А. А. Самарский, А. В. Гулин

ЧИСЛЕННЫЕ МЕТОДЫ

Допущено Государственным комитетом СССР
по народному образованию
в качестве учебного пособия
для студентов вузов, обучающихся по специальности
«Прикладная математика»



МОСКВА «НАУКА»
ГЛАВНАЯ РЕДАКЦИЯ
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ
1989

ББК 22.19
С17
УДК 519.6(075.8)

Самарский А. А., Гулин А. В. **Численные методы:** Учеб. пособие для вузов.— М.: Наука. Гл. ред. физ-мат. лит., 1989.— 432 с.— ISBN 5-02-013996-3.

Излагаются основные принципы построения и исследования численных методов решения на ЭВМ различных классов математических задач. Наряду с традиционными разделами, такими как интерполирование, численное интегрирование, методы решения задачи Коши для обыкновенных дифференциальных уравнений, большое место в книге занимают разностные методы для уравнений в частных производных и итерационные методы решения сеточных уравнений.

Для студентов, обучающихся по специальности «Прикладная математика» и «Физика», а также для широкого круга специалистов, применяющих ЭВМ для научных расчетов.

Табл. 2. Ил. 16. Библиогр. 46 назв.

Рецензент
доктор физико-математических наук *А. А. Абрамов*

С $\frac{1602120000-045}{053(02)-89}$ 52-89

ISBN 5-02-013996-3

© Издательство «Наука». Главная редакция физико-математической литературы, 1989

ОГЛАВЛЕНИЕ

Предисловие 8

ЧАСТЬ I

ВВЕДЕНИЕ В ЧИСЛЕННЫЕ МЕТОДЫ

§ 1. Математическое моделирование и вычислительный эксперимент . . 11
1. Схема вычислительного эксперимента (11). 2. Вычислительный алгоритм (12). 3. Требования к вычислительным методам (14).

§ 2. Погрешности округления 16
1. Представление вещественных чисел в ЭВМ (16). 2. Округление чисел в ЭВМ (17). 3. Накопление погрешностей округления (19). 4. Разностные уравнения первого порядка (20). 5. Оценки погрешностей округления (22)

§ 3. Разностные уравнения второго порядка 25
1. Задача Коши и краевые задачи для разностных уравнений (25). 2. Однородное разностное уравнение второго порядка с постоянными коэффициентами (26). 3. Однородное разностное уравнение второго порядка с переменными коэффициентами (28). 4. Неоднородное разностное уравнение второго порядка (31).

§ 4. Разностная аппроксимация дифференциальных уравнений . . . 34
1. Сетки и сеточные функции (34). 2. Разностная краевая задача (35). 3. Некоторые разностные тождества (38). 4. Разностная задача на собственные значения (39). 5. Свойства собственных значений и собственных функций (41). 6. Разрешимость и сходимость разностной задачи (43). 7. Метод прогонки (45).

ЧАСТЬ II

ЧИСЛЕННЫЕ МЕТОДЫ АЛГЕБРЫ И АНАЛИЗА

Глава 1. **Прямые методы решения систем линейных алгебраических уравнений** 48

§ 1. Метод Гаусса численного решения систем линейных алгебраических уравнений 49
1. Основная идея метода (49). 2. Расчетные формулы (51). 3. Подсчет числа действий (53).

§ 2. Условия применимости метода Гаусса 54
1. Связь метода Гаусса с разложением матрицы на множители (54). 2. Теорема об LU-разложении (55). 3. Элементарные треугольные матрицы (58).

§ 3. Метод Гаусса с выбором главного элемента 60
1. Основная идея метода (60). 2. Матрицы перестановок (61). 3 Пример (62). 4. Общий вывод (65). 5. Доказательство теоремы 1 (66). 6. Вычисление определителя (67).

§ 4. Обращение матрицы 68

§ 5. Метод квадратного корня 69
1. Факторизация эрмитовой матрицы (69). 2. Пример (70). 3. Общие расчетные формулы (71). 4. Подсчет числа действий (72).

§ 6. Обусловленность систем линейных алгебраических уравнений . . 74
1. Устойчивость системы линейных алгебраических уравнений (74). 2. Число обусловленности (76). 3. Полная оценка относительной погрешности (77). 4. Влияние погрешностей округления при решении систем линейных алгебраических уравнений методом Гаусса (79).

Глава 2. **Итерационные методы решения систем линейных алгебраических уравнений** 82

§ 1. Примеры и канонический вид итерационных методов решения систем линейных алгебраических уравнений 82
1. Итерационные методы Якоби и Зейделя (82). 2. Матричная запись методов Якоби и Зейделя (83). 3. Каноническая форма одношаговых итерационных методов (84).

§ 2. Исследование сходимости итерационных методов 86

§ 3. Необходимое и достаточное условие сходимости стационарных итерационных методов 90
1. Введение (90). 2. Норма матрицы (91). 3. Теорема о сходимости итерационного метода (92). 4. Продолжение доказательства (93).

§ 4. Оценки скорости сходимости стационарных итерационных методов . 95
1. Скорость сходимости итерационного метода (95). 2. Оценки скорости сходимости в случае симметричных матриц A и B (96). 3. Правила действий с матричными неравенствами (98). 4. Доказательство теоремы 1 (100). 5. Оценка погрешности в случае несимметричной матрицы B (102).

§ 5. Многочлены Чебышева 103
1. Многочлен Чебышева на отрезке [—1, 1] (103). 2. Случай произвольного отрезка (105). 3. Другая нормировка многочленов Чебышева (106). 4. Примеры применения многочленов Чебышева (107).

§ 6. Итерационные методы с чебышевским набором параметров . . . 109
1. Явный итерационный метод (109). 2. Численная устойчивость итерационного метода с чебышевским набором параметров (112). 3. Неявный чебышевский итерационный метод (113). 4. Случай, когда точные границы спектра неизвестны (114).

§ 7. Итерационные методы вариационного типа 115
1. Метод минимальных невязок (116). 2. Метод минимальных поправок (118). 3. Метод скорейшего спуска (119). 4. Метод сопряженных градиентов (120). 5. Минимизация погрешности (121). 6. Выбор итерационных параметров в методе сопряженных градиентов (122). 7 Оценка погрешности в методе сопряженных градиентов (126).

Глава 3. **Интерполирование и приближение функций** 127

§ 1. Интерполирование алгебраическими многочленами 127
1. Интерполяционная формула Лагранжа (127). 2. Интерполяционная формула Ньютона (129).

§ 2. Погрешность интерполирования 132
1. Остаточный член интерполяционной формулы (132). 2. Оптимальный выбор узлов интерполирования (134). 3. О сходимости интерполяционного процесса (134).

§ 3. Интерполирование с кратными узлами 136
1. Интерполяционный многочлен Эрмита (136). 2. Пример (138).

§ 4. Интерполирование сплайнами 140
1. Построение кубического сплайна (141). 2. Сходимость процесса интерполирования кубическими сплайнами (143).

§ 5. Другие постановки задач интерполирования и приближения функций 148
1. Примеры (148). 2. Общая постановка задачи интерполирования (151). 3. Наилучшее приближение функции, заданной таблично (152). 4. Сглаживание сеточных функций (154).

§ 6. Наилучшие приближения в гильбертовом пространстве 156
1. Постановка задачи (156). 2. Сведение к алгебраической задаче о минимуме квадратичного функционала (157). 3. Следствия (159).

Глава 4. **Численное интегрирование и дифференцирование** . . . 161

§ 1. Примеры формул численного интегрирования 161
1. Введение (161). 2. Формула прямоугольников (162). 3. Формула трапеций (164). 4. Формула Симпсона (165). 5. Апостериорная оценка погрешности методом Рунге. Автоматический выбор шага интегрирования (168). 6. Экстраполяция Ричардсона (169).

§ 2. Квадратурные формулы интерполяционного типа 172
1. Вывод формул (172). 2. Оценка погрешности (174). 3. Симметричные формулы (175). 4. Формулы Ньютона — Котеса. Численная устойчивость квадратурных формул (178).

§ 3. Метод Гаусса вычисления определенных интегралов 180
1. Постановка задачи (180). 2. Основная теорема (181). 3. Существование и единственность квадратурных формул наивысшей алгебраической степени точности (183). 4. Свойства квадратурных формул Гаусса (184). 5. Частный случай формул Гаусса (185).

§ 4. Численное дифференцирование 186
1. Некорректность операции численного дифференцирования (186). 2. Применение интерполирования (188).

Глава 5. **Решение нелинейных уравнений и систем уравнений** . . . 190
§ 1. Примеры итерационных методов решения нелинейных уравнений . 190
1. Введение (190). 2. Метод простой итерации (191). 3. Метод Ньютона (193). 4 Метод секущих (194). 5. Интерполяционные методы (194). 6. Использование обратной интерполяции (195).
§ 2. Сходимость метода простой итерации 195
1. Теорема о сходимости (195). 2. Метод Эйткена ускорения сходимости (198).
§ 3. Сходимость метода Ньютона 199
1. Простой вещественный корень (199). 2. Кратные корни (202). 3. Односторонние приближения (203). 4 Комплексный корень (205).
§ 4. Итерационные методы для систем нелинейных уравнений . . . 207
1. Общие понятия (207). 2. Сходимость стационарного метода (208). 3. Примеры итерационных методов (209).

Глава 6. **Численные методы решения задачи Коши для обыкновенных дифференциальных уравнений** 214
§ 1. Исходная задача и примеры численных методов ее решения . . 214
1. Постановка исходной задачи (214). 2. Примеры численных методов (214).
§ 2. Методы Рунге — Кутта 218
1. Общая формулировка методов. Семейство методов второго порядка (218). 2. Доказательство сходимости (221). 3. Методы третьего порядка точности (224). 4. Методы четвертого порядка точности (226).
§ 3. Многошаговые разностные методы 230
1. Формулировка методов (230). 2 Погрешность аппроксимации многошаговых методов (231). 3. Устойчивость и сходимость разностных методов (233). 4. Примеры многошаговых разностных методов (235).
§ 4. Сходимость и оценка погрешности многошагового разностного метода 236
1. Уравнение для погрешности (236). 2. Однородное разностное уравнение с постоянными коэффициентами. Частные решения (238). 3. Однородное разностное уравнение с постоянными коэффициентами. Устойчивость по начальным данным (240). 4. Оценка решения неоднородного уравнения (243). 5. Оценки погрешности разностного метода (244).
§ 5. Численное интегрирование жестких систем обыкновенных дифференциальных уравнений 247
1 Условно устойчивые и абсолютно устойчивые разностные методы (247). 2. Понятие жесткой системы дифференциальных уравнений (249). 3. Нелинейные системы дифференциальных уравнений (251). 4. Специальные определения устойчивости (252). 5. Чисто неявные разностные методы (255).

ЧАСТЬ III
РАЗНОСТНЫЕ МЕТОДЫ РЕШЕНИЯ ЗАДАЧ МАТЕМАТИЧЕСКОЙ ФИЗИКИ

Глава 1. **Вводные понятия** 259
§ 1. Примеры разностных аппроксимаций 259
§ 2. Построение разностных схем интегро-интерполяционным методом . 262
1. Построение разностной схемы (262).
§ 3. Исследование аппроксимации и сходимости 265
1. Аппроксимация дифференциального уравнения (265). 2 Аппроксимация граничного условия (267). 3. Уравнение для погрешности (268). 4. Разностные тождества и неравенства (269). 5. Доказательство сходимости (270).
§ 4. Разностные схемы для уравнения теплопроводности 272
1. Исходная задача (272). 2. Явная схема (272). 3. Неявные схемы (276). 4 Уравнения с переменными коэффициентами и нелинейные уравнения (279).
§ 5. Трехслойные разностные схемы 283
1 Разностные схемы для уравнения колебаний (283). 2 Трехслойные схемы для уравнения теплопроводности (285).
§ 6. Основные понятия теории разностных схем: аппроксимация, сходимость, устойчивость 286
1 Введение (286). 2. Погрешность аппроксимации и погрешность схемы (287). 3. Корректность разностной схемы. Сходимость. Связь между устойчивостью и сходимостью (290).

Глава 2. **Принцип максимума для разностных схем** 291
§ 1. Разностная аппроксимация задачи Дирихле для уравнения Пуассона 291
1. Постановка разностной задачи (291). 2. Канонический вид разностного уравнения (293).
§ 2. Принцип максимума для разностных схем. Основные теоремы . . 294
1. Исходные предположения (294). 2. Принцип максимума и его следствия (295). 3. Теорема сравнения Устойчивость по граничным условиям (298). 4. Примеры (299).
§ 3. Доказательство устойчивости и сходимости разностной задачи Дирихле для уравнения Пуассона 300
1. Устойчивость по граничным условиям (300). 2. Устойчивость по правой части и сходимость (302).
§ 4. Примеры применения принципа максимума 304
§ 5. Монотонные разностные схемы для уравнений второго порядка, содержащих первые производные 308

Глава 3. **Метод разделения переменных** 310
§ 1. Разностная задача на собственные значения 311
1. Оператор второй разностной производной (311). 2. Задача на собственные значения (312). 3. Свойства собственных значений и собственных функций (313). 4 Операторные неравенства (315).
§ 2. Задача на собственные значения для пятиточечного разностного оператора Лапласа 317
1. Самосопряженность (317). 2. Оценка собственных чисел. Положительность оператора (318).
§ 3. Исследование устойчивости и сходимости схемы с весами для уравнения теплопроводности 320
1. Исходная задача и разностная схема (320). 2. Устойчивость схемы по начальным данным (322). 3. Устойчивость по правой части и сходимость (324). 4 Схема с весами для двумерного уравнения теплопроводности (326). 5. Асимптотическая устойчивость (328).
§ 4. Решение разностного уравнения второго порядка методом Фурье . 332
§ 5. Быстрое дискретное преобразование Фурье 334
§ 6. Решение разностного уравнения Пуассона с использованием быстрого преобразования Фурье 337

Глава 4. **Теория устойчивости разностных схем** 339
§ 1. Разностные схемы как операторные уравнения 339
1. Представление разностных схем в виде операторных уравнений (339). 2. Корректность операторных уравнений (342). 3. Операторы первой разностной производной (347)
§ 2. Канонический вид и условия устойчивости двуслойных разностных схем 349
1. Канонический вид двуслойных разностных схем (349). 2. Устойчивость разностных схем (351). 3. Теоремы об устойчивости по начальным данным (354). 4. Несамосопряженные разностные схемы (359).
§ 3. Канонический вид и условия устойчивости трехслойных разностных схем 362
1 Канонический вид (362). 2. Эквивалентность трехслойной схемы двуслойной (363). 3. Устойчивость по начальным данным (364). 4. Примеры (366).
§ 4. Об экономичных методах решения многомерных нестационарных задач математической физики 369
1. Недостатки обычных разностных методов (369). 2. Пример метода переменных направлений (372). 3 Абсолютная устойчивость продольно-поперечной схемы (373). 4 Понятие суммарной аппроксимации (376).

Глава 5. **Прямые и итерационные методы решения сеточных уравнений** 378
§ 1. Модельная задача 378
1. Введение (378) 2 Модельная задача (379). 3. Применение методов Якоби и Зейделя (381). 4. Метод верхней релаксации (384).
§ 2. Применение явного итерационного метода с оптимальным набором параметров 389
1. Явный итерационный метод с чебышевскими параметрами (389). 2 Применение к модельной задаче (390). 3. Применение чебышевского метода к разностным аппроксимациям уравнений эллиптического типа (391).

§ 3. Попеременно-треугольный итерационный метод 394
1. Алгебраическая теория (394). 2. Применение к модельной задаче (398). 3. Попеременно-треугольный метод с чебышевскими итерационными параметрами (401). 4. Модифицированный попеременно-треугольный итерационный метод (402).
§ 4. Итерационный метод переменных направлений 404
1. Формулировка метода и исследование сходимости (404). 2. Пример (406). 3. Случай прямоугольной области (408).
§ 5. Метод матричной прогонки 411
1. Введение (411). 2. Запись разностного уравнения Пуассона в виде системы векторных уравнений (412). 3 Алгоритм матричной прогонки (414). 4. Устойчивость матричной прогонки (415).
§ 6. Метод редукции 418
1. Вывод основных формул (418). 2. Обращение матриц (421). 3. Вычисление правых частей (423). 4 Формулировка и обсуждение алгоритма (424).

Список литературы 426

Предметный указатель 428

ПРЕДИСЛОВИЕ

В книге излагаются основы численных методов решения задач алгебры, анализа, обыкновенных дифференциальных уравнений и уравнений математической физики. Книга предназначена для студентов вузов, специализирующихся в области прикладной математики. Она может оказаться полезной также студентам других специальностей, желающим получить представление о методах решения математических задач с помощью ЭВМ. Книга основана на курсе лекций, который читался в течение ряда лет студентам факультета вычислительной математики и кибернетики Московского университета.

В курсах численных методов изучаются вопросы построения, применения и теоретического обоснования алгоритмов приближенного решения различных классов математических задач. В настоящее время большинство вычислительных алгоритмов ориентировано на использование быстродействующих ЭВМ, что существенно влияет на отбор учебного материала и на характер его изложения. Следует отметить некоторые особенности предмета численных методов. Во-первых, для численных методов характерна множественность, т. е. возможность решить одну и ту же задачу различными методами. Во-вторых, вновь возникающие естественно-научные задачи и быстрое развитие вычислительной техники вынуждают переоценивать значение существующих алгоритмов и приводят к созданию новых. Перечисленные особенности предмета, его обширность и неоднородность делают иллюзорной попытку изложить предмет «во всей полноте и строгости». Поэтому авторы настоящей книги поставили перед собой задачу собрать минимальный материал, достаточный для дальнейшей работы выпускников вузов в области применения и создания вычислительных методов.

Вычислительный алгоритм естественно рассматривать как необходимую составную часть вычислительного эксперимента — эффективного метода решения крупных естественно-научных и народнохозяйственных задач. С этих позиций и ведется изложение численных методов в данной книге. Рассматриваются только те

методы, которые выдержали испытание практикой и применяются для решения реальных задач. Наибольшее внимание уделяется фундаментальным разделам численных методов — численному решению систем линейных алгебраических уравнений и разностным методам решения задач математической физики. В то же время авторы сознают, что многие интересные и важные методы изложены недостаточно полно или совсем не вошли в книгу. За рамками книги остались такие этапы вычислительного эксперимента, как построение математической модели, программирование и организация вычислений. В тех случаях, когда подробное изложение численного метода оказывалось слишком громоздким, содержало много выкладок или опиралось на труднодоступный студентам математический аппарат, авторы предпочитали ограничиться характерными примерами.

Книга состоит из трех частей. Часть I является вводной, в ней дается представление о месте численных методов в общем процессе математического моделирования и вычислительного эксперимента, а также рассматриваются на уровне примеров некоторые вычислительные алгоритмы. В части II излагаются традиционные разделы численных методов, такие как прямые и итерационные методы решения систем линейных алгебраических уравнений, интерполирование, численное интегрирование, решение нелинейных уравнений, методы решения задачи Коши для обыкновенных дифференциальных уравнений. Может возникнуть вопрос, зачем нужно столь подробно излагать методы, для большинства из которых уже давно существует хорошо зарекомендовавшая себя программная реализация? Дело в том, что сознательное использование существующих программ и тем более создание новых улучшенных версий вряд ли возможно без изучения самих методов и связанных с ними теоретических представлений. В части III рассматриваются разностные методы решения задач математической физики. Здесь большое внимание уделяется принципам построения разностных схем для различных задач, исследованию их устойчивости и сходимости, методам решения сеточных уравнений.

Для чтения части II требуется знание алгебры, анализа и обыкновенных дифференциальных уравнений в объеме одного-двух курсов вузовского обучения. Часть III предполагает знакомство с постановкой типичных задач математической физики. Каких-либо специальных предварительных сведений из области вычислительной математики не требуется, хотя могут оказаться полезными отдельные главы из книг

Тихонов А. Н., Костомаров Д. П. Вводные лекции по прикладной математике.— М.: Наука, 1984.

Самарский А. А. Введение в численные методы.— 2-е изд.— М.: Наука, 1987.

Предполагается, что одновременно с изучением данного курса читатель овладевает навыками решения задач с помощью ЭВМ, а также участвует в работе студенческого семинара по численным методам.

Более подробное изложение отдельных разделов курса можно найти в книгах:

Самарский А. А. Теория разностных схем.— 2-е изд.— М.: Наука, 1983.

Самарский А. А., Николаев Е. С. Методы решения сеточных уравнений.— М.: Наука, 1978.

Самарский А. А., Попов Ю. П. Разностные методы решения задач газовой динамики.— 2-е изд.— М.: Наука, 1980.

Авторы приносят глубокую благодарность декану факультета вычислительной математики и кибернетики МГУ академику А. Н. Тихонову, при активном участии которого обсуждались вопросы преподавания численных методов.

Считаем также своим приятным долгом выразить благодарность нашим товарищам и сотрудникам по работе В. Б. Андрееву, Т. Н. Галишниковой, Л. М. Дегтяреву, Н. И. Ионкину, Н. Н. Калиткину, Д. П. Костомарову, Е. С. Николаеву, Ю. П. Попову, А. П. Фаворскому, И. В. Фрязинову за полезное обсуждение и сделанные замечания по содержанию книги.

А. А. Самарский, А. В. Гулин

ЧАСТЬ I
ВВЕДЕНИЕ В ЧИСЛЕННЫЕ МЕТОДЫ

§ 1. Математическое моделирование и вычислительный эксперимент

1. Схема вычислительного эксперимента. Эффективное решение крупных естественно-научных и народнохозяйственных задач сейчас невозможно без применения быстродействующих электронно-вычислительных машин (ЭВМ). В настоящее время выработалась технология исследования сложных проблем, основанная на построении и анализе с помощью ЭВМ математических моделей изучаемого объекта. Такой метод исследования называют *вычислительным экспериментом.*

Пусть, например, требуется исследовать какой-то физический объект, явление, процесс. Тогда схема вычислительного эксперимента выглядит так, как показано на рис. 1. Формулируются основные законы, управляющие данным объектом исследования (I) и строится соответствующая *математическая модель* (II), представляющая обычно запись этих законов в форме системы уравнений (алгебраических, дифференциальных, интегральных и т. д.).

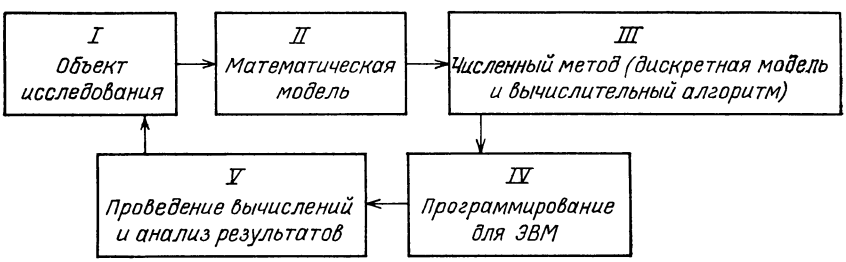


Рис. 1. Схема вычислительного эксперимента

При выборе физической и, следовательно, математической модели мы пренебрегаем факторами, не оказывающими существенного влияния на ход изучаемого процесса. Типичные математические модели, соответствующие физическим явлениям, формулируются в виде уравнений математической физики. Большинство реальных процессов описывается нелинейными уравнениями и лишь в первом приближении (при малых значениях параметров, малых отклонениях от равновесия и др.) эти уравнения можно заменить линейными.

После того как задача сформулирована в математической форме, необходимо найти ее решение. Но что значит решить математическую задачу? Только в исключительных случаях удается найти решение в явном виде, например в виде ряда. Иногда утверждение «задача решена» означает, что доказано существование и единственность решения. Ясно, что этого недостаточно для практических приложений. Необходимо еще изучить качественное поведение решения и найти те или иные количественные характеристики.

Именно на этом этапе требуется привлечение ЭВМ и, как следствие, развитие численных методов (см. III на рис. 1). Под *численным методом* здесь понимается такая интерпретация математической модели («дискретная модель»), которая доступна для реализации на ЭВМ. Например, если математическая модель представляет собой дифференциальное уравнение, то численным методом может быть аппроксимирующее его разностное уравнение совместно с алгоритмом, позволяющим отыскать решение этого разностного уравнения. Результатом реализации численного метода на ЭВМ является число или таблица чисел. Отметим, что в настоящее время помимо собственно численных методов имеются также методы, которые позволяют проводить на ЭВМ аналитические выкладки. Однако аналитические методы для ЭВМ не получили пока достаточно широкого распространения.

Чтобы реализовать численный метод, необходимо составить программу для ЭВМ (см. IV на рис. 1) или воспользоваться готовой программой. После отладки программы наступает этап проведения вычислений и анализа результатов (V). Полученные результаты изучаются с точки зрения их соответствия исследуемому явлению и при необходимости вносятся исправления в численный метод и уточняется математическая модель.

Такова в общих чертах схема вычислительного эксперимента. Его основу составляет триада: *модель — метод* (*алгоритм*) *— программа*. Опыт решения крупных задач показывает, что метод математического моделирования и вычислительный эксперимент соединяют в себе преимущества традиционных теоретических и экспериментальных методов исследования. Можно указать такие крупные области применения вычислительного эксперимента, как энергетика, аэрокосмическая техника, обработка данных натурного эксперимента, совершенствование технологических процессов.

2. Вычислительный алгоритм. Предметом данной книги является изложение вопросов, отражающих лишь один из этапов вычислительного эксперимента, а именно этап построения и исследования численного метода. Таким образом, здесь не обсуждаются исходные задачи и их математическая постановка, не рассматриваются вопросы программирования и организации вычислений, интерпретации результатов. Предварительные понятия о проблематике математического моделирования и вычислительного эксперимента читатель может получить из книг [36, 40].

Необходимо подчеркнуть, что процесс исследования исходного объекта методом математического моделирования и вычислительного эксперимента неизбежно носит приближенный характер, потому что на каждом этапе вносятся те или иные погрешности. Так, построение математической модели связано с упрощением исходного явления, недостаточно точным заданием коэффициентов уравнения и других входных данных. По отношению к численному методу, реализующему данную математическую модель, указанные погрешности являются *неустранимыми*, поскольку они неизбежны в рамках данной модели.

При переходе от математической модели к численному методу возникают погрешности, называемые *погрешностями метода*. Они связаны с тем, что всякий численный метод воспроизводит исходную математическую модель приближенно. Наиболее типичными погрешностями метода являются *погрешность дискретизации* и *погрешность округления*. Поясним причины возникновения таких погрешностей.

Обычно построение численного метода для заданной математической модели разбивается на два этапа: а) формулировка дискретной задачи, б) разработка вычислительного алгоритма, позволяющего отыскать решение дискретной задачи. Например, если исходная математическая задача сформулирована в виде системы дифференциальных уравнений, то для численного решения необходимо заменить ее системой конечного, может быть, очень большого числа линейных или разностных алгебраических уравнений. В этом случае говорят, что проведена *дискретизация исходной математической задачи*. Простейшим примером дискретизации является построение *разностной схемы* путем замены дифференциальных выражений конечно-разностными отношениями. В общем случае дискретную модель можно рассматривать как конечномерный аналог исходной математической задачи. Ясно, что решение дискретизированной задачи отличается от решения исходной задачи. Разность соответствующих решений и называется *погрешностью дискретизации*. Обычно дискретная модель зависит от некоторого параметра (или множества параметров) дискретизации, при стремлении которого к нулю должна стремиться к нулю и погрешность дискретизации. При этом число алгебраических уравнений, составляющих дискретную модель, неограниченно возрастает. В случае разностных методов таким параметром является шаг сетки.

Как уже отмечалось, дискретная модель представляет собой систему большого числа алгебраических уравнений. Невозможно найти решение такой системы точно и в явном виде. Поэтому приходится использовать тот или иной численный алгоритм решения системы алгебраических уравнений. Входные данные этой системы, а именно коэффициенты и правые части, задаются в ЭВМ не точно, а с округлением. В процессе работы алгоритма погрешности округления обычно накапливаются, и в результате решение, полученное на ЭВМ, будет отличаться от точного решения дискрети-

зированной задачи. Результирующая погрешность называется *погрешностью округления* (иногда ее называют *вычислительной погрешностью*). Величина этой погрешности определяется двумя факторами: точностью представления вещественных чисел в ЭВМ и чувствительностью данного алгоритма к погрешностям округления.

Алгоритм называется *устойчивым*, если в процессе его работы вычислительные погрешности возрастают незначительно, и *неустойчивым* — в противоположном случае. При использовании неустойчивых вычислительных алгоритмов накопление погрешностей округления приводит в процессе счета к переполнению арифметического устройства ЭВМ.

Итак, следует различать погрешности модели, метода и вычислительную. Какая же из этих трех погрешностей является преобладающей? Ответ здесь неоднозначен. Видимо, типичной является ситуация, возникающая при решении задач математической физики, когда погрешность модели значительно превышает погрешность метода, а погрешностью округления в случае устойчивых алгоритмов можно пренебречь по сравнению с погрешностью метода. С другой стороны, при решении, например, систем обыкновенных дифференциальных уравнений возможно применение столь точных методов, что их погрешность будет сравнима с погрешностью округления. В общем случае нужно стремиться, чтобы все указанные погрешности имели один и тот же порядок. Например, нецелесообразно пользоваться разностными схемами, имеющими точность 10^{-6}, если коэффициенты исходных уравнений задаются с точностью 10^{-2}.

3. Требования к вычислительным методам. Одной и той же математической задаче можно поставить в соответствие множество различных дискретных моделей. Однако далеко не все из них пригодны для практической реализации. Вычислительные алгоритмы, предназначенные для быстродействующих ЭВМ, должны удовлетворять многообразным и зачастую противоречивым требованиям. Попытаемся здесь сформулировать основные из этих требований в общих чертах. Далее в частях II и III книги эти требования конкретизируются при рассмотрении алгоритмов численного решения типичных математических задач.

Можно выделить две группы требований к численным методам. Первая группа связана с адекватностью дискретной модели исходной математической задаче, и вторая группа — с реализуемостью численного метода на ЭВМ. К первой группе относятся такие требования, как сходимость численного метода, выполнение дискретных аналогов законов сохранения, качественно правильное поведение решения дискретной задачи.

Поясним эти требования. Предположим, что дискретная модель математической задачи представляет собой систему большого, но конечного числа алгебраических уравнений. Обычно, чем точнее мы хотим получить решение, тем больше уравнений приходится брать. Говорят, что численный метод *сходится*, если при неограни-

ченном увеличении числа уравнений решение дискретной задачи стремится к решению исходной задачи.

Поскольку реальная ЭВМ может оперировать лишь с конечным числом уравнений, на практике сходимость, как правило, не достигается. Поэтому важно уметь оценивать погрешность метода в зависимости от числа уравнений, составляющих дискретную модель. По этой же причине стараются строить дискретную модель таким образом, чтобы она правильно отражала качественное поведение решения исходной задачи даже при сравнительно небольшом числе уравнений.

Например, дискретной моделью задачи математической физики может быть разностная схема. Для ее построения область изменения независимых переменных заменяется дискретным множеством точек — *сеткой*, а входящие в исходное уравнение производные заменяются на сетке конечно-разностными отношениями. В результате получаем систему алгебраических уравнений относительно значений искомой функции в точках сетки. Число уравнений этой системы равно числу точек сетки. Известно, что дифференциальные уравнения математической физики являются следствиями интегральных законов сохранения. Поэтому естественно требовать, чтобы для разностной схемы выполнялись аналоги таких законов сохранения. Разностные схемы, удовлетворяющие этому требованию, называются *консервативными*. Оказалось, что при одном и том же числе точек сетки консервативные разностные схемы более правильно отражают поведение решения исходной задачи, чем неконсервативные схемы.

Сходимость численного метода тесно связана с его корректностью. Предположим, что исходная математическая задача поставлена корректно, т. е. ее решение существует, единственно и непрерывно зависит от входных данных. Тогда дискретная модель этой задачи должна быть построена таким образом, чтобы свойство корректности сохранилось. Таким образом, в понятие *корректности численного метода* включаются свойства однозначной разрешимости соответствующей системы уравнений и ее устойчивости по входным данным. Под *устойчивостью* понимается непрерывная зависимость решения от входных данных, равномерная относительно числа уравнений, составляющих дискретную модель.

Вторая группа требований, предъявляемых к численным методам, связана с возможностью реализации данной дискретной модели на данной ЭВМ, т. е. с возможностью получить на ЭВМ решение соответствующей системы алгебраических уравнений за приемлемое время. Основным препятствием для реализации корректно поставленного алгоритма является ограниченный объем оперативной памяти ЭВМ и ограниченные ресурсы времени счета. Реальные вычислительные алгоритмы должны учитывать эти обстоятельства, т. е. они должны быть экономичными как по числу арифметических действий, так и по требуемому объему памяти.

§ 2. Погрешности округления

1. Представление вещественных чисел в ЭВМ. Одним из источников вычислительных погрешностей является приближенное представление вещественных чисел в ЭВМ, обусловленное конечностью разрядной сетки. Хотя исходные данные представляются в ЭВМ с большой точностью, накопление погрешностей округления в процессе счета может привести к значительной результирующей погрешности, а некоторые алгоритмы могут оказаться и вовсе непригодными для реального счета на ЭВМ.

Напомним о способах представления чисел в ЭВМ и связанных с ними погрешностях округления. Более подробно этот круг вопросов рассматривается в [6, 8, 15, 29].

При ручном счете пользуются десятичной системой счисления. Например, запись 103,67 определяет число

$$1\cdot 10^2+0\cdot 10^1+3\cdot 10^0+6\cdot 10^{-1}+7\cdot 10^{-2}.$$

Здесь 10 — основание системы счисления, запятая отделяет дробную часть числа от целой, 1, 0, 3, 6, 7 — числа из базисного набора {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}, с помощью которого можно представить любое вещественное число.

ЭВМ работают, как правило, в двоичной системе, когда любое число записывается в виде последовательности нулей и единиц. Например, запись 0, 0101 в двоичной системе определяет число

$$0\cdot 2^0+0\cdot 2^{-1}+1\cdot 2^{-2}+0\cdot 2^{-3}+1\cdot 2^{-4}.$$

Как двоичная, так и десятичная системы относятся к позиционным системам счисления. В *позиционной системе с основанием* r запись

$$a=\pm a_n a_{n-1}\ldots a_0,\ a_{-1}a_{-2}\ldots \tag{1}$$

означает, что

$$a=\pm\left(a_n r^n+a_{n-1}r^{n-1}+\ldots+a_0 r^0+a_{-1}r^{-1}+a_{-2}r^{-2}+\ldots\right).$$

Будем считать далее, что r — целое число, большее единицы. Каждое из чисел a_i может принимать одно из значений $\{0, 1, \ldots, r-1\}$. Числа a_i называются *разрядами*, например: a_3 — третий разряд перед запятой, a_{-2} — второй разряд после запятой.

Запись вещественного числа в виде (1) называется также его представлением в форме *числа с фиксированной запятой*. В ЭВМ чаще всего используется представление чисел в форме *с плавающей запятой*, т. е. в виде

$$a=Mr^p, \tag{2}$$

где r — основание системы счисления, p — целое число (положительное, отрицательное или нуль) и

$$r^{-1}\leqslant|M|<1. \tag{3}$$

Число M представляется в форме числа с фиксированной запятой и называется *мантиссой числа* a. Число p называется *порядком числа* a. В виде (2) можно единственным образом представить

любое вещественное число кроме нуля. Единственность обеспечивается условием нормировки (3).

Например, число 103,67 в форме с плавающей запятой имеет вид $0{,}10367 \cdot 10^3$, т. е. $M=0{,}10367$, $p=3$. Двоичное число $0{,}0101 = 0{,}101 \cdot 2^{-1}$ имеет в двоичной системе мантиссу $M=0{,}101$ и порядок $p=-1$.

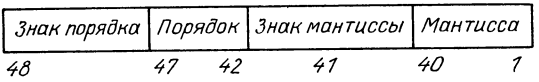


Рис. 2. Разрядная сетка

В ЭВМ для записи каждого числа отводится фиксированное число разрядов (*разрядная сетка*). Например, в ЭВМ БЭСМ-6 для записи числа, представленного в форме с плавающей запятой, отводится 48 двоичных разрядов, которые распределяются следующим образом: в разрядах с 1 по 40 помещается абсолютное значение мантиссы, в 41 разряде — знак мантиссы, в разрядах от 42 до 47 — абсолютная величина порядка, в 48 разряде — знак порядка (см. рис. 2). Отсюда легко найти диапазон чисел, представимых в ЭВМ БЭСМ-6. Поскольку максимальное значение порядка в двоичной системе равно 111 111=63 и мантисса не превосходит единицы, то с помощью указанной разрядной сетки можно представить числа, абсолютная величина которых лежит примерно в диапазоне от 2^{-63} до 2^{63}, т. е. от 10^{-19} до 10^{19}.

Ту же 48-разрядную сетку можно использовать для представления чисел с фиксированной запятой. Пусть, например, разряды с 1 по 24 отводятся для записи дробной части числа и разряды с 25 по 47 — для записи целой части числа. Тогда максимальное число, которое можно представить с помощью данной разрядной сетки, будет равно

$$\underbrace{11\ldots1}_{\text{23 разряда}},\ \underbrace{11\ldots1}_{\text{24 разряда}} < 2^{23} \approx 10^7.$$

Следовательно, в данном случае диапазон допустимых чисел в 10^{12} раз меньше, чем при использовании представления с плавающей запятой. Возможностью существенного увеличения диапазона допустимых чисел при той же разрядной сетке и объясняется преимущественное использование в ЭВМ представления чисел в форме с плавающей запятой. Комплексное число представляется в ЭВМ в виде пары вещественных чисел.

2. Округление чисел в ЭВМ. Будем считать в дальнейшем, что вещественные числа представляются в ЭВМ в форме с плавающей запятой. Минимальное положительное число M_0, которое может быть представлено в ЭВМ с плавающей запятой, называется *машинным нулем*. Мы видим, что для ЭВМ БЭСМ-6 число $M_0 \approx 10^{-19}$. Число $M_\infty = M_0^{-1}$ называется *машинной бесконечностью*. Все вещественные числа, которые могут быть представлены в данной ЭВМ, расположены по абсолютной величине в диапазоне от M_0 до M_∞. Если в процессе счета какой-либо задачи появится вещест-

венное число, меньшее по модулю чем M_0, то ему присваивается нулевое значение. Так, на ЭВМ БЭСМ-6 в результате перемножения двух чисел 10^{-11} и 10^{-10} получим нуль. При появлении в процессе счета вещественного числа, большего по модулю чем M_∞, происходит так называемое переполнение разрядной сетки, после чего ЭВМ прекращает счет задачи. Отметим, что нуль и целые числа представляются в ЭВМ особым образом, так что они могут выходить за пределы диапазона $M_0 \div M_\infty$.

Из-за конечности разрядной сетки в ЭВМ можно представить точно не все числа из диапазона $M_0 \div M_\infty$, а лишь конечное множество чисел. Число a, не представимое в ЭВМ точно, подвергается *округлению*, т. е. оно заменяется близким ему числом $\tilde{a}$, представимым в ЭВМ точно. Точность представления чисел в ЭВМ с плавающей запятой характеризуется *относительной погрешностью*

$$|a-\tilde{a}|/|a|.$$

Величина относительной погрешности зависит от способа округления. Простейшим, но не самым точным способом округления является отбрасывание всех разрядов мантиссы числа a, которые выходят за пределы разрядной сетки.

Найдем границу относительной погрешности при таком способе округления. Пусть для записи мантиссы в ЭВМ отводится t двоичных разрядов. Предположим, что надо записать число, представленное в виде бесконечной двоичной дроби

$$a = \pm 2^p \left(\frac{a_1}{2} + \frac{a_2}{2^2} + \ldots + \frac{a_t}{2^t} + \frac{a_{t+1}}{2^{t+1}} + \ldots \right), \tag{4}$$

где каждое из a_j равно 0 или 1. Отбрасывая все лишние разряды, получим округленное число

$$\tilde{a} = \pm 2^p \left(\frac{a_1}{2} + \frac{a_2}{2^2} + \ldots + \frac{a_t}{2^t} \right).$$

Таким образом, для погрешности округления

$$a - \tilde{a} = \pm 2^p \left(\frac{a_{t+1}}{2^{t+1}} + \frac{a_{t+2}}{2^{t+2}} + \ldots \right)$$

справедлива оценка

$$|a-\tilde{a}| \leqslant 2^p \frac{1}{2^{t+1}} \left(1 + \frac{1}{2} + \frac{1}{2^2} + \ldots \right) = 2^{p-t}.$$

Далее заметим, что из условия нормировки $|M| \geqslant 0{,}5$ (см. (3)) следует, что в разложении (4) всегда $a_1 = 1$. Поэтому $|a| \geqslant 2^p \cdot 2^{-1} = 2^{p-1}$, и для относительной погрешности округления получим оценку

$$\frac{|a-\tilde{a}|}{|a|} \leqslant 2^{-t+1}.$$

При более точных способах округления можно уменьшить погрешность по крайней мере в два раза и добиться, чтобы выполнялась оценка

$$\frac{|a-\tilde{a}|}{|a|} \leqslant 2^{-t}. \tag{5}$$

Итак, *относительная точность в ЭВМ с плавающей запятой определяется числом разрядов t, отводимых для записи мантиссы. Можно считать, что точное число a и отвечающее ему округленное число $\tilde{a}$ связаны равенством*

$$\tilde{a}=a(1+\varepsilon), \tag{6}$$

где $|\varepsilon|\leqslant 2^{-t}$. Число 2^{-t} называют иногда *машинным эпсилоном.* Оно характеризует относительную точность представления чисел в ЭВМ. Для ЭВМ БЭСМ-6 имеем $t=40$, $2^{-t}\approx 10^{-12}$, т. е. относительная точность представления чисел составляет 12 десятичных знаков.

Соотношение (6) справедливо лишь в случае $|a|\geqslant M_0$, где M_0 — машинный нуль. Если же число a мало, а именно $|a|<M_0$, то полагается $\tilde{a}=0$, что соответствует $\varepsilon=-1$ в формуле (6).

3. Накопление погрешностей округления. В процессе проведения вычислений погрешности округления могут накапливаться, так как выполнение каждой из четырех арифметических операций вносит некоторую погрешность.

Будем в дальнейшем обозначать округленное в системе с плавающей запятой число, соответствующее точному числу x, через $\mathrm{fl}(x)$ (от английского floating — плавающий). Считается, что выполнение каждой арифметической операции вносит относительную погрешность не большую, чем 2^{-t}. Это предположение можно записать в виде

$$\mathrm{fl}(a*b)=a*b(1+\varepsilon), \tag{7}$$

где звездочка означает любую из операций $+$, $-$, $\times$, $:$, и $|\varepsilon|\leqslant 2^{-t}$. Если результат выполнения арифметической операции является машинным нулем, то в формуле (7) надо положить $\varepsilon=-1$.

Может показаться, что предположение (7) не обосновано, так как согласно (6) каждое из чисел a и b записывается с относительной погрешностью 2^{-t}, следовательно, погрешность результата может достигнуть 2^{-t+1}. Однако ЭВМ обладает возможностью проводить промежуточные вычисления с двойной точностью, т. е. с мантиссой, содержащей $2t$ разрядов, причем округлению до t разрядов подвергается лишь окончательный результат. Это обстоятельство позволяет добиться выполнения соотношения (7).

Для оценки влияния погрешностей округления на результат того или иного вычислительного алгоритма очень часто используется предположение о том, что *результат вычислений, искаженный погрешностями округления, совпадает с результатом точного выполнения этого же алгоритма, но с иными входными данными.*

Рассмотрим, например, процесс вычисления суммы

$$z=y_1+y_2+y_3 \tag{8}$$

трех положительных чисел. Пусть сначала находится сумма y_1+y_2 Тогда согласно (7) получим

$$z_1=\mathrm{fl}(y_1+y_2)=(y_1+y_2)(1+\varepsilon_1), \qquad |\varepsilon_1|\leqslant 2^{-t}.$$

Затем в результате сложения z_1 и y_3 получим число

$$\tilde{z}=\mathrm{fl}(z_1+y_3)=(z_1+y_3)(1+\varepsilon_2),$$

где $|\varepsilon_2| \leqslant 2^{-t}$. Таким образом, вместо точного значения суммы z получаем приближенное значение

$$\tilde{z}=(y_1+y_2)(1+\varepsilon_1)(1+\varepsilon_2)+y_3(1+\varepsilon_2).$$

Отсюда видно, что результат выполнения алгоритма (8), искаженный погрешностями округления, совпадает с результатом точного выполнения того же алгоритма (8), примененного к другим исходным данным

$$\tilde{y}_i=(1+\varepsilon_1)(1+\varepsilon_2)y_i, \quad i=1, 2, \qquad \tilde{y}_3=(1+\varepsilon_2)y_3.$$

На этом же примере видно, что результирующая погрешность зависит от порядка выполнения операций, так что вычисление суммы (8) в обратном порядке $(y_3+y_2)+y_1$ может привести к другому результату.

Приведенный пример имеет чисто иллюстративное значение, так как число слагаемых в сумме (8) невелико, а погрешности ε_i малы. Практический интерес представляют оценки результирующей погрешности в зависимости от числа выполненных арифметических действий n. Однако прежде чем перейти к получению таких оценок, необходимо познакомиться с методами решения разностных уравнений.

4. Разностные уравнения первого порядка. Предположим, что надо вычислить сумму

$$z_n=\sum_{j=1}^{n} y_j. \tag{9}$$

Тогда вычисления организуются обычно следующим образом. Задается начальное значение $z_0=0$ и затем последовательно, начиная с $j=1$, находятся числа z_j, связанные рекуррентным соотношением

$$z_j=z_{j-1}+y_j, \quad j=1, 2, \ldots, n, \quad z_0=0. \tag{10}$$

Для вычисления произведения

$$z_n=\prod_{j=1}^{n} y_j \tag{11}$$

достаточно задать начальное значение $z_0=1$ и воспользоваться рекуррентными соотношениями

$$z_j=y_j z_{j-1}, \quad j=1, 2, \ldots, n, \quad z_0=1. \tag{12}$$

Уравнения (10) и (11) являются частными случаями *линейного разностного уравнения первого порядка*

$$z_j=q_j z_{j-1}+\varphi_j, \quad j=1, 2, \ldots, n, \tag{13}$$

где q_j, φ_j — заданные числа, z_j — искомые числа. Для уравнения (13) рассматривается задача с начальными условиями или *задача Коши*, которая состоит в отыскании всех z_j, $j=1, 2, \ldots, n$, при за-

данном начальном значении z_0. Ясно, что решение задачи Коши для разностного уравнения (13) существует и единственно.

Коэффициенты q_j, правые части φ_j и искомое решение z_j уравнения (13) можно рассматривать как функции целочисленного аргумента j, т. е. $q_j=q(j)$, $\varphi_j=\varphi(j)$, $z_j=z(j)$.

Нам потребуется прежде всего записать решение уравнения (13) в явном виде. Подставляя в (13) вместо z_{j-1} выражение

$$z_{j-1}=q_{j-1}z_{j-2}+\varphi_{j-1},$$

получим

$$z_j=q_jq_{j-1}z_{j-2}+\varphi_j+q_j\varphi_{j-1}.$$

Теперь можно подставить сюда выражение для z_{j-2}, затем — для z_{j-3} и т. д. В результате получим формулу, в которой z_j выражается через z_{j-l}, φ_{j-l+1}, φ_{j-l+2}, ..., φ_j. Эта формула имеет вид

$$z_j=Q_{jl}z_{j-l}+\sum_{k=j-l+1}^{j}Q_{j,j-k}\varphi_k,\quad l=1,2,\ldots,j-1,j,\quad j=1,2,\ldots,n, \tag{14}$$

где

$$Q_{jl}=\begin{cases}1, & l=0,\\ q_jq_{j-1}\ldots q_{j-l+1}, & 1\leqslant l\leqslant j.\end{cases} \tag{15}$$

Строго доказать формулу (14) можно индукцией по числу j при каждом фиксированном l. Нам потребуется формула (14) при $l=j$, т. е.

$$z_j=Q_{jj}z_0+\sum_{k=1}^{j}Q_{j,j-k}\varphi_k,\quad j=1,2,\ldots,n, \tag{16}$$

где согласно (15)

$$Q_{l,j-k}=\begin{cases}1, & k=j,\\ q_jq_{j-1}\ldots q_{k+1}, & 0\leqslant k\leqslant j-1.\end{cases} \tag{17}$$

В частности, если (13) является уравнением с постоянными коэффициентами, т. е. $q_j=q$ для всех j, то из (16) получим

$$z_j=q^jz_0+\sum_{k=1}^{j}q^{j-k}\varphi_k,\quad j=1,2,\ldots,n. \tag{18}$$

Явную формулу (16) можно использовать для получения различных оценок решения z_j через начальные данные z_0, заданные коэффициенты q_j и правые части φ_j.

Лемма 1. *Если для некоторого $q\geqslant 0$ выполнены неравенства*

$$|q_j|\leqslant q,\quad j=1,2,\ldots,n, \tag{19}$$

то для решения уравнения (13) *справедливы оценки*

$$|z_j|\leqslant q^j|z_0|+\sum_{k=1}^{j}q^{j-k}|\varphi_k|,\quad j=1,2,\ldots,n. \tag{20}$$

Доказательство. Из (17) и (19) получаем, что

$$|Q_{j,j-k}| \leqslant q^{j-k}, \quad k=0, 1, \ldots, j.$$

Отсюда и из (16) следуют оценки (20).

Замечание. Оценки (20) неулучшаемы в том смысле, что для уравнения (13) с постоянными коэффициентами и положительными z_0, φ_k, $k=1, 2, \ldots, J$, неравенства (20) выполняются согласно (18) со знаком равенства.

5. Оценки погрешностей округления. Приведем примеры оценок погрешностей округления, возникающих в результате выполнения вычислительных алгоритмов. Нас будет интересовать в основном зависимость результирующей погрешности от числа арифметических действий n и от величины $\varepsilon=2^{-t}$, определяемой разрядностью ЭВМ.

Пример 1. Вычисление произведения

$$z_n = \prod_{j=1}^{n} y_j$$

n вещественных чисел проводится по формуле

$$z_j=y_j z_{j-1}, \quad j=1, 2, \ldots, n, \quad z_0=1. \tag{21}$$

Предположим, что в результате округления вместо точного значения z_{j-1} получено приближенное значение $\tilde{z}_{j-1}$. Тогда согласно (7) вместо $y_j\tilde{z}_{j-1}$ получим величину

$$\mathrm{fl}(y_j\tilde{z}_{j-1})=y_j\tilde{z}_{j-1}(1+\varepsilon_j),$$

где $|\varepsilon_j| \leqslant \varepsilon=2^{-t}$. Таким образом, вместо z_j получаем

$$\tilde{z}_j=(1+\varepsilon_j)y_j\tilde{z}_{j-1},$$

т. е. приближенное значение $\tilde{z}_j$ удовлетворяет рекуррентному соотношению

$$\tilde{z}_j=\tilde{y}_j\tilde{z}_{j-1}, \quad j=1, 2, \ldots, n, \quad \tilde{z}_0=1, \tag{22}$$

где $\tilde{y}_j=y_j(1+\varepsilon_j)$. Результирующая погрешность равна

$$z_n-\tilde{z}_n = \prod_{j=1}^{n} y_j - \prod_{j=1}^{n} (1+\varepsilon_j)\, y_j,$$

поэтому относительная погрешность есть

$$\frac{z_n-\tilde{z}_n}{z_n} = 1-\prod_{j=1}^{n}(1+\varepsilon_j).$$

Для оценки относительной погрешности заметим, что

$$|1+\varepsilon_j| \leqslant 1+\varepsilon, \quad j=1, 2, \ldots, n, \quad \varepsilon=2^{-t},$$

поэтому с точностью до величин второго порядка малости относительно ε можно считать, что

$$\left|\frac{z_n-\tilde{z}_n}{z_n}\right| \leqslant n\varepsilon = n2^{-t}. \tag{23}$$

При выводе оценки (23) предполагалось, что $\varepsilon=2^{-t}$, т. е. при перемножении не возникает чисел, меньших машинного нуля или больших машинной бесконечности. Однако может оказаться, что на каком-то этапе вычислений в качестве промежуточного результата будет получен либо машинный нуль M_0, либо машинная бесконечность M_∞. Поскольку оба указанных случая приводят к неверному окончательному результату, необходимо видоизменить вычислительный алгоритм. Оказывается, что здесь существенным является порядок действий.

Пусть, например, $M_0=2^{-p}$ и $M_\infty=2^p$ при некотором $p>0$. Предположим, что надо перемножить пять чисел $y_1=2^{p/2}$, $y_2=2^{p/4}$, $y_3=2^{3p/4}$, $y_4=2^{-p/2}$, $y_5=2^{-3p/4}$. Каждое из этих чисел и их произведение $2^{p/4}$ принадлежат допустимому диапазону чисел (M_0, M_∞). Однако произведение $y_1y_2y_3=2^{3p/2}>M_\infty$, поэтому при указанном порядке действий дальнейшее выполнение алгоритма становится невозможным. Если проводить вычисление в порядке $y_5y_4y_3y_2y_1$, то получим $y_5y_4=2^{-5p/4}<M_0$, следовательно, $\mathrm{fl}(y_5y_4)=0$ и все произведение окажется равным нулю, т. е. получим неверный результат. В данном примере к верному результату приводит вычисление произведения в порядке

$$y_5y_3y_1y_4y_2.$$

В случае произвольного числа n сомножителей можно предложить следующий алгоритм вычисления произведения (см. [6]). Предположим, что

$$|y_1|\leqslant|y_2|\leqslant\ldots\leqslant|y_n|,$$

причем $|y_1|\leqslant 1$, $|y_n|\geqslant 1$. Будем сначала проводить умножение в порядке $y_1y_ny_{n-1}\ldots$ до тех пор, пока впервые не получим число, большее единицы. Затем полученное частичное произведение будем последовательно умножать на y_2, y_3 и т. д. до тех пор, пока новое частичное произведение не станет меньше единицы. Процесс повторяется до тех пор, пока все оставшиеся сомножители будут либо только большими единицы по модулю, либо только меньшими. Далее умножение проводится в произвольном порядке.

Пример 2. Рассмотрим процесс вычисления суммы

$$z_n=y_1+y_2+\ldots+y_n. \tag{24}$$

Для простоты изложения предположим, что все y_i положительны и больше машинного нуля. Тогда в процессе вычислений не может появиться нулевого результата. Алгоритм вычисления суммы (24) состоит в решении разностного уравнения (10) при начальном значении $z_0=0$.

Получим уравнение, которому удовлетворяет приближенное решение $\tilde{z}_j$. Предположим, что вместо точного значения z_{j-1} в результате накопления погрешностей округления получено приближенное значение $\tilde{z}_{j-1}$. Тогда согласно (7) вместо z_j получим число

$$\tilde{z}_j=\mathrm{fl}(\tilde{z}_{j-1}+y_j)=(1+\varepsilon_j)(\tilde{z}_{j-1}+y_j),$$

где $|\varepsilon_j|\leqslant 2^{-t}$.

Таким образом, приближенное значение $\tilde{z}_j$ удовлетворяет разностному уравнению

$$\tilde{z}_j=q_j\tilde{z}_{j-1}+\tilde{y}_j,\quad j=1, 2, \ldots, n,\quad \tilde{z}_0=0, \tag{25}$$

где $q_j=1+\varepsilon_j$, $\tilde{y}_j=(1+\varepsilon_j)y_j$. Можно считать, что уравнение (25) получено из исходного уравнения (10) путем внесения возмущений в коэффициенты и в правые части, причем для каждого j возмущение пропорционально ε_j и не превосходит 2^{-t}.

Оценим теперь результирующую погрешность $\tilde{z}_n-z_n$. Для этого выпишем в явном виде решения уравнений (10) и (25), предполагая, что $z_0=\tilde{z}_0=0$. Согласно (16), (18) имеем

$$z_n=\sum_{k=1}^{n} y_k, \quad \tilde{z}_n=\sum_{k=1}^{n} Q_{n,n-k}\tilde{y}_k,$$

где $\tilde{y}_k=q_k y_k$. Поэтому для погрешности получим следующее выражение:

$$\tilde{z}_n-z_n=\sum_{k=1}^{n} E_{nk}y_k, \tag{26}$$

где

$$E_{nk}=q_kQ_{n,n-k}-1=\begin{cases} q_n-1, & k=n, \\ q_nq_{n-1}\ldots q_{k+1}q_k-1, & k=1, 2, \ldots, n-1. \end{cases} \tag{27}$$

Коэффициент E_{nk} в формуле (26) указывает, какую долю погрешности вносит k-е слагаемое суммы (24) в общую погрешность. Покажем, что чем меньше номер k, тем большая погрешность вносится за счет y_k. Для этого оценим приближенно величины E_{nk}. Так как $q_j=1+\varepsilon_j$ и $|\varepsilon_j|<\varepsilon=2^{-t}$, то $|q_n|\leqslant 1+\varepsilon$, $|q_nq_{n-1}\ldots q_{k+1}q_k|\leqslant (1+\varepsilon)^{n-k+1}$. Отбрасывая величины второго порядка малости относительно ε, можно считать, что

$$|q_nq_{n-1}\ldots q_k|\leqslant 1+(n-k+1)\varepsilon,$$

и тогда

$$|E_{nk}|\leqslant (n-k+1)\varepsilon, \quad k=1, 2, \ldots, n. \tag{28}$$

Из формулы (26) легко получить оценку относительной погрешности $|\tilde{z}_n-z_n|/|z_n|$. Заметим сначала, что для положительных $y_1, \ldots, y_n$ последовательность z_j, определенная согласно (10), неотрицательная и монотонно возрастающая, т. е.

$$0\leqslant z_{k-1}\leqslant z_k, \quad k=1, 2, \ldots, n.$$

Поэтому для $y_k=z_k-z_{k-1}$ справедливо неравенство

$$0\leqslant y_k\leqslant |z_k|+|z_{k-1}|\leqslant 2|z_n|, \quad k=1, 2, \ldots, n.$$

Отсюда и из (26) получим оценку

$$|\tilde{z}_n-z_n|\leqslant 2|z_n|\sum_{k=1}^{n}|E_{nk}|.$$

Учитывая приближенное неравенство (28), приходим к следующей

оценке относительной погрешности:

$$\left|\frac{\tilde{z}_n - z_n}{z_n}\right| \leqslant \varepsilon n(n+1), \qquad \varepsilon = 2^{-t}.$$

Следовательно, относительная погрешность, возникающая при суммировании n положительных чисел, оценивается примерно как $n^2 2^{-t}$, где t — число разрядов, отводимое для записи мантиссы. Например, при $2^{-t} = 10^{-12}$, $n = 10^3$ получаем, что результирующая относительная погрешность не превзойдет 10^{-6}.

§ 3. Разностные уравнения второго порядка

1. Задача Коши и краевые задачи для разностных уравнений. В п. 4 § 2 рассматривалась задача Коши для разностного уравнения первого порядка. Обратимся теперь к линейным разностным уравнениям второго порядка

$$a_j y_{j-1} - c_j y_j + b_j y_{j+1} = -f_j, \tag{1}$$

где a_j, b_j, c_j, f_j — заданные коэффициенты и правая часть и y_j — искомое решение. Индекс j в уравнении (1) пробегает некоторое допустимое множество J целых чисел. Например,

$$J = \{0, 1, 2, \ldots\}, \quad J = \{1, 2, \ldots, N-1\}, \quad J = \{0, \pm 1, \pm 2, \ldots\},$$

где $N > 1$ — заданное целое число. Всюду в дальнейшем будем предполагать, что $b_j \neq 0$, $a_j \neq 0$ для всех допустимых j.

Коэффициенты, правую часть и решение уравнения (1) следует рассматривать как функции целочисленного аргумента $j \in J$, т. е. $y_j = y(j)$, $f_j = f(j)$ и т. д.

Уравнение (1) имеет бесконечное множество решений. Каждое отдельное решение называется *частным решением* уравнения (1). *Общим решением* уравнения (1) называется такое двухпараметрическое семейство решений, которое содержит любое частное решение. В пп. 3, 4 будет показано, каким образом строится общее решение уравнения (1).

Для того чтобы из совокупности всех решений уравнения (1) выделить единственное, необходимо задать те или иные дополнительные условия.

Задача Коши состоит в отыскании решения y_j, $j = 0, 1, 2, \ldots$, уравнения (1), удовлетворяющего при $j = 0, 1$ заданным начальным условиям

$$y_0 = \mu_1, \qquad y_1 = \mu_2. \tag{2}$$

Если $b_j \neq 0$ для всех допустимых j, то уравнение (1) можно разрешить относительно y_{j+1}, т. е. записать в виде

$$y_{j+1} = -\frac{a_j}{b_j} y_{j-1} + \frac{c_j}{b_j} y_j - \frac{f_j}{b_j}. \tag{3}$$

Отсюда следует, что задача Коши имеет единственное решение.

Более общая постановка задачи Коши состоит в отыскании при всех $j=0, \pm 1, \pm 2, \ldots$ решения уравнения (1), удовлетворяющего условиям $y_{j_0}=\mu_1$, $y_{j_0+1}=\mu_2$ с заданными j_0, μ_1, μ_2. Если $a_j \neq 0$, $b_j \neq 0$ для всех j, то такая задача имеет единственное решение.

Краевая задача состоит в отыскании решения уравнения

$$a_j y_{j-1} - c_j y_j + b_j y_{j+1} = -f_j, \quad j=1, 2, \ldots, N-1, \tag{4}$$

удовлетворяющего дополнительным условиям

$$y_0 = \varkappa_1 y_1 + \mu_1, \quad y_N = \varkappa_2 y_{N-1} + \mu_2, \tag{5}$$

где $\varkappa_i$, μ_i, $i=1, 2$ — заданные числа. В частности, при $\varkappa_1=\varkappa_2=0$ получаем *краевые условия первого рода*

$$y_0 = \mu_1, \quad y_N = \mu_2, \tag{6}$$

а при $\varkappa_1=\varkappa_2=1$ — *краевые условия второго рода.* Достаточные условия существования единственного решения краевой задачи (4), (5), а также алгоритм построения этого решения будут указаны в п. 7 § 4.

2. Однородное разностное уравнение второго порядка с постоянными коэффициентами. Рассмотрим разностное уравнение

$$a y_{j-1} - c y_j + b y_{j+1} = 0, \quad a \neq 0, \quad b \neq 0, \tag{7}$$

с вещественными коэффициентами a, b, c, не зависящими от j.

Будем искать частные решения уравнения (7) в виде

$$y_j = q^j, \tag{8}$$

где q — число, подлежащее определению. Подставляя (8) в (7), получим квадратное уравнение

$$bq^2 - cq + a = 0, \tag{9}$$

которое называется *характеристическим уравнением, соответствующим разностному уравнению* (7).

В зависимости от знака дискриминанта $c^2 - 4ab$ могут представиться три различных случая. Если $c^2 > 4ab$, то корни

$$q_1 = \frac{c + \sqrt{c^2 - 4ab}}{2b}, \quad q_2 = \frac{c - \sqrt{c^2 - 4ab}}{2b} \tag{10}$$

уравнения (9) вещественны и различны. В этом случае разностное уравнение (7) имеет частные решения

$$y_j^{(1)} = q_1^j, \quad y_j^{(2)} = q_2^j. \tag{11}$$

Если $c^2 < 4ab$, то корни q_1 и q_2 комплексно сопряжены. Функции (11) и в этом случае являются решениями разностного уравнения (7), однако удобнее представить q_1 в тригонометрической форме: $q_1 = r(\cos\varphi + i \sin\varphi)$, где

$$r = \sqrt{\frac{a}{b}}, \quad \sin\varphi = \frac{\sqrt{4ab - c^2}}{2\sqrt{ab}}, \quad \cos\varphi = \frac{c}{2\sqrt{ab}}. \tag{12}$$

В качестве решений уравнения (7) можно взять функции

$$y_j^{(1)} = r^j \cos(j\varphi), \quad y_j^{(2)} = r^j \sin(j\varphi).$$

Наконец, если $c^2=4ab$, то уравнение (9) имеет кратный корень $q=c/(2b)$, а разностное уравнение имеет частные решения

$$y_j^{(1)} = q^j, \quad y_j^{(2)} = jq^j. \tag{13}$$

Построим теперь решение задачи Коши

$$ay_{j-1} - cy_j + by_{j+1} = 0, \quad j = 1, 2, \ldots, \tag{14}$$

$$y_0 = \mu_1, \quad y_1 = \mu_2, \tag{15}$$

исходя из найденных частных решений (11). В силу линейности и однородности уравнения (7) любая линейная комбинация

$$y_j = \alpha_1 q_1^j + \alpha_2 q_2^j \tag{16}$$

также является его решением. Подберем параметры α_1 и α_2 таким образом, чтобы удовлетворялись начальные условия (15):

$$\alpha_1 + \alpha_2 = \mu_1, \quad \alpha_1 q_1 + \alpha_2 q_2 = \mu_2. \tag{17}$$

Решая систему (17), находим

$$\alpha_1 = \frac{\mu_1 q_2 - \mu_2}{q_2 - q_1}, \quad \alpha_2 = \frac{\mu_2 - \mu_1 q_1}{q_2 - q_1}. \tag{18}$$

Подставляя (18) в (16) и собирая коэффициенты при μ_1, μ_2, получим, что решение задачи Коши (14), (15) в случае $c^2>4ab$ имеет вид

$$y_j = \frac{q_1 q_2 (q_1^{j-1} - q_2^{j-1})}{q_2 - q_1}\mu_1 + \frac{q_2^j - q_1^j}{q_2 - q_1}\mu_2, \quad j = 0, 1, 2, \ldots, \tag{19}$$

где $q_{1,2}$ определены согласно (10).

В таком же виде представляется и решение задачи Коши (14), (15) в случае $c^2<4ab$. Заметим, что в этом случае

$$\frac{q_1 q_2 (q_1^{j-1} - q^{j-1})}{q_2 - q_1} = -r^j \frac{\sin((j-1)\varphi)}{\sin\varphi},$$

$$\frac{q_2^j - q_1^j}{q_2 - q_1} = r^{j-1} \frac{\sin(j\varphi)}{\sin\varphi},$$

где r и φ определены согласно (12). Поэтому решение задачи Коши можно записать в виде

$$y_j = -r^j \frac{\sin((j-1)\varphi)}{\sin\varphi}\mu_1 + r^{j-1} \frac{\sin(j\varphi)}{\sin\varphi}\mu_2. \tag{20}$$

В случае $c^2=4ab$, используя частные решения (13), можно представить решение задачи Коши (14), (15) в виде

$$y_j = -(j-1)q^j\mu_1 + jq^{j-1}\mu_2, \tag{21}$$

где $q=c/(2b)$.

Аналогичным образом строится решение краевой задачи

$$ay_{j-1}-cy_j+by_{j+1}=0, \quad j=1, 2, \ldots, N-1, \tag{22}$$

$$y_0=\mu_1, \quad y_N=\mu_2. \tag{23}$$

Если $c^2\neq 4ab$, то

$$y_j=\frac{(q_2^{N-j}-q_1^{N-j})(q_1q_2)^j}{q_2^N-q_1^N}\mu_1+\frac{q_2^j-q_1^j}{q_2^N-q_1^N}\mu_2, \tag{24}$$

где $q_{1,2}$ определены согласно (10).

Если же $c^2=4ab$, то

$$y_j=\left(1-\frac{j}{N}\right)q^j\mu_1+\frac{j}{N}q^{-(N-j)}\mu_2, \tag{25}$$

где $q=c/(2b)$.

3. Однородное разностное уравнение второго порядка с переменными коэффициентами. Для уравнения с переменными коэффициентами (1) существует теория, аналогичная теории линейных дифференциальных уравнений, а именно: общее решение однородного уравнения

$$a_jy_{j-1}-c_jy_j+b_jy_{j+1}=0 \tag{26}$$

представляется в виде линейной комбинации двух его линейно независимых решений, а общее решение неоднородного уравнения (1) — в виде суммы частного решения неоднородного уравнения и общего решения однородного уравнения.

Изучим более подробно свойства разностного уравнения (26). Будем считать сейчас, что $J=\{0, \pm 1, \pm 2, \ldots\}$, т. е. уравнение (26) определено для всех целых j. Заметим прежде всего, что если u_j и v_j — два решения уравнения (26), то и любая линейная комбинация $\alpha_1u_j+\alpha_2v_j$ также является решением. Этот факт следует из линейности и однородности уравнения (26).

Для дальнейшего потребуются понятия линейной зависимости и линейной независимости функций, заданных на множестве J. Две функции u_j и v_j целочисленного аргумента $j\in J$ называются *линейно зависимыми*, если существуют постоянные α_1, α_2, одновременно не равные нулю и такие, что выполнено равенство

$$\alpha_1u_j+\alpha_2v_j=0 \text{ для всех } j\in J. \tag{27}$$

Если же из условия (27) следует, что $\alpha_1=\alpha_2=0$, то функции u_j, v_j называются *линейно независимыми*.

Линейная зависимость решений u_j, v_j характеризуется значениями определителей

$$w_j[u, v]=\begin{vmatrix} u_j & v_j \\ u_{j+1} & v_{j+1} \end{vmatrix}, \tag{28}$$

являющимися аналогами определителя Вронского.

Лемма 1. *Если функции u_j, v_j линейно зависимы, то $w_j=0$ для всех $j\in J$.*

Действительно, согласно (27) для всех $j \in J$ выполняются равенства

$$u_j\alpha_1 + v_j\alpha_2 = 0,$$
$$u_{j+1}\alpha_1 + v_{j+1}\alpha_2 = 0, \tag{29}$$

где $\alpha_1^2 + \alpha_2^2 \neq 0$. Рассматривая (29) при каждом фиксированном j как однородную систему линейных алгебраических уравнений относительно α_1, α_2 и учитывая, что $\alpha_1^2 + \alpha_2^2 \neq 0$, получим, что определитель w_j этой системы равен нулю.

Для решений однородного уравнения (26) справедливо утверждение, обратное лемме 1.

Лемма 2. *Если u_j, v_j — линейно независимые решения однородного уравнения* (26) *и $a_j \neq 0$, $b_j \neq 0$ для всех j, то определитель $w_j[u, v]$ не обращается в нуль ни в одной точке $j \in J$.*

Доказательство проведем от противного. Предположим, что найдется точка $j_0 \in J$, для которой $w_{j_0}[u, v] = 0$. Рассмотрим систему уравнений

$$u_{j_0}\alpha_1 + v_{j_0}\alpha_2 = 0,$$
$$u_{j_0+1}\alpha_1 + v_{j_0+1}\alpha_2 = 0 \tag{30}$$

относительно неизвестных α_1, α_2. Поскольку определитель этой системы равен нулю, существует нетривиальное решение $\{\alpha_1, \alpha_2\}$. Образуем с помощью этого решения $\{\alpha_1, \alpha_2\}$ функцию

$$z_j = \alpha_1 u_j + \alpha_2 v_j \tag{31}$$

и покажем, что $z_j = 0$ для всех j.

Поскольку u_j и v_j — решения однородного уравнения (26), функция (31) также является его решением, т. е. удовлетворяет уравнению

$$a_j z_{j-1} - c_j z_j + b_j z_{j+1} = 0. \tag{32}$$

Кроме того, согласно (30) выполнены условия $z_{j_0} = z_{j_0+1} = 0$.

По предположению коэффициенты a_j, b_j отличны от нуля для всех j. Следовательно, для уравнения (32) можно рассмотреть задачи Коши

$$z_{j+1} = \frac{c_j}{b_j} z_j - \frac{a_j}{b_j} z_{j-1}, \quad j = j_0 + 1, j_0 + 2, \ldots,$$
$$z_{j_0} = z_{j_0+1} = 0, \tag{33}$$
$$z_{j-1} = \frac{c_j}{a_j} z_j - \frac{b_j}{a_j} z_{j+1}, \quad j = j_0, j_0 - 1, j_0 - 2, \ldots,$$
$$z_{j_0} = z_{j_0+1} = 0. \tag{34}$$

Из рекуррентных соотношений (33), (34) получаем, что $z_j = 0$ для всех $j = 0, \pm 1, \pm 2, \ldots$ Последнее означает, что $\alpha_1 u_j + \alpha_2 v_j = 0$ для всех j, причем $\alpha_1^2 + \alpha_2^2 \neq 0$. Следовательно, функции u_j, v_j линейно зависимы, что противоречит предположению леммы 2.

Следствие 1. *Определитель* (28), *составленный для двух решений уравнения* (26), *или тождественно по* j *равен нулю, или отличен от нуля для всех* j.

Любая система из двух линейно независимых решений уравнения (26) называется *фундаментальной системой.*

Теорема 1. *Уравнение* (26) *с* $a_j \neq 0$, $b_j \neq 0$, $j \in J$, *всегда имеет фундаментальную систему.*

Доказательство. Фундаментальную систему образуют, например, решения u_j и v_j следующих задач Коши:

$$a_j u_{j-1} - c_j u_j + b_j u_{j+1} = 0,$$
$$j = 0, \pm 1, \pm 2, \ldots, \quad u_0 = 0, \quad u_1 = 1;$$
$$a_j v_{j-1} - c_j v_j + b_j v_{j+1} = 0,$$
$$j = 0, \pm 1, \pm 2, \ldots, \quad v_0 = 1, \quad v_1 = 0.$$

Действительно,

$$w_0 = \begin{vmatrix} u_0 & u_1 \\ v_0 & v_1 \end{vmatrix} \neq 0,$$

и согласно следствию 1 $w_j[u, v] \neq 0$ для всех j. Но тогда согласно лемме 1 функции u_j, v_j линейно независимы.

Теорема 2. *Если* u_j, v_j — *фундаментальная система решений уравнения* (26), *то его общее решение имеет вид*

$$y_j = \alpha_1 u_j + \alpha_2 v_j, \tag{35}$$

где α_1 *и* α_2 — *произвольные постоянные.*

Доказательство. Пусть y_j — любое решение уравнения (26) и u_j, v_j — два заданных линейно независимых решения. Надо показать, что найдутся постоянные α_1 и α_2, для которых справедливо (35). Пусть y_0 и y_1 — значения решения y_j в точках $j = 0$ и $j = 1$ соответственно. Выберем постоянные α_1 и α_2 из условий

$$u_0 \alpha_1 + v_0 \alpha_2 = y_0, \quad u_1 \alpha_1 + v_1 \alpha_2 = y_1. \tag{36}$$

Определитель этой системы $w_0[u, v] \neq 0$, так как u и v — линейно независимые решения. Следовательно, при заданных y_0, y_1 система (36) имеет единственное решение $\{\alpha_1, \alpha_2\}$. В силу единственности решения задачи Коши функция (35), построенная с помощью найденных постоянных α_1 и α_2, совпадает с заданным решением y_j.

Следствие. *Любые три решения однородного уравнения* (26) *линейно зависимы.*

Пусть u_j, v_j, y_j — любые решения уравнения (26). Если u_j и v_j линейно зависимы, то утверждение доказано. Если же u_j и v_j линейно независимы, то они образуют фундаментальную систему и согласно теореме 2 решение y_j представляется в виде линейной комбинации u_j и v_j.

В качестве упражнения предлагается проверить, что частные решения (11) уравнения (7) с постоянными коэффициентами будут линейно независимы при $q_1 \neq q_2$ и линейно зависимы — при $q_1 = q_2$. В последнем случае линейно независимыми будут решения

(13). Заметим, что вследствие предположения $a \neq 0$ характеристическое уравнение (9) не имеет нулевых корней.

4. Неоднородное разностное уравнение второго порядка. Обратимся снова к неоднородному уравнению

$$a_j y_{j-1} - c_j y_j + b_j y_{j+1} = -f_j. \tag{37}$$

Уравнение

$$a_j y_{j-1} - c_j y_j + b_j y_{j+1} = 0 \tag{38}$$

называется *однородным уравнением, соответствующим уравнению* (37).

Теорема 3. *Общее решение неоднородного уравнения* (37) *есть сумма какого-либо его частного решения и общего решения соответствующего однородного уравнения.*

Доказательство. Пусть Y_j — какое-либо частное решение неоднородного уравнения (37) и u_j, v_j — линейно независимые решения соответствующего однородного уравнения (38). Тогда общее решение однородного уравнения (38) имеет вид $\alpha_1 u_j + \alpha_2 v_j$, где α_1 и α_2 — произвольные постоянные. Непосредственной подстановкой проверяется, что функция

$$y_j = Y_j + \alpha_1 u_j + \alpha_2 v_j \tag{39}$$

является решением неоднородного уравнения (37). Остается доказать, что функция (39) является общим решением, т. е. что при соответствующем выборе параметров α_1, α_2 любое решение уравнения (37) можно записать в виде (39). Пусть z_j — любое решение уравнения (37). Оно однозначно определяется заданием начальных условий z_0 и z_1. Поэтому для совпадения y_j, определенного согласно (39), с заданным решением z_j достаточно потребовать $y_0 = z_0$, $y_1 = z_1$, т. е.

$$\alpha_1 u_0 + \alpha_2 v_0 = z_0 - Y_0,$$
$$\alpha_1 u_1 + \alpha_2 v_1 = z_1 - Y_1.$$

Рассматривая эти условия как систему уравнений относительно α_1, α_2, получаем, что она имеет единственное решение, поскольку определитель

$$\begin{vmatrix} u_0 & v_0 \\ u_1 & v_1 \end{vmatrix} = w_0[u, v]$$

отличен от нуля в силу линейной независимости решений u_j, v_j. Теорема 3 доказана.

Частное решение неоднородного уравнения (37) можно построить, если известны линейно независимые решения u_j, v_j соответствующего однородного уравнения (38). Для такого построения применяется *метод вариации постоянных.*

Напомним метод вариации постоянных на примере дифференциального уравнения

$$y''(x) = -f(x). \tag{40}$$

Пусть $u(x)$, $v(x)$ — линейно независимые решения соответствующего однородного уравнения, т. е.

$$u''(x) = 0, \quad v''(x) = 0. \tag{41}$$

Будем искать решение неоднородного уравнения (40) в виде

$$y(x)=\alpha(x)u(x)+\beta(x)v(x), \tag{42}$$

где $\alpha(x)$, $\beta(x)$ — функции, подлежащие определению. Для нахождения функций $\alpha(x)$, $\beta(x)$ необходимо получить два уравнения. Первое из них получается из требования, чтобы производная $y'(x)$ имела вид

$$y'(x)=\alpha(x)u'(x)+\beta(x)v'(x), \tag{43}$$

которое, очевидно, эквивалентно требованию

$$\alpha'(x)u(x)+\beta'(x)v(x)=0. \tag{44}$$

Второе уравнение, связывающее $\alpha(x)$ и $\beta(x)$, получается в результате подстановки (42) в исходное уравнение (40). Учитывая (43), (41), получим

$$y''(x)=\alpha u''(x)+\beta v''(x)+\alpha'(x)u'(x)+\beta'(x)u'(x)=\alpha'(x)u'(x)+\beta'(x)v'(x).$$

Следовательно, уравнение (40) будет выполнено, если

$$\alpha'(x)u'(x)+\beta'(x)v'(x)=-f(x). \tag{45}$$

Из системы уравнений (44), (45) найдем

$$\alpha'(x)=\frac{f(x)v(x)}{u(x)v'(x)-u'(x)v(x)}, \quad \beta'(x)=\frac{-f(x)u(x)}{u(x)v'(x)-u'(x)v(x)}. \tag{46}$$

Знаменатель в полученных выражениях отличен от нуля, так как он является определителем Вронского для линейно независимых решений однородного уравнения. Из выражений (46) коэффициенты $\alpha(x)$, $\beta(x)$ находятся в квадратурах.

Обращаясь к разностному уравнению (37), будем искать его решение в виде

$$y_j=\alpha_j u_j+\beta_j v_j, \tag{47}$$

где u_j, v_j — линейно независимые решения соответствующего однородного уравнения (38) и α_j, β_j — искомые функции. Потребуем по аналогии с (43), чтобы разность $y_{j+1}-y_j$ представлялась в виде

$$y_{j+1}-y_j=\alpha_j(u_{j+1}-u_j)+\beta_j(v_{j+1}-v_j). \tag{48}$$

Такое требование эквивалентно выполнению условия

$$(\alpha_{j+1}-\alpha_j)u_{j+1}+(\beta_{j+1}-\beta_j)v_{j+1}=0. \tag{49}$$

Далее, из (48) получим

$$y_j-y_{j-1}=\alpha_{j-1}(u_j-u_{j-1})+\beta_{j-1}(v_j-v_{j-1})$$

или

$$y_j-y_{j-1}=\alpha_j(u_j-u_{j-1})+\beta_j(v_j-v_{j-1})-\varphi_j, \tag{50}$$

где $\varphi_j=(\alpha_j-\alpha_{j-1})(u_j-u_{j-1})+(\beta_j-\beta_{j-1})(v_j-v_{j-1})$.

Для дальнейшего удобно представить уравнение (37) в виде

$$(a_j-c_j+b_j)y_j+b_j(y_{j+1}-y_j)-a_j(y_j-y_{j-1})=-f_j. \tag{51}$$

Подставляя в (51) выражения (47), (48), (50) и собирая коэффициенты при α_j, β_j, получим

$$\alpha_j[(a_j-c_j+b_j)u_j+b_j(u_{j+1}-u_j)-a_j(u_j-u_{j-1})]+$$
$$+\beta_j[(a_j-c_j+b_j)v_j+b_j(v_{j+1}-v_j)-a_j(v_j-v_{j-1})]+a_j\varphi_j=-f_j.$$

Выражения, стоящие в квадратных скобках, равны нулю, потому что u_j и v_j являются решениями однородного уравнения (38). Следовательно, уравнение (37) будет выполнено, если потребовать $a_j\varphi_j=-f_j$, т. е.

$$(\alpha_j-\alpha_{j-1})(u_j-u_{j-1})+(\beta_j-\beta_{j-1})(v_j-v_{j-1})=-\frac{f_j}{a_j}. \tag{52}$$

Поскольку индекс j произволен, уравнение (49) можно переписать в виде

$$(\alpha_j-\alpha_{j-1})u_j+(\beta_j-\beta_{j-1})v_j=0. \tag{53}$$

Решая систему уравнений (52), (53), получим

$$\begin{aligned}\alpha_j-\alpha_{j-1}&=\frac{v_j}{v_ju_{j-1}-u_jv_{j-1}}\,\frac{f_j}{a_j},\\ \beta_j-\beta_{j-1}&=-\frac{u_j}{v_ju_{j-1}-u_jv_{j-1}}\,\frac{f_j}{a_j}.\end{aligned} \tag{54}$$

Знаменатель полученных выражений совпадает с определителем $w_{j-1}[u, v]$ (см. (28)) и согласно лемме 2 не обращается в нуль ни в одной точке j. В результате суммирования каждого из уравнений (54) получим

$$\begin{aligned}\alpha_j&=\alpha_0+\sum_{k=1}^{j}\frac{v_k}{v_ku_{k-1}-u_kv_{k-1}}\,\frac{f_k}{a_k},\\ \beta_j&=\beta_0+\sum_{k=1}^{j}\frac{-u_k}{v_ku_{k-1}-u_kv_{k-1}}\,\frac{f_k}{a_k}.\end{aligned}$$

Подставляя найденные выражения для α_j, β_j в формулу (47), получаем общее решение неоднородного уравнения (37) в виде

$$y_j=\alpha_0u_j+\beta_0v_j+\sum_{k=1}^{j}\frac{\begin{vmatrix}u_j & v_j\\ u_k & v_k\end{vmatrix}}{\begin{vmatrix}u_{k-1} & v_{k-1}\\ u_k & v_k\end{vmatrix}}\,\frac{f_k}{a_k}, \tag{55}$$

где α_0, β_0 — произвольные постоянные и u_j, v_j — линейно независимые решения однородного уравнения (38).

Отметим, что сумма

$$z_j=\alpha_0u_j+\beta_0v_j$$

является общим решением однородного уравнения (38), а сумма

$$Y_j=\sum_{k=1}^{j}\frac{\begin{vmatrix}u_j & v_j\\ u_k & v_k\end{vmatrix}}{\begin{vmatrix}u_{k-1} & v_{k-1}\\ u_k & v_k\end{vmatrix}}\,\frac{f_k}{a_k} \tag{56}$$

— частным решением неоднородного уравнения (37), соответствующим значениям $\alpha_0=\beta_0=0$. Следовательно, функция (55) является общим решением неоднородного уравнения (37).

В заключение параграфа отметим, что многие понятия и результаты, относящиеся к разностным уравнениям второго порядка, можно обобщить и на разностные уравнения произвольного порядка (см., например, [35]).

§ 4. Разностная аппроксимация дифференциальных уравнений

1. Сетки и сеточные функции. Для численного решения дифференциальных уравнений, обыкновенных и в частных производных, часто применяется метод сеток или *разностный метод*. В настоящем параграфе поясняются основные идеи разностного метода на самых простых примерах. Систематическое изложение теории разностных методов содержится в ч. III (см. также [32]).

Сеткой на отрезке $[a, b]$ называется любое конечное множество точек этого отрезка. Функция, определенная в точках сетки, называется *сеточной функцией*. Будем обозначать через ω_N сетку, удовлетворяющую условиям

$$a=x_0<x_1<x_2<\cdots<x_{N-1}<x_N=b, \tag{1}$$

и через f_i — значение сеточной функции $f(x)$ в точке $x_i\in\omega_N$, т. е. $f_i=f(x_i)$. Точки $x_i\in\omega_N$ называются *узлами сетки* ω_N. *Равномерной сеткой* на $[a, b]$ называется множество точек

$$\omega_h=\{x_i=a+ih, \quad i=0, 1, \ldots, N\}, \tag{2}$$

где $h=(b-a)/N$ — *шаг сетки*.

Рассмотрим задачу о приближенном вычислении производных функции $u(x)$, определенной и непрерывной на отрезке $[a, b]$. Будем считать, что $u(x)$ обладает необходимой по ходу изложения гладкостью. Введем согласно (2) сетку ω_h и обозначим

$$u_i=u(x_i), \quad u_{\overline{x},i}=(u_i-u_{i-1})/h,$$
$$u_{x,i}=(u_{i+1}-u_i)/h, \quad u_{\underset{\circ}{x},i}=(u_{i+1}-u_{i-1})/(2h).$$

Выписанные здесь разностные отношения называются, соответственно, *левой, правой* и *центральной разностными производными функции* $u(x)$ *в точке* $x=x_i$. Если точка x_i фиксирована, а шаг h стремится к нулю (при этом $i\to\infty$), то каждое из упомянутых разностных отношений стремится к значению производной функции $u(x)$ в точке x_i. Поэтому в качестве приближенного значения $u'(x)$ можно взять любое из этих разностных отношений.

Нетрудно получить выражение для погрешности, возникающей при замене дифференциального выражения разностным. Рассмотрим, например, левую разностную производную в точке $x=x_i$ и запишем ее в виде

$$u_{\overline{x},i}=\frac{u(x)-u(x-h)}{h}.$$

По формуле Тейлора получим

$$u(x-h)=u(x)-hu'(x)+\frac{h^2}{2}u''(\zeta), \quad \zeta\in(x-h, x),$$

следовательно,

$$u_{\bar{x},i}=u'(x_i)-\frac{h}{2}u''(\zeta_i). \tag{3}$$

Погрешность $u_{\bar{x},i}-u'(x_i)$, возникающая при замене дифференциального выражения $u'(x)$ разностным выражением $u_{\bar{x},i}$, называется *погрешностью аппроксимации*. Из разложения (3) видно, что погрешность аппроксимации является величиной $O(h)$ при $h\to 0$. В этом случае говорят, что имеет место *аппроксимация первого порядка*.

Приведем разложения, аналогичные (3), для других разностных отношений:

$$u_{x,i}=u'(x_i)+\frac{h}{2}u''(\zeta_i^{(1)}), \quad \zeta_i^{(1)}\in(x_i, x_{i+1}), \tag{4}$$

$$u_{\overset{\circ}{x},i}=u'(x_i)+\frac{h^2}{6}u'''(\zeta_i^{(2)}), \quad \zeta_i^{(2)}\in(x_{i-1}, x_{i+1}). \tag{5}$$

Из разложения (5) видно, что центральная разностная производная аппроксимирует $u'(x)$ со вторым порядком и, следовательно, является более точным приближением к $u'(x)$, чем левая или правая разностные производные. В дальнейшем наряду с (3)—(5) будем использовать менее детальную запись тех же разложений, а именно

$$u_{\bar{x},i}=u_i'+O(h), \quad u_{x,i}=u_i'+O(h), \quad u_{\overset{\circ}{x},i}=u_i'+O(h^2).$$

Вторую производную $u''(x)$ можно приближенно заменить в точке $x_i\in\omega_h$ *второй разностной производной*

$$u_{\bar{x}x,i}=\frac{1}{h}(u_{x,i}-u_{\bar{x},i})=\frac{u_{i+1}-2u_i+u_{i-1}}{h^2}. \tag{6}$$

Разложение по формуле Тейлора приводит к следующему выражению для погрешности:

$$u_{\bar{x}x,i}-u''(x_i)=\frac{h^2}{12}u^{\mathrm{IV}}(\zeta_i), \tag{7}$$

т. е. имеет место аппроксимация второго порядка.

Мы привели простейшие примеры аппроксимации дифференциальных выражений разностными на равномерной сетке. В общем случае погрешность, возникающая в результате замены дифференциального выражения разностным, зависит как от распределения узлов сетки, так и от гладкости функции.

2. Разностная краевая задача. Первая краевая задача для уравнения

$$u''(x)=-f(x) \tag{8}$$

состоит в отыскании функции $u(x)$, дважды непрерывно диффе-

ренцируемой на интервале (a, b), непрерывной на отрезке $[a, b]$, удовлетворяющей уравнению (8) при $x \in (a, b)$ и дополнительным условиям

$$u(a) = \mu_1, \quad u(b) = \mu_2, \tag{9}$$

где μ_1, μ_2 — заданные числа.

Нетрудно построить решение задачи (8), (9) в виде квадратур. Представим $u(x)$ в виде суммы двух функций:

$$u(x) = v(x) + w(x),$$

где

$$v''(x) = 0, \quad x \in (a, b), \qquad v(a) = \mu_1, \quad v(b) = \mu_2, \tag{10}$$

$$w''(x) = -f(x), \quad x \in (a, b), \qquad w(a) = w(b) = 0. \tag{11}$$

Решением задачи (10) является линейная функция

$$v(x) = \frac{b-x}{b-a}\mu_1 + \frac{x-a}{b-a}\mu_2. \tag{12}$$

Далее, интегрируя уравнение (11), получим

$$w'(t) = w'(a) - \int_a^t f(s)\, ds.$$

Интегрируя еще раз предыдущее соотношение и учитывая условие $w(a) = 0$, получим

$$w(x) = (x-a)\, w'(a) - \int_a^x \left(\int_a^t f(s)\, ds \right) dt.$$

Из условия $w(b) = 0$ получаем, что

$$w'(a) = \frac{1}{b-a} \int_a^b \left(\int_a^t f(s)\, ds \right) dt,$$

и, следовательно,

$$w(x) = \frac{x-a}{b-a} \int_a^b \left(\int_a^t f(s)\, ds \right) dt - \int_a^x \left(\int_a^t f(s)\, ds \right) dt. \tag{13}$$

Решение краевой задачи (8), (9) есть сумма функций (12) и (13).

Для численного решения задачи (8), (9) введем на отрезке $[a, b]$ равномерную сетку с шагом h согласно (2) и заменим $u''(x_i)$ второй разностной производной $u_{\bar{x}x,i}$. Тогда вместо дифференциального уравнения (8) получим разностное уравнение второго порядка

$$\frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} = -f_i. \tag{14}$$

Это уравнение можно записать для $i = 1, 2, \ldots, N-1$, т. е. во всех внутренних точках сетки ω_h. В граничных точках в соответствии с (9) следует положить

$$u_0 = \mu_1, \quad u_N = \mu_2. \tag{15}$$

Таким образом, применение разностного метода позволяет заменить исходную дифференциальную задачу (8), (9) системой из $(N-1)$ линейных алгебраических уравнений (14), (15) относительно неизвестных $u_1, u_2, \ldots, u_{N-1}$. Система уравнений (14), (15) называется *разностной схемой* или *разностной краевой задачей*, соответствующей исходной дифференциальной задаче (8)—(9). В дальнейшем, чтобы не было путаницы в обозначениях, будем через $u(x)$ обозначать решение дифференциальной задачи и через $y_i = y(x_i)$ — решение разностной задачи.

Итак, мы получили разностную схему

$$\frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} = -f_i, \quad i = 1, 2, \ldots, N-1, \tag{16}$$

$$y_0 = \mu_1, \quad y_N = \mu_2.$$

В связи с этой разностной схемой возникают следующие проблемы, которые типичны для разностных методов вообще. Во-первых, необходимо убедиться, что система линейных алгебраических уравнений (16) имеет единственное решение, и указать алгоритм, позволяющий получить это решение. И, во-вторых, надо показать, что при стремлении шага сетки h к нулю решение разностной задачи будет сходиться к решению исходной дифференциальной задачи. Вопросы разрешимости и сходимости разностной задачи (16) будут исследованы в п. 6.

Построим по аналогии с (13) точное решение разностной задачи (16). Представим y_i в виде суммы

$$y_i = v_i + w_i, \quad i = 0, 1, \ldots, N,$$

где

$$v_{\bar{x}x,i} = 0, \quad i = 1, 2, \ldots, N-1, \quad v_0 = \mu_1, \quad v_N = \mu_2, \tag{17}$$

$$w_{\bar{x}x,j} = -f_j, \quad j = 1, 2, \ldots, N-1, \quad w_0 = w_N = 0. \tag{18}$$

Запишем (17) подробнее:

$$v_{i-1} - 2v_i + v_{i+1} = 0, \quad i = 1, 2, \ldots, N-1,$$

$$v_0 = \mu_1, \quad v_N = \mu_2,$$

и заметим, что соответствующее характеристическое уравнение

$$q^2 - 2q + 1 = 0$$

имеет кратный корень $q = 1$. Поэтому согласно (25) из § 3, решение разностной краевой задачи (17) имеет вид

$$v_i = \left(1 - \frac{i}{N}\right)\mu_1 + \frac{i}{N}\mu_2.$$

Учитывая, что

$$\frac{i}{N} = \frac{x_i - a}{b - a},$$

можно записать v_i в виде, аналогичном (12), т. е.

$$v_i = \frac{b - x_i}{b - a}\mu_1 + \frac{x_i - a}{b - a}\mu_2. \tag{19}$$

Найдем явное выражение для w_i. Для этого перепишем уравнение (18) в виде

$$w_{\bar{x},j+1} - w_{\bar{x},j} = -hf_j, \quad j = 1, 2, \ldots, N-1,$$

и просуммируем по j от 1 до k. Тогда получим

$$w_{\bar{x},k+1} = w_{\bar{x},1} - \sum_{j=1}^{k} hf_j$$

или

$$w_{k+1} - w_k = hw_{\bar{x},1} - h\sum_{j=1}^{k} hf_j, \quad k = 1, 2, \ldots, N-1.$$

Суммируя последнее уравнение по k от 1 до $i-1$ и учитывая, что $w_0 = 0$, получим

$$w_i = ihw_{\bar{x},1} - \sum_{k=1}^{i-1} h \sum_{j=1}^{k} hf_j.$$

Отсюда и из условия $w_N = 0$ находим

$$w_{\bar{x},1} = \frac{1}{b-a} \sum_{k=1}^{N-1} h \sum_{j=1}^{k} hf_j,$$

следовательно,

$$w_i = \frac{x_i - a}{b-a} \sum_{k=1}^{N-1} h \sum_{j=1}^{k} hf_j - \sum_{k=1}^{i-1} h \sum_{j=1}^{k} hf_j, \quad i = 2, 3, \ldots, N-1, \quad (20)$$

$$w_1 = \frac{x_1 - a}{b-a} \sum_{k=1}^{N-1} h \sum_{j=1}^{k} hf_j.$$

Формула (20) является разностным аналогом формулы (13).

3. Некоторые разностные тождества. Для сеточных функций выполняются разностные аналоги некоторых формул дифференциального и интегрального исчисления. Для простоты изложения будем рассматривать равномерную сетку (2). Разностными аналогами формулы дифференцирования произведения $(uv)' = u'v + uv'$ являются тождества

$$(yv)_{\bar{x},i} = y_i v_{\bar{x},i} + v_{i-1} y_{\bar{x},i},$$

$$(yv)_{x,i} = y_i v_{x,i} + y_{x,i} v_{i+1}. \quad (21)$$

Суммируя (21) по i от 1 до $N-1$, получим

$$y_N v_N - y_1 v_1 = \sum_{i=1}^{N-1} hy_i v_{x,i} + \sum_{i=1}^{N-1} hy_{x,i} v_{i+1}$$

или

$$\sum_{i=1}^{N-1} hy_i v_{x,i} = -\sum_{i=2}^{N} hy_{\bar{x},i} v_i + y_N v_N - y_1 v_1.$$

Учитывая, что

$$y_1 v_1 = v_1 (y_1 - y_0) + v_1 y_0 = h v_1 y_{\bar{x},1} + v_1 y_0,$$

получим

$$\sum_{i=1}^{N-1} h y_i v_{x,i} = - \sum_{i=1}^{N} h v_i y_{\bar{x},i} + y_N v_N - y_0 v_1.$$

Обозначая

$$(w, z) = \sum_{i=1}^{N-1} h w_i z_i,$$

$$(w, z] = \sum_{i=1}^{N} h w_i z_i, \tag{22}$$

перепишем последнее тождество в виде

$$(y, v_x) = - (v, y_{\bar{x}}] + y_N v_N - y_0 v_1. \tag{23}$$

Тождество (23) является разностным аналогом формулы интегрирования по частям

$$\int_a^b y(x) v'(x)\, dx = - \int_a^b v(x) y'(x)\, dx + y(b) v(b) - y(a) v(a)$$

и называется *формулой суммирования по частям.*

4. Разностная задача на собственные значения. Задача на собственные значения

$$u''(x) + \lambda u(x) = 0, \quad a < x < b, \quad u(a) = u(b) = 0 \tag{24}$$

имеет решение

$$\lambda_k = \left(\frac{\pi k}{b-a} \right)^2, \quad u_k(x) = \sin \frac{\pi k (x-a)}{b-a}, \quad k = 1, 2, \ldots$$

Рассмотрим на равномерной сетке (2) разностный аналог задачи (24),

$$\frac{y_{j-1} - 2y_j + y_{j+1}}{h^2} + \lambda^{(h)} y_j = 0, \quad j = 1, 2, \ldots, N-1, \tag{25}$$

$$y_0 = y_N = 0, \quad hN = b - a, \quad y_j = y(x_j), \quad x_j = a + jh.$$

Система уравнений (25) представляет собой задачу на собственные значения

$$Ay = \lambda^{(h)} y$$

для симметричной матрицы

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \ldots & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \ldots & 0 & -1 & 2 \end{bmatrix}$$

порядка $N-1$. Поэтому существует ровно $N-1$ вещественных собственных значений $\lambda_k^{(h)}$, $k=1, 2, \ldots, N-1$, матрицы A. Построим в явном виде собственные значения и собственные функции задачи (25).

Перепишем разностное уравнение (25) в виде

$$y_{j-1}-(2-\mu)y_j+y_{j+1}=0, \quad \mu=h^2\lambda^{(h)}, \tag{26}$$

и рассмотрим отвечающее (26) характеристическое уравнение

$$q^2-(2-\mu)q+1=0. \tag{27}$$

Общее решение уравнения (26) имеет вид

$$y_j=c_1q_1^j+c_2q_2^j, \tag{28}$$

где c_1, c_2 — произвольные постоянные и q_1, q_2 — корни уравнения (27). Из граничных условий $y_0=y_N=0$ получаем

$$c_1+c_2=0, \quad c_1q_1^N+c_2q_2^N=0.$$

Эта однородная система уравнений имеет нетривиальное решение при условии

$$q_1^N=q_2^N.$$

Учитывая, что $q_1q_2=1$, приходим к условию

$$q_1^{2N}=1. \tag{29}$$

Отсюда, представляя q_1 в тригонометрической форме

$$q_1=\rho e^{i\varphi},$$

получим $\rho=1$ и

$$\varphi=\frac{\pi k}{N}, \quad k=1, 2, \ldots, N-1. \tag{30}$$

С другой стороны, из уравнения (27) имеем

$$q_1=1-\frac{\mu}{2}+\sqrt{\left(1-\frac{\mu}{2}\right)^2-1},$$

следовательно,

$$\cos\varphi=1-0{,}5\mu,$$

и из (30) получим

$$\mu=2(1-\cos\varphi)=4\sin^2\frac{\varphi}{2}=4\sin^2\frac{\pi k}{2N}.$$

Таким образом, собственные числа задачи (25) имеют вид

$$\lambda_k^{(h)}=\frac{4}{h^2}\sin^2\frac{\pi k}{2N}, \quad k=1, 2, \ldots, N-1, \tag{31}$$

где $hN=b-a$.

Собственные функции y_j вычисляются согласно (28), где $c_2=-c_1$. Так как $q_1q_2=1$, то

$$y_j=c_1(q_1^j-q_2^j)=c_1(q_1^j-q_1^{-j})=c_1(e^{ij\varphi}-e^{-ij\varphi}),$$

где φ определено согласно (30). Полагая $c_1 = -0{,}5i$, получим

$$y_j^{(k)} = \sin\frac{\pi kj}{N}, \quad k, j = 1, 2, \ldots, N-1. \tag{32}$$

Собственные функции (32) определены с точностью до произвольного постоянного (не зависящего от j) множителя.

Интересно сопоставить решения дифференциальной (24) и разностной (25) задач на собственные значения. Значения собственных функций (32) разностной задачи совпадают в точках сетки со значениями собственных функций дифференциальной задачи. Спектр дифференциальной задачи не ограничен, т. е. $\lambda_k \to \infty$ при $k \to \infty$, в то время как спектр разностной задачи ограничен сверху при каждом фиксированном шаге h числом $4h^{-2}$. Для каждого фиксированного номера $k \leqslant k_0$, где k_0 не зависит от h, собственные значения $\lambda_k^{(h)}$ разностной задачи сходятся при $h \to 0$ к соответствующему собственному значению λ_k дифференциальной задачи, т. е.

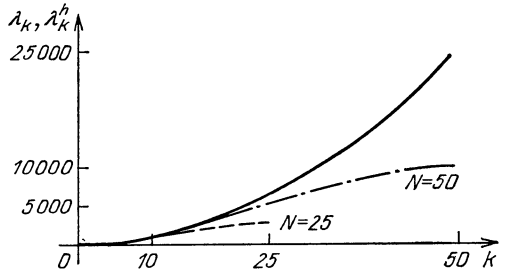


Рис. 3. Собственные значения дифференциальной задачи (сплошная черта) и разностной схемы

$$\lim_{h\to 0} \frac{4}{h^2}\sin^2\frac{\pi kh}{2(b-a)} = \left(\frac{\pi k}{b-a}\right)^2 = \lambda_k.$$

При этом собственные значения разностной задачи (25) всегда меньше соответствующих собственных значений дифференциальной задачи (24). Погрешность $\lambda_k - \lambda_k^{(h)}$ минимальна для малых номеров k и сильно возрастает с ростом k. На рис. 3 изображены графики λ_k (сплошная черта) и $\lambda_k^{(h)}$ в зависимости от номера k для значений $a=0$, $b=1$, $N=25$ и $N=50$.

5. Свойства собственных значений и собственных функций. Перечислим свойства собственных значений и собственных функций разностной задачи (25). Прежде всего из (31) видно, что

$$0 < \lambda_1^{(h)} < \lambda_2^{(h)} < \ldots < \lambda_k^{(h)} < \lambda_{k+1}^{(h)} < \ldots < \lambda_{N-1}^{(h)} < \frac{4}{h^2}.$$

Последнее неравенство неулучшаемо, так как

$$\lambda_{N-1}^{(h)} = \frac{4}{h^2}\cos^2\frac{\pi h}{2(b-a)}$$

и $\cos^2\dfrac{\pi h}{2(b-a)} \to 1$ при $h \to 0$. Оценку снизу для наименьшего

собственного значения λ_1 можно уточнить. Обозначая $\alpha = \pi h/(2(b-a))$, получим

$$\lambda_1^{(h)} = \lambda_1 \left(\frac{\sin\alpha}{\alpha}\right)^2,$$

где $\lambda_1 = \left(\frac{\pi}{b-a}\right)^2$ — наименьшее собственное значение дифференциальной задачи. Не ограничивая общности, можно предположить, что $h \leqslant (b-a)/3$. Тогда $\alpha \leqslant \pi/6$, и поскольку функция $\sin\alpha/\alpha$ монотонно убывает при $\alpha \in [0, \pi/6]$, получим

$$\left(\frac{\sin\alpha}{\alpha}\right)^2 \geqslant \left(\frac{1}{2}\frac{6}{\pi}\right)^2 = \frac{9}{\pi^2},$$

т. е.

$$\lambda_1^{(h)} \geqslant 9/(b-a)^2. \tag{33}$$

Таким образом, наименьшее собственное значение задачи (25) отделено от нуля константой $\delta_1 = 9/(b-a)^2$, не зависящей от h.

Покажем теперь, что собственные функции (32) задачи (25), отвечающие различным собственным значениям, ортогональны в смысле скалярного произведения

$$(u, v) = \sum_{j=1}^{N-1} u_j v_j h. \tag{34}$$

Запишем уравнение (25) для функций $y^{(k)}$ и $y^{(l)}$ в виде

$$y^{(k)}_{\overline{x}x,j} + \lambda_k^{(h)} y_j^{(k)} = 0, \tag{35}$$

$$y^{(l)}_{\overline{x}x,j} + \lambda_l^{(h)} y_j^{(l)} = 0. \tag{36}$$

Умножим уравнение (35) скалярно на $y^{(l)}$, уравнение (36) — на $y^{(k)}$ и вычтем из первого полученного равенства второе. Тогда будем иметь

$$(y^{(k)}_{\overline{x}x}, y^{(l)}) - (y^{(l)}_{\overline{x}x}, y^{(k)}) = (\lambda_l^{(h)} - \lambda_k^{(h)})(y^{(k)}, y^{(l)}). \tag{37}$$

Из разностного аналога формулы интегрирования по частям (23), учитывая условия $y_0^{(l)} = y_N^{(l)} = 0$, получим

$$(y^{(k)}_{\overline{x}x}, y^{(l)}) = -(y^{(k)}_{\overline{x}}, y^{(l)}_{\overline{x}}]$$

и точно так же

$$(y^{(l)}_{\overline{x}x}, y^{(k)}) = -(y^{(l)}_{\overline{x}}, y^{(k)}_{\overline{x}}] = -(y^{(k)}_{\overline{x}}, y^{(l)}_{\overline{x}}].$$

Следовательно, левая часть равенства (37) обращается в нуль, и поскольку $\lambda_l^{(h)} \neq \lambda_k^{(h)}$ при $k \neq l$, получаем

$$(y^{(k)}, y^{(l)}) = 0, \text{ если } k \neq l.$$

Множество функций

$$y = (y_0, y_1, y_2, \ldots, y_{N-1}, y_N), \quad y_j = y(x_j),$$

заданных на сетке (2) и удовлетворяющих нулевым граничным

условиям $y_0=y_N=0$, образует $(N-1)$-мерное линейное пространство H относительно покоординатного сложения и умножения на число. Собственные функции $y^{(k)}$, $k=1, 2, \ldots, N-1$, задачи (25) ортогональны и, следовательно, линейно независимы в H. Тем самым множество собственных функций задачи (25) образует ортогональный базис в H. Нетрудно показать, что

$$\|y^{(k)}\|^2=\sum_{j=1}^{N-1} h\,(y_j^{(k)})^2=0{,}5\,(b-a)$$

для всех $k=1, 2, \ldots, N-1$. Следовательно, множество собственных функций $\mu^{(k)}$, $k=1, 2, \ldots,$ с координатами

$$\mu_j^{(k)}=\sqrt{\frac{2}{b-a}}\sin\frac{\pi kj}{N}, \quad j=1, 2, \ldots, N-1,$$

образует в H ортонормированный базис. Любой элемент $y\in H$ можно единственным образом представить в виде разложения

$$y=\sum_{k=1}^{N-1} c_k\mu^{(k)}.$$

6. Разрешимость и сходимость разностной задачи. Обратимся к исследованию разностной задачи (16). Прежде всего теперь можно утверждать, что система линейных алгебраических уравнений (16) имеет единственное решение. Действительно, в предыдущем пункте показано, что матрица системы (16) не имеет нулевых собственных значений. Поэтому отвечающая (16) однородная система уравнений

$$y_{i-1}-2y_i+y_{i+1}=0, \quad i=1, 2, \ldots, N-1, \quad y_0=y_N=0$$

имеет только тривиальное решение и, следовательно, неоднородная система (16) имеет единственное решение.

Исследуем сходимость при $h\to 0$ решения разностной задачи (16) к решению исходной дифференциальной задачи (8)—(9). Обозначим через

$$z_i=y_i-u(x_i), \quad x_i\in\omega_h,$$

погрешность в точке x_i, т. е. разность между решениями задач (25) и (8)—(9). Подставляя в (16) вместо y_i сумму $z_i+u(x_i)$, $i=1, 2, \ldots, N-1$, получим, что погрешность удовлетворяет разностному уравнению

$$\frac{z_{i-1}-2z_i+z_{i+1}}{h^2}=-\psi_i, \quad i=1, 2, \ldots, N-1, \quad z_0=z_N=0, \tag{38}$$

где

$$\psi_i=u_{\overline{xx},i}+f_i. \tag{39}$$

Сеточная функция ψ_i называется *погрешностью аппроксимации* или *невязкой* разностной схемы (16) на решении задачи (8)—(9).

Записывая ψ_i в виде

$$\psi_i = (u_{\bar{x}x,i} - u''(x_i)) + (u''(x_i) + f_i)$$

и учитывая, что согласно (8) выполняется равенство $u''(x_i)+f_i=0$, получим

$$\psi_i = u_{\bar{x}x,i} - u''(x_i).$$

Разложение по формуле Тейлора показывает, что если $u^{\mathrm{IV}}(x)$ ограничена, то $\psi_i = O(h^2)$ при $h \to 0$. По этой причине говорят, что разностная схема (16) имеет *второй порядок аппроксимации на решении исходной задачи* (8)—(9). Наша ближайшая цель — показать, что схема (16) сходится, т. е. $z_i \to 0$ при $h \to 0$ и, более того, имеет второй порядок точности, т. е. $z_i = O(h^2)$.

Воспользуемся возможностью получить решение задачи (38) в явном виде. Уравнение (38) отличается от изученного ранее уравнения (18) только обозначениями. Поэтому согласно (20) решение задачи (38) представляется в виде

$$z_i = \frac{x_i - a}{b-a} \sum_{k=1}^{N-1} h \sum_{j=1}^{k} h\psi_j - \sum_{k=1}^{i-1} h \sum_{j=1}^{k} h\psi_j, \quad j = 2, 3, \ldots, N, \qquad (40)$$

$$z_1 = h(b-a)^{-1} \sum_{k=1}^{N-1} h \sum_{j=1}^{k} h\psi_j.$$

Из разложения по формуле Тейлора

$$u_{\bar{x}x,j} = u''(x_j) + \frac{h^2}{12} u^{\mathrm{IV}}(\zeta_j)$$

и ограниченности $u^{\mathrm{IV}}(x)$ следует, что существует постоянная $M_1 > 0$, не зависящая от h и от j и такая, что

$$|\psi_j| \leqslant M_1 h^2, \quad j = 1, 2, \ldots, N-1.$$

Поэтому из формулы (40) следует оценка

$$|z_i| \leqslant (M_1 h^2) \left[\frac{x_i - a}{b-a} h^2 \frac{(N-1)N}{2} + h^2 \frac{i(i-1)}{2} \right],$$

т. е.

$$|z_i| \leqslant (M_1 h^2) \left[\frac{i}{N} (N-1)N + i(i-1) \right] \frac{h^2}{2}.$$

Выражение в квадратных скобках равно $i(N+i-2)$ и оценивается сверху числом $2N^2$. Таким образом,

$$|z_i| \leqslant (M_1 h^2) N^2 h^2 = M_1 (b-a)^2 h^2,$$

т. е. $|z_i| = O(h^2)$ при $h \to 0$, $i = 1, 2, \ldots, N-1$. В этом случае говорят, что *схема имеет второй порядок точности.*

Отметим, что приведенный здесь способ исследования сходимости, основанный на явном представлении решения разностной задачи, непригоден для более сложных задач. Другие методы исследования сходимости разностных схем излагаются в части III.

7. Метод прогонки. Система уравнений (16) представляет собой частный случай систем линейных алгебраических уравнений

$$Ay=f$$

с трехдиагональной матрицей $A=[a_{ij}]$, т. е. с матрицей, все элементы которой, не лежащие на главной и двух побочных диагоналях, равны нулю ($a_{ij}=0$ при $j>i+1$ и $j<i-1$).

В общем случае системы линейных алгебраических уравнений с трехдиагональной матрицей имеют вид

$$a_jy_{j-1}-c_jy_j+b_jy_{j+1}=-f_j, \quad j=1, 2, \ldots, N-1, \tag{41}$$

$$y_0=\varkappa_1y_1+\mu_1, \quad y_N=\varkappa_2y_{N-1}+\mu_2. \tag{42}$$

Для численного решения систем с трехдиагональными матрицами применяется *метод прогонки*, который представляет собой вариант метода последовательного исключения неизвестных. Особенно широкое применение метод прогонки получил при решении систем разностных уравнений, возникающих при аппроксимации краевых задач для дифференциальных уравнений второго порядка.

Приведем вывод расчетных формул метода прогонки. Будем искать решение системы (41) в виде

$$y_j=\alpha_{j+1}y_{j+1}+\beta_{j+1}, \quad j=0, 1, \ldots, N-1, \tag{43}$$

где α_{j+1}, β_{j+1} — неизвестные пока коэффициенты. Отсюда найдем

$$y_{j-1}=\alpha_jy_j+\beta_j=\alpha_j(\alpha_{j+1}y_{j+1}+\beta_{j+1})+\beta_j=$$
$$=\alpha_j\alpha_{j+1}y_{j+1}+(\alpha_j\beta_{j+1}+\beta_j), \quad j=1, 2, \ldots, N-1.$$

Подставляя полученные выражения для y_j, y_{j-1} в уравнение (41), приходим при $j=1, 2, \ldots, N-1$ к уравнению

$$[\alpha_{j+1}(a_j\alpha_j-c_j)+b_j]y_{j+1}+[\beta_{j+1}(a_j\alpha_j-c_j)+a_j\beta_j+f_j]=0.$$

Последнее уравнение будет выполнено, если коэффициенты α_{j+1}, β_{j+1} выбрать такими, чтобы выражения в квадратных скобках обращались в нуль. А именно, достаточно положить

$$\alpha_{j+1}=\frac{b_j}{c_j-\alpha_ja_j}, \quad \beta_{j+1}=\frac{a_j\beta_j+f_j}{c_j-\alpha_ja_j}, \quad j=1, 2, \ldots, N-1. \tag{44}$$

Соотношения (44) представляют собой нелинейные разностные уравнения первого порядка. Для их решения необходимо задать начальные значения α_1, β_1. Эти начальные значения находим из требования эквивалентности условия (43) при $j=0$, т. е. условия $y_0=\alpha_1y_1+\beta_1$, первому из уравнений (42). Таким образом, получаем

$$\alpha_1=\varkappa_1, \quad \beta_1=\mu_1. \tag{45}$$

Нахождение коэффициентов α_{j+1}, β_{j+1} по формулам (44), (45) называется *прямой прогонкой*. После того как прогоночные коэффициенты α_{j+1}, β_{j+1}, $j=0, 1, \ldots, N-1$, найдены, решение системы (41), (42) находится по рекуррентной формуле (43), начиная с

$j=N-1$. Для начала счета по этой формуле требуется знать y_N, которое определяется из уравнений

$$y_N=\varkappa_2 y_{N-1}+\mu_2, \quad y_{N-1}=\alpha_N y_N+\beta_N$$

и равно $(\varkappa_2\beta_N+\mu_2)/(1-\varkappa_2\alpha_N)$. Нахождение y_j по формулам

$$y_j=\alpha_{j+1}y_{j+1}+\beta_{j+1}, \quad j=N-1,\ N-2,\ \ldots,\ 0,$$

$$y_N=\frac{\varkappa_2\beta_N+\mu_2}{1-\varkappa_2\alpha_N} \tag{46}$$

называется *обратной прогонкой*. Алгоритм решения системы (41), (42), определяемый по формулам (44)—(46), называется *методом прогонки*. Применяются и другие варианты метода прогонки (см. [32]).

Метод прогонки можно применять, если знаменатели выражений (44), (46) не обращаются в нуль. Покажем, что *для возможности применения метода прогонки достаточно потребовать, чтобы коэффициенты системы* (41), (42) *удовлетворяли условиям*

$$a_j\neq 0, \quad b_j\neq 0, \quad |c_j|\geqslant |a_j|+|b_j|, \quad j=1,\ 2,\ \ldots,\ N-1, \tag{47}$$

$$|\varkappa_1|\leqslant 1, \quad |\varkappa_2|<1. \tag{48}$$

Заметим, что числа a_j, b_j, c_j, $\varkappa_1$, $\varkappa_2$ могут быть комплексными.

Сначала докажем по индукции, что при условиях (47), (48) модули прогоночных коэффициентов α_j, $j=1,\ \ldots,\ N-1$, не превосходят единицы. Согласно (45), (48) имеем $|\alpha_1|=|\varkappa_1|\leqslant 1$. Предположим, что $|\alpha_j|\leqslant 1$ для некоторого j и докажем, что $|\alpha_{j+1}|\leqslant 1$. Из оценок

$$|c_j-\alpha_j a_j|\geqslant\big||c_j|-|\alpha_j||a_j|\big|\geqslant\big||c_j|-|a_j|\big|$$

и условий (47) получаем

$$|c_j-\alpha_j a_j|\geqslant |b_j|>0,$$

т. е. знаменатели выражений (44) не обращаются в нуль. Более того,

$$|\alpha_{j+1}|=\frac{|b_j|}{|c_j-\alpha_j a_j|}\leqslant 1.$$

Следовательно, $|\alpha_j|\leqslant 1$, $j=1,\ 2,\ \ldots,\ N$. Далее, учитывая второе из условий (48) и только что доказанное неравенство $|\alpha_N|\leqslant 1$, имеем

$$|1-\varkappa_2\alpha_N|\geqslant 1-|\varkappa_2||\alpha_N|\geqslant 1-|\varkappa_2|>0,$$

т. е. не обращается в нуль и знаменатель в выражении для y_N.

К аналогичному выводу можно прийти и в том случае, когда условия (47), (48) заменяются условиями

$$a_j\neq 0, \quad b_j\neq 0, \quad |c_j|>|a_j|+|b_j|, \quad j=1,\ 2,\ \ldots,\ N-1, \tag{49}$$

$$|\varkappa_1|\leqslant 1, \quad |\varkappa_2|\leqslant 1. \tag{50}$$

В этом случае из предположения $|\alpha_j|\leqslant 1$ следует

$$|c_j-\alpha_j a_j|\geqslant\big||c_j|-|a_j|\big|>|b_j|, \quad |\alpha_{j+1}|<1,$$

т. е. все прогоночные коэффициенты, начиная со второго, по модулю строго меньше единицы. При этом $|1-\varkappa_2\alpha_N| \geqslant 1-|\varkappa_2||\alpha_N| \geqslant \geqslant 1-|\alpha_N|>0$.

Таким образом при выполнении условий (47), (48) (так же как и условий (49), (50)) система (41)—(42) эквивалентна системе (44)—(46). Поэтому условия (47), (48) (или условия (49), (50)) гарантируют существование и единственность решения системы (41), (42) и возможность нахождения этого решения методом прогонки. Кроме того, доказанные неравенства $|\alpha_j| \leqslant 1$, $j=1, 2, \ldots, N$, обеспечивают устойчивость счета по рекуррентным формулам (46). Последнее означает, что погрешность, внесенная на каком-либо шаге вычислений, не будет возрастать при переходе к следующим шагам. Действительно, пусть в формуле (46) при $j=j_0+1$ вместо y_{j_0+1} вычислена величина $\tilde{y}_{j_0+1}=y_{j_0+1}+\delta_{j_0+1}$. Тогда на следующем шаге вычислений, т. е. при $j=j_0$, вместо $y_{j_0}=\alpha_{j_0+1}y_{j_0+1}+\beta_{j_0+1}$ получим величину $\tilde{y}_{j_0}=\alpha_{j_0+1}(y_{j_0+1}+\delta_{j_0+1})+\beta_{j_0+1}$ и погрешность окажется равной

$$\delta_{j_0}=\tilde{y}_{j_0}-y_{j_0}=\alpha_{j_0+1}\delta_{j_0+1}.$$

Отсюда получим, что $|\delta_{j_0}| \leqslant |\alpha_{j_0+1}||\delta_{j_0+1}| \leqslant |\delta_{j_0+1}|$, т. е. погрешность не возрастает.

Отметим, что для разностной краевой задачи (16), записанной в виде

$$y_{j-1}-2y_j+y_{j+1}=-h^2f_j, \quad j=1, 2, \ldots, N-1,$$

имеем $a_j=b_j=1$, $c_j=2$, $\varkappa_1=\varkappa_2=0$. Поэтому выполнены условия устойчивости (47), (48) и решение задачи (16) можно отыскивать методом прогонки.

ЧАСТЬ II

ЧИСЛЕННЫЕ МЕТОДЫ АЛГЕБРЫ И АНАЛИЗА

ГЛАВА 1

ПРЯМЫЕ МЕТОДЫ РЕШЕНИЯ СИСТЕМ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

В главах 1, 2 рассматриваются численные методы решения систем линейных алгебраических уравнений

$$Ax=f, \tag{1}$$

где A — матрица $m\times m$, $x=(x_1, x_2, \ldots, x_m)^T$ — искомый вектор, $f=(f_1, f_2, \ldots, f_m)^T$ — заданный вектор. Предполагается, что определитель матрицы A отличен от нуля, так что решение x существует и единственно. Для большинства вычислительных задач характерным является большой порядок матрицы A. Из курса алгебры известно, что систему (1) можно решить по крайней мере двумя способами: либо по формулам Крамера, либо методом последовательного исключения неизвестных (методом Гаусса). При больших m первый способ, основанный на вычислении определителей, требует порядка $m!$ арифметических действий, в то время как метод Гаусса — только $O(m^3)$ действий. Поэтому метод Гаусса в различных вариантах широко используется при решении на ЭВМ задач линейной алгебры.

Методы численного решения системы (1) делятся на две группы: прямые методы и итерационные методы. В *прямых* (или точных) *методах* решение x системы (1) находится за конечное число арифметических действий. Примером прямого метода является метод Гаусса. Отметим, что вследствие погрешностей округления при решении задач на ЭВМ прямые методы на самом деле не приводят к точному решению системы (1) и называть их точными можно лишь отвлекаясь от погрешностей округления. Сопоставление различных прямых методов проводится обычно по числу арифметических действий (а еще чаще — по асимптотике при больших m числа арифметических действий), необходимых для получения решения. При прочих равных условиях предпочтение отдается методу с меньшим числом действий.

Итерационные методы (их называют также *методами последовательных приближений*) состоят в том, что решение x системы (1) находится как предел при $n\to\infty$ последовательных приближений $x^{(n)}$, где n — номер итерации. Как правило, за конечное число итераций этот предел не достигается. Обычно задается некоторое ма-

лое число $\varepsilon>0$ (точность) и вычисления проводятся до тех пор, пока не будет выполнена оценка

$$\|x^{(n)}-x\|<\varepsilon. \tag{2}$$

Число итераций $n=n(\varepsilon)$, которое необходимо провести для получения заданной точности ε (т. е. для выполнения оценки (2)), для многих методов можно найти из теоретических рассмотрений. Качество различных итерационных процессов можно сравнивать по необходимому числу итераций $n(\varepsilon)$.

К решению систем линейных алгебраических уравнений сводится подавляющее большинство задач вычислительной математики. В настоящее время предложено колоссальное количество алгоритмов решения задач линейной алгебры (см. [8, 35]), большинство из которых рассчитано на матрицы A специального вида (трехдиагональные, симметричные, ленточные, большие разреженные матрицы).

Прямые методы, которые рассматриваются в гл. 1, не предполагают, что матрица A имеет какой-либо специальный вид. На практике они применяются для матриц умеренного порядка (порядка ста). Итерационные методы, рассмотренные в гл. 2, можно применять и для матриц высокого порядка, однако их сходимость не очень быстрая. Более совершенные прямые и итерационные методы, учитывающие структуру матрицы, излагаются в части III.

§ 1. Метод Гаусса численного решения систем линейных алгебраических уравнений

1. Основная идея метода. В ближайших двух главах рассматриваются численные методы решения системы линейных алгебраических уравнений

$$Ax=f, \tag{1}$$

где A — вещественная квадратная матрица порядка m, а f — заданный и x — искомый векторы. Будем предполагать, что определитель матрицы A отличен от нуля. Тогда для каждого вектора f система (1) имеет единственное решение.

Запишем систему (1) в развернутом виде

$$\begin{aligned} a_{11}x_1+a_{12}x_2+\ldots+a_{1m}x_m&=f_1,\\ a_{21}x_1+a_{22}x_2+\ldots+a_{2m}x_m&=f_2,\\ \ldots\ldots\ldots\ldots\ldots\ldots&\\ a_{m1}x_1+a_{m2}x_2+\ldots+a_{mm}x_m&=f_m. \end{aligned} \tag{2}$$

Метод Гаусса решения системы (2) состоит в последовательном исключении неизвестных $x_1, x_2, \ldots, x_m$ из этой системы. Предположим, что $a_{11}\neq 0$. Поделив первое уравнение на a_{11}, получим

$$x_1+c_{12}x_2+\cdots+c_{1m}x_m=y_1, \tag{3}$$

где

$$c_{1j}=\frac{a_{1j}}{a_{11}},\quad j=2,\ldots,m,\qquad y_1=\frac{f_1}{a_{11}}.$$

Рассмотрим теперь оставшиеся уравнения системы (2):

$$a_{i1}x_1+a_{i2}x_2+\ldots+a_{im}x_m=f_i, \quad i=2, 3, \ldots, m. \tag{4}$$

Умножим (3) на a_{i1} и вычтем полученное уравнение из i-го уравнения системы (4), $i=2, \ldots, m$. В результате получим следующую систему уравнений:

$$\begin{gathered} x_1+c_{12}x_2+\ldots+c_{1j}x_j+\ldots+c_{1m}x_m=y_1, \\ a_{22}^{(1)}x_2+\ldots+a_{2j}^{(1)}x_j+\ldots+a_{2m}^{(1)}x_m=f_2^{(1)}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{m2}^{(1)}x_2+\ldots+a_{mj}^{(1)}x_j+\ldots+a_{mm}^{(1)}x_m=f_m^{(1)}. \end{gathered} \tag{5}$$

Здесь обозначено

$$a_{ij}^{(1)}=a_{ij}-c_{1j}a_{i1}, \quad f_i^{(1)}=f_i-y_1a_{i1}, \quad i, j=2, 3, \ldots, m. \tag{6}$$

Матрица системы (5) имеет вид

$$\begin{bmatrix} 1 & c_{12} & \ldots & c_{1m} \\ 0 & a_{22}^{(1)} & \ldots & a_{2m}^{(1)} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & a_{m2}^{(1)} & \ldots & a_{mm}^{(1)} \end{bmatrix}.$$

Матрицы такой структуры принято обозначать так:

$$\begin{bmatrix} 1 & \times & \ldots & \times \\ 0 & \times & \ldots & \times \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \times & \ldots & \times \end{bmatrix},$$

где крестиками обозначены ненулевые элементы. В системе (5) неизвестное x_1 содержится только в первом уравнении, поэтому в дальнейшем достаточно иметь дело с укороченной системой уравнений

$$\begin{gathered} a_{22}^{(1)}x_2+\ldots+a_{2j}^{(1)}x_j+\ldots+a_{2m}^{(1)}x_m=f_2^{(1)}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{m2}^{(1)}x_2+\ldots+a_{mj}^{(1)}x_j+\ldots+a_{mm}^{(1)}x_m=f_m^{(1)}. \end{gathered} \tag{7}$$

Тем самым мы осуществили первый шаг метода Гаусса. Если $a_{22}^{(1)}\neq 0$, то из системы (7) совершенно аналогично можно исключить неизвестное x_2 и прийти к системе, эквивалентной (2) и имеющей матрицу следующей структуры:

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

При этом первое уравнение системы (5) остается без изменения.

Исключая таким же образом неизвестные $x_3, x_4, \ldots, x_m$, придем окончательно к системе уравнений вида

$$\begin{aligned} x_1 + c_{12}x_2 + \ldots + c_{1m}x_m &= y_1, \\ x_2 + \ldots + c_{2m}x_m &= y_2, \\ \ldots\ldots\ldots\ldots\ldots & \\ x_{m-1} + c_{m-1,m}x_m &= y_{m-1}, \\ x_m &= y_m, \end{aligned} \tag{8}$$

эквивалентной исходной системе (2).

Матрица этой системы

$$C = \begin{bmatrix} 1 & c_{12} & \ldots & c_{1,m-1} & c_{1m} \\ 0 & 1 & \ldots & c_{2,m-1} & c_{2m} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 & c_{m-1,m} \\ 0 & 0 & \ldots & 0 & 1 \end{bmatrix}. \tag{9}$$

содержит нули всюду ниже главной диагонали. Матрицы такого вида называются *верхними треугольными матрицами. Нижней треугольной* называется такая матрица, у которой равны нулю все элементы, расположенные выше главной диагонали.

Получение системы (8) составляет *прямой ход метода Гаусса. Обратный ход* заключается в нахождении неизвестных $x_1, x_2, \ldots, x_m$ из системы (8). Поскольку матрица системы имеет треугольный вид, можно последовательно, начиная с x_m, найти все неизвестные. Действительно, $x_m = y_m$, $x_{m-1} = y_{m-1} - c_{m-1\,m}x_m$ и т. д. Общие формулы обратного хода имеют вид

$$x_i = y_i - \sum_{j=i+1}^{m} c_{ij}x_j, \quad i = m-1, \ldots, 1, \qquad x_m = y_m. \tag{10}$$

2. Расчетные формулы. При реализации на ЭВМ прямого хода метода Гаусса нет необходимости действовать с переменными $x_1, x_2, \ldots, x_m$. Достаточно указать алгоритм, согласно которому исходная матрица A преобразуется к треугольному виду (9), и указать соответствующее преобразование правых частей системы. Получим эти общие формулы. Пусть осуществлены первые $k-1$ шагов, т. е. уже исключены переменные $x_1, x_2, \ldots, x_{k-1}$. Тогда имеем систему

$$\begin{aligned} x_1 + c_{12}x_2 + \ldots + c_{1k}x_k + \ldots + c_{1m}x_m &= y_1, \\ x_2 + \ldots + c_{2k}x_k + \ldots + c_{2m}x_m &= y_2, \\ \ldots\ldots\ldots\ldots\ldots & \\ x_{k-1} + c_{k-1,k}x_k + \ldots + c_{k-1,m}x_m &= y_{k-1}, \\ a_{kk}^{(k-1)}x_k + \ldots + a_{km}^{(k-1)}x_m &= f_k^{(k-1)}, \\ \ldots\ldots\ldots\ldots\ldots & \\ a_{mk}^{(k-1)}x_k + \ldots + a_{mm}^{(k-1)}x_m &= f_m^{(k-1)}. \end{aligned} \tag{11}$$

Рассмотрим k-е уравнение этой системы

$$a_{kk}^{(k-1)}x_k + \ldots + a_{km}^{(k-1)}x_m = f_k^{(k-1)}$$

и предположим, что $a_{kk}^{(k-1)} \neq 0$. Поделив обе части этого уравнения на $a_{kk}^{(k-1)}$, получим

$$x_k + c_{k,k+1}x_{k+1} + \ldots + c_{km}x_m = y_k, \tag{12}$$

где

$$c_{kj} = \frac{a_{kj}^{(k-1)}}{a_{kk}^{(k-1)}}, \quad j = k+1,\ k+2,\ \ldots,\ m,$$

$$y_k = \frac{f_k^{(k-1)}}{a_{kk}^{(k-1)}}.$$

Далее, умножим уравнение (12) на $a_{ik}^{(k-1)}$ и вычтем полученное соотношение из i-го уравнения системы (11), где $i = k+1, k+2, \ldots, m$. В результате последняя группа уравнений системы (11) примет вид

$$\begin{aligned}
&x_k + c_{k,k+1}x_{k+1} + \ldots + c_{km}x_m = y_k,\\
&a_{k+1,k+1}^{(k)}x_{k+1} + \ldots + a_{k+1,m}^{(k)}x_m = f_{k+1}^{(k)},\\
&\ldots\ldots\ldots\ldots\ldots\ldots\\
&a_{m,k+1}^{(k)}x_{k+1} + \ldots + a_{mm}^{(k)}x_m = f_m^{(k)},
\end{aligned}$$

где

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - a_{ik}^{(k-1)}c_{kj}, \quad i, j = k+1,\ k+2,\ \ldots,\ m,$$

$$f_i^{(k)} = f_i^{(k-1)} - a_{ik}^{(k-1)}y_k, \quad i = k+1,\ k+2,\ \ldots,\ m.$$

Таким образом, в прямом ходе метода Гаусса коэффициенты уравнений преобразуются по следующему правилу:

$$a_{kj}^{(0)} = a_{kj}, \quad k, j = 1, 2, \ldots, m,$$

$$c_{kj} = \frac{a_{kj}^{(k-1)}}{a_{kk}^{(k-1)}}, \quad j = k+1,\ k+2,\ \ldots, m, \quad k = 1, 2, \ldots, m, \tag{13}$$

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - a_{ik}^{(k-1)}c_{kj}, \tag{14}$$

$$i, j = k+1,\ k+2,\ \ldots,\ m, \quad k = 1, 2, \ldots, m-1.$$

Вычисление правых частей системы (8) осуществляется по формулам

$$f_k^{(0)} = f_k, \quad y_k = \frac{f_k^{(k-1)}}{a_{kk}^{(k-1)}}, \quad k = 1, 2, \ldots, m, \tag{15}$$

$$f_i^{(k)} = f_i^{(k-1)} - a_{ik}^{(k-1)}y_k, \quad i = k+1,\ k+2,\ \ldots,\ m. \tag{16}$$

Коэффициенты c_{ij} и правые части y_i, $i = 1, 2, \ldots, m$, $j = i+1, i+2, \ldots$

..., m, хранятся в памяти ЭВМ и используются при осуществлении обратного хода по формулам (10).

Основным ограничением метода является предположение о том, что все элементы $a_{kk}^{(k-1)}$, на которые проводится деление, отличны от нуля. Число $a_{kk}^{(k-1)}$ называется *ведущим элементом на k-м шаге исключения.* Даже если какой-то ведущий элемент не равен нулю, а просто близок к нему, в процессе вычислений может происходить сильное накопление погрешностей. Выход из этой ситуации состоит в том, что в качестве ведущего элемента выбирается не $a_{kk}^{(k-1)}$, а другое число (т. е. на k-м шаге исключается не x_k, а другое переменное x_j, $j \neq k$). Наиболее последовательно такая стратегия выбора ведущих элементов осуществлена в методе Гаусса с выбором главного элемента (см. § 3).

3. Подсчет числа действий. Подсчитаем число арифметических действий, необходимых для решения системы (2) с помощью метода Гаусса. Поскольку выполнение операций умножения и деления на ЭВМ требует гораздо больше времени, чем выполнение сложения и вычитания, ограничимся подсчетом числа умножений и делений. Читатель по аналогии может самостоятельно найти требуемое число действий сложения и вычитания.

1. Вычисление коэффициентов c_{kj}, $k=1, 2, \ldots, m$, $j=k+1, k+2, \ldots, m$, по формулам (13) требует

$$\sum_{k=1}^{m} (m-k) = 1 + 2 + \ldots + (m-1) = \frac{m(m-1)}{2}$$

делений.

2. Вычисление всех коэффициентов $a_{ij}^{(k)}$ по формулам (14) требует

$$\sum_{k=1}^{m-1} (m-k)^2 = 1^2 + 2^2 + \ldots + (m-1)^2 = \frac{(m-1)\,m\,(2m-1)}{6}$$

умножений.

Таким образом, вычисление ненулевых элементов c_{ij} треугольной матрицы C требует

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

операций умножения и деления. При больших m это число действий равно приблизительно $m^3/3$.

3. Вычисление правых частей y_k по формулам (15) требует m делений, а нахождение $f_i^{(k)}$ по формулам (16)

$$\sum_{k=1}^{m} (m-k) = \frac{m(m-1)}{2}$$

умножений. Следовательно, вычисление правых частей преобразованной системы (8) требует

$$m+\frac{m(m-1)}{2}=\frac{m(m+1)}{2}$$

действий умножения и деления.

В итоге для осуществления прямого хода метода Гаусса необходимо выполнить

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

действий, из которых основное число действий (порядка $m^3/3$) приходится на вычисление элементов матрицы C.

4. Для осуществления обратного хода метода Гаусса по формулам (10) требуется

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

умножений.

Итак, для реализации метода Гаусса требуется выполнить

$$\frac{m(m+1)(2m+1)}{6}+\frac{m(m-1)}{2}=\frac{m(m^2+3m-1)}{3}$$

действий умножения и деления. Подчеркнем, что основное время расчета затрачивается на осуществление прямого хода. Для больших m число действий умножения и деления в методе Гаусса близко к $m^3/3$. Это означает, что на вычисление одного неизвестного тратится в среднем $m^2/3$ действий. По затратам времени и необходимой машинной памяти метод Гаусса пригоден для решения систем уравнений (2) общего вида с числом неизвестных m порядка 100.

§ 2. Условия применимости метода Гаусса

1. Связь метода Гаусса с разложением матрицы на множители. В предыдущем параграфе было показано, что метод Гаусса преобразует исходную систему уравнений

$$Ax=f \tag{1}$$

в эквивалентную систему

$$Cx=y, \tag{2}$$

где C — верхняя треугольная матрица с единицами на главной диагонали. Выясним теперь, как связаны между собой векторы правых частей f и y. Для этого обратимся к формулам (16) из § 1, из которых последовательно получим

$$f_1=a_{11}y_1,\quad f_2=a_{21}y_1+a_{22}^{(1)}y_2,\ \ldots$$

и вообще

$$f_j=b_{j1}y_1+b_{j2}y_2+\ldots+b_{jj}y_j,\qquad j=1,\ 2,\ \ldots,\ m, \tag{3}$$

где b_{ji} — числовые коэффициенты, причем $b_{jj}=a_{jj}^{(j-1)}$. Соотношения (3) можно записать в матричном виде

$$f=By, \tag{4}$$

где B — нижняя треугольная матрица с элементами $a_{jj}^{(j-1)}$, $j=1, 2, \ldots, m$, $(a_{11}^{(0)}=a_{11})$ на главной диагонали. Напомним, что основное допущение при формулировке метода Гаусса состояло в том, что все $a_{jj}^{(j-1)}\neq 0$. Поэтому на диагонали матрицы B стоят ненулевые элементы, и, следовательно, матрица B имеет обратную.

Подставляя в уравнение (2) выражение для y в виде $y=B^{-1}f$, приходим к уравнению

$$Cx=B^{-1}f,$$

или, что то же самое, к уравнению

$$BCx=f. \tag{5}$$

Сопоставляя (5) с уравнением (1), приходим к выводу, что в результате применения метода Гаусса получено разложение исходной матрицы A в произведение $A=BC$, где B — нижняя треугольная матрица с ненулевыми элементами на главной диагонали и C — верхняя треугольная матрица с единичной главной диагональю.

Теперь мы имеем право трактовать метод Гаусса следующим образом. Пусть заданы матрицы A и вектор f. Сначала проводится разложение A в произведение двух треугольных матриц, $A=BC$. Затем последовательно решаются две системы уравнений

$$By=f, \tag{6}$$

$$Cx=y \tag{7}$$

с треугольными матрицами, откуда и находится искомый вектор x. Разложение $A=BC$ соответствует прямому ходу метода Гаусса, а решение системы (6)—(7) — обратному ходу. Заметим, что в алгоритме, изложенном в § 1, разложение $A=BC$ и решение системы (6) проводится одновременно.

Далее, следуя стандартным обозначениям, нижние треугольные матрицы будем обозначать буквой L (от английского lower — нижний) и верхние треугольные — буквой U (от английского upper — верхний).

2. Теорема об *LU*-разложении. Обозначим через Δ_j угловой минор порядка j матрицы A, т. е.

$$\Delta_1=a_{11},\ \Delta_2=\det\begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix},\ \ldots,\ \Delta_m=\det A.$$

Теоретическое обоснование возможности разложения матрицы в произведение двух треугольных матриц содержит следующая

Теорема 1 (теорема об *LU*-разложении). *Пусть все угловые миноры матрицы A отличны от нуля, $\Delta_j\neq 0$, $j=1, 2, \ldots, m$. Тогда*

матрицу A можно представить, причем единственным образом, в виде произведения

$$A=LU, \tag{8}$$

где L — нижняя треугольная матрица с ненулевыми диагональными элементами и U — верхняя треугольная матрица с единичной диагональю.

Доказательство. Докажем сформулированное утверждение сначала для матриц второго порядка. Будем искать разложение матрицы

$$A=\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

в виде

$$A=\begin{bmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{bmatrix}\begin{bmatrix} 1 & u_{12} \\ 0 & 1 \end{bmatrix},$$

где l_{11}, l_{21}, l_{22}, u_{12}— неизвестные пока числа. Для их нахождения придем к системе уравнений

$$l_{11}=a_{11}, \quad l_{11}u_{12}=a_{12}, \quad l_{21}=a_{21},$$
$$l_{21}u_{12}+l_{22}=a_{22},$$

которая имеет единственное решение

$$l_{11}=a_{11}, \quad u_{12}=a_{12}/a_{11}, \quad l_{21}=a_{21},$$
$$l_{22}=\frac{a_{11}a_{22}-a_{21}a_{12}}{a_{11}}.$$

По предположению теоремы $a_{11}\neq 0$, $a_{11}a_{22}\neq a_{21}a_{12}$, следовательно, элементы l_{11} и l_{22} отличны от нуля.

Дальнейшее доказательство проведем методом индукции. Пусть утверждение теоремы справедливо для матриц порядка k—1; докажем, что оно справедливо и для матриц порядка k. Представим матрицу A порядка k в виде

$$A=\left[\begin{array}{ccc|c} a_{11} & \dots & a_{1,k-1} & a_{1k} \\ \dots & \dots & \dots & \dots \\ a_{k-1,1} & \dots & a_{k-1,k-1} & a_{k-1,k} \\ \hline a_{k1} & \dots & a_{k,k-1} & a_{kk} \end{array}\right] \tag{9}$$

и обозначим

$$A_{k-1}=\begin{bmatrix} a_{11} & \dots & a_{1,k-1} \\ \dots & \dots & \dots \\ a_{k-1,1} & \dots & a_{k-1,k-1} \end{bmatrix}, \quad a_{k-1}=\begin{bmatrix} a_{1k} \\ \dots \\ a_{k-1,k} \end{bmatrix},$$
$$b_{k-1}=(a_{k1}, \dots, a_{k,k-1}).$$

Согласно предположению индукции существует требуемое разложение матрицы A_{k-1}, т. е.

$$A_{k-1}=L_{k-1}U_{k-1},$$

где L_{k-1}, U_{k-1}— соответственно нижняя и верхняя треугольные мат-

рицы, обладающие указанными в теореме свойствами. Будем искать разложение матрицы (9) в виде

$$A = \begin{bmatrix} L_{k-1} & 0 \\ l_{k-1} & l_{kk} \end{bmatrix} \begin{bmatrix} U_{k-1} & u_{k-1} \\ 0 & 1 \end{bmatrix}, \tag{10}$$

где l_{k-1}, u_{k-1} — неизвестные пока векторы,

$$l_{k-1} = (l_{k1}, l_{k2}, \ldots, l_{k,k-1}), \quad u_{k-1} = (u_{1k}, u_{2k}, \ldots, u_{k-1,k})^T.$$

Перемножая матрицы в правой части уравнения (10) и учитывая (9), приходим к системе уравнений

$$L_{k-1}u_{k-1} = a_{k-1}, \tag{11}$$

$$l_{k-1}U_{k-1} = b_{k-1}, \tag{12}$$

$$l_{k-1}u_{k-1} + l_{kk} = a_{kk}. \tag{13}$$

Из предположения индукции следует существование матриц L_{k-1}^{-1}, U_{k-1}^{-1}. Поэтому из (11) и (12) получим

$$u_{k-1} = L_{k-1}^{-1}a_{k-1}, \quad l_{k-1} = b_{k-1}U_{k-1}^{-1}$$

и, далее,

$$l_{kk} = a_{kk} - l_{k-1}u_{k-1}.$$

Таким образом, LU-разложение матрицы A порядка k существует. Остается доказать, что $l_{kk} \neq 0$. Рассмотрим определитель матрицы A. Из разложения (10) следует, что

$$\det A = (\det L_{k-1})\, l_{kk}\, (\det U_{k-1}) = (\det L_{k-1})\, l_{kk}.$$

По условию теоремы $\det A \neq 0$, следовательно, $l_{kk} \neq 0$. Тем самым индукция завершена и доказана возможность требуемого разложения.

Покажем теперь, что такое разложение единственно. Предположим, что матрицу A можно разложить двумя способами:

$$A = L_1U_1 = L_2U_2.$$

Тогда $L_2 = L_1U_1U_2^{-1}$ и

$$U_1U_2^{-1} = L_1^{-1}L_2. \tag{14}$$

Матрица в левой части уравнения (14) является верхней треугольной, а в правой части — нижней треугольной. Такое равенство возможно лишь в случае, если матрицы $U_1U_2^{-1}$ и $L_1^{-1}L_2$ диагональные. Но на диагонали матрицы $U_1U_2^{-1}$ (а следовательно, и матрицы $L_1^{-1}L_2$) стоят единицы, следовательно, эти матрицы единичные:

$$U_1U_2^{-1} = L_1^{-1}L_2 = E.$$

Отсюда получаем $U_1 = U_2$, $L_1 = L_2$, т. е. разложение (8) единственно. Теорема об LU-разложении полностью доказана.

Замечание. Если хотя бы один из угловых миноров матрицы A равен нулю, то указанное LU-разложение невозможно. Это легко видеть на примере матриц второго порядка.

Следствие. *Метод Гаусса можно применять тогда и только тогда, когда все угловые миноры матрицы A отличны от нуля.*

3. Элементарные треугольные матрицы. Мы уже видели, что метод Гаусса приводит к разложению исходной матрицы в произведение двух треугольных. Более детально описать структуру этих треугольных матриц можно с помощью так называемых элементарных треугольных матриц.

Матрица L_j называется *элементарной нижней треугольной матрицей*, если она имеет вид

$$L_j = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ 0 & \dots & l_{jj} & & 0 & & \\ 0 & \dots & l_{j+1,j} & 1 & & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \dots & l_{mj} & 0 & \dots & 1 & \end{bmatrix}.$$

В матрице L_j все элементы главной диагонали кроме l_{jj} равны единице. Из остальных элементов отличными от нуля могут быть только элементы j-го столбца, расположенные ниже l_{jj}. Обратной к L_j является элементарная нижняя треугольная матрица

$$L_j^{-1} = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ 0 & \dots & l_{jj}^{-1} & & 0 & & \\ 0 & \dots & -l_{j+1,j}l_{jj}^{-1} & 1 & & & \\ 0 & \dots & -l_{j+2,j}l_{jj}^{-1} & 0 & \ddots & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \\ 0 & \dots & -l_{mj}l_{jj}^{-1} & 0 & \dots & 0 & 1 \end{bmatrix}.$$

Рассмотрим для наглядности сначала систему $Ax=f$, состоящую из трех уравнений:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= f_1, \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= f_2, \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= f_3. \end{aligned} \tag{15}$$

После первого шага исключения по методу Гаусса преобразованная система принимает вид

$$\begin{aligned} x_1 \quad & + \frac{a_{12}}{a_{11}}x_2 \quad + \frac{a_{13}}{a_{11}}x_3 = \frac{f_1}{a_{11}}, \\ & \left(a_{22} - \frac{a_{21}a_{12}}{a_{11}}\right)x_2 + \left(a_{23} - \frac{a_{21}a_{13}}{a_{11}}\right)x_3 = f_2 - \frac{a_{12}}{a_{11}}f_1, \\ & \left(a_{32} - \frac{a_{31}a_{12}}{a_{11}}\right)x_2 + \left(a_{33} - \frac{a_{13}a_{31}}{a_{11}}\right)x_3 = f_3 - \frac{a_{31}}{a_{11}}f_1. \end{aligned} \tag{16}$$

Отсюда видно, что матрица A_1 системы (16) получается из исход-

ной матрицы A путем умножения A слева на элементарную матрицу

$$L_1 = \begin{bmatrix} 1/a_{11} & 0 & 0 \\ -a_{21}/a_{11} & 1 & 0 \\ -a_{31}/a_{11} & 0 & 1 \end{bmatrix}, \tag{17}$$

так что $A_1 = L_1 A$. При этом систему (16) можно записать в виде

$$L_1 A x = L_1 f.$$

Матрицу (17) будем называть элементарной треугольной матрицей, соответствующей первому шагу исключения метода Гаусса. Перепишем систему (16) в виде

$$\begin{aligned} x_1 + c_{12}x_2 + c_{13}x_3 &= y_1, \\ a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 &= f_2^{(1)}, \\ a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 &= f_3^{(1)} \end{aligned} \tag{18}$$

и осуществим второй шаг метода Гаусса, т. е. исключим неизвестное x_2 из последнего уравнения. Тогда получим систему вида

$$\begin{aligned} x_1 + c_{12}x_2 + c_{13}x_3 &= y_1, \\ x_2 + c_{23}x_3 &= y_2, \\ a_{33}^{(2)}x_3 &= f_3^{(2)}. \end{aligned} \tag{19}$$

Нетрудно видеть, что переход от (18) к (19) осуществляется путем умножения системы (18) на элементарную треугольную матрицу

$$L_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/a_{22}^{(1)} & 0 \\ 0 & -a_{32}^{(1)}/a_{22}^{(1)} & 1 \end{bmatrix}. \tag{20}$$

Таким образом, после второго шага исключения мы приходим к системе

$$L_2 L_1 A x = L_2 L_1 f, \tag{21}$$

где матрицы L_1 и L_2 определены согласно (17), (20). Наконец, умножая (21) на матрицу

$$L_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1/a_{33}^{(2)} \end{bmatrix},$$

получаем систему

$$L_3 L_2 L_1 A x = L_3 L_2 L_1 f, \tag{22}$$

матрица которой $U = L_3 L_2 L_1 A$ является верхней треугольной матрицей с единичной главной диагональю. Отсюда следует, в частности, что $A = LU$, где $L = L_1^{-1} L_2^{-1} L_3^{-1}$ — нижняя треугольная матрица. Таким образом, LU-разложение матрицы A может быть получено с помощью элементарных треугольных матриц: сначала строятся матрицы L_1, L_2, L_3 и вычисляется $U = L_3 L_2 L_1 A$ и затем находится $L =$

$= L_1^{-1}L_2^{-1}L_3^{-1}$. Отметим, что матрицы L_k^{-1} имеют простой вид:

$$L_1^{-1} = \begin{bmatrix} a_{11} & 0 & 0 \\ a_{21} & 1 & 0 \\ a_{31} & 0 & 1 \end{bmatrix}, \quad L_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & a_{22}^{(1)} & 0 \\ 0 & a_{32}^{(1)} & 1 \end{bmatrix},$$

$$L_3^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & a_{33}^{(2)} \end{bmatrix}, \quad L = \begin{bmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22}^{(1)} & 0 \\ a_{31} & a_{32}^{(1)} & a_{33}^{(2)} \end{bmatrix},$$

причем на диагонали матрицы L расположены ведущие элементы метода исключения.

