

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ  
имени М.В. ЛОМОНОСОВА



Факультет  
вычислительной математики  
и кибернетики



**В.Б. АНДРЕЕВ**

# ЧИСЛЕННЫЕ МЕТОДЫ

Учебное пособие

МОСКВА

2013

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ  
имени М.В. ЛОМОНОСОВА  
Факультет вычислительной математики и кибернетики

**В.Б. Андреев**

# **Численные методы**

*Учебное пособие*



---

МОСКВА – 2013

УДК 519.6(075.8)

ББК 22.193я73

A65

*Печатается по решению Редакционно-издательского совета  
факультета вычислительной математики и кибернетики МГУ имени М.В. Ломоносова*

Рецензенты:

*А.В. Разулин – профессор; Н.В. Соснин – профессор*

**Андреев В.Б.**

**A65 Численные методы: Учебное пособие.** – М.: Издательский отдел факультета ВМиК МГУ имени М.В. Ломоносова (лицензия ИД N 05899 от 24.09.2001 г.); МАКС Пресс, 2013. – 336 с.  
ISBN 978-5-89407-514-3  
ISBN 978-5-317-04669-9

Учебное пособие посвящено изложению тех разделов вычислительной математики, которые на факультете ВМК МГУ имени М.В. Ломоносова изучаются на третьем курсе. Основными из указанных разделов являются вычислительная линейная алгебра и численные методы решения дифференциальных уравнений. В разделе линейной алгебры представлены прямые и итерационные методы решения систем линейных алгебраических уравнений с квадратной невырожденной матрицей и методы решения задачи на собственные значения. Далее изложены методы решения задачи Коши для обыкновенных дифференциальных уравнений, разностные схемы для двукточечных краевых задач и уравнений теплопроводности и колебаний струны.

*Ключевые слова:* численный метод, разностная схема, линейная алгебра, задача Коши, ОДУ, уравнение теплопроводности, уравнение колебаний струны, QR-алгоритм, методы Рунге-Кутты.

УДК 519.6(075.8)

ББК 22.193я73

**Andreev V.B.**

**Numerical Methods: Textbook.** – М.: МАКС Press, 2013. – 336 p.

This manual is devoted to those parts of computational mathematics, which are studied at the Faculty of Computational Mathematics and Cybernetics of Lomonosov Moscow State University by the third-year students. Main topics are numerical linear algebra and numerical methods for solving differential equations. The linear algebra section includes direct and iterative methods for solving systems of linear algebraic equations with square nonsingular matrices and methods for solving eigenvalue problems. Further, some numerical methods for solving the Cauchy problem for ordinary differential equations and difference schemes for two-point boundary value problems and for the heat equation and the wave equation are considered.

*Key words:* numerical method, difference scheme, linear algebra, Cauchy problem, ODE, heat equation, wave equation, QR-algorithm, Runge-Kutta methods.

---

*Учебное издание*  
**АНДРЕЕВ Владимир Борисович**  
**ЧИСЛЕННЫЕ МЕТОДЫ**  
*Учебное пособие*

Издательский отдел Факультета вычислительной математики и кибернетики МГУ имени М.В. Ломоносова

Лицензия ИД N 05899 от 24.09.01 г.

119992, ГСП-2, Москва, Ленинские горы, МГУ имени М.В. Ломоносова, 2-й учебный корпус

Напечатано с готового оригинал-макета в издательстве ООО «МАКС Пресс»

Лицензия ИД N 00510 от 01.12.99 г.

Подписано в печать 06.12.2013 г. Формат 60х90 1/16. Усл.печ.л. 21,0. Тираж 150 экз. Заказ 455.

119992, ГСП-2, Москва, Ленинские горы, МГУ им. М.В. Ломоносова,

2-й учебный корпус, 527 к. Тел. 8(495)939-3890/91. Тел./Факс 8(495)939-3891.

ISBN 978-5-89407-514-3

ISBN 978-5-317-04669-9

© Факультет ВМК МГУ имени М.В. Ломоносова, 2013

© Издательство ООО «МАКС Пресс», 2013

© Андреев В.Б., 2013

# Оглавление

|   |           |
|---|-----------|
| Предисловие   | vii       |
| Вычислительная линейная алгебра                         | 2         |
| <b>I Прямые методы решения линейных систем</b>          | <b>4</b>  |
| <b>1 Метод Гаусса и треугольное разложение матрицы</b>  | <b>5</b>  |
| 1.1 Метод исключения Гаусса . . . . .                   | 5         |
| 1.2 $LU$ разложение матрицы. . . . .                    | 8         |
| 1.3 Метод Холецкого (квадратных корней) . . . . .       | 15        |
| 1.4 Обращение матрицы . . . . .                         | 18        |
| <b>2 Методы <math>QR</math>-факторизации</b>            | <b>20</b> |
| 2.1 Метод вращений Гивенса . . . . .                    | 21        |
| 2.2 Метод отражений Хаусхолдера . . . . .               | 27        |
| <b>3 Ленточные методы</b>                               | <b>34</b> |
| 3.1 Метод прогонки . . . . .                            | 34        |
| 3.2 Ленточные матрицы . . . . .                         | 35        |
| 3.3 Ленточный вариант треугольного разложения . . . . . | 37        |
| 3.4 Оценка трудоемкости . . . . .                       | 40        |
| 3.5 Несимметричная ленточность . . . . .                | 44        |
| 3.6 Ленточный вариант метода Холецкого . . . . .        | 46        |
| 3.7 Метод блочного исключения . . . . .                 | 47        |

|            |  |            |
|------------|--|------------|
| <b>4</b>   | <b>Устойчивость вычислительных алгоритмов</b>                      | <b>51</b>  |
| 4.1        | Введение . . . . .   | 51         |
| 4.2        | Примеры плохо обусловленных систем . . . . .                       | 53         |
| 4.3        | Возмущение матрицы коэффициентов . . . . .                         | 56         |
| 4.4        | Пример хорошо обусловленной системы . . . . .                      | 58         |
| 4.5        | Метод Гаусса с выбором главного элемента . . . . .                 | 60         |
| <b>II</b>  | <b>Итерационные методы решения линейных систем</b>                 | <b>66</b>  |
| <b>5</b>   | <b>Простая итерация и чебышевский итерационный метод</b>           | <b>67</b>  |
| 5.1        | Одношаговые итерационные методы . . . . .                          | 68         |
| 5.2        | Неявные методы . . . . .   | 69         |
| 5.3        | Чебышевский итерационный метод . . . . .                           | 71         |
| 5.4        | Об устойчивости . . . . .  | 77         |
| <b>6</b>   | <b>Метод наискорейшего спуска</b>                                  | <b>78</b>  |
| 6.1        | Метод наискорейшего спуска . . . . .                               | 78         |
| 6.2        | Неулучшаемость оценки . . . . .                                    | 81         |
| <b>7</b>   | <b>Метод сопряженных градиентов</b>                                | <b>84</b>  |
| 7.1        | Построение метода . . . . .  | 84         |
| 7.2        | Оценка скорости сходимости . . . . .                               | 88         |
| 7.3        | Вспомогательные утверждения . . . . .                              | 90         |
| 7.4        | Окончательные соотношения . . . . .                                | 94         |
| 7.5        | Метод сопряженных градиентов с предобуславливателем . . . . .      | 95         |
| 7.6        | Неполное разложение разреженных матриц . . . . .                   | 97         |
| <b>III</b> | <b>Задача на собственные значения</b>                              | <b>100</b> |
| <b>8</b>   | <b>Степенной метод и обратные итерации</b>                         | <b>101</b> |
| 8.1        | Постановка задачи . . . . .  | 101        |
| 8.2        | Степенной метод . . . . .  | 104        |
| 8.2.1      | Нахождение максимального по модулю собственного значения . . . . . | 104        |

|           |   |            |
|-----------|---|------------|
| 8.2.2     | Пример . . . . .  | 108        |
| 8.2.3     | Нахождение второго по величине модуля собственного значения . . . . . | 109        |
| 8.3       | Обратные итерации . . . . .   | 110        |
| 8.4       | Итерации с отношением Рэля . . . . .                                  | 110        |
| <b>9</b>  | <b><math>QR</math> - алгоритм</b>                                     | <b>114</b> |
| 9.1       | Ускорение сходимости $QR$ -алгоритма . . . . .                        | 123        |
|           | <b>Численные методы математического анализа</b>                       | <b>132</b> |
| <b>10</b> | <b>Разностные уравнения</b>   | <b>133</b> |
| 10.1      | Разностные уравнения первого порядка . . . . .                        | 135        |
| 10.2      | Разностные уравнения $k$ -го порядка . . . . .                        | 138        |
| 10.3      | Системы разностных уравнений . . . . .                                | 141        |
| 10.4      | Задача на собственные значения . . . . .                              | 142        |
| <b>11</b> | <b>Ортогональные многочлены</b>                                       | <b>146</b> |
| 11.1      | Общие ортогональные многочлены . . . . .                              | 146        |
| 11.2      | Многочлены Чебышева первого рода . . . . .                            | 150        |
| 11.3      | Свойства многочленов Чебышева . . . . .                               | 152        |
| 11.4      | Многочлены Лежандра . . . . .   | 155        |
| <b>12</b> | <b>Численное дифференцирование</b>                                    | <b>156</b> |
| 12.1      | Введение . . . . .  | 156        |
| 12.2      | Метод неопределенных коэффициентов . . . . .                          | 160        |
| 12.3      | Использование интерполяционных формул . . . . .                       | 162        |
| 12.4      | О корректности численного дифференцирования . . . . .                 | 167        |
| <b>13</b> | <b>Методы решения нелинейных уравнений</b>                            | <b>169</b> |
| 13.1      | Метод бисекции . . . . .  | 170        |
| 13.2      | Метод простых итераций . . . . .                                      | 171        |
| 13.3      | Метод Ньютона . . . . .   | 174        |
| 13.4      | Метод секущих . . . . .   | 180        |

|  |   |            |
|--|---|------------|
| 13.5   | Глобальная сходимость. . . . .                        | 187        |
| 13.6   | Системы нелинейных уравнений . . . . .                | 188        |
| 13.6.1   | Метод Ньютона . . . . .                               | 191        |
| 13.6.2   | Аналоги метода секущих . . . . .                      | 192        |
| 13.6.3   | Метод Бройдена . . . . .                              | 193        |
| <b>Численные методы решения дифференциальных уравнений</b> |   | <b>196</b> |
| <b>IV</b>  | <b>Численное решение задачи Коши для ОДУ</b>          | <b>197</b> |
| <b>14</b>  | <b>Введение</b>                                       | <b>198</b> |
| 14.1   | Примеры численных методов . . . . .                   | 198        |
| 14.2   | Аппроксимация. . . . .                                | 201        |
| <b>15</b>  | <b>Методы Рунге-Кутты</b>                             | <b>204</b> |
| 15.1   | Общая концепция . . . . .                             | 204        |
| 15.2   | Одноэтапные методы Рунге-Кутты . . . . .              | 207        |
| 15.3   | Методы третьего порядка аппроксимации . . . . .       | 210        |
| 15.4   | Двухэтапные неявные методы третьего порядка . . . . . | 214        |
| 15.5   | Явные двухэтапные методы . . . . .                    | 217        |
| 15.6   | Двухэтапный метод четвертого порядка . . . . .        | 218        |
| 15.7   | Явные трехэтапные методы третьего порядка . . . . .   | 219        |
| 15.8   | Более общие методы Рунге-Кутты . . . . .              | 222        |
| 15.9   | Сходимость методов Рунге-Кутты . . . . .              | 223        |
| <b>16</b>  | <b>Линейные многошаговые методы</b>                   | <b>226</b> |
| 16.1   | Методы Адамса . . . . .                               | 226        |
| 16.2   | Формулы дифференцирования назад . . . . .             | 230        |
| 16.3   | Общие линейные многошаговые методы . . . . .          | 232        |
| 16.4   | Погрешность аппроксимации методов Адамса . . . . .    | 234        |
| 16.5   | Поучительный пример . . . . .                         | 235        |

|   |            |
|---|------------|
| <b>17 Устойчивость</b>  | <b>238</b> |
| 17.1 Нуль-устойчивость . . . . .  | 238        |
| 17.2 Жесткие задачи . . . . .   | 242        |
| 17.3 A-устойчивость . . . . .   | 245        |
| 17.4 Устойчивость методов Рунге-Кутты . . . . .                                       | 251        |
| <br>  |            |
| <b>V Двухточечные краевые задачи</b>  | <b>255</b> |
| <br>  |            |
| <b>18 Элементы теории разностных схем</b>   | <b>256</b> |
| 18.1 Введение . . . . .   | 256        |
| 18.2 Основные понятия теории разностных схем . . . . .                                | 258        |
| 18.3 Разрешимость и сходимость . . . . .  | 261        |
| 18.4 Метод баланса (конечных объемов) . . . . .                                       | 265        |
| 18.5 Аппроксимация граничных условий . . . . .  | 267        |
| 18.6 Исследование погрешности аппроксимации . . . . .                                 | 269        |
| 18.7 Уравнения с разрывными коэффициентами . . . . .                                  | 272        |
| 18.8 Неравномерная сетка . . . . .  | 276        |
| 18.9 Априорные оценки и оценка точности . . . . .                                     | 279        |
| 18.10 Аппроксимация производной . . . . .   | 284        |
| 18.11 Уравнение конвекции-диффузии . . . . .  | 285        |
| <br>  |            |
| <b>19 Сингулярно возмущенные уравнения</b>  | <b>287</b> |
| 19.1 Осцилляции решения . . . . .   | 287        |
| 19.2 Четырехточечная схема . . . . .  | 291        |
| 19.3 Монотонная схема Самарского . . . . .  | 292        |
| 19.4 О равномерной по $\epsilon$ сходимости . . . . .                                 | 294        |
| <br>  |            |
| <b>20 Численные методы для задач с негладкими решениями</b>                           | <b>295</b> |
| <br>  |            |
| <b>VI Численные методы для дифференциальных уравнений<br/>с частными производными</b> | <b>300</b> |
| <br>  |            |
| <b>21 Разностные методы для уравнения теплопроводности</b>                            | <b>301</b> |
| 21.1 Устойчивость по начальным данным . . . . .                                       | 305        |



|   |            |
|---|------------|
| 21.2 Устойчивость по правой части . . . . .               | 309        |
| 21.3 Устойчивость в смысле максимума модуля . . . . .     | 311        |
| 21.4 Сеточное преобразование Фурье . . . . .              | 312        |
| 21.5 Устойчивость по начальным данным . . . . .           | 314        |
| <b>22 Разностные схемы для уравнения колебаний струны</b> | <b>316</b> |
| 22.1 Аппроксимация . . . . .                              | 316        |
| 22.2 Устойчивость по начальным данным . . . . .           | 318        |
| <b>Литература</b>   | <b>323</b> |
| <b>Предметный указатель</b>                               | <b>325</b> |

# Предисловие

Численные методы или, что то же самое, вычислительные методы, методы вычислений есть раздел математики, называемый также вычислительной математикой, который изучает методы (приближенные) решения различных математических задач. Под словом изучает понимается как разработка этих методов, так и исследование их свойств. Изучением методов решения математических задач занимаются и другие разделы математики. В чем же специфика математики вычислительной? Пусть, например, требуется вычислить определенный интеграл от действительной функции одной переменной. В математическом анализе — тоже разделе математики — в частности изучаются методы построения первообразных таких функций, и если средствами математического анализа первообразная интересующей нас функции может быть найдена, то рассматриваемая задача элементарно решается при помощи формулы Ньютона-Лейбница. Однако хорошо известно, что первообразные в элементарных и даже в специальных функциях выражаются далеко не всегда, а потому наша задача этим средствами тоже решена может быть не всегда. Здесь и наступает черед вычислительной математики, которая предлагает искать решение (приближенное) поставленной задачи при помощи квадратурных формул. Похожая ситуация возникает при решении обыкновенных дифференциальных уравнений, дифференциальных уравнений с частными производными и во многих других случаях.

Данная книга состоит из трех частей. Первая часть посвящена задачам линейной алгебры. Здесь изложены прямые и некоторые итерационные методы решения систем линейных алгебраических уравнений с квадратной невырожденной матрицей, а также методы решения задачи на собственные значения. Во второй части кратко рассмотрено численное дифференцирование и методы решения нелинейных уравнений и систем. Здесь же изложен некоторый вспомогательный аппарат. Третья часть посвящена численному решению дифференциальных уравнений. Рассмотрены три класса задач: задача Коши для обыкновенных диффе-

ренциальных уравнений первого порядка, двухточечные краевые задачи для уравнения второго порядка и задачи для простейших уравнений с частными производными. Для задачи Коши изложены явные и неявные методы Рунге-Кутты и многошаговые методы. Разностные методы для двухточечных задач изложены и исследованы на равномерной, неравномерной и сгущающихся сетках. Из уравнений с частными производными рассмотрены одномерное нестационарное уравнение теплопроводности и уравнение колебаний струны.

Столь специфический отбор материала связан с тем, что именно эти разделы численных методов читаются на третьем курсе ВМК МГУ.

Автор приносит искреннюю благодарность И.Г. Белухиной за ее труд по оформлению книги.

# Вычислительная линейная алгебра

Вычислительная линейная алгебра есть раздел численных методов и линейной алгебры, который занимается разработкой и исследованием практических алгоритмов решения на компьютерах задач линейной алгебры. Вычислительная линейная алгебра находит широкое применение в большинстве числовых вычислений во всех областях приложений, таких как физика, техника, химия, финансы и т.д. Имеются две основные задачи матричных вычислений:

- \* решение систем линейных уравнений,
- \* вычисление собственных значений и собственных векторов матриц.

Конечно, в линейной алгебре имеются и другие важные задачи, но эти две преобладают, и именно они будут изучены в данном курсе.

В силу теоремы Кронекера - Капелли система линейных алгебраических уравнений

$$Ax = b$$

разрешима тогда и только тогда, когда ранг матрицы  $A$  равен рангу расширенной матрицы  $[Ab]$ . Это заведомо так, если матрица  $A$  квадратная и невырожденная, т.е.  $\det A \neq 0$ . В этом случае система не только разрешима при любых  $b$ , но и имеет единственное решение (разрешима однозначно). Именно этот случай мы и будем изучать.

Методы решения систем линейных алгебраических уравнений делятся на две группы. К первой группе принадлежат так называемые прямые методы — алгоритмы, позволяющие получить решение за конечное число арифметических действий. Сюда относятся известное правило Крамера нахождения решения при помощи определителей, метод исключения Гаусса, метод прогонки — метод решения систем с трехдиагональными матрицами. Существуют и другие методы, из которых отметим метод Холецкого (метод квадратных корней), применяемый к системам с симметричными положительно определенными матрицами, метод вращений и метод отражений.

Вторую группу составляют приближенные методы, в частности, итерационные. В итерационных методах решение системы получается как предел при стремлении числа итераций  $n$  к бесконечности. При конечных  $n$ , как правило, получаются лишь приближенные решения.

Прямые и итерационные методы имеют свою область применения: если размерность системы не слишком велика, то часто предпочтительнее

использовать прямые методы. Итерационные методы выгодны для систем большого порядка. Особенно в случае матриц специального вида.

# I

## Прямые методы решения линейных систем

# 1

## Метод исключения Гаусса и треугольное ( $LU$ ) разложение матрицы

Исключение Гаусса безусловно знакомо читателю хотя бы из курса линейной алгебры, а, может быть, и из школьного курса. Этот метод является простейшим методом решения системы линейных алгебраических уравнений. Он широко используется при ручном решении таких систем и в то же время является стандартным методом для их решения на компьютерах.

### 1.1 Метод исключения Гаусса

Система  $Ax = b$  в развернутой форме имеет вид

$$\sum_{j=1}^n a_{ij}x_j = b_i, \quad i = 1, \dots, n. \quad (1.1)$$

*Метод Гаусса* или *метод последовательного исключения неизвестных* состоит в том, что неизвестные  $x_j$ ,  $j = 1, \dots, n - 1$  последовательно исключаются из соответствующих уравнений системы (1.1), в результате чего она преобразуется к эквивалентной системе с треугольной матрицей, решение которой просто находится.

Напомним требуемые рассуждения для случая  $n = 3$ . При этом значении  $n$  система (1.1) примет вид

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1, \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2, \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3. \end{aligned}$$



Снабдим коэффициенты матрицы этой системы и элементы ее правой части верхним индексом  $(0)$ , т.е. пусть  $a_{ij} =: a_{ij}^{(0)}$ ,  $b_j =: b_j^{(0)}$ . Тогда

$$\begin{aligned} a_{11}^{(0)} x_1 + a_{12}^{(0)} x_2 + a_{13}^{(0)} x_3 &= b_1^{(0)}, \\ a_{21}^{(0)} x_1 + a_{22}^{(0)} x_2 + a_{23}^{(0)} x_3 &= b_2^{(0)}, \\ a_{31}^{(0)} x_1 + a_{32}^{(0)} x_2 + a_{33}^{(0)} x_3 &= b_3^{(0)}. \end{aligned} \quad (1.2)$$

*Первый шаг.* Предположим, что  $a_{11} = a_{11}^{(0)} \neq 0$ , и поделим первое уравнение системы на этот коэффициент. Затем умножим полученное уравнение на  $a_{21}^{(0)}$  и результат вычтем из второго уравнения

$$\left( a_{21}^{(0)} - \frac{a_{21}^{(0)}}{a_{11}^{(0)}} a_{11}^{(0)} \right) x_1 + \left( a_{22}^{(0)} - \frac{a_{21}^{(0)}}{a_{11}^{(0)}} a_{12}^{(0)} \right) x_2 + \left( a_{23}^{(0)} - \frac{a_{21}^{(0)}}{a_{11}^{(0)}} a_{13}^{(0)} \right) x_3 = b_2^{(0)} - \frac{a_{21}^{(0)}}{a_{11}^{(0)}} b_1^{(0)}.$$

Коэффициент при  $x_1$  в преобразованном втором уравнении равен нулю, и поэтому система (1.2) после этого преобразования принимает вид

$$\begin{aligned} a_{11}^{(0)} x_1 + a_{12}^{(0)} x_2 + a_{13}^{(0)} x_3 &= b_1^{(0)}, \\ a_{22}^{(1)} x_2 + a_{23}^{(1)} x_3 &= b_2^{(1)}, \\ a_{31}^{(0)} x_1 + a_{32}^{(0)} x_2 + a_{33}^{(0)} x_3 &= b_3^{(0)}, \end{aligned} \quad (1.3)$$

где

$$a_{2j}^{(1)} = a_{2j}^{(0)} - l_{21} a_{1j}^{(0)}, \quad l_{21} = a_{21}^{(0)} / a_{11}^{(0)}, \quad b_2^{(1)} = b_2^{(0)} - l_{21} b_1^{(0)}.$$

Теперь преобразованное первое уравнение с единичным коэффициентом при  $x_1$  умножим на  $a_{31}^{(0)}$  и вычтем из третьего уравнения

$$\left( a_{31}^{(0)} - \frac{a_{31}^{(0)}}{a_{11}^{(0)}} a_{11}^{(0)} \right) x_1 + \left( a_{32}^{(0)} - \frac{a_{31}^{(0)}}{a_{11}^{(0)}} a_{12}^{(0)} \right) x_2 + \left( a_{33}^{(0)} - \frac{a_{31}^{(0)}}{a_{11}^{(0)}} a_{13}^{(0)} \right) x_3 = b_3^{(0)} - \frac{a_{31}^{(0)}}{a_{11}^{(0)}} b_1^{(0)}.$$

Коэффициент при  $x_1$  равен нулю и в этом уравнении, а исходная система преобразуется к виду

$$\begin{aligned} a_{11}^{(0)} x_1 + a_{12}^{(0)} x_2 + a_{13}^{(0)} x_3 &= b_1^{(0)}, \\ a_{22}^{(1)} x_2 + a_{23}^{(1)} x_3 &= b_2^{(1)}, \\ a_{32}^{(1)} x_2 + a_{33}^{(1)} x_3 &= b_3^{(1)}, \end{aligned} \quad (1.4)$$

где

$$a_{3j}^{(1)} = a_{3j}^{(0)} - l_{31} a_{1j}^{(0)}, \quad l_{31} = a_{31}^{(0)} / a_{11}^{(0)}, \quad b_3^{(1)} = b_3^{(0)} - l_{31} b_1^{(0)}.$$



Вычисления по этим формулам называют *обратным ходом метода Гаусса* или *обратной подстановкой*.

**Замечание 1.1.** В формулах (1.6) при преобразованиях системы (1.1) первое уравнение осталось без изменения. С равным успехом может быть использован и другой вариант исключения, когда первое уравнение (1.1) делится на  $a_{11}$ , а вместо (1.6) получается система с единичными коэффициентами при  $x_j$  в  $j$ -ом уравнении.

**Замечание 1.2.** Вычисления по формулам (1.9), (1.10), а, следовательно, и по формулам (1.7), (1.8) возможны лишь тогда, когда все числа

$$a_{ii}^{(i-1)} \neq 0, \quad i = 1, \dots, n. \quad (1.11)$$

Необходимые и достаточные условия выполнения (1.11) устанавливаются в доказываемой ниже теореме 1.2.

## 1.2 $LU$ разложение матрицы.

Покажем, что метод Гаусса эквивалентен разложению матрицы  $A$  системы (1.1) в произведение нижней  $L$  и верхней  $U$  треугольных матриц с последующим решением вспомогательных систем с этими матрицами. Сделаем это на примере матрицы третьего порядка из (1.2). Для этого введем в рассмотрение матрицы

$$E_{21} = \begin{bmatrix} 1 & 0 & 0 \\ -l_{21} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad E_{31} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -l_{31} & 0 & 1 \end{bmatrix}, \quad L_2^{-1} := E_{32} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -l_{32} & 1 \end{bmatrix},$$

где  $l_{ik}$  определяются соотношениями (1.9). Легко проверить, что переход от системы (1.2) к системе (1.3) может быть осуществлен умножением матрицы  $A^{(0)}$  системы (1.2) и ее правой части  $b^{(0)}$  на матрицу  $E_{21}$ , а от (1.3) к (1.4) — умножением соответствующей матрицы и вектора на матрицу  $E_{31}$ . Очевидно также, что

$$L_1^{-1} := E_{31}E_{21} = E_{21}E_{31} = \begin{bmatrix} 1 & 0 & 0 \\ -l_{21} & 1 & 0 \\ -l_{31} & 0 & 1 \end{bmatrix}$$

и, следовательно, переход от (1.2), минуя (1.3), сразу к (1.4) осуществляется умножением матрицы системы (1.2) и ее правой части на  $L_1^{-1}$ .

Аналогично (будем говорить только о матрицах)

$$A^{(2)} = L_2^{-1}A^{(1)} = L_2^{-1}L_1^{-1}A^{(0)},$$

где  $A^{(k)}$  — матрицы, получаемые на  $k$ -ом шаге, и, следовательно,

$$A = A^{(0)} = L_1L_2A^{(2)}. \quad (1.12)$$

Легко проверить, что

$$L_1 = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & 0 & 1 \end{bmatrix}, \quad L_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & l_{32} & 1 \end{bmatrix},$$

т.е. при обращении  $L_k$  меняется только знак перед поддиагональными элементами  $l_{ik}$ ,  $i > k$ . Далее,

$$L := L_1L_2 = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix},$$

а матрица  $A^{(2)}$  из (1.12), т.е. матрица системы (1.5), есть верхняя треугольная матрица. Обозначая ее через  $U$  и принимая во внимание вышесказанное, приходим к искомому разложению

$$A = LU, \quad (1.13)$$

где  $L$  — нижняя треугольная матрица с единичной главной диагональю.

Все вышесказанное остается справедливым и в общем случае матрицы  $A$  порядка  $n$ . Теперь матрицы  $L$  и  $U$  из (1.13) имеют вид

$$L = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ l_{21} & 1 & 0 & \dots & 0 \\ l_{31} & l_{32} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ l_{n1} & l_{n2} & l_{n3} & \dots & 1 \end{bmatrix}, \quad U = \begin{bmatrix} u_{11} & u_{12} & u_{13} & \dots & u_{1n} \\ 0 & u_{22} & u_{23} & \dots & u_{2n} \\ 0 & 0 & u_{33} & \dots & u_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & u_{nn} \end{bmatrix}, \quad (1.14)$$

где элементы матрицы  $l_{ik}$  матрицы  $L$  вычисляются по формулам (1.9), а элементы  $u_{kj} := a_{kj}^{(k-1)}$  матрицы  $U$  — по формулам (1.7).

Имея разложение (1.13), систему (1.1) можно переписать в виде

$$Ax = LUx = Ly = b, \quad Ux = y,$$

после чего решение системы (1.1) распадается на решение двух систем с треугольными матрицами

$$Ly = b \quad \text{и} \quad Ux = y. \quad (1.15)$$

Решение первой из этих систем заменяет преобразование вектора правой части системы (1.1) по формулам (1.8) прямого хода метода Гаусса. Решение же  $x$  второй системы определяется формулами (1.10) обратной подстановки, где  $b_i^{(i-1)} = y_i$ , а  $a_{ij}^{(i-1)} = u_{ij}$ .

**Замечание 1.3.** Соотношения (1.7) содержат формулы для  $u_{kj} = a_{kj}^{(k-1)}$  и промежуточные значения, которые тоже нужно запоминать и хранить. Мы сейчас преобразуем эти формулы к такому виду, при котором хранение промежуточных значений не требуется.

Пусть матрицы  $L$  и  $U$  имеют вид (1.14), т.е. их элементы

$$l_{ik} = 0 \quad \text{при} \quad k > i, \quad (1.16)$$

а

$$u_{kj} = 0 \quad \text{при} \quad k > j. \quad (1.17)$$

Поскольку  $LU = A$ , то по правилу умножения матриц находим, что

$$a_{ij} = \sum_{k=1}^n l_{ik} u_{kj}. \quad (1.18)$$

Преобразуем эту формулу двумя способами. В силу (1.14), (1.16)

$$\sum_{k=1}^n l_{ik} u_{kj} = \sum_{k=1}^{i-1} l_{ik} u_{kj} + l_{ii}^{=1} u_{ij} + \sum_{k=i+1}^n l_{ik}^{=0} u_{kj} = \sum_{k=1}^{i-1} l_{ik} u_{kj} + u_{ij},$$

а в силу (1.17)

$$\sum_{k=1}^n l_{ik} u_{kj} = \sum_{k=1}^{j-1} l_{ik} u_{kj} + l_{ij} u_{jj} + \sum_{k=j+1}^n l_{ik} \overset{0}{u}_{kj} = \sum_{k=1}^{j-1} l_{ik} u_{kj} + l_{ij} u_{jj}.$$

Отсюда и из (1.18) имеем

$$\begin{aligned} u_{ij} &= \left[ a_{ij} - \sum_{k=1}^{i-1} l_{ik} u_{kj} \right] \quad i = 1, \dots, n; \quad j = i, \dots, n; \\ l_{ij} &= \frac{1}{u_{jj}} \left[ a_{ij} - \sum_{k=1}^{j-1} l_{ik} u_{kj} \right] \quad j = 1, \dots, n; \quad i = j + 1, \dots, n. \end{aligned} \quad (1.19)$$

Очевидно, что реализация формул (1.19) возможна только тогда, когда все  $u_{jj} = a_{jj}^{(j-1)}$  отличны от нуля (ср. с (1.11)).

**Замечание 1.4.** Формулы (1.19) устроены так, что нельзя сначала вычислить все  $u_{ij}$ , а затем все  $l_{ij}$  или наоборот. Можно предложить следующий порядок вычислений по этим формулам:

$$\begin{aligned} u_{1j} &= a_{1j}, & j &= 1, 2, \dots, n; \\ l_{i1} &= a_{i1}/u_{11}, & i &= 2, 3, \dots, n; \\ u_{2j} &= a_{2j} - l_{21}u_{1j}, & j &= 2, 3, \dots, n; \\ l_{i2} &= (a_{i2} - l_{i1}u_{12})/u_{22}, & i &= 3, 4, \dots, n; \end{aligned} \tag{1.20}$$

и т.д., т.е. чередовать вычисление строк матрицы  $U$  и столбцов матрицы  $L$ .

После построения матриц  $L$  и  $U$  решение систем (1.15) с треугольными матрицами находятся по формулам

$$y_i = b_i - \sum_{k=1}^{i-1} l_{ik}y_k, \quad i = 1, 2, \dots, n, \tag{1.21}$$

(вычисления ведутся сверху вниз)— прямая подстановка,

$$x_k = \frac{1}{u_{kk}} \left[ y_k - \sum_{j=k+1}^n u_{kj}x_j \right], \quad k = n, n-1, \dots, 1 \tag{1.22}$$

(вычисления ведутся снизу вверх) — обратная подстановка.

Одной из важнейших характеристик любого численного метода является его трудоемкость. Под трудоемкостью метода, предназначенного для решения системы (1.1), обычно понимают число арифметических действий, необходимых для нахождения искомого решения. Часто в трудоемкость метода включают лишь действия умножения и деления, как наиболее трудоемкие операции с точки зрения работы компьютера. Так будем поступать и мы.

Легко видеть, что для вычислений по формулам (1.19) (для получения

треугольного разложения) требуется

$$\begin{aligned}
 Q &= \sum_{i=1}^n \sum_{j=i}^n (i-1) + \sum_{j=1}^n \sum_{i=j+1}^n j = \sum_{i=1}^n [(i-1)(n-i+1) + i(n-i)] = \\
 &= 2 \sum_{i=1}^n [(n+1)i - i^2] - n(n+1) = n(n+1)^2 - \frac{n(n+1)(2n+1)}{3} - \\
 &- n(n+1) = \frac{n(n^2-1)}{3} = \frac{n^3}{3} + O(n) \approx \frac{n^3}{3}
 \end{aligned} \tag{1.23}$$

действий умножения и деления.

Для вычислений по формулам (1.21) и (1.22) имеем соответственно

$$\overset{\circ}{q} = \sum_{i=1}^n (i-1) = \frac{n(n-1)}{2} \quad \text{и} \quad \bar{q} = \sum_{k=1}^n (n-k+1) = \frac{n(n+1)}{2}, \tag{1.24}$$

т.е. общее число действий для решения систем (1.15) по формулам (1.21), (1.22) есть

$$q = \overset{\circ}{q} + \bar{q} = n^2. \tag{1.25}$$

**Замечание 1.5.** Из формул (1.23) и (1.25) следует, что при больших  $n$  основной объем работы, которую нужно выполнить для решения системы (1.1) описанным методом, падает на преобразование коэффициентов матрицы системы, т.е. на построение треугольного разложения, в то время как преобразование вектора правой части (решение первой системы (1.15)) и на отыскание самого решения трудозатраты сравнительно невелики. В связи с этим при больших  $n$  решение нескольких систем с различными правыми частями и одной и той же матрицей оказывается по трудоемкости практически таким же как и решение одной системы.

Выясним теперь условия, при которых вычисления по формулам (1.19)-(1.22) возможны, т.е. все  $u_{jj}$  отличны от нуля.

**Теорема 1.1.** Пусть  $A$  — невырожденная матрица,  $L$  — нижняя треугольная матрица с единичной главной диагональю, а  $U$  — невырожденная верхняя треугольная матрица. Тогда, если  $A = LU$ , то это представление единственно.

Для доказательства теоремы 1.1 нам потребуется

**Лемма 1.1.** Пусть  $T$  есть нижняя треугольная матрица (соответственно, верхняя треугольная матрица). Тогда ее обратная (если она существует) также является нижней треугольной матрицей (соответственно, верхней треугольной) с диагональными элементами, равными обратным величинам диагональных элементов  $T$ . Пусть  $T'$  — другая нижняя треугольная матрица (соответственно, верхняя треугольная). Тогда произведение  $TT'$  есть также нижняя треугольная матрица (соответственно, верхняя треугольная) с диагональными элементами, равными произведениям соответствующих диагональных элементов  $T$  и  $T'$ .

**Упражнение 1.1.** Доказать лемму 1.1.

**Доказательство** теоремы 1.1. Пусть  $A = L_1U_1 = L_2U_2$ . Тогда

$$L_1 = L_2U_2U_1^{-1} \quad \text{и} \quad L_2^{-1}L_1 = U_2U_1^{-1}.$$

Слева стоит произведение нижних треугольных матриц, а справа — верхних. Поэтому произведение есть диагональная матрица  $D$ , т.е.  $L_2^{-1}L_1 = D$ . Отсюда находим, что  $L_1 = L_2D$ . Поскольку главные диагонали  $L_1$  и  $L_2$  единичные, а главная диагональ  $L_2D$  совпадает с главной диагональю  $D$ , то  $D = I$ . Отсюда  $L_1 = L_2$  и  $U_1 = U_2$ . Теорема доказана.

**Замечание 1.6.** Элементы  $l_{ik}$  и  $u_{kj}$  фигурирующих в теореме 1.1 матриц  $L$  и  $U$  определяются из нелинейной системы (1.18), содержащей  $n^2$  уравнений. Самых же элементов у двух треугольных матриц  $n(n+1)$ , т.е. на  $n$  больше, чем уравнений. Недостающие уравнения для однозначного определения матриц  $L$  и  $U$  как раз и задаются условием  $l_{ii} = 1$ ,  $i = 1, 2, \dots, n$ .

**Теорема 1.2.** Пусть  $A$  — квадратная невырожденная матрица,  $L$  — нижняя треугольная матрица с единичной главной диагональю, а  $U$  — невырожденная верхняя треугольная матрица. Разложение  $A = LU$  существует тогда и только тогда, когда все угловые миноры матрицы  $A$  отличны от нуля.

Напомним, что угловыми минорами матрицы  $A$  называются величины

$$\Delta_1 = a_{11}, \quad \Delta_2 = \det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad \dots, \quad \Delta_n = \det[A].$$



**Доказательство.** 1°. (Необходимость)

Пусть разложение  $A = LU$  существует. Тогда по теореме 1.1 оно единственно. Представим матрицы  $A$ ,  $L$  и  $U$  в блочном виде

$$A = \begin{bmatrix} A_m & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad L = \begin{bmatrix} L_m & 0 \\ L_{21} & L_{22} \end{bmatrix}, \quad U = \begin{bmatrix} U_m & U_{12} \\ 0 & U_{22} \end{bmatrix},$$

где  $A_m$ ,  $L_m$ ,  $U_m$  и  $A_{22}$ ,  $L_{22}$ ,  $U_{22}$  — квадратные матрицы размерностей  $m \times m$  и  $(n - m) \times (n - m)$  соответственно, а  $m$  — произвольное число.

Разложение  $A = LU$  в блочном представлении имеет вид

$$\begin{bmatrix} A_m & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} L_m & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} U_m & U_{12} \\ 0 & U_{22} \end{bmatrix} = \begin{bmatrix} L_m U_m & L_m U_{12} \\ L_{21} U_m & L_{21} U_{12} + L_{22} U_{22} \end{bmatrix} \quad (1.26)$$

Отсюда следует, что

$$A_m = L_m U_m. \quad (1.27)$$

Поскольку матрица  $U$  треугольная и невырожденная, то все ее диагональные элементы отличны от нуля. Поэтому невырождена и треугольная матрица  $U_m$ . Тем самым

$$\Delta_m = \det[A_m] = \det[L_m] \det[U_m] = u_{11} \dots u_{mm} \neq 0$$

при  $m = 1, \dots, n$ .

2°. (Достаточность) Пусть теперь  $\Delta_1 \Delta_2 \dots \Delta_n \neq 0$ . Для доказательства существования треугольного разложения воспользуемся методом полной математической индукции по порядку системы  $n$ . При  $n = 1$  матрица  $A = a_{11} = \Delta_1 \neq 0$ , матрица  $L = 1$  и поэтому  $U = u_{11} = a_{11} = \det U \neq 0$ . Существование искомого разложения при  $n = 1$  доказано.

Пусть  $A_k$  — матрица порядка  $k$  и разложение  $A_k = L_k U_k$  существует с  $\det U_k \neq 0$  при  $k = 1, \dots, m - 1$ . Докажем, что существует и  $A_m = L_m U_m$ , причем  $\det U_m \neq 0$ . Пусть  $a_{\cdot m} = [a_{1m} \dots a_{m-1m}]^T$  — столбец,  $a_{m\cdot} = [a_{m1} \dots a_{mm-1}]$  — строка и разложение  $A_m$  будем искать в виде

$$A_m = \begin{bmatrix} A_{m-1} & a_{\cdot m} \\ a_{m\cdot} & a_{mm} \end{bmatrix} = \begin{bmatrix} L_{m-1} & 0 \\ l_{m\cdot} & 1 \end{bmatrix} \begin{bmatrix} U_{m-1} & u_{\cdot m} \\ 0 & u_{mm} \end{bmatrix} = \begin{bmatrix} L_{m-1} U_{m-1} & L_{m-1} u_{\cdot m} \\ l_{m\cdot} U_{m-1} & l_{m\cdot} u_{\cdot m} + u_{mm} \end{bmatrix}.$$

Отсюда следует, что неизвестный столбец  $u_{\cdot m}$ , неизвестная строка  $l_{m\cdot}$  и неизвестный элемент  $u_{mm}$  определяются следующими соотношениями:

$$\begin{aligned} L_{m-1} u_{\cdot m} &= a_{\cdot m} \\ l_{m\cdot} U_{m-1} &= a_{m\cdot} \quad \Rightarrow \quad U_{m-1}^T l_{m\cdot}^T = a_{m\cdot}^T, \\ u_{mm} &= a_{mm} - l_{m\cdot} u_{\cdot m}. \end{aligned} \quad (1.28)$$

Поскольку  $L_{m-1}$  и  $U_{m-1}$  невырожденные, из первого и второго соотношений (1.28) можно найти  $u_{.m}$  и  $l_{m.}$ , соответственно, после чего третье соотношение дает  $u_{mm}$ . Существование разложения (1.26) для  $m$  доказано. Осталось доказать, что  $\det U_m \neq 0$ . Но с учетом (1.27)

$$0 \neq \Delta_m = \det A_m = \det U_m,$$

что и требовалось доказать. Теорема полностью доказана.

### 1.3 Метод Холецкого (квадратных корней)

Вновь обратимся к системе

$$Ax = b. \quad (1.29)$$

На этот раз будем предполагать, что матрица  $A$  симметрична и положительно определена, т.е.

$$A = A^T \quad \text{и} \quad A > 0. \quad (1.30)$$

Последнее означает, что квадратичная форма  $x^T Ax > 0$  для любого ненулевого вектора  $x$ . Напомним, что симметричная матрица имеет только действительные собственные значения, а положительно определенная — только положительные. В силу критерия Сильвестра необходимым и достаточным условием положительной определенности матрицы  $A$  является положительность всех ее угловых миноров  $\Delta_i > 0$ ,  $i = 1, \dots, n$ .

Построим алгоритм решения системы (1.29), который использует свойства (1.30) матрицы  $A$ . Это будет *метод Холецкого*. Основой метода Холецкого является

**Теорема 1.3.** *Если  $A = A^T > 0$ , то существует единственное разложение*

$$A = LL^T, \quad (1.31)$$

где  $L$  — нижняя треугольная матрица с положительными диагональными элементами.

**Определение 1.1.** Разложение (1.31) называется *разложением Холецкого*, а матрица  $L$  — *множителем Холецкого*.

**Доказательство теоремы.** Сначала докажем единственность. Пусть существуют два разложения

$$A = L_1 L_1^T = L_2 L_2^T.$$

Обращая матрицы  $L_1^T$  и  $L_2$ , будем иметь

$$L_2^{-1}L_1 = L_2^T (L_1^T)^{-1}.$$

Примем во внимание, что  $(AB)^T = B^T A^T$ , а для невырожденных матриц  $(AB)^{-1} = B^{-1}A^{-1}$ . Тогда

$$(L_1^{-1}L_2)^{-1} = L_2^{-1}L_1 = L_2^T (L_1^T)^{-1} = (L_1^{-1}L_2)^T. \quad (1.32)$$

В силу леммы 1.1 обратная к нижней треугольной матрице есть нижняя треугольная матрица и произведение таких матриц есть снова нижняя треугольная матрица. Из сказанного следует, что в левой части (1.32) стоит нижняя треугольная матрица, а справа — верхняя. Равенство (1.32) возможно только тогда, когда обе матрицы диагональные. Но диагональная матрица совпадает со своей транспонированной. Поэтому из (1.32) следует, что

$$(L_1^{-1}L_2)^{-1} = L_1^{-1}L_2 = D.$$

Это соотношение утверждает, что диагональная матрица  $D$  совпадает со своей обратной, что возможно только в том случае, если у этой матрицы диагональными элементами являются числа  $\pm 1$ . Поскольку  $L_2 = L_1 D$ , а диагональные элементы  $L_1$  и  $L_2$  положительны, то диагональные элементы  $D$  тоже должны быть положительны, т.е.  $D \equiv I$  и, следовательно,  $L_2 = L_1$ . Единственность доказана.

Построим теперь формулы для вычисления элементов  $L$ , откуда и будет следовать существование. Так как  $a_{ij} = a_{ji}$ , а  $l_{ij} = 0$  при  $i < j$ , то будем считать, что

$$i \geq j.$$

Тогда

$$a_{ij} = \sum_{k=1}^n l_{ik} l_{kj}^T = \sum_{k=1}^n l_{ik} l_{jk} = \sum_{k=1}^{j-1} l_{ik} l_{jk} + l_{ij} l_{jj} + \sum_{k=j+1}^n l_{ik} l_{jk} \stackrel{0}{=} \sum_{k=1}^{j-1} l_{ik} l_{jk} + l_{ij} l_{jj}.$$

При  $i = j$  находим, что

$$l_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2}, \quad j = 1, \dots, n. \quad (1.33)$$

Далее,

$$l_{ij} = \frac{1}{l_{jj}} \left[ a_{ij} - \sum_{k=1}^{j-1} l_{ik} l_{jk} \right], \quad i = j+1, \dots, n, \quad j = 1, \dots, n-1. \quad (1.34)$$

Вычисления можно вести по столбцам  $j = 1, \dots, n$  для  $i = j + 1, \dots, n$ .

$$j = 1: \quad l_{11} = \sqrt{a_{11}}, \quad l_{i1} = \frac{a_{i1}}{l_{11}}, \quad i = 2, \dots, n$$

$$j = 2: \quad l_{22} = \sqrt{a_{22} - l_{21}^2}, \quad l_{i2} = \frac{1}{l_{22}} [a_{i2} - l_{i1}l_{21}], \quad i = 3, \dots, n$$

и т.д.

Осталось доказать, что все  $l_{jj}$  положительны, т.е. положительны подкоренные выражения. Докажем, что

$$l_{jj} = \sqrt{\Delta_j / \Delta_{j-1}}, \quad \text{где } \Delta_0 = 1.$$

Пусть, как раньше,

$$A = \begin{bmatrix} A_j & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad L = \begin{bmatrix} L_j & 0 \\ L_{21} & L_{22} \end{bmatrix}.$$

Тогда

$$A_j = L_j L_j^T.$$

Отсюда

$$\Delta_j = \det A_j = (\det L_j)^2 = \left( \prod_{k=1}^j l_{kk} \right)^2.$$

Аналогично

$$\Delta_{j-1} = \left( \prod_{k=1}^{j-1} l_{kk} \right)^2$$

и, следовательно,

$$l_{jj}^2 = \Delta_j / \Delta_{j-1} > 0, \quad j = 1, \dots, n.$$

Теорема доказана.

**Упражнение 1.2.** Показать, что для реализации формул (1.33), (1.34) при всех  $i$  и  $j$  требуется

$$Q = \frac{n(n+1)(n+2)}{6} \approx \frac{n^3}{6} \tag{1.35}$$

операций умножения, деления и извлечения корня.

**Замечание 1.7.** Из (1.35) следует, что разложение Холецкого в два раза более экономично, чем треугольное разложение.

Обратимся теперь к решению системы (1.29). Поскольку  $Ax = LL^T x = b$ , то, полагая  $L^T x = y$ , получим  $Ly = b$ . При этом

$$\begin{aligned} y_i &= \frac{1}{l_{ii}} \left[ b_i - \sum_{k=1}^{i-1} l_{ik} y_k \right], \quad i = 1, \dots, n, \\ x_i &= \frac{1}{l_{ii}} \left[ y_i - \sum_{k=i+1}^n l_{ki} x_k \right], \quad i = n, \dots, 1. \end{aligned} \tag{1.36}$$

(Ср. с (1.21), (1.22)).

**Замечание 1.8.** Очень часто используется модификация разложения Холецкого, называемая  $LDL^T$ -разложением. Суть ее в том, что вместо разложения (1.31) строится разложение матрицы  $A$  вида

$$A = LDL^T,$$

где  $L$  — попережнему нижняя треугольная матрица, но в отличие от (1.31) ее диагональные элементы равны 1, а  $D$  — диагональная матрица. Достоинство  $LDL^T$ -разложения состоит в том, что при его вычислении не требуется находить квадратные корни, а потому оно существует не только для положительно определенных матриц. Условием существования такого разложения для симметричной матрицы является отличие от нуля всех ее угловых миноров.

**Упражнение 1.3.** Показать, что для элементов матриц  $L$  и  $D$  из  $LDL^T$ -разложения имеют место соотношения

$$\begin{aligned} d_j &= a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2 d_k, \quad j = 1, 2, \dots, n, \\ l_{ij} &= \left[ a_{ij} - \sum_{k=1}^{j-1} l_{ik} d_k l_{jk} \right] / d_j = \left[ a_{ij} - \sum_{k=1}^{j-1} l_{ik} m_{jk} \right] / d_j, \quad \begin{aligned} i &= j+1, \dots, n, \\ j &= 1, \dots, n-1, \end{aligned} \\ m_{jk} &= l_{jk} d_k, \quad k = 1, \dots, j-1. \end{aligned} \tag{1.37}$$

## 1.4 Обращение матрицы

Опишем одну из возможных процедур вычисления  $A^{-1}$ , основанную на использовании  $LU$ -разложения. Пусть  $A = LU$ . Тогда  $A^{-1} = U^{-1}L^{-1}$  или

$$UA^{-1} = L^{-1}.$$

Воспользуемся этим соотношением для вычисления  $A^{-1}$ . Допустим сначала, что  $L^{-1}$  — известна. Обозначим  $A^{-1} = X$ ,  $L^{-1} = Y$ . Пусть  $x_{.j}$  и  $y_{.j}$  —  $j$ -е столбцы матриц  $X$  и  $Y$ , соответственно, т.е.  $X = [x_{.1}x_{.2}\dots x_{.n}]$ ,  $x_{.j} = [x_{1j}x_{2j}\dots x_{nj}]^T$ . Тогда получим  $n$  систем вида

$$Ux_{.j} = y_{.j}, \quad j = 1, \dots, n \quad (1.38)$$

с треугольной матрицей, решения которых могут быть найдены по формулам (1.22) и

$$x_{kj} = \frac{1}{u_{kk}} \left[ y_{kj} - \sum_{m=k+1}^n u_{km}x_{mj} \right], \quad k = n, n-1, \dots, 1.$$

Найдем теперь  $L^{-1} = Y$ . Поскольку  $LY = I$ , то имеем  $n$  систем

$$Ly_{.j} = e_j, \quad j = 1, \dots, n. \quad (1.39)$$

Заметим, что матрица  $L^{-1}$  — нижняя треугольная, и поэтому у столбца  $y_{.j}$  первые  $j$  элемента известны и равны нулю для  $i = 1, \dots, j-1$  и единице для  $i = j$ , т.е.  $y_{.j}^T = [0 \dots 0 \ 1 \ y_{j+1j} \dots y_{nj}] = [0^T \ 1 \ \bar{y}_{.j}^T]$ . Отсюда следует, что для отыскания истинных неизвестных вектора  $y_{.j}$  нужно решить систему с треугольной матрицей размеров  $(n-j) \times (n-j)$  относительно  $\bar{y}_{.j}$ . Для этого можно воспользоваться формулами типа (1.22).

Оценим объем работы по вычислению  $A^{-1}$ . В силу (1.23) факторизация  $A = LU$  требует  $\approx n^3/3$  умножений, решение одной системы (1.38) (см. (1.25)) —  $\approx n^2/2$ , а всех  $\approx n^3/2$ , решение всех систем (1.39)

$$\sum_{j=1}^n \frac{(n-j+1)^2}{2} = \frac{1}{2} \sum_{j=1}^n j^2 = \frac{n(n+1)(2n+1)}{12} \approx n^3/6.$$

Складывая, находим, что для вычисления матрицы  $A^{-1}$  при помощи описанного алгоритма требуется  $\approx n^3$  умножений. Это всего лишь в три раза больше, чем для решения системы (1.29).

## 2

# Методы $QR$ -факторизации

Основная идея  $QR$ -факторизации снова состоит в сведении линейной системы к треугольной. Однако на этот раз матрица раскладывается в произведение не верхней и нижней треугольных матриц, как это было при  $LU$ -разложении или разложении Холецкого, а в произведение верхней треугольной матрицы  $R$  и ортогональной матрицы  $Q$ , которая тоже легко обращается, ибо  $Q^{-1} = Q^T$ .

Для того, чтобы решить линейную систему

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

мы должны сделать три шага:

1° выполнить факторизацию  $A$ , т.е. найти такую ортогональную матрицу  $Q$ , для которой  $Q^T A = R$  есть верхняя треугольная

2° вычислить  $Q^T b$

3° выполнить обратную подстановку, т.е. решить треугольную систему  $Rx = Q^T b$ .

Имеет место

**Теорема 2.1** (О  $QR$ -факторизации). *Пусть  $A$  есть действительная невырожденная матрица. Тогда существует единственная пара  $(Q, R)$ , где  $Q$  — ортогональная матрица, а  $R$  — верхняя треугольная с положительными элементами на главной диагонали, такая, что*

$$A = QR.$$

**Доказательство.** Здесь мы докажем только единственность разложения, а доказательство существования будет осуществлено в разделе 2.1 путем конструктивного построения матриц  $Q$  и  $R$ .

Для доказательства единственности этой факторизации мы предположим, что существуют две факторизации

$$A = Q_1 R_1 = Q_2 R_2.$$

Тогда  $Q_2^T Q_1 = R_2 R_1^{-1}$  есть верхняя треугольная матрица с положительными диагональными элементами (см. лемму 1.1). Пусть  $R = R_2 R_1^{-1}$ . Тогда

$$R R^T = (Q_2^T Q_1)(Q_2^T Q_1)^T = I,$$

т.е.  $R$  есть множитель Холецкого при факторизации единичной матрицы, который определяется однозначно, и, следовательно,  $R = I$ . Поэтому  $R_1 = R_2$  и  $Q_1 = Q_2$ . Единственность факторизации доказана.

**Замечание 2.1.** В приведенном выше доказательстве единственности  $QR$ -факторизации принципиальным было предположение о положительности диагональных элементов  $R$ . Исследуем множественность разложений при отказе от указанного предположения. Рассмотрим две  $QR$ -факторизации некоторой невырожденной матрицы

$$A = Q_1 R_1 = Q_2 R_2.$$

И в этом случае для  $R = R_2 R_1^{-1}$  справедливо равенство  $R R^T = I$ , откуда следует, что  $R^{-1} = R^T$ , т.е.  $R = D = R^T$ , где  $D$  — диагональная матрица. Но тогда  $R^2 = D^2 = I$  и  $d_{ii} = \pm 1$ . Поскольку  $Q_2^T Q_1 = R_2 R_1^{-1} = R = D$ , то

$$R_2 = D R_1, \quad Q_1 = Q_2 D, \quad (Q_2 = Q_1 D).$$

Другими словами,  $QR$ -факторизация действительной невырожденной матрицы единственна с точностью до умножения каждого  $k$ -го столбца  $Q$  и каждой  $k$ -ой строки  $R$  на множитель  $d_k = \pm 1$ . В комплексном случае таким множителем является комплексное число  $e^{il}$ , где  $l$  — действительное число.

## 2.1 Метод вращений Гивенса

Построим метод вращений решения системы (2.1), который позволяет получить представление матрицы  $A$  в виде произведения ортогональной матрицы  $Q$  и верхней треугольной матрицы  $R$ . Это построение завершит доказательство теоремы 2.1 и в части существования  $QR$ -разложения.





слева на матрицу  $T_{12}$  — матрицу вращения, имеющую вид

$$T_{12} = \begin{bmatrix} c & s & & & \\ -s & c & & & \\ & & 1 & & \\ & 0 & & \ddots & \\ & & & & 1 \end{bmatrix}.$$

Для исключения неизвестного  $x_1$  из третьего уравнения системы (2.4) новое первое уравнение и третье уравнение, которое пока не было затронуто преобразованиями, заменяются их линейными комбинациями с новыми коэффициентами  $c$ ,  $s$  и  $-s$ ,  $c$ , соответственно. Будем их теперь обозначать через  $c_{13}$  и  $s_{13}$ , а решению (2.3) системы (2.2) присвоим обозначения  $c_{12}$  и  $s_{12}$ . При этом коэффициенты  $c_{13}$  и  $s_{13}$  находятся из системы

$$-s_{13}a_{11}^{(1)} + c_{13}a_{31} = 0, \quad c_{13}^2 + s_{13}^2 = 1 \quad (2.6)$$

и суть

$$c_{13} = \frac{a_{11}^{(1)}}{\sqrt{(a_{11}^{(1)})^2 + a_{31}^2}}, \quad s_{13} = \frac{a_{31}}{\sqrt{(a_{11}^{(1)})^2 + a_{31}^2}},$$

а коэффициенты преобразованных первого и третьего уравнений находятся по формулам

$$\begin{aligned} a_{1j}^{(2)} &= c_{13}a_{1j}^{(1)} + s_{13}a_{3j}, & a_{3j}^{(1)} &= -s_{13}a_{1j}^{(1)} + c_{13}a_{3j}, & j &= 2, 3, \dots, n, \\ b_1^{(2)} &= c_{13}b_1^{(1)} + s_{13}b_3, & b_3^{(1)} &= -s_{13}b_1^{(1)} + c_{13}b_3. \end{aligned}$$

Это преобразование эквивалентно умножению слева матрицы системы (2.4) и ее вектора правой части на матрицу

$$T_{13} = \begin{bmatrix} c_{13} & 0 & s_{13} & & \\ 0 & 1 & 0 & & 0 \\ -s_{13} & 0 & c_{13} & & \\ & & & 1 & \\ & 0 & & & \ddots \\ & & & & & 1 \end{bmatrix}$$

и приводит к тому, что коэффициент при  $x_1$  в преобразованном третьем уравнении обращается в нуль.





а

$$c_{kl} = \frac{a_{kk}^{(l-2)}}{\sqrt{\left(a_{kk}^{(l-2)}\right)^2 + \left(a_{lk}^{(k-1)}\right)^2}}, \quad s_{kl} = \frac{a_{lk}^{(k-1)}}{\sqrt{\left(a_{kk}^{(l-2)}\right)^2 + \left(a_{lk}^{(k-1)}\right)^2}}. \quad (2.9)$$

В матричной записи полученная система имеет вид

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

где

$$A^{(n-1)} = T_{n-1}A^{(n-2)}, \quad b^{(n-1)} = T_{n-1}b^{(n-2)}, \quad T_{n-1} = T_{n-1n}.$$

Обозначим через  $R$  полученную верхнюю треугольную матрицу  $A^{(n-1)}$ . Она связана с исходной матрицей  $A$  равенством

$$R = TA, \quad (2.10)$$

где  $T = T_{n-1}T_{n-2} \dots T_1$ .

**Замечание 2.3.** В силу соображений, аналогичных изложенным в замечании 2.2, все элементы  $a_{kk}^{(n-1)}$ ,  $k = 1, 2, \dots, n$  главной диагонали матрицы  $R$  положительны.

**Замечание 2.4.** Действие матрицы  $T_{kl}^T$  на вектор  $x$  эквивалентно его повороту против хода часовой стрелки в координатной плоскости  $Ox_kx_l$  на угол  $\varphi_{kl}$  такой, что

$$\cos \varphi_{kl} = c_{kl}, \quad \sin \varphi_{kl} = s_{kl}.$$

Существование такого угла гарантируется соотношениями (2.9). Эта геометрическая интерпретация и дала название методу.

С учетом (2.2), (2.6) и т.д. легко видеть, что

$$T_{kl}T_{kl}^T = I,$$

т.е. матрицы  $T_{kl}$  ортогональные. Произведение ортогональных матриц есть матрица ортогональная, и поэтому  $T$  есть ортогональная матрица, равно как и  $T^T = T^{-1} = Q$ . Отсюда и из (2.10) находим, что

$$A = QR, \quad (2.11)$$

где  $Q$  — ортогональная матрица, а  $R$  — верхняя треугольная с положительными элементами на главной диагонали.

**Упражнение 2.3.** Показать, что для построения разложения (2.11) с использованием формул (2.8), (2.9), требуется  $\approx 4n^3/3$  действий умножения.

**Замечание 2.5.** Принимая во внимание результаты вычислений из упражнения 2.3, заключаем, что метод вращений примерно в четыре раза более трудоемок, чем метод Гаусса.

## 2.2 Метод отражений Хаусхолдера

Рассмотрим еще один метод, который дает разложение матрицы  $A$  в произведение ортогональной и верхней треугольной матриц. Это будет *метод отражений*.

Пусть  $w$  — некоторый вектор (столбец) единичной длины

$$\|w\|_2^2 = (w, w) = w^T w = 1. \quad (2.12)$$

Введем в рассмотрение матрицу

$$U = I - 2ww^T, \quad (2.13)$$

которую назовем *матрицей отражения* или *матрицей Хаусхолдера*, и изучим ее свойства.

1°. Матрица  $U$  симметрична, т.е.

$$U = U^T. \quad (2.14)$$

В самом деле, так как

$$(ww^T)^T = (w^T)^T w^T = ww^T,$$

то  $ww^T$  есть симметричная матрица, а в силу (2.13) вместе с ней симметричной является и матрица  $U$ .

2°. Матрица  $U$  есть ортогональная матрица, т.е.

$$U^{-1} = U^T. \quad (2.15)$$

Принимая во внимание (2.14), (2.13) и (2.12), имеем

$$UU^T = UU = (I - 2ww^T)(I - 2ww^T) = I - 4ww^T + 4w \underbrace{w^T w}_1 w^T = I,$$

что и означает справедливость (2.15).

3°. Число  $\lambda = -1$  является однократным собственным значением матрицы  $U$ , которому отвечает собственный вектор  $w$  из (2.13). Число  $\lambda = 1$  является  $(n - 1)$ -кратным собственным значением матрицы  $U$ , которому отвечает  $(n - 1)$ -мерное собственное подпространство, состоящее из всех векторов  $v$ , ортогональных  $w$ .

В силу (2.14), (2.15) имеем  $U^2 = UU^T = I$ . Так как все собственные значения матрицы  $I$  равны 1, то собственные значения матрицы  $U$  удовлетворяют соотношению  $\lambda_U^2 = 1$ , и, следовательно, равны либо  $+1$ , либо  $-1$ .

Далее, принимая во внимание (2.13), (2.12), находим, что

$$Uw = (I - 2ww^T)w = w - 2w \underbrace{w^T w}_{\parallel 1} = -w. \quad (2.16)$$

Наконец, пусть

$$(v, w) = w^T v = 0.$$

Тогда

$$Uv = (I - 2ww^T)v = v - 2w \underbrace{w^T v}_{\parallel 0} = v, \quad (2.17)$$

что и требовалось доказать.

4°. Вектор  $Uy$  есть зеркальное отражение вектора  $y$  относительно плоскости, ортогональной вектору  $w$ .

Пусть  $y$  — произвольный вектор. Представим его в виде

$$y = z + v, \quad (2.18)$$

где  $z$  — проекция  $y$  на  $w$ , т.е.  $z = (w, y)w$ , а вектор  $v$  ортогонален  $w$ :  $(w, v) = 0$ . Принимая во внимание (2.16), (2.17), находим, что

$$Uy = U(z + v) = -z + v \quad (2.19)$$

(см. рис.1)

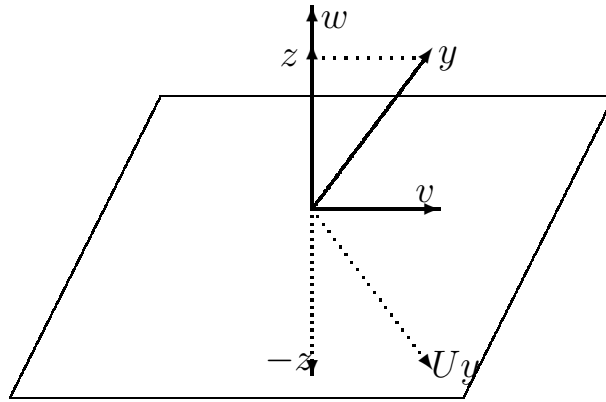


Рис. 1

5°. Векторы  $y - Uy$  и  $y + Uy$  ортогональны друг другу. При этом, если  $y = z + v$ , как в (2.18), то

$$y - Uy = 2z, \quad z \parallel w, \quad (2.20)$$

$$y + Uy = 2v, \quad v \perp w. \quad (2.21)$$

Утверждения следуют из (2.18), (2.19).

6°. Пусть  $x$  и  $y$  — произвольные векторы одинаковой длины. Тогда, если

$$w = \frac{y \pm x}{\|y \pm x\|_2}, \quad (2.22)$$

то матрица отражения  $U$ , построенная по вектору  $w$ , переводит вектор  $y$  в вектор, коллинеарный вектору  $x$ .

