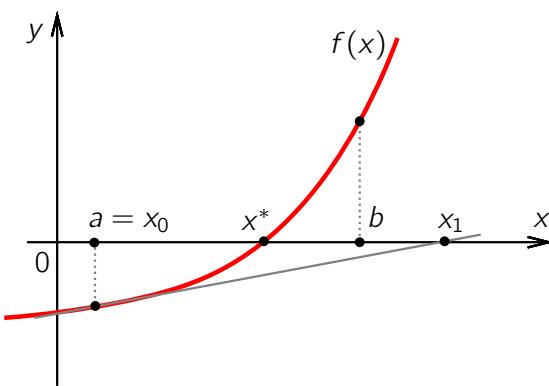


Рисунок 4.4

Что касается вопроса о прекращении итерационного процесса по совпадению двух соседних приближений с требуемой точностью, то это правило, как и в случае метода итераций, не является абсолютно строгим.



С чисто формальной точки зрения метод Ньютона можно трактовать как [метод простой итерации](#) с выбором функции $\varphi(x)$ в виде

$$\varphi(x) = x - \frac{f(x)}{f'(x)},$$

и поэтому формально все результаты раздела [4.1.3](#) могут быть перенесены на этот случай (так, например, формальное достаточное условие сходимости метода Ньютона может выглядеть следующим образом: $|\varphi'(x)| = \left| \frac{f(x)f''(x)}{[f'(x)]^2} \right| < 1$ при $x \in \Delta$).

Мы, однако, вновь проведем исследование поведения погрешности вблизи корня. Легко видеть, что некоторые «неприятности» могут возникнуть в случае $f'(x^*) = 0$ (ведь значение $f'(x_n)$ стоит в знаменателе формулы [\(4.18\)](#)). Посмотрим, что в этом случае будет происходить. Для погрешности $\varepsilon_n = x^* - x_n$ из [\(4.18\)](#) получим уравнение

$$\varepsilon_{n+1} = \varepsilon_n + \frac{f(x^* - \varepsilon_n)}{f'(x^* - \varepsilon_n)}. \quad (4.21)$$

Если x^* – корень исходного уравнения кратности m , то

$$f(x) = (x - x^*)^m a_m + (x - x^*)^{m+1} a_{m+1} + \dots,$$

$$f'(x) = m(x - x^*)^{m-1} a_m + (m+1)(x - x^*)^m a_{m+1} + \dots,$$

где $a_k = \frac{1}{k!} f^{(k)}(x^*)$, $k = m, m+1, \dots$. Тогда

$$f(x^* - \varepsilon_n) \approx (-1)^m \varepsilon_n^m a_m \left(1 - \varepsilon_n \frac{a_{m+1}}{a_m} \right),$$

$$f'(x^* - \varepsilon_n) \approx (-1)^{m-1} m \varepsilon_n^{m-1} a_m \left(1 - \varepsilon_n \frac{m+1}{m} \frac{a_{m+1}}{a_m} \right),$$

$$\frac{1}{f'(x^* - \varepsilon_n)} \approx \frac{(-1)^{m-1}}{m \varepsilon_n^{m-1} a_m} \cdot \frac{1}{1 - \varepsilon_n \frac{m+1}{m} \frac{a_{m+1}}{a_m}} \approx \frac{(-1)^{m-1}}{m \varepsilon_n^{m-1} a_m} \left(1 + \varepsilon_n \frac{m+1}{m} \frac{a_{m+1}}{a_m} \right).$$

Подставляя полученные разложения в [\(4.21\)](#), будем иметь:

$$\frac{f(x^* - \varepsilon_n)}{f'(x^* - \varepsilon_n)} \approx -\frac{\varepsilon_n}{m} \left(1 - \varepsilon_n \frac{a_{m+1}}{a_m} \right) \left(1 + \varepsilon_n \frac{m+1}{m} \frac{a_{m+1}}{a_m} \right) \approx -\frac{\varepsilon_n}{m} \left(1 + \frac{\varepsilon_n}{m} \frac{a_{m+1}}{a_m} \right)$$



и

$$\varepsilon_{n+1} \approx \varepsilon_n - \frac{\varepsilon_n}{m} \left(1 + \frac{\varepsilon_n}{m} \frac{a_{m+1}}{a_m} \right) = \varepsilon_n \left(1 - \frac{1}{m} \right) - \varepsilon_n^2 \cdot \frac{1}{m^2} \frac{a_{m+1}}{a_m}.$$

Отсюда, если $m = 1$, то

$$\varepsilon_{n+1} \approx -\varepsilon_n^2 \frac{a_2}{a_1} = -\frac{1}{2} \frac{f''(x^*)}{f'(x^*)} \varepsilon_n^2,$$

т.е. $\varepsilon_{n+1} = O(\varepsilon_n^2)$ и следует ожидать квадратичной сходимости.

Если же $m > 1$, то $\varepsilon_{n+1} \approx -\varepsilon_n \left(1 - \frac{1}{m} \right)$, т.е. между ε_{n+1} и ε_n теперь существует линейная связь. Чем больше m , тем ближе $q = 1 - \frac{1}{m}$ к 1 и, следовательно, тем медленнее сходимость метода. Докажем теперь строгое утверждение о сходимости метода Ньютона.

Теорема 4.3. Если:

- 1) функция $f(x)$ определена и дважды непрерывно дифференцируема на отрезке S_0 вещественной оси с концами в точках x_0 и $x_0 + 2h_0$, где $h_0 = -\frac{f(x_0)}{f'(x_0)}$, при этом на концах отрезка S_0 $f(x)f'(x) \neq 0$;
- 2) выполняется неравенство $2|h_0| M \leq |f'(x_0)|$, где $M = \max_{x \in S_0} |f''(x)|$,

то:

- 1) внутри отрезка S_0 уравнение $f(x) = 0$ имеет корень x^* и при этом единственный;
- 2) последовательность приближений $\{x_n\}$, $n = 1, 2, \dots$ к корню x^* этого уравнения может быть построена по методу Ньютона с начальным приближением x_0 ;
- 3) последовательность приближений сходится к этому корню;
- 4) скорость сходимости характеризуется неравенством

$$|x^* - x_{n+1}| \leq |x_{n+1} - x_n| \leq \frac{M}{2|f'(x_n)|} |x_n - x_{n-1}|^2, \quad n = 1, 2, \dots \quad (4.22)$$

[\[Доказательство\]](#)



Отметим теперь, что оценка (4.22) как раз и означает квадратичную сходимость, причем, как видим, практическая точность вычислений обеспечивается за счет сравнения двух соседних приближений. Из оценки (4.22) можно получить следующую априорную оценку:

$$\begin{aligned} |x^* - x_{n+1}| &\leq |x_{n+1} - x_n| \leq \alpha |x_n - x_{n-1}|^2 \leq \alpha \alpha^2 |x_{n-1} - x_{n-2}|^2 \leq \dots \leq \\ &\leq \alpha \alpha^2 \alpha^2 \dots \alpha^{2^{n-1}} |x_1 - x_0|^{2^n} = \alpha^{2^n - 1} |x_1 - x_0|^{2^n} = \frac{1}{\alpha} (\alpha |x_1 - x_0|)^{2^n}, \end{aligned} \quad (4.23)$$

где $\alpha = \max_{x \in S_0} \left| \frac{f''(x)}{2f'(x)} \right|$. О таком максимуме можно говорить, так как $|f'(x)| > 0$.

Отсюда количество итераций, необходимое для достижения требуемой точности, удовлетворяет неравенству

$$n \geq \log_2 \frac{\ln(\alpha \epsilon)}{\ln(\alpha |x_1 - x_0|)} \quad (4.24)$$

(сравните с (4.8)).

Отметим также, что (4.23) очень похожа на ту оценку, которая получается из приближенного равенства, связывающего погрешности на двух соседних шагах.

При определенных условиях метод Ньютона дает возможность построить монотонную последовательность приближений. Эти условия дают следующая

Теорема 4.4. Пусть $f(x) \in C^2[a, b]$, и на $[a, b]$ уравнение $f(x) = 0$ имеет единственный корень. Если для любого $x \in [a, b]$ выполняется неравенство $f'(x)f''(x) > 0$, причем $f'(x)$ и $f''(x)$ сохраняют знак, то последовательность $\{x_n\}$, построенная по методу Ньютона при $x_0 = b$, монотонно убывает и сходится к x^* .

[\[Доказательство\]](#)

Замечание 4.2. Аналогичное утверждение имеет место и в случае $f'(x)f''(x) < 0$, $x_0 = a$ при сохранении всех остальных условий теоремы 2, только $\{x_n\}$ в этом случае будет монотонно возрастающей.

Замечание 4.3. Метод Ньютона может применяться и для отыскания комплексных корней уравнения $f(z) = 0$ с сохранением обозначенных свойств алгоритма.



4.1.6. Видоизменения метода Ньютона

Случай кратных корней

Упрощение вычислений

Дискретный вариант метода Ньютона

Метод секущих

Метод хорд и комбинированные методы

Случай кратных корней

Анализируя формулу, связывающую погрешность двух соседних приближений метода Ньютона в случае корня кратности m , легко получить алгоритм, который будет обладать квадратичной сходимостью и в этом случае. Действительно, рассматривая алгоритм

$$x_{n+1} = x_n - k \frac{f(x_n)}{f'(x_n)}, \quad n = 0, 1, \dots \quad (4.25)$$

где k – параметр, подлежащий определению, для погрешностей получим:

$$\varepsilon_{n+1} \approx \left(1 - \frac{k}{m}\right) \varepsilon_n - \varepsilon_n^2 \frac{k}{m^2} \frac{a_{m+1}}{a_m}.$$

Если теперь положить $k = m$, то

$$\varepsilon_{n+1} \approx -\varepsilon_n^2 \frac{1}{m} \frac{a_{m+1}}{a_m}$$

и сходимость снова будет квадратичной.

Упрощение вычислений



Заманчиво отказаться от вычисления последовательности $f'(x_n)$ с тем, чтобы уменьшить объем работы. Вычислим только одно значение — $f'(x_0)$. Такое видоизменение называется *методом Ньютона с постоянной производной (касательной)* и имеет вид

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)}, \quad n = 0, 1, \dots \quad (4.26)$$

Его геометрическая интерпретация приведена на рисунке (смысл ее состоит в том, что приближение к корню осуществляется по семейству прямых, параллельных касательной в точке, абсцисса которой есть начальное приближение к корню).

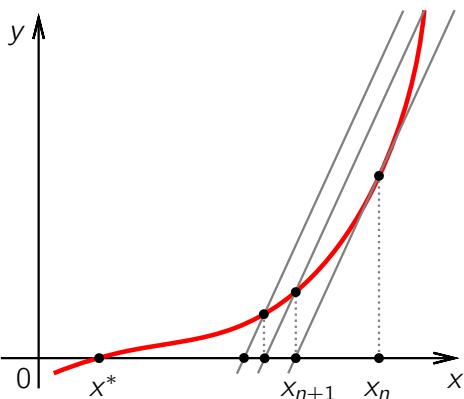


Рисунок 4.5

Исследование поведения погрешности вблизи корня дает

$$\varepsilon_{n+1} = \varepsilon_n + \frac{f(x^* - \varepsilon_n)}{f'(x_0)},$$

откуда $\varepsilon_{n+1} \approx \left(1 - \frac{f'(x^*)}{f'(x_0)}\right) \varepsilon_n$.



Скорость сходимости определяется, как видим, числом $q = 1 - \frac{f'(x^*)}{f'(x_0)}$. Обычно $\frac{f'(x^*)}{f'(x_0)}$ близко к единице и q мало. Тем не менее, скорость сходимости – линейная.

Дискретный вариант метода Ньютона

В случае, если аналитическое вычисление производной $f'(x_n)$ по каким-либо причинам является нежелательным, используют ее замену с помощью разностного отношения. При этом, выбрав некоторое приращение $h_n \approx \varepsilon_M$, можем записать

$$f'(x_n) = \frac{f(x_n + h_n) - f(x_n)}{h_n} + O(h_n).$$

Таким образом, вместо (4.18) получим формулу

$$x_{n+1} = x_n - \frac{f(x_n) h_n}{f(x_n + h_n) - f(x_n)}, \quad n = 0, 1, \dots \quad (4.27)$$

Это *дискретный вариант метода Ньютона*. Можно показать, что сходимость в этом случае остается квадратичной. Заметим также, что приращение h_n можно брать не зависящим от номера итерации. Платой за не вычисление производной (аналитически) является увеличение объема работы примерно в два раза.

Метод секущих

В этом случае в качестве приращения h_n из дискретного варианта метода Ньютона используется разность двух предыдущих итерационных приближений, т.е. производная вычисляется по формуле

$$f'(x_n) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}, \quad n = 1, 2, \dots$$

и, следовательно, алгоритм метода принимает вид

$$x_{n+1} = x_n - f(x_n) \cdot \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}, \quad n = 1, 2, \dots \quad (4.28)$$

Этот алгоритм требует предварительного вычисления двух значений – x_1 и x_0 , и только после этого можно продолжать вычисления. Подобные алгоритмы, как мы уже отмечали ранее, называют двухшаговыми



(по аналогии можно рассматривать и k -шаговые алгоритмы). При выборе начальных приближений также необходимо придерживаться условий типа Фурье (см. неравенство (4.20)).

Заметим, что формулу (4.28) можно переписать и в другом, математически эквивалентном, виде:

$$x_{n+1} = \frac{x_{n-1}f(x_n) - x_nf(x_{n-1})}{f(x_n) - f(x_{n-1})}, \quad n = 1, 2, \dots$$

Однако эта запись менее привлекательна, чем (4.28) ввиду того, что мы ведем вычисления с округлением: если x_{n-1} и x_n близки между собой, то неизбежны потери значащих цифр при вычислениях, поэтому всегда следует отдавать предпочтение алгоритмам, работающим с поправками.

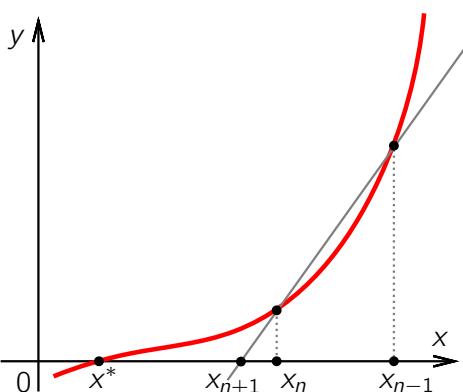


Рисунок 4.6

Рассмотрим геометрическую интерпретацию метода (4.28). Запишем уравнение прямой, проходящей через точки $(x_{n-1}, f(x_{n-1}))$ и $(x_n, f(x_n))$

$$\frac{y - f(x_n)}{x - x_n} = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}, \quad (*)$$



и найдем точку ее пересечения с осью Ox . Полагая в (*) $y = 0$, получим:

$$x = x_n - f(x_n) \cdot \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}.$$

Таким образом, здесь к корню мы приближаемся по секущей, проходящей через точки двух предыдущих приближений. В силу этого метод (4.28) называют еще [методом секущих](#).

Выясним картину поведения погрешности при таком видоизменении метода Ньютона. Для этого установим связь между погрешностями ε_{n+1} , ε_n и ε_{n-1} . Имеем следующее уравнение, их связывающее:

$$\varepsilon_{n+1} = \varepsilon_n - f(x^* - \varepsilon_n) \cdot \frac{\varepsilon_n - \varepsilon_{n-1}}{f(x^* - \varepsilon_n) - f(x^* - \varepsilon_{n-1})}, \quad (**)$$

Разложим $f(x_n)$ и $f(x_{n-1})$ в ряд Тейлора (предполагая, что x^* – простой корень):

$$f(x^* - \varepsilon_n) \approx -\varepsilon_n f'(x^*) + \frac{\varepsilon_n^2}{2} f''(x^*),$$

$$f(x^* - \varepsilon_{n-1}) \approx -\varepsilon_{n-1} f'(x^*) + \frac{\varepsilon_{n-1}^2}{2} f''(x^*).$$

Подставляя эти выражения в (**), получим:

$$\varepsilon_{n+1} \approx \varepsilon_n - \frac{(\varepsilon_n - \varepsilon_{n-1}) \varepsilon_n [f'(x^*) - \frac{\varepsilon_n}{2} f''(x^*)]}{(\varepsilon_n - \varepsilon_{n-1}) f'(x^*) - \frac{1}{2} (\varepsilon_n^2 - \varepsilon_{n-1}^2) f''(x^*)} =$$

$$= \varepsilon_n - \frac{\varepsilon_n [f'(x^*) - \frac{\varepsilon_n}{2} f''(x^*)]}{f'(x^*) - \frac{1}{2} (\varepsilon_n + \varepsilon_{n-1}) f''(x^*)} =$$

$$= \frac{-\frac{1}{2} \varepsilon_n \varepsilon_{n-1} f''(x^*)}{f'(x^*) - \frac{1}{2} (\varepsilon_n + \varepsilon_{n-1}) f''(x^*)} \approx [f'(x^*) \neq 0] \approx -\frac{1}{2} \varepsilon_n \varepsilon_{n-1} \frac{f''(x^*)}{f'(x^*)}.$$

Отсюда видно, что этот алгоритм ближе к [методу Ньютона](#), чем алгоритм с [постоянной производной](#). Заметим, что если бы корень x^* был кратным, то ε_{n+1} была бы только порядка ε_n .



Итак, для метода Ньютона имеем:

$$\varepsilon_{n+1} \approx -\frac{1}{2} \frac{f''(x^*)}{f'(x^*)} \varepsilon_n^2, \quad (4.29)$$

Для метода секущих имеем:

$$\varepsilon_{n+1} \approx -\frac{1}{2} \frac{f''(x^*)}{f'(x^*)} \varepsilon_n \varepsilon_{n-1} \quad (4.30)$$

Соотношения (4.29), (4.30) представляют собой нелинейные разностные уравнения. Если соотношения типа (4.29) мы как-то исследовали ранее, то (4.30) встречается впервые.

Исследуем сейчас эти соотношения более подробно с единых позиций. Пусть $E_n \approx |\varepsilon_n|$, $\alpha = \frac{1}{2} \left| \frac{f''(x^*)}{f'(x^*)} \right|$. Тогда вместо (4.30) получим соотношение

$$E_{n+1} = \alpha E_n E_{n-1}, \quad (***)$$

Специальный вид нелинейности позволяет легко перейти к линейному разностному уравнению. Прологарифмировав (***) и введя обозначения $a = \ln \alpha$, $\xi_n = \ln E_n$, получим:

$$\xi_{n+1} - \xi_n - \xi_{n-1} = a.$$

Найдем его общее решение: частное решение, очевидно, имеет вид $\xi_n^* = -a$, а корни соответствующего характеристического уравнения $\lambda^2 - \lambda - 1 = 0$ равны $\lambda_1 = \frac{1+\sqrt{5}}{2}$, $\lambda_2 = \frac{1-\sqrt{5}}{2}$.

Тогда общее решение определяется формулой

$$\xi_n = C_1 \lambda_1^n + C_2 \lambda_2^n - a, \quad n = 0, 1, \dots$$

(здесь C_i – произвольные постоянные, определяемые из начальных условий). Возвращаясь к первоначальным переменным, получим:

$$|\varepsilon_n| \approx E_n = \frac{1}{\alpha} e^{C_1 \lambda_1^n} \cdot e^{C_2 \lambda_2^n}, \quad n = 0, 1, \dots \quad (4.31)$$



Формула (4.31) показывает, что сходимость метода секущих определяется условием $C_1 < 0$ (учитывая численные значения λ_1 и λ_2). Найдем C_1 . Записывая формулу общего решения при $n = 0$ и $n = 1$, получим:

$$\begin{cases} C_1\lambda_1 + C_2\lambda_2 = a + \xi_1, \\ C_1 + C_2 = a + \xi_0. \end{cases}$$

Отсюда $C_1 = \frac{a + \xi_1 - (a + \xi_0)\lambda_2}{\lambda_1 - \lambda_2} = \frac{a + \xi_1 - (a + \xi_0)\lambda_2}{\sqrt{5}}$.

Тогда условие $C_1 < 0$ дает

$$a + \xi_1 - (a + \xi_0)\lambda_2 < 0$$

или, учитывая, что $\lambda_1\lambda_2 = -1$,

$$(a + \xi_1)\lambda_1 + (a + \xi_0) < 0.$$

Отсюда следует:

$$\xi_0 + \xi_1\lambda_1 + a(1 + \lambda_1) < 0.$$

Возвращаясь к прежним обозначениям, получим

$$\ln E_0 + \lambda_1 \ln E_1 + (1 + \lambda_1) \ln \alpha < 0$$

или

$$\ln(E_0 \cdot E_1^{\lambda_1} \cdot \alpha^{1+\lambda_1}) < 0$$

и, наконец,

$$|\varepsilon_0| |\varepsilon_1|^{\frac{1+\sqrt{5}}{2}} \cdot \left| \frac{1}{2} \frac{f''(x^*)}{f'(x^*)} \right|^{1+\frac{1+\sqrt{5}}{2}} < 1.$$

Полученное неравенство представляет собой условие сходимости метода секущих.

Проделав аналогичные выкладки для уравнения (4.29), получим:

$$E_{n+1} = \alpha E_n^2; \quad \xi_{n+1} - 2\xi_n = a; \quad \lambda - 2 = 0; \quad \lambda_1 = 2; \quad \xi_n = C_1\lambda_1^n - a;$$



Вверх

Назад

Вперёд

Пред.

След.

Указатель

Помощь

Экран

Отсюда

$$|\varepsilon_n| \approx E_n = \frac{1}{\alpha} e^{C_1 \lambda_1^n}. \quad (4.32)$$

Наконец, $\xi_0 = C_1 - a$ и, следовательно, $C_1 = a + \xi_0 = \ln(|\varepsilon_0| \cdot \alpha)$. Таким образом, сходимость метода Ньютона определяется условием ($C_1 < 0$):

$$\frac{1}{2} \left| \frac{f''(x^*)}{f'(x^*)} \right| \cdot |\varepsilon_0| < 1.$$

При этом на каждом шаге метода погрешность возводится в квадрат ($\lambda_1 = 2$), а в методе секущих – в степень $\lambda_1 = \frac{1+\sqrt{5}}{2} \approx 1.62$, т.е. метод секущих сходится медленнее.

Замечание 4.4. Разностное уравнение (4.30) можно было бы исследовать и напрямую, без сведения его к линейному, а именно: будем искать решение в виде $\varepsilon_{n+1} = a^\alpha \varepsilon_n^\beta$, где $a = -\frac{1}{2} \frac{f''(x^*)}{f'(x^*)}$. Тогда, поскольку записанное соотношение предполагается выполняющимся при любых значениях n , $\varepsilon_n = a^\alpha \varepsilon_{n-1}^\beta$. Поэтому с одной стороны $\varepsilon_{n+1} = a^\alpha \left(a^\alpha \varepsilon_{n-1}^\beta \right)^\beta$, а с другой, в силу (4.30), $-\varepsilon_{n+1} = a \cdot a^\alpha \varepsilon_{n-1}^\beta \cdot \varepsilon_{n-1}$. Сравнивая параметры правых частей двух последних равенств, приходим к системе уравнений для определения параметров α и β :

$$\begin{cases} \alpha\beta + \alpha = \alpha + 1, \\ \beta^2 = \beta + 1, \end{cases} \quad \text{решая которую, приходим к полученным выше результатам.}$$

Метод хорд и комбинированные методы

Очередной модификацией метода Ньютона является [метод хорд](#). Аналитически его алгоритм выглядит следующим образом:

$$x_{n+1} = x_n - f(x_n) \cdot \frac{x_n - x_0}{f(x_n) - f(x_0)}, \quad n = 1, 2, \dots \quad (4.33)$$

Геометрически очередным приближением будет абсцисса точки пересечения хорды, проходящей через точки $(x_0, f(x_0))$ и $(x_n, f(x_n))$.



Приближения x_0 и x_1 считаются заданными. Несмотря на то, что в вычислениях участвуют два значения — x_n и x_0 , метод хорд является одношаговым. По скорости сходимости он заметно уступает методу Ньютона и методу секущих. Действительно, аналогично методу секущих можно записать:

$$\varepsilon_{n+1} \approx -\frac{1}{2} \frac{f''(x^*)}{f'(x^*)} \varepsilon_0 \varepsilon_n.$$

Отсюда следует, что скорость сходимости будет равна скорости сходимости геометрической прогрессии со знаменателем $q = -\frac{1}{2} \frac{f''(x^*)}{f'(x^*)} \varepsilon_0$.

Сам по себе метод хорд применяется редко, но он может быть полезным в комбинации с другими методами, так как позволяет при соответствующем выборе начальных данных получить двусторонние приближения к корню. Так, например, произойдет, если выбрать $f(x_1) \cdot f(x_0) < 0$, а затем комбинировать метод хорд и метод Ньютона: в качестве начального значения для метода хорд (x_0) брать новое значение x_1 , вычисленное по методу Ньютона.

Геометрическую интерпретацию можно видеть на рисунке.

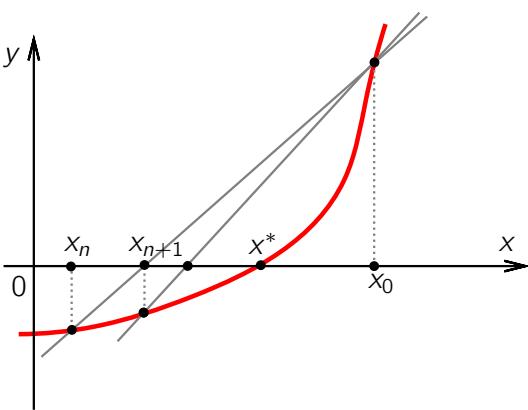


Рисунок 4.7



Аналитически описанный алгоритм выглядит следующим образом:

$$x'_n = x'_{n-1} - \frac{f(x'_{n-1})}{f'(x'_{n-1})}, \quad (4.34)$$

$$x_{n+1} = x_n - f(x_n) \cdot \frac{x_n - x'_n}{f(x_n) - f(x'_n)}, \quad n = 1, 2, \dots$$

Заметим еще раз, что все рассмотренные одношаговые методы можно рассматривать как частные случаи метода итерации. Действительно, ранее мы отмечали преобразование $x + \psi(x) f(x) \equiv \varphi(x)$, приводящее уравнение $f(x) = 0$ к виду, удобному для итерации. С этой точки зрения имеем:

если $\psi(x) = -\frac{1}{f'(x_0)}$, то получается **видоизменение метода Ньютона с постоянной производной**;

если $\psi(x) = -\frac{k}{f'(x)}$, то получаем **видоизменение метода Ньютона на случай кратных корней**;

если $\psi(x) = -\frac{x - x_0}{f(x) - f(x_0)}$, то получим **метод хорд**.

Эта связь полезна тем, что, как мы уже не раз отмечали, позволяет применять результаты по сходимости метода итерации к этим методам, хотя это и приводит, как правило, к несколько завышенным результатам.



4.1.7. Методы отыскания корней алгебраических уравнений

Метод Лобачевского

Метод Лина разложения многочлена на множители

Метод Лина выделения линейного множителя

Метод Лобачевского

Идея рассматриваемого ниже метода принадлежит Н.И. Лобачевскому и предложена им в 1834 г. Метод не требует предварительного отделения корней и позволяет найти сразу все корни.

Итак, рассмотрим алгебраическое уравнение

$$P(x) = a_0x^n + a_1x^{n-1} + \dots + a_n = 0, \quad a_0 \neq 0. \quad (4.35)$$

Для определенности считаем коэффициенты a_i вещественными. Предположим также, что все корни вещественны и удовлетворяют соотношению

$$|x_1| \gg |x_2| \gg \dots \gg |x_n|, \dots \frac{|x_{k+1}|}{|x_k|} \ll 1, \quad k = 1, \dots, n-1. \quad (4.36)$$

Запишем соотношения Виета, связывающие корни уравнения (4.35) с его коэффициентами a_i :

$$\left\{ \begin{array}{l} x_1 + x_2 + \dots + x_n = -\frac{a_1}{a_0}, \\ x_1x_2 + x_1x_3 + \dots + x_{n-1}x_n = \frac{a_2}{a_0}, \\ x_1x_2x_3 + x_1x_2x_4 + \dots + x_{n-2}x_{n-1}x_n = -\frac{a_3}{a_0}, \\ \dots \\ x_1x_2 \dots x_n = (-1)^n \frac{a_n}{a_0}. \end{array} \right. \quad (4.37)$$



В случае выполнения условий (4.36) (говорят, что в этом случае корни сильно разделены в смысле отношений) в левых частях соотношений Виета (4.37) главными членами будут первые слагаемые; тогда вместо точных можно записать приближенные соотношения (причем здесь уже можно уточнить смысл понятия «сильно разделены»: когда остальными слагаемыми в пределах принятой точности можно пренебречь):

$$\left\{ \begin{array}{l} x_1 \approx -\frac{a_1}{a_0}, \\ x_1 x_2 \approx \frac{a_2}{a_0}, \\ x_1 x_2 x_3 \approx -\frac{a_3}{a_0}, \\ \dots \\ x_1 x_2 \dots x_n = (-1)^n \frac{a_n}{a_0}. \end{array} \right.$$

Отсюда можно найти приближенные значения корней:

$$x_i \approx -\frac{a_i}{a_{i-1}}, \quad i = 1, \dots, n. \quad (4.38)$$

Если же требования сильной разделенности корней не выполняются, то и соотношения (4.38) также не будут справедливы. В то же время, если построить новое уравнение, корни которого будут высокими степенями соответствующих корней заданного уравнения, то можно надеяться получить уравнение с сильно разделенными корнями. Лобачевским был предложен способ построения уравнения, корни которого являются квадратами корней исходного.

Запишем исходное уравнение в виде

$$P(x) = a_0 (x - x_1)(x - x_2) \dots (x - x_n) = 0.$$

Параллельно рассмотрим многочлен

$$P^*(x) = a_0 (x + x_1)(x + x_2) \dots (x + x_n),$$

корни которого отличаются от корней исходного лишь знаками.



Многочлен $P^*(x)$ можно записать в развернутом виде, не имея сведений о его корнях:

$$P^*(x) = a_0 x^n - a_1 x^{n-1} + a_2 x^{n-2} - \cdots + (-1)^n a_n = 0.$$

По многочленам $P(x)$ и $P^*(x)$ легко построить многочлен $P_1(y)$, корнями которого являются x_i^2 , $i = \overline{1, n}$. Для этого достаточно перемножить многочлены $P(x)$ и $P^*(x)$ и произвести замену $y = x^2$:

$$P(x) P^*(x) = a_0^2 (x^2 - x_1^2) (x^2 - x_2^2) \dots (x^2 - x_n^2).$$

Вычислим коэффициенты многочлена $P_1(y)$. Для этого перемножим многочлены $P(x)$ и $P^*(x)$ почленно:

$$P(x) P^*(x) = (a_0 x^n + a_1 x^{n-1} + a_2 x^{n-2} + \cdots + a_n) (a_0 x^n - a_1 x^{n-1} + a_2 x^{n-2} - \cdots + (-1)^n a_n).$$

Отсюда находим (коэффициенты $P_1(y)$ будем обозначать $a_i^{(1)}$):

$$\left\{ \begin{array}{l} a_0^{(1)} = a_0^2, \\ a_1^{(1)} = 2a_0 a_2 - a_1^2, \\ a_2^{(1)} = 2a_0 a_4 - 2a_1 a_3 + a_2^2, \\ a_3^{(1)} = 2a_0 a_6 - 2a_1 a_5 + 2a_2 a_4 - a_3^2, \\ \dots \\ a_n^{(1)} = (-1)^n a_n^2. \end{array} \right. \quad (4.39)$$

На основании этих соотношений можно построить последовательность многочленов

$$P_k(x) = a_0^{(k)} x^n + a_1^{(k)} x^{n-1} + \cdots + a_n^{(k)},$$



корнями которых будут $x_i^{2^k}$. На некотором шаге мы получим сильную разделенность корней, и тогда для их нахождения можно будет воспользоваться формулами типа (4.38), которые будут теперь иметь вид

$$x_i^{2^k} \approx -\frac{a_i^{(k)}}{a_{i-1}^{(k)}}.$$

Отсюда, извлекая корни степени 2^k , найдем модули корней, а их знаки определим подстановкой в многочлен.

Исследуем вопрос о том, сколько шагов описанного процесса (в литературе он носит название процесса квадрирования) нужно произвести, чтобы получить «сильную разделенность». Напомним, что пока мы предполагаем корни простыми.

Пусть процесс квадрирования проведен k раз и построен многочлен $P_k(x)$ такой, что его корни достаточно разделены. Тогда имеем:

$$\left\{ \begin{array}{l} x_1^{2^k} \approx -\frac{a_1^{(k)}}{a_0^{(k)}}, \\ x_1^{2^k} x_2^{2^k} \approx -\frac{a_2^{(k)}}{a_0^{(k)}}, \\ x_1^{2^k} x_2^{2^k} x_3^{2^k} \approx -\frac{a_3^{(k)}}{a_0^{(k)}}, \\ \dots \\ x_1^{2^k} x_2^{2^k} \dots x_n^{2^k} = (-1)^n \frac{a_n^{(k)}}{a_0^{(k)}} \end{array} \right. , \quad (*)$$



Сделаем еще один шаг квадрирования. От этого разделенность корней может лишь усилиться. Поэтому

$$\left\{ \begin{array}{l} x_1^{2^{k+1}} \approx -\frac{a_1^{(k+1)}}{a_0^{(k+1)}}, \\ x_1^{2^{k+1}} x_2^{2^{k+1}} \approx \frac{a_2^{(k+1)}}{a_0^{(k+1)}}, \\ x_1^{2^{k+1}} x_2^{2^{k+1}} x_3^{2^{k+1}} \approx -\frac{a_3^{(k+1)}}{a_0^{(k+1)}}, \\ \dots \\ x_1^{2^{k+1}} x_2^{2^{k+1}} \dots x_n^{2^{k+1}} = (-1)^n \frac{a_n^{(k+1)}}{a_0^{(k+1)}}. \end{array} \right. \quad (**)$$

Так как $a_0^{(k+1)} = [a_0^{(k)}]^2$, то из (*) и (**) имеем:

$$x_1^{2^{k+1}} = (x_1^{2^k})^2,$$

что равносильно равенству

$$-\frac{a_1^{(k+1)}}{a_0^{(k+1)}} \approx \left[-\frac{a_1^{(k)}}{a_0^{(k)}} \right]^2 = \frac{\left[a_1^{(k)} \right]^2}{\left[a_0^{(k)} \right]^2},$$

откуда следует, что $a_1^{(k+1)} \approx -[a_1^{(k)}]^2$. Аналогично получаем: $a_2^{(k+1)} \approx +[a_2^{(k)}]^2, \dots$

Таким образом, условием того, что достигнута требуемая степень разделенности корней, является следующая связь между коэффициентами многочленов $P_k(x)$ и $P_{k+1}(x)$:

$$a_i^{(k+1)} \approx (-1)^i \left[a_i^{(k)} \right]^2, \quad i = \overline{1, n} \quad (4.40)$$

Тогда модули корней могут быть найдены по формулам

$$|x_i| \approx \sqrt[2^{k+1}]{-\frac{a_i^{(k+1)}}{a_{i-1}^{(k+1)}}}, \quad i = 1, \dots, n. \quad (4.41)$$



В случае, если между корнями имеются более сложные связи (например, кратные корни, комплексно-сопряженные пары и т.п.), требуется и более сложный анализ по их определению.

Так, например, если $|x_1| > |x_2| = |x_3| > |x_4| > \dots > |x_n|$ (x_2 и x_3 – вещественные), то вместо (*) получим равенства

$$\left\{ \begin{array}{l} x_1^{2^k} \approx -\frac{a_1^{(k)}}{a_0^{(k)}}, \\ 2x_1^{2^k} x_2^{2^k} \approx \frac{a_2^{(k)}}{a_0^{(k)}}, \\ x_1^{2^k} x_2^{2^k} x_3^{2^k} \approx -\frac{a_3^{(k)}}{a_0^{(k)}}, \\ \dots \\ x_1^{2^k} x_2^{2^k} \dots x_n^{2^k} = (-1)^n \frac{a_n^{(k)}}{a_0^{(k)}}. \end{array} \right.$$

откуда следует, что формулы (4.41) сохраняются за исключением $i = 2$, вместо которого будет равенство

$$a_2^{(k+1)} \approx \frac{1}{2} \left[a_2^{(k)} \right]^2.$$

Аналогичный анализ можно провести и в случае комплексно-сопряженной пары, хотя здесь более удобно разделить эту пару при помощи сдвига, а именно: рассмотрев многочлен $R(x) = P(x - \xi)$, где ξ – случайно выбранное комплексное число. Для $R(x)$ соответствующая пара корней уже будет иметь различные модули.

Отметим также, что в ходе вычислений после нескольких квадрирований обычно возникают большие числа, что может привести к переполнению, от которого следует страховаться введением масштабирующих множителей.

Замечание 4.5. Метод в общем случае обладает квадратичной сходимостью.

Замечание 4.6. Существуют видоизменения метода Лобачевского, более удобные с точки зрения практического использования.



Метод Лина разложения многочлена на множители

В литературе рассматриваемый ниже метод носит также название «метод предпоследнего остатка».

Изложим вначале общую его схему. Пусть нужно решить уравнение $P(x) = a_0x^n + a_1x^{n-1} + \dots + a_n = 0$, $a_0 \neq 0$ и $a_i \in R$. Чем выше степень уравнения, тем труднее его решать. Если бы удалось представить $P(x)$ в виде $P(x) = Q(x)R(x)$, где степени многочленов $Q(x)$ и $R(x)$ меньше n , то это было бы существенным шагом вперед в решении исходного уравнения. Из алгебры известно, что алгебраическое уравнение с вещественными коэффициентами может быть разложено на множители первой и второй степени (а значит, и в произведение пары сомножителей, сумма степеней которых равна n).

Попробуем выделить в качестве множителя многочлен $Q(x)$ произвольной степени $m < n$:

$$Q(x) = x^m + q_1x^{m-1} + \dots + q_m.$$

Поиск такого множителя будем проводить методом последовательных приближений (итераций). Для этого построим начальное приближение

$$Q_0(x) = x^m + q_1^{(0)}x^{m-1} + \dots + q_m^{(0)}.$$

Это можно сделать достаточно просто, если удалось отделить m корней исходного уравнения. Выбрав $Q_0(x)$, делим $P(x)$ на $Q_0(x)$ до предпоследнего остатка (степень которого, вообще говоря, равна m):

$$P(x) = B_0(x)Q_0(x) + C_0(x).$$

Пусть старший коэффициент многочлена – остатка $C_0(x)$ отличен от нуля ($c_0 \neq 0$) (ибо в противном случае метод неприменим). По остатку составляем многочлен $Q_1(x) = \frac{1}{c_0}C_0(x)$, т.е.

$$P(x) = B_0(x)Q_0(x) + c_0Q_1(x).$$

На следующем шаге принимаем $Q_1(x)$ за исходное приближение и повторяем процесс:

$$P(x) = B_1(x)Q_1(x) + c_1Q_2(x), \quad c_1 \neq 0,$$

.....

$$P(x) = B_k(x)Q_k(x) + c_kQ_{k+1}(x), \quad c_k \neq 0.$$



В результате мы получим две последовательности полиномов ($B_k(x)$ и $Q_k(x)$) и числовую последовательность c_k , $r = 0, 1, \dots$. Если эти последовательности окажутся сходящимися, то, переходя к пределу при $k \rightarrow \infty$, получим:

$$P(x) = B(x)Q(x) + cQ(x).$$

Тем самым имеем: $P(x) = Q(x)R(x)$, где $R(x) = Q(x) + c$.

На практике следует выполнять итерации до тех пор, пока в пределах принятой точности не станут выполняться равенства $q_i^{(k)} = q_i^{(k+1)}$, $i = \overline{1, n}$.

Метод Лина выделения линейного множителя

Пусть $P(x)$ имеет по крайней мере один вещественный корень x^* (иначе вещественный множитель первой степени выделить нельзя). Тогда $Q(x) = x - x^*$.

Будем считать, что корень x^* отделен и указано начальное приближение x_0 . Тогда

$$Q_0(x) = x - x_0,$$

$$P(x) = B_0(x)Q_0(x) + C_0(x),$$

Где

$$B_0(x) = b_0^{(0)}x^{n-1} + b_1^{(0)}x^{n-2} + \dots + b_{n-2}^{(0)}x,$$

$$C_0(x) = b_{n-1}^{(0)}(x - x_0) + b_n^{(0)}.$$

Пусть $b_{n-1}^{(0)} \neq 0$. Тогда

$$Q_1(x) = x - x_0 + \frac{b_n^{(0)}}{b_{n-1}^{(0)}} = x - x_1,$$

Где

$$x_1 = x_0 - \frac{b_n^{(0)}}{b_{n-1}^{(0)}}.$$



Аналогично получим:

$$x_{k+1} = x_k - \frac{b_n^{(k)}}{b_{n-1}^{(k)}}, \quad b_{n-1}^{(k)} \neq 0, \quad k = 0, 1, \dots,$$

причем $Q_k(x) = x - x_k$.

В итоге имеем:

$$P(x) = \left(b_0^{(k)} x^{n-1} + b_1^{(k)} x^{n-2} + \dots + b_{n-2}^{(k)} x + b_{n-1}^{(k)} \right) (x - x_k) + b_n^{(k)}.$$

Сравнивая это выражение с первоначальным, получим следующие формулы для вычисления коэффициентов $b_i^{(k)}$:

$$\begin{cases} a_0 = b_0^{(k)}, \\ a_1 = b_1^{(k)} - b_0^{(k)} x_k, \\ a_2 = b_2^{(k)} - b_1^{(k)} x_k, \\ \dots \\ a_n = b_n^{(k)} - b_{n-1}^{(k)} x_k. \end{cases}$$

Отсюда находим:

$$\begin{cases} b_0^{(k)} = a_0, \\ b_1^{(k)} = a_1 + a_0 x_k, \\ b_2^{(k)} = a_2 + a_1 x_k + a_0 x_k^2, \\ \dots \\ b_{n-1}^{(k)} = a_{n-1} + a_{n-2} x_k + \dots + a_1 x_k^{n-2} + a_0 x_k^{n-1}, \\ b_n^{(k)} = a_n + a_{n-1} x_k + \dots + a_1 x_k^{n-1} + a_0 x_k^n. \end{cases}$$



Легко видеть, что последние два равенства можно переписать в виде

$$b_n^{(k)} = P(x_k),$$

$$b_{n-1}^{(k)} = \frac{P(x_k) - a_n}{x_k}.$$

Следовательно, имеем алгоритм для определения последовательности $\{x_k\}$:

$$x_{k+1} = x_k - \frac{b_n^{(k)}}{b_{n-1}^{(k)}} = x_k - \frac{x_k P(x_k)}{P(x_k) - a_n}, \quad k = 0, 1, \dots$$

или (4.42)

$$x_{k+1} = -\frac{a_n}{a_0 x_k^n + \dots + a_{n-1}}, \quad k = 0, 1, \dots$$

Если $x_k \xrightarrow{k \rightarrow \infty} x^*$, то x^* – корень исходного уравнения.

Замечание 4.7. Описанный выше алгоритм Лина достаточно часто применяют для выделения квадратичного множителя.



4.2. Приближенное решение систем численных уравнений

4.2.1. Метод простой итерации и его видоизменения

4.2.2. Метод Ньютона и его видоизменения

4.2.3. Другие подходы к решению нелинейных систем

Мы переходим к изучению методов численного решения систем нелинейных уравнений. В общем виде такую систему можно записать следующим образом:

$$f_i(x_1, \dots, x_n) = 0, \quad i = \overline{1, n} \quad (4.43)$$

$$f(x) = 0, \text{ где } f = (f_1, \dots, f_n)^T, \quad x = (x_1, \dots, x_n)^T.$$

Здесь будут иметь место те же основные этапы решения, что и в рассмотренном ранее [одномерном](#) случае:

- 1) отделение корня и выбор начального приближения x_0 ;
- 2) построение последовательности приближений;
- 3) контроль сходимости.

Практически тем же самым будет и набор алгоритмов, которые могут быть применены для решения обозначенной задачи (как, впрочем, и теоремы о сходимости). Правда, при этом уровень технической сложности исполнения будет намного выше.

Так, например, проблема отделения корня и удачного выбора начального приближения x_0 превращается в самостоятельную и очень сложную проблему, не имеющую в общем случае сколько-нибудь удовлетворительного решения, что приводит к повышению роли *исследования*.

Мы рассмотрим лишь методы уточнения этого приближения.



4.2.1. Метод простой итерации и его видоизменения

Методы Зейделя и Гаусса-Зейделя

Как и ранее, метод состоит в следующем: система (4.43) преобразуется к виду, удобному для итераций (каноническому виду)

$$x = \varphi(x) \quad (4.44)$$

или в скалярном виде

$$x_i = \varphi_i(x_1, \dots, x_n), \quad i = \overline{1, n}.$$

Сделать это принципиально возможно теми же способами, что и ранее, например, с помощью тождественного преобразования $x = x + \psi(x)f(x)$. После этого приближение к решению $x^{(*)}$ (или, что то же, к неподвижной точке отображения $\varphi : R^n \rightarrow R^n$) осуществляется по правилу

$$x^{(k+1)} = \varphi(x^{(k)}), \quad k = 0, 1, \dots \quad (4.45)$$

или в координатной форме

$$x_i^{(k+1)} = \varphi_i(x_1^{(k)}, \dots, x_n^{(k)}), \quad i = \overline{1, n}; \quad k = 0, 1, \dots$$

Это и есть *метод простой итерации для систем численных уравнений*.

Выясним приближенную картинку поведения вектора ошибки вблизи корня. Пусть, как и ранее, $\varepsilon^{(k)} = x^{(*)} - x^{(k)}$. Тогда, согласно (4.45),

$$x^{(*)} - \varepsilon^{(k+1)} = \varphi(x^{(*)} - \varepsilon^{(k)}).$$

Разлагая в ряд Тейлора в окрестности корня, имеем:

$$x^{(*)} - \varepsilon^{(k+1)} = \varphi(x^{(*)}) - \frac{\partial \varphi}{\partial x}(x^{(*)})\varepsilon^{(k)} + o(\|\varepsilon^{(k)}\|)$$



или

$$\varepsilon^{(k+1)} \approx \frac{\partial \varphi(x^{(*)})}{\partial x} \varepsilon^{(k)} \quad (4.46)$$

Здесь $\frac{\partial \varphi}{\partial x} = \begin{pmatrix} \frac{\partial \varphi_1}{\partial x_1} & \dots & \frac{\partial \varphi_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial \varphi_n}{\partial x_1} & \dots & \frac{\partial \varphi_n}{\partial x_n} \end{pmatrix}$ – матрица Якоби отображения φ . Естественно, выполняя соответствующие операции, мы предполагаем возможность их осуществления.

Формула (4.46) указывает, какие изменения претерпевает вектор ошибки на одном шаге итерации. Напомним, что аналогичные формулы мы имели и в случае метода простой итерации для систем линейных алгебраических уравнений (там $\frac{\partial \varphi}{\partial x} = B$), и в случае метода простой итерации для одного нелинейного уравнения. Эта аналогия позволяет утверждать, что если все собственные значения матрицы Якоби по модулю или какая-либо ее норма будут меньше единицы, то при удачном выборе начального приближения метод простой итерации будет сходиться, причем со скоростью геометрической прогрессии.

Сформулируем сейчас строгое утверждение о сходимости метода простой итерации для систем, аналогичное соответствующей теореме 4.2.

Теорема 4.5. Если:

- 1) $\varphi(x)$ определена и непрерывна в области $\Omega(x^{(0)}, \delta) = \{x : \rho(x^{(0)}, x) \leq \delta\}$;
- 2) Отображение $\varphi(x)$ является сжимающим в этой области;
- 3) Справедливо неравенство $\frac{m}{1-q} \leq \delta$, где $\rho(\varphi(x^{(0)}, x^{(0)})) \leq m$,

то:

- 1) Последовательность приближений (4.45) с начальным приближением $x^{(0)}$ может быть построено;
- 2) В $\Omega(x^{(0)}, \delta)$ уравнение $x = \varphi(x)$ имеет решение и притом единственное;
- 3) Последовательность $\{x^{(n)}\}$ сходится к этому решению;
- 4) Скорость сходимости определяется неравенством $\rho(x^{(*)}, x^{(n)}) \leq \frac{m}{1-q} q^n$.

[[Доказательство](#)]



Методы Зейделя и Гаусса-Зейделя

По аналогии с методом Зейделя для систем линейных алгебраических уравнений на основе [метода простой итерации](#) можем записать итерационный алгоритм

$$\left\{ \begin{array}{l} x_1^{(k+1)} = \varphi_1(x_1^{(k)}, x_2^{(k)}, \dots, x_n^{(k)}), \\ x_2^{(k+1)} = \varphi_2(x_1^{(k+1)}, x_2^{(k)}, \dots, x_n^{(k)}), \\ \dots \\ x_n^{(k+1)} = \varphi_n(x_1^{(k+1)}, x_2^{(k+1)}, \dots, x_{n-1}^{(k+1)}, x_n^{(k)}) \end{array} \right. , \quad k = 0, 1, 2, \dots \quad (4.47)$$

или

$$x_i^{(k+1)} = \varphi_i(x_1^{(k+1)}, \dots, x_{i-1}^{(k+1)}, x_i^{(k)}, \dots, x_n^{(k)}), \quad i = \overline{1, n}; \quad k = 0, 1, \dots$$

Это [метод Зейделя](#) для системы (4.44). Скорость его сходимости, как и скорость сходимости метода простой итерации (если, конечно, она имеет место), линейна, т.е. равна скорости сходимости некоторой геометрической прогрессии. Достаточное условие сходимости выглядит аналогично линейному случаю: $\left\| \frac{\partial \varphi}{\partial x} \right\| < 1$ для всех $x \in \Omega(x^{(0)}, \delta)$.

В практике вычислений достаточно часто рассматривают и другую разновидность метода Зейделя, которую в литературе называют [методом Гаусса-Зейделя](#). В отличие от (4.47) она не требует предварительного преобразования системы (4.43) к каноническому виду и имеет вид

$$f_i(x_1^{(k+1)}, \dots, x_i^{(k+1)}, x_{i+1}^{(k+1)}, \dots, x_n^{(k)}) = 0, \quad i = \overline{1, n}; \quad k = 0, 1, \dots$$



или

$$\begin{cases} f_1 \left(x_1^{(k+1)}, x_2^{(k)}, \dots, x_n^{(k)} \right) = 0, \\ f_2 \left(x_1^{(k+1)}, x_2^{(k+1)}, \dots, x_n^{(k)} \right) = 0, \\ \dots \\ f_n \left(x_1^{(k+1)}, x_2^{(k+1)}, \dots, x_n^{(k+1)} \right) = 0, \end{cases} \quad k = 0, 1, \dots \quad (4.48)$$

Нахождение каждого нового значения $x_i^{(k+1)}$ требует решения, вообще говоря, нелинейного уравнения:

$$f_i \left(x_1^{(k+1)}, \dots, x_{i-1}^{(k+1)}, x_i^{(k+1)}, x_{i+1}^{(k)}, \dots, x_n^{(k)} \right) = 0 \quad (4.49)$$

Здесь $x_j^{(k)}$ при ($j > i$) известны как значения, найденные на предыдущей итерации, а $x_j^{(k+1)}$ при ($j < i$) – как значения, найденные уже в ходе вычисления текущего приближения.

Для решения уравнения (4.49) (поскольку это одно уравнение с одним неизвестным) можно применять любой из известных итерационных методов (например, из числа [описанных нами ранее](#)). Таким образом, имеем два вложенных итерационных процесса: один – внешний – задается формулами (4.48), а другой – внутренний – формулами соответствующего итерационного метода решения уравнения (4.49).

Например, если для решения уравнения (4.49) применяется [метод Ньютона](#), то соответствующие формулы внутреннего итерационного процесса имеют вид

$$x_i^{(k+1)} = x_i^{(k)} - \frac{f_i \left(x_1^{(k+1)}, \dots, x_{i-1}^{(k+1)}, x_i^{(k+1)}, x_{i+1}^{(k)}, \dots, x_n^{(k)} \right)}{\frac{\partial f_i}{\partial x_i} \left(x_1^{(k+1)}, \dots, x_{i-1}^{(k+1)}, x_i^{(k+1)}, x_{i+1}^{(k)}, \dots, x_n^{(k)} \right)}, \quad s = 0, 1, \dots \quad (4.50)$$

Внутренний итерационный процесс проводят до сходимости последовательно для всех компонент (т.е. для $i = 1, \dots, n$) и только вычислив все $x_i^{(k+1)}$, переходят к следующему шагу внешних итераций. В качестве начального приближения можно брать значение соответствующей компоненты, полученное на предыдущей внешней итерации: $x_i^{(0(k+1))} = x_i^{(k)}$.



4.2.2. Метод Ньютона и его видоизменения

Другой версией решения нелинейных систем, не требующей предварительного преобразования к виду, удобному для итерации, как и в случае одного нелинейного уравнения, является [метод Ньютона](#). Его основная идея, как это уже отмечалось, — линеаризация. Изложим ее подробнее.

Как и в случае одного уравнения, введем вектор ошибки $\varepsilon^{(k)} = x^{(*)} - x^{(k)}$. Тогда для его определения имеем задачу

$$f\left(x^{(k)} + \varepsilon^{(k)}\right) = 0.$$

Разлагая левую часть по формуле Тейлора и ограничиваясь лишь линейными членами, будем иметь

$$f\left(x^{(k)}\right) + \frac{\partial f\left(x^{(k)}\right)}{\partial x} \varepsilon^{(k)} \approx 0.$$

Если теперь, как и ранее, ввести обозначение $\Delta x^{(k)} \approx \varepsilon^{(k)}$, то в итоге получим систему линейных алгебраических уравнений для определения поправок:

$$\frac{\partial f\left(x^{(k)}\right)}{\partial x} \Delta x^{(k)} = -f\left(x^{(k)}\right) \quad (4.51)$$

Если матрица Якоби $\frac{\partial f\left(x^{(*)}\right)}{\partial x}$ окажется невырожденной, то из (4.51) можно единственным образом найти вектор поправок $\Delta x^{(k)}$ и, следовательно, построить следующее приближение:

$$x^{(k+1)} = x^{(k)} + \Delta x^{(k)} \quad (4.52)$$

Отметим, что решение системы (4.51) можно искать с помощью любого (вообще говоря, наиболее подходящего) метода решения систем линейных алгебраических уравнений, как прямого, так и итерационного. В последнем случае вновь, как и в методе [Гаусса-Зейделя](#), получаем вложенные итерационные процессы.

Рассмотрим приближенную картину поведения вектора ошибки, считая его координаты малыми по модулю. Так как

$$\Delta x^{(k)} = x^{(k+1)} - x^{(k)} = \varepsilon^{(k)} - \varepsilon^{(k+1)},$$



$$f(x^{(k)}) = f(x^{(*)} - \varepsilon^{(k)}) = f(x^{(*)}) - \frac{\partial f(x^{(*)})}{\partial x} \varepsilon^{(k)} + \frac{1}{2} \left(\frac{\partial^2 f(x^{(*)})}{\partial x^2} \varepsilon^{(k)}, \varepsilon^{(k)} \right) + \dots \approx$$

$$\approx -\frac{\partial f(x^{(*)})}{\partial x} \varepsilon^{(k)} + \frac{1}{2} \left(\frac{\partial^2 f(x^{(*)})}{\partial x^2} \varepsilon^{(k)}, \varepsilon^{(k)} \right),$$

$$\frac{\partial f(x^{(k)})}{\partial x} = \frac{\partial f(x^{(*)} - \varepsilon^{(k)})}{\partial x} = \frac{\partial f(x^{(*)})}{\partial x} - \frac{\partial^2 f(x^{(*)})}{\partial x^2} \varepsilon^{(k)} + \dots \approx \frac{\partial f(x^{(*)})}{\partial x} - \frac{\partial^2 f(x^{(*)})}{\partial x^2} \varepsilon^{(k)}$$

то, подставив эти выражения в (4.51), будем иметь:

$$\left(\frac{\partial f(x^{(*)})}{\partial x} - \frac{\partial^2 f(x^{(*)})}{\partial x^2} \varepsilon^{(k)} \right) (\varepsilon^{(k)} - \varepsilon^{(k+1)}) \approx \frac{\partial f(x^{(*)})}{\partial x} \varepsilon^{(k)} - \frac{1}{2} \left(\frac{\partial^2 f(x^{(*)})}{\partial x^2} \varepsilon^{(k)}, \varepsilon^{(k)} \right)$$

или

$$\left(\frac{\partial f(x^{(*)})}{\partial x} - \frac{\partial^2 f(x^{(*)})}{\partial x^2} \varepsilon^{(k)} \right) \varepsilon^{(k+1)} \approx -\frac{1}{2} \left(\frac{\partial^2 f(x^{(*)})}{\partial x^2} \varepsilon^{(k)}, \varepsilon^{(k)} \right).$$

Домножая обе части последнего равенства на $\frac{\partial f(x^{(*)})}{\partial x} + \frac{\partial^2 f(x^{(*)})}{\partial x^2} \varepsilon^{(k)}$, получим:

$$\left(\left(\frac{\partial f(x^{(*)})}{\partial x} \right)^2 - \left(\frac{\partial^2 f(x^{(*)})}{\partial x^2} \varepsilon^{(k)} \right)^2 \right) \varepsilon^{(k+1)} \approx -\frac{1}{2} \left(\frac{\partial f(x^{(*)})}{\partial x} + \frac{\partial^2 f(x^{(*)})}{\partial x^2} \varepsilon^{(k)} \right) \left(\frac{\partial^2 f(x^{(*)})}{\partial x^2} \varepsilon^{(k)}, \varepsilon^{(k)} \right).$$

Наконец, отбрасывая члены порядка $o(\|\varepsilon^{(k)}\|^2)$ и домножая обе части слева на $\left(\frac{\partial f(x^{(*)})}{\partial x} \right)^{-2}$, приходим к приближенному равенству

$$\varepsilon^{(k+1)} \approx -\frac{1}{2} \left[\frac{\partial f(x^{(*)})}{\partial x} \right]^{-1} \left(\frac{\partial^2 f(x^{(*)})}{\partial x^2} \varepsilon^{(k)}, \varepsilon^{(k)} \right),$$



из которого следует, что ошибка на $(k + 1)$ -м шаге квадратично зависит от ошибки на k -м шаге, что означает возможную квадратичную сходимость.

Строгих теорем о сходимости метода Ньютона существует несколько. Сформулируем одну из них.

Теорема 4.6. Пусть в шаре $\Omega(x^{(*)}, \delta)$ выполнены условия

1)

$$\left\| \left[\frac{\partial f(x)}{\partial x} \right]^{-1} \right\| \leq a_1 x \in \Omega(x^{(*)}, \delta) \quad (4.53)$$

2)

$$\left\| f(x_1) - f(x_2) - \frac{\partial f(x_2)}{\partial x} (x_1 - x_2) \right\| \leq a_2 \|x_1 - x_2\|^2 x_1, \quad x_1, x_2 \in \Omega(x^{(*)}, \delta) \quad (4.54)$$

3) $x^{(0)} \in \Omega(x^{(*)}, b)$, где $b = \min \{ \delta, c^{-1} \}$, а $c = a_1 a_2$,

Тогда метод Ньютона (4.51) сходится в $\Omega(x^{(*)}, b)$ и имеет место оценка погрешности

$$\|x^{(k)} - x^{(*)}\| \leq c^{-1} \left(c \|x^{(0)} - x^{(*)}\| \right)^{2^k}.$$

В отличие от рассматривавшейся нами ранее одномерной теоремы здесь существование решения $x^{(*)}$ предполагается.

[Доказательство]

Замечание 4.8. Если $f \in C^2(\Omega(x^{(*)}, \delta))$, то соотношение (4.54) автоматически выполняется.

Как и в случае одного нелинейного уравнения, можно говорить о видоизменениях метода Ньютона.

Так, например, [метод Ньютона с постоянной матрицей Якоби](#)

$$\frac{\partial f(x^{(0)})}{\partial x} \Delta x^{(k)} = -f(x^{(k)}); \quad x^{(k+1)} = x^{(k)} + \Delta x^{(k)}, \quad k = 0, 1, \dots \quad (4.55)$$

применяют с целью уменьшения объема вычислений на одном шаге. При этом для каждого значения k необходимо решать систему линейных алгебраических уравнений с одной и той же матрицей. Поэтому может



Вверх

Назад

Вперёд

Пред.

След.

Указатель Помощь Экран

оказаться целесообразным нахождение $\left[\frac{\partial f(x^{(0)})}{\partial x} \right]^{-1}$ (и нахождение поправки вместо (4.55) по формуле

$\Delta x^{(k)} = - \left[\frac{\partial f(x^{(0)})}{\partial x} \right]^{-1} f(x^{(k)})$, либо нахождение соответствующих компонент *LU-разложения*. Конечно же, мы должны помнить, что видоизменение (4.55) обладает более медленной сходимостью по сравнению с базовым методом (4.51).

Помимо рассмотренной выше модификации в настоящее время достаточно широкое применение находят методы, основное предназначение которых – избежать явного вычисления производных. Среди последних различают следующие.

Дискретный метод Ньютона. Задаем некоторый векторный параметр $h^{(k)} \in R^n$ (вообще говоря, свой для каждого значения k). Тогда частные производные можно заменить, например, следующим образом:

$$\frac{\partial f_i(x_1, \dots, x_n)}{\partial x_j} \approx \frac{f_i(x_1, \dots, x_j, \dots, x_n) - f_i(x_1, \dots, x_j - h_j^{(k)}, \dots, x_n)}{h_j^{(k)}} \quad (4.56)$$

(чаще всего h_j выбирают порядка ε_M) и, следовательно, матрица Якоби заменяется некоторой матрицей $J(x^{(k)}, h^{(k)})$. Тогда вместо (4.51) в алгоритме *дискретного метода Ньютона* для каждого k решается система

$$J(x^{(k)}, h^{(k)}) \Delta x^{(k)} = -f(x^{(k)}). \quad (4.57)$$

Метод секущих. В этом случае в качестве $h^{(k)}$ берут $h^{(k)} = x^{(k)} - x^{(k-1)}$. Как следствие, формулы для аппроксимации производных будут иметь вид

$$\frac{\partial f_i(x_1^{(k)}, \dots, x_n^{(k)})}{\partial x_j} \approx \frac{f_i(x_1^{(k)}, \dots, x_j^{(k)}, \dots, x_n^{(k)}) - f_i(x_1^{(k)}, \dots, x_j^{(k-1)}, \dots, x_n^{(k)})}{x_j^{(k)} - x_j^{(k-1)}}, \quad i, j = \overline{1, n} \quad (4.58)$$

При этом на каждом шаге *метода секущих* решается система типа (4.57).



Другое направление совершенствования метода Ньютона связано с тем, чтобы обойти трудности, возникающие при обращении матриц. Это привело к появлению алгоритмов, в которых аппроксимации матриц, обратных к матрицам Якоби, на каждом шаге пересчитываются с минимизацией вычислительных затрат по некоторым рекуррентным формулам (например, с помощью специально подобранных матриц единичного ранга). В литературе такие алгоритмы получили название квазиньютоновских.



4.2.3. Другие подходы к решению нелинейных систем

Методы вариационного типа

Методы продолжения по параметру

Методы вариационного типа

Подобно случаю систем линейных алгебраических уравнений для решения нелинейных систем могут применяться методы типа [рассмотренных ранее методов спуска](#), которые мы получали исходя из замены задачи решения системы задачей вариационного типа: минимизации некоторого функционала.

Итак, в этом случае строится некоторый вещественный функционал, который обращается в нуль на любом решении системы (4.43) и положителен во всех остальных случаях, например,

$$Q(x) = \|f(x)\|^2 = \sum_{i=1}^n f_i^2(x_1, \dots, x_n) \quad (4.59)$$

Затем каким-либо образом отыскиваются точки, доставляющие минимум этому функционалу. Приведем примеры соответствующих алгоритмов.

Пусть известно некоторое приближение $x^{(k)}$ к точке минимума, а направление поиска минимума задается с помощью некоторого вектора $g^{(k)}$, т.е.

$$x = x^{(k)} + t g^{(k)} \quad (4.60)$$

где $t > 0$ – некоторый числовой параметр, который выбирают таким образом, чтобы выполнялось условие

$$Q(x) < Q\left(x^{(k)}\right) \quad (4.61)$$

Найдя каким-либо образом параметр $t = t_k$, следующее приближение строят по правилу

$$x^{(k+1)} = x^{(k)} + t_k g^{(k)}.$$



В качестве t_k можно, например, взять

$$t_k = \arg \min_t Q\left(x^{(k)} + tg^{(k)}\right),$$

решив для этого задачу минимизации функции одной независимой переменной.

Если шаг (4.60) разбить на n подшагов, на каждом из которых направление поиска задавать вектором e_i – ортом i -й координатной оси, то получим алгоритм *метода покоординатного спуска*, который может быть сформулирован следующим образом: при известном значении $x^{(k)}$ для $i = \overline{1, n}$ находим:

$$x^{(k+1)} = {}^i x^{(k+1)} + t_i^{(k)} e_i,$$

где

$$t_i^{(k)} = \arg \min_t Q\left({}^i x^{(k+1)} + te_i\right), \quad {}^0 x^{(k+1)} = x^{(k)}, \quad x^{(k+1)} = {}^n x^{(k+1)}.$$

Таким образом, один шаг метода покоординатного спуска состоит в последовательной минимизации функционала $Q(x)$ вдоль каждой из координатных осей.

Если же в качестве $g^{(k)}$ в (3.2) выбрать вектор, скорость изменения $Q(x)$ вдоль которого максимальна (это приводит к равенству $g^{(k)} = -\text{grad}Q(x^{(k)})$), то в итоге получим *метод наискорейшего (градиентного) спуска*:

$$x^{(k+1)} = x^{(k)} - t_k \text{grad} Q\left(x^{(k)}\right), \quad (4.62)$$

где t_k определяется из условия

$$t_k = \arg \min_t Q\left(x^{(k)} - t \text{grad} Q\left(x^{(k)}\right)\right).$$

Методы продолжения по параметру

Как мы уже отмечали ранее, большинство рассмотренных методов обеспечивает сходимость к решению только лишь в том случае, если начальное приближение достаточно близко к $x^{(*)}$. Поэтому методы, о которых пойдет речь ниже, можно рассматривать, как попытку расширить область сходимости, или иначе, как способ получения достаточно близких начальных приближений.



Их сущность заключается в замене задачи нахождения решения системы $f(x) = 0$ специально построенной последовательностью задач, каждая из которых незначительно отличается от предыдущей. Последовательность строится таким образом, что первая система имеет известное решение $x^{(0)}$, а последняя совпадает с исходной системой.

Поскольку соседние системы последовательности отличаются незначительно, то решение предыдущей окажется хорошим начальным приближением к решению последующей, и, таким образом, можно ожидать, что выбранный итерационный процесс окажется сходящимся. Тогда, переходя от одной задачи последовательности к другой, в конце процесса находим решение исходной системы.

Таким образом, вместо системы $f(x) = 0$ рассмотрим семейство систем

$$H(x, t) = 0, \quad t \in [0; 1] \quad (4.63)$$

зависящее от параметра t . При этом решение системы $H(x, 0) = 0$ известно ($x^{(0)}$), а решение системы $H(x, 1) = 0$ совпадает с решением системы $f(x) = 0$. Простейшим примером (4.63) может служить

$$H(x, t) = tf(x) + (1 - t)f_0(x),$$

где система $f_0(x) = 0$ имеет известное решение. Тогда рассматриваем разбиение отрезка $[0; 1]$ точками $0 = t_0 < t_1 < \dots < t_N = 1$ и последовательно решаем системы

$$H(x, t_i) = 0, \quad i = 1, \dots, N \quad (4.64)$$

применив какой-либо итерационный метод, использующий в качестве начального приближения i -го уравнения решение $(i - 1)$ -го уравнения: $x^{(k+1)}(t_i) = x(t_{i-1})$. Если разность $t_{i+1} - t_i$ достаточно мала, можно надеяться на сходимость (либо регулировать последнюю за счет выбора указанной величины).



Глава 5

Приближение функций

- 5.1. Общая информация
- 5.2. Наилучшие приближения функций
- 5.3. Интерполярование
- 5.4. Приближение сплайнами

5.1. Общая информация

На практике часто бывает необходимо многократно вычислять значение некоторой функции $f(x)$ скалярного либо векторного аргумента, например, значения [элементарных функций](#) e^x , $\ln x$, $\sin x$, $\cos x$ и других (особенно это касается работы на компьютере). Запоминать и хранить таблицы значений таких функций, а затем тратить время на выборку нужного значения из таблицы нецелесообразно. Поэтому часто для нахождения значений функции $f(x)$ с заданной точностью ε ее заменяют другой, легко вычисляемой функцией $\varphi(x, a) \in \Phi(x, a)$ – некоторому подмножеству пространства, которому принадлежит $f(x)$. Здесь $a = (a_0, a_1, \dots, a_n)$ – векторный параметр, а значения $\varphi(x, a)$ вычисляются проще. В зависимости от способа оценки близости $f(x)$ и $\varphi(x, a)$ получаются различные способы приближения (наилучшие либо интерполяционные). Кроме того, аппроксимационные задачи можно естественным образом классифицировать исходя из того, как класс приближающих функций $\Phi(x, a)$ зависит от параметров a_k и как он зависит от переменных x_i .

Определение. Аппроксимационная задача называется [линейной](#), если множество Φ линейно относительно параметров a_k (например, является линейным подпространством, натянутым на заданные базисные функции $\varphi_k(x)$, $k = \overline{0, n}$); в противном случае задача называется [нелинейной](#).

По зависимости от аргументов x_i наиболее часто употребляются следующие частные случаи:

1. Полиномиальное приближение многочленами одной или нескольких независимых переменных. В случае двух независимых переменных x, y класс Φ может, например, состоять из функций

$$\varphi(x, y) = \sum_{i,j=0}^n a_{ij} x^i y^j.$$

(Вообще говоря, степени многочленов по каждой из переменных могут быть различными).

2. Экспоненциальное приближение. В этом случае Φ состоит из функций вида

$$\varphi(x) = \sum_{i=0}^m \sum_{j=1}^n a_{ij} e^{b_{ij} x_j}.$$



Если b_{ij} – фиксированные постоянные, то имеем линейное приближение, в противном случае – нелинейное; a_{ij} – либо постоянные, подлежащие определению, либо многочлены заданной степени (также подлежащие определению).

3. Тригонометрическое приближение. В этом случае используются функции

$$\varphi(x) = \sum_{i=0}^m \sum_{j=1}^n a_{ij} \begin{Bmatrix} \sin \\ \cos \end{Bmatrix} (x_j - c_{ij}).$$

4. Дробно-рациональное приближение.

$$\varphi(x) = \frac{\sum_{i=0}^m a_i u_i(x)}{\sum_{i=0}^m b_i v_i(x)},$$

Где a_i и b_i – свободные параметры, а $u_i(x)$ и $v_i(x)$ – фиксированные функции.



5.2. Наилучшие приближения функций

5.2.1. Введение

5.2.2. Наилучшие приближения в гильбертовом пространстве

5.2.3. Наилучшее равномерное приближение



5.2.1. Введение

Приведем общую постановку задачи. При этом будем оставаться в рамках линейных приближений. Пусть R – *линейное нормированное пространство* и $f \in R$ – элемент, который требуется приблизить. Возьмем в R $n+1$ линейно независимых элементов φ_i ($i = \overline{0, n}$) и образуем $(n+1)$ -мерное линейное подпространство Φ_n всевозможных линейных комбинаций (*обобщенных многочленов*)

$$\varphi = \sum_{i=0}^n c_i \varphi_i \quad (5.1)$$

с действительными коэффициентами c_i , $i = 0, 1, \dots, n$.

Определение. Величина

$$\Delta(f) = \inf_{\varphi \in \Phi_n} \|f - \varphi\| \quad (5.2)$$

называется *наилучшим приближением* элемента f на множестве Φ_n .

Соответственно, возникает вопрос: существует ли в множестве Φ_n элемент φ_0 , для которого выполняется соотношение (5.2) и если да, то как его найти.

Определение. Элемент $\varphi_0 \in \Phi_n$, такой, что

$$\|f - \varphi_0\| = \Delta(f),$$

называется *элементом наилучшего приближения* для f на Φ_n , или *проекцией* f на Φ_n .

Теорема 5.1. Для любого $f \in R$ в Φ_n существует *элемент наилучшего приближения*, причем множество всех элементов наилучшего приближения выпукло. [\[Доказательство\]](#)

Определение. Нормированное пространство R называется *строго нормированным*, если в нем равенство $\|f + g\| = \|f\| + \|g\|$ возможно только при условии $f = \lambda g$, $\lambda > 0$.

Теорема 5.2. В строго нормированном пространстве R *элемент наилучшего приближения* единственен.

[\[Доказательство\]](#)



5.2.2. Наилучшие приближения в гильбертовом пространстве

Наилучшее среднеквадратичное приближение функций алгебраическими многочленами
Метод наименьших квадратов

Пусть R – *гильбертово пространство*, H – его линейное подпространство, $f \in R$. Так как гильбертово пространство является *строго нормированным*, то, согласно общей теории, изложенной выше, в H существует единственный *элемент наилучшего приближения*. Обозначив его h_0 , поставим задачу: найти h_0 .

Теорема 5.3. Для того чтобы h_0 был элементом наилучшего приближения к f в подпространстве H , необходимо и достаточно, чтобы выполнялось соотношение $f - h_0 \perp H$, т.е. $(f - h_0, h) = 0$ для всех $h \in H$.

[[Доказательство](#)]

Рассмотрим теперь вопрос о построении элемента наилучшего приближения. Пусть подпространство Φ_n порождено элементами $\varphi_0, \varphi_1, \dots, \varphi_n$, а Φ_0 – *элемент наилучшего приближения* к $f \in R$ в Φ_n . Так как

$$\Phi_0 = \sum_{i=0}^n c_i \varphi_i,$$

то в силу (5.3) теоремы 3 задача равносильна отысканию коэффициентов c_0, c_1, \dots, c_n таких, чтобы выполнялось равенство

$$(f - \Phi_0, \varphi) = 0 \quad \forall \varphi \in \Phi_n.$$

Последнее же условие равносильно системе из $(n+1)$ условий вида

$$(f - \Phi_0, \varphi_j) = 0, \quad j = 0, 1, \dots, n.$$



Последние равенства представляют собой систему линейных алгебраических уравнений, которые в развернутом виде можно записать следующим образом:

$$\left\{ \begin{array}{l} c_0(\varphi_0, \varphi_0) + c_1(\varphi_1, \varphi_0) + \cdots + c_n(\varphi_n, \varphi_0) = (f, \varphi_0), \\ c_0(\varphi_0, \varphi_1) + c_1(\varphi_1, \varphi_1) + \cdots + c_n(\varphi_n, \varphi_1) = (f, \varphi_1), \\ \quad \dots \\ c_0(\varphi_0, \varphi_n) + c_1(\varphi_1, \varphi_n) + \cdots + c_n(\varphi_n, \varphi_n) = (f, \varphi_n). \end{array} \right. \quad (5.3)$$

Матрица системы (5.3) $G_{n+1} = [G(\varphi_0, \varphi_1, \dots, \varphi_n)]$ называется **матрицей Грамма** системы элементов $\varphi_0, \varphi_1, \dots, \varphi_n$. Так как $(\varphi_i, \varphi_j) = \overline{(\varphi_j, \varphi_i)}$, то матрица Грамма является эрмитовой.

Лемма 5.1. *Если система элементов $\varphi_0, \varphi_1, \dots, \varphi_n$ линейно независима, то матрица G_{n+1} положительно определена.*

Доказательство. Пусть $c = (c_0, c_1, \dots, c_n)^T$ – произвольный вектор с вещественными коэффициентами. Тогда

$$\left\| \sum_{i=0}^n c_i \varphi_i \right\|^2 = \left(\sum_{i=0}^n c_i \varphi_i, \sum_{i=0}^n c_i \varphi_i \right) = \sum_{i,j=0}^n c_i c_j (\varphi_i, \varphi_j) = (G_{n+1} c, c) \geqslant 0.$$

С другой стороны, если $\varphi_0, \varphi_1, \dots, \varphi_n$ линейно независимы, то равенство $\left\| \sum_{i=0}^n c_i \varphi_i \right\| = 0$ возможно только в том случае, когда все $c_i = 0$, $i = 0, 1, \dots, n$. Таким образом, $(G_{n+1} c, c) > 0$ для всех $c \neq 0$ и, согласно **определению**, G_{n+1} положительно определена. \square

Так как G_{n+1} положительно определена, то ее определитель отличен от нуля и, следовательно, система (5.3) имеет единственное решение (по сути, мы получили еще одно доказательство существования и единственности элемента наилучшего приближения). Заметим, что в рассматриваемом случае достаточно несложно получить формулу для величины **наилучшего приближения**. Действительно, имеем:

$$\Delta^2(f) = \|f - \Phi_0\|^2 = (f - \Phi_0, f - \Phi_0) = (f - \Phi_0, f) - (f - \Phi_0, \Phi_0)$$



Последнее слагаемое здесь равно нулю в силу теоремы 5.3. Поэтому

$$\Delta^2(f) = (f - \Phi_0, f) = (f, f) - (\Phi_0, f) = (f, f) - c_0(\varphi_0, f) - c_1(\varphi_1, f) - \cdots - c_n(\varphi_n, f).$$

Рассматривая полученное уравнение совместно с системой (5.3), можем записать:

$$\left\{ \begin{array}{l} c_0(\varphi_0, \varphi_0) + c_1(\varphi_1, \varphi_0) + \cdots + c_n(\varphi_n, \varphi_0) - 1 \cdot (f, \varphi_0) = 0, \\ c_0(\varphi_0, \varphi_1) + c_1(\varphi_1, \varphi_1) + \cdots + c_n(\varphi_n, \varphi_1) - 1 \cdot (f, \varphi_1) = 0, \\ \dots \\ c_0(\varphi_0, \varphi_n) + c_1(\varphi_1, \varphi_n) + \cdots + c_n(\varphi_n, \varphi_n) - 1 \cdot (f, \varphi_n) = 0, \\ c_0(\varphi_0, f) + c_1(\varphi_1, f) + \cdots + c_n(\varphi_n, f) - 1 \cdot ((f, f) - \Delta^2(f)) = 0. \end{array} \right.$$

Полученную систему соотношений можно рассматривать как систему линейных алгебраических уравнений относительно неизвестных c_0, c_1, \dots, c_n и -1 . Так как эта система имеет ненулевое решение, то ее матрица вырождена, т.е.

$$\left| \begin{array}{ccccc} (\varphi_0, \varphi_0) & (\varphi_1, \varphi_0) & \cdots & (\varphi_n, \varphi_0) & (f, \varphi_0) \\ (\varphi_0, \varphi_1) & (\varphi_1, \varphi_1) & \cdots & (\varphi_n, \varphi_1) & (f, \varphi_1) \\ & & \dots & & \\ (\varphi_0, \varphi_n) & (\varphi_1, \varphi_n) & \cdots & (\varphi_n, \varphi_n) & (f, \varphi_n) \\ (\varphi_0, f) & (\varphi_1, f) & \cdots & (\varphi_n, f) & -\Delta^2(f) + (f, f) \end{array} \right| = 0.$$



Представляя элементы последнего столбца в виде суммы $(f, \varphi_i) = (f, \varphi_i) + 0$, перепишем полученное равенство в виде суммы

$$\begin{array}{c}
 \left| \begin{array}{ccccc}
 (\varphi_0, \varphi_0) & (\varphi_1, \varphi_0) & \cdots & (\varphi_n, \varphi_0) & 0 \\
 (\varphi_0, \varphi_1) & (\varphi_1, \varphi_1) & \cdots & (\varphi_n, \varphi_1) & 0 \\
 & & \cdots & & \\
 (\varphi_0, \varphi_n) & (\varphi_1, \varphi_n) & \cdots & (\varphi_n, \varphi_n) & 0 \\
 (\varphi_0, f) & (\varphi_1, f) & \cdots & (\varphi_n, f) & -\Delta^2(f)
 \end{array} \right| + \\
 + \left| \begin{array}{ccccc}
 (\varphi_0, \varphi_0) & (\varphi_1, \varphi_0) & \cdots & (\varphi_n, \varphi_0) & (f, \varphi_0) \\
 (\varphi_0, \varphi_1) & (\varphi_1, \varphi_1) & \cdots & (\varphi_n, \varphi_1) & (f, \varphi_1) \\
 & & \cdots & & \\
 (\varphi_0, \varphi_n) & (\varphi_1, \varphi_n) & \cdots & (\varphi_n, \varphi_n) & (f, \varphi_n) \\
 (\varphi_0, f) & (\varphi_1, f) & \cdots & (\varphi_n, f) & (f, f)
 \end{array} \right| = 0.
 \end{array}$$

Отсюда, разлагая первый из определителей по последнему столбцу, найдем:

$$\Delta^2(f) G(\varphi_0, \varphi_1, \dots, \varphi_n) = G(\varphi_0, \varphi_1, \dots, \varphi_n, f)$$

и, следовательно,

$$\Delta^2(f) = \frac{G(\varphi_0, \varphi_1, \dots, \varphi_n, f)}{G(\varphi_0, \varphi_1, \dots, \varphi_n)} \quad (5.4)$$

Таким образом, для построения наилучшего приближения в гильбертовом пространстве необходимо:

- 1) Выбрать систему $\varphi_0, \varphi_1, \dots, \varphi_n$ базисных элементов подпространства Φ_n ;



- 2) Составить и решить систему (5.3). Ее решения будут коэффициентами линейной комбинации функций $\varphi_0, \varphi_1, \dots, \varphi_n$, задающей элемент наилучшего приближения.

При практическом построении наилучшего приближения нужно проявлять известную осторожность при выполнении первого пункта сформулированного выше алгоритма, поскольку при неудачном выборе матрица Грамма может оказаться плохо обусловленной со всеми вытекающими отсюда последствиями. Это, в конечном итоге, определяет и выбор параметра n , поскольку вместо ожидаемой сходимости при $n \rightarrow \infty$ с ростом n можно получать все более плохие результаты. Впрочем, построение элемента наилучшего приближения заметно упрощается, если $\varphi_0, \varphi_1, \dots, \varphi_n$ – ортонормированная система элементов, так как в этом случае система (5.3) примет вид

$$\left\{ \begin{array}{l} c_0 = (f, \varphi_0), \\ c_1 = (f, \varphi_1), \\ \dots \\ c_n = (f, \varphi_n). \end{array} \right. \quad (5.5)$$

Соответствующая величина $\Delta(f)$ также вычисляется проще:

$$\Delta^2(f) = (f, f) - c_0(\varphi_0, f) - c_1(\varphi_1, f) - \dots - c_n(\varphi_n, f) =$$

$$= (f, f) - c_0\bar{c_0} - c_1\bar{c_1} \dots - c_n\bar{c_n} = \|f\|^2 - \sum_{i=0}^n |c_i|^2,$$

т.е.

$$\Delta(f) = \sqrt{\|f\|^2 - \sum_{i=0}^n |c_i|^2} \quad (5.6)$$



Ортогонализация Грамма–Шмидта. В общем случае ортонормированный базис можно построить, используя известную процедуру *ортогонализации Грамма–Шмидта*. Напомним, в чем ее смысл. Пусть имеется система линейно независимых элементов $\varphi_0, \varphi_1, \dots, \varphi_n$ гильбертова пространства R . Тогда можно построить такую ортонормированную систему g_0, g_1, \dots, g_n , что элементы ее будут линейными комбинациями элементов $\varphi_0, \varphi_1, \dots, \varphi_n$. Будем строить систему g_0, g_1, \dots, g_n последовательно. Положим $g_0 = \frac{\varphi_0}{\|\varphi_0\|}$ (деление возможно, так как линейно независимая система не содержит нулевого элемента). Рассмотрим далее элемент $\psi_1 = \varphi_1 - \alpha_{10}g_0$ и подберем его так, чтобы выполнялось равенство $(\psi_1, g_0) = 0$. Получим:

$$0 = (\psi_1, g_0) = (\varphi_1, g_0) - \alpha_{10}(g_0, g_0) = (\varphi_1, g_0) - \alpha_{10}.$$

Отсюда следует, что $\alpha_{10} = (\varphi_1, g_0)$. Очевидно, $\|\psi_1\| \neq 0$, так как в противном случае выполнялось бы соотношение $\varphi_1 + \alpha g_0 = \varphi_1 + \alpha \frac{\varphi_0}{\|\varphi_0\|} = 0$, что невозможно в силу линейной независимости элементов φ_0 и φ_1 . Положим $g_1 = \frac{\psi_1}{\|\psi_1\|}$. Тогда $\|g_1\| = 1$ и $(g_0, g_1) = 0$. Пусть уже построены элементы g_0, g_1, \dots, g_k такие, что $\|g_0\| = \|g_1\| = \dots = \|g_k\| = 1$, $(g_i, g_j) = 0$ при $i \neq j$ и элемент g_i является линейной комбинацией элементов $\varphi_0, \varphi_1, \dots, \varphi_i$. Построим элемент

$$\psi_{k+1} = \varphi_{k+1} - \alpha_{k+1,0}g_0 - \alpha_{k+1,1}g_1 - \dots - \alpha_{k+1,k}g_k$$

и подберем числа $\alpha_{k+1,i}$ так, чтобы выполнялись равенства $(\psi_{k+1}, g_i) = 0$, $i = \overline{0, k}$. Получим:

$$0 = (\psi_{k+1}, g_i) = (\varphi_{k+1}, g_i) - \alpha_{k+1,i}(g_i, g_i) = (\varphi_{k+1}, g_i) - \alpha_{k+1,i}$$

т.е.

$$\alpha_{k+1,i} = (\varphi_{k+1}, g_i), \quad i = 0, \dots, k.$$

Так как ψ_{k+1} есть линейная комбинация $\varphi_0, \varphi_1, \dots, \varphi_{k+1}$, то $\psi_{k+1} \neq 0$. Поэтому можно положить $g_{k+1} = \frac{\psi_{k+1}}{\|\psi_{k+1}\|}$. По индукции искомая последовательность может быть построена при любых n .

Замечание 5.1. Коэффициенты c_i в случае использования для построения элемента наилучшего приближения обобщенного многочлена по ортогональной системе являются коэффициентами Фурье элемента f по системе $\varphi_0, \varphi_1, \dots, \varphi_n$.



Наилучшее среднеквадратичное приближение функций алгебраическими многочленами

Возьмем сейчас в качестве R пространство вещественнонезначимых функций, интегрируемых с квадратом на отрезке $[a, b]$ по весу $p(x)$ (которое будем обозначать $L_2(p)[a, b]$). Функция $p(x)$ удовлетворяет условиям:

- 1) $p(x) \geq 0$ на $[a, b]$;
- 2) $p(x)$ обращается в нуль на $[a, b]$ не более чем на множество меры нуль.

Скалярное произведение в таком пространстве можно задать формулой

$$(f, g) = \int_a^b p(x) f(x) g(x) dx \quad (5.7)$$

В качестве системы функций $\varphi_0, \varphi_1, \dots, \varphi_n$ выберем систему $\varphi_i(x) = x^i, i = \overline{0, n}$. Обобщенный многочлен $\varphi = \sum_{i=0}^n c_i \varphi_i$ в этом случае превратится в обычный [алгебраический многочлен](#)

$$\varphi = P_n(x) = \sum_{i=0}^n c_i x^i. \quad (5.8)$$

Согласно общей теории, изложенной выше, существует (и притом единственный) многочлен (5.8), который дает [наилучшее приближение](#) функции $f(x) \in L_2(p)[a, b]$ в смысле метрики этого пространства, т.е. такой многочлен $P_n^*(x)$, для которого

$$\begin{aligned} \Delta^2(f) &= \|f(x) - P_n(x)\|^2 = \int_a^b p(x) [f(x) - P_n^*(x)]^2 dx = \\ &= \inf_{P_n(x)} \int_a^b p(x) [f(x) - P_n(x)]^2 dx. \end{aligned}$$

Такой многочлен называют *многочленом наилучшего среднеквадратичного приближения*. Если ввести обозначения

$$s_i = \int_a^b p(x) x^i dx, \quad m_i = \int_a^b p(x) f(x) x^i dx, \quad (5.9)$$

то коэффициенты многочлена наилучшего среднеквадратичного приближения могут быть найдены как решение системы (5.3), которая в этом случае примет вид

$$\left\{ \begin{array}{l} c_0 s_0 + c_1 s_1 + \cdots + c_n s_n = m_0, \\ c_0 s_1 + c_1 s_2 + \cdots + c_n s_{n+1} = m_1, \\ \dots \\ c_0 s_n + c_1 s_{n+1} + \cdots + c_n s_{2n} = m_n, \end{array} \right. \quad (5.10)$$

которая имеет (в соответствии с общей теорией) единственное решение.

Заметим, однако, что уже частный случай $p(x) \equiv 1$, $[a, b] = [0; 1]$ приводит к следующим числовым значениям: $s_i = \int_0^1 x^i dx = \frac{1}{i+1}$. Тогда матрица системы (5.10) будет иметь вид

$$G_{n+1} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n+1} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{n+2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \frac{1}{n+1} & \frac{1}{n+2} & \frac{1}{n+3} & \cdots & \frac{1}{2n} \end{pmatrix}$$

и представляет собой знаменитую матрицу Гильберта, известную среди вычислителей своей плохой обусловленностью (при $n = 11$ $\|G_n\| \sim 10^{16}$). Так что большие значения n в описанном алгоритме категорически противопоказаны (при работе в режиме данных float уже при $n = 7$ результаты становятся неузнаваемыми). Как мы уже отмечали выше, система (5.10) значительно упростится, если в качестве системы

$\varphi_0, \varphi_1, \dots, \varphi_n$ выбрать ортонормированную в смысле скалярного произведения (5.7) систему, которую можно построить в соответствии с описанным выше алгоритмом Грамма-Шмидта, примененным к системе $\varphi_i(x) = x^i, i = \overline{0, n}$. Однако в рассматриваемом случае процедура ортогонализации является, вообще говоря, трехэлементной, т.е. ортогональные многочлены удовлетворяют простым трехчленным рекуррентным соотношениям, которые и можно использовать для построения соответствующих систем.

Рекуррентные соотношения для ортогональных многочленов. Обозначим ортогональные многочлены $Q_0(x), Q_1(x), \dots, Q_n(x), \dots$. Тогда многочлен $xQ_n(x)$ имеет степень $n + 1$ и его можно представить в виде

$$xQ_n(x) = \alpha_0 Q_0(x) + \alpha_1 Q_1(x) + \dots + \alpha_{n+1} Q_{n+1}(x) \quad (5.11)$$

Умножим обе части этого равенства скалярно на $Q_i(x), i = \overline{0, n+1}$ (т.е. умножим обе части на произведение $p(x)Q_i(x)$ и проинтегрируем по отрезку $[a, b]$). Получим:

$$\int_a^b p(x) Q_n(x) [xQ_i(x)] dx = \sum_{j=0}^{n+1} \alpha_j \int_a^b p(x) Q_i(x) Q_j(x) dx. \quad (*)$$

В правой части записанного равенства в силу ортогональности функций $Q(x)$ останется только одно слагаемое, соответствующее значению индекса суммирования, равному i . В левой же части при любом значении $i \leq n-2$ интеграл будет равен нулю по той же причине $Q_n(x)$ ортогонален к любому (!) многочлену меньшей степени. Поэтому имеем:

$$0 = \alpha_i \int_a^b p(x) Q_i^2(x) dx, \quad i = \overline{0, n-2},$$

откуда $\alpha_i = 0, i = \overline{0, n-2}$. Следовательно, (5.11) примет вид

$$\alpha_{n+1} Q_{n+1}(x) + (\alpha_n - x) Q_n(x) + \alpha_{n-1} Q_{n-1}(x) = 0 \quad (5.12)$$

Коэффициенты в (5.12) находим, полагая в (*) $i = n-1, n, n+1$. Тогда

$$\alpha_{n-1} = \frac{\int_a^b p(x) x Q_{n-1}(x) Q_n(x) dx}{\int_a^b p(x) Q_{n-1}^2(x) dx},$$



$$\alpha_n = \frac{\int_a^b p(x) x Q_n^2(x) dx}{\int_a^b p(x) Q_n^2(x) dx},$$

$$\alpha_{n+1} = \frac{\int_a^b p(x) x Q_n(x) Q_{n+1}(x) dx}{\int_a^b p(x) Q_{n+1}^2(x) dx}.$$

Если система $Q_i(x)$ нормирована, т.е. $\int_a^b p(x) Q_i^2(x) dx = 1, i = 0, 1, \dots$, то выражения для $\alpha_{n-1}, \alpha_n, \alpha_{n+1}$ упростятся:

$$\alpha_{n-1} = \int_a^b p(x) Q_{n-1}(x) Q_n(x) dx, \alpha_n = \int_a^b p(x) Q_n^2(x) dx, \alpha_{n+1} = \int_a^b p(x) Q_n(x) Q_{n+1}(x) dx. \quad (5.13)$$

Если обозначить $a_{i,k} = \int_a^b p(x) Q_i(x) Q_k(x) dx$, то рекуррентная формула (5.12) для нормированных многочленов будут иметь вид

$$a_{n,n+1} Q_{n+1}(x) + (a_{n,n} - x) Q_n(x) + a_{n-1,n} Q_{n-1}(x) = 0. \quad (5.14)$$

Она имеет смысл при $n \geq 1$, но если формально положить $Q_{-1}(x) \equiv 0$, то она будет иметь смысл и при $n = 0$. Таким образом, зная два начальных члена последовательности $\{Q_n(x)\}$, можно по трехчленным рекуррентным формулам построить всю последовательность. Приведем сейчас соответствующий пример. Пусть $p(x) \equiv 1$, $[a, b] = [-1; 1]$. Тогда $Q_{-1}(x) \equiv 0$; $Q_0(x) = c$. Так как $\|Q_0\|^2 = \int_{-1}^1 Q_0^2(x) dx = 2c^2 = 1$, то $c = \frac{1}{\sqrt{2}}$, т.е. $Q_0(x) = \frac{1}{\sqrt{2}}$. Полагая в (5.14) $n = 0$, получим:

$$a_{01} Q_1(x) + (a_{00} - x) Q_0(x) = 0. \quad (5.15)$$

Так как в соответствии с (5.13) $a_{00} = \int_{-1}^1 x Q_0^2(x) dx = \frac{1}{2} \int_{-1}^1 x dx = 0$, то отсюда следует, что $a_{01} Q_1(x) = x Q_0(x)$, или $Q_1(x) = \frac{x}{a_{01} \sqrt{2}}$. Используя условие нормировки, имеем:

$$\|Q_1\|^2 = \int_{-1}^1 Q_1^2(x) dx = \frac{1}{2a_{01}^2} \int_{-1}^1 x^2 dx = \frac{1}{3a_{01}^2} = 1.$$



Вверх

Назад

Вперёд

Пред.

След.

Указатель

Помощь

Экран

Тогда $a_{01} = \frac{1}{\sqrt{3}}$ и $Q_1(x) = \sqrt{\frac{3}{2}}x$. Аналогично, полагая в (5.14) $n = 1$ и учитывая, что $a_{11} = \int_{-1}^1 x Q_1^2(x) dx = 0$, найдем:

$$a_{12} Q_2(x) = x Q_1(x) - a_{01} Q_0(x) = \sqrt{\frac{3}{2}}x^2 - \frac{1}{\sqrt{6}} = \frac{3x^2 - 1}{\sqrt{6}},$$

т.е.

$$Q_2(x) = \frac{3x^2 - 1}{a_{12}\sqrt{6}}.$$

Еще раз учитывая условие нормировки, получим:

$$\|Q_2\|^2 = \int_{-1}^1 Q_2^2(x) dx = \frac{1}{6a_{12}^2} \int_{-1}^1 (3x^2 - 1)^2 dx = \frac{1}{3a_{12}^2} \int_0^1 (3x^2 - 1)^2 dx = \frac{4}{15a_{12}^2} = 1,$$

откуда $a_{12} = \frac{2}{\sqrt{15}}$ и $Q_2(x) = \sqrt{\frac{5}{2}} \frac{3x^2 - 1}{2}$. Процесс, очевидно, легко продолжить, получая при этом любое необходимое количество членов последовательности. В то же время следует отметить, что в специальной литературе, посвященной ортогональным многочленам, известны аналитические представления соответствующих систем (так, например, для разобранного примера соответствующая система – система многочленов Лежандра).

Метод наименьших квадратов

Рассмотрим сейчас вкратце вопрос о построении наилучшего среднеквадратичного приближения для таблично заданной функции (соответствующий алгоритм в литературе называют методом наименьших квадратов). Прежде всего, заметим, что вместо формулы (5.7) скалярное произведение задать и в виде

$$(f, g) = \int_a^b f(x) g(x) d\alpha(x), \quad (5.16)$$



где $\alpha(x)$ – неубывающая функция и интеграл понимается в смысле Лебега-Стилтьеса. Тогда, если $\alpha(x)$ – непрерывно дифференцируемая функция, то (5.16) и (5.7) эквивалентны. Если же $\alpha(x)$ кусочно постоянна и имеет скачки p_i в точках $x_i \in [a, b]$, то скалярное произведение (5.16) сводится к сумме

$$(f, g) = \sum_{i=0}^N p_i f(x_i) g(x_i),$$

которая задает скалярное произведение функций дискретного аргумента. Именно этот вариант следует использовать для поиска наилучшего среднеквадратичного приближения таблично заданных функций. Естественно при этом, что все основные расчетные формулы остаются теми же, что и выше. Таким образом, алгоритм метода наименьших квадратов состоит в построении системы (5.10) и нахождении ее решения, которое и даст коэффициенты искомого приближения. При этом вместо формул (5.9) для определения элементов расширенной матрицы системы (5.10) необходимо использовать формулы

$$s_i = \sum_{j=0}^N p_j x_j^i, \quad m_i = \sum_{j=0}^N p_j f(x_j) x_j^i.$$

5.2.3. Наилучшее равномерное приближение

Многочлен наилучшего равномерного приближения

Примеры наилучшего равномерного приближения

Если норма в линейном нормированном пространстве определена не через скалярное произведение, то нахождение элемента наилучшего приближения существенно усложняется. Пусть R – пространство непрерывных вещественных функций, определенных на отрезке $[a, b]$ вещественной оси, с нормой $\|f\| = \sup_{x \in [a, b]} |f(x)|$. На основании результатов, изложенных ранее, мы можем утверждать, что [элемент наилучшего приближения](#) всегда существует.

Но полученное там же достаточное условие единственности элемента наилучшего приближения здесь неприменимо, поскольку пространство $C[a, b]$ не является строго нормированным. По литературе известны результаты, касающиеся единственности элемента наилучшего равномерного приближения на подпространстве обобщенных многочленов. Мы же здесь более подробно исследуем лишь случай, соответствующий выбору базисных функций $\varphi_i(x) = x^i$, когда обобщенный многочлен превращается в обычный алгебраический многочлен степени n :

$$Q_n(x) = \sum_{i=0}^n c_i x^i.$$

Многочлен наилучшего равномерного приближения

Определение. Многочлен Q_n^0 такой, что

$$\|Q_n^0 - f\| = \inf_{Q \in \Pi_n} \|Q - f\|,$$

называется [многочленом наилучшего равномерного приближения](#) степени n для функции f . Здесь Π_n – множество всех многочленов степени не выше n .

Поставим задачу, аналогичную решенной ранее для гильбертовых пространств: выяснить отличительные признаки многочлена наилучшего равномерного приближения. Вначале получим оценку снизу для величины $\Delta_n(f)$ наилучшего равномерного приближения на подпространстве многочленов степени n .



Теорема 5.4 (Валле-Пуссена). Пусть существуют $n + 2$ точек $x_0 < x_1 < \dots < x_{n+1}$ отрезка $[a, b]$ такие, что

$$\operatorname{sign} \left((f(x_i) - Q_n(x_i)) (-1)^i \right) = \text{Const},$$

что означает: при переходе от точки x_i к следующей точке x_{i+1} величина $f(x_i) - Q_n(x_i)$ меняет знак. Тогда

$$\Delta_n(f) \geq \mu = \min_{0 \leq i \leq n+1} |f(x_i) - Q_n(x_i)|.$$

[\[Доказательство\]](#)

Теперь сформулируем и докажем «отличительный признак» многочлена наилучшего равномерного приближения.

Теорема 5.5 (Чебышева). Для того чтобы многочлен $Q_n(x)$ был [многочленом наилучшего равномерного приближения](#) непрерывной на отрезке $[a, b]$ функции $f(x)$, необходимо и достаточно существование на этом отрезке по крайней мере $n + 2$ точек $x_0 < x_1 < \dots < x_{n+1}$ таких, что для них выполняются соотношения

$$f(x_i) - Q_n(x_i) = \alpha (-1)^i \|f - Q_n\| \quad i = 0, \dots, n + 1,$$

причем $\alpha = 1$ (или $\alpha = -1$) одновременно для все i .

[\[Доказательство\]](#)

Замечание 5.2. Точки x_0, x_1, \dots, x_{n+1} называют [точками чебышевского альтернанса](#), а теорему (5.5) – [теоремой о чебышевском альтернансе](#).

Теорема 5.6. [Многочлен наилучшего равномерного приближения](#) непрерывной функции единственен.

[\[Доказательство\]](#)

Примеры наилучшего равномерного приближения

Рассмотрим сейчас некоторые простейшие случаи построения многочленов наилучшего равномерного приближения.

Пример 5.1. Приблизить непрерывную на отрезке $[a, b]$ функцию $f(x)$ многочленом нулевой степени.



Решение. Пусть

$$\sup_{x \in [a,b]} f(x) = f(x_1) = M; \quad \inf_{x \in [a,b]} f(x) = f(x_2) = m.$$

Тогда многочлен $Q_0(x) = \frac{m+M}{2}$ является многочленом наилучшего равномерного приближения, а x_1 и x_2 – точками чебышевского альтернанса. Действительно,

$$f(x_1) - Q_0(x_1) = M - \frac{m+M}{2} = \frac{M-m}{2};$$

$$f(x_2) - Q_0(x_2) = m - \frac{m+M}{2} = -\frac{M-m}{2}.$$

Таким образом, $\Delta_0(f) = \frac{M-m}{2}$. □

Пример 5.2. Приблизить выпуклую на отрезке $[a, b]$ функцию $f(x)$ многочленом первой степени $Q_1(x) = c_0 + c_1x$.

Решение. Вследствие выпуклости функции $f(x)$ разность $f(x) - (c_0 + c_1x)$ может иметь только одну внутреннюю точку экстремума, поэтому точки a и b являются точками чебышевского альтернанса. Пусть d – третья точка альтернанса. Согласно теореме Чебышева имеем равенства

$$\left\{ \begin{array}{l} f(a) - (c_0 + c_1a) = \alpha L, \\ f(d) - (c_0 + c_1d) = -\alpha L, \\ f(b) - (c_0 + c_1b) = \alpha L. \end{array} \right. \quad (***)$$

Вычитая из третьего уравнения первое, получим:

$$f(b) - f(a) = c_1(b-a),$$



откуда

$$c_1 = \frac{f(b) - f(a)}{b - a}.$$

Для определения неизвестных d , L , c_0 , c_1 и $\alpha \in \{-1, 1\}$ получено всего три уравнения. Однако следует вспомнить, что точка d является точкой экстремума разности $f(x) - (c_0 + c_1x)$. Если $f(x)$ – дифференцируемая функция, то для определения d имеем уравнение (четвертое вместе с системой (***)

$$f'(d) - c_1 = 0.$$

Найдя отсюда d , можно определить c_0 , сложив первое и второе уравнения (***):

$$f(a) + f(d) = 2c_0 + c_1(a + d),$$

откуда

$$c_0 = \frac{1}{2} [f(a) + f(d) - c_1(a + d)].$$

Геометрически описанная процедура решения системы (***) выглядит следующим образом (см. рис. 5.1):

- 1) проводим секущую через точки $(a, f(a))$ и $(b, f(b))$. Для нее тангенс угла наклона равен c_1 ;
- 2) проводим касательную к кривой $y = f(x)$, параллельную секущей, построенной на предыдущем шаге (это равносильно нахождению точки d);
- 3) проводим прямую, равноудаленную от построенных секущей и касательной, которая и будет искомой.

□

Пример 5.3. Показать, что многочлен наилучшего равномерного приближения для нечетной функции тоже является нечетной функцией (в случае отрезка $[a, b]$ общего вида $f(x)$ должна быть нечетной относительно середины отрезка).

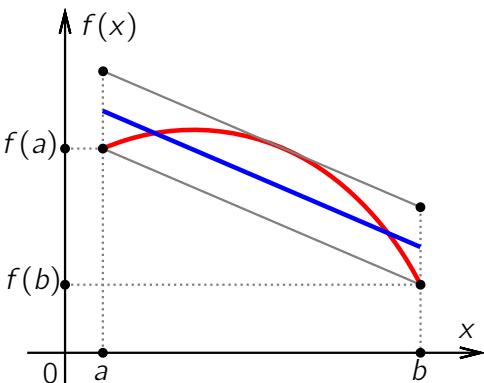


Рисунок 5.1

Решение. Действительно, пусть $Q_n(x)$ – многочлен наилучшего приближения для $f(x)$. Тогда $|f(x) - Q_n(x)| \leq \Delta_n(f)$ для всех $x \in [-1; 1]$. Заменяя x на $(-x)$, получаем:

$$|f(-x) - Q_n(-x)| \leq \Delta_n(f)$$

или (поскольку $f(-x) = -f(x)$) $|-f(x) - Q_n(-x)| \leq \Delta_n(f)$ для всех $x \in [-1; 1]$, откуда

$$|f(x) - (-Q_n(-x))| \leq \Delta_n(f).$$

Поэтому $-Q_n(-x)$ также является многочленом наилучшего равномерного приближения. В силу единственности $Q_n(x) = -Q_n(-x)$, т.е. многочлен $Q_n(x)$ является нечетной функцией. \square

В заключение отметим, что справедлива теорема, дающая оценку скорости сходимости последовательности многочленов наилучшего равномерного приближения к $f(x)$.

Теорема 5.7. Если $f(x) \in C^p[a, b]$, причем производная $f^{(p)}(x)$ удовлетворяет условию Липшица с константой C_p , то $\Delta_n(f) < \frac{C_p}{n^{p+1}}$.



Оценка, приведенная в теореме, говорит о том, что $\Delta_n(f)$ при достаточно гладкой функции $f(x)$ стремится к нулю очень быстро.

Замечание 5.3. Для построения многочленов наилучшего равномерного приближения в общем случае используются специальные итерационные алгоритмы.



5.3. Интерполярование

[5.3.1. Введение](#)

[5.3.2. Задача интерполяирования по значениям функции](#)

[5.3.3. Основные представления алгебраического интерполяционного многочлена](#)

[5.3.4. Остаток интерполяирования](#)

[5.3.5. Минимизация остатка интерполяирования](#)

[5.3.6. Интерполярование при равноотстоящих узлах](#)

[5.3.7. Некоторые правила интерполяирования при равноотстоящих узлах](#)

[5.3.8. Интерполярование с кратными узлами](#)

[5.3.9. Сходимость интерполяционного процесса](#)

[5.3.10. Некоторые приложения интерполяирования](#)

[5.3.11. Многомерная алгебраическая интерполяция](#)

5.3.1. Введение

Ранее мы отмечали, что в зависимости от способа оценки близости приближаемой $f(x)$ и приближающей $\varphi(x, a)$ функций получаются различные способы приближения, и в предыдущей главе рассмотрели вопрос о построении наилучших приближений, когда приближающий элемент обеспечивает оптимальную оценку на всем классе функций и на всем пространстве, т.е. мерой близости была та или иная норма.

Сейчас же мы в качестве «меры близости» рассмотрим совпадение на заданном множестве точек значений приближающей и приближаемой функций, а также некоторых величин, связанных с ними.

Такой подход к приближению будем называть *интерполяцией*. Значения приближаемой функции и другие величины, связанные с ней и используемые при построении интерполяционного приближения, будем называть исходными данными интерполяции. Наиболее часто в этом качестве используются значения функции и (или) значения ее производных до некоторого порядка включительно.

Дадим общую постановку задачи такого интерполяирования:

Пусть

- 1) в k_0 различных точках $x_0^{(0)}, x_1^{(0)}, \dots, x_{k_0-1}^{(0)}$ известны значения функции: $f(x_i^{(0)}), i = \overline{0, k_0 - 1}$;
- 2) в k_1 различных точках $x_0^{(1)}, x_1^{(1)}, \dots, x_{k_1-1}^{(1)}$ известны значения функции: $f'(x_i^{(1)}), i = \overline{0, k_1 - 1}$
- 3) ... и т.д. ...
- 4) в k_m различных точках $x_0^{(m)}, x_1^{(m)}, \dots, x_{k_m-1}^{(m)}$ известны значения функции: $f^{(m)}(x_i^{(m)}), i = \overline{0, k_m - 1}$.

Общее число исходных данных равно $k_0 + k_1 + \dots + k_m \stackrel{\text{def}}{=} n + 1$. Требуется построить такую функцию $F(x)$ что

$$F^{(j)}(x_i^{(j)}) = f^{(j)}(x_i^{(j)}), \quad j = \overline{0, m}; \quad i = \overline{0, k_j - 1}. \quad (5.17)$$

Эту функцию $F(x)$ называют *интерполирующей*. Таким образом, система (5.17) задает условия, определяющие интерполирующую функцию.

Помимо сформулированной выше постановки задачи интерполяирования достаточно часто в приложениях рассматривается и случай, когда сетки узлов, в которых задаются производные различных порядков, совпадают. Таким образом, указанная задача выглядит так:



- 1) в точке x_0 известны значения производных функции $f(x)$ до некоторого порядка $k_0 - 1$ включительно:
 $f^{(j)}(x_0)$, $j = \overline{0, k_0 - 1}$;
- 2) в точке x_1 известны значения производных функции $f(x)$ до некоторого порядка $k_1 - 1$ включительно:
 $f^{(j)}(x_1)$, $j = \overline{0, k_1 - 1}$ и т.д.
- 3) в точке x_m известны значения производных функции $f(x)$ до некоторого порядка $k_m - 1$ включительно:
 $f^{(j)}(x_m)$, $j = \overline{0, k_m - 1}$.

Общее число исходных данных, как и выше, равно $k_0 + k_1 + \dots + k_m \stackrel{\text{def}}{=} n + 1$. Требуется построить такую функцию $F(x)$ что

$$F^{(j)}(x_i) = f^{(j)}(x_i), \quad i = \overline{0, m}; \quad j = \overline{0, k_i - 1}. \quad (5.18)$$

Указанную постановку мы далее будем рассматривать более подробно. В литературе она носит название **интерполяции Эрмита**. Частным случаем его является случай, когда для всех i $k_i = 1$, т.е. случай, когда среди входных данных отсутствуют значения производных функции $f(x)$. Такое интерполярование называют **интерполяцией по значениям функции** или простым интерполярованием.

5.3.2. Задача интерполяирования по значениям функции

Согласно данной выше общей постановке задачи в этом случае нам известны значения $y_k = f(x_k)$, $k = \overline{0, n}$. Необходимо построить новую функцию $\varphi(x)$ такую, что

$$\varphi(x_k) = f(x_k), \quad k = \overline{0, n}, \quad (5.19)$$

т.е. функцию, график которой проходил бы через $(n+1)$ заданную точку (x_k, y_k) , и которая была бы, кроме того, более удобной с вычислительной точки зрения.

Как мы уже отмечали в общей постановке задачи о приближении функций, часто в качестве $\varphi(x)$ берут известную функцию, зависящую от $(n+1)$ числового параметра, т.е. $\varphi(x) = g(x, a_0, a_1, \dots, a_n)$. Тогда для построения функции g на основании условий (5.19) получаем систему из $(n+1)$ нелинейного уравнения с $(n+1)$ неизвестным, решив которую, можно определить значения параметров a_i :

$$g(x_k, a_0, a_1, \dots, a_n) = f(x_k), \quad k = 0, 1, \dots, n. \quad (5.20)$$

Вид функции g обычно выбирают из некоторых дополнительных соображений (например, с целью добиться максимального качественного соответствия в свойствах). Если же никакой дополнительной информации о функции f не имеется, то g можно выбирать таким образом, чтобы система (5.20) решалась по возможности проще, например, была линейной. Последнее требование можно удовлетворить, если в качестве g взять обобщенный многочлен по заданной системе координатных функций $\varphi_k(x)$:

$$g(x, a_0, a_1, \dots, a_n) = \sum_{i=0}^n a_i \varphi_i(x) := Q_n(x).$$

Система для определения параметров a_0, a_1, \dots, a_n в этом случае будет линейной и примет вид

$$\left\{ \begin{array}{l} a_0 \varphi_0(x_0) + a_1 \varphi_1(x_0) + \cdots + a_n \varphi_n(x_0) = f(x_0), \\ a_0 \varphi_0(x_1) + a_1 \varphi_1(x_1) + \cdots + a_n \varphi_n(x_1) = f(x_1), \\ \dots \\ a_0 \varphi_0(x_n) + a_1 \varphi_1(x_n) + \cdots + a_n \varphi_n(x_n) = f(x_n). \end{array} \right. \quad (5.21)$$

Определитель этой системы Δ имеет вид

$$\Delta = \begin{vmatrix} \varphi_0(x_0) & \cdots & \varphi_n(x_0) \\ \cdots & \cdots & \cdots \\ \varphi_0(x_n) & \cdots & \varphi_n(x_n) \end{vmatrix}.$$

Если $\Delta \neq 0$, то при любых значениях $f(x_j)$ система (5.21) будет иметь решение, и притом единственное. Тогда выражение для коэффициентов a_i можно, используя правило Крамера, записать в виде $a_i = \frac{\Delta_i}{\Delta}$, где Δ_i – определитель, получающийся из Δ путем замены i -го столбца столбцом свободных членов $f(x_j)$. Следовательно, обобщенный многочлен $Q_n(x)$, интерполирующий функцию (x) по ее значениям, будет иметь вид

$$Q_n(x) = \frac{\Delta_0}{\Delta} \varphi_0(x) + \frac{\Delta_1}{\Delta} \varphi_1(x) + \cdots + \frac{\Delta_n}{\Delta} \varphi_n(x). \quad (5.22)$$

Функцию $Q_n(x)$ можно представить и в другой форме. Для этого разложим определитель Δ_i по элементам i -го столбца:

$$\Delta_i = \sum_{j=0}^n f(x_j) \Delta_{ij},$$

где Δ_{ij} – соответствующие алгебраические дополнения. Подставляя последнее выражение в (5.22) и собирая вместе члены с одинаковыми $f(x_j)$, будем иметь:

$$Q_n(x) = \Phi_0(x) f(x_0) + \Phi_1(x) f(x_1) + \cdots + \Phi_n(x) f(x_n). \quad (5.23)$$

Здесь функции $\Phi_i(x)$ являются, очевидно, линейными комбинациями координатных функций $\varphi_k(x)$. Они не зависят от интерполируемой функции $f(x)$ и целиком определяются функциями $\varphi_i(x)$ и сеткой узлов интерполяции (в силу чего $\Phi_i(x)$ называют *функцией влияния i-го узла*).

Заметим, что при любой системе значений $f(x_j)$ должны выполняться равенства

$$f(x_j) = \Phi_0(x_j) f(x_0) + \Phi_1(x_j) f(x_1) + \cdots + \Phi_n(x_j) f(x_n), \quad j = 0, 1, \dots, n.$$

Отсюда следует, что функции $\Phi_i(x)$ удовлетворяют условиям

$$\Phi_i(x_j) = \delta_{ij}^j. \quad (5.24)$$

Обсудим теперь вопрос о том, какие условия нужно наложить на систему $\{\varphi_i(x)\}$ для того, чтобы определитель Δ не обращался в нуль. Для целей интерполяирования важно использовать одну и ту же систему $\{\varphi_i(x)\}$ при различных совокупностях точек x_0, x_1, \dots, x_n (узлов интерполяирования). Поэтому следует отыскивать условия того, что $\Delta \neq 0$ ни при какой системе точек x_0, x_1, \dots, x_n , $x_i \neq x_j$, $x_i \in [a, b]$. Свойства линейной независимости функций $\varphi_i(x)$ уже становится недостаточно, хотя это условие и является необходимым. Так, например, функции 1 и $\sin x$ линейно независимы, но, если выбрать $x_2 = \pi - x_1$, то

$$\Delta = \begin{vmatrix} 1 & \sin x_1 \\ 1 & \sin x_2 \end{vmatrix} = 0.$$

Если $\Delta = 0$ для какой-то системы чисел x_0, x_1, \dots, x_n , то это означает, что существуют такие постоянные c_0, c_1, \dots, c_n , не все равные нулю, для которых линейная комбинация $c_0\varphi_0(x) + c_1\varphi_1(x) + \dots + c_n\varphi_n(x)$ обращается в нуль в точках x_0, x_1, \dots, x_n , т.е. эти точки являются корнями соответствующего обобщенного многочлена, который, таким образом, на отрезке $[a, b]$ имеет $(n+1)$ различный корень. Поэтому к системе функций $\{\varphi_i(x)\}$ необходимо предъявить требование, запрещающее подобное.

Определение. Система функций $\{\varphi_i(x)\}$, обобщенный многочлен степени n по которой имеет на отрезке $[a, b]$ не более n различных корней, называется [системой функций Чебышева](#).

Определение. Систему функций $\{\varphi_i(x)\}$ будем называть [полной в классе \$F\$ функций \$f\(x\)\$](#) , если для любой функции $f(x) \in F$ и любого $\varepsilon > 0$ существует натуральное число N такое, что при любом $n > N$ найдется набор параметров a_0, a_1, \dots, a_n – коэффициентов обобщенного многочлена степени $Q_n(x)$ по системе $\{\varphi_i(x)\}$ такой, что при всех $x \in [a, b]$ выполняется неравенство $|f(x) - Q_n(x)| < \varepsilon$.

Таким образом, система $\{\varphi_i(x)\}$ должна удовлетворять следующим требованиям:

- 1) Она должна быть системой Чебышева;



2) Она должна быть полной в рассматриваемом классе функций (например, С-полной).

Первое из этих требований позволяет построить интерполяционное приближение при любом расположении узлов интерполяирования на отрезке $[a, b]$, а второе оставляет, по крайней мере, теоретическую возможность построения интерполяционного приближения, обеспечивающего любую наперед заданную точность.

Простейшим примером системы функций, удовлетворяющих сформулированным требованиям, является система $\varphi_i(x) = x^i$, обобщенным многочленом по которой будет обычный алгебраический многочлен. Первое из свойств для такой системы следует из основной теоремы алгебры, а второе (для класса функций $C[a, b]$ – из теореме Вейерштрасса). Аналогично, система $\{1, \sin x, \sin 2x, \dots\}$ является полной в классе непрерывных 2π -периодических функций и системой Чебышева на отрезке $[-\frac{\pi}{2}; \frac{\pi}{2}]$.



5.3.3. Основные представления алгебраического интерполяционного многочлена

Представление алгебраического интерполяционного многочлена в форме Лагранжа

Разделенные разности и их свойства

Представление алгебраического интерполяционного многочлена в форме Ньютона

Сейчас мы более подробно займемся изучением проблемы алгебраического интерполяирования по значениям функции. В этом случае в качестве системы координатных функций рассматривается система $\varphi_i(x) = x^i$. Как следует из приведенных выше рассуждений, задача такого интерполяирования всегда разрешима, причем единственным образом, хотя этот факт можно доказать и непосредственно, поскольку определитель Δ в этом случае имеет вид

$$\Delta = \begin{vmatrix} 1 & x_0 & \cdots & x_0^n \\ 1 & x_1 & \cdots & x_1^n \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_n & \cdots & x_n^n \end{vmatrix}$$

И представляет из себя хорошо известный [определитель Вандермонда](#), и поэтому отличен от нуля, так как $x_i \neq x_j$ при $i \neq j$.

Представление алгебраического интерполяционного многочлена в форме Лагранжа

Как мы видели выше для построения алгебраического [интерполяционного многочлена](#) $P_n(x)$ достаточно решить систему (5.21), в которой $\varphi_i(x) = x^i$. Однако сейчас мы дадим более простое решение этой задачи не требующее решения указанной системы. Согласно (5.23) искомый интерполяционный многочлен $P_n(x)$ может быть представлен в виде

$$P_n(x) = \sum_{i=0}^n \Phi_i(x) f(x_i), \quad (5.25)$$

где функции влияния $\Phi_i(x)$ удовлетворяют условиям (5.54) и являются в нашем случае алгебраическими многочленами степени n (как линейные комбинации функций x^i , $i = \overline{0, n}$). В силу условий (5.54) имеем: узлы $x_0, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n$ являются корнями многочлена $\Phi_i(x)$. А поскольку это многочлен степени n , то это все его корни. Поэтому

$$\Phi_i(x) = c_i(x - x_0) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n),$$

где c_i – некоторая неизвестная постоянная, которую мы можем определить, воспользовавшись оставшимся из условий (5.54): $\Phi_i(x_i) = 1$, откуда

$$c_i = \frac{1}{(x_i - x_0) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)} = \frac{1}{\omega'_{n+1}(x_i)},$$

где для сокращения записи введено обозначение

$$\omega_{n+1}(x) = (x - x_0)(x - x_1) \cdots (x - x_n).$$

Следовательно,

$$\Phi_i(x) = \frac{\omega_{n+1}(x)}{(x - x_i)\omega'_{n+1}(x_i)}$$

и

$$P_n(x) \stackrel{\text{def}}{=} L_n(x) = \sum_{i=0}^n \Phi_i(x) f(x_i) = \sum_{i=0}^n \frac{\omega_{n+1}(x)}{(x - x_i)\omega'_{n+1}(x_i)} f(x_i). \quad (5.26)$$

Мы получили *интерполяционный многочлен в форме Лагранжа*. Очевидно, в его помощь удобно интерполировать различные функции по одной и той же таблице узлов, так как коэффициенты влияния в этом случае могут быть вычислены заранее и не требуют пересчета.

Разделенные разности и их свойства

Из курса математического анализа известен такой способ замены функции $f(x)$ как разложение ее в ряд Тейлора. Сейчас мы получим интерполяционный аналог этой формулы. Для этого нам понадобятся объекты,

которые являются дискретным аналогом производной. Таким обобщением является понятие раздёленной разности.

По определению *раздёленные разности* нулевого порядка функции $f(x)$ совпадают со значениями $f(x_i)$; разности первого порядка определяются равенствами

$$f(x_i, x_j) = \frac{f(x_j) - f(x_i)}{x_j - x_i},$$

разности второго порядка – равенствами

$$f(x_i, x_j, x_k) = \frac{f(x_j, x_k) - f(x_i, x_j)}{x_k - x_i}$$

и вообще, разности $(k+1)$ -го порядка определяются через разности k -го порядка по формуле

$$f(x_0, x_1, \dots, x_{k+1}) = \frac{f(x_1, x_2, \dots, x_{k+1}) - f(x_0, x_1, \dots, x_k)}{x_{k+1} - x_0}.$$

Получим сейчас представление раздёленных разностей через значения функции, для которой они строятся. Имеет место

Лемма 5.2.

$$f(x_0, x_1, \dots, x_k) = \sum_{j=0}^k \frac{f(x_j)}{\omega'_k(x_j)}. \quad (5.27)$$

Доказательство. Будем проводить его методом математической индукции. При $k = 0$ (5.27) превращается в равенство $f(x_0) = f(x_0)$; при $k = 1$ имеем:

$$f(x_0, x_1) = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

и

$$\sum_{j=0}^k \frac{f(x_j)}{\omega'_k(x_j)} = \frac{f(x_0)}{x_0 - x_1} + \frac{f(x_1)}{x_1 - x_0} = \frac{f(x_1) - f(x_0)}{x_1 - x_0},$$

т.е. (5.27) также справедлива.

Пусть равенство (5.27) доказано при всех $k \leq l$. Тогда по определению разделенной разности и в соответствии с индуктивным предположением имеем:

$$\begin{aligned} f(x_0, x_1, \dots, x_{l+1}) &= \frac{f(x_1, \dots, x_{l+1}) - f(x_0, \dots, x_l)}{x_{l+1} - x_0} = \\ &= \frac{1}{x_{l+1} - x_0} \left(\sum_{j=1}^{l+1} \frac{f(x_j)}{\omega'_{l+1,1}(x_j)} - \sum_{j=0}^l \frac{f(x_j)}{\omega'_{l+1,0}(x_j)} \right). \end{aligned}$$

Приводя подобные, соберем коэффициенты при $f(x_j)$ для всех значений j . Получим: для $j = 0$:

$$-\frac{1}{x_{l+1} - x_0} \cdot \frac{1}{\omega'_{l+1,0}(x_0)} = -\frac{1}{x_{l+1} - x_0} \cdot \frac{1}{(x_0 - x_1) \cdots (x_0 - x_l)} = \frac{1}{\omega'_{l+2}(x_0)},$$

для $j = l + 1$:

$$\frac{1}{x_{l+1} - x_0} \cdot \frac{1}{\omega'_{l+1,1}(x_{l+1})} = -\frac{1}{x_{l+1} - x_0} \cdot \frac{1}{(x_{l+1} - x_1) \cdots (x_{l+1} - x_l)} = \frac{1}{\omega'_{l+2}(x_{l+1})};$$

для $1 \leq j \leq l$:

$$\begin{aligned} &\frac{1}{x_{l+1} - x_0} \cdot \left[\frac{1}{\omega'_{l+1,1}(x_j)} - \frac{1}{\omega'_{l+1,0}(x_j)} \right] = \\ &= \frac{1}{x_{l+1} - x_0} \cdot \left[\frac{1}{(x_j - x_1) \dots (x_j - x_{j-1})(x_j - x_{j+1}) \dots (x_j - x_{l+1})} - \frac{1}{(x_j - x_1) \dots (x_j - x_{j-1})(x_j - x_{j+1}) \dots (x_j - x_{l+1})} \right] = \\ &= \frac{1}{x_{l+1} - x_0} \cdot \frac{(x_j - x_0) \dots (x_j - x_{l+1})}{(x_j - x_0) \dots (x_j - x_{j-1})(x_j - x_{j+1}) \dots (x_j - x_{l+1})} = \frac{1}{\omega'_{l+2}(x_0)}. \end{aligned}$$

Полученные равенства завершают доказательство. □

Обсудим сейчас основные свойства разделенных разностей:



Свойства разделенных разностей

1. Разделенная разность любого порядка есть линейный функционал своего аргумента f , т.е. если $g(x) = \alpha f(x) + \beta h(x)$, то $g(x_0, x_1, \dots, x_k) = \alpha f(x_0, x_1, \dots, x_k) + \beta h(x_0, x_1, \dots, x_k)$.
2. Разделенная разность есть симметрическая функция своих аргументов x_0, \dots, x_k (т.е. не меняется при любой их перестановке).
3. Разделенная разность первого порядка от алгебраического многочлена степени n есть алгебраический многочлен степени $n - 1$ от тех значений аргумента, по которым эта разность составляется.

В силу линейности последнее свойство достаточно установить для функции $P_n(x) = x^n$. В этом случае имеем:

$$f(x_i, x_j) = \frac{x_j^n - x_i^n}{x_j - x_i} = x_j^{n-1} + x_j^{n-2}x_i + \dots + x_jx_i^{n-2} + x_i^{n-1}$$

(т.е. [свойство 3](#) справедливо). Из свойства 3 непосредственно вытекает, что разделенная разность порядка n от алгебраического многочлена степени n есть постоянная, а все разделенные разности более высокого порядка равны нулю.

Вычисление разделенных разностей обычно оформляют в виде таблицы (в компьютерном варианте это матрица), которая при ручном счете имеет вид

 $x_0, f(x_0)$ $f(x_0, x_1)$ $x_1, f(x_1)$ $f(x_0, x_1, x_2)$ $f(x_1, x_2)$ \vdots $x_2, f(x_2)$ $f(x_0, x_1, x_2)$ $f(x_0, \dots, x_{n-1})$ \vdots $f(x_2, x_3)$ \vdots $f(x_0, \dots, x_n)$ \vdots \vdots \vdots $f(x_1, \dots, x_n)$ \vdots \vdots \vdots \vdots $x_{n-1}, f(x_{n-1})$ \vdots $f(x_{n-2}, x_{n-1}, x_n)$ $f(x_{n-1}, x_n)$ $x_n, f(x_n)$

Представление алгебраического интерполяционного многочлена в форме Ньютона

Используя раздёленные разности, можно получить другую форму записи интерполяционного многочлена. Проделаем это.

Пусть $P_k(x)$ – алгебраический многочлен степени k , интерполирующий функцию $f(x)$ по узлам x_0, x_1, \dots, x_k . Запишем тождественное равенство

$$P_n(x) = P_0(x) + [P_1(x) - P_0(x)] + [P_2(x) - P_1(x)] + \cdots + [P_n(x) - P_{n-1}(x)]. \quad (5.28)$$

5.3. Интерполярование

5.3.3. Представления алгебраического интерполяционного многочлена

Вверх Назад Вперёд Пред. След. Указатель Помощь Экран

Очевидно, разность $P_k(x) - P_{k-1}(x)$ есть многочлен степени k , причем он обращается в нуль в узлах интерполяирования x_0, x_1, \dots, x_{k-1} , общих для $P_k(x)$ и $P_{k-1}(x)$. Следовательно,

$$P_k(x) - P_{k-1}(x) = A_k(x - x_0) \cdots (x - x_{k-1}) = A_k \omega_k(x).$$

Подставляя в последнее равенство значение $x = x_k$ и учитывая, что $P_k(x_k) = f(x_k)$, получим:

$$f(x_k) - P_{k-1}(x_k) = A_k(x_k - x_0) \cdots (x_k - x_{k-1}) = A_k \omega_k(x_k) = A_k \omega'_{k+1}(x_k).$$

Отсюда

$$A_k = \frac{f(x_k)}{\omega'_{k+1}(x_k)} - \frac{P_{k-1}(x_k)}{\omega_k(x_k)}.$$

Заменяя здесь $P_{k-1}(x_k)$ его значением, вычисленным по формуле Лагранжа (5.26) и учитывая, что $(x_k - x_j) \omega'_k(x_j) = -\omega'_{k+1}(x_j)$, а также формулу (5.27), найдем:

$$\begin{aligned} A_k &= \frac{f(x_k)}{\omega'_{k+1}(x_k)} - \sum_{j=0}^{k-1} \frac{\omega_k(x_k)}{(x_k - x_j) \omega'_k(x_j) \omega_k(x_k)} f(x_j) = \frac{f(x_k)}{\omega'_{k+1}(x_k)} + \sum_{j=0}^{k-1} \frac{f(x_j)}{\omega'_{k+1}(x_j)} = \\ &= \sum_{j=0}^k \frac{f(x_j)}{\omega'_{k+1}(x_j)} = f(x_0, x_1, \dots, x_k). \end{aligned}$$

Таким образом,

$$P_k(x) - P_{k-1}(x) = (x - x_0) \cdots (x - x_{k-1}) f(x_0, x_1, \dots, x_k).$$

Подставляя это соотношение в (5.28), окончательно получим:

$$P_n(x) = f(x_0) + (x - x_0) f(x_0, x_1) + \cdots + (x - x_0) \cdots (x - x_{n-1}) f(x_0, x_1, \dots, x_n) \quad (5.29)$$

Это **интерполяционный многочлен в форме Ньютона**. В отличие от **представления Лагранжа** здесь при добавлении нового узла в таблицу (при интерполяции одной и той же функции) не нужно пересчитывать все слагаемые.



5.3.4. Остаток интерполяирования

Остаточный член в форме Ньютона

Остаточный член в форме Лагранжа

Под *остатком интерполяирования* будем понимать разность

$$r_n(x) = f(x) - P_n(x).$$

Очевидно, таким образом определенный остаток интерполяирования (его величина) зависит от следующих факторов:

- 1) свойств интерполируемой функции f ;
- 2) выбора узлов интерполяирования x_0, x_1, \dots, x_n ;
- 3) выбора точки интерполяирования x .

Получим сейчас представления остатков интерполяирования для многочленов в форме Лагранжа и Ньютона (*напомним здесь, что в силу единственности решения задачи алгебраического интерполяирования по значениям функции это всего лишь различные представления одного и того же многочлена*).

Остаточный член в форме Ньютона

Рассмотрим разделенную разность $f(x, x_0, x_1, \dots, x_n)$. Применив к ней формулу (5.27), можем записать:

$$\begin{aligned}f(x, x_0, x_1, \dots, x_n) &= \frac{f(x)}{(x - x_0) \cdots (x - x_n)} + \frac{f(x_0)}{(x_0 - x)(x_0 - x_1) \cdots (x_0 - x_n)} + \cdots + \\&+ \frac{f(x_n)}{(x_n - x)(x_n - x_0) \cdots (x_n - x_{n-1})} = \frac{f(x)}{\omega_{n+1}(x)} + \sum_{j=0}^n \frac{f(x_j)}{(x_j - x)\omega'_{n+1}(x_j)}.\end{aligned}$$

Отсюда имеем:

$$f(x) = \sum_{j=0}^n \frac{\omega_{n+1}(x)}{(x - x_j) \omega'_{n+1}(x_j)} f(x_j) + \omega_{n+1}(x) f(x, x_0, x_1, \dots, x_n)$$

и, согласно определению (учитывая, что в последнем соотношении первое слагаемое есть не что иное, как интерполяционный многочлен в форме Лагранжа),

$$r_n(x) = f(x) - P_n(x) = \omega_{n+1}(x) f(x, x_0, x_1, \dots, x_n). \quad (5.30)$$

Таким образом, мы получили *представление остатка интерполяирования в форме Ньютона*.

Остаточный член в форме Лагранжа

При выводе формулы (5.30) мы практически не использовали никаких дифференциальных свойств функции $f(x)$. Поэтому получили выражение для остаточного члена, практическое использование которого затруднительно.

Пусть теперь $f(x) \in C^{n+1}[a, b]$. Получим выражение для остатка интерполяирования в этом случае. Для этого введем вспомогательную функцию

$$\varphi(t) = f(t) - P_n(t) - K\omega_{n+1}(t),$$

где K – некоторая постоянная.

Очевидно, $\varphi(x_0) = \varphi(x_1) = \dots = \varphi(x_n) = 0$. Подберем K таким образом, чтобы $\varphi(x)$, где x – та точка, в которой мы получаем выражение для остатка, также обращалась в нуль, т.е.

$$\varphi(x) = f(x) - P_n(x) - K\omega_{n+1}(x) = r_n(x) - K\omega_{n+1}(x) = 0.$$

Отсюда

$$K = \frac{r_n(x)}{\omega_{n+1}(x)}, \quad (5.31)$$

причем знаменатель здесь отличен от нуля, поскольку точка интерполяирования x предполагается отличной от узлов интерполяирования, т.е. $x \neq x_i, i = \overline{0, n}$.



Таким образом, функция $\varphi(t) \in C^{n+1}[a, b]$ обращается в нуль на отрезке $[a, b]$ по крайней мере в $n + 2$ различных точках: x, x_0, x_1, \dots, x_n . Следовательно, на основании теоремы Ролля (между двумя нулями функции лежит по крайней мере один нуль производной) производная $\varphi'(t)$ обращается в нуль по крайней мере в $(n + 1)$ точке интервала (a, b) . Применяя теорему Ролля к $\varphi'(t)$, получим, что существует, по крайней мере, n точек на (a, b) , в которых обращается в нуль функция $\varphi''(t)$ и т.д., существует, по крайней мере, одна точка ξ на промежутке (a, b) такая, что $\varphi^{(n+1)}(\xi) = 0$. Но

$$\varphi^{(n+1)}(t) = f^{(n+1)}(t) - P_n(t) - K\omega_{n+1}^{(n+1)}(t) = f^{(n+1)}(t) - K(n+1)!$$

Положив здесь $t = \xi$, получим:

$$f^{(n+1)}(\xi) - K(n+1)! = 0, \quad \xi \in [a, b]$$

или

$$K = \frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

Отсюда, используя (5.31), окончательно имеем:

$$r_n(x) = \omega_{n+1}(x) \frac{f^{(n+1)}(\xi)}{(n+1)!}, \quad \xi \in [a, b]. \quad (5.32)$$

Это и есть остаточный член в форме Лагранжа.

Очевидно, для корректности применения теоремы Ролля достаточно положить

$$a = \min\{x, x_0, \dots, x_n\}; \quad b = \max\{x, x_0, \dots, x_n\}.$$

Формула (5.32) позволяет решить задачу об оценке величины погрешности интерполяирования в любой точке рассматриваемого отрезка:

$$|r_n(x)| \leq |\omega_{n+1}(x)| \frac{\max_{x \in [a, b]} |f^{(n+1)}(x)|}{(n+1)!}.$$



Кроме того, сравнивая выражения (5.30) и (5.32), получим для $f(x) \in C^n[a, b]$ связь между раздделенной разностью порядка n от функции $f(x)$ и ее производной n -го порядка:

$$f(x_0, x_1, \dots, x_n) = \frac{f^{(n)}(\xi)}{n!}, \quad \xi \in [a, b],$$

где a и b определены выше.

5.3.5. Минимизация остатка интерполяирования

Функцию $f(x)$ будем полагать $(n+1)$ раз непрерывно дифференцируемой и $|f^{(n+1)}(x)| \leq M$ для всех $x \in [a, b]$. Пусть также все узлы интерполяирования расположены на отрезке $[a, b]$, т.е. $x_i \in [a, b], i = \overline{0, n}$.

Очевидно, за меру погрешности в данной точке можно взять величину $|r_n(x)|$. Если рассматривать x на всем отрезке $[a, b]$, то величина погрешности будет $\max_{x \in [a, b]} |r_n(x)|$. Если же рассматривать и произвольные

функции $f(x) \in C^{n+1}[a, b]$, для которых $|f^{(n+1)}(x)| \leq M$ при всех $x \in [a, b]$, то мерой погрешности будет величина $\sup_f \max_{x \in [a, b]} |r_n(x)|$.

Каждая из этих величин зависит от выбора узлов интерполяирования. Задача минимизации остатка интерполяирования на отрезке $[a, b]$ для рассматриваемого класса сводится к следующему вопросу: как сделать, чтобы величина $\sup_f \max_{x \in [a, b]} |r_n(x)|$ была минимальной.

На основании представления остатка интерполяирования в форме Лагранжа (5.32) имеем:

$$|r_n(x)| \leq \frac{M}{(n+1)!} |\omega_{n+1}(x)|.$$

Поэтому

$$\max_{a \leq x \leq b} |r_n(x)| \leq \frac{M}{(n+1)!} \max_{a \leq x \leq b} |\omega_{n+1}(x)|.$$

Обе эти оценки точные в том смысле, что в рассматриваемом классе функций существуют $f(x)$, для которых они достигаются, например, $f(x) = \frac{M}{(n+1)!} x^{n+1} + c_1 x^n + \dots + c_{n+1}$, где $c_i, i = 1, n+1$ – произвольные постоянные. Следовательно,

$$\sup_f \max_{a \leq x \leq b} |r_n(x)| \leq \frac{M}{(n+1)!} \max_{a \leq x \leq b} |\omega_{n+1}(x)|.$$

Таким образом, мы решим задачу о минимизации остатка интерполяирования на заданном классе функций, если среди всего множества приведенных алгебраических многочленов степени $(n+1)$, имеющих различные действительные корни из отрезка $[a, b]$, мы выберем такой, максимум модуля которого будет минимальным, т.е. задача свелась к построению приведенного алгебраического многочлена степени $(n+1)$,



наименее уклоняющегося от нуля на отрезке $[a, b]$, корни которого вещественны, различны и все принадлежат $[a, b]$. Приведем сейчас решение этой задачи для отрезка $[-1, 1]$.

Итак, нужно построить многочлен $P_n(x)$ вида

$$P_n(x) = x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n, \quad (5.33)$$

удовлетворяющий условию

$$\Delta_n = \Delta_n(P_n) = \min_{P_n^*(x) \in (5.33)} \max_{x \in [-1, 1]} |0 - P_n^*(x)|.$$

Очевидно, с учетом вида (5.33) многочлена $P_n(x)$ задача эквивалентна построению [многочлена наилучшего равномерного приближения](#) степени $(n - 1)$ к функции $f(x) = x^n$ на отрезке $[-1, 1]$.

В соответствии с [теоремой о чебышевском альтернансе](#) для того чтобы многочлен $P_n(x)$ был решением поставленной задачи, необходимо и достаточно существование по меньшей мере $(n + 1)$ точек x_0, x_1, \dots, x_n на отрезке $[-1, 1]$, в которых $P_n(x)$ принимает с чередующимися знаками значение Δ_n .

Покажем сначала, что таких точек должно быть ровно $(n + 1)$. Действительно, для того чтобы непрерывная функция более чем в $(n + 1)$ последовательных точках отрезка $[-1, 1]$ могла принимать отличные от нуля значения Δ_n с чередующимися знаками, она должна обратиться в нуль на этом отрезке более чем в n точках. А поскольку полином $P_n(x)$ отличен от тождественно нулевого, то на отрезке $[-1, 1]$ он может обратиться в нуль не более чем в n точках. Поэтому искомый многочлен $P_n(x)$ на $[-1, 1]$ значение Δ_n принимает ровно $(n + 1)$ раз.

Охарактеризуем эти точки. Если $P_n(x)$ во внутренней точке отрезка $[-1, 1]$ принимает экстремальное значение, то производная $P'_n(x)$ в этой точке обращается в нуль. Но степень многочлена $P'_n(x)$ равна $(n - 1)$ и, следовательно, производная искомого многочлена может обратиться в нуль лишь в $(n - 1)$ точках. Поэтому искомый многочлен имеет $(n - 1)$ внутренних экстремальных точек на $[-1, 1]$ и, следовательно, два краевых экстремума, т.е.

$$|P_n(-1)| = |P_n(1)| = \Delta_n.$$

Обозначив через ω_j , $j = \overline{1, n}$ корни многочлена $P_n(x)$ и через x_j , $j = \overline{0, n}$ его точки экстремума, можем записать:

$$P_n(\omega_j) = 0, \quad j = \overline{1, n}; \quad (5.34)$$

$$|P_n(x_j)| = \Delta_n, \quad j = \overline{0, n}.$$

При этом легко видеть, что указанные множества точек связаны следующими соотношениями:

$$-1 = x_n < \omega_n < x_{n-1} < \omega_{n-1} < \cdots < \omega_1 < x_0 = 1. \quad (5.35)$$

Кроме того, так как $\lim_{x \rightarrow +\infty} P_n(x) = +\infty$ и все корни $P_n(x)$ лежат на отрезке $[-1, 1]$, то $P_n(1) = \Delta_n$ и, следовательно, справедливы равенства

$$P_n(x_j) = (-1)^j \Delta_n, \quad j = \overline{0, n}. \quad (5.36)$$

Изложенные выше соображения позволяют установить функциональное соотношение, которому удовлетворяет искомый полином $P_n(x)$, а именно: справедлива

Лемма 5.3. Полином $P_n(x)$ вида (5.33), наименее уклоняющийся от нуля на отрезке $[-1, 1]$, удовлетворяет дифференциальному уравнению

$$(1 - x^2) (P'_n(x))^2 = n^2 (\Delta_n^2 - P_n^2(x)). \quad (5.37)$$

Доказательство. Ранее мы доказали, что точки x_1, x_2, \dots, x_{n-1} – простые нули полинома $P'_n(x)$. С другой стороны, эти же точки являются двукратными нулями полинома $\Delta_n^2 - P_n^2(x)$. Действительно, возводя в квадрат равенства (5.36), получим: $P_n^2(x) - \Delta_n^2 = 0$ при $x = x_j$, $j = \overline{0, n}$, а поскольку $(P_n^2(x) - \Delta_n^2)' = 2P_n(x) \cdot P'_n(x)$, то, подставляя в это соотношение $x = x_j$, $j = \overline{1, n-1}$, найдем, что эти точки являются и нулями производной полинома $\Delta_n^2 - P_n^2(x)$.

Кроме того, точки $x_0 = 1$ и $x_n = -1$ – простые нули этого же полинома. Поэтому полиномы $(1 - x^2) (P'_n(x))^2$ и $\Delta_n^2 - P_n^2(x)$ степени $2n$ имеют одни и те же нули и, следовательно, отличаются только постоянным множителем, т.е. имеет место равенство

$$(1 - x^2) (P'_n(x))^2 = C (\Delta_n^2 - P_n^2(x)).$$

Приравнивая коэффициенты при старших степенях x у обоих полиномов, находим: $C = n^2$. □

Построим теперь $P_n(x)$, проинтегрировав уравнение (5.37). Это уравнение, помимо неизвестной функции $P_n(x)$, содержит еще неизвестный параметр Δ_n . Поэтому при построении будем использовать всю известную информацию о $P_n(x)$.

Рассмотрим сначала уравнение (5.37) на отрезке $[-1, 1]$. В этом случае $|P_n(x)| \leq \Delta_n$ и, следовательно, из левой и правой частей (5.37) можно извлечь корень, переписав уравнение в виде

$$\pm \frac{dP_n}{\sqrt{\Delta_n^2 - P_n^2}} = n \frac{dx}{\sqrt{1-x^2}}, \quad -1 \leq x \leq 1. \quad (5.38)$$

Исследуем левую часть (5.38) на каждом из отрезков, концами которых являются соседние точки экстремума. Если $P_n(x_{j+1}) = \Delta_n$, то при изменении x от x_{j+1} до x_j функция $P_n(x)$ убывает от Δ_n до $-\Delta_n$. При этом дифференциал dP_n отрицателен и в левой части (5.38) следует взять знак « $-$ ». Аналогично, если $P_n(x_{j+1}) = -\Delta_n$, то следует взять знак « $+$ ». Учитывая (5.36), получим, что на отрезке $[x_{j+1}; x_j]$ уравнение (5.38) должно быть записано в виде

$$(-1)^j \frac{dP_n}{\sqrt{\Delta_n^2 - P_n^2}} = n \frac{dx}{\sqrt{1-x^2}}, \quad x \in [x_{j+1}; x_j], \quad j = 0, 1, \dots, n-1. \quad (5.39)$$

Получим теперь явное выражение для $P_n(x)$ на отрезке $[-1, 1]$. Пусть x – любая точка отрезка $[-1, 1]$ и для определенности пусть, например, x принадлежит отрезку $[x_{k+1}; x_k]$ для некоторого значения k , удовлетворяющего неравенству $0 \leq k \leq n-1$.

Проинтегрируем правую часть уравнения (5.39) по x в пределах от x до 1. Получим:

$$n \int_x^1 \frac{dx}{\sqrt{1-x^2}} = n \arcsin x \left| \begin{array}{l} 1 \\ x \end{array} \right. = n \left(\frac{\pi}{2} - \arcsin x \right) = n \arccos x.$$

Проинтегрируем теперь левую часть (5.39). Когда x меняется от x_{j+1} до x_j , функция $P_n(x)$ меняется

от $P_n(x_{j+1}) = (-1)^{j+1} \Delta_n$ до $P_n(x_j) = (-1)^j \Delta_n$. Поэтому, выполняя замену переменной интегрирования x на $(-1)^j x$, будем иметь:

$$\begin{aligned} (-1)^j \int_{P_n(x_{j+1})}^{P_n(x_j)} \frac{dP_n}{\sqrt{\Delta_n^2 - P_n^2}} &= \int_{-\Delta_n}^{\Delta_n} \frac{dP_n}{\sqrt{\Delta_n^2 - P_n^2}} = \\ &= \arcsin \left. \frac{P_n}{\Delta_n} \right|_{-\Delta_n}^{\Delta_n} = \arcsin 1 - \arcsin(-1) = \pi. \end{aligned}$$

Далее, при интегрировании левой части (5.39) от $P_n(x)$ до $P_n(x_k)$ получим:

$$\begin{aligned} (-1)^k \int_{P_n(x)}^{P_n(x_k)} \frac{dP_n}{\sqrt{\Delta_n^2 - P_n^2}} &= \int_{(-1)^k P_n(x)}^{\Delta_n} \frac{dP_n}{\sqrt{\Delta_n^2 - P_n^2}} = \\ &= \arcsin \left. \frac{P_n}{\Delta_n} \right|_{(-1)^k P_n(x)}^{\Delta_n} = \arccos(-1)^k \frac{P_n(x)}{\Delta_n}. \end{aligned}$$

Так как $\int_x^1 = \int_x^{x_k} + \sum_{j=0}^{k-1} \int_{x_{j+1}}^{x_j}$, то окончательно будем иметь:

$$n \arccos x = k\pi + \arccos(-1)^k \frac{P_n(x)}{\Delta_n}. \quad (5.40)$$

Отсюда найдем:

$$P_n(x) = \Delta_n \cos(n \arccos x), \quad |x| \leq 1. \quad (5.41)$$

Полагая в (5.41) $x = \omega_k \in [x_{k+1}; x_k]$, найдем корни полинома $P_n(x)$:

$$\omega_k = \cos \frac{(2k+1)\pi}{2n}, \quad k = 0, 1, \dots, n-1.$$

Формула (5.41) определяет $P_n(x)$ для всех $x \in [-1, 1]$. Найдем вид $P_n(x)$ для $|x| \geq 1$ и определим Δ_n . Для этого заметим, что

$$\omega_{n-k-1} = \cos\left(\pi - \frac{2k+1}{2n}\pi\right) = -\omega_k, \quad k = 0, 1, \dots, n-1.$$

Отсюда следует, что четность многочлена $P_n(x)$ совпадает с четностью его степени n . Таким образом, выполняется равенство $P_n(-x) = (-1)^n P_n(x)$ и, следовательно, достаточно определить $P_n(x)$ для $x \geq 1$.

Исследуем уравнение (5.37) при $x \geq 1$. В этом случае его необходимо переписать следующим образом:

$$(x^2 - 1)(P'_n(x))^2 = n^2 (P_n^2(x) - \Delta_n^2), \quad x \geq 1.$$

Так как $x \geq 1$, то $P_n(x) \geq \Delta_n$ и функция возрастает. Поэтому, извлекая корень, получим:

$$\frac{dP_n}{\sqrt{P_n^2 - \Delta_n^2}} = n \frac{dx}{\sqrt{x^2 - 1}}.$$

При интегрировании правой части этого уравнения от 1 до x левая часть будет интегрироваться от Δ_n до $P_n(x)$. Поэтому имеем:

$$\begin{aligned} \int_{\Delta_n}^{P_n(x)} \frac{dP_n}{\sqrt{P_n^2 - \Delta_n^2}} &= \ln \left(\frac{P_n(x)}{\Delta_n} + \sqrt{\frac{P_n^2(x)}{\Delta_n^2} - 1} \right) = \operatorname{arch} \frac{P_n(x)}{\Delta_n} = \\ &= n \int_1^x \frac{dx}{\sqrt{x^2 - 1}} = n \ln \left(x + \sqrt{x^2 - 1} \right) = n \operatorname{arch} x. \end{aligned}$$

Отсюда

$$P_n(x) = \Delta_n \operatorname{ch}(n \operatorname{arch} x), \quad x \geq 1.$$

При $x \leq -1$, воспользовавшись свойством четности, можем записать:

$$P_n(x) = (-1)^n \Delta_n \operatorname{ch}(n \operatorname{arch}(-x)), \quad x \leq -1.$$



Таким образом, осталось найти Δ_n . Заметив, что старший коэффициент многочлена $\cos(n \arccos x)$ равен 2^{n-1} , а у $P_n(x) = 1$, получим: $\Delta_n = \frac{1}{2^{n-1}}$. Поэтому окончательно имеем:

$$P_n(x) = \frac{1}{2^{n-1}} T_n(x),$$

где

$$T_n(x) = \begin{cases} \cos(n \arccos x) & \text{если } |x| \leq 1, \\ \operatorname{ch}(n \operatorname{arch} x) & \text{если } x \geq 1, \\ (-1)^n \operatorname{ch}(n \operatorname{arch}(-x)) & \text{если } x \leq -1 \end{cases} \quad (5.42)$$

Полином $T_n(x)$ называется *полиномом Чебышева первого рода* степени n .

Отметим, что на основании тождества

$$\cos(n+1)\alpha + \cos(n-1)\alpha = 2\cos\alpha\cos n\alpha,$$

Полагая в последнем $\alpha = \arccos x$, получим рекуррентную формулу, связывающую три последовательных многочлена Чебышева (учитывая, что $T_0(x) = \cos(\arccos(0 \cdot x)) = 1$; $T_1(x) = \cos(\arccos x) = x$)

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \quad |x| \leq 1, \quad n = 1, 2, \dots \quad (5.43)$$

(естественно, полином $T_n(x)$ как полином имеет единое выражение на всей области определения, (3.13) дают лишь его различные функциональные представления на различных частях области определения). Из (3.14) имеем:

$$T_2(x) = 2x^2 - 1,$$

$$T_3(x) = 4x^3 - 3x,$$

$$T_4(x) = 8x^4 - 8x^2 + 1,$$

.....

$$T_n(x) = 2^{n-1}x^n - \dots$$



Таким образом, для отрезка $[-1, 1]$ поставленная нами задача минимизации остатка интерполяирования полностью решена. Искомым многочленом $\omega_{n+1}(x)$ будет многочлен $\omega_{n+1}(x) = \frac{1}{2^n} T_{n+1}(x)$, а искомыми узлами интерполяирования – его корни:

$$x_k = \cos \frac{(2k+1)\pi}{2(n+1)}, \quad k = \overline{0, n};$$

при этом $\sup_f \max_{-1 \leq x \leq 1} |r_n(x)| = \frac{M}{(n+1)! \cdot 2^n}$.

Если интерполярование производится на произвольном отрезке $[a, b]$, то линейной заменой $x = \frac{a+b}{2} + \frac{b-a}{2}t$ его можно перевести в отрезок $[-1, 1]$. Тогда узлами интерполяирования будут

$$x_k = \frac{a+b}{2} + \frac{b-a}{2} \cos \frac{(2k+1)\pi}{2(n+1)}, \quad k = \overline{0, n},$$

а сам многочлен будет иметь вид (здесь мы используем обратное преобразование из $[-1, 1]$ в $[a, b]$, которое имеет вид $t = \frac{2}{b-a} (2x - b - a)$)

$$\omega_{n+1}(x) = \frac{(b-a)^{n+1}}{2^{2n+1}} T_{n+1}\left(\frac{2x-b-a}{b-a}\right).$$

Соответственно, для оценки погрешности получим оценку

$$|f(x) - P_n(x)| \leq \frac{M(b-a)^{n+1}}{(n+1)! \cdot 2^{2n+1}}.$$

Возможна и другая постановка задачи о минимизации остатка интерполяирования, а именно: таблица узлов предполагается фиксированной. Будем считать, что $f(x)$ – функция из рассматриваемого класса и задана таблично. Как выбрать узлы интерполяирования (из заданной таблицы), чтобы результат интерполяирования в точке x , отличной от узлов таблицы, был наилучшим?

Решение этой задачи основано на отыскании таких узлов, при которых достигает минимума выражение

$$\sup_f |r_n(x)| = \frac{M}{(n+1)!} |\omega_{n+1}(x)|, \quad x \in [a, b].$$



Вверх

Назад

Вперёд

Пред.

След.

Указатель

Помощь

Экран

В качестве узлов, очевидно, необходимо выбирать те табличные значения аргумента, которые расположены ближе всего к точке интерполяирования (предполагается, что используется не вся таблица, а интерполярование производится с помощью многочлена наименьшей возможной степени).



5.3.6. Интерполярование при равноотстоящих узлах

Конечные разности

Естественно ожидать, что если промежутки между последовательными узлами интерполяирования равны, то полученные ранее представления интерполяционного многочлена упростятся. Так оно и есть на самом деле.

Мы сейчас основное внимание уделим преобразованию [многочлена Ньютона](#). Для этих целей нам понадобится специальный математический аппарат.

Конечные разности

Объекты, вынесенные в подзаголовок, являются нужным для наших целей рабочим инструментом и играют роль, сходную с ролью дифференциала при табличном задании функций.

Итак, пусть $f(x)$ задана таблично в точках, отстоящих друг от друга на равных расстояниях:

$$y_k = f(x_0 + kh), h > 0, \quad k = 0, 1, 2, \dots$$

(Правыми) [конечными разностями](#) первого порядка назовем числа

$$\Delta y_0 = y_1 - y_0, \quad \Delta y_1 = y_2 - y_1, \dots, \quad \Delta y_{n-1} = y_n - y_{n-1}.$$

По конечным разностям первого порядка рекурсивно строятся конечные разности второго порядка, по разностям второго порядка – разности третьего порядка и т.д.:

$$\Delta^2 y_0 = \Delta y_1 - \Delta y_0, \quad \Delta^2 y_1 = \Delta y_2 - \Delta y_1, \dots, \quad \Delta^2 y_{n-2} = \Delta y_{n-1} - \Delta y_{n-2},$$

.....

$$\Delta^k y_j = \Delta^{k-1} y_{j+1} - \Delta^{k-1} y_j.$$

Свойства конечных разностей вполне аналогичны (за исключением симметрии) свойствам соответствующих разделенных разностей.



Свойства конечных разностей

1. (Линейность). Если $g(x) = \alpha f(x) + \beta h(x)$, то $\Delta^k g(x) = \alpha \Delta^k f(x) + \beta \Delta^k h(x)$.

2. Конечная разность первого порядка от произвольного алгебраического многочлена степени n есть алгебраический многочлен степени $n - 1$.

Первое свойство представляется достаточно очевидным и может быть легко доказано по индукции.

Свойство 2, в силу свойства 1, достаточно проверить лишь для многочлена $P_n(x) = x^n$. В этом же случае имеем:

$$\Delta P_n(x_0) = (x_0 + h)^n - (x_0)^n = P_{n-1}(x_0).$$

Очевидным следствием свойства 2 будет такое утверждение: конечная разность порядка n от произвольного алгебраического многочлена степени n есть величина постоянная, а все конечные разности более высокого порядка равны нулю. Выше мы дали рекурсивное определение (правых) конечных разностей всех порядков. Установим сейчас их связь со значениями исходной функции $f(x)$. Пусть E – оператор сдвига аргумента на шаг, т.е. $Ey_i = y_{i+1}$ или $Ef(x) = f(x + h)$. Тогда произвольная степень оператора E может быть определена равенством $E^\alpha f(x) = f(x + \alpha h)$. Поэтому можем записать:

$$\Delta y_0 = y_1 - y_0 = E y_0 - y_0 = (E - I) y_0,$$

где $Iy_0 = y_0$, т.е. I – тождественный оператор.

Далее:

$$\Delta^2 y_0 = \Delta y_1 - \Delta y_0 = (E - I) y_1 - (E - I) y_0 = (E - I)(y_1 - y_0) = (E - I)^2 y_0.$$

По индукции легко видеть, что

$$\Delta^k y_0 = (E - I)^k y_0. \quad (5.44)$$



(5.44) представляет собой формулу, выражающую конечную разность произвольного порядка непосредственно через значения функции, например,

$$\begin{aligned}\Delta^2 y_0 &= y_2 - 2y_1 + y_0, \\ \Delta^3 y_0 &= y_3 - 3y_2 + 3y_1 - y_0, \\ &\dots\end{aligned}\tag{5.45}$$

Столь же просто может быть найдено выражение любого значения y_k функции через начальное ее значение y_0 и начальные значения конечных разностей Δy_0 , $\Delta^2 y_0$, В самом деле, по определению имеем:

$$y_1 = y_0 + \Delta y_0 = (I + \Delta) y_0,$$

где Δ – оператор конечной разности;

$$y_2 = y_1 + \Delta y_1 = (I + \Delta) y_1 = (I + \Delta)^2 y_0$$

и т.д., по индукции получим:

$$y_k = (I + \Delta)^k y_0.\tag{5.46}$$

В частных случаях можем записать:

$$\begin{aligned}y_2 &= y_0 + 2\Delta y_0 + \Delta^2 y_0, \\ y_3 &= y_0 + 3\Delta y_0 + 3\Delta^2 y_0 + \Delta^3 y_0, \\ &\dots\end{aligned}$$

Наконец, установим связь между конечными и разделенными разностями:

$$f(x_0, x_0 + h) = \frac{f(x_0 + h) - f(x_0)}{h} = \frac{y_1 - y_0}{h} = \frac{\Delta y_0}{1! \cdot h},$$

$$\begin{aligned} f(x_0, x_0 + h, x_0 + 2h) &= \frac{f(x_0 + h, x_0 + 2h) - f(x_0, x_0 + h)}{2h} = \\ &= \frac{\frac{\Delta y_1}{1! \cdot h} - \frac{\Delta y_0}{1! \cdot h}}{2h} = \frac{\Delta^2 y_0}{2! \cdot h^2}. \end{aligned}$$

Очевидно, по индукции получим:

$$f(x_0, x_0 + h, \dots, x_0 + kh) = \frac{\Delta^k y_0}{k! \cdot h^k}. \quad (5.47)$$

Так как ранее мы получили формулу $f(x_0, x_1, \dots, x_k) = \frac{f^{(k)}(\xi)}{k!}$, связывающую разделенные разности с производными, то с ее помощью из (5.47) будем иметь:

$$\Delta^k y_0 = h^k f^{(k)}(\xi), \quad \xi \in [x_0; x_0 + kh],$$

т.е. если $f^{(k)}(\xi) \neq 0$, то $\Delta^k y_0 = O(h^k)$; если же $f^{(k)}(\xi) = 0$, то $\Delta^k y_0 = o(h^k)$.

И первое, и второе означает, что для достаточно гладкой функции ее конечные разности уже не очень высокого порядка оказываются достаточно малыми величинами, если вычисления проводятся без погрешностей и шаг h не очень велик (например, меньше единицы).