Запись метода Гаусса в виде (22) детально описывает процесс исключения.

Все сказанное выше переносится без изменения и на системы уравнений произвольного порядка (2). Процесс исключения можно записать формулой

$$L_m L_{m-1} \dots L_1 Ax = L_m L_{m-1} \dots L_1 f, \qquad (23)$$

где элементарная нижняя треугольная матрица L_k на k-м шаге исключения имеет вид

$$L_k = \begin{bmatrix} 1 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1/a_{kk}^{(k-1)} & 0 & \dots & 0 \\ 0 & \dots & -a_{k+1,k}^{(k-1)}/a_{kk}^{(k-1)} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & -a_{mk}^{(k-1)}/a_{kk}^{(k-1)} & 0 & \dots & 1 \end{bmatrix}.$$

Матрица L_k осуществляет исключение неизвестного x_k из уравнений с номерами $k+1$, $k+2$, ..., m. Матрицы L_k^{-1} существуют и имеют вид

$$L_k^{-1} = \begin{bmatrix} 1 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & a_{kk}^{(k-1)} & 0 & \dots & 0 \\ 0 & \dots & a_{k+1\,k}^{(k-1)} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & a_{mk}^{(k-1)} & 0 & \dots & 1 \end{bmatrix}.$$

§ 3. Метод Гаусса с выбором главного элемента

1. Основная идея метода. Может оказаться, что система

$$Ax = f \qquad (1)$$

имеет единственное решение, хотя какой-либо из угловых миноров матрицы A равен нулю. Кроме того, заранее обычно неизвестно, все ли угловые миноры матрицы A отличны от нуля. В этих случа-

ях обычный метод Гаусса может оказаться непригодным. Избежать указанных трудностей позволяет метод Гаусса с выбором главного элемента. Основная идея метода состоит в том, чтобы на очередном шаге исключать не следующее по номеру неизвестное, а то неизвестное, коэффициент при котором является наибольшим по модулю. Таким образом, в качестве ведущего элемента здесь выбирается *главный, т. е. наибольший по модулю элемент*. Тем самым, если $\det A \neq 0$, то в процессе вычислений не будет происходить деление на нуль.

Различные варианты метода Гаусса с выбором главного элемента проиллюстрируем на примере системы из двух уравнений

$$a_{11}x_1+a_{12}x_2=f_1, \quad a_{21}x_1+a_{22}x_2=f_2. \tag{2}$$

Предположим, что $|a_{12}|>|a_{11}|$. Тогда на первом шаге будем исключать переменное x_2. Такой прием эквивалентен тому, что система (2) переписывается в виде

$$a_{12}x_2+a_{11}x_1=f_1, \quad a_{22}x_2+a_{21}x_1=f_2 \tag{3}$$

и к (3) применяется первый шаг обычного метода Гаусса. Указанный способ исключения называется *методом Гаусса с выбором главного элемента по строке*. Он эквивалентен применению обычного метода Гаусса к системе, в которой на каждом шаге исключения проводится соответствующая перенумерация переменных.

Применяется также метод Гаусса с выбором главного элемента по столбцу. Предположим, что $|a_{21}|>|a_{11}|$. Перепишем систему (2) в виде

$$a_{21}x_1+a_{22}x_2=f_2, \quad a_{11}x_1+a_{12}x_2=f_1$$

и к новой системе применим на первом шаге обычный метод Гаусса. Таким образом, *метод Гаусса с выбором главного элемента по столбцу* эквивалентен применению обычного метода Гаусса к системе, в которой на каждом шаге исключения проводится соответствующая перенумерация уравнений.

Иногда применяется и метод Гаусса с выбором главного элемента по всей матрице, когда в качестве ведущего выбирается максимальный по модулю элемент среди всех элементов матрицы системы.

2. Матрицы перестановок. В предыдущем параграфе было показано, что обычный метод Гаусса можно записать в виде

$$L_mL_{m-1}\dots L_1Ax=L_mL_{m-1}\dots L_1f,$$

где L_k, $k=1, 2, \dots, m$,— элементарные нижние треугольные матрицы. Чтобы получить аналогичную запись метода Гаусса с выбором главного элемента, нам необходимо познакомиться с матрицами перестановок.

Определение 1. *Матрицей перестановок* P называется квадратная матрица, у которой в каждой строке и в каждом столбце только один элемент отличен от нуля и равен единице.

Определение 2. *Элементарной матрицей перестановок* P_{kl} называется матрица, полученная из единичной матрицы перестановкой k-й и l-й строк.

Например, элементарными матрицами перестановок третьего порядка являются матрицы

$$P_{12}=\begin{bmatrix}0&1&0\\1&0&0\\0&0&1\end{bmatrix},\quad P_{13}=\begin{bmatrix}0&0&1\\0&1&0\\1&0&0\end{bmatrix},\quad P_{23}=\begin{bmatrix}1&0&0\\0&0&1\\0&1&0\end{bmatrix}.$$

Отметим следующие свойства элементарных матриц перестановок, вытекающие непосредственно из их определения.

1°. Произведение двух (а следовательно, и любого числа) элементарных матриц перестановок является матрицей перестановок (не обязательно элементарной).

2°. Для любой квадратной матрицы A матрица $P_{kl}A$ отличается от A перестановкой k-й и l-й строк.

3°. Для любой квадратной матрицы A матрица AP_{kl} отличается от A перестановкой k-го и l-го столбцов.

3. Пример. Поясним применение элементарных матриц перестановок для описания метода Гаусса с выбором главного элемента по столбцу. Рассмотрим следующий пример системы третьего порядка:

$$\begin{aligned}x_1+\ x_2+\ x_3&=f_1,\\2x_1\qquad+\ x_3&=f_2,\\5x_2+3x_3&=f_3.\end{aligned}\tag{4}$$

Система имеет вид (1), где

$$A=\begin{bmatrix}1&1&1\\2&0&1\\0&5&3\end{bmatrix}.\tag{5}$$

Максимальный элемент первого столбца матрицы A находится во второй строке. Поэтому в системе (4) надо поменять местами первую и вторую строки и перейти к эквивалентной системе

$$\begin{aligned}2x_1\qquad+\ x_3&=f_2,\\x_1+\ x_2+\ x_3&=f_1,\\5x_2+3x_3&=f_3.\end{aligned}\tag{6}$$

Систему (6) можно записать в виде

$$P_{12}Ax=P_{12}f,\tag{7}$$

т. е. она получается из системы (4) путем умножения на матрицу перестановок

$$P_{12}=\begin{bmatrix}0&1&0\\1&0&0\\0&0&1\end{bmatrix}.$$

Далее, к системе (6) надо применить первый шаг обычного метода исключения Гаусса. Этот шаг, как мы видели, эквивалентен

умножению системы (7) на элементарную нижнюю треугольную матрицу (см. (17) из § 2)

$$L_1 = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ -\frac{1}{2} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

В результате от (7) перейдем к системе

$$L_1 P_{12} A x = L_1 P_{12} f \tag{8}$$

или, в развернутом виде,

$$\begin{aligned} x_1 \quad + \frac{1}{2} x_3 &= \frac{f_2}{2}, \\ x_2 + \frac{1}{2} x_3 &= f_1 - \frac{f_2}{2}, \\ 5x_2 + \; 3x_3 &= f_3. \end{aligned} \tag{9}$$

Из последних двух уравнений системы (9) надо теперь исключить переменное x_2. Поскольку максимальным элементом первого столбца укороченной системы

$$\begin{aligned} x_2 + \frac{1}{2} x_3 &= f_1 - \frac{f_2}{2}, \\ 5x_2 + 3x_3 &= f_3 \end{aligned} \tag{10}$$

является элемент второй строки, делаем в (10) перестановку строк и тем самым от системы (9) переходим к эквивалентной системе

$$\begin{aligned} x_1 + \quad \frac{1}{2} x_3 &= \frac{f_2}{2}, \\ 5x_2 + \; 3x_3 &= f_3, \\ x_2 + \frac{1}{2} x_3 &= f_1 - \frac{f_2}{2}, \end{aligned} \tag{11}$$

которую можно записать в матричном виде как

$$P_{23} L_1 P_{12} A x = P_{23} L_1 P_{12} f. \tag{12}$$

Таким образом, система (12) получена применением элементарной матрицы перестановок

$$P_{23} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

к системе (8).

Далее, к системе (11) надо применить второй шаг исключения обычного метода Гаусса. Это эквивалентно умножению системы (11) на элементарную треугольную матрицу

$$L_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/5 & 0 \\ 0 & -1/5 & 1 \end{bmatrix}.$$

В результате получим систему

$$L_2P_{23}L_1P_{12}Ax=L_2P_{23}L_1P_{12}f \tag{13}$$

или

$$\begin{aligned} x_1 \quad &+\frac{1}{2}x_3=\frac{f_2}{2}, \\ x_2+&\frac{3}{5}x_3=\frac{1}{5}f_3, \\ -&\frac{1}{10}x_3=f_1-\frac{f_2}{2}-\frac{1}{5}f_3. \end{aligned} \tag{14}$$

Заключительный шаг прямого хода метода Гаусса состоит в замене последнего уравнения системы (14) уравнением

$$x_3=-10\left(f_1-\frac{f_2}{2}-\frac{1}{5}f_3\right),$$

что эквивалентно умножению (13) на матрицу

$$L_3=\begin{bmatrix}1&0&0\\0&1&0\\0&0&-10\end{bmatrix}.$$

Таким образом, для рассмотренного примера процесс исключения Гаусса с выбором главного элемента по столбцу записывается в виде

$$L_3L_2P_{23}L_1P_{12}Ax=L_3L_2P_{23}L_1P_{12}f. \tag{15}$$

По построению матрица

$$U=L_3L_2P_{23}L_1P_{12}A \tag{16}$$

является верхней треугольной матрицей с единичной главной диагональю.

Отличие от обычного метода Гаусса состоит в том, что в качестве сомножителей в (16) наряду с элементарными треугольными матрицами L_k могут присутствовать элементарные матрицы перестановок P_{kl}.

Покажем еще, что из (16) следует разложение

$$PA=LU, \tag{17}$$

где L — нижняя треугольная матрица, имеющая обратную, и P — матрица перестановок. Для этого найдем матрицу

$$\tilde{L}_1=P_{23}L_1P_{23}. \tag{18}$$

По свойству 2° матрица $P_{23}L_1$ получается из матрицы L_1 перестановкой второй и третьей строк,

$$P_{23}L_1=\begin{bmatrix}\frac{1}{2}&0&0\\0&0&1\\-1/2&1&0\end{bmatrix}.$$

Матрица $\tilde{L}_1$ согласно свойству 3° получается из $P_{23}L_1$ перестановкой второго и третьего столбцов,

$$\tilde{L}_1=\begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \\ -1/2 & 0 & 1 \end{bmatrix},$$

т. е. $\tilde{L}_1$ — нижняя треугольная матрица, имеющая обратную.

Из (18), учитывая равенство $P_{23}^{-1}=P_{23}$, получим

$$\tilde{L}_1P_{23}=P_{23}L_1. \tag{19}$$

Отсюда и из (16) видим, что

$$U=L_3L_2\tilde{L}_1P_{23}P_{12}A=L^{-1}PA,$$

где обозначено $P=P_{23}P_{12}$, $L=\tilde{L}^{-1}L_2^{-1}L_3^{-1}$. Поскольку P — матрица перестановок и L — нижняя треугольная матрица, свойство (17) доказано. Оно означает, что метод Гаусса с выбором главного элемента по столбцу эквивалентен обычному методу Гаусса, примененному к матрице PA, т. е. к системе, полученной из исходной системы перестановкой некоторых уравнений.

4. Общий вывод. Результат, полученный здесь для очень частного примера, справедлив и в случае общей системы уравнений (1). А именно, метод Гаусса с выбором главного элемента по столбцу можно записать в виде

$$L_mL_{m-1}P_{m-1,j_{m-1}}L_{m-2}\ldots L_2P_{2,j_2}L_1P_{1,j_1}Ax=$$
$$=L_mL_{m-1}P_{m-1,j_{m-1}}L_{m-2}\ldots L_2P_{2,j_2}L_1P_{1,j_1}f, \tag{20}$$

где P_{k,j_k} — элементарные матрицы перестановок такие, что $k\leqslant j_k\leqslant m$ и L_k — элементарные треугольные матрицы.

Отсюда, используя соотношения перестановочности, аналогичные (19), можно показать, что метод Гаусса с выбором главного элемента эквивалентен обычному методу Гаусса, примененному к системе

$$PAx=Pf, \tag{21}$$

где P — некоторая матрица перестановок.

Теоретическое обоснование метода Гаусса с выбором главного элемента содержится в следующей теореме.

Теорема 1. *Если* $\det A\neq 0$, *то существует матрица перестановок* P *такая, что матрица* PA *имеет отличные от нуля угловые миноры.*

Доказательство теоремы 1 приведено в п. 5.

Следствие. *Если* $\det A\neq 0$, *то существует матрица перестановок* P *такая, что справедливо разложение*

$$PA=LU, \tag{22}$$

где L *— нижняя треугольная матрица с отличными от нуля диагональными элементами и* U *— верхняя треугольная матрица с единичной главной диагональю.*

Следует подчеркнуть, что в методе Гаусса с выбором главного элемента матрица P не задается заранее, а строится в процессе исключения, как это было показано в примере из п. 3. Как правило, не требуется знать эту матрицу в явном виде.

5. Доказательство теоремы 1. Докажем теорему 1 индукцией по числу m — порядку матрицы A. Пусть $m=2$, т. е.

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

Если $a_{11} \neq 0$, то утверждение теоремы 1 выполняется при $P=E$, где E — единичная матрица второго порядка. Если $a_{11}=0$, то $a_{21} \neq 0$, так как $\det A \neq 0$. При этом у матрицы

$$P_{12}A = \begin{bmatrix} a_{21} & a_{22} \\ a_{11} & a_{12} \end{bmatrix}$$

все угловые миноры отличны от нуля.

Пусть утверждение теоремы верно для любых квадратных матриц порядка $m-1$. Покажем, что оно верно и для матриц порядка m. Разобьем матрицу A порядка m на блоки

$$A = \begin{bmatrix} A_{m-1} & a_{m-1} \\ b_{m-1} & a_{mm} \end{bmatrix},$$

где

$$A_{m-1} = \begin{bmatrix} a_{11} & \dots & a_{1,m-1} \\ \dots & \dots & \dots \\ a_{m-1,1} & \dots & a_{m-1,m-1} \end{bmatrix}, \quad a_{m-1} = \begin{bmatrix} a_{1m} \\ a_{2m} \\ \dots \\ a_{m-1,m} \end{bmatrix},$$

$$b_{m-1} = (a_{m1}, a_{m2}, \dots, a_{m,m-1}).$$

Достаточно рассмотреть два случая: $\det A_{m-1} \neq 0$ и $\det A_{m-1}=0$. В первом случае по предположению индукции существует матрица перестановок P_{m-1} порядка $m-1$ такая, что $P_{m-1}A_{m-1}$ имеет отличные от нуля угловые миноры. Тогда для матрицы перестановок

$$P = \begin{bmatrix} P_{m-1} & 0 \\ 0 & 1 \end{bmatrix}$$

имеем

$$PA = \begin{bmatrix} P_{m-1}A_{m-1} & P_{m-1}a_{m-1} \\ b_{m-1} & a_{mm} \end{bmatrix},$$

причем $\det(PA) = \pm \det A \neq 0$. Тем самым все угловые миноры матрицы PA отличны от нуля.

Рассмотрим второй случай, когда $\det A_{m-1}=0$. Так как $\det A \neq 0$, найдется хотя бы один отличный от нуля минор порядка $m-1$ матрицы A, полученный вычеркиванием последнего столбца и какой-либо строки. Пусть, например:

$$\begin{vmatrix} a_{11} & \dots & a_{1,m-1} \\ \dots & \dots & \dots \\ a_{l-1,1} & \dots & a_{l-1,m-1} \\ a_{l+1,1} & \dots & a_{l+1,m-1} \\ \dots & \dots & \dots \\ a_{m1} & \dots & a_{m,m-1} \end{vmatrix} \neq 0, \tag{23}$$

где $l \neq m$. Переставляя в матрице A строки с номерами l и m, получим матрицу $P_{lm}A$, у которой угловой минор порядка m—1 имеет вид

$$\begin{vmatrix} a_{11} & \dots & a_{1,m-1} \\ \dots & \dots & \dots \\ a_{l-1,1} & \dots & a_{l-1,m-1} \\ a_{m1} & \dots & a_{m,m-1} \\ a_{l+1,1} & \dots & a_{l+1,m-1} \\ \dots & \dots & \dots \\ a_{m-1,1} & \dots & a_{m-1,m-1} \end{vmatrix}$$

и отличается от (23) только перестановкой строк. Следовательно, этот минор не равен нулю и мы приходим к рассмотренному выше случаю.

6. Вычисление определителя. В большинстве существующих стандартных программ одновременно с решением системы линейных алгебраических уравнений (1) вычисляется определитель матрицы A. Пусть в процессе исключения найдено разложение (22), т. е. построены матрицы L и U. Тогда

$$\det(PA) = \det L \det U = \det L = l_{11} l_{22} \dots l_{mm},$$

т. е. произведение диагональных элементов матрицы L равно определителю матрицы PA. Поскольку матрицы PA и A отличаются только перестановкой строк, определитель матрицы PA может отличаться от определителя матрицы A только знаком. А именно, $\det(PA) = \det A$, если число перестановок четно, и $\det(PA) = -\det A$, если число перестановок нечетно. Таким образом, для вычисления определителя необходимо знать, сколько перестановок было осуществлено в процессе исключения.

Если матрица A вырождена, то при использовании метода Гаусса с выбором главного элемента по столбцу на некотором шаге исключения k все элементы k-го столбца, находящиеся ниже главной диагонали и на ней, окажутся равными нулю.

Действительно, рассмотрим укороченную систему (см. (11) из § 1), которая получается на k-м шаге исключения:

$$\begin{aligned} a_{kk}^{(k-1)} x_k + \dots + a_{km}^{(k-1)} x_m &= f_k^{(k-1)}, \\ a_{k+1,k}^{(k-1)} x_k + \dots + a_{k+1,m}^{(k-1)} x_m &= f_{k+1}^{(k-1)}, \\ \dots \dots \dots \dots \dots \dots \dots & \\ a_{mk}^{(k-1)} x_k + \dots + a_{mm}^{(k-1)} x_m &= f_m^{(k-1)}. \end{aligned} \qquad (24)$$

При решении системы (24) могут возникнуть два случая: 1) хотя бы один из коэффициентов $a_{kk}^{(k-1)}, a_{k+1,k}^{(k)}, \dots, a_{mk}^{(k-1)}$ отличен от нуля; 2) $a_{kk}^{(k-1)} = a_{k+1,k}^{(k-1)} = \dots = a_{mk}^{(k-1)} = 0$. Если для всех $k = 1, 2, \dots, m$ реализуется первый случай, то систему (1) можно решить методом Гаусса с выбором главного элемента по столбцу, и, следовательно, $\det A \neq 0$. Если же $\det A = 0$, то при некотором k реализуется второй случай. При этом дальнейшее исключение становится невозможным и программа должна выдать информацию о том, что определитель матрицы равен нулю.

§ 4. Обращение матрицы

Нахождение матрицы, обратной матрице A, эквивалентно решению матричного уравнения

$$AX=E, \tag{1}$$

где E — единичная матрица и X — искомая квадратная матрица. Пусть $A=[a_{ij}]$, $X=[x_{ij}]$. Уравнение (1) можно записать в виде системы m^2 уравнений

$$\sum_{k=1}^{m} a_{ik}x_{kj}=\delta_{ij}, \quad i, j=1, 2, \ldots, m, \tag{2}$$

где $\delta_{ij}=1$ при $i=j$ и $\delta_{ij}=0$ при $i\neq j$.

Для дальнейшего важно заметить, что система (2) распадается на m независимых систем уравнений с одной и той же матрицей A, но с различными правыми частями. Эти системы имеют вид

$$Ax^{(j)}=\delta^{(j)}, \quad j=1, 2, \ldots, m, \tag{3}$$

где $x^{(j)}=(x_{1j}, x_{2j}, \ldots, x_{mj})^T$, у вектора $\delta^{(j)}$ равна единице j-я компонента и равны нулю остальные компоненты.

Например, для матрицы второго порядка система (2) распадается на две независимые системы

$$a_{11}x_{11}+a_{12}x_{21}=1, \quad a_{11}x_{12}+a_{12}x_{22}=0,$$

и

$$a_{21}x_{11}+a_{22}x_{21}=0 \quad a_{21}x_{12}+a_{22}x_{22}=1.$$

Для решения систем (3) используется метод Гаусса (обычный или с выбором главного элемента). Рассмотрим применение метода Гаусса без выбора главного элемента. Поскольку все системы (3) имеют одну и ту же матрицу A, достаточно один раз совершить прямой ход метода Гаусса, т. е. получить разложение $A=LU$ и запомнить матрицы L и U. Для этого, как мы знаем (см. § 1), требуется сделать $m(m^2-1)/3$ действий умножения и деления.

Обратный ход осуществляется путем решения систем уравнений

$$Ly^{(j)}=\delta^{(j)}, \quad y^{(j)}=(y_{1j}, y_{2j}, \ldots, y_{mj})^T, \tag{4}$$

$$Ux^{(j)}=y^{(j)}, \quad j=1, 2, \ldots, m, \tag{5}$$

с треугольными матрицами L и U. Решение системы (5) при каждом j требует $0{,}5\,m(m-1)$ действий. Для решения системы (4) надо еще добавить m делений на диагональные элементы матрицы L, так что здесь потребуется $0{,}5\,m(m+1)$ умножений и делений. Всего при каждом j на обратный ход затрачивается $0{,}5(m-1)m+0{,}5(m+1)m=m^2$ действий, а для всех j требуется m^3 действий.

Можно сократить число действий, принимая во внимание специальный вид правых частей системы (4). Запишем подробнее пер-

вые $j-1$ уравнений системы (4):

$$l_{11}y_{1j}=0,$$
$$l_{21}y_{1j}+l_{22}y_{2j}=0,$$
$$\dots\dots\dots\dots$$
$$l_{j-1,1}y_{1j}+l_{j-1,2}y_{2j}+\dots+l_{j-1,j-1}y_{j-1,j}=0.$$

Учитывая невырожденность матрицы L, отсюда получим

$$y_{1j}=y_{2j}=\dots=y_{j-1,j}=0.$$

При этом оставшиеся уравнения системы (4) имеют вид

$$l_{jj}y_{jj}=1,$$
$$l_{ij}y_{jj}+l_{i,j+1}y_{j+1,j}+\dots+l_{ii}y_{ij}=0,$$
$$i=j+1,\ j+2,\ \dots,\ m.$$

Отсюда последовательно находятся неизвестные y_{ij} по формулам

$$y_{ij}=-\frac{\sum\limits_{k=j}^{i-1}l_{ik}y_{kj}}{l_{ii}},\qquad i=j+1,j+2,\dots,m, \tag{6}$$

$$y_{jj}=1/l_{jj}. \tag{7}$$

Подсчитаем число умножений и делений, необходимое для проведения вычислений по формулам (6). При фиксированном i для вычислений по формуле (6) требуется 1 деление и $i-j$ умножений. Вычисления по формулам (6), (7) при фиксированном j потребуют

$$1+\sum_{i=j+1}^{m}(i-j+1)=\frac{(m-j+2)(m-j+1)}{2}$$

действий. Наконец, решение указанным способом систем (4) при всех $j=1,\ 2,\ \dots,\ m$ потребует

$$\frac{1}{2}\sum_{j=1}^{m}(m-j+2)(m-j+1)=\frac{1}{2}\sum_{k=1}^{m}k(k+1)=\frac{m(m+1)(m+2)}{6}$$

действий. Общее число действий умножения и деления, необходимое для обращения матрицы указанным способом,

$$\frac{m(m^2-1)}{3}+\frac{m^2(m-1)}{2}+\frac{m(m+1)(m+2)}{6}=m^3.$$

Тем самым обращение матрицы требует не намного больше времени, чем решение системы уравнений.

§ 5. Метод квадратного корня

1. Факторизация эрмитовой матрицы. Метод предназначен для решения систем уравнений

$$Ax=f \tag{1}$$

с симметричной (в комплексном случае — эрмитовой) матрицей. Он основан на разложении матрицы A в произведение

$$A=S^*DS, \tag{2}$$

где S — верхняя треугольная матрица с положительными элементами на главной диагонали, S^* — транспонированная к ней (или комплексно сопряженная) матрица, D — диагональная матрица, на диагонали которой находятся числа, равные ± 1.

Возможность представления (2) можно получить как следствие теоремы об LU-разложении (см. § 2). Пусть все угловые миноры матрицы A отличны от нуля. Тогда справедливо разложение $A=LU$, где L — нижняя треугольная матрица, имеющая обратную, и U — верхняя треугольная с единичной диагональю.

Представим матрицу L в виде произведения $L=MK$, где M — нижняя треугольная матрица с единичной главной диагональю и K — диагональная матрица, главная диагональ которой совпадает с главной диагональю матрицы L, т. е.

$$K=\operatorname{diag}[l_{11}, l_{22}, \ldots, l_{mm}]. \tag{3}$$

По условию диагональные элементы матрицы L отличны от нуля, и, следовательно, разложение $L=MK$ существует. Тогда

$$A=MKU, \tag{4}$$

где M и U — треугольные матрицы с единичной главной диагональю и K — диагональная матрица, имеющая обратную. Из условия $A^*=A$ получаем $U^*K^*M^*=MKU$ и

$$K^{-1}M^{-1}U^*K^*=UM^{*-1}. \tag{5}$$

Матрица, находящаяся в левой части равенства (5), является нижней треугольной, а в правой части — верхней треугольной. Поэтому из равенства (5) следует, что обе матрицы UM^{*-1} и $K^{-1}M^{-1}U^*K^*$ являются диагональными.

Далее, поскольку матрица UM^{*-1} имеет единичную главную диагональ, она является единичной матрицей, $UM^{*-1}=E$, т. е. $U=M^*$. Отсюда и из (4) получаем разложение

$$A=MKM^*. \tag{6}$$

Представим матрицу K, определенную согласно (3) в виде произведения трех диагональных матриц:

$$K=|K|^{1/2}D|K|^{1/2},$$

где обозначено

$$|K|^{1/2}=\operatorname{diag}[\sqrt{|l_{11}|}, \sqrt{|l_{22}|}, \ldots, \sqrt{|l_{mm}|}],$$
$$D=\operatorname{diag}[\operatorname{sign} l_{11}, \operatorname{sign} l_{22}, \ldots, \operatorname{sign} l_{mm}].$$

Тогда из (6) получим разложение (2), где $S=|K|^{1/2}M^*$ — верхняя треугольная матрица с положительными элементами на главной диагонали.

2. **Пример.** Если разложение (2) получено, то решение системы (1) сводится к последовательному решению двух систем уравнений с треугольными матрицами

$$S^*Dy=f, \tag{7}$$
$$Sx=y. \tag{8}$$

Покажем на примере матриц второго порядка как можно получить разложение (2). Пусть A — действительная симметричная матрица

$$A=\begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix}.$$

Будем искать S и D в виде

$$S=\begin{bmatrix} s_{11} & s_{12} \\ 0 & s_{22} \end{bmatrix}, \quad D=\begin{bmatrix} d_{11} & 0 \\ 0 & d_{22} \end{bmatrix},$$

где каждое из чисел d_{11}, d_{22} может быть либо $+1$, либо -1. Тогда

$$S^*DS=\begin{bmatrix} s_{11}^2 d_{11} & s_{11}s_{12}d_{11} \\ s_{11}s_{12}d_{11} & s_{12}^2 d_{11}+s_{22}^2 d_{22} \end{bmatrix}$$

и из условия (2) получаем три уравнения:

$$s_{11}^2 d_{11}=a_{11}, \quad s_{11}s_{12}d_{11}=a_{12}, \quad s_{12}^2 d_{11}+s_{22}^2 d_{22}=a_{22}.$$

Из первого уравнения находим $d_{11}=\operatorname{sign} a_{11}$, $s_{11}=\sqrt{|a_{11}|}$. Далее, если $a_{11}\neq 0$, то

$$s_{12}=a_{12}/(s_{11}d_{11}),$$

и, наконец,

$$s_{22}^2 d_{22}=a_{22}-s_{12}^2 d_{11},$$

т. е.

$$d_{22}=\operatorname{sign}(a_{22}-s_{12}^2 d_{11}), \quad s_{22}=\sqrt{|a_{22}-s_{12}^2 d_{11}|}.$$

3. Общие расчетные формулы. Получим разложение (2) в случае эрмитовой матрицы A произвольного порядка m. Если $S=[s_{ij}]$ и $D=\operatorname{diag}[d_{11}, \ldots, d_{mm}]$, то элемент матрицы DS, имеющий индексы (i, j), равен

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

Кроме того, $S^*=[\bar{s}_{ji}]$, поэтому

$$(S^*DS)_{ij}=\sum_{l=1}^{m} \bar{s}_{li}d_{ll}s_{lj},$$

где $\bar{s}_{li}$ — число, комплексно сопряженное s_{li}. Из условия (2) получаем уравнения

$$\sum_{l=1}^{m} \bar{s}_{li}d_{ll}s_{lj}=a_{ij}, \qquad i, j=1, 2, \ldots, m. \tag{9}$$

Так как матрица A эрмитова, можно, не ограничивая общности, считать, что в системе (9) выполняется неравенство $i\leqslant j$. Перепишем (9) в виде

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

и заметим, что $\bar{s}_{li}=0$ для $l>i$. Таким образом, получим систему уравнений

$$s_{ii}s_{ij}d_{ii}+\sum_{l=1}^{i-1} \bar{s}_{li}s_{lj}d_{ll}=a_{ij}, \qquad i\leqslant j. \tag{10}$$

В частности, при $i=j$ получаем

$$|s_{ii}|^2 d_{ii} = a_{ii} - \sum_{l=1}^{i-1} |s_{li}|^2 d_{ll},$$

т. е.

$$d_{ii} = \operatorname{sign}\left(a_{ii} - \sum_{l=1}^{i-1} |s_{li}|^2 d_{ll}\right), \tag{11}$$

$$s_{ii} = \left(\left|a_{ii} - \sum_{l=1}^{i-1} |s_{li}|^2 d_{ll}\right|\right)^{1/2}. \tag{12}$$

Далее, при $i<j$ из (10) получим

$$s_{ij} = \frac{a_{ij} - \sum_{l=1}^{i-1} \bar{s}_{li} s_{lj} d_{ll}}{s_{ii} d_{ii}}. \tag{13}$$

По формулам (11)—(13) находятся рекуррентно все ненулевые элементы матриц D и S.

4. Подсчет числа действий. Метод квадратного корня применяется обычно к системам с положительно определенной эрмитовой матрицей A. В этом случае из (6) следует положительность диагональных элементов матрицы K, и тем самым $D=E$ в разложении (2). Если предположить дополнительно, что A — действительная матрица, то из (11)—(13) получим следующие расчетные формулы:

$$s_{11} = \sqrt{a_{11}}, \quad s_{ii} = \sqrt{a_{ii} - \sum_{l=1}^{i-1} s_{li}^2}, \quad i = 2, 3, \ldots, m, \tag{14}$$

$$s_{1j} = a_{1j}/s_{11}, \quad j = 2, 3, \ldots, m, \tag{15}$$

$$s_{ij} = \frac{a_{ij} - \sum_{l=1}^{i-1} s_{li} s_{lj}}{s_{ii}}, \quad j = 2, 3, \ldots, m, \quad i = 2, 3, \ldots, j-1. \tag{16}$$

Подсчитаем сначала число умножений. Вычисления по формуле (14) требуют

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

умножений. Вычисления по формуле (16) при каждом фиксированном j требуют

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

умножений, а всего здесь требуется

$$\sum_{j=2}^{m} \frac{(j-2)(j-1)}{2} = \frac{1}{2}\sum_{k=1}^{m-1} k(k-1) = \frac{m(m-1)(m-2)}{6}$$

умножений. Следовательно, общее число умножений

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

Число делений, необходимых для вычислений по формулам (14)—(16), совпадает с числом наддиагональных элементов матрицы S и равно $m(m-1)/2$.

Таким образом, общее число действий умножения и деления, необходимое для факторизации $A=S^*S$,

$$\frac{m(m^2-1)}{6} + \frac{m(m-1)}{2} = \frac{m(m-1)(m+4)}{6} \approx \frac{m^3}{6}.$$

При больших m это число примерно в два раза меньше числа умножений и делений в прямом ходе метода Гаусса. Такое сокращение числа действий объясняется тем, что A — симметричная матрица. Заметим, что данный метод требует m операций извлечения корня.

Если матрица A факторизована в виде $A=S^*S$, то обратный ход метода квадратного корня состоит в последовательном решении двух систем уравнений

$$S^*y=f, \quad Sx=y.$$

Решения этих систем находятся по рекуррентным формулам

$$y_i = \frac{f_i - \sum_{j=1}^{i-1} s_{ji}y_j}{s_{ii}}, \quad i=2,3,\ldots,m, \tag{17}$$

$$y_1 = f_1/s_{11},$$

$$x_i = \frac{y_i - \sum_{j=i+1}^{m} s_{ij}x_j}{s_{ii}}, \quad i=m-1, m-2, \ldots, 1, \tag{18}$$

$$x_m = y_m/s_{mm}.$$

Вычисления по каждой из формул (17), (18) требуют m делений и $0{,}5\,m(m-1)$ умножений. Следовательно, всего для обратного хода требуется $m(m+1)$ операций умножения и деления. Всего метод квадратного корня при факторизации $A=S^*S$ требует

$$m(m+1) + \frac{m(m-1)(m+4)}{6} = \frac{m(m^2+9m+2)}{6}$$

операций умножения и деления и m операций извлечения квадратного корня.

§ 6. Обусловленность систем линейных алгебраических уравнений

1. Устойчивость системы линейных алгебраических уравнений. При использовании численных методов для решения тех или иных математических задач необходимо различать свойства самой задачи и свойства вычислительного алгоритма, предназначенного для ее решения. Для каждой математической задачи принято рассматривать вопрос о ее корректности. Говорят, что задача поставлена *корректно*, если ее решение существует и единственно и если оно непрерывно зависит от входных данных. Последнее свойство называется также *устойчивостью* относительно входных данных. Сформулированное здесь общее и не очень четкое определение корректности должно уточняться при переходе к изучению конкретных классов математических задач. Так, хорошо известны определения и методы исследования корректности задачи Коши для систем обыкновенных дифференциальных уравнений, корректная постановка типичных задач математической физики.

Корректность исходной математической задачи еще не гарантирует хороших свойств численного метода ее решения. Поэтому для корректно поставленных задач свойства численных методов должны изучаться особо.

Отметим, что часто возникает необходимость численного решения некорректно поставленных задач. Этот круг вопросов подробно освещен в книге [38].

В настоящем параграфе вопросы корректности исходной задачи и численных алгоритмов ее решения рассматриваются на примере системы линейных алгебраических уравнений

$$Ax=f \tag{1}$$

с квадратной матрицей A порядка m. Хорошо известно, что для каждого m-мерного вектора f решение x задачи (1) существует тогда и только тогда, когда $\det A\neq 0$. В этом случае можно определить матрицу A^{-1}, обратную матрице A, и записать решение в виде

$$x=A^{-1}f. \tag{2}$$

Для того чтобы убедиться в корректности задачи (1), необходимо еще установить непрерывную зависимость решения от входных данных. В связи с этим возникает по крайней мере два вопроса. Первый: что считать входными данными задачи (1), и второй: в каком смысле следует понимать непрерывную зависимость? Ответ на первый вопрос очевиден: входными данными являются правая часть f и элементы a_{ij}, $i, j=1, 2, \ldots, m$, матрицы A. Соответственно различают *устойчивость по правой части* (когда возмущается только правая часть f, а матрица A остается неизменной) и *коэффициентную устойчивость* (когда возмущается только матрица A, а правая часть f остается неизменной).

Чтобы можно было говорить о непрерывной зависимости, необходимо ввести на множестве m-мерных векторов ту или иную мет-

рику. Будем считать, что решение и правая часть задачи (1) принадлежат линейному пространству H, состоящему из m-мерных векторов (вещественных или комплексных — безразлично). Введем в H норму $\|\cdot\|$; конкретный вид этой нормы сейчас не имеет значения, важно лишь, чтобы выполнялись все аксиомы нормы:

$\|x\|>0$ для любого $0\neq x\in H$, $\|0\|=0$;

$\|\alpha x\|=|\alpha|\,\|x\|$ для любого числа α и любого $x\in H$;

$\|x+y\|\leqslant\|x\|+\|y\|$ для любых x, $y\in H$.

Нормой матрицы A, подчиненной данной норме вектора, называется число

$$\|A\|=\sup_{0\neq x\in H}\frac{\|Ax\|}{\|x\|}. \tag{3}$$

Из определений следует, что $\|Ax\|\leqslant\|A\|\|x\|$ для всех $x\in H$, $\|A+B\|\leqslant\|A\|+\|B\|$, $\|AB\|\leqslant\|A\|\|B\|$ для любых матриц A, B; $\|E\|=1$, где E — единичная матрица. Наряду с основной системой уравнений (1) рассмотрим «возмущенную систему»

$$A\tilde{x}=\tilde{f}, \tag{4}$$

которая отличается от (1) правой частью. Будем предполагать пока, что в матрицу A возмущений не вносится. Нас интересует, насколько сильно может измениться решение x в результате изменения правой части. Обозначим

$$\delta x=\tilde{x}-x,\quad \delta f=\tilde{f}-f.$$

Говорят, что *система* (1) *устойчива по правой части*, если при любых f, $\tilde{f}$ справедлива оценка

$$\|\delta x\|\leqslant M_1\|\delta f\|, \tag{5}$$

где $M_1>0$ — постоянная, не зависящая от правых частей f, $\tilde{f}$. Оценка (5) выражает факт непрерывной зависимости решения от правой части, т. е. показывает, что $\|\delta x\|\to 0$ при $\|\delta f\|\to 0$. Наличие устойчивости очень важно при численном решении систем уравнений, поскольку почти никогда нельзя задать правую часть f точно, на самом деле вместо вектора f задается какой-то близкий ему вектор $\tilde{f}$. Погрешность $\delta f=\tilde{f}-f$ возникает, например, в результате погрешностей округления.

Легко показать, что если $\det A\neq 0$, то система (1) устойчива по правой части. Действительно, из (1) и (4) следует уравнение для погрешности

$$A(\delta x)=\delta f,$$

откуда получаем

$$\delta x=A^{-1}(\delta f),$$

$$\|\delta x\|\leqslant\|A^{-1}\|\|\delta f\|, \tag{6}$$

т. е. выполняется неравенство (5) с константой $M_1=\|A^{-1}\|$. Заметим, что чем ближе к нулю определитель матрицы A, тем больше постоянная M_1 и, следовательно, тем сильнее погрешность правой части может исказить искомое решение.

2. Число обусловленности. В оценку (5) входят *абсолютные погрешности* решения $\delta x=\tilde{x}-x$ и правой части $\delta f=\tilde{f}-f$. При решении системы (1) на ЭВМ с плавающей запятой более естественными характеристиками являются *относительные погрешности*

$$\frac{\|\delta f\|}{\|f\|}, \quad \frac{\|\delta x\|}{\|x\|}.$$

Получим оценку, выражающую относительную погрешность решения через относительную погрешность правой части. Для этого используем неравенство

$$\|f\| \leqslant \|A\| \|x\|, \tag{7}$$

которое следует из (1). Перемножив (6) и (7), придем к требуемой оценке

$$\frac{\|\delta x\|}{\|x\|} \leqslant M_A \frac{\|\delta f\|}{\|f\|}, \tag{8}$$

где

$$M_A = \|A^{-1}\| \|A\|. \tag{9}$$

Число M_A, входящее в эту оценку, называется *числом обусловленности матрицы* A и характеризует степень зависимости относительной погрешности решения от относительной погрешности правой части. Матрицы с большим числом обусловленности M_A называются *плохо обусловленными матрицами.* При численном решении систем с плохо обусловленными матрицами возможно сильное накопление погрешностей.

Отметим следующие свойства числа обусловленности:

1°. $M_A \geqslant 1$.

2°. $M_A \geqslant |\lambda_{\max}(A)| / |\lambda_{\min}(A)|$,

где $\lambda_{\max}(A)$ и $\lambda_{\min}(A)$ — соответственно наибольшее и наименьшее по модулю собственные числа матрицы A.

3°. $M_{AB} \leqslant M_A M_B$.

Докажем свойство 2°. Число $\rho(A) = |\lambda_{\max}(A)|$ называется *спектральным радиусом матрицы* A. Покажем сначала, что для любой нормы вектора подчиненная ей норма матрицы удовлетворяет неравенству

$$\rho(A) \leqslant \|A\|. \tag{10}$$

Рассмотрим собственный вектор y матрицы A, отвечающий наибольшему по модулю собственному значению. Справедливо равенство

$$Ay = \lambda_{\max}(A) y,$$

из которого следует, что

$$\|Ay\| = |\lambda_{\max}(A)| \|y\|.$$

С другой стороны, $\|Ay\| \leqslant \|A\| \|y\|$, и, следовательно, $|\lambda_{\max}(A)| \leqslant \|A\|$, т. е. получаем (10).

Поскольку $\lambda_{\min}^{-1}(A)$ является максимальным по модулю собственным значением матрицы A^{-1}, для него выполняется неравенство

$$|\lambda_{\min}(A)|^{-1} \leqslant \|A^{-1}\|.$$

Отсюда и из (10) следует свойство 2°. Заметим, что правая часть неравенства 2° не зависит от выбора нормы.

Существуют нормы и матрицы, для которых 2° выполняется со знаком равенства. Пусть H — вещественное пространство со скалярным произведением

$$(y, v) = \sum_{i=1}^{m} y_i v_i$$

и нормой $\|y\| = \sqrt{(y, y)}$. Тогда норма симметричной матрицы A совпадает с ее спектральным радиусом:

$$\|A\| = \rho(A). \tag{11}$$

Действительно, пусть $\{\mu_k\}_{k=1}^m$ — ортонормированный базис собственных векторов матрицы A и $x \in H$ — любой вектор. Разлагая x по базису $\{\mu_k\}$, получим

$$x = \sum_{k=1}^{m} c_k \mu_k, \quad \|x\|^2 = \sum_{k=1}^{m} c_k^2$$

и, далее,

$$Ax = \sum_{k=1}^{m} c_k \lambda_k \mu_k, \quad \|Ax\|^2 = \sum_{k=1}^{m} c_k^2 \lambda_k^2.$$

Отсюда имеем

$$\|Ax\|^2 \leqslant \rho^2(A) \|x\|^2,$$

т. е. $\|A\| \leqslant \rho(A)$. Сопоставляя это неравенство с (10), получаем (11).

Точно так же $\|A^{-1}\| = \rho(A^{-1}) = |\lambda_{\min}(A)|^{-1}$, и, следовательно,

$$M_A = |\lambda_{\max}(A)| / |\lambda_{\min}(A)| \tag{12}$$

для симметричной матрицы A и среднеквадратичной нормы вектора, $\|x\| = \sqrt{(x, x)}$.

Свойство 1° следует непосредственно из 2°, а свойство 3° — из аналогичного свойства матричных норм, $\|AB\| \leqslant \|A\| \|B\|$.

3. Полная оценка относительной погрешности. Предположим теперь, что в системе (1) возмущены как правая часть, так и коэффициенты. Рассмотрим наряду с (1) возмущенную систему

$$\tilde{A}\tilde{x} = \tilde{f} \tag{13}$$

и обозначим $\delta A = \tilde{A} - A$, $\delta x = \tilde{x} - x$, $\delta f = \tilde{f} - f$.

Справедлива следующая теорема (см. [6]).

Теорема 1. *Пусть матрица A имеет обратную и пусть выполнено условие*

$$\|\delta A\| < \|A^{-1}\|^{-1}. \tag{14}$$

Тогда матрица $\tilde{A} = A + \delta A$ имеет обратную и справедлива следующая оценка относительной погрешности:

$$\frac{\|\delta x\|}{\|x\|} \leqslant \frac{M_A}{1 - M_A \left(\dfrac{\|\delta A\|}{\|A\|}\right)} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta f\|}{\|f\|}\right). \tag{15}$$

При доказательстве будет использована

Лемма 1. *Пусть C — квадратная матрица, удовлетворяющая условию $\|C\|<1$ и E — единичная матрица. Тогда существует матрица $(E+C)^{-1}$, причем*

$$\|(E+C)^{-1}\| \leqslant \frac{1}{1-\|C\|}. \tag{16}$$

Доказательство. Для любого вектора x имеем

$$\|(E+C)x\| = \|x+Cx\| \geqslant \|x\| - \|Cx\| \geqslant \|x\| - \|C\|\|x\| = (1-\|C\|)\|x\| = \delta\|x\|,$$

где $\delta = 1-\|C\| > 0$. Поэтому однородное уравнение $(E+C)x=0$ имеет только нулевое решение: если $(E+C)x \neq 0$, то $0 \geqslant \delta\|x\|$ и $x=0$. Поэтому в неравенстве

$$\|(E+C)x\| \geqslant \delta\|x\| \tag{17}$$

можем обозначить $(E+C)x=y$, $x=(E+C)^{-1}y$ и переписать (17) в виде $\|y\| \geqslant \delta\|(E+C)^{-1}y\|$. Отсюда получим

$$\|(E+C)^{-1}y\| \leqslant \frac{1}{\delta}\|y\| = \frac{1}{1-\|C\|}\|y\|,$$

следовательно, выполнено (16).

Доказательство теоремы 1. Докажем, что существует матрица, обратная матрице $A+\delta A$. Имеем

$$\tilde{A} = A+\delta A = A(E+A^{-1}\delta A) = A(E+C),$$

где $C = A^{-1}\delta A$. По условию (14) имеем

$$\|C\| = \|A^{-1}\delta A\| \leqslant \|A^{-1}\|\|\delta A\| < 1, \tag{18}$$

поэтому согласно лемме 1 существует $(E+C)^{-1}$, а следовательно, и $\tilde{A}^{-1}$.

Нам осталось получить оценку (15). Представим $\delta x = \tilde{x} - x$ в виде $\delta x = \tilde{A}^{-1}\tilde{f} - A^{-1}f = \tilde{A}^{-1}(\tilde{f}-f) + (\tilde{A}^{-1}-A^{-1})f$. Согласно (1) имеем $f = Ax$, поэтому

$$\delta x = (\tilde{A}^{-1}A - E)x + \tilde{A}^{-1}\delta f. \tag{19}$$

Дальнейшие выкладки связаны с оценками норм матриц $\tilde{A}^{-1}A-E$ и $\tilde{A}^{-1}$. Обозначим

$$\Delta = \tilde{A}^{-1}A - E. \tag{20}$$

Имеем

$$\Delta = \tilde{A}^{-1}A - E = (A+\delta A)^{-1}A - E = [A(E+A^{-1}\delta A)]^{-1}A - E =$$

$$= (E+C)^{-1} - E, \quad C = A^{-1}\delta A,$$

$$\Delta = (E+C)^{-1}(E-(E+C)) = -(E+C)^{-1}C.$$

По лемме 1 получим

$$\|\Delta\| \leqslant \frac{\|C\|}{1-\|C\|} \leqslant \frac{\|A^{-1}\|\|\delta A\|}{1-\|A^{-1}\|\|\delta A\|}, \tag{21}$$

так как $\|C\| \leqslant \|A^{-1}\|\|\delta A\|$. Далее,

$$\tilde{A}^{-1} = (A+\delta A)^{-1} = [A(E+A^{-1}\delta A)]^{-1} = (E+C)^{-1}A^{-1}$$

и

$$\|\tilde{A}^{-1}\| \leqslant \frac{\|A^{-1}\|}{1-\|A^{-1}\|\|\delta A\|}.$$

Возвращаясь к (19), получим

$$\|\delta x\| \leqslant \frac{\|A^{-1}\|\,\|\delta A\|\,\|x\|}{1-\|A^{-1}\|\,\|\delta A\|} + \frac{\|A^{-1}\|\,\|\delta f\|}{1-\|A^{-1}\|\,\|\delta A\|} =$$

$$= \frac{\|A^{-1}\|}{1-\|A^{-1}\|\,\|\delta A\|}\,(\|x\|\,\|\delta A\| + \|\delta f\|).$$

Сделаем преобразование

$$\|\delta f\| = \frac{\|\delta f\|}{\|f\|}\|f\| = \frac{\|\delta f\|}{\|f\|}\|Ax\| \leqslant \frac{\|\delta f\|}{\|f\|}\|A\|\,\|x\|.$$

Тогда получим неравенство

$$\|\delta x\| \leqslant \frac{\|A^{-1}\|\,\|x\|}{1-\|A^{-1}\|\,\|\delta A\|}\left(\frac{\|\delta A\|}{\|A\|}\|A\| + \frac{\|\delta f\|}{\|f\|}\|A\|\right),$$

которое совпадает с (15). Теорема 1 доказана.

4. Влияние погрешностей округления при решении систем линейных алгебраических уравнений методом Гаусса. Метод Гаусса относится к прямым методам решения, поэтому он должен был бы давать точное решение задачи (1) за конечное число действий. Однако при задании на ЭВМ исходной информации (матрицы A, правой части f) неизбежно вносятся погрешности округления. Поэтому при проведении вычислений на ЭВМ точное решение почти никогда не достигается. Результирующая погрешность вычислений тем больше, чем выше порядок матрицы. Кроме того, точность вычислений зависит и от вида матрицы. Например, велика погрешность при решении систем уравнений с определителем, близким к нулю.

Не надо думать, однако, что накопление погрешностей округления делает непригодным метод Гаусса или другие численные методы. Ведь обычно требуется знать решение не абсолютно точно, а лишь с какой-то степенью точности. Важно, чтобы результирующая погрешность вычислений находилась в пределах заданной точности. Для этого необходимо проводить анализ влияния погрешностей округления на точность алгоритма.

Поскольку полное проведение такого анализа весьма трудоемко и выходит за рамки данной книги, ограничимся лишь основными сведениями. Подробное изложение этого круга вопросов можно найти в книге [6].

Для большинства вычислительных алгоритмов влияние погрешностей округления можно учесть, рассматривая возмущенную систему (13). Считают, что решение системы (1), искаженное погрешностями округления, совпадает с точным решением некоторой системы (13). Иначе говоря, процесс решения системы (1) с учетом погрешностей округления эквивалентен точному решению некоторой возмущенной системы (13).

Предположим для простоты, что правая часть f задается точно. Пусть в результате погрешностей округления вместо точного ре-

шения системы (1) получено точное решение возмущенной системы

$$\tilde{A}\tilde{x}=f. \tag{22}$$

В этом случае матрица $\delta A=\tilde{A}-A$ называется *матрицей эквивалентных возмущений.* Каждому вычислительному алгоритму отвечает своя матрица эквивалентных возмущений. Если известна оценка нормы матрицы δA, то погрешность, возникшую в результате погрешностей округления, можно оценить согласно (15), а именно

$$\frac{\|\tilde{x}-x\|}{\|x\|}\leqslant\frac{M_A}{1-M_A\dfrac{\|\delta A\|}{\|A\|}}\frac{\|\delta A\|}{\|A\|}. \tag{23}$$

Отсюда видно, что на точность решения влияют два фактора: число обусловленности матрицы A и эквивалентное возмущение $\|\delta A\|/\|A\|$. Чем больше числа M_A и $\|\delta A\|/\|A\|$, тем меньше точность решения. Подчеркнем, что число обусловленности не связано с каким-либо численным алгоритмом, а характеризует только свойства исходной системы (1). Величина эквивалентного возмущения определяется численным алгоритмом. Тем самым при рассмотрении конкретных алгоритмов необходимо получать оценки соответствующих эквивалентных возмущений.

Так, в результате факторизации $A=LU$ по методу Гаусса решение системы (1) сводится к решению двух систем

$$Ly=f,\quad Ux=y$$

с треугольными матрицами. Погрешности округления приводят к тому, что вместо решения x системы $Ux=y$ получаем решение $\tilde{x}$ возмущенной системы $\tilde{U}\tilde{x}=\tilde{y}$. Точно так же вместо решения y системы $Ly=f$ получаем решение $\tilde{y}$ системы $\tilde{L}\tilde{y}=f$. Таким образом, можно считать, что вместо исходной системы (1) точно решается возмущенная система $\tilde{A}\tilde{x}=f$, где $\tilde{A}=\tilde{L}\tilde{U}$. Чтобы найти матрицу $\tilde{A}$, надо выписать все формулы метода Гаусса, внести в них погрешности округления и получить тем самым матрицы $\tilde{L}$ и $\tilde{U}$. Далее следует оценить норму матрицы $\delta A=LU-\tilde{L}\tilde{U}$. Опуская все эти весьма трудоемкие выкладки, приведем лишь окончательный результат (см. [6]).

Пусть t — число разрядов мантиссы в двоичном представлении чисел на ЭВМ с плавающей запятой, так что относительная погрешность округления действительного числа есть величина порядка 2^{-t}. Тогда для эквивалентного возмущения метода Гаусса верна оценка

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

где m — порядок матрицы.

Таким образом, накопление погрешностей округления при решении систем линейных алгебраических уравнений методом Гаусса

на ЭВМ с плавающей запятой приводит к тому, что искомое решение определяется с относительной погрешностью

$$\frac{\|\tilde{x}-x\|}{\|x\|}=O(M_A\cdot m\cdot 2^{-t}). \tag{24}$$

Отметим, что для машины БЭСМ-6 число 2^{-t} равно приблизительно 10^{-12}.

Рассмотрим пример применения оценки (24). Пусть методом конечных разностей решается краевая задача

$$u''(x)=-f(x),\quad 0<x<1,\quad u(0)=u(1)=0. \tag{25}$$

Введем равномерную сетку

$$\omega_h=\{x_i=ih,\ i=0,\ 1,\ \ldots,\ N,\ hN=1\}$$

и заменим (25) разностной схемой второго порядка точности

$$-\frac{u_{i+1}-2u_i+u_{i-1}}{h^2}=f_i,\quad i=1,\ 2,\ \ldots,\ N-1,\qquad u_0=u_N=0. \tag{26}$$

Казалось бы, чем меньше возьмем шаг h, тем с большей точностью получим решение задачи (25). Однако порядок $m=N-1$ системы линейных алгебраических уравнений (26) обратно пропорционален шагу h. Значит, уменьшение шага h приведет согласно (24) к увеличению погрешностей округления, и при некотором значении h погрешности округления могут превзойти погрешность разностного метода, пропорциональную h^2. Оценка (24) позволяет выбрать порядок шага h, при котором погрешность округления еще не превосходит погрешности метода. Остановимся на этом подробнее.

Поскольку матрица A системы (26) симметрична и положительно определена, ее число обусловленности равно отношению максимального собственного значения к минимальному, т. е.

$$M_A=\frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}.$$

Как известно (см. п. 4 § 4 ч. I), для данной матрицы

$$\lambda_{\min}(A)=\frac{4}{h^2}\sin^2\frac{\pi h}{2},\qquad \lambda_{\max}(A)=\frac{4}{h^2}\cos^2\frac{\pi h}{2},$$

поэтому

$$M_A=\operatorname{ctg}^2\frac{\pi h}{2}\approx\frac{4}{\pi^2h^2}=O\left(\frac{1}{h^2}\right).$$

Отсюда видно, что при малых h система (26) плохо обусловлена. Далее, $m=O(h^{-1})$ и

$$M_Am=O\left(\frac{1}{h^3}\right).$$

Поэтому правая часть равенства (24) оценивается как $O(2^{-t}h^{-3})$. Для того чтобы погрешность округления имела тот же порядок, что и погрешность разностного метода, достаточно потребовать $2^{-t}h^{-3}=O(h^2)$, т. е. $h=O(2^{-t/5})$. В частности, при $2^{-t}=10^{-12}$ приходим к выводу о том, что нецелесообразно брать шаг h меньше, чем 0,001, так как в противном случае накопление погрешностей округления может оказаться слишком велико.

ГЛАВА 2

ИТЕРАЦИОННЫЕ МЕТОДЫ РЕШЕНИЯ СИСТЕМ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

§ 1. Примеры и канонический вид итерационных методов решения систем линейных алгебраических уравнений

1. Итерационные методы Якоби и Зейделя. Перейдем к изучению итерационных методов решения систем линейных алгебраических уравнений. Будем рассматривать систему

$$Ax=f, \tag{1}$$

где матрица $A=[a_{ij}]$, $i, j=1, 2, \ldots, m$, имеет обратную, $x=(x_1, x_2, \ldots, x_m)^T$, $f=(f_1, f_2, \ldots, f_m)^T$.

Рассмотрим сначала два примера итерационных методов. Для их построения предварительно преобразуем систему (1) к виду

$$x_i=-\sum_{j=1}^{i-1}\frac{a_{ij}}{a_{ii}}x_j-\sum_{j=i+1}^{m}\frac{a_{ij}}{a_{ii}}x_j+\frac{f_i}{a_{ii}},\quad i=1, 2, \ldots, m \tag{2}$$

(при этом предполагается, что все a_{ii} отличны от нуля).

Условимся, как обычно, считать значение суммы равным нулю, если верхний предел суммирования меньше нижнего. Так, уравнение (2) при $i=1$ имеет вид

$$x_1=-\sum_{j=2}^{m}\frac{a_{1j}}{a_{11}}x_j+\frac{f_1}{a_{11}}.$$

В дальнейшем верхний индекс будет указывать номер итерации, например

$$x^n=(x_1^n, x_2^n, \ldots, x_m^n)^T,$$

где x_i^n — n-я итерация i-й компоненты вектора x.

В *методе Якоби* исходят из записи системы в виде (2), причем итерации определяются следующим образом:

$$x_i^{n+1}=-\sum_{j=1}^{i-1}\frac{a_{ij}}{a_{ii}}x_j^n-\sum_{j=i+1}^{m}\frac{a_{ij}}{a_{ii}}x_j^n+\frac{f_i}{a_{ii}}, \tag{3}$$

$$n=0, 1, \ldots, n_0,\quad i=1, 2, \ldots, m.$$

Начальные значения x_i^0, $i=1, 2, \ldots, m$ задаются произвольно. Окончание итераций определяется либо заданием максимального числа итераций n_0, либо условием

$$\max_{1\leqslant i\leqslant m}|x_i^{n+1}-x_i^n|<\varepsilon,$$

где $\varepsilon>0$ — заданное число. Позже в § 2 будет показано, что при определенных условиях на матрицу A метод Якоби сходится, т. е. $\|x^n-x\|\to 0$ при $n\to\infty$ (здесь x — точное решение системы (1), а x^n — приближенное решение, полученное на n-й итерации).

Итерационный *метод Зейделя* имеет вид

$$x_i^{n+1} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^{n+1} - \sum_{j=i+1}^{m} \frac{a_{ij}}{a_{ii}} x_j^n + \frac{f_i}{a_{ii}},$$

$$i=1, 2, \ldots, m, \quad n=0, 1, \ldots, n_0. \tag{4}$$

Чтобы понять, как находятся отсюда значения x_i^{n+1}, $i=1, 2, \ldots, m$, запишем подробнее первые два уравнения системы (4):

$$x_1^{n+1} = -\sum_{j=2}^{m} \frac{a_{1j}}{a_{11}} x_j^n + \frac{f_1}{a_{11}}, \tag{5}$$

$$x_2^{n+1} = -\frac{a_{21}}{a_{22}} x_1^{n+1} - \sum_{j=3}^{m} \frac{a_{2j}}{a_{22}} x_j^n + \frac{f_2}{a_{22}}. \tag{6}$$

Первая компонента x_1^{n+1} вектора x^{n+1} находится из уравнения (5) явным образом, для ее вычисления нужно знать вектор x^n и значение f_1. При нахождении x_2^{n+1} из уравнения (6) используются только что найденное значение x_1^{n+1} и известные значения x_j^n, $j=3, \ldots, m$, с предыдущей итерации. Таким образом, компоненты x_i^{n+1} вектора x^{n+1} находятся из уравнения (4) последовательно, начиная с $i=1$.

2. Матричная запись методов Якоби и Зейделя. Для исследования сходимости итерационных методов удобнее записывать их не в координатной, а в матричной форме. Представим матрицу A системы (1) в виде суммы трех матриц

$$A = A_1 + D + A_2, \tag{7}$$

где $D = \operatorname{diag}[a_{11}, a_{22}, \ldots, a_{mm}]$ — диагональная матрица с той же главной диагональю, что и матрица A, матрица A_1 — нижняя треугольная и матрица A_2 — верхняя треугольная с нулевыми главными диагоналями.

Например, при $m=3$ матрицы A_1, A_2, D имеют вид

$$A_1 = \begin{bmatrix} 0 & 0 & 0 \\ a_{21} & 0 & 0 \\ a_{31} & a_{32} & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & a_{12} & a_{13} \\ 0 & 0 & a_{23} \\ 0 & 0 & 0 \end{bmatrix}, \quad D = \begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix}.$$

Представление системы (1) в форме (2) эквивалентно ее записи в виде матричного уравнения

$$x = -D^{-1}A_1 x - D^{-1}A_2 x + D^{-1}f.$$

Отсюда видно, что метод Якоби (3) в векторной записи выглядит следующим образом:

$$x^{n+1} = -D^{-1}A_1 x^n - D^{-1}A_2 x^n + D^{-1}f,$$

или, что то же самое,

$$Dx^{n+1} + (A_1 + A_2)x^n = f. \tag{8}$$

Метод Зейделя (4) записывается в виде

$$x^{n+1}=-D^{-1}A_1x^{n+1}-D^{-1}A_2x^n+D^{-1}f$$

или

$$(D+A_1)x^{n+1}+A_2x^n=f. \tag{9}$$

Учитывая (7), методы (8) и (9) можно переписать соответственно в виде

$$D(x^{n+1}-x^n)+Ax^n=f, \tag{10}$$

$$(D+A_1)(x^{n+1}-x^n)+Ax^n=f. \tag{11}$$

Из этой записи видно, что если итерационный метод сходится, то он сходится к решению исходной системы уравнений.

Очень часто для ускорения сходимости в итерационные методы вводят числовые параметры, которые зависят, вообще говоря, от номера итерации. Например, в методы (10), (11) можно ввести *итерационные параметры* τ_{n+1} следующим образом:

$$D\frac{x^{n+1}-x^n}{\tau_{n+1}}+Ax^n=f,$$

$$(D+A_1)\frac{x^{n+1}-x^n}{\tau_{n+1}}+Ax^n=f.$$

Способ выбора итерационных параметров выясняется при исследовании сходимости. В теории итерационных методов существует два круга вопросов: а) при каких значениях параметров метод сходится, б) при каких значениях параметров сходимость будет наиболее быстрой (соответствующие параметры называются оптимальными). В дальнейшем (см. § 4) мы подробнее остановимся на этих вопросах в связи с конкретными итерационными методами.

Приведенные выше методы Якоби и Зейделя относятся к *одношаговым итерационным методам*, когда для нахождения x^{n+1} требуется помнить только одну предыдущую итерацию x^n. Иногда используются и *многошаговые итерационные методы*, в которых x^{n+1} определяется через значения x^k на двух и более предыдущих итерациях, т. е. $x^{n+1}=F[x^n, x^{n-1}, \ldots, x^{n-l}]$.

3. Каноническая форма одношаговых итерационных методов. На примере методов Якоби и Зейделя видно, что один и тот же итерационный метод можно записать многими различными способами. Поэтому целесообразно ввести какую-то стандартную форму записи итерационных методов. Условимся прежде всего записывать итерационный метод не в координатной форме, а в матричной, Теперь x_n будет обозначать вектор, полученный в результате n-й итерации.

Канонической формой одношагового итерационного метода решения системы (1) называется его запись в виде

$$B_{n+1}\frac{x_{n+1}-x_n}{\tau_{n+1}}+Ax_n=f, \quad n=0, 1, \ldots, n_0. \tag{12}$$

Здесь B_{n+1} — матрица, задающая тот или иной итерационный метод, τ_{n+1} — итерационный параметр. Предполагается, что задано начальное приближение x_0 и что существуют матрицы B_n^{-1}, $n=1, 2, \ldots, n_0-1$. Тогда из уравнения (12) можно последовательно определить все x_n, $n=1, 2, \ldots, n_0$. Для нахождения x_{n+1} по известным f и x_n достаточно решить систему уравнений

$$B_{n+1}x_{n+1}=F_n,$$

где $F_n=(B_{n+1}-\tau_{n+1}A)x_n+\tau_{n+1}f$.

Итерационный метод называют *явным* (*неявным*), если $B_n=E$ ($B_n\neq E$), где E — единичная матрица. Как правило, неявные итерационные методы имеет смысл применять лишь в том случае, когда каждую матрицу B_n обратить легче, чем исходную матрицу A (т. е. когда решение системы уравнений с матрицей B_n требует меньше машинной памяти или времени или алгоритмически проще, чем решение исходной системы). Например, в методе Зейделя приходится обращать треугольную матрицу. В дальнейшем (см. § 4) будет показано, что преимуществом неявных методов является более быстрая сходимость.

Итерационный метод (12) называется *стационарным*, если $B_{n+1}=B$ и $\tau_{n+1}=\tau$ не зависят от номера итерации, и *нестационарным* — в противоположном случае.

Приведем еще несколько примеров итерационных методов. *Методом простой итерации* называют явный метод

$$\frac{x_{n+1}-x_n}{\tau}+Ax_n=f \tag{13}$$

с постоянным параметром τ. Явный метод

$$\frac{x_{n+1}-x_n}{\tau_{n+1}}+Ax_n=f \tag{14}$$

с переменным параметром τ_{n+1} называется *итерационным методом Ричардсона*. Для методов (13), (14) известен способ выбора оптимальных итерационных параметров в том случае, когда A — симметричная положительно определенная матрица (см. § 6).

Обобщением метода Зейделя (11) является *метод верхней релаксации*

$$(D+\omega A_1)\frac{x_{n+1}-x_n}{\omega}+Ax_n=f, \tag{15}$$

где $\omega>0$ — заданный числовой параметр. В § 2 будет показано, что в случае симметричной положительно определенной матрицы A метод (15) сходится при $0<\omega<2$.

Для получения расчетных формул перепишем (15) в виде

$$(E+\omega D^{-1}A_1)x_{n+1}=((1-\omega)E-\omega D^{-1}A_2)x_n+\omega D^{-1}f.$$

В покомпонентной записи получим

$$x_i^{n+1}+\omega\sum_{j=1}^{i-1}\frac{a_{ij}}{a_{ii}}x_j^{n+1}=$$

$$=(1-\omega)\,x_i^n-\omega\sum_{j=i+1}^{m}\frac{a_{ij}}{a_{ii}}x_j^n+\omega\frac{f_i}{a_{ii}},\qquad i=1, 2, \ldots, m.$$

Отсюда последовательно, начиная с $i=1$, находим все x_i^{n+1}:

$$x_1^{n+1} = (1-\omega)\, x_1^n - \omega \sum_{j=2}^{m} \frac{a_{1j}}{a_{11}} x_j^n + \omega \frac{f_1}{a_{11}},$$

$$x_2^{n+1} = -\omega \frac{a_{21}}{a_{11}} x_1^{n+1} + (1-\omega) x_2^n - \omega \sum_{j=3}^{m} \frac{a_{2j}}{a_{22}} x_j^n + \omega \frac{f_2}{a_{22}}$$

и т. д.

§ 2. Исследование сходимости итерационных методов

Рассмотрим систему линейных алгебраических уравнений

$$Ax = f \tag{1}$$

с невырожденной действительной матрицей A и одношаговый стационарный итерационный метод, записанный в каноническом виде

$$B \frac{x_{n+1} - x_n}{\tau} + Ax_n = f, \quad n = 0, 1, \ldots, \tag{2}$$

где x_0 задан.

Говорят, что итерационный метод (2) *сходится*, если $\|x_n - x\| \to 0$ при $n \to \infty$. Под нормой вектора x будем понимать сейчас среднеквадратичную норму

$$\|x\| = \left(\sum_{j=1}^{m} x_j^2 \right)^{1/2}.$$

Решение x системы (1) будем рассматривать как элемент m-мерного евклидова пространства H со скалярным произведением

$$(u, v) = \sum_{j=1}^{m} u_j v_j.$$

При формулировке условий сходимости будут использоваться матричные неравенства. Для действительной матрицы C неравенство $C > 0$ означает, что $(Cx, x) > 0$ для всех $x \in H$, $x \neq 0$. Из неравенства $C > 0$ следует, что существует константа $\delta > 0$ такая, что $(Cx, x) \geqslant \delta \|x\|^2$.

Действительно, если $C > 0$ — симметричная матрица, то все ее собственные значения положительны и в качестве δ можно взять минимальное собственное значение. Если $C > 0$ — несимметричная матрица, то для любого $x \in H$, $x \neq 0$ имеем

$$(Cx, x) = \frac{1}{2} [(Cx, x) + (x, C^* x)] > 0,$$

где C^* — матрица, транспонированная к C. Поэтому в качестве δ можно взять минимальное собственное значение матрицы $C_0 = 0{,}5(C + C^*)$. Из оценки $(Cx, x) \geqslant \delta \|x\|^2$ следует, что существует матрица C^{-1}. Неравенство $C \geqslant 0$ означает, что $(Cx, x) \geqslant 0$ для всех $x \in H$. Если $C \geqslant 0$, то C^{-1} может и не существовать.

Перейдем к исследованию сходимости итерационного метода (2). *Погрешность метода на n-й итерации* характеризуется вектором $z_n = x_n - x$, который согласно (1), (2) удовлетворяет однородному уравнению

$$B\frac{z_{n+1}-z_n}{\tau} + Az_n = 0, \qquad n = 0, 1, \ldots, \qquad z_0 = x_0 - x. \tag{3}$$

Теорема 1. *Пусть A — симметричная положительно определенная матрица, $\tau > 0$ и пусть выполнено неравенство*

$$B - 0{,}5\tau A > 0. \tag{4}$$

Тогда итерационный метод (2) *сходится.*

Доказательство. Достаточно показать, что среднеквадратичная норма решения z_n уравнения (3) стремится к нулю при $n \to \infty$ и при любой начальной погрешности z_0. Покажем сначала, что при условии (4) числовая последовательность $J_n = (Az_n, z_n)$ является невозрастающей. Из уравнения (3) найдем

$$z_{n+1} = (E - \tau B^{-1}A)z_n, \quad Az_{n+1} = (A - \tau AB^{-1}A)z_n,$$

откуда получим

$$(Az_{n+1}, z_{n+1}) = (Az_n, z_n) - \tau(AB^{-1}Az_n, z_n) - \\ - \tau(Az_n, B^{-1}Az_n) + \tau^2(AB^{-1}Az_n, AB^{-1}Az_n).$$

Вследствие симметричности матрицы A имеем

$$(AB^{-1}Az_n, z_n) = (Az_n, B^{-1}Az_n),$$

поэтому

$$(Az_{n+1}, z_{n+1}) = (Az_n, z_n) - 2\tau((B - 0{,}5\tau A)B^{-1}Az_n, B^{-1}Az_n). \tag{5}$$

Отсюда, учитывая условие (4), получаем неравенство

$$(Az_{n+1}, z_{n+1}) \leqslant (Az_n, z_n).$$

Таким образом, числовая последовательность $J_n = (Az_n, z_n)$ монотонна и ограничена снизу нулем. Следовательно, существует

$$\lim_{n\to\infty} J_n = J. \tag{6}$$

Далее, из положительной определенности матрицы $B - 0{,}5\tau A$ следует существование константы $\delta > 0$ такой, что

$$((B - 0{,}5\tau A)B^{-1}Az_n, B^{-1}Az_n) \geqslant \delta\|B^{-1}Az_n\|^2.$$

Отсюда и из (5) получаем неравенство

$$J_{n+1} - J_n + 2\delta\tau\|B^{-1}Az_n\|^2 \leqslant 0.$$

Переходя в этом неравенстве к пределу при $n \to \infty$ и учитывая (6), убеждаемся в том, что существует

$$\lim_{n\to\infty}\|w_n\| = 0,$$

где $w_n = B^{-1}Az_n$. Наконец, замечая, что A — положительно опреде

ленная и, следовательно, обратимая матрица, получим

$$z_n = A^{-1}Bw_n, \quad \|z_n\| \leqslant \|A^{-1}B\| \|w_n\|$$

и тем самым

$$\lim_{n\to\infty} \|z_n\| = 0.$$

Теорема 1 доказана.

Замечание. Как показано в [32, с. 527], при условиях теоремы 1 для погрешности $z_n = x_n - x$ итерационного метода (2) справедлива оценка

$$\|z_n\|_A \leqslant \rho^n \|z_0\|_A,$$

где $\rho \in (0, 1)$, $\|z_n\|_A = (Az_n, z_n)^{1/2}$. Эта оценка означает, что метод сходится со скоростью геометрической прогрессии со знаменателем ρ. Константа $\rho = (1 - 2\tau\delta_* \delta/\|B\|^2)^{1/2}$, где δ — минимальное собственное значение матрицы A и δ_* — минимальное собственное значение матрицы $0{,}5\,(B^* + B - \tau A)$.

Применим теорему 1 к конкретным итерационным методам, рассмотренным в предыдущем параграфе. Метод Якоби имеет следующий канонический вид:

$$D(x_{n+1} - x_n) + Ax_n = f, \tag{7}$$

где $D = \operatorname{diag}[a_{11}, a_{22}, \ldots, a_{mm}]$. Таким образом, в данном случае $B = D$, $\tau = 1$.

Следствие 1. *Пусть A — симметричная положительно определенная матрица с диагональным преобладанием, т. е.*

$$a_{ii} > \sum_{j \neq i} |a_{ij}|, \qquad i = 1, 2, \ldots, m. \tag{8}$$

Тогда метод Якоби сходится.

Доказательство. Условие сходимости (4) в данном случае имеет вид $A < 2D$. Покажем, что это матричное неравенство следует из неравенств (8). Рассмотрим положительно определенную квадратичную форму

$$(Ax, x) = \sum_{i,j=1}^{m} a_{ij} x_i x_j$$

и воспользуемся оценками

$$(Ax, x) \leqslant \frac{1}{2} \sum_{i,j=1}^{m} |a_{ij}| x_i^2 + \frac{1}{2} \sum_{i,j=1}^{m} |a_{ij}| x_j^2 =$$

$$= \frac{1}{2} \sum_{i,j=1}^{m} |a_{ij}| x_i^2 + \frac{1}{2} \sum_{i,j=1}^{m} |a_{ji}| x_i^2.$$

Из условий симметричности и положительной определенности матрицы A имеем $a_{ji} = a_{ij}$, $a_{ii} > 0$, $i, j = 1, 2, \ldots, m$, и поэтому предыдущая оценка приводит к неравенству

$$(Ax, x) \leqslant \sum_{i,j=1}^{m} |a_{ij}| x_i^2 = \sum_{i=1}^{m} x_i^2 \left(\sum_{j \neq i} |a_{ij}| + a_{ii} \right). \tag{9}$$

Перепишем условие (8) в виде

$$a_{ii} + \sum_{j \neq i} |a_{ij}| < 2a_{ii}, \quad i = 1, 2, \ldots, m.$$

Тогда из неравенства (9) получим

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

что и требовалось.

Следствие 2. *Пусть A — симметричная положительно определенная матрица. Тогда метод верхней релаксации*

$$(D + \omega A_1) \frac{x_{n+1} - x_n}{\omega} + Ax_n = f$$

сходится при условии $0 < \omega < 2$. В частности, метод Зейделя ($\omega = 1$) сходится.

Доказательство. Метод верхней релаксации приводится к каноническому виду (2) с $B = D + \omega A_1$, $\tau = \omega$. Напомним, что исходная матрица A представляется в виде суммы $A = D + A_1 + A_2$, где A_1 — нижняя треугольная, A_2 — верхняя треугольная и D — диагональная матрицы (см. (7) из § 1). Для симметричной матрицы A матрица A_2 является транспонированной к A_1, поэтому

$$(Ax, x) = (Dx, x) + (A_1 x, x) + (A_2 x, x) = (Dx, x) + 2(A_1 x, x).$$

Условие сходимости (4) принимает вид

$$(Bx, x) - 0{,}5\omega (Ax, x) =$$

$$= ((D + \omega A_1)x, x) - 0{,}5\omega ((Dx, x) + 2(A_1 x, x)) = (1 - 0{,}5\omega)(Dx, x) > 0$$

и выполняется при $0 < \omega < 2$.

Рассмотрим еще вопрос о сходимости метода простой итерации

$$\frac{x_{n+1} - x_n}{\tau} + Ax_n = f \tag{10}$$

с симметричной положительно определенной матрицей A. Согласно (4) метод сходится при условии

$$E - 0{,}5\tau A > 0. \tag{11}$$

Какие ограничения на параметр τ накладывает условие (11)? Пусть λ_i, $i = 1, 2, \ldots, m$, — собственные значения матрицы A, расположенные в порядке возрастания. Условие (11) эквивалентно тому, что все собственные значения матрицы $E - 0{,}5\tau A$ положительны. Достаточно потребовать положительности минимального собственного числа этой матрицы, равного $1 - 0{,}5\tau\lambda_m$. Таким образом, итерационный метод (10) сходится, если

$$\tau < 2/\lambda_{\max}, \tag{12}$$

где $\lambda_{\max}$ — максимальное собственное число матрицы A.

Условие (12) и необходимо для сходимости метода (10), т. е. если (12) нарушено, то найдется начальное приближение x_0, при котором $\|x_n - x\| \not\to 0$ при $n \to \infty$.

Докажем последнее утверждение. Возьмем в качестве начального приближения вектор $x_0 = x + \mu$, где x — точное решение задачи (1), а μ — собственный вектор матрицы A, отвечающий собственному числу $\lambda_{max} = \lambda_m$, т. е. $A\mu = \lambda_m \mu$. При таком выборе начального приближения имеем $z_0 = x_0 - x = \mu$. Из уравнения (3) при $B = E$ получим

$$z_n = (E - \tau A)^n z_0 = (E - \tau A)^n \mu$$

и, следовательно, $z_n = (1 - \tau\lambda_m)^n \mu$, $\|z_n\| = |1 - \tau\lambda_m|^n \|\mu\|$.

Если $\tau = 2\lambda_m^{-1}$, то $\|z_n\| = \|\mu\| \not\to 0$ при $n \to \infty$. Если же $\tau > 2\lambda_m^{-1}$, то $|1 - \tau\lambda_m| > 1$ и $\|z_n\| \to \infty$ при $n \to \infty$. Таким образом, условие (12) необходимо и достаточно для сходимости метода простой итерации (10).

В заключение параграфа отметим, что теория итерационных методов не заканчивается исследованием сходимости. При наличии хотя бы двух итерационных методов возникает вопрос о том, какой из этих методов сходится быстрее, т. е. для какого метода погрешность $\|x_n - x\|$ станет меньше заданного числа ε при меньшем числе итераций n. Сюда же примыкает вопрос о нахождении итерационных параметров, минимизирующих число итераций, необходимых для получения заданной точности. Этот круг вопросов будет подробно рассмотрен в следующих параграфах.

§ 3. Необходимое и достаточное условие сходимости стационарных итерационных методов

1. Введение. Некоторые итерационные методы решения систем линейных алгебраических уравнений уже рассматривались в § 1, 2. Напомним необходимые для дальнейшего сведения.

Пусть дана система уравнений

$$Ax = f, \tag{1}$$

где $A = [a_{ij}]$, $i, j = 1, 2, \ldots, m$,— вещественная квадратная матрица, имеющая обратную, и $x = (x_1, x_2, \ldots, x_m)^T$, $f = (f_1, f_2, \ldots, f_m)^T$. Канонической формой одношагового итерационного метода называется его запись в виде

$$B_{n+1} \frac{x_{n+1} - x_n}{\tau_{n+1}} + Ax_n = f, \quad n = 0, 1, \ldots, \tag{2}$$

где n — номер итерации, x_0 — заданное начальное приближение, $x_n = (x_1^n, x_2^n, \ldots, x_m^n)^T$. Матрицы B_{n+1} и числа $\tau_{n+1} > 0$ задают тот или иной конкретный итерационный метод.

В настоящем параграфе подробно рассматриваются стационарные итерационные методы

$$B \frac{x_{n+1} - x_n}{\tau} + Ax_n = f, \tag{3}$$

в которых матрица B и числовой параметр τ не зависят от номера итерации n.

Погрешность итерационного метода (3) $v_n = x_n - x$, где x — точное решение системы (1), удовлетворяет уравнению

$$B\frac{v_{n+1}-v_n}{\tau} + Av_n = 0, \quad n = 0, 1, \ldots, \quad v_0 = x_0 - x, \tag{4}$$

которое отличается от уравнения (3) лишь тем, что является однородным.

Сходимость итерационного метода (3) означает, что $v_n \to 0$ в некоторой норме при $n \to \infty$. Переписывая уравнение (4) в разрешенной относительно v_{n+1} форме

$$v_{n+1} = Sv_n, \tag{5}$$

где

$$S = E - \tau B^{-1}A, \tag{6}$$

видим, что свойство сходимости итерационного метода целиком определяется матрицей S. Необходимые и достаточные условия сходимости в терминах матрицы S приведены ниже в п. 3. Матрица S называется *матрицей перехода от n-й итерации к* $(n+1)$*-й.*

2. Норма матрицы. При исследовании сходимости будем рассматривать векторы x_n и x как элементы m-мерного линейного пространства H, в котором введена норма $\|x\|$ вектора x. Нормой матрицы A, подчиненной данной норме вектора, называется число

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

Норму вектора в пространстве H можно ввести различным образом. Нам прежде всего потребуется норма

$$\|x\|_C = \max_{1 \leqslant i \leqslant m} |x_i|.$$

Подчиненная ей норма матрицы A выражается через элементы матрицы A следующим образом:

$$\|A\|_C = \max_{1 \leqslant i \leqslant m} \sum_{j=1}^{m} |a_{ij}|.$$

Докажем это утверждение. Для любого вектора x справедливо неравенство

$$\|Ax\|_C = \max_{1 \leqslant i \leqslant m} \left| \sum_{j=1}^{m} a_{ij}x_j \right| \leqslant \max_{1 \leqslant j \leqslant m} |x_j| \max_{1 \leqslant i \leqslant m} \sum_{j=1}^{m} |a_{ij}| =$$

$$= \left(\max_{1 \leqslant i \leqslant m} \sum_{j=1}^{m} |a_{ij}| \right) \|x\|_C,$$

т. е.

$$\|Ax\|_C \leqslant \left(\max_{1 \leqslant i \leqslant m} \sum_{j=1}^{m} |a_{ij}| \right) \|x\|_C. \tag{7}$$

Чтобы завершить доказательство, достаточно построить вектор $x_0 = (x_1^0, x_2^0, \ldots, x_m^0)^T$, для которого выполняется равенство

$$\| Ax_0 \|_C = \left(\max_{1 \leqslant i \leqslant m} \sum_{j=1}^{m} | a_{ij} | \right) \| x_0 \|_C. \tag{8}$$

Пусть функция

$$\varphi_i = \sum_{j=1}^{m} | a_{ij} |, \qquad i = 1, 2, \ldots, m,$$

достигает максимума при $i = k$, т. е.

$$\max_{1 \leqslant i \leqslant m} \sum_{j=1}^{m} | a_{ij} | = \sum_{j=1}^{m} | a_{kj} |. \tag{9}$$

Рассмотрим вектор x_0, имеющий координаты

$$x_j^0 = \begin{cases} 1, & \text{если } a_{kj} \geqslant 0, \\ -1, & \text{если } a_{kj} < 0. \end{cases} \tag{10}$$

Очевидно, что $\| x_0 \|_C = 1$. Оценим снизу выражение для $\| Ax_0 \|_C$. Имеем

$$\| Ax_0 \|_C = \max_{1 \leqslant i \leqslant m} \left| \sum_{j=1}^{m} a_{ij} x_j^0 \right| \geqslant \left| \sum_{j=1}^{m} a_{kj} x_j^0 \right|.$$

Далее, исходя из определения (10) вектора x_0, получим

$$\left| \sum_{j=1}^{m} a_{kj} x_j^0 \right| = \sum_{j=1}^{m} a_{kj} x_j^0 = \sum_{j=1}^{m} | a_{kj} |,$$

и, следовательно,

$$\| Ax_0 \|_C \geqslant \sum_{j=1}^{m} | a_{kj} | = \max_{1 \leqslant i \leqslant m} \sum_{j=1}^{m} | a_{ij} |.$$

Последнее равенство справедливо в силу (9). Тем самым нашли вектор x_0, для которого

$$\| Ax_0 \| \geqslant \max_{1 \leqslant i \leqslant m} \sum_{j=1}^{m} | a_{ij} | \, \| x_0 \|_C.$$

Поскольку для каждого вектора x справедливо противоположное неравенство (7), заключаем, что для x_0 справедливо равенство (8).

3. Теорема о сходимости итерационного метода. Справедлива

Теорема 1. *Итерационный метод* (3) *сходится при любом начальном приближении тогда и только тогда, когда все собственные значения матрицы* $S = E - \tau B^{-1} A$ *по модулю меньше единицы.*

Доказательство. Представим уравнение (4) для погрешности $v_n = x_n - x$ в виде (5)—(6). Докажем сначала необходимость условий теоремы 1. Предположим, что матрица S имеет собственное число s, для которого $|s| > 1$, и покажем, что в этом случае можно так подобрать начальное приближение x_0, чтобы погрешность $v_n = x_n - x$ неограниченно возрастала при $n \to \infty$. Пусть μ — собственный вектор матрицы S, отвечающий собственному числу

s, $|s|>1$. Возьмем в качестве начального приближения вектор $x_0=x+\mu$, так что начальная погрешность $v_0=\mu$. Тогда из уравнения (5) получим

$$v_n=S^nv_0=s^nv_0=s^n\mu$$

и $\|v_n\|=|s|^n\|\mu\|\to\infty$ при $n\to\infty$. Если $|s|=1$, то $\|v_n\|=\|\mu\|\not\to 0$ при $n\to\infty$.

Доказательство достаточности условий теоремы 1 проведем сначала в предположении, что матрица S имеет m линейно независимых собственных векторов. Пусть s_k, $k=1, 2, \ldots, m$,— собственные числа матрицы S и μ_k, $k=1, 2, \ldots, m$,— отвечающие им линейно независимые собственные векторы. Разложим начальную погрешность $v_0=x_0-x$ по векторам μ_k:

$$v_0=\sum_{k=1}^{m} c_k\mu_k.$$

Тогда получим

$$v_n=S^nv_0=\sum_{k=1}^{m} c_ks_k^n\mu_k.$$

В любой норме справедлива оценка

$$\|v_n\|\leqslant\rho^n\sum_{k=1}^{m}|c_k|\|\mu_k\|, \tag{11}$$

где $\rho=\max\limits_{1\leqslant k\leqslant m}|s_k|$—спектральный радиус матрицы S. Из оценки (11) в силу предположения теоремы 1 о том, что $\rho<1$, и следует сходимость метода.

4. Продолжение доказательства. В общем случае, когда система собственных векторов матрицы S не является полной, доказательство достаточности условий теоремы 1 проводится с помощью приведения S к жордановой форме. Напомним (см. [12], стр. 147), что для любой квадратной матрицы S порядка m существует невырожденная матрица P такая, что матрица $\widetilde{S}=P^{-1}SP$ имеет жорданову каноническую форму

$$\widetilde{S}=\begin{bmatrix}\widetilde{S}_1 & 0 & \ldots & 0\\ 0 & \widetilde{S}_2 & \ldots & 0\\ \cdot & \cdot & \cdot & \cdot\\ 0 & 0 & \ldots & \widetilde{S}_l\end{bmatrix},$$

где $\widetilde{S}_k$ либо собственное число матрицы S, либо жорданова клетка, т. е. матрица вида

$$\widetilde{S}_k=\begin{bmatrix}s_k & 1 & 0 & 0 & \ldots & 0\\ 0 & s_k & 1 & 0 & \ldots & 0\\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\ 0 & 0 & 0 & 0 & \ldots & 1\\ 0 & 0 & 0 & 0 & \ldots & s_k\end{bmatrix},$$

а s_k — собственные числа матрицы S.

Помимо обычной жордановой формы нам потребуется еще так называемая модифицированная жорданова форма матрицы S. Она строится следующим образом.

Применим к матрице S преобразование подобия $D^{-1}SD$ с диагональной матрицей $D=\operatorname{diag}[1, \varepsilon, \ldots, \varepsilon^{m-1}]$, где ε — любое положительное число. Нетрудно убедиться, что матрица

$$\hat{S}=D^{-1}SD$$

имеет ту же блочно-диагональную структуру, что и матрица S, однако жордановы клетки имеют теперь следующий вид:

$$\hat{S}_k=\begin{bmatrix} s_k & \varepsilon & 0 & 0 & \ldots & 0 \\ 0 & s_k & \varepsilon & 0 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \ldots & \varepsilon \\ 0 & 0 & 0 & 0 & \ldots & s_k \end{bmatrix}.$$

Матрицы S и $\hat{S}$ связаны равенством

$$\hat{S}=Q^{-1}SQ, \quad Q=PD. \tag{12}$$

Матрица $\hat{S}$ имеет в каждой строке не более двух отличных от нуля элементов, поэтому

$$\|\hat{S}\|_C \leqslant \rho(S)+\varepsilon, \tag{13}$$

где $\rho(S)$ — спектральный радиус матрицы S, т. е.

$$\rho(S)=\max_{1\leqslant k\leqslant m}|s_k|.$$

Напомним, что согласно (10) из § 6 гл. 1 подчиненная норма матрицы удовлетворяет неравенству

$$\|S\| \geqslant \rho(S). \tag{14}$$

Покажем теперь, что можно найти такую норму вектора, для которой подчиненная норма матрицы станет сколь угодно близкой к ее спектральному радиусу. Точнее, справедливо следующее утверждение.

Лемма 1. *Для любого $\varepsilon>0$ существует норма $\|\cdot\|_*$ вектора такая, что для подчиненной нормы матрицы справедливо неравенство*

$$\|S\|_* \leqslant \rho(S)+\varepsilon. \tag{15}$$

Доказательство. Воспользуемся преобразованием (12) и определим норму вектора $\|\cdot\|_*$ равенством

$$\|y\|_*=\|Q^{-1}y\|_C$$

для любого вектора $y\in H$. Для подчиненной нормы матрицы S имеем

$$\|S\|_*=\sup_{y\neq 0}\frac{\|Sy\|_*}{\|y\|_*}=\sup_{y\neq 0}\frac{\|Q^{-1}Sy\|_C}{\|Q^{-1}y\|_C}.$$

Обозначая $Q^{-1}y=x$ и учитывая (12), (13), получим отсюда

$$\|S\|_* = \sup_{x\neq 0} \frac{\|Q^{-1}SQx\|_C}{\|x\|_C} = \sup_{x\neq 0} \frac{\|\hat{S}x\|_C}{\|x\|_C} = \|\hat{S}\|_C \leqslant \rho(S) + \varepsilon,$$

что и требовалось.

Завершим доказательство теоремы 1. Из уравнения (5) получим

$$v_n = S^n v_0, \qquad n=0, 1, \ldots \tag{16}$$

Пусть $\|\cdot\|_*$ — норма, для которой выполнено неравенство (15). По условию теоремы $\rho(S)<1$, поэтому существует $\varepsilon>0$ такое, что $\|S\|_* \leqslant \rho(S)+\varepsilon \leqslant q<1$. Из (16) получим оценку

$$\|v_n\|_* \leqslant \|S\|_*^n \|v_0\|_* \leqslant q^n \|v_0\|_*, \tag{17}$$

из которой следует, что $\|v_n\|_* \to 0$ при любых начальных приближениях.

§ 4. Оценки скорости сходимости стационарных итерационных методов

1. Скорость сходимости итерационного метода. При практическом использовании итерационных методов важен не только сам факт сходимости, но и скорость, с которой приближенное решение сходится к точному. Так как при численном решении всегда осуществляется конечное число итераций, необходимо знать, во сколько раз уменьшается начальная погрешность после проведения заданного числа итераций. Ответить на эти вопросы позволяет анализ оценок погрешности итерационного метода.

В предыдущем параграфе при доказательстве теоремы 1 была получена оценка (17), которую можно переписать в виде

$$\|x_n - x\|_* \leqslant q^n \|x_0 - x\|_*, \qquad n=0, 1, \ldots, \tag{1}$$

где $q\in(0, 1)$. Если для погрешности итерационного метода выполняются оценки вида (1), то говорят, что метод сходится со скоростью геометрической прогрессии со знаменателем q.

Используя оценку (1), можно определить число итераций, достаточное для того, чтобы начальная погрешность уменьшилась в заданное число раз. Действительно, зададим произвольное $\varepsilon>0$ и потребуем, чтобы $q^n<\varepsilon$, т. е. чтобы

$$n \geqslant n_0(\varepsilon) = \frac{\ln(1/\varepsilon)}{\ln(1/q)}. \tag{2}$$

Тогда из (1) получим, что

$$\|x_n - x\|_* \leqslant \varepsilon \|x_0 - x\|_*,$$

т. е. после проведения $n_0(\varepsilon)$ итераций начальная погрешность $\|x_0 - x\|_*$ уменьшилась в ε^{-1} раз. Целая часть числа $n_0(\varepsilon)$ называется *минимальным числом итераций, необходимым для получения заданной точности* ε.

Выражение $\ln(1/q)$, находящееся в знаменателе числа $n_0(\varepsilon)$, называется *скоростью сходимости итерационного метода.* Скорость сходимости целиком определяется свойствами матрицы перехода S и не зависит ни от номера итерации n, ни от выбора начального приближения x_0, ни от задаваемой точности ε. Качество различных итерационных методов сравнивают обычно по их скорости сходимости: чем выше скорость сходимости, тем лучше метод.

2. Оценки скорости сходимости в случае симметричных матриц A и B. Продолжим изучение итерационных методов решения систем линейных алгебраических уравнений

$$Ax=f. \tag{3}$$

Будем по-прежнему рассматривать стационарные одношаговые итерационные методы

$$B\frac{x_{n+1}-x_n}{\tau}+Ax_n=f. \tag{4}$$

Теорема о сходимости, доказанная в предыдущем параграфе, имеет принципиальное теоретическое значение и накладывает минимальные ограничения на матрицы A и B. Однако ее непосредственное применение к конкретным итерационным методам не всегда возможно, так как отыскание или исследование спектра матрицы $S=E-\tau B^{-1}A$ является, как правило, более трудной задачей, чем решение системы (3).

В настоящем параграфе будет доказана теорема, в которой условия сходимости формулируются в виде легко проверяемых матричных неравенств, связывающих матрицы A и B. Аналогичная теорема о сходимости была доказана в § 2, однако там не были получены оценки скорости сходимости.

Будем рассматривать решение x системы (3) и последовательные приближения x_n как элементы конечномерного линейного пространства H, а матрицы A, B и другие — как операторы, действующие в пространстве H. Предположим, что в H введены скалярное произведение (y, v) и норма $\|y\|=\sqrt{(y, y)}$. Для двух симметричных матриц A и B неравенство $A\geqslant B$ означает, что $(Ax, x)\geqslant (Bx, x)$ для всех $x\in H$. В случае симметричной положительно определенной матрицы D будем обозначать $\|y\|_D=\sqrt{(Dy, y)}$.

Теорема 1. *Пусть A и B — симметричные положительно определенные матрицы, для которых справедливы неравенства*

$$\gamma_1 B\leqslant A\leqslant\gamma_2 B, \tag{5}$$

где γ_1, γ_2 — *положительные постоянные*, $\gamma_2>\gamma_1$. *При*

$$\tau=2/(\gamma_1+\gamma_2) \tag{6}$$

итерационный метод (4) *сходится и для погрешности справедливы оценки*

$$\|x_n-x\|_A\leqslant\rho^n\|x_0-x\|_A, \quad n=0, 1, \ldots, \tag{7}$$

$$\|x_n-x\|_B\leqslant\rho^n\|x_0-x\|_B, \quad n=0, 1, \ldots, \tag{8}$$

где $\|v\|_A=\sqrt{(Av, v)}$, $\|v\|_B=\sqrt{(Bv, v)}$ *и*

$$\rho=\frac{1-\xi}{1+\xi}, \quad \xi=\frac{\gamma_1}{\gamma_2}. \tag{9}$$

Доказательство теоремы 1 будет дано в п. 4. Сделаем необходимые замечания и приведем следствия из этой теоремы.

Рассмотрим обобщенную задачу на собственные значения

$$A\mu=\lambda B\mu. \tag{10}$$

Если для матриц A и B выполнены неравенства (5), то из (10) для любого собственного вектора получим неравенства

$$\gamma_1(B\mu, \mu)\leqslant(A\mu, \mu)=\lambda(B\mu, \mu)\leqslant\gamma_2(B\mu, \mu).$$

Отсюда следует, что

$$\gamma_1\leqslant\lambda_{\min}(B^{-1}A), \quad \gamma_2\geqslant\lambda_{\max}(B^{-1}A), \tag{11}$$

где $\lambda_{\min}(B^{-1}A)$ и $\lambda_{\max}(B^{-1}A)$ — минимальное и максимальное собственные числа задачи (10).

Таким образом, наиболее точными константами, с которыми выполняются неравенства (5), являются константы

$$\gamma_1=\lambda_{\min}(B^{-1}A), \quad \gamma_2=\lambda_{\max}(B^{-1}A).$$

В этом случае параметр

$$\tau_0=\frac{2}{\lambda_{\min}(B^{-1}A)+\lambda_{\max}(B^{-1}A)}$$

называется оптимальным итерационным параметром, так как он минимизирует величину

$$\rho=\frac{1-\xi}{1+\xi}, \quad \xi=\frac{\gamma_1}{\gamma_2}$$

на множестве всех положительных γ_1, γ_2, удовлетворяющих условиям (11).

Пусть $\lambda_{\min}(A)$ и $\lambda_{\max}(A)$ — соответственно минимальное и максимальное собственные значения матрицы A.

Следствие 1. *Если* $A^T=A>0$, *то для метода простой итерации*

$$\frac{x_{n+1}-x_n}{\tau}+Ax_n=f \tag{12}$$

при $\tau=\tau_0=\dfrac{2}{\lambda_{\min}(A)+\lambda_{\max}(A)}$ *справедлива оценка*

$$\|x_n-x\|\leqslant\rho_0^n\|x_0-x\|, \tag{13}$$

где $\rho_0=\dfrac{1-\xi}{1+\xi}$, $\xi=\dfrac{\lambda_{\min}(A)}{\lambda_{\max}(A)}$.

Следствие 2. *Для симметричной матрицы* A *и* $\tau_0 = 2/(\lambda_{\min}(A)+\lambda_{\max}(A))$ *справедливо равенство*

$$\|E-\tau_0 A\| = \rho_0,$$

где $\rho_0 = \dfrac{1-\xi}{1+\xi}$, $\xi = \dfrac{\lambda_{\min}(A)}{\lambda_{\max}(A)}$.

В приложениях часто встречаются задачи с плохо обусловленной матрицей A, когда отношение $\lambda_{\max}(A)/\lambda_{\min}(A)$ велико. В этом случае число ρ_0 близко к единице и метод простой итерации сходится медленно. Оценим число итераций $n_0(\varepsilon)$, которое требуется в случае малых ξ для достижения заданной точности ε, т. е. для получения оценки

$$\|x_n - x\| \leqslant \varepsilon \|x_0 - x\|.$$

Из условия $\rho_0{}^n < \varepsilon$ получаем, что $n \geqslant n_0(\varepsilon)$, где

$$n_0(\varepsilon) = \frac{\ln(1/\varepsilon)}{\ln(1/\rho_0)},$$

и при малых ξ имеем

$$n_0(\varepsilon) \approx \frac{\ln(1/\varepsilon)}{2\xi} = O\left(\frac{1}{\xi}\right). \tag{14}$$

Таким образом, метод простой итерации (12) в случае малых ξ является медленно сходящимся методом. Ускорить сходимость итерационных методов можно двумя способами: во-первых, за счет применения неявных итерационных методов (4), когда $B \neq E$, и, во-вторых, оставаясь в классе явных методов, можно выбрать $\tau = \tau_n$ зависящим от номера итерации и таким, чтобы уменьшить общее число итераций. Применяется и комбинация этих двух способов, т. е. используются неявные итерационные методы с переменными итерационными параметрами.

Использование неявных итерационных методов (4) объясняется тем, что при соответствующем выборе матрицы B отношение $\xi = \dfrac{\lambda_{\min}(B^{-1}A)}{\lambda_{\max}(B^{-1}A)}$ для обобщенной задачи на собственные значения (10) будет больше, чем отношение $\dfrac{\lambda_{\min}(A)}{\lambda_{\max}(A)}$.

3. Правила действий с матричными неравенствами. Прежде чем переходить к доказательству теоремы 1, приведем необходимые для дальнейшего сведения из линейной алгебры.

1) *Если* A *— вещественная симметричная матрица, то существует ортогональная матрица* Q *(т. е.* $Q^T = Q^{-1}$*) такая, что* $A = Q^T \Lambda Q$, *где* Λ *— диагональная матрица. На главной диагонали матрицы* Λ *находятся собственные значения матрицы* A. Доказательство см. в [12, с. 156].

2) *Для симметричной матрицы* A *неравенство* $A \geqslant 0$ $(A > 0)$ *эквивалентно неотрицательности (положительности) всех ее собственных значений.*

Доказательство. Используя свойство 1), получим для любого $x \in H$, что

$$(Ax, x) = (Q^T \Lambda Q x, x) = (\Lambda Q x, Qx) = \sum_{i=1}^{m} \lambda_i y_i^2,$$

где λ_i — собственные числа матрицы A и y_i — i-я компонента вектора $y = Qx$. Отсюда сразу следует, что если все $\lambda_i \geqslant 0$ ($\lambda_i > 0$), то $(Ax, x) \geqslant 0$ для любого $x \in H$ ($(Ax, x) > 0$ для любого $x \neq 0$). Обратно, пусть λ_j — любое собственное число матрицы A. Зададим вектор y, у которого все компоненты кроме j-й равны нулю, а $y_j = 1$. Так как матрица $Q^{-1} = Q^T$ существует, для заданного вектора y найдется вектор $x \in H$ такой, что $Qx = y$. Но тогда

$$0 \leqslant (Ax, x) = (\Lambda y, y) = \lambda_j,$$

т. е. $\lambda_j \geqslant 0$.

3) *Если* $A^T = A > 0$, *то существует* A^{-1}.

Доказательство. Согласно 2) все собственные числа матрицы A положительны, следовательно, $\det A \neq 0$ и существует A^{-1}.

4) *Для симметричной матрицы* S *и любого числа* $\rho > 0$ *эквивалентны следующие* матричные неравенства:

$$-\rho E \leqslant S \leqslant \rho E, \tag{15}$$

$$S^2 \leqslant \rho^2 E. \tag{16}$$

Доказательство. Согласно свойству 2) условие (15) эквивалентно неравенствам

$$|s_k| \leqslant \rho, \quad k = 1, 2, \ldots, m,$$

где s_k — собственные числа матрицы S. Отсюда получаем $s_k^2 \leqslant \rho^2$, $k = 1, 2, \ldots, m$, что в свою очередь эквивалентно (16).

5) *Если* $A^T = A$ *и* $A \geqslant 0$ ($A > 0$), *то существует матрица* B, *обладающая следующими свойствами:*

$$B^2 = A, \quad B^T = B, \quad B \geqslant 0 \quad (B > 0). \tag{17}$$

Эта матрица называется *квадратным корнем из матрицы* A и обозначается $A^{1/2}$.

Доказательство. Пусть λ_i — собственные числа матрицы A, $i = 1, 2, \ldots, m$. Согласно свойству 1) существует ортогональная матрица Q такая, что

$$QAQ^T = \Lambda = \operatorname{diag}[\lambda_1, \lambda_2, \ldots, \lambda_m].$$

Поскольку все λ_i неотрицательны, можно определить матрицу $\Lambda^{1/2}$ как

$$\Lambda^{1/2} = \operatorname{diag}[\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_m}].$$

Тогда матрица $B = Q^T \Lambda^{1/2} Q$ обладает свойствами (17).

6) *Пусть* $A^T = A$ *и* L — *невырожденная матрица. Тогда эквивалентны неравенства*

$$A \geqslant 0, \quad L^T A L \geqslant 0.$$

Аналогично, эквивалентны строгие неравенства

$$A>0, \quad L^TAL>0.$$

Доказательство. Для любого $x\in H$ имеем $(L^TALx, x)=(ALx, Lx)$. Значит, $L^TAL\geqslant 0$, если $A\geqslant 0$. Докажем обратное. Так как L^{-1} существует, любой $x\in H$ можно представить в виде $x=Ly$, где $y=L^{-1}x$. Тогда получим

$$(Ax, x)=(L^TALy, y)\geqslant 0,$$

т. е. $A\geqslant 0$.

7) *Если A и B — симметричные и L — невырожденная матрицы, то эквивалентны неравенства*

$$A\geqslant B, \quad L^TAL\geqslant L^TBL.$$

Доказательство следует немедленно из (6).

8) *Пусть $C^T=C>0$ и α, β — любые действительные числа. Тогда эквивалентны неравенства*

$$\alpha C\geqslant \beta E, \quad \alpha E\geqslant \beta C^{-1}.$$

Доказательство. Согласно 5) существует матрица $C^{1/2}=(C^{1/2})^T>0$. Используя свойство 7), перейдем от первого неравенства ко второму с помощью следующей цепочки эквивалентных неравенств:

$$\alpha (C^{-1/2})\, C\, (C^{-1/2})\geqslant \beta\, (C^{-1/2})\, (C^{-1/2}),$$
$$\alpha (C^{-1/2}C^{1/2})\, (C^{1/2}C^{-1/2})\geqslant \beta C^{-1},$$
$$\alpha E\geqslant \beta C^{-1}.$$

9) *Пусть $A^T=A>0$, $B^T=B>0$, α и β — любые действительные числа. Тогда эквивалентны неравенства*

$$\alpha A\geqslant \beta B, \quad \alpha B^{-1}\geqslant \beta A^{-1}.$$

Доказательство. Умножая первое неравенство слева и справа на $B^{-1/2}$, перейдем к эквивалентному неравенству

$$\alpha C\geqslant \beta E, \quad C=B^{-1/2}AB^{-1/2}.$$

Согласно 8) последнее неравенство эквивалентно неравенству $\alpha E\geqslant \beta C^{-1}$, т. е.

$$\alpha E\geqslant \beta B^{1/2}A^{-1}B^{1/2},$$

умножая которое слева и справа на $B^{-1/2}$, получим $\alpha B^{-1}\geqslant \beta A^{-1}$.

4. Доказательство теоремы 1. Уравнение для погрешности $v_n=x_n-x$ имеет вид

$$B\frac{v_{n+1}-v_n}{\tau}+Av_n=0, \quad n=0, 1, \ldots, \tag{18}$$

$$v_0=x_0-x,$$

откуда получим

$$v_{n+1}=Sv_n, \quad S=E-\tau B^{-1}A. \tag{19}$$

Лемма 1. *Пусть A и B — симметричные положительно определенные матрицы и $\rho>0$ — число. Матричные неравенства*

$$\frac{1-\rho}{\tau}B\leqslant A\leqslant\frac{1+\rho}{\tau}B \tag{20}$$

необходимы и достаточны для того, чтобы при любых $v_0\in H$ для решения задачи (18) *выполнялась оценка*

$$\|v_{n+1}\|_A\leqslant\rho\|v_n\|_A,\quad n=0,\ 1,\ \ldots \tag{21}$$

Доказательство. Оценку (21) можно записать в виде

$$\|w_{n+1}\|\leqslant\rho\|w_n\|, \tag{22}$$

где $w_n=A^{1/2}v_n$, $\|w_n\|=\sqrt{(w_n,\ w_n)}$. Из (19) получим, что функция w_n удовлетворяет уравнению

$$w_{n+1}=Sw_n, \tag{23}$$

где $S=A^{1/2}SA^{-1/2}=E-\tau C$, $C=A^{1/2}B^{-1}A^{1/2}$. Для решения этого уравнения в силу симметричности матрицы S имеем

$$\|w_{n+1}\|^2=(Sw_n,\ Sw_n)=(S^2w_n,\ w_n).$$

Тем самым оценка (22) эквивалентна неравенству

$$S^2\leqslant\rho^2E \tag{24}$$

и остается доказать эквивалентность неравенств (20) и (24).

Согласно свойству 4) из п. 3, неравенство (24) эквивалентно двум матричным неравенствам

$$-\rho E\leqslant S\leqslant\rho E$$

или

$$\frac{1-\rho}{\tau}E\leqslant C\leqslant\frac{1+\rho}{\tau}E.$$

Так как $C=A^{1/2}B^{-1}A^{1/2}$ — симметричная положительно определенная матрица, согласно свойству 8) из п. 3, в этих неравенствах можно перейти к обратным матрицам, т. е. записать, что

$$\frac{1-\rho}{\tau}C^{-1}\leqslant E\leqslant\frac{1+\rho}{\tau}C^{-1}.$$

Подставляя сюда выражение для C, получим

$$\frac{1-\rho}{\tau}A^{-1/2}BA^{-1/2}\leqslant E\leqslant\frac{1+\rho}{\tau}A^{-1/2}BA^{-1/2}.$$

Умножая последние неравенства слева и справа на $A^{1/2}$ (см. свойство 6) из п. 3), приходим к неравенствам (20). Лемма 1 доказана.

Лемма 2. *При тех же условиях что и в лемме* 1, *неравенства* (20) *необходимы и достаточны для выполнения оценки*

$$\|v_{n+1}\|_B\leqslant\rho\|v_n\|_B,\quad n=0,\ 1,\ \ldots \tag{25}$$

Доказательство проводится почти так же, как и в лемме 1, только в качестве w_n надо взять вектор $B^{1/2}v_n$, а в качестве C — матрицу $B^{-1/2}AB^{-1/2}$.

Для доказательства теоремы 1 теперь достаточно заметить, что матричные неравенства (5) можно переписать в виде (20), где

$$\rho=\frac{1-\xi}{1+\xi},\quad \xi=\frac{\gamma_1}{\gamma_2},\quad \tau=\frac{2}{\gamma_1+\gamma_2}.$$

После этого замечания утверждение теоремы 1 следует из лемм 1 и 2.

5. Оценка погрешности в случае несимметричной матрицы B. Пусть задана любая симметричная положительно определенная матрица D. Обозначим через S матрицу перехода итерационного метода (4), т. е. $S=E-\tau B^{-1}A$, и через v_n — погрешность метода, $v_n=x_n-x$. Для исследования сходимости итерационных методов в случае несимметричных матриц A и B может оказаться полезной следующая простая

Лемма 3. *Если B^{-1} существует, то для выполнения оценок*

$$\|v_{n+1}\|_D\leqslant\rho\|v_n\|_D,\quad n=0,\ 1,\ \ldots \tag{26}$$

необходимо и достаточно выполнение матричного неравенства

$$\rho^2D\geqslant S^TDS. \tag{27}$$

Доказательство. Учитывая уравнение для погрешности (19), перепишем (26) в виде

$$(DSv_n,\ Sv_n)\geqslant\rho^2(Dv_n,\ v_n)$$

или

$$\rho^2(Dv_n,\ v_n)\geqslant(S^TDSv_n,\ v_n),\quad n=0,\ 1,\ \ldots$$

Так как v_0 произвольно, отсюда следует (27).

Обратно, если выполнено (27), то

$$\|v_{n+1}\|_D^2=(Dv_{n+1},\ v_{n+1})=(S^TDSv_n,\ v_n)\leqslant\rho^2(Dv_n,\ v_n)=\rho^2\|v_n\|_D^2,$$

т. е. приходим к (26).

В следующей теореме сформулированы достаточные условия сходимости метода (4) в случае несимметричной матрицы B.

Теорема 2. *Пусть A — симметричная положительно определенная матрица и B — невырожденная матрица. Если выполнено матричное неравенство*

$$\frac{B^T+B}{2}-\frac{\tau}{2}A\geqslant\frac{1-\rho^2}{2\tau}B^TA^{-1}B \tag{28}$$

с константой $\rho\in(0,\ 1)$, не зависящей от n, то итерационный метод (4) сходится и для погрешности справедлива оценка

$$\|x_n-x\|_A\leqslant\rho^n\|x_0-x\|_A. \tag{29}$$

Доказательство. Достаточно показать, что выполнены условия леммы 3 при $D=A$. Запишем неравенство (27) для $D=A$

в виде

$$\rho^2 A \geqslant (E-\tau A B^T)A(E-\tau B^{-1}A).$$

Раскрывая скобки в правой части этого неравенства, получим

$$\tau A(B^{T-1}+B^{-1})A \geqslant (1-\rho^2)A+\tau^2 AB^{T-1}AB^{-1}A. \qquad (30)$$

Согласно свойству 7) матричных неравенств можно умножить каждую часть неравенства (30) справа на матрицу $L=A^{-1}B$ и одновременно слева — на матрицу $L^T=B^TA^{-1}$. Тогда получим эквивалентное (30) неравенство

$$\tau(B+B^T) \geqslant (1-\rho^2)B^TA^{-1}B+\tau^2 A,$$

которое совпадает с (28). Таким образом, из (28) следует (27) и согласно лемме 3 — оценка (29). Поскольку $\rho \in (0, 1)$, из оценки (29) получаем, что $\|x_n-x\|_A \to 0$ при $n\to\infty$, т. е. метод (4) сходится. Теорема 2 доказана.

Замечание. Лемма 3 и теорема 2 остаются справедливыми и в случае комплексных матриц A и B, если только заменить S^T и B^T на матрицы S^* и B^*, комплексно сопряженные с матрицами S и B. В частности, условие (28) принимает вид

$$B_0 - 0{,}5\tau A \geqslant \frac{1-\rho^2}{2\tau} B^*A^{-1}B, \qquad (31)$$

где $B_0=0{,}5(B+B^*)$.

§ 5. Многочлены Чебышева

1. Многочлен Чебышева на отрезке $[-1, 1]$. В ряде вопросов численного анализа, связанных с проблемой минимизации погрешности вычислительного алгоритма, нашли применение многочлены, наименее уклоняющиеся от нуля.

Рассмотрим следующую задачу: среди всех многочленов степени n со старшим коэффициентом 1 найти такой многочлен $T_n(x)$, для которого величина

$$\max_{x\in[-1,1]} |T_n(x)|$$

является минимальной. Многочлен, обладающий этим свойством, называется *многочленом, наименее уклоняющимся от нуля на отрезке* $[-1, 1]$ или *многочленом Чебышева.* В этом параграфе будет показано, что функция

$$T_n(x) = 2^{1-n}\cos(n \arccos x) \qquad (1)$$

является многочленом Чебышева.

Рассмотрим сначала функцию

$$P_n(x) = \cos(n \arccos x), \qquad (2)$$

которая отличается от $T_n(x)$ только постоянным множителем. Проводя преобразование

$$\cos((n+1)\arccos x)+\cos((n-1)\arccos x) = 2\cos(n \arccos x)\cos(\arccos x) = 2xP_n(x),$$

убеждаемся в том, что справедливо рекуррентное соотношение

$$P_{n+1}(x) - 2xP_n(x) + P_{n-1}(x) = 0. \tag{3}$$

Кроме того, согласно (2) имеем $P_0(x) = 1$, $P_1(x) = x$. Отсюда и из (3) по индукции легко доказать, что $P_n(x)$ — многочлен степени n со старшим коэффициентом 2^{n-1}, $n = 1, 2, \ldots$ Следовательно, $T_n(x)$ — многочлен степени n со старшим коэффициентом 1.

Замечание. Для вещественных x правая часть выражения (1) определена только при $|x| \leqslant 1$. Если $|x| \geqslant 1$, то многочлен $T_n(x)$ доопределяется формулой

$$T_n(x) = 2^{-n}((x + \sqrt{x^2 - 1})^n + (x - \sqrt{x^2 - 1})^n).$$

Возможность такого доопределения объясняется тем, что для любого комплексного числа z справедливо тождество

$$\cos(n \arccos z) = 0{,}5((z + \sqrt{z^2 - 1})^n + (z - \sqrt{z^2 - 1})^n).$$

Корни многочлена $T_n(x)$ расположены в точках

$$x_k = \cos\frac{(2k+1)\pi}{2n}, \quad k = 0, 1, \ldots, n-1, \tag{4}$$

а экстремумы — в точках

$$x'_k = \cos\frac{k\pi}{n}, \quad k = 0, 1, \ldots, n, \tag{5}$$

причем

$$T_n(x'_k) = (-1)^k 2^{1-n}, \quad k = 0, 1, \ldots, n. \tag{6}$$

Следовательно,

$$\max_{x \in [-1,1]} |T_n(x)| = 2^{1-n}. \tag{7}$$

Докажем теперь, что среди всех многочленов степени n со старшим коэффициентом 1 многочлен $T_n(x)$ наименее уклоняется от нуля на отрезке $[-1, 1]$. Пусть $Q_n(x)$ — любой многочлен степени n со старшим коэффициентом 1. Обозначим

$$\|Q_n\| = \max_{x \in [-1,1]} |Q_n(x)|.$$

Лемма 1. *Пусть существует система точек*

$$-1 \leqslant x'_n \leqslant x'_{n-1} < \ldots < x'_1 < x'_0 \leqslant 1 \tag{8}$$

такая, что

$$|Q_n(x'_k)| = \|Q_n\|, \quad k = 0, 1, \ldots, n, \tag{9}$$

причем числа $Q_n(x'_k)$ имеют чередующиеся знаки. Тогда среди всех многочленов степени n со старшим коэффициентом 1 многочлен $Q_n(x)$ наименее уклоняется от нуля.

Доказательство. Предположим обратное, т. е. что существует многочлен $\bar{Q}_n(x)$ степени n со старшим коэффициентом 1, для которого $\|\bar{Q}_n\| < \|Q_n\|$ и, следовательно,

$$|\bar{Q}_n(x)| < \|Q_n\| \tag{10}$$

для всех $x \in [-1, 1]$. Рассмотрим функцию $R(x) = Q_n(x) - \overline{Q}_n(x)$, которая является многочленом степени $n-1$, отличным от тождественного нуля. Согласно условию леммы числа $Q_n(x_k')$ имеют чередующиеся знаки. Пусть для определенности

$$Q_n(x_k') = (-1)^k \|Q_n\|, \quad k = 0, 1, \ldots, n$$

(случай, когда $Q_n(x_k') = (-1)^{k+1}\|Q_n\|$ рассматривается аналогично). Тогда

$$R(x_k') = (-1)^k \|Q_n\| - \overline{Q}_n(x_k'), \quad k = 0, 1, \ldots, n$$

и согласно (10) получим, что многочлен $R(x)$ на отрезке $[-1, 1]$ меняет знак n раз, т. е. имеет n корней. Но это невозможно потому, что $R(x)$ — многочлен степени $n-1$, отличный от тождественного нуля. Полученное противоречие и доказывает лемму 1.

Замечание. Справедливо утверждение, обратное лемме 1: если $Q_n(x)$ — многочлен со старшим коэффициентом 1, наименее уклоняющийся от нуля на $[-1, 1]$, то найдется система точек (8), для которой выполняются равенства (9), причем числа $Q_n(x_k')$ имеют чередующиеся знаки.

На доказательстве этого утверждения останавливаться не будем.

Согласно (6), (7) многочлен $T_n(x)$ удовлетворяет всем условиям леммы 1, поэтому он наименее уклоняется от нуля на отрезке $[-1, 1]$ среди всех многочленов степени n со старшим коэффициентом 1.

2. Случай произвольного отрезка. Иногда требуется найти многочлен, наименее уклоняющийся от нуля на заданном отрезке $[a, b]$, среди всех многочленов степени n со старшим коэффициентом 1. Эта задача сводится к предыдущей с помощью замены

$$t = \frac{2}{b-a}x - \frac{b+a}{b-a},$$

переводящей отрезок $a \leqslant x \leqslant b$ в отрезок $-1 \leqslant t \leqslant 1$. При такой замене многочлена Чебышева

$$T_n(t) = 2^{1-n}\cos(n \arccos t) \tag{11}$$

преобразуется к виду

$$F_n(x) = 2^{1-n}\cos\left(n \arccos \frac{2x-(b+a)}{b-a}\right),$$

причем коэффициент при x^n оказывается равным $2^n/(b-a)^n$. Следовательно, многочленом, наименее уклоняющимся от нуля на $[a, b]$, среди всех многочленов степени n со старшим коэффициентом 1 является многочлен

$$T_n(x) = \frac{(b-a)^n}{2^{2n-1}}\cos\left(n \arccos \frac{2x-(b+a)}{b-a}\right). \tag{12}$$

Корни этого многочлена расположены в точках

$$x_k = \frac{a+b}{2} + \frac{b-a}{2}\cos\frac{(2k+1)\pi}{2n}, \quad k = 0, 1, \ldots, n-1, \tag{13}$$

а его максимальное отклонение от нуля равно

$$\max_{x\in[a,b]} |T_n(x)| = \frac{(b-a)^n}{2^{2n-1}}. \tag{14}$$

3. Другая нормировка многочленов Чебышева. В теории итерационных методов возникает следующая задача: найти многочлен $P_n(x)$ степени n, наименее уклоняющийся от нуля на $[a, b]$, среди всех многочленов степени n, принимающих при $x=0$ значение 1. Ясно, что искомый многочлен отличается от многочлена (12) только нормировкой, т. е.

$$P_n(x) = \frac{T_n(x)}{T_n(0)}. \tag{15}$$

Будем считать в дальнейшем, что $T_n(0)\neq 0$.

Если $T_n(0)=0$, то задача не имеет решения в классе многочленов заданной степени n. Например, для многочлена первой степени $P_1(x)=ax+1$ имеем

$$\max_{x\in[-1,1]} |P_1(x)| = 1+|a|$$

и минимум достигается при $a=0$. Но при этом $P_1(x)$ перестает быть многочленом первой степени.

Из (12), (15) получим при $b>a>0$, что

$$P_n(x) = p_n \cos\left(n \arccos \frac{2x-(b+a)}{b-a}\right), \tag{16}$$

где

$$p_n = \left(\cos\left(n \arccos \frac{b+a}{a-b}\right)\right)^{-1}.$$

Обозначая

$$\xi = \frac{a}{b}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \tag{17}$$

получим

$$p_n = \left(\cos\left(n \arccos\left(-\frac{1}{\rho_0}\right)\right)\right)^{-1}. \tag{18}$$

Для дальнейших преобразований воспользуемся тождествами

$$\cos(n \arccos(-z)) = (-1)^n \cos(n \arccos z) =$$
$$= (-1)^n 0{,}5\left((z+\sqrt{z^2-1})^n + (z-\sqrt{z^2-1})^n\right). \tag{19}$$

При $z=\rho_0^{-1}$ имеем

$$z-\sqrt{z^2-1} = \frac{1}{\rho_0} - \sqrt{\frac{1}{\rho_0^2}-1} = \frac{1-\sqrt{1-\rho_0^2}}{\rho_0}$$

и, подставляя сюда выражение для ρ_0 из (17), получим

$$z-\sqrt{z^2-1} = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad z+\sqrt{z^2-1} = \frac{1+\sqrt{\xi}}{1-\sqrt{\xi}}.$$

Отсюда и из (18), (19) получаем

$$p_n = (-1)^n \left[\frac{1}{2}\left(\rho_1^n + \frac{1}{\rho_1^n}\right)\right]^{-1} = (-1)^n \frac{2\rho_1^n}{1+\rho_1^{2n}},$$

где $\rho_1 = (1-\sqrt{\xi})/(1+\sqrt{\xi})$. Таким образом, приходим к следующему выводу: среди всех многочленов степени n, принимающих при $x=0$ значение 1, наименее уклоняется от нуля на отрезке $[a, b]$ многочлен

$$P_n(x) = (-1)^n q_n \cos\left(n \arccos \frac{2x-(b+a)}{b-a}\right), \tag{20}$$

где

$$q_n = \frac{2\rho_1^n}{1+\rho_1^{2n}}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \xi = \frac{a}{b}, \tag{21}$$

$$b > a > 0.$$

Корни многочлена (20) расположены в точках

$$x_k = \frac{a+b}{2} + \frac{b-a}{2} \cos \frac{(2k-1)}{2n} \pi, \quad k = 1, 2, \ldots, n. \tag{22}$$

4. Примеры применения многочленов Чебышева.

Пример 1. В теории интерполирования возникает следующая задача. Рассмотрим многочлен

$$\omega(x) = (x-x_0)(x-x_1)\cdots(x-x_n)$$

степени $n+1$. Требуется так подобрать числа x_k (среди которых нет совпадающих чисел), принадлежащие заданному отрезку $[a, b]$, чтобы минимизировать величину

$$\max_{x \in [a,b]} |\omega(x)|.$$

Поскольку старший коэффициент многочлена $\omega(x)$ равен 1, для решения данной задачи достаточно потребовать, чтобы $\omega(x)$ совпал с многочленом Чебышева

$$T_{n+1}(x) = \frac{(b-a)^{n+1}}{2^{2n+1}} \cos\left((n+1) \arccos \frac{2x-(b+a)}{b-a}\right)$$

(см. (12)). Условие $|\omega(x)| \equiv |T_{n+1}(x)|$ будет выполнено тогда и только тогда, когда совпадут все корни многочленов $\omega(x)$ и $T_{n+1}(x)$. Корнями многочлена $\omega(x)$ являются числа $x_0, x_1, \ldots, x_n$, а корни $T_{n+1}(x)$ определяются согласно (13) формулами

$$\tilde{x}_k = \frac{a+b}{2} + \frac{b-a}{2} \cos \frac{(2k+1)\pi}{2(n+1)}, \quad k = 0, 1, \ldots, n. \tag{23}$$

Таким образом, если задать точки x_k по правилу

$$x_k = \frac{a+b}{2} + \frac{b-a}{2} \cos \frac{(2k+1)\pi}{2(n+1)}, \quad k = 0, 1, \ldots, n, \tag{24}$$

то величина отклонения многочлена $\omega(x)$ от нуля окажется минимальной и равной

$$\max_{x\in[a,b]} |\omega(x)| = \frac{(b-a)^{n+1}}{2^{2n+1}}.$$

Заметим, что для минимизации отклонения многочлена $\omega(x)$ от нуля не обязательно точки x_k, $k=0, 1, \ldots, n$, располагать в том порядке, который указан формулами (24). Важно лишь, чтобы множество точек $\{x_k\}_{k=0}^n$ совпало с множеством $\{\widetilde{x}_k\}_{k=0}^n$ корней многочлена Чебышева. Такое множество точек $\{x_k\}_{k=0}^n$ естественно назвать оптимальным. Если множество $\{x_k\}_{k=0}^n$ оптимальное, то любая перестановка его элементов приводит также к оптимальному множеству. Потребуем, например, чтобы выполнялось условие

$$a\leqslant x_0<x_1<\cdots<x_n\leqslant b.$$

Тогда для оптимальности множества $\{x_k\}_{k=0}^n$ достаточно положить $x_k=\widetilde{x}_{n-k}$, $k=0, 1, \ldots, n$, т. е.

$$x_k = \frac{a+b}{2} - \frac{b-a}{2}\cos\frac{(2k+1)\pi}{2(n+1)}, \quad k=0, 1, \ldots, n.$$

Пример 2. При построении оптимальных итерационных параметров рассматривается следующая задача. Для многочлена

$$f_n(\lambda) = (1-\tau_1\lambda)(1-\tau_2\lambda)\ldots(1-\tau_n\lambda) \tag{25}$$

подобрать параметры $\tau_k>0$, $k=1, 2, \ldots, n$, так, чтобы минимизировать величину

$$\max_{0<\gamma_1\leqslant\lambda\leqslant\gamma_2} |f_n(\lambda)|.$$

Многочлен (25) удовлетворяет условию $f_n(0)=1$. Поэтому данная задача решается с помощью многочлена Чебышева (20). Корни многочлена (25)

$$\lambda_k = \tau_k^{-1}, \quad k=1, 2, \ldots, n,$$

должны совпадать с корнями многочлена

$$P_n(\lambda) = (-1)^n q_n \cos\left(n\arccos\frac{2\lambda-(\gamma_1+\gamma_2)}{\gamma_2-\gamma_1}\right), \tag{26}$$

где

$$q_n = \frac{2\rho_1^n}{1+\rho_1^{2n}}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2}. \tag{27}$$

Согласно (22) корни многочлена (26) расположены в точках

$$\widetilde{\lambda}_k = \frac{\gamma_1+\gamma_2}{2} + \frac{\gamma_2-\gamma_1}{2}\cos\frac{(2k-1)\pi}{2n}, \quad k=1, 2, \ldots, n. \tag{28}$$

Следовательно, если выбрать

$$\tau_k^{-1} = \frac{\gamma_1+\gamma_2}{2} + \frac{\gamma_2-\gamma_1}{2}\cos\frac{(2k-1)\pi}{2n}, \quad k=1, 2, \ldots, n, \tag{29}$$

то отклонение $f_n(\lambda)$ от нуля окажется минимальным и равным

$$\max_{\lambda\in[\gamma_1,\gamma_2]} |f_n(\lambda)| = q_n,$$

где q_n определено согласно (27). Здесь остается в силе замечание, сделанное в конце предыдущего примера. А именно, для оптимальности набора параметров $\{\tau_k\}_{k=1}^n$ не обязательно выбирать τ_k согласно (29), достаточно, чтобы множество $\{\tau_k^{-1}\}_{k=1}^n$ совпадало с множеством $\{\widetilde{\lambda}_k\}_{k=1}^n$ корней многочлена Чебышева (26).

§ 6. Итерационные методы с чебышевским набором параметров

1. Явный итерационный метод. Рассмотрим систему линейных уравнений

$$Ax=f \tag{1}$$

с симметричной положительно определенной матрицей A. Будем решать эту систему с помощью явного нестационарного итерационного метода

$$\frac{x_{k+1}-x_k}{\tau_{k+1}} + Ax_k = f, \quad k=0, 1, \ldots, \tag{2}$$

где x_0 задан.

Поставим задачу об оптимальном выборе итерационных параметров, т. е. о нахождении положительных чисел $\tau_1, \tau_2, \ldots, \tau_n$, для которых норма погрешности x_n-x на n-й итерации минимальна. В дальнейшем в этом параграфе будем использовать среднеквадратичную норму

$$\|x\| = \sqrt{(x, x)} = \sum_{j=1}^{m} |x^j|^2,$$

где x^j — j-я координата вектора x.

Точная формулировка и решение задачи оптимизации итерационного метода (2) содержатся в следующей теореме.

Теорема 1. *Пусть A — симметричная положительно определенная матрица, $\lambda_{\min}(A)>0$ и $\lambda_{\max}(A)>0$ — ее наименьшее и наибольшее собственные значения. Пусть задано число итераций n. Среди методов вида (2) наименьшую погрешность $\|x_n-x\|$ имеет метод, для которого*

$$\tau_k = \frac{\tau_0}{1+\rho_0 t_k}, \quad k=1, 2, \ldots, n, \tag{3}$$

где

$$\tau_0 = \frac{2}{\lambda_{\min}(A)+\lambda_{\max}(A)}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad \xi = \frac{\lambda_{\min}(A)}{\lambda_{\max}(A)},$$

$$t_k = \cos\frac{(2k-1)\pi}{2n}, \quad k=1, 2, \ldots, n. \tag{4}$$

Если выбрать τ_k *согласно* (3), (4), *то для погрешности будет справедлива оценка*

$$\|x_n - x\| \leqslant q_n \|x_0 - x\|, \tag{5}$$

где

$$q_n = \frac{2\rho_1^n}{1+\rho_1^{2n}}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \xi = \frac{\lambda_{\min}(A)}{\lambda_{\max}(A)}. \tag{6}$$

Доказательство. Для погрешности $z_k = x_k - x$ получаем уравнение

$$\frac{z_{k+1} - z_k}{\tau_{k+1}} + Az_k = 0, \quad k = 0, 1, \ldots, n-1, \quad z_0 = x_0 - x. \tag{7}$$

Из уравнения (7) получим, что

$$z_k = (E - \tau_k A)(E - \tau_{k-1}A) \ldots (E - \tau_1 A) z_0$$

для $k = 1, 2, \ldots, n$ и, в частности,

$$z_n = T_n z_0,$$

где

$$T_n = (E - \tau_n A)(E - \tau_{n-1}A) \ldots (E - \tau_1 A). \tag{8}$$

Поскольку T_n — симметричная матрица, ее норма совпадает с ее спектральным радиусом (см. § 6 гл. 1) и выполняется оценка

$$\|z_n\| \leqslant |\nu| \|z_0\|, \tag{9}$$

где ν — максимальное по модулю собственное значение матрицы T_n. Оценка (9) неулучшаема, т. е. найдется вектор z_0, для которого она выполняется со знаком равенства. Для доказательства теоремы остается подобрать параметры $\tau_1, \tau_2, \ldots, \tau_n$ так, чтобы минимизировать $|\nu|$.

Пусть λ_k, $k = 1, 2, \ldots, m$, — собственные числа матрицы A. Не ограничивая общности, можно считать, что

$$0 < \lambda_{\min}(A) = \lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_m = \lambda_{\max}(A). \tag{10}$$

Согласно (8) имеем

$$|\nu| = \max_{1 \leqslant k \leqslant m} |(1 - \tau_1\lambda_k)(1 - \tau_2\lambda_k) \ldots (1 - \tau_n\lambda_k)|. \tag{11}$$

Очевидно, что

$$|\nu| \leqslant \max_{\lambda_{\min}(A) \leqslant \lambda \leqslant \lambda_{\max}(A)} |f_n(\lambda)|,$$

где $f_n(\lambda) = (1-\tau_1\lambda)(1-\tau_2\lambda) \ldots (1-\tau_n\lambda)$.

Таким образом, приходим к задаче о нахождении

$$\min_{\tau_1, \tau_2, \ldots, \tau_n} \max_{\lambda_{\min}(A) \leqslant \lambda \leqslant \lambda_{\max}(A)} |f_n(\lambda)|,$$

уже рассмотренной в примере 2 из п. 4 § 5. Полученные в этом примере формулы (29) для параметров τ_k совпадают с формулами (3), (4), а величина отклонения при данных параметрах равна $|\nu| = q_n$, где q_n определяется согласно (6). Теорема 1 доказана.

Итерационный метод (2) с параметрами τ_k, определенными согласно (3), (4), называется *явным итерационным методом с чебышевским набором параметров.*

Замечание. В случае $n=1$ метод (2)—(4) совпадает с методом простой итерации

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

где

$$\tau=\tau_0=\frac{2}{\lambda_{\min}(A)+\lambda_{\max}(A)}.$$

Из теоремы 1 следует, что $\tau=\tau_0$ является оптимальным значением параметра τ в методе простой итерации. Оценка (5), (6) в случае $n=1$ принимает вид

$$\|x_1-x\|\leqslant\rho_0\|x_0-x\|.$$

Точно так же для любой итерации k справедлива оценка

$$\|x_{k+1}-x\|\leqslant\rho_0\|x_k-x\|,$$

откуда

$$\|x_k-x\|\leqslant\rho_0^k\|x_0-x\|.$$

Эта оценка погрешности метода простой итерации была получена другим способом в § 4 (оценка (13) из § 4).

Подсчитаем число итераций, достаточное для получения заданной точности ε при использовании явного метода с чебышевским набором параметров. Из оценки (5) получим, что

$$\|x_n-x\|<\varepsilon\|x_0-x\|,$$

если $q_n<\varepsilon$, где q_n определено согласно (6). Таким образом приходим к неравенству

$$\frac{1+\rho_1^{2n}}{\rho_1^n}>\frac{2}{\varepsilon},$$

решая которое относительно $z=\rho_1^{-n}>1$, получим

$$\frac{1}{\rho_1^n}>\frac{1+\sqrt{1-\varepsilon^2}}{\varepsilon}.$$

Последнее неравенство будет выполнено, если потребовать $1/\rho_1^n\geqslant 2/\varepsilon$, т. е.

$$n\geqslant n_0(\varepsilon)=\frac{\ln(2/\varepsilon)}{\ln(1/\rho_1)}. \tag{12}$$

В наиболее неблагоприятном случае, когда отношение $\xi=\lambda_{\min}(A)/\lambda_{\max}(A)$ мало, имеем

$$\ln\frac{1}{\rho_1}=\ln\left(\frac{1+\sqrt{\xi}}{1-\sqrt{\xi}}\right)\approx 2\sqrt{\xi}$$

и из (12) получим следующее приближенное выражение для числа итераций:

$$n_0(\varepsilon) = \frac{\ln(2/\varepsilon)}{2\sqrt{\xi}}. \qquad (13)$$

Таким образом, при малых ξ явный итерационный метод с чебышевским набором параметров требует для достижения заданной точности ε числа итераций $n_0(\xi) = O(1/\sqrt{\xi})$. Именно в этом состоит его преимущество перед методом простой итерации, для которого согласно (14) из § 4 имеем $n_0(\xi) = O(1/\xi)$.

2. Численная устойчивость итерационного метода с чебышевским набором параметров. Как уже отмечалось в примере 2 из п. 4 § 5 оценка погрешности (5) остается одной и той же при различном упорядочении набора итерационных параметров (3). Теоретически эти параметры можно использовать в любом порядке. Например, можно взять их в том порядке, как это указано в формуле (3). Можно использовать параметры в обратном порядке, т. е. положить

$$\tau_k = \frac{\tau_0}{1 + \rho_0 t_{n-k+1}}, \quad k = 1, 2, \ldots, n.$$

Однако при практическом применении метода было обнаружено, что порядок выбора параметров существенно влияет на численную устойчивость метода. Оказалось, что использование параметров в произвольном порядке может привести к недопустимо сильному возрастанию вычислительных погрешностей. Дело в том, что рассматриваемый метод, вообще говоря, не гарантирует монотонного убывания погрешности от итерации к итерации. Запишем уравнение для погрешности (7) в виде

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

Норма оператора перехода $E - \tau_{k+1}A$ данного итерационного метода может оказаться больше единицы для нескольких соседних итераций, что и приведет к возрастанию погрешности. Иногда вычислительная погрешность возрастает настолько сильно, что происходит переполнение арифметического устройства ЭВМ.

Здесь можно провести аналогию с вычислением произведения нескольких чисел. Рассмотрим следующий пример. Пусть на некоторой ЭВМ машинным нулем является число $M_0 = 10^{-p}$, а машинной бесконечностью — число $M_\infty = 10^p$, где $p > 0$. Попытаемся вычислить на этой ЭВМ произведение пяти чисел $10^{p/2}$, $10^{p/4}$, $10^{-p/2}$, $10^{3p/4}$, $10^{-3p/4}$. Это произведение равно $10^{p/4}$ и принадлежит допустимому интервалу чисел (M_0, M_∞).

Однако результат вычисления на ЭВМ будет зависеть от того, в каком порядке перемножаются данные числа. При перемножении в порядке убывания

$$10^{3p/4} \cdot 10^{p/2} \cdot 10^{p/4} \cdot 10^{-p/2} \cdot 10^{-3p/4}$$

уже выполнение первого умножения приводит к переполнению, так

как $10^{5p/4} > M_\infty$. После этого вычисления прекращаются, и мы просто не сможем вычислить все произведение. При перемножении в порядке возрастания

$$10^{-3p/4} \cdot 10^{-p/2} \cdot 10^{p/4} \cdot 10^{p/2} \cdot 10^{3p/4}$$

после первого умножения получаем число $10^{-5p/4} < M_0$, которое полагается равным нулю, следовательно, равным нулю оказывается и все произведение. Если же расположить сомножители в таком порядке:

$$10^{-3p/4} 10^{p/2} 10^{3p/4} 10^{-p/2} 10^{p/4},$$

то удается довести вычисления до конца и получить правильный результат.

В настоящее время известен алгоритм построения такого упорядоченного набора итерационных параметров (3), для которого итерационный метод (2) является устойчивым. Подробное изложение этого алгоритма можно найти в [32].

3. Неявный чебышевский итерационный метод. Скорость сходимости явного метода (2), (3) зависит, как мы видели, от отношения $\xi = \lambda_{\min}(A)/\lambda_{\max}(A)$ минимального собственного числа матрицы A к максимальному: чем больше ξ, тем выше скорость сходимости. Рассмотрим теперь неявный итерационный метод

$$B\frac{x_{k+1} - x_k}{\tau_{k+1}} + Ax_k = f, \quad k = 0, 1, \ldots, \ x_0 \text{ задан}, \tag{14}$$

с симметричной положительно определенной матрицей B и переменными параметрами τ_{k+1}. Как и в случае стационарных методов (см. § 4), скорость сходимости метода (14) будет определяться уже не отношением собственных чисел матрицы A, а отношением $\xi = \lambda_{\min}(B^{-1}A)/\lambda_{\max}(B^{-1}A)$ минимального и максимального собственных чисел обобщенной задачи

$$A\mu = \lambda B\mu. \tag{15}$$

При соответствующем выборе матрицы B это отношение будет больше чем $\lambda_{\min}(A)/\lambda_{\max}(A)$, а следовательно, итерационный метод (14) будет сходиться быстрее, чем явный метод (2).

Теория неявных итерационных методов легко сводится к теории явного метода. Для неявного чебышевского метода справедлива

Теорема 2. *Пусть A и B — симметричные положительно определенные матрицы, а $\lambda_{\min}(B^{-1}A)$, $\lambda_{\max}(B^{-1}A)$ — соответственно наименьшее и наибольшее собственные значения задачи* (15). *Пусть задано число итераций n. Метод* (14) *имеет минимальную погрешность $\|x_n - x\|_B$, если параметры τ_k определить согласно* (3), (4), *где*

$$\xi = \frac{\lambda_{\min}(B^{-1}A)}{\lambda_{\max}(B^{-1}A)}.$$

При этом справедлива оценка

$$\|x_n - x\|_B \leqslant q_n \|x_0 - x\|_B, \tag{16}$$

где

$$q_n = \frac{2\rho_1^n}{1+\rho_1^{2n}}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \xi = \frac{\lambda_{\min}(B^{-1}A)}{\lambda_{\max}(B^{-1}A)}. \tag{17}$$

Доказательство. Погрешность $z_k = x_k - x$ удовлетворяет однородному уравнению

$$B\frac{z_{k+1}-z_k}{\tau_{k+1}} + Az_k = 0, \quad k = 0, 1, \ldots, \quad z_0 = x_0 - x. \tag{18}$$

Умножим уравнение (18) на матрицу $B^{-1/2}$ и обозначим $v_k = B^{1/2}z_k$. Тогда получим уравнение

$$\frac{v_{k+1}-v_k}{\tau_{k+1}} + Cv_k = 0, \quad k = 0, 1, \ldots, \quad v_0 = B^{1/2}(x_0 - x), \tag{19}$$

где $C = B^{-1/2}AB^{-1/2}$. Уравнение (19) отличается только обозначениями от уравнения (7), которому удовлетворяет погрешность явного итерационного метода. Поэтому нам остается лишь проверить выполнение условий теоремы 1 по отношению к матрице C.

Матрица C является симметричной и положительно определенной, причем ее спектр совпадает со спектром обобщенной задачи на собственные значения (15). В частности, $\lambda_{\min}(B^{-1}A)$ является минимальным собственным числом $\lambda_{\min}(C)$ матрицы C, а $\lambda_{\max}(B^{-1}A)$ — максимальным. Следуя теореме 1, выберем параметры τ_k согласно (3), (4), где $\xi = \lambda_{\min}(C)/\lambda_{\max}(C)$. Тогда для решения уравнения (19) будет выполняться оценка

$$\|v_n\| \leqslant q_n \|v_0\|.$$

Подставляя сюда $v_k = B^{1/2}z_k$, $k = 0, n$, получим

$$\|z_n\|_B \leqslant q_n \|z_0\|_B,$$

т. е. придем к требуемой оценке (16). Теорема 2 доказана.

Замечание. При условиях теоремы 2 наряду с оценкой (16) справедлива и оценка

$$\|x_n - x\|_A \leqslant q_n \|x_0 - x\|_A.$$

Для доказательства достаточно переписать уравнение (18) в виде (19), где $v_k = A^{1/2}z_k$, $C = A^{1/2}B^{-1}A^{1/2}$, и повторить рассуждения теоремы 2.

Так же как и в случае явного метода, численная устойчивость неявного итерационного метода зависит от способа упорядочения итерационных параметров. Алгоритм построения устойчивого набора итерационных параметров тот же, что и для явного метода.

4. Случай, когда точные границы спектра неизвестны. В теоремах 1 и 2 фигурируют точные границы спектра матриц A и $B^{-1}A$ соответственно. Очень часто минимальные и максимальные собственные значения не известны точно, а известны лишь оценки для них. Например, если выполнены матричные неравенства

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B \tag{20}$$

с некоторыми константами $\gamma_2 > \gamma_1 > 0$, то можно утверждать, что

$$\lambda_{\min}(B^{-1}A) \geqslant \gamma_1, \quad \lambda_{\max}(B^{-1}A) \leqslant \gamma_2.$$

Приведем теорему о сходимости итерационного метода (14) при условиях (20).

Теорема 3. *Пусть A и B — симметричные положительно определенные матрицы, удовлетворяющие условию* (20). *Пусть задано число итераций n. Если параметры τ_k определить согласно* (3), (4), *где $\xi = \gamma_1/\gamma_2$, то для погрешности будет справедлива оценка*

$$\|x_n - x\|_B \leqslant q_n \|x_0 - x\|_B, \tag{21}$$

где

$$q_n = \frac{2\rho_1^n}{1+\rho_1^{2n}}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2}. \tag{22}$$

Доказательство. Как и при доказательстве теоремы 2, запишем уравнение для погрешности в виде (19). Из (19) получим

$$v_n = T_n v_0,$$

где $T_n = (E - \tau_n C)(E - \tau_{n-1}C) \ldots (E - \tau_1 C)$. Отсюда следует $\|v_n\| \leqslant \leqslant \|T_n\| \|v_0\|$.

Пусть c_j — j-е собственное число матрицы C, $j = 1, 2, \ldots, m$. Так как T_n — симметричная матрица, норма $\|T_n\|$ совпадает с максимумом из модулей ее собственных значений

$$\max_{1 \leqslant j \leqslant m} |(1 - \tau_n c_j)(1 - \tau_{n-1} c_j) \ldots (1 - \tau_1 c_j)|$$

и не превосходит величины

$$\max_{0 < \gamma_1 \leqslant c \leqslant \gamma_2} |(1 - \tau_n c)(1 - \tau_{n-1} c) \ldots (1 - \tau_1 c)|. \tag{23}$$

Выбирая τ_k, $k = 1, 2, \ldots, n+1$, согласно (3), (4) при $\xi = \gamma_1/\gamma_2$, мы минимизируем величину (23) и приходим к оценке $\|T_n\| \leqslant q_n$, где q_n определено согласно (22). Наконец, замечая, что $\|v_n\| = \|x_n - x\|_B$, приходим к требуемой оценке (21).

Замечание. Хотя теорема 3 и не гарантирует оптимальности итерационного метода, из оценки (21) следует сходимость метода, причем при малых ξ число итераций, достаточных для достижения заданной точности ε, оценивается как

$$n_0(\xi) = \frac{\ln(2/\varepsilon)}{2\sqrt{\xi}}.$$

§ 7. Итерационные методы вариационного типа *)

В предыдущих параграфах рассматривались такие итерационные методы решения системы

$$Ax = f, \tag{1}$$

*) При первом чтении этот параграф можно пропустить.