В самом деле, пусть матрица отражения  $U$  обладает указанным свойством, т.е.  $Uy = -\sigma x$ . Найдем вектор  $w$ , образующий эту матрицу. Согласно свойству 5° (см. (2.20)) вектор  $y - Uy$  коллинеарен вектору  $w$ , а поскольку  $\|w\|_2 = 1$ , то

$$w = \frac{y - Uy}{\|y - Uy\|_2} = \frac{y + \sigma x}{\|y + \sigma x\|_2}.$$

В силу 2° матрица  $U$  ортогональна и поэтому

$$\|Uy\|_2 = \|y\|_2 := \|x\|_2 = |\sigma| \|x\|_2,$$

т.е.  $\sigma = \pm 1$ , что и приводит к (2.22)

**Замечание 2.6.** Если  $y = x$ , то при использовании в (2.22) нижнего знака ( $-$ ) знаменатель обратится в нуль. Это же произойдет при выборе верхнего знака ( $+$ ), если  $y = -x$ . Чтобы избежать этих неприятностей,



целесообразно знак в (2.22) выбирать совпадающим со знаком  $\text{sign}(x, y)$ , т.е. полагать

$$w = \frac{y + \text{sign}(x, y) x}{\|y + \text{sign}(x, y) x\|_2}. \quad (2.23)$$

Если вектор  $y$  не равен  $\pm x$ , а лишь близок к одному из них, то и в этом случае при выборе (2.23) мы будем избавлены от неприятностей, связанных с делением на малое число.

Воспользуемся матрицей отражения для приведения квадратной матрицы к треугольному виду. На первом шаге приведения рассмотрим в качестве вектора  $y$  из свойства 6° первый столбец матрицы  $A$

$$y_1 = [a_{11} \ a_{21} \ \dots \ a_{n1}]^T, \quad (2.24)$$

а в качестве  $x$  — вектор, коллинеарный  $e_1 = [1 \ 0 \ \dots \ 0]^T$ . Если  $a_{21} = a_{31} = \dots = a_{n1} = 0$ , то переходим к следующему шагу, положив  $A^{(1)} = A$ ,  $U_1 = I$  и введя обозначения  $a_{ij}^{(1)} = a_{ij}$ . В противном случае умножим матрицу  $A$  слева на матрицу отражения

$$U_1 = I - 2w_1 w_1^T = I_n - 2w_1 w_1^T, \quad (2.25)$$

где вектор  $w_1$  вычисляется согласно формуле (2.23)

$$w_1 = \frac{y_1 + \text{sign } a_{11} \|y_1\|_2 e_1}{\|y_1 + \text{sign } a_{11} \|y_1\|_2 e_1\|_2}. \quad (2.26)$$

В результате получим матрицу

$$A^{(1)} = U_1 A,$$

в первом столбце которой стоят нули во всех позициях, кроме первой. Этим заканчивается первый этап.

Пусть мы уже осуществили  $l \geq 1$  шагов и пришли к матрице  $A^{(l)}$  с элементами  $a_{ij}^{(l)}$  такими, что  $a_{ij}^{(l)} = 0$  при  $i > j$ ,  $j = 1, \dots, l$ . В пространстве  $\mathbb{R}_{n-l}$  векторов размерности  $n - l$  рассмотрим вектор

$$y_{l+1} = [a_{l+1 \ l+1}^{(l)} \ \dots \ a_{n \ l+1}^{(l)}]^T.$$

Если  $a_{l+2 \ l+1}^{(l)} = \dots = a_{n \ l+1}^{(l)} = 0$ , то переходим к следующему шагу, положив

$$A^{(l+1)} = A^{(l)}, \quad U_{l+1} = I.$$

В противном случае строим матрицу отражения

$$V_{l+1} = I_{n-l} - 2w_{l+1}w_{l+1}^T$$

(размеры матрицы  $V_{l+1}$  и вектора  $w_{l+1}$  равны  $(n-l)$ ), переводящую вектор  $y_{l+1}$  в вектор, коллинеарный  $e_{l+1} = [1\ 0\ \dots\ 0]^T \in \mathbb{R}_{n-l}$ , и переходим к матрице

$$A^{(l+1)} = U_{l+1}A^{(l)},$$

где

$$U_{l+1} = \begin{bmatrix} I_l & 0 \\ 0 & V_{l+1} \end{bmatrix}. \quad (2.27)$$

Если матрицу  $A^{(l)}$  записать в блочно-треугольном виде

$$A^{(l)} = \begin{bmatrix} A_{11}^{(l)} & A_{12}^{(l)} \\ 0 & A_{22}^{(l)} \end{bmatrix},$$

где  $A_{11}^{(l)}$  — верхняя треугольная матрица размера  $l \times l$ , то будем иметь

$$A^{(l+1)} = \begin{bmatrix} I_l & 0 \\ 0 & V_{l+1} \end{bmatrix} \begin{bmatrix} A_{11}^{(l)} & A_{12}^{(l)} \\ 0 & A_{22}^{(l)} \end{bmatrix} = \begin{bmatrix} A_{11}^{(l)} & A_{12}^{(l)} \\ 0 & V_{l+1}A_{22}^{(l)} \end{bmatrix} = \begin{bmatrix} A_{11}^{(l+1)} & A_{12}^{(l+1)} \\ 0 & A_{22}^{(l+1)} \end{bmatrix}, \quad (2.28)$$

где  $A_{11}^{(l+1)}$  — верхняя треугольная матрица размера  $(l+1) \times (l+1)$ . После  $(n-1)$  шагов мы приходим к матрице

$$A^{(n-1)} = U_{n-1}U_{n-2}\dots U_1A,$$

имеющей верхнюю треугольную форму. Обозначим

$$U_{n-1}\dots U_1 = U.$$

Тогда

$$A^{(n-1)} = UA, \quad A = U^T A^{(n-1)}.$$

Если нужно решить систему (2.1), то после описанных преобразований приходим к эквивалентной системе

$$A^{(n-1)}x = Ub \quad (2.29)$$

с треугольной матрицей.

Оценим трудоемкость  $QR$ -факторизации матрицы  $A$  методом отражений. Заметим сначала, что при построении метода отражений не использовались никакие предположения о невырожденности  $A$ . Поэтому

метод будет работать и в том случае, когда матрицей  $A$  является прямоугольная матрица, образованная  $p \leq n$  столбцами высоты  $n$ . (Эта матрица становится квадратной, если к ней добавить недостающие  $(n-p)$  (нулевых) столбцов). Принимая это во внимание, оценим трудоемкость  $QR$ -факторизации прямоугольной матрицы, образованной  $p$  столбцами  $a_j$  и имеющей вид

$$A = [a_1 \ a_2 \ \dots \ a_p]$$

Первый шаг метода отражений состоит в умножении матрицы отражений  $U_1$  на матрицу  $A$ . В общем случае такое перемножение требует  $n^2p$  умножений. Однако в нашем случае одна из матриц (матрица  $U_1$ ) имеет специальную структуру (2.13), и умножение на такую матрицу можно выполнить более эффективно. Именно,

$$U_1 A = (I_n - 2w_1 w_1^T) A = A - 2w_1 (w_1^T A) = A - 2w_1 [w_1^T a_1 \ \dots \ w_1^T a_p].$$

Скалярное произведение вектора  $w_1$  с одним из столбцов  $a_j$  требует  $n$  умножений, а общее число умножений для нахождения  $(w_1^T A)$  равно  $np$ . Построение матрицы

$$(2w_1)(w_1^T A)$$

требует еще  $np$  умножений. Само построение по формуле (2.26) вектора  $w_1$  и умножение его на 2 требует еще  $O(n)$  умножений, но при больших  $n$  и  $p$  этими трудозатратами можно пренебречь. Итак, на первом шаге метода отражений требуется приблизительно  $2np$  умножений. Принимая во внимание соотношения (2.28), находим, что при построении матрицы  $A^{(l+1)}$  реально требуется перемножение только матрицы отражений  $V_{l+1}$  размера  $(n-l) \times (n-l)$  с прямоугольной матрицей  $A_{22}^{(l)}$  размера  $(n-l) \times (p-l)$ . Тем самым, для построения матрицы  $A^{(l+1)}$  требуется приблизительно

$$2(n-l)(p-l)$$

умножений. Суммируя теперь все трудозатраты, находим, что трудоемкость построения  $QR$ -факторизации методом отражений приблизительно есть

$$2 \sum_{l=0}^{p-1} (n-l)(p-l) \approx np^2 - \frac{1}{3}p^3. \quad (2.30)$$

Тем самым, трудоемкость  $QR$ -факторизации квадратной матрицы методом отражений приблизительно равна

$$2n^3/3,$$

что в два раза меньше трудоемкости метода вращений и в два раза больше трудоемкости  $LU$ -факторизации.







**Пример 3.5.** У матрицы на рис. 5

$$\begin{bmatrix} * & 0 & * & & & \\ 0 & * & 0 & * & & \\ * & 0 & * & 0 & * & \\ & * & 0 & * & 0 & * \\ & & * & 0 & * & 0 \\ & & & * & 0 & * \end{bmatrix}$$

Рис. 5.

полуширина  $p = 2$ , ширина 5 и всего три ненулевых диагонали.

**Определение 3.2.** Рисунок, на котором (звездочками) отмечены позиции, где только и могут располагаться ненулевые элементы матрицы  $A$ , называется ее портретом.

### 3.3 Ленточный вариант треугольного разложения

Модифицируем алгоритм исключения Гаусса на случай ленточных матриц, т.е. заранее отбросим те вычисления, которые заведомо приводят к нулевым элементам. Это позволит нам сэкономить в трудозатратах на решение системы. Обратимся сразу к варианту, основанному на треугольном разложении матрицы  $A$ .

Нам потребуется

**Лемма 3.1.** Если полуширина матрицы  $A$  равна  $p$ , то и в треугольном разложении  $A = LU$  полуширина  $L$  ( $U$ ) равна  $p$ .

**Доказательство.** По условию леммы

$$a_{ij} = 0 \quad \text{при} \quad |i - j| > p. \quad (3.7)$$

Докажем, что

$$l_{ij} = 0 \quad \text{при} \quad i - j > p. \quad (3.8)$$

Для доказательства применим метод полной математической индукции по номерам столбцов матрицы  $L$ . При  $j = 1$  из (1.19) находим, что

$$l_{i1} = a_{i1}/u_{11}, \quad i = 2, \dots, n,$$

а если принять во внимание (3.7), то это соотношение примет вид

$$l_{i1} = \begin{cases} a_{i1}/u_{11}, & i = 2, \dots, p + 1, \\ 0, & i = p + 2, \dots, n. \end{cases}$$



Отсюда и следует (3.8) для  $j = 1$ .

Пусть теперь утверждение (3.8) верно для столбцов матрицы  $L$  с номерами  $k = 1, 2, \dots, j - 1$ , т.е.

$$l_{ik} = 0 \quad \text{при} \quad i - k > p, \quad k = 1, 2, \dots, j - 1. \quad (3.9)$$

Докажем справедливость (3.8) для  $j$ -ого столбца. В силу (3.7) из (1.19) следует, что при  $i - j > p$

$$l_{ij} = -\frac{1}{u_{jj}} \sum_{k=1}^{j-1} l_{ik} u_{kj}, \quad i - j > p. \quad (3.10)$$

Оценим разность индексов  $i - k$  у первых сомножителей произведений, стоящих под знаком суммы в (3.10). Принимая во внимание ограничения на  $i$  и  $k$ , диктуемые (3.10), будем иметь

$$i - k > p + j - k \geq p + 1.$$

Но тогда в силу (3.9) первые сомножители в сумме (3.10) обращаются в нуль и соотношение (3.8) установлено.

В силу (3.8) из (1.19) находим, что

$$l_{j+p,j} = a_{j+p,j}/u_{jj}.$$

и, следовательно, полуширина  $L$  совпадает с полушириной  $A$ . Для матрицы  $U$  доказательство аналогичное. Лемма доказана.

Преобразуем формулы (1.19)-(1.22) на случай ленточной матрицы  $A$ . Сначала выясним, для каких значений индексов  $i$  и  $j$  нужно проводить вычисления по формулам (1.19). Так как в силу леммы 3.1

$$u_{ij} = 0 \quad \text{при} \quad j - i > p, \quad (3.11)$$

то элементы  $u_{ij}$ , которые только и нужно вычислять, имеют индексы, подчиненные условию  $j - i \leq p$ , т.е.

$$j = i, i + 1, \dots, \min [n, p + i].$$

Аналогично

$$l_{ij} = 0 \quad \text{при} \quad i - j > p \quad (3.12)$$

и  $l_{ij}$  нужно вычислять только для

$$i = j + 1, j + 2, \dots, \min [n, p + j].$$

Теперь преобразуем суммы в (1.19). В силу (3.11) ненулевые слагаемые в суммах (1.19) могут быть только при  $j - k \leq p$ , т.е. при

$$k \geq j - p,$$

а в силу (3.12) — только при  $i - k \leq p$ , т.е. при

$$k \geq i - p.$$

Объединяя эти неравенства и принимая во внимание, что  $k$  — натуральное, будем иметь

$$k \geq \max[1, i - p, j - p].$$

Поскольку в формулах для  $u_{ij}$  индексы  $i, j$  подчинены ограничению  $j \geq i$ , то в этих формулах

$$k \geq \max[1, j - p].$$

В формулах же для  $l_{ij}$  наоборот  $i > j$  и поэтому в них

$$k \geq \max[1, i - p].$$

С учетом сказанного, для ленточной матрицы  $A$  с полушириной  $p$  формулы (1.19) принимают вид

$$\begin{aligned} u_{ij} &= \left[ a_{ij} - \sum_{k=\max[1, j-p]}^{i-1} l_{ik} u_{kj} \right] & i &= 1, \dots, n, \\ & & j &= i, \dots, \min[n, p+i], \\ l_{ij} &= \frac{1}{u_{jj}} \left[ a_{ij} - \sum_{k=\max[1, i-p]}^{j-1} l_{ik} u_{kj} \right] & j &= 1, \dots, n, \\ & & i &= j+1, \dots, \min[n, p+j]. \end{aligned} \quad (3.13)$$

Преобразуем формулы (1.21) и (1.22). В формулах (1.21) в силу (3.12)  $i - k \leq p$ , а в формулах (1.22) в силу (3.11)  $j - k \leq p$ . Поэтому

$$\begin{aligned} y_i &= b_i - \sum_{k=\max[1, i-p]}^{i-1} l_{ik} y_k, & i &= 1, \dots, n, \\ x_k &= \frac{1}{u_{kk}} \left[ y_k - \sum_{j=k+1}^{\min[n, k+p]} u_{kj} x_j \right], & k &= n, \dots, 1. \end{aligned} \quad (3.14)$$



Здесь  $Q$  — трудоемкость, нижний индекс — номер формулы, а аргумент — объект вычислений.

Далее

$$\begin{aligned} Q_{16}(u_{ij}) &= i - 1 - j + p + 1 = i - j + p, \\ Q_{16}(U) &= \sum_{j=p+1}^n \sum_{i=j-p}^j (i - j + p), \\ Q_{18}(l_{ij}) &= j - 1 - i + p + 1 + 1 = j - i + p + 1, \\ Q_{18}(L) &= \sum_{i=p+1}^n \sum_{j=i-p}^{i-1} (j - i + p + 1) = \sum_{j=p+1}^n \sum_{i=j-p}^{j-1} (i - j + p + 1), \end{aligned}$$

и, следовательно,

$$\begin{aligned} Q_{16}(U) + Q_{18}(L) &= \sum_{j=p+1}^n \left[ p + \sum_{i=j-p}^{j-1} (i - j + p + i - j + p + 1) \right] = \\ &= 2 \sum_{j=p+1}^n \left[ p + \sum_{k=1}^{p-1} k \right] = (n - p)p(p - 1). \end{aligned}$$

Отсюда и из (3.19) находим, что общее число умножений и делений при построении  $LU$ -разложения есть

$$Q = \frac{p(p+1)}{3}(3n - 2p - 1) = n(p^2 + p) - \frac{2}{3}p^3 + O(p^2). \quad (3.20)$$

**Замечание 3.1.** Полуширина полной матрицы  $p = n - 1$ . Подставляя это значение  $p$  в найденное выражение, получим выражение для трудоемкости треугольного разложения, совпадающее с (1.23). Полагая же здесь  $p = 1$ , получим (3.5).

Обратимся к формулам (3.14).

$$\begin{aligned} Q_{14}(y_i) &= \begin{bmatrix} i - 1, & i = 1, \dots, p, \\ i - 1 - i + p + 1 = p, & i = p + 1, \dots, n, \end{bmatrix} \\ Q_{14}(y) &= \sum_{i=1}^n Q_{14}(y_i) = \frac{p(p-1)}{2} + p(n-p) = \frac{p(2n-p-1)}{2}, \\ Q_{14}(x) &= \frac{p(2n-p-1)}{2} + n \end{aligned}$$

и, следовательно,

$$q = Q_{14}(x) + Q_{14}(y) = p(2n - p - 1) + n = (2p + 1)n - p(p + 1) \quad (3.21)$$

(ср. с (1.25) при  $p = n - 1$  и (3.6) при  $p = 1$ ).

Проанализируем формулы (3.20), (3.21). Рассмотрим три случая.

1°.  $p = O(n)$ , например,  $p = \alpha n$ ,  $\alpha < 1$ . Тогда

$$Q \approx \alpha^2 n^3 - \frac{2}{3} \alpha^3 n^3 = \alpha^2 \left(1 - \frac{2\alpha}{3}\right) n^3.$$

Легко проверить, что при  $0 < \alpha < 1$  коэффициент при  $n^3$  меньше  $1/3$ .

2°.  $p = o(n)$ , но  $p \rightarrow \infty$  при  $n \rightarrow \infty$ . В этом случае

$$Q \approx p^2 n, \quad q \approx 2pn.$$

В частности, при  $p = \sqrt{n}$

$$Q \approx n^2, \quad q \approx 2n^{3/2}.$$

3°.  $p = \text{const}$  ( $p = 1, 2, \dots$ ).

$$Q \approx p(p+1)n, \quad q \approx (2p+1)n.$$

При  $p = 1$   $Q < q$ , при  $p \geq 2$   $Q > q$ .

**Упражнение 3.2.** Показать, что в обозначениях (3.2) соотношения (3.3), (3.4) совпадают с (3.13), (3.14).

Приведем пример важной системы, полуширина ленты матрицы которой есть  $O(\sqrt{n})$ .

**Пример 3.6.** Среди уравнений математической физики одним из важнейших является уравнение Пуассона, которое в декартовых координатах на плоскости  $Oxy$  имеет вид

$$\Delta u := \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = -f(x, y), \quad (x, y) \in \Omega \subset \mathbb{R}^2.$$

Одной из основных задач для этого уравнения является задача Дирихле, когда на границе  $\partial\Omega$  области  $\Omega$  задается значение искомой функции

$$u(x, y) \Big|_{\partial\Omega} = g(x, y).$$

Пусть областью  $\Omega$  является единичный квадрат  $\Omega = (0, 1)^2$ , а граничная функция  $g(x, y) = 0$ . Приближенное решение этой задачи будем искать методом конечных разностей. Для этого в области  $\Omega$  введем квадратную сетку с шагом  $h = 1/N$ , образованную точками пересечения двух семейств прямых:  $x = ih$ ,  $i = 1, \dots, N - 1$  и  $y = jh$ ,  $j = 1, \dots, N - 1$  (см. рис. 7). Координаты этих точек (узлов сетки) будем



### 3.5 Несимметричная ленточность

**Определение 3.3.** Говорят, что матрица  $A$  ленточная с лентой нижней полуширины  $p_1$  и верхней полуширины  $p_2$ , если  $a_{ij} = 0$  при  $i - j > p_1$  и  $j - i > p_2$ .

**Лемма 3.2.** Пусть матрица  $A$  имеет нижнюю полуширину  $p_1$  и верхнюю полуширину  $p_2$ . Тогда в треугольном разложении  $A = LU$  полуширина  $L$  равна  $p_1$ , а полуширина  $U$  —  $p_2$ .

Доказательство этой леммы почти полностью повторяет доказательство леммы 3.1.

**Упражнение 3.4.** Показать, что при несимметричной ленточности в разложении  $A = LU$  элементы матриц  $L$  и  $U$  вычисляются по формулам

$$\begin{aligned} u_{ij} &= \left[ a_{ij} - \sum_{k=\varkappa_{ij}}^{i-1} l_{ik} u_{kj} \right] & i = 1, \dots, n, \\ & & j = i, \dots, \min[n, p_2 + i], \\ l_{ij} &= \frac{1}{u_{ij}} \left[ a_{ij} - \sum_{k=\varkappa_{ij}}^{j-1} l_{ik} u_{kj} \right] & j = 1, \dots, n, \\ & & i = j + 1, \dots, \min[n, p_1 + j], \end{aligned} \quad (3.22)$$

где

$$\varkappa_{ij} = \max[1, i - p_1, j - p_2].$$

Для оценки трудоемкости  $LU$ -разложения в рассматриваемом случае соотношения (3.22) можно переписать в виде, аналогичном (3.15)-(3.18).

При  $p_2 \geq p_1$  будем иметь

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj}, \quad i = 1, \dots, p_1, \quad j = i, \dots, p_2,$$

$$u_{ij} = a_{ij} - \sum_{k=j-p_2}^{i-1} l_{ik}u_{kj}, \quad j = p_2 + 1, \dots, n, \quad i = j - p_2, \dots, j - p_2 + p_1 - 1,$$

$$u_{ij} = a_{ij} - \sum_{k=i-p_1}^{i-1} l_{ik}u_{kj}, \quad \begin{array}{l} i = p_1 + 1, \dots, p_2, \quad j = p_2 + 1, \dots, n, \\ j = i, \dots, p_2, \quad i = j - p_2 + p_1, \dots, j, \end{array}$$

$$l_{ij} = \frac{1}{u_{ij}} \left[ a_{ij} - \sum_{k=1}^{j-1} l_{ik}u_{kj} \right], \quad j = 1, \dots, p_1, \quad i = j + 1, \dots, p_1,$$

$$l_{ij} = \frac{1}{u_{ij}} \left[ a_{ij} - \sum_{k=i-p_1}^{j-1} l_{ik}u_{kj} \right], \quad i = p_1 + 1, \dots, n, \quad j = i - p_1, \dots, i - 1.$$

Число умножений и делений, необходимых для реализации этих формул, задается величиной

$$Q(p_1, p_2) = p_1 \left[ (p_2 + 1)n - \frac{p_1^2 + 3p_2^2 + 3p_1 + 3p_2 + 2}{6} \right], \quad p_1 \leq p_2,$$

которая при  $p_1 = p_2 = p$ , естественно, совпадает с (3.20). Особый интерес с точки зрения дальнейших приложений представляет случай  $p_2 = n - 1$ . В этом случае

$$Q(p_1, n - 1) = p_1 \left[ \frac{n(n + 1)}{2} - \frac{(p_1 + 1)(p_1 + 2)}{6} \right].$$

Несмотря на то, что при конечном  $p_1$  матрица  $A$  содержит  $O(n^2)$  ненулевых элементов, трудоемкость  $LU$ -разложения оценивается величиной  $O(n^2)$ , а не  $O(n^3)$ , как в общем случае.

**Определение 3.4.** Квадратная ленточная матрица с нижней полушириной  $p_1 = 1$  и верхней полушириной  $p_2 = n - 1$  называется верхней матрицей Хессенберга.

Матрицы Хессенберга играют важную роль в методах решения задачи на собственные значения. Трудоемкость факторизации матрицы Хессенберга есть

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



**Замечание 3.2.** Если  $p_1 > p_2$ , то

$$Q = (p_2 + 1) \left( p_1 n - \frac{3p_1^2 + p_2^2 + 3p_1 - p_2}{6} \right).$$

Тем самым, при любых  $p_1$  и  $p_2$  трудоемкость  $LU$ -разложения оценивается величиной  $O(p_1 p_2 n)$ .

### 3.6 Ленточный вариант метода Холецкого

Как и в случае треугольного разложения можно построить разложение Холецкого в ленточном варианте. Пусть  $A$  — симметричная положительно определенная ленточная матрица с полушириной  $p$ . Справедлива

**Лемма 3.3.** Если полуширина матрицы  $A = A^T > 0$  равна  $p$ , то и полуширина множителя Холецкого  $L$  равна  $p$ .

На доказательстве этой леммы мы не останавливаемся, ибо оно почти дословно повторяет доказательство аналогичной леммы 3.1.

В силу леммы 3.3 ненулевыми элементами  $l_{ik}$  матрицы  $L$  могут быть только те, у которых индексы подчинены условиям

$$j - p \leq k \leq j \quad (l_{jk} \neq 0). \quad (3.23)$$

Поэтому формула (1.33) принимает вид

$$l_{jj} = \sqrt{a_{jj} - \sum_{k=\max[1, j-p]}^{j-1} l_{jk}^2}, \quad j = 1, \dots, n. \quad (3.24)$$

В силу (3.23) в формулах (1.34)  $i - p \leq k \leq i$  и  $i - p \leq j \leq i$ . Поэтому формулы (1.34) принимают вид

$$l_{ij} = \frac{1}{l_{jj}} \left[ a_{ij} - \sum_{k=\max[1, i-p]}^{j-1} l_{ik} l_{jk} \right], \quad \begin{array}{l} i = j + 1, \dots, \min[n, p + j], \\ j = 1, \dots, n. \end{array} \quad (3.25)$$

Ленточный вариант разложения Холецкого построен.

Для отыскания решения системы (1.29) нужно еще преобразовать формулы (1.36). Они принимают вид

$$\begin{aligned} y_i &= \frac{1}{l_{ii}} \left[ b_i - \sum_{k=\max[1, i-p]}^{i-1} l_{ik} y_k \right], \quad i = 1, \dots, n, \\ x_i &= \frac{1}{l_{ii}} \left[ y_i - \sum_{k=i+1}^{\min[n, i+p]} l_{ki} x_k \right], \quad i = n, \dots, 1. \end{aligned} \quad (3.26)$$

Эти формулы полностью аналогичны формулам (3.14).

**Упражнение 3.5.** Показать, что число действий умножения, деления и извлечения корня, необходимых для реализации формул (3.24)-(3.25), есть

$$Q = \frac{(p+1)(p+2)(3n-2p)}{6}.$$

### 3.7 Метод блочного исключения (метод частичного исключения неизвестных)

В методе исключения Гаусса из системы (1.29) последовательно исключаются неизвестные — компоненты вектора  $x^T = [x_1, x_2, \dots, x_n]$ . В ряде случаев бывает полезным процедуру исключения неизвестных произвести блочно. Пусть

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad (3.27)$$

где  $A_{11}$  — квадратная невырожденная матрица размеров  $m \times m$ , а  $b_1$  и  $x_1$  —  $m$ -мерные векторы. С учетом (3.27) система (1.29) принимает вид

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

или после блочного перемножения

$$\begin{aligned} A_{11}x_1 + A_{12}x_2 &= b_1, \\ A_{21}x_1 + A_{22}x_2 &= b_2. \end{aligned} \quad (3.28)$$

Из первого уравнения (3.28) находим, что

$$x_1 = A_{11}^{-1}(b_1 - A_{12}x_2). \quad (3.29)$$

Подставляя это представление  $x_1$  во второе уравнение (3.28), получим

$$A_{21}A_{11}^{-1}(b_1 - A_{12}x_2) + A_{22}x_2 = b_2$$

или после преобразования

$$(A_{22} - A_{21}A_{11}^{-1}A_{12})x_2 = b_2 - A_{21}A_{11}^{-1}b_1.$$

В результате система (3.28) преобразовалась к системе

$$\begin{aligned} A_{11}x_1 + A_{12}x_2 &= b_1, \\ (A_{22} - A_{21}A_{11}^{-1}A_{12})x_2 &= b_2 - A_{21}A_{11}^{-1}b_1. \end{aligned} \quad (3.30)$$

(Неизвестные  $x_1$  исключены из второй группы уравнений).

Из (3.30) вроде бы следует, что для реализации блочного исключения нужно вычислять  $A_{11}^{-1}$ . На самом деле явно это делать вовсе не обязательно. Принимая во внимание (3.29), введем следующие обозначения:

$$A_{11}^{-1}b_1 = \overset{\circ}{x}_1, \quad A_{11}^{-1}A_{12} = Z_{12}. \quad (3.31)$$

Тогда вторая группа уравнений (3.30) примет вид

$$(A_{22} - A_{21}Z_{12})x_2 = (b_2 - A_{21}\overset{\circ}{x}_1). \quad (3.32)$$

Соотношения (3.31) можно переписать в виде систем уравнений

$$A_{11}\overset{\circ}{x}_1 = b_1, \quad A_{11}Z_{12} = A_{12}, \quad (3.33)$$

а из (3.29) и (3.31) находим, что

$$x_1 = \overset{\circ}{x}_1 - Z_{12}x_2. \quad (3.34)$$

Итак, чтобы найти решение системы (3.28) нужно:

- 1° решить  $(n - m + 1)$  систем (3.33) с матрицей  $A_{11}$  для отыскания вектора  $\overset{\circ}{x}_1$  и столбцов матрицы  $Z_{12}$ ,
- 2° по найденным  $\overset{\circ}{x}_1$  и  $Z_{12}$  сформировать матрицу и правую часть системы (3.32) и решить полученную систему — найти вектор  $x_2$ ,
- 3° найти вектор  $x_1$  по формулам (3.34).

**Замечание 3.3.** В трактовке (3.32),(3.33) метода блочного исключения фактически исключенными оказываются не неизвестные  $x_1$ , а неизвестные  $x_2$ . Из системы (3.28) как бы исключается часть неизвестных (именно  $x_2$ ), затем она решается относительно оставшихся неизвестных (3.33) (но не полностью — нужен еще шаг (3.34)) и лишь потом находится  $x_2$  из (3.32). Отсюда второе название метода — метод частичного исключения неизвестных (исключение  $x_2$ ).

**Пример 3.7.** Пусть матрица  $A$  имеет портрет, изображенный на рис. 9. Матрица  $A$  не является ленточной, хотя ее подматрица  $A_{11}$ , расположенная в первых  $(n - 1)$  строках и  $(n - 1)$  столбцах является ленточной с полушириной  $p = 2$ . Для решения системы (1.29) с такой матрицей не годится ленточный вариант исключения Гаусса, а применение общего метода требует  $O(n^3)$  умножений и делений. Но если можно применить алгоритм блочного исключения,

$$\begin{bmatrix} * & * & * & & & & & & & * \\ * & * & * & * & & & & & & * \\ * & * & * & * & * & & & & & * \\ & * & * & * & * & * & & & & * \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ & & & & & & * & * & * & * \\ * & * & * & * & * & \dots & * & * & * & * \end{bmatrix}$$

Рис. 9

то для решения двух систем (3.33) с пятидиагональными матрицами с использованием ленточного варианта исключения потребуется  $O(n)$  действий. Столько же действий потребуется для вычислений по формулам (3.32) и (3.34). В результате система будет решена за  $O(n)$  действий.

**Пример 3.8.** Матрица имеет портрет, изображенный на рис. 10.

$$\begin{bmatrix} * & * & * & * & * & * & \dots & * & * & * \\ * & * & * & & & & & & & * \\ * & * & * & * & & & & & & * \\ * & & * & * & * & & & & & * \\ * & & & * & * & * & & & & * \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ * & & & & & & & * & * & * \\ * & * & * & * & * & * & \dots & * & * & * \end{bmatrix}$$

Рис. 10

Переставляя первую строку на последнее место и то же делая с первым столбцом, получим матрицу с портретом, изображенным на рис. 11.

$$\begin{bmatrix} * & * & & & & & & & * & * \\ * & * & * & & & & & & * & * \\ & * & * & * & & & & & * & * \\ & & * & * & * & & & & * & * \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ & & & & & & & & * & * & * & * \\ * & * & * & * & * & \dots & * & * & * & * \\ * & * & * & * & * & \dots & * & * & * & * \end{bmatrix}$$

Рис. 11

Теперь в качестве  $A_{11}$  следует выбрать трехдиагональную матрицу размеров  $(n - 2) \times (n - 2)$ , стоящую в левом верхнем углу.

# 4

## Устойчивость вычислительных алгоритмов линейной алгебры

### 4.1 Введение

Исследуем вопрос об устойчивости решения линейной системы по отношению к возмущению правой части. Пусть рассматривается система

$$Ax = b \quad (4.1)$$

с квадратной невырожденной матрицей и система с возмущенной правой частью

$$A\tilde{x} = \tilde{b}. \quad (4.2)$$

Обозначим  $\tilde{b} - b = \delta b$ ,  $\tilde{x} - x = \delta x$  и оценим  $\delta x$  через  $\delta b$ . Вычитая (4.1) из (4.2), будем иметь

$$A\delta x = \delta b \quad \Rightarrow \quad \delta x = A^{-1}\delta b. \quad (4.3)$$

Пусть  $\|\cdot\|$  — некоторая норма вектора. В линейной алгебре наиболее часто используются следующие *нормы*

$$\|x\|_1 = \sum_{i=1}^n |x_i|, \quad \|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}, \quad \|x\|_\infty = \max_i |x_i|.$$

Как известно, *норма* матрицы, *подчиненная* векторной норме  $\|\cdot\|$  (операторная норма), определяется соотношением

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\|. \quad (4.4)$$

Указанным векторным нормам подчинены следующие матричные нормы:

$$\|A\|_1 = \max_j \sum_{i=1}^n |a_{ij}|, \quad \|A\|_2 = \sqrt{\lambda_{\max}(AA^T)}, \quad \|A\|_\infty = \max_i \sum_{j=1}^n |a_{ij}|.$$

Очевидно, что для симметричных матриц

$$A = A^T, \quad \|\cdot\|_\infty = \|\cdot\|_1, \quad \text{а } \|A\|_2 = |\lambda_{\max}|,$$

ибо  $Ax = \lambda x$ ,  $A^2x = \lambda Ax = \lambda^2x$ .

**Упражнение 4.1.** Доказать, что подчиненные нормы задаются именно этими соотношениями.

Из определения (4.4) матричной нормы, в частности, следует, что

$$\frac{\|Ax\|}{\|x\|} \leq \|A\| \quad \Rightarrow \quad \|Ax\| \leq \|A\| \|x\|. \quad (4.5)$$

Применяя это неравенство ко второму соотношению (4.3), будем иметь

$$\|\delta x\| \leq \|A^{-1}\| \|\delta b\|. \quad (4.6)$$

Соотношение (4.6) дает оценку абсолютной погрешности решения через абсолютную погрешность правой части. При этом множителем (коэффициентом усиления) выступает норма обратной матрицы. Чем больше эта норма, тем на меньшую точность мы можем рассчитывать.

Получим теперь оценку относительной погрешности. Из (4.1) в силу (4.5)

$$\|b\| \leq \|A\| \|x\|. \quad (4.7)$$

Деля (4.6) на (4.7), получим

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\| \|A^{-1}\| \frac{\|\delta b\|}{\|b\|}. \quad (4.8)$$

Это и есть оценка относительной погрешности. Здесь коэффициентом усиления выступает число

$$\|A\| \|A^{-1}\| =: \text{cond } A = \kappa(A) = \kappa, \quad (4.9)$$

называемое *числом обусловленности* матрицы  $A$ .

**Замечание 4.1.** Если вместо (4.1) решается задача  $A^{-1}x = b$ , то оценка относительной погрешности будет даваться тем же неравенством (4.8).

Если число обусловленности матрицы  $A$  большое, то про матрицу  $A$  говорят, что она *плохо обусловлена*. В противном случае говорят о хорошо обусловленной матрице. Поскольку  $AA^{-1} = I$ , то  $\|A\| \|A^{-1}\| \geq 1$ , т.е. число обусловленности не может быть меньше единицы. Имея систему с хорошо обусловленной матрицей (хорошо обусловленную систему), мы вправе рассчитывать на то, что при небольших возмущениях правой части возмущение решения не будет слишком велико.

## 4.2 Примеры плохо обусловленных систем

**Пример 4.1.**

$$\begin{aligned} x_1 &= 1, \\ x_1 + 0.01x_2 &= 1. \end{aligned} \tag{4.10}$$

Очевидно, что эта система невырождена и ее единственным решением является вектор  $[1, 0]^T$ . Возмутим правую часть системы (4.10) и найдем решение возмущенной задачи

$$\begin{cases} \tilde{x}_1 &= 1, \\ \tilde{x}_1 + 0.01\tilde{x}_2 &= 1.01, \end{cases} \quad \delta b_2 = 0.01. \tag{4.11}$$

Очевидно, что решением этой системы является вектор  $\tilde{x} = [1, 1]^T$ , который мало похож на невозмущенный вектор  $x$ , ибо

$$\delta x_2 = 1, \quad \delta x_1 = 0, \quad \|\delta x\|_1 = 1.$$

Это значение абсолютной погрешности решения полностью согласуется с оценками (4.6), (4.8), ибо

$$\begin{aligned} \|A\|_1 &= \max_j \sum_{i=1}^2 |a_{ij}| = 2, \quad A^{-1} = \begin{bmatrix} 1 & 0 \\ -100 & 100 \end{bmatrix}, \quad \|A^{-1}\|_1 = 101, \\ \|x\|_1 &= 1, \quad \|b\|_1 = 2, \quad \|\delta b\|_1 = 0.01 \end{aligned}$$

и, следовательно, в силу (4.6)

$$\|\delta x\|_1 \leq 101 \cdot 0.01 = 1.01,$$

а в силу (4.8)

$$\frac{\|\delta x\|_1}{\|x\|_1} \leq 2 \cdot 101 \cdot \frac{0.01}{2} = 1.01.$$

При заданном (4.11) уровне погрешности правой части обусловленность матрицы этой системы ( $\kappa = 202$ ) следует признать плохой.



**Упражнение 4.2.** Пусть искомая и возмущенная системы суть

$$\begin{aligned} x_1 &= 1, & \tilde{x}_1 &= 1, \\ -100 x_1 + 100x_2 &= 0, & -100 \tilde{x}_1 + 100\tilde{x}_2 &= 0.01. \end{aligned}$$

Исследовать ошибку, вызванную этим возмущением, и сравнить полученный результат с (4.6) и (4.8).

**Пример 4.2.**

$$A = \begin{bmatrix} 1 & -1 & -1 & \dots & -1 & -1 \\ 0 & 1 & -1 & \dots & -1 & -1 \\ 0 & 0 & 1 & \dots & -1 & -1 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & -1 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} -1 \\ -1 \\ -1 \\ \dots \\ -1 \\ 1 \end{bmatrix}.$$

В развернутом виде система запишется так

$$\begin{aligned} x_1 - x_2 - x_3 - \dots - x_n &= -1, \\ x_2 - x_3 - \dots - x_n &= -1, \\ \dots & \\ x_{n-1} - x_n &= -1, \\ x_n &= 1. \end{aligned} \tag{4.12}$$

Очевидно, что решением системы (4.12) является вектор

$$x = [0, 0, \dots, 0, 1]^T.$$

Легко видеть, что  $\det A = 1$ .

Возмутим последнюю компоненту вектора  $b$

$$\tilde{b} = [-1, -1, \dots, -1, 1 + \varepsilon]^T$$

и оценим погрешность решения.

Вычитая из возмущенной системы систему (4.12), для погрешности решения получим

$$\begin{aligned} \delta x_1 - \delta x_2 - \dots - \delta x_n &= 0, \\ \delta x_2 - \dots - \delta x_n &= 0, \\ \dots & \\ \delta x_{n-1} - \delta x_n &= 0, \\ \delta x_n &= \varepsilon. \end{aligned}$$

Отсюда находим, что

$$\begin{aligned}\delta x_n &= \varepsilon, & \delta x_{n-1} &= \varepsilon, & \delta x_{n-2} &= \delta x_n + \delta x_{n-1} = 2\varepsilon, \\ \delta x_{n-3} &= \delta x_n + \delta x_{n-1} + \delta x_{n-2} = 4\varepsilon = 2^2\varepsilon.\end{aligned}$$

Погрешность в каждой из последующих компонент, начиная с  $\delta x_{n-2}$ , удваивается, так что

$$\delta x_{n-k} = \delta x_n + \delta x_{n-1} + \cdots + \delta x_{n-(k-1)} = 2^{k-1}\varepsilon,$$

а

$$\delta x_1 = 2^{n-2}\varepsilon.$$

Таким образом,

$$\|\delta x\|_\infty = 2^{n-2}|\varepsilon|, \quad \|x\|_\infty = 1, \quad \|\delta b\|_\infty = |\varepsilon|, \quad \|b\|_\infty = 1, \quad \|A\|_\infty = n.$$

Поскольку в силу (4.8)

$$\frac{\|\delta x\|_\infty}{\|x\|_\infty} \leq \kappa \frac{\|\delta b\|_\infty}{\|b\|_\infty},$$

а в рассматриваемом случае

$$\frac{\|\delta x\|_\infty}{\|x\|_\infty} = 2^{n-2}|\varepsilon|,$$

то

$$\kappa = \|A^{-1}\|_\infty \|A\|_\infty \geq 2^{n-2}$$

и, следовательно,

$$\|A^{-1}\|_\infty \geq n^{-1}2^{n-2}.$$

При  $n = 102$ ,  $\|A\|_\infty = 102$ ,  $\kappa \geq 2^{100} > 10^{30}$ ,  $\|A^{-1}\|_\infty > 10^{27}$ . Если  $\varepsilon = 10^{-15}$ , то  $\|\delta x\|_\infty > 10^{15}$ . Матрица рассматриваемой системы очень плохо обусловлена.

**Упражнение 4.3.** Найти матрицу, обратную к матрице системы (4.12)

Понятие числа обусловленности введено нами только для невырожденных матриц. Условие  $\det A = 0$  означает вырожденность матрицы  $A$ , и может сложиться впечатление, что, если  $\det A \approx 0$ , то матрица плохо обусловлена. Однако, прямой связи между величиной определителя матрицы  $A$  и ее обусловленностью нет. Так, определитель матрицы из примера (4.2) равен единице, а

$$\kappa \geq 2^{n-2}.$$

С другой стороны, хорошо обусловленная матрица может иметь очень маленький определитель. Например, у матрицы

$$A = \begin{bmatrix} 10^{-1} & & & \\ & 10^{-1} & & 0 \\ & & \ddots & \\ 0 & & & 10^{-1} \end{bmatrix}$$

$\kappa = 1$ , хотя  $\det A = 10^{-n}$ .

### 4.3 Возмущение матрицы коэффициентов

Исследуем теперь вопрос о том, как влияет возмущение коэффициентов матрицы  $A$  на погрешность, приобретаемую решением.

**Теорема 4.1.** *Если  $A$  невырождена и*

$$\|\delta A\|/\|A\| < \kappa^{-1}(A), \quad (4.13)$$

*то и  $A + \delta A$  невырождена.*

**Доказательство.** Условие (4.13) в силу определения (4.9) можно переписать в виде

$$\|\delta A\| < 1/\|A^{-1}\| \quad \text{или} \quad \|\delta A\| \|A^{-1}\| < 1. \quad (4.14)$$

Для доказательства теоремы достаточно показать, что, если  $A + \delta A$  вырождена, то

$$\|\delta A\| \|A^{-1}\| \geq 1. \quad (4.15)$$

Пусть  $A + \delta A$  вырождена. Тогда нуль является ее собственным значением, и существует ненулевой вектор (собственный вектор)  $y$  такой, что

$$(A + \delta A)y = 0.$$

Обращая в этом соотношении матрицу  $A$ , которая по предположению теоремы является невырожденной, найдем, что

$$y = -A^{-1}\delta Ay.$$

Отсюда

$$\|y\| = \|A^{-1}\delta Ay\| \leq \|A^{-1}\| \|\delta A\| \|y\|.$$

Поскольку  $\|y\| > 0$ , то, сокращая полученное неравенство на  $\|y\|$ , будем иметь

$$1 \leq \|A^{-1}\| \|\delta A\|,$$

что совпадает с (4.15) и противоречит (4.14). Теорема доказана.

Предположим теперь, что у системы (4.1) возмущен не только вектор правой части  $b$ , но и сама матрица  $A$ , т.е. пусть

$$(A + \delta A)(x + \delta x) = b + \delta b. \quad (4.16)$$

Оценим возмущение  $\delta x$ .

**Теорема 4.2.** Пусть матрица  $A$  системы (4.1) невырождена и для ее возмущения  $\delta A$  справедлива оценка (4.13). Тогда для относительной погрешности решения справедлива оценка

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\varkappa}{1 - \varkappa \frac{\|\delta A\|}{\|A\|}} \left( \frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right). \quad (4.17)$$

**Доказательство.** Из (4.16)

$$Ax + \delta Ax + A\delta x + \delta A\delta x = b + \delta b.$$

Вычитая отсюда (4.1), получим

$$A\delta x = \delta b - \delta Ax - \delta A\delta x,$$

или

$$\delta x = A^{-1}[\delta b - \delta Ax - \delta A\delta x].$$

Отсюда

$$\|\delta x\| \leq \|A^{-1}\|(\|\delta b\| + \|\delta A\| \|x\| + \|\delta A\| \|\delta x\|).$$

Разрешим это неравенство относительно  $\|\delta x\|$

$$(1 - \|A^{-1}\| \|\delta A\|) \|\delta x\| \leq \|A^{-1}\|(\|\delta b\| + \|\delta A\| \|x\|).$$

В силу (4.13) (см. также (4.14)) коэффициент при  $\|\delta x\|$  положителен и, следовательно,

$$\|\delta x\| \leq \frac{\|A^{-1}\|(\|\delta b\| + \|\delta A\| \|x\|)}{1 - \|A^{-1}\| \|\delta A\|}. \quad (4.18)$$

Но

$$1 - \|A^{-1}\| \|\delta A\| = 1 - \varkappa \frac{\|\delta A\|}{\|A\|}.$$

Учитывая это и деля (4.18) на  $\|x\|$ , будем иметь

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\|A^{-1}\| \|A\| \left( \frac{\|\delta b\|}{\|A\| \|x\|} + \frac{\|\delta A\|}{\|A\|} \right)}{1 - \kappa \frac{\|\delta A\|}{\|A\|}}. \quad (4.19)$$

Поскольку  $\|A\| \|x\| \geq \|b\|$ , то из (4.19) вытекает (4.17). Теорема доказана.

#### 4.4 Пример хорошо обусловленной системы

Рассмотрим систему (4.1) с матрицей

$$A = \begin{bmatrix} 10^{-4} & 1 \\ 1 & 1 \end{bmatrix}. \quad (4.20)$$

Легко проверить, что

$$A^{-1} = (1 - 10^{-4}) \begin{bmatrix} -1 & 1 \\ 1 & -10^{-4} \end{bmatrix}$$

и, следовательно,

$$\kappa(A) = \|A\|_{\infty} \|A^{-1}\|_{\infty} = 4(1 - 10^{-4})^{-1} \approx 4.$$

Таким образом, матрица (4.20) хорошо обусловлена, и мы вправе надеяться, что система (4.1) с матрицей (4.20) может быть решена численно с хорошей точностью. Пусть

$$b = [1 \quad 2]^T. \quad (4.21)$$

Тогда решение системы (4.1), (4.20), (4.21) есть

$$x = \frac{1}{1 - 10^{-4}} \begin{bmatrix} 1 \\ 1 - 2 \cdot 10^{-4} \end{bmatrix} \approx \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \quad (4.22)$$

Будем решать эту систему при помощи  $LU$  разложения матрицы  $A$  с использованием трехрядной десятичной арифметики с плавающей точкой. В силу (1.19)

$$\begin{aligned} u_{11} &= a_{11}, & u_{12} &= a_{12}, \\ l_{21} &= a_{21}/u_{11}, & u_{22} &= a_{22} - l_{21}u_{12}. \end{aligned}$$

Поэтому приближенные вычисления будут использоваться только три раза: при делении, при умножении и при вычитании

$$\begin{aligned}\widetilde{l}_{21} &= \text{fl}(1/10^{-4}) = 10^4 = l_{21}, \\ \widetilde{l}_{21}u_{12} &= \text{fl}(10^4 \cdot 1) = 10^4 = l_{21}u_{12}, \quad \widetilde{u}_{22} = \text{fl}(1 - 10^4) = \text{fl}(-9999) = -10^4.\end{aligned}$$

Итак,

$$\widetilde{L} = L + \delta L = \begin{bmatrix} 1 & 0 \\ 10^4 & 1 \end{bmatrix} = L, \quad \widetilde{U} = U + \delta U = \begin{bmatrix} 10^{-4} & 1 \\ 0 & -10^4 \end{bmatrix}.$$

При этом

$$\widetilde{A} = A + \delta A = \widetilde{L}\widetilde{U} = \begin{bmatrix} 10^{-4} & 1 \\ 1 & 0 \end{bmatrix},$$

в то время как

$$A = \begin{bmatrix} 10^{-4} & 1 \\ 1 & 1 \end{bmatrix},$$

и, следовательно,

$$\delta A = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix},$$

т.е.  $\|\delta A\| \sim \|A\|$ , а это означает, что мы фактически решаем систему с совсем другой матрицей. В самом деле, решая систему  $\widetilde{L}\widetilde{y} = b$  при помощи формул (1.21), находим, что

$$\widetilde{y}_1 = b_1 = 1, \quad \widetilde{y}_2 = b_2 - l_{21}\widetilde{y}_1 = \text{fl}(2 - 10^4 \cdot 1) = \text{fl}(-9998) = -10^4,$$

решение системы  $\widetilde{U}\widetilde{x} = \widetilde{y}$  по формулам (1.22) дает

$$\begin{aligned}\widetilde{x}_2 &= \widetilde{y}_2/\widetilde{u}_{22} = \text{fl}(-10^4/(-10^4)) = 1, \\ \widetilde{x}_1 &= [\widetilde{y}_1 - \widetilde{u}_{12}\widetilde{x}_2]/\widetilde{u}_{11} = \text{fl}((1 - 1)/10^{-4}) = 0.\end{aligned}$$

Тем самым,

$$\widetilde{x} = x + \delta x = [0 \quad 1]^T,$$

что, как и ожидалось, мало похоже на точное решение (4.22).

Отметим, что индикатором неблагополучия являются и оценки

$$\kappa(\widetilde{L}) \approx 10^8, \quad \kappa(\widetilde{U}) \approx 10^8$$

при  $\kappa(A) \approx 4$ .

В чем же причина появления столь значительной погрешности? Говорить о накоплении ошибок округления не приходится, равно как и о

плохой обусловленности системы. Действительная причина состоит в том, что метод исключения Гаусса в том виде, в каком он был описан, является неустойчивым методом. Чтобы определить, в чем именно его слабость, рассмотрим более внимательно процедуру вычисления  $\tilde{u}_{22}$ , где и появились первые округления. При вычитании из  $a_{22}$  большого числа  $l_{21}u_{12}$  и последующего округления элемент  $a_{22}$  был полностью утерян, что и привело к большой погрешности, а большое вычитаемое образовалось из-за большого элемента  $l_{21}$ , чему способствовала малость главного элемента  $u_{11} = a_{11}$ .

## 4.5 Метод Гаусса с выбором главного элемента

Из-за отмеченной неустойчивости метод Гаусса в вычислительной практике обычно применяется в сочетании с некоторой схемой выбора главного элемента. Например, схема выбора главного элемента по столбцу состоит в следующем. Перед началом первого шага среди коэффициентов  $a_{11}, a_{21}, \dots, a_{n1}$ , образующих первый столбец матрицы  $A$ , выбирается коэффициент с наибольшим модулем; пусть это будет  $a_{k,1}$ . Если  $k > 1$ , то в системе (4.1) переставляются 1-е и  $k$ -е уравнения, при  $k = 1$  перестановка не нужна. После этой предварительной работы обычным образом проводится 1-й шаг прямого хода. До начала 2-го шага среди коэффициентов  $a_{22}^{(1)}, a_{31}^{(1)}, \dots, a_{n,2}^{(1)}$  (т.е. во втором столбце текущей матрицы) выбирается коэффициент  $a_{l2}^{(1)}$  с наибольшим модулем. В случае  $l > 2$  переставляются 2-е и  $l$ -е уравнения, затем выполняется 2-й шаг. И т.д.

Опишем этот алгоритм формально в терминах треугольного разложения. Для этого нам потребуется определить матрицы перестановок.

**Определение 4.1.** Матрицей перестановок  $P$  называется матрица, полученная из единичной произвольной перестановкой строк.

**Лемма 4.1.** Пусть  $P, P_1$  и  $P_2$  — матрицы перестановок порядка  $n$ , а  $A$  — произвольная  $n \times n$  матрица. Тогда

1°  $PA$  отличается от  $A$  только порядком строк.  $AP$  отличается от  $A$  только порядком столбцов,

2°  $P^{-1} = P^T$ ,

3°  $\det P = \pm 1$ ,

4°  $P_1P_2$  также является матрицей перестановок.

**Упражнение 4.4.** Доказать лемму 4.1.

**Теорема 4.3.** Для невырожденной матрицы  $A$  существуют перестановки, задаваемые матрицей  $P$ , нижняя треугольная матрица  $L$  с единичной главной диагональю и невырожденная верхняя треугольная матрица  $U$ , такие, что  $PA = LU$ .

**Доказательство.** Проведем доказательство при помощи метода полной математической индукции по порядку  $n$ . Утверждение очевидно для  $1 \times 1$  матриц:  $P = L = 1$  и  $U = A$ .

Предположим, что утверждение верно для матриц порядка  $n - 1$ . Поскольку невырожденная матрица  $A$  в каждом столбце должна иметь ненулевые элементы, выберем матрицу перестановок  $P_1$  так, чтобы элемент первой строки и первого столбца матрицы  $P_1A$ , который обозначим через  $a_{11}$ , был отличен от нуля. Затем представим  $P_1A$  в блочном виде

$$P_1A = \begin{bmatrix} a_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

где  $A_{22}$  — матрица порядка  $n - 1$ , а  $A_{21}$  и  $A_{12}^T$  — матрицы-столбцы высоты  $n - 1$ . И, наконец, выполним блочное треугольное разложение

$$\begin{bmatrix} a_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ L_{21} & I \end{bmatrix} \begin{bmatrix} u_{11} & U_{12} \\ 0 & A_{22}^{(1)} \end{bmatrix} = \begin{bmatrix} u_{11} & U_{12} \\ L_{21}u_{11} & L_{21}U_{12} + A_{22}^{(1)} \end{bmatrix}.$$

Сравнивая первую и последнюю матрицы в этом соотношении, найдем, что

$$\begin{aligned} u_{11} = a_{11} \neq 0, \quad U_{12} = A_{12}, \quad L_{21}u_{11} = A_{21} &\Rightarrow L_{21} = A_{21}/u_{11}, \\ A_{22}^{(1)} = A_{22} - L_{21}U_{12}. & \end{aligned} \quad (4.23)$$

Теперь, для того чтобы к  $A_{22}^{(1)}$  можно было применить предположение индукции, нужно убедиться, что  $\det A_{22}^{(1)} \neq 0$ . В самом деле, поскольку  $\det [P_1A] = \pm \det A \neq 0$ , а

$$\det [P_1A] = \det \begin{bmatrix} 1 & 0 \\ L_{21} & I \end{bmatrix} \cdot \det \begin{bmatrix} u_{11} & U_{12} \\ 0 & A_{22}^{(1)} \end{bmatrix} = 1 \cdot (u_{11} \cdot \det A_{22}^{(1)}),$$

то  $\det A_{22}^{(1)} \neq 0$ .

Итак, по предположению индукции, найдется матрица перестановок  $\tilde{P}$ , такая что  $\tilde{P}A_{22}^{(1)} = L^{(1)}U^{(1)}$ , где  $L^{(1)}$  — нижняя треугольная матрица с единичной главной диагональю, а  $U^{(1)}$  — невырожденная верхняя



треугольная матрица. Преобразовывая это равенство с использованием леммы 4.1 и подставляя результат в написанное выше блочное  $2 \times 2$ -разложение, получим

$$\begin{aligned} P_1 A &= \begin{bmatrix} 1 & 0 \\ L_{21} & I \end{bmatrix} \begin{bmatrix} u_{11} & U_{12} \\ 0 & \tilde{P}^T L^{(1)} U^{(1)} \end{bmatrix} = \\ &= \begin{bmatrix} 1 & 0 \\ L_{21} & I \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \tilde{P}^T L^{(1)} \end{bmatrix} \begin{bmatrix} u_{11} & U_{12} \\ 0 & U^{(1)} \end{bmatrix} = \\ &= \begin{bmatrix} 1 & 0 \\ L_{21} & \tilde{P}^T L^{(1)} \end{bmatrix} \begin{bmatrix} u_{11} & U_{12} \\ 0 & U^{(1)} \end{bmatrix} = \\ &= \begin{bmatrix} 1 & 0 \\ 0 & \tilde{P}^T \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \tilde{P} L_{21} & L^{(1)} \end{bmatrix} \begin{bmatrix} u_{11} & U_{12} \\ 0 & U^{(1)} \end{bmatrix}, \end{aligned}$$

что и дает требуемое разложение  $A$ :

$$\left( \begin{bmatrix} 1 & 0 \\ 0 & \tilde{P}^T \end{bmatrix}^T P_1 \right) A = P_2 P_1 A = P A = \begin{bmatrix} 1 & 0 \\ \tilde{P} L_{21} & L^{(1)} \end{bmatrix} \begin{bmatrix} u_{11} & U_{12} \\ 0 & U^{(1)} \end{bmatrix}.$$

Теорема доказана.

**Замечание 4.2.** Теорема остается в силе, если вместо перестановок строк матрицы  $A$  использовать перестановки столбцов:  $A \Rightarrow AP$ . Более того, перестановки строк и столбцов можно использовать одновременно:  $A \Rightarrow P_1 A P_2$ .

**Замечание 4.3.** Из (4.23) и (1.20) следует, что матрицы  $L_{21}$  и  $U_{12}$  суть поддиагональная часть первого столбца и наддиагональная часть первой строки матриц  $L$  и  $U$ , соответственно, в  $LU$  разложении матрицы  $P_1 A$ , а матрица  $A_{22}^{(1)}$  совпадает с ведущей подматрицей, получающейся после первого шага исключения Гаусса, коэффициенты которой определяются формулами (1.7), (1.9) при  $k = 1$ . (Для  $n = 3$  см. (1.4) и формулы для  $a_{ij}^{(1)}$ ).

**Упражнение 4.5.** Пусть матрица перестановок  $n$ -го порядка  $P_2$  имеет вид  $\begin{bmatrix} 1 & 0 \\ 0 & \tilde{P}_2 \end{bmatrix}$ , где  $\tilde{P}_2$  — матрица перестановок порядка  $n - 1$ . Доказать, что

$$P_2 P_1 A = \begin{bmatrix} 1 & 0 \\ \tilde{P}_2 L_{21} & I \end{bmatrix} \begin{bmatrix} u_{11} & U_{12} \\ 0 & \tilde{P}_2 A_{22}^{(1)} \end{bmatrix}.$$

Что достигается выбором ведущего элемента по столбцу? Согласно (1.9)  $l_{ik} = \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}$ , и мы можем теперь гарантировать, что множители  $l_{ij}$  всех шагов (элементы матрицы  $L$ ) по абсолютной величине ограничены единицей. Формулы (1.7) показывают, что, во-первых, добавки  $l_{ik}a_{kj}^{(k-1)}$  к текущим значениям коэффициентов имеют тот же порядок величины, что и сами коэффициенты, во-вторых, за один шаг уровень коэффициентов матрицы может вырасти не более, чем в два раза. Действительно, согласно (1.7)

$$|a_{ij}^{(k)}| = |a_{ij}^{(k-1)} - l_{ik}a_{kj}^{(k-1)}| \leq |a_{ij}^{(k-1)}| + |a_{kj}^{(k-1)}| \leq 2 \max_{ij} |a_{ij}^{(k-1)}|,$$

$$\rho = \frac{\max |u_{kj}|}{\max a_{kj}} \leq 2^{n-1}. \quad (4.24)$$

**Упражнение 4.6.** Решить систему (4.1), (4.20), (4.21) методом Гаусса с выбором главного элемента по столбцу на 3-разрядном десятичном калькуляторе.

Иногда используется и другая схема выбора главного элемента, а именно, схема выбора по строке. Здесь до начала 1-го шага определяется наибольший по модулю среди коэффициентов  $a_{11}, a_{12}, \dots, a_{1n}$ . Пусть им будет коэффициент  $a_{1k}$ . Если,  $k > 1$ , то производится перенумерация неизвестных: 1-е и  $k$ -е неизвестные меняются номерами. Это соответствует перестановке столбцов матрицы системы. При  $k = 1$  перестановка не нужна. Теперь обычным образом проводится 1-й шаг прямого хода. И т.д.

Переходя к выбору главного элемента по столбцу, мы получили для системы (4.1) приближенное решение хорошего качества. Но это не значит, что описанные схемы с выбором главного элемента придают методу Гаусса гарантированную устойчивость. Хотя обычно схемы с выбором главного элемента по столбцу или по строке действительно обеспечивают устойчивое вычисление.

В каких же случаях утрачивается устойчивость? Чтобы понять это, заметим, что во многих численных методах ошибки промежуточных вычислений в совокупности равносильны тому, как если бы тем же методом точно решали исходную задачу, предварительно изменив ее входные данные. Это относится и к методу Гаусса. Можно показать, что решение линейной системы (4.1), вычисленное методом Гаусса (с той или иной

схемой выбора главного элемента или вообще без выбора) при наличии ошибок округления точно удовлетворяет измененному уравнению

$$(A + \delta A)\tilde{x} = b. \quad (4.25)$$

В пояснение сказанного рассмотрим умножение двух треугольных матриц с учетом ошибок округления

$$\begin{aligned} \widetilde{AB} &= \begin{bmatrix} a_{11} & a_{12} \\ 0 & a_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ 0 & b_{22} \end{bmatrix} = \\ &= \begin{bmatrix} a_{11}b_{11}(1 + \varepsilon_1) & (a_{11}b_{12}(1 + \varepsilon_2) + a_{12}b_{22}(1 + \varepsilon_3))(1 + \varepsilon_4) \\ 0 & a_{22}b_{22}(1 + \varepsilon_5) \end{bmatrix} \end{aligned}$$

Если положить

$$\tilde{A} = \begin{bmatrix} a_{11} & a_{12}(1 + \varepsilon_3)(1 + \varepsilon_4) \\ 0 & a_{22}(1 + \varepsilon_5) \end{bmatrix}, \quad \tilde{B} = \begin{bmatrix} b_{11}(1 + \varepsilon_1) & b_{12}(1 + \varepsilon_2)(1 + \varepsilon_4) \\ 0 & b_{22} \end{bmatrix},$$

то легко проверить, что

$$\widetilde{AB} = \tilde{A}\tilde{B}.$$

Для элементов матрицы возмущения в (4.25) справедлива оценка

$$|A - \tilde{L}\tilde{U}| \leq 3n\eta(|A| + |\tilde{L}||\tilde{U}|) + O(\eta^2). \quad (4.26)$$

Итак, при прочих равных условиях ошибки  $\delta A$ , вносимые методом в исходную информацию, тем больше, чем больший рост элементов матриц допускается в прямом ходе. Если в версии метода Гаусса из первой лекции рост элементов может быть сколь угодно велик, то в схемах с выбором главного элемента он ограничен. Действительно, за шаг максимальный модуль элемента матрицы может вырасти самое большее в два раза (см. (4.24)). Так как маловероятно, чтобы возмущение происходило на каждом шаге, то обычно коэффициент роста  $g(A)$  невелик. В этом случае благодаря множителю  $\eta$  в правой части (4.26) матрицу-возмущение  $\delta A$  можно считать малой по сравнению с  $A$ .

Приведем примеры матриц, преобразование которых при помощи метода Гаусса с выбором главного элемента приводит к максимально возможному увеличению коэффициентов промежуточных матриц прямого хода.

**Пример 4.3.** Выбор по столбцу

$$A = \begin{bmatrix} 1 & 0 & 0 & 1 \\ -1 & 1 & 0 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & -1 & 1 & 2 \\ 0 & -1 & -1 & 2 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & -1 & 4 \end{bmatrix},$$

$$A_3 = U = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 8 \end{bmatrix}, \quad L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ -1 & -1 & 1 & 0 \\ -1 & -1 & -1 & 1 \end{bmatrix}.$$

**Упражнение 4.7.** Показать, что выбор главного элемента по строке для матрицы

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

приводит к максимальному росту элементов промежуточных матриц.

**Замечание 4.4.** Возможна еще одна схема выбора главного элемента: выбор по всей матрице. В этом случае гарантируется полная устойчивость метода Гаусса, однако сама процедура выбора такого главного элемента очень трудоемка — для ее реализации требуется  $O(n^3)$  действий, что сравнимо с трудоемкостью самого метода и, следовательно, существенно удорожает решение. Отметим, что при выборе главного элемента по столбцу или по строке требуется лишь  $O(n^2)$  дополнительных операций.

## II

# Итерационные методы решения линейных систем

## 5

# Простая итерация и чебышевский итерационный метод

В предыдущих лекциях для системы линейных алгебраических уравнений

$$Ax = b \quad (5.1)$$

с квадратной невырожденной матрицей были рассмотрены четыре прямых метода отыскания решения:

- а) метод Гаусса ( $LU$ -разложение, треугольное разложение) и его модификация с выбором ведущего элемента,
- б) метод Холецкого, применяемый в случае симметричной положительно определенной матрицы,
- в) метод вращений,
- г) метод отражений.