Предположим теперь, что табличные значения y_i вычислены приближенно, т.е. $y_i = \tilde{y}_i + \delta_i$, где для всех значений i справедливы неравенства $|\delta_i| \leq \delta$. Посмотрим, как отразятся эти ошибки на состоянии конечной разности k -го порядка:

$$\Delta^k y_0 = (E - I)^k y_0 = (E - I)^k (\tilde{y}_0 + \delta_0) = (E - I)^k \tilde{y}_0 + (E - I)^k \delta_0.$$

Таким образом, видим, что погрешность конечной разности k -го порядка есть величина

$$(E - I)^k \delta_0 = \delta_k - C_k^1 \delta_{k-1} + C_k^2 \delta_{k-2} - \dots + (-1)^{k-1} C_k^{k-1} \delta_1 + (-1)^k C_k^k \delta_0.$$

Отсюда следует, что

$$|(E - I)^k \delta_0| \leq \delta (1 + C_k^1 + C_k^2 + \dots + C_k^{k-1} + C_k^k) = \delta (1 + 1)^k = 2^k \delta.$$



Окончательно имеем: в наихудшем варианте погрешность может изменяться по закону геометрической прогрессии со знаменателем 2 и, следовательно, быстро возрастать.

Все это, с учетом убывания точных значений конечных разностей, приводит к тому, что при некотором k_0 точные значения конечных разностей становятся сопоставимыми по величине с погрешностями, им соответствующими:

$$\Delta^{k_0} y_i \approx 2^{k_0} \delta,$$

а с дальнейшим ростом k погрешности становятся доминирующими, т.е. при $k > k_0$ таблица конечных разностей уже не несет никакой полезной информации и, следовательно, не должна использоваться в вычислениях (в частности, при построении интерполяционных многочленов). Таблицу конечных разностей, свободную от указанного недостатка, называют *правильной*.



5.3.7. Некоторые правила интерполяирования при равноотстоящих узлах

Интерполярование в начале таблицы

Интерполярование в конце таблицы

Интерполярование внутри таблицы

Исходя из соображений минимизации остатка интерполяирования и поведения вычислительной погрешности в таблице конечных разностей естественно придерживаться следующих принципов:

- 1) в качестве узлов интерполяирования нужно брать узлы, ближайшие к точке интерполяирования;
- 2) значения функции f в узлах, ближайших к точке интерполяирования, должны привлекаться через разности по возможности более низкого порядка.

Таким образом, если рассматриваемая таблица будет таблицей конечных размеров, то можно указать следующие три типа правил:

1. Если точка интерполяирования x лежит в начале таблицы, то естественно привлекать узлы правее ее, включая x_0 ;
2. Если точка интерполяирования x лежит в конце таблицы, то естественно привлекать узлы левее ее, включая x_n ;
3. Если точка интерполяирования x лежит внутри таблицы, то:
 - (a) если она лежит вблизи одного из узлов, то естественно привлекать узлы по одному справа и слева, начиная от этого узла;
 - (b) если она лежит вблизи одного из узлов, то естественно привлекать узлы по одному справа и слева, начиная от этого узла;
 - (c) если она лежит посередине между двумя соседними узлами, то число узлов естественно брать четным.

Рассмотрим эти случаи более подробно.



Интерполярование в начале таблицы

Считая, что для достижения целей достаточно интерполяирования с помощью многочлена k -й степени, привлекаем узлы $x_0, x_0 + h, \dots, x_0 + kh$. Запишем интерполяционный многочлен в форме Ньютона:

$$\begin{aligned} P_k(x) = & f(x_0) + (x - x_0) f(x_0, x_0 + h) + \\ & + (x - x_0)(x - x_0 - h) f(x_0, x_0 + h, x_0 + 2h) + \dots + \\ & + (x - x_0)(x - x_0 - h) \cdots (x - x_0 - kh + h) f(x_0, x_0 + h, \dots, x_0 + kh). \end{aligned}$$

Произведем замену переменной по формуле $t = \frac{x - x_0}{h}$. Так как по предположению $x \in [x_0; x_0 + h]$, то $t \in [0; 1]$. Тогда, учитывая формулу (5.47), после несложных преобразований получим:

$$P_k(x) = P_k(x_0 + th) = y_0 + \frac{t}{1!} \Delta y_0 + \frac{t(t-1)}{2!} \Delta^2 y_0 + \dots + \frac{t(t-1)\cdots(t-k+1)}{k!} \Delta^k y_0. \quad (5.48)$$

Это **интерполяционный многочлен Ньютона для начала таблицы**.

Рассмотрим сейчас остаток интерполяирования. Если $f(x) \in C^{k+1}[a, b]$, то, согласно (5.32),

$$r_k(x) = \omega_{k+1}(x) \frac{f^{(k+1)}(\xi)}{(k+1)!}.$$

Отсюда, учитывая замену переменных, имеем:

$$r_k(x) = r_k(x_0 + th) = h^{k+1} \frac{t(t-1)\cdots(t-k)}{(k+1)!} f^{(k+1)}(\xi).$$

Интерполярование в конце таблицы

В этом случае в качестве узлов интерполяирования берем $x_n, x_n - h, \dots, x_n - kh$. Тогда

$$\begin{aligned} P_k(x) = & f(x_n) + (x - x_n) f(x_n, x_n - h) + \\ & + (x - x_n)(x - x_n + h) f(x_n, x_n - h, x_n - 2h) + \dots + \end{aligned}$$

$$+ (x - x_n) (x - x_n + h) \cdots (x - x_n + kh - h) f(x_n, x_n - h, \dots, x_n - kh).$$

Произведем замену переменной по формуле $t = \frac{x - x_n}{h}$. Так как по предположению $x \in [x_n - h; x_n]$, то $t \in [-1; 0]$. Тогда, учитывая, что

$$f(x_n, x_n - h, \dots, x_n - ih) = \frac{\Delta^i y_{n-i}}{i! \cdot h^i}$$

в силу формулы (5.47), после несложных преобразований получим:

$$P_k(x) = P_k(x_n + th) = y_n + \frac{t}{1!} \Delta y_{n-1} + \frac{t(t+1)}{2!} \Delta^2 y_{n-2} + \cdots + \frac{t(t+1)\cdots(t+k-1)}{k!} \Delta^k y_{n-k}. \quad (5.49)$$

Мы получили *интерполяционный многочлен Ньютона для конца таблицы*. В этом случае

$$r_k(x) = r_k(x_n + th) = h^{k+1} \frac{t(t+1)\cdots(t+k)}{(k+1)!} f^{(k+1)}(\xi).$$

Интерполярование внутри таблицы

Случай 1. Пусть x_n – внутренний узел таблицы. Предположим, что точка интерполяирования x лежит вблизи x_n с той или другой стороны. Табличные узлы для интерполяирования здесь разумно привлекать в следующем порядке: сначала взять x_n , а затем брать пары точек $(x_n + h, x_n - h)$, $(x_n + 2h, x_n - 2h)$, ..., $(x_n + kh, x_n - kh)$. Число привлекаемых для интерполяирования узлов будет нечетным и равным $2k + 1$.

Интерполяционный многочлен Ньютона при таком порядке узлов запишется следующим образом:

$$\begin{aligned} P_{2k}(x) &= f(x_n) + (x - x_n) f(x_n, x_n + h) + \\ &+ (x - x_n) (x - x_n - h) f(x_n, x_n + h, x_n - h) + \cdots + \\ &+ (x - x_n) (x - x_n - h) (x - x_n + h) f(x_n, x_n + h, x_n - h, x_n + 2h) + \\ &+ (x - x_n) (x - x_n - h) \cdots (x - x_n + kh - h) f(x_n, x_n + h, x_n - h, \dots, x_n - kh + h, x_n + kh) + \\ &+ (x - x_n) (x - x_n - h) \cdots (x - x_n + kh - h) (x - x_n - kh) f(x_n, x_n + h, x_n - h, \dots, x_n + kh, x_n - kh). \end{aligned}$$

Вводя замену переменной $t = \frac{x-x_n}{h}$ и учитывая, что $f(x_n) = y_n$, $f(x_n, x_n + h) = \frac{\Delta y_n}{1! \cdot h}$,
 $f(x_n, x_n + h, x_n - h) = \frac{\Delta^2 y_{n-1}}{2! \cdot h^2}$, $f(x_n, x_n + h, x_n - h, x_n + 2h) = \frac{\Delta^3 y_{n-1}}{3! \cdot h^3}$, ..., получим:

$$\begin{aligned}
 P_{2k}(x) &= P_{2k}(x_n + th) = y_n + \frac{t}{1!} \Delta y_n + \frac{t(t-1)}{2!} \Delta^2 y_{n-1} + \frac{t(t-1)(t+1)}{3!} \Delta^3 y_{n-1} + \\
 &+ \frac{t(t-1)(t+1)(t-2)}{4!} \Delta^4 y_{n-2} + \dots + \frac{t(t-1)(t+1) \cdots (t-k+1)(t+k-1)}{(2k-1)!} \Delta^{2k-1} y_{n-k+1} + \\
 &+ \frac{t(t-1)(t+1) \cdots (t-k+1)(t+k-1)(t-k)}{(2k)!} \Delta^{2k} y_{n-k}.
 \end{aligned}$$

Чтобы придать членам правой части симметричный вид, приведем сначала равенство к виду

$$\begin{aligned}
 P_{2k}(x) &= P_{2k}(x_n + th) = y_n + \frac{t}{1!} \left(\Delta y_n - \frac{1}{2} \Delta^2 y_{n-1} \right) + \frac{t^2}{2!} \Delta^2 y_{n-1} + \frac{t(t^2-1^2)}{3!} \left(\Delta^3 y_{n-1} - \frac{1}{2} \Delta^4 y_{n-2} \right) - \\
 &+ \frac{t^2(t^2-1^2)}{4!} \Delta^4 y_{n-2} + \dots + \frac{t(t^2-1^2) \cdots (t^2-(k-1)^2)}{(2k-1)!} \left(\Delta^{2k-1} y_{n-k+1} - \frac{1}{2} \Delta^{2k} y_{n-k} \right) + \\
 &+ \frac{t^2(t^2-1) \cdots (t^2-(k-1)^2)}{(2k)!} \Delta^{2k} y_{n-k}. \quad (*)
 \end{aligned}$$

Преобразуем теперь разности в скобках, исключив конечные разности четного порядка:

$$\Delta y_n - \frac{1}{2} \Delta^2 y_{n-1} = \Delta y_n - \frac{\Delta y_n - \Delta y_{n-1}}{2} = \frac{\Delta y_n + \Delta y_{n-1}}{2},$$

$$\Delta^3 y_{n-1} - \frac{1}{2} \Delta^4 y_{n-2} = \Delta^3 y_{n-1} - \frac{\Delta^3 y_{n-1} - \Delta^3 y_{n-2}}{2} = \frac{\Delta^3 y_{n-1} + \Delta^3 y_{n-2}}{2};$$

.....



Подставляя полученные выражения в (*), окончательно будем иметь:

$$\begin{aligned}
 P_{2k}(x) = P_{2k}(x_n + th) &= y_n + \frac{t}{1!} \cdot \frac{\Delta y_n + \Delta y_{n-1}}{2} + \frac{t^2}{2!} \Delta^2 y_{n-1} + \frac{t(t^2 - 1^2)}{3!} \cdot \frac{\Delta^3 y_{n-1} + \Delta^3 y_{n-2}}{2} + \\
 &+ \frac{t^2(t^2 - 1^2)}{4!} \Delta^4 y_{n-2} + \cdots + \frac{t(t^2 - 1^2) \cdots (t^2 - (k-1)^2)}{(2k-1)!} \cdot \frac{\Delta^{2k-1} y_{n-k+1} + \Delta^{2k-1} y_{n-k}}{2} + \\
 &+ \frac{t^2(t^2 - 1) \cdots (t^2 - (k-1)^2)}{(2k)!} \Delta^{2k} y_{n-k}.
 \end{aligned} \tag{5.50}$$

Это *интерполяционная формула Ньютона – Стирлинга для интерполяирования внутри таблицы*.

По аналогии со случаями (5.48), (5.49) получим представление остатка интерполяирования:

$$r_{2k}(x) = r_{2k}(x_n + th) = h^{2k+1} \frac{t(t^2 - 1^2) \cdots (t^2 - k^2)}{(2k+1)!} f^{(2k+1)}(\xi).$$

Случай 2. Наконец, рассмотрим еще одно правило, предназначенное для использования в том случае, когда точка x лежит вблизи середины отрезка между двумя соседними табличными значениями. Пусть это будут значения x_n и $x_n + h$.

Соображения симметрии побуждают строить интерполяционное правило со следующим порядком привлечения узлов: сначала привлекается пара $(x_n, x_n + h)$, затем последовательно $(x_n - h, x_n + 2h)$, $(x_n - 2h, x_n + 3h)$, ..., $(x_n - kh + h, x_n + kh)$. Число узлов будет четным. Интерполяционный многочлен Ньютона при таком расположении узлов примет следующий вид:

$$\begin{aligned}
 P_{2k-1}(x) &= f(x_n) + (x - x_n) f(x_n, x_n + h) + (x - x_n)(x - x_n - h) f(x_n, x_n + h, x_n - h) + \\
 &+ (x - x_n)(x - x_n - h)(x - x_n + h) f(x_n, x_n + h, x_n - h, x_n + 2h) + \cdots + \\
 &+ (x - x_n)(x - x_n - h) \cdots (x - x_n - kh + h) f(x_n, x_n + h, x_n - h, \dots, x_n + kh - h, x_n - kh + h) + \\
 &+ (x - x_n)(x - x_n - h) \cdots (x - x_n - kh + h)(x - x_n + kh - h) f(x_n, x_n + h, \dots, x_n - kh + h, x_n + kh).
 \end{aligned}$$



Вверх

Назад

Вперёд

Пред.

След.

Указатель

Помощь

Экран

После замены переменной $t = \frac{x-x_n}{h}$ и учитывая, что $f(x_n) = y_n$, $f(x_n, x_n + h) = \frac{\Delta y_n}{1! \cdot h}$,
 $f(x_n, x_n + h, x_n - h) = \frac{\Delta^2 y_{n-1}}{2! \cdot h^2}$, $f(x_n, x_n + h, x_n - h, x_n + 2h) = \frac{\Delta^3 y_{n-1}}{3! \cdot h^3}$, ..., получим:

$$\begin{aligned} P_{2k-1}(x) &= P_{2k-1}(x_n + th) = y_n + \frac{t}{1!} \Delta y_n + \frac{t(t-1)}{2!} \Delta^2 y_{n-1} + \frac{t(t-1)(t+1)}{3!} \Delta^3 y_{n-1} + \\ &+ \frac{t(t-1)(t+1)(t-2)}{4!} \Delta^4 y_{n-2} + \dots + \frac{t(t-1)(t+1)\dots(t+k-2)(t-k+1)}{(2k-2)!} \Delta^{2k-2} y_{n-k+1} + \\ &+ \frac{t(t-1)(t+1)\dots(t-k+1)(t+k-1)}{(2k-1)!} \Delta^{2k-1} y_{n-k+1}. \end{aligned}$$

Для приведения членов правой части к виду, симметричному относительно точки $x_n + \frac{1}{2}h$ отделим от четных разностей половины их значений $\frac{1}{2}y_n$, $\frac{1}{2}\Delta^2 y_{n-1}$, $\frac{1}{2}\Delta^4 y_{n-2}$, ... и заменим эти значения при помощи тождеств

$$\frac{1}{2}y_n = \frac{1}{2}(y_{n+1} - \Delta y_n), \quad \frac{1}{2}\Delta^2 y_{n-1} = \frac{1}{2}(\Delta^2 y_n - \Delta^3 y_{n-1}), \dots$$



После приведения подобных членов получим:

$$\begin{aligned} P_{2k-1}(x_n + th) = & \frac{y_n + y_{n+1}}{2} + \frac{t - \frac{1}{2}}{1!} \Delta y_n + \frac{t(t-1)}{2!} \frac{\Delta^2 y_n + \Delta^2 y_{n-1}}{2} + \frac{t(t-1)(t-\frac{1}{2})}{3!} \Delta^3 y_{n-1} + \\ & + \frac{t(t^2 - 1^2)(t-2)}{4!} \frac{\Delta^4 y_{n-1} + \Delta^4 y_{n-2}}{2} + \frac{t(t^2 - 1^2)(t-2)(t-\frac{1}{2})}{5!} \Delta^5 y_{n-2} + \dots + \\ & + \frac{t(t^2 - 1^2) \dots (t^2 - (k-2)^2)(t-k+1)}{(2k-2)!} \cdot \frac{\Delta^{2k-2} y_{n-k+1} + \Delta^{2k-2} y_{n-k+2}}{2} + \\ & + \frac{t(t^2 - 1^2) \dots (t^2 - (k-2)^2)(t-k+1)(t-\frac{1}{2})}{(2k)!} \Delta^{2k} y_{n-k}. \quad (5.51) \end{aligned}$$

Формула (5.51) называется *интерполяционной формулой Ньютона – Бесселя для интерполяирования внутри таблицы*.

Остаток интерполяирования будет выглядеть следующим образом:

$$r_{2k-1}(x) = r_{2k-1}(x_n + th) = h^{2k} \frac{t(t^2 - 1^2) \dots (t^2 - (k-1)^2)(t-k)}{(2k)!} f^{(2k)}(\xi).$$



5.3.8. Интерполярование с кратными узлами

Остаточный член интерполяционной формулы Эрмита

Частные случаи интерполяирования Эрмита

Ранее мы уже формулировали постановку задачи интерполяирования с кратными узлами. Напомним ее. Пусть на отрезке $[a, b]$ заданы $(m + 1)$ различных узлов интерполяирования x_0, x_1, \dots, x_m . Рассмотрим функцию $f(x)$ и будем считать, что в точке x_0 известны значения как самой функции $f(x_0)$, так и ее производные $f'(x_0), \dots, f^{(\alpha_0-1)}(x_0)$; в точке x_1 заданы значения $f(x_1), f'(x_1), \dots, f^{(\alpha_1-1)}(x_1)$ и т.д. Числа $\alpha_0, \alpha_1, \dots, \alpha_m$ называются кратностями соответствующих узлов. Общее число всех исходных данных о функции $f(x)$ обозначим через $n + 1$: $\alpha_0 + \alpha_1 + \dots + \alpha_m = n + 1$.

Поставим задачей найти многочлен

$$P_n(x) = a_0x^n + a_1x^{n-1} + \dots + a_{n-1}x + a_n$$

степени не выше n , удовлетворяющий условиям

$$P_n^{(i)}(x_k) = f^{(i)}(x_k), \quad i = \overline{0, \alpha_k - 1}; \quad k = \overline{0, m}. \quad (5.52)$$

Эти условия, как и в случае простого интерполяирования (по значениям функции), дадут для определения коэффициентов a_0, a_1, \dots, a_n искомого многочлена $P_n(x)$ систему линейных алгебраических уравнений. Чтобы убедиться в существовании и единственности ее решения, достаточно доказать, что соответствующая ей однородная система

$$P_n^{(i)}(x_k) = 0, \quad i = \overline{0, \alpha_k - 1}; \quad k = \overline{0, m} \quad (5.53)$$

имеет только нулевое решение. Но система (5.53) для многочлена $P_n(x)$ говорит о том, что узлы x_0, x_1, \dots, x_m должны быть корнями $P_n(x)$ кратностей не меньше, чем $\alpha_0, \alpha_1, \dots, \alpha_m$ соответственно. Сумма кратностей должна быть больше или равна $\alpha_0 + \alpha_1 + \dots + \alpha_m = n + 1$. Но степень $P_n(x)$ – не выше n . Поэтому иметь сумму кратностей корней, большую n , многочлен $P_n(x)$ может только в том случае, когда он тождественно равен нулю. Тогда все его коэффициенты равны нулю и однородная система (5.53) имеет только нулевое решение.

Таким образом, рассматриваемая интерполяционная задача с кратными узлами разрешима и имеет только одно решение, каковы бы ни были значения $f^{(i)}(x_k)$ в (5.52). Как и в случае алгебраического интерполяирования по значениям функции, для $P_n(x)$ легко выписать явное выражение через определители. Но из тех же соображений, что и ранее, получим явное представление для него. Для этих целей воспользуемся конструкцией типа (5.23) (т.е. представлением через функции влияния). Построим многочлены $H_{ij}(x)$ степени не выше n , удовлетворяющие условиям

$$\begin{cases} H_{ij}(x_k) = H'_{ij}(x_k) = \dots = H_{ij}^{(\alpha_k-1)}(x_k) = 0, & k \neq i, \\ H_{ij}(x_i) = H'_{ij}(x_i) = \dots = H_{ij}^{(j-1)}(x_i) = H_{ij}^{(j+1)}(x_i) = \dots = H_{ij}^{(\alpha_i-1)}(x_i) = 0, \\ H_{ij}^{(j)}(x_i) = 1, & i = \overline{0, m}; \quad j = \overline{0, \alpha_i - 1}. \end{cases} \quad (5.54)$$

Тогда по аналогии с (5.23) можно будет записать следующее представление для интерполяционного многочлена:

$$P_n(x) = \sum_{i=0}^m \sum_{j=0}^{\alpha_i-1} H_{ij}(x) f^{(j)}(x_i). \quad (5.55)$$

Так как $H_{ij}(x)$ обладает в точках $x_0, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_m$ нулями кратностей $\alpha_0, \alpha_1, \dots, \alpha_{i-1}, \alpha_{i+1}, \dots, \alpha_m$ соответственно, а в точке x_i — нулем кратности j , то

$$H_{ij}(x) = (x - x_0)^{\alpha_0} (x - x_1)^{\alpha_1} \cdots (x - x_{i-1})^{\alpha_{i-1}} (x - x_i)^j (x - x_{i+1})^{\alpha_{i+1}} \cdots (x - x_m)^{\alpha_m} \bar{H}_{ij}(x), \quad (5.56)$$

где $\bar{H}_{ij}(x)$ — многочлен степени $\alpha_i - j - 1$, не обращающийся в нуль при $x = x_i$. Запишем его по степеням двучлена $(x - x_i)$:

$$\bar{H}_{ij}(x) = A_{ij}^{(0)} + A_{ij}^{(1)}(x - x_i) + \cdots + A_{ij}^{(\alpha_i-j-1)}(x - x_i)^{\alpha_i-j-1}. \quad (5.57)$$

Вводя в рассмотрение многочлен $\Omega(x) = (x - x_0)^{\alpha_0} (x - x_1)^{\alpha_1} \cdots (x - x_m)^{\alpha_m}$, равенство (5.56) можем переписать в виде

$$H_{ij}(x) = \frac{\Omega(x)}{(x - x_i)^{\alpha_i}} \cdot (x - x_i)^j \cdot \bar{H}_{ij}(x),$$

откуда, учитывая представление (5.56), получим равенство

$$\bar{H}_{ij}(x) = A_{ij}^{(0)} + A_{ij}^{(1)}(x - x_i) + \cdots + A_{ij}^{(\alpha_i - j - 1)}(x - x_i)^{\alpha_i - j - 1} = \frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \cdot \frac{(x - x_i)^j}{H_{ij}(x)}.$$

Подставляя сюда $x = x_i$, получим:

$$A_{ij}^{(0)} = \lim_{x \rightarrow x_i} \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \cdot \frac{H_{ij}(x)}{(x - x_i)^j} \right].$$

Первое отношение непрерывно при $x = x_i$. Следовательно,

$$\lim_{x \rightarrow x_i} \frac{(x - x_i)^{\alpha_i}}{\Omega(x)} = \frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \Bigg|_{x = x_i}.$$

Предел второго отношения найдем по правилу Лопитала (хотя оно так же, как и первое, является непрерывным при $x = x_i$, воспользоваться равенством, аналогичным предыдущему, мы не можем, поскольку пока не знаем явного представления для $H_{ij}(x)$):

$$\lim_{x \rightarrow x_i} \frac{H_{ij}(x)}{(x - x_i)^j} = \lim_{x \rightarrow x_i} \frac{H_{ij}^{(j)}(x)}{j!}.$$

Таким образом, в силу последнего из условий (5.54) данный предел равен $\frac{1}{j!}$.

Следовательно,

$$A_{ij}^{(0)} = \frac{1}{j!} \cdot \frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \Bigg|_{x = x_i} \quad (5.58)$$

Аналогично для всех оставшихся коэффициентов представления (5.57) имеем:

$$A_{ij}^{(k)} = \frac{1}{k!} \lim_{x \rightarrow x_i} \frac{d^k}{dx^k} \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \cdot \frac{H_{ij}(x)}{(x - x_i)^j} \right]. \quad (5.59)$$

Применим правило Лейбница для дифференцирования произведения:

$$\frac{d^k}{dx^k} \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \cdot \frac{H_{ij}(x)}{(x - x_i)^j} \right] = \sum_{p=0}^k C_k^p \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \right]^{(p)} \cdot \left[\frac{H_{ij}(x)}{(x - x_i)^j} \right]^{(k-p)}. \quad (5.60)$$

Как и выше, производные $\left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \right]^{(p)}$ непрерывны в точке $x = x_i$. Поэтому

$$\lim_{x \rightarrow x_i} \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \right]^{(p)} = \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \right]^{(p)} \Big|_{x=x_i}.$$

Для отыскания $\lim_{x \rightarrow x_i} \left[\frac{H_{ij}(x)}{(x - x_i)^j} \right]^{(k-p)}$ воспользуемся тем же приемом, что и для отыскания коэффициентов $A_{ij}^{(k)}$. Записывая разложение многочлена $H_{ij}(x)$ по степеням двучлена $(x - x_i)$, получим:

$$H_{ij}(x) = B_{ij}^{(0)}(x - x_i)^j + B_{ij}^{(1)}(x - x_i)^{j+1} + \cdots + B_{ij}^{(n-j)}(x - x_i)^n.$$

Учитывая вторую и третью строки соотношений (5.54), имеем:

$$B_{ij}^{(0)} = \frac{1}{j!} \cdot H_{ij}^{(j)}(x_i) = \frac{1}{j!}, \quad B_{ij}^{(s)} = \frac{1}{(j+s)!} \cdot H_{ij}^{(j+s)}(x_i) = 0, \quad s = 1, \dots, \alpha_i - j - 1.$$

Поэтому

$$\frac{H_{ij}(x)}{(x - x_i)^j} = B_{ij}^{(0)} + B_{ij}^{(\alpha_i - j)}(x - x_i)^{\alpha_i - j} + \cdots + B_{ij}^{(n-j)}(x - x_i)^{n-j}$$

и, поскольку $k - p \leq k \leq \alpha_i - j - 1$,

$$\lim_{x \rightarrow x_i} \left[\frac{H_{ij}(x)}{(x - x_i)^j} \right]^{(k-p)} = \begin{cases} B_{ij}^{(0)} = \frac{1}{j!}, & \text{если } k - p = 0, \\ 0, & \text{если } k - p \neq 0. \end{cases}$$

Тогда из (5.59), (5.60) находим:

$$A_{ij}^{(k)} = \frac{1}{k!} \cdot \lim_{x \rightarrow x_i} \frac{d^k}{dx^k} \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \cdot \frac{H_{ij}(x)}{(x - x_i)^j} \right] = \frac{1}{k!} \cdot \frac{1}{j!} \cdot \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \right]^{(k)} \Bigg|_{x=x_i}$$

Подставляя найденные коэффициенты в (5.57), получим явное выражение для $\bar{H}_{ij}(x)$:

$$\bar{H}_{ij}(x) = \frac{1}{j!} \sum_{k=0}^{\alpha_i-j-1} \frac{1}{k!} \cdot \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \right]^{(k)} \Bigg|_{x=x_i} \cdot (x - x_i)^k.$$

Следовательно,

$$\begin{aligned} H_{ij}(x) &= \frac{\Omega(x)}{(x - x_i)^{\alpha_i-j}} \bar{H}_{ij}(x) = \\ &= \frac{1}{j!} \cdot \frac{\Omega(x)}{(x - x_i)^{\alpha_i-j}} \cdot \sum_{k=0}^{\alpha_i-j-1} \frac{1}{k!} \cdot \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \right]^{(k)} \Bigg|_{x=x_i} \cdot (x - x_i)^k. \end{aligned}$$

Наконец, формула (5.23) дает искомое явное представление интерполяционного многочлена:

$$P_n(x) = \sum_{i=0}^m \sum_{j=0}^{\alpha_i-1} \sum_{k=0}^{\alpha_i-j-1} \frac{1}{j!} \cdot \frac{1}{k!} \cdot \frac{\Omega(x)}{(x - x_i)^{\alpha_i-j}} \cdot \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \right]^{(k)} \Bigg|_{x=x_i} \cdot (x - x_i)^k \cdot f^{(j)}(x_i). \quad (5.61)$$

Формула (5.61) определяет *интерполяционный многочлен Эрмита*.

Замечание 5.4. Введем понятие разделенных разностей с кратными узлами:

$$f \left(\underbrace{x_0, x_0, \dots, x_0}_{j+1} \right) = \lim_{\varepsilon \rightarrow 0} f(x_{00}^\varepsilon, x_{01}^\varepsilon, \dots, x_{0j}^\varepsilon),$$

где все узлы x_{0i}^ε различны и $\lim_{\varepsilon \rightarrow 0} x_{0i}^\varepsilon = x_0$ для всех $i = \overline{0, j}$;

$$f \left(\underbrace{x_0, \dots, x_0}_{j_0}; \dots; \underbrace{x_p, \dots, x_p}_{j_p} \right) = \\ = \frac{f \left(\underbrace{x_0, \dots, x_0}_{j_0-1}; \dots; \underbrace{x_p, \dots, x_p}_{j_p} \right) - f \left(\underbrace{x_0, \dots, x_0}_{j_0}; \dots; \underbrace{x_p, \dots, x_p}_{j_p-1} \right)}{x_p - x_0}.$$

Тогда, записав по узлам $x_{00}^\varepsilon, \dots, x_{0\alpha_0-1}^\varepsilon; x_{10}^\varepsilon, \dots, x_{1\alpha_1-1}^\varepsilon; \dots; x_{m0}^\varepsilon, \dots, x_{m\alpha_m-1}^\varepsilon$ обычный интерполяционный многочлен в форме Ньютона (все узлы различны!) и переходя в нем к пределу при $\varepsilon \rightarrow 0$, можем получить другое представление многочлена Эрмита:

$$P_n(x) = f(x_0) + (x - x_0) f(x_0, x_0) + \dots + (x - x_0)^{\alpha_0-1} f(x_0, \dots, x_0) + (x - x_0)^{\alpha_0} f(x_0, \dots, x_0; x_1) + \\ + (x - x_0)^{\alpha_0} (x - x_1) f(x_0, \dots, x_0; x_1, x_1) + \dots + (x - x_0)^{\alpha_0} (x - x_1)^{\alpha_1-1} f(x_0, \dots, x_0; x_1, \dots, x_1) + \dots + (*) \\ + (x - x_0)^{\alpha_0} (x - x_1)^{\alpha_1} \dots (x - x_m)^{\alpha_m-1} f(x_0, \dots, x_0; x_1, \dots, x_1; \dots; x_m, \dots, x_m).$$

Остаточный член интерполяционной формулы Эрмита

Пусть как и в случае интерполяирования по значениям функции $f(x) \in C^{n+1}[a, b]$. По аналогии с выводом остатка интерполяирования в форме Лагранжа рассмотрим вспомогательную функцию

$$\varphi(t) = f(t) - P_n(t) - K\Omega(t),$$

где K – некоторая постоянная. Функция $\varphi(t)$ имеет нуль x_0 кратности α_0 , нуль x_1 кратности α_1 и т.д. нуль x_m кратности α_m . Подберем постоянную K так, чтобы $\varphi(t)$ обратилась в нуль в точке x , для которой мы проводим интерполярование. Тогда

$$K = \frac{f(x) - P_n(x)}{\Omega(x)} = \frac{r_n(x)}{\Omega(x)}.$$

В таком случае функция $\varphi(t)$ будет иметь на отрезке $[a, b]$ в общей сложности не менее $(n + 2)$ нулей (каждый считаем столько раз, какова его кратность). На основании теоремы Ролля производная $\varphi'(t)$ обратится в нуль в $(m + 1)$ различных точках в интервалах между x, x_0, \dots, x_m и, кроме того, будет иметь нули кратностей $\alpha_0 - 1, \dots, \alpha_m - 1$ в точках x_0, \dots, x_m , т.е. всего $(n + 1)$ нулей на отрезке $[a, b]$.

Рассуждая аналогично, получим, что вторая производная $\varphi''(t)$ будет иметь по крайней мере n нулей и т.д., производная порядка $(n + 1)$ на отрезке $[a, b]$ будет иметь по крайней мере один нуль, т.е. на отрезке $[a, b]$ найдется по крайней мере одна точка ξ такая что $\varphi^{(n+1)}(\xi) = 0$, а так как

$$\varphi^{(n+1)}(\xi) = f^{(n+1)}(\xi) - K \cdot (n+1)!,$$

то отсюда $K = \frac{f^{(n+1)}(\xi)}{(n+1)!}$ и окончательно, приравнивая два различных представления для K , получим:

$$r_n(x) = f(x) - P_n(x) = \Omega(x) \frac{f^{(n+1)}(\xi)}{(n+1)!}. \quad (5.62)$$

Частные случаи интерполяирования Эрмита

- Пусть $\alpha_0 = \alpha_1 = \dots = \alpha_m = 2$, т.е. кратность всех узлов интерполяирования равна двум. Тогда, очевидно, $\Omega(x) = \omega_{m+1}^2(x)$ и формула (5.61) примет вид

$$\begin{aligned} P_n(x) &= \sum_{i=0}^m \left[\frac{\omega_{m+1}^2(x)}{(x - x_i)^2} \cdot \left[\frac{x - x_i}{\omega_{m+1}(x)} \right]^2 \right] \Bigg|_{x=x_i} + \\ &+ \frac{\omega_{m+1}^2(x)}{x - x_i} \cdot \left[\left(\frac{x - x_i}{\omega_{m+1}(x)} \right)^2 \right]' \Bigg|_{x=x_i} \cdot f(x_i) + \\ &+ \sum_{i=0}^m \frac{\omega_{m+1}^2(x)}{x - x_i} \cdot \left[\frac{x - x_i}{\omega_{m+1}(x)} \right]^2 \Bigg|_{x=x_i} \cdot f'(x_i). \end{aligned}$$

Так как

$$\left. \frac{x - x_i}{\omega_{m+1}(x)} \right|_{x=x_i} = \frac{1}{\omega'_{m+1}(x_i)},$$

а

$$\begin{aligned} & \left[\left(\frac{x - x_i}{\omega_{m+1}(x)} \right)^2 \right]' \Big|_{x=x_i} = 2 \lim_{x \rightarrow x_i} \frac{x - x_i}{\omega_{m+1}(x)} \cdot \lim_{x \rightarrow x_i} \frac{\omega_{m+1}(x) - (x - x_i) \omega'_{m+1}(x)}{\omega_{m+1}^2(x)} = \\ & = 2 \cdot \frac{1}{\omega'_{m+1}(x_i)} \cdot \lim_{x \rightarrow x_i} \frac{\omega'_{m+1}(x) - \omega'_{m+1}(x) - (x - x_i) \omega''_{m+1}(x)}{2\omega_{m+1}(x) \omega'_{m+1}(x)} = \\ & = \frac{1}{\omega'_{m+1}(x_i)} \cdot \frac{-\omega''_{m+1}(x_i)}{\omega'_{m+1}(x_i)} \cdot \lim_{x \rightarrow x_i} \frac{x - x_i}{\omega_{m+1}(x)} = -\frac{\omega''_{m+1}(x_i)}{[\omega'_{m+1}(x_i)]^3}, \end{aligned}$$

то

$$P_n(x) = \sum_{i=0}^m \left[\frac{\omega_{m+1}(x)}{(x - x_i) \omega'_{m+1}(x_i)} \right]^2 \left\{ \left[1 - (x - x_i) \frac{\omega''_{m+1}(x_i)}{\omega'_{m+1}(x_i)} \right] f(x_i) + (x - x_i) f'(x_i) \right\}. \quad (5.63)$$

Из (5.62) получаем соответствующее представление для остатка:

$$r_n(x) = \omega_{m+1}^2(x) \frac{f^{(2m+2)}(\xi)}{(2m+2)!}. \quad (5.64)$$

2. Другой интересный частный случай имеет место, если рассмотреть один $(n+1)$ -кратный узел. В этом случае $\Omega(x) = (x - x_0)^{n+1}$; $n = 0$, $\alpha_0 = n+1$; $\frac{(x-x_0)^{n+1}}{\Omega(x)} = 1$. Таким образом, формула (5.61), с учетом

$$\left[\frac{(x - x_0)^{n+1}}{\Omega(x)} \right]^{(k)} = \begin{cases} 1 & \text{при } k = 0, \\ 0 & \text{при } k > 0 \end{cases},$$



будет иметь вид

$$P_n(x) = \sum_{j=0}^n \frac{1}{j!} \cdot (x - x_0)^j f^{(j)}(x_0), \quad (5.65)$$

т.е. представляет собой всем хорошо знакомый отрезок ряда Тейлора. Остаток, как и положено, будет

$$r_n(x) = (x - x_0)^{n+1} \frac{f^{(n+1)}(\xi)}{(n+1)!}, \quad (5.66)$$

т.е. является стандартным остатком ряда Тейлора в форме Лагранжа.



5.3.9. Сходимость интерполяционного процесса

Пусть на отрезке $[a, b]$ рассматривается функция $f(x)$ с конечными значениями (ограничимся случаем простого алгебраического интерполяирования по значениям функции).

Рассмотрим таблицу узлов интерполяирования (причем все $x_i^{(k)} \in [a, b]$)

$$X = \left\{ \begin{array}{cccc} x_0^{(0)} & & & \\ x_0^{(1)} & x_1^{(1)} & & \\ x_0^{(2)} & x_1^{(2)} & x_2^{(2)} & \\ \dots & \dots & & \\ x_0^{(n)} & x_1^{(n)} & \dots & x_n^{(n)} \\ \dots & \dots & & \end{array} \right\}.$$

По каждой строке этой таблицы будем строить интерполяционный многочлен. Такой процесс называется **интерполяционным**. Если последовательность интерполяционных многочленов сходится к $f(x)$ при $n \rightarrow \infty$ ($P_n(x) \xrightarrow{n \rightarrow \infty} f(x)$), то говорят, что **интерполяционный процесс сходится**. Сходимость, очевидно, зависит от свойств функции $f(x)$ и от свойств таблицы узлов X .

Приведем без доказательства сводку основных результатов, касающихся сходимости интерполяционных процессов.

Естественно ожидать, что в случае произвольной таблицы X требования к функции $f(x)$ должны быть самыми жесткими. Оказывается, что в этом случае она должна быть *регулярной* в некоторой области комплексной плоскости. Справедлива

Теорема 5.8. *Если $f(x)$ регулярна в замкнутой области σ , то для любой таблицы X узлов интерполяирования из отрезка $[a, b]$ соответствующий интерполяционный процесс будет сходиться равномерно по x на этом отрезке.*

Если таблица X узлов интерполяирования становится равномерной, то это приводит к уменьшению размеров области, в которой требуется регулярность функции $f(x)$. Пусть отрезок $[a, b]$ приведен линейной заменой к отрезку $[0; 1]$. Тогда границей соответствующей области должен быть контур, задаваемый уравнением $\int_0^1 \ln \frac{1}{|t-z|} dt = 0$.

Теорема 5.9. Если $f(x)$ регулярна в замкнутой области θ , то соответствующий интерполяционный процесс будет сходиться равномерно относительно x на отрезке $[0; 1]$ в случае таблицы узлов, равномерно расположенных на $[0; 1]$.

Самые слабые условия накладываются на функцию $f(x)$ в случае, если таблица X узлов составлена по корням многочленов Чебышева: $x_k^{(n)} = \cos \frac{(2k+1)\pi}{2(n+1)}$, $k = 0, n$. Тогда справедлива

Теорема 5.10. В случае таблицы X узлов Чебышева рассматриваемый интерполяционный процесс сходится равномерно относительно x на отрезке $[-1, 1]$ для любой абсолютно непрерывной функции $f(x)$.

Замечание 5.5. Функция $f(x)$ называется *абсолютно непрерывной*, если она представима в виде $f(x) = \int_{-1}^x \varphi(t) dt$, где $\varphi(t)$ – абсолютно интегрируемая функция.

В то же время, если функция $f(x)$ только непрерывна на отрезке $[-1, 1]$, то Бернштейн показал, что существует пример функции, для которой интерполяционный процесс по равномерной таблице узлов не сходится почти ни в одной точке из отрезка $[-1, 1]$ (так, для функции $f(x) = |x|$ сходимость имеет место лишь в точках $-1, 0, 1$). Более того, справедлива

Теорема 5.11 (Флобер). Не существует такой таблицы узлов X из отрезка $[a, b]$, для которой соответствующий интерполяционный процесс был бы равномерно сходящимся по x на отрезке $[a, b]$ для любой непрерывной функции $f(x)$.

С другой стороны, для каждой конкретной непрерывной на $[a, b]$ функции $f(x)$ задача о выборе соответствующей таблицы узлов разрешима, о чем говорит

Теорема 5.12 (Мартинкевич). Для каждой непрерывной на отрезке $[a, b]$ функции $f(x)$ можно указать такую таблицу узлов X из этого отрезка, чтобы соответствующий интерполяционный процесс сходился равномерно по x на отрезке $[a, b]$.



Причем последнее утверждение связано со сходимостью последовательности полиномов наилучшего равномерного приближения к функции $f(x)$.

Приведенные выше результаты говорят о том, что практическое использование интерполяционных многочленов высокой степени представляется не слишком целесообразным.



5.3.10. Некоторые приложения интерполяирования

Приближенное вычисление производных

Применение интерполяирования к решению уравнений

Приближенное вычисление производных

К приближенному вычислению производных (или к численному дифференцированию) приходится прибегать в том случае, когда функция $f(x)$, для которой нужно найти производную, задана таблично или же функциональная зависимость x и $f(x)$ имеет очень сложное аналитическое выражение. В первом случае методы дифференциального исчисления вообще не применимы, а во втором случае их использование вызывает значительные трудности.

В этих случаях вместо функции $f(x)$ используют интерполирующую функцию $\varphi(x)$ и считают производную от $f(x)$ приближенно равной производной от $\varphi(x)$. При этом производная от $f(x)$ будет найдена с некоторой погрешностью. Поэтому наряду с установлением правил вычисления производных одной из основных задач является оценка погрешности, допускаемой при этих вычислениях.

Итак, пусть для $f(x) \in C^{n+1} [a, b]$ известны ее значения в точках x_0, x_1, \dots, x_n . По этим данным нужно найти значение производной порядка m в любой точке $x \in [a, b]$.

Для решения поставленной задачи в соответствии с изложенной выше идеей прибегнем к алгебраическому интерполяированию. Пусть $P_n(x)$ – соответствующий интерполяционный многочлен. Тогда

$$f(x) = P_n(x) + r_n(x).$$

Вычисляя производные порядка m от обеих частей равенства, получим:

$$f^{(m)}(x) = P_n^{(m)}(x) + r_n^{(m)}(x).$$

Если пренебречь здесь величиной $r_n^{(m)}(x)$, получим численное выражение для нахождения нужной производной:

$$f^{(m)}(x) \approx P_n^{(m)}(x),$$

погрешность которого равна $r_n^{(m)}(x)$. При применении правила мы должны, очевидно, считать $m \leq n$, так как все производные от $P_n(x)$ порядка выше n равны нулю тождественно.

В общем случае существенно упростить выражение для остатка не удается, но если $x \in (\alpha; \beta)$ – наименьшему отрезку, на котором находятся узлы x_0, x_1, \dots, x_n , то справедливо представление

$$r_n^{(m)}(x) = \omega_{n+1}^{(m)}(x) \frac{f^{(n+1)}(\xi)}{(n+1)!}. \quad (5.67)$$

Для доказательства этого соотношения, как и ранее, введем вспомогательную функцию

$$\varphi(t) = f(t) - P_n(t) - K\omega_{n+1}(t).$$

По построению $\varphi(t)$ имеет на отрезке $[\alpha; \beta]$ по крайней мере $(n+1)$ корень. Тогда по теореме Ролля производная $\varphi^{(m)}(t)$ внутри отрезка $[\alpha; \beta]$ имеет по крайней мере $n-m+1$ корень. Выберем теперь постоянную K так, чтобы при $t=x$ производная $\varphi^{(m)}(t)$ также обращалась в нуль, т.е.

$$f^{(m)}(x) - P_n^{(m)}(x) - K\omega_{n+1}^{(m)}(x) = 0.$$

Убедимся, что такой выбор возможен, т.е. что $\omega_{n+1}^{(m)}(x) \neq 0$. Так как $\omega_{n+1}(t)$ на $[\alpha; \beta]$ имеет ровно $(n+1)$ корень, то все корни $\omega_{n+1}^{(m)}(t)$ при $m > 0$ будут лежать внутри $(\alpha; \beta)$, а поскольку $x \in (\alpha; \beta)$, то $\omega_{n+1}^{(m)}(x) \neq 0$. Следовательно,

$$K = \frac{r_n^{(m)}(x)}{\omega_{n+1}^{(m)}(x)}.$$

Тогда $\varphi^{(m)}(t)$ имеет, по крайней мере, $n-m+2$ корня и по теореме Ролля существует точка $\xi \in (a, b)$, в которой выполняется равенство $\varphi^{(n+1)}(\xi) = 0$, откуда и следует доказываемое равенство (5.67).

Приведем примеры конкретных правил для вычисления производных. При этом будем исходить из интерполяционной формулы Ньютона для неравных промежутков:

$$\begin{aligned} f(x) \approx & f(x_0) + (x - x_0) f(x_0, x_1) + (x - x_0)(x - x_1) f(x_0, x_1, x_2) + \cdots + \\ & + (x - x_0) \cdots (x - x_1) f(x_0, \dots, x_n). \end{aligned}$$



Для сокращения записи обозначим $\alpha_i = x - x_i$. Тогда, дифференцируя обе части последнего равенства, получим:

$$f'(x) \approx f(x_0, x_1) + (\alpha_0 + \alpha_1) f(x_0, x_1, x_2) + (\alpha_0 \alpha_1 + \alpha_0 \alpha_2 + \alpha_1 \alpha_2) f(x_0, x_1, x_2, x_3) + \dots,$$

$$f''(x) \approx 2! f(x_0, x_1, x_2) + 2(\alpha_0 + \alpha_1 + \alpha_2) f(x_0, x_1, x_2, x_3) + \dots,$$

(5.68)

$$f'''(x) \approx 3! f(x_0, x_1, x_2, x_3) + 3!(\alpha_0 + \alpha_1 + \alpha_2 + \alpha_3) f(x_0, x_1, x_2, x_3, x_4) + \dots,$$

.....

$$f^{(k)}(x) \approx k! [f(x_0, x_1, \dots, x_k) + (\alpha_0 + \alpha_1 + \dots + \alpha_k) f(x_0, x_1, \dots, x_{k+1}) + \dots].$$

Формулы (5.68) несколько упрощаются, если в качестве точки x брать один из узлов интерполяирования.

Совершенно аналогично могут быть получены выражения производных через конечные разности в случае равноотстоящих узлов.

Если, например, исходить из [правил Ньютона для начала таблицы](#)

$$y(x_0 + th) \approx y_0 + \frac{t}{1!} \Delta y_0 + \frac{t(t-1)}{2!} \Delta^2 y_0 + \frac{t(t-1)(t-2)}{3!} \Delta^3 y_0 + \dots,$$

то получим следующие выражения для производных:

$$hy'(x_0 + th) \approx \Delta y_0 + \frac{2t-1}{2!} \Delta^2 y_0 + \frac{3t^2-6t+2}{3!} \Delta^3 y_0 + \dots,$$

$$h^2 y''(x_0 + th) \approx \Delta^2 y_0 + \frac{6(t-1)}{3!} \Delta^3 y_0 + \dots,$$

.....



Вверх

Назад

Вперёд

Пред.

След.

Указатель

Помощь

Экран

При $x = x_0$ переменная t примет значение, равное нулю, и тогда

$$hy'(x_0) \approx \Delta y_0 - \frac{1}{2}\Delta^2 y_0 + \frac{1}{3}\Delta^3 y_0 - \dots,$$

$$h^2 y''(x_0) \approx \Delta^2 y_0 - \Delta^3 y_0 + \dots,$$

.....

Выражая конечные разности через значения функции и приводя подобные, получим:

$$y'(x_0) \approx \frac{y_1 - y_0}{h}, \quad ()$$

$$y'(x_0) \approx \frac{-y_2 + 4y_1 - 3y_0}{2h}, \quad ()$$

.....

Аналогичные выражения могут быть получены и для производных более высокого порядка.

Применение интерполяирования к решению уравнений

Вначале заметим, что описанные ранее алгоритмы методов хорд и секущих могут быть интерпретированы следующим образом: решая уравнение $f(x) = 0$, мы заменяем функцию $f(x)$ интерполяционным многочленом первой степени $P_1(x)$ по двум узлам: x_0 и x_n для первого и x_{n-1} , x_n для второго. После этого решается уравнение $P_1(x) = 0$ и его корень принимается за следующее приближение к корню.

Применение аналогичного приема к трем последним приближениям (x_{n-2}, x_{n-1}, x_n) приводит к *методу парабол*. Записывая интерполяционный многочлен Ньютона по указанным узлам, получим:

$$f(x_n) + (x - x_n)f(x_{n-1}, x_n) + (x - x_n)(x - x_{n-1})f(x_{n-2}, x_{n-1}, x_n) = 0. \quad (5.69)$$

Корень этого квадратного уравнения, наиболее близкий к x_n , принимается за следующее приближение.

Увеличение степени аппроксимирующего многочлена приводит, однако, не только к увеличению скорости сходимости, но и к необходимости решать алгебраические уравнения все более высокой степени, что представляется весьма существенным недостатком. От него можно избавиться, используя интерполяирование обратной функции. Поясним это. Наряду с функцией $y = f(x)$ рассмотрим обратную функцию $x = F(y)$, определенную в окрестности, содержащей единственный корень x^* уравнения $f(x) = 0$. Тогда, очевидно, этот корень можно вычислить по формуле $x^* = F(0)$. При этом таблицу пар чисел $(x_{n-i}, y_{n-i} = f(x_{n-i}))$, $i = \overline{0, k}$, можно рассматривать и как таблицу значений функции $y = f(x)$, так и как таблицу значений функции $x = F(y)$, т.е. как таблицу пар $(y_{n-i}, x_{n-i} = F(y_{n-i}))$, $i = \overline{0, k}$. Проинтегрируем функцию $F(y)$ по значениям x_{n-i} в точках y_{n-i} :

$$F(y) \approx P_k(y) = \sum_{i=0}^k \frac{\omega_{k+1}(y)}{(y - y_{n-i}) \omega'_{k+1}(y_{n-i})} x_{n-i}, \quad \omega_{k+1}(y) = (y - y_n) \dots (y - y_{n-k}).$$

Полагая здесь $y = 0$, найдем очередное приближение x_{n+1} :

$$x_{n+1} = P_k(0) = - \sum_{i=0}^k \frac{\omega_{k+1}(0)}{y_{n-i} \omega'_{k+1}(y_{n-i})} x_{n-i}. \quad (5.70)$$

В частности, при $k = 1$ это приводит к методу секущих:

$$x_{n+1} = - \left(\frac{y_{n-1}}{y_n - y_{n-1}} x_n + \frac{y_n}{y_{n-1} - y_n} x_{n-1} \right) = \frac{x_{n-1}f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})}.$$

Замечание 5.6. Интерполяционный подход к решению уравнений может быть обобщен и в сторону привлечения интерполяирования с кратными узлами (в частности, один двукратный узел приводит к алгоритму метода Ньютона).

5.3.11. Многомерная алгебраическая интерполяция

[Интерполяционный многочлен первой степени](#)

[Обобщение интерполяционных формул Ньютона на случай функций многих переменных](#)

[Последовательная интерполяция на прямоугольной сетке](#)

[Общий случай интерполирования на треугольнике](#)

Интерполяирование функций многих переменных значительно сложнее, чем функций одной переменной. Это вызвано не только тем, что рассуждения становятся более громоздкими в силу наличия большого числа переменных, но и рядом принципиальных трудностей.

Ограничимся сейчас случаем двух переменных (в том числе и потому, что двумерный случай наиболее широко распространен в практике), хотя принципиальных трудностей при обобщении на большее число измерений не существует.

Итак, пусть на плоскости (x, y) заданы $n+1$ точек $(x_0, y_0), \dots, (x_n, y_n)$. Будем разыскивать многочлен $P(x, y)$ относительно x и y возможно низшей степени, который бы в этих точках принимал, соответственно, значения z_0, z_1, \dots, z_n .

Если многочлен записать в виде

$$P(x, y) = a_{00} + a_{10}x + a_{01}y + a_{20}x^2 + a_{11}xy + a_{02}y^2 + \dots + a_{m0}x^m + \dots + a_{0m}y^m = \sum_{i+j=0}^m a_{ij}x^i y^j, \quad (5.71)$$

то, подставляя данные координаты точек и приравнивая левую часть соответствующему значению z_i , получим систему из $(n+1)$ линейных уравнений относительно $\frac{(m+1)(m+2)}{2}$ неизвестных a_{ij} :

$$\sum_{i+j=0}^m a_{ij}x_k^i y_k^j = z_k, \quad k = \overline{0, n}. \quad (5.72)$$

В общем случае уравнения (5.72) независимы. Поэтому, если не накладывать на $P(x, y)$ никаких дополнительных условий, то должно выполняться соотношение, связывающее число узлов и степень многочлена (матрица системы (5.72) должна быть квадратной)

$$n+1 = \frac{(m+1)(m+2)}{2}. \quad (5.73)$$



Это – первое принципиальное затруднение, состоящее в том, что мы уже не можем решать поставленную задачу при произвольном количестве узлов интерполяирования (возможные значения для n (в скобках указаны степень многочлена): 0 (0), 2 (1), 5 (2), 9 (3) и т.д.). Конечно, если число узлов не соответствует условию (5.73), то можно часть коэффициентов многочлена $P(x, y)$ задавать принудительно (в частности, нулями). Однако для выбора этих коэффициентов редко можно дать разумное обоснование.

Далее, рассмотрим определитель системы (5.72) (при выполнении условия (5.73)). При $m = 1$ ($n = 2$) этот определитель принимает вид

$$\begin{vmatrix} 1 & x_0 & y_0 \\ 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \end{vmatrix}.$$

Очевидно, он будет обращаться в нуль, если три узловые точки лежат на одной прямой. Аналогично, при $m > 1$ соответствующий определитель будет обращаться в нуль, если все узлы лежат на одной кривой m -го порядка.

Это – второе принципиальное затруднение: узлы интерполяирования не могут быть расположены произвольно. При этом проверка того, что определители не обращаются в нуль, чрезвычайно затруднительна.

Третье принципиальное затруднение возникает при оценке остаточных членов. Теорема Ролля, которой мы пользовались ранее, для ныне рассматриваемого нами случая действовать не будет.

Рассмотрим некоторые частные случаи двумерной интерполяции.

Интерполяционный многочлен первой степени

Пусть даны три узла: (x_0, y_0) , (x_1, y_1) , (x_2, y_2) . Таким образом, требование на количество входных данных выполнено. Будем, как и в случае интерполяирования функций одной переменной, многочлен $P_1(x, y)$ искать в виде линейной комбинации значений в узлах, коэффициентами которой являются [функции влияния узлов](#):

$$P_1(x, y) = z_0 \cdot P_{10}(x, y) + z_1 \cdot P_{11}(x, y) + z_2 \cdot P_{12}(x, y), \quad (5.74)$$

где $P_{1i}(x, y)$ – многочлены первой степени, равные единице в точке $Q_i(x_i, y_i)$ и обращающиеся в нуль в остальных двух точках. Введем в рассмотрение векторы

$$\begin{cases} r_k = (x - x_k) \vec{i} + (y - y_k) \vec{j}, \\ r_{kl} = (x_k - x_l) \vec{i} + (y_k - y_l) \vec{j}, \\ r_{kl}^* = (y_k - y_l) \vec{i} - (x_k - x_l) \vec{j}. \end{cases} \quad (5.75)$$

Тогда скалярное произведение (r_1, r_{12}^*) будет многочленом первой степени. Этот многочлен обращается в нуль в точке $Q_1(x_1, y_1)$, так как при этом первый множитель обращается в нуль. Он обращается в нуль также и в точке $Q_2(x_2, y_2)$, так как векторы r_{21} и r_{12}^* ортогональны. В точке $Q_0(x_0, y_0)$ рассматриваемое скалярное произведение будет равно нулю в том и только в том случае, когда точки лежат на одной прямой.

Таким образом, за многочлен $P_{10}(x, y)$ можно принять выражение $P_{10}(x, y) = \frac{(r_1, r_{12}^*)}{(r_{01}, r_{12}^*)}$. Аналогично получим:

$P_{11}(x, y) = \frac{(r_2, r_{20}^*)}{(r_{12}, r_{20}^*)}$, $P_{12}(x, y) = \frac{(r_0, r_{01}^*)}{(r_{20}, r_{01}^*)}$. Следовательно, искомый интерполяционный многочлен, учитывая (5.74), может быть записан в виде

$$P_1(x, y) = z_0 \frac{(r_1, r_{12}^*)}{(r_{01}, r_{12}^*)} + z_1 \frac{(r_2, r_{20}^*)}{(r_{12}, r_{20}^*)} + z_2 \frac{(r_0, r_{01}^*)}{(r_{20}, r_{01}^*)}. \quad (5.76)$$

Обобщение интерполяционных формул Ньютона на случай функций многих переменных



Легко видеть, что увеличение числа произвольным образом расположенных узлов ведет к усложнению вида интерполяционного многочлена. Поэтому рассмотрим $\frac{(n+1)(n+2)}{2}$ узлов, расположенных специальным образом:

$$(x_0, y_0), (x_1, y_0), \dots, (x_{n-1}, y_0), (x_n, y_0)$$

$$(x_0, y_1), (x_1, y_1), \dots, (x_{n-1}, y_1),$$

.....

$$(x_0, y_{n-1}), (x_1, y_{n-1}),$$

$x_i \neq x_j, y_i \neq y_j$ при $i \neq j$

$$(x_0, y_n)$$

Значения x_i и y_j могут быть произвольными, так что взаимное расположение узлов может быть достаточно общим. Проверим, что нет кривой n -го порядка, проходящей через все эти узлы. В самом деле, если бы такая кривая имелась, то она содержала бы точки, расположенные в первой строке таблицы. Таких точек $(n+1)$ и все они лежат на одной прямой. Следовательно, вся прямая также принадлежала бы кривой порядка n . В этом случае вся кривая порядка n распадается на прямую и кривую порядка $(n-1)$, через остальные $\frac{n(n+1)}{2}$ точек. Для нее можно было бы провести аналогичные рассуждения. Продолжая этот процесс, мы в конце концов пришли бы к заключению, что три точки $(x_0, y_{n-1}), (x_1, y_{n-1}), (x_0, y_n)$ лежат на одной прямой. Этого заведомо нет; следовательно, выбранные нами узлы не могут лежать на одной кривой порядка n .

Построим теперь интерполяционный многочлен по нашим узлам. Обозначим его через $P_n(x, y)$, а $P_n(x_i, y_i)$ через z_{ij} . Если рассмотреть только те из выбранных нами узлов, для которых $i+j < n$, то на тех же основаниях мы можем построить интерполяционный многочлен $P_{n-1}(x, y)$ степени $(n-1)$, принимающий в точках (x_i, y_j) , $i + j < n$, значения z_{ij} . Образуем разность $P_n(x, y) - P_{n-1}(x, y)$. Она будет являться

многочленом степени не выше n , обращающимся в нуль в точках (x_i, y_i) , $i + j < n$. Будем разыскивать ее в виде

$$P_n(x, y) - P_{n-1}(x, y) = A_{n0}(x - x_0) \cdots (x - x_{n-1}) + A_{n-1,1}(x - x_0) \cdots (x - x_{n-2})(y - y_0) + \cdots + \\ + A_{0n}(y - y_0) \cdots (y - y_{n-1}). \quad (5.77)$$

При этом постоянные $A_{n-i,i}$ определяются однозначно, поскольку

$$P_n(x_i, y_{n-i}) - P_{n-1}(x_i, y_{n-i}) = A_{i,n-i}(x_i - x_0) \cdots (x_i - x_{i-1})(y_{n-i} - y_0) \cdots (y_{n-i} - y_{n-i-1}).$$

В силу единственности представления интерполяционного многочлена по выбранным нами узлам это будет единственным значением разности. Тогда из (5.77) получим:

$$P_n(x, y) = P_{n-1}(x, y) + \sum_{i=0}^n A_{n-i,i}(x - x_0) \cdots (x - x_{n-i-1})(y - y_0) \cdots (y - y_{i-1}).$$

Поступая точно так же с $P_{n-1}(x, y)$, затем с $P_{n-2}(x, y)$ и т.д., получим:

$$P_n(x, y) = A_{00} + A_{10}(x - x_0) + A_{01}(y - y_0) + A_{20}(x - x_0)(x - x_1) + A_{11}(x - x_0)(y - y_0) + \\ + A_{02}(y - y_0)(y - y_1) + A_{n0}(x - x_0) \cdots (x - x_{n-1}) + \cdots + A_{0n}(y - y_0) \cdots (y - y_{n-1}). \quad (5.78)$$

Выразим теперь коэффициенты A_{ij} через значения функции $z_{kl} = f(x_k, y_l)$. Полагая в (5.78) $y = y_0$, будем иметь:

$$P_n(x, y_0) = A_{00} + A_{10}(x - x_0) + A_{20}(x - x_0)(x - x_1) + \cdots + A_{n0}(x - x_0) \cdots (x - x_{n-1}).$$

Это – интерполяционный многочлен относительно x , принимающий в точках (x_i, y_0) значения $f(x_i, y_0)$. Следовательно, в силу единственности представления интерполяционного многочлена для функций одной независимой переменной $A_{i0} = f(x_0, x_1, \dots, x_i; y_0)$, $i = \overline{0, n}$.

При $y = y_1$ наш интерполяционный многочлен примет вид

$$\begin{aligned}
 P_n(x, y_1) &= [A_{00} + A_{01}(y_1 - y_0)] + [A_{10} + A_{11}(y_1 - y_0)](x - x_0) + \\
 &\quad [A_{20} + A_{21}(y_1 - y_0)](x - x_0)(x - x_1) + \dots + \\
 &\quad + [A_{n-1,0} + A_{n-1,1}(y_1 - y_0)](x - x_0) \dots (x - x_{n-2}) + A_{n0}(x - x_0) \dots (x - x_{n-1})
 \end{aligned}$$

Этот интерполяционный многочлен относительно x должен в точках (x_i, y_1) , $i = \overline{0, n-1}$, принимать значения $f(x_i, y_1)$. Последнее слагаемое при этих значениях x обращается в нуль. Следовательно, все члены правой части, кроме последнего, дают интерполяционный многочлен Ньютона степени $(n-1)$, принимающий в точках (x_i, y_1) , $i = \overline{0, n-1}$ значения $f(x_i, y_1)$. Таким образом,

$$A_{k0} + A_{k1}(y_1 - y_0) = f(x_0, \dots, x_k; y_1), \quad k = \overline{0, n-1}.$$

Отсюда

$$A_{k1} = \frac{f(x_0, \dots, x_k; y_1) - f(x_0, \dots, x_k; y_0)}{y_1 - y_0} = f(x_0, \dots, x_k; y_0, y_1), \quad k = \overline{0, n-1}.$$

Вообще, если мы уже знаем, что $A_{ki} = f(x_0, \dots, x_k; y_0, \dots, y_i)$ для всех $i < m$, то, рассматривая $P_n(x, y_m)$, получим:

$$\begin{aligned}
 P_n(x, y_m) &= [A_{00} + A_{01}(y_m - y_0) + \dots + A_{0m}(y_m - y_0) \dots (y_m - y_{m-1})] + \\
 &\quad + [A_{10} + A_{11}(y_m - y_0) + \dots + A_{1m}(y_m - y_0) \dots (y_m - y_{m-1})](x - x_0) + \dots + \\
 &\quad + [A_{n-m,0} + A_{n-m,1}(y_m - y_0) + \dots + A_{n-m,m}(y_m - y_0) \dots (y_m - y_{m-1})](x - x_0) \dots (x - x_{n-m-1}) + \\
 &\quad + [A_{n-m+1,0} + A_{n-m+1,1}(y_m - y_0) + \dots + A_{n-m+1,m-1}(y_m - y_0) \dots (y_m - y_{m-2})](x - x_0) \dots (x - x_{n-m}) + \\
 &\quad + \dots + [A_{n-1,0} + A_{n-1,1}(y_m - y_0)](x - x_0) \dots (x - x_{n-2}) + A_{n0}(x - x_0) \dots (x - x_{n-1}).
 \end{aligned}$$



Этот интерполяционный многочлен относительно x должен в точках (x_i, y_m) , $i = \overline{0, n-m}$, принимать значения $f(x_i, y_m)$. Последние m слагаемых при этих значениях x обращаются в нуль. Следовательно, все члены правой части, кроме них, дают интерполяционный многочлен Ньютона степени $(n-m)$, принимающий в точках (x_i, y_m) , $i = \overline{0, n-m}$ значения $f(x_i, y_m)$. Таким образом,

$$A_{k0} + A_{k1}(y_m - y_0) + \cdots + A_{km}(y_m - y_0) \cdots (y_m - y_{m-1}) = f(x_0, \dots, x_k; y_m),$$

$$k = \overline{0, n-m}.$$

Рассматривая это выражение как функцию переменной y_m , получаем: левая часть равенства есть многочлен степени m относительно переменной y_m , принимающий в узлах y_0, \dots, y_{m-1} те же значения, что и функция $f(x_0, \dots, x_k; y_m)$. Следовательно,

$$A_{km} = f(x_0, \dots, x_k; y_0, \dots, y_m), \quad k = \overline{0, n-m}.$$

Таким образом, интерполяционную формулу мы можем записать в виде

$$P_n(x, y) = \sum_{k=0}^n \sum_{i+j=k} (x - x_0) \cdots (x - x_{i-1})(y - y_0) \cdots (y - y_{j-1}) f(x_0, \dots, x_i; y_0, \dots, y_j). \quad (5.79)$$

Формула (5.79) представляет собой обобщение [интерполяционной формулы Ньютона](#) для неравноотстоящих промежутков на случай двух независимых переменных.

Последовательная интерполяция на прямоугольной сетке

Рассмотрим еще один вариант расположения узлов двумерной интерполяции, в котором построение интерполяционного многочлена особенно просто.

Итак, пусть задана прямоугольная таблица узлов:

$$(x_0, y_0) \quad \cdots \quad (x_n, y_0)$$

$$\vdots \qquad \cdots \qquad \vdots$$

$$(x_0, y_m) \quad \cdots \quad (x_n, y_m)$$

Предполагаются также известными значения функции $f(x, y)$ в этих узлах.

Возможен следующий способ приближенного определения значения функции f в некоторой точке (x, y) , не совпадающей с узлами интерполяции. Сначала интерполируем нашу функцию как функцию одной переменной x при фиксированных значениях y_i ($i = 0, 1, \dots, m$). При этом мы каждый раз используем одну строку заданной таблицы узлов. Таким образом, мы можем найти приближенное значение $f(x, y_i)$. Затем по найденным значениям $f(x, y_i)$ путем интерполяции по переменной y находим значение $f(x, y)$. Посмотрим, как будет выглядеть интерполяционная формула при таком способе интерполяции. Применим интерполяционную формулу Ньютона для неравных промежутков, можем записать:

$$f(x, y) \approx f(x, y_0) + (y - y_0) f(x; y_0, y_1) + (y - y_0)(y - y_1) f(x; y_0, y_1, y_2) + \dots +$$

$$+ (y - y_0) \dots (y - y_{m-1}) f(x; y_0, y_1, \dots, y_m).$$

Снова применим [интерполяционную формулу Ньютона](#) для неравных промежутков для приближения каждого из сомножителей вида $f(x; y_0, y_1, \dots, y_k)$; рассматривая эту разделившую разность как функцию переменной x , имеем:

$$f(x; y_0, y_1, \dots, y_k) \approx f(x_0; y_0, y_1, \dots, y_k) + (x - x_0) f(x_0, x_1; y_0, y_1, \dots, y_k) + \dots +$$

$$+ (x - x_0) \dots (x - x_{n-1}) f(x_0, \dots, x_n; y_0, y_1, \dots, y_k), \quad k = 0, 1, \dots, m.$$

Подставляя эти выражения в предыдущую формулу, получим:

$$f(x, y) \approx P_{n,m}(x, y) = \sum_{i=0}^n \sum_{j=0}^m (x - x_0) \dots (x - x_{i-1})(y - y_0) \dots (y - y_{j-1}) f(x_0, \dots, x_i; y_0, \dots, y_j). \quad (5.80)$$



Замечание 5.7. В формулах (5.79), (5.80) при $i = 0$ и $j = 0$ множители будут иметь вид $x - x_{-1}$ и $y - y_{-1}$. Будем считать их равными единице.

Очевидно, столь же просто может быть получена результирующая формула и в случае, если на каждом этапе использовать [многочлен Лагранжа](#). Она будет иметь вид

$$P_{n,m}(x, y) = \sum_{i=0}^n \sum_{j=0}^m \frac{\omega_{n+1}(x) \omega_{m+1}(y)}{(x - x_i)(y - y_j) \omega'_{n+1}(x_i) \omega'_{m+1}(y_j)} f(x_i, y_j). \quad (5.81)$$

Напоследок заметим, что по рассмотренной таблице узлов интерполяционный многочлен, вообще говоря, определяется неоднозначно в силу невыполнения условия типа (5.73). Однозначность формулам (5.80), (5.81) сама процедура интерполяции.

Общий случай интерполяирования на треугольнике

Очень часто в приложениях возникает задача интерполяирования функции в области треугольной формы (частный случай ее мы рассмотрели выше).

В этом случае сетку естественно задавать следующим образом: каждую сторону треугольника разобьем на n равных частей и через точки деления проведем прямые, параллельные сторонам треугольника (сами стороны треугольника также включаем в это множество прямых). Тогда множество узлов – это точки пересечения соответствующих прямых. Очевидно, число узлов в этом случае будет равно

$$m = \frac{(n+1)(n+2)}{2}$$

и мы можем построить единственный интерполяционный многочлен степени n , принимающий в указанных узлах заданные значения.

Возьмем некоторую фиксированную точку $Q(x_i, y_j)$. Тогда среди прямых, с помощью которых мы построили сетку, имеется ровно n прямых, удовлетворяющих условию: существует не более одной вершины треугольника такой, что $Q(x_i, y_j) = Q_k$ и эта вершина лежат по одну сторону от такой прямой. На рисунке (точка Q_k выделена кружком): для вершины B – это прямые, параллельные стороне AC , с номерами $0, 1, \dots, j-1$ (всего – j штук); для вершины C – это прямые, параллельные стороне AB , с номерами $0, 1, \dots, i-1$ (всего – i штук); для вершины A – это прямые, параллельные стороне BC , с номерами

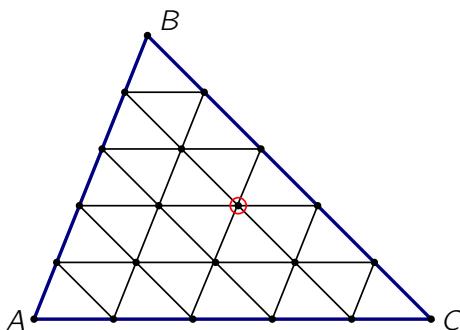


Рисунок 5.2

$j + i + 1, \dots, n$ (всего $n - (j + i)$ штук). При этом каждая точка из множества узлов, отличная от узла Q_k , лежит на одной из таких прямых.

Пусть $L_{k,i}(x) = 0$, $k = 1, m$, $i = 1, n$ – уравнения этих прямых. Тогда функция

$$\Phi_k(x, y) = \prod_{i=1}^n \frac{L_{k,i}(x, y)}{L_{k,i}(Q_k)} \quad (5.82)$$

является многочленом степени n и принимает значение, равное единице в точке Q_k , и нуль во всех остальных узлах.

Таким образом, она является функцией влияния и поэтому многочлен степени n

$$P(x, y) = \sum_{k=1}^m \Phi_k(x, y) f(Q_k) \quad (5.83)$$

будет искомым.

В качестве примера рассмотрим сейчас более подробно случай $n = 1$.



Тогда для точки A означенной выше прямой будет прямая, на которой лежит сторона BC . Ее уравнение имеет вид

$$L_A(x, y) = \begin{vmatrix} x - x_B & y - y_B \\ x_C - x_B & y_C - y_B \end{vmatrix}.$$

Поэтому

$$\Phi_A(x, y) = \begin{vmatrix} x - x_B & y - y_B \\ x_C - x_B & y_C - y_B \\ x_A - x_B & y_A - y_B \\ x_C - x_B & y_C - y_B \end{vmatrix}.$$

Аналогично для точки B это будет прямая, на которой лежит сторона AC , с уравнением

$$L_B(x, y) = \begin{vmatrix} x - x_C & y - y_C \\ x_A - x_C & y_A - y_C \end{vmatrix}$$

и

$$\Phi_B(x, y) = \begin{vmatrix} x - x_C & y - y_C \\ x_A - x_C & y_A - y_C \\ x_B - x_C & y_B - y_C \\ x_A - x_C & y_A - y_C \end{vmatrix},$$



а для точки C – прямая, на которой лежит сторона AB , с уравнением

$$L_C(x, y) = \begin{vmatrix} x - x_B & y - y_B \\ x_A - x_B & y_A - y_B \end{vmatrix}$$

и

$$\Phi_C(x, y) = \frac{\begin{vmatrix} x - x_B & y - y_B \\ x_A - x_B & y_A - y_B \\ x_C - x_B & y_C - y_B \\ x_A - x_B & y_A - y_B \end{vmatrix}}{\begin{vmatrix} x_A - x_B & y_A - y_B \\ x_C - x_B & y_C - y_B \\ x_A - x_B & y_A - y_B \end{vmatrix}}.$$

Следовательно, сам интерполяционный многочлен первой степени примет вид

$$P_1(x, y) = \Phi_A(x, y) f(A) + \Phi_B(x, y) f(B) + \Phi_C(x, y) f(C). \quad (5.84)$$

Замечание 5.8. Если сравнить (5.84) и (5.76), то, очевидно, с точностью до обозначений они совпадают. Это и естественно, поскольку три точки, не лежащие на одной прямой, всегда определяют единственную плоскость.



5.4. Приближение сплайнами

- 5.4.1. Общее определение сплайна. Простейшие примеры
- 5.4.2. Интерполяционный кубический сплайн
- 5.4.3. Кусочно-кубическая интерполяция со сглаживанием
- 5.4.4. Интерполяционный бикубический сплайн
- 5.4.5. Приближение кривых и поверхностей

5.4.1. Общее определение сплайна. Простейшие примеры

Интерполяционный сплайн первой степени

Ранее мы отмечали неудовлетворительное состояние дел со сходимостью интерполяционного процесса, и, в частности, приводили пример Бернштейна, показывающий, что последовательность интерполяционных многочленов по равноточным узлам на отрезке $[-1, 1]$ не сходится для непрерывной функции $f(x) = |x|$. Еще более любопытен пример, восходящий к Рунге и состоящий в том, что указанный интерполяционный процесс не сходится на отрезке $[-1, 1]$ даже для сколь угодно раз дифференцируемой функции $f(x) = \frac{1}{1+25x^2}$.

Поэтому на практике для того, чтобы достаточно хорошо приблизить функцию, вместо построения интерполяционного многочлена высокой степени сейчас значительно чаще используют кусочно-функциональные приближения (как правило, в случае, когда речь идет о многочленных приближениях, то эти многочлены оказываются не очень высокой степени).

Перейдем к описанию предмета исследования.

Разобьем отрезок $[a, b]$, на котором ищется приближение к функции $f(x)$, на N частей точками x_j : $a = x_0 < x_1 < \dots < x_{N-1} < x_N = b$.

По определению положим: $x_j - x_{j-1} = h_j > 0$, $j = 1, \dots, N$. Соответствующее разбиение далее будем обозначать Δ .

Определение. Разобьем отрезок $[a, b]$, на котором ищется приближение к функции $f(x)$, на N частей точками x_j : $a = x_0 < x_1 < \dots < x_{N-1} < x_N = b$.

По определению положим: $x_j - x_{j-1} = h_j > 0$, $j = 1, \dots, N$. Соответствующее разбиение далее будем обозначать Δ .

Назовем *полиномиальным сплайном порядка m дефекта k* на разбиении Δ (обозначение $S_{\Delta}^m(x)$) функцию, которая:

1. На каждом из отрезков $[x_{i-1}; x_i]$, $i = \overline{1, N}$ является алгебраическим многочленом степени m , т.е.

$$S_{\Delta}^m(x) = P_{im}(x) = a_{i0} + a_{i1}x + \dots + a_{im}x^m, \quad x \in [x_{i-1}; x_i], \quad i = \overline{1, N}; \quad (5.85)$$

2. Является функцией класса $C^{m-k} [a, b]$, т.е. во всех внутренних узлах разбиения $\Delta S_{\Delta}^m(x)$ удовлетворяет условию непрерывности производных до порядка $m - k$ включительно:

$$(S_{\Delta}^m(x_i + 0))^{(j)} = (S_{\Delta}^m(x_i - 0))^{(j)}, \quad i = \overline{1, N-1}; \quad j = \overline{0, m-k}. \quad (5.86)$$

В случае $k = 1$ вместо слов «полиномиальный сплайн дефекта 1» говорят «полиномиальный сплайн» или просто «сплайн». В дальнейшем мы будем иметь дело именно с такими объектами.

Таким образом, для того чтобы задать сплайн, необходимо указать значения коэффициентов a_{ij} (всего их $- N \times (m+1)$ штук). При этом требования (5.86) предстаивают для этих целей $(N-1) \times m$ условий. Следовательно, если использовать для построения сплайна только его определение, то в итоге останется $N+m$ свободных параметров.

Однако данное выше определение пока никаким образом не связывает сплайн и функцию f , которую мы собираемся приближать.

Поэтому естественно оставшиеся $N+m$ параметров использовать для выполнения соответствующей привязки (ее наличие далее будем подчеркивать, добавляя в число аргументов сплайна, помимо независимой переменной, еще и функцию $f: S_{\Delta}^m(f; x)$). Способы такой привязки могут быть различными. И один из них – интерполяционный. В этом случае в дополнение к условиям (5.85), (5.86) требуют также совпадения значений сплайна и приближаемой функции $f(x)$ в узлах разбиения Δ , т.е.

$$S_{\Delta}^m(f; x_i) = f(x_i), \quad i = \overline{0, N}. \quad (5.87)$$

Такие сплайны мы будем называть **интерполяционными**. При этом, однако, необходимо иметь в виду, что все равно $m-1$ параметр остается свободным. Как распорядиться остающейся свободой, мы рассмотрим далее в конкретных случаях.

Интерполяционный сплайн первой степени

В качестве примера проведем построение **интерполяционного сплайна первой степени**.

Тогда, очевидно, условие (5.85) примет вид

$$S_{\Delta}^1(f; x) = a_{i0} + a_{i1}x, \quad x \in [x_{i-1}; x_i], \quad i = \overline{1, N},$$



Вверх

Назад

Вперёд

Пред.

След.

Указатель

Помощь

Экран

а условия (5.86) и (5.87) можно записать следующим образом

$$\begin{cases} P_{i1}(x_i) = f(x_i), & i = \overline{1, N}, \\ P_{i1}(x_{i-1}) = f(x_{i-1}), \end{cases} \quad (5.88)$$

Обратим внимание на то, что здесь реализованы и условия непрерывности, и условия интерполяционные. Так как $m = 1$, то свободных параметров в этом случае не будет. Распишем (5.88) подробнее:

$$\begin{cases} P_{i1}(x_i) = a_{i0} + a_{i1}x_i = f(x_i), \\ P_{i1}(x_{i-1}) = a_{i0} + a_{i1}x_{i-1} = f(x_{i-1}). \end{cases}$$

Отсюда находим:

$$a_{i1} = \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} = f(x_{i-1}, x_i); \quad a_{i0} = f(x_{i-1}) - a_{i1}x_{i-1}, \quad i = \overline{1, N},$$

т.е.

$$S_{\Delta}^1(f; x) = f(x_{i-1}) + (x - x_{i-1})f(x_{i-1}, x_i), \quad x \in [x_{i-1}; x_i]. \quad (5.89)$$

Таким образом, на каждом из отрезков разбиения Δ сплайн является хорошо нам известным **интерполяционным многочленом** с узлами интерполяции x_{i-1} и x_i .



5.4.2. Интерполяционный кубический сплайн

Сходимость процесса интерполирования кубическими сплайнами

Экстремальное свойство интерполяционных кубических сплайнов

В течение многих лет чертежники использовали длинные гибкие рейки из дерева или какого-либо другого материала в качестве лекала, проводя с их помощью плавные кривые через заданные точки. Эти рейки (англ. *spline*) закрепляют на месте, привешивая к ним в некоторых точках свинцовые грузила. Изменяя положения сплайна и грузил, при достаточном количестве грузил можно добиться того, чтобы сплайн проходил через заданные точки.

Если рассматривать рейку как тонкую балку, то для нее справедлив закон Бернулли-Эйлера:

$$M(x) = EI \cdot \left(\frac{1}{R(x)} \right),$$

где M – изгибающий момент, E – модуль Юнга, I – геометрический момент инерции и R – радиус кривизны кривой, совпадающей с деформированной осью балки. При незначительных изгиbach справедливо соотношение $R(x) \approx \frac{1}{y''(x)}$. Отсюда имеем уравнение

$$y''(x) = \frac{1}{EI} M(x).$$

При этом изгибающий момент $M(x)$ изменяется линейно между точками закрепления грузил. Проинтегрировав записанное уравнение, получим пример [интерполяционного кубического сплайна](#).

Этот исторический экскурс поясняет происхождение термина «сплайн», а также некоторые моменты, которые можно использовать при построении одного из математических алгоритмов нахождения интерполяционного кубического сплайна. Опишем этот алгоритм подробнее.

Пусть

$$M_j = \frac{d^2}{dx^2} S_{\Delta}^3(f; x_j), \quad j = \overline{0, N}.$$

В силу линейности второй производной кубического сплайна на каждом из отрезков $[x_{j-1}; x_j]$ можно записать для нее следующее представление:

$$\frac{d^2}{dx^2} S_{\Delta}^3(f; x) = P''_{i3}(x) = M_{j-1} \frac{x_j - x}{h_j} + M_j \frac{x - x_{j-1}}{h_j}, \quad x \in [x_{j-1}; x_j], \quad j = \overline{1, N}. \quad (5.90)$$

Этим соотношением мы учитываем два факта: непрерывность второй производной во всех внутренних узлах разбиения Δ , а также принятие второй производной в узлах разбиения значений M_j .

Проинтегрируем соотношение (5.90) дважды (при этом константы интегрирования будем брать в специальном виде, обеспечивающем простоту их нахождения; в частности, при первом интегрировании вводим сразу обе произвольные постоянные):

$$\frac{d}{dx} S_{\Delta}^3(f; x) = -M_{j-1} \frac{(x_j - x)^2}{2h_j} + M_j \frac{(x - x_{j-1})^2}{2h_j} + \frac{A_j - B_j}{h_j}, \quad x \in [x_{j-1}; x_j], \quad j = \overline{1, N}; \quad (5.91)$$

$$S_{\Delta}^3(f; x) = M_{j-1} \frac{(x_j - x)^3}{6h_j} + M_j \frac{(x - x_{j-1})^3}{6h_j} + A_j \frac{x - x_{j-1}}{h_j} + B_j \frac{x_j - x}{h_j}, \quad x \in [x_{j-1}; x_j], \quad j = \overline{1, N}.$$

Постоянные A_j и B_j найдем, используя условия интерполяции (5.87) и непрерывности сплайна в узлах, которые вновь можно объединить в условия вида (5.88):

$$\begin{cases} S_{\Delta}^3(f; x_j) = f(x_j), \\ S_{\Delta}^3(f; x_{j-1}) = f(x_{j-1}), \end{cases} \quad j = \overline{1, N}.$$

Подставляя во второе из соотношений (5.91) соответствующие значения аргумента, найдем:

$$\begin{cases} f(x_j) = M_j \frac{h_j^2}{6} + A_j, \\ f(x_{j-1}) = M_{j-1} \frac{h_j^2}{6} + B_j, \end{cases} \quad j = \overline{1, N}.$$

Отсюда

$$\begin{cases} A_j = f(x_j) - M_j \frac{h_j^2}{6}, \\ B_j = f(x_{j-1}) - M_{j-1} \frac{h_j^2}{6}, \end{cases} \quad j = \overline{1, N}.$$

Подставляя найденные значения A_j , B_j в (5.91), можем записать:

$$\frac{d}{dx} S_\Delta^3(f; x) = -M_{j-1} \frac{(x_j - x)^2}{2h_j} + M_j \frac{(x - x_{j-1})^2}{2h_j} + \frac{f_j - f_{j-1}}{h_j} - \frac{M_j - M_{j-1}}{6} h_j; \quad (5.92)$$

$$S_\Delta^3(f; x) = M_{j-1} \frac{(x_j - x)^3}{6h_j} + M_j \frac{(x - x_{j-1})^3}{6h_j} + \left(f_j - M_j \frac{h_j^2}{6} \right) \frac{x - x_{j-1}}{h_j} + \left(f_{j-1} - M_{j-1} \frac{h_j^2}{6} \right) \frac{x_j - x}{h_j}.$$

Эти соотношения имеют место для $x \in [x_{j-1}; x_j]$, $j = \overline{1, N}$.

Используем теперь оставшееся из условий, определяющих сплайн – непрерывность первой производной в узлах разбиения Δ . Так как

$$\frac{d}{dx} S_\Delta^3(f; x_j + 0) = P'_{j+1,3}(x_j) = -M_j \frac{h_{j+1}}{2} + \frac{f_{j+1} - f_j}{h_{j+1}} - \frac{M_{j+1} - M_j}{6} h_{j+1} =$$

$$= -M_j \frac{h_{j+1}}{3} - M_{j+1} \frac{h_{j+1}}{6} + \frac{f_{j+1} - f_j}{h_{j+1}},$$

а

$$\frac{d}{dx} S_\Delta^3(f; x_j - 0) = P'_{j,3}(x_j) = M_j \frac{h_j}{2} + \frac{f_j - f_{j-1}}{h_j} - \frac{M_j - M_{j-1}}{6} h_j =$$

$$= M_j \frac{h_j}{3} + M_{j-1} \frac{h_j}{6} + \frac{f_j - f_{j-1}}{h_j},$$

то из условия их равенства получаем:

$$\frac{h_j}{6} M_{j-1} + \frac{h_j + h_{j+1}}{3} M_j + \frac{h_{j+1}}{6} M_{j+1} = \frac{f_{j+1} - f_j}{h_{j+1}} - \frac{f_j - f_{j-1}}{h_j}, \quad j = \overline{1, N-1}. \quad (5.93)$$



Как и следует из общей теории, в соотношениях (5.93) не достает двух уравнений, связывающих M_j между собой, для взаимно однозначного соответствия. Чаще всего различают следующие способы задания дополнительных (граничных) условий:

- 1) $M_0 = M_N = 0$ (так называемые естественные граничные условия);
- 2) $M_0 = f''(x_0) = f''(a)$; $M_N = f''(x_N) = f''(b)$ (если известны соответствующие значения $f''(a)$ и $f''(b)$);
- 3) если известны значения $f'(a)$ и $f'(b)$, то, используя первое из равенств (5.92), получим:

$$\frac{d}{dx} S_{\Delta}^3(f; a) = -M_0 \frac{h_1}{2} + \frac{f_1 - f_0}{h_1} - \frac{M_1 - M_0}{6} h_1 = f'(a)$$

или

$$2M_0 + M_1 = \frac{6}{h_1} \left(\frac{f_1 - f_0}{h_1} - f'(a) \right) = 6f(x_0, x_0, x_1).$$

Аналогично для точки $x = b$ условие $\frac{d}{dx} S_{\Delta}^3(f; b) = f'(b)$ приводит к уравнению

$$M_{N-1} + 2M_N = \frac{6}{h_N} \left(f'(b) - \frac{f_N - f_{N-1}}{h_N} \right) = 6f(x_{N-1}, x_N, x_N).$$

При любом из рассмотренных типов граничных условий (заметим, что существуют и другие) задача (5.93) + условия представляет собой систему линейных алгебраических уравнений с трехдиагональной матрицей, имеющую, в силу строгого диагонального доминирования последней, единственное решение.

Таким образом, интерполяционный кубический сплайн всегда может быть построен, и при том единственным образом (с точностью до задания дополнительных условий). Решение указанной выше системы (5.93) с дополнительными условиями может быть найдено, например, с помощью экономичного алгоритма разностной прогонки, который в данном случае будет устойчив. После нахождения величин M_j сплайн и его производные определяются формулами (5.90), (5.92).

Сходимость процесса интерполирования кубическими сплайнами

Как уже указывалось выше, одним из мотивов перехода к использованию кусочных приближений явилась плохая сходимость процесса интерполирования с помощью обычных алгебраических многочленов. Изучим сейчас вопрос о сходимости процесса интерполяционного приближения с помощью рассматриваемых нами ныне объектов. Конкретные исследования проведем на примере изучения сходимости процесса интерполирования кубическими сплайнами, алгоритм построения которых изложен выше.

Вначале установим вспомогательный результат, касающийся оценки нормы обратной матрицы. Везде далее будем рассматривать **кубическую норму матрицы**. Напомним, что она является подчиненной по отношению к привычной нам **максимум-норме вектора**.

Справедлива

Лемма 5.4. Пусть матрица A имеет диагональное доминирование. Тогда имеет место неравенство

$$\|A^{-1}\| \leq \left(\min_{1 \leq i \leq n} \left\{ |a_{ii}| - \sum_{j \neq i} |a_{ij}| \right\} \right)^{-1}.$$

Доказательство. Пусть $x \neq 0$ – произвольный вектор. Для него выберем номер координаты k такой, что $\|x\| = \max_{1 \leq i \leq n} |x_i| = |x_k|$ (т.е. зафиксируем компоненту, на которой достигается максимум-норма). Рассмотрим теперь вектор $y = Ax$. Тогда для него справедлива цепочка неравенств:

$$\begin{aligned} \|y\| &= \|Ax\| = \max_{1 \leq i \leq n} \left| \sum_{j=1}^n a_{ij} x_j \right| \geq \left| \sum_{j=1}^n a_{kj} x_j \right| \geq |a_{kk} x_k| - \left| \sum_{j \neq k} a_{kj} x_j \right| \geq \\ &\geq |a_{kk}| |x_k| - \sum_{j \neq k} |a_{kj}| |x_k| = |x_k| \left(|a_{kk}| - \sum_{j \neq k} |a_{kj}| \right) = \|x\| \left(|a_{kk}| - \sum_{j \neq k} |a_{kj}| \right) \geq \\ &\geq \|x\| \min_{1 \leq i \leq n} \left(|a_{ii}| - \sum_{j \neq i} |a_{ij}| \right). \end{aligned}$$

Так как в силу диагонального доминирования матрицы A матрица A^{-1} существует и, следовательно, $x = A^{-1}y$, то для нормы матрицы A^{-1} получаем:

$$\begin{aligned}\|A^{-1}\| &= \sup_{y \neq 0} \frac{\|A^{-1}y\|}{\|y\|} = \sup_{y \neq 0} \frac{\|x\|}{\|y\|} \leqslant \sup_{y \neq 0} \frac{\|x\|}{\|x\| \min_{1 \leq i \leq n} (|a_{ii}| - \sum_{j \neq i} |a_{ij}|)} = \\ &= \left(\min_{1 \leq i \leq n} \left(|a_{ii}| - \sum_{j \neq i} |a_{ij}| \right) \right)^{-1}.\end{aligned}$$

□

Перейдем теперь к исследованию процесса интерполяции с помощью кубических сплайнов. Будем рассматривать наиболее общий из трех типов дополнительных условий – условия типа 3.

Вначале несколько преобразуем полученную нами ранее систему уравнений. Для этого разделим каждое из уравнений (5.93) на $\frac{h_j+h_{j+1}}{6}$ и введем обозначения

$$\mu_j = \frac{h_j}{h_j + h_{j+1}}, \quad \lambda_j = \frac{h_{j+1}}{h_j + h_{j+1}} = 1 - \mu_j.$$

С учетом этих обозначений система для определения «моментов» M_j при граничных условиях типа в) примет вид

$$\begin{cases} 2M_0 + M_1 = 6f(x_0, x_0, x_1), \\ \mu_j M_{j-1} + 2M_j + \lambda_j M_{j+1} = 6f(x_{j-1}, x_j, x_{j+1}), \quad j = \overline{1, N-1}, \\ M_{N-1} + 2M_N = 6f(x_{N-1}, x_N, x_N). \end{cases}$$

или в матричной форме

$$AM = d, \tag{5.94}$$

где

$$A = \begin{pmatrix} 2 & 1 & & & \\ \mu_1 & 2 & \lambda_1 & & \\ & \mu_2 & 2 & \lambda_2 & \\ & \dots & \dots & & \\ & \mu_{N-1} & 2 & \lambda_{N-1} & \\ & 1 & 1 & & \end{pmatrix}; \quad d = \begin{pmatrix} d_0 \\ d_1 \\ d_2 \\ \dots \\ d_{N-1} \\ d_N \end{pmatrix}, \quad \begin{cases} d_0 = 6f(x_0, x_0, x_1), \\ d_j = 6f(x_{j-1}, x_j, x_{j+1}), \quad j = \overline{1, N-1}, \\ d_N = 6f(x_{N-1}, x_N, x_N). \end{cases} \quad (5.95)$$

Пусть также $h = \max_{1 \leq j \leq N} \{h_j\}$, а $\omega(h; f)$ – модуль непрерывности функции $f(x)$ на отрезке $[a, b]$, т.е.

$$\omega(h; f) = \sup_{\substack{|x_1 - x_2| \leq h \\ x_1, x_2 \in [a, b]}} |f(x_1) - f(x_2)|.$$

Нас будет интересовать степень близости функции $f(x)$ (и ее производных) к сплайну (и его производным) на отрезке $[a, b]$. Естественно при этом, что эта величина будет зависеть от свойств гладкости функции $f(x)$. Сформулируем и докажем один из результатов.

Теорема 5.13. Пусть $f(x) \in C^2[a, b]$ и интерполирующий ее сплайн $S_\Delta^3(f; x)$ определяется системой (5.94), (5.95) и формулами (5.90), (5.92). Тогда имеют место соотношения

$$f^{(p)}(x) - \frac{d^p}{dx^p} S_\Delta^3(f; x) = O(h^{2-p}) \omega(h; f''), \quad p = 0, 1, 2. \quad (5.96)$$

[[Доказательство](#)]

Из установленных оценок, в частности, следует равномерная сходимость при $h \rightarrow 0$ сплайна и его первой и второй производных к $f(x)$ и ее первой и второй производной.

Замечание 5.9. Если $f''(x)$ удовлетворяет условию Липшица с константой K , то, очевидно,

$$\omega(h; f) = \sup_{\substack{|x_1 - x_2| \leq h \\ x_1, x_2 \in [a, b]}} |f(x_1) - f(x_2)| \leq \sup_{|x_1 - x_2| \leq h} K|x_1 - x_2| \leq Kh.$$

Следовательно, оценки (5.96) примут вид

$$\left| f^{(p)}(x) - \frac{d^p}{dx^p} S_\Delta^3(f; x) \right| = O(h^{3-p}), \quad p = 0, 1, 2.$$

Замечание 5.10. Если $f(x) \in C^4[a, b]$, то путем несколько более сложных рассуждений можно установить оценки

$$\left| f^{(p)}(x) - \frac{d^p}{dx^p} S_\Delta^3(f; x) \right| = O(h^{4-p}), \quad p = 0, 1, 2, 3.$$

Экстремальное свойство интерполяционных кубических сплайнов

Кубические сплайн-функции обладают очень важным свойством, которое, по сути, и обеспечивает высокую эффективность сплайн-интерполяции. А именно: рассмотрим на отрезке $[a, b]$ класс $W_2^2[a, b]$, состоящий из функций, имеющих интегрируемые с квадратом вторые производные. Поставим задачу отыскания интерполяционной функции $u \in W_2^2[a, b]$, $u(x_0) = f_k$, $k = \overline{0, N}$, которая минимизирует функционал

$$\Phi(u) = \int_a^b [u''(x)]^2 dx \tag{5.97}$$

на классе $W_2^2[a, b]$. Утверждается, что минимум такого функционала достигается на кусочно-кубической сплайн-функции $S_\Delta^3(f; x)$ с краевыми условиями типов 1 либо 3.

В самом деле, рассмотрим величину

$$\Phi(u - S_\Delta^3) = \int_a^b \left(u'' - \frac{d^2}{dx^2} S_\Delta^3 \right)^2 dx.$$

Имеем:

$$\Phi(u - S_{\Delta}^3) = \int_a^b \left\{ \left[(u'')^2 - \left(\frac{d^2}{dx^2} S_{\Delta}^3 \right)^2 \right] - 2 \left[u'' - \frac{d^2}{dx^2} S_{\Delta}^3 \right] \frac{d^2}{dx^2} S_{\Delta}^3 \right\} dx = \\ (5.98)$$

$$= \Phi(u) - \Phi(S_{\Delta}^3) - 2 \int_a^b \left(u'' - \frac{d^2}{dx^2} S_{\Delta}^3 \right) \frac{d^2}{dx^2} S_{\Delta}^3 dx.$$

Преобразуем последнее слагаемое правой части (5.98), интегрируя по частям:

$$\int_a^b \left(u'' - \frac{d^2}{dx^2} S_{\Delta}^3 \right) \frac{d^2}{dx^2} S_{\Delta}^3 dx = \left(u' - \frac{d}{dx} S_{\Delta}^3 \right) \frac{d^2}{dx^2} S_{\Delta}^3 \Big|_a^b - \int_a^b \left(\left(u' - \frac{d}{dx} S_{\Delta}^3 \right) \frac{d^3}{dx^3} S_{\Delta}^3 \right) dx.$$

Так как третья производная кубического сплайна есть кусочно-постоянная функция (т.е. $\frac{d^3}{dx^3} S_{\Delta}^3(f; x) = C_j$ при $x \in [x_{j-1}; x_j]$, $j = \overline{1, N}$), а двойная подстановка обращается в нуль на граничных условиях типа 1 или 3, то далее последнее соотношение далее преобразуется следующим образом:

$$\int_a^b \left(u'' - \frac{d^2}{dx^2} S_{\Delta}^3 \right) \frac{d^2}{dx^2} S_{\Delta}^3 dx = \left(u' - \frac{d}{dx} S_{\Delta}^3 \right) \frac{d^2}{dx^2} S_{\Delta}^3 \Big|_a^b - \int_a^b \left(\left(u' - \frac{d}{dx} S_{\Delta}^3 \right) \frac{d^3}{dx^3} S_{\Delta}^3 \right) dx = \\ = - \sum_{j=1}^N \int_a^{x_j} \left(u' - \frac{d}{dx} S_{\Delta}^3 \right) C_j dx = - \sum_{j=1}^N C_j [u(x) - S_{\Delta}^3(f; x)] \Big|_{x_{j-1}}^{x_j} = 0.$$

Поэтому из (5.98) имеем:

$$\Phi(u - S_{\Delta}^3) = \Phi(u) - \Phi(S_{\Delta}^3).$$



Следовательно,

$$\Phi(S_{\Delta}^3) = \Phi(u) - \Phi(u - S_{\Delta}^3) \leq \Phi(u)$$

для всех функций u , удовлетворяющих поставленным условиям: $u \in W_2^2[a, b]$, $u(x_k) = f_k$, $k = \overline{0, N}$.

Таким образом, на кусочно-кубической функции $S_{\Delta}^3(f; x)$ реализуется минимум функционала (5.97).

Заметим, что минимизирующая функция единственна, так как если $g(x)$ – другая минимизирующая функция, то $\Phi(g - S_{\Delta}^3) = 0$ и, следовательно, $g''(x) = \frac{d^2}{dx^2} S_{\Delta}^3(x)$ почти всюду на отрезке $[a, b]$. Отсюда следует, что $g(x)$ и $S_{\Delta}^3(f; x)$ отличаются только лишь на линейную функцию, т.е. $g(x) = S_{\Delta}^3(f; x) + \alpha x + \beta$. Поскольку же $g(x_k) = S_{\Delta}^3(f; x_k) + \alpha x_k + \beta$, $k = \overline{0, N}$, то $\alpha = \beta = 0$.

Иногда установленное экстремальное свойство берется в качестве определения интерполяционного кубического сплайна как функции, минимизирующей функционал (5.97). Однако при этом сужается множество рассматриваемых сплайнов, так как при этом учитываются не все типы граничных условий.

5.4.3. Кусочно-кубическая интерполяция со сглаживанием

На практике часто приходится иметь дело со случаем, когда значения f_k в узлах разбиения Δ заданы с некоторой погрешностью (например, они могут быть результатами некоторых измерений и тогда означенная погрешность есть погрешность измерительного инструмента). Если погрешность исходных данных относительно велика, то это крайне неблагоприятно влияет на поведение интерполяционного сплайна и особенно его производных. В частности, график сплайна обычно в этом случае приобретает ярко выраженные осцилляции. Поэтому возникает вопрос, нельзя ли построить сплайн, проходящий вблизи заданных значений, но более «гладкий», чем интерполяционный. Такие сплайны будем называть *сглаживающими*.

Итак, рассмотрим вопрос о построении таких функций. Потребуем, чтобы искомая функция $g(x)$ минимизировала на классе $W_2^2[a, b]$ функционал

$$\Phi_1(u) = \int_a^b (u''(x))^2 dx + \sum_{j=0}^N p_j [u(x_j) - f_j]^2, \quad (5.99)$$

где p_j – некоторые положительные числа.

В функционале $\Phi_1(u)$ скомбинированы интерполяционные условия прохождения кривой вблизи заданных значений и условия минимальности «изгиба» функции. Чем больше весовые коэффициенты p_j , тем больший вклад в $\Phi_1(u)$ вносят интерполяционные условия, тем ближе к заданным значениям проходит сглаживающая функция.

Покажем, что решением вариационной задачи (5.99) является кубический сплайн, т.е. функция, удовлетворяющая определению с $m = 3$ и краевым условием $\frac{d^2}{dx^2} S_\Delta^3(f; a) = \frac{d^2}{dx^2} S_\Delta^3(f; b) = 0$.

Действительно, пусть $u_0 \in W_2^2[a, b]$ – решение задачи. Построим сплайн $S_\Delta^3(f; x)$ такой, что $S_\Delta^3(f; x_j) = u_0(x_j)$, $j = \overline{0, N}$. Тогда второе слагаемое в (5.99) будет одинаковым для $S_\Delta^3(f; x)$ и для $u_0(x)$. Поэтому

$$\int_a^b (u_0'')^2 dx = \int_a^b \left(\frac{d^2}{dx^2} S_\Delta^3 \right)^2 dx.$$

Но как показано в разделе 5.4.2, $S_\Delta^3(f; x)$ – единственная функция, дающая при интерполировании $u_0(x)$ минимум функционалу $\Phi(u)$. Поэтому $u_0(x) \equiv S_\Delta^3(f; x)$.

Итак, минимум функционала (5.99) достаточно искать в классе кубических сплайнов. А поскольку кубический сплайн однозначно определяется множеством его значений μ_j , $j = \overline{0, N}$ в узлах разбиения Δ , то минимизация функционала $\Phi_1(u)$ сводится к нахождению минимума некоторой функции от переменных μ_0, \dots, μ_N .

Так как в силу (5.90)

$$\frac{d^2}{dx^2} S_{\Delta}^3(f; x) = P''_{i3}(x) = M_{j-1} \frac{x_j - x}{h_j} + M_j \frac{x - x_{j-1}}{h_j}, \quad x \in [x_{j-1}; x_j], \quad j = \overline{1, N},$$

то

$$\Phi_1(S_{\Delta}^3) = \sum_{j=1}^N \int_{x_{j-1}}^{x_j} \left(M_{j-1} \frac{x_j - x}{h_j} + M_j \frac{x - x_{j-1}}{h_j} \right)^2 dx + \sum_{j=0}^N p_j (\mu_j - \tilde{f}_j)^2. \quad (5.100)$$

Производя в (5.100) интегрирование, получим:

$$\begin{aligned} & \sum_{j=1}^N \int_{x_{j-1}}^{x_j} \left(M_{j-1} \frac{x_j - x}{h_j} + M_j \frac{x - x_{j-1}}{h_j} \right)^2 dx = \\ & = \sum_{j=1}^N \int_{x_{j-1}}^{x_j} \left[M_{j-1}^2 \frac{(x_j - x)^2}{h_j^2} + 2M_{j-1}M_j \frac{(x_j - x)(x - x_{j-1})}{h_j^2} + M_j^2 \frac{(x - x_{j-1})^2}{h_j^2} \right] dx = \\ & = \sum_{j=1}^N \left(M_{j-1}^2 \frac{h_j}{3} + M_{j-1}M_j \frac{h_j}{3} + M_j^2 \frac{h_j}{3} \right). (*) \end{aligned}$$

Так как $M_0 = M_N = 0$, то, выполняя замены индексов, имеем:

$$\sum_{j=1}^N M_{j-1}^2 \frac{h_j}{3} = \sum_{j=1}^{N-1} M_j^2 \frac{h_{j+1}}{3},$$

$$\sum_{j=1}^N \left(M_{j-1}^2 \frac{h_j}{3} + M_j^2 \frac{h_j}{3} \right) = \sum_{j=1}^{N-1} M_j^2 \frac{h_j + h_{j+1}}{3},$$

$$\sum_{j=1}^N M_{j-1} M_j \frac{h_j}{3} = \sum_{j=1}^{N-1} M_{j-1} M_j \frac{h_j}{6} + \sum_{j=1}^{N-1} M_j M_{j+1} \frac{h_j}{6}.$$

Подставляя эти выражения в (*), получим:

$$\begin{aligned} \sum_{j=1}^N \int_{x_{j-1}}^{x_j} \left(M_{j-1} \frac{x_j - x}{h_j} + M_j \frac{x - x_{j-1}}{h_j} \right)^2 dx &= \sum_{j=1}^{N-1} \left(M_{j-1}^2 \frac{h_j}{3} + M_{j-1} M_j \frac{h_j}{3} + M_j^2 \frac{h_j}{3} \right) = \\ &= \sum_{j=1}^{N-1} \left(\frac{h_j}{6} M_{j-1} M_j + \frac{h_j + h_{j+1}}{3} M_j^2 + \frac{h_{j+1}}{6} M_j M_{j+1} \right) = \\ &= \sum_{j=1}^{N-1} M_j \left(\frac{h_j}{6} M_{j-1} + \frac{h_j + h_{j+1}}{3} M_j + \frac{h_{j+1}}{6} M_{j+1} \right). \end{aligned}$$

Заметим, что систему (5.93) с естественными граничными условиями типа 1 можно переписать в виде

$$\begin{cases} \frac{h_1+h_2}{3} M_1 + \frac{h_2}{6} M_2 = \frac{1}{h_1} f_0 - \left(\frac{1}{h_1} + \frac{1}{h_2} \right) f_1 + \frac{1}{h_2} f_2, \\ \frac{h_j}{6} M_{j-1} + \frac{h_j+h_{j+1}}{3} M_j + \frac{h_{j+1}}{6} M_{j+1} = \frac{1}{h_j} f_{j-1} - \left(\frac{1}{h_j} + \frac{1}{h_{j+1}} \right) f_j + \frac{1}{h_{j+1}} f_{j+1}, \quad j = \overline{2, N-2}; \\ \frac{h_{N-1}}{6} M_{N-1} + \frac{h_{N-1}+h_N}{3} M_N = \frac{1}{h_{N-1}} f_{N-2} - \left(\frac{1}{h_{N-1}} + \frac{1}{h_N} \right) f_{N-1} + \frac{1}{h_N} f_N \end{cases}$$

или в матричной форме

$$BM = Hf, \quad (5.101)$$

$$\text{где } B = \begin{pmatrix} \frac{h_1+h_2}{3} & \frac{h_2}{6} & & & \\ \frac{h_2}{6} & \frac{h_2+h_3}{3} & \frac{h_3}{6} & & \\ & \dots & & & \\ & & \frac{h_{N-2}}{6} & \frac{h_{N-2}+h_{N-1}}{3} & \frac{h_{N-1}}{6} \\ & & & \frac{h_{N-1}}{6} & \frac{h_{N-1}+h_N}{3} \end{pmatrix} - (N-1) \times (N-1)\text{-матрица; } f = \begin{pmatrix} f_0 \\ f_1 \\ \dots \\ f_N \end{pmatrix} -$$

$(N+1) \times 1$ - матрица (вектор-столбец);

$$H = \begin{pmatrix} \frac{1}{h_1} & -\left(\frac{1}{h_1} + \frac{1}{h_2}\right) & \frac{1}{h_2} & & \\ & \frac{1}{h_2} & -\left(\frac{1}{h_2} + \frac{1}{h_3}\right) & \frac{1}{h_3} & \\ & & \dots & & \\ & & & \frac{1}{h_{N-1}} & -\left(\frac{1}{h_{N-1}} + \frac{1}{h_N}\right) \frac{1}{h_N} \end{pmatrix} - (N-1) \times (N+1)\text{-матрица.}$$

Учитывая (5.101), результаты интегрирования запишем в виде

$$\sum_{j=1}^N \int_{x_{j-1}}^{x_j} \left(M_{j-1} \frac{x_j - x}{h_j} + M_j \frac{x - x_{j-1}}{h_j} \right)^2 dx = (BM, M).$$

Поэтому соотношение (5.100) примет вид

$$\Phi_1(S_\Delta^3) = (BM, M) + \sum_{j=0}^N p_j (\mu_j - f_j)^2. \quad (5.102)$$

Поскольку наш сплайн в узлах разбиения Δ принимает значения μ_j , система (5.101) для него примет вид

$$BM = H\mu \quad (5.103)$$



(что означает, что M линейно выражается через μ). Поэтому $\Phi_1(S_\Delta^3)$ – положительная форма от μ . Ее экстремумом может быть только минимум, необходимым условием которого является

$$\frac{\partial \Phi_1}{\partial \mu_i} = \frac{\partial}{\partial \mu_i} (BM, M) + 2p_i (\mu_i - f_i) = 0, \quad i = \overline{0, N}.$$

Но матрица B не зависит от μ . Поэтому в силу (5.103) имеем:

$$\begin{aligned} \frac{\partial}{\partial \mu_i} (BM, M) &= 2 \left(\frac{\partial (BM)}{\partial \mu_i}, M \right) = 2 \left(\frac{\partial (H\mu)}{\partial \mu_i}, M \right) = 2 \left(H \frac{\partial \mu}{\partial \mu_i}, M \right) = \\ &= 2 \left(\frac{\partial \mu}{\partial \mu_i}, H^T M \right) = 2 (H^T M)_i. \end{aligned}$$

Таким образом, в векторной форме условие минимума имеет вид

$$H^T M + P\mu = Pf, \quad (5.104)$$

где $P = \text{diag}\{p_0, \dots, p_N\}$.

Умножив равенство (5.104) слева на HP^{-1} , получим:

$$HP^{-1}H^T M + H\mu = Hf$$

или (с учетом (5.103))

$$(B + HP^{-1}H^T) M = Hf. \quad (5.105)$$

Матрица системы (5.105) пятидиагональна, симметрична и положительно определена (как сумма положительной и неотрицательной). Ее можно решать каким-либо известным методом (например, [методом Гаусса](#) или пятидиагональной прогонки, хотя, вероятно, наиболее разумным выбором будет [метод квадратного корня](#)). После того как вектор M определен, необходимо найти вектор μ сеточных значений сплайна по формуле, которая легко следует из (5.104):

$$\mu = f - P^{-1}H^T M, \quad (5.106)$$



а затем по формулам (5.90), (5.92) восстановить сплайн (и его производные).

По сравнению с интерполяционным сплайном построение сглаживающего сплайна требует значительно большего объема работы. Поэтому при решении вопроса о том, каким сплайном пользоваться, нужно учитывать в первую очередь погрешность исходных данных, а также те требования, которые предъявляются к сплайну (так, например, практически бесполезно применение сглаживающих сплайнов, когда исходная информация задана с точностью порядка ε_M).



5.4.4. Интерполяционный бикубический сплайн

Возможны различные обобщения сплайн-функций на случай многих независимых переменных. При этом важное значение имеет форма области, в которой осуществляется приближение, а также то, каким образом производится ее разбиение на подобласти.

Рассмотрим сейчас наиболее простой случай, когда (для двух независимых переменных) областью задания является прямоугольник, а разбиение также осуществляется на прямоугольники.

Итак, пусть в прямоугольной области $\Omega = [a, b] \times [c; d]$ введена сетка линий $\Delta = \Delta_x \times \Delta_y$, где

$$\Delta_x = \{x = x_i, j = \overline{1, N}, a = x_0 < x_1 < \dots < x_N = b\};$$

$$\Delta_y = \{y = y_j, j = \overline{1, M}, c = y_0 < y_1 < \dots < y_M = d\},$$

делящая область на прямоугольные ячейки $\Omega_{ij} = \{(x, y) : x \in [x_{i-1}; x_i], y \in [y_{j-1}; y_j]\}$, $i = \overline{1, N}$; $j = \overline{1, M}$.

Тогда по аналогии со [случаем одной независимой переменной](#) можно дать следующее

Определение. Назовем функцию $S_{\Delta}^{n,m}(x, y)$ [полиномиальным сплайном степени \$n\$ по переменной \$x\$ и степени \$m\$ по переменной \$y\$](#) с линиями склейки на сетке Δ , если:

- На каждой ячейке Ω_{ij} $S_{\Delta}^{n,m}(x, y)$ является многочленом степени n по переменной x и степени m по переменной y , т.е.

$$S_{\Delta}^{n,m}(x, y) = \sum_{k=0}^n \sum_{l=0}^m a_{kl}^{ij} (x - x_i)^k (y - y_j)^l, \quad i = \overline{1, N}; \quad j = \overline{1, M}; \quad (5.107)$$

- 2.

$$S_{\Delta}^{n,m}(x, y) \in C^{n-1, m-1}(\Omega). \quad (5.108)$$



Если к отмеченным двум условиям добавить требование, чтобы в узлах разбиения Δ значения сплайна $S_{\Delta}^{n,m}(x, y)$ совпадали со значениями приближаемой функции (в этом случае к аргументам сплайна будем добавлять еще и f и писать $S_{\Delta}^{n,m}(f; x, y)$), т.е.

$$S_{\Delta}^{n,m}(f; x_i, y_j) = f(x_i, y_j) \stackrel{\text{def}}{=} f_{ij}, \quad i = \overline{0, N}; \quad j = \overline{0, M}, \quad (5.109)$$

то мы получим интерполяционный сплайн степени n по переменной x и степени m по переменной y .

Рассмотрим сейчас более подробно алгоритм построения интерполяционного бикубического сплайна ($n = m = 3$).

Очевидно, что, как и в случае одной независимой переменной условий (5.108), (5.109) недостаточно для однозначного определения сплайна $S_{\Delta}^{3,3}(f; x, y)$. Поэтому в качестве дополнительных условий будем брать условия типа 1 (точнее, их двумерный аналог):

$$\left. \frac{\partial^2}{\partial \nu^2} S_{\Delta}^{3,3} \right|_{\Gamma} = 0, \quad (5.110)$$

где ν – внешняя нормаль к границе Γ области Ω .

Принципиально построение ничем не отличается от одномерного случая. Вспомним, что для вычисления одномерной сплайн-функции в любой точке по формулам типа (5.92) необходимо знать значения самой функции и ее производных второго порядка в узловых точках, а для того чтобы найти эти вторые производные, нужно один раз решить линейную алгебраическую систему с трехдиагональной матрицей (типа (5.93) с дополнительными условиями). Какие же предварительные вычисления нужно проделать, чтобы потом по явным формулам вычислять функцию в любой точке в двумерном случае?

Чтобы ответить на этот вопрос, прибегнем к уже известной нам технологии повторного интерполяирования. Запишем, считая u параметром, интерполяционный кубический сплайн по переменной x (верхние

индексы в обозначении $S_{\Delta}^{3,3}(f; x, y)$ сплайна далее будем опускать), пользуясь для этих целей формулами типа (5.92):

$$S_{\Delta}(f; x, y) = \frac{\partial^2 S_{\Delta}(f; x_{i-1}, y)}{\partial x^2} \frac{(x_i - x)^3}{6h_i} + \frac{\partial^2 S_{\Delta}(f; x_i, y)}{\partial x^2} \frac{(x - x_{i-1})^3}{6h_i} + \\ + \left(S_{\Delta}(f; x_i, y) - \frac{\partial^2 S_{\Delta}(f; x_i, y)}{\partial x^2} \frac{h_i^2}{6} \right) \frac{x - x_{i-1}}{h_i} + \left(S_{\Delta}(f; x_{i-1}, y) - \frac{\partial^2 S_{\Delta}(f; x_{i-1}, y)}{\partial x^2} \frac{h_i^2}{6} \right) \frac{x_i - x}{h_i}, \quad (5.111)$$

$$x \in [x_{i-1}; x_i], \quad i = \overline{1, N}.$$

Чтобы пользоваться этой формулой, мы должны, очевидно, уметь вычислять значения $S_{\Delta}(f; x_i, y)$ и для всех значений $i = \overline{0, N}$.

Поскольку эти выражения являются функциями только одной переменной, то для их нахождения можно воспользоваться интерполяционными кубическими сплайнами по переменной y , а именно:

$$S_{\Delta}(f; x_i, y) = \frac{\partial^2 S_{\Delta}(f; x_i, y_{j-1})}{\partial y^2} \frac{(y_j - y)^3}{6\tau_j} + \frac{\partial^2 S_{\Delta}(f; x_i, y_j)}{\partial y^2} \frac{(y - y_{j-1})^3}{6\tau_j} + \\ + \left(S_{\Delta}(f; x_i, y_j) - \frac{\partial^2 S_{\Delta}(f; x_i, y_j)}{\partial y^2} \frac{\tau_j^2}{6} \right) \frac{y - y_{j-1}}{\tau_j} + \left(S_{\Delta}(f; x_i, y_{j-1}) - \frac{\partial^2 S_{\Delta}(f; x_i, y_{j-1})}{\partial y^2} \frac{\tau_j^2}{6} \right) \frac{y_j - y}{\tau_j}, \quad (5.112)$$

$$y \in [y_{j-1}; y_j], \quad j = \overline{1, M}.$$

Учитывая, что здесь $S_\Delta(f; x_i, y_j) = f(x_i, y_j) = f_{ij}$, для определения величин $L_{ij} := \frac{\partial^2 S_\Delta(f; x_i, y_j)}{\partial y^2}$ имеем набор систем типа (5.94) (всего их будет $(N+1)$):

$$\begin{cases} \mu_j^* L_{i,j-1} + 2L_{i,j} + \lambda_j^* L_{i,j+1} = 6f(x_i; y_{j-1}, y_j, y_{j+1}), & j = \overline{1, M-1}, \\ L_{i,0} = L_{i,M} = 0, & i = \overline{0, N}. \\ \mu_j^* = \frac{\tau_j}{\tau_j + \tau_{j+1}}; \quad \lambda_j^* = \frac{\tau_{j+1}}{\tau_j + \tau_{j+1}}, & \end{cases} \quad (5.113)$$

Сплайн, аналогичный (5.112), построим для функции $\frac{\partial^2 S_\Delta(f; x_i, y)}{\partial x^2}$:

$$\frac{\partial^2 S_\Delta(f; x_i, y)}{\partial x^2} = \frac{\partial^4 S_\Delta(f; x_i, y_{j-1})}{\partial x^2 \partial y^2} \frac{(y_j - y)^3}{6\tau_j} + \frac{\partial^4 S_\Delta(f; x_i, y_j)}{\partial x^2 \partial y^2} \frac{(y - y_{j-1})^3}{6\tau_j} + \quad (4.8)$$

$$+ \left(\frac{\partial^2 S_\Delta(f; x_i, y_j)}{\partial x^2} - \frac{\partial^2 S_\Delta(f; x_i, y_{j-1})}{\partial y^2} \frac{\tau_j^2}{6} \right) \frac{y - y_{j-1}}{\tau_j} + \left(\frac{\partial^2 S_\Delta(f; x_i, y_{j-1})}{\partial x^2} - \frac{\partial^2 S_\Delta(f; x_i, y_{j-1})}{\partial y^2} \frac{\tau_j^2}{6} \right) \frac{y_j - y}{\tau_j}, \quad (5.114)$$

$$y \in [y_{j-1}; y_j], \quad j = \overline{1, M}.$$

При этом для нахождения величин $K_{i,j} := \frac{\partial^4 S_\Delta(f; x_i, y_j)}{\partial x^2 \partial y^2}$ необходимо решить системы

$$\begin{cases} \mu_j^* K_{i,j-1} + 2K_{i,j} + \lambda_j^* K_{i,j+1} = 6 \frac{\partial^2 S_\Delta(f; x_i; y_{j-1}, y_j, y_{j+1})}{\partial x^2}, & j = \overline{1, M-1}, \\ K_{i,0} = K_{i,M} = 0, & i = \overline{0, N} \end{cases} \quad (5.115)$$

общим числом $(N+1)$, правыми частями которых являются **разделенные разности** второго порядка от второй производной по переменной x сплайна в узлах разбиения Δ . Чтобы их найти, достаточно решить системы (здесь $M_{i,j} := \frac{\partial^2 S_\Delta(f; x_i, y_j)}{\partial x^2}$)

$$\left\{ \begin{array}{l} \mu_i M_{i-1,j} + 2M_{i,j} + \lambda_i M_{i+1,j} = 6f(x_{i-1}, x_i, x_{i+1}; y_j), \quad i = \overline{1, N-1}, \\ M_{0,j} = M_{N,j} = 0, \\ \mu_i = \frac{h_i}{h_i+h_{i+1}}; \quad \lambda_i = \frac{h_{i+1}}{h_i+h_{i+1}}, \end{array} \right. \quad j = \overline{0, M}. \quad (5.116)$$

общим числом ($M + 1$), построив тем самым интерполяционный кубический сплайн $S_\Delta(f; x, y_j)$.

Произведя упорядочивание работ, получим следующий алгоритм построения интерполяционного бикубического сплайна на прямоугольной сетке:

1. Решаем $(M + 1)$ линейную систему (5.116), из которых находим величины $M_{i,j} = \frac{\partial^2 S_\Delta(f; x_i, y_j)}{\partial x^2}$, $i = \overline{0, N}$; $j = \overline{0, M}$;
2. Решаем $(N + 1)$ линейную систему (5.115), из которых находим величины $K_{i,j} = \frac{\partial^4 S_\Delta(f; x_i, y_j)}{\partial x^2 \partial y^2}$, $i = \overline{0, N}$; $j = \overline{0, M}$; В результате выполнения этих двух этапов построен сплайн (5.114).
3. Решаем $(N + 1)$ линейную систему (5.113), из которых находим величины $L_{i,j} = \frac{\partial^2 S_\Delta(f; x_i, y_j)}{\partial y^2}$, $i = \overline{0, N}$; $j = \overline{0, M}$; После этого построен сплайн (5.112).
4. Значение интерполяционного бикубического сплайна в точке $(x, y) \in [x_{i-1}; x_i] \times [y_{j-1}; y_j]$ вычисляем по формуле (5.111) с использованием формул (5.112) и (5.114).

Таким образом, прежде чем приступить к расчету функции $S_\Delta(f; x, y)$ в интересующих нас точках необходимо решить один раз $(M + 1) + (N + 1) + (N + 1) = 2N + M + 3$ линейных систем, а для расчета $S_\Delta^{3,3}(f; x, y)$ в одной точке области нужно пять раз выполнить расчеты по формулам, определяющим сплайн: дважды – по формулам (5.112) (при $x = x_i$ и $x = x_{i-1}$), дважды – по формулам (5.114) (при $x = x_i$ и $x = x_{i-1}$) и один раз – по формулам (5.111).

Замечание 5.11. Подставив (5.112) и (5.114) в (5.111), можно получить явное полиномиальное выражение для $S_\Delta^{3,3}(f; x, y)$ в каждой ячейке разбиения, но для хранения коэффициентов многочлена потребуется в четыре раза больше памяти (хотя при описанном выше способе организации работы мы примерно в четыре раза больше вычисляем).



Замечание 5.12. При построении бикубического сплайна можно поменять порядок приближения по независимым переменным местами. Тогда нужно будет решить $N + 2M + 3$ систем, но итоговый результат не изменится.

Описанный алгоритм может быть обобщен на многомерные области типа параллелепипеда.

Распространение алгоритма на другие типы краевых областей производится очевидным образом. Кроме того, для интерполяционных бикубических сплайнов оказывается справедливым экстремальное свойство, аналогичное рассмотренному нами в одномерном случае, а также имеет место оценка погрешности (в случае, если $f(x, y) \in C^{4,4}(\Omega)$)

$$\frac{\partial^{\alpha+\beta} f(x, y)}{\partial x^\alpha \partial y^\beta} - \frac{\partial^{\alpha+\beta} S_{\Delta}^{3,3}(f; x, y)}{\partial x^\alpha \partial y^\beta} = O(h^{4-\alpha} + \tau^{4-\beta}), \quad \alpha, \beta \in \{0, 1, 2, 3\}.$$



5.4.5. Приближение кривых и поверхностей

Интерполяция кривых сплайнами

С помощью рассмотренных ранее [сплайнов одной переменной](#) можно приближать лишь такие плоские кривые, которые в выбранной системе координат (не обязательно прямоугольной декартовой) описываются функциональной зависимостью вида $y = f(x)$. Однако не все кривые могут быть представлены подобным образом. Более универсальным способом является параметрическое задание их координат в виде двух функций $x = x(u)$ и $y = y(u)$ некоторого параметра u .

При интерполяции кривой, заданной параметрически, естественно ввести разбиение на промежутки изменения параметра u : $u_0 < u_1 < \dots < u_N$, затем вычислить соответствующие значения координат точек на кривой $(x_i = x(u_i), y = y(u_i))$, и построить для функций $x(u)$ и $y(u)$ интерполяционные сплайны $S(x; u)$, $S(y; u)$. Совокупность этих двух сплайнов называется [интерполяционным параметрическим сплайном](#). В зависимости от вида функций $S(x; u)$, $S(y; u)$ будем говорить о параметрических линейных, кубических и т.п. сплайнах. В качестве меры погрешности приближения проще всего взять величину

$$R(u) = \sqrt{(S(x; u) - x(u))^2 + (S(y; u) - y(u))^2}.$$

Главной особенностью практических задач о приближении кривых является то, что заданы бывают только упорядоченные массивы точек на них, а информация о способе параметризации, которая необходима для построения сплайнов, отсутствует.

Аналогичные проблемы, только технически еще более сложные, возникают и при приближении поверхностей.

Интерполяция кривых сплайнами

$(x_0, y_0)(x_1, y_1)(x_{i+1}, y_{i+1})(x_N, y_N)(x_0, y_0)(x_1, y_1)(x_{i+1}, y_{i+1})(x_N, y_N)$ Пусть на некоторой кривой L задана последовательность точек $P_i(x_i, y_i)$, $i = \overline{0, N}$. Как уже отмечалось ранее, основной проблемой является отсутствие параметризации. В то же время в анализе известен способ введения параметра, называемый естественной параметризацией (в качестве параметра принимается длина дуги кривой). Поэтому в данном случае мы рас-

смотрим аналог естественной параметризации: введем ее по суммарной длине хорд d_i , соединяющих точки P_{i-1} и P_i ($d_i = |P_{i-1}P_i|$).

Если обозначить новый параметр через \tilde{s} , то сетка узлов интерполяции будет такой:

$$\tilde{\Delta} : 0 = \tilde{s}_0 < \tilde{s}_1 < \cdots < \tilde{s}_N,$$

где

$$\tilde{s}_i = \sum_{k=0}^{i-1} d_k, \quad d_k = \sqrt{(x_{k+1} - x_k)^2 + (y_{k+1} - y_k)^2}.$$

При этом параметр \tilde{s} пробегает отрезок $[0; \tilde{s}_N]$.

Интерполяционный параметрический сплайн первой степени. Согласно [полученным ранее формулам](#) на промежутке между точками P_{i-1} и P_i рассматриваемый сплайн задается соотношениями

$$\left\{ \begin{array}{l} S_1(x; \tilde{s}) = x_i \frac{\tilde{s} - \tilde{s}_{i-1}}{d_{i-1}} + x_{i-1} \frac{\tilde{s}_i - \tilde{s}}{d_{i-1}}, \\ \qquad \qquad \qquad i = \overline{1, N}. \\ S_1(y; \tilde{s}) = y_i \frac{\tilde{s} - \tilde{s}_{i-1}}{d_{i-1}} + y_{i-1} \frac{\tilde{s}_i - \tilde{s}}{d_{i-1}}, \end{array} \right. \quad (5.117)$$

Из (5.117) следует равенство

$$\frac{S'_1(y; \tilde{s})}{S'_1(x; \tilde{s})} = \frac{y_i - y_{i-1}}{x_i - x_{i-1}}, \quad x_i \neq x_{i-1}, \quad (5.118)$$

которое используется для приближенного вычисления наклона касательной к кривой L между точками P_{i-1} и P_i . Если $x_i = x_{i+1}$, то это означает, что данное звено сплайна параллельно оси Oy .

Геометрически параметрический сплайн первой степени представляет собой ломаную, состоящую из отрезков прямых, соединяющих точки P_i .

Интерполяционный параметрический сплайн третьей степени. Вновь в качестве параметра берем суммарную длину хорд \tilde{s} . В этом случае, в соответствии с формулами, полученными ранее, можем записать:

$$S(x; \tilde{s}) = \tilde{M}_i \frac{(\tilde{s} - \tilde{s}_{i-1})^3}{6d_{i-1}} + \tilde{M}_{i-1} \frac{(\tilde{s}_i - \tilde{s})^3}{6d_{i-1}} + \left(x_i - \frac{d_{i-1}^2}{6} \tilde{M}_i \right) \frac{\tilde{s} - \tilde{s}_{i-1}}{d_{i-1}} + \left(x_{i-1} - \frac{d_{i-1}^2}{6} \tilde{M}_{i-1} \right) \frac{\tilde{s}_i - \tilde{s}}{d_{i-1}}, \quad (5.119)$$

$$s \in [\tilde{s}_{i-1}; \tilde{s}_i], \quad i = \overline{1, N},$$

где величины \tilde{M}_i определяются из системы

$$\begin{cases} \tilde{\mu}_i \tilde{M}_{i-1} + 2\tilde{M}_i + \tilde{\lambda}_i \tilde{M}_{i+1} = \frac{6}{d_{i-1} + d_i} \left(\frac{x_{i+1} - x_i}{d_i} - \frac{x_i - x_{i-1}}{d_{i-1}} \right), & i = \overline{1, N-1}, \\ \tilde{M}_0 = \tilde{M}_N = 0, \end{cases}$$

и аналогично

$$S(y; \tilde{s}) = \tilde{L}_i \frac{(\tilde{s} - \tilde{s}_{i-1})^3}{6d_{i-1}} + \tilde{L}_{i-1} \frac{(\tilde{s}_i - \tilde{s})^3}{6d_{i-1}} + \left(y_i - \frac{d_{i-1}^2}{6} \tilde{L}_i \right) \frac{\tilde{s} - \tilde{s}_{i-1}}{d_{i-1}} + \left(y_{i-1} - \frac{d_{i-1}^2}{6} \tilde{L}_{i-1} \right) \frac{\tilde{s}_i - \tilde{s}}{d_{i-1}},$$

$$y \in [\tilde{s}_{i-1}; \tilde{s}_i], \quad i = \overline{1, N}, \quad \begin{cases} \tilde{\mu}_i \tilde{L}_{i-1} + 2\tilde{L}_i + \tilde{\lambda}_i \tilde{L}_{i+1} = \frac{6}{d_{i-1} + d_i} \left(\frac{y_{i+1} - y_i}{d_i} - \frac{y_i - y_{i-1}}{d_{i-1}} \right), & i = \overline{1, N-1}, \\ \tilde{L}_0 = \tilde{L}_N = 0. \end{cases} \quad (5.120)$$

В обеих системах $\tilde{\mu}_i = \frac{d_{i-1}}{d_{i-1} + d_i}$, $\tilde{\lambda} = 1 - \tilde{\mu}_i$.

При определенных требованиях к функциям $x(s)$ и $y(s)$ (s – естественный параметр – длина дуги), а также к кривой L можно получить оценки скорости сходимости, аналогичные полученным ранее. Можно заметить также, что параметрический сплайн не изменяется при переходе к новому параметру $\bar{s} = \gamma \tilde{s}$, где $\gamma > 0$ – произвольный числовой параметр. Поэтому в некоторых случаях удобно полагать $\gamma = \tilde{s}_N^{-1}$. Такую параметризацию называют нормированной по суммарной длине хорд.



Глава 6

Приближенное вычисление интегралов

- 6.1. Вычисление определенного интеграла
- 6.2. Вычисление кратных интегралов



6.1. Вычисление определенного интеграла

- 6.1.1. Задача численного интегрирования: постановка, основные понятия
- 6.1.2. Интерполяционные квадратурные формулы
- 6.1.3. Квадратурные формулы Ньютона–Котеса
- 6.1.4. Примеры квадратурных формул с равноотстоящими узлами
- 6.1.5. Оценка погрешности квадратурных формул
- 6.1.6. Квадратурные формулы наивысшей алгебраической степени точности
- 6.1.7. Квадратурные формулы, содержащие наперед заданные узлы
- 6.1.8. Квадратурные формулы с равными коэффициентами
- 6.1.9. Нестандартные приемы интегрирования



6.1.1. Задача численного интегрирования: постановка, основные понятия

Пусть $f(x)$ – интегрируемая на отрезке $[a, b]$ функция. Ставится задача вычисления определенного интеграла $I = \int_a^b f(x) dx$. Если для функции $f(x)$ можно найти аналитическое выражение первообразной $F(x)$, то интеграл I можно вычислить, используя формулу Ньютона-Лейбница:

$$I = \int_a^b f(x) dx = F(b) - F(a).$$

Однако, как правило, выразить первообразную $F(x)$ через элементарные функции не удается. Поэтому приходится прибегать к приближенному вычислению интеграла. Очевидно, одним из простейших приближенных алгоритмов, которые теоретически можно использовать для этих целей, является вычисление интеграла непосредственно по определению, с помощью интегральных сумм, в качестве одной из которых можно взять, например, такую:

$$S_n = \sum_{i=0}^n f(x_i) \Delta x_i.$$

Как следует из теории, таким образом интеграл I можно найти с любой наперед заданной точностью. Однако практически этот прием мало пригоден из-за медленной сходимости.

Поэтому для построения формул приближенного вычисления интеграла, как правило, используют следующий прием: функцию $f(x)$ заменяют близкой к ней функцией $\varphi(x)$, интеграл от которой просто может быть вычислен аналитически (с помощью формулы Ньютона-Лейбница) (например, алгебраическим многочленом). Успех приближения, как мы помним, зависит от свойств гладкости приближаемой функции. В силу этого подынтегральную функцию чаще всего представляют в виде произведения двух сомножителей, один из которых – достаточно гладкая функция (подлежащая в дальнейшем упрощающей замене), а вторая содержит основные особенности подынтегрального выражения и легко интегрируется, т.е. рассматривают интегралы вида $I = \int_a^b p(x) f(x) dx$. В этом выражении $p(x)$ – фиксированная, не эквивалентная нулю функция (ее мы далее будем называть весовой или просто весом), а $f(x)$ – достаточно гладкая функция, которую далее будем называть интегрируемой.



Пример:

$$\int_{-1}^1 \frac{dx}{\sqrt{1-x^4}} = \int_{-1}^1 \frac{1}{\sqrt{1-x^2}} \frac{1}{\sqrt{1+x^2}} dx.$$

Здесь $p(x) = \frac{1}{\sqrt{1-x^2}}$, и $f(x) = \frac{1}{\sqrt{1+x^2}}$.

Если использовать интерполяционный способ замены $f(x)$ (причем линейный по параметрам), то приближенная формула для вычисления интеграла ([квадратурная формула](#)) будет выглядеть следующим образом:

$$I = \int_a^b p(x) f(x) dx \approx \sum_{k=0}^n A_k f(x_k). \quad (6.1)$$

При этом линейную комбинацию, стоящую в правой части соотношения (6.1), будем называть [квадратурной суммой](#), A_k – ее [коэффициентами](#), а x_k – [узлами](#).

При фиксированном n квадратурная сумма зависит от $2(n+1)$ параметров A_k и x_k ($k = \overline{0, n}$). Их выбор может осуществляться из следующих двух основных соображений: [повышения степени точности квадратурного правила](#) и [минимизации остатка квадратурной формулы на классах функций](#).

Повышение степени точности квадратурного правила. Поскольку формула (6.1) получается путем замены интегрируемой функции $f(x)$ некоторым обобщенным многочленом

$$Q_n(x) = c_0 \varphi_0(x) + c_1 \varphi_1(x) + \cdots + c_n \varphi_n(x)$$

с последующим его интегрированием, то можно ожидать, что если мы выбором узлов x_k и коэффициентов A_k в (6.1) достигнем хорошей точности в интегрировании функций $\varphi_i(x)$, то формула (6.1) должна будет также дать хороший результат (по точности) при вычислении интеграла от всякой функции $f(x)$ из рассматриваемого класса. Эти несложные соображения имеют, разумеется, только наводящее значение и погрешность построенной квадратурной формулы должна быть подвергнута точному анализу и оценке. Но они позволяют указать простой принцип выбора x_k и A_k : будем стремиться выбором x_k и A_k добиваться того, чтобы формула (6.1) давала точный результат для возможно большего числа первых функций $\varphi_i(x)$.



Определение. Говорят, что **квадратурная формула** имеет *степень точности m относительно системы функций $\{\varphi_i(x)\}$* , если она точна на первых m функциях $\varphi_0(x), \dots, \varphi_m(x)$ и не точна на функции $\varphi_{m+1}(x)$, т.е. выполняются соотношения

$$\begin{cases} \int_a^b p(x) \varphi_i(x) dx = \sum_{k=0}^n A_k \varphi_i(x_k), & i = \overline{0, m}, \\ \int_a^b p(x) \varphi_{m+1}(x) dx \neq \sum_{k=0}^n A_k \varphi_{m+1}(x_k). \end{cases} \quad (6.2)$$

Если в качестве системы $\{\varphi_i(x)\}$ взять систему алгебраических многочленов, в частности, систему степеней, т.е. положить $\varphi_i(x) = x^i$, то из предыдущего определения следует определение **алгебраической степени точности квадратурной суммы**:

Определение. Говорят, что **квадратурная формула** имеет *алгебраическую степень точности, равную m* , если она точна для всевозможных многочленов степени m и существует хотя бы один многочлен степени $(m+1)$, для которого формула точной не является.

Требование точности для всевозможных многочленов степени m равносильно требованию точности на любой базисной системе пространства многочленов степени m , и в частности, на системе $\{x^i\}$. Поэтому из **определения** с необходимостью следует выполнение системы соотношений

$$\begin{cases} \int_a^b p(x) x^i dx = \sum_{k=0}^n A_k x_k^i, & i = \overline{0, m}, \\ \int_a^b p(x) x^{m+1} dx \neq \sum_{k=0}^n A_k x_k^{m+1}. \end{cases} \quad (6.3)$$

Систему (6.3) можно использовать как для отыскания алгебраической степени точности заданной квадратурной формулы, так и для построения квадратурной формулы методом неопределенных коэффициентов. В последнем случае первые $(m+1)$ соотношений системы (6.3) рассматриваются как система уравнений (в общем случае нелинейных) относительно неизвестных A_k, x_k (или части их).



Минимизация остатка квадратурной формулы на классах функций. *Остатком квадратурной формулы* естественно называть величину

$$R_n(f) = \int_a^b p(x) f(x) dx - \sum_{k=0}^n A_k f(x_k). \quad (6.4)$$

Если при этом $f(x)$ принадлежит заданному классу F функций, то, очевидно, можно получить такую характеристику остатка: $\sup_{f \in F} |R_n(f)|$, а в качестве задачи поставить задачу поиска минимума этой характеристики.

При этом необходимо иметь в виду следующие соображения: как правило, значения функции $f(x)$ в узлах x_k известны с некоторой погрешностью ε_k , $k = \overline{0, n}$. Эти погрешности при вычислении квадратурной суммы повлекут за собой ошибку $\sum_{k=0}^n A_k \varepsilon_k$. Если считать, что для всех значений k погрешности ε_k ограничены по модулю сверху, т.е. $|\varepsilon_k| \leq \varepsilon$, то для погрешности квадратурной суммы получим оценку

$$\left| \sum_{k=0}^n A_k \varepsilon_k \right| \leq \varepsilon \sum_{k=0}^n |A_k|.$$

Поэтому необходимо подбирать коэффициенты A_k таким образом, чтобы величина $\sum_{k=0}^n |A_k|$ была по возможности меньшей.

Пусть $p(x) \geq 0$ на отрезке $[a, b]$ и квадратурная формула (6.1) имеет алгебраическую степень точности, не меньшую нуля. Тогда

$$\sum_{k=0}^n A_k = \int_a^b p(x) dx > 0.$$

С другой стороны,

$$\sum_{k=0}^n |A_k| \geq \sum_{k=0}^n A_k,$$

причем равенство имеет место только в том случае, когда все A_k положительны. Поэтому условие знакопостоянства коэффициентов квадратурной суммы обеспечивает наименьшую оценку вычислительной погрешности квадратурной формулы, т.е. ее устойчивость.



Помимо отмеченных двух основных соображений в основу выбора параметров квадратурных формул могут быть положены и некоторые другие (например, распределение той же погрешности по заданному закону и т.п.).



6.1.2. Интерполяционные квадратурные формулы

Как мы уже отмечали, для построения [квадратурных формул](#) чаще всего пользуются интерполированием интегрируемой функции, при этом наиболее употребительный класс приближающих функций – алгебраические многочлены.

Выберем на отрезке интегрирования $[a, b]$ ($n + 1$) произвольных различных точек x_0, x_1, \dots, x_n и проинтерполируем функцию $f(x)$ по ее значениям в этих точках. Интерполяционный многочлен в данном случае удобнее брать в [форме Лагранжа](#):

$$f(x) = P_n(x) + r_n(x),$$

где

$$P_n(x) = \sum_{k=0}^n \Phi_k(x) f(x_k) = \sum_{k=0}^n \frac{\omega_{n+1}(x)}{(x - x_k) \omega'_{n+1}(x_k)} f(x_k),$$

а $r_n(x)$ – [остаток интерполирования](#).

Отсюда получим:

$$I = \int_a^b p(x) f(x) dx = \sum_{k=0}^n A_k f(x_k) + R_n(f), \quad (6.5)$$

причем

$$A_k = \int_a^b p(x) \Phi_k(x) dx = \int_a^b p(x) \frac{\omega_{n+1}(x)}{(x - x_k) \omega'_{n+1}(x_k)} dx, \quad (6.6)$$

а

$$R_n(f) = \int_a^b p(x) r_n(x) dx.$$

Если остаток интерполирования $r_n(x)$ мал, то и величина $R_n(f)$ также будет малой и, следовательно, в (6.5) ей можно пренебречь. В итоге получим квадратурную формулу (6.1).



Определение. Квадратурные формулы, коэффициенты которых вычисляются по формулам (6.6), называют **интерполяционными**.

Интерполяционные квадратурные формулы могут быть охарактеризованы следующей простой теоремой.

Теорема 6.1. Для того чтобы квадратурная формула была интерполяционной, необходимо и достаточно, чтобы она была точной для всевозможных многочленов степени не выше n (т.е. имела алгебраическую степень точности, равную n). [Доказательство]

Из доказанной теоремы следует, что любая квадратурная формула, точная для многочленов степени не выше n и имеющая $(n+1)$ узел, является интерполяционной.

С помощью имеющихся представлений для остатка интерполирования $r_n(x)$ мы можем получить различные представления для остатка квадратурной формулы $R_n(f)$.

Пусть, например, $f(x) \in C^{n+1}[a, b]$. Тогда остаток интерполирования в форме Лагранжа имеет вид

$$r_n(x) = \omega_{n+1}(x) \frac{f^{(n+1)}(\xi)}{(n+1)!}, \quad \xi \in [a, b].$$

Отсюда получаем:

$$R_n(f) = \frac{1}{(n+1)!} \int_a^b p(x) \omega_{n+1}(x) f^{(n+1)}(\xi) dx. \quad (6.7)$$

Если допустить, что $|f^{(n+1)}(x)| \leq M$ для всех $x \in [a, b]$, то для остатка квадратурной формулы получим оценку сверху

$$|R_n(f)| \leq \frac{M}{(n+1)!} \int_a^b |p(x) \omega_{n+1}(x)| dx, \quad (6.8)$$

которая является вычислимой и может быть, следовательно, использована для практической (априорной) оценки погрешности численного интегрирования. Заметим, что, как и в случае интерполирования, оценка (6.8) является достижимой.



6.1.3. Квадратурные формулы Ньютона–Котеса

Среди [интерполяционных квадратурных формул](#) ранее других были построены [формулы Ньютона–Котеса](#). Они относятся к случаю равноотстоящих узлов.

Отрезок $[a, b]$ разделим на n равных частей длины $h = \frac{b-a}{n}$ и точки деления $x_k = a + kh$, $k = \overline{0, n}$, примем за узлы интерполяционной квадратурной формулы. Саму формулу запишем в виде

$$\int_a^b p(x) f(x) dx \approx (b-a) \sum_{k=0}^n B_k^n f(a+kh), \quad (6.9)$$

где

$$B_k^n = \frac{1}{b-a} A_k = \frac{1}{b-a} \int_a^b p(x) \frac{\omega_{n+1}(x)}{(x-a-kh)\omega'_{n+1}(a+kh)} dx, \quad k = \overline{0, n} \quad (6.10)$$

Введем новую переменную по формуле $x = a + th$, B.5. $t = \frac{x-a}{h}$. Тогда имеем:

$$0 \leq t \leq n; \quad b-a = nh; \quad x-a-kh = th-kh = h(t-k); \quad dx = hdt;$$

$$\omega_{n+1}(x) = \omega_{n+1}(a+th) = h^{n+1} t(t-1) \cdots (t-n);$$

$$\omega'_{n+1}(x_k) = \omega'_{n+1}(a+kh) = (x_k - a)(x_k - a - h) \cdots (x_k - a - (k-1)h) \cdot (x_k - a - (k+1)h) \cdots$$

$$\cdots (x_k - a - nh) = kh \cdot (k-1)h \cdots h \cdot (-1) \cdot h \cdot (-2) \cdot h \cdots (-1)(n-k) \cdot h = (-1)^{n-k} h^n k! \cdot (n-k)!.$$



Подставляя эти выражения в формулу (6.10), получим:

$$\begin{aligned} B_k^n &= \frac{(-1)^{n-k}}{k! \cdot (n-k)! \cdot (b-a)} \int_0^n p(a+th) \frac{h^{n+1} t(t-1)\cdots(t-n)}{h(t-k) \cdot h^n} \cdot h dt = \\ &= \frac{(-1)^{n-k}}{n \cdot k! \cdot (n-k)!} \int_0^n p(a+th) \frac{t(t-1)\cdots(t-n)}{(t-k)} dt, \quad k = \overline{0, n}. \end{aligned} \quad (6.11)$$

Отметим некоторые дополнительные свойства коэффициентов (6.11) и квадратурной формулы (6.9) в случае постоянной весовой функции ($p(x) \equiv 1$).

Свойства квадратурных формул Ньютона–Котеса

1. $B_k^n = B_{n-k}^n$, т.е. равноотстоящие от концов суммы коэффициенты равны. [\[Доказательство\]](#)
2. Квадратурная формула (6.9) точна для любой функции $f(x)$, нечетной относительно середины отрезка $[a, b]$, т.е. функции, для которой выполняется соотношение $f\left(x - \frac{a+b}{2}\right) = -f\left(\frac{a+b}{2} - x\right)$. [\[Доказательство\]](#)
3. При четном значении n (тогда общее количество узлов нечетно и, как следствие, точка $x = \frac{a+b}{2}$ также является узлом) квадратурная формула (6.9) имеет алгебраическую степень точности, равную $(n+1)$. [\[Доказательство\]](#)

Замечание 6.1. Свойство 3, по сути, означает, что мы проводим интерполяцию по n простым узлам x_k и одному двукратному $x^* = \frac{a+b}{2}$, поскольку квадратурные формулы, получающиеся в означенном случае и при интерполяции по $(n+1)$ простому узлу, получаются одинаковыми. Это имеет, следовательно, значение при исследовании остатка.

Было установлено также, что при $n = 8$ в формуле (6.9) появляются отрицательные коэффициенты, а при $n \geq 10$ среди B_k^n обязательно будут отрицательные, причем при больших n – сколь угодно большие по модулю (в силу того, что сумма всех таких коэффициентов постоянна и равна единице). Поэтому

$$\sum_{k=0}^n |B_k^n| \xrightarrow{n \rightarrow \infty} \infty,$$



что приводит к катастрофическому накоплению вычислительной погрешности. В силу этого квадратурные формулы Ньютона–Котеса с большим количеством узлов, как правило, не используются. Для достижения же высокой точности процедуры приближенного интегрирования исходный отрезок $[a, b]$ разбивают на отрезки небольшой длины. Тогда на каждом из них можно получить хороший результат уже при небольших значениях n . Получаемые таким образом квадратурные формулы называют *составными* или *обобщенными*.



6.1.4. Примеры квадратурных формул с равноотстоящими узлами

Квадратурные формулы с одним узлом

Формула трапеций

Формула Симпсона

Квадратурные формулы с одним узлом

Начнем рассмотрение со случая $n = 0$ (один узел интерполирования), который по формальным причинам не охватывается рассмотренными выше [формулами Ньютона-Котеса](#). Получающиеся квадратурные формулы, исходя из геометрических соображений, носят название [формул прямоугольников](#).

Формула левых прямоугольников. Единственным узлом интерполирования в данном случае предполагается левый конец отрезка интегрирования, т.е. $x_0 = a$. Интерполяционным многочленом для функции $f(x)$ по этим данным будет $P_0(x) = f(a)$. Подставляя это выражение вместо $f(x)$ под знак интеграла и выполняя интегрирование, получим искомую квадратурную формулу:

$$\int_a^b f(x) dx \approx (b - a) f(a). \quad (6.12)$$

Заметим, что с геометрической точки зрения в правой части формулы (6.12) мы имеем площадь прямоугольника, одна сторона которого равна длине отрезка интегрирования, а вторая – значению интегрируемой функции $f(x)$ на левом конце данного отрезка (отсюда – название квадратурной формулы).

Если $f(x) \in C^1 [a, b]$, то остаток интерполирования в форме Лагранжа имеет вид $r_0(x) = (x - a) f'(ξ)$, где $ξ \in [a, b]$ (и зависит от x).

Следовательно, $R_0^n(f) = \int_a^b r_0(x) dx = \int_a^b (x - a) f'(\xi) dx$. Чтобы упростить это выражение, восполь-

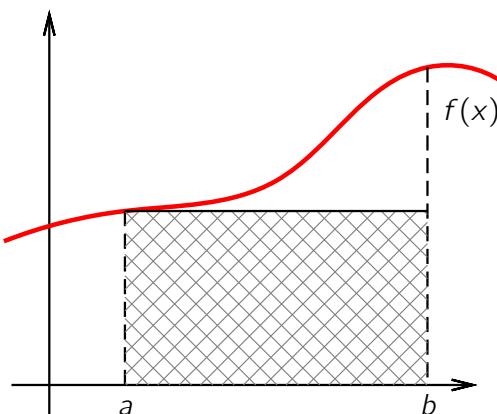


Рисунок 6.1

зумеся теоремой о среднем. Поскольку функция $(x - a)$ сохраняет знак на отрезке интегрирования (неотрицательна), то $(\eta \in [a, b])$

$$R_0^n(f) = f'(\eta) \int_a^b (x - a) dx = \frac{(b - a)^2}{2} f'(\eta). \quad (6.13)$$

Таким образом, формула (6.12) — *квадратурная формула левых прямоугольников*, а (6.13) — ее остаток. Легко видеть, что *алгебраическая степень точности* формулы (6.12) равна нулю, а с помощью (6.13) можно получить оценку сверху для величины погрешности численного интегрирования. Однако управлять величиной этой погрешности, очевидно, не представляется возможным. Поэтому построим сейчас на базе (6.12) составную формулу левых прямоугольников.

Разобьем отрезок $[a, b]$ на N частей длины $h = \frac{b-a}{N}$, воспользуемся свойством аддитивности интеграла



и на каждом из отрезков получившегося разбиения применим квадратурную формулу левых прямоугольников (6.12):

$$\int_a^b f(x) dx = \sum_{k=0}^{N-1} \int_{x_k}^{x_{k+1}} f(x) dx \approx h \sum_{k=0}^{N-1} f(x_k) = h \sum_{k=0}^{N-1} f(a + kh). \quad (6.14)$$

Это и есть *составная (обобщенная) квадратурная формула левых прямоугольников*. Очевидно, для ее остатка по аналогии с (6.13) можно записать представление

$$R_L(f) = \frac{h^2}{2} \sum_{k=0}^{N-1} f'(\xi_k), \quad \xi_k \in [x_k, x_{k+1}].$$

Упростим полученное представление остатка. Поскольку функция $f'(x)$ непрерывна на отрезке $[a, b]$, то в соответствии с теоремой Вейерштрасса она достигает на нем своих наименьшего и наибольшего значений, т.е. $m = \min_{x \in [a, b]} f'(x)$, $M = \max_{x \in [a, b]} f'(x)$, причем $m \leq f'(x) \leq M$ для всех $x \in [a, b]$. Следовательно,

$$Nm \leq \sum_{k=0}^{N-1} f'(\xi_k) \leq NM \text{ и } m \leq \frac{1}{N} \sum_{k=0}^{N-1} f'(\xi_k) \leq M.$$

Тогда по теореме о промежуточном значении непрерывной функции существует точка $\eta \in [a, b]$, в которой выполняется соотношение $f'(\eta) = \frac{1}{N} \sum_{k=0}^{N-1} f'(\xi_k)$. Поэтому

$$R_L(f) = \frac{h^2}{2} \sum_{k=0}^{N-1} f'(\xi_k) = \frac{h}{2} \cdot \frac{b-a}{N} \sum_{k=0}^{N-1} f'(\xi_k) = h \frac{b-a}{2} f'(\eta) = \frac{(b-a)^2}{2N} f'(\eta), \quad \eta \in [a, b] \quad (6.15)$$

Очевидно, степень точности не повысилась, но $R_L(f) \xrightarrow[N \rightarrow \infty]{} 0$ и, таким образом, у нас появляется рычаг, с помощью которого можно воздействовать на величину погрешности.

Замечание 6.2. Точки разбиения не обязаны быть, вообще говоря, равнотстоящими. В этом более общем случае составная формула левых прямоугольников будет иметь вид

$$\int_a^b f(x) dx = \sum_{k=0}^{N-1} \int_{x_k}^{x_{k+1}} f(x) dx \approx \sum_{k=0}^{N-1} (x_{k+1} - x_k) f(x_k) = \sum_{k=0}^{N-1} h_{k+1} f(x_k), \quad (6.16)$$



а ее остаток будет таким:

$$R_n(f) = \sum_{k=0}^{N-1} \frac{h_{k+1}^2}{2} f'(\xi_k)$$

и упростить его до вида, аналогичного представлению (6.16), не удается.

В дальнейшем замечаний по поводу неравномерной сетки более делать не будем.

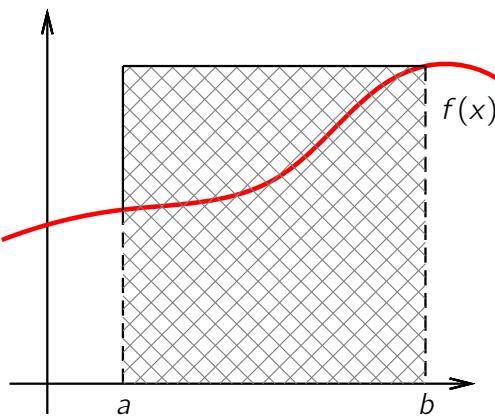


Рисунок 6.2

Формула правых прямоугольников. Здесь единственный узел интерполяции $x_0 = b$. Если $f(x) \in C^1[a, b]$, то по аналогии с рассуждениями [предыдущего пункта](#) получим аналогичную по точности квадратурной формуле левых прямоугольников формулу [правых прямоугольников](#):

$$\int_a^b f(x) dx \approx (b - a) f(b) \quad (6.17)$$



и ее остаток

$$R_0^n(f) = f'(\eta) \int_a^b (x - b) dx = -\frac{(b-a)^2}{2} f'(\eta), \quad \eta \in [a, b]. \quad (6.18)$$

Так же легко получается *составная формула правых прямоугольников*

$$\int_a^b f(x) dx = \sum_{k=1}^N \int_{x_{k-1}}^{x_k} f(x) dx \approx h \sum_{k=1}^N f(x_k) = h \sum_{k=1}^N f(a + kh) \quad (6.19)$$

и ее остаток

$$R_n(f) = -h \frac{b-a}{2} f'(\eta) = -\frac{(b-a)^2}{2N} f'(\eta), \quad \eta \in [a, b]. \quad (6.20)$$

Заметим, что в случае, когда производная $f'(x)$ сохраняет знак на отрезке $[a, b]$, квадратурные формулы правых и левых прямоугольников дают оценку значения нужного значения интеграла с двух сторон.

Формула средних прямоугольников Как и в двух предыдущих вариантах, для приближения интегрируемой функции используется единственный узел. В этом случае – это середина отрезка интегрирования, т.е. $x_0 = \frac{a+b}{2}$. Интегрирование соответствующего интерполяционного многочлена дает *квадратурную формулу средних прямоугольников*

$$\int_a^b f(x) dx \approx (b-a) f\left(\frac{a+b}{2}\right). \quad (6.21)$$

Для вывода же формулы остатка воспользуемся замечанием из предыдущего пункта. Симметричное расположение единственного узла дает возможность повысить алгебраическую степень точности формулы (6.21) до единицы и считать интерполяционную замену интегрируемой функции интерполированием с кратным узлом. *Остаток соответствующей формулы Эрмита* в случае $f(x) \in C^2[a, b]$ имеет вид

$$r_0(x) = \left(x - \frac{a+b}{2}\right)^2 \frac{f''(\xi)}{2}.$$

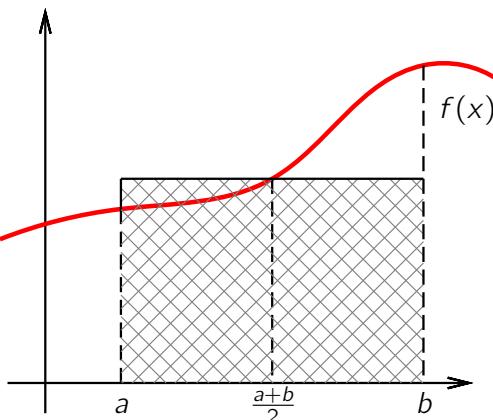


Рисунок 6.3

Поэтому

$$R_0^c(f) = \int_a^b \left(x - \frac{a+b}{2} \right)^2 \frac{f''(\xi)}{2} dx = \frac{f''(\eta)}{2} \int_a^b \left(x - \frac{a+b}{2} \right)^2 dx = \frac{(b-a)^3}{24} f''(\eta), \quad \eta \in [a, b]. \quad (6.22)$$

По аналогии с предыдущими двумя случаями получаем *составную формулу средних прямоугольников*

$$\int_a^b f(x) dx = \sum_{k=0}^{N-1} \int_{x_k}^{x_{k+1}} f(x) dx \approx h \sum_{k=0}^{N-1} f\left(\frac{x_k + x_{k+1}}{2}\right) = h \sum_{k=0}^{N-1} f\left(a + kh + \frac{h}{2}\right) \quad (6.23)$$

и ее остаток

$$R_c(f) = h^2 \frac{b-a}{24} f''(\eta) = \frac{(b-a)^3}{24N^2} f''(\eta), \quad \eta \in [a, b]. \quad (6.24)$$



Формула трапеций

Квадратурная формула трапеций (малая) получается как частный случай формулы Ньютона-Котеса (6.9) (или непосредственно) при $n = 1$ и имеет вид

$$\int_a^b f(x) dx \approx \frac{b-a}{2} (f(a) + f(b)). \quad (6.25)$$

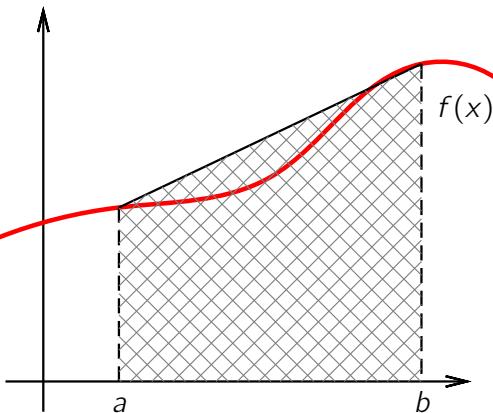


Рисунок 6.4

Остаток интерполяции многочленом первой степени при $f(x) \in C^2[a, b]$ имеет вид

$$r_1(x) = (x-a)(x-b) \frac{f''(\xi)}{2}.$$



Вверх

Назад

Вперёд

Пред.

След.