в которых для задания итерационных параметров требовалось знать границы γ_1 и γ_2 собственных значений матрицы A. Рассмотрим теперь итерационные методы вида

$$B\frac{x_{k+1}-x_k}{\tau_{k+1}}+Ax_k=f, \tag{2}$$

в которых параметры τ_{k+1} выбираются из условия минимума погрешности $\|x_{k+1}-x\|_D$ при заданной погрешности $\|x_k-x\|_D$. Здесь D — заданная симметричная положительно определенная матрица, $\|v\|_D=\sqrt{(Dv, v)}$. В зависимости от выбора матриц D и B получим различные итерационные методы. Скорость сходимости таких методов не выше, чем у чебышевского итерационного метода. Преимуществом их является то, что они не требуют знания границ спектра матрицы A.

1. Метод минимальных невязок. Рассмотрим систему (1) с симметричной положительно определенной матрицей A. Обозначим через

$$r_k=Ax_k-f \tag{3}$$

невязку, которая получается при подстановке приближенного значения x_k, полученного на k-й итерации, в уравнение (1). Заметим, что погрешность $z_k=x_k-x$ и невязка r_k связаны равенством $Az_k=r_k$.

Рассмотрим явный итерационный метод

$$\frac{x_{k+1}-x_k}{\tau_{k+1}}+Ax_k=f \tag{4}$$

и перепишем его в виде

$$x_{k+1}=x_k-\tau_{k+1}r_k. \tag{5}$$

Методом минимальных невязок называется итерационный метод (4), в котором параметр τ_{k+1} выбирается из условия минимума $\|r_{k+1}\|$ при заданной норме $\|r_k\|$. Получим явное выражение для итерационного параметра τ_{k+1}. Из (5) получаем

$$Ax_{k+1}=Ax_k-\tau_{k+1}Ar_k$$

и, следовательно,

$$r_{k+1}=r_k-\tau_{k+1}Ar_k, \tag{6}$$

т. е. невязка r_k удовлетворяет тому же уравнению, что и погрешность $z_k=x_k-x$.

Возводя обе части уравнения (6) скалярно в квадрат, получим

$$\|r_{k+1}\|^2=\|r_k\|^2-2\tau_{k+1}(r_k, Ar_k)+\tau_{k+1}^2\|Ar_k\|^2. \tag{7}$$

Отсюда видно, что $\|r_{k+1}\|$ достигает минимума, если

$$\tau_{k+1}=\frac{(Ar_k, r_k)}{\|Ar_k\|^2}. \tag{8}$$

Таким образом, в методе минимальных невязок переход от k-й итерации к $(k+1)$-й осуществляется следующим образом. По найденному значению x_k вычисляется вектор невязки $r_k=Ax_k-f$ и по

формуле (8) находится параметр τ_{k+1}. Затем по формуле (5) досчитывается вектор x_{k+1}.

Метод минимальных невязок (5), (8) сходится с той же скоростью, что и метод простой итерации с оптимальным параметром τ. Справедлива

Теорема 1. *Пусть A — симметричная положительно определенная матрица. Для погрешности метода минимальных невязок выполняется оценка*

$$\|A(x_n - x)\| \leqslant \rho_0^n \|A(x_0 - x)\|, \quad n = 0, 1, \ldots, \tag{9}$$

где

$$\rho_0 = \frac{1-\xi}{1+\xi}, \quad \xi = \frac{\lambda_{\min}(A)}{\lambda_{\max}(A)}. \tag{10}$$

Доказательство. Рассмотрим тождество (7). При заданном векторе r_k правая часть этого тождества достигает минимума, если τ_{k+1} выбрать согласно (8). При любом другом значении τ_{k+1} правая часть тождества (7) может только увеличиться. Поэтому, полагая в (7) $\tau_{k+1} = \tau_0$, где

$$\tau_0 = \frac{2}{\lambda_{\min}(A) + \lambda_{\max}(A)}, \tag{11}$$

получим неравенство

$$\|r_{k+1}\|^2 \leqslant \|(E - \tau_0 A) r_k\|^2$$

и, следовательно,

$$\|r_{k+1}\| \leqslant \|E - \tau_0 A\| \|r_k\|. \tag{12}$$

Согласно следствию 2 теоремы 1 из § 4 имеем $\|E - \tau_0 A\| = \rho_0$, поэтому при всех k справедливо неравенство

$$\|r_{k+1}\| \leqslant \rho_0 \|r_k\|, \tag{13}$$

или, что то же самое, неравенство

$$\|A(x_{k+1} - x)\| \leqslant \rho_0 \|A(x_k - x)\|.$$

Отсюда и следует оценка (9).

Замечание. Используя доказательство теоремы 1, можно получить полезное неравенство

$$(y, y)^2 \leqslant (Ay, y)(A^{-1}y, y) \leqslant \frac{1}{4}\left(\sqrt{\xi} + \frac{1}{\sqrt{\xi}}\right)^2 (y, y)^2, \tag{14}$$

справедливое для симметричных положительно определенных матриц A и любого вектора $y \neq 0$. Для доказательства (14) запишем тождество (7) при τ_{k+1}, определенном согласно (8). Тогда получим

$$\|r_{k+1}\|^2 = \|r_k\|^2 - \frac{(Ar_k, r_k)}{\|Ar_k\|^2}.$$

Учитывая неравенство (13), получаем

$$0 \leqslant \|r_k\|^2 - \frac{(Ar_k, r_k)^2}{\|Ar_k\|^2} \leqslant \rho_0^2 \|r_k\|^2$$

или

$$(Ar_k, r_k)^2 \leqslant (r_k, r_k)(Ar_k, r_k),$$

$$(Ar_k, r_k)^2 \geqslant (1 - \rho_0^2)(r_k, r_k)(Ar_k, Ar_k).$$

Сделав замену $A^{1/2}r_k=y$ и учитывая, что $1-\rho_0^2=4\xi/(1+\xi)^2$, получим соответственно неравенства

$$(y, y)^2 \leqslant (A^{-1}y, y)(Ay, y),$$

$$(y, y)^2 \geqslant \frac{4\xi}{(1+\xi)^2}(A^{-1}y, y)(Ay, y),$$

которые совпадают с (14). Обратно, если непосредственно доказать неравенства (14), то из них можно вывести утверждение теоремы 1.

2. Метод минимальных поправок. Запишем неявный итерационный метод (2) в виде

$$x_{k+1}=x_k-\tau B^{-1}r_k, \tag{15}$$

где $r_k=Ax_k-f$ — невязка. Вектор

$$w_k=B^{-1}r_k$$

называется *поправкой* на $(k+1)$-й итерации. Поправка w_k удовлетворяет тому же уравнению, что и погрешность $z_k=x_k-x$ неявного метода, т. е. уравнению

$$B\frac{w_{k+1}-w_k}{\tau_{k+1}}+Aw_k=0. \tag{16}$$

Будем предполагать, что B — симметричная положительно определенная матрица. *Методом минимальных поправок* называется неявный итерационный метод (2), в котором параметр τ_{k+1} выбирается из условия минимума нормы $\|w_{k+1}\|_B^2=(Bw_{k+1}, w_{k+1})^{1/2}$ при заданном векторе w_k. В случае $B=E$ метод минимальных поправок совпадает с методом минимальных невязок.

Найдем выражение для итерационного параметра τ_{k+1}. Перепишем (16) в виде

$$w_{k+1}=w_k-\tau_{k+1}B^{-1}Aw_k$$

и вычислим

$$\|w_{k+1}\|_B^2=\|w_k\|_B^2-2\tau_{k+1}(Aw_k, w_k)+\tau_{k+1}^2(B^{-1}Aw_k, Aw_k).$$

Отсюда следует, что $\|w_{k+1}\|_B^2$ будет минимальной, если положить

$$\tau_{k+1}=\frac{(Aw_k, w_k)}{(B^{-1}Aw_k, Aw_k)}. \tag{17}$$

Для реализации метода минимальных поправок требуется на каждой итерации решить систему уравнений $Bw_k=r_k$, откуда найдем поправку w_k, и решить систему уравнений $Bv_k=Aw_k$, откуда найдем вектор $v_k=B^{-1}Aw_k$, необходимый для вычисления параметра τ_{k+1}.

Скорость сходимости метода минимальных поправок определяется границами спектра обобщенной задачи на собственные значения

$$Ax=\lambda Bx. \tag{18}$$

Теорема 2. *Пусть A, B — симметричные положительно определенные матрицы и $\lambda_{\min}(B^{-1}A)$, $\lambda_{\max}(B^{-1}A)$ — наименьшее и наи-*

большее собственные значения задачи (18). *Для погрешности метода минимальных поправок выполняется оценка*

$$\|A(x_n - x)\|_{B^{-1}} \leqslant \rho_0^n \|A(x_0 - x)\|_{B^{-1}}, \quad n = 0, 1, \ldots \tag{19}$$

где

$$\rho_0 = \frac{1-\xi}{1+\xi}, \quad \xi = \frac{\lambda_{\min}(B^{-1}A)}{\lambda_{\max}(B^{-1}A)}.$$

Доказательство. Перепишем уравнение для поправки (16) в виде

$$\frac{v_{k+1} - v_k}{\tau_{k+1}} + Cv_k = 0, \tag{20}$$

где $v_k = B^{1/2}w_k$, $C = B^{-1/2}AB^{-1/2}$. Выражение (17) для итерационного параметра τ_{k+1} принимает вид

$$\tau_{k+1} = \frac{(Cv_k, v_k)}{\|Cv_k\|^2}. \tag{21}$$

Уравнения (20) и (21) — это те же самые уравнения (6), (8), которые возникают в методе минимальных невязок. Поэтому можно воспользоваться теоремой 1 и записать оценку (13), которая теперь примет вид

$$\|v_{k+1}\| \leqslant \rho_0 \|v_k\|, \tag{22}$$

где

$$\rho_0 = \frac{1-\xi}{1+\xi}, \quad \xi = \frac{\lambda_{\min}(C)}{\lambda_{\max}(C)}.$$

Заметим, что $\lambda_{\min}(C) = \lambda_{\min}(B^{-1}A)$ и $\lambda_{\max}(C) = \lambda_{\max}(B^{-1}A)$. Кроме того,

$$\|v_k\| = \|B^{1/2}w_k\| = \|B^{-1/2}r_k\| = \|r_k\|_{B^{-1}} = \|A(x_k - x)\|_{B^{-1}}.$$

Отсюда и следует оценка (19).

3. Метод скорейшего спуска. Рассмотрим явный метод (4) и выберем итерационный параметр τ_{k+1} из условия минимума $\|z_{k+1}\|_A$ при заданном векторе z_k, где $z_{k+1} = x_{k+1} - x$. Поскольку погрешность z_k удовлетворяет уравнению

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

имеем

$$\|z_{k+1}\|_A^2 = \|z_k\|_A^2 - 2\tau_{k+1}(Az_k, Az_k) + \tau_{k+1}^2(A^2z_k, Az_k).$$

Следовательно, $\|z_{k+1}\|_A^2$ будет минимальной, если положить

$$\tau_{k+1} = \frac{(Az_k, Az_k)}{(A^2z_k, Az_k)}.$$

Так как величина $z_k=x_k-x$ неизвестна (поскольку неизвестно точное решение x), надо учесть, что $Az_k=r_k=Ax_k-f$, и вычисление τ_{k+1} проводить по формуле

$$\tau_{k+1}=\frac{(r_k, r_k)}{(Ar_k, r_k)}. \tag{23}$$

Так же, как и в теореме 1, доказывается, что метод скорейшего спуска сходится с той же скоростью, что и метод простой итерации с оптимальным параметром $\tau=\tau_0$. Для погрешности метода скорейшего спуска справедлива оценка

$$\|x_n-x\|_A\leqslant\rho_0^n\|x_0-x\|_A, \qquad n=0, 1, \ldots, \tag{24}$$

где $\rho_0=\dfrac{1-\xi}{1+\xi}$, $\xi=\dfrac{\lambda_{\min}(A)}{\lambda_{\max}(A)}$.

Неявным методом скорейшего спуска называется метод (2), в котором параметр τ_{k+1} выбирается из условия минимума $\|z_{k+1}\|_A$. Так как погрешность $z_k=x_k-x$ удовлетворяет уравнению

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

получим

$$\|z_{k+1}\|_A^2=\|z_k\|_A^2-2\tau_{k+1}(Az_k, B^{-1}Az_k)+\tau_{k+1}^2(AB^{-1}Az_k, B^{-1}Az_k),$$

или

$$\|z_{k+1}\|_A^2=\|z_k\|_A^2-2\tau_{k+1}(r_k, w_k)+\tau_{k+1}^2(Aw_k, w_k).$$

Следовательно, $\|z_{k+1}\|_A^2$ будет минимальной, если положить

$$\tau_{k+1}=\frac{(r_k, w_k)}{(Aw_k, w_k)}. \tag{25}$$

При этом для неявного метода скорейшего спуска справедлива оценка (24), где

$$\xi=\frac{\lambda_{\min}(B^{-1}A)}{\lambda_{\max}(B^{-1}A)}.$$

4. Метод сопряженных градиентов. Метод сопряженных градиентов является двухшаговым итерационным методом, т. е. для нахождения новой итерации x_{n+1} используются две предыдущие итерации x_n и x_{n-1}. Следовательно, здесь возрастает требуемый объем памяти, нужно помнить не только вектор x_n, но и x_{n-1}. Применение двухшаговых методов оправдывается тем, что при правильном выборе итерационных параметров скорость сходимости будет выше, чем у одношаговых методов. Например, рассматриваемый далее метод сопряженных градиентов при любом начальном приближении сходится за конечное число итераций.

Пусть A — матрица системы (1) и B — симметричная положительно определенная матрица. Рассмотрим следующий класс не-

явных двухшаговых итерационных методов:

$$B\frac{(x_{k+1}-x_k)+(1-\alpha_{k+1})(x_k-x_{k-1})}{\tau_{k+1}\alpha_{k+1}}+Ax_k=f,\quad k=1,2,\ldots \qquad (26)$$

Здесь α_{k+1}, τ_{k+1} — итерационные параметры, которые будут определены далее. Для начала счета необходимо задать два начальных приближения x_0 и x_1. Начальное приближение x_0 будем задавать произвольно, а вектор x_1 вычислять по одношаговой формуле, которая получается из (26) при $k=0$, $\alpha_1=1$, т. е. по формуле

$$B\frac{x_1-x_0}{\tau_1}+Ax_0=f. \qquad (27)$$

Если параметры α_{k+1}, τ_{k+1} найдены, то новое приближение x_{k+1} выражается через два предыдущих значения x_k и x_{k-1} по формуле

$$x_{k+1}=\alpha_{k+1}x_k+(1-\alpha_{k+1})x_{k-1}-\tau_{k+1}\alpha_{k+1}w_k, \qquad (28)$$

где $w_k=B^{-1}r_k$, $r_k=Ax_k-f$.

5. Минимизация погрешности. Перейдем к вопросу о выборе итерационных параметров в методе (26). Для погрешности $z_k=x_k-x$ получим уравнения

$$z_{k+1}=\alpha_{k+1}(E-\tau_{k+1}B^{-1}A)z_k+(1-\alpha_{k+1})z_{k-1},\quad k=1,2,\ldots,$$
$$z_1=(E-\tau_1B^{-1}A)z_0.$$

Введем, как и ранее, вспомогательную функцию $v_k=A^{1/2}z_k$, для которой $\|v_k\|=\|z_k\|_A$. Функция v_k удовлетворяет уравнениям

$$v_{k+1}=\alpha_{k+1}(E-\tau_{k+1}C)v_k+(1-\alpha_{k+1})v_{k-1},\quad k=1,2,\ldots, \qquad (29)$$
$$v_1=(E-\tau_1C)v_0, \qquad (30)$$

где $C=A^{1/2}B^{-1}A^{1/2}$. Будем считать, что A и B — симметричные положительно определенные матрицы, удовлетворяющие неравенствам

$$\gamma_1B\leqslant A\leqslant\gamma_2B,\quad \gamma_2>\gamma_1>0. \qquad (31)$$

Тогда $C^*=C>0$, причем

$$\gamma_1E\leqslant C\leqslant\gamma_2E. \qquad (32)$$

Исключая последовательно из уравнений (29), (30) векторы $v_1, v_2, \ldots, v_{k-1}$, найдем, что

$$v_k=P_k(C)v_0, \qquad (33)$$

где $P_k(C)$ — многочлен степени k от оператора C, удовлетворяющий условию $P_k(0)=E$.

Поставим задачу выбрать итерационные параметры τ_k, α_k так, чтобы при любом $n=1, 2, \ldots$ была бы минимальной $\|v_n\|=\|z_n\|_A$. Обратим внимание на отличие от постановки задачи, возникающей при построении оптимального чебышевского набора итерационных параметров (см. § 6). Там при фиксированном n требовалось найти параметры, минимизирующие $\|z_n\|_A$. Теперь же требуется большее — минимизировать $\|z_n\|_A$ при каждом n.

Параметр τ_1 находится из условия минимума $\|v_1\|$ при заданном векторе v_0. Так же, как и в методе скорейшего спуска, получаем

$$\tau_1 = \frac{(Cv_0, v_0)}{\|Cv_0\|^2}. \tag{34}$$

Отметим, что при таком выборе τ_1 выполняется равенство $(Cv_1, v_0) = 0$, т. е. векторы v_1 и v_0 ортогональны в смысле скалярного произведения

$$(u, v)_C = (Cu, v).$$

Далее, рассмотрим погрешность

$$v_k = P_k(C)\, v_0,$$

возникающую на k-й итерации, и запишем многочлен $P_k(C)$ в виде

$$P_k(C) = E + \sum_{i=1}^{k} a_i^{(k)} C^i, \tag{35}$$

где $a_i^{(k)}$ — числовые коэффициенты, определяемые параметрами α_i, τ_i, $i = 1, 2, \ldots, k$. Тогда

$$v_k = v_0 + \sum_{i=1}^{k} a_i^{(k)} C^i v_0, \qquad k = 1, 2, \ldots \tag{36}$$

Найдем условия, которым должны удовлетворять коэффициенты $a_i^{(n)}$, минимизирующие $\|v_n\|^2$. Из (36) получим

$$\|v_n\|^2 = \sum_{i,j=1}^{n} a_i^{(n)} a_j^{(n)} (C^i v_0, C^j v_0) + 2 \sum_{j=1}^{n} a_j^{(n)} (v_0, C^j v_0) + \|v_0\|^2, \tag{37}$$

т. е. $\|v_n\|^2$ является многочленом второй степени по переменным $a_1^{(n)}, \ldots, a_n^{(n)}$. Приравнивая к нулю частные производные $\partial \|v_n\|^2 / \partial a_j^{(n)}$, $j = 1, 2, \ldots, n$, приходим к системе уравнений

$$\sum_{i=1}^{n} a_i^{(n)} (C^i v_0, C^j v_0) + (C^j v_0, v_0) = 0, \tag{38}$$

решение которой $a_i^{(n)}$, $i = 1, 2, \ldots, n$, и обращает в минимум $\|v_n\|^2$.

6. Выбор итерационных параметров в методе сопряженных градиентов. Целью дальнейших рассуждений является нахождение параметров α_k, τ_k, $k = 1, 2, \ldots, n$, для которых выполнены условия (38). Заметим прежде всего, что (38) можно записать в виде

$$(C^j v_0, v_n) = 0, \qquad j = 1, 2, \ldots, n. \tag{39}$$

Лемма 1. *Условия* (39) *эквивалентны условиям*

$$(Cv_j, v_n) = 0, \qquad j = 0, 1, \ldots, n-1. \tag{40}$$

Доказательство. Согласно (36) имеем

$$Cv_j = Cv_0 + \sum_{i=1}^{j} a_i^{(j)} C^{i+1} v_0,$$

поэтому

$$(Cv_j, v_n) = (Cv_0, v_n) + \sum_{i=1}^{j} a_i^{(j)} (C^{i+1}v_0, v_n) =$$

$$= (Cv_0, v_n) + \sum_{i=2}^{j+1} a_{i-1}^{(j)} (C^i v_0, v_n). \quad (41)$$

Пусть выполнены условия (39). Тогда, если $j+1 \leqslant n$ (т. е. $j \leqslant n-1$), то

$$(Cv_0, v_n) = 0, \quad (C^2 v_0, v_n) = 0, \ldots, (C^{j+1}v_0, v_n) = 0.$$

Поэтому из (41) при $j \leqslant n-1$ получим

$$(Cv_j, v_n) = 0.$$

Итак, условия (40) следуют из (39). Покажем, что верно и обратное, т. е. из (40) следует (39). Доказательство проведем индукцией по числу j. Условие (40) при $j=0$ совпадает с условием (39) при $j=1$. Предположим, что условия (39) выполнены для $j=1, 2, \ldots, k$, и покажем, что они выполнены и для $j=k+1$, где $k \leqslant n-1$.

Из (40) при $j=k$ получим, учитывая (36),

$$0 = (Cv_k, v_n) = \left(Cv_0 + \sum_{i=1}^{k} a_i^{(k)} C^{i+1} v_0, v_n \right) =$$

$$= (Cv_0, v_n) + \sum_{i=2}^{k+1} a_{i-1}^{(k)} (C^i v_0, v_n). \quad (42)$$

По предположению индукции условия (39) выполнены при $j=1, 2, \ldots, k$. Поэтому из (42) получим

$$a_k^{(k)} (C^{k+1}v_0, v_n) = 0.$$

Поскольку $a_k^{(k)} \neq 0$ (так как $P_k(C)$ — многочлен степени k), отсюда получаем $(C^{k+1}v_0, v_n) = 0$, т. е. условия (39) выполнены и при $j=k+1$.

Лемма 1 доказана. Она потребуется для построения оптимальных итерационных параметров в методе (26).

Заметим, что число n в лемме 1 предполагалось фиксированным, в то время как при постановке задачи оптимизации мы требовали, чтобы $\|v_n\|$ была минимальной при любом $n=1, 2, \ldots$ Поэтому оптимальные параметры надо отыскивать не из условий (40) при фиксированном n, а из условий

$$(Cv_j, v_n) = 0, \quad n = 1, 2, \ldots, \quad j = 0, 1, \ldots, n-1. \quad (43)$$

Если такие параметры будут найдены, то это будет означать, что построена ортогональная (в смысле скалярного произведения $(u, v)_C = (Cu, v)$) система векторов $v_0, v_1, \ldots, v_k, \ldots$ Поскольку пространство решений системы (1) имеет размерность m, постро-

енная ортогональная система будет содержать не более m векторов. Это означает, что начиная с некоторого n ($n \leqslant m$) погрешности v_n обратятся в нуль, т. е. метод сойдется за конечное число итераций.

Перейдем к построению итерационных параметров, для которых выполнены условия (43). Параметры α_1 и τ_1 найдены согласно (34):

$$\alpha_1 = 1, \quad \tau_1 = \frac{(Cv_0, v_0)}{\|Cv_0\|^2}. \tag{44}$$

Пусть параметры $\tau_1, \tau_2, \ldots, \tau_k, \alpha_1, \alpha_2, \ldots, \alpha_k$ уже выбраны оптимальным образом. Тогда согласно (43) выполняются условия

$$(Cv_j, v_i) = 0, \quad i = 1, 2, \ldots, k, \quad j = 0, 1, \ldots, i-1. \tag{45}$$

Построим оптимальные параметры τ_{k+1}, α_{k+1}. Согласно лемме 1 при $n = k+1$ должны выполняться условия

$$(Cv_j, v_{k+1}) = 0, \quad j = 0, 1, \ldots, k. \tag{46}$$

Часть из этих условий, а именно условия (46) при $j = 0, 1, \ldots, k-2$, следует из (45). Действительно, согласно (29) имеем

$$(v_{k+1}, Cv_j) = \alpha_{k+1}(v_k, Cv_j) - \alpha_{k+1}\tau_{k+1}(Cv_k, Cv_j) + (1 - \alpha_{k+1})(v_{k-1}, Cv_j).$$

Из (45) при $i = k$ и $i = k-1$ получим, соответственно,

$$(Cv_j, v_k) = 0, \quad j = 0, 1, \ldots, k-1,$$
$$(Cv_j, v_{k-1}) = 0, \quad j = 0, 1, \ldots, k-2.$$

Поэтому

$$(v_{k+1}, Cv_j) = -\alpha_{k+1}\tau_{k+1}(Cv_k, Cv_j) \tag{47}$$

при $j = 0, 1, \ldots, k-2$.

Покажем, что для этих же значений j выполняются равенства $(Cv_k, Cv_j) = 0$. Запишем уравнение (29) при $k = j$:

$$v_{j+1} = \alpha_{j+1}(E - \tau_{j+1}C)v_j + (1 - \alpha_{j+1})v_{j-1}$$

и найдем отсюда

$$Cv_j = \frac{1}{\tau_{j+1}} v_j - \frac{1}{\alpha_{j+1}\tau_{j+1}} [v_{j+1} - (1 - \alpha_{j+1}) v_{j-1}].$$

Умножая предыдущее соотношение скалярно на Cv_k и учитывая симметричность матрицы C, получим

$$(Cv_j, Cv_k) =$$
$$= \frac{1}{\tau_{j+1}}(v_j, Cv_k) - \frac{1}{\alpha_{j+1}\tau_{j+1}}[(v_{j+1}, Cv_k) - (1 - \alpha_{j+1})(v_{j-1}, Cv_k)] =$$
$$= \frac{1}{\tau_{j+1}}(Cv_j, v_k) - \frac{1}{\alpha_{j+1}\tau_{j+1}}[(Cv_{j+1}, v_k) - (1 - \alpha_{j+1})(Cv_{j-1}, v_k)].$$

Из (45) при $i = k$ имеем

$$(Cv_j, v_k) = 0, \quad j = 0, 1, \ldots, k-1,$$
$$(Cv_{j+1}, v_k) = 0, \quad j = 0, 1, \ldots, k-2,$$
$$(Cv_{j-1}, v_k) = 0, \quad j = 1, 2, \ldots, k.$$

Следовательно $(Cv_j, Cv_k) = 0$ при $j = 0, 1, \ldots, k-2$, и согласно (47) имеем $(v_{k+1}, Cv_j) = 0$, $j = 0, 1, \ldots, k-2$.

Итак, из всех условий (46) остаются лишь два:

$$(Cv_{k-1}, v_{k+1}) = 0, \tag{48}$$

$$(Cv_k, v_{k+1}) = 0. \tag{49}$$

Подставляя в (48) значение v_{k+1} из (29), получим

$$0 = \alpha_{k+1}(v_k, Cv_{k-1}) - \alpha_{k+1}\tau_{k+1}(Cv_k, Cv_{k-1}) + (1 - \alpha_{k+1})(v_{k-1}, Cv_{k-1}).$$

Согласно (45) при $i = k$, $j = k-1$ имеем $(Cv_{k-1}, v_k) = 0$, так что предыдущее уравнение принимает вид

$$-\alpha_{k+1}\tau_{k+1}(Cv_k, Cv_{k-1}) + (1-\alpha_{k+1})(Cv_{k-1}, v_{k-1}) = 0. \tag{50}$$

Далее, подставляя в (49) значение v_{k+1} из (29), получим

$$0 = \alpha_{k+1}(v_k, Cv_k) - \alpha_{k+1}\tau_{k+1}(Cv_k, Cv_k) + (1 - \alpha_{k+1})(v_{k-1}, Cv_k).$$

Последнее слагаемое в этом тождестве равно нулю, так как согласно (45) при $i = k$, $j = k-1$ имеем

$$(v_{k-1}, Cv_k) = (Cv_{k-1}, v_k) = 0.$$

Таким образом, приходим к тождеству

$$\alpha_{k+1}[(Cv_k, v_k) - \tau_{k+1}(Cv_k, Cv_k)] = 0,$$

из которого находим значение параметра τ_{k+1}:

$$\tau_{k+1} = \frac{(Cv_k, v_k)}{\|Cv_k\|^2}. \tag{51}$$

Обратимся теперь к уравнению (50) и исключим из него выражение (Cv_k, Cv_{k-1}). Из уравнения (29) получим

$$Cv_{k-1} = \frac{1}{\tau_k} v_{k-1} - \frac{1}{\alpha_k \tau_k}[v_k - (1 - \alpha_k) v_{k-2}], \tag{52}$$

следовательно,

$$(Cv_k, Cv_{k-1}) = \frac{1}{\tau_k}(Cv_k, v_{k-1}) - \frac{1}{\alpha_k \tau_k}[(Cv_k, v_k) - (1 - \alpha_k)(Cv_k, v_{k-2})].$$

Согласно (45) имеем

$$(Cv_k, v_{k-1}) = (v_k, Cv_{k-1}) = 0,$$
$$(Cv_k, v_{k-2}) = (v_k, Cv_{k-2}) = 0.$$

Поэтому

$$(Cv_k, Cv_{k-1}) = -\frac{1}{\alpha_k \tau_k}(Cv_k, v_k).$$

Подставляя это выражение в (50), получим

$$\frac{\alpha_{k+1}\tau_{k+1}}{\alpha_k \tau_k} \frac{(Cv_k, v_k)}{(Cv_{k-1}, v_{k-1})} + 1 - \alpha_{k+1} = 0.$$

Отсюда приходим к рекуррентной формуле для параметров α_{k+1}:

$$\alpha_{k+1}=\left[1-\frac{\tau_{k+1}}{\tau_k}\frac{1}{\alpha_k}\frac{(Cv_k, v_k)}{(Cv_{k-1}, v_{k-1})}\right]^{-1}, \quad k=1, 2, \ldots, \quad \alpha_1=1. \quad (53)$$

Формулами (51), (53) задаются выражения для итерационных параметров в методе сопряженных градиентов. Скалярные произведения, входящие в эти выражения, вычисляются в процессе итераций. Учитывая, что

$$v_k=A^{1/2}z_k, \quad C=A^{1/2}B^{-1}A^{1/2}, \quad z_k=x_k-x, \quad Az_k=Ax_k-f=r_k, \quad B^{-1}r_k=w_k,$$

получим $Cv_k=A^{1/2}w_k$, $(Cv_k, v_k)=(w_k, r_k)$, $(Cv_k, Cv_k)=(Aw_k, w_k)$.

Поэтому окончательно приходим к следующим формулам для определения итерационных параметров в методе сопряженных градиентов:

$$\tau_{k+1}=\frac{(w_k, r_k)}{(Aw_k, w_k)}, \quad k=0, 1, \ldots, \quad (54)$$

$$\alpha_{k+1}=\left[1-\frac{\tau_{k+1}}{\tau_k}\frac{1}{\alpha_k}\frac{(w_k, r_k)}{(Aw_{k-1}, w_{k-1})}\right]^{-1}, \quad k=1, 2, \ldots, \quad \alpha_1=1. \quad (55)$$

7. Оценка погрешности в методе сопряженных градиентов. Выше отмечалось, что в методе сопряженных градиентов точное решение системы уравнений (1) получается за конечное число итераций, равное порядку системы. Если порядок системы велик, то может оказаться полезной и оценка погрешности. Эта оценка не хуже, чем в одношаговом итерационном методе с чебышевским набором параметров. Действительно, из выражения для погрешности (33) получаем

$$\|v_n\|=\|x_n-x\|_A\leqslant\|P_n(C)\|\,\|x_0-x\|_A.$$

Поскольку $P_n(C)$ — многочлен степени n от оператора C, удовлетворяющий условию $P_n(0)=E$, выполняется оценка

$$\|P_n(C)\|\leqslant\|T_n(C)\|=\frac{2\rho_1^n}{1+\rho_1^{2n}}, \quad \rho_1=\frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \xi=\frac{\gamma_1}{\gamma_2},$$

где T_n — многочлен Чебышева, наименее уклоняющийся от нуля на $[\gamma_1, \gamma_2]$, $T_n(0)=1$.

Таким образом, для погрешности метода сопряженных градиентов справедлива оценка

$$\|x_n-x\|_A\leqslant q_n\|x_0-x\|_A,$$

где $q_n=2\rho_1^n/(1+\rho_1^{2n})$.

ГЛАВА 3

ИНТЕРПОЛИРОВАНИЕ И ПРИБЛИЖЕНИЕ ФУНКЦИЙ

В настоящей главе излагаются вычислительные аспекты некоторых задач теории приближения функций. Задача интерполирования состоит в том, чтобы по значениям функции $f(x)$ в нескольких точках отрезка восстановить ее значения в остальных точках этого отрезка. Разумеется, такая задача допускает сколь угодно много решений. Задача интерполирования возникает, например, в том случае, когда известны результаты измерения $y_k=f(x_k)$ некоторой физической величины $f(x)$ в точках x_k, $k=0, 1, \ldots, n$, и требуется определить ее значения в других точках. Интерполирование используется также при сгущении таблиц, когда вычисление значений $f(x)$ трудоемко. Иногда возникает необходимость приближенной замены или аппроксимации данной функции другими функциями, которые легче вычислить. В частности, рассматривается задача о наилучшем приближении в нормированном пространстве H, когда заданную функцию $f\in H$ требуется заменить линейной комбинацией φ заданных элементов из H так, чтобы отклонение $\|f-\varphi\|$ было минимальным. Результаты и методы теории интерполирования и приближения функций нашли широкое применение в численном анализе, например при выводе формул численного дифференцирования и интегрирования, при построении сеточных аналогов задач математической физики.

§ 1. Интерполирование алгебраическими многочленами

1. Интерполяционная формула Лангранжа. Пусть на отрезке $a\leqslant x\leqslant b$ заданы точки x_k, $k=0, 1, \ldots, n$ (*узлы интерполирования*), в которых известны значения функции $f(x)$. Задача *интерполирования алгебраическими многочленами* состоит в том, чтобы построить многочлен

$$L_n(x)=a_0+a_1x+\ldots+a_nx^n \tag{1}$$

степени n, значения которого в заданных точках x_k, $k=0, 1, \ldots, n$, совпадают со значениями функции $f(x)$ в этих точках.

Для любой непрерывной функции $f(x)$ сформулированная задача имеет единственное решение. Действительно, для отыскания коэффициентов $a_0, a_1, \ldots, a_n$ получаем систему линейных уравнений

$$a_0+a_1x_i+a_2x_i^2+\ldots+a_nx_i^n=f(x_i), \quad i=0, 1, \ldots, n, \tag{2}$$

определитель которой (определитель Вандермонда [12, с. 33]) отличен от нуля, если среди точек x_i, $i=0, 1, \ldots, n$, нет совпадающих.

Многочлен $L_n(x)$, удовлетворяющий условиям

$$L_n(x_i)=f(x_i), \quad i=0, 1, \ldots, n, \tag{3}$$

называется *интерполяционным многочленом* для функции $f(x)$, построенным по узлам $\{x_i\}_0^n$.

Решение системы (2) можно записать различным образом. Наиболее употребительна запись интерполяционного многочлена в форме Лагранжа и в форме Ньютона.

Интерполяционная формула Лагранжа позволяет представить многочлен $L_n(x)$ в виде линейной комбинации

$$L_n(x) = \sum_{k=0}^{n} c_k(x) f(x_k) \tag{4}$$

значений функции $f(x)$ в узлах интерполирования.

Найдем явное выражение для коэффициентов $c_k(x)$. Из условий интерполирования (3) получаем

$$\sum_{k=0}^{n} c_k(x_i) f(x_k) = f(x_i), \qquad i = 0, 1, \ldots, n.$$

Эти соотношения будут выполнены, если на функции $c_k(x)$ наложить условия

$$c_k(x_i) = \begin{cases} 0, & i \neq k, \\ 1, & i = k, \ i = 0, 1, \ldots, n, \end{cases}$$

которые означают, что каждая из функций $c_k(x)$, $k=0, 1, \ldots, n$, имеет не менее n нулей на $[a, b]$. Поскольку $L_n(x)$ — многочлен степени n, коэффициенты $c_k(x)$ естественно искать также в виде многочленов степени n, а именно в виде

$$c_k(x) = \lambda_k (x-x_0)(x-x_1)\ldots(x-x_{k-1})(x-x_{k+1})\ldots(x-x_n).$$

Из условия $c_k(x_k) = 1$ находим

$$\lambda_k^{-1} = (x_k - x_0)(x_k - x_1) \ldots (x_k - x_{k-1})(x_k - x_{k+1}) \ldots (x_k - x_n).$$

Таким образом, коэффициенты $c_k(x)$ интерполяционного многочлена (4) находятся по формулам

$$c_k(x) = \frac{\prod\limits_{j \neq k} (x - x_j)}{\prod\limits_{j \neq k} (x_k - x_j)}. \tag{5}$$

Часто коэффициенты $c_k(x)$ записывают в другом виде. Введем многочлен $\omega(x)$ степени $n+1$:

$$\omega(x) = (x-x_0)(x-x_1)\ldots(x-x_{k-1})(x-x_k)(x-x_{k+1})\ldots(x-x_n) \tag{6}$$

и вычислим его производную в точке x_k:

$$\omega'(x_k) = (x_k-x_0)\ldots(x_k-x_{k-1})(x_k-x_{k+1})\ldots(x_k-x_n).$$

Тогда получим, что

$$c_k(x) = \frac{\omega(x)}{(x - x_k)\,\omega'(x_k)}.$$

Итак, *интерполяционный многочлен Лагранжа* имеет вид

$$L_n(x) = \sum_{k=0}^{n} \frac{\omega(x)}{(x - x_k)\,\omega'(x_k)} f(x_k) \tag{7}$$

или, более подробно,

$$L_n(x) = \sum_{k=0}^{n} \frac{\prod\limits_{j \neq k} (x - x_j)}{\prod\limits_{j \neq k} (x_k - x_j)} f(x_k). \tag{8}$$

2. Интерполяционная формула Ньютона. Эта формула позволяет выразить интерполяционный многочлен $L_n(x)$ через значение $f(x)$ в одном из узлов и через разделенные разности функции $f(x)$, построенные по узлам $x_0, x_1, \ldots, x_n$. Она является разностным аналогом формулы Тейлора

$$f(x) = f(x_0) + (x - x_0) f'(x_0) + \frac{(x - x_0)^2}{2!} f''(x_0) + \ldots$$

Сначала приведем необходимые сведения о разделенных разностях. Пусть в узлах $x_k \in [a, b]$, $k = 0, 1, \ldots, n$, известны значения функции $f(x)$. Предположим, что среди точек x_k, $k = 0, 1, \ldots, n$, нет совпадающих. *Разделенными разностями первого порядка* называются отношения

$$f(x_i, x_j) = \frac{f(x_j) - f(x_i)}{x_j - x_i}, \qquad i, j = 0, 1, \ldots, n, \quad i \neq j.$$

Будем рассматривать разделенные разности, составленные по соседним узлам, т. е. выражения $f(x_0, x_1), f(x_1, x_2), \ldots, f(x_{n-1}, x_n)$.

По этим разделенным разностям первого порядка можно построить *разделенные разности второго порядка*:

$$f(x_0, x_1, x_2) = \frac{f(x_1, x_2) - f(x_0, x_1)}{x_2 - x_0},$$

$$f(x_1, x_2, x_3) = \frac{f(x_2, x_3) - f(x_1, x_2)}{x_3 - x_1}, \ldots,$$

$$f(x_{n-2}, x_{n-1}, x_n) = \frac{f(x_{n-1}, x_n) - f(x_{n-2}, x_{n-1})}{x_n - x_{n-2}}.$$

Аналогично определяются разделенные разности более высокого порядка. Например, если известны разности k-го порядка

$$f(x_j, x_{j+1}, \ldots, x_{j+k}), \quad f(x_{j+1}, x_{j+2}, \ldots, x_{j+k+1}),$$

то *разделенная разность* $(k+1)$*-го порядка* определяется как

$$f(x_j, x_{j+1}, \ldots, x_{j+k}, x_{j+k+1}) = \frac{f(x_{j+1}, x_{j+2}, \ldots, x_{j+k+1}) - f(x_j, x_{j+1}, \ldots, x_{j+k})}{x_{j+k+1} - x_j}.$$

При вычислении разделенных разностей принято записывать их в виде таблицы

x_0	$f(x_0)$				
		$f(x_0, x_1)$			
x_1	$f(x_1)$		$f(x_0, x_1, x_2)$		
		$f(x_1, x_2)$	·		
x_2	$f(x_2)$	·	·	·	$f(x_0, x_1, \ldots x_n)$
·	·	·	·		
·	·	·	$f(x_{n-2}, x_{n-1}, x_n)$		
·	·	$f(x_{n-1}, x_n)$			
x_n	$f(x_n)$				

Разделенная разность k-го порядка следующим образом выражается через значения функции $f(x)$ в узлах:

$$f(x_j, x_{j+1}, \ldots, x_{j+k}) = \sum_{i=j}^{j+k} \frac{f(x_i)}{\prod\limits_{\substack{l=j \\ l\neq i}}^{j+k} (x_i - x_l)}. \tag{9}$$

Эту формулу можно доказать методом индукции. Нам потребуется частный случай формулы (9):

$$f(x_0, x_1, \ldots, x_n) = \sum_{i=0}^{n} \frac{f(x_i)}{\prod\limits_{\substack{l=0, \\ l\neq i}}^{n} (x_i - x_l)} =$$

$$= \sum_{i=0}^{n} \frac{f(x_i)}{(x_i - x_0)(x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)}. \tag{10}$$

Интерполяционным многочленом Ньютона называется многочлен

$$P_n(x) = f(x_0) + (x - x_0) f(x_0, x_1) + (x - x_0)(x - x_1) f(x_0, x_1, x_2) + \ldots$$
$$\ldots + (x - x_0)(x - x_1) \ldots (x - x_{n-1}) f(x_0, x_1, \ldots, x_n). \tag{11}$$

Покажем, что многочлен $P_n(x)$ совпадает с многочленом Лагранжа (8). Рассмотрим наряду с $L_n(x)$ многочлены $L_0(x) = f(x_0)$,

$L_1(x), \ldots, L_{n-1}(x)$ и представим $L_n(x)$ в виде

$$L_n(x) = L_0(x) + \sum_{j=1}^{n} (L_j(x) - L_{j-1}(x)). \tag{12}$$

Из условий интерполяции (3) получаем, что

$$L_{j-1}(x_k) = L_j(x_k) = f(x_k)$$

при $k=0, 1, \ldots, j-1$ и $j=1, 2, \ldots, n$. Следовательно, разность $L_j(x) - L_{j-1}(x)$ представляет собой алгебраический многочлен степени j, который обращается в нуль в точках $x_0, x_1, \ldots, x_{j-1}$, т. е.

$$L_j(x) - L_{j-1}(x) = A_j(x-x_0)(x-x_1)\ldots(x-x_{j-1}), \tag{13}$$

где A_j — числовой коэффициент. Этот коэффициент находится из условия

$$L_j(x_j) - L_{j-1}(x_j) = A_j(x_j-x_0)(x_j-x_1)\ldots(x_j-x_{j-1}),$$

откуда, учитывая условие $L_j(x_j) = f(x_j)$, получаем

$$A_j = \frac{f(x_j) - L_{j-1}(x_j)}{(x_j-x_0)\ldots(x_j-x_{j-1})}. \tag{14}$$

Подставим в (14) вместо слагаемого $L_{j-1}(x_j)$ его значение, вычисленное согласно (8), т. е.

$$L_{j-1}(x_j) = \sum_{k=0}^{j-1} f(x_k) \frac{(x_j-x_0)(x_j-x_1)\ldots(x_j-x_{k-1})(x_j-x_{k+1})\ldots(x_j-x_{j-1})}{(x_k-x_0)(x_k-x_1)\ldots(x_k-x_{k-1})(x_k-x_{k+1})\ldots(x_k-x_{j-1})}.$$

Тогда получим

$$A_j = \frac{f(x_j)}{(x_j-x_0)\ldots(x_j-x_{j-1})} - \sum_{k=0}^{j-1} \frac{f(x_k)}{(x_j-x_k)} \cdot \frac{1}{(x_k-x_0)\ldots(x_k-x_{k-1})(x_k-x_{k+1})\ldots(x_k-x_{j-1})} = \sum_{k=0}^{j} \frac{f(x_k)}{(x_k-x_0)\ldots(x_k-x_{k-1})(x_k-x_{k+1})\ldots(x_k-x_{j-1})(x_k-x_j)}.$$

Сопоставляя это выражение с (10), видим, что A_j совпадает с разделенной разностью j-го порядка:

$$A_j = f(x_0, x_1, \ldots, x_j).$$

Отсюда и из (12), (13) приходим к интерполяционной формуле Ньютона

$$L_n(x) = f(x_0) + (x-x_0) f(x_0, x_1) + (x-x_0)(x-x_1) f(x_0, x_1, x_2) + \ldots + (x-x_0)(x-x_1)\ldots(x-x_{n-1}) f(x_0, x_1, \ldots, x_n). \tag{15}$$

Подчеркнем еще раз, что формулы (8) и (15) представляют собой различную запись одного и того же многочлена

$$a_0+a_1x+a_2x^2+\ldots+a_nx^n,$$

удовлетворяющего условиям интерполяции (2).

Интерполяционную формулу Ньютона удобнее применять в том случае, когда интерполируется одна и та же функция $f(x)$, но число узлов интерполяции постепенно увеличивается. Если узлы интерполяции фиксированы и интерполируется не одна, а несколько функций, то удобнее пользоваться формулой Лагранжа.

Замечание. При выводе формулы (15) не предполагалось, что узлы $x_0, x_1, \ldots, x_n$ расположены в каком-то определенном порядке. Поэтому роль точки x_0 в формуле (15) может играть любая из точек $x_0, x_1, \ldots, x_n$. Соответствующее множество интерполяционных формул можно получить из (15) перенумераций узлов. Например, тот же самый многочлен $L_n(x)$ можно представить в виде

$$\begin{aligned} L_n(x) = f(x_n) &+ (x-x_n)\,f(x_n, x_{n-1}) + \\ &+ (x-x_n)(x-x_{n-1})\,f(x_n, x_{n-1}, x_{n-2}) + \ldots \\ \ldots &+ (x-x_n)(x-x_{n-1})\ldots(x-x_1)\,f(x_n, x_{n-1}, \ldots, x_0). \end{aligned} \tag{16}$$

Если $x_0<x_1<x_2<\ldots<x_n$, то (15) называется *формулой интерполирования вперед*, а (16) — *формулой интерполирования назад.*

§ 2. Погрешность интерполирования

1. Остаточный член интерполяционной формулы. Заменяя функцию $f(x)$ интерполяционным многочленом $L_n(x)$, мы допускаем погрешность

$$r_n(x)=f(x)-L_n(x),$$

которая называется *погрешностью интерполирования* или, что то же самое, *остаточным членом интерполяционной формулы.* Ясно, что в узлах интерполирования эта погрешность равна нулю. Оценим погрешность в любой точке $x\in[a, b]$. Для этого рассмотрим вспомогательную функцию

$$g(s)=f(s)-L_n(s)-K\omega(s), \tag{1}$$

где $s\in[a, b]$, K — постоянная и

$$\omega(s)=(s-x_0)(s-x_1)\ldots(s-x_n). \tag{2}$$

Пусть требуется оценить $r_n(x)$ в заданной точке $x\in[a, b]$, не являющейся узлом интерполирования. Выберем постоянную K из условия $g(x)=0$. Для этого достаточно положить

$$K=\frac{f(x)-L_n(x)}{\omega(x)}.$$

Предположим, что $f(s)$ имеет $n+1$ непрерывную производную на отрезке $a\leqslant s\leqslant b$. Функция $g(s)$ имеет не менее $n+2$ нулей на этом отрезке, а именно в точках x, x_k, $k=0, 1, \ldots, n$. Поэтому производная $g'(s)$ имеет не менее чем $n+1$ нулей на $[a, b]$, $g''(s)$ —

не менее n нулей и т. д., функция $g^{(n+1)}(s)$ по крайней мере один раз обращается в нуль на $[a, b]$. Тем самым существует точка $\xi \in [a, b]$, в которой $g^{(n+1)}(\xi) = 0$.

Поскольку

$$g^{(n+1)}(s) = f^{(n+1)}(s) - (n+1)!K,$$

получаем

$$f^{(n+1)}(\xi) = \frac{f(x) - L_n(x)}{\omega(x)}(n+1)!$$

Таким образом доказано, что *погрешность интерполирования можно представить в виде*

$$f(x) - L_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega(x), \tag{3}$$

где $\xi \in [a, b]$ *и* $\omega(x)$ *— многочлен, определенный согласно* (2).

Отсюда следует оценка

$$|f(x) - L_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!}|\omega(x)|, \tag{4}$$

где $M_{n+1} = \sup\limits_{x \in [a,b]} |f^{(n+1)}(x)|$. В частности, если $f(x)$ — алгебраический многочлен степени n, то интерполирование, проведенное по любым точкам $x_0, x_1, \ldots, x_n$, осуществляется точно, т. е. $L_n(x) \equiv f(x)$.

Замечание. Наряду с интерполированием применяют и *экстраполирование*, т. е. вычисление значений функции $f(x)$ в точках $x \in [a, b]$ по приближенной формуле $f(x) \approx L_n(x)$, где $L_n(x)$ — интерполяционный многочлен. Однако погрешность экстраполирования обычно оказывается существенно большей, чем погрешность интерполирования. К этому выводу можно прийти, рассматривая поведение многочлена $\omega(x)$ внутри и вне отрезка $[a, b]$.

Поскольку многочлены Лагранжа и Ньютона отличаются только формой записи, представление погрешности в виде (3) справедливо как для формулы Лагранжа, так и для формулы Ньютона. Однако погрешность интерполирования можно представить и в другом виде. Для этого рассмотрим разделенную разность

$$f(x, x_0, x_1, \ldots, x_n) = \frac{f(x)}{(x-x_0)(x-x_1)\ldots(x-x_n)} +$$

$$+ \frac{f(x_0)}{(x_0-x)(x_0-x_1)\ldots(x_0-x_n)} + \ldots + \frac{f(x_n)}{(x_n-x)(x_n-x_0)\ldots(x_n-x_{n-1})},$$

имеющую порядок $n+1$. Отсюда найдем

$$f(x) = f(x_0)\frac{(x-x_1)(x-x_2)\ldots(x-x_n)}{(x_0-x_1)(x_0-x_2)\ldots(x_0-x_n)} + \ldots$$

$$\ldots + f(x_n)\frac{(x-x_0)(x-x_1)\ldots(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)\ldots(x_n-x_{n-1})} +$$

$$+ (x-x_0)(x-x_1)\ldots(x-x_n)f(x, x_0, x_1, \ldots, x_n) =$$

$$= L_n(x) + (x-x_0)(x-x_1)\ldots(x-x_n)f(x, x_0, \ldots, x_n).$$

Таким образом, погрешность интерполяционной формулы можно представить в виде

$$f(x)-L_n(x)=\omega(x)f(x, x_0, x_1, \ldots, x_n). \tag{5}$$

Сопоставляя (3) и (5), видим, что существует точка $\xi\in[a, b]$, для которой

$$f(x, x_0, x_1, \ldots, x_n)=\frac{f^{(n+1)}(\xi)}{(n+1)!}. \tag{6}$$

Формула (6) устанавливает связь между разделенной разностью порядка $n+1$ и $(n+1)$-й производной функции $f(x)$.

2. Оптимальный выбор узлов интерполирования. Величину $|\omega(x)|$, входящую в оценку (4), можно минимизировать за счет выбора узлов интерполирования. Задача состоит в том, чтобы подобрать узлы $x_k\in[a, b]$, $k=0, 1, \ldots, n$, так, чтобы минимизировать величину

$$\max_{x\in[a,b]}|(x-x_0)(x-x_1)\ldots(x-x_n)|.$$

Эта задача уже рассматривалась в примере 1 из § 5 гл. 2. Она решается, как мы знаем, с помощью многочлена Чебышева

$$T_{n+1}(x)=\frac{(b-a)^{n+1}}{2^{2n+1}}\cos\left((n+1)\arccos\frac{2x-(b+a)}{b-a}\right), \tag{7}$$

причем в качестве узлов интерполирования надо взять корни многочлена (7), т. е. точки

$$x_k=\frac{a+b}{2}+\frac{b-a}{2}\cos\frac{(2k+1)\pi}{2(n+1)}, \quad k=0, 1, \ldots, n. \tag{8}$$

При этом

$$\max_{x\in[a,b]}|\omega(x)|=\frac{(b-a)^{n+1}}{2^{2n+1}}$$

и оценка (4) примет вид

$$|f(x)-L_n(x)|\leqslant\frac{M_{n+1}}{(n+1)!}\frac{(b-a)^{n+1}}{2^{2n+1}}. \tag{9}$$

3. О сходимости интерполяционного процесса. Возникает вопрос, будет ли стремиться к нулю погрешность интерполирования $f(x)-L_n(x)$, если число узлов n неограниченно увеличивать. Ответ, вообще говоря, отрицательный.

Сформулируем определение сходимости интерполяционного процесса. Множество точек x_i, $i=0, 1, \ldots, n$, таких, что

$$a\leqslant x_0<x_1<\ldots<x_n\leqslant b$$

назовем *сеткой* на отрезке $[a, b]$ и обозначим через Ω_n. До сих пор предполагалось, что число узлов интерполяции фиксировано. Переходя к изучению сходимости интерполяционного процесса, необхо-

димо рассмотреть последовательность сеток с возрастающим числом узлов, а именно последовательность

$$\Omega_0 = \{x_0^{(0)}\},\ \Omega_1 = \{x_0^{(1)}, x_1^{(1)}\},\ \ldots,\ \Omega_n = \{x_0^{(n)}, x_1^{(n)}, \ldots, x_n^{(n)}\},\ \ldots$$

Пусть функция $f(x)$ определена и непрерывна на $[a, b]$. Тогда можно задать последовательность интерполяционных многочленов $L_n[f(x)]$, построенных для функции $f(x)$ по ее значениям в узлах сетки Ω_n.

Говорят, что интерполяционный процесс для функции $f(x)$ *сходится в точке* $x^* \in [a, b]$, если существует

$$\lim_{n\to\infty} L_n[f(x^*)] = f(x^*).$$

Кроме поточечной сходимости рассматривается сходимость в различных нормах. Например, *равномерная сходимость на отрезке* $[a, b]$ означает, что

$$\max_{x\in[a,b]} |f(x) - L_n[f(x)]| \to 0$$

при $n\to\infty$.

Свойство сходимости или расходимости интерполяционного процесса зависит как от выбора последовательности сеток, так и от гладкости функции $f(x)$.

Известны примеры несложных функций, для которых интерполяционный процесс расходится. Так, последовательность интерполяционных многочленов, построенных для непрерывной функции $f(x) = |x|$ по равноотстоящим узлам на отрезке $[-1, 1]$, не сходится к функции $|x|$ ни в одной точке отрезка $[-1, 1]$, кроме точек -1, 0, 1 (пример С. Н. Бернштейна, см. [24, с. 519]). На рис. 4 в качестве иллюстрации изображен график многочлена $L_9(x)$ при $0 \leqslant x \leqslant 1$, построенного для функции $|x|$ по равноотстоящим узлам на отрезке $[-1, 1]$.

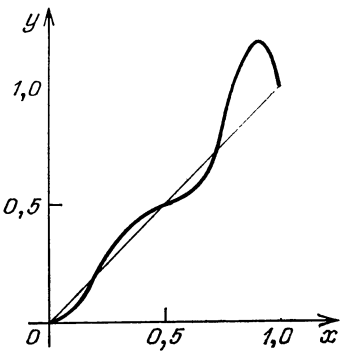


Рис. 4. График интерполяционного многочлена для функции $y = |x|$

Более общее утверждение содержится в теореме Фабера (доказательство см. в [24, с. 515]): *какова бы ни была последовательность сеток* Ω_n, *найдется непрерывная на* $[a, b]$ *функция* $f(x)$ *такая, что последовательность интерполяционных многочленов* $L_n[f(x)]$ *не сходится к* $f(x)$ *равномерно на отрезке* $[a, b]$.

Для заданной непрерывной функции $f(x)$ можно добиться сходимости за счет выбора расположения узлов интерполяции. Справедлива теорема Марцинкевича (см. [24, с. 519]): *если* $f(x)$ *непрерывна на* $[a, b]$, *то найдется такая последовательность*

сеток, для которой соответствующий интерполяционный процесс сходится равномерно на $[a, b]$.

Заметим, что построить такие сетки чрезвычайно сложно и, кроме того, для каждой функции требуется своя сетка. В практике вычислений избегают пользоваться интерполяционными многочленами высокой степени. Вместо этого применяется кусочнополиномиальная интерполяция, пример которой будет рассмотрен в § 4.

§ 3. Интерполирование с кратными узлами

1. Интерполяционный многочлен Эрмита. В предыдущих параграфах предполагалось, что в узлах интерполяции заданы только значения функции $f(x)$. Более общая постановка задачи интерполирования состоит в следующем.

В узлах $x_k \in [a, b]$, $k=0, 1, \ldots, m$, среди которых нет совпадающих узлов, заданы значения функции $f(x_k)$ и ее производных $f^{(i)}(x_k)$ до порядка N_k-1 включительно, $i=1, 2, \ldots, N_k-1$. Таким образом, в каждой точке x_k, $k=0, 1, \ldots, m$, известны

$$f(x_k), f'(x_k), \ldots, f^{(N_k-1)}(x_k)$$

и, следовательно, всего известно $N_0+N_1+\ldots+N_m$ величин. Требуется построить алгебраический многочлен $H_n(x)$ степени $n=N_0+N_1+\ldots+N_m-1$, для которого

$$H_n^{(i)}(x_k)=f^{(i)}(x_k), \quad k=0, 1, \ldots, m, \quad i=0, 1, \ldots, N_k-1. \qquad (1)$$

Многочлен $H_n(x)$, удовлетворяющий условиям (1), называется *интерполяционным многочленом Эрмита* для функции $f(x)$. Число N_k называется *кратностью узла* x_k.

Докажем, что интерполяционный многочлен Эрмита существует и единствен. Условия интерполяции (1) представляют собой систему линейных алгебраических уравнений относительно коэффициентов $a_0, a_1, \ldots, a_n$ многочлена

$$H_n(x)=a_0+a_1x+\ldots+a_nx^n.$$

Число уравнений этой системы равно числу неизвестных и равно $N_0+N_1+\ldots+N_m$. Поэтому достаточно показать, что однородная система

$$H_n^{(i)}(x_k)=0, \quad k=0, 1, \ldots, m, \quad i=0, 1, \ldots, N_k-1, \qquad (2)$$

имеет только тривиальное решение $a_0=a_1=\ldots=a_n=0$.

Группа условий (2) при фиксированном k и $i=0, 1, \ldots, N_k-1$ означает, что число x_k является корнем кратности N_k многочлена $H_n(x)$. Таким образом, многочлен $H_n(x)$ имеет всего с учетом кратности не менее $N_0+N_1+\ldots+N_m=n+1$ корня на $[a, b]$. Поскольку степень $H_n(x)$ равна n, этот многочлен тождественно равен нулю, следовательно, равны нулю его коэффициенты и однородная система уравнений (2) имеет единственное решение $a_0=a_1=\ldots=a_n=0$. Неоднородная система (1) однозначно разрешима при любых правых частях.

Поскольку значения $f^{(i)}(x_k)$, $k=0, 1, \ldots, m$, $i=0, 1, \ldots, N_k-1$, входят только в правую часть системы (1), коэффициенты a_j многочлена $H_n(x)$ выражаются линейно через значения $f^{(i)}(x_k)$, и этот многочлен можно представить в виде линейной комбинации

$$H_n(x)=\sum_{k=0}^{m}\sum_{i=0}^{N_k-1} c_{ki}(x)\,f^{(i)}(x_k),$$

где $c_{ki}(x)$ — многочлены степени n.

Ввиду громоздкости выражений для $c_{ki}(x)$ мы их не приводим.

Получим представление для погрешности интерполирования

$$r_n(x)=f(x)-H_n(x).$$

Для этого рассмотрим, как и в § 2, вспомогательную функцию

$$g(s)=f(s)-H_n(s)-K\omega(s), \tag{3}$$

где K — постоянная и

$$\omega(s)=(s-x_0)^{N_0}(s-x_1)^{N_1}\ldots(s-x_m)^{N_m}. \tag{4}$$

Постоянную K выберем так, чтобы в точке интерполирования x выполнялось условие $g(x)=0$, т. е. положим

$$K=\frac{f(x)-H_n(x)}{\omega(x)}.$$

Узлы x_k являются корнями кратности N_k функции $g(s)$, $k=1, 2, \ldots, m$. Кроме того, точка $x\in[a, b]$ является корнем $g(s)$. Таким образом, функция $g(s)$ имеет с учетом кратности $N_0+N_1+\ldots+N_m+1=n+2$ корня на отрезке $[a, b]$. По теореме Ролля производная $g'(s)$ имеет по крайней мере один нуль между двумя соседними корнями функции $g(s)$. Следовательно, $g'(s)$ имеет не менее $m+1$ корня на $[a, b]$ в точках, не совпадающих ни с одной из точек $x_0, x_1, \ldots, x_m, x$. Кроме того, $g'(s)$ имеет в точке x_k корень кратности N_k-1, $k=0, 1, \ldots, m$. Таким образом, $g'(s)$ имеет с учетом кратности не менее

$$(N_0-1)+\ldots+(N_m-1)+(m+1)=N_0+N_1+\ldots+N_m=n+1$$

корней на $[a, b]$. Аналогично $g''(s)$ имеет не менее n корней и т. д. Производная $g^{(n+1)}(s)$ по крайней мере один раз обращается в нуль на $[a, b]$, т. е. существует точка $\xi\in[a, b]$, в которой $g^{(n+1)}(\xi)=0$. Из (3) имеем

$$g^{(n+1)}(s)=f^{(n+1)}(s)-K\omega^{(n+1)}(s).$$

Так как $\omega(s)$ — многочлен степени $n+1$ со старшим коэффициентом 1, имеем $\omega^{(n+1)}(s)=(n+1)!$ Поэтому из условия $g^{(n+1)}(\xi)=0$ получаем, учитывая выражение для K, следующее представление для погрешности интерполирования:

$$f(x)-H_n(x)=\frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)^{N_0}(x-x_1)^{N_1}\ldots(x-x_m)^{N_m}. \tag{5}$$

2. **Пример.** Пусть $x_0<x_1<x_2$ — точки, в которых заданы значения $f(x_0)=f_0$, $f(x_1)=f_1$, $f'(x_1)=f_1'$, $f(x_2)=f_2$. Требуется построить многочлен третьей степени $H_3(x)$ такой, что

$$H_3(x_0)=f_0, \quad H_3(x_1)=f_1, \quad H_3'(x_1)=f_1', \quad H_3(x_2)=f_2. \tag{6}$$

Будем искать его в виде

$$H_3(x)=c_0(x)f_0+c_1(x)f_1+c_2(x)f_2+b_1(x)f_1',$$

где $c_0(x)$, $c_1(x)$, $c_2(x)$, $b_1(x)$ — многочлены третьей степени. Ясно, что $H_3(x)$ будет искомым интерполяционным многочленом, если потребовать

$$\begin{array}{llll} c_0(x_0)=1, & c_1(x_0)=0, & c_2(x_0)=0, & b_1(x_0)=0, \\ c_0(x_1)=0, & c_1(x_1)=1, & c_2(x_1)=0, & b_1(x_1)=0, \\ c_0(x_2)=0, & c_1(x_2)=0, & c_2(x_2)=1, & b_1(x_2)=0, \\ c_0'(x_1)=0, & c_1'(x_1)=0, & c_2'(x_1)=0, & b_1'(x_1)=1. \end{array}$$

Найдем многочлены третьей степени, удовлетворяющие перечисленным требованиям. Поскольку многочлен $c_0(x)$ имеет кратный корень в точке x_1 и простой корень в точке x_2, его можно искать в виде

$$c_0(x)=K(x-x_1)^2(x-x_2).$$

Из условия $c_0(x_0)=1$ находим

$$K=\frac{1}{(x_0-x_1)^2(x_0-x_2)}.$$

Таким образом,

$$c_0(x)=\frac{(x-x_1)^2(x-x_2)}{(x_0-x_1)^2(x_0-x_2)}. \tag{7}$$

Аналогично получаем

$$c_2(x)=\frac{(x-x_0)(x-x_1)^2}{(x_2-x_0)(x_2-x_1)^2}, \tag{8}$$

$$b_1(x)=\frac{(x-x_0)(x-x_1)(x-x_2)}{(x_1-x_2)(x_1-x_0)}. \tag{9}$$

Далее, многочлен $c_1(x)$ будем искать в виде

$$c_1(x)=(x-x_0)(x-x_2)(\alpha x+\beta),$$

где α и β — постоянные, подлежащие определению. Из условия $c_1(x_1)=1$ находим

$$\alpha x_1+\beta=\frac{1}{(x_1-x_0)(x_1-x_2)}. \tag{10}$$

Условие $c_1'(x_1)=0$ приводит к уравнению

$$(x_1-x_0)(x_1-x_2)\alpha+(\alpha x_1+\beta)(2x_1-x_0-x_2)=0. \tag{11}$$

Из уравнений (10) и (11) находим

$$\alpha = -\frac{2x_1 - x_0 - x_2}{(x_1 - x_0)^2 (x_1 - x_2)^2},$$

$$\beta = \frac{1}{(x_1 - x_0)(x_1 - x_2)}\left(1 + \frac{(2x_1 - x_0 - x_2)\, x_1}{(x_1 - x_0)(x_1 - x_2)}\right).$$

Таким образом,

$$c_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)}\left(1 - \frac{(x - x_1)(2x_1 - x_0 - x_2)}{(x_1 - x_0)(x_1 - x_2)}\right). \qquad (12)$$

Искомый интерполяционный многочлен $H_3(x)$ имеет вид

$$H_3(x) = \frac{(x - x_1)^2 (x - x_2)}{(x_0 - x_1)^2 (x_0 - x_2)} f(x_0) +$$
$$+ \left(1 - \frac{(x - x_1)(2x_1 - x_0 - x_2)}{(x_1 - x_0)(x_1 - x_2)}\right) \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} f(x_1) +$$
$$+ \frac{(x - x_0)(x - x_1)^2}{(x_2 - x_0)(x_2 - x_1)^2} f(x_2) + \frac{(x - x_0)(x - x_1)(x - x_2)}{(x_1 - x_2)(x_1 - x_0)} f'(x_1). \qquad (13)$$

Согласно (5) погрешность интерполирования в случае многочлена (13) можно записать в виде

$$f(x) - H_3(x) = \frac{f^{IV}(\xi)}{24}(x - x_0)(x - x_1)^2 (x - x_2), \qquad (14)$$

где $\xi \in (x_0, x_2)$.

Интерполяционный многочлен Эрмита можно построить путем предельного перехода в многочленах Лагранжа и Ньютона. Поясним это на том же примере. Наряду с узлами x_0, x_1, x_2 введем узел x_3 (отличный от x_0, x_1, x_2) и построим по узлам x_0, x_1, x_2, x_3 интерполяционный многочлен Лагранжа

$$L_3(x) = \frac{(x - x_0)(x - x_1)(x - x_2)}{(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)} f(x_3) +$$
$$+ \frac{(x - x_0)(x - x_2)(x - x_3)}{(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)} f(x_1) + \frac{(x - x_1)(x - x_2)(x - x_3)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)} f(x_0) +$$
$$+ \frac{(x - x_0)(x - x_1)(x - x_3)}{(x_2 - x_0)(x_2 - x_1)(x_2 - x_3)} f(x_2). \qquad (15)$$

Получим многочлен (13) путем предельного перехода в (15). Зафиксируем точки x, x_0, x_1, x_2 и устремим x_3 к x_1. Тогда последние два слагаемые перейдут в пределе в выражение

$$\frac{(x - x_1)^2 (x - x_2)}{(x_0 - x_1)^2 (x_0 - x_2)} f(x_0) + \frac{(x - x_0)(x - x_1)^2}{(x_2 - x_0)(x_2 - x_1)^2} f(x_2). \qquad (16)$$

Первые два слагаемые в (15) объединим следующим образом:

$$\frac{(x - x_0)(x - x_2)}{x_3 - x_1}\left(\frac{(x - x_1) f(x_3)}{(x_3 - x_0)(x_3 - x_2)} - \frac{(x - x_3) f(x_1)}{(x_1 - x_0)(x_1 - x_2)}\right).$$

Поскольку при указанном предельном переходе величины $x_3 - x_1$ и

$$\frac{(x - x_1) f(x_3)}{(x_3 - x_0)(x_3 - x_2)} - \frac{(x - x_3) f(x_1)}{(x_1 - x_0)(x_1 - x_2)} \qquad (17)$$

стремятся к нулю, можно раскрыть неопределенность вида 0/0, воспользовав-

шись правилом Лопиталя. Дифференцируя выражение (17) по x_3, получим функцию

$$\alpha(x_3) = \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)} - \frac{(x - x_1)(2x_3 - x_0 - x_2)}{(x_3 - x_0)^2 (x_3 - x_2)^2} f(x_3) + \\ + \frac{(x - x_1)(x_3 - x_0)(x_3 - x_2)}{(x_3 - x_0)^2 (x_3 - x_2)^2} f'(x_3),$$

так что

$$\lim_{x_3 \to x_1} \alpha(x_3) = \frac{1}{(x_1 - x_0)(x_1 - x_2)} \left[1 - \frac{(x - x_1)(2x_1 - x_0 - x_2)}{(x_1 - x_0)(x_1 - x_2)}\right] f(x_1) + \\ + \frac{(x - x_1) f'(x_1)}{(x_1 - x_0)(x_1 - x_2)}.$$

Итак, первые два слагаемые в (15) переходят в пределе в выражение

$$\left[1 - \frac{(x - x_1)(2x_1 - x_0 - x_2)}{(x_1 - x_0)(x_1 - x_2)}\right] \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} f(x_1) + \\ + \frac{(x - x_0)(x - x_1)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} f'(x_1).$$

Отсюда и из (16) получаем многочлен (10).

§ 4. Интерполирование сплайнами

Интерполирование многочленом Лагранжа или Ньютона на всем отрезке $[a, b]$ с использованием большого числа узлов интерполяции часто приводит к плохому приближению, что объясняется сильным накоплением погрешностей в процессе вычислений. Кроме того, из-за расходимости процесса интерполяции увеличение числа узлов не обязано приводить к повышению точности. Для того чтобы избежать больших погрешностей, весь отрезок $[a, b]$ разбивают на частичные отрезки и на каждом из частичных отрезков приближенно заменяют функцию $f(x)$ многочленом невысокой степени (так называемая *кусочно-полиномиальная интерполяция*).

Одним из способов интерполирования на всем отрезке является интерполирование с помощью сплайн-функций. Сплайн-функцией или сплайном называют кусочно-полиномиальную функцию, определенную на отрезке $[a, b]$ и имеющую на этом отрезке некоторое число непрерывных производных.

Слово «сплайн» (английское spline) означает гибкую линейку, используемую для проведения гладких кривых через заданные точки плоскости. Мы не будем придавать слову «сплайн» какого-либо определенного технического смысла.

Преимуществом сплайнов перед обычной интерполяцией является, во-первых, их сходимость и, во-вторых, устойчивость процесса вычислений.

В этом параграфе будет рассмотрен частный, но распространенный в вычислительной практике случай, когда сплайн определяется с помощью многочленов третьей степени (кубический сплайн).

1. Построение кубического сплайна. Пусть на $[a, b]$ задана непрерывная функция $f(x)$. Введем сетку

$$a=x_0<x_1<\ldots<x_{N-1}<x_N=b$$

и обозначим $f_i=f(x_i)$, $i=0, 1, \ldots, N$.

Сплайном, соответствующим данной функции $f(x)$ и данным узлам $\{x_i\}_{i=0}^N$, называется функция $s(x)$, удовлетворяющая следующим условиям:

а) на каждом сегменте $[x_{i-1}, x_i]$, $i=1, 2, \ldots, N$, функция $s(x)$ является многочленом третьей степени;

б) функция $s(x)$, а также ее первая и вторая производные непрерывны на $[a, b]$;

в) $s(x_i)=f(x_i)$, $i=0, 1, \ldots, N$.

Последнее условие называется *условием интерполирования*, а сплайн, определяемый условиями а)—в), называется также *интерполяционным кубическим сплайном.*

Докажем существование и единственность сплайна, определяемого перечисленными условиями. Приведенное ниже доказательство содержит также способ построения сплайна.

На каждом из отрезков $[x_{i-1}, x_i]$, $i=1, 2, \ldots, N$, будем искать функцию $s(x)=s_i(x)$ в виде многочлена третьей степени

$$s_i(x)=a_i+b_i(x-x_i)+\frac{c_i}{2}(x-x_i)^2+\frac{d_i}{6}(x-x_i)^3, \qquad (1)$$
$$x_{i-1}\leqslant x\leqslant x_i, \quad i=1, 2, \ldots, N,$$

где a_i, b_i, c_i, d_i— коэффициенты, подлежащие определению. Поясним смысл введенных коэффициентов. Имеем

$$s_i'(x)=b_i+c_i(x-x_i)+\frac{d_i}{2}(x-x_i)^2,$$
$$s_i''(x)=c_i+d_i(x-x_i),\ s_i'''(x)=d_i,$$

поэтому

$$a_i=s_i(x_i),\quad b_i=s_i'(x_i),\quad c_i=s_i''(x_i),\quad d_i=s_i'''(x_i).$$

Из условий интерполирования $s(x_i)=f(x_i)$, $i=1, 2, \ldots, N$, получаем, что

$$a_i=f(x_i), \quad i=1, 2, \ldots, N.$$

Доопределим, кроме того, $a_0=f(x_0)$.

Далее, требование непрерывности функции $s(x)$ приводит к условиям

$$s_i(x_i)=s_{i+1}(x_i), \quad i=1, 2, \ldots, N-1.$$

Отсюда, учитывая выражения для функций $s_i(x)$, получаем при $i=0, 1, \ldots, N-1$ уравнения

$$a_i=a_{i+1}+b_{i+1}(x_i-x_{i+1})+\frac{c_{i+1}}{2}(x_i-x_{i+1})^2+\frac{d_{i+1}}{6}(x_i-x_{i+1})^3.$$

Обозначая $h_i=x_i-x_{i-1}$, перепишем эти уравнения в виде

$$h_ib_i-\frac{h_i^2}{2}c_i+\frac{h_i^3}{6}d_i=f_i-f_{i-1},\quad i=1,2,\ldots,N. \tag{2}$$

Условия непрерывности первой производной

$$s_i'(x_i)=s_{i+1}'(x_i),\quad i=1,2,\ldots,N-1,$$

приводят к уравнениям

$$c_ih_i-\frac{d_i}{2}h_i^2=b_i-b_{i-1},\quad i=2,3,\ldots,N. \tag{3}$$

Из условия непрерывности второй производной получаем уравнения

$$d_ih_i=c_i-c_{i-1},\ i=2,3,\ldots,N. \tag{4}$$

Объединяя (2) – (4), получим систему $3N-2$ уравнений относительно $3N$ неизвестных b_i, c_i, d_i, $i=1,2,\ldots,N$.

Два недостающих уравнения получают, задавая те или иные граничные условия для $s(x)$. Предположим, например, что функция $f(x)$ удовлетворяет условиям $f''(a)=f''(b)=0$. Тогда естественно требовать, чтобы $s''(a)=s''(b)=0$. Отсюда получаем $s_1''(x_0)=0$, $s_N''(x_N)=0$, т. е. $c_1-d_1h_1=0$, $c_N=0$.

Заметим, что условие $c_1-d_1h_1=0$ совпадает с уравнением (4) при $i=1$, если положить $c_0=0$. Таким образом, приходим к замкнутой системе уравнений для определения коэффициентов кубического сплайна:

$$h_id_i=c_i-c_{i-1},\quad i=1,2,\ldots,N,\quad c_0=c_N=0, \tag{5}$$

$$h_ic_i-\frac{h_i^2}{2}d_i=b_i-b_{i-1},\quad i=2,3,\ldots,N, \tag{6}$$

$$h_ib_i-\frac{h_i^2}{2}c_i+\frac{h_i^3}{6}d_i=f_i-f_{i-1},\quad i=1,2,\ldots,N. \tag{7}$$

Убедимся в том, что эта система имеет единственное решение. Исключим из (5)—(7) переменные b_i, d_i, $i=1,2,\ldots,N-1$, и получим систему, содержащую только c_i, $i=1,2,\ldots,N-1$. Для этого рассмотрим два соседних уравнения (7):

$$b_i=\frac{h_i}{2}c_i-\frac{h_i^2}{6}d_i+\frac{f_i-f_{i-1}}{h_i},$$

$$b_{i-1}=\frac{h_{i-1}}{2}c_{i-1}-\frac{h_{i-1}^2}{6}d_{i-1}+\frac{f_{i-1}-f_{i-2}}{h_{i-1}}$$

и вычтем второе уравнение из первого. Тогда получим

$$b_i-b_{i-1}=$$
$$=\frac{1}{2}(h_ic_i-h_{i-1}c_{i-1})-\frac{1}{6}(h_i^2d_i-h_{i-1}^2d_{i-1})+\frac{f_i-f_{i-1}}{h_i}-\frac{f_{i-1}-f_{i-2}}{h_{i-1}}.$$

Подставляя найденное выражение для $b_i - b_{i-1}$ в правую часть уравнения (6), получим

$$h_i c_i + h_{i-1} c_{i-1} - \frac{h_{i-1}^2}{3} d_{i-1} - \frac{2h_i^2}{3} d_i = 2\left(\frac{f_i - f_{i-1}}{h_i} - \frac{f_{i-1} - f_{i-2}}{h_{i-1}}\right). \quad (8)$$

Далее, из уравнения (5) получаем

$$h_i^2 d_i = h_i (c_i - c_{i-1}), \quad h_{i-1}^2 d_{i-1} = h_{i-1} (c_{i-1} - c_{i-2})$$

и, подставляя эти выражения в (8), приходим к уравнению

$$h_{i-1} c_{i-2} + 2(h_{i-1} + h_i) c_{i-1} + h_i c_i = 6\left(\frac{f_i - f_{i-1}}{h_i} - \frac{f_{i-1} - f_{i-2}}{h_{i-1}}\right).$$

Окончательно для определения коэффициентов c_i получаем систему уравнений

$$h_i c_{i-1} + 2(h_i + h_{i+1}) c_i + h_{i+1} c_{i+1} = 6\left(\frac{f_{i+1} - f_i}{h_{i+1}} - \frac{f_i - f_{i-1}}{h_i}\right), \quad (9)$$

$$i = 1, 2, \ldots, N-1, \quad c_0 = c_N = 0.$$

В силу диагонального преобладания система (9) имеет единственное решение. Так как матрица системы трехдиагональная, решение легко найти методом прогонки, которая в данном случае устойчива (см. п. 7 § 4 ч. I). По найденным коэффициентам c_i коэффициенты b_i и d_i определяются с помощью явных формул

$$d_i = \frac{c_i - c_{i-1}}{h_i}, \quad b_i = \frac{h_i}{2} c_i - \frac{h_i^2}{6} d_i + \frac{f_i - f_{i-1}}{h_i}, \quad (10)$$

$$i = 1, 2, \ldots, N.$$

Таким образом, доказано, что существует единственный кубический сплайн, определяемый условиями а) — в) и граничными условиями $s''(a) = s''(b) = 0$. Заметим, что можно рассматривать и другие граничные условия.

2. Сходимость процесса интерполирования кубическими сплайнами *). Покажем, что интерполирование кубическими сплайнами является сходящимся процессом, т. е. при неограниченном увеличении числа узлов N соответствующая последовательность сплайн-функций сходится к интерполируемой функции $f(x)$. Оценки погрешности интерполяции $r(x) = f(x) - s(x)$ зависят от выбора сеток и от гладкости $f(x)$. Для простоты изложения будем рассматривать сейчас последовательность равномерных сеток

$$\omega_h = \{x_i = a + ih, \quad i = 0, 1, \ldots, N\}$$

---

*) Изучение этого раздела не обязательно для понимания дальнейшего материала. Более подробное изложение см. в [20].

с шагом $h=(b-a)/N$. В этом случае основная система уравнений (9) принимает вид

$$c_{i-1}+4c_i+c_{i+1}=6f_{\bar{x}x,i}, \quad i=1, 2, \ldots, N-1, \tag{11}$$

$$c_0=c_N=0,$$

где обозначено $f_{\bar{x}x,i}=(f_{i-1}-2f_i+f_{i+1})/h^2$.

От функции $f(x)$ будем требовать существования непрерывной на $[a, b]$ четвертой производной, $f(x)\in C^{(4)}[a, b]$. Кроме того, предположим, что выполнены граничные условия $f''(a)=f''(b)=0$ и такие же условия для сплайнов. Обозначим

$$\|g(x)\|_{C[a,b]}=\max_{x\in[a,b]}|g(x)|, \quad M_4=\|f^{(4)}(x)\|_{C[a,b]}.$$

Пусть $s_h(x)$ — кубический сплайн, построенный для функции $f(x)$ на сетке ω_h. В следующей теореме приведены оценки погрешности интерполяции для функции $f(x)$ и ее производных $f'(x)$, $f''(x)$.

Теорема 1. *Для $f\in C^{(4)}[a, b]$ справедливы оценки*

$$\|f(x)-s_h(x)\|_{C[a,b]}\leqslant M_4h^4, \tag{12}$$

$$\|f'(x)-s_h'(x)\|_{C[a,b]}\leqslant M_4h^3, \tag{13}$$

$$\|f''(x)-s_h''(x)\|_{C[a,b]}\leqslant M_4h^2. \tag{14}$$

Из этих оценок следует, что при $h\to 0$ (т. е. при $N\to\infty$) последовательности $s_h^{(i)}(x)$, $i=0, 1, 2$, сходятся соответственно к функциям $f^{(i)}(x)$, $i=0, 1, 2$.

Для доказательства теоремы 1 потребуется следующая лемма 1, в которой даны оценки погрешности $f''(x_i)-s_h''(x_i)$ в узлах сетки. Будем обозначать

$$\|\varphi(x)\|_{C(\omega_h)}=\max_{x_i\in\omega_h}|\varphi(x_i)|.$$

Лемма 1. *Для $f(x)\in C^{(4)}[a, b]$ справедливы оценки*

$$\|f''(x)-s_h''(x)\|_{C(\omega_h)}\leqslant\frac{3M_4}{4}h^2. \tag{15}$$

Доказательство. Поскольку $s_h''(x_i)=c_i$, где c_i — решение системы (11), достаточно получить оценку для погрешности $z_i=c_i-f''(x_i)$, $i=1, 2, \ldots, N-1$. Подставляя $c_i=z_i+f_i''$ в (11), получаем для z_i уравнения

$$z_{i-1}+4z_i+z_{i+1}=\psi_i, \quad i=1, 2, \ldots, N-1, \quad z_0=z_N=0, \tag{16}$$

где

$$\psi_i=6f_{\bar{x}x,i}-(f_{i+1}''+4f_i''+f_{i-1}''). \tag{17}$$

Оценим решение системы уравнений (16) через правые части ψ_i. Для этого перепишем уравнение (16) в виде

$$4z_i=-z_{i-1}-z_{i+1}+\psi_i$$

и воспользуемся неравенствами

$$4|z_i| \leqslant |z_{i-1}| + |z_{i+1}| + |\psi_i| \leqslant 2\|z\|_{C(\omega_h)} + \|\psi\|_{C(\omega_h)}.$$

Так как это неравенство справедливо при всех i, оно выполняется и в той точке $x_i = x_{i_0}$, в которой достигается максимум $|z_i|$, т. е. в точке, где

$$|z_{i_0}| = \|z\|_{C(\omega_h)}.$$

Поэтому выполняется неравенство

$$4\|z\|_{C(\omega_h)} \leqslant 2\|z\|_{C(\omega_h)} + \|\psi\|_{C(\omega_h)},$$

т. е.

$$\|f''(x) - s_h''(x)\|_{C(\omega_h)} \leqslant \frac{1}{2}\|\psi\|_{C(\omega_h)}. \tag{18}$$

Для того чтобы получить отсюда неравенство (15), осталось оценить $\|\psi\|_{C(\omega_h)}$, где ψ_i определено согласно (17). Перепишем ψ_i в виде

$$\psi_i = 6(f_{\bar{x}x,i} - f_i'') - h^2 (f'')_{\bar{x}x,i} \tag{19}$$

и воспользуемся разложениями (см. п. 1 § 4 ч. I)

$$f_{\bar{x}x,i} = f_i'' + \frac{h^2}{12} f^{\mathrm{IV}}(\xi_i), \qquad \xi_i \in (x_{i-1}, x_{i+1}),$$

$$(f'')_{\bar{x}x\,i} = f^{\mathrm{IV}}(\zeta_i), \qquad \zeta_i \in (x_{i-1}, x_{i+1}),$$

справедливыми для $f(x) \in C^{(4)}[a, b]$. Тогда из (19) получим

$$\psi_i = \frac{h^2}{2} f^{\mathrm{IV}}(\xi_i) - h^2 f^{\mathrm{IV}}(\zeta_i),$$

т. е. при любом $i = 1, 2, \ldots, N-1$ справедлива оценка

$$|\psi_i| \leqslant \frac{3}{2} h^2 \max_{x \in [a,b]} |f^{\mathrm{IV}}(x)|,$$

или

$$\|\psi\|_{C(\omega_h)} \leqslant 1{,}5 h^2 M_4,$$

где $M_4 = \max\limits_{x \in [a,b]} |f^{\mathrm{IV}}(x)|$. Отсюда и из (18) получаем требуемое неравенство (15). Лемма 1 доказана.

Перейдем к доказательству теоремы 1. Получим сначала оценку (14) погрешности $f''(x) - s_h''(x)$, возникающей при интерполировании $f(x)$ кубическим сплайном $s_h(x)$. Рассмотрим отрезок $[x_{i-1}, x_i]$, где i — любое из чисел $1, 2, \ldots, N$. На этом отрезке $s_h''(x) = c_i + d_i(x - x_i)$ и согласно (4) имеем

$$s_h''(x) = c_i + \frac{c_i - c_{i-1}}{h}(x - x_i)$$

или

$$s_h''(x) = \alpha c_i + (1-\alpha)\, c_{i-1},$$

где

$$\alpha = \frac{x - x_{i-1}}{h}, \quad 1 - \alpha = \frac{x_i - x}{h}. \tag{20}$$

Для сеточных функций v, определенных на сетке ω_h, обозначим

$$v_i^{(\alpha)} = \alpha v_i + (1-\alpha)\, v_{i-1},$$

так что

$$s_h''(x) = c_i^{(\alpha)}.$$

Воспользовавшись тождеством

$$f''(x) = (f_i'')^{(\alpha)} + (f''(x) - f_i'')^{(\alpha)},$$

представим погрешность $f''(x) - s_h''(x)$ в виде

$$f''(x) - s_h''(x) = (f_i'' - c_i)^{(\alpha)} + (f''(x) - f_i'')^{(\alpha)}.$$

Отсюда получаем неравенство

$$|f''(x) - s_h''(x)| \leqslant |(f_i'' - c_i)^{(\alpha)}| + |(f''(x) - f_i'')^{(\alpha)}|, \tag{21}$$

справедливое для любого $x \in [x_{i-1}, x_i]$. Оценим отдельно каждое из двух слагаемых в правой части неравенства (21). Для первого слагаемого, учитывая лемму 1, имеем

$$|(f_i'' - c_i)^{(\alpha)}| = |\alpha (f_i'' - c_i) + (1-\alpha)(f_{i-1}'' - c_{i-1})| \leqslant$$
$$\leqslant \alpha \|f''(x) - s_h''(x)\|_{C(\omega_h)} + (1-\alpha)\|f''(x) - s_h(x)\|_{C(\omega_h)} \leqslant \frac{3}{4} M_4 h^2,$$

т. е.

$$|(f_i'' - c_i)^{(\alpha)}| \leqslant \frac{3}{4} M_4 h^2. \tag{22}$$

Далее, рассмотрим выражение

$$(f''(x) - f_i'')^{(\alpha)} = \alpha (f''(x) - f_i'') + (1-\alpha)(f''(x) - f_{i-1}'').$$

По формуле Тейлора имеем

$$f''(x) - f_i'' = (x - x_i) f'''(x) - \frac{(x-x_i)^2}{2} f^{\mathrm{IV}}(\xi_i),$$

$$f''(x) - f_{i-1}'' = (x - x_{i-1}) f'''(x) - \frac{(x-x_{i-1})^2}{2} f^{\mathrm{IV}}(\zeta_i),$$

где $\xi_i, \zeta_i \in (x_{i-1}, x_i)$. Отсюда и из (20) получим

$$(f''(x) - f_i'')^{(\alpha)} = [\alpha (x - x_i) + (1-\alpha)(x - x_{i-1})] f'''(x) -$$
$$- \frac{1}{2}[\alpha (x-x_i)^2 f^{\mathrm{IV}}(\xi_i) + (1-\alpha)(x-x_{i-1})^2 f^{\mathrm{IV}}(\zeta_i)] =$$
$$= -\frac{(x-x_{i-1})(x_i - x)}{2h}[(x_i - x) f^{\mathrm{IV}}(\xi_i) + (x - x_{i-1}) f^{\mathrm{IV}}(\zeta_i)],$$

так что

$$|(f''(x)-f_i'')^{(\alpha)}|\leqslant$$

$$\leqslant\frac{1}{2h}\max_{x_{i-1}\leqslant x\leqslant x_i}[(x-x_{i-1})(x_i-x)]\{M_4(x_i-x)+M_4(x-x_{i-1})\}=$$

$$=0{,}5M_4\frac{h^2}{4}=\frac{M_4h^2}{8}.$$

Итак, второе слагаемое в правой части неравенства (21) оценивается следующим образом:

$$|(f''(x)-f_i'')^{(\alpha)}|\leqslant\frac{M_4h^2}{8}. \tag{23}$$

Подставляя оценки (22) и (23) в неравенство (21), получим

$$|f''(x)-s_h''(x)|\leqslant\frac{7M_4h^2}{8} \tag{24}$$

для любого $x\in[x_{i-1}, x_i]$. Поскольку неравенство (24) справедливо для любого $i=1, 2, \ldots, N$, из него следует оценка (14).

Докажем теперь оценку (13). Рассмотрим на отрезке $[x_{i-1}, x_i]$ функцию $r(x)=f(x)-s_h(x)$. По определению сплайна имеем $r(x_{i-1})=r(x_i)=0$, следовательно, найдется точка $\xi\in(x_{i-1}, x_i)$, в которой $r'(\xi)=0$. Поэтому

$$|r'(x)|=|r'(x)-r'(\xi)|=|r''(\zeta)(x-\xi)|\leqslant|r''(\zeta)|h,$$

где $\zeta\in(x_{i-1}, x_i)$. Таким образом,

$$|f'(x)-s_h'(x)|\leqslant|f''(\zeta)-s_h''(\zeta)|h,$$

и, учитывая (14), получим неравенство

$$|f'(x)-s_{h3}''(x)|\leqslant M_4h^3,$$

из которого следует оценка (13).

Осталось получить оценку (12). Пусть x — любая точка из интервала (x_{i-1}, x_i). Введем функцию

$$g(t)=f(t)-s_h(t)-K(t-x_{i-1})(t-x_i), \tag{25}$$

где $t\in[x_{i-1}, x_i]$ и K — постоянная, выбираемая из условия $g(x)=0$, т. е.

$$K=\frac{f(x)-s_h(x)}{(x-x_{i-1})(x-x_i)}.$$

Имеем $g(x_{i-1})=g(x_i)=g(x)=0$. Поэтому найдется хотя бы одна точка $\xi\in(x_{i-1}, x_i)$, в которой $g''(\xi)=0$. Поскольку

$$g''(t)=f''(t)-s_h''(t)-2K,$$

получим

$$f''(\xi)-s_h''(\xi)=2K,$$

т. е.

$$f(x)-s_h(x)=\frac{f''(\xi)-s_h''(\xi)}{2}(x-x_{i-1})(x-x_i).$$

Отсюда и из (14) получаем неравенство

$$|f(x)-s_{i}(x)|\leqslant\frac{1}{2}\|f''(x)-s_h''(x)\|_{C(\omega_h)}\frac{h^2}{4}\leqslant\frac{M_4h^4}{8},$$

которое и приводит к оценке (12). Теорема 1 доказана.

§ 5. Другие постановки задач интерполирования и приближения функций

1. Примеры. Во многих случаях возникает необходимость приближенной замены данной функции другими, более простыми функциями. Одним из способов такой замены является интерполяция алгебраическими многочленами, подробно рассмотренная в предыдущих параграфах. Однако не всякую функцию целесообразно приближать алгебраическими многочленами. Отметим в виде примеров несколько других способов интерполирования.

Пример 1. *Тригонометрическая интерполяция.* Если $f(x)$ — периодическая функция с периодом l, то естественно строить приближения с помощью функций

$$\varphi_k(x)=a_k\cos\frac{\pi kx}{l}+b_k\sin\frac{\pi kx}{l},\qquad k=0,\ 1,\ \ldots,\ n.$$

Таким образом, *тригонометрическая интерполяция* состоит в замене $f(x)$ тригонометрическим многочленом

$$T_n(x)=\sum_{k=0}^{n}\varphi_k(x)=a_0+\sum_{k=1}^{n}\left(a_k\cos\frac{\pi kx}{l}+b_k\sin\frac{\pi kx}{l}\right),$$

коэффициенты которого отыскиваются из системы уравнений

$$T_n(x_j)=f(x_j),\qquad j=1,\ 2,\ \ldots,\ 2n+1,$$

где $x_0<x_1<\ldots<x_{2n+1}$, $x_{2n+1}-x_0=l$.

Пример 2. *Приближение рациональными функциями.* Пусть значения функции $f(x)$ заданы в точках $x_0<x_1<\ldots<x_n$. Требуется построить функцию

$$\varphi_{kl}(x)=\frac{a_kx^k+a_{k-1}x^{k-1}+\ldots+a_0}{x^l+b_{l-1}x^{l-1}+\ldots+b_0}\qquad(1)$$

(k, l — заданы), для которой

$$\varphi_{kl}(x_j)=f(x_j),\qquad j=0,\ 1,\ \ldots,\ n.\qquad(2)$$

Уравнения (2) представляют собой систему из $n+1$ уравнения относительно $k+l+1$ неизвестного $a_0,\ a_1,\ \ldots,\ a_k,\ b_0,\ b_1,\ \ldots,\ b_{l-1}$. Будем требовать, чтобы число уравнений равнялось числу неизвестных, т. е. $n=k+l$. Тогда придем к системе линейных уравнений

$$\sum_{i=0}^{k}a_ix_j^i-f_j\sum_{i=0}^{l-1}b_ix_j^i=f_jx_j^l,\qquad j=0,\ 1,\ \ldots,\ k+l,\qquad(3)$$

в которой неизвестными являются величины a_i, $i=0, 1, \ldots, k$, и b_i, $i=0, 1, \ldots, l-1$.

Пример 3. *Дробно-линейная интерполяция.* Пусть значения функции $f(x)$ заданы в трех узлах, а именно в точках x_{i-1}, x_i, x_{i+1}, причем $x_{i-1}<x_i<x_{i+1}$. Построим функцию

$$\varphi(x)=\frac{a_1x+a_0}{x+b_0}, \tag{4}$$

для которой $\varphi(x_j)=f(x_j)$, $j=i-1, i, i+1$. Данная задача является частным случаем задачи, сформулированной в предыдущем примере, когда $k=l=1$. Поэтому для определения коэффициентов a_0, a_1, b_0 можно воспользоваться системой уравнений (3), которая в данном случае примет вид

$$\begin{aligned} a_0+a_1x_{i-1}-b_0f_{i-1}&=x_{i-1}f_{i-1},\\ a_0+a_1x_i-b_0f_i&=x_if_i,\\ a_0+a_1x_{i+1}-b_0f_{i+1}&=x_{i+1}f_{i+1}. \end{aligned} \tag{5}$$

Найдем в явном виде решение этой системы. Обозначим

$$h_i=x_i-x_{i-1},\quad h_{i+1}=x_{i+1}-x_i,\quad \hbar_i=0{,}5\,(h_{i+1}+h_i),$$
$$f_{\bar{x},i}=(f_i-f_{i-1})/h_i,\quad f_{x,i}=(f_{i+1}-f_i)/h_{i+1},$$
$$f_{\bar{x}\hat{x},i}=(f_{x,i}-f_{\bar{x},i})\,\hbar_i.$$

Применяя последовательное исключение неизвестных, приведем систему (5) к треугольному виду

$$\begin{aligned} a_0+a_1x_i-b_0f_i&=x_if_i,\\ a_1-b_0f_{\bar{x},i}&=(xf)_{\bar{x},i},\\ -b_0f_{\bar{x}\hat{x},i}&=(xf)_{\bar{x}\hat{x},i}. \end{aligned} \tag{6}$$

Если $f_{\bar{x}\hat{x},i}\neq 0$, то из (6) последовательно найдем

$$b_0=-\frac{(xf)_{\bar{x}\hat{x},i}}{f_{\bar{x}\hat{x},i}},\quad a_1=f_i-\frac{2f_{\bar{x},i}f_{x,i}}{f_{\bar{x}\hat{x},i}},$$
$$a_0=\frac{2x_if_{\bar{x},i}f_{x,i}-f_i(xf)_{\bar{x}\hat{x},i}}{f_{\bar{x}\hat{x},i}}.$$

При проведении вычислений по этим формулам может оказаться полезным тождество

$$(xf)_{\bar{x}\hat{x},i}=x_if_{\bar{x}\hat{x},i}+\frac{f_{i+1}-f_{i-1}}{\hbar_i}.$$

Используя приближение с помощью рациональных функций, необходимо следить за тем, чтобы на отрезке интерполирования знаменатель выражения (1) не обращался в нуль. Другой опасностью является такой неудачный выбор узлов интерполирования, при котором числитель выражения (1) делится без остатка на зна-

менатель. В последнем случае дробно-линейная функция (4) вырождается в константу.

В качестве примера рассмотрим функцию $f(x)=kx^2$, $k\neq 0$. Для нее $f_{\hat{x}\hat{x}}=2k\neq 0$ и система (6) имеет единственное решение

$$b_0=-(x_{i-1}+x_i+x_{i+1}),$$
$$a_1=-k(x_{i-1}x_i+x_{i-1}x_{i+1}+x_ix_{i+1}),$$
$$a_0=kx_{i-1}x_ix_{i+1},$$

причем

$$a_1b_0-a_0=k(x_{i+1}+x_{i-1})[x_i(x_{i-1}+x_i+x_{i+1})+x_{i+1}x_{i-1}].$$

Следовательно, приближение $f(x)$ с помощью функции (4) невозможно вблизи точки $x=-b_0=x_{i-1}+x_i+x_{i+1}$. Кроме того, условие $a_1b_0-a_0\neq 0$ приводит к следующим ограничениям на расположение узлов интерполирования:

$$x_{i+1}+x_{i-1}\neq 0,\quad x_i(x_{i-1}+x_i+x_{i+1})+x_{i+1}x_{i-1}\neq 0.$$

Пример 4. *Двумерная интерполяция.* На плоскости xOy заданы три точки $A_i(x_i, y_i)$, $i=1, 2, 3$, не лежащие на одной прямой. Требуется, используя значения $u_i=u(x_i, y_i)$ функции $u(x, y)$ в этих точках, построить аппроксимацию производных $\partial u/\partial x$, $\partial u/\partial y$. Для решения этой задачи воспользуемся линейной интерполяцией, т. е. будем считать, что

$$u(x, y)=a(x-x_1)+b(y-y_1)+c. \tag{7}$$

Тогда получим, что $\partial u/\partial x=a$, $\partial u/\partial y=b$, т. е. при интерполяции функции $u(x, y)$ с помощью линейной функции производные заменяются константами. Явные выражения для коэффициентов a, b, c нетрудно найти из условий интерполирования $u(x_i, y_i)=u_i$. Действительно, из условия $u(x_1, y_1)=u_1$ получаем, что $c=u_1$. Далее, решая систему

$$a(x_2-x_1)+b(y_2-y_1)=u_2-u_1,$$
$$a(x_3-x_1)+b(y_3-y_1)=u_3-u_1,$$

получим

$$a=\frac{1}{\Delta}\begin{vmatrix} u_2-u_1 & y_2-y_1 \\ u_3-u_1 & y_3-y_1 \end{vmatrix},\quad b=\frac{1}{\Delta}\begin{vmatrix} x_2-x_1 & u_2-u_1 \\ x_3-x_1 & u_3-u_1 \end{vmatrix}, \tag{8}$$

где

$$\Delta=\begin{vmatrix} x_2-x_1 & y_2-y_1 \\ x_3-x_1 & y_3-y_1 \end{vmatrix}.$$

Выражения (8) и задают искомые приближения к производным $\partial u/\partial x$, $\partial u/\partial y$.

Определитель Δ данной системы не равен нулю, так как по условию точки A_1, A_2, A_3 не лежат на одной прямой.

Заметим, что соотношения (7), (8) можно записать в виде

$$\begin{vmatrix} u-u_1 & x-x_1 & y-y_1 \\ u_2-u_1 & x_2-x_1 & y_2-y_1 \\ u_3-u_1 & x_3-x_1 & y_3-y_1 \end{vmatrix}=0,$$

т. е. в виде уравнения плоскости, проходящей через три заданные точки (x_i, y_i, u_i), $i=1, 2, 3$.

2. Общая постановка задачи интерполирования. Пусть на отрезке $[a, b]$ задана система функций

$$\varphi_0(x), \varphi_1(x), \ldots, \varphi_n(x) \tag{9}$$

и введена сетка

$$a \leqslant x_0 < x_1 < \ldots < x_n \leqslant b. \tag{10}$$

Образуем линейную комбинацию

$$\varphi(x) = c_0\varphi_0(x) + c_1\varphi_1(x) + \ldots + c_n\varphi_n(x) \tag{11}$$

с числовыми коэффициентами $c_0, c_1, \ldots, c_n$. Задача интерполирования функции $f(x)$ системой функций (9) на сетке (10) состоит в нахождении коэффициентов $c_0, c_1, \ldots, c_n$, для которых выполнены условия

$$\varphi(x_j) = f(x_j), \quad j=0, 1, \ldots, n. \tag{12}$$

Интерполирование алгебраическими многочленами является частным случаем сформулированной задачи, когда $\varphi_k(x) = x^k$, $k=0, 1, \ldots, n$. Возникает вопрос о существовании и единственности решения общей задачи интерполирования. Запишем систему (12) более подробно:

$$\begin{aligned}
&c_0\varphi_0(x_0) + c_1\varphi_1(x_0) + \ldots + c_n\varphi_n(x_0) = f(x_0),\\
&c_0\varphi_0(x_1) + c_1\varphi_1(x_1) + \ldots + c_n\varphi_n(x_1) = f(x_1),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&c_0\varphi_0(x_n) + c_1\varphi_1(x_n) + \ldots + c_n\varphi_n(x_n) = f(x_n).
\end{aligned}$$

Для того чтобы эта система имела единственное решение, необходимо и достаточно, чтобы определитель матрицы

$$A = \begin{bmatrix}
\varphi_0(x_0) & \varphi_1(x_0) & \ldots & \varphi_n(x_0)\\
\varphi_0(x_1) & \varphi_1(x_1) & \ldots & \varphi_n(x_1)\\
\ldots & \ldots & \ldots & \ldots\\
\varphi_0(x_n) & \varphi_1(x_n) & \ldots & \varphi_n(x_n)
\end{bmatrix} \tag{13}$$

был отличен от нуля. Более того, поскольку узлы $x_0, x_1, \ldots, x_n$ могут быть как угодно расположены на $[a, b]$, лишь бы среди них не было совпадающих, необходимо потребовать, чтобы $\det A \neq 0$ при любом расположении узлов. Выполнение или невыполнение этого требования зависит от выбора системы функций $\{\varphi_k(x)\}_{k=0}^n$.

Система функций $\{\varphi_k(x)\}_{k=0}^n$ называется *системой Чебышева* на $[a, b]$, если определитель матрицы (13) отличен от нуля при любом расположении узлов $x_k \in [a, b]$, $k=0, 1, \ldots, n$, когда среди этих узлов нет совпадающих. Таким образом, общая задача интерполирования однозначно разрешима, если $\{\varphi_k(x)\}_{k=0}^n$ чебышевская система функций. Функция $\varphi(x)$, определенная согласно (11) и удовлетворяющая условиям интерполяции (12), называется *обобщенным интерполяционным многочленом* по системе $\{\varphi_k(x)\}_{k=0}^n$.

Система алгебраических многочленов $\varphi_k(x)=x^k$, $k=0, 1, \ldots, n$, является чебышевской системой на любом отрезке $[a, b]$. Система тригонометрических многочленов $\varphi_k(x)$ (см. пример 1) является чебышевской системой на отрезке периодичности.

Приведем примеры систем функций, не являющихся чебышевскими. Пусть на отрезке $[-1, 1]$ задана система функций

$$\varphi_0(x)=1,\ \varphi_1(x)=\begin{cases}0, & -1\leqslant x\leqslant 0,\\ x, & 0\leqslant x\leqslant 1.\end{cases}$$

Если в качестве узлов интерполирования взять, например, точки $x_0=-\frac{3}{4}$, $x_1=-\frac{1}{2}$, то получим

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

т. е. данная система не является чебышевской на $[-1, 1]$.

Менее тривиальным является пример системы

$$\varphi_0(x)=1,\quad \varphi_1(x)=x^2-1/4,\qquad x\in[-1, 1].$$

Выбрав в качестве узлов интерполирования корни функции $\varphi_1(x)$, т. е. точки $x_0=-0{,}5$, $x_1=0{,}5$, придем к той же самой матрице A.

Вообще, из (13) видно, что если какая-либо из функций $\varphi_0, \varphi_1, \ldots, \varphi_n$ обращается на отрезке $[a, b]$ в нуль более чем n раз, то система не является чебышевской. Действительно, если, например, $\varphi_j(x_k)=0$ для некоторого j и для $k=0, 1, \ldots, n$, то, выбирая точки $x_0, x_1, \ldots, x_n$ в качестве узлов интерполирования, получим, что j-й столбец матрицы A содержит только нулевые элементы.

Можно доказать, что справедливо следующее утверждение (см., например, [4]). Для того чтобы система $\{\varphi_k(x)\}_{k=0}^n$ была чебышевской на $[a, b]$, необходимо и достаточно, чтобы любой обобщенный многочлен по этой системе, у которого хотя бы один из коэффициентов отличен от нуля, имел на $[a, b]$ не более n нулей. Иногда это свойство принимается за определение чебышевской системы.

3. Наилучшее приближение функции, заданной таблично. Пусть значения функции $f(x)$ и функций $\varphi_j(x)$, $j=0, 1, \ldots, n$, из системы (9) известны в точках $x_k\in[a, b]$, $k=0, 1, \ldots, m$. Если $m>n$, то задача интерполирования становится переопределенной. В этом случае можно рассматривать *задачу о наилучшем приближении*, которая формулируется следующим образом.

Введем обобщенный многочлен (11) и будем рассматривать его значения только в узлах x_k, т. е.

$$\varphi(x_k)=c_0\varphi_0(x_k)+c_1\varphi_1(x_k)+\ldots+c_n\varphi_n(x_k),\quad k=0, 1, \ldots, m.$$

Образуем разности

$$r_k=\varphi(x_k)-f(x_k),\quad k=0, 1, \ldots, m,$$

характеризующие отклонение в узлах x_k точного значения функ-

ции $f(x)$ от ее приближенного значения, полученного с помощью обобщенного многочлена (11). Для вектора погрешностей

$$r=(r_0, r_1, \ldots, r_m)^T$$

можно ввести ту или иную норму, например,

$$\|r\|=\left(\sum_{k=0}^{m} r_k^2\right)^{1/2}=\left(\sum_{k=0}^{m} (\varphi(x_k)-f(x_k))^2\right)^{1/2}, \tag{14}$$

или

$$\|r\|=\max_{0\leqslant k\leqslant m}|r_k|=\max_{0\leqslant k\leqslant m}|\varphi_k-f(x_k)|. \tag{15}$$

Задача о наилучшем приближении функции $f(x)$, заданной таблично, состоит в нахождении коэффициентов $c_0, c_1, \ldots, c_n$, минимизирующих норму вектора r. В зависимости от выбора нормы получим различные задачи. Так, норме (14) соответствует задача о наилучшем среднеквадратичном приближении, а норме (15) — задача о наилучшем равномерном приближении функции, заданной таблично.

Если $m=n$, то независимо от выбора нормы решение $c=(c_0, c_1, \ldots, c_n)^T$ задачи о наилучшем приближении совпадает с решением задачи интерполирования. Действительно, в этом случае требование $\|r\|=0$ приводит к условиям

$$\varphi(x_k)=f(x_k), \quad k=0, 1, \ldots, n,$$

т. е. к задаче интерполирования.

Пример 1. Построим наилучшее среднеквадратичное приближение для случая $n=1$, $m=2$, когда заданы $f_i=f(x_i)$, $i=0, 1, 2$. Обозначим $h_0=x_1-x_0$, $h_1=x_2-x_1$ и будем искать обобщенный многочлен $\varphi(x)$ в виде

$$\varphi(x)=c_0+c_1(x-x_1).$$

Тогда для $r(x)=\varphi(x)-f(x)$ получим, что $\|r\|^2=F(c_0, c_1)$, где

$$F(c_0, c_1)=(c_0-c_1h_0-f_0)^2+(c_0-f_1)^2+(c_0+c_1h_1-f_2)^2.$$

Точку минимума $F(c_0, c_1)$ найдем из условий $F'_{c_0}=F'_{c_1}=0$, которые приводят к системе уравнений

$$3c_0+(h_1-h_0)c_1=f_0+f_1+f_2,$$
$$(h_1-h_0)c_0+(h_0^2+h_1^2)c_1=h_1f_2-h_0f_0.$$

Отсюда получим

$$c_0=\alpha_0f_0+(1-\alpha_0-\alpha_2)f_1+\alpha_2f_2,$$
$$c_1=\beta\frac{f_2-f_1}{h_1}+(1-\beta)\frac{f_1-f_0}{h_0},$$

где

$$\alpha_0=\frac{h_1(h_0+h_1)}{2(h_0^2+h_1^2+h_1h_0)}, \quad \alpha_2=\frac{h_0(h_0+h_1)}{2(h_0^2+h_1^2+h_1h_0)}, \quad \beta=\frac{h_1(2h_1+h_0)}{2(h_0^2+h_1^2+h_1h_0)}.$$

Вводя обозначения $\hbar=0{,}5(h_1+h_2)$, $f=f_1$, $f_x=(f_2-f_1)/h_1$, $f_{\bar{x}}=$

$= (f_1 - f_0)/h_0$, $f_{\bar{x}\hat{x}} = (f_x - f_{\bar{x}})/\hbar$, можно записать коэффициенты c_0 и c_1 в виде

$$c_0 = f + \frac{h_1 h_0 \hbar^2}{h_0^2 + h_1^2 + h_1 h_0} f_{\bar{x}\hat{x}}, \tag{16}$$

$$c_1 = \frac{1}{2(h_0^2 + h_1^2 + h_1 h_0)} ((2h_1 + h_0) h_1 f_x + (2h_0 + h_1) h_0 f_{\bar{x}}). \tag{17}$$

Если $h_1 = h_2 = h$, то

$$c_0 = \frac{1}{3}(f_0 + f_1 + f_2), \quad c_1 = \frac{f_2 - f_0}{2h}. \tag{18}$$

Оценим погрешность полученного приближения. Проводя элементарные выкладки, получим с учетом (16), (17), что

$$r_1 = c_0 - f_1 = \frac{h_1 h_0 \hbar^2}{h_0^2 + h_1^2 + h_1 h_0} f_{\bar{x}\hat{x}},$$

$$r_0 = c_0 - c_1 h_0 - f_0 = -\frac{h_1}{2\hbar} r_1,$$

$$r_2 = c_0 - c_1 h_1 - f_2 = -\frac{h_0}{2\hbar} r_1.$$

Отсюда имеем

$$\|r\|^2 = r_0^2 + r_1^2 + r_2^2 = \frac{h_1^2 + h_0^2 + h_1 h_0}{2\hbar^2} r_1^2 = \frac{h_1^2 h_0^2 \hbar^2}{2(h_0^2 + h_1^2 + h_1 h_0)} (f_{\bar{x}\hat{x}})^2,$$

следовательно,

$$\|r\| = \|f - \varphi\| = \frac{h_1 h_0 \hbar}{(2(h_0^2 + h_1^2 + h_1 h_0))^{1/2}} |f_{\bar{x}\hat{x}}|.$$

Согласно (6) из § 2 существует точка $\xi \in (x_0, x_2)$, для которой $f_{\bar{x}\hat{x}} = f''(\xi)$. Поэтому окончательно можно записать

$$\|f - \varphi\| = \frac{h_1 h_0 \hbar}{\sqrt{2(h_0^2 + h_1^2 + h_1 h_0)}} |f''(\xi)|.$$

В частности, на равномерной сетке, когда $h_1 = h_2 = h$, получим

$$\|f - \varphi\| = \frac{h^2}{\sqrt{6}} |f''(\xi)|,$$

т. е. погрешность имеет второй порядок по h.

4. Сглаживание сеточных функций. Пусть имеется таблица значений $\{f_i\}_{i=0}^N$ функции $f(x)$, полученная, например, путем измерения некоторой физической величины или с помощью численных расчетов. Может оказаться, что $f(x)$ сильно меняется на отдельных участках. В этом случае иногда целесообразно применить *процедуру сглаживания*, т. е. приближенно заменить $f(x)$ другой, более гладкой функцией $\varphi(x)$.

Для построения сглаженных функций можно воспользоваться среднеквадратичными приближениями, рассмотренными в преды-

дущем пункте. Согласно (18) получаем, что многочлен $\varphi^{(i)}(x)$ наилучшего среднеквадратичного приближения, построенный по значениям f_{i-1}, f_i, f_{i+1}, имеет вид

$$\varphi^{(i)}(x) = \frac{f_{i-1}+f_i+f_{i+1}}{3} + \frac{f_{i+1}-f_{i-1}}{2h}(x-x_i),$$

причем

$$\varphi^{(i)}(x_i) = \frac{f_{i-1}+f_i+f_{i+1}}{3}, \quad i=1, 2, \ldots, N-1. \tag{19}$$

Доопределим $\varphi^{(0)}(x_0)=f_0$, $\varphi^{(N)}(x_N)=f_N$ и обозначим $\varphi_i=\varphi^{(i)}(x_i)$, $i=0, 1, \ldots, N$.

Процедура сглаживания по формулам (19) состоит в замене сеточной функции $\{f_i\}_{i=0}^N$ сеточной функцией $\{\varphi_i\}_{i=0}^N$, определенной согласно (19). То, что такая замена действительно осуществляет сглаживание, можно иллюстрировать примером, приведенным в таблице.

i	0	1	2	3	4	5	6	7	8	9	10	11	12
f_i	1	1	1	1	0	0	0	0	10	0	0	0	0
φ_i	1	1	1	$\frac{2}{3}$	$\frac{1}{3}$	0	0	$\frac{10}{3}$	$\frac{10}{3}$	$\frac{10}{3}$	0	0	0

Здесь функция f_i имеет две особенности: разрыв при $i=3$ и выброс при $i=8$. Сглаживание приводит к размазыванию разрыва, а также к размазыванию выброса и уменьшению его амплитуды. На участках гладкости $f(x)$ функция $\varphi(x)$ также остается гладкой. Для наглядности читателю предлагается построить графики функций $f(x)$ и $\varphi(x)$.

В рассмотренном случае сглаживание свелось к *осреднению* функции $f(x)$ по трем соседним точкам. Можно проводить осреднение и по большему числу точке, например по пяти точкам, когда

$$\varphi_i = \sum_{j=-2}^{2} \alpha_j f_{i+j}, \quad \sum_{j=-2}^{2} \alpha_j = 1.$$

Чтобы выяснить, почему осреднение приводит к сглаживанию, вернемся к рассмотренному примеру. Будем считать, что $f(x)$ задана на равномерной сетке

$$\omega_h=\{x_i=ih,\ i=0, 1, \ldots, N,\ hN=l\},$$

причем $f_0=f_N=0$. Сглаживание по формулам (19) приводит к функции

$$\varphi_i = \frac{f_{i-1}+f_i+f_{i+1}}{3} = f_i + \frac{h^2}{3} f_{\bar{x}x,i}, \tag{20}$$
$$i=1, 2, \ldots, N-1, \quad \varphi_0=\varphi_N=0,$$

т. е. к осреднению $f(x)$ по трем соседним точкам. Таким образом, можно считать, что процедура осреднения представляет собой замену сеточной функции f

сеточной функцией Tf, где $T=E+\frac{h^2}{3}\Lambda$, E — единичный оператор, Λ — оператор второй разностной производной. Будем называть T *оператором осреднения.*

В п. 5 § 4 ч. I показано, что любую сеточную функцию f, для которой $f_0=f_N=0$, можно представить в виде разложения

$$f(x)=\sum_{k=1}^{N-1} c_k\mu_k(x), \qquad x\in\omega_h, \tag{21}$$

где $\mu_k(x)$ — собственные функции оператора Λ:

$$\Lambda\mu_k(x)+\lambda_k\mu_k(x)=0. \tag{22}$$

Собственные функции и собственные числа оператора Λ можно выписать в явном виде (см. п. 4 § 4 ч. I):

$$\lambda_k=\frac{4}{h^2}\sin^2\frac{\pi k}{2N}, \quad \mu_k(x_j)=\sqrt{\frac{2}{l}}\sin\frac{\pi kj}{N},$$
$$k=1, 2, \ldots, N-1, \quad j=0, 1, \ldots, N.$$

Применяя к f оператор T, получим согласно (21), (22) разложение

$$Tf(x)=\sum_{k=1}^{N-1} c_k t_k\mu_k(x), \tag{23}$$

где $t_k=1-\frac{h^2}{3}\lambda_k=1-\frac{4}{3}\sin^2\frac{\pi k}{2N}$ — собственные значения оператора T.

Коэффициент t_k в разложении (23) характеризует влияние оператора осреднения T на k-ю гармонику. Для низкочастотных гармоник, когда k/N мало, имеем $\sin^2\frac{\pi k}{2N}\approx 0$ и t_k близко к единице. Для больших k, когда $k/N\approx 1$, имеем $\sin^2\frac{\pi k}{2N}\approx 1$ и $|t_k|\approx 1/3$. Таким образом, оператор T не подавляет низкочастотные гармоники и уменьшает амплитуду высокочастотных гармоник примерно в три раза. Этим и объясняется эффект сглаживания.

§ 6. Наилучшие приближения в гильбертовом пространстве

1. Постановка задачи. В п. 3 § 5 рассматривалась задача о приближении функции, заданной таблично. Однако задачу о приближении функций можно сформулировать и в более общем виде, а именно в терминах теории приближений в линейных нормированных пространствах.

Пусть дано линейное нормированное пространство H, может быть бесконечномерное, и в нем задана конечная система линейно независимых элементов

$$\varphi_k\in H, \qquad k=0, 1, \ldots, n. \tag{1}$$

Требуется приближенно заменить заданный элемент $f\in H$ линейной комбинацией

$$\varphi=c_0\varphi_0+c_1\varphi_1+\ldots+c_n\varphi_n. \tag{2}$$

Элемент φ, определенный согласно (2), называется *обобщенным многочленом*, построенным по системе элементов (1).

Будем рассматривать *задачу о наилучшем приближении*, состоящую в том, чтобы для заданного $f \in H$ среди всех линейных комбинаций вида (2) найти такой обобщенный многочлен φ, для которого отклонение

$$\|f-(c_0\varphi_0+c_1\varphi_1+\ldots+c_n\varphi_n)\| \tag{3}$$

было бы минимальным. Элемент

$$\bar{\varphi}=\bar{c}_0\varphi_0+\bar{c}_1\varphi_1+\ldots+\bar{c}_n\varphi_n,$$

дающий решение этой задачи, называется *элементом наилучшего приближения*.

Известно (см., например, [2]), что при весьма общих предположениях элемент наилучшего приближения существует и единствен. В зависимости от выбора пространства H, нормы $\|\cdot\|$ и системы $\{\varphi_k\}_{k=0}^n$ можно получить ту или иную конкретную задачу о наилучшем приближении.

Рассмотрим более подробно задачу о наилучшем приближении в том случае, когда H — вещественное гильбертово пространство со скалярным произведением $(f, g)_H$ и нормой $\|f\|_H=\sqrt{(f, f)_H}$. Типичным примером гильбертова пространства является пространство $L_2(a, b)$ вещественных функций $f(x)$, интегрируемых с квадратом на $[a, b]$, причем

$$(f, g)_{L_2}=\int_a^b f(x)\,g(x)\,dx, \quad \|f\|_{L_2}=\left(\int_a^b |f(x)|^2\,dx\right)^{1/2}. \tag{4}$$

Пусть задана конечная система линейно независимых элементов $\varphi_k \in H$, $k=0, 1, \ldots, n$. В данном случае задача о наилучшем приближении состоит в том, чтобы для заданного элемента $f \in H$ найти обобщенный многочлен

$$\bar{\varphi}=\bar{c}_0\varphi_0+\bar{c}_1\varphi_1+\ldots+\bar{c}_n\varphi_n, \tag{5}$$

для которого отклонение

$$\|f-\bar{\varphi}\|_H=(f-\varphi, f-\varphi)_H^{1/2} \tag{6}$$

является минимальным среди всех обобщенных многочленов вида

$$\varphi=c_0\varphi_0+c_1\varphi_1+\ldots+c_n\varphi_n.$$

2. Сведение к алгебраической задаче о минимуме квадратичного функционала. Покажем, что сформулированная задача имеет единственное решение. Перепишем равенство (6) в виде

$$\|f-\bar{\varphi}\|_H^2=\sum_{k,l=0}^n c_k c_l\,(\varphi_k, \varphi_l)_H-2\sum_{k=0}^n c_k\,(f, \varphi_k)_H+\|f\|_H^2. \tag{7}$$

Пусть $A=[a_{kl}]$ — матрица с элементами

$$a_{kl}=(\varphi_k, \varphi_l)_H, \qquad k, l=0, 1, \ldots, n \tag{8}$$

и c, $\hat{f}$ — векторы,

$$c=(c_0, c_1, \ldots, c_n)^T, \quad \hat{f}=(f_0, f_1, \ldots, f_n)^T,$$

где $f_i=(f, \varphi_i)_H$, $i=0, 1, \ldots, n$. Обозначая через

$$(u, v)=\sum_{i=0}^{n} u_i v_i$$

скалярные произведения векторов u и v, можно записать тождество (7) в виде

$$\|f-\bar{\varphi}\|_H^2=(Ac, c)-2(\hat{f}, c)+\|f\|_H^2. \tag{9}$$

Отсюда видно, что задача о нахождении наилучшего приближения в гильбертовом пространстве H сводится к минимизации функционала

$$F(c)=(Ac, c)-2(\hat{f}, c), \tag{10}$$

определенного на множестве вещественных $(n+1)$-мерных векторов.

Отметим основные свойства матрицы A. Прежде всего, A — симметричная матрица, поскольку $a_{kl}=(\varphi_k, \varphi_l)_H=(\varphi_l, \varphi_k)_H=a_{lk}$. Кроме того, A — положительно определенная матрица.

Докажем последнее свойство, исходя из тождества (7). При $f=0$ из (7) получим

$$(Ac, c)=\|\bar{\varphi}\|_H^2=\left\|\sum_{k=0}^{n} c_k\varphi_k\right\|_H^2 \geqslant 0$$

для любого вектора c. Предположим, что $(Ay, y)=0$ для некоторого $y=(y_0, y_1, \ldots, y_n)^T$. Тогда для обобщенного многочлена

$$\varphi=y_0\varphi_0+y_1\varphi_1+\ldots+y_n\varphi_n$$

имеем $\|\varphi\|_H^2=(Ay, y)=0$, следовательно, $\varphi=\sum_{k=0}^{n} y_k\varphi_k=0$. Отсюда в силу линейной независимости элементов $\varphi_0, \varphi_1, \ldots, \varphi_n$ получим, что $y_0=y_1=\ldots=y_n=0$. Таким образом, $(Ac, c)>0$ для всех $c\neq 0$, т. е. A — положительно определенная матрица. Заметим, что положительно определенными являются и матрицы всех угловых миноров A.

Следующая теорема сводит проблему минимизации квадратичного функционала (10) к решению некоторой системы линейных алгебраических уравнений.

Теорема 1. *Пусть A — симметричная положительно определенная матрица и $\hat{f}$ — заданный вектор. Тогда функционал* (10) *имеет единственную точку минимума $\bar{c}$. Вектор $\bar{c}$ удовлетворяет системе уравнений*

$$A\bar{c}=\hat{f}. \tag{11}$$

Доказательство. Заметим прежде всего, что система (11) имеет единственное решение, поскольку A — положительно определенная матрица. Остается доказать, что вектор $\bar{c}$ минимизирует функционал (10) тогда и только тогда, когда он является решением системы (11). Докажем сначала достаточность. Для любых векторов v и $\bar{c}$ имеем

$$F(\bar{c}+v) = (A(\bar{c}+v), \bar{c}+v) - 2(\hat{f}, \bar{c}+v) =$$
$$= (A\bar{c}, \bar{c}) - 2(\hat{f}, \bar{c}) + 2(A\bar{c}, v) - 2(\hat{f}, v) + (Av, v),$$

т. е.

$$F(\bar{c}+v) = F(\bar{c}) + 2(A\bar{c} - \hat{f}, v) + (Av, v). \tag{12}$$

Предположим, что $\bar{c}$ является решением системы (11). Тогда из (12) получим

$$F(\bar{c}+v) = F(\bar{c}) + (Av, v).$$

В силу положительной определенности матрицы A отсюда следует неравенство

$$F(\bar{c}+v) > F(\bar{c})$$

для любого ненулевого вектора v. Это и означает, что $\bar{c}$ — точка минимума функционала $F(c)$.

Докажем необходимость условия (11). Надо показать, что если $\bar{c}$ — точка минимума функционала (10), то выполнено уравнение (11). Для этого воспользуемся тождеством (12), в котором положим $v=\lambda y$, где λ — действительное число и y — произвольный вектор. Тогда получим

$$F(\bar{c}+\lambda y) = F(\bar{c}) + 2\lambda(A\bar{c} - \hat{f}, y) + \lambda^2(Ay, y).$$

Рассмотрим выражение в правой части этого тождества как функцию λ и обозначим

$$g(\lambda) = \lambda^2(Ay, y) + 2\lambda(A\bar{c} - \hat{f}, y) + F(\bar{c}).$$

Поскольку $\bar{c}$ — точка минимума функционала $F(c)$, при любых y и λ выполняется неравенство $F(\bar{c}+\lambda y) \geqslant F(\bar{c})$, т. е. $g(\lambda) \geqslant g(0)$. Таким образом, $\lambda=0$ является точкой минимума $g(\lambda)$ и, следовательно, $g'(0)=0$. Отсюда получаем, что

$$g'(0) = 2(A\bar{c} - \hat{f}, y) = 0$$

и в силу произвольности вектора y приходим к выводу, что $A\bar{c}-\hat{f}=0$. Теорема 1 доказана.

3. Следствия. Более подробно систему (11) можно записать в виде

$$\sum_{l=0}^{n} \bar{c}_l(\varphi_k, \varphi_l)_H = (f, \varphi_k)_H, \qquad k = 0, 1, \ldots, n. \tag{13}$$

Таким образом, элемент наилучшего приближения в пространстве H имеет вид (5), где коэффициенты $\bar{c}_k$, $k=0, 1, \ldots, n$, отыскиваются

из системы (13). Из сказанного выше ясно, что алгоритм построения элемента наилучшего приближения в гильбертовом пространстве состоит в следующем:

1) вычисление элементов $a_{kl}=(\varphi_k, \varphi_l)_H$, $k, l=0, 1, \ldots, n$, матрицы A;

2) вычисление правых частей $(f, \varphi_k)_H$, $k=0, 1, \ldots, n$;

3) решение системы (13);

4) вычисление суммы $\bar{\varphi}=\sum_{k=0}^{n} \bar{c}_k \varphi_k$.

Как правило, каждый из этапов этого алгоритма осуществляется приближенно, с помощью численных методов. Например, в случае пространства L_2 необходимо уметь вычислять интегралы

$$(\varphi_k, f)_{L_2}=\int_a^b \varphi_k(x) f(x)\, dx,$$

что можно сделать, вообще говоря, лишь приближенно.

Оценим теперь отклонение $\|f-\bar{\varphi}\|_H$, которое получается в результате использования наилучшего приближения в гильбертовом пространстве. Докажем сначала, что справедлива

Лемма 1. *Если $\bar{\varphi}$ — элемент наилучшего приближения в H, то*

$$(f-\bar{\varphi}, \bar{\varphi})_H=0, \tag{14}$$

т. е. погрешность $f-\bar{\varphi}$ ортогональна элементу наилучшего приближения.

Доказательство. Из (11) имеем $(A\bar{c}, \bar{c})=(\hat{f}, \bar{c})$. Как показано ранее, $(A\bar{c}, \bar{c})=\|\bar{\varphi}\|_H^2$. Далее,

$$(\hat{f}, \bar{c})=\sum_{k=0}^{n} \bar{c}_k (f, \varphi_k)_H=\left(f, \sum_{k=0}^{n} \bar{c}_k \varphi_k\right)_H=(f, \bar{\varphi})_H.$$

Таким образом, приходим к тождеству

$$(f, \bar{\varphi})_H=\|\bar{\varphi}\|_H^2,$$

совпадающему с (14).

Следствие. Если $\bar{\varphi}$ — элемент наилучшего приближения в H, то

$$\|f-\bar{\varphi}\|_H^2=\|f\|_H^2-\|\bar{\varphi}\|_H^2. \tag{15}$$

Доказательство следует из тождества

$$\|f-\bar{\varphi}\|_H^2=\|f\|_H^2-2(f, \bar{\varphi})_H+\|\bar{\varphi}\|_H^2$$

и равенства (14).

Наиболее часто среднеквадратичные приближения используются в том случае, когда система $\{\varphi_k\}_{k=0}^{n}$ ортонормирована, т. е.

$$(\varphi_k, \varphi_l)_H=\begin{cases} 0, & k\neq l, \\ 1, & k=l. \end{cases}$$

Тогда система (13) решается в явном виде,

$$\bar{c}_k = (f, \varphi_k)_H, \quad k=0, 1, \ldots, n, \tag{16}$$

а погрешность приближения определяется формулой

$$\|f - \bar{\varphi}\|_H^2 = \|f\|_H^2 - \sum_{k=0}^{n} \bar{c}_k^2. \tag{17}$$

Числа $\bar{c}_k$, определенные согласно (16), называются *коэффициентами Фурье* элемента $f \in H$ по ортонормированной системе $\{\varphi_k\}_{k=0}^n$, а обобщенный многочлен

$$\bar{\varphi} = \sum_{k=0}^{n} \bar{c}_k \varphi_k$$

называется *многочленом Фурье.*

ГЛАВА 4

ЧИСЛЕННОЕ ИНТЕГРИРОВАНИЕ И ДИФФЕРЕНЦИРОВАНИЕ

§ 1. Примеры формул численного интегрирования

1. Введение. В настоящем параграфе рассматриваются способы приближенного вычисления определенных интегралов

$$I = \int_a^b f(x)\, dx, \tag{1}$$

основанные на замене интеграла конечной суммой

$$I_n = \sum_{k=0}^{n} c_k f(x_k), \tag{2}$$

где c_k — числовые коэффициенты и x_k — точки отрезка $[a, b]$, $k = 0, 1, \ldots, n$. Приближенное равенство

$$\int_a^b f(x)\, dx \approx \sum_{k=0}^{n} c_k f(x_k)$$

называется *квадратурной формулой*, а сумма вида (2) — *квадратурной суммой.* Точки x_k называются *узлами квадратурной формулы*, а числа c_k — *коэффициентами квадратурной формулы.* Разность

$$\Psi_n = \int_a^b f(x)\, dx - \sum_{k=0}^{n} c_k f(x_k)$$

называется *погрешностью квадратурной формулы.* Погрешность зависит как от расположения узлов, так и от выбора коэффициентов.

При оценке погрешности в приводимых ниже примерах функция $f(x)$ предполагается достаточно гладкой.

Введем на $[a, b]$ равномерную сетку с шагом h, т. е. множество точек

$$\omega_h = \{x_i = a + ih,\ i = 0, 1, \ldots, N,\ hN = b - a\},$$

и представим интеграл (1) в виде суммы интегралов по частичным отрезкам:

$$\int_a^b f(x)\,dx = \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} f(x)\,dx. \tag{3}$$

Для построения формулы численного интегрирования на всем отрезке $[a, b]$ достаточно построить квадратурную формулу для интеграла

$$\int_{x_{i-1}}^{x_i} f(x)\,dx \tag{4}$$

на частичном отрезке $[x_{i-1}, x_i]$ и воспользоваться свойством (3).

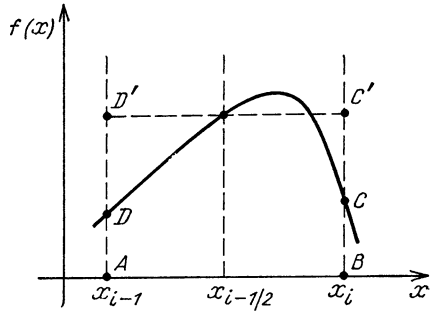


Рис. 5. Геометрический смысл формулы прямоугольников

2. Формула прямоугольников. Заменим интеграл (4) выражением $f(x_{i-1/2})h$, где $x_{i-1/2} = x_i - 0{,}5\,h$.

Геометрически такая замена означает, что площадь криволинейной трапеции $ABCD$ заменяется площадью прямоугольника $ABC'D'$ (см. рис. 5). Тогда получим формулу

$$\int_{x_{i-1}}^{x_i} f(x)\,dx \approx f(x_{i-1/2})\,h, \tag{5}$$

которая называется *формулой прямоугольников на частичном отрезке* $[x_{i-1}, x_i]$.

Погрешность метода (5) определяется величиной

$$\psi_i = \int_{x_{i-1}}^{x_i} f(x)\,dx - f(x_{i-1/2})\,h,$$

которую легко оценить с помощью формулы Тейлора. Действительно, запишем ψ_i в виде

$$\psi_i = \int_{x_{i-1}}^{x_i} (f(x) - f(x_{i-1/2}))\,dx \tag{6}$$

и воспользуемся разложением

$$f(x) = f(x_{i-1/2}) + (x - x_{i-1/2})\,f'(x_{i-1/2}) + \frac{(x - x_{i-1/2})^2}{2} f''(\zeta_i),$$

где $\zeta_i = \zeta_i(x) \in [x_{i-1}, x_i]$. Тогда из (6) получим

$$\psi_i = \int_{x_{i-1}}^{x_i} \frac{(x - x_{i-1/2})^2}{2} f''(\zeta_i)\, dx.$$

Обозначая $M_{2,i} = \max\limits_{x \in [x_{i-1}, x_i]} |f''(x)|$, оценим ψ_i следующим образом:

$$|\psi_i| \leqslant M_{2,i} \int_{x_{i-1}}^{x_i} \frac{(x - x_{i-1/2})^2}{2}\, dx = \frac{h^3}{24} M_{2,i}.$$

Таким образом, для погрешности формулы прямоугольников на частичном отрезке справедлива оценка

$$|\psi_i| \leqslant \frac{h^3}{24} M_{2,i}, \tag{7}$$

т. е. формула имеет погрешность $O(h^3)$ при $h \to 0$.

Заметим, что оценка (7) является неулучшаемой, т. е. существует функция $f(x)$, для которой (7) выполняется со знаком равенства. Действительно, для $f(x) = (x - x_{i-1/2})^2$ имеем $M_{2,i} = 2$, $f(x_{i-1/2}) = 0$ и

$$\int_{x_{i-1}}^{x_i} f(x)\, dx - f(x_{i-1/2})\, h = \frac{h^3}{12} = \frac{h^3}{24} M_{2,i}.$$

Суммируя равенства (5) по i от 1 до N, получим *составную формулу прямоугольников*

$$\int_a^b f(x)\, dx \approx \sum_{i=1}^{N} f(x_{i-1/2})\, h. \tag{8}$$

Погрешность этой формулы

$$\Psi = \int_a^b f(x)\, dx - \sum_{i=1}^{N} f(x_{i-1/2})\, h$$

равна сумме погрешностей по всем частичным отрезкам,

$$\Psi = \sum_{i=1}^{N} \psi_i = \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} \frac{(x - x_{i-1/2})^2}{2} f''(\zeta_i)\, dx.$$

Отсюда, обозначая $M_2 = \max\limits_{x \in [a,b]} |f''(x)|$, получим

$$|\Psi| \leqslant \frac{M_2 N h^3}{24} = \frac{h^2 (b-a)}{24} M_2, \tag{9}$$

т. е. погрешность формулы прямоугольников на всем отрезке есть величина $O(h^2)$.

В этом случае говорят, что квадратурная формула имеет *второй порядок точности.*

Замечание. Возможны формулы прямоугольников и при ином расположении узлов, например такие формулы:

$$\int_a^b f(x)\,dx \approx \sum_{i=1}^{N} hf(x_{i-1}), \quad \int_a^b f(x)\,dx \approx \sum_{i=1}^{N} hf(x_i).$$

Однако из-за нарушения симметрии погрешность таких формул является величиной $O(h)$.

3. Формула трапеций. На частичном отрезке эта формула имеет вид

$$\int_{x_{i-1}}^{x_i} f(x)\,dx \approx \frac{f(x_{i-1})+f(x_i)}{2}h \tag{10}$$

и получается путем замены подынтегральной функции $f(x)$ интерполяционным многочленом первой степени, построенным по узлам x_{i-1}, x_i, т. е. функцией

$$L_{1,i}(x) = \frac{1}{h}\left((x-x_{i-1})f(x_i) - (x-x_i)f(x_{i-1})\right).$$

Для оценки погрешности достаточно вспомнить (см. п. 1 § 2 гл. 3), что

$$f(x) - L_{1,i}(x) = \frac{(x-x_{i-1})(x-x_i)}{2} f''(\zeta_i(x)).$$

Отсюда получим

$$\psi_i = \int_{x_{i-1}}^{x_i} f(x)\,dx - \frac{f(x_{i-1})+f(x_i)}{2}h =$$

$$= \int_{x_{i-1}}^{x_i} (f(x) - L_{1,i}(x))\,dx = \int_{x_{i-1}}^{x_i} \frac{(x-x_{i-1})(x-x_i)}{2} f''(\zeta_i(x))\,dx$$

и, следовательно,

$$|\psi_i| \leqslant \frac{M_{2,i}h^3}{12}. \tag{11}$$

Оценка (11) неулучшаема, так как в ней достигается равенство, например, для $f(x) = (x-x_i)^2$.

Составная формула трапеций имеет вид

$$\int_a^b f(x)\,dx \approx \sum_{i=1}^{N} \frac{f(x_i)+f(x_{i-1})}{2}h =$$

$$= h(0{,}5f_0 + f_1 + \ldots + f_{N-1} + 0{,}5f_N), \tag{12}$$

где $f_i = f(x_i)$, $i = 0, 1, \ldots, N$, $hN = b-a$.

Погрешность этой формулы оценивается следующим образом:

$$|\Psi| \leqslant \frac{h^2 (b-a)}{12} M_2, \qquad M_2 = \max_{x \in [a,b]} |f''(x)|.$$

Таким образом, формула трапеций имеет, так же как и формула прямоугольников, второй порядок точности, $\Psi = O(h^2)$, но ее погрешность оценивается величиной в два раза большей (см. (9)).

4. Формула Симпсона. При аппроксимации интеграла (4) заменим функцию $f(x)$ параболой, проходящей через точки $(x_j, f(x_j))$, $j=i-1, i-0{,}5, i$, т. е. представим приближенно $f(x)$ в виде

$$f(x) \approx L_{2,i}(x), \qquad x \in [x_{i-1}, x_i],$$

где $L_{2,i}(x)$ — интерполяционный многочлен Лагранжа второй степени,

$$L_{2,i}(x) = \frac{2}{h^2} \{(x - x_{i-1/2})(x - x_i) f_{i-1} - \\ - 2(x - x_{i-1})(x - x_i) f_{i-1/2} + (x - x_i)(x - x_{i-1/2}) f_i\}. \quad (13)$$

Проводя интегрирование, получим

$$\int_{x_{i-1}}^{x_i} L_{2,i}(x)\, dx = \frac{h}{6} (f_{i-1} + 4f_{i-1/2} + f_i), \qquad h = x_i - x_{i-1}.$$

Таким образом, приходим к приближенному равенству

$$\int_{x_{i-1}}^{x_i} f(x)\, dx \approx \frac{h}{6} (f_{i-1} + 4f_{i-1/2} + f_i), \quad (14)$$

которое называется *формулой Симпсона* или *формулой парабол*.

На всем отрезке $[a, b]$ формула Симпсона имеет вид

$$\int_a^b f(x)\, dx \approx \sum_{i=1}^{N} \frac{h}{6} (f_{i-1} + 4f_{i-1/2} + f_i) = \\ = \frac{h}{6} [f_0 + f_N + 2(f_1 + f_2 + \ldots + f_{N-1}) + 4(f_{1/2} + f_{3/2} + \ldots + f_{N-1/2})].$$

Чтобы не использовать дробных индексов, можно обозначить

$$x_i = a + 0{,}5hi, \quad f_i = f(x_i), \quad i = 0, 1, \ldots, 2N, \quad hN = b - a$$

и записать формулу Симпсона в виде

$$\int_a^b f(x)\, dx \approx \frac{b-a}{6N} [f_0 + f_{2N} + 2(f_2 + f_4 + \ldots + f_{2N-2}) + \\ + 4(f_1 + f_3 + \ldots + f_{2N-1})]. \quad (15)$$

Прежде чем переходить к оценке погрешности формулы (14), заметим, что она является точной для любого многочлена третьей

степени, т. е. имеет место точное равенство

$$\int_{x_{i-1}}^{x_i} f(x)\,dx = \frac{h}{6}(f_{i-1} + 4f_{i-1/2} + f_i),$$

если $f(x) = a_0 + a_1x + a_2x^2 + a_3x^3$. Это утверждение нетрудно проверить непосредственно, что и предоставляется сделать читателю.

Для оценки погрешности формулы Симпсона воспользуемся интерполяционным многочленом Эрмита. Построим многочлен третьей степени $H_3(x)$ такой, что

$$H_3(x_{i-1}) = f(x_{i-1}), \qquad H_3(x_{i-1/2}) = f(x_{i-1/2}),$$
$$H_3'(x_{i-1/2}) = f'(x_{i-1/2}), \qquad H_3(x_i) = f(x_i).$$

Из § 3 гл. 3 известно, что такой многочлен существует и единствен. Он построен в явном виде в примере из п. 2 § 3 гл. 3. Однако нам даже не потребуется явный вид многочлена $H_3(x)$. Вспоминая, что формула Симпсона точна для любого многочлена третьей степени, получим

$$\int_{x_{i-1}}^{x_i} H_3(x)\,dx = \frac{h}{6}(H_3(x_{i-1}) + 4H_3(x_{i-1/2}) + H_3(x_i)) =$$
$$= \frac{h}{6}(f_{i-1} + 4f_{i-1/2} + f_i). \qquad (16)$$

Представим теперь $f(x)$ в виде

$$f(x) = H_3(x) + r_i(x), \qquad x \in [x_{i-1}, x_i], \qquad (17)$$

где $r_i(x)$ — погрешность интерполирования многочленом Эрмита $H_3(x)$. Интегрируя (17) и учитывая (16), получим

$$\int_{x_{i-1}}^{x_i} f(x)\,dx - \frac{h}{6}(f_{i-1} + 4f_{i-1/2} + f_i) = \int_{x_{i-1}}^{x_i} r_i(x)\,dx. \qquad (18)$$

Согласно (14) из § 3 гл. 3 имеем

$$r_i(x) = \frac{f^{\mathrm{IV}}(\zeta_i)}{24}(x - x_i)(x - x_{i-1/2})^2(x - x_{i-1}),$$

поэтому из (18) для погрешности ψ_i формулы (14) получаем оценку

$$|\psi_i| \leqslant \frac{M_{4,i}}{24}\int_{x_{i-1}}^{x_i} (x - x_i)(x - x_{i-1/2})^2(x - x_{i-1})\,dx,$$

где $M_{4,i} = \sup\limits_{x \in [x_{i-1}, x_i]} |f^{\mathrm{IV}}(x)|$.

Вычисляя интеграл, приходим окончательно к оценке

$$|\psi_i| \leqslant \frac{h^5}{2880} M_{4,i}. \qquad (19)$$

Погрешность составной формулы Симпсона (15) оценивается так:

$$|\Psi| \leqslant \frac{h^4 (b-a)}{2880} M_4, \quad hN = b-a, \quad M_4 = \sup_{x \in [a,b]} |f^{IV}(x)|.$$

Отсюда видно, что формула Симпсона существенно точнее, чем формулы прямоугольников и трапеций. На частичном отрезке она имеет точность $O(h^5)$, а на всем отрезке — $O(h^4)$.

Приведем вывод формулы Симпсона, основанный на методе экстраполяции. *Метод экстраполяции* состоит в следующем. Проведем два расчета по формуле трапеций (12), первый расчет с шагом h, когда вычисляется сумма

$$I_h = \sum_{i=1}^{N} \frac{f_i + f_{i-1}}{2} h, \quad f_i = f(x_i), \quad x_i = a + ih,$$

$$h = (b-a)/N,$$

и второй расчет — с шагом $0{,}5h$, когда вычисляется сумма

$$I_{h/2} = \sum_{i=1}^{N} \left[\left(\frac{f_{i-1} + f_{i-1/2}}{2} \right) + \left(\frac{f_{i-1/2} + f_i}{2} \right) \right] \frac{h}{2} =$$

$$= \sum_{i=1}^{N} (f_{i-1} + 2f_{i-1/2} + f_i) \frac{h}{4}, \qquad f_{i-1/2} = f(x_i - 0{,}5h).$$

Используя разложение по формуле Тейлора, можно показать, что для достаточно гладкой функции $f(x)$ справедливо равенство

$$I_h = I + c_1 h^2 + O(h^4),$$

где I — исходный интеграл (1) и c_1 — постоянная, не зависящая от h. Точно так же

$$I_{h/2} = I + c_1 \left(\frac{h}{2} \right)^2 + O(h^4).$$

Отсюда получим, что

$$I_{h/2} - \frac{1}{4} I_h = \frac{3}{4} I + O(h^4),$$

т. е. выражение

$$J_h = \frac{4}{3} I_{h/2} - \frac{1}{3} I_h$$

совпадает с интегралом I с точностью до величин $O(h^4)$.

В данном примере не обязательно проводить расчет на двух сетках, так как можно построить явное выражение для суммы J_h. Действительно,

$$4I_{h/2} - I_h = \sum_{i=1}^{N} (f_{i-1} + 2f_{i-1/2} + f_i) h - \sum_{i=1}^{N} \frac{f_i + f_{i-1}}{2} h =$$

$$= \sum_{i=1}^{N} \frac{f_{i-1} + 4f_{i-1/2} + f_i}{2} h,$$

$$J_h = \sum_{i=1}^{N} \frac{f_{i-1} + 4f_{i-1/2} + f_i}{6} h.$$

Таким образом, снова получим квадратурную формулу Симпсона.

5. Апостериорная оценка погрешности методом Рунге. Автоматический выбор шага интегрирования. Величина погрешности численного интегрирования зависит как от шага сетки h, так и от гладкости подынтегральной функции $f(x)$. Например, в оценку (11), наряду с h, входит величина

$$M_{2,i} = \max_{x \in [x_{i-1}, x_i]} |f''(x)|,$$

которая может сильно меняться от точки к точке и, вообще говоря, заранее неизвестна. Если величина погрешности велика, то ее можно уменьшить путем измельчения сетки на данном отрезке $[x_{i-1}, x_i]$. Для этого прежде всего надо уметь апостериорно, т. е. после проведения расчета, оценивать погрешность.

Апостериорную оценку погрешности можно осуществить *методом Рунге*, который мы поясним сначала на примере формулы трапеций. Пусть отрезок $[a, b]$ разбит на частичные отрезки $[x_{i-1}, x_i]$, $i=1, 2, \ldots, N$, $x_0=a$, $x_N=b$, имеющие, может быть, разную длину $h_i=x_i-x_{i-1}$. На каждом частичном отрезке применяется формула трапеций

$$I_i = \int_{x_{i-1}}^{x_i} f(x)\,dx \approx \frac{f_i + f_{i-1}}{2} h_i = I_{h,i}.$$

Согласно (11) имеем

$$I_i - I_{h,i} \approx c_i h_i^3, \tag{20}$$

где константа c_i зависит от гладкости $f(x)$ и заранее неизвестна. Измельчим на отрезке $[x_{i-1}, x_i]$ сетку в два раза и повторим расчет с шагом $0{,}5h$, т. е. вычислим сумму

$$I_{h/2,i} = (f_{i-1} + 2f_{i-1/2} + f_i)\frac{h_i}{4}.$$

Тогда согласно (20) будем иметь

$$I_i - I_{h/2,i} \approx c_i \left(\frac{h_i}{2}\right)^3. \tag{21}$$

Из соотношений (20), (21) можно исключить константу c_i и получить оценку погрешности, которая содержит лишь известные величины $I_{h,i}$, $I_{h/2,i}$:

$$I_i - I_{h,i} \approx \frac{8}{7}(I_{h/2,i} - I_{h,i}),$$

$$I_i - I_{h/2,i} = \frac{1}{8}(I_i - I_{h,i}) \approx \frac{1}{7}(I_{h/2,i} - I_{h,i}).$$

Разумеется, метод Рунге можно применять и для оценки погрешности других квадратурных формул. Пусть какая-то квадратурная формула имеет на частичном отрезке порядок точности m, т. е. $I_i - I_{h,i} \approx c_i h_i^m$. Тогда

$$I_i - I_{h/2,i} \approx c_i \left(\frac{h_i}{2}\right)^m,$$

откуда получим

$$I_i - I_{h,i} \approx 2^m (I_i - I_{h/2,i}), \tag{22}$$

$$I_i - I_{h/2,i} \approx \frac{I_{h/2,i} - I_{h,i}}{2^m - 1}. \tag{23}$$

Возможность апостериорно оценивать погрешность позволяет вычислять интеграл (1) с заданной точностью $\varepsilon > 0$ путем автоматического выбора шага интегрирования h_i. Пусть используется составная квадратурная формула

$$I \approx I_h = \sum_{i=1}^{N} I_{h,i},$$

где $I_{h,i}$ — квадратурная сумма на частичном отрезке, причем на каждом частичном отрезке используется одна и та же квадратурная формула (например, формула трапеций, Симпсона и др.). Проведем на каждом частичном отрезке $[x_{i-1}, x_i]$ все вычисления дважды, один раз — с шагом h_i и второй раз — с шагом $0{,}5h_i$ и оценим погрешность по правилу Рунге (23).