Все эти методы позволяют в принципе (при отсутствии ошибок округления) найти точное решение системы (5.1) за конечное число действий. Это число было оценено нами величиной  $O(n^3)$ , где  $n$  — порядок системы. Если матрица  $A$  системы имеет ленточную структуру с полушириной ленты  $p$  много меньшей  $n$ , то ленточные варианты первых двух методов позволяют найти точное решение с затратой  $O(p^2n)$  действий.

В этом параграфе мы рассмотрим другой класс методов решения системы (5.1) — итерационных. Эти методы, как правило, если и позволяют найти точное решение системы (5.1), то только как предел при стремлении числа итераций (а, следовательно, и действий) к бесконечности. Однако для широкого класса задач, встречающихся в приложениях, те или иные

итерационные методы могут оказаться предпочтительнее с точки зрения используемых ресурсов, чем описанные прямые.

## 5.1 Одношаговые итерационные методы

Из курса "Введение в численные методы" известно, что многие одношаговые итерационные методы могут быть записаны в так называемой канонической форме

$$B \frac{x^{k+1} - x^k}{\tau} + Ax^k = b, \quad k = 0, 1, \dots, \quad (5.2)$$

где  $B$  — некоторая матрица, определяющая итерационный метод, а  $\tau$  — итерационный параметр. В частности, в виде (5.2) могут быть записаны метод Якоби, метод Гаусса-Зейделя, метод последовательной верхней релаксации, метод простых итераций. Предположим, что

$$A = A^T > 0, \quad B = B^T > 0. \quad (5.3)$$

При этих предположениях в указанном курсе доказано, что если

$$B > \frac{\tau}{2}A, \quad (5.4)$$

т.е. если для любого ненулевого вектора  $x$  справедливо неравенство

$$(Bx, x) > \frac{\tau}{2}(Ax, x),$$

то итерационный метод (5.2) является сходящимся.

Для метода простых итераций, который имеет вид (5.2) с  $B = I$ , кроме того, дана и оценка скорости сходимости. Именно, если

$$\tau = 2/(\lambda_1 + \lambda_n), \quad (5.5)$$

где  $\lambda_1$  и  $\lambda_n$ , соответственно, наименьшее и наибольшее собственные значения матрицы  $A$ , то для погрешности итераций справедлива оценка

$$\|x - x^k\| \leq \rho \|x - x^{k-1}\|, \quad \rho = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} = \frac{1 - \lambda_1/\lambda_n}{1 + \lambda_1/\lambda_n} < 1, \quad (5.6)$$

где  $\|\cdot\|$  — евклидова длина.

Из (5.6) следует, что если  $\lambda_n \gg \lambda_1$ , то число  $\rho$  очень близко к единице, а скорость сходимости итераций очень низкая. Но при указанном выборе нормы вектора

$$\|A\| = \max_i \lambda_i(A) = \lambda_n, \quad \|A^{-1}\| = \max_i \lambda_i(A^{-1}) = \lambda_1^{-1}$$

и

$$\lambda_n/\lambda_1 = \|A\| \|A^{-1}\| = \text{cond } A = \varkappa(A) = \varkappa.$$

Таким образом, если матрица  $A$  плохо обусловлена, а это типичная ситуация, то метод простых итераций сходится очень медленно.

## 5.2 Неявные методы

Какие есть пути увеличения скорости сходимости итерационных методов? Изучим влияние матрицы  $B$  из (5.2) на скорость сходимости. В силу (5.3) существует матрица  $B^{1/2}$  такая, что

$$B^{1/2} = (B^{1/2})^T > 0 \quad \text{и} \quad B^{1/2} B^{1/2} = B.$$

Эта матрица называется квадратным корнем из матрицы  $B$ .

Напомним построение матрицы  $B^{1/2}$ . Пусть  $\lambda$  — собственное значение матрицы  $B$ , а  $\xi$  — отвечающий ему собственный вектор, т.е.  $B\xi = \lambda\xi$ . Перенумеруем все собственные значения матрицы  $B$  и введем в рассмотрение диагональную матрицу  $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ , образованную этими собственными значениями, и ортогональную матрицу  $\Xi = [\xi_1, \xi_2, \dots, \xi_n]$ , образованную ортонормированными собственными векторами  $\xi_i$  матрицы  $B$ , упорядоченными в соответствии с нумерацией собственных значений. Поскольку  $\Xi\Lambda = [\lambda_1\xi_1 \lambda_2\xi_2 \dots \lambda_n\xi_n]$ , то  $B\Xi = \Xi\Lambda$ . Отсюда следует, что  $B = \Xi\Lambda\Xi^T$ .

Очевидно, что матрица  $\Lambda^{1/2} = \text{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots, \sqrt{\lambda_n})$ . Поэтому  $B^{1/2} = \Xi\Lambda^{1/2}\Xi^T$ .

Введем обозначение

$$B^{1/2}x^k = y^k, \quad x^k = (B^{1/2})^{-1}y^k = B^{-1/2}y^k. \quad (5.7)$$

В новых обозначениях (5.2) можно переписать так

$$B^{1/2} \frac{y^{k+1} - y^k}{\tau} + AB^{-1/2}y^k = b,$$

а после применения к этому соотношению матрицы  $B^{-1/2}$ , получим

$$\frac{y^{k+1} - y^k}{\tau} + B^{-1/2}AB^{-1/2}y^k = B^{-1/2}b =: f. \quad (5.8)$$

Обозначая

$$B^{-1/2}AB^{-1/2} = C, \quad (5.9)$$



будем иметь соотношения

$$\frac{y^{k+1} - y^k}{\tau} + Cy^k = f, \quad k = 0, 1, \dots, \quad (5.10)$$

которые по форме полностью совпадают с (5.2). Очевидно, что  $C = C^T > 0$ , и поэтому для итерационного метода (5.10) (а это есть метод простых итераций) при

$$\tau = \frac{2}{\lambda_1(C) + \lambda_n(C)} \quad (5.11)$$

справедлива оценка

$$\|y^{k+1} - y\| \leq \rho(C) \|y^k - y\|, \quad \rho(C) = \frac{\lambda_n(C) - \lambda_1(C)}{\lambda_n(C) + \lambda_1(C)} = \frac{1 - \kappa^{-1}(C)}{1 + \kappa^{-1}(C)} < 1, \quad (5.12)$$

где  $\lambda_1(C)$  и  $\lambda_n(C)$  — соответственно минимальное и максимальное собственные значения матрицы  $C$ :

$$C\xi = \lambda(C)\xi. \quad (5.13)$$

Здесь  $\xi$  — собственные векторы матрицы  $C$ . Подставляя в (5.13) представление  $C$  из (5.9), получим

$$B^{-1/2}AB^{-1/2}\xi = \lambda(C)\xi,$$

а, обозначая

$$B^{-1/2}\xi = \eta, \quad \xi = B^{1/2}\eta$$

и применяя к последней задаче матрицу  $B^{1/2}$ , будем иметь

$$A\eta = \lambda(C)B\eta. \quad (5.14)$$

Таким образом,  $\lambda_1(C)$  и  $\lambda_n(C)$  суть минимальное и максимальное собственные значения обобщенной задачи на собственные значения (5.14).

Подставляя в (5.12)  $y^k$  из (5.7) и обозначая  $(Bx, x) = \|x\|_B^2$ , получим

$$\|x^{k+1} - x\|_B \leq \rho(C) \|x^k - x\|_B.$$

Итак, если  $B = B^T > 0$  такова, что

$$\lambda_n/\lambda_1 > \lambda_n(C)/\lambda_1(C),$$

то итерационный метод (5.2), (5.11) с этой матрицей  $B$  будет сходиться быстрее (в смысле  $B$ -нормы), чем метод простых итераций. Наибольшую скорость сходимости мы получим, выбирая  $B = A$ . Тогда

$$C = I, \quad \lambda_1(C) = \lambda_n(C) = 1, \quad \tau = 1$$

и (5.2) принимает вид

$$Ax^{k+1} = b,$$

что с точностью до обозначений совпадает с (5.1). Метод сходится за одну итерацию. Лучшего быть не может. Но мы пришли к тому, от чего хотели уйти: нам снова нужно решать систему (5.1). Отсюда следует, что на выбор матрицы  $B$  нужно наложить весьма серьезные ограничения — матрица  $B$  должна быть относительно легко обратима. Таковыми, например, являются диагональные и треугольные матрицы. Хотя последние и не являются симметричными, это неудобство легко устранить, выбирая в качестве  $B$  подходящее произведение треугольных матриц. Однако скорость сходимости метода (5.2) при таком выборе  $B$  из очевидных соображений остается сравнительно низкой.

В настоящее время неизвестно регулярных способов хорошего выбора матрицы  $B$  для произвольной  $A$ . Все удачные находки так или иначе связаны со спецификой матрицы  $A$ .

### 5.3 Чебышевский итерационный метод

Другой путь увеличения скорости сходимости итерационного метода (5.2) состоит в том, чтобы вместо одного итерационного параметра  $\tau$  использовать несколько — свой на каждой итерации. Итерационные методы такого типа называются нестационарными и имеют вид

$$B \frac{x^{k+1} - x^k}{\tau_{k+1}} + Ax^k = b, \quad k = 0, 1, \dots \quad (5.15)$$

или с учетом (5.7), (5.9)

$$\frac{y^{k+1} - y^k}{\tau_{k+1}} + Cy^k = f, \quad k = 0, 1, \dots \quad (5.16)$$

Укажем один из возможных способов выбора итерационных параметров  $\tau_k$ . Пусть

$$z^k = y^k - y$$

— погрешность решения после  $k$  итераций, а

$$Cy = f. \quad (5.17)$$

Вычитая (5.17) из (5.16), получим задачу для  $z^k$ :

$$\frac{z^{k+1} - z^k}{\tau_{k+1}} + Cz^k = 0, \quad k = 0, 1, \dots, \quad z^0 = y^0 - y. \quad (5.18)$$

Разрешим это соотношение относительно  $z^{k+1}$ . Получим

$$z^{k+1} = (I - \tau_{k+1}C)z^k,$$

и, следовательно,

$$z^k = \prod_{j=1}^k (I - \tau_j C) z^0, \quad (5.19)$$

т.е.

$$z^k = P_k(C)z^0,$$

где

$$P_k(t) = \prod_{j=1}^k (1 - \tau_j t) = 1 + a_1^{(k)}t + \dots + a_k^{(k)}t^k. \quad (5.20)$$

Из (5.19) находим, что

$$\|z^k\| = \|y^k - y\| = \left\| \prod_{j=1}^k (I - \tau_j C) z^0 \right\| \leq \left\| \prod_{j=1}^k (I - \tau_j C) \right\| \|z^0\|. \quad (5.21)$$

Но

$$\left\| \prod_{j=1}^k (I - \tau_j C) \right\| = \max_l |\lambda_l(P_k(C))| = \max_l |P_k(\lambda_l(C))|. \quad (5.22)$$

Поскольку мы хотим, чтобы итерации сходились как можно быстрее при любом начальном приближении, то можно поставить задачу о минимизации  $\|P_k(C)\|$  в зависимости от итерационных параметров  $\tau_j$ ,  $j = \overline{1, k}$ . В силу (5.20), (5.22) эта задача эквивалентна задаче построения многочлена  $P_k(t)$  степени  $k$  с единичным свободным членом, который в точках спектра матрицы  $C$  наиболее близок к нулю. Однако поставленная задача практически не разрешима. Тем не менее, вместо нее можно поставить близкую задачу о построении  $P_k(t)$ , наименее отклоняющегося от нуля не на спектре, а на отрезке  $[\lambda_1, \lambda_n]$ , где этот спектр расположен. Эта задача много проще. Найдем это решение.

Среди многочленов степени  $k$  таких, что  $Q_k(0) = 1$  (см. (5.20)), требуется найти многочлен  $P_k(t)$ , максимум модуля которого на  $[\lambda_1, \lambda_n]$  минимален.

Линейной заменой переменной  $t = ax + b$  переведем отрезок  $[\lambda_1, \lambda_n]$  в отрезок  $[1, -1]$ . Имеем

$$\left. \begin{aligned} \lambda_1 &= a + b \\ \lambda_n &= -a + b \end{aligned} \right\}, \quad b = \frac{\lambda_n + \lambda_1}{2}, \quad a = -\frac{\lambda_n - \lambda_1}{2},$$

т.е.

$$t = \frac{\lambda_n + \lambda_1}{2} - \frac{\lambda_n - \lambda_1}{2} x = \frac{\lambda_1 + \lambda_n}{2} \left[ 1 - \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} x \right] = \frac{1}{\tau_0} [1 - \rho_0 x], \quad (5.23)$$

где

$$\tau_0 = \frac{2}{\lambda_n + \lambda_1}, \quad \rho_0 = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} = \frac{1 - \varkappa^{-1}}{1 + \varkappa^{-1}} < 1. \quad (5.24)$$

Заметим, что  $\tau_0$  совпадает с  $\tau$  из (5.5), (5.11) для стационарного итерационного процесса, а  $\rho_0$  совпадает с  $\rho$  из (5.6), (5.12) и характеризует скорость сходимости этого процесса.

Пусть

$$P_k(t) = \widehat{P}_k(x). \quad (5.25)$$

Поскольку  $t = 0$  отвечает точка  $x = \rho_0^{-1}$  (см. (5.23)), то должно быть

$$P_k(0) = \widehat{P}_k(\rho_0^{-1}) = 1. \quad (5.26)$$

Наша задача свелась к отысканию многочлена  $\widehat{P}_k(x)$ , который наименее отклоняется от нуля на  $[-1, 1]$  и удовлетворяет условию (5.26). Похожая задача решена в гл. 11. Оттуда мы знаем, что среди многочленов степени  $k$  вида  $x^k + \dots$  наименее отклоняется от нуля на  $[-1, 1]$  многочлен

$$\overline{T}_k(x) = \frac{1}{2^{k-1}} T_k(x),$$

где  $T_k(x)$  — многочлен Чебышева первого рода. Разумеется, сам многочлен Чебышева  $T_k(x)$  является наименее отклоняющимся от нуля на  $[-1, 1]$  среди многочленов вида  $P_k(x) = 2^{k-1}x^k + \dots$

Пусть

$$T_k(\rho_0^{-1}) = q_k^{-1}. \quad (5.27)$$

Построим многочлен  $\widetilde{T}_k(x) = q_k T_k(x)$ . Очевидно, что этот многочлен является наименее отклоняющимся от нуля на  $[-1, 1]$  среди многочленов вида

$$P_k(x) = 2^{k-1} q_k x^k + \dots \quad (5.28)$$

и, кроме того многочлен  $\widetilde{T}_k(x)$  удовлетворяет условию (5.26). Покажем, что  $\widetilde{T}_k(x)$  и есть интересующий нас многочлен. Поскольку  $\widetilde{T}_k(x)$  дает решение задачи минимизации среди многочленов, удовлетворяющих двум условиям (5.28) и (5.26), а не одному (5.26), то, вообще говоря, может существовать другой многочлен, удовлетворяющий (5.26), максимум модуля которого на  $[-1, 1]$  меньше, чем у  $\widetilde{T}_k(x)$ . Пусть этот многочлен есть

$P_k(x)$ . Поскольку в силу свойств 4° многочленов Чебышева многочлен  $\tilde{T}_k(x)$  в  $(k+1)$  точке (11.21) отрезка  $[-1, 1]$  принимает с чередующимися знаками максимальное значение модуля, то в этих же точках многочлен  $\tilde{T}_k(x) - P_k(x)$  также имеет различные знаки. Поэтому на  $[-1, 1]$  существует  $k$  точек, где многочлен  $k$ -ой степени  $\tilde{T}_k(x) - P_k(x)$  обращается в нуль. Но этот многочлен обращается в нуль и в  $(k+1)$ -ой точке  $\rho_0^{-1} \notin [-1, 1]$ , что невозможно, если  $P_k(x) \neq \tilde{T}_k(x)$ . Тем самым,

$$\hat{P}_k(x) = q_k T_k(x), \quad (5.29)$$

и нестационарный итерационный метод построен.

Найдем формулы для итерационных параметров  $\tau_j$ . Из (5.25), (5.29) и (5.23) следует, что нули полиномов  $P_k(t)$  и  $T_k\left(\frac{1-\tau_0 t}{\rho_0}\right)$  совпадают. Так как полином  $P_k(t)$  имеет нули в точках  $t = 1/\tau_j$ ,  $j = \overline{1, k}$ , а нулями полинома Чебышева  $T_k(x)$  являются числа (11.21)

$$x_j = \cos \frac{(2j-1)\pi}{2k}, \quad j = 1, 2, \dots, k,$$

то с учетом (5.23) находим, что

$$\tau_j = \frac{\tau_0}{1 - \rho_0 \mu_j}, \quad j = 1, 2, \dots, k, \quad (5.30)$$

где

$$\mu_j \in \mathfrak{M}_k = \left\{ \cos \frac{2i-1}{2k}\pi, \quad i = 1, 2, \dots, k \right\}. \quad (5.31)$$

Оценим скорость сходимости построенного итерационного процесса, который называется *чебышевским итерационным методом*.

В силу свойства 4° многочленов Чебышева  $\max_{[-1,1]} |T_k(x)| = 1$ . Отсюда с учетом (5.18) и (5.25) находим, что

$$\max_{[\lambda_1, \lambda_n]} |P_k(t)| = \max_{[-1,1]} \|\hat{P}_k(x)\| = q_k.$$

Поэтому следствием (5.21), (5.22) будет оценка

$$\|z^k\| \leq q_k \|z^0\|. \quad (5.32)$$

Проанализируем зависимость  $q_k$  от  $k$  и числа обусловленности  $\varkappa$  матрицы  $C$ . Из свойства 6° многочленов Чебышева находим, что  $T_k(\rho_0^{-1}) = \operatorname{ch} k \operatorname{Arch} \rho_0^{-1}$ , и поэтому, с учетом (5.27),

$$\operatorname{ch} k \operatorname{Arch} \frac{1}{\rho_0} = \frac{1}{q_k}.$$

Отсюда

$$\begin{aligned} k \operatorname{Arch} \frac{1}{\rho_0} &:= k \ln \frac{1 + \sqrt{1 - \rho_0^2}}{\rho_0} = \ln \left[ \frac{1 + \sqrt{1 - \rho_0^2}}{\rho_0} \right]^k = \\ &= \operatorname{Arch} \frac{1}{q_k} = \ln \frac{1 + \sqrt{1 - q_k^2}}{q_k} \end{aligned}$$

и, следовательно,

$$\frac{1 + \sqrt{1 - q_k^2}}{q_k} = \left[ \frac{1 + \sqrt{1 - \rho_0^2}}{\rho_0} \right]^k =: \rho_1^{-k}. \quad (5.33)$$

Преобразовывая это соотношение к квадратному относительно  $q_k$  уравнению и решая его, находим, что

$$q_k = \frac{2\rho_1^{-k}}{1 + \rho_1^{-2k}} = \frac{2\rho_1^k}{1 + \rho_1^{2k}}, \quad (5.34)$$

где, согласно (5.33), (5.24)

$$\rho_1 = \frac{\rho_0}{1 + \sqrt{1 - \rho_0^2}} = \frac{\frac{1 - \varkappa^{-1}}{1 + \varkappa^{-1}}}{1 + \sqrt{1 - \left(\frac{1 - \varkappa^{-1}}{1 + \varkappa^{-1}}\right)^2}} = \frac{1 - \varkappa^{-1}}{1 + \varkappa^{-1} + 2\varkappa^{-1/2}} = \frac{1 - \varkappa^{-1/2}}{1 + \varkappa^{-1/2}}. \quad (5.35)$$

Из формул (5.30), (5.31) для итерационных параметров  $\tau_j$  видно, что для их вычисления требуется задать число итераций  $k$ . Обычно в качестве условия окончания итерационного процесса берется неравенство

$$\|z^k\| \leq \varepsilon \|z^0\|$$

и числом итераций называется наименьшее из чисел  $k$ , для которого это неравенство выполняется. Из (5.32) следует, что для чебышевского итерационного метода число итераций находится из неравенства  $q_k \leq \varepsilon$ . Решим это неравенство. В силу представления (5.34), исследуемое неравенство преобразуется к виду

$$\frac{2\rho_1^k}{1 + \rho_1^{2k}} \leq \varepsilon.$$

Это неравенство эквивалентно неравенству

$$\left( \rho_1^k - \frac{1 + \sqrt{1 - \varepsilon^2}}{\varepsilon} \right) \left( \rho_1^k - \frac{1 - \sqrt{1 - \varepsilon^2}}{\varepsilon} \right) \geq 0.$$

Поскольку  $\rho_1 < 1$ , то первый сомножитель отрицателен, и должно выполняться условие

$$\rho_1^k \leq \frac{1 - \sqrt{1 - \varepsilon^2}}{\varepsilon} = \frac{\varepsilon}{1 + \sqrt{1 - \varepsilon^2}}.$$

Отсюда

$$k \geq \frac{\ln \frac{1 + \sqrt{1 - \varepsilon^2}}{\varepsilon}}{\ln 1/\rho_1}.$$

Поскольку  $\varepsilon$  обычно мало, то пользуются следующей формулой

$$k \geq \frac{\ln 2/\varepsilon}{\ln 1/\rho_1}. \quad (5.36)$$

Сравним скорости сходимости чебышевского итерационного метода с оптимальным методом простых итераций. Из (5.6) следует, что для оптимального метода простых итераций

$$\|x - x^k\| \leq q \|x - x^{k-1}\|,$$

где

$$q = \frac{1 - \kappa^{-1}}{1 + \kappa^{-1}} = \rho_0$$

и, следовательно,

$$\|x - x^k\| \leq q^k \|x - x^0\|.$$

Если мы и здесь потребуем, чтобы

$$\|x - x^k\| \leq \varepsilon \|x - x^0\|,$$

то для числа итераций  $k$  будем иметь

$$k \ln \frac{1}{\rho_0} \geq \ln \frac{1}{\varepsilon},$$

т.е.

$$k \geq \frac{\ln 1/\varepsilon}{\ln 1/\rho_0}. \quad (5.37)$$

Для плохо обусловленной матрицы  $\kappa^{-1} \ll 1$  и поэтому

$$\frac{1}{\rho_0} = 1 + 2\kappa^{-1} + O(\kappa^{-2}).$$

Отсюда и из (5.37)

$$k \approx \frac{1}{2} \varkappa \ln 1/\varepsilon. \quad (5.38)$$

Для чебышевского итерационного метода в силу (5.36), (5.35), будем иметь

$$\rho_1^{-1} = \frac{1 + \varkappa^{-1/2}}{1 - \varkappa^{-1/2}} = 1 + 2\varkappa^{-1/2} + O(\varkappa^{-1})$$

и

$$k \approx \frac{1}{2} \varkappa^{1/2} \ln 2/\varepsilon.,$$

что много лучше, чем (5.38).

## 5.4 Об устойчивости

К сожалению, вычисления по формулам (5.15) при произвольном использовании итерационных параметров не являются устойчивыми. Связано это с тем обстоятельством, что не все сомножители  $(I - \tau_j C)$  разрешающего оператора из (5.19) имеют ограниченную единичной норму. При компьютерных вычислениях всегда присутствуют ошибки округления, и последовательное использование в вычислениях матриц  $(I - \tau_j C)$  с превосходящими единичную норму приводит к накоплению этих погрешностей и неустойчивости вычислительного процесса. Проиллюстрировать это можно на простом примере перемножения нескольких чисел, среди которых есть очень большие и очень маленькие. Пусть  $M_0 = 10^{-p}$  — машинный ноль, а  $M_\infty = 10^p$  — бесконечность. Пусть требуется перемножить следующие числа

$$10^{3p/4}, \quad 10^{p/2}, \quad 10^{p/4}, \quad 10^{-p/2}, \quad 10^{-3p/4}.$$

Очевидно, что искомое произведение равно  $10^{p/4} \in [M_0, M_\infty]$ . Однако, если эти числа перемножать в естественном порядке, то уже после первого умножения получится бесконечность

$$\underbrace{10^{3p/4} \cdot 10^{p/2}} \cdot 10^{p/4} \cdot \underbrace{10^{-p/2} \cdot 10^{-3p/4}} = M_\infty.$$

Перемножение в обратном порядке тоже не приводит к правильному результату — первое же умножение приводит к нулю. Правильный результат получится лишь в том случае, если порядок сомножителей мы будем контролировать, например, так  $10^{-3p/4} 10^{p/2} 10^{3p/4} 10^{-p/2} 10^{p/4}$ . Итерационные параметры (5.30) также можно упорядочить (см. [9]) так, чтобы итерации были устойчивыми.



# 6

## Метод наискорейшего спуска

### 6.1 Метод наискорейшего спуска

Вновь обратимся к решению системы

$$Ax = b, \quad A = A^T > 0. \quad (6.1)$$

Будем для простоты использовать явный нестационарный метод

$$\frac{x^{k+1} - x^k}{\tau_{k+1}} + Ax^k = b, \quad k = 0, 1, \dots \quad (6.2)$$

В предыдущей лекции был указан способ выбора итерационных параметров  $\tau_k$ , использующий априорную информацию о расположении спектра матрицы  $A$ . Сейчас мы рассмотрим другой способ выбора этих параметров.

Пусть, как обычно,  $z^k = x^k - x$ . Тогда

$$z^{k+1} = z^k - \tau_{k+1}Az^k. \quad (6.3)$$

Вычислим  $A$ -норму погрешности  $z^{k+1}$  и выразим ее через  $z^k$ . Используя (6.3), находим, что

$$\begin{aligned} \|z^{k+1}\|_A^2 &= (Az^{k+1}, z^{k+1}) = (A(z^k - \tau_{k+1}Az^k), z^k - \tau_{k+1}Az^k) = \\ &= \|z^k\|_A^2 - 2\tau_{k+1} (Az^k, Az^k) + \tau_{k+1}^2 (A^2z^k, Az^k). \end{aligned}$$

Выберем  $\tau_{k+1}$  из условия минимума  $\|z^{k+1}\|_A^2$ . Дифференцируя по  $\tau_{k+1}$  и приравнявая производную нулю, найдем, что

$$-2 (Az^k, Az^k) + 2\tau_{k+1} (A^2z^k, Az^k) = 0,$$

т.е.

$$\tau_{k+1} = \frac{(Az^k, Az^k)}{(A^2z^k, Az^k)}. \quad (6.4)$$

Казалось бы, что это соотношение не позволяет найти интересующий нас параметр  $\tau_{k+1}$ , поскольку  $z^{k+1} = x^k - x$  не известна. Но нам она и не нужна. Нам нужна

$$Az^k = Ax^k - Ax = Ax^k - b = -r^k.$$

Величина  $r^k$  называется *невязкой*. Тем самым,

$$\tau_{k+1} = \frac{(r^k, r^k)}{(Ar^k, r^k)}, \quad r^k = b - Ax^k. \quad (6.5)$$

Метод (6.2), (6.5) называется методом скорейшего спуска.

Дадим геометрическую интерпретацию этого метода, которая и объяснит его название. Пусть

$$J(x) = \frac{1}{2}(Ax, x) - (b, x) \quad (6.6)$$

— квадратичная функция  $n$  переменных  $x_1, x_2, \dots, x_n$ . Поставим задачу об отыскании точки минимума этой функции. Для решения этой задачи нужно найти первые производные (6.6) по  $x_1, x_2, \dots, x_n$  и приравнять их нулю. Это и будут уравнения для нахождения точки минимума. Перепишем (6.6) в координатном виде

$$J(x) = \frac{1}{2} \sum_{k,j=1}^n a_{kj}x_kx_j - \sum_{k=1}^n b_kx_k$$

и продифференцируем по  $x_i$

$$\frac{\partial J(x)}{\partial x_i} = \frac{1}{2} \sum_{j=1}^n a_{ij}x_j + \frac{1}{2} \sum_{k=1}^n a_{ki}x_k - b_i = \sum_{j=1}^n a_{ij}x_j - b_i, \quad i = 1, \dots, n.$$

**Замечание 6.1.**  $\text{grad } J = Ax - b$ .

Из математического анализа известно, что функция наиболее быстро убывает в направлении анти-градиента.

Итак, задача отыскания точки минимума функции (6.6) эквивалентна решению системы (6.1) с симметричной матрицей. Если мы найдем способ

приближенного нахождения точки минимума функции (6.6), то мы будем иметь метод приближенного нахождения решения системы (6.1).

Построим метод минимизации (6.6). Применительно к (6.6)

$$\nabla J(x) = Ax - b,$$

и процесс минимизации принимает вид

$$x^{k+1} = x^k - \alpha \nabla J(x^k) = x^k - \alpha(Ax^k - b) = x^k + \alpha r^k \quad (6.7)$$

или

$$\frac{x^{k+1} - x^k}{\alpha} + Ax^k = b,$$

что совпадает с (6.2) с точностью до обозначения итерационного параметра. Для определения значения  $\alpha$  рассмотрим

$$J(x^{k+1}) = J(x^k - \alpha \nabla J(x^k)) \quad (6.8)$$

как функцию  $\alpha$  и найдем такое значение  $\alpha$ , при котором  $J$  принимает наименьшее значение. В нашем случае

$$\begin{aligned} J(x^k - \alpha \nabla J(x^k)) &= \frac{1}{2} (A(x^k - \alpha(Ax^k - b)), x^k - \alpha(Ax^k - b)) - \\ &- (b, x^k - \alpha(Ax^k - b)) = \frac{1}{2} (A(x^k + \alpha r^k), x^k + \alpha r^k) - (b, x^k + \alpha r^k). \end{aligned}$$

Дифференцируя по  $\alpha$  и приравнивая производную нулю, находим, что

$$(A(x^k + \alpha r^k), r^k) - (b, r^k) = 0,$$

откуда

$$\alpha = \alpha_{k+1} = \frac{(r^k, r^k)}{(Ar^k, r^k)},$$

что совпадает с ранее полученным значением (6.5) итерационного параметра. Тем самым, оба метода совпадают, а второй метод дает им название.

Имеет место

**Теорема 6.1.** *Итерации по методу скорейшего спуска (6.2), (6.5) сходятся не медленнее, чем в оптимальном методе простых итераций. Именно*

$$\|x^k - x\|_A \leq \left( \frac{1 - \kappa^{-1}}{1 + \kappa^{-1}} \right)^k \|x^0 - x\|_A,$$

где  $\kappa$  — число обусловленности матрицы  $A$ .

**Доказательство.** Из (6.3)

$$\|z^{k+1}\|_A = \|(I - \tau_{k+1}A)z^k\|_A,$$

причем правая часть принимает минимальное значение именно при  $\tau_{k+1}$  из (6.5). Тем самым, при любом другом значении  $\tau_{k+1}$  правая часть будет только больше, и, следовательно,

$$\|z^{k+1}\|_A^2 \leq \|(I - \tau A)z^k\|_A^2$$

для любого  $\tau$  и, в частности, для

$$\tau = \frac{2}{\lambda_1 + \lambda_n}$$

— итерационного параметра метода простых итераций.

Но

$$\begin{aligned} \|(I - \tau A)z^k\|_A^2 &= ((I - \tau A)z^k, (I - \tau A)Az^k) = \\ &= \left( (I - \tau A)A^{1/2}z^k, (I - \tau A)A^{1/2}z^k \right) = \\ &= \|(I - \tau A)A^{1/2}z^k\|^2 \leq \|I - \tau A\| \|z^k\|_A^2, \end{aligned}$$

а в силу (8.22), (8.27), (8.28), при  $k = 1$

$$\|I - \tau A\| \leq \max_{\lambda \in [\lambda_1, \lambda_n]} |1 - \tau\lambda| = q_1 = \rho_0 = \frac{1 - \varkappa^{-1}}{1 + \varkappa^{-1}}.$$

Собирая оценки для всех  $k$ , получим утверждение теоремы.

Из теоремы 6.1 следует, что нестационарный метод (6.2), (6.5) сравним по скорости сходимости с методом простых итераций, и, казалось бы, мы с этим методом не продвинулись вперед. Однако, у этих методов имеется существенное различие. Для использования метода простых итераций требуется информация о границах спектра матрицы  $A$ . В случае же метода (6.2), (6.5) такая информация не требуется.

## 6.2 Неулучшаемость оценки

Покажем на примере, что полученная оценка сходимости достигается, если начальное приближение задано специальным образом. Допустим, что  $x^0$  таково, что

$$x^0 - x = z^0 = c \left( \sqrt{\frac{\lambda_n}{\lambda_1}} \xi_1 + \sqrt{\frac{\lambda_1}{\lambda_n}} \xi_n \right),$$

где  $\lambda_1$  и  $\lambda_n$  суть минимальное и максимальное собственные значения матрицы  $A$  из (6.1), а  $\xi_1$  и  $\xi_n$  — отвечающие им ортонормированные собственные векторы. Тогда

$$\begin{aligned} Az^0 &= c \sqrt{\lambda_1 \lambda_n} (\xi_1 + \xi_n), \\ A^2 z^0 &= c \sqrt{\lambda_1 \lambda_n} (\lambda_1 \xi_1 + \lambda_n \xi_n), \\ (Az^0, Az^0) &= c^2 \lambda_1 \lambda_n 2, \\ (A^2 z^0, A^2 z^0) &= c^2 \lambda_1 \lambda_n (\lambda_1 + \lambda_n). \end{aligned}$$

В силу (6.4)

$$\tau_1 = \frac{(Az^0, Az^0)}{(A^2 z^0, A^2 z^0)} = \frac{2}{\lambda_1 + \lambda_n},$$

а в силу (6.3)

$$\begin{aligned} z^1 &= c \sqrt{\frac{\lambda_n}{\lambda_1}} \xi_1 + c \sqrt{\frac{\lambda_1}{\lambda_n}} \xi_n - \frac{2}{\lambda_1 + \lambda_n} c \sqrt{\lambda_1 \lambda_n} (\xi_1 + \xi_n) = \\ &= c \left[ \sqrt{\frac{\lambda_n}{\lambda_1}} \xi_1 \left( 1 - \frac{2\lambda_1}{\lambda_1 + \lambda_n} \right) + \sqrt{\frac{\lambda_1}{\lambda_n}} \xi_n \left( 1 - \frac{2\lambda_n}{\lambda_1 + \lambda_n} \right) \right] = \\ &= c \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} \left[ \sqrt{\frac{\lambda_n}{\lambda_1}} \xi_1 - \sqrt{\frac{\lambda_1}{\lambda_n}} \xi_n \right] = \\ &= c \rho \left[ \sqrt{\frac{\lambda_n}{\lambda_1}} \xi_1 - \sqrt{\frac{\lambda_1}{\lambda_n}} \xi_n \right]. \end{aligned}$$

Поскольку

$$\|z^0\|_A^2 = (Az^0, z^0) = c^2 \sqrt{\lambda_1 \lambda_n} \left( \sqrt{\frac{\lambda_n}{\lambda_1}} + \sqrt{\frac{\lambda_1}{\lambda_n}} \right) = c^2 (\lambda_1 + \lambda_n),$$

а

$$\begin{aligned} Az^1 &= c \rho \left( \sqrt{\lambda_1 \lambda_n} \xi_1 - \sqrt{\lambda_1 \lambda_n} \xi_n \right), \\ \|z^1\|_A^2 &= c^2 \rho^2 (\lambda_n + \lambda_1) = \rho^2 \|z^0\|_A^2, \end{aligned}$$

то

$$\|z^1\|_A = \rho \|z^0\|_A.$$

Делая следующую итерацию, найдем, что

$$z^2 = \rho^2 z^0$$

и т.д. Отсюда вытекает, что

$$\|z^k\|_A = \rho^k \|z^0\|_A,$$

т.е. полученная оценка точная.

Следует, однако, заметить, что такие плохие начальные приближения в реальных задачах практически не встречаются, и итерации, особенно на начальном этапе, сходятся много быстрее. По мере увеличения числа итераций скорость сходимости уменьшается и выходит на ту, которая гарантируется оценкой. Имея хорошее начальное приближение, можно получить приближенное решение с хорошей точностью при существенно меньших трудозатратах.

# 7

## Метод сопряженных градиентов

Построим другой метод минимизации функции

$$J(x) = \frac{1}{2}(Ax, x) - (b, x), \quad (7.1)$$

точка минимума которой совпадает с решением системы

$$Ax = b, \quad A = A^T > 0. \quad (7.2)$$

### 7.1 Построение метода

В методе скорейшего спуска на каждом шаге происходила одномерная минимизация вдоль направления, задаваемого антиградиентом, который совпадает с невязкой  $r^k = b - Ax^k$ . Рассмотрим теперь последовательную минимизацию  $J(x)$  вдоль совокупности направлений  $\{p^1, p^2, \dots\}$ , которые не обязаны совпадать с направлениями невязок  $\{r^0, r^1, \dots\}$ .

Пусть направления  $p^1, p^2, \dots$  заданы, и

$$x^{k+1} = x^k + \alpha_{k+1}p^{k+1}, \quad k = 0, 1, \dots \quad (7.3)$$

Поскольку

$$\begin{aligned} J(x+y) &= \frac{1}{2}((x+y), A(x+y)) - (b, x+y) = \\ &= \frac{1}{2}\|x\|_A^2 + \frac{1}{2}\|y\|_A^2 + (Ax, y) - (b, x) - (b, y) = \\ &= J(x) + (Ax - b, y) + \frac{1}{2}\|y\|_A^2, \end{aligned} \quad (7.4)$$

то

$$J(x^{k+1}) = J(x^k + \alpha p^{k+1}) = J(x^k) - \alpha(r^k, p^{k+1}) + \frac{\alpha^2}{2}(Ap^{k+1}, p^{k+1}), \quad (7.5)$$

а, дифференцируя это выражение по  $\alpha$  и приравнивая производную нулю, находим итерационный параметр (ср. с (6.5))

$$\alpha_{k+1} = \frac{(r^k, p^{k+1})}{(p^{k+1}, Ap^{k+1})}. \quad (7.6)$$

Подставляя (7.6) в (7.5), найдем, что

$$\begin{aligned} J(x^{k+1}) &= J(x^k) - \frac{(r^k, p^{k+1})}{(p^{k+1}, Ap^{k+1})}(r^k, p^{k+1}) + \frac{1}{2} \frac{(r^k, p^{k+1})^2}{(p^{k+1}, Ap^{k+1})^2} (Ap^{k+1}, p^{k+1}) = \\ &= J(x^k) - \frac{1}{2} \frac{(r^k, p^{k+1})^2}{(p^{k+1}, Ap^{k+1})}, \end{aligned} \quad (7.7)$$

т.е. на  $(k+1)$  итерации действительно будет происходить уменьшение функции  $J(x)$ , если выполнено условие

$$(r^k, p^{k+1}) \neq 0. \quad (7.8)$$

**Замечание 7.1.** Без ограничения общности можно предполагать, что

$$x^0 = 0. \quad (7.9)$$

Если бы нам было известно хорошее приближение  $\tilde{x}$ , то, делая замену  $x = \tilde{x} + z$ , мы бы нашли, что  $Az + A\tilde{x} = b$  и  $Az = b - A\tilde{x}$ . Тем самым, для  $z$  начальным приближением было бы  $z^0 = 0$ .

Из (7.3) следует, что при начальном приближении (7.9) векторы  $x^k$  являются линейными комбинациями векторов  $p^1, p^2, \dots, p^k$ , т.е.

$$x^k \in \text{span}\{p^1, p^2, \dots, p^k\}. \quad (7.10)$$

При выборе направлений  $p^i$  наша задача состоит в том, чтобы гарантировать сходимость и добиться скорости сходимости бóльшей, чем у метода скорейшего спуска. Представляется, что наилучшим способом выбора  $p^i$  был бы такой, при котором  $x^{k+1}$  минимизировал бы функцию  $J(x)$  не только по направлению  $p^{k+1}$ , но и по всему подпространству  $\text{span}\{p^1, p^2, \dots, p^{k+1}\} \subset \mathbb{R}^n$ , т.е.

$$J(x^{k+1}) = \min_{x \in \text{span}\{p^1, p^2, \dots, p^{k+1}\}} J(x). \quad (7.11)$$

Если бы такой выбор  $p^i$  удалось осуществить, то это не только гарантировало бы сходимость, но привело бы к конечности итерационного процесса,



ибо при  $k + 1 = n$  и линейно независимых  $p^i$  задача (7.11) представляет собой исходную задачу глобальной минимизации, и, следовательно,  $Ax^n = b$ .

Попытаемся решить поставленную задачу. Пусть

$$P_k = [p^1 p^2 \dots p^k]$$

есть  $(n \times k)$ -матрица, столбцами которой являются искомые направления. Пусть  $x = P_k y + \alpha p^{k+1} \in \text{im } P_{k+1}$ ,<sup>1</sup>  $y \in \mathbb{R}^k$ ,  $\alpha \in \mathbb{R}$ . Тогда (см. (7.5))

$$J(x) = J(P_k y) + \alpha(A P_k y, p^{k+1}) - \alpha(b, p^{k+1}) + \frac{\alpha^2}{2}(A p^{k+1}, p^{k+1}). \quad (7.12)$$

Если бы в (7.12) отсутствовал "перекрестный" член

$$\alpha(P_k y, A p^{k+1}),$$

то задача минимизации  $J(x)$  на  $\text{span} \{p^1, p^2, \dots, p^{k+1}\} = \text{im } P_{k+1}$ , т.е. задача (7.11), распалась бы на минимизацию по  $\text{im } P_k$ , где решение  $x^k$  предполагается известным, и простую минимизацию для определения скалярной величины  $\alpha$ .

В самом деле, пусть выполнены условия

$$(p^i, A p^j) = 0, \quad i \neq j. \quad (7.13)$$

(Векторы, удовлетворяющие условию (7.13), называются  $A$ -сопряженными или

$A$ -ортогональными.) Определим вектор  $x^k \in \text{im } P_k$  и  $\alpha_{k+1} \in \mathbb{R}$  следующим образом

$$J(x^k) = \min_y J(P_k y), \quad \alpha_{k+1} = \frac{(b, p^{k+1})}{(p^{k+1}, A p^{k+1})}. \quad (7.14)$$

Тогда (см. (7.12))

$$\min_{y, \alpha} J(P_k y + \alpha p^{k+1}) = \min_y J(P_k y) + \min_{\alpha} \left\{ \frac{\alpha^2}{2}(A p^{k+1}, p^{k+1}) - \alpha(b, p^{k+1}) \right\}$$

находится при  $P_k y = x^k$  и  $\alpha = \alpha_{k+1}$  из (7.14). На самом деле  $\alpha_{k+1}$  из (7.14) совпадает с (7.6), ибо в силу (7.10) и (7.13)

$$(A p^{k+1}, x^k) = 0$$

---

<sup>1</sup>Напомним, что множество всех векторов  $x$ , представимых в виде  $x = By$ , называется образом матрицы  $B$  и обозначается  $\text{im } B$ .

и, следовательно,

$$(p^{k+1}, b) = (p^{k+1}, b - Ax^k + Ax^k) = (p^{k+1}, r^k), \quad (7.15)$$

что вместе с (7.14) приводит к (7.6).

Итак, для реализации задуманного метода нужно последовательно находить  $A$ -сопряженные векторы  $p^1, p^2, \dots, p^{k+1}$ , для которых выполнено условие (7.8), и проводить вычисления по формуле (7.3) с параметром  $\alpha_{k+1}$  из (7.6).

Обратимся к наиболее целесообразному выбору векторов  $p^{k+1}$ . При выборе  $p^{k+1}$  наша цель состоит в быстрой минимизации функции  $J(x)$ , и в силу (7.7) мы должны максимизировать

$$\frac{(r^k, p^{k+1})^2}{(p^{k+1}, Ap^{k+1})}. \quad (7.16)$$

**Замечание 7.2.** Эта величина не зависит от длины вектора  $p^{k+1}$ , а зависит только от его направления. Поэтому при отыскании  $p^{k+1}$  достаточно ограничиться нахождением его направления.

Поскольку  $p^{k+1}$  должен еще удовлетворять условиям  $A$ -сопряженности (7.13), т.е. быть ортогональным к  $\{Ap^1, Ap^2, \dots, Ap^k\} = \text{im } AP_k$ , то искомый вектор  $p^{k+1}$  должен принадлежать ортогональному дополнению подпространства  $\text{im } AP_k$  — подпространству  $(\text{im } AP_k)^\perp$

$$p^{k+1} \in (\text{im } AP_k)^\perp.$$

Пусть  $r^k = r_{\parallel}^k + r_{\perp}^k$ , где  $r_{\parallel}^k \in \text{im } (AP_k)$ , а  $r_{\perp}^k \in (\text{im } (AP_k))^\perp$ . Тогда

$$(r^k, p^{k+1}) = (r_{\parallel}^k + r_{\perp}^k, p^{k+1}) = (r_{\perp}^k, p^{k+1}) = \|r_{\perp}^k\| \|p^{k+1}\| \cos(r_{\perp}^k, p^{k+1}),$$

и искомый максимум выражения (7.16) будет достигаться при достижении максимума последним сомножителем, т.е. при  $|\cos(r_{\perp}^k, p^{k+1})| = 1$ . Это будет так, если, например,

$$p^{k+1} = r_{\perp}^k \in (\text{im } (AP_k))^\perp \quad (7.17)$$

— ортогональной проекции  $r^k$  на  $(\text{im } (AP_k))^\perp$ . Отметим, что отсюда следует соотношение

$$p^1 = r^0. \quad (7.18)$$

Построение процесса минимизации  $J(x)$  в первом приближении будет закончено, если принять во внимание, что имеет место

**Теорема 7.1.** *Два последовательных направления спуска в методе сопряженных градиентов связаны соотношением*

$$p^{k+1} = r^k + \beta_{k+1}p^k. \quad (7.19)$$

Доказательство этой теоремы мы отложим на потом, а сейчас заметим, что поскольку векторы  $p^k$  и  $p^{k+1}$  должны быть  $A$ -сопряжены, то для параметра  $\beta_{k+1}$  из (7.19) имеет место представление

$$\beta_{k+1} = -\frac{(r^k, Ap^k)}{(p^k, Ap^k)}. \quad (7.20)$$

Итак, метод сопряженных градиентов состоит в вычислениях по следующим формулам

$$\begin{aligned} r^k &= b - Ax^k, & k &= 0, 1, \dots, \\ p^{k+1} &= r^k + \beta_{k+1}p^k, & k &= 1, 2, \dots, & p^1 &= r^0, \\ x^{k+1} &= x^k + \alpha_{k+1}p^{k+1}, & k &= 0, 1, \dots, & x^0 &= 0, \\ \alpha_{k+1} &= (r^k, p^{k+1})/(p^{k+1}, Ap^{k+1}), & k &= 0, 1, \dots, \\ \beta_{k+1} &= -(Ap^k, r^k)/(Ap^k, p^k), & k &= 1, 2, \dots, \end{aligned} \quad (7.21)$$

## 7.2 Оценка скорости сходимости

Дадим оценку скорости сходимости метода сопряженных градиентов. Имеет место

**Теорема 7.2.** *Метод сопряженных градиентов (7.21) сходится не хуже, чем чебышевский итерационный метод, т.е.*

$$\|x^k - x\|_A \leq 2 \left( \frac{1 - \varkappa^{-1/2}}{1 + \varkappa^{-1/2}} \right)^k \|x\|_A,$$

где  $\varkappa$  — число обусловленности матрицы  $A$ .

**Доказательство.** Отметим, что минимизация  $J(x^k)$  ведет к минимизации

$\|x^k - x\|_A$ . В самом деле, пусть

$$z^k = x^k - x.$$

Тогда, подставляя  $x^k = x + z^k$  в  $J(x^k)$  и принимая во внимание (7.4), будем иметь

$$J(x^k) = \frac{1}{2} \|z^k\|_A^2 + J(x). \quad (7.22)$$

Установим теперь связь между  $z^k$  на последовательных итерациях. Из третьего соотношения (7.21) находим, что

$$p^{k+1} = (x^{k+1} - x^k) / \alpha_{k+1}.$$

Подставим это представление  $p^{k+1}$  во второе соотношение (7.21)

$$\frac{x^{k+1} - x^k}{\alpha_{k+1}} - \beta_{k+1} \frac{x^k - x^{k-1}}{\alpha_k} = b - Ax^k.$$

Отсюда находим, что

$$\frac{z^{k+1} - z^k}{\alpha_{k+1}} - \beta_{k+1} \frac{z^k - z^{k-1}}{\alpha_k} + Az^k = 0.$$

Далее

$$z^1 = z^0 + \alpha_1 p^1 = z^0 + \alpha_1 r^0 = z^0 + \alpha_1 b = z^0 + \alpha_1 Ax = z^0 - \alpha_1 Az^0.$$

Тем самым (не путать  $P_k$  и  $P_k(A)$ ),

$$z^k = P_k(A)z^0, \quad P_k(0) = 1$$

и

$$\|z^k\|_A = \|P_k(A)z^0\|_A.$$

Но по построению  $x^k$  и с учетом (7.22)

$$\|z^k\|_A = \min_{Q_k} \|Q_k(A)z^0\|_A, \quad Q_k(0) = 1$$

и, следовательно, для любого  $Q_k(A)$

$$\begin{aligned} \|z^k\|_A &\leq \|Q_k(A)z^0\|_A = \|Q_k(A)A^{1/2}z^0\|_2 \leq \|Q_k(A)\| \|z^0\|_A \leq \\ &\leq \max_{\lambda_1 \leq t \leq \lambda_n} |Q_k(t)| \|z^0\|_A = \max_{y \in [-1, 1]} \left| Q_k \left( \frac{\lambda_n + \lambda_1}{2} - \frac{\lambda_n - \lambda_1}{2} y \right) \right| \|z^0\|_A = \\ &= \max_{y \in [-1, 1]} |\hat{Q}_k(y)| \|z^0\|_A. \end{aligned}$$

Если положить  $\hat{Q}_k(y) = q_k T_k(y)$  (см. гл. 11), то

$$\|z^k\|_A \leq q_k \|z^0\|_A = \frac{2\rho_1^k}{1 + 2\rho_1^{2k}} \|z^0\|_A \leq 2 \left( \frac{1 - \varkappa^{-1/2}}{1 + \varkappa^{-1/2}} \right)^k \|z^0\|_A.$$

Теорема доказана.

### 7.3 Вспомогательные утверждения

Чтобы описание метода сопряженных градиентов (7.21) было корректным, нужно доказать теорему 7.1 (о связи между двумя последовательными направлениями спуска). Для этого нам понадобится ряд вспомогательных утверждений.

**Лемма 7.1.** Пусть  $p^1, p^2, \dots, p^k$  суть ненулевые  $A$ -сопряженные векторы. Тогда, либо существует ненулевой  $A$ -сопряженный к ним вектор  $p^{k+1}$ , удовлетворяющий условию (7.8), либо  $r^k = 0$ .

**Доказательство.** Пусть не существует такого вектора  $p^{k+1}$ ,  $A$ -сопряженного с  $p^1, p^2, \dots, p^k$ , для которого выполнено условие (7.8), т.е. для любого вектора  $p \perp \text{im } AP_k$  (или  $Ap \perp \text{im } P_k$ , или  $Ap \in (\text{im } P_k)^\perp$ ) справедлива цепочка равенств

$$0 = (r^k, p) = (b - Ax^k, p) = (b, p) = (Ap, x).$$

Тем самым, решение  $x \in \text{im } P_k$ . Но  $x^k$  минимизирует  $J(x)$  на  $\text{im } P_k$  и, следовательно,  $x^k = x$ , т.е.  $r^k = 0$ .

Отсюда же вытекает, что, если  $r^k \neq 0$ , то существует  $p^{k+1}$  из (7.19), для которого выполнено условие (7.8). Лемма доказана.

**Замечание 7.3.** Лемма 7.1 утверждает, что либо мы на  $k$ -ой итерации закончили вычисления, получив точное решение задачи (7.2), либо имеем возможность вычисления продолжить.

**Теорема 7.3.** После  $k$  итераций метода сопряженных градиентов при каждом  $j = 1, 2, \dots, k$

$$\begin{aligned} \text{span } \{p^1, p^2, \dots, p^j\} &= \text{span } \{r^0, r^1, \dots, r^{j-1}\} = \\ &= \mathcal{K}_j(A, b) := \text{span } \{b, Ab, \dots, A^{j-1}b\}. \end{aligned} \quad (7.23)$$

**Доказательство** проведем методом полной математической индукции. При  $j = 1$  соотношения (7.23) имеют место, ибо в силу выбора (7.9) начального приближения  $r^0 = b$ , а из (7.18)  $p^1 = r^0$ . Предположим, что (7.23) справедливы при некотором  $j$ , удовлетворяющем неравенству  $1 \leq j < k$ . Докажем их справедливость при  $j + 1$ .

В качестве первого шага покажем, что

$$\text{span } \{p^1, p^2, \dots, p^{j+1}\} \subset \text{span } \{r^0, r^1, \dots, r^j\}. \quad (7.24)$$

В силу (7.17)

$$p^{j+1} = r_{\perp}^j = r^j - r_{\parallel}^j, \quad r_{\parallel}^j \in \text{im}[AP_j]$$

и, следовательно,

$$p^{j+1} = r^j - AP_j y_j, \quad y_j \in \mathbb{R}^j. \quad (7.25)$$

Из третьего соотношения (7.21) вытекает, что

$$Ax^j = Ax^{j-1} + \alpha_j Ap^j.$$

Вычитая из обеих частей этого равенства по  $b$ , будем иметь

$$r^j = r^{j-1} - \alpha_j Ap^j \quad (7.26)$$

и, следовательно,

$$Ap^j = -(r^j - r^{j-1})/\alpha_j.$$

Подставляя это представление в (7.25), находим, что

$$p^{j+1} = r^j + \left[ \frac{r^1 - r^0}{\alpha_1} \frac{r^2 - r^1}{\alpha_2} \dots \frac{r^j - r^{j-1}}{\alpha_j} \right] y_j = r^j + R_{j+1} \tilde{y}_j \in \text{im } R_{j+1},$$

где  $R_{j+1} = [r^0 r^1 \dots r^j]$ , а  $\tilde{y}_j \in \mathbb{R}^{j+1}$ . Отсюда и из предположения индукции следует включение (7.24).

Теперь установим включение

$$\text{span} \{r^0, r^1, \dots, r^j\} \subset \mathcal{K}_{j+1}(A, b). \quad (7.27)$$

По предположению индукции векторы  $p^j \in \mathcal{K}_j(A, b)$ . Поэтому

$$Ap^j \in \mathcal{K}_{j+1}(A, b).$$

Примем во внимание предположение индукции  $r^{j-1} \in \mathcal{K}_j(A, b)$ . Тогда из (7.26) найдем, что

$$r^j \in \mathcal{K}_{j+1}(A, b).$$

Вновь принимая во внимание предположение индукции, будем иметь желаемое включение (7.27).

Итак, вместо равенства (7.23) мы пока имеем только включения (7.24), (7.27) пространств, размерность каждого из которых не превышает  $j + 1$ . Поскольку векторы  $p^1, p^2, \dots, p^{j+1}$  ненулевые и  $A$ -сопряженные, то

$$\dim \text{span} \{p^1, p^2, \dots, p^{j+1}\} = j + 1.$$

Отсюда и из включений (7.24), (7.27) следует искомое равенство (7.23).

**Определение 7.1.** Подпространства  $\mathcal{K}_j(A, b)$  называются подпространствами Крылова.

**Лемма 7.2.** После  $k$  итераций по методу сопряженных градиентов невязка  $r^j$  ортогональна всем векторам спуска  $p^1, \dots, p^j$ , т.е.

$$P_j^T r^j = 0, \quad j = 1, \dots, k, \quad (7.28)$$

**Доказательство.** В силу (7.10) существует вектор  $y_j \in \mathbb{R}^j$  такой, что  $x^j = P_j y_j$ . Поэтому

$$\begin{aligned} J(x^j) &= J(P_j y_j) = \frac{1}{2}(AP_j y_j, P_j y_j)_n - (b, P_j y_j)_n = \\ &= \frac{1}{2}[P_j y_j]^T [AP_j y_j] - [P_j y_j]^T b = \frac{1}{2}y_j^T P_j^T AP_j y_j - y_j^T P_j^T b = \\ &= \frac{1}{2}([P_j^T AP_j]y_j, y_j)_j - (P_j^T b, y_j)_j, \end{aligned}$$

т.е.  $y_j$  есть решение задачи минимизации с матрицей  $P_j^T AP_j$  и вектором  $P_j^T b$ , и, следовательно, вектор  $y_j$  является решением следующей системы

$$[P_j^T AP_j]y_j = P_j^T b.$$

Отсюда

$$P_j^T r^j = P_j^T (b - Ax^j) = P_j^T (b - AP_j y_j) = 0.$$

Лемма доказана.

**Теорема 7.4.** После  $k$  шагов метода сопряженных градиентов невязки  $r^0, r^1, \dots, r^k$  взаимно ортогональны.

**Доказательство.** В силу теоремы 7.3

$$p^j \in \text{span} \{r^0, r^1, \dots, r^{j-1}\}.$$

Это означает, что  $p^1$  выражается только через  $r^0$ ,  $p^2$  — только через  $r^0$  и  $r^1$  и т.д., т.е.

$$P_j = [p^1 p^2 \dots p^j] = [r^0 r^1 \dots r^{j-1}] U_j =: R_j U_j, \quad (7.29)$$

где  $U_j$  — верхняя треугольная  $(j \times j)$ -матрица.

Поскольку векторы  $p^1, p^2, \dots, p^j$   $A$ -ортогональны, а в силу леммы 7.1 и ненулевые, то они линейно независимы. Тогда в силу теоремы 7.3 и

векторы  $r^0, r^1, \dots, r^{j-1}$  линейно независимы. Поэтому  $U_j$  — невырожденная матрица. Подставляя представление (7.29) матрицы  $P_j$  в (7.28), будем иметь

$$0 = U_j^T R_j^T r^j = U_j^T [(r^0, r^j) (r^1, r^j) \dots (r^{j-1}, r^j)]^T.$$

Рассматривая это соотношение как систему линейных однородных уравнений относительно  $(r^i, r^j)$  с невырожденной матрицей, приходим к заключению, что  $(r^i, r^j) = 0$  при  $i \neq j$ . Теорема доказана.

Мы теперь имеем все необходимое для того, чтобы доказать теорему 7.1, т.е.

$$p^{k+1} = r^k + \beta_{k+1} p^k. \quad (7.30)$$

**Доказательство** теоремы 7.1. В силу (7.17)

$$p^{k+1} = r_{\perp}^k = r^k - r_{\parallel}^k, \quad r_{\parallel}^k \in \text{im } AP_k,$$

т.е.

$$p^{k+1} = r^k - \sum_{j=1}^k c_{k+1,j} A p^j.$$

Поскольку в силу теоремы 7.3 вектор  $p^j \in \mathcal{K}_j(A, b)$ , то

$$A p^j \in \mathcal{K}_{j+1}(A, b), \quad (7.31)$$

а, снова используя теорему 7.3 находим, что  $A p^j \in \text{span} \{p^1, p^2, \dots, p^{j+1}\}$  и, следовательно,

$$p^{k+1} = r^k - \sum_{j=1}^{k+1} d_{k+1,j} p^j$$

или

$$(1 + d_{k+1,k+1}) p^{k+1} = r^k - \sum_{j=1}^k d_{k+1,j} p^j. \quad (7.32)$$

Коэффициент  $(1 + d_{k+1,k+1})$  при  $p^{k+1}$  нулю не равен, ибо в противном случае

$$r^k = \sum_{j=1}^k d_{k+1,j} p^j = P_k [d_{k+1,1} d_{k+1,2} \dots d_{k+1,k}]^T = P_k d_k$$

и с учетом леммы 7.2

$$\|r^k\|^2 = r^{kT} r^k = r^{kT} P_k d_k = (r^{kT} P_k d_k)^T = d_k^T (P_k^T r^k) = 0,$$

т.е.  $r^k = 0$ , что в силу леммы 7.1 возможно лишь по завершении итераций.



Так как вектор  $p^{k+1}$  из (7.32) должен быть  $A$ -ортогонален векторам  $p^m$ ,  $m = 1, 2, \dots, k$ , то

$$(r^k, Ap^m) = \sum_{j=0}^k d_{k+1,j}(p^j, Ap^m) = d_{k+1,m}(p^m, Ap^m), \quad m = 1, \dots, k,$$

и, следовательно

$$d_{k+1,m} = \frac{(r^k, Ap^m)}{(Ap^m, p^m)}, \quad m = 1, 2, \dots, k. \quad (7.33)$$

Снова обращаясь к (7.31) и теореме 7.3, находим, что

$$Ap^m \in \text{span} \{r^0, r^1, \dots, r^m\},$$

т.е.  $Ap^m = R_{m+1}l_{m+1}$ , а по теореме 7.4

$$(r^k, Ap^m) = \sum_{i=0}^m l_{m+1,i}(r^k, r^i) = 0, \quad m = 1, 2, \dots, k-1.$$

Следовательно,

$$d_{k+1,j} = 0, \quad j = 1, 2, \dots, k-1,$$

а (7.32) принимает вид

$$(1 + d_{k+1,k+1})p^{k+1} = r^k - d_{k+1,k}p^k,$$

где  $d_{k+1,k}$  определяется соотношением (7.33) (ср. с (7.20)), что с точностью до длины вектора  $p^{k+1}$  (см. Замечание 7.2) совпадает с (7.19). Теорема доказана.

## 7.4 Окончательные соотношения

Преобразуем соотношения (7.21) метода сопряженных градиентов. В этих соотношениях наиболее трудоемкими являются две операции: вычисление векторов  $Ax^k$  и  $Ap^k$ . Однако операцию вычисления вектора  $Ax^k$  можно исключить. Поскольку этот вектор используется только при вычислении невязки  $r^k$ , то можно заменить первую из формул (7.21) на (7.26)

$$r^k = r^{k-1} - \alpha_k Ap^k, \quad k = 1, 2, \dots, \quad r^0 = b. \quad (7.34)$$

Преобразуем еще формулы для вычисления параметров  $\alpha_{k+1}$  и  $\beta_{k+1}$ . Подставляя второе из соотношений (7.21) в числитель четвертого и принимая во внимание лемму 7.2, найдем, что

$$\alpha_{k+1} = (r^k, r^k)/(p^{k+1}, Ap^{k+1}), \quad k = 0, 1, \dots \quad (7.35)$$

Далее, заменяя здесь  $k + 1$  на  $k$  и подставляя полученное выражение для  $(p^k, Ap^k)$  в последнее из соотношений (7.21), будем иметь

$$\beta_{k+1} = -\alpha_k \frac{(Ap^k, r^k)}{(r^{k-1}, r^{k-1})}.$$

Теперь подставим сюда вместо  $Ap^k$  его выражение из (7.34). Принимая во внимание теорему 7.4 об ортогональности невязок, найдем, что

$$\beta_{k+1} = \frac{(r^k, r^k)}{(r^{k-1}, r^{k-1})}, \quad k = 1, 2, \dots \quad (7.36)$$

С учетом (7.34)-(7.36) формулы метода сопряженных градиентов (7.21) преобразуются к виду

$$\begin{aligned} r^k &= r^{k-1} - \alpha_k Ap^k, \quad k = 1, 2, \dots, \quad r^0 = b - Ax_0, \\ \beta_{k+1} &= \|r^k\|^2 / \|r^{k-1}\|^2, \quad k = 1, 2, \dots, \\ p^{k+1} &= r^k + \beta_{k+1} p^k, \quad k = 1, 2, \dots, \quad p^1 = r^0, \\ \alpha_{k+1} &= \|r^k\|^2 / (p^{k+1}, Ap^{k+1}), \quad k = 0, 1, \dots, \\ x^{k+1} &= x^k + \alpha_{k+1} p^{k+1}, \quad k = 0, 1, \dots, \quad x^0. \end{aligned} \quad (7.37)$$

Легко проверить, что вычисления можно проводить в следующем порядке

$$\begin{aligned} r^0 &= b - Ax_0, \quad p^1 = r^0, \quad Ap^1, \quad \alpha_1, \quad x^1, \\ r^1, \quad \beta_2, \quad p^2, \quad Ap^2, \quad \alpha_2, \quad x^2 \quad \dots \end{aligned}$$

## 7.5 Метод сопряженных градиентов с предобуславливателем

Вновь обратимся к решению системы (7.2), но теперь для улучшения обусловленности матрицы системы домножением ее на  $B^{-1}$ , где  $B = B^T > 0$ . После этого система (7.2) примет вид

$$B^{-1}Ax = B^{-1}b.$$

Поскольку матрица этой системы, вообще говоря, не является симметричной, применять непосредственно к ней метод сопряженных градиентов нельзя. Нужно ее сначала преобразовать. Пусть  $B = B^{1/2}B^{1/2}$ , где  $B^{1/2}$  — симметричная положительно определенная матрица — квадратный корень из  $B$ . Тогда

$$B^{-1/2}B^{-1/2}AB^{-1/2}B^{1/2}x = B^{-1/2}B^{-1/2}b.$$

Пусть

$$B^{1/2}x = y, \quad B^{-1/2}b = f, \quad B^{-1/2}AB^{-1/2} = C = C^T > 0. \quad (7.38)$$

Тогда для отыскания неизвестного вектора  $y$  получим систему

$$Cy = f,$$

к которой можно применить метод сопряженных градиентов (7.37), снабдив векторы невязки  $r^k$  и спуска  $p^k$  тильдами:

$$\begin{aligned} \tilde{r}^k &= \tilde{r}^{k-1} - \alpha_k C \tilde{p}^k, \\ \tilde{p}^{k+1} &= \tilde{r}^k + \beta_{k+1} \tilde{p}^k, \\ y^{k+1} &= y^k + \alpha_{k+1} \tilde{p}^{k+1}, \\ \beta_{k+1} &= \frac{(\tilde{r}^k, \tilde{r}^k)}{(\tilde{r}^{k-1}, \tilde{r}^{k-1})}, \\ \alpha_{k+1} &= \frac{(\tilde{r}^k, \tilde{r}^k)}{(C \tilde{p}^{k+1}, \tilde{p}^{k+1})}. \end{aligned} \quad (7.39)$$

Установим связь этих соотношений с (7.37). По определению и с учетом (7.38)

$$\tilde{r}^k = f - Cy^k = B^{-1/2}b - B^{-1/2}AB^{-1/2}B^{1/2}x^k = B^{-1/2}(b - Ax^k) = B^{-1/2}r^k. \quad (7.40)$$

Далее, векторы  $\tilde{p}^m$  являются  $C$ -сопряженными, так что

$$(C \tilde{p}^k, \tilde{p}^j) = (B^{-1/2}AB^{-1/2} \tilde{p}^k, \tilde{p}^j) = (AB^{-1/2} \tilde{p}^k, B^{-1/2} \tilde{p}^j) = 0, \quad j \neq k,$$

и поэтому можно положить

$$B^{-1/2} \tilde{p}^k = p^k. \quad (7.41)$$

Из (7.40) и (7.41) вытекает, что соотношение для  $\tilde{r}^k$  (7.39) эквивалентно соотношению для  $r^k$  в (7.37). Применяя матрицу  $B^{-1/2}$  ко второму соотношению (7.39) и принимая во внимание (7.41), найдем, что

$$p^{k+1} = B^{-1}r^k + \beta_{k+1}p^k.$$

Положим

$$Bz^k = r^k. \quad (7.42)$$

Тогда соотношение для нахождения  $p^{k+1}$  примет вид

$$p^{k+1} = z^k + \beta_{k+1}p^k.$$

С учетом (7.38) и (7.41) находим, что третьи соотношения в (7.39) и (7.37) эквивалентны.