Указатель Помощь Экран

Поэтому (вновь используя факт знакопостоянства одного из сомножителей подынтегральной функции) имеем

$$R_1^T(f) = \int_a^b (x-a)(x-b) \frac{f''(\xi)}{2} dx = \frac{f''(\eta)}{2} \int_a^b (x-a)(x-b) dx = -\frac{(b-a)^3}{12} f''(\eta), \quad \eta \in [a, b] \quad (6.26)$$

По аналогии с предыдущими случаями получаем *составную формулу трапеций*

$$\int_a^b f(x) dx = \sum_{k=0}^{N-1} \int_{x_k}^{x_{k+1}} f(x) dx \approx \frac{h}{2} \sum_{k=0}^{N-1} [f(x_k) + f(x_{k+1})] = h \left[\frac{f(a) + f(b)}{2} + \sum_{k=1}^{N-1} f(x_k) \right] \quad (6.27)$$

и ее остаток

$$R_T(f) = -h^2 \frac{b-a}{12} f''(\eta) = -\frac{(b-a)^3}{12N^2} f''(\eta), \quad \eta \in [a, b]. \quad (6.28)$$

Пара квадратурных формул трапеций и средних прямоугольников также при определенных условиях (знакопостоянство второй производной интегрируемой функции на отрезке интегрирования) дает двустороннее приближение.

Формула Симпсона

Указанная квадратурная формула получается как частный случай [квадратурной формулы Ньютона-Котеса](#) при $n = 2$. Таким образом, интерполирование интегрируемой функции $f(x)$ в этом случае осуществляется по трем узлам, равномерно, включая концы, расположенным на отрезке интегрирования: $x_0 = a$, $x_1 = \frac{a+b}{2}$, $x_2 = b$. По формуле (6.11) при $p(x) \equiv 1$ (учитывая свойство 1) найдем:

$$B_0^2 = B_2^2 = \frac{(-1)^{2-0}}{2 \cdot 0! \cdot 2!} \int_0^2 \frac{t(t-1)(t-2)}{t} dt = \frac{1}{4} \int_0^2 (t^2 - 3t + 2) dt = \frac{1}{6}.$$

$$B_1^2 = \frac{(-1)^{2-1}}{2 \cdot 1! \cdot 1!} \int_0^2 \frac{t(t-1)(t-2)}{t-1} dt = -\frac{1}{2} \int_0^2 (t^2 - 2t) dt = \frac{4}{6}.$$



Следовательно, *квадратурная формула Симпсона* имеет вид

$$\int_a^b f(x) dx \approx \frac{b-a}{6} \left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right). \quad (6.29)$$

Ее алгебраическая степень точности равна 3.

Остаток формулы Симпсона получим, используя замечание [6.1. Остаток многочлена Эрмита](#) в нашем случае имеет вид

$$r_2(x) = (x-a) \left(x - \frac{a+b}{2} \right)^2 (x-b) \frac{f^{(4)}(\xi)}{4!},$$

если $f(x) \in C^4[a, b]$. При этом, очевидно, многочленный сомножитель остатка знакопостоянен. Поэтому интегрирование данного остатка с применением теоремы о среднем дает

$$\begin{aligned} R_2^C(f) &= \int_a^b (x-a) \left(x - \frac{a+b}{2} \right)^2 (x-b) \frac{f^{(4)}(\xi)}{4!} dx = \frac{f^{(4)}(\eta)}{4!} \int_a^b (x-a) \left(x - \frac{a+b}{2} \right)^2 (x-b) dx = \\ &= -\left(\frac{b-a}{2}\right)^5 \frac{f^{(4)}(\eta)}{90} = -\frac{(b-a)^5}{2880} f^{(4)}(\eta). \end{aligned} \quad (6.30)$$

Составная формула Симпсона в литературе встречается в двух вариантах:

а) отрезок $[a, b]$ разбивается на произвольное количество (N) частей и на каждой из этих частей интеграл заменяется формулой типа [\(6.29\)](#). В итоге получим квадратурную формулу

$$\int_a^b f(x) dx = \sum_{k=0}^{N-1} \int_{x_k}^{x_{k+1}} f(x) dx \approx \frac{h}{6} \sum_{k=0}^{N-1} \left[f(x_k) + 4f\left(\frac{x_k+x_{k+1}}{2}\right) + f(x_{k+1}) \right]. \quad (6.31)$$

Ее остаток по аналогии с рассмотренными выше случаями с использованием формулы [\(6.30\)](#) и имеет вид

$$R_C(f) = -h^4 \frac{b-a}{2880} f^{(4)}(\eta) = -\frac{(b-a)^5}{2880 N^4} f^{(4)}(\eta), \quad \eta \in [a, b]. \quad (6.32)$$



6) Во второй версии отрезок интегрирования разбивается на четное количество ($2N$) частей, и интеграл заменяется формулой Симпсона на каждой паре отрезков разбиения (т.е. на отрезке длиной $2h$). Такой вариант составной формулы Симпсона выглядит следующим образом:

$$\begin{aligned} \int_a^b f(x) dx &= \sum_{k=0}^{N-1} \int_{x_{2k}}^{x_{2k+2}} f(x) dx \approx \frac{2h}{6} \sum_{k=0}^{N-1} [f(x_{2k}) + 4f(x_{2k+1}) + f(x_{2k+2})] = \\ &= \frac{h}{3} [f(a) + f(b)] + \frac{4h}{3} [f(a+h) + f(a+3h) + \dots + f(b-h)] + \\ &\quad + \frac{2h}{3} [f(a+2h) + f(a+4h) + \dots + f(b-2h)], \end{aligned} \tag{6.33}$$

а ее остаток примет вид

$$R_C(f) = -h^4 \frac{b-a}{180} f^{(4)}(\eta) = -\frac{(b-a)^5}{2880 N^4} f^{(4)}(\eta), \quad \eta \in [a, b]. \tag{6.34}$$



6.1.5. Оценка погрешности квадратурных формул

Учет избыточной гладкости интегрируемой функции

Правило Рунге

Заметим, что все полученные выше представления остатков квадратурных формул (как простых, так и составных) позволяют решить задачу об априорной оценке погрешности приближенного значения интеграла, доставляемого соответствующей квадратурной формулой. Кроме того, в случае составных формул с их помощью можно выбрать параметры заданной квадратурной формулы (число N частей, на которое необходимо разбивать отрезок интегрирования или шаг разбиения) таким образом, чтобы модуль остатка был не более некоторой заданной величины ε (пользовательского требования к точности).

Так, например, если исходная квадратурная формула есть составная формула [левых](#) или [правых](#) [прямоугольников](#), то $|R_0(f)| \leq \frac{(b-a)^2 M_1}{2N} \leq \varepsilon$, где $M_k = \max_{x \in [a,b]} |f^{(k)}(x)|$, ε – величина пользовательского требования к точности, откуда для величины N получаем оценку

$$N_{\Pi} \geq \frac{(b-a)^2 M_1}{2\varepsilon}. \quad (6.35)$$

Аналогично для составной формулы [средних](#) [прямоугольников](#)

$$N_c \geq \sqrt{\frac{(b-a)^3 M_2}{24\varepsilon}}, \quad (6.36)$$

для [формулы трапеций](#)

$$N_t \geq \sqrt{\frac{(b-a)^3 M_2}{12\varepsilon}}, \quad (6.37)$$

для формулы Симпсона (6.32)

$$N_C \geq \sqrt{\frac{(b-a)^5 M_4}{2880\varepsilon}}. \quad (6.38)$$



Однако существенным недостатком указанного подхода к вычислению интеграла с заданной точностью являемся необходимость аналитического вычисления оценок норм производных от интегрируемой функции (в том числе достаточно высоких порядков), что далеко не всегда представляет собой простую задачу.

Учет избыточной гладкости интегрируемой функции

Все приведенные выше представления для остатков квадратурных формул были получены в предположении, что интегрируемая функция принадлежит вполне определенному классу гладкости. Естественно, если гладкость оказывается недостаточной, то соответствующее представление также не имеет места. В то же время, если функция обладает большей по сравнению с минимально необходимой степенью гладкости, то это позволяет при сохранении порядка выделить из погрешности составной квадратурной формулы некоторое количество последовательных главных частей.

Покажем, как это можно сделать, на примере [составной формулы средних прямоугольников](#). Пусть, скажем, $f(x) \in C^4[a, b]$. Тогда остаток формулы средних прямоугольников есть величина порядка $O(h^2)$ (см. [\(6.24\)](#)). В то же время, раскладывая интегрируемую функцию на отрезке $[x_k, x_{k+1}]$ в ряд Тейлора в точке $x_{k+\frac{1}{2}}$ и учитывая, что интеграл по указанному отрезку от нечетных степеней разности $(x - x_{k+\frac{1}{2}})$ равен нулю, можем записать:

$$\begin{aligned} & \int_{x_k}^{x_{k+1}} f(x) dx = \\ &= \int_{x_k}^{x_{k+1}} [f\left(x_{k+\frac{1}{2}}\right) + \left(x - x_{k+\frac{1}{2}}\right) f'\left(x_{k+\frac{1}{2}}\right) + \\ &+ \frac{\left(x - x_{k+\frac{1}{2}}\right)^2}{2} f''\left(x_{k+\frac{1}{2}}\right) + \frac{\left(x - x_{k+\frac{1}{2}}\right)^3}{6} f'''\left(x_{k+\frac{1}{2}}\right) + \\ &+ \frac{\left(x - x_{k+\frac{1}{2}}\right)^4}{24} f^{IV}(\xi)] d\xi = \end{aligned}$$



$$= hf \left(x_{k+\frac{1}{2}} \right) + \frac{h^3}{24} f'' \left(x_{k+\frac{1}{2}} \right) + \frac{h^5}{1920} f^{IV} (\eta_k).$$

Тогда

$$I = \int_a^b f(x) dx = \sum_{k=0}^{N-1} \int_{x_k}^{x_{k+1}} f(x) dx = h \sum_{k=0}^{N-1} f \left(x_{k+\frac{1}{2}} \right) + \frac{h^2}{24} \left[h \sum_{k=0}^{N-1} f'' \left(x_{k+\frac{1}{2}} \right) \right] + \frac{h^4 (b-a)}{1920} f^{IV} (\eta).$$

Заметим теперь, что стоящее в квадратных скобках в правой части последнего равенства выражение есть не что иное, как [квадратурная формула средних прямоугольников](#) для вычисления интеграла $\int_a^b f''(x) dx$, т.е.

$$h \sum_{k=0}^{N-1} f'' \left(x_{k+\frac{1}{2}} \right) = \int_a^b f''(x) dx + O(h^2).$$

Следовательно,

$$\begin{aligned} I &= \int_a^b f(x) dx = h \sum_{k=0}^{N-1} f \left(x_{k+\frac{1}{2}} \right) + \frac{h^2}{24} \left(\int_a^b f''(x) dx + \frac{h^2}{24} f^{IV} (\eta_1) \right) + \frac{h^4 (b-a)}{1920} f^{IV} (\eta) = \\ &\quad (6.39) \end{aligned}$$

$$= h \sum_{k=0}^{N-1} f \left(x_{k+\frac{1}{2}} \right) + \frac{h^2}{24} \int_a^b f''(x) dx + O(h^4),$$

или

$$I = Q^C(h, f) + Ch^2 + O(h^4),$$

где через $Q^C(h, f)$ обозначена составная формула средних прямоугольников (6.23), а

$$C = \int_a^b f''(x) dx.$$



Таким образом, из остатка квадратурной формулы выделена первая главная часть. С помощью аналогичных рассуждений указанный процесс, вообще говоря, можно продолжить, предполагая еще более высокую гладкость интегрируемой функции. Аналогичные разложения также могут быть получены и для других квадратурных формул.

Представление (6.37) можно достаточно просто использовать с целью аналитического улучшения составной формулы средних прямоугольников. Вычисляя константу C , можем записать квадратурную формулу

$$I = \int_a^b f(x) dx \approx Q_1(h, f) = h \sum_{k=0}^{N-1} f\left(x_{k+\frac{1}{2}}\right) + \frac{h^2}{24} (f'(b) - f'(a)), \quad (6.40)$$

погрешность которой является величиной порядка $O(h^4)$ (а алгебраическая степень точности равна трем).

Правило Рунге

Выше мы показали, что погрешность составных квадратурных формул при условии дополнительной гладкости интегрируемой функции может быть разложена в ряд по последовательным главным частям. Этот факт можно использовать для апостериорных оценок погрешности. Способ, о котором пойдет речь ниже, носит название [правила Рунге](#) и может быть применен не только для оценки погрешности приближенного вычисления интегралов, но и для оценки погрешности любых других алгоритмов, допускающих упомянутое разложение погрешности по последовательным главным частям. Главная его идея заключается в том, чтобы по нескольким приближенным значениям искомой величины, полученным при различных значениях параметров вычислительного процесса (например, шага сетки h) вычислить параметры главных частей разложения остатка (такие как константы одной или нескольких последовательных главных частей, показатели степени шага сетки в этих главных частях и т.п.).

Покажем, как это можно технически реализовать в простейшем случае на примере оценки погрешности квадратурной формулы.

Итак, пусть имеет место разложение

$$R(h, f) = Ch^m + O(h^{m+p}).$$



Тогда, если I – искомый интеграл, а $Q(h, f)$ – аппроксимирующая его [квадратурная сумма](#), то

$$I = Q(h, f) + R(h, f) \approx Q(h, f) + Ch^m.$$

Выполнив вычисление интеграла с помощью рассматриваемой квадратурной суммы с двумя различными значениями параметра h , можем записать приближенную систему, состоящую из двух уравнений, неизвестными которой являются константа C и точное значение искомого интеграла I :

$$\begin{cases} I \approx Q(h_1, f) + Ch_1^m, \\ I \approx Q(h_2, f) + Ch_2^m. \end{cases}$$

Отсюда, исключая I , находим:

$$C \approx \frac{Q(h_2, f) - Q(h_1, f)}{h_1^m - h_2^m}.$$

Следовательно,

$$R(h_1, f) \approx h_1^m \frac{Q(h_2, f) - Q(h_1, f)}{h_1^m - h_2^m}. \quad (6.41)$$

Таким образом, мы получили выражение для вычисления главной части остатка квадратурной формулы, по которой можно проводить практическую оценку погрешности полученного приближенного значения интеграла $Q(h_1, f)$.

Вычисленное значение предполагает дальнейшую реакцию (программным путем) на его абсолютную величину в сравнении с задаваемой пользовательской величиной погрешности ε . В случае преобладания первой очевидным рецептом является дальнейшее измельчение сетки узлов (т.е. уменьшение величины шага h). При этом наиболее удобным способом организации работы служит выбор $h_2 = \frac{h_1}{2}$. Тогда в случае необходимости уменьшения шага полагают $h_1 = h_2$ и повторяют процесс вычисления интеграла. В этом случае значение $Q(h_1, f)$ оказывается полученным на предыдущем шаге процесса и, таким образом, оказывается возможной существенная экономия в объеме вычислений. Интеграл считается вычисленным с заданной пользователем точностью в том случае, когда вычисленная по формуле (6.41) величина остатка (ее модуль) оказывается меньше пользовательского требования ε .



Заметим, что вычисленная величина главной части остатка позволяет также и уточнять само приближенное значение интеграла. Это можно сделать, например, так:

$$I \approx Q(h_1, f) + h_1^m \frac{Q(h_2, f) - Q(h_1, f)}{h_1^m - h_2^m}. \quad (6.42)$$

Также следует отметить, что правило Рунге может использоваться и в том случае, когда величина m априори неизвестна (например, в случае, когда интегрируемая функция не обладает достаточной для выписывания очередного члена разложения гладкостью). Кроме того, возможна и такая организация работы, которая позволяет вычислить сразу несколько последовательных главных частей. При этом, конечно же, необходимо выполнять расчеты не на двух, а на большем числе вложенных сеток.



6.1.6. Квадратурные формулы наивысшей алгебраической степени точности

Тождество Кристоффеля–Дарбу

Остаток квадратурных формул типа Гаусса

Некоторые частные случаи квадратурных формул типа Гаусса

Выше мы получили следующий результат: в случае произвольного расположения узлов x_k квадратурную формулу

$$\int_a^b p(x) f(x) dx \approx \sum_{k=0}^n A_k f(x_k)$$

за счет выбора коэффициентов A_k можно сделать точной для всех алгебраических многочленов до степени n включительно. Заметим, что использование свойства симметрии в расположении узлов (формулы [средних прямоугольников](#) и [Симпсона](#)) приводит к увеличению алгебраической степени точности на единицу.

Поставим задачу: выяснить, чего можно достичь в смысле повышения [алгебраической степени точности](#) за счет специального расположения узлов. Так как число узлов равно $(n+1)$, то можно надеяться за счет их выбора увеличить алгебраическую степень точности на $(n+1)$, т.е. увеличить ее до $n + (n+1) = 2n+1$.

Установим сейчас условия, при которых квадратурная формула будет иметь алгебраическую степень точности, равную $2n+1$. При этом вместо узлов x_k нам будет удобнее рассматривать многочлен $\omega_{n+1}(x) = (x - x_0) \cdots (x - x_n)$.

Теорема 6.2 (критерий квадратурных формул НАСТ). Для того чтобы [квадратурная формула](#) с $(n+1)$ узлами была точной для любых алгебраических многочленов до степени $(2n+1)$ включительно, необходимо и достаточно выполнение условий:

1) квадратурная формула должна быть [интерполяционной](#), т.е.

$$A_k = \int_a^b p(x) \frac{\omega_{n+1}(x)}{(x - x_k) \omega'_{n+1}(x_k)} dx, \quad k = \overline{0, n}. \quad (6.43)$$



2) многочлен $w_{n+1}(x)$ должен быть ортогонален по данному весу $p(x)$ на отрезке $[a, b]$ ко всем многочленам степени не выше n , т.е.

$$\int_a^b p(x) w_{n+1}(x) Q_m(x) dx = 0, \quad m \leq n. \quad (6.44)$$

[\[Доказательство\]](#)

Выясним теперь, когда требования, сформулированные в [теореме 6.2](#), выполнимы. Вопрос фактически сводится к нахождению такого многочлена $w_{n+1}(x)$, который удовлетворяет соотношениям [\(6.44\)](#), причем его корни действительны, различны и все лежат на отрезке $[a, b]$.

Теорема 6.3. Если весовая функция $p(x)$ сохраняет знак на отрезке $[a, b]$, то приведенный многочлен $w_{n+1}(x)$ степени $(n + 1)$, ортогональный на данном отрезке по весу $p(x)$ ко всем многочленам меньшей степени, существует и единственен для любого фиксированного n . При этом все его корни действительны, различны и лежат внутри данного отрезка.

[\[Доказательство\]](#)

Следствие 6.1. Если весовая функция $p(x)$ знакопостоянна на отрезке интегрирования $[a, b]$, то квадратурная формула вида [\(6.1\)](#) с $(n + 1)$ узлами, точная для любого многочлена до степени $2n + 1$ включительно, существует для любого фиксированного значения n .

Возникает вопрос: будет ли степень $2n + 1$ наивысшей? Ответ на него дает

Теорема 6.4. Если $p(x)$ знакопостоянна на отрезке интегрирования, то ни при каком выборе узлов и коэффициентов [квадратурной формулы](#) с $(n + 1)$ узлами не может быть точной для любого алгебраического многочлена степени $2n + 2$.

[\[Доказательство\]](#)

Отметим также, что справедлива

Теорема 6.5. Если квадратурная формула вида [\(6.1\)](#) точна для всевозможных многочленов степени $2n$, то при знакопостоянной весовой функции весовой функции $p(x)$ все ее коэффициенты A_k имеют один и тот же знак (совпадающий со знаком $p(x)$).

[\[Доказательство\]](#)



Следовательно, квадратурные формулы типа Гаусса имеют все коэффициенты одного знака, что равносильно их вычислительной устойчивости. Получим сейчас более удобные формулы для их вычисления.

Лемма 6.1 (тождество Кристоффеля-Дарбу). Для системы ортонормированных многочленов $Q_i(x)$, $i = 0, 1, \dots, n, \dots$ справедливо тождество

$$\sum_{i=0}^n Q_i(t) Q_i(x) = a_{n,n+1} \frac{Q_{n+1}(t) \cdot Q_n(x) - Q_{n+1}(x) \cdot Q_n(t)}{t - x}. \quad (6.45)$$

Тождество Кристоффеля–Дарбу

Ранее мы получили трехчленное рекуррентное соотношение, связывающее ортонормированные многочлены:

$$a_{i,i+1} Q_{i+1}(x) + a_{i,i} Q_i(x) + a_{i-1,i} Q_{i-1}(x) = x Q_i(x), \quad (*)$$

где

$$a_{i,k} = \int_a^b p(x) Q_i(x) Q_k(x) dx.$$

Умножим соотношение (*) на $Q_i(t)$. Получим:

$$x Q_i(x) \cdot Q_i(t) = a_{i,i+1} Q_{i+1}(x) \cdot Q_i(t) + a_{i,i} Q_i(x) \cdot Q_i(t) + a_{i-1,i} Q_{i-1}(x) \cdot Q_i(t). \quad (**)$$

Поменяем в последнем равенстве ролями переменные x и t :

$$t Q_i(t) \cdot Q_i(x) = a_{i,i+1} Q_{i+1}(t) \cdot Q_i(x) + a_{i,i} Q_i(t) \cdot Q_i(x) + a_{i-1,i} Q_{i-1}(t) \cdot Q_i(x).$$

Теперь, вычитая из последнего равенства равенство (**), получим:

$$(t - x) Q_i(t) \cdot Q_i(x) = a_{i,i+1} [Q_{i+1}(t) \cdot Q_i(x) - Q_i(t) \cdot Q_{i+1}(x)] - \\ - a_{i-1,i} [Q_i(t) \cdot Q_{i-1}(x) - Q_i(x) \cdot Q_{i-1}(t)].$$



Просуммировав полученное равенство по i от 0 до n , будем иметь:

$$(t - x) \sum_{i=0}^n Q_i(x) Q_i(t) = a_{n,n+1} [Q_{n+1}(t) \cdot Q_n(x) - Q_{n+1}(x) \cdot Q_n(t)],$$

или

$$\sum_{i=0}^n Q_i(t) Q_i(x) = a_{n,n+1} \frac{Q_{n+1}(t) \cdot Q_n(x) - Q_{n+1}(x) \cdot Q_n(t)}{t - x}.$$

Заметим, что если записать $Q_n(x)$ в виде

$$Q_n(x) = c_n x^n + \dots,$$

то равенство (*) примет вид

$$\begin{aligned} a_{i,i+1} (c_{i+1} x^{i+1} + \dots) + a_{i,i} (c_i x^i + \dots) + a_{i-1,i} (c_i x^{i-1} + \dots) \\ = x (c_i x^i + \dots) \end{aligned}$$

Отсюда, приравнивая коэффициенты при x^{i+1} , получим:

$$a_{i,i+1} c_{i+1} = c_i,$$

т.е.

$$a_{i,i+1} = \frac{c_i}{c_{i+1}}, \quad i = 0, 1, \dots$$

С учетом полученного соотношения тождество (Д.8) можно переписать в виде

$$\sum_{i=0}^n Q_i(t) Q_i(x) = \frac{c_n}{c_{n+1}} \frac{Q_{n+1}(t) \cdot Q_n(x) - Q_{n+1}(x) \cdot Q_n(t)}{t - x}. \quad (6.46)$$

Теперь преобразуем формулу (6.43) для вычисления коэффициентов квадратурных формул наивысшей алгебраической степени точности, умножив ее числитель и знаменатель на c_{n+1} и учитывая, что $c_{n+1}\omega_{n+1}(x) = Q_{n+1}(x)$:

$$A_k = \int_a^b p(x) \frac{\omega_{n+1}(x)}{(x - x_k) \omega'_{n+1}(x_k)} dx = \int_a^b p(x) \frac{Q_{n+1}(x)}{(x - x_k) Q'_{n+1}(x_k)} dx. \quad (6.47)$$

Положим в (6.46) $t = x_k$ и учтем, что x_k – корень многочлена $Q_{n+1}(x)$. Тогда $Q_{n+1}(x_k) = 0$ и (6.46) перепишется в виде

$$\sum_{i=0}^n Q_i(x) Q_i(x_k) = \frac{c_n}{c_{n+1}} \cdot Q_n(x_k) \cdot \frac{Q_{n+1}(x)}{x - x_k}.$$

Умножим последнее равенство на $p(x)$ и проинтегрируем по отрезку $[a, b]$:

$$\sum_{i=0}^n Q_i(x_k) \int_a^b p(x) Q_i(x) dx = \frac{c_n}{c_{n+1}} \cdot Q_n(x_k) \cdot \int_a^b p(x) \frac{Q_{n+1}(x)}{x - x_k} dx, \quad (***)$$

Так как система многочленов $Q_i(x)$ является ортонормированной, то

$$\int_a^b p(x) Q_i(x) dx = \begin{cases} 0, & \text{если } i \geq 1, \\ 1, & \text{если } i = 0. \end{cases}$$

Следовательно, равенство (***) с учетом (6.47) примет вид

$$1 = \frac{c_n}{c_{n+1}} Q_n(x_k) \int_a^b p(x) \frac{Q_{n+1}(x)}{x - x_k} dx = \frac{c_n}{c_{n+1}} \cdot Q_n(x_k) \cdot Q'_{n+1}(x_k) \cdot A_k.$$

Отсюда

$$A_k = \frac{c_{n+1}}{c_n} \cdot \frac{1}{Q_n(x_k) \cdot Q'_{n+1}(x_k)}, \quad k = \overline{0, n}. \quad (6.48)$$



Формула (6.48) несколько более удобна для вычисления коэффициентов квадратурных формул типа Гаусса по сравнению с (6.43), поскольку не требует вычисления интегралов (однако требует знания систем ортонормированных многочленов).

Заметим, что в формуле (6.48) можно осуществить обратный переход от системы ортонормированных многочленов $Q_n(x)$ к приведенным ортогональным многочленам $\omega_n(x)$. Действительно, учитывая, что $c_k \omega_k(x) = Q_k(x)$, умножим числитель и знаменатель (6.48) на c_{n+1} . Тогда

$$A_k = \frac{1}{c_n^2 \cdot \omega_n(x_k) \cdot \omega'_{n+1}(x_k)}, \quad k = \overline{0, n},$$

а поскольку

$$1 = \|Q_n(x)\|^2 = c_n^2 \|\omega_n(x)\|^2,$$

то

$$A_k = \frac{\|\omega_n(x)\|^2}{\omega_n(x_k) \cdot \omega'_{n+1}(x_k)}, \quad k = \overline{0, n}. \quad (6.49)$$

Остаток квадратурных формул типа Гаусса

Наконец, установим формулу для вычисления остатка квадратурных формул типа Гаусса. Построим для функции $f(x)$ интерполяционный многочлен Эрмита степени не выше $2n + 1$ с двукратными узлами x_0, x_1, \dots, x_n :

$$f(x) = P_{2n+1}(x) + r_{2n+1}(x),$$

где

$$r_{2n+1}(x) = \omega_{n+1}^2(x) \cdot \frac{f^{(2n+2)}(\xi)}{(2n+2)!},$$

если $f(x) \in C^{2n+2}[a, b]$.

Тогда

$$\int_a^b p(x) f(x) dx = \int_a^b p(x) P_{2n+1}(x) dx + \int_a^b p(x) r_{2n+1}(x) dx.$$



С другой стороны, так как алгебраическая степень точности формулы равна $2n + 1$, то

$$\int_a^b p(x) P_{2n+1}(x) dx = \sum_{k=0}^n A_k P_{2n+1}(x_k) = \sum_{k=0}^n A_k f(x_k).$$

Следовательно,

$$\begin{aligned} R_n(f) &= \int_a^b p(x) r_{2n+1}(x) dx = \int_a^b p(x) \omega_{n+1}^2(x) \frac{f^{(2n+2)}(\xi)}{(2n+2)!} dx \stackrel{\text{т. о среднем}}{=} \\ &\stackrel{\text{т. о среднем}}{=} \frac{f^{(2n+2)}(\eta)}{(2n+2)!} \int_a^b p(x) \omega_{n+1}^2(x) dx = \frac{f^{(2n+2)}(\eta)}{(2n+2)!} \cdot \|\omega_{n+1}(x)\|^2. \quad (6.50) \end{aligned}$$

Некоторые частные случаи квадратурных формул типа Гаусса

Рассмотрим сейчас некоторые наиболее часто встречающиеся в приложениях случаи квадратурных формул типа Гаусса.

Формулы Гаусса (Гаусса–Лежандра). Пусть $p(x) \equiv 1$. Отрезок интегрирования считаем конечным, а $f(x)$ – достаточно гладкой функцией (именно этот случай был подробно рассмотрен Гауссом).

Всякий конечный отрезок $[a, b]$ линейной заменой переменной может быть преобразован в отрезок $[-1, 1]$, и мы будем считать, что интеграл приведен к виду

$$I = \int_{-1}^1 f(x) dx. \quad (*)$$



Построим в явном виде систему ортогональных многочленов. Пусть $\omega_n(x)$ – n -й приведенный многочлен данной системы, а $q(x)$ – произвольный многочлен степени не выше $n - 1$. Введем обозначения

$$\varphi_1(x) = \int_{-1}^x \omega_n(x) dx, \quad \varphi_{i+1}(x) = \int_{-1}^x \varphi_i(x) dx, \quad i = 1, \dots, n-1. \quad (**)$$

Тогда, последовательно интегрируя по частям, будем иметь:

$$0 = \int_{-1}^1 \omega_n(x) q(x) dx = \varphi_1(x) q(x) \Big|_{-1}^1 - \int_{-1}^1 \varphi_1(x) q'(x) dx =$$

$$= [\varphi_1(x) q(x) - \varphi_2(x) q'(x)] \Big|_{-1}^1 + \int_{-1}^1 \varphi_2(x) q''(x) dx = \dots =$$

$$= \left[\varphi_1(x) q(x) - \varphi_2(x) q'(x) + \dots + (-1)^{n-1} \varphi_n(x) q^{(n-1)}(x) \right] \Big|_{-1}^1 +$$

$$+ (-1)^n \int_{-1}^1 \varphi_n(x) q^{(n)}(x) dx. \quad (6.51)$$

Интегральное слагаемое в правой части полученного равенства, очевидно, равно нулю, поскольку согласно предположению $q(x)$ – многочлен степени не выше $n - 1$. Точно так же равны нулю и все оставшиеся



слагаемые на нижнем пределе двойной подстановки, так как в силу равенств (***) $\varphi_i(-1) = 0$, $i = \overline{1, n}$. Отсюда в силу произвольности многочлена $q(x)$ следует, что

$$\varphi_i(1) = 0, \quad i = \overline{1, n}.$$

Таким образом, многочлен степени $2n$ $\varphi_n(x)$ обладает корнями кратности n при $x = \pm 1$ и, значит,

$$\varphi_n(x) = C(x+1)^n(x-1)^n = C(x^2 - 1)^n,$$

где C – некоторая постоянная (старший коэффициент). Тогда

$$\omega_n(x) = C \frac{d^n}{dx^n} (x^2 - 1)^n.$$

Постоянную C подбираем таким образом, чтобы многочлен $\omega_n(x)$ был равен единице. Так как

$$\omega_n(x) = C \frac{d^n}{dx^n} (x^2 - 1)^n = C \frac{d^n}{dx^n} (x^{2n} - \dots) =$$

$$= C \cdot (2n) \cdot (2n-1) \cdot \dots \cdot (n+1) (x^n - \dots) = C \cdot \frac{(2n)!}{n!} \cdot (x^n - \dots),$$

то

$$C = \frac{n!}{(2n)!},$$

т.е.

$$\omega_n(x) = \frac{n!}{(2n)!} \frac{d^n}{dx^n} (x^2 - 1)^n, \quad n = 0, 1, \dots \quad (6.52)$$

Вычислим сейчас норму многочлена $\omega_n(x)$, применяя, как и выше, последовательное интегрирование по частям:

$$\int_{-1}^1 \omega_n^2(x) dx = C^2 \int_{-1}^1 \frac{d^n}{dx^n} (x^2 - 1)^n \cdot \frac{d^n}{dx^n} (x^2 - 1)^n dx =$$



$$\begin{aligned}
 &= C^2 \left[\frac{d^n}{dx^n} (x^2 - 1)^n \cdot \frac{d^{n-1}}{dx^{n-1}} (x^2 - 1)^n \right] \Big|_{-1}^1 - \int_{-1}^1 \frac{d^{n-1}}{dx^{n-1}} (x^2 - 1)^n \cdot \frac{d^{n+1}}{dx^{n+1}} (x^2 - 1)^n dx = \\
 &= C^2 \left[\frac{d^n}{dx^n} (x^2 - 1)^n \cdot \frac{d^{n-1}}{dx^{n-1}} (x^2 - 1)^n - \frac{d^{n+1}}{dx^{n+1}} (x^2 - 1)^n \cdot \frac{d^{n-2}}{dx^{n-2}} (x^2 - 1)^n \right] \Big|_{-1}^1 + \\
 &\quad + \int_{-1}^1 \frac{d^{n-2}}{dx^{n-2}} (x^2 - 1)^n \cdot \frac{d^{n+2}}{dx^{n+2}} (x^2 - 1)^n dx = \dots = \\
 &= C^2 \left[\frac{d^n}{dx^n} (x^2 - 1)^n \cdot \frac{d^{n-1}}{dx^{n-1}} (x^2 - 1)^n - \frac{d^{n+1}}{dx^{n+1}} (x^2 - 1)^n \cdot \right. \\
 &\quad \left. \cdot \frac{d^{n-2}}{dx^{n-2}} (x^2 - 1)^n + \dots + (-1)^{n-1} \frac{d^{2n-1}}{dx^{2n-1}} (x^2 - 1)^n \cdot (x^2 - 1)^n \right] \Big|_{-1}^1 + \\
 &\quad + (-1)^n \int_{-1}^1 \frac{d^{2n}}{dx^{2n}} (x^2 - 1)^n \cdot (x^2 - 1)^n dx = (-1)^n C^2 \cdot (2n)! \cdot \\
 &\quad \cdot \int_{-1}^1 (x - 1)^n (x + 1)^n dx = (-1)^n C^2 \cdot (2n)! \cdot \left[\frac{(x - 1)^n (x + 1)^{n+1}}{n + 1} \right] \Big|_{-1}^1 - \\
 &\quad - \frac{n}{n + 1} \int_{-1}^1 (x - 1)^{n-1} (x + 1)^{n+1} dx] = \\
 &= (-1)^n C^2 \cdot (2n)! \cdot \left[\frac{(x - 1)^n (x + 1)^{n+1}}{n + 1} - \frac{n}{n + 1} \frac{(x - 1)^{n-1} (x + 1)^{n+2}}{n + 2} \right] \Big|_{-1}^1 +
 \end{aligned}$$



$$\begin{aligned}
 & + \frac{n(n-1)}{(n+1)(n+2)} \int_{-1}^1 (x-1)^{n-2} (x+1)^{n+2} dx] = (-1)^n C^2 \cdot (2n)! \cdot \\
 & \cdot \left[\frac{(x-1)^n (x+1)^{n+1}}{n+1} - \frac{n}{n+1} \frac{(x-1)^{n-1} (x+1)^{n+2}}{n+2} + \dots + \right. \\
 & \left. + (-1)^{n-1} \frac{n! \cdot (x-1) (x+1)^{2n}}{(n+1) \cdot \dots \cdot (2n)} \right]_{-1}^1 + \\
 & + (-1)^n C^2 \cdot (2n)! \cdot \frac{n!}{(n+1) \cdot \dots \cdot (2n)} \int_{-1}^1 (x+1)^{2n} dx = \\
 & = (-1)^n C^2 \cdot (2n)! \cdot \frac{n! \cdot n!}{(2n)!} \cdot \frac{(x+1)^{2n+1}}{2n+1} \Big|_{-1}^1 = \frac{(n!)^2}{((2n)!)^2} \cdot (n!)^2 \cdot \frac{2^{2n+1}}{2n+1}.
 \end{aligned}$$

Таким образом,

$$\|\omega_n(x)\| = \frac{(n!)^2 \cdot 2^n}{(2n)!} \cdot \sqrt{\frac{2}{2n+1}} \quad (6.53)$$

и, следовательно, многочлены $Q_n(x)$ ортонормированной системы будут иметь вид

$$Q_n(x) = \frac{\omega_n(x)}{\|\omega_n(x)\|} = \frac{n!}{(2n)!} \cdot \sqrt{\frac{2n+1}{2}} \cdot \frac{(2n)!}{(n!)^2 \cdot 2^n} \omega_n(x) = \sqrt{\frac{2n+1}{2}} \cdot \frac{1}{n! \cdot 2^n} \frac{d^n}{dx^n} (x^2 - 1)^n. \quad (6.54)$$

Окончательно можем сформулировать следующий результат: квадратурная формула наивысшей алгебраической степени точности для вычисления интеграла (*) ([квадратурная формула Гаусса–Лежандра](#)) имеет вид

$$I = \int_{-1}^1 f(x) dx \approx \sum_{k=0}^n A_k f(x_k) \quad (6.55)$$



где x_k – корни многочлена степени $(n + 1)$, определяемого формулой (6.52), а коэффициенты A_k могут быть вычислены по формулам (6.48) или (6.49), в которых многочлены определяются по формулам (6.54) или (6.52) соответственно.

При этом, учитывая формулы (6.50) и (6.53), можем записать представление ее остатка:

$$\begin{aligned} R_n(f) &= \frac{f^{(2n+2)}(\eta)}{(2n+2)!} \cdot \| \omega_{n+1}(x) \|^2 = \frac{f^{(2n+2)}(\eta)}{(2n+2)!} \cdot \frac{((n+1)!)^4}{((2n+2)!)^2} \cdot \frac{2^{2n+3}}{2n+3} = \\ &= \frac{f^{(2n+2)}(\eta) \cdot 2^{2n+3}}{(2n+2)! \cdot (2n+3)} \cdot \left(\frac{((n+1)!)^2}{(2n+2)!} \right)^2. \end{aligned} \quad (6.56)$$

Замечание 6.3. В теории ортогональных многочленов чаще вместо системы многочленов $\omega_n(x)$ (или $Q_n(x)$), определенных выше, используют многочлены, отличающиеся от них постоянным множителем и имеющие вид

$$P_n(x) = \frac{1}{2^n \cdot n!} \frac{d^n}{dx^n} (x^2 - 1)^n \quad (6.57)$$

норма которых равна $\| P_n(x) \| = \sqrt{\frac{2n+1}{2}}$. Эти многочлены называются **многочленами Лежандра**. В терминах многочленов Лежандра формулу для вычисления квадратурных коэффициентов можно записать в следующем виде

$$A_k = \frac{2}{(n+1) P_n(x_k) P'_{n+1}(x_k)}, \quad k = \overline{0, n}.$$

Если же при этом воспользоваться известным в теории многочленов Лежандра соотношением

$$(1 - x^2) P'_n(x) = n [P_{n-1}(x) - x P_n(x)],$$

то формулы для вычисления коэффициентов можно еще более упростить и привести к виду

$$A_k = \frac{2}{(1 - x_k^2) [P'_{n+1}(x_k)]^2}, \quad k = \overline{0, n}. \quad (6.58)$$

Замечание 6.4. На «пользовательском» уровне узлы и коэффициенты квадратурной формулы Гаусса удобнее брать из соответствующих таблиц.



Формулы Гаусса–Якоби. Рассмотрим $p(x) = (b - x)^\alpha (x - a)^\beta$. Как видно из записи интеграла с указанной весовой функцией, соответствующая квадратурная формула предназначена для приближенного интегрирования функций, на концах отрезка интегрирования имеющих степенные особенности.

Вновь, как и в предыдущем случае, можно ограничиться рассмотрением интеграла

$$I = \int_{-1}^1 (1 - x)^\alpha (1 + x)^\beta f(x) dx.$$

Здесь α и β – вещественные параметры, причем $\alpha, \beta > -1$ (последнее требование необходимо для сходимости интеграла).

Практически дословно повторяя рассуждения предыдущего пункта, можно было бы построить систему многочленов, ортогональных по весу $p(x) = (1 - x)^\alpha (1 + x)^\beta$ на отрезке $[-1, 1]$. Однако мы этого больше делать не будем и воспользуемся готовыми результатами. Искомой системой многочленов является система **многочленов Якоби**

$$P_n^{(\alpha, \beta)}(x) = \frac{(-1)^n}{2^n \cdot n!} \cdot (1 - x)^{-\alpha} (1 + x)^{-\beta} \frac{d^n}{dx^n} [(1 - x)^{\alpha+n} (1 + x)^{\beta+n}] \quad (6.59)$$

Таким образом, *квадратурная формула Гаусса–Якоби* будет иметь вид

$$I = \int_{-1}^1 (1 - x)^\alpha (1 + x)^\beta f(x) dx \approx \sum_{k=0}^n A_k f(x_k), \quad (6.60)$$

где узлы x_k будут корнями многочленов $P_{n+1}^{(\alpha, \beta)}(x)$, для коэффициентов A_k можно получить формулы

$$A_k = 2^{\alpha+\beta+1} \cdot \frac{\Gamma(\alpha + n + 2) \cdot \Gamma(\beta + n + 2)}{(n + 1)! \cdot \Gamma(\alpha + \beta + n + 2) \cdot (1 - x_k^2) \cdot \left[\frac{d}{dx} P_{n+1}^{(\alpha, \beta)}(x_k) \right]^2}, \quad k = \overline{0, n}, \quad (6.61)$$

а для остатка справедливо представление

$$R_n(f) = \frac{f^{(2n+2)}(\eta)}{(2n + 2)!} \cdot 2^{\alpha+\beta+2n+1}.$$



$$\cdot \frac{(n+1)! \cdot \Gamma(\alpha + n + 2) \cdot \Gamma(\beta + n + 2) \cdot \Gamma(\alpha + \beta + n + 2)}{(\alpha + \beta + 2n + 3) \cdot \Gamma(\alpha + \beta + 2n + 3)}. \quad (6.62)$$

В формулах (6.61), (6.62) $\Gamma(x)$ обозначает Γ -функцию Эйлера.

Частными случаями многочленов Якоби являются рассмотренные выше [многочлены Лежандра](#) (соответствуют случаю $\alpha = \beta = 0$), а также многочлены – наименее отклоняющиеся от нуля – [многочлены Чебышева первого рода](#) $T_n(x)$ (при $\alpha = \beta = -\frac{1}{2}$). В этом случае квадратурная формула наивысшей алгебраической степени точности выглядит особо просто. Поэтому рассмотрим ее несколько подробнее.

Ее узлы – корни многочлена Чебышева $T_{n+1}(x)$ – имеют вид:

$$x_k = \cos \frac{2k+1}{2n+2} \pi, \quad k = \overline{0, n}. \quad (6.63)$$

Помимо этого, квадратурная формула наивысшей алгебраической степени точности, соответствующая весовой функции $p(x) = \frac{1}{\sqrt{1-x^2}}$ обладает еще одним замечательным свойством. Пользуясь формулой (6.61), вычислим ее коэффициенты:

$$A_k = -\frac{\Gamma^2(n + \frac{3}{2})}{(n+1)! \cdot \Gamma(n+1) \cdot (1-x_k^2) \cdot \left[\frac{d}{dx} P_{n+1}^{(-\frac{1}{2}, -\frac{1}{2})}(x_k) \right]^2}.$$

Поскольку $P_{n+1}^{(-\frac{1}{2}, -\frac{1}{2})}(x) = C_{n+1} T_{n+1}(x)$ и

$$T'_{n+1}(x_k) = [\cos((n+1)\arccos x_k)]' \Big|_{x=x_k} = \sin[(n+1)\arccos x_k] \cdot \frac{n+1}{\sqrt{1-x_k^2}} = \frac{(-1)^k (n+1)}{\sqrt{1-x_k^2}},$$

то

$$(1-x_k^2) \left[\frac{d}{dx} P_{n+1}^{(-\frac{1}{2}, -\frac{1}{2})}(x_k) \right]^2 = C_{n+1}^2 (n+1)^2$$



и, следовательно,

$$A_k = \frac{\Gamma^2(n + \frac{3}{2})}{(n+1)! \cdot \Gamma(n+1) \cdot C_{n+1}^2 \cdot (n+1)^2}, \quad k = \overline{0, n}.$$

Правая часть полученного равенства не зависит от номера k и поэтому все коэффициенты A_k будут одинаковы. Обозначим общую величину их буквой A . Численное значение A , конечно же, может быть найдено на основании последнего равенства, но проще это сделать, если воспользоваться тем, что квадратурная формула должна дать точный результат в случае $f(x) \equiv 1$ (т.е. иметь [алгебраическую степень точности](#) не менее нуля) и, стало быть, должно выполняться равенство

$$\sum_{k=0}^n A_k = (n+1)A = \int_{-1}^1 \frac{dx}{\sqrt{1-x^2}} = \pi.$$

Отсюда имеем:

$$A = \frac{\pi}{n+1}.$$

Таким образом, окончательно получаем: рассматриваемая квадратурная формула ([формула Гаусса–Чебышева](#)) имеет вид

$$I = \int_{-1}^1 \frac{f(x)}{\sqrt{1-x^2}} dx \approx \frac{\pi}{n+1} \sum_{k=0}^n f\left(\cos \frac{2k+1}{2n+2}\pi\right), \quad (6.64)$$

а ее остаток после несложных преобразований формулы (6.62) с использованием свойств функции Эйлера примет вид

$$R_n(f) = \frac{\pi}{2^{2n+1} \cdot (2n+2)!} f^{(2n+2)}(\eta). \quad (6.65)$$

Формулы Гаусса–Лагерра. Рассмотрим интегралы вида

$$I = \int_0^\infty x^\alpha e^{-x} f(x) dx.$$



Ортогональными на полуоси $[0, \infty)$ по весу $p(x) = x^\alpha e^{-x}$ ($\alpha > -1$) являются **многочлены Чебышева–Лагерра**

$$L_n^{(\alpha)}(x) = (-1)^n x^{-\alpha} e^x \frac{d^n}{dx^n} (x^{\alpha+n} e^{-x}) \quad (6.66)$$

Таким образом, узлы **квадратурной формулы Гаусса–Лагерра** будут корнями многочлена вида (6.66) $L_{n+1}^{(\alpha)}(x)$, коэффициенты будут вычисляться по формулам

$$A_k = \frac{\Gamma(n+2) \cdot \Gamma(n+\alpha+2)}{x_k \left[\frac{d}{dx} L_{n+1}^{(\alpha)}(x) \right]^2}, \quad k = \overline{0, n}, \quad (6.67)$$

а остаток будет иметь вид

$$R_n(f) = \frac{\Gamma(n+2) \cdot \Gamma(n+\alpha+2)}{(2n+2)!} \cdot f^{(2n+2)}(\eta). \quad (6.68)$$

Формулы Гаусса–Эрмита. Эти формулы являются квадратурными формулами наивысшей алгебраической степени точности для вычисления интегралов вида

$$I = \int_{-\infty}^{\infty} e^{-x^2} f(x) dx.$$

Систему многочленов, ортогональных на всей числовой оси $(-\infty, +\infty)$ по весу $p(x) = e^{-x^2}$, образуют **многочлены Чебышева–Эрмита**

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2}. \quad (6.69)$$

Значит, узлы **квадратурной формулы Гаусса–Эрмита** являются корнями многочлена $H_{n+1}(x)$, коэффициенты могут быть вычислены по формуле

$$A_k = \frac{2^{n+1} \cdot (n+1)! \cdot \sqrt{\pi}}{\left[\frac{d}{dx} H_{n+1}^2(x_k) \right]^2}, \quad k = \overline{0, n}, \quad (6.70)$$



а остаток имеет вид

$$R_n(f) = \frac{(n+1)! \cdot \sqrt{\pi}}{2^{n+1} \cdot (2n+2)!} \cdot f^{(2n+2)}(\eta). \quad (6.71)$$

Еще раз напомним, что для всех рассмотренных выше случаев (а также и для многих других) составлены таблицы узлов и коэффициентов для различных значений n .



6.1.7. Квадратурные формулы, содержащие наперед заданные узлы

Формулы с предписанными узлами частного вида (формулы типа Маркова)

В прикладных задачах иногда возникает необходимость построения таких квадратурных формул, часть узлов которых задается заранее, другая же часть может быть взята произвольно и выбором их можно распоряжаться для тех или иных целей.

Рассмотрим квадратурную формулу

$$\int_a^b p(x) f(x) dx \approx \sum_{l=1}^m B_l f(a_l) + \sum_{k=0}^{n-m} A_k f(x_k) \quad (6.72)$$

в которой m узлов a_1, \dots, a_m ($m \leq n+1$) фиксированы. Она содержит $2n - m + 2$ параметров A_k , x_k ($k = 0, \dots, n-m$) и B_l ($l = 1, \dots, m$). Попытаемся их выбрать так, чтобы равенство (6.72) было точным для многочленов возможно более высокой степени.

Введем два многочлена, связанных с узлами a_l и x_k :

$$\Omega_m(x) = (x - a_1) \cdot \dots \cdot (x - a_m) ,$$

$$\omega_{n-m+1}(x) = (x - x_0) \cdot \dots \cdot (x - x_{n-m}) .$$

За счет выбора коэффициентов A_k и B_l формулу (6.72) можно сделать точной для многочленов степени n . Для этого ее достаточно сделать [интерполяционной](#). Достичь же того, чтобы (6.72) была точной для многочленов более высокой степени, можно только за счет специального подбора узлов x_k .

Справедлива теорема, аналогичная [теореме 6.2](#) (причем смысл ее состоит в том, чтобы многочлен $\Omega_m(x)$ в «состав» весовой функции).

Теорема 6.6. Для того чтобы квадратурная формула (6.72) была точной для многочленов степени $2n - m + 1$, необходимо и достаточно, чтобы выполнялись следующие условия:



1) она была интерполяционной, т.е.

$$A_k = \int_a^b p(x) \frac{\omega_{n-m+1}(x) \Omega_m(x)}{(x - x_k) \omega'_{n-m+1}(x_k) \Omega_m(x_k)} dx, \quad k = 0, \dots, n-m, \quad (6.73)$$

$$B_l = \int_a^b p(x) \frac{\omega_{n-m+1}(x) \Omega_m(x)}{(x - a_l) \omega_{n-m+1}(a_l) \Omega'_m(a_l)} dx, \quad l = 1, \dots, m;$$

2) многочлен $\omega_{n-m+1}(x)$ был ортогонален на отрезке $[a, b]$ по весу $p(x) \Omega_m(x)$ ко всем многочленам степени не выше $n - m$, т.е.

$$\int_a^b p(x) \Omega_m(x) Q_p(x) \omega_{n-m+1}(x) dx = 0, \quad p \leq n - m. \quad (6.74)$$

[\[Доказательство\]](#)

Далее будем предполагать, что многочлен $\omega_{n-m+1}(x)$ существует, т.е. квадратурная формула вида (6.72), обладающая алгебраической степенью точности, равной $2n - m + 1$, может быть построена.

Получим в этом случае представление остатка. Для этого, как и выше, выполним интерполирование функции $f(x)$ на отрезке $[a, b]$ по следующим условиям:

$$P_{2n-m+1}(a_l) = f(a_l), \quad l = 1, \dots, m,$$

$$P_{2n-m+1}(x_k) = f(x_k), \quad P'_{2n-m+1}(x_k) = f'(x_k), \quad k = 0, \dots, n - m.$$

Тогда остаток такого интерполирования может быть представлен в виде

$$r(x) = \Omega_m(x) \omega_{n-m+1}^2(x) \frac{f^{(2n-m+2)}(\xi)}{(2n-m+2)!}, \quad \xi \in [a, b].$$



Интегрируя равенство $f(x) = P_{2n-m+1}(x) + r(x)$, получим квадратурную формулу (4.1) и

$$R_n(f) = \int_a^b p(x) r(x) dx = \int_a^b p(x) \Omega_m(x) \omega_{n-m+1}^2(x) \frac{f^{(2n-m+2)}(\xi)}{(2n-m+2)!} dx. \quad (6.75)$$

Точно таким же образом, как и в предыдущем параграфе, можно получить формулы для вычисления коэффициентов A_k , не требующие вычисления интегралов:

$$A_k = \frac{c_{n-m+1}}{c_{n-m}} \cdot \frac{1}{\Pi_{n-m}(x_k) \cdot \Pi'_{n-m+1}(x_k) \cdot \Omega_m(x_k)}, \quad k = 0, \dots, n-m, \quad (6.76)$$

где $\{\Pi_i(x)\}$ – система многочленов, ортонормированная на отрезке $[a, b]$ по весу $p(x) \Omega_m(x)$.

Таким образом, мы вторично сталкиваемся с необходимостью находить систему многочленов, ортогональных на отрезке $[a, b]$ по весу $p(x) \Omega_m(x)$. В некоторых случаях здесь может оказаться полезной теорема о преобразовании ортогональной системы многочленов при умножении веса на многочлен.

Теорема 6.7. Пусть $\Omega_m(x) = (x - a_1) \cdot \dots \cdot (x - a_m)$ и существуют и единственны системы приведенных многочленов $\tilde{P}_s(x)$ и $\tilde{\Pi}_s(x)$, ортогональные на отрезке $[a, b]$ по весу $p(x)$ и $p(x) \Omega_m(x)$ соответственно. Тогда

$$\tilde{\Pi}_{n-m}(x) = \frac{1}{\Delta \cdot \Omega_m(x)} \cdot \begin{vmatrix} \tilde{P}_n(x) & \tilde{P}_n(a_1) & \dots & \tilde{P}_n(a_m) \\ \tilde{P}_{n-1}(x) & \tilde{P}_{n-1}(a_1) & \dots & \tilde{P}_{n-1}(a_m) \\ \vdots & \vdots & \dots & \vdots \\ \tilde{P}_{n-m}(x) & \tilde{P}_{n-m}(a_1) & \dots & \tilde{P}_{n-m}(a_m) \end{vmatrix}, \quad (6.77)$$

где

$$\Delta = \begin{vmatrix} \tilde{P}_{n-1}(a_1) & \dots & \tilde{P}_{n-1}(a_m) \\ \vdots & \dots & \vdots \\ \tilde{P}_{n-m}(a_1) & \dots & \tilde{P}_{n-m}(a_m) \end{vmatrix}.$$


[\[Доказательство\]](#)

Представляя данный определитель в виде суммы двух определителей, а затем разлагая один из них по первому столбцу, получим (6.77).

Формулы с предписанными узлами частного вида (формулы типа Маркова)

Ограничимся здесь рассмотрением следующих случаев квадратурных формул типа (6.72):

- 1) $m = 1$ и $a_1 = a$;
- 2) $m = 1$ и $a_1 = b$;
- 3) $m = 2$ и $a_1 = a$, $a_2 = b$.

Во всех этих случаях при знакопостоянной весовой функции $p(x)$ произведение $p(x)\Omega_m(x)$ также знакопостоянно на отрезке $[a, b]$ и, следовательно, квадратурные формулы, имеющие алгебраическую степень точности, равную $2n - m + 1$, всегда могут быть построены (см. теорему 6.3).

Случай 1 и 2. Итак, пусть имеем случай 1 (случай 2 сводится к нему линейной заменой переменной интегрирования $x = a + b - t$ и отдельно рассматриваться не будет).

Тогда

$$I = \int_a^b p(x)f(x)dx = Bf(a) + \sum_{k=0}^{n-1} A_k f(x_k) + R(f) \quad (6.78)$$

где

$$A_k = \frac{c_n}{c_{n-1}} \cdot \frac{1}{b(x_k-a)\Pi_{n-1}(x_k)\Pi'_n(x_k)}, \quad k = 0, 1, \dots, n-1, \quad (6.79)$$

$$B = \frac{1}{\Pi(a)} \int_a^b p(x)\Pi_n(x)dx.$$

При этом оказывается, что все коэффициенты положительны в случае положительной весовой функции.



Алгебраическая степень точности квадратурной формулы (6.78) равна $2n$, а ее остаток имеет вид

$$R(f) = \frac{f^{(2n+1)}(\eta)}{(2n+1)!} \cdot \int_a^b p(x)(x-a)\omega_n^2(x)dx. \quad (6.80)$$

Рассмотрим сейчас несколько подробнее случай $p(x) \equiv 1$. Как и ранее, отрезок $[a, b]$ приведем к отрезку $[-1, 1]$. Тогда квадратурная формула (6.78) примет вид

$$I = \int_{-1}^1 f(x)dx = Bf(-1) + \sum_{k=0}^{n-1} A_k f(x_k) + R(f). \quad (6.81)$$

При этом $\Omega_1(x) = 1+x$ и, следовательно, многочлен $\omega_n(x)$ будет только коэффициентом отличаться от многочлена Якоби $P_n^{(0,1)}(x)$, а коэффициенты A_k только множителем $\frac{1}{\Omega_1(x_k)}$ будут отличаться от соответствующих коэффициентов формулы наивысшей алгебраической степени точности:

$$A_k = \frac{4}{(1+x_k)(1-x_k^2) \left[\frac{d}{dx} P_n^{(0,1)}(x_k) \right]^2}, \quad k = 0, 1, \dots, n-1, \quad (6.82)$$

$$B = \int_{-1}^1 \frac{\omega_n(x)}{\omega_n(-1)} dx = \frac{1}{P_n^{(0,1)}(-1)} \int_{-1}^1 P_n^{(0,1)}(x) dx = \frac{2}{(n+1)^2},$$

$$R(f) = \frac{2}{n+1} \left[\frac{2^n \cdot n! \cdot (n+1)!}{(2n+1)!} \right]^2 \cdot \frac{f^{(2n+1)}(\eta)}{(2n+1)!} \quad (6.83)$$

Случай 3. В этом случае имеем $\Omega_2(x) = (x-a)(b-x)$.



В соответствии с теоремой 6.7 имеем формулу для вычисления системы многочленов, ортогональных по весу $p(x)\Omega_2(x)$:

$$\Pi_{n-2}(x) = \frac{K_{n-2}}{(x-a)(x-b)} \begin{vmatrix} P_n(x) & P_n(a) & P_n(b) \\ P_{n-1}(x) & P_{n-1}(a) & P_{n-1}(b) \\ P_{n-2}(x) & P_{n-2}(a) & P_{n-2}(b) \end{vmatrix},$$

где $P_n(x)$ – многочлены, образующие ортогональную систему по весу $p(x)$, а K_{n-2} – некоторая константа.
Все коэффициенты квадратурной формулы

$$\int_a^b p(x) f(x) dx \approx B_1 f(a) + B_2 f(b) + \sum_{k=0}^{n-2} A_k f(x_k)$$

знакопостоянны (положительном весе $p(x)$),

$$R(f) = \frac{f^{(2n)}(\eta)}{(2n)!} \cdot \int_a^b p(x)(x-a)(x-b)\omega_{n-1}^2(x) dx.$$

Вновь, как и выше, рассмотрим несколько подробнее случай $p(x) \equiv 1$ и $[a, b] = [-1, 1]$. В этом случае квадратурная формула примет вид

$$I = \int_{-1}^1 f(x) dx = B_1 f(-1) + B_2 f(1) + \sum_{k=0}^{n-2} A_k f(x_k) + R(f). \quad (6.84)$$

При этом, поскольку $\Omega_2(x) = 1 - x^2$, то многочлен $\omega_{n-1}(x)$ будет только коэффициентом отличаться от многочлена Якоби $P_n^{(1,1)}(x)$, а коэффициенты A_k только множителем $\frac{1}{\Omega_2(x_k)}$ будут отличаться от соответствующих коэффициентов формулы наивысшей алгебраической степени точности:



$$A_k = 8 \cdot \frac{n}{n+1} \cdot \frac{1}{(1-x_k^2)^2 \left[\frac{d}{dx} P_{n-1}^{(1,1)}(x_k) \right]^2}, \quad k = 0, 1, \dots, n-2, \quad (6.85)$$

$$B_1 = B_2 = \frac{2}{n(n+1)}, \quad (6.86)$$

$$R(f) = \frac{8n}{(2n+1)(n+1)} \cdot \left[\frac{2^{n-1} \cdot (n-1)! \cdot (n+1)!}{(2n)!} \right]^2 \cdot \frac{f^{(2n)}(\eta)}{(2n)!}. \quad (6.87)$$



6.1.8. Квадратурные формулы с равными коэффициентами

В приложениях достаточно удобными могут оказаться квадратурные формулы, все коэффициенты которых одинаковы, т.е. квадратурные формулы, имеющие вид

$$I = \int_a^b p(x) f(x) dx \approx C_{n+1} \sum_{k=0}^n f(x_k). \quad (6.88)$$

Их называют *квадратурными формулами Чебышева*. Требование точного выполнения равенства (6.88) при $f(x) \equiv 1$ приводит к уравнению

$$\int_a^b p(x) dx = (n+1) C_{n+1},$$

откуда

$$C_{n+1} = \frac{1}{n+1} \int_a^b p(x) dx. \quad (6.89)$$

Если, кроме того, потребовать, чтобы равенство (6.88) точно выполнялось для $f(x)$, равных x, x^2, \dots, x^{n+1} , то для нахождения узлов x_i получим систему алгебраических уравнений

$$\left\{ \begin{array}{l} x_0 + x_1 + \cdots + x_n = \frac{1}{C_{n+1}} \int_a^b p(x) x dx, \\ x_0^2 + x_1^2 + \cdots + x_n^2 = \frac{1}{C_{n+1}} \int_a^b p(x) x^2 dx, \\ \dots \\ x_0^{n+1} + x_1^{n+1} + \cdots + x_n^{n+1} = \frac{1}{C_{n+1}} \int_a^b p(x) x^{n+1} dx. \end{array} \right. \quad (6.90)$$

Таким образом, для построения квадратурной формулы вида (6.88) достаточно по формуле (6.89) найти коэффициент C_{n+1} , а затем, решив систему (6.90), вычислить узлы искомой формулы. Однако,



учитывая нелинейность системы (6.90), при практическом построении квадратурной формулы вида (6.88) удобнее искать не узлы x_i , а коэффициенты a_i многочлена $\omega_{n+1}(x)$:

$$\omega_{n+1}(x) = x^{n+1} + a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n.$$

Воспользуемся соотношениями Ньютона, связывающими коэффициенты a_i многочлена $\omega_{n+1}(x)$ и степенные суммы его корней $S_k = x_0^k + x_1^k + \cdots + x_n^k$. Тогда вместо формул (6.90) получим систему для определения величин a_i :

$$\left\{ \begin{array}{l} S_1 + a_0 = 0, \\ S_2 + a_0 \cdot S_1 + 2a_1 = 0, \\ \dots \\ S_n + a_0 \cdot S_{n-1} + a_1 \cdot S_{n-2} + \cdots + n a_{n-1} = 0, \\ S_{n+1} + a_0 \cdot S_n + a_1 \cdot S_{n-1} + \cdots + (n+1) a_n = 0. \end{array} \right. \quad (6.91)$$

Формулы (6.91) позволяют последовательно найти все коэффициенты a_i по известным (см. (6.91)) значениям S_k . Далее, решив уравнение $\omega_{n+1}(x) = 0$, найдем все узлы x_k , $k = \overline{0, n}$.

В соответствии с общей схемой алгебраическая степень точности квадратурной формулы (6.88) будет не менее чем $(n+1)$. При этом существует единственная квадратурная формула наивысшей алгебраической степени точности, имеющая равные коэффициенты — рассмотренная ранее [квадратурная формула Гаусса–Чебышева](#).

Сейчас рассмотрим несколько более подробно случай $p(x) \equiv 1$ и $[a, b] = [-1, 1]$, т.е. квадратурные формулы вида

$$I = \int_{-1}^1 f(x) dx \approx C_{n+1} \sum_{k=0}^n f(x_k). \quad (6.92)$$



В этом случае из (6.89) следует, что

$$C_{n+1} = \frac{1}{n+1} \int_{-1}^1 dx = \frac{2}{n+1},$$

а так как

$$\int_{-1}^1 x^k dx = \frac{1 + (-1)^k}{k+1} = \begin{cases} 0, & \text{если } k \text{ нечетное,} \\ \frac{2}{k+1}, & \text{если } k \text{ четное,} \end{cases}$$

то система (6.90) примет вид

$$\left\{ \begin{array}{l} S_1 = x_0 + x_1 + \cdots + x_n = 0, \\ S_2 = x_0^2 + x_1^2 + \cdots + x_n^2 = \frac{n+1}{3}, \\ S_3 = x_0^3 + x_1^3 + \cdots + x_n^3 = 0, \\ S_4 = x_0^4 + x_1^4 + \cdots + x_n^4 = \frac{n+1}{5}, \\ \dots\dots \\ S_{n+1} = x_0^{n+1} + x_1^{n+1} + \cdots + x_n^{n+1} = \frac{n+1}{2} \cdot \frac{1+(-1)^{n+1}}{n+2}. \end{array} \right.$$



Следовательно, формулы для нахождения коэффициентов a_i примут вид

$$\left\{ \begin{array}{l} a_0 = 0, \\ \frac{n+1}{3} + 2a_1 = 0, \\ a_2 = 0, \\ \frac{n+1}{5} + \frac{n+1}{3}a_1 + 4a_3 = 0, \\ \dots \end{array} \right.$$

Таким образом, все коэффициенты a_i четных номеров равны нулю и, следовательно, многочлен $\omega_{n+1}(x)$ будет иметь либо только четные, либо только нечетные степени x . Поэтому его корни располагаются на отрезке $[-1, 1]$ симметрично относительно точки $x = 0$ (а значит, в случае, если $x = 0$ – узел, т.е. при $n = 2m$ алгебраическая степень точности формулы (6.92) будет не ниже $(n + 2)$).

Рассмотрим примеры таких квадратурных формул.

- Пусть $n = 0$. Тогда $a_0 = 0$, $\omega_1(x) = x$. Следовательно, $x_0 = 0$ и, учитывая, что $C_1 = 2$, имеем квадратурную формулу

$$I = \int_{-1}^1 f(x) dx \approx 2f(0),$$

алгебраическая степень точности которой равна 1.

- Пусть $n = 1$. Тогда $C_2 = 1$, для определения коэффициентов a_i имеем систему

$$\left\{ \begin{array}{l} a_0 = 0, \\ 2a_1 + \frac{2}{3} = 0, \end{array} \right.$$



откуда $a_0 = 0$, $a_1 = -\frac{1}{3}$, т.е. $\omega_2(x) = x^2 - \frac{1}{3}$. Следовательно, $x_0 = -\frac{1}{\sqrt{3}}$, $x_1 = \frac{1}{\sqrt{3}}$ и квадратурная формула будет иметь вид

$$I = \int_{-1}^1 f(x) dx \approx f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right).$$

Замечание 6.5. Было установлено, что при $n > 8$ для весовой функции $p(x) \equiv 1$ квадратурных формул Чебышева не существует, поскольку среди корней многочлена $\omega_{n+1}(x)$, построенного описанным выше способом, обязательно появляются комплексные. Лишь сравнительно недавно (1966 г.) были найдены весовые функции $p(x)$, для которых квадратурные формулы Чебышева существуют при любых n .



6.1.9. Нестандартные приемы интегрирования

Метод Филона

Повышение гладкости интегрируемой функции

Случай бесконечных пределов интегрирования

Как мы уже отмечали ранее, использование априорной информации о свойствах и характере поведения функции может существенно улучшить качество приближения. Так, периодические (или близкие к ним) функции более естественно приближать тригонометрическими многочленами, функции, имеющие экспоненциальное поведение, – многочленами от экспонент, и т.д.

Такой подход может оказать существенную помощь и при построении квадратурных формул.

Метод Филона

В радиотехнических задачах часто встречаются функции $f(x)$, описывающие высокочастотные колебания $e^{i\omega x}$ с модулированной амплитудой. Это – быстропеременные функции и их производные $f^{(p)}(x) \sim \omega^p$ велики. Поэтому при интегрировании их по «штатным» квадратурным формулам приходится брать настолько мелкий шаг, чтобы выполнялось условие $wh \ll 1$, т.е. чтобы одна осцилляция содержала бы достаточно большое число узлов интегрирования. А это приводит к большому объему вычислений.

Для уменьшения объема вычислений используем априорные сведения о подынтегральной функции. Представим ее в виде $f(x) = y(x)e^{i\omega x}$, где частота ω известна, а амплитуда $y(x)$ мало меняется за период основного колебания. Выбирая для $y(x)$ несложные полиномиальные аппроксимации, можем получить квадратурные формулы, называемые *формулами Филона* (по сути, речь идет о том, что мы рассматриваем $e^{i\omega x}$ как весовую функцию).

Построим, например, аналог *квадратурной формулы средних прямоугольников*. Для этого при вычислении интеграла поциальному интервалу сетки заменим амплитуду ее значением в середине интервала:

$$y(x) \approx y_{k-\frac{1}{2}}, \quad x \in [x_{k-1}, x_k].$$



При этом для остатка может быть записано приближенное представление (получаемое путем разложения в ряд Тейлора)

$$r(x) = y(x) - y_{k-\frac{1}{2}} \approx \left(x - x_{k-\frac{1}{2}} \right) y'_{k-\frac{1}{2}}.$$

Тогда для вычисления интеграла получим формулу

$$\begin{aligned} I &= \int_a^b e^{i\omega x} y(x) dx = \sum_{k=1}^n \int_{x_{k-1}}^{x_k} e^{i\omega x} y(x) dx \approx \sum_{k=1}^n y_{k-\frac{1}{2}} \int_{x_{k-1}}^{x_k} e^{i\omega x} dx = \\ &= \sum_{k=1}^n y_{k-\frac{1}{2}} \frac{e^{i\omega x_k} - e^{i\omega x_{k-1}}}{i\omega} = \sum_{k=1}^n y_{k-\frac{1}{2}} e^{-\frac{1}{2}} \frac{e^{i\frac{\omega}{2} h_k} - e^{-i\frac{\omega}{2} h_k}}{2i\frac{\omega}{2}} = \frac{2}{\omega} \sum_{k=1}^n f_{k-\frac{1}{2}} \sin\left(\frac{\omega}{2} h_k\right), \end{aligned} \quad (6.93)$$

а для ее остатка – выражение

$$\begin{aligned} R &= \int_a^b r(x) e^{i\omega x} dx \approx \sum_{k=1}^n y'_{k-\frac{1}{2}} \int_{x_{k-1}}^{x_k} \left(x - x_{k-\frac{1}{2}} \right) e^{i\omega x} dx = \sum_{k=1}^n y'_{k-\frac{1}{2}} \left[\frac{x-x_{k-\frac{1}{2}}}{i\omega} e^{i\omega x} \Big|_{x_{k-1}}^{x_k} - \frac{1}{i\omega} \int_{x_{k-1}}^{x_k} e^{i\omega x} dx \right] = \\ &= \sum_{k=1}^n y'_{k-\frac{1}{2}} \left[-\frac{2}{i\omega^2} e^{-\frac{1}{2}} \sin\left(\frac{\omega}{2} h_k\right) + \frac{h_k}{2} \cdot \frac{1}{i\omega} (e^{i\omega x_k} + e^{i\omega x_{k-1}}) \right] = \frac{2i}{\omega^2} \sum_{k=1}^n y'_{k-\frac{1}{2}} \left[\sin\frac{\omega h_k}{2} - \frac{h_k}{2} \cos\frac{\omega h_k}{2} \right] e^{-\frac{1}{2}} \end{aligned} \quad (6.94)$$

Несложно видеть, что при $h \rightarrow 0$ (6.93) и (6.94) переходят в обобщенную формулу средних прямоугольников (6.23) и ее остаток соответственно.

Для построения формул Филона высокого порядка приходится использовать более сложные многочленные аппроксимации.

Повышение гладкости интегрируемой функции

Как правило, все особенности подынтегрального выражения стараются включить в весовую функцию. В литературе этот способ носит название *мультипликативного способа* выделения особенностей.



Вторым способом ослабления особенностей является *аддитивный*. Его суть состоит в следующем. Функцию $f(x)$ представляют в виде $f_1(x) + f_2(x)$, где $f_1(x)$ содержит все или почти все особенности $f(x)$ и при этом интеграл $I_1 = \int_a^b p(x) f_1(x) dx$ вычисляется точно, а $f_2(x)$ имеет ослабленные особенности и для нее с большим успехом применимы квадратурные формулы.

Рассмотрим пример:

$$I = \int_0^{\frac{\pi}{2}} \ln \sin x dx.$$

Подынтегральная функция имеет логарифмическую особенность на левом конце отрезка интегрирования. Поэтому представляет ее в виде

$$\ln \sin x = \ln x + \ln \frac{\sin x}{x}.$$

Тогда

$$I = I_1 + I_2 = \int_0^{\frac{\pi}{2}} \ln x dx + \int_0^{\frac{\pi}{2}} \ln \frac{\sin x}{x} dx.$$

При этом

$$I_1 = \frac{\pi}{2} \left(\ln \frac{\pi}{2} - 1 \right),$$

а в I_2 подынтегральная функция не имеет особенностей и I_2 может быть вычислен, например, по квадратурной формуле Симпсона.

Рассмотрим далее несколько подробнее случай $p(x) \equiv 1$ и алгебраических особенностей. Пусть $f(x) = (x - x_0)^\alpha \varphi(x)$, где $\alpha > -1$, $x_0 \in [a, b]$, а $\varphi(x)$ – достаточно гладкая. При $\alpha < 0$ $f(x)$ имеет алгебраическую особенность, а при $\alpha > 0$ и нецелых производные от $f(x)$ начиная с некоторого порядка будут иметь особенности.

Разложим $\varphi(x)$ в ряд Тейлора:

$$\varphi(x) \approx \varphi(x_0) + \frac{x - x_0}{1!} \varphi'(x_0) + \cdots + \frac{(x - x_0)^{k-1}}{(k-1)!} \varphi^{(k-1)}(x_0)$$



и представим $f(x)$ в виде $f(x) = f_1(x) + f_2(x)$, где

$$f_1(x) = (x - x_0)^\alpha \left[\varphi(x_0) + \frac{x-x_0}{1!} \varphi'(x_0) + \cdots + \frac{(x-x_0)^{k-1}}{(k-1)!} \varphi^{(k-1)}(x_0) \right],$$

$$f_2(x) = f(x) - f_1(x).$$

Интеграл от $f_1(x)$ вычисляется точно, а вычисление интеграла от $f_2(x)$ с помощью квадратурных формул должно дать более хороший результат, так как $f_2(x)$ имеет более высокий порядок гладкости (конкретно: порядок выше на k единиц, поскольку справедливо соотношение $f_2(x) = (x - x_0)^\alpha \left[\frac{\varphi^{(k)}(\xi)}{k!} (x - x_0)^k \right]$).

Случай бесконечных пределов интегрирования

Основными подходами, предназначенными для решения указанной задачи, являются следующие:

- 1) Введение замены переменных, превращающей пределы интегрирования в конечные. Например, для интеграла $\int_a^\infty f(x) dx$, $a > 0$ замена $x = \frac{a}{1-t}$ превращает полуправую $[a, +\infty)$ в отрезок $[0, 1]$. Если после замены подынтегральная функция вместе с некоторым числом производных остается ограниченной, то интеграл можно найти стандартными (описанными выше) способами.
- 2) «Обрезание» бесконечного предела. Пользуясь свойством аддитивности интеграла, представляем его в виде

$$\int_a^{+\infty} f(x) dx = \int_a^b f(x) dx + \int_b^{+\infty} f(x) dx \approx \int_a^b f(x) dx.$$

Очевидно, качество последнего приближенного равенства существенным образом зависит от выбора величины b . Поэтому данный подход требует корректной аналитической оценки и учета величины отброшенного слагаемого (при дальнейшем стандартном вычислении оставленного). Фактически данный прием хорошо комбинировать с применением асимптотических оценок для отбрасываемого члена.

- 3) Применение квадратурных формул наивысшей алгебраической степени точности со специальными весовыми функциями (см. формулы Чебышева-Лягерра и Чебышева-Эрмита).



- 4) Построение специальных *нелинейных* квадратурных формул, применимых на бесконечном интервале.



6.2. Вычисление кратных интегралов

6.2.1. Кубатурные формулы, основанные на сведении кратного интеграла к повторному

6.2.2. Простейшие кубатурные формулы

При решении задачи о приближенном вычислении кратных интегралов чаще других применяются следующие два подхода:

1) сведение задачи к последовательному приближенному вычислению цепочки однократных интегралов. Его идеологической базой является известный в анализе прием сведения кратных интегралов к повторным. При этом, естественно, поскольку речь идет о приближенном вычислении однократных интегралов, то полностью работает вся изложенная в предыдущей главе теория квадратурных формул;

2) построение специальных приближенных формул, непосредственно решающих задачу о вычислении кратного интеграла минуя описанную выше промежуточную стадию.

С точки зрения общей теории при использовании второго подхода к решению задачи о приближенном вычислении кратных интегралов можно пользоваться практически теми же самыми основными понятиями, которые вводились в предыдущей главе. Рассмотрим постановку задачи подробнее.

Пусть необходимо в n -мерном пространстве E_n вычислить интеграл по некоторой области Ω

$$I = \int_{\Omega} p(x) f(x) dx, \quad (6.95)$$

где $x = (x_1, x_2, \dots, x_n)$, $dx = dx_1 dx_2 \dots dx_n$; $p(x)$ – (как и ранее) весовая функция, в состав которой мы будем включать все или основные «неприятности» подынтегрального выражения. Более того, по опыту предыдущей главы будем $p(x)$ считать такой, что существуют интегралы (их мы будем называть моментами) $\mu_{\alpha_1 \dots \alpha_n} = \int_{\Omega} P(x) x_1^{\alpha_1} \dots x_n^{\alpha_n} dx$.

Тогда приближенная формула для вычисления интеграла (6.95) (по аналогии с одномерным случаем) будет иметь вид

$$I = \int_{\Omega} p(x) f(x) dx \approx \sum_{k=0}^n A_k f(x^{(k)}) \quad (6.96)$$



и называться *кубатурной формулой*, A_k – ее коэффициенты, а $x^{(k)} = (x_i^{(k)}, x_2^{(k)}, \dots, x_n^{(k)})$ – узлы.

Если при подстановке в (6.96) вместо $f(x)$ любого алгебраического многочлена от n переменных до степени m включительно равенство превращается в точное (и уже не является таковым при степени многочлена, равной $m+1$), то m – алгебраическая степень точности кубатурной формулы (6.96).

Для построения кубатурных формул можно пользоваться всеми теми же приемами, которые мы рассматривали в разделе 6.1. Это:

- 1) непосредственное использование определения алгебраической степени точности;
- 2) интерполяционная замена интегрируемой функции с последующим точным вычислением интеграла.

При этом правомочна постановка задач о построении кубатурных формул с минимальным числом узлов и максимально возможной алгебраической степенью точности. Известные на настоящий момент результаты теории кубатурных формул аналогичны результатам, изложенным выше. Так, имеет место понятие «*интерполяционная кубатурная формула*» и соответствующая теорема-критерий, а также теоремы о распределении узлов кубатурных формул, обладающих экстремальными характеристиками. При этом распределение узлов связано с поверхностями, определяемыми системами ортогональных многочленов.

Далее, учитывая, что все сказанное в той или иной мере справедливо для произвольного n , мы будем рассматривать вопросы, связанные с вычислением двукратных интегралов.



6.2.1. Кубатурные формулы, основанные на сведении кратного интеграла к повторному

[Повторное интегрирование по прямоугольной области](#)

[Интегралы по криволинейной трапеции](#)

[Повторное интегрирование по прямоугольной области](#)

Как известно из курса анализа, вычисление кратных интегралов может быть осуществлено путем повторного вычисления однократных интегралов. Поэтому, как уже отмечалось выше, одним из простейших путей получения формул для приближенного вычисления кратных интегралов является повторное применение полученных нами ранее квадратурных формул для вычисления однократных интегралов.

Проиллюстрируем это на примере вычисления двойного интеграла по прямоугольнику:

$$I = \int_a^b \int_c^d f(x, y) dx dy. \quad (6.97)$$

Запишем интеграл (6.97) в виде

$$I = \int_a^b dx \int_c^d f(x, y) dy. \quad (6.98)$$

Применяя для вычисления внешнего интеграла квадратурную формулу средних прямоугольников, можем записать:

$$I = \int_a^b dx \int_c^d f(x, y) dy \approx (b - a) \int_c^d f\left(\frac{a + b}{2}, y\right) dy.$$

Вычислив теперь оставшийся интеграл также по формуле средних прямоугольников, окончательно получим:

$$I \approx (b - a)(d - c) f\left(\frac{a + b}{2}, \frac{c + d}{2}\right) = S \cdot f\left(\frac{a + b}{2}, \frac{c + d}{2}\right). \quad (6.99)$$



В качестве других примеров рассмотрим варианты кубатурных формул, получаемых на основе применения других известных вариантов квадратурных формул при повторном интегрировании.

- Кубатурная формула трапеций:

$$\begin{aligned} I = \int_a^b dx \int_c^d f(x, y) dy &\approx \frac{b-a}{2} \left[\int_c^d f(a, y) dy + \int_c^d f(b, y) dy \right] \approx \\ &\approx \frac{b-a}{2} \cdot \left[\frac{d-c}{2} (f(a, c) + f(a, d)) + \frac{d-c}{2} (f(b, c) + f(b, d)) \right] \approx \\ &\approx \frac{S}{4} \cdot [f(a, c) + f(a, d) + f(b, c) + f(b, d)] . \end{aligned} \quad (6.100)$$

- Кубатурная формула Симпсона:

$$\begin{aligned} I = \int_a^b dx \int_c^d f(x, y) dy &\approx \frac{b-a}{6} \cdot \left[\int_c^d f(a, y) dy + 4 \int_c^d f\left(\frac{a+b}{2}, y\right) dy + \int_c^d f(b, y) dy \right] \approx \\ &\approx \frac{S}{36} \cdot [f(a, c) + f(a, d) + f(b, c) + f(b, d)] + \\ &+ \frac{S}{9} \cdot [f\left(a, \frac{c+d}{2}\right) + f\left(\frac{a+b}{2}, c\right) + f\left(\frac{a+b}{2}, d\right) + f\left(b, \frac{c+d}{2}\right) + 4f\left(\frac{a+b}{2}, \frac{c+d}{2}\right)] . \end{aligned} \quad (6.101)$$

Общая схема построения указанного типа кубатурных формул может быть получена, если воспользоваться формулами повторного интерполирования. Например, используя представление соответствующего **интерполяционного многочлена в форме Лагранжа**

$$P_{n,m}(x, y) = \sum_{i=0}^n \sum_{j=0}^m \frac{\omega_{n+1}(x) \omega_{m+1}(y)}{(x - x_i)(y - y_j) \omega'_{n+1}(x_i) \omega'_{m+1}(y_j)} f(x_i, y_j),$$



получим:

$$I = \int_a^b \int_c^d f(x, y) dx dy \approx \sum_{i=0}^n \sum_{j=0}^m f(x_i, y_j) \int_a^b \frac{\omega_{n+1}(x)}{(x - x_i) \omega'_{n+1}(x_i)} dx \int_c^d \frac{\omega_{m+1}(y)}{(y - y_j) \omega'_{m+1}(y_j)} dy. \quad (6.102)$$

Интегралы по криволинейной трапеции

Рассмотрим теперь случай, когда область интегрирования Ω не является прямоугольником, но удовлетворяет условиям, при которых может быть осуществлено сведение к повторному интегралу без разбиения ее на подобласти (для этого достаточно, чтобы контур области пересекался прямыми, параллельными координатным осям, только в двух точках).

Тогда

$$I = \iint_{\Omega} f(x, y) dx dy = \int_a^b dx \int_{y_1(x)}^{y_2(x)} f(x, y) dy = \int_a^b F(x) dx.$$

Выбор правила для вычисления интеграла I , таким образом, должен быть согласован со свойствами функции $f(x, y)$ и, во-вторых, со свойствами области интегрирования Ω .

Если предположить, что $f(x, y)$ является достаточно гладкой всюду в Ω , то интеграл $F(x)$ может быть вычислен по одному из известных правил с постоянным весом, например, по [правилу Гаусса](#), [Симпсона](#) и т.п. Форма области оказывает влияние только на границы интегрирования $y_1(x)$ и $y_2(x)$. Отрезок $[y_1(x), y_2(x)]$ можно привести к каноническому, например, к отрезку $[0, 1]$ с помощью подстановки $y = y_1(x) + (y_2(x) - y_1(x))\eta$. Тогда

$$F(x) = (y_2(x) - y_1(x)) \int_0^1 f(x, y_1(x) + (y_2(x) - y_1(x))\eta) d\eta = (y_2(x) - y_1(x)) \Phi(x).$$

Выделившийся при замене в интеграле $I = \int_a^b F(x) dx$ множитель $y_2(x) - y_1(x)$ является естественной весовой функцией. Поэтому при вычислении интеграла $I = \int_a^b (y_2(x) - y_1(x)) \Phi(x) dx$ можно воспользоваться любой квадратурной формулой, построенной для веса $p(x) = y_2(x) - y_1(x)$, например, квадратурной формулой наивысшей алгебраической степени точности.



Такой полный учет формы области, вероятно, неразумно делать, так как каждой области Ω будет отвечать свой вес $p(x)$ и поэтому пришлось бы использовать большое число узлов и коэффициентов.

Можно упростить задачу на основании следующих простых соображений. Рассмотрим две весовые функции, отличающиеся друг от друга достаточно гладким множителем $\rho(x)$, не обращающимся в нуль на отрезке интегрирования $[a, b]$: $q(x) = \rho(x)p(x)$. Тогда следует ожидать, что квадратурные формулы, соответствующие этим двум весовым функциям $p(x)$ и $q(x)$, будут близки по точности.

А теперь вспомним о [весовой функции Якоби](#) $q(x) = (x - a)^\beta(b - x)^\alpha$. Она зависит от двух параметров α и β и их часто можно подобрать таким образом, чтобы отношение

$$\rho(x) = \frac{y_2(x) - y_1(x)}{(b - x)^\alpha (x - a)^\beta}, \quad a \leq x \leq b,$$

было ограничено сверху и снизу положительными числами. В этом случае можно воспользоваться весом Якоби, преобразовав интеграл I к виду

$$I = \int_a^b (b - x)^\alpha (x - a)^\beta \psi(x) dx.$$

Например, если область интегрирования имеет форму, изображенную на рисунке 6.5, причем контур области λ имеет в точке A с прямой $x = a$ соприкосновение первого порядка, то можно считать $\alpha = 0$, $\beta = 0,5$ и за весовую функцию принять $p(x) = \sqrt{x - a}$. Интеграл

$$I = \int_a^b \sqrt{x - a} \cdot \psi(x) dx$$

может быть вычислен с помощью формул (6.60), (6.61).

Аналогично, в случае, если область интегрирования имеет вид, изображенный на рисунке 6.6 и контур

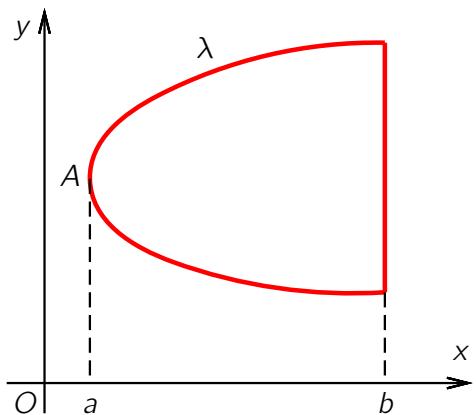


Рисунок 6.5

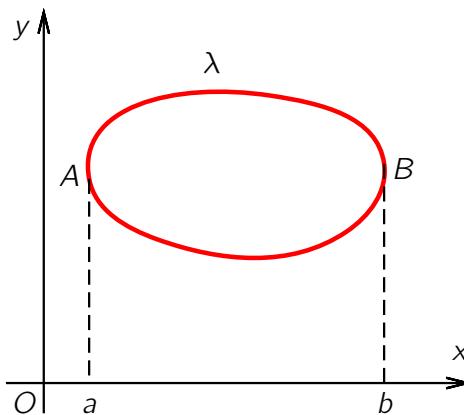


Рисунок 6.6

λ имеет с прямыми $x = a$ и $x = b$ соприкосновение первого порядка, то за весовую функцию можно принять $p(x) = \sqrt{(x - a)(b - x)}$, и к вычислению интеграла

$$I = \int_a^b \sqrt{(x - a)(b - x)} \cdot \psi(x) dx$$

также применить формулы (6.60), (6.61).



6.2.2. Простейшие кубатурные формулы

[Кубатурные формулы на прямоугольнике](#)

[Кубатурные формулы на треугольнике](#)

Несмотря на множество общих моментов, отмеченных во введении к данной главе, проблема вычисления кратных интегралов существенно сложнее по сравнению с аналогичной проблемой вычисления определенного интеграла, и в первую очередь, благодаря тому, что гораздо более сложной может быть область интегрирования (такой, например, является область с криволинейной, пусть и достаточно гладкой, границей). Поэтому изучение вопроса проведем на примерах наиболее простых областей.

Кубатурные формулы на прямоугольнике

Начнем с простейшей области интегрирования – прямоугольника: $\Omega = [a, b] \times [c, d]$ (весовую функцию $p(x, y)$ будем считать тождественной равной единице). Таким образом, речь идет о вычислении интеграла

$$I = \int_a^b \int_c^d f(x, y) dx dy. \quad (6.103)$$

Очевидно, замена независимых переменных

$$\begin{cases} x = \frac{b-a}{2}u + \frac{b+a}{2}, \\ y = \frac{d-c}{2}v + \frac{d+c}{2} \end{cases}$$

переводит рассматриваемый интеграл в интеграл по квадрату $\Omega_1 = [-1; 1] \times [-1, 1]$:

$$I = \frac{(b-a)(d-c)}{4} \cdot \int_{-1}^1 \int_{-1}^1 f_1(u, v) du dv = \frac{S(\Omega)}{4} \cdot \int_{-1}^1 \int_{-1}^1 f_1(u, v) du dv.$$



Учитывая это, далее более подробно рассмотрим вычисление интеграла

$$I = \int_{-1}^1 \int_{-1}^1 f(x, y) dx dy \quad (6.104)$$

Кубатурная формула средних. Самой простой кубатурной формулой будет, естественно, кубатурная формула с одним узлом, т.е. формула вида

$$I = \int_{-1}^1 \int_{-1}^1 f(x, y) dx dy \approx A_0 f(x_0, y_0). \quad (6.105)$$

Выберем узел (x_0, y_0) и коэффициент A_0 таким образом, чтобы обеспечить для рассматриваемой конструкции максимально возможную алгебраическую степень точности. Так как

$$\int_{-1}^1 \int_{-1}^1 dx dy = 4, \quad \int_{-1}^1 \int_{-1}^1 x dx dy = \int_{-1}^1 \int_{-1}^1 y dx dy = 0, \quad (6.106)$$

то для определения параметров кубатурной формулы (6.105) получим, пользуясь определением алгебраической степени точности, систему уравнений

$$\begin{cases} A_0 = 4, \\ A_0 x_0 = 0, \\ A_0 y_0 = 0, \end{cases}$$

решив которую, найдем: $A_0 = 4$, $(x_0, y_0) = (0, 0)$. Следовательно, искомая кубатурная формула имеет вид

$$I = \int_{-1}^1 \int_{-1}^1 f(x, y) dx dy \approx 4f(0, 0). \quad (6.107)$$



Алгебраическая степень точности построенной формулы – не менее единицы (как несложно проверить – в точности равна 1). Найдем ее остаток.

Разлагая интегрируемую функцию $f(x, y)$ в ряд Тейлора в окрестности точки $(0, 0)$, получим:

$$f(x, y) = f(0, 0) + x \frac{\partial f(0, 0)}{\partial x} + y \frac{\partial f(0, 0)}{\partial y} + \frac{x^2}{2} \frac{\partial^2 f(0, 0)}{\partial x^2} + xy \frac{\partial^2 f(0, 0)}{\partial x \partial y} + \frac{y^2}{2} \frac{\partial^2 f(0, 0)}{\partial y^2} + \dots$$

Тогда

$$\begin{aligned} R_0(f) &= \int_{-1}^1 \int_{-1}^1 f(x, y) dx dy - 4f(0, 0) = \\ &= \int_{-1}^1 \int_{-1}^1 \left[x \frac{\partial f(0, 0)}{\partial x} + y \frac{\partial f(0, 0)}{\partial y} + \frac{x^2}{2} \frac{\partial^2 f(0, 0)}{\partial x^2} + xy \frac{\partial^2 f(0, 0)}{\partial x \partial y} + \frac{y^2}{2} \frac{\partial^2 f(0, 0)}{\partial y^2} + \dots \right] dx dy = \\ &= \frac{2}{3} \left[\frac{\partial^2 f(0, 0)}{\partial x^2} + \frac{\partial^2 f(0, 0)}{\partial y^2} \right] + \dots \end{aligned} \quad (6.108)$$

Заметим, что применительно к вычислению интеграла (6.103) кубатурная формула (6.107) будет выглядеть следующим образом:

$$I = \int_a^b \int_c^d f(x, y) dx dy \approx S(\Omega) \cdot f\left(\frac{a+b}{2}, \frac{c+d}{2}\right), \quad (6.109)$$

а ее остаток примет вид

$$R_0(f) = \frac{S(\Omega)}{24} \cdot \left[(b-a)^2 \frac{\partial^2 f\left(\frac{a+b}{2}, \frac{c+d}{2}\right)}{\partial x^2} + (d-c)^2 \frac{\partial^2 f\left(\frac{a+b}{2}, \frac{c+d}{2}\right)}{\partial y^2} \right] + \dots \quad (6.110)$$

Учитывая, что точка $(\bar{x}, \bar{y}) = \left(\frac{a+b}{2}, \frac{c+d}{2}\right)$ является центром прямоугольника Ω , формулу (6.109) (или (6.107)), в точности совпадающую с формулой (6.99), называют *кубатурной формулой средних*.

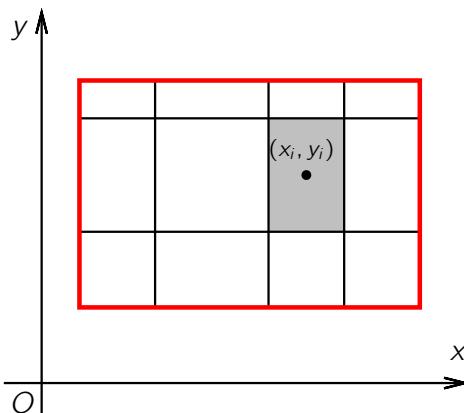


Рисунок 6.7

По аналогии с одномерным случаем легко получить *составную (обобщенную) формулу средних*. Разбивая область интегрирования на прямоугольные ячейки и применяя на каждой ячейке формулу средних (6.109), получим:

$$I = \int_a^b \int_c^d f(x, y) dx dy \approx \sum_i S_i f(\bar{x}_i, \bar{y}_i). \quad (6.111)$$

Здесь S_i – площадь i -й ячейки, а (\bar{x}_i, \bar{y}_i) – координаты ее центра. Сетка, вообще говоря, не обязана быть равномерной по каждому направлению. Если же она таковой является, то формула (6.111) будет иметь несколько более простой вид:

$$I = \int_a^b \int_c^d f(x, y) dx dy \approx \sum_{i=-1}^{N_x} \sum_{j=1}^{N_y} h_x h_y f(\bar{x}_i, \bar{y}_i) \quad (6.112)$$



где отрезок $[a, b]$ разбивается на N_x равных частей, $[c, d]$ – на N_y , т.е.

$$h_x = \frac{b-a}{N_x}, \quad h_y = \frac{d-c}{N_y}, \quad \bar{x}_i = a + \left(i - \frac{1}{2}\right) \cdot h_x, \quad \bar{y}_j = c + \left(j - \frac{1}{2}\right) \cdot h_y.$$

Для каждой ячейки Ω_{ij} остаток вычисляется по формуле (6.101), т.е.

$$R_{ij}(f) = \frac{S(\Omega_{ij})}{24} \cdot \left[h_x^2 \frac{\partial^2 f(\bar{x}_i, \bar{y}_j)}{\partial x^2} + h_y^2 \frac{\partial^2 f(\bar{x}_i, \bar{y}_j)}{\partial y^2} \right] + \dots$$

Суммируя эти выражения по всем ячейкам сетки, получим погрешность составной кубатурной формулы средних:

$$R_0^C(f) = \frac{1}{24} \left[h_x^2 \int_a^b \int_c^d \frac{\partial^2 f(x, y)}{\partial x^2} dx dy + h_y^2 \int_a^b \int_c^d \frac{\partial^2 f(x, y)}{\partial y^2} dx dy \right] + \dots = O(h_x^2 + h_y^2).$$

Таким образом, составная формула средних имеет второй порядок точности.

Замечание 6.6. Формулу средних можно теоретически достаточно просто обобщить на случай более сложной области интегрирования Ω . В этом случае (6.107) примет вид

$$I = \iint_{\Omega} f(x, y) dx dy \approx S(\Omega) \cdot f(\bar{x}, \bar{y}) \quad (6.113)$$

где $S(\Omega)$ – площадь области Ω , а (\bar{x}, \bar{y}) – координаты ее центра тяжести, т.е.

$$S(\Omega) = \iint_{\Omega} dx dy; \quad \bar{x} = \frac{1}{S(\Omega)} \iint_{\Omega} x dx dy; \quad \bar{y} = \frac{1}{S(\Omega)} \iint_{\Omega} y dx dy \quad (6.114)$$

Очевидно, формула (6.113) будет, как и в случае прямоугольной области, иметь алгебраическую степень точности, равную единице, что легко проверить непосредственно.

Замечание 6.7. При любом другом расположении единственного узла кубатурная формула будет иметь алгебраическую степень точности, равную нулю.



Другие кубатурные формулы по прямоугольной области. Вновь возвращаясь к интегралу (6.104), рассмотрим другие примеры кубатурных формул. Естественно, целью поставим построение формул, имеющих более высокую алгебраическую степень точности по сравнению с формулой средних. Как уже отмечалось выше, важную роль играет распределение узлов кубатурной формулы. Вначале исследуем возможности простого количественного увеличения их числа с расположением в «естественных» местах области интегрирования.

Придерживаясь идеологии интерполяционной замены и вспоминая основные способы построения интерполяционного многочлена для функции двух независимых переменных, возьмем в качестве узлов интерполяции точки $(-1, -1)$, $(1, -1)$ и $(-1, 1)$. Тогда

$$f(x, y) \approx P_1(x, y) = f(-1, -1) + (x+1)f(-1, 1; -1) + (y+1)f(-1; -1, 1)$$

Отсюда

$$\begin{aligned} I &= \int_{-1}^1 \int_{-1}^1 f(x, y) dx dy \approx 4f(-1, -1) + 2f(-1, 1; -1) + 2f(-1; -1, 1) = \\ &= 4f(-1, -1) + 4\frac{f(1, -1) - f(-1, -1)}{2} + 4\frac{f(-1, 1) - f(-1, -1)}{2} = 2 \cdot [f(-1, 1) + f(1, -1)] . \end{aligned} \quad (6.115)$$

Таким образом, де-факто получилась кубатурная формула с двумя узлами, имеющая алгебраическую степень точности, равную единице.

Можно показать, что использование двух узлов не приводит к повышению алгебраической степени точности и в то же время выбор в качестве узлов любой пары точек, лежащих на прямой, проходящей через центр симметрии, позволяет построить кубатурную формулу с алгебраической степенью точности, равной единице.

Точно так же и использование четырех узлов, расположенных в соответствии с «естественными эстетическими» соображениями, не приводит к повышению алгебраической степени точности. Убедимся в этом.

Вначале расположим узлы в вершинах квадрата. В этом случае кубатурную формулу проще всего стро-



ить, прибегая к процедуре повторного интерполяирования. Очевидно, формула будет аналогична формуле (6.100) (поскольку интерполирование в каждом направлении проводится по двум узлам):

$$I = \int_{-1}^1 \int_{-1}^1 f(x, y) dx dy \approx f(-1, -1) + f(-1, 1) + f(1, -1) + f(1, 1). \quad (6.116)$$

Несложно видеть, что алгебраическая степень точности построенной кубатурной формулы действительно будет равна единице, поскольку условие точности не выполняется ни для одного из элементарных многочленов второй степени, кроме xy .

В качестве второго примера рассмотрим случай, когда те же четыре узла кубатурной формулы расположены на серединах сторон квадрата, т.е. в точках $(-1, 0)$, $(0, 1)$, $(1, 0)$ и $(0, -1)$. В данном случае интерполяционный многочлен строить несколько затруднительно ввиду его неоднозначности (по количеству узлов). Поэтому прибегнем непосредственно к определению алгебраической степени точности (и, как следствие, к методу неопределенных коэффициентов). Таким образом, будем искать кубатурную формулу вида

$$I = \int_{-1}^1 \int_{-1}^1 f(x, y) dx dy \approx A_0 f(-1, 0) + A_1 f(0, 1) + A_2 f(1, 0) + A_3 f(0, -1). \quad (6.117)$$

Учитывая формулы (6.106), а также равенство $\int_{-1}^1 \int_{-1}^1 x^2 dx dy = \frac{4}{3}$, для определения коэффициентов формулы (6.117) получим систему уравнений

$$\left\{ \begin{array}{l} A_0 + A_1 + A_2 + A_3 = 4, \\ -A_0 + A_2 = 0, \\ A_1 - A_3 = 0, \\ A_0 + A_2 = \frac{4}{3}, \end{array} \right.$$



из которой легко находим: $A_0 = A_2 = \frac{2}{3}$, $A_1 = A_3 = \frac{4}{3}$. Таким образом, (6.117) примет вид

$$I = \int_{-1}^1 \int_{-1}^1 f(x, y) dx dy \approx \frac{2}{3} [f(-1, 0) + 2f(0, 1) + f(1, 0) + 2f(0, -1)].$$

Заметим, что в этом случае выполняется также и уравнение, дающее точность на многочленах вида xy . Но в то же время $\int_{-1}^1 \int_{-1}^1 y^2 dx dy = \frac{4}{3}$, а $A_1 + A_3 = \frac{8}{3} \neq \frac{4}{3}$. Так что алгебраическая степень точности построенной кубатурной формулы остается равной единице, хотя и удовлетворяется на одно уравнение больше по сравнению с рассмотренным выше случаем.

В общем случае, вспоминая, что многочлен второй степени от двух независимых переменных зависит от шести коэффициентов, третьей – от десяти и т.д., делаем вывод, что кубатурные формулы интерполяционного типа повышенной алгебраической степени точности будут достаточно трудоемкими. В то же время, пример формулы средних показывает, что возможно улучшение ситуации (формула средних вместо штатных трех узлов, расположенных почти произвольно (не на одной прямой!) содержит всего один, но специальный узел). В общем случае решение задачи минимизации количества узлов при заданной алгебраической степени точности связано с их расположением на некоторой алгебраической кривой (см., например, [12]). Рассмотрим некоторые частные случаи.

Вначале попытаемся построить кубатурную формулу с тремя узлами, имеющую алгебраическую степень точности, равную двум. Эта формула будет иметь вид

$$I = \int_{-1}^1 \int_{-1}^1 f(x, y) dx dy \approx A_0 f(x_0, y_0) + A_1 f(x_1, y_1) + A_2 f(x_2, y_2). \quad (6.118)$$

Учитывая сказанное выше, расположим узлы специальным образом: равномерно на окружности неко-



торого радиуса r с центром в начале координат. Тогда, очевидно, для описания положения узлов удобно воспользоваться полярной системой координат. В итоге получим:

$$x_0 = r \cos \varphi, \quad y_0 = r \sin \varphi;$$

$$x_1 = r \cos \left(\varphi + \frac{2\pi}{3} \right), \quad y_1 = r \sin \left(\varphi + \frac{2\pi}{3} \right);$$

$$x_2 = r \cos \left(\varphi - \frac{2\pi}{3} \right), \quad y_2 = r \sin \left(\varphi - \frac{2\pi}{3} \right).$$

Как следствие, система уравнений для определения параметров кубатурной формулы (6.118) примет вид

$$\left\{ \begin{array}{l} A_0 + A_1 + A_2 = 4, \\ r(A_0 \cos \varphi + A_1 \cos(\varphi + \frac{2\pi}{3}) + A_2 \cos(\varphi - \frac{2\pi}{3})) = 0, \\ r(A_0 \sin \varphi + A_1 \sin(\varphi + \frac{2\pi}{3}) + A_2 \sin(\varphi - \frac{2\pi}{3})) = 0, \\ r^2(A_0 \cos^2 \varphi + A_1 \cos^2(\varphi + \frac{2\pi}{3}) + A_2 \cos^2(\varphi - \frac{2\pi}{3})) = \frac{4}{3}, \\ r^2(A_0 \sin^2 \varphi + A_1 \sin^2(\varphi + \frac{2\pi}{3}) + A_2 \sin^2(\varphi - \frac{2\pi}{3})) = \frac{4}{3}, \\ r^2(A_0 \cos \varphi \sin \varphi + A_1 \cos(\varphi + \frac{2\pi}{3}) \sin(\varphi + \frac{2\pi}{3}) + \\ \quad + A_2 \cos(\varphi - \frac{2\pi}{3}) \sin(\varphi + \frac{2\pi}{3})) = 0. \end{array} \right. \quad (6.119)$$

Складывая пятое и четвертое уравнения данной системы, получим:

$$r^2(A_0 + A_1 + A_2) = \frac{8}{3},$$

откуда, с учетом первого уравнения, найдем: $r = \sqrt{\frac{2}{3}}$.



После этого исключим из второго и третьего уравнений неизвестное A_0 . Умножая второе уравнение на $\sin \varphi$, а третье – на $\cos \varphi$ и вычитая полученные уравнения друг из друга, получим:

$$\begin{aligned} A_1 \left[\sin \varphi \cos \left(\varphi + \frac{2\pi}{3} \right) - \cos \varphi \sin \left(\varphi + \frac{2\pi}{3} \right) \right] + \\ + A_2 \left[\sin \varphi \cos \left(\varphi - \frac{2\pi}{3} \right) - \cos \varphi \sin \left(\varphi - \frac{2\pi}{3} \right) \right] = 0 \end{aligned}$$

или

$$-A_1 \sin \frac{2\pi}{3} + A_2 \sin \frac{2\pi}{3} = 0,$$

откуда следует, что $A_1 = A_2$. Подставляя это равенство во второе уравнение (6.119), будем иметь:

$$A_0 \cos \varphi + A_1 \left[\cos \left(\varphi + \frac{2\pi}{3} \right) + \cos \left(\varphi - \frac{2\pi}{3} \right) \right] = 0.$$

Полученное уравнение равносильно уравнению

$$A_0 \cos \varphi + 2A_1 \cos \varphi \cos \frac{2\pi}{3} = 0,$$

т.е. $A_0 = A_1 = A_2$. Тогда из первого уравнения (6.119) имеем: $A_0 = A_1 = A_2 = \frac{4}{3}$.

После этого остается заметить, что в силу тригонометрических соотношений

$$\cos \alpha + \cos \left(\alpha + \frac{2\pi}{3} \right) + \cos \left(\alpha - \frac{2\pi}{3} \right) = 0,$$

$$\sin \alpha + \sin \left(\alpha + \frac{2\pi}{3} \right) + \sin \left(\alpha - \frac{2\pi}{3} \right) = 0$$

все уравнения системы (6.119) при найденных значениях A_0 , A_1 , A_2 и r обращаются в тождества при любых значениях аргумента φ .



Таким образом, получаем однопараметрическое семейство кубатурных формул с тремя узлами второго порядка точности:

$$\begin{aligned} I &= \int_{-1}^1 \int_{-1}^1 f(x, y) dx dy \approx \\ &\approx \frac{4}{3} \left(f\left(\sqrt{\frac{2}{3}} \cos \varphi, \sqrt{\frac{2}{3}} \sin \varphi\right) + f\left(\sqrt{\frac{2}{3}} \cos\left(\varphi + \frac{2\pi}{3}\right), \sqrt{\frac{2}{3}} \sin\left(\varphi + \frac{2\pi}{3}\right)\right) + \right. \\ &\quad \left. + f\left(\sqrt{\frac{2}{3}} \cos\left(\varphi - \frac{2\pi}{3}\right), \sqrt{\frac{2}{3}} \sin\left(\varphi - \frac{2\pi}{3}\right)\right) \right) \end{aligned}$$

Придавая φ конкретные значения, получим частные случаи кубатурных формул:

1) $\varphi = 0$:

$$I = \int_{-1}^1 \int_{-1}^1 f(x, y) dx dy \approx \frac{4}{3} \left(f\left(\frac{\sqrt{6}}{3}, 0\right) + f\left(-\frac{\sqrt{6}}{6}, \frac{\sqrt{2}}{2}\right) + f\left(-\frac{\sqrt{6}}{6}, -\frac{\sqrt{2}}{2}\right) \right). \quad (6.120)$$

2) $\varphi = \frac{\pi}{4}$:

$$I = \int_{-1}^1 \int_{-1}^1 f(x, y) dx dy \approx \frac{4}{3} \left(f\left(\frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}\right) + f\left(-\frac{\sqrt{3}+3}{6}, \frac{3-\sqrt{3}}{6}\right) + f\left(-\frac{3-\sqrt{3}}{6}, -\frac{3+\sqrt{3}}{6}\right) \right). \quad (6.121)$$



Аналогичное расположение четырех узлов приводит к кубатурной формуле, алгебраическая степень точности которой равна 3:

$$\begin{aligned} I = \int_{-1}^1 \int_{-1}^1 f(x, y) dx dy \approx & f\left(\sqrt{\frac{2}{3}} \cos \varphi, \sqrt{\frac{2}{3}} \sin \varphi\right) + f\left(-\sqrt{\frac{2}{3}} \sin \varphi, \sqrt{\frac{2}{3}} \cos \varphi\right) + \\ & + f\left(-\sqrt{\frac{2}{3}} \cos \varphi, \sqrt{\frac{2}{3}} \sin \varphi\right) + f\left(\sqrt{\frac{2}{3}} \sin \varphi, -\sqrt{\frac{2}{3}} \cos \varphi\right). \end{aligned} \quad (6.122)$$

Действительно, в этом случае координаты узлов будут иметь вид

$$x_0 = r \cos \varphi, \quad y_0 = r \sin \varphi;$$

$$x_1 = r \cos\left(\varphi + \frac{\pi}{2}\right) = -r \sin \varphi, \quad y_1 = r \sin\left(\varphi + \frac{\pi}{2}\right) = r \cos \varphi;$$

$$x_2 = r \cos(\varphi + \pi) = -\cos \varphi, \quad y_2 = r \sin(\varphi + \pi) = -r \sin \varphi;$$

$$x_3 = r \cos\left(\varphi + \frac{3\pi}{2}\right) = r \sin \varphi, \quad y_3 = r \sin\left(\varphi + \frac{3\pi}{2}\right) = -r \cos \varphi.$$



Система уравнений для определения параметров кубатурной формулы с учетом соотношений $\int_{-1}^1 \int_{-1}^1 x^i y^j dx dy = 0$, если $i + j = 3$ и $i \geq 0, j \geq 0$ примет вид

$$\left\{ \begin{array}{l} A_0 + A_1 + A_2 + A_3 = 4, \\ r(A_0 \cos \varphi - A_1 \sin \varphi - A_2 \cos \varphi + A_3 \sin \varphi) = 0, \\ r(A_0 \sin \varphi + A_1 \cos \varphi - A_2 \sin \varphi - A_3 \cos \varphi) = 0, \\ r^2(A_0 \cos^2 \varphi + A_1 \sin^2 \varphi + A_2 \cos^2 \varphi + A_3 \sin^2 \varphi) = \frac{4}{3}, \\ r^2(A_0 \sin^2 \varphi + A_1 \cos^2 \varphi + A_2 \sin^2 \varphi + A_3 \cos^2 \varphi) = \frac{4}{3}, \\ r^2(A_0 \cos \varphi \sin \varphi - A_1 \cos \varphi \sin \varphi + A_2 \cos \varphi \sin \varphi - A_3 \cos \varphi \sin \varphi) = 0, \\ r^3(A_0 \cos^3 \varphi - A_1 \sin^3 \varphi - A_2 \cos^3 \varphi + A_3 \sin^3 \varphi) = 0, \\ r^3(A_0 \sin^3 \varphi + A_1 \cos^3 \varphi - A_2 \sin^3 \varphi - A_3 \cos^3 \varphi) = 0, \\ r^3(A_0 \cos^2 \varphi \sin \varphi + A_1 \cos \varphi \sin^2 \varphi - A_2 \cos^2 \varphi \sin \varphi - A_3 \cos \varphi \sin^2 \varphi) = 0, \\ r^3(A_0 \cos \varphi \sin^2 \varphi - A_1 \cos^2 \varphi \sin \varphi - A_2 \cos \varphi \sin^2 \varphi + A_3 \cos^2 \varphi \sin \varphi) = 0. \end{array} \right. (*)$$

Как и ранее, складывая четвертое и пятое уравнения данной системы, найдем: $r = \sqrt{\frac{2}{3}}$. Умножая второе уравнение на $\sin \varphi$, а третье – на $\cos \varphi$ и вычитая, получим:

$$-A_1 + A_3 = 0,$$

т.е.

$$A_1 = A_3.$$

Аналогично, умножая второе уравнение системы на $\cos \varphi$, а третье – на $\sin \varphi$ и складывая, получим: $A_0 = A_2$.



Подставляя полученные соотношения в шестое уравнение системы, перепишем его в виде

$$r^2 \cos \varphi \sin \varphi (2A_0 - 2A_1) = 0,$$

откуда

$$A_0 = A_1,$$

и, таким образом,

$$A_0 = A_1 = A_2 = A_3 = 1.$$

Теперь остается проверить, что все уравнения системы (*) обращаются в тождество независимо от величины φ .

Вновь, как и выше выпишем некоторые частные случаи формулы (6.122):

1) $\varphi = 0$:

$$I = \int_{-1}^1 \int_{-1}^1 f(x, y) dx dy \approx f\left(\frac{\sqrt{6}}{3}, 0\right) + f\left(0, \frac{\sqrt{6}}{3}\right) + f\left(-\frac{\sqrt{6}}{3}, 0\right) + f\left(0, -\frac{\sqrt{6}}{3}\right). \quad (6.123)$$

2) $\varphi = \frac{\pi}{4}$:

$$I = \int_{-1}^1 \int_{-1}^1 f(x, y) dx dy \approx f\left(\frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}\right) + f\left(-\frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}\right) + f\left(-\frac{\sqrt{3}}{3}, -\frac{\sqrt{3}}{3}\right) + f\left(\frac{\sqrt{3}}{3}, -\frac{\sqrt{3}}{3}\right). \quad (6.124)$$

Замечание 6.8. Несмотря на то, что в полученных кубатурных формулах второго и третьего порядков точности параметр φ произволен, повысить степень точности за счет его выбора, как легко проверить, не удается.



Кубатурные формулы на треугольнике

Другим интересным типом области является треугольник. Он хорош не только тем, что для него можно явно (по координатам вершин) указать параметры соответствующей кубатурной формулы, но и как тип элементарной ячейки, которая может служить основой для построения составных кубатурных формул при интегрировании по произвольной области (при этом предполагается освоенным процесс *триангуляции области*).

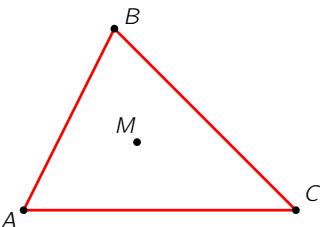


Рисунок 6.8

Формула средних на треугольнике. Перейдем к рассмотрению простейших кубатурных формул, которые построим путем интерполяции подынтегральной функции с последующим интегрированием. Заметим, что площадь треугольника может быть вычислена по формуле

$$S = \frac{1}{2} |g(A, B, C)|, g(A, B, C) = \begin{vmatrix} x_A & y_A & 1 \\ x_B & y_B & 1 \\ x_C & y_C & 1 \end{vmatrix}, \quad (6.125)$$

а координаты центра тяжести (это точка пересечения медиан) находятся по формуле

$$\bar{x} = x_M = \frac{x_A + x_B + x_C}{3}; \quad \bar{y} = y_M = \frac{y_A + y_B + y_C}{3}. \quad (6.126)$$



Тогда (6.113), (6.125), (6.126) – кубатурная формула средних для треугольника. При этом обобщенную формулу (6.113) можно применять к областям, ограниченным ломаной линией.

Получим сейчас аналог формулы трапеций для треугольника. Выбирая в качестве узлов интерполяирования вершины треугольника A , B и C и учитывая, что функция g из формулы (6.125) будет положительной, если расположение ее аргументов соответствует обходу вершин против часовой стрелки (например, $g(A, C, B) > 0$, $g(A, B, C) < 0$), заменим функцию $f(x, y)$ ее интерполяционным многочленом первой степени, который запишем в виде (Q – произвольная точка плоскости, имеющая координаты (x, y))

$$P_1(x, y) = \frac{1}{g(A, C, B)} [f(A)g(Q, C, B) + f(B)g(A, C, Q) + f(C)g(A, Q, B)]. \quad (6.127)$$

Так как $g(Q, C, B)$, $g(A, C, Q)$ и $g(A, Q, B)$ являются линейными функциями аргументов x и y , то для упрощения вычисления интегралов от них можно воспользоваться кубатурной формулой средних (6.123), (6.125), (6.126), которая имеет алгебраическую степень точности, равную единице, и, следовательно, в нашем случае будет давать точное значение каждого из интегралов. Поэтому (M – центр тяжести треугольника)

$$\begin{aligned} \iint_{\Delta} f(x, y) dx dy &\approx \iint_{\Delta} P_1(x, y) dx dy = \\ &= \frac{1}{g(A, C, B)} \iint_{\Delta} [f(A)g(Q, C, B) + f(B)g(A, C, Q) + f(C)g(A, Q, B)] dx dy = \\ &= \frac{1}{g(A, C, B)} \cdot S_{\Delta} \cdot [f(A)g(M, C, B) + f(B)g(A, C, M) + f(C)g(A, M, B)] \end{aligned}$$

Теперь осталось заметить, что в силу (6.125) $g(A, C, B) = 2S_{\Delta}$, а по свойству медиан треугольника

$$g(M, C, B) = g(A, C, M) = g(A, M, B) = \frac{2}{3}S_{\Delta}.$$

Следовательно,

$$\iint_{\Delta} f(x, y) dx dy \approx \frac{1}{2S_{\Delta}} \cdot S_{\Delta} \cdot [f(A) + f(B) + f(C)] \cdot \frac{2}{3}S_{\Delta} = \frac{S_{\Delta}}{3} [f(A) + f(B) + f(C)]. \quad (6.128)$$



Как видим, здесь также нет принципиальных проблем, однако возникают значительные технические трудности при вычислении интегралов от многочленов по произвольному треугольнику. Эти трудности можно преодолеть, если от произвольного треугольника перейти к некоторому стандартному. Достичь этого можно, если ввести так называемые **барицентрические (симплексные)** координаты $(\lambda_1, \lambda_2, \lambda_3)$ произвольной точки (x, y) плоскости по формулам (A, B, C – вершины треугольника)

$$\begin{cases} \lambda_1 + \lambda_2 + \lambda_3 = 1, \\ x_A \lambda_1 + x_B \lambda_2 + x_C \lambda_3 = x, \\ y_A \lambda_1 + y_B \lambda_2 + y_C \lambda_3 = y. \end{cases} \quad (*)$$

Отсюда следует, что если точки A, B, C не лежат на одной прямой (т.е. исходный треугольник не является вырожденным), то

$$\begin{vmatrix} 1 & 1 & 1 \\ x_A & x_B & x_C \\ y_A & y_B & y_C \end{vmatrix} = g(A, B, C) \neq 0$$

и, следовательно, барицентрические координаты $(\lambda_1, \lambda_2, \lambda_3)$ произвольной точки Q плоскости будут иметь вид

$$\lambda_1 = \frac{g(Q, B, C)}{g(A, B, C)}, \quad \lambda_2 = \frac{g(A, Q, C)}{g(A, B, C)}, \quad \lambda_3 = \frac{g(A, B, Q)}{g(A, B, C)}. \quad (6.129)$$

Заметим, что введенные таким образом координаты обладают следующими свойствами.

Свойства барицентрических координат:

1. Если точка Q лежит внутри треугольника, то $\lambda_i > 0$, $i = \overline{1, 3}$;
2. Если точка Q лежит на стороне треугольника, то одна из координат равна нулю;



3. Вершины треугольника имеют координаты $A(1; 0; 0)$, $B(0; 1; 0)$, $C(0; 0; 1)$, центр тяжести — $M\left(\frac{1}{3}; \frac{1}{3}; \frac{1}{3}\right)$.

Первые три свойства означают, что исходный треугольник ABC общего положения взаимно однозначным образом отображается в плоскости, определяемой любыми двумя из трех переменных $\lambda_i > 0$, $i = \overline{1, 3}$, в стандартный равнобедренный прямоугольный треугольник с единичными катетами и вершиной прямого угла, расположенного в начале координат. Доказательство их представляется достаточно очевидным и основывается на формулах (6.129).

$$\begin{aligned} & \iint_{\Delta} \lambda_1^{\alpha_1}(x, y) \cdot \lambda_2^{\alpha_2}(x, y) \cdot \lambda_3^{\alpha_3}(x, y) dx dy = \\ & = S_{\Delta} \cdot \frac{\alpha_1! \cdot \alpha_2! \cdot \alpha_3! \cdot 2!}{(\alpha_1 + \alpha_2 + \alpha_3 + 2)!}, \quad \alpha_i \geqslant 0, \quad \alpha_i \in \mathbb{Z}, \quad i = \overline{1, 3}. \end{aligned} \quad (6.130)$$

Докажем (6.130). Опуская предположение $\alpha_i \in \mathbb{Z}$, перейдем в интегrale к переменным λ_1 , λ_2 (при этом помним, что $\lambda_3 = 1 - \lambda_1 - \lambda_2$ в силу первого из уравнений (*)). С учетом отмеченного второе и третье уравнения (*) примут вид

$$\begin{cases} x = x_C + (x_A - x_C)\lambda_1 + (x_B - x_C)\lambda_2, \\ y = y_C + (y_A - y_C)\lambda_1 + (y_B - y_C)\lambda_2. \end{cases}$$

Следовательно, для якобиана преобразования имеем:

$$J = \frac{D(x, y)}{D(\lambda_1, \lambda_2)} = \begin{vmatrix} \frac{\partial x}{\partial \lambda_1} & \frac{\partial x}{\partial \lambda_2} \\ \frac{\partial y}{\partial \lambda_1} & \frac{\partial y}{\partial \lambda_2} \end{vmatrix} = \begin{vmatrix} x_A - x_C & x_B - x_C \\ y_A - y_C & y_B - y_C \end{vmatrix} = \begin{vmatrix} 1 & 1 & 1 \\ x_A & x_B & x_C \\ y_A & y_B & y_C \end{vmatrix} = g(A, B, C).$$



Таким образом, $|J| = 2! \cdot S_\Delta$, и в результате замены вычисляемый интеграл преобразуется в интеграл

$$I = S_\Delta \cdot 2! \cdot \iint_{\substack{\lambda_1 \geq 0, \lambda_2 \geq 0 \\ \lambda_1 + \lambda_2 \leq 1}} \lambda_1^{\alpha_1} \cdot \lambda_2^{\alpha_2} \cdot (1 - \lambda_1 - \lambda_2)^{\alpha_3} d\lambda_1 d\lambda_2.$$

Переходя в последнем интеграле к повторному, запишем его в виде

$$I = S_\Delta \cdot 2! \cdot \int_0^1 \lambda_1^{\alpha_1} d\lambda_1 \int_0^{1-\lambda_1} \lambda_2^{\alpha_2} (1 - \lambda_1 - \lambda_2)^{\alpha_3} d\lambda_2.$$

Сделав во внутреннем интеграле замену переменной по формуле $\lambda_2 = (1 - \lambda_1) t$, в итоге получим:

$$\begin{aligned} I &= S_\Delta \cdot 2! \cdot \int_0^1 \lambda_1^{\alpha_1} d\lambda_1 \int_0^{1-\lambda_1} \lambda_2^{\alpha_2} (1 - \lambda_1 - \lambda_2)^{\alpha_3} d\lambda_2 = S_\Delta \cdot 2! \cdot \\ &\quad \cdot \int_0^1 \lambda_1^{\alpha_1} (1 - \lambda_1)^{\alpha_1 + \alpha_3 + 1} d\lambda_1 \int_0^1 t^{\alpha_2} (1 - t)^{\alpha_3} dt = \\ &= S_\Delta \cdot 2! \cdot B(\alpha_1 + 1, \alpha_1 + \alpha_3 + 2) \cdot B(\alpha_2 + 1, \alpha_3 + 1) = \\ &= S_\Delta \cdot 2! \cdot \frac{\Gamma(\alpha_1 + 1) \cdot \Gamma(\alpha_2 + 1) \cdot \Gamma(\alpha_3 + 1)}{\Gamma(\alpha_1 + \alpha_2 + \alpha_3 + 3)} \end{aligned}$$

В случае целых значений α_i полученное значение интеграла совпадает с (6.130).

Использование барицентрических координат позволяет значительно упростить построение кубатурных формул. Действительно, в случае, рассмотренном выше, формула (6.127), дающая представление интерполяционного многочлена первой степени, примет вид

$$P_1(x, y) = P_1(\lambda_1, \lambda_2, \lambda_3) = \lambda_1 \cdot f(A) + \lambda_2 \cdot f(B) + \lambda_3 \cdot f(C).$$

После этого использование формул (6.130) при вычислении интегралов сразу же приводит к (6.128).



Кубатурная формула повышенной АСТ. Построим сейчас кубатурную формулу более высокой алгебраической степени точности. Для этого заменим функцию $f(x, y)$ интерполяционным многочленом второй степени по узлам, расположенным в вершинах треугольника и на серединах его сторон.

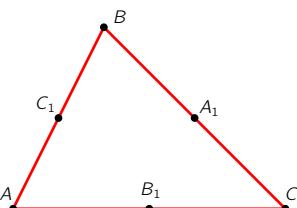


Рисунок 6.9

Согласно формулам (5.82), (5.83) (при этом функции $L_{k,i}(Q)$ в (5.82) будем обозначать точками, через которые проходят соответствующие прямые) соответствующий интерполяционный многочлен примет вид

$$\begin{aligned}
 P_2(x, y) = P_2(\lambda_1, \lambda_2, \lambda_3) &= \frac{BC(Q)}{BC(A)} \cdot \frac{B_1C_1(Q)}{B_1C_1(A)} \cdot f(A) + \frac{AC(Q)}{AC(B)} \cdot \frac{A_1C_1(Q)}{A_1C_1(A)} \cdot f(B) + \\
 &+ \frac{AB(Q)}{AB(C)} \cdot \frac{A_1B_1(Q)}{A_1B_1(C)} \cdot f(C) + \frac{AB(Q)}{AB(A_1)} \cdot \frac{AC(Q)}{AC(A_1)} \cdot f(A_1) + \\
 &+ \frac{BC(Q)}{BC(B_1)} \cdot \frac{AB(Q)}{AB(B_1)} \cdot f(B_1) + \frac{BC(Q)}{BC(C_1)} \cdot \frac{AC(Q)}{AC(C_1)} \cdot f(C_1). \tag{6.131}
 \end{aligned}$$

Уравнение прямой, проходящей через точки B и C , в барицентрических координатах имеет вид

$$\lambda_1 = 0.$$

Поэтому $BC(A) = 1$; $BC(B_1) = BC(C_1) = \frac{1}{2}$.



Аналогично $B_1C_1(Q) = \lambda_1 - \frac{1}{2}$ и $B_1C_1(A) = 1 - \frac{1}{2} = \frac{1}{2}$; $AC(Q) = \lambda_2$ и $AC(B) = 1$; $AC(A_1) = AC(C_1) = \frac{1}{2}$; $A_1C_1(Q) = \lambda_2 - \frac{1}{2}$ и $A_1C_1(B) = 1 - \frac{1}{2} = \frac{1}{2}$; $AB(Q) = \lambda_3$ и $AB(C) = 1$; $AB(A_1) = AB(B_1) = \frac{1}{2}$; $A_1B_1(Q) = \lambda_3 - \frac{1}{2}$ и $A_1B_1(C) = 1 - \frac{1}{2} = \frac{1}{2}$.

Учитывая эти равенства, (6.131) перепишем в виде

$$\begin{aligned} P_2(\lambda_1, \lambda_2, \lambda_3) &= \lambda_1(2\lambda_1 - 1) \cdot f(A) + \lambda_2(2\lambda_2 - 1) \cdot f(B) + \\ &+ \lambda_3(2\lambda_3 - 1) \cdot f(C) + 4\lambda_2\lambda_3 \cdot f(A_1) + \\ &+ 4\lambda_1\lambda_3 \cdot f(B_1) + 4\lambda_1\lambda_2 \cdot f(C_1) \end{aligned}$$

Остается проинтегрировать этот многочлен по треугольнику ABC . Используя (6.130), получим:

$$\iint_{\Delta} f(x, y) dx dy \approx \frac{S_{\Delta}}{3} [f(A_1) + f(B_1) + f(C_1)]. \quad (6.132)$$

Используя формулы (6.130), легко показать, что данная кубатурная формула имеет алгебраическую степень точности, равную 2. Несмотря на то, что интерполяция велось по шести узлам, полученная кубатурная формула содержит всего три узла (и это количество узлов является минимально возможным для формул с алгебраической степенью точности, равной 2).

Замечание 6.9. Кубатурная формула с тремя узлами, обладающая алгебраической степенью точности, равной 2, не является единственной.

Действительно, несложно проверить, что в трехмерном (относительно переменных $\lambda_1, \lambda_2, \lambda_3$) пространстве узлы полученной кубатурной формулы лежат на сфере с центром в точке $M(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ (центре тяжести треугольника) радиуса $r = \frac{1}{\sqrt{6}}$. Используя гипотезу о симметрии (в частности, это означает, что для всех узлов, лежащих на сфере, коэффициенты кубатурной формулы одинаковы) можно непосредственно по определению алгебраической степени точности построить однопараметрическое семейство таких формул. Будем искать такую формулу в виде (опять-таки используя [симплексные координаты](#))

$$\iint_{\Delta} f(x, y) dx dy \approx S_{\Delta} \cdot [A_0 f(\lambda_1^0, \lambda_2^0, \lambda_3^0) + A_1 f(\lambda_1^1, \lambda_2^1, \lambda_3^1) + A_2 f(\lambda_1^2, \lambda_2^2, \lambda_3^2)]. \quad (6.133)$$



Используя для проверки алгебраической степени точности многочлены $1, \lambda_1, \lambda_2, \lambda_1^2, \lambda_2^2, \lambda_1\lambda_2$ и формулы (6.130) для вычисления интегралов, получим для определения параметров формулы (6.133) систему уравнений

$$\left\{ \begin{array}{l} A_0 + A_1 + A_2 = 1, \\ A_0\lambda_1^0 + A_1\lambda_1^1 + A_2\lambda_1^2 = \frac{1}{3}, \\ A_0\lambda_2^0 + A_1\lambda_2^1 + A_2\lambda_2^2 = \frac{1}{3}, \\ A_0(\lambda_1^0)^2 + A_1(\lambda_1^1)^2 + A_2(\lambda_1^2)^2 = \frac{1}{6}, \\ A_0(\lambda_2^0)^2 + A_1(\lambda_2^1)^2 + A_2(\lambda_2^2)^2 = \frac{1}{6}, \\ A_0\lambda_1^0\lambda_2^0 + A_1\lambda_1^1\lambda_2^1 + A_2\lambda_1^2\lambda_2^2 = \frac{1}{12}. \end{array} \right. \quad (6.134)$$

Следуя гипотезе о равенстве коэффициентов, положим $A_0 = A_1 = A_2 = \frac{1}{3}$. Тогда оставшиеся уравнения перепишутся в виде

$$\left\{ \begin{array}{l} \lambda_1^0 + \lambda_1^1 + \lambda_1^2 = 1, \\ \lambda_2^0 + \lambda_2^1 + \lambda_2^2 = 1, \\ (\lambda_1^0)^2 + (\lambda_1^1)^2 + (\lambda_1^2)^2 = \frac{1}{2}, \\ (\lambda_2^0)^2 + (\lambda_2^1)^2 + (\lambda_2^2)^2 = \frac{1}{2}, \\ \lambda_1^0\lambda_2^0 + \lambda_1^1\lambda_2^1 + \lambda_1^2\lambda_2^2 = \frac{1}{4}. \end{array} \right. \quad (6.135)$$

Полагая в первом уравнении $\lambda_1^2 = \alpha$, выразим из него λ_1^1 : $\lambda_1^1 = 1 - \alpha - \lambda_1^0$. Подставив это выражение в третье уравнение, получим:

$$(\lambda_1^0)^2 + (1 - \alpha - \lambda_1^0)^2 = \frac{1}{2} - \alpha^2,$$



Вверх

Назад

Вперёд

Пред.

След.

Указатель

Помощь

Экран

откуда

$$\lambda_1^0 = \frac{1 - \alpha \pm \sqrt{2\alpha - 3\alpha^2}}{2}$$

при условии $0 \leq \alpha \leq \frac{2}{3}$.Пусть для определенности $\lambda_1^0 = \frac{1-\alpha+\sqrt{2\alpha-3\alpha^2}}{2}$. Тогда с учетом введенных обозначений

$$\lambda_1^1 = \frac{1 - \alpha - \sqrt{2\alpha - 3\alpha^2}}{2}, \quad \lambda_1^2 = \alpha.$$

Аналогично, используя второе и четвертое уравнения последней системы, найдем:

$$\lambda_2^0 = \frac{1 - \beta + \sqrt{2\beta - 3\beta^2}}{2}, \quad \lambda_2^1 = \frac{1 - \beta - \sqrt{2\beta - 3\beta^2}}{2}, \quad \lambda_2^2 = \beta, \quad 0 \leq \beta \leq \frac{2}{3}.$$

Подставляя найденные выражения в последнее (пятое) уравнение системы, получим связь между параметрами α и β :

$$\frac{(1-\alpha)(1-\beta)+(1-\alpha)\sqrt{2\beta-3\beta^2}+(1-\beta)\sqrt{2\alpha-3\alpha^2}+\sqrt{2\alpha-3\alpha^2}\sqrt{2\beta-3\beta^2}}{4} +$$

$$+ \frac{(1-\alpha)(1-\beta)-(1-\alpha)\sqrt{2\beta-3\beta^2}-(1-\beta)\sqrt{2\alpha-3\alpha^2}+\sqrt{2\alpha-3\alpha^2}\sqrt{2\beta-3\beta^2}}{4} + \alpha\beta = \frac{1}{4}$$

или

$$2(1 - \alpha - \beta + \alpha\beta) + 2\sqrt{2\alpha - 3\alpha^2}\sqrt{2\beta - 3\beta^2} + 4\alpha\beta = 1.$$

Уединяя радикал и приводя подобные, перепишем это уравнение в виде

$$2\sqrt{2\alpha - 3\alpha^2}\sqrt{2\beta - 3\beta^2} = (2\alpha - 1) - 2(3\alpha - 1)\beta. \quad (*)$$

Заметим также, что если выбрать в формулах для λ_1^0 и λ_2^0 (а значит, и для λ_1^1 и λ_2^1) знаки перед радикалами «в противофазе», то уравнение, связывающее α и β будет иметь вид

$$-2\sqrt{2\alpha - 3\alpha^2}\sqrt{2\beta - 3\beta^2} = (2\alpha - 1) - 2(3\alpha - 1)\beta. \quad (**)$$



Поскольку при избавлении от иррациональности уравнения (*) и (**) переходят в одно и то же уравнение:

$$4(2\alpha - 3\alpha^2)(2\beta - 3\beta^2) = (2\alpha - 1)^2 - 4(2\alpha - 1)(3\alpha - 1)\beta + 4(3\alpha - 1)^2,$$

то это означает, что система (6.135) имеет решения при любых отмеченных выше допустимых значениях переменных α и β .

Приводя подобные, получим квадратное уравнение относительно переменной β :

$$4\beta^2 - 4\beta(1 - \alpha) + (2\alpha - 1)^2 = 0.$$

Его корни –

$$\beta_1 = \frac{1 - \alpha + \sqrt{2\alpha - 3\alpha^2}}{2}, \quad \beta_2 = \frac{1 - \alpha - \sqrt{2\alpha - 3\alpha^2}}{2}.$$

Положим $\beta = \beta_1 = \frac{1 - \alpha + \sqrt{2\alpha - 3\alpha^2}}{2}$. Тогда

$$\begin{aligned} 2\beta - 3\beta^2 &= \frac{4(1 - \alpha) + 4\sqrt{2\alpha - 3\alpha^2} - 3((1 - \alpha)^2 + 2(1 - \alpha)\sqrt{2\alpha - 3\alpha^2} + 2\alpha - 3\alpha^2)}{4} = \\ &= \frac{1 - 4\alpha + 6\alpha^2 - 2(1 - 3\alpha)\sqrt{2\alpha - 3\alpha^2}}{4} = \frac{(1 - 6\alpha + 9\alpha^2) + 2\alpha - 3\alpha^2 - 2(1 - 3\alpha)\sqrt{2\alpha - 3\alpha^2}}{4} = \\ &= \left(\frac{1 - 3\alpha - \sqrt{2\alpha - 3\alpha^2}}{2} \right)^2 \end{aligned}$$

Поэтому (выбираем один из вариантов)

$$\begin{aligned} \lambda_2^0 &= \frac{1 - \beta + \sqrt{2\beta - 3\beta^2}}{2} = \frac{1 - \alpha - \sqrt{2\alpha - 3\alpha^2} + 1 - 3\alpha - \sqrt{2\alpha - 3\alpha^2}}{4} = \\ &= \frac{1 - \alpha - \sqrt{2\alpha - 3\alpha^2}}{2}, \end{aligned}$$



$$\lambda_2^1 = \frac{1 - \beta - \sqrt{2\beta - 3\beta^2}}{2} = \frac{1 - \alpha - \sqrt{2\alpha - 3\alpha^2} - (1 - 3\alpha - \sqrt{2\alpha - 3\alpha^2})}{4} = \alpha.$$

Таким образом, имеем следующее однопараметрическое семейство решений системы (6.135) (в том, что надлежащая комбинация знаков подобрана удачно, убеждаемся непосредственной проверкой):

$$\begin{aligned}\lambda_1^0 &= \frac{1-\alpha+\sqrt{2\alpha-3\alpha^2}}{2}; \quad \lambda_1^1 = \frac{1-\alpha-\sqrt{2\alpha-3\alpha^2}}{2}; \quad \lambda_1^2 = \alpha; \\ 0 &\leq \alpha \leq \frac{2}{3}. \end{aligned}\tag{6.136}$$

$$\lambda_2^0 = \frac{1-\alpha-\sqrt{2\alpha-3\alpha^2}}{2}; \quad \lambda_2^1 = \alpha; \quad \lambda_2^2 = \frac{1-\alpha+\sqrt{2\alpha-3\alpha^2}}{2};$$

Отсюда, в частности, полагая $\alpha = 0$, находим:

$$\lambda_1^0 = \frac{1}{2}; \quad \lambda_1^1 = \frac{1}{2}; \quad \lambda_1^2 = 0;$$

$$\lambda_2^0 = \frac{1}{2}; \quad \lambda_2^1 = 0; \quad \lambda_2^2 = \frac{1}{2}.$$

Таким образом, получаем формулу (6.132).

Аналогично, полагая $\alpha = \frac{2}{3}$, находим:

$$\lambda_1^0 = \frac{1}{6}; \quad \lambda_1^1 = \frac{1}{6}; \quad \lambda_1^2 = \frac{2}{3};$$

$$\lambda_2^0 = \frac{1}{6}; \quad \lambda_2^1 = \frac{2}{3}; \quad \lambda_2^2 = \frac{1}{6}.$$



В результате получаем кубатурную формулу

$$\lambda_1^0 = \frac{1-\alpha+\sqrt{2\alpha-3\alpha^2}}{2}; \quad \lambda_1^1 = \frac{1-\alpha-\sqrt{2\alpha-3\alpha^2}}{2}; \quad \lambda_1^2 = \alpha; \\ 0 \leq \alpha \leq \frac{2}{3} \quad (6.137)$$

$$\lambda_2^0 = \frac{1-\alpha-\sqrt{2\alpha-3\alpha^2}}{2}; \quad \lambda_2^1 = \alpha; \quad \lambda_2^2 = \frac{1-\alpha+\sqrt{2\alpha-3\alpha^2}}{2};$$

Заметим, что среди решений (6.136) существует бесконечно много с рациональными компонентами.

Несмотря на то, что, как показано выше, существует, по крайней мере, однопараметрическое семейство кубатурных формул с тремя узлами, обладающих алгебраической степенью точности, равной 2, среди них нет таких, степень точности которых была бы равной 3. В то же время добавление еще одного, четвертого, узла позволяет эту задачу решить.

Действительно, расположим три узла, как и выше, на рассмотренной там же сфере (при этом соответствующие коэффициенты кубатурной формулы будем считать, вновь следуя гипотезе о симметрии, равными), а в качестве четвертого узла рассмотрим центр указанной сферы. Таким образом, кубатурную формулу будем искать в виде

$$\iint_{\Delta} f(x, y) dx dy \approx S_{\Delta} \cdot \left[A(f(\lambda_1^0, \lambda_2^0, \lambda_3^0) + f(\lambda_1^1, \lambda_2^1, \lambda_3^1) + f(\lambda_1^2, \lambda_2^2, \lambda_3^2)) + Bf\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \right]. \quad (6.138)$$



Как и выше, для определения параметров формулы, добавив условия точности на многочленах λ_1^3 , $\lambda_1^2\lambda_2$, $\lambda_1\lambda_2^2$, λ_2^3 , получим систему уравнений

$$\left\{ \begin{array}{l} 3A + B = 1, \\ A(\lambda_1^0 + \lambda_1^1 + \lambda_1^2) + \frac{1}{3}B = \frac{1}{3}, \\ A(\lambda_2^0 + \lambda_2^1 + \lambda_2^2) + \frac{1}{3}B = \frac{1}{3}, \\ A((\lambda_1^0)^2 + (\lambda_1^1)^2 + (\lambda_1^2)^2) + \frac{1}{9}B = \frac{1}{6}, \\ A((\lambda_2^0)^2 + (\lambda_2^1)^2 + (\lambda_2^2)^2) + \frac{1}{9}B = \frac{1}{6}, \\ A(\lambda_1^0\lambda_2^0 + \lambda_1^1\lambda_2^1 + \lambda_1^2\lambda_2^2) + \frac{1}{9}B = \frac{1}{12}, \\ A((\lambda_1^0)^3 + (\lambda_1^1)^3 + (\lambda_1^2)^3) + \frac{1}{27}B = \frac{1}{10}, \\ A((\lambda_1^0)^2\lambda_2^0 + (\lambda_1^1)^2\lambda_2^1 + (\lambda_1^2)^2\lambda_2^2) + \frac{1}{27}B = \frac{1}{30}, \\ A(\lambda_1^0(\lambda_2^0)^2 + \lambda_1^1(\lambda_2^1)^2 + \lambda_1^2(\lambda_2^2)^2) + \frac{1}{27}B = \frac{1}{30}, \\ A((\lambda_2^0)^3 + (\lambda_2^1)^3 + (\lambda_2^2)^3) + \frac{1}{27}B = \frac{1}{10}. \end{array} \right. \quad (6.139)$$

С учетом сделанного выше предположения о симметрии система будет переопределенной. Выразив из первого уравнения (6.139) B через A , рассмотрим отдельно две системы, первая из которых состоит из второго, четвертого и седьмого, а вторая – из третьего, пятого и десятого уравнений исходной. Неиз-



вестными первой из них будут $\lambda_1^0, \lambda_1^1, \lambda_1^2$, а второй – $\lambda_2^0, \lambda_2^1, \lambda_2^2$. А и в том, и в другом случае считаем параметром. Поэтому указанные системы запишем в виде

$$\left\{ \begin{array}{l} \lambda_1^0 + \lambda_1^1 + \lambda_1^2 = 1, \\ (\lambda_1^0)^2 + (\lambda_1^1)^2 + (\lambda_1^2)^2 = \frac{1}{3} + \frac{1}{18A}, \\ (\lambda_1^0)^3 + (\lambda_1^1)^3 + (\lambda_1^2)^3 = \frac{1}{9} + \frac{1}{270A} \end{array} \right. \quad \left\{ \begin{array}{l} \lambda_2^0 + \lambda_2^1 + \lambda_2^2 = 1, \\ (\lambda_2^0)^2 + (\lambda_2^1)^2 + (\lambda_2^2)^2 = \frac{1}{3} + \frac{1}{18A}, \\ (\lambda_2^0)^3 + (\lambda_2^1)^3 + (\lambda_2^2)^3 = \frac{1}{9} + \frac{1}{270A}. \end{array} \right. \quad (6.140)$$

Обе эти системы как системы относительно указанных выше неизвестных имеют, очевидно, одно и то же множество решений (поскольку с точностью до обозначений неизвестных совпадают). Поэтому рассмотрим несколько подробнее первую из них. Переходя в неё к новым переменным, которые являются элементарными симметрическими многочленами от переменных $\lambda_1^0, \lambda_1^1, \lambda_1^2$ (т.е. $\sigma_1 = \lambda_1^0 + \lambda_1^1 + \lambda_1^2, \sigma_2 = \lambda_1^0\lambda_1^1 + \lambda_1^0\lambda_1^2 + \lambda_1^1\lambda_1^2, \sigma_3 = \lambda_1^0\lambda_1^1\lambda_1^2$), получим:

$$\left\{ \begin{array}{l} \sigma_1 = 1, \\ \sigma_1^2 - 2\sigma_2 = \frac{1}{3} + \frac{1}{18A}, \\ \sigma_1^3 - 3\sigma_1\sigma_2 + 3\sigma_3 = \frac{1}{9} + \frac{17}{270A}. \end{array} \right.$$

Отсюда $\sigma_1 = 1, \sigma_2 = \frac{1}{3} - \frac{1}{36A}, \sigma_3 = \frac{1}{27} \left(1 - \frac{11}{60A}\right)$. Таким образом, $\lambda_1^0, \lambda_1^1, \lambda_1^2$ являются корнями кубического уравнения

$$t^3 - t^2 + \left(\frac{1}{3} - \frac{1}{36A}\right)t - \frac{1}{27} \left(1 - \frac{11}{60A}\right) = 0 \quad (6.141)$$

(t может быть любым из неизвестных $\lambda_1^0, \lambda_1^1, \lambda_1^2$).

Как уже отмечалось выше, решения $\lambda_2^0, \lambda_2^1, \lambda_2^2$ второй из систем (6.140) также будут удовлетворять уравнению (6.141). В то же время понятно, что перестановка корней данного уравнения, определяющая значения $\lambda_1^0, \lambda_1^1, \lambda_1^2$, должна отличаться от аналогичной перестановки, определяющей $\lambda_2^0, \lambda_2^1, \lambda_2^2$ (т.е. эти решения должны быть отличными друг от друга), ибо в противном случае правая часть шестого из



уравнений системы (2.32) будет совпадать с правой частью четвертого (или пятого) уравнения, а левая – от нее отличаться, и, таким образом, (6.139) окажется несовместной.

Таким образом, для того чтобы выяснить какими могут быть соответствующие решения систем (2.32), необходимо исследовать шестое, восьмое и девятое уравнения этой системы в следующих пяти случаях (указываем подстановки, верхняя из которых корни первой из систем (6.140), а нижняя – соответствующие им по номеру корни второй):

$$\begin{pmatrix} \lambda_1^0 & \lambda_1^1 & \lambda_1^2 \\ \lambda_2^1 & \lambda_2^0 & \lambda_2^2 \end{pmatrix}, \quad \begin{pmatrix} \lambda_1^0 & \lambda_1^1 & \lambda_1^2 \\ \lambda_2^1 & \lambda_2^2 & \lambda_2^0 \end{pmatrix}, \quad \begin{pmatrix} \lambda_1^0 & \lambda_1^1 & \lambda_1^2 \\ \lambda_2^0 & \lambda_2^2 & \lambda_2^1 \end{pmatrix},$$

$$\begin{pmatrix} \lambda_1^0 & \lambda_1^1 & \lambda_1^2 \\ \lambda_2^2 & \lambda_2^0 & \lambda_2^1 \end{pmatrix}, \quad \begin{pmatrix} \lambda_1^0 & \lambda_1^1 & \lambda_1^2 \\ \lambda_2^2 & \lambda_2^1 & \lambda_2^0 \end{pmatrix}.$$

Рассмотрим подробнее первый случай. Так как $\lambda_1^0 = \lambda_2^1$, $\lambda_1^1 = \lambda_2^0$, $\lambda_1^2 = \lambda_2^2$, то система из шестого, восьмого и девятого уравнений (6.139) примет вид

$$\begin{cases} 2\lambda_1^0\lambda_1^1 + (\lambda_1^2)^2 = \frac{1}{3} - \frac{1}{36A}, \\ (\lambda_1^0)^2\lambda_1^1 + \lambda_1^0(\lambda_1^1)^2 + (\lambda_1^2)^3 = \frac{1}{9} - \frac{1}{270A}, \\ \lambda_1^0(\lambda_1^1)^2 + (\lambda_1^0)^2\lambda_1^1 + (\lambda_1^2)^3 = \frac{1}{9} - \frac{1}{270A}. \end{cases} \quad (6.142)$$

Как видим, второе и третье уравнения в данной системе совпадают. Поэтому в дальнейшем оставляем только одно из них. В то же время, правая часть первого уравнения совпадает с найденным ранее значением σ_2 , т.е. справедливо соотношение, связывающее значения λ_1^0 , λ_1^1 , λ_1^2 :

$$\lambda_1^0\lambda_1^1 + \lambda_1^0\lambda_1^2 + \lambda_1^1\lambda_1^2 = 2\lambda_1^0\lambda_1^1 + (\lambda_1^2)^2$$

или

$$(\lambda_1^0 - \lambda_1^2)(\lambda_1^2 - \lambda_1^1) = 0.$$



Таким образом, рассматриваемый случай распадается на два. Пусть вначале $\lambda_1^0 = \lambda_1^1$. Тогда, во-первых, (6.142) примет вид

$$\begin{cases} 2\lambda_1^0\lambda_1^1 + (\lambda_1^0)^2 = \frac{1}{3} - \frac{1}{36A}, \\ \lambda_1^0\lambda_1^1(\lambda_1^0 + \lambda_1^1) + (\lambda_1^0)^3 = \frac{1}{9} - \frac{1}{270A}. \end{cases} \quad (6.143)$$

а во-вторых, уравнение (6.141) имеет двукратный корень, т.е. корень λ_1^0 является также корнем и производной многочлена, стоящего в левой части (6.141). Поэтому

$$3(\lambda_1^0)^2 - 2\lambda_1^0 = -\left(\frac{1}{3} - \frac{1}{36A}\right).$$

Подставляя сюда вместо правой части левую часть первого из уравнений (6.143), имеем:

$$3(\lambda_1^0)^2 - 2\lambda_1^0 = -2\lambda_1^0\lambda_1^1 - (\lambda_1^0)^2,$$

откуда

$$\lambda_1^0 = \frac{1 - \lambda_1^1}{2}.$$

С учетом найденного соотношения (6.143) примет вид

$$\begin{cases} \lambda_1^1(1 - \lambda_1^1) + \frac{(1 - \lambda_1^1)^2}{4} = \frac{1}{3} - \frac{1}{36A}, \\ \frac{\lambda_1^0(1 - \lambda_1^1)(1 + \lambda_1^1)}{4} + \frac{(1 - \lambda_1^1)^3}{8} = \frac{1}{9} - \frac{1}{270A} \end{cases}$$

или

$$\begin{cases} \frac{(1 - \lambda_1^1)(3\lambda_1^1 + 1)}{4} = \frac{1}{3} - \frac{1}{36A}, \\ \frac{(1 - \lambda_1^1)(3(\lambda_1^1)^2 + 1)}{8} = \frac{1}{9} - \frac{1}{270A}. \end{cases}$$

Умножая первое уравнение на $\frac{1}{15}$, а второе – на $-\frac{1}{2}$ и складывая, исключим неизвестное A :

$$\frac{(1 - \lambda_1^1)(3\lambda_1^1 + 1)}{60} - \frac{(1 - \lambda_1^1)(3(\lambda_1^1)^2 + 1)}{16} = -\frac{1}{30}.$$



Избавляясь от знаменателя, перепишем это уравнение в виде

$$(5\lambda_1^1 - 3)(3\lambda_1^1 - 1)^2 = 0.$$

Отсюда $\lambda_1^1 = \frac{3}{5}$. Тогда $A = \frac{25}{48}$, $\lambda_1^0 = \lambda_1^2 = \frac{1}{5}$, $B = 1 - 3A = -\frac{27}{48}$.

Таким образом, искомая кубатурная формула в данном случае имеет вид

$$\iint_{\Delta} f(x, y) dx dy \approx \frac{S_{\Delta}}{48} \cdot \left[25 \left(f\left(\frac{1}{5}, \frac{3}{5}, \frac{1}{5}\right) + f\left(\frac{3}{5}, \frac{1}{5}, \frac{1}{5}\right) + f\left(\frac{1}{5}, \frac{1}{5}, \frac{3}{5}\right) \right) - 27f\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \right]. \quad (6.144)$$



Глава 7

Численное решение интегральных уравнений

- 7.1. Введение
- 7.2. Методы решения интегральных уравнений Фредгольма второго рода
- 7.3. Проекционные методы решения интегральных уравнений
- 7.4. Решение интегральных уравнений первого рода



7.1. Введение

Данный раздел является первым из разделов, посвященных численному решению задач, искомой величиной в которых является некоторая функция (функциональных уравнений).

В достаточно общем виде *интегральное уравнение* с одной неизвестной функцией может быть записано в форме

$$\Phi \left(x, u(x), \int_a^t f(x, s, u(s)) ds \right) = 0, \quad (7.1)$$

где $u(x)$ – искомая функция, $t \in \{b, x\}$.

Если $t = b$, то говорят, что уравнение принадлежит виду Фредгольма, а если $t = x$, то виду Вольтерра.

В приложениях в столь общем виде интегральные уравнения встречаются достаточно редко. Гораздо чаще приходится иметь дело с линейными уравнениями. В первом из случаев они будут иметь вид

$$A(x) u(x) + \int_a^b K(x, s) u(s) ds = F(x), \quad (7.2)$$

а во втором –

$$A(x) u(x) + \int_a^x K(x, s) u(s) ds = F(x). \quad (7.3)$$

Если коэффициент $A(x)$ тождественно равен нулю на отрезке $[a, b]$, то неизвестная функция $u(x)$ входит в соответствующее уравнение только под знаком интеграла. В этом случае уравнение носит название интегрального уравнения первого рода. Таким образом, имеем *интегральное уравнение Фредгольма* первого рода

$$\int_a^b K(x, s) u(s) ds = F(x) \quad (7.4)$$



и [интегральное уравнение Вольтерра](#) первого рода

$$\int_a^x K(x, s) u(s) ds = F(x). \quad (7.5)$$

Задача решения этих уравнений имеет принципиальные трудности, поскольку относится к разряду некорректных.

Если же $A(x) \neq 0$ для любого $x \in [a, b]$, то уравнения (7.2), (7.3) могут быть приведены соответственно к виду

$$u(x) - \lambda \int_a^b K(x, s) u(s) ds = f(x) \quad (7.6)$$

и

$$u(x) - \lambda \int_a^x K(x, s) u(s) ds = f(x). \quad (7.7)$$

Формулы (7.6) и (7.7) определяют [линейные интегральные уравнения Фредгольма и Вольтерра второго рода](#) соответственно. Формально положив в (7.7) $K(x, s) \equiv 0$ при $s > x$, можно свести уравнение Вольтерра к уравнению Фредгольма, рассматривая его как частный случай последнего. Однако возникающая при этом разрывность ядра $K(x, s)$ при $x = s$ делает указанную процедуру нежелательной и, вообще говоря, теория интегральных уравнений Фредгольма и Вольтерра существенно отличаются, как и методы их численного решения, к изучению которых мы и переходим.



7.2. Методы решения интегральных уравнений Фредгольма второго рода

- 7.2.1. Метод механических квадратур
- 7.2.2. Метод замены ядра на вырожденное
- 7.2.3. Метод последовательных приближений
- 7.2.4. Методы решения интегральных уравнений Вольтерра второго рода



7.2.1. Метод механических квадратур

Оценка погрешности метода механических квадратур

Рассмотрим интегральное уравнение Фредгольма второго рода

$$u(x) - \lambda \int_a^b K(x, s) u(s) ds = f(x). \quad (7.8)$$

Выберем на отрезке $[a, b]$ $(n+1)$ точек $a \leq x_0 < x_1 < \dots < x_n \leq b$ и заменим в уравнении (7.8) интеграл некоторой [квадратурной суммой](#), причем точки x_k будут ее узлами. Тогда вместо (7.8) получим равенство

$$u(x) - \lambda \sum_{k=0}^n K(x, x_k) u(x_k) = f(x) + \lambda \rho(x), \quad (7.9)$$

где A_k – коэффициенты выбранной квадратурной формулы, а $\rho(x)$ – ее остаток. Соответствующие выражения для них мы получали [ранее](#). Например, если в качестве квадратурной формулы используется [составная формула средних прямоугольников](#) на равномерной сетке, то

$$x_k = a + \left(k + \frac{1}{2}\right) h, \quad k = \overline{0, n-1}; \quad h = \frac{b-a}{n}, \quad A_k = h,$$

$$\rho(x) = \frac{h^2}{24} (b-a) \frac{\partial^2 (K(x, \eta) u(\eta))}{\partial s^2}.$$

Если теперь в равенстве (7.9) положить последовательно $x = x_i$, $i = \overline{0, n}$, то в результате получится система линейных алгебраических уравнений для нахождения точных значений решения в узлах:

$$u(x_i) - \lambda \sum_{k=0}^n K(x_i, x_k) u(x_k) = f(x_i) + \lambda \rho(x_i), \quad i = \overline{0, n}. \quad (7.10)$$



Остаток $\rho(x)$ квадратурной формулы обычно мал по сравнению с самой величиной квадратурной суммы, поэтому, отбрасывая в (7.10) малые величины $\lambda\rho(x_i)$, получим систему линейных алгебраических уравнений

$$y_i - \lambda \sum_{k=0}^n K_{ik} y_k = f_i, \quad i = \overline{0, n} \quad (7.11)$$

(здесь использованы обозначения $y_i \approx u(x_i)$, $K_{ik} = K(x_i, x_k)$, $f_i = f(x_i)$).

Решения y_i системы (7.11) будут являться некоторыми приближениями к точным решениям задачи (7.8) в узлах x_i , $i = \overline{0, n}$. Найти их можно, используя для этих целей любой (или наиболее подходящий) из методов решения систем линейных алгебраических уравнений (например, [метод Гаусса](#)).

Замечание 7.1. Зная величины y_i , $i = \overline{0, n}$, можно легко восстановить приближенное решение во всех точках отрезка $[a, b]$. Для этих целей естественно воспользоваться формулой

$$y(x) = f(x) + \lambda \sum_{k=0}^n K(x, x_k) y_k. \quad (7.12)$$

Как известно из теории систем линейных алгебраических уравнений, в случае, если определитель матрицы системы (7.11) отличен от нуля, т.е.

$$\Delta(\lambda) = \begin{vmatrix} 1 - \lambda A_0 K_{00} & -\lambda A_1 K_{01} & \dots & -\lambda A_n K_{0n} \\ -\lambda A_0 K_{10} & 1 - \lambda A_1 K_{11} & \dots & -\lambda A_n K_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ -\lambda A_0 K_{n0} & -\lambda A_1 K_{n1} & \dots & 1 - \lambda A_n K_{nn} \end{vmatrix} \neq 0,$$

то решение ее существует и единствено. Если же $\Delta(\lambda) = 0$, то значения λ , при которых это соотношение выполняется, будут приближениями к соответствующим собственным значениям ядра. Если для этих λ положить в (7.11) $f_i = 0$, $i = \overline{0, n}$, то отсюда можно найти и собственные функции ядра интегрального уравнения (7.8) (естественно, приближенно), т.е. метод механических квадратур может быть использован для решения проблемы собственных значений.



Оценка погрешности метода механических квадратур

Проведем исследование погрешности [метода механических квадратур](#). Будем предполагать, что $\Delta(\lambda) \neq 0$. При численном решении системы (7.11) каждое уравнение удовлетворяется, вообще говоря, с некоторой погрешностью (округлений). Обозначим ее через $-\delta_i$. Тогда для величин y_i , $i = \overline{0, n}$ вместо (7.11) будут выполняться равенства

$$y_i - \lambda \sum_{k=0}^n K_{ik} y_k = f_i - \delta_i, \quad i = \overline{0, n}. \quad (7.13)$$

Отсюда, используя правило Крамера, можем записать:

$$y_i = \frac{1}{\Delta(\lambda)} \sum_{k=0}^n \Delta_{ki} (f_k - \delta_k), \quad i = \overline{0, n}, \quad (7.14)$$

где Δ_{ki} – алгебраическое дополнение элемента $f_k - \delta_k$ определителя Δ_i .

Рассмотрим погрешность приближенного решения $\varepsilon_i = u(x_i) - y_i$ и получим для нее оценку. Ранее мы видели, что точные значения $u(x_i)$ удовлетворяют системе (7.10). Поэтому, аналогично (7.14), может записать:

$$u(x_i) = \frac{1}{\Delta(\lambda)} \sum_{k=0}^n \Delta_{ki} (f_k + \lambda \rho(x_k)), \quad i = \overline{0, n}. \quad (7.15)$$

Вычитая из (7.15) (7.14), получим:

$$\varepsilon_i = \frac{1}{\Delta(\lambda)} \sum_{k=0}^n \Delta_{ki} [\lambda \rho(x_k) + \delta_k], \quad i = \overline{0, n}. \quad (7.16)$$

Пусть теперь для погрешности $\rho(x)$ квадратурной формулы верна оценка $|\rho(x)| \leq \rho = \rho(n)$ (ранее мы видели, что такая оценка всегда может быть получена и зависит от производных некоторого порядка



интегрируемой функции). Предположим также, что для вычислительной погрешности также имеет место оценка $|\delta_i| \leq \delta = \delta(n)$. Тогда из (7.16) следует, что

$$|\varepsilon_i| \leq \frac{|\lambda| \rho + \delta}{|\Delta(\lambda)|} \sum_{k=0}^n |\Delta_{ki}| = B(|\lambda| \rho + \delta), \quad (7.17)$$

где $B = \frac{\sum_{k=0}^n |\Delta_{ki}|}{|\Delta(\lambda)|}$.