Если для заданного $\varepsilon > 0$ будут выполняться неравенства

$$|I_i - I_{h/2,i}| \approx \frac{|I_{h/2,i} - I_{h,i}|}{2^m - 1} \leqslant \frac{\varepsilon h_i}{b-a}, \quad i = 1, 2, \ldots, N, \tag{24}$$

то получим

$$|I - I_{h/2}| \leqslant \frac{\varepsilon}{b-a} \sum_{i=1}^{N} h_i = \varepsilon,$$

т. е. будет достигнута заданная точность ε.

Если же на каком-то из частичных отрезков оценка (24) не будет выполняться, то шаг на этом отрезке надо измельчить еще в два раза и снова оценить погрешность. Измельчение сетки на данном отрезке следует проводить до тех пор, пока не будет достигнута оценка вида (24). Заметим, что для некоторых функций $f(x)$ такое измельчение может продолжаться слишком долго. Поэтому в соответствующей программе следует предусмотреть ограничение сверху на число измельчений, а также возможность увеличения ε.

Таким образом, автоматический выбор шага интегрирования приводит к тому, что интегрирование ведется с крупным шагом на участках плавного изменения функции $f(x)$ и с мелким шагом — на участках быстрого изменения $f(x)$. Это позволяет при заданной точности ε уменьшить количество вычислений значений $f(x)$ по сравнению с расчетом на сетке с постоянным шагом. Подчеркнем, что для нахождения сумм $I_{h/2,i}$ не надо пересчитывать значения $f(x)$ во всех узлах, достаточно вычислять $f(x)$ только в новых узлах.

6. Экстраполяция Ричардсона. Способ повышения точности квадратурной формулы, рассмотренный в конце п. 4, можно обобщить на случай многократного измельчения сетки.

Предположим, что для вычисления интеграла (1) отрезок $[a, b]$ разбит на N равных отрезков длины $h=(b-a)/N$ и на каждом частичном отрезке применяется одна и та же квадратурная формула. Тогда исходный интеграл I заменяется некоторой квадратурной суммой I_h, причем возникающая погрешность зависит от шага сетки h. Для некоторых квадратурных формул удается получить разложение погрешности I_h-I по степеням h. Предположим, что для данной квадратурной суммы I_h существует разложение

$$I_h = I + a_1 h^{\alpha_1} + a_2 h^{\alpha_2} + \ldots + a_m h^{\alpha_m} + O(h^{\alpha_{m+1}}), \qquad (25)$$

где $0<\alpha_1<\alpha_2<\ldots<\alpha_m<\alpha_{m+1}$ и коэффициенты a_i не зависят от h. Подчеркнем, что получение подобных разложений является трудной задачей анализа и здесь не рассматривается. Явный вид коэффициентов a_i нам не потребуется, однако величины α_i предполагаются известными.

Вычислим приближенно значение интеграла I по данной квадратурной формуле на последовательности сеток с шагами $h_0=h$, h_1, h_2, ..., h_m. Для определенности будем предполагать, что сетка измельчается по геометрической прогрессии, т. е. $h_k=q^k h_0$, $k=0, 1, \ldots, m$, где $q\in(0, 1)$. Вычисляя квадратурную сумму I_h при различных значениях h, получим величины I_{h_k}, $k=0, 1, \ldots, m$, причем согласно (25) будем иметь

$$I_{h_k} = I + a_1 h_k^{\alpha_1} + a_2 h_k^{\alpha_2} + \ldots + a_m h_k^{\alpha_k} + O(h_k^{\alpha_{m+1}}). \qquad (26)$$

Обозначим $I^{(0)}=I$, $I^{(1)}_{h_k}=I_{h_k}$. Исключая коэффициент a_1 из соотношений

$$I^{(1)}_{h_{k-1}} = I + a_1 h_{k-1}^{\alpha_1} + O(h_{k-1}^{\alpha_2}),$$
$$I^{(1)}_{h_k} = I + a_1 h_k^{\alpha_1} + O(h_k^{\alpha_2}),$$

получим

$$I = I^{(2)}_{h_{k-1}} + O(h_{k-1}^{\alpha_2}), \qquad (27)$$

где обозначено

$$I^{(2)}_{h_{k-1}} = I^{(1)}_{h_{k-1}} + \frac{I^{(1)}_{h_k} - I^{(1)}_{h_{k-1}}}{1-q^{\alpha_1}}. \qquad (28)$$

По формулам (28) можно вычислить величины $I^{(2)}_{h_k}$, $k=0, 1, \ldots, m-1$. Согласно (27), они дают более точное, чем $I^{(1)}_{h_k}$, приближение к интегралу I. Этот процесс повышения точности можно продолжить, вычисляя величины $I^{(j)}_{h_k}$ с помощью рекуррентных соотношений

$$I^{(j+1)}_{h_{k-1}} = I^{(j)}_{h_{k-1}} + \frac{I^{(j)}_{h_k} - I^{(j)}_{h_{k-1}}}{1-q^{\alpha_j}}, \qquad (29)$$
$$j=1, 2, \ldots, m, \quad k=1, 2, \ldots, m-j+1,$$
$$I^{(1)}_{h_k} = I_{h_k}, \quad k=0, 1, \ldots, m.$$

Лемма 1. *Пусть для квадратурной суммы I_h справедливо разложение* (25) *и сетка измельчается по правилу $h_k=q^k h$, $k=0, 1, \ldots, m$. Тогда для величин $I^{(j)}_{h_{k-1}}$, определенных согласно* (29), *справедливы разложения*

$$I^{(j)}_{h_{k-1}} = I + b_j^{(j)} h_{k-1}^{\alpha_j} + b_{j+1}^{(j)} h_{k-1}^{\alpha_{j+1}} + \ldots + b_m^{(j)} h_{k-1}^{\alpha_m} + O\left(h_{k-1}^{\alpha_{m+1}}\right),$$
$$j=1, 2, \ldots, m, \quad k=1, 2, \ldots, m-j+1, \tag{30}$$

где коэффициенты $b_i^{(j)}$ не зависят от сетки.

Доказательство. Проведем его индукцией по j. При $j=1$ равенство (30) выполняется с $b_i^{(j)}=a_i$ согласно (25). Предположим, что равенство (30) выполняется при $j=l$ и докажем, что оно выполняется при $j=l+1$.

Имеем

$$I^{(l)}_{h_{k-1}} = I + \sum_{i=l}^{m} b_i^{(l)} h_{k-1}^{\alpha_i} + O\left(h_{k-1}^{\alpha_{m+1}}\right),$$
$$I^{(l)}_{h_k} = I + \sum_{i=l}^{m} b_i^{(l)} h_k^{\alpha_i} + O\left(h_k^{\alpha_{m+1}}\right) \tag{31}$$

и, следовательно,

$$I^{(l)}_{h_k} - I^{(l)}_{h_{k-1}} = \sum_{i=l}^{m} b_i^{(l)} \left(h_k^{\alpha_i} - h_{k-1}^{\alpha_i}\right) + O\left(h_{k-1}^{\alpha_{m+1}}\right).$$

Далее, подставляя полученную разность в (29) при $j=l$ и учитывая (31), получим

$$I^{(l+1)}_{h_{k-1}} = I + \sum_{i=l}^{m} b_i^{(l)} h_{k-1}^{\alpha_i} + \frac{1}{1-q^{\alpha_l}} \sum_{i=l}^{m} b_i^{(l)} \left(h_k^{\alpha_i} - h_{k-1}^{\alpha_i}\right) + O\left(h_{k-1}^{\alpha_{m+1}}\right),$$

т. е.

$$I^{(l+1)}_{h_{k-1}} = I + \sum_{i=l}^{m} b_i^{(l)} h_{k-1}^{\alpha_i} \left(\frac{q^{\alpha_l} - q^{\alpha_i}}{1-q^{\alpha_l}}\right) + O\left(h_{k-1}^{\alpha_{m+1}}\right).$$

Отсюда получаем

$$I^{(l+1)}_{h_{k-1}} = I + \sum_{i=l+1}^{m} b_i^{(l+1)} h_{k-1}^{\alpha_i} + O\left(h_{k-1}^{\alpha_{m+1}}\right),$$

т. е. равенство (30) выполняется с $j=l+1$, причем

$$b_i^{(l+1)} = \frac{q^{\alpha_i} - q^{\alpha_l}}{1-q^{\alpha_l}} b_i^{(l)}, \quad i=l+1, \ldots, m.$$

Лемма 1 доказана.

Из леммы следует, что суммы $I_{h_k}^{(j)}$ совпадают с интегралом I с точностью до величин $O(h^{k\alpha_j})$, т. е. порядок точности повышается по сравнению с исходной формулой в α_j/α_1 раз.

Изложенный метод повышения точности называется *методом экстраполяции Ричардсона*. Его можно применять не только к квадратурным формулам, но и к самым различным сеточным функциям, если только для них существуют асимптотические разложения по степеням h. Подробное изложение метода экстраполяции по Ричардсону содержится в книге [23]. Применительно к формуле трапеций данный метод называется *методом Ромберга*. Существуют стандартные программы вычисления определенных интегралов методом Ромберга. Пример, приведенный в конце п. 4, является частным случаем метода (29), когда I_h — квадратурная сумма, соответствующая методу трапеций, $m=1$, $q=0{,}5$.

Отметим еще, что для формулы трапеций

$$I \approx I_h = \sum_{i=1}^{N} 0{,}5\,(f_i + f_{i-1})\,h, \quad f_i = f(x_i),$$

разложение (25) имеет вид

$$I_h = \int_a^b f(x)\,dx + \frac{h^2}{12}(f'(b) - f'(a)) - \frac{h^4}{720}(f'''(b) - f'''(a)) +$$
$$+ \frac{h^6}{30240}(f^{(5)}(b) - f^{(5)}(a)) + \ldots + c_{2r}h^{2r}(f^{(2r-1)}(b) - f^{(2r-1)}(a)) + O(h^{2r+2}). \tag{32}$$

Здесь коэффициенты c_{2r} совпадают с коэффициентами разложения функции

$$G(h) = \frac{h}{2}\,\frac{e^h+1}{e^h-1}$$

в ряд Тейлора:

$$G(h) = 1 + c_2h^2 + c_4h^4 + \ldots + c_{2r}h^{2r} + \ldots$$

Доказательство формулы (32), называемой *формулой Эйлера*, можно найти, например, в [2, с. 165].

§ 2. Квадратурные формулы интерполяционного типа

1. Вывод формул. Будем рассматривать формулы приближенного вычисления интегралов

$$\int_a^b \rho(x) f(x)\,dx, \tag{1}$$

где $\rho(x) > 0$ — заданная интегрируемая функция (так называемая весовая функция) и $f(x)$ — достаточно гладкая функция. Рассма-

триваемые далее формулы имеют вид

$$\int_a^b \rho(x) f(x)\, dx \approx \sum_{k=0}^{n} c_k f(x_k), \tag{2}$$

где $x_k \in [a, b]$ и c_k — числа, $k=0, 1, \ldots, n$.

В отличие от предыдущего параграфа, не будем разбивать отрезок $[a, b]$ на частичные отрезки, а получим квадратурные формулы путем замены $f(x)$ интерполяционным многочленом сразу на всем отрезке $[a, b]$. Полученные таким образом формулы называются *квадратурными формулами интерполяционного типа.* Как правило, точность этих формул возрастает с увеличением числа узлов интерполирования. Рассмотренные в § 1 формулы прямоугольников, трапеций и Симпсона являются частными случаями квадратурных формул интерполяционного типа, когда $n=0, 1, 2$, $\rho(x) \equiv 1$.

Получим выражения для коэффициентов квадратурных формул интерполяционного типа. Пусть на отрезке $[a, b]$ заданы узлы интерполирования x_k, $k=0, 1, \ldots, n$. Предполагается, что среди этих узлов нет совпадающих, в остальном они могут быть расположены как угодно на $[a, b]$.

Заменяя в интеграле (1) функцию $f(x)$ интерполяционным многочленом Лагранжа

$$L_n(x) = \sum_{k=0}^{n} \frac{\omega(x)}{(x - x_k)\,\omega'(x_k)} f(x_k), \tag{3}$$

где

$$\omega(x) = \prod_{j=0}^{n} (x - x_j), \quad \omega'(x_k) = \prod_{j \neq k} (x_k - x_j),$$

получим приближенную формулу (2), где

$$c_k = \int_a^b \frac{\rho(x)\,\omega(x)}{(x - x_k)\,\omega'(x_k)}\, dx, \qquad k = 0, 1, \ldots, n. \tag{4}$$

Таким образом, *формула* (2) *является квадратурной формулой интерполяционного типа тогда и только тогда, когда ее коэффициенты вычисляются по правилу* (4).

Приведем пример квадратурной формулы, не являющейся формулой интерполяционного типа. Рассмотрим интеграл

$$\int_0^1 f(x)\, dx \tag{5}$$

и выберем в качестве узлов точки $x_0=0$, $x_1=0{,}5$, $x_2=1$.

Квадратурная формула интерполяционного типа, построенная по заданным узлам, совпадает с формулой Симпсона

$$\int_0^1 f(x)\, dx \approx \frac{1}{6}\left(f(0) + 4f(0{,}5) + f(1)\right). \tag{6}$$

Заменим теперь в (5) функцию $f(x)$ многочленом $\varphi(x)$ наилучшего среднеквадратичного приближения первой степени. Согласно (18) из § 5 гл. 3 этот многочлен имеет вид

$$\varphi(x) = (f(1) - f(0))(x - 0{,}5) + \frac{1}{3}(f(0) + f(0{,}5) + f(1)).$$

Отсюда приходим к квадратурной формуле

$$\int_0^1 f(x)\,dx \approx \frac{1}{3}(f(0) + f(0{,}5) + f(1)), \tag{7}$$

не совпадающей с (6).

2. Оценка погрешности. Получим выражение для погрешности квадратурной формулы интерполяционного типа. Представим функцию $f(x)$ в виде

$$f(x) = L_n(x) + r_{n+1}(x),$$

где $L_n(x)$ — интерполяционный многочлен для $f(x)$, построенный по узлам $x_0, x_1, \ldots, x_n$ и $r_{n+1}(x)$ — погрешность интерполирования. Тогда получим

$$\int_a^b \rho(x) f(x)\,dx = \int_a^b \rho(x) L_n(x)\,dx + \int_a^b \rho(x) r_{n+1}(x)\,dx =$$

$$= \sum_{k=0}^{n} c_k f(x_k) + \int_a^b \rho(x) r_{n+1}(x)\,dx.$$

Таким образом, погрешность ψ_{n+1} квадратурной формулы (2), (4) равна

$$\psi_{n+1} = \int_a^b \rho(x) r_{n+1}(x)\,dx, \tag{8}$$

где $r_{n+1}(x)$ — погрешность интерполирования.

Вспоминая выражение для погрешности интерполирования (см. (3) из § 2 гл. 3)

$$r_{n+1}(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}\omega(x),$$

получаем

$$\psi_{n+1} = \frac{1}{(n+1)!}\int_a^b \rho(x)\omega(x) f^{(n+1)}(\xi(x))\,dx. \tag{9}$$

Отсюда приходим к следующей оценке погрешности квадратурной формулы интерполяционного типа:

$$|\psi_{n+1}| \leqslant \frac{M_{n+1}}{(n+1)!}\int_a^b \rho(x)|\omega(x)|\,dx, \tag{10}$$

где $M_{n+1} = \max\limits_{x\in[a,b]} |f^{(n+1)}(x)|$. Из формулы (10) видно, что справедливо следующее утверждение.

Квадратурная формула интерполяционного типа, построенная по $n+1$ узлу $x_0, x_1, \ldots, x_n$, является точной для любого многочлена степени n, т. е. если $f(x)$ — многочлен степени n и c_k — коэффициенты, вычисленные согласно (4), то имеет место точное равенство

$$\int_a^b \rho(x) f(x)\, dx = \sum_{k=0}^{n} c_k f(x_k). \tag{11}$$

Справедливо и обратное утверждение.

Теорема 1. *Если квадратурная формула*

$$\int_a^b \rho(x) f(x)\, dx \approx \sum_{k=0}^{n} d_k f(x_k) \tag{12}$$

точна для любого многочлена степени n, то она является квадратурной формулой интерполяционного типа.

Доказательство. Достаточно показать, что $d_k = c_k$, где c_k определены согласно (4), $k=0, 1, \ldots, n$. Рассмотрим многочлены

$$\varphi_k(x) = \frac{\omega(x)}{(x-x_k)\,\omega'(x_k)}, \qquad k = 0, 1, \ldots, n,$$

имеющие степень n, и вычислим интегралы

$$I_k = \int_a^b \rho(x)\, \varphi_k(x)\, dx.$$

По условию теоремы справедливы точные равенства

$$I_k = \sum_{l=0}^{n} d_l \varphi_k(x_l).$$

Поскольку

$$\varphi_k(x_l) = \begin{cases} 0, & k \neq l, \\ 1, & k = l, \end{cases}$$

получаем $I_k = d_k$, $k=0, 1, \ldots, n$.

С другой стороны, согласно (4) имеем

$$I_k = \int_a^b \rho(x)\, \frac{\omega(x)}{(x-x_k)\,\omega'(x_k)} = c_k.$$

Таким образом, $d_k = c_k$, $k=0, 1, \ldots, n$, что и требовалось.

3. Симметричные формулы. Для некоторых квадратурных формул оценка погрешности (10) является грубой, так как она не учитывает симметрии формул.

Рассмотрим, например, формулу Симпсона

$$\int_{-1}^{1} f(x)\, dx \approx \frac{1}{3}\left(f(-1) + 4f(0) + f(1)\right) \tag{13}$$

для функции $f(x)=x^3$. В данном случае имеем $n=2$, $f^{(n+1)}(x)=6$, поэтому согласно (9) погрешность можно представить в виде

$$\psi_3=\int_{-1}^{1}\omega(x)\,dx,$$

где $\omega(x)=x(x^2-1)$. Благодаря симметричному расположению узлов имеем $\psi_3=0$. В то же время правая часть неравенства (10), равная

$$\int_{-1}^{1}|\omega(x)|\,dx=\frac{1}{2},$$

отлична от нуля. Таким образом, оценка (10) не является точной для формулы Симпсона.

Квадратурная формула (2) называется *симметричной*, если

1) n четно;

2) узлы расположены симметрично относительно середины отрезка $[a, b]$, т. е.

$$\frac{a+b}{2}-x_k=x_{n-k}-\frac{a+b}{2},\qquad k=0, 1, \ldots, n/2; \tag{14}$$

3)
$$c_k=c_{n-k},\qquad k=0, 1, \ldots, n/2. \tag{15}$$

Свойство (15) коэффициентов квадратурной формулы определяется не только симметричным расположением узлов, но и симметрией весовой функции $\rho(x)$. Говорят, что $\rho(x)$ — четная функция относительно середины отрезка $[a, b]$, если

$$\rho\left(\frac{a+b}{2}+x\right)=\rho\left(\frac{a+b}{2}-x\right) \tag{16}$$

для всех $x\in[0, (b-a)/2]$.

Лемма 1. *Если c_k определены согласно* (4) *и n четно, то соотношения* (15) *являются следствием условий* (14), (16).

Доказательство. Покажем сначала, что

$$\omega(x)=\prod_{k=0}^{n}(x-x_k) \tag{17}$$

является нечетной функцией относительно точки

$$x_{n/2}=(a+b)/2.$$

Имеем

$$\omega(x)=(x-x_{n/2})\prod_{k=0}^{n/2-1}(x-x_k)(x-x_{n-k}),$$

откуда, учитывая условия (14), получим

$$\omega(x)=(x-x_{n-2})\prod_{k=0}^{n/2-1}[(x-x_{n/2})^2-(x_k-x_{n/2})^2]. \tag{18}$$

Следовательно, при любом t

$$\omega(x_{n/2}+t)=t\prod_{k=0}^{n/2-1}[t^2-(x_k-x_{n/2})^2]=-\omega(x_{n/2}-t),$$

т. е. функция $\omega(x)$ нечетна относительно точки $x_{n/2}$.

Из формулы (18) следует также, что $\omega'(x)$ — четная функция относительно точки $x_{n/2}$, поэтому

$$\omega'(x_{n-k})=\omega'(x_k). \tag{19}$$

Рассмотрим теперь разность

$$c_k-c_{n-k}=\int_a^b \rho(x)\,\omega(x)\,\mu(x)\,dx, \tag{20}$$

где

$$\mu(x)=\frac{1}{(x-x_k)\,\omega'(x_k)}-\frac{1}{(x-x_{n-k})\,\omega'(x_{n-k})}.$$

Учитывая (19) видим, что

$$\mu(x)=\frac{1}{\omega'(x_k)}\cdot\frac{x_n-x_{n-k}}{(x-x_k)(x-x_{n-k})}=\frac{1}{\omega'(x_k)}\cdot\frac{x_k-x_{n-k}}{[(x-x_{n/2})^2-(x_{n/2}-x_k)^2]},$$

откуда следует четность $\mu(x)$ относительно точки $x_{n/2}$.

Таким образом, подынтегральная функция в (20) нечетна относительно середины отрезка $[a, b]$, и, следовательно, интеграл равен нулю. Лемма 1 доказана.

Покажем теперь, что наличие симметрии повышает точность квадратурных формул. Справедлива

Теорема 2. *Пусть $\rho(x)$ — четная функция относительно точки $(a+b)/2$ и пусть выполнены условия* (14), *где n — четное число. Тогда, если квадратурная формула интерполяционного типа* (2) *точна для любого многочлена степени n, то она точна и для любого многочлена степени $n+1$.*

Доказательство. Достаточно показать, что формула точна для многочлена

$$f(x)=(x-x_{n/2})^{n+1},\qquad x_{n/2}=0{,}5(a+b).$$

Поскольку

$$\int_a^b \rho(x)\,(x-x_{n/2})^{n+1}\,dx=0$$

вследствие нечетности подынтегральной функции, необходимо доказать, что

$$I_n=\sum_{k=0}^{n} c_k f(x_k)=0.$$

Представим I_n в виде суммы $I_n^{(1)}+I_n^{(2)}$, где

$$I_n^{(1)}=\sum_{k=0}^{n/2-1} c_k\,(x_k-x_{n/2})^{n+1},$$

$$I_n^{(2)}=\sum_{k=n/2+1}^{n} c_k\,(x_k-x_{n/2})^{n+1}.$$

Из условий (14) получим

$$I_n^{(2)} = \sum_{k=n/2+1}^{n} c_k (x_{n/2} - x_{n-k})^{n+1}$$

или

$$I_n^{(2)} = \sum_{l=0}^{n/2-1} c_{n-l} (x_{n/2} - x_l)^{n+1} = -\sum_{k=0}^{n/2-1} c_{n-k} (x_k - x_{n/2})^{n+1}.$$

Последнее равенство справедливо в силу того, что n — четное число.

Таким образом, получаем

$$I_n = I_n^{(1)} + I_n^{(2)} = \sum_{k=0}^{n/2-1} (c_k - c_{n-k}) (x_k - x_{n/2})^{n+1}.$$

Согласно лемме 1 имеем $c_k = c_{n-k}$, $k=0, 1, \ldots, n/2-1$, т. е. $I_n = 0$, что и завершает доказательство теоремы 2.

4. Формулы Ньютона — Котеса. Численная устойчивость квадратурных формул. *Формулами Ньютона — Котеса* называются квадратурные формулы интерполяционного типа, построенные на равномерной сетке, когда $x_k - x_{k-1} = h$, $k = 1, 2, \ldots, n$.

Различают два типа формул Ньютона — Котеса: формулы замкнутого типа и формулы открытого типа. В *формулах замкнутого типа* $x_0 = a$ и $x_n = b$, а в *формулах открытого типа* хотя бы один из узлов x_0 или x_n не совпадает с соответствующей граничной точкой отрезка $[a, b]$. Для простоты изложения рассмотрим лишь случай формул замкнутого типа, когда $x_k = a + kh$, $k = 0, 1, \ldots, n$, $hn = b - a$.

В случае равномерной сетки можно упростить выражения для коэффициентов квадратурных формул. В формуле (4) сделаем замену $x = a + th$, $0 \leqslant t \leqslant n$. Простые выкладки, которые мы опускаем, показывают, что в результате замены формула (4) примет вид

$$c_k = (b - a) b_k^{(n)},$$

где

$$b_k^{(n)} = \frac{(-1)^{n-k}}{k!\,(n-k)!} \left(\frac{1}{n} \int_0^n \rho(a + th) \frac{t(t-1)\ldots(t-n)}{t-k} dt \right). \qquad (21)$$

Отметим, что формулы Ньютона — Котеса с $n \geqslant 10$ редко используются из-за их численной неустойчивости, приводящей к резкому возрастанию вычислительной погрешности. Причиной такой неустойчивости является то, что коэффициенты формул Ньютона — Котеса при больших n имеют различные знаки, а именно при $n \geqslant 10$, $\rho(x) \equiv 1$ существуют как положительные, так и отрицательные коэффициенты.

Остановимся подробнее на значении знакопостоянства коэффициентов для устойчивости вычислений.

Рассмотрим квадратурную сумму

$$I_n = \sum_{k=0}^{n} c_k f(x_k). \qquad (22)$$

Предположим, что значения функции $f(x)$ вычисляются с некоторой погрешностью, т. е. вместо точного значения получаем приближенное значение $\tilde{f}(x_k) = f(x_k) + \delta_k$. Тогда вместо I_n получим сумму

$$\tilde{I}_n = \sum_{k=0}^{n} c_k (f(x_k) + \delta_k) = I_n + \delta I_n,$$

где

$$\delta I_n = \sum_{k=0}^{n} c_k \delta_k. \tag{23}$$

Поскольку квадратурная формула (2), (4) точна для $f(x) \equiv 1$, имеем

$$\sum_{k=0}^{n} c_k = \int_a^b \rho(x)\, dx,$$

т. е. при $\rho(x) > 0$ сумма

$$\sum_{k=0}^{n} c_k = M \tag{24}$$

ограничена числом $M > 0$, не зависящим от n.

Предположим теперь, что все коэффициенты c_k неотрицательны. Тогда из (23), (24) получим оценку

$$|\delta I_n| \leqslant \sum_{k=0}^{n} |c_k|\,|\delta_k| = \sum_{k=0}^{n} c_k |\delta_k| \leqslant \left(\max_{0 \leqslant k \leqslant n} |\delta_k|\right) M, \tag{25}$$

которая означает, что при больших n погрешность в вычислении квадратурной суммы (22) имеет тот же порядок, что и погрешность в вычислении функции. В этом случае говорят, что сумма (22) вычисляется устойчиво.

Если коэффициенты c_k имеют различные знаки, то может оказаться, что сумма

$$\sum_{k=0}^{n} |c_k|$$

не является равномерно ограниченной по n и, следовательно, погрешность в вычислении I_n неограниченно возрастает с ростом n. В этом случае вычисления по формуле (22) будут неустойчивы и пользоваться такой формулой при больших n нельзя.

Таким образом, если необходимо сосчитать интеграл (1) более точно, то имеются две возможности. Во-первых, можно разбить весь отрезок $[a, b]$ на несколько частичных отрезков и на каждом из частичных отрезков применить формулу Ньютона — Котеса с небольшим числом узлов. Полученные таким образом формулы называются *составными квадратурными формулами.* Они часто применяются на практике, хотя и не являются достаточно экономичными, поскольку требуют многократного вычисления значений функции

$f(x)$. Во-вторых, можно попытаться выбрать узлы квадратурной формулы так, чтобы полученная формула имела большую точность, чем формула Ньютона — Котеса с тем же числом узлов. В следующем параграфе рассматривается один из методов, основанных на выборе узлов квадратурной формулы, а именно метод Гаусса. Он приводит к квадратурным формулам с положительными коэффициентами при любых n и существенно более точным, нежели формулы Ньютона — Котеса.

§ 3. Метод Гаусса вычисления определенных интегралов

1. Постановка задачи. В предыдущем параграфе предполагалось, что узлы квадратурных формул заданы заранее. Было показано, что если использовать n узлов интерполяции, то получим квадратурные формулы, точные для алгебраических многочленов степени n—1. Оказывается, что за счет выбора узлов можно получить квадратурные формулы, которые будут точными и для многочленов степени выше n—1. Рассмотрим следующую задачу: построить квадратурную формулу

$$\int_a^b \rho(x) f(x)\, dx \approx \sum_{k=1}^{n} c_k f(x_k), \tag{1}$$

которая при заданном n была бы точна для алгебраического многочлена возможно большей степени. Обратим внимание, что здесь в отличие от § 2 для удобства изложения нумерация узлов начинается с $k=1$.

В настоящем параграфе будет показано, что такие квадратурные формулы существуют. Они называются *квадратурными формулами наивысшей алгебраической степени точности* или *формулами Гаусса*. Эти формулы точны для любого алгебраического многочлена степени $2n$—1.

Итак, потребуем, чтобы квадратурная формула (1) была точна для любого алгебраического многочлена степени m. Это эквивалентно требованию, чтобы формула была точна для функций $f(x) = x^\alpha$, $\alpha = 0, 1, \ldots, m$. Отсюда получаем условия

$$\sum_{k=1}^{n} c_k x_k^\alpha = \int_a^b \rho(x)\, x^\alpha\, dx, \qquad \alpha = 0, 1, \ldots, m, \tag{2}$$

которые представляют собой нелинейную систему $m+1$ уравнений относительно $2n$ неизвестных

$$c_1, c_2, \ldots, c_n; \quad x_1, x_2, \ldots, x_n.$$

Для того чтобы число уравнений равнялось числу неизвестных, надо потребовать $m = 2n-1$. В дальнейшем будет показано, что система (2) при $m = 2n-1$ имеет единственное решение. Однако сначала рассмотрим несколько частных случаев, когда решение системы (2) можно найти непосредственно.

Пусть $\rho(x) \equiv 1$, $a = -1$, $b = 1$. При $n = 1$ получаем $m = 1$ и система (2) принимает вид

$$c_1 = \int_{-1}^{1} dx = 2, \quad c_1 x_1 = \int_{-1}^{1} x\, dx = 0,$$

т. е. приходим к известной формуле прямоугольников

$$\int_{-1}^{1} f(x)\, dx \approx 2f(0),$$

которая точна для любого многочлена первой степени.

При $n = 2$, $m = 3$ система (2) записывается в виде

$$c_1 + c_2 = 2, \quad c_1 x_1 + c_2 x_2 = 0,$$
$$c_1 x_1^2 + c_2 x_2^2 = \frac{2}{3}, \quad c_1 x_1^3 + c_2 x_2^3 = 0.$$

Отсюда находим

$$c_1 = c_2 = 1, \quad x_1 = -\frac{1}{\sqrt{3}}, \quad x_2 = \frac{1}{\sqrt{3}},$$

т. е. получаем квадратурную формулу

$$\int_{-1}^{1} f(x)\, dx \approx f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right),$$

которая точна для любого алгебраического многочлена третьей степени.

2. Основная теорема. Возвращаясь к рассмотрению квадратурных формул (1) общего вида, введем многочлен

$$\omega(x) = (x - x_1)(x - x_2) \ldots (x - x_n). \tag{3}$$

Будем предполагать, что $\rho(x) > 0$.

Теорема 1. *Квадратурная формула* (1) *точна для любого многочлена степени* $m = 2n - 1$ *тогда и только тогда, когда выполнены два условия:*

1) *многочлен* $\omega(x)$ *ортогонален с весом* $\rho(x)$ *любому многочлену* $q(x)$ *степени меньше* n, *т. е.*

$$\int_a^b \rho(x)\, \omega(x)\, q(x)\, dx = 0; \tag{4}$$

2) *формула* (1) *является квадратурной формулой интерполяционного типа, т. е.*

$$c_k = \int_a^b \rho(x)\, \frac{\omega(x)}{(x - x_k)\, \omega'(x_k)}\, dx, \quad k = 1, 2, \ldots, n. \tag{5}$$

Доказательство. Необходимость. Пусть формула (1) точна для любого многочлена степени $m = 2n - 1$. Это значит,

что она точна и для многочлена $\omega(x)q(x)$, имеющего степень не выше $2n-1$, т. е.

$$\int_a^b \rho(x)\,\omega(x)\,q(x)\,dx = \sum_{k=1}^{n} c_k \omega(x_k)\,q(x_k) = 0.$$

Требование (5) выполняется в силу теоремы 1 из § 2 (если квадратурная формула (1) точна для любого многочлена степени $n-1$, то она является формулой интерполяционного типа).

Достаточность. Пусть $f(x)$ — любой многочлен степени $2n-1$. Согласно теореме о делении многочленов, его можно представить в виде

$$f(x) = \omega(x)\,q(x) + r(x),$$

где $q(x)$ и $r(x)$ — многочлены, имеющие степень не выше $n-1$. При этом

$$\int_a^b \rho(x) f(x)\,dx = \int_a^b \rho(x)\,\omega(x)\,q(x)\,dx + \int_a^b \rho(x)\,r(x)\,dx = \int_a^b \rho(x)\,r(x)\,dx.$$

Последнее равенство справедливо в силу предположения (4). Далее, поскольку $r(x)$ — многочлен степени не выше $n-1$ и формула (1) является формулой интерполяционного типа, она точна для $r(x)$, т. е.

$$\int_a^b \rho(x)\,r(x)\,dx = \sum_{k=1}^{n} c_k r(x_k) = \sum_{k=1}^{n} c_k\,(f(x_k) - \omega(x_k)\,q(x_k)) = \sum_{k=1}^{n} c_k f(x_k).$$

Таким образом,

$$\int_a^b \rho(x) f(x)\,dx = \sum_{k=1}^{n} c_k f(x_k),$$

т. е. формула (1) точна для любого многочлена степени $2n-1$. Теорема 1 доказана.

Отметим, что использование теоремы 1 существенно упрощает построение формул Гаусса.

Условие (4) эквивалентно требованиям

$$\int_a^b \rho(x)\,\omega(x)\,x^{\alpha}dx = 0, \qquad \alpha = 0, 1, \ldots, n-1, \tag{6}$$

которые представляют собой систему n уравнений относительно n неизвестных $x_1, x_2, \ldots, x_n$. Таким образом, для построения формул Гаусса достаточно найти узлы $x_1, x_2, \ldots, x_n$ из соотношений ортогональности (6) и затем вычислить коэффициенты c_k согласно (5).

Теорема 1 не гарантирует существования и единственности решения системы (6). Надо доказать еще существование и единственность многочлена $\omega(x)$ степени n, ортогонального всем многочленам степени меньшей n, а также убедиться в том, что все корни такого многочлена расположены на отрезке $[a, b]$.

3. Существование и единственность квадратурных формул наивысшей алгебраической степени точности. Представим искомый многочлен (3) в виде

$$\omega(x)=a_0+a_1x+\ldots+a_{n-1}x^{n-1}+x^n. \tag{7}$$

Тогда условия ортогональности (6) примут вид

$$\int_a^b \rho(x)\,(a_0+a_1x+\ldots+a_{n-1}x^{n-1}+x^n)\,x^\alpha dx=0, \tag{8}$$

$$\alpha=0,\ 1,\ \ldots,\ n-1.$$

Условия (8) представляют собой систему линейных алгебраических уравнений относительно коэффициентов $a_0,\ a_1,\ \ldots,\ a_{n-1}$. Покажем, что соответствующая однородная система уравнений

$$\int_a^b \rho(x)\,(a_0+a_1x+\ldots+a_{n-1}x^{n-1})\,x^\alpha dx=0, \tag{9}$$

$$\alpha=0,\ 1,\ \ldots,\ n-1,$$

имеет единственное решение $a_0=a_1=\ldots=a_{n-1}=0$. Для этого умножим уравнение (9) на a_α и просуммируем по всем α. Тогда получим

$$\sum_{\alpha=0}^{n-1} a_\alpha \int_a^b \left(\rho(x)\sum_{k=0}^{n-1} a_kx^k\right)x^\alpha dx=\int_a^b \rho(x)\left[\sum_{\alpha=0}^{n-1}\sum_{k=0}^{n-1} a_\alpha a_k x^k x^\alpha\right]dx=0,$$

т. е.

$$\int_a^b \rho(x)\left(\sum_{l=0}^{n-1} a_l x^l\right)^2 dx=0. \tag{10}$$

Если хотя бы один из коэффициентов a_l, $l=0,\ 1,\ \ldots,\ n-1$, отличен от нуля, то функция

$$\left(\sum_{l=0}^{n-1} a_l x^l\right)^2$$

может обратиться в нуль на $[a,\ b]$ лишь в конечном числе точек. Отсюда и из условия $\rho(x)>0$ следует, что равенство (10) возможно лишь в случае

$$a_0=a_1=\ldots=a_{n-1}=0.$$

Тем самым неоднородная система (8) имеет единственное решение. Следовательно, существует единственный многочлен $\omega(x)$ степени n со старшим коэффициентом 1, ортогональный с весом $\rho(x)>0$ любому многочлену степени $n-1$.

Теорема 2. *Если $\omega(x)$ — многочлен степени n, ортогональный на $[a,\ b]$ с весом $\rho(x)>0$ любому многочлену степени меньше n, то все его корни различны и расположены на $[a,\ b]$.*

Доказательство. Предположим, что многочлен $\omega(x)$ имеет $m\geqslant 0$ различных корней нечетной кратности на $[a,\ b]$. Очевидно,

что $m \leqslant n$. Теорема 2 будет доказана, если покажем, что $m=n$. Обозначая эти корни через $\xi_1, \xi_2, \ldots, \xi_m$, представим $\omega(x)$ в виде

$$\omega(x) = (x-\xi_1)^{\alpha_1}(x-\xi_2)^{\alpha_2}\ldots(x-\xi_m)^{\alpha_m} r(x),$$

где $\alpha_1, \alpha_2, \ldots, \alpha_m$ — нечетные числа и функция $r(x)$ не меняет знак на $[a, b]$. Вычислим интеграл

$$I = \int_a^b \rho(x)\,\omega(x)(x-\xi_1)\ldots(x-\xi_m)\,dx = \int_a^b \rho(x)(x-\xi_1)^{\alpha_1+1}\ldots(x-\xi_m)^{\alpha_m+1} r(x)\,dx. \quad (11)$$

Поскольку $\alpha_1+1, \ldots, \alpha_m+1$ четные числа и $r(x)$ знакопостоянна на $[a, b]$, интеграл (11) отличен от нуля. С другой стороны, если $m<n$, то

$$q(x) = (x-\xi_1)(x-\xi_2)\ldots(x-\xi_m)$$

— многочлен степени меньше n и по условию теоремы имеем $I=0$. Следовательно, $m=n$, что и доказывает теорему 2.

Из теорем 1 и 2 следует, что для любого n существует, притом единственная, квадратурная формула, точная для любого многочлена степени $2n-1$.

4. Свойства квадратурных формул Гаусса. Нетрудно показать, что $2n-1$ — наивысшая точность формулы Гаусса, т. е. что существует многочлен степени $2n$, для которого эта формула не является точной. Действительно, для многочлена (3) имеем

$$\int_a^b \rho(x)\,\omega^2(x)\,dx > 0,$$

но

$$\sum_{k=1}^n c_k \omega^2(x_k) = 0.$$

Докажем теперь, что при любом n коэффициенты c_k формул Гаусса положительны. Рассмотрим многочлены

$$\varphi_i(x) = \left(\frac{\omega(x)}{(x-x_i)\,\omega'(x_i)}\right)^2, \qquad i = 1, 2, \ldots, n,$$

имеющие степень $2n-2$ и обладающие свойством

$$\varphi_i(x_k) = \delta_{ik}.$$

Так как для этих многочленов формула Гаусса точна, справедливы равенства

$$\int_a^b \rho(x)\,\varphi_i(x)\,dx = \sum_{k=1}^n c_k \varphi_i(x_k) = c_i,$$

откуда и следует, что $c_i>0$, $i=1, 2, \ldots, n$.

В п. 4 § 2 отмечалось, что свойство положительности коэффициентов чрезвычайно важно для устойчивости вычислений и позволяет использовать формулы с большим числом узлов n. На практике применяются формулы Гаусса с числом узлов до 100.

Для погрешности формул Гаусса справедливо представление

$$\psi_n(f) = \frac{1}{(2n)!}\int_a^b \rho(x)\,\omega^2(x)\,f^{(2n)}(\xi)\,dx, \tag{12}$$

где $\xi \in (a, b)$.

Не приводя доказательства (см. [16, т. 1, с. 248]), отметим лишь, что оно основано на использовании интерполяционного многочлена Эрмита $H(x)$ с двукратными узлами

$$H(x_k) = f(x_k), \quad H'(x_k) = f'(x_k), \quad k = 1, 2, \ldots, n.$$

5. Частный случай формул Гаусса. *Формулами Эрмита* называются формулы Гаусса для вычисления интеграла

$$\int_{-1}^{1} \frac{f(x)\,dx}{\sqrt{1-x^2}}, \tag{13}$$

т. е. когда $a = -b = -1$, $\rho(x) = (1-x^2)^{-1/2}$.

Чтобы определить узлы соответствующей квадратурной формулы, надо, согласно теореме 1, найти многочлен (3), для которого

$$\int_{-1}^{1} \frac{\omega(x)\,q(x)}{\sqrt{1-x^2}}\,dx = 0 \tag{14}$$

для любого многочлена $q(x)$ степени меньше n. Можно показать (см. [2, с. 117]), что таковым является многочлен Чебышева

$$\omega(x) = T_n(x) = \frac{1}{2^{n-1}}\cos(n \arccos x). \tag{15}$$

Поэтому узлами квадратурной формулы Эрмита являются корни этого многочлена

$$x_k = \cos\frac{(2k-1)\pi}{2n}, \quad k = 1, 2, \ldots, n. \tag{16}$$

Соответствующие коэффициенты вычисляются по формулам (5)

$$c_k = \int_{-1}^{1} \frac{T_n(x)\,dx}{\sqrt{1-x^2}\,T_n'(x_k)\,(x-x_k)} \tag{17}$$

и оказываются равными

$$c_k = \pi/n, \quad k = 1, 2, \ldots, n.$$

Таким образом, формулы Эрмита имеют вид

$$\int_{-1}^{1} \frac{f(x)\,dx}{\sqrt{1-x^2}} \approx \frac{\pi}{n}\sum_{k=1}^{n} f(x_k), \tag{18}$$

где x_k — корни многочлена Чебышева, определенные согласно (16).

§ 4. Численное дифференцирование

1. Некорректность операции численного дифференцирования. Задача численного дифференцирования состоит в приближенном вычислении производных функции $u(x)$ по заданным в конечном числе точек значениям этой функции. Простейшие примеры формул численного дифференцирования рассматривались в п. 1 § 4 ч. I. Напомним эти примеры. Пусть на $[a, b]$ введена сетка

$$\omega_h = \{x_i = a + ih,\ i = 0, 1, \ldots, N,\ hN = b - a\}$$

и определены значения $u_i = u(x_i)$ функции $u(x)$ в точках сетки. В качестве приближенного значения $u'(x_i)$ можно взять, например, любое из следующих разностных отношений:

$$u_{\overline{x},i} = \frac{u_i - u_{i-1}}{h}, \quad u_{x,i} = \frac{u_{i+1} - u_i}{h}, \quad u_{\overset{\circ}{x},i} = \frac{u_{i+1} - u_{i-1}}{2h}.$$

Возникающая в результате такой замены погрешность характеризуется разложениями

$$u_{\overline{x},i} = u'(x_i) - \frac{h}{2} u''(\xi_i^{(1)}), \tag{1}$$

$$u_{x,i} = u'(x_i) + \frac{h}{2} u''(\xi_i^{(2)}), \tag{2}$$

$$u_{\overset{\circ}{x},i} = u'(x_i) + \frac{h^2}{6} u'''(\xi_i^{(3)}), \tag{3}$$

где $\xi_i^{(j)}$, $j = 1, 2, 3$,— точки из интервала (x_{i-1}, x_{i+1}).

Вторую производную в точке x_i можно заменить отношением

$$u_{\overline{x}x,i} = \frac{1}{h}(u_{x,i} - u_{\overline{x},i}) = \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2},$$

при этом

$$u_{\overline{x}x,i} = u''(x_i) + \frac{h^2}{12} u^{\mathrm{IV}}(x_i) + O(h^4). \tag{4}$$

Четвертая производная $u^{\mathrm{IV}}(x_i)$ с точностью до величины $O(h^2)$ аппроксимируется разностным отношением

$$u_{\overline{x}x\overline{x}x,i} = \frac{1}{h^2}(u_{\overline{x}x,i+1} - 2u_{\overline{x}x,i} + u_{\overline{x}x,i-1}) =$$

$$= \frac{1}{h^4}(u_{i+2} - 4u_{i+1} + 6u_i - 4u_{i-1} + u_{i-2}).$$

Как правило, значения функции $u(x)$ в точках сетки ω_h вычисляются не точно, а с каким-то приближением. Например, элементарные трансцендентные функции вычисляются с помощью рядов, причем ряды заменяются конечными суммами. Другим источником погрешностей являются погрешности округления. Оказывается, что погрешность, возникающая при вычислении разностных отношений, намного превосходит погрешность в задании значений функции $u(x)$ и даже может неограниченно возрастать при стремлении шага

сетки h к нулю. Поэтому операцию вычисления разностных отношений называют некорректной. Поясним причину некорректности на примере вычисления разностного отношения $u_{\bar{x},i} = (u_i - u_{i-1})/h$.

Разностное отношение $u_{\bar{x},i}$ хорошо приближает $u'(x_i)$ только в том случае, когда шаг h достаточно мал. Требование малости величины h, находящейся в знаменателе разностного отношения, как раз и является причиной некорректности операции численного дифференцирования. Действительно, пусть вместо точного значения u_i, u_{i-1} вычислены приближенные значения $\tilde{u}_i = u_i + \delta_i$, $\tilde{u}_{i-1} = u_{i-1} + \delta_{i-1}$. Тогда вместо $u_{\bar{x},i}$ будет вычислена величина $u_{\bar{x},i} + (\delta_i - \delta_{i-1})/h$. Следовательно, погрешность в вычислении первой разностной производной окажется равной $\delta_{\bar{x},i} = (\delta_i - \delta_{i-1})/h$.

В дальнейшем погрешности такого рода будем называть погрешностями округления (хотя их реальная природа может быть иной).

Пусть известна граница δ погрешностей δ_i, δ_{i-1}, т. е. $|\delta_i| \leqslant \delta$, $|\delta_{i-1}| \leqslant \delta$. Тогда

$$|\delta_{\bar{x},i}| \leqslant 2\delta/h, \tag{5}$$

причем эта оценка достигается при $\delta_i = -\delta_{i-1} = \delta$. Из оценки (5) видно, что вследствие малости h погрешность, возникающая при вычислении первой разностной производной, значительно превосходит погрешность вычисления самой функции $u(x)$. Если δ не зависит от h, то погрешность $\delta_{\bar{x},i}$ неограниченно возрастает при $h \to 0$.

Сказанное не означает, что нельзя пользоваться формулами численного дифференцирования. Чтобы не происходило существенного понижения точности, надо следить за тем, чтобы погрешность округления имела тот же порядок, что и погрешность аппроксимации. Например, из (1) следует, что погрешность аппроксимации при замене $u'(x)$ отношением $u_{\bar{x},i}$ не превосходит величины $0{,}5hM_2$, где $M_2 = \max\limits_{x \in [a,b]} |u''(x)|$. Естественно потребовать, чтобы и погрешность округления $\delta_{\bar{x},i}$ была бы сравнима с погрешностью аппроксимации, например

$$2\delta/h \leqslant M_2 h/2, \tag{6}$$

где M_2 не зависит от h. Это означает, что погрешность δ при вычислении значений функции $u(x_i)$ должна быть величиной $O(h^2)$. С другой стороны, неравенство (6) показывает, что если величина δ задана и мы не можем ее менять, то вычисления надо проводить не с произвольно малым шагом h, а с шагом, удовлетворяющим условию $h \geqslant h_0$, где $h_0 = 2\sqrt{\delta/M_2}$.

При вычислении производных более высокого порядка, когда в знаменатель разностного отношения входит h^k, $k > 1$, влияние неточности в задании $u(x_i)$ сказывается еще сильнее. Например, при вычислении разностного отношения $u_{\bar{x}x\bar{x}x,i}$ погрешность округления является величиной $O(\delta h^{-4})$, где δ — граница погрешности округления функции $u(x)$. В этом случае для того чтобы погрешность округления $\delta_{\bar{x}x\bar{x}x,i}$ была сравнима с погрешностью аппроксимации,

надо потребовать, чтобы $h \geqslant h_0$, где $h_0 = O(\delta^{1/2})$, либо проводить вычисление $u(x_i)$ с погрешностью $\delta = O(h^6)$. Например, если $\delta \approx 10^{-12}$, то шаг h надо брать примерно равным 0,01. При этом погрешность аппроксимации и погрешность округления будут примерно равными 10^{-4}.

Вычисление производной $u'(x)$ по заданной функции $u(x)$ также является некорректной операцией в том смысле, что для ограниченной функции $u(x)$ производная $u'(x)$ может быть сколь угодно большой. Например, для $u(x) = \sin \omega x$ имеем $\max\limits_{x \in [a,b]} |u(x)| \leqslant 1$ и $\max\limits_{x \in [a,b]} |u'(x)| = |\omega| \to \infty$ при $\omega \to \infty$.

Строгие определения корректности математической задачи и способы решения некорректных задач изложены в книге [38].

2. **Применение интерполирования.** Многие формулы численного дифференцирования можно получить как следствие интерполяционных формул. Для этого достаточно заменить функцию $u(x)$ ее интерполяционным многочленом $L_n(x)$ и вычислить производные многочлена $L_n(x)$, используя его явное представление. В отличие от п. 1 рассмотрим неравномерную сетку

$$\omega_h = \{a = x_0 < x_1 < x_2 < \ldots < x_N = b\}$$

и обозначим через $h_i = x_i - x_{i-1}$, $i = 1, 2, \ldots, N$, шаги этой сетки. В качестве примера получим формулы численного дифференцирования, основанные на использовании многочлена Лагранжа $L_{2,i}(x)$, построенного для функции $u(x)$ по трем точкам x_{i-1}, x_i, x_{i+1}. Многочлен $L_{2,i}(x)$ имеет вид

$$L_{2,i}(x) = \frac{(x - x_i)(x - x_{i+1})}{h_i (h_i + h_{i+1})} u_{i-1} - \frac{(x - x_{i-1})(x - x_{i+1})}{h_i h_{i+1}} u_i + \frac{(x - x_{i-1})(x - x_i)}{h_{i+1}(h_i + h_{i+1})} u_{i+1}. \tag{7}$$

Отсюда получим

$$L'_{2,i}(x) = \frac{(2x - x_i - x_{i+1})}{h_i (h_i + h_{i+1})} u_{i-1} - \frac{(2x - x_{i-1} - x_{i+1})}{h_i h_{i+1}} u_i + \frac{(2x - x_{i-1} - x_i)}{h_{i+1}(h_i + h_{i+1})} u_{i+1}.$$

Это выражение можно принять за приближенное значение $u'(x)$ в любой точке $x \in [x_{i-1}, x_{i+1}]$. Его удобнее записать в виде

$$L'_{2,i}(x) = \frac{1}{\hbar_i} \left[(x - x_{i-1/2}) \frac{u_{i+1} - u_i}{h_{i+1}} + (x_{i+1/2} - x) \frac{u_i - u_{i-1}}{h_i} \right], \tag{8}$$

где $\hbar_i = 0{,}5(h_i + h_{i+1})$, $x_{i-1/2} = x_i - 0{,}5h_i$. В частности, при $x = x_i$ получим

$$L'_{2,i}(x_i) = \frac{1}{2} \left(\frac{h_i}{\hbar_i} \frac{u_{i+1} - u_i}{h_{i+1}} + \frac{h_{i+1}}{\hbar_i} \frac{u_i - u_{i-1}}{h_i} \right), \tag{9}$$

и если сетка равномерна, $h_{i+1} = h_i = h$, то приходим к центральной разностной производной, $L'_{2,i}(x_i) = u_{\overset{\circ}{x},i}$.

При использовании интерполяционного многочлена первой степени точно таким же образом можно получить односторонние разностные производные $u_{\overline{x},i}$ и $u_{x,i}$.

Далее, вычисляя вторую производную многочлена $L_{2,i}(x)$, получим приближенное выражение для $u''(x)$ при $x \in [x_{i-1}, x_{i+1}]$:

$$u''(x) \approx L''_{2,i}(x) = \frac{1}{h_i}\left(\frac{u_{i+1}-u_i}{h_{i+1}} - \frac{u_i - u_{i-1}}{h_i}\right). \qquad (10)$$

На равномерной сетке это выражение совпадает со второй разностной производной $u_{\overline{x}x,i}$. Ясно, что для приближенного вычисления дальнейших производных уже недостаточно многочлена $L_{2,i}(x)$, надо привлекать многочлены более высокого порядка и тем самым увеличивать число узлов, участвующих в аппроксимации.

Порядок погрешности аппроксимации зависит как от порядка интерполяционного многочлена, так и от расположения узлов интерполирования. Получим выражение для погрешности аппроксимации, возникающей при замене $u'(x)$ выражением $L'_{2,i}(x)$. Будем считать, что $x \in [x_{i-1}, x_{i+1}]$ и что величины h_i, h_{i+1} имеют один и тот же порядок малости при измельчении сетки. По формуле Тейлора в предположении ограниченности $u^{IV}(x)$ получим

$$u_{i+k} = u(x) + (x_{i+k} - x)\,u'(x) + \frac{(x_{i+k}-x)^2}{2}u''(x) + \frac{(x_{i+k}-x)^3}{6}u'''(x) + O(h^4),$$

где $k = 0, \pm 1$, $h = \max\{h_i, h_{i+1}\}$. Отсюда приходим к следующим разложениям разностных отношений:

$$\frac{u_i - u_{i-1}}{h_i} = u'(x) - (x - x_{i-1/2})\,u''(x) + \left(\frac{(x-x_{i-1/2})^2}{2} + \frac{h_i^2}{24}\right)u'''(x) + O(h^3), \qquad (11)$$

$$\frac{u_{i+1} - u_i}{h_{i+1}} = u'(x) + (x_{i+1/2} - x)\,u''(x) + \left(\frac{(x_{i+1/2}-x)^2}{2} + \frac{h_{i+1}^2}{24}\right)u'''(x) + O(h^3). \qquad (12)$$

Подставляя (11) и (12) в выражение для разностной производной (8) и приводя подобные члены, получим

$$L'_{2,i}(x) = u'(x) - \left[\frac{(x-x_i)^2}{2} - \frac{(h_{i+1}-h_i)(x-x_i)}{3} - \frac{h_i h_{i+1}}{6}\right]u'''(x) + O(h^3), \quad x \in (x_{i-1}, x_{i+1}).$$

Отсюда видно, что разностное выражение (8) аппроксимирует $u'(x)$ со вторым порядком. Несколько хуже обстоит дело с выражением (10), аппроксимирующим вторую производную. Из (4) видно,

что на равномерной сетке в точке $x=x_i$ имеет место аппроксимация $O(h^2)$. Покажем, что на неравномерной сетке ($h_i \neq h_{i+1}$) погрешность аппроксимации будет иметь только первый порядок. Подставляя разложения (11), (12) в выражение (10) для $L''_{2,i}(x)$, получим

$$L''_{2,i}(x) = u''(x) + \left(x_i - x + \frac{h_{i+1} - h_i}{3}\right) u'''(x) + O(h^2).$$

Здесь даже на равномерной сетке второй порядок аппроксимации имеет место лишь в точке $x=x_i$, а относительно других точек (например, точек $x=x_{i-1}$ и $x=x_{i+1}$) выполняется аппроксимация только первого порядка.

ГЛАВА 5

РЕШЕНИЕ НЕЛИНЕЙНЫХ УРАВНЕНИЙ И СИСТЕМ УРАВНЕНИЙ

§ 1. Примеры итерационных методов решения нелинейных уравнений

1. Введение. Пусть задана функция $f(x)$ действительного переменного. Требуется найти корни уравнения

$$f(x) = 0 \tag{1}$$

или, что то же самое, нули функции $f(x)$.

Уже на примере алгебраического многочлена известно, что нули $f(x)$ могут быть как действительными, так и комплексными. Поэтому более точная постановка задачи состоит в нахождении корней уравнения (1), расположенных в заданной области комплексной плоскости. Можно рассматривать также задачу нахождения действительных корней, расположенных на заданном отрезке. Иногда, пренебрегая точностью формулировок, будем говорить, что требуется решить уравнение (1).

Задача нахождения корней уравнения (1) обычно решается в два этапа. На первом этапе изучается расположение корней (в общем случае на комплексной плоскости) и проводится их разделение, т. е. выделяются области в комплексной плоскости, содержащие только один корень. Кроме того, изучается вопрос о кратности корней. Тем самым находятся некоторые начальные приближения для корней уравнения (1). На втором этапе, используя заданное начальное приближение, строится итерационный процесс, позволяющий уточнить значение отыскиваемого корня.

Не существует каких-то общих регулярных приемов решения задачи о расположении корней произвольной функции $f(x)$. Наиболее полно изучен вопрос о расположении корней алгебраических многочленов

$$f(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_m x^m. \tag{2}$$

Например известно, что если для многочлена (2) с действительными коэффицентами выполнены неравенства

$$f(c) > 0,\ f'(c) > 0,\ \ldots,\ f^{(m)}(c) > 0,$$

то положительные корни $f(x)$ не превосходят числа c. Действительно, из формулы Тейлора

$$f(x)=f(c)+(x-c)f'(c)+\frac{(x-c)^2}{2!}f''(c)+\ldots+\frac{(x-c)^m}{m!}f^{(m)}(c)$$

получаем, что $f(x)>0$ при $x\geqslant c$.

Численные методы решения нелинейных уравнений являются, как правило, итерационными методами, которые предполагают задание достаточно близких к искомому решению начальных данных.

Прежде чем переходить к изложению конкретных итерационных методов, отметим два простых приема отделения действительных корней уравнения (1). Предположим, что $f(x)$ определена и непрерывна на $[a, b]$.

Первый прием состоит в том, что вычисляется таблица значений функции $f(x)$ в заданных точках $x_k\in[a, b]$, $k=0, 1, \ldots, n$. Если обнаружится, что при некотором k числа $f(x_k)$, $f(x_{k+1})$ имеют разные знаки, то это будет означать, что на интервале (x_k, x_{k+1}) уравнение (1) имеет по крайней мере один действительный корень (точнее, имеет нечетное число корней на (x_k, x_{k+1})). Затем можно разбить интервал (x_k, x_{k+1}) на более мелкие интервалы и с помощью аналогичной процедуры уточнить расположение корня.

Более регулярным способом отделения действительных корней является *метод бисекции* (*деления пополам*). Предположим, что на (a, b) расположен лишь один корень x_* уравнения (1). Тогда $f(a)$ и $f(b)$ имеют различные знаки. Пусть для определенности $f(a)>0$, $f(b)<0$. Положим $x_0=0{,}5(a+b)$ и вычислим $f(x_0)$. Если $f(x_0)<0$, то искомый корень находится на интервале (a, x_0), если же $f(x_0)>0$, то $x_*\in(x_0, b)$. Далее, из двух интервалов (a, x_0) и (x_0, b) выбираем тот, на границах которого функция $f(x)$ имеет различные знаки, находим точку x_1 — середину выбранного интервала, вычисляем $f(x_1)$ и повторяем указанный процесс. В результате получаем последовательность интервалов, содержащих искомый корень x_*, причем длина каждого последующего интервала вдвое меньше, чем предыдущего. Процесс заканчивается, когда длина вновь полученного интервала станет меньше заданного числа $\varepsilon>0$, и в качестве корня x_* приближенно принимается середина этого интервала.

Заметим, что если на (a, b) имеется несколько корней, то указанный процесс сойдется к одному из корней, но заранее неизвестно, к какому именно. Можно использовать прием выделения корней: если корень $x=x_*$ кратности m найден, то рассматривается функция

$$g(x)=f(x)/(x-x_*)^m$$

и для нее повторяется процесс нахождения корня.

2. Метод простой итерации. Он состоит в том, что уравнение (1) заменяется эквивалентным уравнением

$$x=s(x) \tag{3}$$

и итерации образуются по правилу

$$x_{n+1}=s(x_n), \quad n=0, 1, \ldots, \tag{4}$$

причем задается начальное приближение x_0. Для сходимости большое значение имеет выбор функции $s(x)$. Эту функцию можно задавать различными способами, однако обычно она берется в виде

$$s(x)=x+\tau(x)f(x), \tag{5}$$

причем функция $\tau(x)$ не меняет знака на том отрезке, где отыскивается корень. В § 2 будет показано, что метод простой итерации сходится при надлежащем выборе начального приближения x_0, если $|s'(x_*)|<1$, где x_* — корень уравнения (1).

Отметим, что в форме метода простой итерации (4) можно записать, по существу, любой одношаговый итерационный метод.

В частности, если $\tau(x)=\tau=\text{const}$, то получим *метод релаксации*

$$\frac{x_{n+1}-x_n}{\tau}=f(x_n), \quad n=0, 1, \ldots, \tag{6}$$

для которого $s'(x)=1+\tau f'(x)$, и метод сходится при условии

$$-2<\tau f'(x_*)<0. \tag{7}$$

Если в некоторой окрестности корня выполняются условия

$$f'(x)<0, \quad 0<m_1<|f'(x)|<M_1, \tag{8}$$

то метод релаксации сходится при $\tau\in(0, 2/M_1)$.

Чтобы выбрать оптимальный параметр τ в методе релаксации, рассмотрим уравнение для погрешности $z_n=x_n-x_*$. Подставляя $x_n=x_*+z_n$ в (6), получим уравнение

$$\frac{z_{n+1}-z_n}{\tau}=f(x_*+z_n).$$

По теореме о среднем имеем

$$f(x_*+z_n)=f(x_*)+z_nf'(x_*+\theta z_n)=z_nf'(x_*+\theta z_n),$$

где $\theta\in(0, 1)$. Таким образом, для погрешности метода релаксации выполняется уравнение

$$\frac{z_{n+1}-z_n}{\tau}=f'(x_*+\theta z_n)\,z_n.$$

Отсюда приходим к оценке

$$|z_{n+1}|\leqslant|1+\tau f'(x_*+\theta z_n)|\cdot|z_n|\leqslant\max_x|1+\tau f'(x_*+\theta z_n)|\cdot|z_n|,$$

и если выполнены условия (8), то

$$|z_{n+1}|\leqslant\max\{|1-\tau M_1|, |1-\tau m_1|\}|z_n|.$$

Таким образом, задача выбора оптимального параметра сводится к нахождению τ, для которого функция

$$q(\tau)=\max\{|1-\tau M_1|, |1-\tau m_1|\}$$

принимает минимальное значение.

Из рассмотрения графика функции $q(\tau)$ видно, что точка минимума определяется условием

$$|1-\tau M_1| = |1-\tau m_1|$$

и равна

$$\tau = \tau_0 = 2/(M_1+m_1).$$

При этом значении τ имеем

$$q(\tau_0) = \rho_0 = \frac{1-\xi}{1+\xi}, \quad \xi = \frac{m_1}{M_1},$$

так что для погрешности справедлива оценка

$$|z_n| \leqslant \rho_0^n |z_0|, \qquad n = 0, 1, \ldots$$

3. Метод Ньютона. Пусть начальное приближение x_0 известно. Заменим $f(x)$ отрезком ряда Тейлора

$$f(x) \approx H_1(x) = f(x_0) + (x-x_0) f'(x_0)$$

и за следующее приближение x_1 возьмем корень уравнения $H_1(x) = 0$, т. е.

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Вообще, если итерация x_k известна, то следующее приближение x_{k+1} в *методе Ньютона* определяется по правилу

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \qquad k = 0, 1, \ldots \tag{9}$$

Метод Ньютона называют также *методом касательных*, так как новое приближение x_{k+1} является абсциссой точки пересечения касательной, проведенной в точке $(x_k, f(x_k))$ к графику функции $f(x)$, с осью Ox.

Исследование сходимости метода Ньютона будет проведено в § 3. Здесь отметим без доказательства лишь две особенности этого метода. Во-первых, метод имеет *квадратичную сходимость*, т. е. в отличие от линейных задач погрешность на следующей итерации пропорциональна квадрату погрешности на предыдущей итерации: $x_{k+1}-x_* = O((x_k-x_*)^2)$.

И, во-вторых, такая быстрая сходимость метода Ньютона гарантируется лишь при очень хороших, т. е. близких к точному решению, начальных приближениях. Если начальное приближение выбрано неудачно, то метод может сходиться медленно, либо не сойдется вообще.

Модифицированный метод Ньютона

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_0)}, \qquad k = 0, 1, \ldots \tag{10}$$

применяют в том случае, когда хотят избежать многократного вычисления производной $f'(x)$. Метод (10) предъявляет меньше тре-

бований к выбору начального приближения x_0, однако обладает *лишь линейной сходимостью*, т. е. $x_{k+1}-x_* = O(x_k-x_*)$.

Метод (10) гарантирует отсутствие деления на нуль, если $f'(x_0)\neq 0$.

4. Метод секущих. Этот метод получается из метода Ньютона (9) заменой $f'(x_k)$ разделенной разностью $\dfrac{f(x_k)-f(x_{k-1})}{x_k-x_{k-1}}$, вычисленной по известным значениям x_k и x_{k-1}. В результате получаем итерационный метод

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k)-f(x_{k-1})} f(x_k), \qquad k = 1, 2, \ldots, \tag{11}$$

который в отличие от ранее рассмотренных методов является *двухшаговым*, т. е. новое приближение x_{k+1} определяется двумя предыдущими итерациями x_k и x_{k-1}. В методе (11) необходимо задавать два начальных приближения x_0 и x_1.

Геометрическая интерпретация метода секущих состоит в следующем. Через точки $(x_{k-1}, f(x_{k-1}))$, $(x_k, f(x_k))$ проводится прямая, абсцисса точки пересечения этой прямой с осью Ox и является новым приближением x_{k+1}. Иначе говоря, на отрезке $[x_{k-1}, x_k]$ функция $f(x)$ интерполируется многочленом первой степени и за очередное приближение x_{k+1} принимается корень этого многочлена.

5. Интерполяционные методы. Идея *интерполяционных методов* состоит в том, что нахождение корней уравнения (1) заменяется нахождением корней интерполяционного многочлена, построенного для $f(x)$. Интерполяционный метод первого порядка приводит к методу секущих. Интерполяционный метод второго порядка называется *методом парабол*. Метод Ньютона (9) можно получить, заменяя $f(x)$ интерполяционным многочленом Эрмита первой степени.

Получим формулы метода парабол. Пусть приближения x_{k-2}, x_{k-1}, x_k известны. Построим интерполяционный многочлен Ньютона (см. (11) из § 1 гл. 3)

$$P_2(x) = f(x_k) + (x-x_k) f(x_k, x_{k-1}) + (x-x_k)(x-x_{k-1}) f(x_k, x_{k-1}, x_{k-2})$$

и обозначим $z = x - x_k$. Тогда уравнение $P_2(x) = 0$ примет вид

$$az^2 + bz + c = 0, \tag{12}$$

где $a = f(x_k, x_{k-1}, x_{k-2})$, $b = f(x_k, x_{k-1}) + (x_k - x_{k-1}) f(x_k, x_{k-1}, x_{k-2})$, $c = f(x_k)$.

Решая уравнение (12), получим два, может быть комплексных, корня, $z^{(1)}$ и $z^{(2)}$, по которым вычислим $x^{(1)} = x_k + z^{(1)}$, $x^{(2)} = x_k + z^{(2)}$. В качестве следующего приближения в методе парабол выбирается то из значений $x^{(1)}$, $x^{(2)}$, которое ближе к x_k, т. е. отвечающее минимальному по модулю корню уравнения (12). Метод парабол удобен тем, что позволяет получить комплексные корни уравнения (7), пользуясь вещественными начальными приближениями x_0, x_1, x_2.

6. Использование обратной интерполяции. Ряд итерационных методов можно получить с помощью интерполирования функции $x=\varphi(y)$, обратной $f(x)$. Заметим, что если x_* — корень уравнения $f(x)=0$, то $\varphi(0)=x_*$. Таким образом, задача нахождения корня x_* сводится к вычислению значения $\varphi(0)$.

Предположим, что известны приближения $x_0, x_1, \ldots, x_n$ к корню x_*. Тогда можно вычислить $y_i=f(x_i)$, $i=0, 1, \ldots, n$, и считать, что по переменной y заданы узлы $y_0, y_1, \ldots, y_n$ и в них известны значения $x_0=\varphi(y_0), \ldots, x_n=\varphi(y_n)$. По данным $(y_i, \varphi(y_i))$, $i=0, 1, \ldots, n$, строится интерполяционный многочлен $L_n(y)$ для функции $\varphi(y)$ и в качестве следующего приближения x_{n+1} берется $L_n(0)$.

Линейная обратная интерполяция ($n=1$) приводит к методу секущих. Квадратичная обратная интерполяция ($n=2$) приводит к методу

$$x_{k+1}=x_k-x_k\varphi(y_k, y_{k-1})+x_{k-1}x_k\varphi(y_k, y_{k-1}, y_{k-2}),$$

отличному от метода парабол. Здесь $\varphi(y_k, y_{k-1})$ и $\varphi(y_k, y_{k-1}, y_{k-2})$ — разделенные разности первого и второго порядков соответственно.

Сделаем следующее замечание. Перечисленные выше итерационные методы в случае сходимости позволяют при заданных начальных приближениях найти лишь один из корней уравнения (1). Чтобы отыскать другие корни, надо менять начальные приближения. Может оказаться, что и при других начальных данных метод сходится к тому же корню $x=x_*$. Тогда целесообразно отделить этот корень, т. е. применить итерационный метод к $g(x)=f(x)/(x-x_*)$.

§ 2. Сходимость метода простой итерации

1. Теорема о сходимости. Перепишем уравнение

$$f(x)=0 \tag{1}$$

в эквивалентном виде

$$x=s(x) \tag{2}$$

и рассмотрим метод простой итерации

$$x_{k+1}=s(x_k), \quad k=0, 1, \ldots, x_0 \text{ задан.} \tag{3}$$

Говорят, что итерационный метод *сходится*, если последовательность $\{x_k\}$ имеет предел при $k\to\infty$.

В следующей теореме формулируются условия на функцию $s(x)$, гарантирующие существование и единственность решения уравнения (2) и сходимость метода простой итерации к этому решению. Напомним, что функция $s(x)$ называется липшиц-непрерывной с постоянной q на множестве X, если для всех x', $x''\in X$ выполняется неравенство

$$|s(x')-s(x'')|\leqslant q|x'-x''|. \tag{4}$$

В дальнейшем в качестве X будем брать отрезок

$$U_r(a)=\{x:\ |x-a|\leqslant r\} \tag{5}$$

длины $2r$ с серединой в точке a.

Теорема 1. *Если $s(x)$ липшиц-непрерывна с постоянной $q\in(0, 1)$ на отрезке $U_r(a)$, причем*

$$|s(a)-a|\leqslant(1-q)r, \tag{6}$$

то уравнение (2) имеет на отрезке $U_r(a)$ единственное решение x_* и метод простой итерации (3) сходится к x_* при любом начальном приближении $x_0\in U_r(a)$. Для погрешности справедлива оценка

$$|x_k-x_*|\leqslant q^k|x_0-x_*|, \quad k=0, 1, 2, \ldots \tag{7}$$

Доказательство. Сначала докажем по индукции, что $x_k\in U_r(a)$, $k=1, 2, \ldots$, т. е. что метод простой итерации не выводит за пределы того множества, на котором $s(x)$ липшиц-непрерывна с постоянной $q\in(0, 1)$. Предположим, что $x_j\in U_r(a)$ при некотором $j\geqslant 0$, и докажем, что тогда $x_{j+1}\in U_r(a)$. Из равенства

$$x_{j+1}-a=s(x_j)-a=(s(x_j)-s(a))+(s(a)-a)$$

получим

$$|x_{j+1}-a|\leqslant|s(x_j)-s(a)|+|s(a)-a|.$$

Учитывая условие липшиц-непрерывности, предположение индукции и условие (6), имеем

$$|s(x_j)-s(a)|\leqslant q|x_j-a|\leqslant qr,$$
$$|x_{j+1}-a|\leqslant qr+(1-q)r\leqslant r,$$

т. е. $x_{j+1}\in U_r(a)$.

Оценим теперь разность двух соседних итераций $x_{j+1}-x_j$. Имеем

$$x_{j+1}-x_j=s(x_j)-s(x_{j-1}),$$

и поскольку все точки x_j, $j=1, 2, \ldots$, находятся на отрезке $U_r(a)$, получаем оценку

$$|x_{j+1}-x_j|\leqslant q|x_j-x_{j-1}|$$

и, следовательно,

$$|x_{j+1}-x_j|\leqslant q^j|x_1-x_0|, \quad j=1, 2, \ldots \tag{8}$$

Оценка (8) позволяет доказать фундаментальность последовательности $\{x_k\}$. Действительно, пусть p — любое натуральное число. Тогда

$$x_{k+p}-x_k=\sum_{j=1}^{p}(x_{k+j}-x_{k+j-1}),$$

и согласно (8) имеем

$$|x_{k+p}-x_k|\leqslant|x_1-x_0|\sum_{j=1}^{p}q^{k+j-1}=q^k\frac{1-q^p}{1-q}|x_1-x_0|\leqslant\frac{q^k}{1-q}|x_1-x_0|,$$

т. е.

$$|x_{k+p}-x_k|\leqslant\frac{q^k}{1-q}|x_1-x_0|, \quad k, p=1, 2, \ldots \tag{9}$$

Поскольку правая часть неравенства (9) стремится к нулю при $k\to\infty$ и не зависит от p, последовательность $\{x_k\}$ является фундаментальной. Следовательно, существует

$$\lim_{k\to\infty} x_k = x_* \in U_r(a).$$

Переходя в (3) к пределу при $k\to\infty$ и учитывая непрерывность функции $s(x)$, получим $x_* = s(x_*)$, т. е. x_* — решение уравнения (2).

Предположим, что x_*' — какое-то решение уравнения (2), принадлежащее отрезку $U_r(a)$. Тогда

$$x_* - x_*' = s(x_*) - s(x_*')$$

и по условию теоремы

$$|x_* - x_*'| \leqslant q\,|x_* - x_*'|.$$

Так как $q<1$, последнее неравенство может выполняться лишь при $x_*' = x_*$, т. е. решение единственно.

Докажем оценку погрешности (7). Из уравнения (3) получим

$$x_{k+1} - x_* = s(x_k) - x_* = s(x_k) - s(x_*),$$

и так как $x_k,\ x_* \in U_r(a)$, приходим к неравенству

$$|x_{k+1} - x_*| \leqslant q\,|x_k - x_*|, \tag{10}$$

справедливому для всех $k=0, 1, \ldots$, из которого и следует оценка (7). Теорема 1 доказана.

Замечание 1. Если для погрешности какого-либо итерационного метода выполняется неравенство

$$|x_k - x_*| \leqslant M_1 q^k |x_0 - x_*|,$$

где $q\in(0, 1)$ и M_1 не зависит от k, то говорят, что метод сходится линейно со скоростью геометрической прогрессии со знаменателем q. Такая терминология объясняется тем, что при $k\to\infty$ погрешность убывает как q^k.

Замечание 2. Зафиксируем в неравенстве (9) индекс k и устремим p к бесконечности. Тогда получим оценку погрешности

$$|x_k - x_*| \leqslant \frac{q^k}{1-q}\,|s(x_0) - x_0|, \qquad k = 1, 2, \ldots \tag{11}$$

В правую часть оценки (11) входят только известные величины, в то время как оценка (7) содержит заранее неизвестное значение x_*.

Приведем следствия из теоремы 1, содержащие более удобные для проверки условия сходимости.

Будем предполагать, что $s(x)$ непрерывно дифференцируема на отрезке $U_r(a)$.

Следствие 1. *Если*

$$|s'(x)| \leqslant q < 1 \tag{12}$$

для $x\in U_r(a)$, *выполнено условие* (6) *и* $x_0\in U_r(a)$, *то уравнение* (2) *имеет единственное решение* $x_*\in U_r(a)$, *метод* (3) *сходится и справедлива оценка* (7).

Действительно, из (12) следует (4) с $q\in(0, 1)$.

Следствие 2. *Пусть уравнение* (2) *имеет решение* x_*, *функция* $s(x)$ *непрерывно дифференцируема на отрезке*

$$U_r(x_*) = \{x: \ |x - x_*| \leqslant r\} \tag{13}$$

и $|s'(x_*)| < 1$. *Тогда существует* $\varepsilon > 0$ *такое, что на отрезке* $U_\varepsilon(x_*)$ *уравнение* (2) *не имеет других решений и метод* (3) *сходится, если только* $x_0 \in U_\varepsilon(x_*)$.

Доказательство. Поскольку $s(x)$ непрерывно дифференцируема на отрезке $U_r(x_*)$ и $|s'(x_*)| < 1$, найдутся числа $q \in (0, 1)$ и $\varepsilon \in (0, r]$ такие, что

$$|s'(x)| \leqslant q < 1$$

для всех $x \in U_\varepsilon(x_*)$.

2. Метод Эйткена ускорения сходимости. Предположим, что какой-либо итерационный метод имеет линейную сходимость, т. е.

$$x_k - x_* \approx aq^k, \quad q \in (0, 1), \quad k = 1, 2, \ldots$$

Числа a, q, x_* заранее неизвестны, но их можно найти, используя три последовательных итерации x_k, x_{k+1}, x_{k+2}. Составим уравнения

$$x_k - x_* = aq^k, \quad x_{k+1} - x_* = aq^{k+1}, \quad x_{k+2} - x_* = aq^{k+2}$$

(здесь равенства надо понимать как приближенные), из которых найдем

$$\begin{aligned} \Delta x_k &= x_{k+1} - x_k = aq^k (q - 1), \\ \Delta^2 x_k &= \Delta x_{k+1} - \Delta x_k = aq^k (q - 1)^2, \\ x_* &= x_{k+2} - \frac{(\Delta x_{k+1})^2}{\Delta^2 x_k}. \end{aligned} \tag{14}$$

Метод Эйткена ускорения сходимости состоит в том, что после вычисления x_k, x_{k+1}, x_{k+2} производится пересчет по формуле

$$y_{k+1} = x_{k+2} - \frac{(\Delta x_{k+1})^2}{\Delta^2 x_k} \tag{15}$$

и значение y_{k+1} принимается за новое приближение. Если бы равенство (14) выполнялось точно, то y_{k+1} совпало бы с точным решением x_*. В общем случае y_{k+1} дает лучшее приближение к x_*, чем очередная итерация x_{k+2}. Подчеркнем, что главным предположением здесь является требование линейной сходимости основного итерационного метода. В случае методов, имеющих более высокую скорость сходимости (например метода Ньютона), ускорение по Эйткену в форме (15) неэффективно.

На практике не обязательно проводить пересчет по формуле (15) на каждой итерации k. Употребительны методы, в которых такой пересчет осуществляется циклически, т. е. через определенное число основных итераций.

С помощью метода Эйткена на основе известных итерационных методов можно получить иногда новые итерационные методы, об-

ладающие более высокой сходимостью. Рассмотрим, например, метод релаксации

$$\frac{x_{k+1}-x_k}{\tau}+f(x_k)=0 \tag{16}$$

(см. (6) из § 1), который имеет линейную сходимость, если

$$M_1>f'(x)>0, \quad 0<\tau<2/M_1.$$

Предположим, что при некотором k получены значения x_k, x_{k+1}, x_{k+2}. Вычислим согласно (15) величину

$$y_{k+1}=x_{k+2}-\frac{(x_{k+2}-x_{k+1})^2}{x_{k+2}-2x_{k+1}+x_k} \tag{17}$$

и исключим из (17) с помощью (16) величины x_{k+1}, x_{k+2}. Имеем

$$x_{k+1}=x_k-\tau f(x_k),$$
$$x_{k+2}=x_{k+1}-\tau f(x_{k+1})=x_k-\tau f(x_k)-\tau f(x_k-\tau f(x_k)),$$

следовательно,

$$y_{k+1}=x_k-\tau\frac{f^2(x_k)}{f(x_k)-f(x_k-\tau f(x_k))}.$$

Проведенные построения позволяют предположить, что одношаговый итерационный метод

$$y_{k+1}=y_k-\tau\frac{f^2(y_k)}{f(y_k)-f(y_k-\tau f(y_k))} \tag{18}$$

обладает более быстрой сходимостью, чем исходный метод релаксации (16). Действительно, как показано, например, в [25], метод (18) при $\tau=1$ (*метод Стеффенсена*) имеет квадратичную сходимость.

§ 3. Сходимость метода Ньютона

1. Простой вещественный корень. Предположим, что уравнение

$$f(x)=0 \tag{1}$$

имеет простой вещественный корень $x=x_*$, так что $f(x_*)=0$, $f'(x_*)\neq 0$. Будем предполагать, что $f(x)$ дважды непрерывно дифференцируема в окрестности корня x_*. Исследуем сходимость метода Ньютона

$$x_{k+1}=x_k-\frac{f(x_k)}{f'(x_k)}, \quad k=0, 1, \ldots \tag{2}$$

Заметим прежде всего, что (2) можно рассматривать как частный случай метода простой итерации

$$x_{k+1}=s(x_k), \quad k=0, 1, \ldots, \tag{3}$$

для которого

$$s(x)=x-\frac{f(x)}{f'(x)}. \tag{4}$$

В § 2 было показано, что для сходимости метода (3) достаточно потребовать, чтобы в некоторой окрестности искомого корня выполнялось неравенство

$$|s'(x)| \leqslant q < 1. \tag{5}$$

Для функции (4) имеем

$$s'(x) = \frac{f(x) f''(x)}{(f'(x))^2},$$

и если x_* — корень $f(x)$, то $s'(x_*) = 0$. Поэтому найдется окрестность корня, в которой выполнено неравенство (5). Тем самым при надлежащем выборе начального приближения метод Ньютона сходится. Однако следствием малости $s'(x)$ в окрестности x_* является не просто сходимость, а сходимость существенно более быстрая, чем в общем случае метода простой итерации. В следующей теореме доказано, что метод Ньютона имеет квадратичную сходимость, т. е. что он сходится и погрешность на $(k+1)$-й итерации пропорциональна квадрату погрешности на k-й итерации.

Теорема 1. *Пусть x_* — простой вещественный корень уравнения* (1) *и пусть $f'(x) \neq 0$ в окрестности*

$$U_r(x_*) = \{x: |x - x_*| < r\}.$$

Предположим, что $f''(x)$ непрерывна в $U_r(x_*)$ и

$$0 < m_1 = \inf_{x \in U_r(x_*)} |f'(x)|, \quad M_2 = \sup_{x \in U_r(x_*)} |f''(x)|, \tag{6}$$

причем

$$\frac{M_2 |x_0 - x_*|}{2m_1} < 1. \tag{7}$$

Тогда если $x_0 \in U_r(x_*)$, то метод Ньютона (2) *сходится, причем для погрешности справедлива оценка*

$$|x_k - x_*| \leqslant q^{2^k - 1} |x_0 - x_*|, \tag{8}$$

где

$$q = \frac{M_2 |x_0 - x_*|}{2m_1} < 1. \tag{9}$$

Доказательство. Из уравнения (2) получим

$$x_{k+1} - x_* = x_k - x_* - \frac{f(x_k)}{f'(x_k)}$$

или

$$x_{k+1} - x_* = \frac{F(x_k)}{f'(x_k)}, \tag{10}$$

где

$$F(x) = (x - x_*) f'(x) - f(x). \tag{11}$$

Заметим, что $F(x_*) = 0$ и

$$F'(x) = (x - x_*) f''(x). \tag{12}$$

Далее, воспользовавшись тождеством

$$F(x_k) = F(x_*) + \int_{x_*}^{x_k} F'(t)\,dt$$

и выражением (12) для $F'(t)$, получим

$$F(x_k) = \int_{x_*}^{x_k} (t - x_*) f''(t)\,dt.$$

Так как функция $t - x_*$ не меняет знак на отрезке интегрирования, можно воспользоваться формулой среднего значения и записать, что

$$F(x_k) = \int_{x_*}^{x_k} (t - x_*) f''(t)\,dt = f''(\xi_k) \int_{x_*}^{x_k} (t - x_*)\,dt = \frac{(x_k - x_*)^2}{2} f''(\xi_k),$$

где $\xi_k = \theta_k x_k + (1 - \theta_k) x_*$, $|\theta_k| < 1$. Обращаясь к (10), получим

$$x_{k+1} - x_* = \frac{f''(\xi_k)(x_k - x_*)^2}{2f'(x_k)}, \tag{13}$$

т. е. погрешность на $(k+1)$-й итерации пропорциональна квадрату погрешности на k-й итерации.

Докажем оценку (8) по индукции. При $k = 0$ из (13) получим

$$x_1 - x_* = \frac{f''(\xi_0)(x_0 - x_*)^2}{2f'(x_0)}. \tag{14}$$

По условию теоремы $x_0 \in U_r(x_*)$, и поэтому согласно первому из условий (6) имеем $|f'(x_0)| \geqslant m_1 > 0$. Кроме того, $\xi_0 = \theta_0 x_0 + (1 - \theta_0) x_*$,

$$\xi_0 - x_* = \theta_0 (x_0 - x_*), \quad |\xi_0 - x_*| \leqslant |\theta_0|\,|x_0 - x_*| < r,$$

т. е. $\xi_0 \in U_r(x_*)$. Но тогда согласно (6) $|f''(\xi_0)| \leqslant M_2$. Таким образом, приходим к оценке

$$|x_1 - x_*| \leqslant \frac{M_2 (x_0 - x_*)^2}{2m_1} = q\,|x_0 - x_*|,$$

совпадающей с оценкой (8) при $k = 1$.

Предположим, что оценка (8) выполняется при $k = l \geqslant 1$, и докажем, что она выполняется и при $k = l + 1$. При $k = l$ выражение (13) принимает вид

$$x_{l+1} - x_* = \frac{f''(\xi_l)(x_l - x_*)^2}{2f'(x_l)}. \tag{15}$$

Покажем, что x_l, $\xi_l \in U_r(x_*)$. Действительно, из (8) при $k = l$ имеем

$$|x_l - x_*| \leqslant q^{2^l - 1}\,|x_0 - x_*| < |x_0 - x_*| < r,$$

т. е. $x_l \in U_r(x_*)$. Кроме того,

$$\xi_l - x_* = \theta_l (x_l - x_*), \quad |\theta_l| < 1,$$

и, следовательно, $\xi_l \in U_r(x_*)$.

Теперь можно воспользоваться условиями (6) и оценить

$$|f'(x_l)| \geqslant m_1 > 0, \quad |f''(\xi_l)| \leqslant M_2.$$

Отсюда и из (15) получим

$$|x_{l+1} - x_*| \leqslant \frac{M_2 (x_l - x_*)^2}{2m_1}.$$

Из этого неравенства и из неравенства (8), имеющего следующий вид при $k=l$:

$$|x_l - x_*|^2 \leqslant q^{2^{l+1}-2} |x_0 - x_*|^2,$$

получим оценку

$$|x_{l+1} - x_*| \leqslant \left(\frac{M_2 |x_0 - x_*|}{2m_1}\right) q^{2^{l+1}-2} |x_0 - x_*| = q^{2^{l+1}-1} |x_0 - x_*|,$$

т. е. оценку (8) при $k=l+1$. Из оценки (8) следует сходимость метода (2), так как для $q \in (0, 1)$ правая часть неравенства (8) стремится к нулю при $k \to \infty$. Теорема 1 доказана.

Замечания. 1. Условие (7) означает, что начальное приближение надо брать достаточно близко к искомому корню.

2. Выполнение равенства (13) означает, что метод имеет квадратичную сходимость.

3. Поскольку x_* заранее неизвестен, иногда трудно проверить условие $x_0 \in U_r(x_*)$. Но если известно, что $|f'(x)| \geqslant m_1 > 0$ в некоторой окрестности корня, то для оценки близости начального приближения к корню можно воспользоваться неравенством

$$|x_0 - x_*| \leqslant |f(x_0)|/m_1. \tag{16}$$

Действительно,

$$f(x_0) = f(x_0) - f(x_*) = (x_0 - x_*) f'(\xi),$$

откуда и следует (16).

2. Кратные корни. Говорят, что x_* является корнем кратности p, если

$$f(x_*) = f'(x_*) = \ldots = f^{(p-1)}(x_*) = 0, \quad f^{(p)}(x_*) \neq 0$$

Будем предполагать сейчас, что $f^{(p+1)}(x)$ непрерывна в окрестности корня x_* кратности p. В случае корня кратности p квадратичную сходимость имеет *метод Ньютона с параметром*

$$f'(x_k) \frac{x_{k+1} - x_k}{\tau} + f(x_k) = 0, \tag{17}$$

где $\tau = p$. Справедлива

Теорема 2. *Пусть x_* — корень кратности p уравнения (1) и в окрестности*

$$U_r(x_*) = \{x: |x - x_*| < r\}$$

производная $f^{(p)}(x)$ отлична от нуля.

Пусть $f^{(p+1)}(x)$ непрерывна в $U_r(x_*)$ и

$$0 < m_p = \inf_{x \in U_r(x_*)} |f^{(p)}(x)|, \quad M_{p+1} = \sup_{x \in U_r(x_*)} |f^{(p+1)}(x)|,$$

причем

$$\frac{M_{p+1}\,|\,x_0 - x_*|}{m_p p\,(p+1)} < 1.$$

Тогда если $x_0 \in U_r(x_*)$, *то метод* (17) *при* $\tau = p$ *сходится, причем для погрешности справедлива оценка*

$$|\,x_k - x_*\,| \leqslant q^{2^k - 1}\,|\,x_0 - x_*\,|,$$

где

$$q = \frac{M_{p+1}\,|\,x_0 - x_*\,|}{m_p p\,(p+1)} < 1.$$

Доказательство теоремы 2 мало отличается от доказательства теоремы 1 (см. [25]). Для погрешности $x_{k+1} - x_*$ метода (17) с $\tau = p$ получаем выражение (10), где

$$F(x) = (x - x_*) f'(x) - p f(x).$$

При этом

$$F^{(m)}(x_*) = 0, \quad m = 0, 1, \ldots, p-1, p.$$

Применяя формулу Тейлора с остаточным членом в интегральной форме, получим, что

$$F(x_k) = \frac{1}{(p-1)!} \int_{x_*}^{x_k} (t - x_*)(x_k - t)^{p-1} f^{(p+1)}(t)\,dt.$$

Далее, воспользовавшись формулой среднего значения, получим представление $F(x_k)$ в виде

$$F(x_k) = \frac{f^{(p+1)}(\xi_k^{(1)})}{(p+1)!} (x_k - x_*)^{p+1}.$$

Для оценки знаменателя выражения (10) используется формула Тейлора с остаточным членом в форме Лагранжа. В результате получаем, что

$$f'(x_k) = \frac{(x_k - x_*)^{p-1}}{(p-1)!} f^{(p)}(\xi_k^{(2)}).$$

Далее повторяются те же рассуждения, что и при доказательстве теоремы 1.

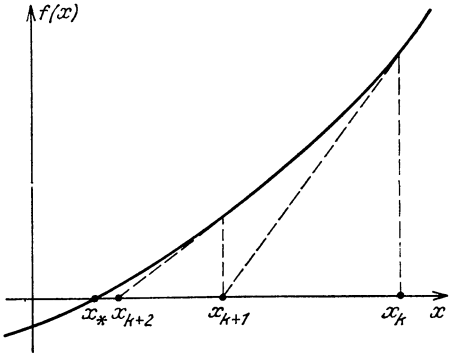


Рис. 6. Монотонная сходимость метода Ньютона

3. Односторонние приближения. Если вокрестности корня x_* производная функции $f(x)$ сохраняет знак и монотонна, то приближения x_k, получаемые в методе Ньютона, сходятся к x_* с одной стороны. Это означает, что последовательность $\{x_k\}$ либо монотонно убывает, так что $x_* < x_{k+1} < x_k$ для всех k, либо монотонно воз-

растает, так что $x_k < x_{k+1} < x_*$ для всех k. Монотонная сходимость метода Ньютона хорошо видна на рис. 6. Важное свойство монотонности метода Ньютона более точно сформулировано в теоремах 3, 4. В этих теоремах предполагается, что на отрезке $[a, b]$ уравнение (1) имеет единственный корень x_* и функция $f(x)$ дважды непрерывно дифференцируема.

Теорема 3. *Пусть для всех* $x \in [a, b]$ *либо*

$$f'(x) > 0, \quad f''(x) > 0, \tag{18}$$

либо

$$f'(x) < 0, \quad f''(x) < 0. \tag{19}$$

Тогда последовательность $\{x_k\}$, *определенная согласно* (2) *с* $x_0 = b$, *монотонно убывает и сходится к* x_*.

Теорема 4. *Пусть для всех* $x \in [a, b]$ *либо*

$$f'(x) > 0, \quad f''(x) < 0,$$

либо

$$f'(x) < 0, \quad f''(x) > 0.$$

Тогда последовательность $\{x_k\}$, *определенная согласно* (2) *с* $x_0 = a$, *монотонно возрастает и сходится к* x_*.

Поскольку формулировки и доказательства теорем 3, 4 совершенно аналогичны, ограничимся доказательством теоремы 3.

Доказательство теоремы 3. Монотонность последовательности $\{x_k\}$ докажем по индукции. По условию $x_0 = b$. Предположим, что для некоторого $k \geqslant 0$ выполняются неравенства

$$x_* < x_k \leqslant b, \tag{20}$$

и докажем, что тогда

$$x_* < x_{k+1} < x_k. \tag{21}$$

Перепишем уравнение (2) в виде

$$x_k - x_{k+1} = \frac{f(x_k) - f(x_*)}{f'(x_k)}$$

и воспользуемся формулой конечных приращений Лагранжа. Тогда получим

$$x_k - x_{k+1} = \frac{(x_k - x_*) f'(\xi_k)}{f'(x_k)}, \tag{22}$$

где $\xi_k \in (x_*, x_k)$. Пусть выполнено условие (18). Тогда

$$0 < \frac{f'(\xi_k)}{f'(x_k)} < 1, \tag{23}$$

причем последнее неравенство является следствием монотонного возрастания $f'(x)$. Те же самые неравенства (23) выполняются и в случае условий (19). Таким образом,

$$0 < \frac{(x_k - x_*) f'(\xi_k)}{f'(x_k)} < x_k - x_*,$$

и из (22) получим

$$0<x_k-x_{k+1}<x_k-x_*,$$

т. е. получим требуемые неравенства (21). Таким образом, последовательность $\{x_k\}$ монотонно убывает и ограничена снизу числом x_*. Поэтому данная последовательность имеет предел, который в силу непрерывности функции $f(x)$ и условия $f'(x_*)\neq 0$ совпадает с корнем x_* уравнения (1). Теорема 3 доказана.

Сделаем замечания относительно скорости сходимости метода Ньютона при условиях теорем 3 и 4. Если начальное приближение x_0 выбрано достаточно близко к искомому корню, так что выполняется условие (7), то согласно теореме 1 метод имеет квадратичную сходимость и для погрешности справедлива оценка (8).

Если же условие (7) не выполнено, то на начальных итерациях погрешность будет убывать более медленно. Однако в силу сходимости последовательности $\{x_k\}$ найдется номер $k=k_0$, для которого очередное приближение x_{k_0} удовлетворит неравенству

$$\frac{M_2|x_{k_0}-x_*|}{2m_1}<1,$$

и с этого момента сходимость станет квадратичной.

4. Комплексный корень. Пусть $f(z)$ — функция комплексного переменного $z=x+iy$ и $z_*=x_*+iy_*$ — простой нуль $f(z)$. Будем считать, что $f(z)$ аналитична в некоторой окрестности z_*. Тогда можно рассматривать метод Ньютона

$$z_{k+1}=z_k-\frac{f(z_k)}{f'(z_k)}, \quad k=0, 1, \ldots \tag{24}$$

Сходимость метода (24) устанавливается в теореме 5, которая обобщает на комплексный случай теорему 1.

Теорема 5. *Пусть z_* — простой корень уравнения $f(z)=0$ и пусть $f(z)$ аналитична в круге*

$$U_r(z_*)=\{z:\ |z-z_*|<r\}.$$

Предположим, что

$$\inf_{z\in U_r(z_*)}|f'(z)|=m_1>0, \quad \sup_{z\in U_r(z_*)}|f''(z)|=M_2, \tag{25}$$

причем

$$\frac{M_2|z_0-z_*|}{2m_1}<1. \tag{26}$$

Тогда если $z_0\in U_r(z_*)$, то метод (24) сходится, причем для погрешности справедлива оценка

$$|z_k-z_*|\leqslant q^{2^k-1}|z_0-z_*|, \tag{27}$$

где

$$q=\frac{M_2|z_0-z_*|}{2m_1}<1. \tag{28}$$

Доказательство. Для погрешности получаем уравнение

$$z_{k+1} - z_* = \frac{F(z_k)}{f'(z_k)}, \tag{29}$$

где

$$F(z) = (z - z_*) f'(z) - f(z), \; F'(z) = (z - z_*) f''(z).$$

Воспользовавшись формулой Ньютона — Лейбница, получим

$$F(z_k) = F(z_*) + \int_{z_*}^{z_k} F'(z)\, dz,$$

т. е.

$$F(z_k) = \int_{z_*}^{z_k} (z - z_*) f''(z)\, dz. \tag{30}$$

Докажем оценку (27) по индукции. При $k=0$ из (29) получим

$$z_1 - z_* = \frac{F(z_0)}{f'(z_0)}. \tag{31}$$

Так как $z_0 \in U_r(z_*)$, имеем согласно (25), что $|f'(z_0)| \geqslant m_1 > 0$. Далее, оценим

$$F(z_0) = \int_{z_*}^{z_0} (z - z_*) f''(z)\, dz. \tag{32}$$

Для этого сделаем в интеграле (32) замену переменной

$$z = tz_0 + (1-t) z_*$$

и перепишем его в виде

$$F(z_0) = (z_0 - z_*)^2 \int_0^1 t f''(tz_0 + (1-t) z_*)\, dt. \tag{33}$$

Имеем

$$z - z_* = t(z_0 - z_*), \quad |z - z_*| \leqslant |z_0 - z_*| < r,$$

т. е. $z = tz_0 + (1-t) z_* \in U_r(z)$, и согласно (25) выполняется оценка

$$|f''(tz_0 + (1-t) z_*)| \leqslant M_2.$$

Отсюда и из (33) получаем оценку

$$|F(z_0)| \leqslant M_2 |z_0 - z_*|^2 \int_0^1 t\, dt = 0{,}5 M_2 |z_0 - z_*|^2.$$

Учитывая (31), получим неравенство

$$|z_1 - z_*| \leqslant \frac{M_2 |z_0 - z_*|^2}{2m_1},$$

которое совпадает с неравенством (27) при $k=1$.

Предположим, что оценка (27) выполняется при $k = l \geqslant 1$, и докажем, что она выполняется и при $k = l+1$. Заметим прежде всего,

что из оценки (27) при $k=l$ следует, что $z_l \in U_r(z_*)$. Поэтому согласно условию (25) имеем $|f'(z_l)| \geqslant m_1 > 0$.

Далее, оценим

$$F(z_l) = \int_{z_*}^{z_l} (z - z_*) f''(z)\, dz.$$

Учитывая, что $z_l \in U_r(z_*)$, можно получить оценку этого интеграла так же, как и оценку $F(z_0)$, а именно

$$|F(z_l)| \leqslant \frac{M_2}{2} |z_l - z_*|^2.$$

Тогда из (29) при $k=l$ получим

$$|z_{l+1} - z_*| \leqslant \frac{M_2 |z_l - z_*|^2}{2m_1},$$

и, учитывая (27) при $k=l$, придем к неравенству (27) при $k=l+1$. Теорема 5 доказана.

Заметим, что условие сходимости (26) означает близость на комплексной плоскости начального приближения z_0 к искомому корню z_*. В частности, это условие может не выполняться для вещественных начальных приближений.

При численной реализации метода Ньютона можно пользоваться комплексной арифметикой, однако иногда бывает удобнее разделить в формулах (24) действительные и мнимые части и проводить вычисления только с вещественными числами.

§ 4. Итерационные методы для систем нелинейных уравнений

1. Общие понятия. Рассмотрим систему нелинейных уравнений

$$\begin{aligned} &f_1(x_1, x_2, \ldots, x_m) = 0, \\ &f_2(x_1, x_2, \ldots, x_m) = 0, \\ &\ldots\ldots\ldots\ldots \\ &f_m(x_1, x_2, \ldots, x_m) = 0, \end{aligned} \tag{1}$$

где f_i, $i=1, 2, \ldots, m$,— функции вещественных переменных $x_1, \ldots, x_m$. В дальнейшем систему (1) будем рассматривать как операторное уравнение в некотором линейном пространстве H размерности m. Обозначим

$$x = (x_1, x_2, \ldots, x_m)^T \in H,$$

$$F(x) = (f_1(x), f_2(x), \ldots, f_m(x))^T$$

и запишем (1) в виде операторного уравнения

$$F(x) = 0, \tag{2}$$

где F: $H \to H$ — отображение, нелинейное, вообще говоря, из H в H.

Многие одношаговые итерационные методы для решения системы (2) можно записать в виде

$$B_{k+1}\frac{x^{k+1}-x^k}{\tau_{k+1}}+F(x^k)=0,\quad k=0,\ 1,\ \ldots,\ x_0\ \text{задан},\tag{3}$$

где k — номер итерации,

$$x^k=(x_1^k,\ x_2^k,\ \ldots,\ x_m^k)^T,$$

τ_{k+1} — числовые параметры, B_{k+1} — матрица $m\times m$, имеющая обратную.

Если F — линейный оператор, то (3) совпадает с канонической формой одношагового итерационного метода (см. § 1 гл. 2), т. е. в виде (3) можно записать любой одношаговый метод для линейной системы уравнений. В случае нелинейной системы (1) возможны методы, содержащие новую итерацию x^{k+1} нелинейно, и тем самым не представимые в виде (3). Однако мы по-прежнему будем называть канонической формой запись итерационного метода в виде (3).

Для нахождения x^{k+1} по известному x^k из уравнения (3) необходимо решить систему линейных алгебраических уравнений

$$B_{k+1}x^{k+1}=g(x^k),\tag{4}$$

где $g(x^k)=B_{k+1}x^k-\tau_{k+1}F(x^k)$. Метод (3) называется *явным*, если $B_{k+1}=E$ для всех $k=0,\ 1,\ \ldots$, и *неявным* — в противном случае. Метод (3) называется *стационарным*, если B и τ не зависят от номера итерации k. Систему линейных уравнений (4) можно решать либо прямым, либо итерационным методом. В последнем случае итерации, приводящие к решению системы (4), называются *внутренними итерациями*, а итерации (3) — *внешними итерациями*.

2. Сходимость стационарного метода. Остановимся кратко на вопросе о сходимости метода (3). Предположим, что метод (3) — стационарный, т. е. B и τ не зависят от k. Тогда уравнение (3) можно переписать в виде

$$x^{k+1}=S(x^k),\tag{5}$$

а исходное уравнение (2) — в виде

$$x=S(x),\tag{6}$$

где $S(x)=x-\tau B^{-1}F(x)$.

Будем считать, что H — конечномерное линейное нормированное пространство, т. е. что определен функционал $\|x\|$, удовлетворяющий всем аксиомам нормы.

Точка $x_*\in H$, для которой $S(x_*)=x_*$, называется *неподвижной точкой оператора* S. Очевидно, что точка x_* является решением операторного уравнения (2) тогда и только тогда, когда она является неподвижной точкой оператора S. Таким образом, отыскание корней уравнения (2) эквивалентно отысканию неподвижных точек оператора S.

Говорят, что S является *сжимающим оператором на множестве* $K\subseteq H$ *с коэффициентом сжатия* q, если существует число $q\in(0,\ 1)$ такое, что для любых $x',\ x''\in K$ выполняется неравенство

$$\|S(x')-S(x'')\|\leqslant q\|x'-x''\|.$$

Теперь мы в состоянии сформулировать теорему, которая называется *принципом сжимающих отображений* и содержит условия сходимости метода простой итерации

$$x^{k+1}=S(x^k) \tag{7}$$

в конечномерном линейном нормированном пространстве H. Она является многомерным аналогом теоремы 1 из § 2.

Теорема 1. *Пусть оператор S определен на множестве*

$$\overline{U}_r(a)=\{x\in H:\ \|x-a\|\leqslant r\}$$

и является сжимающим оператором на этом множестве с коэффициентом сжатия q, причем

$$\|S(a)-a\|\leqslant(1-q)r, \qquad 0<q<1. \tag{8}$$

Тогда в $\overline{U}_r(a)$ оператор S имеет единственную неподвижную точку x_* и итерационный метод (7) сходится к x_* при любом $x^0\in\overline{U}_r(a)$. Для погрешности справедливы оценки

$$\|x^k-x_*\|\leqslant q^k\|x^0-x_*\|, \tag{9}$$

$$\|x^k-x_*\|\leqslant\frac{q^k}{1-q}\|S(x^0)-x^0\|. \tag{10}$$

Доказательство теоремы 1 можно найти в [42].

3. Примеры итерационных методов.

Пример 1. *Метод релаксации* представляет собой частный случай метода (3), когда $B_{k+1}=E$, $\tau_{k+1}=\tau$. Это стационарный итерационный метод, который можно записать в виде

$$x^{k+1}=S(x^k),$$

где

$$S(x)=x-\tau F(x).$$

Метод сходится, если $\|S'(x_*)\|<1$. В данном случае $S'(x)=E-\tau F'(x)$ и

$$F'(x)=\begin{bmatrix} \dfrac{\partial f_1(x)}{\partial x_1} & \dfrac{\partial f_1(x)}{\partial x_2} & \cdots & \dfrac{\partial f_1(x)}{\partial x_m} \\ \dfrac{\partial f_2(x)}{\partial x_1} & \dfrac{\partial f_2(x)}{\partial x_2} & \cdots & \dfrac{\partial f_2(x)}{\partial x_m} \\ \cdots & \cdots & \cdots & \cdots \\ \dfrac{\partial f_m(x)}{\partial x_1} & \dfrac{\partial f_m(x)}{\partial x_2} & \cdots & \dfrac{\partial f_m(x)}{\partial x_m} \end{bmatrix}. \tag{11}$$

Пример 2. *Метод Пикара.* Пусть $F(x)$ представляется в виде

$$F(x)=Ax+G(x),$$

где A — матрица $m\times m$. Тогда итерации можно определить следующим образом:

$$Ax^{k+1}+G(x^k)=0.$$

Итерационный метод можно переписать в виде

$$A(x^{k+1}-x^k)+F(x^k)=0,$$

т. е. в канонической форме (3) с $B_{k+1}=A$, $\tau_{k+1}=1$. Можно и здесь ввести итерационный параметр и рассматривать более общий метод

$$A\frac{x^{k+1}-x^k}{\tau}+F(x^k)=0.$$

Пример 3. *Метод Ньютона* для системы уравнений (1) строится следующим образом.

Пусть приближение $x^k=(x_1^k, x_2^k, \ldots, x_m^k)^T$ уже известно. Выпишем разложение функции $f_i(x_1, x_2, \ldots, x_m)$ по формуле Тейлора в точке x^k,

$$f_i(x_1, x_2, \ldots, x_m)=f_i(x_1^k, x_2^k, \ldots, x_m^k)+(x_1-x_1^k)\frac{\partial f_i(x^k)}{\partial x_1}+$$

$$+(x_2-x_2^k)\frac{\partial f_i(x^k)}{\partial x_2}+\ldots+(x_m-x_m^k)\frac{\partial f_i(x^k)}{\partial x_m}+O(|x-x^k|^2),$$

и отбросим величины второго порядка малости. Тогда система (1) заменится системой уравнений

$$\sum_{j=1}^{m}(x_j-x_j^k)\frac{\partial f_i(x^k)}{\partial x_j}+f_i(x^k)=0, \quad i=1, 2, \ldots, m, \tag{12}$$

линейной относительно приращений $x_j-x_j^k$, $j=1, 2, \ldots, m$. Решение $x=(x_1, x_2, \ldots, x_m)^T$ системы (12) примем за следующее приближение и обозначим через

$$x^{k+1}=(x_1^{k+1}, \ldots, x_m^{k+1})^T.$$

Таким образом, итерационный метод Ньютона для (1) определяется системой уравнений

$$\sum_{j=1}^{m}(x_j^{k+1}-x_j^k)\frac{\partial f_i(x^k)}{\partial x_j}+f_i(x^k)=0, \quad i=1, 2, \ldots, m, \tag{13}$$

из которой последовательно, начиная с заданного $x^0=(x_1^0, \ldots, x_m^0)^T$, находятся векторы x^k, $k=1, 2, \ldots$

Систему (13) можно записать в векторном виде

$$F'(x^k)(x^{k+1}-x^k)+F(x^k)=0, \; k=0, 1, \ldots, \; x^0 \text{ задан}, \tag{14}$$

где матрица $F'(x)$ определена согласно (11). Таким образом, метод Ньютона имеет канонический вид (3), где

$$B_{k+1}=F'(x^k), \quad \tau_{k+1}=1.$$

Для реализации метода Ньютона необходимо существование матриц $(F'(x^k))^{-1}$, обратных $F'(x^k)$. По поводу сходимости метода Ньютона для систем уравнений можно сказать то же, что и в слу-

чае одного уравнения, а именно, метод имеет квадратичную сходимость, если начальное приближение выбрано достаточно хорошо.

Приведем без доказательства одну из теорем о сходимости метода Ньютона.

Пусть E^m — множество m-мерных вещественных векторов с нормой $\|x\| = \left(\sum_{i=1}^{m} x_j^2\right)^{1/2}$, $\|A\|$ — норма матрицы A, подчиненная данной норме вектора. Обозначим

$$U_r(x^0) = \{x \in E^m: \|x - x^0\| < r\}$$

и предположим, что в шаре $U_r(x^0)$ функции $f_i(x)$, $i = 1, 2, \ldots, m$, непрерывно дифференцируемы.

Теорема 2. *Предположим, что в $U_r(x^0)$ матрица $F'(x)$ удовлетворяет условию Липшица с постоянной L,* т. е.

$$\|F'(x^1) - F'(x^2)\| \leqslant L\|x^1 - x^2\|$$

для любых $x^1, x^2 \in U_r(x^0)$. Пусть в $U_r(x^0)$ матрица $(F'(x))^{-1}$ существует, причем элементы ее непрерывны и

$$\|(F'(x))^{-1}\| \leqslant M.$$

Если начальное приближение x^0 таково, что $\|F(x_0)\| \leqslant \eta$ и

$$q = \frac{M^2 L \eta}{2} < 1,$$

причем

$$M\eta \sum_{k=0}^{\infty} q^{2^k - 1} < r,$$

то система уравнений (2) *имеет решение $x_* \in \overline{U}_r(x_0)$, к которому сходится метод Ньютона* (14). *Оценка погрешности дается неравенством*

$$\|x^k - x_*\| \leqslant M\eta \frac{q^{2^k - 1}}{1 - q^{2^k}}.$$

Доказательство теоремы 2 можно найти в [42].

Пример 4. *Модифицированный метод Ньютона* имеет вид

$$F'(x^0)(x^{k+1} - x^k) + F(x^k) = 0 \tag{15}$$

и обладает линейной сходимостью. Упрощение в численной реализации по сравнению с обычным методом Ньютона состоит в том, что матрицу $F'(x)$ надо обращать не на каждой итерации, а лишь один раз. Возможно циклическое применение модифицированного метода Ньютона, когда $F'(x)$ обращается через определенное число итераций.

Пример 5. *Метод Ньютона с параметром* имеет вид

$$F'(x^k)\frac{x^{k+1} - x^k}{\tau_{k+1}} + F(x^k) = 0. \tag{16}$$

Рассмотренные до сих пор методы являлись линейными относительно новой итерации x^{k+1}. Возможны и нелинейные методы, ког-

да для вычисления x^{k+1} приходится решать нелинейные системы уравнений. Приведем примеры таких методов.

Пример 6. *Нелинейный метод Якоби* для системы (1) имеет вид

$$f_i(x_1^k, x_2^k, \ldots, x_{i-1}^k, x_i^{k+1}, x_{i+1}^k, \ldots, x_m^k) = 0, \quad i = 1, 2, \ldots, m. \tag{17}$$

Здесь для отыскания x^{k+1} необходимо решить m независимых скалярных уравнений. Для решения скалярного уравнения можно применить какой-либо из итерационных методов, рассмотренных в § 1, причем не обязательно применять один и тот же метод для всех уравнений.

Пример 7. *Нелинейный метод Зейделя* состоит в последовательном решении уравнений

$$f_i(x_1^{k+1}, x_2^{k+1}, \ldots, x_i^{k+1}, x_{i+1}^k, \ldots, x_m^k) = 0 \tag{18}$$

относительно переменной x_i^{k+1}, $i = 1, 2, \ldots, m$.

Большое распространение получили гибридные методы, когда внешние итерации осуществляются одним методом, а внутренние — другим. При этом число внутренних итераций может быть фиксированным и не очень большим, так что внутренние итерации не доводятся до сходимости. В результате получается некоторый новый метод, сочетающий свойства исходных методов. Приведем примеры таких методов.

Пример 8. *Внешние итерации — по Зейделю и внутренние — по Ньютону.* Здесь в качестве основной (внешней) итерации выбирается нелинейный метод Зейделя (18), а для нахождения x_i^{k+1} используется метод Ньютона. Обозначим $y_i = x_i^{k+1}$. Тогда итерации определяются следующим образом:

$$\frac{\partial f_i}{\partial x_i}(x_1^{k+1}, x_2^{k+1}, \ldots, x_{i-1}^{k+1}, y_i^s, x_{i+1}^k, \ldots, x_m^k)(y_i^{s+1} - y_i^s) + f(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, y_i^s, x_{i+1}^k, \ldots, x_m^k) = 0,$$

$$s = 0, 1, \ldots, l, \quad y_i^0 = x_i^k, \quad y_i^{l+1} = x_i^{k+1}, \quad i = 1, 2, \ldots, m. \tag{19}$$

Здесь индексом s обозначен номер внутренней итерации.

Иногда в (19) делают всего одну внутреннюю итерацию, полагая $l = 0$, $y_i^0 = x_i^k$, $y_i^1 = x_i^{k+1}$. Тогда приходят к следующему итерационному методу:

$$\frac{\partial f_i}{\partial x_i}(x_1^{k+1}, x_2^{k+1}, \ldots, x_{i-1}^{k+1}, x_i^k, x_{i+1}^k, \ldots, x_m^k)(x_i^{k+1} - x_i^k) + f_i(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i^k, x_{i+1}^k, \ldots, x_m^k) = 0, \quad k = 0, 1, \ldots \tag{20}$$

В частности, при $m=2$ метод (20) принимает вид

$$\frac{\partial f_1(x_1^k, x_2^k)}{\partial x_1}(x_1^{k+1}-x_1^k)+f_1(x_1^k, x_2^k)=0,$$

$$\frac{\partial f_2}{\partial x_2}(x_1^{k+1}, x_2^k)(x_2^{k+1}-x_2^k)+f_2(x_1^{k+1}, x_2^k)=0. \tag{21}$$

Пример 9. *Внешние итерации — по Ньютону и внутренние — по Зейделю.* Запишем метод Ньютона для системы (2) в виде

$$F'(x^k)(x^{k+1}-x^k)+F(x^k)=0, \tag{22}$$

где $F'(x^k)=(a_{ij})$, $a_{ij}=\dfrac{\partial f_i(x^k)}{\partial x_j}$, $i, j=1, 2, \ldots, m$. Для решения системы линейных уравнений (22) воспользуемся методом Зейделя. Напомним (см. § 1 гл. 2), что для линейной системы

$$Aw+F=0 \tag{23}$$

метод Зейделя строится следующим образом. Матрица A представляется в виде суммы $A=A_-+D+A_+$, где матрицы A_-, A_+, D соответственно нижняя треугольная, верхняя треугольная и диагональная. Итерации метода Зейделя строятся по правилу

$$(A_-+D)w^{s+1}+A_+w^s+F=0, \quad s=0, 1, \ldots, l, \tag{24}$$

и система (24) решается путем обращения нижней треугольной матрицы A_-+D.

В случае системы (22) надо положить $A=F'(x^k)$, вычислить последовательно векторы w^s согласно (24), начиная с $w^0=0$, и положить $w^{l+1}=x^{k+1}-x^k$, так что $x^{k+1}=x^k+w^{l+1}$.

Заметим, что итерации по Зейделю можно осуществлять и относительно вектора x^{k+1}.

Пусть в (24) совершается только одна итерация, т. е. $l=0$. Тогда, учитывая, что $w^0=0$, $w^1=x^{k+1}-x^k$, получим метод

$$(A_-+D)(x^{k+1}-x^k)+F(x^k)=0, \tag{25}$$

где A_-+D — «нижняя треугольная» часть матрицы Якоби (11), вычисленной при $x=x^k$.

В частности, при $m=2$ метод (25) принимает вид

$$\frac{\partial f_1}{\partial x_1}(x_1^k, x_2^k)(x_1^{k+1}-x_1^k)+f_1(x_1^k, x_2^k)=0,$$

$$\frac{\partial f_2}{\partial x_1}(x_1^k, x_2^k)(x_1^{k+1}-x_1^k)+\frac{\partial f_2}{\partial x_2}(x_1^k, x_2^k)(x_2^{k+1}-x_2^k)+f_2(x_1^k, x_2^k)=0. \tag{26}$$

Сопоставление (21) и (26) показывает, что методы, рассмотренные в двух последних примерах, не совпадают.

ГЛАВА 6

ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ ЗАДАЧИ КОШИ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

§ 1. Исходная задача и примеры численных методов ее решения

1. Постановка исходной задачи. Будем рассматривать задачу Коши для системы обыкновенных дифференциальных уравнений

$$\frac{du(t)}{dt} = f(t, u), \quad t > 0, \quad u(0) = u^{(0)} \tag{1}$$

или, подробнее,

$$\frac{du_i(t)}{dt} = f_i(t, u_1, u_2, \ldots, u_m), \quad t > 0, \quad i = 1, 2, \ldots, m, \tag{2}$$

$$u_i(0) = u_i^{(0)}, \quad i = 1, 2, \ldots, m. \tag{3}$$

Хорошо известны условия, гарантирующие существование и единственность решения задачи Коши (см. [39, с. 49]). Предположим, что функции f_i, $i = 1, 2, \ldots, m$, непрерывны по всем аргументам в замкнутой области

$$D = \{|t| \leqslant a, \quad |u_i - u_i^{(0)}| \leqslant b, \quad i = 1, 2, \ldots, m\}.$$

Из непрерывности функций f_i следует их ограниченность, т. е. существование константы $M > 0$ такой, что всюду в D выполняются неравенства $|f_i| \leqslant M$, $i = 1, 2, \ldots, m$.

Предположим, кроме того, что в D функции f_i удовлетворяют условию Липшица по аргументам $u_1, u_2, \ldots, u_m$, т. е.

$$|f_i(t, u_1', u_2', \ldots, u_m') - f_i(t, u_1'', u_2'', \ldots, u_m'')| \leqslant$$
$$\leqslant L\{|u_1' - u_1''| + |u_2' - u_2''| + \ldots + |u_m' - u_m''|\}$$

для любых точек $(t, u_1', \ldots, u_m')$ и $(t, u_1'', u_2'', \ldots, u_m'')$ области D.

Если выполнены сформулированные выше предположения, то существует единственное решение

$$u_1 = u_1(t), \; u_2 = u_2(t), \ldots, u_m = u_m(t)$$

системы (2), определенное при $|t| \leqslant t_0 = \min(a, b/M)$ и принимающее при $t = 0$ заданные начальные значения (3).

При исследовании численных методов для задачи Коши будем заранее предполагать, что ее решение существует, единственно и обладает необходимыми свойствами гладкости.

2. Примеры численных методов. Существуют две группы численных методов решения задачи Коши: многошаговые разностные методы и методы Рунге — Кутта. Приведем примеры и поясним основные понятия, возникающие при использовании численных методов. Для простоты изложения будем рассматривать сейчас одно

уравнение

$$\frac{du}{dt} = f(t, u), \quad t > 0, \quad u(0) = u_0. \tag{4}$$

Введем по переменному t *равномерную сетку с шагом* $\tau > 0$, т. е. рассмотрим множество точек

$$\omega_\tau = \{t_n = n\tau, \ n = 0, 1, 2, \ldots\}.$$

Будем обозначать через $u(t)$ точное решение задачи (4), а через $y_n = y(t_n)$ — приближенное решение. Заметим, что приближенное решение является *сеточной функцией*, т. е. определено только в точках сетки ω_τ.

Пример 1. *Метод Эйлера.* Уравнение (4) заменяется разностным уравнением

$$\frac{y_{n+1} - y_n}{\tau} - f(t_n, y_n) = 0, \quad n = 0, 1, 2, \ldots, \quad y_0 = u_0. \tag{5}$$

Решение этого уравнения находится явным образом по рекуррентной формуле

$$y_{n+1} = y_n + \tau f(t_n, y_n), \quad n = 0, 1, \ldots, y_0 = u_0.$$

При использовании приближенных методов основным является вопрос о сходимости. Понятие сходимости приближенного метода можно сформулировать по-разному. Применительно к разностным методам, к которым относится и метод Эйлера (5), наибольшее распространение получило понятие *сходимости при* $\tau \to 0$. Оно означает следующее. Фиксируем точку t и построим последовательность сеток ω_τ таких, что $\tau \to 0$ и $t_n = n\tau = t$ (тогда необходимо $n \to \infty$). Говорят, что *метод* (5) *сходится в точке* t, если $|y_n - u(t_n)| \to 0$ при $\tau \to 0$, $t_n = t$.

Метод *сходится на* отрезке $(0, T]$, если он сходится в каждой точке $t \in (0, T]$.

Говорят, что *метод имеет* p*-й порядок точности*, если существует число $p > 0$ такое, что $|y_n - u(t_n)| = O(\tau^p)$ при $\tau \to 0$.

Получим уравнение, которому удовлетворяет *погрешность метода* $z_n = y_n - u(t_n)$. Подставляя $y_n = z_n + u_n$ в (5), получим

$$\frac{z_{n+1} - z_n}{\tau} = f(t_n, u_n + z_n) - \frac{u_{n+1} - u_n}{\tau}. \tag{6}$$

Правую часть уравнения (6) можно представить в виде суммы

$$\psi_n^{(1)} + \psi_n^{(2)},$$

где

$$\psi_n^{(1)} = -\frac{u_{n+1} - u_n}{\tau} + f(t_n, u_n),$$

$$\psi_n^{(2)} = f(t_n, u_n + z_n) - f(t_n, u_n).$$

Функция $\psi_n^{(1)}$ называется *невязкой* или *погрешностью аппроксимации разностного уравнения* (5) *на решении исходного уравнения* (4). Видно, что невязка представляет собой результат подста-

новки точного решения $u=u(t)$ в левую часть разностного уравнения (5). Если бы приближенное решение y_n совпадало с точным $u(t_n)$, то невязка равнялась бы нулю. Говорят, что *разностный метод аппроксимирует исходное дифференциальное уравнение*, если $\psi_n^{(1)}\to 0$ при $\tau\to 0$. Разностный метод имеет *p-й порядок аппроксимации*, если $\psi_n^{(1)}=O(\tau^p)$. В дальнейшем будет показано, что при очень общих предположениях порядок точности разностного метода совпадает с порядком аппроксимации.

Функция

$$\psi_n^{(2)}=f(t_n, u_n+z_n)-f(t_n, u_n)$$

обращается в нуль, если правая часть f не зависит от решения u. В общем случае $\psi_n^{(2)}$ пропорциональна погрешности z_n, так как по формуле конечных приращений имеем

$$\psi_n^{(2)}=\frac{\partial f}{\partial u}(t_n, u_n+\theta z_n)\, z_n, \qquad |\theta|\leqslant 1.$$

Порядок аппроксимации метода Эйлера (5) нетрудно найти, используя разложение по формуле Тейлора. Поскольку

$$\frac{u_{n+1}-u_n}{\tau}=u'(t_n)+O(\tau),$$

то в силу уравнения (4)

$$\psi_n^{(1)}=-u'(t_n)+f(t_n, u_n)+O(\tau)=O(\tau),$$

т. е. метод Эйлера имеет первый порядок аппроксимации. При выводе предполагалась ограниченность $u''(t)$.

Пример 2. *Симметричная схема.* Уравнение (4) заменяется разностным уравнением

$$\frac{y_{n+1}-y_n}{\tau}-\frac{1}{2}(f(t_n, y_n)+f(t_{n+1}, y_{n+1}))=0, \quad n=0, 1, \ldots, y_0=u_0. \quad (7)$$

Данный метод более сложен в реализации, чем метод Эйлера (5), так как новое значение y_{n+1} определяется по найденному ранее y_n путем решения уравнения

$$y_{n+1}-0{,}5\tau f(t_{n+1}, y_{n+1})=F_n,$$

где $F_n=y_n+0{,}5\tau f(t_n, y_n)$. По этой причине метод называется *неявным*. Преимуществом метода (7) по сравнению с (5) является более высокий порядок точности.

Для невязки

$$\psi_n^{(1)}=-\frac{u_{n+1}-u_n}{\tau}+\frac{1}{2}(f(t_n, y_n)+f(t_{n+1}, y_{n+1}))$$

справедливо разложение

$$\psi_n^{(1)}=-u_n'-\frac{\tau}{2}u_n''+O(\tau^2)+\frac{1}{2}(u_n'+u_{n+1}')=$$

$$=-u_n'-\frac{\tau}{2}u_n''+\frac{1}{2}(u_n'+u_n'+\tau u_n''+O(\tau^2)),$$

т. е. $\psi_n^{(1)} = O(\tau^2)$. Таким образом, метод (7) имеет второй порядок аппроксимации. Из результатов § 3 будет следовать, что он имеет и второй порядок точности.

Приведенные примеры представляют собой простейшие случаи *разностных методов*, или, как их еще называют, *разностных схем*. *Методы Рунге — Кутта* отличаются от разностных методов тем, что в них допускается вычисление правых частей $f(t, u)$ не только в точках сетки, но и в некоторых промежуточных точках.

Пример 3. *Метод Рунге — Кутта второго порядка точности.* Предположим, что приближенное значение y_n решения исходной задачи в момент $t = t_n$ уже известно. Для нахождения $y_{n+1} = y(t_{n+1})$ поступим следующим образом. Сначала, используя схему Эйлера

$$\frac{y_{n+1/2} - y_n}{0{,}5\tau} = f(t_n, y_n), \tag{8}$$

вычислим промежуточное значение $y_{n+1/2}$, а затем воспользуемся разностным уравнением

$$\frac{y_{n+1} - y_n}{\tau} = f(t_n + 0{,}5\tau, y_{n+1/2}), \tag{9}$$

из которого явным образом найдем искомое значение y_{n+1}.

Для исследования невязки подставим промежуточное значение $y_{n+1/2} = y_n + 0{,}5\tau f_n$, где $f_n = f(t_n, y_n)$, в уравнение (9). Тогда получим разностное уравнение

$$\frac{y_{n+1} - y_n}{\tau} - f(t_n + 0{,}5\tau, y_n + 0{,}5\tau f_n) = 0, \tag{10}$$

невязка которого равна

$$\psi_n^{(1)} = -\frac{u_{n+1} - u_n}{\tau} + f(t_n + 0{,}5\tau, u_n + 0{,}5\tau f(t_n, u_n)). \tag{11}$$

Имеем

$$\frac{u_{n+1} - u_n}{\tau} = u_n' + 0{,}5\tau u_n'' + O(\tau^2),$$

$$f(t_n + 0{,}5\tau, u_n + 0{,}5\tau f(t_n, u_n)) = f(t_n, u_n) +$$

$$+ 0{,}5\tau \left(\frac{\partial f(t_n, u_n)}{\partial t} + 0{,}5\tau f(t_n, u_n) \frac{\partial f(t_n, u_n)}{\partial u} \right) = f(t_n, u_n) + 0{,}5\tau u_n'',$$

так как в силу (4) справедливо равенство

$$\frac{d^2u}{dt^2} = \frac{\partial f}{\partial t} + f\frac{\partial f}{\partial u}.$$

Таким образом, метод (10) имеет второй порядок погрешности аппроксимации, $\psi_n^{(1)} = O(\tau^2)$, и в отличие от (7) является явным.

Реализация метода (10) в виде двух этапов (8), (9) называется *методом предиктор — корректор* (предсказывающе-исправляющим), поскольку на первом этапе (8) приближенное значение пред-

сказывается с невысокой точностью $O(\tau)$, а на втором этапе (9) это предсказанное значение исправляется, так что результирующая погрешность имеет второй порядок по τ.

Тот же самый метод (10) можно реализовать несколько иначе. А именно, сначала вычислим последовательно функции

$$k_1=f(t_n, y_n), \quad k_2=f(t_n+0{,}5\tau, y_n+0{,}5\tau k_1),$$

а затем найдем y_{n+1} из уравнения $(y_{n+1}-y_n)/\tau=k_2$.

Такая форма реализации метода (10) называется *методом Рунге — Кутта.* Поскольку требуется вычислить две промежуточные функции, k_1 и k_2, данный метод относится к *двухэтапным методам.* В следующем параграфе будут рассмотрены более общие m-этапные методы Рунге — Кутта, позволяющие получить бо́льшую точность.

§ 2. Методы Рунге — Кутта

1. Общая формулировка методов. Семейство методов второго порядка. По-прежнему рассматриваем задачу Коши для одного уравнения

$$\frac{du}{dt}=f(t, u), \quad t>0, \quad u(0)=u_0. \tag{1}$$

Явный m-этапный метод Рунге — Кутта состоит в следующем. Пусть решение $y_n=y(t_n)$ уже известно. Задаются числовые коэффициенты a_i, b_{ij}, $i=2, 3, \ldots, m$, $j=1, 2, \ldots, m-1$, σ_i, $i=1, 2, \ldots, m$, и последовательно вычисляются функции

$$k_1=f(t_n, y_n), \; k_2=f(t_n+a_2\tau, \; y_n+b_{21}\tau k_1),$$
$$k_3=f(t_n+a_3\tau, \; y_n+b_{31}\tau k_1+b_{32}\tau k_2), \; \ldots,$$
$$k_m=f(t_n+a_m\tau, \; y_n+b_{m1}\tau k_1+b_{m2}\tau k_2+\ldots+b_{m,m-1}\tau k_{m-1}).$$

Затем из формулы

$$\frac{y_{n+1}-y_n}{\tau}=\sum_{i=1}^{m}\sigma_i k_i \tag{2}$$

находится новое значение $y_{n+1}=y(t_{n+1})$.

Коэффициенты a_i, b_{ij}, σ_i выбираются из соображений точности. Например, для того чтобы уравнение (2) аппроксимировало исходное уравнение (1), необходимо потребовать $\sum\limits_{i=1}^{m}\sigma_i=1$. Отметим, что методы Рунге — Кутта при $m>5$ не используются.

Остановимся более подробно на отдельных методах. При $m=1$ получаем схему Эйлера, рассмотренную в примере 1 из предыдущего параграфа. При $m=2$ получаем семейство методов

$$k_1=f(t_n, y_n), \quad k_2=f(t_n+a_2\tau, \; y_n+b_{21}\tau k_1),$$
$$y_{n+1}=y_n+\tau(\sigma_1 k_1+\sigma_2 k_2). \tag{3}$$

Исследуем погрешность аппроксимации методов (3) в зависимости от выбора параметров. Исключая из последнего уравнения

функции k_1 и k_2, получаем

$$\frac{y_{n+1}-y_n}{\tau}=\sigma_1 f(t_n, y_n)+\sigma_2 f(t_n+a_2\tau, y_n+b_{21}\tau f(t_n, y_n)). \qquad (4)$$

По определению погрешностью аппроксимации или невязкой метода (3) называется выражение

$$\psi_n^{(1)}=-\frac{u_{n+1}-u_n}{\tau}+\sigma_1 f(t_n, u_n)+\sigma_2 f(t_n+a_2\tau, u_n+b_{21}f(t_n, u_n)), \qquad (5)$$

полученное заменой в (4) приближенного решения y_n точным решением $u_n=u(t_n)$.

Найдем порядок погрешности аппроксимации в предположении достаточной гладкости решения $u(t)$ и функции $f(t, u)$. Для этого разложим все величины, входящие в выражение (5), по формуле Тейлора в точке t_n. Имеем

$$\frac{u_{n+1}-u_n}{\tau}=u'(t_n)+\frac{\tau}{2}u''(t_n)+O(\tau^2),$$

$$f(t_n+a_2\tau, u_n+b_{21}\tau f_n)=f_n+a_2\tau\frac{\partial f_n}{\partial t}+b_{21}\tau f_n\frac{\partial f_n}{\partial u}+O(\tau^2),$$

где $f_n=f(t_n, u_n)$, $\frac{\partial f_n}{\partial u}=\frac{\partial f}{\partial u}(t_n, u_n)$. Далее, согласно уравнению (1), получим

$$u''=\frac{\partial f}{\partial t}+\frac{\partial f}{\partial u}u'=\frac{\partial f}{\partial t}+\frac{\partial f}{\partial u}f.$$

Поэтому

$$\psi_n^{(1)}=-u_n'+(\sigma_1+\sigma_2)f_n+$$
$$+\tau\left[(\sigma_2 b_{21}-0{,}5)f_n\frac{\partial f_n}{\partial u}+(\sigma_2 a_2-0{,}5)\frac{\partial f_n}{\partial t}\right]+O(\tau^2). \qquad (6)$$

Отсюда видно, что методы (3) имеют первый порядок аппроксимации, если $\sigma_1+\sigma_2=1$.

Если же дополнительно потребовать $\sigma_2 a_2=\sigma_2 b_{21}=0{,}5$, то получим методы второго порядка аппроксимации. Таким образом, имеется однопараметрическое семейство двухэтапных методов Рунге — Кутта второго порядка аппроксимации. Это семейство методов можно записать в виде

$$\frac{y_{n+1}-y_n}{\tau}=(1-\sigma)f(t_n, y_n)+\sigma f(t_n+a\tau, y_n+a\tau f(t_n, y_n)), \qquad (7)$$

где $\sigma a=0{,}5$.

В частности, при $\sigma=1$, $a=0{,}5$ получаем метод, рассмотренный в примере 3 предыдущего параграфа. При $\sigma=0{,}5$, $a=1$ получаем другой метод второго порядка:

$$k_1=f(t_n, y_n),\quad k_2=f(t_n+\tau, y_n+\tau k_1),$$
$$y_{n+1}=y_n+0{,}5\tau(k_1+k_2).$$

Двухэтапных методов третьего порядка аппроксимации не существует. Чтобы убедиться в этом, достаточно рассмотреть уравнение $u'=u$. Для него двухэтапный метод Рунге – Кутта (7) принимает вид

$$\frac{y_{n+1}-y_n}{\tau}=(1+\tau\sigma a)\,y_n$$

и погрешность аппроксимации равна

$$\psi_n^{(1)}=-\frac{u_{n+1}-u_n}{\tau}+(1+\tau\sigma a)\,u_n.$$

Разлагая $\psi_n^{(1)}$ по формуле Тейлора и учитывая, что $u'''=u''=u'=u$, получим

$$\psi_n^{(1)}=\tau(\sigma a-0{,}5)\,u-\frac{\tau^2}{6}u(t_n+\theta\tau),\qquad 0\leqslant\theta\leqslant 1.$$

Отсюда видно, что наивысший достижимый порядок аппроксимации равен двум.

Приведем примеры методов Рунге – Кутта более высокого порядка аппроксимации.

Метод третьего порядка:

$$k_1=f(t_n,\,y_n),\quad k_2=f\left(t_n+\frac{\tau}{2},\,y_n+\frac{\tau}{2}k_1\right),$$

$$k_3=f(t_n+\tau,\,y_n-\tau k_1+2\tau k_2),\quad \frac{y_{n+1}-y_n}{\tau}=\frac{1}{6}(k_1+4k_2+k_3).$$

Метод третьего порядка:

$$k_1=f(t_n,\,y_n),\quad k_2=f\left(t_n+\frac{\tau}{3},\,y_n+\frac{\tau k_1}{3}\right),$$

$$k_3=f\left(t_n+\frac{2\tau}{3},\,y_n+\frac{2\tau k_2}{3}\right),\quad \frac{y_{n+1}-y_n}{\tau}=\frac{1}{4}(k_1+3k_3).$$

Метод четвертого порядка:

$$k_1=f(t_n,\,y_n),\quad k_2=f\left(t_n+\frac{\tau}{2},\,y_n+\frac{\tau k_1}{2}\right),$$

$$k_3=f\left(t_n+\frac{\tau}{2},\,y_n+\frac{\tau k_2}{2}\right),\quad k_4=f(t_n+\tau,\,y_n+\tau k_3),$$

$$\frac{y_{n+1}-y_n}{\tau}=\frac{1}{6}(k_1+2k_2+2k_3+k_4).$$

Метод четвертого порядка:

$$k_1=f(t_n,\,y_n),\quad k_2=f\left(t_n+\frac{\tau}{4},\,y_n+\frac{\tau k_1}{4}\right),$$

$$k_3=f\left(t_n+\frac{\tau}{2},\,y_n+\frac{\tau k_2}{2}\right),$$

$$k_4=f(t_n+\tau,\,y_n+\tau k_1-2\tau k_2+2\tau k_3),$$

$$\frac{y_{n+1}-y_n}{\tau}=\frac{1}{6}(k_1+4k_3+k_4)$$

Приведенные здесь методы являются частными случаями методов Рунге — Кутта третьего и четвертого порядков, рассмотренных подробнее в пп. 3 и 4.

2. Доказательство сходимости. Докажем, что методы Рунге — Кутта сходятся и порядок их точности совпадает с порядком аппроксимации.

Выпишем уравнение, которому удовлетворяет погрешность $z_n = y_n - u(t_n)$. Основное уравнение метода Рунге — Кутта имеет вид

$$\frac{y_{n+1} - y_n}{\tau} = \sum_{i=1}^{m} \sigma_i k_i(y), \tag{8}$$

где

$$k_i(y) = f\left(t_n + a_i\tau,\ y_n + \sum_{j=1}^{i-1} \tau b_{ij} k_j(y)\right), \quad i = 2, 3, \ldots, m, \tag{9}$$

$$k_1(y) = f(t_n, y_n).$$

Подставим в левую часть уравнения (8) вместо y_l выражения $u_l + z_l$ при $l = n$, $n+1$, а в правой части этого уравнения добавим и вычтем сумму

$$\sum_{i=1}^{m} \sigma_i k_i(u),$$

где

$$k_i(u) = f\left(t_n + a_i\tau,\ u_n + \sum_{j=1}^{i-1} \tau b_{ij} k_j(u)\right), \quad i = 2, 3, \ldots, m,$$

$$k_1(u) = f(t_n, u_n). \tag{10}$$

Тогда уравнение (8) примет вид

$$\frac{z_{n+1} - z_n}{\tau} = \psi_n^{(1)} + \psi_n^{(2)}, \tag{11}$$

где

$$\psi_n^{(1)} = -\frac{u_{n+1} - u_n}{\tau} + \sum_{i=1}^{m} \sigma_i k_i(u) \tag{12}$$

есть по определению погрешность аппроксимации метода (8), (9) на решении исходной задачи (1) (невязка) и

$$\psi_n^{(2)} = \sum_{i=1}^{m} \sigma_i (k_i(y) - k_i(u)). \tag{13}$$

Будем рассматривать (11) как уравнение для погрешности метода. Оно выполняется для $n = 0, 1, \ldots$ Поскольку начальные значения y_0 задаются точно, $y_0 = u(0)$, величина z_0 равна нулю. Будем считать, что задача (1) решается на ограниченном отрезке времени $0 < t \leqslant T$, и, следовательно, при любых n и τ выполняется неравенство $t_n = n\tau \leqslant T$.

Предположим, что в рассматриваемой области изменения переменных (t, u) функция $f(t, u)$ удовлетворяет условию Липшица по u с константой L, не зависящей от t. При этих предположениях оценим сначала функцию $\psi_n^{(2)}$, а затем и решение z_{n+1} уравнения (11)

Из выражений (9), (10), используя условие Липшица, получим

$$|k_i(y) - k_i(u)| \leqslant L\left(|y_n - u_n| + \sum_{j=1}^{i-1} \tau |b_{ij}|\,|k_j(y) - k_j(u)|\right),$$

$$i=2, 3, \ldots, m, \quad |k_1(y) - k_1(u)| \leqslant L|y_n - u_n|.$$

Обозначим

$$r_i = |k_i(y) - k_i(u)|, \quad i = 1, 2, \ldots, m,$$
$$b = \max_{\substack{2\leqslant i\leqslant m \\ 1\leqslant j\leqslant i-1}} |b_{ij}|, \quad g = L|y_n - u_n|. \tag{14}$$

Тогда согласно предыдущему неравенству будем иметь

$$r_i \leqslant Lb\sum_{j=1}^{i-1} \tau r_j + g, \quad i = 2, 3, \ldots, m, \quad r_1 \leqslant g,$$

или, что то же самое,

$$r_{i+1} \leqslant Lb\sum_{j=0}^{i} \tau r_j + g, \quad i = 0, 1, \ldots, m-1, \quad r_0 = 0. \tag{15}$$

Лемма 1. *Из неравенств* (15) *при* $Lb\tau > 0$ *следуют оценки*

$$r_i \leqslant \rho^{i-1} g, \quad i = 1, 2, \ldots, m, \tag{16}$$

где $\rho = 1 + Lb\tau$.

Доказательство. Оценка (16) при $i=1$ совпадает с оценкой (15) для $i=0$. Пусть неравенство (16) выполнено для $i = 1, 2, \ldots, k$. Покажем, что оно выполнено и для $i=k+1$. Из (15) при $i=k$ получим

$$r_{k+1} \leqslant Lb\sum_{j=1}^{k} \tau r_j + g.$$

Согласно предположению индукции имеем

$$r_j \leqslant \rho^{j-1} g, \quad j = 1, 2, \ldots, k,$$

следовательно,

$$r_{k+1} \leqslant \left(Lb\tau \sum_{j=1}^{k} \rho^{j-1} + 1\right) g = \left(Lb\tau \frac{\rho^k - 1}{\rho - 1} + 1\right) g = \rho^k g,$$

что и требовалось.

Оценим теперь функцию $\psi_n^{(2)}$, определенную согласно (13). Из (14), (16) следует неравенство

$$|\psi_n^{(2)}| \leqslant \sum_{i=1}^{m} |\sigma_i|\,|r_i| \leqslant \sigma g \sum_{i=1}^{m} \rho^{i-1} \leqslant \sigma g m \rho^{m-1},$$

где $\sigma = \max\limits_{1\leqslant i\leqslant m} |\sigma_i|$, $\rho = 1+Lb\tau$, $g = L|z_n|$.

Итак, окончательно имеем следующую оценку для $\psi_n^{(2)}$:

$$|\psi_n^{(2)}| \leqslant \sigma L m (1+Lb\tau)^{m-1} |z_n|. \tag{17}$$

Таким образом, при возрастании погрешности $|z_n|$ величина $|\psi_n^{(2)}|$ растет не быстрее первой степени погрешности.

Теперь уже несложно оценить погрешность $z_n = y_n - u(t_n)$. Из уравнения (11) имеем

$$z_{n+1} = z_n + \tau\psi_n^{(2)} + \tau\psi_n^{(1)},$$

откуда, учитывая (17), получаем неравенство

$$|z_{n+1}| \leqslant (1+\alpha\tau)|z_n| + \tau|\psi_n^{1)}|, \qquad n = 0, 1, \ldots, \tag{18}$$

где

$$\alpha = \alpha(\tau) = \sigma L m (1+Lb\tau)^{m-1}. \tag{19}$$

Заметим, что $\alpha(\tau) \to \sigma L m$ при $\tau \to 0$. Если $\tau \leqslant \tau_0$, то $\alpha(\tau) \leqslant \sigma L m e^{Lb(m-1)\tau_0}$, т. е. $\alpha(\tau)$ ограничена равномерно по τ. В качестве τ_0 с большим загрублением можно взять T.

Из неравенства (18) следует оценка

$$|z_{n+1}| \leqslant (1+\alpha\tau)^{n+1}|z_0| + \sum_{j=0}^{n} \tau(1+\alpha\tau)^{n-j}|\psi_j^{(1)}|, \tag{20}$$

которую легко доказать по индукции.

Загрубляя оценку (20) и учитывая, что $z_0 = 0$, получим

$$|z_{n+1}| \leqslant (n+1)\tau(1+\alpha\tau)^n \max_{0\leqslant j\leqslant n} |\psi_j^{(1)}| \leqslant t_{n+1} e^{\alpha t_n} \max_{0\leqslant j\leqslant n} |\psi_j^{(1)}|,$$

где $t_n = n\tau \leqslant T$.

Таким образом, доказана

Теорема 1. *Пусть правая часть уравнения* (1) $f(t, u)$ *удовлетворяет условию Липшица по второму аргументу с константой* L. *Пусть* $\psi_j^{(1)}$ *— невязка метода Рунге — Кутта* (2), *определенная согласно* (12). *Тогда для погрешности метода при* $n\tau \leqslant T$ *справедлива оценка*

$$|y_n - u(t_n)| \leqslant T e^{\alpha T} \max_{0\leqslant j\leqslant n-1} |\psi_j^{(1)}|, \tag{21}$$

где

$$\alpha = \sigma L m (1+Lb\tau)^{m-1},$$

$$\sigma = \max_{1\leqslant i\leqslant m} |\sigma_i|, \quad b = \max_{\substack{2\leqslant i\leqslant m \\ 1\leqslant j\leqslant i-1}} |b_{ij}|.$$

Следствие. *Если метод Рунге — Кутта аппроксимирует исходное уравнение, то он сходится при* $\tau \to 0$, *причем порядок точности совпадает с порядком аппроксимации.*

Доказательство этого утверждения сразу следует из оценки (21) и замечания о равномерной ограниченности $\alpha(\tau)$.

3. Методы третьего порядка точности. При решении обыкновенных дифференциальных уравнений часто используются методы третьего и четвертого порядка точности. Приведем вывод таких методов. Сначала рассмотрим трехэтапный метод

$$\begin{gathered} k_1 = f(t_n, y_n), \quad k_2 = f(t_n + a_2\tau, y_n + b_{21}\tau k_1), \\ k_3 = f(t_n + a_3\tau, y_n + b_{31}\tau k_1 + b_{32}\tau k_2), \\ y_{n+1} = y_n + \tau(\sigma_1 k_1 + \sigma_2 k_2 + \sigma_3 k_3). \end{gathered} \tag{22}$$

Выясним, каким условиям должны удовлетворять параметры a_i, b_{ij}, σ_i для того, чтобы данный метод имел третий порядок аппроксимации. Погрешность аппроксимации метода (22) дается выражением

$$\psi_n^{(1)} = -\frac{u_{n+1} - u_n}{\tau} + \sigma_1 k_1 + \sigma_2 k_2 + \sigma_3 k_3, \tag{23}$$

где

$$\begin{gathered} k_1 = f(t_n, u_n), \; k_2 = f(t_n + a_2\tau, \; u_n + b_{21}\tau k_1), \\ k_3 = f(t_n + a_3\tau, \; u_n + b_{31}\tau k_1 + b_{32}\tau k_2) \end{gathered} \tag{24}$$

и $u_n = u(t_n)$ — решение исходного уравнения (1).