Осталось преобразовать формулы для  $\beta_k$  и  $\alpha_k$ , выразив эти величины через  $r^k$ ,  $p^k$  и  $z^k$ . С учетом (7.40) и (7.42)

$$\beta_{k+1} = \frac{(B^{-1/2}r^k, B^{-1/2}r^k)}{(B^{-1/2}r^{k-1}, B^{-1/2}r^{k-1})} = \frac{(B^{-1}r^k, r^k)}{(B^{-1}r^{k-1}, r^{k-1})} = \frac{(z^k, r^k)}{(z^{k-1}, r^{k-1})},$$

а

$$\alpha_{k+1} = \frac{(z^k, r^k)}{(Ap^{k+1}, p^k)}.$$

В новых терминах соотношения (7.39) принимают вид

$$\begin{aligned} r^k &= r^{k-1} - \alpha_k p^k, \\ p^{k+1} &= z^k + \beta_k p^k, \\ x^{k+1} &= x^k + \alpha_{k+1} p^{k+1}, \\ \beta_{k+1} &= \frac{(z^k, r^k)}{(z^{k-1}, r^{k-1})}, \\ \alpha_{k+1} &= \frac{(z^k, r^k)}{(Ap^k, p^k)}. \end{aligned}$$

## 7.6 Неполное разложение разреженных матриц

Одним из способов задания предобуславливателя в методе сопряженных градиентов является использование матрицы неполного разложения Холецкого. Здесь мы опишем чуть более общий алгоритм неполного  $LU$  разложения ( $ILLU$ -алгоритм).

**Определение 7.2.** Портретом разреженной матрицы  $A$  называется множество пар индексов  $(i, j)$  таких, что  $a_{ij} \neq 0$ , т.е.

$$P_A := \{(i, j) \mid a_{ij} \neq 0\}.$$

Для матрицы  $A$  строится разложение

$$A = LU + R$$

такое, что

- $P_U \subset P_A, \quad P_L \subset P_A.$
- $\forall (i, j) \in P_A : \quad [LU]_{ij} = [A]_{ij}.$
- $P_A \cap P_R = \emptyset.$

Покажем, как это разложение можно построить. Предположим, что уже найдены  $k-1$  строк матриц  $L$  и  $U$  в неполном разложении и требуется найти  $k$ -е строки. Если

$$A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}, \quad U = \begin{bmatrix} U_1 \\ U_2 \end{bmatrix}, \quad R = \begin{bmatrix} R_1 \\ R_2 \end{bmatrix}, \quad \text{а} \quad L = \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix},$$

где матрицы  $A_1$ ,  $U_1$  и  $R_1$  имеют по  $k$  строк и  $n$  столбцов, а  $L_{11}$  — квадратная ( $k \times k$ ) матрица, то из

$$\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} + \begin{bmatrix} R_1 \\ R_2 \end{bmatrix}$$

следует, что

$$A_1 = L_{11}U_1 + R_1.$$

Перепишем это соотношение в виде

$$\begin{bmatrix} A_{11} & A_{12} \\ \bar{a}_{21} & \bar{a}_{22} \end{bmatrix} = \begin{bmatrix} \tilde{L}_{11} & 0 \\ \bar{l}_{21} & 1 \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ 0 & \bar{u}_{22} \end{bmatrix} + \begin{bmatrix} R_{11} & R_{12} \\ \bar{r}_{21} & \bar{r}_{22} \end{bmatrix},$$

где надчеркнутыми строчными буквами обозначены строки соответствующей длины. Из равенства матриц в левой и правой частях следует, что искомые строки  $\bar{l}_{21}$  и  $\bar{u}_{22}$  должны удовлетворять уравнениям

$$\begin{aligned} \bar{l}_{21}U_{11} &= \bar{a}_{21} - \bar{r}_{21}, \\ \bar{u}_{22} &= \bar{a}_{22} - \bar{l}_{21}U_{12} - \bar{r}_{22}. \end{aligned}$$

Согласно формуле (1.19) для элементов нижней треугольной матрицы  $L$  из  $LU$ -разложения

$$l_{kj} = \frac{1}{u_{jj}} \left[ a_{kj} - r_{kj} - \sum_{i=1}^{j-1} l_{ki}u_{ij} \right].$$

Если  $a_{kj} = 0$ , то  $l_{kj} = 0$ , и это соотношение определяется  $r_{kj}$ . В противном случае, т.е. при  $a_{kj} \neq 0$  элемент  $r_{kj} = 0$ , и эта формула работает обычным образом.

Аналогично имеем

$$u_{kj} = a_{kj} - r_{kj} - \sum_{i=1}^{k-1} l_{ki} u_{ij}.$$

Искомые элементы строки  $\bar{u}_{22}$  находятся из этого соотношения при помощи тех же рассуждений, что и выше.

## III

# Задача на собственные значения

# 8

## Степенной метод и обратные итерации

### 8.1 Постановка задачи

Пусть  $A$  — квадратная матрица с действительными или комплексными коэффициентами, и требуется найти собственные векторы и собственные значения этой матрицы. Напомним

**Определение 8.1.** Число  $\lambda$  называется *собственным значением* матрицы  $A$ , если однородная система

$$A\xi = \lambda\xi \quad (8.1)$$

имеет нетривиальное решение  $\|\xi\| \neq 0$ . Это нетривиальное решение называется *собственным вектором* матрицы  $A$ , отвечающим собственному значению  $\lambda$ .

Собственные значения являются нулями *характеристического многочлена*

$$\det [A - \lambda I] = 0,$$

степень которого совпадает с порядком матрицы и есть  $n$ . Тем самым, у каждой квадратной матрицы существует  $n$  собственных значений, действительных или комплексных, простых или кратных. С собственными векторами ситуация сложнее: их число может быть от 1 до  $n$ .

**Определение 8.2.** Матрицы  $A$  и  $B$  называются *подобными*, если существует невырожденная матрица  $S$  (матрица подобия) такая, что  $B = S^{-1}AS$ . Матрицы  $A$  и  $B$  называются унитарно (ортогонально) подобными, если матрица  $S$  унитарная (ортогональная).



Подобные матрицы имеют одинаковый набор собственных значений. Если  $y$  — собственный вектор матрицы  $B$ , подобной  $A$ , то собственный вектор матрицы  $A$  имеет вид

$$x = Sy.$$

**Теорема 8.1** (Каноническая форма Жордана). *Любая матрица  $A$  преобразованием подобия  $S^{-1}AS$  с подходящей матрицей подобия  $S$  может быть приведена к нормальной (Жордановой) форме, т.е. к такой двухдиагональной матрице, у которой на главной диагонали стоят собственные значения, а на наддиагонали — числа ноль или единица*

$$\begin{bmatrix} \lambda_1 & \sigma_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & \lambda_2 & \sigma_2 & 0 & \dots & 0 & 0 \\ 0 & 0 & \lambda_3 & \sigma_3 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & \lambda_{n-1} & \sigma_{n-1} \\ 0 & 0 & 0 & 0 & \dots & 0 & \lambda_n \end{bmatrix}.$$

**Теорема 8.2** (Каноническая форма Шура). *Для произвольной матрицы  $A \in \mathbb{C}^{n \times n}$  найдется унитарная матрица  $Q$  и верхнетреугольная матрица  $T$  такие, что  $Q^H A Q = T$ . Собственными значениями матрицы  $A$  являются диагональные элементы матрицы  $T$ .*

**Теорема 8.3** (Вещественная каноническая форма Шура). *Для любой матрицы  $A \in \mathbb{R}^{n \times n}$  существует ортогональная матрица  $Q \in \mathbb{R}^{n \times n}$  такая, что*

$$Q^T A Q = \begin{bmatrix} R_{11} & R_{12} & \dots & R_{1n} \\ 0 & R_{22} & \dots & R_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & R_{nn} \end{bmatrix},$$

где каждый блок  $R_{ii}$  является либо действительным числом, либо матрицей второго порядка, имеющей комплексно сопряженные собственные значения.

**Определение 8.3.** Матрица  $A$  называется матрицей простой структуры (или диагонализуемой), если ее жордановой формой является диагональная матрица.

- Матрица простой структуры имеет ровно  $n$  линейно независимых собственных векторов. Про такую матрицу еще говорят, что она имеет полный набор собственных векторов.

- Если все собственные значения матрицы  $A$  различны, то она заведомо имеет простую структуру.
- Симметричная матрица имеет простую структуру, и поэтому у нее имеется  $n$  линейно-независимых собственных векторов. Ее собственные векторы, отвечающие различным собственным значениям, ортогональны в смысле обычного скалярного произведения

$$y^T x = (x, y) = \sum_{i=1}^n x_i y_i,$$

а собственные векторы, отвечающие кратному собственному значению (собственному значению кратности  $m$  отвечает  $m$  линейно-независимых собственных векторов), могут быть ортогонализированы.

- Матрица  $A^T$  имеет те же собственные значения, что и матрица  $A$ , а собственные векторы  $\xi_i$  и  $\eta_j$  матриц  $A$  и  $A^T$ , соответственно, отвечающие различным собственным значениям, ортогональны (образуют биортогональную систему). Векторы  $\xi_i$  иногда называют правыми собственными векторами матрицы  $A$ , а  $\eta_j$  — левыми, ибо  $\eta_j^T A = \lambda_j^T$ .
- Собственные векторы матриц  $A$  и  $A^{-1}$  совпадают, а собственные значения связаны соотношением  $\lambda_i(A^{-1}) = \lambda_i^{-1}(A)$ .
- Собственные векторы матриц  $A$  и  $B = A + \alpha I$  совпадают, а собственные значения связаны соотношениями  $\lambda_i(B) = \lambda_i(A) + \alpha$ .
- Число

$$r(x) = \frac{x^T A x}{\|x\|_2^2}$$

называется *отношением Релея*. Если  $x = \xi_i$ , где  $\xi_i$  — собственный вектор матрицы  $A$ , то отношение Релея равно собственному значению, отвечающему этому собственному вектору

$$\lambda_i = \frac{\xi_i^T A \xi_i}{\xi_i^T \xi_i}.$$

Задача нахождения всех собственных значений и собственных векторов называется *полной проблемой собственных значений*. Эта проблема в общем случае довольно сложна.

Наряду с полной проблемой собственных значений существуют частичные проблемы собственных значений, отыскание решений которых много проще.

К последним относятся:

- 1) Задача отыскания максимального или минимального по модулю собственного значения и, быть может, отвечающего ему собственного вектора.
- 2) Задача отыскания двух наибольших по модулю собственных значений и соответствующих собственных векторов.
- 3) Задача отыскания собственного значения, наиболее близкого к заданному числу.

Этими задачами мы сначала и займемся.

## 8.2 Степенной метод

Степенной метод хорошо подходит для приближенного вычисления экстремальных собственных значений матрицы, т.е. собственных значений, имеющих наибольший или наименьший модуль, равно как и отвечающих им собственных векторов.

### 8.2.1 Нахождение максимального по модулю собственного значения

Пусть матрица  $A \in \mathbb{R}^{n \times n}$  является диагонализуемой. Это, в частности, означает, что матрица  $A$  имеет  $n$  линейно независимых собственных векторов, совокупность которых образует базис в  $\mathbb{C}^n$ . Будем считать, что эти векторы нормированы, и будем обозначать их через  $\xi_i$ ,  $i = 1, \dots, n$ . Предположим также, что собственные значения  $\lambda_i$  матрицы  $A$ , которым отвечают собственные векторы  $\xi_i$ , расположены в порядке убывания их модулей, причем максимальный модуль имеет только одно собственное значение  $\lambda_1$ , т.е.

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \dots \geq |\lambda_n|. \quad (8.2)$$

В этом случае  $\lambda_1$  называется *доминирующим собственным значением*.

**Замечание 8.1.** При выполнении условий (8.2) собственное значение  $\lambda_1$  является действительным.

*Степенным методом* называется следующий итерационный процесс: начиная с произвольного начального вектора единичной евклидовой длины  $\xi^{(0)} \in \mathbb{R}^n$ , строятся последовательности

$$\begin{aligned} x^{(k)} &= A\xi^{(k-1)}, \\ \xi^{(k)} &= x^{(k)} / \|x^{(k)}\|_2, \\ \lambda^{(k)} &= [\xi^{(k)}]^T A\xi^{(k)}, \quad k = 1, 2, \dots \end{aligned} \quad (8.3)$$

Изучим вопрос о сходимости метода (8.3). Из первых двух соотношений (8.3) при  $k = 1$  следует, что

$$\xi^{(1)} = A\xi^{(0)} / \|A\xi^{(0)}\|_2,$$

а индукцией по  $k$  находим, что

$$\xi^{(k)} = \frac{A^k \xi^{(0)}}{\|A^k \xi^{(0)}\|_2}, \quad k = 1, 2, \dots, \quad (8.4)$$

Это соотношение объясняет роль степеней  $A$  в изучаемом методе.

Разложим вектор  $\xi^{(0)}$  по базису  $\xi_i$ ,  $i = 1, 2, \dots, n$

$$\xi^{(0)} = \sum_{i=1}^n c_i \xi_i, \quad c_i \in \mathbb{C}, \quad i = 1, 2, \dots, n \quad (8.5)$$

и предположим, что

$$c_1 \neq 0. \quad (8.6)$$

Поскольку  $A\xi_i = \lambda_i \xi_i$ , то

$$\begin{aligned} A^k \xi^{(0)} &= \sum_{i=1}^n c_i \lambda_i^k \xi_i = c_1 \lambda_1^k \left( \xi_1 + \sum_{i=2}^n \frac{c_i}{c_1} \left( \frac{\lambda_i}{\lambda_1} \right)^k \xi_i \right) = \\ &= c_1 \lambda_1^k \left( \xi_1 + \zeta^{(k)} \right), \end{aligned} \quad (8.7)$$

где

$$\zeta^{(k)} = \sum_{i=2}^n \frac{c_i}{c_1} \left( \frac{\lambda_i}{\lambda_1} \right)^k \xi_i. \quad (8.8)$$

Так как в силу (8.2)  $|\lambda_i/\lambda_1| < 1$  для  $i = 2, 3, \dots, n$ , то при возрастании  $k$  вектор  $A^k \xi^{(0)}$  (а с ним вместе и вектор  $\xi^{(k)}$  из (8.4)) будет стремиться принять направление собственного вектора  $\xi_1$ , в то время как компоненты

разложения по другим направлениям  $\xi_i$  будут убывать. Используя (8.4) и (8.7), получим

$$\xi^{(k)} = \frac{c_1 \lambda_1^k (\xi_1 + \zeta^{(k)})}{\|c_1 \lambda_1^k (\xi_1 + \zeta^{(k)})\|_2} = \pm \frac{\xi_1 + \zeta^{(k)}}{\|\xi_1 + \zeta^{(k)}\|_2}. \quad (8.9)$$

Имеет место

**Теорема 8.4.** Пусть  $A \in \mathbb{R}^{n \times n}$  — диагонализуемая матрица, собственные значения которой подчиняются условиям (8.2). Тогда, если выполнено условие (8.6), то существует постоянная  $c > 0$  такая, что

$$\|\pm \xi^{(k)} - \xi_1\|_2 + |\lambda^{(k)} - \lambda_1| \leq c |\lambda_2/\lambda_1|^k, \quad k \geq 1, \quad (8.10)$$

где  $\xi^{(k)}$  и  $\lambda^{(k)}$  определяются соотношениями (8.3).

**Доказательство.** Из (8.9) следует, что

$$\pm \xi^{(k)} - \xi_1 = \frac{\xi_1 + \zeta^{(k)}}{\|\xi_1 + \zeta^{(k)}\|_2} - \xi_1 = \frac{1 - \|\xi_1 + \zeta^{(k)}\|_2}{\|\xi_1 + \zeta^{(k)}\|_2} \xi_1 + \frac{\zeta^{(k)}}{\|\xi_1 + \zeta^{(k)}\|_2}.$$

Поскольку в силу неравенства треугольника

$$\|\xi_1\|_2 - \|\zeta^{(k)}\|_2 \leq \|\xi_1 + \zeta^{(k)}\|_2 \leq \|\xi_1\|_2 + \|\zeta^{(k)}\|_2,$$

а  $\|\xi_1\|_2 = 1$ , то

$$|1 - \|\xi_1 + \zeta^{(k)}\|_2| \leq \|\zeta^{(k)}\|_2.$$

Вновь используя неравенство треугольника, находим, что

$$\|\pm \xi^{(k)} - \xi_1\|_2 \leq 2 \frac{\|\zeta^{(k)}\|_2}{\|\xi_1 + \zeta^{(k)}\|_2}.$$

Наконец, из (8.2), (8.6) и (8.8) следует, что

$$\|\zeta^{(k)}\|_2 \leq \sum_{i=2}^n |c_i/c_1| |\lambda_2/\lambda_1|^k,$$

а эта оценка вместе с предыдущей приводит к оценке

$$\|\pm \xi^{(k)} - \xi_1\|_2 \leq c |\lambda_2/\lambda_1|^k. \quad (8.11)$$

Далее, в силу (8.11)

$$\xi^{(k)} = \pm \xi_1 + O(|\lambda_2/\lambda_1|^k)$$

и, следовательно,

$$\lambda^{(k)} = \xi^{(k)} A \xi^{(k)} = \lambda_1 + O(|\lambda_2/\lambda_1|^k).$$

Теорема доказана.

**Теорема 8.5.** Пусть  $A$  — действительная симметричная матрица, собственные значения которой подчинены условиям (8.2). Тогда

$$|\lambda_1 - \lambda^{(k)}| \leq (|\lambda_1| + |\lambda_n|) \operatorname{tg}^2 \theta |\lambda_2/\lambda_1|^{2k},$$

где  $\lambda^{(k)}$  определяется соотношениями (8.3), а  $\theta$  — угол между векторами  $\xi^{(0)}$  и  $\xi_1$ .

**Доказательство.** В силу симметрии матрицы  $A$  ее собственные векторы  $\xi_i$ ,  $i = 1, 2, \dots, n$  ортогональны. Поэтому вектор  $\xi_1$  ортогонален вектору  $\zeta^{(k)}$  из (8.8) равно как и вектору  $A\zeta^{(k)}$ . В силу сказанного и с учетом (8.9), соотношение (8.3) принимает вид

$$\lambda^{(k)} = \frac{[\xi_1 + \zeta^{(k)}]^T A [\xi_1 + \zeta^{(k)}]}{[\xi_1 + \zeta^{(k)}]^T [\xi_1 + \zeta^{(k)}]} = \frac{\lambda_1 + [\zeta^{(k)}]^T A [\zeta^{(k)}]}{1 + [\zeta^{(k)}]^T [\zeta^{(k)}]},$$

и, следовательно,

$$\begin{aligned} |\lambda_1 - \lambda^{(k)}| &= \frac{|[\zeta^{(k)}]^T A [\zeta^{(k)}] - \lambda_1 [\zeta^{(k)}]^T [\zeta^{(k)}]|}{1 + \|\zeta^{(k)}\|_2^2} = \\ &= \left| \frac{\lambda_2}{\lambda_1} \right|^{2k} \frac{\sum_{i=2}^n (c_i/c_1)^2 (\lambda_i/\lambda_1)^{2k} |\lambda_i - \lambda_1|}{1 + \|\zeta^{(k)}\|_2^2} \leq \\ &\leq (|\lambda_1| + |\lambda_n|) \left( \frac{\lambda_2}{\lambda_1} \right)^{2k} \frac{\sum_{i=2}^n c_i^2}{c_1^2} = (|\lambda_1| + |\lambda_n|) \left( \frac{\lambda_2}{\lambda_1} \right)^{2k} \frac{1 - c_1^2}{c_1^2}, \end{aligned}$$

ибо вектор  $\xi^{(0)}$ , коэффициенты разложения по базису  $\xi_i$ ,  $i = 1, 2, \dots, n$ , которого являются числа  $c_i$ ,  $i = 1, 2, \dots, n$ , нормирован. Но

$$c_1 = \xi_1^T \xi^{(0)} = \cos \theta,$$

что и завершает доказательство теоремы.

**Замечание 8.2.** Если условие  $c_1 = 0$  не выполнено (априори проверить это условие нельзя), то это еще не значит, что итерационный процесс (8.3) с начальным приближением (8.5) не приведет к результату. При достаточно большом числе итераций за счет ошибок округления может появиться ненулевая компонента  $c_1$ , и итерационный процесс выйдет в конце концов на нужное решение. Но при этом нужно иметь в виду, что если  $|\lambda_3| \ll |\lambda_2|$ , то итерации очень быстро выйдут на второе собственное значение и второй собственный вектор, и можно обмануться, приняв их за искомые величины. Это не так вероятно, если  $|\lambda_2|$  и  $|\lambda_3|$  не слишком сильно различаются, а требуемая точность достаточно велика. Итерации

в этом случае будут сходиться достаточно медленно, и их потребуется много для получения требуемой точности. За это время погрешности округления накопятся, и может сформироваться новая точка притяжения итерационного процесса —  $(\lambda_1, \xi_1)$ . Если нет уверенности в правильности найденного собственного значения, следует провести еще один или несколько расчетов с другими начальными приближениями.

**Замечание 8.3.** 1) Подтверждением того, что  $\lambda_1$  не является кратным собственным значением, и что нет собственного значения  $(-\lambda_1)$ , служит сходимость итерационного процесса к одному и тому же собственному вектору при различных начальных приближениях.

2) Если при различных начальных векторах  $\xi^{(0)}$  значения  $\lambda^{(k)}$  сходятся к одному и тому же числу, а последовательности векторов  $\xi^{(k)}$  приводят к неколлинеарным векторам, то это обстоятельство служит подтверждением того, что максимальное по модулю собственное значение является кратным. Если требуется найти собственное подпространство, или нужно определить кратность найденного собственного значения, нужно проводить вычисления с различными начальными приближениями до тех пор, пока перестанут получаться векторы, линейно-независимые с уже найденными.

### 8.2.2 Пример

Применим степенной метод для отыскания максимального по модулю собственного значения матрицы

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

В качестве начального приближения возьмем вектор  $x^{(0)} = [0, 1]^T$ . Тогда

$$\begin{aligned} x^{(1)} &= \begin{bmatrix} 2 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, & x^{(2)} &= \begin{bmatrix} 2 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}, \\ x^{(3)} &= \begin{bmatrix} 2 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} 3 \\ -2 \end{bmatrix}, \dots \end{aligned}$$

Очевидно, что  $x^{(k)} = [k(-k+1)]^T$ ,  $x^{(k+1)} = [(k+1)(-k)]^T$ . Тогда  $(x^{(k)}, x^{(k)}) = 2k^2 - 2k + 1$ ,  $(x^{(k)}, x^{(k+1)}) = 2k^2$ , и, следовательно,

$$\lambda^{(k+1)} = \frac{2k^2}{2k^2(1 - 1/k + O(k^{-2}))} = 1 + \frac{1}{k} + O\left(\frac{1}{k^2}\right).$$

Сходимость к собственному значению  $\lambda = 1$  очень медленна и не похожа на ту, которую мы имели в теореме 8.4.

Выясним, с чем это связано. Решая задачу на собственные значения, находим, что рассматриваемая матрица имеет двукратное собственное значение  $\lambda = 1$ , которому отвечает единственный собственный вектор  $[1, -1]^T$ . Тем самым, эта матрица не является матрицей простой структуры, как это было в теореме 8.4, а ее жордановой формой является клетка порядка два. Приведенный пример показывает, что степенной метод не отказывается работать и в том случае, когда максимальному по модулю кратному собственному значению отвечает жорданова клетка, но скорость сходимости резко падает от скорости сходимости геометрической прогрессии к скорости сходимости гармонического ряда.

### 8.2.3 Нахождение второго по величине модуля собственного значения

Пусть

$$|\lambda_1| > |\lambda_2| > |\lambda_3| \geq \dots \geq |\lambda_n|.$$

Будем считать, что  $\lambda_1$ ,  $\xi_1$  и  $\eta_1$  ( $A^T \eta_1 = \lambda_1 \eta_1$ ) известны, причем  $\|\xi_1\| = 1$ ,  $(\eta_1, \xi_1) = 1$ . Найти  $\lambda_1$ ,  $\xi_1$  и  $\eta_1$  можно описанным выше способом. Пусть  $x^{(0)}$  — произвольный вектор, такой, что  $(x^{(0)}, \eta_1) \neq 0$ . Тогда

$$x^{(0)} = c_1 \xi_1 + c_2 \xi_2 + \dots + c_n \xi_n, \quad c_1 = (x^{(0)}, \eta_1), \quad c_2 \neq 0.$$

Построим вектор

$$y^{(0)} = x^{(0)} - (x^{(0)}, \eta_1) \xi_1 = c_2 \xi_2 + c_3 \xi_3 + \dots + c_n \xi_n$$

и вектор

$$\xi_2^{(0)} = y^{(0)} / \|y^{(0)}\|_2.$$

Итерационный процесс будем осуществлять по формулам

$$\begin{aligned} x^{(k)} &= A \xi^{(k-1)}, \\ y^{(k)} &= x^{(k)} - (x^{(k)}, \eta_1) \xi_1, \\ \xi_2^{(k)} &= y^{(k)} / \|y^{(k)}\|_2, \\ \lambda_2^{(k)} &= \xi_2^{(k)T} A \xi_2^{(k)}. \end{aligned}$$

Тогда

$$\begin{aligned} \lambda_2^{(k)} &= \lambda_2 + O(|\lambda_3/\lambda_2|^k), \\ \xi_2^{(k)} &= \pm \xi_2 + O(|\lambda_3/\lambda_2|^k). \end{aligned}$$



### 8.3 Обратные итерации

В этом разделе мы будем искать приближения к собственному значению матрицы  $A \in \mathbb{R}^{n \times n}$ , которое является ближайшим к заданному числу  $\mu \in \mathbb{R}$ , где  $\mu \notin \sigma(A)$ . Для этого степенной метод (8.3) следует применить к матрице  $M_\mu^{-1} := (A - \mu I)^{-1}$ , определив так называемые *обратные итерации*. Число  $\mu$  здесь называется *сдвигом*.

Собственные значения матрицы  $M_\mu^{-1}$  суть  $\mu_i = (\lambda_i - \mu)^{-1}$ . Предположим, что существует такое целое число  $m$ , что

$$|\lambda_m - \mu| < |\lambda_i - \mu| \quad \forall i = 1, 2, \dots, n, \quad i \neq m. \quad (8.12)$$

Это означает, что собственное значение  $\lambda_m$ , ближайшее к  $\mu$ , является единственным собственным значением с таким свойством. Более того, (8.12) показывает, что  $\mu_m$  есть собственное значение  $M_\mu^{-1}$  с наибольшим модулем; в частности, если  $\mu = 0$ , то  $\lambda_m$  превращается в собственное значение  $A$  с минимальным модулем.

По заданному произвольному начальному вектору  $\xi^{(0)} \in \mathbb{R}^n$  единичной евклидовой длины построим последовательности

$$\begin{aligned} (A - \mu I)y^{(k)} &= \xi^{(k-1)}, \\ \xi^{(k)} &= y^{(k)} / \|y^{(k)}\|_2, \\ \lambda^{(k)} &= \left[ \xi^{(k)} \right]^T A \left[ \xi^{(k)} \right]. \end{aligned} \quad (8.13)$$

Заметим, что собственные векторы матрицы  $M_\mu$  те же самые, что и у матрицы  $A$ . Поэтому отношение Релея вычисляется по матрице  $A$  (а не по матрице  $M_\mu^{-1}$ ). Основное отличие от метода (8.3) состоит в том, что на каждом шаге  $k$  нужно решать линейную систему с матрицей  $M_\mu$ . Вычислительные затраты состоят из однократной  $LU$  факторизации матрицы  $M_\mu$  стоимостью  $O(m^3)$  действий при  $k = 1$  и решения на каждом шаге двух систем с треугольными матрицами с затратами  $O(n^2)$  действий.

Хотя этот метод более дорогой, чем степенной метод (8.3), обратные итерации позволяют найти приближение к любому собственному значению матрицы  $A$  (ближайшему к сдвигу  $\mu$ ).

### 8.4 Итерации с отношением Рэля

Будем предполагать, что матрица  $A$  является не только действительной, но и симметричной. Это предположение обеспечивает действительность

собственных значений и ортогональность собственных векторов, которые мы будем предполагать нормированными.

Обратимся к отношению Рэля

$$r(x) = \frac{x^T Ax}{x^T x}$$

и покажем, что собственные векторы  $\xi_j$  матрицы  $A$  являются стационарными точками функции  $r(x)$ , т.е.  $\text{grad } r(x) \big|_{x=\xi_j} = 0$ . Дифференцируя  $r(x)$  по  $x_j$ , находим, что

$$\begin{aligned} \frac{\partial r(x)}{\partial x_j} &= \frac{\partial / \partial x_j (x^T Ax)}{x^T x} - \frac{(x^T Ax) \partial / \partial x_j (x^T x)}{(x^T x)^2} = \\ &= 2 \frac{(Ax)_j}{x^T x} - \frac{(x^T Ax) 2x_j}{(x^T x)^2} = \frac{2}{x^T x} \left( Ax - \frac{x^T Ax}{x^T x} x \right)_j = \\ &= \frac{2}{x^T x} (Ax - r(x)x)_j. \end{aligned}$$

Собирая эти производные в  $n$ -мерный вектор, получим градиент  $r(x)$ , который будем обозначать как  $\nabla r(x)$ . Имеем

$$\nabla r(x) = \frac{2}{x^T x} (Ax - r(x)x).$$

Как было уже замечено,  $r(\xi_j) = \lambda_j$  и, следовательно,

$$\nabla r(\xi_j) = 0.$$

Пусть вектор  $x$  является приближением к собственному вектору  $\xi_j$ . Оценим близость  $r(x)$  к собственному значению  $\lambda_j$ . Поскольку  $r(x)$  является гладкой функцией  $x$  всюду, кроме  $x = 0$ , то, раскладывая  $r(x)$  в точке  $\xi_j$  по формуле Тейлора, найдем, что

$$r(x) = r(\xi_j) + (\nabla r(\xi_j))^T (x - \xi_j) + O(\|x - \xi_j\|^2).$$

Поэтому

$$|r(x) - \lambda_j| = O(\|x - \xi_j\|^2).$$

Этот факт без доказательства уже был отмечен в формулировке теоремы 8.5.

Алгоритм итераций с отношением Рэля базируется на обратных итерациях, где для улучшения сходимости на каждой итерации используется

новый сдвиг — найденное на предыдущем шаге приближение к собственному значению.

Начиная с произвольного начального вектора  $\xi^{(0)}$  единичной длины находится число

$$\lambda^{(0)} = \left(v^{(0)}\right)^T A v^{(0)},$$

и для  $k = 1, 2, \dots$  решается система

$$(A - \lambda^{(k-1)}I)w = v^{(k-1)},$$

решение которой нормируется

$$\xi^{(k)} = w/\|w\|$$

и подставляется в отношение Рэля

$$\lambda^{(k)} = (\xi^{(k)})^T A \xi^{(k)}.$$

**Теорема 8.6.** *Итерации с отношением Рэля сходятся к собственной паре  $(\lambda_j, \xi_j)$  для почти всех начальных векторов  $\xi^{(0)}$ . В малой окрестности  $(\lambda_j, \xi_j)$  эта сходимость кубичная, т.е. при  $k \rightarrow \infty$*

$$\|\xi^{(k+1)} - (\pm)\xi_j\| = O\left(\|\xi^{(k)} - (\pm)\xi_j\|^3\right)$$

и

$$\left|\lambda^{(k+1)} - \lambda_j\right| = O\left(\left|\lambda^{(k)} - \lambda_j\right|^3\right).$$

**Пример 8.1.** Рассмотрим симметричную матрицу

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 4 \end{bmatrix}$$

и пусть  $v^{(0)} = [1 \ 1 \ 1]^T/\sqrt{3}$  — начальный вектор. Три последовательные итерации с отношением Рэля дает следующие приближения к собственному значению

$$\lambda^{(0)} = 5, \quad \lambda^{(1)} = 5.2131\dots, \quad \lambda^{(2)} = 5.214319743184.$$

Действительная величина собственного значения, отвечающего собственному вектору, ближайшему к  $v^{(0)}$  есть  $\lambda = 5.214319743377$ . Тем самым, после всего лишь трех итераций с отношением Рэля мы получили результат с десятью верными знаками.

В рассматриваемом методе при отыскании очередного приближения к собственному вектору приходится решать систему все более близкую к вырожденной, т.е. плохо обусловленной. Как следует из предыдущего анализа, решение систем с плохо обусловленными матрицами, вообще говоря, не может быть слишком успешным. Выясним, к каким негативным последствиям это может привести в рассматриваемом нами случае.

Пусть  $\mu$  — число очень близкое к одному из изолированных собственных значений матрицы  $A$ . Обозначим это число через  $\lambda_j$  и рассмотрим систему

$$(A - \mu I)x = b, \quad (8.14)$$

где  $b$  — некоторый вектор. Разложим векторы  $b$  и  $x$  по собственным векторам матрицы  $A$

$$b = \sum_{j=1}^n d_j \xi_j, \quad x = \sum_{j=1}^n c_j \xi_j.$$

Подставляя эти разложения в (8.14), найдем, что

$$c_j = d_j / (\lambda_j - \mu),$$

и, следовательно,

$$x = \sum_{j=1}^n \frac{d_j}{\lambda_j - \mu} \xi_j = \frac{d_J}{\lambda_J - \mu} \xi_J + \sum_{j \neq J} \frac{d_j}{\lambda_j - \mu} \xi_j.$$

Поскольку  $\sum_{j \neq J} |d_j / (\lambda_j - \mu)| = O(1)$ , то ортогональная к  $\xi_J$  составляющая правой части  $b$  практически не оказывает влияние на решение  $x = O((\lambda_J - \mu)^{-1})$ , которое сосредоточено около вектора  $\xi_J$ . То же самое происходит с погрешностью, если система решается устойчивым методом: почти вся погрешность (большая) сосредоточена вдоль вектора  $\xi_J$ . Поскольку нас интересует не сам вектор  $x$ , а лишь направление, которое он задает, то, принимая во внимание, что  $\|x\| = |d_J / (\lambda_J - \mu)| + O(1)$ , находим, что

$$x / \|x\| = \pm \xi_J + O(\lambda_J - \mu)$$

даже для произвольного вектора  $b$ .

## 9

# $QR$ - алгоритм

В этом разделе мы рассмотрим некоторую технику нахождения *всех* собственных значений матрицы  $A$ . Основная идея, которая здесь используется, состоит в сведении матрицы  $A$  при помощи подходящих преобразований к форме, для которой вычисление собственных значений проще, чем для исходной матрицы.

Метод, о котором пойдет речь, есть  $QR$ -алгоритм. Здесь он будет рассмотрен только для случая действительных матриц.

Для заданной ортогональной матрицы  $Q_0$ , начиная с  $A_0 = Q_0^T A Q_0$ , где  $A \in \mathbb{R}^{n \times n}$ , построить последовательность

$$A_k = R_k Q_k, \quad k = 1, 2, \dots, \quad (9.1)$$

где  $R_k$  и  $Q_k$  — матрицы из  $QR$  разложения матрицы

$$A_{k-1} = Q_k R_k. \quad (9.2)$$

Для каждого шага  $k$  первая стадия итерации состоит в факторизации матрицы  $A_{k-1}$ , т.е. в представлении ее в виде произведения ортогональной матрицы  $Q_k$  и верхней треугольной матрицы  $R_k$ . Вторая стадия — простое перемножение матриц.

В силу того, что

$$\begin{aligned} A_k &= R_k Q_k = (Q_k^T Q_k) R_k Q_k = Q_k^T [Q_k R_k] Q_k = Q_k^T A_{k-1} Q_k = \\ &= [Q_0 Q_1 \dots Q_k]^T A [Q_0 Q_1 \dots Q_k], \end{aligned}$$

каждая матрица  $A_k$  ортогонально подобна  $A$ .

В базовой версии  $QR$ -алгоритма полагают  $Q_0 = I$ , что приводит к  $A_0 = A$ . При некоторых ограничениях на  $A$  матрицы  $A_k$  по форме сходятся к верхней треугольной матрице, на главной диагонали которой

стоят собственные числа. Понимать это нужно так, что поддиагональные элементы стремятся к нулю, диагональные к собственным числам, а наддиагональные могут куда не стремиться.

Следует отметить, что описанный алгоритм очень трудоемкий, ибо на каждой итерации нужно осуществлять  $QR$  факторизацию матрицы. Поэтому, прежде чем более подробно обсуждать вопрос о сходимости  $QR$ -алгоритма, модифицируем его базовую версию.

Сначала напомним введенное нами ранее определение матрицы Хессенберга.

**Определение 9.1.** Матрица  $A$  имеет верхнюю форму Хессенберга, если  $a_{ij} = 0$  для любых  $i > j + 1$ .

Это означает, что матрица имеет почти треугольную форму.

$$\begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ & * & * & * & * \\ & & * & * & * \\ & & & * & * \end{bmatrix}$$

**Теорема 9.1.** *Всякая действительная квадратная матрица при помощи ортогонального преобразования подобия может быть приведена к форме Хессенберга.*

**Доказательство.** Построим требуемое преобразование. Этот алгоритм очень похож на алгоритм  $QR$ -разложения матрицы при помощи отражений. Приведение осуществляется за  $n - 2$  шага. На первом шаге необходимыми нулями заполнится первый столбец, на втором шаге — второй и т.д.

Выполним первый шаг. Запишем матрицу  $A$  в виде

$$A = \begin{bmatrix} a_{11} & c^T \\ b & \hat{A} \end{bmatrix}.$$

Пусть  $\hat{U}_1$  — матрица отражения, переводящая  $(n - 1)$ -мерный вектор  $b$  в вектор  $[t_1 \ 0 \ \dots \ 0]^T$ , и пусть

$$U_1 = \begin{bmatrix} 1 & 0 \\ 0 & \hat{U}_1 \end{bmatrix}. \quad (9.3)$$

Тогда

$$U_1 A = \begin{bmatrix} 1 & 0 \\ 0 & \widehat{U}_1 \end{bmatrix} \begin{bmatrix} a_{11} & c^T \\ b & \widehat{A} \end{bmatrix} = \begin{bmatrix} a_{11} & c^T \\ \widehat{U}_1 b & \widehat{U}_1 \widehat{A} \end{bmatrix} = \left[ \begin{array}{c|c} a_{11} & c^T \\ \hline t_1 & \\ 0 & \\ \vdots & \widehat{U}_1 \widehat{A} \\ 0 & \end{array} \right].$$

Далее, при вычислении  $U_1 A U_1^{-1}$  вспомним, что  $U_1^{-1} = U_1^T = U_1$ , и, следовательно,

$$\begin{aligned} A^{(1)} &= U_1 A U_1 = \begin{bmatrix} a_{11} & c^T \\ \widehat{U}_1 b & \widehat{U}_1 \widehat{A} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \widehat{U}_1 \end{bmatrix} = \begin{bmatrix} a_{11} & c^T \widehat{U}_1^T \\ \widehat{U}_1 b & \widehat{U}_1 \widehat{A} \widehat{U}_1 \end{bmatrix} = \\ &= \left[ \begin{array}{c|c} a_{11} & c^T \widehat{U}_1 \\ \hline t_1 & \\ 0 & \\ \vdots & \widehat{U}_1 \widehat{A} \widehat{U}_1 \\ 0 & \end{array} \right] = \left[ \begin{array}{c|ccc} a_{11} & * & \dots & * \\ \hline t_1 & & & \\ 0 & & & \\ \vdots & & \widehat{A}_1 & \\ 0 & & & \end{array} \right]. \end{aligned} \quad (9.4)$$

Матрица  $A^{(1)}$  ортогонально подобна матрице  $A$  и имеет первый столбец такого же вида, как у матрицы Хессенберга ( $a_{ii}^{(i)} = 0$  при  $i = 3, 4, \dots, n$ ). Отметим, что домножение матрицы  $U_1 A$  на  $U_1$  справа не меняет портрет матрицы  $U_1 A$ , приобретенный  $A$  после умножения ее на  $U_1$  слева.

Теперь построим матрицу отражения  $\widehat{U}_2 \in \mathbb{R}^{(n-2) \times (n-2)}$ , переводящую вектор  $[a_{32}^{(1)} a_{42}^{(1)} \dots a_{n2}^{(1)}]^T$  в вектор, коллинеарный вектору  $[10 \dots 0]^T \in \mathbb{R}^{n-2}$ , и введем матрицу

$$U_2 = \begin{bmatrix} I_2 & 0 \\ 0 & \widehat{U}_2 \end{bmatrix}.$$

Выполняя преобразование подобия  $U_2 A^{(1)} U_2 = A^{(2)}$ , убеждаемся в том, что портрет первых двух столбцов  $A^{(2)}$  совпадает с портретом тех же столбцов матрицы Хессенберга. Продолжая аналогичные преобразования, на  $(n-2)$  шаге придем к матрице  $A^{(n-2)}$ , имеющей форму Хессенберга. Теорема доказана.

**Замечание 9.1.** Если бы мы попытались взять такую матрицу отражения  $U_1$ , которая оставляет в первом столбце только один ненулевой элемент, то при умножении  $U_1 A$  на  $U_1$  справа нам не удалось бы сохранить структуру первого столбца, т.е. полученные после первого умножения  $A$

на  $U_1$  слева нули. Именно благодаря форме (9.3) матрицы  $U_1$  операция умножения на  $U_1$  справа не затрагивает нули в первом столбце (9.4).

**Замечание 9.2.** Для симметричной матрицы  $A$  ортогонально подобная ей матрица тоже симметрична и, следовательно, в этом случае матрица Хессенберга будет тридиагональной.

**Теорема 9.2.** Пусть  $A_m$  — невырожденная верхняя хессенбергова матрица и  $A_{m+1}$  получена из  $A_m$  посредством одной  $QR$ -итерации (9.1)–(9.2). Тогда  $A_{m+1}$  также имеет верхнюю форму Хессенберга.

**Доказательство.** Для построения  $A_{m+1}$  нам нужно сначала построить разложение  $A_m = Q_m R_m$ , которое можно переписать в виде  $Q_m = A_m R_m^{-1}$ . Ранее было показано (упражнение 1.1), что обратная к невырожденной верхней треугольной матрице есть верхняя треугольная матрица, и произведение верхних треугольных матриц тоже верхняя треугольная. Поэтому  $R_m^{-1}$  является верхней треугольной. Аналогично доказывается, что произведение верхней треугольной и верхней хессенберговой матрицы в любом порядке есть верхняя хессенбергова матрица. Поэтому  $Q_m$  есть верхняя хессенбергова матрица, и  $A_{m+1} = R_m Q_m$  тоже верхняя хессенбергова. Теорема доказана.

Модифицируем базовую версию  $QR$ -алгоритма, положив начальную матрицу  $A_0$  равной матрице Хессенберга  $A^{(n-2)}$  из теоремы 9.1. Будем эту матрицу обозначать через  $H$ . При таком выборе  $A_0$  в силу теоремы 9.2 на каждом шаге  $QR$ -алгоритма будем снова получать матрицы Хессенберга.

Какой выигрыш в трудозатратах при реализации  $QR$ -алгоритма дает эта модификация? Приведение матрицы  $A$  к хессенберговой форме снова требует  $O(n^3)$  итераций. Но теперь это единичная акция, осуществляемая перед началом  $QR$ -итераций. Сами же  $QR$ -итерации становятся дешевле. Чтобы увидеть это, опишем процесс  $QR$ -факторизации хессенберговой матрицы.

Пусть  $H$  — верхняя хессенбергова матрица,  $QR$ -факторизацию которой нужно найти. К верхней треугольной матрице  $R$  эту матрицу можно привести при помощи  $n-1$  плоских вращений, которые преобразуют  $n-1$  поддиагональных элементов в нули. Обозначим их через  $T_1, T_2, \dots, T_{n-1}$ . Тогда, если

$$T = T_{n-1} T_{n-2} \dots T_1 =: Q^T,$$

то

$$R = TH = Q^T H \quad \text{или} \quad H = QR.$$



QR-факторизация матрицы  $H$  построена. Очевидно, что реализация одного вращения  $T_i$  требует  $O(n)$  действий, и, следовательно, QR-факторизация осуществляется за  $O(n^2)$  действий.

Теперь нужно осуществить второй этап QR-шага — вычислить  $H_1 = RQ$ . Поскольку  $H_1^T = Q^T R^T = T R^T$ , то и эта процедура, приводящая к хессенберговой матрице, осуществляется за  $O(n^2)$  действий.

Мы показали, что один шаг QR-алгоритма требует  $O(n^2)$  действий.

Для того, чтобы выяснить причины, по которым QR-алгоритм действительно может сходиться туда, куда нужно, рассмотрим другой метод, называемый *методом одновременных итераций* или *методом итерирования подпространства*, и исследуем его сходимость. Затем установим связь между этим методом и QR-алгоритмом.

Суть метода одновременных итераций состоит в том, чтобы применить степенной метод не к одному вектору  $\xi^{(0)}$ , а к совокупности нескольких векторов  $\xi^{(0)}, \eta^{(0)}, \dots, \zeta^{(0)}$ , которые мы объединим в матрицу

$$X_0 = [\xi^{(0)} \eta^{(0)} \dots \zeta^{(0)}]$$

размерности  $n \times p$ ,  $p > 1$ . Линейная оболочка  $\text{span} \{\xi^{(0)}, \eta^{(0)}, \dots, \zeta^{(0)}\}$  образует линейное подпространство. После одной итерации

$$AX_0 = X_1 = [A\xi^{(0)} A\eta^{(0)} \dots A\zeta^{(0)}] = [\xi^{(1)} \eta^{(1)} \dots \zeta^{(1)}]$$

начальное подпространство переходит в другое подпространство  $\text{span} \{\xi^{(1)}, \eta^{(1)}, \dots, \zeta^{(1)}\}$ . Если исходные векторы были линейно независимые, то размерность  $\text{span} \{\xi^{(0)}, \eta^{(0)}, \dots, \zeta^{(0)}\}$  есть  $p$ . Размерность подпространства  $\text{span} \{\xi^{(1)}, \eta^{(1)}, \dots, \zeta^{(1)}\}$  будет тоже  $p$ , по крайней мере, если матрица  $A$  невырождена, что мы и будем предполагать. Тогда и при последующих итерациях мы будем получать подпространства размерности  $p$ . Однако, из анализа степенного метода мы знаем, что итерации  $A^k x$  любого вектора  $x$  (при некоторых ограничениях) стараются принять направление собственного вектора  $\xi_1$ , отвечающего доминирующему собственному значению. Поэтому и  $A^k \xi^{(0)}, A^k \eta^{(0)}, \dots, A^k \zeta^{(0)}$  стремятся принять это направление. Тем не менее, указанные векторы остаются линейно независимыми, и, следовательно, могут рассматриваться как элементы базиса в  $\text{span} \{A^k \xi^{(0)}, A^k \eta^{(0)}, \dots, A^k \zeta^{(0)}\}$ . Но базис этот будет плохой, ибо углы между его элементами стремятся к нулю при  $k \rightarrow \infty$ . Более того, длины элементов этого базиса стремятся либо к нулю, либо к бесконечности в зависимости от того,  $|\lambda_1| < 1$  или  $|\lambda_1| > 1$ . В степенном

методе после каждой итерации мы вектор нормировали, и этого было достаточно. В рассматриваемом же случае одного нормирования мало: векторы нужно еще ортогонализировать. Поэтому в методе одновременных итераций (почему он есть и метод итерирования подпространств, ясно из вышесказанного) начальные векторы берут ортонормированными (отсюда третье название метода — *метод ортогональных итераций*). Матрицу, составленную из этих векторов, обозначим через  $Z_0$ . Ее размер есть  $n \times p$ . Сами же итерации задаются следующими правилами: на каждой итерации  $k$  вычисляется  $n \times p$  матрица

$$Y_{k+1} = AZ_k, \quad k = 0, 1, \dots, \quad (9.5)$$

которая затем подвергается  $QR$ -факторизации

$$Y_{k+1} = Z_{k+1}R_{k+1} \quad (9.6)$$

для получения  $Z_{k+1}$ .

**Некоторое отступление.** В гл. 2 мы доказали существование  $QR$ -факторизации для любой квадратной матрицы путем построения конкретных разложений при помощи вращений и отражений. В нашем же случае матрица  $Y_{k+1}$  прямоугольная размера  $n \times p$ . Чтобы свести задачу к предыдущей, дополним  $Y_{k+1}$  нулевой матрицей до квадратной размера  $n \times n$ . Опуская индексы  $k + 1$ , получим соотношение

$$Y = ZR,$$

где все матрицы — квадратные, причем  $R$  — верхняя треугольная, а  $Z$  — ортогональная. Если обозначить через  $y_i$  и  $z_i$  столбцы матриц  $Y$  и  $Z$  и обратить матрицу  $Z$ , то предыдущее соотношение примет вид

$$R = \begin{bmatrix} z_1^T \\ z_2^T \\ \vdots \\ z_n^T \end{bmatrix} [y_1 y_2 \dots y_p 0 \dots 0] = \begin{bmatrix} z_1^T y_1 & z_1^T y_2 & \dots & z_1^T y_p & 0 & \dots & 0 \\ & z_2^T y_2 & \dots & z_2^T y_p & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ & & & z_p^T y_p & 0 & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & \dots & \dots & 0 \end{bmatrix}$$

Отсюда следует, что в соотношении (9.6) матрица  $Z_{k+1}$  есть прямоугольная матрица размера  $n \times p$ , столбцами которой являются первые  $p$  столбцов ортогональной матрицы из  $QR$ -разложения матрицы  $Y$ , а  $R$  — верхняя треугольная матрица размера  $p \times p$ .

Поскольку матрица  $Z_0$  имеет ранг  $p$ , а матрица  $A$  невырождена, то и матрица  $Y_1$  имеет тот же ранг, равно как и матрица  $Z_1$ . Поэтому ранг  $p$  имеют и матрицы  $Z_{k+1}$  и  $Y_{k+1}$ . Следовательно, матрица  $R_{k+1}$ , размер которой есть  $p \times p$ , невырождена, и соотношение

$$Z_{k+1} = Y_{k+1}R_{k+1}^{-1}$$

представляет собой не что иное как процедуру ортогонализации с нормированием столбцов матрицы  $Y_{k+1}$ . При этом первый столбец  $Y_{k+1}$  только нормируется, второй ортогонализируется к первому и нормируется и т.д. (ср. с процедурой ортогонализации Грама-Шмидта).

Проведем анализ этого метода. Предположим, что  $|\lambda_p| > |\lambda_{p+1}|$ . Если  $p = 1$ , то и метод, и его анализ такие же, как в случае степенного метода. (Первый столбец  $Y_{k+1}$  только нормируется). При  $p > 1$  из (9.6), (9.5) имеем

$$\text{span} \{Z_{k+1}\} = \text{span} \{Y_{k+1}\} = \text{span} \{AZ_k\}.$$

Поэтому

$$\text{span} \{Z_k\} = \text{span} \{A^k Z_0\}.$$

Но  $A$  — диагонализируемая матрица, и, следовательно, существует матрица подобия  $S$  такая, что

$$A = S\Lambda S^{-1},$$

где  $\Lambda$  — диагональная матрица из собственных значений матрицы  $A$ . Тогда

$$A^k = S\Lambda^k S^{-1}$$

и

$$A^k Z_0 = S\Lambda^k S^{-1} Z_0.$$

Пусть  $S^{-1}Z_0 = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix}$ , где  $C_1$  — квадратная матрица порядка  $p$ , а  $\Lambda =$

$\begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix}$  с диагональными матрицами  $\Lambda_1$  и  $\Lambda_2$  порядков  $p$  и  $n - p$ , соответственно. Пусть, кроме того,  $S = [S_1 S_2]$ , где  $S_1$  и  $S_2$  — прямоугольные матрицы порядков  $n \times p$  и  $n \times (n - p)$ , соответственно. Тогда

$$\begin{aligned} S\Lambda^k S^{-1} Z_0 &= [S_1 S_2] \begin{bmatrix} \Lambda_1^k & \\ & \Lambda_2^k \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = [S_1 S_2] \begin{bmatrix} (\Lambda_1^k C_1) \\ (\Lambda_2^k C_2) \end{bmatrix} = \\ &= [S_1 \Lambda_1^k C_1 + S_2 \Lambda_2^k C_2] = \lambda_p^k [S_1 (\lambda_p^{-1} \Lambda_1)^k C_1 + S_2 (\lambda_p^{-1} \Lambda_2)^k C_2]. \end{aligned}$$

Поскольку

$$\lambda_p^{-1}\Lambda_2 = \begin{bmatrix} \lambda_{p+1}/\lambda_p & & \\ & \dots & \\ & & \lambda_n/\lambda_p \end{bmatrix},$$

а  $|\lambda_{p+1}|/|\lambda_p| < 1$ , то  $(\lambda_p^{-1}\Lambda_2)^k$  при  $k \rightarrow \infty$  стремится к нулевой матрице, а вместе с ней к нулевой матрице стремится и матрица  $S_2(\lambda_p^{-1}\Lambda_2)^k C_2$ . Поэтому

$$\begin{aligned} \text{span} \{Z_k\} &= \text{span} \{S\Lambda^k S^{-1}Z_0\} = \\ &= \text{span} \{S_1(\lambda_p^{-1}\Lambda_1)^k C_1 + S_2(\lambda_p^{-1}\Lambda_2)^k C_2\} \implies \\ &\implies \text{span} \{S_1(\lambda_p^{-1}\Lambda_1)^k C_1\}, \end{aligned}$$

т.е. подпространство  $\text{span} \{Z_k\}$  сходится к  $\{S_1(\lambda_p^{-1}\Lambda_1)^k C_1\}$ . Далее, поскольку модули всех диагональных элементов диагональной матрицы  $\lambda_p^{-1}\Lambda_1$  не меньше 1, то матрица  $(\lambda_p^{-1}\Lambda_1)^k$  невырождена. Если предположить, что и матрица  $C_1$  является невырожденной, то

$$\text{span} \{S_1[(\lambda_p^{-1}\Lambda_1)^k C_1]\} = \text{span} \{S_1\},$$

и, следовательно,

$$\text{span} \{Z_k\} \implies \text{span} \{S_1\} = \text{span} \{s_1 s_2 \dots s_p\},$$

где  $s_j$   $j = 1, \dots, p$ , суть собственные векторы, отвечающие собственным значениям  $\lambda_1, \lambda_2, \dots, \lambda_p$ .

Итак, мы показали, что если  $|\lambda_p| > |\lambda_{p+1}|$  и  $\det C_1 \neq 0$ , то  $\text{span} \{Z_k\}$  сходится к инвариантному подпространству, натянутому на первые  $p$  собственных векторов матрицы  $A$ . Легко проверить, что при  $p = 1$  условие  $\det C_1 \neq 0$  переходит в условие (8.6) степенного метода.

Заметим, что, если в итерациях этого алгоритма следить лишь за первыми  $p'$  столбцами матриц  $Z_k$ , где  $p' < p$ , то они будут теми же самыми, что и при итерировании только первых  $p'$  столбцов матрицы  $Z_0$ , а не  $p$  столбцов, как в первом случае. Иными словами, итерация по существу происходит одновременно для всех значений  $p' = 1, 2, \dots, p$ . Поэтому, если модули всех собственных значений попарно различны, то из проведенного анализа сходимости вытекает, что линейная оболочка первых  $p'$  столбцов матрицы  $Z_k$  при любом  $p' \leq p$  сходится к  $\text{span} \{s_1, \dots, s_{p'}\}$ .

**Теорема 9.3.** Пусть одновременные итерации с  $p = n$  и  $Z_0 = I$  применяются к матрице  $A$ . Если все собственные значения матрицы имеют

попарно различные модули, и все угловые миноры матрицы  $S$  невырождены, то матрицы  $A_k = Z_k^T A Z_k$  сходятся к верхней форме Шура матрицы  $A$ , т.е. к верхней треугольной матрице, на главной диагонали которой расположены собственные значения. Они упорядочены по убыванию абсолютных величин.

**Доказательство.** Предположение об отличии от нуля угловых миноров матрицы  $S$  обеспечивает невырожденность матрицы  $C_1$ , требуемую предыдущим анализом. Геометрически это условие означает, что ни один вектор из инвариантного подпространства  $\text{span}\{s_1, \dots, s_j\}$  не ортогонален к подпространству  $\text{span}\{e_1, \dots, e_j\}$ , натянутому на первые  $j$  столбцов матрицы  $Z_0 = I$ . Заметим теперь, что  $Z_k$  — квадратная ортогональная матрица. Поэтому матрицы  $A$  и  $A_k = Z_k^T A Z_k$  подобны. Положим

$$Z_k = [Z_{1k} \ Z_{2k}],$$

где  $Z_{1k}$  состоит из  $p$  столбцов. Тогда

$$\begin{aligned} A_k &= Z_k^T A Z_k = \begin{bmatrix} Z_{1k}^T \\ Z_{2k}^T \end{bmatrix} A [Z_{1k} \ Z_{2k}] = \\ &= \begin{bmatrix} Z_{1k}^T A Z_{1k} & Z_{1k}^T A Z_{2k} \\ Z_{2k}^T A Z_{1k} & Z_{2k}^T A Z_{2k} \end{bmatrix}. \end{aligned}$$

Поскольку подпространство  $\text{span}\{Z_{1k}\}$  сходится к инвариантному подпространству матрицы  $A$ , то и  $\text{span}\{AZ_{1k}\}$  сходится к тому же самому подпространству, а  $Z_{2k}^T A Z_{1k}$  сходится к  $Z_{2k}^T Z_{1k} = 0$ . Учитывая, что сказанное справедливо для каждого  $p < n$ , заключаем, что все поддиагональные элементы матрицы  $A_k$  стремятся к нулю. Таким образом,  $A_k$  сходится к верхней треугольной матрице, т.е. к матрице Шура.

**Теорема 9.4.** Пусть  $A_k$  — матрицы, вычисляемые в QR-алгоритме (9.1), (9.2). Тогда

$$A_k = Z_k^T A Z_k,$$

где  $Z_k$  — матрицы, вычисляемые одновременными итерациями (9.5), (9.6), исходя из  $Z_0 = I$ . Таким образом, если все собственные значения имеют различные модули, то матрицы  $A_k$  сходятся к верхней форме Шура.

**Доказательство.** Проведем доказательство индукцией по номеру итерации  $k$ . При  $k = 0$  матрица  $A_0 = A$ , а  $Z_0 = I$ . Так что

$$A_0 = Z_0^T A Z_0.$$

Пусть

$$A_k = Z_k^T A Z_k.$$

Согласно (9.5), (9.6)

$$A Z_k = Z_{k+1} R_{k+1},$$

где  $Z_{k+1}$  и  $R_{k+1}$  — соответственно ортогональная и верхнетреугольная матрицы. Но тогда

$$A_k = Z_k^T A Z_k = Z_k^T Z_{k+1} R_{k+1}.$$

Поскольку  $Z_k$  и  $Z_{k+1}$  — ортогональные матрицы, то

$$Q_{k+1} = Z_k Z_{k+1}$$

— тоже ортогональная, и предшествующее соотношение принимает вид:

$$A_k = Q_{k+1} R_{k+1},$$

т.е. мы нашли QR-факторизацию (9.2) матрицы  $A_k$ . Теперь, принимая во внимание, что

$$R_{k+1} = Z_{k+1}^T A Z_k,$$

как и в (9.1), будем иметь

$$\begin{aligned} Z_{k+1}^T A Z_{k+1} &= Z_{k+1}^T A (Z_k Z_k^T) Z_{k+1} = (Z_{k+1}^T A Z_k) Z_k^T Z_{k+1} = \\ &= R_{k+1} Z_k^T Z_{k+1} = R_{k+1} Q_{k+1} =: A_{k+1}. \end{aligned}$$

Теорема доказана.

## 9.1 Ускорение сходимости QR-алгоритма

Пусть  $H$  — матрица Хессенберга, полученная из матрицы  $A$ , с тем, чтобы применить к ней QR-алгоритм для отыскания собственных значений.

**Определение 9.2.** Матрица Хессенберга называется неразложимой, если все ее поддиагональные элементы  $h_{i+1,i}$ ,  $i = 1, 2, \dots, n-1$  отличны от нуля.

Если для некоторого  $p$ ,  $1 \leq p \leq n-1$ , элемент  $h_{p+1,p} = 0$ , то матрица Хессенберга может быть представлена в блочно-треугольном виде

$$H = \begin{bmatrix} H_{11} & H_{12} \\ & H_{22} \end{bmatrix},$$

где  $H_{11}$  и  $H_{22}$  — матрицы Хессенберга порядков  $p$  и  $n - p$ , соответственно. Поскольку объединение спектров матриц  $H_{11}$  и  $H_{22}$  совпадает со спектром матрицы  $H$ , то запланированные QR-итерации с целью экономии затрат следует проводить не с  $H$ , а по отдельности с  $H_{11}$  и  $H_{22}$ . Поэтому в дальнейших рассуждениях будем предполагать, что матрица  $H$  является неразложимой.

**Замечание 9.3.** Если вышеуказанное число  $p = n - 1$ , т.е. нулевым поддиагональным элементом матрицы  $H$  является  $h_{n,n-1}$ , то порядок матрицы  $H_{22}$  равен единице, и ее единственное собственное значение, которое также является минимальным по модулю собственным значением матрицы  $H$ , совпадает с ее единственным элементом  $h_{n,n}$ . Отсюда можно предположить, что, если  $h_{n,n-1}$  мал, то  $h_{n,n}$  близок к минимальному по модулю собственному значению матрицы  $H$ .

Рассмотрим последовательность итераций QR-алгоритма для хессенберговой матрицы  $H$  с элементами  $h_{i,j}$ . Пусть  $\lambda_1, \lambda_2, \dots, \lambda_n$  — собственные значения матрицы  $H$ , упорядоченные так, что

$$|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_n|.$$

Если все неравенства строгие, то, в силу теоремы 9.4 о сходимости QR-алгоритма, элементы

$$h_{i+1,i}^{(k)}, \quad i = 1, 2, \dots, n - 1$$

матрицы  $H_k$  стремятся к нулю при  $k \rightarrow \infty$ . При этом скорость стремления к нулю является линейной и определяется отношениями  $|\lambda_{i+1}/\lambda_i|$ . Если для каких-то  $i$  модули собственных чисел  $\lambda_i$  и  $\lambda_{i+1}$  отличаются мало, то соответствующие элементы  $h_{i+1,i}^{(k)}$  к нулю будут стремиться медленно. Однако, мы можем улучшить скорость сходимости, уменьшая одно или несколько отношений  $|\lambda_{i+1}/\lambda_i|$ . Сделать это можно при помощи сдвига собственных значений.

Сдвинутая матрица, т.е. матрица  $H - \rho I$  имеет собственные значения

$$\lambda_1 - \rho, \lambda_2 - \rho, \dots, \lambda_n - \rho. \quad (9.7)$$

Если мы перенумеруем собственные значения таким образом, что

$$|\lambda_1 - \rho| \geq |\lambda_2 - \rho| \geq \dots \geq |\lambda_n - \rho|, \quad (9.8)$$

то отношениями собственных чисел, отвечающими  $H - \rho I$ , будут

$$|(\lambda_{i+1} - \rho)/(\lambda_i - \rho)|, \quad i = 1, 2, \dots, n - 1.$$

На самом деле, сделать действительно малым можно лишь отношение

$$|(\lambda_n - \rho)/(\lambda_{n-1} - \rho)|.$$

При условии  $\lambda_n \neq \lambda_{n-1}$  его можно сделать сколь угодно близким к нулю, если выбрать  $\rho$  очень близким к  $\lambda_n$ .