Отметим, что значения $\Delta(\lambda)$ и Δ_{ki} могут быть найдены в процессе вычислений и поэтому B является вычислимой константой в отличие от ρ , способ нахождения которой еще предстоит указать.

Рассмотрим остаточный член $\rho(x)$ более подробно.

$$\begin{aligned} \rho(x) &= \int_a^b K(x, s) u(s) ds - \sum_{k=0}^n A_k K(x, x_k) u(x_k) = \left[u(x) = f(x) + \lambda \int_a^b K(x, s) u(s) ds \right] = \\ &= \int_a^b K(x, s) \left[f(s) + \lambda \int_a^b K(s, t) u(t) dt \right] - \sum_{k=0}^n A_k K(x, x_k) \left[f(x_k) + \lambda \int_a^b K(x_k, t) u(t) dt \right] = \\ &= \int_a^b K(x, s) f(s) ds - \sum_{k=0}^n K(x, x_k) f(x_k) + \lambda \int_a^b \left[\int_a^b K(x, s) K(s, t) ds - \sum_{k=0}^n A_k K(x, x_k) K(x_k, t) \right] u(t) dt = \\ &\quad = \rho_f(x) + \lambda \int_a^b \rho_K(x, t) u(t) dt, \quad (7.18) \end{aligned}$$

где

$$\rho_f(x) = \int_a^b K(x, s) f(s) ds - \sum_{k=0}^n A_k K(x, x_k) f(x_k)$$

$$\rho_K(x, t) = \int_a^b K(x, s) K(s, t) ds - \sum_{k=0}^n A_k K(x, x_k) K(x_k, t).$$



Так как $\rho_f(x)$ и $\rho_K(x, t)$ могут быть вычислены, поскольку $f(x)$ и $K(x, s)$ известны, то из (7.18) получим:

$$|\rho(x_i)| \leq \rho \leq \max_{x \in [a, b]} |\rho(x)| \leq \rho_f + |\lambda| (b - a) \rho_K H, \quad (7.19)$$

где $H = \max_{x \in [a, b]} |u(x)|$.

Подставляя (7.19) в (7.17), найдем:

$$|\varepsilon_i| \leq B [|\lambda| (\rho_f + |\lambda| (b - a) \rho_K H) + \delta], \quad i = \overline{0, n}. \quad (7.20)$$

Оценка (7.20) позволяет во многих случаях сделать заключение о сходимости вычислительного процесса. Однако ее недостатком является наличие величины H – максимума модуля неизвестной функции.

Оценим H через вычислимые величины. Определив приближенное решение во всех точках отрезка $[a, b]$ по формуле (7.12), для погрешности $\varepsilon(x)$ получим:

$$\varepsilon(x) = u(x) - y(x) = \lambda \left[\sum_{k=0}^n A_k K(x, x_k) \varepsilon_k + \rho(x) \right].$$

Отсюда, используя (7.19), (7.20), будем иметь (в предположении, что использованная квадратурная формула имеет положительные коэффициенты):

$$\begin{aligned} |\varepsilon(x)| &\leq |\lambda| \left[\left| \sum_{k=0}^n A_k K(x, x_k) \varepsilon_k \right| + |\rho(x)| \right] \leq |\lambda| \left[\max_{0 \leq k \leq n} |\varepsilon_k| + \max_{(x,s)} |K(x, s)| \cdot \left| \sum_{k=0}^n A_k \right| + |\rho(x)| \right] \leq \\ &\leq |\lambda| \{ (b - a) MB [|\lambda| (\rho_f + |\lambda| (b - a) \rho_K H) + \delta] + \rho_f + |\lambda| (b - a) \rho_K H \}. \end{aligned}$$

Здесь использовано обозначение $M = \max_{(x,s)} |K(x, s)|$. Пусть теперь $\tilde{H} = \max_{x \in [a, b]} |y(x)|$. Эта величина вычислима. Так как $u(x) = y(x) + \varepsilon(x)$, то отсюда $H \leq \tilde{H} + \max_{x \in [a, b]} |\varepsilon(x)|$, и, применяя оценку для $|\varepsilon(x)|$, имеем:

$$H \leq \tilde{H} + |\lambda| \{ (b - a) MB [|\lambda| (\rho_f + |\lambda| (b - a) \rho_K H) + \delta] + \rho_f + |\lambda| (b - a) \rho_K H \}$$



или

$$H \left[1 - |\lambda|^3 (b-a)^2 MB \rho_K - |\lambda|^2 (b-a) \rho_K \right] \leq \tilde{H} + |\lambda| [\rho_f + (b-a) MB (|\lambda| \rho_f + \delta)].$$

Отсюда при выполнении условия

$$1 - |\lambda|^2 (b-a) \rho_K [|\lambda| (b-a) MB + 1] > 0$$

вытекает неравенство для H :

$$H \leq \frac{\tilde{H} + |\lambda| [\rho_f + (b-a) MB (|\lambda| \rho_f + \delta)]}{1 - |\lambda|^2 (b-a) \rho_K [|\lambda| (b-a) MB + 1]}. \quad (7.21)$$

Таким образом, совместно оценки (7.20), (7.21) являются вычислимыми.

Учитывая приведенные оценки, можно сформулировать следующие соображения, касающиеся выбора квадратурной формулы для замены интеграла в (7.8): естественно выбирать такую квадратурную формулу, чтобы ее остаток $\rho(x)$ был по возможности малым. Этого можно достичь двумя путями:

- 1) за счет увеличения числа узлов;
- 2) за счет повышения алгебраической степени точности квадратурной формулы.

При этом следует учитывать, что увеличение числа узлов ведет к увеличению объема работы при решении системы (7.11), увеличение же алгебраической степени точности квадратурной формулы даст эффект лишь тогда, когда интегрируемые функции обладают достаточными свойствами гладкости (в первую очередь это касается ядра $K(x, s)$).

Заметим также, что приведенные выше оценки погрешности, хоть и являются вычислимыми, однако требуют очень серьезной аналитической работы. Поэтому практическое представление о погрешности можно составить, используя расчеты на вложенных сетках (аналог правила Рунге).



7.2.2. Метод замены ядра на вырожденное

Разложение ядра в ряд Тейлора

Использование ортогональных разложений

Интерполяционные способы замены ядра

Способ Бэтмена

Определение. Ядро $K(x, s)$ называется **вырожденным**, если оно может быть представлено в виде

$$K(x, s) = \sum_{i=0}^n \alpha_i(x) \beta_i(s). \quad (7.22)$$

Системы $\alpha_i(x)$ и $\beta_i(s)$ ($i = \overline{0, n}$) в (7.22) естественно считать линейно независимыми, так как в противном случае число слагаемых в (7.22) можно было бы уменьшить.

Примеры вырожденных ядер:

$$1) K(x, s) = e^{x+s} = e^x \cdot e^s = \alpha_0(x) \cdot \beta_0(s);$$

$$2) K(x, s) = \sin(x + s) = \sin x \cdot \cos s + \cos x \cdot \sin s = \alpha_0(x) \cdot \beta_0(s) + \alpha_1(x) \cdot \beta_1(s).$$

Для вырожденных ядер **уравнение Фредгольма второго рода** решается в аналитическом виде за конечное число действий.

Действительно, перепишем (7.8) в виде

$$u(x) = \lambda \int_a^b K(x, s) u(s) ds + f(x) = \lambda \int_a^b [\sum_{i=0}^n \alpha_i(x) \beta_i(s)] u(s) ds + f(x) =$$

$$= \lambda \sum_{i=0}^n \alpha_i(x) \int_a^b \beta_i(s) u(s) ds + f(x) = f(x) + \lambda \sum_{i=0}^n C_i \alpha_i(x),$$



т.е. фактически мы нашли вид точного решения

$$u(x) = f(x) + \lambda \sum_{i=0}^n C_i \alpha_i(x), \quad (7.23)$$

где

$$C_i = \int_a^b \beta_i(s) u(s) ds, \quad i = \overline{0, n}. \quad (7.24)$$

Чтобы найти C_i , подставим (7.23) в (7.24):

$$C_i = \int_a^b \beta_i(s) \left[f(s) + \lambda \sum_{j=0}^n C_j \alpha_j(s) \right] ds, \quad i = \overline{0, n}$$

или

$$C_i - \lambda \sum_{j=0}^n C_j a_{ij} = b_i, \quad i = \overline{0, n}, \quad (7.25)$$

где

$$b_i = \int_a^b \beta_i(s) f(s) ds; \quad a_{ij} = \int_a^b \beta_i(s) \alpha_j(s) ds, \quad i = \overline{0, n}; \quad j = \overline{0, n}. \quad (7.26)$$

Таким образом, для определения коэффициентов C_i формулы (7.23) получаем систему линейных алгебраических уравнений. Если определитель ее

$$\Delta(\lambda) = \begin{vmatrix} 1 - \lambda a_{00} & -\lambda a_{01} & \dots & -\lambda a_{0n} \\ -\lambda a_{10} & 1 - \lambda a_{11} & \dots & -\lambda a_{1n} \\ \dots & & & \\ -\lambda a_{n0} & -\lambda a_{n1} & \dots & 1 - \lambda a_{nn} \end{vmatrix}$$



отличен от нуля, то мы найдем единственным образом набор констант C_i и, следовательно, построим точное решение $u(x)$. Случай $\Delta(\lambda) = 0$ соответствует собственным значениям.

Изложенное выше позволяет указать метод приближенного решения интегрального уравнения (7.8), основной идеей которого является замена ядра исходного интегрального уравнения $K(x, s)$ близким к нему вырожденным ядром $\tilde{K}(x, s)$ и последующее решение этого уравнения с вырожденным ядром изложенным выше способом.

Укажем несколько способов такой замены.

Разложение ядра в ряд Тейлора

- Если ядро $K(x, s)$ обладает достаточной гладкостью по переменной x на отрезке $[a, b]$, то в качестве вырожденного ядра можно взять соответствующей длины отрезок ряда Тейлора по x :

$$\tilde{K}(x, s) = \sum_{i=0}^n \frac{(x - x_0)^i}{i!} \frac{\partial^i K(x_0, s)}{\partial x^i}, \quad (7.27)$$

где x_0 – некоторая точка из отрезка $[a, b]$ (ее выбор может быть подчинен, например, требованию минимизации остатка ряда).

Очевидно, в данном случае ядро имеет вид (7.22), в котором

$$\alpha_i(x) = (x - x_0)^i, \quad \beta_i(s) = \frac{1}{i!} \frac{\partial^i K(x_0, s)}{\partial x^i}.$$

- Если ядро $K(x, s)$ достаточно гладкое по переменной s на $[a, b]$, то аналогично можем применить разложение в ряд Тейлора по переменной s :

$$\tilde{K}(x, s) = \sum_{i=0}^m \frac{(s - s_0)^i}{i!} \frac{\partial^i K(x, s_0)}{\partial s^i}. \quad (7.28)$$

При таком способе замены

$$\alpha_i(x) = \frac{1}{i!} \frac{\partial^i K(x, s_0)}{\partial s^i}, \quad \beta_i(s) = (s - s_0)^i.$$



3. Для построения вырожденного ядра можно также использовать конечный отрезок двойного ряда Тейлора:

$$\tilde{K}(x, s) = \sum_{i=0}^n \frac{1}{i!} \left[(x - x_0) \frac{\partial}{\partial x} + (s - s_0) \frac{\partial}{\partial s} \right]^i K(x_0, s_0), \quad (x_0, s_0 \in [a, b]). \quad (7.29)$$

В последнем случае, очевидно, нам всегда гарантировано точное вычисление интегралов во второй из формул (7.26).

Использование ортогональных разложений

Рассмотрим этот прием на примере применения ряда Фурье. Известно, что непрерывная на отрезке $[-l, l]$ функция допускает разложение в ряд Фурье (например, по косинусам, если она четна). Поэтому можно положить (в предположении, что ядро $K(x, s)$ непрерывно по x и четно по этой же переменной)

$$\tilde{K}(x, s) = \frac{1}{2} a_0(s) + \sum_{i=1}^n a_i(s) \cos \frac{i\pi x}{l}, \quad (7.30)$$

где

$$a_i(s) = \frac{2}{l} \int_0^l K(x, s) \cos \frac{i\pi x}{l} dx, \quad i = \overline{0, n}.$$

При этом следует иметь в виду, что исходный отрезок $[a, b]$ линейной заменой может быть превращен в нужный. Аналогичные формулы можно записать, если использовать отрезки рядов Фурье по переменной s , либо по обеим переменным.

Интерполяционные способы замены ядра

Рассмотрим для примера применение алгебраического интерполирования по значениям функции. Выбрав на отрезке $[a, b]$ ($n+1$) точек $a \leq x_0 < x_1 < \dots < x_n \leq b$ (узлов интерполирования), можем записать:

$$\tilde{K}(x, s) = \sum_{i=0}^n \frac{\omega_{n+1}(x)}{(x - x_i) \omega'_{n+1}(x_i)} K(x_i, s). \quad (7.31)$$



Аналогично может быть использована замена ядра интерполяционным многочленом по переменной s :

$$\tilde{K}(x, s) = \sum_{j=0}^m \frac{\omega_{m+1}(s)}{(s - s_j) \omega'_{m+1}(s_j)} K(x, s_j). \quad (7.32)$$

Иногда также целесообразным бывает использование интерполирования по обеим переменным. Так, например, использование повторного интерполирования приводит к формуле

$$\tilde{K}(x, s) = \sum_{i=0}^n \sum_{j=0}^m \frac{\omega_{n+1}(x) \omega_{m+1}(s)}{(x - x_i)(s - s_j) \omega'_{n+1}(x_i) \omega'_{m+1}(s_j)} K(x_i, s_j). \quad (7.33)$$

При этом, очевидно, вместо представлений Лагранжа могут быть использованы и другие.

Способ Бэтмена

Аппроксимирующее ядро $\tilde{K}(x, s)$ предлагается определять с помощью равенства

$$\begin{vmatrix} \tilde{K}(x, s) & K(x, s_0) & \dots & K(x, s_n) \\ K(x_0, s) & K(x_0, s_0) & \dots & K(x_0, s_n) \\ & & \ddots & \\ K(x_n, s_0) & K(x_n, s_0) & \dots & K(x_n, s_n) \end{vmatrix} = 0,$$

где $x_0, \dots, x_n, s_0, \dots, s_n$ – некоторые точки из отрезка $[a, b]$. Представляя элементы первого столбца записанного определителя в виде

$$\tilde{K}(x, s) + 0, 0 + K(x_0, s), \dots, 0 + K(x_n, s)$$



и разлагая определитель в сумму двух определителей, получим:

$$\tilde{K}(x, s) = -\frac{1}{\Delta} \begin{vmatrix} 0 & K(x, s_0) & \dots & K(x, s_n) \\ K(x_0, s) & K(x_0, s_0) & \dots & K(x_0, s_n) \\ & \dots & & \\ K(x_n, s_0) & K(x_n, s_0) & \dots & K(x_n, s_n) \end{vmatrix}, \quad (7.34)$$

где

$$\Delta = \begin{vmatrix} K(x_0, s_0) & \dots & K(x_0, s_n) \\ & \dots & \\ K(x_n, s_0) & \dots & K(x_n, s_n) \end{vmatrix}. \quad (7.35)$$



7.2.3. Метод последовательных приближений

Вновь рассмотрим интегральное уравнение Фредгольма второго рода (7.8). Будем искать его решение в виде степенного ряда

$$u(x) = \sum_{i=0}^{\infty} \lambda^i \varphi_i(x), \quad (7.36)$$

где λ – числовой параметр из уравнения (7.8), а функции $\varphi_i(x)$ подлежат определению.

Подставим ряд (7.36) в исходное интегральное уравнение (7.8).

$$\sum_{i=0}^{\infty} \lambda^i \varphi_i(x) - \lambda \int_a^b K(x, s) \sum_{i=0}^{\infty} \lambda^i \varphi_i(s) ds = f(x).$$

Меняя порядок суммирования и интегрирования (в предположении, что ряд (7.36) сходится) и приравнивая коэффициенты при одинаковых степенях λ , получим рекуррентные соотношения, позволяющие последовательно находить (быть может, приближенно) функциональные коэффициенты ряда (7.36):

$$\begin{cases} \varphi_0(x) = f(x), \\ \varphi_i(x) = \int_a^b K(x, s) \varphi_{i-1}(s) ds, \quad i = 1, 2, \dots \end{cases} \quad (7.37)$$

Таким образом, алгоритм построения последовательности приближений определен. Исследуем его сходимость. Пусть в области $R = [a, b] \times [a, b]$ выполняется неравенство $|K(x, s)| \leq M$, и на отрезке $[a, b]$ – неравенство $|f(x)| \leq N$. Тогда из формул (7.37) последовательно получим:

$$|\varphi_0(x)| = |f(x)| \leq N,$$

$$|\varphi_1(x)| = \left| \int_a^b K(x, s) \varphi_0(s) ds \right| \leq \int_a^b |K(x, s)| |\varphi_0(s)| ds \leq NM(b-a),$$



$$|\varphi_2(x)| = \left| \int_a^b K(x, s) \varphi_1(s) ds \right| \leq \int_a^b |K(x, s)| |\varphi_1(s)| ds \leq NM^2(b-a)^2,$$

.....

$$|\varphi_i(x)| \leq NM^i(b-a)^i, \quad i = 0, 1, 2, \dots$$

Учитывая полученные оценки, видим, что ряд (7.36) мажорируется числовым рядом $N \sum_{i=0}^{\infty} (|\lambda| M(b-a))^i$, представляющим собой геометрическую прогрессию, и, следовательно, сходящимся при выполнении условия

$$|\lambda| M(b-a) < 1. \quad (7.38)$$

Таким образом, если параметры исходного интегрального уравнения будут удовлетворять условию (7.38), то ряд (7.36) равномерно сходится на отрезке $[a, b]$. Тогда в качестве приближенного решения можно взять

$$y(x) = y_n(x) = \sum_{i=0}^n \lambda^i \varphi_i(x). \quad (7.39)$$

Оценим погрешность такого решения (в предположении, что все интегралы в (7.37) вычисляются точно). Имеем:

$$|\varepsilon_n(x)| = |u(x) - y_n(x)| = \left| \sum_{i=n+1}^{\infty} \lambda^i \varphi_i(x) \right| \leq N |\lambda|^{n+1} M^{n+1} (b-a)^{n+1} (1 + |\lambda| M(b-a) + \dots) =$$

$$(7.40)$$

$$= N [|\lambda| M(b-a)]^{n+1} \frac{1}{1-|\lambda|M(b-a)}, \quad x \in [a, b].$$

Из этой оценки следует равномерная сходимость $\varepsilon_n(x)$ к нулю. Отсюда видно также, что $y_n(x) \xrightarrow[n \rightarrow \infty]{} u(x)$ по крайней мере со скоростью геометрической прогрессии.

Заметим, однако, что все нужные интегралы в (7.37), как правило, вычисляются приближенно, поэтому оценка истинной погрешности будет несколько отличаться от полученной.



Следует также иметь в виду, что метод последовательных приближений может употребляться и в другой форме, несколько более удобной с точки зрения машинной реализации. Действительно, перепишем (7.8) в виде

$$u(x) = \lambda \int_a^b K(x, s) u(s) ds + f(x).$$

Получим (см. также [аналогичный метод решения алгебраических уравнений](#)) вид, удобный для итерации. Выбирая в качестве начального приближения произвольную функцию $y_0(x)$, построим последовательность приближений

$$y_{n+1}(x) = \lambda \int_a^b K(x, s) y_n(s) ds + f(x), \quad n = 0, 1, \dots, \quad (7.41)$$

которая при $y_0(x) \equiv 0$ будет полностью совпадать с (7.37), (7.39).

В то же время, запись метода последовательных приближений в форме (7.41) позволяет трактовать условие сходимости (7.40) как условие сжимаемости отображения $\varphi(u) = \lambda \int_a^b K(x, s) u(s) ds$. С другой стороны, процедура исследования сходимости метода последовательных приближений в форме (7.37), (7.39) говорит о том, что условие (7.40) является достаточным.



7.2.4. Методы решения интегральных уравнений Вольтерра второго рода

[Метод механических квадратур](#)

[Метод последовательных приближений](#)

Метод механических квадратур

Как мы уже отмечали, чисто формально уравнение Вольтерра можно считать частным случаем интегрального уравнения Фредгольма, у которого $K(x, s) \equiv 0$ при $s > x$. Поэтому алгоритм метода механических квадратур, рассмотренный нами [выше](#), может рассматриваться и как алгоритм решения интегрального уравнения Вольтерра второго рода. Однако на таком пути мы встретимся с теоретическими трудностями при оценке погрешности, так как при таком подходе ядро $\tilde{K}(x, s)$ оказывается разрывной функцией и для величины ρ_K не может быть получено сколько-нибудь удовлетворительных оценок.

Поэтому повторим вывод алгоритма, несколько его видоизменив. Рассмотрим интегральное уравнение Вольтерра второго рода

$$u(x) - \lambda \int_a^x K(x, s) u(s) ds = f(x), \quad x \in [a, b]. \quad (7.42)$$

Выберем на отрезке $[a, b]$ ($n + 1$) точек $a \leq x_0 < x_1 < \dots < x_n \leq b$ и рассмотрим уравнение (7.42) в этих точках:

$$u(x_i) - \lambda \int_a^{x_i} K(x_i, s) u(s) ds = f(x_i), \quad i = \overline{0, n}. \quad (7.43)$$

Интеграл в (7.43) представляет собой интеграл с постоянными (!) пределами. Заменим его квадратурной суммой, использующей значения подынтегральной функции в точках x_0, x_1, \dots, x_i :

$$u(x_i) - \lambda \sum_{k=0}^i A_K^{(i)} K(x_i, x_k) u(x_k) = f(x_i) + \lambda \rho^{(i)}(x_i). \quad (7.44)$$



Здесь $A_k^{(i)}$ – коэффициенты выбранной квадратурной формулы (принципиально для каждого значения i возможно использование своей квадратурной формулы), а $\rho^{(i)}(x)$ – ее остаток. Отбрасывая в (7.44) остаточный член, получим систему линейных алгебраических уравнений для определения приближенных значений решения интегрального уравнения (7.42) в узлах выбранной сетки (используем те же обозначения, что и в формуле (7.11)):

$$y_i - \lambda \sum_{k=0}^i K_{ik} y_k = f_i, \quad i = \overline{0, n},$$

или в развернутом виде

$$\left\{ \begin{array}{l} \left(1 - \lambda A_0^{(0)} K_{00}\right) y_0 = f_0, \\ -\lambda A_0^{(1)} K_{10} y_0 + \left(1 - \lambda A_1^{(1)} K_{11}\right) y_1 = f_1, \\ \dots \\ -\lambda A_0^{(n)} K_{n0} y_0 - \dots - \lambda A_{n-1}^{(n)} K_{nn-1} y_{n-1} + \left(1 - \lambda A_n^{(n)} K_{nn}\right) y_n = f_n. \end{array} \right. \quad (7.45)$$

Матрица данной системы является нижней треугольной. Поэтому решение системы (7.45) по сути представляет собой обратный ход метода Гаусса.

Метод последовательных приближений

Аналогично описанному [выше](#) методу для уравнения Фредгольма решение уравнения (7.42) может быть получено в виде ряда

$$u(x) = \sum_{i=0}^{\infty} \lambda^i \varphi_i(x), \quad (7.46)$$



где $\varphi_i(x)$ на сей раз определяются по формулам

$$\begin{cases} \varphi_0(x) = f(x), \\ \varphi_i(x) = \int_a^x K(x, s) \varphi_{i-1}(s) ds, \quad i = 1, 2, \dots \end{cases} \quad (7.47)$$

Таким образом, до сих пор все фактически совпадает с алгоритмом метода последовательных приближений для интегральных уравнений Фредгольма второго рода.

Исследуем сходимость полученного алгоритма, используя для этих целей тот же прием построения мажоранты, что и выше.

Если, как и ранее, предположить, что в области $R = [a, b] \times [a, b]$ выполняется неравенство $|K(x, s)| \leq M$, и на отрезке $[a, b]$ — неравенство $|f(x)| \leq N$, то

$$|\varphi_0(x)| = |f(x)| \leq N, \quad a \leq x \leq b,$$

$$|\varphi_1(x)| = \left| \int_a^x K(x, s) \varphi_0(s) ds \right| \leq \int_a^x |K(x, s)| |\varphi_0(s)| ds \leq NM(x-a), \quad a \leq x \leq b,$$

$$|\varphi_2(x)| = \left| \int_a^x K(x, s) \varphi_1(s) ds \right| \leq \int_a^x |K(x, s)| |\varphi_1(s)| ds \leq NM^2 \int_a^x (s-a) ds \leq NM^2 \frac{(x-a)^2}{2!}, \quad a \leq x \leq b,$$

.....

$$|\varphi_i(x)| \leq NM^i \frac{(x-a)^i}{i!}, \quad a \leq x \leq b, \quad i = 0, 1, 2, \dots$$

Таким образом, ряд (7.12) в случае интегрального уравнения Вольтерра второго рода будет мажорироваться степенным рядом $N \sum_{i=0}^{\infty} \frac{|(\lambda|M(x-a))|^i}{i!}$, который, как известно, сходится при любом x и λ (к функции $N e^{|\lambda|M(x-a)}$), т.е. в отличие от интегральных уравнений Фредгольма второго рода сходимость метода последовательных приближений для уравнения (7.42) не накладывает ограничений на количественные характеристики параметров исходной задачи.



Для погрешности $\varepsilon_n(x)$ приближенного решения

$$y_n(x) = \sum_{i=0}^n \lambda^i \varphi_i(x)$$

получим:

$$|\varepsilon_n(x)| = \left| \sum_{i=n+1}^{\infty} \lambda^i \varphi_i(x) \right| \leq N \frac{|\lambda|^{n+1} M^{n+1} (x-a)^{n+1}}{(n+1)!} \left(1 + \frac{|\lambda|M(x-a)}{n+2} + \dots \right) \leq [q = |\lambda| M(b-a)] \leq$$

$$\leq N \frac{q^{n+1}}{(n+1)!} \left(1 + \frac{q}{n+2} + \frac{q^2}{(n+2)(n+3)} + \dots \right) < N \frac{q^{n+1}}{(n+1)!} \left[1 + \frac{q}{n+2} + \frac{q^2}{(n+2)^2} + \dots \right] =$$

$$= N \frac{q^{n+1}}{(n+1)!} \cdot \frac{1}{1 - \frac{q}{n+2}}, \quad a \leq x \leq b, \quad n > q - 2.$$



7.3. Проекционные методы решения интегральных уравнений

7.3.1. Метод моментов и метод Галеркина

7.3.2. Другие проекционные методы



7.3.1. Метод моментов и метод Галеркина

Связь метода Галеркина с заменой ядра вырожденным

Зададим две системы функций:

1) $\varphi_0(x), \varphi_1(x), \dots$

Эта система составляет базис подпространств, в которые проектируется решение исходного интегрального уравнения. Поэтому требования, предъявляемые к ней, должны быть такими:

- При любом значении i функции $\varphi_i(x)$ непрерывны;
- При любом конечном значении n система $\{\varphi_i(x)\}_{i=1}^n$ является линейно независимой;
- Система $\{\varphi_i(x)\}$ обладает свойством C -полноты на множестве непрерывных функций (это означает следующее: для любой функции $F(x) \in C[a, b]$ и любого $\varepsilon > 0$ существует n и набор коэффициентов c_0, c_1, \dots, c_n такие, что $|F(x) - \sum_{i=0}^n c_i \varphi_i(x)| < \varepsilon$ для всех $x \in [a, b]$).

2) $\psi_0(x), \psi_1(x), \dots$

Эту систему будем использовать для возможно лучшего в каком-то смысле выполнения исходного уравнения на приближенном решении. Будем предполагать все функции этой системы непрерывными, линейно независимыми, а саму систему – замкнутой на множестве $C[a, b]$. Применительно к нашим целям это означает, что если $\int_a^b f(x) \psi_i(x) dx = 0, i = 0, 1, \dots$, то отсюда с необходимостью следует, что $f(x) \equiv 0$ (используя терминологию скалярного произведения, это можно переписать следующим образом: $(f, \psi_i) = 0, i = 0, 1, \dots \Leftrightarrow f \equiv 0$).

Составим линейную комбинацию

$$u_n(x) = f(x) + \sum_{i=0}^n c_i \varphi_i(x). \quad (7.48)$$

Она содержит $(n+1)$ произвольных коэффициентов. Опишем требования, на основании которых можно осуществить их выбор.



Переписав исходную задачу в виде

$$Lu \equiv u(x) - \lambda \int_a^b K(x, s) u(s) ds = f(x), \quad (7.49)$$

в силу замкнутости системы $\{\psi_j(x)\}$ получим требование: чтобы $u_n(x)$ была решением уравнения (7.49), т.е. чтобы имело место тождество $Lu_n - f \equiv 0$, необходимо и достаточно выполнение бесконечного множества равенств

$$(Lu_n - f, \psi_j) = 0, \quad j = 0, 1, 2, \dots$$

Однако в нашем распоряжении имеется лишь $(n + 1)$ коэффициентов c_i , выбором которых мы можем распоряжаться. Следовательно, мы имеем возможность удовлетворить только первым $(n + 1)$ из выписанных условий:

$$(Lu_n - f, \psi_j) = 0, \quad j = 0, 1, \dots, n. \quad (7.50)$$

Условия (7.50) дают систему линейных алгебраических уравнений для определения коэффициентов c_i . Запишем ее более подробно. Так как оператор L линейный, то $Lu_n = Lf + \sum_{i=0}^n c_i L\varphi_i$. Поэтому из (7.50) получим:

$$\left(Lf + \sum_{i=0}^n c_i L\varphi_i - f, \psi_j \right) = 0, \quad j = 0, 1, \dots, n$$

или (учитывая также и линейность скалярного произведения)

$$\sum_{i=0}^n c_i (L\varphi_i, \psi_j) = (f - Lf, \psi_j), \quad j = 0, 1, \dots, n.$$

Таким образом, система для нахождения коэффициентов линейной комбинации (7.48), определяющей приближенное решение, имеет вид

$$\sum_{i=0}^n a_{ji} c_i = b_j, \quad j = 0, 1, \dots, n, \quad (7.51)$$



где

$$\begin{aligned}
 a_{ji} &= (L\varphi_i, \psi_j) = \int_a^b \left[\varphi_i(x) - \lambda \int_a^b K(x, s) \varphi_i(s) ds \right] \psi_j(x) dx = \\
 &= \int_a^b \varphi_i(x) \psi_j(x) dx - \lambda \int_a^b \int_a^b K(x, s) \varphi_i(s) \psi_j(x) ds dx,
 \end{aligned} \tag{7.52}$$

$$\begin{aligned}
 b_j &= (f - Lf, \psi_j) = \int_a^b \left[f(x) - f(x) + \lambda \int_a^b K(x, s) f(s) ds \right] \psi_j(x) dx = \\
 &= \lambda \int_a^b \int_a^b K(x, s) f(s) \psi_j(x) ds dx.
 \end{aligned}$$

При сформулированных выше требованиях, предъявляемых к системам функций $\{\varphi_i(x)\}$ и $\{\psi_i(x)\}$ система (7.51), (7.52) будет иметь единственное решение. Найдя его тем или иным способом, построим приближенное решение $u_n(x)$.

Если системы $\{\varphi_i(x)\}$ и $\{\psi_i(x)\}$ совпадают, то получаем алгоритм метода Галеркина. В дальнейшем будем говорить именно о нем.

Связь метода Галеркина с заменой ядра вырожденным

Покажем, что применение метода Галеркина равносильно замене ядра $K(x, s)$ вырожденным, строящимся некоторым специальным образом. Действительно, предполагая ортонормированность системы $\{\varphi_i(x)\}$ (если это не так, то всегда можно применить процедуру ортогонализации), разложим ядро $K(x, s)$ как функцию переменной x в ряд Фурье по этой системе и за $\tilde{K}(x, s)$ примем n -ю частичную сумму этого ряда. Получим:

$$\tilde{K}(x, s) = \sum_{i=0}^n \beta_i(s) \varphi_i(x),$$



где

$$\beta_i(s) = \int_a^b K(x, s) \varphi_i(x) dx.$$

Если теперь для уравнения

$$u(x) - \lambda \int_a^b \tilde{K}(x, s) u(s) ds = f(x)$$

записать (в соответствии с теорией уравнений с вырожденным ядром) точное решение

$$u(x) = f(x) + \lambda \sum_{i=0}^n c_i \varphi_i(x),$$

то для определения коэффициентов c_i имеем систему (см. также формулы (7.58), (7.59))

$$c_i - \lambda \sum_{j=0}^n a_{ij} c_j = b_i, \quad i = 0, 1, \dots, n,$$

где

$$b_i = \int_a^b f(s) \beta_i(s) ds = \int_a^b \int_a^b K(x, s) f(s) \varphi_i(x) ds dx,$$

$$a_{ij} = \int_a^b \beta_i(s) \varphi_j(s) ds = \int_a^b \int_a^b K(x, s) \varphi_i(s) \varphi_j(x) ds dx,$$

совпадающую с (7.51), (7.52) при $\psi_i(x) \equiv \varphi_i(x)$, $i = 0, 1, \dots$.

Замечание 7.2. Возможна организация работы по описанным выше методам с несколько иным представлением приближенного решения, немного отличающимся от (7.48):

$$u_n(x) = \sum_{i=0}^n c_i \varphi_i(x). \tag{7.53}$$



В этом случае система для определения коэффициентов c_i также будет иметь вид (7.51) с той лишь разницей, что для коэффициентов b_j необходимо использовать формулу

$$b_j = (f, \psi_j) = \int_a^b f(x) \psi_j(x) dx. \quad (7.54)$$



7.3.2. Другие проекционные методы

[Метод наименьших квадратов](#)

[Метод коллокации](#)

Метод наименьших квадратов

Рассмотрим вновь интегральное уравнение (7.49) предыдущего параграфа. Легко видеть, что решение этой задачи эквивалентно задаче нахождения минимума функционала

$$J(u) = (Lu - f, Lu - f), \quad (7.55)$$

(где, по-прежнему, $(f, g) = \int_a^b f(x)g(x)dx$ обозначает скалярное произведение) и, следовательно, может быть заменено последней, причем, в соответствии с общей идеей проекционных методов, минимум будем искать в подпространствах конечной размерности.

Задавая систему функций $\{\varphi_i(x)\}$, описанную в предыдущем параграфе (с такими же свойствами), приближенное решение будем искать в виде

$$u_n(x) = \sum_{i=0}^n c_i \varphi_i(x). \quad (7.56)$$

Функционал (7.55) на приближенном решении примет вид

$$J(u_n) = \left(\sum_{i=0}^n c_i L\varphi_i - f, \sum_{i=0}^n c_i L\varphi_i - f \right)$$

и будет являться функцией переменных c_0, c_1, \dots, c_n (коэффициентов комбинации (7.56)).

Записывая необходимое условие минимума первого порядка, получим:

$$\frac{\partial J(u_n)}{\partial c_j} = 2 \left(\sum_{i=0}^n c_i L\varphi_i - f, L\varphi_j \right) = 0, \quad j = 0, 1, \dots, n.$$



Таким образом, имеем систему линейных алгебраических уравнений для определения коэффициентов c_0, c_1, \dots, c_n :

$$\sum_{i=0}^n a_{ji} c_i = b_j, \quad j = 0, 1, \dots, n, \quad (7.57)$$

где

$$a_{ji} = (L\varphi_i, L\varphi_j) = \int_a^b \left[\varphi_i(x) - \lambda \int_a^b K(x, s) \varphi_i(s) ds \right] \left[\varphi_j(x) - \lambda \int_a^b K(x, s) \varphi_j(s) ds \right] dx, \quad (7.58)$$

$$b_j = (f, L\varphi_j) = \int_a^b f(x) \left[\varphi_j(x) - \lambda \int_a^b K(x, s) \varphi_j(s) ds \right] dx.$$

Метод коллокации

Как и выше, решаем уравнение $Lu(x) = f(x)$. В соответствии с общей идеологией проекционных методов приближенное решение будем искать в виде (7.56) с использованием введенной ранее системы функций $\{\varphi_i(x)\}$.

По сути, конечным «продуктом» метода является опять-таки система линейных алгебраических уравнений. Отличия состоят лишь в способе ее получения (и, как следствие – в виде элементов матрицы и свободных членов). Здесь для этих целей используется идея обращения в нуль невязки уравнения (7.49) для приближенного решения на выбранном множестве точек. Таким образом, требуется выполнение соотношений

$$Lu_j(x_j) = f(x_j), \quad j = 0, 1, \dots, n \quad (7.59)$$

или

$$\sum_{i=0}^n a_{ji} c_i = f(x_j), \quad j = 0, 1, \dots, n, \quad (7.60)$$

где

$$a_{ji} = L\varphi_i(x_j) = \varphi_i(x_j) - \lambda \int_a^b K(x_j, s) \varphi_i(s) ds. \quad (7.61)$$



Чтобы система (7.61) имела единственное решение, необходимо потребовать отлиния от нуля ее определителя:

$$\Delta = \begin{vmatrix} L\varphi_0(x_0) & \dots & L\varphi_n(x_0) \\ \dots & \dots & \dots \\ L\varphi_0(x_n) & \dots & L\varphi_n(x_n) \end{vmatrix} \neq 0.$$

Как мы помним, в этом случае теоретически достаточно выполнения требования (которое, однако, на практике совсем не просто выполнить), чтобы система $\{L\varphi_i(x)\}$ являлась [системой функций Чебышева](#) на отрезке $[a, b]$. Это можно считать дополнительными ограничениями, накладываемыми на систему $\{\varphi_i(x)\}$.

В качестве общего замечания отметим следующее:

- 1) Для решения интегральных уравнений можно применять и другие идеи (например, использовать для этих целей [сплайн-приближения](#));
- 2) Практически все изложенные методы (за исключением, разве что, [метода замены ядра на в вырожденное](#), в котором явно используется линейность задачи) можно применять и для решения других типов интегральных уравнений (в том числе – нелинейных) с соответствующими изменениями в алгоритме.



7.4. Решение интегральных уравнений первого рода

[7.4.1. Основные определения и примеры](#)

[7.4.2. Метод регуляризации решения некорректных задач](#)



7.4.1. Основные определения и примеры

Большинство [некорректных](#) задач может быть приведено к операторному уравнению первого рода, имеющему вид

$$Au = f, \quad u \in U, \quad f \in F, \quad (7.62)$$

в котором по заданному оператору A (не обязательно линейному), действующему из пространства U в пространство F , и по заданному элементу $f \in F$ требуется определить решение $u \in U$.

В частном случае имеем

1. Интегральное уравнение Фредгольма первого рода:

$$\int_a^b K(x, s) u(s) ds = f(x), \quad c \leq x \leq d; \quad (7.63)$$

2. Интегральное уравнение Вольтерра первого рода

$$\int_a^x K(x, s) u(s) ds = f(x), \quad c \leq x \leq; \quad (7.64)$$

По [определению](#) корректность задачи связана с наличием обратного оператора A^{-1} , определенного и непрерывного на всем пространстве F .

Простейшим примером некорректно поставленной задачи могут служить системы линейных алгебраических уравнений. Запишем соответствующую задачу в матричном виде

$$Au = f.$$

Теоретически возможны следующие ситуации:

- 1) A – квадратная матрица, причем $\det A = 0$. Тогда решений либо не существует вовсе, либо их будет бесконечно много. Таким образом, нарушаются либо условие а), либо условие б);



- 2) A – квадратная матрица, но $\det A \neq 0$. Тогда система имеет единственное решение, но при нарушении условия в) имеем плохо обусловленную задачу, сложность решения которой отмечалась ранее;
- 3) A – прямоугольная матрица. Тогда, как это следует из общей теории линейных систем, также нарушаются условия а) и б) из определения корректности.

Покажем сейчас, что интегральное уравнение Фредгольма первого рода (7.63) также является некорректной задачей.

Рассмотрим более простой для исследования частный случай. Пусть ядро $K(x, s)$ вещественно и симметрично, т.е. $K(x, s) = K(s, x)$. Предположим также, что $K(x, s)$ и $f(x)$ непрерывны. Тогда, как известно (см., например, [2]), существует полная ортонормированная система собственных функций $\varphi_i(x)$ оператора A :

$$A\varphi_i(x) = \int_a^b K(x, s) \varphi_i(s) ds = \lambda_i \varphi_i(x), \quad i = 0, 1, \dots$$

$$(\varphi_i, \varphi_j) = \int_a^b \varphi_i(s) \varphi_j(s) ds = \delta_i^j.$$

При этом ядро $K(x, s)$ раскладывается в сходящийся ряд по собственным функциям (ряд Фурье)

$$K(x, s) = \sum_{i=0}^{\infty} \lambda_i \varphi_i(x) \varphi_i(s),$$

где сходимость ряда в правой части понимается в L_2 -норме:

$$\|K(x, s)\| = \sqrt{\int_a^b \int_c^d |K(x, s)|^2 dx ds}.$$

Отсюда, в частности, следует, что $\|K\|^2 = \sum_{i=0}^{\infty} |\lambda_i|^2$ и, следовательно, $\lambda_i \xrightarrow{i \rightarrow \infty} 0$.



Рассмотрим случай, когда $\lambda_i \neq 0$ при $0 \leq i \leq N$ и все $\lambda_i = 0$ при $i > N$. Тогда ядро имеет вид

$$K(x, s) = \sum_{i=0}^N \lambda_i \varphi_i(x) \varphi_i(s),$$

т.е. является вырожденным и уравнение (7.63) может быть переписано в виде

$$\int_a^b K(x, s) u(s) ds = \sum_{i=0}^N \lambda_i \int_a^b \varphi_i(x) \varphi_i(s) u(s) ds = \sum_{i=0}^N \lambda_i (\varphi_i, u) \varphi_i(x) = f(x).$$

Отсюда следует, что задача может иметь решение только в том случае, когда $f(x)$ является линейной комбинацией функций $\varphi_0(x), \dots, \varphi_N(x)$, т.е. записывается в виде

$$f(x) = \sum_{i=0}^N f_i \varphi_i(x).$$

Легко видеть, что решением в этом случае является функция

$$u(x) = u_0(x) = \sum_{i=0}^N \frac{f_i}{\lambda_i} \varphi_i(x).$$

Действительно, если искать решение $u(x)$ в виде (в полном согласии с теорией уравнений с вырожденным ядром)

$$u(x) = \sum_{i=0}^N C_i \varphi_i(x),$$

то

$$(u, \varphi_k) = C_k$$

и тогда

$$\int_a^b K(x, s) u(s) ds = \sum_{i=0}^N \lambda_i (\varphi_i, u) \varphi_i(x) = \sum_{i=0}^N C_i \lambda_i \varphi_i(x) = f(x) = \sum_{i=0}^N f_i \varphi_i(x),$$



откуда $C_i = \frac{f_i}{\lambda_i}$.

В то же время, любая функция $u(x)$, представимая в виде

$$u(x) = u_0(x) + \sum_{i=N+1}^{\infty} C_i \varphi_i(x),$$

где $\sum_{i=N+1}^{\infty} |C_i|^2 < +\infty$, также будет решением уравнения (7.63).

Таким образом, в рассматриваемом случае задача (7.63) может не иметь решения; в случае же, когда это решение существует, оно не единственno.

Совершенно аналогично можно рассмотреть случай, когда все собственные значения λ_i отличны от нуля. Решение в этом случае представимо в виде ряда

$$u(x) = \sum_{i=0}^{\infty} \frac{f_i}{\lambda_i} \varphi_i(x),$$

где, опять-таки, f_i – коэффициенты разложения $f(x)$ в ряд по собственным функциям $\varphi_i(x)$. Если $f(x) \in L_2[a, b]$, то такой ряд будет сходиться в норме пространства L_2 , т.е. решение в этом случае будет существовать и окажется единственным (два решения могут не совпадать на множестве меры нуль).

Но в то же время, если правые части уравнения (7.63) $f^n(x)$, $n = 0, 1, 2, \dots$ сходятся к некоторой функции $f(x)$ в пространстве $L_2[a, b]$, т.е.

$$\| f^n - f \| ^2 = \sum_{i=0}^{\infty} (f_i^n - f_i)^2 \xrightarrow{n \rightarrow \infty} 0,$$

то норма разности соответствующих решений уравнения (7.63), выражаемая равенством

$$\| u^n - u \| ^2 = \sum_{i=0}^{\infty} \frac{(f_i^n - f_i)^2}{\lambda_i^2},$$

не только не обязана стремиться к нулю, но и может быть бесконечно большой. В этом легко убедиться, положив $f^n(x) = f(x) + \sqrt{|\lambda_n|} \varphi_n(x)$. Тогда

$$\| f^n - f \| ^2 = \left\| \sqrt{|\lambda_n|} \varphi_n \right\|^2 = |\lambda_n| \xrightarrow{n \rightarrow \infty} 0.$$



В то же время

$$\|u^n - u\|^2 = \left\| u + \frac{\sqrt{|\lambda_n|}}{|\lambda_n|} \varphi_n - u \right\|^2 = \frac{1}{|\lambda_n|} \xrightarrow{n \rightarrow \infty} \infty.$$

Следовательно, устойчивость решений отсутствует.

Таким образом, в рассмотренных частных случаях возможны нарушения всех трех условий корректности. Отметим также, что для уравнения (7.64) справедливы аналогичные результаты. При этом, в частности, простейшее интегральное уравнение Вольтерра первого рода

$$\int_a^x u(s) ds = f(x) - f(a) \quad (7.65)$$

эквивалентно задаче нахождения производной, поскольку решением (7.65) (в случае если $f(x)$ дифференцируема!) является функция $u(x) = f'(x)$. О некорректности последней мы уже упоминали ранее.



7.4.2. Метод регуляризации решения некорректных задач

[Вариационный метод регуляризации](#)

[Выбор параметров регуляризации](#)

[Уравнение Эйлера](#)

[Замечание о решении плохо обусловленных линейных систем](#)

Как следует из изложенного выше, непосредственно решать некорректно поставленные задачи при неточно заданной правой части бессмысленно. Если $\bar{f}(x)$ задана с погрешностью $\delta f(x)$, то соответствующее решение $u_\delta(x)$ или не существует, или отличается от искомого решения $\bar{u}(x)$ на величину $\delta u(x)$, которая может быть большой.

Даже если $f(x)$ задана точно, но отыскание решения выполняется численными методами, то неизбежно вносится погрешность метода и округления. Это снова приводит к большой погрешности решения $\delta u(x)$.

Однако никто не обязывает нас непосредственно решать исходную задачу

$$Au = f \tag{7.66}$$

с возмущенной правой частью. Всегда можно попытаться заменить эту задачу некоторой «близкой» задачей, решение которой будет «близко» к $u(x)$. Символически запишем измененную задачу в виде

$$A_\alpha u_\alpha = f, \tag{7.67}$$

где $\alpha > 0$ – некоторый параметр (параметр регуляризации), а ее решение будем обозначать $u_\alpha(x)$.

Определение. Оператор A_α называют [регуляризирующим](#), если:

- 1) задача (7.67) является корректно поставленной в классе правых частей F при любом $\alpha > 0$;
- 2) существуют такие функции $\alpha(\delta)$ и $\delta(\varepsilon)$, что если $\|f - \bar{f}\|_F \leq \delta(\varepsilon)$, то $\|u_{\alpha(\delta)} - \bar{u}\|_U \leq \varepsilon$.



Таким образом, если найден регуляризующий оператор A_α , то задача (7.67) имеет решения при любых $f \in F$, в том числе и отличающихся от \bar{f} на любого вида погрешность δf ; эта задача устойчива, так что ее можно решать обычными численными методами. При правильно подобранном параметре регуляризации α ее решение $u_\alpha(x)$ достаточно мало отличается от нужного нам решения $\bar{u}(x)$ исходной задачи (7.66).

Вариационный метод регуляризации

Рассмотрим уравнение Фредгольма первого рода

$$\int_a^b K(x, s) u(s) ds = f(x), \quad c \leq x \leq d. \quad (7.68)$$

Будем считать, что ядро его непрерывно и таково, что в случае $f(x) \equiv 0$ имеет только тривиальное решение $u(x) \equiv 0$. Тогда при любой правой части $f(x) \in F$ решение либо единственное, либо не существует. Тем самым интегральный оператор

$$A(x, u(s)) = \int_a^b K(x, s) u(s) ds \quad (7.69)$$

отображает U в F взаимно однозначно.

Исходную задачу (7.68) можно заменить эквивалентной вариационной задачей

$$\int_c^d [A(x, u(s)) - f(x)]^2 dx \rightarrow \min. \quad (7.70)$$

Рассмотрим измененную задачу

$$M(\alpha, f(x), u(s)) = \int_c^d [A(x, u(s)) - f(x)]^2 dx + \alpha \Omega(u(s)) \rightarrow \min, \quad (7.71)$$



где $\Omega(u(s))$ – так называемый *тихоновский стабилизатор*. Чаще всего в качестве $\Omega(u)$ берут функционал

$$\Omega(u) = \Omega_n(u) = \|u\|_{W_2^n}^2 = \int_a^b \left[\sum_{k=0}^n p_k(s) \left(\frac{d^k u(s)}{ds^k} \right)^2 \right] ds$$

при некотором значении n (здесь W_2^n – пространство Соболева, а все весовые функции $p_k(s)$ непрерывны и неотрицательны).

Теорема 7.1. Задача (7.71) имеет решение $u_\alpha(x)$ при любых $f(x) \in F$ и $\alpha > 0$.

[[Доказательство](#)]

Теорема 7.2. Алгоритм (7.71) является регуляризующим для задачи (7.69).

[[Доказательство](#)]

Следствие 7.1. Задача (7.71) корректно поставлена.

[[Доказательство](#)]

Замечание 7.3. Сходимость в пространстве W_2^n означает, что n -я производная от сходится среднеквадратично, а сама функция и все остальные (до порядка $(n - 1)$ включительно) – равномерно. Таким образом, использование стабилизатора $\Omega_n(u)$ обеспечивает слабую регуляризацию при $n = 0$, сильную при $n = 1$ и $(n - 1)$ -го порядка гладкости при $n > 1$.

Выбор параметров регуляризации

В ряде прикладных задач известно, что правые части имеют характерную погрешность $\|\tilde{f} - f\|$ порядка некоторой заданной величины δ . Если при этом выбрать α настолько малым, что нарушится критерий (Д.12), то устойчивость расчетов станет недостаточной, так что регуляризованное решение \tilde{u}_α будет заметно «разболтанным». Если же α настолько велико, что не соблюден критерий (Д.15), то регуляризованное решение \tilde{u}_α будет чрезмерно слажено, что также нежелательно.

Вдобавок непосредственно проверить выполнение критериев (Д.12), (Д.15) не удается, поскольку $\beta(\varepsilon)$ неизвестно. Поэтому оптимальный выбор параметра α является сложной задачей.

Обычно на практике проводят расчеты с несколькими значениями параметра α , составляющими геометрическую прогрессию (например, $10^{-1}, 10^{-2}, \dots$) (или по какому-либо другому закону), из полученных результатов выбирают наилучший либо визуально, либо по какому-либо критерию правдоподобия.



Такое поведение характерно для некорректных задач. Например, приближенное вычисление производной на сетке с шагом h с помощью численного дифференцирования приводит к следующему: характерной погрешностью метода является величина вида Ch^m , а полная (включая погрешность входных данных) – $E = CH^m + \frac{\delta}{h}$. Следовательно, оптимальным значением величины шага является значение $h = h_{\text{опт}} = \sqrt[m+1]{\frac{\delta}{Cm}}$.

Выбор n . Аналогично предыдущим рассуждениям можно отметить, что при чрезмерно больших значениях n регуляризованное решение сильно сглаживается. Значение $n = 0$ обеспечивает лишь среднеквадратичную сходимость \tilde{u}_α к \bar{u} . Поэтому наиболее часто используют значение $n = 1$.

Уравнение Эйлера

Учитывая явный вид операторов A и Ω , перепишем задачу (7.71) следующим образом:

$$\alpha \sum_{k=0}^n \int_a^b p_k(s) \left[u^{(k)}(s) \right]^2 ds + \int_c^d \left[\int_a^b K(x, s) u(s) ds - f(x) \right]^2 dx \rightarrow \min. \quad (7.72)$$

Из теории вариационного исчисления известно, что функция $\tilde{u}(x)$, доставляющая решение задачи (7.72), удовлетворяет

текущему уравнению Эйлера, смысл которого аналогичен необходимому условию минимума первого порядка для функций: первая вариация равна нулю. Полагая $u \sim u_1 := u + \delta u$, обнуляем те слагаемые, которые содержат первые степени δ :

$$0 = \alpha \sum_{k=0}^n \int_a^b p_k(s) u^{(k)}(s) \delta u^{(k)}(s) ds + \int_c^d \left[\int_a^b K(x, \eta) u(\eta) d\eta - f(x) \right] \int_a^b K(x, s) \delta u(s) ds dx. \quad (7.73)$$



Интегралы, стоящие под знаком суммы, будем вычислять последовательным интегрированием по частям:

$$\int_a^b p_k(s) u^{(k)}(s) \delta u^{(k)}(s) ds = \delta u^{(k-1)}(s) p_k(s) u^{(k)}(s) \Big|_a^b - \int_a^b \delta u^{(k-1)}(s) \frac{d}{ds} [p_k(s) u^{(k)}(s)] ds =$$

$$= \sum_{j=0}^{k-1} (-1)^j \delta u^{(k-1-j)}(s) \frac{d^j}{ds^j} [p_k(s) u^{(k)}(s)] \Big|_a^b + (-1)^k \int_a^b \delta u(s) \frac{d^k}{ds^k} [p_k(s) u^{(k)}(s)] ds.$$

Подставляя это выражение в уравнение вариации (7.73), получим:

$$\alpha \sum_{k=0}^n \sum_{j=0}^{k-1} (-1)^j \delta u^{(k-1-j)}(s) \frac{d^j}{ds^j} [p_k(s) u^{(k)}(s)] \Big|_a^b + \alpha \sum_{k=0}^n (-1)^k \int_a^b \delta u(s) \frac{d^k}{ds^k} [p_k(s) u^{(k)}(s)] ds +$$

$$+ \int_c^d \left[\int_a^b K(x, \eta) u(\eta) d\eta - f(x) \right] \int_a^b K(x, s) \delta u(s) ds dx = 0.$$

Теперь поменяем порядок суммирования (что равносильно собиранию коэффициентов при $\delta u^{(k)}$):

$$\alpha \sum_{j=1}^n \delta u^{(j-1)}(s) \sum_{k=j}^n (-1)^{k-j} \frac{d^{k-j}}{ds^{k-j}} [p_k(s) u^{(k)}(s)] \Big|_a^b + \alpha \sum_{k=0}^n (-1)^k \int_a^b \delta u(s) \frac{d^k}{ds^k} [p_k(s) u^{(k)}(s)] ds +$$

$$+ \int_c^d \left[\int_a^b K(x, \eta) u(\eta) d\eta \right] \left[\int_a^b K(x, s) \delta u(s) ds \right] dx = \int_c^d f(x) \int_a^b K(x, s) \delta u(s) ds dx.$$

Введем обозначение $q_j(u) = \sum_{k=j}^n (-1)^{k-j} \frac{d^{k-j}}{ds^{k-j}} [p_k(s) u^{(k)}(s)]$.

Тогда, полагая

$$q_j(u(a)) = q_j(u(b)) = 0, \quad j = \overline{1, n} \tag{7.74}$$



и приравнивая коэффициенты при $\delta u(s)$ под знаками интеграла справа и слева (эта процедура эквивалентна выбору в качестве вариации δ -функции), получим:

$$\alpha \sum_{k=0}^n (-1)^k \frac{d^k}{ds^k} \left[p_k(s) u^{(k)}(s) \right] + \int_a^b \left[\int_c^d K(x, \eta) K(x, s) dx \right] u(\eta) d\eta = \int_c^d K(x, s) f(x) dx.$$

Вводя обозначения

$$Q(s, \eta) = \int_c^d K(x, \eta) K(x, s) dx, \quad \Phi(s) = \int_c^d K(x, s) f(x) dx, \quad (7.75)$$

последнее уравнение перепишем в виде

$$\alpha \sum_{k=0}^n (-1)^k \frac{d^k}{ds^k} \left[p_k(s) u^{(k)}(s) \right] + \int_a^b Q(s, \eta) u(\eta) d\eta = \Phi(s). \quad (7.76)$$

(7.74) – (7.76) и представляет собой искомое уравнение Эйлера и является интегро-дифференциальным уравнением, ядро $Q(x, s)$ которого определено на квадрате $[a, b] \times [a, b]$, симметрично и непрерывно, а правая часть $\Phi(s)$ непрерывна.

В частном случае слабой регуляризации (при $n = 0$) (7.74) – (7.76) превращается в

$$\alpha u(s) + \int_a^b Q(s, \eta) u(\eta) d\eta = \Phi(s), \quad a \leq s \leq b. \quad (7.77)$$

т.е. в этом случае регуляризованная задача представляет собой обычное интегральное уравнение Фредгольма второго рода, способы решения которого мы изучали выше.

Замечание о решении плохо обусловленных линейных систем

Описанным выше способом можно решать и системы линейных алгебраических уравнений. Рассмотрим вкратце эту ситуацию.



Пусть задана линейная система

$$Au = f,$$

где u и f – конечномерные векторы.

Выбирая в [стабилизирующем функционале](#) $\Omega_n(u)$ $n = 0$, по аналогии с изложенным выше, запишем задачу минимизации вариационного функционала

$$M(\alpha, f, u) = \|Au - f\|^2 + \alpha \|u\|^2 \rightarrow \min(\|u\|^2 = (a, a)). \quad (7.78)$$

Формально $n = 0$ соответствует слабой регуляризации, но в конечномерном пространстве все нормы эквивалентны. Поэтому сходимость регуляризованного решения (при $\alpha \rightarrow 0$) будет равномерной.

Поскольку (7.78) является квадратичной формой относительно u , то нахождение минимума последней сводится к решению линейной алгебраической системы

$$(A^T A + \alpha E) u = A^T f.$$

Благодаря слагаемому αE эта система хорошо обусловлена (по крайней мере, при не слишком малых $\alpha > 0$). Поэтому ее можно решать с помощью стандартных алгоритмов.

Описанный алгоритм может быть применен также к решению систем с вырожденной матрицей A .



Глава 8

Методы решения задачи Коши для обыкновенных дифференциальных уравнений

- [8.1. Общая информация](#)
- [8.2. Одношаговые методы решения задачи Коши](#)
- [8.3. Многошаговые методы решения задачи Коши](#)
- [8.4. Понятие устойчивости численных методов решения задачи Коши](#)
- [8.5. Жесткие задачи и методы их решения](#)



8.1. Общая информация

8.1.1. Введение

8.1.2. Классификация методов решения задачи Коши



8.1.1. Введение

С помощью обыкновенных дифференциальных уравнений можно описывать движение системы взаимодействующих материальных точек (динамика системы материальных точек), концентрации реагирующих веществ в химических превращениях (задачи химической кинетики), задачи теории электрических цепей, сопротивления материалов и т.п. Учитывая, что в большинстве своем — это задачи динамики (т.е. изменения состояния некоторой системы с течением времени), независимую переменную будем обозначать буквой t .

Конкретная прикладная задача может приводить к дифференциальному уравнению любого порядка или к системе таких уравнений (опять-таки, любого порядка). При этом каждое из уравнений может иметь различный вид (в том числе и задаваться неявно). Мы в нашем курсе будем, однако, основное внимание уделять некоторому частному, но достаточно важному и широко распространенному классу задач, а именно: мы будем рассматривать уравнения, разрешенные относительно старшей производной.

Так, например, уравнение n -го порядка, разрешенное относительно старшей производной, выглядит следующим образом:

$$u^{(n)}(t) = f(t, u, u', \dots, u^{(n-1)}), \quad (8.1)$$

где f — некоторая заданная функция от $(n+1)$ аргументов.

В то же время, хорошо известно, что задачу (1) с помощью замены $u^{(k)}(t) = u_{k+1}(t)$ можно свести к эквивалентной системе обыкновенных дифференциальных уравнений первого порядка

$$\begin{cases} u'_k(t) = u_{k+1}(t), & k = 1, \dots, n-1, \\ u'_n(t) = f(t, u_1, \dots, u_n). \end{cases}$$

Аналогично произвольную систему обыкновенных дифференциальных уравнений любого порядка можно заменить некоторой эквивалентной системой уравнений первого порядка (естественно, с увеличением ее размерности по сравнению с исходной).

Учитывая сказанное, как правило, рассматривают системы уравнений первого порядка

$$u'_k(t) = f_k(t, u_1, \dots, u_n), \quad k = \overline{1, n},$$



которые в векторной форме имеют вид

$$u'(t) = f(t, u), \quad (8.2)$$

(здесь $u = (u_1, \dots, u_n)^T$, $f = (f_1, \dots, f_n)^T$).

Известно, что система имеет множество решений, которое в общем случае зависит от n параметров $C = (C_1, \dots, C_n)^T$: $u(t) = u(t; C)$.

Для определения значений этих параметров, т.е. для выделения конкретного решения необходимо задать n дополнительных ограничений на функции $u_k(t)$. Эти ограничения, вообще говоря, могут иметь самый разнообразный характер, хотя достаточно часто в качестве таковых используются значения указанных функций (или линейных комбинаций от них) в определенных точках промежутка, на котором поставлена задача.

Различают три основных типа задач для обыкновенных дифференциальных уравнений. Это:

- 1) Задачи Коши (или начальные задачи);
- 2) краевые задачи (или задачи с граничными условиями);
- 3) задачи на собственные значения (задачи Штурма-Лиувилля).

Задача Коши имеет дополнительные условия вида

$$u'_k(t_0) = u_k^0, \quad k = \overline{1, n}, \quad (8.3)$$

т.е. заданы значения всех функций в одной и той же точке $t = t_0$.



8.1.2. Классификация методов решения задачи Коши

Для простоты изложения основных идей вычислительных методов решения задачи Коши в дальнейшем будем рассматривать (если это не оговорено особо) случай одного обыкновенного дифференциального уравнения первого порядка. Как правило, эти идеи, равно как и полученные с их помощью алгоритмы, легко переносятся на случай систем вида (8.2) (а следовательно, и на случай уравнений высших порядков).

Итак, пусть на отрезке $t_0 \leq t \leq T$ требуется найти решение $u(t)$ дифференциального уравнения

$$u'(t) = f(t, u), \quad (8.4)$$

удовлетворяющее начальному условию

$$u(t_0) = u_0 \quad (8.5)$$

Условия существования и единственности решения поставленной задачи Коши будем считать выполнеными. Будем также предполагать, что функция $f(t, u)$ обладает необходимой по ходу изложения гладкостью.

Существующие приближенные алгоритмы можно (с определенной долей условности) разделить на

- 1) аналитические (когда решение получается в виде аналитически заданной функции);
- 2) численные (когда решение находят в виде таблицы значений в узлах заданной, либо параллельно построенной сетки).

Простейшим из аналитических методов является метод последовательных приближений или [метод Пикара](#). Этот метод позволяет получать в аналитическом виде последовательность приближений $u_m(t)$, $m = 0, 1, \dots$ к решению $u(t)$ по следующему правилу:

$$u_m(t) = u_0 + \int_{t_0}^t f(x, u_{m-1}(x)) dx, \quad t_0 \leq t \leq T, \quad m = 1, 2, \dots,$$

$$u_0(t) \equiv u_0.$$



Метод Пикара, однако, редко используется в практике вычислений. Одним из его существенных недостатков является необходимость выполнения операции интегрирования при осуществлении каждой итерации.

Несколько более широкое распространение в практике получил другой аналитический метод, основанный на идее разложения решения рассматриваемой задачи в ряд — [метод рядов](#). Особенно часто для этих целей используют ряд Тейлора. В этом случае вычислительные правила строятся особенно просто. Приближенное решение $u_m(t)$ исходной задачи ищется в виде

$$u_m(t) = \sum_{i=0}^m \frac{(t - t_0)^i}{i!} u^{(i)}(t_0), \quad t_0 \leq t \leq T, \quad (8.6)$$

где

$$u^{(0)}(t_0) = u(t_0) = u_0, \quad u^{(1)}(t_0) = u'(t_0) = f(t_0, u_0),$$

а значения $u^{(i)}(t_0)$, $i = 2, 3, \dots, m$ находят по формулам, полученным последовательным дифференцированием уравнения (8.4):

$$u^{(2)}(t_0) = u''(t_0) = f_t(t_0, u_0) + f_u(t_0, u_0) \cdot f(t_0, u_0),$$

$$u^{(3)}(t_0) = u'''(t_0) = f_{tt}(t_0, u_0) + 2f_{tu}(t_0, u_0) \cdot f(t_0, u_0) + f_{uu}(t_0, u_0) \cdot f^2(t_0, u_0) + \dots \quad (8.7)$$

$$+ f_u(t_0, u_0)(f_t(t_0, u_0) + f_u(t_0, u_0) \cdot f(t_0, u_0)),$$

.....

Для значений t , близких к t_0 , метод рядов (8.6) — (8.7) при достаточно большом m дает хорошее приближение к точному решению $u(t)$ задачи (8.4) — (8.5). Однако с увеличением расстояния $t - t_0$ погрешность приближенного равенства $u(t) \approx u_m(t)$, вообще говоря, возрастает по абсолютной величине, и



правило (8.6) становится вовсе непригодным, когда t выходит за пределы области сходимости соответствующего ряда.

Более предпочтительными в таких случаях будут численные методы решения задачи Коши, позволяющие в узлах сетки $t_0 < t_1 < t_2 < \dots < t_N = T$ последовательно находить значения $y_j \approx u(t_j)$, $j = 1, 2, \dots, N$ приближенного решения.

Большинство численных методов решения рассматриваемой задачи Коши можно записать в виде

$$y_{j+1} = F(y_{j-q}, y_{j-q+1}, \dots, y_j, y_{j+1}, \dots, y_{j+s}), \quad (8.8)$$

где F — некоторая известная функция указанных аргументов, определяемая способом построения метода и зависящая от вида уравнения (8.4) и избранной сетки узлов. При $q = 0$ и $s \in \{0, 1\}$ такие вычислительные правила называют *одношаговыми*, а при $q \geq 1$ или $s > 1$ — *многошаговыми*.

Как одношаговые, так и многошаговые методы вида (8.8) называют явными в случае $s = 0$ и неявными при $s \geq 1$. При $s > 1$ многошаговые правила часто называют методами с забеганием вперед.

Если правило (8.8) является одношаговым, то вычисления по нему можно начинать со значения $j = 0$ и проводить до значения $j = N - 1$ включительно. В случае же многошаговых методов указанного вида, вообще говоря, нарушается однородность вычислительного процесса, и для нахождения первых q значений y_1, \dots, y_q и последних $s - 1$ значений требуется применение специальных (отличных от базового) правил. В этом смысле одношаговые правила оказываются предпочтительнее. Удобнее пользоваться ими и в том случае, когда шаг сетки $\tau_j = t_{j+1} - t_j$ не является постоянным для всех значений j (например, когда величина τ выбирается компьютером автоматически по результатам вычислений).

Лучше работают одношаговые методы и в областях резкого изменения функций. В то же время, в экономичности решения доступных им задач, как правило, выигрывают многошаговые методы.

Прежде чем перейти к рассмотрению конкретных вычислительных правил, остановимся на некоторых понятиях, используемых в теории и практике решения дифференциальных уравнений.

Определение. Невязку численного метода (8.8) на точном решении задачи (8.4)

$$r(t_j, \tau) = u(t_{j+1}) - F(u(t_{j-q}), \dots, u(t_j), u(t_{j+1}), \dots, u(t_{j+s})) \quad (8.9)$$

будем называть *локальной погрешностью* метода (8.8).



Определение. Величину

$$\psi(t_j, \tau) = \frac{u(t_{j+1}) - u(t_j)}{\tau} - \frac{F(u(t_{j-q}), \dots, u(t_j), u(t_{j+1}), \dots, u(t_{j+s})) - u(t_j)}{\tau} \equiv \frac{r(t_j, \tau)}{\tau}$$

будем называть *погрешностью аппроксимации* дифференциальной задачи (8.4) разностной задачей (8.8).

Если при этом $\psi(t_j, \tau) = O(\tau^p)$, $p \geq 1$, то метод (8.8) называют *методом p-го порядка точности*.



8.2. Одношаговые методы решения задачи Коши

- 8.2.1. Пошаговый вариант метода рядов
- 8.2.2. Способ Рунге–Кутта построения одношаговых методов
- 8.2.3. Явные методы Рунге–Кутта
- 8.2.4. Способ последовательного повышения порядка точности построения одношаговых методов
- 8.2.5. Практический контроль погрешности приближенного решения
- 8.2.6. Сходимость одношаговых методов решения задачи Коши



8.2.1. Пошаговый вариант метода рядов

Прежде всего еще раз отметим, что основная отличительная черта численных методов — это тот факт, что они представляют собой алгоритмы вычисления приближенных (а иногда — точных) значений искомого решения $u(t)$ на некоторой выбранной сетке значений аргумента t_j . Они не позволяют найти общее решение системы, а дают какое-либо частное. При этом одношаговые методы для нахождения решения в очередном узле сетки t_{j+1} используют информацию о задаче только из отрезка $[t_j; t_{j+1}]$.

Итак, будем считать, что процесс решения задачи (8.4), (8.5) доведен до некоторой точки t_j ($0 \leq j < N$) (таким образом, нам известны значения y_k , $k = 0, 1, \dots, j$).

Построим сейчас простейший вычислительный алгоритм для нахождения решения в точке $t_{j+1} = t_j + \tau_j$ сетки. Поскольку при построении одношаговых методов используется информация о решаемой задаче лишь в пределах одного шага интегрирования, то можно без ущерба для понимания не писать индекс j , обозначающий номер шага процесса.

Для решения поставленной задачи воспользуемся формулой (8.6), положив в ней вместо t_0 t_j , а вместо $t - t_{j+1}$. В результате получим:

$$y_{j+1} = \sum_{i=0}^m \frac{\tau^i}{i!} y_j^{(i)}, \quad (8.10)$$

где производные $y_j^{(i)}$ вычисляются, как и ранее, по формулам (8.7), в которых вместо t_0 будет стоять t_j , а вместо u — y .

Очевидно, решение в (8.10) разложено по последовательным главным частям (естественно, при достаточно малых τ) и по величине поправки (т.е. очередного слагаемого в сумме) можно судить о том, с какой локальной погрешностью получено интересующее нас значение.

Конкретными примерами методов типа (8.10) могут служить:

1. $m = 1$ — метод первого порядка:

$$y_{j+1} = y_j + \tau f(t_j, y_j) \quad (8.11)$$

(другие названия метода (8.11) — [явный метод Эйлера](#), метод ломаных);



2. $m = 2$ — метод второго порядка:

$$y_{j+1} = y_j + \tau f(t_j, y_j) + \frac{\tau^2}{2} (f_t(t_j, y_j) + f(t_j, y_j) f_u(t_j, y_j)). \quad (8.12)$$

Поменяв точку, в окрестности которой строится разложение, с t_0 на t_j , мы в какой-то степени ослабили первый из недостатков [метода рядов](#), связанный со сходимостью. В то же время, второй недостаток, связанный с необходимостью нахождения большого числа различных функций $\binom{m(m+1)}{2}$ здесь не только не исчезает, а наоборот, усиливается, поскольку делать это теперь нужно на каждом шаге. Поэтому при $m > 1$ указанный метод применяется редко.



8.2.2. Способ Рунге–Кутта построения одношаговых методов

Проинтегрировав уравнение (8.4) по отрезку $[t_j; t_{j+1}]$, получим равенство

$$u(t_j + \tau) = u(t_j) + \int_{t_j}^{t_{j+1}} f(x, u(x)) dx, \quad (8.13)$$

которое связывает значения решения рассматриваемого уравнения в двух соседних узлах сетки. Указав эффективный способ вычисления интеграла в (8.13), мы получим одно из приближенных правил численного интегрирования уравнения (8.4). В силу требования одношаговости конструируемых правил при нахождении значения указанного интеграла мы можем использовать информацию о функции $f(t, u(t))$ лишь из отрезка $[t_j; t_j + \tau]$. По постановке задачи значение этой функции в точке t_j нам известно. Поэтому для вычисления интеграла можно применить, например, [квадратурную формулу левых прямоугольников](#)

$$\int_a^b \varphi(x) dx \approx (b - a) \varphi(a).$$

В результате получим метод (первого порядка точности),

$$y_{j+1} = y_j + \tau f(t_j, y_j),$$

полностью совпадающий с (8.11).

Аналогично, применив для вычисления интеграла в (8.13) формулу правых прямоугольников, получим [неявный метод Эйлера](#)

$$y_{j+1} = y_j + \tau f(t_{j+1}, y_{j+1}), \quad (8.14)$$

который также является методом первого порядка.

Воспользовавшись для замены интеграла в (8.13) формулой трапеций, получим одношаговый метод численного интегрирования — [неявный метод трапеций](#)

$$y_{j+1} = y_j + \frac{\tau}{2} (f(t_j, y_j) + f(t_{j+1}, y_{j+1})), \quad (8.15)$$



являющийся методом второго порядка.

Заметим, что методы (8.14) и (8.15) (как и любой другой неявный метод) требуют для определения искомого решения y_{j+1} на каждом шаге решать нелинейное уравнение, что далеко не всегда просто.

Встает вопрос: как построить явные одношаговые методы, более точные, чем (8.11), и не использующие производных от функции f . В способе Рунге–Кутта предлагается использовать следующий специальный прием приближенного вычисления интеграла в (8.13).

Введем обозначение $\Delta u = u(t_j + \tau) - u(t_j)$ и сделаем в (8.13) замену переменной интегрирования по формуле $x = t_j + \alpha\tau$. Тогда (8.13) можно переписать в виде

$$\Delta u = \tau \int_0^1 f(t_j + \alpha\tau, u(t_j + \alpha\tau)) d\alpha. \quad (8.16)$$

Зададим три набора параметров: $(a_{ij})_{i,j=1}^s$ (матрица), $(b_1, \dots, b_s)^T$ и $(c_1, \dots, c_s)^T$ (векторы). При помощи параметров c_i и a_{ij} построим величины

$$k_1 = f(t_j + c_1\tau, y_j + \tau \sum_{l=1}^s a_{1l} k_l),$$

$$k_2 = f(t_j + c_2\tau, y_j + \tau \sum_{l=1}^s a_{2l} k_l),$$

.....

$$k_s = f(t_j + c_s\tau, y_j + \tau \sum_{l=1}^s a_{sl} k_l).$$

Каждая из величин $k_i = f(t_j + c_i\tau, y_j + \tau \sum_{l=1}^s a_{il} k_l)$, вообще говоря, не равна значению $f(t_j + c_i\tau, u(t_j + c_i\tau))$, однако при соответствующем выборе параметров их можно надеяться сделать близкими. А это, в свою очередь, дает основание надеяться при помощи параметров b_i составить



такую линейную комбинацию величин k_i ($i = \overline{1, s}$), которая будет являться аналогом [квадратурной суммы](#) и позволит вычислить приближенное значение приращения:

$$\Delta y = \tau \sum_{i=1}^s b_i k_i$$

или

$$y_{j+1} = y_j + \tau \sum_{i=1}^s b_i k_i. \quad (8.18)$$

Формулы (8.17), (8.18) определяют [s-стадийный метод Рунге–Кутта](#).

Как видим, метод однозначно определяется своими параметрами, которые традиционно размещают в таблицу вида

$$\begin{array}{c|ccccc} & c_1 & a_{11} & \cdots & a_{1s} \\ \hline c & \vdots & \vdots & \ddots & \vdots \\ A & & & & \\ \hline b^T & c_s & a_{s1} & \cdots & a_{ss} \\ & & b_1 & \cdots & b_s \end{array}. \quad (8.19)$$

Эта таблица носит название [таблицы Бутчера](#).

При этом, если $a_{in} = 0$ при $n \geq i$ для всех i (и $c_1 = 0$), то вектор k_i может быть вычислен явным образом по значениям k_1, \dots, k_{i-1} . Поэтому такие методы называют [явными](#) методами Рунге–Кутта.

Если $a_{in} = 0$ при $n > i$ и хотя бы при одном значении i $a_{ii} \neq 0$, то получающиеся методы носят название [диагонально неявных](#), и, наконец, если среди элементов матрицы A имеются отличные от нуля и выше ее главной диагонали, то такие методы называют просто [неявными](#).



8.2.3. Явные методы Рунге–Кутта

Методы первого порядка точности

Методы второго порядка точности

Условия порядка для явных методов Рунге–Кутта

Методы третьего и четвертого порядка точности

Начнем рассмотрение явных методов Рунге–Кутта.

Итак, построить метод Рунге–Кутта — значит указать конкретную [таблицу Бутчера](#) (или, что то же самое, — конкретные наборы параметров c , A и b).

Естественно выбирать их таким образом, чтобы конструируемый метод (8.18), (8.17) имел по возможности более высокий [порядок точности](#). В предположении, что правая часть уравнения (8.4) (функция f) является достаточно гладкой, можно записать разложение [локальной погрешности](#) формулы (8.18) $r(t_j, \tau)$ в ряд Тейлора с остаточным членом в форме Лагранжа:

$$r(t_j, \tau) = \Delta u - \tau \sum_{i=1}^s b_i k_i = \sum_{l=0}^k \frac{\tau^l}{l!} r^{(l)}(t_j, 0) + \frac{\tau^{k+1}}{(k+1)!} r^{(k+1)}(t_j, \theta\tau), \quad 0 < \theta < 1.$$

Если теперь подобрать параметры c , A и b так, чтобы выполнялись условия

$$r^{(l)}(t_j, 0) = 0, \quad l = 0, 1, \dots, k, \tag{8.20}$$

то локальная погрешность метода примет вид

$$r(t_j, \tau) = \frac{\tau^{k+1}}{(k+1)!} r^{(k+1)}(t_j, \theta\tau), \tag{8.21}$$

т.е. метод будет методом k -го порядка точности.

Практически при построении методов может оказаться более целесообразной следующая схема действий: составляют разложение по степеням τ величины

$$u(t_j + \tau) = u(t_j) + \tau u'(t_j) + \frac{\tau^2}{2} u''(t_j) + \dots = u + \tau f + \frac{\tau^2}{2} (f_t + f_u f) + \dots$$



Аналогичное разложение составляется для правой части формулы (8.18) (на точном решении задачи (8.4)), т.е. для комбинации

$$u(t_j) + \tau \sum_{i=1}^s b_i k_i = u(t_j) + \tau \sum_{i=1}^s b_i f(t_j + c_i \tau, u(t_j) + \tau (a_{i1} k_1 + \dots + a_{ii-1} k_{i-1})).$$

После этого требуют, чтобы полученные разложения совпадали до членов с возможно более высокими степенями τ для произвольной функции f .

При произвольных s и k систему уравнений для определения параметров c , A и b записать достаточно сложно и мы этим займемся немного позже. А пока рассмотрим применение указанного подхода к построению конкретных примеров методов невысоких порядков точности.

Методы первого порядка точности

Зададимся минимальным значением $s = 1$, что равнозначно введению лишь одного параметра b_1 . Равенство (8.18) в этом случае будет иметь вид

$$y_{j+1} = y_j + \tau b_1 k_1 = y_j + \tau f(t_j, y_j),$$

а погрешность $r(t_j, \tau)$ примет вид

$$r(t_j, \tau) = u(t_j + \tau) - u(t_j) - \tau b_1 f(t_j, u(t_j)).$$

Учитывая простоту конструкции, непосредственно вычислим производные от локальной погрешности:

$$r'(t_j, \tau) = u'(t_j + \tau) - b_1 f(t_j, u(t_j)),$$

$$r''(t_j, \tau) = u''(t_j + \tau).$$

Так как $r''(t_j, \tau)$ не зависит от b_1 , то уже при $l = 2$ условие (8.20) в случае произвольной функции f удовлетворено быть не может. Поэтому $k = 1$ и система (8.20) принимает вид

$$r'(t_j, 0) = u'(t_j) - b_1 f(t_j, u(t_j)) = (1 - b_1) f(t_j, u(t_j)) = 0.$$



Отсюда следует, что $b_1 = 1$. В этом случае получим хорошо уже знакомый нам [явный метод Эйлера](#), таблица Бутчера которого имеет вид

0	0
	1

Его локальная погрешность, согласно (8.21), будет иметь вид

$$r(t_j, \tau) = \frac{\tau^2}{2} r''(t_j, \theta\tau) = \frac{\tau^2}{2} u''(t_j + \theta\tau).$$

Методы второго порядка точности

Положим $s = 2$ (при $s = 1$, как мы видели, явных методов порядка выше первого получить нельзя). Тогда

$$y_{j+1} = y_j + \tau(b_1 k_1 + b_2 k_2).$$

Выполняя указанные выше разложения правой и левой частей последней формулы на точном решении в ряды по степеням τ , имеем:

$$\begin{aligned} u(t_j + \tau) &= u(t_j) + \tau u'(t_j) + \frac{\tau^2}{2} u''(t_j) + \frac{\tau^3}{6} u'''(t_j) + O(\tau^4) = \\ &= u + \tau f + \frac{\tau^2}{2} (f_t + f_u f) + \frac{\tau^3}{6} (f_{tt} + 2f_{tu}f + f_{uu}f^2 + f_u(f_t + f_u f)) + O(\tau^4). \end{aligned} \tag{*}$$

$$\begin{aligned} u(t_j) + \tau(b_1 k_1 + b_2 k_2) &= u(t_j) + \tau b_1 f(t_j, u(t_j)) + \tau b_2 f(t_j + c_2 \tau, u(t_j) + \tau a_{21} f(t_j, u(t_j))) = \\ &= u + \tau b_1 f + \tau b_2 \left[f + c_2 \tau f_t + \tau a_{21} f_u f + \frac{c_2^2 \tau^2}{2} f_{tt} + c_2 a_{21} \tau^2 f_{tu} f + \frac{a_{21}^2 \tau^2}{2} f_{uu} f^2 \right] + O(\tau^4) = \\ &= \tau(b_1 + b_2) f + \tau^2 b_2 (c_2 f_t + a_{21} f_u f) + \frac{\tau^3 b_2}{2} (c_2^2 f_{tt} + 2c_2 a_{21} f_{tu} f + a_{21}^2 f_{uu} f^2) + O(\tau^4). \end{aligned} \tag{**}$$

Приравняем в правых частях разложений (*) и (**) коэффициенты при одинаковых степенях τ и сомножителях, зависящих от f , одинакового вида.

Тем самым на выбор четырех параметров c_2 , a_{21} , b_1 и b_2 будут наложены три условия:

$$\begin{cases} b_1 + b_2 = 1, \\ 2b_2 c_2 = 1, \\ 2b_2 a_{21} = 1. \end{cases} \quad (8.22)$$

Непосредственно из разложений следует, что в случае $s = 2$ для произвольных f нельзя добиться совпадения всех членов с множителем τ^3 за счет выбора введенных параметров. Поэтому при $s = 2$ максимальный порядок точности правил типа Рунге–Кутта равен 2. С другой стороны, система (8.22), очевидно, имеет бесчисленное множество решений, например, такое: $c_2 = a_{21} = \alpha$; $b_2 = \frac{1}{2\alpha}$; $b_1 = 1 - \frac{1}{2\alpha}$, где в качестве α может быть взято, вообще говоря, любое отличное от нуля число (хотя требование одношаговости, более естественными будут $\alpha \in (0; 1]$). Таким образом, существует однопараметрическое семейство методов второго порядка точности, задаваемое [таблицей Бутчера](#)

0	0	0
α	α	0
		1 – $\frac{1}{2\alpha}$
		$\frac{1}{2\alpha}$

или

$$\begin{cases} y_{j+1} = y_j + \frac{\tau}{2\alpha} ((2\alpha - 1) k_1 + k_2), \\ k_1 = f(t_j, y_j), \\ k_2 = f(t_j + \alpha\tau, y_j + \tau\alpha k_1) \end{cases} \quad (8.23)$$

Наиболее употребительными из формул (8.23) являются их частные случаи при $\alpha = \frac{1}{2}$, т.е.

0	0	0
$\frac{1}{2}$	$\frac{1}{2}$	0
		0
		1

или

$$\begin{cases} y_{j+1} = y_j + \tau k_2, \\ k_1 = f(t_j, y_j), \\ k_2 = f(t_j + \frac{\tau}{2}, y_j + \frac{\tau}{2} k_1) \end{cases} \quad (8.24)$$

(аналог [квадратурной формулы средних прямоугольников](#)), и $\alpha = 1$:

0	0	0
1	1	0
		$\frac{1}{2}$
		$\frac{1}{2}$

или

$$\begin{cases} y_{j+1} = y_j + \frac{\tau}{2} (k_1 + k_2), \\ k_1 = f(t_j, y_j), \\ k_2 = f(t_j + \tau, y_j + \tau k_1) \end{cases} \quad (8.25)$$

(аналог [квадратурной формулы трапеций](#)).



Локальная погрешность любого из методов типа (8.23), как следует из разложений (*) и (**), может быть представлена в виде

$$r(t_j, \tau) = \frac{\tau^3}{6} [f_{tt}(1 - 3c_2^2 b_2) + 2ff_{tu}(1 - 3c_2 a_{21} b_2) + f^2 f_{uu}(1 - 3a_{21}^2 b_2) + f_u(f_t + f_u f)] + O(\tau^4). \quad (8.26)$$

Исходя из этого, свободный параметр α иногда выбирают таким образом, чтобы в этом представлении обратилась в нуль хотя бы часть слагаемых. Так как $c_2 = a_{21} = \alpha$, $b_2 = \frac{1}{2\alpha}$, то в этом случае получим:

$$1 - 3\alpha^2 \cdot \frac{1}{2\alpha} = 0,$$

откуда $\alpha = \frac{2}{3}$. При таком выборе α (8.26) существенно упростится:

$$r(t_j, \tau) = \frac{\tau^3}{6} f_u(f_t + f_u f) + O(\tau^4),$$

а соответствующий метод Рунге–Кутта будет иметь вид

	0	0
$\frac{2}{3}$	$\frac{2}{3}$	0
	$\frac{1}{4}$	$\frac{3}{3}$

или

$$\begin{cases} y_{j+1} = y_j + \frac{\tau}{4}(k_1 + 3k_2), \\ k_1 = f(t_j, y_j), \\ k_2 = f(t_j + \frac{2\tau}{3}, y_j + \frac{2\tau}{3}k_1). \end{cases} \quad (8.27)$$

Условия порядка для явных методов Рунге–Кутта

Изучим сейчас общую структуру условий, определяющих порядок метода. Для упрощения вывода преобразуем уравнение (8.4) к автономной форме путем добавления t к зависимым переменным:

$$\begin{pmatrix} t \\ u \end{pmatrix}' = \begin{pmatrix} 1 \\ f(t, u) \end{pmatrix}. \quad (8.28)$$

Таким образом, вместо одного уравнения (в скалярном случае) мы имеем уже систему. Поэтому в дальнейшем (в этом пункте) будем иметь дело с системами, при этом компоненты векторов мы будем обозначать верхними заглавными буквами. Тогда автономную систему общего вида можно записать так:

$$(u^J)' = f^J(u^1, \dots, u^n), \quad J = 1, \dots, n \quad (8.29)$$



Зафиксируем также в схеме (8.18) узел сетки (т.е. будем полагать $t_j = t_0$) и сделаем запись формул (8.18) более симметричной, перейдя от функций $k_i = f(g_i)$ к их аргументам:

$$\begin{cases} g_i^J = y_0^J + \sum_{j=1}^{i-1} a_{ij} \tau f^J(g_j^1, \dots, g_j^n), & i = 1, \dots, s, \\ y_1^J = y_0^J + \sum_{j=1}^s b_j \tau f^J(g_j^1, \dots, g_j^n). \end{cases} \quad (8.30)$$

В частности, если система (8.29) получается из (8.28), то (8.30) при $J = 1$ дает

$$g_j^1 = y_0^1 + \sum_{j=1}^{i-1} a_{ij} \tau. \quad (*)$$

Обычно в явных методах Рунге–Кутта параметры c_i таблицы Бутчера удовлетворяют условию (и мы это видели в частных случаях)

$$c_i = \sum_j a_{ij}. \quad (8.31)$$

С учетом последнего условия равенство (*) примет вид

$$g_j^1 = y_0^1 + \sum_{j=1}^{i-1} a_{ij} \tau = t_0 + c_i \tau,$$

т.е. будет соответствовать «штатному» виду формул (8.18) для неавтономной системы.

Таким образом, если выполнено условие (8.31), то для вывода условий порядка достаточно рассмотреть автономную систему (8.29).

Как мы уже отмечали ранее, для получения условий порядка нужно сравнивать ряды Тейлора для u_1^J и выражения, стоящего в правой части формул (8.30) (естественно, при подстановке туда точного решения и вместо y). Для этой цели вычислим сначала значения производных u_1^J и g_i^J по τ при $\tau = 0$. Ввиду внешнего сходства обеих формул (8.30) достаточно проделать это для g_i^J . В правые части этих формул входят выражения вида $\tau \varphi(\tau)$ и мы воспользуемся формулой Лейбница

$$(\tau \varphi(\tau))^{(q)} \Big|_{\tau=0} = q \cdot (\varphi(\tau))^{(q-1)} \Big|_{\tau=0}. \quad (8.32)$$



Имеем

$$q = 0:$$

$$(g_i^J)^{(0)} \Big|_{\tau=0} = u_0^J, \quad (\Pi.0)$$

$$q = 1:$$

$$(g_i^J)^{(1)} \Big|_{\tau=0} = \sum_j a_{ij} f^J, \quad (\Pi.1)$$

$$q = 2: \text{ Так как}$$

$$(f^J(g_j))^{(1)} = \sum_K f_K^J(g_j) \cdot (g_j^K)^{(1)}, \quad (\Phi.1)$$

где $f_K^J = \frac{\partial f^J}{\partial u^K}$ (формула производной сложной функции), то

$$\begin{aligned} (g_i^J)^{(2)} \Big|_{\tau=0} &= [\varphi(\tau) = \sum_j a_{ij} f^J(g_j)] = 2 \sum_j a_{ij} (f^J(g_j))^{(1)} \Big|_{\tau=0} = 2 \sum_j a_{ij} \sum_K f_K^J(g_j) (g_j^K)^{(1)} \Big|_{\tau=0} = \\ &= 2 \sum_j a_{ij} \sum_K f_K^J \sum_L a_{jk} f^K = 2 \sum_{j,k} a_{ij} a_{jk} \sum_K f_K^J f^K. \end{aligned} \quad (\Pi.2)$$

Аналогично

$$(g_i^J)^{(3)} \Big|_{\tau=0} = 3 \sum_j a_{ij} (f^J(g_j))^{(2)} \Big|_{\tau=0}. \quad (8.33)$$

Продифференцировав [\(Φ.1\)](#), получим:

$$(f^J(g_j))^{(2)} = \left(\sum_K f_K^J(g_j) (g_j^K)^{(1)} \right)^{(1)} = \sum_{K,L} f_{KL}^J(g_j) (g_j^L)^{(1)} (g_j^K)^{(1)} + \sum_K f_K^J(g_j) (g_j^K)^{(2)}. \quad (\Phi.2)$$

Подставляя это выражение в [\(8.33\)](#), будем иметь:

$$\begin{aligned} (g_i^J)^{(3)} \Big|_{\tau=0} &= 3 \sum_j a_{ij} \left[\sum_{K,L} f_{KL}^J(g_j) (g_j^L)^{(1)} (g_j^K)^{(1)} + \sum_K f_K^J(g_j) (g_j^K)^{(2)} \right] \Big|_{\tau=0} = \\ &= 3 \sum_{j,k,l} a_{ij} a_{jk} a_{jl} \sum_{K,L} f_{KL}^J f^K f^L + 6 \sum_{j,k,l} a_{ij} a_{jk} a_{kl} \sum_{K,L} f_K^J f_L^K f^L. \end{aligned} \quad (\Pi.3)$$



Для правой части последней из формул (8.30) ввиду симметрии будут справедливы те же формулы (П.1) — (П.3), если в них заменить a_{ij} на b_j .

Найдем теперь производные точного решения:

$$(u^J)^{(1)} = f^J(u), \quad (\text{T.1})$$

$$(u^J)^{(2)} = \sum_K f_K^J(u) \cdot (u^K)^{(1)} = \sum_K f_K^J f^K, \quad (\text{T.2})$$

$$\begin{aligned} (u^J)^{(3)} &= (\sum_K f_K^J f^K)^{(1)} = \sum_K \sum_L (f_{KL}^J f^L f^K + f_K^J f_L^K f^L) = \\ &= \sum_{K,L} f_{KL}^J f^K f^L + \sum_{K,L} f_K^J f_L^K f^L. \end{aligned} \quad (\text{T.3})$$

Учитывая полученные формулы, легко выписать условия, при которых метод Рунге–Кутта имеет третий порядок:

$$\left\{ \begin{array}{l} \sum_j b_j = 1, \\ 2 \sum_{j,k} b_j a_{jk} = 1, \\ 3 \sum_{j,k,l} b_j a_{jk} a_{jl} = 1, \\ 6 \sum_{j,k,l} b_j a_{jk} a_{kl} = 1. \end{array} \right. \quad (8.34)$$

Если к ним теперь добавить условия (8.31), то мы получим условия третьего порядка методов Рунге–Кутта в том виде, в котором они встречаются в литературе:

$$\left\{ \begin{array}{l} \sum_j b_j = 1, \\ \sum_j b_j c_j = \frac{1}{2}, \\ \sum_j b_j c_j^2 = \frac{1}{3}, \\ \sum_{j,k} b_j a_{jk} c_k = \frac{1}{6}, \\ \sum_j a_{ij} = c_i. \end{array} \right. \quad (8.35)$$



Методы третьего и четвертого порядка точности

Методы третьего порядка. Построим трехстадийный метод третьего порядка. Положив в условиях (8.35) $s = 3$, получаем:

$$\left\{ \begin{array}{l} b_1 + b_2 + b_3 = 1, \\ b_2 c_2 + b_3 c_3 = \frac{1}{2}, \\ b_2 c_2^2 + b_3 c_3^3 = \frac{1}{3}, \\ b_3 a_{32} c_2 = \frac{1}{6}, \\ c_2 = a_{21}, \\ c_3 = a_{31} + a_{32}. \end{array} \right.$$

Одним из алгоритмов третьего порядка точности, получающихся в результате решения этой системы, является, например, такой:

$$\begin{array}{c|cc} 0 & & \\ \hline \frac{1}{3} & \frac{1}{3} & \\ \frac{2}{3} & 0 & \frac{2}{3} \\ \hline & \frac{1}{4} & 0 & \frac{3}{4} \end{array} . \quad (8.36)$$



Методы четвертого порядка. Условия четвертого порядка можно получить аналогично [условиям третьего порядка](#). Более эффективным, однако, является использование для этих целей аппарата помеченных деревьев (см. [17]). Мы же сразу приведем условия четвертого порядка без вывода:

$$\left\{ \begin{array}{l} \sum_j b_j = 1, \\ \sum_{j,k} b_j a_{jk} = \frac{1}{2}, \\ \sum_{j,k,l} b_j a_{jk} a_{jl} = \frac{1}{3}, \\ \sum_{j,k,l} b_j a_{jk} a_{kl} = \frac{1}{6}, \\ \sum_{j,k,l,m} b_j a_{jk} a_{jl} a_{lm} = \frac{1}{4}, \\ \sum_{j,k,l,m} b_j a_{jk} a_{jm} a_{kl} = \frac{1}{8}, \\ \sum_{j,k,l,m} b_j a_{jk} a_{kl} a_{km} = \frac{1}{12}, \\ \sum_{j,k,l,m} b_j a_{jk} a_{kl} a_{lm} = \frac{1}{24} \end{array} \right. \quad \text{или, с учетом (8.31),} \quad \left\{ \begin{array}{l} \sum_j b_j = 1, \\ \sum_j b_j c_j = \frac{1}{2}, \\ \sum_j b_j c_j^2 = \frac{1}{3}, \\ \sum_{j,k} b_j a_{jk} c_k = \frac{1}{6}, \\ \sum_j b_j c_j^3 = \frac{1}{4}, \\ \sum_{j,k} b_j c_j a_{jk} c_k = \frac{1}{8}, \\ \sum_{j,k} b_j a_{jk} c_k^2 = \frac{1}{12}, \\ \sum_{j,k,l} b_j a_{jk} a_{kl} c_l = \frac{1}{24}. \end{array} \right. \quad (8.37)$$



Отсюда при $s = 4$ имеем систему

$$\left\{ \begin{array}{l} b_1 + b_2 + b_3 + b_4 = 1, \\ b_2 c_2 + b_3 c_3 + b_4 c_4 = \frac{1}{2}, \\ b_2 c_2^2 + b_3 c_3^2 + b_4 c_4^2 = \frac{1}{3}, \\ b_2 c_2^3 + b_3 c_3^3 + b_4 c_4^3 = \frac{1}{4}, \\ b_3 a_{32} c_2 + b_4 (a_{42} c_2 + a_{43} c_3) = \frac{1}{6}, \\ b_3 a_{32} c_2 c_3 + b_4 c_4 (a_{42} c_2 + a_{43} c_3) = \frac{1}{8}, \\ b_3 a_{32} c_2^2 + b_4 (a_{42} c_2^2 + a_{43} c_3^2) = \frac{1}{12}, \\ b_4 a_{43} a_{32} c_2 = \frac{1}{24}, \\ c_2 = a_{21}, \\ c_3 = a_{31} + a_{32}, \\ c_4 = a_{41} + a_{42} + a_{43}, \end{array} \right.$$

одним из решений которой является метод, записанный ниже:

$$\begin{array}{c|ccccc} & 0 & & & & \\ \hline \frac{1}{2} & & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & & \frac{1}{2} & & \\ \hline 1 & 0 & 0 & 1 & & \\ \hline & \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6} & \end{array} \quad (8.38)$$

Именно этот метод в технической литературе и называют методом Рунге–Кутта.



Меню



Вверх

Назад

Вперёд

Пред.

След.

Указатель

Помощь

Экран

8.2.4. Способ последовательного повышения порядка точности построения одношаговых методов

[Методы первого порядка точности](#)

[Методы второго порядка точности](#)

[Методы третьего порядка точности](#)

Как и в [способе Рунге–Кутта](#), при построении методов численного решения задачи Коши (8.4), (8.5) будем исходить из интегрального соотношения (8.13), которое сейчас перепишем в виде

$$u(t_j + \tau) = u(t_j) + \tau \int_0^1 z_j(\alpha) d\alpha, \quad (8.39)$$

где $z_j(\alpha) = f(t_j + \alpha\tau, u(t_j + \alpha\tau))$.

Заменим интеграл в (8.39) [квадратурной суммой](#) $\sum_{i=0}^q A_i z_j(\alpha_i)$. Тогда будем иметь:

$$u(t_j + \tau) \approx u(t_j) + \tau \sum_{i=0}^q A_i z_j(\alpha_i) = u(t_j) + \tau \sum_{i=0}^q A_i f(t_j + \alpha_i \tau, u(t_j + \alpha_i \tau)). \quad (8.40)$$

Выбор параметров $A_i, \alpha_i, i = 0, 1, \dots, q$, в этом приближенном равенстве будем осуществлять, например, исходя из требования, чтобы квадратурная формула

$$\int_0^1 z_j(\alpha) d\alpha \approx \sum_{i=0}^q A_i z_j(\alpha_i) \quad (8.41)$$

была точной для всевозможных алгебраических многочленов до степени $k-1$ ($0 < k \leq 2q+2$) включительно. Это приводит к следующей системе из k уравнений с $2q+2$ неизвестными A_i, α_i ($i = 0, 1, \dots, q$):

$$\begin{cases} \sum_{i=0}^q A_i = 1, \\ \sum_{i=0}^q A_i \alpha_i^j = \frac{1}{j+1}, \quad j = 1, \dots, k-1. \end{cases} \quad (8.42)$$



Заметим, что последняя система может быть получена и исходя из требования, чтобы разложения по степеням τ обеих частей приближенного равенства

$$u(t_j + \tau) \approx u(t_j) + \sum_{i=0}^q A_i f(t_j + \alpha_i \tau, u(t_j + \alpha_i \tau)) = u(t_j) + \sum_{i=0}^q A_i u'(t_j + \alpha_i \tau)$$

совпадали до членов с τ^k включительно. Тогда, очевидно, локальная погрешность формулы (8.40) будет иметь вид

$$r(t_j, \tau) = \tau^{k+1} u^{(k+1)}(t_j) \left[\frac{1}{(k+1)!} - \frac{1}{k!} \sum_{i=0}^q A_i \alpha_i^k \right] + O(\tau^{k+2}). \quad (8.43)$$

Так как весовая функция в случае интеграла $\int_0^1 z_j(\alpha) d\alpha$ равна 1, то квадратурная формула вида (8.41), имеющая алгебраическую степень точности, равную $2q + 1$ (формула наивысшей алгебраической степени точности), может быть построена и притом единственным образом для любого значения $q \geq 0$. Поэтому при $k = 2q + 2$ система (8.42) имеет единственное решение, при этом $0 < A_i \leq 1$, $0 < \alpha_i < 1$, $i = \overline{0, q}$. Следовательно, при $1 \leq k \leq 2q + 2$ у этой системы существует хотя бы одно решение. Таким образом, приближенное равенство (8.40) может быть построено (т.е. вопрос о разрешимости системы здесь целиком решается на базе теории квадратурных формул).

Если бы в (8.40) все значения $u(t_j + \alpha_i \tau)$, $i = \overline{0, q}$, были известны точно, то это приближенное равенство позволяло бы найти искомое значение $u(t_j + \tau)$ соответствующего решения задачи Коши по известному значению $u(t_j)$ этого решения с локальной ошибкой порядка τ^{k+1} .

Хотя точными значениями $u(t_j + \alpha_i \tau)$ мы не располагаем, но, подобно (8.40), заменив там τ на $\alpha_i \tau$, нетрудно указать правила для их приближенного вычисления через значения $u(t_j + \alpha_i \beta_{in} \tau)$, для нахождения которых, в свою очередь, можно построить подобные же рекурсивные формулы. При этом следует иметь в виду, что наличие множителя τ перед суммой в формуле (8.40) позволяет находить значения $u(t_j + \alpha_i \tau)$ с локальной ошибкой порядка τ^k , значения $u(t_j + \alpha_i \beta_{in} \tau)$ — с ошибкой порядка τ^{k-1} , и т.д., понижая на каждом шаге рекурсии требования к порядку точности на единицу. Параметры соответствующих приближенных равенств должны удовлетворять системе уравнений типа (8.42), в которой с понижением требований к точности на порядок следует уменьшать на единицу и количество уравнений (отбрасывая при этом последнее из них). При этом часто бывает целесообразным уменьшать и число q , определяющее количество подлежащих выбору параметров.



Следуя такой схеме действий, придем, наконец, к приближенным равенствам

$$u(t_j + \alpha_i \beta_{in} \dots \gamma_{in...l} \tau) \approx u(t_j) + \alpha_i \beta_{in} \dots \gamma_{in...l} \tau f(t_j, u(t_j)), \quad (8.44)$$

на которых процесс замыкается. Погрешность таких равенств будет, очевидно, величиной порядка τ^2 . Они получаются из равенств типа (8.40) в случае, когда квадратурная формула (8.41) является простейшей формулой левых прямоугольников.

Приведем сейчас примеры методов, полученных описанным выше способом, условившись предварительно о следующих обозначениях:

$$[k] y_{j+\alpha} = u(t_j + \alpha \tau) + O(\tau^k), \quad [k] f_{j+\alpha} = f\left(t_j + \alpha \tau, [k] y_{j+\alpha}\right).$$

Методы первого порядка точности

Система (8.42) в этом случае вырождается в единственное уравнение

$$\sum_{i=0}^q A_i = 1. \quad (8.45)$$

Параметры α_i , $i = \overline{0, q}$, могут принимать, вообще говоря, любые фиксированные значения. Однако для случая одношаговых методов выбор этих параметров должен быть подчинен ограничению $0 \leq \alpha_i \leq 1$. Положив в (8.45), например, $q = 0$, найдем: $A_0 = 1$. Задавая теперь $\alpha_0 = 0$, получим формулу [явного метода Эйлера](#):

$$[2] y_{j+1} = [2] y_j + \tau [2] f_j. \quad (8.46)$$

Методы второго порядка точности

В этом случае требование (8.45) нужно дополнить условием

$$\sum_{i=0}^q A_i \alpha_i = \frac{1}{2}. \quad (8.47)$$



При $q = 0$ система (8.45), (8.47) имеет единственное решение $A_0 = 1$, $\alpha_0 = \frac{1}{2}$, что приводит к следующему вычислительному правилу:

$$\begin{cases} {}^{[2]}y_{j+\frac{1}{2}} = {}^{[3]}y_j + \frac{\tau}{2} {}^{[3]}y_j, \\ {}^{[3]}y_{j+1} = {}^{[3]}y_j + \tau f_{j+\frac{1}{2}}. \end{cases} \quad (8.48)$$

При $q = 1$ система (8.45), (8.47) примет вид

$$\begin{cases} A_0 + A_1 = 1, \\ A_0\alpha_0 + A_1\alpha_1 = \frac{1}{2}. \end{cases}$$

Выбрав, например, $\alpha_0 = 0$, $\alpha_1 = 1$, найдем: $A_0 = A_1 = \frac{1}{2}$ и получим **неявный метод трапеций**

$${}^{[3]}y_{j+1} = {}^{[3]}y_j + \frac{\tau}{2} \left({}^{[3]}f_j + {}^{[3]}f_{j+1} \right). \quad (8.49)$$

Используя формулу типа (8.44) (или, что то же самое, явный метод Эйлера), (8.49) можно преобразовать в **явный метод трапеций**:

$$\begin{cases} {}^{[2]}y_{j+1} = {}^{[3]}y_j + \tau {}^{[3]}f_j, \\ {}^{[3]}y_{j+1} = {}^{[3]}y_j + \frac{\tau}{2} \left({}^{[3]}f_j + {}^{[2]}f_{j+1} \right). \end{cases} \quad (8.50)$$

Заметим, что наряду с (8.50) можно записать и более экономичный вариант явного метода трапеций, требующий, в отличие от (8.50), вычисления всего одного значения правой части исходной задачи на каждый узел сетки, кроме первого:

$$\begin{cases} {}^{[2]}y_{j+1} = {}^{[3]}y_j + \tau {}^{[2]}f_j, \\ {}^{[3]}y_{j+1} = {}^{[3]}y_j + \frac{\tau}{2} \left({}^{[2]}f_j + {}^{[2]}f_{j+1} \right). \end{cases} \quad (8.51)$$



Методы третьего порядка точности

К уравнениям (8.45), (8.47) добавляется еще одно:

$$\sum_{i=0}^q A_i \alpha_i^2 = \frac{1}{3}. \quad (8.52)$$

Положив $q = 1$ (при $q = 0$ система, очевидно, несовместна), получим:

$$\begin{cases} A_0 + A_1 = 1, \\ A_0 \alpha_0 + A_1 \alpha_1 = \frac{1}{2}, \\ A_0 \alpha_0^2 + A_1 \alpha_1^2 = \frac{1}{3}. \end{cases}$$

Отсюда, положив $\alpha_0 = 0$, находим: $\alpha_1 = \frac{2}{3}$, $A_1 = \frac{3}{4}$, $A_0 = \frac{1}{4}$. Используя для вычисления $y_{j+\frac{2}{3}}^{[3]}$ формулы (8.48) (с заменой τ на $\frac{2}{3}\tau$), окончательно будем иметь:

$$\begin{cases} [2] y_{j+\frac{1}{3}} = [4] y_j + \frac{\tau}{3} [4] f_j, \\ [3] y_{j+\frac{2}{3}} = [4] y_j + \frac{2\tau}{3} [2] f_{j+\frac{1}{3}}, \\ [4] y_{j+1} = [4] y_j + \frac{\tau}{4} \left([4] f_j + 3 [3] f_{j+\frac{2}{3}} \right). \end{cases} \quad (8.53)$$



При $q = 2$ можно построить, например, такой вычислительный алгоритм третьего порядка точности, базирующийся на [квадратурной формуле Симпсона](#):

$$\left\{ \begin{array}{l} [2] \\ y_{j+\frac{1}{4}} = y_j + \frac{\tau}{4} f_j, \\ [3] \\ y_{j+\frac{1}{2}} = y_j + \frac{\tau}{2} f_{j+\frac{1}{4}}, \\ [3] \\ y_{j+1} = y_j + \tau f_{j+\frac{1}{2}}, \\ [4] \\ y_{j+1} = y_j + \frac{\tau}{6} \left(f_j + 4 f_{j+\frac{1}{2}} + f_{j+1} \right). \end{array} \right. \quad (8.54)$$

Здесь i может принимать значения 3 или 4. При $i = 4$ построенное правило на один узел сетки требует четырехкратного обращения к блоку нахождения значений правой части исходного уравнения. В случае же $i = 3$ точность результата, вообще говоря, несколько понижается (при сохранении порядка), однако в основном счете число обращений к блоку вычисления значений функции $f(t, u)$ сокращается до трех на узел сетки.

Отметим также, что построенный численный метод, как и правила (8.50) и (8.51), имеют предсказывающе-исправляющий характер. Приближенное значение величины $u(t_{j+1})$, найденное с локальной погрешностью порядка τ^3 , уточняется затем по формуле, локальная погрешность которой имеет четвертый порядок. Сравнение значений y_{j+1} и $y_{j+1}^{[4]}$ дает практическую возможность по ходу вычислений без дополнительных вычислительных затрат составить представление о локальной точности полученного приближения к $u(t_{j+1})$. Такое сравнение, в частности, может быть положено в основу правила автоматического выбора шага интегрирования.

На примере приведенных вычислительных правил легко видеть, что описанный способ построения методов численного интегрирования дифференциальных уравнений удовлетворяет принципу модульности, когда сложные вычислительные правила компонуются на основе более простых типовых расчетных формул.



Меню

8.2.5. Практический контроль погрешности приближенного решения

[Правило Рунге](#)

[Использование вложенных методов](#)

В ходе расчетов всегда желательно иметь представление о том, сколь далеко полученное приближенное решение от истинного решения исходной задачи и в соответствии с величиной оценки погрешности выбирать шаг численного интегрирования. Большинство из известных методик, применяемых для этих целей, оценивают главный член погрешности метода.

[Правило Рунге](#)

Как мы видели, главный член [погрешности аппроксимации](#) метода k -го порядка имеет вид $\tau^k \rho(t)$, т.е.

$$u(t_j + \tau) = y(t_j + \tau) + \tau^k \rho(t_j).$$

Тогда, проведя расчеты из одной и той же точки t_j с двумя различными шагами τ_1 и τ_2 , получим:

$$\begin{cases} u(t_j + \tau) \approx y_{\tau_1}(t_j + \tau) + \tau_1^k \rho(t_j), \\ u(t_j + \tau) \approx y_{\tau_2}(t_j + \tau) + \tau_2^k \rho(t_j), \end{cases}$$

откуда

$$\rho(t_j) \approx \frac{y_{\tau_1} - y_{\tau_2}}{\tau_2^k - \tau_1^k}. \quad (8.55)$$

Эта формула дает возможность после проведения вычислений до точки $t_j + \tau$ с шагами τ_1 и τ_2 получить приближенные значения величин погрешностей для каждого из приближенных значений решения y_{τ_1} и y_{τ_2} . Кроме того, при заданной границе ε допустимой погрешности на основании приближенного равенства

$$\varepsilon \approx |\rho(t_j)| \tau_\varepsilon^k$$



по результатам этих вычислений можно выбрать практически более приемлемое при данных требованиях к точности значение шага:

$$\tau_\varepsilon = \sqrt[k]{\varepsilon \left| \frac{\tau_2^k - \tau_1^k}{y_{\tau_1} - y_{\tau_2}} \right|}. \quad (8.56)$$

В практике вычислений достаточно часто в качестве τ_1 и τ_2 выбирают τ и $\frac{\tau}{2}$ соответственно (прием, предложенный еще Рунге). В этом случае формулы (8.55) и (8.56) принимают вид

$$\rho(t_j) \approx \frac{y_{\frac{\tau}{2}} - y_\tau}{\tau^k \left(1 - \frac{1}{2^k}\right)},$$

$$\tau_\varepsilon = \frac{\tau}{2} \sqrt[k]{\frac{(2^k - 1)\varepsilon}{\left| y_{\frac{\tau}{2}} - y_\tau \right|}},$$

а сам способ оценки носит название *двойного пересчета*.

Замечание 8.1. Значение решения в очередной точке сетки только тогда следует считать найденным (вместе с координатой узла сетки), когда значение $\rho(t_j)$, найденное по формуле (8.55) (или ее аналогу) будет в пределах допустимой погрешности. И только после этого следует приступать к очередному шагу.

Использование вложенных методов

Как мы [уже отмечали](#), способ последовательного повышения порядка точности часто приводит к тому, что приближенное решение получается в одной и той же точке, но с разным порядком погрешности относительно шага сетки. Такими, например, являются явный метод трапеций (8.50) и метод третьего порядка точности (8.54).

Тогда разность между этими величинами дает возможность вычислить главный член локальной погрешности метода более низкого порядка (или, если разделить ее на τ , то слово «локальной» можно опустить). Далее действия должны быть такими же, как и в описанном [выше](#) подходе Рунге.

Вложенные методы Рунге–Кутта



Методы Рунге–Кутта, в изложенной там схеме не предоставляют возможности оценки погрешности, аналогичной изложенной выше. С другой стороны, принципиальная возможность использовать идею вложенности существует.

В общем случае нам нужно найти такую таблицу Бутчера

	0				
c_2	a_{21}				
c_3	a_{31}	a_{32}			
\vdots	\vdots	\vdots	\ddots		
c_s	a_{s1}	a_{s2}	\cdots	$a_{s,s-1}$	
	b_1	b_2	\cdots	b_{s-1}	b_s
	\hat{b}_1	\hat{b}_2	\cdots	\hat{b}_{s-1}	\hat{b}_s

(8.57)

чтобы величина

$$y_{j+1} = y_j + \tau (b_1 k_1 + \cdots + b_s k_s)$$

имела порядок p , а величина

$$y_{j+1} = y_j + \tau (\hat{b}_1 k_1 + \cdots + \hat{b}_s k_s)$$

порядок q (обычно $q = p - 1$ или $q = p + 1$).

Поясним эту идею на конкретном примере, построив вложенные формулы порядков 2 и 3. Тогда таблица (8.58) будет иметь вид (полагаем $s = 3$, так как при $s = 2$ явных методов Рунге–Кутта третьего порядка построить нельзя)

	0			
c_2	a_{21}			
c_3	a_{31}	a_{32}		
	b_1	b_2	b_3	
	\hat{b}_1	\hat{b}_2	\hat{b}_3	

(8.58)



Параметры этой таблицы должны удовлетворять условиям

$$\begin{cases} b_1 + b_2 + b_3 = 1, \\ b_2 c_2 + b_3 c_3 = \frac{1}{2} \end{cases} \quad (\text{второй порядок}) \text{ и}$$

$$\begin{cases} \hat{b}_1 + \hat{b}_2 + \hat{b}_3 = 1, \\ \hat{b}_2 c_2 + \hat{b}_3 c_3 = \frac{1}{2}, \\ \hat{b}_2 c_2^2 + \hat{b}_3 c_3^2 = \frac{1}{3}, \\ \hat{b}_3 a_{32} c_2 = \frac{1}{6} \end{cases} \quad (\text{третий порядок}).$$

Выбрав $c_2 = 1$ и $b_3 = 0$, из первых двух уравнений получим: $b_2 = b_1 = \frac{1}{2}$. Осталось четыре уравнения с пятью неизвестными. Если положить $c_2 = \frac{1}{2}$, то $\hat{b}_1 = \frac{1}{6}$, $\hat{b}_2 = \frac{1}{6}$, $\hat{b}_3 = \frac{4}{6}$ и $a_{32} = \frac{1}{4}$. Таким образом, получившийся метод имеет вид

0			
1	1		
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	
$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0
$\frac{1}{6}$	$\frac{1}{6}$	$\frac{4}{6}$	

(8.59)

Его аббревиатура в литературе — RKF2(3) по первым буквам фамилий: Рунге–Кутта–Фельберга. Автором данного алгоритма — Фельбергом — были построены и другие варианты вложенных формул различных порядков точности.



8.2.6. Сходимость одношаговых методов решения задачи Коши

Исследуем сейчас вопрос о сходимости одношаговых методов решения задачи Коши, частные случаи построения которых мы рассмотрели выше.

Предположим, что исходное дифференциальное уравнение (8.4) имеет единственное решение на отрезке $t_0 \leq t \leq T$ не только при начальных данных (8.5), но и в случае любых начальных данных вида $u(\xi) = \eta$, где

$$(\xi, \eta) \in D = \{(\xi, \eta) : t_0 \leq \xi \leq T; u(\xi, T, u(T, t_0, u_0) - \varepsilon) \leq \eta \leq u(\xi, T, u(T, t_0, u_0) + \varepsilon)\},$$

а через $u(t, \xi, \eta)$ обозначено значение в точке t решения $u(t)$ уравнения (8.4) при начальных данных $u(\xi) = \eta$. Величина ε при этом может быть истолкована как верхняя граница допустимой погрешности в нахождении решения задачи (8.4), (8.5) в точке $t = T$ (см. рис. 8.1).

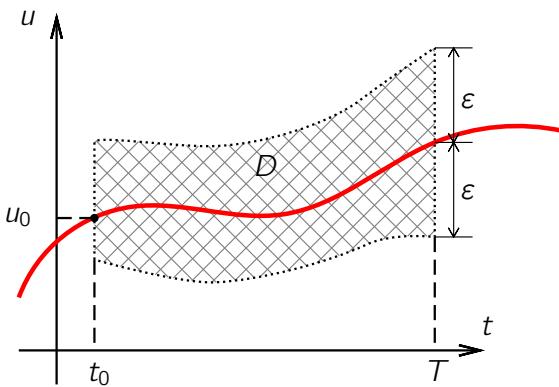


Рисунок 8.1

Вначале докажем вспомогательное утверждение, с помощью которого можно находить производные от решения по начальным данным.



Лемма 8.1. Имеет место соотношение

$$\frac{\partial u(t, \xi, \eta)}{\partial \eta} = \exp \int_{\xi}^t \frac{\partial f(x, u(x, \xi, \eta))}{\partial u} dx. \quad (8.60)$$

[[Доказательство](#)]

Пусть $M > 0$ таково, что

$$\left| \frac{\partial u(t, \xi, \eta)}{\partial \eta} \right| \leq M\xi, \quad \eta \in D, \quad \xi \leq t. \quad (*)$$

Будем считать, что $u(t, \xi, \eta)$ имеет все необходимые по ходу изложения производные по t , порядок которых определяется конструкцией избранного метода

$$y_{j+1} = F(y_j), \quad j = 0, 1, \dots, N-1. \quad (8.61)$$

Здесь, как и ранее, y_j — приближенное значение для $u(t_i, t_0, u_0)$, полученное при условии точного выполнения всех предусмотренных в (8.61) операций.

В действительности же мы вместо (8.61) имеем соотношения вида

$$\tilde{y}_{j+1} = F(\tilde{y}_j) - \delta_{j+1}, \quad j = 0, 1, \dots, N-1. \quad (8.62)$$

Разность $\varepsilon_0 = y_0 - \tilde{y}_0$ будем называть *погрешностью начального условия*, а величину $-\delta_{j+1}$ — *погрешностью округления* на $(j+1)$ -м шаге вычислительного процесса. При отсутствии ошибок начальных данных и округлений величину

$$\varepsilon_j = u(t_j, t_0, u_0) - y_j = u(t_j, t_0, u_0) - u(t_j, t_j, y_j)$$

обычно называют *погрешностью метода*.

На практике интерес представляет величина *погрешности приближенного решения*

$$\tilde{\varepsilon}_j = u(t_j, t_0, u_0) - \tilde{y}_j = u(t_j, t_0, u_0) - u(t_j, t_j, \tilde{y}_j). \quad (8.63)$$



Погрешность формулы (8.61) на каждом шаге реального вычислительного процесса можно определить посредством равенства

$$u(t_{j+1}, t_j, \tilde{y}_j) = F(u(t_j, t_j, \tilde{y}_j)) + r_{j+1} = F(\tilde{y}_j) + r_{j+1}. \quad (8.64)$$

Здесь r_{j+1} — фактически — [локальная погрешность](#) метода. Тогда, вычитая из (8.64) (8.62), получим:

$$u(t_{j+1}, t_j, \tilde{y}_j) - \tilde{y}_{j+1} = r_{j+1} + \delta_{j+1}. \quad (8.65)$$

Эта формула дает возможность оценить реально допускаемое отклонение на каждом шаге вычислительного процесса.

Теперь вернемся к соотношению (8.63), дающему [глобальную погрешность](#):

$$\tilde{\varepsilon}_j = u(t_j, t_0, y_0) - u(t_j, t_j, \tilde{y}_j) = u(t_j, t_0, y_0) - u(t_j, t_j, \tilde{y}_0) + \sum_{i=1}^j [u(t_j, t_{i-1}, \tilde{y}_{i-1}) - u(t_j, t_i, \tilde{y}_i)].$$

Отсюда, поскольку $u(t_j, t_{i-1}, \tilde{y}_{i-1}) = u(t_j, t_i, u(t_i, t_{i-1}, \tilde{y}_{i-1}))$, то

$$\tilde{\varepsilon}_j = u(t_j, t_0, y_0) - u(t_j, t_j, \tilde{y}_0) + \sum_{i=1}^j [u(t_j, t_i, u(t_i, t_{i-1}, \tilde{y}_{i-1})) - u(t_j, t_i, \tilde{y}_i)].$$

Последнее соотношение, используя формулу Лагранжа о конечном приращении, перепишем в виде

$$\begin{aligned} \tilde{\varepsilon}_j &= (y_0 - \tilde{y}_0) \frac{\partial u(t_j, t_0, \tilde{y}_0)}{\partial \eta} + \sum_{i=1}^j [u(t_i, t_{i-1}, \tilde{y}_{i-1}) - \tilde{y}_i] \frac{\partial u(t_j, t_i, \tilde{y}_i)}{\partial \eta} = \\ &= \tilde{\varepsilon}_0 \frac{\partial u(t_j, t_0, \tilde{y}_0)}{\partial \eta} + \sum_{i=1}^j (r_i + \delta_i) \frac{\partial u(t_j, t_i, \tilde{y}_i)}{\partial \eta}. \end{aligned} \quad (8.66)$$

Таким образом, отсюда, учитывая соотношение (*), получаем оценку

$$\begin{aligned} |\tilde{\varepsilon}_j| &\leq \left(\varepsilon_0 + \sum_{i=1}^j (|r_i| + |\delta_i|) \right) M \leq (\varepsilon_0 + j(r + \delta)) M \leq (\varepsilon_0 + N(r + \delta)) M = \left(\varepsilon_0 + \frac{(r + \delta)(T - t_0)}{\tau} \right) M, \\ &j = 1, 2, \dots, N. \end{aligned}$$



На основании этой оценки можно утверждать, что если $\varepsilon_0 \rightarrow 0$, $\frac{\delta}{\tau} \rightarrow 0$ и $\frac{r}{\tau} \rightarrow 0$ при $\tau \rightarrow 0$, то в любой точке отрезка $[t_0; T]$ приближенное решение задачи Коши, полученное с помощью одношагового метода (8.61), будет сходиться к точному решению этой задачи.

В частности, когда $\tilde{\varepsilon}_0 = 0$, $\delta_i = 0$ имеем условие сходимости в виде $\frac{r}{\tau} \rightarrow 0$. При этом обычно для локальной погрешности справедливо представление $r = O(\tau^{p+1})$ для некоторого $p > 0$. Поэтому данное условие автоматически выполняется, причем скорость сходимости будет иметь p -й порядок.

Отметим, что соотношение (8.66) с учетом (8.60) позволяет провести более подробный анализ зависимости погрешности от свойств исходного дифференциального уравнения. Так, например, если $\frac{\partial f}{\partial u} > 0$, то $\frac{\partial u}{\partial \eta} > 1$ и влияние погрешностей начальных данных и округления растет, а в случае $\frac{\partial f}{\partial u} < 0$ — ослабевает.



8.3. Многошаговые методы решения задачи Коши

[8.3.1. Введение](#)

[8.3.2. Методы Адамса](#)

[8.3.3. Общие линейные многошаговые методы](#)



8.3.1. Введение

Рассмотрим сейчас другой класс численных методов решения задачи Коши, отказавшись от условия одноступенчатости.

Вновь воспользуемся интегральным соотношением типа (8.13) предыдущей главы, которое сейчас будет удобно переписать в виде

$$u(t_{j+1}) = u(t_j) + \int_{t_j}^{t_{j+1}} u'(x) dx. \quad (8.67)$$

Предполагая известными (уже найденными) решения y_{j-k}, \dots, y_j , заменим функцию $u'(x)$, стоящую под знаком интеграла, некоторым ее интерполяционным приближением по выписанным выше значениям:

$$u'(x) \approx \varphi(x, y_{j-k}, \dots, y_j), \quad (8.68)$$

где φ — некоторая известная функция.

Очевидно, после такой замены интеграл может быть непосредственно вычислен. В результате получим некоторый [явный многошаговый метод](#) решения задачи Коши.



8.3.2. Методы Адамса

[Экстраполяционные методы Адамса](#)
[Интерполяционные методы Адамса](#)

Экстраполяционные методы Адамса

Чаще всего в (8.68) используется многочленное приближение. Ранее мы рассматривали наиболее часто используемые представления соответствующих интерполяционных многочленов. Одним из них является, например, [интерполяционный многочлен в форме Лагранжа](#):

$$u'(x) \approx L_k(x) = \sum_{i=0}^k \frac{\omega_{k+1}(x)}{(x - t_{j-i}) \omega'_{k+1}(t_{j-i})} y'(t_{j-i}) = \sum_{i=0}^k \frac{\omega_{k+1}(x)}{(x - t_{j-i}) \omega'_{k+1}(t_{j-i})} f(t_{j-i}, y_{j-i}). \quad (8.69)$$

Подставляя это выражение в (8.67), получим численный метод вида

$$y_{j+1} = y_j + \sum_{i=0}^k \beta_i f(t_{j-i}, y_{j-i}), \quad (8.70)$$

где

$$\beta_i = \int_{t_j}^{t_{j+1}} \frac{\omega_{k+1}(x)}{(x - t_{j-i}) \omega'_{k+1}(t_{j-i})} dx, \quad i = \overline{0, k}. \quad (8.71)$$

Формулы (8.70), (8.71) определяют [k-шаговый экстраполяционный метод Адамса](#) на неравномерной сетке.

Очевидно, с ростом k растет как порядок точности рассматриваемых методов, так и их сложность. В частном случае, при $k = 1$, получим:

$$u'(x) \approx L_1(x) = \frac{x - t_{j-1}}{t_j - t_{j-1}} f_j + \frac{x - t_j}{t_{j-1} - t_j} f_{j-1}$$



или, используя обозначение $t_j - t_{j-1} = \tau_{j-1}$

$$u'(x) \approx L_1(x) = \frac{x - t_{j-1}}{\tau_{j-1}} f_j - \frac{x - t_j}{\tau_{j-1}} f_{j-1}.$$

Отсюда

$$\begin{aligned} y_{j+1} &= y_j + \int_{t_j}^{t_{j+1}} \left(\frac{x - t_{j-1}}{\tau_{j-1}} f_j - \frac{x - t_j}{\tau_{j-1}} f_{j-1} \right) dx = \\ &= y_j + \frac{1}{2\tau_{j-1}} \left\{ \left[(t_{j+1} - t_{j-1})^2 - (t_j - t_{j-1})^2 \right] f_j - (t_{j+1} - t_j)^2 f_{j-1} \right\} = y_j + \frac{(\tau_j + \tau_{j-1})^2 - \tau_{j-1}^2}{2\tau_{j-1}} f_j - \frac{\tau_j^2}{2\tau_{j-1}} f_{j-1}. \end{aligned}$$

Экстраполяционные методы Адамса на равномерной сетке

Если сетка узлов равномерна, т.е. когда для всех значений j $\tau_j = const = \tau$, то предлагаемые методы имеют значительно более простой вид. Дадим их подробное описание.

Делая вновь, как и в способе Рунге–Кутта, замену переменных в интегральном тождестве (8.67), перепишем его в виде

$$u(t_{j+1}) = u(t_j) + \tau \int_0^1 u'(t_j + \alpha\tau) d\alpha. \quad (8.72)$$

Тогда описанная выше процедура приведет к методу

$$y_{j+1} = y_j + \tau \sum_{i=0}^k \beta_i f_{j-i}, \quad (8.73)$$

где

$$\beta_i = \frac{(-1)^i}{i! \cdot (k-i)!} \int_0^1 \frac{\alpha(\alpha+1)\dots(\alpha+k)}{\alpha+i} d\alpha, \quad i = \overline{0, k}. \quad (8.74)$$

В частности, имеем:

1) при $k = 0$:

$$\beta_0 = \int_0^1 d\alpha = 1$$

и (8.73) примет вид

$$y_{j+1} = y_j + \tau f_j.$$

2) при $k = 1$:

$$\beta_0 = \int_0^1 (\alpha + 1) d\alpha = \frac{3}{2}; \quad \beta_1 = - \int_0^1 \alpha d\alpha = -\frac{1}{2}$$

и

$$y_{j+1} = y_j + \frac{\tau}{2} (3f_j - f_{j-1}).$$

3) при $k = 2$:

$$\beta_0 = \frac{1}{2} \int_0^1 (\alpha + 1)(\alpha + 2) d\alpha = \frac{23}{12}; \quad \beta_1 = - \int_0^1 \alpha(\alpha + 2) d\alpha = -\frac{4}{3}; \quad \beta_2 = \frac{1}{2} \int_0^1 \alpha(\alpha + 1) d\alpha = \frac{5}{12}$$

и

$$y_{j+1} = y_j + \frac{\tau}{12} (23f_j - 16f_{j-1} + 5f_{j-2}).$$

Заметим, что более известно другое представление экстраполяционных методов Адамса, базирующееся на представлении интерполяционного многочлена в форме Ньютона. Поскольку интерполирование ведется по узлам $t_j, t_{j-1}, \dots, t_{j-k}$, то естественно применять интерполяционную формулу Ньютона для конца таблицы:

$$u'(t_j + \alpha\tau) \approx u'(t_j) + \frac{\alpha}{1!} \Delta u'(t_{j-1}) + \frac{\alpha(\alpha+1)}{2!} \Delta^2 u'(t_{j-2}) + \dots + \frac{\alpha(\alpha+1)\dots(\alpha+k-1)}{k!} \Delta^k u'(t_{j-k}), \quad (8.75)$$

где остаток интерполирования имеет вид

$$r_k(t_j + \alpha\tau) = \tau^{k+1} \frac{\alpha(\alpha+1)\dots(\alpha+k)}{(k+1)!} u^{(k+2)}(\xi), \quad t_{j-k} \leq \xi \leq t_{j+1}.$$



Подставляя вместо u' ее представление в (8.72) и выполняя интегрирование, получим экстраполяционный метод Адамса в виде

$$y_{j+1} = y_j + \tau \sum_{i=0}^k C_i \Delta^i f_{j-i}, \quad (8.76)$$

где

$$C_i = \frac{1}{i!} \int_0^1 \alpha (\alpha + 1) \dots (\alpha + i - 1) d\alpha. \quad (8.77)$$

Очевидно, локальная погрешность метода (8.76), (8.77), учитывая (8.75), будет иметь вид

$$r_k(t_j, \tau) = \tau^{k+2} \int_0^1 \frac{\alpha (\alpha + 1) \dots (\alpha + k)}{(k+1)!} u^{(k+2)}(\xi) d\alpha = C_{k+1} \tau^{k+2} u^{(k+2)}(\xi'), \quad (8.78)$$

т.е. метод (8.76), (8.77) является методом $(k+1)$ -го порядка точности. Из (8.77) следует, что

$$C_0 = \int_0^1 d\alpha = 1; \quad C_1 = \frac{1}{1!} \int_0^1 \alpha d\alpha = \frac{1}{2}; \quad C_2 = \frac{1}{2!} \int_0^1 \alpha (\alpha + 1) d\alpha = \frac{5}{12}; \dots$$

Таким образом, (8.76) может быть переписан в виде

$$y_{j+1} = y_j + \tau \left(f_j + \frac{1}{2} \Delta f_{j-1} + \frac{5}{12} \Delta^2 f_{j-2} + \dots + C_k \Delta^k f_{j-k} \right).$$

При этом приближенное решение вновь (как и в пошаговом варианте метода рядов) оказывается разложенным по последовательным главным частям.

Многошаговые методы характеризуются известной неоднородностью вычислительного процесса. Первые k значений y_j (начало таблицы) должны быть вычислены каким-либо другим способом (например, с помощью одношаговых методов).



Интерполяционные методы Адамса

Все изложенное в предыдущем пункте можно было бы повторить при одном лишь отличии: интерполирование функции $u'(x)$ проводить не по узлам $t_j, t_{j-1}, \dots, t_{j-k}$, а по узлам $t_{j+1}, t_j, \dots, t_{j-k}$.

Мы, однако, вновь более подробно остановимся на случае равномерной сетки узлов, так как тогда расчетные формулы будут иметь наиболее простой вид.

Сделаем в (8.67) замену переменной $x = t_{j+1} + \alpha\tau$. Получим:

$$u(t_{j+1}) = u(t_j) + \tau \int_{-1}^0 u'(t_{j+1} + \alpha\tau) d\alpha. \quad (8.79)$$

Вновь проинтерполируем подынтегральную функцию по формуле Ньютона для конца таблицы:

$$u'(t_{j+1} + \alpha\tau) \approx u'(t_{j+1}) + \frac{\alpha}{1!} \Delta u'(t_j) + \frac{\alpha(\alpha+1)}{2!} \Delta^2 u'(t_{j-1}) + \dots + \frac{\alpha(\alpha+1)\dots(\alpha+k)}{k!} \Delta^k u'(t_{j-k}).$$

Остаток интерполирования имеет вид

$$\rho_{k+1}(t_{j+1} + \alpha\tau) = \tau^{k+2} \frac{\alpha(\alpha+1)\dots(\alpha+k+1)}{(k+2)!} u^{(k+3)}(\xi), \quad t_{j-k} \leq \xi \leq t_{j+1}.$$

Выполнив интегрирование, получим **интерполяционный метод Адамса** вида

$$y_{j+1} = y_j + \tau \sum_{i=0}^{k+1} C_i^* \Delta^i f_{j+1-i}, \quad (8.80)$$

где

$$C_i^* = \frac{1}{i!} \int_{-1}^0 \alpha(\alpha+1)\dots(\alpha+i-1) d\alpha, \quad (8.81)$$

а для **локальной погрешности метода** справедливо представление

$$r_k(t_j, \tau) = \tau^{k+3} \int_0^1 \frac{\alpha(\alpha+1)\dots(\alpha+k+1)}{(k+2)!} u^{(k+3)}(\xi) d\alpha = C_{k+2}^* \tau^{k+3} u^{(k+3)}(\xi'), \quad t_{j-k} \leq \xi' \leq t_{j+1}, \quad (8.82)$$



В частности, из (8.81) следует, что

$$C_0^* = \int_{-1}^0 d\alpha = 1; \quad C_1^* = \frac{1}{1!} \int_{-1}^0 \alpha d\alpha = -\frac{1}{2}; \quad C_2^* = \frac{1}{2!} \int_{-1}^0 \alpha(\alpha+1) d\alpha = -\frac{1}{12}; \dots,$$

т.е. (8.80) можно переписать в виде

$$y_{j+1} = y_j + \tau \left(f_{j+1} - \frac{1}{2} \Delta f_j - \frac{1}{12} \Delta^2 f_{j-1} - \dots + C_{k+1}^* \Delta^{k+1} f_{j-k} \right). \quad (8.83)$$

Замечание 8.2. Ранее мы получили два представления для [экстраполяционных методов Адамса](#) (через конечные разности и значения функции). Здесь также легко получить второе (через значения функции) представление. Для этого достаточно либо просто заменить в (8.80) конечные разности их [выражениями через значения функции](#), либо, как и выше, интерполяционный многочлен брать в [форме Лагранжа](#).

В любом варианте получим такое семейство методов:

$y_{j+1} = y_j + \tau f_{j+1}$ — метод первого порядка ([неявный метод Эйлера](#));

$y_{j+1} = y_j + \frac{\tau}{2} (f_{j+1} + f_j)$ — метод второго порядка ([неявный метод трапеций](#));

$y_{j+1} = y_j + \frac{\tau}{12} (5f_{j+1} + 8f_j - f_{j-1})$ — метод третьего порядка;

.....

Заметим также, что в компьютерной реализации последние представления используются чаще, нежели представления через конечные разности.



8.3.3. Общие линейные многошаговые методы

Рассмотренные выше семейства [экстраполяционных](#) и [интерполяционных](#) методов Адамса являются частными случаями более общего семейства методов вида

$$\sum_{i=-1}^k a_i y_{j-i} = \tau \sum_{i=-1}^k b_i f(t_{j-i}, y_{j-i}), \quad (8.84)$$

которые носят название [линейных многошаговых методов](#) (ЛММ) без старших производных.

Очевидно, для методов Адамса $a_{-1} = 1$, $a_0 = -1$, $a_1 = \dots = a_k = 0$.

Получим сейчас общий вид условий, которым должны удовлетворять коэффициенты a_i и b_i метода (8.84). Для этих целей воспользуемся той же идеей, что и при рассмотрении [методов Рунге–Кутта](#): порядок метода должен быть максимальным. Запишем выражение для локальной погрешности метода (8.84):

$$r(t_j, \tau) = \sum_{i=-1}^k [a_i u(t_j - i\tau) - \tau b_i f(t_j - i\tau, u(t_j - i\tau))]. \quad (8.85)$$

Поскольку

$$u(t_j - i\tau) = u_j + \frac{(-i)\tau}{1!} u'_j + \frac{(-i)^2 \tau^2}{2!} u''_j + \dots,$$

$$f(t_j - i\tau, u(t_j - i\tau)) = u'(t_j - i\tau) = u'_j + \frac{(-i)\tau}{1!} u''_j + \frac{(-i)^2 \tau^2}{2!} u'''_j + \dots,$$

то, подставив эти разложения в (8.85), будем иметь:

$$r(t_j, \tau) = \left(\sum_{i=-1}^k a_i \right) u_j - \frac{\tau}{1!} \sum_{i=-1}^k (ia_i + b_i) u'_j + \frac{\tau^2}{2!} \sum_{i=-1}^k (i^2 a_i + 2ib_i) u''_j + \dots + \frac{(-\tau)^l}{l!} \sum_{i=-1}^k i^{l-1} (ia_i + lb_i) u_j^{(l)} + \dots.$$



Меню

Отсюда видно, что метод (8.84) будет иметь порядок точности p , если выполнены условия

$$\begin{cases} \sum_{i=-1}^k a_i = 0, \\ \sum_{i=-1}^k (ia_i + lb_i) i^{l-1} = 0, \quad l = 1, 2, \dots, p. \end{cases} \quad (8.86)$$

Используя условия порядка (8.86), можно на основе конструкции (8.84) получить большинство из используемых ныне в вычислительной практике линейных многошаговых методов.

Замечание 8.3. Аналогичную (8.84) конструкцию можно получить, если отказаться от «запрета» на использование старших производных от решения. Она будет иметь вид

$$\sum_{i=-1}^k a_i y_{j-i} = \sum_{l=0}^p \tau^{l+1} \sum_{i=-1}^k b_{li} f^{(l)}(t_{j-i}, y_{j-i}). \quad (8.87)$$

В литературе эта конструкция носит название обобщенных линейных многошаговых методов.

Замечание 8.4. Чаще оказывается целесообразным часть параметров метода отдать не на достижение максимального порядка аппроксимации, а на то, чтобы добиться выполнения некоторых других важных свойств (например, расширения области устойчивости).



8.4. Понятие устойчивости численных методов решения задачи Коши

8.4.1. Корневое условие

8.4.2. Устойчивость на модельном уравнении

Вновь, если не оговорено особо, будем предполагать, что рассматривается случай одного обыкновенного дифференциального уравнения первого порядка.

Ранее мы рассматривали достаточно общие определения, касающиеся таких понятий как «корректность», «устойчивость», «сходимость» численных методов. Естественно, такие общие определения очень часто нуждаются в конкретизации, для того чтобы их можно было реально применить к исследованию конкретных свойств того или иного алгоритма.

С этих позиций обсудим сейчас термин «устойчивость» применительно к методам решения задачи Коши.



8.4.1. Корневое условие

Прежде всего, заметим, что все численные методы решения задачи Коши представляют собой *разностные уравнения* различных порядков: одношаговые методы — первого порядка, многошаговые (k -шаговые) — k -го порядка.

Далее, вполне очевидным представляется тот факт, что численный метод должен более или менее адекватно отражать истинные свойства решения исследуемой (решаемой) дифференциальной задачи и причем, желательно, в достаточно широком диапазоне правых частей уравнения (т.е. функций $f(t, u)$).

В частности, положив $f(t, u) \equiv 0$, получим следующую запись линейного многошагового метода (8.84):

$$a_{-1}y_{j+1} + a_0y_j + \cdots + a_ky_{j-k} = 0,$$

т.е. будем иметь линейное разностное уравнение порядка $(k+1)$ с постоянными коэффициентами. Его общее решение может быть записано в виде

$$y_j = \sum_{i=-1}^k C_i q_i^j,$$

где C_i — коэффициенты, определяемые из дополнительных (например, начальных) условий, aq_i , $i = \overline{0, k}$ — корни *характеристического уравнения*

$$a_{-1}q^{k+1} + a_0q^k + \cdots + a_k = 0. \quad (8.88)$$

Отсюда следует, что если $|q_i| > 1$ хотя бы для некоторого значения i , то $|y_j| \xrightarrow{j \rightarrow \infty} \infty$, в то время как решение исходного уравнения $u' = f(t, u)$ при $f(t, u) \equiv 0$, очевидно, представляет собой константу.

Аналогичная ситуация имеет место и в случае, если $|q_{i_0}| = 1$, но кратность этого корня выше единицы, так как тогда в конструкции общего решения перед сомножителем $q_{i_0}^j$ появляется многочленный по j коэффициент, учитывающий кратность.

Определение. Будем говорить, что численный метод удовлетворяет *условию корней*, если все корни q_0, \dots, q_k характеристического уравнения (8.88) лежат внутри или на границе единичного круга комплексной плоскости, причем на границе круга нет кратных корней.



Очевидно, метод, не удовлетворяющий условию корней (или корневому условию), для вычислений не пригоден. Это следует иметь в виду при построении различных алгоритмов. Заметим, что рассмотренные нами ранее классы методов (как [Адамса](#), так и одношаговые [типа Рунге–Кутта](#) и [последовательного повышения порядка точности](#)) корневому условию удовлетворяют, поскольку для них характеристическое уравнение (8.88) будет иметь вид

$$a_{-1}q + a_0 = 0$$

при $a_{-1} = 1$, $a_0 = -1$, т.е.

$$q - 1 = 0$$

и имеет единственный корень $q = 1$.



8.4.2. Устойчивость на модельном уравнении

Однако не все численные методы, удовлетворяющие [корневому условию](#), всегда пригодны для расчетов. Понятно, что при произвольной правой части дифференциального уравнения (функции f) вряд ли возможно получить сколько-нибудь эффективных оценок по этому поводу. Поэтому в отличие от такого свойства численных методов как [аппроксимация](#), о которой мы говорили выше, и которая может быть исследована принципиально при любой правой части, другие свойства приближенных методов, характеризующие [качественное поведение](#) приближенного решения, могут быть исследованы только на задачах определенного вида — [модельных уравнениях](#).

В частности, при исследовании свойства устойчивости в качестве такого модельного уравнения чаще всего рассматривают уравнение

$$u'(t) = \lambda u(t), \quad (8.89)$$

где λ — произвольное комплексное число с отрицательной вещественной частью ($\operatorname{Re} \lambda < 0$). Отчасти выбор в качестве модели уравнения (8.88) объясняется тем, что оно представляет собой линейное (однородное) приближение к задаче общего вида (вспомним, что исследование устойчивости по Ляпунову для дифференциальных уравнений проводится на линейном приближении).

Отметим, что поскольку точное решение уравнения (8.88) имеет вид $u(t) = Ce^{\lambda t}$, то при указанных значениях λ задача (8.88) устойчива, причем $u(t) \xrightarrow[t \rightarrow \infty]{} 0$ и в частности, при любом значении шага τ выполняется условие

$$|u(t + \tau)| \leq |u(t)|,$$

т.е. модуль решения монотонно не возрастает. Естественно было бы потребовать качественно такого же поведения и от приближенного решения, доставляемого тем или иным численным методом.

Легко видеть, что применение линейного многошагового метода (8.84) к решению уравнения (8.88) приводит к разностному уравнению вида

$$\sum_{i=-1}^k (a_i - z b_i) y_{j-i} = 0, \quad (8.90)$$

где $z = \tau\lambda$.



Аналогичные разностные уравнения (но только первого порядка) получаются, если к решению уравнения (8.88) применить соответствующий метод Рунге–Кутта или последовательного повышения порядка точности.

Определение. Численный метод решения задачи Коши будем называть *устойчивым* при некотором значении z , если при данном значении z устойчиво соответствующее ему разностное уравнение (8.90), получающееся вследствие применения исследуемого метода к решению модельного уравнения (8.89).

Очевидно, для того чтобы метод был устойчивым, достаточно, чтобы все корни соответствующего характеристического уравнения по модулю не превосходили единицы.

Определение. *Областью устойчивости* численного метода будем называть множество всех точек z комплексной плоскости, для которых данный метод *устойчив*.

Определение. *Интервалом устойчивости* численного метода будем называть пересечение *области устойчивости* с вещественной осью координат.

Рассмотрим некоторые примеры.

1. Явный метод Эйлера:

$$y_{j+1} = y_j + \tau f_j.$$

Применяя его к уравнению (8.88), будем иметь:

$$y_{j+1} = (1 + z) y_j.$$

Единственный корень соответствующего характеристического уравнения здесь $q = 1 + z$. Поэтому условие устойчивости примет вид

$$|1 + z| \leq 1. \quad (8.91)$$

Если λ — комплексное число, то, записав z в виде $z = x + iy$, перепишем (8.91) в виде $|1 + x + iy| \leq 1$ или $(1 + x)^2 + y^2 \leq 1$.

Очевидно, последнее неравенство, описывающее *область устойчивости* явного метода Эйлера, задает в комплексной плоскости круг единичного радиуса с центром в точке $(-1, 0)$. Пересечением данной



области с вещественной осью будет отрезок $[-2; 0]$, который и будет [интервалом устойчивости](#) явного метода Эйлера. (Заметим, что интервал устойчивости на самом деле искать гораздо проще, чем область, поскольку для этого достаточно просто решить неравенство (8.91) над полем вещественных чисел).

2. Неявный метод Эйлера:

$$y_{j+1} = y_j + \tau f_{j+1},$$

откуда

$$(1 + z) y_{j+1} = y_j$$

или

$$y_{j+1} = \frac{1}{1 - z} y_j$$

. Тогда $q = \frac{1}{1-z}$ и условие устойчивости примет вид $\left| \frac{1}{1-z} \right| \leq 1$ или $|1 - z| \geq 1$.

Таким образом, [областью устойчивости](#) неявного метода Эйлера является внешность круга единичного радиуса с центром в точке $(1; 0)$. [Интервалом устойчивости](#) является вся числовая прямая, кроме промежутка $(0; 2)$. Следовательно, при $\operatorname{Re} \lambda < 0$ шаг численного интегрирования τ может быть любым (в то время как для явного Эйлера от ограничен сверху величиной $\frac{2}{|\lambda|}$).

Определение. Численный метод будем называть [A-устойчивым](#), если его [область устойчивости](#) содержит всю левую полуплоскость $\operatorname{Re} z < 0$.

Сущность данного определения состоит в том, что A-устойчивый метод является абсолютно (т.е. при любых $\tau > 0$) устойчивым, если устойчиво решение исходного дифференциального уравнения. Отметим, что [неявный метод Эйлера](#) является A-устойчивым, а [явный](#) — нет. Заметим, однако, что неявный метод Эйлера является устойчивым и в той области, где исходная задача является неустойчивой (!).

Теорема 8.1. *Не существует явных линейных A-устойчивых методов.*

Это означает, что при использовании таких методов всегда будут иметь место ограничения на выбор допустимой величины шага τ , подобные ограничению, характерному для явного метода Эйлера. В то же время среди неявных линейных методов A-устойчивые, как мы видели, существуют.

Подводя итоги, можем сформулировать следующую процедуру исследования устойчивости численных методов решения задачи Коши:



- 1) Применяя исследуемый метод к решению уравнения (8.89), получаем разностное уравнение, которому удовлетворяет приближенное решение;
- 2) Записываем соответствующее [характеристическое уравнение](#);
- 3) Находим корни характеристического уравнения (q_i , $i = 1, \dots, k$);
- 4) Решение системы неравенств $|q_i| \leq 1$, $i = 1, \dots, k$, дает искомую область устойчивости.

Заметим, однако, что практическая реализация изложенного алгоритма может натолкнуться на значительные технические трудности (особенно это касается случаев комплексных λ для многостадийных методов, а также методов многошаговых). Поэтому практически для построения областей устойчивости используют прием, который носит название *метод множества точек границы* и состоит в следующем: точка z комплексной плоскости будет принадлежать границе области устойчивости, если при данном значении z выполняется равенство $\max_i |q_i| = 1 \stackrel{\text{def}}{=} |q^*|$ или $q^* = e^{i\varphi}$, $\varphi \in [0; 2\pi]$ (здесь i — мнимая единица). Решая записанное уравнение относительно z (возможно, при фиксированных значениях φ из указанного промежутка), мы получаем множество точек, составляющих границу области устойчивости. Далее остается определить (например, путем подстановки), по какую сторону границы находится сама область.

Примеры исследования устойчивости.

1. Метод последовательного повышения порядка точности второго порядка

$$\begin{cases} y_{j+\frac{1}{2}} = y_j + \frac{\tau}{2} f_j, \\ y_{j+1} = y_j + \tau f_{j+\frac{1}{2}}. \end{cases}$$

Применяя данный метод к решению уравнения (8.89), получим:

$$y_{j+1} = \left(1 + z + \frac{z^2}{2}\right) y_j.$$



Отсюда

$$q = 1 + z + \frac{z^2}{2} = e^{i\varphi}$$

и

$$z = z(\varphi) = -1 \pm \sqrt{2e^{i\varphi} - 1}.$$

Кривая $z(\varphi)$ и есть граница области устойчивости, а сама область есть внутренность данной кривой.

2. Экстраполяционный метод Адамса второго порядка.

$$y_{j+1} = y_j + \frac{\tau}{2} (3f_j - f_{j-1}).$$

Разностное уравнение имеет вид

$$y_{j+1} - \left(1 + \frac{3}{2}z\right)y_j + \frac{z}{2}y_{j-1} = 0.$$

Тогда характеристическое уравнение будет таким: $q^2 - \left(1 + \frac{3}{2}z\right)q + \frac{z}{2} = 0$.

Подставим сюда $q = e^{i\varphi}$ и найдем границу области устойчивости:

$$z(\varphi) = 2 \frac{e^{2i\varphi} - e^{i\varphi}}{3e^{i\varphi} - 1}. \quad (*)$$

Замечание 8.5. Для определения интервала устойчивости вдоль вещественной оси для линейных многошаговых методов в уравнение типа (*) достаточно подставить $\varphi = \pi$. Так, для указанного выше метода получим левую границу интервала

$$z(\pi) = 2 \cdot \frac{1 + 1}{3 \cdot (-1) - 1} = -1,$$

т.е. экстраполяционный метод Адамса второго порядка устойчив на отрезке $z \in [-1; 0]$.

Замечание 8.6. В случае систем обыкновенных дифференциальных уравнений можно показать, что соответствующие условия устойчивости будут иметь такой же вид, как и в случае одного уравнения, но с заменой параметра λ на максимальное по модулю собственное значение матрицы Якоби системы.



8.5. Жесткие задачи и методы их решения

8.5.1. Явление жесткости

8.5.2. Методы, применяемые для решения жестких систем



8.5.1. Явление жесткости

Задачи, называемые жесткими, весьма разнообразны, и дать математически строгое определение жесткости непросто. Поэтому в литературе можно встретить различные определения жесткости, отличающиеся степенью строгости. Сущность же явления жесткости состоит в том, что решение, которое необходимо вычислить, меняется медленно, однако в любой его окрестности существуют быстро затухающие возмущения. Наличие таких возмущений затрудняет получение медленно меняющегося решения численным способом. При этом жесткими могут как скалярные дифференциальные уравнения, так и, что встречается особенно часто, системы обыкновенных дифференциальных уравнений. Приведем вначале некоторые примеры.

1. Скалярное уравнение

$$\begin{cases} u'(t) = \lambda u(t) + F'(t) - \lambda F(t), & t > 0, \quad \lambda \ll 0, \\ u(0) = u_0, \end{cases} \quad (8.92)$$

где $F(t)$ — медленно меняющаяся функция, зависящая только от t (например, $F(t) = \text{th } t$). Решение задачи (8.92) имеет вид

$$u(t) = F(t) + e^{\lambda t} [u_0 - F(0)].$$

Так как $\lambda \ll 0$, то ясно, что уже после очень небольшого отрезка времени второе слагаемое в решении практически отсутствует и, таким образом, на большей части отрезка интегрирования преобладает медленно меняющаяся функция, которая принципиально может быть достаточно хорошо приближенно описана на сетке с крупным шагом τ . В то же время, если применить к решению задачи (8.92) явный метод Эйлера (или любой другой явный метод типа Рунге–Кутта), то легко видеть, что допустимый шаг интегрирования, как и в случае модельного уравнения (8.89), будет определяться величиной $\tau \leq -\frac{2}{\lambda}$, т.е. будет очень малым.

2. Рассмотрим теперь систему из двух независимых уравнений

$$\begin{cases} u'_1(t) = -\lambda_1 u_1(t), \\ u'_2(t) = -\lambda_2 u_2(t), \quad t > 0, \quad \lambda_2 \gg \lambda_1 > 0. \end{cases} \quad (8.93)$$



Эта система имеет решение $u(t) = (u_1(t), u_2(t))^T = (u_1^0 e^{-\lambda_1 t}, u_2^0 e^{-\lambda_2 t})^T$. При выписанных услови-ях на λ_1 и λ_2 , очевидно, компонента $u_2(t)$ решения затухает гораздо быстрее, чем $u_1(t)$ и, начиная с некоторого момента t поведение вектора $u(t)$ почти полностью определяется компонентой $u_1(t)$. Однако при решении системы (8.93) численным методом величина шага интегрирования, как правило, определяется компонентой $u_2(t)$, не существенной с точки зрения поведения решения системы. Например, используя тот же явный метод Эйлера, мы из первого уравнения имеем ограничение на шаг $\tau \leq \frac{2}{\lambda_1}$, а из второго — $\tau \leq \frac{2}{\lambda_2}$ и, таким образом, ясно, что для решения системы (8.93) как цельного математического объекта шаг τ ограничен величиной $\frac{2}{\lambda_2}$.

Такая же ситуация типична и при решении любой системы обыкновенных дифференциальных уравнений вид

$$u'(t) = Au(t), \quad (8.94)$$

если матрица этой системы имеет большой разброс собственных значений.

Определение. Система обыкновенных дифференциальных уравнений (8.94) с постоянной $(n \times n)$ -матрицей A называется **жесткой**, если:

1) $\operatorname{Re} \lambda_k < 0$, $k = \overline{1, n}$ (т.е. задача устойчива);

2) отношение $S = \frac{\max_{1 \leq k \leq n} |\operatorname{Re} \lambda_k|}{\min_{1 \leq k \leq n} |\operatorname{Re} \lambda_k|}$ велико (например, $S > 10$). Число S иногда называют коэффициентом жесткости системы (8.94).

Если в (8.94) матрица A будет зависеть от t , то, очевидно, и $S = S(t)$, т.е. коэффициент жесткости может меняться с течением времени.

Поскольку система нелинейных обыкновенных дифференциальных уравнений вида $u'(t) = f(t, u(t))$ может быть в окрестности некоторого известного решения $v(t)$ заменена линейной системой

$$u'(t) = f_u(t, v + \theta(u - v))u + b(t),$$

где f_u — матрица Якоби системы, а $b(t) = f(t, v) - f_u(t, v + \theta(u - v))v$, то понятие жесткости для нелинейных систем может быть определено аналогично. Заметим, однако, что за пределами класса систем линейных обыкновенных дифференциальных уравнений с постоянной матрицей полагаться на спектр как на



Вверх

Назад

Вперёд

Пред.

След.

Указатель

Помощь

Экран

источник надежной информации о распространении погрешности уже нельзя (это показывают известные из литературы примеры (см., например, [7])).



8.5.2. Методы, применяемые для решения жестких систем

[Неявные методы Рунге–Кутта](#)

[Формулы дифференцирования назад](#)

[Реализация неявных методов](#)

Учитывая все сказанное выше, можно сделать вывод, что для решения жестких задач наиболее пригодны те численные методы, которые требуют наиболее слабых ограничений на величину шага численного интегрирования из соображений устойчивости. В настоящее время наиболее часто для этих целей используют либо неявные методы, среди которых, как мы видели, встречаются [A-устойчивые](#) (правда, при этом не следует думать, что все неявные методы будут в этом смысле хорошими), либо методы, специально сконструированные для решения задач конкретного вида.

Неявные методы Рунге–Кутта

Важным классом одношаговых методов, применяемых для решения жестких задач, являются неявными методами неявные [методы Рунге–Кутта](#). Пользуясь полученными там же [условиями порядка](#), рассмотрим подробнее некоторые частные случаи.

Одностадийные методы. Эти методы имеют вид

c_1	a_{11}
	b_1

В отличие от аналогичных явных методов Рунге–Кутта, они зависят от трех параметров. Поэтому здесь возможно построение методов не только первого порядка.

1) *Методы первого порядка.* При выводе условий порядка мы фактически не пользовались свойством явности. И действительно, эти условия оказываются универсальными и в нашем случае имеют вид

$$\left\{ \begin{array}{l} b_1 = 1, \\ c_1 = a_{11}. \end{array} \right. \quad (8.95)$$



Таким образом, имеем два уравнения с тремя неизвестными. Полагая $c_1 = a_{11} = \alpha$, получим однопараметрическое семейство методов первого порядка

$$\begin{array}{c|c} \alpha & \alpha \\ \hline & 1 \end{array}$$

или в развернутом виде

$$\begin{cases} y_{j+1} = y_j + \tau k_1, \\ k_1 = f(t_j + \alpha\tau, y_j + \alpha\tau k_1). \end{cases} \quad (8.96)$$

При выработке рекомендаций по выбору параметра вспомним, что основной целью при конструировании неявных методов является, по сути дела, расширение **области устойчивости** (вплоть до А-устойчивости). Поэтому проведем исследование **устойчивости** построенного семейства. Полагая $f = \lambda y$, имеем:

$$k_1 = \lambda (y_j + \alpha\tau k_1),$$

откуда

$$k_1 = \frac{\lambda}{1 - \alpha z} y_j$$

и, следовательно,

$$y_{j+1} = y_j + \frac{z}{1 - \alpha z} y_j,$$

т.е.

$$y_{j+1} = \frac{1 + (1 - \alpha) z}{1 - \alpha z} y_j.$$

Неравенство

$$\left| \frac{1 + (1 - \alpha) z}{1 - \alpha z} \right| \leq 1$$

равносильно неравенству

$$|1 + (1 - \alpha) z|^2 \leq |1 - \alpha z|^2$$

или

$$(1 + (1 - \alpha) x)^2 + (1 - \alpha)^2 y^2 \leq (1 - \alpha x)^2 + \alpha^2 y^2,$$



где x и y — соответственно вещественная и мнимая части комплексного числа $z = \tau\lambda$. Приводя в последнем неравенстве подобные, перепишем его в виде

$$2x + (1 - 2\alpha)x^2 + (1 - 2\alpha)y^2 \leq 0. \quad (8.97)$$

Теперь остается рассмотреть три случая:

- а) $\alpha = \frac{1}{2}$. В этом случае (8.97) превращается в неравенство $x \leq 0$, что равносильно [А-устойчивости](#) (8.96), так как область устойчивости в точности совпадает с левой полуплоскостью;
- б) $\alpha > \frac{1}{2}$. В этом случае (8.97) может быть переписано в виде

$$\left(x - \frac{1}{2\alpha - 1}\right)^2 + y^2 \geq \left(\frac{1}{2\alpha - 1}\right)^2.$$

Последнее же означает, что областью устойчивости является внешность круга радиуса $\frac{1}{2\alpha - 1}$ с центром в точке $(\frac{1}{2\alpha - 1}, 0)$ и, как легко видеть, эта область целиком содержит всю левую полуплоскость, т.е. для всех рассматриваемых α метод (8.96) также является А-устойчивым;

- в) $\alpha < \frac{1}{2}$. В этом случае (8.97) следует переписать в виде

$$\left(x + \frac{1}{1 - 2\alpha}\right)^2 + y^2 \leq \left(\frac{1}{1 - 2\alpha}\right)^2.$$

Следовательно, областью устойчивости является внутренность круга радиуса $\frac{1}{1 - 2\alpha}$ с центром в точке $(-\frac{1}{1 - 2\alpha}, 0)$, т.е. (8.96) в этом случае является условно устойчивым.