Применяя разложение по формуле Тейлора для функции двух переменных и учитывая, что $k_1 = f(t_n, u_n)$, получим

$$k_2 = f + a_2\tau f_t + b_{21}\tau f_t f_u + \frac{\tau^2}{2}[a_2^2 f_{tt} + 2a_2 b_{21} f f_{tu} + b_{21}^2 f^2 f_{uu}] + O(\tau^3). \tag{25}$$

Здесь значения функции $f(t, u)$ и ее частных производных берутся при $t = t_n$, $u = u_n$. Точно так же

$$\begin{aligned} k_3 = f + a_3\tau f_t + (b_{31}k_1 + b_{32}k_2)\,\tau f_u + \\ + \frac{\tau^2}{2}[a_3^2 f_{tt} + 2a_3(b_{31}k_1 + b_{32}k_2) f_{tu} + (b_{31}k_1 + b_{32}k_2)^2 f_{uu}] + O(\tau^3). \end{aligned}$$

Подставляя сюда

$$k_1 = f, \quad k_2 = f + a_2\tau f_t + b_{21}\tau f f_u + O(\tau^2),$$

получим

$$\begin{aligned} k_3 = f + \tau[a_3 f_t + (b_{31} + b_{32}) f f_u] + \frac{\tau^2}{2}[a_3^2 f_{tt} + 2a_3(b_{31} + b_{32}) f f_{tu} + \\ + (b_{31} + b_{32})^2 f^2 f_{uu} + 2b_{32}a_2 f_t f_u + 2b_{32}b_{21} f (f_u)^2] + O(\tau^3). \end{aligned} \tag{26}$$

Далее, из разложений (24)—(26) следует

$$\begin{aligned} \sigma_1 k_1 + \sigma_2 k_2 + \sigma_3 k_3 = (\sigma_1 + \sigma_2 + \sigma_3) f + \\ + \tau[(\sigma_2 a_2 + \sigma_3 a_3) f_t + (\sigma_2 b_{21} + \sigma_3 (b_{31} + b_{32})) f f_u] + \\ + \frac{\tau^2}{2}[(\sigma_2 a_2^2 + \sigma_3 a_3^2) f_{tt} + 2(\sigma_2 a_2 b_{21} + \sigma_3 a_3 (b_{31} + b_{32})) f f_{tu} + \\ + ((\sigma_2 b_{21}^2 + \sigma_3 (b_{31} + b_{32})^2) f^2 f_{uu} + 2\sigma_3 b_{32} a_2 f_t f_u + 2\sigma_3 b_{32} b_{21} f (f_u)^2] + O(\tau^3). \end{aligned} \tag{27}$$

Получим теперь разложение по степеням τ разностного отношения $(u_{n+1}-u_n)/\tau$, входящего в выражение для погрешности аппроксимации (23). При этом учтем, что в силу уравнения (1) справедливы следующие соотношения:

$$u'=f, \quad u''=f_t+ff_u,$$
$$u'''=f_{tt}+2ff_{tu}+f^2f_{uu}+f_uf_t+f(f_u)^2. \tag{28}$$

Тогда будем иметь

$$\frac{u_{n+1}-u_n}{\tau}=u_n'+\frac{\tau}{2}u_n''+\frac{\tau^2}{6}u_n'''+O(\tau^3)=$$
$$=f+\frac{\tau}{2}(f_t+ff_u)+\frac{\tau^2}{6}[f_{tt}+2ff_{tu}+f^2f_{uu}+f_uf_t+f(f_u)^2]+O(\tau^3).$$

Отсюда и из (27) получаем следующее разложение выражения для погрешности аппроксимации (23):

$$\psi_n^{(1)}=(\sigma_1+\sigma_2+\sigma_3-1)f+$$
$$+\frac{\tau}{2}\{[2(\sigma_2a_2+\sigma_3a_3)-1]f_t+[2(\sigma_2b_{21}+\sigma_3(b_{31}+b_{32}))-1]ff_u\}+$$
$$+\frac{\tau^2}{6}\{[3(\sigma_2a_2^2+\sigma_3a_3^2)-1]f_{tt}+[6(\sigma_2a_2b_{21}+\sigma_3a_3(b_{31}+b_{32}))-2]ff_{tu}+$$
$$+[3(\sigma_2b_{21}^2+\sigma_3(b_{31}+b_{32})^2)-1]f^2f_{uu}+$$
$$+(6\sigma_3b_{32}a_2-1)f_uf_t+(6\sigma_3b_{32}b_{21}-1)f(f_u)^2\}+O(\tau^3).$$

Приравнивая нулю коэффициенты при τ^j, $j=0, 1, 2$, получаем условия третьего порядка аппроксимации:

$$\sigma_1+\sigma_2+\sigma_3=1,$$
$$\sigma_2a_2+\sigma_3a_3=\sigma_2b_{21}+\sigma_3(b_{31}+b_{32})=0{,}5,$$
$$\sigma_2a_2^2+\sigma_3a_3^2=\sigma_2a_2b_{21}+\sigma_3a_3(b_{31}+b_{32})=\sigma_2b_{21}^2+\sigma_3(b_{31}+b_{32})^2=\frac{1}{3},$$
$$\sigma_3b_{32}a_2=\sigma_3b_{32}b_{21}=\frac{1}{6}.$$

После проведения эквивалентных преобразований эту систему уравнений можно записать в более простом виде:

$$\sigma_2a_2+\sigma_3a_3=\frac{1}{2}, \quad \sigma_2a_2^2+\sigma_3a_3^2=\frac{1}{3},$$
$$a_2=b_{21}, \quad a_3=b_{31}+b_{32}, \quad \sigma_3b_{32}a_2=\frac{1}{6}, \tag{29}$$
$$\sigma_1=1-\sigma_2-\sigma_3, \quad \sigma_3\neq0, \quad a_2\neq0.$$

Исключим с помощью (29) из выражений (22) коэффициенты b_{ij}. Тогда получим метод

$$k_1=f(t_n, y_n), \quad k_2=f(t_n+a_2\tau, y_n+a_2\tau k_1),$$
$$k_3=f\left(t_n+a_3\tau, y_n+a_3\tau k_1+\frac{\tau(k_2-k_1)}{6\sigma_3a_2}\right), \tag{30}$$
$$\frac{y_{n+1}-y_n}{\tau}=\sigma_1k_1+\sigma_2k_2+\sigma_3k_3, \quad \sigma_1=1-\sigma_2-\sigma_3,$$

который имеет третий порядок аппроксимации при условиях

$$\sigma_2 a_2 + \sigma_3 a_3 = \frac{1}{2}, \quad \sigma_2 a_2^2 + \sigma_3 a_3^2 = \frac{1}{3}, \quad \sigma_3 \neq 0, \quad a_2 \neq 0. \tag{31}$$

Таким образом, в общем случае существует двухпараметрическое семейство трехэтапных методов Рунге — Кутта, имеющих третий порядок аппроксимации. Задавая a_2 и a_3 в качестве свободных параметров, получим из (31)

$$\sigma_2 = \frac{\frac{1}{2} a_3 - \frac{1}{3}}{a_2 (a_3 - a_2)}, \qquad \sigma_3 = \frac{\frac{1}{3} - \frac{1}{2} a_2}{a_3 (a_3 - a_2)}. \tag{32}$$

Кроме того, система (31) имеет два однопараметрических семейства решений, определяемых условиями

$$a_2 = a_3 = \frac{2}{3}, \quad \sigma_2 + \sigma_3 = \frac{3}{4}, \quad \sigma_3 \neq 0, \tag{33}$$

$$a_2 = \frac{2}{3}, \quad a_3 = 0, \quad \sigma_2 = \frac{3}{4}, \quad \sigma_3 \neq 0 \quad \text{любое}. \tag{34}$$

Например, полагая $a_2 = \frac{1}{2}$, $a_3 = 1$, получим из (30), (32) следующий метод третьего порядка аппроксимации:

$$\begin{gathered} k_1 = f(t_n, y_n), \quad k_2 = f\left(t_n + \frac{\tau}{2}, \; y_n + \frac{\tau}{2} k_1\right), \\ k_3 = f(t_n + \tau, \; y_n - \tau k_1 + 2\tau k_2), \\ \frac{y_{n+1} - y_n}{\tau} = \frac{1}{6}(k_1 + 4k_2 + k_3). \end{gathered} \tag{35}$$

4. **Методы четвертого порядка точности.** Рассмотрим теперь четырехэтапный метод

$$\begin{gathered} k_1 = f(t_n, y_n), \quad k_2 = f(t_n + a_2\tau, \; y_n + b_{21}\tau k_1), \\ k_3 = f(t_n + a_3\tau, \; y_n + b_{31}\tau k_1 + b_{32}\tau k_2), \\ k_4 = f(t_n + a_4\tau, \; y_n + b_{41}\tau k_1 + b_{42}\tau k_2 + b_{43}\tau k_3), \\ y_{n+1} = y_n + \tau(\sigma_1 k_1 + \sigma_2 k_2 + \sigma_3 k_3 + \sigma_4 k_4). \end{gathered} \tag{36}$$

Погрешность аппроксимации метода (36) равна по определению

$$\psi_n^{(1)} = -\frac{u_{n+1} - u_n}{\tau} + \sigma_1 k_1 + \sigma_2 k_2 + \sigma_3 k_3 + \sigma_4 k_4, \tag{37}$$

где функции k_i, $i = 1, 2, 3, 4$, получаются из (36) путем замены y_n на точное решение $u_n = u(t_n)$.

Чтобы построить схемы четвертого порядка аппроксимации, необходимо разложить функции, входящие в (37), по формуле Тейлора до величин третьего порядка по τ включительно и приравнять нулю коэффициенты при степенях τ^n, $n = 0, 1, 2, 3$. Необходимо при этом учесть соотношения (28) и аналогичное выражение для u^{IV}. Опуская выкладки, приведем систему уравнений, которой должны удовлетворять коэффициенты метода (36) для того, чтобы данный метод

имел четвертый порядок аппроксимации:

$$\sigma_1+\sigma_2+\sigma_3+\sigma_4=1,$$

$$b_{21}=a_2,\quad b_{31}+b_{32}=a_3,\ b_{41}+b_{42}+b_{43}=a_4; \tag{38}$$

$$\sigma_2a_2+\sigma_3a_3+\sigma_4a_4=\frac{1}{2},$$

$$\sigma_2a_2^2+\sigma_3a_3^2+\sigma_4a_4^2=\frac{1}{3}, \tag{39}$$

$$\sigma_2a_2^3+\sigma_3a_3^3+\sigma_4a_4^3=\frac{1}{4};$$

$$(\sigma_3b_{32}+\sigma_4b_{42})\,a_2+\sigma_4b_{43}a_3=\frac{1}{6},$$

$$(\sigma_3b_{32}+\sigma_4b_{42})\,a_2^2+\sigma_4b_{43}a_3^2=\frac{1}{12}, \tag{40}$$

$$\sigma_3b_{32}a_2a_3+\sigma_4b_{42}a_2a_4+\sigma_4b_{43}a_3a_4=\frac{1}{8};$$

$$\sigma_4b_{43}b_{32}a_2=\frac{1}{24}. \tag{41}$$

Система (38)—(41) состоит из одиннадцати уравнений и содержит тринадцать неизвестных. Выберем в качестве независимых параметров неизвестные a_2 и a_3 и выразим остальные величины через эти неизвестные.

Для этого сначала разрешим группу уравнений (39) относительно переменных σ_2, σ_3, σ_4. Определитель δ этой системы

$$\delta=a_2a_3a_4(a_3-a_2)(a_4-a_2)(a_4-a_3). \tag{42}$$

Это означает, что мы не рассматриваем значения параметров a_2, a_3, a_4, удовлетворяющие хотя бы одному из условий

$$a_2=0,\quad a_3=0,\quad a_4=0,\quad a_2=a_3,\quad a_2=a_4,\quad a_3=a_4.$$

Заметим сразу же, что $a_2\neq0$, $\sigma_4\neq0$ согласно (41).

Предположим, что $\delta\neq0$. Случай $\delta=0$ будет рассмотрен позже.

При $\delta\neq0$ система (39) имеет следующее решение:

$$\sigma_2=\frac{6a_3a_4-4a_3-4a_4+3}{12a_2(a_3-a_2)(a_4-a_2)}, \tag{43}$$

$$\sigma_3=-\frac{6a_2a_4-4a_2-4a_4+3}{12a_3(a_3-a_2)(a_4-a_3)}, \tag{44}$$

$$\sigma_4=\frac{6a_2a_3-4a_2-4a_3+3}{12a_4(a_4-a_2)(a_4-a_3)}. \tag{45}$$

Точно так же, решая систему (40), получим

$$\sigma_3b_{32}=\frac{4a_4-3}{24a_2(a_4-a_3)}, \tag{46}$$

$$\sigma_4b_{42}=-\frac{2(1-2a_2)(a_4-a_3)-(3-4a_3)(a_3-a_2)}{24(a_3-a_2)(a_4-a_3)a_2}, \tag{47}$$

$$\sigma_4b_{43}=\frac{1-2a_2}{12(a_3-a_2)a_3}. \tag{48}$$

Учтем теперь соотношение (41). Прежде всего заметим, что $\sigma_3 \neq 0$. Действительно, при $\sigma_3 = 0$ из (44) и (46) получаем

$$a_4 = \frac{3}{4}, \quad a_2 = 0,$$

но в силу (41) имеем $a_2 \neq 0$. Таким образом, уравнение (41) эквивалентно уравнению

$$(\sigma_3 b_{32})(\sigma_4 b_{43}) a_2 = \frac{\sigma_3}{24}.$$

Подставляя сюда выражения для $\sigma_3 b_{32}$, $\sigma_4 b_{43}$, σ_3 из (44), (46), (48) и приводя подобные члены, получим уравнение $a_2(1-a_4)=0$, из которого следует, что $a_4 = 1$.

Таким образом, при $\delta \neq 0$ система (38)—(41) имеет следующее двухпараметрическое семейство решений:

$$a_4 = 1, \quad \sigma_2 \triangleq \frac{2a_3 - 1}{12a_2(a_3 - a_2)(1 - a_2)},$$

$$\sigma_3 = -\frac{2a_2 - 1}{12a_3(a_3 - a_2)(1 - a_3)},$$

$$\sigma_4 = \frac{6a_2a_3 - 4a_2 - 4a_3 + 3}{12(1 - a_2)(1 - a_3)},$$

$$b_{42} = -\frac{4a_3^2 - a_2 - 5a_3 + 2}{24\sigma_4 a_2(a_3 - a_2)(1 - a_3)},$$

$$b_{43} = \frac{1 - 2a_2}{12\sigma_4 a_3(a_3 - a_2)},$$

$$b_{41} = 1 - b_{42} - b_{43},$$

$$b_{31} = a_2 - b_{32},$$

$$b_{21} = a_2,$$

$$\sigma_1 = 1 - \sigma_2 - \sigma_3 - \sigma_4.$$

Здесь, как уже отмечалось, $\sigma_3 \neq 0$, $\sigma_4 \neq 0$, т. е. $a_2 \neq 0{,}5$, $6a_2a_3 - 4a_2 - 4a_3 + 3 \neq 0$. Приведенное выше решение справедливо при $\delta \neq 0$, т. е. когда параметры a_2, a_3, a_4 удовлетворяют условиям

$$a_i \neq 0, \quad i = 2, 3, 4, \quad a_2 \neq a_3, \quad a_2 \neq a_4, \quad a_3 \neq a_4.$$

Рассмотрим систему (38)—(41) при тех значениях параметров a_2, a_3, a_4, когда $\delta = 0$. При $a_2 = 0$ система не имеет решения вследствие (41). При $a_3 = 0$ система (40) принимает вид

$$(\sigma_3 b_{32} + \sigma_4 b_{42}) a_2 = \frac{1}{6}, \quad (\sigma_3 b_{32} + \sigma_4 b_{42}) a_2^2 = \frac{1}{12}, \tag{49}$$

$$\sigma_4 a_2 a_4 b_{42} = \frac{1}{8},$$

откуда следует, что $a_2 = \frac{1}{2}$ и

$$\sigma_3 b_{32} + \sigma_4 b_{42} = \frac{1}{3}. \tag{50}$$

Далее, система (39) при $a_2 = \frac{1}{2}$, $a_3 = 0$ принимает вид

$$\sigma_2 + 2\sigma_4 a_4 = 1, \quad \sigma_2 + 4\sigma_4 a_4^2 = \frac{4}{3}, \quad \sigma_2 + 8\sigma_4 a_4^3 = 2$$

и имеет единственное решение

$$a_4 = 1, \quad \sigma_2 = \frac{2}{3}, \quad \sigma_4 = \frac{1}{6}.$$

Подставляя эти значения a_4, σ_2, σ_4 и $a_2 = \frac{1}{2}$ в уравнения (49), (50), получаем

$$b_{42} = \frac{3}{2}, \quad b_{32} = \frac{1}{12\sigma_3}.$$

Кроме того, из (41) имеем

$$b_{43} = \frac{1}{24\sigma_4 b_{32} a_2} = 6\sigma_3.$$

Таким образом, система (38) имеет следующее семейство решений, зависящее от параметра $\sigma_3 \neq 0$:

$$a_2 = \frac{1}{2}, \quad a_3 = 0, \quad a_4 = 1,$$

$$\sigma_2 = \frac{2}{3}, \quad \sigma_4 = \frac{1}{6},$$

$$b_{32} = \frac{1}{12\sigma_3}, \quad b_{42} = \frac{3}{2}, \quad b_{43} = 6\sigma_3,$$

$$b_{41} = -\frac{1}{2} - 6\sigma_3, \quad b_{31} = -\frac{1}{12\sigma_3}, \quad b_{21} = \frac{1}{2}, \quad \sigma_1 = \frac{1}{6} - \sigma_3.$$

Точно так же при условии $a_2 = a_3$ система (38)—(41) имеет решение

$$a_2 = a_3 = \frac{1}{2}, \quad a_4 = 1,$$

$$\sigma_2 = \frac{2}{3} - \sigma_3, \quad \sigma_4 = \frac{1}{6},$$

$$b_{32} = \frac{1}{6\sigma_3}, \quad b_{42} = 1 - 3\sigma_3, \quad b_{43} = 3\sigma_3,$$

$$b_{41} = 0, \quad b_{31} = \frac{1}{2} - \frac{1}{6\sigma_3}, \quad b_{21} = \frac{1}{2}, \quad \sigma_1 = \frac{1}{6},$$

зависящее от параметра $\sigma_3 \neq 0$.

При $a_2 = a_4$ имеем решение

$$a_2 = a_4 = 1, \quad a_3 = \frac{1}{2}, \quad \sigma_2 = \frac{1}{6} - \sigma_4, \quad \sigma_3 = \frac{2}{3},$$

$$b_{32} = \frac{1}{8}, \quad b_{42} = -\frac{1}{6\sigma_4}, \quad b_{43} = \frac{1}{3\sigma_4},$$

$$b_{41} = 1 - \frac{1}{6\sigma_4}, \quad b_{31} = \frac{3}{8}, \quad b_{21} = 1,$$

$$\sigma_1 = \frac{1}{6},$$

зависящее от параметра $\sigma_4 \neq 0$.

Определитель δ обращается в нуль еще в двух случаях: при $a_4=0$ и $a_3=a_4$. Оказывается, что в этих случаях система (38)—(41) не имеет решения. Пусть, например, $a_4=0$. Тогда из первых двух уравнений системы (39) получим

$$\sigma_2=\frac{3a_2-2}{6a_2(a_3-a_2)},\quad \sigma_3=\frac{2-3a_2}{6a_3(a_3-a_2)}.$$

При этом последнее уравнение системы (39) приводит к условию

$$6a_2a_3-4a_2-4a_3+3=0. \tag{51}$$

Аналогично находим, что система (40), (41) разрешима относительно b_{32}, b_{42}, b_{43} только при условии

$$6a_2a_3-6a_2-4a_3+3=0. \tag{52}$$

Из (51), (52) находим $a_2=0$, что невозможно в силу (41). Точно так же доказывается, что не существует решений с $a_3=a_4$.

§ 3. Многошаговые разностные методы

1. Формулировка методов. Для решения задачи Коши

$$\frac{du}{dt}=f(t,u),\quad t>0,\qquad u(0)=u_0 \tag{1}$$

введем сетку

$$\omega_\tau=\{t_n=n\tau,\quad n=0,1,\ldots\}$$

с постоянным шагом $\tau>0$. Обозначим через $y_n=y(t_n)$, $f_n=f(t_n, y_n)$ функции, определенные на сетке ω_τ. *Линейным m-шаговым разностным методом* называется система разностных уравнений

$$\frac{a_0y_n+a_1y_{n-1}+\ldots+a_my_{n-m}}{\tau}=b_0f_n+b_1f_{n-1}+\ldots+b_mf_{n-m}, \tag{2}$$
$$n=m,\ m+1,\ldots,$$

где a_k, b_k — числовые коэффициенты, не зависящие от n, $k=0, 1, \ldots, m$, причем $a_0\neq 0$.

Уравнение (2) следует рассматривать как рекуррентное соотношение, выражающее новое значение $y_n=y(t_n)$ через найденные ранее значения $y_{n-1}, y_{n-2}, \ldots, y_{n-m}$.

Расчет начинается с $n=m$, т. е. с уравнения

$$\frac{a_0y_m+a_1y_{m-1}+\ldots+a_my_0}{\tau}=b_0f_m+b_1f_{m-1}+\ldots+b_mf_0.$$

Отсюда видно, что для начала расчета необходимо задать m начальных значений $y_0, y_1, \ldots, y_{m-1}$. Значение y_0 определяется исходной задачей (1), а именно полагают $y_0=u_0$. Величины $y_1, y_2, \ldots, y_{m-1}$ можно вычислить, например, с помощью метода Рунге — Кутта. В дальнейшем будем предполагать, что начальные значения $y_0, y_1, \ldots, y_{m-1}$ заданы.

Из уравнения (2) видно, что в отличие от методов Рунге — Кутта многошаговые разностные методы допускают вычисление правых частей только в точках основной сетки ω_τ.

Метод (2) называется *явным*, если $b_0=0$, и, следовательно, искомое значение y_n выражается явным образом через предыдущие значения $y_{n-1}, y_{n-2}, \ldots, y_{n-m}$. В противном случае (т. е. когда $b_0\neq 0$) метод называется *неявным*. Тогда для нахождения y_n приходится решать нелинейное уравнение

$$\frac{a_0}{\tau} y_n - b_0 f(t_n, y_n) = F[y_{n-1}, y_{n-2}, \ldots, y_{n-m}],$$

где

$$F[y_{n-1}, y_{n-2}, \ldots, y_{n-m}] = \sum_{k=1}^{m} \left(b_k f_{n-k} - \frac{a_k}{\tau} y_{n-k}\right).$$

Обычно это уравнение решают методом Ньютона, выбирая начальное приближение $y_n^{(0)}$ равным y_{n-1}.

Заметим, что коэффициенты уравнения (2) определены с точностью до множителя. Чтобы устранить этот произвол, будем считать, что выполнено условие

$$\sum_{k=0}^{m} b_k = 1, \tag{3}$$

означающее, что правая часть разностного уравнения (2) аппроксимирует правую часть дифференциального уравнения (1).

В практике вычислений наибольшее распространение получили методы Адамса, которые представляют собой частный случай многошаговых методов (2), когда производная $u'(t)$ аппроксимируется только по двум точкам, t_n и t_{n-1}, т. е.

$$a_0=-a_1=1, \quad a_k=0, \quad k=2, 3, \ldots, m.$$

Таким образом, *методы Адамса* имеют вид

$$\frac{y_n - y_{n-1}}{\tau} = \sum_{k=0}^{m} b_k f_{n-k}. \tag{4}$$

В случае $b_0=0$ методы Адамса называются *явными*, в случае $b_0\neq 0$ — *неявными*.

При изучении разностных методов (2) мы рассмотрим прежде всего, как влияет выбор коэффициентов a_k, b_k на погрешность аппроксимации, а затем исследуем тесно связанные между собой вопросы устойчивости и сходимости.

2. Погрешность аппроксимации многошаговых методов. *Погрешностью аппроксимации на решении* или *невязкой разностного метода* (2) называется функция

$$\psi_n = -\sum_{k=0}^{m} \frac{a_k}{\tau} u_{n-k} + \sum_{k=0}^{m} b_k f(t_{n-k}, u_{n-k}), \tag{5}$$

получающаяся в результате подстановки точного решения $u(t)$ дифференциальной задачи (1) в разностное уравнение (2).

Выясним вопрос о порядке погрешности аппроксимации при $\tau \to 0$ в зависимости от выбора коэффициентов a_k, b_k, $k=0, 1, \ldots, m$. Будем предполагать при этом, что все рассматриваемые функции обладают необходимой гладкостью.

Разлагая функции $u_{n-k}=u(t_n-k\tau)$ в точке $t=t_n$ по формуле Тейлора, получим

$$u_{n-k} = \sum_{l=0}^{p} \frac{(-k\tau)^l u^{(l)}(t_n)}{l!} + O(\tau^{p+1}),$$

$$f(t_{n-k}, u_{n-k}) = u'(t_n - k\tau) =$$

$$= \sum_{l=0}^{p-1} \frac{(-k\tau)^l u^{(l+1)}(t_n)}{l!} + O(\tau^p), \quad k = 1, 2, \ldots, m.$$

Подставляя эти разложения в выражение (5) для погрешности аппроксимации, будем иметь

$$\psi_n = -\sum_{k=0}^{m} \frac{a_k}{\tau} \left(\sum_{l=0}^{p} \frac{(-k\tau)^l u^{(l)}(t_n)}{l!} \right) +$$

$$+ \sum_{k=0}^{m} b_k \left(\sum_{l=0}^{p-1} \frac{(-k\tau)^l u^{(l+1)}(t_n)}{l!} \right) + O(\tau^p) =$$

$$= -\sum_{l=0}^{p} \left(\sum_{k=0}^{m} \frac{a_k}{\tau} \cdot \frac{(-k\tau)^l u^{(l)}(t_n)}{l!} \right) +$$

$$+ \sum_{l=1}^{p} \left(\sum_{k=0}^{m} b_k \frac{(-k\tau)^{l-1} u^{(l)}(t_n)}{(l-1)!} \right) + O(\tau^p).$$

После очевидных преобразований приходим к разложению

$$\psi_n = -\left(\sum_{k=0}^{m} \frac{a_k}{\tau} \right) u(t_n) +$$

$$+ \sum_{l=1}^{p} \left(\sum_{k=0}^{m} (-k\tau)^{l-1} \left(a_k \frac{k}{l} + b_k \right) \right) \frac{u^{(l)}(t_n)}{(l-1)!} + O(\tau^p). \quad (6)$$

Отсюда видно, что погрешность аппроксимации имеет порядок p, если выполнены условия

$$\sum_{k=0}^{m} a_k = 0, \quad (7)$$

$$\sum_{k=0}^{m} k^{l-1} (k a_k + l b_k) = 0, \quad l = 1, 2, \ldots, p. \quad (8)$$

Вместе с условием нормировки (3) уравнения (7), (8) образуют систему из $p+2$ линейных алгебраических уравнений относи-

тельно $2(m+1)$ неизвестных

$$a_0, a_1, \ldots, a_m, b_0, b_1, \ldots, b_m.$$

Можно несколько упростить эту систему. А именно, рассмотрим уравнение (8) при $l=1$,

$$\sum_{k=0}^{m} ka_k + \sum_{k=0}^{m} b_k = 0$$

и учтем условие нормировки (3). Тогда получим уравнение

$$\sum_{k=0}^{m} ka_k = -1.$$

Окончательно получаем систему уравнений

$$\sum_{k=1}^{m} ka_k = -1, \tag{9}$$

$$\sum_{k=1}^{m} k^{l-1}(ka_k + lb_k) = 0, \quad l = 2, 3, \ldots, p,$$

которая содержит p уравнений и $2m$ неизвестных $a_1, a_2, \ldots, a_m$, $b_1, b_2, \ldots, b_m$. Коэффициенты a_0 и b_0 вычисляются по формулам

$$a_0 = -\sum_{k=1}^{m} a_k, \quad b_0 = 1 - \sum_{k=1}^{m} b_k. \tag{10}$$

Для того чтобы система (9) не была переопределена, необходимо потребовать, чтобы $p \leqslant 2m$. Это требование означает, что *порядок аппроксимации линейных m-шаговых разностных методов не может превосходить $2m$.*

Итак, наивысший достижимый порядок аппроксимации неявных m-шаговых методов равен $2m$, а явных — $2m-1$.

Заметим, что если в системе (8) отбросить последние n уравнений, $n=1, 2, \ldots, p-1$, то получим условия, обеспечивающие порядок аппроксимации $p-n$.

Для методов Адамса (4) условия p-го порядка аппроксимации (9) принимают вид

$$l\sum_{k=1}^{m} k^{l-1}b_k = 1, \quad l = 2, 3, \ldots, p, \quad b_0 = 1 - \sum_{k=1}^{m} b_k. \tag{11}$$

Отсюда видно, что *наивысший порядок аппроксимации m-шагового метода Адамса равен $m+1$, а наивысший порядок аппроксимации явного метода Адамса ($b_0=0$) равен m.*

3. Устойчивость и сходимость разностных методов. Оказывается, что методы наивысшего порядка аппроксимации практически непригодны для расчетов, так как они неустойчивы. Подробно вопросы устойчивости и сходимости разностных методов будут

рассмотрены в следующем параграфе, а сейчас ограничимся изложением самых необходимых сведений.

Рассмотрим наряду с (2) однородное разностное уравнение

$$a_0v_n+a_1v_{n-1}+\ldots+a_mv_{n-m}=0, \quad n=m,\ m+1,\ \ldots, \tag{12}$$

и будем искать решения уравнения (12), имеющие вид $v_n=q^n$, где q — число, подлежащее определению. Тогда для нахождения q получаем уравнение

$$a_0q^m+a_1q^{m-1}+\ldots+a_{m-1}q+a_m=0, \tag{13}$$

которое называется *характеристическим уравнением разностного метода* (2).

Говорят, что метод (2) удовлетворяет *условию корней*, если все корни $q_1, q_2, \ldots, q_m$ характеристического уравнения (13) лежат внутри или на границе единичного круга комплексной плоскости, причем на границе единичного круга нет кратных корней. Разностный метод (2), удовлетворяющий условию корней, называют *устойчивым методом*. Существует определенное ограничение на порядок аппроксимации устойчивого метода. Приведем без доказательства следующее утверждение.

Пусть метод (2) *удовлетворяет условию корней и имеет порядок аппроксимации* p. *Тогда* $p\leqslant m+1$ *при* m *нечетном и* $p\leqslant m+2$ *при* m *четном. Для явных* m*-шаговых устойчивых методов порядок аппроксимации не превосходит* m.

В § 4 будет доказана следующая теорема о связи между устойчивостью и сходимостью разностного метода (2) (см. теорему 2 из § 4).

Пусть метод (2) *удовлетворяет условию корней и* $|f_u(t, u)|\leqslant L$ *при* $0\leqslant t\leqslant T$. *Тогда при* $m\tau\leqslant t_n=n\tau\leqslant T$, $n\geqslant m$ *и всех достаточно малых* τ *выполнена оценка*

$$|y_n-u(t_n)|\leqslant M\Big(\max_{0\leqslant j\leqslant m-1}|y_j-u(t_j)|+\max_{0\leqslant k\leqslant n-m}|\psi_k|\Big), \tag{14}$$

где ψ_k — *погрешность аппроксимации,* $y_j-u(t_j)$, $j=0,\ 1,\ \ldots,\ m-1$ — *погрешности в задании начальных условий и* M — *константа, зависящая от* L, T *и не зависящая от* n.

Из оценки (14) следует, что если начальные погрешности $y_j-u(t_j)$, $j=0,\ 1,\ \ldots,\ m-1$, и погрешность аппроксимации ψ_k, $k=0,\ 1,\ \ldots,\ n-m$, являются величинами $O(\tau^p)$, $p>0$, то и $y_n-u(t_n)=O(\tau^p)$ при $n\geqslant m$, т. е. метод сходится и имеет p-й порядок точности.

Таким образом, исследование сходимости метода (2) сводится к анализу погрешности аппроксимации и проверке условия корней.

Заметим, что методы Адамса

$$\frac{y_n-y_{n-1}}{\tau}=\sum_{k=0}^{m}b_kf(t_{n-k}, y_{n-k})$$

всегда удовлетворяют условию корней, так как для них $a_0=-a_1=1$, т. е. $q=q_1=1$.

Простым примером метода, не удовлетворяющего условию корней, является явный двухшаговый метод

$$\frac{y_n + 4y_{n-1} - 5y_{n-2}}{6\tau} = \frac{2f_{n-1} + f_{n-2}}{3},$$

имеющий третий порядок аппроксимации.

4. Примеры многошаговых разностных методов. Наивысший порядок аппроксимации явных m-шаговых методов Адамса

$$\frac{y_n - y_{n-1}}{\tau} = b_1 f_{n-1} + b_2 f_{n-2} + \ldots + b_m f_{n-m} \tag{15}$$

равен m. Согласно (11) условия m-го порядка аппроксимации имеют вид

$$\sum_{k=1}^{m} k^{l-1} b_k = \frac{1}{l}, \quad l = 1, 2, \ldots, m. \tag{16}$$

Решая систему (16), можно найти коэффициенты метода наивысшего порядка (15), (16) при каждом конкретном m. Так, при $m=1$ получаем метод Эйлера

$$\frac{y_n - y_{n-1}}{\tau} = f_{n-1}.$$

При $m=2, 3, 4, 5$ получаем соответственно следующие методы m-го порядка аппроксимации:

$$\frac{y_n - y_{n-1}}{\tau} = \frac{3}{2} f_{n-1} - \frac{1}{2} f_{n-2}, \quad m=2,$$

$$\frac{y_n - y_{n-1}}{\tau} = \frac{1}{12}(23f_{n-1} - 16f_{n-2} + 5f_{n-3}), \quad m=3,$$

$$\frac{y_n - y_{n-1}}{\tau} = \frac{1}{24}(55f_{n-1} - 59f_{n-2} + 37f_{n-3} - 9f_{n-4}), \quad m=4,$$

$$\frac{y_n - y_{n-1}}{\tau} = \frac{1}{720}(1901f_{n-1} - 2774f_{n-2} + 2616f_{n-3} - 1274f_{n-4} + 251f_{n-5}), \quad m=5.$$

Для неявных m-шаговых методов Адамса

$$\frac{y_n - y_{n-1}}{\tau} = b_0 f_n + b_1 f_{n-1} + \ldots + b_m f_{n-m} \tag{17}$$

наивысший порядок аппроксимации равен $m+1$. Коэффициенты метода (17) наивысшего порядка находятся из системы (11) с $p=m+1$. При $m=1$ получаем метод второго порядка аппроксимации

$$\frac{y_n - y_{n-1}}{\tau} = \frac{1}{2}(f_n + f_{n-1}), \quad p=2,$$

называемый методом трапеций. При $m=2$, 3, 4 получаем соответственно следующие методы $(m+1)$-го порядка аппроксимации:

$$\frac{y_n - y_{n-1}}{\tau} = \frac{1}{12}(5f_n + 8f_{n-1} - f_{n-2}), \quad p=3,$$

$$\frac{y_n - y_{n-1}}{\tau} = \frac{1}{24}(9f_n + 19f_{n-1} - 5f_{n-2} + f_{n-3}), \quad p=4,$$

$$\frac{y_n - y_{n-1}}{\tau} = \frac{1}{720}(251f_n + 646f_{n-1} - 264f_{n-2} + 106f_{n-3} - 19f_{n-4}), \quad p=5.$$

Выписанные выше неявные методы содержат искомое значение нелинейно, поэтому для их реализации необходимо применять итерационные методы. Например, для неявного метода Адамса четвертого порядка используется итерационный метод

$$\frac{y_n^{(s+1)} - y_{n-1}}{\tau} = \frac{1}{24}(9f(t_n, y_n^{(s)}) + 19f(t_{n-1}, y_{n-1}) - 5f(t_{n-2}, y_{n-2}) + f(t_{n-3}, y_{n-3})), \quad (18)$$

где s — номер итерации, $s=0, 1, \ldots$ В качестве начального значения $y_n^{(0)}$ можно взять решение, полученное с помощью явного метода Адамса третьего порядка, т. е. метода

$$\frac{y_n^{(0)} - y_{n-1}}{\tau} = \frac{1}{12}(23f(t_{n-1}, y_{n-1}) - 16f(t_{n-2}, y_{n-2}) + 5f(t_{n-3}, y_{n-3})). \quad (19)$$

Записывая (18) в виде

$$y_n^{(s+1)} = \frac{3\tau}{8} f(t_n, y_n^{(s)}) + F,$$

получаем, что если $\left|\frac{\partial f}{\partial y}\right| \leqslant M$, то итерационный метод сходится при условии $\frac{3\tau M}{8} < 1$, которое выполнено при достаточно малом τ.

Если в (18) ограничиться только одной итерацией $s=0$, то получим метод, называемый методом предиктор — корректор (предсказывающе-исправляющий).

§ 4. Сходимость и оценка погрешности многошагового разностного метода *)

1. Уравнение для погрешности. Для задачи Коши

$$\frac{du}{dt} = f(t, u), \quad t>0, \quad u(0)=u_0 \quad (1)$$

*) При первом чтении этот параграф можно опустить.

рассмотрим m-шаговый разностный метод

$$\frac{a_0y_n+a_1y_{n-1}+\ldots+a_my_{n-m}}{\tau}=b_0f(t_n, y_n)+$$
$$+b_1f(t_{n-1}, y_{n-1})+\ldots+b_mf(t_{n-m}, y_{n-m}), \quad (2)$$

где $n=m, m+1, \ldots$, заданы начальные значения $y_0, y_1, \ldots, y_{m-1}$.

В настоящем параграфе выясняются условия, при которых сходится метод (2) и даются оценки погрешности $z_n=y_n-u(t_n)$ в любой момент времени $t_n=n\tau$, $n\geqslant m$, через начальные погрешности $z_0, z_1, \ldots, z_{m-1}$ и через погрешность аппроксимации.

Получим уравнение, которому удовлетворяет погрешность $z_n=y_n-u(t_n)$. Подставляя в левую часть уравнения (2) вместо y_j выражения $u(t_j)+z_j$, $j=n, n-1, \ldots, n-m$, получим

$$\frac{a_0z_n+a_1z_{n-1}+\ldots+a_mz_{n-m}}{\tau}=-\frac{a_0u_n+a_1u_{n-1}+\ldots+a_mu_{n-m}}{\tau}+$$
$$+b_0f(t_n, y_n)+b_1f(t_{n-1}, y_{n-1})+\ldots+b_mf(t_{n-m}, y_{n-m}).$$

Далее, добавим к правой части этого уравнения и вычтем из нее выражение

$$b_0f(t_n, u_n)+b_1f(t_{n-1}, u_{n-1})+\ldots+b_mf(t_{n-m}, u_{n-m}).$$

Тогда уравнение для погрешности примет вид

$$\frac{a_0z_n+a_1z_{n-1}+\ldots+a_mz_{n-m}}{\tau}=\psi_{n-m}+\varphi_{n-m}, \quad (3)$$
$$n=m, m+1, \ldots,$$

где через ψ_{n-m} обозначена погрешность аппроксимации

$$\psi_{n-m}=-\frac{a_0u_n+a_1u_{n-1}+\ldots+a_mu_{n-m}}{\tau}+$$
$$+b_0f(t_n, u_n)+b_1f(t_{n-1}, u_{n-1})+\ldots+b_mf(t_{n-m}, u_{n-m}) \quad (4)$$

и через φ_{n-m} — функция

$$\varphi_{n-m}=b_0(f(t_n, y_n)-f(t_n, u_n))+b_1(f(t_{n-1}, y_{n-1})-f(t_{n-1}, u_{n-1}))+\ldots$$
$$\ldots+b_m(f(t_{n-m}, y_{n-m})-f(t_{n-m}, u_{n-m})). \quad (5)$$

Погрешность аппроксимации ψ_{n-m} оценивалась в п. 2 § 3, где были найдены условия p-го порядка аппроксимации. В частности, при выполнении этих условий $\psi_{n-m}\to 0$ при $\tau\to 0$.

Функция φ_{n-m}, входящая в правую часть уравнения (3), зависит нелинейно от погрешности z_j, $j=n, n-1, \ldots, n-m$. Вид нелинейности определяется функцией $f(t, u)$. В дальнейшем будем предполагать, что $f(t, u)$ удовлетворяет условию Липшица по второму аргументу, т. е.

$$|f(t, u_1)-f(t, u_2)|\leqslant L|u_1-u_2| \quad (6)$$

для всех t, u_1, u_2 из рассматриваемой области. Тогда из (5) следует,

что для функции φ_{n-m} выполнена оценка

$$|\varphi_{n-m}| \leqslant bL(|z_n| + |z_{n-1}| + \ldots + |z_{n-m+1}| + |z_{n-m}|), \qquad (7)$$

где $b = \max(|b_0|, |b_1|, \ldots, |b_m|)$.

В дальнейшем будет получена оценка решения z_n уравнения (3) через $z_0, z_1, \ldots, z_{m-1}$ и ψ_j, $j=0, 1, \ldots, n-m$, из которой будет следовать сходимость метода (2). Предварительно нам потребуются некоторые сведения из теории разностных уравнений.

2. Однородное разностное уравнение с постоянными коэффициентами. Частные решения. Рассмотрим разностное уравнение

$$a_0 v_n + a_1 v_{n-1} + \ldots + a_m v_{n-m} = 0, \quad n = m, m+1, \ldots, \qquad (8)$$

коэффициенты которого $a_0, a_1, \ldots, a_m$ не зависят от n.

Будем искать частные решения уравнения (8), имеющие вид $v_n = q^n$, где q — число, подлежащее определению. Подставляя $v_{n-k} = q^{n-k}$, $k=0, 1, \ldots, m$, в (8) и сокращая на q^{n-m}, получим уравнение

$$a_0 q^m + a_1 q^{m-1} + \ldots + a_{m-1} q + a_m = 0, \qquad (9)$$

которое называется характеристическим уравнением, соответствующим разностному уравнению (8). Многочлен

$$F(q) = a_0 q^m + a_1 q^{m-1} + \ldots + a_{m-1} q + a_m \qquad (10)$$

называется характеристическим многочленом разностного уравнения (8).

Таким образом, разностное уравнение (8) имеет решение q^n тогда и только тогда, когда q является корнем характеристического уравнения (10). Более того, если корень q имеет кратность $r \geqslant 1$, то разностное уравнение (8) имеет частные решения $v_n = n^j q^n$, $j=0, 1, \ldots, r-1$.

Докажем последнее утверждение. Подставляя $v_{n-k} = (n-k)^j q^{n-k}$ в (8) и сокращая на q^{n-m}, получаем уравнение

$$\sum_{k=0}^{m} a_k q^{m-k} (n-k)^j = 0, \qquad (11)$$

которое при $j=0$ совпадает с характеристическим уравнением (9). Представим многочлен

$$\sum_{k=0}^{m} a_k q^{m-k} (n-k)^j \qquad (12)$$

в виде линейной комбинации характеристического многочлена (10) и его производных $F^{(i)}(q)$, $i=1, 2, \ldots, j-1$.

Для этого воспользуемся сначала разложением по формуле бинома Ньютона,

$$(n-k)^j = [(n-m) + (m-k)]^j = \sum_{l=0}^{j} C_j^l (n-m)^l (m-k)^{j-l},$$

где $C_j^l = \dfrac{j(j-1)\ \ldots\ (j-l+1)}{l!}$, $C_j^0 = 1$. Тогда получим

$$\sum_{k=0}^{m} a_k (n-k)^l q^{m-k} = \sum_{l=0}^{l} C_j^l (n-m)^l F_{j-l}(q), \tag{13}$$

где

$$F_{j-l}(q) = \sum_{k=0}^{m} (m-k)^{j-l} a_k q^{m-k}, \qquad l = 0, 1, \ldots, j, \tag{14}$$

причем $F_0(q) \equiv F(q)$.

Далее, обозначим $z = m-k$, $f_{j-l}(z) = z^{j-l}$ и запишем многочлен $F_{j-l}(q)$ в виде

$$F_{j-l}(q) = \sum_{k=0}^{m} f_{j-l}(z)\, a_k q^{m-k}.$$

Интерполируем функцию $f_{j-l}(z)$ алгебраическим многочленом степени $j-l$ по узлам $z_i = i$, $i = 0, 1, \ldots, j-l$. В данном случае погрешность интерполяции тождественно равна нулю, так как $f_{j-l}(z) = z^{j-l}$ — многочлен степени $j-l$. Поэтому согласно интерполяционной формуле Ньютона имеем

$$f_{j-l}(z) = f_{j-l}(z_0) + \sum_{i=1}^{j-l} f_{j-l}(z_0, z_1, \ldots, z_i)(z-z_0)\ \ldots\ (z-z_{i-1}),$$

где $f_{j-l}(z_0, \ldots, z_i)$ — разделенная разность i-го порядка функции z^{j-l}, построенная по узлам $z_\alpha = \alpha$, $\alpha = 0, 1, \ldots, i$. При $l < j$ имеем $f_{j-l}(z_0) = z_0^{j-l} = 0$ и

$$f_{j-l}(z) = \sum_{i=1}^{j-l} f_{j-l}(z_0, \ldots, z_i)(z-z_0)\ \ldots\ (z-z_{i-1}).$$

Подставляя сюда $z = m-k$, получим при $l < j$

$$(m-k)^{j-l} = \sum_{i=1}^{j-l} f_{j-l}(z_0, \ldots, z_i)(m-k)(m-k-1)\ \ldots\ (m-k-(i-1)).$$

Итак,

$$F_{j-l}(q) = \sum_{k=0}^{m} a_k q^{m-k} \sum_{i=1}^{j-l} d_{i,j-l}(m-k)(m-k-1)\ \ldots\ (m-k-(i-1)),$$

$$l = 0, 1, \ldots, j-1, \tag{15}$$

где обозначено $d_{i,\,j-l} = f_{j-l}(z_0, z_1, \ldots, z_i)$.

С другой стороны, для производных $F^{(l)}(q)$ многочлена (10) имеем

$$q^l F^{(l)}(q) = \sum_{k=0}^{m} (m-k)(m-k-1)\ \ldots\ (m-k-(i-1))\, a_k q^{m-k}, \tag{16}$$

$$i = 1, 2, \ldots, m.$$

Сопоставляя (15) и (16), получим

$$F_{j-l}(q) = \sum_{i=1}^{j-l} d_{i,\,j-l} q^i F^{(i)}(q), \qquad l = 0, 1, \ldots, j-1.$$

Поэтому тождество (13) можно переписать в виде

$$\sum_{k=0}^{m} a_k (n-k)^j q^{m-k} = (n-m)^j F(q) + \sum_{l=0}^{j-1} C_j^l (n-m)^l \sum_{i=1}^{j-l} d_{i,j-l} q^i F^{(i)}(q) =$$

$$= (n-m)^j F(q) + \sum_{i=1}^{j} q^i F^{(i)}(q) \left(\sum_{l=0}^{j-i} C_j^l d_{i,j-l} (n-m)^l \right).$$

Итак, получено следующее представление многочлена (12) в виде линейной комбинации характеристического многочлена $F(q)$ и его производных:

$$\sum_{k=0}^{m} a_k q^{m-k} (n-k)^j = (n-m)^j F(q) + \sum_{i=1}^{j} b_i q^i F^{(i)}(q), \qquad (17)$$

где $b_i = \sum_{l=0}^{j-i} C_j^l (n-m)^l d_{i,j-l}$, а $d_{i,j-l}$ — разделенная разность функции z^{j-l}, построенная по узлам $z_0=0$, $z_1=1$, ..., $z_i=i$.

Если q — корень кратности r характеристического уравнения (9), то $F(q)=0, \ldots, F^{(r-1)}(q)=0$ и правая часть уравнения (17) обращается в нуль при $j=1$, $j=2$, ..., $j=r-1$. Следовательно, функции q, nq^n, ..., $n^{r-1}q^n$ являются решениями разностного уравнения (8).

3. Однородное разностное уравнение с постоянными коэффициентами. Устойчивость по начальным данным. Задача Коши для уравнения (8) состоит в отыскании сеточной функции v_n, удовлетворяющей при всех $n \geqslant m$ уравнению (8) и принимающей при $n=0, 1, \ldots, m-1$ заданные начальные значения $v_0, v_1, \ldots, v_{m-1}$.

В дальнейшем будем считать, что $a_0 \neq 0$. Тогда уравнение (8) можно разрешить относительно v_n:

$$v_n = -\frac{a_m}{a_0} v_{n-m} - \frac{a_{m-1}}{a_0} v_{n-m+1} - \ldots - \frac{a_1}{a_0} v_{n-1}.$$

Отсюда следует, что при $a_0 \neq 0$ решение задачи Коши существует и единственно.

Говорят, что уравнение (8) *устойчиво по начальным данным*, если существует постоянная M_1, не зависящая от n и такая, что при любых начальных данных $v_0, v_1, \ldots, v_{m-1}$ для его решения выполняется оценка

$$|v_n| \leqslant M_1 \max_{0 \leqslant j \leqslant m-1} |v_j|, \qquad n = m, m+1, \ldots \qquad (18)$$

Тем самым устойчивость означает равномерную по n ограниченность решения задачи Коши.

Оказывается, что устойчивость или неустойчивость уравнения (8) по начальным данным целиком определяется расположением корней характеристического уравнения (9).

Будем говорить, что выполнено *условие корней*, если все корни $q_1, \ldots, q_m$ характеристического уравнения (9) лежат внутри или

на границе единичного круга комплексной плоскости, причем на границе единичного круга нет кратных корней. Справедлива

Теорема 1. *Условие корней необходимо и достаточно для устойчивости уравнения* (8) *по начальным данным.*

Доказательство. Докажем сначала необходимость. Пусть уравнение (8) имеет корень q, для которого $|q|>1$. Задавая в качестве начальных данных функции $v_j=q^j$, $j=0, 1, \ldots, m-1$, получим решение $v_n=q^n$, $n\geqslant m$, неограниченно возрастающее при $n\to\infty$. Для такого решения невозможна оценка вида (18) с константой M_1, не зависящей от n. Следовательно, условие $|q_k|\leqslant 1$, $k=1, 2, \ldots, m$, необходимо для устойчивости.

Пусть уравнение (9) имеет корень q кратности $r>1$, для которого $|q|=1$. Тогда разностное уравнение (8) имеет решение $n^{r-1}q^n$, растущее при $n\to\infty$ как n^{r-1}, и, следовательно, в этом случае оценка (18) также невозможна.

Прежде чем переходить к доказательству достаточности условий теоремы 1, необходимо провести некоторые вспомогательные построения.

Запишем (8) в виде эквивалентной системы уравнений

$$v_{n-m+1}=v_{n-m+1},\ \ldots,\ v_{n-1}=v_{n-1},$$

$$v_n=-\frac{a_m}{a_0}v_{n-m}-\frac{a_{m-1}}{a_0}v_{n-m+1}-\ldots-\frac{a_1}{a_0}v_{n-1}$$

и представим эту систему в векторной форме

$$V_n=SV_{n-1},\quad n=m,\ m+1,\ldots, \tag{19}$$

где $V_n=(v_{n-m+1}, v_{n-m+2}, \ldots, v_n)^T$,

$$S=\begin{bmatrix} 0 & 1 & 0 & \cdot & 0 \\ 0 & 0 & 1 & \cdot & 0 \\ \cdot\cdot\cdot & \cdot\cdot\cdot\cdot & \cdot\cdot\cdot\cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & 1 \\ -\frac{a_m}{a_0} & -\frac{a_{m-1}}{a_0} & -\frac{a_{m-2}}{a_0} & \cdot & -\frac{a_1}{a_0} \end{bmatrix}. \tag{20}$$

Начальный вектор $V_{m-1}=(v_0, v_1, \ldots, v_{m-1})^T$ задан.

Нетрудно проверить, что множество собственных чисел матрицы S совпадает с множеством корней характеристического уравнения (9).

Лемма 1. *Если выполнено условие корней, то существует норма* $\|\cdot\|_*$ *вектора такая, что для подчиненной нормы матрицы* S *справедливо неравенство* $\|S\|_*\leqslant 1$.

Доказательство. Заметим, что данная лемма уточняет лемму 1 из § 3 гл. 2, поэтому их доказательства похожи (см. [2, с. 338]). С помощью преобразования подобия

$$\check{S}=QSQ^{-1} \tag{21}$$

приведем S к модифицированной жордановой форме

$$\hat{S}=\begin{bmatrix}\hat{S}_1 & & 0\\ & \hat{S}_2 & \\ & \ddots & \\ 0 & & \hat{S}_l\end{bmatrix},$$

где $\hat{S}_k$ — либо число, либо жорданова клетка

$$\hat{S}_k=\begin{bmatrix}q_k & \varepsilon & & 0\\ & \ddots & \ddots & \\ & & q_k & \varepsilon\\ 0 & & & q_k\end{bmatrix},$$

q_k — собственное число матрицы S, $\varepsilon>0$ — любое число. Оценим норму

$$\|\hat{S}\|_C=\max_{1\leqslant i\leqslant m}\sum_{j=1}^{m}|\hat{s}_{ij}|=\max_{1\leqslant i<m}(|\hat{s}_{ii}|+|\hat{s}_{ii+1}|).$$

Здесь $\hat{s}_{ii}$ совпадает с одним из корней характеристического уравнения, а внедиагональный элемент

$$\hat{s}_{i,i+1}=\begin{cases}0 \text{ в случае простого корня } \hat{s}_{ii},\\ \varepsilon \text{ в случае кратного корня } \hat{s}_{ii}.\end{cases}$$

Если $\hat{s}_{i,i+1}=0$ для некоторого i, то по условию леммы $|s_{ii}|\leqslant 1$. Если же $\hat{s}_{ii+1}=\varepsilon$, то $\hat{s}_{ii}$ — кратный корень, и согласно условию корней выполняется строгое неравенство $|\hat{s_{ii}}|<1$. Но тогда при достаточно малом ε имеем

$$|\hat{s}_{ii}|+|\hat{s}_{ii+1}|<1.$$

Таким образом, выбирая ε достаточно малым, получим $\|\hat{S}\|_C\leqslant 1$. Введем норму вектора

$$\|y\|_*=\|Qy\|_c, \tag{22}$$

где Q определено согласно (21). Тогда получим $\|S\|_*=\|\hat{S}\|_C\leqslant 1$. Лемма 1 доказана.

Завершим теперь доказательство теоремы 1. Покажем, что условие корней достаточно для устойчивости уравнения (8) по начальным данным.

Из уравнения (19), учитывая лемму 1, получим неравенство

$$\|V_n\|_*\leqslant\|S\|_*\|V_{n-1}\|_*\leqslant\|V_{n-1}\|_*,$$

и, следовательно,

$$\|V_n\|_*\leqslant\|V_{m-1}\|_*, \quad n=m,\ m+1,\ldots \tag{23}$$

По определению (22) нормы $\|\cdot\|_*$ имеем

$$\|V\|_*=\|QV\|_C\leqslant\|Q\|_c\|V\|_C.$$

С другой стороны, для любой невырожденной матрицы Q справедливо тождество

$$V=Q^{-1}QV,$$

из которого следует оценка

$$\|V\|_c\leqslant\|Q^{-1}\|_c\|QV\|_c=\|Q^{-1}\|_c\|V\|_*.$$

Таким образом, если норма $\|\cdot\|_*$ определена равенством (22), то выполняются оценки

$$(\|Q^{-1}\|_c)^{-1}\|V_c\|\leqslant\|V\|_*\leqslant\|Q\|_c\|V\|_c. \tag{24}$$

Из (23) и (24) получаем

$$\|V_n\|_c\leqslant M_1\|V_{m-1}\|_c, \tag{25}$$

где $M_1=\|Q^{-1}\|_c\|Q\|_c$. Заметим, что константа M_1 не зависит от n. Из (25) следует неравенство

$$|v_n|\leqslant M_1\max_{0\leqslant j\leqslant m-1}|v_j|, \tag{26}$$

означающее устойчивость уравнения (8) по начальным данным. Теорема 1 полностью доказана.

4. Оценка решения неоднородного уравнения. Рассмотрим задачу Коши для неоднородного разностного уравнения

$$a_0y_n+a_1y_{n-1}+\ldots+a_my_{n-m}=\tau g_{n-m}, \tag{27}$$

где $n=m,\ m+1,\ldots$, величины $y_0,\ y_1,\ \ldots,\ y_{m-1}$ заданы, правая часть g_k, $k=0,\ 1,\ \ldots$ — заданная функция. Если $a_0\neq0$, то для каждой правой части решение задачи Коши существует и единственно. Оно может быть найдено по рекуррентной формуле

$$y_n=-\frac{a_m}{a_0}y_{n-m}-\frac{a_{m-1}}{a_0}y_{n-m+1}-\ldots-\frac{a_1}{a_0}y_{n-1}+\frac{\tau g_{n-m}}{a_0},$$
$$n=m,\ m+1,\ \ldots, \tag{28}$$

исходя из заданных начальных условий $y_0,\ y_1,\ \ldots,\ y_{m-1}$ и известной правой части q_k.

В предыдущем пункте получена оценка решения однородного уравнения через начальные данные, означающая устойчивость по начальным данным. Получим теперь аналогичную оценку решения неоднородного уравнения (27) через начальные данные и правую часть.

Основным результатом здесь будет доказательство того, что *если однородное уравнение* (8) *устойчиво по начальным данным, то для неоднородного уравнения* (27) *справедлива оценка*

$$|y_n|\leqslant M_1\max_{0\leqslant j\leqslant m-1}|y_j|+M_2\sum_{k=0}^{n-m}\tau|g_k|, \tag{29}$$

где M_1 и M_2 не зависят от n.

Выполнение оценки (29) означает по определению устойчивость уравнения (27) по правой части. Таким образом, будет показано,

что из устойчивости по начальным данным следует устойчивость по правой части.

Представим уравнение (27) в векторном виде

$$Y_n = SY_{n-1} + \tau G_{n-1}, \qquad n = m, m+1, \ldots, \tag{30}$$

где $Y_n = (y_{n-m+1}, y_{n-m+2}, \ldots, y_n)^T$,

$$G_{n-1} = \left(0, 0, \ldots, 0, \frac{g_{n-m}}{a_0}\right)^T, \tag{31}$$

матрица S определена согласно (20).

Пусть уравнение (8) устойчиво по начальным данным. Тогда согласно теореме 1 выполнено условие корней. При этом условии в соответствии с леммой 1 для некоторой нормы матрицы S справедливо неравенство $\|S\|_* \leqslant 1$. Таким образом, из (30) получаем неравенства

$$\|Y_k\|_* \leqslant \|Y_{k-1}\|_* + \tau \|G_{k-1}\|_*, \qquad k = m, m+1, \ldots$$

Суммируя эти неравенства по k от m до n, получим

$$\|Y_n\|_* \leqslant \|Y_{m-1}\|_* + \sum_{k=m-1}^{n-1} \tau \|G_k\|_*. \tag{32}$$

Используем далее неравенство (24), устанавливающее эквивалентность норм $\|\cdot\|_*$ и $\|\cdot\|_C$. Тогда из (32) получим

$$(\|Q^{-1}\|_C)^{-1} \|Y_n\|_C \leqslant \|Q\|_C \left(\|Y_{m-1}\|_C + \sum_{k=m-1}^{n-1} \tau \|G_k\|_C\right). \tag{33}$$

Учитывая специальный вид (31) вектора G_k, имеем

$$\|G_k\|_C = \frac{|g_{k+1-m}|}{|a_0|},$$

и, следовательно,

$$\sum_{k=m-1}^{n-1} \tau \|G_k\|_C = \frac{1}{a_0} \sum_{k=0}^{n-m} \tau |g_k|.$$

Отсюда и из (33) получаем требуемое неравенство

$$\|y_n\|_C \leqslant M_1 \max_{0 \leqslant j \leqslant m-1} |y_j| + M_2 \sum_{k=0}^{n-m} \tau |g_k|, \tag{34}$$

где $M_1 = \|Q^{-1}\|_C \|Q\|_C$, $M_2 = M_1 a_0^{-1}$. Напомним, что Q — матрица, преобразующая S согласно (21) к модифицированной жордановой форме.

5. Оценки погрешности разностного метода. Вернемся к уравнению для погрешности (3), полученному в п. 1. Нам нужно оценить погрешность метода z_n через погрешность аппроксимации ψ_k, $k = 0, 1, \ldots, n-m$. Если бы функция φ_k равнялась нулю, $k = 0, 1, \ldots$

..., $n-m$, то достаточно было бы воспользоваться оценкой (34). Однако наличие функции φ_k, зависящей от решения z и характеризующей нелинейность задачи, усложняет получение требуемых оценок. Следуя [2, с. 484], докажем, что справедлива

Теорема 2. *Пусть все корни характеристического уравнения* (9) *лежат внутри или на границе единичного круга, причем на границе нет кратных корней. Пусть* $|f_u(t, u)| \leqslant L$ *для* $t \in (0, T]$. *Тогда при*

$$\tau \leqslant \tau_0 = \frac{|a_0|}{2|b_0|L} \tag{35}$$

для решения уравнения (3) *справедлива оценка*

$$|z_n| \leqslant M_1 e^{c_1 T} \max_{0 \leqslant j \leqslant m-1} |z_j| + M_2 \sum_{k=0}^{n-m} \tau |\psi_k|, \quad n = m, m+1, \ldots, \tag{36}$$

где

$$M_1 = \|Q\|_c \|Q^{-1}\|_c,$$

$$M_2 = \frac{2M_1}{|a_0|}, \quad c_1 = M_1 v L, \quad v = \frac{2}{a_0^2} \sum_{k=1}^{m} (|a_0|\,|b_k| + |a_k|\,|b_0|).$$

Доказательство. Согласно формуле конечных приращений Лагранжа имеем

$$f(t_{n-k}, y_{n-k}) - f(t_{n-k}, u_{n-k}) = l_{n-k} z_{n-k},$$

где

$$l_{n-k} = f_u(t_{n-k}, \tilde{u}_{n-k}), \quad \tilde{u}_{n-k} = y_{n-k} + \theta z_{n-k}, \quad 0 \leqslant \theta \leqslant 1, \quad k = 0, 1, 2, \ldots, m.$$

Представим функцию φ_{n-m}, определенную согласно (5), в виде

$$\varphi_{n-m} = b_0 l_0 z_n + \sum_{k=1}^{m} b_k l_{n-k} z_{n-k}. \tag{37}$$

Подставляя (37) в уравнение (3), получим

$$(a_0 - b_0 l_0 \tau) z_n = \sum_{k=1}^{m} (-a_k + \tau b_k l_{n-k}) z_{n-k} + \tau \psi_{n-m}. \tag{38}$$

Из условия (35) следует, что

$$|a_0 - \tau b_0 l_0| \geqslant \frac{|a_0|}{2} > 0, \tag{39}$$

поэтому уравнение (38) можно разрешить относительно z_n:

$$z_n = \sum_{k=1}^{m} \frac{-a_k + \tau b_k l_{n-k}}{a_0 - \tau b_0 l_0} z_{n-k} + \tau \frac{\psi_{n-m}}{a_0 - \tau b_0 l_0}. \tag{40}$$

Добавляя к правой части уравнения (40) и вычитая из нее выражение

$$\sum_{k=1}^{m} \frac{a_k}{a_0} z_{n-k},$$

перепишем (40) в виде

$$z_n = -\sum_{k=1}^{m} \frac{a_k}{a_0} z_{n-k} + \tau \sum_{k=1}^{m} v_{nk} z_{n-k} + \tau \frac{\psi_{n-m}}{a_0 - \tau b_0 l_0}, \tag{41}$$

где

$$v_{nk} = \frac{a_0 b_k l_{n-k} - a_k b_0 l_0}{a_0 (a_0 - \tau b_0 l_0)}. \tag{42}$$

Представим уравнение (41) в векторной форме. Для этого введем векторы

$$Z_n = (z_{n-m+1}, z_{n-m+2}, \ldots, z_n)^T,$$

$$\Psi_{n-1} = \left(0, \ldots, 0, \frac{\psi_{n-m}}{a_0 - \tau b_0 l_0}\right)^T$$

и матрицы

$$V_{n-1} = \begin{bmatrix} 0 & \ldots & 0 \\ \cdot & \cdot & \cdot \\ 0 & \ldots & 0 \\ v_{nm} & \ldots & v_{n1} \end{bmatrix}, \quad S = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ \cdot & \cdot & \cdot & \ldots & \cdot \\ 0 & 0 & 0 & \ldots & 1 \\ -\frac{a_m}{a_0} & -\frac{a_{m-1}}{a_0} & -\frac{a_{m-2}}{a_0} & \ldots & -\frac{a_1}{a_0} \end{bmatrix}.$$

Тогда получим, что уравнение (41) эквивалентно векторному уравнению

$$Z_n = SZ_{n-1} + \tau V_{n-1} Z_{n-1} + \tau \Psi_{n-1}. \tag{43}$$

Для оценки решения уравнения (43) используем лемму 1 из п. 3. Согласно этой лемме, $\|S\|_* \leqslant 1$ в некоторой норме $\|\cdot\|_*$. Поэтому из (43) получим

$$\|Z_n\|_* \leqslant \|Z_{n-1}\|_* + \tau \|V_{n-1}\|_* \|Z_{n-1}\|_* + \tau \|\Psi_{n-1}\|_*. \tag{44}$$

Покажем, что $\|V_{n-1}\|_*$ ограничена числом, не зависящим от n и τ. Согласно (39), (42) имеем

$$|v_{nk}| \leqslant 2 \frac{|a_0||b_k| + |a_k||b_0|}{|a_0|^2} L, \quad k = 1, 2, \ldots, m,$$

поэтому

$$\|V_{n-1}\|_C = \sum_{k=1}^{m} |v_{nk}| \leqslant vL,$$

где

$$v = \frac{2}{a_0^2} \sum_{k=1}^{m} (|a_0||b_k| + |a_k||b_0|). \tag{45}$$

Для любого вектора x имеем

$$\|V_{n-1}x\|_* = \|QV_{n-1}x\|_C = \|(QV_{n-1}Q^{-1})(Qx)\|_C \leqslant$$

$$\leqslant \|QV_{n-1}Q^{-1}\|_C \|Qx\|_C \leqslant M_1 \|V_{n-1}\|_C \|x\|_*,$$

где $M_1=\|Q\|_C\|Q^{-1}\|_C$. Следовательно,

$$\|V_{n-1}\|_*\leqslant M_1\|V_{n-1}\|_C\leqslant M_1 vL.$$

Подставляя эту оценку в (44), приходим к неравенству

$$\|Z_k\|_*\leqslant(1+\tau c_1)\|Z_{k-1}\|_*+\tau\|\Psi_{k-1}\|_*, \tag{46}$$

где $k=m, m+1, \ldots, c_1=M_1 vL$. Из (46) получим

$$\|Z_n\|_*\leqslant(1+\tau c_1)^{n-m}\|Z_{m-1}\|_*+\tau\sum_{k=m}^{n}\|\Psi_{k-1}\|_*\leqslant$$

$$\leqslant e^{c_1 t_{n-m}}\|Z_{m-1}\|_*+\tau\sum_{k=m}^{n}\|\Psi_{k-1}\|_*, \tag{47}$$

где $t_{n-m}=\tau(n-m)\leqslant T$.

Далее, учитывая (24), (39), имеем

$$\|Z_n\|_*\geqslant(\|Q^{-1}\|_C)^{-1}\|Z_n\|_C\geqslant(\|Q^{-1}\|_C)^{-1}|z_n|,$$

$$\|Z_{m-1}\|_*\leqslant\|Q\|_C\|Z_{m-1}\|_C=\|Q\|_C\max_{0\leqslant j\leqslant m-1}|z_j|,$$

$$\|\Psi_{k-1}\|_*\leqslant\|Q\|_C\|\Psi_{k-1}\|_C=\frac{\|Q\|_C}{|a_0-\tau b_0 l_0|}|\psi_{k-m}|\leqslant\frac{2\|Q\|_C}{|a_0|}|\psi_{k-m}|.$$

Подставляя эти неравенства в (47) и обозначая $M_1=\|Q\|_C\times$ $\times\|Q^{-1}\|_C$, $M_2=M_1\dfrac{2}{|a_0|}$, получим оценку

$$|z_n|\leqslant M_1 e^{c_1 T}\max_{0\leqslant j\leqslant m-1}|z_j|+M_2\sum_{k=0}^{n-m}\tau|\psi_k|,$$

совпадающую с (36). Теорема 2 доказана.

Следствие. *Если* $0\leqslant n\tau\leqslant T$, *выполнено условие корней,* $|z_j|\to 0$ *при* $\tau\to 0$, $j=0, 1, \ldots, m-1$, *и разностное уравнение* (2) *аппроксимирует исходное уравнение* (1), *то решение разностной задачи* (2) *сходится при* $\tau\to 0$ *к решению исходной задачи* (1).

Доказательство следует немедленно из оценки (36), если учесть, что

$$\sum_{k=0}^{n-m}\tau|\psi_k|\leqslant T\max_{0\leqslant k\leqslant n-m}|\psi_k|.$$

§ 5. Численное интегрирование жестких систем обыкновенных дифференциальных уравнений

1. Условно устойчивые и абсолютно устойчивые разностные методы. Для задачи Коши

$$\frac{du}{dt}=f(t,u),\quad t>0,\qquad u(0)=u_0 \tag{1}$$

будем рассматривать разностные методы вида

$$\sum_{k=0}^{m} \frac{a_k}{\tau} y_{n-k} = \sum_{k=0}^{m} b_k f(t_{n-k}, y_{n-k}), \qquad n = m,\ m+1,\ \ldots \tag{2}$$

В § 4 показано, что устойчивость и сходимость метода определяются расположением корней характеристического уравнения

$$\sum_{k=0}^{m} a_k q^{m-k} = 0. \tag{3}$$

А именно, требуется, чтобы все корни удовлетворяли условию $|q| \leqslant 1$, причем корни q, для которых $|q| = 1$, не должны быть кратными.

Эти условия устойчивости являются очень общими и не могут учесть многие характерные свойства решений исходной дифференциальной задачи (1) и аппроксимирующего ее разностного метода (2). Они означают лишь, что все решения однородного разностного уравнения, соответствующего (2), остаются ограниченными при $n \to \infty$.

В частности, при таком подходе коэффициенты b_k, $k = 0, 1, \ldots, m$, входящие в правую часть уравнения (2), никак не влияют на устойчивость.

Предположим, однако, что заранее известна та или иная характерная особенность в поведении решения исходной дифференциальной задачи. Тогда естественно требовать, чтобы эта особенность сохранялась и у решения разностного уравнения. Такое требование приведет к сужению класса допустимых разностных методов. В настоящем параграфе будут рассмотрены методы, предназначенные для расчета асимптотически устойчивых решений уравнения (1).

Рассмотрим сначала характерный пример. Уравнение

$$\frac{du}{dt} = \lambda u, \quad t > 0, \qquad u(0) = u_0, \tag{4}$$

где $\lambda < 0$, имеет решение

$$u(t) = u_0 e^{\lambda t},$$

монотонно убывающее при $t \to \infty$. При любых $\tau > 0$ для решения этого уравнения справедливо неравенство

$$|u(t+\tau)| \leqslant |u(t)|, \tag{5}$$

означающее устойчивость решения $u(t)$.

Естественно требовать, чтобы и для решения разностной задачи, аппроксимирующей (4), выполнялось бы неравенство, аналогичное (5). Рассмотрим с этой точки зрения метод Эйлера

$$\frac{y_{n+1} - y_n}{\tau} = \lambda y_n, \qquad n = 0, 1, \ldots \tag{6}$$

Из уравнения (6) получаем

$$y_{n+1} = q y_n, \quad q = 1 + \tau\lambda.$$

Оценка вида (5), т. е. неравенство

$$|y_{n+1}| \leqslant |y_n|, \quad n=1, 2, \ldots \tag{7}$$

для метода (6) будет выполнено тогда и только тогда, когда $|q| \leqslant 1$. В случае $\lambda<0$ это условие эквивалентно следующему ограничению на шаг τ:

$$0<\tau \leqslant \frac{2}{|\lambda|}. \tag{8}$$

Таким образом, разностный метод (6) устойчив в смысле выполнения оценки (7), если шаг τ удовлетворяет неравенству (8).

Разностный метод (2) называется *абсолютно устойчивым*, если он устойчив при любых $\tau>0$, и *условно устойчивым*, если он устойчив при некоторых ограничениях на шаг τ. Мы видели, что метод Эйлера (6) условно устойчив при условии (8). Примером абсолютно устойчивого метода для уравнения (4) с $\lambda<0$ является *неявный метод Эйлера*

$$\frac{y_{n+1}-y_n}{\tau} = \lambda y_{n+1},$$

для которого $|q| = |(1-\tau\lambda)^{-1}| < 1$ при любых $\tau>0$.

Приведенные здесь простые примеры характерны тем, что и для более общих асимптотически устойчивых систем дифференциальных уравнений явные разностные методы являются условно устойчивыми, а среди неявных методов существуют абсолютно устойчивые методы.

Условная устойчивость является недостатком явного метода, так как вынуждает брать слишком мелкий шаг τ. Например, если $\lambda = -200$, то условие (8) выполнено при $\tau \leqslant 0{,}01$, и для того чтобы вычислить решение $u(t)$ при $t=1$, надо сделать сто шагов по методу Эйлера. Неявный метод лишен этого недостатка, однако его применение к задаче (1) приводит к необходимости решения на каждом шаге системы алгебраических уравнений, в общем случае нелинейной.

2. Понятие жесткой системы дифференциальных уравнений. Многие из рассмотренных в § 1—4 численных методов интегрирования обыкновенных дифференциальных уравнений переносятся без изменений на системы дифференциальных уравнений. Однако в случае численного решения систем уравнений могут появиться дополнительные трудности, связанные с разномасштабностью процессов, описываемых данной системой.

Поясним характер возникающих трудностей на примере системы, состоящей из двух независимых уравнений

$$\frac{du_1}{dt} + a_1 u_1 = 0, \quad \frac{du_2}{dt} + a_2 u_2 = 0, \quad t>0, \tag{9}$$

где a_1 и a_2 — положительные постоянные.

Система (9) имеет решение

$$u_1(t) = u_1(0)\, e^{-a_1 t}, \quad u_2(t) = u_2(0)\, e^{-a_2 t},$$

монотонно убывающее с ростом t. Предположим, что a_2 гораздо больше, чем a_1. Тогда компонента $u_2(t)$ затухает гораздо быстрее, чем $u_1(t)$, и, начиная с некоторого t, поведение решения $u(t)=\{u_1(t),\ u_2(t)\}$ почти полностью определяется компонентой $u_1(t)$. Однако оказывается, что при решении системы (9) разностным методом шаг интегрирования τ определяется, как правило, компонентой $u_2(t)$, не существенной с точки зрения поведения решения системы. Например, метод Эйлера

$$\frac{u_1^{n+1}-u_1^n}{\tau}+a_1u_1^n=0,\quad \frac{u_2^{n+1}-u_2^n}{\tau}+a_2u_2^n=0, \tag{10}$$

где $u_i^n=u_i(t_n)$, $i=1,\ 2$, будет устойчив, если шаг τ удовлетворяет одновременно двум неравенствам $\tau a_1\leqslant 2$, $\tau a_2\leqslant 2$. Поскольку $a_2\gg a_1>0$, условие устойчивости приводит к ограничению $\tau\leqslant 2/a_2$.

Приведенный пример может показаться искусственным, так как ясно, что каждое из уравнений системы (9) следует решать независимо одно от другого со своим шагом интегрирования τ_j, $j=1,\ 2$, $\tau_1\leqslant 2/a_1$, $\tau_2\leqslant 2/a_2$. Однако аналогичные трудности возникают и при решении любой системы обыкновенных дифференциальных уравнений

$$\frac{du}{dt}=Au, \tag{11}$$

если матрица этой системы имеет большой разброс собственных чисел.

Предположим, например, что матрицу A системы (11) можно привести преобразованием подобия $Q^{-1}AQ$ к диагональному виду. Тогда замена $u=Qv$ преобразует систему (11) в систему независимых уравнений

$$\frac{dv}{dt}=Q^{-1}AQv,$$

матрица которой имеет те же собственные числа, что и матрица A.

Сформулируем теперь определение жесткой системы уравнений. Рассмотрим сначала систему (11) с постоянной, т. е. не зависящей от t матрицей A. Система дифференциальных уравнений (11) с постоянной матрицей $A(m\times m)$ называется *жесткой*, если

1) $$\operatorname{Re}\lambda_k<0,\qquad k=1,\ 2,\ \ldots,\ m$$

(т. е. система асимптотически устойчива по Ляпунову),

2) отношение

$$s=\frac{\max\limits_{1\leqslant k\leqslant m}|\operatorname{Re}\lambda_k|}{\min\limits_{1\leqslant k\leqslant m}|\operatorname{Re}\lambda_k|}$$

велико.

Число s называется *числом жесткости* системы (11). Второе требование не указывает границу для s, начиная с которой система становится жесткой.

Если матрица A зависит от t, то $\lambda_k = \lambda_k(t)$, $k = 1, 2, \ldots, m$. При каждом t можно определить число жесткости

$$s(t) = \frac{\max\limits_{1 \leqslant k \leqslant m} |\operatorname{Re} \lambda_k(t)|}{\min\limits_{1 \leqslant k \leqslant m} |\operatorname{Re} \lambda_k(t)|}.$$

В этом случае свойство жесткости может зависеть от длины отрезка интегрирования. Система

$$\frac{du}{dt} = A(t)\, u$$

называется *жесткой на интервале* $(0, T)$, если $\operatorname{Re} \lambda_k(t) < 0$, $k = 1, 2, \ldots, m$, для всех $t \in (0, T)$ и число $\sup\limits_{t \in (0,T)} s(t)$ велико.

Так же, как и в случае системы (10), нетрудно прийти к следующему выводу. Решение жесткой системы содержит как быстро убывающие, так и медленно убывающие составляющие. Начиная с некоторого $t > 0$ решение системы почти полностью определяется медленно убывающей составляющей. Однако при использовании явных разностных методов быстро убывающая составляющая отрицательно влияет на устойчивость, что вынуждает брать шаг интегрирования τ слишком мелким.

Выход из этой парадоксальной ситуации был найден в применении неявных абсолютно устойчивых разностных методов. Например, систему (10) можно решать с помощью неявного метода Эйлера

$$\frac{u_1^{n+1} - u_1^n}{\tau} + a_1 u_1^{n+1} = 0, \quad \frac{u_2^{n+1} - u_2^n}{\tau} + a_2 u_2^{n+1} = 0,$$

который устойчив при всех $\tau > 0$. Поэтому шаг интегрирования τ здесь можно выбирать, руководствуясь лишь соображениями точности, а не устойчивости.

3. Нелинейные системы дифференциальных уравнений. Обобщим понятие жесткости на случай нелинейной системы

$$\frac{du}{dt} = f(t, u), \qquad t > 0, \tag{12}$$

где

$$u(t) = (u_1(t), u_2(t), \ldots, u_m(t))^T,$$

$$f(t, u) = (f_1(t, u), f_2(t, u), \ldots, f_m(t, u))^T.$$

Зафиксируем какое-либо решение $v(t)$ системы (12) и образуем разность $z(t) = u(t) - v(t)$ между произвольным решением системы (12) и данным решением $v(t)$. Эта разность удовлетворяет следующей системе уравнений:

$$\frac{dz_k}{dt} = f_k(t, v(t) + z(t)) - f_k(t, v(t)), \qquad k = 1, 2, \ldots, m. \tag{13}$$

Будем рассматривать $z(t)$ как малое возмущение, внесенное в основное решение $v(t)$.

Проведем разложение по формуле Тейлора в правой части системы (13). Так как

$$f_k(t, u) = f_k(t, u_1, u_2, \ldots, u_m),$$

имеем

$$f_k(t, v+z) - f_k(t, v) = \sum_{j=1}^{m} \frac{\partial f_k(t, v)}{\partial u_j} z_j(t) + o(|z|),$$

где через $o(|z|)$ обозначены величины более высокого, чем первый, порядка малости по z. В результате разложения система (13) примет вид

$$\frac{dz}{dt} = A(t, v(t))\, z(t) + o(|z|), \tag{14}$$

где через $A(t, v(t)) = \dfrac{\partial f(t, v(t))}{\partial u}$ обозначена матрица с элементами

$$a_{ij}(t, v(t)) = \frac{\partial f_i(t, v(t))}{\partial u_j}, \quad i, j = 1, 2, \ldots, m.$$

Отбрасывая в (14) величины $o(|z|)$, получим так называемую систему уравнений первого приближения

$$\frac{dw(t)}{dt} = A(t, v(t))\, w(t). \tag{15}$$

Система (15) является системой линейных дифференциальных уравнений относительно $w(t)$, так как функция $v(t)$ задана.

Определение жесткости системы нелинейных дифференциальных уравнений связано как с данным фиксированным решением $v(t)$, так и с длиной отрезка интегрирования.

Пусть $\lambda_k(t)$, $k=1, 2, \ldots, m$,— собственные числа матрицы $A(t, v(t))$.

Число жесткости $s(t)$ определяется как

$$s(t) = \frac{\max\limits_{1 \leqslant k \leqslant m} |\operatorname{Re} \lambda_k(t)|}{\min\limits_{1 \leqslant k \leqslant m} |\operatorname{Re} \lambda_k(t)|}.$$

Система (12) называется *жесткой на решении* $v(t)$ *и на данном интервале* $0 < t < T$, если

1) $\operatorname{Re} \lambda_k(t) < 0$, $k=1, 2, \ldots, m$, для всех $t \in (0, T)$,

2) число $\sup\limits_{t \in (0,T)} s(t)$ велико.

4. Специальные определения устойчивости. При исследовании разностных методов для жестких систем уравнений обычно рассматривают уравнение

$$\frac{du}{dt} = \lambda u, \tag{16}$$

где λ — произвольное комплексное число. Свойства различных разностных методов изучают и сопоставляют на примере модельного уравнения (16). Для того чтобы уравнение (16) действительно моделировало исходную систему (11), необходимо рассматривать его при всех таких λ, которые являются собственными числами матрицы A.

Разностный метод (2), примененный к уравнению (16), имеет вид

$$\sum_{k=0}^{m} (a_k - \mu b_k) y_{n-k} = 0, \quad n = m, m+1, \ldots, \tag{17}$$

где $\mu = \tau\lambda$ — комплексный параметр.

Если искать решения уравнения (17), имеющие вид $y_n = q^n$, то для q получим характеристическое уравнение

$$\sum_{k=0}^{m} (a_k - \mu b_k) q^{m-k} = 0, \tag{18}$$

отличающееся от уравнения (3) тем, что его коэффициенты зависят от параметра $\mu = \tau\lambda$. При малых μ корни уравнений (3) и (18) близки. Однако в дальнейшем мы не будем делать предположений относительно малости μ.

Кроме обычного определения устойчивости разностного метода (все корни характеристического уравнения (18) не превосходят по модулю единицу), в случае жестких систем используют и другие, более узкие определения устойчивости. Здесь мы рассмотрим два таких определения: A-устойчивый метод и $A(\alpha)$-устойчивый метод.

Предварительно введем следующее понятие. *Областью устойчивости разностного метода* (2) называется множество всех точек комплексной плоскости $\mu = \tau\lambda$, для которых данный метод, примененный к уравнению (16), является устойчивым.

Рассмотрим, например, явный метод Эйлера

$$\frac{y_{n+1} - y_n}{\tau} = f(t_n, y_n).$$

В применении к уравнению (16) этот метод принимает вид

$$y_{n+1} = (1 + \mu) y_n, \quad \mu = \tau\lambda.$$

Условие устойчивости $|1 + \mu| \leqslant 1$ для комплексного $\mu = \mu_0 + i\mu_1$ означает, что $(\mu_0 + 1)^2 + \mu_1^2 \leqslant 1$. Тем самым область устойчивости данного метода представляет собой круг единичного радиуса с центром в точке $(-1, 0)$.

Для неявного метода Эйлера

$$\frac{y_{n+1} - y_n}{\tau} = f(t_{n+1}, y_{n+1})$$

областью устойчивости является внешность круга единичного радиуса с центром в точке $(1, 0)$.

Разностный метод называется *A-устойчивым*, если область его устойчивости содержит левую полуплоскость $\operatorname{Re}\mu<0$. Отметим, что уравнение (16) асимптотически устойчиво при $\operatorname{Re}\lambda<0$. Поэтому сущность приведенного определения состоит в том, что A-устойчивый разностный метод является абсолютно устойчивым (устойчивым при любых $\tau>0$), если устойчиво решение исходного дифференциального уравнения.

Нетрудно видеть, что неявный метод Эйлера является A-устойчивым, а явный метод Эйлера — не является.

Рассмотрим еще одношаговый метод второго порядка точности

$$\frac{y_{n+1}-y_n}{\tau}=0{,}5\,(f(t_{n+1},\ y_{n+1})+f(t_n,\ y_n)). \tag{19}$$

Для уравнения (16) метод принимает вид

$$y_{n+1}=qy_n, \qquad q=\frac{1+0{,}5\mu}{1-0{,}5\mu}.$$

Отсюда видно, что $|q|\leqslant 1$ тогда и только тогда, когда $\operatorname{Re}\mu\leqslant 0$. Следовательно, метод (19) является A-устойчивым.

При решении жестких систем уравнений было бы желательно пользоваться именно A-устойчивыми разностными методами, так как условия их устойчивости не накладывают ограничений на шаг τ. Оказывается, однако, что класс A-устойчивых методов весьма узок. В частности, *среди методов вида* (2) *не существует явных A-устойчивых методов.*

Для доказательства запишем характеристическое уравнение (18) в виде

$$\mu=\frac{a_0q^m+a_1q^{m-1}+\ldots+a_m}{b_0q^m+b_1q^{m-1}+\ldots+b_m}. \tag{20}$$

Если (2) — явный m-шаговый метод, то $b_0=0$, $a_0\neq 0$. Могут оказаться равными нулю и другие коэффициенты b_k, но не все, так как по условию $b_0+b_1+\ldots+b_m=1$ (см. (3) из § 3). Пусть $b_0=b_1=\ldots=b_{j-1}=0$, $b_j\neq 0$, $0<j\leqslant m$. Тогда из (20) получим

$$\mu=\frac{a_0q^m+a_1q^{m-1}+\ldots+a_m}{b_jq^{m-j}+\ldots+b_m}.$$

Отсюда видно, что при больших q функция $\mu(q)$ ведет себя как $\frac{a_0}{b_j}q^j$, $j\geqslant 1$.

Следовательно, для любого достаточно большого по модулю числа μ (в том числе и для μ, лежащих в левой полуплоскости) найдется корень q уравнения (18) с $|q|\geqslant 1$.

Доказано (см. [37] и указанную там литературу), что *среди неявных линейных многошаговых методов нет A-устойчивых методов, имеющих порядок точности выше второго.* Примером A-устойчивости метода второго порядка точности является симметричная схема (19).

В связи с этим было введено еще несколько определений устойчивости, которые являются менее ограничительными, чем определение A-устойчивости.

Разностный метод называется $A(\alpha)$-*устойчивым*, если область его устойчивости содержит угол

$$|\arg(-\mu)|<\alpha, \quad \mu=\tau\lambda.$$

В частности, $A\left(\frac{\pi}{2}\right)$-устойчивость совпадает с A-устойчивостью.

Доказано, что *ни для какого* α *не существует явного* $A(\alpha)$-*устойчивого линейного многошагового метода.* Построены $A(\alpha)$-устойчивые неявные методы третьего и четвертого порядка точности. К ним относятся, в частности, чисто неявные многошаговые разностные схемы, у которых правая часть $f(t, u)$ вычисляется только при $t=t_n$, а производная $u'(t)$ аппроксимируется в точке t_n по нескольким предыдущим точкам. Например, схема

$$\frac{25y_n-48y_{n-1}+36y_{n-2}-16y_{n-3}+3y_{n-4}}{12\tau}=f(t_n, y_n) \tag{21}$$

имеет четвертый порядок точности и $A(\alpha)$-устойчива при некотором $\alpha>0$.

5. Чисто неявные разностные методы. В настоящее время при интегрировании жестких систем уравнений широко используется метод Гира [37], в основу которого положены чисто неявные многошаговые разностные методы высокого порядка точности.

Разностный метод

$$\sum_{k=0}^{m} a_k y_{n-k}=\tau f(t_n, y_n) \tag{22}$$

называется *чисто неявным.* Он является частным случаем метода (2), когда

$$b_1=b_2=\ldots=b_m=0, \quad b_0=1.$$

Для отыскания y_n получаем из (22) нелинейное уравнение

$$a_0y_n-\tau f(t_n, y_n)=-\sum_{k=1}^{m} a_k y_{n-k}, \tag{23}$$

которое можно решать тем или иным итерационным методом.

Условия p-го порядка аппроксимации (9), (10) из § 3 в случае метода (22) принимают вид

$$a_0=-\sum_{k=1}^{m} a_k, \quad \sum_{k=1}^{m} ka_k=-1, \quad \sum_{k=1}^{m} k^l a_k=0, \quad l=2, 3, \ldots, p. \tag{24}$$

Отсюда видно, что наивысший достижимый порядок аппроксимации чисто неявного m-шагового метода равен m. Упомянутый выше метод Гира использует чисто неявные схемы наивысшего порядка аппроксимации. Система уравнений (24) для определения

коэффициентов $a_1, \ldots, a_m$ метода наивысшего порядка имеет вид

$$\begin{aligned} a_1+2a_2+\ldots+ma_m&=-1,\\ a_1+2^2a_2+\ldots+m^2a_m&=a_0,\\ &\cdots\cdots\cdots\cdots\\ a_1+2^ma_2+\ldots+m^ma_m&=0. \end{aligned} \tag{25}$$

Эта система однозначно разрешима, так как ее определитель отличен от нуля.

При $m=1$ метод (23), (25) совпадает с неявным методом Эйлера. При $m=2$ и $m=3$ получаем методы

$$\frac{3}{2}y_n-2y_{n-1}+\frac{1}{2}y_{n-2}=\tau f(t_n, y_n), \tag{26}$$

$$\frac{11}{6}y_n-3y_{n-1}+\frac{3}{2}y_{n-2}-\frac{1}{3}y_{n-3}=\tau f(t_n, y_n), \tag{27}$$

имеющие, соответственно, второй и третий порядок точности. При $m=4$ из (23), (25) получим схему (21). Для практических расчетов используются аналогичные методы вплоть до десятого порядка точности.

Важно отметить, что чисто неявные разностные методы обладают хорошими свойствами устойчивости, позволяющими использовать их для решения жестких систем уравнений.

Рассмотрим более подробно метод второго порядка (26) и найдем область его устойчивости. Для модельного уравнения (4) метод (26) принимает вид

$$\frac{3}{2}y_n-2y_{n-1}+\frac{1}{2}y_{n-2}=\mu y_n, \tag{28}$$

где $\mu=\tau\lambda$. Ему соответствует характеристическое уравнение

$$\left(\frac{3}{2}-\mu\right)q^2-2q+\frac{1}{2}=0. \tag{29}$$

Нам нужно найти множество точек G комплексной плоскости $\mu=\mu_0+i\mu_1$, для которых оба корня $q_{1,2}(\mu)$ уравнения (29) не превосходят по модулю единицу. Границей области G является множество таких точек μ, для которых $|q|=1$.

Выразим из уравнения (29) параметр μ через переменное q, т. е. запишем

$$\mu=\frac{3}{2}-2q^{-1}+\frac{1}{2}q^{-2}. \tag{30}$$

Отсюда видно, что если $|q|=1$, т. е. $q=e^{-i\varphi}$, то

$$\mu=\frac{3}{2}-2e^{i\varphi}+\frac{1}{2}e^{2i\varphi}. \tag{31}$$

При изменении аргумента φ от 0 до 2π точка μ описывает замкнутую кривую Γ, симметричную относительно действительной оси

(см. рис. 7). Для точек $\mu(q)$, расположенных снаружи от этой кривой, выполнено условие $|q|<1$, поэтому область устойчивости G метода (28) представляет собой внешность кривой Г. Точки, расположенные внутри Г, составляют область неустойчивости. Обозначая $x=\cos\varphi$, можно переписать (31) в виде

$$\mu=(1-x)^2\pm i\sqrt{1-x^2}(2-x),$$

откуда следует, что вся кривая Г расположена в правой полуплоскости. Поэтому область устойчивости метода (26) целиком содержит левую полуплоскость и тем самым метод (26) является A-устойчивым.

Исследуем аналогичным образом область устойчивости метода четвертого порядка (21).

Записывая характеристическое уравнение в виде

$$\mu=\frac{1}{12}(25-48q^{-1}+36q^{-2}-16q^{-3}+3q^{-4})$$

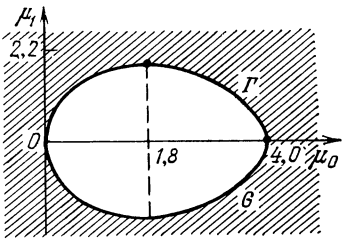


Рис. 7. Граница устойчивости метода (26)

и полагая $q=e^{-i\varphi}$, находим уравнение границы, разделяющей области устойчивости и неустойчивости:

$$\mu=-\frac{2}{3}(1-x)^3(3x+1)\pm i\frac{\sqrt{1-x^2}}{3}(6x^3-16x^2+15x-8),$$

где $x=\cos\varphi$ (см. рис. 8). В отличие от предыдущего примера, имеются точки границы, расположенные в левой полуплоскости. Поэтому метод (21) не является A-устойчивым.

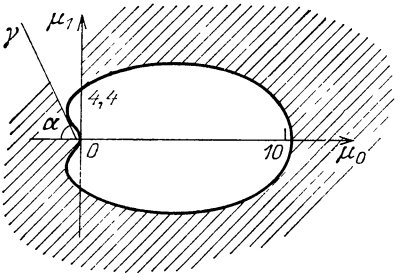


Рис. 8. Граница устойчивости метода (21)

Найдем теперь значение α, при котором метод (21) $A(\alpha)$-устойчив. Для этого достаточно найти угол α, который образует касательная γ (см. рис. 8), проходящая через точку (0, 0), с отрицательным направлением оси μ_0.

Обозначим

$$\mu_0(x)=-\frac{2}{3}(1-x)^3(3x+1),$$

$$\mu_1(x)=\frac{\sqrt{1-x^2}}{3}(6x^3-16x^2+15x-8). \qquad (32)$$

Условие касания γ с границей определяется следующим уравнением относительно параметра x:

$$\frac{\mu_1(x)}{\mu_0(x)}=\frac{\mu_1'(x)}{\mu_0'(x)}. \qquad (33)$$

Из (32) получим

$$\mu_0'(x) = 8x(1-x)^2,$$
$$\mu_1'(x) = \frac{-8x^4+16x^3-4x^2-8x+5}{\sqrt{1-x^2}}. \tag{34}$$

Уравнение (33) после подстановки в него выражений для μ_0, μ_0', μ_1, μ_1' из (32), (34) и приведения подобных членов сводится к линейному уравнению $x-0{,}2=0$. При $x=0{,}2$ из (32) получим

$$\mu_0 = -\frac{4^5}{3\cdot 5^4}, \quad \mu_1 = -\frac{\sqrt{24}\cdot 699}{3\cdot 5^4},$$

так что

$$\operatorname{tg}\alpha = \frac{\mu_1}{\mu_0} = \frac{\sqrt{6}\cdot 699}{512} > \sqrt{6}.$$

Поэтому метод (21) $A(\alpha)$-устойчив, где $\alpha = \operatorname{arctg}\sqrt{6} \approx 68°$.

В заключение отметим, что подробное изложение численных методов решения жестких систем дифференциальных уравнений содержится в книгах [26, 37].

ЧАСТЬ III

РАЗНОСТНЫЕ МЕТОДЫ РЕШЕНИЯ ЗАДАЧ МАТЕМАТИЧЕСКОЙ ФИЗИКИ

Здесь излагаются приближенные методы решения краевых задач для уравнений с частными производными. В основе этих методов лежит сведение дифференциальной задачи к системе линейных алгебраических уравнений путем замены дифференциального оператора разностным.

ГЛАВА 1

ВВОДНЫЕ ПОНЯТИЯ

§ 1. Примеры разностных аппроксимаций

Различные способы приближенной замены одномерных дифференциальных уравнений разностными изучались в § 4 ч. I. Напомним примеры разностных аппроксимаций и введем необходимые обозначения. Будем рассматривать равномерную сетку с шагом h, т. е. множество точек

$$\omega_h=\{x_i=ih,\ i=0,\ \pm 1,\ \pm 2,\ \ldots\}.$$

Пусть $u(x)$ — достаточно гладкая функция, заданная на отрезке $[x_{i-1},\ x_{i+1}]$. Обозначим $u_i=u(x_i)$, $u_{x,i}=\dfrac{u_{i+1}-u_i}{h}$, $u_{\bar{x},i}=\dfrac{u_i-u_{i-1}}{h}$, $u_{\overset{\circ}{x},i}=\dfrac{u_{i+1}-u_{i-1}}{2h}$.

Разностные отношения $u_{x,i}$, $u_{\bar{x},i}$, $u_{\overset{\circ}{x},i}$ называются соответственно правой, левой и центральной разностными производными функции $u(x)$ в точке x_i. Каждое из этих разностных отношений аппроксимирует $u'(x)$ в точке x_i, т. е. при фиксированном x_i и при $h\to 0$ (тем самым при $i\to\infty$) пределом этих отношений является $u'(x_i)$. Проводя разложение по формуле Тейлора, получим

$$u_{x,i}-u'(x_i)=0{,}5hu''(x_i)+O(h^2),$$
$$u_{\bar{x},i}-u'(x_i)=-0{,}5hu''(x_i)+O(h^2),$$
$$u_{\overset{\circ}{x},i}-u'(x_i)=O(h^2).$$

Отсюда видно, что левая и правая разностные производные аппроксимируют $u'(x)$ с первым порядком по h, а центральная раз-

ностная производная — со вторым порядком. Нетрудно показать, что вторая разностная производная

$$u_{\bar{x}x,i} = \frac{u_{x,i} - u_{\bar{x},i}}{h} = \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2}$$

аппроксимирует $u''(x_i)$ со вторым порядком по h, причем справедливо разложение

$$u_{\bar{x}x,i} - u''(x_i) = \frac{h^2}{12} u^{\mathrm{IV}}(x_i) + O(h^4).$$

Рассмотрим дифференциальное выражение

$$Lu = \frac{d}{dx}\left(k(x)\frac{du}{dx}\right) \tag{1}$$

с переменным коэффициентом $k(x)$. Заменим выражение (1) разностным отношением

$$L_h u = (au_{\bar{x}})_{x,i} = \frac{1}{h}\left(a_{i+1}\frac{u_{i+1} - u_i}{h} - a_i \frac{u_i - u_{i-1}}{h}\right), \tag{2}$$

где $a = a(x)$ — функция, определенная на сетке ω_h. Найдем условия, которым должна удовлетворять функция $a(x)$ для того, чтобы отношение $(au_{\bar{x}})_{x,i}$ аппроксимировало $(ku')'$ в точке x_i со вторым порядком по h. Подставляя в (2) разложения

$$u_{x,i} = u_i' + \frac{h}{2}u_i'' + \frac{h^2}{6}u_i''' + O(h^3),$$
$$u_{\bar{x},i} = u_i' - \frac{h}{2}u_i'' + \frac{h^2}{6}u_i''' + O(h^3),$$

где $u_i' = u'(x_i)$, получим

$$L_h u = \frac{a_{i+1} - a_i}{h}u_i' + \frac{a_{i+1} + a_i}{2}u_i'' + \frac{h(a_{i+1} - a_i)}{6}u_i''' + O(h^2).$$

С другой стороны,

$$Lu = (ku')' = ku'' + k'u',$$

т. е.

$$L_h u - Lu(x_i) = \left(\frac{a_{i+1} - a_i}{h} - k_i'\right)u_i' + \left(\frac{a_{i+1} + a_i}{2} - k_i\right)u_i'' + \frac{h(a_{i+1} - a_i)}{6}u_i''' + O(h^2).$$

Отсюда видно, что $L_h u - Lu = O(h^2)$, если выполнены условия

$$\frac{a_{i+1} - a_i}{h} = k'(x_i) + O(h^2), \quad \frac{a_{i+1} + a_i}{2} = k_i + O(h^2). \tag{3}$$

Условия (3) называются *достаточными условиями второго порядка аппроксимации*. При их выводе предполагалось, что функция $u(x)$ имеет непрерывную четвертую производную и $k(x)$ — диффе-

ренцируемая функция. Нетрудно показать, что условиям (3) удовлетворяют, например, следующие функции:

$$a_i = 0{,}5\,(k(x_i) + k(x_{i-1})), \quad a_i = k(x_i - 0{,}5h), \quad a_i = \sqrt{k(x_i)\,k(x_{i-1})}.$$

Заметим, что если положить $a_i = k(x_i)$, то получим только первый порядок аппроксимации. В § 2 будет рассмотрен регулярный метод получения разностных аппроксимаций вида (2).

В качестве следующего примера рассмотрим разностную аппроксимацию оператора Лапласа

$$Lu = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}. \qquad (4)$$

Введем на плоскости (x_1, x_2) прямоугольную сетку с шагом h_1 по направлению x_1 и с шагом h_2 по направлению x_2, т. е. множество точек

$$\omega_h = \{(x_1^i, x_2^j) \mid x_1^i = ih_1,\ x_2^j = jh_2;\ i, j = 0, \pm 1, \pm 2, \ldots\}$$

(см. рис. 9), и обозначим

$$u_{\bar{x}_1 x_1, ij} = \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{h_1^2},$$

$$u_{\bar{x}_2 x_2, ij} = \frac{u_{i,j+1} - 2u_{ij} + u_{i,j-1}}{h_2^2}.$$

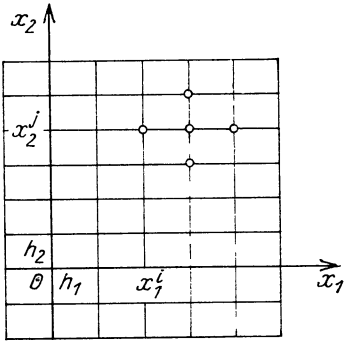


Рис. 9. Сетка ω_h и пятиточечный шаблон

Из предыдущих рассуждений следует, что разностное выражение

$$L_h u_{ij} = u_{\bar{x}_1 x_1, ij} + u_{\bar{x}_2 x_2, ij} \qquad (5)$$

аппроксимирует дифференциальное выражение (4) со вторым порядком, т. е. $L_h u_{ij} - Lu(x_1^i, x_2^j) = O(h_1^2) + O(h_2^2)$. Более того, для функций $u(x_1, x_2)$, обладающих непрерывными шестыми производными, справедливо разложение

$$L_h u_{ij} - Lu(x_1^i, x_2^j) = \frac{h_1^2}{12}\frac{\partial^4 u(x_1^i, x_2^j)}{\partial x_1^4} + \frac{h_2^2}{12}\frac{\partial^4 u(x_1^i, x_2^j)}{\partial x_2^4} + O(h_1^4 + h_2^4).$$

Разностное выражение (5) называется *пятиточечным разностным оператором Лапласа*, так как оно содержит значения функции $u(x_1, x_2)$ в пяти точках сетки, а именно в точках (x_1^i, x_2^j), $(x_1^{i\pm1}, x_2^j)$, $(x_1^i, x_2^{j\pm1})$ (см. рис. 9). Указанное множество точек называется *шаблоном* разностного оператора. Возможны разностные аппроксимации оператора Лапласа и на шаблонах, содержащих большее число точек.

§ 2. Построение разностных схем интегро-интерполяционным методом

1. Построение разностной схемы. Задачи математической физики формулируются в виде основного дифференциального уравнения и дополнительных (начальных, граничных) условий, обеспечивающих существование и единственность решения. Типичными примерами являются: задача Дирихле для уравнения Лапласа, задача Коши и смешанные задачи для уравнений параболического и гиперболического типов. Под *разностной схемой* понимают совокупность разностных уравнений, аппроксимирующих основное уравнение и дополнительные условия исходной дифференциальной задачи. Существуют различные способы построения разностных схем. В этом параграфе будет рассмотрен один из способов, носящий название *интегро-интерполяционного метода* (или *метода баланса*) построения разностных схем.

В качестве примера рассмотрим применение интегро-интерполяционного метода к построению разностной схемы следующей краевой задачи для обыкновенного дифференциального уравнения второго порядка:

$$(k(x)u'(x))'-q(x)u(x)+f(x)=0, \quad 0<x<l, \tag{1}$$

$$-k(0)u'(0)+\beta u(0)=\mu_1, \quad u(l)=\mu_2, \tag{2}$$

где $k(x)$, $q(x)$, $f(x)$ — заданные достаточно гладкие функции, удовлетворяющие условиям $k(x)\geqslant k_*>0$, $q(x)\geqslant 0$, и $\beta\geqslant 0$, μ_1, μ_2 — заданные числа. При сформулированных предположениях существует и единственно решение $u(x)$ задачи (1), (2). Будем считать, что это решение является достаточно гладким.

Уравнение (1) можно трактовать как уравнение установившегося распределения температуры $u(x)$ в стержне длины l, на одном конце которого ($x=l$) поддерживается заданная температура μ_2, а на другом ($x=0$) происходит теплообмен с окружающей средой по закону Ньютона (см. [41]). При этом $k(x)$ — коэффициент теплопроводности, $-k(x)u'(x)$ — тепловой поток, коэффициенты $q(x)$, $f(x)$ характеризуют плотность тепловых источников.

Для построения разностной схемы введем прежде всего на отрезке $[0, l]$ равномерную сетку с шагом h, т. е. множество точек

$$\omega_h=\{x_i=ih,\ i=0, 1, \ldots, N,\ hN=l\}.$$

Обозначим $x_{i\pm 1/2}=x_i\pm 0{,}5h$, $w(x)=k(x)u'(x)$, $w_{i\pm 1/2}=w(x_{i\pm 1/2})$ и проинтегрируем уравнение (1) на отрезке $[x_{i-1/2}, x_{i+1/2}]$. Тогда получим уравнение

$$w_{i+1/2}-w_{i-1/2}-\int\limits_{x_{i-1/2}}^{x_{i+1/2}} q(x)u(x)\,dx+\int\limits_{x_{i-1/2}}^{x_{i+1/2}} f(x)\,dx=0, \tag{3}$$

которое представляет собой уравнение баланса тепла на отрезке

$[x_{i-1/2}, x_{i+1/2}]$. Далее, заменим интеграл

$$\int_{x_{i-1/2}}^{x_{i+1/2}} q(x)\,u(x)\,dx$$

его приближенным значением $u_i \int_{x_{i-1/2}}^{x_{i+1/2}} q(x)\,dx$ и введем обозначения

$$d_i = \frac{1}{h}\int_{x_{i-1/2}}^{x_{i+1/2}} q(x)\,dx, \quad \varphi_i = \frac{1}{h}\int_{x_{i-1/2}}^{x_{i+1/2}} f(x)\,dx. \tag{4}$$

В результате вместо уравнения (3) получим уравнение

$$\frac{w_{i+1/2} - w_{i-1/2}}{h} - d_i u_i + \varphi_i = 0. \tag{5}$$

Выразим теперь $w_{i\pm 1/2}$ через значения функции $u(x)$ в точках сетки. Для этого проинтегрируем соотношение $u'(x) = w(x)/k(x)$ на отрезке $[x_{i-1}, x_i]$. Тогда получим

$$u_i - u_{i-1} = \int_{x_{i-1}}^{x_i} \frac{w(x)}{k(x)}\,dx \approx w_{i-1/2}\int_{x_{i-1}}^{x_i} \frac{dx}{k(x)}.$$

Обозначая

$$a_i = \left(\frac{1}{h}\int_{x_{i-1}}^{x_i} \frac{dx}{k(x)}\right)^{-1}, \tag{6}$$

получим $w_{i-1/2} \approx a_i \dfrac{u_i - u_{i-1}}{h} = a_i u_{\overline{x},i}$, $w_{i+1/2} = a_{i+1} u_{x,i}$.

Подставляя эти выражения в уравнение (5), получим разностное уравнение, содержащее значения искомой функции в точках x_i, $x_{i\pm 1}$:

$$\frac{1}{h}\left(a_{i+1}u_{x,i} - a_i u_{\overline{x},i}\right) - d_i u_i + \varphi_i = 0$$

или в сокращенной записи

$$(a u_{\overline{x}})_{x,i} - d_i u_i + \varphi_i = 0. \tag{7}$$

Это уравнение по своему построению является разностным аналогом основного дифференциального уравнения (1). Записывая уравнение (7) во всех точках сетки, в которых оно определено, т. е. при $i = 1, 2, \ldots, N-1$, получим систему из $N-1$ линейных алгебраических уравнений относительно $N+1$ неизвестных $u_0, u_1, \ldots, u_N$. Два недостающих уравнения получаются путем аппроксимации граничных условий (2). Одним из этих уравнений является условие $u_N =$

$=\mu_2$, а второе может быть получено интегро-интерполяционным методом. Для этого проинтегрируем основное уравнение (1) на отрезке $[0, x_{1/2}]$, где $x_{1/2}=0{,}5h$:

$$w_{1/2} - w_0 - \int_0^{x_{1/2}} q(x)\,u(x)\,dx + \int_0^{x_{1/2}} f(x)\,dx = 0. \tag{8}$$

Полагая, как и ранее, $w_{i-1/2} \approx a_i u_{\overline{x},i}$, получим при $i=1$, что $w_{1/2} = a_1 u_{\overline{x},1}$. Выражение для w_0 следует из граничного условия при $x=0$: $w_0 = -\mu_1 + \beta u_0$. Наконец, полагая

$$\int_0^{x_{1/2}} q(x)\,u(x)\,dx \approx u_0 \int_0^{x_{1/2}} q(x)\,dx,$$

получим из тождества (8) разностное уравнение

$$a_1 u_{\overline{x},1} - \beta u_0 + \mu_1 - u_0 \int_0^{x_{1/2}} q(x)\,dx + \int_0^{x_{1/2}} f(x)\,dx = 0. \tag{9}$$

Обозначая

$$d_0 = \frac{1}{0{,}5h} \int_0^{x_{1/2}} q(x)\,dx, \quad \varphi_0 = \frac{1}{0{,}5h} \int_0^{x_{1/2}} f(x)\,dx,$$

перепишем уравнение (9) в виде

$$-a_1 u_{x,0} + (\beta + 0{,}5hd_0)\,u_0 = \mu_1 + 0{,}5h\varphi_0.$$

Из этой записи видно, что полученное уравнение является разностным аналогом граничного условия $-k(0)u'(0) + \beta u(0) = \mu_1$.

В дальнейшем решение разностной задачи в отличие от решения дифференциальной задачи будем обозначать буквой y, так что $y_i = y(x_i)$, $x_i \in \omega_h$. Объединяя все уравнения (7), (9), получаем следующую разностную схему для задачи (1), (2):

$$\begin{gathered} (ay_{\overline{x}})_{x,i} - d_i y_i + \varphi_i = 0, \quad i = 1, 2, \ldots, N-1, \\ -a_1 y_{x,0} + (\beta + 0{,}5hd_0)\,y_0 = \mu_1 + 0{,}5h\varphi_0, \quad y_N = \mu_2. \end{gathered} \tag{10}$$

При анализе разностной схемы (10), как впрочем и любой другой разностной схемы, возникают следующие вопросы: а) существование и единственность решения системы линейных алгебраических уравнений (10); б) каким методом надо отыскивать это решение; в) какое отношение имеет система разностных уравнений (10) к исходной задаче (1), (2), иначе говоря, переходит ли разностное уравнение (10) в уравнение (1), если шаг сетки h стремится к нулю? Это вопрос об аппроксимации дифференциальной задачи (1), (2) разностной схемой (10); г) сходится ли решение $y(x)$ разностной задачи к решению $u(x)$ дифференциальной задачи при $h \to 0$?

На первые два вопроса можно ответить немедленно. Разностная задача (10) является типичным примером задачи, которая решает-

ся методом прогонки, изложенным в п. 7 § 4 ч. I. Систему уравнений (10) можно записать в виде

$$A_i y_{i-1} - C_i y_i + B_i y_{i+1} = -F_i, \quad i = 1, 2, \ldots, N-1,$$

$$y_0 = \varkappa_1 y_1 + \widetilde{\mu}_1, \quad y_N = \varkappa_2 y_2 + \mu_2,$$

где

$$A_i = a_i,\ B_i = a_{i+1},\ C_i = a_i + a_{i+1} + h^2 d_i,\ F_i = h^2 \varphi_i,\ \varkappa_2 = 0,$$

$$\varkappa_1 = \frac{1}{1 + h a_1^{-1} (\beta + 0{,}5 h d_0)}, \quad \widetilde{\mu}_1 = \frac{h (\mu_1 + 0{,}5 h \varphi_0)}{a_1}.$$

Из условий $a_i > 0$, $\beta \geqslant 0$, $d_i \geqslant 0$ следует, что $C_i \geqslant A_i + B_i > 0$, т. е. выполнены условия устойчивости прогонки. Поэтому разностная задача (10) однозначно разрешима и ее можно решать методом прогонки по формулам (44)—(46) из § 4 ч. I.

Вопросы аппроксимации и сходимости обсуждаются в следующем параграфе.

§ 3. Исследование аппроксимации и сходимости

1. Аппроксимация дифференциального уравнения. В § 2 рассматривалась краевая задача

$$(k(x) u'(x))' - q(x) u(x) + f(x) = 0, \quad 0 < x < l, \tag{1}$$

$$-k(0) u'(0) + \beta u(0) = \mu_1, \quad u(l) = \mu_2, \tag{2}$$

$$k(x) \geqslant c_1 > 0, \quad \beta \geqslant 0,$$

для которой интегро-интерполяционным методом была построена разностная схема

$$(a y_{\bar{x}})_{x,i} - d_i y_i + \varphi_i = 0, \quad i = 1, 2, \ldots, N-1, \tag{3}$$

$$-a_1 y_{x,0} + \widetilde{\beta} y_0 = \widetilde{\mu}_1, \quad y_N = \mu_2, \tag{4}$$

где

$$a_i = \left(\frac{1}{h} \int_{x_{i-1}}^{x_i} \frac{dx}{k(x)} \right)^{-1}, \quad i = 1, 2, \ldots, N, \tag{5}$$

$$d_i = \frac{1}{h} \int_{x_{i-1/2}}^{x_{i+1/2}} q(x)\, dx, \quad \varphi_i = \frac{1}{h} \int_{x_{i-1/2}}^{x_{i+1/2}} f(x)\, dx, \quad i = 1, 2, \ldots, N-1, \tag{6}$$

$$\widetilde{\beta} = \beta + 0{,}5 h d_0, \quad \widetilde{\mu}_1 = \mu_1 + 0{,}5 h \varphi_0,$$

$$d_0 = \frac{1}{0{,}5h} \int_0^{0{,}5h} q(x)\, dx, \quad \varphi_0 = \frac{1}{0{,}5h} \int_0^{0{,}5h} f(x)\, dx.$$

Обозначим через $Lu(x)$ левую часть уравнения (1) и через $L_h y_i$ — левую часть уравнения (3), т. е.

$$Lu(x) = (k(x) u')' - q(x) u(x) + f(x), \qquad L_h y_i = (a y_{\bar{x}})_{x,i} - d_i y_i + \varphi_i.$$

Пусть $v(x)$ — достаточно гладкая функция и $v(x_i)$ — ее значение в точке x_i сетки

$$\omega_h=\{x_i=ih,\ i=0,\ 1,\ \ldots,\ N,\ hN=l\}. \tag{7}$$

Говорят, что *разностный оператор L_h аппроксимирует дифференциальный оператор L в точке $x=x_i$*, если разность $L_hv_i-Lv(x_i)$ стремится к нулю при $h\to 0$. В этом случае говорят также, что разностное уравнение (3) аппроксимирует дифференциальное уравнение (1).

Чтобы установить наличие аппроксимации, достаточно разложить по формуле Тейлора в точке $x=x_i$ значения $v_{i\pm1}=v(x_i\pm h)$, входящие в разностное выражение L_hv_i. Большая часть этой работы проделана в § 1, где показано, что при условиях

$$\frac{a_{i+1}-a_i}{h}=k'(x_i)+O(h^2),\quad \frac{a_{i+1}+a_i}{2}=k(x_i)+O(h^2) \tag{8}$$

выполняется соотношение

$$(av_{\bar{x}})_{x,i}-(k(x)v'(x))'\,|_{x=x_i}=O(h^2).$$

Если кроме того, докажем, что

$$d_i=q(x_i)+O(h^2),\quad \varphi_i=f(x_i)+O(h^2), \tag{9}$$

то тем самым будет установлено, что оператор L_h аппроксимирует L со вторым порядком по h, т. е.

$$L_hv_i-Lv(x_i)=O(h^2),\qquad i=1,\ 2,\ \ldots,\ N-1. \tag{10}$$

Итак, доказательство второго порядка аппроксимации сводится к проверке условий (8), (9) для коэффициентов (5), (6). Проверим сначала выполнение условий (8). Обозначая $p(x)=k^{-1}(x)$, получим

$$\frac{1}{a_i}=\frac{1}{h}\int\limits_{x_{i-1}}^{x_i}p(x)\,dx=p_{i-1/2}+\frac{h^2}{12}p''_{i-1/2}+O(h^4),$$

следовательно,

$$a_i=k_{i-1/2}-\frac{h^2}{12}\frac{p''_{i-1/2}}{p^2_{i-1/2}}+O(h^4)=k_{i-1/2}-\frac{h^2}{12}\frac{p''_i}{p_i^2}+O(h^3).$$

Аналогично

$$a_{i+1}=k_{i+1/2}-\frac{h^2}{12}\frac{p''_i}{p_i^2}+O(h^3).$$

Отсюда получим

$$\frac{a_{i+1}+a_i}{2}=k(x_i)+O(h^2),$$

$$\frac{a_{i+1}-a_i}{h}=\frac{k_{i+1/2}-k_{i-1/2}}{h}+O(h^2)=k'(x_i)+O(h^2),$$

т. е. условия (8) выполнены. Условия (9) выполнены в силу того,

что замена интегралов (6) значениями q_i, f_i соответствует приближенному вычислению этих интегралов по формуле прямоугольников с узлом в середине отрезка интегрирования.

2. Аппроксимация граничного условия. Исследуем погрешность аппроксимации разностного граничного условия (4). Обозначим $l_h v(0) = -a_1 v_{x,0} + \widetilde{\beta} v_0$. Если $v(x)$ — произвольная достаточно гладкая функция, то очевидно

$$l_h v(0) = -k(0) v'(0) + \beta v(0) + O(h),$$

т. е. имеет место аппроксимация первого порядка по h. Однако если $v = u(x)$ — решение задачи (1), (2), то разностное граничное условие (4) имеет второй порядок аппроксимации, т. е.

$$-a_1 u_{x,0} + \widetilde{\beta} u_0 - \widetilde{\mu}_1 = -k(0) u'(0) + \beta u(0) - \mu_1 + O(h^2).$$

Докажем последнее утверждение. Используя разложения

$$u_{x,0} = (u_1 - u_0)/h = u'(x_{1/2}) + O(h^2), \qquad x_{1/2} = 0{,}5h,$$
$$a_1 = k_{1/2} + O(h^2),$$

получим

$$a_1 u_{x,0} = w_{1/2} + O(h^2) = w_0 + 0{,}5h w_0' + O(h^2) =$$
$$= k(0) u'(0) + 0{,}5h (ku')'(0) + O(h^2).$$

Отсюда имеем

$$l_h u(0) = -k(0) u'(0) - 0{,}5h (ku')'(0) + \beta u_0 - \widetilde{\mu}_1 + O(h^2) =$$
$$= [-k(0) u'(0) + \beta u(0) - \mu_1] + 0{,}5h [-(ku')'(0) + d_0 u_0 - \varphi_0] + O(h^2)$$

Учитывая граничное условие (2), получаем

$$l_h u(0) = 0{,}5h [-(ku')'(0) + d_0 u_0 - \varphi_0] + O(h^2).$$

Выражение, стоящее в квадратных скобках, преобразуем, учитывая уравнение (1), к виду

$$-(ku')'(0) + d_0 u_0 - \varphi_0 = -(ku')'(0) + q(0) u(0) - f(0) +$$
$$+ (d_0 - q(0)) u_0 - (f(0) - \varphi_0) = (d_0 - q(0)) u_0 - (f(0) - \varphi_0).$$

Из соотношений

$$d_0 = \frac{1}{0{,}5h} \int_0^{x_{1/2}} q(x)\, dx = q(0) + O(h), \qquad \varphi_0 = f(0) + O(h)$$

получаем

$$l_h u(0) = -a_1 u_{x,0} + \widetilde{\beta} u_0 - \widetilde{\mu}_1 = O(h^2),$$

что и требовалось доказать.

Таким образом, при достаточной гладкости коэффициентов $k(x)$, $q(x)$, $f(x)$ и решения $u(x)$ разностная схема (10) аппроксимирует исходную задачу (2) со вторым порядком по h.

При практическом использовании разностной схемы для нахождения ее коэффициентов не обязательно вычислять интегралы (4), (6) точно. Можно воспользоваться коэффициентами, полученными путем замены этих интегралов квадратурными формулами, имеющими точность $O(h^2)$ и выше. Например, в результате применения формулы прямоугольников получим следующие коэффициенты: $a_i = k(x_i - 0{,}5h)$, $d_i = q(x_i)$, $\varphi_i = f(x_i)$.

Применяя формулу трапеций, получим

$$a_i = \frac{2k_i k_{i-1}}{k_i + k_{i-1}}, \quad d_i = \frac{q_{i-1/2} + q_{i+1/2}}{2}, \quad \varphi_i = \frac{f_{i-1/2} + f_{i+1/2}}{2}.$$

Представление коэффициентов разностной схемы в виде интегралов (4), (6) оказывается полезным при исследовании сходимости в случае разрывных функций $k(x)$, $q(x)$, $f(x)$ (см. [32]).

3. Уравнение для погрешности. Решение $y_i = y(x_i)$ разностной задачи (3), (4) зависит от шага h сетки, $y(x_i) = y_h(x_i)$. По существу, мы имеем семейство решений $\{y_h(x_i)\}$, зависящее от параметра h. Говорят, что решение $y_h(x)$ разностной задачи *сходится* к решению $u(x)$ исходной дифференциальной задачи, если при $h \to 0$ погрешность $y_h(x_i) - u(x_i)$, $i = 0, 1, \ldots, N$, стремится к нулю в некоторой норме. В настоящем параграфе в качестве такой нормы будем брать норму в сеточном пространстве $C(\omega_h)$, т. е. положим $\|y_h - u\|_{C(\omega_h)} = \max\limits_{x_i \in \omega_h} |y_h(x_i) - u(x_i)|$. Говорят, что разностная схема *имеет m-й порядок точности* (или *сходится с порядком m*), если

$$\|y_h - u\|_{C(\omega_h)} \leqslant Mh^m,$$

где $m > 0$, $M > 0$ — константы, не зависящие от h.

Выше было установлено, что схема (3), (4) имеет второй порядок аппроксимации. Докажем теперь, что эта схема имеет и второй порядок точности. Для этого прежде всего выпишем уравнение, которому удовлетворяет погрешность $z_i = y_i - u(x_i)$. Подставим $y_i = z_i + u(x_i)$ в уравнения (3), (4). Тогда получим уравнения

$$(az_{\bar{x}})_{x,i} - d_i z_i = -\psi_i, \quad i = 1, 2, \ldots, N-1, \tag{11}$$

$$-a_1 z_{x,0} + \tilde{\beta} z_0 = \nu_1, \quad z_N = 0, \tag{12}$$

где обозначено

$$\psi_i = -(au_{\bar{x}})_{x,i} + d_i u_i + \varphi_i, \quad \nu_1 = a_1 u_{x,0} - \tilde{\beta} u_0 + \tilde{\mu}_1.$$

Функция ψ_i, входящая в правую часть уравнения (11), называется *погрешностью аппроксимации дифференциального уравнения* (1) *разностным уравнением* (3) *на решении задачи* (1), (2). В п. 1 было показано, что $\psi_i = O(h^2)$ при $h \to 0$, $i = 1, 2, \ldots, N-1$. Аналогично, величина ν_1 является по определению *погрешностью аппроксимации краевого условия* (2) *разностным краевым условием* (4) *на решении задачи* (1), (2), причем $\nu_1 = O(h^2)$. Таким образом, структура уравнений для погрешности (11), (12) та же, что и у разностной схемы (3), (4), отличаются только правые части.

Чтобы доказать сходимость разностной схемы, оценим решение задачи (11), (12) через правые части ψ_i, ν_1, т. е. получим неравенство вида

$$\|z\|_{C(\omega_h)} \leqslant M_1(\|\psi\|_{C(\omega_h)} + |\nu_1|) \tag{13}$$

с константой M_1, не зависящей от h. Из этого неравенства и будет следовать, что $\|z\|_{C(\omega_h)} = O(h^2)$.

Отметим, что неравенства вида (13), называемые априорными оценками, нашли широкое применение в теории разностных схем. Поскольку структура уравнений для погрешности (11), (12) та же, что и у разностной схемы (3), (4), а отличаются только правые части, то оценка (13) выполняется одновременно с аналогичной оценкой

$$\|y\|_{C(\omega_h)} \leqslant M_1(\|\varphi\|_{C(\omega_h)} + |\tilde{\mu}_1|)$$

для разностной схемы (3), (4) при $\mu_2 = 0$. Последняя оценка выражает устойчивость решения разностной задачи по правым частям φ и $\tilde{\mu}_1$.

4. Разностные тождества и неравенства. Для того чтобы доказать неравенство (13), нам потребуются некоторые разностные тождества и неравенства. Будем рассматривать сеточные функции, заданные на сетке (7). Обозначим $y_i = y(x_i)$, $x_i \in \omega_h$, $y_{x,i} = (y_{i+1} - y_i)/h$, $y_{\bar{x},i} = (y_i - y_{i-1})/h$, $(y, v) = \sum_{i=1}^{N-1} y_i v_i h$, $(y, v] = \sum_{i=1}^{N} y_i v_i h$.

Справедливо следующее разностное тождество:

$$(y, v_x) = -(v, y_{\bar{x}}] + y_N v_N - y_0 v_1. \tag{14}$$

Действительно,

$$\begin{aligned}(y, v_x) = \sum_{i=1}^{N-1} y_i v_{x,i} h = \sum_{i=1}^{N-1} y_i (v_{i+1} - v_i) = \sum_{i=2}^{N} y_{i-1} v_i - \sum_{i=1}^{N-1} y_i v_i = \\ = \sum_{i=1}^{N} y_{i-1} v_i - y_0 v_1 - \sum_{i=1}^{N} y_i v_i + y_N v_N = \\ = -\sum_{i=1}^{N} v_i (y_i - y_{i-1}) + y_N v_N - y_0 v_1,\end{aligned}$$

что и требовалось доказать. Тождество (14) называется *формулой суммирования по частям.*

Подставляя в (14) вместо v выражение $az_{\bar{x}}$ и вместо y функцию z, получаем *первую разностную формулу Грина*

$$(z, (az_{\bar{x}})_x) = -(a, (z_{\bar{x}})^2] + a_N z_{\bar{x},N} z_N - a_1 z_{x,0} z_0. \tag{15}$$

Здесь $(a, (z_{\bar{x}})^2] = \sum_{i=1}^{N} a_i (z_{\bar{x},i})^2 h$. В частности, если $z_N = 0$ (как в

задаче (11), (12)), то получим

$$(z, (az_{\bar{x}})_x) = -(a, z_{\bar{x}}^2] - a_1 z_{x,0} z_0. \tag{16}$$

Обозначим

$$\|z_{\bar{x}}]\|^2 = \sum_{i=1}^{N} (z_{\bar{x},i})^2 h$$

и докажем, что *для любой сеточной функции* z_i, *удовлетворяющей условию* $z_N = 0$, *справедливо неравенство*

$$\|z\|^2_{C(\omega_h)} \leqslant l \|z_{\bar{x}}]\|^2. \tag{17}$$

Для доказательства воспользуемся тождеством

$$z_i = -\sum_{j=i+1}^{N} h z_{\bar{x},j} = -\sum_{j=i+1}^{N} \sqrt{h}\,(\sqrt{h} z_{\bar{x},j}), \qquad i = 0, 1, \ldots, N-1,$$

и применим неравенство Коши — Буняковского

$$\left| \sum_{j=i+1}^{N} a_j b_j \right|^2 \leqslant \left(\sum_{j=i+1}^{N} a_j^2 \right) \left(\sum_{j=i+1}^{N} b_j^2 \right).$$

Тогда получим

$$|z_i|^2 \leqslant \left(\sum_{j=i+1}^{N} h \right) \left(\sum_{j=i+1}^{N} h z_{\bar{x},j}^2 \right) = (l - x_i) \sum_{j=i+1}^{N} h z_{\bar{x},j}^2 \leqslant l \sum_{j=1}^{N} h z_{\bar{x},j}^2,$$

откуда сразу следует неравенство (17).

5. Доказательство сходимости. Возвращаясь к доказательству сходимости схемы (3), (4), получим тождество, которому удовлетворяет погрешность $z_i = y_i - u(x_i)$. Для этого умножим уравнение (11) на hz_i и просуммируем по i от 1 до $N-1$. Тогда получим

$$((az_{\bar{x}})_x, z) - (d, z^2) = -(\psi, z)$$

Отсюда, применяя разностную формулу Грина (16), получим

$$(a, z_{\bar{x}}^2] + a_1 z_{x,0} z_0 + (d, z^2) = (\psi, z).$$

Далее, согласно (12) имеем

$$a_1 z_{x,0} z_0 = \widetilde{\beta} z_0^2 - \nu_1 z_0,$$

следовательно, справедливо тождество

$$(a, z_{\bar{x}}^2] + \widetilde{\beta} z_0^2 + (d, z^2) = (\psi, z) + \nu_1 z_0. \tag{18}$$

Из этого тождества и будет сейчас выведено требуемое неравенство вида (13).

Заметим прежде всего, что если

$$k(x) \geqslant c_1 > 0, \qquad \beta \geqslant 0, \qquad q(x) \geqslant 0,$$

то коэффициенты разностной схемы (3), (4) удовлетворяют неравенствам

$$a_i \geqslant c_1 > 0, \quad \widetilde{\beta} \geqslant 0, \quad d_i \geqslant 0. \tag{19}$$

Это утверждение сразу следует из явного представления коэффициентов (5), (6).

Воспользовавшись (19), оценим слагаемые, входящие в левую часть тождества (18), следующим образом:

$$(a, z_{\bar{x}}^2] = \sum_{i=1}^{N} a_i z_{\bar{x},i}^2 h \geqslant c_1 \sum_{i=1}^{N} z_{\bar{x},i}^2 h = c_1 \| z_{\bar{x}}] \|^2,$$

$$\widetilde{\beta} z_0^2 \geqslant 0, \quad (d, z^2) \geqslant 0.$$

Тогда придем к неравенству

$$c_1 \| z_{\bar{x}}] \|^2 \leqslant |(\psi, z)| + |\nu_1| |z_0|. \tag{20}$$

Оценим сверху правую часть этого неравенства. Будем иметь

$$(\psi, z)| + |\nu_1| |z_0| \leqslant \sum_{i=1}^{N-1} |\psi_i| |z_i| h + |\nu_1| |z_0| \leqslant$$

$$\leqslant \| z \|_{C(\omega_h)} \left(\sum_{i=1}^{N-1} |\psi_i| h + |\nu_1| \right).$$

Подставляя эту оценку в (20) и учитывая неравенство (17), получим

$$\frac{c_1}{l} \| z \|_{C(\omega_h)}^2 \leqslant \left(\sum_{i=1}^{N-1} |\psi_i| h + |\nu_1| \right) \| z \|_{C(\omega_h)},$$

т. е.

$$\| z \|_{C(\omega_h)} \leqslant \frac{l}{c_1} \left(\sum_{i=1}^{N-1} |\psi_i| h + |\nu_1| \right).$$

Окончательно

$$\| z \|_{C(\omega_h)} \leqslant \frac{l}{c_1} (l \| \psi \|_{C(\omega_h)} + |\nu_1|). \tag{21}$$

Поскольку $\| \psi \|_{C(\omega_h)} = O(h^2)$, $|\nu_1| = O(h^2)$, из неравенства (21) следует, что погрешность $z_i = y_i - u(x_i)$ также является величиной $O(h^2)$ при $h \to 0$. Итак, справедливо следующее утверждение.

Пусть $k(x)$ — непрерывно дифференцируемая и $q(x)$, $f(x)$ — непрерывные функции при $x \in [0, l]$, решение $u(x)$ задачи (1), (2) *обладает непрерывными четвертыми производными. Пусть коэффициенты разностной схемы* (3), (4) *удовлетворяют условиям* (8), (9), (19). *Тогда решение разностной задачи* (3), (4) *сходится при $h \to 0$ к решению исходной дифференциальной задачи* (1), (2) *со вторым*

порядком по h, так что выполняется оценка

$$\|y-u\|_{C(\omega_h)} \leqslant Mh^2,$$

где M — постоянная, не зависящая от h.

Замечания. 1. Из доказательства видно, что конкретный вид коэффициентов (5), (6) не влияет на справедливость высказанного утверждения, важно лишь выполнение условий (8), (9), (19). 2. Можно ослабить требования на гладкость коэффициентов $k(x)$, $q(x)$, $f(x)$ и решения $u(x)$, однако при этом априорные оценки вида (21) становятся бесполезными, так как норма $\|\psi\|_{C(\omega_h)}$ может и не стремиться к нулю. Доказательство сходимости в классе разрывных коэффициентов и в случае неравномерных сеток, основанное на оценках погрешности $y_i-u(x_i)$ через слабые нормы погрешности аппроксимации ψ_i, например нормы вида $\|\psi\|=\sum\limits_{i=1}^{N-1} h\left|\sum\limits_{j=1}^{i} h\psi_j\right|$, можно найти в [32].

§ 4. Разностные схемы для уравнения теплопроводности

1. Исходная задача. Будем рассматривать следующую первую краевую задачу для уравнения теплопроводности с постоянными коэффициентами. В области $\{0<x<1,\ 0<t\leqslant T\}$ требуется найти решение уравнения

$$\frac{\partial u}{\partial t}=\frac{\partial^2 u}{\partial x^2}+f(x,t), \tag{1}$$

удовлетворяющее начальному условию

$$u(x,\ 0)=u_0(x) \tag{2}$$

и граничным условиям

$$u(0,t)=\mu_1(t),\quad u(1,t)=\mu_2(t). \tag{3}$$

Здесь $u_0(x)$, $\mu_1(t)$, $\mu_2(t)$ — заданные функции. Известно (см. [41]), что при определенных предположениях гладкости решение задачи (1)—(3) существует и единственно. В дальнейшем при исследовании аппроксимации разностных схем будем предполагать, что решение $u(x,t)$ обладает необходимым по ходу изложения числом производных по x и по t. Решение задачи (1)—(3) удовлетворяет принципу максимума и тем самым непрерывно зависит от начальных и граничных данных.

2. Явная схема. Как всегда, для построения разностной схемы надо прежде всего ввести сетку в области изменения независимых переменных и задать шаблон, т. е. множество точек сетки, участвующих в аппроксимации дифференциального выражения. Введем сетку по переменному x такую же, как и в § 3, т. е.

$$\omega_h=\{x_i=ih,\ i=0,\ 1,\ \ldots,\ N,\ hN=1\},$$

и сетку по переменному t с шагом τ, которую обозначим

$$\omega_\tau=\{t_n=n\tau,\ n=0,\ 1,\ \ldots,\ K,\ K\tau=T\}.$$

Точки (x_i, t_n), $i=0, 1, \ldots, N$, $n=0, 1, \ldots, K$, образуют узлы пространственно-временно́й сетки $\omega_{h,\tau}=\omega_h\times\omega_\tau$ (см. рис. 10). Узлы (x_i, t_n), принадлежащие отрезкам $l_0=\{0\leqslant x\leqslant 1, t=0\}$, $l_1=\{x=0, 0\leqslant t\leqslant T\}$, $l_2=\{x=1, 0\leqslant t\leqslant T\}$, называются *граничными узлами* сетки $\omega_{h,\tau}$, а остальные узлы — *внутренними*. На рис. 10 граничные узлы обозначены крестиками, а внутренние — кружочками.

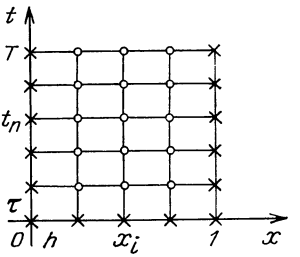


Рис. 10. Пространственно-временна́я сетка $\omega_{h,\tau}$

Слоем называется множество всех узлов сетки $\omega_{h,\tau}$, имеющих одну и ту же временну́ю координату. Так, n-м слоем называется множество узлов

$$(x_0, t_n), (x_1, t_n), \ldots, (x_N, t_n).$$

Для функции $y(x, t)$, определенной на сетке $\omega_{h,\tau}$, введем обозначения $y_i^n=y(x_i, t_n)$,

$$y_{t,i}^n=\frac{y_i^{n+1}-y_i^n}{\tau}, \quad y_{\overline{x}x,i}^n=\frac{y_{i+1}^n-2y_i^n+y_{i-1}^n}{h^2}. \tag{4}$$

Иногда для упрощения записи индексы i и n будем опускать, обозначая $y_t=y_{t,i}^n$, $y_{\overline{x}x}=y_{\overline{x}x,i}^n$.

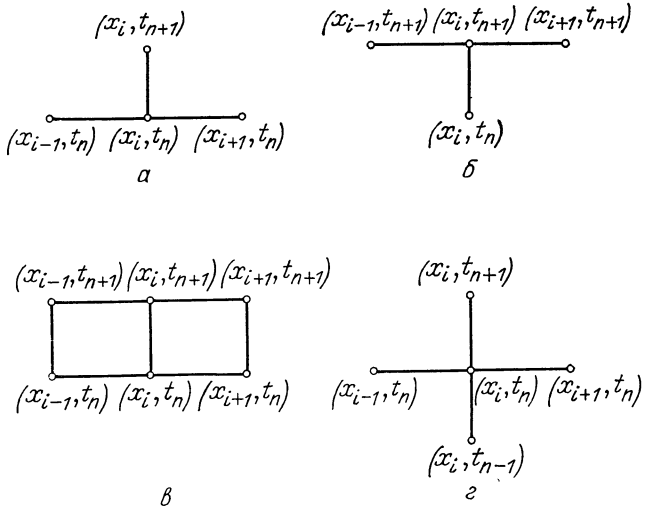


Рис. 11. Шаблоны разностных схем: *а* — явная схема; *б* — чисто неявная схема; *в* — симметричная схема; *г* — трехслойная схема

Чтобы аппроксимировать уравнение (1) в точке (x_i, t_n), введем шаблон, изображенный на рис. 11, *а* и состоящий из четырех узлов $(x_{i\pm1}, t_n)$, (x_i, t_n), (x_i, t_{n+1}). Производную $\partial u/\partial t$ заменим в точке (x_i, t_n) разностным отношением $y_{t,i}^n$, а производную $\partial^2 u/\partial x^2$ — второй разностной производной $y_{\overline{x}x,i}^n$. Правую часть $f(x, t)$ заменим при-

ближенно сеточной функцией φ_i^n, в качестве φ_i^n можно взять одно из следующих выражений:

$$f(x_i, t_n),\ \frac{1}{h}\int\limits_{x_{i-1/2}}^{x_{i+1/2}} f(x, t_n)\,dx,\ \frac{1}{h\tau}\int\limits_{t_n}^{t_{n+1}} dt \int\limits_{x_{i-1/2}}^{x_{i+1/2}} f(x, t)\,dx.$$

В результате получим разностное уравнение

$$\frac{y_i^{n+1} - y_i^n}{\tau} = \frac{y_{i+1}^n - 2y_i^n + y_{i-1}^n}{h^2} + \varphi_i^n, \tag{5}$$

которое аппроксимирует исходное дифференциальное уравнение в точке (x_i, t_n) с первым порядком по τ и вторым порядком по h при условии, что разность $\varphi_i^n - f(x_i, t_n)$ имеет тот же порядок малости.

Под разностной схемой понимается совокупность разностных уравнений, аппроксимирующих основное дифференциальное уравнение во всех внутренних узлах сетки и дополнительные (начальные и граничные) условия — в граничных узлах сетки. Разностную схему по аналогии с дифференциальной задачей будем называть также разностной задачей. В данном случае разностная схема имеет вид

$$\frac{y_i^{n+1} - y_i^n}{\tau} = \frac{y_{i+1}^n - 2y_i^n + y_{i-1}^n}{h^2} + \varphi_i^n,$$

$$i = 1, 2, \ldots, N-1, \quad n = 0, 1, \ldots, K-1, \quad hN = 1, \quad K\tau = T, \tag{6}$$

$$y_0^n = \mu_1(t_n), \quad y_N^n = \mu_2(t_n), \quad n = 0, 1, 2, \ldots, K,$$

$$y_i^0 = u_0(x_i), \quad i = 0, 1, \ldots, N.$$

Эта схема представляет собой систему линейных алгебраических уравнений с числом уравнений, равным числу неизвестных. Находить решение такой системы следует по слоям. Решение на нулевом слое задано начальными условиями $y_i^0 = u_0(x_i)$, $i = 0, 1, \ldots, N$. Если решение y_i^n, $i = 0, 1, \ldots, N$, на слое n уже найдено, то решение y_i^{n+1} на слое $n+1$ находится по явной формуле

$$y_i^{n+1} = y_i^n + \tau\,(y_{\bar{x}x,i}^n + \varphi_i^n), \quad i = 1, 2, \ldots, N-1, \tag{7}$$

а значения $y_0^{n+1} = \mu_1(t_{n+1})$, $y_N^{n+1} = \mu_2(t_{n+1})$ доопределяются из граничных условий. По этой причине схема (6) называется *явной разностной схемой*. Несколько позже мы познакомимся и с неявными схемами, в которых для нахождения y_i^{n+1} при заданных y_i^n требуется решать систему уравнений.

Погрешность разностной схемы (6) определяется как разность $z_i^n = y_i^n - u(x_i, t_n)$ между решением задачи (6) и решением исходной задачи (1)—(3). Подставляя в (6) $y_i^n = z_i^n + u(x_i, t_n)$, получим

уравнение для погрешности

$$\frac{z_i^{n+1}-z_i^n}{\tau}=\frac{z_{i+1}^n-2z_i^n+z_{i-1}^n}{h^2}+\psi_i^n,$$
$$i=1, 2, \ldots, N-1, \quad n=0, 1, \ldots, K-1, \quad hN=1, \quad K\tau=T, \tag{8}$$
$$z_0^n=z_N^n=0, \quad n=1, 2, \ldots, K, \quad z_i^0=0, \quad i=0, 1, \ldots, N,$$

где $\psi_i^n=-u_{t,i}^n+u_{\overline{xx},i}^n+\varphi_i^n$ — *погрешность аппроксимации разностной схемы* (6) *на решении задачи* (1)—(3), $\psi_i^n=O(\tau+h^2)$. Можно оценить решение z_i^n уравнения (8) через правую часть ψ_i^n и доказать тем самым сходимость разностной схемы (6) с первым порядком по τ и вторым — по h. Однако это исследование мы отложим до § 3 гл. 3, а сейчас на примере схемы (6) продемонстрируем один распространенный прием исследования разностных схем с постоянными коэффициентами, называемый *методом гармоник*. Хотя данный метод не является достаточно обоснованным, в частности не учитывает влияния граничных условий и правых частей, он позволяет легко найти необходимые условия устойчивости и сходимости разностных схем. Покажем, например, что *явную схему* (6) *можно применять лишь при условии* $\tau\leqslant 0{,}5h^2$, означающем, что шаг по времени надо брать достаточно малым.

Рассмотрим уравнение

$$\frac{y_j^{n+1}-y_j^n}{\tau}=\frac{y_{j+1}^n-2y_j^n+y_{j-1}^n}{h^2}, \tag{9}$$

т. е. однородное уравнение, соответствующее (5). Будем искать частные решения уравнения (9), имеющие вид

$$y_j^n(\varphi)=q^n e^{ijh\varphi}, \tag{10}$$

где i — мнимая единица, φ — любое действительное число и q — число, подлежащее определению. Подставляя (10) в уравнение (9) и сокращая на $e^{ijh\varphi}$, получим

$$\frac{q-1}{\tau}=\frac{e^{ih\varphi}-2+e^{-ih\varphi}}{h^2},$$

откуда найдем

$$q=1-4\gamma\sin^2\frac{h\varphi}{2}, \qquad \gamma=\frac{\tau}{h^2}. \tag{11}$$

Начальные условия $y_j^0(\varphi)=e^{ijh\varphi}$, соответствующие решениям вида (10) (их называют гармониками), ограничены. Если для некоторого φ множитель q станет по модулю больше единицы, то решение вида (10) будет неограниченно возрастать при $n\to\infty$. В этом случае разностное уравнение (9) называется *неустойчивым*, поскольку нарушается непрерывная зависимость его решения от начальных условий. Если же $|q|\leqslant 1$ для всех действительных φ, то

все решения вида (10) ограничены при любом n и разностное уравнение (9) называется *устойчивым*. В случае неустойчивости найти решение разностной задачи (6) по формулам (7) практически невозможно, так как погрешности (например погрешности округления), внесенные в начальный момент времени, будут неограниченно возрастать при увеличении n. Такие разностные схемы называются *неустойчивыми*.

Для уравнения (9) неравенство $|q| \leqslant 1$ выполняется согласно (11) при всех φ тогда и только тогда, когда $\gamma \leqslant 0{,}5$. Таким образом, использование схемы (6) возможно лишь при выполнении условия $\tau \leqslant 0{,}5h^2$. Разностные схемы, устойчивые лишь при некотором ограничении на отношение шагов по пространству и по времени, называются *условно устойчивыми*. Следовательно, схема (6) условно устойчива, причем условие устойчивости имеет вид $\tau/h^2 \leqslant 0{,}5$. Условно устойчивые схемы для уравнений параболического типа используются редко, так как они накладывают слишком сильное ограничение на шаг по времени. Действительно, пусть, например, $h = 10^{-2}$. Тогда шаг τ не должен превосходить $0{,}5 \cdot 10^{-4}$, и для того чтобы вычислить решение y_i^n при $t = 1$, надо взять число шагов по времени $n = \tau^{-1} \geqslant 2 \cdot 10^4$, т. е. провести не менее $2 \cdot 10^4$ вычислений по формулам (7). В следующем пункте будет показано, что многие неявные схемы лишены этого недостатка и являются устойчивыми при любых шагах h и τ. Такие схемы называются *абсолютно устойчивыми*.

3. Неявные схемы. *Чисто неявной разностной схемой* для уравнения теплопроводности (схемой с опережением) называется разностная схема, использующая шаблон (x_i, t_n), $(x_{i\pm1}, t_{n+1})$, (x_i, t_{n+1}) (см. рис. 11, *б*) и имеющая вид

$$\frac{y_i^{n+1} - y_i^n}{\tau} = \frac{y_{i+1}^{n+1} - 2y_i^{n+1} + y_{i-1}^{n+1}}{h^2} + \varphi_i^n, \tag{12}$$

$$i = 1, 2, \ldots, N-1, \quad n = 0, 1, \ldots, K-1,$$

$$y_0^{n+1} = \mu_1(t_{n+1}), \quad y_N^{n+1} = \mu_2(t_{n+1}), \quad n = 0, 1, \ldots, K-1,$$

$$y_i^0 = u_0(x_i), \quad i = 0, 1, \ldots, N.$$

Здесь $\varphi_i^n = f(x_i, t_{n+1}) + O(\tau + h^2)$. Схема имеет первый порядок аппроксимации по τ и второй — по h. Решение системы (12) находится, как и в случае явной схемы, по слоям, начиная с $n = 1$. Однако теперь, в отличие от явной схемы, для нахождения y_i^{n+1} по известным y_i^n требуется решить систему уравнений

$$\gamma y_{i+1}^{n+1} - (1 + 2\gamma) y_i^{n+1} + \gamma y_{i-1}^{n+1} = -F_i^n, \quad i = 1, 2, \ldots, N-1,$$
$$y_0^{n+1} = \mu_1(t_{n+1}), \quad y_N^{n+1} = \mu_2(t_{n+1}), \tag{13}$$

где $\gamma = \tau/h^2$, $F_i^n = y_i^n + \tau\varphi_i^n$. Эту систему можно решать методом прогонки (см. п. 7 § 4 ч. I), так как условия устойчивости прогонки выполнены.

Для исследования устойчивости разностной схемы (12) будем искать частные решения уравнения

$$\frac{y_j^{n+1}-y_j^n}{\tau}=\frac{y_{j+1}^{n+1}-2y_j^{n+1}+y_{j-1}^{n+1}}{h^2},$$

имеющие вид (10). Тогда получим

$$q=\left(1+4\gamma\sin^2\frac{h\varphi}{2}\right)^{-1},\qquad \gamma=\frac{\tau}{h^2},$$

следовательно, $|q|\leqslant 1$ при любых φ, τ, h. Таким образом, схема (12) абсолютно устойчива, т. е. устойчива при любых шагах τ и h. *Абсолютная устойчивость является основным преимуществом неявных схем.* Теперь уже не надо брать шаг τ слишком малым, можно взять, например, $\tau=h=10^{-2}$. Величина шагов сетки τ, h определяется теперь необходимой точностью расчета, а не соображениями устойчивости.

Шеститочечной симметричной схемой называется разностная схема

$$\frac{y_i^{n+1}-y_i^n}{\tau}=\frac{1}{2}\left(y_{\overline{x}x,i}^{n+1}+y_{\overline{x}x,i}^n\right)+\varphi_i^n, \tag{14}$$

для которой начальные и граничные условия задаются так же, как и в схеме (12). Эта схема использует шеститочечный шаблон, изображенный на рис. 11, *в*.

Предлагаем читателю самостоятельно доказать, что эта схема имеет второй порядок аппроксимации как по h, так и по τ (если только $\varphi_i^n=f(x_i, t_n+0{,}5\tau)+O(\tau^2+h^2)$), она абсолютно устойчива и ее можно решать методом прогонки.