Заметим, что  $\lambda_i$  в (9.7) — это, вообще говоря, не  $\lambda_i$  в (9.8), ибо мы после сдвига произвели перенумерацию собственных значений. Поэтому в роли  $\lambda_n$  из (9.8) может выступать любое  $\lambda_i$  из (9.7), к которому априори мы можем найти хорошее приближение.

Если мы это сделаем и применим QR-алгоритм к сдвинутой матрице  $H - \rho I$ , то в новом итерационном процессе элемент  $h_{n,n-1}^{(k)}$  будет стремиться к нулю очень быстро. Как только он станет достаточно малым, его можно будет рассматривать практически равным нулю. Полученная в результате QR-итераций матрица  $H_k$  приближает форму Шура матрицы  $H - \rho I$ . Форму Шура матрицы  $H$  будет приближать матрица

$$H_k + \rho I = \left[ \begin{array}{ccc|c} \widehat{H}^k & & & * \\ & & & \vdots \\ & & & * \\ - & - & - & - \\ 0 & \dots & 0 & h_{n,n}^{(k)} \end{array} \right]$$

Эта матрица является блочно-треугольной, и поэтому ее собственными значениями будут собственные значения матрицы  $\widehat{A}_k$  и число  $h_{n,n}$ .

Если помимо  $h_{n,n}$  требуется найти и остальные собственные значения матрицы  $A$ , то дальнейшие QR-итерации можно осуществлять с хессенберговой матрицей  $\widehat{H}_k$  размера  $(n-1) \times (n-1)$ , что в конце концов приведет к существенному уменьшению трудозатрат.

Чтобы применить эти соображения, нужно иметь хорошие приближения к собственным числам. Как можно найти эти приближения? Предположим, что сначала мы делаем несколько QR-итераций без сдвига. Через некоторое время матрицы начнут приобретать треугольный вид, а элементы главной диагонали будут приближаться к собственным значениям. В частности,  $h_{n,n}^{(k)}$  будет приближаться к  $\lambda_n$  — наименьшему по модулю собственному значению матрицы  $A$ . Поэтому разумно в какой-то момент положить  $\rho = h_{n,n}^{(k)}$  (сдвиг Рэля) и последующие итерации выполнять для сдвинутой матрицы  $H_k - \rho I$ . В действительности можно сделать еще лучше. С каждым шагом мы получаем все лучшие приближения к  $\lambda_n$ .



Нет никаких причин, мешающих нам обновлять часто значение сдвига, чтобы улучшить скорость сходимости. На самом деле, мы можем выбирать новый сдвиг на каждом шаге. Именно это и делает *QR-алгоритм со сдвигом*:

$$A_{k-1} - \rho_{k-1}I = Q_k R_k, \quad R_k Q_k + \rho_{k-1}I = A_k.$$

Эта последовательность генерирует матрицы, подобные  $A$ . В самом деле,

$$\begin{aligned} A_k &= R_k Q_k + \rho_{k-1}I = Q_k^T [Q_k R_k Q_k + \rho_{k-1}Q_k] = \\ &= Q_k^T [Q_k R_k + \rho_{k-1}I] Q_k = Q_k^T A_{k-1} Q_k. \end{aligned}$$

На каждом шаге  $\rho_{k-1}$  выбирается так, чтобы оно являлось приближением собственного значения, появляющегося в правом нижнем углу матрицы.

**Теорема 9.5.** Пусть  $\mu$  — собственное значение неприводимой хессенберговой матрицы  $H$  размера  $n \times n$ . Тогда, если

$$H - \mu I = QR,$$

то у матрицы

$$\tilde{H} = RQ + \mu I$$

в последней строке

$$h_{n,n-1} = 0, \quad h_{n,n} = \mu.$$

**Доказательство.** Так как  $H$  — неприводимая хессенбергова матрица, то первые  $(n-1)$  ее столбцов линейно независимы (нижний левый угловой минор  $H$  является определителем верхней треугольной матрицы с главной диагональю из ненулевых чисел). То же самое верно и для матрицы  $H - \mu I$  при любом  $\mu$ . Поэтому, если

$$QR = H - \mu I$$

есть  $QR$ -разложение, то  $r_{i,i} \neq 0$ ,  $i = 1, \dots, n-1$ . Но, если матрица  $H - \mu I$  вырожденная, то

$$r_{1,1} \dots r_{nn} = 0,$$

и, следовательно,  $r_{n,n} = 0$ . Значит, у матрицы  $R$  последняя строка нулевая, равно как и у матрицы  $RQ$ . Поэтому матрица

$$\tilde{H} = RQ + \mu I$$

имеет последнюю строку из нулей и число  $\mu$  в последней позиции. Теорема доказана.

Доказанная теорема говорит о том, что, если мы сдвигаем на точное собственное значение, то в точной арифметике исчерпывание происходит на первом шаге.

Опыт показывает, что не надо ждать, пока  $a_{nn}^{(m)}$  станет хорошим приближением для  $\lambda_n$ : ничто не мешает нам делать сдвиги на первых же шагах.

**Пример.** Рассмотрим матрицу

$$A = \begin{bmatrix} 8 & 2 \\ 2 & 5 \end{bmatrix},$$

у которой собственные значения равны

$$\lambda_1 = 9 \quad \text{и} \quad \lambda_2 = 4.$$

Когда QR-алгоритм без сдвига применяется к  $A$ , элемент  $a_{2,1}^{(k)}$  сходится к нулю линейно с коэффициентом  $\lambda_2/\lambda_1 = 4/9$ . Если же применить QR-алгоритм со сдвигом Рэля, то следует положить  $\rho_0 = 5$  и выполнить QR-шаг с матрицей  $A - \rho_0 I$ . Собственными значениями этой матрицы являются  $\lambda_1 - 5 = 4$  и  $\lambda_2 - 5 = -1$ . Отношение  $|\lambda_2 - \rho_0|/|\lambda_1 - \rho_0| = 1/4$ , что меньше  $4/9$ , и можно ожидать улучшения сходимости для шага со сдвигом. Далее,

$$A_0 - \rho_0 I = Q_1 R_1,$$

где

$$Q_1 = \frac{1}{\sqrt{3}} \begin{bmatrix} 3 & 2 \\ 2 & -3 \end{bmatrix}, \quad R_1 = \frac{1}{\sqrt{3}} \begin{bmatrix} 13 & 6 \\ 0 & 4 \end{bmatrix}.$$

Таким образом,

$$A_1 = R_1 Q_1 + \rho_0 I = \frac{1}{13} \begin{bmatrix} 51 & 8 \\ 8 & -12 \end{bmatrix} \approx \begin{bmatrix} 8.9231 & 0.6154 \\ 0.6154 & 4.0769 \end{bmatrix}.$$

На четвертом шаге

$$A_4 = \begin{bmatrix} \overbrace{9.0 \dots 0}^{14} & 0.0 \dots 0 \\ 0.0 \dots 0 & 4.0 \dots 0 \end{bmatrix}$$

Приведем другой пример, показывающий, что стратегия Рэля не всегда срабатывает.

**Пример.** Рассмотрим матрицу

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix},$$

собственными значениями которой являются числа

$$\lambda_1 = 3 \quad \text{и} \quad \lambda_2 = 1.$$

Сдвиг Рэлея равен  $\rho = 2$ , что находится ровно посередине между собственными значениями. Сдвинутая матрица  $A - \rho I$  имеет собственные значения  $\pm 1$ , которые одинаковы по абсолютной величине. Поскольку  $A - \rho I$  ортогональная, ее  $QR$ -множители суть  $Q_1 = A - \rho I$  и  $R_1 = I$  (В силу "единственности"  $QR$ -разложения).  $QR$ -алгоритм с таким сдвигом сходиться не будет.

Поскольку итерации со сдвигом Рэлея иногда оказываются несостоятельными, предпочтительнее оказывается другой сдвиг — сдвиг Уилкинсона, определяемый как собственное значение последней  $2 \times 2$  подматрицы

$$\begin{bmatrix} a_{n-1,n-1}^{(k-1)} & a_{n-1,n}^{(k-1)} \\ a_{n,n-1}^{(k-1)} & a_{n,n}^{(k-1)} \end{bmatrix},$$

ближайшее к  $a_{n,n}^{(k-1)}$ . Есть надежда, что этот сдвиг лучше сдвига Рэлея. По крайней мере, для случая симметричных матриц известно, что  $QR$ -алгоритм с этим сдвигом всегда сходится.

Для общих матриц все еще остаются очень специфические случаи, когда и сдвиг Уилкинсона терпит неудачу.

**Пример.** Пусть

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

Эта верхняя хессенбергова матрица является ортогональной, сдвиг Уилкинсона для нее равен нулю, а  $QR$ -разложение ничего не меняет, как и в предыдущем примере.

Для подавляющего числа матриц стратегия сдвига Уилкинсона работает очень хорошо. Опыт показывает, что обычно требуется от пяти до десяти  $QR$ -циклов, прежде чем появится первое собственное значение. Поэтому в среднем последующие собственные значения будут выявляться за меньшее число итераций. Обычно большая часть последующих собственных значений появляется после одного-двух шагов. В среднем требуется от трех до пяти итераций на собственное значение. Для симметричных матриц дело обстоит еще лучше: требуется лишь от двух до трех итераций на собственное значение. В силу этого обстоятельства  $QR$ -алгоритм со сдвигом Уилкинсона некоторые называют "прямым" методом.

По ходу выполнения  $QR$ -итераций иногда случается, что один из поддиагональных элементов (вне последней строки) становится (практически) равным нулю. Для действительно больших матриц — это обычное явление. Всякий раз, как это происходит, задачу можно разложить, т.е. можно разбить на две меньшие задачи. Пусть, например,  $a_{i-1,i}^{(k)} = 0$ . Тогда  $A_k$  имеет вид

$$A_k = \begin{bmatrix} B_{11} & B_{12} \\ 0 & D_{22} \end{bmatrix},$$

где  $B_{11} \in R^{i \times i}$ ,  $B_{22} \in R^{j \times j}$ ,  $i + j = n$ . Теперь можно найти собственные значения  $B_{11}$  и  $B_{22}$ .

Исследованный нами ранее способ ускорения сходимости при помощи сдвигов, которые мы впредь будем называть одинарными сдвигами, не поможет нам, если на очереди стоит комплексное собственное значение, а сдвиги мы делаем действительными (как при стратегии Рэлея). Если же сдвиги делать комплексными (которые могут появиться из стратегии Уилкинсона), то нужно вроде бы переходить к комплексной арифметике, чего делать по известным причинам не хочется. Поскольку комплексное собственное значение у действительной матрицы появляется в паре с комплексно сопряженным собственным значением (они имеют одинаковые модули), то естественно эту пару попытаться находить одновременно. И сделать это можно с использованием двойного сдвига.

Пусть  $A \in \mathbb{R}^{n \times n}$  — действительная неразложимая верхняя хессенбергова матрица. Рассмотрим пару  $QR$ -шагов со сдвигами  $\rho$  и  $\tau$ , которые могут быть комплексными:

$$\begin{aligned} A - \rho I &= Q_\rho R_\rho, & R_\rho Q_\rho + \rho I &= \check{A}, \\ \check{A} - \tau I &= Q_\tau R_\tau, & R_\tau Q_\tau + \tau I &= \hat{A} \end{aligned} \quad (9.9)$$

**Теорема 9.6** (теорема "единственности"). Пусть  $Q$  и  $P$  — действительные ортогональные матрицы, преобразующие действительную матрицу  $A$  в неразложимую матрицу Хессенберга

$$H_Q = Q^T A Q, \quad H_P = P^T A P. \quad (9.10)$$

Если первые столбцы матриц  $Q$  и  $P$  совпадают, то найдется диагональная ортогональная матрица  $D$  (т.е. диагональная матрица с  $\pm 1$  на главной диагонали) такая, что

$$P = QD, \quad H_D = DH_Q D. \quad (9.11)$$

Иными словами, матрицы  $Q$  и  $P$  могут различаться лишь знаками столбцов, а хессенберговы матрицы  $H_Q$  и  $H_P$  — только знаками внедиагональных элементов.

**Доказательство.** Пусть  $D = Q^T P$ . Поскольку  $Q$  и  $P$  — ортогональные матрицы, то  $D$  тоже ортогональная, т.е.  $D^{-1} = D^T$ . Если мы еще докажем, что  $D$  является верхней треугольной матрицей, то отсюда будет следовать ее диагональность в виде

$$D = \text{diag} \{ \pm 1, \pm 1, \dots, \pm 1 \}.$$

Итак, нужно доказать, что  $D$  есть верхняя треугольная матрица. Принимая во внимание (9.10), находим, что

$$H_Q D = (H_Q Q^T) P = Q^T (A P) = (Q^T P) H_P = D H_P,$$

т.е.  $H_Q D = D H_P$ . Отсюда следует, что (если обозначить через  $W_j$  столбец с номером  $j$  матрицы  $W$ )

$$H_Q D_j = (H_Q D)_j = (D H_P)_j. \quad (9.12)$$

Пусть  $d_{i,k}$  и  $h_{k,j}$  суть элементы матриц  $D$  и  $H_P$ , соответственно. Тогда для элемента  $(i, j)$  матрицы  $D H_P$  имеет место представление

$$(D H_P)_{i,j} = \sum_{k=1}^n d_{i,k} h_{k,j} = \sum_{k=1}^{j+1} d_{i,k} h_{k,j},$$

а для ее  $j$ -го столбца —

$$(D H_P)_j = \sum_{k=1}^{j+1} h_{k,j} D_k = h_{j+1,j} \cdot D_{j+1} + \sum_{k=1}^j h_{k,j} D_k.$$

Поэтому, с учетом (9.12)

$$h_{j+1,j} \cdot D_{j+1} = H_Q D_j - \sum_{k=1}^j h_{k,j} D_k. \quad (9.13)$$

Напомним, что  $h_{j+1,j}$  отличны от нуля, ибо матрица  $H_P$  неразложимая. В силу определения матрицы  $D$  ее первый столбец  $D_1 = Q^T P_1$ . Так как строки  $Q^T$  ортогональны, а  $Q_1^T = P_1$ , то

$$D_1 = [10 \dots 0]^T.$$

Мы доказали, что первый столбец матрицы  $D$  имеет вид первого столбца верхней треугольной матрицы. Пусть это верно для  $j$ -го столбца  $D_j$ . Докажем, что это будет верно и для  $D_{j+1}$ . Для этого обратимся к его представлению (9.13). Поскольку лишь первые  $j$  элементов столбца  $D_j$  отличны от нуля, то

$$h_{j+1,j}D_{j+1} = \sum_{i=1}^j d_{i,j}(H_Q)_i - \sum_{k=1}^j h_{k,j}D_k.$$

Отсюда и следует, что

$$D_{j+1} = [d_{1,j+1}d_{2,j+1} \dots d_{j+1,j+1}0 \dots 0]^T.$$

Утверждаемые теоремой свойства матрицы  $D$  установлены. Из первого соотношения (9.11) и второго соотношения (9.10) следует второе соотношение (9.11). Поскольку теперь

$$h_{ij} = d_{ii}(h_Q)_{ij}d_{jj},$$

то теорема полностью доказана.

# Численные методы математического анализа

# 10

## Разностные уравнения

Пусть  $y(n) = y_n$  — функция целочисленного аргумента  $n \in \mathbb{Z}$ . Будем ее называть сеточной функцией. Обозначим через  $\nabla$  (набла) оператор левой конечной разности (разности назад), т.е.

$$\nabla y_n = y_n - y_{n-1}. \quad (10.1)$$

Степень оператора  $\nabla$  определим рекуррентным образом

$$\nabla^k = \nabla(\nabla^{k-1}). \quad (10.2)$$

Пусть  $a(n)$ ,  $b(n)$ ,  $c(n)$ ,  $d(n)$  и  $f(n)$  — заданные сеточные функции. Рассмотрим уравнение

$$a(n)\nabla^3 y_n + b(n)\nabla^2 y_n + c(n)\nabla y_n + d(n)y_n = f(n) \quad (10.3)$$

относительно сеточной функции  $y_n$ . Уравнение (10.3) называется *разностным уравнением*. Разностные уравнения являются аналогами дифференциальных уравнений и в значительной степени повторяют свойства последних. Как и в случае дифференциальных уравнений, важным является понятие порядка разностного уравнения. Если  $a(n) \neq 0$ , то казалось бы естественным объявить порядком уравнения (10.3) число три. Однако при таком определении порядка разностного уравнения нас ждут неприятности. Чтобы убедиться в этом, положим в (10.3)  $a(n) = 1$ ,  $b(n) = 0$ ,  $c(n) = -3$ ,  $d(n) = 2$ . В результате получим уравнение

$$\nabla^3 y_n - 3\nabla y_n + 2y_n = f(n). \quad (10.4)$$

Принимая во внимание (10.1) и (10.2), находим, что

$$\nabla y_n = y_n - y_{n-1}, \quad \nabla^3 y_n = y_n - 3y_{n-1} + 3y_{n-2} - y_{n-3},$$



а, подставляя эти выражения в (10.4), будем иметь

$$y_n - 3y_{n-1} + 3y_{n-2} - y_{n-3} - 3y_n + 3y_{n-1} + 2y_n = f(n)$$

или

$$3y_{n-2} - y_{n-3} = f(n). \quad (10.5)$$

Вводя новый индекс  $m = n - 2$ , уравнение (10.5) преобразуем к виду

$$3y_m - y_{m-1} = f(m + 2). \quad (10.6)$$

Это уравнение эквивалентно уравнению (10.4), и назвать его разностным уравнением третьего порядка просто не поворачивается язык. И дело, конечно, не просто в названии. От удачно введенного определения зависит простота последующих утверждений, использующих это определение. Поскольку запись (10.3) не содержит явным образом информации о числе, которым следовало бы определить порядок разностного уравнения, то будем разностное уравнение записывать в виде

$$\Phi(n, y_n, y_{n-1}, \dots, y_{n-k}) = 0. \quad (10.7)$$

**Определение 10.1.** Уравнение (10.7) называется разностным уравнением.

**Определение 10.2.** Разностное уравнение (10.7), если оно явно зависит от  $y_n$  и от  $y_{n-k}$ , называется уравнением  $k$ -го порядка.

**Определение 10.3.** Разностное уравнение  $k$ -го порядка называется линейным, если оно линейно зависит от  $y_n, y_{n-1}, \dots, y_{n-k}$ .

Мы будем изучать только линейные разностные уравнения, которые будем записывать в виде

$$\sum_{j=0}^k \alpha_j(n) y_{n-j} = f(n), \quad n \in \mathbb{Z}. \quad (10.8)$$

Пока мы предполагаем, что уравнение (10.8) задано при всех  $n \in \mathbb{Z}$ . Уравнение (10.8) будет уравнением  $k$ -го порядка, если коэффициенты  $\alpha_0(n)$  и  $\alpha_k(n)$  не обращаются в нуль ни при одном  $n \in \mathbb{Z}$ .

**Определение 10.4.** Сеточная функция  $y_n$ ,  $n \in \mathbb{Z}$  называется решением уравнения (10.8), если при подстановке ее в (10.8) последнее превращается в тождество.

**Определение 10.5.** Сеточная функция  $y_n$ ,  $n \in \mathbb{Z}$  называется общим решением разностного уравнения (10.8), если в ней содержится любое решение (10.8).

Для того, чтобы определить какое-либо решение уравнения (10.8) (частное решение) достаточно указать его значения в любых  $k$  последовательных точках, например,  $n_0, n_0 + 1, \dots, n_0 + k - 1$ .

## 10.1 Линейные разностные уравнения первого порядка

Эти уравнения имеют вид

$$\alpha_0(n)y_n + \alpha_1(n)y_{n-1} = f(n). \quad (10.9)$$

Поскольку  $\alpha_0(n) \neq 0$ , то на этот коэффициент уравнение можно поделить. Пусть  $\alpha_1(n)/\alpha_0(n) = -q_n \neq 0$ , а  $f(n)/\alpha_0(n)$  снова обозначим через  $f(n) = f_n$ . Тогда разностное уравнение первого порядка (10.9) можно переписать так

$$y_n = q_n y_{n-1} + f_n. \quad (10.10)$$

Разрешить разностное уравнение — значит выразить  $y_n$  через известные величины. Чтобы можно было решить (10.10), нужно задать начальное условие

$$y_0 = a. \quad (10.11)$$

Используя теперь рекуррентные соотношения (10.10), можно последовательно определить  $y_n$  при всех последующих значениях  $n$ :

$$\begin{aligned} y_1 &= q_1 y_0 + f_1 = q_1 a + f_1, \\ y_2 &= q_2 y_1 + f_2 = q_2(q_1 a + f_1) + f_2 = q_1 q_2 a + q_2 f_1 + f_2 \end{aligned}$$

и т.д.

Часто бывает полезно иметь не рекуррентное соотношение для последовательного вычисления решения, а некоторую формулу, представляющую решение. Найдем представление решения уравнения (10.10). Для этого рассмотрим сначала отвечающее ему однородное уравнение

$$y_n = q_n y_{n-1} \quad (10.12)$$

и найдем его решение. Имеем

$$\begin{aligned} y_1 &= q_1 y_0, \\ y_2 &= q_2 y_1, \\ &\dots\dots\dots \\ y_n &= q_n y_{n-1}. \end{aligned}$$

Перемножая последовательно полученные равенства и сокращая левую и правую части на  $y_1 y_2 \dots y_{n-1}$ , получим

$$y_n = q_1 \dots q_n y_0 = y_0 \prod_{j=1}^n q_j. \quad (10.13)$$

Величина  $y_0$  есть начальное значение  $y_n$  и является произвольной постоянной. Решение однородного уравнения (10.12) найдено.

**Замечание 10.1.** Напомним, что если линейное однородное дифференциальное уравнение первого порядка записать в виде  $y' = P(x)y$ , то его общее решение примет вид

$$y(x) = c \exp \left\{ \int_0^x P(\xi) d\xi \right\}.$$

Обратимся теперь к неоднородному уравнению (10.10). Его решение будем искать, используя решение однородного уравнения (10.12), методом вариации постоянной. Пусть

$$\overset{\circ}{y}_n = \prod_{j=1}^n q_j. \quad (10.14)$$

Это — решение уравнения (10.12), а  $c \overset{\circ}{y}_n$  — его общее решение. Заставим коэффициент  $c$  зависеть от  $n$  и в таком виде будем искать решение уравнения (10.10)

$$y_n = c_n \overset{\circ}{y}_n. \quad (10.15)$$

Подставляя (10.15) в (10.10), получим

$$c_n \overset{\circ}{y}_n = q_n c_{n-1} \overset{\circ}{y}_{n-1} + f_n.$$

Из (10.12)  $\overset{\circ}{y}_n = q_n \overset{\circ}{y}_{n-1}$  и поэтому

$$c_n \overset{\circ}{y}_n = c_{n-1} \overset{\circ}{y}_n + f_n,$$

т.е.

$$c_n = c_{n-1} + f_n / \overset{\circ}{y}_n.$$

Отсюда

$$\begin{aligned} c_1 &= c_0 + f_1 / \overset{\circ}{y}_1, \\ c_2 &= c_1 + f_2 / \overset{\circ}{y}_2, \\ &\dots\dots\dots \\ c_n &= c_{n-1} + f_n / \overset{\circ}{y}_n, \end{aligned}$$

Складывая эти соотношения, находим, что

$$c_n = \sum_{k=1}^n \frac{f_k}{\overset{\circ}{y}_k} + c_0,$$

а принимая во внимание (10.14), будем иметь

$$c_n = \sum_{k=1}^n f_k \prod_{j=1}^k q_j^{-1} + c_0.$$

Подставляя это выражение в (10.15), получим общее решение неоднородного уравнения (10.10)

$$y_n = \prod_{j=1}^n q_j \left( c + \sum_{k=1}^n f_k \prod_{j=1}^k q_j^{-1} \right). \quad (10.16)$$

**Замечание 10.2.** Напомним, что если линейное неоднородное дифференциальное уравнение первого порядка записать в виде  $y' = P(x)y + f(x)$ , то его общее решение примет вид

$$y(x) = \exp \left\{ \int_0^x P(\xi) d\xi \right\} \left( c + \int_0^x f(\eta) \exp \left\{ - \int_0^\eta P(\xi) d\xi \right\} d\eta \right).$$

Если коэффициент  $q_n = \text{const} = q$ , то из (10.16) находим, что

$$y_n = q^n \left( c + \sum_{k=1}^n f_k q^{-k} \right), \quad (10.17)$$

а если и  $f_n = \text{const} = f$ , то при  $q \neq 1$

$$y_n = q^n \left( c + f \sum_{k=1}^n q^{-k} \right) = q^n \left( c + f \frac{q^{-1} - q^{-n-1}}{1 - q^{-1}} \right) = cq^n + f \frac{1 - q^n}{1 - q}. \quad (10.18)$$

## 10.2 Линейные разностные уравнения $k$ -го порядка с постоянными коэффициентами

Если коэффициенты  $\alpha_j(n)$  из (10.8) не зависят от  $n$ , то мы имеем разностное уравнение с постоянными коэффициентами

$$\sum_{j=0}^k \alpha_j y_{n-j} = f_n, \quad n \in \mathbb{Z}, \quad \alpha_0 \alpha_k \neq 0. \quad (10.19)$$

Решение отвечающего (10.19) однородного уравнения

$$\sum_{j=0}^k \alpha_j y_{n-j} = 0 \quad (10.20)$$

можно искать в виде

$$y_n = q^n, \quad q \neq 0 \quad (\text{ср. с } y(x) = e^{\lambda x}), \quad (10.21)$$

где  $q = \text{const} \neq 0$ . Подставляя (10.21) в (10.20), получим

$$q^{n-k} \sum_{j=0}^k \alpha_j q^{k-j} = 0.$$

На  $q^{n-k}$  можно сократить, в результате чего для отыскания  $q$  получим алгебраическое уравнение степени  $k$

$$\alpha_0 q^k + \alpha_1 q^{k-1} + \dots + \alpha_{k-1} q + \alpha_k = 0, \quad (10.22)$$

называемое *характеристическим уравнением*, отвечающим разностному уравнению (10.20).

Характеристическое уравнение (10.22) имеет ровно  $k$  корней, включая кратные и комплексные. Обозначим их через

$$q_1, q_2, \dots, q_k. \quad (10.23)$$

Очевидно, что сеточные функции

$$q_l^n, \quad l = 1, \dots, k \quad (10.24)$$

являются решениями разностного уравнения (10.20).

Имеет место

**Теорема 10.1.** Если корни (10.23) характеристического уравнения (10.22) простые, то решения (10.24) разностного уравнения (10.20) линейно независимы, а общее решение этого уравнения имеет вид

$$y_n = \sum_{l=1}^k c_l q_l^n.$$

**Доказательство.** Проведем доказательство линейной независимости (10.24) при  $k = 2$ . Допустим противное, т.е. пусть

$$c_1 q_1^n + c_2 q_2^n \equiv 0, \quad |c_1| + |c_2| \neq 0.$$

Но тогда и

$$c_1 q_1^{n-1} + c_2 q_2^{n-1} = 0.$$

Рассмотрим эти два тождества как систему уравнений для определения  $c_1$  и  $c_2$ . Находим, что определитель этой системы

$$\Delta = \begin{vmatrix} q_1^n & q_2^n \\ q_1^{n-1} & q_2^{n-1} \end{vmatrix} = (q_1 q_2)^{n-1} (q_1 - q_2) \neq 0$$

и, следовательно, система имеет лишь тривиальное решение  $c_1 = c_2 = 0$ . Это противоречит предположению, что и доказывает теорему.

**Замечание 10.3.** Если комплексное число  $q = |q|e^{i\varphi}$ ,  $\varphi \neq m\pi$ ,  $m \in \mathbb{Z}$  является корнем характеристического уравнения (10.22), коэффициенты которого действительны, то число  $\bar{q} = |q|e^{-i\varphi}$ , комплексно сопряженное к  $q$ , также является корнем характеристического уравнения (10.22), а наряду с комплексными решениями разностного уравнения (10.20)

$$q^n \quad \text{и} \quad \bar{q}^n \tag{10.25}$$

решениями указанного разностного уравнения будут и действительная и мнимая части решений (10.25), т.е.

$$|q|^n \cos n\varphi, \quad |q|^n \sin n\varphi. \tag{10.26}$$

Решения (10.26), как и (10.25), линейно независимы.

**Пример 10.1.** Найдём общее решение разностного уравнения

$$y_n - 2 \operatorname{ch} \alpha y_{n-1} + y_{n-2} = 0, \quad \alpha \neq 0.$$

Характеристическое уравнение этого разностного уравнения имеет вид

$$q^2 - 2 \operatorname{ch} \alpha q + 1 = 0,$$

а его корни суть

$$q_{1,2} = \operatorname{ch} \alpha \pm \sqrt{\operatorname{ch}^2 \alpha - 1} = e^{\pm \alpha}.$$

Эти корни различные, и поэтому

$$y_n = c_1 e^{\alpha n} + c_2 e^{-\alpha n}.$$

**Теорема 10.2.** *Если  $q$  есть корень характеристического уравнения (10.22) кратности  $s \geq 1$ , то сеточная функция*

$$P_{s-1}(n)q^n,$$

где  $P_{s-1}(n)$  — произвольный многочлен, степень которого не выше  $s-1$ , является решением разностного уравнения (10.20). При этом решения

$$n^l q^n, \quad l = 0, \dots, s-1$$

линейно независимы.

**Доказательство.** Доказательство проведем для случая  $s = k = 2$ . Покажем сначала, что  $nq^n$  есть решение уравнения (10.20) при  $k = 2$ . Имеем

$$\begin{aligned} \alpha_0 n q^n + \alpha_1 (n-1) q^{n-1} + \alpha_2 (n-2) q^{n-2} &= \\ &= q^{n-2} [(\alpha_0 q^2 + \alpha_1 q + \alpha_2)(n-2) + q(2\alpha_0 q + \alpha_1)] = \\ &= q^{n-1} (2\alpha_0 q + \alpha_1) = 0, \end{aligned}$$

ибо  $2\alpha_0 q + \alpha_1 = (\alpha_0 q^2 + \alpha_1 q + \alpha_2)'$  и обязано обращаться в нуль на корне кратности два. Линейная независимость решений  $q^n$  и  $nq^n$  доказывается так же, как и в предыдущей теореме.

**Теорема 10.3.** *Если правая часть  $f(n)$  неоднородного разностного уравнения (10.19) имеет вид  $P_m(n) \overset{\circ}{q}^n$ , где  $P_m(n)$  — многочлен степени  $m$ , а  $\overset{\circ}{q}$  является  $s$ -кратным,  $s \geq 0$ , корнем характеристического уравнения (10.22), то уравнение (10.19) имеет решение вида*

$$y(n) = n^s Q_m(n) \overset{\circ}{q}^n. \quad (10.27)$$

Доказательство смотри, например, в [9].

**Теорема 10.4.** *Если правая часть  $f(n)$  неоднородного разностного уравнения (10.19) представима в виде  $f_1(n) + f_2(n)$ , а  $y_n^{(k)}$ ,  $k = 1, 2$ , есть решение этого уравнения с правой частью  $f_k^{(n)}$ , то  $y_n = y_n^{(1)} + y_n^{(2)}$ .*

**Упражнение 10.1.** Найти общее решение неоднородного разностного уравнения

$$y_n - 2y_{n-1} + y_{n-2} = n(1 + 2^n).$$

**Ответ.**

$$y_n = c_1 + c_2 n + \frac{1}{6}(n+3)n^2 + (n-2)2^{n+2}.$$

### 10.3 Системы разностных уравнений

Рассмотрим систему двух линейных однородных разностных уравнений первого порядка с двумя неизвестными функциями

$$\begin{aligned} u_n &= a_{11}u_{n-1} + a_{12}v_{n-1}, \\ v_n &= a_{21}u_{n-1} + a_{22}v_{n-1}, \end{aligned} \quad n \in \mathbb{Z}, \quad (10.28)$$

где  $a_{ij}$ ,  $i, j = 1, 2$  — постоянные. Введем в рассмотрение вектор-функцию

$$y(n) = [u_n \ v_n]^T$$

и матрицу

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

которую будем предполагать невырожденной,  $\det A \neq 0$ . Используя введенные обозначения, систему (10.28) можно переписать в векторном виде

$$y(n) = Ay(n-1), \quad \det A \neq 0, \quad \text{или} \quad A^{-1}y(n) = y(n-1). \quad (10.29)$$

В записи (10.29) можно забыть, что  $y(n)$  был двумерный вектор, а  $A$  — матрица второго порядка. Будем мыслить систему (10.29) как систему  $m$ -го порядка. Решение этой системы будем искать в виде

$$y(n) = \xi q^n \quad (10.30)$$

где  $q = \text{const} \neq 0$ , а  $\xi$  — ненулевой  $m$ -мерный вектор. Подставляя (10.30) в (10.29), получим

$$\xi q^n = A\xi q^{n-1},$$

а сокращая на  $q^{n-1} \neq 0$ , приходим к системе

$$\xi q = A\xi.$$



Эта система однородна, и, чтобы у нее были нетривиальные решения, определитель ее матрицы должен быть равен нулю

$$|A - qI| = 0. \quad (10.31)$$

Уравнение (10.31) — *характеристическое уравнение* системы (10.29) — является алгебраическим уравнением  $m$ -ой степени. Если все его корни  $q_k$  — собственные значения матрицы  $A$  — различны, то соответствующие собственные векторы  $\xi_k$  линейно независимы, и общее решение системы (10.29) принимает вид

$$y(n) = \sum_{k=1}^m c_k \xi_k q_k^n. \quad (10.32)$$

Матрица  $A$  может иметь полный набор линейно независимых собственных векторов и при наличии кратных корней характеристического уравнения (10.31). И в этом случае общее решение системы (10.29) имеет вид (10.32). Если же у канонической формы матрицы  $A$  имеются жордановы клетки, то для отыскания общего решения системы (10.29) нужно поступать так же, как и в случае систем дифференциальных уравнений. На этом мы останавливаться не будем.

**Упражнение 10.2.** Найти общее решение системы (10.29) с матрицей

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

**Ответ.**

$$y(n) = \left\{ c_1 \begin{bmatrix} -\sin \frac{3\pi n}{4} \\ \cos \frac{3\pi n}{4} \end{bmatrix} + c_2 \begin{bmatrix} \cos \frac{3\pi n}{4} \\ \sin \frac{3\pi n}{4} \end{bmatrix} \right\} q^{n/2}.$$

## 10.4 Разностная задача на собственные значения

До сих пор мы обсуждали вопросы отыскания общего решения разностного уравнения, которое зависит от  $k$  произвольных постоянных, если уравнение имеет порядок  $k$ . Чтобы выделить единственное решение разностного уравнения, как и в случае дифференциального уравнения  $k$ -го порядка, нужно задать  $k$  линейно независимых начальных или граничных условий. Как и для дифференциального уравнения, для разностного уравнения можно поставить задачу на собственные значения.



Характеристическое уравнение, отвечающее уравнению (10.36), есть

$$q^2 - 2 \cos \frac{\alpha}{N} q + 1 = 0,$$

а его корни суть

$$q_{1,2} = \cos \frac{\alpha}{N} \pm \sqrt{\cos^2 \frac{\alpha}{N} - 1} = \cos \frac{\alpha}{N} \pm i \sin \frac{\alpha}{N} = e^{\pm \alpha i / N}.$$

Тем самым,

$$y_n = c_1 \sin \frac{\alpha n}{N} + c_2 \cos \frac{\alpha n}{N} \quad (10.37)$$

есть общее решение уравнения (10.36). Потребуем, чтобы это решение удовлетворяло граничным условиям (10.34). Будем иметь

$$y_0 = c_2 = 0, \quad y_N = c_1 \sin \alpha = 0. \quad (10.38)$$

Поскольку  $c_1$  не может быть нулевым, то отсюда находим, что  $\sin \alpha = 0$  и, следовательно,

$$\alpha = k\pi, \quad k \in \mathbb{Z}. \quad (10.39)$$

Из (10.35) находим

$$\lambda = \lambda_k = 2 \left( 1 - \cos \frac{k\pi}{N} \right) = 4 \sin^2 \frac{k\pi}{2N}, \quad k \in \mathbb{Z}, \quad (10.40)$$

а из (10.38) —

$$y_n = y_n^{(k)} = c_1 \sin \frac{k\pi n}{N}. \quad (10.41)$$

При  $k = 0$  решение  $y_n^{(0)} \equiv 0$  и, следовательно, число  $\lambda_0 = 0$  не является собственным значением. При  $k = N$  решение  $y_n^{(N)} = c \sin N\pi \equiv 0$ , и  $\lambda_N$  тоже не может быть собственным значением. Собственные значения

$$\lambda_k = 4 \sin^2 \frac{k\pi}{2N}, \quad k = 1, \dots, N-1 \quad (10.42)$$

различны, ибо функция  $\sin t$  при  $0 < t < \pi/2$  является монотонной. Поскольку изучаемая задача эквивалентна алгебраической задаче на собственные значения для матрицы  $(N-1)$  порядка, то соотношения (10.42) задают все собственные значения задачи (10.33). Собственные функции

$$y_n^{(k)} = c_1 \sin \frac{k\pi n}{N}, \quad k = 1, \dots, N-1 \quad (10.43)$$

ортогональны. Подсчитаем их нормы

$$\begin{aligned} \|y_n^{(k)}\|^2 &= \left(y_n^{(k)}, y_n^{(k)}\right) = c_1^2 \sum_{n=1}^{N-1} \sin^2 \frac{k\pi n}{N} = \\ &= c_1^2 \sum_{n=1}^{N-1} \frac{1 - \cos \frac{2k\pi n}{N}}{2} = c_1^2 \left[ \frac{N-1}{2} - \frac{1}{2} \sum_{n=1}^{N-1} \cos \frac{2k\pi n}{N} \right]. \end{aligned}$$

Далее,

$$\sum_{n=1}^{N-1} \cos \frac{2k\pi n}{N} = \operatorname{Re} \sum_{n=1}^{N-1} \left( e^{\frac{2ik\pi}{N}} \right)^n = \operatorname{Re} \frac{e^{2ik\pi} - e^{2ik\pi/N}}{e^{2ik\pi/N} - 1} = -1.$$

Тем самым,

$$\|y_n^{(k)}\|^2 = c_1^2 N/2 = 1 \quad \text{при} \quad c_1 = \sqrt{2/N}, \quad (10.44)$$

а ортонормированные собственные функции суть

$$y_n^{(k)} = \sqrt{2/N} \sin \frac{k\pi n}{N}, \quad k = 1, \dots, N-1. \quad (10.45)$$

Из (10.42) следует, что для всех собственных значений предположение (10.34) выполнено. Поэтому в рассмотрении противоположного предположения смысла нет.

**Упражнение 10.3.** Решить следующую задачу на собственные значения

$$\begin{aligned} -y_{n+1} + 2y_n - y_{n-1}, \quad n = 1, \dots, N-1, \\ -y_1 + y_0 + \frac{\lambda}{2}y_0 = 0, \quad y_N - y_{N-1} + \frac{\lambda}{2}y_N = 0. \end{aligned}$$

**Ответ.**

$$\lambda_k = 4 \sin^2 \frac{k\pi}{2N}, \quad k = 0, \dots, N, \quad y_n^{(k)} = \sqrt{2/N} \cos \frac{k\pi n}{N}.$$

# 11

## Ортогональные многочлены

### 11.1 Общие ортогональные многочлены

Функцию  $\rho(x) \neq 0$  будем называть весовой функцией на интервале  $(-1, 1)$ , если на этом интервале она неотрицательна и интегрируема.

Пусть на  $(-1, 1)$  задана последовательность многочленов

$$P_0(x), P_1(x), \dots, P_n(x), \dots, \quad (11.1)$$

в которой каждый многочлен  $P_n(x)$  имеет степень  $n$ . Если для любых двух многочленов из этой последовательности выполняется условие

$$(P_m, P_n) := \int_{-1}^1 \rho(x) P_m(x) P_n(x) dx = 0, \quad m \neq n,$$

то многочлены (11.1) называются ортогональными на  $(-1, 1)$  с весом  $\rho(x)$ .

**Лемма 11.1.** *Если в системе из  $(n + 1)$  ортогональных многочленов*

$$P_0(x), P_1(x), \dots, P_n(x)$$

*каждый многочлен  $P_k(x)$  имеет степень  $k$ , то всякий многочлен  $Q_n(x)$  степени  $n$  можно единственным образом представить в виде*

$$Q_n(x) = a_0 P_0(x) + a_1 P_1(x) + \dots + a_n P_n(x). \quad (11.2)$$

**Доказательство.** Пусть ортогональные многочлены  $P_k(x)$  имеют вид

$$P_k(x) = c_0^{(k)} + c_1^{(k)} x + \dots + c_k^{(k)} x^k, \quad c_k^{(k)} \neq 0,$$

а многочлен

$$Q_n(x) = c_0 + c_1 x + \dots + c_n x^n.$$



где числа  $\mu_n$  и  $a_k$  подлежат определению. Умножая это соотношение скалярно на  $P_m(x)$ ,  $m = 0, \dots, n-1$ , находим, что

$$0 = \mu_n (x^n, P_m(x)) + a_m, \quad m = 0, \dots, n-1,$$

т.е.

$$a_m = -\mu_n (x^n, P_m(x))$$

и, следовательно,

$$P_n(x) = \mu_n \left[ x^n - \sum_{k=0}^{n-1} (x^n, P_k) x^k \right]$$

есть произвольный ортогональный многочлен степени  $n$ . Умножая его скалярно на самого себя и требуя нормированности, находим

$$1 = \mu_n^2 \underbrace{\left( \left( x^n - \sum_{k=1}^{n-1} (x^n, P_k) x^k \right)^2, 1 \right)}_0.$$

Отсюда определяем  $\mu_n > 0$ . Лемма доказана.

**Лемма 11.3.** Если  $P_n(x)$  принадлежит совокупности ортогональных многочленов, то для всякого многочлена  $Q_m(x)$  степени  $m < n$

$$(Q_m, P_n(x)) = 0, \quad m < n. \quad (11.3)$$

**Доказательство.** В силу леммы 11.1

$$Q_m(x) = a_0 P_0(x) + a_1 P_1(x) + \dots + a_m P_m(x).$$

Подставляя это представление в (11.3), получаем утверждение леммы.

**Лемма 11.4.** Если весовая функция  $\rho(x)$  четная, то каждый ортогональный многочлен  $P_n(x)$  содержит только те степени  $x$ , которые имеют одинаковую с номером  $n$  четность, т.е.

$$P_n(-x) \equiv (-1)^n P_n(x). \quad (11.4)$$

**Доказательство.** Пусть  $\rho(x) = \rho(-x)$  и

$$\int_{-1}^1 \rho(x) P_n(x) P_m(x) dx = 0, \quad m = 0, \dots, n-1.$$

Заменой переменной интегрирования  $x = -t$  эти условия приводятся к виду

$$\int_{-1}^1 \rho(t) P_n(-t) P_m(-t) dt = 0, \quad m = 0, \dots, n-1,$$

т.е.  $P_m(-t)$  тоже ортогональные многочлены. Но в силу леммы 11.2 любой ортогональный многочлен определен с точностью до множителя, и поэтому

$$P_n(-x) = c_n P_n(x).$$

Отсюда в частности следует, что

$$(-1)^n a_n x^n = c_n a_n x^n,$$

т.е.  $c_n = (-1)^n$  и поэтому

$$P_n(-x) = (-1)^n P_n(x).$$

Лемма доказана.

**Теорема 11.1.** *Для любых трех последовательных ортогональных многочленов справедлива рекуррентная формула*

$$\alpha_n P_{n+1}(x) = (x - \beta_n) P_n(x) - \gamma_n P_{n-1}(x). \quad (11.5)$$

**Доказательство.** Перепишем (11.5) в виде

$$x P_n(x) = \alpha_n P_{n+1}(x) + \beta_n P_n(x) + \gamma_n P_{n-1}(x).$$

В левой части этого равенства стоит многочлен степени  $n+1$ . В силу леммы 11.1 он может быть разложен по многочленам  $P_0, \dots, P_{n+1}$

$$x P_n(x) = \sum_{k=0}^{n+1} a_k^{(n+1)} P_k(x), \quad (11.6)$$

где в силу замечания 11.1

$$a_k^{(n+1)} = \frac{(x P_n, P_k)}{\|P_k\|^2}. \quad (11.7)$$

Но тогда с учетом (11.6)

$$\begin{aligned} a_k^{(n+1)} &= \frac{1}{\|P_k\|^2} (P_n, x P_k) = \frac{1}{\|P_k\|^2} \left( P_n, \sum_{j=0}^{k+1} a_j^{(k+1)} P_j \right) = \\ &= \frac{1}{\|P_k\|^2} \sum_{j=0}^{k+1} a_j^{(k+1)} (P_n, P_j). \end{aligned}$$



Поскольку

$$(P_n, P_0) = 0, \quad (P_n, P_1) = 0, \dots, (P_n, P_{n-1}) = 0,$$

то при  $k + 1 \leq n - 1$  коэффициенты  $a_k^{n+1} = 0$  и поэтому

$$xP_n(x) = a_{n+1}^{(n+1)}P_{n+1}(x) + a_n^{(n+1)}P_n(x) + a_{n-1}^{(n+1)}P_{n-1}(x), \quad (11.8)$$

т.е.

$$\alpha_n = a_{n+1}^{(n+1)}, \quad \beta_n = a_n^{(n+1)}, \quad \gamma_n = a_{n-1}^{(n+1)}. \quad (11.9)$$

Теорема доказана.

**Теорема 11.2.** *Все нули ортогонального многочлена  $P_n(x)$  действительны, различны и расположены на интервале  $(-1, 1)$ .*

**Доказательство.** Достаточно показать, что многочлен  $P_n(x)$  на  $(-1, 1)$  меняет знак  $n$  раз. Допустим противное, т.е. что многочлен  $P_n(x)$  меняет знак только в точках  $\xi_1, \xi_2, \dots, \xi_m$ , где  $m < n$ . Тогда многочлен

$$Q_m(x) = (x - \xi_1)(x - \xi_2) \dots (x - \xi_m)$$

также меняет знак только в этих точках. Следовательно, произведение  $P_n(x)Q_m(x) \not\equiv 0$  сохраняет знак на  $(-1, 1)$  и следовательно

$$\int_{-1}^1 \rho(x)P_n(x)Q_m(x)dx \neq 0,$$

что противоречит лемме 11.3. Противоречие снимается, если  $n = m$ . Теорема доказана.

**Замечание 11.2.** В силу доказанной теоремы для нулей  $x_k^{(n)}$  ортогонального многочлена  $P_n(x)$  имеют место неравенства

$$-1 < x_1^{(n)} < x_2^{(n)} < \dots < x_k^{(n)} < \dots < x_n^{(n)} < 1. \quad (11.10)$$

## 11.2 Многочлены Чебышева первого рода

Рассмотрим следующее однородное разностное уравнение второго порядка с постоянными коэффициентами

$$y_{n+1} - 2xy_n + y_{n-1} = 0. \quad (11.11)$$

Здесь  $x$  — параметр. Поставим для (11.11) начальные условия

$$y_0 = 1, \quad y_1 = x. \quad (11.12)$$

Тогда

$$\begin{aligned} y_2 &= 2x \cdot x - 1 = 2x^2 - 1, \\ y_3 &= 2x(2x^2 - 1) - x = 4x^3 - 3x, \\ y_4 &= 8x^4 - 8x^2 + 1, \dots \end{aligned} \quad (11.13)$$

Очевидно, что значение решения задачи (11.11), (11.12) в узле  $n$  есть многочлен от  $x$  степени  $n$ .

Найдем решение задачи (11.11), (11.12) в явном виде. Характеристическое уравнение разностного уравнения (11.11) имеет вид

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

а его корни суть

$$q_1 = q = x + \sqrt{x^2 - 1} \quad \text{и} \quad q_2 = 1/q. \quad (11.14)$$

Поэтому общее решение уравнения (11.11) есть

$$y_n = c_1 q^n + c_2 q^{-n}.$$

Полагая здесь  $n = 0$  и  $n = 1$  и принимая во внимание начальные условия (11.12), находим, что

$$\begin{aligned} y_0 &= c_1 + c_2 = 1, \\ y_1 &= c_1 q + c_2 q^{-1} = x. \end{aligned} \quad (11.15)$$

Преобразуем второе из уравнений (11.15) с учетом (11.14) и первого уравнения,

$$\begin{aligned} y_1 &= c_1(x + \sqrt{x^2 - 1}) + c_2(x - \sqrt{x^2 - 1}) = \\ &= (c_1 + c_2)x + (c_1 - c_2)\sqrt{x^2 - 1} = x + (c_1 - c_2)\sqrt{x^2 - 1} = x. \end{aligned}$$

Отсюда следует, что  $c_1 = c_2$ , а с учетом (11.15) находим, что

$$c_1 = c_2 = 1/2,$$

и поэтому

$$y_n = \frac{q^n + q^{-n}}{2} \quad (11.16)$$

есть решение задачи (11.11), (11.12).

Как было замечено раньше, это есть многочлен от  $x$  степени  $n$ . Пусть  $|x| < 1$ . Тогда в силу (11.14)

$$q = x + i\sqrt{1-x^2}$$

и, следовательно,  $|q| = 1$ . Пусть  $q = e^{i\varphi}$ . Тогда

$$\begin{aligned} x &= \cos \varphi, \quad \varphi = \arccos x, \\ y_n &= \frac{e^{in\varphi} + e^{-in\varphi}}{2} = \cos n\varphi = \cos[n \arccos x]. \end{aligned} \quad (11.17)$$

**Определение 11.1.** Алгебраические многочлены

$$T_n(x) = \cos[n \arccos x], \quad |x| < 1, \quad n = 0, 1, \dots \quad (11.18)$$

называются *многочленами Чебышева* первого рода.

Они принадлежат к семейству ортогональных многочленов. Определим весовую функцию  $\rho(x)$ , при которой многочлены  $T_n(x)$  будут ортогональными. Из (11.17) следует, что

$$d\varphi = -\frac{dx}{\sqrt{1-x^2}}, \quad x = 1 \text{ при } \varphi = 0 \quad \text{и} \quad x = -1 \text{ при } \varphi = \pi.$$

Принимая теперь во внимание (11.18), находим, что при  $m \neq n$

$$0 = \int_0^\pi \cos m\varphi \cos n\varphi d\varphi = \int_{-1}^1 \frac{1}{\sqrt{1-x^2}} T_m(x) T_n(x) dx.$$

Тем самым

$$\rho(x) = (1-x^2)^{-1/2}. \quad (11.19)$$

### 11.3 Свойства многочленов Чебышева

1°. При четном  $n$  многочлен  $T_n(x)$  является четной функцией  $x$ , а при нечетном  $n$  — нечетной.

**Доказательство** следует из (11.19) и леммы 11.4.

2°. Коэффициент при старшей степени многочлена  $T_n(x)$  для  $n \geq 1$  равен  $2^{n-1}$ , т.е.  $\mu_n = 2^{n-1}$ , а

$$T_n(x) = 2^{n-1}x^n + \dots$$

**Доказательство.** См. рекуррентную формулу (11.11).

3°. Нули многочлена  $T_n(x)$  расположены в точках

$$x_k = \cos \frac{(2k-1)\pi}{2n}, \quad k = 1, 2, \dots, n. \quad (11.20)$$

**Доказательство.** Из (11.18) находим, что

$$n \arccos x_k = -\frac{\pi}{2} + k\pi = \frac{(2k-1)\pi}{2}$$

или

$$\arccos x_k = \frac{(2k-1)\pi}{2n},$$

т.е.

$$x_k = \cos \frac{(2k-1)\pi}{2n}, \quad k = 1, 2, \dots, n.$$

Так как функции  $T_n(x)$  являются либо четными, либо нечетными, то нули  $T_n(x)$  расположены симметрично относительно начала координат

$$x_{n+1-k} = -x_k = -\cos \frac{(2k-1)\pi}{2n}.$$

4°.  $\max_{[-1,1]} |T_n(x)| = 1$ , причем

$$T_n(x_m) = (-1)^m,$$

где

$$x_m = \cos \frac{m\pi}{n}, \quad m = 0, \dots, n. \quad (11.21)$$

**Доказательство** очевидно.

5°. Среди всех многочленов степени  $n$  с единичным коэффициентом при старшей степени многочлен

$$\bar{T}_n(x) = \frac{1}{2^{n-1}} T_n(x), \quad n \geq 1$$

на  $[-1, 1]$  имеет наименьшее значение максимума модуля.

**Доказательство.** Допустим противное, т.е. допустим существование такого многочлена  $\bar{P}_n(x) = x^n + \dots$ , что

$$\max_{[-1,1]} |\bar{P}_n(x)| < \max_{[-1,1]} |\bar{T}_n(x)|. \quad (11.22)$$

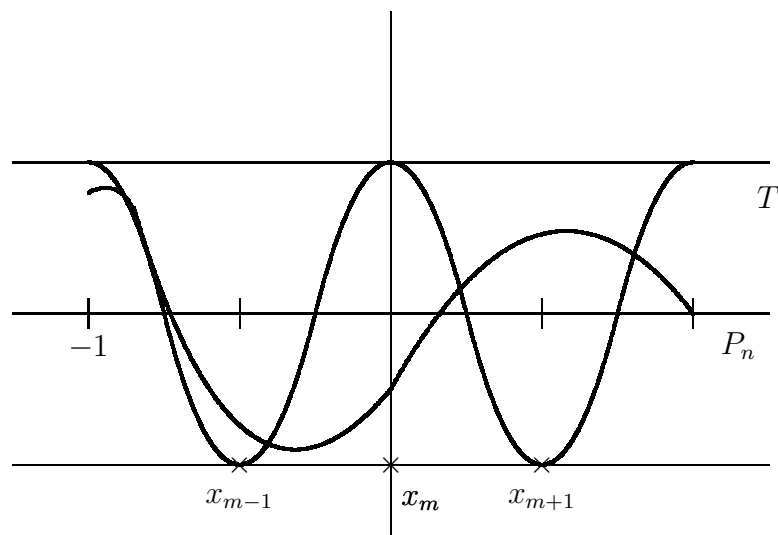


Рис. 1

Тогда  $\bar{T}_n(x) - \bar{P}_n(x) \neq 0$  и это есть многочлен степени не выше  $(n - 1)$ . Более того, в  $(n + 1)$  точке (11.21) этот многочлен принимает отличные от нуля значения с чередующимися знаками.

Но это означает, что алгебраический многочлен  $\bar{T}_n(x) - \bar{P}_n(x)$  степени меньшей  $n$  обращается в нуль по крайней мере в  $n$  точках, что невозможно.

**Замечание 11.3.** Можно доказать, что если  $\bar{P}_n(x) = x^n + \dots$ ,  $n \geq 1$ , и

$$\max_{[-1,1]} |\bar{P}_n(x)| = 2^{-n+1},$$

то  $\bar{P}_n(x) \equiv \bar{T}_n(x) = 2^{-n-1}T_n(x)$ .

Благодаря свойству 5° многочлены Чебышева  $T_n(x)$  называются многочленами, наименее уклоняющимися от нуля.

6°. Если  $x > 1$ , то

$$T_n(x) = \operatorname{ch} n \operatorname{Arch} x,$$

где

$$\operatorname{Arch} x = \ln(x + \sqrt{x^2 - 1}).$$

**Доказательство.** В силу (11.16)

$$\begin{aligned} T_n(x) &= \frac{q^n + q^{-n}}{2} = \frac{e^{n \ln q} + e^{-n \ln q}}{2} = \\ &= \operatorname{ch} n \ln q = \operatorname{ch} n \ln(x + \sqrt{x^2 - 1}). \end{aligned}$$

**Замечание 11.4.**  $\text{Arch } x$  — обратная функция к  $\text{ch } x$ .

**Упражнение 11.1.** Доказать, что

$$\begin{aligned}\text{ch Arch } x &= x, \\ \text{Arch ch } x &= x.\end{aligned}$$

## 11.4 Многочлены Лежандра

Многочленами Лежандра называются многочлены, которые ортогональны друг другу на  $[-1, 1]$  с весом  $\rho(x) \equiv 1$ . Обозначаются они через  $P_n(x)$

$$\int_{-1}^1 P_m(x)P_n(x)dx = 0, \quad m \neq n.$$

Если  $P_0(x) \equiv 1$ , то  $P_1(x) \equiv x$ .

Трехточечное рекуррентное соотношение для многочленов Лежандра имеет вид

$$(n+1)P_{n+1}(x) - (2n+1)xP_n(x) + nP_{n-1}(x) = 0$$

и следовательно

$$\begin{aligned}P_2(x) &= \frac{1}{2}(3x^2 - 1), & P_3(x) &= \frac{1}{2}(5x^3 - 3x), \\ P_4(x) &= \frac{1}{8}(35x^4 - 30x^2 + 3), \\ P_5(x) &= \frac{1}{8}(63x^5 - 70x^3 + 15x)\end{aligned}$$

и т.д. Общий вид  $P_n(x)$  задается формулой Родрига

$$P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n.$$

# 12

## Численное дифференцирование

### 12.1 Введение

Численное дифференцирование применяется, если функция задана таблицей или если ее трудно продифференцировать аналитически. Допустим, что в окрестности некоторой точки  $x$  у функции  $f(x)$  существует производная. По определению

$$f'(x) = \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}.$$

Если отказаться от предельного перехода, то можно положить

$$f'(x) \approx \frac{f(x + \Delta x) - f(x)}{\Delta x}. \quad (12.1)$$

Это и есть простейшая формула численного дифференцирования. Оценим ее погрешность в предположении, что значения функции  $f(x)$  вычисляются точно, и она дважды непрерывно дифференцируема. Используя формулу Тейлора, находим, что

$$\begin{aligned} \frac{f(x + \Delta x) - f(x)}{\Delta x} &= \frac{f(x) + \Delta x f'(x) + \frac{(\Delta x)^2}{2} f''(\xi) - f(x)}{\Delta x} = \\ &= f'(x) + \frac{\Delta x}{2} f''(\xi), \quad \xi \in (x, x + \Delta x). \end{aligned} \quad (12.2)$$

Отсюда заключаем, что формула (12.1) для функции  $f(x) \in C^2$  имеет погрешность первого порядка малости относительно  $\Delta x$ .

Однако, в силу формулы конечных приращений Лагранжа,

$$\frac{f(x + \Delta x) - f(x)}{\Delta x} = f'(\xi), \quad \xi \in (x, x + \Delta x), \quad (12.3)$$

т.е. в промежутке между  $x$  и  $x + \Delta x$  существует такая точка, что отношение из левой части (12.3), которое будем называть разностным отношением, совпадает со значением производной в этой точке. Конечно, положение  $\xi$  зависит от функции  $f$  и, вообще говоря, не известно. Тем не менее, если функция  $f(x)$  является многочленом второй степени  $P_2(x)$ , то точка  $\xi$  из (12.3) расположена точно посередине между  $x$  и  $x + \Delta x$ , т.е.  $\xi = x + \Delta x/2$ . В самом деле, разностное отношение постоянной равно нулю как и производная, разностное отношение линейной функции совпадает с ее производной в любой точке, а

$$\frac{(x + \Delta x)^2 - x^2}{\Delta x} = 2(x + \Delta x/2).$$

Это обстоятельство наводит на мысль, что и для других функций разностное отношение из (12.3) будет лучше аппроксимировать производную в точке  $x + \Delta x/2$ , чем в точке  $x$ . В самом деле, раскладывая значения  $f(x + \Delta x)$  и  $f(x)$  по формуле Тейлора в точке  $x + \Delta x/2$ , найдем, что

$$\begin{aligned} f(x + \Delta x) = & f(x + \Delta x/2) + \frac{\Delta x}{2} f'(x + \Delta x/2) + \frac{(\Delta x)^2}{8} f''(x + \Delta x/2) + \\ & + \frac{1}{3!} \left( \frac{\Delta x}{2} \right)^3 f'''(x + \theta_1 \Delta x), \end{aligned}$$

а

$$\begin{aligned} f(x) = & f(x + \Delta x/2) - \frac{\Delta x}{2} f'(x + \Delta x/2) + \frac{(\Delta x)^2}{8} f''(x + \Delta x/2) - \\ & - \frac{1}{3!} \left( \frac{\Delta x}{2} \right)^3 f'''(x + \theta_2 \Delta x) \end{aligned}$$

и, следовательно,

$$\frac{f(x + \Delta x) - f(x)}{\Delta x} = f'(x + \Delta x/2) + \frac{(\Delta x)^2}{4!} f'''(\eta).$$

Тем самым, для функций из  $C^3$  формула

$$f'(x + \Delta x/2) \approx \frac{f(x + \Delta x) - f(x)}{\Delta x} \quad (12.4)$$

имеет погрешность второго порядка малости относительно  $\Delta x$ .

Пусть  $x_i = x_0 + ih$ , где  $i \in \mathbb{Z}$ , а  $h > 0$  — шаг сетки. Тогда, полагая в (12.2)  $x = x_i$ , а  $\Delta x = h$ , получим

$$\frac{f(x_{i+1}) - f(x_i)}{h} = f'(x_i) + \frac{h}{2} f''(\xi_i). \quad (12.5)$$



Если же в (12.2) положить  $\Delta x = -h$  и снова  $x = x_i$ , то

$$\frac{f(x_{i-1}) - f(x_i)}{-h} = \frac{f(x_i) - f(x_{i-1})}{h} = f'(x_i) - \frac{h}{2}f''(\tilde{\xi}_i). \quad (12.6)$$

Из (12.5), (12.6) вытекает, что и приближенная формула

$$f'(x_i) \approx \frac{f(x_{i+1}) - f(x_i)}{h} \quad (12.7)$$

и приближенная формула

$$f'(x_i) \approx \frac{f(x_i) - f(x_{i-1})}{h} \quad (12.8)$$

являются формулами первого порядка точности. Однако их погрешности, вообще говоря, имеют разные знаки. Поэтому есть надежда, что у полусуммы правых частей (12.7), (12.8) погрешность будет иметь бóльший порядок малости относительно  $h$  (при бóльшей гладкости). В самом деле, используя формулу Тейлора, находим, что

$$\begin{aligned} & \frac{1}{2} \left[ \frac{f(x_{i+1}) - f(x_i)}{h} + \frac{f(x_i) - f(x_{i-1})}{h} \right] = \frac{f(x_{i+1}) - f(x_{i-1})}{2h} = \\ & = \left[ f_i + hf'_i + \frac{h^2}{2}f''_i + \frac{h^3}{6}f'''(\bar{\xi}_i) - \left( f_i - hf'_i + \frac{h^2}{2}f''_i - \frac{h^3}{6}f'''(\bar{\xi}_i) \right) \right] / 2h = \\ & = f'_i + \frac{h^2}{6} \frac{f'''(\bar{\xi}_i) + f'''(\bar{\xi}_i)}{2} = f'_i + \frac{h^2}{6}f'''(\xi_i), \quad \xi_i \in (\bar{\xi}_i, \bar{\xi}_i) \subset (x_{i-1}, x_{i+1}). \end{aligned} \quad (12.9)$$

(Мы здесь для сокращения письма использовали обозначение  $f_i := f(x_i)$ .)

Отсюда заключаем, что для  $f(x) \in C^3$  формула

$$f'(x_i) \approx \frac{f(x_{i+1}) - f(x_{i-1})}{2h} \quad (12.10)$$

(ср. с (12.4)) имеет погрешность  $O(h^2)$ .

Теперь вычтем из (12.5) соотношение (12.6)

$$\frac{f(x_{i+1}) - 2f(x_i) + f(x_{i-1}))}{h} = h \frac{f''(\xi_i) + f''(\tilde{\xi}_i)}{2} = f''(\tilde{\xi}_i)h.$$

Следовательно,

$$\frac{f_{i+1} - 2f_i + f_{i-1}}{h^2} = f''(\tilde{\xi}_i), \quad \tilde{\xi}_i \in (x_{i-1}, x_{i+1}), \quad (12.11)$$

т.е. левая часть этого соотношения аппроксимирует вторую производную функции  $f(x)$ . Исследуем погрешность этой аппроксимации в точке  $x_i$ :

$$\begin{aligned} \frac{f_{i+1} - 2f_i + f_{i-1}}{h^2} &= \frac{1}{h^2} \left[ f_i + hf'_i + \frac{h^2}{2!}f''_i + \frac{h^3}{3!}f'''_i + \frac{h^4}{4!}f^{IV}(\bar{\xi}_i) - \right. \\ &\quad \left. - 2f_i + f_i - hf'_i + \frac{h^2}{2!}f''_i - \frac{h^3}{3!}f'''_i + \frac{h^4}{4!}f^{IV}(\bar{\xi}_i) \right] = \\ &= f''_i + \frac{h^2}{12} \frac{f^{IV}(\bar{\xi}_i) + f^{IV}(\bar{\bar{\xi}}_i)}{2} = f''(x_i) + \frac{h^2}{12}f^{IV}(\xi_i). \end{aligned} \quad (12.12)$$

Отсюда следует, что левая часть соотношения (12.12) аппроксимирует вторую производную функции  $f(x) \in C^4$  с погрешностью  $O(h^2)$ .