Таким образом, проведенный анализ показывает, что выбор параметра α в (8.96) должен быть подчинен условию $\alpha \geq \frac{1}{2}$.

2) *методы второго порядка*. В этом случае к уравнениям (8.95), рассмотренным выше, добавляется еще одно уравнение:

$$b_1 c_1 = \frac{1}{2}.$$

Получившаяся система из трех уравнений с тремя неизвестными, как легко видеть, имеет единственное



решение: $b_1 = 1$; $c_1 = a_{11} = \frac{1}{2}$. Таким образом, единственный одностадийный метод второго порядка имеет вид

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$$

или в развернутом виде

$$\begin{cases} y_{j+1} = y_j + \tau k_1, \\ k_1 = f\left(t_j + \frac{\tau}{2}, y_j + \frac{\tau}{2} k_1\right). \end{cases} \quad (8.98)$$

Очевидно, полученный метод является частным случаем семейства (8.96), соответствующим случаю $\alpha = \frac{1}{2}$, и потому является [A-устойчивым](#).

Двухстадийные методы. В соответствии с общей схемой эти методы в [форме Бутчера](#) имеют вид

$$\begin{array}{c|cc} c_1 & a_{11} & a_{12} \\ \hline c_2 & a_{21} & a_{22} \\ \hline b_1 & b_2 \end{array}$$

и, таким образом, зависят от восьми произвольных параметров. Учитывая это, исследование начнем со случая

a) *Методы второго порядка.* Условия порядка в этом случае будут иметь вид

$$\begin{cases} b_1 + b_2 = 1, \\ b_1 c_1 + b_2 c_2 = \frac{1}{2}, \\ c_1 = a_{11} + a_{12}, \\ c_2 = a_{21} + a_{22}. \end{cases} \quad (8.99)$$

В этом варианте свободных параметров все равно достаточно много (четыре), и мы, не задаваясь целью провести полное исследование, распорядимся их выбором с целью упрощения реализации получаемых методов. Для этого зададим верхнюю строку таблицы Бутчера нулевой (т.е. положим $a_{11} = a_{12} = c_1 = 0$). Кроме того, положим $c_2 = \alpha$, $a_{22} = \beta$. Тогда оставшиеся неизвестные однозначно определяются из системы



(8.99): $a_{21} = \alpha - \beta$, $b_2 = \frac{1}{2\alpha}$, $b_1 = \frac{2\alpha - 1}{2\alpha}$. Таким образом, получаем двухпараметрическое семейство методов второго порядка точности

0	0	0
α	$\alpha - \beta$	β
	$1 - \frac{1}{2\alpha}$	$\frac{1}{2\alpha}$

или в развернутой форме

$$\begin{cases} y_{j+1} = y_j + \frac{\tau}{2\alpha} [(2\alpha - 1) k_1 + k_2] , \\ k_1 = f(t_j, y_j) , \\ k_2 = f(t_j + \alpha\tau, y_j + \tau[(\alpha - \beta) k_1 + \beta k_2]) . \end{cases} \quad (8.100)$$

Заметим, что построенный метод является своеобразным аналогом семейства [явных двухстадийных методов Рунге–Кутта второго порядка](#).

Вновь анализ возможного выбора параметров проведем на основе требования [устойчивости](#). Применяя (8.100) к модельному уравнению, будем иметь: $k_1 = \lambda y_j$. Тогда для определения k_2 получим уравнение

$$k_2 = \lambda [y_j + \tau(\alpha - \beta)\lambda y_j + \tau\beta k_2] ,$$

откуда

$$k_2 = \lambda y_j \cdot \frac{1 + (\alpha - \beta)z}{1 - \beta z} .$$

Следовательно,

$$\begin{aligned} y_{j+1} &= y_j + \frac{\tau}{2\alpha} \left[(2\alpha - 1) \lambda y_j + \lambda y_j \cdot \frac{1 + (\alpha - \beta)z}{1 - \beta z} \right] = y_j \left[1 + \frac{z}{2\alpha} \left(2\alpha - 1 + \frac{1 + (\alpha - \beta)z}{1 - \beta z} \right) \right] = \\ &= y_j \left[1 + \frac{z}{2\alpha} \cdot \frac{2\alpha - 1 - \beta(2\alpha - 1)z + 1 + (\alpha - \beta)z}{1 - \beta z} \right] = y_j \left[1 + \frac{z}{2\alpha} \cdot \frac{2\alpha + \alpha z(1 - 2\beta)}{1 - \beta z} \right] = \\ &= y_j \left[1 + \frac{z + z^2(\frac{1}{2} - \beta)}{1 - \beta z} \right] = y_j \cdot \frac{1 + (1 - \beta)z + (\frac{1}{2} - \beta)z^2}{1 - \beta z} . \end{aligned}$$



Отсюда видим, что требование [A-устойчивости](#) может быть выполнено лишь при значении параметра β , равном $\frac{1}{2}$ (так как в противном случае степень числителя множителя перехода будет выше степени знаменателя, и, следовательно, при больших по модулю значениях z множитель перехода будет заведомо больше единицы по модулю, при $\beta = \frac{1}{2}$ мы получим исследованный выше случай). Таким образом, в (8.100) параметр β следует полагать равным $\frac{1}{2}$. Отметим в этом варианте два частных значения α :

- 1) $\alpha = 1$. В этом случае получаем метод,

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline \frac{1}{2} & \frac{1}{2} \end{array}$$

являющийся аналогом неявного метода трапеций, о котором мы упоминали ранее.

- 2) $\alpha = \frac{1}{2}$. Получающийся в этом случае метод

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \hline 0 & 1 \end{array}$$

по сути, является методом (8.98) (т.е. одностадийным!) (аналог неявной формулы средних прямоугольников).

- б) Методы третьего порядка. Условия порядка имеют вид

$$\left\{ \begin{array}{l} b_1 + b_2 = 1, \\ b_1 c_1 + b_2 c_2 = \frac{1}{2}, \\ b_1 c_1^2 + b_2 c_2^2 = \frac{1}{3}, \\ b_1 (a_{11} c_1 + a_{12} c_2) + b_2 (a_{21} c_1 + a_{22} c_2) = \frac{1}{6}, \\ c_1 = a_{11} + a_{12}, \\ c_2 = a_{21} + a_{22}. \end{array} \right.$$

Таким образом, здесь шесть уравнений и восемь неизвестных. Исходя из соображений простоты даль-



нейшой реализации метода, положим $a_{12} = 0$, $a_{11} = a_{22}$. Тогда, с учетом пятого уравнения, систему можно переписать в виде

$$\begin{cases} b_1 + b_2 = 1, \\ b_1 c_1 + b_2 c_2 = \frac{1}{2}, \\ b_1 c_1^2 + b_2 c_2^2 = \frac{1}{3}, \\ b_1 c_1^2 + b_2 (a_{21} c_1 + c_1 c_2) = \frac{1}{6}, \\ c_2 = a_{21} + c_1. \end{cases}$$

Подставляя в четвертое уравнение вместо a_{12} его выражение через c_1 и c_2 , получим систему из четырех уравнений с четырьмя неизвестными:

$$\begin{cases} b_1 + b_2 = 1, \\ b_1 c_1 + b_2 c_2 = \frac{1}{2}, \\ b_1 c_1^2 + b_2 c_2^2 = \frac{1}{3}, \\ b_1 c_1^2 + b_2 (2c_1 c_2 - c_1^2) = \frac{1}{6}. \end{cases}$$

Вычитая из второго уравнения последней системы первое, умноженное на c_1 , из третьего — второе, умноженное на c_1 , и из четвертого — третье, перепишем систему в виде

$$\begin{cases} b_1 + b_2 = 1, \\ b_2 (c_2 - c_1) = \frac{1}{2} - c_1, \\ b_2 c_2 (c_2 - c_1) = \frac{1}{3} - \frac{1}{2} c_1, \\ b_2 (c_2 - c_1)^2 = \frac{1}{6}. \end{cases} (*)$$

Разделив третье уравнение полученной системы на второе, получим:

$$c_2 = \frac{\frac{1}{3} - \frac{1}{2} c_1}{\frac{1}{2} - c_1}. \quad (8.101)$$

Проделав аналогичную операцию с четвертым и вторым уравнениями, найдем:

$$c_2 - c_1 = \frac{\frac{1}{6}}{\frac{1}{2} - c_1}$$



или

$$c_2 = \frac{\frac{1}{2}c_1 - c_1^2 + \frac{1}{6}}{\frac{1}{2} - c_1}. \quad (8.102)$$

Приравнивая теперь правые части формул (8.101) и (8.102), получим уравнение

$$\frac{1}{3} - \frac{1}{2}c_1 = \frac{1}{2}c_1 - c_1^2 + \frac{1}{6}$$

или

$$c_1^2 - c_1 + \frac{1}{6} = 0.$$

Отсюда

$$c_1 = \frac{3 \pm \sqrt{3}}{6}.$$

Для определенности положим $c_1 = \frac{3-\sqrt{3}}{6}$. Тогда из (8.101) следует, что $c_2 = \frac{3+\sqrt{3}}{6}$. Следовательно (четвертое из уравнений системы (*)), $b_2 = \frac{1}{2}$. Наконец, $b_1 = \frac{1}{2}$ и $a_{21} = c_2 - c_1 = \frac{\sqrt{3}}{3}$.

Таким образом, искомый метод третьего порядка будет иметь вид

$\frac{3-\sqrt{3}}{6}$	$\frac{3-\sqrt{3}}{6}$	0
$\frac{3+\sqrt{3}}{6}$	$\frac{\sqrt{3}}{3}$	$\frac{3-\sqrt{3}}{6}$
<hr/>	<hr/>	<hr/>
$\frac{1}{2}$	$\frac{1}{2}$	

или в развернутой форме

$$\begin{cases} y_{j+1} = y_j + \frac{\tau}{2} [k_1 + k_2], \\ k_1 = f \left(t_j + \frac{3-\sqrt{3}}{6} \tau, y_j + \frac{3-\sqrt{3}}{6} \tau k_1 \right), \\ k_2 = f \left(t_j + \frac{3+\sqrt{3}}{6} \tau, y_j + \tau \left[\frac{\sqrt{3}}{3} k_1 + \frac{3-\sqrt{3}}{6} k_2 \right] \right). \end{cases} \quad (8.103)$$

Метод (8.103) представляет собой пример диагонально неявного метода.



Формулы дифференцирования назад

Положим в конструкции (8.95) общих линейных многошаговых методов $b_0 = b_1 = \dots = b_k = 0$. Тогда получим конструкцию следующего вида:

$$\sum_{i=-1}^k a_i y_{n-i} = \tau f(t_{j+1}, y_{j+1}). \quad (8.104)$$

Именно эта конструкция в литературе и носит название *формул дифференцирования назад*. Смысл ее состоит в том, чтобы не использовать в алгоритме вычисление производных от решения (или правых частей дифференциальной задачи) в тех точках сетки, значение в которых уже известно, причем приближенно. В этом случае можно надеяться на достижение более сильных свойств устойчивости, чем, скажем, для рассматривавшихся нами ранее (тоже неявных!) *интерполяционных методов Адамса*.

Условия порядка (8.97) для (8.104) примут вид

$$\begin{cases} \sum_{i=-1}^k a_i = 0, \\ \sum_{i=-1}^k i^l a_i = (-1)^l, \quad l = 1, 2, \dots, p. \end{cases} \quad (8.105)$$

Построим конкретные примеры таких методов:

- При $p = 1$ имеем два уравнения. Поэтому достаточно положить $k = 0$. Тогда (8.105) примет вид

$$\begin{cases} a_{-1} + a_0 = 0, \\ -a_{-1} = -1, \end{cases}$$

откуда $a_{-1} = 1$, $a_0 = -1$. Соответствующий метод дифференцирования назад первого порядка будет иметь выглядеть следующим образом:

$$y_{j+1} - y_j = \tau f(t_{j+1}, y_{j+1}).$$

Легко узнат в нем изучавшийся нами ранее *неявный метод Эйлера* (который, как мы помним, является *А-устойчивым*).



2. Пусть теперь $p = 2$. В этом случае уравнений будет три и k достаточно положить равным 2. Тогда (8.105) примет вид

$$\begin{cases} a_{-1} + a_0 + a_1 = 0, \\ -a_{-1} + a_1 = -1, \\ a_{-1} + a_1 = 2 \end{cases}$$

и имеет решение $a_1 = \frac{1}{2}$, $a_{-1} = \frac{3}{2}$, $a_0 = -2$, что приводит к методу дифференцирования назад второго порядка

$$\frac{3}{2}y_{j+1} - 2y_j + \frac{1}{2}y_{j-1} = \tau f(t_{j+1}, y_{j+1}). \quad (8.106)$$

Реализация неявных методов

Непосредственно по виду всех рассмотренных нами выше неявных методов можно сделать вывод, что для того, чтобы превратить их в алгоритмы для вычисления приближенного решения задачи Коши в очередной точке сеточной области, необходимо «приложить» к каждому из них некоторый алгоритм решения соответствующего уравнения или системы. При этом неявные многошаговые методы представляют собой, как правило (мы других типов не рассматривали), одно уравнение, если исходная задача — это задача Коши для одного уравнения, и систему уравнений, если исходная задача — это задача Коши для системы уравнений. В то же время неявные одношаговые методы (типа Рунге–Кутта) практически всегда представляют собой системы уравнений. Способы решения таких задач мы рассматривали ранее. Формально любой из них может быть применен и в нашем случае. Однако не все получающиеся при этом алгоритмы будут одинаково успешными. Поясним ситуацию (а заодно и продемонстрируем, как выглядят соответствующие алгоритмы) на примере неявного метода Эйлера, применяемого к решению одного нелинейного уравнения. Как мы помним, формула его имеет вид

$$y_{j+1} = y_j + \tau f(t_{j+1}, y_{j+1}). \quad (8.107)$$

Метод итерации. Простейшим из методов, применяемых для решения нелинейных уравнений, является метод итераций. Технически в нашем случае его применять особенно удобно, ибо формула (8.107) фактиче-



ски дает канонический вид нелинейного уравнения, готовый к применению метода итераций. Следовательно, для нахождения приближенного значения y_{j+1} можно записать итерационный процесс

$$y_{j+1}^{k+1} = y_j + \tau f\left(t_{j+1}, \overset{k}{y}_{j+1}\right), \quad k = 0, 1, \dots \quad (8.108)$$

В качестве начального приближения $\overset{0}{y}_{j+1}$ в простейшем варианте может быть взято значение решения в предыдущем узле сетки, т.е. $\overset{0}{y}_{j+1} = y_j$. Иногда для предсказания начального приближения используют некоторый явный метод решения задачи Коши (например, в нашем случае — **явный метод Эйлера**, вполне согласованный с (8.107) по порядку, т.е. $\overset{0}{y}_{j+1} = y_j + \tau f(t_j, y_j)$).

Итерации продолжают до достижения сходимости (т.е., например, до выполнения неравенства $|y_{j+1}^{k+1} - \overset{k}{y}_{j+1}| \leq \varepsilon$, где ε — заданная величина погрешности, которая должна быть, очевидно, согласована с пользовательским требованием к точности нахождения решения исходной задачи Коши (по крайней мере, быть не больше последней)). В качестве критерия для остановки итерационного процесса можно использовать и соответствующий аналог относительной погрешности.

Вспомним, однако, что метод итерации имеет ограничения на область сходимости. В частности, достаточное условие сходимости может иметь вид $\left| \frac{\partial \varphi(y_{j+1})}{\partial y} \right| < 1$, где $\varphi(y)$ — функция, стоящая в правой части канонического представления. В нашем случае это ограничение примет вид $\tau \left| \frac{\partial f(t_{j+1}, y_{j+1})}{\partial y} \right| < 1$ и дает ограничение на допустимую величину шага сетки, аналогичную соответствующему ограничению, накладываемому явным методом Эйлера.

Таким образом, метод итераций уничтожает одно из главных достоинств неявных методов — их возможную **A-устойчивость**, — и поэтому, несмотря на простоту, не может быть рекомендован к широкому использованию (по крайней мере, при решении жестких задач).



Метод Ньютона. Переписав (8.107) в виде $g(y_{j+1}) = 0$, где $g(y_{j+1}) = y_{j+1} - y_j - \tau f(t_{j+1}, y_{j+1})$, можем записать для нахождения неизвестной величины y_{j+1} итерационный метод Ньютона:

$$y_{j+1}^{k+1} = y_j^k - \frac{y_{j+1}^k - y_j - \tau f(t_{j+1}, y_{j+1}^k)}{1 - \tau \frac{\partial f(t_{j+1}, y_{j+1}^k)}{\partial y}}, \quad k = 0, 1, \dots \quad (8.109)$$

Выбор начального приближения может быть осуществлен аналогично рассмотренному выше для метода итераций. При этом, однако, необходимо иметь в виду, что в случае жестких задач использование явных алгоритмов для предсказания начального приближения следует осуществлять с известной долей осторожности. Контроль сходимости также осуществляется аналогично.

Напомним, что сходимость метода Ньютона практически регламентируется только качеством выбора начального приближения и поэтому это — один из лучших способов реализации неявных численных методов.

Замечание 8.7. Ограничившись одной итерацией метода Ньютона (8.109) при выборе $y_{j+1}^0 = y_j$, получим простейший пример нелинейного явного численного метода решения задачи Коши:

$$y_{j+1} = y_j + \frac{\tau f(t_{j+1}, y_j)}{1 - \tau \frac{\partial f(t_{j+1}, y_j)}{\partial y}}. \quad (8.110)$$

Несложно проверить, что в этом случае получается явный **A-устойчивый** метод.



Меню

Глава 9

Методы решения граничных задач для обыкновенных дифференциальных уравнений

- 9.1. Введение
- 9.2. Методы, основанные на сведении к решению задач Коши
- 9.3. Проекционные методы решения граничных задач



9.1. Введение

Граничные или многоточечные задачи представляют собой более сложный тип задач по сравнению с задачами Коши. Здесь и сам характер постановки задач более общий. Задаются, как правило, не значения искомой функции или ее производных, а лишь связи между этими значениями. При решении подобных задач более сложен как сам процесс поиска решения, так и вопросы существования и единственности решения.

Будем считать, что на отрезке $[a, b]$ задано обыкновенное дифференциальное уравнение n -го порядка (можно рассматривать систему из n уравнений)

$$u^{(n)} = f(x, u, u', \dots, u^{(n-1)}),$$

выбраны k различных точек $x_1 < x_2 < \dots < x_k$ и заданы некоторые n связей (условий)

$$v_j \left[u(x_1), u'(x_1), \dots, u^{(n-1)}(x_1), \dots, u(x_k), u'(x_k), \dots, u^{(n-1)}(x_k) \right] = 0, \quad j = 1, 2, \dots, n.$$

Тогда говорят, что для нашего уравнения поставлена *многоточечная (k -точечная) задача*. В частном случае, когда $k = 1$, а функции v_j имеют тривиальный вид, мы получаем задачу Коши.

Сейчас (и, как правило, далее) рассмотрим случай $k = 2$ и $x_1 = a$, $x_2 = b$. В этом случае задача называется *граничной* (или краевой).

Если исходное дифференциальное уравнение или хотя бы одно из дополнительных условий нелинейны, то мы имеем нелинейную граничную задачу. В противном случае задача называется линейной. В достаточно общем виде линейную граничную задачу можно записать следующим образом:

$$Lu \equiv u^{(n)}(x) + p_1(x)u^{(n-1)}(x) + \dots + p_{n-1}(x)u'(x) + p_n(x)u(x) = f(x) \quad (9.1)$$

$$l_j(u) \equiv \sum_{i=0}^{n-1} \left[\alpha_{ij}u^{(i)}(a) + \beta_{ij}u^{(i)}(b) \right] = A_j, \quad j = 1, 2, \dots, n. \quad (9.2)$$



9.2. Методы, основанные на сведении к решению задач Коши

- 9.2.1. Метод редукции граничных задач к задачам Коши
- 9.2.2. Метод стрельбы для линейных граничных задач
- 9.2.3. Метод дифференциальной прогонки
- 9.2.4. Метод стрельбы для нелинейных граничных задач



9.2.1. Метод редукции граничных задач к задачам Коши

Будем иметь в виду линейную граничную задачу (9.1), (9.2). Сейчас нас будет интересовать процедура сведения решения такой задачи к решению задач с начальными условиями. Предлагаемый ниже алгоритм существенно использует общий вид решения дифференциального уравнения (9.1).

Как известно из теории дифференциальных уравнений, общее решение линейного уравнения (9.1) может быть записано в виде

$$u(x) = u_0(x) + \sum_{i=1}^n C_i u_i(x), \quad (9.3)$$

где $u_0(x)$ — некоторое частное решение неоднородного уравнения (9.1), т.е.

$$L u_0(x) = f(x),$$

а $u_i(x)$, $i = \overline{1, n}$ — решения соответствующего однородного уравнения, образующие линейно независимую систему:

$$L u_i(x) = 0, \quad i = \overline{1, n},$$

C_i — произвольные постоянные, конкретные значения которых определяются дополнительными условиями (9.2).

Если бы нам были известны функции $u_0(x)$ и $u_i(x)$, $i = \overline{1, n}$, то те значения произвольных постоянных C_i , которые соответствуют искомому решению, были бы легко найдены из условий (9.2), а именно:

$$l_j(u_0) + \sum_{i=1}^n C_i \cdot l_j(u_i) = A_j, \quad j = 1, 2, \dots, n, \quad (9.4)$$

и, таким образом, имеем для определения названных констант систему из n линейных алгебраических уравнений с n неизвестными.

Если определитель матрицы системы (9.4) отличен от нуля, то она будет иметь единственное решение и, следовательно, мы получим единственный набор C_i . В противном же случае будем иметь либо бесконечное множество решений, либо неразрешимость в зависимости от ранга расширенной матрицы граничных условий.



Весь вопрос решения граничной задачи, таким образом, сводится к тому, как найти частное решение неоднородного уравнения $u_0(x)$ и систему $u_i(x)$. Поскольку речь идет о численном решении соответствующей задачи и мы на данный момент для дифференциальных уравнений знакомы лишь с методами решения задач Коши, то естественно в качестве искомых функций взять решения некоторых задач Коши.

Учитывая, что $u_0(x)$ — произвольное частное решение неоднородного уравнения, задачу Коши для его определения можно взять, например, в виде (использовав простейшие начальные условия)

$$\begin{cases} Lu_0(x) = f(x), \\ u_0(a) = 0, \\ \dots \\ u_0^{(n-1)}(a) = 0. \end{cases} \quad (9.5)$$

Аналогичным образом, поскольку единственным ограничением для системы $u_i(x)$ является ее линейная независимость, то соответствующий набор начальных условий для их определения должен быть подчинен условию: определитель Вронского искомой системы в точке a должен быть отличен от нуля. Поэтому простейший тип таких начальных условий приводит к задачам вида

$$\begin{cases} Lu_i(x) = 0, \\ u_i^{(j)}(a) = \delta_i^j, \quad j = \overline{0, n-1}; \quad i = \overline{1, n}. \end{cases} \quad (9.6)$$

В этом случае определитель Вронского есть определитель единичной матрицы, что обеспечивает выполнение сформулированных выше требований.

Таким образом, мы свели решение граничной задачи к решению $(n+1)$ задач Коши (9.5), (9.6) и системы линейных алгебраических уравнений (9.4) для определения постоянных интегрирования C_i .

Очевидно, что граничные условия могут быть и нелинейными, что, в конечном итоге, приведет к необходимости решать вместо линейной системы (9.4) соответствующую систему нелинейных уравнений.

Замечание 9.1. Все указанные выше задачи Коши могут быть решены численно с использованием изучавшихся нами ранее методов. При этом сетки, на которых отыскиваются решения задач (9.5), (9.6) должны иметь непустое пересечение узловых точек (в оптимуме — совпадать), ибо только на этом пересечении и может быть построена соответствующая линейная комбинация (9.3). Это накладывает некоторые ограничения на процедуры поиска решений задач Коши с апостериорной оценкой погрешности.



Замечание 9.2. Описанный выше алгоритм может быть применен к поиску решения не только задачи (9.1), (9.2), но и любой другой, общее решение которой имеет вид (9.3) (например, граничной задачи для системы линейных обыкновенных дифференциальных уравнений).



9.2.2. Метод стрельбы для линейных граничных задач

[Уравнение второго порядка случай](#)

[Общий линейный случай](#)

Заметим, что в изложенном [выше](#) алгоритме метода редукции фигурируют слова «произвольный» по отношению к построению частного решения неоднородного уравнения и линейно независимой системы решений однородного уравнения. Если отказаться от этой произвольности в пользу того, чтобы некоторые комбинации найденных частных решений удовлетворяли части граничных условий, то это может позволить сократить общее количество решаемых задач Коши.

[Уравнение второго порядка случай](#)

Изучим вначале применение этой идеи на частном примере задачи

$$\begin{cases} Lu(x) \equiv u''(x) + p(x)u'(x) + q(x)u(x) = f(x), & a \leq x \leq b, \\ u(a) = A, \quad u(b) = B. \end{cases} \quad (9.7)$$

Учитывая простоту граничных условий, можно предложить следующее: рассмотрим частное решение $u_0(x)$ неоднородного уравнения, которое удовлетворяет левому граничному условию. Значение $u'_0(a)$ при этом может задаваться произвольно. Таким образом, функция $u_0(x)$ находится как решение задачи

$$\begin{cases} Lu_0(x) = f(x), \\ u_0(a) = A, \\ u'_0(a) = \eta_0, \quad \eta_0 — \text{любое}. \end{cases} \quad (9.8)$$

Рассмотрим также частное решение однородного уравнения $u_1(x)$, удовлетворяющее условиям

$$\begin{cases} Lu_1(x) = 0, \\ u_1(a) = 0, \\ u'_1(a) = \eta_1, \quad \eta_1 \neq 0 — \text{любое}. \end{cases} \quad (9.9)$$



Тогда составленная из функций $u_0(x)$ и $u_1(x)$ линейная комбинация

$$u(x) = u_0(x) + C u_1(x) \quad (9.10)$$

заведомо удовлетворяет дифференциальному уравнению (9.7) и левому граничному условию при любом значении произвольной постоянной C . Выбором же последней следует распорядиться таким образом, чтобы удовлетворить второму граничному условию. Исходя из этого, получится уравнение

$$u_0(b) + C u_1(b) = B,$$

которое в случае $u_1(b) \neq 0$ имеет решение

$$C = \frac{B - u_0(b)}{u_1(b)}$$

и, таким образом, решение задачи (9.7) может быть вычислено по формуле (9.10). По сравнению с классическим вариантом [метода редукции](#) здесь необходимо решать всего две задачи Коши.

Общий линейный случай

Перейдем теперь к рассмотрению общего случая краевой задачи для системы линейных обыкновенных дифференциальных уравнений первого порядка. Итак, пусть имеем краевую задачу

$$\begin{cases} u'(x) = A(x)u(x) + f(x), & a < x < b \\ Bu(a) = c, \\ Du(b) = d, \end{cases} \quad (9.11)$$

где u, f, c, d — векторы размерностей соответственно $n, n, n-r, r$, а A, B, D — матрицы размерностей $n \times n, (n-r) \times n, r \times n$. В дальнейшем будем предполагать, что ранг матрицы B равен $n-r$, а ранг матрицы D равен r . Рассмотрим линейную систему алгебраических уравнений

$$Bu = c \quad (9.12)$$



задающую граничное условие на левом конце отрезка интегрирования. Так как по предположению ранг матрицы B равен $n - r$, то общее решение системы (9.12) может быть записано в виде

$$u_0 + \sum_{i=1}^r C_i u_i,$$

где u_0 — произвольное решение неоднородной системы (9.12), а u_1, u_2, \dots, u_r — произвольная система из r линейно независимых решений однородной системы $Bu = 0$. Пусть $u_1, u_2, \dots, u_r, u_0$ — какой-либо набор таких векторов. Тогда при помощи численного интегрирования найдем частное решение неоднородной системы

$$\begin{cases} u'_0(x) = A(x)u_0(x) + f(x), \\ u_0(a) = u_0 \end{cases} \quad (9.13)$$

и решения однородных систем

$$\begin{cases} u'_j(x) = A(x)u_j(x), \\ u_j(a) = u_j, \quad j = \overline{1, r}. \end{cases} \quad (9.14)$$

Теперь заметим, что всякая функция вида

$$u(x) = u_0(x) + \sum_{j=1}^r C_j u_j(x) \quad (9.15)$$

при любых значениях произвольных постоянных C_j удовлетворяет исходной системе (9.11) и левому граничному условию. Таким образом, остается определить эти постоянные, удовлетворив правому граничному условию:

$$D \left(u_0(b) + \sum_{j=1}^r C_j u_j(b) \right) = d. \quad (9.16)$$

Выражение (9.16) представляет собой систему из r линейных алгебраических уравнений с r неизвестными. Матрица D этой системы невырождена (в соответствии с предположением ее ранг равен r), поэтому задача имеет единственное решение.

Окончательно алгоритм решения граничной задачи (9.11) может выглядеть следующим образом:



- 1) Находим частное решение линейной неоднородной системы алгебраических уравнений (9.12), соответствующей левому граничному условию, и r линейно независимых частных решений соответствующей (9.12) линейной однородной системы;
- 2) Решаем задачи Коши (9.13), (9.14) (в количестве $(r + 1)$ штук), начальными условиями которых являются найденные на первом этапе частные решения системы (9.12) и соответствующей ей однородной;
- 3) Решаем систему линейных алгебраических уравнений (9.16) относительно произвольных постоянных C_j ;
- 4) По формуле (9.15) определяем решение исходной граничной задачи.

9.2.3. Метод дифференциальной прогонки

Другим алгоритмом, применяемым для решения линейных граничных задач, является метод дифференциальной прогонки. Технику применения одной из простейших разновидностей этого алгоритма рассмотрим на примере задачи

$$\begin{cases} Lu(x) \equiv u''(x) + p(x)u'(x) + q(x)u(x) = f(x), & a \leq x \leq b, \\ \alpha_0 u(a) + \alpha_1 u'(a) = A, \\ \beta_0 u(b) + \beta_1 u'(b) = B. \end{cases} \quad (9.17)$$

Существенным моментом здесь (как, впрочем, и в задачах, решавшихся в предыдущем параграфе) является то, что граничные условия разделены по концам.

Как мы уже отмечали ранее, общее решение уравнения (9.17) имеет вид

$$u(x) = u_0(x) + C_1 u_1(x) + C_2 u_2(x).$$

Выделим теперь то подмножество решений, которое удовлетворяет одному из граничных условий (для определенности — левому). В итоге получим однопараметрическое семейство решений, которое можно рассматривать как общее решение некоторого линейного дифференциального уравнения первого порядка

$$u'(x) + P(x)u(x) = F(x) \quad (9.18)$$

Мы не знаем коэффициентов этого уравнения, но и $P(x)$, и $F(x)$ можно найти. Для этого следует учесть, что любое решение уравнения (9.18) есть решение исходного уравнения второго порядка (9.18) и, кроме того, удовлетворяет левому граничному условию из (9.17). Удовлетворим сначала первому требованию. Так как из (9.18) следует, что

$$u'(x) = F(x) - P(x)u(x),$$

то

$$u''(x) = F'(x) - P'(x)u(x) - P(x)u'(x) = F'(x) - P'(x)u(x) - P(x)[F(x) - P(x)u(x)].$$

Подставив эти выражения в исходное уравнение (9.17), будем иметь:

$$F'(x) - P'(x)u(x) - P(x)[F(x) - P(x)u(x)] + p(x)[F(x) - P(x)u(x)] + q(x)u(x) = f(x).$$

Поскольку мы хотим выполнения этого равенства при любых $u(x)$, то, приравнивая коэффициенты при $u(x)$ в левой и правой части последнего равенства, а также — отдельно — свободные члены, получим:

$$\begin{cases} P'(x) + [p(x) - P(x)]P(x) = q(x), \\ F'(x) + [p(x) - P(x)]F(x) = f(x). \end{cases} \quad (9.19)$$

Таким образом, для определения коэффициентов уравнения (9.18) получим систему обыкновенных дифференциальных уравнений первого порядка (9.19).

Чтобы выделить конкретные $P(x)$ и $F(x)$, удовлетворим левому граничному условию из (9.17): для любой функции $u(x)$ должно выполняться равенство

$$\alpha_0 u(a) + \alpha_1 [F(a) - P(a)u(a)] = A.$$

Как и выше, приравнивая коэффициенты при $u(a)$ и свободные члены, найдем:

$$\begin{cases} P(a) = \frac{\alpha_0}{\alpha_1}, \\ F(a) = \frac{A}{\alpha_1}. \end{cases} \quad (9.20)$$

Отсюда следует, что при $\alpha_1 \neq 0$ (это — одно из условий применимости рассматриваемого варианта метода прогонки) мы для определения функций $P(x)$ и $F(x)$ получаем задачу Коши (9.19), (9.20). Решив данную задачу, мы построим уравнение (9.18). Теперь можно будет потребовать, чтобы его решение удовлетворяло второму из граничных условий задачи (9.17). На основании этого требования получим систему линейных алгебраических уравнений

$$\begin{cases} \beta_0 u(b) + \beta_1 u'(b) = B, \\ P(b)u(b) + u'(b) = F(b). \end{cases} \quad (9.21)$$



Если $\Delta = \begin{vmatrix} \beta_0 & \beta_1 \\ P(b) & 1 \end{vmatrix} \neq 0$, то мы отсюда единственным образом найдем значения $u(b)$ и $u'(b)$:

$$\left\{ \begin{array}{l} u(b) = \frac{\begin{vmatrix} B & \beta_1 \\ F(b) & 1 \end{vmatrix}}{\Delta}, \\ u'(b) = \frac{\begin{vmatrix} \beta_0 & B \\ P(b) & F(b) \end{vmatrix}}{\Delta}. \end{array} \right. \quad (9.22)$$

После этого, решая уравнение (9.18) с $u(b)$ из (9.22) в качестве начального (точнее, конечного) условия, найдем решение задачи (9.17) (заметим, что последняя задача Коши решается в противоположном направлении (от правого конца отрезка к левому), что предполагает использование в формулах численных методов, применяемых для этих целей, отрицательного значения шага).

Замечание 9.3. Описанный вариант метода дифференциальной прогонки носит название метода *левой прогонки*, поскольку мы строили уравнение (9.18) удовлетворяющим левому граничному условию. Если же $\alpha_1 = 0$, то нам не удастся воспользоваться формулами (9.20) для вычисления начальных условий для функций $P(x)$ и $F(x)$. В этом случае можно (если $\beta_1 \neq 0$) построить формулы типа (9.20), получить их из соображений удовлетворения правому граничному условию. Таким образом, внося необходимые корректизы в дальнейшие рассуждения, придем к методу *правой прогонки*. Если же $\alpha_1 = \beta_1 = 0$, то метод дифференциальной прогонки для данной задачи не применим.

Замечание 9.4. На аналогичной идеологии могут быть построены варианты метода дифференциальной прогонки и для решения линейных граничных задач более общего вида (например, типа (9.11) из предыдущего параграфа).



9.2.4. Метод стрельбы для нелинейных граничных задач

Вначале рассмотрим нелинейную граничную задачу для системы обыкновенных дифференциальных уравнений первого порядка:

$$\begin{cases} u'(x) = f(x, u(x)), \\ B(u(a)) = 0, \\ D(u(b)) = 0. \end{cases} \quad (9.23)$$

Здесь $u = (u_1, u_2, \dots, u_n)^T$, $B = (b_1, b_2, \dots, b_{n-r})^T$, $D = (d_1, d_2, \dots, d_r)^T$, причем $u(x)$ — искомая вектор-функция, а B и D — заданные нелинейные вектор-функции указанного количества аргументов. Основой для создания алгоритма решения задачи (9.23) путем сведения к решению задачи Коши может служить очень простой факт: решение задачи зависит от начальных данных. Поэтому, если добавить, например, недостающие на левом конце отрезка интегрирования r уравнений связи вида $g_i(u(a)) = \eta_i$, $i = \overline{1, r}$, (здесь $g_i(u(a))$ — заданные функции векторного аргумента, а η_i — заданные числовые параметры) таким образом, чтобы система нелинейных уравнений

$$\begin{cases} b_i(u(a)) = 0, \quad i = \overline{1, n-r}, \\ g_j(u(a)) = \eta_j, \quad j = \overline{1, r} \end{cases} \quad (9.24)$$

позволяла однозначно определить вектор начальных данных $u(a)$ как функцию параметров η : $u(a) = \omega(\eta)$, $\eta = (\eta_1, \dots, \eta_r)$, то решение системы уравнений (9.23) с найденными начальными условиями также будет функцией вектора параметров η : $u(x) = u(x, \eta)$. Эта функция, очевидно, удовлетворяет левым граничным условиям. Поэтому вектор параметров η может быть определен путем подстановки в правое граничное условие:

$$\psi(\eta) \stackrel{\text{def}}{=} D(u(b, \eta)) = 0 \quad (9.25)$$

Таким образом, формально задача свелась к решению нелинейной системы уравнений относительно вектора параметров.

Основная проблема, возникающая при решении системы (9.25), состоит в том, что мы не знаем аналитического вида функциональной зависимости $\psi(\eta)$, но в то же время имеем техническую возможность



при фиксированном значении вектора параметров η вычислять значение функции ψ . Для этого необходимо решить систему (9.24) (этим самым мы задаем начальные условия для исходной системы (9.23)), затем с найденными начальными условиями решить (численно) задачу Коши и найденное решение подставить в правое граничное условие. Это и будет искомое значение функции ψ . Учитывая сказанное, в качестве метода для решения системы (9.25) следует выбирать, с одной стороны, достаточно быстро сходящийся, а с другой — тот, алгоритм которого «в состоянии» обходиться без вычисления производных от функции ψ (последняя задача не относится к разряду неразрешимых, но является очень трудоемкой).

Опишем алгоритм (метода стрельбы) более подробно на примере системы вида (9.23) при $n = 2$, $r = 1$, т.е.

$$\begin{cases} u'(x) = f(x, u(x), v(x)), \\ v'(x) = g(x, u(x), v(x)), \\ \varphi(u(a), v(a)) = 0, \\ \psi(u(b), v(b)) = 0 \end{cases} \quad (9.26)$$

В соответствии с изложенной выше общей схемой выберем произвольно значение $u(a) = \eta$, рассмотрим левое граничное условие как алгебраическое уравнение

$$\varphi(\eta, v(a)) = 0$$

и определим удовлетворяющее ему значение $v(a) = \xi(\eta)$. Возьмем значения $u(a) = \eta$ и $v(a) = \xi(\eta)$ в качестве начальных условий задачи Коши для системы (9.26) и проинтегрируем полученную задачу любым подходящим численным методом. При этом получим решение $u(x; \eta)$, $v(x; \eta)$, зависящее от η как от параметра.

Значение $\xi(\eta)$ выбрано так, что найденное решение задачи Коши удовлетворяет левому граничному условию задачи (9.26). Однако второму граничному условию это решение, вообще говоря, не удовлетворяет: при его подстановке левая часть правого граничного условия, рассматриваемая как функция параметра η

$$\bar{\psi}(\eta) = \psi(u(b; \eta), v(b; \eta)) \quad (9.27)$$

не обратится в нуль. Таким образом, необходимо каким-либо образом менять числовые значения параметра η , пока не подберем такое значение, для которого $\bar{\psi}(\eta) \approx 0$ с требуемой точностью, т.е., как и было



отмечено в общей схеме, решение граничной задачи (9.26) в конечном итоге сводится к нахождению корня алгебраического уравнения

$$\bar{\psi}(\eta) = 0 \quad (9.28)$$

Простейшим из методов его решения, которые целесообразно применять в данном случае, является [метод дихотомии](#). При его реализации делают пробные «выстрелы» — расчеты с наугад (если нет каких-либо специальных соображений) выбранными значениями η_i до тех пор, пока среди величин $\bar{\psi}(\eta_i)$ не окажутся разные по знаку. Пара таких значений η_i, η_{i+1} образует «вилку». С математической точки зрения мы таким образом отделим некоторый корень уравнения (9.28)). Далее, последовательно деляя отрезок $[\eta_i; \eta_{i+1}]$ пополам до получения нужной точности, производим «пристрелку» (уточнение) параметра η . Благодаря этому процессу весь метод получил название метода стрельбы.

Однако нахождение каждого нового значения функции $\bar{\psi}(\eta)$ требует численного интегрирования системы (9.26), т.е. достаточно трудоемко. Поэтому корень уравнения (9.28) желательно находить с помощью метода, обладающего более быстрой сходимостью, нежели дихотомия. Часто таковым является разностный аналог метода Ньютона — метод секущих. В этом случае первые два расчета делают с наудачу выбранными значениями η_0, η_1 , а следующие вычисляют по формуле

$$\eta_{i+1} = \eta_i - \frac{(\eta_i - \eta_{i-1}) \bar{\psi}(\eta_i)}{\bar{\psi}(\eta_i) - \bar{\psi}(\eta_{i-1})}, \quad i = 1, 2, \dots \quad (9.29)$$

Сходимость итерационного процесса (9.29), очевидно, зависит от выбора η_0 и η_1 .

Замечание 9.5. Очевидно, при решении общей задачи (9.23) приведенные здесь соображения становятся значительно сложнее с точки зрения технической их реализации.



Меню



Вверх

Назад

Вперёд

Пред.

След.

Указатель

Помощь

Экран

9.3. Проекционные методы решения граничных задач

9.3.1. Введение

9.3.2. Вариационные методы решения граничных задач

9.3.3. Метод моментов и метод Галеркина решения граничных задач

9.3.4. Метод наименьших квадратов решения граничных задач

9.3.5. Метод коллокации решения граничных задач



9.3.1. Введение

С идеологией данной группы методов решения операторных уравнений мы познакомились при изучении [методов решения интегральных уравнений](#). Напомним, что эти методы называют проекционными по той причине, что первоначальное пространство, в котором поставлена исходная задача, проектируют в некоторое подпространство более простой структуры, где и отыскивают приближение. Часто элементами подпространства являются известные координатные функции, и остается только подобрать коэффициенты линейной комбинации. Приближенное решение здесь получается в аналитическом виде.



Меню

9.3.2. Вариационные методы решения граничных задач

Вариационная задача для операторных уравнений

Граничная задача для обыкновенного дифференциального уравнения второго порядка как операторное уравнение

Метод Ритца нахождения минимума функционала

Сходимость минимизирующей последовательности к минимизирующей функции

Построение минимизирующей последовательности по Ритцу

Вариационная задача для операторных уравнений

Пусть H — некоторое вещественное гильбертово пространство и A — линейный оператор, определенный на множестве H_A , всюду плотном в H . Рассмотрим операторное уравнение

$$Au = f, \quad (9.30)$$

где $f \in H$ — заданный элемент, а $u \in H_A$ — искомый.

Оператор A будем предполагать положительным и самосопряженным. Тогда уравнение (9.30) не может иметь более одного решения.

Действительно, пусть имеются два решения $u_1 \neq u_2$. Вычитая равенства $Au_1 = f$ и $Au_2 = f$, получим: $A(u_1 - u_2) = 0$ и, следовательно, $(A(u_1 - u_2), u_1 - u_2) = 0$. Поскольку по условию $A > 0$, то отсюда с необходимостью следует, что $u_1 - u_2 = 0$, т.е. $u_1 = u_2$, что противоречит предположению.

Таким образом, если задача (9.30) при сделанных относительно оператора A предположениях имеет решение, то это решение будет единственным.

Теорема 9.1. Если уравнение (9.30) имеет некоторое решение u_1 , то это решение доставляет минимум функционала

$$J(u) = (Au, u) - 2(f, u) \quad (9.31)$$

[[Доказательство](#)]



Меню

Часть III. Теоретические материалы

Глава 9. Методы решения граничных задач для ОДУ

9.3. Проекционные методы решения граничных задач

9.3.2. Вариационные методы решения граничных задач

Справедливо и обратное утверждение.

Теорема 9.2. Если найдется такой элемент $u_1 \in H_A$, который доставляет минимум функционалу $J(u)$, то этот элемент будет решением уравнения (9.30). [\[Доказательство\]](#)

Граничная задача для обыкновенного дифференциального уравнения второго порядка как операторное уравнение

Вернемся сейчас к граничной задаче для обыкновенного дифференциального уравнения второго порядка (с краевыми условиями первого либо третьего рода) и выясним, какой вид она должна иметь для того, чтобы непосредственно из общей теории следовала ее равносильность некоторой вариационной задаче типа (9.31).

Как уже отмечалось в общей теории, оператор A должен быть самосопряженным и положительным. Пусть исходное дифференциальное уравнение имеет вид

$$Lu(x) \equiv p_0(x) u''(x) + p_1(x) u'(x) + p_2(x) u(x) = f(x), \quad x \in [a, b].$$

Тогда условие самосопряженности оператора L имеет вид

$$(Lu, v) = (u, Lv),$$

или

$$\int_a^b [(p_0 u'' + p_1 u' + p_2 u) v - (p_0 v'' + p_1 v' + p_2 v) u] dx \equiv 0,$$

причем это соотношение выполняется для любых допустимых функций u и v .

Выполняя в слагаемых, содержащих вторые производные, интегрирование по частям, имеем:

$$(Lu, v) - (u, Lv) = p_0 (u'v - uv') \left| \begin{array}{l} b \\ a \end{array} \right. + \int_a^b \left[-u' (p_0 v)' + v' (p_0 u)' + p_1 (u'v - uv') \right] dx =$$

$$= p_0 (u'v - uv') \left| \begin{array}{l} b \\ a \end{array} \right. + \int_a^b (p_0' - p_1) (uv' - u'v) dx = 0.$$



Отсюда, учитывая, что $p_0(x) \neq 0$, а также то, что $u(x)$ и $v(x)$ — произвольные допустимые функции, получим:

$$(u'v - uv')(b) = (u'v - uv')(a) = 0 \quad (9.32)$$

$$p'_0(x) - p_1(x) = 0 \quad (9.33)$$

Рассмотрим теперь по отдельности случаи граничных условий первого и третьего рода (условия второго рода — частный случай условий третьего рода).

Итак, пусть вначале граничные условия имеют вид

$$\begin{cases} u(a) = A, \\ u(b) = B. \end{cases} \quad (*)$$

Тогда из условия (9.32) следует, что допустимые функции должны обращаться на границе в нуль, т.е. $A = B = 0$. Кроме того, условие (9.33) означает, что

$$p_0(x)u''(x) + p_1(x)u'(x) = (p_0(x)u'(x))'.$$

Таким образом, первая краевая задача, исходя из условия самосопряженности дифференциального оператора, будет иметь вид

$$\begin{cases} Lu(x) \equiv (p_0(x)u'(x))' + p_2(x)u(x) = f(x), & a \leq x \leq b, \\ u(a) = u(b) = 0. \end{cases}$$

Выясним теперь, какие условия на коэффициенты $p_0(x)$ и $p_2(x)$ накладывает требование положительности оператора L . Поскольку для всех отличных от тождественного нуля функций $u(x)$ должно выполняться неравенство $(Lu, u) > 0$, то имеем:

$$\begin{aligned} (Lu, u) &= \int_a^b \left[(p_0u')' + p_2u \right] u dx = p_0u' \cdot u \Big|_a^b + \int_a^b \left[-p_0 \cdot (u')^2 + p_2u^2 \right] dx = \\ &= \int_a^b \left[-p_0(x) \cdot (u'(x))^2 + p_2(x)u(x)^2 \right] dx > 0. \end{aligned}$$



Отсюда следует, что (поскольку $u(x)$ — произвольная допустимая функция)

$$\begin{cases} p_0(x) \leq c_0 < 0, \\ p_2(x) \geq 0, \end{cases} \quad \text{для всех } x \in [a, b].$$

Учитывая полученные результаты, в дальнейшем первую краевую задачу с самосопряженным положительным оператором будем записывать в виде

$$\begin{cases} Lu(x) \equiv -(p(x) u'(x))' + q(x) u(x) = -f(x), \quad a \leq x \leq b, \\ u(a) = u(b) = 0. \end{cases} \quad (9.34)$$

где $p(x) \geq p_0 > 0$, $p(x) \in C^1[a, b]$ и $q(x) \in C[a, b]$, $q(x) \geq 0$, $f(x) \in C[a, b]$.

Заметим, что в этом случае функционал (9.31) примет вид

$$\begin{aligned} J(u) &= (Lu, u) - 2(f, u) = \int_a^b \left[-(p(x) u'(x))' + q(x) u(x) + 2f(x) \right] u(x) dx = \\ &= \int_a^b \left[p(x) (u'(x))^2 + q(x) u^2(x) + 2f(x) u(x) \right] dx, \end{aligned} \quad (9.35)$$

т.е. задача (9.34) равносильна задаче минимизации функционала (9.35).

Случай неоднородных граничных условий (*) может быть сведен к рассмотренному. Действительно, зафиксируем некоторую функцию $v(x)$, удовлетворяющую условиям (*), и представим решение исходной задачи с неоднородными условиями в виде $u(x) = u_1(x) + v(x)$. Тогда функция $u_1(x)$ удовлетворяет нулевым граничным условиям и уравнению

$$Lu_1(x) = Lu(x) - Lv(x) = -f(x) - Lv(x) = -f(x) + (p(x) v'(x))' - q(x) v(x).$$



Меню

Часть III. Теоретические материалы

Глава 9. Методы решения граничных задач для ОДУ

9.3. Проекционные методы решения граничных задач

9.3.2. Вариационные методы решения граничных задач

Решение этой задачи равносильно минимизации функционала (9.35)

$$\begin{aligned}
 J(u_1) &= \int_a^b \left[p(x)(u'_1(x))^2 + q(x)u_1^2(x) + 2\left(f(x) - (p(x)v'(x))' + q(x)v(x)\right)u_1(x) \right] dx = \\
 &= \int_a^b \left[p(u'_1)^2 + qu_1^2 + 2fu_1 - 2(pv')' u_1 + 2qvu_1 \right] dx = \left[\begin{array}{l} \text{интегрируем слагаемое} \\ (pv')' u_1 \text{ по частям} \end{array} \right] = \\
 &= -2pv'u_1 \left| \frac{b}{a} \right. + \int_a^b \left[p(u'_1)^2 + qu_1^2 + 2fu_1 + 2pv'u'_1 + 2qvu_1 \right] dx = \\
 &= \int_a^b \left[p(u'_1 + v')^2 + q(u_1 + v)^2 + 2f(u_1 + v) \right] dx - \int_a^b \left[p(v')^2 + qv^2 + 2fv \right] dx = J(u) - J(v).
 \end{aligned}$$

Так как $v(x)$ — фиксированная функция, то минимизация функционала $J(u_1)$ равносильна минимизации функционала $J(u)$

Рассмотрим теперь граничные условия третьего рода

$$\begin{cases} \alpha_0 u(a) + \alpha_1 u'(a) = A, \\ \beta_0 u(b) + \beta_1 u'(b) = B. \end{cases} \quad (9.36)$$

и выясним, когда выполняется условие (9.32). Так как из (9.36) следует, что ($\alpha_1 \neq 0$, $\beta_1 \neq 0$)

$$u'(a) = \frac{A - \alpha_0 u(a)}{\alpha_1}, \quad u'(b) = \frac{B - \beta_0 u(b)}{\beta_1} \quad (9.37)$$

(заметим, что точно таким же граничным условиям удовлетворяет и функция $v(x)$), то (9.32) преобразуется к виду

$$\begin{aligned}
 \frac{B - \beta_0 u(b)}{\beta_1} v(b) - u(b) \frac{B - \beta_0 v(b)}{\beta_1} &= \frac{A - \alpha_0 u(a)}{\alpha_1} v(a) - u(a) \frac{A - \alpha_0 v(a)}{\alpha_1} \Rightarrow \\
 \Rightarrow \frac{B}{\beta_1} (v(b) - u(b)) &= \frac{A}{\alpha_1} (v(a) - u(a)) \equiv 0,
 \end{aligned}$$



откуда, учитывая произвольность допустимых функций, непосредственно получаем: $A = B = 0$, т.е. граничные условия (9.36) должны быть однородного типа (с нулевыми правыми частями). При этом (9.37) перепишутся в виде

$$u'(a) = -\frac{\alpha_0}{\alpha_1}u(a), \quad u'(b) = -\frac{\beta_0}{\beta_1}u(b) \quad (9.38)$$

Рассматривая дифференциальный оператор в виде (9.34), выясним условия положительной определенности (помимо $p(x) \geq p_0 > 0$ и $q(x) \geq 0$):

$$(Lu, u) = \int_a^b \left[-(p(x)u'(x))' + q(x)u(x) \right] u(x) dx = -p(x)u'(x)u(x) \Big|_a^b + \int_a^b [p(u')^2 + qu^2] dx.$$

Второе слагаемое при отмеченных условиях положительно, а первое с учетом (9.38) примет вид

$$p(b) \cdot \frac{\beta_0}{\beta_1} \cdot u^2(b) - p(a) \cdot \frac{\alpha_0}{\alpha_1} \cdot u^2(a).$$

Отсюда непосредственно следует:

$$\frac{\beta_0}{\beta_1} \geq 0, \quad \frac{\alpha_0}{\alpha_1} \leq 0.$$

Таким образом, при выполнении найденных условий задача

$$\begin{cases} Lu(x) \equiv -(p(x)u'(x))' + q(x)u(x) = -f(x), \quad a \leq x \leq b, \\ \alpha_0u(a) + \alpha_1u'(a) = 0, \\ \beta_0u(b) + \beta_1u'(b) = 0 \end{cases} \quad (9.39)$$

равносильна задаче минимизации функционала $J(u)$, который теперь примет вид

$$\begin{aligned} J(u) = (Lu, u) - 2(f, u) = p(b)\frac{\beta_0}{\beta_1}u^2(b) - p(a)\frac{\alpha_0}{\alpha_1}u^2(a) + \\ + \int_a^b [p(x)(u'(x))^2 + q(x)u^2(x) + 2f(x)u(x)] dx. \end{aligned} \quad (9.40)$$



Метод Ритца нахождения минимума функционала

После того как для заданной граничной задачи построен функционал $J(u)$, минимизация которого эквивалентна отысканию решения исходной задачи, встает вопрос о том, каким образом элемент, доставляющий минимум, может быть найден.

Такая задача (минимизации функционала) имеет смысл и самостоятельно, без привязки к дифференциальным уравнениям. Поэтому основные понятия и идеи рассмотрим на примере задачи минимизации функционала несколько более общего вида

$$J(u) = \int_a^b F(x, u, u') dx \quad (9.41)$$

на множестве функций $u(x) \in C^1[a, b]$, удовлетворяющих граничным условиям (*).

Предположим, что множество значений функционала $J(u)$ ограничено снизу и существует такая допустимая функция $u^*(x)$, что

$$J(u^*) = \min_u J(u) = m.$$

Тогда, если существует последовательность допустимых функций $u_n(x)$, $n = 0, 1, \dots$, для которой соответствующая ей последовательность функционалов $J(u_n)$ сходится к минимуму m , т.е.

$$m_n = J(u_n) \xrightarrow{n \rightarrow \infty} m = J(u^*) ,$$

то такая последовательность называется **минимизирующей**.

Из сходимости последовательности функционалов, вообще говоря, не всегда следует сходимость последовательности их аргументов, т.е. из того, что последовательность $\{u_n(x)\}$ — минимизирующая, еще не следует, что она сходится к $u^*(x)$. Эта сходимость будет иметь место только при определенных условиях, которым должен быть подчинен способ построения минимизирующей последовательности.

В качестве n -го приближения к $u^*(x)$ будем брать n -й член некоторой последовательности $\{u_n(x)\}$. Способ, который мы рассмотрим далее, принадлежит В. Ритцу и был предложен им в 1908 г. Его основная идея состоит в замене задачи нахождения минимума функционала более простой задачей поиска минимума функции.



Рассмотрим семейство функций

$$u_n(x) = \varphi(x, a_1, a_2, \dots, a_n), \quad n = 1, 2, \dots, \quad (9.42)$$

где φ — некоторая заданная функция, а a_1, a_2, \dots, a_n — числовые параметры. Будем считать, что при любых конечных значениях параметров a_i каждая функция этого семейства удовлетворяет условиям:

- 1) $\varphi(x, a_1, a_2, \dots, a_n) \in C^1[a, b]$;
- 2) $\varphi(a, a_1, a_2, \dots, a_n) = A$; $\varphi(b, a_1, a_2, \dots, a_n) = B$.

Выписанные условия означают, что все функции рассматриваемого семейства являются допустимыми.

Легко видеть, что значение функционала (9.41) на функции $u_n(x)$ представляет собой некоторую функцию, аргументами которой являются числовые параметры a_1, a_2, \dots, a_n , т.е.

$$J(u_n) = \int_a^b F(x, u_n, u'_n) dx = \Phi(a_1, a_2, \dots, a_n).$$

Таким образом, задача об отыскании минимума функционала $J(u_n)$ по всевозможным функциям $u_n(x)$ свелась к задаче отыскания минимума функции $\Phi(a_1, a_2, \dots, a_n)$. При этом

$$J(u_n^*) = \min_{a_1, \dots, a_n} \Phi(a_1, a_2, \dots, a_n) = m_n = \Phi(a_1^*, a_2^*, \dots, a_n^*).$$

Записав необходимые условия минимума первого порядка, получим следующую систему уравнений для определения параметров $a_1^*, a_2^*, \dots, a_n^*$:

$$\frac{\partial \Phi(a_1^*, a_2^*, \dots, a_n^*)}{\partial a_i} = 0, \quad i = 1, 2, \dots, n \quad (9.43)$$

Если система (9.43) (в общем случае нелинейная) имеет решение, то найденный набор a_i^* мы и возьмем в качестве искомого. Тем самым будет построена последовательность Ритца. При этом, очевидно, $J(u_n) \geq J(u_n^*)$.

Если при этом взятое нами семейство функций $\{u_n(x)\}$ вида (9.42) будет достаточно широким и будет хорошо отражать свойства класса допустимых функций, то можно надеяться, что построенная последовательность будет минимизирующей, т.е.

$$\lim_{n \rightarrow \infty} J(u_n^*) = m = J(u^*)$$



Теорема 9.3. Если функция $F(x, u, u')$ непрерывна в области $a \leq x \leq b; -\infty < u, u' < +\infty$, а семейство функций (9.42) расширяется с увеличением p и обладает свойством C^1 -полноты, то построенная по Ритцу последовательность $\{u_n^*(x)\}$ — минимизирующая.

[[Доказательство](#)]

Сходимость минимизирующей последовательности к минимизирующей функции

Вновь возвратимся к нашей конкретной граничной задаче

$$\begin{cases} Lu(x) \equiv -(p(x)u'(x))' + q(x)u(x) = -f(x), \\ u(a) = A, \\ u(b) = B. \end{cases} \quad (9.44)$$

Теорема 9.4. Если выполняются условия:

- 1) $p(x) \geq p_0 > 0$ и $p(x) \in C^1[a, b]$;
- 2) $q(x) \geq 0$ и $q(x), f(x) \in C[a, b]$;
- 3) последовательность функций $\{u_n(x)\}$ является минимизирующей для вариационной задачи (9.35), то эта последовательность функций будет равномерно сходящейся на отрезке $[a, b]$ к $u^{*(x)}$ — решению граничной задачи (9.44).

[[Доказательство](#)]

Построение минимизирующей последовательности по Ритцу

На практике часто с целью упрощения системы (9.43) функции (9.42) берут в виде обобщенных полиномов. Тогда построение минимизирующей по методу Ритца может выглядеть следующим образом. Сначала выбирается последовательность координатных функций $\{\varphi_k(x)\}$, $k = 0, 1, 2, \dots$, удовлетворяющих условиям:

- 1) $\varphi_k(x) \in C^1[a, b]$, $k = 0, 1, 2, \dots$;
- 2) для функции $\varphi_0(x)$ должны выполняться граничные условия, т.е. $\varphi_0(a) = A$, $\varphi_0(b) = B$; другие функции $\varphi_k(x)$ должны удовлетворять однородным граничным условиям такого же типа, т.е. в нашем случае $\varphi_k(a) = \varphi_k(b) = 0$, $k = 1, 2, \dots$;
- 3) при любом конечном p функции $\varphi_1(x), \dots, \varphi_n(x)$ линейно независимы;



4) образованное по $\{\varphi_k(x)\}$ семейство $\{u_n(x)\}$, где

$$u_n(x) = \varphi_0(x) + \sum_{k=1}^n a_k \varphi_k(x),$$

обладает свойством C^1 -полноты.

Если систему функций $\{\varphi_k(x)\}$ выбрать указанным образом, то при любом выборе параметров a_1, \dots, a_n функция $u_n(x)$ будет непрерывно дифференцируемой и удовлетворять граничным условиям, т.е. — допустимой.

Заметим также, что при таком выборе $u_n(x)$ функционал $J(u_n)$ будет квадратичным и поэтому вместо системы (9.43) для определения параметров a_1, \dots, a_n часто рассматривают другую систему:

$$\frac{1}{2} \frac{\partial J(u_n)}{\partial a_i} = 0, \quad i = 1, \dots, n \quad (9.45)$$

Распишем эту систему более подробно:

$$\frac{1}{2} \frac{\partial J(u_n)}{\partial a_i} = \int_a^b (pu'_n\varphi'_i + qu_n\varphi_i + f\varphi_i) dx = 0, \quad i = \overline{1, n}, \quad (*)$$

или

$$\int_a^b \left[p \left(\varphi'_0 + \sum_{j=1}^n a_j \varphi'_j \right) \varphi'_i + q \left(\varphi_0 + \sum_{j=1}^n a_j \varphi_j \right) \varphi_i + f\varphi_i \right] dx = 0, \quad i = \overline{1, n}.$$

Поменяв местами порядок суммирования и интегрирования и собирая коэффициенты при a_j , получим:

$$\sum_{j=1}^n a_j \int_a^b (p\varphi'_i\varphi'_j + q\varphi_i\varphi_j) dx + \int_a^b (p\varphi'_0\varphi'_i + q\varphi_0\varphi_i + f\varphi_i) dx = 0, \quad i = \overline{1, n}$$



или, если ввести обозначения

$$\alpha_{ij} = \int_a^b (p\varphi_i' \varphi_j' + q\varphi_i \varphi_j) dx, \quad \beta_i = \int_a^b (p\varphi_0' \varphi_i' + q\varphi_0 \varphi_i + f\varphi_i) dx, \quad i = \overline{1, n}; \quad j = \overline{1, n} \quad (9.46)$$

то система (9.45) примет вид

$$\sum_{j=1}^n \alpha_{ij} a_j + \beta_i = 0, \quad i = \overline{1, n} \quad (9.47)$$

Посмотрим, когда она разрешима и определена. Рассмотрим соответствующую однородную систему:

$$\sum_{j=1}^n \alpha_{ij} a_j = 0, \quad i = \overline{1, n}.$$

Учитывая (*), запишем ее в виде

$$\int_a^b (pz_n' \varphi_i' + qz_n \varphi_i) dx = 0, \quad i = \overline{1, n},$$

где

$$z_n(x) = \sum_{j=1}^n a_j \varphi_j(x).$$

Умножим теперь i -е уравнение системы на a_i и просуммируем все уравнения по i от единицы до n . В итоге получим:

$$\int_a^b \left(p(z_n')^2 + qz_n^2 \right) dx = 0.$$

Так как интеграл равен нулю и подынтегральная функция неотрицательна, то отсюда имеем:

$$p(z_n')^2 + qz_n^2 \equiv 0,$$



а следовательно, поскольку каждое слагаемое неотрицательно, $p(z'_n)^2 \equiv 0$ и $qz_n^2 \equiv 0$.

Далее, в силу того что $p(x) > 0$, то $z'_n(x) \equiv 0$, т.е. $z_n(x) = \text{const}$ или, поскольку $z_n(a) = z_n(b) = 0$, то $z_n(x) \equiv 0$.

Таким образом, мы получили, что

$$\sum_{j=1}^n a_j \varphi_j(x) = z_n(x) \equiv 0.$$

Отсюда, в силу линейной независимости системы $\{\varphi_i(x)\}$ следует, что $a_1 = a_2 = \dots = a_n = 0$, т.е. однородная система имеет лишь тривиальное решение, а поэтому соответствующая ей неоднородная система (9.46), (9.47) разрешима, причем единственным образом.

Легко видеть, что последовательность $\{u_n(x)\}$ будет минимизирующей, поскольку все условия Теоремы 9.3 при таком выборе $\varphi_i(x)$ будут выполнены.

В качестве системы функций $\{\varphi_i(x)\}$ могут быть взяты следующие системы:

$$1) \quad \varphi_i(x) = (x - a)^i (b - x) \quad i = \overline{1, n} \quad (9.48)$$

$$2) \quad \varphi_k(x) = \sin k\pi \frac{x - a}{b - a}, \quad k = \overline{1, n} \quad (9.49)$$



9.3.3. Метод моментов и метод Галеркина решения граничных задач

Пусть при $x \in [a, b]$ задано дифференциальное уравнение

$$F(x, u, u', u'') = 0 \quad (9.50)$$

с граничными условиями

$$\begin{cases} u(a) = A, \\ u(b) = B. \end{cases} \quad (9.51)$$

Будем считать, что граничная задача (9.50), (9.51) имеет на отрезке $[a, b]$ единственное решение, принадлежащее классу $C^2[a, b]$.

Следуя общей идеи метода моментов, рассмотрим две системы функций:

- 1) Система функций $\{\psi_k(x)\}$, $k = 1, \dots, \infty$, подчиненная условиям
 - (a) $\psi_k(x) \in C[a, b]$, $k = 1, 2, \dots$;
 - (b) функции $\psi_k(x)$ образуют замкнутую систему, т.е. из того что

$$\int_a^b f(x) \psi_k(x) dx = 0, \quad k = 1, 2, \dots$$

должно с необходимостью следовать, что $f(x) \equiv 0$, если $f(x) \in C[a, b]$.

- 2) Система функций $\{\varphi_k(x)\}$, $k = 0, 1, \dots$, удовлетворяющая условиям:
 - (a) $\varphi_k(x) \in C^2[a, b]$, $k = 0, 1, \dots$;
 - (b) при любом конечном n функции $\varphi_0(x), \dots, \varphi_n(x)$ линейно независимы на $[a, b]$;
 - (c) функция $\varphi_0(x)$ удовлетворяет граничным условиям (9.51) а $\varphi_1(x), \dots, \varphi_n(x)$ удовлетворяют требованию $\varphi_k(a) = \varphi_k(b) = 0$, $k = 1, 2, \dots$;



(d) функции $\varphi_k(x)$ образуют в классе $C^2[a, b]$ полную систему.

Перейдем к построению приближенного решения граничной задачи (9.50), (9.51) по методу моментов. Функцию F в (9.50) будем считать непрерывной по всем аргументам в области $\{a \leq x \leq b, -\infty < u, u', u'' < \infty\}$.

Составим линейную комбинацию

$$u_n(x) = \varphi_0(x) + \sum_{k=1}^n a_k \varphi_k(x) \quad (9.52)$$

где a_k — некоторые параметры. В силу выбора функций $\varphi_k(x)$ $u_n(x)$ удовлетворяет граничным условиям (9.51) при любых значениях параметров a_k . Обсудим сейчас проблему выбора последних.

Подставим приближенное решение (9.52) в исходное дифференциальное уравнение (9.50):

$$F(x, u_n, u'_n, u''_n) = \delta(x, a_1, a_2, \dots, a_n) = \delta_n(x).$$

Здесь функция $\delta_n(x)$ играет роль невязки на приближенном решении. При любом n $\delta_n(x)$ непрерывны, так как непрерывна функция F . Очевидно, выбор параметров a_1, \dots, a_n следует осуществлять таким образом, чтобы величина $\delta_n(x)$ была близка к нулю.

Ясно, что если бы удалось удовлетворить бесконечное число требований

$$\int_a^b \delta_n(x) \psi_i(x) dx = 0, \quad i = 1, 2, \dots \quad (9.53)$$

то отсюда в силу замкнутости системы функций $\{\psi_k(x)\}$ следовало бы, что $\delta_n(x) = 0$, т.е. $u_n(x)$ — точное решение задачи (9.50), (9.51). Однако за счет выбора конечного числа параметров бесконечному числу требований практически невозможно. Поэтому было бы разумно параметры a_1, \dots, a_n выбирать, удовлетворяя первым из n требований системы (9.53), т.е.

$$\int_a^b \delta_n(x) \psi_i(x) dx = 0, \quad i = 1, 2, \dots, n \quad (9.54)$$



В этом случае можно надеяться, что $\delta_n(x)$ будет близко к нулю и при выполнении некоторых дополнительных условий $\delta_n(x) \xrightarrow{n \rightarrow \infty} 0$. Можно также ожидать, что $u_n(x)$, определяемое по формуле (9.52), также будет близко к $u(x)$.

Таким образом, в методе моментов приближенное решение ищется в виде (9.52), причем параметры a_1, \dots, a_n определяются из системы уравнений (9.54).

Если уравнение (9.50) будет линейным, т.е.

$$F(x, u, u', u'') \equiv Lu - f(x) = 0$$

где

$$Lu(x) \equiv u''(x) + p(x)u(x) + q(x)u(x) \quad (9.55)$$

то запись системы (9.54) упростится. В этом случае мы получим систему

$$\sum_{k=1}^n c_{ki}a_k - D_i = 0, \quad i = \overline{1, n} \quad (9.56)$$

где

$$c_{ki} = \int_a^b L\varphi_k(x)\psi_i(x)dx, \quad D_i = \int_a^b [f(x) - L\varphi_0(x)]dx \quad (9.57)$$

В частном случае, когда системы функций $\{\psi_i(x)\}$ и $\{\varphi_i(x)\}$ совпадают (при этом все сформулированные к ним ранее требования остаются в силе), мы получим алгоритм метода Галеркина.

Можно также показать, что в случае, когда уравнение (9.30) является уравнением Эйлера, т.е.

$$F(x, u, u', u'') \equiv \frac{d}{dx} \frac{\partial f(x, u, u')}{\partial u'} - \frac{\partial f(x, u, u')}{\partial u} = 0,$$

системы уравнений для определения параметров по методу Галеркина и по методу Ритца будут совпадать, т.е. при одинаковом n два этих метода дают одинаковое приближенное решение, определяемое по формуле (9.52).



9.3.4. Метод наименьших квадратов решения граничных задач

Этот метод, как и рассмотренные выше, применим к любым функциональным уравнениям. В частности, мы уже применяли его к решению интегральных уравнений. Там же мы сформулировали основные его черты как представителя семейства проекционных методов и отличия от других представителей данного семейства. Рассмотрим сейчас его применение к решению дифференциальных уравнений на примере задачи (9.50), (9.51). Предположения относительно функции F оставим прежними. Приближенное решение, как и ранее, будем искать в виде (9.52), т.е.

$$u_n(x) = \varphi_0(x) + \sum_{k=1}^n a_k \varphi_k(x), \quad (9.58)$$

причем $\varphi_k(x)$ удовлетворяют тем же четырем требованиям, что и в методе моментов.

Вновь составим невязку

$$\delta_n(x) = \delta(x, a_1, \dots, a_n) = F(x, u_n, u'_n, u''_n).$$

Однако на сей раз выбор параметров a_1, \dots, a_n подчиним другому требованию: будем добиваться среднеквадратичной малости невязки $\delta_n(x)$ на отрезке $[a, b]$ (или, что то же самое, квадрата L_2 -нормы невязки). Минимизация ее эквивалентна минимизации функционала

$$J(u_n) = \int_a^b F^2(x, u_n, u'_n, u''_n) dx = \Phi(a_1, \dots, a_n).$$

Записывая необходимые условия минимума первого порядка, получим систему уравнений

$$\frac{\partial \Phi(a_1, \dots, a_n)}{\partial a_i} = 0, \quad i = \overline{1, n}, \quad (9.59)$$

из которой находим (если это возможно) значения параметров a_1, \dots, a_n . Исследуем подробнее линейный случай, т.е. случай, когда уравнение (9.50) имеет вид (9.55). Тогда система (9.59) может быть записана в виде

$$\sum_{j=1}^n c_{ij} a_j - b_i = 0, \quad i = \overline{1, n}, \quad (9.60)$$



где

$$c_{ij} = \int_a^b L\varphi_i(x) L\varphi_j(x) dx, \quad i, j = \overline{1, n}, \quad (9.61)$$

$$b_i = \int_a^b (f(x) - L\varphi_0(x)) \varphi_i(x) dx, \quad i = \overline{1, n}.$$

Разрешимость системы (9.60), (9.61) зависит не только от свойств системы функций $\{\varphi_i(x)\}$, но также и от природы рассматриваемой граничной задачи, в частности, от того, имеет ли однородная граничная задача

$$\begin{cases} Lu(x) = 0, \\ u(a) = u(b) = 0 \end{cases}$$

только нулевое решение.



9.3.5. Метод коллокации решения граничных задач

Вновь рассмотрим граничную задачу (9.50), (9.51), систему функций $\{\varphi_k(x)\}$, удовлетворяющую прежним требованиям, а приближенное решение ищем в виде

$$u_n(x) = \varphi_0(x) + \sum_{k=1}^n a_k \varphi_k(x).$$

В соответствии с идеологией метода коллокации, изложенной ранее на примере применения метода для решения интегральных уравнений, потребуем, чтобы невязка $\delta_n(x)$ была мала в следующем смысле: чтобы в некоторых заданных точках отрезка $[a, b]$ x_1, \dots, x_n эта невязка обращалась в нуль:

$$\delta_n(x_i) = 0, \quad i = \overline{1, n}. \quad (9.62)$$

В итоге получаем систему (в общем случае нелинейных) уравнений для определения параметров a_1, \dots, a_n приближенного решения.

Если исходная задача имеет вид (9.55) (т.е. становится линейной), то и система (9.62) также станет линейной:

$$\sum_{i=1}^n a_i L\varphi_i(x_j) = f(x_j) - L\varphi_0(x_j), \quad j = \overline{1, n}. \quad (9.63)$$

Для разрешимости последней необходимо выполнение условия

$$\begin{vmatrix} L\varphi_1(x_1) & \cdots & L\varphi_n(x_1) \\ \vdots & \ddots & \vdots \\ L\varphi_1(x_n) & \cdots & L\varphi_n(x_n) \end{vmatrix} \neq 0. \quad (9.64)$$

Требование (9.63), как легко видеть, равносильно тому, чтобы система функций $\{L\varphi_i(x)\}$ была [системой функций Чебышева](#) на отрезке $[a, b]$.

Поскольку метод коллокации можно рассматривать и как решение задачи об интерполяции функции $f(x)$ обобщенным многочленом, построенным по системе функций $\{L\varphi_i(x)\}$ на заданном множестве



узлов x_1, \dots, x_n , то задача выбора последних также имеет немаловажное значение. В то же время, как мы помним, существуют большие проблемы со [сходимостью интерполяционных процессов](#). Поэтому в целом, несмотря на простоту системы (9.25), метод коллокации в изложенном виде применяется сравнительно редко.

Замечание 9.6. Все изложенные выше алгоритмы проекционно-вариационного типа можно рассматривать с точки зрения теории приближения функций с той лишь разницей, что вместо совпадения на множестве точек (как в методе коллокации) рассматриваются интегральные аналоги этих условий. Так, например, метод наименьших квадратов — это, фактически построение [наилучшего среднеквадратичного приближения](#) к функции $f(x)$ и т.п.



Меню

Глава 10

Численные методы математической физики

- [10.1. Основные понятия теории разностных схем](#)
- [10.2. Способы построения разностных схем](#)
- [10.3. Методы исследования устойчивости разностных схем](#)
- [10.4. Разностные схемы для стационарных задач математической физики](#)
- [10.5. Итерационные методы решения разностных задач](#)
- [10.6. Численные методы решения задач математической физики в областях сложной формы](#)



Для решения граничных задач (причем не только для обыкновенных дифференциальных уравнений, но и для уравнений с частными производными) помимо изученных нами ранее алгоритмов достаточно широко и уже давно применяется еще один подход, обладающий известной универсальностью: подход, позволяющий достаточно несложным образом свести решение указанной выше задачи к решению системы уравнений, неизвестными которой, как правило, являются значения приближенного решения на заданном каким-либо способом множестве точек (узлов). Получающиеся алгоритмы называются сеточными, а метод их получения — методом сеток (или конечных разностей). Раздел численных методов, посвященный теории метода сеток, носит название теории разностных схем. Далее мы познакомимся с ее основными моментами.



10.1. Основные понятия теории разностных схем

[10.1.1. Сетки](#)

[10.1.2. Сеточные функции](#)

[10.1.3. Разностная аппроксимация дифференциальных операторов](#)

[10.1.4. Погрешность аппроксимации на сетке](#)

[10.1.5. Постановка разностной задачи](#)

[10.1.6. Сходимость и точность разностных схем](#)

[10.1.7. Повышение порядка аппроксимации разностных схем](#)

[10.1.8. Математический аппарат теории разностных схем](#)



Меню

10.1.1. Сетки

[Равномерная сетка на отрезке](#)

[Неравномерная сетка на отрезке](#)

[Сетка в прямоугольнике](#)

[Сетка в криволинейной ортогональной системе координат](#)

[Пространственно-временная сетка в прямоугольнике](#)

[Прямоугольная сетка в области сложной формы](#)

[Треугольная сетка в области сложной формы](#)

[Сетка на криволинейном четырехугольнике](#)

Система алгебраических уравнений, заменяющая исходную дифференциальную задачу и зависящая от шага замены как от параметра, обычно называется разностной схемой.

Для того чтобы написать разностную схему, приближенно описывающую рассматриваемую дифференциальную задачу, необходимо совершить следующие два шага:

- 1) Заменить область непрерывного изменения аргумента областью дискретного его изменения;
- 2) Заменить дифференциальные операторы некоторыми разностными операторами, а также сформулировать аналогичным образом разностные аналоги краевых условий.

Остановимся на этих вопросах подробнее.

При численном решении той или иной математической задачи (связанной с решением функциональных уравнений) мы, очевидно, не можем воспроизводить решение для всех значений аргумента, изменяющегося внутри некоторой области евклидова пространства. Естественно поэтому выбрать в этой области некоторое конечное подмножество точек и приближенное решение искать только в этих точках. Такое множество точек мы в дальнейшем будем называть *сеткой*. Отдельные точки этого множества будем называть *узлами*. Функцию, определенную в узлах сетки, будем называть *сеточной функцией*.

Таким образом, мы заменили область непрерывного изменения аргумента сеткой, т.е. областью дискретного изменения аргумента. Иными словами, мы осуществили *аппроксимацию* пространства решений исходной дифференциальной задачи пространством сеточных функций.



Меню

Свойства приближенного (разностного) решения, и в частности, его близость к точному решению, зависят от выбора сетки.

Рассмотрим сейчас примеры наиболее часто используемых типов сеточных областей.

Равномерная сетка на отрезке

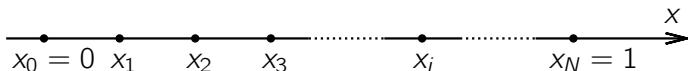


Рисунок 10.1

Рассмотрим стандартный отрезок $[0, 1]$ и разобьем его на заданное число N равных частей. Расстояние между соседними узлами $x_i - x_{i-1} = h = \frac{1}{N}$ назовем *шагом сетки*, а точки деления $x_i = ih$ примем в качестве узлов сетки. Множество всех узлов x_i и составляет равномерную сетку на отрезке $[0, 1]$, которую в дальнейшем будем обозначать ω_h :

$$\omega_h = \left\{ x_i = ih; \quad i = 1, 2, \dots, N-1; \quad h = \frac{1}{N} \right\}.$$

В это множество можно включать и граничные узлы. Обозначение такой сетки — $\bar{\omega}_h$:

$$\bar{\omega}_h = \left\{ x_i = ih; \quad i = 0, 1, 2, \dots, N; \quad h = \frac{1}{N} \right\}.$$

На отрезке $[0, 1]$ вместо функции непрерывного аргумента $y(x)$ будем рассматривать функцию дискретного аргумента $y_h(x_i)$. Значения этой функции вычисляются только в узлах x_i , а сама функция зависит от шага сетки h как от параметра.

Неравномерная сетка на отрезке

Вновь рассмотрим отрезок $[0, 1]$. Вводя произвольные точки $0 < x_1 < x_2 < \dots < x_{N-1} < 1$, разобьем его на N частей. Множество узлов

$$\hat{\omega}_h := \{x_i, \quad i = 0, \dots, N; \quad x_0 = 0, \quad x_N = 1\}$$

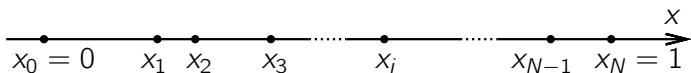


Рисунок 10.2

образует неравномерную сетку на отрезке $[0, 1]$. Расстояние между соседними узлами — шаг сетки — равно $h_i = x_i - x_{i-1}$ и зависит от номера i узла, т.е. является сеточной функцией. Шаги сетки \hat{h}_h удовлетворяют условию нормировки

$$\sum_{i=1}^N h_i = 1.$$

Сетка в прямоугольнике

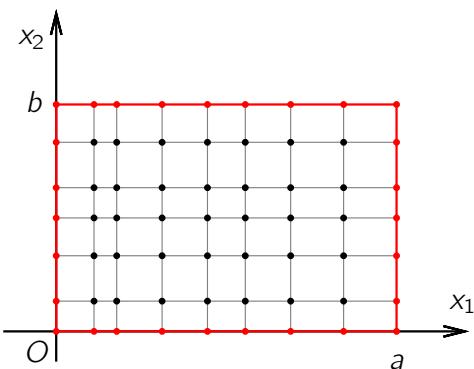


Рисунок 10.3



Пусть исходная область G является прямоугольником с основанием a и высотой b , основание которого лежит на оси Ox_1 , а одна из боковых сторон — на оси Ox_2 , т.е.

$$\bar{G} = \{(x_1, x_2) : 0 \leq x_1 \leq a; 0 \leq x_2 \leq b\}.$$

Разобьем отрезки $[0, a]$ и $[0, b]$ на N_1 и N_2 частей соответственно. Пусть точки деления на оси Ox_1 имеют координаты x_{1,i_1} , а на оси Ox_2 — x_{2,i_2} , причем

$$0 = x_{1,0} < x_{1,1} < \cdots < x_{1,N_1-1} < x_{1,N_1} = a,$$

$$0 = x_{2,0} < x_{2,1} < \cdots < x_{2,N_2-1} < x_{2,N_2} = b.$$

Через точки деления проведем два семейства прямых

$$x_1 = x_{1,i_1}, \quad i_1 = 0, 1, \dots, N_1;$$

$$x_2 = x_{2,i_2}, \quad i_2 = 0, 1, \dots, N_2,$$

параллельных соответствующим координатным осям.

В качестве узлов сетки возьмем точки пересечения этих прямых. Общее число узлов равно $(N_1 + 1) \cdot (N_2 + 1)$ и все они принадлежат прямоугольнику \bar{G} . Распределение узлов характеризуется векторным параметром $h = \{h_{1,1}, \dots, h_{1,N_1}; h_{2,1}, \dots, h_{2,N_2}\}$, составленным из шагов по каждому направлению:

$$h_{1,i_1} = x_{1,i_1} - x_{1,i_1-1}, \quad i_1 = 1, \dots, N_1;$$

$$h_{2,i_2} = x_{2,i_2} - x_{2,i_2-1}, \quad i_2 = 1, \dots, N_2.$$

Если все шаги сетки как по направлению x_1 , так и по направлению x_2 , равны между собой, т.е. $h_{1,1} = h_{1,2} = \cdots = h_{1,N_1} =: h_1 = \frac{a}{N_1}$, $h_{2,1} = h_{2,2} = \cdots = h_{2,N_2} =: h_2 = \frac{b}{N_2}$, то сетка называется *равномерной* и обозначается

$$\bar{\omega}_h = \bar{\omega}_{h_1 h_2} = \left\{ (x_{1,i_1}, x_{2,i_2}) : x_{1,i_1} = i_1 h_1; x_{2,i_2} = i_2 h_2; h_1 = \frac{a}{N_1}, h_2 = \frac{b}{N_2}; i_1 = \overline{0, N_1}, i_2 = \overline{0, N_2} \right\}.$$



В противном случае сетка называется *неравномерной* и обозначается

$$\hat{\omega}_h = \hat{\omega}_{h_1 h_2} = \hat{\omega}_{h_1} \times \hat{\omega}_{h_2} = \{(x_{1,i_1}, x_{2,i_2}), \quad i_1 = \overline{1, N_1}, \quad i_2 = \overline{1, N_2}; \quad x_{1,0} = 0, \quad x_{1,N_1} = a; \quad x_{2,0} = 0, \quad x_{2,N_2} = b\}.$$

Если $h_1 = h_2$, то сетку называют *квадратной*.

Неравномерные сетки бывают эффективны в случае сильно неоднородного решения. Узлы сетки в этом случае сгущают в зоне сильно изменяющегося решения за счет уменьшения их плотности на участках, где решение меняется слабо, не увеличивая их общего количества.

Сетка в криволинейной ортогональной системе координат

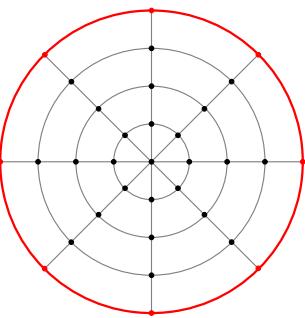


Рисунок 10.4

В качестве примера рассмотрим область \bar{G} , имеющую вид круга радиуса R . Двумерную задачу в такой области удобно формулировать в полярных координатах (r, φ) , поместив полюс в центр круга. Тогда $\bar{G} = \{(r, \varphi) : 0 \leq r \leq R, 0 \leq \varphi < 2\pi\}$.

Проведем два семейства кривых, параллельных координатным линиям:

$$r = r_i, \quad i = \overline{1, N_1};$$

$$\varphi = \varphi_j, \quad j = \overline{1, N_2 - 1}.$$



Кривые первого семейства являются концентрическими окружностями, а второе семейство образуют лучи, исходящие из полюса. Узлами рассматриваемой сетки будут точки пересечения данных линий, т.е. множество точек $\omega_{r\varphi} = \{(r_i, \varphi_j), i = \overline{0, N_1}, j = \overline{0, N_2 - 1}; r_0 = 0, r_{N_1} = R; \varphi_0 = 0 = \varphi_{N_2}\}$.

Пространственно-временная сетка в прямоугольнике

Сетка строится аналогично [сетке в прямоугольнике](#), но в других обозначениях: рассматривается область $\overline{G} = \{(x, t) : 0 \leq x \leq a, 0 \leq t \leq T\}$ и равномерная сетка на ней имеет вид

$$\bar{\omega}_{h\tau} = \left\{ (x_i, t_j) : x_i = ih, t_j = j\tau; h = \frac{a}{N_1}, \tau = \frac{T}{N_2}; i = \overline{0, N_1}, j = \overline{0, N_2} \right\},$$

а неравномерная —

$$\hat{\omega}_{h\tau} = \left\{ (x_i, t_j) : x_i = x_{i-1} + h_i, t_j = t_{j-1} + \tau_j; i = \overline{0, N_1}, j = \overline{0, N_2}; x_0 = 0, x_{N_1} = a; t_0 = 0, t_{N_2} = T \right\}.$$

Прямоугольная сетка в области сложной формы

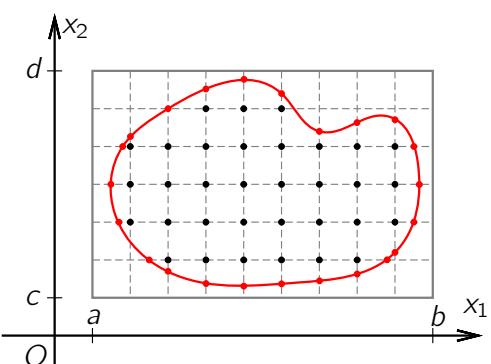


Рисунок 10.5

Пусть в плоскости Ox_1x_2 задана область сложной формы G с границей Γ . Заключим область \bar{G} в прямоугольник $\bar{\Pi} = [a, b] \times [c, d]$ (этот прямоугольник может быть как в некотором смысле минимальным, когда на каждой стороне прямоугольника существуют точки области \bar{G} , принадлежащие этой стороне, так и любым прямоугольником, содержащим в себе описанный минимальный). После этого зададим на данном прямоугольнике сетку $\bar{\omega}_h$. Те узлы сетки $\bar{\omega}_h$, которые принадлежат области \bar{G} , а также точки пересечения прямых, образующих сетку $\bar{\omega}_h$ с границей Γ , и составляют искомую прямоугольную сетку $\bar{\Omega}_h$ в области сложной формы. Несмотря на то, что исходная сетка $\bar{\omega}_h$ равномерна по каждому направлению, построенная сетка $\bar{\Omega}_h$ таковой может не оказаться. Легко видеть, что равномерность может быть нарушена вблизи границы.

Треугольная сетка в области сложной формы

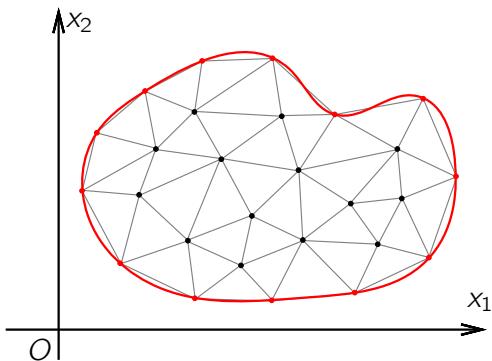


Рисунок 10.6

Вновь рассмотрим область \bar{G} сложной формы, описанную выше. Выберем на границе Γ области множество точек, являющихся узлами ломаной. Эта ломаная будет границей некоторого многоугольника Π , которым мы приближенно заменим исходную область \bar{G} . Многоугольник Π покроем множеством треугольником, каждая пара которых либо вовсе не имеет общих точек, либо имеет общую вершину, либо имеет

общую сторону. Таким образом, получим треугольную сетку в области сложной формы. Все узлы (вершины треугольников) здесь можно занумеровать одним индексом: $\xi_p = (x_{1,p}, x_{2,p})$, $p = \overline{0, N}$. Эти узлы выбираются внутри и на границе области \bar{G} , вообще говоря, произвольно, исходя из структуры решения и требований точности.

Сетка на криволинейном четырехугольнике

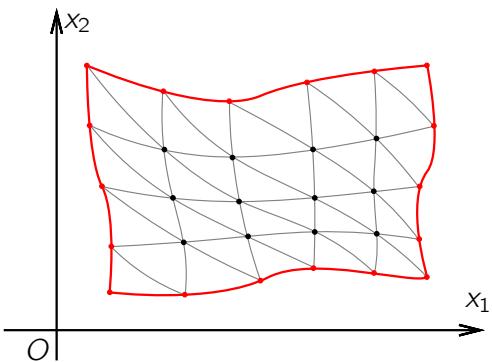


Рисунок 10.7

Если же область \bar{G} представляет собой некоторый криволинейный четырехугольник, то в этом случае узлы удобно нумеровать двумя индексами. Сначала выбираем узлы на границе — равное количество на противоположных сторонах криволинейного четырехугольника. Противоположные точки соединяют двумя семействами попарно непересекающихся кривых, включая границы (сравнить: процесс построения параметрических сплайнов от двух переменных). Точки пересечения этих семейств образуют узлы: $\xi_{i_1 i_2} = (x_{1,i_1}, x_{2,i_2})$, $i_1 = \overline{0, N_1}$; $i_2 = \overline{0, N_2}$. Элементами сетки являются криволинейные треугольники.

Аналогичным образом строятся сетки и в задачах с количеством независимых переменных $n > 2$.

10.1.2. Сеточные функции

Итак, область \bar{G} изменения аргумента x мы заменяем сеткой $\bar{\omega}_h$, т.е. конечным множеством точек $x_i \in \bar{G}$. Вместо функций $u(x)$ непрерывного аргумента $x \in \bar{G}$ будем рассматривать *сеточные функции* $y(x_i)$, т.е. функции точки x_i , являющейся узлом сетки $\bar{\omega}_h$. Сеточную функцию можно представить в виде вектора. Если перенумеровать все узлы в некотором порядке (x_1, x_2, \dots, x_N) , то значения сеточной функции в этих узлах можно рассматривать как компоненты вектора $Y = (y_1, y_2, \dots, y_N)$. Если область \bar{G} , в которой построена сетка, конечна, то размерность N вектора Y также конечна. В случае неограниченной области G сетка состоит из бесконечного числа узлов и, следовательно, размерность вектора Y также бесконечна.

Обычно рассматриваются множества сеток $\{\omega_h\}$, зависящие от шага h как от параметра. Поэтому и сеточные функции $y_h(x)$ зависят от параметра h (или от числа узлов N в случае равномерной сетки). Если сетка ω_h неравномерна, то под h следует понимать вектор $h = (h_1, h_2, \dots, h_N)$. Это же замечание относится и к случаю, когда область G многомерна, т.е. $x = (x_1, x_2, \dots, x_p)$; тогда $h = (h_1, h_2, \dots, h_p)$, если сетка ω_h равномерна по каждому из аргументов x_1, x_2, \dots, x_p .

Функции $u(x)$ непрерывного аргумента $x \in G$ являются элементами некоторого функционального пространства H_0 . Множество сеточных функций $y_h(x)$ образует пространство H_h . Таким образом, используя метод конечных разностей, мы заменяем пространство H_0 функций $u(x)$ непрерывного аргумента пространством H_h сеточных функций $y_h(x)$.

Рассматривая множество сеток $\{\omega_h\}$, мы получаем множество пространств сеточных функций $\{H_h\}$, зависящих от параметра h . В линейном пространстве H_h вводится норма $\|\cdot\|_h$, являющаяся сеточным аналогом нормы $\|\cdot\|_0$ в исходном пространстве H_0 .

Укажем простейшие типы норм в пространстве H_h сеточных функций для случая сетки $\bar{\omega}_h = \{x_i = ih, i = \overline{0, N}, h = \frac{1}{N}\}$ на отрезке $[0, 1]$ (индекс h у y_h тогда будем опускать):

- Сеточный аналог нормы в пространстве C ($H_0 = C$):

$$\|y\|_C = \max_{x \in \bar{\omega}_h} |y(x)| \quad \text{или} \quad \|y\|_C = \max_{0 \leq i \leq N} |y_i|.$$

- Сеточные аналоги нормы в пространстве L_2 ($H_0 = L_2$):

$$\|y\| = \left(\sum_{i=1}^{N-1} hy_i^2 \right)^{\frac{1}{2}} \quad \text{или} \quad \|y\| = \left(\sum_{i=1}^N hy_i^2 \right)^{\frac{1}{2}}.$$

Пусть $u(x)$ — решение исходной непрерывной задачи, $u \in H_0$, y_h — решение приближенной (разностной) задачи, $y_h \in H_h$. Основной интерес для теории приближенных методов представляет оценка близости y_h к u . Однако y_h и u являются элементами различных функциональных пространств. Поэтому для изучения вопроса о близости y_h к u принципиально имеются две возможности:

- Сеточная функция y_h , заданная в узлах сетки $\omega_h(G)$, доопределяется (например, с помощью интерполяции) во всех остальных точках области G . В результате получаем функцию $\tilde{y}(x, h)$ непрерывного аргумента $x \in G$. Разность $\tilde{y}(x, h) - u(x)$ принадлежит пространству H_0 , а близость y_h к u характеризуется числом $\|\tilde{y}(x, h) - u(x)\|_0$ (см., например, [оценку погрешности метода механических квадратур](#));
- Пространство H_0 отображается на пространство H_h . Каждой функции $u(x) \in H_0$ ставится в соответствие сеточная функция $u_h(x)$, $x \in \omega_h$ так что $u_h = P_h u \in H_h$, где P_h — линейный оператор из H_0 в H_h . Это соответствие можно осуществить по-разному, выбирая различные операторы P_h . Если, например, $u(x)$ — непрерывная функция, то можно положить (и чаще всего делается именно так) $u_h(x) = u(x)$, $x \in \omega_h$. Иногда определяют $u_h(x)$, $x \in \omega_h$ как интегральное среднее значение $u(x)$ по некоторой окрестности (например, диаметра $O(h)$) данного узла x_i . Имея сеточную функцию u_h , образуем разность $y_h - u_h \in H_h$. Близость y_h к u будет, следовательно, характеризоваться числом $\|y_h - u_h\|_h$. При этом естественно требовать, чтобы норма $\|\cdot\|_h$ аппроксимировала норму $\|\cdot\|_0$:

$$\lim_{h \rightarrow 0} \|u_h\|_h = \|u\|_0 \quad \text{для всех } u \in H_0.$$

Это условие будем называть [условием согласованности норм](#) в H_h и H_0 .

При изучении вопроса о близости y_h к u чаще всего используется второй подход.



10.1.3. Разностная аппроксимация дифференциальных операторов

[Способ неопределенных коэффициентов](#)

[Способ численного дифференцирования](#)

После того как область G заменена сеточной областью ω_h , можно переходить к следующему этапу: замене дифференциального оператора его разностным аналогом. Такую замену обычно называют [аппроксимацией дифференциального оператора разностным оператором](#).

Рассмотрим этот вопрос несколько подробнее. Итак, пусть задан линейный дифференциальный оператор L , действующий на функцию $u = u(x)$. Для того чтобы аппроксимировать его в любой точке сетки ω_h разностным оператором L_h , действующим на сеточную функцию u_h , необходимо вначале указать (выбрать) [шаблон](#), т.е. множество узлов $\mathbb{W}(x)$ сетки, которое будет непосредственно использоваться при аппроксимации оператора L оператором L_h в точке $x \in \omega_h$. Обычно выбор шаблона $\mathbb{W}(x)$ зависит от порядка производных, входящих в оператор L , а также от некоторых других моментов.

Сама же аппроксимация может быть осуществлена двумя способами:

- методом неопределенных коэффициентов;
- методом численного дифференцирования.

Прежде чем рассматривать эти способы, напомним, что разность

$$\psi(x) = L_h u(x) - L u(x), \quad x \in \omega_h$$

называется [погрешностью аппроксимации](#) дифференциального оператора L разностным оператором L_h в точке $x \in \omega_h$.

Кроме того, будем говорить, что L_h [аппроксирует дифференциальный оператор \$L\$ с порядком \$m > 0\$](#) в точке $x \in \omega_h$, если $\psi(x) = O(|h|^m)$.

Рассмотрим теперь подробнее способы построения разностных операторов.

[Способ неопределенных коэффициентов](#)

Выбрав шаблон $\mathbb{W}(x)$, разностную аппроксимацию $L_h u(x)$ будем искать в виде линейной комбинации значений функции u в точках этого шаблона:



$$L_h u(x) = \sum_{\xi \in \mathbb{W}(x)} A_h(x, \xi) u(\xi), \quad (10.1)$$

где $A_h(x, \xi)$ — неизвестные коэффициенты, выбор которых осуществляется таким образом, чтобы [погрешность аппроксимации](#) $\psi(x)$ имела в точке x заданный (чаще всего — максимально возможный в данной ситуации) порядок. Практически это осуществляется путем разложения (при естественном предположении законности этой операции) погрешности аппроксимации $\psi(x)$ в ряд Тейлора

$$\psi(x) = \sum_{\xi \in \mathbb{W}(x)} A_h(x, \xi) u(\xi) - L u(x) = \sum_{|j| \geq 0} B_h^{(j)}(x) u^{(j)}(x)$$

и приравниванием к нулю заданного (максимального) количества первых членов разложения. После этого, решив получившуюся систему линейных алгебраических уравнений, найдем коэффициенты $A_h(x, \xi)$ и по формуле (10.1) запишем искомый разностный оператор.

Несложно заметить, что для аппроксимации дифференциального оператора, содержащего производную порядка k по некоторой независимой переменной, необходимо использовать шаблон, содержащий не менее $(k+1)$ точек вдоль координатного направления по данной независимой переменной, поскольку при разложении функции $u(x)$ в ряд Тейлора по данной переменной производная k -го порядка будет находиться на $(k+1)$ -м месте.

Приведем примеры построения простейших разностных аппроксимаций.

Пример 10.1. Аппроксимация дифференциального оператора

$$L u(x) = \frac{du(x)}{dx} = u'(x)$$

разностным.

Решение. Необходимый для аппроксимации шаблон должен содержать, по крайней мере, две точки, одна из которых — точка аппроксимации x .



Правая разностная производная. Выберем в качестве шаблона $\mathbb{W}(x)$ узлы x и $x+h$, т.е. $\mathbb{W}(x) = \{x, x+h\}$. Тогда, согласно (10.1), разностный оператор L_h будем искать в виде

$$L_h u(x) = a_0 u(x) + a_1 u(x+h).$$

Запишем погрешность аппроксимации $\psi(x)$

$$\psi(x) = a_0 u(x) + a_1 u(x+h) - u'(x)$$

и разложим ее в ряд Тейлора в окрестности точки x :

$$\begin{aligned} \psi(x) &= a_0 u(x) + a_1 \left[u(x) + hu'(x) + \frac{h^2}{2} u''(x) + \dots \right] - u'(x) = \\ &= (a_0 + a_1) u(x) + (ha_1 - 1) u'(x) + \frac{h^2}{2} u''(x) + \dots . \end{aligned}$$

Приравнивая к нулю первые коэффициенты разложения, получим:

$$\begin{cases} a_0 + a_1 = 0, \\ ha_1 - 1 = 0, \end{cases}$$

откуда $a_1 = \frac{1}{h}$, $a_0 = -\frac{1}{h}$.

Таким образом,

$$L_h u := u_x = \frac{u(x+h) - u(x)}{h} \tag{10.2}$$

Аппроксимация (10.2) носит название *правой разностной производной*. При этом

$$\psi(x) = \frac{h^2}{2} a_1 u''(x) + \dots = \frac{h^2}{2} \cdot \frac{1}{h} u''(x) + \dots = \frac{h}{2} u''(x) + \dots = O(h),$$

т.е. правая разностная производная аппроксимирует исходный дифференциальный оператор первой производной с первым порядком.



Левая разностная производная. Выбрав в качестве шаблона $\mathbb{W}(x)$ множество узлов $\mathbb{W}(x) = \{x - h, x\}$, точно так же легко получить следующую аппроксимацию:

$$L_h u := u_{\bar{x}} = \frac{u(x) - u(x - h)}{h} \quad (10.3)$$

Формула (10.3) определяет *левую разностную производную*, причем

$$\psi(x) = u_{\bar{x}}(x) - u'(x) = O(h).$$

Центральная разностная производная. На шаблоне из трех точек ($\mathbb{W}(x) = \{x - h, x, x + h\}$) можно получить (если не требовать максимального порядка аппроксимации) однопараметрическое семейство разностных операторов

$$L_h^{(\sigma)} u = \sigma u_x + (1 - \sigma) u_{\bar{x}}, \quad (10.4)$$

где σ — любое вещественное число. При этом

$$\psi(x) = L_h^{(\sigma)} u(x) - u(x) = (2\sigma - 1) \frac{h}{2} u''(x) + O(h^2).$$

Отсюда следует, что при любом $\sigma \neq \frac{1}{2}$ разностный оператор $L_h^{(\sigma)}$ и имеет первый порядок аппроксимации. В то же время при $\sigma = \frac{1}{2}$, как легко видеть, *погрешность аппроксимации* становится величиной второго порядка. Сам разностный оператор в этом случае принимает вид

$$L_h^{(0.5)} u = \frac{u_x + u_{\bar{x}}}{2} := u_{\circlearrowright} = \frac{u(x + h) - u(x - h)}{2h} \quad (10.5)$$

и называется *центральной разностной производной*, а его погрешность —

$$\psi(x) = u_{\circlearrowright}(x) - u'(x) = \frac{h^2}{6} u'''(x) + \dots = O(h^2).$$





Способ численного дифференцирования

Формальная схема данного способа выглядит следующим образом: выбрав шаблон $\mathbb{W}(x)$, заменяем на этом шаблоне функцию $u(x)$ **интерполяционным многочленом** (или в общем случае интерполяционной функцией заданного вида):

$$u(x) = P(x) + r(x).$$

После этого применим к последнему равенству дифференциальный оператор L , аппроксимацию которого мы ищем:

$$Lu(x) = LP(x) + Lr(x) \quad (10.6)$$

Заметим, что равенство (10.6) справедливо для всех значений x , а не только в узлах сетки. Поэтому, рассмотрев его в интересующем нас узле $x \in \omega_h$, получим:

$$\begin{aligned} L_h u(x) &= LP(x), \\ \psi(x) &= -Lr(x). \end{aligned} \quad (10.7)$$

Получим описанным способом аппроксимацию [примера 10.1](#).

Правая разностная производная. Здесь $\mathbb{W}(x_i) = \{x_i, x_i + h\}$. Тогда

$$P(x) = P_1(x) = u(x_i) + (x - x_i) u(x_i, x_i + h);$$

$$P'_1(x) = u(x_i, x_i + h) = \frac{u(x_i + h) - u(x_i)}{h} = u_{x,i}.$$

Так как

$$r(x) = \frac{\omega_2(x) u''(\xi)}{2!},$$

то

$$\psi(x_i) = -r'(x_i) = -\frac{1}{2!} \left[\omega'_2(x) u''(\xi) + \omega_2(x) \frac{d}{dx} u''(\xi) \right] \Big|_{x=x_i} = -\frac{1}{2!} \omega'_2(x_i) u''(\xi) = \frac{h}{2} u''(\xi),$$

т.е. получили результаты, полностью согласующиеся с [предыдущими](#).

Пример 10.2. Аппроксимация дифференциального оператора

$$Lu(x) = u''(x) = \frac{d^2u(x)}{dx^2}.$$

Решение. Здесь уже минимально необходимое количество узлов шаблона равно трем. Выбрав в качестве такого Ш(x) = { $x - h$, x , $x + h$ }, любым из описанных выше способов построим разностный оператор

$$L_h u(x) := u_{\bar{x}x} = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2}, \quad (10.8)$$

который носит название *второй разностной производной*. При этом *погрешность аппроксимации* имеет вид

$$\psi(x) = \frac{h^2}{12} u^{IV}(x) + O(h^4) = O(h^2) \quad (10.9)$$

т.е., является величиной второго порядка, а не первого, как следовало бы ожидать. Этот факт объясняется совпадением точки аппроксимации с центром симметрии шаблона. \square

Пример 10.3. Аппроксимация дифференциального оператора

$$Lu(x) = u^{IV}(x) = \frac{d^4u(x)}{dx^4}.$$

Решение. Выбрав минимально необходимый шаблон из пяти точек вида (вновь выбираем множество узлов максимально симметричным) Ш(x) = { $x - 2h$, $x - h$, x , $x + h$, $x + 2h$ }, найдем следующий разностный оператор:

$$L_h u(x) := u_{\bar{x}x\bar{x}x} = \frac{u(x+2h) - 4u(x+h) + 6u(x) - 4u(x-h) + u(x-2h)}{h^4} \quad (10.10)$$

Полученный разностный оператор называется *четвертой разностной производной*.

Проводя соответствующие разложения в ряд Тейлора, можем записать

$$u_{\bar{x}x\bar{x}x} = u^{IV}(x) + \frac{h^2}{6} u^{VI}(x) + O(h^4),$$



откуда

$$\psi(x) = \frac{h^2}{6} u^{VI}(x) + O(h^4) = O(h^2),$$

т.е. четвертая разностная производная также имеет второй порядок аппроксимации. \square

Замечание 10.1. Символы, используемые для обозначения второй, четвертой и т.д. разностных производных, в действительности не просто являются обозначениями, но и предписывают применение в указанном порядке и указанное число раз операторов правой и левой разностных производных. В самом деле, например,

$$\begin{aligned} u_{\bar{x}x}(x) = (u_{\bar{x}}(x))_x &= \frac{u_{\bar{x}}(x+h) - u_{\bar{x}}(x)}{h} = \frac{\frac{u(x+h) - u(x)}{h} - \frac{u(x) - u(x-h)}{h}}{h} = \\ &= \frac{u(x+h) - 2u(x) + u(x-h)}{h^2}. \end{aligned}$$

Разложение погрешности аппроксимации в ряд по степеням h в принципе можно использовать для повышения порядка аппроксимации. Действительно, из (10.9) имеем:

$$u_{\bar{x}x} - u'' = \frac{h^2}{12} u^{IV} + O(h^4) = [\text{используем формулу (10.10)}] =$$

$$= \frac{h^2}{12} [u_{\bar{x}x\bar{x}x} + O(h^2)] + O(h^4) = \frac{h^2}{12} u_{\bar{x}x\bar{x}x} + O(h^4).$$

Отсюда следует, что оператор $L_h u = u_{\bar{x}x} - \frac{h^2}{12} u_{\bar{x}x\bar{x}x}$, определенный на пятиточечном шаблоне $\mathbb{W}(x) = \{x-2h, x-h, x, x+h, x+2h\}$, аппроксимирует оператор $L u = u''$ с четвертым порядком.

Принципиально такой процесс повышения порядка аппроксимации можно продолжить и дальше и получить любой порядок аппроксимации в классе достаточно гладких функций. При этом количество узлов шаблона, естественно, возрастает. Однако указанный прием повышения порядка аппроксимации не всегда можно рекомендовать для практического применения, так как качество получающихся при этом разностных операторов ухудшается (увеличивается объем работы, могут возникнуть проблемы с устойчивостью и т.п.).

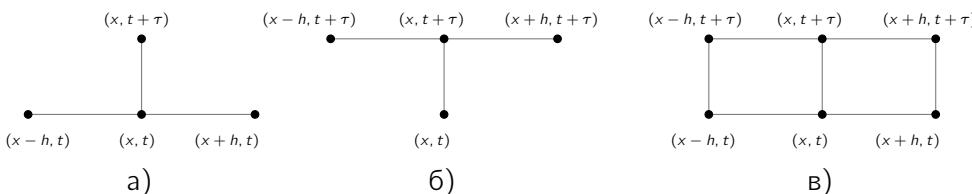


Пример 10.4. Аппроксимация дифференциального оператора

$$Lu = \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2}, \quad u = u(x, t).$$

Решение. Прежде чем приступить к построению соответствующих разностных операторов, заметим, что, зная аппроксимации типа (10.2), (10.3) для первой производной и (10.8) для второй, можно записать разностные аппроксимации большинства дифференциальных операторов, использующихся в приложениях (по крайней мере, простейшие).

Сконструируем теперь шаблон для аппроксимации интересующего нас дифференциального оператора Lu . Ранее мы видели, что для аппроксимации оператора первой производной необходимо, как минимум, две точки, а второй — три. Поэтому наш шаблон в простейшем случае может иметь один из следующих видов:



Используя шаблон а), можем, очевидно, записать такой разностный оператор:

$$L_{ht}^{(0)} u = \frac{u(x, t+\tau) - u(x, t)}{\tau} - \frac{u(x+h, t) - 2u(x, t) + u(x-h, t)}{h^2} \quad (10.11)$$

Для сокращения записи в дальнейшем будем использовать следующие обозначения:

$$u = u(x, t), \quad \hat{u} = u(x, t+\tau), \quad \check{u} = u(x, t-\tau).$$

В этих обозначениях формула (10.11) может быть переписана в виде

$$L_{ht}^{(0)} u = u_t - u_{xx}. \quad (10.12)$$



Используя шаблон б), аналогично можем записать

$$L_{h\tau}^{(1)} u = u_t - \hat{u}_{\bar{x}x}. \quad (10.13)$$

Взяв линейную комбинацию операторов (10.12) и (10.13), получим однопараметрическое семейство разностных операторов

$$L_{h\tau}^{(\sigma)} u = u_t - (\sigma \hat{u}_{\bar{x}x} + (1 - \sigma) u_{\bar{x}x}), \quad (10.14)$$

определенных при $\sigma \neq 0$ и $\sigma \neq 1$ на шеститочечном шаблоне в) (случай $\sigma = 0$ дает разностный оператор (10.12), а $\sigma = 1$ — (10.13)).

Для оценки порядка разностной аппроксимации воспользуемся формулами

$$u_t = \frac{\partial u(x, t)}{\partial t} + \frac{\tau}{2} \frac{\partial^2 u(x, t)}{\partial t^2} + O(\tau^2) = \frac{\partial u(x, t + \tau)}{\partial t} - \frac{\tau}{2} \frac{\partial^2 u(x, t + \tau)}{\partial t^2} + O(\tau^2) = \frac{\partial u\left(x, t + \frac{\tau}{2}\right)}{\partial t} + O(\tau^2),$$

$$u_{\bar{x}x} = \frac{\partial^2 u(x, t)}{\partial x^2} + \frac{h^2}{12} \frac{\partial^4 u(x, t)}{\partial x^4} + O(h^4) = \frac{\partial^2 u\left(x, t + \frac{\tau}{2}\right)}{\partial x^2} - \frac{\tau}{2} \frac{\partial^3 u\left(x, t + \frac{\tau}{2}\right)}{\partial x^2 \partial t} + O(\tau^2 + h^2),$$

$$\hat{u}_{\bar{x}x} = \frac{\partial^2 u(x, t + \tau)}{\partial x^2} + \frac{h^2}{12} \frac{\partial^4 u(x, t + \tau)}{\partial x^4} + O(h^4) = \frac{\partial^2 u\left(x, t + \frac{\tau}{2}\right)}{\partial x^2} + \frac{\tau}{2} \frac{\partial^3 u\left(x, t + \frac{\tau}{2}\right)}{\partial x^2 \partial t} + O(\tau^2 + h^2).$$

Подставляя эти разложения в (10.12), (10.13) и (10.14), получим:

$$L_{h\tau}^{(0)} u = \frac{\partial u(x, t)}{\partial t} + \frac{\tau}{2} \frac{\partial^2 u(x, t)}{\partial t^2} + O(\tau^2) - \frac{\partial^2 u(x, t)}{\partial x^2} - \frac{h^2}{12} \frac{\partial^4 u(x, t)}{\partial x^4} + O(h^4) = L u(x, t) + O(\tau + h^2);$$

$$L_{h\tau}^{(1)} u = \frac{\partial u(x, t + \tau)}{\partial t} - \frac{\tau}{2} \frac{\partial^2 u(x, t + \tau)}{\partial t^2} + O(\tau^2) - \frac{\partial^2 u(x, t + \tau)}{\partial x^2} - \frac{h^2}{12} \frac{\partial^4 u(x, t + \tau)}{\partial x^4} + O(h^4) = \\ = L u(x, t + \tau) + O(\tau + h^2);$$



Верх

Назад

Вперёд

Пред.

След.

Указатель

Помощь

Экран

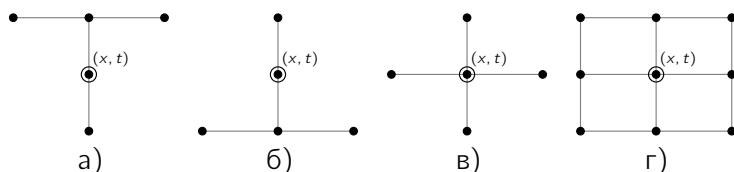
$$\begin{aligned}
 L_{ht}^{(\sigma)} &= \frac{\partial u\left(x, t + \frac{\tau}{2}\right)}{\partial t} + O(\tau^2) - \sigma \left(\frac{\partial^2 u\left(x, t + \frac{\tau}{2}\right)}{\partial x^2} + \frac{\tau}{2} \frac{\partial^3 u\left(x, t + \frac{\tau}{2}\right)}{\partial x^2 \partial t} + O(\tau^2 + h^2) \right) - \\
 &- (1 - \sigma) \left(\frac{\partial^2 u\left(x, t + \frac{\tau}{2}\right)}{\partial x^2} - \frac{\tau}{2} \frac{\partial^3 u\left(x, t + \frac{\tau}{2}\right)}{\partial x^2 \partial t} + O(\tau^2 + h^2) \right) = \frac{\partial u\left(x, t + \frac{\tau}{2}\right)}{\partial t} - \frac{\partial^2 u\left(x, t + \frac{\tau}{2}\right)}{\partial x^2} + \\
 &+ (1 - 2\sigma) \frac{\tau}{2} \frac{\partial^3 u\left(x, t + \frac{\tau}{2}\right)}{\partial x^2 \partial t} + O(\tau^2 + h^2) = Lu\left(x, t + \frac{\tau}{2}\right) + (1 - 2\sigma) \frac{\tau}{2} \frac{\partial^3 u\left(x, t + \frac{\tau}{2}\right)}{\partial x^2 \partial t} + O(\tau^2 + h^2).
 \end{aligned}$$

Таким образом, оператор $L_{ht}^{(\sigma)}$ аппроксимирует оператор L со вторым порядком по h при любом значении параметра σ , с первым порядком по τ при $\sigma \neq 0.5$ (в том числе при $\sigma = 0$ и при $\sigma = 1$) и со вторым порядком по τ при $\sigma = 0.5$. \square

Пример 10.5. Аппроксимация дифференциального оператора

$$Lu = \frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2}.$$

Решение. В этом случае при конструировании шаблона необходимо учесть, что в операторе L присутствует вторая производная по t . Таким образом, минимально возможными будут следующие конфигурации шаблона:





Соответствующие разностные операторы будут иметь вид:

$$L_{ht}^{(1,0)} u = u_{\bar{t}t} - \hat{u}_{\bar{x}x} \quad \text{на шаблоне а), здесь } u_{\bar{t}t} = \frac{\hat{u} - 2u + \check{u}}{\tau^2}; \quad (10.15)$$

$$L_{ht}^{(0,0)} u = u_{\bar{t}t} - u_{\bar{x}x} \quad \text{на шаблоне б);} \quad (10.16)$$

$$L_{ht}^{(0,1)} u = u_{\bar{t}t} - \check{u}_{\bar{x}x} \quad \text{на шаблоне в).} \quad (10.17)$$

На девятиточечном шаблоне г) можно записать (по аналогии с [примером 10.4](#), комбинируя разностные операторы (10.15) — (10.17)) двухпараметрическое семейство разностных операторов

$$L_{ht}^{(\sigma_1, \sigma_2)} u = u_{\bar{t}t} - (\sigma_1 \hat{u}_{\bar{x}x} + (1 - \sigma_1 - \sigma_2) u_{\bar{x}x} + \sigma_2 \check{u}_{\bar{x}x}) \quad (10.18)$$

частными случаями которого являются операторы (10.15) — (10.17) (соответствующие значения σ_1 и σ_2 использованы в их обозначениях).

С помощью разложений, аналогичных использованным в [примере 10.5](#), несложно показать, что оператор (10.16) имеет погрешность аппроксимации $O(\tau^2 + h^2)$. Этот же порядок имеет и оператор (10.18) при $\sigma_1 = \sigma_2 = \sigma$, где σ — любое число. \square

Следует заметить, что параметры σ_1 и σ_2 , так же, как и параметр σ в операторе (10.14), управляют не только порядком аппроксимации, но и устойчивостью соответствующей разностной схемы.

Пример 10.6. Аппроксимация дифференциального оператора

$$Lu = u''.$$

Решение. Вновь вернемся к дифференциальному оператору примера 2, но на сей раз рассмотрим сетку \hat{w}_h и, следовательно, для построения разностной аппроксимации зададим *нерегулярный* трехточечный шаблон $\mathbb{W}(x) = \{x - h_-, x, x + h_+\}$. Наряду с использованными нами ранее обозначениями для правой и левой разностных производных, которые на данном шаблоне запишутся следующим образом:

$$u_{\bar{x}} = \frac{u(x) - u(x - h_-)}{h_-}, \quad u_x = \frac{u(x + h_+) - u(x)}{h_+},$$



введем еще и такие:

$$\hbar = \frac{h_+ + h_-}{2} \quad \text{и} \quad u_{\bar{x}} = \frac{u(x + h_+) - u(x)}{\hbar}.$$

Тогда для разностной аппроксимации рассматриваемого дифференциального оператора любым из описанных выше способов может быть получено следующее выражение

$$L_h u = \frac{1}{\hbar} \left[\frac{u(x + h_+) - u(x)}{h_+} - \frac{u(x) - u(x - h_-)}{h_-} \right] = \frac{u_x - u_{\bar{x}}}{\hbar} =: u_{\bar{x}\bar{x}} \quad (10.19)$$

Заметим, что при $h_+ = h_- = h$ оператор (10.19) совпадает с (10.8). Так как

$$u_x = u'(x) + \frac{h_+}{2} u''(x) + \frac{h_+^2}{6} u'''(x) + O(h_+^3),$$

$$u_{\bar{x}} = u'(x) - \frac{h_-}{2} u''(x) + \frac{h_-^2}{6} u'''(x) + O(h_-^3),$$

то

$$L_h u = u''(x) + \frac{h_+^2 - h_-^2}{6\hbar} u'''(x) + O(\hbar^2) = u''(x) + \frac{h_+ - h_-}{3} u'''(x) + O(\hbar^2).$$

Таким образом, разностный оператор (10.19) при $h_+ \neq h_-$ имеет первый порядок аппроксимации. □



10.1.4. Погрешность аппроксимации на сетке

До сих пор мы рассматривали локальную разностную аппроксимацию (аппроксимацию в точке). Обычно же требуется оценка порядка разностной аппроксимации на всей сетке (заметим: эти порядки *могут не совпадать*).

Пусть ω_h — сетка в некоторой области G p -мерного пространства, H_h — линейное пространство сеточных функций, заданных на ω_h , H_0 — пространство гладких функций $u(x)$, $\|\cdot\|_0$ — норма в H_0 , $\|\cdot\|_h$ — норма в H_h . Как и ранее, будем предполагать, что:

- 1) существует оператор проектирования $P_h : P_h u = u_h \in H_h$ для всех функций $u \in H_0$;
- 2) нормы $\|\cdot\|_h$ и $\|\cdot\|_0$ согласованы, т.е. $\lim_{|h| \rightarrow 0} \|u_h\|_h = \|u\|_0$.

Рассмотрим некоторый оператор $L : H_0 \rightarrow H_0$ и оператор $L_h : H_h \rightarrow H_h$. Назовем *погрешностью аппроксимации дифференциального оператора* L разностным оператором L_h сеточную функцию

$$\psi_h = L_h u_h - (Lu)_h,$$

где $u_h = P_h u$, $(Lu)_h = P_h(Lu)$, а u — произвольный элемент из H_0 .

Если $\|\psi_h\|_h \xrightarrow{|h| \rightarrow 0} 0$, то будем говорить, что разностный оператор L_h аппроксимирует дифференциальный оператор L на сетке ω_h .

Если

$$\|\psi_h\|_h = \|L_h u_h - (Lu)_h\|_h = O(|h|^m) \quad (10.20)$$

или, что то же самое,

$$\|L_h u_h - (Lu)_h\|_h \leq M |h|^m,$$

где M — не зависящая от $|h|$ константа, а $m > 0$ то будем говорить, что разностный оператор L_h *аппроксимирует дифференциальный оператор L на сетке ω_h с порядком m*.

Под символом $|h|$ здесь понимается следующее:



а) если $h = (h_1, \dots, h_p)$ (в случае p -мерного пространства), то, например, $|h| = \sqrt{h_1^2 + \dots + h_p^2}$. При этом может оказаться, что аппроксимации по каждому из h_α , $\alpha = 1, \dots, p$ различны по порядку. Тогда вместо (10.20) можно записать неравенство

$$\|L_h u_h - (Lu)_h\|_h \leq M \sum_{\alpha=1}^p h_\alpha^{m_\alpha},$$

где $m_\alpha > 0$ — порядок аппроксимации по α -й компоненте. Если теперь положить $m = \min \{m_1, m_2, \dots, m_p\}$, то получим оценку (10.20).

б) если сетка ω_h — одномерная и неравномерная, т.е. $h = (h_1, \dots, h_N)$, где N — число узлов, то, например, $|h| = \max_{1 \leq i \leq N} h_i$, или так же, как в предыдущем случае.

Рассмотрим примеры.

Пример 10.7. Разностная аппроксимация на неравномерной сетке ([пример 10.6](#)).

Решение. Здесь

$$Lu = \frac{d^2u}{dx^2}; \quad u \in H_0 = C^4[0, 1]; \quad \hat{\omega}_h = \{x_i, \quad i = \overline{0, N}; \quad x_0 = 0, \quad x_N = 1\}; \quad (L_h u)_i = u_{\bar{x}\bar{x}, i},$$

причем

$$\psi_i = \frac{h_{i+1}}{h_i} u''_i + O(\hbar_i^2), \quad i = 1, 2, \dots, N-1.$$

Отсюда видно, что оператор $L_h u$ имеет в сеточной норме C первый порядок аппроксимации:

$$\|\psi_h\|_C = \max_{1 \leq i \leq N-1} |\psi_i| = O(h), \quad h = \max_{1 \leq i \leq N} h_i.$$

В сеточной L_2 -норме также получим первый порядок аппроксимации:

$$\|\psi_h\|_{L_2} = \left(\sum_{i=1}^{N-1} \hbar_i \psi_i^2 \right)^{\frac{1}{2}} = O(h), \quad h = \max_{1 \leq i \leq N} h_i.$$



Однако в норме

$$\|\psi_h\|_{(-1)} = \left[\sum_{i=1}^{N-1} h_i \left(\sum_{k=1}^i \hbar_k \psi_k \right)^2 \right]^{\frac{1}{2}} \quad (10.21)$$

ψ_h имеет второй порядок, т.е. $\|\psi_h\|_{(-1)} = O(h^2)$, где $h = \max_{1 \leq i \leq N} h_i$. Действительно,

$$\psi_i = \frac{h_{i+1}^2 - h_i^2}{6\hbar_i} u'''_i + O(\hbar_i^2)$$

и, так как $u'''_i = u'''_{i+1} + O(h_{i+1})$, то

$$\psi_i = \frac{h_{i+1}^2 u'''_{i+1} - h_i^2 u'''_i}{6\hbar_i} + \overset{\circ}{\psi}_i = \overset{\circ}{\psi}_i + \psi_i^*,$$

где $\psi_i^* = O(h^2)$ в любой норме. Главный же член разложения имеет так называемый *дивергентный* вид. Поэтому

$$S_i = \sum_{k=1}^i \hbar_k \overset{\circ}{\psi}_k = \frac{1}{6} \sum_{k=1}^i (h_{k+1}^2 u'''_{k+1} - h_k^2 u'''_k) = \frac{h_{i+1}^2 u'''_{i+1} - h_i^2 u'''_i}{6},$$

т.е. $|S| \leq Mh^2$.

Следовательно,

$$\|\overset{\circ}{\psi}_h\|_{(-1)} = \left(\sum_{i=1}^{N-1} h_i S_i^2 \right)^{\frac{1}{2}} = O(h^2),$$

а так как

$$\|\psi_h\| \leq \|\overset{\circ}{\psi}_h\|_{(-1)} + \|\psi_h^*\|_{(-1)} = O(h^2),$$

то $\|\psi_h\|_{(-1)} = O(h^2)$. □



Этот пример показывает, что исследование [локальной аппроксимации](#) может быть недостаточным для суждения о качестве разностного оператора. Выбор же подходящей нормы всякий раз должен быть предметом изучения, поскольку связан со структурой исходного дифференциального оператора.

Если ищется решение нестационарного уравнения (например, теплопроводности или колебаний), то переменная t выделяется по физическому смыслу (время). Поэтому говорят о порядке аппроксимации отдельно по временной переменной t и отдельно по пространственной (пространственным) переменной x . При этом чаще всего используют нормы

$$\|y\|_{h\tau} = \max_{t \in \omega_\tau} \|y(t)\|_h \quad (10.22)$$

или

$$\|y\|_{h\tau} = \left[\sum_{t \in \omega_\tau} \tau \|y(t)\|_h^2 \right]^{\frac{1}{2}} \quad (10.23)$$

Пример 10.8. Аппроксимация оператора

$$Lu = \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2}, \quad L_{h\tau}u = u_t - u_{xx}.$$

Решение. Тогда в случае, если $u(x, t)$ имеет четыре непрерывных производных по x и две — по t , то (см. [пример 10.4](#))

$$\psi_{h\tau}(x, t) = L_{h\tau}u - Lu = O(\tau + h^2).$$

Отсюда следует, что $L_{h\tau}$ аппроксимирует исходный оператор L со вторым порядком по x и первым по t в любой из норм, рассмотренных выше. □

10.1.5. Постановка разностной задачи

До сих пор мы занимались приближенной заменой дифференциальных операторов разностными. Однако дифференциальные задачи в целом помимо собственно дифференциальных уравнений включают в себя еще и дополнительные условия, которые и обеспечивают выделение из всей совокупности возможных решений единственного. Поэтому при формулировке разностной задачи, помимо аппроксимации дифференциального уравнения, необходимо описывать в разностном виде еще и эти дополнительные условия. Совокупность разностных уравнений, аппроксимирующих дифференциальное уравнение и дополнительные условия, называют *разностной схемой*.

При этом используется две формы записи разностных схем (безындексная и индексная), которыми необходимо научиться пользоваться. Вновь обратимся к примерам.

Пример 10.9. Задача Коши для обыкновенного дифференциального уравнения первого порядка:

$$\begin{cases} u'(t) = f(t), & t > 0, \\ u(0) = u_0. \end{cases} \quad (10.24)$$

Решение. Выберем равномерную сетку $\omega_\tau = \{t_j = j\tau; \tau > 0; j = 0, 1, \dots\}$. Тогда в соответствие дифференциальной задаче можно поставить разностную схему, которая в безындексной форме имеет вид

$$\begin{cases} y_t = \varphi, \\ y(0) = u_0 \end{cases} \quad (10.25)$$

а в индексной

$$\begin{cases} \frac{y^{j+1} - y^j}{\tau} = \varphi^j, & j = 0, 1, \dots \\ y^0 = u_0. \end{cases} \quad (10.26)$$

При этом правую часть φ можно задавать различными способами, лишь бы выполнялось условие $\varphi - f = O(\tau)$, например, $\varphi(t) = f(t)$ или $\varphi(t) = \frac{f(t) + f(t+\tau)}{2}$ при $t \in \omega_\tau$, что в индексной форме выглядит соответственно $\varphi^j = f(t_j)$ или $\varphi^j = \frac{f(t_j) + f(t_{j+1})}{2}$ при $j = 0, 1, \dots$



Для нахождения решения, как это следует из (10.26), получаем рекуррентную формулу

$$\begin{cases} y^{j+1} = y^j + \tau \varphi^j, & j = 0, 1, \dots \\ y^0 = u_0. \end{cases} \quad (10.27)$$

□

Пример 10.10. Задача Коши для системы обыкновенных дифференциальных уравнений первого порядка.

$$\begin{cases} \frac{du(t)}{dt} + Au(t) = 0, & t > 0, \\ u(0) = u_0, \end{cases} \quad (10.28)$$

где A — $n \times n$ -матрица, $u = (u_1, \dots, u_n)^T$.

Решение. Выберем сетку так же, как и в [примере 10.9](#). Тогда разностная схема может иметь вид ([явная схема Эйлера](#))

$$\begin{cases} y_t + Ay = 0, \\ y(0) = u_0 \end{cases} \quad \text{или} \quad \begin{cases} \frac{y^{j+1} - y^j}{\tau} + Ay^j = 0, & j = 0, 1, \dots \\ y^0 = u_0 \end{cases} \quad (10.29)$$

Здесь $y^j = (y_1^j, \dots, y_n^j)^T$.

Решения данной разностной задачи могут быть найдены по рекуррентным формулам типа (10.27):

$$\begin{cases} y^{j+1} = y^j - \tau Ay^j, & j = 0, 1, \dots \\ y^0 = u_0. \end{cases}$$

□

Пример 10.11. Краевая задача для обыкновенного дифференциального уравнения второго порядка:

$$\begin{cases} u''(x) = -f(x), & 0 < x < 1, \\ u(0) = \mu_0, \\ u(1) = \mu_1. \end{cases} \quad (10.30)$$



Решение. Вновь выберем равномерную сетку $\bar{\omega}_h$. Тогда разностная схема может иметь вид

$$\begin{cases} y_{\bar{x}x} = -\varphi, \quad x \in \omega_h, \\ y(0) = \mu_0, \\ y(1) = \mu_1 \end{cases} \quad \text{или} \quad \begin{cases} \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} = -\varphi_i, \quad i = 1, \dots, N-1, \\ y_0 = \mu_0, \\ y_N = \mu_1. \end{cases} \quad (10.31)$$

Сеточная функция φ , как и в [примере 10.9](#), может быть выбрана лишь исходя из условия $\varphi - f = O(h^2)$. Решение задачи (10.31) может быть найдено (напомним: это — система линейных алгебраических уравнений (!) с трехдиагональной матрицей) с помощью [метода разностной прогонки](#). \square

Пример 10.12. Первая краевая задача для уравнения теплопроводности:

$$\begin{cases} Lu = \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = f(x, t), \quad 0 < x < 1, \quad 0 < t \leq T, \\ u(x, 0) = u_0(x), \\ u(0, t) = \mu_0(t), \\ u(1, t) = \mu_1(t). \end{cases} \quad (10.32)$$

Решение. Выбрав равномерную сетку $\bar{\omega}_{h\tau} = \bar{\omega}_h \times \bar{\omega}_\tau$ и простейший четырехточечный шаблон (см. [пример 10.4 а\)](#)), получим разностную схему

$$\begin{cases} y_t = y_{\bar{x}x} + \varphi, \quad (x, t) \in \omega_{h\tau}, \\ y(x, 0) = u_0(x), \quad x \in \bar{\omega}_h, \\ y(0, t) = \mu_0(t), \quad t \in \bar{\omega}_\tau, \\ y(1, t) = \mu_1(t), \quad t \in \bar{\omega}_\tau \end{cases} \quad (10.33)$$

или в индексной форме

$$\begin{cases} \frac{y_i^{j+1} - y_i^j}{\tau} = \frac{y_{i+1}^j - 2y_i^j + y_{i-1}^j}{h^2} + \varphi_i^j, \quad i = \overline{1, N_1 - 1}; \quad j = \overline{0, N_2 - 1}; \\ y_i^0 = u_0(x_i), \quad i = \overline{0, N_1}, \\ y_0^j = \mu_0(t_j), \quad j = \overline{0, N_2}, \\ y_{N_1}^j = \mu_1(t_j), \quad j = \overline{0, N_2}. \end{cases}$$



Функция φ здесь снова выбирается из условия, аналогичного предыдущим: $\varphi - f = O(\tau + h^2)$, например, $\varphi_i^j = f(x_i, t_j)$ или $\varphi_i^j = f\left(x_i, t_{j+\frac{1}{2}}\right)$ и т.п.

Схема (10.33) — пример *явной разностной схемы*: значения решения на верхнем временном слое y^{j+1} определяются через значения решения на предыдущем слое по явным рекуррентным формулам:

$$y_i^{j+1} = y_i^j + \tau \left(y_{\bar{x}, i}^j + \varphi_i^j \right), \quad i = \overline{1, N_1 - 1}; \quad j = \overline{0, N_2 - 1}.$$

Значения y_i^0 , y_0^j и $y_{N_1}^j$ при этом известны из начального и граничных условий.

Аналогичным образом, выбрав для аппроксимации дифференциального оператора L в (10.32) шаблон примера 10.4 б), можем записать *неявную разностную схему*:

$$\begin{cases} y_t = \hat{y}_{\bar{x}} + \varphi, & (x, t) \in \omega_{h\tau}, \\ y(x, 0) = u_0(x), & x \in \bar{\omega}_h, \\ y(0, t) = \mu_0(t), & t \in \bar{\omega}_\tau, \\ y(1, t) = \mu_1(t), & t \in \bar{\omega}_\tau \end{cases} \quad (10.34)$$

или в индексной форме

$$\begin{cases} \frac{y_i^{j+1} - y_i^j}{\tau} = \frac{y_{i+1}^{j+1} - 2y_i^{j+1} + y_{i-1}^{j+1}}{h^2} + \varphi_i^j, & i = \overline{1, N_1 - 1}; \quad j = \overline{0, N_2 - 1}; \\ y_i^0 = u_0(x_i), & i = \overline{0, N_1}, \\ y_0^j = \mu_0(t_j), & j = \overline{0, N_2}, \\ y_{N_1}^j = \mu_1(t_j), & j = \overline{0, N_2}. \end{cases}$$

Значения решения на верхнем временном слое y^{j+1} вновь определяются через значения решения на предыдущем слое, но на сей раз путем решения соответствующей системы линейных алгебраических уравнений с трехдиагональной матрицей (например, как в примере 10.11, с помощью *метода разностной прогонки*).





10.1.6. Сходимость и точность разностных схем

Итак, пусть дифференциальной задаче

$$Lu = f(x), \quad x \in G, \quad (10.35a)$$

$$Iu = \mu(x), \quad x \in \Gamma \quad (10.35b)$$

на сетке $\omega_h + \gamma_h$ поставлена в соответствие разностная схема

$$L_h y_h = \varphi, \quad x \in \omega_h, \quad (10.36a)$$

$$I_h y_h = \chi_h(x), \quad x \in \gamma_h. \quad (10.36b)$$

Основной целью всякого приближенного метода является получение решения исходной непрерывной задачи с заданной точностью $\varepsilon > 0$ за конечное число действий $N(\varepsilon)$. Чтобы выяснить принципиальную возможность приближения решения u задачи (10.35) решением y_h задачи (10.36), сравним y_h и $u(x)$. Это сравнение, как обычно, будем проводить в пространстве H_h .

Пусть u_h — значение функции $u(x)$ на сетке $\bar{\omega}_h$. Рассмотрим погрешность разностной схемы (10.36): $z_h = y_h - u_h$ и выпишем задачу для ее определения. Подставляя $y_h = z_h + u_h$ в разностные уравнения (10.36a) и (10.36b), получим:

$$\begin{cases} L_h z_h + L_h u_h = \varphi_h, \\ I_h z_h + I_h u_h = \chi_h \end{cases}$$

или

$$\begin{cases} L_h z_h = \varphi_h - L_h u_h := \psi_h, \quad x \in \omega_h, \\ I_h z_h = \chi_h - I_h u_h := \nu_h, \quad x \in \gamma_h. \end{cases} \quad (10.37)$$

Правые части задачи (10.37) (ψ_h и ν_h) называются соответственно погрешностью аппроксимации уравнения (10.35a) разностным уравнением (10.36a) и граничных условий (10.35b) — разностными граничными условиями (10.36b) на решении дифференциальной задачи (10.35).

Для оценки погрешности схемы z_h и погрешности аппроксимации ψ_h и ν_h введем на множестве сеточных функций нормы $\|\cdot\|_{(1_h)}$, $\|\cdot\|_{(2_h)}$ и $\|\cdot\|_{(3_h)}$ соответственно.



Будем говорить, что решение разностной задачи (10.36) *сходится* к решению задачи (10.35) (или, что то же самое: схема (10.36) *сходится*), если

$$\| z_h \|_{(1_h)} = \| y_h - u_h \|_{(1_h)} \xrightarrow{|h| \rightarrow 0} 0.$$

Разностная схема *сходится со скоростью* $O(|h|^n)$ или *имеет n-й порядок точности*, если при достаточно малом $|h| \leq h_0$ выполняется неравенство

$$\| z_h \|_{(1_h)} = \| y_h - u_h \|_{(1_h)} \leq M |h|^n,$$

где M — константа, не зависящая от h и $n > 0$.

Говорят также, что разностная схема (10.36) *обладает n-м порядком аппроксимации*, если

$$\| \psi_h \|_{(2_h)} = O(|h|^n), \quad \| \nu_h \|_{(3_h)} = O(|h|^n).$$

Обозначая f_h и $(Lu)_h$ значения $f(x)$ и $Lu(x)$ на сетке ω_h и учитывая, что $(f - Lu)_h = 0$, запишем ψ_h в виде

$$\psi_h = \varphi_h - L_h u_h = (u_h - L_h u_h) - (f_h - (Lu)_h) = (\varphi_h - f_h) + ((Lu)_h - L_h u_h) = \psi_h^{(1)} + \psi_h^{(2)} \quad (10.38)$$

Таким образом, погрешность аппроксимации схемы ψ_h складывается из погрешности аппроксимации $\psi_h^{(1)} = \varphi_h - f_h$ правой части и *погрешности аппроксимации* $\psi_h^{(2)} = (Lu)_h - L_h u_h$ дифференциального оператора.

Так как ψ_h есть погрешность аппроксимации в классе решений дифференциального уравнения, то условие $\| \psi_h \|_{(2_h)} = O(|h|^n)$ может быть выполнено, если $\psi_h^{(1)}$ и $\psi_h^{(2)}$ не имеют по отдельности n -го порядка.

Возникает вопрос: как зависит порядок точности схемы от порядка аппроксимации на решении? Прежде чем дать на него ответ, напомним понятие корректной постановки задачи применительно к разностным схемам:

- 1) решение y_h разностной задачи существует и единственно для всех входных данных φ_h из некоторого допустимого семейства;
- 2) решение y_h непрерывно зависит от φ_h , причем эта зависимость равномерна по h .



Второе условие корректности означает, что существует константа $M > 0$, не зависящая от h и такая, что при достаточно малом $|h| \leq h_0$ выполняется неравенство

$$\|y_h - \tilde{y}_h\|_{(1_h)} \leq M \|\varphi_h - \tilde{\varphi}_h\|_{(2_h)} \quad (10.39)$$

где \tilde{y}_h — решение задачи с правой частью $\tilde{\varphi}_h$.

Второе свойство (непрерывной зависимости решения разностной задачи от входных данных), выражаемое неравенством (10.38), называется *устойчивостью (по входным данным)*.

Теперь ответ на поставленный выше вопрос дает следующая

Теорема 10.1 (Лакса). *Если линейная разностная схема устойчива и аппроксимирует исходную дифференциальную задачу, то она сходится, причем порядок точности схемы определяется ее порядком аппроксимации.*

[[Доказательство](#)]



10.1.7. Повышение порядка аппроксимации разностных схем

Как уже отмечалось выше, скорость сходимости разностной схемы, если последняя устойчива, совпадает с порядком ее аппроксимации на решении исходной дифференциальной задачи, причем, как это следует из формулы (Д.17), порядок аппроксимации может быть более высоким, чем порядок аппроксимации дифференциального оператора разностным. Этот факт может быть использован для повышения порядка аппроксимации разностной схемы без увеличения геометрических размеров шаблона. Рассмотрим этот прием на примерах.

Пример 10.13. Рассмотрим задачу [примера 10.9](#):

$$\begin{cases} u'(t) = f(t), \\ u(0) = u_0. \end{cases} \quad (10.40)$$

Решение. Ранее для рассматриваемой задачи была построена разностная схема

$$\begin{cases} y_t = \varphi, & t \in \omega_\tau \\ y(0) = u_0. \end{cases} \quad (10.41)$$

Найдем невязку разностного уравнения на решении $u(t)$ уравнения (10.40):

$$\psi_\tau(t) = u_t(t) - \varphi(t).$$

Так как

$$u(t + \tau) = u(t) + \tau u'(t) + \frac{\tau^2}{2} u''(t) + \frac{\tau^3}{6} u'''(t) + O(\tau^4)$$

и

$$u'(t) = f(t), \quad u''(t) = f'(t), \quad u'''(t) = f''(t), \dots,$$

то

$$\psi_\tau(t) = u'(t) + \frac{\tau}{2} u''(t) + \frac{\tau^2}{6} u'''(t) - \varphi(t) + O(\tau^3).$$

Отсюда видим, что:



- 1) выбирая $\varphi(t) = f(t) + O(\tau)$ (например, $\varphi(t) = f(t)$ (в индексной форме $\varphi^j = f^j$)), получим разностную схему первого порядка аппроксимации;
- 2) если же выбрать $\varphi(t) = f(t) + \frac{\tau}{2}f'(t) + O(\tau^2)$ (например, $\varphi(t) = f(t + \frac{\tau}{2})$ (в индексной форме $\varphi^j = f^{j+\frac{1}{2}}$), или $\varphi(t) = f(t) + \frac{\tau}{2}f_t(t)$ (в индексной форме $\varphi^j = f^j + \frac{\tau}{2}f_t^j$), или $\varphi(t) = \frac{f(t)+f(t+\tau)}{2}$ (в индексной форме $\varphi^j = \frac{f^j+f^{j+1}}{2}$) и т.п.), то разностная схема становится схемой второго порядка;
- 3) аналогично, выбрав $\varphi(t) = f(t) + \frac{\tau}{2}f'(t) + \frac{\tau^2}{6}f''(t) + O(\tau^3)$

(например, $\varphi(t) = f(t) + \frac{\tau}{2}f_t(t) + \frac{\tau^2}{6}f_{tt}(t)$ (в индексной форме $\varphi^j = f^j + \frac{\tau}{2}f_t^j + \frac{\tau^2}{6}f_{tt}^j$) или $\varphi(t) = \frac{5f(t+\tau)+8f(t)-f(t-\tau)}{12}$ (в индексной форме $\varphi^j = \frac{5f^{j+1}+8f^j-f^{j-1}}{12}$)), то получим схему третьего порядка.

□

Пример 10.14. Пусть дифференциальная задача имеет вид

$$\begin{cases} u''(x) - qu(x) = -f(x), & 0 < x < 1, \quad q = \text{const}, \\ u(0) = \mu_0, \\ u(1) = \mu_1. \end{cases} \quad (10.42)$$

Решение. На равномерной сетке $\bar{\omega}_h$ запишем для нее трехточечную разностную схему

$$\begin{cases} y_{\bar{x}x} - dy = -\varphi, & x \in \omega_h, \\ y(0) = \mu_0, \\ y(1) = \mu_1. \end{cases} \quad (10.43)$$

Невязка разностных граничных условий на решении задачи (10.42) здесь, очевидно, равна нулю (ибо последние не содержат производных). Поэтому рассмотрим невязку разностного уравнения (точкой, в которой это делается, мы будем считать произвольный узел сетки ω_h):

$$\psi_h = u_{\bar{x}x} - du + \varphi.$$

Так как

$$u_{\bar{x}x} = u'' + \frac{h^2}{12} u^{IV} + O(h^4)$$

а

$$u'' = qu - f,$$

то

$$\begin{aligned} \psi_h &= u'' + \frac{h^2}{12} u^{IV} - du + \varphi + O(h^4) = qu - f + \frac{h^2}{12} u^{IV} - du + \varphi + O(h^4) = \\ &= (q - d)u + (\varphi - f) + \frac{h^2}{12} u^{IV} + O(h^4). \end{aligned}$$

Отсюда видим, что $\psi_h = O(h^2)$ при $d = q + O(h^2)$, $\varphi = f + O(h^2)$. Продифференцировав дважды равенство $u'' = qu - f$, будем иметь:

$$u^{IV} = qu'' - f'' = q(qu - f) - f''.$$

Тогда невязку можно переписать в виде

$$\begin{aligned} \psi_h &= (q - d)u + (\varphi - f) + \frac{h^2}{12}q(qu - f) - \frac{h^2}{12}f'' + O(h^4) = \\ &= \left[\varphi - f - \frac{h^2}{12}(f'' + qf) \right] + \left[\left(q + \frac{h^2}{12}q^2 \right) - d \right] u + O(h^4). \end{aligned}$$

Таким образом, если положить

$$d = q + \frac{h^2}{12}q^2 + O(h^4); \quad \varphi = f + \frac{h^2}{12}(qf + f'') + O(h^4) = f + \frac{h^2}{12}(qf + f_{\bar{x}x}) + O(h^4),$$

то получим трехточечную разностную схему повышенного (четвертого) порядка аппроксимации на решении исходного дифференциального уравнения (10.42). \square

Пример 10.15. Третья краевая задача для обыкновенного дифференциального уравнения второго порядка

$$\begin{cases} u''(x) - qu(x) = -f(x), \quad 0 < x < 1, \quad q = \text{const}, \\ u'(0) = \sigma_0 u(0) - \mu_0, \\ u(1) = \mu_1. \end{cases} \quad (10.44)$$



Решение. Вновь выбрав равномерную сетку $\bar{\omega}_h$, разностное уравнение запишем в виде

$$y_{\bar{x}x} - dy = -\varphi, \quad (10.45)$$

где $\varphi = f + O(h^2)$, $d = q + O(h^2)$.

Краевое условие при $x = 1$ удовлетворяется точно:

$$y(1) = \mu_1 \quad (10.46)$$

Производную $u'(0)$ заменим правой разностной производной $y_x(0)$ (использование для этих целей левой разностной производной приведет к тому, что одна из используемых точек шаблона (конкретно $x = -h$) выйдет за пределы области, в которой определена задача). Тогда краевое условие при $x = 0$ запишется в виде

$$y_x(0) = \sigma_0 y(0) - \mu_0 (I_h y = \mu_0), \quad (10.47)$$

причем разностный оператор I_h определен на двухточечном шаблоне $\{0, h\}$, что имеет важное значение при реализации разностной схемы (10.45) — (10.47), поскольку для [стандартного варианта метода разностной прогонки](#) граничные условия должны быть не более чем двухточечными.

Выше (см. [пример 10.14](#)) мы видели, что указанный выбор сеточных функций φ и d приводит к тому, что погрешность аппроксимации разностного уравнения (10.45) $\psi_h = O(h^2)$. С другой стороны, так как

$$u(h) = u(0) + hu'(0) + \frac{h^2}{2}u''(0) + O(h^3),$$

то

$$\begin{aligned} \nu_h(0) &= u_x(0) - \sigma_0 u(0) + \mu_0 = u'(0) + \frac{h}{2}u''(0) + O(h^2) - \sigma_0 u(0) + \mu_0 = \\ &= [u'(0) - (\sigma_0 u(0) - \mu_0)] + \frac{h}{2}u''(0) + O(h^2) = \frac{h}{2}u''(0) + O(h^2), \end{aligned}$$

т.е. $\nu_h(0) = O(h)$.

Таким образом, согласно общей схеме рассуждений, порядок аппроксимации разностной схемы (10.45) — (10.47) равен единице.



Подправим разностное граничное условие (10.47) таким образом, чтобы получить второй порядок аппроксимации. Для этого, введя вместо фиксированных коэффициентов σ_0 и μ_0 из граничного условия исходной задачи подлежащие выбору сеточные коэффициенты (параметры) $\bar{\sigma}_0$ и $\bar{\mu}_0$, будем искать его в виде

$$y_x(0) = \bar{\sigma}_0 y(0) - \bar{\mu}_0 \quad (10.48)$$

Проделав выкладки, аналогичные проведенным выше, получим:

$$\nu_h(0) = u'(0) - (\bar{\sigma}_0 u(0) - \bar{\mu}_0) + \frac{h}{2} u''(0) + O(h^2).$$

Из исходного дифференциального уравнения (10.44) (предполагая его выполняющимся и при $x = 0$) найдем: $u''(0) = qu(0) - f(0)$, а из первого граничного условия — $u'(0) = \sigma_0 u(0) - \mu_0$. Тогда

$$\begin{aligned} \nu_h(0) &= \sigma_0 u(0) - \mu_0 - (\bar{\sigma}_0 u(0) - \bar{\mu}_0) + \frac{h}{2} (qu(0) - f(0)) + O(h^2) = \\ &= [\sigma_0 - (\bar{\sigma}_0 - \frac{h}{2}q)] u(0) + [\bar{\mu}_0 - (\mu_0 + \frac{h}{2}f(0))] + O(h^2) = \\ &= [\bar{\mu}_0 - (\mu_0 + \frac{h}{2}f(0))] - [\bar{\sigma}_0 - (\sigma_0 + \frac{h}{2}q)] u(0) + O(h^2). \end{aligned}$$

Отсюда следует, что, выбрав

$$\bar{\sigma}_0 = \sigma + \frac{h}{2}q, \quad \bar{\mu}_0 = \mu_0 + \frac{h}{2}f(0) \quad (10.49)$$

получим разностное граничное условие (10.48) (на том же двухточечном шаблоне) второго порядка аппроксимации. \square

Пример 10.16. Третья краевая задача для уравнения теплопроводности

$$\left\{ \begin{array}{l} \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x, t), \quad 0 < x < 1, \quad 0 < t \leq T, \\ u(x, 0) = u_0(x), \quad 0 \leq x \leq 1, \\ \frac{\partial u(0, t)}{\partial x} = \sigma_0 u(0, t) - \mu_0(t), \quad 0 \leq t \leq T, \\ u(1, t) = \mu_1(t), \quad 0 \leq t \leq T. \end{array} \right. \quad (10.50)$$



Решение. На сетке $\bar{\omega}_{h\tau}$, используя простейший четырехточечный шаблон ([пример 10.4а](#)), напишем явную разностную схему

$$\begin{cases} y_t = y_{\bar{x}x} + \varphi, & (x, t) \in \bar{\omega}_{h\tau} \\ y(x, 0) = u_0(x), & x \in \bar{\omega}_h, \\ y(1, t) = \mu_1(t), & t \in \bar{\omega}_\tau. \end{cases} \quad (10.51)$$

Эта разностная схема при условии $\varphi(x, t) = f(x, t) + O(\tau + h^2)$ аппроксимирует дифференциальную задачу ([10.50](#)) (кроме первого граничного условия) с погрешностью $O(\tau + h^2)$. Однако, заменяя в условии при $x = 0$ производную $\frac{\partial u}{\partial x}$ правой разностной производной (как и в предыдущем примере), получим разностное граничное условие

$$y_x(0, t) = \sigma_0 y(0, t) - \mu_0(t), \quad t \in \bar{\omega}_\tau \quad (10.52)$$

имеющее, как легко видеть, лишь первый порядок аппроксимации по переменной x .

Получим сейчас разностное граничное условие с порядком аппроксимации $O(\tau + h^2)$, не увеличивая числа точек шаблона по x . Как и в [примере 10.15](#) будем искать требуемую аппроксимацию в виде

$$y_x(0, t) = \bar{\sigma}_0 y(0, t) - \bar{\mu}_0(t), \quad t \in \bar{\omega}_\tau \quad (10.53)$$

где $\bar{\sigma}_0$ и $\bar{\mu}_0(t)$ — сеточные функции, подлежащие определению.

Исследуем погрешность аппроксимации:

$$\begin{aligned} \nu(0, t) &= u_x(0, t) - \bar{\sigma}_0 u(0, t) + \bar{\mu}_0(t) = \frac{\partial u(0, t)}{\partial x} + \frac{h}{2} \frac{\partial^2 u(0, t)}{\partial x^2} + O(h^2) - \bar{\sigma}_0 u(0, t) + \bar{\mu}_0(t) = \\ &= \sigma_0 u(0, t) - \mu_0(t) + \frac{h}{2} \frac{\partial^2 u(0, t)}{\partial x^2} + O(h^2) - \bar{\sigma}_0 u(0, t) + \bar{\mu}_0(t) = \left[\frac{\partial^2 u(0, t)}{\partial x^2} = \frac{\partial u(0, t)}{\partial t} - f(0, t) \right] = \\ &= (\sigma_0 - \bar{\sigma}_0) u(0, t) - \mu_0(t) + \bar{\mu}_0(t) + \frac{h}{2} \frac{\partial u(0, t)}{\partial t} - \frac{h}{2} f(0, t) + O(h^2) = (\sigma_0 - \bar{\sigma}_0) u(0, t) + \\ &\quad + (\bar{\mu}_0(t) - \mu_0(t) + \frac{h}{2} u_{\bar{x}}(0, t) - \frac{h}{2} f(0, t)) + O(\tau^2 + h^2). \end{aligned}$$

При проведении записанных выкладок использованы как исходное дифференциальное уравнение (в предположении, что оно выполняется при $x = 0$), так и граничное условие на левом конце отрезка изменения



переменной x . При этом был применен прием изменения направления дифференцирования, который позволил вторую производную по x заменить первой производной по t (последнюю затем заменяют разностным аналогом).

Анализируя полученное выражение, приходим к выводу, что при

$$\bar{\sigma}_0 = \sigma, \quad \bar{\mu}_0(t) = \mu_0(t) + \frac{h}{2}f(0, t) - \frac{h}{2}y_{\bar{t}}(0, t) \quad (10.54)$$

разностное граничное условие имеет погрешность аппроксимации порядка $O(\tau^2 + h^2)$.

Таким образом, в безындексной форме явная разностная схема, аппроксимирующая задачу (10.50) с погрешностью порядка $O(\tau + h^2)$, имеет вид

$$\begin{cases} y_t = y_{\bar{x}x} + \varphi, \quad (x, t) \in \omega_{ht}, \\ y(x, 0) = u_0(x), \quad x \in \bar{\omega}_h, \\ y_x(0, t) = \sigma_0 y(0, t) - \mu_0(t) - \frac{h}{2}f(0, t) + \frac{h}{2}y_{\bar{t}}(0, t), \quad t \in \omega_\tau, \\ y(1, t) = \mu_1(t), \quad t \in \omega_\tau. \end{cases} \quad (10.55)$$

Распишем ее в индексной форме и укажем способ реализации:

$$\begin{cases} \frac{y_i^{j+1} - y_i^j}{\tau} = \frac{y_{i+1}^j - 2y_i^j + y_{i-1}^j}{h^2} + \varphi_i^j, \quad i = \overline{1, N_1 - 1}; \quad j = \overline{0, N_2 - 1}, \\ y_i^0 = u_0(x_i), \quad i = \overline{0, N_1}, \\ \frac{y_1^{j+1} - y_0^{j+1}}{h} - \frac{h}{2} \frac{y_0^{j+1} - y_0^j}{\tau} = \sigma_0 y_0^{j+1} - \mu_0(t_{j+1}) - \frac{h}{2}f(0, t_{j+1}), \quad j = \overline{0, N_2 - 1}, \\ y_{N_1}^{j+1} = \mu_1(t_{j+1}), \quad j = \overline{0, N_2 - 1}. \end{cases}$$

Таким образом, решение на слое t_{j+1} может быть определено по рекуррентным формулам

$$\begin{cases} y_i^{j+1} = y_i^j + \frac{\tau}{h^2} \left(y_{i+1}^j - 2y_i^j + y_{i-1}^j \right) + \tau \varphi_i^j, \quad i = \overline{1, N_1 - 1}, \\ y_0^{j+1} = \frac{1}{\frac{1}{h} + \frac{h}{2\tau} + \sigma_0} \left(\frac{1}{h} y_1^{j+1} + \frac{h}{2\tau} y_0^j + \mu_0(t_{j+1}) + \frac{h}{2} f(0, t_{j+1}) \right), \quad j = \overline{0, N_2 - 1}, \\ y_{N_1}^{j+1} = \mu_1(t_{j+1}), \quad j = \overline{0, N_2 - 1}. \end{cases}$$



Начальное значение решения (при $t = 0$) определяется из начального условия:

$$y_i^0 = u_0(x_i), \quad i = \overline{0, N_1}.$$

Аналогичным образом может быть построена неявная разностная схема с погрешностью порядка $O(\tau + h^2)$. \square

Пример 10.17. Первая краевая задача для уравнения колебаний струны

$$\begin{cases} \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + f(x, t), \quad 0 < x < 1, \quad 0 < t \leq T, \\ u(x, 0) = u_0(x), \quad 0 \leq x \leq 1, \\ \frac{\partial u(x, 0)}{\partial t} = u_1(x), \quad 0 \leq x \leq 1, \\ u(0, t) = \mu_0(t), \quad 0 \leq t \leq T, \\ u(1, t) = \mu_1(t), \quad 0 \leq t \leq T. \end{cases} \quad (10.56)$$

Решение. На сетке $\omega_{h\tau}$, используя шаблон «крест» ([пример 10.5 в](#))), запишем явную разностную схему:

$$\begin{cases} y_{\bar{t}\bar{t}} = y_{\bar{x}\bar{x}} + f, \quad (x, t) \in \omega_{h\tau}, \\ y(x, 0) = u_0(x), \quad x \in \bar{\omega}_h, \\ y_t(x, 0) = u_1(x), \quad x \in \bar{\omega}_h, \\ y(0, t) = \mu_0(t), \quad t \in \omega_\tau, \\ y(1, t) = \mu_1(t), \quad t \in \omega_\tau. \end{cases} \quad (10.57)$$

Разностное уравнение имеет погрешность аппроксимации $O(\tau^2 + h^2)$, первое начальное условие и оба граничных аппроксимируются точно. Второе начальное условие имеет погрешность аппроксимации

$$v_1(x, 0) = u_t(x, 0) - u_1(x) = \frac{\partial u(x, 0)}{\partial t} + \frac{\tau}{2} \frac{\partial^2 u(x, 0)}{\partial t^2} + O(\tau^2) - u_1(x) = \frac{\tau}{2} \frac{\partial^2 u(x, 0)}{\partial t^2} + O(\tau^2) = O(\tau).$$

Чтобы обеспечить разностной схеме второй порядок аппроксимации по времени, поднимем порядок аппроксимации второго начального условия, не расширяя шаблона. Как и в предыдущих примерах, будем искать новое разностное условие в виде

$$y_t(x, 0) = \bar{u}_1(x),$$



где $\bar{u}_1(x)$ — неизвестная пока сеточная функция.

Имеем:

$$\begin{aligned} \nu_1(x, 0) &= u_t(x, 0) - \bar{u}_1(x) = \frac{\partial u(x, 0)}{\partial t} + \frac{\tau}{2} \frac{\partial^2 u(x, 0)}{\partial t^2} + O(\tau^2) - \bar{u}_1(x) = \\ &= \left[\frac{\partial^2 u(x, 0)}{\partial t^2} = \frac{\partial^2 u(x, 0)}{\partial x^2} + f(x, 0) = u''_0(x) + f(x, 0) \right] = u_1(x) + \frac{\tau}{2} (u''_0(x) + f(x, 0)) + \\ &= +O(\tau^2) - \bar{u}_1(x). \end{aligned}$$

Отсюда следует, что в качестве $\bar{u}_1(x)$ можно взять, например,

$$\bar{u}_1(x) = u_1(x) + \frac{\tau}{2} (u''_0(x) + f(x, 0))$$

(здесь можно использовать вместо $u''_0(x)$ (это — известная функция) и ее разностный аналог).