Обобщением трех рассмотренных схем является однопараметрическое *семейство схем с весами.* Зададим произвольный действительный параметр σ и определим разностную схему

$$\begin{gathered}\frac{y_i^{n+1}-y_i^n}{\tau}=\sigma y_{\overline{x}x,i}^{n+1}+(1-\sigma)\,y_{\overline{x}x,i}^n+\varphi_i^n,\\ i=1, 2, \ldots, N-1,\quad n=0, 1, \ldots, K-1,\\ y_0^{n+1}=\mu_1(t_{n+1}),\ y_N^{n+1}=\mu_2(t_{n+1}),\ n=0, 1, \ldots, K-1,\\ y_i^0=u_0(x_i),\ i=0, 1, \ldots, N.\end{gathered} \tag{15}$$

При $\sigma=0$ получим отсюда явную схему, при $\sigma=1$ — чисто неявную схему и при $\sigma=0{,}5$ — симметричную схему (14). Исследуем погрешность аппроксимации схемы (15) на решении исходной задачи (1) — (3). Представим решение задачи (15) в виде $y_i^n=u(x_i, t_n)+z_i^n$, где $u(x_i, t_n)$ — точное решение дифференциальной задачи (1) — (3). Тогда для погрешности получим систему

уравнений

$$\frac{z_i^{n+1}-z_i^n}{\tau}=\sigma z_{\bar{x}x,i}^{n+1}+(1-\sigma)z_{\bar{x}x,i}^n+\psi_i^n, \tag{16}$$

$$i=1, 2, \ldots, N-1, \quad n=0, 1, \ldots, K-1,$$

$$z_0^{n+1}=z_N^{n+1}=0,\ n=0, 1, \ldots, K-1,\ z_i^0=0,\ i=0, 1, \ldots, N.$$

Сеточная функция ψ_i^n, входящая в правую часть уравнения (16) и равная

$$\psi_i^n=\sigma u_{\bar{x}x,i}^{n+1}+(1-\sigma)u_{\bar{x}x,i}^n-u_{t,i}^n+\varphi_i^n, \tag{17}$$

называется погрешностью аппроксимации схемы (15) на решении задачи (1) – (3). Получим первые члены разложения функции ψ_i^n по степеням h и τ. Будем разлагать все функции, входящие в выражение для ψ_i^n, по формуле Тейлора в точке $(x_i,\ t_n+0{,}5\tau)$. Учитывая разложения $u_{t,i}^n=\dot{u}(x_i, t_{n+1/2})+O(\tau^2)$, $u_{\bar{x}x,i}=u''(x_i)+\frac{h^2}{12}u^{\mathrm{IV}}(x_i)+O(h^4)$, где $u''=\partial^2u/\partial x^2$, $\dot{u}=\partial u/\partial t$, $t_{n+1/2}=t_n+0{,}5\tau$, получим

$$\psi_i^n=\sigma\left(u''(x_i, t_{n+1})+\frac{h^2}{12}u^{\mathrm{IV}}(x_i, t_{n+1})\right)+(1-\sigma)\left(u''(x_i, t_n)+\right.$$
$$\left.+\frac{h^2}{12}u^{\mathrm{IV}}(x_i, t_n)\right)-\dot{u}(x_i, t_{n+1/2})+\varphi_i^n+O(\tau^2)+O(h^4).$$

Отсюда, проводя разложение в точке $(x_i, t_{n+1/2})$ и обозначая $u=u(x_i, t_{n+1/2})$, будем иметь

$$\psi_i^n=\sigma\left(u''+\frac{\tau}{2}\dot{u}''+\frac{h^2}{12}u^{\mathrm{IV}}\right)+(1-\sigma)\left(u''-\frac{\tau}{2}\dot{u}''+\frac{h^2}{12}u^{\mathrm{IV}}\right)-$$
$$-\dot{u}+\varphi_i^n+O(\tau^2)+O(h^4)$$

и, перегруппировывая слагаемые, получим, что

$$\psi_i^n=(u''-\dot{u}+\varphi_i^n)+(\sigma-0{,}5)\tau\dot{u}''+\frac{h^2}{12}u^{\mathrm{IV}}+O(\tau^2+h^4).$$

Учитывая уравнение (1) $u''-\dot{u}=-f$ и следствие из него $u^{\mathrm{IV}}-\dot{u}''=-f''$, окончательно можем записать, что

$$\psi_i^n=\left[(\sigma-0{,}5)\tau+\frac{h^2}{12}\right]\dot{u}''+\varphi_i^n-f(x_i, t_{n+1/2})-$$
$$-\frac{h^2}{12}f''(x_i, t_{n+1/2})+O(\tau^2+h^4). \tag{18}$$

Из формулы (18) можно сделать следующие выводы. Если

$$\sigma=\sigma_*=\frac{1}{2}-\frac{h^2}{12\tau},\quad \varphi_i^n=f(x_i, t_{n+1/2})+\frac{h^2}{12}f''(x_i, t_{n+1/2})+O(\tau^2+h^4),$$

то схема (15) имеет второй порядок аппроксимации по τ и четвертый – по h. Такая схема называется *схемой повышенного порядка*

аппроксимации. Если

$$\sigma = 0{,}5, \quad \varphi_i^n = f(x_i, t_{n+1/2}) + O(\tau^2 + h^2),$$

то схема (15) имеет второй порядок аппроксимации по τ и по h. При остальных значениях σ и при $\varphi_i^n = f(x_i, t_{n+1}) + O(\tau + h^2)$ схема (15) имеет первый порядок аппроксимации по τ и второй — по h.

Опуская выкладки, отметим, что если искать решение уравнения (15) с $\varphi_i^n \equiv 0$ в виде (10), то получим

$$q = \frac{1 - 4\gamma(1-\sigma)\sin^2\frac{h\varphi}{2}}{1 + 4\gamma\sigma\sin^2\frac{h\varphi}{2}}$$

и $|q| \leqslant 1$ при всех φ, если

$$\sigma \geqslant \frac{1}{2} - \frac{h^2}{4\tau}. \tag{19}$$

Отсюда видно, в частности, что *все схемы с* $\sigma \geqslant 0{,}5$ *абсолютно устойчивы.* Схема повышенного порядка аппроксимации ($\sigma = \sigma_*$) также абсолютно устойчива, что проверяется непосредственно.

При $\sigma \neq 0$ разностная схема (15) является неявной схемой. Для нахождения решения y_i^{n+1} по заданным y_i^n требуется решать систему уравнений

$$\sigma\gamma y_{i+1}^{n+1} - (1 + 2\sigma\gamma) y_i^{n+1} + \sigma\gamma y_{i-1}^{n+1} = -F_i^n, \quad i = 1, 2, \ldots, N-1, \tag{20}$$

$$y_0^{n+1} = \mu_1(t_{n+1}), \quad y_N^{n+1} = \mu_2(t_{n+1}),$$

где

$$\gamma = \frac{\tau}{h^2}, \quad F_i^n = y_i^n + (1-\sigma)\tau y_{\overline{x}x,i}^n + \tau\varphi_i^n.$$

Система (20) решается методом прогонки. Условия устойчивости прогонки (условия (47), (48) из § 4 ч. I) при $\sigma \neq 0$ сводятся к неравенству

$$|1 + 2\sigma\gamma| \geqslant 2|\sigma|\gamma$$

и выполнены при $\sigma \geqslant -1/(4\gamma)$. Последнее неравенство следует из условия устойчивости (19) разностной схемы.

4. Уравнения с переменными коэффициентами и нелинейные уравнения. Рассмотрим первую краевую задачу для уравнения теплопроводности с переменными коэффициентами

$$\rho(x, t)\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(k(x, t)\frac{\partial u}{\partial x}\right) + f(x, t), \quad 0 < x < 1, \quad 0 < t \leqslant T, \tag{21}$$

$$u(x, 0) = u_0(x), \quad u(0, t) = \mu_1(t), \quad u(1, t) = \mu_2(t),$$

где $\rho(x, t)$, $k(x, t)$, $f(x, t)$ — достаточно гладкие функции, удовлетворяющие условиям

$$0 < c_1 \leqslant k(x, t) \leqslant c_2, \quad \rho(x, t) \geqslant c_3 > 0. \tag{22}$$

Дифференциальное выражение $Lu = \frac{\partial}{\partial x}\left(k(x, t)\frac{\partial u}{\partial x}\right)$ при каждом фиксированном t аппроксимируем в точке (x_i, t) так же, как и в стационарном случае (см. § 1), разностным отношением

$$\Lambda(t)\, y_i = (a(x_i, t)\, y_{\bar{x}})_{x,i} = \\ = \frac{1}{h}\left[a(x_{i+1}, t)\frac{y_{i+1} - y_i}{h} - a(x_i, t)\frac{y_i - y_{i-1}}{h}\right], \qquad (23)$$

где разностный коэффициент теплопроводности $a(x_i, t)$ должен удовлетворять условиям второго порядка аппроксимации

$$a(x_{i+1}, t) + a(x_i, t) = 2k(x_i, t) + O(h^2),$$

$$\frac{a(x_{i+1}, t) - a(x_i, t)}{h} = k'(x_i, t) + O(h^2).$$

Наиболее употребительны следующие выражения для $a(x_i, t)$:

$$a(x_i, t) = 0{,}5\,(k(x_i, t) + k(x_{i-1}, t)), \quad a(x_i, t) = k\left(x_i - \frac{h}{2}, t\right),$$

$$a(x_i, t) = \frac{2k(x_{i-1}, t)\, k(x_i, t)}{k(x_{i-1}, t) + k(x_i, t)}.$$

Разностная схема с весами для задачи (21) имеет вид

$$\rho(x_i, t)\frac{y_i^{n+1} - y_i^n}{\tau} = \Lambda(t)\,(\sigma y_i^{n+1} + (1 - \sigma)\, y_i^n) + f(x_i, t), \; i = 1, 2, \ldots, N - 1, \qquad (24)$$

$$y_0^n = \mu_1(t_n), \quad y_N^n = \mu_2(t_n), \quad y_i^0 = u_0(x_i).$$

Здесь в качестве t можно взять любое значение $t \in [t_n, t_{n+1}]$, например $t = t_n + 0{,}5\tau$. Если в уравнении (24) $t = t_n + 0{,}5\tau$, $\sigma = 0{,}5$, то схема (24) имеет второй порядок аппроксимации по τ и по h. При остальных значениях σ и t выполняется первый порядок аппроксимации по τ и второй — по h.

При исследовании устойчивости разностных схем с переменными коэффициентами иногда применяется *принцип замороженных коэффициентов*, сводящий задачу к уравнению с постоянными коэффициентами. Рассмотрим явную схему, соответствующую уравнению (24) с $\sigma = 0$ и $f(x_i, t) \equiv 0$, т. е. схему

$$\rho(x_i, t)\frac{y_i^{n+1} - y_i^n}{\tau} = (a(x_i, t)\, y_{\bar{x}}^n)_{x,i}. \qquad (25)$$

Предположим, что коэффициенты $\rho(x_i, t)$, $a(x_i, t)$ — постоянные, $\rho(x_i, t) \equiv \rho = \text{const}$, $a(x_i, t) \equiv a = \text{const}$. Тогда уравнение (25) можно записать в виде

$$\rho\frac{y_i^{n+1} - y_i^n}{\tau} = a y_{\bar{x}x,i}$$

или

$$\frac{y_i^{n+1}-y_i^n}{\tau'}=y_{\bar{x}x,i},\qquad \tau'=\frac{\tau a}{\rho}.$$

Из п. 2 известно, что последнее уравнение устойчиво при $\tau'\leqslant 0{,}5h^2$, т. е. при

$$\frac{\tau a}{\rho}\leqslant\frac{h^2}{2}. \tag{26}$$

Принцип замороженных коэффициентов утверждает, что схема (25) устойчива, если условие (26) выполнено при всех допустимых значениях $a(x_i, t)$, $\rho(x_i, t)$, т. е. если при всех x, t выполнены неравенства

$$\frac{\tau a(x_i, t)}{\rho(x_i, t)}\leqslant\frac{h^2}{2}. \tag{27}$$

Если известно, что $0<c_1\leqslant a(x_i, t)\leqslant c_2$, $\rho(x_i, t)\geqslant c_3>0$, то неравенство (27) будет выполнено при

$$\frac{\tau}{h^2}\leqslant\frac{c_3}{2c_2}.$$

Строгое обоснование устойчивости схемы (25) будет дано в примере 2 из § 4 гл. 2.

Если параметр $\sigma\geqslant 0{,}5$, то из принципа замороженных коэффициентов следует абсолютная устойчивость схемы (24).

Рассмотрим теперь первую краевую задачу для нелинейного уравнения теплопроводности

$$\frac{\partial u}{\partial t}=\frac{\partial}{\partial x}\left(k(u)\frac{\partial u}{\partial x}\right)+f(u). \tag{28}$$

В случае нелинейных уравнений, когда заранее неизвестны пределы изменения функции $k(u)$, избегают пользоваться явными схемами. Чисто неявная схема, линейная относительно y_i^{n+1}, $i=1, 2, \ldots, N-1$, имеет вид

$$\frac{y_i^{n+1}-y_i^n}{\tau}=\frac{1}{h}\left(a_{i+1}\frac{y_{i+1}^{n+1}-y_i^{n+1}}{h}-a_i\frac{y_i^{n+1}-y_{i-1}^{n+1}}{h}\right)+f(y_i^n), \tag{29}$$

где $a_i=0{,}5\,(k(y_i^n)+k(y_{i-1}^n))$. Эта схема абсолютно устойчива, имеет первый порядок аппроксимации по τ и второй — по h. Решение y_i^{n+1}, $i=1, 2, \ldots, N-1$, находится методом прогонки. Заметим, что схему (29) можно записать в виде

$$y_{t,i}^n=\frac{1}{2}\left((ky_{\bar{x}})_{x,i}+(ky_x)_{\bar{x},i}\right)+f(y_i^n),$$

где $k_i=k(y_i^n)$.

Часто используется нелинейная схема

$$\frac{y_i^{n+1}-y_i^n}{\tau}=\frac{1}{h}\left(a(y_{i+1}^{n+1})\frac{y_{i+1}^{n+1}-y_i^{n+1}}{h}-a(y_i^{n+1})\frac{y_i^{n+1}-y_{i-1}^{n+1}}{h}\right)+f(y_i^{n+1}),$$

$$a(y_i^{n+1})=\frac{k(y_i^{n+1})+k(y_{i-1}^{n+1})}{2}. \qquad (30)$$

Для реализации этой схемы необходимо применить тот или иной итерационный метод. Например такой:

$$\frac{y_i^{(s+1)}-y_i^n}{\tau}=\frac{1}{h}\left(a(y_{i+1}^{(s)})\frac{y_{i+1}^{(s+1)}-y_i^{(s+1)}}{h}-a(y_i^{(s)})\frac{y_i^{(s+1)}-y_{i-1}^{(s+1)}}{h}\right)+f(y_i^{(s)}), \qquad (31)$$

$$s=0,\ 1,\ \ldots,\ M-1,\quad y_i^{(0)}=y_i^n,\quad y_i^{(M)}=y_i^{n+1}.$$

Здесь s — номер итерации. Как видим, нелинейные коэффициенты берутся с предыдущей итерации, а в качестве начального приближения для y_i^{n+1} выбирается y_i^n. Это начальное приближение тем лучше, чем меньше шаг τ. Число итераций M задается из соображений точности. В задачах с гладкими коэффициентами при $k(u)\geqslant c_1>0$ часто бывает достаточно провести две — три итерации. Значения $y_i^{(s+1)}$ на новой итерации находятся из системы (31) методом прогонки. При $M=1$ итерационный метод (31) совпадает с разностной схемой (29).

Для приближенного решения нелинейного уравнения (28) применяются также схемы предиктор — корректор второго порядка точности, аналогичные методу Рунге — Кутта для обыкновенных дифференциальных уравнений (см. п. 2 § 1 гл. 6 ч. II). Здесь переход со слоя n на слой $n+1$ осуществляется в два этапа. Приведем пример такой схемы. На первом этапе решается неявная линейная система уравнений

$$\frac{y_i^{n+1/2}-y_i^n}{0{,}5\tau}=(a(y_i^n)\,y_{\overline{x}}^{n+1/2})_{x,i}+f(y_i^n),\quad i=1,\ 2,\ \ldots,\ N-1,$$

$$y_0^{n+1/2}=\mu_1(t_n+0{,}5\tau),\quad y_N^{n+1/2}=\mu_2(t_n+0{,}5\tau),$$

из которой находятся промежуточные значения $y_i^{n+1/2}$, $i=0,\ 1,\ \ldots,\ N$. Затем на втором этапе используется симметричная шеститочечная схема для уравнения (28), в которой нелинейные коэффициенты $a(y)$, $f(y)$ вычисляются при $y=y_i^{n+1/2}$, т. е. схема

$$\frac{y_i^{n+1}-y_i^n}{\tau}=\frac{1}{2}((a(y_i^{n+1/2})\,y_{\overline{x}}^{n+1})_{x,i}+(a(y_i^{n+1/2})\,y_{\overline{x}}^{n})_{x,i})+f(y_i^{n+1/2}),$$

$$i=1,\ 2,\ \ldots,\ N-1,$$

$$y_0^{n+1}=\mu_1(t_{n+1}),\quad y_N^{n+1}=\mu_2(t_{n+1}).$$

§ 5. Трехслойные разностные схемы

1. Разностные схемы для уравнения колебаний. Рассмотрим первую краевую задачу для уравнения колебаний

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 1, \quad 0 < t \leqslant T, \tag{1}$$

$$u(0, t) = \mu_1(t), \quad u(1, t) = \mu_2(t), \quad 0 \leqslant t \leqslant T, \tag{2}$$

$$u(x, 0) = u_0(x), \quad \frac{\partial u}{\partial t}(x, 0) = \bar{u}_0(x), \quad 0 \leqslant x \leqslant 1. \tag{3}$$

Известно (см., например, [41]), что эта задача поставлена корректно, т. е. ее решение существует, оно единственно и непрерывно зависит от начальных и граничных данных.

Будем использовать ту же сетку $\omega_{h\tau}$, что и в § 4, т. е. $\omega_{h\tau} = \omega_h \times \omega_\tau$,

$$\omega_h = \{x_i = ih, \ i = 0, 1, \ldots, N, \ hN = 1\},$$
$$\omega_\tau = \{t_n = n\tau, \ n = 0, 1, \ldots, K, \ K\tau = T\}.$$

Очевидно, минимальный шаблон, на котором можно аппроксимировать уравнение (1), это пятиточечный шаблон, изображенный на рис 11, *г*. Таким образом, в отличие от схем для уравнения теплопроводности, в которых использовалось только два временны́х слоя (слои n и $n+1$), здесь требуется использовать три слоя: $n-1$, n, $n+1$. Такие схемы называются *трехслойными*. Их применение предполагает, что при нахождении значений y_i^{n+1} на верхнем слое значения на предыдущих слоях y_i^{n-1}, y_i^n, $i=0, 1, \ldots, N$ хранятся в памяти ЭВМ.

Видимо, простейшей разностной аппроксимацией уравнения (1) и граничных условий (2) является следующая система уравнений:

$$\frac{y_i^{n+1} - 2y_i^n + y_i^{n-1}}{\tau^2} = \frac{y_{i+1}^n - 2y_i^n + y_{i-1}^n}{h^2}, \tag{4}$$

$$i = 1, 2, \ldots, N-1, \quad n = 1, 2, \ldots, K-1,$$

$$y_0^{n+1} = \mu_1(t_{n+1}), \quad y_N^{n+1} = \mu_2(t_{n+1}), \quad n = 0, 1, \ldots, K-1. \tag{5}$$

Разностное уравнение (4) имеет второй порядок погрешности аппроксимации по τ и по h. Решение y_i^{n+1} выражается явным образом через значения на предыдущих слоях:

$$y_i^{n+1} = 2y_i^n - y_i^{n-1} + \gamma(y_{i+1}^n - 2y_i^n + y_{i-1}^n),$$

$$i = 1, 2, \ldots, N-1, \quad \gamma = \tau^2/h^2, \quad n = 1, 2, \ldots, K-1. \tag{6}$$

Для начала счета по формулам (6) должны быть заданы значения y_i^0, y_i^1, $i=0, 1, \ldots, N-1, N$. Из первого начального условия (3) сразу получаем

$$y_i^0 = u_0(x_i), \quad i = 1, 2, \ldots, N-1. \tag{7}$$

Простейшая замена второго из начальных условий (3) уравнением $(y_i^1 - y_i^0)/\tau = \bar{u}_0(x_i)$ имеет лишь первый порядок аппроксимации по τ. Поскольку уравнение (2) аппроксимирует основное уравнение (1) со вторым порядком, желательно, чтобы и разностное начальное условие также имело второй порядок аппроксимации. Чтобы добиться этого, воспользуемся разложением

$$\frac{u(x, \tau) - u(x, 0)}{\tau} = \frac{\partial u(x, 0)}{\partial t} + \frac{\tau}{2}\frac{\partial^2 u(x, 0)}{\partial t^2} + O(\tau^2)$$

и учтем, что в силу дифференциального уравнения (1) выполняется равенство

$$\frac{\partial^2 u(x, 0)}{\partial t^2} = \frac{\partial_2 u(x, 0)}{\partial x^2} = u_0''(x).$$

Таким образом,

$$\frac{\partial u(x, 0)}{\partial t} = \frac{u(x, \tau) - u(x, 0)}{\tau} - \frac{\tau}{2} u_0''(x) + O(\tau^2),$$

и, следовательно, разностное уравнение

$$\frac{y_i^1 - y_i^0}{\tau} = \bar{u}_0(x_i) + \frac{\tau}{2} u_{0\,\bar{x}x,i}, \qquad i = 1, 2, \ldots, N-1, \tag{8}$$

аппроксимирует второе из условий (3) со вторым порядком по τ и по h.

Совокупность уравнений (4), (5), (7), (8) составляет разностную схему, аппроксимирующую исходную задачу (1) – (3).

Покажем еще один способ получения уравнения (8). Уравнение

$$\frac{y_i^1 - y_i^{-1}}{2\tau} = \bar{u}_0(x_i), \qquad y_i^{-1} = y(x_i, -\tau) \tag{9}$$

аппроксимирует уравнение $u_t'(0, x) = \bar{u}_0(x)$ со вторым порядком. Чтобы найти значения y_i^{-1} запишем уравнение (4) при $n = 0$:

$$\frac{y_i^1 - 2y_i^0 + y_i^{-1}}{\tau^2} = y_{\bar{x}x,i}^0,$$

и учтем, что $y_i^0 = u_0(x_i)$. Отсюда получим $y_i^{-1} = -y_i^1 + 2u_0(x_i) + \tau^2 u_{0\bar{x}x,i}$. Подставляя это выражение для y^{i-1} в уравнение (9), приходим к уравнению (8).

Для исследования устойчивости будем так же, как и в § 4, искать решение уравнения (4) в виде

$$y_j^n = q^n e^{ijh\varphi}. \tag{10}$$

Подставляя это выражение в (4) и сокращая на $e^{ijh\varphi}$, получим для q квадратное уравнение

$$q^2 - 2\left(1 - 2\gamma \sin^2 \frac{h\varphi}{2}\right) q + 1 = 0, \qquad \gamma = \frac{\tau^2}{h^2}. \tag{11}$$

Будем считать разностное уравнение (4) устойчивым, если оба корня уравнения (11) не превосходят по модулю единицу. Пусть q_1 и q_2 — корни этого уравнения. Если оба корня действительные, то поскольку $q_1q_2=1$, найдется φ, для которого один из корней меньше единицы по модулю, а второй — больше единицы. Если же корни комплексно сопряженные, то $|q_1|=|q_2|=1$. Таким образом, разностное уравнение (4) устойчиво, если при всех действительных φ выполняется неравенство $\left(1-2\gamma\sin^2\frac{h\varphi}{2}\right)^2\leqslant 1$, т. е. $\gamma\sin^2\frac{h\varphi}{2}\leqslant 1$. Последнее неравенство выполняется при всех φ, если

$$\tau\leqslant h. \tag{12}$$

Строгое обоснование устойчивости схемы (4) будет дано в § 3 гл. 4.

2. Трехслойные схемы для уравнения теплопроводности. Хотя трехслойные схемы для уравнения теплопроводности

$$\frac{\partial u}{\partial t}=\frac{\partial^2 u}{\partial x^2} \tag{13}$$

применяются значительно реже двухслойных, их иногда используют для повышения порядка аппроксимации или для улучшения устойчивости. Приведем несколько примеров трехслойных схем для уравнения (13). На первый взгляд кажется очень естественным заменить уравнение (13) явной симметричной схемой второго порядка аппроксимации

$$\frac{y_j^{n+1}-y_j^{n-1}}{2\tau}=\frac{y_{j+1}^n-2y_j^n+y_{j-1}^n}{h^2}. \tag{14}$$

Однако эта схема совершенно непригодна для использования на быстродействующих ЭВМ, поскольку при любых шагах τ и h она является неустойчивой. Если искать ее решение в виде (10), то получим уравнение

$$q^2+8\gamma\sin^2\frac{h\varphi}{2}\,q-1=0,\qquad \gamma=\frac{\tau}{h^2},$$

один из корней которого по модулю всегда больше единицы.

Если в уравнении (14) заменить значение y_j^n на полусумму $0{,}5\,(y_j^{n+1}+y_j^{n-1})$, то получим схему

$$\frac{y_j^{n+1}-y_j^{n-1}}{2\tau}=\frac{y_{j+1}^n-y_j^{n+1}-y_j^{n-1}+y_{j-1}^n}{h^2}, \tag{15}$$

которая интересна тем, что является абсолютно устойчивой, но обладает условной аппроксимацией. Обозначим

$$y_{\overset{\circ}{t},i}^n=\frac{y_i^{n+1}-y_i^{n-1}}{2\tau},\qquad y_{\overline{tt},i}^n=\frac{y_i^{n+1}-2y_i^n+y_i^{n-1}}{\tau^2}.$$

Тогда уравнение (15) можно переписать в виде

$$y^n_{\overset{\circ}{t},i} + \frac{\tau^2}{h^2} y^n_{\overline{t}t,i} = y^n_{\overline{x}x,i}.$$

Отсюда легко получить, что уравнение (15) аппроксимирует исходное дифференциальное уравнение (13) лишь при условии, что $\tau^2/h^2 \to 0$ при $\tau \to 0$, $h \to 0$. Погрешность аппроксимации является величиной $O(\tau^2 + h^2 + \tau^2/h^2)$. Если же положить, например, $\tau = h$, то (15) будет аппроксимировать уравнение гиперболического типа

$$\frac{\partial u}{\partial t} + \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}.$$

§ 6. Основные понятия теории разностных схем: аппроксимация, сходимость, устойчивость

1. Введение. Ранее мы уже встречались с перечисленными в заглавии понятиями в связи с самыми различными примерами разностных схем для конкретных дифференциальных уравнений. В настоящем параграфе дается изложение основных понятий теории разностных схем и выясняется связь между ними для линейных разностных схем самого общего вида, безотносительно к конкретной структуре исходного дифференциального уравнения и аппроксимирующей его разностной схемы.

Пусть дана исходная дифференциальная задача, которую мы запишем в виде

$$Lu(x) = f(x), \qquad (1)$$

где $x \in G$, G — область m-мерного пространства, $f(x)$ — заданная функция, L — линейный дифференциальный оператор. Предполагается, что дополнительные условия (типа начальных и граничных условий) учтены оператором L и правой частью f.

В качестве простейшего примера задачи (1) читатель может рассмотреть первую краевую задачу $-u''(x) = f(x)$, $0 < x < 1$, $u(0) = u(1) = 0$, хотя в общем случае уравнение (1) может быть многомерным, в том числе и нестационарным уравнением. Существенно в дальнейшем лишь требование линейности оператора L.

Для построения разностной схемы прежде всего вводится *сетка* G_h — конечное множество точек, принадлежащих G, плотность распределения которых характеризуется параметром h — *шагом сетки*. В общем случае параметр h — вектор, причем определена $|h|$ — длина вектора h. Обычно сетка G_h выбирается так, что при $|h| \to 0$ множество G_h стремится заполнить всю область $\overline{G}$. Функция, определенная в точках сетки G, называется *сеточной функцией*.

Пример 1. На отрезке $G = [a, b]$ введем произвольную неравномерную сетку G_h, т. е. множество точек

$$G_h = \{x_i \in [a, b] \mid x_0 = a < x_1 < \ldots < x_N = b\}.$$

Обозначим $h_i=x_i-x_{i-1}$, $i=1, 2, \ldots, N$. Тогда $h=(h_1, \ldots, h_n)$, $|h|=\max\limits_{1\leqslant i\leqslant N} h_i$.

Можно определить также $|h|=\left(\sum\limits_{i=1}^{N} h_i^2\right)^{1/2}$.

Пример 2. На плоскости (x, t) рассматривается область $G=\{0<x<1, 0<t\leqslant T\}$. Сетка G_h состоит из точек (x_i, t_n), где $x_i=ih$, $i=0, 1, \ldots, N$, $hN=1$, $t_n=n\tau$, $n=0, 1, \ldots, K$, $K\tau=T$. Эта сетка использовалась при аппроксимации уравнения теплопроводности в § 4. Здесь можно положить $|\bar{h}|=\sqrt{h^2+\tau^2}$, либо $|\bar{h}|=\sqrt{h^2+\tau}$.

После введения сетки G_h следует заменить в уравнении (1) дифференциальный оператор L разностным оператором L_h, правую часть $f(x)$ — сеточной функцией $\varphi_h(x)$. В результате получим систему разностных уравнений

$$L_h y_h(x)=\varphi_h(x), \quad x\in G_h, \tag{2}$$

которая называется *разностной схемой* или *разностной задачей*. В отличие от дифференциального уравнения решение разностной задачи будем обозначать буквой y.

2. Погрешность аппроксимации и погрешность схемы. Перейдем к изложению основных понятий теории разностных схем: аппроксимации, корректности (устойчивости) и сходимости. Прежде чем давать формальные определения, заметим, что свойство аппроксимации означает близость разностного оператора к дифференциальному. Отсюда еще не следует, вообще говоря, близость решений дифференциального и разностного уравнений. Свойство устойчивости разностной схемы является ее внутренним свойством, не зависящим от того, аппроксимирует ли эта схема какое-либо дифференциальное уравнение (см. [30]). Оказывается, однако, что если разностная схема аппроксимирует корректно поставленную задачу и устойчива, то ее решение сходится при $|h|\to 0$ к решению исходной дифференциальной задачи.

Будем считать, что решение $u(x)$ задачи (1) принадлежит линейному нормированному пространству $\mathscr{B}_0$, $\|\cdot\|_0$ — норма в $\mathscr{B}_0$. Например, $\mathscr{B}_0=C[a, b]$, $\|u\|_0=\max\limits_{x\in[a,b]} |u(x)|$. Аналогично считаем, что сеточные функции $y_h(x)$, $\varphi_h(x)$ являются элементами линейного нормированного пространства $\mathscr{B}_h$ (*пространства сеточных функций*) с нормой $\|\cdot\|_h$. По существу, имеем семейство линейных нормированных пространств, зависящее от параметра h.

Чтобы иметь возможность сравнивать функции из различных пространств, вводится *оператор проектирования* p_h: $\mathscr{B}_0\to\mathscr{B}_h$. Это, по определению, линейный оператор, сопоставляющий каждой функции из $\mathscr{B}_0$ некоторую функцию из $\mathscr{B}_h$. Для функции $u\in\mathscr{B}_0$ обозначим через u_h ее проекцию на пространство $\mathscr{B}_h$, т. е. $u_h(x)=p_h u(x)$.

Приведем примеры операторов проектирования.

Пример 3. Пусть $\mathscr{B}_0$ — пространство непрерывных функций на $[0, 1]$ и G_h — равномерная сетка с шагом h:

$$G_h=\{x_i=ih, \ i=0, 1, \ldots, N, \ hN=1\}.$$

Тогда в качестве оператора проектирования можно взять оператор вычисления значения функции в данной точке сетки. Этот оператор определяется следующим образом:

$$(p_h u)(x_i) = u(x_i),\ i=0,\ 1,\ \ldots,\ N.$$

Пример 4. Пусть $\mathscr{B}_0$ — пространство функций, интегрируемых на $[0,\ 1]$, и G_h — та же сетка, что и в предыдущем примере. Тогда в качестве оператора проектирования можно взять оператор осреднения

$$(p_h u)(x_i) = \frac{1}{h}\int\limits_{x_i-0,5h}^{x_i+0,5h} u(x)\,dx, \qquad i=1,\ 2,\ \ldots,\ N-1,$$

$$(p_h u)(x_0) = \frac{1}{0,5h}\int\limits_{0}^{0,5h} u(x)\,dx,\ (p_h u)(x_N) = \frac{1}{0,5h}\int\limits_{x_N-0,5h}^{x_N} u(x)\,dx.$$

В дальнейшем будем требовать, чтобы нормы в $\mathscr{B}_h$ были *согласованы* с нормой в исходном пространстве $\mathscr{B}_0$. Это означает, что для любой $u\in\mathscr{B}_0$ выполняется условие

$$\lim_{|h|\to 0}\|p_h u\|_h = \|u\|_0. \tag{3}$$

Требование согласования норм обеспечивает единственность предела сеточных функций при $|h|\to 0$. Действительно, если для $u,\ v\in\mathscr{B}_0$ имеем $\lim\limits_{|h|\to 0}\|y_h-p_h u\|_h=0$, $\lim\limits_{|h|\to 0}\|y_h-p_h v\|_h=0$, то согласно (3)

$$\|p_h u-p_h v\|_h=\|(p_h u-y_h)+(y_h-p_h v)\|_h\leqslant\|p_h u-y_h\|_h+\|y_h-p_h v\|_h$$

и

$$\|u-v\|_0=\lim_{|h|\to 0}\|p_h(u-v)\|_h=0,$$

т. е. $u=v$.

Пример 5. Сеточная норма

$$\|y\|_h=\left(\sum_{i=0}^{N} h\,|y_i|^2\right)^{1/2}, \qquad hN=1,$$

согласована с нормой в L_2

$$\|y\|=\left(\int\limits_0^1 |y(x)|^2\,dx\right)^{1/2}.$$

Сеточная норма

$$\|y\|_h=\left(\sum_{i=0}^{N}|y_i|^2\right)^{1/2}, \qquad hN=1$$

не согласована ни с одной из норм для функций непрерывного аргумента, так как ряд $\sum\limits_{i=0}^{\infty}|y_i|^2$ может расходиться. Норма

$$\|y\|_h=\max_{0\leqslant i\leqslant N}|y_i|$$

согласована с нормой в C.

Пусть $u(x)$ — решение исходной задачи (1) и $y_h(x)$ — решение разностной задачи (2).

Определение 1. Сеточная функция $z_h(x)=y_h(x)-p_hu(x)$, $x\in G_h$, называется *погрешностью разностной схемы* (2).

Подставим $y_h(x)=p_hu(x)+z_h(x)$ в уравнение (2). Тогда получим, что погрешность $z_h(x)$ удовлетворяет уравнению

$$L_hz_h(x)=\psi_h(x), \quad x\in G_h, \tag{4}$$

где

$$\psi_h(x)=\varphi_h(x)-L_h(p_hu(x))\equiv\varphi_h(x)-L_hu_h(x). \tag{5}$$

Определение 2. Сеточная функция $\psi_h(x)$, определенная формулой (5), называется *погрешностью аппроксимации разностной задачи* (2) *на решении исходной дифференциальной задачи* (1).

Преобразуем выражение для $\psi_h(x)$. Проектируя уравнение (1) на сетку G_h, получим

$$p_hLu(x)=p_hf(x)$$

или, учитывая принятые обозначения,

$$(Lu)_h(x)=f_h(x). \tag{6}$$

Из (5) и (6) получаем

$$\psi_h(x)=[(Lu)_h(x)-L_hu_h(x)]+(\varphi_h(x)-f_h(x)),$$

т. е.

$$\psi_h(x)=\psi_{h,1}(x)+\psi_{h,2}(x),$$

где

$$\psi_{h,1}(x)=(Lu)_h(x)-L_hu_h(x), \quad \psi_{h,2}=\varphi_h(x)-f_h(x). \tag{7}$$

Определение 3. Функции $\psi_{h,1}(x)$ и $\psi_{h,2}(x)$ называются, соответственно, *погрешностью аппроксимации дифференциального оператора L разностным оператором L_h* и *погрешностью аппроксимации правой части.*

Определение 4. Говорят, что разностная задача (2) *аппроксимирует* исходную задачу (1), если $\|\psi_h\|_h\to0$ при $|h|\to0$. Разностная схема имеет *k-й порядок аппроксимации*, если существуют постоянные $k>0$, $M_1>0$, не зависящие от h и такие, что

$$\|\psi_h\|_h\leqslant M_1|h|^k.$$

Аналогично определяются погрешность аппроксимации и порядок погрешности аппроксимации правых частей и дифференциального оператора.

Замечание. Мы видели, что погрешность аппроксимации на решении представляется в виде суммы погрешностей аппроксимации дифференциального оператора и правой части. Однако порядок погрешности аппроксимации на решении ψ может оказаться выше, чем порядок погрешности аппроксимации оператора ψ_1 и правой части ψ_2 в отдельности. Нетрудно, например, показать, что разностное уравнение

$$y_{\bar{x}x,i}=-\varphi_i, \quad \varphi_i=f_i+\frac{h^2}{12}f_i''$$

имеет четвертый порядок аппроксимации на решении дифференциального уравнения

$$u''(x)=-f(x),$$

хотя дифференциальный оператор и правая часть аппроксимируются лишь со вторым порядком.

3. Корректность разностной схемы. Сходимость. Связь между устойчивостью и сходимостью. По аналогии с дифференциальным случаем вводится понятие корректности разностной задачи.

Определение 5. Разностная схема (2) называется *корректной*, если 1) ее решение существует и единственно при любых правых частях $\varphi_h \in \mathcal{B}_h$ и 2) существует постоянная $M_2>0$, не зависящая от h и такая, что при любых $\varphi_h \in \mathcal{B}_h$ справедлива оценка

$$\|y_h\|_h \leqslant M_2 \|\varphi_h\|_h. \tag{8}$$

Свойство 2), означающее непрерывную зависимость, равномерную относительно h, решения разностной задачи от правой части, называется *устойчивостью* разностной схемы. Заметим, что требование 1) эквивалентно существованию оператора L_h^{-1}, обратного оператору L_h, а требование 2) эквивалентно равномерной по h ограниченности оператора L_h^{-1}.

Основным вопросом теории разностных схем, как впрочем и других приближенных методов, является вопрос о сходимости. Сформулируем строго понятие сходимости.

Определение 6. Решение разностной задачи (2) *сходится* к решению дифференциальной задачи (1), если при $|h|\to 0$

$$\|y_h - p_h u\|_h \to 0.$$

Разностная схема имеет *k-й порядок точности*, если существуют постоянные $k>0$, $M_3>0$, не зависящие от h и такие, что

$$\|y_h - p_h u\|_h \leqslant M_3 |h|^k.$$

Часто для краткости просто говорят «разностная схема сходится», подразумевая сходимость решения разностной задачи к решению дифференциальной задачи.

Справедлива следующая теорема о связи устойчивости и сходимости.

Пусть дифференциальная задача (1) *поставлена корректно, разностная схема* (2) *является корректной и аппроксимирует исходную задачу* (1). *Тогда решение разностной задачи* (2) *сходится к решению исходной задачи* (1), *причем порядок точности совпадает с порядком аппроксимации.*

Доказательство следует прямо из определений. Действительно, уравнение для погрешности (4) имеет ту же структуру, что и разностная задача (2). Поэтому из требования корректности следует оценка

$$\|z_h\|_h \leqslant M_2 \|\psi_h\|_h. \tag{9}$$

Поскольку константа M_2 не зависит от h, получаем, что при $\|\psi_h\|_h \to$

$\to 0$ норма погрешности z_h также стремится к нулю, т. е. схема сходится. Если $\|\psi_h\|_h \leqslant M_1 |h|^k$, то из (9) получим

$$\|z_h\|_h \leqslant M_1 M_2 |h|^k,$$

т. е. разностная схема имеет k-й порядок точности.

Значение приведенной выше теоремы состоит в том, что она позволяет разделить изучение сходимости на два отдельных этапа: доказательство аппроксимации и доказательство устойчивости. Обычно более сложным этапом является исследование устойчивости, которое состоит в получении оценок вида (8), называемых *априорными оценками.*

Замечание. Теорема доказана в предположении, что решение y_h и правая часть φ_h измеряются в одной и той же норме. Однако, изменив соответствующие определения, можно легко показать, что теорема остается справедливой и в том случае, когда решение измеряется в одной норме, а правая часть — в другой (см., например, [32]).

ГЛАВА 2

ПРИНЦИП МАКСИМУМА ДЛЯ РАЗНОСТНЫХ СХЕМ

В § 1, 3 изучается разностная схема для уравнения Пуассона в двумерной прямоугольной области. В § 1 формулируется разностная краевая задача и вводится каноническая форма записи разностных схем, удобная для применения принципа максимума. Такая каноническая форма пригодна не только для разностного уравнения Пуассона, но и вообще для любого линейного разностного уравнения. В § 2 излагаются основные теоремы принципа максимума для разностных схем, записанных в канонической форме. В § 3 принцип максимума применяется к исследованию сходимости разностной аппроксимации задачи Дирихле. В § 4, 5 приводятся примеры применения принципа максимума к другим стационарным и нестационарным разностным задачам.

§ 1. Разностная аппроксимация задачи Дирихле для уравнения Пуассона

1. Постановка разностной задачи. Рассмотрим задачу Дирихле для уравнения Пуассона: найти непрерывную в $\overline{G} = G \cup \Gamma$ функцию $u(x_1, x_2)$, удовлетворяющую уравнению

$$\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = -f(x_1, x_2), \quad x = (x_1, x_2) \in G, \tag{1}$$

и граничному условию

$$u(x) = \mu(x), \quad x \in \Gamma,$$

где G — прямоугольник,

$$G=\{0<x_1<l_1,\ 0<x_2<l_2\},$$

Γ — его граница, $f(x)$, $\mu(x)$ — заданные функции. Предполагаем, что $f(x)$, $\mu(x)$ таковы, что решение задачи (1) существует, единственно и является достаточно гладкой функцией. При $f\equiv 0$ получаем задачу Дирихле для уравнения Лапласа

$$\frac{\partial^2 u}{\partial x_1^2}+\frac{\partial^2 u}{\partial x_2^2}=0,\quad x\in G;\qquad u(x)=\mu(x),\quad x\in\Gamma. \tag{2}$$

Одним из основных свойств задачи (2) является выполнение принципа максимума: непрерывное в $\bar{G}$ и отличное от константы решение $u(x_1, x_2)$ может достигать своего максимального по модулю значения только на границе Γ. Отсюда следует, что справедлива оценка

$$\max_{(x_1,x_2)\in\bar{G}}|u(x_1, x_2)|\leqslant\max_{(x_1,x_2)\in\Gamma}|\mu(x_1, x_2)|,\quad \bar{G}=G\cup\Gamma,$$

означающая устойчивость задачи (2) по граничным данным. В следующих параграфах аналогичные оценки будут получены и для некоторых разностных схем, аппроксимирующих уравнение (2). Эти оценки помогут установить сходимость разностной схемы для уравнения Пуассона (1).

Введем в $\bar{G}$ прямоугольную сетку Ω с шагами h_1 по направлению x_1 и h_2 — по направлению x_2, так что $h_1=l_1/N_1$, $h_2=l_2/N_2$, где N_1 и N_2 — целые числа. Обозначим $x_1^i=ih_1$, $x_2^j=jh_2$. Сетка Ω состоит из совокупности узлов $x_{ij}=(x_1^i, x_2^j)$, $i=0, 1, \ldots, N_1$, $j=0, 1, \ldots, N_2$. Для функций y, определенных на Ω, обозначим

$$y_{ij}=y(x_{ij}),\ y_{\bar{x}_1x_1,ij}=(y_{i+1,j}-2y_{ij}+y_{i-1,j})/h_1^2,$$

$$y_{\bar{x}_2x_2,ij}=(y_{i,j+1}-2y_{ij}+y_{i,j-1})/h_2^2.$$

Задаче Дирихле (1) сопоставим следующую разностную схему:

$$y_{\bar{x}_1x_1,ij}+y_{\bar{x}_2x_2,ij}=-f_{ij},\quad i=1, 2, \ldots, N_1-1,\quad j=1, 2, \ldots, N_2-1, \tag{3}$$

$$y_{i,0}=\mu(x_1^i, 0),\quad y_{i,N_2}=\mu(x_1^i, l_2),\qquad i=1, 2, \ldots, N_1-1, \tag{4}$$

$$y_{0,j}=\mu(0, x_2^j),\quad y_{N_1,j}=\mu(l_1, x_2^j),\qquad j=1, 2, \ldots, N_2-1.$$

Точки x_{ij}, в которых записываются уравнения (3), принадлежат подмножеству

$$\omega=\{x_{ij}\,|\,i=1, 2, \ldots, N_1-1,\ j=1, 2, \ldots, N_2-1\}$$

сетки Ω, называемому *множеством внутренних точек сетки* Ω. Совокупность точек

$$\gamma=\{x_{0j}, x_{N_1j}\}_{j=1}^{N_2-1}\cup\{x_{i0}, x_{iN_2}\}_{i=1}^{N_1-1},$$

в которых заданы разностные граничные условия (4), называется *границей сетки* Ω. На рис. 12 внутренние точки отмечены кружочками, а граничные — крестиками. Отметим, что угловые точки $(0, 0)$, $(l_1, 0)$, $(0, l_2)$, (l_1, l_2) не участвуют в данной аппроксимации и поэтому не относятся ни к внутренним, ни к граничным точкам.

По поводу разностной схемы (3), (4) можно задать обычные вопросы о существовании и единственности ее решения, о сходимости при $h_1 \to 0$, $h_2 \to 0$, о способах решения. Эти вопросы рассматриваются в следующих параграфах. Здесь мы ограничимся лишь очевидными замечаниями о том, что построенная разностная схема имеет второй порядок погрешности аппроксимации по h_1 и по h_2 и что она представляет собой систему линейных алгебраических уравнений относительно y_{ij}, состоящую из $(N_1-1)\times(N_2-1)$ уравнений и стольких же неизвестных.

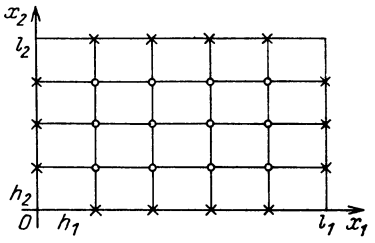


Рис. 12. Прямоугольная сетка

2. Канонический вид разностного уравнения. Для дальнейшего исследования удобно записать уравнение (3) в виде, разрешенном относительно y_{ij}, а точнее в виде

$$\left(\frac{2}{h_1^2}+\frac{2}{h_2^2}\right)y_{ij}=\frac{y_{i+1,j}+y_{i-1,j}}{h_1^2}+\frac{y_{i,j+1}+y_{i,j-1}}{h_2^2}+f_{ij}. \tag{5}$$

Обозначим через x точку x_{ij} — центральную точку шаблона, на котором аппроксимируется уравнение (1), а через $Ш(x)$ — весь этот шаблон, т. е. совокупность пяти точек x_{ij}, $x_{i\pm1,j}$, $x_{i,j\pm1}$. Назовем *окрестностью точки* x и обозначим через $Ш'(x)$ все точки шаблона $Ш(x)$ за исключением точки x, т. е. $Ш'(x)$ — это четыре точки $x_{i\pm1,j}$, $x_{i,j\pm1}$. Тогда уравнение (5) можно записать в виде

$$A(x)\,y(x)=\sum_{\xi\in Ш'(x)}B(x,\xi)\,y(\xi)+F(x), \tag{6}$$

где коэффициенты $A(x)$, $B(x, \xi)$ и правая часть $F(x)$ определены следующим образом:

$$A(x)=\frac{2}{h_1^2}+\frac{2}{h_2^2},\quad B(x, x_{i\pm1,j})=\frac{1}{h_1^2},$$
$$B(x, x_{i,j\pm1})=\frac{1}{h_2^2},\quad F(x)=f(x_{ij}). \tag{7}$$

Обратим внимание на свойства этих коэффициентов: $A(x)>0$, $B(x, \xi)>0$, $A(x)=\sum\limits_{\xi\in Ш'(x)}B(x, \xi)$. Запись разностного уравнения в виде (6) называется *канонической формой разностного уравне-*

ния. Она применима не только к уравнению (5), но и к любому линейному разностному уравнению. Разумеется, в каждом конкретном случае необходимо задать множества $Ш'(x)$ и определить коэффициенты $A(x)$, $B(x, \xi)$ и правую часть $F(x)$.

Уравнение (6) определено при $x \in \omega$, т. е. только во внутренних точках сетки Ω. Поэтому к нему требуется добавить еще граничные условия (4). Заметим, однако, что если при $x \in \gamma$ считать $Ш'(x)$ пустым множеством, то уравнение (6) принимает вид $A(x)y(x)=F(x)$, $x \in \gamma$, и представляет собой запись граничных условий $y(x)=\mu(x)$ при $x \in \gamma$, причем $F(x)=A(x)\mu(x)$.

Итак, разностную схему (3), (4) можно записать в виде системы уравнений (6), где x пробегает все множество Ω. Во всех внутренних точках $x \in \omega$ выполняются условия

$$A(x)>0, \quad B(x, \xi)>0 \text{ для всех } \xi \in Ш'(x), \tag{8}$$

$$A(x)=\sum_{\xi \in Ш'(x)} B(x, \xi).$$

В граничных точках $x \in \gamma$ имеем $Ш'(x)$ — пустое множество и $A(x)>0$.

В следующем параграфе изучаются свойства общей разностной схемы (6) безотносительно к конкретному виду ее коэффициентов, важно лишь, чтобы выполнялись условия, аналогичные (8). Полученные выводы окажется возможным применить не только к разностной схеме (3), (4), но и к более широким классам разностных схем.

§ 2. Принцип максимума для разностных схем. Основные теоремы

1. Исходные предположения. В предыдущем параграфе на примере уравнения Пуассона была введена каноническая форма записи разностной схемы

$$A(x)y(x)=\sum_{\xi \in Ш'(x)} B(x, \xi)y(\xi)+F(x), \qquad x \in \Omega. \tag{1}$$

Поясним теперь, как следует понимать уравнение (1) в общем случае. Пусть в n-мерном евклидовом пространстве задано конечное множество точек — сетка Ω. Каждой точке $x \in \Omega$ сопоставим один и только один шаблон $Ш(x)$ — любое подмножество Ω, содержащее данную точку x. Окрестностью точки x назовем множество $Ш'(x)=Ш(x) \setminus \{x\}$. Заметим, что $Ш'(x)$ может быть и пустым множеством. Пусть заданы функции $A(x)$, $B(x, \xi)$, $F(x)$, определенные при любых $x \in \Omega$, $\xi \in \Omega$ и принимающие вещественные значения. Далее, каждой точке $x \in \Omega$ соотносится одно и только одно уравнение вида (1), в котором $y(x)$ — искомая сеточная функция. В результате получаем систему линейных алгебраических уравнений с числом уравнений, равным числу неизвестных. Эту систему уравнений и будем называть разностной схемой.

Введем понятие связной сетки. Сетку Ω будем называть *связной* сеткой, если для любых двух ее узлов x_0, x_0' таких, что по крайней мере один из узлов имеет непустую окрестность, существует такое множество узлов $x_i \in \Omega$, $i = 1, 2, \ldots, m$, что $x_1 \in Ш'(x_0)$, $x_2 \in Ш'(x_1)$, $\ldots$, $x_m \in Ш'(x_{m-1})$, $x_0' \in Ш'(x_m)$, т. е. каждый последующий узел принадлежит окрестности предыдущего. Аналогичным образом определяется понятие связности любого подмножества из Ω. Наглядный смысл требования связности состоит в том, чтобы от любого узла $x_0 \in \Omega$ можно было перейти к любому другому узлу $x_0' \in \Omega$, пользуясь только заданными шаблонами.

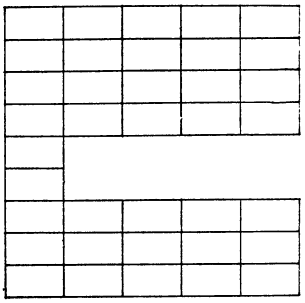


Рис. 13. Несвязная сетка

На рис. 13 изображен пример сеточной области, не являющейся связной (шаблон предполагается пятиточечным, таким, как при аппроксимации уравнения Пуассона).

Определим сеточный оператор L формулами

$$Ly(x) = A(x)\,y(x) - \sum_{\xi \in Ш'(x)} B(x, \xi)\,y(\xi) \tag{2}$$

и обозначим

$$D(x) = A(x) - \sum_{\xi \in Ш'(x)} B(x, \xi). \tag{3}$$

Тогда задачу (1) можно записать в виде

$$Ly(x) = F(x), \quad x \in \Omega. \tag{4}$$

Заметим, что выражение $Ly(x)$ можно представить также в виде

$$Ly(x) = D(x)\,y(x) + \sum_{\xi \in Ш'(x)} B(x, \xi)\,(y(x) - y(\xi)).$$

Будем говорить, что в точке $x \in \Omega$ выполнены *условия положительности коэффициентов*, если

$$A(x) > 0, \quad B(x, \xi) > 0 \text{ для всех } \xi \in Ш'(x), \quad D(x) \geqslant 0. \tag{5}$$

2. Принцип максимума и его следствия. Сформулируем теперь основную теорему настоящего параграфа (см. [33]). Наряду с сеткой Ω будем рассматривать какое-либо ее подмножество ω и обозначим

$$\overline{\omega} = \bigcup_{x \in \omega} Ш(x).$$

Для наглядности читатель может представить себе, что Ω — это сетка, введенная в § 1 при аппроксимации уравнения Пуассона в прямоугольнике, а ω — множество ее внутренних узлов. Очевидно,

что при этом $\omega=\Omega$. В общем же случае требуемые свойства множеств Ω и ω сформулированы в приведенной ниже теореме. Заметим, что в этой теореме функция $y(x)$ не обязана являться решением задачи (4), используются только свойства оператора L.

Теорема 1 (принцип максимума). *Пусть сетка Ω и ее подмножество ω являются связными, причем $\bar{\omega}\subseteq\Omega$. Пусть в ω выполнены условия положительности коэффициентов* (5). *Тогда, если функция $y(x)$, заданная на Ω, не является постоянной на $\bar{\omega}$ и*

$$Ly(x)\leqslant 0 \text{ при всех } x\in\omega \tag{6}$$

(либо $Ly(x)\geqslant 0$ при всех $x\in\omega$), то $y(x)$ не может принимать наибольшего положительного (соответственно наименьшего отрицательного) значения на ω среди всех ее значений на $\bar{\omega}$.

Доказательство. Пусть выполнено условие (6). Будем доказывать теорему от противного. Допустим, что в точке $x_0\in\omega$ функция $y(x)$ принимает наибольшее положительное значение, т. е.

$$y(x_0)=\max_{x\in\bar{\omega}} y(x)>0. \tag{7}$$

В этой точке выражение

$$Ly(x_0)=D(x_0)\,y(x_0)+\sum_{\xi\in Ш'(x_0)} B(x_0,\xi)\,(y(x_0)-y(\xi)) \tag{8}$$

неотрицательно. Действительно, согласно условиям (5) и предположению (7) имеем $D(x_0)\geqslant 0$, $y(x_0)>0$, $B(x_0,\xi)>0$, $y(x_0)\geqslant y(\xi)$, так что $Ly(x_0)\geqslant 0$. С другой стороны, из условия (6) следует, что $Ly(x_0)\leqslant 0$. Таким образом, если выполнено (7) в точке $x_0\in\omega$, то $Ly(x_0)=0$. Но тогда, учитывая неотрицательность всех слагаемых правой части выражения (8), получим

$$D(x_0)y(x_0)=0,\quad B(x_0,\xi)\,(y(x_0)-y(\xi))=0,\quad \xi\in Ш'(x_0).$$

Отсюда, в силу предположения $y(x_0)>0$ и условия $B(x_0,\xi)>0$ следует

$$y(\xi)=y(x_0) \text{ для всех } \xi\in Ш'(x_0). \tag{9}$$

Далее, поскольку $y(x)\not\equiv\text{const}$ в $\bar{\omega}$, найдется точка $x_0'\in\bar{\omega}$, в которой $y(x_0')<y(x_0)$. Из предположения о связности сетки ω вытекает существование системы узлов $x_1, x_2, \ldots, x_m$, принадлежащих ω и удовлетворяющих условиям

$$x_1\in Ш'(x_0),\ x_2\in Ш'(x_1),\ \ldots,\ x_m\in Ш'(x_{m-1}),$$
$$x_0'\in Ш'(x_m).$$

Из условия (7) и доказанного свойства (9) получаем $y(x_1)=y(x_0)$. Следовательно, относительно точки x_1 можно повторить все предыдущие рассуждения и доказать, что

$$y(\xi)=y(x_1) \text{ для всех } \xi\in Ш'(x_1).$$

Аналогично докажем, что

$$y(x_1)=y(x_2)=\ldots=y(x_m)=y(x_0).$$

Оценим величину

$$Ly(x_m)=D(x_m)\,y(x_m)+\sum_{\xi\in Ш'(x_m)} B(x_m,\xi)\,(y(x_m)-y(\xi)).$$

Из условий (5), равенства $y(x_m)=y(x_0)$ и предположения (7) получаем строгое неравенство

$$Ly(x_m)\geqslant B(x_m,x_0')\,(y(x_0)-y(x_0''))>0,$$

которое противоречит условию (6). Таким образом, допущение (7) неверно. Случай, когда $Ly(x)\geqslant 0$, для всех $x\in\omega$ сводится к рассмотренному случаю путем замены y на $-y$. Теорема 1 доказана.

Замечание. Принцип максимума остается справедливым и в том случае, когда $\omega=\Omega$. Предполагается при этом, что $\overline{\omega}=\bigcup\limits_{x\in\omega} Ш(x)=\Omega$. В дальнейшем, не оговаривая это особо, будем считать сетку Ω связной.

Следствие 1. *Если при всех* $x\in\Omega$

а) *выполнены условия положительности коэффициентов* (5),

б) $Ly(x)\leqslant 0\,(Ly(x)\geqslant 0)$, *и найдется хотя бы один узел* $x_0\in\Omega$, в котором

$$D(x_0)>0,\quad x_0\in\Omega, \tag{10}$$

то $y(x)\leqslant 0\,(y(x)\geqslant 0)$ *для всех* $x\in\Omega$.

Доказательство. Если $y(x)\not\equiv\text{const}$ при $x\in\Omega$, то утверждение следует из принципа максимума. Действительно, предполагая, что $y(x)>0$ хотя бы в одной точке $x\in\Omega$, мы допускаем существование в Ω положительного максимума функции $y(x)$, что противоречит принципу максимума. Если $y(x)\equiv\text{const}$ при $x\in\Omega$, то в точке x_0, для которой $D(x_0)>0$, имеем

$$Ly(x_0)=D(x_0)\,y(x_0)+\sum_{\xi\in Ш'(x_0)} B(x_0,\xi)\,(y(x_0)-y(\xi))=D(x_0)\,y(x_0)\leqslant 0,$$

откуда получим $y(x)\equiv y(x_0)\leqslant 0$.

Точку $x\in\Omega$ назовем *граничной точкой*, если $Ш'(x)$ — пустое множество, $Ш'(x)=\varnothing$. Если сетка Ω содержит хотя бы одну граничную точку x_0, то

$$D(x_0)=A(x_0)-\sum_{\xi\in Ш'(x_0)} B(x,\xi)=A(x_0)>0$$

и можно применять следствие 1.

Теперь мы уже в состоянии сформулировать достаточные условия однозначной разрешимости задачи (1).

Следствие 2. *Пусть коэффициенты оператора* L *удовлетворяют условиям* (5) *при каждом* $x\in\Omega$ *и условию* (10). *Тогда задача* (1) *имеет единственное решение.*

Доказательство. Ранее отмечалось, что задача (1) представляет собой систему линейных алгебраических уравнений, в которой число уравнений равно числу неизвестных. Поэтому достаточно показать, что однородное уравнение $Ly(x)=0$, $x\in\Omega$, имеет только тривиальное решение $y(x)\equiv 0$. Поскольку условия $Ly(x)\geqslant 0$ и $Ly(x)\leqslant 0$ в данном случае выполнены, из следствия 1 заключаем, что в каждой точке $x\in\Omega$ одновременно выполняются неравенства $y(x)\geqslant 0$ и $y(x)\leqslant 0$. Но это справедливо лишь тогда, когда $y(x)\equiv 0$ на Ω.

Предоставляем читателю возможность самостоятельно убедиться в том, что разностная схема, аппроксимирующая задачу Дирихле для уравнения Пуассона (см. § 1), удовлетворяет всем условиям теоремы 1 и ее следствий и тем самым имеет единственное решение. Для того чтобы доказать непрерывную зависимость решения от правой части и от граничных условий, полученной теоремы недостаточно. Докажем еще несколько утверждений, следующих из принципа максимума.

3. Теорема сравнения. Устойчивость по граничным условиям. Наряду с задачей (4) рассмотрим задачу

$$LY(x)=\overline{F}(x), \quad x\in\Omega, \tag{11}$$

отличающуюся от (4) правой частью.

Теорема 2 (теорема сравнения). *Пусть при всех* $x\in\Omega$ *выполнены условия положительности коэффициентов* (5) *и выполнено условие* (10). *Тогда, если*

$$|F(x)|\leqslant\overline{F}(x) \text{ для всех } x\in\Omega,$$

то $|y(x)|\leqslant Y(x)$ *для всех* $x\in\Omega$.

Доказательство. Для функций $v(x)=Y(x)+y(x)$ и $w(x)=Y(x)-y(x)$ имеем

$$Lv(x)=\overline{F}(x)+F(x)\geqslant 0, \quad Lw(x)=\overline{F}(x)-F(x)\geqslant 0.$$

Согласно следствию 1 имеем $v(x)\geqslant 0$, $w(x)\geqslant 0$, т. е.

$$-Y(x)\leqslant y(x)\leqslant Y(x),$$

что и требовалось.

Сформулируем первую краевую задачу для уравнения (1). Пусть γ — множество граничных точек сетки Ω, т. е. точек $x\in\Omega$, для которых $Ш'(x)=\varnothing$. Множество точек сетки Ω, не являющихся граничными, назовем множеством внутренних узлов и обозначим через ω. Таким образом, $\Omega=\omega\cup\gamma$. В граничном узле $x\in\gamma$ уравнение (1) принимает вид

$$A(x)y(x)=F(x)$$

или, что то же самое,

$$y(x)=\mu(x), \tag{12}$$

где $\mu(x)=F(x)/A(x)$ — заданная функция. Первая краевая задача состоит в том, чтобы найти сеточную функцию $y(x)$, удовлетворяющую уравнению (1) при $x\in\omega$ и условию (12) при $x\in\gamma$. Уже

отмечалось, что при условиях теоремы 1 первая краевая задача имеет единственное решение.

Переформулируем теорему сравнения на случай первой краевой задачи.

Рассмотрим две задачи:

$$Ly(x)=F(x), \quad x\in\omega; \quad y(x)=\mu(x), \quad x\in\gamma, \tag{13}$$

$$LY(x)=\overline{F}(x), \quad x\in\omega; \quad Y(x)=\overline{\mu}(x), \quad x\in\gamma. \tag{14}$$

Если при $x\in\omega$ *выполнены условия* (5) *и*

$$|F(x)|\leqslant\overline{F}(x), \quad x\in\omega, \quad |\mu(x)|\leqslant\overline{\mu}(x), \quad x\in\gamma,$$

то

$$|y(x)|\leqslant Y(x) \text{ при всех } x\in\Omega.$$

Функция $Y(x)$, фигурирующая в теореме сравнения, называется *мажорантной функцией* для решения $y(x)$ задачи (4). Для получения оценки решения $y(x)$ обычно строят вспомогательную задачу (11) или (14) так, чтобы можно было легко найти ее решение $Y(x)$ и затем применяют теорему сравнения.

Теорема сравнения позволяет легко доказать устойчивость решения первой краевой задачи по граничным условиям. Рассмотрим однородное уравнение (13) с неоднородным граничным условием

$$Ly(x)=0, \quad x\in\omega; \quad y(x)=\mu(x), \quad x\in\gamma. \tag{15}$$

Следствие 3 (устойчивость по граничным условиям). *Пусть при* $x\in\omega$ *выполнены условия* (5). *Тогда для решения задачи* (15) *справедлива оценка*

$$\max_{x\in\omega}|y(x)|\leqslant\max_{x\in\gamma}|\mu(x)|. \tag{16}$$

Доказательство. Наряду с задачей (15) рассмотрим задачу

$$LY(x)=0, \quad x\in\omega; \quad Y(x)=\alpha, \quad x\in\gamma, \tag{17}$$

где $\alpha=\max\limits_{x\in\gamma}|\mu(x)|$. Все условия теоремы сравнения выполнены, поэтому $|y(x)|\leqslant Y(x)$.

Далее, для функции $v(x)=\alpha-Y(x)$ имеем

$$Lv(x)=L\alpha-LY(x)=L\alpha=D(x)\alpha\geqslant 0$$

и $v(x)=0$ при $x\in\omega$ и, согласно следствию 1, получим $v(x)\geqslant 0$, т. е. $Y(x)\leqslant\alpha$. Но тогда при всех $x\in\Omega$ имеем $|y(x)|\leqslant Y(x)\leqslant\alpha$, откуда и следует (16).

4. Примеры. Приведем несколько простых примеров.

Пример 1. Рассмотрим задачу

$$u''(x)=-f(x), \quad 0<x<1, \quad u'(0)=0, \quad u(1)=0.$$

По аналогии с § 2 гл. 1 построим разностную схему второго

порядка аппроксимации

$$y_{\bar{x}x,i} = -f(x_i),\ i=1, 2, \ldots, N-1,\ -y_{x,0} = 0{,}5hf_0,\ y_N = 0. \quad (18)$$

Запишем схему (18) в каноническом виде:

$$y_0 = y_1 + 0{,}5h^2 f_0,\ y_i = 0{,}5(y_{i-1} + y_{i+1}) + \frac{h^2}{2} f_i,\quad i=1, 2, \ldots, N-1,$$

$$y_N = 0.$$

Сетка Ω состоит из узлов $x_i = ih$, $i=0, 1, \ldots, N$, и имеет одну граничную точку $x=x_N$. Окрестность $Ш'(x_0)$ узла x_0 состоит из одного узла $x=x_1$. Окрестность $Ш'(x_i)$ узла x_i при $i=1, 2, \ldots, N-1$ состоит из двух узлов x_{i-1}, x_{i+1}. Сетка, очевидно, является связной. Свойства положительности коэффициентов (5) выполнены, причем $D(x_i)=0$ при $i=1, 2, \ldots, N-1$, $D(x_N)=A(x_N)=1$. Таким образом, к разностной схеме (18) можно применять принцип максимума и его следствия.

Пример 2. Для уравнения

$$u''(x) = -f(x),\quad 0<x<1,\quad u'(0)=u'(1)=0 \quad (19)$$

строится разностная схема

$$y_{\bar{x}x,i} = -f(x_i),\quad i=1, 2, \ldots, N-1,$$
$$-y_{x,0} = 0{,}5hf_0,\quad y_{\bar{x},N} = 0{,}5hf_N. \quad (20)$$

Канонический вид этой схемы

$$y_0 = y_1 + 0{,}5h^2 f_0,\quad y_i = 0{,}5(y_{i-1} + y_{i+1}) + 0{,}5h^2 f_i,$$
$$i=1, 2, \ldots, N-1,\ y_N = y_{N-1} + 0{,}5h^2 f_N.$$

Окрестности узлов x_0, x_N сетки Ω состоят каждая из одного узла, а окрестности точек x_i, $i=1, 2, \ldots, N-1$, — из двух узлов. Граничных точек сетка не имеет. Условия $A(x)>0$, $B(x, \xi)>0$, $D(x)=0$ выполнены в каждой точке сетки. Нет ни одной точки x_0 сетки Ω, в которой выполнялось бы строгое неравенство $D(x_0)>0$. Поэтому нельзя применять следствия 1 и 2 и утверждать о существовании и единственности решения задачи (20). И действительно, решение задачи (20) (так же как и исходной дифференциальной задачи (19)) не единственно: наряду с $y(x)$ решением является функция $v(x) = y(x) + \alpha$, где α — любая постоянная.

§ 3. Доказательство устойчивости и сходимости разностной задачи Дирихле для уравнения Пуассона

1. Устойчивость по граничным условиям. В § 1 рассматривалась задача Дирихле для уравнения Пуассона

$$\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = -f(x_1, x_2),\quad x = (x_1, x_2) \in G,$$
$$u(x) = \mu(x),\quad x \in \Gamma, \quad (1)$$

в прямоугольнике $G=\{0<x_1<l_1,\ 0<x_2<l_2\}$ с границей Г. Была введена сетка

$$\Omega=\{x_{ij}=(x_1^i,\ x_2^j),\ i=0,\ 1,\ \ldots,\ N_1,\ j=0,\ 1,\ \ldots,\ N_2\},$$

где $x_1^i=ih_1$, $x_2^j=jh_2$, $h_1=l_1/N_1$, $h_2=l_2/N_2$, и построена разностная схема второго порядка аппроксимации

$$y_{\overline{x}_1x_1,ij}+y_{\overline{x}_2x_2,ij}=-f_{ij},\quad x_{ij}\in\omega,$$
$$y_{ij}=\mu(x_1^i,\ x_2^j),\quad x_{ij}\in\gamma. \tag{2}$$

Здесь ω — множество внутренних узлов сетки Ω и γ — множество граничных узлов:

$$\omega=\{x_{ij}=(x_1^i,\ x_2^j),\quad i=1,\ 2,\ \ldots,\ N_1-1,\ j=1,\ 2,\ \ldots,\ N_2-1\},$$
$$\gamma=\left\{x_{0j},\ x_{N_1j}\right\}_{j=1}^{N_2-1}\cup\left\{x_{i0},\ x_{iN_2}\right\}_{i=1}^{N_1-1}.$$

Разностная схема (2) приводится к каноническому виду

$$A(x)y(x)=\sum_{\xi\in Ш'(x)}B(x,\xi)y(\xi)+F(x),\ x\in\omega, \tag{3}$$
$$y(x)=\mu(x),\quad x\in\gamma, \tag{4}$$

где для $x=x_{ij}\in\omega$ окрестность $Ш'(x)$ состоит из четырех узлов $x_{i\pm1,j}$, $x_{i,j\pm1}$ и

$$A(x)=\frac{2}{h_1^2}+\frac{2}{h_2^2},\quad B(x,\ x_{i\pm1,j})=\frac{1}{h_1^2},\quad B(x,\ x_{i,j\pm1})=\frac{1}{h_2^2}, \tag{5}$$
$$F(x)=f(x_{ij}).$$

Обозначая, как и ранее,

$$L(x)y(x)=A(x)y(x)-\sum_{\xi\in Ш'(x)}B(x,\xi)y(\xi),$$

запишем задачу (3), (4) в виде

$$L(x)y(x)=F(x),\quad x\in\omega,\quad y(x)=\mu(x),\quad x\in\gamma. \tag{6}$$

В настоящем параграфе будет получена оценка решения задачи (2) через правую часть f и граничные условия μ, означающая устойчивость этой задачи, и будет показано, что при $h_1\to0$, $h_2\to0$ норма погрешности

$$\|y-u\|_{C(\Omega)}=\max_{x\in\Omega}|y(x)-u(x)|$$

стремится к нулю. Тем самым будет доказана сходимость разностной схемы.

Запишем задачу (2) в виде (6) и представим ее решение $y(x)$ в виде суммы $y(x)=\tilde{y}(x)+\bar{y}(x)$, где $\tilde{y}(x)$ — решение однородного

уравнения с неоднородным граничным условием:

$$L\tilde{y}(x)=0, \quad x\in\omega, \quad \tilde{y}(x)=\mu(x), \quad x\in\gamma, \tag{7}$$

и $\bar{y}(x)$ — решение неоднородного уравнения с однородным граничным условием:

$$L\bar{y}(x)=F(x), \quad x\in\omega, \quad \bar{y}(x)=0, \quad x\in\gamma. \tag{8}$$

Отметим, что для задачи (6) выполняются все условия принципа максимума, поэтому можно воспользоваться результатами § 2. В частности, к задаче (7) можно применить следствие 3 из § 2, которое приводит к оценке

$$\|\tilde{y}\|_{C(\Omega)}\leqslant\|\mu\|_{C(\gamma)}, \tag{9}$$

где

$$\|y\|_{C(\Omega)}=\max_{x\in\Omega}|y(x)|, \quad \|\mu\|_{C(\gamma)}=\max_{x\in\gamma}|\mu(x)|.$$

2. Устойчивость по правой части и сходимость. Оценить решение неоднородного уравнения (8), пользуясь только результатами § 2, невозможно. Однако можно легко построить мажорантную функцию для решения задачи (8) и применить затем теорему сравнения. Рассмотрим функцию

$$Y(x)=K(l_1^2+l_2^2-x_1^2-x_2^2), \tag{10}$$

где K — пока произвольная положительная постоянная, а l_1, l_2 — длины сторон прямоугольника G. Ясно, что $Y(x)\geqslant 0$ при всех $x\in\Omega$. Обозначим

$$D(x)=A(x)-\sum_{\xi\in Ш'(x)}B(x,\xi)$$

и вычислим выражение

$$LY(x)=D(x)Y(x)+\sum_{\xi\in Ш(x)}B(x,\xi)(Y(x)-Y(\xi))$$

для функции (10) в любой точке $x\in\omega$. Заметим, что по построению

$$LY(x)=-Y_{\bar{x}_1x_1}-Y_{\bar{x}_2x_2}.$$

Кроме того, для функции (10) справедливы равенства

$$Y_{\bar{x}_1x_1}=\frac{\partial^2Y}{\partial x_1^2}=-2K, \quad Y_{\bar{x}_2x_2}=\frac{\partial^2Y}{\partial x_2^2}=-2K.$$

Таким образом, $LY(x)=4K$ и можно считать, что функция $Y(x)$ является решением краевой задачи

$$LY(x)=\bar{F}(x), \quad x\in\omega, \quad Y(x)=\bar{\mu}(x), \quad x\in\gamma, \tag{11}$$

где $\bar{F}(x)=4K$ и $\mu(x)\geqslant 0$ — значение функции (10) при $x\in\gamma$. Если положить

$$K=\frac{1}{4}\|F\|_{C(\omega)},$$

то по отношению к задачам (8), (11) будут выполнены все условия теоремы сравнения (см. аналогичные задачи (13), (14) в § 2). Из теоремы сравнения следует оценка

$$\|\bar{y}\|_{C(\Omega)} \leqslant \max_{x\in\Omega} Y(x) \leqslant K(l_1^2 + l_2^2).$$

Отсюда, учитывая выбор константы K, получим

$$\|\bar{y}\|_{C(\Omega)} \leqslant \frac{l_1^2 + l_2^2}{4} \|F\|_{C(\omega)}. \tag{12}$$

Из неравенства треугольника и оценок (9), (12) следует оценка решения задачи (2)

$$\|y\|_{C(\Omega)} \leqslant \frac{l_1^2 + l_2^2}{4} \|f\|_{C(\omega)} + \|\mu\|_{C(\gamma)}. \tag{13}$$

Поскольку константы, входящие в оценку (13), не зависят от шагов сетки h_1 и h_2, данная оценка выражает собой устойчивость разностной схемы по правой части f и по граничным условиям μ. Отметим геометрический смысл константы $l_1^2 + l_2^2$ — это квадрат диаметра области G.

Тем самым полностью доказана корректность (однозначная разрешимость и устойчивость) разностной схемы (2). Перейдем теперь к исследованию сходимости разностной схемы и к оценкам погрешности.

Обозначим $z_{ij} = y_{ij} - u(x_1^i, x_2^j)$, где y_{ij} — решение разностной задачи (2) и $u(x_1, x_2)$ — решение исходной дифференциальной задачи (1). Подставляя $y_{ij} = z_{ij} + u_{ij}$ в уравнение (2), получим, что погрешность удовлетворяет уравнению

$$\left.\begin{aligned} z_{\bar{x}_1 x_1, ij} + z_{\bar{x}_2 x_2, ij} &= -\psi_{ij}, \quad x_{ij} \in \omega, \\ z_{ij} &= 0, \quad x_{ij} \in \gamma, \end{aligned}\right\} \tag{14}$$

где $\psi_{ij} = u_{\bar{x}_1 x_1, ij} + u_{\bar{x}_2 x_2, ij} + f_{ij}$ — погрешность аппроксимации на решении задачи (1). Если четвертые производные решения $u(x_1, x_2)$ ограничены, то погрешность аппроксимации является величиной второго порядка малости относительно $h = (h_1^2 + h_2^2)^{1/2}$, т. е. существует постоянная M_1, не зависящая от h_1 и h_2 и такая, что

$$\|\psi\|_{C(\omega)} \leqslant M_1 (h_1^2 + h_2^2). \tag{15}$$

Заметим, что задача (14) отличается от разностной схемы (2) только правыми частями в основном уравнении и в граничных условиях. Поэтому для решения задачи (14) справедлива оценка, аналогичная (13), а именно оценка

$$\|z\|_{C(\Omega)} \leqslant \frac{l_1^2 + l_2^2}{4} \|\psi\|_{C(\omega)}.$$

Отсюда и из (15) получаем неравенство

$$\|z\|_{C(\Omega)} \leqslant M_2 (h_1^2 + h_2^2), \tag{16}$$

где $M_2 = 0{,}25 M_1 (l_1^2 + l_2^2)$ – постоянная, не зависящая от h_1 и h_2. Из оценки (16) и следует, что схема (2) сходится и имеет второй порядок точности.

§ 4. Примеры применения принципа максимума

В этом параграфе будем рассматривать разностные уравнения

$$Ly(x) = F(x), \quad x \in \Omega, \tag{1}$$

где

$$Ly(x) = A(x) y(x) - \sum_{\xi \in Ш'(x)} B(x, \xi) y(\xi), \tag{2}$$

причем оператор L удовлетворяет условию положительности коэффициентов

$$A(x) > 0, \quad B(x, \xi) > 0, \quad D(x) = A(x) - \sum_{\xi \in Ш'(x)} B(x, \xi) \geqslant 0. \tag{3}$$

Линейный оператор L называется *монотонным оператором*, если из условия $Ly(x) \geqslant 0$ для всех $x \in \Omega$ следует, что $y(x) \geqslant 0$ для всех $x \in \Omega$. Поэтому разностные схемы, удовлетворяющие при всех $x \in \Omega$ условиям (3), называются *монотонными разностными схемами*. Схемы, для которых условия (3) не выполнены хотя бы в одной точке $x \in \Omega$, называются *немонотонными*.

В § 2 было показано, что условия (3) обеспечивают монотонность оператора L, выполнение принципа максимума и корректность разностной задачи (1) в сеточной норме C:

$$\|y\|_{C(\Omega)} = \max_{x \in \Omega} |y(x)|.$$

Разумеется, отсюда не следует, что немонотонная схема обязательно некорректна. Подчеркнем, что выполнение условий (3) (наряду с другими условиями, сформулированными в теоремах из § 2) является достаточным условием корректности.

Приведем несколько примеров монотонных разностных схем для нестационарных уравнений.

Пример 1. Рассмотрим схемы с весами для уравнения теплопроводности

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 1, \quad 0 < t \leqslant T, \tag{4}$$

$$u(x, 0) = u_0(x), \quad u(0, t) = \mu_1(t), \quad u(1, t) = \mu_2(t).$$

Эти схемы подробно выписывались в § 4 из гл. 1 (см. схему (15) при $\varphi_i^n \equiv 0$ из § 4 гл. 1), поэтому мы не будем формулировать разностную задачу в полной постановке, а приведем только одно

уравнение

$$\frac{y_i^{n+1}-y_i^n}{\tau}=\sigma y_{\overline{x}x,i}^{n+1}+(1-\sigma)\,y_{\overline{x}x,i}^n. \tag{5}$$

Найдем, при каких значениях параметров τ, h, σ схема (5) будет монотонной. Чтобы записать уравнение (5) в виде (1), (2), разрешим его относительно y_i^{n+1}. Тогда получим уравнение

$$y_i^{n+1}=\sigma\gamma\,(y_{i+1}^{n+1}-2y_i^{n+1}+y_{i-1}^{n+1})+(1-\sigma)\,\gamma\,(y_{i+1}^n-2y_i^n+y_{i-1}^n)+y_i^n,$$
$$\gamma=\tau/h^2$$

или

$$(1+2\sigma\gamma)\,y_i^{n+1}=$$
$$=(1-2\,(1-\sigma)\,\gamma)\,y_i^n+\sigma\gamma\,(y_{i+1}^{n+1}+y_{i-1}^{n+1})+(1-\sigma)\,\gamma\,(y_{i+1}^n+y_{i-1}^n). \tag{6}$$

Отсюда видно, что в каждом узле $x=(x_i, t_{n+1})$ шаблон $Ш(x)$ состоит из шести точек, а окрестность $Ш'(x)$ точки (x_i, t_{n+1}) состоит из пяти точек $(x_{i\pm1}, t_{n+1})$, (x_i, t_n), $(x_{i\pm1}, t_n)$. Условия положительности коэффициентов (3) сводятся к неравенствам $0<\sigma<1$, $\sigma>1-1/(2\gamma)$. Заметим, что схема останется монотонной и в том случае, если эти неравенства заменить на нестрогие, т. е. потребовать

$$0\leqslant\sigma\leqslant1,\quad \sigma\geqslant1-1/(2\gamma). \tag{7}$$

Действительно, выполнение одного из условий (7) со знаком равенства означает лишь, что окрестность $Ш'(x)$ состоит не из пяти, а из меньшего числа узлов. Например, при $\sigma=0$ (явная схема) окрестность $Ш'(x)$ состоит из точек (x_i, t_n), $(x_{i\pm1}, t_n)$ и условие монотонности (7) принимает вид

$$\frac{\tau}{h^2}\leqslant\frac{1}{2}. \tag{8}$$

Если $\sigma=0$, $\tau/h^2=0{,}5$, то два из трех неравенств (7) выполнены со знаком равенства. В этом случае надо считать, что окрестность $Ш'(x)$ состоит из двух узлов $(x_{i\pm1}, t_n)$.

Итак, схема с весами (5) является монотонной при условиях (7), а чисто неявная схема ($\sigma=1$) монотонна при любых τ и h. Шеститочечная симметричная схема ($\sigma=0{,}5$) монотонна при условии $\tau\leqslant h^2$. В § 4 из гл. 1 отмечалось, что необходимым условием устойчивости схемы (5) является условие

$$\sigma\geqslant\frac{1}{2}-\frac{1}{4\gamma}. \tag{9}$$

Сопоставляя с (7), видим, что монотонность является, вообще говоря, более сильным требованием, чем просто устойчивость. В следующей главе будет показано, что условие (9) достаточно для устойчивости схемы (5), однако не в сеточной норме C, а в среднеквадратичной норме.

С помощью принципа максимума можно исследовать также устойчивость разностных схем с переменными коэффициентами.

Пример 2. Рассмотрим схему (25) из § 4 гл. 1:

$$\rho_i^n \frac{y_i^{n+1}-y_i^n}{\tau} = (ay_{\bar{x}}^n)_{x,i},$$

где $a=a_i^n$, $0<c_1\leqslant a_i^n\leqslant c_2$, $\rho_i^n \geqslant c_3>0$. Переписывая это разностное уравнение в виде

$$\frac{\rho_i^n}{\tau} y_i^{n+1} = \frac{1}{h^2}(a_{i+1}^n y_{i+1}^n + a_i^n y_{i-1}^n) + \left(\frac{\rho_i^n}{\tau} - \frac{a_{i+1}^n + a_i^n}{h^2}\right) y_i^n , \qquad (10)$$

получаем, что схема монотонна при условии

$$\frac{\tau(a_{i+1}^n + a_i^n)}{h^2} \leqslant \rho_i^n, \quad i=1, 2, \ldots, N-1, \quad n=0, 1, \ldots, K-1, \qquad (11)$$

(ср. с (27) из § 4 гл. 1), которое и является условием устойчивости данной схемы. Оно будет выполнено, если потребовать

$$\frac{\tau c_2}{h^2 c_3} \leqslant \frac{1}{2} .$$

Последнее неравенство совпадает с условием устойчивости, полученным в § 4 гл. 1 при помощи принципа замороженных коэффициентов.

Получим априорную оценку решения задачи (10) через начальные значения y_i^0 при условии (11). Предположим, что $y_0^n = y_N^n = 0$, $n=0, 1, \ldots, K$, и обозначим

$$\|y^n\|_{C(\omega_h)} = \max_{1\leqslant i\leqslant N-1} |y_i^n| .$$

Тогда в силу неотрицательности коэффициентов уравнения (10) получим

$$\frac{\rho_i^n}{\tau} |y_i^{n+1}| \leqslant \frac{\rho_i^n}{\tau} \|y^n\|_{C(\omega_h)} ,$$

и, следовательно,

$$\|y^{n+1}\|_{C(\omega_h)} \leqslant \|y^n\|_{C(\omega_h)} \leqslant \ldots \leqslant \|y^0\|_{C(\omega_h)} .$$

Рассмотрим теперь краевую задачу для уравнения первого порядка

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0, \quad x>0, \quad t>0,$$

$$u(x, 0)=u_0(x), \quad x>0, \quad u(0, t)=\mu_1(t), \quad t>0. \qquad (12)$$

Известно, что решение $u(x, t)$ этой задачи переносится по характеристикам $t=x+\text{const}$ с начальной прямой, т. е. $u(x, t)=u_0(x-t)$, если $x>t$ и $u(x, t)=\mu_1(t-x)$, если $x<t$.

Пример 3. В квадранте $x>0$, $t>0$ введем сетку с шагом h по x и шагом τ по t и обозначим $x_i=ih$, $i=0, 1, \ldots$, $t_n=n\tau$, $n=0, 1, \ldots$, $y_i^n=y(x_i, t_n)$. Одна из простейших схем для уравнения (12) имеет вид

$$\frac{y_i^{n+1}-y_i^n}{\tau}+\frac{y_i^n-y_{i-1}^n}{h}=0, \quad i=1, 2, \ldots, \quad n=0, 1, 2, \ldots, \tag{13}$$

$$y_i^0=u_0(x_i), \quad i=0, 1, \ldots, \quad y_0^n=\mu_1(t_n), \quad n=1, 2, \ldots$$

Записывая уравнение (13) в виде, разрешенном относительно y_i^{n+1}, получим

$$y_i^{n+1}=(1-\gamma)\, y_i^n+\gamma y_{i-1}^n, \quad \gamma=\frac{\tau}{h}.$$

Отсюда следует, что схема (13) монотонна при условии $\tau\leqslant h$. Пользуясь приемом, изложенным в § 4 гл. 1, можно показать, что условие $\tau\leqslant h$ и необходимо для устойчивости схемы (13).

Другая схема

$$\frac{y_i^{n+1}-y_i^n}{\tau}+\frac{y_{i+1}^n-y_i^n}{h}=0$$

немонотонна при любых τ и h. Более того, эта схема абсолютно неустойчива.

Явная схема

$$\frac{y_i^{n+1}-y_i^n}{\tau}+\frac{y_{i+1}^n-y_{i-1}^n}{2h}=0,$$

имеющая второй порядок аппроксимации по h, также немонотонна и абсолютно неустойчива. Если в последней схеме заменить y_i^n на полусумму $0{,}5\,(y_{i+1}^n+y_{i-1}^n)$, то получим разностную схему

$$\frac{y_i^{n+1}-0{,}5\,(y_{i+1}^n+y_{i-1}^n)}{\tau}+\frac{y_{i+1}^n-y_{i-1}^n}{2h}=0, \tag{14}$$

которая монотонна при $\tau\leqslant h$. Однако указанная замена ухудшает аппроксимацию, погрешность аппроксимации схемы (14) является величиной $O(\tau+h^2)+O(h^2/\tau)$. В этом легко убедиться, если записать схему (14) в виде

$$y_{t,i}^n+y_{\overset{\circ}{x},i}^n=\frac{h^2}{2\tau}\, y_{\overline{x}x,i}^n,$$

где

$$y_{\overset{\circ}{x},i}^n=(y_{i+1}^n-y_{i-1}^n)/(2h).$$

Пример 4. Рассмотрим еще одну схему для уравнения (12):

$$y_{t,i}^n+y_{\overset{\circ}{x},i}^n=0{,}5h\nu_0 y_{\overline{x}x,i}^n. \tag{15}$$

Здесь $\nu_0>0$ — постоянная, не зависящая от τ и h. При $\nu_0=0$ получаем абсолютно неустойчивую схему. Введение искусственного

добавка $0{,}5h\nu_0\ y^n_{\overline{x}x,i}$ в правую часть уравнения делает схему условно устойчивой, понижая одновременно порядок аппроксимации по h до первого. Схемы, аналогичные (15) и аппроксимирующие уравнения газовой динамики, называются схемами с искусственной вязкостью (см. [36]). Записывая уравнение (15) в виде

$$y_i^{n+1} = 0{,}5\gamma(\nu_0 - 1)\, y_{i+1}^n + (1 - \nu_0\gamma)\, y_i^n + 0{,}5\gamma(\nu_0 + 1)\, y_{i-1}^n,$$

получаем, что условия монотонности (3) выполнены при $\nu_0 \geqslant 1$, $\gamma \leqslant \nu_0^{-1}$. Таким образом, чем больше коэффициент искусственной вязкости ν_0, тем слабее ограничение на шаги сетки, вызванное требованием устойчивости. Надо помнить, однако, что введение искусственной вязкости может существенно исказить поведение истинного решения задачи (12). Поэтому при практических расчетах коэффициент вязкости ν_0 берут не слишком большим.

§ 5. Монотонные разностные схемы для уравнений второго порядка, содержащих первые производные

Рассмотрим обыкновенное дифференциальное уравнение второго порядка

$$u''(x) + r(x)u'(x) = -f(x), \quad 0 < x < 1, \quad u(0) = u(1) = 0 \qquad (1)$$

и поставим задачу построить для него разностную схему, имеющую второй порядок аппроксимации и монотонную при любых шагах сетки h. Очевидная схема второго порядка аппроксимации, которая получается заменой $u'(x)$ центральной разностной производной, является монотонной лишь при достаточно малых h. Действительно, такая схема имеет вид

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + r_i \frac{y_{i+1} - y_{i-1}}{2h} = -f_i$$

или

$$\frac{2}{h^2} y_i = \left(\frac{1}{h^2} + \frac{r_i}{2h}\right) y_{i+1} + \left(\frac{1}{h^2} - \frac{r_i}{2h}\right) y_{i-1} + f_i.$$

Условия положительности коэффициентов сводятся к неравенствам $0{,}5h|r_i| < 1$ и выполняются, если $h \leqslant 2/(\max\limits_{0 \leqslant i \leqslant N} |r_i|)$. Схема будет монотонной при любых h только в случае $r(x) \equiv 0$.

Прежде чем построить требуемую схему для уравнения (1), рассмотрим несколько частных случаев. Предположим, что $r(x) \geqslant 0$ для всех $x \in (0, 1)$ и рассмотрим схему с односторонней разностью

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + r_i \frac{y_{i+1} - y_i}{h} = -f_i.$$

Эта схема имеет первый порядок аппроксимации и монотонна при любых h. Действительно, записывая ее в виде

$$\left(\frac{2}{h^2}+\frac{r_i}{h}\right)y_i=\left(\frac{1}{h^2}+\frac{r_i}{h}\right)y_{i+1}+\frac{1}{h^2}y_{i-1}+f_i$$

и учитывая неотрицательность $r(x)$, убеждаемся в том, что условия положительности коэффициентов выполнены при всех h. Точно так же, если $r(x)\leqslant 0$, то схема

$$\frac{y_{i+1}-2y_i+y_{i-1}}{h^2}+r_i\frac{y_i-y_{i-1}}{h}=-f_i$$

монотонна при любых h и имеет первый порядок аппроксимации.

В общем случае представим функцию $r(x)$ в виде суммы $r(x)=r_+(x)+r_-(x)$, где

$$r_+(x)=0{,}5(r(x)+|r(x)|)\geqslant 0,\quad r_-(x)=0{,}5(r(x)-|r(x)|)\leqslant 0. \quad (2)$$

Схема с «направленными разностями»

$$\frac{y_{i+1}-2y_i+y_{i-1}}{h^2}+r_+(x_i)\frac{y_{i+1}-y_i}{h}+r_-(x_i)\frac{y_i-y_{i-1}}{h}=-f_i \quad (3)$$

является, как нетрудно видеть, монотонной при любых h, но имеет первый порядок аппроксимации. Изучим подробнее асимптотику погрешности аппроксимации

$$\psi_i=u_{\overline{x}x,i}+r_+(x_i)u_{x,i}+r_-(x_i)u_{\overline{x},i}+f_i \quad (4)$$

этой разностной схемы. Пользуясь разложением по формуле Тейлора, получим

$$u_{\overline{x}x,i}=u''(x_i)+O(h^2),\quad u_{x,i}=u'(x_i)+\frac{h}{2}u''(x_i)+O(h^2),$$

$$u_{\overline{x},i}=u'(x_i)-\frac{h}{2}u''(x_i)+O(h^2).$$

Подставляя эти разложения в выражение (4) и приводя подобные члены, имеем

$$\psi_i=(u_i''+f_i)+(r_+(x_i)+r_-(x_i))u'(x_i)+0{,}5h(r_+-r_-)u_i''+O(h^2),$$

откуда, учитывая (1) и (2), получим

$$\psi_i=0{,}5h|r(x_i)|u_i''+O(h^2).$$

Отсюда видно, что несколько измененная по сравнению с (3) схема

$$(1-0{,}5h|r(x_i)|)\frac{y_{i+1}-2y_i+y_{i-1}}{h^2}+r_+(x_i)\frac{y_{i+1}-y_i}{h}+$$
$$+r_-(x_i)\frac{y_i-y_{i-1}}{h}=-f_i$$

имеет второй порядок аппроксимации. Порядок аппроксимации не

уменьшится, если коэффициент $1-\frac{h}{2}|r(x_i)|$ заменим с точностью до $O(h^2)$ положительным коэффициентом

$$\varkappa_i = \frac{1}{1+0{,}5h|r(x_i)|}. \tag{5}$$

Таким образом, разностная схема

$$\varkappa_i y_{\bar{x}x,i} + r_+(x_i)\, y_{x,i} + r_-(x_i)\, y_{\bar{x},i} = -f_i \tag{6}$$

имеет второй порядок аппроксимации на решении уравнения (1). Записывая схему (6) в виде

$$\left(\frac{2\varkappa_i}{h^2} + \frac{r_+(x_i)}{h} - \frac{r_-(x_i)}{h}\right) y_i = \\ = \left(\frac{\varkappa_i}{h^2} + \frac{r_+(x_i)}{h}\right) y_{i+1} + \left(\frac{\varkappa_i}{h^2} - \frac{r_-(x_i)}{h}\right) y_{i-1} + f_i,$$

убеждаемся в том, что она монотонна при любых τ и h.

Для параболического уравнения

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + r(x)\frac{\partial u}{\partial x}$$

монотонной при любых τ и h схемой является чисто неявная схема

$$\frac{y_i^{n+1} - y_i^n}{\tau} = \varkappa_i y_{\bar{x}x,i}^{n+1} + r_+(x_i)\, y_{x,i}^{n+1} + r_-(x_i)\, y_{\bar{x},i}^{n+1}, \tag{7}$$

где $\varkappa_i$ определяется согласно (5). Схема (7) имеет аппроксимацию $O(\tau+h^2)$.

ГЛАВА 3

МЕТОД РАЗДЕЛЕНИЯ ПЕРЕМЕННЫХ

Метод разделения переменных успешно применяется для построения решений разностных схем, главным образом с постоянными коэффициентами, и для исследования сходимости. В основе метода лежит разложение решения разностной задачи по системе ее собственных функций. Требование полноты системы собственных функций сильно сужает класс рассматриваемых задач, и мы ограничиваемся в этой главе лишь задачами с самосопряженными операторами типа разностного оператора Лапласа. В § 1, 2 изучаются спектральные свойства разностных операторов, далее в § 3 методом разделения переменных проводится исследование устойчивости и сходимости разностных схем для уравнения теплопроводности. В остальных параграфах рассматриваются экономичные методы нахождения решений разностных краевых задач с постоянными коэффициентами, основанные на методе разделения переменных.

§ 1. Разностная задача на собственные значения

1. Оператор второй разностной производной. Каждую разностную краевую задачу можно рассматривать как операторное уравнение с операторами, действующими в некотором линейном конечномерном пространстве (пространстве сеточных функций). Рассмотрим, например, разностное уравнение

$$y_{\bar{x}x,i} = -f_i, \quad i = 1, 2, \ldots, N-1, \quad y_0 = \mu_1, \quad y_N = \mu_2 \tag{1}$$

на сетке

$$\omega_h = \{x_i = ih, \ i = 0, 1, \ldots, N, \ hN = l\}. \tag{2}$$

Исключая из системы уравнений (1) с помощью граничных условий значения $y_0 = \mu_1$ и $y_N = \mu_2$, придем к эквивалентной системе уравнений

$$\frac{-y_{i-1} + 2y_i - y_{i+1}}{h^2} = f_i, \quad i = 2, 3, \ldots, N-2, \tag{3}$$

$$\frac{2y_1 - y_2}{h^2} = \tilde{f}_1, \quad \frac{-y_{N-2} + 2y_{N-1}}{h^2} = \tilde{f}_{N-1},$$

где $\tilde{f}_1 = f_1 + \mu_1/h^2$, $\tilde{f}_{N-1} = f_{N-1} + \mu_2/h^2$.

Рассмотрим множество векторов $y = (y_1, y_2, \ldots, y_{N-1})^T$, $y_i = y(x_i)$, $x_i \in \omega_h$ и определим на этом множестве оператор A формулами

$$(Ay)_i = -y_{\bar{x}x,i}, \quad i = 2, 3, \ldots, N-2,$$

$$(Ay)_1 = \frac{2y_1 - y_2}{h^2}, \quad (Ay)_N = \frac{-y_{N-2} + 2y_{N-1}}{h^2}. \tag{4}$$

Тогда систему (3) можно записать в виде операторного уравнения

$$Ay = f, \tag{5}$$

где $f = (\tilde{f}_1, f_2, \ldots, f_{N-2}, \tilde{f}_{N-1})^T$. Отметим, что уравнение (5) учитывает как правую часть разностной схемы (1), так и ее граничные условия.

Итак, разностная задача (1) порождает разностный оператор (4). Оператор (4) определен на множестве функций, заданных только во внутренних точках сетки ω_h, т. е. при $i = 1, 2, \ldots, N-1$. Удобнее, однако, считать, что оператор A определен на подпространстве H функций, заданных на всей сетке ω_h и обращающихся в нуль на границе: $y_0 = y_N = 0$. При этом оператор A задается единообразными формулами

$$(Ay)_i = -y_{\bar{x}x,i}, \quad i = 1, 2, \ldots, N-1, \quad y_0 = y_N = 0 \tag{6}$$

во всей области определения.

Оператор A, определенный согласно (6), будем называть *оператором второй разностной производной*. Изучим свойства этого оператора.

2. Задача на собственные значения. Задача на собственные значения для оператора A состоит в том, чтобы найти числа λ (*собственные числа* или *собственные значения*) такие, для которых уравнение

$$Ay=\lambda y \tag{7}$$

имеет нетривиальные решения (*собственные функции*), и найти собственные функции. Заметим, что по существу уравнение (7) для оператора (6) представляет собой алгебраическую задачу на собственные значения для матрицы

$$A=\frac{1}{h^2}\begin{bmatrix} 2 & -1 & 0 & 0 & \dots & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \dots & 0 & -1 & 2 \end{bmatrix},$$

которая называется матрицей разностного оператора (6). Матрица A является трехдиагональной симметричной матрицей порядка $N-1$. Известно, что у такой матрицы существуют $N-1$ действительных собственных чисел и столько же линейно независимых собственных функций.

В случае оператора (6) можно выписать все собственные числа и отвечающие им собственные функции в явном виде. Эти вопросы уже рассматривались в § 4 ч. I. Напомним полученные там результаты. Запишем уравнение (7) в виде

$$-y_{\bar{x}x,i}=\lambda y_i, \quad i=1, 2, \dots, N-1, \quad y_0=y_N=0, \quad hN=l, \tag{8}$$

или подробнее

$$y_{i+1}+y_{i-1}=(2-h^2\lambda)y_i, \quad i=1, 2, \dots, N-1, \quad y_0=y_N=0. \tag{9}$$

Разностная задача (8) представляет собой аппроксимацию дифференциальной задачи

$$-u''(x)=\lambda u(x), \quad 0<x<l, \quad u(0)=u(l)=0, \tag{10}$$

решением которой являются собственные числа

$$\lambda_k=\left(\frac{\pi k}{l}\right)^2, \quad k=1, 2, \dots$$

и отвечающие им собственные функции

$$u_k(x)=\sin\frac{\pi k x}{l}, \quad k=1, 2, \dots$$

Попытаемся поэтому искать собственные функции задачи (8) в виде

$$y_k(x_i)=\sin\frac{\pi k x_i}{l}, \quad x_i=ih, \quad k=1, 2, \dots \tag{11}$$

Граничные условия $y_0=y_N=0$ при этом выполнены. Подставляя (11) в уравнение (9), получим уравнение

$$\sin\frac{\pi k(x_i+h)}{l}+\sin\frac{\pi k(x_i-h)}{l}=(2-h^2\lambda)\sin\frac{\pi kx_i}{l}$$

или

$$2\sin\frac{\pi kx_i}{l}\cos\frac{\pi kh}{l}=(2-h^2\lambda)\sin\frac{\pi kx_i}{l}.$$

Отсюда видно, что функция (11) является собственной функцией оператора (6), если

$$2\cos\frac{\pi kh}{l}=2-h^2\lambda,$$

т. е.

$$\lambda=\lambda_k=\frac{4}{h^2}\sin^2\frac{\pi kh}{2l}.$$

При $k=1, 2, \ldots, N-1$ получаем $N-1$ различных действительных собственных чисел λ_k и отвечающих им собственных функций. Итак, решение задачи (8) имеет вид

$$\lambda_k=\frac{4}{h^2}\sin^2\frac{\pi kh}{2l},\quad y_k(x_i)=\sin\frac{\pi kx_i}{l},$$

$$x_i=ih,\quad i=0, 1, \ldots, N,\quad k=1, 2, \ldots, N-1,\quad hN=l. \qquad (12)$$

3. Свойства собственных значений и собственных функций. Справедливы неравенства

$$0<\lambda_1<\lambda_2<\ldots<\lambda_k<\lambda_{k+1}<\ldots<\lambda_{N-1}<\frac{4}{h^2}.$$

Для минимального собственного числа λ_1 в п. 5 § 4 ч. I была получена оценка снизу $\lambda_1\geqslant\delta_1>0$ константой $\delta_1=9/l^2$, не зависящей от h.

Переходя к изучению свойств собственных функций, введем в пространстве H (напомним, что H — линейное пространство функций, заданных на ω_h и удовлетворяющих условию $y_0=y_N=0$) скалярное произведение

$$(y, v)=\sum_{i=1}^{N-1}y_iv_ih$$

и норму

$$\|y\|=\sqrt{(y, y)}=\left(\sum_{i=1}^{N-1}y_i^2\right)^{1/2}.$$

Оператор A второй разностной производной (6) является самосопряженным в H оператором, т. е. $(Ay, v)=(y, Av)$ для всех $y, v\in H$. Это сразу следует из тождеств, называемых разностными формулами Грина (см. также (15) из § 3 гл. 1). Имеем по

определению

$$(Ay, v) = -\sum_{i=1}^{N-1} y_{\bar{x}x,i} v_i h = -\sum_{i=1}^{N-1} y_{x,i} v_i + \sum_{i=1}^{N-1} y_{\bar{x},i} v_i =$$

$$= \sum_{i=1}^{N-1} y_{\bar{x},i} v_i - \sum_{i=2}^{N} y_{\bar{x},i} v_{i-1} = \sum_{i=1}^{N} y_{\bar{x},i} v_{\bar{x},i} h.$$

Последнее равенство получено с учетом условия $v_0 = v_N = 0$. Меняя y и v местами, получим

$$(Av, y) = \sum_{i=1}^{N} v_{\bar{x},i} y_{\bar{x},i} h = (v, Ay). \tag{13}$$

Следствием самосопряженности оператора (6) является ортогональность его собственных функций, отвечающих различным собственным числам. Действительно, пусть

$$Ay_k = \lambda_k y_k, \quad Ay_m = \lambda_m y_m, \quad \lambda_k \neq \lambda_m.$$

Тогда

$$(Ay_k, y_m) = \lambda_k (y_k, y_m), \quad (Ay_m, y_k) = \lambda_m (y_m, y_k)$$

и в силу самосопряженности имеем

$$0 = (Ay_k, y_m) - (Ay_m, y_k) = (\lambda_k - \lambda_m)(y_m, y_k).$$

Отсюда и получим, что $(y_m, y_k) = 0$, если $\lambda_k \neq \lambda_m$. Таким образом, система собственных функций (12) образует ортогональный базис в пространстве H.

Замечание. Поскольку все собственные числа различны, то условие $\lambda_k \neq \lambda_m$ эквивалентно условию $k \neq m$. Таким образом доказано, что сумма

$$\sum_{i=1}^{N-1} \sin \frac{\pi k x_i}{l} \sin \frac{\pi m x_i}{l} h, \qquad hN = l,$$

обращается в нуль при $k \neq m$. Это замечание потребуется в § 2 при изучении собственных функций пятиточечного разностного оператора Лапласа.

Вычислим квадрат нормы собственной функции $y_k(x)$. В случае дифференциальной задачи (10) имеем

$$\|u\|^2 = \int_0^l \sin^2 \frac{\pi k x}{l} dx = \frac{l}{2}.$$

Покажем, что и в разностном случае собственная функция $y_k(x)$ имеет ту же самую норму $\sqrt{l/2}$. По определению имеем

$$\|y\|^2 = \sum_{i=1}^{N-1} h \sin^2 \frac{\pi k x_i}{l} = \sum_{i=1}^{N} h \sin^2 \frac{\pi k x_i}{l}.$$

Преобразуем выражение, стоящее под знаком суммы

$$\sin^2 \frac{\pi k x_i}{l} = \frac{1}{2}\left(1 - \cos \frac{2\pi k x_i}{l}\right),$$

и учтем тождество

$$\sin\frac{2\pi k\,(x_i+0{,}5h)}{l}-\sin\frac{2\pi k\,(x_{i-1}+0{,}5h)}{l}=2\sin\frac{\pi kh}{l}\cos\frac{2\pi kx_i}{l}.$$

Тогда получим

$$\|y\|^2=\frac{hN}{2}-\frac{1}{4\sin(\pi kh/l)}\left(\sin\frac{2\pi k\,(x_N+0{,}5h)}{l}-\sin\frac{\pi kh}{l}\right)=\frac{l}{2}.$$

Таким образом, система собственных функций

$$\mu_k(x)=\sqrt{\frac{2}{l}}\sin\frac{\pi kx}{l},\quad k=1,2,\ldots,N-1,$$

$$x=x_i,\quad i=1,2,\ldots,N-1,\quad hN=l, \tag{14}$$

образует ортонормированный базис в пространстве H.

4. Операторные неравенства. Любой элемент $y\in H$ можно разложить по базису, т. е. единственным образом представить в виде суммы

$$y(x)=\sum_{k=1}^{N-1}c_k\mu_k(x),\qquad x\in\omega_h, \tag{15}$$

где $c_k=(y,\mu_k)$ — коэффициенты Фурье.

Из (15) и ортонормированности системы $\{\mu_k\}$ следует тождество

$$\|y\|^2=(y,y)=\sum_{k=1}^{N-1}c_k^2.$$

Используя разложение (15), получим

$$Ay(x)=\sum_{k=1}^{N-1}c_kA\mu_k(x)=\sum_{k=1}^{N-1}c_k\lambda_k\mu_k(y)$$

и

$$(Ay,y)=\sum_{k=1}^{N-1}c_k^2\lambda_k. \tag{16}$$

Из тождества (16) следуют важные неравенства

$$\lambda_1\|y\|^2\leqslant(Ay,y)\leqslant\lambda_{N-1}\|y\|^2, \tag{17}$$

справедливые для любого $y\in H$. Учитывая доказанные выше оценки для собственных чисел, получим из (17) неравенства

$$\delta\|y\|^2\leqslant(Ay,y)\leqslant\frac{4}{h^2}\|y\|^2,\qquad\delta=\frac{9}{l^2}. \tag{18}$$

Из (18) следует в частности, что $(Ay,y)>0$ для всех $y\in H$, $y\neq0$. Операторы, обладающие этим свойством, называются *положительными операторами*. Неравенство $A>0$ будет означать, что A — положительный оператор. В дальнейшем мы будем часто использо-

вать и другие операторные неравенства. Неравенство $A \geqslant 0$ означает, что $(Ay, y) \geqslant 0$ для всех $y \in H$. Для двух операторов A и B неравенство $A \geqslant B$ означает, что $A - B \geqslant 0$. В этих обозначениях свойство (18) можно записать в виде операторных неравенств

$$\delta E \leqslant A \leqslant \frac{4}{h^2} E,$$

где E — единичный оператор.

Рассмотрим теперь разностный оператор Λ с переменными коэффициентами, определяемый формулами

$$(\Lambda y)_i = -(a y_{\bar{x}})_{x,i}, \quad i = 1, 2, \ldots, N-1, \quad y_0 = y_N = 0, \tag{19}$$

где $0 < c_1 \leqslant a_i \leqslant c_2$. Покажем, что Λ — самосопряженный и положительный в H оператор и получим оценки, аналогичные (18). Для любых $y, v \in H$ имеем

$$(\Lambda y, v) = -\sum_{i=1}^{N-1} (a y_{\bar{x}})_{x,i} v_i h = -\sum_{i=1}^{N-1} a_{i+1} y_{x,i} v_i + \sum_{i=1}^{N-1} a_i y_{\bar{x},i} v_i =$$

$$= \sum_{i=1}^{N-1} a_i y_{\bar{x},i} v_i - \sum_{i=2}^{N} a_i y_{\bar{x},i} v_{i-1} = \sum_{i=1}^{N} a_i y_{\bar{x},i} v_{\bar{x},i} h.$$

Отсюда видно, что $(\Lambda y, v) = (y, \Lambda v)$ и

$$(\Lambda y, y) = \sum_{i=1}^{N} a_i y_{\bar{x},i}^2 h. \tag{20}$$

Из тождества (20) получаем неравенства

$$c_1 \sum_{i=1}^{N} y_{\bar{x},i}^2 h \leqslant (\Lambda y, y) \leqslant c_2 \sum_{i=1}^{N} y_{\bar{x},i}^2 h.$$

Согласно (13) имеем

$$(Ay, y) = \sum_{i=1}^{N} y_{\bar{x},i}^2 h,$$

где A — оператор, определенный согласно (6).

Поэтому последние неравенства можно переписать в виде

$$c_1 (Ay, y) \leqslant (\Lambda y, y) \leqslant c_2 (Ay, y)$$

или в виде операторных неравенств

$$c_1 A \leqslant \Lambda \leqslant c_2 A. \tag{21}$$

Два оператора, A и Λ, называются *энергетически эквивалентными операторами*, если выполнены неравенства вида (21) с положительными постоянными c_1, c_2. Название объясняется тем, что в приложениях выражение (Ay, y) представляет собой энергию самосопряженного положительного оператора A. Константы c_1, c_2 на-

зываются константами эквивалентности операторов A и Λ. Норма $\|y\|_A=\sqrt{(Ay, y)}$ называется *энергетической нормой*, порожденной оператором A.

Из (21) и (18) получаем неравенства

$$c_1\delta E\leqslant\Lambda\leqslant\frac{4c_2}{h^2}E,\quad \delta=\frac{9}{l^2},$$

из которых следует, что спектр оператора (19) принадлежит отрезку $[c_1\delta, 4c_2/h^2]$.

Свойства матричных неравенств, доказанные в п. 3 § 4 гл. 2 ч. II, остаются справедливыми и для операторных неравенств в конечномерном пространстве H со скалярным произведением, если только заменить в соответствующих неравенствах транспонированные матрицы A^T, B^T на сопряженные операторы A^*, B^*.

В частности, если $A^*=A>0$, то существует квадратный корень $A^{1/2}$ из оператора A, который является самосопряженным положительным оператором. Если L — обратимый оператор, то операторные неравенства

$$A\geqslant B,\quad L^*AL\geqslant L^*BL$$

эквивалентны. Если $C^*=C>0$ и α, β — любые вещественные числа, то эквивалентны неравенства

$$\alpha C\geqslant\beta E,\quad \alpha E\geqslant\beta C^{-1}.$$

§ 2. Задача на собственные значения для пятиточечного разностного оператора Лапласа

1. Самосопряженность. В § 1 гл. 2 изучалась разностная задача Дирихле для уравнения Пуассона в прямоугольнике. Эта задача порождает оператор, который называется пятиточечным разностным оператором Лапласа и определяется следующим образом:

$$(Ay)_{ij}=-y_{\bar{x}_1x_1,ij}-y_{\bar{x}_2x_2,ij},\qquad (1)$$

$$i=1, 2, \ldots, N_1-1,\quad j=1, 2, \ldots, N_2-1,\quad h_1N_1=l_1,\quad h_2N_2=l_2,$$

$$y(x_{ij})=0,\ \text{если}\ x_{ij}\in\gamma.$$

Прямоугольная сетка $\Omega=\omega\cup\gamma$ с шагами h_1 и h_2 состоит из узлов

$$x_{ij}=(x_1^{(i)}, x_2^{(j)}),\quad x_1^{(i)}=ih_1,\ x_2^{(j)}=jh_2,$$

$$i=0, 1, \ldots, N_1,\quad j=0, 1, \ldots, N_2,\quad h_\alpha N_\alpha=l_\alpha,\quad \alpha=1, 2,$$

ω — множество внутренних узлов сетки Ω и γ — граница Ω.

Предположение о том, что $y_{ij}=0$ на γ, не является дополнительным ограничением в случае задачи Дирихле, поскольку можно считать, что неоднородные граничные условия учтены правой частью операторного уравнения $Ay=f$.

Введем линейное пространство H функций, определенных на сетке Ω и обращающихся в нуль на γ. Это конечномерное пространство размерности $(N_1-1)(N_2-1)$. Определим в H скалярное

произведение и норму

$$(y, v) = \sum_{i=1}^{N_1-1} h_1 \sum_{j=1}^{N_2-1} h_2 y_{ij} v_{ij}, \qquad \|y\| = \sqrt{(y, y)}.$$

Будем считать, что оператор A действует в пространстве H.

Покажем, что оператор A, определенный согласно (1), является самосопряженным и положительным в H оператором. Пусть $y, v \in H$. Так же как и в § 1, можно доказать, что при каждом $j = 1, 2, \ldots, N_2-1$ справедливо тождество

$$-\sum_{i=1}^{N_1-1} h_1 y_{\bar{x}_1 x_1, ij} v_{ij} = \sum_{i=1}^{N_1} h_1 y_{\bar{x}_1, ij} v_{\bar{x}_1, ij} \tag{2}$$

и при каждом $i = 1, 2, \ldots, N_1-1$ — тождество

$$-\sum_{j=1}^{N_2-1} h_2 y_{\bar{x}_2 x_2, ij} v_{ij} = \sum_{j=1}^{N_2} h_2 y_{\bar{x}_2, ij} v_{\bar{x}_2, ij}, \tag{3}$$

где

$$y_{\bar{x}_1, ij} = (y_{ij} - y_{i-1, j})/h_1, \quad y_{\bar{x}_2, ij} = (y_{ij} - y_{i, j-1})/h_2.$$

Из (2) и (3) получим

$$-\sum_{j=1}^{N_2-1} h_2 \sum_{i=1}^{N_1-1} h_1 y_{\bar{x}_1 x_1, ij} v_{ij} - \sum_{i=1}^{N_1-1} h_1 \sum_{j=1}^{N_2-1} h_2 y_{\bar{x}_2 x_2, ij} v_{ij} =$$

$$= \sum_{j=1}^{N_2-1} h_2 \sum_{i=1}^{N_1} h_1 y_{\bar{x}_1, ij} + \sum_{i=1}^{N_1-1} h_1 \sum_{j=1}^{N_2} h_2 y_{\bar{x}_2, ij} v_{\bar{x}_2, ij},$$

т. е.

$$(Ay, v) = \sum_{i=1}^{N_1} h_1 \sum_{j=1}^{N_2-1} h_2 y_{\bar{x}_1, ij} v_{\bar{x}_1, ij} + \sum_{i=1}^{N_1-1} h_1 \sum_{j=1}^{N_2} h_2 y_{\bar{x}_2, ij} v_{\bar{x}_2, ij}. \tag{4}$$

Поскольку функции y и v входят в правую часть тождества (4) равноправно, получаем, что $(Ay, v) = (y, Av)$ для любых $y, v \in H$. Итак, оператор A — самосопряженный.

2. Оценка собственных чисел. Положительность оператора. Рассмотрим теперь для этого оператора задачу на собственные значения

$$Ay = \lambda y$$

или, более подробно,

$$y_{\bar{x}_1 x_1, ij} + y_{\bar{x}_2 x_2, ij} + \lambda y_{ij} = 0, \qquad x_{ij} \in \omega, \tag{5}$$

$$y_{ij} = 0, \qquad x_{ij} \in \gamma.$$

Так как A — самосопряженный в H оператор, существуют $(N_1-1) \times (N_2-1)$ действительных собственных чисел, а система собственных функций образует в H ортогональный базис.

Выпишем в явном виде решение задачи (5). Рассмотрим два набора чисел

$$\lambda_{k_1} = \frac{4}{h_1^2} \sin^2 \frac{\pi k_1 h_1}{2l_1}, \quad k_1 = 1, 2, \ldots, N_1 - 1,$$

$$\lambda_{k_2} = \frac{4}{h_2^2} \sin^2 \frac{\pi k_2 h_2}{2l_2}, \quad k_2 = 1, 2, \ldots, N_2 - 1,$$

и образуем всевозможные суммы вида

$$\lambda_{k_1 k_2} = \lambda_{k_1} + \lambda_{k_2}, \quad k_\alpha = 1, 2, \ldots, N_\alpha - 1, \quad \alpha = 1, 2. \tag{6}$$

Далее, рассмотрим системы функций

$$\mu_{k_1}(x_1^{(i)}) = \sqrt{\frac{2}{l_1}} \sin \frac{\pi k_1 x_1^{(i)}}{l_1}, \quad k_1 = 1, 2, \ldots, N_1 - 1,$$

$$x_1^{(i)} = ih_1, \; i = 0, 1, \ldots, N_1, \; h_1 N_1 = l_1$$

и

$$\mu_{k_2}(x_2^{(j)}) = \sqrt{\frac{2}{l_2}} \sin \frac{\pi k_2 x_2^{(j)}}{l_2}, \quad k_2 = 1, 2, \ldots, N_2 - 1,$$

$$x_2^{(j)} = jh_2, \; j = 0, 1, \ldots, N_2, \; h_2 N_2 = l_2$$

и образуем всевозможные произведения вида

$$\mu_k(x_{ij}) = \mu_{k_1}(x_1^{(i)}) \mu_{k_2}(x_2^{(j)}), \tag{7}$$

$$k = (k_1, k_2), \quad k_1 = 1, 2, \ldots, N_1 - 1, \quad k_2 = 1, 2, \ldots, N_2 - 1,$$

$$x_{ij} = (x_1^{(i)}, x_2^{(j)}) \in \Omega.$$

Учитывая результаты § 1, относящиеся к одномерной задаче на собственные значения, можно простой подстановкой чисел (6) и функций (7) в уравнение (5) убедиться в том, что при каждом $k = (k_1, k_2)$ число

$$\lambda_k = \lambda_{k_1 k_2} = \frac{4}{h_1^2} \sin^2 \frac{\pi k_1 h_1}{2l_1} + \frac{4}{h_2^2} \sin^2 \frac{\pi k_2 h_2}{2l_2} \tag{8}$$

является собственным числом пятиточечного разностного оператора Лапласа (1), которому отвечает собственная функция

$$\mu_k(x_{ij}) = \mu_{k_1 k_2}(x_{ij}) = \frac{2}{\sqrt{l_1 l_2}} \sin \frac{\pi k_1 x_1^{(i)}}{l_1} \sin \frac{\pi k_2 x_2^{(j)}}{l_2}. \tag{9}$$

Индексы $k = (k_1, k_2)$ и $m = (m_1, m_2)$ назовем совпадающими, и обозначим $k = m$, если $k_1 = m_1$, $k_2 = m_2$ и несовпадающими ($k \neq m$) — в противном случае. Покажем, что при $k \neq m$ функции μ_k и μ_m являются ортогональными, т. е. $(\mu_k, \mu_m) = 0$ при $k \neq m$. По

определению имеем

$$(\mu_k, \mu_m) = \sum_{i=1}^{N_1-1} h_1 \sum_{j=1}^{N_2-1} h_2 \mu_{k_1}(x_1^{(i)})\,\mu_{k_2}(x_2^{(j)})\,\mu_{m_1}(x_1^{(i)})\,\mu_{m_2}(x_2^{(j)}) =$$

$$= \left(\sum_{i=1}^{N_1-1} h_1 \mu_{k_1}(x_1^{(i)})\,\mu_{m_1}(x_1^{(i)})\right)\left(\sum_{j=1}^{N_2-1} h_2 \mu_{k_2}(x_2^{(j)})\,\mu_{m_2}(x_2^{(j)})\right).$$

Согласно замечанию на стр. 314 в § 1, по крайней мере одна из сумм, стоящих в круглых скобках, равна нулю при $k \neq m$. Аналогично доказывается, что норма функции (9) равна единице. Таким образом, система функций $\{\mu_{k_1k_2}(x_{ij})\}_{k_1,k_2=1}^{N_1-1,N_2-1}$ образует ортонормированный базис в пространстве H и числа $\lambda_{k_1k_2}$, определенные согласно (8), составляют при $k_1 = 1, 2, \ldots, N_1-1$, $k_2 = 1, 2, \ldots, N_2-1$ весь спектр оператора A.

Пользуясь результатами § 1, относящимися к оценкам собственных чисел одномерной задачи, получаем, что все собственные числа (8) удовлетворяют неравенствам

$$\frac{9}{l_1^2} + \frac{9}{l_2^2} \leqslant \lambda_k \leqslant \frac{4}{h_1^2} + \frac{4}{h_2^2}. \tag{10}$$

Наименьшее и наибольшее собственные числа

$$\lambda_{\min} = \frac{4}{h_1^2}\sin^2\frac{\pi h_1}{2l_1} + \frac{4}{h_2^2}\sin^2\frac{\pi h_2}{2l_2}, \quad \lambda_{\max} = \frac{4}{h_1^2}\cos^2\frac{\pi h_1}{2l_1} + \frac{4}{h_2^2}\cos^2\frac{\pi h_2}{2l_2}. \tag{11}$$

Так же, как и в § 1, получаем оценки для энергии оператора

$$\lambda_{\min}\|y\|^2 \leqslant (Ay, y) \leqslant \lambda_{\max}\|y\|^2.$$

Заметим, что согласно (4),

$$(Ay, y) = \sum_{i=1}^{N_1} h_1 \sum_{j=1}^{N_2-1} h_2 (y_{\bar{x}_1,ij})^2 + \sum_{i=1}^{N_1-1} h_1 \sum_{j=1}^{N_2} h_2 (y_{\bar{x}_2,ij})^2.$$

§ 3. Исследование устойчивости и сходимости схемы с весами для уравнения теплопроводности

1. Исходная задача и разностная схема. Схема с весами для уравнения теплопроводности рассматривалась в § 4 гл. 1, где была исследована ее погрешность аппроксимации и найдены необходимые условия устойчивости. В настоящем параграфе дано полное исследование устойчивости и сходимости схемы с весами и получены оценки погрешности.

Будем рассматривать первую краевую задачу для уравнения теплопроводности

$$\begin{aligned} &\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x, t), \quad 0 < x < l, \quad 0 < t \leqslant T, \\ &u(0, t) = u(l, t) = 0, \quad 0 < t \leqslant T, \\ &u(x, 0) = u_0(x), \quad 0 \leqslant x \leqslant l. \end{aligned} \tag{1}$$

Введем сетку $\omega_{h\tau}=\omega_h\times\omega_\tau$, где

$$\omega_h=\{x_i=ih,\ i=0,\ 1,\ \ldots,\ N,\ hN=l\},$$
$$\omega_\tau=\{t_n=n\tau,\ n=0,\ 1,\ \ldots,\ K,\ K\tau=T\},$$

и обозначим $y_i^n=y(x_i,\ t_n)$,

$$y_{t,i}^n=\frac{y_i^{n+1}-y_i^n}{\tau},\quad y_{\overline{x}x,i}^n=\frac{y_{i+1}^n-2y_i^n+y_{i-1}^n}{h^2}.$$

Дифференциальную задачу (1) заменим на сетке $\omega_{h\tau}$ разностной задачей

$$y_{t,i}^n=\sigma y_{\overline{x}x,i}^{n+1}+(1-\sigma)\,y_{\overline{x}x,i}^n+\varphi_i^n, \tag{2}$$
$$i=1,\ 2,\ \ldots,\ N-1,\quad n=0,\ 1,\ \ldots,\ K-1,$$

где σ — число и φ_i^n — сеточная функция, заменяющая функцию $f(x,\ t)$. К уравнениям (2) следует добавить разностные начальные и граничные условия

$$y_0^n=y_N^n=0,\quad n=1,\ 2,\ \ldots,\ K-1,$$
$$y_i^0=u_0(x_i),\quad i=0,\ 1,\ \ldots,\ N. \tag{3}$$

Разностная задача (2), (3) называется *схемой с весами для уравнения теплопроводности.* Точность этой разностной схемы характеризуется погрешностью $z_i^n=y_i^n-u(x_i,\ t_n)$. Для погрешности получаем задачу

$$z_{t,i}^n=\sigma z_{\overline{x}x,i}^{n+1}+(1-\sigma)\,z_{\overline{x}x,i}^n+\psi_i^n,$$
$$i=1,\ 2,\ \ldots,\ N-1,\quad n=0,\ 1,\ \ldots,\ K-1, \tag{4}$$
$$z_0^n=z_N^n=0,\quad z_i^0=0,\quad i=0,\ 1,\ \ldots,\ N,$$

где $\psi_i^n=-u_{t,i}^n+\sigma u_{\overline{x}x,i}^{n+1}+(1-\sigma)\,u_{\overline{x}x,i}^n+\varphi_i^n$ — погрешность аппроксимации схемы (2), (3) на решении задачи (1). В § 4 гл. 1 было показано, что при надлежащем выборе φ_i^n справедливы соотношения

$$\psi_i^n=O(\tau^2+h^4)\ \text{при}\ \sigma=\sigma_*=\frac{1}{2}-\frac{h^2}{12\tau},$$
$$\psi_i^n=O(\tau^2+h^2)\ \text{при}\ \sigma=0{,}5,$$
$$\psi_i^n=O(\tau+h^2)\ \text{при}\ \sigma\neq\sigma_*,\quad \sigma\neq 0{,}5.$$

В настоящем параграфе мы получим оценки решения разностной задачи (2), (3) через начальные данные y_i^0 и правую часть φ_i^n, выражающие устойчивость схемы по начальным данным и по правой части. Из этих оценок будут сразу следовать оценки погрешности z_i^n через погрешность аппроксимации ψ_i^n, характеризующие сходимость и точность схемы (2), (3).

Для того чтобы схемой (2), (3) можно было пользоваться, необходимо, чтобы уравнение (2) было разрешимо относительно y^{n+1}. Введем оператор, уже рассмотренный в § 1, а именно оператор второй разностной производной

$$(Ay)_i = -y_{\bar{x}x,i}, \quad i = 1, 2, \ldots, N-1, \quad y_0 = y_N = 0. \tag{5}$$

Тогда схему (2), (3) можно записать в операторном виде

$$\frac{y^{n+1} - y^n}{\tau} + \sigma A y^{n+1} + (1-\sigma) A y^n = \varphi^n, \quad y^0 = u^0, \tag{6}$$

где $y^n = (y_1^n, y_2^n, \ldots, y_{N-1}^n)^T$, $\varphi^n = (\varphi_1^n, \ldots, \varphi_{N-1}^n)^T$, $u^0 = (u_0(x_1), u_0(x_2), \ldots, u_0(x_{N-1}))^T$, или, что то же самое, в виде

$$(E + \sigma\tau A) y^{n+1} = (E - (1-\sigma)\tau A) y^n + \tau\varphi^n,$$

где E — единичный оператор. Разрешимость уравнения (6) относительно y^{n+1} эквивалентна обратимости оператора $B = E + \sigma\tau A$. Оператор B будет иметь обратный, если потребовать

$$1 + \sigma\tau\lambda_k > 0, \qquad k = 1, 2, \ldots, N-1, \tag{7}$$

где $\lambda_k > 0$ — собственные числа оператора (5). В дальнейшем всегда будем предполагать, что неравенство (7) выполнено. Заметим, впрочем, что условие разрешимости (7) следует из полученного далее условия устойчивости схемы (2), (3).

2. Устойчивость схемы по начальным данным. Переходя к исследованию устойчивости схемы (2), (3), будем искать ее решение в виде разложения по ортонормированному базису $\{\mu_k\}_{k=1}^{N-1}$ собственных функций оператора (5). Явный вид собственных чисел и собственных функций μ_k дается формулами (12), (14) из § 1. При каждом n решение $y_i^n = y(x_i, t_n)$ можно представить в виде

$$y(x_i, t_n) = \sum_{k=1}^{N-1} c_k(t_n)\, \mu_k(x_i). \tag{8}$$

Правая часть φ_i^n уравнения (2) также допускает разложение

$$\varphi(x_i, t_n) = \sum_{k=1}^{N-1} \hat{\varphi}_k(t_n)\, \mu_k(x_i). \tag{9}$$

Здесь $c_k(t_n)$, $\hat{\varphi}_k(t_n)$ — коэффициенты Фурье функций $y(x_i, t_n)$, $\varphi(x_i, t_n)$ соответственно. Подставляя (8) и (9) в уравнение (2) и учитывая, что $(\mu_k(x))_{\bar{x}x,i} = -\lambda_k \mu_k(x_i)$, получим

$$\sum_{k=1}^{N-1} \mu_k(x_i) \left[\frac{c_k(t_{n+1}) - c_k(t_n)}{\tau} + \sigma\lambda_k c_k(t_{n+1}) + \right.$$
$$\left. + (1-\sigma)\lambda_k c_k(t_n) - \hat{\varphi}_k(t_n) \right] = 0.$$

В силу линейной независимости функций $\mu_k(x)$ отсюда следует ра-

венство нулю выражений, стоящих в квадратных скобках, т. е.

$$\frac{c_k(t_{n+1})-c_k(t_n)}{\tau}+\sigma\lambda_k c_k(t_{n+1})+(1-\sigma)\lambda_k c_k(t_n)=\hat{\varphi}_k(t_n), \qquad (10)$$

$$n=0, 1, \ldots, K-1, \quad k=1, 2, \ldots, N-1.$$

Уравнение (10) при каждом k представляет собой разностное уравнение первого порядка относительно $c^{(n)}=c_k(t_n)$. Чтобы выделить единственное решение, надо задать начальное условие $c_k(0)=(y^0, \mu_k)$.

Из уравнения (10) получаем

$$c_k(t_{n+1})=q_k c_k(t_n)+\frac{\tau}{1+\sigma\tau\lambda_k}\hat{\varphi}_k(t_n), \qquad (11)$$

где

$$q_k=\frac{1-(1-\sigma)\tau\lambda_k}{1+\sigma\tau\lambda_k}. \qquad (12)$$

Выражение, стоящее в знаменателе, положительно согласно (7). Учитывая (8) и (11), представим решение y_i^{n+1} задачи (2), (3) в виде

$$y_i^{n+1}=\sum_{k=1}^{N-1}\left[q_k c_k(t_n)+\frac{\tau}{1+\sigma\tau\lambda_k}\hat{\varphi}_k(t_n)\right]\mu_k(x_i). \qquad (13)$$

Обозначая

$$\bar{y}_i^{n+1}=\sum_{k=1}^{N-1} q_k c_k(t_n)\mu_k(x_i), \qquad (14)$$

$$\tilde{y}_i^{n+1}=\sum_{k=1}^{N-1}\frac{\tau}{1+\sigma\tau\lambda_k}\hat{\varphi}_k(t_n)\mu_k(x_i), \qquad (15)$$

получим, что $y_i^{n+1}=\bar{y}_i^{n+1}+\tilde{y}_i^{n+1}$.

Оценим по отдельности нормы функций $\bar{y}^{n+1}$ и $\tilde{y}^{n+1}$. Из (14) в силу ортонормированности базиса $\{\mu_k\}$ получаем

$$\|\bar{y}^{n+1}\|^2=\sum_{i=1}^{N-1}(\bar{y}_i^{n+1})^2 h=\sum_{k=1}^{N-1} q_k^2 (c_k(t_n))^2$$

и, следовательно,

$$\|\bar{y}^{n+1}\|\leqslant\left(\sum_{k=1}^{N-1}(c_k(t_n))^2\right)^{1/2}\max_{1\leqslant k\leqslant N-1}|q_k|=\|y^n\|\max_{1\leqslant k\leqslant N-1}|q_k|.$$

Потребуем, чтобы выполнялось условие

$$|q_k|\leqslant 1, \quad k=1, 2, \ldots, N-1. \qquad (16)$$

Нетрудно видеть, что (16) эквивалентно условию

$$\sigma\geqslant\frac{1}{2}-\frac{1}{\tau\lambda_{N-1}}, \qquad (17)$$

где $\lambda_{N-1}=\frac{4}{h^2}\cos^2\frac{\pi h}{2l}$ — наибольшее собственное значение оператора (5). Условие (17) будет выполнено, если потребовать

$$\sigma\geqslant\frac{1}{2}-\frac{h^2}{4\tau}. \tag{18}$$

Заметим, что из (17) при любом $k=1, 2, \ldots, N-1$ следует неравенство

$$1+\sigma\tau\lambda_k\geqslant\frac{\tau\lambda_k}{2}>0,$$

т. е. неравенство (7).

Итак, если выполнено (18), то справедлива оценка

$$\|\bar{y}^{n+1}\|\leqslant\|y^n\|. \tag{19}$$

По существу, эта оценка означает устойчивость схемы (2), (3) по начальным данным. Действительно, если в уравнении (2) $\varphi_i^n\equiv 0$, то $y_i^{n+1}=\bar{y}_i^{n+1}$, и из (19) получаем

$$\|y^{n+1}\|\leqslant\|y^n\|\leqslant\|y^{n-1}\|\leqslant\ldots\leqslant\|y^0\|,$$

что означает устойчивость схемы (2), (3) по начальным данным в норме

$$\|y\|=\left(\sum_{i=1}^{N-1}hy_i^2\right)^{1/2}. \tag{20}$$

Таким образом, приходим к следующему выводу. *Если параметры схемы* (2), (3) *связаны неравенством* (18), *то схема устойчива по начальным данным и при любых* $y^0\in H$ *для решения задачи* (2), (3) *с* $\varphi_i^n\equiv 0$ *справедлива оценка*

$$\|y^{n+1}\|\leqslant\|y^0\|,\quad n=0, 1, \ldots, K-1,$$

где норма $\|y\|$ *определена согласно* (20).

Заметим, что неравенство (18) совпадает с полученным в § 4 гл. 1 необходимым условием устойчивости схемы (2), (3).

3. Устойчивость по правой части и сходимость. Чтобы оценить функцию $\tilde{y}^{n+1}$ (см. (15)), усилим условие (18) и потребуем выполнения неравенства

$$\sigma\geqslant\frac{1}{2}-\frac{(1-\varepsilon)\,h^2}{4\tau} \tag{21}$$

с постоянной $\varepsilon\in(0, 1)$. Тогда $\sigma\geqslant\frac{1}{2}-\frac{1-\varepsilon}{\tau\lambda_{N-1}}$, и при любом $k=1, 2, \ldots, N-1$ получим

$$1+\sigma\tau\lambda_k\geqslant\frac{\tau\lambda_k}{2}+1-\frac{(1-\varepsilon)\,\lambda_k}{\lambda_{N-1}}>1-\frac{(1-\varepsilon)\,\lambda_{N-1}}{\lambda_{N-1}}=\varepsilon>0,$$

т. е. $1+\sigma\tau\lambda_k \geqslant \varepsilon > 0$. Из разложения (15) получаем

$$\|\tilde{y}^{n+1}\|^2 = \sum_{k=1}^{N-1} \frac{\tau^2}{(1+\sigma\tau\lambda_k)^2} (\hat{\varphi}_k(t_n))^2 \leqslant \frac{\tau^2}{\varepsilon^2} \sum_{k=1}^{N-1} (\hat{\varphi}_k(t_n))^2,$$

следовательно,

$$\|\tilde{y}^{n+1}\| \leqslant \frac{\tau}{\varepsilon} \|\varphi^n\|. \tag{22}$$

Если $\sigma \geqslant 0$, то условие (21) становится лишним, так как $1+\sigma\tau\lambda_k \geqslant 1$ и оценка (22) выполняется с $\varepsilon = 1$. Из неравенства треугольника

$$\|y^{n+1}\| \leqslant \|\bar{y}^{n+1}\| + \|\tilde{y}^{n+1}\|$$

и оценок (19), (22) получаем неравенство

$$\|y^{n+1}\| \leqslant \|y^n\| + \frac{\tau}{\varepsilon} \|\varphi^n\|, \tag{23}$$

справедливое при $n = 0, 1, \ldots, K-1$. Суммирование (23) по n приводит к оценке

$$\|y^{n+1}\| \leqslant \|y^0\| + \frac{1}{\varepsilon} \sum_{j=0}^{n} \tau \|\varphi^j\|, \tag{24}$$

которая означает устойчивость задачи (2), (3) по начальным данным и по правой части. Из оценки (24), учитывая условие $\tau n \leqslant T$, получим

$$\|y^{n+1}\| \leqslant \|y^0\| + \frac{T}{\varepsilon} \max_{0 \leqslant j \leqslant n} \|\varphi^j\|. \tag{25}$$

Итак, *если выполнено условие* (21) *с* $\varepsilon \in (0, 1)$, *то схема* (2), (3) *устойчива по начальным данным и по правой части, причем для ее решения справедлива оценка* (25). *Если* $\sigma \geqslant 0$ *и выполнено условие* (18), *то справедлива оценка* (25) *с* $\varepsilon = 1$.

Из оценки (25) и требования аппроксимации следует сходимость схемы (2), (3). Для задачи (4) оценка (25) принимает вид

$$\|z^{n+1}\| \leqslant \frac{T}{\varepsilon} \max_{0 \leqslant j \leqslant n} \|\psi^j\|. \tag{26}$$

Следовательно, $\|z^{n+1}\|$ имеет тот же порядок малости, что и погрешность аппроксимации. В частности, при $\sigma = \sigma_* = \frac{1}{2} - \frac{h^2}{12\tau}$, $\varphi_i^n = f(x_i, t_{n+1/2}) + \frac{h^2}{12} f''(x_i, t_{n+1/2}) + O(\tau^2+h^4)$ имеем $\|\psi^j\| = O(\tau^2+h^4)$, а условие устойчивости (21) выполнено с $\varepsilon = 2/3$, поэтому $\|z^{n+1}\| = O(\tau^2+h^4)$, т. е. схема имеет второй порядок точности по τ и четвертый — по h. Если $\sigma = 0{,}5$, $\varphi_i^n = f(x_i, t_{n+1/2}) + O(\tau^2+h^2)$, то условие устойчивости выполнено при любых τ и h и $\|z^{n+1}\| = O(\tau^2+h^2)$. При остальных значениях σ имеем $\|z^{n+1}\| = O(\tau+h^2)$, если выполнено (21) с $\varepsilon \in (0, 1)$, или если $\sigma \geqslant 0$ и выполнено (18).

4. Схема с весами для двумерного уравнения теплопроводности. Пусть область G — прямоугольник $\{0<x_\alpha<l_\alpha,\ \alpha=1,\ 2\}$ с границей Γ. В области $Q=G\times(0,\ T]$ рассмотрим первую краевую задачу для двумерного уравнения теплопроводности

$$\begin{aligned}&\frac{\partial u}{\partial t}=\frac{\partial^2 u}{\partial x_1^2}+\frac{\partial^2 u}{\partial x_2^2},\qquad (x_1,\ x_2,\ t)\in Q,\\ &u(x_1,\ x_2,\ 0)=u_0(x_1,\ x_2),\qquad (x_1,\ x_2)\in\bar{G},\\ &u(x_1,\ x_2,\ t)=0,\qquad (x_1,\ x_2,\ t)\in\Gamma\times(0,\ T].\end{aligned}\tag{27}$$

Оператор Лапласа аппроксимируем так же, как и в § 2, пятиточечным разностным оператором. Для этого введем сетку Ω_h по пространственным переменным следующим образом:

$$\Omega_h=\{x_{ij}=(x_1^{(i)},\ x_2^{(j)})\mid x_1^{(i)}=ih_1,\ x_2^{(j)}=jh_2,\ i=0,\ 1,\ \ldots,\ N_1,\ j=0,\ 1,\ \ldots,\ N_2,\ h_\alpha N_\alpha=l_\alpha,\ \alpha=1,\ 2\}.$$

Множество точек сетки Ω_h, принадлежащих Γ, будем обозначать через γ_h, а множество внутренних точек — через ω_h, так что $\Omega_h=\omega_h\cup\gamma_h$. Определим на Ω_h разностный оператор

$$\begin{aligned}&Ay_{ij}=-y_{\bar{x}_1x_1,ij}-y_{\bar{x}_2x_2,ij},\qquad x_{ij}\in\omega_h,\\ &y_{ij}=0,\quad x_{ij}\in\gamma_h.\end{aligned}\tag{28}$$

По переменной t введем равномерную сетку

$$\omega_\tau=\{t_n=n\tau,\ n=0,\ 1,\ \ldots,\ K,\ K\tau=T\}$$

и обозначим $y_{ij}^n=y(x_1^{(i)},\ x_2^{(j)},\ t_n)$. Дифференциальную задачу (27) аппроксимируем разностной схемой с весами

$$\begin{aligned}&\frac{y_{ij}^{n+1}-y_{ij}^n}{\tau}+\sigma Ay_{ij}^{n+1}+(1-\sigma)Ay_{ij}^n=0,\\ &i=1,\ 2,\ \ldots,\ N_1-1,\quad j=1,\ 2,\ \ldots,\ N_2-1,\ n=0,\ 1,\ \ldots,\ K-1,\\ &y_{ij}^0=u_0(x_1^{(i)},\ x_2^{(j)}),\qquad x_{ij}\in\Omega_h,\\ &y_{ij}^{n+1}=0,\quad x_{ij}\in\gamma_h,\quad n=1,\ 2,\ \ldots,\ K.\end{aligned}\tag{29}$$

При $\sigma=0$ получаем явную схему, для которой решение y_{ij}^{n+1} выражается явным образом через значения y_{ij}^n. Если $\sigma\neq0$, то схема неявная и для нахождения y_{ij}^{n+1} требуется решить систему двумерных разностных уравнений. Методы решения таких систем будут изложены в гл. 5, а один из методов рассматривается в § 6 настоящей главы. Схема имеет второй порядок аппроксимации по h и первый по τ (за исключением случая $\sigma=0{,}5$, когда по τ также второй порядок аппроксимации).

Рассмотрим вопрос об устойчивости схемы (29) по начальным данным. Исследование устойчивости проводится точно так же, как

и в одномерном случае с помощью метода разделения переменных. На самом деле даже нет необходимости повторять проделанные ранее выкладки. Достаточно заметить, что основные результаты об устойчивости схемы (6) не зависели от конкретного вида оператора A, а использовали только следующие его свойства:

1) существование полной ортонормированной системы собственных функций,

2) положительность всех собственных чисел и знание верхней границы $\lambda_{\max}$ спектра.

При этих условиях было доказано, что схема (6) устойчива по начальным данным, если весовой множитель σ удовлетворяет неравенству

$$\sigma \geqslant \frac{1}{2} - \frac{1}{\tau\lambda_{\max}}. \tag{30}$$

Обратим внимание на то, что схема (29) имеет тот же вид, что и схема (6) с $\varphi^n \equiv 0$, однако оператор A определяется теперь по-иному, а именно в соответствии с формулами (28). Как было показано в § 2, оператор (28) обладает перечисленными выше свойствами 1) и 2), причем для него

$$\lambda_{\max} = \frac{4}{h_1^2}\cos^2\frac{\pi h_1}{2l_1} + \frac{4}{h_2^2}\cos^2\frac{\pi h_2}{2l_2} < \frac{4}{h_1^2} + \frac{4}{h_2^2}.$$

Условие устойчивости (30) будет выполнено, если

$$\sigma \geqslant \frac{1}{2} - \frac{1}{\tau\Delta}, \quad \Delta = \frac{4}{h_1^2} + \frac{4}{h_2^2}. \tag{31}$$

Таким образом, *схема* (29) *устойчива по начальным данным при условии* (31). Устойчивость здесь понимается как выполнение при любых начальных данных оценки

$$\|y^n\| \leqslant \|y^0\|, \qquad n = 0, 1, \ldots, K-1,$$

где

$$\|y^n\|^2 = \sum_{i=1}^{N_1-1} h_1 \sum_{j=1}^{N_2-1} h_2 (y_{ij}^n)^2.$$

Аналогично исследуются устойчивость схемы (29) по правой части и ее сходимость. Если $\sigma = 0{,}5$, то схема (29) имеет второй порядок точности по τ и по h, при остальных σ — первый порядок точности по τ и второй — по h.

Условие (31) становится более наглядным, если сетка Ω_h — квадратная, т. е. $h_1 = h_2 = h$. Тогда неравенство (31) принимает вид

$$\sigma \geqslant \frac{1}{2} - \frac{h^2}{8\tau}. \tag{32}$$

В частности, явная схема ($\sigma = 0$) устойчива при условии $\frac{\tau}{h^2} \leqslant \frac{1}{4}$, которое является еще более жестким, чем в одномерном случае.

Неявные схемы с $\sigma \geqslant 0{,}5$ абсолютно устойчивы, однако в отличие от одномерного случая решение неявных двумерных разностных уравнений представляет значительные трудности.

5. Асимптотическая устойчивость. Проведение расчетов на современных быстродействующих ЭВМ предъявляет к разностным схемам наряду с обычными требованиями аппроксимации, устойчивости и сходимости ряд дополнительных требований. Эти требования сводятся к тому, что разностная схема должна хорошо моделировать характерные свойства исходного дифференциального уравнения в условиях, когда шаги сетки остаются конечными. Так, при решении уравнений параболического типа на больших отрезках времени существенное значение имеет свойство асимптотической устойчивости разностной схемы. Поясним понятие асимптотической устойчивости на примере разностных схем для уравнения теплопроводности

$$\begin{gathered}\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < l, \quad t > 0, \\ u(0, t) = u(l, t) = 0, \quad t > 0, \\ u(x, 0) = u_0(x), \quad 0 \leqslant x \leqslant l.\end{gathered} \tag{33}$$

Как известно, решение этой задачи можно записать в виде ряда

$$u(x, t) = \sqrt{\frac{2}{l}} \sum_{k=1}^{\infty} c_k \sin \frac{\pi k x}{l} e^{-\lambda_k t}, \tag{34}$$

где $\lambda_k = \frac{\pi^2 k^2}{l^2}$ — собственные значения оператора второй производной и

$$c_k = \sqrt{\frac{2}{l}} \int_0^l u_0(x) \sin \frac{\pi k x}{l} dx$$

— коэффициенты Фурье функции $u_0(x)$. С ростом t гармоники

$$u_k = c_k e^{-\lambda_k t} \sin \frac{\pi k x}{l}$$

при $k > 1$ затухают быстрее, чем первая гармоника, так что при больших значениях t имеем

$$u(x, t) \approx c_1 \sqrt{\frac{2}{l}} e^{-\lambda_1 t} \sin \frac{\pi x}{l}. \tag{35}$$

Для среднеквадратичной нормы решения задачи (33)

$$\|u(t)\| = \left(\int_0^l u^2(x, t) dx \right)^{1/2}$$

из (34) следует оценка

$$\|u(t)\| \leqslant e^{-\lambda_1 t} \|u(0)\|, \quad \lambda_1 = \frac{\pi^2}{l^2}. \tag{36}$$

Далеко не всякая разностная схема, устойчивая и аппроксимирующая задачу (33), обладает свойствами, аналогичными (35), (36). Рассмотрим семейство схем с весами

$$y^n_{t,i} = \sigma y^{n+1}_{\overline{xx},i} + (1-\sigma)\, y^n_{\overline{xx},i}, \quad i = 1, 2, \ldots, N-1, \quad n = 1, 2, \ldots, \tag{37}$$

$$y^n_0 = y^n_N = 0, \quad y^0_i = u_0(x_i),$$

аппроксимирующее задачу (33). Потребуем, чтобы для решения разностной задачи (37) выполнялась оценка

$$\|y_n\| \leqslant e^{-\delta t_n} \|y_0\|, \tag{38}$$

где $t_n = n\tau$, $\|y\| = \left(\sum_{i=1}^{N-1} h\,(y^n_i)^2 \right)^{1/2}$ и $\delta = \delta(\tau, h) \to \lambda_1$ при $\tau \to 0$, $h \to 0$.

Свойство, выраженное неравенством (38), будем называть асимптотической устойчивостью разностной схемы. Заметим, что устойчивость в обычном смысле определяется как выполнение оценки (38) с $\delta = 0$.

Получим условия асимптотической устойчивости схемы с весами (37). Как было показано ранее, из представления (14) для решения задачи (37) следует оценка

$$\|y^{n+1}\| \leqslant \max_{1 \leqslant k \leqslant N-1} |q_k| \, \|y^n\|, \tag{39}$$

где

$$q_k = 1 - \frac{\tau \lambda^{(h)}_k}{1 + \sigma\tau\lambda^{(h)}_k}, \quad k = 1, 2, \ldots, N-1 \tag{40}$$

и $\lambda^{(h)}_k$ — собственные значения оператора второй разностной производной,

$$0 < \lambda^{(h)}_1 < \lambda^{(h)}_2 < \ldots < \lambda^{(h)}_{N-1}. \tag{41}$$

Устойчивость по начальным данным в обычном смысле обеспечивается условием

$$|q_{N-1}| \leqslant 1, \tag{42}$$

которое можно записать в виде

$$\sigma \geqslant \frac{1}{2} - \frac{1}{\tau \lambda^{(h)}_{N-1}}. \tag{43}$$

Исследование асимптотической устойчивости схемы с весами (37) основано на следующей лемме.