**Замечание 12.1.** Мы уже трижды (в (12.9), (12.12) и в формуле после соотношения (12.10)) воспользовались утверждением о том, что для непрерывной функции  $0.5(f(x)+f(y)) = f(z)$ , где  $z \in (x, y)$ . Докажем это утверждение в более общем виде. Пусть  $f(x) \in C[a, b]$ ,  $m = \min_{[a,b]} f(x)$ ,  $M = \max_{[a,b]} f(x)$ ,  $x_1, x_2 \in [a, b]$ ,  $\alpha > 0$ ,  $\beta > 0$ . Тогда

$$\eta = \frac{\alpha f(x_1) + \beta f(x_2)}{\alpha + \beta} = f(x_3), \quad x_3 \in [a, b].$$

В самом деле,

$$m \leq \frac{\alpha f(x_1) + \beta f(x_2)}{\alpha + \beta} = \eta \leq M.$$

и по теореме о промежуточных значениях  $\eta = f(x_3)$ .

Введем следующие обозначения

$$f_{x,i} := \frac{f_{i+1} - f_i}{h}, \quad f_{\bar{x},i} := \frac{f_i - f_{i-1}}{h}, \quad f_{\bar{x},i}^{\circ} := \frac{1}{2}(f_{x,i} + f_{\bar{x},i}).$$

Тогда

$$f_{\bar{x}x,i} := \frac{f_{i+1} - 2f_i + f_{i-1}}{h^2}.$$

Отсюда и из (12.5), (12.6), (12.9), (12.12) находим, что

$$\begin{aligned} f_{x,i} &= f'_i + O(h), \\ f_{\bar{x},i} &= f'_i + O(h), \\ f_{\bar{x},i}^{\circ} &= f'_i + O(h^2), \\ f_{\bar{x}x,i} &= f''_i + O(h^2). \end{aligned} \quad (12.13)$$

## 12.2 Метод неопределенных коэффициентов

Рассмотренные простейшие формулы численного дифференцирования были построены из тех или иных эвристических соображений. Существуют и регулярные способы построения формул численного дифференцирования. Один из них — метод неопределенных коэффициентов.

Будем строить формулу для приближенного вычисления  $k$ -ой производной в следующем виде

$$f^{(k)}(x) \approx \sum_{j=0}^n c_j f(x_j), \quad k \leq n \quad (12.14)$$

и выберем  $c_j$  из тех условий, чтобы формула была точна на многочленах некоторой степени. Рассмотрим

**Пример 12.1.** Пусть

$$f'(h) \approx c_0 f(0) + c_1 f(h) + c_2 f(2h). \quad (12.15)$$

Потребуем, чтобы формула была точна на линейных функциях. Подставляя в (12.15)  $f(x) \equiv 1$  и  $f(x) \equiv x$  и требуя выполнения точного равенства, будем иметь

$$\begin{aligned} 0 &= c_0 + c_1 + c_2, \\ 1 &= c_1 h + c_2 2h. \end{aligned}$$

Принимая  $c_0$  за параметр, для  $c_1$  и  $c_2$  получим систему

$$\begin{aligned} c_1 + c_2 &= -c_0, \\ hc_1 + 2hc_2 &= 1. \end{aligned} \quad (12.16)$$

Определитель этой системы равен  $h$  и поэтому

$$c_1 = \begin{vmatrix} -c_0 & 1 \\ 1 & 2h \end{vmatrix} / h = (-2hc_0 - 1)/h, \quad c_2 = \begin{vmatrix} 1 & -c_0 \\ h & 1 \end{vmatrix} / h = (1 + c_0 h)/h. \quad (12.17)$$

Мы построили однопараметрическое семейство трехточечных формул численного нахождения первой производной

$$f'(h) \approx c_0 f(0) - \frac{1 + 2c_0 h}{h} f(h) + \frac{1 + c_0 h}{h} f(2h).$$

При  $c_0 = 0$  имеем

$$f'(h) \approx \frac{f(2h) - f(h)}{h} = f_x(h).$$

При  $c_0 = -1/h$

$$f'(h) \approx \frac{f(h) - f(0)}{h} = f_{\bar{x}}(h).$$

Это уже известные нам формулы.

Потребуем теперь, чтобы (12.15) была точна на многочленах второй степени. Тогда к уравнениям (12.16) добавится еще одно уравнение

$$2h = c_1 h^2 + c_2 4h^2.$$

Подставляя сюда  $c_1$  и  $c_2$  из (12.17), получим

$$-(1 + 2c_0 h)h + 4(1 + c_0 h)h = 2h,$$

откуда находим  $c_0 = -1/2h$ . Подставляя это значение в (12.17), будем иметь  $c_1 = 0$ ,  $c_2 = 1/2h$ , и поэтому

$$f'(h) \approx \frac{f(2h) - f(0)}{2h} \equiv f'_x(h).$$

И эта формула нам тоже уже известна.

Построим теперь новую формулу.

**Пример 12.2.** Пусть (ср. с (12.15))

$$f'(2h) \approx c_0 f(0) + c_1 f(h) + c_2 f(2h). \quad (12.18)$$

Будем требовать, чтобы формула (12.18) была точна на многочленах второй степени. Подставляя в (12.18) последовательно  $f(x) \equiv 1$ ,  $f(x) \equiv x$  и  $f(x) \equiv x^2$  и требуя выполнения точного равенства, получим систему

$$\begin{aligned} c_0 + c_1 + c_2 &= 0, \\ hc_1 + 2hc_2 &= 1, \\ h^2 c_1 + 4h^2 c_2 &= 4h. \end{aligned} \quad (12.19)$$

Первые два уравнения (12.19) совпадают с (12.16). Поэтому  $c_1$  и  $c_2$  выражаются через  $c_0$  при помощи (12.17). Подставляя (12.17) в последнее уравнение (12.19), находим, что

$$-(2hc_0 + 1)h + 4h(1 + hc_0) = 4h$$

и, следовательно,  $c_0 = 1/2h$ , а с учетом (12.17)

$$c_1 = -2/h, \quad c_2 = 3/(2h).$$

Тем самым,

$$f'(2h) \approx \frac{f_0 - 4f_1 + 3f_2}{2h} = \frac{f_2 - f_1}{h} + \frac{f_0 - 2f_1 + f_2}{2h} = f_{\bar{x},2} + \frac{h}{2}f_{\bar{x}\bar{x},2}. \quad (12.20)$$

Эта новая формула для вычисления первой производной.

**Упражнение 12.1.** Показать, что для  $f(x) \in C^3$  формула (12.20) имеет погрешность  $O(h^2)$ . Сравнить погрешность этой формулы с погрешностью из (12.9).

### 12.3 Использование интерполяционных формул

Наиболее универсальный способ построения формул численного дифференцирования основан на использовании интерполяционных формул.

Многочлен  $n$ -ой степени  $L_n(x)$  называется интерполяционным многочленом (интерполянт) Лагранжа функции  $f(x)$ , если

$$L_n(x_k) = f(x_k), \quad k = 0, 1, \dots, n,$$

где  $x_k$ ,  $k = 0, 1, \dots, n$  — узлы интерполяции. Если функция  $f(x)$  такова, что  $f(x_k) = 0$  при  $k \neq i$ , а  $f(x_i) = 1$ , то ее интерполянт по этим узлам является многочлен

$$\begin{aligned} & \frac{(x - x_0)(x - x_1) \dots (x - x_{i-1})(x - x_{i+1}) \dots (x - x_n)}{(x_i - x_0)(x_i - x_1) \dots (x_i - x_{i-1})(x_i - x_{i+1}) \dots (x_i - x_n)} \equiv \\ & \equiv \prod_{\substack{k=0 \\ k \neq i}}^n \frac{x - x_k}{x_i - x_k} =: p_i(x). \end{aligned}$$

В общем же случае интерполяционный многочлен Лагранжа имеет вид

$$L_n(x) \equiv \sum_{i=0}^n f_i p_i(x),$$

причем

$$f(x) = L_n(x) + R_n(x),$$

где  $R_n(x)$  — остаточный член, для которого имеет место представление

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{k=0}^n (x - x_k), \quad \xi \in (x_0, x_n).$$

Полагая

$$f^{(m)}(x) \approx L_n^{(m)}(x), \quad 0 \leq m \leq n, \quad (12.21)$$

получим формулу численного дифференцирования.

**Пример 12.3.** Пусть  $n = 1$ ,  $x_1 = x_0 + h$ . Тогда

$$L_1(x) = f_0 \frac{x - x_1}{-h} + f_1 \frac{x - x_0}{h}, \quad L'_1(x) = \frac{f_0}{-h} + \frac{f_1}{h} = \frac{f_1 - f_0}{h}.$$

Полагая здесь  $x = 0$ , получим формулу (12.7), полагая  $x = h$  — формулу (12.8).

**Пример 12.4.** Пусть теперь  $n = 2$ ,  $x_0 = 0$ ,  $x_1 = h$ ,  $x_2 = 2h$ . Тогда

$$L_2(x) = f_0 \frac{x - x_1}{x_0 - x_1} \frac{x - x_2}{x_0 - x_2} + f_1 \frac{x - x_0}{x_1 - x_0} \frac{x - x_2}{x_1 - x_2} + f_2 \frac{x - x_0}{x_2 - x_0} \frac{x - x_1}{x_2 - x_1},$$

$$L'_2(x) = f_0 \frac{2x - (x_2 + x_1)}{(-h)(-2h)} + f_1 \frac{2x - (x_2 + x_0)}{h(-h)} + f_2 \frac{2x - (x_0 + x_1)}{2h \cdot h}.$$

Отсюда

$$L'_2(x_2) = L'_2(2h) = \frac{f_0 - 4f_1 + 3f_2}{2h}, \quad (\text{ср. с (12.20)})$$

$$L'_2(x_1) = L'_2(h) = \frac{f_2 - f_0}{2h}, \quad (\text{ср. с (12.10)})$$

$$L'_2(x_0) = L'_2(0) = \frac{-3f_0 + 4f_1 - f_2}{2h}.$$

Последнее соотношение задает новую формулу; ее погрешность на функциях из  $C^3$  есть  $O(h^2)$ .

**Пример 12.5.** Пусть  $n = 2$ , а  $x_0 = 0$ ,  $x_1 = h_1$ ,  $x_2 = x_1 + h_2$ ,  $m = 2$ . Легко проверить, что

$$L''_2(x) = \frac{2}{-h_1(-h_1 - h_2)} f_0 + \frac{2}{h_1(-h_2)} f_1 + \frac{2}{(h_1 + h_2)h_2} f_2.$$

Если  $h_1 = h_2 = h$ , то

$$L''_2(x) = \frac{f_0 - 2f_1 + f_2}{h^2} = f_{\bar{x}\bar{x}}(h). \quad (12.22)$$

В противном случае

$$L''_2(x) = \frac{1}{h_1(h_1 + h_2)/2} (f_0 - f_1) + \frac{1}{h_2(h_1 + h_2)/2} (f_2 - f_1) =$$

$$= \left( \frac{f_2 - f_1}{h_2} - \frac{f_1 - f_0}{h_1} \right) \frac{1}{(h_1 + h_2)/2} = \frac{2}{h_1 + h_2} [f_x(x_1) - f_{\bar{x}}(x_1)]. \quad (12.23)$$

**Упражнение 12.2.** Доказать, что соотношение (12.23) при  $h_1 \neq h_2$  аппроксимирует производную  $f''(x_1)$  с погрешностью не выше  $O(h_1 + h_2)$ . При какой гладкости  $f(x)$  такая погрешность достигается?

На самом деле для построения формул численного дифференцирования удобнее использовать интерполяционный многочлен Лагранжа в форме Ньютона. Чтобы получить эту форму, запишем интерполяционный многочлен в виде

$$L_n(x) = L_0(x) + (L_1(x) - L_0(x)) + (L_2(x) - L_1(x)) + \dots + (L_n(x) - L_{n-1}(x)),$$

где  $L_0(x) = \text{const} = A_0$ . Многочлен  $L_k(x) - L_{k-1}(x)$  имеет степень  $k$  и обращается в нуль при  $x = x_0, x_1, \dots, x_{k-1}$ . Поэтому

$$L_k(x) - L_{k-1}(x) = A_k \prod_{j=0}^{k-1} (x - x_j) = A_k \omega_{k-1}(x),$$

а

$$L_n(x) = A_0 + \sum_{k=1}^n A_k \omega_{k-1}(x). \quad (12.24)$$

Для упрощения дальнейших рассуждений будем предполагать, что узлы интерполяции являются равноотстоящими, т.е.  $\Delta x_k := x_{k+1} - x_k = h$ . Пусть, кроме того,  $\Delta^j := \Delta(\Delta^{j-1})$ .

Найдем коэффициенты  $A_k$  из представления (12.24). Для этого последовательно положим  $x = x_0$ ,  $x = x_1$  и т.д. и примем во внимание, что  $L_n(x_j) = f(x_j)$ . Поскольку

$$\begin{aligned} \omega_{k-1}(x_0) &= 0, & k &= 1, 2, \dots, n, \\ \omega_{k-1}(x_1) &= 0, & k &= 2, \dots, n \text{ и т.д.,} \end{aligned}$$

то

$$L_n(x_0) = A_0$$

и, следовательно,

$$A_0 = f(x_0) = f_0.$$

Далее,

$$L_n(x_1) = f_0 + A_1(x_1 - x_0) = f_0 + hA_1$$

и, следовательно,

$$A = \frac{f_1 - f_0}{h} = \frac{\Delta f_0}{h}.$$

Аналогично,

$$L_n(x_2) = f_0 + \frac{\Delta f_0}{h}(x_2 - x_0) + A_2(x_2 - x_0)(x_1 - x_0) = f_0 + 2\Delta f_0 + 2h^2 A_2.$$

Поэтому

$$A_2 = \frac{f_2 - f_0 - 2\Delta f_0}{2h^2} = \frac{f_2 - 2f_1 + f_0}{2h^2} = \frac{\Delta^2 f_0}{2h^2}.$$

Индукцией можно показать, что

$$A_k = \frac{\Delta^k f_0}{k!h^k}.$$

Поэтому

$$\begin{aligned} L_n(x) = f_0 + \frac{\Delta f_0}{h}(x - x_0) + \frac{\Delta^2 f_0}{2h^2}(x - x_0)(x - x_1) + \dots + \\ + \frac{\Delta^n f_0}{n!h^n}(x - x_0) \dots (x - x_{n-1}). \end{aligned} \quad (12.25)$$

Это представление и называется *интерполяционным многочленом Ньютона для равных промежутков*. (Ср. с формулой Тейлора для функции  $f(x)$ .)

Интерполяционный многочлен в форме Ньютона может быть записан и для неравных промежутков, т.е. когда, вообще говоря,  $x_{k+1} - x_k \neq x_k - x_{k-1}$ , однако в этом случае его структура будет много сложнее.

Формулы численного дифференцирования и теперь определяются соотношением (12.21), однако при использовании (12.25) мы почти сразу получаем явное представление. Например, при  $n = m = 2$  находим, что

$$f''(x) \approx \frac{\Delta^2 f_0}{h^2} = f_{\bar{x}x}(h),$$

что совпадает с формулой (12.22)

Для построения формул численного дифференцирования можно использовать не только интерполяционные многочлены Лагранжа, но и интерполяционные многочлены Эрмита. Такие формулы полезны, когда в узлах заданы не только значения функции, но и значения производных, а производные нужно знать в других точках. Формула численного дифференцирования и в этом случае выглядит аналогично (12.21). Если

$$f(x) = H_n(x) + R_n(x),$$

где  $H_n(x)$  — интерполяционный многочлен Эрмита, а  $R_n(x)$  — остаточный член, то

$$f^{(m)}(x) \approx H_n^{(m)}(x), \quad 1 \leq m \leq n. \quad (12.26)$$



**Пример 12.6.** Пусть интерполяционный многочлен Эрмита имеет степень три и написан по значениям функции и ее первой производной в двух узлах  $x_0 = 0$  и  $x_1 = h$ , т.е.

$$H_3(x) = p_{00}(x)f_0 + p_{01}(x)f'_0 + p_{10}(x)f_1 + p_{11}(x)f'_1,$$

где

$$\begin{aligned} p_{00}(x) &= \frac{(2x+h)(x-h)^2}{h^3}, & p_{01}(x) &= \frac{x(x-h)^2}{h^2}, \\ p_{10}(x) &= \frac{x^2(3h-2x)}{h^3}, & p_{11}(x) &= \frac{x^2(x-h)}{h^2}. \end{aligned}$$

Тогда формула для приближенного нахождения первой производной в точке  $x$  примет вид

$$H'_3(x) = p'_{00}(x)f_0 + p'_{01}(x)f'_0 + p'_{10}(x)f_1 + p'_{11}(x)f'_1.$$

Если принять во внимание, что

$$\begin{aligned} p'_{00}(x) &= 6x(x-h)/h^3, & p'_{01}(x) &= \frac{3x^2 - 4hx + h^2}{h^2}, \\ p'_{10}(x) &= -6x(x-h)/h^3, & p'_{11}(x) &= \frac{3x^2 - 2hx}{h^2}, \end{aligned}$$

то, например,  $H'_3(0) = f'_0$ , а

$$H'_3(h/2) = \frac{3}{2} \frac{f_1 - f_0}{h} - \frac{f'_0 + f'_1}{4}.$$

Далее, для приближенного вычисления второй производной в точке  $x$  имеем формулу

$$H''_3(x) = p''_{00}(x)f_0 + p''_{01}(x)f'_0 + p''_{10}(x)f_1 + p''_{11}(x)f'_1.$$

Принимая во внимание, что

$$\begin{aligned} p''_{00}(x) &= 6(2x-h)/h^3, & p''_{01}(x) &= 2(3x-2h)/h^2, \\ p''_{10}(x) &= 6(-2x+h)/h^3, & p''_{11}(x) &= 2(3x-h)/h^2, \end{aligned}$$

находим, например,

$$H''_3(0) = \frac{6(f_1 - f_0)}{h^2} - \frac{4}{h}f'_0 - \frac{2}{h}f'_1.$$

Элементарные вычисления показывают, что

$$6(f_1 - f_0)/h^2 - 4f'_0/h - 2f'_1/h = f''_0 - \frac{h^2}{12}f^{IV}(\xi).$$

## 12.4 О корректности численного дифференцирования

В формулах численного дифференцирования линейные комбинации значений функции  $f(x)$  в узлах  $x_i$  делятся на  $h^m$ , где  $m$  — порядок вычисляемой производной. Поскольку сами значения функции, как правило, задаются или вычисляются не точно, то при малых  $h$  неустранимые погрешности оказывают существенное влияние на точность численного дифференцирования.

Пусть  $\delta_i$  — величина погрешности, с которой вычисляется значение функции  $f(x)$  в узле  $x_i$ , т.е. вычисляемое приближенное значение есть

$$\tilde{f}_i = f_i + \delta_i.$$

Будем предполагать, что  $|\delta_i| \leq \delta$ .

Пусть для приближенного вычисления первой производной используется формула (12.10). Тогда, с учетом (12.9),

$$\begin{aligned} f'(x_i) &\approx \frac{\tilde{f}_{i+1} - \tilde{f}_{i-1}}{2h} = \frac{f_{i+1} + \delta_{i+1} - f_{i-1} - \delta_{i-1}}{2h} = \\ &= \frac{f_{i+1} - f_{i-1}}{2h} + \frac{\delta_{i+1} - \delta_{i-1}}{2h} = f'(x_i) + \frac{h^2}{6} f'''(\xi_i) + (\delta_{i+1} - \delta_{i-1})/(2h). \end{aligned}$$

Отсюда находим, что для полной погрешности этой формулы  $\varepsilon_1 = (\tilde{f}_{i+1} - \tilde{f}_{i-1})/(2h) - f'(x_i)$  справедлива оценка

$$|\varepsilon_1| \leq \frac{h^2}{6} M_3 + \delta/h, \quad (12.27)$$

где  $M_3 = \max_{x \in [x_{i+1}, x_{i-1}]} |f'''(x)|$ . Из этой оценки следует, что при уменьшении  $h$  полная погрешность убывает только до определенного предела, после чего начинает расти. Если, например,  $\delta$  сравнима с  $h$ , то *мы не можем найти приближенное значение производной*, ибо погрешность будет  $O(1)$ . Чтобы вычисленное значение можно было рассматривать как приближенное значение производной, нужно, чтобы  $h$  было много больше  $\delta$ . Наивысшую точность мы получим при том  $h$ , при котором правая часть (12.27) достигает минимума по  $h$ . Указанное значение

$$h = h_1 = \sqrt[3]{3\delta/M_3}.$$

При этом

$$\varepsilon_1 = \frac{3}{2} \left( \frac{M_3}{3} \right)^{1/3} \delta^{2/3}.$$

Если при тех же предположениях о точности вычисления значений  $f_i$  воспользоваться формулой из левой части (12.12), дающее приближенное значение второй производной, то полная погрешность  $\varepsilon_2 = (\tilde{f}_{i+1} - 2\tilde{f}_i + \tilde{f}_{i-1})/h^2 - f''(x_i)$  оценится так

$$|\varepsilon_2| \leq \frac{h^2}{12} M_4 + \frac{4\delta}{h^2},$$

где  $M_4 = \max_{x_{i-1} \leq x \leq x_{i+1}} |f^{IV}|$ . При этом оптимальное значение  $h = h_2 = 2(3\delta/M_4)^{1/4}$ , а  $\varepsilon_2 = 2\sqrt{M_4/3} \delta^{1/2}$ .

Следует, однако, заметить, что предельная точность при приближенном вычислении производных не всегда ниже, чем точность, с которой задана сама функция. Пусть, например,  $\tilde{f}_i = f_i + \delta v_i$ , где  $v_i$  — некоторая "гладкая" функция, т.е. такая, что, например,  $|v_{x,i}| \leq M$ . Тогда для формулы (12.10) полная погрешность будет оцениваться следующим образом:

$$|\varepsilon_1| \leq \frac{h^2}{6} M_3 + M\delta$$

и, если  $h/\sqrt{\delta} = O(1)$ , то  $|\varepsilon_1| = O(\delta)$ .

## Методы решения нелинейных уравнений

Пусть задана непрерывная функция  $f(x)$  действительной переменной  $x$ , и требуется найти ее нули, т.е. корни уравнения

$$f(x) = 0. \quad (13.1)$$

При такой формулировке задача весьма неопределенна, ибо корней может не быть вовсе, или их может быть бесконечно много. Обычно задача формулируется более конкретно с дополнительными указаниями. Например, отыскание корней на заданном интервале. Поскольку не существует регулярных методов отыскания точных значений корней уравнения (13.1), то речь должна идти об итерационных методах нахождения приближенного решения. (Только если  $f(x)$  представляет собой многочлен не выше 4-ой степени, имеются методы представления его нулей в виде радикалов.)

Чтобы воспользоваться тем или иным итерационным методом, нужно иметь начальное приближение к корню. Для этого нужно, по крайней мере, изучить расположение корней и выделить области, где имеется единственный корень. В противном случае мы должны с использованием того или иного итерационного процесса уточнить значения корней или найти их с требуемой точностью.

Способы локализации корней (выделение областей, где имеется единственный корень) многообразны, и указать универсальный метод не представляется возможным. Иногда отрезки локализации известны заранее, а иногда определяются из физических соображений. В простых ситуациях хороший результат может дать графический метод; широко применяют построение таблиц функции  $f(x)$  вида  $y_i = f(x_i)$ ,  $i = \overline{1, n}$  для обнаружения перемен знака.

### 13.1 Метод бисекции (метод деления отрезка пополам)

Пусть  $f(x) \in C[a, b]$  и  $f(a)f(b) < 0$ . Последнее означает, что на  $[a, b]$  имеется, по крайней мере, один корень уравнения (13.1). (Условие существования решения.) Предположим, что решение единственное, т.е.  $x^* \in (a, b)$  — единственный корень уравнения (13.1) на  $[a, b]$ . Положим  $a_0 = a$ ,  $b_0 = b$ , найдем середину отрезка  $[a_0, b_0]$

$$x_0 = \frac{a_0 + b_0}{2}$$

и примем эту величину за приближенное значение  $x^*$ . Так как положение корня  $x^*$  на отрезке  $[a_0, b_0]$  неизвестно, то можно лишь утверждать, что погрешность этого приближения не превосходит половины длины  $[a_0, b_0]$ :

$$|x_0 - x^*| \leq \frac{b_0 - a_0}{2}.$$

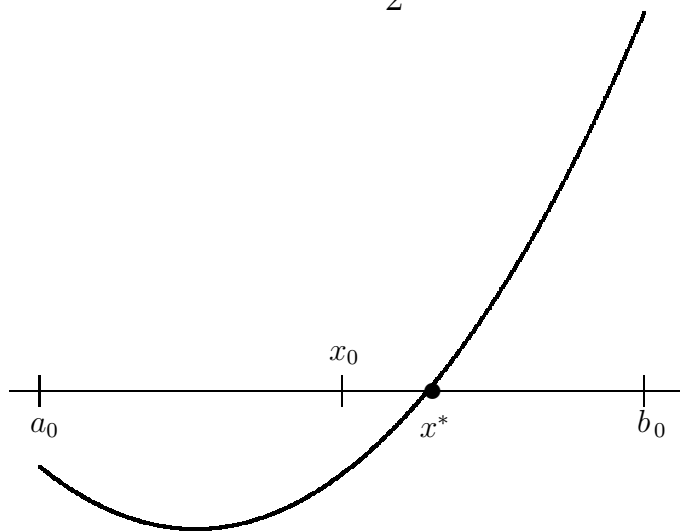


Рис. 1

Вычислим  $f(x_0)$ . Если  $f(x_0) = 0$ , то  $x^* = x_0$ , и вычисления на этом заканчиваются. Если  $f(x_0) \neq 0$ , то знак  $f(x_0)$  совпадает либо со знаком  $f(a_0)$ , либо со знаком  $f(b_0)$ . Пусть для определенности  $f(a_0) < 0$ ,  $f(b_0) > 0$ . Из двух отрезков  $[a_0, x_0]$  и  $[x_0, b_0]$  выберем тот, на концах которого  $f(x)$  принимает значения с противоположными знаками. Обозначим этот отрезок через  $[a_1, b_1]$ , где

$$a_1 = a_0, \quad b_1 = x_0 \quad \text{при} \quad f(x_0) > 0$$

и

$$a_1 = x_0, \quad b_1 = b_0 \quad \text{при} \quad f(x_0) < 0.$$

Отрезок  $[a_1, b_1]$  имеет вдвое меньшую длину, чем  $[a_0, b_0]$ ,  $f(a_1)f(b_1) < 0$  и  $x^* \in (a_1, b_1)$ , причем  $|x_0 - x^*| \leq \frac{b_0 - a_0}{2}$ . Найдем середину отрезка  $[a_1, b_1]$  и т.д. Пусть

$$x_k = \frac{a_k + b_k}{2}, \quad k = 1, 2, \dots,$$

и всегда

$$|x_k - x^*| \leq \frac{b_k - a_k}{2} = \frac{b - a}{2^{k+1}}.$$

Процесс деления отрезка пополам продолжается до тех пор, пока длина нового отрезка  $[a_k, b_k]$  не станет меньше  $2\varepsilon$ , где  $\varepsilon$  — требуемая точность в определении приближенного значения корня. Тогда

$$x_k = \tilde{x}, \quad |\tilde{x} - x^*| \leq \varepsilon,$$

т.е. изложенный метод позволяет найти приближенное решение с *гарантированной* точностью. Скорость сходимости метода не ниже скорости сходимости к нулю геометрической прогрессии со знаменателем  $1/2$ . Каждая итерация уменьшает погрешность не менее, чем в два раза.

**Пример 13.1.** Для того, чтобы уменьшить первоначальную локализацию в  $10^6$  раз, нужно сделать 20 итераций, ибо  $2^{20} = 1048576$ .

Метод деления отрезка пополам является *глобально* сходящимся итерационным методом. Какова бы ни была функция  $f(x) \in C[a, b]$  и каков бы ни был отрезок  $[a, b]$ , на котором имеется один нуль функции  $f(x)$ , итерационный процесс обязательно к этому нулю сойдется.

## 13.2 Метод простых итераций

Чтобы применить метод простых итераций для решения нелинейного уравнения (13.1), необходимо преобразовать это уравнение к следующему виду:

$$x = \varphi(x). \tag{13.2}$$

Это можно сделать многими различными способами, некоторые из которых будут изложены позже. Пусть, например,

$$\varphi(x) = x + \tau(x)f(x), \tag{13.3}$$

где  $\tau(x)$  — произвольная непрерывная знакоопределенная функция.

Выбирая некоторое начальное приближение  $x_0$ , построим итерационный процесс

$$x_{k+1} = \varphi(x_k), \quad k = 0, 1, \dots \quad (13.4)$$

Итерационный процесс (13.4) называется *методом простых итераций*.

**Теорема 13.1.** Пусть  $x^*$  — корень уравнения (13.2). Тогда, если  $|\varphi'(x)| \leq q < 1$  для  $x \in [x^* - \delta, x^* + \delta]$ , то при любом начальном приближении  $x_0 \in [x^* - \delta, x^* + \delta]$  метод простых итераций сходится со скоростью геометрической прогрессии, знаменателем которой является число  $q$ . При этом

$$|x_k - x^*| \leq q^k |x_0 - x^*|.$$

**Доказательство.** В силу (13.4)

$$x_k = \varphi(x_{k-1}),$$

а в силу (13.2)

$$x^* = \varphi(x^*).$$

Вычитая из первого соотношения второе и используя формулу конечных приращений Лагранжа, с учетом условий теоремы найдем, что

$$\begin{aligned} |x_k - x^*| &= |\varphi(x_{k-1}) - \varphi(x^*)| = |\varphi'(\xi_k)| |x_{k-1} - x^*| \leq \\ &\leq q |x_{k-1} - x^*| \leq q^k |x_0 - x^*|. \end{aligned}$$

Здесь  $\xi_k \in (x_{k-1}, x^*) \subset [x^* - \delta, x^* + \delta]$ . Теорема доказана.

**Определение 13.1.** Говорят, что последовательность  $x_k$ ,  $k = 0, 1, \dots$  сходится линейно при  $k \rightarrow \infty$ , если

$$|x_k| \leq q |x_{k-1}|, \quad q = \text{const} < 1,$$

т.е. если она сходится со скоростью геометрической прогрессии. Эта последовательность *сходится сверхлинейно*, если

$$|x_k| \leq q_k |x_{k-1}|, \quad \lim_{k \rightarrow \infty} q_k = 0.$$

Дадим геометрическую иллюстрацию итерационного процесса (13.4). Изобразим на плоскости  $Oxy$  прямую  $y = x$  и кривую  $y = \varphi(x)$ . Пусть сначала  $0 < \varphi'(x) < 1$ .

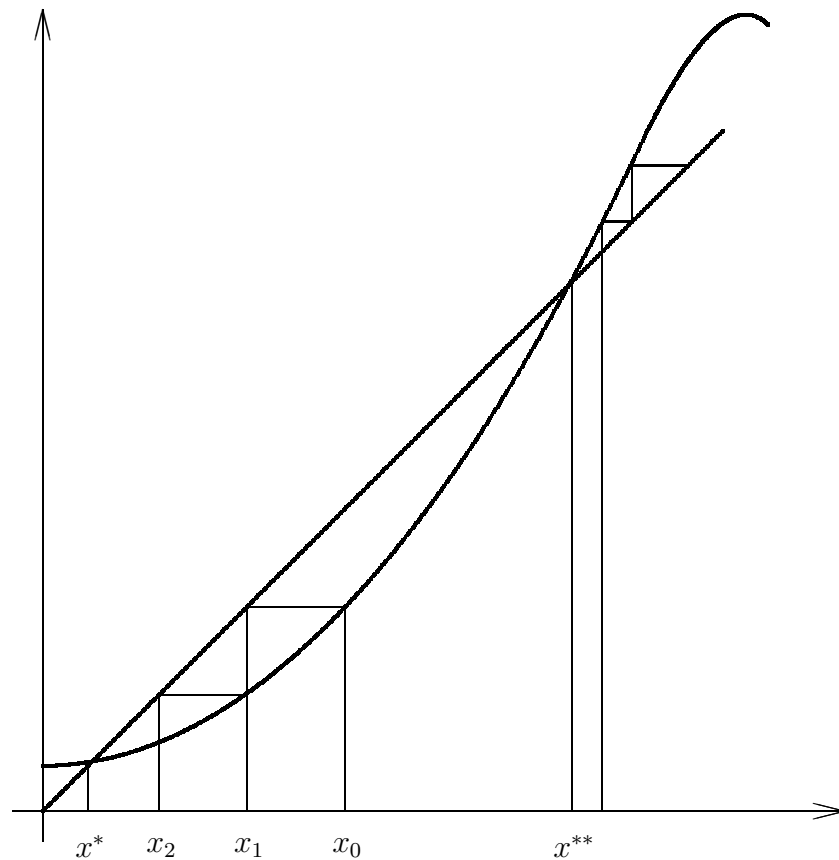


Рис. 2

Из рисунка 2 видно, что при  $0 < \varphi'(x) \leq q < 1$  последовательность  $x_k$  монотонно сходится к  $x^*$ , причем с той стороны, с которой расположено начальное приближение.

Если  $\varphi'(x^{**}) > 1$ , то итерации не сходятся к  $x^{**}$ .

При  $-1 < \varphi'(x) < 0$  приближения двусторонние (см. рис. 3). В этом случае по двум последовательным приближениям легко судить о достигнутой точности

$$|x_k - x^*| < |x_k - x_{k-1}|.$$

Можно также увидеть, что сходимость тем быстрее, чем меньше  $|\varphi'|$ . Если  $\varphi'(x^*) \ll 1$ , то процесс сходимости ускоряется по мере приближения к корню.



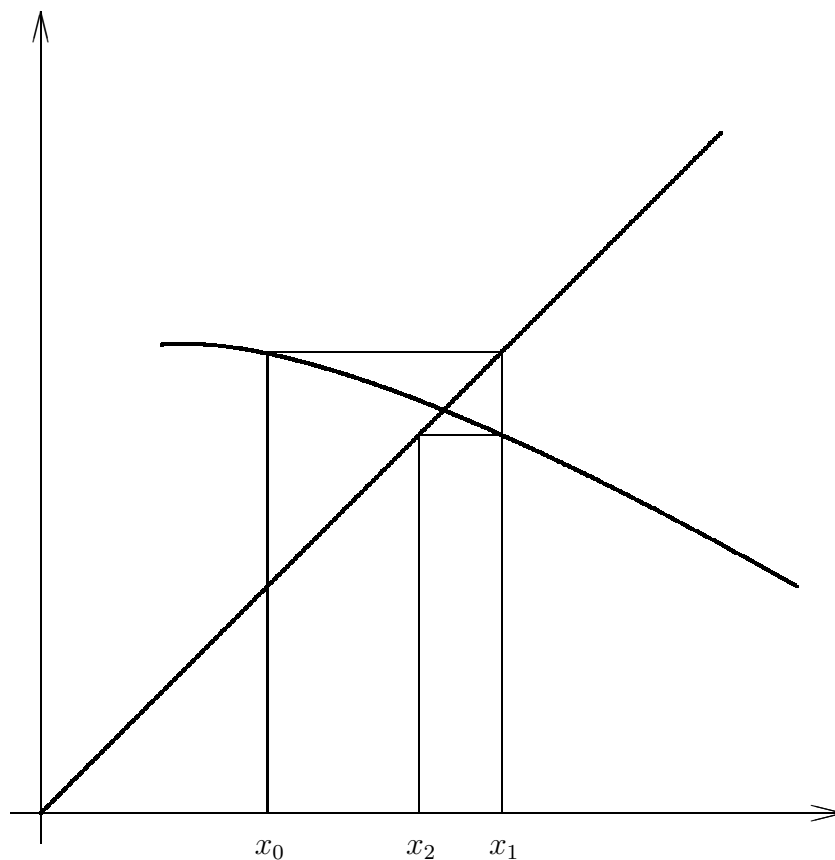


Рис. 3

### 13.3 Метод Ньютона

Пусть  $k$ -е приближение  $x_k$  к решению уравнения (13.1) найдено. Разложим функцию  $f(x)$  в точке  $x_k$  по формуле Тейлора

$$f(x) = f(x_k) + (x - x_k)f'(x_k) + \frac{(x - x_k)^2}{2}f''(\xi_k), \quad \xi_k \in (x, x_k) \quad (13.5)$$

или

$$\Delta f := f(x) - f(x_k) = df + O(d^2 f), \quad df := f'dx = f'(x_k)(x - x_k).$$

Заменим приближенно приращение функции ее дифференциалом

$$\Delta f \approx df$$

или

$$f(x) \approx f(x_k) + (x - x_k)f'(x_k) =: P_1(x).$$

Приравняем теперь функцию  $P_1(x)$ , являющуюся приближением к  $f(x)$ , нулю

$$P_1(x) = 0 : \quad f(x_k) + (x - x_k)f'(x_k) = 0$$

и найдем корень полученного уравнения

$$x - x_k = -\frac{f(x_k)}{f'(x_k)} \quad \Rightarrow \quad x = x_k - \frac{f(x_k)}{f'(x_k)}.$$

Этот корень и примем за новое приближение. Итак, алгоритм метода Ньютона следующий:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, \dots, \quad x_0 - \text{задано.} \quad (13.6)$$

*Геометрическая интерпретация метода Ньютона* такова. Как известно из математического анализа,  $f'(x_k)$  есть тангенс угла наклона касательной к кривой  $y = f(x)$  в точке  $x = x_k$ . Прямая

$$y = P_1(x) \quad (13.7)$$

имеет тот же наклон, что и касательная к  $f(x)$  в точке  $x_k$ . Более того, в точке  $x = x_k$  значения  $P_1(x)$  и  $f(x)$  совпадают, и, следовательно, (13.7) есть уравнение касательной к кривой  $y = f(x)$  в точке  $x = x_k$ . Тем самым,

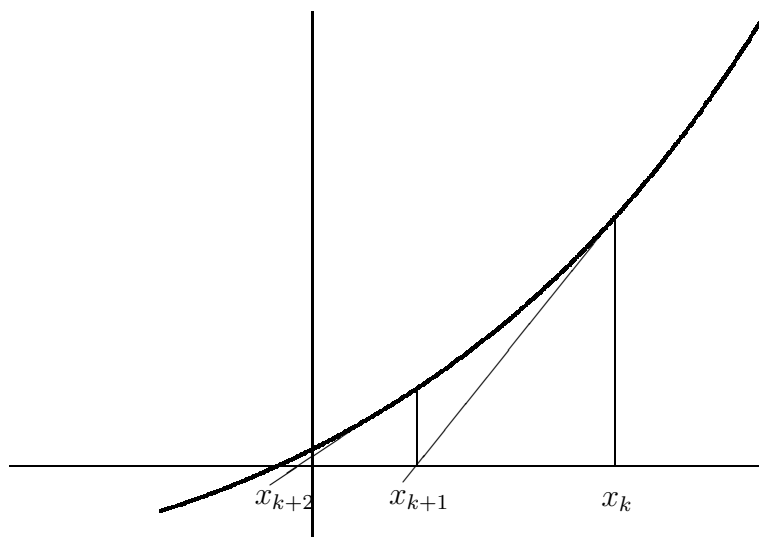


Рис. 4

в методе Ньютона на каждой итерации кривая  $y = f(x)$  заменяется касательной в точке  $x_k$ , и вместо уравнения  $f(x) = 0$  решается уравнение  $P_1(x) = 0$ .

**Связь метода Ньютона с методом простых итераций.** При применении метода простых итераций к решению уравнения (13.1) оно сначала преобразовывалось к виду (13.2), где функция  $\varphi(x)$  определялась соотношением (13.3), а итерации проводились по формуле (13.4). Сравнивая (13.4) и (13.6), находим, что

$$\varphi(x) = x - \frac{f(x)}{f'(x)}, \quad (13.8)$$

т.е.  $\tau(x)$  из (13.3) есть  $[-f'(x)]^{-1}$ . Как следует из теоремы 13.1, для сходимости метода простых итераций производная функции  $\varphi(x)$  в окрестности корня  $x^*$  должна быть по модулю меньше единицы. Из (13.8)

$$\varphi'(x) = 1 - \frac{(f'(x))^2 - f(x)f''(x)}{(f'(x))^2} = \frac{f(x)f''(x)}{(f'(x))^2}. \quad (13.9)$$

Если  $x^*$  — простой корень уравнения (13.1), то  $f'(x^*) \neq 0$ , а  $\varphi'(x^*) = 0$ , и существует окрестность  $x^*$ , где  $|\varphi'(x)| < 1$ . Поэтому метод Ньютона всегда сходится, если начальное условие выбрано удачно.

**Оценка скорости сходимости метода Ньютона.** Полагая в (13.5)  $x = x^*$ , получим

$$0 = f(x_k) + (x^* - x_k)f'(x_k) + \frac{(x^* - x_k)^2}{2}f''(\xi_k^*), \quad \xi_k^* \in (x^*, x_k).$$

В силу (13.6)

$$0 = (x_{k+1} - x_k)f'(x_k) + f(x_k). \quad (13.10)$$

Вычитая это соотношение из предыдущего, получим

$$0 = (x^* - x_{k+1})f'(x_k) + \frac{1}{2}(x^* - x_k)^2f''(\xi_k^*).$$

Отсюда

$$(x_{k+1} - x^*) = \frac{1}{2} \frac{f''(\xi_k^*)}{f'(x_k)} (x_k - x^*)^2. \quad (13.11)$$

Будем предполагать, что

$$|f'(x)| \geq m_1, \quad |f''(x)| \leq M_2. \quad (13.12)$$

Тогда

$$|x_{k+1} - x^*| \leq \frac{M_2}{2m_1} |x_k - x^*|^2. \quad (13.13)$$

Домножая левую и правую части этого неравенства на  $M_2/(2m_1)$ , получим, что

$$\beta_{k+1} = \frac{M_2}{2m_1} |x_{k+1} - x^*| \leq \left[ \frac{M_2}{2m_1} (x_k - x^*) \right]^2 = \beta_k^2,$$

т.е.

$$\beta_{k+1} \leq \beta_k^2, \quad (13.14)$$

где

$$\beta_k = \frac{M_2}{2m_1} |x_k - x^*|. \quad (13.15)$$

Из (13.14) имеем

$$\beta_1 \leq \beta_0^2, \quad \beta_2 \leq \beta_1^2 \leq \beta_0^4 = \beta_0^{2^2}, \quad \beta_3 \leq \beta_2^2 \leq \beta_0^{2^3}$$

и вообще

$$\beta_k^2 \leq \beta_0^{2^k}.$$

Принимая во внимание (13.15), находим, что

$$|x_k - x^*| \leq \frac{2m_1}{M_2} \left[ \frac{M_2}{2m_1} |x_0 - x^*| \right]^{2^k}. \quad (13.16)$$

Для сходимости достаточно, чтобы

$$\frac{M_2}{2m_1} |x_0 - x^*| \leq q < 1. \quad (13.17)$$

Итак, доказана

**Теорема 13.2.** Пусть  $f(x) \in C^2[x^* - \delta, x^* + \delta]$ , где  $x^*$  — простой корень уравнения (13.1), и при  $x \in [x^* - \delta, x^* + \delta]$  справедливы оценки (13.12). Тогда, если начальное приближение  $x_0 \in [x^* - \delta, x^* + \delta]$  таково, что справедливо (13.17), то метод Ньютона (13.6) сходится с квадратичной скоростью, и справедлива оценка (13.16).

**Пример 13.2.** Пусть требуется найти корень степени  $p$  из числа  $a > 0$ . Тогда

$$f(x) := x^p - a$$

$$x_{k+1} = x_k - \frac{x_k^p - a}{px_k^{p-1}} = \frac{p-1}{p}x_k + \frac{a}{px_k^{p-1}}. \quad (13.18)$$

При  $p = 2$  и  $a = 3$

$$x_{k+1} = \frac{x_k}{2} + \frac{3}{2x_k}.$$

Пусть  $x_0 = 2$ . Тогда

$$x_1 = \frac{7}{4} = 1.75, \quad x_2 = \frac{97}{56} \approx 1.7321, \quad x_3 = \frac{97}{112} + \frac{84}{97} \approx 1.73205080,$$

а  $x^* = 1.732050808\dots$ . Отсюда следует, что если в  $x_0$  один верный знак, то в  $x_1$  — два, в  $x_2$  — четыре и т.д. Грубо говоря, число верных знаков после каждой итерации удваивается.

**Замечание 13.1.** Сходимость метода Ньютона установлена при условии, что корень  $x^*$  является простым. Ну, а что будет, если корень окажется кратным? Чтобы ответить на этот вопрос, исследуем  $f f'' / (f')^2$  из (13.9). Если корень  $x^*$  имеет кратность  $p > 1$ , то

$$\begin{aligned} f(x) &= a(x - x^*)^p + O((x - x^*)^{p+1}) \\ f'(x) &= ap(x - x^*)^{p-1} + O((x - x^*)^p) \\ f''(x) &= ap(p-1)(x - x^*)^{p-2} + O((x - x^*)^{p-1}). \end{aligned}$$

Отсюда

$$\begin{aligned} \varphi'(x) &= \frac{f(x)f''(x)}{[f'(x)]^2} = \\ &= \frac{a^2 p(p-1)(x - x^*)^{2p-2} + O((x - x^*)^{2p-1})}{p^2 a^2 (x - x^*)^{2p-2} + O((x - x^*)^{2p-1})} = \frac{p-1}{p} + O(x - x^*) \end{aligned} \quad (13.19)$$

и в малой окрестности  $x^*$  имеем  $|\varphi'(x)| < 1$ . Тем самым, метод Ньютона будет сходиться и к кратному корню, но эта сходимость не будет квадратичной; она будет всего лишь линейной.

**Пример 13.3.** Пусть  $f(x) = x^2$ . Здесь  $x^* = 0$  есть двукратный корень. Метод Ньютона приводит к соотношению

$$x_{k+1} = x_k - \frac{x_k^2}{2x_k} = \frac{1}{2}x_k.$$

Эти соотношения представляют собой геометрическую прогрессию со знаменателем  $q = 1/2$ .

Возникает вопрос, а нельзя ли увеличить скорость сходимости к кратному корню? Ответ на этот вопрос положительный. Это можно сделать путем следующего обобщения метода Ньютона. Итерации будем вести по формуле

$$x_{k+1} = x_k - \tau \frac{f(x_k)}{f'(x_k)},$$

где значение параметра  $\tau$  определяется кратностью искомого корня. Найдем это значение. Имеем

$$\varphi(x) = x - \tau \frac{f(x)}{f'(x)}, \quad \varphi'(x) = 1 - \tau \frac{f'^2 - f f''}{f'^2} = (1 - \tau) + \tau \frac{f f''}{f'^2}.$$

Отсюда с учетом (13.19)

$$\varphi'(x) = (1 - \tau) + \tau \left[ \frac{p-1}{p} + O(x - x^*) \right] = 1 - \tau + \tau \frac{p-1}{p} + O(x - x^*).$$

Выберем  $\tau$  из условия, что  $\varphi'(x^*) = 0$ . Тогда  $\tau = p$ , и обобщенный метод Ньютона

$$x_{k+1} = x_k - p \frac{f(x_k)}{f'(x_k)},$$

где  $p$  — кратность искомого корня, обладает скоростью сходимости метода Ньютона к простому корню.

**Пример 13.4.** Пусть  $f(x) = x^2(x+1)$  и  $x^* = 0$  — двукратный корень. Обобщенный метод Ньютона принимает вид

$$x_{k+1} = \frac{x_k^2}{2 + 3x_k}.$$

Если  $x_0 = 1$ , то  $x_1 = \frac{1}{5} = 0.2$ ,  $x_2 = \frac{1}{65} = 0.015$ ,  $x_3 = 0.0000115$  и т.д.

Соотношения (13.13) и (13.16) установленные при доказательстве теоремы 13.2, представляют собой так называемые *априорные оценки* точности приближенного решения  $x_k$ . Эти оценки справедливы вне зависимости от того, проводилось ли реальное вычисление  $x_k$  или нет. Их ценность состоит в том, что они указывают скорость убывания погрешности при увеличении  $k$ . Для непосредственной же оценки точности получаемого приближения они непригодны, ибо и в левой, и в правой частях неравенств присутствует неизвестное точное решение  $x^*$ .

Наряду с априорными оценками можно установить и *апостериорные оценки*, воспользоваться которыми можно только *после* того, как мы найдем соответствующее приближение. Апостериорные оценки используются для непосредственной оценки точности полученного решения.

**Теорема 13.3.** В условиях теоремы 13.2

$$|x_{k+1} - x^*| \leq \frac{M_2}{2m_1} |x_{k+1} - x_k|^2.$$

**Доказательство.** Полагая в (13.5)  $x = x_{k+1}$ , найдем, что

$$f(x_{k+1}) = f(x_k) + (x_{k+1} - x_k)f'(x_k) + \frac{(x_{k+1} - x_k)^2}{2}f''(\xi_k).$$

Отсюда в силу (13.10) следует, что

$$|f(x_{k+1})| = \frac{(x_{k+1} - x_k)^2}{2}f''(\xi_k) \leq \frac{(x_{k+1} - x_k)^2}{2}M_2.$$

С другой стороны, на основании формулы конечных приращений и (13.1)

$$f(x_{k+1}) - f(x^*) = f'(\eta_{k+1})(x_{k+1} - x^*) = f(x_{k+1}).$$

Отсюда и из первого соотношения вытекает искомая оценка. Теорема доказана.

**Упражнение 13.1.** Доказать, что, если  $f(x^*) = 0$ ,  $f'(x^*) \neq 0$ ,  $f''(x^*) = 0$ ,  $f'''(x^*) \neq 0$ , то сходимость метода Ньютона будет кубической.

## 13.4 Метод секущих

Основным достоинством метода Ньютона, которое делает его очень привлекательным, является высокая скорость сходимости. К недостаткам следует отнести необходимость вычисления на каждом шаге итераций производной. Вторым недостатком — сильная зависимость результативности метода от начального приближения: если начальное приближение оказалось неудачным (недостаточно близким к искомому решению), метод просто расходится. Первый недостаток в какой-то мере может быть преодолен путем замены производной разностным отношением. Именно, заменим в (13.5) производную  $f'(x_k)$  на разностное отношение

$$\frac{\Delta f_k}{\Delta x_k} = \frac{f_k - f_{k-1}}{x_k - x_{k-1}},$$

где  $f_k = f(x_k)$ . В результате будем иметь

$$x_{k+1} = x_k - f_k \frac{x_k - x_{k-1}}{f_k - f_{k-1}}, \quad k = 1, 2, \dots \quad (13.20)$$

Отметим, что этот метод двухшаговый: чтобы найти  $x_{k+1}$ , нужно знать  $x_k$  и  $x_{k-1}$ . В частности, для того, чтобы начать итерации, также требуются значения начальных приближений  $x_0$  и  $x_1$ .

Обратимся к геометрической интерпретации метода (13.20). Пусть

$$L_1(x) = f_{k-1} \frac{x - x_k}{x_{k-1} - x_k} + f_k \frac{x - x_{k-1}}{x_k - x_{k-1}} \quad (13.21)$$

— интерполяционный многочлен Лагранжа первой степени, построенный по значениям функции  $f(x)$  в узлах  $x_{k-1}$  и  $x_k$ . Рассмотрим прямую

$$y = L_1(x)$$

и найдем ее нуль, т.е. решение уравнения  $L_1(x) = 0$ . Принимая во внимание (13.21), находим, что

$$f_{k-1}(x - x_k) = f_k(x - x_{k-1}) \equiv f_k(x - x_k) + f_k(x_k - x_{k-1}).$$

Отсюда

$$x = x_k - f_k \frac{x_k - x_{k-1}}{f_k - f_{k-1}},$$

что совпадает с  $x_{k+1}$  из (13.20). Отсюда следует, что если в методе Ньютона кривая  $y = f(x)$  всякий раз заменяется касательной в точке  $x_k$ , то в методе (13.20) кривая  $y = f(x)$  заменяется секущей, пересекающей  $y = f(x)$  при  $x = x_{k-1}$  и  $x = x_k$ . Эта геометрическая интерпретация метода (13.20) и дает ему название — метод секущих.

Обратимся к оценке скорости сходимости метода секущих. Сначала заметим, что в силу (13.21) и с учетом формулы конечных приращений

$$\begin{aligned} L_1(x_{k+1}) - L_1(x^*) &= \\ &= f_{k-1} \frac{x_{k+1} - x_k}{x_{k-1} - x_k} + f_k \frac{x_{k+1} - x_{k-1}}{x_k - x_{k-1}} - f_{k-1} \frac{x^* - x_k}{x_{k-1} - x_k} - f_k \frac{x^* - x_{k-1}}{x_k - x_{k-1}} = \\ &= f_{k-1} \frac{x_{k+1} - x^*}{x_{k-1} - x_k} + f_k \frac{x_{k+1} - x^*}{x_k - x_{k-1}} = \frac{f_k - f_{k-1}}{x_k - x_{k-1}} (x_{k+1} - x^*) = \\ &= f'(\xi_k)(x_{k+1} - x^*), \quad \xi_k \in (x_{k-1}, x_k) \end{aligned}$$

и поэтому

$$x_{k+1} - x^* = \frac{L_1(x_{k+1}) - L_1(x^*)}{f'(\xi_k)} = -\frac{L_1(x^*)}{f'(\xi_k)}, \quad (13.22)$$

ибо по построению  $L_1(x_{k+1}) = 0$ . При вычислении  $L_1(x^*)$  воспользуемся формулой для погрешности интерполяции

$$f(x) - L_1(x) = \frac{1}{2} f''(\eta_k)(x - x_{k-1})(x - x_k), \quad \eta_k \in (x, x_{k-1}, x_k),$$



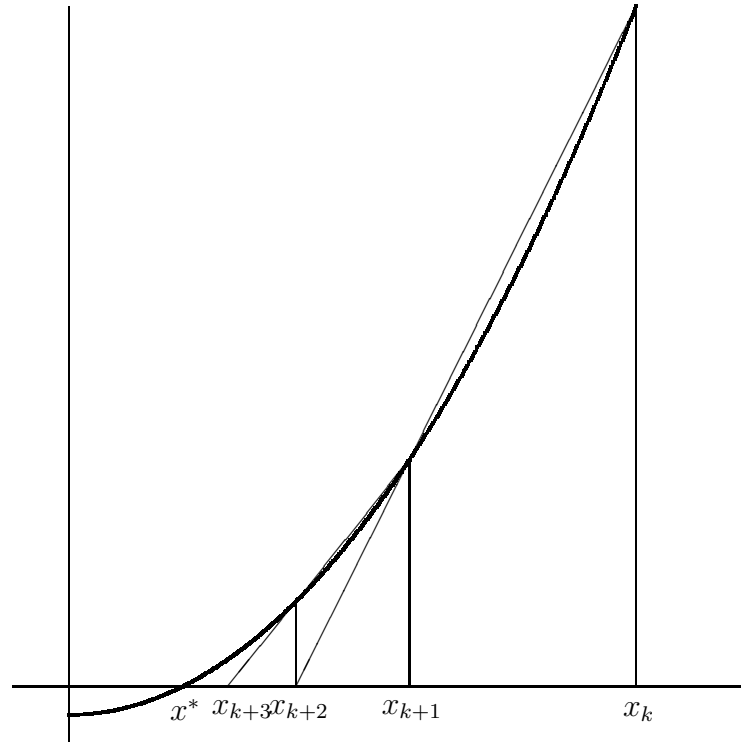


Рис. 5

где  $(x, x_{k-1}, x_k)$  — открытый интервал, концами которого являются крайние из указанных трех точек. Полагая здесь  $x = x^*$  и принимая во внимание, что  $f(x^*) = 0$ , находим искомое выражение, подставляя которое в (13.22), с учетом вышесказанного будем иметь

$$x_{k+1} - x^* = \frac{1}{2} \frac{f''(\eta_k)}{f'(\xi_k)} (x_k - x^*)(x_{k-1} - x^*). \quad (13.23)$$

Как и при изучении метода Ньютона, будем предполагать, что

$$|f'(x)| \geq m_1, \quad |f''(x)| \leq M_2. \quad (13.24)$$

Тогда из (13.23) будем иметь оценку

$$|x_{k+1} - x^*| \leq \frac{M_2}{2m_1} |x_k - x^*| |x_{k-1} - x^*|.$$

Домножая теперь обе части этого неравенства на  $M_2/(2m_1)$  и обозначая

$$\frac{M_2}{2m_1} |x_k - x^*| = \Delta_k, \quad (13.25)$$

найдем, что

$$\Delta_{k+1} \leq \Delta_k \Delta_{k-1}. \quad (13.26)$$

Отсюда следует, по крайней мере, сверхлинейная сходимость метода секущих. Для того, чтобы получить более точную оценку скорости сходимости, мы сначала воспользуемся не вполне строгими рассуждениями. Сравним (13.26) с оценкой (13.14), из которой следует квадратичная сходимость метода Ньютона. Скорость сходимости метода секущих, конечно, ниже. Возможно,

$$\Delta_{k+1} \leq \Delta_k^\nu, \quad \text{где } 1 < \nu < 2. \quad (13.27)$$

Мы не будем пытаться доказывать эту оценку, а установим оценку типа (13.27) для мажоранты погрешности  $\Delta_k$ . Заметим, что, если

$$z_{k+1} = z_k z_{k-1}, \quad z_0 \geq \Delta_0, \quad z_1 \geq \Delta_1, \quad (13.28)$$

то

$$\Delta_{k+1} \leq z_{k+1}. \quad (13.29)$$

Оценив убывание  $z_k$ , мы получим и оценку  $\Delta_k$ . Для того, чтобы оценить  $z_k$  — решение нелинейного разностного уравнения второго порядка (13.28), выясним, нет ли среди решений этого уравнения таких, которые одновременно являются решениями уравнения первого порядка

$$z_{k+1} = z_k^\nu. \quad (13.30)$$

при некотором  $\nu$ . Если это так, то  $z_k = z_{k-1}^\nu$ , и, следовательно,

$$z_{k-1} = z_k^{1/\nu}.$$

Подставляя это соотношение и (13.30) в (13.28), получим

$$z_k^\nu = z_k^{1+1/\nu}.$$

Приравнивая степени  $z_k$  в левой и правой частях, находим, что

$$\nu = 1 + 1/\nu, \quad \text{т.е. } \nu^2 - \nu - 1 = 0. \quad (13.31)$$

Отсюда вытекает, что наша гипотеза верна при

$$\nu_{1,2} = \frac{1 \pm \sqrt{5}}{2}.$$

Корню  $\nu = (1 - \sqrt{5})/2$  отвечает неубывающее решение уравнения (13.30), которое не может описывать сходящийся итерационный процесс. Поэтому следует взять

$$\nu = \frac{1 + \sqrt{5}}{2} \approx 1.6180339. \quad (13.32)$$

Итак, вместо (13.27) имеем (13.29), (13.30), (13.32), что позволяет говорить о сходимости метода секущих со скоростью (13.32). Как и ожидалось, эта скорость меньше, чем у метода Ньютона.

**Пример 13.5.** Пусть требуется найти корень функции  $f(x) = x^2 - a$ . Метод секущих применительно к этой функции принимает вид

$$x_{k+1} = x_k - \frac{x_k^2 - a}{x_k^2 - x_{k-1}^2} = \frac{x_k x_{k-1} + a}{x_k + x_{k-1}}.$$

Если  $a = 2$  и положить  $x_0 = x_1 = 2$ , то

$$x_2 = \frac{3}{2}, \quad x_3 = \frac{7}{5}, \quad x_4 = \frac{41}{29} \approx 1.41379, \quad x_5 = \frac{577}{408} \approx 1.4142156$$

при  $x^* = 1.41421356 \dots$

**Замечание 13.2.** Нелинейное разностное уравнение (13.30) имеет очевидное решение

$$z_k = z_0^{\nu^k}, \quad (13.33)$$

для которого, в частности,  $z_1 = z_0^{\nu}$ . Но в итерационном методе (13.20) используются два начальных условия, и поэтому величина  $z_1$  не должна зависеть от  $z_0$ . Мы вынуждены констатировать, что найденное решение (13.33) не совсем правильно описывает этот процесс. То, что мы не получили решения уравнения (13.28), удовлетворяющего обоим начальным условиям, не должно вызывать удивления: мы ведь нашли решение, общее для (13.28) и (13.30), а интересующее нас решение может (13.30) и не удовлетворять.

Вернемся к задаче (13.28) и найдем ее решение. Логарифмируя уравнение (13.28), будем иметь

$$\ln z_{k+1} = \ln z_k + \ln z_{k-1}.$$

Обозначим

$$\ln z_k = y_k, \quad (13.34)$$

и получим для  $y_k$  линейное разностное уравнение с постоянными коэффициентами

$$y_{k+1} = y_k + y_{k-1}.$$

Его характеристическое уравнение есть

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

(сравни с (13.31)) с корнями

$$q_1 = \frac{1 + \sqrt{5}}{2}, \quad q_2 = \frac{1 - \sqrt{5}}{2}, \quad q_1 + q_2 = 1, \quad \frac{q_2}{q_1} = \frac{-3 + \sqrt{5}}{2} \approx -0.38.$$

Поэтому

$$y_k = c_1 q_1^k + c_2 q_2^k. \quad (13.35)$$

Удовлетворяя начальным условиям (13.28), будем иметь

$$\begin{aligned} c_1 + c_2 &= y_0 = \ln z_0, \\ q_1 c_1 + q_2 c_2 &= y_1 = \ln z_1. \end{aligned}$$

Отсюда находим, что

$$\begin{aligned} c_1 &= \frac{y_1 - y_0 q_2}{\sqrt{5}} = \frac{y_1 - y_0 + y_0 q_1}{\sqrt{5}}, \\ c_2 &= -\frac{y_1 - y_0 q_1}{\sqrt{5}} = -\frac{y_1 - y_0 + y_0 q_2}{\sqrt{5}} \end{aligned}$$

и, следовательно,

$$\begin{aligned} y_k &= \frac{y_1 + (q_1 - 1)y_0}{\sqrt{5}} q_1^k - \frac{y_1 + (q_2 - 1)y_0}{\sqrt{5}} q_2^k = \\ &= \frac{\ln(z_1 z_0^{q_1 - 1})}{\sqrt{5}} q_1^k - \frac{\ln(z_1 z_0^{q_2 - 1})}{\sqrt{5}} q_2^k, \end{aligned}$$

а с учетом (13.34)

$$\begin{aligned} z_k &= \exp \left\{ \ln(z_1 z_0^{q_1 - 1}) \frac{q_1^k}{\sqrt{5}} - \ln(z_1 z_0^{q_2 - 1}) \frac{q_2^k}{\sqrt{5}} \right\} = \\ &= \left( z_1 z_0^{q_1 - 1} \right)^{q_1^k / \sqrt{5}} / \left( z_1 z_0^{q_2 - 1} \right)^{q_2^k / \sqrt{5}}. \end{aligned}$$

Отсюда, в частности, следует, что, если

$$z_1 = z_0^{q_1},$$

то

$$z_1 z_0^{q_2-1} = z_0^{q_1+q_2-1} = 1, \quad z_1 z_0^{q_1-1} = z_0^{2q_1-1} = z_0^{\sqrt{5}}$$

и, следовательно,

$$z_k = z_0^{q_1^k},$$

что совпадает с (13.33), ибо  $q_1 = \nu$ .

В более же реалистическом случае, когда  $z_1 = z_0$ ,

$$z_k = z_0^{(q_1^{k+1} - q_2^{k+1})/\sqrt{5}} = z_0^{q_1^{k+1}(1 - (q_2/q_1)^{k+1})/\sqrt{5}}.$$

Это соотношение дает представление погрешности на  $k$ -ой итерации через погрешности  $z_0$  и  $z_1 = z_0$ .

Небезынтересен вопрос о скорости убывания погрешности на двух соседних итерациях. Пусть

$$z_{k+1} = z_k^{\nu_k}.$$

Отсюда

$$\nu_k = (\ln z_{k+1})/(\ln z_k) = y_{k+1}/y_k,$$

а с учетом (13.35)

$$\begin{aligned} \nu_k &= \frac{y_{k+1}}{y_k} = \frac{c_1 q_1^{k+1} + c_2 q_2^{k+1}}{c_1 q_1^k + c_2 q_2^k} = \\ &= q_1 \frac{1 + \frac{c_2}{c_1} (q_2/q_1)^{k+1}}{1 + \frac{c_2}{c_1} (q_2/q_1)^k} = q_1 \left[ 1 - \frac{c_2}{c_1} \left( \frac{q_2}{q_1} \right)^k + O \left( \frac{q_2}{q_1} \right)^{k+1} \right] = \\ &= q_1 + O \left( (q_2/q_1)^k \right). \end{aligned}$$

Отсюда следует, что при любых начальных приближениях убывание погрешности в малой окрестности корня происходит со скоростью  $\sim q_1$ .

**Замечание 13.3.** Мы уже отмечали, что скорость сходимости метода секущих ниже, чем метода Ньютона. И, тем не менее, метод секущих может оказаться более предпочтительным по сравнению с методом Ньютона. Для реализации каждой итерации в методе Ньютона нужно вычислять значение функции и ее производной в новой точке. В методе секущих на каждой итерации нужно знать только одно новое значение функции. Если эти операции трудоемкие, то две итерации по методу секущих могут быть сравнимы по трудоемкости с одной итерацией по методу Ньютона, а это приводит к большему уменьшению начальной погрешности.