Таким образом, разностная схема второго порядка аппроксимации в безиндексной форме имеет вид

$$\begin{cases} y_{\bar{x}t} = y_{\bar{x}x} + f, \quad (x, t) \in \omega_{h\tau}, \\ y(x, 0) = u_0(x), \quad x \in \bar{\omega}_h, \\ y_t(x, 0) = u_1(x) + \frac{\tau}{2} (u''_0(x) + f(x, 0)), \quad x \in \bar{\omega}_h, \\ y(0, t) = \mu_0(t), \quad t \in \omega_\tau, \\ y(1, t) = \mu_1(t), \quad t \in \omega_\tau, \end{cases} \quad (10.58)$$

а в индексной:

$$\begin{cases} \frac{y_i^{j+1} - 2y_i^j + y_i^{j-1}}{\tau^2} = \frac{y_{i+1}^j - 2y_i^j + y_{i-1}^j}{h^2} + f_i^j, \quad i = \overline{1, N_1 - 1}; \quad j = \overline{0, N_2 - 1}, \\ y_i^0 = u_0(x_i), \quad i = \overline{0, N_1}, \\ \frac{y_i^1 - y_i^0}{\tau} = u_1(x_i) + \frac{\tau}{2} (u''_0(x_i) + f_i^0), \quad i = \overline{0, N_1}, \\ y_0^{j+1} = \mu_0(t_{j+1}), \quad j = \overline{0, N_2 - 1}, \\ y_{N_1}^{j+1} = \mu_1(t_{j+1}), \quad j = \overline{0, N_2 - 1}. \end{cases}$$



Решение соответствующей разностной задачи может быть найдено по формулам

$$\begin{cases} y_i^0 = u_0(x_i), \quad i = \overline{0, N_1}, \\ y_i^1 = y_i^0 + \tau [u_1(x_i) + \frac{\tau}{2} (u''_0(x_i) + f_i^0)], \quad i = \overline{0, N_1}, \\ y_i^{j+1} = 2y_i^j - y_i^{j-1} + \frac{\tau^2}{h^2} (y_{i+1}^j - 2y_i^j + y_{i-1}^j) + \tau^2 f_i^j, \quad i = \overline{1, N_1 - 1}, \\ y_0^{j+1} = \mu_0(t_{j+1}), \quad j = \overline{0, N_2 - 1}, \\ y_{N_1}^{j+1} = \mu_1(t_{j+1}), \quad j = \overline{0, N_2 - 1}. \end{cases}$$

□



Меню

10.1.8. Математический аппарат теории разностных схем

Некоторые разностные формулы

Отыскание собственных функций и собственных значений на примере простейшей разностной задачи

Разностные аналоги теорем вложения

Метод энергетических неравенств

Как уже отмечалось ранее, основной задачей теории разностных схем является получение априорных оценок решения разностной задачи через ее входные данные. Для этих целей нам в дальнейшем потребуются различные формулы преобразования разностных выражений, а также знание общей методики работы с разностными выражениями.

Некоторые разностные формулы

Получим сейчас простейшие формулы, проводя аналогию с соответствующими формулами дифференциального исчисления.

Формулы разностного дифференцирования произведения. Как известно, в дифференциальном исчислении существует формула дифференцирования произведения

$$(uv)' = u'v + uv'.$$

Для сеточных функций ранее мы ввели (на двухточечном шаблоне) два типа разностных производных — [правые](#) и [левые](#). Соответственно этому имеется и две формулы разностного дифференцирования произведения:

$$(uv)_x = u_x v + u^{(+1)} v_x = u_x v^{(+1)} + u v_x \quad (10.59)$$

и

$$(uv)_{\bar{x}} = u_{\bar{x}} v + u^{(-1)} v_{\bar{x}} = u_{\bar{x}} v^{(-1)} + u v_{\bar{x}}, \quad (10.60)$$

где $f^{(\pm 1)} = f(x \pm h)$.



Докажем, например, первое из этих равенств. Записывая его в индексной форме, получим:

$$\frac{u_{i+1}v_{i+1} - u_i v_i}{h} = \frac{u_{i+1} - u_i}{h} v_i + u_{i+1} \frac{v_{i+1} - v_i}{h}.$$

Справедливость последнего равенства очевидна.

Формулы суммирования по частям. Эти формулы являются разностными аналогами формулы интегрирования по частям

$$\int_0^1 uv' dx = uv \Big|_0^1 - \int_0^1 u' v dx.$$

Для сеточных функций, как и в предыдущем случае, имеются формулы двух типов:

$$(u, v_x) = u_N v_N - u_0 v_1 - (u_{\bar{x}}, v] \quad (10.61)$$

и

$$(u, v_{\bar{x}}) = u_N v_{N-1} - u_0 v_0 - [u_x, v], \quad (10.62)$$

где

$$(u, v) = \sum_{i=1}^{N-1} h u_i v_i, \quad (u, v] = \sum_{i=1}^N h u_i v_i, \quad [u, v) = \sum_{i=0}^{N-1} h u_i v_i.$$

Докажем, например, формулу (10.61):

$$\begin{aligned} (u, v_x) &= \sum_{i=1}^{N-1} (uv_x)_i h = [uv_x = (uv)_x - u_x v^{(+1)}] = \sum_{i=1}^{N-1} (uv)_{x,i} h - \sum_{i=1}^{N-1} u_{x,i} v_{i+1} h = \\ &= u_N v_N - u_1 v_1 - \sum_{i=1}^{N-1} u_{\bar{x},i+1} v_{i+1} h = u_N v_N - u_1 v_1 - \sum_{i=1}^N u_{\bar{x},i+1} v_{i+1} h + u_{\bar{x},1} v_1 h = \\ &= u_N v_N - u_1 v_1 + u_1 v_1 - u_0 v_1 - \sum_{i=1}^N u_{\bar{x},i+1} v_{i+1} h = u_N v_N - u_0 v_1 - (u_{\bar{x}}, v), \end{aligned}$$

что и требовалось доказать.



Первая разностная формула Грина. Равенство

$$\int_0^1 u(kv')' dx = - \int_0^1 ku'v' dx + kuv' \Big|_0^1$$

в дифференциальном исчислении обычно называют первой формулой Грина.

Для сеточных функций аналог первой формулы Грина можно получить, пользуясь формулами суммирования по частям. Подставляя в (10.61) $u = z$, $v = ay_{\bar{x}}$, получим

$$(z, (ay_{\bar{x}})_x) = -(ay_{\bar{x}}, z_{\bar{x}}] + a_N y_{\bar{x},N} z_N - a_1 y_{x,0} z_0. \quad (10.63)$$

Формула (10.63) — *первая разностная формула Грина*.

Отметим некоторые частные случаи, имеющие более простой вид и часто используемые на практике. Если $z_0 = z_N = 0$, то первая разностная формула Грина имеет вид

$$(z, (ay_{\bar{x}})) = -(ay_{\bar{x}}, z_{\bar{x}}] \quad (z, \Lambda y) = -(ay_{\bar{x}}, z_{\bar{x}}], \quad \Lambda y = (ay_{\bar{x}})_x; \quad (10.64)$$

при $z = y$ отсюда получаем:

$$(\Lambda y, y) = - \left(a, (y_{\bar{x}})^2 \right], \quad (10.65)$$

что на практике может быть использовано для исследования знакопостоянства разностного оператора Λ (при знакопостоянстве сеточной функции a).

Вторая разностная формула Грина. В интегральном исчислении вторая формула Грина имеет вид

$$\int_0^1 u(kv')' dx - \int_0^1 v(ku')' dx = k(uv' - vu') \Big|_0^1.$$

Чтобы получить ее разностный аналог, запишем на основании (10.63) соотношение

$$(y, (az_{\bar{x}})_x) = -(az_{\bar{x}}, y_{\bar{x}}] + a_N z_{\bar{x},N} y_N - a_1 z_{x,0} y_0 \quad (10.66)$$



Теперь, вычитая из (10.63) (10.63), будем иметь:

$$(z, (ay_{\bar{x}})_x) - (y, (az_{\bar{x}})_x) = a_N (zy_{\bar{x}} - yz_{\bar{x}})_N - a_1 (zy_x - yz_x)_0 \quad (10.67)$$

Формула (10.67) — *вторая разностная формула Грина*. Из нее, в частности, в случае, когда сеточные функции y и z обращаются в ноль при $x = 0$ и $x = 1$, непосредственно следует равенство

$$(z, \Lambda y) = (y, \Lambda z),$$

которое означает самосопряженность введенного выше разностного оператора Λ .

Неравенство Коши–Буняковского и ε -неравенство. Напомним здесь известные из курса анализа неравенства. Одно из них — *неравенство Коши–Буняковского* — имеет вид

$$|(u, v)| \leq \|u\| \cdot \|v\|,$$

где (u, v) — скалярное произведение в некотором линейном пространстве (в том числе, и в пространстве сеточных функций), а $\|u\| = \sqrt{(u, u)}$. В нашем случае под скалярным произведением будем понимать любое из введенных ранее скалярных произведений в пространстве сеточных функций. Второе из неравенств — *ε -неравенство* — имеет вид

$$|ab| \leq \varepsilon a^2 + \frac{b^2}{4\varepsilon},$$

где ε — любое положительное число. Из него, в частности, получаем неравенство

$$|(u, v)| \leq \|u\| \cdot \|v\| \leq \varepsilon \|u\|^2 + \frac{1}{4\varepsilon} \|v\|^2.$$

Отыскание собственных функций и собственных значений на примере простейшей разностной задачи

Применение известного из курса уравнений в частных производных метода разделения переменных в теории разностных схем приводит к появлению разностных задач на собственные значения.

Рассмотрим сейчас задачу об отыскании собственных значений для простейшего разностного оператора.



Предварительно напомним основные факты, связанные с простейшей задачей об отыскании собственных функций и собственных значений для дифференциального оператора второй производной. В математической формулировке задача выглядит следующим образом: найти, при каких значениях числового параметра λ существуют нетривиальные решения краевой задачи

$$\begin{cases} u''(x) + \lambda u(x) = 0, & 0 < x < l, \\ u(0) = u(l) = 0 \end{cases} \quad (10.68)$$

и указать эти решения.

Известно следующее:

- 1) Нетривиальные решения задачи (10.68) — собственные функции $u_k(x)$ — и отвечающие им собственные значения λ_k — выражаются следующим образом:

$$u_k(x) = \sqrt{\frac{2}{l}} \sin \frac{k\pi x}{l}, \quad \lambda_k = \frac{k^2\pi^2}{l^2}, \quad k = 1, 2, \dots \quad (10.69)$$

- 2) Собственные функции $u_k(x)$ образуют ортонормированную систему:

$$\int_0^l u_k(x) u_m(x) dx = \delta_k^m;$$

- 3) Если $f(x)$ дважды непрерывно дифференцируема и удовлетворяет однородным краевым условиям ($f(0) = f(l) = 0$), то она представима в виде равномерно сходящегося ряда:

$$f(x) = \sum_{k=1}^{\infty} f_k u_k(x),$$

где

$$f_k = \int_0^l f(x) u_k(x) dx,$$



Меню

причем

$$\|f\|^2 = \int_0^l f^2(x) dx = \sum_{k=1}^{\infty} f_k^2.$$

Поставим в соответствие дифференциальной задаче (10.68) на равномерной сетке $\bar{\omega}_h [0, l]$ разностную задачу

$$\begin{cases} y_{\bar{x}\bar{x}} + \lambda y = 0, & x \in \omega_h, \\ y(0) = y(l) = 0 \end{cases} \quad (10.70)$$

об отыскании нетривиальных решений — собственных функций $y_k(x)$ и соответствующих им собственных значений.

Перейдем в (10.70) к индексной форме записи:

$$y_{i+1} - 2 \left(1 - \frac{h^2 \lambda}{2}\right) y_i + y_{i-1} = 0, \quad i = \overline{1, N-1}. \quad (10.71)$$

Эта формула представляет собой разностное уравнение второго порядка с постоянными коэффициентами. Заметим, что при наличии у соответствующего ему характеристического уравнения

$$r^2 - 2 \left(1 - \frac{h^2 \lambda}{2}\right) r + 1 = 0 \quad (10.72)$$

вещественных корней построить нетривиальное решение задачи (10.70) (как и в случае исходной дифференциальной задачи (10.68)!) не удается. Поэтому необходимым условием разрешимости задачи (10.70) является отрицательность дискриминанта квадратного уравнения (10.72), которая, как легко видеть, имеет место при $\lambda \in (0, \frac{4}{h^2})$. Поскольку при этом также $r_1 r_2 = 1$, то решение уравнения (10.71) будем искать (что непосредственно следует из общей теории разностных уравнений) в виде $y = \sin \alpha x$, где постоянная α подлежит определению. Поскольку в этом случае

$$y_{i+1} + y_{i-1} = [x_i = x] = \sin \alpha (x + h) + \sin \alpha (x - h) = 2 \sin \alpha x \cos \alpha h,$$



то из (10.71) следует:

$$2 \sin \alpha x \cos \alpha h = 2 \left(1 - \frac{h^2 \lambda}{2}\right) \sin \alpha x.$$

Так как мы ищем нетривиальное решение, т.е. $\sin \alpha x$ отлично от тождественного нуля, то отсюда следует

$$1 - \frac{h^2 \lambda}{2} = \cos \alpha h,$$

т.е.

$$\lambda = \frac{2}{h^2} (1 - \cos \alpha h) = \frac{4}{h^2} \sin^2 \frac{\alpha h}{2}.$$

Значение параметра α выберем так, чтобы функция $y = \sin \alpha x$ удовлетворяла граничным условиям задачи (10.70): $y(0) = y(l) = 0$. При $x = 0 \sin \alpha x = 0$ при любых значениях α , а при $x = l$ имеем:

$$\sin \alpha l = 0,$$

откуда

$$\alpha l = k\pi, \quad k = 1, \dots, N - 1.$$

Тогда

$$\alpha = \frac{k\pi}{l} = \alpha_k, \quad k = 1, \dots, N - 1.$$

Таким образом, мы получили собственные функции и собственные значения задачи (10.70). Перечислим их свойства.

Свойства собственных функций и собственных значений задачи (10.70)

1. Множество собственных функций и собственных значений задачи (10.70) имеет вид

$$y^{(k)}(x) = \sin \frac{k\pi x}{l}, \quad \lambda_k = \frac{4}{h^2} \sin^2 \frac{k\pi h}{2l}, \quad k = 1, \dots, N - 1. \quad (10.73)$$



2. Собственные значения λ_k перенумерованы в порядке возрастания, причем

$$0 < \lambda_1 = \frac{4}{h^2} \sin^2 \frac{\pi h}{2l} < \lambda_2 < \dots < \lambda_{N-1} = \frac{4}{h^2} \sin^2 \frac{\pi(N-1)h}{2l} = \frac{4}{h^2} \cos^2 \frac{\pi h}{2l} < \frac{4}{h^2}. \quad (10.74)$$

Отсюда, в частности, следует, что все собственные значения задачи (10.70) положительны.

3. Собственные функции задачи (10.70) $y^{(k)}(x)$, $y^{(m)}(x)$, отвечающие различным собственным значениям, ортогональны: $(y^{(k)}, y^{(m)}) = 0$ при $k \neq m$. [\[Доказательство\]](#)

4. $\|y^{(k)}\| = \sqrt{\frac{l}{2}}$. [\[Доказательство\]](#)

5. Набор сеточных функций

$$\mu^{(k)}(x) = \sqrt{\frac{2}{l}} y^{(k)}(x) = \sqrt{\frac{2}{l}} \sin \frac{k\pi x}{l}, \quad k = \overline{1, N-1}, \quad (10.75)$$

образует ортонормированную систему (это следует из предыдущего свойства).

6. Пусть на сетке $\bar{\omega}_h$ задана функция $f(x)$, причем $f(0) = f(l) = 0$. Тогда она представима в виде

$$f(x) = \sum_{k=1}^{N-1} f_k \mu^{(k)}(x), \quad (10.76)$$

где $f_k = (f(x), \mu^{(k)}(x))$, причем справедливо равенство

$$\|f\|^2 = \sum_{k=1}^{N-1} f_k^2. \quad (10.77)$$

[\[Доказательство\]](#)



Разностные аналоги теорем вложения

При оценке различных свойств разностных схем часто используются неравенства, связывающие нормы в различных функциональных пространствах, соответствующие простейшим теоремам вложения Соболева.

Лемма 10.1. Для всякой сеточной функции $y(x)$, заданной на сетке $\bar{\omega}_h$ и обращающейся в нуль при $x = 0$ и $x = l$, справедливо неравенство

$$\|y\|_C \leq \frac{1}{2} \|y_{\bar{x}}\| \quad (10.78)$$

[[Доказательство](#)]

Замечание 10.2. Для $\bar{\omega}_h(0, l)$ неравенство (10.78) следует переписать в виде

$$\|y\|_C \leq \frac{\sqrt{l}}{2} \|y_{\bar{x}}\| \quad (10.79)$$

Замечание 10.3. Если $y(0) \cdot y(l) \neq 0$, то (10.79), вообще говоря, неверно.

Замечание 10.4. Для случая неравномерной сетки (10.78) остается в силе.

Лемма 10.2. Для всякой функции $y(x)$, заданной на сетке $\bar{\omega}_h(0, l)$ и обращающейся в нуль при $x = 0$ и $x = l$, справедливы оценки

$$\frac{h^2}{4} \|y_{\bar{x}}\|^2 \leq \|y\|^2 \leq \frac{l^2}{8} \|y_{\bar{x}}\|^2. \quad (10.80)$$

[[Доказательство](#)]

Метод энергетических неравенств

Одним из общих и весьма эффективных способов получения априорных оценок является [метод энергетических неравенств](#). Приведем пример использования данного метода для получения априорных оценок применительно к разностным задачам.



Пусть имеем модельную задачу

$$\begin{cases} u''(x) + f(x) = 0, \quad 0 < x < 1, \\ u(0) = u(1) = 0. \end{cases} \quad (10.81)$$

Введем на отрезке $[0, 1]$ равномерную сетку $\bar{\omega}_h$ и заменим (10.81) разностной схемой

$$\begin{cases} y_{\bar{x}x} + f(x) = 0, \quad x \in \omega_h, \\ y(0) = y(1) = 0. \end{cases} \quad (10.82)$$

Умножив разностное уравнение (10.82) скалярно на исковую функцию y , получим:

$$(y_{\bar{x}x}, y) + (f, y) = 0.$$

Применив к первому слагаемому первую разностную формулу Грина, перепишем полученное равенство в виде

$$\|y_{\bar{x}}\|^2 = (f, y) \quad (10.83)$$

Согласно неравенству Коши — Буняковского $|(f, y)| \leq \|f\| \cdot \|y\|$, а в силу леммы 10.2 имеем оценку $\|y\|^2 \leq \frac{l^2}{8} \|y_{\bar{x}}\|^2$ или $\|y\| \leq \frac{l}{2\sqrt{2}} \|y_{\bar{x}}\|$. Поэтому для скалярного произведения получим оценку сверху вида

$$|(f, y)| \leq \frac{1}{2\sqrt{2}} \|f\| \cdot \|y_{\bar{x}}\|.$$

Таким образом, используя лемму 10.1, из (10.83) последовательно будем иметь:

$$\|y_{\bar{x}}\|^2 = (f, y) \leq \frac{1}{2\sqrt{2}} \|f\| \cdot \|y_{\bar{x}}\|,$$

$$2\|y\|_C \leq \|y_{\bar{x}}\| \leq \frac{1}{2\sqrt{2}} \|f\|$$

и, наконец,

$$\|y\|_C \leq \frac{1}{4\sqrt{2}} \|f\|. \quad (10.84)$$



Выражение (10.84) дает априорную оценку решения разностной задачи (10.82) через входные данные.

Покажем, как ее можно использовать для оценки скорости сходимости разностной схемы (10.84).

Запишем уравнение для погрешности. Если $z = y - u$, то для z имеем задачу

$$\begin{cases} z_{xx} + \psi(x) = 0, & x \in \omega_h, \\ z(0) = z(1) = 0, \end{cases} \quad (10.85)$$

где $\psi(x)$ — погрешность аппроксимации разностной схемы на решении задачи (10.81). Согласно (10.84) можем записать:

$$\|z\|_C \frac{1}{4\sqrt{2}} \|\psi\|,$$

а так как $\psi = O(h^2)$, то, следовательно, $\|z\|_C = \|y - u\|_C \leq Mh^2$, т.е. решение разностной задачи (7.23) сходится к решению дифференциальной задачи (10.81) со скоростью $O(h^2)$.

В заключение отметим, что метод энергетических неравенств является достаточно универсальным.



Меню



10.2. Способы построения разностных схем

10.2.1. Требования, предъявляемые к разностным схемам

10.2.2. Интегро-интерполяционный метод построения разностных схем

10.2.3. Вариационно-проекционные подходы к построению разностных схем



10.2.1. Требования, предъявляемые к разностным схемам

[Однородные разностные схемы](#)

[Консервативные разностные схемы](#)

Выше мы приводили примеры [разностных аппроксимаций для дифференциальных операторов различных порядков](#), а также разностных схем для дифференциальных уравнений [первого и второго порядка](#), в том числе с [граничными условиями, содержащими производные от искомого решения](#).

При этом в случае дифференциальных уравнений с переменными коэффициентами задача построения разностной схемы может существенно усложниться. На заданном шаблоне мы можем построить бесчисленное множество разностных схем, эквивалентных по порядку аппроксимации. Так, например, для дифференциального уравнения

$$\frac{d^2u}{dx^2} - q(x)u = -f(x)$$

на трехточечном шаблоне можно построить однопараметрическое семейство разностных аппроксимаций

$$L_h y_i = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} - d_i y_i = -f_i, \quad d_i = \alpha q_{i-1} + (1 - 2\alpha) q_i + \alpha q_{i+1},$$

имеющих при любом вещественном значении параметра α второй порядок аппроксимации. Аналогичную конструкцию можно использовать и для аппроксимации функции $f(x)$. Кроме того, к коэффициенту d_i (равно как и к f_i) можно без нарушения порядка аппроксимации добавлять слагаемые вида βh^2 , где β – произвольное не зависящее от шага h число.

Таким образом, возникает задача выбора разностных схем из множества допустимых схем, заданных на некотором шаблоне и имеющих один и тот же порядок аппроксимации. Для этого необходимо сформулировать требования, которые следует предъявлять к разностным схемам. Интуитивно понятно, что любой приближенный метод должен давать возможность найти численное решение с заданной точностью ε за конечное число действий $Q(\varepsilon)$. Естественно поэтому стремиться минимизировать величину $Q(\varepsilon)$, т.е. найти оптимальный метод.

При фиксированном методе решения системы объем вычислений зависит от ее порядка. Он тем меньше, чем крупнее шаг сетки. Однако уменьшение числа узлов сетки приводит к уменьшению точности схемы.

Поэтому желательно иметь схему с возможно более высоким порядком точности (который зависит от гладкости коэффициентов дифференциального уравнения, начальных и граничных условий). Практически это означает, что надо искать схемы с минимальным шаблоном, имеющие максимально возможный на этом шаблоне порядок аппроксимации.

Итак, количественные требования к семейству разностных схем могут выглядеть следующим образом:

- 1) определенный [порядок аппроксимации](#);
- 2) максимальный [порядок точности](#) на всем классе решаемых задач;
- 3) экономичность, т.е. минимум операций при машинной реализации сеточных уравнений.

Необходимо выделить также следующие качественные характеристики разностных схем:

- 1) схема должна быть [однородной](#), т.е. сеточные уравнения для любой задачи из рассматриваемого класса K и любой сетки в любом узле должны записываться единообразно, по одному и тому же закону;
- 2) система разностных уравнений должна быть разрешимой на любой допустимой сетке и для любой задачи из рассматриваемого класса K ;
- 3) схема должна быть [сходящейся](#) для любой задачи из рассматриваемого класса K .

Однородные разностные схемы

Однородность разностной схемы означает, что все ее коэффициенты являются функционалами коэффициентов дифференциальной задачи, зависящими от шага сетки h как от параметра и не зависящими от узла сетки и от выбора коэффициентов задачи.

Формальная схема может выглядеть следующим образом. Пусть заданы:

- а) целочисленный шаблон $\mathbb{W} = \{-m_1, -m_1 + 1, \dots, -1, 0, 1, \dots, m_2\}$, где $m_1 > 0$ и $m_2 > 0$ — целые числа, на котором определяется сеточная функция $\bar{y}(j)$, $j \in \mathbb{W}$;
- б) шаблон $\Sigma = \{-m_1 \leq s \leq m_2\}$, на котором определена вектор-функция $\bar{k}(s)$ коэффициентов исходной дифференциальной задачи (концы шаблонов \mathbb{W} и Σ могут и не совпадать).



Обозначим через $A_j^h(\bar{k}(s))$, $F^h(\bar{k}(s))$, $j \in (-m_1, m_2)$, $s \in \Sigma$, шаблонные функционалы. Рассматривается функционал

$$\Phi^h(\bar{y}(j)) = \sum_{j=-m_1}^{m_2} A_j^h(\bar{k}(s)) \bar{y}(j) + F^h(\bar{k}(s))$$

и от него осуществляется переход к однородной схеме следующим образом: полагая $\bar{y}(j) = y^h(x_i + jh)$, $\bar{k}(s) = k(x_i + sh)$ и пользуясь выражением для Φ^h , получаем однородную разностную схему

$$(L_h y^h + F^h)_i = \sum_{i=-m_1}^{m_2} A_j^h(k(x_i + sh)) y^h(x_i + jh) + F^h(k(x_i + sh)) = 0,$$

где $y^h(x_i)$ — сеточная функция, $k(x)$ — вектор-функция непрерывного аргумента.

Семейство однородных схем задано, если заданы шаблонные функционалы $A_j^h(\bar{k}(s))$ и $F^h(\bar{k}(s))$, $j = -m_1, m_2$. Производств в их выборе должен быть ограничен требованиями разрешимости, аппроксимации определенного порядка, экономичности.

Проиллюстрируем понятие однородности на примере трехточечных схем для задачи

$$\begin{cases} \frac{d}{dx} \left(k(x) \frac{du(x)}{dx} \right) - q(x) u(x) = -f(x), & 0 < x < 1, \\ u(0) = \mu_0, \quad u(1) = \mu_1, \quad k(x) \geq c > 0, \quad q(x) \geq 0. \end{cases} \quad (10.86)$$

На сетке $\bar{\omega}_h$ рассмотрим трехточечный шаблон $\{x_{i-1}, x_i, x_{i+1}\}$, так что ($= \{-1, 0, 1\}$) и $m_1 = m_2 = 1$. Пусть коэффициентный шаблон имеет вид $\Sigma = \{-1 \leq s \leq 1\}$ и $A^h(\bar{k}(s))$, $B^h(\bar{k}(s))$, $F^h(\bar{k}(s))$ — шаблонные функционалы. Тогда однородная разностная схема будет иметь вид

$$\begin{cases} \frac{1}{h} \left(b_i \frac{y_{i+1} - y_i}{h} - a_i \frac{y_i - y_{i-1}}{h} \right) - d_i y_i = -\varphi_i, & i = 1, \dots, N-1, \\ y_0 = \mu_0, \quad y_N = \mu_1, \end{cases} \quad (10.87)$$

причем коэффициенты ее вычисляются во всех узлах $x_i \in \omega_h$ и для любых $k(x)$, $q(x)$, $f(x)$ одинаково:

$$a_i = A^h(k(x_i + sh)), \quad b_i = B^h(k(x_i + sh)), \quad d_i = F^h(q(x_i + sh)), \quad \varphi_i = F^h(f(x_i + sh)).$$



Для простоты здесь каждый из коэффициентов разностного уравнения зависит только от соответствующего коэффициента дифференциального уравнения (причем для d_i и φ_i эта зависимость одинакова). В общем случае A^h , B^h и F^h — нелинейные функционалы, но мы далее будем предполагать их линейными и не зависящими от h .

Найдем сейчас условия, при которых разностная схема (10.87) имеет второй порядок аппроксимации. Так как

$$\frac{u_{i+1} - u_i}{h} = u'_i + \frac{h}{2} u''_i + \frac{h^2}{6} u'''_i + O(h^3),$$

$$\frac{u_i - u_{i-1}}{h} = u'_i - \frac{h}{2} u''_i + \frac{h^2}{6} u'''_i + O(h^3),$$

и исходное дифференциальное уравнение, раскрыв скобки в его левой части, можно переписать в виде $k u'' + k' u' - q u + f = 0$, то

$$\begin{aligned}\psi_i &= \frac{1}{h} \left(b_i \frac{u_{i+1} - u_i}{h} - a_i \frac{u_i - u_{i-1}}{h} \right) - d_i u_i + \varphi_i - (k_i u''_i + k'_i u'_i - q_i u_i + f_i) = \\ &= \left(\frac{b_i + a_i}{2} - k_i \right) u''_i + \left(\frac{b_i - a_i}{h} - k'_i \right) u'_i - (d_i - q_i) u_i + \varphi_i - f_i + O(h^2).\end{aligned}$$

Таким образом, схема (10.87) будет иметь второй порядок аппроксимации, если

$$\left\{ \begin{array}{l} \frac{b_i + a_i}{2} = k_i + O(h^2); \\ \frac{b_i - a_i}{h} = k'_i + O(h^2); \\ d_i = q_i + O(h^2); \\ \varphi_i = f_i + O(h^2) \end{array} \right. \quad (10.88)$$

Воспользовавшись разложениями

$$k(x + sh) = k(x) + shk'(x) + \frac{s^2 h^2}{2} k''(x) + O(h^3),$$

$$f(x + sh) = f(x) + shf'(x) + O(h^2);$$

$$q(x + sh) = q(x) + shq'(x) + O(h^2),$$



получим:

$$\begin{aligned} a_i &= A(k(x_i + sh)) = A\left[k_i + shk'_i + \frac{s^2 h^2}{2} k''_i + O(h^3)\right] = \\ &= A(1)k_i + hk'_i A(s) + \frac{h^2}{2} k''_i A(s^2) + O(h^3); \end{aligned}$$

$$\begin{aligned} b_i &= B(k(x_i + sh)) = B\left[k_i + shk'_i + \frac{s^2 h^2}{2} k''_i + O(h^3)\right] = \\ &= B(1)k_i + hk'_i B(s) + \frac{h^2}{2} k''_i B(s^2) + O(h^3); \end{aligned}$$

$$d_i = F(q(x_i + sh)) = F[q_i + shq'_i + O(h^2)] = F(1)q_i + hq'_i F(s) + O(h^2);$$

$$\varphi_i = F(f(x_i + sh)) = F[f_i + shf'_i + O(h^2)] = F(1)f_i + hf'_i F(s) + O(h^2).$$

Тогда условия (10.88) перепишутся в виде

$$\left\{ \begin{array}{l} \frac{A(1)+B(1)}{2}k_i + hk'_i \frac{A(s)+B(s)}{2} + O(h^2) = k_i + O(h^2); \\ \frac{B(1)-B(1)}{h}k_i + hk'_i \frac{B(s)-A(s)}{h} + \frac{h^2}{2} k''_i \frac{B(s^2)-A(s^2)}{h} + O(h^2) = k'_i + O(h^2); \\ F(1)q_i + hq'_i F(s) + O(h^2) = q_i + O(h^2); \\ F(1)f_i + hf'_i F(s) + O(h^2) = f_i + O(h^2). \end{array} \right.$$

Отсюда, приравнивая коэффициенты при одинаковых степенях h , получаем:

$$\left\{ \begin{array}{l} \frac{A(1)+B(1)}{2} = 1; \\ \frac{B(1)-A(1)}{h} = 0; \\ \frac{A(s)+B(s)}{2} = 0; \\ B(s) - A(s) = 1; \\ F(1) = 1; \\ F(s) = 0; \\ \frac{B(s^2)-A(s^2)}{2} = 0; \end{array} \right. , \quad (*)$$



Решая данную систему, находим:

$$\left\{ \begin{array}{l} A(1) = B(1) = F(1) = 1; \\ B(s) = \frac{1}{2}; \\ A(s) = -\frac{1}{2}; \\ F(s) = 0; \\ B(s^2) = A(s^2) \end{array} \right. \quad (10.89)$$

Требование разрешимости системы разностных уравнений (10.87) будет выполнено, если $a_i > 0$, $b_i > 0$, $a_i + b_i + h^2 d_i \geq a_i + b_i$ (это ведь не что иное как условие применимости [метода разностной прогонки](#)). Для выполнения этих условий достаточно потребовать, чтобы функционалы A , B и F были положительны. Экономичность разностной схемы гарантируется алгоритмом разностной прогонки.

В простейшем случае A , B и F представляют собой линейные комбинации значений функций $\bar{k}(s)$ и $\bar{f}(s)$ в конечном числе точек на шаблоне Σ , например

$$\left\{ \begin{array}{l} A(\bar{k}(s)) = \alpha_{-1}\bar{k}(-1) + \alpha_0\bar{k}(0) + \alpha_1\bar{k}(1), \\ B(\bar{k}(s)) = \beta_{-1}\bar{k}(-1) + \beta_0\bar{k}(0) + \beta_1\bar{k}(1), \\ F(\bar{k}(s)) = \gamma_{-1}\bar{k}(-1) + \gamma_0\bar{k}(0) + \gamma_1\bar{k}(1), \end{array} \right.$$

так что

$$\left\{ \begin{array}{l} a_i = \alpha_{-1}k_{i-1} + \alpha_0k_i + \alpha_1k_{i+1}, \\ b_i = \beta_{-1}k_{i-1} + \beta_0k_i + \beta_1k_{i+1}, \\ d_i = \gamma_{-1}q_{i-1} + \gamma_0q_i + \gamma_1q_{i+1}, \\ \varphi_i = \gamma_{-1}f_{i-1} + \gamma_0f_i + \gamma_1f_{i+1}. \end{array} \right.$$



В этом случае условия второго порядка аппроксимации (10.89) могут быть переписаны в виде системы соотношений, связывающих коэффициенты α_i , β_i , γ_i :

$$\left\{ \begin{array}{l} \alpha_{-1} + \alpha_0 + \alpha_1 = 1, \\ \beta_{-1} + \beta_0 + \beta_1 = 1, \\ \gamma_{-1} + \gamma_0 + \gamma_1 = 1, \\ -\gamma_{-1} + \gamma_1 = 0, \\ \\ -\beta_{-1} + \beta_1 = \frac{1}{2}, \\ -\alpha_{-1} + \alpha_1 = -\frac{1}{2}, \\ \beta_{-1} + \beta_1 = \alpha_{-1} + \alpha_1. \end{array} \right.$$

Таким образом, при обсуждаемом способе задания шаблонных функционалов существует двухпараметрическое семейство трехточечных однородных разностных схем второго порядка.

Замечание 10.5. Чтобы разностная схема (10.88) имела *первый порядок аппроксимации* (т.е. вообще аппроксимировала задачу) в рассмотренной выше системе (*) достаточно оставить только четыре уравнения

$$\left\{ \begin{array}{l} \frac{A(1)+B(1)}{2} = 1, \\ \frac{B(1)-A(1)}{h} = 0, \\ B(s) - A(s) = 1, \\ F(1) = 1, \end{array} \right.$$

откуда следует:

$$\left\{ \begin{array}{l} A(1) = B(1) = F(1) = 1, \\ B(s) - A(s) = 1. \end{array} \right. \quad (10.90)$$

Консервативные разностные схемы

Помимо формальных требований разрешимости, аппроксимации, экономичности необходимо обеспечить также сходимость. Вообще говоря, это следует из аппроксимации и устойчивости ([теорема Лакса](#)). Однако в реальном вычислительном процессе шаг сетки не должен быть слишком малым.



Чтобы получить хорошее приближение на реальных сетках, необходимо, как показывает практика, пользоваться схемами, хорошо отражающими основные свойства дифференциальных уравнений.

Уравнения же математической физики выражают, как правило, законы сохранения в дифференциальной форме. Так, например, уравнение

$$\frac{d}{dx} \left(k(x) \frac{du}{dx} \right) = -f(x), \quad 0 < x < 1 \quad (10.91)$$

можно трактовать как уравнение стационарного распределения температуры $u(x)$ в стержне $0 < x < 1$ с коэффициентом теплопроводности $k(x)$. Интегрируя это уравнение по переменной x от $x^{(1)}$ до $x^{(2)}$, получим закон сохранения тепла на отрезке $x^{(1)} \leq x \leq x^{(2)}$:

$$W(x^{(2)}) - W(x^{(1)}) = \int_{x^{(1)}}^{x^{(2)}} f(x) dx, \quad W(x) = -k(x) \frac{du}{dx} \quad (10.92)$$

Слева в равенстве (10.92) стоит разность тепловых потоков на концах отрезка, справа — количество выделившегося (поглотившегося) тепла.

Определение. Разностные схемы, которые выражают законы сохранения на сетке, называют **консервативными** разностными схемами.

Поясним смысл консервативности на примере разностной схемы (10.87) для задачи (10.91). Так как $q(x) = 0$, то (10.87) имеет вид

$$\frac{1}{h} \left(b_i \frac{y_{i+1} - y_i}{h} - a_i \frac{y_i - y_{i-1}}{h} \right) = -\varphi_i$$

или

$$\frac{1}{h} \left(a_{i+1} \frac{y_{i+1} - y_i}{h} - a_i \frac{y_i - y_{i-1}}{h} \right) = -\varphi_i - \frac{b_i - a_{i+1}}{h} \cdot \frac{y_{i+1} - y_i}{h}.$$

Просуммировав по сетке от i_1 до i_2 ($x^{(1)} = i_1 h$; $x^{(2)} = i_2 h$), получим:

$$W_{i_2+1}^h - W_{i_1}^h = \sum_{i=i_1}^{i_2} h \varphi_i + \sum_{i=i_1}^{i_2} (b_i - a_{i+1}) \frac{y_{i+1} - y_i}{h} \quad (10.93)$$



Вверх

Назад

Вперёд

Пред.

След.

Указатель

Помощь

Экран

где $W_i^h = -a_i \frac{y_i - y_{i-1}}{h}$. Формула (10.93) — сеточный аналог закона сохранения (10.92).

В правую часть равенства (10.92) входит величина дисбаланса

$$D = \sum_{i=i_1}^{i_2} (b_i - a_{i+1}) \frac{y_{i+1} - y_i}{h},$$

которая обратится в нуль для любых сеточных функций y_i только при условии

$$b_i = a_{i+1} \quad (10.94)$$

Условие (10.94) — необходимое и достаточное условие консервативности разностной схемы (в данном случае при условии $y_0 = y_N = 0$ оно совпадает с условием самосопряженности оператора Λu).

Консервативность является важным свойством. Покажем, что консервативность является необходимым условием сходимости разностной схемы в случае простейшей задачи

$$\begin{cases} \frac{d}{dx} (k(x) \frac{du}{dx}) = 0, & 0 < x < 1, \\ u(0) = 1, \quad u(1) = 0 \end{cases} \quad (10.95)$$

в классе кусочно-постоянных коэффициентов:

$$k(x) = \begin{cases} k_1, & \text{если } 0 < x < \xi, \\ k_2, & \text{если } \xi < x < 1, \end{cases} \quad (10.96)$$

где ξ — иррациональное число: $\xi = x_n + \theta h$, $0 < \theta < 1$.

Как известно, точное решение такой задачи в точках разрыва коэффициентов удовлетворяет условиям сопряжения (непрерывности температуры и теплового потока):

$$\begin{cases} [u] = u(\xi + 0) - u(\xi - 0) = 0, \\ [ku'] = k_2 u'(\xi + 0) - k_1 u'(\xi - 0) = 0. \end{cases}$$

Исследование начнем с нахождения аналитического решения поставленной задачи.



Поскольку на промежутке $[0, \xi)$ уравнение имеет вид $u''(x) = 0$, то $u(x) = C_0 + C_1x$, $x \in [0, \xi)$, а так как при этом $u(0) = 1$, то $C_0 = 1$. Таким образом, $u(x) = 1 - \gamma_0x$, $x \in [0, \xi)$.

Аналогично при $x \in (\xi, 1]$ также $u''(x) = 0$, откуда $u(x) = C_2 + C_3x$, $x \in (\xi, 1]$ и в силу граничного условия на правом конце отрезка $C_2 + C_3 = 0$, т.е. $u(x) = \delta_0(1 - x)$, $x \in (\xi, 1]$.

Таким образом,

$$u(x) = \begin{cases} 1 - \gamma_0x, & 0 \leq x < \xi, \\ \delta_0(1 - x), & \xi < x \leq 1. \end{cases}$$

В точке $x = \xi$ функцию $u(x)$ доопределим ее предельным значением, которое, в силу первого из условий сопряжения, существует. При этом само условие примет вид

$$1 - \gamma_0\xi = \delta_0(1 - \xi).$$

Аналогично второе условие сопряжения, учитывая значения коэффициента теплопроводности, может быть переписано в виде

$$-k_1\gamma_0 = -k_2\delta_0.$$

Следовательно, получаем систему из двух уравнений для определения параметров γ_0 и δ_0

$$\begin{cases} 1 - \gamma_0\xi = \delta_0(1 - \xi), \\ k_1\gamma_0 = k_2\delta_0. \end{cases}$$

Из второго уравнения этой системы получаем: $\delta_0 = \frac{k_1}{k_2}\gamma_0$. Подставляя это выражение в первое уравнение, находим:

$$\gamma_0 \left[\frac{k_1}{k_2} (1 - \xi) + \xi \right] = 1.$$

Отсюда

$$\gamma_0 = \frac{1}{\Delta_0},$$

где

$$\Delta_0 = \frac{k_1}{k_2}(1 - \xi) + \xi.$$



Применим теперь для решения задачи (10.95), (10.96) однородную разностную схему вида (10.87):

$$\begin{cases} \frac{1}{h} \left(b_i \frac{y_{i+1} - y_i}{h} - a_i \frac{y_i - y_{i-1}}{h} \right) = 0, & i = \overline{1, N-1}, \\ y_0 = 1, \\ y_N = 0. \end{cases} \quad (10.97)$$

Здесь, как и выше, $b_i = B(k(x_i + sh))$, $a_i = A(k(x_i + sh))$.

Так как рассматриваемая разностная схема обладает аппроксимацией (по крайней мере, первого порядка), то должны выполняться условия (10.90). Следовательно, на всех отрезках длины $2h$, не содержащих точки разрыва коэффициента теплопроводности, коэффициенты схемы a_i , b_i будут постоянны ($A(k_1) = k_1 \cdot A(1) = k_1$ и т.п.), т.е. справедливы равенства

$$\begin{cases} a_i = b_i = k_1, & 0 < i < n, \\ a_i = b_i = k_2, & n+1 < i < N. \end{cases}$$

Таким образом, рассматриваемое разностное уравнение при $i = \overline{1, n-1}$ и $i = \overline{n+2, N-1}$ может быть переписано в виде

$$y_{i+1} - 2y_i + y_{i-1} = 0$$

и его решение, учитывающее граничные условия по аналогии с точным может быть определено в виде

$$y_i = \begin{cases} 1 - \gamma x_i, & 0 \leq i \leq n, \\ \delta(1 - x_i), & n+1 \leq i \leq N. \end{cases} \quad (10.98)$$

Запишем теперь разностное уравнение (10.97) при двух оставшихся значениях индекса i : при $i = n$ и $i = n+1$:

$$\begin{cases} b_n(y_{n+1} - y_n) - a_n(y_n - y_{n-1}) = 0, \\ b_{n+1}(y_{n+2} - y_{n+1}) - a_{n+1}(y_{n+1} - y_n) = 0. \end{cases} \quad (10.99)$$

Умножая первое из этих уравнений на a_{n+1} , а второе — на b_n и складывая, получим:

$$b_{n+1}b_n(y_{n+2} - y_{n+1}) = a_n a_{n+1}(y_n - y_{n-1}) \quad (10.100)$$



Поскольку из (10.98) следует, что

$$y_n - y_{n-1} = (1 - \gamma x_n) - (1 - \gamma x_{n-1}) = -\gamma (x_n - x_{n-1}) = -\gamma h, \quad (10.101)$$

$$y_{n+2} - y_{n+1} = \delta (1 - x_{n+2}) - \delta (1 - x_{n+1}) = -\delta (x_{n+2} - x_{n+1}) = -\delta h,$$

то из (10.100) находим:

$$b_{n+1} b_n (-\delta h) = a_n a_{n+1} (-\gamma h),$$

т.е.

$$\delta = \frac{a_n a_{n+1}}{b_n b_{n+1}} \gamma \quad (10.102)$$

Подставляя найденную связь, например, в первое из уравнений системы (10.99) и учитывая, что

$$y_{n+1} - y_n = \delta (1 - x_{n+1}) - (1 - \gamma x_n),$$

будем иметь:

$$b_n \left(\gamma \frac{a_n a_{n+1}}{b_n b_{n+1}} (1 - x_{n+1}) - 1 + \gamma x_n \right) + a_n \gamma h = 0,$$

или

$$\gamma \left[\frac{a_n a_{n+1}}{b_n b_{n+1}} (1 - x_{n+1}) + x_n + \frac{a_n}{b_n} h \right] = 1,$$

т.е.

$$\gamma = \frac{1}{\frac{a_n}{b_n} h + x_n + \frac{a_n a_{n+1}}{b_n b_{n+1}} (1 - x_{n+1})} \quad (10.103)$$

Таким образом, разностное решение в узлах сетки определяется однозначно формулами (10.98), (10.102), (10.103). Распространим его на весь отрезок $[0, 1]$ путем линейной интерполяции. Применим для этих целей, например, [интерполяционный многочлен в форме Ньютона](#):

$$\tilde{y}(x, h) = y_i + \frac{x - x_i}{h} (y_{i+1} - y_i), \quad x \in [x_i, x_{i+1}].$$



Тогда, используя формулы (10.98) и (10.101), при всех $i \leq n - 1$ будем иметь

$$\tilde{y}(x, h) = 1 - \gamma x_i + \frac{x - x_i}{h} (-\gamma h) = 1 - \gamma x,$$

а при $i \geq n + 1$ —

$$\tilde{y}(x, h) = \delta(1 - x_i) + \frac{x - x_i}{h} (-\delta h) = \delta(1 - x).$$

На отрезке $[x_n, x_{n+1}]$ доопределение выполним естественным образом: слева от точки ξ — как для $i \leq n - 1$, а справа — как для $i \geq n + 1$. Таким образом,

$$\tilde{y}(x; h) = \begin{cases} 1 - \gamma x, & 0 \leq x \leq \xi, \\ \delta(1 - x), & \xi \leq x \leq 1. \end{cases}$$

Сходимость найденного приближенного решения к точному решению $u(x)$ задачи, т.е. выполнение соотношения $\tilde{y}(x; h) \xrightarrow[h \rightarrow 0]{} u(x)$, равносильна, учитывая вид найденного выше точного решения, выполнению соотношений

$$\gamma \xrightarrow[h \rightarrow 0]{} \gamma_0, \quad \delta \xrightarrow[h \rightarrow 0]{} \delta_0.$$

Последние же соотношения с учетом вида всех входящих в них констант, выполняются при условии

$$\frac{a_n a_{n+1}}{b_n b_{n+1}} \xrightarrow[h \rightarrow 0]{} \frac{k_1}{k_2}$$

или

$$R_n = \frac{b_n b_{n+1}}{k_2} - \frac{a_n a_{n+1}}{k_1} \xrightarrow[h \rightarrow 0]{} 0 \tag{10.104}$$

Чтобы проверить полученное условие сходимости, зададим конкретный вид шаблонных функционалов $A(k(x_i + sh))$ и $B(k(x_i + sh))$, ибо входящие в него коэффициенты разностного уравнения только лишь приведены выше условиями аппроксимации не определяются. Пусть, например, это будут рассмотренные ранее линейные комбинации значений функции в трех соседних узлах сетки, т.е.

$$\begin{cases} a_i = \alpha_{-1} k_{i-1} + \alpha_0 k_i + \alpha_1 k_{i+1}, \\ b_i = \beta_{-1} k_{i-1} + \beta_0 k_i + \beta_1 k_{i+1}. \end{cases}$$



Тогда обсуждавшиеся выше условия аппроксимации (10.90) примут вид

$$\begin{cases} \alpha_{-1} + \alpha_0 + \alpha_1 = 1, \\ \beta_{-1} + \beta_0 + \beta_1 = 1, \\ \beta_1 - \beta_{-1} = 1 + \alpha_1 - \alpha_{-1}. \end{cases} \quad (10.105)$$

При этом из соображений разрешимости, как отмечалось ранее, все коэффициенты α_i и β_i должны быть неотрицательными.

При таком выборе функционалов получим:

$$\begin{cases} a_i = k_1, & i = \overline{1, n-1}, \\ a_i = k_2, & i = \overline{n+2, N-1}, \\ a_n = \alpha_{-1}k_1 + \alpha_0k_1 + \alpha_1k_2 = (\alpha_{-1} + \alpha_0)k_1 + \alpha_1k_2 = (1 - \alpha_1)k_1 + \alpha_1k_2, \\ a_{n+1} = \alpha_{-1}k_1 + \alpha_0k_2 + \alpha_1k_2 = \alpha_{-1}k_1 + (\alpha_0 + \alpha_1)k_2 = \alpha_{-1}k_1 + (1 - \alpha_{-1})k_2, \end{cases}$$

$$\begin{cases} b_i = k_1, & i = \overline{1, n-1}, \\ b_i = k_2, & i = \overline{n+2, N-1}, \\ b_n = \beta_{-1}k_1 + \beta_0k_1 + \beta_1k_2 = (\beta_{-1} + \beta_0)k_1 + \beta_1k_2 = (1 - \beta_1)k_1 + \beta_1k_2, \\ b_{n+1} = \beta_{-1}k_1 + \beta_0k_2 + \beta_1k_2 = \beta_{-1}k_1 + (\beta_0 + \beta_1)k_2 = \beta_{-1}k_1 + (1 - \beta_{-1})k_2. \end{cases}$$

Отсюда видим, что поскольку коэффициенты a_n , a_{n+1} , b_n , b_{n+1} не зависят от шага h , то полученное выше условие сходимости (10.104) равносильно равенству

$$R_n = \frac{b_n b_{n+1}}{k_2} - \frac{a_n a_{n+1}}{k_1} = 0.$$

Подставляя сюда полученные выражения коэффициентов a_n , a_{n+1} , b_n , b_{n+1} , будем иметь:

$$\begin{aligned} \frac{1}{k_2} [(1 - \beta_1) k_1 + \beta_1 k_2] [\beta_{-1} k_1 + (1 - \beta_{-1}) k_2] - \frac{1}{k_1} [(1 - \alpha_1) k_1 + \alpha_1 k_2] [\alpha_{-1} k_1 + (1 - \alpha_{-1}) k_2] = \\ = \left[\text{введем обозначение } t = \frac{k_1}{k_2} \right] = \\ = \frac{k_2^2}{k_2} [(1 - \beta_1) t + \beta_1] [\beta_{-1} t + (1 - \beta_{-1})] - \frac{k_2^2}{k_1} [(1 - \alpha_1) t + \alpha_1] [\alpha_{-1} t + (1 - \alpha_{-1})] = \\ = \frac{k_2^2}{k_1} \{t [(1 - \beta_1) t + \beta_1] [\beta_{-1} t + (1 - \beta_{-1})] - [(1 - \alpha_1) t + \alpha_1] [\alpha_{-1} t + (1 - \alpha_{-1})]\} = 0. \end{aligned}$$

Раскрывая скобки и отбрасывая отличный от нуля коэффициент $\frac{k_2^2}{k_1}$, рассмотрим это равенство как многочлен по переменной t :

$$\begin{aligned} \beta_{-1} (1 - \beta_1) t^3 + [\beta_1 \beta_{-1} + (1 - \beta_1) (1 - \beta_{-1}) - \alpha_{-1} (1 \alpha_1)] t^2 + \\ + [\beta_1 (1 - \beta_{-1}) - (1 - \alpha_1) (1 - \alpha_{-1}) - \alpha_1 \alpha_{-1}] t - \alpha_1 (1 - \alpha_{-1}) = 0. \end{aligned}$$

Требуя тождественного по t равенства нулю, приравниваем нулю коэффициенты при всех степенях t . В итоге получим систему уравнений, связывающую параметры α и β :

$$\begin{cases} \beta_{-1} (1 - \beta_1) = 0, \\ \alpha_1 (1 - \alpha_{-1}) = 0, \\ \beta_1 \beta_{-1} + (1 - \beta_1) (1 - \beta_{-1}) - \alpha_{-1} (1 - \alpha_1) = 0, \\ \beta_1 (1 - \beta_{-1}) - (1 - \alpha_1) (1 - \alpha_{-1}) - \alpha_1 \alpha_{-1} = 0. \end{cases} \quad (10.106)$$

Исследуем решения этой системы. Из первого уравнения имеем: либо $\beta_{-1} = 0$, либо $\beta_1 = 1$. В первом случае вновь возможны две версии:

a) $\alpha_{-1} = 1$. Тогда оставшиеся уравнения системы перепишутся в виде

$$\begin{cases} 1 - \beta_1 - 1 + \alpha_1 = 0, \\ \beta_1 - \alpha_1 = 0, \end{cases}$$



откуда $\beta_1 = \alpha_1$;

б) $\alpha_1 = 0$. Аналогично предыдущему случаю имеем: оставшиеся уравнения отличаются только знаком и, следовательно, $\beta_1 = 1 - \alpha_{-1}$;

Во втором случае также имеем две версии:

в) $\alpha_{-1} = 1$. Тогда оставшиеся уравнения системы перепишутся в виде

$$\begin{cases} \beta_{-1} - 1 + \alpha_1 = 0, \\ 1 - \beta_{-1} - \alpha_1 = 0, \end{cases}$$

откуда $\beta_{-1} = 1 - \alpha_1$;

г) $\alpha_1 = 0$. Аналогично предыдущему случаю имеем: оставшиеся уравнения отличаются только знаком и, следовательно, $\beta_{-1} = \alpha_{-1}$.

Вспомним теперь, что помимо системы (10.106) коэффициенты α_i и β_i должны также удовлетворять системе условий порядка (10.105), а также условию положительности соответствующих функционалов.

В случае а) система (10.105) примет вид

$$\begin{cases} 1 + \alpha_0 + \alpha_1 = 1, \\ 0 + \beta_0 + \beta_1 = 1, \\ \beta_1 - 0 = 1 + \alpha_1 - 1, \end{cases}$$

откуда, учитывая, что $\alpha_i \geq 0$, имеем: $\alpha_0 = \alpha_1 = 0$, $\beta_1 = 0$, $\beta_0 = 1$. Таким образом, в этом варианте

$$a_i = k_{i-1}, \quad b_i = k_i = a_{i+1}.$$

В случае б) аналогично имеем:

$$\begin{cases} \alpha_{-1} + \alpha_0 + 0 = 1, \\ 0 + \beta_0 + \beta_1 = 1, \\ \beta_1 - 0 = 1 + 0 - \alpha_{-1}, \end{cases}$$

откуда $\alpha_0 = 1 - \alpha_{-1}$, $\beta_0 = \alpha_{-1}$, $\beta_1 = 1 - \alpha_{-1}$, т.е.

$$a_i = \alpha_{-1} k_{i-1} + (1 - \alpha_{-1}) k_i, \quad b_i = \alpha_{-1} k_i + (1 - \alpha_{-1}) k_{i+1} = a_{i+1}.$$

В случае в):

$$\begin{cases} \alpha_{-1} + \alpha_0 + 0 = 1, \\ \beta_{-1} + \beta_0 + 1 = 1, \\ 1 - \beta_{-1} = 1 + 0 - \alpha_{-1}, \end{cases}$$

откуда $\alpha_{-1} = 0$, $\alpha_0 = 1$, $\beta_{-1} = \beta_0 = 0$, т.е.

$$a_i = k_i, \quad b_i = k_{i+1} = a_{i+1}.$$

Наконец, в случае г):

$$\begin{cases} 1 + \alpha_0 + \alpha_1 = 1, \\ \beta_{-1} + \beta_0 + 1 = 1, \\ 1 - \beta_{-1} = 1 + \alpha_1 - 1, \end{cases}$$

откуда $\beta_0 = \beta_{-1} = 0$, $\alpha_1 = 1$, $\alpha_0 = -1$. Полученное решение противоречит требованию положительности функционала $A(k(x_i + sh))$.

Окончательно получаем, что во всех случаях коэффициенты разностной схемы удовлетворяют условию консервативности (10.94).

В то же время простейшая разностная схема второго порядка, полученная путем раскрытия скобок в уравнении (10.95) с последующей заменой производных разностными отношениями, условию сходимости (10.104) не удовлетворяет.

Действительно, переходя к уравнению

$$ku'' + k'u' = 0$$

и выполняя замену производных, получим разностную схему

$$k_i \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + \frac{k_{i+1} - k_{i-1}}{2h} \cdot \frac{y_{i+1} - y_{i-1}}{2h} = 0 \quad (10.107)$$

или

$$\frac{1}{h} \left(b_i \frac{y_{i+1} - y_i}{h} - a_i \frac{y_i - y_{i-1}}{h} \right) = 0,$$



где

$$a_i = k_i - \frac{k_{i+1} - k_{i-1}}{4}, \quad b_i = k_i + \frac{k_{i+1} - k_{i-1}}{4}.$$

Отсюда, во-первых, следует, что такие коэффициенты не всегда будут положительны. В самом деле, так как $a_n = k_1 - \frac{k_2 - k_1}{4} = \frac{5k_1 - k_2}{4}$, то отсюда следует ограничение $\frac{k_1}{k_2} > \frac{1}{5}$. С другой стороны, $b_{n+1} = k_2 + \frac{k_2 - k_1}{4} = \frac{5k_2 - k_1}{4}$, откуда $\frac{k_1}{k_2} < 5$. Таким образом, разрешимость разностной схемы (10.107) гарантирована только при $\frac{k_1}{k_2} \in (\frac{1}{5}; 5)$.

Проверим теперь выполнение условия сходимости (10.104):

$$R_n = \frac{b_n b_{n+1}}{k_2} - \frac{a_n a_{n+1}}{k_1} = \frac{(3k_1 + k_2)(5k_2 - k_1)}{16k_2} - \frac{(5k_1 - k_2)(3k_2 + k_1)}{16k_1} = \frac{3(k_2 - k_1)^3}{16k_1 k_2} = 0.$$

Последнее же равенство возможно только лишь в случае $k_1 = k_2$.

Таким образом, консервативность разностной схемы является необходимым условием ее сходимости в классе кусочно-непрерывных коэффициентов.

Замечание 10.6. Можно показать, что это — также и достаточное условие.



10.2.2. Интегро-интерполяционный метод построения разностных схем

Апроксимация и сходимость построенной разностной схемы

Рассмотрим сейчас один из способов построения разностных схем для дифференциального уравнения с переменными коэффициентами, который позволяет автоматически удовлетворить требованию консервативности при наличии у исходного дифференциального оператора требования самосопряженности. Ранее мы выяснили, что соответствующее дифференциальное уравнение второго порядка должно иметь вид

$$Lu(x) \equiv (k(x) u'(x))' - q(x) u(x) = -f(x), \quad 0 < x < 1 \quad (10.108)$$

Будем рассматривать уравнение (10.108) как уравнение стационарного распределения тепла в стержне. Для него справедлив закон сохранения тепла (уравнение баланса), который на отрезке $[x^{(1)}, x^{(2)}]$ имеет вид

$$W(x^{(1)}) - W(x^{(2)}) - \int_{x^{(1)}}^{x^{(2)}} q(x) u(x) dx + \int_{x^{(1)}}^{x^{(2)}} f(x) dx = 0 \quad (10.109)$$

Это уравнение, очевидно, может быть получено путем интегрирования исходного дифференциального уравнения (10.108) по указанному отрезку. Здесь $W(x) = -k(x) \frac{du(x)}{dx}$ — тепловой поток, $k(x) > 0$ — коэффициент теплопроводности, $u(x)$ — температура.

Воспользуемся уравнением (10.109) для написания разностной схемы, аппроксимирующей дифференциальное уравнение (10.108). Пусть на отрезке $[0, 1]$ задана сетка $\bar{\omega}_h$ и $x_{i-0.5} = x_i - 0.5h$, $x_{i+0.5} = x_i + 0.5h$. Запишем уравнение баланса (10.109) для отрезка $[x_{i-0.5}, x_{i+0.5}]$:

$$W_{i-0.5} - W_{i+0.5} - \int_{x_{i-0.5}}^{x_{i+0.5}} q(x) u(x) dx + \int_{x_{i-0.5}}^{x_{i+0.5}} f(x) dx = 0 \quad (10.110)$$

Чтобы построить разностную схему, аппроксимируем тепловой поток и первый из интегралов в



(10.110). Заменим функцию $u(x)$ на отрезке $[x_{i-0.5}, x_{i+0.5}]$ интерполяционным многочленом нулевой степени: $u(x) \approx P_0(x) = u(x_i) =: u_i$, $x \in [x_{i-0.5}, x_{i+0.5}]$. Тогда

$$\int_{x_{i-0.5}}^{x_{i+0.5}} q(x) u(x) dx \approx h d_i u_i, \quad 345 \quad d_i = \frac{1}{h} \int_{x_{i-0.5}}^{x_{i+0.5}} q(x) u(x) dx, \quad i = \overline{1, N-1} \quad (10.111)$$

Теперь займемся тепловым потоком. Выразив производную от температуры:

$$\frac{du(x)}{dx} = -\frac{W(x)}{k(x)},$$

проинтегрируем последнее равенство по отрезку $[x_{i-1}, x_i]$:

$$u_{i-1} - u_i = \int_{x_{i-1}}^{x_i} \frac{W(x)}{k(x)} dx \quad (10.112)$$

Так как в интегральное соотношение (10.110) тепловой поток входит в полуцелых точках, то, по аналогии с проделанной выше операцией по интерполированию температуры проинтерполируем $W(x)$ с помощью многочлена нулевой степени: $W(x) \approx P_0^*(x) = W_{i-0.5}$ при $x \in [x_{i-1}, x_i]$. Тогда из (10.112) получим:

$$u_i - u_{i-1} \approx -W_{i-0.5} \cdot \int_{x_{i-1}}^{x_i} \frac{dx}{k(x)}$$

или

$$W_{i-0.5} \approx -a_i \frac{u_i - u_{i-1}}{h} = -a_i u_{\bar{x},i} \quad (10.113)$$

где

$$a_i = \left[\frac{1}{h} \int_{x_{i-1}}^{x_i} \frac{dx}{k(x)} \right]^{-1}, \quad i = \overline{1, N} \quad (10.114)$$

Подставляя в (10.110) выражения (10.111) и (10.113), получим:

$$a_{i+1}u_{\bar{x},i+1} - a_iu_{\bar{x},i} - hd_iu_i + h\varphi_i \approx 0,$$

где

$$\varphi_i = \frac{1}{h} \int_{x_{i-0.5}}^{x_{i+0.5}} f(x) dx \quad (10.115)$$

Разделив полученное приближенное равенство на h и переходя к точному равенству для приближенных значений решения в узлах сетки, получим разностную схему:

$$\frac{1}{h} \left(a_{i+1} \frac{y_{i+1} - y_i}{h} - a_i \frac{y_i - y_{i-1}}{h} \right) - d_i y_i = -\varphi_i, \quad i = \overline{1, N-1}$$

или

$$(ay_{\bar{x}})_x - dy = -\varphi, \quad x \in \omega_h, \quad (10.116)$$

коэффициенты которой вычисляются по формулам (10.111), (10.114), (10.115).

Займемся теперь аппроксимацией граничных условий третьего рода для уравнения (10.108). Пусть это условие задано в точке $x = 0$ и имеет вид

$$u'(0) = \tilde{\kappa}_0 u(0) - \tilde{g}_0.$$

По предположению коэффициент $k(x)$ в (2.1) больше нуля. Поэтому перепишем данное условие в несколько более удобном виде

$$k(0)u'(0) = \kappa_0 u(0) - g_0. \quad (10.117)$$

Запись граничного условия в виде (10.117) более естественна, так как теперь в левой его части с точностью до знака стоит величина теплового потока. Для аппроксимации полученного условия запишем уравнение баланса (10.109) на отрезке $[0, \frac{h}{2}]$:

$$W_0 - W_{0.5} - \int_0^{\frac{h}{2}} q(x) u(x) dx + \int_0^{\frac{h}{2}} f(x) dx = 0. \quad (10.118)$$



Теперь из (10.117) имеем: $W_0 = -k(0)u'(0) = -\kappa_0 u(0) + g_0$. Для аппроксимации $W_{0.5}$ воспользуемся формулой (10.113) при $i = 1$, т.е. $W_{0.5} \approx -a_1 u_{\bar{x},1} = -a_1 u_{x,0}$, а также проинтерполируем функцию $u(x)$, стоящую под знаком первого интеграла, с помощью многочлена нулевой степени: $u(x) \approx P_0(x) = u_0$ при $x \in [0, \frac{h}{2}]$. С учетом сказанного получим разностную аппроксимацию граничного условия (10.117):

$$a_1 u_{x,0} = \left(\kappa_0 + \frac{h}{2} d_0 \right) y_0 - \left(g_0 + \frac{h}{2} \varphi_0 \right) \quad (10.119)$$

где

$$d_0 = \frac{2}{h} \int_0^{\frac{h}{2}} q(x) dx, \quad \varphi_0 = \frac{2}{h} \int_0^{\frac{h}{2}} f(x) dx \quad (10.120)$$

Аналогичным образом, записав граничное условие на правом конце отрезка в виде

$$-k(1)u'(1) = \kappa_1 u(1) - g_1 \quad (10.121)$$

и используя уравнение баланса (10.109), записанное для отрезка $[1 - \frac{h}{2}, 1]$, получим его разностную аппроксимацию в виде

$$-a_N u_{\bar{x},N} = \left(\kappa_1 + \frac{h}{2} d_N \right) y_N - \left(g_1 + \frac{h}{2} \varphi_N \right) \quad (10.122)$$

где

$$d_N = \frac{2}{h} \int_{1-\frac{h}{2}}^1 q(x) dx, \quad \varphi_N = \frac{2}{h} \int_{1-\frac{h}{2}}^1 f(x) dx \quad (10.123)$$

Таким образом, дифференциальная задача (10.108), (10.117), (10.121) аппроксимирована разностной схемой (10.116), (10.119), (10.122).

Аппроксимация и сходимость построенной разностной схемы

Для простоты изложения рассмотрим случай, когда коэффициенты исходной дифференциальной задачи обладают достаточной гладкостью.



Определимся с видом шаблонных функционалов. Легко видеть, что

$$\frac{1}{a_i} = A \left(\frac{1}{k(x_i+sh)} \right) = \frac{1}{h} \int_{x_{i-1}}^{x_i} \frac{dx}{k(x)} = \int_{-1}^0 \frac{ds}{k(x_i+sh)}, \quad \frac{1}{a_{i+1}} = B \left(\frac{1}{k(x_i+sh)} \right),$$

$$d_i = F(q(x_i + sh)) = \int_{-0.5}^{0.5} q(x_i + sh) ds, \quad \varphi_i = F(f(x_i + sh)) = \int_{-0.5}^{0.5} f(x_i + sh) ds, \quad i = \overline{1, N-1}.$$

Поэтому

$$A(1) = \int_{-1}^0 ds = 1, \quad A(s) = \int_{-1}^0 s ds = -\frac{1}{2}, \quad B(s) = A(s+1) = \frac{1}{2}, \quad B(s^2) - A(s^2) = \int_0^1 s^2 ds - \int_{-1}^0 s^2 ds = 0,$$

$$F(1) = \int_{-0.5}^{0.5} ds = 1, \quad F(s) = \int_{-0.5}^{0.5} s ds = 0.$$

Отсюда, учитывая условия (10.89), делаем вывод, что разностное уравнение (10.116) имеет второй порядок аппроксимации.

Для левого граничного условия непосредственно получаем:

$$\nu(0) = a_1 u_{x,0} - \left(\kappa_0 + \frac{h}{2} d_0 \right) u(0) + \left(g_0 + \frac{h}{2} \varphi_0 \right),$$

а поскольку

$$\begin{aligned} a_1 u_{x,0} &= \left[k\left(\frac{h}{2}\right) + O(h^2) \right] \cdot \frac{u(h)-u(0)}{h} = \left[k\left(\frac{h}{2}\right) + O(h^2) \right] \cdot \left[u'\left(\frac{h}{2}\right) + O(h^2) \right] = \\ &= -W\left(\frac{h}{2}\right) + O(h^2) = - \left[W(0) + \frac{h}{2} W'(0) + O(h^2) \right] = k(0) u'(0) + \frac{h}{2} (k u')'(0) + O(h^2), \end{aligned}$$

то, используя уравнение (10.108), имеем:

$$\begin{aligned} \nu(0) &= \kappa_0 u(0) - g_0 + \frac{h}{2} (q(0) u(0) - f(0)) - \left(\kappa_0 + \frac{h}{2} d_0 \right) u(0) + g_0 + \frac{h}{2} \varphi_0 = \\ &= \frac{h}{2} (q(0) - d_0) u(0) + \frac{h}{2} (\varphi_0 - f(0)) + O(h^2). \end{aligned}$$



Так как

$$q(0) - d_0 = q(0) - \frac{2}{h} \int_0^{\frac{h}{2}} q(x) dx = q(0) - \frac{2}{h} \left[\frac{h}{2} q(0) + O(h^2) \right] = O(h)$$

и аналогично $\varphi_0 - f(0) = O(h)$, то $\nu(0) = O(h^2)$, т.е. левое разностное граничное условие аппроксимирует условие (10.117) со вторым порядком.

Аналогичным образом показывается, что правое граничное условие также имеет второй порядок.

Таким образом, разностная схема, построенная с помощью рассмотренного варианта метода баланса, в классе достаточно гладких коэффициентов обладает вторым порядком аппроксимации.

Исследуем теперь вопрос о сходимости данной разностной схемы (сделав, однако, предположение о том, что на правом конце отрезка задано условие первого рода, так как это несколько упрощает выкладки). Тогда задача для погрешности $z_h = y_h - u_h$ будет выглядеть следующим образом:

$$\begin{cases} (az_{\bar{x}})_x - dz = -\psi, & x \in \omega_h, \\ a_1 z_{x,0} = \tilde{\kappa}_0 z_0 - \nu(0), \\ z_N = 0. \end{cases}$$

Умножим разностное уравнение скалярно на сеточную функцию z :

$$((az_{\bar{x}})_x, z) - (d, z^2) = -(\psi, z).$$

Применяя первую разностную формулу Грина и учитывая условие $z_N = 0$, получим:

$$-(a, z_{\bar{x}}^2) - a_1 z_{x,0} z_0 - (d, z^2) = -(\psi, z)$$

или, так как $a_1 z_{x,0} = \tilde{\kappa}_0 z_0 - \nu(0)$,

$$-(a, z_{\bar{x}}^2) - \tilde{\kappa}_0 z_0^2 - (d, z^2) = -(\psi, z) - \nu(0) z_0.$$

Умножая на -1 , перепишем последнее соотношение в виде

$$(a, z_{\bar{x}}^2) + \tilde{\kappa}_0 z_0^2 + (d, z^2) = (\psi, z) + \nu(0) z_0 \quad (10.124)$$

По предположению $k(x) \geq c_1 > 0$, $q(x) \geq 0$, $\kappa_0 \geq 0$. Поэтому

$$c_1 \|z_{\bar{x}}\|^2 \geq c_1 \|z_{\bar{x}}\|^2, \quad (d, z^2) \geq 0, \quad \tilde{\kappa}_0 z_0^2 \geq 0.$$

Следовательно, из (10.124) получаем:

$$c_1 \|z_{\bar{x}}\|^2 \leq |(\psi, z)| + |\nu(0)| |z_0|.$$

Оценим сверху правую часть этого неравенства:

$$|(\psi, z)| + |\nu(0)| |z_0| \leq \sum_{i=1}^{N-1} h |\psi_i| |z_i| + |\nu(0)| |z_0| \leq \|z\|_C \left(\|\psi\|_C \sum_{i=1}^{N-1} h + |\nu(0)| \right) = \|z\|_C (\|\psi\|_C + |\nu(0)|).$$

С другой стороны, для сеточной функции z , обращающейся в нуль при $x = l$ (т.е. $z_N = 0$) справедлив аналог теоремы вложения $\|z\|_C \leq \sqrt{l} \|z_{\bar{x}}\|$. Поэтому имеем:

$$c_1 \|z_{\bar{x}}\|^2 \geq \frac{c_1}{l} \|z\|_C^2 = c_1 \|z\|_C^2$$

и, следовательно,

$$c_1 \|z\|_C^2 \leq \|z\|_C (\|\psi\|_C + |\nu(0)|),$$

откуда

$$\|z\|_C \leq \frac{1}{c_1} (\|\psi\|_C + |\nu(0)|).$$

Из полученного неравенства следует сходимость данной разностной схемы в норме C со вторым порядком в классе гладких коэффициентов.

Заметим, что в данном доказательстве важны, по сути, только формулы, задающие второй порядок аппроксимации. Поэтому для практических целей удобно иметь возможно более простые формулы для нахождения сеточных функций a , d , φ , использующие значения коэффициентов исходного уравнения $k(x)$, $q(x)$, $f(x)$ в отдельных точках. Обычно используют шаблон из одной или двух точек, полагая, например,

$$a_i = k_{i-0.5} = k\left(x_i - \frac{h}{2}\right) \quad \text{здесь} \quad (A(\bar{k}(s)) = \bar{k}(-0.5)),$$

$$d_i = q_i, \quad \varphi_i = f_i \quad (F(\bar{f}(s)) = \bar{f}(0))$$



или

$$a_i = \frac{k_i + k_{i-1}}{2} \quad \left(A(\bar{k}(s)) = \frac{1}{2} (\bar{k}(-1) + \bar{k}(0)) \right),$$

или

$$a_i = \frac{2k_i k_{i-1}}{k_i + k_{i-1}} \quad \left(\frac{1}{A(\bar{k}(s))} = \frac{1}{2} \left(\frac{1}{\bar{k}(-1)} + \frac{1}{\bar{k}(0)} \right) \right).$$

Фактически речь идет о замене интегралов в формулах (10.111), (10.114), (10.115) квадратурными формулами, обладающими необходимой точностью (в том числе — нелинейными).

Замечание 10.7. Все полученные при этом разностные схемы будут сходиться (но не в норме $C(!)$) также и в классе кусочно-гладких коэффициентов, однако скорость сходимости станет равной единице, и только исходная разностная схема с интегральным представлением коэффициентов сохраняет второй порядок. Поэтому ее называют *наилучшей* консервативной однородной разностной схемой.



10.2.3. Вариационно-проекционные подходы к построению разностных схем

[Метод Ритца построения разностных схем](#)

[Метод аппроксимации квадратичного функционала](#)

[Метод Галеркина построения разностных схем](#)

[Метод аппроксимации интегрального тождества](#)

Метод Ритца построения разностных схем

Рассмотренный выше способ построения разностных схем автоматически приводит к консервативным разностным схемам, если оператор исходной дифференциальной задачи является самосопряженным. Изложим сейчас еще один способ, позволяющий добиться такого же эффекта.

Ранее мы рассматривали классический вариант [метода Ритца решения граничных задач](#). Напомним, что решение всякого операторного уравнения с самосопряженным положительным оператором может быть сведено к эквивалентной задаче отыскания функции, доставляющей минимум некоторому функционалу (функционалу Ритца). Так, например, нахождение решения краевой задачи

$$\begin{cases} (k(x)u'(x))' - q(x)u(x) = -f(x), \quad 0 < x < 1, \quad k(x) \geq c_1 > 0, \quad q(x) \geq 0, \\ k(0)u'(0) = \kappa_0 u(0) - g_0, \quad \kappa_0 \geq 0, \\ -k(1)u'(1) = \kappa_1 u(1) - g_1, \quad \kappa_1 \geq 0, \end{cases} \quad (10.125)$$

эквивалентно задаче отыскания функции $u(x)$, доставляющей минимум функционалу

$$J(u) = \frac{1}{2} [u, u] - \int_0^1 f(x)u(x) dx - g_0 u(0) - g_1 u(1) \quad (10.126)$$

где

$$[u, v] = \int_0^1 [k(x)u'(x)v'(x) + q(x)u(x)v(x)] dx + \kappa_0 u(0)v(0) + \kappa_1 u(1)v(1) \quad (10.127)$$



для которого уравнение (10.125) есть уравнение Эйлера.

Известно, что если входные параметры задачи удовлетворяют указанным в (10.125) условиям, то минимум функционала (10.126) существует и соответствующий элемент принадлежит пространству $W_2^1[0, 1]$. В соответствии с идеей Ритца построим последовательность конечномерных подпространств $V_n \in W_2^1$ и вместо того, чтобы искать минимум на W_2^1 , будем искать его на V_n . Пусть размерность подпространства V_n равна n и $\eta_i^{(n)}$, $i = \overline{0, n-1}$ — базис этого подпространства, т.е. любой элемент u_n этого подпространства представим в виде

$$u_n = \sum_{i=0}^{n-1} a_i \eta_i^{(n)} \quad (10.128)$$

Подставляя записанное представление для u_n вместо u в функционал $I(u)$, получим функцию n переменных a_0, a_1, \dots, a_{n-1} . Так как мы желаем получить минимум этой функции, то числа a_i должны удовлетворять системе уравнений

$$\frac{\partial J(u_n)}{\partial a_i} = 0, \quad i = \overline{0, n-1} \quad (10.129)$$

Решив эту систему, мы получим определенные значения параметров a_0, a_1, \dots, a_{n-1} , дающие $I(u_n)$ абсолютный минимум, а затем по формуле (10.128) получим требуемое приближенное решение. Найдем вид системы (10.128), исходя из конкретного функционала (10.126):

$$J(u_n) = \frac{1}{2} \left[\sum_{i=0}^{n-1} a_i \eta_i^{(n)}, \sum_{i=0}^{n-1} a_i \eta_i^{(n)} \right] - \int_0^1 f(x) \sum_{i=0}^{n-1} a_i \eta_i^{(n)}(x) dx - g_0 \sum_{i=0}^{n-1} a_i \eta_i^{(n)}(0) - g_1 \sum_{i=0}^{n-1} a_i \eta_i^{(n)}(1).$$

Тогда

$$\frac{\partial J(u_n)}{\partial a_j} = \sum_{i=0}^{n-1} a_i [\eta_i^{(n)}, \eta_j^{(n)}] - \int_0^1 f(x) \eta_j^{(n)}(x) dx - g_0 \eta_j^{(n)}(0) - g_1 \eta_j^{(n)}(1) = 0$$

или

$$\sum_{j=0}^{n-1} \alpha_{ij} a_j = \beta_i, \quad i = \overline{0, n-1} \quad (10.130)$$



где

$$\alpha_{ij} = \left[\eta_i^{(n)}, \eta_j^{(n)} \right], \quad \beta_i = \int_0^1 f(x) \eta_i^{(n)}(x) dx + g_0 \eta_i^{(n)}(0) + g_1 \eta_i^{(n)}(1) \quad (10.131)$$

Поскольку наша цель — построение разностной схемы с помощью метода Ритца, то достичь этой цели можно с помощью специального подбора координатных функций.

Пусть на отрезке $[0, 1]$ задана сетка $\bar{\omega}_h$ и $N = n - 1$. Тогда система (10.130) будет иметь вид трехточечной разностной схемы, если матрица этой системы будет трехдиагональной, т.е. если коэффициенты α_{ij} будут равны нулю при $|i - j| > 1$. При этих условиях система (10.130) будет классической разностной схемой, если в качестве параметров в разложении (10.128) будут выбраны значения функции u_n в узлах сетки $\bar{\omega}_h$.

Матрица системы (10.130) будет трехдиагональной, если базисные функции подпространства V_n при $|i - j| \geq 2$ будут ортогональны в смысле скалярного произведения (10.127). Это условие, очевидно, будет выполнено, если в качестве $\eta_i^{(N+1)}(x)$ взять функции, которые отличны от нуля только при $|x - x_i| \leq h$, $x \in [0, 1]$ (такие функции называются *функциями с конечным носителем или финитными*). Так как значения u_{N+1} в узле являются коэффициентами разложения, то должны выполняться условия $\eta_i^{(N+1)}(x_i) = 1$. Простейшими функциями указанного вида, принадлежащими W_2^1 , являются функции

$$\eta_i^{(N+1)}(x) = \begin{cases} 0, & \text{если } x \notin [x_{i-1}, x_{i+1}], \\ \frac{x - x_{i-1}}{h}, & \text{если } x \in [x_{i-1}, x_i], \quad i = \overline{1, N-1}, \\ \frac{x_{i+1} - x}{h}, & \text{если } x \in [x_i, x_{i+1}], \end{cases}$$

$$\eta_0^{(N+1)}(x) = \begin{cases} \frac{h-x}{h}, & \text{если } x \in [0, h], \\ 0, & \text{если } x \notin [0, h], \end{cases} \quad (10.132)$$

$$\eta_N^{(N+1)}(x) = \begin{cases} \frac{x-1+h}{h}, & \text{если } x \in [1-h, 1], \\ 0, & \text{если } x \notin [1-h, 1]. \end{cases}$$

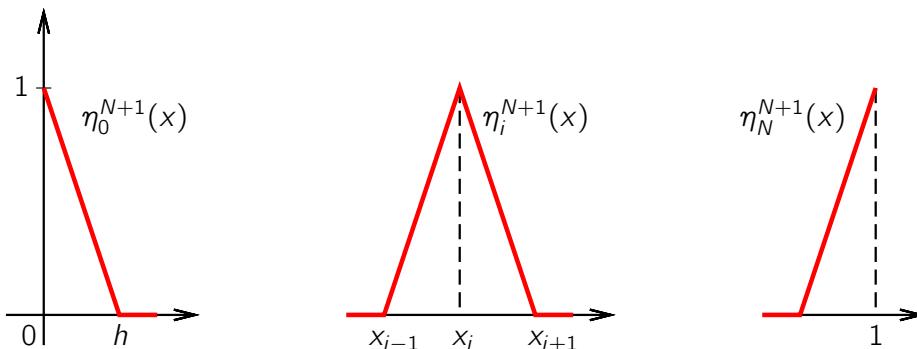


Рисунок 10.8

Если координатные функции $\eta_i^{(N+1)}(x)$ выбрать по формулам (10.132), то система (10.130) примет вид

$$\begin{cases} \alpha_{ii-1}y_{i-1} + \alpha_{ii}y_i + \alpha_{ii+1}y_{i+1} = \beta_i, & i = \overline{1, N-1}, \\ \alpha_{00}y_0 + \alpha_{01}y_1 = \beta_0, \\ \alpha_{NN-1}y_{N-1} + \alpha_{NN}y_N = \beta_N. \end{cases} \quad (10.133)$$

Пользуясь формулами (10.131), (10.132), вычислим коэффициенты α_{ii} и β_i :

$$\alpha_{ii} = \frac{1}{h^2} \left[\int_{x_{i-1}}^{x_{i+1}} k(x) dx + \int_{x_{i-1}}^{x_i} q(x)(x - x_{i-1})^2 dx + \int_{x_i}^{x_{i+1}} q(x)(x_{i+1} - x)^2 dx \right], \quad i = \overline{1, N-1},$$

$$\alpha_{00} = \frac{1}{h^2} \left[\int_0^h k(x) dx + \int_0^h q(x)(x - h)^2 dx \right] + \kappa_0,$$



$$\alpha_{NN} = \frac{1}{h^2} \left[\int_{1-h}^1 k(x) dx + \int_{1-h}^1 q(x)(x-1+h)^2 dx \right] + \kappa_1,$$

$$\alpha_{ii+1} = \alpha_{i+1i} = \frac{1}{h^2} \left[- \int_{x_i}^{x_{i+1}} k(x) dx + \int_{x_i}^{x_{i+1}} q(x)(x_{i+1}-x)(x-x_i) dx \right], \quad i = \overline{1, N-1},$$

$$\beta_i = \frac{1}{h} \left[\int_{x_{i-1}}^{x_i} f(x)(x-x_{i-1}) dx + \int_{x_i}^{x_{i+1}} f(x)(x_{i+1}-x) dx \right], \quad i = \overline{1, N-1},$$

$$\beta_0 = \frac{1}{h} \int_0^h f(x)(h-x) dx + g_0,$$

$$\beta_N = \frac{1}{h} \int_{1-h}^1 f(x)(x-1+h) dx + g_1.$$

Полученную систему уравнений (10.133) легко записать в стандартном для однородных консервативных схем виде (10.116), (10.119), (10.122). Действительно, разностное уравнение (10.116) в развернутом виде выглядит следующим образом:

$$\frac{a_i}{h^2} y_{i-1} - \left(\frac{a_i + a_{i+1}}{h^2} + d_i \right) y_i + \frac{a_{i+1}}{h^2} y_{i+1} = -\varphi_i.$$

Чтобы получить такой вид из (10.133), очевидно, необходимо положить

$$a_i = -h\alpha_{ii-1}, \quad \varphi_i = \frac{1}{h}\beta_i, \quad d_i = \frac{1}{h}(\alpha_{ii-1} + \alpha_{ii} + \alpha_{ii+1}).$$

Аналогично разбираемся и с граничными условиями. Поэтому при

$$\begin{cases} a_i = \frac{1}{h} \left[\int_{x_{i-1}}^{x_i} k(x) dx - \int_{x_{i-1}}^{x_i} q(x)(x_i-x)(x-x_{i-1}) dx \right], \quad i = \overline{1, N}, \\ d_i = \frac{1}{h^2} \left[\int_{x_{i-1}}^{x_i} q(x)(x-x_{i-1}) dx - \int_{x_i}^{x_{i+1}} q(x)(x_{i+1}-x) dx \right], \quad i = \overline{1, N-1}, \\ d_0 = \frac{2}{h^2} \int_0^h q(x)(h-x) dx; \quad d_N = \frac{2}{h^2} \int_{1-h}^1 q(x)(x-1+h) dx, \\ \varphi_i = \frac{1}{h^2} \left[\int_{x_{i-1}}^{x_i} f(x)(x-x_{i-1}) dx - \int_{x_i}^{x_{i+1}} f(x)(x_{i+1}-x) dx \right], \quad i = \overline{1, N-1}, \\ \varphi_0 = \frac{2}{h^2} \int_0^h f(x)(h-x) dx; \quad \varphi_N = \frac{2}{h^2} \int_{1-h}^1 f(x)(x-1+h) dx \end{cases} \quad (10.134)$$



система (10.133) превратится в стандартную разностную схему

$$\begin{cases} (ay_{\bar{x}})_x - dy = -\varphi, \quad x \in \omega_h, \\ a_1 y_{x,0} = (\kappa_0 + \frac{h}{2} d_0) y_0 - (g_0 + \frac{h}{2} \varphi_0), \\ -a_N y_{\bar{x},N} = (\kappa_1 + \frac{h}{2} d_N) y_N - (g_1 + \frac{h}{2} \varphi_N). \end{cases} \quad (10.135)$$

Метод аппроксимации квадратичного функционала

В [предыдущем пункте](#) мы использовали эквивалентность задачи (10.125) задаче об отыскании минимума функционала (10.126), (10.127), который в «сборном» виде будет выглядеть следующим образом:

$$J(u) = \int_0^1 \left[k(x) (u'(x))^2 + q(x) u^2(x) \right] dx - 2 \int_0^1 f(x) u(x) dx + \kappa_0 u^2(0) + \kappa_1 u^2(1) - 2g_0 u(0) - 2g_1 u(1) \quad (10.136)$$

для построения разностной схемы методом Ритца. При этом в качестве стандартной идеи использовалась идея аппроксимации пространства, в котором отыскивался минимум функционала, подпространствами конечной размерности.

Сейчас же мы поступим по-другому: аппроксимируем не пространство, а сам функционал (10.136). Для этого заменим на сетке $\bar{\omega}_h$ интегралы, входящие в выражение (10.136), [квадратурными формулами](#), предварительно переписав его в виде

$$J(u) = \sum_{i=1}^N \int_{x_{i-1}}^{x_i} k(x) (u'(x))^2 dx + \sum_{i=1}^N \int_{x_{i-1}}^{x_i} (q(x) u^2(x) - 2f(x) u(x)) dx + \kappa_0 u^2(0) + \kappa_1 u^2(1) - 2g_0 u(0) - 2g_1 u(1).$$

После этого аппроксимируем интегралы следующим образом:

$\int_{x_{i-1}}^{x_i} k(x) (u'(x))^2 dx \approx a_i (u_{\bar{x},i})^2 \cdot h$, — аналог формулы [средних прямоугольников](#),

$\int_{x_{i-1}}^{x_i} [q(x) u^2(x) - 2f(x) u(x)] dx \approx \frac{h}{2} [(q_i u_i^2 - 2f_i u_i) + (q_{i-1} u_{i-1}^2 - 2f_{i-1} u_{i-1})]$ — [формула трапеций](#).



Здесь a_i — некоторый функционал, зависящий от коэффициента $k(x)$ на отрезке $[x_{i-1}, x_i]$, например,

$$a_i = \frac{1}{h} \int_{x_{i-1}}^{x_i} k(x) dx \text{ или } a_i = k_{i-0.5} \text{ и т.п.}$$

Тогда вместо $J(u)$ получим функционал

$$\begin{aligned} J_h(y) = \sum_{i=1}^N h a_i y_{\bar{x}, i}^2 + \sum_{i=1}^{N-1} h (q_i y_i^2 - 2 f_i y_i) + \frac{h}{2} [q_0 y_0^2 + q_N y_N^2 - 2 f_0 y_0 - 2 f_N y_N] + \\ + \kappa_0 y_0^2 + \kappa_1 y_N^2 - 2 g_0 y_0 - 2 g_N y_N \end{aligned}$$

или

$$\begin{aligned} J_h(y) = \sum_{i=1}^N h a_i y_{\bar{x}, i}^2 + \sum_{i=1}^{N-1} h (q_i y_i^2 - 2 f_i y_i) + (\kappa_0 + \frac{h}{2} q_0) y_0^2 + (\kappa_1 + \frac{h}{2} q_N) y_N^2 - \\ - 2(g_0 + \frac{h}{2} f_0) y_0 - 2(g_1 + \frac{h}{2} f_N) y_N. \end{aligned}$$

В результате имеем: $J_h(y)$ есть функция $(N+1)$ переменных y_i , и для того чтобы найти уравнения, определяющие точку ее минимума, необходимо приравнять нулю первые производные этой функции по переменным y_i :

$$\frac{\partial J_h(y)}{\partial y_i} = 0, \quad i = \overline{0, N}.$$

Тогда при $i = \overline{1, N-1}$ получим следующие уравнения:

$$2h a_{i+1} y_{\bar{x}, i+1} \cdot \left(-\frac{1}{h}\right) + 2h a_i y_{\bar{x}, i} \cdot \frac{1}{h} + 2h q_i y_i - 2h f_i = 0, \quad i = \overline{1, N-1}.$$

Аналогично при $i = 0$ будем иметь

$$2h a_1 y_{\bar{x}, 1} \cdot \left(-\frac{1}{h}\right) + 2 \left(\kappa_0 + \frac{h}{2} q_0\right) y_0 - 2 \left(g_0 + \frac{h}{2} f_0\right),$$

а при $i = N$ —

$$2h a_N y_{\bar{x}, N} \cdot \frac{1}{h} + 2 \left(\kappa_1 + \frac{h}{2} q_N\right) y_N - 2 \left(g_1 + \frac{h}{2} f_N\right).$$

От полученных соотношений очевиден переход к трехточечной разностной схеме

$$\begin{cases} (ay_{\bar{x}})_x - qy = -f, \quad x \in \omega_h, \\ a_1 y_{x,0} = (\kappa_0 + \frac{h}{2} q_0) y_0 - (g_0 + \frac{h}{2} f_0), \\ -a_N y_{\bar{x},N} = (\kappa_1 + \frac{h}{2} q_N) y_N - (g_1 + \frac{h}{2} f_N). \end{cases} \quad (10.137)$$

Разностная схема (10.137) при надлежащем выборе a_i имеет второй порядок аппроксимации.

Замечание 10.8. Описанную процедуру можно организовать и таким образом, чтобы, аналогично (10.137), ее коэффициенты были представимы в некоторой интегральной форме.

Метод Галеркина построения разностных схем

Очень близким к изложенному выше [методу Ритца](#), но имеющим несколько более широкую область применимости, является другой проекционный метод — метод Галеркина (Бубнова — Галеркина). Этот метод применим, в частности, и тогда, когда задача не является самосопряженной. Рассмотрим технику его использования для построения разностных схем на примере задачи

$$\begin{cases} \frac{d}{dx} \left(k(x) \frac{du(x)}{dx} \right) + r(x) \frac{du(x)}{dx} - q(x) u(x) = -f(x), \quad 0 < x < 1, \\ k(0) \frac{du(0)}{dx} = \kappa_0 u(0) - g_0, \\ -k(1) \frac{du(1)}{dx} = \kappa_1 u(1) - g_1. \end{cases} \quad (10.138)$$

В соответствии с [общей идеей метода Галеркина](#) коэффициенты a_i приближенного решения

$$u_n = \sum_{i=0}^{n-1} a_i \eta_i^{(n)}$$

находятся из условия ортогональности невязки $Lu_n - f$ ко всем базисным функциям $\eta_i^{(n)}$. В случае задачи (3.14) равенство $(Lu - f, v) = 0$ принимает вид (первое слагаемое под знаком интеграла, содержащее

вторые производные, интегрируем по частям и пользуемся граничными условиями при вычислении двойной подстановки)

$$\int_0^1 [k(x) u'(x) v'(x) - r(x) u'(x) v(x) + q(x) u(x) v(x) - f(x) v(x)] dx + \\ (10.139)$$

$$+ \kappa_0 u(0) v(0) + \kappa_1 u(1) v(1) - g_0 v(0) - g_1 v(1) = 0.$$

Выберем такое же подпространство координатных функций, как и в методе Ритца: оно определяется формулами (10.132). Тогда приближенное решение $u_{N+1}(x)$ примет вид

$$u_{N+1}(x) = \sum_{j=0}^N y_j \eta_j^{(N+1)}(x) \quad (10.140)$$

где y_j — приближенные решения задачи (3.14) в узлах сетки $\bar{\omega}_h$. Подставляя (10.140) в (10.139) и выбирая в качестве поверочной функции $v(x)$ $\eta_i^{(N+1)}(x)$, $i = \overline{0, N}$, имеем:

$$\sum_{j=0}^N \left\{ \int_0^1 \left[k(x) y_j \frac{d}{dx} \eta_j^{(N+1)}(x) \frac{d}{dx} \eta_i^{(N+1)}(x) - r(x) y_j \frac{d}{dx} \eta_j^{(N+1)}(x) \eta_i^{(N+1)}(x) + q(x) y_j \eta_j^{(N+1)}(x) \eta_i^{(N+1)}(x) - f(x) \frac{d}{dx} \eta_i^{(N+1)}(x) \right] dx + \kappa_0 y_j \eta_j^{(N+1)}(0) \eta_i^{(N+1)}(0) + \kappa_1 y_j \eta_j^{(N+1)}(1) \eta_i^{(N+1)}(1) - g_0 \eta_i^{(N+1)}(0) - g_1 \eta_i^{(N+1)}(1) \right\} = 0.$$

Учитывая вид координатных функций, отсюда получаем систему уравнений

$$\begin{cases} \alpha_{ii-1} y_{i-1} + \alpha_{ii} y_i + \alpha_{ii+1} y_{i+1} = \beta_i, & i = \overline{1, N-1}, \\ \alpha_{00} y_0 + \alpha_{01} y_1 = \beta_0, \\ \alpha_{NN-1} y_{N-1} + \alpha_{NN} y_N = \beta_N, \end{cases} \quad (10.141)$$



где

$$\alpha_{ii-1} = \frac{1}{h^2} \left[- \int_{x_{i-1}}^{x_i} k(x) dx + \int_{x_{i-1}}^{x_i} r(x)(x - x_{i-1}) dx + \int_{x_{i-1}}^{x_i} q(x)(x_i - x)(x - x_{i-1}) dx \right],$$

$$\alpha_{ii} = \frac{1}{h^2} \left[\int_{x_{i-1}}^{x_{i+1}} k(x) dx - \int_{x_{i-1}}^{x_i} r(x)(x - x_{i-1}) dx + \int_{x_i}^{x_{i+1}} r(x)(x_{i+1} - x) dx + \int_{x_{i-1}}^{x_i} q(x)(x - x_{i-1})^2 dx + \right.$$

$$\left. + \int_{x_i}^{x_{i+1}} q(x)(x_{i+1} - x)^2 dx \right],$$

$$\alpha_{ii+1} = \frac{1}{h^2} \left[- \int_{x_i}^{x_{i+1}} k(x) dx - \int_{x_i}^{x_{i+1}} r(x)(x_{i+1} - x) dx + \int_{x_i}^{x_{i+1}} q(x)(x - x_i)(x_{i+1} - x) dx \right], \quad i = \overline{1, N-1},$$

$$\alpha_{00} = \frac{1}{h^2} \left[\int_0^h k(x) dx + \int_0^h r(x)(h - x) dx + \int_0^h q(x)(h - x)^2 dx \right] + \kappa_0,$$

$$\alpha_{01} = \frac{1}{h^2} \left[- \int_0^h k(x) dx - \int_0^h r(x)(h - x) dx + \int_0^h q(x)x(h - x) dx \right],$$

$$\alpha_{NN-1} = \frac{1}{h^2} \left[- \int_{1-h}^1 k(x) dx + \int_{1-h}^1 r(x)(x - 1 + h) dx + \int_{1-h}^1 q(x)(1 - x)(x - 1 + h) dx \right],$$

$$\alpha_{NN} = \frac{1}{h^2} \left[\int_{1-h}^1 k(x) dx - \int_{1-h}^1 r(x)(x - 1 + h) dx + \int_{1-h}^1 q(x)(x - 1 + h)^2 dx \right] + \kappa_1,$$

$$\beta_i = \frac{1}{h} \left[\int_{x_{i-1}}^{x_i} f(x)(x - x_{i-1}) dx + \int_{x_i}^{x_{i+1}} f(x)(x_{i+1} - x) dx \right], \quad i = \overline{1, N-1},$$

$$\beta_0 = \frac{1}{h} \int_0^h f(x)(h - x) dx + g_0,$$

$$\beta_N = \frac{1}{h} \int_{1-h}^1 f(x)(x - 1 + h) dx + g_1.$$



Точно так же, как мы это поделали в методе Ритца, систему (10.141) можно привести к стандартному безындексному виду

$$\begin{cases} (ay_{\bar{x}})_x + b^+ y_x + b^- y_{\bar{x}} - dy = -\varphi, \quad x \in \omega_h \\ (a_1 + b_0^+ h) y_{x,0} = (\kappa_0 + \frac{h}{2} d_0) y_0 - (g_0 + \frac{h}{2} \varphi_0), \\ -(a_N - b_N^+ h) y_{\bar{x},N} = (\kappa_1 + \frac{h}{2} d_N) y_N - (g_1 + \frac{h}{2} \varphi_N), \end{cases} \quad (10.142)$$

где

$$\left\{ \begin{array}{l} a_i = \frac{1}{h} \left[\int_{x_{i-1}}^{x_i} k(x) dx - \int_{x_{i-1}}^{x_i} q(x)(x_i - x)(x - x_{i-1}) dx \right], \quad i = \overline{1, N}, \\ b_i^+ = \frac{1}{h^2} \int_{x_i}^{x_{i+1}} r(x)(x_{i+1} - x) dx, \quad i = \overline{0, N-1}, \\ b_i^- = \frac{1}{h^2} \int_{x_{i-1}}^{x_{i+1}} r(x)(x - x_{i-1}) dx, \quad i = \overline{1, N}, \\ d_i = \frac{1}{h^2} \left[\int_{x_{i-1}}^{x_i} q(x)(x - x_{i-1}) dx - \int_{x_i}^{x_{i+1}} q(x)(x_{i+1} - x) dx \right], \quad i = \overline{1, N-1}, \\ d_0 = \frac{2}{h^2} \int_0^h q(x)(h-x) dx; \quad d_N = \frac{2}{h^2} \int_{1-h}^1 q(x)(x-1+h) dx, \\ \varphi_i = \frac{1}{h^2} \left[\int_{x_{i-1}}^{x_i} f(x)(x - x_{i-1}) dx - \int_{x_i}^{x_{i+1}} f(x)(x_{i+1} - x) dx \right], \quad i = \overline{1, N-1}, \\ \varphi_0 = \frac{2}{h^2} \int_0^h f(x)(h-x) dx; \quad \varphi_N = \frac{2}{h^2} \int_{1-h}^1 f(x)(x-1+h) dx. \end{array} \right.$$

Разностная схема (10.142) имеет второй порядок аппроксимации.

Метод аппроксимации интегрального тождества

Этот метод находится в таком же отношении к [методу Галеркина](#), как [метод аппроксимации квадратичного функционала](#) к [методу Ритца](#).

Для построения разностной схемы на сетке $\bar{\omega}_h$ аппроксимируем интегральное тождество (10.139) сумматорным тождеством для сеточных функций (поэтому метод также называют методом сумматорных тождеств), используя технику, изложенную в [пункте 10.2.3](#): переписав тождество в виде

$$\begin{aligned} J(u, v) = \sum_{i=1}^N & \left\{ \int_{x_{i-1}}^{x_i} k(x) u'(x) v'(x) dx + \int_{x_{i-1}}^{x_i} [q(x) u(x) v(x) - f(x) v(x) - r(x) u'(x) v(x)] dx + \right. \\ & \left. + \kappa_0 u(0) v(0) + \kappa_1 u(1) v(1) - g_0 v(0) - g_1 v(1) \right\} = 0, \end{aligned}$$



заменим первый интеграл под знаком суммы некоторым аналогом **квадратурной формулы средних прямогоугольников** (при этом производные аппроксимируем **левыми разностными**), а второй — **квадратурной формулой трапеций** (производную заменяем **центральной разностной**):

$$\int_{x_{i-1}}^{x_i} k(x) u'(x) v'(x) dx \approx h a_i y_{\bar{x},i} v_{\bar{x},i},$$

$$\int_{x_{i-1}}^{x_i} [q(x) u(x) v(x) - f(x) v(x) - r(x) u'(x) v(x)] dx \approx \frac{h}{2} \left[q_{i-1} y_{i-1} v_{i-1} - f_{i-1} v_{i-1} + q_i y_i v_i - f_i v_i - r_i y_{\bar{x},i} v_i \right].$$

В результате получим сумматорное тождество

$$J_h(y_h, v_h) = \sum_{i=1}^N h a_i y_{\bar{x},i} v_{\bar{x},i} + \sum_{i=1}^{N-1} h \left[q_i y_i v_i - f_i v_i - r_i y_{\bar{x},i} v_i \right] + \frac{h}{2} q_0 y_0 v_0 - \frac{h}{2} f_0 v_0 - \frac{h}{2} r_0 y_{\bar{x},1} v_0 +$$

$$+ \frac{h}{2} q_N y_N v_N - \frac{h}{2} f_N v_N - \frac{h}{2} r_N y_{\bar{x},N} v_N + \kappa_0 y_0 v_0 + \kappa_1 y_N v_N - g_0 v_0 - g_1 v_N =$$

$$= \sum_{i=1}^N h a_i y_{\bar{x},i} v_{\bar{x},i} + \sum_{i=1}^{N-1} h \left[q_i y_i v_i - f_i v_i - r_i y_{\bar{x},i} v_i \right] + (\kappa_0 + \frac{h}{2} q_0) y_0 v_0 + (\kappa_1 + \frac{h}{2} q_N) y_N v_N -$$

$$- \frac{h}{2} r_0 y_{\bar{x},0} v_0 - \frac{h}{2} r_N y_{\bar{x},N} v_N - (g_0 + \frac{h}{2} f_0) v_0 - (g_1 + \frac{h}{2} f_N) v_N = 0.$$

В этом тождестве v — произвольная сеточная функция. Выбирая ее равной единице в одном из узлов сетки и равной нулю в остальных, получим уравнение в той точке, где v отлична от нуля. Перебирая таким образом все узлы, получим разностную схему

$$\begin{cases} (ay_{\bar{x}})_x + r(x) y_{\bar{x}} - qy = -f(x), & x \in \omega_h \\ (a_1 + \frac{h}{2} r(0)) y_{\bar{x},0} = (\kappa_0 + \frac{h}{2} q(0)) y_0 - (g_0 + \frac{h}{2} f(0)), \\ -(a_N - \frac{h}{2} r(1)) y_{\bar{x},N} = (\kappa_1 + \frac{h}{2} q(1)) y_N - (g_1 + \frac{h}{2} f(1)), \end{cases} \quad (10.143)$$

имеющую второй порядок аппроксимации.

Замечание 10.9. Проекционные подходы к построению разностных схем, разобранные нами в данном параграфе, в современной литературе достаточно часто относят к методам конечных элементов.



Замечание 10.10. Все рассмотренные нами выше способы построения разностных схем могут быть распространены как на случай неравномерной сетки на отрезке, так и на случай функций многих независимых переменных.



10.3. Методы исследования устойчивости разностных схем

[10.3.1. Принцип максимума](#)

[10.3.2. Метод разделения переменных](#)

[10.3.3. Метод энергетических неравенств](#)



Меню

10.3.1. Принцип максимума

Примеры исследования устойчивости с помощью принципа максимума

Монотонные разностные схемы для обыкновенных дифференциальных уравнений второго порядка

Ранее мы отмечали важность такого свойства разностных схем как [устойчивость](#). Ее исследование, как правило, состоит в получении априорных оценок решения разностной задачи через ее входные данные.

Для оценок в равномерной метрике разностных эллиптических и параболических уравнений, а также разностных уравнений переноса, применяется [принцип максимума](#). Он позволяет получить равномерные оценки решения через правую часть уравнения, граничные и начальные данные. Опишем его подробнее.

Пусть Ω — некоторое конечное множество точек $x = (x_1, \dots, x_p)$ p -мерного евклидова пространства (сетка). Пусть также в каждой точке $x \in \Omega$ задан шаблон $\mathbb{W}(x) \subset \Omega$. Через $\mathbb{W}'(x)$, как и ранее, обозначим окрестность точки x , т.е. $\mathbb{W}'(x) = \mathbb{W}(x) \setminus \{x\}$.

Рассмотрим уравнение

$$Sy(x) = F(x), \quad x \in \Omega \quad (10.144)$$

где $y(x)$ — искомая функция, $F(x)$ — заданная сеточная функция, а S — линейный оператор, определяемый формулой

$$Sv(x) = A(x)v(x) - \sum_{\xi \in \mathbb{W}'(x)} B(x, \xi)v(\xi) \quad (10.145)$$

коэффициенты которого $A(x)$ и $B(x, \xi)$ — заданные сеточные функции x и ξ . Будем далее предполагать, что они удовлетворяют условиям

- 1) $A(x) > 0, \quad B(x, \xi) > 0 \quad \text{для всех } x \in \Omega, \quad \xi \in \mathbb{W}'(x);$
 - 2) $D(x) \equiv A(x) - \sum_{\xi \in I'(x)} B(x, \xi) \geq 0.$
- (10.146)

Пусть x — произвольный узел сетки Ω . Тогда возможны два случая:

- a) $\mathbb{W}'(x) = \emptyset;$
- б) $\mathbb{W}'(x)$ содержит хотя бы один узел $\xi \in \Omega$.



Меню

Если имеет место первый случай, т.е. $\mathbb{W}'(\bar{x}) = \emptyset$, то уравнение (10.144) при $x = \bar{x}$ имеет вид

$$A(\bar{x})y(\bar{x}) = F(\bar{x})$$

или

$$y(\bar{x}) = g(\bar{x}).$$

Такую точку будем называть граничным узлом (и писать $\bar{x} \in \gamma$), а остальные узлы, окрестность которых состоит, по крайней мере, из одной точки, — внутренними (обозначаем: множество ω). Согласно сказанному $\omega \cup \gamma = \Omega$. Отметим, что с такой точки зрения в случае краевых условий второго или третьего рода для эллиптических уравнений граничных узлов нет.

Будем предполагать также, что сетка Ω — связная, т.е. для любых двух узлов $\bar{x}, \bar{\bar{x}}$, не являющихся одновременно граничными (например, для определенности, $\bar{x} \in \omega$) можно указать такую последовательность узлов x_1, x_2, \dots, x_m , что каждый последующий узел принадлежит окрестности предыдущего, т.е.

$$x_1 \in \mathbb{W}'(\bar{x}), \quad x_2 \in \mathbb{W}'(x_1), \dots, \quad x_{m+1} \in \mathbb{W}'(x_m), \quad \bar{\bar{x}} \in \mathbb{W}'(x_{m+1}) \quad (10.147)$$

Теорема 10.2 (Принцип максимума). Пусть $y(x)$ — отличная от тождественной постоянной сеточная функция, определенная на связной сетке Ω , и пусть на ω выполняются условия (10.146). Тогда из условия $Sy(x) \leq 0$ ($Sy(x) \geq 0$) на ω следует, что $y(x)$ не может принимать наибольшего положительного (наименьшего отрицательного) значения во внутренних узлах сетки Ω .

[\[Доказательство\]](#)

Замечание 10.11. Возможна несколько более общая формулировка доказанной теоремы: не на всей сетке ω , а на некотором связном подмножестве $\Omega' \subset \Omega$.

Следствие 10.1. Пусть $Sy(x) \leq 0$ ($Sy(x) \geq 0$) на связной сетке Ω и существует, по крайней мере, один узел $x_0 \in \Omega$, для которого

$$D(x_0) > 0. \quad (10.148)$$

Тогда $y(x) \leq 0$ ($y(x) \geq 0$) на сетке Ω .

[\[Доказательство\]](#)

Следствие 10.2. Пусть оператор S удовлетворяет на сетке Ω условиям (10.146), (10.146). Тогда задача (10.144), (10.145) имеет единственное решение.

[\[Доказательство\]](#)



Теорема 10.3 (теорема сравнения). Пусть $y(x)$ — решение задачи (10.144) — (10.146), (10.148) , а $\bar{y}(x)$ — решение той же задачи с правой частью $\bar{F}(x)$. Тогда из условия $|F(x)| \leq \bar{F}(x)$ следует, что $|y(x)| \leq |\bar{y}(x)|$ на Ω .

Теорема сравнения позволяет сразу получить оценку решения первой краевой задачи в случае однородного уравнения. Имеет место

Следствие 10.3. Для решения задачи $\begin{cases} Sy(x) = 0, & x \in \omega, \\ y(x) = \mu(x), & x \in \gamma \end{cases}$ имеет место априорная оценка

$$\max_{x \in \omega} |y(x)| \leq \max_{x \in \gamma} |\mu(x)| \quad \text{или} \quad \|y\|_{\bar{C}} \leq \|\mu\|_{C_\gamma}. \quad (10.149)$$

[[Доказательство](#)]

С помощью доказанных утверждений можно получить оценки и для решения неоднородной задачи.

Теорема 10.4. Если $D(x) > 0$ для всех $x \in \Omega$, то для решения задачи (10.144) — (10.146) верна априорная оценка

$$\max_{x \in \Omega} |y(x)| \leq \max_{x \in \Omega} \frac{|F(x)|}{D(x)} \quad \text{или} \quad \|y\|_{\bar{C}} \leq \left\| \frac{F}{D} \right\|_{\bar{C}}. \quad (10.150)$$

[[Доказательство](#)]

Аналогично доказывается

Теорема 10.5. Пусть сетка Ω разбита на два непересекающихся непустых подмножества Ω' и Ω'' , причем Ω' — связное. Если $F(x) \equiv 0$ на Ω' , а на Ω'' $F(x)$ отлична от тождественного нуля и $D(x) > 0$, то для решения задачи (10.144) — (10.146) справедлива оценка

$$\|y\|_{\bar{C}} = \max_{x \in \Omega} |y(x)| \leq \max_{x \in \Omega'} \frac{|F(x)|}{D(x)} \quad (10.151)$$



Примеры исследования устойчивости с помощью принципа максимума

Фактически исследование состоит в том, чтобы:

- 1) привести задачу к виду (10.144) — (10.145);
- 2) проверить условия (10.146), (10.148).

При решении первой части задачи в качестве ориентировки следует помнить, что коэффициент $A(x)$ должен быть *диагональным* элементом матрицы при записи задачи в матрично-векторной форме.

Пример 10.18. Задача Коши для уравнения переноса

$$\begin{cases} \frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = 0, & t > 0, \quad -\infty < x < +\infty, \\ u(x, 0) = u_0(x), & a = \text{const} > 0. \end{cases}$$

Решение. Зададим сетку $\omega_{ht} = \omega_h \times \omega_t$ и на ней запишем простейшую разностную схему, аппроксимирующую поставленную задачу

$$\begin{cases} y_t + ay_x = 0, \\ y(x, 0) = u_0(x) \end{cases}$$

или в индексной форме

$$\begin{cases} \frac{y_k^{j+1} - y_k^j}{\tau} + a \frac{y_k^j - y_{k-1}^j}{h} = 0, & j = 0, 1, \dots \\ y_k^0 = u_0(x_k). \end{cases}$$

Решение данной разностной задачи, очевидно, должно находиться послойно. Единственное же значение на $(j+1)$ -м временном слое, подлежащее определению — y_k^{j+1} . Поэтому коэффициент при нем — искомый диагональный элемент, т.е. в канонической записи (10.145) $x = (x_k, t_{j+1})$. Поэтому схему перепишем в виде

$$\frac{1}{\tau} y_k^{j+1} = \left(\frac{1}{\tau} - \frac{a}{h} \right) y_k^j + \frac{a}{h} y_{k-1}^j.$$

Поэтому

$$A(x) = \frac{1}{\tau}, \quad B_1 = \frac{1}{\tau} - \frac{a}{h}, \quad B_2 = \frac{a}{h}, \quad D(x) = A(x) - (B_1 + B_2) = \frac{1}{\tau} - \left(\frac{1}{\tau} - \frac{a}{h} + \frac{a}{h} \right) = 0.$$

Следовательно, условия (10.146) примут вид

$$\begin{cases} A(x) = \frac{1}{\tau} > 0, \\ B_1 = \frac{1}{\tau} - \frac{a}{h} \geq 0, \\ B_2 = \frac{a}{h} \geq 0, \\ D(x) = 0 \geq 0. \end{cases}$$

Заметим, что для коэффициентов $B(x, \xi)$ мы используем в условиях нестрогие неравенства, поскольку равенство нулю того или иного коэффициента, по сути, означает отсутствие в шаблоне соответствующего узла. В полученной системе первое, третье и четвертое неравенства (учитывая знак коэффициента a) выполняются автоматически, а второе приводит к ограничению, связывающему допустимые шаги сетки: $\frac{a\tau}{h} \leq 1$ (в литературе его называют *условием Куранта*). При выполнении найденного условия легко (на основании [следствия 10.3](#)) получить оценку разностного решения: $\|y\|_{\bar{C}} \leq \|u_0\|_{C_\gamma}$. \square

Пример 10.19. Первая краевая задача для уравнения теплопроводности

$$\begin{cases} \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x, t), \quad 0 < x < 1, \quad t > 0, \\ u(x, 0) = u_0(x), \quad 0 \leq x \leq 1, \\ u(0, t) = \mu_0(t), \quad t \geq 0, \\ u(1, t) = \mu_1(t), \quad t \geq 0. \end{cases}$$

Решение. На сетке $\bar{\omega}_{h\tau}$ запишем явную разностную схему:

$$\begin{cases} y_t = y_{\bar{x}x} + \varphi, \quad (x, t) \in \omega_{h\tau}, \\ y(x, 0) = u_0(x), \quad x \in \bar{\omega}_h, \\ y(0, t) = \mu_0(t), \quad t \in \omega_\tau, \\ y(1, t) = \mu_1(t), \quad \omega_\tau. \end{cases}$$

Расписав разностное уравнение в индексной форме, получим:

$$\frac{y_i^{j+1} - y_i^j}{\tau} = \frac{y_{i+1}^j - 2y_i^j + y_{i-1}^j}{h^2} + \varphi_i^j.$$



Отсюда, учитывая послойный принцип реализации и единственный узел сетки на верхнем временному слое, находим: $x = (x_i, t_{j+1})$. Поэтому, умножив уравнение в индексной форме на τ , перепишем его в виде

$$y_j^{j+1} = \left(1 - \frac{2\tau}{h^2}\right) y_i^j + \frac{\tau}{h^2} \left(y_{i+1}^j + y_{i-1}^j\right) + \tau \varphi_i^j.$$

Таким образом, $A(x) = 1 > 0$, $B_2 = B_3 = \frac{\tau}{h^2} > 0$. Неравенство $B_1 = 1 - \frac{2\tau}{h^2} \geq 0$ приводит к ограничению $\tau \leq \frac{h^2}{2}$ (условие Куранта для уравнения теплопроводности), а последнее из условий

$$D(x) = 1 - \left(1 - \frac{2\tau}{h^2} + \frac{\tau}{h^2} + \frac{\tau}{h^2}\right) \equiv 0$$

очевидным образом выполняется. Следовательно, в равномерной метрике явная разностная схема для уравнения теплопроводности устойчива при выполнении условия $\tau \leq \frac{h^2}{2}$. \square

Монотонные разностные схемы для обыкновенных дифференциальных уравнений второго порядка

Используя полученные в начале параграфа результаты, несложно исследовать и конкретные разностные схемы в случае граничных задач для обыкновенного дифференциального уравнения второго порядка.

Пусть, например, исходная дифференциальная задача имеет вид

$$\begin{cases} \frac{d}{dx} \left(k(x) \frac{du(x)}{dx} \right) - q(x) u(x) = -f(x), \quad 0 < x < 1, \\ k(x) \geq k_0 > 0, \quad q(x) \geq 0, \\ u(0) = \mu_0, \\ u(1) = \mu_1. \end{cases} \quad (10.152)$$

Запишем для нее однородную консервативную разностную схему

$$\begin{cases} (ay_{\bar{x}})_x - dy = -\varphi, \quad x \in \omega_h, \\ y(0) = \mu_0, \\ y(1) = \mu_1. \end{cases} \quad (10.153)$$



Расписывая ее в индексной форме, имеем (уравнения, описывающие граничные условия, в данной задаче тривиальны):

$$\frac{1}{h} \left(a_{i+1} \frac{y_{i+1} - y_i}{h} - a_i \frac{y_i - y_{i-1}}{h} \right) - d_i y_i = -\varphi_i.$$

Исходя из принципа «точка x должна соответствовать диагональному элементу», получаем: $x = x_i$. Поэтому, собрав подобные, перепишем наше разностное уравнение в виде

$$\left(\frac{1}{h^2} (a_{i+1} + a_i) + d_i \right) y_i - \left(\frac{1}{h^2} a_{i+1} y_{i+1} + \frac{1}{h^2} a_i y_{i-1} \right) = \varphi_i,$$

откуда

$$A(x) = \frac{1}{h^2} (a_{i+1} + a_i) + d_i, \quad B_1 = \frac{1}{h^2} a_{i+1}, \quad B_2 = \frac{1}{h^2} a_i, \quad D(x) = d_i.$$

Очевидно, если функционалы, с помощью которых вычисляются коэффициенты разностной схемы, сохраняют свойства коэффициентов исходной дифференциальной задачи (в частности, положительность), то записанная разностная схема будет удовлетворять условиям принципа максимума при всех значениях параметра h . Такие разностные схемы называют [монотонными](#).

Не следует думать, что монотонность — «врожденное» свойство разностных схем. Чтобы убедиться в обратном, рассмотрим задачу более общего, нежели (10.152), вида

$$\begin{cases} Lu(x) \equiv \frac{d}{dx} \left(k(x) \frac{du(x)}{dx} \right) + r(x) \frac{du(x)}{dx} - q(x) u(x) = -f(x), \quad 0 < x < 1, \\ k(x) \geq k_0 > 0, \quad q(x) \geq 0, \quad |r(x)| \leq C, \\ u(0) = \mu_0, \\ u(1) = \mu_1. \end{cases} \quad (10.154)$$

Для этой задачи легко записать разностную схему второго порядка, заменив производную $\frac{du}{dx}$ в слагаемом $r(x) \frac{du}{dx}$ центральной разностной производной:

$$\begin{cases} (ay_x)_x + ry_{xx} - qy = -f, \quad x \in \omega_h, \\ y(0) = \mu_0, \\ y(1) = \mu_1. \end{cases} \quad (10.155)$$



Проделывая выкладки, аналогичные приведенным выше, получаем:

$$\left(\frac{a_{i+1} + a_i}{h^2} + q_i \right) y_i - \left(\frac{a_{i+1}}{h^2} + \frac{r_i}{2h} \right) y_{i+1} - \left(\frac{a_i}{h^2} - \frac{r_i}{2h} \right) y_{i-1} = f_i,$$

т.е.

$$A(x) = \frac{a_{i+1} + a_i}{h^2} + q_i, \quad B_1 = \frac{a_{i+1}}{h^2} + \frac{r_i}{2h}, \quad B_2 = \frac{a_i}{h^2} - \frac{r_i}{2h}, \quad D(x) = q_i.$$

Отсюда видим, что сеточные коэффициенты $A(x)$ и $D(x)$ удовлетворяют условиям [принципа максимума](#) при всех h , в то время как неотрицательность коэффициентов B_1 и B_2 приводит к ограничению на шаг сетки вида $h \leq \frac{2k(x)}{|r(x)|}$, которое становится достаточно обременительным, если $|r(x)| \gg 1$.

В то же время, если воспользоваться для аппроксимации $\frac{du}{dx}$ в слагаемом $r(x) \frac{du}{dx}$ односторонними производными ([правой](#) при $r(x) \geq 0$ и [левой](#) при $r(x) \leq 0$: так называемая [аппроксимация против потока](#)), то полученная разностная схема

$$\begin{cases} (ay_{\bar{x}})_x + r^+ y_x + r^- y_{\bar{x}} - qy = -f, & x \in \omega_h, \\ y(0) = \mu_0, \\ y(1) = \mu_1 \end{cases} \quad (10.156)$$

будет монотонной, но ее порядок равен единице.

Построим монотонную схему второго порядка точности, содержащую односторонние производные, учитывающие знак коэффициента $r(x)$. Для этого, как оказывается, достаточно написать монотонную схему с односторонними производными типа (10.156) для уравнения с возмущенными коэффициентами

$$\tilde{L}u(x) \equiv \kappa \frac{d}{dx} \left(k(x) \frac{du(x)}{dx} \right) + r(x) \frac{du(x)}{dx} - q(x) u(x) = -f(x) \quad (10.157)$$

где $\kappa = \frac{1}{1+R}$, $R = \frac{h|r|}{2k}$ — разностное число Рейнольдса.

Аппроксимируем слагаемое $r(x) \frac{du(x)}{dx}$ выражением

$$(ru')_i = \left(\frac{r}{k} (ku') \right)_i \sim b_i^+ a_{i+1} u_{x,i} + b_i^- a_i u_{\bar{x},i},$$



где $b_i^\pm = F(\tilde{r}^\pm(x_i + sh))$, $\tilde{r}^\pm = \frac{r_i^\pm}{k}$, а F — шаблонный функционал, используемый для вычисления коэффициентов d и φ разностной схемы (например, можно просто положить $b_i^+ = \frac{r_i^+}{k_i} = \frac{r_i + |r_i|}{2k_i}$, $b_i^- = \frac{r_i^-}{k_i} = \frac{r_i - |r_i|}{2k_i}$).

В результате получаем однородную разностную схему

$$\begin{cases} \kappa(ay_{\bar{x}})_x + b^+ a^{(+1)} y_x + b^- ay_{\bar{x}} - dy = -\varphi, & x \in \omega_h, \\ a^{(+1)} = a(x + h), \quad \kappa = \frac{1}{1+R}, \quad R = \frac{h|r|}{2k}, \\ y(0) = \mu_0, \\ y(1) = \mu_1 \end{cases} \quad (10.158)$$

Приводя (10.158) к каноническому виду по изложенной ранее схеме, имеем:

$$B_1 = \frac{\kappa a_{i+1}}{h^2} + \frac{b_i^+ a_{i+1}}{h} > 0, \quad B_2 = \frac{\kappa a_i}{h^2} - \frac{b_i^- a_i}{h} > 0, \quad A(x) = B_1 + B_2 + d_i > 0, \quad D(x) = d_i \geq 0.$$

Таким образом, разностная схема (10.158) является монотонной.

Погрешность аппроксимации этой схемы

$$\psi = \kappa(au_{\bar{x}})_x + b^+ a^{(+1)} u_x + b^- au_{\bar{x}} - du + \varphi - (Lu + f)$$

представим в виде суммы

$$\psi = \psi^{(1)} + \psi^{(2)},$$

$$\psi^{(1)} = [(au_{\bar{x}})_x - du + \varphi] - [(ku')' - qu + f],$$

$$\psi^{(2)} = [(\kappa - 1)(au_{\bar{x}})_x + b^+ a^{(+1)} u_x + b^- au_{\bar{x}}] - ru'.$$



Как мы помним, для достаточно гладких функций $\psi^{(1)} = O(h^2)$. В то же время,

$$b^+ = \tilde{r}^+ + O(h^2), \quad b^- = \tilde{r}^- + O(h^2), \quad k\tilde{r}^\pm = r^\pm, \quad r^+ + r^- = r, \quad r^+ - r^- = |r|,$$

$$au_{\bar{x}} = ku' - \frac{h}{2}(ku')' + O(h^2), \quad a^{(+1)}u_x = ku' + \frac{h}{2}(ku')' + O(h^2),$$

$$(au_{\bar{x}})_x = (ku')' + O(h^2).$$

Поэтому

$$\begin{aligned} b^+ a^{(+1)}u_x + b^- au_{\bar{x}} &= [\tilde{r}^+ + O(h^2)] \cdot [ku' + \frac{h}{2}(ku')' + O(h^2)] + [\tilde{r}^- + O(h^2)] \cdot [ku' - \frac{h}{2}(ku')' + O(h^2)] = \\ &= (\tilde{r}^+ + \tilde{r}^-)ku' + \frac{h}{2}(ku')'(\tilde{r}^+ - \tilde{r}^-) + O(h^2) = ru' + \frac{h}{2}(ku')' \cdot \frac{|r|}{k} + O(h^2). \end{aligned}$$

Следовательно,

$$\begin{aligned} \psi^{(2)} &= [\kappa - 1 = \frac{1}{1+R} - 1 = -\frac{R}{1+R}] = -\frac{R}{1+R}(ku')' + ru' + \frac{h}{2}(ku')' \cdot \frac{|r|}{k} + O(h^2) - ru' = \\ &= (ku')' \cdot (R - \frac{R}{1+R}) + O(h^2) = (ku')' \cdot \frac{R^2}{1+R} + O(h^2) = O(h^2), \end{aligned}$$

так как $R = \frac{h|r|}{2k} = O(h)$.

Таким образом, построенная разностная схема (10.158) имеет второй порядок и является монотонной. Ее целесообразно использовать в случае быстро меняющейся функции $r(x)$.

Замечание 10.12. Монотонную разностную схему второго порядка несложно написать, если от уравнения (10.154) перейти к уравнению

$$\frac{d}{dx} \left(\mu(x) k(x) \frac{du(x)}{dx} \right) - \mu(x) q(x) u(x) = -\mu(x) f(x),$$

где $\mu(x) = \exp \left(\int \frac{r(x)}{k(x)} dx \right)$, т.е. преобразовав его к самосопряженному виду.



10.3.2. Метод разделения переменных

Этот метод применяется для строгого обоснования многих линейных разностных схем и нестрогого исследования большинства нелинейных задач. Теоретические основы метода и его практическое применение традиционно изучаются в курсе «Уравнений математической физики».

Технически поиск частного решения уравнения в виде произведения функций, каждая из которых зависит только от одной независимой переменной, приводит к необходимости решать задачу на собственные значения для некоторого дифференциального оператора. И, таким образом, решение задачи получается в виде ряда по собственным функциям данных операторов. При этом, конечно же, хорошо, если данная система оказывается ортогональной (или, более того, ортонормированной).

В разностном варианте технически и теоретически все остается таким же: ищем частное решение в виде произведения функций, каждая из которых зависит от одной (своей) независимой переменной, переходим к задаче на собственные значения. После ее решения можно делать некоторые заключения об исследуемых свойствах решения разностной задачи (в частности, об устойчивости).

Конечно, вместо поиска конкретных систем собственных функций и собственных значений для каждого разностного оператора можно пользоваться и некоей универсальной системой сеточных функций, которая являлась бы ортогональной системой собственных функций любого оператора разностного дифференцирования на равномерной сетке (по аналогии с тем, как мы проводим разложение в ряд Фурье на всей числовой прямой функций, периодических с периодом l , или периодически продолжая их на всю числовую прямую).

В этом случае также можно сеточную функцию y_h , определенную на сетке $\bar{w}_h = \{x_k = kh, k = 0, 1, \dots, N; h = \frac{l}{N}\}$ доопределить на всей числовой прямой с координатами $x_k = kh, k = 0, \pm 1, \pm 2, \dots$ так, чтобы получилась l -периодическая сеточная функция. Множество таких функций обозначим M_h .

В пространстве M_h введем скалярное произведение по формуле

$$(v_h, y_h) = \sum_{s=0}^{N-1} h v_h(x_s) \bar{y}_h(x_s).$$

Тогда примером системы линейно независимых l -периодических ортогональных в M_h функций являются функции $\mu_k(x) = \exp(ik\frac{2\pi}{l}x)$, $k = 0, \pm 1, \pm 2, \dots, \pm \frac{N-1}{2}$ (или $k = 0, 1, \dots, N-1$).

Действительно,

$$(\mu_k(x), \mu_m(x)) = \sum_{s=0}^{N-1} h \exp\left(ik\frac{2\pi}{l}x_s\right) \cdot \exp\left(-im\frac{2\pi}{l}x_s\right) = \sum_{s=0}^{N-1} h \exp\left(i(k-m)\frac{2\pi}{l}sh\right) = Nh\delta_k^m = l\delta_k^m.$$

Следовательно, любую функцию из M_h можно разложить в «сумму Фурье»

$$y_h(x) = \sum_{k=0}^{N-1} a_k \mu_k(x).$$

Заметим также, что $\mu_k(x)$ являются собственными функциями операторов правой и левой разностных производных. Действительно,

$$\begin{aligned} (\mu_k(x))_x &= \frac{\mu_k(x+h) - \mu_k(x)}{h} = \frac{\exp(ik\frac{2\pi}{l}x) \cdot \exp(ik\frac{2\pi}{l}h) - \exp(ik\frac{2\pi}{l}x)}{h} = \\ &= \mu_k(x) \cdot \frac{\exp(ik\frac{2\pi}{l}h) - 1}{h}. \end{aligned}$$

Видим отсюда, что собственным значением оператора правой разностной производной, соответствующим собственной функции $\mu_k(x)$, является $\lambda_k = \frac{\exp(ik\frac{2\pi}{l}h) - 1}{h}$. Аналогичный результат имеет место для левой разностной производной. Следовательно, функции $\mu_k(x)$ образуют полную систему собственных функций для любого разностного оператора L_h вида

$$L_h y_h = \sum_{s,q} a_{sq} D^s \bar{D}^q y_h,$$

где

$$D^s y_h = y_{\underbrace{xx\dots x}_s}, \quad \bar{D}^q y_h = y_{\underbrace{\bar{x}\bar{x}\dots \bar{x}}_q}.$$

Это позволяет изучать вопросы исследования устойчивости разностных схем с использованием данной системы функций.



Рассмотрим применение к линейным двухслойным разностным схемам, записываемым в каноническом виде

$$By_t + Ay = \varphi \quad (10.159)$$

где B и A — некоторые разностные операторы, действующие по пространственной переменной x .

При фиксированной правой части погрешность z приближенного решения удовлетворяет однородному уравнению

$$B\hat{z} = (B - \tau A)z \quad (10.160)$$

Будем, в соответствии с изложенным выше, искать частное решение в виде

$$z(x_s, t_j) = q_k^j \exp\left(ik\frac{2\pi}{l}x_s\right) \quad (10.161)$$

При этом, очевидно, $\hat{z} = q_k z$, так что q_k есть множитель роста k -й гармоники при переходе с одного временного слоя на другой. Подставляя (10.161) в (10.160), получим уравнение для определения q_k :

$$q_k^{j+1} \lambda_k(B) \exp\left(ik\frac{2\pi}{l}x_s\right) = q_k^j [\lambda_k(B) - \tau \lambda_k(A)] \exp\left(ik\frac{2\pi}{l}x_s\right).$$

Отсюда

$$q_k = 1 - \tau \frac{\lambda_k(A)}{\lambda_k(B)}$$

(при этом, естественно, исходная разностная схема должна быть разностной схемой с постоянными коэффициентами).

Теперь остается оценить «степень роста». Имеет место

Теорема 10.6 (признак устойчивости). *Разностная схема (10.159) с постоянными коэффициентами устойчива по начальным данным, если для всех k выполняется неравенство*

$$|q_k| \leq 1 + C\tau, \quad C. \quad (10.162)$$

[[Доказательство](#)]



Замечание 10.13. Фактически константа C не должна быть слишком большой, поэтому при проверке сформулированного признака обычно полагают $C = 0$.

Следствие 10.4. Если хотя бы для одного k величину $|q_k|$ нельзя мажорировать величиной $1 + C\tau$, то схема неустойчива.

Замечание 10.14. Практическое использование метода разделения переменных обычно состоит в следующем:

- 1) полагают $y_k^j = q^j e^{ik\varphi}$, где $\varphi \in [0, 2\pi)$ (по сути, используется обозначение $\varphi = \frac{2\pi}{l}x_s$);
- 2) подставляя это выражение в исследуемую разностную схему, находят q ;
- 3) проверяют условие $|q| \leq 1$.

Пример 10.20. Исследуем описанным способом разностную схему для уравнения переноса, изученную нами в предыдущем параграфе с помощью принципа максимума:

$$y_t + ay_{\bar{x}} = 0.$$

Решение. Расписав разностное уравнение в индексной форме, имеем:

$$\frac{y_k^{j+1} - y_k^j}{\tau} + a \frac{y_k^j - y_{k-1}^j}{h} = 0 \quad (10.163)$$

Пусть теперь $y_k^j = q^j e^{ik\varphi}$. Подставляя это выражение в (10.163), получаем:

$$\frac{q^{j+1}e^{ik\varphi} - q^j e^{ik\varphi}}{\tau} + a \frac{q^j e^{ik\varphi} - q^{j-1} e^{i(k-1)\varphi}}{h} = 0.$$

Отсюда, сократив на $q^j e^{ik\varphi}$, находим:

$$q = 1 - \frac{a\tau}{h} + \frac{a\tau}{h} e^{-i\varphi}$$

или (полагая $\gamma = \frac{a\tau}{h}$)

$$q = 1 - \gamma + \gamma e^{-i\varphi} = 1 - \gamma (1 - e^{-i\varphi}).$$



Следовательно,

$$|q|^2 = (1 - \gamma + \gamma \cos \varphi)^2 + \gamma^2 \sin^2 \varphi = 1 - 2\gamma + 2\gamma \cos \varphi - 2\gamma^2 \cos \varphi + 2\gamma^2.$$

Поэтому неравенство $|q|^2 \leq 1$ может быть переписано в виде

$$-\gamma(1 - \cos \varphi) + \gamma^2(1 - \cos \varphi) \leq 0$$

или

$$\gamma(1 - \cos \varphi)(1 - \gamma) \geq 0,$$

откуда, учитывая положительность коэффициента a , получаем ограничение $\gamma \leq 1$, которое совпадает с условием Куранта, полученным нами ранее. \square

Замечание 10.15. Полученное совпадение, вообще говоря, случайно, поскольку речь идет об устойчивости в различных нормах.



10.3.3. Метод энергетических неравенств

Ранее мы приводили простейший пример использования метода для получения априорных оценок. Сейчас используем его для получения критерия устойчивости двуслойных разностных схем, записанных в канонической форме

$$By_t + Ay = \varphi. \quad (10.164)$$

Для упрощения выкладок детальное изложение проведем в предположении, что:

- 1) операторы A и B не зависят от t (постоянны);
- 2) $B > 0$ (положителен);
- 3) $A = A^* > 0$ (положителен и самосопряжен).

Умножив уравнение (10.164) скалярно на сеточную функцию $2\tau y_t$, получим:

$$2\tau (By_t, y_t) + 2\tau (Ay, y_t) = 2\tau (\varphi, y_t) \quad (10.165)$$

Так как

$$y = \frac{\hat{y} + y}{2} - \frac{\hat{y} - y}{2} = \frac{1}{2}(\hat{y} + y) - \frac{\tau}{2}y_t,$$

то (10.165) перепишем в виде

$$2\tau \left(\left(B - \frac{\tau}{2} A \right) y_t, y_t \right) + (A(\hat{y} + y), \hat{y} - y) = 2\tau (\varphi, y_t).$$

Поскольку

$$(A(\hat{y} + y), \hat{y} - y) = (A\hat{y}, \hat{y}) - (Ay, y),$$

то отсюда имеем:

$$2\tau \left(\left(B - \frac{\tau}{2} A \right) y_t, y_t \right) + (A\hat{y}, \hat{y}) = (Ay, y) + 2\tau (\varphi, y_t) \quad (10.166)$$

Формула (10.166) определяет *энергетическое тождество* для разностной схемы (10.164).

**Теорема 10.7. Условие**

$$B \geq \frac{\tau}{2} A \quad (10.167)$$

необходимо и достаточно для устойчивости в H_A по начальным данным разностной схемы (10.164), т.е. для выполнения неравенства

$$\|y^j\|_A \leq \|y^0\|_A, \quad j = 1, 2, \dots \quad (10.168)$$

(здесь $\|y\|_A = \sqrt{(Ay, y)}$).

[\[Доказательство\]](#)

Замечание 10.16. Условие (10.167) достаточно для устойчивости схемы (10.164), если $B = B(t)$ — несамосопряженный положительный оператор.

Замечание 10.17. Если исходное семейство разностных схем другое (т.е. операторы A и B удовлетворяют некоторым условиям, отличным от сформулированных выше) (например, оба положительные, самосопряженные и постоянные и т.п.), то аналогичным путем могут быть установлены и другие условия (чаще всего — достаточные) устойчивости (например, в норме $\|\cdot\|_B$)

Пример 10.21. Вновь обратимся к явной разностной схеме для уравнения теплопроводности с нулевыми граничными условиями и нулевой правой частью:

$$\begin{cases} \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, & 0 < x < 1, \quad t > 0, \\ u(x, 0) = u_0(x), & 0 \leq x \leq 1, \\ u(0, t) = u(1, t) = 0. \end{cases}$$

Решение. Соответствующая разностная схема на сетке $\bar{\omega}_{ht}$ имеет вид

$$\begin{cases} y_t = y_{\bar{x}x}, & (x, t) \in \omega_{ht}, \\ y(x, 0) = u_0(x), & x \in \bar{\omega}_h, \\ y(0, t) = y(1, t) = 0, & t \in \omega_\tau. \end{cases}$$

Приведем ее к виду (10.164):

$$y_t - y_{\bar{x}x} = 0$$



или

$$By_t + Ay = 0,$$

где

$$B = E, \quad A = -\Lambda.$$

Условие (10.167) теперь примет вид

$$E - \frac{\tau}{2}A \geq 0.$$

Так как $A \leq \|A\|E$, то неравенство можно усилить:

$$E - \frac{\tau}{2}A \geq E - \frac{\tau}{2}\|A\|E = \left(1 - \frac{\tau}{2}\|A\|\right)E \geq 0.$$

Отсюда следует:

$$1 - \frac{\tau}{2}\|A\| \geq 0$$

или

$$\tau \leq \frac{2}{\|A\|}.$$

Так как оператор A — самосопряженный, то $\|A\| = \lambda_{N-1} = \frac{4}{h^2} \cos^2 \frac{\pi h}{2} < \frac{4}{h^2}$. Поэтому полученное условие устойчивости можно записать в более удобном виде $\tau \leq \frac{h^2}{2}$. \square

Замечание 10.18. В случае самосопряженных операторов B и A вместо операторного неравенства (10.167) можно перейти к системе числовых неравенств вида

$$\lambda_k(B) \geq \frac{\tau}{2} \lambda_k(A).$$

Последнее вполне аналогично условию спектральной устойчивости, полученному нами выше для метода разделения переменных.



10.4. Разностные схемы для стационарных задач математической физики

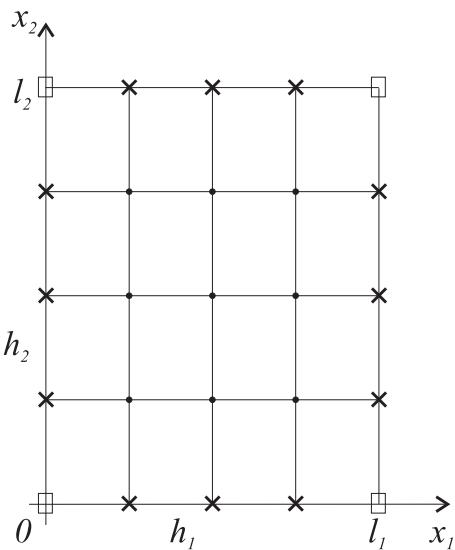
10.4.1. Разностная задача Дирихле для уравнения Пуассона в прямоугольной области

10.4.2. Консервативная схема для задачи Дирихле

Большой класс физических задач составляют стационарные задачи, решение которых не зависит от времени, а определяется только пространственными переменными.



10.4.1. Разностная задача Дирихле для уравнения Пуассона в прямоугольной области



Пусть $\bar{G} = \{0 \leq x_1 \leq l_1, 0 \leq x_2 \leq l_2\}$ — прямоугольник со сторонами l_1 и l_2 и границей Γ . В области $\bar{G} = G \cup \Gamma$ рассмотрим задачу Дирихле для уравнения Пуассона:

$$\left\{ \begin{array}{l} Lu \equiv \nabla^2 u \equiv \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = -f(x), \quad x = (x_1, x_2) \in G, \quad u = u(x) \\ u \Big|_{\Gamma} = \mu(x), \quad x \in \Gamma. \end{array} \right. \quad (10.169)$$



Построим в \bar{G} равномерную сетку $\omega_h = \omega_h \cup \gamma_h$ с шагами $h_1 = l_1/N_1$, $h_2 = l_2/N_2$ и узлами $x_i = (x_{1,i_1}, x_{2,i_2})$, $i_1 = \overline{0, N_1}$, $i_2 = \overline{0, N_2}$.

Во внутренних узлах заменим оператор Лапласа разностным оператором

$$Lu \sim \Delta u = u_{\bar{x}_1 x_1} + u_{\bar{x}_2 x_2}, \quad x \in \omega_h.$$

Правую часть $f(x)$ аппроксимируем сеточной функцией $\varphi(x)$ такой, что $\varphi(x) = f(x) + O(|h|^2)$, $x \in \omega_h$. Например, $\varphi(x) = f(x)$.

Исходной дифференциальной задаче (10.169) поставим в соответствие разностную задачу:

$$\begin{cases} \Lambda y \equiv y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2} = -\varphi(x), & x \in \omega_h, \\ y(x) = \mu(x), & x \in \gamma_h. \end{cases} \quad (10.170)$$

Разностная задача (10.170) является СЛАУ с пятидиагональной матрицей и может быть решена любым подходящим методом решения СЛАУ, например: [МПИ](#), [Зейделя](#), [релаксации](#).

Оценим [погрешность аппроксимации](#) в точке $x = (x_1, x_2) \in \omega_h$ в классе достаточно гладких функций:

$$\begin{aligned} \psi(x) &= u_{\bar{x}_1 x_1} + u_{\bar{x}_2 x_2} + \varphi(x) = Lu + \frac{1}{12} h_1^2 \frac{\partial^4 u}{\partial x_1^4} + \frac{1}{12} h_2^2 \frac{\partial^4 u}{\partial x_2^4} + O(|h|^4) + \varphi = \\ &= (\varphi - f) + \frac{1}{12} h_1^2 \frac{\partial^4 u}{\partial x_1^4} + \frac{1}{12} \frac{\partial^4 u}{\partial x_2^4} + O(|h|^4) = O(|h|^2). \end{aligned}$$

Таким образом,

$$\|\psi\|_{\omega_h} = O(|h|^2). \quad (10.171)$$

Поскольку граничные условия аппроксимируются точно, то это означает, что разностная схема (10.170) имеет [второй порядок аппроксимации](#).

Исследуем [устойчивость](#) и [сходимость](#) разностной задачи (10.170). Для этого представим ее в [канонической форме принципа максимума](#):

$$\begin{cases} -\Lambda y(x) \equiv \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) y(x) - \left(\frac{1}{h_1^2} y^{(-1_1)} + \frac{1}{h_1^2} y^{(+1_1)} + \frac{1}{h_2^2} y^{(-1_2)} + \frac{1}{h_2^2} y^{(+1_2)} \right) = \\ = \varphi(x), \quad x \in \omega_h, \\ y(x) = \mu(x), \quad x \in \gamma_h \end{cases} \quad (10.172)$$

Схема приведена к виду

$$A(x)y(x) - \sum_{\xi \in \text{III}'(x)} B(x, \xi)y(\xi) = F(x), \quad x \in \bar{\omega}_h,$$

где

$$A(x) = \frac{2}{h_1^2} + \frac{2}{h_2^2} > 0, \quad F(x) = \varphi(x),$$

$$B(x, x^{(\pm 1_1)}) = \frac{1}{h_1^2} > 0, \quad B(x, x^{(\pm 1_2)}) = \frac{1}{h_2^2} > 0,$$

$$D(x) = A(x) - \sum_{\xi \in \text{III}'(x)} B(x, \xi) = 0 \text{ при } x \in \omega_h,$$

$$A(x) = 1 > 0, \quad F(x) = \mu(x), \quad \text{III}'(x) = \emptyset,$$

$$D(x) = A(x) = 1 > 0 \text{ при } x \in \gamma_h.$$

Следовательно, коэффициенты схемы (10.172) удовлетворяют условиям [принципа максимума](#) при любых h_1, h_2 .

Согласно [следствию 10.3 принципа максимума](#), схема (10.172) устойчива по граничным условиям. С помощью [теоремы сравнения](#) можно доказать, что схема (10.172) абсолютно устойчива. Для этого выберем функцию

$$\bar{y}(x) = c_1 [l^2 - (x_1^2 + x_2^2)] + c_2, \quad l^2 = l_1^2 + l_2^2 \quad (10.173)$$

в качестве мажоранты, где $c_1, c_2 = \text{const} \geq 0$, которые мы выберем позже.

Вычислим

$$\begin{aligned} \bar{\varphi}(x) &= -\Lambda \bar{y}(x) = -\bar{y}_{\bar{x}_1 x_1} - \bar{y}_{\bar{x}_2 x_2} = c_1 ((x_1^2)_{\bar{x}_1 x_1} + (x_2^2)_{\bar{x}_2 x_2}) = \\ &= c_1 \left[\frac{(x_1 - h_1)^2 - 2x_1^2 + (x_1 + h_1)^2}{h_1^2} + \frac{(x_2 - h_2)^2 - 2x_2^2 + (x_2 + h_2)^2}{h_2^2} \right] = 4c_1, \quad x \in \omega_h. \end{aligned}$$

$$\bar{\mu}(x) = \bar{y}(x) \geq c_2, \quad x \in \gamma_h.$$



Выбирая

$$4c_1 = \|\varphi\|_{\omega_h} = \max_{x \in \omega_h} |\varphi(x)| \geq |\varphi(x)|,$$

$$c_2 = \|\mu\|_{\gamma_h} = \max_{x \in \gamma_h} |\mu(x)| \geq |\mu(x)|,$$

будем иметь задачу

$$\begin{cases} \Lambda \bar{y}(x) = -\bar{\varphi}(x), & x \in \omega_h, \\ \bar{y}(x) = \bar{\mu}(x), & x \in \gamma_h, \end{cases}$$

где $\bar{\varphi}(x) \geq |\varphi(x)|$ при $x \in \omega_h$, $\bar{\mu}(x) \geq |\mu(x)|$ при $x \in \gamma_h$.

По теореме сравнения справедлива оценка

$$|y(x)| \leq \bar{y}(x) \quad \forall x \in \bar{\omega}_h$$

для решения задачи (10.172), или

$$|y(x)| \leq \bar{y}(x) \leq c_1 l^2 + c_2 = \frac{l^2}{4} \|\varphi(x)\|_{\omega_h} + \|\mu\|_{\gamma_h}. \quad (10.174)$$

Неравенство (10.174) означает, что схема (10.172), т. е. (10.170), устойчива при любых h_1, h_2 , то есть абсолютно устойчива.

По теореме сходимости из (10.171), (10.174) следует, что

$$\|z\|_{\bar{\omega}_h} = \|y - u\|_{\bar{\omega}_h} = O(h^2),$$

т. е. схема (10.170) имеет второй порядок точности (схема (10.170) равномерно сходится со скоростью $O(|h|^2)$).



10.4.2. Консервативная схема для задачи Дирихле

Продемонстрируем построение [консервативной](#) разностной схемы с применением [интегро-интерполяционного метода](#) на примере задачи Дирихле в прямоугольнике $G = G \cup \Gamma$ для стационарного двумерного уравнения теплопроводности:

$$\begin{cases} Lu \equiv \nabla \cdot (k \nabla u) \equiv \frac{\partial}{\partial x_1} \left(k \frac{\partial u}{\partial x_1} \right) + \frac{\partial}{\partial x_2} \left(k \frac{\partial u}{\partial x_2} \right) = -f(x), & x = (x_1, x_2) \in G, \\ u \Big|_{\Gamma} = \mu(x), & x \in \Gamma. \end{cases} \quad (10.175)$$

Здесь $k = k(x) > 0$ — коэффициент теплопроводности, $f(x)$ — функция, описывающая мощность и распределение внутренних источников (стоков) тепла в области G , $u(x)$ — искомая температура в точке $x \in \bar{G}$.

Для построения консервативной схемы необходимо сначала найти выражение для закона сохранения, соответствующего уравнению (10.175).

С этой целью проинтегрируем уравнение (10.175) по произвольной площади (объему) $D \subseteq \bar{G}$:

$$\iint_D \left[\sum_{\alpha=1}^2 \frac{\partial}{\partial x_\alpha} \left(k \frac{\partial u}{\partial x_\alpha} \right) + f \right] dx_1 dx_2 = 0.$$

Переходя от двойного интеграла к контурному по формуле Грина (Остроградского), имеем

$$\oint_C (L_1 u dx_2 - L_2 u dx_1) + \iint_D f dx_1 dx_2 = 0, \quad (10.176)$$

где C — граница области D ,

$$L_\alpha u = k \frac{\partial u}{\partial x_\alpha}, \quad \alpha = 1, 2.$$

Замечание 10.19. Уравнение баланса (10.176) можно записать по-другому:

$$\oint_C q dC = \iint_D f dx_1 dx_2,$$



где $q = -k \frac{\partial u}{\partial n}$ — тепловой поток в точке $x \in C$, n — внешняя нормаль к контуру C .

Первый интеграл определяет количество тепла, которое проходит через контур C в единицу времени, то есть разность между количеством тепла, поступившего в область D и количеством тепла, вышедшего из области D через контур C .

Второй интеграл представляет собой количество тепла, выделенного (поглощённого) в объёме D за счёт внутренних источников (стоков). Следовательно, соотношение (10.176) выражает закон сохранения тепла в области $D \subseteq \bar{G}$.

Замечание 10.20. Уравнения (10.175) и (10.176) эквивалентны лишь при условии, что функции $L_1 u$, $L_2 u$ непрерывно дифференцируемы.

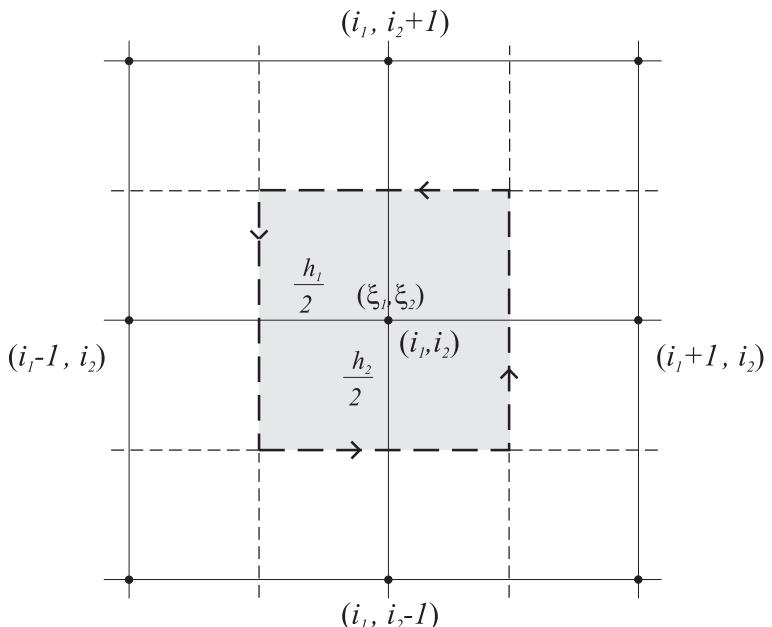
Введём в прямоугольнике \bar{G} равномерную сетку $\bar{\omega}_h = \omega_h \cup \gamma_h$ с шагами h_1, h_2 и узлами $x_i = (x_{1,i}, x_{2,i})$, $x_{1,i} = i_1 h_1$, $i_1 = \overline{0, N_1}$, $x_{2,i} = i_2 h_2$, $i_2 = \overline{0, N_2}$.

Чтобы построить консервативную схему в узле $x_i = (x_{1,i}, x_{2,i})$, $i_\alpha = \overline{1, N_\alpha - 1}$, выберем в качестве области интегрирования D элементарную площадку D_i (контрольный объем), ограниченную линиями

$$\begin{aligned} x_1 &= x_1^{(-0.5)} = x_{1,i-\frac{1}{2}} = x_{1,i} - \frac{1}{2}h_1, \\ x_1 &= x_1^{(+0.5)} = x_{1,i+\frac{1}{2}} = x_{1,i} + \frac{1}{2}h_1, \\ x_2 &= x_2^{(-0.5)} = x_{2,i-\frac{1}{2}} = x_{2,i} - \frac{1}{2}h_2, \\ x_2 &= x_2^{(+0.5)} = x_{2,i+\frac{1}{2}} = x_{2,i} + \frac{1}{2}h_2. \end{aligned}$$

Интеграл по контуру в (10.176) тогда трансформируется следующим образом:

$$\oint_C (L_1 u dx_2 - L_2 u dx_1) = - \int_{x_1^{(-0.5)}}^{x_1^{(+0.5)}} L_2 u \Big|_{x_2=x_2^{(-0.5)}} dx_1 + \int_{x_2^{(-0.5)}}^{x_2^{(+0.5)}} L_1 u \Big|_{x_1=x_1^{(+0.5)}} dx_2 - \\ - \int_{x_1^{(+0.5)}}^{x_1^{(-0.5)}} L_2 u \Big|_{x_2=x_2^{(+0.5)}} dx_1 + \int_{x_2^{(+0.5)}}^{x_2^{(-0.5)}} L_1 u \Big|_{x_1=x_1^{(-0.5)}} dx_2.$$





Следовательно, для контрольного объёма D , уравнение баланса (10.176) примет вид

$$h_1 \left(\int_{x_2^{(-0.5)}}^{x_2^{(+0.5)}} L_1 u \Big|_{x_1=x_1^{(-0.5)}} dx_2 \right)_{x_1} + h_2 \left[\int_{x_1^{(-0.5)}}^{x_1^{(+0.5)}} L_2 u \Big|_{x_2=x_2^{(-0.5)}} dx_1 \right]_{x_2} + \int_{x_1^{(-0.5)}}^{x_1^{(+0.5)}} \int_{x_2^{(-0.5)}}^{x_2^{(+0.5)}} f dx_1 dx_2 = 0, \quad (10.177)$$

где $[\dots]_{x_1}, [\dots]_{x_2}$ — правые разностные производные.

Теперь, следуя интегро-интерполяционному методу, аппроксимируем уравнение баланса (10.177) путём интерполяции подынтегральных функций. От способа интерполяции зависят устойчивость, экономичность и аппроксимационные свойства разностной схемы.

Аппроксимируем подынтегральные функции $L_1 u, L_2 u$ некоторыми постоянными значениями на отрезке интегрирования, предполагая, что температура $u(x_1, x_2)$ и тепловые потоки $L_1 u = k \frac{\partial u}{\partial x_1}, L_2 u = k \frac{\partial u}{\partial x_2}$ являются непрерывными функциями, а коэффициент теплопроводности $k(x_1, x_2)$ и производные $\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}$ могут содержать разрывы I рода.

Имеем для первого интеграла:

$$L_1 u = k \frac{\partial u}{\partial x_1} \Rightarrow \frac{\partial u}{\partial x_1} = \frac{1}{k} L_1 u \text{ при } x_1 = x_1^{(-0.5)}, x_2^{(-0.5)} \leq x_2 \leq x_2^{(+0.5)}.$$

Проинтегрируем полученное соотношение по x_1 от $x_1 = x_1^{(-1)}$ до $x_1 = x_1^0$ при любом фиксированном $x_2^{(-0.5)} \leq x_2 \leq x_2^{(+0.5)}$:

$$u(x_1^0, x_2) - u(x_1^{(-1)}, x_2) = \int_{x_1^{(-1)}}^{x_1^0} \frac{1}{k} L_1 u dx_1.$$



Пользуясь непрерывностью $L_1 u$, сделаем приближенную замену:

$$u(x_1^0, x_2) - u(x_1^{(-1)}, x_2) \approx L_1 u \Big|_{x_1=x_1^{(-0.5)}} \int_{x_1^{(-1)}}^{x_1^0} \frac{dx_1}{k} \text{ при } \forall x_2^{(-0.5)} \leq x_2 \leq x_2^{(+0.5)} \Rightarrow$$

$$L_1 u \Big|_{x_1=x_1^{(-0.5)}} \approx [u(x_1^0, x_2) - u(x_1^{(-1)}, x_2)] \left(\int_{x_1^{(-1)}}^{x_1^0} \frac{dx_1}{k} \right)^{-1} \text{ при } \forall x_2^{(-0.5)} \leq x_2 \leq x_2^{(+0.5)} \Rightarrow$$

$$\int_{x_2^{(-0.5)}}^{x_2^{(+0.5)}} L_1 u \Big|_{x_1=x_1^{(-0.5)}} dx_2 \approx \int_{x_2^{(-0.5)}}^{x_2^{(+0.5)}} [u(x_1^0, x_2) - u(x_1^{(-1)}, x_2)] \left(\int_{x_1^{(-1)}}^{x_1^0} \frac{dx_1}{k} \right)^{-1} dx_2 \approx$$

$$\approx (u - u^{(-1)}) \int_{x_2^{(-0.5)}}^{x_2^{(+0.5)}} \left(\int_{x_1^{(-1)}}^{x_1^0} \frac{dx_1}{k} \right)^{-1} dx_2 = h_2 u_{\bar{x}_1} a_1, \quad a_1 = \frac{h_1}{h_2} \int_{x_2^{(-0.5)}}^{x_2^{(+0.5)}} \left(\int_{x_1^{(-1)}}^{x_1^0} \frac{dx_1}{k} \right)^{-1} dx_2,$$

$$\text{где } u = u(x_1^0, x_2^0), u^{(-1)} = u(x_1^{(-1)}, x_2^0).$$

Аналогично,

$$\int_{x_1^{(-0.5)}}^{x_1^{(+0.5)}} L_2 u \Big|_{x_2=x_2^{(-0.5)}} dx_1 \approx h_1 u_{\bar{x}_2} a_2, \quad a_2 = \frac{h_2}{h_1} \int_{x_1^{(-0.5)}}^{x_1^{(+0.5)}} \left(\int_{x_2^{(-1)}}^{x_2^0} \frac{dx_2}{k} \right)^{-1} dx_1.$$

Используя эти приближения в уравнении баланса (10.177), делённом на $h_1 h_2$, получаем консервативную разностную схему:

$$\begin{cases} \Lambda y \equiv (a_1 y_{\bar{x}_1})_{x_1} + (a_2 y_{\bar{x}_2})_{x_2} = -\varphi, & x = (x_1^0, x_2^0) \in \omega_h, \\ y(x) = \mu(x), & x \in \gamma_h, \end{cases} \quad (10.178)$$



где

$$\varphi = \frac{1}{h_1 h_2} \int_{x_1^{(-0.5)}}^{x_1^{(+0.5)}} \int_{x_2^{(-0.5)}}^{x_2^{(+0.5)}} f dx_1 dx_2,$$

$$a_1 = \frac{h_1}{h_2} \int_{x_2^{(-0.5)}}^{x_2^{(+0.5)}} \left(\int_{x_1^{(-1)}}^{x_1^0} \frac{dx_1}{k} \right)^{-1} dx_2,$$

$$a_2 = \frac{h_2}{h_1} \int_{x_1^{(-0.5)}}^{x_1^{(+0.5)}} \left(\int_{x_2^{(-1)}}^{x_2^0} \frac{dx_2}{k} \right)^{-1} dx_1. \quad (10.179)$$

Схема (10.178), (10.179) есть разностное представление интегрального закона сохранения (10.176) для контрольного объёма D_i .

Можно показать, что консервативная схема (4), (5) имеет второй порядок аппроксимации как в случае гладких коэффициентов k, f , так и в случае, если эти коэффициенты имеют разрывы I рода.