Лемма 1. *Пусть величины q_k определены согласно* (40), (41) *и параметр σ удовлетворяет условию*

$$1 + \sigma\tau\lambda^{(h)}_{N-1} > 0. \tag{44}$$

Если выполнено неравенство

$$q_1 + q_{N-1} > 0, \tag{45}$$

то $q_1 \in (0, 1)$ *и*

$$|q_k| < q_1, \qquad k = 2, 3, \ldots, N-1. \tag{46}$$

Доказательство. Из (41), (44) следует, что

$$1 + \sigma\tau\lambda_k^{(h)} > 1 - \frac{\lambda_k^{(h)}}{\lambda_{N-1}^{(h)}} \geqslant 0,$$

т. е.

$$1 + \sigma\tau\lambda_k^{(h)} > 0, \qquad k = 1, 2, \ldots, N-1.$$

Неравенство (46) эквивалентно выполнению двух неравенств: $q_1 - q_k > 0$ и $q_1 + q_k > 0$. Согласно (40) имеем

$$q_1 - q_k = \frac{\tau(\lambda_k^{(h)} - \lambda_1^{(h)})}{(1 + \sigma\tau\lambda_1^{(h)})(1 + \sigma\tau\lambda_k^{(h)})} > 0, \qquad k = 2, 3, \ldots, N-1.$$

Далее, $q_1 + q_k = (q_1 + q_{N-1}) + (q_k - q_{N-1})$, откуда, учитывая условие (45), получим

$$q_1 + q_k > q_k - q_{N-1} = \frac{\tau(\lambda_{N-1}^{(h)} - \lambda_k^{(h)})}{(1 + \sigma\tau\lambda_k^{(h)})(1 + \sigma\tau\lambda_{N-1}^{(h)})} \geqslant 0,$$

следовательно,

$$q_1 + q_k > 0, \qquad k = 1, 2, \ldots, N-1.$$

Тем самым неравенство (46) выполнено. Из него следует, что $q_1 > 0$. Неравенство $q_1 < 1$ следует из (40), (41), (44). Лемма 1 доказана.

Заметим, что для неотрицательных σ условия (44) всегда выполнены.

Следствие 1. *Если выполнены условия леммы* 1, *то для решения разностной задачи* (37) *справедлива оценка*

$$\|y^n\| \leqslant \rho^n \|y^0\|, \tag{47}$$

где

$$\rho = q_1 = \frac{1 - (1 - \sigma)\tau\lambda_1^{(h)}}{1 + \sigma\tau\lambda_1^{(h)}}. \tag{48}$$

Доказательство следует немедленно из оценок (39), (45).

Следствие 2. *Если выполнены условия леммы* 1 *и параметр* σ *не зависит от* τ *и* h, *то схема с весами* (37) *асимптотически устойчива.*

Доказательство. Согласно (38) достаточно показать, что

$$\rho^n = e^{-\delta t_n}, \tag{49}$$

где $\delta = \delta(h, \tau) \to \pi^2/l^2$ при $\tau \to 0$, $h \to 0$. Переписывая равенство (49) в виде

$$\delta = \frac{1}{\tau} \ln \frac{1}{\rho}, \tag{50}$$

получим из (48), что

$$\delta = \frac{1}{\tau} \ln \frac{1}{1 - \dfrac{\tau\lambda_1^{(h)}}{1 + \sigma\tau\lambda_1^{(h)}}} = \frac{1}{\tau}\left(\frac{\mu_1}{1+\sigma\mu_1} + O(\mu_1^2)\right), \tag{51}$$

где $\mu_1 = \tau\lambda_1^{(h)}$. Заметим, что

$$\lambda_1^{(h)} = \frac{4}{h^2}\sin^2\frac{\pi h}{2l} \to \frac{\pi^2}{l^2}$$

при $h \to 0$ и $\tau\lambda_1^{(h)} \to 0$ при $\tau \to 0$, $h \to 0$.

Поэтому из (51) получим

$$\delta = \frac{\lambda_1^{(h)}}{1 + \sigma\tau\lambda_1^{(h)}} + O(\tau)$$

и $\lim\limits_{\tau, h \to 0} \delta(h, \tau) = \pi^2/l^2$, что и требовалось.

Не представляет труда более подробно выписать асимптотику величины $\delta(h, \tau)$ при $\tau \to 0$. Имеем

$$\delta(h, \tau) = \frac{1}{\tau}\ln\frac{1}{\rho} = \lambda_1^{(h)} - (\sigma^2 - (1-\sigma)^2)\frac{\tau(\lambda_1^{(h)})^2}{2} + (\sigma^3 + (1-\sigma)^3)\frac{\tau^2(\lambda_1^{(h)})^3}{3} -$$

$$- (\sigma^4 - (1-\sigma)^4)\frac{\tau^3(\lambda_1^{(h)})^4}{4} + (\sigma^5 + (1-\sigma)^5)\frac{\tau^4(\lambda_1^{(h)})^5}{5} + O(\tau^5),$$

откуда видно, в частности, что $\delta(h, \tau) = \lambda_1^{(h)} + O(\tau^2)$ при $\sigma = 0{,}5$.

Таким образом, неравенство (45) представляет собой условие асимптотической устойчивости схемы с весами (37). Его можно переписать в виде неравенства

$$\frac{\tau\lambda_1^{(h)}}{1 + \sigma\tau\lambda_1^{(h)}} + \frac{\tau\lambda_{N-1}^{(h)}}{1 + \sigma\tau\lambda_{N-1}^{(h)}} < 2, \tag{52}$$

где $\lambda_1^{(h)} = \frac{4}{h^2}\sin^2\frac{\pi h}{2l}$, $\lambda_{N-1}^{(h)} = \frac{4}{h^2}\cos^2\frac{\pi h}{2l}$.

Заметим, что из (52) и (44) следует неравенство

$$\frac{\tau\lambda_{N-1}^{(h)}}{1 + \sigma\tau\lambda_{N-1}^{(h)}} \leqslant 2,$$

совпадающее с (17) и обеспечивающее устойчивость схемы (37) в обычном смысле.

Из (52) получаем, что явная схема ($\sigma = 0$) асимптотически устойчива при условии $\tau < 0{,}5\,h^2$. Чисто неявная схема ($\sigma = 1$) асимптотически устойчива при любых τ и h. Симметричная схема ($\sigma = 0{,}5$) асимптотически устойчива при условии

$$\tau < 2/\sqrt{\lambda_1^{(h)}\lambda_{N-1}^{(h)}} \approx hl/\pi. \tag{53}$$

Таким образом, симметричная схема, будучи абсолютно устойчивой в обычном смысле, является условно асимптотически устойчивой при условии (53).

Асимптотическая устойчивость разностной схемы тесно связана с ее точностью. Нарушение асимптотической устойчивости приводит к потере точности схемы на больших временах. Так, в [32] показано, что если положить

$$\tau = m\frac{2}{\sqrt{\lambda_1^{(h)}\lambda_{N-1}^{(h)}}}, \quad m > 1,$$

то при больших t решение $y(t_n, x_i)$ симметричной разностной схемы имеет асимптотику

$$y(t_n, x_i) \approx c_{N-1}e^{-\lambda_1 t_n/(4m^2)}(-1)^{i-1}\sin\frac{\pi x_i}{l}.$$

Сопоставляя с асимптотикой (35) решения исходной задачи, видим, что решение полностью искажается.

Отметим, что в [32] предложена разностная схема для уравнения теплопроводности, обладающая безусловной асимптотической устойчивостью и имеющая второй порядок точности, однако данная схема не принадлежит семейству схем с весами (37).

§ 4. Решение разностного уравнения второго порядка методом Фурье

Рассмотрим разностную схему

$$y_{\bar{x}x,i} = -f_i, \quad i = 1, 2, \ldots, N-1, \quad y_0 = y_N = 0 \tag{1}$$

и построим ее решение в виде разложения по базису собственных функций оператора

$$(Ay)_i = -y_{\bar{x}x,i}, \quad i = 1, 2, \ldots, N-1, \quad hN = l, \quad y_0 = y_N = 0. \tag{2}$$

Оператор (2) подробно изучался в § 1, где было показано, что он имеет полную ортонормированную систему собственных функций

$$\mu_k(x_i) = \sqrt{\frac{2}{l}}\sin\frac{\pi k x_i}{l}, \quad k = 1, 2, \ldots, N-1, \quad i = 1, 2, \ldots, N-1.$$

Соответствующие собственные числа оператора A имеют вид

$$\lambda_k = \frac{4}{h^2}\sin^2\frac{\pi k h}{2l}.$$

Поэтому можно искать решение задачи (1) в виде

$$y_j = y(x_j) = \sum_{k=1}^{N-1} c_k\mu_k(x_j), \quad j = 1, 2, \ldots, N-1, \tag{3}$$

где c_k — неизвестные пока коэффициенты.

Разложим правую часть уравнения (1) в сумму Фурье, т. е. представим ее в виде

$$f_j = \sum_{k=1}^{N-1} \hat{f}_k \mu_k(x_j),$$

где

$$\hat{f}_k = (f, \mu_k) = \sum_{j=1}^{N-1} f_j \mu_k(x_j) h, \quad k = 1, 2, \ldots, N-1. \tag{4}$$

Подставляя разложения (3), (4) в уравнение (1) при $i=j$, получим

$$\sum_{k=1}^{N-1} c_k (\mu_k(x))_{\overline{x}x,j} = -\sum_{k=1}^{N-1} \hat{f}_k \mu_k(x_j),$$

откуда, учитывая соотношение $(\mu_k(x))_{\overline{x}x,j} = -\lambda_k \mu_k(x_j)$ и линейную независимость функций $\mu_k(x)$, приходим к уравнениям

$$c_k \lambda_k = \hat{f}_k, \quad k = 1, 2, \ldots, N-1.$$

Отсюда находим значения коэффициентов Фурье функции $y(x_j)$:

$$c_k = \hat{f}_k / \lambda_k, \quad k = 1, 2, \ldots, N-1. \tag{5}$$

Таким образом, приходим к следующему алгоритму решения разностной краевой задачи (1). Сначала по заданной правой части f_j и известным собственным функциям $\mu_k(x_j)$ вычисляем по формулам (4) коэффициенты Фурье правой части. Затем по формулам (5), пользуясь тем, что собственные числа известны в явном виде, находим коэффициенты Фурье c_k искомого решения $y(x_j)$. И, наконец, вычисляя суммы (3), находим решение $y(x_j)$.

Подсчитаем число умножений и делений, необходимое для нахождения указанным способом решения задачи (1). Для вычисления $\hat{f}_k$ при каждом k требуется $N-1$ умножений, а вычисление всех $\hat{f}_k$, $k=1, 2, \ldots, N-1$, требует $(N-1)^2$ умножений. Следует подчеркнуть, что здесь и далее мы предполагаем все функции $\mu_k(x_j)$ и числа λ_k уже вычисленными и хранящимися в памяти машины. Вычисление коэффициентов c_k по формулам (5) требует $N-1$ делений. Вычисление y_j при фиксированном j по формулам (3) требует $N-1$ умножений, а вычисление всех y_j, $j=1, 2, \ldots, N-1$ требует $(N-1)^2$ умножений. Таким образом, весь алгоритм осуществляется за $2(N-1)^2$ умножений и $N-1$ делений. Вспомним, что уравнение (1) можно решить методом прогонки (см. п. 7 § 4 ч. I) по формулам

$$\alpha_{i+1} = \frac{1}{2-\alpha_i}, \quad i = 1, 2, \ldots, N-1, \quad \alpha_1 = 0,$$

$$\beta_{i+1} = \alpha_{i+1}(\beta_i + h^2 f_i), \quad i = 1, 2, \ldots, N-1, \quad \beta_1 = 0,$$

$$y_i = \alpha_{i+1} y_{i+1} + \beta_{i+1}, \quad i = N-1, \ N-2, \ldots, 1, \quad y_N = 0$$

всего за $2(N-1)$ умножений и $N-1$ деление.

Следовательно, предложенный здесь метод Фурье неэкономичен, он требует $O(N^2)$ действий вместо $O(N)$ действий в методе прогонки. Пользоваться таким методом для решения одномерных разностных краевых задач нецелесообразно. Однако данный метод в сочетании с методом быстрого вычисления сумм вида (3) нашел применение при решении двумерных разностных уравнений с постоянными коэффициентами. Эти вопросы рассматриваются в следующих параграфах.

§ 5. Быстрое дискретное преобразование Фурье

Наибольшее число действий в методе Фурье требуется для вычисления сумм (3), (4). Эти суммы имеют вид

$$s_j = \sum_{k=1}^{N-1} z_k \sin \frac{\pi kj}{N}, \quad j = 1, 2, \ldots, N-1, \tag{1}$$

где z_k — заданные числа.

Для непосредственного вычисления всех s_j, $j=1, 2, \ldots, N-1$, требуется, как уже отмечалось, $(N-1)^2$ умножений. Сейчас мы, следуя [21], изложим метод вычисления сумм вида (1), требующий $O(N \ln N)$ действий умножения. Такое ускоренное вычисление сумм основано на том, что среди чисел $\sin \frac{\pi kj}{N}$, $k=1, 2, \ldots, N-1$, $j=1, 2, \ldots, N-1$, есть много одинаковых. Поэтому можно перегруппировать слагаемые и уменьшить тем самым число умножений.

Для дальнейшего удобно преобразовать сумму (1) к виду

$$s_j = \sum_{k=1}^{N-1} z_k \sin \frac{2\pi kj}{2N} = \sum_{k=0}^{2N-1} z_k \sin \frac{2\pi kj}{2N},$$

где полагаем $z_0 = z_N = z_{N+1} = \ldots = z_{2N-1} = 0$. Обозначим $M=2N$ и рассмотрим более общую задачу о вычислении суммы

$$v_j = \sum_{k=0}^{M-1} z_k e^{i\frac{2\pi kj}{M}} = \sum_{k=0}^{M-1} z_k w^{kj},$$

где

$$w = e^{i\frac{2\pi}{M}} \tag{2}$$

и i — мнимая единица. Итак, будем вычислять суммы

$$v_j = \sum_{k=0}^{M-1} z_k w^{kj}, \quad j = 0, 1, \ldots, M-1. \tag{3}$$

Ясно, что $\operatorname{Im} v_j = s_j$ для действительных z_k. Для дальнейшего существенным является условие $M=2^m$, $m>0$, означающее, что число точек сетки $N=2^{m-1}$ является степенью двойки.

Представим число k в формуле (3) в двоичной системе $k = k_0 + 2k_1 + 2^2k_2 + \ldots + 2^{m-1}k_{m-1}$, где k_i либо нуль, либо единица.

Обозначим

$$k=(k_{m-1}, k_{m-2}, \ldots, k_0), \quad z_k=z(k_0, k_1, \ldots, k_{m-1}).$$

Тогда сумму (3) можно записать в виде

$$v_j = \sum_{k_0,k_1,\ldots,k_{m-1}} z(k_0, k_1, \ldots, k_{m-1})\, w^{(k_0+2k_1+\ldots+2^{m-1}k_{m-1})j} =$$

$$= \sum_{k_0=0}^{1} w^{k_0 j}\left[\sum_{k_1=1}^{1} w^{2k_1 j} \ldots \sum_{k_{m-1}=0}^{1} w^{2^{m-1}k_{m-1}j} z(k_0, k_1, \ldots, k_{m-1})\right]. \quad (4)$$

Преобразуем внутреннюю сумму

$$\sum_{k_{m-1}=0}^{1} w^{2^{m-1}k_{m-1}j} z(k_0, k_1, \ldots, k_{m-1}). \quad (5)$$

Представим число j в двоичной системе $j=j_0+2j_1+\ldots+2^{m-1}j_{m-1}$. Тогда получим

$$w^{2^{m-1}k_{m-1}j} = (w^{2^{m-1}k_{m-1}j_0})\,(w^{2^{m-1}k_{m-1}2j_1}) \ldots (w^{2^{m-1}k_{m-1}j_{m-1}}). \quad (6)$$

В этом произведении все сомножители начиная со второго равны единице. Действительно, $w^{2^{m-1}k_{m-1}2j_1} = w^{2^m k_{m-1}j_1} = w^{Mk_{m-1}j_1}$ и, вспоминая выражение (2) для w, получим $w^M=1$, $w^{2^{m-1}k_{m-1}2j_1}=1$, поскольку $k_{m-1}j_1$ равно либо нулю, либо единице.

Точно так же равны единице и последующие сомножители в произведении (6). Таким образом, $w^{2^{m-1}k_{m-1}j} = w^{2^{m-1}k_{m-1}j_0}$ и сумму (5) можно записать в виде

$$\sum_{k_{m-1}=0}^{1} w^{2^{m-1}k_{m-1}j_0} z(k_0, k_1, \ldots, k_{m-1}).$$

Обозначая эту сумму через $z_1(j_0, k_1, \ldots, k_{m-2})$, запишем предпоследнюю сумму в выражении (4) в виде

$$\sum_{k_{m-2}=0}^{1} w^{2^{m-2}k_{m-2}j} z_1(j_0, k_1, k_2, \ldots, k_{m-2}). \quad (7)$$

Представим число $2^{m-2}k_{m-2}j$ в виде $2^{m-2}k_{m-2}(j_0+2j_1)+2^m k_{m-2}(j_2+\ldots+2^{m-3}j_{m-1})$, откуда получим $w^{2^{m-2}k_{m-2}j} = w^{2^{m-2}k_{m-2}(j_0+2j_1)}$. Следовательно, сумма (7) равна

$$z_2(j_0, j_1, k_1, \ldots, k_{m-3}) = \sum_{k_{m-2}=0}^{1} w^{2^{m-2}k_{m-2}(j_0+2j_1)} z_1(j_0, k_1, \ldots, k_{m-2}).$$

Далее, аналогичный процесс последовательного вычисления сумм

продолжается до тех пор, пока в (4) не исчерпаются все суммы. Последняя сумма имеет вид $v_j = \sum_{k_0=0}^{1} w^{k_0 j} z_{m-1}(j_0, j_1, \ldots, j_{m-2}, k_0)$.

Итак, можно предложить следующий алгоритм вычисления сумм вида (3), называемый *алгоритмом быстрого дискретного преобразования Фурье*. Числа k и j представляются в двоичной системе

$$k = k_0 + 2k_1 + 2^2 k_2 + \ldots + 2^{m-1} k_{m-1},$$
$$j = j_0 + 2j_1 + 2^2 j_2 + \ldots + 2^{m-1} j_{m-1},$$

где k_i, j_i — либо нуль, либо единица. Далее обозначается $z_k = z(k_0, k_1, \ldots, k_{m-1})$ и последовательно вычисляются суммы, состоящие каждая из двух слагаемых:

$$z_1(j_0, k_0, k_1, \ldots, k_{m-2}) = \sum_{k_{m-1}=0}^{1} w^{2^{m-1} k_{m-1} j_0} z(k_0, k_1, \ldots, k_{m-1}),$$

$$z_2(j_0, j_1, k_0, k_1, \ldots, k_{m-3}) = \sum_{k_{m-2}=0}^{1} w^{2^{m-2} k_{m-2}(j_0 + 2j_1)} z_1(j_0, k_0, k_1, \ldots, k_{m-2}),$$

. . . . . .

$$z_{m-1}(j_0, j_1, \ldots, j_{m-2}, k_0) = \sum_{k_1=0}^{1} w^{2k_1(j_0 + 2j_1 + \ldots + 2^{m-2} j_{m-2})} \times$$
$$\times z_{m-2}(j_0, j_1, \ldots, j_{m-2}, k_0, k_1),$$

$$v_j = \sum_{k_0=0}^{1} w^{k_0 j} z_{m-1}(j_0, j_1, \ldots, j_{m-2}, k_0).$$

Подсчитаем число умножений, необходимое для нахождения всех сумм v_j, $j = 0, 1, \ldots, M-1$, при указанном способе вычислений. Функция $z_1(j_0, k_0, k_1, \ldots, k_{m-2})$ используется только при вычислении функции $z_2(j_0, j_1, k_0, k_1, \ldots, k_{m-3})$. При этом необходимо вычислить значения $z_1(j_0, k_0, \ldots, k_{m-2})$ дважды: при $k_{m-2} = 0$ и $k_{m-2} = 1$. Вычисление $z_1(j_0, k_0, \ldots, k_{m-2})$ при каждом значении k_{m-2} требует двух умножений. Следовательно, общее число умножений, требуемое для вычисления $z_1(j_0, k_0, \ldots, k_{m-2})$, равно четырем. Такое же число умножений требуется для вычисления каждой из сумм $z_l(j_0, j_1, \ldots, j_{l-1}, k_0, k_1, \ldots, k_{m-l-1})$. Всего имеется m таких сумм. Поэтому число умножений, необходимое для вычисления v_j при каждом фиксированном j, равно $4m$, а для вычисления всех v_j, $j = 0, 1, \ldots, M$, это число равно $4mM = 4M \log_2 M$. При больших $N = 2^{m-1}$ это приводит к значительному сокращению числа умножений по сравнению с числом умножений $(N-1)^2$, требуемому при непосредственном вычислении сумм вида (1). Так, при $N = 128$ число умножений будет почти в два раза меньше.

§ 6. Решение разностного уравнения Пуассона с использованием быстрого преобразования Фурье

В § 1 гл. 2 рассматривалась разностная аппроксимация задачи Дирихле для уравнения Пуассона

$$y_{\overline{x}_1 x_1, ij} + y_{\overline{x}_2 x_2, ij} = -f_{ij}, \quad x_{ij} \in \omega_h, \tag{1}$$

$$y_{ij} = 0, \quad x_{ij} \in \gamma_h.$$

Здесь ω_h — множество внутренних и γ_h — множество граничных узлов сетки

$$\Omega_h = \{x_{ij} = (x_1^{(i)}, x_2^{(j)}) \mid x_1^{(i)} = ih_1, \; x_2^{(j)} = jh_2,$$
$$i = 0, 1, \ldots, N_1, \; j = 0, 1, \ldots, N_2, \; h_\alpha N_\alpha = l_\alpha, \; \alpha = 1, 2\}.$$

Разностная схема (1) представляет собой систему большого числа линейных алгебраических уравнений, матрица которой является сильно разреженной, т. е. содержит значительно больше нулевых элементов, чем ненулевых (в данном случае в каждой строке матрицы не более пяти ненулевых элементов). Решать такие системы уравнений с помощью численных методов, предназначенных для систем общего вида, нецелесообразно, а часто даже и невозможно из-за большого размера матрицы. Поэтому развиты специальные методы, прямые и итерационные, пригодные для решения двумерных разностных уравнений. Они подробно рассматриваются в гл. 5, а сейчас мы познакомимся с одним способом решения задачи (1), сочетающим одномерную прогонку с методом Фурье. Заметим, что предположение об однородности граничных условий не ограничивает общности, так как неоднородные условия Дирихле можно включить в правую часть подобно тому, как это было сделано в начале § 1 для одномерного случая.

Рассмотрим одномерную задачу на собственные значения

$$\frac{\mu(j+1) - 2\mu(j) + \mu(j-1)}{h_2^2} + \lambda\mu(j) = 0, \quad j = 1, 2, \ldots, N_2 - 1,$$
$$h_2 N_2 = l_2, \quad \mu_0 = \mu_{N_2} = 0, \quad \mu(j) = \mu(x_2^{(j)}). \tag{2}$$

В § 1 было показано, что задача (2) имеет следующее решение:

$$\mu(j) = \mu_k(j) = \sqrt{\frac{2}{l_2}} \sin \frac{\pi k x_2^{(j)}}{l_2}, \quad \lambda_k = \frac{4}{h_2^2} \sin^2 \frac{\pi k h_2}{2l_2}, \tag{3}$$
$$k = 1, 2, \ldots, N_2 - 1.$$

Зафиксируем какое-либо значение индекса i, $0 < i < N_1$, и будем рассматривать y_{ij}, f_{ij} как функции, зависящие только от j, $j = 1, 2, \ldots, N_2 - 1$. Тогда можно разложить y_{ij}, f_{ij} по собственным функциям задачи (2), т. е. представить их в виде

$$y_{ij} = \sum_{k=1}^{N_2-1} c_k(i)\,\mu_k(j), \quad f_{ij} = \sum_{k=1}^{N_2-1} \hat{f}_k(i)\,\mu_k(j). \tag{4}$$

Подставляя разложения (4) в уравнение (1), получим

$$\sum_{k=1}^{N_2-1} \mu_k(j)\,(c_k(i))_{\overline{x}_1x_1,i} + \sum_{k=1}^{N_2-1} c_k(i)\,(\mu_k(j))_{\overline{x}_2x_2,j} = -\sum_{k=1}^{N_2-1} \hat{f}_k(i)\,\mu_k(j),$$

откуда, учитывая уравнение (2) и линейную независимость μ_k, приходим к уравнениям $(c_k(i))_{\overline{x}_1x_1,i} - \lambda_k c_k(i) + \hat{f}_k(i) = 0$. Таким образом, для нахождения коэффициента c_k, $k = 1, 2, \ldots, N_2-1$, в разложении (4) получаем систему разностных уравнений второго порядка

$$\frac{c_k(i+1) - 2c_k(i) + c_k(i-1)}{h_1^2} - \lambda_k c_k(i) + \hat{f}_k(i) = 0, \tag{5}$$

$$i = 1, 2, \ldots, N_1-1, \quad c_k(0) = c_k(N_1) = 0.$$

Здесь числа λ_k заданы согласно (3), а значения $\hat{f}_k(i)$ вычисляются по правилу

$$\hat{f}_k(i) = \sum_{j=1}^{N_2-1} h_2 f_{ij}\mu_k(j), \qquad i = 1, 2, \ldots, N_1 - 1. \tag{6}$$

Уравнение (5) решается методом прогонки

$$\alpha_{i+1}^{(k)} = \frac{1}{2 + h_1^2\lambda_k - \alpha_i^{(k)}}, \quad \beta_{i+1}^{(k)} = \alpha_{i+1}^{(k)}\,(\beta_i^{(k)} + h_1^2\hat{f}_k(i)),$$

$$i = 1, 2, \ldots, N_1 - 1, \quad \alpha_1^{(k)} = \beta_1^{(k)} = 0,$$

$$c_k(i) = \alpha_{i+1}^{(k)} c_k(i+1) + \beta_{i+1}, \quad i = N_1 - 1, N_1 - 2, \ldots, 1, \quad c_k(N_1) = 0.$$

Таким образом, рассматриваемый алгоритм решения задачи (1) состоит в следующем. Сначала по формулам (6) вычисляются коэффициенты Фурье правой части f_{ij}. При каждом фиксированном i суммы вида (6) можно вычислить для $k = 1, 2, \ldots, N_2-1$ с помощью быстрого дискретного преобразования Фурье за число действий $O(N_2 \ln N_2)$, а вычисление этих сумм для всех $i = 1, 2, \ldots, N_1-1$ потребует $O(N_1 N_2 \ln N_2)$ действий. Затем надо решить методом прогонки уравнения (5) для $k = 1, 2, \ldots, N_2-1$, что потребует $O(N_1 N_2)$ действий. Наконец, зная коэффициенты Фурье $c_k(i)$, можно восстановить решение y_{ij} по формулам

$$y_{ij} = \sum_{k=1}^{N_2-1} c_k(i)\,\mu_k(j), \quad i = 1, 2, \ldots, N_1 - 1, \quad j = 1, 2, \ldots, N_2-1,$$

которые аналогичны формулам (6) и требуют того же числа действий $O(N_1 N_2 \ln N_2)$.

Следовательно, изложенный здесь алгоритм может быть реализован за число действий $O(N_1 N_2 \ln N_2)$. Для сравнения отметим, что обычный метод исключения Гаусса потребовал бы $O(N^6)$ действий и, кроме того, громадной машинной памяти.

Недостатком данного метода является необходимость построения в явном виде собственных чисел и собственных функций одно-

мерной задачи. В случае, когда решение задачи на собственные значения в явном виде выписать невозможно (например, для краевых условий третьего рода или в случае переменных неразделяющихся коэффициентов), данный метод неприменим.

Заметим еще, что рассмотренный метод можно применять и для решения неявных разностных уравнений, возникающих при аппроксимации двумерных нестационарных задач, подобных тем, которые рассматривались в п. 4 § 3. В этом случае уравнения, аналогичные (1), приходится решать многократно (на каждом временном слое), поэтому особенно важной становится экономия числа действий, которую обеспечивает данный метод.

ГЛАВА 4

ТЕОРИЯ УСТОЙЧИВОСТИ РАЗНОСТНЫХ СХЕМ

В главе 3 уже изучалась устойчивость разностных схем, аппроксимирующих уравнение теплопроводности. В настоящей главе изучается устойчивость двуслойных и трехслойных линейных разностных схем общего вида. Разностные схемы рассматриваются независимо от тех или иных исходных уравнений и определяются как операторные уравнения с операторами, действующими в евклидовом пространстве. Условия устойчивости формулируются в виде операторных неравенств. Применение теории к исследованию устойчивости конкретных разностных схем состоит в приведении этих схем к каноническому виду и проверке выполнения операторных неравенств.

Параграф 1 носит вспомогательный характер, в нем на примерах поясняется, что разностную схему можно рассматривать как операторное уравнение; при этом корректность схемы определяется не структурой разностного оператора, а его общими свойствами, такими, как самосопряженность и положительная определенность. В § 2, 3 излагаются элементы теории устойчивости двуслойных и трехслойных разностных схем, а в § 4 теория устойчивости применяется к исследованию экономичных разностных схем для многомерных задач математической физики.

§ 1. Разностные схемы как операторные уравнения

1. Представление разностных схем в виде операторных уравнений. Разностные схемы возникают в результате аппроксимаций той или иной задачи математической физики и предназначены для ее приближенного решения. Поэтому в теории разностных схем важное место занимают вопросы аппроксимации дифференциальных уравнений разностными и сходимости решений разностных задач к решениям исходных дифференциальных задач. Однако будучи построенной, разностная схема превращается в самостоятельный математический объект и может изучаться вне связи с породившей ее дифференциальной задачей. При этом отпадают проблемы

аппроксимации и сходимости и остается лишь проблема корректности разностной схемы, т. е. ее разрешимости и устойчивости.

Разностная схема представляет собой систему линейных алгебраических уравнений. Ее всегда можно записать в векторной форме

$$Ay=\varphi, \tag{1}$$

где A — матрица системы, y — искомый вектор и φ — заданный вектор, определяемый правыми частями разностных уравнений и дополнительными (начальными и граничными) условиями. Такая запись наиболее удобна для стационарных разностных задач, в случае же двуслойных и трехслойных разностных схем будем использовать другую форму записи (см. § 2, 3).

Уравнение (1) можно рассматривать также как операторное уравнение, где A — линейный оператор, действующий в конечномерном пространстве H, y — искомый элемент этого пространства и $\varphi\in H$ — заданный элемент. Для разностных схем характерно, что каждая схема определяет не одно уравнение (1), а целое семейство уравнений

$$A_h y_h=\varphi_h, \tag{2}$$

зависящее от шага сетки h. При каждом допустимом значении h оператор A_h действует в конечномерном пространстве H_h. Размерность пространства H_h зависит от шага сетки h и, как правило, неограниченно возрастает при $h\to 0$.

Приведем несколько примеров записи разностной схемы в виде операторного уравнения (2). Чтобы записать конкретную схему в виде (2), надо ввести соответствующим образом пространства H_h, определить операторы A_h и задать правые части φ_h. Следующий пример уже рассматривался в § 1 гл. 3.

Пример 1. На сетке

$$\Omega_h=\{x_i=ih,\ i=0,\ 1,\ \ldots,\ N,\ hN=l\}$$

рассматривается разностная схема

$$y_{\overline{x}x,i}=-f_i,\quad i=1,\ 2,\ \ldots,\ N-1,\qquad y_0=\mu_1,\ y_N=\mu_2. \tag{3}$$

Перепишем систему (3) в виде

$$\begin{gathered}\frac{2y_1-y_2}{h^2}=\widetilde{f}_1,\quad -y_{\overline{x}x,i}=f_i,\qquad i=2,\ 3,\ \ldots,\ N-2,\\ \frac{-y_{N-2}+2y_{N-1}}{h^2}=\widetilde{f}_{N-1},\end{gathered} \tag{4}$$

где $\widetilde{f}_1=f_1+\mu_1/h^2$, $\widetilde{f}_{N-1}=f_{N-1}+\mu_2/h^2$. Введем пространства H_{N-1} размерности $N-1$, состоящие из функций $y(x)$, заданных для $x\in\omega_h$,

$$\omega_h=\{x_i=ih,\ i=1,\ 2,\ \ldots,\ N-1,\ hN=l\}.$$

Определим в H_{N-1} оператор A и вектор φ следующим образом:

$$(Ay)_1 = \frac{2y_1 - y_2}{h^2}, \quad (Ay)_i = -y_{\bar{x}x,i}, \quad i = 2, 3, \ldots, N-2, \quad (Ay)_{N-1} = \frac{-y_{N-2} + 2y_{N-1}}{h^2}, \tag{5}$$

$$\varphi_1 = f_1, \quad \varphi_i = f_i, \quad i = 2, 3, \ldots, N-2, \quad \varphi_{N-1} = f_{N-1}. \tag{6}$$

Тогда разностную схему (4) (или, что то же самое, разностную схему (3)) можно записать в операторной форме (1). Матрица этого оператора является симметричной и трехдиагональной. Например, для случая $N=6$ она имеет вид

$$A = \frac{1}{h^2}\begin{bmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix}.$$

Возможно и несколько иное определение оператора A, позволяющее записать выражения для его компонент единообразно во всех точках сетки ω_h. Пусть H^0_{N-1} — подпространство функций, заданных на сетке Ω_h и обращающихся в нуль при $i=0$, $i=N$. Введем оператор A, действующий из H^0_{N-1} в H_{N-1} и определенный формулами

$$(Ay)_i = -y_{\bar{x}x,i}, \quad i = 1, 2, \ldots, N-1, \quad y_0 = y_N = 0, \tag{7}$$

и зададим вектор φ согласно (6). Тогда по-прежнему разностную схему (3) можно записать в виде $Ay=\varphi$, где $y \in H^0_{N-1}$, $\varphi \in H_{N-1}$. Такое определение оператора A мы уже использовали в § 1 гл. 3. Подчеркнем, что формулы (5) и (7) определяют, по существу, один и тот же оператор.

Разностные схемы для многомерных задач также можно представить в операторной форме (1).

Пример 2. В области $G(0<x_\alpha<l_\alpha, \alpha=1, 2)$ введем сетку

$$\Omega_h = \{x_{ij} = (x_1^{(i)}, x_2^{(j)}) \mid x_1^{(i)} = ih_1, x_2^{(j)} = jh_2, \\ i = 0, 1, \ldots, N_1, j = 0, 1, \ldots, N_2, h_1N_1 = l_1, h_2N_2 = l_2\}.$$

Пусть γ_h — множество узлов сетки Ω_h, принадлежащих границе области G и ω_h — множество внутренних узлов сетки Ω_h. Рассмотрим разностную схему, аппроксимирующую уравнение Пуассона

$$y_{\bar{x}_1x_1,ij} + y_{\bar{x}_2x_2,ij} = -f_{ij}, \quad \text{если } x_{ij} \in \omega_h, \tag{8}$$

$$y_{ij} = 0, \quad \text{если } x_{ij} \in \gamma_h.$$

Пусть $H(\Omega_h)$ и $H(\omega_h)$ — линейные пространства функций, заданных соответственно на сетках Ω_h и ω_h. Введем также подпространство $H^0(\Omega_h)$ функций, заданных на Ω_h и равных нулю на γ_h. Размер-

ность этого подпространства совпадает с числом внутренних узлов сетки Ω_h и равна $(N_1-1)(N_2-1)$. Задаче (8) соответствует оператор A, действующий из $H^0(\Omega_h)$ в $H(\omega_h)$ и определенный формулами

$$(Ay)_{ij} = -y_{\overline{x}_1x_1,ij} - y_{\overline{x}_2x_2,ij}, \quad \text{если } x_{ij} \in \omega_h; \quad y_{ij} = 0, \quad \text{если } x_{ij} \in \gamma_h. \tag{9}$$

Разностную схему (8) можно записать в виде (1), где оператор A определен согласно (9), а компоненты вектора $\varphi \in H(\omega_h)$ задаются формулами $\varphi_{ij} = \varphi(x_{ij}) = f_{ij}$, если $x_{ij} \in \omega_h$. Свойства оператора (9) подробно изучались в § 2 гл. 3.

Можно было бы, так же как и в примере 1, определить оператор A как оператор, действующий из $H(\omega_h)$ в $H(\omega_h)$. Однако в данном случае это привело бы к громоздким формулам. Например, только на одной части границы при $i=1$ надо было бы задать значения оператора A формулами

$$(Ay)_{1j} = \frac{2y_{1j} - y_{2j}}{h_1^2} - y_{\overline{x}_2x_2,ij}, \qquad j = 2, 3, \ldots, N_2 - 2,$$

$$(Ay)_{11} = \frac{2y_{11} - y_{21}}{h_1^2} + \frac{2y_{11} - y_{12}}{h_2^2},$$

$$(Ay)_{1N_2-1} = \frac{2y_{1N_2-1} - y_{2N_2-1}}{h_1^2} + \frac{2y_{1N_2-1} - y_{1N_2-2}}{h_2^2}.$$

Поэтому удобнее использовать определение (9) оператора A.

Если краевые условия (8) неоднородные, то, по аналогии с примером 1, меняем правую часть в приграничных узлах.

2. Корректность операторных уравнений. Рассмотрим семейство операторных уравнений (2), где A_h — линейный оператор, действующий в конечномерном линейном пространстве H_h. Предположим, что в пространстве H_h заданы нормы $\|\cdot\|_{(1_h)}$ и $\|\cdot\|_{(2_h)}$, в которых измеряются, соответственно, решение уравнения (2) и его правая часть. В соответствии с определениями, введенными в § 6 гл. 1, будем называть уравнение (2) *корректным*, если

1) решение уравнения (2) существует и единственно при любых $\varphi_h \in H_h$,

2) существует константа $M_1 > 0$, не зависящая от h и такая, что при любых $\varphi_h \in H_h$ выполняется оценка

$$\|y_h\|_{(1_h)} \leqslant M_1 \|\varphi_h\|_{(2_h)}. \tag{10}$$

Как уже отмечалось, условие 1) эквивалентно существованию оператора A_h^{-1}, а условие 2) — равномерной по h ограниченности A_h^{-1}.

Замечание. Свойство 2), выраженное оценкой (10), называется *устойчивостью разностной схемы* (2). Вообще, устойчивость какой-либо задачи означает, что при небольшом изменении входных данных решение изменяется мало. Таким образом, для исследования устойчивости необходимо рассматривать уравнение, которому удовлетворяет погрешность, возникающая в результате возму-

щения входных данных. Однако в случае линейного оператора A_h структура уравнения для погрешности та же, что и основного уравнения (2). Поэтому при исследовании устойчивости достаточно ограничиться оценкой (10). Действительно, рассмотрим наряду с (2) уравнение

$$A_h v_h = \varphi_h^{(1)},$$

отличающееся от (2) правой частью. Для погрешности $z_h = y_h - v_h$ получим уравнение

$$A_h z_h = \delta\varphi_h,$$

где $\delta\varphi_h = \varphi_h - \varphi_h^{(1)}$ — возмущение правой части. Если выполнена оценка (10), то

$$\| z_h \|_{(1_h)} \leqslant M_1 \| \delta\varphi_h \|_{(2_h)},$$

следовательно, $\| z_h \|_{(1_h)} \to 0$ при $\| \delta\varphi_h \|_{(2_h)} \to 0$ и задача (2) устойчива.

Нетрудно получить некоторые достаточные условия корректности. Предположим, что H_h — вещественное конечномерное пространство, в котором введены скалярное произведение $(y, v)_h$ и норма $\|y\|_h = \sqrt{(y, y)_h}$. Справедливо следующее утверждение.

Если существует постоянная $\delta > 0$, не зависящая от h и такая, что при любом $v_h \in H_h$ выполнено неравенство

$$(A_h v_h, v_h)_h \geqslant \delta \| v_h \|_h^2, \tag{11}$$

то уравнение (2) *корректно и для его решения выполняется оценка*

$$\| y_h \|_h \leqslant \delta^{-1} \| \varphi_h \|_h. \tag{12}$$

Чтобы доказать существование и единственность решения уравнения (2), достаточно убедиться в том, что однородное уравнение

$$A_h z_h = 0 \tag{13}$$

имеет только тривиальное решение $z_h = 0$.

Пусть z_h — решение уравнения (13). Тогда согласно (11) имеем

$$\delta \| z_h \|_h^2 \leqslant (A_h z_h, z_h) = 0,$$

откуда получаем $\|z_h\|_h = 0$ и, следовательно, $z_h = 0$.

Докажем оценку (12). Согласно условию (11) для решения уравнения (2) справедливо неравенство

$$\delta \| y_h \|_h^2 \leqslant (\varphi_h, y_h)_h.$$

Применяя неравенство Коши — Буняковского, получаем, что

$$\delta \| y_h \|_h^2 \leqslant \| \varphi_h \|_h \| y_h \|_h.$$

Отсюда немедленно следует оценка (12).

Отметим связь условия (11) с оценками собственных значений оператора A_h. Если выполнено условие (11), то все собственные значения оператора A_h — действительные числа, причем для минимального собственного числа выполняется оценка

$$\lambda_{\min}^{(h)} \geqslant \delta > 0. \tag{14}$$

Действительно, пусть λ — любое собственное число оператора A_h и μ — отвечающая ему собственная функция, $A_h\mu=\lambda\mu$. Тогда согласно (11) имеем

$$(\lambda\mu, \mu)_h \geqslant \delta \|\mu\|_h^2$$

и, следовательно, $\lambda \geqslant \delta$.

Для самосопряженного оператора A_h верно и обратное: из условия (14) следует выполнение неравенства (11) при любых $v_h \in H_h$. В данном случае любой элемент $v_h \in H_h$ можно разложить по ортонормированной системе $\{\mu_k\}$ собственных векторов оператора A_h:

$$v_h = \sum_k c_k \mu_k,$$

и получить, что

$$(A_h v_h, v_h) = \sum_k \lambda_k c_k^2 \geqslant \lambda_{\min}^{(h)} \|v_h\|_h^2 \geqslant \delta \|v_h\|_h^2.$$

Таким образом, можно сформулировать еще один признак корректности.

Пусть A_h — самосопряженный оператор и $\lambda_{\min}^{(h)}$ — его минимальное собственное число. Если выполнена оценка (14) *с постоянной $\delta > 0$, не зависящей от h, то уравнение* (2) *корректно и для его решения справедлива оценка* (12).

Вернемся к примерам, рассмотренным в п. 1. Введем в пространстве H_{N-1}^0 (см. пример 1) скалярное произведение

$$(y, v) = \sum_{i=1}^{N-1} y_i v_i h$$

и норму

$$\|y\| = \left(\sum_{i=1}^{N-1} y_i^2 h \right)^{1/2}.$$

Тогда, как было показано в § 1 гл. 3, оператор (7) является самосопряженным в H_{N-1}^0 и для его минимального собственного числа справедливо неравенство (14) с константой $\delta = 9/l^2$. Таким образом, разностная задача (3) корректна и для ее решения выполняется оценка (12), где функция φ определена согласно (6).

В случае схемы из примера (2) скалярное произведение и норма в $H^0(\Omega_h)$ определяются как

$$(y, v) = \sum_{i=1}^{N_1-1} h_1 \sum_{j=1}^{N_2-1} h_2 y_{ij} v_{ij}, \qquad \|y\| = \sqrt{(y, y)}.$$

Оператор (9) является самосопряженным и для его минимального собственного числа выполнена оценка (14) с константой $\delta = 9/l_1^2 + 9/l_2^2$ (см. § 2 гл. 3). Следовательно, разностная схема (8) кор-

ректна и для ее решения справедлива оценка

$$\|y\| \leqslant (9/l_1^2 + 9/l_2^2)^{-1} \|f\|.$$

Иногда оценок вида (12), в которых решение и правая часть вычисляются в одной и той же норме, бывает недостаточно для доказательства сходимости и выяснения порядка точности разностной схемы. В то же время оценки вида (10) со специально подобранной нормой правой части $\|\varphi_h\|_{(2_h)}$ позволяют получить правильное представление о порядке точности разностной схемы.

Приведем соответствующий пример.

Пример 3. В § 3 гл. 1 изучалась разностная схема для задачи

$$(k(x)u')' - q(x)u(x) + f(x) = 0, \qquad 0 < x < l,$$
$$-k(0)u'(0) + \sigma u(0) = \mu_1, \qquad u(l) = \mu_2,$$
$$k(x) \geqslant c_1 > 0, \quad q(x) \geqslant 0, \quad \sigma \geqslant 0.$$

Было показано, что разностная схема (3), (4) из § 3 гл. 1 имеет второй порядок точности. Выпишем уравнение, которому удовлетворяет погрешность $z_i = y_i - u(x_i)$ (сетка Ω_h — та же, что и в примере 1):

$$(az_{\bar{x}})_{x,\,i} - d_i z_i = -\psi_i,$$
$$-a_1 z_{x,\,0} + \widetilde{\sigma} z_0 = \nu_1, \qquad z_N = 0. \tag{15}$$

Здесь

$$\psi_i = O(h^2), \quad \nu_1 = O(h^2), \qquad i = 1, 2, \ldots, N-1,$$
$$a_i \geqslant c_1 > 0, \quad i = 1, 2, \ldots, N, \quad \widetilde{\sigma} = \sigma + 0{,}5\, hd_0,$$
$$d_i \geqslant 0, \quad i = 0, 1, \ldots, N-1.$$

Попытаемся применить условие (11) к оценке решения задачи (15). Запишем схему (15) в операторном виде $Az = \psi$. Для того чтобы матрица оператора A была симметричной, перепишем разностное граничное условие в виде $-\frac{a_1}{h} z_{x,0} + \frac{\widetilde{\sigma}}{h} z_0 = \frac{\nu_1}{h}$. Тогда оператор A и правая часть ψ определяются следующим образом:

$$(Az)_0 = -\frac{a_1}{h} z_{x,0} + \frac{\widetilde{\sigma}}{h} z_0, \qquad z_N = 0,$$
$$(Az)_i = -(az_{\bar{x}})_{x,i} + d_i z_i, \qquad i = 1, 2, \ldots, N-1, \tag{16}$$

$$\psi = \left(\frac{\nu_1}{h}, \psi_1, \psi_2, \ldots, \psi_{N-1}\right)^T. \tag{17}$$

Введем линейное пространство $H_N^{(0)}$ функций, заданных на Ω_h и равных нулю при $i = N$, и зададим скалярное произведение и норму

$$(y, v) = \sum_{i=0}^{N-1} y_i v_i h, \qquad \|y\| = \sqrt{(y, y)}.$$

Вычислим для оператора (16) и произвольного $v \in H_N^{(0)}$ скалярное произведение (Av, v). По определению имеем

$$(Av, v) = -a_1 v_{x,0} v_0 + \widetilde{\sigma} v_0^2 - \sum_{i=1}^{N-1} h\,(av_{\bar{x}})_{x,i} v_i + \sum_{i=1}^{N-1} hd_i v_i^2.$$

Ранее было показано, что при $v_N=0$ справедливо тождество (см. (16) из § 3 гл. 1)

$$a_1 v_{x,0} v_0 + \sum_{i=1}^{N-1} h\,(a v_{\bar{x}})_{x,i} v_i = -\sum_{i=1}^{N} a_i\,(v_{\bar{x},i})^2\,h.$$

Поэтому

$$(Av,\ v) = \sum_{i=1}^{N} a_i\,(v_{\bar{x},i})^2\,h + \widetilde{\sigma} v_0^2 + \sum_{i=1}^{N-1} h d_i v_i^2 \geqslant \widetilde{\sigma} v_0^2 + \sum_{i=1}^{N} h a_i\,(v_{\bar{x},i})^2.$$

Отсюда при $\widetilde{\sigma}\geqslant 0$, $a_i\geqslant c_1>0$, $i=1, 2, \ldots, N$, получим оценку

$$(Av,\ v) \geqslant c_1 \sum_{i=1}^{N} h\,(v_{\bar{x},i})^2. \tag{18}$$

Оценим снизу правую часть неравенства (18) через среднеквадратичную норму

$$\|v\| = \left(\sum_{i=0}^{N-1} h v_i^2\right)^{1/2}. \tag{19}$$

Напомним, что согласно оценке (17) из § 3 гл. 1 при любых $v\in H_N^{(0)}$ справедливо неравенство

$$\sum_{i=1}^{N} h\,(v_{\bar{x},i})^2 \geqslant l^{-1}\,\|v\|^2_{C(\Omega_h)}, \tag{20}$$

где

$$\|v\|_{C(\Omega_h)} = \max_{0\leqslant i\leqslant N-1} |v_i|.$$

С другой стороны, для среднеквадратичной нормы (19) имеем

$$\|v\|^2 \leqslant \left(\max_{0\leqslant i\leqslant N-1} |v_i|^2\right) \sum_{i=0}^{N-1} h = l\,\|v\|^2_{C(\Omega_h)}. \tag{21}$$

Отсюда и из неравенства (20) получим

$$\sum_{i=1}^{N} h\,(v_{\bar{x},i})^2 \geqslant l^{-2}\,\|v\|^2$$

и, учитывая (18), приходим к оценке

$$(Av,\ v)\geqslant c_1 l^{-2}\|v\|^2.$$

Следовательно, для оператора (16) справедливо неравенство (11) с константой $\delta=c_1 l^{-2}$, а для разностной схемы (15) выполняется оценка (12):

$$\|z\| \leqslant c_1^{-1} l^2\,\|\psi\|. \tag{22}$$

Для сеточной функции (17), учитывая, что $\nu_1=O(h^2)$, имеем

$$\|\psi\|^2 = \sum_{i=1}^{N-1} h\psi_i^2 + \nu_1^2/h = O(h^3),$$

так что $\|\psi\|=O(h^{3/2})$ и неравенство (22) приводит к оценке $\|z\|=O(h^{3/2})$. Такое понижение порядка точности по сравнению с доказанной в § 3 гл. 1 точностью $O(h^2)$ вызвано неудачным выбором нормы правой части ψ. Если в качестве нор-

мы взять, например,

$$\|\psi\|_{(1_h)} = \sum_{i=0}^{N-1} h\,|\psi_i|, \tag{23}$$

то для функции (17) получим

$$\|\psi\|_{(1_h)} = |\nu_1| + \sum_{i=1}^{N-1} h\,|\psi_i| = O(h^2). \tag{24}$$

Однако оценка (12) не позволяет использовать норму (23). Поэтому можно поступить следующим образом. Умножая уравнение $Az=\psi$ скалярно на z, получим тождество

$$(Az,\ z) = (\psi,\ z). \tag{25}$$

Оценим правую часть этого тождества следующим образом:

$$(\psi,\ z)\,| = \left|\sum_{i=0}^{N-1} h\psi_i z_i\right| \leqslant \sum_{i=0}^{N-1} h\,|\psi_i|\,|z_i| \leqslant$$

$$\leqslant (\max_{0\leqslant i\leqslant N-1} |z_i|) \sum_{i=0}^{N-1} h\,|\psi_i| = \|z\|_{C(\Omega_h)}\,\|\psi\|_{(1_h)}.$$

Левая часть тождества (25) оценивается снизу согласно неравенствам (18), (20):

$$(Az,\ z) \geqslant c_1 l^{-1}\,\|z\|^2_{C(\Omega_h)}.$$

Таким образом, для схемы (15) справедлива оценка

$$\|z\|_{C(\Omega_h)} \leqslant c_1^{-1} l\,\|\psi\|_{(1_h)}. \tag{26}$$

Из оценки (26), учитывая (24), получаем, что $\|z\|_{C(\Omega_h)} = O(h^2)$. Кроме того, из (26) и (21) получаем, что $\|z\| = O(h^2)$.

3. Операторы первой разностной производной. На сетке

$$\Omega_h = \{x_i = ih,\ i=0,\ 1,\ \ldots,\ N,\ hN = l\}$$

рассмотрим разностное уравнение первого порядка

$$\frac{y_i - y_{i-1}}{h} = \varphi_i, \quad i = 1,\ 2,\ \ldots,\ N, \quad y_0 = \mu_1. \tag{27}$$

Введем пространство H_N функций, заданных на сетке

$$\omega_h = \{x_i = ih,\ i=1,\ 2,\ \ldots,\ N,\ hN = l\},$$

и определим в H_N скалярное произведение

$$(y,\ v) = \sum_{i=1}^{N} y_i v_i h.$$

Зададим оператор A формулами

$$(Ay)_1 = \frac{y_1}{h}, \quad (Ay)_i = \frac{y_i - y_{i-1}}{h}, \quad i = 2,\ 3,\ \ldots,\ N. \tag{28}$$

Тогда уравнение (27) можно записать в виде $Ay=\varphi$, где $\varphi=\left(\varphi_1+\frac{\mu_1}{h}, \varphi_2, \ldots, \varphi_N\right)^T$. Оператор A, определенный формулами (28), называется *оператором левой разностной производной.* Матрица этого оператора имеет вид (для определенности полагаем здесь $N=5$)

$$A=\frac{1}{h}\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix}.$$

Найдем оператор A^*, сопряженный оператору (28). По определению имеем

$$(Ay, v)=\sum_{i=1}^{N}(Ay)_i v_i h = y_1 v_1 + \sum_{i=2}^{N}(y_i - y_{i-1})v_i = \\ = \sum_{i=1}^{N} y_i v_i - \sum_{i=1}^{N-1} y_i v_{i+1} = -\sum_{i=1}^{N-1} y_i \frac{v_{i+1}-v_i}{h} h + y_N \frac{v_N}{h} h.$$

Следовательно, оператор A^* задается формулами

$$(A^*v)_i = -\frac{v_{i+1}-v_i}{h}, \quad i=1, 2, \ldots, N-1, \quad (A^*v)_N = \frac{v_N}{h}. \tag{29}$$

Оператор (29) называется *оператором правой разностной производной.* Матрица оператора (29) является транспонированной по отношению к матрице оператора (28).

Вычислим скалярное произведение (Ay, y) для оператора (28). Обозначим $y_{\bar{x},i}=(y_i-y_{i-1})/h$ и заметим, что справедливо тождество

$$y_{\bar{x},i} y_i = \frac{1}{2}(y^2)_{\bar{x},i} + \frac{h}{2}(y_{\bar{x},i})^2. \tag{30}$$

Тождество (30) доказывается непосредственной проверкой.

Из (28) и (30) получаем

$$(Ay, y) = y_1^2 + \frac{1}{2}\sum_{i=2}^{N}(y^2)_{\bar{x},i} h + \frac{h}{2}\sum_{i=2}^{N}(y_{\bar{x},i})^2 h = \\ = \frac{1}{2}(y_1^2 + y_N^2) + \frac{h}{2}\sum_{i=2}^{N}(y_{\bar{x},i})^2 h.$$

Полагая формально $y_0=0$, получим

$$(Ay, y) = \frac{h}{2}\sum_{i=1}^{N}(y_{\bar{x},i})^2 h + \frac{1}{2}y_N^2 \geqslant \frac{h}{2}\sum_{i=1}^{N} h (y_{\bar{x},i})^2, \quad y_0=0. \tag{31}$$

Из неравенства (31) следует, в частности, что оператор (28) положительный: $(Ay, y)>0$ для всех $y\in H_N$, $y\neq 0$. Действительно, $(Ay, y)\geqslant 0$ для всех $y\in H_N$. Если $(Ay, y)=0$ для некоторого $y=(y_1y_2\ldots y_N)^T$, то $y_1=y_N=0, y_{\bar{x},i}=0$, т. е. $y_i=0$, $i=1, 2, \ldots, N$.

§ 2. Канонический вид и условия устойчивости двухслойных разностных схем

1. Канонический вид двухслойных разностных схем. Общая запись разностных схем в виде операторных уравнений $A_hy_h=\varphi_h$, удобная для стационарных задач, оказывается недостаточно детальной при переходе к нестационарным разностным схемам. Поэтому при исследовании двухслойных и трехслойных разностных схем используются другие канонические формы записи.

Пусть, как и прежде, задано семейство конечномерных линейных пространств H_h, размерность которых зависит от параметра h. Параметр h считаем вектором с нормой $|h|$. В приложениях к конкретным разностным схемам пространство H_h состоит из функций, заданных на пространственной сетке Ω_h, характеризующейся шагом h.

На отрезке $[0, T]$ введем сетку по времени

$$\omega_\tau=\{t_n=n\tau,\ n=0, 1, \ldots, K,\ K\tau=T\}$$

с шагом $\tau>0$ и будем рассматривать функции $y(t_n)\in H_h$ дискретного аргумента $t_n\in\omega_\tau$ со значениями из пространства H_h. Функции $y(t_n)\in H_h$ могут зависеть параметрически от h и τ, $y(t_n)=y_{h,\tau}(t_n)$.

В дальнейшем будем обозначать $y_n=y_{h,\tau}(t_n)$.

Пусть заданы линейные операторы B_1, B_2, действующие в H_h, и функция $\varphi_n\in H_h$. *Двухслойной разностной схемой* называется семейство операторно-разностных уравнений первого порядка

$$B_1y_{n+1}+B_2y_n=\varphi_n,\quad n=0, 1, \ldots, K-1,\quad y_0\in H_h \text{ задан}. \tag{1}$$

Учитывая тождество

$$y_{n+1}=y_n+\tau\frac{y_{n+1}-y_n}{\tau}, \tag{2}$$

получаем, что любую двухслойную разностную схему можно записать на сетке ω_τ в виде

$$B\frac{y_{n+1}-y_n}{\tau}+Ay_n=\varphi_n,\quad n=0, 1, \ldots, K-1,\quad y_0\in H_h \text{ задан}, \tag{3}$$

где A и B — линейные операторы, $A=B_1+B_2$, $B=\tau B_1$.

Каноническим видом (или *канонической формой*) *двухслойной разностной схемы* называется ее запись в виде (3).

Поскольку одну и ту же разностную схему можно записать многими способами, введение единообразной канонической формы записи облегчает анализ и сравнение различных схем. По форме записи схема (3) напоминает абстрактную задачу Коши для

дифференциальных уравнений

$$\frac{du}{dt} + \mathcal{A}u(t) = f(t), \quad t > 0, \quad u(0) = u_0.$$

В случае конкретных разностных схем оператор A обычно представляет собой аппроксимацию пространственного дифференциального оператора $\mathcal{A}$, а оператор B задает ту или иную разностную схему. Поэтому запись схемы в виде (3) часто упрощает проверку аппроксимации. В дальнейшем мы убедимся в том, что условия устойчивости двуслойной разностной схемы удобно формулировать в терминах свойств операторов A и B.

Приведем несколько примеров.

Пример 1. Рассмотрим схему с весами для одномерного уравнения теплопроводности (см. § 4 гл. 1)

$$\frac{y_i^{n+1} - y_i^n}{\tau} = \sigma y_{\bar{x}x,i}^{n+1} + (1 - \sigma) y_{\bar{x}x,i}^n, \tag{4}$$

$$i = 1, 2, \ldots, N-1, \quad n = 0, 1, \ldots, K-1,$$
$$y_0^{n+1} = y_N^{n+1} = 0, \quad y_i^0 = u_0(x_i).$$

Приведем схему (4) к каноническому виду (1). В качестве пространства H_h возьмем множество $H_{N-1}^{(0)}$ действительных функций, заданных на сетке

$$\Omega_h = \{x_i = ih, i = 0, 1, \ldots, N, hN = l\}$$

и обращающихся в нуль при $i=0$, $i=N$ (операции сложения и умножения на число задаются обычным образом, т. е. покоординатно). Определим оператор A (оператор второй разностной производной) формулами

$$(Ay)_i = -y_{\bar{x}x,i}, \quad i = 1, 2, \ldots, N-1, \quad y_0 = y_N = 0. \tag{5}$$

Обозначим через $y_n \in H_{N-1}^{(0)}$ вектор $y_n = (y_1^n, y_2^n, \ldots, y_{N-1}^n)^T$, где $y_i^n = y(x_i, t_n)$. Тогда разностную схему (4) можно записать в операторном виде

$$\frac{y_{n+1} - y_n}{\tau} + \sigma A y_{n+1} + (1 - \sigma) A y_n = 0, \tag{6}$$

который еще не является ее каноническим видом. Для того чтобы перейти к каноническому виду (3), достаточно воспользоваться тождеством (2), откуда получим, что $B = E + \sigma\tau A$.

Таким образом, *разностная схема* (4) *записывается в каноническом виде* (3), *где* $\varphi_n = 0$, *оператор* A *определен согласно* (5) *и* $B = E + \sigma\tau A$.

Пример 2. На той же сетке, что и в примере 1, задана разностная схема

$$y_i^{n+1} = 0{,}5\,(y_{i+1}^n + y_{i-1}^n), \tag{7}$$
$$i = 1, 2, \ldots, N-1, \quad n = 0, 1, \ldots, K-1, \quad y_i^0 = u_0(x_i).$$

Приведем схему (7) к каноническому виду (3). Прежде всего перепишем ее в виде

$$y_i^{n+1} = 0{,}5\,(y_{i+1}^n - 2y_i^n + y_{i-1}^n) + y_i^n$$

или

$$y_i^{n+1} - y_i^n = 0{,}5h^2 y_{\overline{x}x,i}^n.$$

Поделив последнее уравнение на $0{,}5h^2$, убеждаемся в том, что схема (7) представляет собой частный случай схемы с весами (4), когда $\sigma = 1$, $\tau = 0{,}5h^2$. Следовательно, схема (7) имеет канонический вид (3), где оператор A определен согласно (5), $B = E$ и $\tau = 0{,}5h^2$.

Замечание 1. Операторы A, B в схеме (3) могут зависеть от τ, h и t_n, так что $A = A_{h,\tau}(t_n)$, $B = B_{h,\tau}(t_n)$. Функция φ_n зависит, вообще говоря, от τ и h, $\varphi_n = \varphi_{h,\tau}(t_n)$.

Замечание 2. Если сетка ω_τ неравномерная,

$$\omega_\tau = \{t_n,\ n = 0,\ 1,\ \ldots,\ K,\ t_0 = 0,\ t_K = T\}$$

с шагами $\tau_{n+1} = t_{n+1} - t_n$, $n = 0,\ 1,\ \ldots,\ K-1$, то в качестве канонической формы двуслойной схемы можно взять уравнение

$$B\frac{y_{n+1} - y_n}{\tau_{n+1}} + Ay_n = \varphi_n, \qquad n = 0,\ 1,\ \ldots,\ K-1.$$

Замечание 3. Канонический вид двуслойной разностной схемы по форме записи аналогичен одношаговому итерационному методу решения системы линейных алгебраических уравнений

$$Ay = \varphi \tag{8}$$

(см. § 1 гл. 2 ч. II). Такая аналогия не является формальной, поскольку переход от уравнения (8) к итерационному методу

$$B\frac{y_{n+1} - y_n}{\tau} + Ay_n = \varphi \tag{9}$$

можно трактовать как замену стационарного уравнения (8) нестационарным уравнением (9). Отличие итерационного метода (9) от разностной схемы (3) состоит в том, что в уравнении (9) а) A и φ не зависят от n, б) итерационный параметр τ не обязан стремиться к нулю.

2. Устойчивость разностных схем. Как и в случае стационарных задач, разностная схема (3) называется *корректной*, если ее решение $y_n = y_{h,\tau}(t_n)$: а) существует, б) единственно, в) непрерывно (причем равномерно относительно τ и h) зависит от входных данных $\varphi(t_n)$ и y_0. В дальнейшем всегда будет предполагаться, что оператор B^{-1} существует (если $B = B_{h,\tau}(t_n)$, то предполагается существование B^{-1} при всех допустимых значениях h, τ, t_n). Тем самым гарантируется существование и единственность решения задачи (3).

Дадим строгие определения устойчивости. Будем считать, что в H_h заданы две нормы: $\|\cdot\|_{1h}$, в которой измеряется решение $y(t_n) \in H_h$ и $\|\cdot\|_{2h}$, в которой измеряется правая часть $\varphi(t_n)$.

Определение 1. Разностная схема (3) называется *устойчивой*, если существуют постоянные $M_1 > 0$, $M_2 > 0$, не зависящие от h, τ, n и такие, что при любых правых частях $\varphi_{h,\tau}(t_n) \in H_h$ и любых начальных данных $y_0 \in H_h$ для решения уравнения (3) выполняется

оценка

$$\|y_n\|_{1h} \leqslant M_1 \|y_0\|_{1h} + M_2 \max_{0 \leqslant j \leqslant n-1} \|\varphi_j\|_{2h}. \tag{10}$$

Устойчивость, выраженную оценкой (10), называют устойчивостью по начальным данным и по правой части. Используются также понятия устойчивости по начальным данным и устойчивости по правой части. Рассмотрим однородное уравнение

$$B \frac{y_{n+1} - y_n}{\tau} + Ay_n = 0, \quad n = 0, 1, \ldots, K-1, \quad y_0 \in H_h \text{ задан.} \tag{11}$$

Определение 2. Разностная схема (3) называется *устойчивой по начальным данным*, если существует постоянная $M_1 > 0$, не зависящая от h, τ, n и такая, что при любых $y_0 \in H_h$ для решения однородного уравнения (11) справедлива оценка

$$\|y_n\|_{1h} \leqslant M_1 \|y_0\|_{1h}, \quad n = 0, 1, \ldots, K. \tag{12}$$

Рассмотрим теперь неоднородное уравнение (3) с нулевыми начальными данными

$$B \frac{y_{n+1} - y_n}{\tau} + Ay_n = \varphi_n, \quad n = 0, 1, \ldots, K-1, \quad y_0 = 0. \tag{13}$$

Определение 3. Разностная схема (3) называется *устойчивой по правой части*, если существует постоянная $M_2 > 0$, не зависящая от h, τ, n и такая, что при любых $\varphi_{h,\tau}(t_n) \in H_h$ для решения уравнения (13) справедлива оценка

$$\|y_n\|_{1h} \leqslant M_2 \max_{0 \leqslant j \leqslant n-1} \|\varphi_j\|_{2h}. \tag{14}$$

Заметим, что в силу линейности разностной схемы из одновременной устойчивости по начальным данным и по правой части следует устойчивость в смысле определения 1. Более того, покажем, что устойчивость по правой части является в определенном смысле следствием устойчивости по начальным данным.

Определение 4. Разностная схема (3) называется *равномерно устойчивой по начальным данным*, если существуют постоянная $\rho > 0$ и постоянная M_1, не зависящая от h, τ, n, такие, что при любых $n = 0, 1, \ldots, K-1$, $K > 1$, и при всех $y_n \in H_h$ для решения y_{n+1} однородного уравнения (11) справедлива оценка

$$\|y_{n+1}\|_{1h} \leqslant \rho \|y_n\|_{1h}, \tag{15}$$

причем $\rho^n \leqslant M_1$.

В теории разностных схем в качестве константы ρ выбирается обычно одна из величин $\rho = 1$, $\rho = 1 + c_0\tau$ или $\rho = e^{c_0\tau}$, где $c_0 \geqslant 0$ не зависит от h, τ, n. Если, например, $\rho = e^{c_0\tau}$, то $\rho^n = e^{c_0\tau n} = e^{c_0 t_n} \leqslant e^{c_0 T}$, т. е. $M = e^{c_0 T}$, где $T = K\tau$.

Перепишем однородное уравнение (11) в виде

$$y_{n+1} = Sy_n, \quad n = 0, 1, \ldots, K-1, \tag{16}$$

где оператор

$$S=E-\tau B^{-1}A \tag{17}$$

называется *оператором перехода схемы* (3).

Нетрудно видеть, что в силу произвольности $y_n\in H_h$ требование (15) равномерной устойчивости по начальным данным эквивалентно ограниченности нормы оператора S константой ρ:

$$\|S\|\leqslant\rho. \tag{18}$$

Отметим, что оператор S может зависеть от n.

В дальнейшем допустимыми значениями n будем называть числа $n=1, 2, \ldots, K-1$ такие, что $K\tau=T$, где $T>0$ — заданное число. Необходимо отметить, что $K\to\infty$ при $\tau\to 0$.

Теорема 1. *Пусть схема* (3) *равномерно устойчива по начальным данным в норме* $\|\cdot\|_{1h}$. *Тогда схема* (3) *устойчива и по правой части, причем для ее решения выполнена оценка* (10), *где* $\|\varphi_j\|_{2h}=\|B_j^{-1}\varphi_j\|_{1h}$ *и* $M_2=M_1T$.

Доказательство. Перепишем уравнение (3) при $n=j$ в виде

$$y_{j+1}=S_{j+1}y_j+\tau B_j^{-1}\varphi_j$$

и применим неравенство треугольника:

$$\|y_{j+1}\|_{1h}\leqslant\|S_{j+1}\|\,\|y_j\|_{1h}+\tau\|B_j^{-1}\varphi_j\|_{1h}.$$

Из требования равномерной устойчивости по начальным данным в силу оценки (18) получаем неравенство

$$\|y_{j+1}\|_{1h}\leqslant\rho\|y_j\|_{1h}+\tau\|B_j^{-1}\varphi_j\|_{1h},$$

которое приводит к оценке

$$\|y_{n+1}\|_{1h}\leqslant\rho^{n+1}\|y_0\|_{1h}+\sum_{j=0}^{n}\tau\rho^{n-j}\|B_j^{-1}\varphi_j\|_{1h}. \tag{19}$$

Согласно условию равномерной устойчивости по начальным данным имеем, что $\rho^n\leqslant M_1$ для всех допустимых n, в частности $\rho^{n+1}\leqslant M_1$, $\rho^{n-j}\leqslant M_1$. Поэтому из оценки (19) получим

$$\|y_{n+1}\|_{1h}\leqslant M_1\left(\|y_0\|_{1h}+\sum_{j=0}^{n}\tau\|B_j^{-1}\varphi_j\|_{1h}\right).$$

Для завершения доказательства теоремы 1 остается заметить, что

$$\sum_{j=0}^{n}\tau\|B_j^{-1}\varphi_j\|_{1h}\leqslant t_{n+1}\max_{0\leqslant j\leqslant n}\|B_j^{-1}\varphi_j\|_{1h}\leqslant T\max_{0\leqslant j\leqslant n}\|B_j^{-1}\varphi_j\|_{1h}.$$

Имея в виду теорему 1, можно ограничиться изучением равномерной устойчивости по начальным данным. Мы будем рассматривать лишь случай, когда выполняется оценка (15) с константой $\rho=1$.

3. Теоремы об устойчивости по начальным данным. Предположим, что в H_h введены скалярное произведение $(y, v)_h$ и норма $\|y\|_h=\sqrt{(y, y)_h}$. Для упрощения записи индекс h у скалярного произведения и нормы будем в дальнейшем опускать.

Оператор D, действующий в H_h, называется положительным оператором, если $(Dy, y)>0$ для всех $y\in H_h$. Если D — самосопряженный положительный оператор, то можно ввести норму $\|y\|_D=\sqrt{(Dy, y)}$, называемую энергетической нормой, порожденной оператором D. В дальнейшем неравенство $D>0$ ($D\geqslant 0$) означает, что D — положительный (неотрицательный, т. е. $(Dy, y)\geqslant 0$ для всех $y\in H_h$) оператор.

Будем считать сейчас, что H_h — вещественное пространство. Имеет место

Теорема 2. *Пусть в схеме* (3) *оператор A является самосопряженным положительным оператором и не зависит от n. Если выполнено операторное неравенство*

$$B\geqslant 0{,}5\tau A, \tag{20}$$

то схема (3) *равномерно устойчива по начальным данным, причем для решения однородного уравнения* (11) *справедлива оценка*

$$\|y_{n+1}\|_A\leqslant\|y_n\|_A, \qquad n=0, 1, \ldots, K-1. \tag{21}$$

Доказательство. Обозначим $y_t=(y_{n+1}-y_n)/\tau$, $y=y_n$ и умножим уравнение (11) скалярно на y_t. Тогда получим тождество

$$(By_t, y_t)+(Ay, y_t)=0,$$

которое после очевидных преобразований можно записать в виде

$$((B-0{,}5\tau A)y_t, y_t)+(0{,}5\tau Ay_t+Ay, y_t)=0. \tag{22}$$

Замечая, что

$$0{,}5\tau Ay_t+Ay=0{,}5A(y_n+y_{n+1}),$$

перепишем (22) в виде

$$((B-0{,}5\tau A)y_t, y_t)+0{,}5\tau^{-1}(A(y_{n+1}+y_n), y_{n+1}-y_n)=0. \tag{23}$$

Далее, используя условие самосопряженности оператора A, а также его независимость от n и положительность, получим

$$\begin{aligned}(A(y_{n+1}+y_n), y_{n+1}-y_n)&=(Ay_{n+1}, y_{n+1})-(Ay_{n+1}, y_n)+\\&+(Ay_n, y_{n+1})-(Ay_n, y_n)=(Ay_{n+1}, y_{n+1})-(Ay_n, y_n)=\\&=\tau(Ay_n, y_n)_t=\tau(\|y_n\|_A^2)_t.\end{aligned}$$

Отсюда и из (23) приходим к следующему тождеству:

$$((B-0{,}5\tau A)y_t, y_t)+\tau(\|y_n\|_A^2)_t=0. \tag{24}$$

Из условия (20) получаем, что

$$((B-0{,}5\tau A)y_t, y_t)\geqslant 0,$$

поэтому в силу (24) для решения уравнения (11) справедливо

неравенство

$$\tau(\|y_n\|^2_A)_t \leqslant 0,$$

которое совпадает с неравенством (21). Теорема 2 доказана.

Замечание 1. В теореме 2 оператор B может быть несамосопряженным оператором и может как угодно зависеть от n.

Замечание 2. Если A — самосопряженный положительный оператор, то условие (20) является и необходимым условием выполнения оценки (21). Действительно, из (21) и (24) получаем неравенство

$$((B-0{,}5\tau A)y_t, y_t) \geqslant 0. \tag{25}$$

Согласно (11) имеем $y_t = -B^{-1}Ay_n$, следовательно, в силу произвольности y_n и обратимости оператора A неравенство (25) эквивалентно операторному неравенству (20).

Если H_h — комплексное пространство, то справедлива

Теорема 3. *При тех же условиях на оператор A, что и в теореме 2, из неравенства*

$$B^* + B \geqslant \tau A \tag{26}$$

следует оценка (21) для решения уравнения (11).

Доказательство, совершенно аналогичное доказательству теоремы 2, предлагаем провести читателю.

Теоремы 2 и 3 позволяют сформулировать следующее правило исследования устойчивости конкретных двуслойных разностных схем. Прежде всего надо привести разностную схему к каноническому виду (3) и определить тем самым операторы A и B. Затем надо исследовать свойства оператора A. Если этот оператор является самосопряженным и положительным и не зависит от n, то остается проверить выполнение операторного неравенства (20) (в случае комплексного пространства — неравенства (26)). Обычно неравенство (20) приводит к некоторым ограничениям на τ и h, которые и представляют собой условия устойчивости данной разностной схемы.

Приведем примеры исследования устойчивости на основе теоремы 2.

Пример 3. Рассмотрим ту же схему с весами для уравнения теплопроводности, что и в примере 1. Эта схема была приведена к каноническому виду (11), где оператор A определен согласно (5) и $B = E + \sigma\tau A$. В главе 3 показано, что A — самосопряженный и положительный оператор в смысле скалярного произведения

$$(y, v) = \sum_{i=1}^{N-1} y_i v_i h.$$

Для скалярного произведения (Ay, y) справедливо выражение

$$(Ay, y) = \sum_{i=1}^{N} (y_{\overline{x},i})^2 h.$$

Таким образом, оператор A удовлетворяет условиям теоремы 2. Условие устойчивости (20) принимает вид

$$E + \sigma\tau A \geqslant 0{,}5\tau A$$

и означает, что при любом $y \in H_h$ должно выполняться неравенство

$$(\sigma - 0{,}5)\tau(Ay, y) + \|y\|^2 \geqslant 0. \tag{27}$$

Вспоминая неравенство (см. § 1 гл. 3)

$$(Ay, y) \leqslant \lambda_{N-1}\|y\|^2, \qquad \lambda_{N-1} = \frac{4}{h^2}\cos^2\frac{\pi h}{2l},$$

видим, что схема (4) устойчива при условии

$$(\sigma - 0{,}5)\tau + \frac{1}{\lambda_{N-1}} \geqslant 0$$

или

$$\sigma \geqslant \frac{1}{2} - \frac{1}{\tau\lambda_{N-1}}. \tag{28}$$

Достаточным условием устойчивости является неравенство

$$\sigma \geqslant \frac{1}{2} - \frac{h^2}{4\tau}. \tag{29}$$

Напомним, что те же самые условия (28) и (29) были получены ранее методом разделения переменных (см. § 3 гл. 3).

Пример 4. Рассмотрим уравнение теплопроводности в цилиндрических координатах

$$\frac{\partial u}{\partial t} = \frac{1}{x}\frac{\partial}{\partial x}\left(x\frac{\partial u}{\partial x}\right), \qquad 0 < x < l,$$

$$u(x, 0) = u_0(x), \qquad \frac{\partial u(0, t)}{\partial x} = 0, \qquad u(l, t) = 0. \tag{30}$$

Построим разностную схему, имеющую второй порядок аппроксимации по h и первый по τ. Основная трудность состоит в аппроксимации пространственного оператора

$$Lu = \frac{\partial^2 u}{\partial x^2} + \frac{1}{x}\frac{\partial u}{\partial x}$$

и граничного условия в точке $x=0$.

Введем равномерную сетку

$$\Omega_h = \{x_i = ih,\ i = 0, 1, \ldots, N,\ hN = l\}$$

и заменим Lu разностным выражением

$$(L_h u)_i = u_{\overline{x}x,i} + \frac{1}{x_i}u_{\overset{\circ}{x},i}, \qquad i = 1, 2, \ldots, N-1. \tag{31}$$

Ясно, что при такой замене погрешность аппроксимации является величиной $O(h^2)$. Заметим, что разностное выражение (31) можно записать в дивергентном виде

$$(L_h u)_i = \frac{1}{x_i}(a u_{\overline{x}})_{x,i},$$

где

$$a_i = 0{,}5(x_i + x_{i-1}).$$

Далее, чтобы аппроксимировать со вторым порядком граничное условие при $x=0$, воспользуемся разложением

$$u^n_{x,0} = \frac{u^n_1 - u^n_0}{h} = \frac{\partial u(0, t_n)}{\partial x} + \frac{h}{2}\frac{\partial^2 u(0, t_n)}{\partial x^2} + O(h^2). \tag{32}$$

Чтобы исключить из (32) выражение $\partial^2 u(0, t_n)/\partial x^2$, перепишем уравнение (30) в виде

$$\frac{\partial u(x, t)}{\partial t} = \frac{\partial^2 u(x, t)}{\partial x^2} + \frac{1}{x}\frac{\partial u(x, t)}{\partial x}$$

и перейдем к пределу при $x \to 0$. Применяя правило Лопиталя, получим $\lim\limits_{x\to 0}\frac{u'(x)}{x} = u''(0)$, откуда следует, что $\frac{\partial u(0, t)}{\partial t} = 2\frac{\partial^2 u(0, t)}{\partial x^2}$. Отсюда и из (32) получаем

$$\frac{\partial u(0, t_n)}{\partial x} = u^n_{x,0} - \frac{h}{4}\frac{\partial u(0, t_n)}{\partial t} + O(h^2) = u^n_{x,0} - \frac{h}{4}\frac{u^{n+1}_0 - u^n_0}{\tau} + O(h^2 + \tau^2).$$

Таким образом, разностное краевое условие

$$\frac{h}{4}\frac{u^{n+1}_0 - u^n_0}{\tau} = u^n_{x,0} \tag{33}$$

имеет второй порядок аппроксимации на решении задачи (30). Итак, приходим к разностной схеме

$$\frac{y^{n+1}_i - y^n_i}{\tau} = \frac{1}{x_i}(a y^n_{\bar{x}})_{x,i}, \qquad i = 1, 2, \ldots, N-1, \tag{34}$$

$$a_i = 0{,}5(x_i + x_{i-1}), \quad y^n_N = 0, \tag{35}$$

$$\frac{h}{4}\frac{y^{n+1}_0 - y^n_0}{\tau} = y^n_{x,0},$$

которая на решении уравнения (30) имеет аппроксимацию $O(\tau + h^2)$.

Приведем схему (34), (35) к каноническому виду (3). Обозначая

$$y_{t,i} = \frac{y^{n+1}_i - y^n_i}{\tau}, \qquad y^n_i = y,$$

перепишем (34), (35) в виде

$$x_i y_{t,i} = (a y_{\bar{x}})_{x,i}, \qquad a_i = 0{,}5(x_i + x_{i-1}), \tag{36}$$

$$i = 1, 2, \ldots, N-1, \qquad y^n_N = 0,$$

$$y_{t,0} = \frac{4}{h} y_{x,0}. \tag{37}$$

Рассмотрим линейное пространство $H^{(0)}_N$ функций, заданных на сетке Ω и равных нулю при $i=N$. Введем в $H^{(0)}_N$ скалярное произведение и норму

$$(y, v) = \sum_{i=0}^{N-1} y_i v_i h, \qquad \|y\| = \sqrt{(y, y)}.$$

Ясно, что во внутренних точках сетки Ω_h оператор A, соответствующий схеме (36), (37), надо определить следующим образом:

$$(Ay)_i = -(ay_{\bar{x}})_{x,i}, \qquad a_i = 0{,}5\,(x_i + x_{i-1}), \qquad i = 1, 2, \ldots, N-1.$$

Доопределим значение $(Ay)_0$ так, чтобы оператор A был самосопряженным. Если $y_N = v_N = 0$, то справедливо тождество (см. (14) из § 3 гл. 1)

$$\sum_{i=1}^{N-1} (Ay)_i v_i h = -\sum_{i=1}^{N-1} (ay_{\bar{x}})_{x,i} v_i h = \sum_{i=1}^{N} a_i y_{\bar{x},i} v_{\bar{x},i} h + a_1 y_{x,0} v_0.$$

Полагая $(Ay)_0 = -\dfrac{a_1}{h} y_{x,0}$, получим тождество

$$(Ay, v) = \sum_{i=1}^{N} a_i y_{\bar{x},i} v_{\bar{x},i} h,$$

из которого сразу следует самосопряженность оператора A. При этом

$$(Ay, y) = \sum_{i=1}^{N} a_i (y_{\bar{x},i})^2 h, \qquad a_i = 0{,}5\,(x_i + x_{i-1}). \tag{38}$$

Итак, оператор A определяется формулами

$$\begin{aligned} (Ay)_0 &= -\frac{a_1}{h} y_{x,0}, \qquad y_N = 0, \quad a_i = 0{,}5(x_i + x_{i-1}), \\ (Ay)_i &= -(ay_{\bar{x}})_{x,i}, \qquad i = 1, 2, \ldots, N-1. \end{aligned} \tag{39}$$

Заметим, что разностное граничное условие (37) можно записать в виде $\dfrac{h}{8} y_{t,0} + (Ay)_0 = 0$, так как $a_1 = 0{,}5\,h$. Таким образом, разностная схема (36), (37) имеет канонический вид (3), где оператор A определен согласно (39), а оператор B — формулами

$$(By)_0 = \frac{h}{8} y_0, \qquad (By)_i = x_i y_i, \qquad i = 1, 2, \ldots, N-1. \tag{40}$$

Докажем положительную определенность оператора (39). Из тождества (38) следует, что $(Ay, y) \geqslant 0$ для всех $y \in H_N^{(0)}$. Предположим, что $(Ay, y) = 0$ для некоторого $y \in H_N^{(0)}$, и покажем, что тогда $y = 0$. Если $(Ay, y) = 0$, то $y_{x,i} = 0$, $i = 1, 2, \ldots, N$, т. е. $y_0 = y_1 = \ldots = y_N$. Но в силу граничного условия $y_N = 0$, и, следовательно, $y_i = 0$ для всех i. Таким образом, A — самосопряженный положительный оператор и можно применять теорему 2. Условие устойчивости (20) с учетом (38) и (40) приводит к неравенству

$$\frac{h}{8} y_0^2 + \sum_{i=1}^{N-1} x_i y_i^2 h \geqslant 0{,}5\tau \sum_{i=1}^{N} a_i (y_{\bar{x},i})^2 h, \tag{41}$$

которое должно выполняться при каждом $y \in H_N^{(0)}$.

Найдем, при каких соотношениях на шаги τ и h выполняется условие (41). Для этого оценим сверху величину (Ay, y), данную выражением (38). Используя

неравенство $(a+b)^2 \leqslant 2(a^2+b^2)$, получим при $y_N = 0$, что

$$(Ay, y) = \frac{1}{h^2} \sum_{i=1}^{N} a_i (y_i - y_{i-1})^2 h \leqslant \frac{2}{h^2} \left(\sum_{i=1}^{N} a_i y_i^2 h + \sum_{i=1}^{N} a_i y_{i-1}^2 h \right) =$$

$$= \frac{2}{h^2} \left(\sum_{i=1}^{N} (a_i + a_{i+1}) y_i^2 h + a_1 y_0^2 h \right) = \frac{2}{h^2} \sum_{i=1}^{N-1} (a_i + a_{i+1}) y_i^2 h + \frac{2a_1}{h} y_0^2.$$

Отсюда при $a_i = 0{,}5(x_i + x_{i-1})$ имеем $a_i + a_{i+1} = 2x_i$ и, следовательно,

$$\sum_{i=1}^{N} a_i (y_{\overline{x},i})^2 h \leqslant \frac{4}{h^2} \sum_{i=1}^{N-1} x_i y_i^2 h + y_0^2. \tag{42}$$

Неравенство (41) будет выполнено, если потребовать

$$\frac{h^2}{8} y_0^2 + \sum_{i=1}^{N-1} x_i y_i^2 h \geqslant 0{,}5\tau \left(\frac{4}{h^2} \sum_{i=1}^{N-1} x_i y_i^2 h + y_0^2 \right),$$

т. е.

$$\frac{h^2}{8} \left(1 - \frac{4\tau}{h^2}\right) y_0^2 + \left(1 - \frac{\tau}{2h^2}\right) \sum_{i=1}^{N-1} x_i y_i^2 h \geqslant 0.$$

Следовательно, схема (36), (37) устойчива при условии $\tau \leqslant h^2/4$, причем устойчивость имеет место в норме

$$\|y\|_A = \left(\sum_{i=1}^{N} 0{,}5 (x_i + x_{i-1}) (y_{\overline{x},i})^2 h \right)^{1/2}.$$

Отметим, что ухудшение условия устойчивости по сравнению с обычным условием устойчивости явной схемы $\tau \leqslant 0{,}5\ h^2$ произошло лишь за счет разностного граничного условия (37).

4. Несамосопряженные разностные схемы. Рассмотрим двуслойную схему с весами

$$\frac{y_{n+1} - y_n}{\tau} + \sigma A y_{n+1} + (1 - \sigma) A y_n = 0, \tag{43}$$

где A — оператор, действующий в вещественном конечномерном пространстве H_h со скалярным произведением $(\ ,\ )$, а σ — числовой параметр. Схема (43) имеет канонический вид (3), где $B = E + \sigma\tau A$. Как и всегда, предполагаем существование B^{-1}. В отличие от теорем 2 и 3 не будем требовать самосопряженности оператора A. Справедлива

Теорема 4. *Если при любых $v \in H_h$ выполнено неравенство*

$$(\sigma - 0{,}5)\tau \|Av\|^2 + (Av, v) \geqslant 0, \tag{44}$$

то схема (43) *равномерно устойчива по начальным данным и для ее решения справедлива оценка*

$$\|y_{n+1}\| \leqslant \|y_n\|, \quad n = 0, 1, \ldots, K-1, \tag{45}$$

где $\|y_n\| = \sqrt{(y_n, y_n)}$.

Доказательство. Запишем схему (43) в виде

$$y_{n+1}=Sy_n,$$

где $S=E-\tau B^{-1}A$, $B=E+\sigma\tau A$. Оценка (45) эквивалентна тому, что

$$\|Sy_n\|\leqslant\|y_n\|. \tag{46}$$

В силу тождества

$$\|Sy_n\|^2=(y_n-\tau B^{-1}Ay_n,\ y_n-\tau B^{-1}Ay_n)=$$
$$=\|y_n\|^2-2\tau(B^{-1}Ay_n,\ y_n)+\tau^2\|B^{-1}Ay_n\|^2$$

заключаем, что неравенство (46) выполнено тогда и только тогда, когда

$$(B^{-1}Ay_n,\ y_n)\geqslant 0{,}5\tau\|B^{-1}Ay_n\|^2.$$

Учитывая перестановочность операторов A и B, последнее неравенство можно переписать в виде

$$(AB^{-1}y_n,\ y_n)\geqslant 0{,}5\tau\|AB^{-1}y_n\|^2. \tag{47}$$

Обозначим $v=B^{-1}y_n$. Тогда (47) примет вид

$$(Av,\ Bv)\geqslant 0{,}5\tau\|Av\|^2 \text{ или } (Av,\ v+\sigma\tau Av)\geqslant 0{,}5\tau\|Av\|^2.$$

Но это неравенство выполняется в силу условия (44), что и доказывает теорему 4.

Замечание 1. Если A — положительный оператор, то при $\sigma\geqslant 1/2$ схема (43) устойчива при любых τ (абсолютно устойчива).

Замечание 2. Если оператор A зависит от n, то в теореме 4 надо потребовать, чтобы неравенство (44) выполнялось при всех n.

Пример 5. Рассмотрим разностные схемы для уравнения первого порядка

$$\frac{\partial u}{\partial t}+\frac{\partial u}{\partial x}=0,\quad 0<t\leqslant T,\quad 0<x\leqslant l, \tag{48}$$

$$u(0,\ t)=0,\quad u(x,\ 0)=u_0(x).$$

Введем сетку $\omega_{h\tau}=\omega_h\times\omega_\tau$, где

$$\omega_h=\{x_i=ih,\quad i=0,\ 1,\ \ldots,\ N,\quad hN=l\},$$
$$\omega_\tau=\{t_n=n\tau,\quad n=0,\ 1,\ \ldots,\ K,\ K\tau=T\},$$

и аппроксимируем задачу (48) разностной схемой

$$\frac{y_i^{n+1}-y_i^n}{\tau}+\sigma y_{\overline{x},i}^{n+1}+(1-\sigma)y_{\overline{x},i}^n=0,\quad i=1,\ 2,\ \ldots,\ N, \tag{49}$$

$$y_0^n=0,\quad n=0,\ 1,\ \ldots,\ K,\quad y_i^0=u_0(x_i),\quad i=0,\ 1,\ \ldots,\ N,$$

где σ — числовой параметр и

$$y_{\overline{x},i}^n=\frac{y_i^n-y_{i-1}^n}{h}.$$

Введем пространство $H_N^{(0)}$ функций, заданных на сетке ω_h и равных нулю при $i=0$. Определим в $H_N^{(0)}$ скалярное произведение и норму

$$(y, v) = \sum_{i=1}^{N} y_i v_i h, \quad \|y\| = \sqrt{(y, y)}$$

и зададим оператор A формулами

$$(Ay)_i = y_{\bar{x},i}, \quad i = 1, 2, \ldots, N, \quad y_0 = 0. \tag{50}$$

Тогда разностную схему (49) можно записать в виде (43), где $y \in H_N^{(0)}$. Оператор (50) (оператор левой разностной производной) изучался в п. 3 § 1. Было показано, что A — несамосопряженный оператор, для которого при любом $y \in H_N^{(0)}$ выполняется тождество

$$(Ay, y) = \frac{h}{2} \sum_{i=1}^{N} (y_{\bar{x},i})^2 h + \frac{1}{2} y_N^2. \tag{51}$$

Применим теорему 4 к исследованию устойчивости схемы (49). Из определения оператора (50) имеем

$$\|Ay\|^2 = \sum_{i=1}^{N} (y_{\bar{x},i})^2 h,$$

поэтому тождество (51) можно записать в виде

$$(Ay, y) = 0{,}5h \|Ay\|^2 + 0{,}5y_N^2.$$

Тогда условие устойчивости (44) приводит к неравенству

$$(\sigma - 0{,}5)\tau \|Av\|^2 + 0{,}5h \|Av\|^2 + 0{,}5v_N^2 \geqslant 0,$$

которое выполняется при всех $v \in H_N^{(0)}$ тогда и только тогда, когда

$$\sigma \geqslant \frac{1}{2}\left(1 - \frac{h}{\tau}\right). \tag{52}$$

Неравенство (52) и представляет собой условие устойчивости схемы (49). В частности, явная схема ($\sigma=0$) устойчива при условии $\tau \leqslant h$, неявные схемы с $\sigma \geqslant 0{,}5$ устойчивы при любых шагах τ и h.

В семейство схем (49) входит и схема

$$y_{t,i-1}^n + y_{t,i}^n + y_{\bar{x},i}^{n+1} + y_{\bar{x},i}^n = 0, \tag{53}$$

имеющая второй порядок аппроксимации по τ и по h. Действительно,

$$y_{t,i-1}^n = y_{t,i}^n - h y_{t\bar{x},i}^n = y_{t,i}^n - \frac{h}{\tau}(y_{\bar{x},i}^{n+1} - y_{\bar{x},i}^n),$$

поэтому (53) можно переписать в виде

$$y_{t,i}^n + \frac{1}{2}\left(1 - \frac{h}{\tau}\right) y_{\bar{x},i}^{n+1} + \frac{1}{2}\left(1 + \frac{h}{\tau}\right) y_{\bar{x},i}^n = 0,$$

т. е. получаем схему (49) с

$$\sigma = \frac{1}{2}\left(1 - \frac{h}{\tau}\right).$$

При этом значении σ условие устойчивости (52) выполняется для всех τ и h, т. е. схема (53) абсолютно устойчива.

§ 3. Канонический вид и условия устойчивости трехслойных разностных схем

1. Канонический вид. Двуслойная разностная схема определялась в § 2 как разностное уравнение первого порядка по переменному n, коэффициенты которого являются операторами, действующими в линейном пространстве H_h. Естественно определить трехслойную схему как операторно-разностное уравнение второго порядка.

Пусть заданы: сетка

$$\omega_\tau = \{t_n = n\tau,\ n = 0, 1, \ldots, K,\ \tau > 0,\ K\tau = T\},$$

семейство конечномерных линейных пространств $\{H_h\}$, линейные операторы B_0, B_1, B_2, действующие в H_h, и функция $\varphi_n = \varphi(t_n) \in H_h$. *Трехслойной разностной схемой* называется семейство операторно-разностных уравнений второго порядка

$$B_2 y_{n+1} + B_1 y_n + B_0 y_{n-1} = \varphi_n, \qquad n = 1, 2, \ldots, K-1. \tag{1}$$

Операторы B_i, $i = 0, 1, 2$, могут зависеть от h, τ, n, а функция $\varphi_n = \varphi(t_n)$ может зависеть также от τ и h. Мы будем изучать задачу Коши для уравнения (1), состоящую в том, что по заданным элементам y_0, $y_1 \in H_h$ требуется определить решение $y_n = y(t_n)$ для всех последующих значений n, т. е. для $n = 2, 3, \ldots, K$. Чтобы задача Коши была однозначно разрешима, достаточно потребовать существования оператора B_2^{-1}. В дальнейшем мы будем предполагать всегда, что это требование выполнено.

Введем каноническую форму записи трехслойных разностных схем. Определим на сетке ω_τ разностные отношения

$$y_t = \frac{y_{n+1} - y_n}{\tau}, \quad y_{\bar{t}} = \frac{y_n - y_{n-1}}{\tau},$$

$$y_{\mathring{t}} = \frac{1}{2}(y_t + y_{\bar{t}}) = \frac{y_{n+1} - y_{n-1}}{2\tau},$$

$$y_{\bar{t}t} = \frac{y_t - y_{\bar{t}}}{\tau} = \frac{y_{n+1} - 2y_n + y_{n-1}}{\tau^2}$$

и обозначим $y = y_n$. Непосредственной проверкой можно установить справедливость следующих тождеств:

$$\begin{aligned} y_{n-1} &= y - \tau y_{\mathring{t}} + 0{,}5\tau^2 y_{\bar{t}t}, \\ y_{n+1} &= y + \tau y_{\mathring{t}} + 0{,}5\tau^2 y_{\bar{t}t}. \end{aligned} \tag{2}$$

Подставляя эти тождества в уравнение (1) и используя линейность операторов B_i, $i=0, 1, 2$, приходим к уравнению

$$\tau(B_2 - B_0) y_{\overset{\circ}{t}} + 0{,}5\tau^2 (B_2 + B_0) y_{\bar{t}t} + (B_2 + B_1 + B_0) y = \varphi,$$

где $\varphi = \varphi_n$. Обозначим

$$B = \tau(B_2 - B_0), \quad R = 0{,}5(B_2 + B_0), \quad A = B_2 + B_1 + B_0.$$

Тогда уравнение (1) можно записать в виде

$$By_{\overset{\circ}{t}} + \tau^2 R y_{\bar{t}t} + Ay = \varphi. \tag{3}$$

Запись трехслойной схемы в виде (3) называется *каноническим видом трехслойной разностной схемы.* Из предыдущих рассуждений следует, что в каноническом виде (3) можно записать любую трехслойную разностную схему, если операторы B_0, B_1, B_2 линейные, а сетка ω_τ равномерная.

Заметим, что существование B_2^{-1} эквивалентно существованию оператора, обратного оператору $B + 2\tau R$.

2. Эквивалентность трехслойной схемы двуслойной. Покажем, что трехслойную схему всегда можно представить в виде некоторой эквивалентной ей двуслойной схемы. Такое сведение трехслойной схемы к двуслойной аналогично принятой в теории дифференциальных уравнений замене уравнения второго порядка системой двух уравнений первого порядка.

Опуская индекс h, обозначим $H = H_h$ и введем пространство $H^2 = H \oplus H$ — прямую сумму двух экземпляров пространства H. Пространство H^2 определяется как множество векторов вида $y = \{y_1, y_2\}$, где y_1, $y_2 \in H$, а операции сложения и умножения на число вводятся покоординатно, т. е.

$$\alpha y + \beta v = \{\alpha y_1 + \beta v_1, \alpha y_2 + \beta v_2\},$$

где $y = \{y_1, y_2\}$, $v = \{v_1, v_2\}$, α, β — числа. Если в H задано скалярное произведение $(,)_H$, то полагаем

$$(y, v)_{H^2} = (y_1, v_1)_H + (y_2, v_2)_H.$$

Далее, если C_{ij} — линейные операторы, действующие в H, $i, j = 1, 2$, то матрица

$$C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$$

представляет собой оператор, действующий следующим образом: если $x = \{x_1, x_2\} \in H^2$, то $Cx = \{C_{11}x_1 + C_{12}x_2, C_{21}x_1 + C_{22}x_2\}$. Для операторных матриц справедливы те же правила сложения и умножения, что и для обычных матриц, надо только следить за порядком сомножителей.

Возвращаясь к уравнению (1), перепишем его в виде системы

$$\begin{aligned} y_n &= y_n, \\ y_{n+1} &= -B_2^{-1}B_0 y_{n-1} - B_2^{-1}B_1 y_n + B_2^{-1}\varphi_n. \end{aligned}$$

Определим вектор $y^n=\{y_{n-1}, y_n\}$, правую часть $\varphi^n=\{0, B_2^{-1}\varphi_n\}$ и оператор

$$S=\begin{bmatrix} 0 & E \\ -B_2^{-1}B_0 & -B_2^{-1}B_1 \end{bmatrix},$$

действующий в H^2. Тогда уравнение (1) можно записать в виде двухслойной разностной схемы

$$y^{n+1}=Sy^n+\varphi^n, \tag{4}$$

определенной в пространстве H^2.

Учитывая возможность сведения трехслойной схемы к двухслойной, можно многие результаты § 2 перенести на трехслойные разностные схемы. В частности, для трехслойной схемы так же, как и для двухслойной, из равномерной устойчивости по начальным данным следует устойчивость по правой части. Поэтому в дальнейшем мы ограничиваемся изучением устойчивости по начальным данным.

Трехслойную схему можно представить в виде двухслойной схемы не единственным образом. Иногда бывает удобным представить разностную схему (3) в виде двухслойной схемы, записанной в канонической форме

$$\mathscr{B}\frac{y^{n+1}-y^n}{\tau}+\mathscr{A}y^n=\varphi^n, \tag{5}$$

где $y^n\in H^2$ и $\mathscr{A}$, $\mathscr{B}$ — операторы, действующие в H^2. Чтобы получить такое представление, введем векторы

$$\varphi^n=\{\varphi_n, 0\},\ y^n=\{0{,}5(y_n+y_{n-1}),\ y_n-y_{n-1}\}, \tag{6}$$

где y_n — решение, а φ_n — правая часть уравнения (3). Далее, определим операторы

$$\mathscr{A}=\begin{bmatrix} A & 0 \\ 0 & R-\frac{1}{4}A \end{bmatrix},\quad \mathscr{B}=\begin{bmatrix} B+\frac{\tau}{2}A & \tau\left(R-\frac{1}{4}A\right) \\ -\tau\left(R-\frac{1}{4}A\right) & \frac{\tau}{2}\left(R-\frac{1}{4}A\right) \end{bmatrix}. \tag{7}$$

Тогда, как можно убедиться непосредственной проверкой, трехслойная разностная схема (3) представляется в виде двухслойной схемы (5). При этом, если записать схему (5) как систему двух уравнений, то первым будет уравнение

$$By_{\overset{\circ}{t}}+\tau^2Ry_{\bar{t}t}+Ay=\varphi,$$

а вторым — тождество $0=0$.

3. Устойчивость по начальным данным. Будем рассматривать задачу Коши

$$By_{\overset{\circ}{t}}+\tau^2Ry_{\bar{t}t}+Ay_n=0, \tag{8}$$

где $n=1, 2, \ldots, K-1$, y_0, y_1 заданы.

Предполагаем, что существует оператор $(B+2\tau R)^{-1}$, и, следовательно, уравнение (8) однозначно разрешимо относительно y_{n+1}. Будем считать сейчас, что H — конечномерное пространство со скалярным произведением $(\ ,\ )$. Справедлива следующая теорема о равномерной устойчивости схемы (3) по начальным данным.

Теорема 1. *Пусть A и R являются самосопряженными положительными операторами, не зависящими от n. Если выполнены операторные неравенства*

$$R > \frac{1}{4}A, \quad B \geqslant 0, \tag{9}$$

то при любых y_0, $y_1 \in H$ для решения разностной схемы (8) *справедливо неравенство*

$$\|y_{n+1}\|_* \leqslant \|y_n\|_*, \quad n = 1, 2, \ldots, K, \tag{10}$$

где

$$\|y_n\|_*^2 = \frac{1}{4}(A(y_n + y_{n-1}), y_n + y_{n-1}) + \left(\left(R - \frac{1}{4}A\right)(y_n - y_{n-1}), y_n - y_{n-1}\right). \tag{11}$$

Доказательство. Представим схему (8) в виде двуслойной схемы

$$\mathscr{B}\frac{y^{n+1} - y^n}{\tau} + \mathscr{A}y^n = 0, \quad n = 1, 2, \ldots, K-1, \tag{12}$$

где вектор $y^n \in H^2$ и операторы $\mathscr{A}$, $\mathscr{B}$ определены согласно (6), (7). Для схемы (12) задано начальное значение $y^1 = \{0{,}5(y_0 + y_1), y_1 - y_0\}$. Покажем, что схема (12) удовлетворяет всем условиям теоремы 2 из § 2. Из самосопряженности операторов A, R и из операторных неравенств $A > 0$, $R > \frac{1}{4}A$ следует, что оператор $\mathscr{A}$ (см. (7)) является самосопряженным и положительным оператором в пространстве H^2. Поэтому в H^2 можно определить норму $\|v\|_{\mathscr{A}}$, порожденную оператором $\mathscr{A}$. Для вектора $v = \{v_1, v_2\}$ норма $\|v\|_{\mathscr{A}}$ определяется следующим образом:

$$\|v\|_{\mathscr{A}}^2 = (Av_1, v_1) + \left(\left(R - \frac{1}{4}A\right)v_2, v_2\right).$$

Отсюда для решения задачи (5) $y^n = \{0{,}5(y_n + y_{n-1}), y_n - y_{n-1}\}$ имеем

$$\|y^n\|_{\mathscr{A}}^2 = \frac{1}{4}(A(y_n + y_{n-1}), y_n + y_{n-1}) + \left(\left(R - \frac{1}{4}A\right)(y_n - y_{n-1}), y_n - y_{n-1}\right),$$

т. е. $\|y^n\|_{\mathscr{A}}$ совпадает с нормой $\|y_n\|_*$, определенной согласно (11).

Проверим выполнение операторного неравенства

$$\mathscr{B} \geqslant 0{,}5\tau\mathscr{A}, \tag{13}$$

которое согласно теореме 2 из § 2 обеспечивает равномерную устой-

чивость схемы (12) по начальным данным. Из определения (7) операторов $\mathscr{A}$ и $\mathscr{B}$ получаем

$$\mathscr{B}-0{,}5\tau\mathscr{A}=\begin{bmatrix} B & \tau\left(R-\frac{1}{4}A\right) \\ -\tau\left(R-\frac{1}{4}A\right) & 0 \end{bmatrix}.$$

Для любого элемента $v=\{v_1, v_2\}\in H^2$ имеем

$$(\mathscr{B}-0{,}5\tau\mathscr{A})\,v=\left\{Bv_1+\tau\left(R-\frac{1}{4}A\right)v_2,\ -\tau\left(R-\frac{1}{4}A\right)v_1\right\}.$$

Обозначим $(\ ,\ )_{H^2}$, $(\ ,\ )_H$ скалярные произведения в H^2 и в H соответственно. Тогда получим

$$((\mathscr{B}-0{,}5\tau\mathscr{A})\,v,\ v)_{H^2}=(Bv_1,\ v_1)_H+$$

$$+\tau\left(\left(R-\frac{1}{4}A\right)v_2,\ v_1\right)_H-\tau\left(\left(R-\frac{1}{4}A\right)v_1,\ v_2\right)_H=(Bv_1,\ v_1)_H,$$

причем последнее равенство справедливо в силу самосопряженности оператора $R-\frac{1}{4}A$.

Таким образом, из неотрицательности в H оператора B следует выполнение условия устойчивости (13). В силу теоремы 2 из § 2 для решения задачи (12) справедлива оценка $\|y^{n+1}\|_{\mathscr{A}}\leqslant\|y^n\|_{\mathscr{A}}$, которая, как мы показали, совпадает с оценкой (10) для решения задачи (8). Теорема 1 доказана.

Замечание. Пусть H — комплексное пространство. Тогда теорема 1 останется справедливой, если условие $B\geqslant 0$ заменить условием $B^*+B\geqslant 0$.

4. Примеры. Для исследования устойчивости конкретных трехслойных разностных схем надо записать их в каноническом виде (3) и определить, при каких значениях параметров выполняются условия теоремы 1. Приведем несколько примеров исследования устойчивости.

Пример 1. Рассмотрим первую краевую задачу для уравнения колебаний струны

$$\frac{\partial^2 u}{\partial t^2}=\frac{\partial^2 u}{\partial x^2},\quad 0<x<l,\quad 0<t\leqslant T,$$

$$u(x,0)=u_0(x),\ \frac{\partial u(x,0)}{\partial t}=\bar{u}_0(x),\ 0\leqslant x\leqslant l, \tag{14}$$

$$u(0,t)=u(l,t)=0,\quad 0\leqslant t\leqslant T.$$

Введем сетку $\omega_{h,\tau}=\omega_h\times\omega_\tau$, где

$$\omega_h=\{x_i=ih,\ i=0,1,\ldots,N,\ hN=l\},$$

$$\omega_\tau=\{t_n=n\tau,\ n=0,1,\ldots,K,\ K\tau=T\},$$

и сопоставим задаче (14) двухпараметрическое семейство схем с

весами

$$y^n_{\bar{t}t,i} = \sigma_1 \Lambda y^{n+1}_i + (1 - \sigma_1 - \sigma_2)\Lambda y^n_i + \sigma_2 \Lambda y^{n-1}_i,$$
$$n = 1, 2, \ldots, K-1, \quad i = 1, 2, \ldots, N-1, \tag{15}$$
$$y^0_i = u_0(x_i), \qquad y^1_i = \widetilde{u}_0(x_i), \quad y^n_0 = y^n_N = 0.$$

Здесь σ_1, σ_2 — заданные числа,

$$y^n_{\bar{t}t,i} = \frac{y^{n+1}_i - 2y^n_i + y^{n-1}_i}{\tau^2}, \quad \Lambda y^n_i = \frac{y^n_{i+1} - 2y^n_i + y^n_{i-1}}{h^2},$$

а значения $\widetilde{u}_0(x_i)$ подобраны так, чтобы порядок погрешности аппроксимации начального условия $\partial u(x, 0)/\partial t = \overline{u}_0(x)$ совпадал с порядком погрешности аппроксимации основного уравнения (конкретное выражение для $\widetilde{u}_0(x_i)$ нам не потребуется, а способ построения $\widetilde{u}_0(x_i)$ был указан в § 5 гл. 1).

Введем пространство H_h как множество $H^{(0)}_{N-1}$ функций, заданных на сетке ω_h и равных нулю при $i=0$, $i=N$. Определим в $H^{(0)}_{N-1}$ оператор

$$(Ay)_i = -y_{\bar{x}x,i}, \quad i = 1, 2, \ldots, N-1, \quad y_0 = y_N = 0. \tag{16}$$

Тогда разностную схему (15) можно записать в виде

$$y_{\bar{t}t} + \sigma_1 A y_{n+1} + (1 - \sigma_1 - \sigma_2) A y_n + \sigma_2 A y_{n-1} = 0, \tag{17}$$

где $y_n \in H^{(0)}_{N-1}$, $y_n = (y^n_1, y^n_2, \ldots, y^n_{N-1})^T$, $y_{\bar{t}t} = (y_{n+1} - 2y_n + y_{n-1})/\tau^2$. Пользуясь тождествами (2), легко привести схему (17) к каноническому виду (3), где $\varphi = 0$, оператор A определен согласно (16) и

$$B = (\sigma_1 - \sigma_2)\tau A, \quad R = \frac{1}{\tau^2} E + \frac{\sigma_1 + \sigma_2}{2} A. \tag{18}$$

Для выяснения условий устойчивости схемы (15) воспользуемся теоремой 1. Уже неоднократно было показано (см. § 1 гл. 3), что оператор (16) является самосопряженным и положительным оператором в смысле скалярного произведения

$$(y, v) = \sum_{i=1}^{N-1} y_i v_i h,$$

причем его наибольшее собственное значение

$$\lambda_{\max} = \frac{4}{h^2} \cos^2 \frac{\pi h}{2l}$$

оценивается сверху величиной $\Delta = 4/h^2$.

Операторы B и R, определенные формулой (18), также самосопряженные. Согласно теореме 1, для устойчивости схемы (15) достаточно выполнения условий (9). Условие $B \geqslant 0$ приводит к ограничению $\sigma_1 \geqslant \sigma_2$, означающему, что вес нижнего слоя не должен превосходить веса верхнего слоя.

Первое из условий (9), а именно операторное неравенство $R>\frac{1}{4}A$, в данном случае приводит к неравенству

$$\frac{1}{\tau^2}E+\left(\frac{\sigma_1+\sigma_2}{2}-\frac{1}{4}\right)A>0,$$

означающему, что

$$\frac{1}{\tau^2}\|y\|^2+\left(\frac{\sigma_1+\sigma_2}{2}-\frac{1}{4}\right)(Ay, y)>0 \tag{19}$$

для любого отличного от нуля $y\in H^{(0)}_{N-1}$. Поскольку

$$\|y\|^2>\frac{1}{\Delta}(Ay, y), \qquad \Delta=\frac{4}{h^2}, \tag{20}$$

неравенство (19) будет выполнено, если потребовать

$$\frac{1}{\Delta\tau^2}+\frac{\sigma_1+\sigma_2}{2}-\frac{1}{4}>0.$$

Итак, схема (15) устойчива при выполнении условий

$$\sigma_1\geqslant\sigma_2, \quad \frac{\sigma_1+\sigma_2}{2}>\frac{1}{4}\left(1-\frac{1}{\gamma}\right), \quad \gamma=\frac{\tau^2}{h^2}. \tag{21}$$

Следует отметить, что эти неравенства, полученные как достаточные условия устойчивости, на самом деле очень близки к необходимым условиям устойчивости схемы (15). А именно, применяя метод гармоник (см. § 5 гл. 1), можно показать, что для устойчивости схемы (15) необходимо

$$\sigma_1\geqslant\sigma_2, \quad \frac{\sigma_1+\sigma_2}{2}\geqslant\frac{1}{4}\left(1-\frac{1}{\gamma}\right), \quad \gamma=\frac{\tau^2}{h^2}.$$

Частным случаем схемы (15) являются симметричные схемы ($\sigma_1=\sigma_2=\sigma$), которые имеют второй порядок погрешности аппроксимации на решении задачи (14). В этом случае условия устойчивости сводятся к одному неравенству

$$\sigma>\frac{1}{4}\left(1-\frac{1}{\gamma}\right), \quad \gamma=\frac{\tau^2}{h^2}. \tag{22}$$

Например, явная симметричная схема ($\sigma_1=\sigma_2=0$) устойчива при условии $\gamma<1$, т. е. $\tau<h$.

Пример 2. В § 5 гл. 1 уже рассматривалась схема для уравнения теплопроводности

$$\frac{y_i^{n+1}-y_i^{n-1}}{2\tau}=\frac{y_{i+1}^n-(y_i^{n+1}+y_i^{n-1})+y_{i-1}^n}{h^2}, \tag{23}$$

имеющая аппроксимацию $O(\tau^2+h^2)+O\left(\frac{\tau^2}{h^2}\right)$. Покажем, что эта схема абсолютно устойчива. Перепишем ее в виде

$$y_{\overset{\circ}{t}}+\frac{\tau^2}{h^2}y_{\bar{t}t}+Ay=0,$$

где оператор A определен согласно (16).

Тогда получим, что схема (23) имеет канонический вид (3), где $\varphi=0$, $B=E$ и $R=\frac{1}{h^2}E$. Условия устойчивости (9) сводятся к неравенству

$$\frac{1}{h^2}E>\frac{1}{4}A,$$

которое всегда выполнено в силу (20). Тем самым схема (23) абсолютно устойчива.

§ 4. Об экономичных методах решения многомерных нестационарных задач математической физики

1. Недостатки обычных разностных методов. Цель настоящего параграфа дать первоначальное представление о некоторых разностных методах, предназначенных специально для решения нестационарных задач математической физики с числом пространственных переменных, равных двум или трем (такие задачи называют многомерными). Прежде всего поясним необходимость применения специальных методов. В качестве примера рассмотрим двумерное уравнение теплопроводности

$$\begin{aligned}&\frac{\partial u}{\partial t}=\frac{\partial^2 u}{\partial x_1^2}+\frac{\partial^2 u}{\partial x_2^2}, \qquad x=(x_1, x_2)\in G,\\ &u(x, t)=\mu(x, t), \qquad x\in\Gamma, \quad 0<t\leqslant T,\\ &u(x, 0)=u_0(x), \qquad x\in G+\Gamma\end{aligned} \tag{1}$$

в прямоугольнике

$$G=\{0<x_1<l_1, \quad 0<x_2<l_2\}$$

с границей Г.

Введем, как обычно, сетку по времени

$$\omega_\tau=\{t_n=n\tau, \ n=0, 1, \ldots, K-1, \ K\tau=T\}$$

и пространственную сетку

$$\Omega_h=\{x_{ij}=(x_1^{(i)}, x_2^{(j)}), \ x_1^{(i)}=ih_1, \quad x_2^{(j)}=jh_2\},$$

где $i=0, 1, 2, \ldots, N_1$, $j=0, 1, 2, \ldots, N_2$, причем $h_1N_1=l_1$, $h_2N_2=l_2$.

Множество внутренних точек сетки Ω_h (когда $i=1, 2, \ldots, N_1-1$, $j=1, 2, \ldots, N_2-1$) будем обозначать через ω_h, а границу сетки Ω_h — через γ_h. Таким образом, γ_h — это множество точек сетки Ω_h, принадлежащих границе Г прямоугольника G. Будем обозначать $y_{ij}^n=y(x_{ij}, t_n)$, где $x_{ij}\in\Omega_h$, $t_n\in\omega_\tau$.

Как мы знаем (см. § 4 гл. 1), для решения уравнения теплопроводности можно применить либо явную, либо неявную разностную

схему. Рассмотрим сначала явную схему

$$\frac{y_{ij}^{n+1}-y_{ij}^{n}}{\tau}=\Lambda y_{ij}^{n},\ \text{если}\ x_{ij}\in\omega_h,\ t_n\in\omega_\tau, \tag{2}$$

$$y_{ij}^{n+1}=\mu(x_{ij},\ t_{n+1}),\ \text{если}\ x_{ij}\in\gamma_h,\quad t_n\in\omega_\tau,$$

$$y_{ij}^{0}=u_0(x_{ij}),\qquad \text{если}\ x_{ij}\in\Omega_h,\quad n=0,$$

где

$$\Lambda y_{ij}=\Lambda_1 y_{ij}+\Lambda_2 y_{ij},$$

$$\Lambda_1 y_{ij}=y_{\bar{x}_1x_1,ij}=\frac{y_{i+1,j}-2y_{ij}+y_{i-1,j}}{h_1^2}, \tag{3}$$

$$\Lambda_2 y_{ij}=y_{\bar{x}_2x_2,ij}=\frac{y_{i,j+1}-2y_{ij}+y_{i,j-1}}{h_2^2}.$$

Решение разностной схемы (2) находится по слоям с помощью явной формулы

$$y_{ij}^{n+1}=y_{ij}^{n}+\tau\Lambda y_{ij}^{n},\quad n=0,\ 1,\ \ldots,\ K-1,\quad x_{ij}\in\omega_h,$$

причем используются начальные и граничные условия, заданные согласно (2). Таким образом, преимуществом явной схемы является простота нахождения значений y_{ij}^{n+1} решения на верхнем слое. Существенным недостатком этой схемы, не позволяющим использовать ее при практических расчетах, является условная устойчивость. Найдем условие устойчивости по начальным данным схемы (2), опираясь на теорему 2 из § 2. При исследовании устойчивости будем предполагать, что граничные условия $\mu(x,\ t)$ равны нулю.

Введем пространство $H_h^{(0)}$ функций, заданных на сетке Ω_h и равных нулю на γ_h, со скалярным произведением

$$(y,\ v)=\sum_{i=1}^{N_1-1}h_1\sum_{j=1}^{N_2-1}h_2 y_{ij}v_{ij}.$$

Определим в $H_h^{(0)}$ оператор A формулами

$$(Ay)_{ij}=(A_1y)_{ij}+(A_2y)_{ij},\ \text{если}\ x_{ij}\in\omega_h,$$
$$(A_1y)_{ij}=-y_{\bar{x}_1x_1,ij},\ (A_2y)_{ij}=-y_{\bar{x}_2x_2,ij}, \tag{4}$$

$y_{ij}=0$, если $x_{ij}\in\gamma_h$.

Оператор A (пятиточечный разностный оператор Лапласа) изучался в § 2 гл. 3. Было показано, что A — самосопряженный положительный оператор, для которого при любых $y\in H_h^{(0)}$ справедливо неравенство

$$(Ay,\ y)\leqslant\Delta\|y\|^2,\quad \Delta=\frac{4}{h_1^2}+\frac{4}{h_2^2}. \tag{5}$$

Если записать схему (2) с $\mu=0$ как операторное уравнение в пространстве $H_h^{(0)}$, то оно примет вид

$$\frac{y_{n+1}-y_n}{\tau}+Ay_n=0, \tag{6}$$

где $y_n=y(t_n)\in H_h^{(0)}$. Таким образом, схема (2) имеет канонический вид

$$B\frac{y_{n+1}-y_n}{\tau}+Ay_n=0, \tag{7}$$

где A определен согласно (4) и $B=E$ — единичный оператор. Условие устойчивости (см. теорему 2 из § 2)

$$B\geqslant 0{,}5\tau A \tag{8}$$

в данном случае (при $B=E$) означает, что при любых $y\in H_h^{(0)}$ должно выполняться неравенство

$$\|y\|^2\geqslant 0{,}5\tau(Ay,\ y).$$

Отсюда, учитывая (5), получаем, что схема (2) устойчива по начальным данным при условии

$$\tau\left(\frac{1}{h_1^2}+\frac{1}{h_2^2}\right)\leqslant\frac{1}{2}. \tag{9}$$

Это условие накладывает очень жесткое ограничение на шаг по времени τ. Пусть, для определенности, $h_1=h_2=h$. Тогда неравенство (9) примет вид

$$\frac{\tau}{h^2}\leqslant\frac{1}{4}.$$

Если, например, $h=0{,}01$, то устойчивость гарантируется при $\tau\leqslant\tau_0$, где $\tau_0=0{,}25\cdot 10^{-4}$. Предположим, что надо найти решение задачи (1) при $T=1$. Тогда, пользуясь схемой (2), надо совершить не менее чем $n_0=T/\tau_0=40\,000$ шагов по времени. Разумеется, счет с таким мелким шагом неприемлем для практики. По указанной причине при решении уравнений параболического типа избегают пользоваться явными схемами. В случае уравнений гиперболического типа условия устойчивости позволяют взять шаг по времени того же порядка, что и шаг по пространству. Поэтому для гиперболических уравнений явные разностные схемы используются гораздо чаще, чем для параболических.

Рассмотрим теперь неявную схему для уравнения теплопроводности

$$\begin{aligned}
&\frac{y_{ij}^{n+1}-y_{ij}^n}{\tau}=\Lambda y_{ij}^{n+1},\ \text{если}\ x_{ij}\in\omega_h,\quad t_n\in\omega_\tau,\\
&y_{ij}^{n+1}=\mu(x_{ij},\ t_{n+1}),\ \text{если}\ x_{ij}\in\gamma_h,\ t_n\in\omega_\tau,\\
&y_{ij}^0=u_0(x_{ij}),\ \text{если}\ x_{ij}\in\Omega_h,\quad n=0.
\end{aligned} \tag{10}$$

Эта схема устойчива при любых шагах τ и h. Действительно, схему (10) с $\mu \equiv 0$ можно записать как операторное уравнение

$$\frac{y_{n+1} - y_n}{\tau} + Ay_{n+1} = 0,$$

где $y \in H_h^{(0)}$ и оператор A определен согласно (4). Таким образом, неявная схема (10) имеет канонический вид (7), где $B = E + \tau A$, причем условие устойчивости (8) всегда выполнено.

Однако недостатком неявной схемы (10) является необходимость решения на каждом временно́м слое системы уравнений

$$\begin{aligned} y_{ij} - \tau \Lambda y_{ij} &= F_{ij}^n, \quad x_{ij} \in \omega_h, \\ y_{ij} &= \mu_{ij}^{n+1}, \quad x_{ij} \in \gamma_h, \end{aligned} \tag{11}$$

где $y_{ij} = y_{ij}^{n+1}$, $F_{ij}^n = y_{ij}^n$.

Решение подобных систем уравнений представляет значительную трудность. Методы, предназначенные для решения систем линейных алгебраических уравнений общего вида (см. часть II, гл. 2), здесь непригодны из-за слишком большого размера системы. Действительно, если положить, например, $h_1 = h_2 = 0{,}01$ и $l_1 = l_2 = 1$, то число неизвестных y_{ij} в системе (11) окажется равным примерно 10 000. Положение усугубляется еще тем, что систему (11) необходимо решать многократно (на каждом временно́м слое).

Можно предложить приемлемые методы решения, учитывающие специальный вид матрицы системы (11). Один из таких методов рассматривался в § 6 гл. 3, другие прямые и итерационные методы будут изложены в гл. 5. Здесь же мы остановимся на методах решения уравнения (1), которые основаны на сведении многомерной задачи к последовательности одномерных задач. При таком сведении возникают разностные методы, сочетающие положительные стороны явной и неявной схем: абсолютную устойчивость и простоту решения. Начиная с пятидесятых годов, эти методы под различными названиями (методы переменных направлений, дробных шагов, расщепления, локально-одномерные методы) широко применялись для решения многомерных задач математической физики.

2. Пример метода переменных направлений. Рассмотрим подробно одну из разностных схем метода переменных направлений для уравнения (1), называемую *продольно-поперечной разностной схемой* или *схемой Писмена — Рэчфорда*. В этой схеме переход от слоя n к слою $n+1$ осуществляется в два этапа. На первом этапе определяют промежуточные значения $y_{ij}^{n+1/2}$ из системы уравнений

$$\frac{y_{ij}^{n+1/2} - y_{ij}^n}{0{,}5\tau} = \Lambda_1 y_{ij}^{n+1/2} + \Lambda_2 y_{ij}^n, \quad x_{ij} \in \omega_h, \tag{12}$$

а на втором этапе, пользуясь найденными значениями $y_{ij}^{n+1/2}$, находят y_{ij}^{n+1} из системы уравнений

$$\frac{y_{ij}^{n+1} - y_{ij}^{n+1/2}}{0{,}5\tau} = \Lambda_1 y_{ij}^{n+1/2} + \Lambda_2 y_{ij}^{n+1}, \quad x_{ij} \in \omega_h. \tag{13}$$

Здесь разностные отношения $\Lambda_1 v$, $\Lambda_2 v$ определены согласно (3). Уравнение (12) является неявным только по переменному x_1. Поэтому уравнения (12), (13) можно решить последовательным применением одномерных прогонок, сначала по направлению x_1, а затем по направлению x_2. Этим обстоятельством и объясняется название метода.

Остановимся подробнее на алгоритме решения уравнений (12), (13). Перепишем уравнение (12) в виде

$$0{,}5\gamma_1 y^{n+1/2}_{i-1,j} - (1-\gamma_1)\, y^{n+1/2}_{ij} + 0{,}5\gamma_1 y^{n+1/2}_{i+1,j} = -F^n_{ij}, \tag{14}$$

где $\gamma_1 = \tau/h_1^2$, $F^n_{ij} = y^n_{ij} + 0{,}5\tau\Lambda_2 y^n_{ij}$. Уравнение (14) решается при каждом фиксированном $j=1, 2, \ldots, N_2-1$ методом прогонки по переменному i (см. п. 7 § 4 ч. I). Чтобы применить прогонку, надо знать граничные значения $y^{n+1/2}_{0j}$, $y^{n+1/2}_{N_1 j}$, $j=1, 2, \ldots, N_2-1$. На постановке граничных условий для вспомогательной функции $y^{n+1/2}_{ij}$ мы остановимся ниже (см. п. 3). При каждом фиксированном j прогонка по направлению x_1 выполняется за $O(N_1)$ арифметических действий. Следовательно, нахождение всех $y^{n+1/2}_{ij}$ требует $O(N_1N_2)$ арифметических действий.

После того как все $y^{n+1/2}_{ij}$ найдены, решается уравнение (13). Переписывая это уравнение подробнее:

$$0{,}5\gamma_2 y^{n+1}_{i,j-1} - (1-\gamma_2)\, y^{n+1}_{ij} + 0{,}5\gamma_2 y^{n+1}_{i,j+1} = -\Phi^n_{ij},$$
$$\gamma_2 = \tau/h_2^2, \quad \Phi^n_{ij} = y^{n+1/2}_{ij} + 0{,}5\tau\Lambda_2 y^{n+1/2}_{ij}, \tag{15}$$

видим, что при каждом фиксированном $i=1, 2, \ldots, N_1-1$ его можно решить с помощью одномерной прогонки по переменному j. Граничные условия задаются в соответствии с задачей (1):

$$y^{n+1}_{i0} = \mu(x_{i0}, t_{n+1}), \quad y^{n+1}_{iN_2} = \mu(x_{iN_2}, t_{n+1}).$$

Нахождение всех y^{n+1}_{ij} из системы (15) требует $O(N_1N_2)$ арифметических действий.

Таким образом, при $N_1=N_2=N$ нахождение y^{n+1}_{ij} по известным значениям y^n_{ij} с помощью метода переменных направлений требует $O(N^2)$ арифметических действий.

Для сравнения отметим, что решение двумерной неявной схемы (например схемы (10)) с помощью стандартного метода Гаусса потребовало бы $O(N^6)$ действий, поскольку число неизвестных $O(N^2)$. Нахождение решения y^{n+1}_{ij} неявной схемы с помощью быстрого дискретного преобразования Фурье осуществляется за $O(N^2 \log_2 N)$ действий (см. § 6 гл. 3).

3. Абсолютная устойчивость продольно-поперечной схемы. Чтобы исследовать устойчивость продольно-поперечной схемы, исключим сначала из системы (12), (13) промежуточные значения $y^{n+1/2}_{ij}$ и получим эквивалентную разностную схему, связывающую значения неизвестных только на целых слоях n и $n+1$. Затем применим к полученной схеме теорему 2 из § 2 и убедимся в ее абсолютной устойчивости.

Вычтем уравнение (12) из уравнения (13). Тогда придем к уравнению

$$\frac{y_{ij}^{n+1} - 2y_{ij}^{n+1/2} + y_{ij}^{n}}{0{,}5\tau} = \Lambda_2 (y_{ij}^{n+1} - y_{ij}^{n}), \qquad x_{ij} \in \omega_h,$$

из которого получим

$$y_{ij}^{n+1/2} = \frac{y_{ij}^{n+1} + y_{ij}^{n}}{2} - \frac{\tau}{4} \Lambda_2 (y_{ij}^{n+1} - y_{ij}^{n}), \qquad x_{ij} \in \omega_h. \tag{16}$$

Подставляя найденное выражение для $y_{ij}^{n+1/2}$ в уравнение (13), получаем уравнение

$$\left[y_{ij}^{n+1} - 0{,}5 (y_{ij}^{n+1} + y_{ij}^{n}) + \frac{\tau}{4} \Lambda_2 (y_{ij}^{n+1} - y_{ij}^{n}) \right] / (0{,}5\tau) =$$

$$= 0{,}5\Lambda_1 (y_{ij}^{n+1} + y_{ij}^{n}) - \frac{\tau}{4} \Lambda_1 \Lambda_2 (y_{ij}^{n+1} - y_{ij}^{n}) + \Lambda_2 y_{ij}^{n+1},$$

которое после очевидных упрощений приводится к виду

$$\frac{y_{ij}^{n+1} - y_{ij}^{n}}{\tau} = \frac{1}{2} \Lambda (y_{ij}^{n+1} + y_{ij}^{n}) - \frac{\tau}{4} \Lambda_1 \Lambda_2 (y_{ij}^{n+1} - y_{ij}^{n}). \tag{17}$$

Строго говоря, указанная подстановка возможна не во всех точках сетки ω_h. Например, при $i=1$ уравнение (13) имеет вид

$$\frac{y_{1j}^{n+1} - y_{1j}^{n+1/2}}{0{,}5\tau} = \frac{y_{2j}^{n+1/2} - 2y_{1j}^{n+1/2} + y_{0j}^{n+1/2}}{h_1^2} + \Lambda_2 y_{1j}^{n+1}$$

и содержит значения $y_{0j}^{n+1/2}$, пока никак не заданные (подчеркнем, что уравнение (16) определяет $y_{ij}^{n+1/2}$ лишь при $i=1, 2, \ldots, N_1-1$). Аналогичная ситуация имеет место и при $i=N_1-1$, а именно, уравнение (16) при $i=N_1-1$ содержит значения $y_{N_1 j}^{n+1/2}$, не определенные формулой (16). Из вывода уравнения (17) видно, что оно будет справедливым при всех $x_{ij} \in \omega_h$, если доопределить граничные значения $y_{0j}^{n+1/2}$, $y_{N_1 j}^{n+1/2}$ в соответствии с формулами (16), т. е. положить

$$y_{0j}^{n+1/2} = \frac{\mu_{0,j}^{n+1} + \mu_{0j}^{n}}{2} - \frac{\tau}{4} \Lambda_2 (\mu_{0j}^{n+1} - \mu_{0j}^{n}),$$

$$y_{N_1 j}^{n+1/2} = \frac{\mu_{N_1 j}^{n+1} + \mu_{N_1 j}^{n}}{2} - \frac{\tau}{4} \Lambda_2 (\mu_{N_1 j}^{n+1} - \mu_{N_1 j}^{n}),$$

$$j = 1, 2, \ldots, N_2 - 1.$$

Эти граничные значения можно использовать при решении уравнений (14) методом прогонки.

Итак, в результате исключения промежуточных значений $y_{ij}^{n+1/2}$ пришли к разностной схеме (17). Исследуем устойчивость схемы (17) по начальным значениям, предполагая, что $\mu(x, t) \equiv 0$. В этом

случае схему (17) можно переписать в виде

$$\frac{y_{ij}^{n+1}-y_{ij}^{n}}{\tau}=\frac{1}{2}\Lambda\left(y_{ij}^{n+1}+y_{ij}^{n}\right)-\frac{\tau^2}{4}\Lambda_1\Lambda_2\frac{y_{ij}^{n+1}-y_{ij}^{n}}{\tau},$$

$$i=1,\ 2,\ \ldots,\ N_1-1,\quad j=1,\ 2,\ \ldots,\ N_2-1,\quad n=0,\ 1,\ \ldots,\ K-1,$$

$$y_{ij}^{n}=0,\qquad \text{если } x_{ij}\in\gamma_h,\quad n=0,\ 1,\ \ldots,\ K, \tag{18}$$

$$y_{ij}^{0}=u_0\left(x_1^{(i)},\ x_2^{(j)}\right),\ \text{если } x_{ij}\in\omega_h.$$

Введем пространство $H_h^{(0)}$ функций, заданных на Ω_h и равных нулю на γ_h, и определим операторы A_1, A_2, A согласно (4). Тогда схему (18) можно записать в операторном виде

$$\frac{y_{n+1}-y_n}{\tau}+\frac{\tau^2}{4}A_1A_2\frac{y_{n+1}-y_n}{\tau}+\frac{1}{2}A\left(y_n+y_{n+1}\right)=0, \tag{19}$$

$$n=0,\ 1,\ \ldots,\ K-1,\quad y_0 \text{ задан},$$

где $y_n=y(t_n)\in H_h^{(0)}$. Схема (19) имеет канонический вид (7), где

$$B=E+0{,}5\tau A+\frac{\tau^2}{4}A_1A_2=\left(E+0{,}5\tau A_1\right)\left(E+0{,}5\tau A_2\right).$$

Покажем, что оператор A_1A_2 положителен. По определению имеем

$$\left(A_1A_2y\right)_{ij}=y_{\bar{x}_2x_2\bar{x}_1x_1,ij},\qquad i=1,\ 2,\ \ldots,\ N_1-1,\quad j=1,\ 2,\ \ldots,\ N_2-1,$$

так что

$$\left(A_1A_2y,\ y\right)=\sum_{j=1}^{N_2-1}h_2\sum_{i=1}^{N_1-1}h_1y_{\bar{x}_2x_2\bar{x}_1x_1,ij}y_{ij}. \tag{20}$$

Учитывая, что $y_{0j}=y_{N_1j}=0$, преобразуем внутреннюю сумму в (20) по формуле (см. (14) из § 3 гл. 1)

$$\sum_{i=1}^{N_1-1}h_1y_{\bar{x}_2x_2\bar{x}_1x_1,ij}y_{ij}=-\sum_{i=1}^{N_1}h_1y_{\bar{x}_2x_2\bar{x}_1,ij}y_{\bar{x}_1,ij}$$

и запишем (20) в виде

$$\left(A_1A_2y,\ y\right)=-\sum_{i=1}^{N_1}h_1\sum_{j=1}^{N_2-1}h_2y_{\bar{x}_1\bar{x}_2x_2,ij}y_{\bar{x}_1,ij}.$$

Затем, снова применяя формулу (14) из § 3 гл. 1 и учитывая, что $y_{i0}=y_{iN_2}=0$, получим

$$\sum_{j=1}^{N_2-1}h_2y_{\bar{x}_1\bar{x}_2x_2,ij}y_{\bar{x}_1,ij}=-\sum_{j=1}^{N_2}h_2\left(y_{\bar{x}_1\bar{x}_2,ij}\right)^2,$$

так что окончательно будем иметь

$$\left(A_1A_2y,\ y\right)=\sum_{i=1}^{N_1}h_1\sum_{j=1}^{N_2}h_2\left(y_{\bar{x}_1\bar{x}_2,ij}\right)^2,$$

откуда и следует положительность оператора A_1A_2. При этом

$$B = E + 0{,}5\tau A + \frac{\tau^2}{4} A_1A_2 > 0{,}5\tau A,$$

и, следовательно, условие устойчивости (8) выполнено при любых τ, h_1, h_2, т. е. продольно-поперечная схема абсолютно устойчива.

4. Понятие суммарной аппроксимации. В предыдущем пункте мы исследовали устойчивость продольно-поперечной схемы путем исключения промежуточных значений $y_{ij}^{n+1/2}$ и замены исходной схемы переменных направлений эквивалентной ей неявной многомерной схемой (19). Не представляет труда доказать, что при достаточной гладкости решения $u(x, t)$ задачи (1) разностная схема (19) имеет погрешность аппроксимации $O(\tau^2+h^2)$ и сходится в сеточной L_2-норме со вторым порядком по τ и по h. Поэтому справедливо утверждение и о том, что решение y_{ij}^{n+1} продольно-поперечной схемы (12), (13) сходится к решению $u(x, t)$ задачи (1) со вторым порядком, т. е.

$$\|y_{n+1} - u(t_{n+1})\| \leqslant M_1(\tau^2 + h_1^2 + h_2^2), \qquad n = 0, 1, \ldots, K-1,$$

где

$$\|y_{n+1} - u(t_{n+1})\| = \left(\sum_{x_{ij} \in \omega_h} (y_{ij}^{n+1} - u(x_{ij}, t_{n+1}))\, h_1 h_2 \right)^{1/2}$$

и M_1— постоянная, не зависящая от τ, h_1, h_2.

Исключение промежуточных значений упрощает процедуру исследования разностной схемы, однако вносит некоторые неоправданные ограничения. Например, для эквивалентности схемы (12), (13) схеме (19) существенным является предположение о том, что область G — прямоугольник, кроме того, необходимо специальным образом задавать граничные условия для вспомогательной функции $y_{ij}^{n+1/2}$.

Оказывается, что схемы переменных направлений можно исследовать непосредственно, не исключая промежуточных значений $y_{ij}^{n+1/2}$. Для этого надо ввести понятие суммарной аппроксимации, которое мы поясним на примере схемы (12), (13). Пусть $u(x_1, x_2, t)$ — точное решение задачи (1). Представим решение разностной задачи (12), (13) в виде

$$y_{ij}^n = u_{ij}^n + z_{ij}^n, \qquad n = 0, 1, \ldots, K,$$

$$y_{ij}^{n+1/2} = u_{ij}^{n+1/2} + z_{ij}^{n+1/2}, \qquad n = 0, 1, \ldots, K-1,$$

где

$$u_{ij}^n = u(x_1^{(i)}, x_2^{(j)}, t_n), \quad u_{ij}^{n+1/2} = u(x_1^{(i)}, x_2^{(j)}, t_n + 0{,}5\tau).$$

Подставляя указанные выражения для y_{ij}^n, $y_{ij}^{n+1/2}$ в уравнения (12), (13), получим уравнения, которым удовлетворяет погрешность

метода

$$\frac{z_{ij}^{n+1/2}-z_{ij}^{n}}{0{,}5\tau}=\Lambda_1 z_{ij}^{n+1/2}+\Lambda_2 z_{ij}^{n}+\psi_{1,ij}^{n},\qquad x_{ij}\in\omega_h,$$

$$\frac{z_{ij}^{n+1}-z_{ij}^{n+1/2}}{0{,}5\tau}=\Lambda_1 z_{ij}^{n+1/2}+\Lambda_2 z_{ij}^{n+1}+\psi_{2,ij}^{n},\quad x_{ij}\in\omega_h, \tag{21}$$

где

$$\psi_{1,ij}^{n}=-\frac{u_{ij}^{n+1/2}-u_{ij}^{n}}{0{,}5\tau}+\Lambda_1 u_{ij}^{n+1/2}+\Lambda_2 u_{ij}^{n}, \tag{22}$$

$$\psi_{2,ij}^{n}=-\frac{u_{ij}^{n+1}-u_{ij}^{n+1/2}}{0{,}5\tau}+\Lambda_1 u_{ij}^{n+1/2}+\Lambda_2 u_{ij}^{n+1}. \tag{23}$$

Сеточные функции, определенные согласно (22), (23), называются погрешностями аппроксимации уравнений (12), (13) соответственно на решении исходной задачи (1). Разлагая функции, входящие в выражения для ψ_1, ψ_2, по формуле Тейлора в точке $(x_1^{(i)}, x_2^{(j)}, t_n)$, получим

$$\psi_{1,ij}^{n}=-\frac{\tau}{4}\frac{\partial^2 u(x_{ij},t_n)}{\partial t^2}+\frac{\tau}{2}L_1\frac{\partial u(x_{ij},t_n)}{\partial t}+O(\tau^2+h^2),$$

$$\psi_{2,ij}^{n}=-\frac{3}{4}\tau\frac{\partial^2 u(x_{ij},t_n)}{\partial t^2}+\frac{\tau}{2}L_1\frac{\partial u(x_{ij},t_n)}{\partial t}+\tau L_2\frac{\partial u(x_{ij},t_n)}{\partial t}+O(\tau^2+h^2),$$

где $L_\alpha u=\partial^2 u/\partial x_\alpha^2$, $\alpha=1, 2$. Таким образом, каждое из уравнений (12), (13) аппроксимирует исходное уравнение (1) с первым порядком по τ и вторым — по h. Вместе с тем сумма погрешностей аппроксимации $\psi_{ij}^{n}=\psi_{1,ij}^{n}+\psi_{2,ij}^{n}$ имеет второй порядок по τ и по h. Действительно,

$$\psi_{ij}^{n}=-\tau\frac{\partial^2 u(x_{ij},t_n)}{\partial t^2}+\tau(L_1+L_2)\frac{\partial u(x_{ij},t_n)}{\partial t}+O(\tau^2+h^2),$$

и в силу дифференциального уравнения (1) имеем

$$\frac{\partial^2 u}{\partial t^2}=(L_1+L_2)\frac{\partial u}{\partial t},$$

так что $\psi_{ij}^{n}=O(\tau^2+h^2)$.

Поэтому говорят, что схема (12), (13) обладает *суммарной аппроксимацией второго порядка по τ и по h.*

Можно получить оценку решения задачи для погрешности (21) через норму функции $\psi=\psi_1+\psi_2$, из которой будет следовать второй порядок точности схемы (12), (13) (см. [32]).

Приведем другие примеры схем переменных направлений для уравнения (1), обладающих суммарной аппроксимацией. *Локально-одномерная* схема состоит

в последовательном решении уравнений

$$\frac{y_{ij}^{n+1/2}-y_{ij}^{n}}{\tau}=\Lambda_1 y_{ij}^{n+1/2}, \qquad x_{ij}\in\omega_h,$$

$$\frac{y_{ij}^{n+1}-y_{ij}^{n+1/2}}{\tau}=\Lambda_2 y_{ij}^{n+1}, \qquad x_{ij}\in\omega_h. \tag{24}$$

В этой схеме каждое из уравнений в отдельности не аппроксимирует исходное уравнение (1), однако имеет место суммарная аппроксимация $O(\tau+h^2)$. Действительно, в данном случае

$$\psi_1=-\frac{u_{ij}^{n+1/2}-u_{ij}^{n}}{\tau}+\Lambda_1 u_{ij}^{n+1/2}=0{,}5\,\frac{\partial u(x_{ij},t_n)}{\partial t}+L_1 u(x_{ij},t_n)+O(\tau+h^2)=O(1),$$

$$\psi_2=-\frac{u_{ij}^{n+1}-u_{ij}^{n+1/2}}{\tau}+\Lambda_2 u_{ij}^{n+1}=0{,}5\,\frac{\partial u(x_{ij},t_n)}{\partial t}+L_2 u(x_{ij},t_n)+O(\tau+h^2)=O(1)$$

и $\psi_1+\psi_2=O(\tau+h^2)$.

В качестве упражнения читателю предлагается рассмотреть схему

$$\frac{y_{ij}^{n+1/2}-y_{ij}^{n}}{\tau}=\Lambda_1 y_{ij}^{n+1/2}+\Lambda_2 y_{ij}^{n},$$

$$\frac{y_{ij}^{n+1}-y_{ij}^{n+1/2}}{\tau}=\Lambda_2\left(y_{ij}^{n+1}-y_{ij}^{n}\right) \tag{25}$$

и доказать, что она обладает суммарной аппроксимацией. Кроме того, рекомендуется провести исключение промежуточных значений $y_{ij}^{n+1/2}$ из схем (24), (25), получить соответствующие многомерные разностные схемы и сформулировать граничные условия, при которых многомерные схемы эквивалентны исходным схемам переменных направлений. При нахождении граничных условий для $y_{ij}^{n+1/2}$ следует поступать так же, как и в случае схемы (12), (13), т. е. выразить $y_{ij}^{n+1/2}$ из уравнений (24) или (25) через y_{ij}^{n+1}, y_{ij}^{n} и доопределить $y_{ij}^{n+1/2}$ на границе с помощью полученного выражения.

ГЛАВА 5

ПРЯМЫЕ И ИТЕРАЦИОННЫЕ МЕТОДЫ РЕШЕНИЯ СЕТОЧНЫХ УРАВНЕНИЙ

§ 1. Модельная задача

1. Введение. Как мы уже видели, аппроксимация дифференциальных уравнений разностными приводит к системам линейных алгебраических уравнений

$$Ay=f, \tag{1}$$

которые нецелесообразно, а чаще всего и невозможно решать стандартными вычислительными методами линейной алгебры.

Если исходной задачей является краевая задача для обыкновенного дифференциального уравнения, то соответствующую разностную схему можно решить с помощью метода прогонки (см. п. 7 § 4 ч. I). В многомерном случае не существует столь же удобного и экономичного способа решения разностных уравнений, как метод прогонки. Поэтому возникает необходимость в развитии методов, специально предназначенных для решения многомерных разностных краевых задач. Мы будем рассматривать здесь лишь двумерные разностные задачи.

Как и в общем случае систем линейных уравнений, методы решения разностных задач разделяются на прямые и итерационные. Итерационные методы являются более простыми, чем прямые, и в меньшей степени используют структуру матрицы. По этой причине для решения двумерных разностных уравнений первоначально использовались исключительно итерационные методы. Однако в случае разностных задач сходимость таких, например, методов, как метод простой итерации, Зейделя, верхней релаксации, весьма медленная. В настоящее время интенсивно развиваются и прямые методы решения двумерных разностных уравнений. Они применимы, как правило, к уравнениям с разделяющимися переменными, когда область изменения независимых переменных — прямоугольник. Наконец, следует отметить все возрастающее значение неявных итерационных методов, в которых решение на новой итерации находится тем или иным прямым методом. Хотя такие методы алгоритмически сложнее, чем явные, их несомненным преимуществом является существенно более быстрая сходимость.

2. Модельная задача. Методы решения двумерных разностных краевых задач мы будем иллюстрировать в дальнейшем на следующем простом примере. Рассмотрим задачу Дирихле для уравнения Пуассона

$$\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = -f(x), \quad x = (x_1, x_2) \in G,$$

$$u(x) = 0, \quad x = (x_1, x_2) \in \Gamma$$

в единичном квадрате $G(0 < x_1, x_2 < 1)$ с границей Γ.

Введем в G квадратную сетку с шагом h, т. е. множество точек

$$\Omega_h = \{x_{ij} = (x_1^{(i)}, x_2^{(j)})\},$$

где $x_1^{(i)} = ih$, $x_2^{(j)} = jh$, $i, j = 1, 2, \ldots, N$, $hN = 1$. Пусть, как обычно, ω_h — множество внутренних точек и γ_h — множество граничных точек сетки Ω_h.

Заменим исходную дифференциальную задачу разностной задачей

$$y_{\bar{x}_1 x_1, ij} + y_{\bar{x}_2 x_2, ij} = -f_{ij}, \quad x_{ij} \in \omega_h, \tag{2}$$

$$y_{ij} = 0, \quad x_{ij} \in \gamma_h,$$

которую будем рассматривать как модельную при изучении мето-

дов решения сеточных уравнений. Подробнее задачу (2) можно записать в виде системы

$$\frac{y_{i-1,j}-2y_{ij}+y_{i+1,j}}{h^2}+\frac{y_{i,j+1}-2y_{ij}+y_{i,j-1}}{h^2}=-f_{ij}, \qquad (3)$$

$$y_{i0}=y_{iN}=0, \quad y_{0j}=y_{Nj}=0, \quad i, j=1, 2, \ldots, N-1.$$

На этом примере хорошо видны характерные особенности систем уравнений, возникающих при аппроксимации многомерных задач математической физики. Матрицы таких систем характеризуются высоким порядком, сильной разреженностью (т. е. преобладанием нулевых элементов) и большим разбросом собственных чисел.

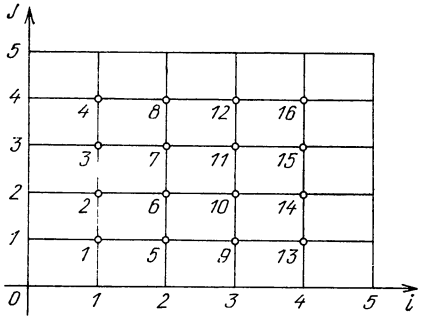


Рис. 14. Одномерная нумерация двумерного массива

Действительно, порядок системы (2) совпадает с числом точек сетки ω_h и равен $(N-1)^2$. Даже при $h=0{,}1$ имеем $(N-1)^2=81$, т. е. матрица системы (2) является квадратной матрицей 81 порядка. На более типичной сетке, когда шаг $h=0{,}01$, порядок системы равен примерно 10 000.

Сильная разреженность видна из того, что каждое уравнение системы (3) содержит не более пяти отличных от нуля коэффициентов. Тем самым отношение числа ненулевых элементов данной матрицы к общему числу ее элементов не превосходит $5/(N-1)^2=O(h^2)$.