## 13.5 Глобальная сходимость. Гибридные алгоритмы

Как уже отмечалось, метод Ньютона является локально сходящимся методом, т.е. его сходимость гарантирована лишь в том случае, если начальное приближение расположено достаточно близко к искомому решению. Обеспечить выполнение этого условия заранее весьма проблематично. Удачным выходом из создавшейся ситуации является объединение локально быстросходящегося метода Ньютона и некоторого глобально, но не слишком быстро сходящегося метода в один алгоритм, называемый *гибридным*. Таким глобально сходящимся методом с линейной скоростью сходимости может быть метод деления отрезка пополам. Однако, поскольку этот метод не имеет аналога в многомерном случае, мы рассмотрим другой метод.

На рис. 6 изображена ситуация, когда из-за неудачного выбора началь-

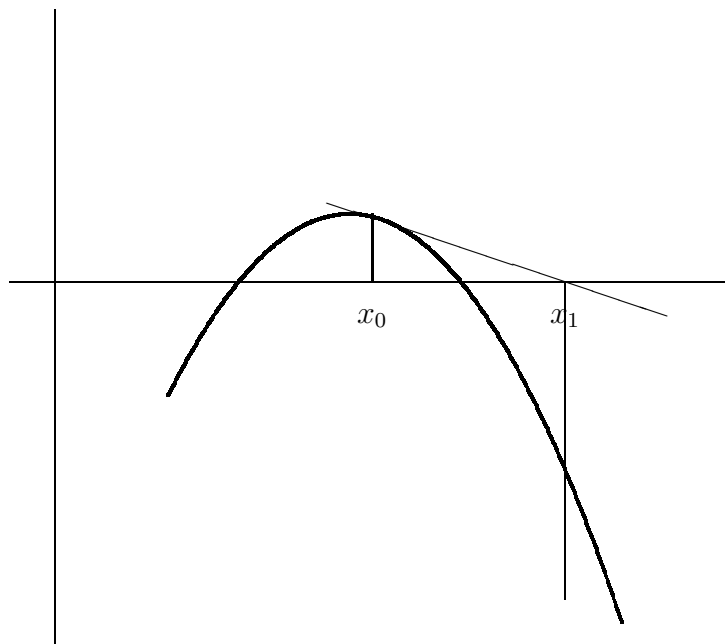


Рис. 6

ного приближения итерационный метод Ньютона начинает расходиться; по крайней мере,

$$|f(x_0)| < |f(x_1)|.$$

Тем не менее, как легко видеть, (геометрически это очевидно) на отрезке

$[x_0, x_1]$  есть точки  $t$  такие, что

$$|f(t)| < |f(x_0)|.$$

На этом факте и основан излагаемый гибридный метод. Именно, сделаем очередную итерацию по методу Ньютона (13.6), однако полученное значение обозначим не  $x_{k+1}$ , а  $z_{k+1}$  — некоторое вспомогательное значение:

$$z_{k+1,1} = x_k - \frac{f_k}{f'_k}.$$

Затем сравним  $f_k$  и  $f(z_{k+1,1})$ . Если

$$|f(z_{k+1,1})| < |f(x_k)|$$

то  $x_{k+1} = z_{k+1,1}$ , и делается новая итерация по методу Ньютона. Если же

$$|f(z_{k+1,1})| \geq |f(x_k)|,$$

то вычисляется

$$z_{k+1,2} = \frac{x_k + x_{k+1}}{2}$$

и сравнивается  $f(z_{k+1,2})$  и  $f(x_k)$ . И т.д.

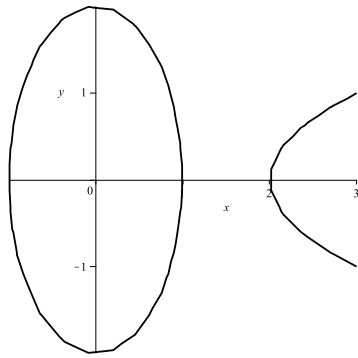
## 13.6 Системы нелинейных уравнений

В качестве примера сначала рассмотрим систему уравнений

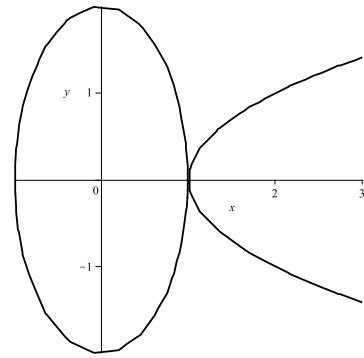
$$\begin{cases} 4x_1^2 + x_2^2 = 4, \\ x_1 - x_2^2 - t = 0. \end{cases}$$

Здесь  $x_1$  и  $x_2$  — неизвестные, а  $t$  — параметр. Первое уравнение задает на плоскости  $Ox_1x_2$  эллипс с полуосями, равными 1 и 2, а второе — параболу. Координаты точек пересечения этих кривых и дают решение системы. Если значение параметра  $t$  изменяется от  $-2$  до  $5$ , то возможны следующие ситуации, изображенные на рисунке 7, расположенном на следующей странице.

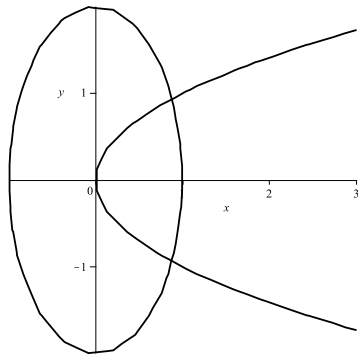
Это пример, иллюстрирующий сложности с локализацией корней в многомерном случае.



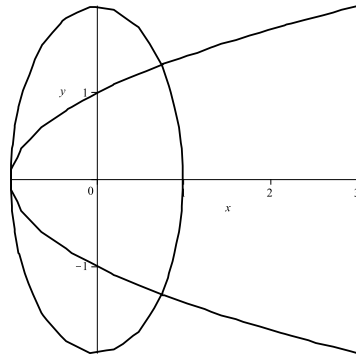
$t = -2$   
Решений нет



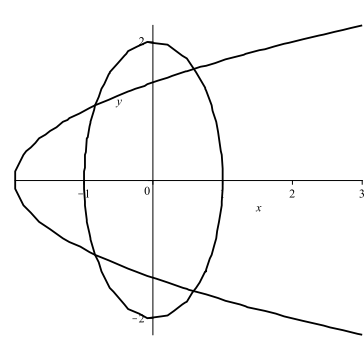
$t = -1$   
Одно решение



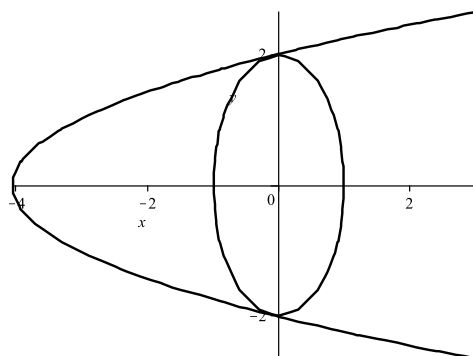
$t = 0$   
Два решения



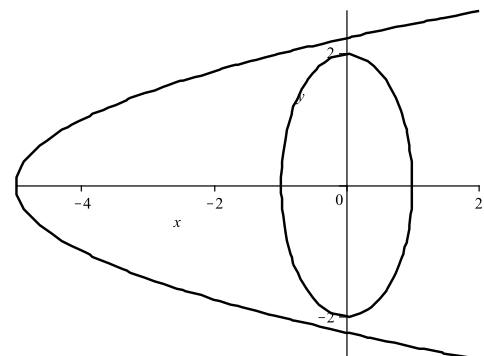
$t = 1$   
Три решения



$t = 2$   
Четыре решения



$t = 65/16$   
Два решения



$t = 5$   
Решений нет

Рис. 7



В общем случае система уравнений имеет вид

$$\begin{aligned} f_1(x_1, x_2, \dots, x_n) &= 0, \\ f_2(x_1, x_2, \dots, x_n) &= 0, \\ &\dots\dots\dots \\ f_n(x_1, x_2, \dots, x_n) &= 0. \end{aligned}$$

Если ввести обозначения  $x = [x_1 \ x_2 \ \dots \ x_n]^T$ ,  $f = [f_1 \ f_2 \ \dots \ f_n]^T$ , то эту систему можно записать в векторном виде

$$f(x) = 0. \quad (13.36)$$

Многие одношаговые итерационные методы решения системы (13.36) могут быть записаны в виде

$$B_k \frac{x^{k+1} - x^k}{\tau_{k+1}} + f(x^k) = 0, \quad k = 0, 1, \dots, \quad (13.37)$$

где  $B_k$  — некоторая невырожденная матрица. Для реализации (13.37) на каждом шаге итераций нужно решать линейную систему

$$B_k x^{k+1} = \varphi(x_k), \quad \varphi(x_k) = B_k x^k - \tau_{k+1} f(x^k). \quad (13.38)$$

Метод (13.37) называется явным, если  $B_k = I$ , и неявным в противном случае. Систему (13.38) можно решать либо прямым методом, либо итерационным. В последнем случае итерации, приводящие к решению линейной системы (13.38), называются внутренними, а итерации по нелинейности (13.37) — внешними.

При  $B_k = I$  и  $\tau_{k+1} = \tau$  из (13.37) имеем

$$x^{k+1} = \varphi(x^k), \quad \varphi(x) = x - \tau f(x), \quad (13.39)$$

т.е. метод простых итераций. В развернутой форме он принимает вид

$$\begin{aligned} x_1^{k+1} &= \varphi_1(x_1^k, x_2^k, \dots, x_n^k), \\ &\dots\dots\dots \\ x_n^{k+1} &= \varphi_n(x_1^k, x_2^k, \dots, x_n^k). \end{aligned}$$

Пусть

$$\varphi' = \begin{bmatrix} \frac{\partial \varphi_1}{\partial x_1} & \frac{\partial \varphi_1}{\partial x_2} & \dots & \frac{\partial \varphi_1}{\partial x_n} \\ \frac{\partial \varphi_2}{\partial x_1} & \frac{\partial \varphi_2}{\partial x_2} & \dots & \frac{\partial \varphi_2}{\partial x_n} \\ \dots\dots\dots \\ \frac{\partial \varphi_n}{\partial x_1} & \frac{\partial \varphi_n}{\partial x_2} & \dots & \frac{\partial \varphi_n}{\partial x_n} \end{bmatrix}$$

— матрица Якоби функции  $\varphi$ . Тогда метод (13.39) сходится, если

$$\|\varphi'\| \leq q < 1.$$

### 13.6.1 Метод Ньютона

Пусть

$$\Delta f = df + O(|dx|^2)$$

и

$$\begin{aligned} \Delta f &\approx df = f'dx \\ dx &= x - x^k, \quad \Delta f = f(x) - f(x^k), \\ f(x) &\approx f(x^k) + f'(x^k)(x - x^k) = 0. \end{aligned}$$

Отсюда

$$f'(x^k)(x^{k+1} - x^k) + f(x^k) = 0, \quad (13.40)$$

или, в разрешенном виде,

$$x^{k+1} = x^k - [f'(x^k)]^{-1}f(x^k). \quad (13.41)$$

Неразрешенная форма более естественна, ибо на каждой итерации требуется решать линейную систему с матрицей  $f'(x^k)$ . Очевидно, что (13.40) есть частный случай (13.37) при  $B_k = f'(x^k)$  и  $\tau_{k+1} = 1$ . Метод очень трудоемок: на каждой итерации требуется вычисление  $n^2$  элементов матрицы  $f'(x^k)$  и решение системы с этой матрицей. Сходимость квадратичная из малой окрестности решения  $x^*$ . Более того, имеет место

**Теорема 13.4.** Пусть функция  $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}^n$  принадлежит  $C^1(D)$ , где  $D$  — открытое выпуклое множество в  $\mathbb{R}^n$ , причем  $x^* \in D$ . Пусть  $[f'(x^*)]^{-1}$  существует и существуют положительные постоянные  $R$ ,  $C$  и  $L$  такие, что  $\|[f'(x^*)]^{-1}\| \leq C$ ,

$$\|f'(x) - f'(y)\| \leq L \|x - y\| \quad \forall x, y \in B(x^*, R).$$

Тогда существует  $r > 0$  такое, что для любого  $x^0 \in B(x^*, r)$  последовательность (13.40) однозначно определена и сходится к  $x^*$ , причем

$$\|x^{k+1} - x^*\| \leq CL \|x^k - x^*\|^2$$

Покажем, как выглядит система (13.40) в развернутом (покоординатном) виде при  $n = 2$ .

$$\begin{bmatrix} \frac{\partial f_1(x_1^k, x_2^k)}{\partial x_1} & \frac{\partial f_1(x_1^k, x_2^k)}{\partial x_2} \\ \frac{\partial f_2(x_1^k, x_2^k)}{\partial x_1} & \frac{\partial f_2(x_1^k, x_2^k)}{\partial x_2} \end{bmatrix} \begin{bmatrix} \delta_1^{k+1} \\ \delta_2^{k+1} \end{bmatrix} + \begin{bmatrix} f_1(x_1^k, x_2^k) \\ f_2(x_1^k, x_2^k) \end{bmatrix} = 0,$$

$$x_1^{k+1} = x_1^k + \delta_1^{k+1},$$

$$x_2^{k+1} = x_2^k + \delta_2^{k+1}.$$

**Пример 13.6.** Рассмотрим нелинейную систему

$$e^{x_1^2 + x_2^2} - 1 = 0,$$

$$e^{x_1^2 - x_2^2} - 1 = 0,$$

которая имеет единственное решение  $x^* = 0$ . Метод Ньютона с начальным приближением  $x^0 = [0.1 \ 0.1]^T$  за 15 итераций приводит к решению  $[0.61 \cdot 10^{-5} \ 0.61 \cdot 10^{-5}]^T$ , что демонстрирует довольно высокую скорость сходимости. Однако при другом начальном приближении  $x^0 = [10 \ 10]^T$  для получения того же результата требуется уже 220 итераций, а при  $x^0 = [20 \ 20]^T$  метод Ньютона для этой системы расходится.

Поскольку итерации по методу Ньютона сходятся не всегда (как в приведенном примере), то и в этом случае его нужно объединять с глобально сходящимся методом, например, так, как это было сделано для одного уравнения.

### 13.6.2 Аналоги метода секущих

Несколько упростить метод Ньютона можно, как и в одномерном случае, отказавшись от точного вычисления матрицы Якоби, т.е. путем замены производных разностными отношениями. Если шаги по переменным  $x_i$  при аппроксимации производных будут постоянными, то построенный метод утратит сверхлинейную скорость сходимости метода Ньютона. Поэтому шаги должны стремиться к нулю при приближении  $x^k$  к  $x^*$ . Если положить

$$h_i^k = x_i^{k-1} - x_i^k, \quad i = 1, \dots, n,$$

и принять

$$\begin{aligned} B(x^k, x^{k-1}) &:= [b_{ij}] = \\ &= \left[ \left\{ f_i(x_1^k, \dots, x_{j-2}^k, x_{j-1}^{k-1}, x_j^k, \dots, x_n^k) - f_i(x_1^k, \dots, x_j^k, \dots, x_n^k) \right\} \frac{1}{x_j^{k-1} - x_j^k} \right] \approx \\ &\approx \left[ \frac{\partial f_i}{\partial x_j}(x^k) \right], \end{aligned}$$

то получим простейший вариант метода секущих

$$B(x^k, x^{k-1})(x^{k+1} - x^k) + F(x^k) = 0.$$

Можно рассчитывать, как и в одномерном случае, на сверхлинейную сходимость с порядком  $\frac{1+\sqrt{5}}{2}$ .

### 13.6.3 Метод Бroyдена

Прежде, чем приступить к построению этого метода, вернемся на некоторое время к одномерному случаю. При построении метода Ньютона функцию  $f(x)$  мы заменили линейной моделью

$$f(x) \approx f(x_k) + f'(x_k)(x - x_k) =: P_1(x),$$

нуль которой и приняли за новое приближение

$$P_1(x) = 0, \quad \Rightarrow \quad x = x_{k+1}.$$

В методе секущих линейная модель была иной:

$$f(x) \approx f(x_k) + q_k(x - x_k) =: L_1(x), \quad (13.42)$$

где

$$q_k = \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}},$$

т.е.

$$f(x_k) - f(x_{k-1}) = q_k(x_k - x_{k-1}). \quad (13.43)$$

Интегрировать соотношение (13.42) можно двояко: рассматривать уравнение  $y = L_1(x)$  как уравнение прямой, аппроксимирующей касательную  $y = P_1(x)$ ; рассматривать  $y = L_1(x)$  как интерполяцию функции  $f(x)$  по точкам  $x^k$  и  $x^{k+1}$ . В одномерном случае обе интерпретации приводят к одним и тем же соотношениям. В многомерном случае дело обстоит

несколько сложнее. Первая интерпретация приводит к дискретным методам Ньютона, которые мы уже обсуждали. при второй интерпретации это соотношение можно рассматривать как уравнение для (однозначного определения  $q_k$ .

В  $n$ -мерном случае линейная модель, аналогичная (13.42), должна выглядеть так:

$$f(x) \approx f(x^k) + Q_k(x - x^k) =: \Psi_k(x), \quad (13.44)$$

где вместо (13.43) мы имеем

$$f(x^k) - f(x^{k-1}) = Q_k(x^k - x^{k-1}). \quad (13.45)$$

Это соотношение уже  $Q_k$  однозначно не определяет: существует много матриц  $Q_k$ , переводящих заданный вектор  $(x^k - x^{k-1})$  в другой заданный вектор  $[f(x^k) - f(x^{k-1})]$ . Поэтому будем строить матрицы  $B_k$  таким образом, чтобы *переход от  $B_{k-1}$  к  $B_k$  требовал минимальных изменений*. Пусть в соответствии с (13.44)

$$f(x^{k-1}) + Q_{k-1}(x - x^{k-1}) = \Psi_{k-1}(x).$$

Вычтем это соотношение из (13.44) и преобразуем с учетом (13.45)

$$\begin{aligned} \Psi_k(x) - \Psi_{k-1}(x) &= f(x^k) + Q_k(x - x^k) - f(x^{k-1}) - Q_{k-1}(x - x^{k-1}) = \\ &= Q_k(x^k - x^{k-1}) + (Q_k - Q_{k-1})x - Q_k x^k + Q_{k-1} x^{k-1} = \\ &= (Q_k - Q_{k-1})(x - x^{k-1}). \end{aligned}$$

Пусть вектор  $t \neq 0$  ортогонален вектору  $(x^k - x^{k-1})$ , т.е.  $(x^k - x^{k-1}, t) = 0$  и пусть

$$x - x^{k-1} = \alpha(x^k - x^{k-1}) + t.$$

Тогда с учетом (13.45)

$$\begin{aligned} \Psi_k(x) - \Psi_{k-1}(x) &= \alpha(Q_k - Q_{k-1})(x^k - x^{k-1}) + (Q_k - Q_{k-1})t = \\ &= \alpha[f(x^k) - f(x^{k-1})] - q_{k-1}(x^k - x^{k-1}) + (Q_k - Q_{k-1})t. \end{aligned}$$

Первые два вектора этого соотношения уже определены, а последний зависит от от определения  $Q_k$ . Пусть  $Q_k$  такова, что

$$(Q_k - Q_{k-1})t = 0 \quad \forall \quad t : (t, x^k - x^{k-1}) = 0.$$

Это будет заведомо так, если положить

$$Q_k - Q_{k-1} = \frac{1}{\|x^k - x^{k-1}\|^2} [(f(x^k) - f(x^{k-1})) - Q_{k-1}(x^k - x^{k-1})] [x^k - x^{k-1}]^T.$$

Итак, построен итерационный метод, называемый методом Бroyдена:

$$Q_k(x^{k+1} - x^k) + f(x^k) = 0, \quad k = 0, 1, \dots$$
$$Q_k = Q_{k-1} + \frac{[(f(x^k) - f(x^{k-1})) - Q_{k-1}(x^k - x^{k-1})][x^k - x^{k-1}]^T}{\|x^k - x^{k-1}\|^2}$$
$$x^0, \quad Q_0$$

Обычно  $Q_0 = f'(x^0)$ .

# Численные методы решения дифференциальных уравнений

## IV

# Численное решение задачи Коши для обыкновенных дифференциальных уравнений



# 14

## Введение

Рассмотрим задачу Коши для обыкновенного дифференциального уравнения первого порядка

$$\frac{du}{dt} = f(t, u), \quad t > 0, \quad u(0) = u_0. \quad (14.1)$$

Из курса дифференциальных уравнений известно, что для однозначной разрешимости задачи (14.1) в некоторой окрестности точки  $t = 0$  достаточно, чтобы функция  $f(t, u)$  была непрерывна в окрестности точки  $(0, u_0)$  и удовлетворяла условию Липшица по второму аргументу. Известны примеры, иллюстрирующие отсутствие решения задачи (14.1) или его неединственность при нарушении указанных условий. Мы всегда будем предполагать, что решение задачи (14.1) существует и единственно. Для дальнейшего нам даже придется предполагать, что искомое решение достаточно гладкое.

### 14.1 Примеры численных методов

Приведем несколько простейших численных методов решения задачи (14.1). Для этого введем на полуоси  $t \geq 0$  равномерную сетку, т.е. множество точек (которые назовем узлами)

$$\omega = \{t_n = n\tau, \quad n = 0, 1, \dots; \quad \tau > 0\}$$

и будем искать приближенное решение задачи (14.1) в узлах  $\omega$ . Величину  $\tau$  будем называть шагом сетки  $\omega$ . Договоримся приближенное решение в узле  $t_n$  обозначать той же буквой, что и решение задачи (14.1), но с индексом  $n$  внизу:  $u_n$ . Тем самым, мы отказываемся от часто используемого обозначения  $u(t_n) = u_n$ ; теперь  $u(t_n)$  — значение точного решения в

узле  $t_n$ , а  $u_n$  — значение приближенного решения в этом узле, и, вообще говоря,  $u(t_n) \neq u_n$ . Наоборот,  $u_n - u(t_n)$  представляет собой погрешность численного метода в узле  $t_n$ , которую нам предстоит оценивать. Данное соглашение не представляется наилучшим, однако остановимся на нем.

Для построения численных методов проинтегрируем уравнение (14.1) от  $t_n$  до  $t_{n+1}$

$$u(t_{n+1}) - u(t_n) = \int_{t_n}^{t_{n+1}} f(t, u(t)) dt \quad (14.2)$$

и заменим приближенно интеграл в правой части этой формулы какой-либо квадратурной формулой. Здесь мы рассмотрим четыре таких формулы.

Построенная в курсе "Введение в численные методы" квадратурная формула прямоугольников представляет интеграл произведением длины отрезка интегрирования и значения подынтегральной функции в середине этого отрезка

$$\int_a^b \varphi(x) dx \approx |b - a| \varphi\left(\frac{a + b}{2}\right). \quad (14.3)$$

Эта квадратурная формула точна на многочленах первой степени, и при малых  $|b - a|$  ее погрешность есть  $O(|b - a|^3)$ .

Наряду с этой квадратурной формулой, которую мы впредь будем называть формулой *центральных прямоугольников*, можно ввести так называемые формулы *левых* и *правых прямоугольников*. Первая из них состоит в представлении интеграла произведением длины отрезка интегрирования и значения подынтегральной функции в левом конце отрезка

$$\int_a^b \varphi(x) dx \approx |b - a| \varphi(a), \quad (14.4)$$

а вторая — произведением длины отрезка интегрирования и значения подынтегральной функции в правом конце отрезка

$$\int_a^b \varphi(x) dx \approx |b - a| \varphi(b). \quad (14.5)$$

Обе эти формулы точны только на многочленах нулевой степени и имеют погрешность  $O(|b - a|^2)$ .

а) **Метод Эйлера.** Заменим интеграл в (14.2) формулой левых прямоугольников (14.4). В результате получим приближенное равенство

$$u(t_{n+1}) - u(t_n) \approx \tau f(t_n, u(t_n)). \quad (14.6)$$

Определим приближенное решение задачи (14.1) как такую сеточную функцию, определенную на  $\omega$ , которая превращает соотношение (14.6) в равенство. Разделив полученное равенство на  $\tau$ , будем иметь

$$\frac{u_{n+1} - u_n}{\tau} = f(t_n, u_n), \quad n = 0, 1, \dots, \quad u_0 = u(0). \quad (14.7)$$

Соотношение (14.7) позволяет рекуррентным образом найти приближенное решение во всех узлах. Численный метод решения задачи (14.1), реализуемый формулами (14.7), называется *методом Эйлера*.

б) **Неявный метод Эйлера.** Заменяем теперь интеграл в (14.2) формулой правых прямоугольников (14.5). Для отыскания приближенного решения получим уравнения

$$\frac{u_{n+1} - u_n}{\tau} = f(t_{n+1}, u_{n+1}), \quad n = 0, 1, \dots, \quad u_0 = u(0). \quad (14.8)$$

Соотношения (14.8) коренным образом отличаются от соотношений (14.7): для отыскания приближенного решения  $u_{n+1}$  теперь нужно решать нелинейные уравнения

$$u_{n+1} - \tau f(t_{n+1}, u_{n+1}) = u_n.$$

Метод (14.8) называется *неявным методом Эйлера*. С точки зрения простоты вычислений он сильно уступает обычному методу Эйлера (14.7). Как будет показано позже, по точности оба метода сравнимы. Еще позже будет установлена существенно бóльшая устойчивость метода (14.8) по сравнению с (14.7).

в) **Метод Рунге.** Заменяем интеграл в (14.2) формулой центральных прямоугольников (14.3)

$$u(t_{n+1}) - u(t_n) \approx \tau f(t_{n+1/2}, u(t_{n+1/2})). \quad (14.9)$$

Использованный нами ранее прием получения численного метода путем превращения приближенного равенства в точное за счет замены  $u(t_n)$  на  $u_n$  здесь напрямую не проходит: в приближенном равенстве фигурирует значение  $f$  при  $u$  в точке  $t_{n+1/2}$ , которая не является узловой. Если же мы все же воспользуемся этим приемом и введем *промежуточное* значение приближенного решения в точке  $t_{n+1/2}$ , то нам потребуется дополнительное уравнение для определения приближенного решения в точке  $t_{n+1/2}$ . Обозначим промежуточное значение приближенного решения через  $u_{n+1/2}$ . Тогда из (14.9)

$$\frac{u_{n+1} - u_n}{\tau} = f(t_{n+1/2}, u_{n+1/2}), \quad (14.10)$$

а для отыскания  $u_{n+1/2}$  напишем, например, соотношение Эйлера (14.7)

$$\frac{u_{n+1/2} - u_n}{\tau/2} = f(t_n, u_n). \quad (14.11)$$

Итак, в методе (14.10), (14.11) вычисление нового приближенного значения искомого решения  $u_{n+1}$  осуществляется поэтапно. Сначала находится промежуточное значение  $u_{n+1/2}$  по формуле (14.11), а затем и само  $u_{n+1}$  из (14.10). Вычисления по обеим формулам явные. *Метод* (14.10), (14.11) был предложен немецким математиком *Рунге* и носит его имя. Будет показано, что точность метода (14.10), (14.11) выше, чем точность методов (14.7) и (14.8); для вычисления интеграла все же использована более точная квадратурная формула.

**Замечание 14.1.** В некоторых учебниках по численным методам метод (14.10), (14.11) называется методом предиктор-корректор (предсказывающе корректирующим).

г) **Метод трапеций.** Наконец, заменим интеграл в (14.2) *формулой трапеций*

$$\int_a^b \varphi(x) dx \approx |b - a| \frac{\varphi(a) + \varphi(b)}{2}.$$

В результате получим

$$\frac{u_{n+1} - u_n}{\tau} = \frac{f(t_n, u_n) + f(t_{n+1}, u_{n+1})}{2}, \quad n = 0, 1, \dots, \quad u_0 = u(0). \quad (14.12)$$

Как и в случае неявного метода Эйлера (14.8), реализация метода (14.12) требует решения нелинейного уравнения

$$u_{n+1} - \frac{\tau}{2} f(t_{n+1}, u_{n+1}) = F(u_n) := u_n + \frac{\tau}{2} f(t_n, u_n).$$

Будет показано, что точность метода (14.12) сравнима с точностью метода Рунге (14.10), (14.11), а по устойчивости он значительно превосходит последний и в этом отношении близок к неявному методу Эйлера (14.8). Метод (14.12) будем называть *методом трапеций*.

## 14.2 Аппроксимация.

**Определение 14.1.** Сеточная функция

$$z_n = u_n - u(t_n), \quad n = 1, 2, \dots$$

называется погрешностью решения.

**Замечание 14.2.** Погрешность решения определена только в узлах основной сетки  $\omega$ , но не в промежуточных узлах.

Выведем уравнение, которому удовлетворяет погрешность решения в методе Эйлера (14.7). Подставив  $u_n = z_n + u(t_n)$  в (14.7), получим

$$\frac{z_{n+1} - z_n}{\tau} + \frac{u(t_{n+1}) - u(t_n)}{\tau} = f(t_n, u(t_n) + z_n). \quad (14.13)$$

Преобразуем правую часть этого соотношения путем разложения по формуле Тейлора

$$f(t_n, u(t_n) + z_n) = f(t_n, u(t_n)) + z_n \frac{\partial f}{\partial u}(t_n, \tilde{u}),$$

где

$$\tilde{u} = u(t_n) + \theta z_n, \quad 0 < \theta < 1.$$

Подставляя это разложение в (14.13) и преобразовывая, найдем, что

$$\frac{z_{n+1} - z_n}{\tau} = \frac{\partial f}{\partial u}(t_n, \tilde{u}) z_n + \psi_n, \quad (14.14)$$

где

$$\psi_n = f(t_n, u(t_n)) - \frac{u(t_{n+1}) - u(t_n)}{\tau}. \quad (14.15)$$

Искомое уравнение получено.

**Определение 14.2.** Сеточная функция  $\psi_n$ , задаваемая соотношением (14.15), называется *погрешностью аппроксимации* дифференциального уравнения (14.1) уравнением (14.7).

**Замечание 14.3.** Погрешность аппроксимации представляет собой разность между правой и левой частями уравнения, определяющего численный метод, если туда вместо приближенного решения подставить точное.

Оценим погрешность аппроксимации метода Эйлера. Используя формулу Тейлора и принимая во внимание уравнение (14.1), в предположении непрерывности второй производной  $u(t)$  из (14.15) будем иметь

$$\begin{aligned} \psi_n &= f(t_n, u(t_n)) - \frac{u(t_n) + \tau u'(t_n) + \frac{\tau^2}{2} u''(t_n + \tilde{\theta}\tau) - u(t_n)}{\tau} = \\ &= [f(t_n, u(t_n)) - u'(t_n)] + \frac{\tau}{2} u''(t_n + \tilde{\theta}\tau) = O(\tau). \end{aligned}$$

Тем самым, метод Эйлера имеет первый порядок аппроксимации.

**Упражнение 14.1.** Исследовать погрешности аппроксимации методов (14.8) и (14.12).

**Указание.** Для упрощения выкладок разложение по формуле Тейлора в методе (14.8) вести в точке  $t_{n+1}$ , а в методе (14.12) — в точке  $t_{n+1/2}$ .

# 15

## Методы Рунге-Кутты

### 15.1 Общая концепция

Основные численные методы решения уравнения

$$\frac{du}{dt} = f(t, u), \quad t > 0, \quad u(0) = u_0 \quad (15.1)$$

и систем таких уравнений, наиболее широко используемые в вычислительной практике, делятся на два больших класса: многошаговые методы и методы Рунге-Кутты. Все приведенные в качестве примеров численные методы относятся к методам Рунге-Кутты, хотя некоторые из них могут трактоваться и как многошаговые (одношаговые).

Сейчас мы опишем общую концепцию методов Рунге-Кутты. Для этого вновь обратимся к интегральному соотношению (14.2), на основе которого мы строили изложенные выше методы. Но прежде сделаем одно допущение относительно уравнения (15.1), которое в дальнейшем существенно облегчит нам жизнь. Будем предполагать, что правая часть  $f$  этого уравнения не зависит явным образом от  $t$ , т.е.  $f \equiv f(u)$  и, следовательно,

$$\frac{du}{dt} = f(u), \quad t > 0, \quad u(0) = u_0. \quad (15.2)$$

Сделанное допущение не является ограничением, ибо все численные методы, построенные для одного уравнения, допускают очевидное распространение на случай системы, т.е., вообще говоря,  $u$  можно считать вектором. Если же  $f$  зависит явным образом от  $t$ , то, обозначив, например,  $t = u_0(t)$  и объявив  $u_0(t)$  новой неизвестной, удовлетворяющей уравнению и начальному условию

$$u_0'(t) = 1, \quad u_0(0) = 0,$$

мы сведем задачу к ранее оговоренному случаю.

Итак, пусть  $f = f(u)$ . Перепишем для этого случая интегральное соотношение (14.2)

$$u(t_{n+1}) - u(t_n) = \int_{t_n}^{t_{n+1}} f(u) dt. \quad (15.3)$$

Сделаем в интеграле (15.3) замену переменной интегрирования, полагая

$$(t - t_n)/\tau = \theta. \quad (15.4)$$

Эта замена переводит отрезок  $[t_n, t_{n+1}]$  в  $[0, 1]$  так, что

$$u(t_{n+1}) - u(t_n) = \tau \int_0^1 f(\hat{u}(\theta)) d\theta, \quad (15.5)$$

где

$$\hat{u}(\theta) = u(t(\theta)).$$

Пусть

$$0 \leq \theta_1 < \theta_2 < \dots < \theta_s \leq 1 \quad (15.6)$$

суть узлы, а  $b_1, b_2, \dots, b_s$  — веса некоторой квадратурной формулы, аппроксимирующей  $\int_0^1 \varphi(\theta) d\theta$ . Используя эту формулу для аппроксимации интеграла в (15.5), будем иметь

$$u(t_{n+1}) - u(t_n) \approx \tau \sum_{i=1}^s b_i f(\hat{u}(\theta_i)). \quad (15.7)$$

Чтобы получить из этого соотношения численный метод, нужно точные значения искомого решения заменить на приближенные, а приближенное равенство — наоборот на точное. Но прежде мы должны ввести дополнительные обозначения. Будем обозначать значение приближенного решения в точке  $t$ , отвечающей узлу квадратурной формулы  $\theta_i$  ( $t = t_n + \tau\theta_i$ ) через  $Y_i$ . Тогда искомое уравнение примет вид

$$u_{n+1} = u_n + \tau \sum_{i=1}^s b_i f(Y_i). \quad (15.8)$$

Чтобы получить уравнение для определения  $Y_i$ , проинтегрируем (15.2) от  $t_n$  до  $t_n + \tau\theta_i$  и сделаем замену (15.4)

$$\hat{u}(\theta_i) - u(t_n) = \int_{t_n}^{t_n + \tau\theta_i} f(u(t)) dt = \tau \int_0^{\theta_i} f(\hat{u}(\theta)) d\theta.$$



Заменим и здесь интеграл квадратурной формулой с теми же узлами (15.6). Эта квадратурная формула будет несколько своеобразной, ибо не все ее узлы будут лежать на отрезке интегрирования. Разумеется, ее веса, вообще говоря, должны быть отличны от  $b_j$  и даже быть своими для каждого  $i$ . Пусть

$$Y_i = u_n + \tau \sum_{j=1}^s a_{ij} f(Y_j), \quad i = 1, 2, \dots, s. \quad (15.9)$$

Соотношения (15.8), (15.9) полностью определяют численный метод.

Итак, для того, чтобы найти приближенное решение  $u_{n+1}$  (когда  $u_n$  уже найдено), сначала нужно решить, вообще говоря, нелинейную систему (15.9) и определить  $Y_i$ ,  $i = 1, 2, \dots, s$ , которые затем следует подставить в (15.8).

**Определение 15.1.** Метод (15.8), (15.9) называется *s-этапным методом Рунге-Кутты*.

Этот метод принято записывать таблицей его коэффициентов, которая называется *таблицей Бутчера*

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \dots & a_{1s} \\ c_2 & a_{21} & a_{22} & \dots & a_{2s} \\ & \dots & \dots & \dots & \dots \\ c_s & a_{s1} & a_{s2} & \dots & a_{ss} \\ \hline & b_1 & b_2 & \dots & b_s \end{array} \quad c_i = \sum_{j=1}^s a_{ij}. \quad (15.10)$$

**Замечание 15.1.** Поскольку  $b_i$  суть весовые коэффициенты квадратурной формулы для интеграла по единичному отрезку, то  $\sum_{i=1}^s b_i = 1$ . Из

аналогичных соображений  $c_i = \sum_{j=1}^s a_{ij} = \theta_i$ .

**Замечание 15.2.** Если бы вместо (15.2) рассматривалось уравнение (15.1), то с учетом замечания 15.1 соотношения (15.8) и (15.9) приняли бы вид

$$u_{n+1} = u_n + \tau \sum_{i=1}^s b_i f(t_n + \tau c_i, Y_i),$$

$$Y_i = u_n + \tau \sum_{j=1}^s a_{ij} f(t_n + \tau c_j, Y_j), \quad i = 1, 2, \dots, s.$$

**Определение 15.2.** Если в таблице Бутчера (15.10) коэффициенты  $a_{ij} = 0$  при  $j \geq i$ , то метод (15.8), (15.9) называется явным  $s$ -этапным методом Рунге-Кутты.

**Определение 15.3.** Если  $a_{ij} = 0$  при  $i > j$  и хотя бы один  $a_{ii} \neq 0$ , то метод (15.8), (15.9) называется диагонально неявным.

**Определение 15.4.** Если  $a_{ij} = 0$  при  $j > i$ , а  $a_{ii} = a$ ,  $i = 1, \dots, s$ , то метод (15.8), (15.9) называется однократно неявным.

Во всех остальных случаях мы говорим о неявных методах Рунге-Кутты.

Коэффициенты в таблице Бутчера (15.10) *при заданных ограничениях* выбираются из соображений максимальной точности численного метода.

## 15.2 Одноэтапные методы Рунге-Кутты

Исследуем одноэтапные ( $s = 1$ ) методы Рунге-Кутты. При  $s = 1$  соотношения (15.9), (15.8) принимают вид

$$Y_1 = u_n + \tau a_{11} f(Y_1), \quad (15.11)$$

$$u_{n+1} = u_n + \tau b_1 f(Y_1). \quad (15.12)$$

Из соображений аппроксимации (квадратурная формула должна быть точной по крайней мере на  $\text{const}$ ) находим, что  $b_1 = 1$ . Если теперь положить  $a_{11} = 0$ , то метод будет явным, причем  $Y_1 = u_n$ , а (15.12) можно переписать в виде

$$\frac{u_{n+1} - u_n}{\tau} = f(u_n).$$

Мы получили метод Эйлера. Тем самым, метод Эйлера есть *явный одноэтапный метод Рунге-Кутты*.

Если взять  $a_{11} = 1$ , то метод (15.11), (15.12) будет неявным. При этом правые части (15.11) и (15.12) совпадают, что приводит к соотношению  $Y_1 = u_{n+1}$ . В этом случае система (15.11), (15.12) преобразуется к виду

$$\frac{u_{n+1} - u_n}{\tau} = f(u_{n+1}).$$

Это неявный метод Эйлера (14.8). Он также является одноэтапным методом Рунге-Кутты.

Исследуем теперь наиболее целесообразный выбор параметров  $b_1$  и  $a_{11}$  с точки зрения минимизации погрешности аппроксимации. Чтобы найти погрешность аппроксимации, перепишем уравнение (15.12) в виде

$$\frac{u_{n+1} - u_n}{\tau} = b_1 f(Y_1) \quad (15.13)$$

(ср. с (14.7), (14.8), (14.10) и (14.12)), а решение уравнения (15.11) обозначим через  $Y_1(u_n)$ . Если, как и выше,  $z_n = u_n - u(t_n)$ , то

$$\frac{z_{n+1} - z_n}{\tau} = b_1 f(Y_1(u(t_n) + z_n)) - \frac{u(t_{n+1}) - u(t_n)}{\tau}.$$

И снова, раскладывая первое слагаемое правой части по формуле Тейлора, находим, что

$$\begin{aligned} \frac{z_{n+1} - z_n}{\tau} &= b_1 \left[ f(Y_1(u(t_n))) + \frac{\partial f}{\partial u}(\tilde{u})z_n \right] - \frac{u(t_{n+1}) - u(t_n)}{\tau} = \\ &= b_1 \frac{\partial f}{\partial Y_1} \frac{\partial Y_1}{\partial u}(\tilde{u})z_n + \psi_n, \end{aligned}$$

где

$$\psi_n = b_1 f(Y_1(u(t_n))) - \frac{u(t_{n+1}) - u(t_n)}{\tau} \quad (15.14)$$

— погрешность аппроксимации, а  $Y_1(u(t_n))$  — решение уравнения (15.11) с  $u(t_n)$  вместо  $u_n$ , т.е.

$$Y_1(u(t_n)) = u(t_n) + \tau a_{11} f(Y_1(u(t_n))). \quad (15.15)$$

**Замечание 15.3.** Погрешность аппроксимации (15.14) представляет собой разность между правой и левой частями уравнения (15.13), если туда вместо приближенного решения подставить точное (ср. с замечанием 14.3).

Разложим погрешность аппроксимации (15.14) по степеням  $\tau$ . Имеем

$$\psi_n = b_1 \left[ f(Y_1) \Big|_{\tau=0} + \tau \frac{df(Y_1)}{d\tau} \Big|_{\tau=0} + \frac{\tau^2}{2} \frac{d^2 f}{d\tau^2} \Big|_{\tau=0} \right] - \left[ u'(t_n) + \frac{\tau}{2} u''(t_n) + \frac{\tau^2}{6} \tilde{u}''' \right].$$

Из (15.15) следует, что  $Y_1|_{\tau=0} = u(t_n)$  и поэтому

$$f(Y_1) \Big|_{\tau=0} = f(u(t_n)).$$

Снова с использованием (15.15) находим, что

$$\frac{df(Y_1)}{d\tau} \Big|_{\tau=0} = \frac{df}{dY_1} \frac{dY_1}{d\tau} \Big|_{\tau=0} = \frac{df}{du}(u(t_n)) a_{11} f(u(t_n)),$$

а из уравнения (15.2)

$$u'(t_n) = f(u(t_n)), \quad u''(t_n) = \frac{df}{dt}(u(t_n)) = \frac{df}{du}(u(t_n)) \frac{du}{dt}(t_n) = \frac{df}{du} f.$$

Поэтому

$$\psi_n = (b_1 - 1)f(u(t_n)) + \tau \left[ b_1 a_{11} - \frac{1}{2} \right] f(u(t_n)) \frac{df}{du}(u(t_n)) + O(\tau^2).$$

Тем самым, для того, чтобы погрешность аппроксимации была  $O(\tau^2)$ , необходимо и достаточно, чтобы выполнялись условия

$$b_1 = 1, \quad a_{11} b_1 = 1/2. \quad (15.16)$$

Отсюда находим

$$b_1 = 1, \quad a_{11} = 1/2$$

и, следовательно, неявный одноэтапный метод Рунге-Кутты

$$\begin{aligned} Y_1 &= u_n + \frac{\tau}{2} f(Y_1), \\ u_{n+1} &= u_n + \tau f(Y_1) \end{aligned} \quad (15.17)$$

имеет второй порядок аппроксимации.

**Замечание 15.4.** Из первого уравнения (15.17) следует, что момент времени, на который  $Y_1$  приближает  $u(t)$ , есть  $t + \tau/2$ , ибо для задачи  $u' = 1$ ,  $u(0) = 0$ , имеющей решение  $u = t$ ,  $Y_1 = u_n + \tau/2 = t_n + \tau/2$ . Тем самым, для уравнения (15.1) метод (15.17) принимает вид (ср. с замечанием 15.2)

$$\begin{aligned} Y_1 &= u_n + \frac{\tau}{2} f(t_{n+1/2}, Y_1), \\ u_{n+1} &= u_n + \tau f(t_{n+1/2}, Y_1). \end{aligned}$$

**Упражнение 15.1.** Показать, что второе из соотношений метода (15.17) получается использованием в (15.7) одноточечной формулы Гаусса (узел — полинома Лежандра первой степени на  $[0, 1]$ , а весовой коэффициент квадратурной формулы — интеграл по  $[0, 1]$  соответствующего весового коэффициента интерполяционного многочлена). При этом для получения первого соотношения (15.17) используется квадратурная формула (с тем же узлом) с весовым коэффициентом, полученным интегрированием по  $[0, \theta_1]$  того же весового коэффициента интерполяционного многочлена.

Соотношения (15.17) можно преобразовать. Исключив  $f(Y_1)$ , найдем, что  $u_{n+1} = 2Y_1 - u_n$ . Выражая отсюда  $Y_1$  и подставляя его во второе уравнение (15.17), получим

$$u_{n+1} = u_n + \tau f\left(\frac{u_{n+1} + u_n}{2}\right).$$

**Замечание 15.5.** Метод (15.17) сильно напоминает метод Рунге (14.10), (14.11). Отличие между ними состоит в том, что здесь промежуточное значение находится по неявной формуле, а в методе Рунге по явной формуле (14.11). Метод (15.17), как мы уже сказали, является одноэтапным (неявным) методом Рунге-Кутты, а метод (14.10), (14.11) — двухэтапным (явным) методом. Подчеркнем, что слову *этап* здесь мы придаем четкий математический смысл.

### 15.3 Методы третьего порядка аппроксимации

Выясним ограничения на коэффициенты таблицы Бутчера (15.10), обеспечивающие третий порядок аппроксимации  $s$ -этапного метода Рунге-Кутты. Для этого нужно исследовать погрешность аппроксимации

$$\psi_n := \psi_n(\tau) := \sum_{i=1}^s b_i f(Y_i(u(t_n))) - \frac{u(t_{n+1}) - u(t_n)}{\tau},$$

где

$$Y_i(u(t_n)) = u(t_n) + \tau \sum_{j=1}^s a_{ij} f(Y_j(u(t_n))) =: Y_i(u(t_n); \tau), \quad i = 1, 2, \dots, s. \quad (15.18)$$

Напомним вывод представления для  $\psi_n$ . Из (15.8)

$$\frac{u_{n+1} - u_n}{\tau} = \sum_{i=1}^n b_i f(Y_i) =: \sum_{i=1}^s b_i f(Y_i(u_n)).$$

Поскольку  $u_n = u(t_n) + z_n$ , то

$$\begin{aligned} \frac{z_{n+1} - z_n}{\tau} &= \sum_{i=1}^s b_i f(Y_i(u(t_n) + z_n)) - \frac{u(t_{n+1}) - u(t_n)}{\tau} = \\ &= \sum_{i=1}^s b_i f(Y_i(u(t_n))) + z_n \sum_{i=1}^s b_i \frac{df(Y_i(u))}{du} \Big|_{u=u(t_n)+\sigma_i z_n} - \frac{u(t_{n+1}) - u(t_n)}{\tau} = \\ &= \left( \sum_{i=1}^s b_i \frac{df(Y_i(u))}{du} \Big|_{u=u(t_n)+\sigma_i z_n} \right) z_n + \psi_n. \end{aligned}$$

Раскладывая  $\psi_n(\tau)$  по  $\tau$  до четвертого порядка, будем иметь

$$\begin{aligned} \psi_n(\tau) &= \sum_{i=1}^s b_i \left[ f(Y_i) \Big|_{\tau=0} + \tau \frac{df(Y_i)}{d\tau} \Big|_{\tau=0} + \frac{\tau^2}{2} \frac{d^2 f(Y_i)}{d\tau^2} \Big|_{\tau=0} + \frac{\tau^3}{6} \frac{d^3 f(Y_i)}{d\tau^3} \Big|_{\tau=0} + \right. \\ &\left. + O(\tau^4) \right] - \left[ u'(t_n) + \frac{\tau}{2} u''(t_n) + \frac{\tau^2}{6} u'''(t_n) + \frac{\tau^3}{24} u^{IV}(t_n) + O(\tau^4) \right]. \end{aligned} \quad (15.19)$$

Поскольку  $f(Y_i(u(t_n)))$  есть сложная функция  $\tau$ , то вычислим сначала производные по  $\tau$  функции  $Y_i(u(t_n); \tau)$  при  $\tau = 0$ . Из (15.18) с учетом (15.10), находим, что

$$\begin{aligned} Y_i \Big|_{\tau=0} &= u(t_n), \\ Y_i' \Big|_{\tau=0} &= \frac{dY_i}{d\tau} \Big|_{\tau=0} = \left[ \sum_{j=1}^s a_{ij} f(Y_j) + \tau \sum_{j=1}^s a_{ij} \frac{df}{dY_j} Y_j' \right] \Big|_{\tau=0} = f(u(t_n)) c_i, \\ Y_i'' \Big|_{\tau=0} &= \left[ 2 \sum_{j=1}^s a_{ij} \frac{df}{dY_j} Y_j' + \tau \sum_{j=1}^s a_{ij} \frac{d^2 f}{dY_j^2} Y_j' Y_j' + \tau \sum_{j=1}^s a_{ij} \frac{df}{dY_j} Y_j'' \right] \Big|_{\tau=0} = \\ &= 2 \frac{df}{du} f(u(t_n)) \sum_{j=1}^s a_{ij} c_j. \end{aligned}$$

Теперь можно найти производные  $f$ :

$$\begin{aligned} f(Y_i(u(t_n))) \Big|_{\tau=0} &= f(u(t_n)), \\ \frac{df(Y_i)}{d\tau} \Big|_{\tau=0} &= \frac{df}{dY_i} Y_i' \Big|_{\tau=0} = \frac{df}{du} f(u(t_n)) c_i, \\ \frac{d^2 f(Y_i)}{d\tau^2} \Big|_{\tau=0} &= \left[ \frac{d^2 f}{dY_i^2} (Y_i')^2 + \frac{df}{dY_i} Y_i'' \right] \Big|_{\tau=0} = \\ &= \frac{d^2 f}{du^2} f^2(u(t_n)) c_i^2 + 2 \left( \frac{df}{du} \right)^2 f(u(t_n)) \sum_{j=1}^s a_{ij} c_j. \end{aligned}$$

Далее, из (15.2)

$$\begin{aligned} u' &= f, \quad u'' = \frac{df}{du} u' = f' f, \quad u''' = f'' u' f + (f')^2 u' = f'' f^2 + (f')^2 f \\ u^{IV} &= f''' u' f^2 + f'' 2f f' u' + 2f' f'' u' f + (f')^3 u' = \\ &= f''' f^3 + 4f'' f' f^2 + (f')^3 f \end{aligned}$$

и, следовательно,

$$\frac{u(t_{n+1}) - u(t_n)}{\tau} = f + \frac{\tau}{2} f' f + \frac{\tau^2}{6} (f'' f^2 + (f')^2 f) + O(\tau^3). \quad (15.20)$$

**Замечание 15.6.** Если бы правая часть  $f$  уравнения в (15.1) явно зависела от  $t$ , то вместо (15.20) при разложении до  $O(\tau^4)$  мы бы имели

$$\begin{aligned} \frac{u(t_{n+1}) - u(t_n)}{\tau} &= \\ &= f + \frac{\tau}{2} (f_t + f_u f) + \frac{\tau^2}{6} (f_{tt} + 2f_{tu} f + f_{ut} f + f_{uu} f f + f_u f_u f) + \\ &+ \frac{\tau^3}{24} (f_{ttt} + 3f_{ttu} f + 3f_{tuu} f f + 3f_{uu} f_t f + 3f_{tu} f_t f + 3f_{tu} f_u f + 2f_u f_{tu} f + \\ &+ f_u f_u f_t + f_u f_{tt} + f_{uuu} f f f + 3f_{uu} f_u f f + f_u f_{uu} f f + f_u f_u f_u f) + O(\tau^4). \end{aligned}$$

Подставляя теперь найденные разложения в (15.19), будем иметь

$$\begin{aligned} \psi_n &= \sum_{i=1}^s b_i \left[ f + \tau f' f c_i + \frac{\tau^2}{2} \left( f'' f^2 c_i^2 + 2f'^2 f \sum_{j=1}^s a_{ij} c_j \right) \right] - \\ &- \left[ f + \frac{\tau}{2} f' f + \frac{\tau^2}{6} (f'' f^2 + f'^2 f) \right] + O(\tau^3). \end{aligned} \quad (15.21)$$

Отсюда, приравнявая нулю коэффициенты при различных степенях  $\tau$ , находим, что условия третьего порядка аппроксимации суть

$$\begin{aligned} \sum_{i=1}^s b_i &= 1, \\ \sum_{i=1}^s b_i c_i &= \frac{1}{2}, \end{aligned} \tag{15.22}$$

$$\begin{aligned} \sum_{i=1}^s b_i c_i^2 &= \frac{1}{3}, \\ \sum_{i,j=1}^s b_i a_{ij} c_j &= \frac{1}{6} \quad (b^T A c = \frac{1}{6}). \end{aligned} \tag{15.23}$$

При этом (15.22) суть условия второго порядка аппроксимации (ср. с (15.16)).

**Замечание 15.7.** Чтобы иметь условия четвертого порядка аппроксимации, к условиям (15.22), (15.23) нужно добавить следующие условия:

$$\begin{aligned} \sum_{i=1}^s b_i c_i^3 &= \frac{1}{4}, \\ \sum_{i,j=1}^s b_i c_i a_{ij} c_j &= \frac{1}{8}, \\ \sum_{i,j=1}^s b_i a_{ij} c_j^2 &= \frac{1}{12}, \\ \sum_{i,j,k=1}^s b_i a_{ij} a_{jk} c_k &= \frac{1}{24}. \end{aligned} \tag{15.24}$$

**Замечание 15.8.** Число условий, накладываемых на коэффициенты  $b_i$  и  $a_{ij}$  метода (15.8), (15.9) (на элементы таблицы Бутчера (15.10)) для получения метода порядка  $p$  приведено в нижеследующей таблице.

Таблица 1.

|               |   |   |   |   |    |    |    |     |     |      |
|---------------|---|---|---|---|----|----|----|-----|-----|------|
| Порядок $p$   | 1 | 2 | 3 | 4 | 5  | 6  | 7  | 8   | 9   | 10   |
| Число условий | 1 | 2 | 4 | 8 | 17 | 37 | 85 | 200 | 486 | 1205 |



**Замечание 15.9.** Условия (15.22) с учетом замечания 15.1 можно трактовать как условия точности квадратурной формулы из (15.7) на линейных функциях.<sup>1</sup> Добавление к этим условиям первого из соотношений (15.23), а затем и первого из соотношений (15.24) на указанную квадратурную формулу накладывает дополнительные условия точности на квадратичных и кубических функциях.

**Упражнение 15.2.** Показать, что метод трапеций (14.12) является неявным двухэтапным методом Рунге-Кутты второго порядка аппроксимации. (Найти все  $b_i$ ,  $a_{ij}$  и показать невыполнение хотя бы одно из условий (15.23))

Ответ:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array} \quad \left( \frac{1}{2}0^2 + \frac{1}{2}1^2 \right) \neq \frac{1}{3}.$$

**Упражнение 15.3.** Показать, что метод Рунге (14.10), (14.11) является явным двухэтапным методом Рунге-Кутты второго порядка.

Ответ:

$$\begin{array}{c|cc} 1/2 & 1/2 & \\ \hline & 0 & 1 \end{array} \quad \left[ 0 \cdot 0 + 1 \cdot \frac{1}{4} \right] \neq \frac{1}{3}.$$

## 15.4 Двухэтапные неявные методы третьего порядка

Положим в (15.22), (15.23) параметр  $s = 2$ . В результате система примет вид

$$\begin{aligned} b_1 + b_2 &= 1, \\ c_1 b_1 + c_2 b_2 &= 1/2, \\ c_1^2 b_1 + c_2^2 b_2 &= 1/3, \\ b_1(a_{11}c_1 + a_{12}c_2) + b_2(a_{21}c_1 + a_{22}c_2) &= 1/6. \end{aligned} \tag{15.25}$$

Эта система содержит четыре уравнения и шесть неизвестных (Если не считать  $c_1$  и  $c_2$ , задаваемые (15.10)). Поэтому, вообще говоря, два из этих неизвестных должны остаться свободными, а остальные выразиться через них. Система (15.25) нелинейная, а для нелинейных систем, вообще говоря, нет регулярных способов отыскания точного решения. Однако,

<sup>1</sup>Ведь  $c_i = \theta_i$ , т.е. координата переменной интегрирования в  $i$ -ом узле.

система (15.25) может быть решена точно. Укажем один из путей, приводящих к решению этой системы.

Для отыскания решения системы (15.25) предположим сначала, что неизвестные  $c_1$  и  $c_2$  найдены, и рассмотрим первые три уравнения (15.25) как систему линейных уравнений относительно  $b_1$  и  $b_2$ . Поскольку эта система переопределена, то для ее разрешимости необходимо обращение в нуль определителя расширенной матрицы, которая является квадратной,

$$\begin{aligned} \begin{vmatrix} 1 & 1 & 1 \\ c_1 & c_2 & 1/2 \\ c_1^2 & c_2^2 & 1/3 \end{vmatrix} &= \frac{1}{3}c_2 + \frac{1}{2}c_1^2 + c_1c_2^2 - c_1^2c_2 - \frac{1}{2}c_2^2 - \frac{1}{3}c_1 = \\ &= (c_2 - c_1) \left[ \frac{1}{3} - \frac{c_1 + c_2}{2} + c_1c_2 \right] = 0. \end{aligned} \quad (15.26)$$

Проанализируем это соотношение. Если бы  $c_1 = c_2$ , то последнее уравнение (15.25) приняло бы вид

$$c_1^2b_1 + c_2^2b_2 = 1/6,$$

что противоречит третьему уравнению (15.25), и поэтому

$$c_1 - c_2 \neq 0. \quad (15.27)$$

Тем самым, из (15.26) и (15.27) следует, что единственным условием разрешимости относительно  $b_1$  и  $b_2$  системы первых трех уравнений (15.25) является условие

$$2 - 3(c_1 + c_2) + 6c_1c_2 = 0$$

или

$$(3 - 6c_1)c_2 = 2 - 3c_1.$$

Поскольку  $c_1 = 1/2$  не удовлетворяет этому уравнению, то

$$c_1 \neq 1/2 \quad (15.28)$$

и можно найти

$$c_2 = \frac{2 - 3c_1}{3(1 - 2c_1)}. \quad (15.29)$$

Разрешим теперь первые два уравнения (15.25) относительно  $b_1$  и  $b_2$  при помощи формул Крамера. Будем иметь

$$\Delta = \begin{vmatrix} 1 & 1 \\ c_1 & c_2 \end{vmatrix} = c_2 - c_1 \neq 0, \quad \Delta_1 = \begin{vmatrix} 1 & 1 \\ 1/2 & c_2 \end{vmatrix} = c_2 - \frac{1}{2}, \quad \Delta_2 = \begin{vmatrix} 1 & 1 \\ c_1 & 1/2 \end{vmatrix} = \frac{1}{2} - c_1,$$

и, следовательно,

$$b_1 = \frac{c_2 - 1/2}{c_2 - c_1} = \frac{1}{4(3c_1^2 - 3c_1 + 1)}, \quad b_2 = \frac{3(1 - 2c_1)^2}{4(3c_1^2 - 3c_1 + 1)}. \quad (15.30)$$

Из (15.29), (15.30) следует, что  $c_1$  можно принять за параметр. В качестве второго параметра возьмем  $a_{12}$ . Тогда

$$a_{11} = c_1 - a_{12}. \quad (15.31)$$

Поскольку

$$a_{21} = c_2 - a_{22}, \quad (15.32)$$

то, подставляя (15.31), (15.32) в последнее из уравнений (15.25), получим

$$b_1[(c_1 - a_{12})c_1 + a_{12}c_2] + b_2[(c_2 - a_{22})c_1 + a_{22}c_2] = 1/6.$$

Принимая во внимание (15.30) и второе из уравнений (15.25), после разрешения полученного соотношения относительно  $a_{22}$ , будем иметь

$$a_{22} = \frac{1/6 - c_1/2 - a_{12}(c_2 - 1/2)}{1/2 - c_1} = \frac{(1 - 3c_1)(1 - 2c_1) - a_{12}}{3(1 - 2c_1)^2}. \quad (15.33)$$

Теперь из (15.32), (15.29) и (15.33) находим, что

$$a_{21} = \frac{1 - 2c_1 + a_{12}}{3(1 - 2c_1)^2}. \quad (15.34)$$

Соотношения (15.29), (15.30), (15.31), (15.33), (15.34) задают двухпараметрическое семейство неявных двухэтапных методов Рунге-Кутты третьего порядка:

$$\begin{array}{c|cc} c_1 & c_1 - a_{12} & a_{12} \\ \hline \frac{2 - 3c_1}{3(1 - 2c_1)} & \frac{1 - 2c_1 + a_{12}}{3(1 - 2c_1)^2} & \frac{(1 - 3c_1)(1 - 2c_1) - a_{12}}{3(1 - 2c_1)^2} \\ \hline & \frac{1}{4(3c_1^2 - 3c_1 + 1)} & \frac{3(1 - 2c_1)^2}{4(3c_1^2 - 3c_1 + 1)} \end{array} \quad (15.35)$$

Если положить,  $a_{12} = 0$ , то получим однопараметрическое семейство *диагонально неявных* двухэтапных методов третьего порядка. Полагая теперь  $c_1 \equiv a_{11} \equiv a_{22}$ , из (15.33) для  $c_1$  получим квадратное уравнение  $6c_1^2 - 6c_1 + 1$  с корнями  $c_1 = \gamma = (3 \pm \sqrt{3})/6$ . Таблица Бутчера этого двухэтапного *однократно неявного* метода третьего порядка имеет вид

$$\begin{array}{c|cc} \theta_1 = \gamma & \gamma & 0 \\ \theta_2 = 1 - \gamma & 1 - 2\gamma & \gamma \\ \hline & 1/2 & 1/2 \end{array} \quad \gamma = \frac{3 \pm \sqrt{3}}{6}. \quad (15.36)$$

При других значениях параметра  $\gamma$  метод (15.36) имеет погрешность аппроксимации  $O(\tau^2)$

## 15.5 ЯВНЫЕ ДВУХЭТАПНЫЕ МЕТОДЫ

В силу определения для явного двухэтапного метода  $a_{11} = a_{12} = a_{22} = 0$  и лишь  $a_{21} \neq 0$ .

**Упражнение 15.4.** Доказать, что при всех  $a_{ij} = 0$  двухэтапный метод вырождается в одноэтапный.

Поскольку двухэтапные методы третьего порядка имеют лишь два свободных параметра, а мы задали три, то рассчитывать на третий порядок у явных двухэтапных методов, вообще говоря, не приходится. Покажем, что так оно и есть.

Принимая  $a_{21}$  за параметр, из условий второго порядка аппроксимации (15.22), которые в нашем случае принимают вид

$$b_1 + b_2 = 1, \quad a_{21}b_2 = 1/2,$$

находим

$$b_1 = \left(1 - \frac{1}{2a_{21}}\right), \quad b_2 = \frac{1}{2a_{21}}. \quad (15.37)$$

Тем самым, явные двухэтапные методы Рунге-Кутты второго проядка образуют однопараметрическое семейство.

Далее, поскольку в рассматриваемом случае наряду с  $a_{11}$ ,  $a_{12}$ ,  $a_{22}$  и  $c_1 = 0$ , то левая часть четвертого из условий (15.25) обращается в нуль и следовательно это условие выполненным быть не может. Мы доказали, что явных двухэтапных методов третьего порядка не существует.

Из (15.21) следует, что погрешность аппроксимации двухэтапного явного метода Рунге-Кутты второго порядка (15.37) имеет вид

$$\psi_n = \tau^2 \left\{ \frac{1}{4} \left( a_{21} - \frac{2}{3} \right) f^2 f'' - \frac{1}{6} f f'^2 \right\} + O(\tau^3).$$

Поэтому при  $a_{21} = 2/3$  главный член погрешности упрощается, и погрешность принимает вид

$$\psi_n = -\frac{\tau^2}{6} f f'^2 + O(\tau^3).$$

Сам же метод в этом случае задается таблицей Бутчера

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 2/3 & 2/3 & 0 \\ \hline & 1/4 & 3/4 \end{array} .$$

Если же положить  $a_{21} = 1$ , то получим *метод Хойна*

$$\begin{aligned} Y_1 &= u_n, & Y_2 &= u_n + \tau f(Y_1), \\ u_{n+1} &= u_n + \frac{\tau}{2}[f(Y_1) + f(Y_2)] \end{aligned}$$

с таблицей Бутчера

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 1/2 & 1/2 \end{array} .$$

При  $a_{21} = 1/2$  приходим к уже известному нам методу Рунге (14.10), (14.11).

**Упражнение 15.5.** Выписать все построенные методы второго порядка и главные члены их погрешностей аппроксимации.

## 15.6 Двухэтапный метод четвертого порядка

Коэффициенты метода четвертого порядка должны удовлетворять еще четырем условиям (15.24). Хотя у двухэтапного метода третьего порядка осталось только два параметра, существует единственный двухэтапный метод четвертого порядка. Его коэффициенты суть

$$\begin{array}{c|cc} 1/2 - \sqrt{3}/6 & 1/4 & 1/4 - \sqrt{3}/6 \\ 1/2 + \sqrt{3}/6 & 1/4 + \sqrt{3}/6 & 1/4 \\ \hline & 1/2 & 1/2 \end{array} \quad (15.38)$$

**Замечание 15.10.** Если бы  $f$  не зависела от  $u$  ( $u' = f(t)$ ,  $u_{n+1} = u_n + \tau \int_0^1 \hat{f}(\theta) d\theta$ ), то двухэтапный метод четвертого порядка получился бы, только если квадратура в (15.8) была квадратурой Гаусса, т.е.  $b_1 = b_2 = 1/2$ , а  $\theta_1$  и  $\theta_2$  — сдвинутые на  $[0, 1]$  нули полинома Лежандра второй степени  $P_2(x) = \frac{1}{2}(3x^2 - 1)$ . Корнями этого полинома являются числа  $x_{1,2} = \pm 1/\sqrt{3}$ . Делая линейную замену, переводящую отрезок  $[-1, 1]$  в

отрезок  $[0, 1]$ , находим, что узлы квадратуры Гаусса на  $[0, 1]$  суть  $\theta_{1,2} = 1/2 \mp \sqrt{3}/6$ , как в (15.38). Остальные коэффициенты получаются, если проинтегрировать по  $[0, \theta_1]$  и  $[0, \theta_2]$  весовые функции интерполяционного полинома Лагранжа с гауссовыми узлами.