Например, в случае гладких коэффициентов будем иметь, используя кубатурную формулу средних в формулах (10.179):

$$\varphi = f + O(|h|^2), \quad a_1 = k^{(-0.5_1)} + O(|h|^2), \quad a_2 = k^{(-0.5_2)} + O(|h|^2).$$

Таким образом, для погрешности аппроксимации схемы (10.178) получаем

$$\begin{aligned} \psi &= (a_1 u_{\bar{x}_1})_{x_1} + (a_2 u_{\bar{x}_2})_{x_2} + \varphi = (k^{(-0.5_1)} u_{\bar{x}_1})_{x_1} + (k^{(-0.5_2)} u_{\bar{x}_2})_{x_2} + f + \\ &+ (O(|h|^2) u_{\bar{x}_1})_{x_1} + (O(|h|^2) u_{\bar{x}_2})_{x_2} + O(|h|^2) = Lu + f + O(|h|^2) = O(|h|^2). \end{aligned}$$

В случае гладких коэффициентов можно, например, положить

$$\varphi = f = f_{i_1 i_2}, \quad a_1 = k^{(-0.5_1)} = k_{i_1 - \frac{1}{2}, i_2}, \quad a_2 = k^{(-0.5_2)} = k_{i_1, i_2 - \frac{1}{2}},$$



сохраняя второй порядок аппроксимации схемы (10.178).

Раскрывая обозначения для разностных производных в схеме (10.178), получим

$$\left\{ \begin{array}{l} \frac{1}{h_1} \left(a_1^{(+1_1)} \frac{y^{(+1_1)} - y}{h_1} - a_1 \frac{y - y^{(-1_1)}}{h_1} \right) + \frac{1}{h_2} \left(a_2^{(+1_2)} \frac{y^{(+1_2)} - y}{h_2} - a_2 \frac{y - y^{(-1_2)}}{h_2} \right) = -\varphi, \\ x \in \omega_h, \\ y(x) = \mu(x), \quad x \in \gamma_h. \end{array} \right.$$

Теперь очевидно, что схема (10.178), (10.179) удовлетворяет [принципу максимума](#) при любых h_1, h_2 .

1. Показать верность последнего утверждения.

Используя это обстоятельство, можно показать, что эта схема *абсолютно устойчива*. Для ее реализации можно применять известные итерационные методы решения СЛАУ.



10.5. Итерационные методы решения разностных задач

[10.5.1. Двухслойные итерационные схемы](#)

[10.5.2. Свойства разностного оператора Лапласа](#)

[10.5.3. Метод простых итераций](#)

[10.5.4. Методы Зейделя и релаксации](#)

Аппроксимация краевых задач математической физики разностными задачами приводит, как правило, к СЛАУ $Ay = f$.

В результате решения этой системы искомая сеточная функция $y(x)$ определяется во всех узлах сетки $\bar{\omega}_h$. Матрица A этой системы имеет большой порядок, равный числу узлов сетки. Если решается p -мерная задача, то число узлов сетки $\bar{\omega}_h$ равно $N = N_1 \cdot N_2 \cdots \cdot N_p$, где N_α — число узлов по переменной x_α , $\alpha = \overline{1, p}$. Таким образом, при $p = 2, 3$ число уравнений может быть очень большим. Например, положив $N_1 = N_2 = 100$ при $p = 2$, имеем $N = 10^4$, а положив $N_1 = N_2 = N_3 = 100$ при $p = 3$, имеем $N = 10^6$ уравнений.

Матрица A , как правило, имеет ленточную структуру, то есть много нулевых элементов. Эта особенность позволяет разрабатывать специальные экономичные алгоритмы.

Как и для всякой СЛАУ, существуют прямые и итерационные методы решения разностных задач. Мы остановимся лишь на итерационных методах, которые являются наиболее универсальным средством решения разностных задач.



Меню

10.5.1. Двухслойные итерационные схемы

Пусть требуется решить разностную задачу

$$\begin{cases} Ay = \varphi(x), & x \in \omega_h, \\ y(x) = \mu(x), & x \in \gamma_h. \end{cases} \quad (10.180)$$

Решение задачи (10.180) ищется в пространстве сеточных функций H_h , определённых на сетке $\bar{\omega}_h = \omega_h \cup \gamma_h$.

Пусть $H_h \subseteq H_h$ — пространство сеточных функций, определённых на сетке $\bar{\omega}_h$ и обращающихся в ноль при $x \in \gamma_h$. Если $\gamma_h = \emptyset$, то $H_h = H_h$.

Определение. Оператор A называется *самосопряжённым* в $\overset{\circ}{H}_h$ ($A = A^*$), если $\forall u, v \in \overset{\circ}{H}_h$ выполняется $(Au, v) = (u, Av)$.

Определение. Оператор A называется *положительно определённым* в $\overset{\circ}{H}_h$ ($A > 0$), если $\forall u \in \overset{\circ}{H}_h$ имеем $(Au, u) \geq 0$, причём $(Au, u) = 0 \Leftrightarrow u = 0$.

Будем обозначать y^n итерационное приближение номер n к точному решению $y(x)$ задачи (10.180).

Определение. *Двухслойной итерационной схемой* будем называть всякий линейный одношаговый итерационный метод решения задачи (10.180), представимый в виде

$$\begin{cases} B \frac{y^{n+1} - y^n}{\tau} + Ay^n = \varphi, & x \in \omega_h, \\ y^{n+1} = \mu(x), & x \in \gamma_h, n = 0, 1, 2, \dots, \end{cases} \quad (10.181)$$

где $y^0 \in H_h$; $\tau > 0$ — параметр релаксации, B — линейный оператор (матрица).

Итерационные методы различаются матрицей B и параметром τ , выбор которых подчиняется требованиям сходимости $y^n \xrightarrow{n \rightarrow \infty} y$ и экономичности, то есть получению решения с заданной точностью ε за минимальное число арифметических действий.



Теорема 10.8 (Достаточное условие сходимости двухслойных итерационных схем). Пусть $A = A^* > 0$ в пространстве сеточных функций $\overset{\circ}{H}_h$. Тогда, если

$$B - \frac{\tau}{2}A > 0 \text{ в пространстве } \overset{\circ}{H}_h, \quad (10.182)$$

то итерационный процесс (10.181) сходится.

[[Доказательство](#)]

При доказательстве мы предполагали, что $\varepsilon^{n+1} - \varepsilon^n \neq 0$, $n = 0, 1, 2, \dots$. Если $\varepsilon^{n+1} - \varepsilon^n = 0$, то в силу (Д.20) получаем $A\varepsilon^n = 0$. Но это возможно тогда и только тогда, когда $\varepsilon^n = y^n - y = 0$, то есть $y^n = y \Rightarrow y^{n+1} = y$ и т.д.



10.5.2. Свойства разностного оператора Лапласа

В качестве примера рассмотрим разностный оператор

$$A = -\Lambda y = -y_{\bar{x}_1 \bar{x}_1} - y_{\bar{x}_2 \bar{x}_2}, \quad x \in \omega_h.$$

Покажем, что он является **самосопряжённым** и **положительно определённым** в пространстве $\overset{\circ}{H}_h$. Пусть $u, v \in \overset{\circ}{H}_h$. Тогда

$$\begin{aligned} (Au, v) &= \sum_{i=1}^{N_1-1} \sum_{j=1}^{N_2-1} \left[\left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) u_{ij} - \frac{1}{h_1^2} (u_{i-1,j} + u_{i+1,j}) - \frac{1}{h_2^2} (u_{i,j-1} + u_{i,j+1}) \right] v_{ij} = \\ &= \sum_{i=1}^{N_1-1} \sum_{j=1}^{N_2-1} \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) u_{ij} v_{ij} - \sum_{i=1}^{N_1-1} \sum_{j=1}^{N_2-1} \frac{1}{h_1^2} u_{ij} v_{i+1,j} - \sum_{i=1}^{N_1-1} \sum_{j=1}^{N_2-1} \frac{1}{h_1^2} u_{ij} v_{i-1,j} - \\ &\quad - \sum_{i=1}^{N_1-1} \sum_{j=1}^{N_2-1} \frac{1}{h_2^2} u_{ij} v_{i,j+1} - \sum_{i=1}^{N_1-1} \sum_{j=1}^{N_2-1} \frac{1}{h_2^2} u_{ij} v_{i,j-1} = \\ &= \sum_{i=1}^{N_1-1} \sum_{j=1}^{N_2-1} \left[\left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) v_{ij} - \frac{1}{h_1^2} (v_{i-1,j} + v_{i+1,j}) - \frac{1}{h_2^2} (v_{i,j-1} + v_{i,j+1}) \right] u_{ij} = (u, Av). \end{aligned}$$

$$\begin{aligned} (Au, u) &= \sum_{i=1}^{N_1-1} \sum_{j=1}^{N_2-1} \left[\left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) u_{ij}^2 - \frac{1}{h_1^2} (u_{i-1,j} + u_{i+1,j}) u_{ij} - \frac{1}{h_2^2} (u_{i,j-1} + u_{i,j+1}) u_{ij} \right] = \\ &= \sum_{j=1}^{N_2-1} \frac{1}{h_1^2} \sum_{i=1}^{N_1} (u_{i-1,j} - u_{ij})^2 + \sum_{i=1}^{N_1-1} \frac{1}{h_2^2} \sum_{j=1}^{N_2} (u_{i,j-1} - u_{ij})^2 \geq 0, \end{aligned}$$

причём $(Au, u) = 0$ тогда и только тогда, когда $u(x) = \text{const}$. Но, поскольку $u \in \overset{\circ}{H}_h$, то $u(x) = 0$. Таким образом, $(Au, u) = 0 \Leftrightarrow u = 0$.



Меню

10.5.3. Метод простых итераций

В качестве примера рассмотрим разностную задачу Дирихле для уравнения Пуассона в прямоугольнике

$$\begin{cases} Ay = -\Lambda y = -y_{x_1 x_1} - y_{x_2 x_2} = \varphi(x), & x \in \omega_h, \\ y(x) = \mu(x), & x \in \gamma_h. \end{cases} \quad (10.183)$$

В индексной форме:

$$\left\{ \begin{array}{l} \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) y_{ij} - \frac{1}{h_1^2} (y_{i-1,j} + y_{i+1,j}) - \frac{1}{h_2^2} (y_{i,j-1} + y_{i,j+1}) = \varphi_{ij}, \\ i = \overline{1, N_1 - 1}, j = \overline{1, N_2 - 1}, \\ y_{0j} = \mu_{0j}, \quad y_{N_1 j} = \mu_{N_1 j}, \quad j = \overline{0, N_2}, \\ y_{i0} = \mu_{i0}, \quad y_{iN_2} = \mu_{iN_2}, \quad i = \overline{0, N_1}. \end{array} \right. \quad (10.184)$$

Матрица СЛАУ (10.184) является трёхдиагональной, причём диагональным элементам соответствуют y_{ij} , $i = \overline{0, N_1}$, $j = \overline{0, N_2}$.

Алгоритм *метода простых итераций*:

$$\left\{ \begin{array}{l} y_{ij}^{n+1} = \frac{1}{\frac{2}{h_1^2} + \frac{2}{h_2^2}} \left(\frac{1}{h_1^2} (y_{i-1,j}^n + y_{i+1,j}^n) + \frac{1}{h_2^2} (y_{i,j-1}^n + y_{i,j+1}^n) + \varphi_{ij} \right), \\ i = \overline{1, N_1 - 1}, j = \overline{1, N_2 - 1}, \\ y_{0j}^{n+1} = \mu_{0j}, \quad y_{N_1 j}^{n+1} = \mu_{N_1 j}, \quad j = \overline{0, N_2}, \\ y_{i0}^{n+1} = \mu_{i0}, \quad y_{iN_2}^{n+1} = \mu_{iN_2}, \quad i = \overline{0, N_1}. \end{array} \right. \quad (10.185)$$

Следствие 10.5. Метод простых итераций (10.185) сходится.

[[Доказательство](#)]



10.5.4. Методы Зейделя и релаксации

Метод Зейделя для задачи (10.183):

$$\begin{cases} y_{ij}^{n+1} = \frac{1}{\frac{2}{h_1^2} + \frac{2}{h_2^2}} \left[\frac{1}{h_1^2} \left(y_{i-1,j}^{n+1} + y_{i+1,j}^n \right) + \frac{1}{h_2^2} \left(y_{i,j-1}^{n+1} + y_{i,j+1}^n \right) + \varphi_{ij} \right], \\ i = \overline{1, N_1 - 1}, j = \overline{1, N_2 - 1}, \\ y_{0j}^{n+1} = \mu_{0j}, y_{N_1 j}^{n+1} = \mu_{N_1 j}, j = \overline{0, N_2}, \\ y_{i0}^{n+1} = \mu_{i0}, y_{iN_2}^{n+1} = \mu_{iN_2}, i = \overline{0, N_1}. \end{cases} \quad (10.186)$$

Метод релаксации:

$$\begin{cases} y_{ij}^{n+1} = (1 - \tau) y_{ij}^n + \tau \frac{1}{\frac{2}{h_1^2} + \frac{2}{h_2^2}} \left[\frac{1}{h_1^2} \left(y_{i-1,j}^{n+1} + y_{i+1,j}^n \right) + \frac{1}{h_2^2} \left(y_{i,j-1}^{n+1} + y_{i,j+1}^n \right) + \varphi_{ij} \right], \\ i = \overline{1, N_1 - 1}, j = \overline{1, N_2 - 1}, \end{cases} \quad (10.187)$$

где τ — параметр релаксации. При $\tau = 1$ этот метод совпадает с методом Зейделя.

Следствие 10.6. Если параметр релаксации τ удовлетворяет условию $0 < \tau < 2$, то метод релаксации (10.187) сходится. [\[Доказательство\]](#)

Следствие 10.7. Метод Зейделя (10.186) сходится.

Замечание 10.21.

Определение. В случае задачи Дирихле (10.183) для уравнения Пуассона в прямоугольнике теория численных методов даёт оптимальное значение параметра релаксации

$$\tau_{opt} = 2 \Big/ \left\{ 1 + \sqrt{1 - \left[1 - \frac{\pi^2}{2} \frac{h_1^2 h_2^2}{h_1^2 + h_2^2} \left(\frac{1}{l_1} + \frac{1}{l_2} \right) \right]^2} \right\}, \quad (10.188)$$



при котором достигается максимальная скорость сходимости метода релаксации (10.187). Очевидно, что $1 < \tau_{opt} < 2$. Поэтому метод релаксации (10.187) при $\tau = \tau_{opt}$ называют *методом верхней релаксации*.

Замечание 10.22. Для того, чтобы получить решение задачи (10.183) с точностью ε по методу верхней релаксации (10.187), (10.188) при $h_1 = h_2 = h = \frac{1}{N}$, $l_1 = l_2 = 1$, требуется

$$n_{min}(\varepsilon) = O(N \ln \frac{1}{\varepsilon})$$

итераций, что значительно меньше, чем по методу Зейделя ($\tau = 1$), когда

$$n_{min}(\varepsilon) = O(N^2 \ln \frac{1}{\varepsilon}).$$

В свою очередь, метод Зейделя является более быстрым, чем метод простых итераций.



10.6. Численные методы решения задач математической физики в областях сложной формы

10.6.1. Метод замены переменных

10.6.2. Разностный метод

10.6.3. Метод конечных элементов

10.6.4. Метод граничных элементов



Меню

Часть III. Теоретические материалы

Глава 10. Численные методы математической физики

10.6. Численное решение задач в областях сложной формы

10.6.1. Метод замены переменных

10.6.1. Метод замены переменных

Метод замены переменных: ищется преобразование, переводящее область сложной формы в простую область — как правило, прямоугольной формы. Затем задача математической физики, переформулированная в новых переменных, решается в прямоугольнике обычным методом конечных разностей на прямоугольной сетке.

Недостатки данного метода:

- 1) чтобы найти преобразование переменных, в общем случае нужно решать дополнительную нелинейную дифференциальную краевую задачу;
- 2) в новых переменных усложняется вид исходной дифференциальной задачи МФ. В частности, в уравнениях возникают смешанные производные, аппроксимация которых, как правило, ухудшает устойчивость разностной схемы.



Меню

10.6.2. Разностный метод

Рассмотрим задачу Дирихле

$$\begin{cases} Lu = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = -f(x), & x = (x_1, x_2) \in G, \\ u|_{\Gamma} = \mu(x), & x \in \Gamma, \end{cases} \quad (10.189)$$

где $\bar{G} = G \cup \Gamma$ — область сложной формы.

Вводится прямоугольная сетка, покрывающая всю область \bar{G} , т. е. строятся два семейства прямых

$$\begin{cases} x_1 = x_{1,i_1}, & i_1 = 0, \pm 1, \pm 2, \dots, \\ x_2 = x_{2,i_2}, & i_2 = 0, \pm 1, \pm 2, \dots, \end{cases}$$

так, что $x_{1,i_1} > x_{1,i_1-1}$, $x_{2,i_2} > x_{2,i_2-1}$.

Определение. Точки пересечения прямых прямоугольной сетки

$$x = x_i = (x_{1,i_1}, x_{2,i_2}),$$

лежащие внутри области \bar{G} , т. е. $x_i \in G$, назовём *внутренними узлами*. Множество внутренних узлов обозначим $\hat{\omega}_h \subset G$.

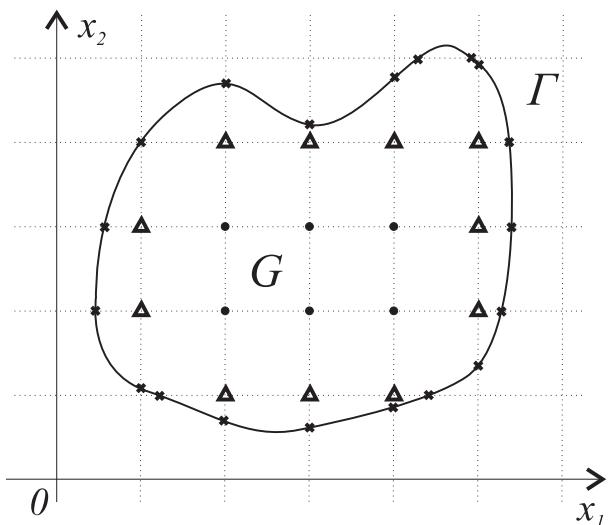
Определение. Точки пересечения этих прямых с границей Γ назовём *граничными узлами*. Это множество обозначим $\hat{\gamma}_h \subset \Gamma$.

Определение. Множество *внутренних* и *граничных* узлов назовём *сеткой* $\hat{\omega}_h = \hat{\omega}_h \cup \hat{\gamma}_h$ в области $\bar{G} = G \cup \Gamma$.

Если область \bar{G} , в которой ищется решение задачи (10.189), имеет криволинейную границу, то построить равномерную *сетку* $\bar{\omega}_h \subset \bar{G}$, вообще говоря, невозможно.

Для любых $x = (x_1, x_2) \in \hat{\omega}_h$ рассмотрим разностные операторы на нерегулярном шаблоне:

$$\begin{aligned} \Lambda_1^* y = y_{\bar{x}_1 \bar{x}_1} &= \frac{1}{h_1} \left(\frac{y^{(+1_1)} - y}{h_1^{(+1)}} - \frac{y - y^{(-1_1)}}{h_1} \right) = \\ &= \frac{1}{h_1} \left(\frac{y(x_1 + h_1^{(+1)}, x_2) - y(x_1, x_2)}{h_1^{(+1)}} - \frac{y(x_1, x_2) - y(x_1 - h_1, x_2)}{h_1} \right), \end{aligned}$$





Меню

Часть III. Теоретические материалы

Глава 10. Численные методы математической физики

10.6. Численное решение задач в областях сложной формы

10.6.2. Разностный метод

$$\begin{aligned}\Lambda_2^* y = y_{\tilde{x}_2 \hat{x}_2} &= \frac{1}{h_2} \left(\frac{y^{(+1_2)} - y}{h_2^{(+1)}} - \frac{y - y^{(-1_2)}}{h_2} \right) = \\ &= \frac{1}{h_2} \left(\frac{y(x_2 + h_2^{(+1)}, x_2) - y(x_1, x_2)}{h_2^{(+1)}} - \frac{y(x_1, x_2) - y(x_1 - h_2, x_2)}{h_2} \right),\end{aligned}$$

где

$$h_\alpha = \frac{1}{2} \left(h_\alpha + h_\alpha^{(+1)} \right),$$

h_1 — расстояние между центральным узлом x шаблона $\Pi(x) \subset \hat{\omega}_h$ и периферийным узлом $x^{(-1_1)} = (x_1^{(-1)}, x_2) \in \Pi'(x) \subset \hat{\omega}_h$, лежащим слева от узла x ,
 $h_1^{(+1)}$ — расстояние между узлом x и узлом $x^{(+1_1)} = (x_1^{(+1)}, x_2) \in \Pi'(x) \subset \hat{\omega}_h$, лежащим справа от узла x .

Аналогично определяются h_2 и $h_2^{(+1)}$.

С помощью этих операторов построим для задачи (10.189) разностную схему

$$\Lambda^* y \equiv y_{\tilde{x}_1 \hat{x}_1} + y_{\tilde{x}_2 \hat{x}_2} = -\varphi(x), \quad x \in \hat{\omega}_h, \quad y(x) = \mu(x), \quad x \in \hat{\gamma}_h. \quad (10.190)$$

Упражнения:

- 1) Показать, что $\forall \varphi(x) = f(x) + O(h^m)$, $m \geq 1$ схема (10.190) имеет первый порядок аппроксимации $\psi(x) = O(h)$, где $h = \max(h_1, h_2)$ на $\hat{\omega}_h$.
- 2) Показать, что разностная схема (10.190) удовлетворяет принципу максимума при любых h_1, h_2 .

Недостатки метода, связанные с прямоугольной сеткой:

- 1) сложность с построением адаптивной сетки;

Определение. Адаптивная сетка — [сетка](#), у которой плотность узлов на участках сильного изменения решения более высокая, чем на участках слабого изменения решения.

- 2) сложности с определением [внутренних](#) и [граничных узлов](#) $\hat{\omega}_h$ в случае, когда граница Γ является подвижной или описывается сеточной функцией.



10.6.3. Метод конечных элементов

Триангуляция области и базисные пирамидальные функции

МКЭ на основе метода Галеркина

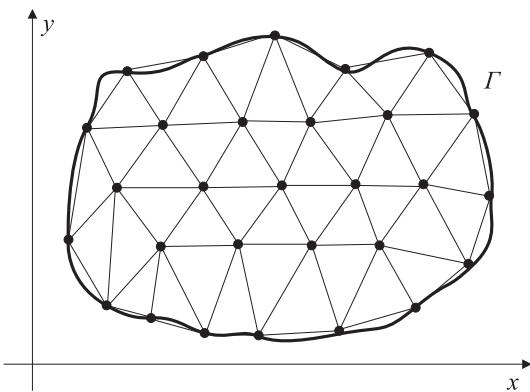
Интегро-интерполяционный метод конечных элементов

Будем рассматривать задачу Дирихле

$$\begin{cases} Lu \equiv \frac{\partial}{\partial x}(k \frac{\partial u}{\partial x}) + \frac{\partial}{\partial y}(k \frac{\partial u}{\partial y}) = -f(\xi), \quad \xi = (x, y) \in G, \quad k(\xi) > 0, \\ u|_{\Gamma} = \mu(\xi), \quad \xi \in \Gamma, \end{cases} \quad (10.191)$$

где $\bar{G} = G \cup \Gamma$ — область сложной формы.

Триангуляция области и базисные пирамидальные функции



В *методе конечных элементов (МКЭ)* область \bar{G} обычно покрывается треугольной сеткой $\bar{\omega}_h = \omega_h \cup \gamma_h$.



Определение. Узлы сетки, произвольно выбранные в области и на границе Γ , соединяются не пересекающимися отрезками так, чтобы каждый внутренний узел был вершиной 6 треугольников (элементов). Такое построение сетки называется *триангуляцией области* \bar{G} .

Так как граничные узлы также соединяются между собой отрезками, то криволинейная граница Γ заменяется ломаной линией.

Пусть в области \bar{G} содержится N внутренних и M граничных узлов, которые мы пронумеруем с помощью одного индекса:

$$\begin{aligned}\xi_i &= (x_i, y_i) \in \omega_h, \quad i = 1, 2, \dots, N, \\ \xi_i &= (x_i, y_i) \in \gamma_h, \quad i = N+1, N+2, \dots, N+M.\end{aligned}$$

Обозначим $\mathcal{M} = \{1, 2, \dots, N, N+1, \dots, N+M\}$ — множество индексов сеточных узлов.

Пусть $\xi_i = (x_i, y_i) \in \omega_h$ — внутренний узел. Обозначим Ω_i шестиугольник, состоящий из треугольных элементов $\Delta_1, \Delta_2, \dots, \Delta_6$, примыкающих к узлу ξ_i , т. е.

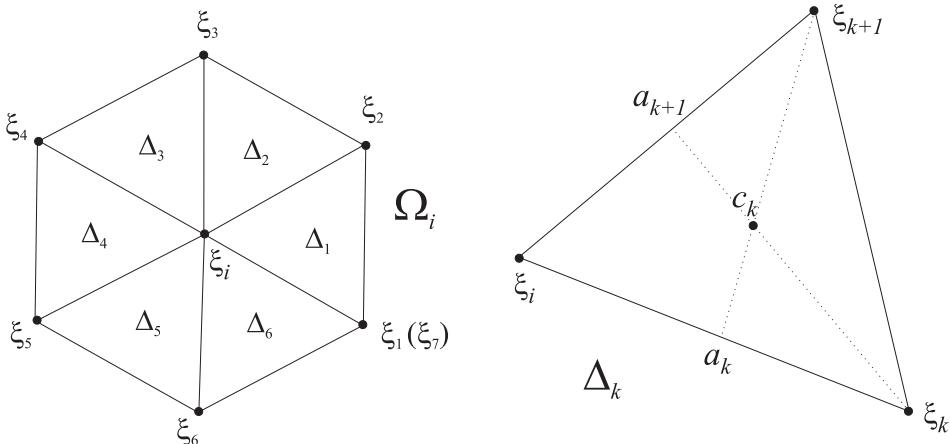
$$\Omega_i = \bigcup_{p=1}^6 \Delta_p \subset \bar{G} \text{ — окрестность узла } \xi_i.$$

Вершинами треугольника являются узлы

$$\xi_i = (x_i, y_i), \quad \xi_p = (x_p, y_p), \quad \xi_{p+1} = (x_{p+1}, y_{p+1}), \quad p = \overline{1, 6}.$$

Площадь треугольника Δ_p вычисляется по формулам:

$$\begin{aligned}S_{\Delta_p} &= \frac{1}{2} \begin{vmatrix} x_p - x_i & y_p - y_i \\ x_{p+1} - x_i & y_{p+1} - y_i \end{vmatrix} = \\ &= \frac{1}{2} \begin{vmatrix} x_{p+1} - x_p & y_{p+1} - y_p \\ x_i - x_p & y_i - y_p \end{vmatrix} = \\ &= \frac{1}{2} \begin{vmatrix} x_i - x_{p+1} & y_i - y_{p+1} \\ x_p - x_{p+1} & y_p - y_{p+1} \end{vmatrix} > 0.\end{aligned}$$



Центр масс треугольника Δ_p (точка пересечения медиан):

$$c_p = (x_{c_p}, y_{c_p}) = \frac{1}{3}(\xi_i + \xi_p + \xi_{p+1}),$$

т. е.

$$x_{c_p} = \frac{1}{3}(x_i + x_p + x_{p+1}), \quad y_{c_p} = \frac{1}{3}(y_i + y_p + y_{p+1}).$$

Обозначим

$$a_p = \frac{1}{2}(\xi_i + \xi_p) — \text{середина стороны } [\xi_i, \xi_p],$$

$$a_{p+1} = \frac{1}{2}(\xi_i + \xi_{p+1}) — \text{середина стороны } [\xi_i, \xi_{p+1}].$$

Таким образом, $x_{a_p} = \frac{1}{2}(x_i + x_p)$, $y_{a_p} = \frac{1}{2}(y_i + y_p)$.

Определение. *Базисными пирамидальными функциями*, заданными на треугольных элементах сетки $\bar{\omega}_h$, называют систему линейно независимых функций $\varphi_1(\xi), \dots, \varphi_N(\xi)$, каждая из которых отлична от нуля лишь в окрестности одного из внутренних узлов $\xi_1, \dots, \xi_N \in \omega_h$.



Меню

Часть III. Теоретические материалы

Глава 10. Численные методы математической физики

10.6. Численное решение задач в областях сложной формы

10.6.3. Метод конечных элементов

Базисная функция $\varphi_i(\xi)$, соответствующая узлу $\xi \in \omega_h$, имеет вид

$$\varphi_i(\xi) = \begin{cases} \varphi_i^{(p)}(\xi), & \xi = (x, y) \in \Delta_p, p = \overline{1, 6}, \\ 0, & \xi = (x, y) \notin \Omega_i, \end{cases}$$

где

$$\varphi_i^{(p)}(\xi_i) = \frac{1}{2S_p} [(x - x_{p+1})(y_p - y_{p+1}) - (y - y_{p+1})(x_p - x_{p+1})], \quad p = \overline{1, 6}$$

— линейная функция, обладающая свойствами

$$\varphi_i^{(p)}(\xi_i) = 1, \quad \varphi_i^{(p)}(\xi_{p+1}) = \varphi_i^{(p)}(\xi_p) = 0.$$

Следовательно,

$$\varphi_i(\xi_i) = 1, \quad \varphi_i(\xi_p) = 0, \quad p = \overline{1, 6}.$$

Определение. *Носителем* функции называют наименьшую замкнутую область, вне которой функция тождественно равна нулю.

Носителем базисной функции $\varphi_i(\xi)$ по построению является шестиугольник Ω_i .

Определение. *Пирамидальной функцией* $\varphi_i(x)$, соответствующей узлу $\xi_i = (x_i, y_i) \in \bar{\omega}_h$, будем называть кусочно-линейную функцию вида

$$\varphi_i(\xi) = \begin{cases} \varphi_i^{(p)}(\xi) = \frac{(x - x_{p+1})(y_p - y_{p+1}) - (y - y_{p+1})(x_p - x_{p+1})}{2S_p}, & \xi = (x, y) \in \Delta_p, p = \overline{1, m}, \\ 0, \xi \notin \Omega_i = \bigcup_{p=1}^m \Delta_p, & \end{cases} \quad i = \overline{1, N + M}. \quad (10.192)$$

Если $\xi_i \in \omega_h$, то $m = 6$, Ω_i — шестиугольник.

Если $\xi_i \in \gamma_h$, то $1 < m < 6$.

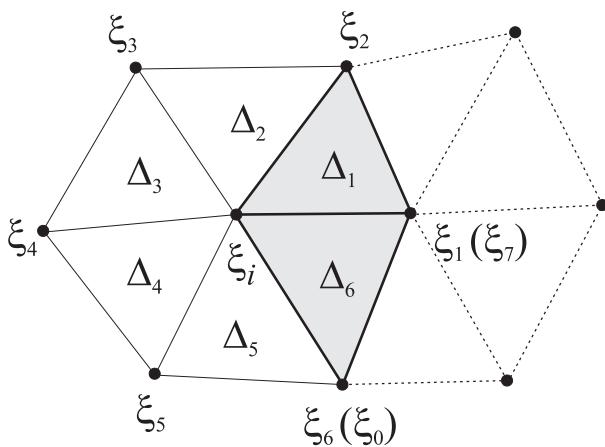
Линейная функция $\varphi_i^{(p)}(\xi)$ определена на треугольном элементе Δ_p и обладает свойствами

$$\varphi_i^{(p)}(\xi_i) = 1, \quad \varphi_i^{(p)}(\xi_p) = 0, \quad \varphi_i^{(p)}(\xi_{p+1}) = 0.$$

Следовательно,

$$\varphi_i(\xi_i) = 1, \quad \varphi_i(\xi_p) = 0, \quad p = \overline{1, m}.$$

График функции $\varphi_i(\xi)$ является боковой поверхностью пирамиды единичной высоты с основанием Ω_i . Пусть узел ξ_i — [внутренний](#). Так как носитель Ω_i есть шестиугольник, пирамида $\varphi_i(\xi)$ — шестигранная. Пусть $\varphi_1(\xi), \dots, \varphi_6(\xi)$ — [пирамидальные функции](#), соответствующие [периферийным узлам](#) ξ_1, \dots, ξ_6 , являющимся вершинами шестиугольника Ω_i .



Очевидно, что [носитель](#) Ω_i [пирамидальной функции](#) $\varphi_i(\xi)$ и [носитель](#) Ω_p [пирамидальной функции](#) $\varphi_p(\xi)$, $p = \overline{1, 6}$ пересекаются на треугольниках Δ_p, Δ_{p-1} , т. е. $\Omega_i \cap \Omega_p = \Delta_p \cup \Delta_{p-1}$.

На треугольнике Δ_p отличны от нуля лишь три базисные функции

$$\varphi_i(\xi) = \varphi_i^{(p)}(\xi), \quad \varphi_p(\xi) = \varphi_p^{(p)}(\xi), \quad \varphi_{p+1}(\xi) = \varphi_{p+1}^{(p)}(\xi), \quad \xi \in \Delta_p,$$

отвечающие вершинам треугольника Δ_p , так как $\Omega_i \cap \Omega_p \cap \Omega_{p+1} = \Delta_p$.

Эти функции на треугольнике Δ_p имеют вид:

$$\left\{ \begin{array}{l} \varphi_i^{(p)}(\xi) = \frac{(x-x_{p+1})(y_p-y_{p+1})-(y-y_{p+1})(x_p-x_{p+1})}{2S_p}, \\ \varphi_p^{(p)}(\xi) = \frac{(x-x_i)(y_{p+1}-y_i)-(y-y_i)(x_{p+1}-x_i)}{2S_p}, \\ \varphi_{p+1}^{(p)}(\xi) = \frac{(x-x_p)(y_i-y_p)-(y-y_p)(x_i-x_p)}{2S_p}, \end{array} \right. \quad (10.193)$$

$$\xi = (x, y) \in \Delta_p, \quad p = \overline{1, 6}$$

и обладают свойствами

$$\left\{ \begin{array}{l} \varphi_i^{(p)}(\xi_i) = 1, \quad \varphi_i^{(p)}(\xi_p) = \varphi_i^{(p)}(\xi_{p+1}) = 0, \\ \frac{\partial \varphi_i^{(p)}}{\partial x} = \frac{y_p - y_{p+1}}{2S_p}, \quad \frac{\partial \varphi_i^{(p)}}{\partial y} = -\frac{x_p - x_{p+1}}{2S_p}, \\ \varphi_p^{(p)}(\xi_p) = 1, \quad \varphi_p^{(p)}(\xi_i) = \varphi_p^{(p)}(\xi_{p+1}) = 0, \\ \frac{\partial \varphi_p^{(p)}}{\partial x} = \frac{y_{p+1} - y_i}{2S_p}, \quad \frac{\partial \varphi_p^{(p)}}{\partial y} = -\frac{x_{p+1} - x_i}{2S_p}, \\ \varphi_{p+1}^{(p)}(\xi_{p+1}) = 1, \quad \varphi_{p+1}^{(p)}(\xi_i) = \varphi_{p+1}^{(p)}(\xi_p) = 0, \\ \frac{\partial \varphi_{p+1}^{(p)}}{\partial x} = \frac{y_i - y_p}{2S_p}, \quad \frac{\partial \varphi_{p+1}^{(p)}}{\partial y} = -\frac{x_i - x_p}{2S_p}. \end{array} \right. \quad (10.194)$$

МКЭ на основе метода Галеркина

Схемы МКЭ часто строят с помощью вариационных методов (Ритца, наименьших квадратов), а также методом Галеркина. При этом линейно независимые [пирамидальные функции](#) $\varphi_i(\xi)$, $i = \overline{1, N}$, соответствующие внутренним узлам $\xi_1, \dots, \xi_N \in \omega_h$, служат в качестве базисных функций.

[Пирамидальные функции](#) $\varphi_{N+1}(\xi), \dots, \varphi_{N+M}(\xi)$, соответствующие граничным узлам $\xi_{N+1}, \dots, \xi_{N+M} \in \gamma_h$, используются для того, чтобы обеспечить выполнение граничных условий задачи (10.191).

Приближенное решение дифференциальной задачи (10.191) на сетке треугольных элементов ищется в аналитическом виде

$$\tilde{u}(\xi) = \bar{u}(\xi) + \sum_{p=1}^N \tilde{u}_p \varphi_p(\xi), \quad \xi \in \bar{G}, \quad (10.195)$$

где $\bar{u}(\xi)$ призвано обеспечить выполнение граничных условий:

$$\bar{u}(\xi) = \sum_{p=N+1}^{N+M} u_p \varphi_p(\xi), \quad u_p = u(\xi_p) = \mu(\xi_p), \quad \xi_p \in \gamma_h,$$

где \tilde{u}_p — неизвестные коэффициенты, требующие определения,

$\varphi_1, \dots, \varphi_N$ — пирамидальные функции вида (10.193), соответствующие внутренним узлам $\xi_1, \dots, \xi_N \in \omega_h$,
 $\varphi_{N+1}, \dots, \varphi_{N+M}$ — [пирамидальные функции](#), соответствующие граничным узлам $\xi_{N+1}, \dots, \xi_{N+M} \in \gamma_h$.

Заметим, что

$$\varphi_i(\xi_i) = 1, \quad \varphi_i(\xi_p) = 0 \quad \forall i = \overline{1, N+M}, \quad p \neq i.$$

В силу этого

$$\bar{u}(\xi_i) = \sum_{p=N+1}^{N+M} u_p \varphi_p(\xi_i) = u_i \varphi_i(\xi_i) = u_i = \mu(\xi_i), \quad i = \overline{N+1, N+M}.$$

Следовательно,

$$\begin{cases} \tilde{u}(\xi_i) = \bar{u}(\xi_i) + \sum_{p=1}^N \tilde{u}_p \varphi_p(\xi_i) = [\varphi_p(\xi_i) = 0] = \bar{u}(\xi_i) = \mu(\xi_i) \text{ при } \xi_i \in \gamma_h, \\ \tilde{u}(\xi_i) = \bar{u}(\xi_i) + \sum_{p=1}^N \tilde{u}_p \varphi_p(\xi_i) = [\bar{u}(\xi_i) = 0] = \tilde{u}_i(\xi_i) \varphi_i(\xi_i) = \tilde{u}_i \text{ при } \xi_i \in \omega_h, \end{cases}$$

т. е. искомая функция (10.195) точно удовлетворяет граничным условиям задачи (10.191) в узлах $\xi_i \in \gamma_h$ и принимает значения \tilde{u}_i во внутренних узлах $\xi_i \in \omega_h$.

Таким образом, неизвестные коэффициенты \tilde{u}_i представляют собой значения [приближенного решения](#) $\tilde{u}(\xi)$ во внутренних узлах $\xi_i, i = \overline{1, N}$. Задачу (10.191) можно считать решённой, если найдены все $\tilde{u}_i, i = \overline{1, N}$.

Для удобства функцию $\tilde{u}(\xi)$ будем рассматривать в виде

$$\tilde{u}(\xi) = \sum_{p=1}^{N+M} \tilde{u}_p \varphi_p(\xi), \quad \xi \in \bar{G}, \quad (10.196)$$

полагая

$$\tilde{u}_p = \mu(\xi_p) \text{ при } p = \overline{N+1, N+M}.$$

Для построения алгоритма метода конечных элементов решения задачи (10.191) воспользуемся [методом Галёркина](#), согласно которому неизвестные коэффициенты $\tilde{u}_1, \dots, \tilde{u}_N$ определяются из условий

$$(L\tilde{u} + f, \varphi_i) = 0, \quad i = \overline{1, N} \quad (10.197)$$

или

$$(L\tilde{u}, \varphi_i) + (f, \varphi_i) = 0, \quad i = \overline{1, N},$$

где

$$Lu = \nabla(k\nabla u) = \frac{\partial}{\partial x}(k \frac{\partial u}{\partial x}) + \frac{\partial}{\partial y}(k \frac{\partial u}{\partial y}), \quad (f, g) = \int_{\bar{G}} f g dxdy.$$

Чтобы ослабить требования на гладкость базисных функций $\varphi_i(\xi)$, выполним интегрирование по частям в области \bar{G} :

$$\begin{aligned} (Lu, \varphi_i) &= (\nabla(k\nabla u), \varphi_i) = \int_{\bar{G}} \nabla(k\nabla u) \cdot \varphi_i dxdy = \\ &= \int_{\Gamma} k \frac{\partial u}{\partial n} \varphi_i d\Gamma - \int_{\bar{G}} k \nabla u \cdot \nabla \varphi_i dxdy = [\varphi_i = 0 \text{ на границе } \Gamma] = \\ &= - \int_{\bar{G}} k \nabla u \cdot \nabla \varphi_i dxdy, \quad i = \overline{1, N}. \end{aligned}$$



Тогда условия Галеркина (10.197) принимают вид

$$\int_{\tilde{G}} k \nabla \tilde{u} \cdot \nabla \varphi_i dx dy = (f, \varphi_i), \quad i = \overline{1, N}. \quad (10.198)$$

Обычно (10.197) называют *прямой формулировкой метода Галёркина*, а (10.198) — *слабой формулировкой метода Галёркина*.

Ввиду того, что $\varphi_i \equiv 0$ при $\xi = (x, y) \notin \Omega_i$, будем иметь

$$\int_{\Omega_i} k \nabla \tilde{u} \cdot \nabla \varphi_i dx dy = (f, \varphi_i), \quad i = \overline{1, N}, \text{ где } (f, \varphi_i) = \int_{\Omega_i} f \varphi_i dx dy.$$

Подставляя сюда (10.196), получим систему линейных алгебраических уравнений относительно $\tilde{u}_1, \dots, \tilde{u}_N$:

$$\sum_{p=1}^{N+M} \tilde{u}_p \int_{\Omega_i} k \nabla \varphi_p \cdot \nabla \varphi_i dx dy = (f, \varphi_i), \quad i = \overline{1, N}, \quad (10.199)$$

где Ω_i — носитель функции $\varphi_i(\xi)$.

Пусть Ω_i — шестиугольник с вершинами $\xi_1 = \xi_7, \xi_2, \xi_3, \xi_4, \xi_5, \xi_6 = \xi_0$, которым соответствуют базисные функции $\varphi_1(\xi) = \varphi_7(\xi), \varphi_2(\xi), \varphi_3(\xi), \varphi_4(\xi), \varphi_5(\xi), \varphi_6(\xi) = \varphi_0(\xi)$. Поскольку на шестиугольнике Ω_i обращаются в нуль все базисные функции, за исключением $\varphi_i, \varphi_1, \dots, \varphi_6$, то в уравнениях (10.199) остаются лишь те слагаемые, которые соответствуют $p = \overline{1, 6}$ и $p = i$. Таким образом, система уравнений (10.199) имеет вид:

$$\tilde{u}_i \int_{\Omega_i} k |\nabla \varphi_i|^2 dx dy + \sum_{p=1}^6 \tilde{u}_p \int_{\Omega_i} (k \nabla \varphi_p \cdot \nabla \varphi_i) dx dy = (f, \varphi_i), \quad i = \overline{1, N}$$

или

$$A_i \tilde{u}_i = \sum_{p=1}^6 B_{ip} \tilde{u}_p + F_i, \quad i = \overline{1, N}, \quad (10.200)$$



где $\tilde{u}_p = \mu(\xi_p)$, если $\xi_p \in \gamma_h$.

$$\begin{aligned}
 B_{ip} &= - \int_{\Omega_i} k \nabla \varphi_p \cdot \nabla \varphi_i dx dy = - \int_{\Delta_p} \underbrace{k (\nabla \varphi_p^{(p)} \cdot \nabla \varphi_i^{(p)})}_{=const} dx dy - \\
 &\quad - \int_{\Delta_{p-1}} \underbrace{k (\nabla \varphi_p^{(p-1)} \cdot \nabla \varphi_i^{(p-1)})}_{=const} dx dy = [r_p = \int_{\Delta_p} k dx dy, \Omega_i \cap \Omega_p = \Delta_{p-1} \cup \Delta_p] = \\
 &= - \left(\frac{\partial \varphi_p^{(p)}}{\partial x} \cdot \frac{\partial \varphi_i^{(p)}}{\partial x} + \frac{\partial \varphi_p^{(p)}}{\partial y} \cdot \frac{\partial \varphi_i^{(p)}}{\partial y} \right) r_p - \left(\frac{\partial \varphi_p^{(p-1)}}{\partial x} \cdot \frac{\partial \varphi_i^{(p-1)}}{\partial x} + \frac{\partial \varphi_p^{(p-1)}}{\partial y} \cdot \frac{\partial \varphi_i^{(p-1)}}{\partial y} \right) r_{p-1} = \\
 &= - \left(\frac{y_{p+1} - y_i}{2S_p} \cdot \frac{y_p - y_{p+1}}{2S_p} + \frac{x_{p+1} - x_i}{2S_p} \cdot \frac{x_p - x_{p+1}}{2S_p} \right) r_p - \\
 &\quad - \left(\frac{y_i - y_{p-1}}{2S_p} \cdot \frac{y_{p-1} - y_p}{2S_p} + \frac{x_i - x_{p-1}}{2S_p} \cdot \frac{x_{p-1} - x_p}{2S_p} \right) r_{p-1} = \\
 &= \frac{1}{4S_p^2} [(x_p - x_{p+1})(x_i - x_{p+1}) + (y_p - y_{p+1})(y_i - y_{p+1})] r_p + \\
 &\quad + \frac{1}{4S_{p-1}^2} [(x_{p-1} - x_p)(x_{p-1} - x_i) + (y_{p-1} - y_p)(y_{p-1} - y_i)] r_{p-1}.
 \end{aligned}$$

$$\begin{aligned}
 F_i = (f, \varphi_i) &= \int_{\Omega_i} f \varphi_i dx dy \approx f_i \int_{\Omega_i} \varphi_i dx dy = \\
 &= \left[\int_{\Omega_i} \varphi_i dx dy - \text{объём пирамиды} \right] = f_i \frac{1}{3} \sum_{p=1}^6 S_p.
 \end{aligned}$$



$$\begin{aligned}
 A_i &= \int_{\Omega_i} k |\nabla \varphi_i|^2 dx dy = \sum_{p=1}^6 \int_{\Delta_p}^{const} k |\nabla \varphi_i^{(p)}|^2 dx dy = \sum_{p=1}^6 |\nabla \varphi_i^{(p)}|^2 r_p = \\
 &= \sum_{p=1}^6 \left[\left(\frac{\partial \varphi_i^{(p)}}{\partial x} \right)^2 + \left(\frac{\partial \varphi_i^{(p)}}{\partial y} \right)^2 \right] r_p = \sum_{p=1}^6 \left[\left(\frac{y_p - y_{p+1}}{2S_p} \right)^2 + \left(\frac{x_p - x_{p+1}}{2S_p} \right)^2 \right] r_p = \\
 &= \sum_{p=1}^6 \frac{1}{4S_p^2} [(x_{p+1} - x_p)^2 + (y_{p+1} - y_p)^2] r_p > 0.
 \end{aligned}$$

Упражнение:

Показать, что $A_i = \sum_{p=1}^6 B_{ip}$.

Замечание 10.23. При построении конечно-элементной схемы (10.200) во внутреннем узле $\xi_i \in \omega_h$ использована упрощённая нумерация вершин шестиугольника Ω_i , т. е. ξ_1, \dots, ξ_6 , не учитывающая зависимости от номера центрального узла ξ_i . Более корректной является нумерация этих вершин индексами $i_1 = i_7, i_2, \dots, i_5, i_6 = i_0 \in \mathcal{M}$. Перенумеровав в схеме (10.200) $\xi_1 = \xi_7$ в $\xi_{i_1} = \xi_{i_7}$, ξ_2 в ξ_{i_2}, \dots, ξ_5 в ξ_{i_5} , $\xi_6 = \xi_0$ в $\xi_{i_6} = \xi_{i_0}$, запишем схему (10.200) в виде

$$\begin{cases} A_i \tilde{u}_i = \sum_{p=1}^6 B_{ip} \tilde{u}_{i_p} + F_i, & i = \overline{1, N}, \\ \tilde{u}_{i_p} = \mu(\xi_{i_p}), & i_p = N + 1, \dots, N + M, \end{cases} \quad (10.201)$$

где

$$\tilde{u}_{i_p} = \tilde{u}(\xi_{i_p}), \quad \xi_{i_p} = (x_{i_p}, y_{i_p}),$$

$$\begin{aligned}
 B_{ip} &= \frac{1}{4S_{i_p}^2} [(x_{i_p} - x_{i_{p+1}})(x_i - x_{i_{p+1}}) + (y_{i_p} - y_{i_{p+1}})(y_i - y_{i_{p+1}})] r_{i_p} + \\
 &\quad + \frac{1}{4S_{i_{p-1}}^2} [(x_{i_{p-1}} - x_{i_p})(x_{i_{p-1}} - x_i) + (y_{i_{p-1}} - y_{i_p})(y_{i_{p-1}} - y_i)] r_{i_{p-1}},
 \end{aligned}$$



$$F_i = \frac{1}{3} f_i \sum_{p=1}^6 S_{i_p}, \quad A_i = \sum_{p=1}^6 B_{ip}, \quad r_{ip} = \int_{\Delta_{ip}} k dxdy.$$

$S_{ip} = \frac{1}{2} [(x_{ip} - x_i)(y_{ip+1} - y_i) - (x_{ip+1} - x_i)(y_{ip} - y_i)] > 0$ — площадь треугольного элемента Δ_{ip} с вершинами $\xi_i = (x_i, y_i)$, $\xi_{ip} = (x_{ip}, y_{ip})$, $\xi_{ip+1} = (x_{ip+1}, y_{ip+1})$.

Замечание 10.24. Можно показать, что если элементы сетки $\bar{\omega}_h$ являются остроугольными треугольниками, то все $B_{ip} > 0$. В этом случае схема (10.201) удовлетворяет [принципу максимума](#), а матрица СЛАУ (10.201) будет иметь [диагональное преобладание](#), на основании которого для ее решения можно применять известные [итерационные методы](#).

Интегро-интерполяционный метод конечных элементов

$$\begin{cases} \frac{\partial}{\partial x}(L_1 u) + \frac{\partial}{\partial x}(L_2 u) + f(\xi) = 0, \\ u|_{\Gamma} = \mu(\xi), \quad \xi \in \Gamma. \end{cases} \quad (10.202)$$

где $\xi = (x, y) \in G$, $k(\xi) > 0$, $L_1 u = k(\xi) \frac{\partial u}{\partial x}$, $L_2 u = k(\xi) \frac{\partial u}{\partial y}$, $\bar{G} = G \cup \Gamma$ — область сложной формы.

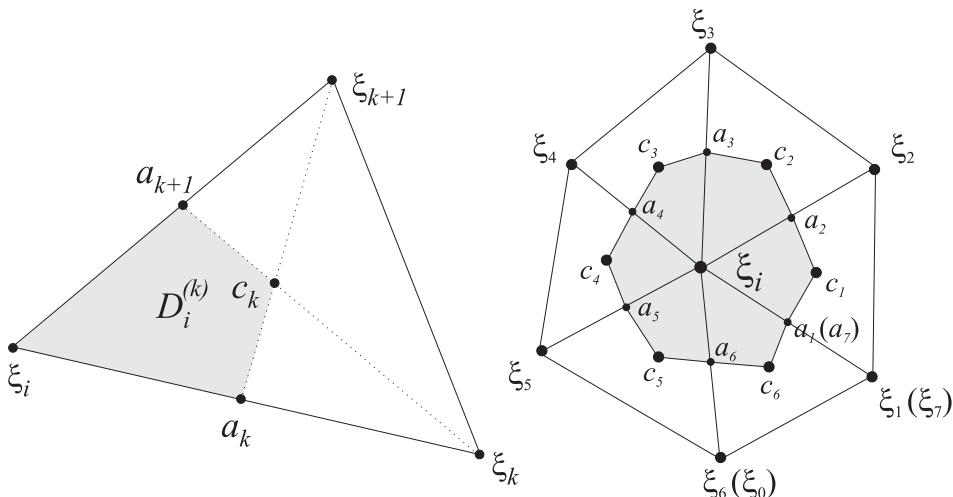
Рассмотрим уравнение баланса, соответствующее уравнению Пуассона (10.191), для произвольной области $D \subseteq \bar{G}$ (C — граница области D) (см. §2 главы 3):

$$\int_C (L_1 u dy - L_2 u dx) + \iint_D f dxdy = 0, \quad (10.203)$$

Пусть $\xi \in \omega_h$ — внутренний узел треугольной сетки, $a_p = \frac{1}{2}(\xi_i + \xi_p)$ — середина стороны $[\xi_i, \xi_p]$ треугольника Δ_p , $c_p = \frac{1}{3}(\xi_i + \xi_p + \xi_{p+1})$ — центр масс треугольника Δ_p .

На каждом шестиугольнике Ω_i выделим контрольный объем $D_i \subset \Omega_i$, соединив на каждом треугольнике Δ_p , $p = \overline{1, 6}$ отрезками последовательно три точки — a_p , c_p и a_{p+1} . Таким образом, на каждом шестиугольнике Ω_i контрольный объем D_i представляет собой двенадцатиугольник, не обязательно выпуклый, с замкнутой границей (ломаной линией)

$$\Gamma_i = a_1 c_1 a_2 c_2 a_3 c_3 a_4 c_4 a_5 c_5 a_6 c_6 a_1.$$



Ту часть контрольного объема и ту часть его границы, которые принадлежат треугольнику Δ_p , обозначим $D_i^{(p)}$ и $\Gamma_i^{(p)}$ соответственно, т. е.

$$\Gamma_i^{(p)} = [a_p, c_p] \cup [c_p, a_{p+1}].$$

Теперь, следуя интегро-интерполяционному методу, запишем уравнение баланса (10.203) для контрольного объема $D_i = \bigcup_{p=1}^6 D_i^{(p)}$ с границей $\Gamma_i = \bigcup_{p=1}^6 \Gamma_i^{(p)}$:

$$\sum_{p=1}^6 \int_{\Gamma_i^{(p)}} (L_1 u dy - L_2 u dx) + \iint_{D_i} f dx dy = 0. \quad (10.204)$$

Следующий этап — аппроксимация уравнения баланса (10.203). Применим простейшие интерполяции для подынтегральных функций:

$$u(\xi) \approx u^{(p)}(\xi) = u_i \varphi_i^{(p)}(\xi) + u_p \varphi_p^{(p)}(\xi) + u_{p+1} \varphi_{p+1}^{(p)}(\xi) \text{ при } \xi \in \Delta_p, p = \overline{1, 6}, \quad (10.205)$$



где $u^{(p)}(\xi)$ — линейная функция, определённая на треугольнике Δ_p и совпадающая с функцией $u(\xi)$ в вершинах треугольника, т. е. на треугольнике Δ_p она представляет собой плоскость, проходящую через точки (ξ_i, u_i) , (ξ_p, u_p) , (ξ_{p+1}, u_{p+1}) , $u_i = u(\xi_i)$, $u_p = u(\xi_p)$, $u_{p+1} = u(\xi_{p+1})$.

Тогда при $\xi = (x, y) \in \Delta_p$ будем иметь

$$\begin{aligned}\frac{\partial u^{(p)}}{\partial x} &= u_i \frac{\partial \varphi_i^{(p)}}{\partial x} + u_p \frac{\partial \varphi_p^{(p)}}{\partial x} + u_{p+1} \frac{\partial \varphi_{p+1}^{(p)}}{\partial x} = \\ &= u_i \frac{y_p - y_{p+1}}{2S_p} + u_p \frac{y_{p+1} - y_i}{2S_p} + u_{p+1} \frac{y_i - y_p}{2S_p} = \\ &= \frac{1}{2S_p} [(u_p - u_i)(y_{p+1} - y_i) - (u_{p+1} - u_i)(y_p - y_i)] = \text{const.}\end{aligned}$$

$$\begin{aligned}\frac{\partial u^{(p)}}{\partial y} &= u_i \frac{\partial \varphi_i^{(p)}}{\partial y} + u_p \frac{\partial \varphi_p^{(p)}}{\partial y} + u_{p+1} \frac{\partial \varphi_{p+1}^{(p)}}{\partial y} = \\ &= -u_i \frac{x_p - x_{p+1}}{2S_p} - u_p \frac{x_{p+1} - x_i}{2S_p} - u_{p+1} \frac{x_i - x_p}{2S_p} = \\ &= -\frac{1}{2S_p} [(u_p - u_i)(x_{p+1} - x_i) - (u_{p+1} - u_i)(x_p - x_i)] = \text{const.}\end{aligned}$$



Отсюда следует:

$$\begin{aligned}
 \int_{\Gamma_i^{(p)}} (L_1 u dy - L_2 u dx) &\approx \frac{\partial u^{(p)}}{\partial x} \int_{\Gamma_i^{(p)}} k dy - \frac{\partial u^{(p)}}{\partial y} \int_{\Gamma_i^{(p)}} k dx = \\
 &= [q_{i_p} = \int_{\Gamma_i^{(p)}} k dy, r_{i_p} = \int_{\Gamma_i^{(p)}} k dx] = \frac{\partial u^{(p)}}{\partial x} q_{i_p} - \frac{\partial u^{(p)}}{\partial y} r_{i_p} = \\
 &= (u_p - u_i) \frac{1}{2S_p} \left[(y_{p+1} - y_p) q_{i_p} + (x_{p+1} - x_p) r_{i_p} \right] + \\
 &\quad + (u_{p+1} - u_i) \frac{1}{2S_p} \left[-(y_p - y_i) q_{i_p} - (x_p - x_i) r_{i_p} \right] \Rightarrow \\
 \sum_{p=1}^6 \int_{\Gamma_i^{(p)}} (L_1 u dy - L_2 u dx) &\approx \sum_{p=1}^6 (u_p - u_i) B_{ip}, \tag{10.206}
 \end{aligned}$$

где

$$\begin{aligned}
 B_{ip} &= \frac{1}{2S_p} [r_{i_p}(x_{p+1} - x_i) + q_{i_p}(y_i - y_{p+1})] + \\
 &\quad + \frac{1}{2S_{p-1}} [(-r_{i_{p-1}}x_{p-1} - x_p) - q_{i_{p-1}}(y_{p-1} - y_i)]. \tag{10.207}
 \end{aligned}$$

$$r_{i_p} = \int_{\Gamma_i^{(p)}} k dx \approx k(c_p) \int_{\Gamma_i^{(p)}} dx = k(c_p)(x_{a_{p-1}} - x_{a_p}) = \frac{1}{2} k \left(\frac{\xi_i + \xi_p + \xi_{p+1}}{3} \right) (x_{p+1} - x_p).$$

Применяя аппроксимацию (10.205), (10.206) к уравнению баланса (10.203), получаем конечно-элементную схему

$$\begin{cases} \sum_{p=1}^6 B_{ip} (\tilde{u}_p - \tilde{u}_i) = -F_i, & \xi_i \in \omega_h, \\ \tilde{u}(\xi_i) = \mu(\xi_i), & \xi_i \in \gamma_h, \end{cases} \tag{10.208}$$



Меню

Часть III. Теоретические материалы

Глава 10. Численные методы математической физики

10.6. Численное решение задач в областях сложной формы

10.6.3. Метод конечных элементов

где $\tilde{u}(\xi)$, $\xi \in \bar{\omega}_h$ — искомая сеточная функция, $\tilde{u}_i \approx u_i = u(\xi_i)$, $\tilde{u}_p \approx u_p = u(\xi_p)$ — приближенные значения искомого решения $u(\xi)$ в узлах ξ_i и ξ_p , коэффициенты B_{ip} вычисляются по формуле (10.206). Правая часть:

$$F_i = \iint_{D_i} f(\xi) dx dy \approx f_i \sum_{p=1}^6 \iint_{D_i^{(p)}} dx dy = f_i \sum_{p=1}^6 \frac{1}{3} S_p. \quad (10.209)$$

Переобозначая периферийные узлы

$$\xi_1 = \xi_7 \rightarrow \xi_{i_1} = \xi_{i_7}, \quad \xi_2 \rightarrow \xi_{i_2}, \dots, \quad \xi_5 \rightarrow \xi_{i_5}, \quad \xi_6 = \xi_0 \rightarrow \xi_{i_6} = \xi_{i_0},$$

запишем схему (10.206)-(10.208) в более корректной форме

$$\begin{cases} A_i \tilde{u}_i = \sum_{p=1}^6 B_{ip} \tilde{u}_{i_p} + F_i, & i = \overline{1, N}, \\ \tilde{u}_{i_p} = \mu(\xi_{i_p}), & i_p = \overline{N+1, N+M}, \end{cases} \quad (10.210)$$

где

$$A_i = \sum_{p=1}^6 B_{ip}, \quad F_i = \frac{1}{3} f_i \sum_{p=1}^6 S_{i_p},$$

Замечание 10.25. Конечно-элементная схема (10.209), полученная интегро-интерполяционным методом, полностью совпадает со схемой (10.201), построенной по методу Галеркина. Поэтому к ней относятся те же замечания, что и к схеме (10.201).

Замечание 10.26. Конечно-элементные схемы, построенные как с помощью интегро-интерполяционного метода, так и по методу Галеркина, являются консервативными.



10.6.4. Метод граничных элементов

[Первая аппроксимация](#)

[Вторая аппроксимация](#)

Область применимости метода граничных элементов — это уравнения математической физики, для которых известно фундаментальное решение. Это, например, уравнения Лапласа, Гельмгольца, однородное уравнение теплопроводности и другие.

Пусть $\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$ — оператор Лапласа в двумерном пространстве, $u = u(\xi)$, $\xi = (x, y)$ — точка пространства. Пусть $\bar{G} = G \cup \Gamma$ — произвольная ограниченная область в двумерном пространстве с границей Γ . Для достаточно гладких в области \bar{G} функций $u(\xi)$, $v(\xi)$ справедлива *вторая формула Грина*:

$$\iint_{\bar{G}} u \nabla^2 v dx dy = \oint_{\Gamma} \left(u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n} \right) d\Gamma + \iint_{\bar{G}} v \nabla^2 u dx dy, \quad (10.211)$$

где n — внешняя нормаль к границе Γ .

Пусть $\xi^0 = (x^0, y^0) \in \bar{G}$ — фиксированная точка (точка наблюдения), $\xi = (x, y) \in \bar{G}$ — текущая точка с переменными координатами x , y .

Определение. Функция $u^*(\xi^0, \xi)$ называется *фундаментальным решением уравнения Лапласа* $\nabla^2 u = 0$, если

$$\nabla^2 u^* = -\pi \alpha \delta(\xi^0, \xi),$$

где δ — дельта-функция Дирака, т. е.

1)

$$\nabla^2 u^*(\xi^0, \xi) = \begin{cases} 0, & \xi \neq \xi^0, \\ +\infty, & \xi = \xi^0. \end{cases}$$

2)

$$\iint_{\Omega} g(\xi) \nabla^2 u^*(\xi^0, \xi) dx dy = -\pi \alpha g(\xi^0)$$

для любой фиксированной точки ξ^0 , любой ее окрестности $\Omega \subseteq \bar{G}$, любой функции $g(\xi) = g(x, y)$, непрерывной в точке ξ^0 , где

$$\alpha = \begin{cases} 1, & \xi^0 \in \Gamma, \\ 2, & \xi^0 \notin \Gamma, \text{ т. е. } \xi^0 \in G. \end{cases}$$

Фундаментальное решение двумерного уравнения Лапласа $\nabla^2 u = 0$ имеет вид

$$u^*(\xi^0, \xi) = \ln \frac{1}{\rho(\xi^0, \xi)} = -\ln \rho(\xi^0, \xi) = -\frac{1}{2} \ln [(x^0 - x)^2 + (y^0 - y)^2], \quad (10.212)$$

где $\rho(\xi^0, \xi) = \sqrt{(x^0 - x)^2 + (y^0 - y)^2}$ — расстояние от точки наблюдения до текущей точки.

Можно показать, что формула Грина (10.211) справедлива и для функции $v(\xi) = u^*(\xi^0, \xi)$, имеющей особенность в точке $\xi = \xi^0$ — разрыв второго рода. С учётом свойств фундаментального решения, получим:

$$-\pi \alpha u(\xi^0) = \oint_{\Gamma} \left(u \frac{\partial u^*}{\partial n} - u^* \frac{\partial u}{\partial n} \right) d\Gamma + \iint_{\bar{G}} u^* \nabla^2 u dx dy \quad \forall \xi^0 \in \bar{G},$$

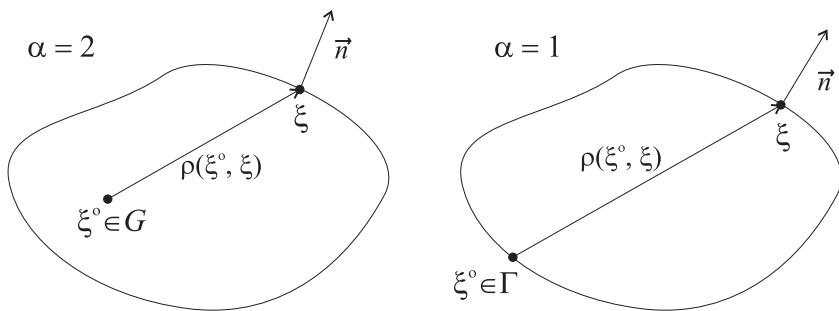
$\alpha = 1$ при $\xi^0 \in \Gamma$, $\alpha = 2$ при $\xi^0 \in G$ ($\xi^0 \notin \Gamma$).

Если $u(\xi)$ — решение уравнения Лапласа $\nabla^2 u = 0$, то

$$-\pi \alpha u(\xi^0) = \oint_{\Gamma} (u q^* - u^* q) d\Gamma, \quad \alpha = \begin{cases} 1, & \xi^0 \in \Gamma, \\ 2, & \xi^0 \notin \Gamma, \end{cases} \quad (10.213)$$

где

$$q = q(\xi) = \left. \frac{\partial u(\xi)}{\partial n} \right|_{\Gamma}, \quad q^* = q^*(\xi^0, \xi) = \left. \frac{\partial u^*(\xi^0, \xi)}{\partial n} \right|_{\Gamma},$$



$$\frac{\partial u^*}{\partial n} = \cos(\widehat{x, n}) \frac{\partial u^*}{\partial x} + \cos(\widehat{y, n}) \frac{\partial u^*}{\partial y} = \cos(\widehat{x, n}) \frac{x - x^0}{\rho^2(\xi^0, \xi)} + \cos(\widehat{y, n}) \frac{y - y^0}{\rho^2(\xi^0, \xi)}.$$

\vec{n} — внешняя нормаль к границе Γ в точке $\xi = (x, y) \in \Gamma$.

Таким образом, решение уравнения Лапласа $\nabla^2 u = 0$ в области $\tilde{G} = G \cup \Gamma$ удовлетворяет интегральному уравнению (10.213).

Метод граничных элементов основан на решении уравнения (10.213).

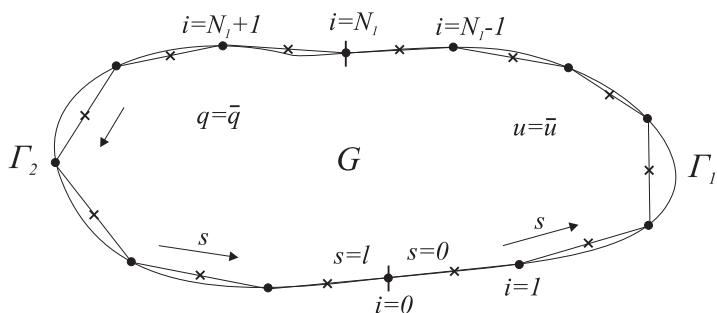
Рассмотрим задачу Дирихле-Неймана для уравнения Лапласа:

$$\begin{cases} \nabla^2 u = 0, \quad \xi = (x, y) \in G, \\ u(\xi) = \bar{u}(\xi), \quad \xi \in \Gamma_1, \\ \frac{\partial u(\xi)}{\partial n} = \bar{q}(\xi), \quad \xi \in \Gamma_2, \quad \Gamma_1 \cup \Gamma_2 = \Gamma. \end{cases} \quad (10.214)$$

Уравнение (10.213) для задачи (10.214) принимает вид

$$-\pi \alpha u(\xi^0) = \int_{\Gamma_1} (\bar{u} q^* - u^* q) d\Gamma + \int_{\Gamma_2} (u q^* - u^* \bar{q}) d\Gamma, \quad (10.215)$$

где неизвестными на Γ_1 являются $q(\xi)$, а на Γ_2 — $u(\xi)$. Левая часть, $\pi \alpha u(\xi^0)$, зависит от того, где находится ξ^0 : $\alpha = 1$, если $\xi^0 \in \Gamma$, $\alpha = 2$, если $\xi^0 \notin \Gamma$.



Если $\xi^0 \in \Gamma$, то уравнение (10.215) удобно рассматривать в виде двух уравнений:

$$\begin{cases} \int_{\Gamma_1} u^* q d\Gamma - \int_{\Gamma_2} u q^* d\Gamma = \pi \bar{u}(\xi^0) + \int_{\Gamma_1} \bar{u} q^* d\Gamma - \int_{\Gamma_2} u^* \bar{q} d\Gamma, & \text{если } \xi^0 \in \Gamma_1, \\ -\pi u(\xi^0) + \int_{\Gamma_1} u^* q d\Gamma - \int_{\Gamma_2} u q^* d\Gamma = \int_{\Gamma_1} \bar{u} q^* d\Gamma - \int_{\Gamma_2} u^* \bar{q} d\Gamma, & \text{если } \xi^0 \in \Gamma_2. \end{cases} \quad (10.216)$$

Присутствующие в левой части уравнений (10.216) $u(\xi)$ и $q(\xi)$ — неизвестные функции, требующие определения. В правой части все функции заданы.

Введём на границе Γ_1 сетку узлов $\xi_i = (x_i, y_i) \in \Gamma_1$, $i = \overline{0, N_1}$, а на границе Γ_2 — сетку узлов $\xi_i = (x_i, y_i) \in \Gamma_2$, $i = \overline{N_1, N_2}$, причём $\xi_0 = \xi_{N_2}$.

Первая аппроксимация

Следуя МКЭ, аппроксимируем контур Γ замкнутой ломаной линией $\tilde{\Gamma}$, соединяя отрезками соседние узлы ξ_{j-1} и ξ_j при $j = \overline{1, N_2}$. Тогда

$$\tilde{\Gamma} = \bigcup_{j=1}^{N_2} \tilde{\Gamma}_j,$$

где $\tilde{\Gamma}_j = [\xi_{j-1}, \xi_j]$ — отрезок, соединяющий точки ξ_{j-1} и ξ_j . Эти отрезки называются *граничными элементами*.

Примем обозначения:

$$\Delta x_j = x_j - x_{j-1}, \quad \Delta y_j = y_j - y_{j-1},$$

$$x_{j-\frac{1}{2}} = \frac{1}{2}(x_{j-1} + x_j) = x_{j-1} + \frac{1}{2}\Delta x_j = x_j - \frac{1}{2}\Delta x_j,$$

$$y_{j-\frac{1}{2}} = \frac{1}{2}(y_{j-1} + y_j) = y_{j-1} + \frac{1}{2}\Delta y_j = y_j - \frac{1}{2}\Delta y_j,$$

$$l_j = \sqrt{\Delta x_j^2 + \Delta y_j^2} — длина отрезка \tilde{\Gamma}_j.$$

Запишем в этих обозначениях параметрические уравнения j -го элемента $\tilde{\Gamma}_j = [\xi_{j-1}, \xi_j]$:

$$x(s) = x'_j(s - s_{j-\frac{1}{2}}) + x_{j-\frac{1}{2}}, \quad y(s) = y'_j(s - s_{j-\frac{1}{2}}) + y_{j-\frac{1}{2}}, \quad (10.217)$$

где

$$x'_j = \frac{\Delta x_j}{l_j}, \quad y'_j = \frac{\Delta y_j}{l_j} \Rightarrow (x'_j)^2 + (y'_j)^2 = 1,$$

s — длина дуги линии $\tilde{\Gamma}$, изменяющаяся от $s = s_0 = 0$ в точке ξ_0 до $s = s_{N_2} = l = \sum_{j=1}^{N_2} l_j$ — в точке $\xi_{N_2} = \xi_0$

после прохождения всего контура $\tilde{\Gamma}$. При этом

$$s_j - s_{j-1} = l_j,$$

$$x(s_{j-1}) = \frac{\Delta x_j}{l_j}(s_{j-1} - s_{j-\frac{1}{2}}) + x_{j-\frac{1}{2}} = x_{j-\frac{1}{2}} - \frac{1}{2}\Delta x_j = x_{j-1},$$

$$x(s_j) = \frac{\Delta x_j}{l_j}(s_j - s_{j-\frac{1}{2}}) + x_{j-\frac{1}{2}} = x_{j-\frac{1}{2}} + \frac{1}{2}\Delta x_j = x_j.$$

Задачу (10.214) будем рассматривать в области $\tilde{G} = \tilde{G} \cup \tilde{\Gamma}$, где \tilde{G} — область, ограниченная ломаной $\tilde{\Gamma}$, т. е.

$$\nabla^2 u = 0, \quad \xi \in \tilde{G}, \quad u(\xi) = \bar{u}(\xi), \quad \xi \in \tilde{\Gamma}_1, \quad q(\xi) = \bar{q}(\xi), \quad \xi \in \tilde{\Gamma}_2.$$



Следуя МКЭ, будем выбирать в качестве точки наблюдения ξ^0 граничные точки $\xi_{i-\frac{1}{2}} \in \tilde{\Gamma}$, $i = \overline{1, N_2}$.

Запишем уравнения (10.216) для границы $\tilde{\Gamma}$:

$$\left\{ \begin{array}{l} \sum_{j=1}^{N_1} \int_{s_{j-1}}^{s_j} u^* q ds - \sum_{j=N_1+1}^{N_2} \int_{s_{j-1}}^{s_j} u q^* ds = \pi \bar{u}_{i-\frac{1}{2}} + \sum_{j=1}^{N_1} \int_{s_{j-1}}^{s_j} \bar{u} q^* ds - \sum_{j=N_1+1}^{N_2} \int_{s_{j-1}}^{s_j} u^* \bar{q} ds, \\ i = \overline{1, N_1}, \\ -\pi u_{i-\frac{1}{2}} + \sum_{j=1}^{N_1} \int_{s_{j-1}}^{s_j} u^* q ds - \sum_{j=N_1+1}^{N_2} \int_{s_{j-1}}^{s_j} u q^* ds = \sum_{j=1}^{N_1} \int_{s_{j-1}}^{s_j} \bar{u} q^* ds - \sum_{j=N_1+1}^{N_2} \int_{s_{j-1}}^{s_j} u^* \bar{q} ds, \\ i = \overline{N_1 + 1, N_2}. \end{array} \right. \quad (10.218)$$

Вторая аппроксимация

Аппроксимируем функции $u(\xi)$, $q(\xi)$, $\bar{u}(\xi)$, $\bar{q}(\xi)$ на каждом элементе константами:

$$u(\xi) \approx u_{j-\frac{1}{2}}, \quad q(\xi) \approx q_{j-\frac{1}{2}}, \quad \bar{u}(\xi) \approx \bar{u}_{j-\frac{1}{2}}, \quad \bar{q}(\xi) \approx \bar{q}_{j-\frac{1}{2}} \text{ при } \xi \in \tilde{\Gamma}_j = [\xi_{j-1}, \xi_j].$$

Тогда система (10.218) принимает вид СЛАУ относительно неизвестных $u_{i-\frac{1}{2}}$ при $i = \overline{1, N_1}$, $q_{i-\frac{1}{2}}$ при $i = \overline{N_1 + 1, N_2}$:

$$\left\{ \begin{array}{l} \sum_{j=1}^{N_1} a_{ij} q_{j-\frac{1}{2}} - \sum_{j=N_1+1}^{N_2} b_{ij} u_{j-\frac{1}{2}} = \pi \bar{u}_{i-\frac{1}{2}} + \sum_{j=1}^{N_1} b_{ij} \bar{u}_{j-\frac{1}{2}} - \sum_{j=N_1+1}^{N_2} a_{ij} \bar{q}_{j-\frac{1}{2}}, \\ i = \overline{1, N_1}, \\ -\pi u_{i-\frac{1}{2}} + \sum_{j=1}^{N_1} a_{ij} q_{j-\frac{1}{2}} - \sum_{j=N_1+1}^{N_2} b_{ij} u_{j-\frac{1}{2}} = \sum_{j=1}^{N_1} b_{ij} \bar{u}_{j-\frac{1}{2}} - \sum_{j=N_1+1}^{N_2} a_{ij} \bar{q}_{j-\frac{1}{2}}, \\ i = \overline{N_1 + 1, N_2}, \end{array} \right. \quad (10.219)$$

где

$$a_{ij} = \int_{s_{j-1}}^{s_j} u^*(\xi_{i-\frac{1}{2}}, \xi) ds, \quad b_{ij} = \int_{s_{j-1}}^{s_j} q^*(\xi_{i-\frac{1}{2}}, \xi) ds.$$

Обозначая неизвестные одной буквой

$$v_i = \begin{cases} q_{i-\frac{1}{2}}, & i = \overline{1, N_1}, \\ u_{i-\frac{1}{2}}, & i = \overline{N_1 + 1, N_2}, \end{cases}$$

запишем систему линейных алгебраических уравнений в общем виде

$$\sum_{j=1}^{N_2} A_{ij} v_j = F_i, \quad i = \overline{1, N_2} \quad (10.220)$$

с коэффициентами

$$A_{ij} = \begin{cases} a_{ij}, & j = \overline{1, N_1}, \\ b_{ij}, & j = \overline{N_1 + 1, N_2}, \end{cases} \quad (i \neq j), \quad A_{ii} = \begin{cases} a_{ii}, & i = \overline{1, N_1}, \\ -\pi - b_{ii}, & i = \overline{N_1 + 1, N_2}, \end{cases}$$

$$F_i = \begin{cases} \pi \bar{u}_{i-\frac{1}{2}} + \sum_{j=1}^{N_1} b_{ij} \bar{u}_{j-\frac{1}{2}} - \sum_{j=N_1+1}^{N_2} a_{ij} \bar{q}_{j-\frac{1}{2}}, & i = \overline{1, N_1}, \\ \sum_{j=1}^{N_1} b_{ij} \bar{u}_{j-\frac{1}{2}} - \sum_{j=N_1+1}^{N_2} a_{ij} \bar{q}_{j-\frac{1}{2}}, & i = \overline{N_1 + 1, N_2}. \end{cases} \quad (10.221)$$

Матрица системы линейных алгебраических уравнений (10.220) является матрицей общего вида, поэтому для ее решения применяется [метод Гаусса](#). Решая по методу Гаусса систему линейных алгебраических уравнений (10.220), найдём приближенное решение задачи (10.214) на границе Γ области \bar{G} .

Займёмся вычислением коэффициентов a_{ij} и b_{ij} , определяющих матрицу и вектор правой части системы (10.220).

$$\begin{aligned} \rho^2(\xi_{i-\frac{1}{2}}, \xi) &= ((x_{i-\frac{1}{2}} - x(s))^2 + ((y_{i-\frac{1}{2}} - y(s))^2 = \\ &= ((x_{i-\frac{1}{2}} - x_{j-\frac{1}{2}} - x'_j(s - s_{j-\frac{1}{2}}))^2 + ((y_{i-\frac{1}{2}} - y_{j-\frac{1}{2}} - y'_j(s - s_{j-\frac{1}{2}}))^2 = \\ &= (s - s_{j-\frac{1}{2}} - c_{ij})^2 + d_{ij}^2, \text{ где} \end{aligned}$$



$$\begin{aligned} c_{ij} &= x'_j \left(x_{i-\frac{1}{2}} - x_{j-\frac{1}{2}} \right) + y'_j \left(y_{i-\frac{1}{2}} - y_{j-\frac{1}{2}} \right), \quad x'_j = \frac{\Delta x_j}{l_j}, \\ d_{ij} &= y'_j \left(x_{i-\frac{1}{2}} - x_{j-\frac{1}{2}} \right) - x'_j \left(y_{i-\frac{1}{2}} - y_{j-\frac{1}{2}} \right), \quad y'_j = \frac{\Delta y_j}{l_j}. \end{aligned} \quad (10.222)$$

$\cos(\widehat{x, n_j}) = y'_j$, $\cos(\widehat{y, n_j}) = x'_j$, откуда

$$\begin{aligned} b_{ij} &= \int_{s_{j-1}}^{s_j} q^* ds = \int_{s_{j-1}}^{s_j} \left[\cos(\widehat{x, n_j}) \frac{x_{i-\frac{1}{2}} - x(s)}{\rho^2(\xi_{i-\frac{1}{2}}, \xi)} + \cos(\widehat{y, n_j}) \frac{y_{i-\frac{1}{2}} - y(s)}{\rho^2(\xi_{i-\frac{1}{2}}, \xi)} \right] ds = \\ &= \int_{s_{j-1}}^{s_j} \frac{y'_j [x_{i-\frac{1}{2}} - x_{j-\frac{1}{2}} - x'_j(s - s_{j-\frac{1}{2}})] - x'_j [y_{i-\frac{1}{2}} - y_{j-\frac{1}{2}} - y'_j(s - s_{j-\frac{1}{2}})]}{(s - s_{j-\frac{1}{2}} - c_{ij})^2 + d_{ij}^2} ds = \\ &= d_{ij} \int_{s_{j-1}}^{s_j} \frac{ds}{(s - s_{j-\frac{1}{2}} - c_{ij})^2 + d_{ij}^2} = \begin{cases} \operatorname{arctg} \frac{c_{ij} + \frac{1}{2}l_j}{d_{ij}} - \operatorname{arctg} \frac{c_{ij} - \frac{1}{2}l_j}{d_{ij}}, & \text{при } d_{ij} \neq 0, \\ 0, & \text{при } d_{ij} = 0. \end{cases} \end{aligned}$$

$$\begin{aligned} a_{ij} &= \int_{s_{j-1}}^{s_j} u^* ds = -\frac{1}{2} \int_{s_{j-1}}^{s_j} \ln \rho^2(\xi_{i-\frac{1}{2}}, \xi) ds = -\frac{1}{2} \int_{s_{j-1}}^{s_j} \ln [(s - s_{j-\frac{1}{2}} - c_{ij})^2 + d_{ij}^2] ds = \\ &= -d_{ij} b_{ij} + l_j + \frac{1}{2} (c_{ij} - \frac{1}{2}l_j) \ln [(c_{ij} - \frac{1}{2}l_j)^2 + d_{ij}^2] - \\ &\quad - \frac{1}{2} (c_{ij} + \frac{1}{2}l_j) \ln [(c_{ij} + \frac{1}{2}l_j)^2 + d_{ij}^2]. \quad (10.223) \end{aligned}$$

По найденным граничным значениям $q_{i-\frac{1}{2}}$, $i = \overline{0, N_1}$ и $u_{i-\frac{1}{2}}$, $i = \overline{N_1 + 1, N_2}$ с помощью явной формулы (10.215) можно вычислить значение $u(\xi^0)$ в любой внутренней точке $\xi^0 = (x^0, y^0) \notin \tilde{\Gamma}$:

$$u(\xi^0) = \frac{-1}{2\pi} \left[\sum_{j=1}^{N_1} \left(b_j \bar{u}_{j-\frac{1}{2}} - a_j q_{j-\frac{1}{2}} \right) \sum_{j=N_1+1}^{N_2} \left(b_j u_{j-\frac{1}{2}} - a_j \bar{q}_{j-\frac{1}{2}} \right) \right], \quad (10.224)$$



где

$$a_{ij} = \int_{s_{j-1}}^{s_j} u^*(\xi^0, \xi) ds = -\frac{1}{2} \int_{s_{j-1}}^{s_j} \ln [(s - s_{j-\frac{1}{2}} - c_j)^2 + d_j^2] ds,$$

$$b_{ij} = \int_{s_{j-1}}^{s_j} q^*(\xi^0, \xi) ds = d_j \int_{s_{j-1}}^{s_j} \frac{ds}{(s - s_{j-\frac{1}{2}} - c_j)^2 + d_j^2},$$

$$\begin{aligned} c_j &= x'_j(x^0 - x_{j-\frac{1}{2}}) + y'_j(y^0 - y_{j-\frac{1}{2}}), \\ d_j &= y'_j(x^0 - x_{j-\frac{1}{2}}) - x'_j(y^0 - y_{j-\frac{1}{2}}). \end{aligned}$$

Достоинства метода:

- 1) понижает на единицу размерность задачи;
- 2) обладает высокой точностью.

Недостатки метода:

- 1) для его реализации обычно применяется метод исключения Гаусса, поэтому число узлов на границе не может быть большим, обычно не более 200;
- 2) пригоден только к уравнениям с известным фундаментальным решением.



Меню



Вверх

Назад

Вперёд

Пред.

След.

Указатель

Помощь

Экран

Предметный указатель

Другие А Б В Г Д Ж З И К Л М Н О П Р С Т У Ф Х Ч Ш Э
Я



Другие

LU-разложение матрицы

QR-разложение матрицы

span

æ

$\delta(u, v)$

\underline{a}_i



A

А-устойчивость

Аддитивный способ ослабления осо-
бенностей

Аппроксимация
дифференциального оператора
разностным



Б

Базисные пирамидальные функции

Барицентрические (симплексные) координаты



B

[Внутренние узлы](#)



Г

Меню



Г

Границочные

узлы

элементы



Д

Меню



Д

Диагональное

преобладание

строгое

Дивергентный вид главного члена

разложения



ЖК

Меню



Ж

[Жесткая система ОДУ](#)



3

Задача

- [границная](#)
- [корректная](#)
- [многоточечная](#)
- [некорректная](#)



И

[Интервал устойчивости](#)

[Интерполяция](#)

Эрмита

простое

[Итерация](#)

порядка k



К

[Квадратурная сумма](#)[Квадратурная формула](#)[Гаусса–Лагерра](#)[Гаусса–Лежандра](#)[Гаусса–Чебышева](#)[Гаусса–Эрмита](#)[Гаусса–Якоби](#)[Ньютона–Котеса](#)[Симпсона](#)[составная](#)[Филона](#)[Чебышева](#)[интерполяционная](#)[левых прямоугольников](#)[составная](#)[правых прямоугольников](#)[составная](#)[составная](#)[средних прямоугольников](#)[составная](#)[типа Гаусса](#)[трапеций](#)[составная](#)[Коэффициенты квадратурной суммы](#)[Кратность собственного значения](#)[алгебраическая](#)[геометрическая](#)[Круг Гершгорина](#)[Кубатурная формула](#)[Симпсона](#)[интерполяционная](#)[средних](#)[составная](#)[трапеций](#)



Меню

Л



Л

Лемма

тождество Кристоффеля-Дарбу

Линейная аппроксимация

Линия склейки



М

Мантисса

Матрица

Грамма

вращения

произвольной размерности
размерности 2

диагонализируемая

нормальная

отражения

положительно определенная

разреженная

Машинная арифметика с плавающей
точкой

Машинное число

Машинный эпсилон

Метод

Адамса

интерполяционный

экстраполяционный

Галеркина

для граничных задач

для интегральных уравнений

Гаусса

Гаусса–Зейделя

для СЛАУ

Данилевского

Зейделя

для разностных задач

для систем

Ньютона

дискретный

для систем

с постоянной матрицей Якоби
с постоянной производной

Пикара

Рунге–Кутта

диагонально–неявный

неявный

явный

Рунге–Кутта–Фельберга

Стеффенсена

ФДН

Чебышева

Эйлера

неявный

явный

Якоби

бисекции (дихотомии, половинного деления)

верхней релаксации

вращений

градиентного спуска

для нелинейных систем

граничных элементов

замены переменных

замены ядра на вырожденное

квадратного корня

коллокации

для граничных задач

для интегральных уравнений

конечных элементов

левой прогонки

линеаризации

линейный многошаговый

механических квадратур

для уравнения Вольтерра

для уравнения Фредгольма

многошаговый

явный

моментов

для граничных задач

наименьших квадратов

для интегральных уравнений

одношаговый

отражений

парабол

покоординатного спуска

последовательных приближения

для уравнения Вольтерра

для уравнения Фредгольма

правой прогонки

прогонки

продолжения по параметру

простой итерации

для СЛАУ

для нелинейных уравнений

для разностных задач



- для систем численных уравнений
- регуляризации
- релаксации
- для разностных задач рядов
- секущих
- для систем сопряженных градиентов спуска (общего вида)
- степенной метод
- трапеций
- неявный
- явный
- хорд
- энергетических неравенств
- Минимизирующая последовательность
- Многочлен
 - Чебышева–Эрмита
 - Якоби
 - Лежандра
 - Чебышева
 - Чебышева–Лагерра
 - алгебраический
 - интерполяционный
 - Ньютона–Бесселя для середины таблицы
 - Ньютона–Стирлинга для середины таблицы
 - Эрмита
 - в форме Лагранжа
 - в форме Ньютона
- в форме Ньютона для конца таблицы
- в форме Ньютона для начала таблицы
- для функции двух переменных в форме Ньютона
- наилучшего равномерного приближения
- наилучшего среднеквадратичного приближения
- обобщённый
- Мультипликативный способ выделения особенностей



Н

[Направление спуска](#)

[Неравенство](#)

ε -неравенство
Коши–Буняковского
треугольника

[Норма](#)

векторная

p -норма

евклидова

максимум-норма

матричная

Фробениуса

индукционная

подчиненная

спектральная

строчная максимум-норма

[Носитель функции](#)



O

[Область устойчивости](#)

[Оператор](#)

[положительно определенный](#)
[простой структуры](#)
[регуляризирующий](#)
[самосопряженный](#)

[Определитель Вандермонда](#)

[Ортогонализация Грамма–Шмидта](#)

[Остаток](#)

[интерполяции](#)
[квадратурной формулы](#)

[Остаточный член](#)

[в форме Лагранжа](#)
[в форме Ньютона](#)

[Отделение корней](#)



П

Пирамидальная функция

Плохо обусловленная задача

Погрешность

аппроксимации

оператора в точке

оператора на сетке

глобальная

локальная

метода

начального условия

округления

абсолютная

относительная

при решении задачи Коши

приближенного решения

задачи Коши

Полная в классе система

Порядок

аппроксимации

оператора в точке

оператора на сетке

разностной схемы

точности метода

точности разностной схемы

Правило

Рунге

для численного интегрирования

для численного решения задачи Коши

округления

Правильная таблица конечных разностей

Приближение

наилучшее

Принцип максимума

Проблема собственных значений

Проекция

Пространство

гильбертово

линейное нормированное

строго нормированное

Процесс

интерполяционный

Прямая формулировка метода Галёркина



P

Разностная

производная

вторая

левая

правая

центральная

четвертая

схема

для нестационарного уравнения теплопроводности

консервативная

монотонная

наилучшая консервативная
однородная

однородная

формула Грина

вторая

первая

Разность

конечная

разделенная



С

Сетка

- квадратная
- прямоугольная
- неравномерная
- равномерная

Сеточная функция

Сжимающее отображение

Система функций Чебышева

Скорость сходимости разностной схемы

Слабая формулировка метода Галёркина

Собственное значение

Собственное подпространство

Собственный вектор

Спектр матрицы

Спектральный радиус

Сплайн

- интерполяционный
- кубический
- параметрический
- первой степени
- полиномиальный
- полиномиальный бикубический
- сглаживающий

Стандарт IEEE 754

Степень точности квадратурной формулы

алгебраическая

Сходимость

- интерполяционного процесса
- разностной схемы



Т

Таблица Бутчера

Теорема

Бауэра–Файка

Валле–Пуссена

Вейерштрасса

Гершгорина

Достаточное условие сходимости двухслойных итерационных схем

Лакса

Мартинкевич

Принцип максимума

Ролля

Флобер

Чебышева

достаточное условие сходимости итерационного процесса общего вида

критерий квадратурных формул
НАСТ

критерий сходимости итерационного процесса общего вида

о QR -разложении

о сходимости метода итерации

о чебышевском альтернансе

признак устойчивости

принцип сжимающих отображений

разложение Холецкого

связь метода Гаусса и LU -

разложения

теорема сравнения

Тихоновский стабилизатор

Точка

чебышевского альтернанса

Триангulationя области



У

Узел

квадратурной суммы

сетки

Уравнение

интегральное

Вольтерра второго рода

Вольтерра первого рода

Фредгольма второго рода

Фредгольма первого рода

разностное

Условие

корней

Куранта

согласованности норм

Устойчивость

метода решения задачи Коши

разностной схемы

по входным данным



Ф

Меню



Ф

[Форма Фробениуса](#)

Формат

[CSC](#)

[CSR](#)

[MSR](#)

координатный

Фундаментальная последовательность

Фундаментальное решение уравнения Лапласа

Функция

абсолютно непрерывная

влияния узла

интерполирующая

с конечным носителем



X

Меню



Вверх

Назад

Вперёд

Пред.

След.

Указатель

Помощь

Экран

X

Характеристический многочлен

Характеристическое уравнение



Ч

Число

обусловленности

задачи

матрицы

с плавающей точкой

денормализованное

нормализованное



Ш

[Шаблон](#)

нерегулярный трехточечный

[Шаг сетки](#)



Э

Меню



Э

Экспонента числа с плавающей точкой

Элемент наилучшего приближения

Элементарная функция

Энергетическое тождество



Меню

Я



Я

Ядро интегрального оператора
вырожденное



Определения

LU-разложение матрицы

QR-разложение матрицы

Строго нормированное пространство

А-устойчивость

Абсолютная и относительная погрешности
округления

Адаптивная сетка

Алгебраическая степень точности квадратур-
ной формулы

Базисные пирамидальные функции

Внутренние узлы

Вырожденное ядро интегрального оператора

Градиент

Границные узлы

Двухслойная итерационная схема

Денормализованное число с плавающей точкой

Диагонализируемая матрица

Диагональное преобладание

Жесткая система ОДУ

Интерполяционная квадратурная формула

Консервативная разностная схема

Корневое условие

Корректно поставленная задача

Круг Гершгорина

Линейная аппроксимация

Локальная погрешность метода

Машинная арифметика с плавающей точкой

Машинные числа

Машинный эпсилон

Метод верхней релаксации

Многочлен наилучшего равномерного прибли-
жения

Наилучшее приближение

Норма Фробениуса

Норма линейного оператора

Нормализованное число с плавающей точкой

Нормальная матрица

Носитель функции

Область устойчивости метода решения задачи

Коши

Область устойчивости метода решения задачи

Коши

Пирамидальная функция



Погрешность аппроксимации
 Полиномиальный бикубический сплайн
 Полиномиальный сплайн
 Полная в классе система
 Положительно определенная матрица
 Положительно определенный оператор
 Правило округления
 Разреженная матрица
 Регуляризирующий оператор
 Самосопряженный оператор
 Сетка
 Сжимающее отображение

Система функций Чебышева
 Собственные значения и собственные векторы
 Спектральный радиус
 Степень точности квадратурной формулы
 Триангуляция области
 Устойчивость метода решения задачи Коши
 Фундаментальная последовательность
 Фундаментальное решение уравнения Лапласа
 Число обусловленности задачи
 Число обусловленности матрицы
 Число с плавающей точкой
 Элемент наилучшего приближения



LU-разложение матрицы

LU-разложением невырожденной матрицы A называется её представление в виде

$$A = LU,$$

где L — нижнетреугольная матрица с единицами на главной диагонали, U — верхнетреугольная матрица.

[\[Перейти к основному тексту\]](#)



QR-разложение матрицы

QR-разложением матрицы A называется её представление в виде

$$A = QR,$$

где Q — ортогональная ($Q^{-1} = Q^T$), а R — верхнетреугольная матрица.

[\[Перейти к основному тексту\]](#)



Пусть $w \in \mathbb{R}^n(\mathbb{C}^n)$, $\|w\|_2 = 1$. Матрица

$$H = H(w) = I - 2ww^T$$

называется *матрицей отражения*. Она задаёт преобразование отражения относительно гиперплоскости с нормалью w .

[\[Перейти к основному тексту\]](#)



Строго нормированное пространство

Нормированное пространство R называется *строго нормированным*, если в нем равенство $\|f + g\| = \|f\| + \|g\|$ возможно только при условии $f = \lambda g$, $\lambda > 0$. [\[Перейти к основному тексту\]](#)



A-устойчивость

Численный метод будем называть *A-устойчивым*, если его *область устойчивости* содержит всю левую полуплоскость $\operatorname{Re} z < 0$.

[[Перейти к основному тексту](#)]



Абсолютная и относительная погрешности округления

Абсолютной погрешностью округления для числа $x \in \mathbb{R}$ в данной МАПТ называется число

$$\Delta(x) = |x - R(x)|,$$

а *относительной погрешностью округления* — число

$$\delta(x) = \frac{|x - R(x)|}{|x|} = \frac{\Delta(x)}{|x|}.$$

[\[Перейти к основному тексту\]](#)



Адаптивная сетка

Адаптивная сетка — [сетка](#), у которой плотность узлов на участках сильного изменения решения более высокая, чем на участках слабого изменения решения.

[[Перейти к основному тексту](#)]



Алгебраическая степень точности квадратурной формулы

Говорят, что [квадратурная формула](#) имеет *алгебраическую степень точности, равную m* , если она точна для всевозможных многочленов степени m и существует хотя бы один многочлен степени $(m+1)$, для которого формула точной не является.

[\[Перейти к основному тексту\]](#)



Базисные пирамидальные функции

Базисными пирамидальными функциями, заданными на треугольных элементах [сетки](#) $\bar{\omega}_h$, называют систему линейно независимых функций $\varphi_1(\xi), \dots, \varphi_N(\xi)$, каждая из которых отлична от нуля лишь в окрестности одного из внутренних узлов $\xi_1, \dots, \xi_N \in \omega_h$.

[[Перейти к основному тексту](#)]



Внутренние узлы

Точки пересечения прямых прямоугольной сетки

$$x = x_i = (x_{1,i_1}, x_{2,i_2}),$$

лежащие внутри области \bar{G} , т. е. $x_i \in G$, назовём *внутренними узлами*. Множество внутренних узлов обозначим $\hat{\omega}_h \subset G$.

[\[Перейти к основному тексту\]](#)



Вырожденное ядро интегрального оператора

Ядро $K(x, s)$ называется *вырожденным*, если оно может быть представлено в виде

$$K(x, s) = \sum_{i=0}^n \alpha_i(x) \beta_i(s).$$

[\[Перейти к основному тексту\]](#)



Градиент

Напомним, что *градиентом* функции n переменных $f : \mathbb{R}^n \rightarrow \mathbb{R}$ называется вектор-функция

$$\operatorname{grad} f = \nabla f : \mathbb{R}^n \rightarrow \mathbb{R}^n,$$

определенная формулой

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n} \right).$$

[\[Перейти к основному тексту\]](#)



Границные узлы

Точки пересечения этих прямых с границей Γ назовём *границными узлами*. Это множество обозначим $\hat{\gamma}_h \subset \Gamma$.

[\[Перейти к основному тексту\]](#)



Двухслойная итерационная схема

Двухслойной итерационной схемой будем называть всякий линейный одношаговый итерационный метод решения задачи (10.180), представимый в виде

$$\begin{cases} B \frac{y^{n+1} - y^n}{\tau} + Ay^n = \varphi, & x \in \omega_h, \\ y^{n+1} = \mu(x), & x \in \gamma_h, \quad n = 0, 1, 2, \dots, \end{cases}$$

где $y^0 \in H_h$; $\tau > 0$ — параметр релаксации, B — линейный оператор (матрица).

[\[Перейти к основному тексту\]](#)



Денормализованное число с плавающей точкой

Вещественные числа вида

$$0.d_1 d_2 \dots d_{p-1} \times \beta^{e_{\min}},$$

где d_i — произвольные β -ичные цифры, называются *денормализованными числами с плавающей точкой* (ДЧПТ). Множество всех ДЧПТ с параметрами β , p , e_{\min} будем обозначать $\mathbb{F}_0(\beta, p, e_{\min})$ либо кратко \mathbb{F}_0 .

[\[Перейти к основному тексту\]](#)



Диагонализируемая матрица

Если квадратная матрица размерности n имеет n линейно независимых собственных векторов, то она называется *диагонализируемой*, а соответствующий ей линейный оператор — *оператором простой структуры*.

[[Перейти к основному тексту](#)]



Диагональное преобладание

Если элементы матрицы A удовлетворяют условиям

$$|a_{ii}| \geq \sum_{j \neq i} |a_{ij}| \quad \forall i = \overline{1, n},$$

то говорят, что такая матрица обладает свойством *диагонального преобладания*. Если неравенство в (2.18) строгое, говорят о *строгом диагональном преобладании*.

[\[Перейти к основному тексту\]](#)



Жесткая система ОДУ

Система обыкновенных дифференциальных уравнений (8.94) с постоянной ($n \times n$)-матрицей A называется **жесткой**, если:

- 1) $\operatorname{Re} \lambda_k < 0$, $k = \overline{1, n}$ (т.е. задача устойчива);
- 2) отношение $S = \frac{\max\limits_{1 \leq k \leq n} |\operatorname{Re} \lambda_k|}{\min\limits_{1 \leq k \leq n} |\operatorname{Re} \lambda_k|}$ велико (например, $S > 10$). Число S иногда называют коэффициентом жесткости системы (8.94).

[\[Перейти к основному тексту\]](#)



Интерполяционная квадратурная формула

Квадратурные формулы, коэффициенты которых вычисляются по формулам [\(6.6\)](#), называют [интерполяционными](#).
[Перейти к основному тексту]



Консервативная разностная схема

Разностные схемы, которые выражают законы сохранения на сетке, называют *консервативными* разностными схемами.

[\[Перейти к основному тексту\]](#)



Корневое условие

Будем говорить, что численный метод удовлетворяет *условию корней*, если все корни q_0, \dots, q_k характеристического уравнения (8.88) лежат внутри или на границе единичного круга комплексной плоскости, причем на границе круга нет кратных корней.

[\[Перейти к основному тексту\]](#)



Корректно поставленная задача

Задача называется *корректно поставленной*, или просто *корректной*, если её решение (а) существует, (б) единственно и (в) непрерывно зависит от начальных данных. Если нарушено хотя бы одно из этих условий, задачу называют *некорректной*.

[[Перейти к основному тексту](#)]



Круг Гершгорина

Кругом Гершгорина для квадратной матрицы A называется замкнутый круг D , на комплексной плоскости с центром в точке a_{ii} и радиусом

$$\rho_i = \sum_{j \neq i} |a_{ij}|.$$

[\[Перейти к основному тексту\]](#)



Линейная аппроксимация

Аппроксимационная задача называется *линейной*, если множество Φ линейно относительно параметров a_k (например, является линейным подпространством, натянутым на заданные базисные функции $\varphi_k(x)$, $k = \overline{0, n}$); в противном случае задача называется нелинейной. [\[Перейти к основному тексту\]](#)



Локальная погрешность метода

Невязку численного метода (8.8) на точном решении задачи (8.4)

$$r(t_j, \tau) = u(t_{j+1}) - F(u(t_{j-q}), \dots, u(t_j), u(t_{j+1}), \dots, u(t_{j+s}))$$

будем называть *локальной погрешностью* метода (8.8).

[\[Перейти к основному тексту\]](#)



Машинная арифметика с плавающей точкой

Машинной арифметикой с плавающей точкой (МАПТ) будем называть множество машинных чисел M в совокупности с правилом округления R .

[[Перейти к основному тексту](#)]



Машинные числа

Машинными числами будем называть элементы множества

$$M = F_0 \cup F_1.$$

[\[Перейти к основному тексту\]](#)



Машинный эпсилон

Машинным эпсилон ε_M для МАПТ называется наименьшее положительное число ε , удовлетворяющее условию

$$R(1 + \varepsilon) > 1.$$

[\[Перейти к основному тексту\]](#)



Метод верхней релаксации

В случае задачи Дирихле (10.183) для уравнения Пуассона в прямоугольнике теория численных методов даёт оптимальное значение параметра релаксации

$$\tau_{opt} = 2 \sqrt{1 - \left[1 - \frac{\pi^2}{2} \frac{h_1^2 h_2^2}{h_1^2 + h_2^2} \left(\frac{1}{l_1} + \frac{1}{l_2} \right) \right]^2},$$

при котором достигается максимальная скорость сходимости метода релаксации (10.187). Очевидно, что $1 < \tau_{opt} < 2$. Поэтому метод релаксации (10.187) при $\tau = \tau_{opt}$ называют *методом верхней релаксации*.

[\[Перейти к основному тексту\]](#)



Многочлен наилучшего равномерного приближения

Многочлен Q_n^0 такой, что

$$\|Q_n^0 - f\| = \inf_{Q \in \Pi_n} \|Q - f\|,$$

называется *многочленом наилучшего равномерного приближения* степени n для функции f . Здесь Π_n — множество всех многочленов степени не выше n .

[[Перейти к основному тексту](#)]



Наилучшее приближение

Величина

$$\Delta(f) = \inf_{\varphi \in \Phi_n} \|f - \varphi\|$$

называется *наилучшим приближением* элемента f на множестве Φ_n .

[\[Перейти к основному тексту\]](#)



Норма Фробениуса

Нормой Фробениуса называется матричная норма $\|\cdot\|_F$, определяемая как

$$\|A\|_F = \sqrt{\sum_{i,j}^n |a_{ij}|^2}.$$

[\[Перейти к основному тексту\]](#)



Норма линейного оператора

Нормой линейного оператора A называют число

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\|.$$

Норма оператора полностью определяется векторной нормой. То есть каждая векторная норма порождает (*индуцирует*) соответствующую ей операторную матричную форму (в этом случае говорят также, что матричная норма *подчинена* векторной). [\[Перейти к основному тексту\]](#)



Нормализованное число с плавающей точкой

Число с плавающей точкой с ненулевым первым разрядом ($d_0 \neq 0$) называется *нормализованным*. Множество всех нормализованных ЧПТ с основанием β , p -разрядной мантиссой, и $e_{\min} \leq e \leq e_{\max}$ условимся обозначать $\mathbb{F}_1(\beta, p, e_{\min}, e_{\max})$ или просто \mathbb{F}_1 .

[[Перейти к основному тексту](#)]



Нормальная матрица

Матрица A , которая может быть диагонализирована унитарным преобразованием подобия, называется [нормальной](#).

[[Перейти к основному тексту](#)]



Носитель функции

Носителем функции называют наименьшую замкнутую область, вне которой функция тождественно равна нулю.

[[Перейти к основному тексту](#)]



Область устойчивости метода решения задачи Коши

Областью устойчивости численного метода будем называть множество всех точек z комплексной плоскости, для которых данный метод [устойчив](#).

[[Перейти к основному тексту](#)]



Область устойчивости метода решения задачи Коши

Интервалом устойчивости численного метода будем называть пересечение [области устойчивости](#) с вещественной осью координат.

[[Перейти к основному тексту](#)]



Пирамидальная функция

Пирамидальной функцией $\varphi_i(x)$, соответствующей узлу $\xi_i = (x_i, y_i) \in \bar{\omega}_h$, будем называть кусочно-линейную функцию вида

$$\varphi_i(\xi) = \begin{cases} \varphi_i^{(p)}(\xi) = \frac{(x-x_{p+1})(y_p-y_{p+1})-(y-y_{p+1})(x_p-x_{p+1})}{2S_p}, \\ \quad \xi = (x, y) \in \Delta_p, \quad p = \overline{1, m}, \\ 0, \quad \xi \notin \Omega_i = \bigcup_{p=1}^m \Delta_p, \end{cases} \quad i = \overline{1, N + M}.$$

[Перейти к основному тексту]



Погрешность аппроксимации

Величину

$$\psi(t_j, \tau) = \frac{u(t_{j+1}) - u(t_j)}{\tau} - \frac{F(u(t_{j-q}), \dots, u(t_j), u(t_{j+1}), \dots, u(t_{j+s})) - u(t_j)}{\tau} \equiv \frac{r(t_j, \tau)}{\tau}$$

будем называть *погрешностью аппроксимации* дифференциальной задачи (8.4) разностной задачей (8.8).

Если при этом $\psi(t_j, \tau) = O(\tau^p)$, $p \geq 1$, то метод (8.8) называют *методом p-го порядка точности*.

[\[Перейти к основному тексту\]](#)



Полиномиальный бикубический сплайн

Назовем функцию $S_{\Delta}^{n,m}(x, y)$ *полиномиальным сплайном степени n по переменной x и степени m по переменной y с линиями склейки на сетке Δ* , если:

- На каждой ячейке Ω_{ij} $S_{\Delta}^{n,m}(x, y)$ является многочленом степени n по переменной x и степени m по переменной y , т.е.

$$S_{\Delta}^{n,m}(x, y) = \sum_{k=0}^n \sum_{l=0}^m a_{kl}^{ij} (x - x_i)^k (y - y_j)^l, \quad i = \overline{1, N}; \quad j = \overline{1, M};$$

- 2.

$$S_{\Delta}^{n,m}(x, y) \in C^{n-1, m-1}(\Omega).$$

[\[Перейти к основному тексту\]](#)



Полиномиальный сплайн

Разобьем отрезок $[a, b]$, на котором ищется приближение к функции $f(x)$, на N частей точками x_j : $a = x_0 < x_1 < \dots < x_{N-1} < x_N = b$.

По определению положим: $x_j - x_{j-1} = h_j > 0$, $j = 1, \dots, N$. Соответствующее разбиение далее будем обозначать Δ .

Назовем *полиномиальным сплайном порядка m дефекта k* на разбиении Δ (обозначение $S_{\Delta}^m(x)$) функцию, которая:

- На каждом из отрезков $[x_{i-1}; x_i]$, $i = \overline{1, N}$ является алгебраическим многочленом степени m , т.е.

$$S_{\Delta}^m(x) = P_{im}(x) = a_{i0} + a_{i1}x + \dots + a_{im}x^m, \quad x \in [x_{i-1}; x_i], \quad i = \overline{1, N};$$

- Является функцией класса $C^{m-k}[a, b]$, т.е. во всех внутренних узлах разбиения Δ $S_{\Delta}^m(x)$ удовлетворяет условию непрерывности производных до порядка $m - k$ включительно:

$$(S_{\Delta}^m(x_i + 0))^{(j)} = (S_{\Delta}^m(x_i - 0))^{(j)}, \quad i = \overline{1, N-1}; \quad j = \overline{0, m-k}.$$

[\[Перейти к основному тексту\]](#)



Полная в классе система

Систему функций $\{\varphi_i(x)\}$ будем называть *полной в классе F функций $f(x)$* , если для любой функции $f(x) \in F$ и любого $\varepsilon > 0$ существует натуральное число N такое, что при любом $n > N$ найдется набор параметров a_0, a_1, \dots, a_n – коэффициентов обобщенного многочлена степени $Q_n(x)$ по системе $\{\varphi_i(x)\}$ такой, что при всех $x \in [a, b]$ выполняется неравенство $|f(x) - Q_n(x)| < \varepsilon$. [\[Перейти к основному тексту\]](#)



Положительно определенная матрица

Квадратная матрица A над полем \mathbb{R} (\mathbb{C}) называется *положительно определённой* ($A > 0$), если

$$(Ax, x) > 0 \quad \forall x \in \mathbb{R}^n (\mathbb{C}^n), x \neq 0.$$

В комплексном случае мы подразумеваем, что все (Ax, x) вещественны.

[\[Перейти к основному тексту\]](#)



Положительно определенный оператор

Оператор A называется *положительно определённым* в $\overset{\circ}{H}_h$ ($A > 0$), если $\forall u \in \overset{\circ}{H}_h$ имеем $(Au, u) \geqslant 0$, причём $(Au, u) = 0 \Leftrightarrow u = 0$.
[Перейти к основному тексту]



Правило округления

Правилом округления для данного множества [чисел с плавающей точкой](#) $\mathbb{F} \subset \mathbb{R}$ будем называть отображение

$$R : \mathbb{R} \rightarrow \mathbb{F}$$

такое, что $R(x) = x$, если $x \in \mathbb{F}$, и $R(x) \approx x$ в противном случае.

[\[Перейти к основному тексту\]](#)



Разреженная матрица

Разреженными называют матрицы, содержащие большой процент нулевых элементов.

[\[Перейти к основному тексту\]](#)



Регуляризирующий оператор

Оператор A_α называют *регуляризирующим*, если:

- 1) задача (7.67) является корректно поставленной в классе правых частей F при любом $\alpha > 0$;
- 2) существуют такие функции $\alpha(\delta)$ и $\delta(\varepsilon)$, что если $\|f - \tilde{f}\|_F \leq \delta(\varepsilon)$, то $\|u_{\alpha(\delta)} - \bar{u}\|_U \leq \varepsilon$.

[Перейти к основному тексту](#)



Самосопряженный оператор

Оператор A называется *самосопряжённым* в $\overset{\circ}{H}_h$ ($A = A^*$), если $\forall u, v \in \overset{\circ}{H}_h$ выполняется $(Au, v) = (u, Av)$.

[\[Перейти к основному тексту\]](#)



Сетка

Множество [внутренних](#) и [границых узлов](#) назовём *сеткой* $\hat{\omega}_h = \hat{\omega}_h \cup \hat{\gamma}_h$ в области $\bar{G} = G \cup \Gamma$.

[\[Перейти к основному тексту\]](#)



Сжимающее отображение

Рассмотрим отображение $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$. Если существует константа $0 \leq \alpha < 1$, такая, что для любых $x, y \in \mathbb{R}^n$

$$\rho(f(x), f(y)) \leq \alpha \rho(x, y),$$

то отображение f называют [сжимающим](#).

[Перейти к основному тексту](#)



Система функций Чебышева

Система функций $\{\varphi_i(x)\}$, обобщенный многочлен степени n по которой имеет на отрезке $[a, b]$ не более n различных корней, называется *системой функций Чебышева*. [\[Перейти к основному тексту\]](#)



Собственные значения и собственные векторы

Пусть A — квадратная матрица над полем \mathbb{C} . Вектор $x \neq 0 \in \mathbb{C}^n$ называется *собственным вектором* матрицы A , если существует $\lambda \in \mathbb{C}$ такое, что

$$Ax = \lambda x.$$

Число λ называется *собственным значением*, соответствующим x . Множество всех собственных значений A называется *спектром* и обозначается $\sigma(A)$.

[\[Перейти к основному тексту\]](#)



Спектральный радиус

Спектральным радиусом $\rho(A)$ называется величина наибольшего по модулю собственного значения матрицы A .

[\[Перейти к основному тексту\]](#)



Степень точности квадратурной формулы

Говорят, что [квадратурная формула](#) имеет *степень точности m относительно системы функций $\{\varphi_i(x)\}$* , если она точна на первых m функциях $\varphi_0(x), \dots, \varphi_m(x)$ и не точна на функции $\varphi_{m+1}(x)$, т.е. выполняются соотношения

$$\begin{cases} \int_a^b p(x) \varphi_i(x) dx = \sum_{k=0}^n A_k \varphi_i(x_k) , & i = \overline{0, m}, \\ \int_a^b p(x) \varphi_{m+1}(x) dx \neq \sum_{k=0}^n A_k \varphi_{m+1}(x_k) . \end{cases}$$

[\[Перейти к основному тексту\]](#)



Триангуляция области

Узлы сетки, произвольно выбранные в области и на границе Γ , соединяются не пересекающимися отрезками так, чтобы каждый внутренний узел был вершиной 6 треугольников (элементов). Такое построение сетки называется *триангуляцией области* \bar{G} .

[\[Перейти к основному тексту\]](#)



Устойчивость метода решения задачи Коши

Численный метод решения задачи Коши будем называть *устойчивым* при некотором значении z , если при данном значении z устойчиво соответствующее ему разностное уравнение (8.90), получающееся вследствие применения исследуемого метода к решению модельного уравнения (8.89). [Перейти к основному тексту]



Фундаментальная последовательность

Последовательность точек (x^k) , $x^k \in \mathbb{R}^n$, называется *фундаментальной*, или *последовательностью Коши*, если

$$\rho(x^k, x^m) \rightarrow 0 \quad \text{при} \quad k, m \rightarrow \infty.$$

[\[Перейти к основному тексту\]](#)



Фундаментальное решение уравнения Лапласа

Функция $u^*(\xi^0, \xi)$ называется *фундаментальным решением уравнения Лапласа* $\nabla^2 u = 0$, если

$$\nabla^2 u^* = -\pi\alpha\delta(\xi^0, \xi),$$

где δ — дельта-функция Дирака, т. е.

1)

$$\nabla^2 u^*(\xi^0, \xi) = \begin{cases} 0, & \xi \neq \xi^0, \\ +\infty, & \xi = \xi^0. \end{cases}$$

2)

$$\iint_{\Omega} g(\xi) \nabla^2 u^*(\xi^0, \xi) dx dy = -\pi\alpha g(\xi^0)$$

для любой фиксированной точки ξ^0 , любой ее окрестности $\Omega \subseteq \bar{G}$, любой функции $g(\xi) = g(x, y)$, непрерывной в точке ξ^0 , где

$$\alpha = \begin{cases} 1, & \xi^0 \in \Gamma, \\ 2, & \xi^0 \notin \Gamma, \text{ т. е. } \xi^0 \in G. \end{cases}$$

[\[Перейти к основному тексту\]](#)



Число обусловленности задачи

Числом обусловленности задачи вычисления $f(x)$ назовём число

$$\alpha(x) = \sup_{\tilde{x} \in M} \frac{\delta(y, \tilde{y})}{\delta(x, \tilde{x})} = \sup_{\tilde{x} \in M} \left(\frac{\|f(x) - f(\tilde{x})\|}{\|f(x)\|} \cdot \frac{\|x\|}{\|x - \tilde{x}\|} \right),$$

где $M \subset X$ — некоторая проколотая окрестность точки x . Если $\alpha(x)$ велико, задачу называют *плохо обусловленной*.

[Перейти к основному тексту]



Число обусловленности матрицы

Числом обусловленности невырожденной матрицы A называется число

$$\alpha(A) = \|A\| \|A^{-1}\|.$$

Если матрица A вырождена, её число обусловленности полагается равным бесконечности.

[\[Перейти к основному тексту\]](#)



Число с плавающей точкой

Пусть $\beta \in \mathbb{N}$ — основание системы счисления, $p \in \mathbb{N}$ — число значащих разрядов, d_i — цифры. Вещественное число вида

$$\pm \underbrace{d_0.d_1d_2 \dots d_{p-1}}_m \times \beta^e, \quad 0 \leq d_i < \beta,$$

называется *числом с плавающей точкой* (ЧПТ). Число $m \in \mathbb{R}$ называют *мантиссой* или значащей частью. Число $e \in \mathbb{Z}$ называют *показателем*, или *экспонентой* (не путать с числом e). [Перейти к основному тексту]



Элемент наилучшего приближения

Элемент $\varphi_0 \in \Phi_n$, такой, что

$$\|f - \varphi_0\| = \Delta(f),$$

называется *элементом наилучшего приближения* для f на Φ_n , или *проекцией* f на Φ_n .

[\[Перейти к основному тексту\]](#)



Доказательства теорем



Теорема 2.1

Воспользуемся методом математической индукции. Преобразование G_1 корректно определено и первый шаг метода Гаусса выполним если (и только если) $a_{11}^{(1)} = a_{11} = |[A]_1| \neq 0$.

Пусть выполнимо k шагов. Это означает, что существует матрица \tilde{G}_k ,

$$\tilde{G}_k = G_k G_{k-1} \dots G_1, \quad \text{и} \quad A^{(k+1)} = \tilde{G}_k A.$$

Нетрудно увидеть, что матрицы \tilde{G}_k имеют блочный вид

$$\left[\begin{array}{c|c} L_k & 0 \\ \hline \boxtimes & I \end{array} \right],$$

где L_k — нижнетреугольная матрица размерности k с единицами на главной диагонали, I — единичная матрица размерности $n - k$.

Запишем равенство $\tilde{G}_k A = A^{(k+1)}$ в блочном виде:

$$\underbrace{\left[\begin{array}{c|c} L_k & 0 \\ \hline \boxtimes & I \end{array} \right]}_{\tilde{G}_k} A = \underbrace{\left[\begin{array}{c|c} U_k & \boxtimes \\ \hline 0 & \boxtimes \end{array} \right]}_{A^{(k+1)}}, \tag{Д.1}$$

где U_k — верхнетреугольная матрица размерности k .

Критерием осуществимости $(k + 1)$ -го шага является условие

$$\theta_{k+1} = a_{k+1,k+1}^{(k+1)} \neq 0,$$

которое гарантирует существование G_{k+1} (см. (2.12), (2.13)).

Из (Д.1) имеем

$$[\tilde{G}_k]_{k+1}[A]_{k+1} = [A^{(k+1)}]_{k+1}.$$



Так как $|\tilde{G}_k|_{k+1} \neq 0$ и θ_{k+1} — единственный ненулевой элемент в последней строке матрицы $[A^{(k+1)}]_{k+1}$, то

$$\theta_{k+1} \neq 0 \Leftrightarrow |[A]_{k+1}| \neq 0.$$

[\[Перейти к основному тексту\]](#)



Теорема 2.2 (связь метода Гаусса и LU-разложения)

⇒ Пусть осуществим базовый алгоритм. Тогда из формулы (Д.1) при $k = n - 1$ получаем $\tilde{G}_{n-1}A = A^{(n)}$, причём по построению \tilde{G}_{n-1} — нижнетреугольная с единичной главной диагональю, а $A^{(n)}$ — верхнетреугольная. Отсюда получаем $A = LU$, где $L = (\tilde{G}_{n-1})^{-1}$, $U = A^{(n)}$.

⇐ Необходимость следует из теоремы 2.1.

[Перейти к основному тексту]



Теорема 2.4

По условию имеем

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

Отсюда $|[A]_1| = a_{11} \neq 0$. Далее рассмотрим $[A]_k$. Эта матрица, очевидно, тоже обладает строгим диагональным преобладанием.

Следовательно, радиус [круга Гершгорина](#) D_i для $[A]_k$ меньше, чем $|a_{ii}|$, поэтому $0 \notin D_i \quad \forall i = \overline{1, k}$.

Значит, по [теореме Гершгорина](#) все собственные значения $[A]_k$ отличны от нуля, и $|[A]_k| \neq 0$.

[\[Перейти к основному тексту\]](#)



Вверх

Назад

Вперёд

Пред.

След.

Указатель

Помощь

Экран

Теорема 2.7 (о QR-разложении)

Доказательство напрямую следует из приведённых выше рассуждений. Если A вырождена, то в ходе метода отражений будем встречать нулевые векторы a_1 . В этом случае нужно просто перейти к следующему шагу.

[\[Перейти к основному тексту\]](#)



Лемма 3.1

Пусть $\|A\| < 1$. Тогда $\rho(A) \leq \|A\| < 1$ и все собственные значения матрицы $I + A$ положительны, то есть $\det(I + A) \neq 0$.

[\[Перейти к основному тексту\]](#)



Теорема 4.2 (о сходимости метода итерации)

Проверим сначала возможность построения последовательности приближений, используя метод математической индукции.

Так как x_0 попадает в исходную область, то $f(x)$ определено. Кроме того, по условию

$$|x_1 - x_0| = |\varphi(x_0) - x_0| \leq m \leq \frac{m}{1-q} \leq \delta.$$

Следовательно, $x_1 \in \Delta$ и $x_2 = \varphi(x_1)$ определено. Поскольку

$$|x_2 - x_1| = |\varphi(x_1) - \varphi(x_0)| \leq q|x_1 - x_0| \leq qm,$$

то, используя неравенство треугольника, получим:

$$|x_2 - x_0| = |x_2 - x_1 + x_1 - x_0| \leq |x_2 - x_1| + |x_1 - x_0| \leq qm + m = m(1 + q) < m(1 + q + q^2 + \dots) = \frac{m}{1-q} \leq \delta.$$

Следовательно, мы получили:

- 1) $x_2 \in \Delta$;
- 2) $|x_2 - x_1| \leq qm$.

Пусть при любом $n = \overline{1, k}$ полученные свойства также имеют место, т.е. $|x_n - x_0| \leq \delta$ и $|x_n - x_{n-1}| \leq q^{n-1}m$ при S_0 . Покажем, что эти соотношения выполняются и при $n = k + 1$. Имеем:

$$|x_{k+1} - x_k| = |\varphi(x_k) - \varphi(x_{k-1})| \leq q|x_k - x_{k-1}| \leq q^k m;$$

$$\begin{aligned} |x_{k+1} - x_0| &\leq |x_{k+1} - x_k| + |x_k - x_{k-1}| + \dots + |x_1 - x_0| \leq q^k m + q^{k-1} m + \dots + qm + m < \\ &< m(1 + q + q^2 + \dots) = \frac{m}{1-q} \leq \delta. \end{aligned}$$



Следовательно, $x_{k+1} \in \Delta$, а значит, исковую последовательность можно построить. Покажем, что эта последовательность сходится. Пользуясь критерием Больцано-Коши, достаточно убедиться в фундаментальности построенной последовательности. Имеем:

$$|x_{n+p} - x_n| \leq |x_{n+p} - x_{n+p-1}| + \dots + |x_{n+1} - x_n| \leq q^{n+p-1}m + \dots + q^n m < mq^n(1 + q + q^2 + \dots) = \frac{mq^n}{1-q} \leq \delta q^n;$$

отсюда, поскольку $q < 1$, следует, что указанная разность может быть сделана сколь угодно малой при достаточно большом n и правая часть неравенства не зависит от p . Таким образом, последовательность $\{x_n\}$ является фундаментальной и, следовательно, сходится, т.е. существует

$$x^* = \lim_{n \rightarrow \infty} x_n.$$

Убедимся, что x^* является решением уравнения $x = \varphi(x)$. Для этого перейдем к пределу в равенстве (4.4). Так как $\varphi(x)$ непрерывна, то переход к пределу под знаком функции дает:

$$x^* = \lim_{n \rightarrow \infty} x_{n+1} = \lim_{n \rightarrow \infty} \varphi(x_n) = \varphi\left(\lim_{n \rightarrow \infty} x_n\right) = \varphi(x^*),$$

т.е. x^* – корень нашего уравнения.

Установим скорость сходимости. Для этого в неравенстве $|x_{n+p} - x_n| \leq mq^n \frac{1}{1-q}$, полученном выше, перейдем к пределу при $p \rightarrow \infty$:

$$\lim_{p \rightarrow \infty} |x_{n+p} - x_n| = |x^* - x_n| \leq \frac{m}{1-q} q^n.$$

Остается проверить единственность корня. Предположим противное: существуют два корня рассматриваемого уравнения – x' и x'' . Тогда

$$|x' - x''| = |\varphi(x') - \varphi(x'')| \leq q|x' - x''|.$$

Следовательно, $|x' - x''|(1 - q) \leq 0$, а поскольку $q < 1$, то отсюда получаем, что $|x' - x''| \leq 0$, т.е. $x' = x''$, что и завершает доказательство теоремы.

[\[Перейти к основному тексту\]](#)



Лемма 4.1

Справедлива следующая цепочка соотношений:

$$|x^* - x_n| = |\varphi(x^*) - \varphi(x_{n-1})| \leq q |x^* - x_{n-1}| = q |x^* - x_n + x_n - x_{n-1}| \leq q (|x^* - x_n| + |x_n - x_{n-1}|).$$

Отсюда, решая получившееся неравенство относительно $|x^* - x_n|$, получим (4.8).

[\[Перейти к основному тексту\]](#)



Лемма 4.2

Пользуясь условием (4.10), имеем:

$$\begin{aligned}\varphi(\varphi(x)) &= \varphi(x^* + \alpha(x - x^*) + o(x - x^*)) = x^* + \alpha(\alpha(x - x^*) + o(x - x^*)) + \\ &+ o(x - x^*) = x^* + \alpha^2(x - x^*) + o(x - x^*) .\end{aligned}$$

Тогда

$$\begin{aligned}\Phi(x) - x^* &= \frac{x\varphi(\varphi(x)) - \varphi^2(x)}{\varphi(\varphi(x)) - 2\varphi(x) + x} - x^* = \frac{(x-x^*)\varphi(\varphi(x)) - \varphi(x)(\varphi(x)-2x^*) - xx^*}{\varphi(\varphi(x)) - 2\varphi(x) + x} = \\ &= \frac{(x-x^*)(x^* + \alpha^2(x-x^*) + o(x-x^*)) - (x^* + \alpha(x-x^*) + o(x-x^*))(-x^* + \alpha(x-x^*) + o(x-x^*)) - xx^*}{x^* + \alpha^2(x-x^*) + o(x-x^*) - 2x^* - 2\alpha(x-x^*) + o(x-x^*) + x} = \\ &= \frac{o(x-x^*)^2}{(x-x^*)(\alpha-1)^2 + o(x-x^*)} = o(x - x^*) .\end{aligned}$$

Отсюда непосредственно следует формула (4.11).

[\[Перейти к основному тексту\]](#)



Теорема 4.3

Рассмотрим вначале возможность построения последовательности $\{x_n\}$. Для этого вновь, как и при доказательстве теоремы о сходимости метода итерации, воспользуемся методом математической индукции.

Первый член последовательности можно построить:

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)},$$

причем $f'(x_0) \neq 0$, что оговорено условием 1). Чтобы говорить о возможности построения x_2 , необходимо проверить, что $x_1 \in S_0$ и что $f'(x_1) \neq 0$. Так как $x_1 = x_0 + h_0$, то x_1 лежит ровно посередине отрезка S_0 . Далее, поскольку

$$|f'(x_1) - f'(x_0)| = \left| \int_{x_0}^{x_1} f''(x) dx \right| \leq M|x_1 - x_0| = M|h_0| \leq M \cdot \frac{|f'(x_0)|}{2M} = \frac{|f'(x_0)|}{2},$$

то

$$|f'(x_1)| = |f'(x_0) - [f'(x_0) - f'(x_1)]| \geq |f'(x_0)| - |f'(x_1) - f'(x_0)| \geq |f'(x_0)| - \frac{|f'(x_0)|}{2} = \frac{|f'(x_0)|}{2},$$

откуда следует, что . Поэтому величина $x_2 = x_1 + h_1$, где $h_1 = -\frac{f(x_1)}{f'(x_1)}$, определена. Наряду с отрезком S_0 рассмотрим теперь отрезок S_1 с концами x_1 и $x_1 + 2h_1$, серединой которого будет x_2 , и покажем, что $S_1 \subset S_0$. Для этого оценим величину $|x^* - x_{n+1}| \leq |h_n| \leq \frac{h_{n-1}^2 M}{2|f'(x_n)|}$, $n = 0, 1, \dots$. По формуле Тейлора получим:

$$|f(x_1)| = \left| f(x_0) + h_0 f'(x_0) + \frac{h_0^2}{2} f''(x_0 + \theta h_0) \right| = \left| \frac{h_0^2}{2} f''(x_0 + \theta h_0) \right| \leq \frac{h_0^2}{2} M.$$

Отсюда

$$|h_1| \leq \frac{h_0^2 M}{2|f'(x_1)|} \leq \frac{h_0^2 M \cdot 2}{2|f'(x_0)|} = \frac{h_0^2 \cdot M}{|f'(x_0)|} \leq \frac{|h_0|}{2}.$$



Тогда длина отрезка S_1 равна $2|h_1| \leq |h_0|$, т.е. $x_1 + 2h_1 = x_0 + h_0 + 2h_1 \leq x_0 + 2h_0 \in S_0$. Таким образом, $S_1 \subset S_0$. Кроме того, заметим, что выполняется неравенство $2|h_1|M \leq |f'(x_1)|$. Действительно, умножив обе части неравенства $|h_1| \leq \frac{|h_0|}{2}$ на $\frac{2M}{|f'(x_1)|}$, получим:

$$\frac{2|h_1|M}{|f'(x_1)|} \leq \frac{2|h_0|M}{2 \cdot |f'(x_1)|} \leq \frac{|f'(x_0)|}{|f'(x_1)|} = 1,$$

т.е. $2|h_1|M \leq |f'(x_1)|$.

Таким образом, на отрезке S_1 функция $f(x)$ обладает всеми теми свойствами, что и на S_0 . Теперь по индукции очевидна возможность построения последовательности $\{x_{n+1}\}$, $n = 0, 1, \dots$ по правилу

$$x_{n+1} = x_n + h_n, \quad n = 0, 1, \dots, \quad h_n = -\frac{f(x_n)}{f'(x_n)}.$$

При этом точка x_{n+1} будет серединой отрезка S_n с концами x_n и $x_n + 2h_n$, принадлежащего отрезку S_{n-1} , определяемому аналогично, и не превосходящего половины длины последнего. Кроме того, будет выполняться неравенство

$$|h_n| \leq \frac{h_{n-1}^2 M}{2|f'(x_n)|}, \tag{*}$$

Таким образом, мы построили последовательность вложенных отрезков $S_0 \supset S_1 \supset \dots$, длины которых с ростом n стремятся к нулю. Поэтому эти отрезки должны стягиваться к точке, а значит, последовательность $\{x_{n+1}\}$, элементы которой являются серединами отрезков S_n , должна сходиться к некоторому значению x^* .

Покажем, что x^* – корень уравнения $f(x) = 0$. Перейдем к пределу в алгоритме Ньютона при $n \rightarrow \infty$ и учтем непрерывность $f(x)$ при условии $f'(x_n) \neq 0$ для всех значений n :

$$\lim_{n \rightarrow \infty} f(x_n) = f\left(\lim_{n \rightarrow \infty} x_n\right) = f(x^*) = 0.$$

Проверим теперь единственность корня x^* . Для этого можно предположить, что $M > 0$ (если $M = 0$,



то $f'(x)$ – линейная функция и метод Ньютона сразу приведет к корню). Кроме того, $f'(x_0) \neq 0$ и $f'(x_0 + 2h_0) \neq 0$. Отсюда следует, что $f'(x) \neq 0$ для любой точки x из отрезка S_0 . Действительно,

$$|f'(x) - f'(x_0)| = \left| \int_{x_0}^x f''(t) dt \right| \leq M|x - x_0| < 2|h_0|M \leq |f'(x_0)|.$$

Отсюда

$$|f'(x) - f'(x_0)| = \left| \int_{x_0}^x f''(t) dt \right| \leq M|x - x_0| < 2|h_0|M \leq |f'(x_0)|.$$

т.е. $f'(x) \neq 0$ для любого x из S_0 .

Это означает, что $f(x)$ на S_0 строго монотонна, а следовательно, уравнение $f(x) = 0$ имеет не более одного корня.

Проверим, наконец, справедливость оценки (4.22). Точка x_{n+1} – середина отрезка S_n длины $2|h_n|$. Кроме того, $x^* \in S_n$. Поэтому

$$|x^* - x_{n+1}| \leq |h_n| \leq \frac{h_{n-1}^2 M}{2|f'(x_n)|}, \quad n = 0, 1, \dots$$

[\[Перейти к основному тексту\]](#)



Теорема 4.4

Пусть для определенности $f'(x) > 0$ и $f''(x) > 0$ для любого $x \in [a, b]$. Монотонность последовательности $\{x_n\}$ докажем по индукции. По условию $x_0 = b > x^*$. Так как $f'(x) > 0$, то $f(b) > 0$ и $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} < x_0$.

Пусть для некоторого $k \geq 0$ выполняются неравенства $x^* < x_k \leq b$. Докажем, что тогда $x^* < x_{k+1} < x_k$. Имеем:

$$x_k - x_{k+1} = \frac{f(x_k)}{f'(x_k)} = \frac{f(x_k) - f(x^*)}{f'(x_k)} = \frac{(x_k - x^*) f'(\xi_k)}{f'(x_k)},$$

где $\xi_k \in (x^*, x_k)$ (формула конечных приращений Лагранжа). При сделанных предположениях $f'(\xi_k) > 0$, $f'(x_k) > 0$ и $f'(x)$ монотонно возрастает на отрезке $[a, b]$. Поэтому $f'(x_k) > f'(\xi_k)$, а значит, $0 < \frac{f'(\xi_k)}{f'(x_k)} < 1$ и, следовательно, $0 < x_k - x_{k+1} < x_k - x^*$. Таким образом, $\{x_n\}$ монотонно убывает и ограничена снизу числом x^* . Поэтому она имеет предел, который в силу непрерывности $f(x)$ и условия $f'(x) \neq 0$ на $[a, b]$ совпадает с корнем x^* .

[\[Перейти к основному тексту\]](#)



Теорема 4.5

Дословно повторяет соответствующее доказательство теоремы 4.2 с заменой знака $|\cdot|$ на $\rho(\cdot, \cdot)$ (учитывая, что модуль – один из частных способов задания расстояния, причем для расстояния справедливо то же неравенство треугольника).

[\[Перейти к основному тексту\]](#)



Теорема 4.6

Индукцией по k покажем, что все $x^{(k)}$ принадлежат шару $\Omega(x^{(*)}, b)$. По условию $x^{(0)} \in \Omega(x^{(*)}, b)$. Пусть при некотором $k = n$ это также справедливо. Тогда, так как $b \leq \delta$, то $x^{(n)} \in \Omega(x^{(*)}, \delta)$. Тогда, подставив в (4.54) $x_1 = x^{(*)}$, $x_2 = x^{(n)}$, получим:

$$\left\| f(x^{(*)}) - f(x^{(n)}) - \frac{\partial f(x^{(n)})}{\partial x} (x^{(*)} - x^{(n)}) \right\| \leq a_2 \|x^{(*)} - x^{(n)}\|^2.$$

Поскольку $f(x^{(n)}) = -\frac{\partial f(x^{(n)})}{\partial x} (x^{(n+1)} - x^{(n)})$, а $f(x^{(*)}) = 0$, то отсюда имеем:

$$\left\| \frac{\partial f(x^{(n)})}{\partial x} (x^{(n+1)} - x^{(*)}) \right\| \leq a_2 \|x^{(n)} - x^{(*)}\|^2.$$

Тогда, воспользовавшись условием (4.53), получим:

$$\begin{aligned} \|x^{(n+1)} - x^{(*)}\| &= \left\| \left[\frac{\partial f(x^{(n)})}{\partial x} \right]^{-1} \left[\frac{\partial f(x^{(n)})}{\partial x} (x^{(n+1)} - x^{(*)}) \right] \right\| \leq \\ &\leq \left\| \left[\frac{\partial f(x^{(n)})}{\partial x} \right]^{-1} \right\| \left\| \left[\frac{\partial f(x^{(n)})}{\partial x} (x^{(n+1)} - x^{(*)}) \right] \right\| \leq a_1 a_2 \|x^{(n)} - x^{(*)}\|^2, \end{aligned}$$

т.е.

$$\|x^{(n+1)} - x^{(*)}\| \leq c \|x^{(n)} - x^{(*)}\|^2 < cb^2 = (cb)b \leq b$$

и, таким образом, $x^{(n+1)} \in \Omega(x^{(*)}, b)$. Следовательно, по индукции все $x^{(k)} \in \Omega(x^{(*)}, b)$ и для них выполняется неравенство

$$\|x^{(k+1)} - x^{(*)}\| \leq c \|x^{(k)} - x^{(*)}\|^2,$$



из которого получаем

$$c \|x^{(k)} - x^{(*)}\| \leq \left(c \|x^{(0)} - x^{(*)}\| \right)^{2^k},$$

а поскольку $c \|x^{(0)} - x^{(*)}\| < cb \leq 1$, то отсюда следует сходимость последовательности приближений.

[\[Перейти к основному тексту\]](#)



Теорема 5.1

Ключевым моментом доказательства существования является использованием того факта, что любая непрерывная на компакте функция достигает на нем своих граней ([теорема Вейерштрасса](#)), причем нижняя грань на специально подобранном компакте даст искомое наилучшее приближение. Итак, поскольку в силу [неравенства треугольника](#)

$$\left\| f - \sum_{i=0}^n c_i^1 \varphi_i \right\| - \left\| f - \sum_{i=0}^n c_i^2 \varphi_i \right\| \leq \left\| \sum_{i=0}^n (c_i^1 - c_i^2) \varphi_i \right\| \leq \sum_{i=0}^n |c_i^1 - c_i^2| \|\varphi_i\|$$

то функция

$$F(c_0, c_1, \dots, c_n) = \|f - \varphi\| = \left\| f - \sum_{i=0}^n c_i \varphi_i \right\|$$

является непрерывной функцией своих аргументов c_i , $i = 0, 1, \dots, n$ при любом $f \in R$. Пусть $|c| = \text{евклидова норма вектора } c = (c_0, c_1, \dots, c_n)$. Функция $F_0(c_0, c_1, \dots, c_n) = \|c_0 \varphi_0 + c_1 \varphi_1 + \dots + c_n \varphi_n\|$ непрерывна на единичной сфере $|c| = 1$ и, поскольку в конечномерном пространстве единичная сфера – компакт, то, по теореме Вейерштрасса, в некоторой ее точке $(c_0^*, c_1^*, \dots, c_n^*)$ достигает своей нижней грани F^* по сфере, причем $F^* \neq 0$, так как равенство $F^* = \|c_0^* \varphi_0 + c_1^* \varphi_1 + \dots + c_n^* \varphi_n\| = 0$ противоречит линейной независимости элементов $\varphi_0, \varphi_1, \dots, \varphi_n$. Тогда для любого $c \neq 0$ справедлива оценка

$$\begin{aligned} \|c_0 \varphi_0 + c_1 \varphi_1 + \dots + c_n \varphi_n\| &= F_0(c_0, c_1, \dots, c_n) = \\ &= |c| F_0\left(\frac{c_0}{|c|}, \frac{c_1}{|c|}, \dots, \frac{c_n}{|c|}\right) \geq |c| \cdot F^*. \end{aligned}$$

Далее, функция $F(c_0, c_1, \dots, c_n)$ непрерывна в замкнутом шаре $|c| \leq \gamma$, а следовательно, в некоторой его точке достигает своей нижней грани по шару F^0 . При этом $F^0 \leq F(0, 0, \dots, 0) = \|f\|$. Вне этого шара выполняются соотношения

$$F(c_0, c_1, \dots, c_n) \geq \|c_0 \varphi_0 + c_1 \varphi_1 + \dots + c_n \varphi_n\| - \|f\| \geq |c| \cdot F^* - \|f\| > \gamma F^* - \|f\|.$$



Если теперь выбрать $\gamma > \frac{2\|f\|}{F^*}$, то из записанного выше неравенства будет следовать, что $F(c_0, c_1, \dots, c_n) > \frac{2\|f\|}{F^*} F^* - \|f\| = \|f\| \geq F^* = F(c_0^0, c_1^0, \dots, c_n^0)$. Таким образом, $F(c_0, c_1, \dots, c_n) \geq F^* = F(c_0^0, c_1^0, \dots, c_n^0)$ при всех возможных значениях c_0, c_1, \dots, c_n , т.е. существует, по крайней мере, один элемент Φ_0 , для которого выполнено соотношение (4). Если теперь $\Phi_0 = \sum_{i=0}^n c_i \varphi_i$, $\Phi_0^* = \sum_{i=0}^n c_i^* \varphi_i$ – два элемента наилучшего приближения, то

$$\|f - \Phi_0\| = \|f - \Phi_0^*\| = \Delta(f).$$

В случае, если $\Delta(f) = 0$, имеем: $f = \Phi_0 = \Phi_0^*$. Поэтому рассмотрим случай $\Delta(f) > 0$. Пусть m – точка отрезка, соединяющего Φ_0 с Φ_0^* , т.е.

$$m = a\Phi_0 + b\Phi_0^*, \quad a, b \geq 0, \quad a + b = 1.$$

Тогда

$$\Delta(f) \leq \|f - m\| = \|a(f - \Phi_0) + b(f - \Phi_0^*)\| \leq a\|f - \Phi_0\| + b\|f - \Phi_0^*\| = \Delta(f).$$

Следовательно, $\|f - m\| = \Delta(f)$, а значит, m также является элементом наилучшего приближения и множество всех элементов наилучшего приближения выпукло. [\[Перейти к основному тексту\]](#)



Теорема 5.2

Пусть существует два элемента наилучшего приближения Φ_0 и Φ_0^* , т.е.

$$\|f - \Phi_0\| = \|f - \Phi_0^*\| = \Delta(f),$$

и пусть $\Delta(f) > 0$, ибо в противном случае $f = \Phi_0 = \Phi_0^*$. В силу выпуклости множества всех элементов наилучшего приближения имеем: $\frac{\Phi_0 + \Phi_0^*}{2}$ – элемент наилучшего приближения, т.е. $\left\| f - \frac{\Phi_0 + \Phi_0^*}{2} \right\| = \Delta(f)$. Но тогда

$$\|(f - \Phi_0) + (f - \Phi_0^*)\| = \|f - \Phi_0\| + \|f - \Phi_0^*\| > 0.$$

Так как R строго нормировано, то существует λ такое, что $f - \Phi_0 = \lambda(f - \Phi_0^*)$. Если $\lambda \neq 1$, то отсюда $f = \frac{\Phi_0^* - \lambda\Phi_0}{1-\lambda}$, т.е. f – элемент пространства, натянутого на Φ_0, Φ_0^* , что невозможно в силу того, что $\Delta(f) > 0$. Следовательно, $\lambda = 1$ и $\Phi_0 = \Phi_0^*$.

[\[Перейти к основному тексту\]](#)



Теорема 5.3

Пусть h_0 – элемент наилучшего приближения и пусть также существует элемент $h_1 \in H$ такой что $(f - h_0, h_1) = \alpha \neq 0$. Можно считать, что $\|h_1\| = 1$, так как в противном случае можно было бы вместо h_1 взять $\frac{h_1}{\|h_1\|}$. Рассмотрим элемент $h_2 = h_0 + \alpha h_1$. Для этого элемента имеем:

$$\begin{aligned} \|f - h_2\|^2 &= (f - h_2, f - h_2) = (f - h_0 - \alpha h_1, f - h_0 - \alpha h_1) = \\ &= (f - h_0, f - h_0) - \alpha (h_1, f - h_0) - \bar{\alpha} (f - h_0, h_1) + \alpha \bar{\alpha} (h_1, h_1) = \\ &= \|f - h_0\|^2 - \alpha \bar{\alpha} - \bar{\alpha} \alpha + \alpha \bar{\alpha} = \|f - h_0\|^2 - |\alpha|^2 \end{aligned}$$

т.е. $\|f - h_2\|^2 < \|f - h_0\|^2$, что невозможно, так как по предположению h_0 – элемент наилучшего приближения. Пусть теперь h – произвольный элемент пространства H и для него выполняется условие $(f - h_0, h) = 0$. Тогда имеем:

$$\begin{aligned} \|f - h, f - h\|^2 &= (f - h_0 + h_0 - h, f - h_0 + h_0 - h) = (f - h_0, f - h_0) + \\ &\quad + (h_0 - h, f - h_0) + (f - h_0, h_0 - h) + (h_0 - h, h_0 - h) . \end{aligned}$$

Поскольку $h_0 - h \in H$, то второе и третье слагаемые обращаются в нуль. Следовательно,

$$\|f - h\|^2 = \|f - h_0\|^2 + \|h - h_0\|^2$$

и, значит, $\|f - h\|^2 > \|f - h_0\|^2$ при всех $h \neq h_0$, т.е. h_0 – элемент наилучшего приближения.

[\[Перейти к основному тексту\]](#)



Теорема 5.4 (Валле-Пуссена)

В случае $\mu = 0$ утверждение теоремы очевидно, поскольку $\Delta_n(f) = \|f - Q_n^0\| \geq 0$. Пусть теперь $\mu > 0$. Предположим противное: для многочлена наилучшего равномерного приближения $Q_n^0(x)$ выполняется неравенство $\|Q_n^0 - f\| = \Delta_n(f) < \mu$. Имеем:

$$\operatorname{sign}(Q_n(x) - Q_n^0(x)) = \operatorname{sign}((Q_n(x) - f(x)) - (Q_n^0(x) - f(x))).$$

Тогда силу предположения в точках x_i первое слагаемое превосходит по модулю второе, ибо $\Delta_n(f) = \|f - Q_n^0\| = \sup_{x \in [a,b]} |f(x) - Q_n^0(x)| < \mu = \min_{0 \leq i \leq n+1} |f(x_i) - Q_n(x_i)|$. Поэтому

$$\operatorname{sign}(Q_n(x_i) - Q_n^0(x_i)) = \operatorname{sign}(Q_n(x_i) - f(x_i))$$

и, следовательно, многочлен $Q_n(x) - Q_n^0(x)$ степени n меняет знак $n + 1$ раз, чего быть не может.

[\[Перейти к основному тексту\]](#)



Теорема 5.5 (Чебышева)

Вначале докажем достаточность. Итак, пусть на отрезке $[a, b]$ существует система точек $x_0 < x_1 < \dots < x_{n+1}$, для которых выполняются условия теоремы. Обозначим через L величину: $L = \|f - Q_n\|$. Тогда в силу (5.1) теоремы 1 имеем: $L = \mu \leq \Delta_n(f)$. С другой же стороны в силу самого определения величины $\Delta_n(f)$ следует, что $\Delta_n(f) \leq \|f - Q_n\| = L$. Сопоставляя оба этих неравенства, получаем, что $\Delta_n(f) = L$ и, таким образом, многочлен $Q_n(x)$ является многочленом наилучшего равномерного приближения функции $f(x)$ на отрезке $[a, b]$.

Необходимость. Пусть теперь многочлен $Q_n(x)$ является многочленом наилучшего равномерного приближения функции $f(x)$ на отрезке $[a, b]$, т.е. $L = \|f - Q_n\| = \inf_{Q_n^o} \|f - Q_n^o\|$. Обозначим через y_1 нижнюю грань точек $x \in [a, b]$, в которых $|f(x) - Q_n(x)| = L$. Из определения величины L и непрерывности нормы следует существование таких точек, причем $|f(y_1) - Q_n(y_1)| = L$. Для определенности далее рассматриваем случай $f(y_1) - Q_n(y_1) = +L$. Обозначим через y_2 нижнюю грань всех точек $x \in (y_1; b]$, в которых $f(x) - Q_n(x) = -L$; последовательно через y_{k+1} обозначим нижнюю грань всех точек $x \in (y_k; b]$, в которых $f(x) - Q_n(x) = (-1)^k L$. Вследствие непрерывности разности $f(x) - Q_n(x)$ при всех k имеем:

$$f(y_{k+1}) - Q_n(y_{k+1}) = (-1)^k L.$$

Продолжим этот процесс до значения $y_m = b$ или значение y_m таково, что при $y_m < x \leq b$ нет точек x таких, что $f(x) - Q_n(x) = (-1)^m L$ (но могут быть точки, в которых $f(x) - Q_n(x) = (-1)^{m-1} L$). Если в результате указанного процесса получилось, что $m \geq n+2$, то утверждение теоремы выполнено. Поэтому предположим противное, т.е. что $m < n+2$. Вследствие непрерывности функции $f(x) - Q_n(x)$ при любом k ($1 < k \leq m$) можно указать точку z_{k-1} такую, что $|f(x) - Q_n(x)| < L$ при $z_{k-1} \leq x < y_k$. Положим $z_0 = a$, $z_m = b$ (на рисунке указанные точки отмечены треугольниками). Согласно проведенным построениям на отрезках $[z_{i-1}; z_i]$, $i = \overline{1, m}$ имеются точки (в частности, точки y_i), в которых $f(x) - Q_n(x) = (-1)^{i-1} L$, и нет точек, в которых $f(x) - Q_n(x) = (-1)^i L$. Положим

$$v(x) = \prod_{j=1}^{m-1} (z_j - x), \quad Q_n^d(x) = Q_n(x) + d v(x),$$



где $d > 0$, и рассмотрим поведение разности

$$f(x) - Q_n^d(x) = f(x) - Q_n(x) - d\nu(x)$$

на отрезках $[z_{i-1}; z_i]$. Для отрезка $[z_0; z_1]$ имеем: на $[z_0; z_1] \nu(x) > 0$, поэтому

$$f(x) - Q_n^d(x) \leq L - d\nu(x) < L.$$

В то же время на этом отрезке выполняется неравенство $f(x) - Q_n(x) > -L$, поэтому при достаточно малых d (например, при $d < d_1 = \frac{\min_{x \in [z_0; z_1]} |f(x) - Q_n(x) + L|}{\max_{x \in [z_0; z_1]} |\nu(x)|}$) на $[z_0; z_1]$ имеем: $f(x) - Q_n^d(x) > -L$. Кроме того,

$$|f(z_1) - Q_n^d(z_1)| = |f(z_1) - Q_n(z_1)| < L.$$

Таким образом, $|f(x) - Q_n^d(x)| < L$ на этом отрезке при достаточно малом d . После проведения аналогичных рассуждений относительно остальных отрезков $[z_{i-1}; z_i]$ мы сможем указать малое d_0 такое, что на всех отрезках выполняется неравенство $|f(x) - Q_n^d(x)| < L$. Получено противоречие с тем, что $Q_n(x)$ – многочлен наилучшего равномерного приближения, а $m < n + 2$.

[\[Перейти к основному тексту\]](#)



Теорема 5.6

Предположим, что существуют два многочлена наилучшего равномерного приближения:

$$Q_n^1(x) \neq Q_n^2(x), \quad \|f - Q_n^1\| = \|f - Q_n^2\| = \Delta_n(f).$$

Так как множество элементов наилучшего приближения выпукло, то многочлен $\frac{1}{2} [Q_n^1(x) + Q_n^2(x)]$ также является многочленом наилучшего равномерного приближения. Пусть x_0, x_1, \dots, x_{n+1} – соответствующие этому многочлену точки чебышевского альтернанса. Тогда

$$\left| \frac{1}{2} [Q_n^1(x_i) + Q_n^2(x_i)] - f(x_i) \right| = \Delta_n(f), \quad i = 0, 1, \dots, n+1$$

или

$$|[Q_n^1(x_i) - f(x_i)] + [Q_n^2(x_i) - f(x_i)]| = 2\Delta_n(f). \quad (**)$$

Так как

$$|Q_n^k(x_i) - f(x_i)| \leq \Delta_n(f), \quad k = 1, 2$$

(в силу того что $|Q_n^k(x) - f(x)| \leq \sup_{x \in [a,b]} |Q_n^k(x) - f(x)| = \Delta_n(f)$, $k = 1, 2$), то соотношение $(**)$ возможно лишь в том случае, когда

$$Q_n^1(x_i) - f(x_i) = Q_n^2(x_i) - f(x_i).$$

Отсюда следует, что два различных многочлена $Q_n^1(x)$ и $Q_n^2(x)$ степени n совпадают в $n+2$ различных точках, чего не может быть. Полученное противоречие доказывает теорему.

[\[Перейти к основному тексту\]](#)



Теорема 5.13

Заметив, что в силу [свойств разделенных разностей](#) справедливы равенства

$$\begin{cases} \frac{1}{3}d_0 = 2f(x_0, x_0, x_1) = f''(\xi_0), \quad \xi_0 \in [x_0; x_1], \\ \frac{1}{3}d_j = 2f(x_{j-1}, x_j, x_{j+1}) = f''(\xi_j), \quad \xi_j \in [x_{j-1}; x_{j+1}], \quad j = \overline{1, N-1}, \\ \frac{1}{3}d_N = 2f(x_{N-1}, x_N, x_N) = f''(\xi_N), \quad \xi_N \in [x_{N-1}; x_N], \end{cases} \quad (\text{Д.2})$$

Преобразуем систему [\(5.94\)](#), вычитая из обеих ее частей вектор $\frac{1}{3}Ad$, к виду

$$A\left(M - \frac{1}{3}d\right) = \left(E - \frac{1}{3}A\right)d. \quad (\text{Д.3})$$

При этом, очевидно, координаты вектора $M - \frac{1}{3}d$ (в силу [\(Д.2\)](#)) представляют собой разности между значениями второй производной сплайна в узлах разбиения Δ и второй производной приближаемой функции в некоторых промежуточных точках. Правые же части системы [\(Д.3\)](#) будут иметь вид

$$\begin{cases} c_0 = d_0 - \frac{1}{3}(Ad)_0 = d_0 - \frac{1}{3}(2d_0 + d_1) = \frac{1}{3}(d_0 - d_1); \\ c_j = d_j - \frac{1}{3}(Ad)_j = d_j - \frac{1}{3}(\mu_j d_{j-1} + 2d_j + \lambda_{j+1}) = \frac{1}{3}(d_j - \mu_j d_{j-1} - \lambda_j d_{j+1}) = \\ = \frac{1}{3}[\mu_j(d_j - d_{j-1}) - \lambda_j(d_{j+1} - d_j)], \quad j = \overline{1, N-1}, \\ c_N = d_N - \frac{1}{3}(Ad)_N = d_N - \frac{1}{3}(d_{N-1} + 2d_N) = \frac{1}{3}(d_N - d_{N-1}). \end{cases}$$

При выводе этих формул мы учли, что $\mu_j + \lambda_j = 1$.

Оценим сейчас компоненты вектора c . Учитывая [\(Д.2\)](#), имеем:

$$|c_0| = \frac{1}{3}|d_0 - d_1| = |f''(\xi_0) - f''(\xi_1)| \leq 2\omega(h; f'')$$

(последнее неравенство можно получить, например, так: поскольку $x_0 \leq \xi_0 \leq x_1$, $x_0 \leq \xi_1 \leq x_2$, то



$|f''(\xi_0) - f''(\xi_1)| \leq |f''(\xi_0) - f''(\frac{\xi_0+\xi_1}{2})| + |f''(\frac{\xi_0+\xi_1}{2}) - f''(\xi_1)| \leq \omega(h; f'') + \omega(h; f'')$ в силу того что $|\xi_1 - \xi_0| \leq 2h$ и, следовательно, $|\xi_0 - \frac{\xi_0+\xi_1}{2}| \leq h$ и $|\xi_1 - \frac{\xi_0+\xi_1}{2}| \leq h$; аналогично,

$$|c_j| = |\mu_j [f''(\xi_j) - f''(\xi_{j-1})] - \lambda_j [f''(\xi_{j+1}) - f''(\xi_j)]| \leq \mu_j |[f''(\xi_j) - f''(\xi_{j-1})]| +$$

$$+ \lambda_j |[f''(\xi_j) - f''(\xi_{j-1})]| \leq \mu_j \cdot 3\omega(h; f'') + \lambda_j \cdot 3\omega(h; f''), \quad j = \overline{1, N-1},$$

поскольку $x_{j-2} \leq \xi_{j-1} \leq x_j, \quad x_{j-1} \leq \xi_j \leq x_{j+1}, \quad x_j \leq \xi_{j+1} \leq x_{j+2}$, т.е. $|\xi_j - \xi_{j-1}| \leq 3h, \quad |\xi_{j+1} - \xi_j| \leq 3h$;

$$|c_N| = \frac{1}{3} |d_N - d_{N-1}| = |f''(\xi_N) - f''(\xi_{N-1})| \leq 2\omega(h; f'').$$

Следовательно, объединяя эти оценки, получим:

$$\|c\| \leq 3\omega(h; f''),$$

а поскольку из системы (Д.3) следует, что $M - \frac{1}{3}d = A^{-1}c$ и в силу леммы $\|A^{-1}\| \leq 1$, то отсюда

$$\left\| M - \frac{1}{3}d \right\| = \|A^{-1}c\| \leq \|A^{-1}\| \cdot \|c\| \leq \|c\| \leq 3\omega(h; f''). \quad (\text{Д.4})$$

Далее, учитывая вид второй производной сплайна (формула (5.90)), получаем:

$$\begin{aligned} \left| f''(x) - \frac{d^2}{dx^2} S_\Delta^3(f; x) \right| &= \left| f''(x) - M_{j-1} \frac{x_j-x}{h_j} - M_j \frac{x-x_{j-1}}{h_j} \right| = \\ &= \left| \frac{x_j-x}{h_j} (f''(x) - M_{j-1}) + \frac{x-x_{j-1}}{h_j} (f''(x) - M_j) \right| \leq \frac{x_j-x}{h_j} |f''(x) - M_{j-1}| + \frac{x-x_{j-1}}{h_j} |f''(x) - M_j| \end{aligned} \quad (\text{Д.5})$$

для всех $x \in [x_{j-1}; x_j], \quad j = \overline{1, N}$.

Оценим теперь последовательно величины $|f''(x) - M_j|$ и $|f''(x) - M_{j-1}|$. Имеем:

$$|f''(x) - M_j| = \left| f''(x) - \frac{1}{3}d_j + \frac{1}{3}d_j - M_j \right| \leq \left| f''(x) - \frac{1}{3}d_j \right| + \left| \frac{1}{3}d_j - M_j \right|.$$



Для первого слагаемого получаем:

$$\left| f''(x) - \frac{1}{3}d_j \right| = |f''(x) - f''(\xi_j)| \leq 2\omega(h; f'')$$

так как $|x - \xi_j| \leq 2h$, а для второго имеем оценку ([Д.4](#)).