Собственные числа матрицы, отвечающей системе (2), найдены в § 2 гл. 3. Для нас существенно сейчас, что отношение наименьшего собственного числа γ_1 к наибольшему собственному числу γ_2 равно

$$\xi=\frac{\gamma_1}{\gamma_2}=\operatorname{tg}^2\frac{\pi h}{2}.$$

Отношение ξ является величиной второго порядка малости при $h\to 0$, а именно

$$\xi=\frac{\pi^4h^2}{4}+O(h^4). \qquad (4)$$

Следствием малости величины ξ является плохая обусловленность системы (2). По этой же причине явные итерационные методы для системы (2) сходятся медленно.

Чтобы записать систему двумерных разностных уравнений в матричном виде (1), надо провести перенумерацию двумерного массива индексов $(i, j)_{i,j=1}^{N-1}$ в одномерный массив. Это можно сделать различными способами. Сопоставим, например, индексу (i, j) двумерного массива индекс k одномерного массива по правилу $k=(N-1)(i-1)+j$ (см. рис. 14). При этом, если i и j меняются в пределах от 1 до $N-1$, то k меняется от 1 до $(N-1)^2$. В результате указанной пе-

ренумерации система уравнений (3) запишется в виде

$$\frac{y_{k+1}-4y_k+y_{k-1}}{h^2}+\frac{y_{k-(N-1)}+y_{k+(N-1)}}{h^2}=-f_k. \qquad (5)$$

Уравнения (5) определены для

$$k=(N-1)(i-1)+j, \qquad i,\ j=2,\ 3,\ \ldots,\ N-2.$$

При остальных значениях k, учитывая нулевые граничные условия, получим следующие уравнения:

$$-4y_1+y_2+y_N=-h^2f_1, \quad k=1 \quad (i=1,\ j=1),$$

$$y_{k-1}-4y_k+y_{k+1}+y_{k+N-1}=-h^2f_k,$$

$$k=2,\ 3,\ \ldots,\ N-2 \quad (i=1,\ j=2,\ 3,\ \ldots,\ N-2),$$

$$y_{N-2}-4y_{N-1}+y_{2(N-1)}=-h^2f_{N-1}, \quad k=N-1 \quad (i=1,\ j=N-1),$$

$$y_{k-(N-1)}-4y_k+y_{k+1}+y_{k+(N-1)}=-h^2f_k, \quad k=(N-1)(i-1)+1$$

$$(i=2,\ 3,\ \ldots,\ N-2,\ j=1),$$

$$-4y_k+y_{k-1}+y_{k-(N-1)}+y_{k+(N-1)}=-h^2f_k, \quad k=(N-1)i \quad (i=2,\ 3,\ \ldots,\ N-2,$$

$$j=N-1),$$

$$y_{k-(N-1)}-4y_k+y_{k+1}=-h^2f_k, \quad k=(N-1)(N-2)+1 \quad (i=N-1,\ j=1),$$

$$y_{k-(N-1)}+y_{k-1}-4y_k+y_{k+1}=-h^2f_k,$$

$$k=(N-1)(N-2)+j \quad (i=N-1,\ j=2,\ 3,\ \ldots,\ N-2),$$

$$y_{k-(N-1)}+y_{k-1}-4y_k=-h^2f_k, \quad k=(N-1)^2 \quad (i=N-1,\ j=N-1).$$

Матрица системы (5) для случая $N=5$ условно изображена на рис. 15, где крестиками отмечены ненулевые элементы. Заметим, что при решении системы (3) нет необходимости записывать ее в виде (5), мы привели такую запись лишь для того, чтобы еще раз продемонстрировать разреженность матрицы и ее ленточную структуру.

3. Применение методов Якоби и Зейделя. Запишем разностное уравнение Пуассона (2) в операторной форме (1), где оператор A определен следующим образом:

$$(Ay)_{ij}=-y_{\overline{x_1x_1},ij}-y_{\overline{x_2x_2},ij}, \qquad x_{ij}\in\omega_h, \qquad (6)$$

$$y_{ij}=0, \quad x_{ij}\in\gamma_h.$$

В дальнейшем будем рассматривать для этого уравнения одношаговые итерационные методы, записанные в каноническом виде (см. § 1 гл. 2 ч. II),

$$B\frac{y_{n+1}-y_n}{\tau_{n+1}}+Ay_n=f. \qquad (7)$$

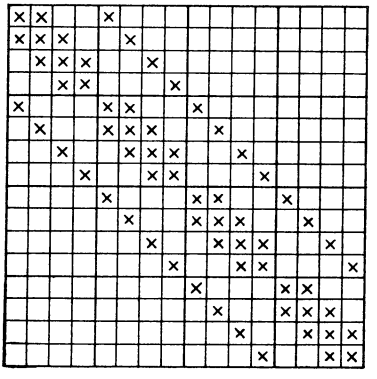


Рис. 15. Структура матрицы системы (5) для $N=5$

Начнем с наиболее простых методов — Якоби и Зейделя. Покажем, что эти методы сходятся, однако их скорость сходимости невысока.

Метод Якоби для системы (3) записывается в виде

$$y_{ij}^{n+1} = \frac{1}{4}(y_{i-1,j}^{n} + y_{i+1,j}^{n} + y_{i,j-1}^{n} + y_{i,j+1}^{n} + h^2 f_{ij}), \qquad x_{ij} \in \omega_h,$$

$$y_{ij}^{n} = 0, \qquad x_{ij} \in \gamma_h. \tag{8}$$

Здесь y_{ij}^{n} — значение решения в точке $x_{ij} \in \Omega_h$ на n-й итерации. В данном случае метод Якоби совпадает с методом простой итерации при оптимальном значении итерационного параметра. Действительно, метод простой итерации $(y_{n+1} - y_n)/\tau + Ay_n = f$ для системы (1) в случае $A^* = A > 0$ обладает наибольшей скоростью сходимости, если $\tau = \tau_0 = 2/(\delta + \Delta)$, где δ, Δ — наименьшее и наибольшее собственные числа матрицы A (см. § 6 гл. 2 ч. II).

Для разностного оператора Лапласа имеем (см. § 2 гл. 3)

$$\delta = \frac{8}{h^2}\sin^2\frac{\pi h}{2}, \quad \Delta = \frac{8}{h^2}\cos^2\frac{\pi h}{2},$$

следовательно, $\tau_0 = h^2/4$. При этом значении параметра метод простой итерации в случае модельной задачи (2) принимает вид

$$\frac{y_{ij}^{n+1} - y_{ij}^{n}}{h^2/4} - y_{\bar{x}_1 x_1, ij}^{n} - y_{\bar{x}_2 x_2, ij}^{n} = f_{ij}, \qquad x_{ij} \in \omega_h,$$

$$y_{ij}^{n} = 0, \qquad x_{ij} \in \gamma_h.$$

Последнее уравнение, как нетрудно видеть, совпадает с уравнением (8).

Скорость сходимости метода (8) как метода простой итерации с оптимальным параметром определяется числом $\rho = \frac{1-\xi}{1+\xi}$, $\xi = \frac{\delta}{\Delta} = \operatorname{tg}^2\frac{\pi h}{2}$.

Число итераций $n_0(\varepsilon)$, необходимых для достижения заданной точности ε, равно

$$n_0(\varepsilon) = \ln\frac{1}{\varepsilon} \Big/ \ln\frac{1}{\rho} = \ln\frac{1}{\varepsilon} \Big/ \ln\left(1 + \frac{2\xi}{1-\xi}\right).$$

При $h \to 0$ имеем $\xi = \frac{\pi^2 h^2}{4}$, $\ln\frac{1}{\rho} \approx 2\xi = \frac{\pi^2 h^2}{2}$, так что

$$n_0(\varepsilon) \approx \frac{2\ln(1/\varepsilon)}{\pi^2 h^2}. \tag{9}$$

Следовательно, метод Якоби требует $O(h^{-2})$ итераций для достижения заданной точности. Это очень медленная сходимость. В настоящее время применяются методы, требующие $O(h^{-1})$ и даже $O(\ln h^{-1})$ итераций для достижения той же точности. С этими методами мы познакомимся в § 2, 3, 4.

Рассмотрим метод Зейделя для системы (3). В общем случае (см. § 1 гл. 2 ч. II) метод Зейделя строится таким образом, чтобы в уравнении с номером i неизвестные, имеющие индекс больший, чем i, вычислялись бы по значениям на n-й итерации. Реализация метода Зейделя для системы (3) приводит к следующему итерационному методу:

$$\frac{y^{n+1}_{i-1,j}-2y^{n+1}_{ij}+y^{n}_{i+1,j}}{h^2}+\frac{y^{n+1}_{i,j-1}-2y^{n+1}_{ij}+y^{n}_{i,j+1}}{h^2}=-f_{ij}, \qquad x_{ij}\in\omega_h,$$

$$y^{n+1}_{ij}=0, \qquad x_{ij}\in\gamma_h. \tag{10}$$

Хотя метод Зейделя является неявным, нахождение значений y^{n+1}_{ij} на новой итерации не представляет труда, поскольку оно сводится к обращению треугольной матрицы. Здесь нужно лишь правильно установить последовательность проведения вычислений. Сначала из уравнения (10), используя известные граничные значения $y^{n+1}_{01}=0$ и $y^{n+1}_{10}=0$, находят y^{n+1}_{11}. Зная y^{n+1}_{11}, можно найти y^{n+1}_{12} и т. д. Таким образом, неизвестные y^{n+1}_{ij} вычисляются в следующем порядке изменения индексов: (1, 1), (1, 2), ..., (1, N—1), (2, 1), (2, 2), ..., (2, N—1), ..., (N—1, 1), (N—1, 2), ..., (N—1, N—1). В этом случае говорят, что вычисления ведутся от левого нижнего угла прямоугольника G к правому верхнему углу.

Метод Зейделя сходится несколько быстрее, чем метод Якоби, однако число итераций, необходимое для достижения заданной точности, здесь также является величиной порядка h^{-2}.

Чтобы доказать последнее утверждение, достаточно построить пример начальных данных, при которых погрешность итерационного метода убывает не быстрее, чем q^n, где $q=1-O(h^2)$. Приведем такой пример. Пусть $\{y_{ij}\}_{i,j=1}^{N-1}$ — решение разностной задачи (2) и $\{y^n_{ij}\}_{i,j=1}^{N-1}$ — приближенное решение, полученное на n-й итерации с помощью метода Зейделя (10). Для погрешности $z^n_{ij}=y^n_{ij}-y_{ij}$ получаем уравнение

$$z^n_{i+1,j}-4z^{n+1}_{ij}+z^{n+1}_{i-1,j}+z^n_{i,j+1}+z^{n+1}_{i,j-1}=0, \qquad x_{ij}\in\omega_h,$$

$$z_{ij}=0, \qquad x_{ij}\in\gamma_h, \tag{11}$$

$$z^0_{ij}=y^0_{ij}-y_{ij}.$$

Будем искать решение задачи (11) в виде

$$z^n_{ij}=q^n v_{ij}, \tag{12}$$

где q — число, а v_{ij} — сеточная функция, не зависящая от n и удовлетворяющая нулевым граничным условиям. Подставляя (12) в (11) и сокращая на q^n, получим систему уравнений

$$v_{i+1,j}-4qv_{ij}+qv_{i-1,j}+v_{i,j+1}+qv_{i,j-1}=0, \qquad x_{ij}\in\omega_h,$$

$$v_{ij}=0, \qquad x_{ij}\in\gamma_h. \tag{13}$$

Система уравнений (13) представляет собой задачу на собственные значения. Будем искать ее решение в виде

$$v^{(k)}_{ij}=s_k^{i+j}\mu^{(k)}_{ij}, \tag{14}$$

где $k=(k_1, k_2)$, $k_\alpha=1, 2, \ldots, N-1$, $\alpha=1, 2$, $\mu_{ij}^{(k)}$ — собственные функции пяти точечного разностного оператора Лапласа, $\mu_{ij}^{(k)}=2\sin(\pi k_1 x_1^{(i)})\sin(\pi k_2 x_2^{(j)})$ и s_k — числа, подлежащие определению.

Подставляя (14) в (13), получим уравнение

$$s_k^2(\mu_{i+1,j}^{(k)}+\mu_{i,j+1}^{(k)})+q(\mu_{i,j-1}^{(k)}+\mu_{i-1,j}^{(k)}-4s_k\mu_{ij}^{(k)})=0.$$

Полагая $q=q_k=s_k^2$, приходим к задаче на собственные значения

$$\mu_{\bar{x}_1x_1,ij}+\mu_{\bar{x}_2x_2,ij}+\lambda\mu_{ij}=0, \qquad x_{ij}\in\omega_h,$$
$$\mu_{ij}=0, \qquad x_{ij}\in\gamma_h,$$

где $\lambda=\lambda_k=4(1-s_k)/h^2$. Выражения для λ_k известны (см. § 2, гл. 3):

$$\lambda_k=\lambda_{k_1k_2}=\frac{4}{h^2}\left(\sin^2\frac{\pi k_1 h}{2}+\sin^2\frac{\pi k_2 h}{2}\right).$$

Выбирая $k=(1, 1)$, получим

$$\lambda=\lambda_1=\delta=\frac{8}{h^2}\sin^2\frac{\pi h}{2}.$$

При этом значении k величина $|q_k|=s_k^2$ достигает максимума, равного

$$|q|=\left(1-\frac{h^2\delta}{4}\right)^2. \tag{15}$$

Таким образом, если положим

$$y_{ij}^0=y_{ij}+v_{ij}^{(1,1)}, \tag{16}$$

где y_{ij} — точное решение задачи (3) и

$$v_{ij}^{(1,1)}=s_{11}^{i+j}\mu_{ij}^{(1,1)}, \quad s_{11}=1-\frac{h^2\delta}{4}=1-2\sin^2\frac{\pi h}{2},$$

то согласно (12) получим $z_{ij}^n=q^n z_{ij}^0$, где выражение для q определено согласно (15). Следовательно, для начальных данных (16) при любой норме выполняется равенство

$$\|y^n-y\|=|q|^n\|y_0-y\|.$$

При малых шагах h имеем

$$q\approx 1-0{,}5\ h^2\delta\approx 1-\pi^2h^2,$$

т е. необходимое число итераций возрастает как h^{-2}.

4. Метод верхней релаксации. Рассмотрим применение метода верхней релаксации к модельной задаче (3). Для общей системы уравнений (1) метод верхней релаксации определяется следующим образом (см. § 1 гл. 2 ч. II). Система (1) представляется в виде

$$(A_-+A_++D)y=f, \tag{17}$$

где A_-, A_+ — нижняя треугольная и верхняя треугольная матрицы с нулевыми диагоналями, D — диагональная матрица. Вводится итерационный параметр $\omega\in(0, 2)$ и итерации определяются фор-

мулами

$$A_-y_{n+1}+A_+y_n+D\left(\frac{1}{\omega}y_{n+1}+\left(1-\frac{1}{\omega}\right)y_n\right)=f. \tag{18}$$

При $\omega=1$ метод (18) совпадает с методом Зейделя

$$(A_-+D)y_{n+1}+A_+y_n=f.$$

В случае модельной задачи (3) представлению (17) соответствует запись системы в виде

$$\frac{y_{i-1,j}+y_{i,j-1}}{h^2}+\frac{y_{i,j+1}+y_{i+1,j}}{h^2}-\frac{4y_{ij}}{h^2}=-f_{ij}, \quad x_{ij}\in\omega_h,$$

$$y_{ij}=0, \quad x_{ij}\in\gamma_h. \tag{19}$$

Метод верхней релаксации определяется уравнениями

$$\frac{y^{n+1}_{i-1,j}+y^{n+1}_{i,j-1}}{h^2}+\frac{y^{n}_{i,j+1}+y^{n}_{i+1,j}}{h^2}-\frac{4}{h^2}\left(\frac{1}{\omega}y^{n+1}_{ij}+\left(1-\frac{1}{\omega}\right)y^{n}_{ij}\right)=-f_{ij},$$

$$x_{ij}\in\omega_h,$$

$$y^n_{ij}=0, \quad x_{ij}\in\gamma_h. \tag{20}$$

Способ нахождения значений y^{n+1}_{ij} на новой итерации тот же, что и в методе Зейделя. А именно, надо записать уравнения (20) в виде

$$y^{n+1}_{i-1,j}+y^{n+1}_{i,j-1}-\frac{4}{\omega}y^{n+1}_{ij}=F^{(n)}_{ij}, \tag{21}$$

где

$$F^{(n)}_{ij}=-(y^n_{i,j+1}+y^n_{i+1,j})+4\left(1-\frac{1}{\omega}\right)y^n_{ij}-h^2f_{ij},$$

и вычислять y^{n+1}_{ij}, начиная от левого нижнего угла области G.

Проведем исследование сходимости метода верхней релаксации (20) и покажем, в частности, что при оптимальном выборе параметра ω число итераций, необходимых для получения заданной точности ε, является величиной $O(h^{-1})$ (а не $O(h^{-2})$, как в методе Зейделя). Основное преимущество метода верхней релаксации перед методом Зейделя как раз и состоит в существенном увеличении скорости сходимости при надлежащем выборе итерационного параметра ω.

Для погрешности $z^n_{ij}=y^n_{ij}-y_{ij}$ метода (20) получаем уравнение

$$\frac{z^{n+1}_{i-1,j}+z^{n+1}_{ij-1}}{h^2}+\frac{z^{n}_{ij+1}+z^{n}_{i+1,j}}{h^2}-\frac{4}{h^2}\left(\frac{1}{\omega}z^{n+1}_{ij}+\left(1-\frac{1}{\omega}\right)z^{n}_{ij}\right)=0, \quad x_{ij}\in\omega_h,$$

$$z^n_{ij}=0, \quad x_{ij}\in\gamma_h,$$

которое приводится к каноническому виду

$$B\frac{z_{n+1}-z_n}{\tau}+Az_n=0,$$

где $z_n=\{z_{ij}^n\}_{i,j=1}^{N-1}$, оператор A определен согласно (6) и

$$B=\frac{2(2-\omega)}{h^2}E+\omega(R_1+R_2),\qquad \tau=\omega, \tag{22}$$

$$(R_1z)_{ij}=\frac{z_{\bar{x}_1,ij}}{h},\quad (R_2z)_{ij}=\frac{z_{\bar{x}_2,ij}}{h}.$$

Операторы A, B, R_1, R_2 определены в пространстве $H_h^{(0)}$ сеточных функций, заданных на Ω_h и равных нулю на γ_h. В $H_h^{(0)}$ введены скалярное произведение

$$(y, v)=\sum_{i,j=1}^{N-1}y_{ij}v_{ij}h^2$$

и норма $\|y\|=\sqrt{(y, y)}$. Оператор A является самосопряженным и положительно определенным оператором в $H_h^{(0)}$. Более того, $A=R^*+R$, где $R=R_1+R_2$.

Таким образом, метод (18) является стационарным итерационным методом с самосопряженным оператором A и несамосопряженным оператором B. Было доказано (см. п. 5 § 2 гл. 2 ч. II), что в этом случае выполнение операторного неравенства

$$B_0-0{,}5\tau A\geqslant\frac{1-\rho^2}{2\tau}B^*A^{-1}B,\quad B_0=0{,}5(B+B^*) \tag{23}$$

с константой $\rho\in(0, 1)$ гарантирует сходимость итерационного метода, причем для погрешности справедлива оценка

$$\|y_n-y\|_A\leqslant\rho^n\|y_0-y\|_A. \tag{24}$$

Проверим выполнение неравенства (23) в случае метода верхней релаксации, когда

$$B=\frac{2(2-\omega)}{h^2}E+\omega R,\quad A=R^*+R,\ \tau=\omega. \tag{25}$$

Прежде всего заметим, что

$$B_0-0{,}5\tau A=\frac{2(2-\omega)}{h^2}E,$$

поэтому неравенство (23) упрощается и принимает вид

$$\frac{2(2-\omega)}{h^2}E\geqslant\frac{1-\rho^2}{2\omega}B^*A^{-1}B.$$

Отсюда с помощью эквивалентных преобразований приходим к не-

равенству

$$\frac{2}{h^2}A \geqslant \frac{1-\rho^2}{2\omega(2-\omega)}BB^*.$$

Подставляя сюда выражение для оператора B из (25) и приводя подобные члены, получаем окончательно

$$RR^* + \frac{4}{h^4\alpha^2}E \leqslant \frac{1+\rho^2}{1-\rho^2}\frac{2}{h^2\alpha}(R+R^*), \tag{26}$$

$$\alpha = \omega/(2-\omega).$$

Найдем константу $\rho^2 \in (0, 1)$, при которой справедливо неравенство (26). Для этого оценим сверху левую часть неравенства (26). При любом $y \in H_h^{(0)}$ имеем

$$(RR^*y, y) = \|R^*y\|^2 = \|R_1^*y + R_2^*y\|^2 \leqslant 2(\|R_1^*y\|^2 + \|R_2^*y\|^2). \tag{27}$$

Из определения (22) операторов R_1, R_2 получаем (см. также п. 3 § 1 гл. 4), что

$$(R_1^*y)_{ij} = -\frac{y_{x_1,ij}}{h}, \quad (R_2^*y)_{ij} = -\frac{y_{x_2,ij}}{h},$$

следовательно,

$$\|R_1^*y\|^2 = \frac{1}{h^2}\sum_{i=0}^{N-1}\sum_{j=1}^{N-1}(y_{x_1,ij})^2h^2,$$

$$\|R_2^*y\|^2 = \frac{1}{h^2}\sum_{i=1}^{N-1}\sum_{j=0}^{N-1}(y_{x_2,ij})^2h^2.$$

Таким образом,

$$\|R_1^*y\|^2 + \|R_2^*y\|^2 = \frac{1}{h^2}(Ay, y)$$

и из (27) следует операторное неравенство

$$RR^* \leqslant \frac{2}{h^2}A. \tag{28}$$

Далее, учитывая неравенство

$$A \geqslant \delta E,$$

где $\delta = \frac{8}{h^2}\sin^2\frac{\pi h}{2}$ — наименьшее собственное значение оператора A, получим

$$\frac{4}{h^4\alpha^2}E \leqslant \frac{4}{h^4\alpha^2\delta}A.$$

Итак, левая часть неравенства (26) оценивается следующим образом:

$$RR^* + \frac{4}{h^4\alpha^2}E \leqslant \left(\frac{2}{h^2} + \frac{4}{h^4\alpha^2\delta}\right)A = \left(\frac{2}{h^2} + \frac{4}{h^4\alpha^2\delta}\right)(R+R^*).$$

Неравенство (26) будет выполнено, если константу ρ^2 подобрать из условия

$$\frac{2}{h^2}+\frac{4}{h^4\alpha^2\delta}=\frac{1+\rho^2}{1-\rho^2}\frac{2}{h^2\alpha}. \tag{29}$$

Решая уравнение (29), получим

$$\rho^2=\rho^2(\alpha)=\frac{1-\mu\alpha(1-\alpha)}{1+\mu\alpha(1+\alpha)}, \tag{30}$$

где

$$\mu=0{,}5\,h^2\delta=4\sin^2\frac{\pi h}{2}.$$

Из (30) видно, что при $\omega\in(0, 2)$ (т. е. при $\alpha>0$) выполняется неравенство $\rho^2(\alpha)<1$. Следовательно, метод верхней релаксации сходится при $\omega\in(0, 2)$.

В случае метода Зейделя имеем $\omega=1$, $\alpha=1$,

$$\rho^2=1/(1+h^2\delta).$$

При малых h получаем $\rho^{-1}\approx 1+0{,}5\,h^2\delta\approx 1+\pi^2h^2$, $\ln\rho^{-1}\approx\pi^2h^2$, так что число итераций $n_0(\varepsilon)$, необходимых для получения заданной точности ε, оказывается равным

$$n_0(\varepsilon)=\frac{\ln\varepsilon^{-1}}{\ln\rho^{-1}}\approx\frac{\ln\varepsilon^{-1}}{\pi^2h^2}.$$

Следовательно, необходимое число итераций пропорционально h^{-2}, что свидетельствует о невысокой скорости сходимости метода Зейделя в случае разностных систем уравнений. Отметим, однако, что требуемое число итераций в методе Зейделя примерно в два раза меньше, чем в методе Якоби (см. (9)).

Обратимся снова к методу верхней релаксации и подберем в выражении (30) параметр α таким образом, чтобы минимизировать $\rho^2(\alpha)$. Нетрудно видеть, что минимум $\rho^2(\alpha)$ достигается при $\alpha=1/\sqrt{\mu}$, т. е. при

$$\omega=\frac{2}{1+\sqrt{\mu}},\quad \mu=4\sin^2\frac{\pi h}{2},$$

и он равен

$$\rho^2\left(\frac{1}{\sqrt{\mu}}\right)=\rho_0^2=\frac{1-0{,}5\sqrt{\mu}}{1+0{,}5\sqrt{\mu}}. \tag{31}$$

Подставляя сюда $\mu=4\sin^2\frac{\pi h}{2}$, получим при малых h, что

$$\rho_0^2=\frac{1-\sin(\pi h/2)}{1+\sin(\pi h/2)}\approx\frac{1-\pi h/2}{1+\pi h/2}\approx 1-\pi h,$$

следовательно, $\ln\rho_0^{-1}\approx 0{,}5\pi h$ и необходимое число итераций $n_0(\varepsilon)$ равно

$$n_0(\varepsilon)\approx\frac{2\ln\varepsilon^{-1}}{\pi h}=O\left(\frac{1}{h}\right). \tag{32}$$

§ 2. Применение явного итерационного метода с оптимальным набором параметров

1. Явный итерационный метод с чебышевскими параметрами. Данный метод подробно рассмотрен в § 6 гл. 2 ч. II применительно к системам линейных алгебраических уравнений

$$Ay=f, \tag{1}$$

с положительно определенной симметричной матрицей A. Напомним необходимые для дальнейшего сведения, относящиеся к данному методу.

Пусть выполнены операторные неравенства

$$\gamma_1 E \leqslant A \leqslant \gamma_2 E, \tag{2}$$

где $\gamma_2 > \gamma_1 > 0$, E — единичная матрица. В качестве γ_1 и γ_2 можно взять, соответственно, наименьшее и наибольшее собственные значения матрицы A. Если точные собственные значения неизвестны, то под γ_1 и γ_2 можно подразумевать их границы, т. е. γ_1 — нижняя (положительная) граница минимального собственного значения и γ_2 — верхняя граница максимального собственного значения.

Явный итерационный метод для системы (1) имеет вид

$$\frac{y_{k+1}-y_k}{\tau_{k+1}} + Ay_k = f, \qquad k=0, 1, \ldots, n-1, \tag{3}$$

где k — номер итерации, y_k — приближенное решение системы (1), полученное на k-й итерации. Предполагается, что задано произвольное начальное приближение y_0. В чебышевском итерационном методе параметры τ_k, $k=1, 2, \ldots, n$, подбираются таким образом, чтобы при заданном числе итераций n минимизировать погрешность $\|y_n-y\|$, возникающую на n-й итерации. Под нормой $\|z\|$ вектора z здесь понимается среднеквадратичная норма

$$\|z\| = \left(\sum_{i=1}^{m} z_i^2 \right)^{1/2}.$$

В § 6 гл. 2 ч. II было показано, что оптимальными являются параметры τ_k, определенные следующим образом:

$$\tau_k = \frac{\tau_0}{1+\rho_0 t_k}, \quad \tau_0 = \frac{2}{\gamma_1+\gamma_2}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \tag{4}$$

$$\xi = \frac{\gamma_1}{\gamma_2}, \quad t_k = \cos\frac{(2k-1)\pi}{2n}, \quad k=1, 2, \ldots, n.$$

Если выбрать τ_k согласно (4), то для погрешности будет справедлива оценка

$$\|y_n-y\| \leqslant q_n \|y_0-y\|, \tag{5}$$

где

$$q_n = \frac{2\rho_1^n}{1+\rho_1^{2n}}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2}. \tag{6}$$

Таким образом, чтобы применить чебышевский итерационный метод к конкретным системам уравнений, нужно

1) убедиться в том, что матрица A симметрична (или доказать, что данная матрица является матрицей самосопряженного оператора);

2) найти границы спектра γ_1 и γ_2 матрицы A,

3) вычислить итерационные параметры τ_k согласно (4) и упорядочить их так, чтобы обеспечить устойчивость метода.

В следующих пунктах рассматривается применение данного метода к разностным аппроксимациям уравнений эллиптического типа.

2. Применение к модельной задаче. Для модельной задачи

$$-\frac{y_{i-1\,j}-2y_{ij}+y_{i+1\,j}}{h^2}-\frac{y_{ij-1}-2y_{ij}+y_{i\,j+1}}{h^2}=f_{ij},$$

$$i,\ j=1,\ 2,\ \ldots,\ N-1,\quad hN=1, \tag{7}$$

$$y_{i0}=y_{iN}=0,\quad i=1,\ 2,\ \ldots,\ N-1,\quad y_{0j}=y_{Nj}=0,\quad j=1,\ 2,\ \ldots,\ N-1,$$

в § 1 было показано, что если записать ее в матричном виде (1), то матрица A будет симметричной, причем ее наименьшее и наибольшее собственные значения определяются формулами

$$\gamma_1=\frac{8}{h^2}\sin^2\frac{\pi h}{2},\quad \gamma_2=\frac{8}{h^2}\cos^2\frac{\pi h}{2}.$$

Следовательно, систему (7) можно решать с помощью чебышевского итерационного метода (3), (4). Вычисление $y_{k+1}=\{y_{ij}^{(k+1)}\}_{i,j=1}^{N-1}$ целесообразно организовать следующим образом. Сначала по известным приближениям $y_{ij}^{(k)}$ находится невязка

$$r_{ij}^{(k)}=Ay_{ij}^{(k)}-f_{ij}=$$

$$=-\left(\frac{y_{i-1,j}^{(k)}-2y_{ij}^{(k)}+y_{i+1,j}^{(k)}}{h^2}+\frac{y_{i,j-1}^{(k)}-2y_{ij}^{(k)}+y_{i,j+1}^{(k)}}{h^2}+f_{ij}\right), \tag{8}$$

$$i,\ j=1,\ 2,\ \ldots,\ N-1,$$

а затем досчитываются значения $y_{ij}^{(k+1)}$ по формуле

$$y_{ij}^{(k+1)}=y_{ij}^{(k)}-\tau_{k+1}r_{ij}^{(k)},\qquad i,\ j=1,\ 2,\ \ldots,\ N-1. \tag{9}$$

При этом полагаем

$$y_{i0}^{(k+1)}=y_{iN}^{(k+1)}=0,\quad y_{0j}^{(k+1)}=y_{Nj}^{(k+1)}=0,\quad i,\ j=1,\ 2,\ \ldots,\ N-1. \tag{10}$$

Скорость сходимости итерационного метода определяется параметром

$$\sqrt{\xi}=\sqrt{\frac{\gamma_1}{\gamma_2}}=\operatorname{tg}\frac{\pi h}{2}.$$

Поскольку шаг сетки h невелик, можно считать, что $\sqrt{\xi}\approx 0{,}5\pi h$.

Оценим число итераций n, необходимое для уменьшения начальной погрешности в $1/\varepsilon$ раз. Из неравенства (5) и выражения (6) для q_n следует оценка

$$\|y_n - y\| \leqslant 2\rho^n \|y_0 - y\|.$$

Поэтому достаточно потребовать $2\rho_1^n \leqslant \varepsilon$, т. е.

$$n \geqslant n_0(\varepsilon) = \ln \frac{2}{\varepsilon} \Big/ \ln \frac{1}{\rho_1}.$$

При малых h имеем

$$\ln \frac{1}{\rho_1} \approx 2\sqrt{\xi} \approx \pi h,$$

следовательно,

$$n_0(\varepsilon) \approx \frac{\ln(2/\varepsilon)}{\pi h}. \tag{11}$$

Основной вывод, который можно отсюда сделать, сводится к следующему: при решении с помощью чебышевского итерационного метода разностных задач, аппроксимирующих уравнения эллиптического типа, число итераций $n_0(\varepsilon)$, необходимых для получения заданной точности ε, является величиной $O(h^{-1})$.

Напомним, что метод простой итерации и метод Зейделя требуют $O(h^{-2})$ итераций, что при $h=0{,}1$ на порядок больше. Порядок числа итераций в чебышевском методе тот же, что и в методе верхней релаксации с оптимальным выбором релаксационного параметра ω.

В данном случае интересно провести сравнение необходимого числа итераций в методе верхней релаксации и в чебышевском методе по числу ε. Согласно (32) из § 1 в методе верхней релаксации необходимое число итераций определяется формулой

$$n_0^{(\text{в.р.})}(\varepsilon) \approx \frac{2\ln \varepsilon^{-1}}{\pi h},$$

в то время как для чебышевского итерационного метода

$$n_0(\varepsilon) \approx \frac{\ln(2\varepsilon^{-1})}{\pi h}.$$

Таким образом,

$$n_0^{(\text{в.р.})}(\varepsilon) - n_0(\varepsilon) \approx \frac{\ln(1/(2\varepsilon))}{\pi h}.$$

Следовательно, метод верхней релаксации требует большего числа итераций.

Естественно требовать, чтобы погрешность ε итерационного метода имела тот же порядок h^2, что и погрешность аппроксимации разностной схемы. Поэтому положим $\varepsilon = 0{,}5\,\alpha h^2$, где $\alpha > 0$ — постоянная, не зависящая от h. Тогда получим

$$n_0^{(\text{в.р.})}(\varepsilon) - n_0(\varepsilon) \approx \frac{\ln 1/(\alpha h^2)}{\pi h} = O\left(\frac{1}{h} \ln \frac{1}{h}\right).$$

3. Применение чебышевского метода к разностным аппроксимациям уравнений эллиптического типа. В случае более общих аппроксимаций уравнений эллиптического типа схема применения чебышевского метода остается той же, что и раньше, однако точ-

ные границы спектра γ_1 и γ_2, как правило, не удается найти в аналитической форме. Поэтому используют те или иные оценки для границ спектра.

В качестве примера рассмотрим аппроксимацию задачи Дирихле для уравнения эллиптического типа

$$\frac{\partial}{\partial x_1}\left(k_1(x_1, x_2)\frac{\partial u}{\partial x_1}\right)+\frac{\partial}{\partial x_2}\left(k_2(x_1, x_2)\frac{\partial u}{\partial x_2}\right)-q(x_1, x_2)u=-f(x_1, x_2) \tag{12}$$

в прямоугольнике

$$G=\{0<x_\alpha<l_\alpha,\ \alpha=1,\ 2\}.$$

На границе Γ прямоугольника G задано условие

$$u(x_1, x_2)=\mu(x_1, x_2), \qquad (x_1, x_2)\in\Gamma. \tag{13}$$

Предполагаем, что при всех $(x_1, x_2)\in G$ выполнены неравенства

$$\begin{gathered}0<c_{1,\alpha}\leqslant k_\alpha(x_1, x_2)\leqslant c_{2,\alpha}, \qquad \alpha=1,\ 2,\\ 0\leqslant d_1\leqslant q(x_1, x_2)\leqslant d_2.\end{gathered} \tag{14}$$

Введем в G прямоугольную сетку Ω с шагами h_1 и h_2 по направлениям x_1, x_2 соответственно и обозначим

$$x_1^{(i)}=ih_1, \quad x_2^{(j)}=jh_2, \quad x_{ij}=(x_1^{(i)}, x_2^{(j)}),$$

$$h_1N_1=l_1, \quad h_2N_2=l_2, \quad y_{ij}=y(x_{ij}),$$

$$i=0,\ 1,\ \ldots,\ N_1, \quad j=0,\ 1,\ \ldots,\ N_2,$$

$$(a_1y_{\bar{x}_1})_{x_1,ij}=\frac{1}{h_1}\left(a_{1,i+1,j}\frac{y_{i+1,j}-y_{ij}}{h_1}-a_{1,ij}\frac{y_{ij}-y_{i-1,j}}{h_1}\right),$$

$$(a_2y_{\bar{x}_2})_{x_2,ij}=\frac{1}{h_2}\left(a_{2,i,j+1}\frac{y_{i,j+1}-y_{ij}}{h_2}-a_{2,ij}\frac{y_{ij}-y_{i,j-1}}{h_2}\right).$$

Обозначим через γ сеточную границу, т. е. пересечение Ω с границей Γ.

Заменим исходную дифференциальную задачу (12), (13) разностной схемой второго порядка аппроксимации

$$(a_1y_{\bar{x}_1})_{x_1,ij}+(a_2y_{\bar{x}_2})_{x_2,ij}-d_{ij}y_{ij}=-f_{ij},$$

$$i=1,\ 2,\ \ldots,\ N_1-1, \quad j=1,\ 2,\ \ldots,\ N_2-1, \tag{15}$$

$$y_{ij}=\mu_{ij}, \text{ если } x_{ij}\in\gamma. \tag{16}$$

Здесь

$$d_{ij}=q_{ij},$$

$$a_{1,ij}=0{,}5\,(k_1(x_1^{(i)}, x_2^{(j)})+k_1(x_1^{(i-1)}, x_2^{(j)})),$$

$$a_{2,ij}=0{,}5\,(k_2(x_1^{(i)}, x_2^{(j)})+k_2(x_1^{(i)}, x_2^{(j-1)})).$$

Покажем, что разностную задачу (15), (16) можно записать в операторной форме (1), где A — самосопряженный оператор, и получим для этого оператора оценки вида (2).

Прежде всего заметим, что, изменив соответствующим образом правую часть уравнения (15), можно считать, что $y_{ij}=0$ при $x_{ij}\in\gamma$. Таким образом, придем к эквивалентной (15), (16) системе уравнений

$$(a_1 y_{\bar{x}_1})_{x_1,ij} + (a_2 y_{\bar{x}_2})_{x_2,ij} - d_{ij}y_{ij} = -\tilde{f}_{ij}, \tag{17}$$

$$i=1, 2, \ldots, N_1-1, \quad j=1, 2, \ldots, N_2-1,$$

$$y_{ij}=0, \text{ если } x_{ij}\in\gamma, \tag{18}$$

где $\tilde{f}_{ij}$ отличается от f_{ij} только в приграничных точках сетки.

Рассмотрим пространство H функций, заданных на сетке Ω и обращающихся в нуль на γ. Определим в H скалярное произведение и норму

$$(y, v) = \sum_{i=1}^{N_1-1} h_1 \sum_{j=1}^{N_2-1} h_2 y_{ij} v_{ij}, \qquad \|y\| = \sqrt{(y, y)}.$$

Далее, зададим в H оператор A формулами

$$(Ay)_{ij} = -(a_1 y_{\bar{x}_1})_{x_1,ij} - (a_2 y_{\bar{x}_2})_{x_2,ij} + d_{ij}y_{ij}, \tag{19}$$

$$i=1, 2, \ldots, N_1-1, \quad j=1, 2, \ldots, N_2-1.$$

Тогда разностную схему (17), (18) можно записать в виде операторного уравнения (1) в пространстве H.

Из разностной формулы Грина (см. (15) из § 3 гл. 1) следует, что для оператора (19) при любых y, $v\in H$ справедливо тождество

$$(Ay, v) = \sum_{i=1}^{N_1} h_1 \sum_{j=1}^{N_2-1} h_2 a_{1,ij} y_{\bar{x}_1,ij} v_{\bar{x}_1,ij} + \\ + \sum_{i=1}^{N_1-1} h_1 \sum_{j=1}^{N_2} h_2 a_{2,ij} y_{\bar{x}_2,ij} v_{\bar{x}_2,ij} + \sum_{i=1}^{N_1-1} h_1 \sum_{j=1}^{N_2-1} h_2 d_{ij} y_{ij} v_{ij}. \tag{20}$$

Отсюда, меняя местами y и v, легко установить, что $(Ay, v) = (y, Av)$ при любых y, $v\in H$. Следовательно, разностной схеме (15), (16) соответствует самосопряженный оператор A.

Далее, полагая в тождестве (20) $y=v$, получим

$$(Ay, y) = \sum_{i=1}^{N_1} h_1 \sum_{j=1}^{N_2-1} h_2 a_{1,ij} (y_{\bar{x}_1,ij})^2 + \\ + \sum_{i=1}^{N_1-1} h_1 \sum_{j=1}^{N_2} h_2 a_{2,ij} (y_{\bar{x}_2,ij})^2 + \sum_{i=1}^{N_1-1} h_1 \sum_{j=1}^{N_2-1} h_2 d_{ij} (y_{ij})^2. \tag{21}$$

Отсюда, учитывая неравенства (14), приходим к оценкам

$$\beta_1 (\overset{\circ}{A}y, y) + d_1\|y\|^2 \leqslant (Ay, y) \leqslant \beta_2 (\overset{\circ}{A}y, y) + d_2\|y\|^2, \tag{22}$$

где

$$(\mathring{A}y, y) = \sum_{i=1}^{N_1} h_1 \sum_{j=1}^{N_2-1} h_2 (y_{\bar{x}_1, ij})^2 + \sum_{i=1}^{N_1-1} h_1 \sum_{j=1}^{N_2} h_2 (y_{\bar{x}_2, ij})^2, \tag{23}$$

$$\beta_1 = c_{1,1} + c_{1,2}, \quad \beta_2 = c_{2,1} + c_{2,2}. \tag{24}$$

Обозначение $(\mathring{A}y, y)$ объясняется тем, что сумма, стоящая в правой части (23), представляет собой скалярное произведение двух векторов y и $\mathring{A}y$, где

$$(\mathring{A}y)_{ij} = -y_{\bar{x}_1 x_1, ij} - y_{\bar{x}_2 x_2, ij}, \quad i = 1, 2, \ldots, N_1 - 1, \quad j = 1, 2, \ldots, N_2 - 1.$$

Поскольку

$$\delta \|y\|^2 \leqslant (\mathring{A}y, y) \leqslant \Delta \|y\|^2,$$

где

$$\delta = \frac{4}{h_1^2} \sin^2 \frac{\pi h_1}{2l_1} + \frac{4}{h_2^2} \sin^2 \frac{\pi h_2}{2l_2}, \tag{25}$$

$$\Delta = \frac{4}{h_1^2} \cos^2 \frac{\pi h_1}{2l_1} + \frac{4}{h_2^2} \cos^2 \frac{\pi h_2}{2l_2} \tag{26}$$

(см. § 1 гл. 3), из (22) следуют операторные неравенства (2) с константами

$$\gamma_1 = \beta_1 \delta + d_1, \quad \gamma_2 = \beta_2 \Delta + d_2. \tag{27}$$

Отсюда по формулам (4) можно вычислить итерационные параметры τ_k и оценить согласно (5), (6) величину погрешности.

Отметим, что приведение разностной схемы (15), (16) к виду (17), (18) потребовалось нам только для того, чтобы определить оператор A и получить оценки его спектра. После того как параметры τ_k найдены, итерации можно проводить непосредственно для схемы (15), (16). Сначала вычисляется невязка

$$r_{ij}^{(k)} = -(a_1 y_{\bar{x}_1}^{(k)})_{x_1, ij} - (a_2 y_{\bar{x}_2}^{(k)})_{x_2, ij} + d_{ij} y_{ij}^{(k)} - f_{ij}, \quad i = 1, 2, \ldots, N_1 - 1, \quad j = 1, 2, \ldots, N_2 - 1,$$

а затем находится новое приближение

$$y_{ij}^{(k+1)} = y_{ij}^{(k)} - \tau_{k+1} r_{ij}^{(k)}, \quad i = 1, 2, \ldots, N_1 - 1, \quad j = 1, 2, \ldots, N_2 - 1.$$

Граничные условия доопределяются согласно (16): $y_{ij}^{(k+1)} = \mu_{ij}$, если $x_{ij} \in \gamma$.

§ 3. Попеременно-треугольный итерационный метод

1. Алгебраическая теория. Пусть дана система линейных алгебраических уравнений

$$Ay = f \tag{1}$$

с симметричной положительно определенной матрицей A порядка

m. Зададим матрицу $R=(r_{ij})$ следующим образом:

$$r_{ij}=\begin{cases} a_{ij}, & \text{если } i>j, \\ 0{,}5\,a_{jj}, & \text{если } i=j, \\ 0, & \text{если } i<j. \end{cases}$$

Тогда матрицу A можно представить в виде суммы $A=R+R^*$, где через R^* обозначена матрица, сопряженная с матрицей R (транспонированная к R в случае действительных матриц и комплексно-сопряженная — в случае комплексных матриц). Ясно, что R — нижняя треугольная матрица и R^* — верхняя треугольная, причем диагонали матриц R и R^* совпадают.

В дальнейшем удобно рассматривать систему уравнений (1) как операторное уравнение с самосопряженным положительным оператором A, действующим в конечномерном евклидовом (унитарном — в комплексном случае) пространстве.

Попеременно-треугольный итерационный метод, который будет рассматриваться в настоящем параграфе, относится к неявным итерационным методам вида

$$B\frac{y_{k+1}-y_k}{\tau}+Ay_k=f \tag{2}$$

с самосопряженным положительным оператором B. А именно, оператор B в попеременно-треугольном итерационном методе определяется как произведение

$$B=(E+\omega R^*)(E+\omega R), \tag{3}$$

где E — единичный оператор и $\omega>0$ — числовой параметр.

В дальнейшем параметры ω и τ будут выбраны исходя из условий сходимости итерационного метода (2), (3). Если ω и τ известны, то новая итерация y_{k+1} находится из уравнения (2) в два этапа. На первом этапе находится промежуточное значение, которое мы обозначим через $y_{k+1/2}$, как решение уравнения

$$(E+\omega R^*)y_{k+1/2}=\varphi_k, \tag{4а}$$

где $\varphi_k=By_k-\tau Ay_k+\tau f$. На втором этапе, используя найденное значение $y_{k+1/2}$, решается относительно y_{k+1} уравнение

$$(E+\omega R)y_{k+1}=y_{k+1/2}. \tag{4б}$$

Решение уравнений (4а), (4б) не представляет труда, поскольку матрицы $E+\omega R^*$ и $E+\omega R$ являются треугольными.

Исследование сходимости попеременно-треугольного метода (2), (3) основано на теореме 1 из § 4 гл. 2 ч. II о сходимости неявных итерационных методов с самосопряженными операторами A, B. В основе этой теоремы лежит предположение о том, что операторы A и B связаны неравенствами

$$\gamma_1 B\leqslant A\leqslant\gamma_2 B, \tag{5}$$

где γ_1 и γ_2 — положительные постоянные. Поэтому нам прежде всего надо доказать неравенства (5) для оператора (3).

Лемма 1. *Пусть существуют положительные постоянные* δ, Δ *такие, что выполнены операторные неравенства*

$$A \geqslant \delta E, \tag{6}$$

$$4R^*R \leqslant \Delta A. \tag{7}$$

Тогда для операторов $A = R^* + R$ *и* $B = (E + \omega R^*)(E + \omega R)$ *справедливы неравенства* (5), *где*

$$\gamma_1 = \left(\frac{1}{\delta} + \omega + \frac{\omega^2 \Delta}{4}\right)^{-1}, \quad \gamma_2 = \frac{1}{2\omega}. \tag{8}$$

Доказательство. Рассмотрим операторы

$$B = B(\omega) = (E + \omega R^*)(E + \omega R) = E + \omega A + \omega^2 R^* R,$$

$$B(-\omega) = (E - \omega R^*)(E - \omega R) = E - \omega A + \omega^2 R^* R.$$

Отсюда получим

$$B(\omega) - B(-\omega) = 2\omega A,$$

следовательно,

$$B = B(\omega) \geqslant 2\omega A,$$

поскольку $B(-\omega) \geqslant 0$. Таким образом, $A \leqslant \gamma_2 B$, где $\gamma_2 = (2\omega)^{-1}$.

Далее, учитывая предположения (6), (7), получим

$$B = E + \omega A + \omega^2 R^* R \leqslant \frac{1}{\delta} A + \omega A + \frac{\omega^2 \Delta}{4} A,$$

т. е. $A \geqslant \gamma_1 B$, где константа γ_1 определена согласно (8). Лемма 1 доказана.

Замечание 1. В качестве константы δ в условии (6) можно взять минимальное собственное значение $\lambda_{\min}(A)$ оператора A или любую положительную постоянную, не превосходящую $\lambda_{\min}(A)$.

Замечание 2. Докажем, что если выполнено условие (7) с некоторой константой $\Delta > 0$, то при $A = R^* + R > 0$ выполняется неравенство $\Delta \geqslant \lambda_{\max}(A)$, где $\lambda_{\max}(A)$ — максимальное собственное значение оператора A. Преобразуем (7) с помощью следующей цепочки эквивалентных преобразований (см. п. 4 § 1 гл. 3):

$$R^*R \leqslant \frac{\Delta}{4}(R^* + R), \quad E \leqslant \frac{\Delta}{4}(R^{-1} + R^{*-1}), \quad RR^* \leqslant \frac{\Delta}{4}(R + R^*).$$

Таким образом, из (7) следует неравенство

$$RR^* + R^*R \leqslant \frac{\Delta}{2}(R + R^*). \tag{9}$$

С другой стороны, воспользовавшись тождеством

$$2(R^*R + RR^*) = (R^* + R)^2 + (R^* - R)(R^* - R)^*,$$

получим

$$R^*R + RR^* \geqslant 0{,}5(R^* + R)^2.$$

Отсюда и из (9) приходим к неравенству

$$(R^* + R)^2 \leqslant \Delta(R^* + R),$$

которое эквивалентно неравенству

$$A=R^*+R\leqslant\Delta E,$$

означающему, что $\lambda_{\max}(A)\leqslant\Delta$. Учитывая замечание 1, видим, что если выполнены неравенства (6), (7) и $\lambda_{\min}(A)\neq\lambda_{\max}(A)$, то $\Delta>\delta$.

Обратимся теперь к исследованию сходимости попеременно-треугольного итерационного метода.

Теорема 1. *Предположим, что $A=R^*+R$ и существуют положительные постоянные δ, Δ, при которых выполнены неравенства $A\geqslant\delta E$, $4R^*R\leqslant\Delta A$. Пусть*

$$\omega=\frac{2}{\sqrt{\delta\Delta}},\quad \tau=\frac{2}{\gamma_1+\gamma_2}, \tag{10}$$

где

$$\gamma_1=\frac{\delta}{2(1+\sqrt{\xi})},\quad \xi=\frac{\delta}{\Delta},\quad \gamma_2=\frac{\sqrt{\delta\Delta}}{4}. \tag{11}$$

Тогда итерационный метод (2), (3) *сходится, причем для погрешности справедлива оценка*

$$\|y_k-y\|_A\leqslant\rho^k\|y_0-y\|_A, \tag{12}$$

где

$$\rho=\frac{1-\sqrt{\xi}}{1+3\sqrt{\xi}}. \tag{13}$$

Доказательство. В лемме 1 установлено, что при любом $\omega>0$ операторы A и B рассматриваемого итерационного метода связаны неравенствами (5), где $\gamma_1=\gamma_1(\omega)$ и $\gamma_2=\gamma_2(\omega)$ определены согласно (8). Поэтому выполнены все предположения теоремы 1 из § 4 гл. 2 ч. II о сходимости стационарных итерационных методов с самосопряженными операторами A и B. Согласно этой теореме, для выполнения оценки (12) с константой

$$\rho=\frac{1-\eta}{1+\eta},\qquad \eta=\frac{\gamma_1(\omega)}{\gamma_2(\omega)}$$

достаточно положить $\tau=2/(\gamma_1+\gamma_2)$. Выберем теперь параметр ω так, чтобы минимизировать ρ. Для этого достаточно найти значение $\omega=\omega_0$, при котором функция

$$f(\omega)=\eta^{-1}=\frac{\gamma_2(\omega)}{\gamma_1(\omega)}$$

достигает максимума. Из формул (8) имеем

$$f(\omega)=\frac{1}{2}+\frac{1}{2}\left(\frac{1}{\omega\delta}+\frac{\omega\Delta}{4}\right),$$

откуда видно, что $f(\omega)$ достигает максимума при $\omega=\omega_0=2/\sqrt{\delta\Delta}$. Подставляя $\omega=\omega_0$ в выражения (8) для γ_1 и γ_2, получим их значения, совпадающие с (11). При этом для константы

$$\rho=\frac{1-\eta}{1+\eta},\qquad \eta=\frac{\gamma_1(\omega_0)}{\gamma_2(\omega_0)}=\frac{2\sqrt{\xi}}{1+\sqrt{\xi}}$$

получаем выражение (13). Теорема 1 доказана.

2. **Применение к модельной задаче.** Рассмотрим модельную задачу

$$-y_{\bar{x}_1 x_1, ij} - y_{\bar{x}_2 x_2, ij} = f_{ij}, \quad i, j = 1, 2, \ldots, N-1, \quad hN = 1, \tag{14}$$
$$y_{i0} = y_{iN} = 0, \quad y_{0j} = y_{Nj} = 0, \quad i, j = 1, 2, \ldots, N-1.$$

Введем пространство H функций, заданных на сетке

$$\Omega = \{x_{ij} = (x_1^{(i)}, x_2^{(j)}), x_1^{(i)} = ih, x_2^{(j)} = jh\}_{i,j=0}^{N}$$

и обращающихся в нуль на ее границе. Определим в H скалярное произведение

$$(y, v) = \sum_{i,j=1}^{N-1} y_{ij} v_{ij} h^2$$

и норму $\|y\| = \sqrt{(y, y)}$. Задача (14) записывается как операторное уравнение (1) в пространстве H, где оператор A определен следующим образом:

$$(Ay)_{ij} = -y_{\bar{x}_1 x_1, ij} - y_{\bar{x}_2 x_2, ij}, \quad i, j = 1, 2, \ldots, N-1. \tag{15}$$

Этот оператор является самосопряженным и положительным. Для того чтобы применить к системе (14) попеременно-треугольный итерационный метод, необходимо представить матрицу оператора (15) в виде $A = R + R^*$, где R — нижняя треугольная матрица, и найти константы δ и Δ, входящие в неравенства (6), (7).

Запишем (15) в виде

$$(Ay)_{ij} = \frac{y_{\bar{x}_1, ij} + y_{\bar{x}_2, ij}}{h} - \frac{y_{x_1, ij} + y_{x_2, ij}}{h}$$

или, более подробно, в виде

$$(Ay)_{ij} = \frac{1}{h}\left(\frac{y_{ij} - y_{i-1,j}}{h} + \frac{y_{ij} - y_{i,j-1}}{h}\right) - \\ - \frac{1}{h}\left(\frac{y_{i+1,j} - y_{ij}}{h} + \frac{y_{i,j+1} - y_{ij}}{h}\right). \tag{16}$$

Тем самым оператор A представлен как сумма двух операторов, $A = R + U$, где

$$(Ry)_{ij} = \frac{1}{h}(y_{\bar{x}_1, ij} + y_{\bar{x}_2, ij}),$$
$$(Uy)_{ij} = -\frac{1}{h}(y_{x_1, ij} + y_{x_2, ij}). \tag{17}$$

Нетрудно понять, что матрица оператора R является нижней треугольной, а матрица оператора U — верхней треугольной. Чтобы убедиться в этом, достаточно записать систему двумерных разностных уравнений (14) в виде одномерной системы (5) из § 1.

Более того, оператор U является сопряженным оператору R в пространстве H. Для доказательства вычислим скалярное произведение (Ry, v), где y и v — любые сеточные функции, заданные на сетке Ω и обращающиеся в нуль на ее границе. По определению оператора R имеем

$$(Ry, v) = \sum_{i,j=1}^{N-1} [(y_{ij} - y_{i-1,j}) + (y_{ij} - y_{i,j-1})] v_{ij} =$$

$$= 2 \sum_{i,j=1}^{N-1} y_{ij} v_{ij} - \sum_{i=0}^{N-2} \sum_{j=1}^{N-1} y_{ij} v_{i+1,j} - \sum_{i=1}^{N-1} \sum_{j=0}^{N-2} y_{ij} v_{i,j+1}.$$

С другой стороны,

$$(y, Uv) = 2 \sum_{i,j=1}^{N-1} y_{ij} v_{ij} - \sum_{i,j=1}^{N-1} y_{ij} v_{i+1,j} - \sum_{i,j=1}^{N-1} y_{ij} v_{i,j+1},$$

и, следовательно,

$$(Ry, v) - (y, Uv) = \sum_{j=1}^{N-1} (y_{N-1,j} v_{Nj} - y_{0j} v_{1j}) + \sum_{i=1}^{N-1} (y_{i,N-1} v_{iN} - y_{i0} v_{i1}).$$

Выражение, стоящее в правой части последнего равенства, равно нулю в силу граничных условий. Таким образом, $(Ry, v) = (y, Uv)$ для любых $y, v \in H$, т. е. $U = R^*$. Искомое разложение $A = R + R^*$ получено.

Докажем теперь неравенства (6), (7). Как уже отмечалось, в качестве константы δ можно взять минимальное собственное значение оператора A, т. е.

$$\delta = \frac{8}{h^2} \sin^2 \frac{\pi h}{2}.$$

Проверим выполнение неравенства (7), которое означает, что

$$4 \|Ry\|^2 \leqslant \Delta (Ay, y) \tag{18}$$

для любого $y \in H$. Как показано в п. 2 § 2 гл. 3, справедливо тождество

$$(Ay, y) = \sum_{i=1}^{N} \sum_{j=1}^{N-1} (y_{\overline{x}_1, ij})^2 h^2 + \sum_{i=1}^{N-1} \sum_{j=1}^{N} (y_{\overline{x}_2, ij})^2 h^2.$$

С другой стороны, из определения (17) оператора R следует, что

$$\|Ry\|^2 = \frac{1}{h^2} \sum_{i,j=1}^{N-1} (y_{\overline{x}_1, ij} + y_{\overline{x}_2, ij})^2 h^2,$$

и поэтому

$$\|Ry\|^2 \leqslant \frac{2}{h^2} \left(\sum_{i,j=1}^{N-1} (y_{\overline{x}_1, ij})^2 h^2 + \sum_{i,j=1}^{N-1} (y_{\overline{x}_2, ij})^2 h^2 \right) \leqslant \frac{2}{h^2} (Ay, y).$$

Таким образом, требуемое неравенство (18) выполнено с константой $\Delta = 8/h^2$.

Заметим, что константа Δ в данном случае незначительно отличается от максимального собственного значения оператора A, которое равно $\frac{8}{h^2}\cos^2\frac{\pi h}{2}$.

Чтобы окончательно задать попеременно-треугольный метод для решения системы (14), надо в соответствии с теоремой 1 определить параметры ω и τ.

Подставляя найденные выражения для δ, Δ в формулы (10), (11), получим

$$\sqrt{\xi} = \sqrt{\frac{\delta}{\Delta}} = \sin\frac{\pi h}{2}, \quad \sqrt{\delta\Delta} = \frac{8}{h^2}\sin\frac{\pi h}{2},$$

$$\eta = \frac{\gamma_1}{\gamma_2} = \frac{2\sin\frac{\pi h}{2}}{1+\sin\frac{\pi h}{2}} \approx \pi h, \tag{19}$$

$$\omega = \frac{h^2}{4\sin\frac{\pi h}{2}} \approx \frac{h}{2\pi}, \quad \tau = \frac{h^2\left(1+\sin\frac{\pi h}{2}\right)}{\sin\frac{\pi h}{2}\left(1+3\sin\frac{\pi h}{2}\right)} \approx \frac{2h}{\pi}.$$

Константа ρ из оценки (12) в данном случае равна

$$\rho = \frac{1-\sin\frac{\pi h}{2}}{1+3\sin\frac{\pi h}{2}} \approx 1 - 2\pi h.$$

Поэтому при малых h число итераций $n_0(\varepsilon)$, необходимых для получения заданной точности ε, оценивается как

$$n_0(\varepsilon) \approx \frac{\ln(1/\varepsilon)}{2\pi h}. \tag{20}$$

Алгоритм нахождения значений $y_{ij}^{(k+1)}$ на новой итерации $k+1$ в соответствии с (4а), (4б) состоит в следующем. На первом этапе решается система уравнений

$$y_{ij}^{(k+1/2)} - \frac{\omega}{h}\left(\frac{y_{i+1,j}^{(k+1/2)} - y_{ij}^{(k+1/2)}}{h} + \frac{y_{i,j+1}^{(k+1/2)} - y_{ij}^{(k+1/2)}}{h}\right) = \varphi_{ij}^{(k)},$$

$$i, j = 1, 2, \ldots, N-1, \tag{21}$$

$$y_{Nj}^{(k+1/2)} = 0, \quad j = 1, 2, \ldots, N-1,$$

$$y_{iN}^{(k+1/2)} = 0, \quad i = 1, 2, \ldots, N-1,$$

где $\varphi_{ij}^{(k)} = (By^{(k)})_{ij} - \tau(Ay^{(k)})_{ij} + \tau f_{ij}$, из которой находятся промежуточ-

ные значения $y^{(k+1/2)}$. На втором этапе решается система уравнений

$$y_{ij}^{(k+1)} + \frac{\omega}{h}\left(\frac{y_{ij}^{(k+1)} - y_{i-1,j}^{(k+1)}}{h} + \frac{y_{ij}^{(k+1)} - y_{i,j-1}^{(k+1)}}{h}\right) = y_{ij}^{(k+1/2)},$$

$$i, j = 1, 2, \ldots, N-1, \quad (22)$$

$$y_{0j}^{(k+1)} = 0, \quad j = 1, 2, \ldots, N-1,$$

$$y_{i0}^{(k+1)} = 0, \quad i = 1, 2, \ldots, N-1.$$

Параметры ω и τ выбираются здесь согласно (19). Уравнения (21) следует начинать решать с точки $i=N-1$, $j=N-1$. В этой точке, учитывая граничные условия

$$y_{N,N-1}^{(k+1/2)} = 0, \quad y_{N-1,N}^{(k+1/2)} = 0,$$

уравнение (21) можно записать в виде

$$y_{N-1,N-1}^{(k+1/2)} + \frac{\omega}{h}\left(\frac{y_{N-1,N-1}^{(k+1/2)}}{h} + \frac{y_{N-1,N-1}^{(k+1/2)}}{h}\right) = \varphi_{N-1,N-1}^{(k)},$$

откуда сразу же найдем $y_{N-1,N-1}^{(k+1/2)}$. Далее, проводим вычисления в точке $i=N-2$, $j=N-1$ и находим $y_{N-2,N-1}^{(k+1/2)}$, затем продвигаемся влево еще на одну точку и т. д. После нахождения $y_{1,N-1}^{(k+1/2)}$ переходим в точку $i=N-1$, $j=N-2$ и т. д. Таким образом вычисления по формулам (21) осуществляются явным образом, причем счет ведется, начиная с правого верхнего угла области G (от точки $i=N-1$, $j=N-1$) и вплоть до левого нижнего угла (до точки $i=1$, $j=1$).

Система уравнений (22) решается аналогично, однако вычисления здесь начинаются в точке $i=1$, $j=1$ и заканчиваются в точке $i=N-1$, $j=N-1$.

3. Попеременно-треугольный метод с чебышевскими итерационными параметрами. Как мы только что видели, попеременно-треугольный итерационный метод с постоянным параметром τ при решении разностных краевых задач требует $O(h^{-1})$ итераций для достижения заданной точности. Покажем теперь, что использование итерационного метода

$$B\frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \ldots, n-1, \quad (23)$$

$$A = R^* + R, \quad B = (E + \omega R^*)(E + \omega R) \quad (24)$$

при соответствующем выборе параметров τ_k, ω позволяет сократить число итераций до $O(h^{-1/2})$.

Воспользуемся теоремой 3 из § 6 гл. 2 ч. II о сходимости неявного чебышевского итерационного метода. Согласно этой теореме, при заданном числе итераций n параметры τ_k выбираются по правилу

$$\tau_k = \frac{\tau_0}{1 + \rho_0 t_k}, \quad k = 1, 2, \ldots, n, \quad (25)$$

где $\tau_0 = \dfrac{2}{\gamma_1+\gamma_2}$, $\rho_0 = \dfrac{1-\eta}{1+\eta}$, $\eta = \dfrac{\gamma_1}{\gamma_2}$, $t_k = \cos\dfrac{(2k-1)\pi}{2n}$,

$k = 1, 2, \ldots, n$, γ_1, γ_2 — константы из неравенства (5).

При этом для малых η число итераций $n_0(\varepsilon)$, необходимых для получения заданной точности ε, примерно равно

$$n_0(\varepsilon) \approx \frac{\ln(2/\varepsilon)}{2\sqrt{\eta}}.$$

Остается заметить, что для операторов (24) константы γ_1 и γ_2 определены согласно (11) и в случае задачи (14) согласно (19) имеем $\eta \approx \pi h$ и поэтому

$$n_0(\varepsilon) \approx \frac{\ln(2/\varepsilon)}{2\sqrt{\pi}\sqrt{h}} = O\left(\frac{1}{\sqrt{h}}\right).$$

При практическом применении данного метода следует использовать итерационные параметры τ_k в том порядке, который обеспечивает вычислительную устойчивость.

4. Модифицированный попеременно-треугольный итерационный метод. Зададим диагональную матрицу D с положительными элементами на диагонали и будем рассматривать итерационный метод (2), где

$$B = (D+\omega R^*)D^{-1}(D+\omega R). \tag{26}$$

Если $D=E$, то получаем рассмотренный ранее попеременно-треугольный итерационный метод. Если $D \neq E$, то приходим к обобщению метода (2), (3), которое при правильном выборе матрицы D позволяет несколько уменьшить число итераций. Дополнительных трудностей при вычислении новой итерации y_{k+1} здесь не возникает. Вместо алгоритма (4а), (4б) можно использовать следующий алгоритм определения y_{k+1}:

$$(D+\omega R^*)\,y_{k+1/2} = \varphi_k, \qquad \varphi_k = By_k - \tau Ay_k + \tau f,$$
$$(D+\omega R)\,y_{k+1} = Dy_{k+1/2}.$$

Таким образом, нахождение y_{k+1} снова сводится к решению двух систем уравнений с треугольными матрицами.

В следующей теореме получена оценка скорости сходимости итерационного метода (2), (26).

Теорема 2. *Пусть $A=R^*+R$. Предположим, что существует самосопряженный положительный оператор D и положительные постоянные δ_D, Δ_D, для которых выполнены неравенства*

$$A \geqslant \delta_D D, \tag{27}$$

$$4R^*D^{-1}R \leqslant \Delta_D A. \tag{28}$$

Положим

$$\omega = \frac{2}{\sqrt{\delta_D\Delta_D}}, \quad \tau = \frac{2}{\gamma_1+\gamma_2}, \tag{29}$$

где

$$\gamma_1 = \frac{\delta_D}{2(1+\sqrt{\xi_D})}, \quad \xi_D = \frac{\delta_D}{\Delta_D}, \quad \gamma_2 = \frac{\sqrt{\delta_D \Delta_D}}{4}. \tag{30}$$

Тогда для погрешности итерационного метода (2), (26) *справедлива оценка*

$$\|y_k - y\|_A \leqslant \rho_D^k \|y_0 - y\|_A, \tag{31}$$

где

$$\rho_D = \frac{1-\sqrt{\xi_D}}{1+3\sqrt{\xi_D}}. \tag{32}$$

Доказательство. Погрешность метода (2), (26) $z_k = y_k - y$ удовлетворяет однородному уравнению

$$(D+\omega R^*) D^{-1} (D+\omega R) \frac{z_{k+1}-z_k}{\tau} + Az_k = 0, \tag{33}$$

$$k = 0, 1, \ldots, \quad z_0 = y_0 - y.$$

Поскольку $D^* = D > 0$, существуют самосопряженные положительные операторы $D^{1/2}$, $D^{-1/2}$. Сделаем в уравнении (33) замену $z_k = D^{-1/2} v_k$ и обозначим

$$R_D = D^{-1/2} R D^{-1/2}, \quad A_D = D^{-1/2} A D^{-1/2}.$$

Тогда после умножения на оператор $D^{-1/2}$ уравнение (33) приводится к виду

$$(E+\omega R_D^*)(E+\omega R_D) \frac{v_{k+1}-v_k}{\tau} + A_D v_k = 0. \tag{34}$$

Так как $A_D = R_D^* + R_D$, уравнение (34) представляет собой уравнение для погрешности немодифицированного попеременно-треугольного итерационного метода (2), (3). При этом $v_k = x_k - x$, где $x = D^{1/2} y$ является решением уравнения $A_D x = D^{-1/2} f$, а $x_k = D^{1/2} y_k$ — приближение к x, полученное на k-й итерации.

Для оценки v_k применим теорему 1. Условия (27), (28) эквивалентны, соответственно, условиям

$$A_D \geqslant \delta_D E, \quad 4R_D^* R_D \leqslant \Delta_D A_D.$$

Выбирая ω и τ согласно (29), (30), получаем, что выполнены условия (10) и (11) теоремы 1. Поэтому для решения уравнения (34) справедлива оценка (12), которая в данном случае принимает вид

$$\|v_k\|_{A_D} \leqslant \rho_D^k \|v_0\|_{A_D},$$

где ρ_D определено согласно (32).

Замечая, что

$$\|v_k\|_{A_D}^2 = (A_D v_k \; v_k) = (D^{-1/2} A D^{-1/2} (D^{1/2} y_k - D^{1/2} y),$$
$$(D^{1/2} y_k - D^{1/2} y)) = (A(y_k - y), y_k - y) = \|y_k - y\|_A^2,$$

приходим к оценке (31). Теорема 2 доказана.

Смысл введения модифицированного попеременно-треугольного метода состоит в том, что при соответствующих D константа ρ_D, входящая в оценку (31), оказывается меньше, чем константа ρ из оценки (12). В [35, с. 425] указан способ выбора диагональной матрицы D, минимизирующей константу ρ_D в случае разностных аппроксимаций уравнений эллиптического типа с переменными коэффициентами.

§ 4. Итерационный метод переменных направлений

1. Формулировка метода и исследование сходимости. Рассмотрим систему линейных алгебраических уравнений

$$Ay=f \tag{1}$$

с невырожденной квадратной матрицей A порядка m и предположим, что $A=A_1+A_2$ представлена в виде суммы двух матриц A_1 и A_2 более простой структуры. Например, в случае разностных аппроксимаций двумерных эллиптических задач матрица A_α аппроксимирует производные только по переменной x_α, $\alpha=1, 2$.

Тогда можно предложить следующий итерационный метод решения системы (1), аналогичный методу переменных направлений для двумерного уравнения теплопроводности (см. § 4 гл. 4).

Переход от k-й итерации к $(k+1)$-й осуществляется в два этапа. На первом этапе находится промежуточное значение $y_{k+1/2}$ как решение системы уравнений

$$\frac{y_{k+1/2}-y_k}{\tau}+A_1y_{k+1/2}+A_2y_k=f. \tag{2}$$

На втором этапе решается система уравнений

$$\frac{y_{k+1}-y_{k+1/2}}{\tau}+A_1y_{k+1/2}+A_2y_{k+1}=f, \tag{3}$$

из которой находится y_{k+1}. Здесь $\tau>0$ — итерационный параметр, предполагается, что задано произвольное начальное приближение y_0.

Записывая уравнения (2), (3) в виде

$$(E+\tau A_1)\,y_{k+1/2}=(E-\tau A_2)\,y_k+\tau f, \tag{4}$$

$$(E+\tau A_2)\,y_{k+1}=(E-\tau A_1)\,y_{k+1/2}+\tau f, \tag{5}$$

убеждаемся в том, что для нахождения y_{k+1} необходимо решить две системы уравнений: первую с матрицей $E+\tau A_1$ и вторую — с матрицей $E+\tau A_2$. Таким образом, метод (2), (3) целесообразно применять лишь тогда, когда матрицы $E+\tau A_\alpha$, $\alpha=1, 2$, гораздо легче обратить, чем исходную матрицу A. Например, в случае разностных аппроксимаций уравнений эллиптического типа системы (4), (5) можно решить последовательным применением одномерных прогонок сначала по направлению x_1 (для системы (4)) и затем — по направлению x_2 (для системы (5)).

Обратимся к исследованию сходимости итерационного метода (2), (3). Будем рассматривать систему (1) как операторное уравнение в конечномерном линейном пространстве H со скалярным произведением (y, v) и нормой $\|y\|=\sqrt{(y, y)}$. Определим погрешности $z_{k+1/2}$, z_{k+1} метода как разности

$$z_{k+1/2}=y_{k+1/2}-y, \quad z_{k+1}=y_{k+1}-y$$

между решениями $y_{k+1/2}$, y_{k+1} систем (2), (3) и решением y исходной системы (1). Введенные погрешности удовлетворяют уравнениям

$$(E+\tau A_1) z_{k+1/2}=(E-\tau A_2) z_k, \tag{6}$$

$$(E+\tau A_2) z_{k+1}=(E-\tau A_1) z_{k+1/2}, \tag{7}$$

из которых можно легко исключить промежуточное значение $z_{k+1/2}$ и получить уравнение, связывающее только z_k и z_{k+1}:

$$(E+\tau A_1)(E+\tau A_2) z_{k+1}=(E-\tau A_1)(E-\tau A_2) z_k. \tag{8}$$

Теорема 1. *Пусть* $A=A_1+A_2$, *где* $A_\alpha=A_\alpha^*>0$, $\alpha=1, 2$, $A_1A_2=A_2A_1$. *Тогда итерационный метод* (2), (3) *сходится при любом* $\tau>0$. *Если*

$$0<\delta E\leqslant A_\alpha\leqslant\Delta E, \quad \alpha=1, 2, \tag{9}$$

то при

$$\tau=1/\sqrt{\delta\Delta} \tag{10}$$

для погрешности справедлива оценка

$$\|y_k-y\|\leqslant\rho_0^k\|y_0-y\|, \tag{11}$$

где

$$\rho_0=\left(\frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}\right)^2, \quad \xi=\frac{\delta}{\Delta}. \tag{12}$$

Доказательство. Запишем уравнение для погрешности в виде

$$z_{k+1}=Sz_k,$$

где

$$S=(E+\tau A_2)^{-1}(E+\tau A_1)^{-1}(E-\tau A_1)(E-\tau A_2). \tag{13}$$

Оператор S является самосопряженным, так как по условию теоремы A_1 и A_2 — самосопряженные перестановочные операторы. Получим оценку для собственных чисел $\lambda_k(S)$, $k=1, 2, \ldots, m$, оператора (13). Любое собственное число можно представить в виде

$$\lambda_k(S)=\frac{(1-\tau\lambda_{k_1}(A_1))(1-\tau\lambda_{k_2}(A_2))}{(1+\tau\lambda_{k_1}(A_1))(1+\tau\lambda_{k_2}(A_2))}, \tag{14}$$

где $\lambda_{k_\alpha}(A_\alpha)$ — собственные числа операторов A_α, $\alpha=1, 2$, $k_\alpha=1, 2, \ldots, m$. Из (14) видно, что при $\tau>0$ все собственные числа

$\lambda_k(S)$ не превосходят по модулю единицу. Следовательно,

$$\|S\| = \max_{1 \leqslant k \leqslant m} |\lambda_k(S)| < 1$$

и метод (2), (3) сходится.

Далее, согласно (14) имеем

$$|\lambda_k(S)| \leqslant \left| \frac{1-\tau\lambda_{k_1}(A_1)}{1+\tau\lambda_{k_1}(A_1)} \right| \left| \frac{1-\tau\lambda_{k_2}(A_2)}{1+\tau\lambda_{k_2}(A_2)} \right|. \tag{15}$$

Из условия (9) получим

$$\delta \leqslant \lambda_{k_\alpha}(A_\alpha) \leqslant \Delta, \qquad \alpha = 1, 2,$$

следовательно,

$$\left| \frac{1-\tau\lambda_{k_\alpha}(A_\alpha)}{1+\tau\lambda_{k_\alpha}(A_\alpha)} \right| \leqslant \max_{\delta \leqslant \lambda_{k_\alpha} \leqslant \Delta} \left\{ \left| \frac{1-\tau\delta}{1+\tau\delta} \right|, \left| \frac{1-\tau\Delta}{1+\tau\Delta} \right| \right\}. \tag{16}$$

Если выбрать τ согласно (10), то получим

$$-\frac{1-\tau\delta}{1+\tau\delta} = \frac{1-\tau\Delta}{1+\tau\Delta} = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}},$$

и поэтому

$$\left| \frac{1-\tau\lambda_{k_\alpha}(A_\alpha)}{1+\tau\lambda_{k_\alpha}(A_\alpha)} \right| \leqslant \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \qquad \alpha = 1, 2.$$

Отсюда и из (15), (16), получаем

$$\|S\| \leqslant \left(\frac{1-\sqrt{\xi}}{1+\sqrt{\xi}} \right)^2 = \rho_0,$$

так что $\|z_{k+1}\| \leqslant \rho_0 \|z_k\| \leqslant \rho_0^k \|z_0\|$. Теорема 1 доказана.

2. Пример. Рассмотрим применение метода к модельной задаче (14) из § 3. В данном случае

$$(A_\alpha y)_{ij} = -y_{\overline{x}_\alpha x_\alpha, ij}, \qquad \alpha = 1, 2, \tag{17}$$

$$i, j = 1, 2, \ldots, N-1, \quad hN = 1,$$

и метод переменных направлений (4), (5) принимает вид

$$y_{ij}^{(k+1/2)} - \tau y_{\overline{x}_1 x_1, ij}^{(k+1/2)} = F_{ij}^{(k)}, \tag{18}$$

$$y_{ij}^{(k+1)} - \tau y_{\overline{x}_2 x_2, ij}^{(k+1)} = \Phi_{ij}^{(k)}, \tag{19}$$

где

$$F_{ij}^{(k)} = y_{ij}^{(k)} + \tau y_{\overline{x}_2 x_2, ij}^{(k)} + \tau f_{ij},$$

$$\Phi_{ij}^{(k)} = y_{ij}^{(k+1/2)} + \tau y_{\overline{x}_1 x_1, ij}^{(k)} + \tau f_{ij}.$$

Уравнение (18) решается при каждом фиксированном $j=1, 2, \ldots, N-1$ с помощью метода прогонки по направлению x_1. Для этого достаточно записать (18) в виде

$$Ay_{i-1\,j}^{(k+1/2)} - Cy_{ij}^{(k+1/2)} + By_{i+1\,j}^{(k+1/2)} = -F_{ij}^{(k)},$$
$$i=1, 2, \ldots, N-1, \quad y_{0j}^{(k+1/2)}=0, \quad y_{Nj}^{(k+1/2)}=0,$$

где $A=B=\tau/h^2$, $C=1+2\tau/h^2$, и применить формулы прогонки (43), (44) из п. 7 § 4 ч. I. Точно так же уравнение (19) записывается в виде

$$Ay_{i,j-1}^{(k+1)} - Cy_{ij}^{(k+1)} + By_{i,j+1}^{(k+1)} = -\Phi_{ij}^{(k)},$$
$$j=1, 2, \ldots, N-1, \quad y_{i0}^{(k+1)}=0, \quad y_{iN}^{(k+1)}=0$$

и при каждом фиксированном $i=1, 2, \ldots, N-1$ решается прогонкой по направлению x_2.

Применим теорему 1 к исследованию сходимости метода (18), (19). Операторы A_1 и A_2, определенные согласно (17), перестановочны, так как разностное выражение

$$(A_2A_1y)_{ij} = y_{\overline{x}_1x_1\overline{x}_2x_2,\,ij}$$

определено во всех внутренних точках сетки Ω, причем

$$(A_1A_2y)_{ij} = y_{\overline{x}_2x_2\overline{x}_1x_1,\,ij} = (A_2A_1y)_{ij}.$$

Таким образом, перестановочность является следствием того, что A_1 и A_2 — операторы с постоянными коэффициентами и область G — прямоугольная. Нарушение хотя бы одного из этих условий приводит, как правило, к нарушению перестановочности. Предположение о перестановочности является, видимо, основным ограничением теоремы 1, не позволяющим применить ее к более общим разностным аппроксимациям уравнений эллиптического типа.

В [32, с. 450] доказана сходимость метода переменных направлений без требования перестановочности. При условиях (9), (10) доказана оценка

$$\|(E+\tau A_2)(y_k-y)\| \leqslant \rho_0^k \|(E+\tau A_2)(y_0-y)\|,$$

где константа ρ_0 определена согласно (12).

Как мы видели ранее (см. § 1,2 гл. 3), операторы A_α, $\alpha=1, 2$, являются самосопряженными в смысле скалярного произведения

$$(y, v) = \sum_{i,\,j=1}^{N-1} y_{ij}v_{ij}h^2,$$

причем для них выполнены операторные неравенства (9), где

$$\delta = \frac{4}{h^2}\sin^2\frac{\pi h}{2}, \quad \Delta = \frac{4}{h^2}\cos^2\frac{\pi h}{2}.$$

Таким образом, при

$$\tau = \frac{h^2}{2\sin(\pi h)} \approx \frac{h}{2\pi}$$

для погрешности будет справедлива оценка (11), (12), где

$$\xi = \operatorname{tg}^2 \frac{\pi h}{2} \approx \frac{\pi^2 h^2}{4}.$$

Число итераций $n_0(\varepsilon)$, необходимых для достижения заданной точности ε, в данном случае равно

$$n_0(\varepsilon) \approx \frac{\ln(1/\varepsilon)}{4\sqrt{\xi}} \approx \frac{\ln(1/\varepsilon)}{2\pi h}.$$

Это примерно столько же итераций, сколько и в попеременно-треугольном итерационном методе. Однако число арифметических действий на каждой итерации здесь больше.

3. Случай прямоугольной области. В теореме 1 предполагалось, что спектр операторов A_1 и A_2 расположен на одном и том же отрезке $[\delta, \Delta]$. Рассмотрим теперь случай, когда вместо (9) выполняются условия

$$0 < \delta_\alpha E \leqslant A_\alpha \leqslant \Delta_\alpha E, \quad \alpha = 1, 2, \tag{20}$$

где δ_α и Δ_α — заданные положительные числа (под δ_α и Δ_α можно понимать наименьшее и, соответственно, наибольшее собственное значение оператора A_α). Типичным примером является разностная аппроксимация уравнения Пуассона на прямоугольной сетке. В случае (20) используется двухпараметрический итерационный метод переменных направлений

$$\frac{y_{k+1/2} - y_k}{\tau_1} + A_1 y_{k+1/2} + A_2 y_k = f, \tag{21}$$

$$\frac{y_{k+1} - y_{k+1/2}}{\tau_2} + A_1 y_{k+1/2} + A_2 y_{k+1} = f, \tag{22}$$

где, вообще говоря, $\tau_1 \neq \tau_2$. Как и в теореме 1, можно доказать, что данный метод сходится при любых $\tau_1 > 0$, $\tau_2 > 0$.

Оптимальные параметры τ_1 и τ_2 задаются следующим образом. Пусть известны константы δ_1, δ_2, Δ_1, Δ_2 в неравенствах (20). Вычислим величины

$$t = \left(\frac{(\Delta_1 - \delta_1)(\Delta_2 - \delta_2)}{(\Delta_1 + \delta_2)(\Delta_2 + \delta_1)}\right)^{1/2}, \quad \varkappa = \left(\frac{\Delta_1 - \delta_1}{\Delta_2 + \delta_1}\right)\frac{\Delta_2}{\Delta_1} \tag{23}$$

и найдем

$$p = \frac{\varkappa - t}{\varkappa + t}, \quad r = \frac{\Delta_1 - \Delta_2 + (\Delta_1 + \Delta_2)p}{2\Delta_1\Delta_2}, \quad q = r + \frac{1-p}{\Delta_1}. \tag{24}$$

Нетрудно проверить, что $0 < t < 1$. Далее, определим

$$\omega = \left(\frac{1+t}{1-t}\right)^{1/2} \tag{25}$$

и, наконец, положим

$$\tau_1 = \frac{q\omega + r}{1 + p\omega}, \quad \tau_2 = \frac{q\omega - r}{1 - p\omega}. \tag{26}$$

Имеет место следующий аналог теоремы 1.

Теорема 2. *Пусть* $A=A_1+A_2$, *где* A_1 *и* A_2 — *перестановочные самосопряженные операторы, удовлетворяющие условиям* (20). *Тогда для погрешности итерационного метода* (21)—(26) *справедлива оценка* (11), *где*

$$\rho_0=\left(\frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}\right)^2, \qquad \xi=\frac{1-t}{1+t}. \tag{27}$$

Доказательство. Запишем уравнение для погрешности метода (21), (22) в виде

$$z_{k+1}=Sz_k, \tag{28}$$

где

$$S=(E+\tau_2A_2)^{-1}(E+\tau_1A_1)^{-1}(E-\tau_2A_1)(E-\tau_1A_2).$$

Оценим собственные числа

$$\lambda_k(S)=\left(\frac{2-\tau_2\lambda_{k_1}(A_1)}{1+\tau_1\lambda_{k_1}(A_1)}\right)\left(\frac{1-\tau_1\lambda_{k_2}(A_2)}{1+\tau_2\lambda_{k_2}(A_2)}\right) \tag{29}$$

оператора S. Для этого сделаем в (29) замену

$$\lambda_{k_1}(A_1)=\frac{\lambda'_{k_1}-p}{q-r\lambda'_{k_1}}, \quad \lambda_{k_2}(A_2)=\frac{\lambda'_{k_2}+p}{q+r\lambda'_{k_2}} \tag{30}$$

с не определенными пока параметрами p, q, r. Тогда получим

$$\lambda_k(S)=\left(\frac{1-\omega_2\lambda'_{k_1}}{1+\omega_1\lambda'_{k_1}}\right)\left(\frac{1-\omega_1\lambda'_{k_2}}{1+\omega_2\lambda'_{k_2}}\right), \tag{31}$$

где

$$\omega_1=\frac{\tau_1-r}{q-\tau_1p}, \quad \omega_2=\frac{\tau_2+r}{q+\tau_2p}.$$

Если выбрать τ_1 и τ_2 согласно (26), то получим, что $\omega_1=\omega_2=\omega$ и

$$\lambda_k(S)=\left(\frac{1-\omega\lambda'_{k_1}}{1+\omega\lambda'_{k_1}}\right)\left(\frac{1-\omega\lambda'_{k_2}}{1+\omega\lambda'_{k_2}}\right). \tag{32}$$

Подберем теперь параметры p, q, r таким образом, чтобы в результате замены (30) оба отрезка

$$\delta_\alpha\leqslant\lambda_{k_\alpha}(A_\alpha)\leqslant\Delta_\alpha, \qquad \alpha=1,2,$$

переходили бы в один и тот же отрезок

$$\delta\leqslant\lambda'_{k_\alpha}\leqslant\Delta, \qquad \alpha=1,2.$$

Для этого достаточно потребовать совпадения граничных точек соответствующих отрезков, т. е. положить

$$\begin{aligned}&\frac{\delta-p}{q-r\delta}=\delta_1, \quad \frac{\delta+p}{q+r\delta}=\delta_2,\\ &\frac{\Delta-p}{q-r\Delta}=\Delta_1, \quad \frac{\Delta+p}{q+r\Delta}=\Delta_2.\end{aligned} \tag{33}$$

Таким образом, приходим к системе четырех уравнений относительно пяти неизвестных p, q, r, δ, Δ. Положим для определенности $\Delta=1$. Тогда после несложных, но громоздких выкладок, которые мы опускаем, получим, что решение системы (33) определяется формулами (23), (24) и

$$\delta=\frac{1-t}{1+t}.$$

Обращаясь к выражению (32) для собственного числа оператора S, видим, что мы пришли к той же задаче, которая возникла при доказательстве теоремы 1, а именно: найти значение ω, которое минимизирует $\|S\|=\max\limits_k|\lambda_k(S)|$ при условии, что $0<\delta\leqslant\lambda_{k_\alpha}\leqslant 1$. Согласно теореме 1 для этого достаточно взять $\omega=\frac{1}{\sqrt{\delta}}=\sqrt{\frac{1+t}{1-t}}$ и тогда получим $\|S\|\leqslant\rho_0$, где $\rho_0=\left(\frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}\right)^2$, $\xi=\delta=\frac{1-t}{1+t}$. Этим и завершается доказательство теоремы 2.

Пример. Рассмотрим разностную аппроксимацию уравнения Пуассона в прямоугольнике G с границей Γ на прямоугольной сетке с шагами h_1 и h_2:

$$\frac{y_{i+1,j}-2y_{ij}+y_{i-1,j}}{h_1^2}+\frac{y_{i,j+1}-2y_{ij}+y_{i,j-1}}{h_2^2}=-f_{ij}, \tag{34}$$

$$i=1,2,\ldots,N_1-1,\quad j=1,2,\ldots,N_2-1,$$

$$h_1N_1=l_1,\quad h_2N_2=l_2,\quad y(x_{ij})=0,\quad x_{ij}\in\Gamma.$$

В данном случае $A=A_1+A_2$, где

$$(A_\alpha y)_{ij}=-y_{\bar{x}_\alpha x_\alpha,ij},\qquad \alpha=1,2.$$

Операторы A_1 и A_2 перестановочны. Они являются самосопряженными и положительно определенными операторами в смысле скалярного произведения

$$(y,v)=\sum_{i=1}^{N_1-1}h_1\sum_{j=1}^{N_2-1}h_2y_{ij}v_{ij}.$$

Как показано в § 1,2 гл. 3, операторы A_α удовлетворяют неравенствам (20), где

$$\delta_\alpha=\frac{4}{h_\alpha^2}\sin^2\frac{\pi h_\alpha}{2l_\alpha},\quad \Delta_\alpha=\frac{4}{h_\alpha^2}\cos^2\frac{\pi h_\alpha}{2l_\alpha},\qquad \alpha=1,2.$$

Таким образом, для решения разностной задачи (34) можно применить итерационный метод (21), (22), в котором итерационные параметры τ_1 и τ_2 выбираются согласно формулам (26). Разумеется, для решения задачи (34) можно использовать итерационный метод (2), (3) с более простым способом выбора итерационного параметра (10), где $\delta=\min(\delta_1,\delta_2)$, $\Delta=\max(\Delta_1,\Delta_2)$. Однако

при этом для достижения той же точности ε потребуется выполнить большее число итераций.

Действительно, согласно теоремам 1 и 2 число итераций при малых ξ пропорционально $\xi^{-1/2}$, где

$$\xi = \xi_1 = \delta/\Delta \quad \text{для метода (2), (3),}$$
$$\xi = \xi_2 = (1-t)/(1+t) \quad \text{для метода (21)—(23).}$$

Из формулы (23) находим

$$t \approx 1 - \frac{1}{2}\left(\frac{\delta_1}{\Delta_1} + \frac{\delta_2}{\Delta_2} + \frac{\delta_1}{\Delta_2} + \frac{\delta_2}{\Delta_1}\right),$$
$$\xi_2 \approx \frac{1}{4}\left(\frac{\delta_1}{\Delta_1} + \frac{\delta_2}{\Delta_2} + \frac{\delta_1}{\Delta_2} + \frac{\delta_2}{\Delta_1}\right).$$

Пусть, для определенности, $\delta_1 < \delta_2$, $\Delta_1 < \Delta_2$. Тогда получим

$$\xi_1 = \delta_1/\Delta_2, \quad \xi_2 = \frac{1}{4}\xi_1'\left(1+\frac{\delta_2}{\delta_1}\right)\left(1+\frac{\Delta_2}{\Delta_1}\right) > \xi_1.$$

Отношение числа итераций $n_0^{(1)}(\varepsilon)$ в методе (2), (3) к числу итераций $n_0^{(2)}(\varepsilon)$ в методе (21)—(23) окажется равным

$$\frac{n_0^{(1)}(\varepsilon)}{n_0^{(2)}(\varepsilon)} \approx \sqrt{\frac{\xi_2}{\xi_1}} = \frac{1}{2}\sqrt{\left(1+\frac{\delta_2}{\delta_1}\right)\left(1+\frac{\Delta_2}{\Delta_1}\right)} > 1.$$

Если, например, $l_2 = 0{,}5\, l_1$, $h_2 = 0{,}5\, h_1$, то получим $\delta_2 = 4\delta_1$, $\Delta_2 = 4\Delta_1$, и, следовательно, $n_0^{(1)}(\varepsilon)/n_0^{(2)}(\varepsilon) \approx 2{,}5$.

Замечание. Существенного ускорения сходимости можно добиться путем использования итерационного метода

$$\frac{y_{k+1/2} - y_k}{\tau_1^{(k+1)}} + A_1 y_{k+1/2} + A_2 y_k = f,$$
$$\frac{y_{k+1} - y_{k+1/2}}{\tau_2^{(k+1)}} + A_1 y_{k+1/2} + A_2 y_{k+1} = f$$

с переменными параметрами $\tau_1^{(k+1)}$, $\tau_2^{(k+1)}$, $k = 0, 1, \ldots, n-1$. Способ выбора итерационных параметров и оценки погрешности подробно изложены в [30, с. 463]. Отметим лишь, что в случае модельной задачи число итераций, необходимых для достижения заданной точности ε, является величиной $O\left(\ln\frac{1}{h}\right)$.

§ 5. Метод матричной прогонки

1. Введение. Матричная прогонка относится к прямым методам решения разностных уравнений. Она применяется к уравнениям, которые можно записать в виде системы векторных уравнений

$$\begin{aligned} -C_0 y_0 + B_0 y_1 &= -F_0, \\ A_i y_{i-1} - C_i y_i + B_i y_{i+1} &= -F_i, \quad i = 1, 2, \ldots, N-1, \\ A_N y_{N-1} - C_N y_N &= -F_N, \end{aligned} \tag{1}$$

где y_i — искомые векторы размерности M, F_i — заданные векторы, A_i, B_i, C_i — заданные квадратные матрицы порядка M.

Матричная прогонка представляет собой обобщение обычной прогонки на случай системы векторных уравнений (1). По сравнению с другими прямыми методами решения разностных задач матричная прогонка более универсальна, так как позволяет решать уравнения с переменными коэффициентами и не накладывает сильных ограничений на вид граничных условий. Однако применение матричной прогонки к решению двумерных разностных задач сталкивается с двумя трудностями: неэкономичностью по числу действий (т. е. большое время счета) и, главным образом, необходимостью в больших ресурсах машинной памяти. Если же матрицы A_i, B_i, C_i имеют относительно невысокий порядок (как это бывает при аппроксимации систем одномерных дифференциальных уравнений), то матричная прогонка ничем не хуже обычной прогонки.

Прежде чем излагать алгоритм, покажем на простом примере, каким образом двумерную разностную задачу можно привести к виду (1).

2. Запись разностного уравнения Пуассона в виде системы векторных уравнений. Пусть в прямоугольнике

$$G=\{0<x_\alpha<l_\alpha,\ \alpha=1,\ 2\}$$

с границей Г требуется найти решение уравнения Пуассона

$$\frac{\partial^2 u}{\partial x_1^2}+\frac{\partial^2 u}{\partial x_2^2}=-f(x_1, x_2), \tag{2}$$

удовлетворяющее условию Дирихле

$$u(x_1, x_2)=\mu(x_1, x_2),\ \text{если}\ (x_1, x_2)\in\Gamma. \tag{3}$$

Введем прямоугольную сетку

$$G_h=\{x_{ij}=(x_1^{(i)}, x_2^{(j)})\},$$

где $x_1^{(i)}=ih_1$, $x_2^{(j)}=jh_2$, $i=0,\ 1,\ \ldots,\ N_1$, $j=0,\ 1,\ \ldots,\ N_2$, $h_1N_1=l_1$, $h_2N_2=l_2$ и заменим задачу (2), (3) разностной схемой

$$\frac{y_{i-1,j}-2y_{ij}+y_{i+1,j}}{h_1^2}+\frac{y_{i,j-1}-2y_{ij}+y_{i,j+1}}{h_2^2}=-f_{ij}, \tag{4}$$

$$\begin{gathered} i=1,\ 2,\ \ldots,\ N_1-1,\quad j=1,\ 2,\ \ldots,\ N_2-1,\\ y_{0j}=\mu_{0j},\quad y_{N_1 j}=\mu_{N_1 j},\quad i=1,\ 2,\ \ldots,\ N_2-1,\\ y_{i0}=\mu_{i0},\quad y_{iN_2}=\mu_{iN_2},\quad i=1,\ 2,\ \ldots,\ N_1-1. \end{gathered} \tag{5}$$

Разностная схема (4), (5) представляет собой систему линейных алгебраических уравнений, в которой неизвестными являются значения y_{ij}, $i=1,\ 2,\ \ldots,\ N_1-1$, $j=1,\ 2,\ \ldots,\ N_2-1$. Число неизвестных равно числу уравнений, т. е. $(N_1-1)(N_2-1)$. Запишем систему (4), (5) в векторном виде (1).

При решении системы (4), (5) матричную прогонку можно проводить как по индексу i, так и по индексу j. Покажем, например, как подготовить систему (4), (5) к виду (1), удобному для применения прогонки по индексу i. Перепишем систему (4) в виде

$$\frac{y_{i-1,j}}{h_1^2} - \left(\frac{2y_{ij}}{h_1^2} - \frac{y_{i,j-1} - 2y_{ij} + y_{i,j+1}}{h_2^2}\right) + \frac{y_{i+1,j}}{h_1^2} = -f_{ij},$$

$$i = 1, 2, \ldots, N_1 - 1, \quad j = 1, 2, \ldots, N_2 - 1$$

и учтем граничные условия

$$y_{i0} = \mu_{i0}, \quad y_{iN_2} = \mu_{iN_2}, \quad i = 1, 2, \ldots, N_1 - 1.$$

Тогда получим систему уравнений

$$\frac{y_{i-1,1}}{h_1^2} - \left(\frac{2y_{i1}}{h_1^2} - \frac{-2y_{i1} + y_{i2}}{h_2^2}\right) + \frac{y_{i+1,1}}{h_1^2} = -f_{i1} - \frac{\mu_{i0}}{h_2^2},$$

$$\frac{y_{i-1,j}}{h_1^2} - \left(\frac{2y_{ij}}{h_1^2} - \frac{y_{i,j-1} - 2y_{ij} + y_{i,j+1}}{h_2^2}\right) + \frac{y_{i+1,j}}{h_1^2} = -f_{ij},$$

$$j = 2, 3, \ldots, N_2 - 2,$$

$$\frac{y_{i-1,N_2-1}}{h_1^2} - \left(\frac{2y_{i,N_2-1}}{h_1^2} - \frac{y_{i,N_2-2} - 2y_{i,N_2-1}}{h_2^2}\right) + \frac{y_{i+1,N_2-1}}{h_1^2} = -f_{i,N_2-1} - \frac{\mu_{iN_2}}{h_2^2},$$

где $i = 1, 2, \ldots, N_1 - 1$. Далее, обозначим через E_2 единичную матрицу порядка $N_2 - 1$ и через Λ_2 — следующую трехдиагональную матрицу того же порядка

$$\Lambda_2 = \frac{1}{h_2^2}\begin{bmatrix} -2 & 1 & 0 & 0 & \cdots & \cdot & \cdot \\ 1 & -2 & 1 & 0 & \cdots & \cdot & \cdot \\ 0 & 1 & -2 & 0 & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 & -2 & 1 \\ \cdot & \cdot & \cdot & \cdot & 0 & 1 & -2 \end{bmatrix}. \tag{6}$$

Ясно, что Λ_2 представляет собой матрицу оператора второй разностной производной по направлению x_2.

Введем для $i = 1, 2, \ldots, N_1 - 1$ векторы

$$y_i = (y_{i1}, y_{i2}, \ldots, y_{i,N_2-1})^T, \tag{7}$$

$$F_i = \left(f_{i1} + \frac{\mu_{i0}}{h_2^2}, f_{i2}, f_{i3}, \ldots, f_{i,N_2-2}, f_{i,N_2-1} + \frac{\mu_{iN_2}}{h_2^2}\right)^T. \tag{8}$$

Тогда предыдущую систему уравнений можно записать в векторном виде

$$\frac{1}{h_1^2} E_2 y_{i-1} - \left(\frac{2}{h_1^2} E_2 - \Lambda_2\right) y_i + \frac{1}{h_1^2} E_2 y_{i+1} = -F_i,$$

$$i = 1, 2, \ldots, N_1 - 1. \tag{9}$$

Эту систему уравнений следует дополнить граничными условиями

$$y_0 = \mu_0, \quad y_{N_1} = \mu_{N_1},$$

где

$$\mu_0=(\mu_{01}, \mu_{02}, \ldots, \mu_{0,N_2-1})^T, \quad \mu_{N_1}=(\mu_{N_1 1}\mu_{N_1 2}\ldots\mu_{N_1,N_2-1})^T.$$

Таким образом, разностная схема (4), (5) записывается в векторном виде (1), где B_0 и A_N— нулевые матрицы, $A_i=B_i=h_1^{-2}E_2$, $C_i=2h_1^{-2}E_2-\Lambda_2$, $i=1, 2, \ldots, N_1-1$.

Может оказаться, что $N_2\gg N_1$, т. е. что число точек сетки по направлению x_2 гораздо больше числа точек по направлению x_1 (например в случае, когда прямоугольник G сильно растянут в направлении x_2). Тогда выгоднее пользоваться прогонкой по индексу j, так как при этом соответствующие матричные коэффициенты будут иметь порядок N_1-1 гораздо меньший, чем N_2-1. Соответствующая система векторных уравнений имеет вид

$$\frac{1}{h_2^2}E_1 y_{j-1}-\left(\frac{2}{h_2^2}E_1-\Lambda_1\right)y_j+\frac{1}{h_2^2}E_1 y_{j+1}=-F_j,$$

$$j=1, 2, \ldots, N_2-1, \quad y_0=\mu_0, \quad y_{N_2}=\mu_{N_2},$$

где E_1— единичная матрица порядка N_1-1, Λ_1— матрица, аналогичная (6) и имеющая порядок N_1-1,

$$y_j=(y_{1j}, y_{2j}, \ldots, y_{N_1-1,j})^T, \quad \mu_0=(\mu_{10}, \mu_{20}, \ldots, \mu_{N_1-1,0})^T,$$

$$\mu_{N_2}=(\mu_1 N_2 \mu_2 N_2 \ldots \mu_{N_1-1,N_2})^T.$$

3. Алгоритм матричной прогонки. Пусть задана система уравнений (1). Формулы матричной прогонки можно получить так же, как и формулы обычной прогонки (см. п. 7 § 4 ч. I), однако при их выводе надо учитывать, что коэффициенты уравнения (1) неперестановочны. Будем искать решение системы (1) в виде

$$y_i=\alpha_{i+1}y_{i+1}+\beta_{i+1}, \quad i=0, 1, \ldots, N-1, \tag{10}$$

где α_{i+1}— квадратные матрицы того же порядка M, что и порядок матриц A_i, B_i, C_i, а β_{i+1}— вектор размерности M. Подставляя $y_i=\alpha_{i+1}y_{i+1}+\beta_{i+1}$ и $y_{i-1}=\alpha_i y_i+\beta_i=\alpha_i\alpha_{i+1}y_{i+1}+(\alpha_i\beta_{i+1}+\beta_i)$ в уравнение

$$A_i y_{i-1}-C_i y_i+B_i y_{i+1}=-F_i,$$

получаем, что это уравнение будет выполнено, если потребовать

$$(A_i\alpha_i-C_i)\alpha_{i+1}+B_{i+1}=0,$$

$$(A_i\alpha_i-C_i)\beta_{i+1}=-(A_i\beta_i+F_i).$$

Отсюда приходим к следующим рекуррентным соотношениям для определения матриц α_{i+1} и векторов β_{i+1}:

$$\alpha_{i+1}=(C_i-A_i\alpha_i)^{-1}B_i, \tag{11}$$

$$\beta_{i+1}=(C_i-A_i\alpha_i)^{-1}(A_i\beta_i+F_i). \tag{12}$$

Здесь $i=1, 2, \ldots, N-1$. Начальные значения α_1 и β_1 задаются в соответствии с уравнением

$$-C_0y_0+B_0y_1=-F_0,$$

которое можно переписать в виде

$$y_0 = C_0^{-1}B_0 y_1 + C_0^{-1}F_0. \tag{13}$$

Сопоставляя (13) с уравнением (10) при $i=0$, получаем

$$\alpha_1 = C_0^{-1}B_0, \quad \beta_1 = C_0^{-1}F_0. \tag{14}$$

После того как все коэффициенты α_i, β_i найдены, векторы y_i, $i=N-1, N-2, \ldots, 1, 0$, определяются последовательно из уравнения (10), начиная с y_{N-1}. Для начала счета надо знать вектор y_N, который определяется из системы двух уравнений

$$A_N y_{N-1} - C_N y_N = -F_N, \quad y_{N-1} = \alpha_N y_N + \beta_N.$$

Отсюда получаем

$$y_N = (C_N - A_N\alpha_N)^{-1}(A_N\beta_N + F_N). \tag{15}$$

Объединяя формулы (10)—(12), (14), (15), приходим к следующему алгоритму матричной прогонки для системы (1):

$$\alpha_{i+1} = (C_i - A_i\alpha_i)^{-1}B_i, \quad i = 1, 2, \ldots, N-1, \quad \alpha_1 = C_0^{-1}B_0, \tag{16}$$

$$\beta_{i+1} = (C_i - A_i\alpha_i)^{-1}(A_i\beta_i + F_i), \quad i = 1, 2, \ldots, N, \quad \beta_1 = C_0^{-1}F_0, \tag{17}$$

$$y_i = \alpha_{i+1}y_{i+1} + \beta_{i+1}, \quad i = N-1, N-2, \ldots, 1, 0, \quad y_N = \beta_{N+1}. \tag{18}$$

При реализации метода матричной прогонки приходится запоминать все матрицы α_i, β_i, $i=1, 2, \ldots, N-1$, что ведет в случае матриц больших размеров к необходимости использования внешней памяти ЭВМ и тем самым к увеличению времени счета.

Кроме того, реализация формул (16) сама по себе требует большого числа действий. В каждой точке i приходится один раз обратить матрицу и сделать два умножения матриц порядка M, что требует $O(M^3)$ арифметических действий. Следовательно, для вычисления всех коэффициентов α_i, $i=1, 2, \ldots, N-1$, требуется $O(M^3N)$ действий. Для модельной задачи, когда $M=N=h^{-1}$, число действий становится величиной $O(h^{-4})$. По указанным причинам (большой объем памяти и значительное число арифметических действий) матричную прогонку сравнительно редко применяют для решения задач математической физики. Однако в тех случаях, когда матрицы A_i, B_i, C_i невысокого порядка (небольшое число точек по направлению x_2), необходимый объем памяти и число действий резко сокращаются и метод можно рекомендовать для практического использования.

4. Устойчивость матричной прогонки. Так же, как и в случае обычной прогонки, возникает вопрос о численной устойчивости метода матричной прогонки. Получим здесь достаточные условия устойчивости в виде требований, предъявляемых к матрицам A_i, B_i, C_i, $i=0, 1, \ldots, N$.

Пусть в системе (1) y_i и F_i — векторы размерности M, A_i, B_i, C_i — квадратные матрицы порядка M (векторы и матрицы могут быть как вещественными, так и комплексными). Будем рассматривать

матрицы A_i, B_i, C_i как линейные операторы, действующие в M-мерном линейном пространстве H (вещественном или комплексном). Предположим, что в H определены нормы вектора $\|\cdot\|$ и подчиненная ей норма матрицы. При доказательстве устойчивости прогонки нам потребуется следующее известное утверждение.

Лемма 1. *Если для данной матрицы A существует константа $\gamma>0$ такая, что для любого $x\in H$ выполнено неравенство*

$$\|Ax\|\geqslant\gamma\|x\|, \quad \gamma>0, \tag{19}$$

то матрица A имеет обратную, причем $\|A^{-1}\|\leqslant\gamma^{-1}$.

Доказательство. Покажем сначала, что все собственные числа матрицы A отличны от нуля и, следовательно, существует A^{-1}. Пусть λ — любое собственное число матрицы A и z — отвечающий ему собственный вектор, т. е. $Az=\lambda z$. Согласно условию (19) имеем

$$\|Az\|=|\lambda|\,\|z\|\geqslant\gamma\|z\|,$$

т. е. $|\lambda|\geqslant\gamma>0$, и тем самым $\lambda\neq0$.

Таким образом, матрица A имеет обратную. Пусть $y\in H$ — любой вектор. Обозначая $x=A^{-1}y$, получим из условия (19), что $\|A^{-1}y\|\leqslant\gamma^{-1}\|y\|$. Следовательно, $\|A^{-1}\|\leqslant\gamma^{-1}$, что и требовалось.

Метод прогонки (16)—(18) будем называть устойчивым, если матрицы $C_i-A_i\alpha_i$ имеют обратные и $\|\alpha_i\|\leqslant1$, $i=1, 2, \ldots, N$.

Из устойчивости прогонки следует однозначная разрешимость системы (1). Действительно, в этом случае, исходя из рекуррентных формул (18), можно представить решение задачи (1) в явной форме в виде конечной суммы с коэффициентами, зависящими от α_i, β_i.

Условия $\|\alpha_i\|\leqslant1$ обеспечивают численную устойчивость счета по формуле (18). Нарушение этих условий не всегда приводит к сильному накоплению погрешности. Однако подробный анализ вычислительной погрешности метода прогонки выходит за рамки данной книги.

Сформулируем теперь теорему об устойчивости матричной прогонки.

Теорема 1. *Пусть A_i, B_i — ненулевые матрицы, $i=1, 2, \ldots, N-1$, и пусть существуют матрицы C_i^{-1}, $i=0, 1, \ldots, N$. Если выполнены неравенства*

$$\|C_i^{-1}A_i\|+\|C_i^{-1}B_i\|\leqslant1, \quad i=1, 2, \ldots, N-1, \tag{20}$$

$$\|C_0^{-1}B_0\|\leqslant1, \quad \|C_N^{-1}A_N\|<1, \tag{21}$$

то матричная прогонка устойчива.

Доказательство. Докажем по индукции, что $\|\alpha_i\|\leqslant1$ и матрицы $C_i-A_i\alpha_i$ имеют обратные, $i=1, 2, \ldots, N$. Неравенство $\|\alpha_1\|\leqslant1$ выполнено в силу первого из условий (21). Предположим, что $\|\alpha_i\|\leqslant1$ для некоторого $i\geqslant1$. Докажем, что тогда $(C_i-A_i\alpha_i)^{-1}$ существует и $\|\alpha_{i+1}\|\leqslant1$. Поскольку $C_i-A_i\alpha_i=C_i(E-C_i^{-1}A_i\alpha_i)$, достаточно доказать существование матрицы $(E-C_i^{-1}A_i\alpha_i)^{-1}$.

Пусть $x \in H$ — любой вектор. Тогда

$$\|(E - C_i^{-1}A_i\alpha_i)x\| \geqslant \|x\| - \|C_i^{-1}A_i\alpha_i x\| \geqslant$$
$$\geqslant \|x\| - \|C_i^{-1}A_i\|\,\|\alpha_i\|\,\|x\| \geqslant (1 - \|C_i^{-1}A_i\|)\,\|x\|.$$

Отсюда и из условий $1 - \|C_i^{-1}A_i\| \geqslant \|C_i^{-1}B_i\|$ (см. (20)) получим

$$\|(E - C_i^{-1}A_i\alpha_i)x\| \geqslant \gamma_i\|x\|, \quad i = 1, 2, \ldots, N-1, \tag{22}$$

где $\gamma_i = \|C_i^{-1}B_i\| > 0$. Неравенство $\gamma_i > 0$ следует из того, что C_i^{-1} — невырожденная матрица и $B_i \neq 0$, и поэтому $C_i^{-1}B_i$ — ненулевая матрица. Из неравенств (22) и леммы 1 следует существование матриц, обратных к $C_i - A_i\alpha_i$, $i = 1, 2, \ldots, N-1$, и оценки

$$\|(E - C_i^{-1}A_i\alpha_i)^{-1}\| \leqslant \|C_i^{-1}B_i\|^{-1}. \tag{23}$$

Таким образом, матрицы α_{i+1}, заданные рекуррентным соотношением (16), существуют. Перепишем выражение для α_{i+1} в виде

$$\alpha_{i+1} = (E - C_i^{-1}A_i\alpha_i)^{-1}(C_i^{-1}B_i).$$

Отсюда и из оценки (23) получим, что

$$\|\alpha_{i+1}\| \leqslant \|(E - C_i^{-1}A_i\alpha_i)^{-1}\|\,\|C_i^{-1}B_i\| \leqslant 1.$$

Итак, по индукции доказано, что $\|\alpha_i\| \leqslant 1$, $i = 1, 2, \ldots, N$. Для завершения доказательства теоремы 1 осталось убедиться в том, что существует матрица, обратная к $C_N - A_N\alpha_N$. Поскольку $\|\alpha_N\| \leqslant 1$, получим, как и ранее, что

$$\|(E - C_N^{-1}A_N\alpha_N)x\| \geqslant (1 - \|C_N^{-1}A_N\|)\,\|x\|$$

для любого $x \in H$. Следовательно, неравенство (22) выполняется и при $i = N$ с константой $\gamma_N = 1 - \|C_N^{-1}A_N\|$. Неравенство $\gamma_N > 0$ выполнено в силу второго из условий (21). Теорема 1 доказана.

Замечание. Матричная прогонка будет устойчивой и в том случае, если вместо (21) потребовать

$$\|C_0^{-1}B_0\| < 1, \quad \|C_N^{-1}B_N\| \leqslant 1. \tag{24}$$

Доказательство проводится так же, как и в теореме 1. Надо заметить только, что в случае (24) выполняются строгие неравенства $\|\alpha_i\| < 1$, $i = 1, 2, \ldots, N$, и

$$\|(E - C_N^{-1}A_N\alpha_N)x\| \geqslant (1 - \|C_N^{-1}A_N\alpha_N\|)\,\|x\|,$$

причем

$$1 - \|C_N^{-1}A_N\alpha_N\| \geqslant 1 - \|\alpha_N\|\,\|C_N^{-1}A_N\| > 1 - \|C_N^{-1}A_N\| \geqslant 0,$$

т. е. $1 - \|C_N^{-1}A_N\alpha_N\| > 0$.

Применим теорему 1 к исследованию устойчивости метода прогонки для разностного уравнения Пуассона (см. п. 2). В этом

случае система (1) принимает вид (9), причем

$$A_i = B_i = h_1^{-2} E_2, \quad C_i = 2h_1^{-2} E_2 - \Lambda_2,$$
$$i = 1, 2, \ldots, N_1 - 1, \quad B_0 = A_N = 0,$$

где матрица Λ_2 определена согласно (6).

Условия устойчивости прогонки (20) принимают вид

$$\|C_i^{-1}\| \leqslant 0{,}5h_1^2, \quad i = 1, 2, \ldots, N_1 - 1,$$

и будут выполнены, если

$$\|C_i y\| \geqslant \frac{2}{h_1^2} \|y\| \tag{25}$$

для любого вектора y размерности $N_2 - 1$.

Выберем в качестве нормы вектора

$$y = (y_1, y_2, \ldots, y_{N_2-1})^T$$

величину $\|y\| = \sqrt{(y, y)}$, где

$$(u, v) = \sum_{j=1}^{N_2-1} h_2 u_j v_j.$$

Тогда

$$\|C_i y\|^2 = \left(\frac{2}{h_1^2} y - \Lambda_2 y, \ \frac{2}{h_1^2} y - \Lambda_2 y\right) =$$
$$= \frac{4}{h_1^4} \|y\|^2 - \frac{4}{h_1^2} (\Lambda_2 y, y) + \|\Lambda_2 y\|^2 \geqslant \frac{4}{h_1^4} \|y\|^2,$$

поскольку

$$(\Lambda_2 y, y) = -\sum_{j=1}^{N_2} (y_{\overline{x}_2, j})^2 h \leqslant 0.$$

Тем самым условие (25) выполнено и матричная прогонка для системы (4)—(5) устойчива.

§ 6. Метод редукции

1. Вывод основных формул. Метод редукции является прямым методом решения системы разностных уравнений, имеющих вид

$$y_{i-1} - Cy_i + y_{i+1} = -F_i, \ i = 1, 2, \ldots, N-1, \tag{1}$$
$$y_0 = \mu_1, \quad y_N = \mu_2, \tag{2}$$

где y_i — искомые векторы размерности M, F_i, μ_1, μ_2 — заданные векторы и C — заданная квадратная матрица порядка M. Принципиальным отличием данной системы от системы (1) из § 5 является независимость матрицы C от индекса i и равенство коэффициентов при y_{i-1} и y_{i+1}.

В основе метода редукции лежит специальный способ исключения неизвестных из системы (1). Запишем уравнение (1) в точках $i-1$ и $i+1$, т. е.

$$y_{i-2}-Cy_{i-1}+y_i=-F_{i-1}, \quad y_i-Cy_{i+1}+y_{i+2}=-F_{i+1},$$

и сложим эти уравнения. Тогда получим

$$y_{i-2}+2y_i-C(y_{i-1}+y_{i+1})+y_{i+2}=-(F_{i-1}+F_{i+1}),$$

откуда, учитывая, что

$$y_{i-1}+y_{i+1}=Cy_i-F_i,$$

придем к уравнению

$$y_{i-2}-(C^2-2E)y_i+y_{i+2}=-(F_{i-1}+CF_i+F_{i+1}), \qquad (3)$$

связывающему значения искомого вектора в узлах одинаковой четности. В частности, если i — четные, то проведено исключение нечетных узлов. Далее этот процесс исключения можно продолжить

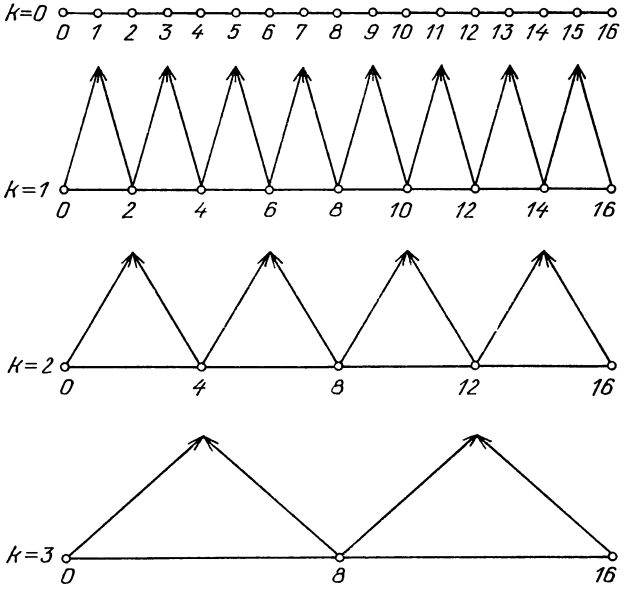


Рис. 16. Порядок исключения неизвестных в методе редукции

аналогичным образом. При этом необходимо предположить, что число узлов N является степенью двойки, $N=2^m$. Прежде чем переходить к случаю произвольного m, рассмотрим для наглядности случай $m=4$, т. е. $N=16$.

Обозначим через k номер этапа исключения неизвестных. При $k=0$ система уравнений совпадает с исходной и содержит значения неизвестных во всех внутренних узлах. На рис. 16 это соответствует верхней горизонтальной черте, где кружочками отмечены номера неизвестных y_i, входящих в систему. На следующем этапе

($k=1$) происходит исключение неизвестных с нечетными номерами, в результате чего получаем систему вида (3), содержащую значения неизвестных только в четных узлах. Этап $k=1$ изображен на рис. 16 второй сверху горизонтальной чертой. Стрелки указывают, какие неизвестные были исключены. На втором этапе ($k=2$) остается каждый четвертый узел и на заключительном этапе ($k=3$) остается только одно уравнение, связывающее y_8, y_0, y_{16}. Поскольку y_0 и y_{16} заданы (см. (2)), из последнего уравнения можно найти y_8. Тем самым начинает осуществляться обратный ход в методе исключения. Зная y_8, можно найти y_4 и y_{12}, далее — все неизвестные с четными номерами и, наконец, все остальные неизвестные.

Вернемся к общему случаю, когда $N=2^m$. Согласно (3), в результате первого этапа исключения ($k=1$) получаем систему уравнений

$$y_{i-2} - C^{(1)} y_i + y_{i+2} = -F_i^{(1)}, \qquad i = 2, 4, 8, \ldots, 2^m - 2, \tag{4}$$

где

$$C^{(1)} = (C^{(0)})^2 - 2E, \quad C^{(0)} = C, \tag{5}$$

$$F_i^{(1)} = F_{i-1} + CF_i + F_{i+1}. \tag{6}$$

По индукции легко доказать, что на k-м этапе исключения, $k=1, 2, \ldots, m$, получаем систему

$$y_{i-2^{k-1}} - C^{(k-1)} y_i + y_{i+2^{k-1}} = -F_i^{(k-1)}, \tag{7}$$

$$i = 2^{k-1}, 3\cdot 2^{k-1}, \ldots, 2^m - 2^{k-1},$$

$$y_0 = \mu_1, \quad y_N = \mu_2,$$

где матрицы $C^{(k-1)}$ и векторы $F_i^{(k-1)}$ находятся из рекуррентных соотношений

$$C^{(k)} = (C^{(k-1)})^2 - 2E, \qquad k = 1, 2, \ldots, m-1,$$

$$C^{(0)} = C, \tag{8}$$

$$F_i^{(k)} = F_{i-2^{k-1}}^{(k-1)} + C^{(k-1)} F_i^{(k-1)} + F_{i+2^{k-1}}^{(k-1)}, \tag{9}$$

$$F_i^{(0)} = F_i, \qquad k = 1, 2, \ldots, m-1,$$

$$i = 2^k, 2\cdot 2^k, 3\cdot 2^k, \ldots, 2^m - 2^k.$$

Таким образом, весь процесс решения состоит из прямого хода и обратного хода. Прямой ход заключается в нахождении матриц $C^{(k)}$ и векторов $F_i^{(k)}$ по формулам (8), (9). Обратный ход состоит в нахождении векторов y_i из системы (7), начиная с $k=m$.

Метод редукции в том виде, как он здесь изложен, не применяется в реальных вычислениях по двум причинам. Во-первых, он неэкономичен из-за того, что на каждом этапе приходится обращать матрицу $C^{(k)}$ общей структуры. Во-вторых, вычисление правых частей по формулам (9) неустойчиво. В следующих пунктах будет показано, как можно устранить указанные недостатки метода редукции.

2. Обращение матриц. Соотношение (8) представляет собой нелинейное разностное уравнение первого порядка для матриц. Его решение можно найти в явном виде.

Рассмотрим сначала числовой аналог уравнения (8), а именно разностное уравнение

$$y_k = y_{k-1}^2 - 2, \quad k = 1, 2, \ldots, \quad y_0 = x, \tag{10}$$

где x — заданное число.

Покажем, что решение $y_k = y_k(x)$ уравнения (10) выражается через многочлен Чебышева, а именно

$$y_k(x) = 2T_{2^k}\left(\frac{x}{2}\right), \quad k = 1, 2, \ldots, \tag{11}$$

где

$$T_n(x) = \begin{cases} \cos(n \arccos x), & \text{если } |x| \leqslant 1, \\ \frac{1}{2}[(x + \sqrt{x^2-1})^n + (x - \sqrt{x^2-1})^n], & \text{если } |x| \geqslant 1. \end{cases}$$

Из выражения для $T_n(x)$ следует, что

$$T_{2n}(x) = 2(T_n(x))^2 - 1.$$

Отсюда при $n = 2^{k-1}$ получаем

$$T_{2^k}\left(\frac{x}{2}\right) = 2\left(T_{2^{k-1}}\left(\frac{x}{2}\right)\right)^2 - 1,$$

т. е. функция (11) удовлетворяет уравнению (10).

Корнями многочлена (11) являются числа

$$x_{l,k} = 2\cos\frac{(2l-1)\pi}{2^{k+1}}, \quad l = 1, 2, \ldots, 2^k.$$

Поэтому функция $y_k(x)$, представляющая собой многочлен относительно x степени 2^k со старшим коэффициентом 1, разлагается в произведение

$$y_k = \prod_{l=1}^{2^k}\left(x - 2\cos\frac{(2l-1)\pi}{2^{k+1}}\right).$$

Приведенные выше выводы имеют место и для матричного уравнения (8). По индукции легко доказывается, что решением $C^{(k)}$ уравнения (8) является многочлен (относительно матрицы C) степени 2^k со старшим коэффициентом 1. Для этого многочлена справедливо представление, аналогичное (11), т. е.

$$C^{(k)} = 2T_{2^k}\left(\frac{C}{2}\right),$$

и справедливо разложение на линейные множители

$$C^{(k)} = \prod_{l=1}^{2^k}\left(C - 2\cos\frac{(2l-1)\pi}{2^{k+1}}E\right). \tag{12}$$

Наличие такого разложения позволяет упростить процедуру обращения матриц. В случае разностных аппроксимаций уравнений эллиптического типа матрица C является трехдиагональной, в то время как $C^{(k)}$ — матрицы общей структуры. Благодаря разложению (12) обращение матрицы $C^{(k)}$ сводится к последовательному обращению трехдиагональных матриц

$$C_{k,l} = C - 2\cos\frac{(2l-1)\pi}{2^{k+1}}E, \quad l = 1, 2, \ldots, 2^k. \tag{13}$$

Действительно, пусть требуется решить уравнение

$$C^{(k)}v = \varphi, \tag{14}$$

где

$$C^{(k)} = \prod_{l=1}^{2^k} C_{k,l}.$$

Рассмотрим для наглядности сначала случай, когда $k=2$. Тогда (14) принимает вид

$$C_{2,1}C_{2,2}C_{2,3}C_{2,4}v = \varphi. \tag{15}$$

Обозначим $v_0=\varphi$, $v_1=C_{2,2}C_{2,3}C_{2,4}v$, $v_2=C_{2,3}C_{2,4}v$, $v_3=C_{2,4}v$, $v_4=v$. Тогда получим, что решение системы (15) сводится к последовательному решению четырех систем уравнений

$$C_{2,1}v_1 = v_0, \quad C_{2,2}v_2 = v_1, \quad C_{2,3}v_3 = v_2, \quad C_{2,4}v_4 = v_3,$$

где $v_0=\varphi$, $v_4=v$.

Точно так же решение системы (14) в общем случае сводится к последовательному решению систем уравнений

$$C_{k,l}v_l = v_{l-1}, \quad l = 1, 2, \ldots, 2^k, \tag{16}$$

где $v_0=\varphi$, $v_{2^k}=v$. Если матрицы $C_{k,l}$ трехдиагональные, то каждую из систем (16) можно решить методом прогонки.

Указанный выше способ обращения матрицы $C^{(k)}$ предполагает определенный порядок выполнения промежуточных этапов: сначала надо обратить матрицу $C_{k,1}$, затем — матрицу $C_{k,2}$ и т. д. Однако, поскольку в матрице $C^{(k)}$ все сомножители перестановочны, можно использовать и другие способы введения промежуточных значений v_l, $l=1, 2, \ldots, 2^k-1$. Так, систему (15) можно записать в виде

$$C_{2,4}C_{2,3}C_{2,2}C_{2,1}v = \varphi$$

и заменить системой уравнений

$$C_{2,4}v_1 = v_0, \quad C_{2,3}v_2 = v_1, \quad C_{2,2}v_3 = v_2, \quad C_{2,1}v_4 = v_3,$$

где $v_0=\varphi$, $v_4=v$. Теоретически при любом порядке выполнения промежуточных этапов мы должны получить одно и то же решение v. Однако, если число промежуточных этапов велико, то при решении систем уравнений (16) на ЭВМ будет происходить накопление погрешностей округления. Рост погрешностей округления зависит от порядка выполнения промежуточных этапов. Здесь имеет место

примерно та же ситуация, что и в итерационном методе с чебышевским набором параметров (см. п. 2 § 6 гл. 2 ч. II). Поэтому при реальном решении системы (14) рекомендуется обращать внимание на порядок выполнения промежуточных этапов. Более подробно этот вопрос рассмотрен в книге [35].

3. Вычисление правых частей. В методе редукции правые части $F_i^{(k-1)}$ уравнения (7) должны удовлетворять соотношению (9). Будем искать $F_i^{(k)}$ в виде

$$F_i^{(k)} = C^{(k)} p_i^{(k)} + q_i^{(k)}, \tag{17}$$

где векторы $p_i^{(k)}$, $q_i^{(k)}$ подлежат определению. Подставляя (17) в (9), получим

$$C^{(k)} p_i^{(k)} + q_i^{(k)} = C^{(k-1)} p_{i-2^k}^{(k-1)} + q_{i-2^k}^{(k-1)} + C^{(k-1)} (C^{(k-1)} p_i^{(k-1)} + q_i^{(k-1)}) + C^{(k-1)} p_{i+2^k}^{(k-1)} + q_{i+2^k}^{(k-1)}.$$

Отсюда, учитывая, что $C^{(k)} = (C^{(k-1)})^2 - 2E$, придем к уравнению

$$(C^{(k-1)})^2 (p_i^{(k)} - p_i^{(k-1)}) + q_i^{(k)} = 2p_i^{(k)} + q_{i-2^k}^{(k-1)} + q_{i+2^k}^{(k-1)} + C^{(k-1)} (p_{i-2^k}^{(k-1)} + p_{i+2^k}^{(k-1)} + q_i^{(k-1)}).$$

Выберем теперь вектор $q_i^{(k)}$ таким, чтобы выполнялось соотношение

$$q_i^{(k)} = 2p_i^{(k)} + q_{i-2^k}^{(k-1)} + q_{i+2^k}^{(k-1)}. \tag{18}$$

Тогда предыдущее уравнение после умножения на $(C^{(k-1)})^{-1}$ примет вид

$$C^{(k-1)} s_i^{(k-1)} = p_{i-2^k}^{(k-1)} + p_{i+2^k}^{(k-1)} + q_i^{(k-1)}, \tag{19}$$

где

$$s_i^{(k-1)} = p_i^{(k)} - p_i^{(k-1)}.$$

Таким образом, вычисление векторов $F_i^{(k)}$ можно заменить нахождением векторов $p_i^{(k)}$, $q_i^{(k)}$ из системы уравнений (17) — (19). Сначала, обращая матрицу $C^{(k-1)}$, находим вектор $s_i^{(k-1)}$, затем вычисляем $p_i^{(k)} = p_i^{(k-1)} + s_i^{(k-1)}$ и затем по формуле (18) вычисляем $q_i^{(k)}$. Правую часть $F_i^{(k-1)}$ уравнения (7) надо заменить согласно (17) выражением

$$F_i^{(k-1)} = C^{(k-1)} p_i^{(k-1)} + q_i^{(k-1)}.$$

Тогда придем к уравнению

$$C^{(k-1)} t_i^{(k-1)} = q_i^{(k-1)} + y_{i-2^k} + y_{i+2^k}, \tag{20}$$

где $t_i^{(k-1)} = y_i - p_i^{(k-1)}$. Из этого уравнения, обращая матрицу $C^{(k-1)}$, находим $t_i^{(k-1)}$ и затем вычисляем $y_i = p_i^{(k-1)} + t_i^{(k-1)}$.

4. Формулировка и обсуждение алгоритма. Сформулируем более детально алгоритм метода редукции.

Вычисления ведутся в цикле по индексу k. Сначала осуществляется прямой ход, т. е. для $k=1, 2, \ldots, m-1$ решаются уравнения

$$C^{(k-1)}s_i^{(k-1)} = p_{i-2^{k-1}}^{(k-1)} + p_{i+2^{k-1}}^{(k-1)} + q_i^{(k-1)} \tag{21}$$

и находятся векторы

$$p_i^{(k)} = p_i^{(k-1)} + s_i^{(k-1)},$$

$$q_i^{(k)} = 2p_i^{(k)} + q_{i-2^{k-1}}^{(k-1)} + q_{i+2^{k-1}}^{(k-1)},$$

где $i=2^k, 2\cdot 2^k, 3\cdot 2^k, \ldots, 2^m-2^k$. Вычисления ведутся, начиная с $k=1$, причем при $k=1$ задаются начальные значения

$$p_i^{(0)}=0, \quad q_i^{(0)}=F_i, \qquad i=1, 2, \ldots, N-1.$$

Решение систем (21) сводится, как это было показано в п. 2, к решению более простых систем уравнений

$$C_{k-1,l}v_{l,i} = v_{l-1,i}, \qquad l=1, 2, \ldots, 2^{k-1}, \tag{22}$$

$$i=2^k, 2\cdot 2^k, \ldots, 2^m-2^k,$$

где

$$v_{0,i} = p_{i-2^{k-1}}^{(k-1)} + p_{i+2^{k-1}}^{(k-1)} + q_i^{(k-1)},$$

$$v_{2^{k-1},i} = s_i^{(k-1)}.$$

Системы (22) решаются при фиксированном l для каждого $i=2^k, 2\cdot 2^k, \ldots, 2^m-2^k$. Поскольку матрица $C_{k-1,l}$ не зависит от i, вычисления целесообразно организовать так, чтобы матрица $C_{k-1,l}$ обращалась один раз.

Подсчитаем число действий, необходимых для решения всех систем вида (22). Предположим, что для решения системы (22) с фиксированными k, l, i требуется q арифметических действий. Тогда при фиксированных k, l для решения систем (22) с $i=2^k, 2\cdot 2^k, \ldots, 2^m-2^k$ требуется $\frac{2^m-2^k}{2^k}q = (2^{m-k}-1)q$ арифметических действий. При фиксированном k и для $l=1, 2, \ldots, 2^{k-1}$, $i=2^k, 2\cdot 2^k, \ldots, 2^m-2^k$, требуется $2^{k-1}(2^{m-k}-1)q=(2^{m-1}-2^{k-1})q$ действий. Наконец, для всех $k=1, 2, \ldots, m-1$ необходимо выполнить

$$q\sum_{k=1}^{m-1}(2^{m-1}-2^{k-1}) = (1+(m-2)\,2^{m-1})\,q$$

арифметических действий. Самое существенное здесь то, что при больших N число действий является величиной $O(qN\log_2 N)$.

После того как все векторы $p_i^{(k)}$, $q_i^{(k)}$ найдены, осуществляется обратный ход метода редукции, т. е. начиная с $k=m$ решаются

уравнения

$$C^{(k-1)}t_i^{(k-1)} = q_i^{(k-1)} + y_{i-2^{k-1}} + y_{i+2^{k-1}} \tag{23}$$

и вычисляются искомые векторы

$$y_i = p_i^{(k-1)} + t_i^{(k-1)},$$

где $k=m, m-1, \ldots, 1$, $i=2^{k-1}, 3\cdot 2^{k-1}, \ldots, N-2^{k-1}, N=2^m$.

Система (23) при фиксированных i, k решается, как и ранее, путем последовательного решения систем уравнений

$$C_{k-1,l}w_{l,i} = w_{l-1,i}, \quad l=1, 2, \ldots, 2^{k-1},$$
$$i=2^{k-1}, 3\cdot 2^{k-1}, 5\cdot 2^{k-1}, \ldots, 2^m-2^{k-1}, \tag{24}$$

где $w_{0,i} = q_i^{(k-1)} + y_{i-2^{k-1}} + y_{i+2^{k-1}}$, $w_{2^{k-1},i} = t_i^{(k-1)}$. Решение системы (23) также требует $O(qN\log_2 N)$ действий.

Рассмотрим применение метода редукции к решению разностного уравнения Пуассона (4), (5) из § 5. Предположим, что сетка содержит $N_1=2^m$ точек по направлению x_1 и N_2 точек — по x_2. Согласно (9) из § 5, это разностное уравнение можно записать в векторном виде (1), (2), где $N=N_1$ и $C=2E_2-h_1^2\Lambda_2$ — трехдиагональная матрица порядка N_2-1. Поэтому системы вида (22), (24) решаются методом прогонки, что требует $q=O(N_2)$ действий. Таким образом, решение разностного уравнения Пуассона осуществляется методом редукции за число действий $O(N_2N_1\log_2 N_1)$. На квадратной сетке, когда $N_1=N_2=N$, число действий является величиной $O(N^2\log_2 N)$, т. е. имеет тот же порядок, что и в методе быстрого дискретного преобразования Фурье. В отличие от последнего, метод редукции не требует знания собственных функций, что позволяет применять его и в более общих случаях, например в случае краевых условий третьего рода.

Метод редукции выгодно отличается от метода матричной прогонки не только числом действий, но и требуемой памятью ЭВМ. В то же время следует еще раз подчеркнуть, что метод редукции можно применять только для решения относительно простых систем уравнений, а именно систем, которые можно записать в виде (1) с постоянной матрицей C. Например, разностные задачи для уравнения

$$\frac{\partial}{\partial x}\left(k_1(x, y)\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(k_2(x, y)\frac{\partial u}{\partial y}\right) = -f(x, y)$$

можно решать методом редукции только в том случае, если коэффициенты k_1, k_2 не зависят от x.

СПИСОК ЛИТЕРАТУРЫ

1. Бабенко К. И. Основы численного анализа.— М.: Наука, 1986.
2. Бахвалов Н. С. Численные методы.— М.: Наука, 1975.
3. Бахвалов Н. С., Жидков Н. П., Кобельков Г. М. Численные методы.— М.: Наука, 1987.
4. Березин И. С., Жидков Н. П. Методы вычислений.— Ч. I.— М.: Наука, 1966. То же.— Ч. 2.— Физматгиз, 1962.
5. Бобков В. В., Городецкий Л. М. Избранные численные методы решения на ЭВМ инженерных и научных задач.— Минск: Изд-во «Университетское», 1985.
6. Воеводин В. В. Вычислительные основы линейной алгебры.— М.: Наука, 1977.
7. Воеводин В. В. Математические модели и методы в параллельных процессах.— М.: Наука, 1986.
8. Воеводин В. В., Кузнецов Ю. А. Матрицы и вычисления.— М.: Наука, 1984.
9. Волков Е. А. Численные методы.— М.: Наука, 1987.
10. Годунов С. К. Решение систем линейных уравнений.— Новосибирск: Наука, 1980.
11. Годунов С. К., Рябенький В. С. Разностные схемы, введение в теорию.— М.: Наука, 1977.
12. Ильин В. А., Позняк Э. Г. Линейная алгебра.— М.: Наука, 1984.
13. Ильин В. П., Кузнецов Ю. И. Трехдиагональные матрицы и их приложения.— М.: Наука, 1985.
14. Калиткин Н. Н. Численные методы.— М.: Наука, 1978.
15. Карпов В. Я. Алгоритмический язык фортран (фортран — Дубна).— М.: Наука, 1976.
16. Крылов В. И., Бобков В. В., Монастырный П. И. Вычислительные методы.— Т. I.— М.: Наука, 1976. То же.— Т. II.— М.: Наука, 1977.
17. Крылов В. И., Бобков В. В., Монастырный П. И. Начала теории вычислительных методов. Интерполирование и интегрирование.— Минск: Наука и техника, 1983.
18. Крылов В. И., Бобков В. В., Монастырный П. И. Начала теории вычислительных методов. Дифференциальные уравнения.— Минск: Наука и техника, 1982.
19. Ляшко И. И., Макаров В. Л., Скоробогатько А. А. Методы вычислений.— Киев: Вища школа, 1977.
20. Макаров В. Л., Хлобыстов В. В. Сплайн-аппроксимация функций.— М.: Высшая школа, 1983.
21. Марчук Г. И., Методы вычислительной математики.— 3-е изд.— М.: Наука, 1989.
22. Марчук Г. И., Агошков В. И. Введение в проекционно-сеточные методы.— М.: Наука, 1981.
23. Марчук Г. И., Шайдуров В. В. Повышение точности решений разностных схем.— М.: Наука, 1979.
24. Натансон И. П. Конструктивная теория функций.— М.: Гостехиздат, 1949.

25. Островский А. М. Решение уравнений и систем уравнений.— М.: ИЛ, 1963.
26. Ракитский Ю. В., Устинов С. М., Черноруцкий И. Г. Численные методы решения жестких систем.— М.: Наука, 1979.
27. Рихтмайер Р., Мортон К. Разностные методы решения краевых задач.— М.: Мир, 1972.
28. Рябенький В. С., Филиппов А. Ф. Об устойчивости разностных уравнений.— М.: Гостехиздат, 1956.
29. Салтыков А. И., Макаренко Г. И. Программирование на языке фортран.— М.: Наука, 1976.
30. Самарский А. А. Введение в теорию разностных схем.— М.: Наука, 1971.
31. Самарский А. А. Введение в численные методы.— 2-е изд.— М.: Наука, 1987.
32. Самарский А. А. Теория разностных схем.— 2-е изд.— М.: Наука, 1983.
33. Самарский А. А., Андреев В. Б. Разностные методы для эллиптических уравнений.— М.: Наука, 1976.
34. Самарский А. А., Гулин А. В. Устойчивость разностных схем.— М.: Наука, 1973.
35. Самарский А. А., Николаев Е. С. Методы решения сеточных уравнений.— М.: Наука, 1978.
36. Самарский А. А., Попов Ю. П. Разностные методы решения задач газовой динамики.— 2-е изд.— М.: Наука, 1980.
37. Современные численные методы решения обыкновенных дифференциальных уравнений.— М.: Мир, 1979.
38. Тихонов А. Н., Арсенин В. Я. Методы решения некорректных задач.— 3-е изд.— М.: Наука, 1986.
39. Тихонов А. Н., Васильева А. Б., Свешников А. Г. Дифференциальные уравнения.— 2-е изд.— М.: Наука, 1985.
40. Тихонов А. Н., Костомаров Д. П. Вводные лекции по прикладной математике.— М.: Наука, 1984.
41. Тихонов А. Н., Самарский А. А. Уравнения математической физики.— 5-е изд.— М.: Наука, 1977.
42. Треногин В. А. Функциональный анализ.— М.: Наука, 1980.
43. Турчак Л. И. Основы численных методов.— М.: Наука, 1987.
44. Уилкинсон Дж. Х. Алгебраическая проблема собственных значений.— М.: Наука, 1970.
45. Фаддеев Д. К. Лекции по алгебре.— М.: Наука, 1984.
46. Форсайт Дж., Малькольм М., Моулер К. Машинные методы математических вычислений.—М.: Мир, 1980.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

Алгоритм быстрого дискретного преобразования Фурье 336
Аппроксимация дифференциального оператора разностным 266
— первого порядка 35
— суммарная 376

Ведущий элемент 53
Вычислительный эксперимент 11

Граница сетки 293
Граничная точка сетки 297

Дискретизация 11

Жесткая система 249

Задача о наилучшем приближении 157

Интегро-интерполяционный метод 262
Интерполирование 127
Интерполяционный многочлен 127
— — Лагранжа 129
— — Ньютона 130
— — обобщенный 151, 156
— — Эрмита 136
Итерационный метод 48
— — верхней релаксации 85
— — двухшаговый явный 208
— — Зейделя 83
— — — нелинейный 212
— — касательных 193
— — минимальных невязок 116
Итерационный метод минимальных поправок 116
— — многошаговый 84
— — неявный 85, 216
— — Ньютона 193, 210
— — — модифицированный 193
— — — с параметром 202, 211
— — одношаговый 84
— — парабол 85, 194
— — переменных направлений 404
— — Пикара 209
— — попеременно-треугольный 394
— — релаксации 209
— — стационарный 85, 208
— — Стеффенсена 199
— — явный 85, 208
— — Якоби 82
— — — нелинейный 212

Каноническая форма одношагового итерационного метода 84
— — разностного уравнения 293
— — разностной схемы двухслойной 349
— — — — трехслойной 363
Квадратурная формула 161
— — Гаусса 180
— — интерполяционного типа 173
— — Ньютона — Котеса 178
— — парабол 165
— — прямоугольников 162, 163
— — составная 179
— — трапеций 164
— — Эрмита 185
Корректность операторных уравнений 342
— разностной схемы 290
— численного метода 15

Мантисса числа 116
Матрица верхняя треугольная 51
— нижняя треугольная 51

Матрица перехода 91
— плохо обусловленная 76
Машинный нуль 17
— эпсилон 19
Метод Адамса 231
— баланса 262
— бисекции 191
— гармоник 275
— Гаусса 51
— — с выбором главного элемента 61
— интегро-интерполяционный 262
— матричной прогонки 411
— последовательных приближений 48
— прогонки 45
— простой итерации 85
— разностный 217
— Ричардсона 85
— Ромберга 172
— Рунге 168
— Рунге — Кутта 217
— Эйлера 215
— — неявный 249
— Эйткена ускорения сходимости 198
— экстраполяции 167, 172
Модель математическая 11
Невязка 43, 116, 215, 231
Норма матрицы 75
— энергетическая 317

Область устойчивости 253
Округление 18
Оператор второй разностной производной 311
— левой разностной производной 348
— монотонный 304
— перехода разностной схемы 353
— положительный 315
— правой разностной производной 348

Погрешность абсолютная 76
— аппроксимации 35, 43, 231
— — на решении 44, 215, 268, 275, 289
— вычислительная 14
— дискретизации 13
— интерполирования 132
— итерационного метода 87
— метода 215
— неустранимая 13
— округления 13, 14
— относительная 18, 76
— разностной схемы 274, 289
— экстраполирования 133
Позиционная система счисления 16
Порядок аппроксимации 44, 216, 289
— точности 44
— — разностного метода 215, 268, 290
Порядок числа 16
Принцип максимума 296
Пространство сеточных функций 287

Разделенная разность 129
Разностная краевая задача 37
— схема 13, 37, 217, 262, 287
— — абсолютно устойчивая 276
— — асимптотически устойчивая 328
— — двуслойная 349
— — консервативная 115
— — локально-одномерная 377
— — монотонная 304
— — неустойчивая 276
— — повышенного порядка аппроксимации 278
— — продольно-поперечная 372
— — с весами 277, 321
— — трехслойная 283, 362
— условно устойчивая 276
— — чисто неявная 276
— — шеститочечная симметричная 277
— — явная 274
— формула Грина 269
Разностный метод 34
— — абсолютно устойчивый 249
— — многошаговый 230
— — условно устойчивый 249
— — чисто неявный 255
— — A-устойчивый 254
— — $A(\alpha)$-устойчивый 255
— оператор Лапласа 261
Разряд числа 16

Сетка 15, 34, 286
— на отрезке 134
— равномерная 34, 215
— разрядная 17
— связная 295
Сеточная функция 34, 215, 286
Слой 273
Скорость сходимости итерационного метода 96
Сплайн 141
Сходимость итерполяционного процесса 135
— квадратичная 193
— при $\tau \to 0$ 215
— разностной схемы 286, 290

Теорема сравнения 298

Узел внутренний 273

Узел граничный 273
— интерполирования 127
— кратный 136
— сетки 34, 273
Устойчивость коэффициентная 74
— разностной схемы 290, 342, 351
— — — по начальным данным 240, 352
— — — по правой части 352

Формула суммирования по частям 39, 269
Функция мажорантная 299

Характеристическое уравнение 26, 234

Число жесткости 250
— обусловленности 76
— с плавающей запятой 16
— с фиксированной запятой 16

Шаблон разностного оператора 261
Шаг сетки 34, 286

САМАРСКИЙ Александр Андреевич,
ГУЛИН Алексей Владимирович

ЧИСЛЕННЫЕ МЕТОДЫ

Заведующий редакцией *Е. Ю. Ходан*
Редактор *Т. Н. Галишникова*
Художественный редактор *Т. Н. Кольченко*
Технические редакторы *Е. В. Морозова, С. Я. Шкляр*
Корректоры: *Т. Е. Егорова, Т. С. Вайсберг*

ИБ № 11740

Сдано в набор 19.07.88. Подписано к печати 09.02.89. Формат 60×90/16. Бумага книжно-журнальная. Гарнитура литературная. Печать высокая. Усл. печ. л. 27. Усл. кр.-отт. 27. Уч.-изд. л. 27,31. Тираж 36000 экз. Заказ № 4624. Цена 1 р. 20 к.

Ордена Трудового Красного Знамени издательство «Наука»
Главная редакция физико-математической литературы
117071 Москва В-71, Ленинский проспект, 15

Вторая типография издательства «Наука»,
121099 Москва, Шубинский пер., 6

Alexander SAMARSKII and *Alexei GOOLIN*

NUMERICAL METHODS

Moskow, Nauka, Main Editorial Board for Physical and Mathematical Literature, 1989

Readership: Applied and computational mathematicians, college teachers and students.

Summary: The material of this book comes from courses that the authors has offered in the Computational Mathematics and Cybernetics Department at Moscow State University. It consists of three parts. Part 1 is of introductory nature. Here the idea of computational experiment as a tool of scientific researches is given, also some theoretical notions related to numerical methods are presented. Part 2 includes such traditional topics as interpolation, numerical integration, numerical linear and non-linear algebra, Runge — Kutta and multistep methods for ordinary differential equations. Part 3 which based on original authors papers presents the theory of difference schemes for partial differential equations including the methods of construction and investigation of difference schemes as well as direct and iteration methods for solving grid equations.

Contents: 1. Mathematical simulation and numerical experiment. 2. Roundoff errors. 3. Two-order difference equations. 4. Direct and iteration methods for solving systems of linear algebraic equations. 5. Interpolation. 6. Solving of nonlinear equations. 7. Numerical integration. 8. Numerical methods for ordinary differential equations. 9. The main notions of the difference schemes theory. 10. The maximume principle and variable dividing for difference schemes. 11. Stability theory of difference schemes. 12. Direct and iteration methods for grid equations.

The authors: Academician A. A. Samarskii is a chief of department of Keldysh Institute of Applied Mathematics, Academy of Science of the USSR, the chairman of the Scientific Council on the problem «Mathematical modelling» Academy of Science of the USSR, the Hero of Socialist Labour, the Lenin and State Prises winner. He is the author of number monographs and textbooks on mathematical physics, theory of difference schemes and numerical methods such as follows.

The equations of mathematical physics (together with A. N. Tichonov), translated into English, German, French. Theory of difference schemes, translated into English. Stability of difference schemes (together with A. V. Goolin). Difference schemes for elliptic equations (together with V. B. Andreev), translated into French.

The difference methods for gas dynamic problems (together with Yu. P. Popov).

Numerical methods for grid equations (together with E. S. Nikolaev), translated into English, French, Italian. D. s. A. V. Goolin is a professor of Computational Mathematics and Cybernetics Department at Moscow State University, a specialist in the field of numerical methods for differential equations.