Поэтому

$$|f''(x) - M_j| \leq 2\omega(h; f'') + 3\omega(h; f'') = 5\omega(h; f''), \quad x \in [x_{j-1}; x_j], \quad j = \overline{1, N}.$$

Отсюда непосредственно следует, что $\|f''(x) - M\| \leq 5\omega(h; f'')$ и, следовательно, из ([Д.5](#)) имеем:

$$\left| f''(x) - \frac{d^2}{dx^2} S_\Delta^3(f; x) \right| \leq \left(\frac{x_j - x}{h_j} + \frac{x - x_{j-1}}{h_j} \right) \cdot 5\omega(h; f'') = 5\omega(h; f''). \quad (\text{Д.6})$$

Таким образом, оценка для вторых производных установлена. Так как по условию сплайн интерполяционный, то выполняются условия

$$\begin{cases} S_\Delta^3(f; x_j) = f(x_j), \\ S_\Delta^3(f; x_{j-1}) = f(x_{j-1}), \end{cases} \quad j = \overline{1, N}.$$

Поэтому в силу теоремы Ролля на отрезке $[x_{j-1}; x_j]$ существует точка ξ_j^* , в которой совпадают значения производных функции $f(x)$ и сплайна $S_\Delta^3(f; x)$, т.е. выполняется равенство $f'(\xi_j^*) = \frac{d}{dx} S_\Delta^3(f; \xi_j^*)$. Следовательно

$$\begin{aligned} \left| f'(x) - \frac{d}{dx} S_\Delta^3(f; x) \right| &= \left| [f'(x) - f'(\xi_j^*)] - \left[\frac{d}{dx} S_\Delta^3(f; x) - \frac{d}{dx} S_\Delta^3(f; \xi_j^*) \right] \right| = \\ &= \left| \int_{\xi_j^*}^x \left[f''(x) - \frac{d^2}{dx^2} S_\Delta^3(f; x) \right] dx \right| \leq \left| \int_{\xi_j^*}^x \left| f''(x) - \frac{d^2}{dx^2} S_\Delta^3(f; x) \right| dx \right| \leq 5\omega(h; f'') \cdot |x - \xi_j^*| \leq \\ &\leq 5h\omega(h; f'') \end{aligned}$$



Аналогично

$$\begin{aligned}
 |f(x) - S_{\Delta}^3(f; x)| &= \left| [f(x) - f(x_{j-1})] - \left[S_{\Delta}^3(f; x) - \frac{d}{dx} S_{\Delta}^3(f; x_{j-1}) \right] \right| = \\
 &= \left| \int_{x_{j-1}}^x \left[f'(x) - \frac{d}{dx} S_{\Delta}^3(f; x) \right] dx \right| \leq \left| \int_{\xi_{j-1}}^x \left| f'(x) - \frac{d}{dx} S_{\Delta}^3(f; x) \right| dx \right| \leq \\
 &\leq 5\omega(h; f'') \cdot \left| \int_{x_{j-1}}^x (x - \xi_j^*) dx \right| = \\
 &= 5\omega(h; f'') \cdot \frac{|\xi_j^* - x_{j-1}|^2}{2} \leq \frac{5}{2} h^2 \omega(h; f'') .
 \end{aligned}$$

Полученные неравенства завершают доказательство.

[\[Перейти к основному тексту\]](#)



Теорема 6.1

Необходимость. Пусть рассматриваемая квадратурная формула является интерполяционной, а $f(x)$ – алгебраический многочлен степени не выше n . Тогда, очевидно, $r_n(x) \equiv 0$ (см., например, [представление остатка интерполирования в форме Лагранжа](#)). Следовательно, и $R_n(f) = 0$, т.е. квадратурная формула точна для таких $f(x)$.

Достаточность. Пусть квадратурная формула точна для всех многочленов до степени n включительно. Докажем, что ее коэффициенты вычисляются по формулам (6.6), т.е. что

$$A_k = \int_a^b p(x) \Phi_k(x) dx, \quad k = \overline{0, n}.$$

Рассмотрим $\int_a^b p(x) \Phi_k(x) dx$. Так как $\Phi_j(x)$ – многочлен степени n , то для него квадратурная формула точна, т.е.

$$\int_a^b p(x) \Phi_j(x) dx = \sum_{k=0}^n A_k \Phi_j(x_k),$$

а поскольку $\Phi_j(x_k) = \delta_j^k$, то отсюда следует, что

$$\int_a^b p(x) \Phi_j(x) dx = \sum_{k=0}^n A_k \Phi_j(x_k) = A_j, \quad j = \overline{0, n}.$$

[\[Перейти к основному тексту\]](#)



Свойства квадратурных формул Ньютона–Котеса



Свойство 1

Действительно,

$$\begin{aligned}
 B_{n-k}^n &= \frac{(-1)^k}{n \cdot k! \cdot (n-k)!} \int_0^n \frac{t(t-1)\cdots(t-n)}{t-n+k} dt = [t=n-z] = \\
 &\int_n^0 \frac{(n-z)(n-z-1)(n-z-2)\cdots(-z)}{-z+k} (-dz) = \frac{(-1)^k}{n \cdot k! \cdot (n-k)!} \cdot \\
 &\cdot \frac{(-1)^k}{n \cdot k! \cdot (n-k)!} \int_0^n \frac{(-1)^{n+1} z(z-1)\cdots(z-n)}{-(z-k)} dz = \frac{(-1)^{n+k}}{n \cdot k! \cdot (n-k)!} \cdot \\
 &\cdot \int_0^n \frac{z(z-1)\cdots(z-n)}{z-k} dz = B_k^n.
 \end{aligned}$$

[\[Перейти к основному тексту\]](#)



Свойство 2

В самом деле, выполняя в интеграле замену переменной по формуле $t = x - \frac{a+b}{2}$ (относительно переменной t функция f будет просто нечетной), имеем: с одной стороны

$$\int_a^b f(x) dx = \int_{-\frac{b-a}{2}}^{\frac{b-a}{2}} f(t) dt = 0,$$

а с другой

$$\sum_{k=0}^n B_k^n f(a + kh) = (B_0^n f(a) + B_n^n f(b)) + (B_1^n f(a + h) + B_{n-1}^n f(b - h)) + \dots = 0$$

(причем в каждой скобке сумма равна нулю как сумма противоположных слагаемых, а то слагаемое, которому нет пары (если такое имеется), само равно нулю, поскольку $f\left(\frac{a+b}{2}\right) = 0$). Таким образом, $\int_a^b f(x) dx = (b - a) \sum_{k=0}^n B_k^n f(a + kh) = 0$.

[\[Перейти к основному тексту\]](#)



Свойство 3

Действительно, так как формула (6.9) является интерполяционной, то она точна для всех алгебраических многочленов до степени n включительно, а в силу свойства 2 она точна и для многочлена $c \left(x - \frac{a+b}{2}\right)^{n+1}$. Следовательно, она будет точной и для базисной функции x^{n+1} .

[\[Перейти к основному тексту\]](#)



Теорема 6.2 (критерий квадратурных формул НАСТ)

Необходимость. Так как квадратурная формула точна для любых многочленов до степени $(2n + 1)$ включительно, то она точна и для многочленов до степени n включительно, а тогда по [теореме 6.1](#) она является интерполяционной, т.е. ее коэффициенты вычисляются по формулам [\(6.43\)](#).

Далее, рассмотрим произвольный многочлен $Q_m(x)$ степени $m \leq n$. Тогда степень многочлена $\omega_{n+1}(x)Q_m(x)$ будет не выше $(2n + 1)$. Следовательно, формула [\(6.1\)](#) точна для такого произведения, т.е. справедливо равенство

$$\int_a^b p(x) \omega_{n+1}(x) Q_m(x) dx = \sum_{k=0}^n A_k \omega_{n+1}(x_k) Q_m(x_k) = 0$$

(последнее равенство цепочки имеет место в силу того, что при любых $k = \overline{0, n}$ узлы x_k квадратурной суммы являются корнями многочлена $\omega_{n+1}(x)$, т.е. $\omega_{n+1}(x_k) = 0$). Таким образом, выполняется соотношение [\(6.44\)](#).

Достаточность. Рассмотрим произвольный алгебраический многочлен $f(x)$ степени не выше $(2n + 1)$. Покажем, что для него квадратурная формула точна. Выполнив деление $f(x)$ на $\omega_{n+1}(x)$ с остатком, получим:

$$f(x) = Q_m(x) \omega_{n+1}(x) + r(x),$$

где степень частного $Q_m(x)$ не превосходит n и степень остатка также будет не выше n . Из записанного равенства в силу того что точки x_k – корни многочлена $\omega_{n+1}(x)$, следует, что $f(x_k) = r(x_k)$, $k = \overline{0, n}$. Тогда

$$\int_a^b p(x) f(x) dx = \int_a^b p(x) \omega_{n+1}(x) Q_m(x) dx + \int_a^b p(x) r(x) dx \stackrel{(6.44)}{=} \int_a^b p(x) r(x) dx$$

$$\stackrel{(6.44)}{=} 0 + \int_a^b p(x) r(x) dx \stackrel{(6.43)}{=} \sum_{k=0}^n A_k r(x_k) = \sum_{k=0}^n A_k f(x_k).$$



Таким образом, требуемое соотношение установлено.

[\[Перейти к основному тексту\]](#)



Теорема 6.3

Запишем многочлен $\omega_{n+1}(x)$ с неопределенными коэффициентами:

$$\omega_{n+1}(x) = x^{n+1} + a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n.$$

Нахождение многочлена $\omega_{n+1}(x)$ эквивалентно нахождению его коэффициентов. Построим систему для нахождения величин a_i , ($i = \overline{0, n}$). Для этих целей воспользуемся условием ортогональности многочлена $\omega_{n+1}(x)$ многочленам 1, x , x^2, \dots, x^n . В итоге получим:

$$\int_a^b p(x) (x^{n+1} + a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n) x^k dx = 0, \quad k = \overline{0, n} \quad (\text{Д.7})$$

Для однозначной разрешимости системы (Д.7) достаточно показать, что соответствующая ей однородная система

$$\int_a^b p(x) (a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n) x^k dx = 0, \quad k = \overline{0, n}, \quad (\text{Д.8})$$

имеет лишь тривиальное решение $a_i = 0$, ($i = \overline{0, n}$).

Умножим k -е уравнение системы (Д.8) на a_{n-k} и просуммируем получившиеся равенства по всем значениям $k = \overline{0, n}$. Тогда будем иметь:

$$\int_a^b p(x) (a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n)^2 dx = 0.$$

В силу того, что весовая функция $p(x)$ сохраняет знак на отрезке $[a, b]$ и $p(x)$ отлична от тождественного нуля, следует, что $a_i = 0$, ($i = \overline{0, n}$). Таким образом, многочлен $\omega_{n+1}(x)$ всегда может быть построен, причем единственным образом.



Рассмотрим теперь корни $\xi_1, \xi_2, \dots, \xi_m$ многочлена $\omega_{n+1}(x)$ нечетной кратности, лежащие внутри отрезка $[a, b]$. Существование хотя бы одного такого корня следует из ортогональности многочлена $\omega_{n+1}(x)$ к $P_0(x) \equiv 1$. Действительно,

$$\int_a^b p(x) \omega_{n+1}(x) dx = 0,$$

а так как весовая функция $p(x)$ сохраняет знак, то $\omega_{n+1}(x)$ обязан на $[a, b]$ знак поменять, ибо в противном случае значение рассмотренного выше интеграла будет отличным от нуля (как интеграла от знакопостоянной функции).

В точках ξ_i многочлен $\omega_{n+1}(x)$, очевидно, меняет знак. Кроме того, $1 \leq m \leq n + 1$. Пусть $m < n + 1$. Тогда по корням $\xi_1, \xi_2, \dots, \xi_m$ построим многочлен степени $m < n + 1$ $Q_m(x) = (x - \xi_1) \cdots (x - \xi_m)$, к которому $\omega_{n+1}(x)$ должен быть ортогонален, т.е.

$$\int_a^b p(x) \omega_{n+1}(x) Q_m(x) dx = 0.$$

Но это равенство невозможно в силу того, что $\omega_{n+1}(x)$ и $Q_m(x)$ имеют одни и те же точки перемены знаков и, таким образом, подынтегральное выражение сохраняет знак на отрезке $[a, b]$. Следовательно, $m = n + 1$ и последнее утверждение теоремы доказано.

[\[Перейти к основному тексту\]](#)



Теорема 6.4

Рассмотрим многочлен $\omega_{n+1}^2(x)$. Его степень равна $2n + 2$. Очевидно, что $\int_a^b p(x) \omega_{n+1}^2(x) dx \neq 0$ ни при каком выборе x_k , так как подынтегральное выражение сохраняет знак на отрезке интегрирования. В то же время $\sum_{k=0}^n A_k \omega_{n+1}^2(x_k) = 0$, т.е.

$$\int_a^b p(x) \omega_{n+1}^2(x) dx \neq \sum_{k=0}^n A_k \omega_{n+1}^2(x_k).$$

Таким образом, если весовая функция $p(x)$ знакопостоянна на отрезке интегрирования $[a, b]$, то наивысшая алгебраическая степень точности квадратурной формулы с $(n + 1)$ узлами равна $2n + 1$. Такие квадратурные формулы называют квадратурными формулами типа Гаусса (или Гаусса-Кристоффеля).

[\[Перейти к основному тексту\]](#)



Теорема 6.5

Рассмотрим многочлен $\Phi_i^2(x)$ степени $2n$. Для него формула точна, поэтому

$$\int_a^b p(x) \Phi_i^2(x) dx = \sum_{k=0}^n A_k \Phi_i^2(x_k) = A_i, \quad i = \overline{0, n}.$$

Таким образом, знаки всех весовых коэффициентов совпадают со знаком весовой функции.

[\[Перейти к основному тексту\]](#)



Теорема 6.6

(аналогично доказательству [теоремы 6.2](#))

Необходимость. Так как квадратурная формула (6.72) точна для всех многочленов до степени $2n-m+1$ включительно и $2n-m+1 \geq 2n-(n+1)+1 = n$, то она точна и для многочленов степени n и, таким образом, является интерполяционной (в силу критерия интерполяционных квадратурных формул).

Кроме того, положив $f(x) = \Omega_m(x) \omega_{n-m+1}(x) Q(x)$, где $Q(x)$ – произвольный многочлен степени не выше $n-m$, имеем: $f(x)$ – многочлен, степень которого не превосходит $2n-m+1$. Следовательно, для такой $f(x)$ квадратурная формула (6.72) точна. Поэтому

$$\int_a^b p(x) \Omega_m(x) \omega_{n-m+1}(x) Q(x) dx = \sum_{l=1}^m B_l \Omega_m(a_l) \omega_{n-m+1}(a_l) \cdot \\ \cdot Q(a_l) + \sum_{k=0}^{n-m} A_k \Omega_m(x_k) \omega_{n-m+1}(x_k) Q(x_k) = 0.$$

Достаточность. Пусть теперь $f(x)$ – произвольный многочлен, степень которого не превосходит $2n-m+1$. Представим его в виде

$$f(x) = \Omega_m(x) \omega_{n-m+1}(x) Q(x) + r(x),$$

где $Q(x)$ – многочлен степени не выше $n-m$, а $r(x)$ – многочлен степени не выше n (остаток от деления $f(x)$ на $\Omega_m(x) \omega_{n-m+1}(x)$). При этом, очевидно,

$$f(a_l) = r(a_l), \quad l = 1, \dots, m$$

$$f(x_k) = r(x_k), \quad k = 0, \dots, n-m.$$



Тогда имеем:

$$\int_a^b p(x) f(x) dx = \int_a^b p(x) \Omega_m(x) \omega_{n-m+1}(x) Q(x) dx + \int_a^b p(x) r(x) dx \stackrel{(6.74)}{=} 0 + \int_a^b p(x) r(x) dx \stackrel{(6.74)}{=} \sum_{l=1}^m B_l r(a_l) + \sum_{k=0}^{n-m} A_k r(x_k) = \sum_{l=1}^m B_l f(a_l) + \sum_{k=0}^{n-m} A_k f(x_k).$$

Полученное равенство завершает доказательство.

[\[Перейти к основному тексту\]](#)



Теорема 6.7

Так как $\Omega_m(x) \tilde{\Pi}_{n-m}(x)$ есть многочлен степени n со старшим членом x^n , то его можно разложить по многочленам системы $\tilde{P}_s(x)$ в виде

$$\Omega_m(x) \tilde{\Pi}_{n-m}(x) = \tilde{P}_n(x) + c_1 \tilde{P}_{n-1}(x) + c_2 \tilde{P}_{n-2}(x) + \dots.$$

Ортогональность $\tilde{\Pi}_{n-m}(x)$ с весом $p(x) \Omega_m(x)$ ко всякому многочлену степени, меньшей $n-m$, равносильна тому, что в указанном разложении должны отсутствовать члены с $\tilde{P}_s(x)$ для $s \leq n-m-1$. Этот факт легко установить, если умножить записанное соотношение на произведение $p(x) \tilde{P}_s(x)$ и проинтегрировать по отрезку $[a, b]$. Тогда получится равенство

$$0 = c_s \int_a^b p(x) \tilde{P}_s^2 dx,$$

из которого немедленно следует: $c_s = 0$. Таким образом,

$$\Omega_m(x) \tilde{\Pi}_{n-m}(x) = \tilde{P}_n(x) + c_1 \tilde{P}_{n-1}(x) + c_2 \tilde{P}_{n-2}(x) + \dots + c_m \tilde{P}_{n-m}(x) \quad (\text{Д.9})$$

Полагая в (Д.9) x поочередно равным a_1, a_2, \dots, a_m , получим систему уравнений для определения коэффициентов разложения c_i :

$$\begin{cases} \tilde{P}_n(a_1) + c_1 \tilde{P}_{n-1}(a_1) + c_2 \tilde{P}_{n-2}(a_1) + \dots + c_m \tilde{P}_{n-m}(a_1) = 0, \\ \dots \\ \tilde{P}_n(a_m) + c_1 \tilde{P}_{n-1}(a_m) + c_2 \tilde{P}_{n-2}(a_m) + \dots + c_m \tilde{P}_{n-m}(a_m) = 0 \end{cases} \quad (\text{Д.10})$$

(эти равенства одновременно служат гарантами того, что правая часть равенства (Д.9) делится на $\Omega_m(x)$ нацело, поскольку свидетельствуют о том, что у $\Omega_m(x)$ и правой части (Д.9) одни и те же корни).

Так как по условию $\tilde{\Pi}_{n-m}(x)$ существует, то система (Д.10) должна иметь единственное решение, т.е. ее определитель, совпадающий с выписанным в условии теоремы Δ , отличен от нуля.



Чтобы получить формулу (6.77), добавим к (Д.10) уравнение (Д.9) и рассмотрим получившуюся систему

$$\left\{ \begin{array}{l} (-\Omega_m(x) \cdot \tilde{\Pi}_{n-m}(x) + \tilde{P}_n(x)) \cdot 1 + c_1 \tilde{P}_{n-1}(x) + c_2 \tilde{P}_{n-2}(x) + \cdots + \\ + c_m \tilde{P}_{n-m}(x) = 0, \\ \tilde{P}_n(a_1) \cdot 1 + c_1 \tilde{P}_{n-1}(a_1) + c_2 \tilde{P}_{n-2}(a_1) + \cdots + c_m \tilde{P}_{n-m}(a_1) = 0, \\ \dots\dots \\ \tilde{P}_n(a_m) \cdot 1 + c_1 \tilde{P}_{n-1}(a_m) + c_2 \tilde{P}_{n-2}(a_m) + \cdots + c_m \tilde{P}_{n-m}(a_m) = 0 \end{array} \right.$$

как однородную систему, состоящую из $(m+1)$ уравнений с неизвестными $1, c_1, \dots, c_m$. Так как данная система имеет ненулевое решение, то ее определитель должен быть равен нулю, т.е.

$$\begin{vmatrix} -\Omega_m(x) \tilde{\Pi}_{n-m}(x) + \tilde{P}_n(x) & \tilde{P}_{n-1}(x) & \dots & \tilde{P}_{n-m}(x) \\ 0 + \tilde{P}_n(a_1) & \tilde{P}_{n-1}(a_1) & \dots & \tilde{P}_{n-m}(a_1) \\ \vdots & \vdots & \dots & \vdots \\ 0 + \tilde{P}_n(a_m) & \tilde{P}_{n-1}(a_m) & \dots & \tilde{P}_{n-m}(a_m) \end{vmatrix} = 0.$$

[\[Перейти к основному тексту\]](#)



Теорема 7.1

При $\alpha > 0$ функционал $M(\alpha, f, u)$ ограничен снизу. Поэтому при данных α и $f(x)$ он имеет точную нижнюю грань: $\bar{M} = \inf_{u \in U} M(\alpha, f, u)$. Выберем некоторую минимизирующую последовательность $u_i(s)$ так, что $\lim_{i \rightarrow \infty} M_i = \bar{M}$, где $M_i = M(\alpha, f, u_i)$. Упорядочим ее так, чтобы M_i не возрастили. Тогда

$$\alpha \Omega(u_i) \leq M_i \leq M_0 = \text{const}$$

или

$$\Omega(u_i) \leq \frac{1}{\alpha} M_0.$$

Таким образом, последовательность $u_i(s)$ принадлежит множеству таких $u(s)$, для которых $\Omega(u) \leq \text{const}$. А это множество, как известно, является компактом в U . Поэтому из последовательности $u_i(s)$ можно выделить подпоследовательность $u_{i_k}(s)$, сходящуюся по норме к некоторой $u_\alpha(s) \in U$. В силу непрерывности функционала $M(\alpha, f, u)$ на этой функции достигает своей точной нижней грани. Тем самым $u_\alpha(s) \in U$ есть решение задачи (7.71).

[\[Перейти к основному тексту\]](#)



Теорема 7.2

Пусть $\bar{u}(s)$ – решение задачи (7.70) с правой частью $\bar{f}(x)$, $\tilde{u}_\alpha(s)$ – решение задачи (7.71) с приближенной правой частью $\tilde{f}(x)$ и $f_\alpha(x) = A(x, \tilde{u}_\alpha(s))$. Поскольку функционал $M(\alpha, \tilde{f}, u)$ достигает минимума на элементе \tilde{u}_α , то справедливо неравенство

$$M(\alpha, \tilde{f}, \tilde{u}_\alpha) \leq M(\alpha, \tilde{f}, \bar{u}).$$

Отсюда получаем:

$$\begin{aligned} \alpha \Omega(\tilde{u}_\alpha) &\leq M(\alpha, \tilde{f}, \tilde{u}_\alpha) \leq M(\alpha, \tilde{f}, \bar{u}) = \int_c^d [A(x, \bar{u}) - \tilde{f}(x)]^2 dx + \alpha \Omega(\bar{u}) = \\ &= \int_c^d [\bar{f}(x) - \tilde{f}(x)]^2 dx + \alpha \Omega(\bar{u}) = \|\bar{f} - \tilde{f}\|_{L_2}^2 + \alpha \Omega(\bar{u}). \end{aligned} \quad (\text{Д.11})$$

Пусть приближенные правые части удовлетворяют условию

$$\|\bar{f} - \tilde{f}\|_{L_2} \leq C\sqrt{\alpha}, \quad (\text{Д.12})$$

где C – некоторая константа. Тогда из неравенства (Д.11) следует, что

$$\Omega(\tilde{u}_\alpha) \leq C^2 + \Omega(\bar{u}) = \text{const}. \quad (\text{Д.13})$$

Значит, \tilde{u}_α принадлежит компактному множеству U_0 функций из U (заметим, что \bar{u} также принадлежит U_0).

Множество F_0 функций $f_\alpha(x)$ есть образ множества U_0 при отображении A . По предположению оператор A непрерывен и таков, что обратное отображение единственно. Поэтому обратное отображение F_0 в компактное множество U_0 при помощи нерегуляризованного оператора A^{-1} будет непрерывным в норме пространства U . Следовательно, по заданному $\varepsilon > 0$ всегда найдется $\beta(\varepsilon)$ такое, что из условия $\|f_\alpha - \bar{f}\| \leq \beta(\varepsilon)$ следует выполнение неравенства $\|\tilde{u}_\alpha - \bar{u}\| \leq \varepsilon$.

Заметим, что

$$\|f_\alpha - \bar{f}\|^2 = \int_c^d [f_\alpha(x) - \bar{f}(x)]^2 dx = \int_c^d [A(x, \tilde{u}_\alpha) - \bar{f}(x)]^2 dx \leq M(\alpha, \tilde{f}, \tilde{u}_\alpha) \leq \alpha(C^2 + \Omega(\bar{u})).$$



Отсюда с учетом условия (Д.12) следует:

$$\|f_\alpha - \tilde{f}\| \leq \|f_\alpha - \tilde{f}\| + \|\tilde{f} - \bar{f}\| \leq \sqrt{\alpha} \left(C + \sqrt{C^2 + \Omega(\bar{u})} \right). \quad (\text{Д.14})$$

Выберем α так, чтобы

$$\alpha \leq \alpha_0(\varepsilon) \equiv \left(\frac{\beta(\varepsilon)}{C + \sqrt{C^2 + \Omega(\bar{u})}} \right)^2. \quad (\text{Д.15})$$

Тогда правая часть неравенства (Д.14) не будет превосходить $\beta(\varepsilon)$, откуда следует, что $\|\tilde{u}_\alpha - \bar{u}\| \leq \varepsilon$.

Таким образом, по заданному ε нашлись $\alpha_0(\varepsilon)$ и $\delta(\alpha) = C\sqrt{\alpha}$ такие, что выполнение неравенств $\alpha \leq \alpha_0(\varepsilon)$ и $\|\tilde{f} - \bar{f}\| \leq \delta(\alpha)$ влечет выполнение неравенства $\|\tilde{u}_\alpha - \bar{u}\| \leq \varepsilon$.

[\[Перейти к основному тексту\]](#)



Следствие 7.1

немедленно следует из того, что заключительная строка доказательства теоремы, по сути, означает непрерывную зависимость решений от правых частей задачи.

[[Перейти к основному тексту](#)]



Лемма 8.1

Так как $u(t, \xi, \eta)$ удовлетворяет исходному уравнению (8.4), то $\frac{\partial u(t, \xi, \eta)}{\partial t} = f(t, u(t, \xi, \eta))$. Продифференцируем последнее соотношение по переменной η :

$$\frac{\partial}{\partial \eta} \frac{\partial u(t, \xi, \eta)}{\partial t} = \frac{\partial f(t, u(t, \xi, \eta))}{\partial u} \cdot \frac{\partial u(t, \xi, \eta)}{\partial \eta}$$

или, меняя порядок дифференцирования слева и вводя обозначение $\frac{\partial u(t, \xi, \eta)}{\partial \eta} = z$:

$$\frac{\partial}{\partial t} z(t) = \frac{\partial f(t, u(t, \xi, \eta))}{\partial \eta} \cdot z(t). \quad (\text{Д.16})$$

Так как $z(\xi) = \frac{\partial}{\partial \eta} u(\xi, \xi, \eta) = \frac{\partial}{\partial \eta} \eta = 1$, то из (Д.16) получим:

$$z(t) = \frac{\partial u(t, \xi, \eta)}{\partial t} = \exp \int_{\xi}^t \frac{\partial f(x, u(x, \xi, \eta))}{\partial u} dx.$$

[\[Перейти к основному тексту\]](#)



Теорема 9.1

Возьмем произвольный элемент $v \in H_A$ и положим $v = u_1 + t$. Тогда

$$\begin{aligned} J(v) &= (Av, v) - 2(f, v) = (A(u_1 + t), u_1 + t) - 2(f, u_1 + t) = \\ &= J(u_1) + 2(Au_1 - f, t) + (At, t) = J(u_1) + (At, t) \geq J(u_1). \end{aligned}$$

[\[Перейти к основному тексту\]](#)



Теорема 9.2

Возьмем произвольный элемент $v \in H_A$ и произвольное вещественное число λ . Тогда $u_1 + \lambda v \in H_A$ и по условию $J(u_1 + \lambda v) \geq J(u_1)$. С другой стороны

$$J(u_1 + \lambda v) = (A(u_1 + \lambda v), u_1 + \lambda v) - 2(f, u_1 + \lambda v) = J(u_1) + 2\lambda(Au_1 - f, v) + \lambda^2(Av, v).$$

Поэтому для всех λ выполняется неравенство

$$\lambda^2(Av, v) + 2\lambda(Au_1 - f, v) \geq 0.$$

Последнее же неравенство справедливо для любых действительных λ только в том случае, когда $(Au_1 - f, v) = 0$.

Так как H_A всюду плотно в H , то отсюда следует, что $Au_1 - f = 0$.

[\[Перейти к основному тексту\]](#)



Теорема 9.3

Пусть $u^*(x)$ — функция, доставляющая минимум функционалу $J(u)$ на классе G допустимых функций. Так как $u^*(x) \in G$ и система функций $\{u_n(x)\}$ обладает свойством C^1 -полноты, то для любого $\delta > 0$ найдутся такие n и $\tilde{a}_1, \dots, \tilde{a}_n$, что для всех x из отрезка $[a, b]$ будут выполнены неравенства

$$|u^*(x) - \tilde{u}_n(x)| \leq \delta, \quad \left| (u^*)'(x) - \tilde{u}'_n(x) \right| \leq \delta.$$

Учитывая непрерывность функции $F(x, u, u')$, можно для любого значения $\varepsilon > 0$ выбрать такое δ , что будут иметь место неравенства

$$0 \leq J(\tilde{u}_n) - J(u^*) \leq \varepsilon. \quad (**)$$

В то же время для последовательности Ритца $\{u_n^*(x)\}$ выполняются неравенства

$$J(u^*) \leq J(u_n^*) \leq J(\tilde{u}_n).$$

Поэтому из неравенств $(**)$ следует, что

$$0 \leq J(u_n^*) - J(u^*) \leq \varepsilon,$$

а отсюда в силу произвольности ε имеем:

$$\lim_{n \rightarrow \infty} J(u_n^*) = J(u^*) = m.$$

[\[Перейти к основному тексту\]](#)



Теорема 9.4

Рассмотрим функцию $\varepsilon_n(x) = u^*(x) - u_n(x)$, где $u_n(x)$ — n -й элемент произвольной минимизирующей последовательности. Так как $\varepsilon_n(a) = 0$, то

$$\varepsilon_n(x) = \int_a^x \varepsilon'_n(t) dt.$$

Отсюда, используя неравенство Коши-Буняковского, получим:

$$\begin{aligned} |\varepsilon_n(x)| &= \left| \int_a^x \varepsilon'_n(t) dt \right| \leqslant \left| \int_a^x 1^2 dt \right|^{\frac{1}{2}} \cdot \left| \int_a^x (\varepsilon'_n(t))^2 dt \right|^{\frac{1}{2}} \leqslant \sqrt{b-a} \left(\int_a^b (\varepsilon'_n(t))^2 dt \right)^{\frac{1}{2}} \leqslant \\ &\leqslant \sqrt{b-a} \left(\int_a^b \frac{p(t)}{\min_{x \in [a,b]} p(x)} (\varepsilon'_n(t))^2 dt \right)^{\frac{1}{2}} \leqslant \sqrt{\frac{b-a}{\min_{x \in [a,b]} p(x)}} \left\{ \int_a^b \left[p(t) (\varepsilon'_n(t))^2 + q(t) \varepsilon_n^2(t) \right] dt \right\}^{\frac{1}{2}} = \\ &= \sqrt{\frac{b-a}{p_0}} \sqrt{J_n(u) - J(u^*)}. \end{aligned}$$

Так как $\{u_n(x)\}$ — минимизирующая последовательность, то $J(u_n) \xrightarrow{n \rightarrow \infty} J(u^*)$, причем независимо от x . Поэтому $\varepsilon_n(x) \xrightarrow{n \rightarrow \infty} 0$ или $u_n(x) \xrightarrow{n \rightarrow \infty} u^*(x)$.

[\[Перейти к основному тексту\]](#)



Теорема 10.1 (Лакса)

Если оператор $\tilde{L}_h = (L_h, I_h)$ линеен и разностная схема (10.36a)- (10.36b) корректна, то на основании соотношения (10.39) можно записать:

$$\| z_h \|_{(1_h)} \leq M \| \tilde{\psi}_h \|_{(2_h)} \quad \text{или} \quad \| z_h \|_{(1_h)} \leq M \left(\| \psi_h \|_{(2_h)} + \| \nu_h \|_{(3_h)} \right) \quad (\text{Д.17})$$

Так как разностная схема аппроксимирует исходную дифференциальную задачу, то отсюда непосредственно получаем утверждение теоремы.

[\[Перейти к основному тексту\]](#)



Свойства собственных функций и собственных значений задачи (10.70)



Свойство 3

Для доказательства этого факта воспользуемся [второй разностной формулой Грина](#), записанной для однородных краевых условий:

$$0 = \left(y_{\bar{x}x}^{(k)}, y^{(m)} \right) - \left(y^{(k)}, y_{\bar{x}x}^{(m)} \right) = (\lambda_k - \lambda_m) \left(y^{(k)}, y^{(m)} \right)$$

Так как $\lambda_k \neq \lambda_m$, то отсюда следует: $(y^{(k)}, y^{(m)}) = 0$.

[\[Перейти к основному тексту\]](#)



Свойство 4

Действительно,

$$\|y^{(k)}\|^2 = \left(y^{(k)}, y^{(k)}\right) = \sum_{s=1}^{N-1} \left(y^{(k)}(x_s)\right)^2 h = \sum_{k=1}^{N-1} h \sin^2 \frac{k\pi x_s}{l} = \sum_{k=1}^{N-1} \frac{h}{2} \left(1 - \cos \frac{2k\pi x_s}{l}\right). \quad (\text{Д.18})$$

Вычислим сумму косинусов: пусть $q_k = \exp\left(i \frac{2k\pi h}{l}\right)$. Тогда

$$q_k^s = \exp\left(i \frac{2k\pi hs}{l}\right) = \exp\left(i \frac{2k\pi x_s}{l}\right); \quad q_k^N = \exp\left(i \frac{2k\pi x_N}{l}\right) = \exp(2k\pi i) = 1$$

и

$$\sum_{s=1}^{N-1} h \cos \frac{2k\pi x_s}{l} = \operatorname{Re} \sum_{s=1}^{N-1} h q_k^s = \operatorname{Re} \left(h \frac{q_k^N - q_k}{q_k - 1} \right) = \operatorname{Re} \left(h \frac{1 - q_k}{q_k - 1} \right) = -h.$$

Тогда из (Д.18) получим:

$$\|y^{(k)}\|^2 = \frac{h}{2} (N - 1 + 1) = \frac{Nh}{2} = \frac{l}{2},$$

что и требовалось установить.

[\[Перейти к основному тексту\]](#)



Свойство 6

Действительно,

$$\|f\|^2 = (f, f) = \left(\sum_{k=1}^{N-1} f_k \mu^{(k)}, \sum_{k=1}^{N-1} f_k \mu^{(k)} \right) = \sum_{k=1}^{N-1} f_k^2,$$

так как $(\mu^{(k)}, \mu^{(m)}) = \delta_k^m$.

[\[Перейти к основному тексту\]](#)



Лемма 10.1

На сетке $\bar{\omega}_h$ справедливо тождество

$$y^2(x) = (1-x)y^2(x) + xy^2(x). \quad (\text{Д.19})$$

Так как $y(0) = y(1) = 0$, то $y^2(x) = \left(\sum_{x'=h}^x y_{\bar{x}}(x') h\right)^2$ или $y^2(x) = \left(\sum_{x'=x+h}^1 y_{\bar{x}}(x') h\right)^2$.

Подставляя эти выражения в [\(Д.19\)](#), получим:

$$y^2(x) = (1-x) \left(\sum_{x'=h}^x y_{\bar{x}}(x') h \right)^2 + x \left(\sum_{x'=x+h}^1 y_{\bar{x}}(x') h \right)^2.$$

Теперь, используя [неравенство Коши-Буняковского](#), отсюда получим:

$$y^2(x) \leq (1-x) \sum_{x'=h}^x h \cdot \sum_{x'=h}^x hy_{\bar{x}}^2(x') + x \sum_{x'=x+h}^1 h \cdot \sum_{x'=x+h}^1 hy_{\bar{x}}^2(x') = x(1-x) \sum_{x'=h}^1 hy_{\bar{x}}^2(x') = x(1-x) \|y_{\bar{x}}\|^2,$$

а так как $x(1-x) \leq \frac{1}{4}$, то $y^2(x) \leq \frac{1}{4} \|y_{\bar{x}}\|^2$ или $\|y\|_C \leq \frac{1}{2} \|y_{\bar{x}}\|$.

[\[Перейти к основному тексту\]](#)



Лемма 10.2

Разложим $y(x)$ по собственным функциям задачи (10.70):

$$y(x) = \sum_{k=1}^{N-1} C_k \mu^{(k)}(x),$$

причём

$$C_k = \left(y(x), \mu^{(k)}(x) \right), \quad \|y\|^2 = \sum_{k=1}^{N-1} C_k^2.$$

В силу первой разностной формулы Грина

$$\|y_{\bar{x}}\|^2 = (-y_{\bar{x}x}, y),$$

а так как $\mu_{\bar{x}x}^{(k)} = -\lambda_k \mu^{(k)}$, то

$$\|y_{\bar{x}}\|^2 = \left(\sum_{k=1}^{N-1} C_k \lambda_k \mu^{(k)}, \sum_{k=1}^{N-1} C_k \mu^{(k)} \right) = \sum_{k=1}^{N-1} C_k^2 \lambda_k.$$

Отсюда получаем:

$$\lambda_1 \|y\| \leq \|y_{\bar{x}}\|^2 \leq \lambda_{N-1} \|y\|,$$

где

$$\lambda_1 = \frac{4}{h^2} \sin^2 \frac{\pi h}{2l}, \quad \lambda_{N-1} = \frac{4}{h^2} \cos^2 \frac{\pi h}{2l}.$$

Оценим λ_1 снизу. Пусть $\alpha = \frac{\pi h}{2l}$. Тогда

$$\lambda_1 = \frac{\pi^2}{l^2} \left(\frac{\sin \alpha}{\alpha} \right)^2.$$



Так как $h \leq \frac{l}{2}$, то $\alpha \in (0, \frac{\pi}{4}]$ и, следовательно, учитывая, что функция $f(\alpha) = \frac{\sin \alpha}{\alpha}$ монотонно убывает на данном промежутке, получаем:

$$\min_{\alpha \in (0, \frac{\pi}{4}]} \frac{\sin \alpha}{\alpha} = f\left(\frac{\pi}{4}\right) = \frac{2\sqrt{2}}{\pi}.$$

Поэтому

$$\lambda_1 \geq \frac{\pi^2}{l^2} \cdot \left(\frac{2\sqrt{2}}{\pi}\right)^2 = \frac{8}{l^2}.$$

С другой стороны, $\lambda_{N-1} < \frac{4}{h^2}$. Таким образом, получаем оценку

$$\frac{8}{l^2} \|y\|^2 \leq \|y_{\bar{x}}\|^2 \leq \frac{4}{h^2} \|y\|^2,$$

откуда непосредственно следует доказываемое неравенство (10.80).

[\[Перейти к основному тексту\]](#)



Теорема 10.2 (Принцип максимума)

Пусть, для определенности, $Sy(x) \leq 0$ и существует узел $\bar{x} \in \omega$, в котором

$$y(\bar{x}) = \max_{x \in \omega} y(x) > 0.$$

Тогда в этом узле

$$Sy(\bar{x}) \equiv A(\bar{x})y(\bar{x}) - \sum_{\xi \in \mathbb{W}'(\bar{x})} B(\bar{x}, \xi)y(\xi) = D(\bar{x})y(\bar{x}) + \sum_{\xi \in \mathbb{W}'(\bar{x})} B(\bar{x}, \xi)(y(\bar{x}) - y(\xi)) \geq 0.$$

Так как по условию теоремы $Sy(\bar{x}) \leq 0$, то отсюда следует, что $Sy(\bar{x}) = 0$, а значит, в силу неравенств $B(\bar{x}, \xi) > 0$, $y(\bar{x}) \geq y(\xi)$, $y(\bar{x}) > 0$ и $D(\bar{x}) \geq 0$ имеем: $D(\bar{x}) = 0$ и $y(\bar{x}) = y(\xi)$ для всех $\xi \in \mathbb{W}'(\bar{x})$.

Поскольку сеточная функция $y(x)$ отлична от тождественной константы при $x \in \omega$, то существует узел $\bar{x} \in \omega$ такой что $y(\bar{x}) < y(\bar{x})$. В силу связности сетки Ω можно указать последовательность узлов x_1, \dots, x_m , удовлетворяющую условиям (1.4). Тогда $y(x_1) = y(\bar{x})$. Повторяя рассуждения, приведенные выше, получим:

$$y(x_1) = y(x_2) = \dots = y(x_m) = y(\bar{x}).$$

Следовательно, для точки x_m получим неравенство

$$Sy(x_m) = D(x_m)y(x_m) + \sum_{\xi \in \mathbb{W}'(x_m)} B(x_m, \xi)(y(x_m) - y(\xi)) \geq B(x_m, \bar{x})(y(x_m) - y(\bar{x})) > 0,$$

которое противоречит условию теоремы.

[Перейти к основному тексту](#)



Следствие 10.1

Пусть $Sy(x) \leq 0$. Если $y(x) \equiv \text{const}$ на Ω , то

$$Sy(x_0) = D(x_0)y(x_0) + \sum_{\xi \in \mathbb{W}'(x_0)} B(x_0, \xi)(y(x_0) - y(\xi)) = D(x_0)y(x_0) \leq 0.$$

Поэтому $y(x) \equiv y(x_0) \leq 0$.

Если же $y(x)$ не является тождественно постоянной, то $y(x) \leq 0$ на основании принципа максимума (в соответствии с последним наибольшее положительное значение функция при указанных условиях может принимать только на границе, а там, поскольку $D(\xi) = 1$ для всех $\xi \in \gamma$, $y(\xi) \leq 0$).

[\[Перейти к основному тексту\]](#)



Следствие 10.2

Убедимся, что однородная задача, соответствующая (10.144), имеет лишь тривиальное решение. Поскольку $Sy(x) = 0 \leqslant 0$, то согласно [следствию 10.1](#) $y(x) \leqslant 0$ для всех $x \in \Omega$. С другой стороны, так как $Sy(x) = 0 \geqslant 0$, то по тому же [следствию](#) $y(x) \geqslant 0$ для всех $x \in \Omega$. Отсюда $y(x) \equiv 0$ для всех $x \in \Omega$.

[\[Перейти к основному тексту\]](#)



Теорема 10.3 (теорема сравнения)

Сложим и вычтем уравнения $Sy(x) = F(x)$ и $S\bar{y}(x) = \bar{F}(x)$. В силу линейности оператора S получим:

$$\begin{cases} S(\bar{y} + y) = \bar{F}(x) + F(x) \geq 0, \\ S(\bar{y} - y) = \bar{F}(x) - F(x) \geq 0. \end{cases}$$

Из первого из этих неравенств с помощью [следствия 10.1](#) получаем, что $\bar{y}(x) + y(x) \geq 0$, а из второго — $\bar{y}(x) - y(x) \geq 0$. Объединяя эти неравенства, получаем: $|y(x)| \leq \bar{y}(x)$.

Таким образом, решение задачи [\(10.144\)](#), [\(10.145\)](#) можно оценить с помощью *мажорантной функции* $\bar{y}(x)$, которая удовлетворяет уравнению $S\bar{y}(x) = \bar{F}(x)$ с правой частью $\bar{F}(x) \geq |F(x)|$ (например, $\bar{F}(x) = \|F(x)\|_C$ или $\bar{F}(x) = |F(x)|$ и т.п.).

[Перейти к основному тексту](#)



Следствие 10.3

Пусть $\bar{y}(x)$ — решение задачи

$$\begin{cases} S\bar{y}(x) = 0, & x \in \omega, \\ \bar{y}(x) = |\mu(x)|, & x \in \gamma. \end{cases}$$

Тогда на основании [теоремы 10.3](#) $|y(x)| \leq \bar{y}(x)$. Поэтому если $\bar{y}(x) \equiv \text{const}$ на Ω , то $\max_{x \in \Omega} \bar{y}(x) = \max_{x \in \gamma} |\mu(x)|$ и, следовательно, неравенство (1.6) справедливо. Если же $\bar{y}(x)$ отлично от тождественной константы, то в силу [теоремы 10.2](#) максимальное значение этой функции может достигаться только на границе, т.е. снова $\|\bar{y}(x)\|_{\mathcal{C}} \leq \|\mu(x)\|_{C_\gamma}$.

[\[Перейти к основному тексту\]](#)



Теорема 10.4

Пусть $\bar{y}(x)$ — решение задачи $S\bar{y}(x) = |F(x)|$, $x \in \Omega$. Тогда по [теореме 10.3](#) $|y(x)| \leq \bar{y}(x)$. Функция $\bar{y}(x)$ достигает наибольшего значения в некотором узле \bar{x} : $\bar{y}(\bar{x}) = \max_{x \in \Omega} \bar{y}(x) > 0$. Тогда

$$Sy(\bar{x}) = D(\bar{x})\bar{y}(\bar{x}) + \sum_{\xi \in \mathbb{W}'(\bar{x})} B(\bar{x}, \xi)(\bar{y}(\bar{x}) - \bar{y}(\xi)) = |F(\bar{x})|$$

и так как $\bar{y}(x) \geq \bar{y}(\xi)$, то $D(\bar{x})\bar{y}(\bar{x}) \leq |F(\bar{x})|$. Отсюда $\bar{y}(x) \leq \frac{|F(\bar{x})|}{D(\bar{x})}$ и поэтому

$$\|y\|_{\bar{C}} \leq \bar{y}(\bar{x}) = \max_{x \in \Omega} \bar{y}(x) = \|\bar{y}\|_{\bar{C}} \leq \frac{|F(\bar{x})|}{D(\bar{x})} \leq \max_{x \in \Omega} \left| \frac{F(x)}{D(x)} \right| = \left\| \frac{F}{D} \right\|_{\bar{C}}.$$

[\[Перейти к основному тексту\]](#)



Теорема 10.6 (признак устойчивости)

Разложим произвольную ошибку начальных данных по системе $\mu_k(x)$:

$$z(x_s, t_0) = \sum_{k=0}^{N-1} a_k \exp\left(ik \frac{2\pi}{l} x_s\right).$$

Тогда, поскольку разностная схема (10.159) линейна, то

$$z(x_s, t_j) = \sum_{k=0}^{N-1} a_k q_k^j \exp\left(ik \frac{2\pi}{l} x_s\right).$$

Следовательно, учитывая ортогональность системы $\{\mu_k(x)\}$, имеем:

$$\begin{aligned} \|z^j\|^2 &= (z^j, z^j) = l \cdot \sum_{k=0}^{N-1} |q_k^j|^2 \cdot |a_k|^2 \leq l \cdot \max_{0 \leq k \leq N-1} |q_k|^{2j} \sum_{k=0}^{N-1} |a_k|^2 = \max_{0 \leq k \leq N-1} |q_k|^{2j} \cdot \|z^0\|^2 \leq \\ &\leq (1 + C\tau)^{2j} \|z^0\|^2 \leq \exp(2Cj\tau) \|z^0\|^2 \leq \exp(2CT) \|z^0\|^2. \end{aligned}$$

[\[Перейти к основному тексту\]](#)



Теорема 10.7

Достаточность. Пусть выполнено условие (10.167). Из энергетического тождества (10.166) при $\varphi = 0$ (мы исследуем устойчивость по начальным данным) следует

$$2\tau \left(\left(B - \frac{\tau}{2} A \right) y_t, y_t \right) + (A\hat{y}, \hat{y}) = (Ay, y).$$

Отсюда в силу (10.167) имеем:

$$(A\hat{y}, \hat{y}) \leq (Ay, y)$$

или

$$\|\hat{y}\|_A^2 \leq \|y\|_A^2,$$

а тогда

$$\|y^{j+1}\|_A \leq \|y^j\|_A \leq \dots \leq \|y^0\|_A.$$

Необходимость. Пусть разностная схема (10.164) устойчива по начальным данным и выполнено неравенство (10.168). Докажем, что отсюда следует операторное неравенство (10.167), т.е. $(Bv, v) \geq \frac{\tau}{2} (Av, v)$ для всех $v \in H$.

Запишем энергетическое тождество при $t = 0$:

$$2\tau \left(\left(B - \frac{\tau}{2} A \right) y_t(0), y_t(0) \right) + (Ay^1, y^1) = (Ay^0, y^0).$$

В силу неравенства (10.168) это тождество может быть выполнено только при

$$2\tau \left(\left(B - \frac{\tau}{2} A \right) y_t(0), y_t(0) \right) = (Ay^0, y^0) - (Ay^1, y^1) \geq 0,$$

т.е.

$$2\tau \left(\left(B - \frac{\tau}{2} A \right) y_t(0), y_t(0) \right) \geq 0., \quad (*)$$

Так как $y^0 \in H$ — произвольный, то и элемент $v = y_t(0) = B^{-1}Ay^0 \in H$ также произведен. В самом деле, задавая любой элемент $v = y_t(0) \in H$, находим $y^0 = -A^{-1}Bv \in H$, так как оператор A^{-1} существует. Таким образом, неравенство (*) выполнено при любых $v = y_t(0) \in H$, т.е. имеет место операторное неравенство (10.167).

[Перейти к основному тексту]

Теорема 10.8 (Достаточное условие сходимости двухслойных итерационных схем)

Рассмотрим итерационную погрешность $\varepsilon^n = y^n - y \in \overset{\circ}{H}_h$. Ввиду (10.180) и (10.181) для неё получаем однородную задачу

$$\begin{cases} B \frac{\varepsilon^{n+1} - \varepsilon^n}{\tau} + A\varepsilon^n = 0, & x \in \omega_h, \\ \varepsilon^{n+1} = 0, & x \in \gamma_h, n = 0, 1, 2, \dots, \\ \varepsilon^0 = y^0 - y. \end{cases} \quad (\text{Д.20})$$

Перепишем уравнение (Д.20) в виде

$$(B - \frac{\tau}{2}A) \frac{\varepsilon^{n+1} - \varepsilon^n}{\tau} + \frac{1}{2}A(\varepsilon^n + \varepsilon^{n+1}) = 0$$

и умножим его скалярно на $2(\varepsilon^{n+1} - \varepsilon^n)$:

$$\begin{aligned} 2 \left((B - \frac{\tau}{2}A) \frac{\varepsilon^{n+1} - \varepsilon^n}{\tau}, \varepsilon^{n+1} - \varepsilon^n \right) + \left(A(\varepsilon^n + \varepsilon^{n+1}), \varepsilon^{n+1} - \varepsilon^n \right) &= 0 \Rightarrow \\ \Rightarrow 2\tau \left((B - \frac{\tau}{2}A) \frac{\varepsilon^{n+1} - \varepsilon^n}{\tau}, \frac{\varepsilon^{n+1} - \varepsilon^n}{\tau} \right) + (A\varepsilon^{n+1}, \varepsilon^{n+1}) - (A\varepsilon^n, \varepsilon^n) + \\ &\quad + (A\varepsilon^n, \varepsilon^{n+1}) - (A\varepsilon^{n+1}, \varepsilon^n) = 0. \end{aligned}$$

Последних два слагаемых взаимно сокращаются в силу свойства самосопряжённости оператора A . Таким образом, имеем:

$$2\tau \left((B - \frac{\tau}{2}A) \frac{\varepsilon^{n+1} - \varepsilon^n}{\tau}, \frac{\varepsilon^{n+1} - \varepsilon^n}{\tau} \right) + (A\varepsilon^{n+1}, \varepsilon^{n+1}) - (A\varepsilon^n, \varepsilon^n) = 0. \quad (\text{Д.21})$$

Отсюда, в силу (10.182), имеем

$$\begin{aligned}
 0 &\leqslant (A\epsilon^{n+1}, \epsilon^{n+1}) = (A\epsilon^n, \epsilon^n) - 2\tau \left((B - \frac{\tau}{2}A) \frac{\epsilon^{n+1} - \epsilon^n}{\tau}, \frac{\epsilon^{n+1} - \epsilon^n}{\tau} \right) < \\
 &< \left[(A\epsilon^n, \epsilon^n) > 0, \text{ так как } \epsilon^n \neq 0, \text{ вторая скобка} > 0 \text{ в силу } \epsilon^{n+1} - \epsilon^n \neq 0 \right] < \\
 &< (A\epsilon^n, \epsilon^n) < (A\epsilon^{n-1}, \epsilon^{n-1}) < \dots < (A\epsilon^0, \epsilon^0).
 \end{aligned}$$

Так как числовая последовательность $(A\epsilon^n, \epsilon^n)$, $n = 0, 1, 2, \dots$ является монотонно убывающей и ограниченной, то она сходится. Покажем, что $(A\epsilon^n, \epsilon^n) \xrightarrow{n \rightarrow \infty} 0$.

Поскольку $(A\epsilon^{n+1}, \epsilon^{n+1}) - (A\epsilon^n, \epsilon^n) \xrightarrow{n \rightarrow \infty} 0$, то из (Д.21) следует

$$\begin{aligned}
 ((B - \frac{\tau}{2}A)(\epsilon^{n+1} - \epsilon^n), \epsilon^{n+1} - \epsilon^n) &\xrightarrow{n \rightarrow \infty} 0 \Leftrightarrow [(10.182)] \epsilon^{n+1} - \epsilon^n \xrightarrow{n \rightarrow \infty} 0 \Rightarrow [(Д.20)] \\
 &\Rightarrow A\epsilon^n \xrightarrow{n \rightarrow \infty} 0 \Rightarrow (A\epsilon^n, \epsilon^n) \xrightarrow{n \rightarrow \infty} 0 \Leftrightarrow \epsilon^n \xrightarrow{n \rightarrow \infty} 0.
 \end{aligned}$$

[\[Перейти к основному тексту\]](#)



Следствие 10.5

Запишем (10.185) в матричной форме:

$$y^{n+1} = \frac{1}{\frac{2}{h_1^2} + \frac{2}{h_2^2}} (\Lambda y^n + \varphi) + y^n,$$

или

$$\left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) (y^{n+1} - y^n) = \Lambda y^n + \varphi.$$

Следовательно, МПИ (10.185) имеет канонический вид (10.181), где

$$B = E, A = -\Lambda, \tau = \frac{1}{\frac{2}{h_1^2} + \frac{2}{h_2^2}}.$$

Поскольку матрица $A = -\Lambda = A^* > 0$ в $\overset{\circ}{H}_h$, то согласно теореме для сходимости МПИ (10.185) достаточно, чтобы выполнялось условие $B - \frac{\tau}{2}A = E + \frac{\tau}{2}\Lambda > 0$, где

$$\tau = \frac{1}{\frac{2}{h_1^2} + \frac{2}{h_2^2}}.$$

Проверим это для $u \in \overset{\circ}{H}_h$:

$$\begin{aligned} ((E + \frac{\tau}{2}\Lambda)u, u) &= (u, u) + \frac{\tau}{2}(\Lambda u, u) = \sum_{i=1}^{N_1-1} \sum_{j=1}^{N_2-1} u_{ij}^2 + \frac{\tau}{2} \sum_i \sum_j \left[\frac{1}{h_1^2} (u_{i-1,j} + u_{i+1,j}) u_{ij} + \right. \\ &\quad \left. + \frac{1}{h_2^2} (u_{i,j-1} + u_{i,j+1}) u_{ij} - \tau u_{ij}^2 \right] = \sum_i \sum_j \left(\frac{1}{2} u_{ij}^2 + \frac{\tau}{h_1^2} u_{i-1,j} u_{ij} + \frac{\tau}{h_2^2} u_{i,j-1} u_{ij} \right) = \\ &= \sum_{i=1}^{N_1-1} \sum_{j=1}^{N_2-1} \left[\frac{1}{2} u_{ij}^2 + \frac{\tau}{2h_1^2} (u_{ij} + u_{i-1,j})^2 + \frac{\tau}{2h_2^2} (u_{ij} + u_{i,j-1})^2 - \frac{1}{2} u_{ij}^2 \right]. \end{aligned}$$



Это выражение либо > 0 , либо $= 0$ (тогда и только тогда, когда $u = 0$).

Следовательно, МПИ (10.185) для задачи (10.180) сходится.

[Перейти к основному тексту]



Следствие 10.6

Представим (10.187) в матричной форме:

$$Dy^{n+1} = (1 - \tau)Dy^n + \tau[\Lambda y^n + \left(\frac{2}{h_1^2} + \frac{2}{h_2^2}\right)y^n + \Lambda^-(y^{n+1} - y^n) + \varphi],$$

где

$$\begin{aligned} D &= \left(\frac{2}{h_1^2} + \frac{2}{h_2^2}\right)E, \quad (\Lambda^- y)_{ij} = \frac{1}{h_1^2}y_{i-1,j} + \frac{1}{h_2^2}y_{i,j-1}, \quad i = \overline{1, N_1 - 1}, \quad j = \overline{1, N_2 - 1} \Rightarrow \\ &\Rightarrow \left(D - \tau\Lambda^-\right)\frac{y^{n+1} - y^n}{\tau} - \Lambda y^n = \varphi. \end{aligned}$$

Следовательно, метод релаксации (10.187) имеет вид (10.180) при $A = -\Lambda$, $B = D - \tau\Lambda^-$. Поскольку $A = -\Lambda = A^* > 0$ в $\overset{\circ}{H}_h$, то, согласно общей теореме о сходимости, достаточно выбирать $\tau > 0$, при которых $B - \frac{1}{2}\tau A > 0$ в $\overset{\circ}{H}_h$.

Пусть $u \in \overset{\circ}{H}_h \Rightarrow$:

$$\begin{aligned} ((B - \frac{1}{2}\tau A)u, u) &= ((D - \tau\Lambda^- + \frac{1}{2}\tau\Lambda)u, u) = (Du, u) + \frac{\tau}{2} \sum_{i=1}^{N_1-1} \sum_{j=1}^{N_2-1} \left[\frac{1}{h_1^2}(u_{i-1,j} + u_{i+1,j}) + \right. \\ &\quad \left. + \frac{1}{h_2^2}(u_{i,j-1} + u_{i,j+1}) - \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right)u_{ij} - \frac{2}{h_1^2}u_{i-1,j} - \frac{2}{h_2^2}u_{i,j-1} \right] u_{ij} = \\ &= \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) \left(1 - \frac{\tau}{2} \right) (u, u) > 0 \text{ при } 0 < \tau < 2. \end{aligned}$$

[\[Перейти к основному тексту\]](#)



Часть IV

Задачи

[Глава 1. Машинная арифметика](#)

[Глава 2. Решение систем линейных алгебраических уравнений](#)

[Глава 3. Решение численных уравнений](#)

[Глава 4. Интерполяция](#)

[Глава 5. Численное интегрирование](#)

[Глава 6. Численные методы математической физики](#)

[Решения и указания](#)

[Ответы к задачам](#)



Глава 1

Машинная арифметика

2. Для указанных значений $\{\beta, p, e_{\min}, e_{\max}\}$ изобразите на числовой прямой соответствующие множества нормализованных и денормализованных чисел с плавающей точкой, вычислите ε_M :
 - а) $\{2, 3, -2, 1\}$,
 - б) $\{3, 2, -1, 1\}$,
 - в) $\{5, 2, -1, 1\}$.
3. Известны три подряд идущих нормализованных машинных числа из некоторой двоичной арифметики с плавающей точкой: 3.75, 4, 4.5. Чему равен ε_M , если округление в арифметике осуществляется до ближайшего машинного числа?
4. Пусть ε_M — машинный эпсилон в некоторой двоичной машинной арифметике с плавающей точкой. Оцените абсолютную погрешность округления $\Delta(x)$ для
 - а) $x = 0.1$,
 - б) $x = \pi$,
 - в) $x = 123.4$,
 - г) $x = 123456.7$.
5. Пусть ε_M — машинный эпсилон в некоторой β -ичной машинной арифметике с плавающей точкой. Оцените абсолютную погрешность округления $\Delta(x)$ для произвольного x .



Глава 2

Решение систем линейных алгебраических уравнений

- 2.1. Метод Гаусса
- 2.2. LU-разложение
- 2.3. Метод квадратного корня
- 2.4. Методы ортогональных преобразований
- 2.5. Итерационные методы



Вверх

Назад

Вперёд

Пред.

След.

Указатель

Помощь

Экран

2.1. Метод Гаусса

6. Выписать расчетную схему метода исключения Гаусса для вычисления определения трехдиагональной матрицы

$$\left(\begin{array}{ccccccc} a_{11} & a_{12} & 0 & 0 & 0 & \dots & 0 \\ a_{21} & a_{22} & a_{23} & 0 & 0 & \dots & 0 \\ 0 & a_{32} & a_{33} & a_{34} & 0 & \dots & 0 \\ 0 & 0 & a_{43} & a_{44} & a_{45} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & a_{n-1,n-2} & a_{n-1,n-1} & a_{n-1,n} \\ 0 & 0 & \dots & 0 & 0 & a_{n,n-1} & a_{n,n} \end{array} \right)$$

Определить вычислительную трудоемкость схемы.

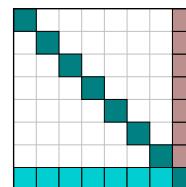
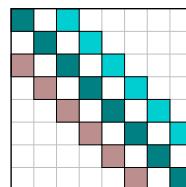
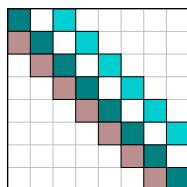
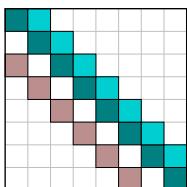
7. Найти методом Гаусса обратные и вычислить число обусловленности в равномерной норме $\|\cdot\|_\infty$ для следующих матриц:

$$1. \begin{pmatrix} 2 & 3 & 3 \\ 1 & 4 & 5 \\ 3 & 7 & 4 \end{pmatrix}, \quad 2. \begin{pmatrix} 3 & 5 & 2 \\ 2 & 3 & 5 \\ 1 & 4 & 3 \end{pmatrix}, \quad 3. \begin{pmatrix} 5 & 2 & -1 \\ 5 & 3 & 4 \\ 2 & -1 & 7 \end{pmatrix}, \quad 4. \begin{pmatrix} 4 & 1 & 1 \\ 3 & -3 & 2 \\ 1 & 4 & 9 \end{pmatrix},$$

$$5. \begin{pmatrix} 6 & 2 & 3 \\ 1 & 3 & -2 \\ 2 & 1 & 9 \end{pmatrix}, \quad 6. \begin{pmatrix} 7 & 1 & 1 \\ 3 & 5 & 0 \\ 1.1 & 0.1 & 3 \end{pmatrix}, \quad 7. \begin{pmatrix} 5 & 2 & 3 \\ 2 & -4 & 1.3 \\ 0.8 & 3 & 1 \end{pmatrix},$$

$$8. \begin{pmatrix} 6 & 0.5 & 2 \\ 3 & 5 & 1.4 \\ 1 & 7 & 3 \end{pmatrix}, \quad 9. \begin{pmatrix} 6 & 0.7 & 1 \\ 1 & 7 & 9 \\ 3 & 0 & 3 \end{pmatrix}, \quad 10. \begin{pmatrix} 3 & 0.4 & 1 \\ 0.4 & 2.5 & 1 \\ 3 & 0 & 8 \end{pmatrix}.$$

8. Составить алгоритм обращения нижнетреугольной матрицы методом Гаусса. Определить вычислительную трудоемкость алгоритма.
9. По аналогии с методом прогонки построить алгоритм решения СЛАУ с матрицей указанной структуры. Матрица задаётся в виде трёх векторов c , d и e . Варианты:



10. Найти решения для следующих систем (с близкими коэффициентами) методом Гаусса. Прокомментировать полученные результаты.

$$1. \begin{cases} x_1 + 2x_2 = 3 \\ x_1 + 2.00001x_2 = 3.00001 \end{cases}$$

$$\text{и} \quad \begin{cases} x_1 + 2x_2 = 3 \\ x_1 + 1.99999x_2 = 3.00001 \end{cases}$$

$$2. \begin{cases} x_1 + 4x_2 = 5 \\ x_1 + 4.00001x_2 = 5.00001 \end{cases}$$

$$\text{и} \quad \begin{cases} x_1 + 4x_2 = 5 \\ x_1 + 3.99999x_2 = 5.00001 \end{cases}$$

$$3. \begin{cases} 2x_1 + 3x_2 = 5 \\ 2x_1 + 3.00001x_2 = 5.00001 \end{cases}$$

$$\text{и} \quad \begin{cases} 2x_1 + 3x_2 = 5 \\ 2x_1 + 2.99999x_2 = 5.00001 \end{cases}$$

11. Рассмотрим плохо обусловленную СЛАУ $Ax = b$. Всегда ли наличие погрешности в векторе b означает большую погрешность в решении? Приведите пример.
12. Исследуйте обусловленность задачи вычисления определителя матрицы размерности 2 относительно погрешности, вносимой в элемент на позиции (1, 1). Приведите примеры хорошо и плохо обусловленной задачи такого типа.

13. Исследуйте обусловленность задачи вычисления определителя произвольной квадратной матрицы A относительно погрешности, вносимой в элемент a_{ij} . Приведите примеры хорошо и плохо обусловленной задачи такого типа.

2.2. LU-разложение

14. Методом LU-разложения решите СЛАУ

$$\text{а) } \left(\begin{array}{ccc|c} 4 & -3 & 0 & 1 \\ 4 & -4 & -2 & 0 \\ 16 & -14 & -1 & 2 \end{array} \right), \quad \text{б) } \left(\begin{array}{ccc|c} 3 & -1 & 4 & 1 \\ -3 & -2 & -9 & -6 \\ -6 & 5 & -4 & 2 \end{array} \right), \quad \text{в) } \left(\begin{array}{ccc|c} 1 & 1 & -1 & 1 \\ -4 & -3 & 3 & -4 \\ -3 & 1 & -3 & -5 \end{array} \right).$$

15. Постройте LU-разложение матрицы вида

$$\left(\begin{array}{ccc} 2 & -1 & & \\ -1 & 2 & -1 & \\ \ddots & \ddots & \ddots & \\ & -1 & 2 & -1 \\ & -1 & 2 & \end{array} \right).$$

16. Данна трёхдиагональная матрица A , как обычно задаваемая тремя векторами c , d и e . Запишите алгоритм построения LU-разложения для такой матрицы, а также алгоритм решения СЛАУ $Ax = b$ с помощью построенного разложения. Оцените сложность алгоритмов.



Вверх

Назад

Вперёд

Пред.

След.

Указатель

Помощь

Экран

2.3. Метод квадратного корня

17. Решите методом квадратного корня СЛАУ

$$\text{а)} \left(\begin{array}{ccc|c} 9 & 3 & 3 & 0 \\ 3 & 5 & -5 & 10 \\ 3 & -5 & 19 & -24 \end{array} \right),$$

$$\text{б)} \left(\begin{array}{ccc|c} 4 & -2 & 8 & 2 \\ -2 & 5 & 2 & 3 \\ 8 & 2 & 34 & 10 \end{array} \right),$$

$$\text{в)} \left(\begin{array}{cccc|c} -9 & -3 & 3 & 3 & 0 \\ -3 & -2 & 3 & 1 & -1 \\ 3 & 3 & 11 & -13 & -10 \\ 3 & 1 & -13 & 24 & 25 \end{array} \right),$$

$$\text{г)} \left(\begin{array}{cccc|c} -4 & 6 & -2 & 8 & 2 \\ 6 & -18 & 6 & -21 & -3 \\ -2 & 6 & 7 & 7 & 1 \\ 8 & -21 & 7 & -9 & 12 \end{array} \right).$$

18. Постройте вычислительный алгоритм решения СЛАУ $Ax = b$ с вещественной симметричной матрицей A , основанный на модифицированном разложении Холецкого: $A = LDL^T$, где L — нижнетреугольная матрица с единицами на главной диагонали, D — диагональная матрица с ненулевыми элементами.
19. Данна симметричная трёхдиагональная матрица A , задаваемая вектором d (главная диагональ) и вектором c (диагональ над и под главной). Запишите алгоритм построения разложения Холецкого $A = R^TDR$ для такой матрицы, а также алгоритм решения СЛАУ $Ax = b$ с помощью построенного разложения. Оцените сложность алгоритмов.



2.4. Методы ортогональных преобразований

20. Методом отражений решите СЛАУ

$$\text{a) } \left(\begin{array}{ccc|c} \frac{2}{3} & 0 & 1 & \frac{5}{3} \\ \frac{2}{3} & -\frac{14}{5} & \frac{13}{5} & \frac{7}{15} \\ \frac{1}{3} & -\frac{2}{5} & \frac{9}{5} & \frac{26}{15} \\ \frac{1}{3} & -\frac{5}{5} & \frac{5}{5} & \frac{15}{15} \end{array} \right),$$

$$\text{б) } \left(\begin{array}{ccc|c} -2 & -\frac{34}{15} & -\frac{14}{3} & -\frac{8}{3} \\ 1 & -\frac{13}{15} & -\frac{8}{3} & -\frac{11}{3} \\ 2 & -\frac{10}{3} & -\frac{1}{3} & -\frac{7}{3} \end{array} \right).$$

21. Методом вращений решите СЛАУ

$$\text{а) } \left(\begin{array}{ccc|c} \frac{1}{\sqrt{2}} & \frac{1}{2} & 1 + \frac{1}{\sqrt{2}} & -1 \\ \frac{1}{\sqrt{2}} & \frac{1}{2} + \sqrt{2} & 1 + \frac{1}{\sqrt{2}} & -1 \\ -1 & -1 + \frac{1}{\sqrt{2}} & -1 + \sqrt{2} & -\sqrt{2} \end{array} \right),$$

$$\text{б) } \left(\begin{array}{ccc|c} 0 & \sqrt{2} & 2\sqrt{2} & \sqrt{2} \\ -1 & -\frac{4}{3} & \frac{11}{3} & 5 \\ 2\sqrt{2} & -\frac{\sqrt{2}}{3} & -\frac{4\sqrt{2}}{3} & -\sqrt{2} \end{array} \right)$$

22. Рассмотрим вектор $w = \frac{1}{\sqrt{n}}(1, \dots, 1)^T \in \mathbb{R}^n$ и матрицу $A = I - 2ww^T$. Найдите A^{2010} .

$$\begin{pmatrix} \frac{\sqrt{3}}{2} & 0 & -\frac{1}{2} \\ 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{\sqrt{3}}{2} \end{pmatrix}$$

23. Рассмотрим матрицу $A = \begin{pmatrix} \frac{\sqrt{3}}{2} & 0 & -\frac{1}{2} \\ 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{\sqrt{3}}{2} \end{pmatrix}$. Найдите A^{2010} .

24. Данна трёхдиагональная матрица A , как обычно задаваемая тремя векторами c , d и e . Запишите алгоритм построения QR -разложения такой матрицы а) методом отражений и б) методом вращений, а также алгоритм решения СЛАУ $Ax = b$ с помощью построенного разложения. Оцените сложность алгоритмов.



2.5. Итерационные методы

25. При каких α, β сходится метод простой итерации $x^{k+1} = Bx^k + c$, где

$$B = \begin{pmatrix} \alpha & \beta & 0 \\ \beta & \alpha & \beta \\ 0 & \beta & \alpha \end{pmatrix}.$$

[\[Ответ\]](#)

26. Пусть матрица в системе $Ax = b$ имеет вид

$$A = \begin{pmatrix} 2 & 0.3 & 0.5 \\ 0.1 & 3 & 0.4 \\ 0.1 & 0.1 & 4.8 \end{pmatrix}.$$

Доказать, что метод простой итерации $x^{k+1} = (E - \tau A)x^k + \tau b$ сходится начиная с любого начального приближения при $0 < \tau < 2/5$.

27. При каких значениях параметра τ метод

$$x^{k+1} = (E - \tau A)x^k + \tau b$$

для системы уравнений $Ax = b$ с матрицей:

$$1) A = \begin{pmatrix} 5 & 0.8 & 4 \\ 2.5 & 2 & 0 \\ 2 & 0.8 & 4 \end{pmatrix}; \quad 2) A = \begin{pmatrix} 2 & 1 & 0.5 \\ 3 & 5 & 1 \\ 1 & 3 & 3 \end{pmatrix};$$



$$3) A = \begin{pmatrix} 1 & 0.5 & 0.3 \\ 1 & 3 & 0 \\ 1 & 1 & 2 \end{pmatrix}; \quad 4) A = \begin{pmatrix} 3 & 1.2 & 0.8 \\ 1.4 & 2 & 0.1 \\ 0.6 & 0.4 & 1 \end{pmatrix}$$

сходится с произвольного начального приближения?

28. Найти области сходимости методов Якоби и Гаусса - Зейделя для систем с матрицами вида

$$A = \begin{pmatrix} \alpha & \beta & 0 \\ \beta & \alpha & \beta \\ 0 & \beta & \alpha \end{pmatrix}.$$

[\[Решение\]](#)

29. Доказать, что для систем линейных уравнений второго порядка ($n = 2$) методы Якоби и Гаусса - Зейделя сходятся и расходятся одновременно.

[\[Решение\]](#)

30. Система $Ax = b$ с матрицей $A = \begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix}$ решается методом Гаусса - Зейделя. Доказать, что:

- 1) если $|a| > 1$, то для некоторого начального приближения итерационный процесс расходится;
- 2) если $|a| < 1$, то итерации сходятся при любом начальном приближении.

31. Найти α, β , при которых метод Гаусса - Зейделя будет сходящимся для систем уравнений с матрицами:

$$\begin{pmatrix} \alpha & 0 & \beta \\ 0 & \alpha & 0 \\ \beta & 0 & \alpha \end{pmatrix}; \quad \begin{pmatrix} \alpha & \beta & 0 \\ \beta & \alpha & 0 \\ 0 & 0 & \alpha \end{pmatrix}; \quad \begin{pmatrix} \alpha & \alpha & 0 \\ \alpha & \beta & \beta \\ 0 & \beta & \alpha \end{pmatrix}.$$



Глава 3

Решение численных уравнений

- 3.1. Метод бисекции
- 3.2. Метод Ньютона



3.1. Метод бисекции

32. Найти приближение x_3 по методу бисекции для функции $f(x) = \sqrt{x} - \cos(x)$ на отрезке $[0, 1]$.
33. Найти приближение x_3 по методу бисекции для функции $f(x) = 3(x + 1)(x - 0.5)(x - 1)$ на интервалах
- 1) $[-2, 1.5]$
 - 2) $[-1.25, 2.5]$.
34. Используя метод бисекции найти корни уравнения $x^3 - 7x^2 + 14x - 6 = 0$ с точностью 10^{-2} на отрезках
- 1) $[0, 1]$
 - 2) $[1, 3.2]$
 - 3) $[3.2, 4]$.
35. Используя метод бисекции найти корни уравнения $x^4 - 2x^3 - 4x^2 + 4x + 4 = 0$ с точностью 10^{-2} на отрезках
- 1) $[-2, -1]$
 - 2) $[0, 2]$
 - 3) $[2, 3]$
 - 4) $[-1, 0]$.
36. Используя метод бисекции найти решение уравнения $x = \tan x$ с точностью 10^{-3} на отрезке $[4, 4.5]$.
37. Используя метод бисекции найти решение уравнения $2 + \cos(e^x - 2) - e^x = 0$ с точностью 10^{-3} на отрезке $[0.5, 1.5]$.



38. Используя метод бисекции найти решения следующих уравнений с точностью 10^{-5}

- 1) $x - 2^{-x} = 0$ $x \in [0, 1]$
- 2) $e^x - x^2 + 3x - 2 = 0$ $x \in [0, 1]$.
- 3) $2x \cos(2x) - (x + 1)^2 = 0$ $x \in [-3, -2]$, $x \in [-1, 0]$
- 4) $x \cos(x) - 2x^2 + 3x - 1 = 0$ $x \in [0.2, 0.3]$, $x \in [1.2, 1.3]$

39. Пусть $f(x) = (x + 2)(x + 1)^2 x(x - 1)^3(x - 2)$. Определить корни, к которым сходится метод бисекции на следующих интервалах

- 1) $[-1.5, 2.5]$
- 2) $[-0.5, 2.4]$
- 3) $[-0.5, 3]$
- 4) $[-3, -0.5]$.

40. Пусть $f(x) = (x + 2)(x + 1)^2 x(x - 1)^3(x - 2)$. Определить корни, к которым сходится метод бисекции на следующих интервалах

- 1) $[-3, 2.5]$
- 2) $[-2.5, 3]$
- 3) $[-1.75, 1.5]$
- 4) $[-1.5, 1.75]$.

41. Найти приближение $\sqrt{3}$ с точностью 10^{-4} используя метод бисекции.

42. Найти приближение $\sqrt[3]{25}$ с точностью 10^{-4} используя метод бисекции.



43. Найти корень уравнения $x^3 + x - 4 = 0$ на отрезке $[1, 4]$ с точностью 10^{-3} . Используя оценку погрешности $|x_n - x| \leq 2^{-n}(b - a)$, $n \geq 1$ найти значение для достижения указанной точности. Здесь $[a, b]$ интервал содержащий корень.
44. Пусть $f(x) = (x - 1)^{10}$ и $x_n = 1 + \frac{1}{n}$. Показать, что $|f(x_n)| < 10^{-3}$ при $n > 1$, но $|x - x_n| < 10^{-3}$ только при $n > 1000$.
45. Показать, что последовательность $x_n = \sum_{k=1}^n \frac{1}{k}$ расходится, хотя $\lim_{n \rightarrow \infty} (x_n - x_{n-1}) = 0$.
46. Пусть $f(x) = \sin(\pi x)$. Показать, что при $-1 < a < 0$ и $2 < b < 3$ метод бисекции сходится
- 1) к 2, если $a + b > 2$
 - 2) к 0, если $a + b < 2$ с) к 1, если $a + b = 2$.



3.2. Метод Ньютона

47. Пусть $f(x) = x^2 - 6$ и $x_0 = 1$. Используя метод Ньютона найти x_2 .
48. Пусть $f(x) = -x^3 - \cos x$ и $x_0 = -1$. Используя метод Ньютона найти x_2 . Можно ли использовать значение $x_0 = 0$?
49. Пусть $f(x) = x^2 - 6$, $x_0 = 3$ и $x_1 = 2$. Найти x_3 методом секущих.
50. Пусть $f(x) = -x^3 - \cos x$, $x_0 = -1$ и $x_1 = 0$. Найти x_3 методом секущих.
51. Используя метод Ньютона найти корни следующих уравнений с точностью 10^{-4}
- 1) $x^3 - 2x^2 - 5 = 0$ на отрезке $[1, 4]$
 - 2) $x^3 + 3x^2 - 1 = 0$ на отрезке $[-3, -2]$
 - 3) $x - \cos x = 0$ на отрезке $[0, \pi/2]$
 - 4) $x - 0.8 - 0.2 \sin x = 0$ на отрезке $[0, \pi/2]$
52. Используя метод Ньютона найти корни следующих уравнений с точностью 10^{-5}
- 1) $e^x + 2^{-x} + 2 \cos x - 6 = 0$ на отрезке $[1, 2]$
 - 2) $\ln(x-1) + \cos(x-1) = 0$ на отрезке $[1.3, 2]$
 - 3) $2x \cos 2x - (x-2)^2 = 0$ на отрезках $[2, 3]$ и $[3, 4]$
 - 4) $(x-2)^2 - \ln x = 0$ на отрезках $[1, 2]$ и $[e, 4]$
 - 5) $e^x - 3x^2 = 0$ на отрезках $[0, 1]$ и $[3, 5]$
 - 6) $\sin x - e^{-x} = 0$ на отрезках $[0, 1]$, $[3, 4]$ и $[6, 7]$



53. Используя метод секущих найти корни уравнений из пункта 6 с точностью 10^{-5}
54. Используя метод Ньютона решить уравнение $\frac{1}{2} + \frac{1}{4}x^2 - x \sin x - \frac{1}{2} \cos 2x = 0$ с точностью 10^{-5} для начальных приближений $x_0 = \pi/2$, $x_0 = 5\pi$ и $x_0 = 10\pi$.
55. Функция $f(x) = 230x^4 + 18x^3 + 9x^2 - 221x - 9$ имеет два вещественных корня на отрезках $[-1, 0]$ и $[0, 1]$. Найти эти корни с точностью 10^{-6}
- 1) методом Ньютона, выбирая в качестве нулевого приближения центр отрезка
 - 2) методом секущих, выбирая в качестве нулевого приближения правый конец отрезка
56. Уравнение $x^2 - 10 \cos x = 0$ имеет два корня ± 1.3793646 . Найти корни методом Ньютона с точностью 10^{-5} и нулевыми приближениями $x_0 = -100$, $x_0 = -50$, $x_0 = -25$, $x_0 = 25$, $x_0 = 50$, $x_0 = 100$.



Глава 4

Интерполяция

- 4.1. Понятие интерполяции
- 4.2. Интерполяционный многочлен в форме Лагранжа
- 4.3. Интерполяционный многочлен в форме Ньютона

4.1. Понятие интерполяции

57. Выписать матрицу для определения коэффициентов многочлена вида

$$P_3(x) = c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1) + c_3(x - x_0)(x - x_1)(x - x_2),$$

58. Построить интерполяционный кубический многочлен

$$P_3(x) = c_0 + c_1x + c_2x^2 + c_3x^3,$$

для которого выполнено $P_3(2) = 1, P_3(3) = 2, P_3(4) = 2, c_2 = 2$.

59. Пусть известны значения функции $f(x)$ в узлах x_1, x_2 . Построить интерполирующую функцию вида $\varphi(x) = a/(x^2 + b)$.

60. Пусть известны значения функции $f(x)$ в узлах x_1, x_2 . Построить интерполирующую функцию вида $\varphi(x) = (a - x)/(x + b)$.

61. Пусть известны значения функции $f(x)$ в узлах x_1, x_2, x_3 . Построить интерполирующую функцию вида $\varphi(x) = (a + bx)/(x + c)$.



4.2. Интерполяционный многочлен в форме Лагранжа

62. Пусть дана таблица значений функции $f(x) = e^{-x}$:

x_i	1.60	1.62	1.63	1.65	1.67
$f(x_i)$	0,2019	0,1979	0,1959	0,1920	0,1882

Требуется:

- a) вычислить значение $f(1.61)$ с помощью линейной интерполяционной формулы Лагранжа, $L_1(1.61)$;
- б) оценить погрешность;
- в) выписать интерполяционную формулу Лагранжа второй степени для вычисления $f(1.61)$.

63. Пусть известна верхняя граница $C = |f^{(n+1)}(x)|$ для функции $f(x)$ на отрезке $[0, 1]$, $x \in [0, 1]$. Оценить возможную и максимальную погрешности при интерполяции многочленом Лагранжа первой степени.

64. Для функции заданной таблично вычислить с помощью многочлена Лагранжа значение функции в заданной точке $(\xi, f(\xi))$.

x_i	1.62	1.64	1.65	1.67	1.68
$f(x_i)$	1.172	1.179	1.182	1.186	1,189

$$f(\xi) = f(1.63)$$

x_i	1.84	1.85	1.87	1.89	1.91
$f(x_i)$	1.225	1.228	1.232	1.236	1,241

$$f(\xi) = f(1.86)$$

x_i	2.23	2.26	2.27	2.29	2.32
$f(x_i)$	1.306	1.3128	1.314	1.318	1,324



$$f(\xi) = f(2.25)$$

4)	x_i	1.60	1.62	1.63	1.65	1.67
	$f(x_i)$	0.2019	0.1979	0.1959	0.1920	0,1882

$$f(\xi) = f(1.64)$$

65. Известен набор узлов $x_0, x_1, x_2, x_3, x_4, x_5$ и значения функции $f(x)$ в этих узлах. Построить интерполяционный многочлен Лагранжа для узлов
а) x_3, x_4 ; б) x_1, x_2, x_3 ; в) x_3, x_4, x_5 .
66. Построить интерполяционный многочлен Лагранжа $L_2(x)$, для которого выполнено
 $L_2(2) = 1, L_2(3) = 2, L_2(4) = 3$.



4.3. Интерполяционный многочлен в форме Ньютона

67. Выписать интерполяционные многочлены Ньютона первой, второй и третьей степени.

68. Пусть дана таблица значений функции $f(x) = e^{-x}$.

x_i	2.80	2.83	2.84	2.86	2.89
$f(x_i)$	0,06081	0,05901	0,05843	0,05727	0,05558

Требуется:

- 1) вычислить значение $f(2.81)$ с помощью линейной интерполяционной формулы Ньютона, $N_1(2.81)$;
- 2) оценить погрешность;
- 3) выписать интерполяционную формулу Ньютона второй степени для вычисления $f(2.81)$.

69. Функция $f(x)$ задана таблицей

x_i	1	3	5	7	8
$f(x_i)$	3	?	2	3	1

Требуется с помощью интерполяционной формулы Ньютона восстановить значение при $x = 3$.

70. Построить интерполяционный многочлен Ньютона для функции $f(x) = 2|x|$ по значениям $-0.5, 0, 0.3$.

71. Построить интерполяционный многочлен Ньютона для функции $f(x) = x^2$ по значениям $-1, -0.2, 0, 0.3, 0.8$.

72. Оценить величину шага интерполяции h для обеспечения точности $\varepsilon \leq 10^{-4}$ интерполяции функции $f(x) = \sqrt{x}$, $x \in [1, 10^3]$ при линейной и квадратичной интерполяции.

73. Оценить погрешность интерполяции функции $f(x) = \sin x$ на отрезке $[0, \pi/4]$ по трем равноотстоящим точкам.



74. Для функции заданной таблично вычислить с помощью многочлена Ньютона значение функции в заданной точке $(\xi, f(\xi))$.

1)	x_i	7.75	7.77	7.79	7.80	7.82
	$f(x_i)$	1.979	1.981	1.982	1.983	1.985

$$f(\xi) = f(7.78)$$

2)	x_i	8.00	8.02	8.04	8.07	8.08
	$f(x_i)$	2.000	2.002	2.003	2.006	2.007

$$f(\xi) = f(8.01)$$

3)	x_i	7.64	7.66	7.67	7.69	7.71
	$f(x_i)$	1.970	1.971	1.972	1.974	1.976

$$f(\xi) = f(7.65)$$

4)	x_i	3.89	3.91	3.92	3.95	3.96
	$f(x_i)$	1.573	1.575	1.577	1.581	1.582

$$f(\xi) = f(3.94)$$

5)	x_i	4.05	4.07	4.08	4.10	4.12
	$f(x_i)$	1.594	1.597	1.598	1.601	1.603

$$f(\xi) = f(4.06)$$

6)	x_i	2.86	2.88	2.89	2.92	2.95
	$f(x_i)$	1.419	1.423	1.424	1.429	1.434



Вверх

Назад

Вперёд

Пред.

След.

Указатель

Помощь

Экран

$$f(\xi) = f(2.87)$$



Глава 5

Численное интегрирование

- 5.1. Простейшие квадратурные формулы
- 5.2. Квадратурные формулы НАСТ
- 5.3. Практическая оценка погрешности



5.1. Простейшие квадратурные формулы

75. Найти оценку погрешности вычисления интеграла

$$\int_0^1 f(x)dx \quad \text{при} \quad f(x) = \frac{1}{1+x^2},$$

по составной квадратурной формуле

$$S(f) = [f(0) + 4f(0.1) + 2f(0.2) + 4f(0.3) + \dots + 4f(0.9) + f(1.0)]/30.$$

[[Ответ](#)]

76. Найти оценку погрешности вычисления интеграла

$$\int_0^1 f(x)dx \quad \text{при} \quad f(x) = \frac{1}{1+x^2},$$

по составной квадратурной формуле

$$S(f) = [f(0) + 2f(0.1) + 2f(0.2) + \dots + 2f(0.9) + f(1.0)]/20.$$

[[Ответ](#)]

77. Оценить минимальное число разбиений отрезка N для вычисления интеграла $\int_0^1 \sin(x^2)dx$ по составной квадратурной формуле трапеций, обеспечивающее точность 10^{-4} .

[[Ответ](#)]



5.2. Квадратурные формулы НАСТ

78. Построить квадратурную формулу вида

$$\int_0^1 f(x)dx \approx c_1 f(0) + c_2 f(2/3),$$

точную для многочленов максимально возможной степени.

79. Построить квадратурную формулу вида

$$\int_0^1 f(x)dx \approx c_1 f(1/2) + c_2 f(2/3),$$

точную для многочленов максимально возможной степени.

80. Для вычисления интеграла $\int_{-1}^1 x^2 f(x)dx$ построить квадратурную формулу вида $S(f) = c_1 f(-1) + c_2 f(x_2) + c_3 f(1)$, точную для многочленов максимально высокой степени.

81. Построить квадратуру Гаусса с одним узлом для вычисления интеграла:

$$1) I(f) = \int_0^1 x f(x)dx;$$

[Ответ]

$$2) I(f) = \int_0^1 e^x f(x)dx.$$

[Ответ]



82. Построить квадратуру Гаусса с двумя узлами для вычисления интеграла:

$$1) I(f) = \int_{-1}^1 x^2 f(x) dx; \quad 2) I(f) = \int_{-\pi/2}^{\pi/2} \cos x f(x) dx.$$

Ответ:

$$1) \frac{1}{3} \left(f\left(\sqrt{\frac{3}{5}}\right) + f\left(\sqrt{-\frac{3}{5}}\right) \right);$$
$$2) f\left(\sqrt{\frac{\pi^2}{4} - 2}\right) + f\left(-\sqrt{\frac{\pi^2}{4} - 2}\right).$$

83. Построить квадратуру Гаусса с тремя узлами для вычисления интеграла:

$$I(f) = \int_{-1}^1 f(x) dx.$$

[Ответ]

84. Построить квадратурную формулу Гаусса с двумя узлами для вычисления интегралов вида

$$\int_0^\pi \sin(x) f(x) dx.$$

[Ответ]

85.

[Ответ]



Построить квадратурную формулу Гаусса с двумя узлами для вычисления интегралов вида

$$\int_0^{\infty} \exp(-x) f(x) dx.$$



5.3. Практическая оценка погрешности

86. Пусть S_h — приближенное значение интеграла, вычисленное по некоторой квадратурной формуле на отрезке длины h , $S_{h/2}$ — значение, вычисленное путем применения той же формулы на двух отрезках длины $h/2$ (составная формула). Используя S_h и $S_{h/2}$, построить квадратурную формулу более высокого порядка.

[Ответ]

87. Пусть M_1, M_2 — количества узлов одной и той же квадратурной формулы S_M с порядком точности $O(1/M^2)$.

1) Доказать справедливость приближенного равенства:

$$I(f) - S_{M_1}(f) \approx \frac{1}{3}(S_{M_2}(f) - S_{M_1}(f)).$$

2) Доказать, что выражение $S_{M_1}(f) + \frac{1}{3}(S_{M_2}(f) - S_{M_1}(f))$ совпадает с квадратурной формулой Симпсона.



Глава 6

Численные методы математической физики

- 6.1. Аппроксимация дифференциальных задач разностными схемами
- 6.2. Исследование устойчивости разностных схем
- 6.3. Реализация разностных схем для уравнений теплопроводности и колебаний струны



6.1. Аппроксимация дифференциальных задач разностными схемами

88. Построить разностную аппроксимацию оператора L методом неопределенных коэффициентов.

Оператор	Шаблон	
1) $Lu(x_1, x_2) = \frac{\partial^2 u(x_1, x_2)}{\partial x_1 \partial x_2}$		[Решение]
2) $Lu(x_0) = u''(x_0)$		
3) $Lu(x_1) = u'(x_1)$		



Оператор	Шаблон
4) $L u(x_2) = u'(x_2)$	

Оператор	Шаблон
5) $L u(x_0) = u''(x_0)$	

Оператор	Шаблон
6) $L u(x_1) = u'(x_1)$	

Оператор	Шаблон
7) $L u(x_2) = u''(x_2)$	



Оператор	Шаблон
8) $L u(x_3) = u'''(x_3)$	

Оператор	Шаблон
9) $L u(x_2) = u'''(x_2)$	

Оператор	Шаблон
10) $L u(x_1, x_2) = \frac{\partial}{\partial}$	

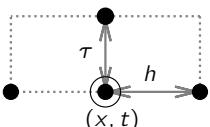
Оператор	Шаблон
11) $L u(x_1, x_2) = \frac{\partial}{\partial}$	



89. Данна линейная дифференциальная задача.

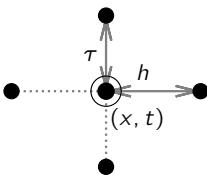
- a) Аппроксимировать задачу разностной схемой на заданном шаблоне. Определить погрешность аппроксимации.
- б) Повысить порядок аппроксимации построенной схемы на минимальном шаблоне, используя вид дифференциальной задачи.

$$1) \begin{cases} \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x, t), & 0 < x < 1, \quad t > 0 \\ u(x, 0) = u_0(x), & 0 \leq x \leq 1, \\ \frac{\partial u(0, t)}{\partial x} = \sigma_0(t)u(0, t) - \mu_0(t), & t \geq 0, \\ u(1, t) = \mu_1(t), & t \geq 0. \end{cases}$$

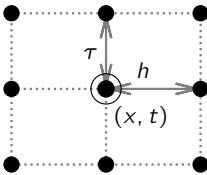


[Решение]

$$2) \begin{cases} \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + f(x, t), & 0 < x < 1, \quad t > 0, \\ u(x, 0) = u_0(x), \quad \frac{\partial u}{\partial t}(x, 0) = u_1(x), & 0 \leq x \leq 1, \\ u(0, t) = \mu_0(t), \quad u(1, t) = \mu_1(t), & t \geq 0. \end{cases}$$

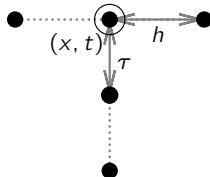


$$3) \begin{cases} \frac{\partial u}{\partial t} = \frac{\partial}{\partial x} \left(k(x, t) \frac{\partial u}{\partial x} \right) + f(x, t), & 0 < x < 1, \quad t > 0, \\ u(x, 0) = u_0(x), & 0 \leq x \leq 1, \\ \frac{\partial u(x, 0)}{\partial x} = \mu_0(t), \quad u(1, t) = \mu_1(t), & t \geq 0. \end{cases}$$

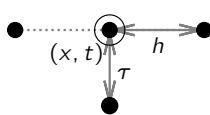




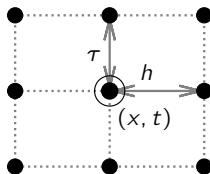
4)
$$\begin{cases} \frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x} (k(x, t) \frac{\partial u}{\partial x}) + f(x, t), & 0 < x < 1, t > 0, \\ u(x, 0) = u_0(x), \quad \frac{\partial u(x, 0)}{\partial t} = u_1(x), & 0 \leq x \leq 1, \\ u(0, t) = \mu_0(t), \quad u(1, t) = \mu_1(t), & t \geq 0. \end{cases}$$



5)
$$\begin{cases} \frac{\partial u}{\partial t} = \frac{\partial}{\partial x} (k(x, t) \frac{\partial u}{\partial x}) + f(x, t), & 0 < x < 1, t > 0, \\ u(x, 0) = u_0(x), & 0 \leq x \leq 1, \\ u(0, t) = \mu_0(t), \quad \frac{\partial u(1, t)}{\partial x} = \mu_1(t), & t \geq 0. \end{cases}$$

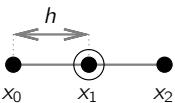


6)
$$\begin{cases} \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + q(x, t)u + f(x, t), & 0 < x < 1, t > 0, \\ u(x, 0) = u_0(x), \quad \frac{\partial u(x, 0)}{\partial t} = u_1(x), & 0 \leq x \leq 1, \\ \frac{\partial u(0, t)}{\partial x} = \sigma_0 u(0, t) - \mu_0(t), \quad u(1, t) = \mu_1(t), & t \geq 0. \end{cases}$$



7)
$$\begin{cases} \frac{\partial u}{\partial t} = \frac{\partial}{\partial x} (k(x, t) \frac{\partial u}{\partial x}) + f(x, t), & 0 < x < 1, t > 0, \\ u(x, 0) = u_0(x), & 0 \leq x \leq 1, \\ u(0, t) = \mu_0(t), \quad \frac{\partial u(1, t)}{\partial x} = \sigma_1 u(1, t) - \mu_1(t), & t \geq 0. \end{cases}$$

8)
$$\begin{cases} u'' + \kappa u' + qu = -f(x), & 0 < x < 1, \\ u'(0) = \sigma_1 u(0) + \mu_1, \quad \kappa = const, \\ -u'(1) = \sigma_2 u(1) + \mu_2, \quad q = const. \end{cases}$$





$$9) \quad \begin{cases} \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + f(x, t), \\ u(x, 0) = u_0(x), \quad \frac{\partial u(x, 0)}{\partial t} = u_1(x), \\ u(0, t) = \mu_1(t), \quad \frac{\partial u(1, t)}{\partial x} = \sigma_2 u(1, t) + \mu_2(t). \end{cases}$$

$$10) \quad \begin{cases} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = -f(x, y), \\ \left. \frac{\partial u(x, y)}{\partial x} \right|_{\Gamma} = \mu(x, y) \end{cases}$$



6.2. Исследование устойчивости разностных схем

90. Исследовать устойчивость разностной схемы с весами

$$\begin{cases} y_t + a[\sigma \hat{y}_x + (1 - \sigma)y_x] = f(x, t + \frac{\tau}{2}), & x \in \omega_h, t \in \omega_\tau; \\ y(x, 0) = u^0(x), & x \in \omega_h; \\ y(0, t) = \mu(t), & t \in \omega_\tau. \end{cases} \quad (6.1)$$

для задачи

$$\begin{cases} \frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = f(x, t), & 0 < x < l, 0 < t \leq T, a = \text{const} \neq 0; \\ u(x, 0) = u^0(x), & 0 \leq x \leq l; \\ u(0, t) = \mu(t), & 0 \leq t \leq T, \end{cases} \quad (6.2)$$

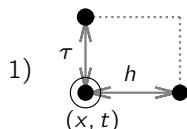
спектральным методом.

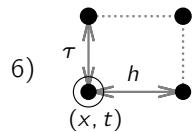
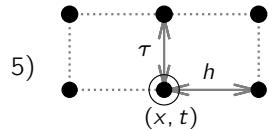
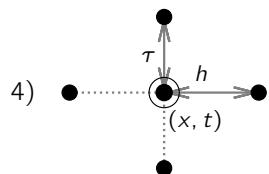
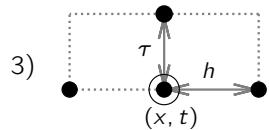
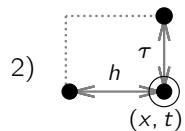
[\[Решение\]](#)

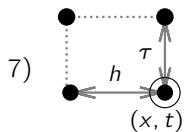
91. Исследовать устойчивость разностной схемы с весами (6.1) для задачи (6.2) с помощью принципа максимума.

[\[Решение\]](#)

92. Для задачи (6.2) построить разностную схему на заданном шаблоне и исследовать ее устойчивость а) спектральным методом и б) с помощью принципа максимума.



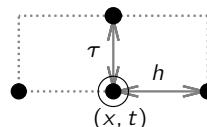




6.3. Реализация разностных схем для уравнений теплопроводности и колебаний струны

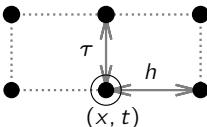
93. Для данной дифференциальной задачи на указанном шаблоне построить разностную схему, найти её порядок аппроксимации, исследовать устойчивость и записать алгоритм машинной реализации.

$$1) \quad \begin{cases} \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x, t), & 0 < x < 1, \quad t > 0, \\ u(x, 0) = u_0(x), & 0 \leq x \leq 1, \\ \frac{\partial u(0, t)}{\partial x} = \sigma_1 u(0, t) + \mu_1(t), & t \geq 0, \quad \sigma_1 > 0, \\ -\frac{\partial u(1, t)}{\partial x} = \sigma_2 u(1, t) + \mu_2(t), & t \geq 0, \quad \sigma_2 > 0. \end{cases} \quad (6.3)$$



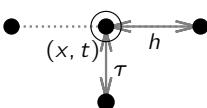
[Решение]

- 2) Третья краевая задача для уравнения теплопроводности (6.3)



[Решение]

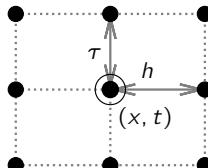
- 3) Третья краевая задача для уравнения теплопроводности (6.3)



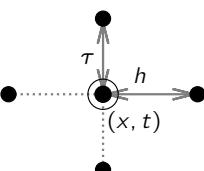


4)

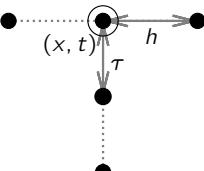
$$\begin{cases} \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + f(x, t) \\ u(x, 0) = u^0(x), \quad \frac{\partial u(x, 0)}{\partial t} = \bar{u}^0(x), \quad 0 \leq x \leq 1, \\ u(0, t) = \mu_0(t), \quad u(1, t) = \mu_1(t), \quad t \geq 0 \end{cases} \quad (6.4)$$



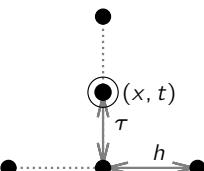
5) Краевая задача для волнового уравнения (6.4).



6) Краевая задача для волнового уравнения (6.4).



7) Краевая задача для волнового уравнения (6.4).





Решения и указания



Часть IV. Задачи

Решения и указания

Глава 2. Решение систем линейных алгебраических уравнений

Меню



Вверх

Назад

Вперёд

Пред.

След.

Указатель
Помощь

Экран

Глава 2. Решение систем линейных алгебраических уравнений



Решение задачи 28

Оператор перехода B в методе Якоби имеет вид $B = -D^{-1}(L + R)$. Рассмотрим задачу на собственные значения $Bx = \lambda x$. Имеем

$$-D^{-1}(L + R)x = \lambda x \Rightarrow (L + \lambda D + R)x = 0 \Rightarrow \det(L + \lambda D + R) = 0.$$

Непосредственные вычисления дают

$$\det \begin{pmatrix} \alpha\lambda & \beta & 0 \\ \beta & \alpha\lambda & \beta \\ 0 & \beta & \alpha\lambda \end{pmatrix} = \alpha\lambda(\alpha^2\lambda^2 - 2\beta^2) = 0.$$

Следовательно,

$$\lambda_1 = 0, \quad \lambda_{2,3}^2 = 2\frac{\beta^2}{\alpha^2} \Rightarrow \left| \frac{\beta_0}{\alpha} \right| < \frac{1}{\sqrt{2}}.$$

Оператор перехода B в методе Зейделя имеет вид $B = -(D + L)^{-1}R$. Рассмотрим задачу на собственные значения $Bx = \lambda x$. Имеем

$$-(D + L)^{-1}Rx = \lambda x \Rightarrow (\lambda L + \lambda D + R)x = 0 \Rightarrow \det(\lambda L + \lambda D + R) = 0.$$

Непосредственные вычисления дают

$$\det \begin{pmatrix} \alpha\lambda & \beta & 0 \\ \beta\lambda & \alpha\lambda & \beta \\ 0 & \beta\lambda & \alpha\lambda \end{pmatrix} = \alpha\lambda^2(\alpha^2\lambda - 2\beta^2) = 0.$$

Следовательно,

$$\lambda_{1,2} = 0, \quad \lambda_3 = 2\frac{\beta^2}{\alpha^2} \Rightarrow \left| \frac{\beta}{\alpha} \right| < \frac{1}{\sqrt{2}}.$$

В данном случае области сходимости методов совпадают.

[\[Вернуться к условию\]](#)



Решение задачи 29

Искомый результат следует из явного представления операторов перехода

$$B_{\text{я}} = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 \end{pmatrix} \Rightarrow \lambda_{1,2}^{\text{я}} = \pm \sqrt{\frac{a_{12}a_{21}}{a_{11}a_{22}}},$$

$$B_3 = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} \\ 0 & \frac{a_{12}a_{21}}{a_{11}a_{22}} \end{pmatrix} \Rightarrow \lambda_1^3 = 0, \lambda_2^3 = \frac{a_{12}a_{21}}{a_{11}a_{22}}.$$

[[Вернуться к условию](#)]



Часть IV. Задачи

Решения и указания

Глава 6. Численные методы математической физики

Меню



Вверх

Назад

Вперёд

Пред.

След.

Указатель Помощь Экран

Глава 6. Численные методы математической физики



Решение задачи 88.1

Искомую разностную аппроксимацию строим в виде линейной комбинации значений функции в узлах шаблона:

$$\begin{aligned} L_h u(x_1, x_2) = & a_0 u(x_1, x_2) + a_1 u(x_1 - h_1, x_2 + h_2) + a_2 u(x_1 - h_1, x_2 - h_2) + a_3 u(x_1 + h_1, x_2 + h_2) + \\ & + a_4 u(x_1 + h_1, x_2 - h_2). \end{aligned}$$

Для определения коэффициентов составим разность разностного и дифференциального операторов в точке аппроксимации и проведем разложение этой разности в ряд по степеням параметров h_1 и h_2 , используя формулу Тейлора для функции двух переменных:

$$\begin{aligned} u(x, y) = & u(x_0, y_0) + [\frac{\partial}{\partial x}(x - x_0) + \frac{\partial}{\partial y}(y - y_0)]u(x_0, y_0) + \\ & + \frac{1}{2!}[\frac{\partial}{\partial x}(x - x_0) + \frac{\partial}{\partial y}(y - y_0)]^2u(x_0, y_0) + \dots \end{aligned}$$

В результате получим:

$$\begin{aligned} \psi(x_1, x_2) = & L_h u(x_1, x_2) - L u(x_1, x_2) = (a_0 + a_1 + a_2 + a_3 + a_4)u(x_1, x_2) + \\ & + h_1(a_3 + a_4 - a_1 - a_2)\frac{\partial u(x_1, x_2)}{\partial x_1} + h_2(a_1 + a_3 - a_2 - a_4)\frac{\partial u(x_1, x_2)}{\partial x_2} + \\ & + \frac{h_1^2}{2}(a_1 + a_2 + a_3 + a_4)\frac{\partial^2 u(x_1, x_2)}{\partial x_1^2} + \frac{h_2^2}{2}(a_1 + a_2 + a_3 + a_4)\frac{\partial^2 u(x_1, x_2)}{\partial x_2^2} + \\ & + [h_1 h_2(a_2 + a_3 - a_1 - a_4) - 1]\frac{\partial^2 u(x_1, x_2)}{\partial x_1 \partial x_2} + \frac{h_1^3}{6}(a_3 + a_4 - a_1 - a_2)\frac{\partial^3 u(x_1, x_2)}{\partial x_1^3} + \\ & + \frac{h_2^3}{6}(a_1 + a_3 - a_2 - a_4)\frac{\partial^3 u(x_1, x_2)}{\partial x_2^3} + \frac{h_1^2 h_2}{2}(a_1 - a_2 + a_3 - a_4)\frac{\partial^3 u(x_1, x_2)}{\partial x_1^2 \partial x_2} + \\ & + \frac{h_1 h_2^2}{2}(-a_1 - a_2 + a_3 + a_4)\frac{\partial^3 u(x_1, x_2)}{\partial x_1 \partial x_2^2} + \dots \end{aligned}$$



Потребуем обращения в нуль первых членов разложения, приравнивая к нулю коэффициенты при производных. Получим систему линейных алгебраических уравнений. Количество линейно независимых уравнений берем равным числу параметров a_i в $L_h u(x_1, x_2)$:

$$\left\{ \begin{array}{l} a_0 + a_1 + a_2 + a_3 + a_4 = 0 \\ a_3 + a_4 - a_1 - a_2 = 0 \\ a_1 + a_3 - a_2 - a_4 = 0 \\ a_1 + a_2 + a_3 + a_4 = 0 \\ a_2 + a_3 - a_1 - a_4 = \frac{1}{h_1 h_2} \end{array} \right. \Rightarrow a_0 = 0; a_1 = a_4 = -\frac{1}{4h_1 h_2}; a_2 = a_3 = \frac{1}{4h_1 h_2}.$$

Таким образом, разностный оператор имеет вид:

$$L_h u(x_1, x_2) = \frac{u(x_1 + h_1, x_2 + h_2) - u(x_1 - h_1, x_2 + h_2) + u(x_1 + h_1, x_2 - h_2) - u(x_1 - h_1, x_2 - h_2)}{4h_1 h_2}.$$

[\[Вернуться к условию\]](#)



Решение задачи 89.1

Пункт а). Строим сетку $\bar{\omega}_{h\tau}$ и на данном шаблоне заменим дифференциальные производные разностными; получим следующую разностную схему:

$$\begin{cases} y_t = y_{\bar{x}\bar{x}} + f(x, t), & (x, t) \in \omega_{h\tau}, \\ y(x, 0) = u_0(x), & x \in \bar{\omega}_h, \\ y_x(0, t) = \sigma_0(t)y(0, t) - \mu_0(t), & t \in \omega_\tau, \\ y(1, t) = \mu_1(t), & t \in \omega_\tau. \end{cases}$$

Определим погрешность аппроксимации дифференциального уравнения разностным как невязку разностного уравнения на точном решении $u(x, t)$:

$$\psi(x, t) = u_t - u_{\bar{x}\bar{x}} - f(x, t) = \frac{u(x, t + \tau) - u(x, t)}{\tau} - f(x, t) - \frac{u(x + h, t) - 2u(x, t) + u(x - h, t)}{h^2},$$

и, выполняя разложение в ряд Тейлора, получим:

$$\psi(x, t) = \frac{\tau}{2} \frac{\partial^2 u(x, t)}{\partial t^2} - \frac{h^2}{12} \frac{\partial^4 u(x, t)}{\partial x^4} + O(\tau^2 + h^4) = O(\tau + h^2),$$

так что $\|\psi\|_{h\tau} = O(\tau + h^2)$. Начальное условие аппроксимируем точно. Для граничного условия на левом конце определим погрешность аппроксимации как невязку разностного краевого условия:

$$\begin{aligned} \nu(0, t) &= u_x(0, t) - \sigma_0(t)u(0, t) + \mu_0(t) = \frac{u(h, t) - u(0, t)}{h} - \sigma_0(t)u(0, t) + \mu_0(t) = \frac{\partial u(0, t)}{\partial x} + \frac{h}{2} \frac{\partial^2 u(0, t)}{\partial x^2} + \\ &+ O(h^2) - \sigma_0(t)u(0, t) + \mu_0(t) = \frac{h}{2} \frac{\partial^2 u(0, t)}{\partial x^2} + O(h^2), \end{aligned}$$

то есть

$$\|\nu\|_{h\tau} = O(h).$$



Граничное условие на правом конце аппроксимируется точно. Таким образом, дифференциальное уравнение аппроксимируется на сетке ω_{ht} с первым порядком по t и вторым – по x , а граничное условие с первым порядком по x . Следовательно, разностная схема аппроксимирует дифференциальную задачу с первым порядком как по переменной t , так и по переменной x .

Пункт б). Как видно из предыдущего пункта, дифференциальное уравнение аппроксимируется простейшей разностной схемой на шаблоне с $\psi(x, t) = O(\tau + h^2)$, а краевое условие – с $\nu(0, t) = O(h)$. Ставится задача: на минимальном шаблоне повысить порядок аппроксимации разностной схемы за счет повышения порядка аппроксимации краевого условия. Будем разностную аппроксимацию краевого условия искать в виде

$$y_x(0, t) = \bar{\sigma}_0(t)y(0, t) - \bar{\mu}_0(t),$$

где $\bar{\sigma}_0(t)$ и $\bar{\mu}_0(t)$ – некоторые сеточные функции, отличные, вообще говоря от $\sigma_0(t)$ и $\mu_0(t)$. Аналогично предыдущему пункту определим погрешность аппроксимации граничного условия:

$$\begin{aligned} \nu(0, t) &= u_x(0, t) - \bar{\sigma}_0(t)u(0, t) + \bar{\mu}_0(t) = \frac{u(x, t) - u(0, t)}{h} - \bar{\sigma}_0(t)u(0, t) + \bar{\mu}_0(t) = \\ &= \frac{\partial u(0, t)}{\partial x} + \frac{h}{2} \frac{\partial^2 u(0, t)}{\partial x^2} + O(h^2) - \bar{\sigma}_0(t)u(0, t) + \bar{\mu}_0(t). \end{aligned}$$

Так как по условию дифференциальной задачи $\frac{\partial u(0, t)}{\partial x} = \sigma_0(t)u(0, t) - \mu_0(t)$, то погрешность аппроксимации перепишется в виде

$$\nu(0, t) = (\sigma_0(t) - \bar{\sigma}_0(t))u(0, t) + \frac{h}{2} \frac{\partial^2 u(0, t)}{\partial x^2} - \mu_0(t) + \bar{\mu}_0(t) + O(h^2).$$

Отсюда видим, что выбор $\bar{\sigma}_0(t) = \sigma_0(t)$, $\bar{\mu}_0(t) = \mu_0(t)$ (как в п.2) приводит к аппроксимации первого порядка. Если же положить $\bar{\sigma}_0(t) = \sigma_0(t)$, $\bar{\mu}_0(t) = \mu_0(t) - \frac{h}{2} \frac{\partial^2 u(0, t)}{\partial x^2} + O(\tau + h^2)$, то получим аппроксимацию граничного условия со вторым порядком по x и первым – по t . Преобразуем теперь выражение для $\bar{\mu}_0(t)$, заменив $\frac{\partial^2 u(0, t)}{\partial x^2}$ разностным отношением. Чтобы удовлетворить требованию минимальности шаблона (в данном случае это две точки в направлении оси Ox), заменим $\frac{\partial^2 u(0, t)}{\partial x^2}$ ее выражением из дифференциального уравнения:

$$\frac{\partial^2 u(0, t)}{\partial x^2} = \frac{\partial u(0, t)}{\partial t} - f(0, t).$$



Тогда

$$\bar{\mu}_0(t) = \mu_0(t) - \frac{h}{2} \left(\frac{\partial u(0, t)}{\partial t} - f(0, t) \right) + O(\tau + h^2) = \mu_0(t) - \frac{h}{2} (u_t(0, t) - f(0, t)).$$

Таким образом, искомая разностная схема повышенного порядка аппроксимации может иметь вид

$$\begin{cases} y_t = y_{\bar{x}\bar{x}} + f(x, t), (x, t) \in \omega_{h\tau}, \\ y(x, 0) = u_0(x), x \in \overline{\omega_h}, \\ y_x(0, t) = \sigma_0(t)y(0, t) + \frac{h}{2}y_t(0, t) - \mu_0(t) - \frac{h}{2}f(0, t), t \in \omega_\tau, \\ y(1, t) = \mu_1(t), t \in \omega_\tau. \end{cases}$$

Она имеет первый порядок аппроксимации по t и второй – по x .

[\[Вернуться к условию\]](#)



Решение задачи 90

Запишем схему (6.1) в индексной форме:

$$\left\{ \begin{array}{l} \frac{y_i^{j+1} - y_i^j}{\tau} + a \left[\sigma \frac{y_i^{j+1} - y_{i-1}^{j+1}}{h} + (1-\sigma) \frac{y_i^j - y_{i-1}^j}{h} \right] = f_i^{j+1/2}, \\ i = 1, 2, \dots, N_1; \quad j = 0, 1, 2, \dots, N_2 - 1; \\ y_i^0 = u^0(x_i), \quad i = 0, 1, 2, \dots, N_1; \quad (h = \frac{1}{N_1}, \quad \tau = \frac{T}{N_2}) \\ y_0^j = \mu(t_j), \quad j = 0, 1, 2, \dots, N_2; \end{array} \right. \quad (\text{P.1})$$

Собирая коэффициенты и обозначая $\gamma = \frac{\sigma T}{h}$, представим разностное уравнение в виде

$$(1 + \sigma\gamma)y_i^{j+1} = \sigma\gamma y_{i-1}^{j+1} + [1 - (1 - \sigma)\gamma]y_i^j + (1 - \sigma)\gamma y_{i-1}^j + \tau f_i^{j+\frac{1}{2}}. \quad (\text{P.2})$$

Используя спектральный метод, исследуем устойчивость схемы (P.1) по начальным данным. С этой целью найдем для уравнения (P.2) при $f \equiv 0$ частное решение в виде гармоники

$$y_i^j = q^j e^{ij\varphi}, \quad i = \sqrt{-1}, \quad (\text{P.3})$$

Подставляя в соответствующее разностное уравнение его решение (P.3), получим

$$(1 + \sigma\gamma)q^{j+1}e^{ij\varphi} = \sigma\gamma q^{j+1}e^{i(j-1)\varphi} + [1 - (1 - \sigma)\gamma]q^j e^{ij\varphi} + (1 - \sigma)\gamma q^j e^{i(j-1)\varphi}$$

или

$$(1 + \sigma\gamma)q = \sigma\gamma q e^{-i\varphi} + 1 - (1 - \sigma)\gamma + (1 - \sigma)\gamma e^{-i\varphi}.$$

Отсюда находим:

$$q = \frac{(1 - \sigma)\gamma + [1 - (1 - \sigma)\gamma]e^{i\varphi}}{e^{i\varphi}(1 + \sigma\gamma) - \sigma\gamma}. \quad (\text{P.4})$$



Гармоника (P.3) с амплитудой q_j , вычисленной по формуле (P.4), является искомым решением однородного аналога уравнения (P.2).

Согласно спектральному методу, схема является устойчивой, если амплитуда q_j не возрастает по модулю с ростом j ни при каких значениях параметра φ . То есть, для устойчивости требуется, чтобы при любых φ выполнялось условие $|q| \leq 1$. Имеем:

$$\begin{aligned}|q|^2 &= \frac{|(1-\sigma)\gamma + [1 - (1-\sigma)\gamma](\cos \varphi + i \sin \varphi)|^2}{|(1+\sigma\gamma)(\cos \varphi + i \sin \varphi) - \sigma\gamma|^2} = \\ &= \frac{(1-\sigma)^2\gamma^2 + 2(1-\sigma)\gamma[1 - (1-\sigma)\gamma] \cos \varphi + [1 - (1-\sigma)\gamma]^2}{(1+\sigma\gamma)^2 - 2(1+\sigma\gamma)\sigma\gamma \cos \varphi + \sigma^2\gamma^2} \leq 1.\end{aligned}$$

Приходим к неравенству

$$(1-\sigma)^2\gamma^2 + [1 - (1-\sigma)\gamma]^2 + 2(1-\sigma)\gamma[1 - (1-\sigma)\gamma] \cos \varphi \leq (1+\sigma\gamma)^2 + \sigma^2\gamma^2 - 2(1+\sigma\gamma)\sigma\gamma \cos \varphi,$$

которое легко преобразуется к виду

$$\gamma[1 + (2\sigma - 1)\gamma](\cos \varphi - 1) \leq 0. \quad (\text{P.5})$$

Так как $\gamma > 0$, $\cos \varphi - 1 \leq 0$ при любых φ , то неравенство (P.5) выполняется, если $1 + (2\sigma - 1)\gamma \geq 0$. Отсюда получаем спектральные условия устойчивости схемы (P.1):

$$\sigma \geq \frac{1}{2}\left(1 - \frac{1}{\gamma}\right). \quad (\text{P.6})$$

[[Вернуться к условию](#)]



Решение задачи 91

Исследуем устойчивость схемы (P.1) с помощью принципа максимума. Этот метод требует, чтобы в каждом узле P разностная схема имела вид:

$$A(P)y(P) = \sum_{Q \in \mathbb{W}'(P)} B(P, Q)y(Q) + F(P) \quad (\text{P.7})$$

где $\mathbb{W}'(P)$ - множество периферийных узлов сеточного шаблона. Согласно принципу максимума схема (P.7) будет устойчивой, если при любом P выполняются условия

$$A(P) > 0, B(P, Q) \geq 0, A(P) \geq \sum_{Q \in \mathbb{W}'(P)} B(P, Q) \quad (\text{P.8})$$

Представим разностное уравнение (P.2) в виде (P.7), полагая

$$P = (x_i, t_{j+1}), Q_1 = (x_i, t_j), Q_2 = (x_{i-1}, t_{j+1}), Q_3 = (x_{i-1}, t_j),$$

$$B(P, Q_1) = 1 - (1 - \sigma)\gamma, B(P, Q_2) = \sigma\gamma, B(P, Q_3) = (1 - \sigma)\gamma,$$

$$F(P) = \tau f_i^{j+\frac{1}{2}}.$$

Решая с этими коэффициентами неравенства (P.8) при $\gamma > 0$, получим следующие условия устойчивости схемы (P.1):

$$0 \leq \sigma \leq 1, \quad \sigma \geq 1 - \frac{1}{\gamma}. \quad (\text{P.9})$$

Как видим, принцип максимума приводит к более жестким условиям, чем спектральный метод.

[\[Вернуться к условию\]](#)



Решение задачи 93.1

Построим на данном шаблоне явную РС с погрешностью аппроксимации $O(\tau + h^2)$ вида

$$\begin{cases} y_t = y_{\bar{x}x} + f(x, t), & (x, t) \in \omega_{h\tau}, \\ y(x, 0) = u_0(x), & x \in \bar{\omega}_h, \\ y_x(0, t) = \sigma_1 y(0, t) + \mu_1(t) + \frac{h}{2}(y_t(0, t) - f(0, t)), & t \in \omega_\tau, \\ -y_x(1, t) = \sigma_2 y(1, t) + \mu_2(t) + \frac{h}{2}(y_t(1, t) - f(1, t)), & t \in \omega_\tau, \end{cases} \quad (\text{P.10})$$

Аппроксимация. Убедимся, что РС (P.10) имеет первый порядок аппроксимации по t и второй по x . Для уравнения погрешность аппроксимации в точке $(x, y) \in \omega_{h\tau}$ равна

$$\begin{aligned} \psi = u_t - u_{\bar{x}x} - f &= \dot{u} + \frac{\tau}{2} \ddot{u} + O(\tau^2) - (u'' + \frac{h^2}{12} u^4 + o(h^4)) - f = \\ &= (\dot{u} - u'' - f) + \frac{\tau}{2} \ddot{u} + O(\tau^2) - \frac{h^2}{12} u^4 + o(h^4) = O(\tau + h^2). \end{aligned} \quad (\text{P.11})$$

Здесь $\dot{u} - u'' - f = 0$ — невязка ДУ на точном решении, $\dot{u} = \frac{\partial u}{\partial t}$, $u'' = \frac{\partial^2 u}{\partial x^2}$.

Для первого граничного условия имеем

$$\begin{aligned} \psi_1 &= u_x(0, t) - \sigma_1 u(0, t) - \mu_1(t) - \frac{h}{2} u_t(0, t) + \frac{h}{2} f(0, t) = \\ &= u'(0, t) + \frac{h}{2} u''(0, t) + O(h^2) - \sigma_1 u(0, t) - \mu_1(t) - \frac{h}{2} (\dot{u}(0, t) + \\ &\quad + O(t)) + \frac{h}{2} f(0, t) = (u'(0, t) - \sigma_1 u(0, t) - \mu_1(t)) + \frac{h}{2} (u''(0, t) - \\ &\quad - \dot{u}(0, t) + f(0, t)) + O(h\tau + h^2) = O(\tau h + h^2). \end{aligned} \quad (\text{P.12})$$

Аналогично, для второго условия $\psi_2 = O(\tau h + h^2)$.



Устойчивость. Явная схема (P.10) для задачи (6.3) устойчива, если для нее выполняется принцип максимума. Запишем схему (P.10) в канонической форме

$$Ay_i^{j+1} = B_1y_{i+1}^j + B_2y_i^j + B_3y_{i-1}^j + \tau f_i^j \quad (\text{P.13})$$

и потребуем выполнения условий

$$A > 0, \quad B_i \geq 0, \quad D = A - \sum B_i \geq 0.$$

Для этого распишем схему (P.10) в индексной форме, получим

$$\begin{cases} \frac{y_i^{j+1} - y_i^j}{\tau} = \frac{y_{i-1}^j - 2y_i^j + y_{i+1}^j}{h^2} + f_i^j, i = \overline{1, N-1}; j = 0, 1, \dots, N-1; \\ y_i^0 = u_0(x_i), i = \overline{0, N}; \\ \frac{y_1^{j+1} - y_1^j}{h} = \sigma_1 y_0^{j+1} + \mu_1(t_{j+1}) + \frac{h}{2} \left(\frac{y_0^{j+1} - y_0^j}{\tau} - f_0^{j+1} \right), j = 0, 1, \dots; \\ -\frac{y_N^{j+1} - y_{N-1}^j}{h} = \sigma_2 y_N^{j+1} + \mu_2(t_{j+1}) + \frac{h}{2} \left(\frac{y_N^{j+1} - y_N^j}{\tau} - f_N^{j+1} \right), j = 0, 1, \dots. \end{cases} \quad (\text{P.14})$$

Соберём коэффициенты в (P.14) и приведём уравнение к виду (P.13)

$$A = 1 > 0, \quad B_1 = B_3 = \gamma > 0, \quad B_2 = 1 - 2\gamma, \quad D = 0, \quad \gamma = \frac{\tau}{h^2}.$$

Условия принципа максимума выполняются, как видим, при $1 - 2\gamma > 0$, отсюда $\gamma \leq \frac{1}{2}$, т.е. условием устойчивости РС (P.10) является ограничение на соотношения шагов по времени и пространственной переменной $\tau \leq \frac{h^2}{2}$.

Для первого граничного условия имеем

$$y_0^{j+1} = \frac{1}{1 + h\sigma_1 + \frac{h^2}{2\tau}} (y_1^{j+1} + \frac{h^2}{2\tau} y_0^j + h\mu_1(t_{j+1}) - \frac{h^2}{2\tau} f_0^{j+1}),$$

отсюда $A = 1 > 0$, $B_1 = \frac{1}{1 + h\sigma_1 + \frac{h^2}{2\tau}} > 0$ при $\sigma_1 > 0$,

$$B_2 = \frac{h^2}{2\tau} B_1 > 0, \quad D = 1 - \frac{1 + \frac{h^2}{2\tau}}{1 + \frac{h^2}{2\tau} + h\sigma_1} > 0.$$



Аналогично будем иметь для второго граничного условия. Таким образом, условиями устойчивости явной разностной схемы (P.10) являются условия

$$\tau \leq \frac{h^2}{2}, \quad \sigma_1 > 0, \quad \sigma_2 > 0. \quad (\text{P.15})$$

Реализация.

- 1) Заполним нулевой слой

$$y_i^0 = U_0(x_i), \quad i = \overline{0, N}; \quad x_i = ih, \quad h = \frac{1}{N};$$

- 2) В цикле по $j = \overline{0, N_t - 1}$ вычисляются следующие значения:

$$y_i^{j+1} = y_i^j + \gamma(y_{i+1}^j - 2y_i^j + y_{i-1}^j) + \tau f_i^j, \quad i = \overline{1, N-1}; \quad \gamma = \frac{\tau}{h^2};$$

$$y_0^{j+1} = \frac{1}{1 + h\sigma_1 + \frac{h^2}{2\tau}}(y_1^{j+1} + \frac{h^2}{2}(\frac{y_0^j}{\tau} - f_0^{j+1}) + h\mu_1(t_{j+1}));$$

$$y_N^{j+1} = \frac{1}{1 + h\sigma_2 + \frac{h^2}{2\tau}}(y_{N-1}^{j+1} + \frac{h^2}{2}(\frac{y_N^j}{\tau} - f_N^{j+1}) + h\mu_2(t_{j+1})).$$

[\[Вернуться к условию\]](#)



Решение задачи 93.2

Неявная разностная схема для задачи (6.3) на шеститочечном шаблоне вида

$$\begin{cases} y_t = \sigma \hat{y}_{\bar{x}x} + (1 - \sigma) y_{\bar{x}x} + f(x, t), & (x, t) \in \omega_{h\tau}, \\ y(x, 0) = U_0(x), & x \in \bar{\omega}_h, \\ \hat{y}_x(0, t) = \sigma_1 \hat{y}(0, t) + \mu_1(t) + \frac{h}{2}(y_t(0, t) - f(0, t)), & t \in \omega_\tau, \\ \hat{y}_{\bar{x}}(1, t) = \sigma_2 \hat{y}(1, t) + \mu_2(t) + \frac{h}{2}(y_t(1, t) - f(1, t)), & t \in \omega_\tau \end{cases} \quad (\text{P.16})$$

$0 \leq \sigma \leq 1$, называется схемой с весами. При $\sigma = 0$ имеем явную разностную схему, при $\sigma = 1$ — чисто неявную.

Аппроксимация и устойчивость. Легко показать, что погрешность аппроксимации РС (P.16) есть $\psi = O(\tau^2 + h^2)$ при $\sigma = 0.5$ и $\psi = O(\tau^2 + h)$ при других значениях σ . Схема (P.16) устойчива в гильбертовой норме при условии $\sigma \geq \frac{1}{2} - \frac{h^2}{4\tau}$, $\sigma \geq 0$, а в норме С при $\tau \leq \frac{h^2}{2(1-\sigma)}$; схема с опережением при $\sigma = 1$ абсолютно устойчива в С, т.е. устойчива при любых h, τ ; симметричная схема при $\sigma = 0.5$ устойчива в С при $\tau \leq h^2$.

Реализация. Для реализации неявных схем, содержащих три неизвестных на верхнем слое в каждом уравнении, применяют обычно метод разностной прогонки. Разностное уравнение

$$\frac{y_i^{j+1} - y_i^j}{\tau} = \sigma \frac{y_{i-1}^{j+1} - 2y_i^{j+1} + y_{i+1}^{j+1}}{h^2} + (1 - \sigma) \frac{y_{i-1}^j - 2y_i^j + y_{i+1}^j}{h^2} + f_i^j$$



и граничные условия приведем к виду, пригодному для разностной прогонки

$$\begin{cases} A_i y_{i-1}^{j+1} - C_i y_{i1}^{j+1} + B_i y_{i+1}^{j+1} = -F_i, i = \overline{1, N-1}, \\ y_0^{j+1} = \kappa_1 y_1^{j+1} + \nu_1, y_N^{j+1} = \kappa_2 y_{N-1}^{j+1} + \nu_2, \end{cases}$$

$$A_i = B_i = \frac{\sigma\tau}{h^2} > 0, \quad C_i = 1 + 2\frac{\sigma\tau}{h^2} > 0,$$

$$F_i = y_i^j + \frac{(1-\sigma)\tau}{h^2} (y_{i-1}^j - 2y_i^j + y_{i+1}^j) + \tau f_{ii}^j,$$

причём $C_i > A_i + B_i$,

$$\kappa_1 = \frac{1}{1 + h\sigma_1 + \frac{h^2}{2\tau}} < 1, \quad \kappa_2 = \frac{1}{1 + h\sigma_2 + \frac{h^2}{2\tau}} < 1,$$

$$\nu_1 = \kappa_1 \left(\frac{h^2}{2} \left(\frac{y_0^j}{\tau} - f(0, t_{j+1}) \right) + h\mu_1(t_{j+1}) \right),$$

$$\nu_2 = \kappa_2 \left(\frac{h^2}{2} \left(\frac{y_{N-1}^j}{\tau} - f(1, t_{j+1}) \right) + h\mu_2(t_{j+1}) \right).$$

Алгоритм имеет следующий вид.

1) Заполнить нулевой слой

$$y_i^0 = U_0(x_i), \quad i = \overline{0, N}.$$

2) Вычислить константы для разностной прогонки

$$A = A_i = B_i = \frac{\sigma\tau}{h^2}; \quad C = C_i = 1 + 2\frac{\sigma\tau}{h^2} \quad \kappa_1, \quad \kappa_2.$$

3) В цикле по $j = \overline{0, N-1}$:

- вычислить ν_1, ν_2 ;

- заполнить массив правых частей

$$F_i = y_i^j + \frac{(1-\sigma)\tau}{h^2} (y_{i-1}^j - 2y_i^j + y_{i+1}^j) + \tau f_{ii}^j, i = \overline{1, N-1}$$

- вычислить y_i^{j+1} , обращаясь к подпрограмме метода разностной прогонки;



Часть IV. Задачи

Решения и указания

Глава 6. Численные методы математической физики

Решение задачи 93.2

Меню



Вверх Назад Вперёд Пред. След. Указатель Помощь Экран

[[Вернуться к условию](#)]



Ответы к задачам



Часть IV. Задачи
Ответы к задачам

Глава 2. Решение систем линейных алгебраических уравнений

Меню



Вверх

Назад

Вперёд

Пред.

След.

Указатель Помощь Экран

Глава 2. Решение систем линейных алгебраических уравнений



Ответ к задаче 25

$\det(B - \lambda E) = (\alpha - \lambda)(\alpha - \lambda - \sqrt{2}\beta)(\alpha - \lambda + \sqrt{2}\beta) = 0, \quad |\alpha| < 1, \quad |\alpha \pm \sqrt{2}\beta| < 1$ [Вернуться к условию]



Часть IV. Задачи

Ответы к задачам

Глава 5. Численное интегрирование

Меню



Вверх

Назад

Вперёд

Пред.

След.

Указатель Помощь Экран

Глава 5. Численное интегрирование



Ответ к задаче 75

$$\|f^{(4)}(x)\| c \frac{1}{2880 \cdot 5^4}$$

[\[Вернуться к условию\]](#)



Часть IV. Задачи

Ответы к задачам

Глава 5. Численное интегрирование

Меню Ответ к задаче 76



Вверх Назад Вперёд Пред. След. Указатель Помощь Экран

Ответ к задаче 76

$$\frac{1}{600}$$

[[Вернуться к условию](#)]



Вверх

Назад

Вперёд

Пред.

След.

Указатель

Помощь

Экран

Ответ к задаче 77

$$N \geq \left\lceil \sqrt{\frac{\theta}{12}} \cdot 10^{12} \right\rceil + 1 \approx 41, \quad \theta = \max(2, 4 \sin 1 - 2 \cos 1) \approx 2.29$$

[\[Вернуться к условию\]](#)



Вверх

Назад

Вперёд

Пред.

След.

Указатель

Помощь

Экран

Ответ к задаче 81.1

$$\frac{1}{2}f\left(\frac{2}{3}\right)$$

[\[Вернуться к условию\]](#)



Ответ к задаче 81.2

$$(e - 1)f \left(\frac{1}{e - 1} \right)$$

[\[Вернуться к условию\]](#)



Ответ к задаче 83

$$\frac{5}{9}f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9}f(0) + \frac{5}{9}f\left(\sqrt{\frac{3}{5}}\right).$$

[\[Вернуться к условию\]](#)



Ответ к задаче 84

$$S_2(f) = f\left(\frac{\pi + \sqrt{\pi^2 - 8}}{2}\right) + f\left(\frac{\pi - \sqrt{\pi^2 - 8}}{2}\right).$$

[\[Вернуться к условию\]](#)



Ответ к задаче 85

$$S_2(f) = \frac{2 + \sqrt{2}}{4} f(2 - \sqrt{2}) + \frac{2 - \sqrt{2}}{4} f(2 + \sqrt{2}).$$

[\[Вернуться к условию\]](#)



Часть IV. Задачи

Ответы к задачам

Тест 5. Численное интегрирование

Меню Ответ к задаче 86



Вверх

Назад

Вперёд

Пред.

След.

Указатель

Помощь

Экран

Ответ к задаче 86

$$S_{h,h/2} = S_{h/2} + \frac{S_{h/2} - S_h}{2^m - 1}.$$

[[Вернуться к условию](#)]



Часть V

Тесты

- Тест 1. Методы решения СЛАУ
- Тест 2. Решение нелинейных уравнений
- Тест 3. Приближение функций
- Тест 4. Численное интегрирование
- Тест 5. Численное решение обыкновенных дифференциальных уравнений



Тест 1

Методы решения СЛАУ

- 1.1. Число обусловленности. Прямые методы решения СЛАУ
- 1.2. Итерационные методы решения СЛАУ. Форматы хранения разреженных матриц



1.1. Число обусловленности. Прямые методы решения СЛАУ

Начало теста.

1. Число обусловленности матрицы A вычисляется по формуле

$$\alpha(A) = \frac{\|A^{-1}\|}{\|A\|}$$

$$\alpha(A) = \frac{\|A\|}{\|A^{-1}\|}$$

$$\alpha(A) = \|A\| \cdot \|A^{-1}\|$$

$$\alpha(A) = |A| \cdot |A^{-1}|$$



2. Если матрица A плохо обусловлена, то

относительная погрешность решения СЛАУ $Ax = b$ может быть намного больше относительной погрешности, вносимой в матрицу A и вектор b .

абсолютная погрешность решения СЛАУ $Ax = b$ может быть намного больше абсолютной погрешности, вносимой в матрицу A и вектор b .

задача решения СЛАУ $Ax = b$ является некорректной.

невозможно точно решить СЛАУ $Ax = b$.



3. Число обусловленности матрицы $A = \begin{pmatrix} 1 & 2 \\ 0 & -7 \end{pmatrix}$ в норме $\|\cdot\|_\infty$ равно .



4. Укажите условие, достаточное для выполнимости метода Гаусса (без выбора главного элемента) при решении СЛАУ $Ax = b$.

Матрица A хорошо обусловлена.

Все главные угловые миноры матрицы A отличны от нуля.

Определитель матрицы A отличен от нуля.

Размерность матрицы A меньше 1000.



5. Система линейных алгебраических уравнений решается методом Гаусса с выбором главного элемента по столбцу. В ходе решения получена следующая промежуточная матрица:

$$\left(\begin{array}{ccccc} 2 & 1 & -1 & 3 & -2 \\ 0 & -1 & 1 & 0 & 5 \\ 0 & 0 & 2 & 1 & 0 \\ 0 & 0 & 4 & -1 & 3 \\ 0 & 0 & -5 & 3 & 1 \end{array} \right).$$

Каким должен быть следующий шаг алгоритма?

Поменять местами 3-ю и 4-ю строки.

Добавить к 5-й строке 3-ю, умноженную на $\frac{5}{2}$.

Вычесть из 4-й строки 3-ю, умноженную на 2.

Поменять местами 3-ю и 5-ю строки.



6. Система линейных алгебраических уравнений решается методом Гаусса с выбором главного элемента по строке. В ходе решения получена следующая промежуточная матрица:

$$\left(\begin{array}{ccccc} 2 & 1 & -1 & 3 & -2 \\ 0 & -1 & 1 & 0 & 5 \\ 0 & 0 & 2 & 1 & 0 \\ 0 & 0 & -1 & -1 & 3 \\ 0 & 0 & 1 & 3 & 1 \end{array} \right).$$

Каким будет следующий шаг алгоритма?

Поменять местами 3-й и 4-й столбцы.

Добавить к 4-й строке 3-ю, разделенную на 2.

Добавить к 5-й строке 4-ю.

Поменять местами 3-й и 5-й столбцы.



7. В ходе решения СЛАУ методом Гаусса получена следующая промежуточная матрица:

$$\begin{pmatrix} 2 & 1 & -1 & 3 & -2 \\ 0 & -1 & 1 & 0 & 5 \\ 0 & 0 & 2 & -3 & 2 \\ 0 & 0 & 0 & -1 & 3 \\ 0 & 0 & -1 & 3 & 1 \end{pmatrix}.$$

Умножение (слева) на какую из указанных матриц будет следующим шагом алгоритма?



$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix}$$
$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0.5 & 0 & 1 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$
$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0.5 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$



8. Метод квадратного корня применяется для решения СЛАУ

- с трехдиагональной матрицей.
- с симметричной матрицей.
- с любой невырожденной матрицей.
- с верхнетреугольной матрицей.



9. Алгоритм решения СЛАУ $Ax = b$ с нижнетреугольной матрицей

$$A = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix},$$

имеет вид

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



10. LU-разложение матрицы $A = \begin{pmatrix} 1 & 2 \\ -1 & 3 \end{pmatrix}$ имеет вид

$$L = \begin{bmatrix} & \\ & \end{bmatrix}, \quad U = \begin{bmatrix} 1 & 2 \\ 0 & 5 \end{bmatrix}.$$



1.2. Итерационные методы решения СЛАУ. Форматы хранения разреженных матриц

Начало теста.

1. Рассмотрим итерационный процесс $x^{k+1} = Bx^k + g$, где B — квадратная матрица, x^k и g — векторы соответствующей размерности. Укажите необходимое условие сходимости этого итерационного процесса.

$$\rho(B) < 1$$

$$\det B < 1$$

$$\|B\|_2 < 1$$

$$\|B\|_\infty < 1$$



2. Рассмотрим итерационный процесс $x^{k+1} = Bx^k + g$, где B — квадратная матрица, x^k и g — векторы соответствующей размерности. Укажите условие, выполнения которого *не достаточно* для сходимости этого итерационного процесса.

$$\rho(B) < 1$$

$$\det B < 1$$

$$\|B\|_2 < 1$$

$$\|B\|_\infty < 1$$



3. Рассмотрим СЛАУ

$$\begin{cases} 5x_1 - 4x_2 + x_3 = 1, \\ 3x_1 + 2x_2 - 3x_3 = 2, \\ x_1 + 4x_3 = -1. \end{cases}$$

Какой из приведенных ниже итерационных процессов соответствует методу Гаусса–Зейделя?

$$\begin{cases} x_1^{k+1} = \frac{1}{5}(1 + 4x_2^k - x_3^k), \\ x_2^{k+1} = \frac{1}{2}(2 - 3x_1^k + 3x_3^k), \\ x_3^{k+1} = -\frac{1}{4}(x_1^{k+1} + 1). \end{cases}$$

$$\begin{cases} x_1^{k+1} = \frac{1}{5}(1 + 4x_2^k - x_3^k), \\ x_2^{k+1} = \frac{1}{2}(2 + 3x_1^k - 3x_3^k), \\ x_3^{k+1} = -\frac{1}{4}(x_1^k + 1). \end{cases}$$

$$\begin{cases} x_1^{k+1} = \frac{1}{5}(1 + 4x_2^{k+1} - x_3^k), \\ x_2^{k+1} = \frac{1}{2}(2 - 3x_1^{k+1} + 3x_3^k), \\ x_3^{k+1} = -\frac{1}{4}(x_1^{k+1} + 1). \end{cases}$$

$$\begin{cases} x_1^{k+1} = \frac{1}{5}(1 + 4x_2^k - x_3^k), \\ x_2^{k+1} = \frac{1}{2}(2 - 3x_1^{k+1} + 3x_3^k), \\ x_3^{k+1} = -\frac{1}{4}(x_1^{k+1} + 1). \end{cases}$$



4. Рассмотрим СЛАУ

$$\begin{cases} 5x_1 - 4x_2 + x_3 = 1, \\ 3x_1 + 2x_2 - 3x_3 = 2, \\ x_1 + 4x_3 = -1. \end{cases}$$

Какой из приведенных ниже итерационных процессов соответствует методу Якоби?

$$\begin{cases} x_1^{k+1} = \frac{1}{5}(1 + 4x_2^k - x_3^k), \\ x_2^{k+1} = \frac{1}{2}(2 - 3x_1^k + 3x_3^k), \\ x_3^{k+1} = -\frac{1}{4}(x_1^{k+1} + 1). \end{cases}$$

$$\begin{cases} x_1^{k+1} = \frac{1}{5}(1 + 4x_2^k - x_3^k), \\ x_2^{k+1} = \frac{1}{2}(2 + 3x_1^k - 3x_3^k), \\ x_3^{k+1} = -\frac{1}{4}(x_1^k + 1). \end{cases}$$

$$\begin{cases} x_1^{k+1} = \frac{1}{5}(1 + 4x_2^{k+1} - x_3^k), \\ x_2^{k+1} = \frac{1}{2}(2 - 3x_1^{k+1} + 3x_3^k), \\ x_3^{k+1} = -\frac{1}{4}(x_1^{k+1} + 1). \end{cases}$$

$$\begin{cases} x_1^{k+1} = \frac{1}{5}(1 + 4x_2^k - x_3^k), \\ x_2^{k+1} = \frac{1}{2}(2 - 3x_1^{k+1} + 3x_3^k), \\ x_3^{k+1} = -\frac{1}{4}(x_1^{k+1} + 1). \end{cases}$$



5. Рассмотрим СЛАУ

$$\begin{cases} x_1 - x_2 = -1, \\ 2x_1 - x_2 + 3x_3 = 1, \\ 2x_2 + 5x_3 = 2. \end{cases}$$

Пусть $x^0 = (1, 1, 1)^T$. Приближенное решение x^1 , вычисленное по методу Якоби, имеет вид

$$x^1 = \begin{pmatrix} 0.6667 \\ 0.6667 \\ 0.6667 \end{pmatrix},$$

а по методу Гаусса–Зейделя —

$$x^1 = \begin{pmatrix} 0.6667 \\ 0.6667 \\ 0.6667 \end{pmatrix}.$$



6. Рассмотрим СЛАУ

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

Пусть $x^0 = (1, 1)^T$. Приближенное решение x^1 , вычисленное по методу градиентного спуска, имеет вид

$$x^1 = \begin{pmatrix} \quad \\ \quad \end{pmatrix}.$$



7. Рассмотрим СЛАУ

$$\begin{cases} x_1 + 2x_2 = 3, \\ 2x_1 - 5x_2 = 2. \end{cases}$$

Охарактеризуйте сходимость итерационных процессов Якоби и Гаусса–Зейделя для этой системы.

Процесс Якоби сходится, процесс Гаусса–Зейделя расходится.

Процесс Якоби расходится, процесс Гаусса–Зейделя сходится.

Оба процесса сходятся.

Оба процесса расходятся.



8. Рассмотрим СЛАУ $Ax = b$. Пусть все собственные значения матрицы A лежат в диапазоне от 1 до 100. Тогда итерационный процесс

$$x^{k+1} = (I - \tau A)x^k + \tau b$$

будет сходиться при всех

$$\tau < \dots$$



9. Представление матрицы $A = \begin{pmatrix} 1 & 0 & 3 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 5 & 7 & 0 & 2 \end{pmatrix}$ в формате CSR имеет вид

AA:

JA:

IA:

,

а в формате MSR —

1 2 3 4 5 6 7 8

AA: X

IJ:

10. Пусть матрица A имеет следующее представление в формате MSR:

	1	2	3	4	5	6	7	8	9	10	11
AA	5	0	7	3	0	X	9	1	8	3	4
IJ	7	8	10	10	11	12	3	4	5	2	2

Тогда

$$A = \begin{pmatrix} & \\ & \\ & \\ & \\ & \end{pmatrix}$$



Тест 2

Решение нелинейных уравнений

Начало теста.

- К корню уравнения $x^7 - 5x^2 + 2 = 0$, расположенному на отрезке $[-2, 1]$ сходится итерационный процесс

$$x_{n+1} = \sqrt[7]{\frac{x_n^7 + 2}{5}}$$

$$x_{n+1} = \sqrt[7]{5x^2 - 2}$$

$$x_{n+1} = x_n + x_n^7 - 5x_n^2 + 2$$

$$x_{n+1} = x_n - x_n^7 + 5x_n^2 - 2$$



2. Один и только один корень уравнения $x \sin 5x = 1$ принадлежит промежутку

$$(0.5, 1)$$

$$\left(\frac{2\pi}{5}, \frac{3\pi}{5} \right)$$

$$\left(\frac{2\pi}{5}, \frac{\pi}{2} \right)$$

$$(1, 2)$$



3. Метод итерации $x_{n+1} + C(x_n^5 - 3x_n + 1)$ сходится к корню уравнения $x^5 = 3x - 1$, расположенному на отрезке $[-2, -1]$ при любом значении C из промежутка

$(-1, 0)$

$(0, 1)$

$(-1, 1)$

$(-0.01, 0)$



4. Какой скоростью сходимости обладает метод секущих вблизи простого корня уравнения $f(x) = 0$?

линейной

квадратичной

кубической

другой ответ



5. Каким образом связаны погрешности двух соседних приближений к корню x^* уравнения $f(x) = 0$, полученных по методу Ньютона?

$$\varepsilon_{n+1} \approx f'(x^*)\varepsilon_n$$

$$\varepsilon_{n+1} \approx f''(x^*)\varepsilon_n^2$$

$$\varepsilon_{n+1} \approx \frac{f''(x^*)}{f'(x^*)}\varepsilon_n^2$$

$$\varepsilon_{n+1} \approx -\frac{1}{2} \frac{f''(x^*)}{f'(x^*)}\varepsilon_n^2$$



6. Алгоритм метода хорд применительно к решению уравнения $x^2 - 2 = 0$ исходя из приближений $x_0 = 1$, $x_1 = 2$ имеет вид

$$x_{n+1} = \frac{x_n + 2}{x_n + 1}$$

$$x_{n+1} = \frac{x_n}{2} + \frac{1}{x_n}$$

$$x_{n+1} = x_n - x_n^2 + 2$$

$$x_{n+1} = x_n - \frac{x_n^2 - 1}{2}$$



7. Алгоритм метода Ньютона применительно к вычислению значения $\sqrt[3]{5}$ путём решения уравнения $x^3 - 5 = 0$ имеет вид

$$x_{n+1} = x_n + 0.5(x_n^3 - 5)$$

$$x_{n+1} = \frac{2}{3}x_n + \frac{5}{3x_n^2}$$

$$x_{n+1} = x_n - 2\frac{x_n^3 - 5}{3x_n^2}$$

$$x_{n+1} = \sqrt{5 - x_n^3 + x_n^2}$$



Вверх

Назад

Вперёд

Пред.

След.

Указатель

Помощь

Экран

Меню

8. Если $(x^{(0)}, y^{(0)}) = (2, 2)$, то приближение $(x^{(1)}, y^{(1)})$ к решению системы

$$\begin{cases} x + y^2 - 2 = 0, \\ x^2 - y = 0, \end{cases}$$

по методу Ньютона равно (,).



9. Функционалом, минимизация которого равносильна решению системы нелинейных уравнений $f_i(x_1, \dots, x_n) = 0, i = \overline{1, n}$, является

$$Q(x_1, \dots, x_n) = \sum_{i=1}^n f_i(x_1, \dots, x_n)$$

$$Q(x_1, \dots, x_n) = \sum_{i=1}^n f_i^2(x_1, \dots, x_n)$$

$$Q(x_1, \dots, x_n) = \prod_{i=1}^n f_i(x_1, \dots, x_n)$$

$$Q(x_1, \dots, x_n) = \prod_{i=1}^n f_i^2(x_1, \dots, x_n)$$



10. Оптимальная длина шага t_k в методе градиентного спуска определяется как решение уравнения

$$Q(x^k - t_k \operatorname{grad} Q(x^{(k)})) = 0$$

$$\frac{d}{dt} Q(x^k - t_k \operatorname{grad} Q(x^{(k)})) = 0$$

$$t_k - \frac{(\operatorname{grad} Q(x^{(k)}), x^{(k)})}{(x^{(k)}, x^{(k)})} = 0$$

$$\frac{(t_k \operatorname{grad} Q(x^{(k)}), x^{(k)})}{(x^{(k)}, x^{(k)})} = 0$$



Тест 3

Приближение функций

Начало теста.

- Интерполяционный многочлен для функции f в форме Лагранжа имеет вид

$$P_n(x) = f(x_0) + (x - x_0)f(x_0, x_1) + \dots + (x - x_0) \dots (x - x_{n-1})f(x_0, \dots, x_n)$$

$$P_n(x) = \sum_{k=0}^n \frac{\omega(x)}{(x - x_k)\omega'(x_k)} f(x_k)$$

$$P_n(x) = \sum_{k=0}^n \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k$$

$$P_n(x) = \sum_{k=0}^n a_k x^k$$



2. Остаток в форме Лагранжа алгебраического интерполяирования по значениям функции f имеет вид

$$r_n(x) = f(x, x_0, x_1, \dots, x_n) \cdot (x - x_0)(x - x_1) \dots (x - x_n)$$

$$r_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} (x - x_0)^{n+1}$$

$$r_n(x) = o(x - x_0)^n$$

$$r_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} (x - x_0)(x - x_1) \dots (x - x_n)$$



3. Интерполяционный многочлен для функции f в форме Ньютона имеет вид

$$P_n(x) = f(x_0) + (x - x_0)f(x_0, x_1) + \dots + (x - x_0) \dots (x - x_{n-1})f(x_0, \dots, x_n)$$

$$P_n(x) = \sum_{k=0}^n \frac{\omega(x)}{(x - x_k)\omega'(x_k)} f(x_k)$$

$$P_n(x) = \sum_{k=0}^n \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k$$

$$P_n(x) = \sum_{k=0}^n a_k x^k$$



4. Остаток в форме Ньютона алгебраического интерполяирования по значениям функции f имеет вид

$$r_n(x) = f(x, x_0, x_1, \dots, x_n) \cdot (x - x_0)(x - x_1) \dots (x - x_n)$$

$$r_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} (x - x_0)^{n+1}$$

$$r_n(x) = o(x - x_0)^n$$

$$r_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} (x - x_0)(x - x_1) \dots (x - x_n)$$



5. Функция $f(x)$ задана таблицей значений

x	0	2	3
<hr/>			
y	1	3	1

Тогда приближённое значение $f(1)$, вычисленное с помощью линейной интерполяции, равно



6. Значение функции $f(x) = \sin x$ в точке $x = 1$ приближённо вычисляется с помощью квадратичной интерполяции по узлам $x_0 = 0$, $x_1 = 0.5$, $x_2 = 2$ (в этих узлах значения $\sin x_i$ заданы точно). Тогда погрешность найденного значения не превосходит .



7. Минимизация остатока интерполяции на отрезке $[-1, 1]$ для заданной достаточно гладкой функции связана с расположением узлов интерполяции в нулях многочленов

Чебышева второго рода

Лежандра

Чебышева второго рода

Эрмита



8. Многочлены Чебышева первого рода связаны рекуррентным соотношением

$$T_{n+1}(x) = \frac{2n+1}{n+1} T_n(x) - \frac{n}{n+1} T_{n-1}(x)$$

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$$

$$T_{n+1}(x) = nxT_n(x) + \frac{n}{n+1} T_{n-1}(x)$$

$$T_{n+1}(x) = \frac{1}{n+1} ((xn+1-x)T_n(x) - nT_{n-1}(x))$$

9. Функция $f(x)$, заданная таблицей

x	0	1	2
<hr/>			
y	2	3	0

интерполируется с помощью кубического сплайна с естественными граничными условиями. Тогда приближённое значение $f(0.5)$ равно .



10. Построенное по методу наименьших квадратов приближение функции $f(x)$, заданной таблицей

x	0	1	2
y	<hr/>		
	2	3	0

с помощью многочлена нулевой степени имеет вид

$$P_0(x) = \dots$$



Тест 4

Численное интегрирование

Начало теста.

1. Алгебраическая степень точности квадратурной формулы

$$\int_0^1 f(x) dx \approx \frac{1}{4} \left(f(0) + 3f\left(\frac{2}{3}\right) \right)$$
 равна

0

1

2

3



2. Если алгебраическая степень точности квадратурной формулы

$$\int_{-1}^1 \frac{f(x)}{\sqrt{1-x^2}} dx \approx A_0 f(-1) + A_1 f(1)$$

равна 1, то значения коэффициентов A_0 и A_1 равны

$$A_0 = A_1 = 1$$

$$A_0 = A_1 = \frac{\pi}{2}$$

$$A_0 = \frac{1}{2}, A_1 = \frac{3}{2}$$

$$A_0 = \frac{\pi}{2}, A_1 = \frac{3\pi}{4}$$



3. Квадратурная формула Ньютона–Котеса для веса $p(x) \equiv 1$ с двумя узлами это

формула средних
прямоугольников

формула трапеций

формула Симпсона

формула левых прямоугольников



4. Априорная оценка количества частей, на которые следует разбить про-

межуток интегрирования для вычисления интеграла $\int_0^1 \cos x dx$ по со-

ставной формуле средних прямоугольников с точностью $\varepsilon \leq 10^{-4}$ имеет вид $N \geq$



5. При вычислении некоторого интеграла по составной формуле трапеций получены следующие результаты: при $h_1 = \frac{1}{10}$ $I_1 = 1.2$, при $h = \frac{1}{20}$ $I_2 = 1.2515$. Тогда оценённая по правилу Рунге погрешность первого из значений равна: .



6. Для того, чтобы квадратурная формула для вычисления интеграла

$$\int_{-1}^1 f(x) dx$$
 была формулой типа Гаусса, её узлы должны:

равномерно располагаться на отрезке

быть корнями многочлена Чебышёва первого рода

быть корнями многочлена Лежандра

быть корнями многочлена Эрмита



7. Единственной квадратурной формулой наивысшей алгебраической степени точности, имеющей равные коэффициенты, является формула, узлы которой —

корни многочлена Чебышева первого рода

корни многочлена Чебышева–Эрмита

корни многочлена Лежандра

корни многочлена Чебышева–Лагерра



- Меню
8. Приближенное значение интеграла $\int_0^2 x^4 dx$, вычисленное по простой квадратурной формуле Симпсона, равно .



Меню

Вверх

Назад

Вперёд

Пред.

След.

Указатель Помощь Экран

9. Оценка погрешности приближённого значения интеграла

$$\int_0^2 x^4 dx,$$

численного по простой квадратурной формуле Симпсона, не превосходит



10. Вычисленное по простой кубатурной формуле средних приближённое значение интеграла $\iint_{\Delta} (x^2+y^2) dx dy$, где Δ — треугольник с вершинами $A(0, 0)$, $B(3, 0)$, $C(0, 3)$, равно .



Тест 5

Численное решение обыкновенных дифференциальных уравнений

Начало теста.

1. Неявным методом Эйлера для уравнения $u'(t) = f(t, u(t))$ является разностное уравнение

$$y_{j+1} = y_j - \tau f(t_{j+1}, y_i)$$

$$y_{j+1} = y_j + \tau f(t_j, y_{j+1})$$

$$y_{j+1} = y_j + \tau f(t_{j+1}, y_{i+1})$$

$$y_{j+1} = y_j + \tau f(t_j, y_j)$$



2. Явным методом Эйлера для уравнения $u'(t) = f(t, u(t))$ является разностное уравнение

$$y_{j+1} = y_j - \tau f(t_{j+1}, y_i)$$

$$y_{j+1} = y_j + \tau f(t_j, y_{j+1})$$

$$y_{j+1} = y_j + \tau f(t_{j+1}, y_{i+1})$$

$$y_{j+1} = y_j + \tau f(t_j, y_j)$$



3. Порядок точности метода $y_{j+1} = y_j + \frac{\tau}{2}(3f_j - f_{j-1})$ равен

1

2

3

4



4. Порядок точности метода $y_{j+1} = y_j + \frac{\tau}{2}(f_j + f_{j+1})$ равен

1

2

3

4



5. Интервал устойчивости явного метода Эйлера, определённый на модельном уравнении $u'(t) = \lambda u(t)$, $\lambda < 0$, равен

$$(-1, 0)$$

$$(-\infty, 0) \cup (2, +\infty)$$

$$(-2, 0)$$

$$(0, 2)$$



6. Интервал устойчивости неявного метода Эйлера, определённый на модельном уравнении $u'(t) = \lambda u(t)$, $\lambda < 0$, равен

$$(-1, 0)$$

$$(-\infty, 0) \cup (2, +\infty)$$

$$(-2, 0)$$

$$(0, 2)$$



7. Отметьте правильную замену 2-й производной в случае равноотстоящих узлов:

$$u''(x_i) = \frac{u(x_{i-1}) - 2u(x_i) + u(x_{i+1})}{h^2} + O(h^2)$$

$$u''(x_i) = \frac{u(x_{i-1}) + 2u(x_i) + u(x_{i+1})}{h^2} + O(h^2)$$

$$u''(x_i) = \frac{u(x_{i-1}) - 2u(x_i) + u(x_{i+1})}{2h} + O(h^2)$$

$$u''(x_i) = \frac{u(x_{i-1}) + 2u(x_i) + u(x_{i+1})}{2h} + O(h^2)$$



8. Найденное методом Галёркина приближенное решение u_1 граничной задачи

$$u'' = 1, \quad u(0) = u(1) = 0$$

с выбором в качестве базисной функции $\varphi_1(x) = x(1 - x)$ имеет вид

$$u_1(x) = \dots \cdot x(1 - x).$$



9. Погрешность аппроксимации разностной схемы

$$\begin{cases} y_{\bar{x}x} + y_{\circ_x} = 0, \\ y(0) = y(1) = 0, \end{cases}$$

вычисленная на решении задачи

$$\begin{cases} u'' + u' = 0, \\ u(0) = u(1) = 0 \end{cases}$$

имеет вид $O(h^{\quad})$.



10. Метод разностной прогонки предназначен для решения

СЛАУ общего вида

СЛАУ с 5-диагональной матрицей

СЛАУ с 3-диагональной матрицей

систем нелинейных уравнений



Рекомендуемая литература

- [1] Алберг Дж., Нильсон Э., Уолш Дж. Теория сплайнов и ее приложения. М.: Мир, 1972.
- [2] А.Б. Антоневич, Я.В. Радыно. Функциональный анализ и интегральные уравнения. 2-е изд., перераб. и доп. – Минск: БГУ, 2003.
- [3] Бахвалов Н. С., Лапин А. В., Чижонков Е. В. Численные методы в задачах и упражнениях. Учеб. пособие. / Под ред. В. А. Садовничего — М.: Высш. шк. 2000. — 190 с.
- [4] Н. С. Бахвалов, Н . П. Жидков, Г. М. Кобельков. Численные методы. — 3-е изд., доп. и перераб. — М.: БИНОМ. Лаборатория знаний, 2004. — 636 с., ил.
- [5] Ващенко Г. В. Вычислительная математика. Основы конечных методов решения систем линейных алгебраических уравнений. Уч. пособие. — Красноярск: СибГТУ, 2005. — 80 с.
- [6] Ващенко Г. В. Вычислительная математика. Основы алгебраической и тригонометрической интерполяции. Уч. пособие. — Красноярск: СибГТУ, 2008. — 64 с.
- [7] Деккер К., Вервер Я. Устойчивость методов Рунге-Кутты для жестких нелинейных дифференциальных уравнений. – М.: Мир, 1988.
- [8] Дэннис Дж., Шнабель Р. Численные методы безусловной оптимизации и решения нелинейных уравнений. – М.: Мир, 1988.
- [9] Завьялов Ю.С., Квасов Б.И., Мирошниченко В.Л. Методы сплайн-функций. М.: Наука, 1980.
- [10] Коллатц Л., Крабс В. Теория приближений. Чебышевские приближения и их приложения. М.: Наука, 1978.



- [11] Каханер Д., Моулер К., Нэш С. Численные методы и программное обеспечение. — М.: Мир, 2001. — 575 с., ил.
- [12] Крылов В. И. Приближенное вычисление интегралов. — М.: Наука, 1967.
- [13] Крылов В. И., Бобков В. В., Монастырный П. И. Вычислительные методы. Т. 1. — М.: Наука, 1976.
- [14] Крылов В. И., Бобков В. В., Монастырный П. И. Вычислительные методы. Т. 2. — М.: Наука, 1977.
- [15] Орtega Дж., Рейнболдт В. Итерационные методы решения нелинейных систем уравнений со многими неизвестными. — М.: Мир, 1975.
- [16] Самарский А. А. Теория разностных схем. — М.: Наука, 1982.
- [17] Хайрер Э., Нёрсетт С., Ваннер Г. Решение обыкновенных дифференциальных уравнений. Нежесткие задачи. — М.: Мир, 1990.
- [18] William H. Press [et al.]. Numerical Recipes in C: the art of scientific computing. — 2nd ed. Cambridge University Press, 1992.