**Упражнение 15.6.** Доказать, что в (15.38)

$$a_{ij} = \int_0^{\theta_j} p_i(\theta) d\theta,$$

где  $p_i(\theta)$  — линейная функция (интерполянты по двум узлам) такая, что  $p_i(\theta_i) = 1$ ,  $p_i(\theta_j) = 0$  при  $i \neq j$ . Убедиться в выполнении условий (15.22)–(15.24).

## 15.7 Явные трехэтапные методы Рунге-Кутты третьего порядка

Рассмотрим более подробно явные трехэтапные методы. В силу определения

$$a_{11} = a_{12} = a_{13} = a_{22} = a_{23} = a_{33} = 0,$$

и указанные методы задаются таблицей

|       |          |          |       |
|-------|----------|----------|-------|
| $c_2$ | $a_{21}$ |          |       |
| $c_3$ | $a_{31}$ | $a_{32}$ |       |
|       | $b_1$    | $b_2$    | $b_3$ |

Условия третьего порядка аппроксимации (15.22), (15.23) в рассматриваемом случае принимают вид

$$\begin{aligned} b_1 + b_2 + b_3 &= 1, \\ b_2 c_2 + b_3 c_3 &= 1/2, \\ b_2 c_2^2 + b_3 c_3^2 &= 1/3, \\ b_3 a_{32} c_2 &= 1/6. \end{aligned} \tag{15.39}$$

Эта система имеет два однопараметрических семейства решений и одно двухпараметрическое. Найдем их. Будем рассматривать второе и третье уравнения системы (15.39) как линейную систему относительно  $b_2$  и  $b_3$ . Эта система может быть как вырожденной (и это приводит к двум однопараметрическим семействам решений), так и невырожденной (двухпараметрическое семейство).

Пусть эта система вырождена, т.е.

$$\begin{vmatrix} c_2 & c_3 \\ c_2^2 & c_3^2 \end{vmatrix} = c_2 c_3 (c_3 - c_2) = 0. \quad (15.40)$$

В силу последнего из уравнений (15.39)  $c_2 \neq 0$ , и поэтому либо

$$c_3 = 0 \quad (15.41)$$

либо

$$c_2 = c_3. \quad (15.42)$$

i) Пусть сначала имеет место (15.41). Тогда второе и третье уравнения (15.39) принимают вид

$$c_2 b_2 = 1/2, \quad c_2^2 b_2 = 1/3$$

и, следовательно,

$$c_2 = 2/3, \quad b_2 = 3/4.$$

Если теперь  $b_3 = b$  принять за параметр, то из последнего уравнения (15.39) находим

$$a_{32} = \frac{1}{4b}.$$

Поскольку  $c_3 = 0$ , то

$$a_{31} = -a_{32} = -\frac{1}{4b}$$

и, наконец, из первого уравнения (15.39)

$$b_1 = \frac{1}{4} - b.$$

Собирая найденные значения, получим таблицу Бутчера

$$\begin{array}{c|ccc} 2/3 & 2/3 & & \\ 0 & -\frac{1}{4b} & \frac{1}{4b} & \\ \hline & 1/4 - b & 3/4 & b \end{array} \quad (15.43)$$

ii) Теперь пусть имеет место (15.42). Снова из второго и третьего уравнений (15.39) находим, что

$$b_2 + b_3 = \frac{1}{2c_2} = \frac{1}{3c_2^2},$$

т.е.

$$a_{21} = c_2 = c_3 = 2/3, \quad b_2 = 3/4 - b,$$

где  $b = b_3$  — параметр. Из последнего уравнения (15.39)

$$a_{32} = \frac{1}{4b},$$

а из первого уравнения

$$b_1 = 1/4.$$

Наконец,

$$a_{31} = \frac{2}{3} - \frac{1}{4b}.$$

Таблица Бутчера рассматриваемого метода имеет вид

$$\begin{array}{c|ccc} 2/3 & 2/3 & & \\ 2/3 & \frac{2}{3} - \frac{1}{4b} & \frac{1}{4b} & \\ \hline & 1/4 & 3/4 - b & b \end{array} \quad (15.44)$$

iii) Если соотношение (15.40) места не имеет, то из второго и третьего уравнений (15.39) находим

$$b_2 = \frac{c_3/2 - 1/3}{c_2(c_3 - c_2)}, \quad b_3 = \frac{1/3 - c_2/2}{c_3(c_3 - c_2)}. \quad (15.45)$$

Привлекая первое уравнение (15.39), найдем, что

$$b_1 = 1 - \frac{3(c_2 + c_3) - 2}{6c_2c_3}, \quad (15.46)$$

а из четвертого

$$a_{32} = \frac{c_3(c_3 - c_2)}{c_2(2 - 3c_2)}. \quad (15.47)$$

Таблица Бутчера этого метода такова

$$\begin{array}{c|cc} c_2 & c_2 & \\ c_3 & \frac{c_3(3c_2 - 3c_2^2 - c_3)}{c_2(2 - 3c_3)} & \frac{c_3(c_3 - c_2)}{c_2(2 - 3c_2)} \\ \hline & \frac{6c_2c_3 - 3(c_2 + c_3) + 2}{6c_2c_3} & \frac{3c_3 - 2}{6c_2(c_3 - c_2)} \quad \frac{2 - 3c_2}{6c_3(c_3 - c_2)} \end{array}$$



Среди явных трехэтапных методов Рунге-Кутты третьего порядка в силу исторических причин наибольшей популярностью пользуются следующие методы из семейства (15.45)-(15.47)

$$\begin{array}{c|cc} 1/2 & 1/2 & \\ \hline 1 & -1 & 2 \\ \hline & 1/6 & 2/3 & 1/6 \end{array}, \quad \begin{array}{c|cc} 1/3 & 1/3 & \\ \hline 2/3 & 0 & 2/3 \\ \hline & 1/4 & 0 & 3/4 \end{array}. \quad (15.48)$$

## 15.8 Более общие методы Рунге-Кутты

Приведем без доказательства несколько теорем о свойствах методов Рунге-Кутты.

**Теорема 15.1.** *Не существует явного  $s$ -этапного метода Рунге-Кутты порядка  $p$ , если  $p > s$ .*

Для  $s = 1, 2$  эта теорема нами фактически доказана.

**Теорема 15.2.** *При  $s \geq 5$  не существует явного  $s$ -этапного метода Рунге-Кутты порядка  $p = s$  (1963г.). При  $s \geq 8$  не существует явного  $s$ -этапного метода Рунге-Кутты порядка  $p = s - 1$  (1965г.). При  $s \geq 10$  — порядка  $p = s - 2$  (1985г.).*

**Теорема 15.3.** *Для данного  $p$  существует явный метод Рунге-Кутты порядка  $p$  с  $s$  этапами, если*

$$s = \begin{cases} (3p^2 - 10p + 24)/8, & p - \text{четное,} \\ (3p^2 - 4p + 9)/8, & p - \text{нечетное.} \end{cases}$$

**Замечание 15.11.** При  $s = 6$  существует явный метод Рунге-Кутты порядка 5. При  $s = 7$  существует явный метод порядка 6. Наивысший порядок, фактически достигнутый для явно построенных явных методов Рунге-Кутты равен 10. При этом число этапов равно 17.

**Теорема 15.4.** *При любом  $s$  существует единственный неявный метод Рунге-Кутты порядка  $p = 2s$ .*

**Замечание 15.12.** Для оптимального метода порядка  $p = 2s$  узлы  $\theta_j$  и веса  $b_j$  суть узлы и веса квадратурной формулы Гаусса, а

$$a_{ij} = \int_0^{\theta_j} p_i(\theta) d\theta,$$

где  $p_j(\theta)$  — многочлен степени  $s$  такой, что  $p_i(\theta_i) = 1$ ,  $p_i(\theta_j) = 0$  при  $i \neq j$ .

Таблица 2. Минимальное число этапов для явных методов различных порядков

| порядок<br>$p$ | минимальное число этапов |     |                 |
|----------------|--------------------------|-----|-----------------|
|                | нижняя граница           | $s$ | верхняя граница |
| 1              |                          | 1   |                 |
| 2              |                          | 2   |                 |
| 3              |                          | 3   |                 |
| 4              |                          | 4   |                 |
| 5              |                          | 6   |                 |
| 6              |                          | 7   |                 |
| 7              |                          | 9   |                 |
| 8              |                          | 11  |                 |
| 9              | 12                       |     | 17              |
| 10             | 13                       |     | 17              |

В вычислительной практике широко используется следующий, очень простой, явный 4-этапный метод 4-го порядка, предложенный в работе Кутты 1901 г.

$$\begin{array}{c|ccc}
 1/2 & 1/2 & & \\
 1/2 & 0 & 1/2 & \\
 1 & 0 & 0 & 1 \\
 \hline
 & 1/6 & 1/3 & 1/3 & 1/6.
 \end{array}$$

**Замечание 15.13.** В большинстве работ, посвященных методам Рунге-Кутты, вместо переменных  $Y_j$  фигурируют  $k_j = f(Y_j)$ . Поэтому, вместо (15.8), (15.9) пишут

$$\begin{aligned}
 k_i &= f\left(u_n + \tau \sum_{j=1}^s a_{ij} k_j\right), \\
 u_{n+1} &= u_n + \tau \sum_{j=1}^s b_j k_j.
 \end{aligned} \tag{15.49}$$

## 15.9 Сходимость методов Рунге-Кутты

Установим оценку погрешности приближенного решения, получаемого при помощи того или иного метода Рунге-Кутты.

Если, как и раньше,

$$z_n = u_n - u(t_n),$$

то из (15.8) находим, что  $z_{n+1}$  удовлетворяет уравнению

$$\frac{z_{n+1} - z_n}{\tau} = \left( \sum_{i=1}^s b_i \frac{df(Y_i(u))}{du} \Big|_{u=u(t_n)+\sigma_i z_n} \right) z_n + \psi_n. \quad (15.50)$$

Прежде чем оценивать  $z_{n+1}$ , оценим коэффициент при  $z_n$  в правой части (15.50). Будем при этом предполагать, что

$$\max_{|u|<\infty} \left| \frac{df(u)}{du} \right| \leq L. \quad (15.51)$$

Тогда

$$\max_{|u|<\infty} \left| \frac{df(Y_j(u))}{du} \right| = \max_{|u|<\infty} \left| \frac{df}{dY_j} \frac{dY_j}{du} \right| \leq L \max_{|u|<\infty} \left| \frac{dY_j}{du} \right|. \quad (15.52)$$

Оценим  $|dY_j/du|$ . Из (15.9) с  $u$  вместо  $u_n$

$$\frac{dY_i}{du} = 1 + \tau \sum_{j=1}^s a_{ij} \frac{df}{dY_j} \frac{dY_j}{du}.$$

Пусть  $Y_{i_0}$  таково, что

$$\max_{|u|<\infty} \left| \frac{dY_{i_0}}{du} \right| = \max_{1 \leq j \leq s} \max_{|u|<\infty} \left| \frac{dY_j}{du} \right|.$$

Тогда с учетом (15.51)

$$\max_{|u|<\infty} \left| \frac{dY_{i_0}}{du} \right| \leq 1 + \tau \sum_{j=1}^s |a_{i_0 j}| L \max_{|u|<\infty} \left| \frac{dY_j}{du} \right| \leq 1 + \tau a s L \max_{|u|<\infty} \left| \frac{dY_{i_0}}{du} \right|,$$

где

$$a = \max_{ij} |a_{ij}|, \quad (15.53)$$

и, следовательно,

$$(1 - \tau a s L) \max_{|u|<\infty} \left| \frac{dY_{i_0}}{du} \right| \leq 1.$$

Будем предполагать, что

$$1 - \tau a s L \geq \frac{1}{2}, \quad \text{т.е.} \quad \tau \leq \frac{1}{2asL}. \quad (15.54)$$

Тогда

$$\left| \frac{dY_j}{du} \right| \leq 2, \quad j = 1, \dots, s. \quad (15.55)$$

Для простоты будем предполагать, что коэффициенты  $b_j$  неотрицательны. Поскольку их сумма равна единице, то с учетом (15.52), (15.55)

$$\left| \sum_{j=1}^s b_j \frac{df(Y_j(u))}{du} \right| \leq 2L.$$

Принимая во внимание эту оценку, из (15.50) находим, что

$$|z_{n+1}| \leq (1 + 2\tau L)|z_n| + \tau|\psi_n|.$$

Разрешим эти неравенства. Поскольку  $z_0 = u_0 - u(0) = 0$ , то

$$\begin{aligned} |z_1| &\leq \tau|\psi_0|, \\ |z_2| &\leq (1 + 2\tau L)|z_1| + \tau|\psi_1|, \\ |z_3| &\leq (1 + 2\tau L)|z_2| + \tau|\psi_2|, \\ &\dots\dots\dots \\ |z_n| &\leq (1 + 2\tau L)|z_{n-1}| + \tau|\psi_{n-1}|. \end{aligned}$$

Подставим теперь оценку  $|z_1|$  в правую часть оценки  $|z_2|$ , а полученную оценку  $|z_2|$  в свою очередь в правую часть оценки  $|z_3|$  и т.д. Получим

$$\begin{aligned} |z_n| &\leq \sum_{j=0}^{n-1} \tau(1 + 2\tau L)^{n-1-j} |\psi_j| \leq (1 + 2\tau L)^n \sum_{j=0}^{n-1} \tau |\psi_j| \leq e^{2\tau LT/\tau} T \max_j |\psi_j| = \\ &= e^{2LT} T \max_j |\psi_j|. \end{aligned} \tag{15.56}$$

Из (15.56) следует

**Теорема 15.5.** *Если метод Рунге-Кутты (15.8), (15.9) аппроксимирует исходное уравнение (15.2) с порядком  $p$ , то при  $\tau \rightarrow 0$  он сходится с тем же порядком.*

# 16

## Линейные многошаговые методы

При изучении методов Рунге-Кутты, используемых для решения задачи Коши

$$\frac{du}{dt} = f(u), \quad t > 0, \quad u(0) = u_0, \quad (16.1)$$

мы не обращали особого внимания на задание начальных условий, ибо это совершенно тривиальная процедура: для того, чтобы начал работать любой из рассмотренных нами методов Рунге-Кутты, нужно задать  $u_0 = u(0)$ , т.е. так же как и для дифференциального уравнения. Обусловлено это тем, что в каждом уравнении связаны между собой значения приближенного решения в двух соседних узлах сетки (не считая промежуточных значений). Другой класс методов составляют так называемые многошаговые методы, в которых уравнения связывают значения приближенного решения в нескольких соседних узлах.

### 16.1 Методы Адамса

Наиболее известными из многошаговых методов и наиболее старыми являются методы Адамса. Опишем эти методы на примере уравнения (16.1). Вновь будем предполагать, что на отрезке интегрирования введена равномерная сетка с шагом  $\tau$ , а уравнение (16.1) проинтегрировано по отрезку между узлами  $t_n$  и  $t_{n+1}$

$$u(t_{n+1}) - u(t_n) = \int_{t_n}^{t_{n+1}} f(u(t)) dt. \quad (16.2)$$

Заменим подынтегральную функцию в (16.2) интерполяционным многочленом Лагранжа по некоторым узлам сетки  $\omega$  (а не по промежуточным

узлам, как это было в методах Рунге-Кутты (!). В зависимости от того, участвует ли узел  $t_{n+1}$  в интерполяции  $f(u(t))$  или нет, различают неявные и явные методы Адамса.

а) **Явные методы Адамса.** Предположим, что  $u(t)$  известна в  $k$  узлах сетки  $\omega$

$$t_n, t_{n-1}, \dots, t_{n+1-k}. \quad (16.3)$$

Построим по этим узлам для подынтегральной функции  $f(u(t))$  из (16.2) интерполяционный многочлен Лагранжа степени  $k - 1$

$$f(u(t)) \approx L_{k-1}(t) := \sum_{j=1}^k p_j(t) f(u(t_{n+1-j})), \quad (16.4)$$

где, как обычно,

$$p_j(t) = \prod_{\substack{i=1 \\ i \neq j}}^k \frac{t - t_{n+1-i}}{t_{n+1-j} - t_{n+1-i}} \quad (16.5)$$

суть весовые функции интерполяционного полинома (многочлены степени  $(k - 1)$ ), обращающиеся в нуль при  $t = t_{n+1-i}$ ,  $i = 1, 2, \dots, j - 1, j + 1, j + 2, \dots, k$  и в единицу при  $t = t_{n+1-j}$ . Подставляя (16.4), (16.5) в (16.2), производя интегрирование и заменяя приближенное равенство на точное, получим следующее уравнение для определения приближенного решения

$$u_{n+1} - u_n = \tau \sum_{j=1}^k b_j f(u_{n+1-j}), \quad (16.6)$$

где после замены переменной интегрирования  $(t - t_n/\tau) = \theta$

$$b_j = \frac{1}{\tau} \int_{t_n}^{t_{n+1}} p_j(t) dt = \int_0^1 \hat{p}_j(\theta) d\theta = \int_0^1 \prod_{\substack{i=1 \\ i \neq j}}^k \frac{\theta - 1 + i}{i - j} d\theta. \quad (16.7)$$

**Определение 16.1.** Численный метод (16.6), (16.7) называется *явным  $k$ -шаговым методом Адамса* (иногда его называют методом Адамса-Бэшфорта).

**Примеры.** 1°.  $k = 1$ .

$$p_1(t) = \hat{p}_1(\theta) = 1, \quad b_1 = 1.$$

2°.  $k = 2$ .

$$\begin{aligned}\hat{p}_1(\theta) &= \theta + 1, & b_1 &= 3/2, \\ \hat{p}_2(\theta) &= -\theta, & b_2 &= -1/2.\end{aligned}$$

3°.  $k = 3$ .

$$\begin{aligned}\hat{p}_1(\theta) &= \frac{1}{2}(\theta + 1)(\theta + 2), & b_1 &= \frac{23}{12}, \\ \hat{p}_2(\theta) &= -\theta(\theta + 2), & b_2 &= -\frac{4}{3}, \\ \hat{p}_3(\theta) &= \frac{\theta(\theta + 1)}{2}, & b_3 &= \frac{5}{12}.\end{aligned}$$

Выпишем уравнения (16.6) для этих частных случаев

$$\begin{aligned}u_{n+1} &= u_n + \tau f(u_n), \\ u_{n+1} &= u_n + \tau \left[ \frac{3}{2}f(u_n) - \frac{1}{2}f(u_{n-1}) \right], \\ u_{n+1} &= u_n + \tau \left[ \frac{23}{12}f(u_n) - \frac{16}{12}f(u_{n-1}) + \frac{5}{12}f(u_{n-2}) \right].\end{aligned}\tag{16.8}$$

**Упражнение 16.1.** Построить явный 4-х-шаговый метод Адамса (16.6).

**Ответ.**

$$u_{n+1} = u_n + \tau \left[ \frac{55}{24}f(u_n) - \frac{59}{24}f(u_{n-1}) + \frac{37}{24}f(u_{n-2}) - \frac{9}{24}f(u_{n-3}) \right].$$

**Замечание 16.1.** Очевидно, что первое из уравнений (16.8) определяет исследованный нами ранее метод Эйлера. Тем самым, метод Эйлера может быть отнесен как к методам Рунге-Кутты, так и к методам Адамса.

Формулы (16.6) получены при интегрировании в пределах от  $t_n$  до  $t_{n+1}$ , в то время как узлы интерполяции располагались на отрезке  $[t_{n+1-k}, t_n]$ , т.е. вне интервала интегрирования (Для подынтегральной функции использовалась экстраполяция). В связи с этим явные методы Адамса иногда называют экстраполяционными методами.

б) **Неявные методы Адамса.** Можно построить и неявные методы Адамса. Для этого к узлам интерполяции (16.3) нужно добавить еще узел  $t_{n+1}$ . В этом случае интерполяционный многочлен (степени  $k$ ) примет вид

$$L_k(t) := \sum_{j=0}^k p_j(t) f(u(t_{n+1-j})),\tag{16.9}$$

а соответствующим ему уравнением будет уравнение

$$u_{n+1} - u_n = \tau \sum_{j=0}^k b_j f(u_{n+1-j}), \quad (16.10)$$

где (ср. с (16.7))

$$b_j = \frac{1}{\tau} \int_{t_n}^{t_{n+1}} p_j(t) dt = \int_0^1 \prod_{\substack{i=0 \\ i \neq j}}^k \frac{\theta - 1 + i}{i - j} d\theta. \quad (16.11)$$

**Определение 16.2.** Численный метод (16.10), (16.11) называется *неявным  $k$ -шаговым методом Адамса* (Иногда его называют методом Адамса-Мултона).

**Примеры.** 4°.  $k = 0$ .

$$\hat{p}_0(\theta) = 1, \quad b_0 = 1.$$

5°.  $k = 1$ .

$$\begin{aligned} \hat{p}_0(\theta) &= \theta, & b_0 &= 1/2, \\ \hat{p}_1(\theta) &= -\theta + 1, & b_1 &= 1/2. \end{aligned}$$

6°.  $k = 2$ .

$$\begin{aligned} \hat{p}_0(\theta) &= \frac{1}{2}\theta(\theta + 1), & b_0 &= \frac{5}{12}, \\ \hat{p}_1(\theta) &= -(\theta^2 - 1), & b_1 &= \frac{2}{3}, \\ \hat{p}_2(\theta) &= \frac{1}{2}\theta(\theta - 1), & b_2 &= -\frac{1}{12}. \end{aligned}$$

Напишем уравнения (16.10) для этих частных случаев

$$\begin{aligned} u_{n+1} &= u_n + \tau f(u_{n+1}), \\ u_{n+1} &= u_n + \frac{\tau}{2} [f(u_{n+1}) + f(u_n)], \\ u_{n+1} &= u_n + \tau \left[ \frac{5}{12} f(u_{n+1}) + \frac{8}{12} f(u_n) - \frac{1}{12} f(u_{n-1}) \right]. \end{aligned} \quad (16.12)$$

**Упражнение 16.2.** Построить неявный 3-х-шаговый метод Адамса (16.10), (16.11).



**Ответ.**

$$u_{n+1} = u_n + \tau \left[ \frac{9}{24}f(u_{n+1}) + \frac{19}{24}f(u_n) - \frac{5}{24}f(u_{n-1}) + \frac{1}{24}f(u_{n-2}) \right].$$

**Замечание 16.2.** Очевидно, что первое из уравнений (16.12), отвечающее  $k = 0$ , является неявным методом Эйлера, а второе уравнение, отвечающее  $k = 1$ , — методом трапеций. Тем самым, эти одношаговые неявные методы Адамса являются и методами Рунге-Кутты.

## 16.2 Формулы дифференцирования назад

Во всех предыдущих случаях, как при построении методов Рунге-Кутты, так и при построении методов Адамса, мы получали численные методы путем интегрирования уравнения (16.1) и замены подынтегральной функции  $f(u)$  в (16.2) интерполяционным многочленом или замены интеграла квадратурной формулой. А можно поступать и иначе: интерполяционным многочленом заменить  $u(t)$ . Тогда для построения численного метода нужно будет выражение интерполяционного многочлена подставить прямо в (16.1). Чтобы получился численный метод, точка  $t_{n+1}$  должна быть в числе узлов интерполяции. Пусть

$$u(t) \approx L_k(t) = \sum_{j=0}^k p_j(t)u(t_{n+1-j}), \quad (16.13)$$

где

$$p_j(t) = \prod_{\substack{i=0 \\ i \neq j}}^k \frac{t - t_{n+1-i}}{t_{n+1-j} - t_{n+1-i}}.$$

Подставляя (16.13) в (16.1), получим приближенное равенство

$$\sum_{j=0}^k p'_j(t)u(t_{n+1-j}) \approx f \left( \sum_{j=0}^k p_j(t)u(t_{n+1-j}) \right).$$

Превратим его в точное равенство в каком-либо узле. В результате получим уравнение для определения приближенного решения. Наиболее интересным является случай, когда указанным узлом является  $t_{n+1}$ . Будем иметь

$$\sum_{j=0}^k p'_j(t_{n+1})u_{n+1-j} = f(u_{n+1}).$$

Как и раньше, сделаем локальную замену переменной  $(t - t_n)/\tau = \theta$ . Тогда

$$p'_j(t) = \frac{dp_j(t)}{dt} = \frac{1}{\tau} \frac{d\hat{p}_j(\theta)}{d\theta} = \frac{1}{\tau} \hat{p}'_j(\theta),$$

где

$$p_j(t) = \hat{p}_j(\theta) = \prod_{\substack{i=0 \\ i \neq j}}^k \frac{\theta - 1 + i}{i - j},$$

и полученный метод принимает вид

$$\sum_{j=0}^k \hat{p}'_j(1) u_{n+1-j} = \tau f(u_{n+1}). \quad (16.14)$$

**Определение 16.3.** Численные методы (16.14) называются *формулами дифференцирования назад*.

Методы (16.14) были предложены в работе Кёртиса и Хиршфельдера (1952) и, в связи с этим, иногда называются их именами. С другой стороны, методы (16.14) были реализованы в популярной в свое время программе, автором которой был Гир (1971). Это обстоятельство послужило тому, что методы (16.14) иногда фигурируют под названием методы Гира.

**Примеры.** 7°.  $k = 1$ .

$$\begin{aligned} \hat{p}_0(\theta) &= \theta, & \hat{p}'_0(1) &= 1, \\ \hat{p}_1(\theta) &= -\theta + 1, & \hat{p}'_1(1) &= -1. \end{aligned}$$

8°.  $k = 2$ .

$$\begin{aligned} \hat{p}_0(\theta) &= \frac{1}{2}\theta(\theta + 1), & \hat{p}'_0(1) &= \frac{3}{2}, \\ \hat{p}_1(\theta) &= 1 - \theta^2, & \hat{p}'_1(1) &= -2, \\ \hat{p}_2(\theta) &= \frac{1}{2}\theta(\theta - 1), & \hat{p}'_2(1) &= \frac{1}{2}. \end{aligned}$$

Выпишем уравнения (16.14) для этих случаев

$$u_{n+1} - u_n = \tau f(u_{n+1}), \quad (16.15)$$

$$\left( \frac{3}{2}u_{n+1} - 2u_n + \frac{1}{2}u_{n-1} \right) = \tau f(u_{n+1}). \quad (16.16)$$

**Упражнение 16.3.** Построить формулу (16.14), отвечающую  $k = 3$ .

**Ответ.**

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

### 16.3 Общие линейные многошаговые методы

Методы Адамса, явные и неявные, и формулы дифференцирования назад являются частными случаями формулы

$$\sum_{j=0}^k \alpha_j u_{n-j} = \tau \sum_{j=0}^k \beta_j f(u_{n-j}), \quad (16.17)$$

где  $\alpha_j$  и  $\beta_j$  — действительные числа. (Обратим внимание на то, что в этой формуле вместо нового неизвестного  $u_{n+1}$  фигурирует  $u_n$ ). Будет предполагать, что

$$\alpha_0 \neq 0, \quad |\alpha_k| + |\beta_k| \neq 0. \quad (16.18)$$

Первое из условий (16.18) обеспечивает разрешимость неявного ( $\beta_0 \neq 0$ ) уравнения (16.17) по крайней мере, для достаточно малого шага  $\tau$ . Второе из условий (16.18) всегда можно считать выполненным, уменьшив при необходимости  $k$ .

**Определение 16.4.** Формула (16.17) называется *линейным многошаговым ( $k$ -шаговым) методом*.

Метод является явным, если  $\beta_0 = 0$ , и неявным в противном случае.

Чтобы линейный многошаговый метод (16.17) можно было использовать для численного решения задачи (16.1), необходимо, чтобы уравнение (16.17) аппроксимировало уравнение (16.1).

**Определение 16.5.** Величина

$$\psi_n = \sum_{j=0}^k \beta_j f(u(t_{n-j})) - \frac{1}{\tau} \sum_{j=0}^k \alpha_j u(t_{n-j}) \quad (16.19)$$

называется погрешностью аппроксимации метода (16.17).

Выясним вопрос о порядке погрешности аппроксимации метода (16.17) при  $\tau \rightarrow 0$ .

**Теорема 16.1.** *Многошаговый метод (16.17) имеет погрешность аппроксимации порядка  $p \leq 2k$  тогда и только тогда, когда выполняются следующие условия*

$$\sum_{j=0}^k \alpha_j = 0, \quad \sum_{j=0}^k (\alpha_j j^q + q\beta_j j^{q-1}) = 0, \quad q = 1, \dots, p. \quad (16.20)$$

**Доказательство.** Разложим  $u(t)$  по формуле Тейлора в точке  $t_n$ :

$$u(t) = \sum_{q=0}^p \frac{(t-t_n)^q}{q!} u^{(q)}(t_n) + O((t-t_n)^{p+1}). \quad (16.21)$$

Так как  $f(u) = u'(t)$ , то, дифференцируя (16.21), получим

$$f(u(t)) = \sum_{q=0}^p q \frac{(t-t_n)^{q-1}}{q!} u^{(q)}(t_n) + O((t-t_n)^p). \quad (16.22)$$

Подставляя теперь разложения (16.21), (16.22) при  $t = t_{n-j}$  в (16.19), будем иметь

$$\begin{aligned} \psi_n &= \sum_{j=0}^k \beta_j \sum_{q=0}^p q \frac{(-j\tau)^{q-1}}{q!} u^{(q)}(t_n) - \\ &- \frac{1}{\tau} \sum_{j=0}^k \alpha_j \sum_{q=0}^p \frac{(-j\tau)^q}{q!} u^{(q)}(t_n) + O(\tau^p) = \\ &= \sum_{q=0}^p \frac{(-\tau)^{q-1}}{q!} u^{(q)}(t_n) \sum_{j=0}^k [\beta_j q j^{q-1} + \alpha_j j^q] + O(\tau^p). \end{aligned}$$

Приравнивая нулю коэффициенты при  $\tau^{q-1}$  для  $q = 0, 1, \dots, p$ , получим (16.20). Теорема доказана.

**Замечание 16.3.** Решение уравнения (16.17) не изменится, если его умножить на какое-либо число, отличное от нуля. Это означает, что его коэффициенты определяются с точностью до множителя (до мультипликативной постоянной). Чтобы устранить этот произвол, пронормируем их, полагая, например,

$$\sum_{j=0}^k \beta_j = 1. \quad (16.23)$$

**Замечание 16.4.** Иногда какие-либо из коэффициентов  $\alpha_j$  и  $\beta_j$  задаются заранее. Если хотя бы один из этих коэффициентов отличен от нуля, то условие (16.23) не используется.

**Замечание 16.5.** Из (16.20), (16.23) имеем  $(p + 2)$  уравнения для  $2(k + 1)$  коэффициентов метода (16.17). Тем самым, максимальный порядок аппроксимации линейного  $k$ -шагового метода есть  $p = 2k$ .

## 16.4 Погрешность аппроксимации методов Адамса

Исследуем вопрос о порядке погрешности аппроксимации методов Адамса. Для этого перепишем сначала явный метод Адамса (16.6), (16.7) в виде (16.17), т.е. заменим  $n + 1$  на  $n$ :

$$u_n - u_{n-1} = \tau \sum_{j=1}^k b_j f(u_{n-j}).$$

Сравнивая это соотношение с (16.17), находим, что

$$\alpha_0 = 1, \alpha_1 = -1, \alpha_2 = \dots = \alpha_k = 0, \beta_0 = 0, b_j = \beta_j, j = 1, \dots, k.$$

Определим, для каких дифференциальных уравнений явные методы Адамса теоретически дают в узлах сетки точное решение. Это произойдет в том случае, когда интерполяционный многочлен  $L_{k-1}(t)$ , определяющий явный метод Адамса, совпадает с  $f(u)$  или с  $f(t, u)$ . Пусть  $f(t, u(t)) = f(t)$ , т.е.  $f$  не зависит от  $u$  и является многочленом степени не выше  $k - 1$ . Тогда  $f(t)$  совпадает со своим интерполяционным многочленом  $L_{k-1}(t)$ , и явный метод Адамса точен для уравнений

$$u' = qt^{q-1}, \quad q = 0, \dots, k.$$

Это означает, что погрешность аппроксимации (16.19) на решениях этих уравнений равна нулю. Подставляя решения этих уравнений  $u = t^q$  в (16.19) при  $n = 0$ , получим

$$\psi_0 = \sum_{j=0}^k \left[ \beta_j q (-\tau j)^{q-1} - \frac{1}{\tau} \alpha_j (-\tau j)^q \right] = 0, \quad q = 0, \dots, k,$$

что совпадает с первыми  $(k + 1)$  уравнениями (16.20). Тем самым, мы доказали, что явный  $k$ -шаговый метод Адамса имеет порядок погрешности аппроксимации не ниже  $k$ . Можно показать, что его порядок аппроксимации в точности равен  $k$ .

**Упражнение 16.4.** Доказать, что порядок аппроксимации неявного  $k$ -шагового метода Адамса не ниже  $k + 1$ .

**Упражнение 16.5.** Доказать, что порядок аппроксимации  $k$ -шаговой формулы дифференцирования назад не ниже  $k$ .

## 16.5 Поучительный пример

Построим двухшаговый явный метод максимального порядка аппроксимации. Согласно ранее сказанному, порядок аппроксимации этого метода должен быть равен трем. Из (16.20), (16.23) имеем

$$\begin{aligned}\alpha_0 + \alpha_1 + \alpha_2 &= 0, \\ \alpha_1 + 2\alpha_2 &= -(\beta_0 + \beta_1 + \beta_2), \\ \alpha_1 + 4\alpha_2 &= -2(\beta_1 + 2\beta_2), \\ \alpha_1 + 8\alpha_2 &= -3(\beta_1 + 4\beta_2), \\ \beta_0 + \beta_1 + \beta_2 &= 1, \\ \beta_0 &= 0.\end{aligned}$$

Разрешая эту линейную систему, находим, что

$$\alpha_0 = \frac{1}{6}, \quad \alpha_1 = \frac{2}{3}, \quad \alpha_2 = -\frac{5}{6}, \quad \beta_1 = \frac{2}{3}, \quad \beta_2 = \frac{1}{3}.$$

Тем самым, метод (16.17) приобретает вид

$$\left(\frac{1}{6}u_n + \frac{4}{6}u_{n-1} - \frac{5}{6}u_{n-2}\right) = \tau \left[\frac{2}{3}f_{n-1} + \frac{1}{3}f_{n-2}\right]. \quad (16.24)$$

Применим этот метод к решению уравнения (16.1) с  $f(u) = \lambda u$ , где  $\lambda = \text{const}$ . Будем при этом предполагать, что начальное значение  $u_0 = 1$ . В этом случае задача (16.1) примет вид

$$u'(t) = \lambda u, \quad u(0) = 1, \quad (16.25)$$

а ее решением будет функция

$$u(t) = e^{\lambda t}. \quad (16.26)$$

Отвечающий (16.25) метод (16.17) можно записать так

$$\sum_{j=0}^k (\alpha_j - \tau \lambda \beta_j) u_{n-j} = 0, \quad (16.27)$$

а применительно к методу (16.24)

$$\frac{1}{6}u_n + \left(\frac{4}{6} - \frac{2}{3}\tau\lambda\right)u_{n-1} + \left(-\frac{5}{6} - \frac{1}{3}\tau\lambda\right)u_{n-2} = 0. \quad (16.28)$$

Это есть линейное однородное разностное уравнение второго порядка с постоянными коэффициентами (см. §6). Найдем его решение. Для этого нужно написать характеристическое уравнение, отвечающее разностному уравнению (16.28), и найти его корни. Искомое характеристическое уравнение имеет вид

$$q^2 + 4(1 - \tau\lambda)q - (5 + 2\tau\lambda) = 0, \quad (16.29)$$

а его корни суть

$$\begin{aligned} q_1 &= -2 + 2\tau\lambda + \sqrt{9 - 6\tau\lambda + 4\tau^2\lambda^2} = 1 + \tau\lambda + O(\tau^2\lambda^2), \\ q_2 &= -2 + 2\tau\lambda - \sqrt{9 - 6\tau\lambda + 4\tau^2\lambda^2} = -5 + O(\tau\lambda). \end{aligned} \quad (16.30)$$

**Упражнение 16.6.** Доказать, что  $q_1 - e^{\tau\lambda} = O(\tau^4\lambda^4)$ .

Поскольку корни (16.30) характеристического уравнения различны, то общее решение разностного уравнения (16.28) имеет вид

$$u_n = c_1q_1^n + c_2q_2^n, \quad (16.31)$$

где  $c_1$  и  $c_2$  — произвольные постоянные.

Рассматриваемый нами метод (16.24) является двухшаговым, и одного начального условия

$$u_0 = 1 \quad (16.32)$$

для его реализации недостаточно. Поскольку точное решение нам известно, то не будем ломать голову над тем, как задать недостающее начальное условие при  $n = 1$ , а просто положим

$$u_1 = u(t_1) = e^{\tau\lambda}. \quad (16.33)$$

Потребуем, чтобы решение (16.31) удовлетворяло условиям (16.32), (16.33). После простых вычислений находим, что искомое решение имеет вид

$$u_n = \frac{e^{\tau\lambda} - q_2}{q_1 - q_2}q_1^n + \frac{q_1 - e^{\tau\lambda}}{q_1 - q_2}q_2^n. \quad (16.34)$$

Изучим поведение этого решения при  $n \rightarrow \infty$ . Пусть  $t = n\tau$  фиксировано, а  $\tau \rightarrow 0$ . Тогда  $n = t/\tau \rightarrow \infty$ . С учетом (16.30) и упражнения 16.6 находим, что

$$\begin{aligned} c_1 &= \frac{e^{\tau\lambda} - q_2}{q_1 - q_2} = \frac{1 + O(\tau) + 5}{6 + O(\tau)} = 1 + O(\tau), \\ c_2 &= \frac{q_1 - e^{\tau\lambda}}{q_1 - q_2} = \frac{O(\tau^4)}{6 + O(\tau)} = O(\tau^4). \end{aligned} \quad (16.35)$$

Далее,

$$q_1^n = [e^{\tau\lambda} + O(\tau^4)]^n = e^{\lambda\tau n} (1 + O(\tau^4))^n = e^{\lambda t} (1 + O(\tau^3)). \quad (16.36)$$

Подставляя теперь (16.35), (16.36), (16.30) в (16.34), будем иметь

$$u_n = [1 + O(\tau)] e^{t\lambda} + O(\tau^4) [-5 + O(\tau)]^n.$$

Проанализируем полученный результат. Первое слагаемое аппроксимирует решение (16.26) задачи (16.25), а второе слагаемое является паразитным. Уже при не слишком больших  $n$  это слагаемое превосходит первое, ибо

$$O(\tau^4) [-5 + O(\tau)]^n = O\left(\left(\frac{t}{n}\right)^4\right) \left(-5 + O\left(\frac{t}{n}\right)\right)^n.$$

Метод (16.28) сходящимся не является.



## Устойчивость

В теореме 15.5 было установлено, что все методы Рунге-Кутты являются сходящимися, а, следовательно, и устойчивыми, по крайней мере при выполнении условия (15.54). Пример из раздела 16.5 показал, что с многошаговыми методами дело обстоит не так благополучно. Поэтому сначала разберемся с хоть какой-нибудь устойчивостью многошаговых методов, а затем введем более жесткие условия устойчивости и изучим оба класса методов с новых позиций.

### 17.1 Нуль-устойчивость

Обратимся к разностному уравнению (16.27) и введем следующие обозначения

$$\rho(\zeta) := \sum_{j=0}^k \alpha_j \zeta^{k-j}, \quad \sigma(\zeta) := \sum_{j=0}^k \beta_j \zeta^{k-j}. \quad (17.1)$$

**Определение 17.1.** Многочлены  $\rho(\zeta)$  и  $\sigma(\zeta)$  из (17.1) называются соответственно первым и вторым производящими многочленами линейного многошагового метода (16.17).

Как уже было отмечено, линейный многошаговый метод (16.17) для уравнения (16.25) принимает вид линейного разностного уравнения с постоянными коэффициентами (16.27). Его характеристическое уравнение есть

$$\rho(q) - \tau \lambda \sigma(q) = 0. \quad (17.2)$$

Применительно к двушаговому методу (16.24)

$$\rho(q) = (q^2 + 4q - 5)/6,$$

а корни уравнения

$$\rho(q) = 0 \quad (17.3)$$

суть

$$q_1 = 1, \quad q_2 = -5,$$

т.е. совпадают с главными членами корней (16.30) характеристического уравнения (16.29). Именно наличие корня  $q_2$  и привело к неустойчивости метода (16.24). Тем самым, корни уравнения (17.3) позволяют судить об устойчивости или неустойчивости метода (16.17). А они связаны с корнями характеристического уравнения (17.2). В силу (16.18),  $\alpha_0 \neq 0$  и, следовательно, степени уравнений (17.2) и (17.3) совпадают. Поэтому характеристическое уравнение (17.2) можно рассматривать как *регулярное возмущение* (при малых  $\tau\lambda$ ) уравнения (17.3) (объяснение терминов: коэффициенты многочлена  $\rho(\zeta)$  суть пределы при  $\tau\lambda \rightarrow 0$  соответствующих коэффициентов характеристического многочлена, и поэтому можно говорить о возмущении; регулярность есть следствие того, что степени возмущенного и невозмущенного многочленов совпадают). Но тогда (в силу регулярности возмущения) корни уравнения (17.3) являются пределами корней уравнения (17.2) при  $\tau\lambda \rightarrow 0$ . Поэтому вопрос о том, будет ли решение уравнения (16.27) неограниченно возрастать при  $n \rightarrow \infty$  (и фиксированном  $t = n\tau$ ), можно решить при анализе корней уравнения (17.3). Отметим, что уравнение (17.3) является характеристическим уравнением для разностного уравнения

$$\sum_{j=0}^k \alpha_j u_{n-j} = 0, \quad (17.4)$$

которое, в свою очередь, получается из (16.27), если в нем положить  $\lambda = 0$ . Это означает, что (17.4) есть линейный многошаговый метод для уравнения

$$u' = 0. \quad (17.5)$$

Тем самым, отбраковка "плохих" (неустойчивых) методов может быть осуществлена при анализе их свойств применительно к уравнению (17.5).

Итак, наличие у уравнения (17.3) корней, модули которых превосходят единицу, приводит к неустойчивости. Однако опасность представляют не только такие корни, но и корни, равные по модулю единице, если они кратные. В самом деле, пусть  $q_1$  — корень характеристического уравнения (17.3) кратности  $s > 1$  такой, что  $|q_1| = 1$ . Тогда сеточная функция

$$P_{s-1}(n)q_1^n$$

будет неограниченным решением уравнения (17.4), в то время как решением уравнения (17.5), которое и аппроксимирует изучаемое уравнение (17.4), есть постоянная.

**Определение 17.2.** Говорят, что линейный многошаговый метод (16.17) удовлетворяет корневому условию, если

- 1) все корни первого производящего многочлена (17.1) расположены в единичном круге  $|\zeta| \leq 1$ ;
- 2) нули  $\rho(\zeta)$ , расположенные на единичной окружности  $|\zeta| = 1$  простые.

**Теорема 17.1.** *Линейный многошаговый метод (16.17), удовлетворяющий корневому условию, устойчив (нуль-устойчив).*

**Замечание 17.1.** Если линейный многошаговый метод (16.17) аппроксимирует какое-либо дифференциальное уравнение, то среди нулей  $\rho(\zeta)$  обязательно есть  $\zeta = 1$ , о чем свидетельствует первое из условий (16.20), являющее собой условие  $\rho(1) = 0$ .

**Примеры.** 1° Явный и неявный методы Адамса. В обоих случаях  $\alpha_0 = 1$ ,  $\alpha_1 = -1$ , а остальные  $\alpha_j = 0$ . Поэтому

$$\rho(q) = q^k - q^{k-1}$$

и, следовательно,

$$q_1 = 1, \quad q_2 = \dots = q_k = 0.$$

Методы Адамса нуль-устойчивы.

2° Двухшаговая формула дифференцирования назад (16.16).

$$\rho(q) = \frac{3}{2}q^2 - 2q + \frac{1}{2},$$

$$q_1 = 1, \quad q_2 = 1/3.$$

Метод нуль-устойчив.

3° Трехшаговая формула дифференцирования назад (см. упражнение 16.3)

$$\rho(q) = \frac{11}{6}q^3 - 3q^2 + \frac{3}{2}q - \frac{1}{3}.$$

Хотя это и многочлен третьей степени, нули его легко находятся, ибо один из его нулей есть  $q_1 = 1$ . Деля  $\rho(q)$  на  $(q - 1)$ , приходим к уравнению

$$\frac{11}{6}q^2 - \frac{7}{6}q + \frac{1}{3} = 0$$

с корнями

$$q_{2,3} = \frac{7 \pm i\sqrt{39}}{22}.$$

Отсюда

$$|q_{2,3}|^2 = \frac{2}{11} < 1.$$

Метод нуль-устойчив.

**Теорема 17.2** (Первый барьер Далквиста). *Порядок  $p$  устойчивого линейного  $k$ -шагового метода подчиняется следующим ограничениям:*

$$\begin{aligned} p &\leq k && \text{для явных методов;} \\ p &\leq k + 1 && \text{для неявных методов при нечетном } k; \\ p &\leq k + 2 && \text{для неявных методов при четном } k. \end{aligned}$$

В качестве иллюстрации второго утверждения теоремы может служить построенный нами в предыдущем параграфе явный двухшаговый метод максимального порядка аппроксимации  $p = 3$ , который оказался неустойчивым.

**Упражнение 17.1.** Построить общий явный устойчивый двухшаговый метод максимального порядка аппроксимации.

**Ответ:**  $\alpha_0$  — параметр метода,

$$\begin{aligned} \alpha_1 &= 1 - 2\alpha_0, & \alpha_2 &= \alpha_0 - 1, \\ \beta_0 &= 0, & \beta_1 &= \frac{1}{2} + \alpha_0, & \beta_2 &= \frac{1}{2} - \alpha_0. \end{aligned}$$

Условие устойчивости:  $1/2 \leq \alpha_0 < \infty$ . При  $\alpha_0 = 1$  имеем явный метод Адамса, при  $\alpha_0 = 1/2$  — метод прямоугольников с шагом  $\tau' = 2\tau$ . При  $\alpha_0 = 1/6$  метод имеет погрешность аппроксимации  $O(\tau^3)$ , но неустойчив.

**Упражнение 17.2.** Построить устойчивый двухшаговый метод максимального порядка аппроксимации.

**Ответ:**

$$\begin{aligned} \alpha_0 &= 1/2, & \alpha_1 &= 0, & \alpha_2 &= -1/2, \\ \beta_0 &= 1/6, & \beta_1 &= 2/3, & \beta_2 &= 1/6. \end{aligned}$$

Этот метод иногда называется методом Симпсона (по аналогии с одноименной квадратурной формулой). Метод имеет четвертый порядок аппроксимации.

## 17.2 Жесткие задачи

При определении нуль-устойчивости многошагового метода мы могли ограничиться изучением простейшего дифференциального уравнения (17.5), ибо производящий многочлен  $\rho(\zeta)$  из (17.1) многошагового метода (16.17), от расположения нулей которого зависит, будет ли метод устойчивым или нет, является характеристическим многочленом именно в применении к уравнению (17.5). Условие нуль-устойчивости предъявляет минимальные требования к численному методу, производя лишь грубую отбраковку абсолютно непригодных для вычислений методов. По существу, нуль-устойчивость метода обеспечивает лишь ограниченность приближенного решения для конечного временного интервала  $[0, T]$  при  $n \rightarrow \infty$ .

Однако имеются задачи, отыскание решений которых при помощи только нуль-устойчивых методов оказывается весьма затруднительным, если не невозможным. Проще всего объяснить возникающие трудности не на примере одного уравнения, а на примере систем уравнений.

Рассмотрим однородную систему линейных дифференциальных уравнений с постоянными коэффициентами

$$\mathbf{u}' = A\mathbf{u}, \quad \mathbf{u}(0) = \mathbf{u}_0, \quad (17.6)$$

где  $\mathbf{u} = [u_1 \ u_2]^T$ , а

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

Найдем и проанализируем решение задачи (17.6). Как обычно, будем его искать в виде

$$\mathbf{u}(t) = \boldsymbol{\xi} e^{\lambda t}, \quad (17.7)$$

где  $\boldsymbol{\xi}$  — двумерный числовой вектор, а  $\lambda$  — постоянная. Подставляя (17.7) в (17.6), находим, что

$$\lambda \boldsymbol{\xi} e^{\lambda t} = e^{\lambda t} A \boldsymbol{\xi},$$

а, сокращая на  $e^{\lambda t}$ , получим следующую задачу на собственные значения:

$$A \boldsymbol{\xi} = \lambda \boldsymbol{\xi}. \quad (17.8)$$

Будем предполагать, что  $A$  — матрица простой структуры, т.е. у нее имеется полный набор собственных векторов. Тогда

$$A \boldsymbol{\xi}_1 = \lambda_1 \boldsymbol{\xi}_1, \quad A \boldsymbol{\xi}_2 = \lambda_2 \boldsymbol{\xi}_2$$

и  $\xi_1$  и  $\xi_2$  линейно независимы.

В рассматриваемом случае общее решение системы (17.6) принимает вид

$$\mathbf{u}(t) = c_1 \xi_1 e^{\lambda_1 t} + c_2 \xi_2 e^{\lambda_2 t}, \quad (17.9)$$

а решение задачи Коши (17.6) получается отсюда при значениях  $c_1$  и  $c_2$ , найденных из алгебраической системы

$$\xi_1 c_1 + \xi_2 c_2 = \mathbf{u}_0. \quad (17.10)$$

Будем для простоты предполагать, что собственные числа  $\lambda_1$  и  $\lambda_2$  действительны. Более существенным для нас будет предположение об их отрицательности

$$\lambda_1 < 0, \quad \lambda_2 < 0. \quad (17.11)$$

В силу сделанных предположений модули компонент  $u_1$  и  $u_2$  решения (17.9) будут стремиться к нулю при  $t \rightarrow \infty$ .

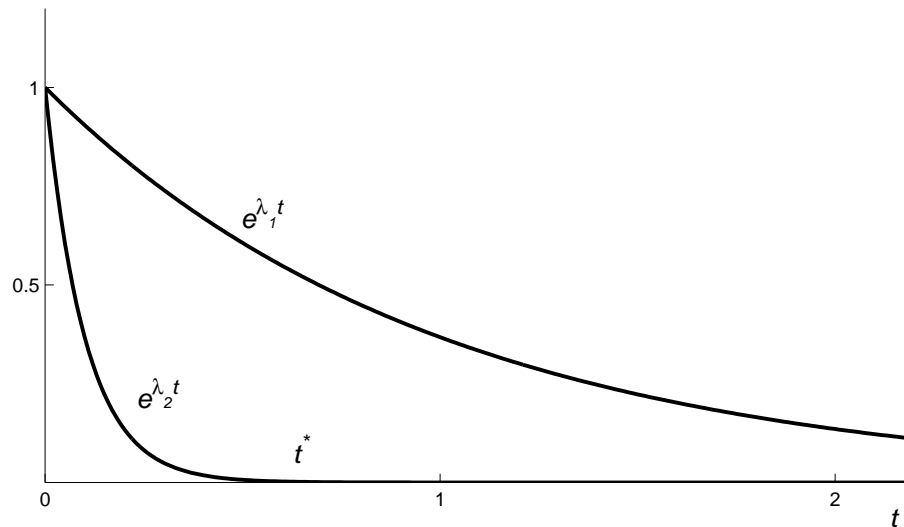


Рис. 1

Предположим теперь дополнительно, что

$$\lambda_1 = O(1), \quad |\lambda_2| \gg |\lambda_1|. \quad (17.12)$$

Так как в этом случае  $e^{\lambda_2 t}$  убывает значительно быстрее  $e^{\lambda_1 t}$ , то через некоторое время  $t^*$  составляющая  $c_2 \xi_2 e^{\lambda_2 t}$  решения (17.9) будет практически равной нулю, и решение будет почти полностью определяться составляющей  $c_1 \xi_1 e^{\lambda_1 t}$ . (см. рис. 1)

В рассматриваемой ситуации естественно было бы ожидать, что и у численного решения задачи (17.6) модули компонент хотя бы не возрастали.

Применим для решения задачи (17.6) метод Эйлера

$$\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\tau} = A\mathbf{u}^n, \quad \mathbf{u}^0 = \mathbf{u}_0. \quad (17.13)$$

Найдем решение задачи (17.13). Искать его будем в виде (см. (10.30))

$$\mathbf{u}^n = \boldsymbol{\xi}q^n, \quad q = \text{const} \neq 0. \quad (17.14)$$

Подставляя (17.14) в (17.13), получим

$$q^n \frac{q - 1}{\tau} \boldsymbol{\xi} = q^n A \boldsymbol{\xi},$$

а после сокращения на  $q^n$  обнаруживаем, что для отыскания  $\boldsymbol{\xi}$  имеем задачу (17.8) с  $\lambda = (q - 1)/\tau$ . Поэтому  $q = 1 + \tau\lambda$ , и решение задачи (17.13) есть

$$\mathbf{u}^n = c_1 \boldsymbol{\xi}_1 (1 + \tau\lambda_1)^n + c_2 \boldsymbol{\xi}_2 (1 + \tau\lambda_2)^n, \quad (17.15)$$

где  $c_1, c_2$  — решение системы (17.10).

Чтобы модули компонент решения (17.15) не возрастали при  $n \rightarrow \infty$ , необходимо и достаточно, чтобы выполнялись условия

$$|1 + \tau\lambda_1| \leq 1, \quad |1 + \tau\lambda_2| \leq 1,$$

что вместе с (17.11) и (17.12) приводит к условию

$$\tau \leq 2/|\lambda_2| \ll 1. \quad (17.16)$$

Ограничение (17.16), вообще говоря, является довольно жестким. Если при  $t \leq t^*$  это ограничение вполне разумно, и даже из соображений аппроксимации и точности нужно требовать  $\tau \ll 2/|\lambda_2|$ , то при  $t > t^*$ , когда вторая составляющая каждой компоненты решения (17.15) вроде бы не должна поставлять новой информации, и желательно было бы увеличить шаг  $\tau$  с той целью, чтобы сэкономить ресурсы и не воспроизводить первую составляющую с излишней точностью. Но тогда придется нарушить условие (17.16), что приведет к резкому возрастанию второй составляющей решения и полной потере точности.

**Определение 17.3.** Система дифференциальных уравнений (17.6) с постоянной матрицей  $A$  порядка  $m$  называется жесткой, если

$$1^\circ \operatorname{Re} \lambda_j < 0 \quad , \quad j = 0, \dots, m,$$

2° отношение

$$S = \frac{\max_j |\operatorname{Re} \lambda_j|}{\min_j |\operatorname{Re} \lambda_j|} \gg 1. \quad (17.17)$$

**Определение 17.4.** Число  $S$  из (17.17) называется коэффициентом жесткости задачи (17.6).

**Замечание 17.2.** Для линейной системы с матрицей  $A$ , зависящей от  $t$ , коэффициент жесткости также зависит от  $t$ , и, если он велик для каких-либо  $t$  из интересующего нас интервала, то система жесткая. Для нелинейных систем жесткость определяется в окрестности какого-либо решения при помощи соответствующей матрицы Якоби.

Применим теперь для решения задачи (17.6) неявный метод Эйлера

$$\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\tau} = A\mathbf{u}^{n+1}.$$

Подставляя сюда (17.14), находим, что

$$q^{n+1} \frac{1 - q^{-1}}{\tau} \boldsymbol{\xi} = q^{n+1} A \boldsymbol{\xi},$$

т.е.  $\lambda\tau = (1 - q^{-1})$ ,  $q = (1 - \tau\lambda)^{-1}$  и

$$\mathbf{u}^n = c_1 \boldsymbol{\xi}_1 (1 - \tau\lambda)^{-n} + c_2 \boldsymbol{\xi}_2 (1 - \tau\lambda_2)^{-n}.$$

Очевидно, что при выполнении условий (17.11) модули компонент  $\mathbf{u}^n$  монотонно убывают при  $n \rightarrow \infty$  при *любом*  $\tau$ , и, следовательно,  $\tau$  можно выбирать только из соображений точности.

Неявный метод Эйлера при решении жестких систем оказался существенно более устойчивым, чем просто метод Эйлера.

Как отобразить методы, пригодные для решения жестких задач? Уже сточить требование устойчивости.

### 17.3 А-устойчивость

Если при определении нуль-устойчивости основной моделью было уравнение (17.5), то теперь следует обратиться к уравнению (16.25). Многошаговый метод (16.17) в применении к линейному однородному уравнению



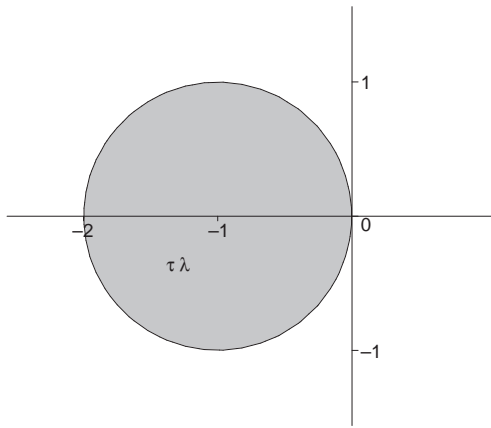


Рис. 2

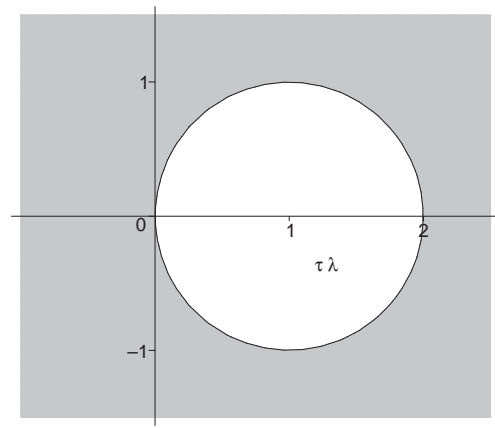


Рис. 3

(16.25) имеет вид (16.27), а характеристическое уравнение этого разностного уравнения задается соотношением (17.2).

**Определение 17.5.** Линейный многошаговый метод (16.17) в применении к уравнению (16.25) называется *абсолютно устойчивым* для данного  $\lambda$  и данного  $\tau$ , если при указанном значении  $\tau\lambda$  все корни характеристического уравнения (17.2) расположены внутри единичного круга.

**Определение 17.6.** Множество всех точек комплексной плоскости  $\tau\lambda$ , для которых линейный многошаговый метод (16.17) в применении к (16.25) абсолютно устойчив, называется *областью абсолютной устойчивости* метода.

**Пример 4°.** Метод Эйлера (14.7). Единственный корень характеристического уравнения этого метода есть  $q = 1 + \tau\lambda$ , условие абсолютной устойчивости имеет вид

$$|1 + \tau\lambda| \leq 1.$$

Тем самым, областью абсолютной устойчивости метода Эйлера является единичный круг с центром в точке  $\tau\lambda = -1$  (см. рис. 2).

**Пример 5°.** Неявный метод Эйлера (14.8). Условие абсолютной устойчивости

$$|q| = |1 - \tau\lambda|^{-1} \leq 1, \quad \text{т.е.} \quad |1 - \tau\lambda| \geq 1.$$

Областью абсолютной устойчивости является внешность единичного круга с центром в точке  $\tau\lambda = 1$  (см. рис. 3).

**Определение 17.7.** Линейный многошаговый метод (16.17) называется *A-устойчивым*, если его область абсолютной устойчивости содержит левую полуплоскость  $\operatorname{Re}(\tau\lambda) < 0$ .

Из приведенных примеров следует, что метод Эйлера не является А-устойчивым, а неявный метод Эйлера А-устойчив.

**Пример 6°.** Метод трапеций (14.12). Применительно к уравнению (16.25) этот метод имеет вид

$$\frac{u^{n+1} - u^n}{\tau} = \lambda \frac{u^{n+1} + u^n}{2},$$

а его характеристическое уравнение есть  $(q - 1)/\tau = \lambda(q + 1)/2$ . Отсюда находим единственный корень

$$q = \frac{1 + \tau\lambda/2}{1 - \tau\lambda/2}$$

и условие абсолютной устойчивости

$$|q| = \left| \frac{1 + \tau\lambda/2}{1 - \tau\lambda/2} \right| \leq 1$$

или

$$|1 + \tau\lambda/2| \leq |1 - \tau\lambda/2|.$$

Пусть  $\tau\lambda = x + iy$ . Тогда условие абсолютной устойчивости примет вид

$$\left| 1 + \frac{x}{2} + i\frac{y}{2} \right| \leq \left| 1 - \frac{x}{2} - i\frac{y}{2} \right|.$$

или

$$\left( 1 + \frac{x}{2} \right)^2 + \frac{y^2}{4} \leq \left( 1 - \frac{x}{2} \right)^2 + \frac{y^2}{4}.$$

Раскрывая скобки, находим, что условие абсолютной устойчивости есть

$$x = \operatorname{Re}(\tau\lambda) < 0.$$

Областью абсолютной устойчивости метода трапеций является левая полуплоскость  $\operatorname{Re}(\tau\lambda) < 0$  (Рис. 4). Метод А-устойчив.

**Теорема 17.3.** Среди линейных многошаговых методов (16.17) не существует явных А-устойчивых методов.

**Теорема 17.4.** Среди неявных линейных многошаговых методов (16.17) не существует А-устойчивых методов, имеющих порядок точности выше второго.

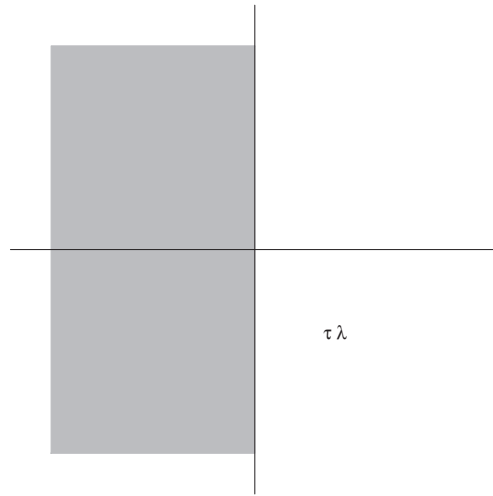


Рис. 4

**Пример 7°.** Двухшаговая формула дифференцирования назад. Этот метод задается соотношением (16.16)

$$\left(\frac{3}{2}u_{n+1} - 2u_n + \frac{1}{2}u_{n-1}\right) = \tau f(u_{n+1}). \quad (17.18)$$

Характеристическое уравнение, отвечающее этому методу в применении к уравнению (16.25) есть

$$\frac{3}{2}q^2 - 2q + \frac{1}{2} - \tau\lambda q^2 = 0. \quad (17.19)$$

Определим область абсолютной устойчивости этого метода. Для этого достаточно найти ее границу, т.е. такое множество комплексной плоскости  $z = \tau\lambda$ , где  $|q(z)| = 1$ . С этой целью выразим из (17.19)  $\tau\lambda$  через  $q$

$$z = \frac{3}{2} - \frac{2}{q} + \frac{1}{2q^2}. \quad (17.20)$$

Поскольку нас интересуют значения  $|q| = 1$ , то пусть  $q = e^{-i\varphi}$ . Отсюда и из (17.20) находим, что граница области абсолютной устойчивости задается уравнением

$$z = \frac{3}{2} - 2e^{i\varphi} + \frac{1}{2}e^{2i\varphi}. \quad (17.21)$$

При изменении аргумента  $\varphi$  от 0 до  $2\pi$  точка  $z$  из (17.21) описывает замкнутую кривую, симметричную относительно действительной оси (функция  $\sin k\varphi$  — нечетная), которая и является границей области абсолютной

устойчивости.

$$\begin{aligned}
 z &= \frac{3}{2} - 2 \cos \varphi + \frac{1}{2} \cos 2\varphi + i(-2 \sin \varphi + \frac{1}{2} \sin 2\varphi) = \\
 &= \frac{3}{2} - 2 \cos \varphi + \cos^2 \varphi - \frac{1}{2} + i(-2 \sin \varphi + \sin \varphi \cos \varphi) = \\
 &= (1 - \cos \varphi)^2 \pm i\sqrt{1 - \cos^2 \varphi}(2 - \cos \varphi) = \\
 &= (1 - t)^2 \pm i\sqrt{1 - t^2}(2 - t), \quad t = \cos \varphi.
 \end{aligned}$$

Отсюда следует, что

$$\operatorname{Re} z = (1 - t)^2 \geq 0,$$

и, следовательно, кривая расположена в правой полуплоскости. Построим ее. Мнимая часть  $z(t)$  обращается в нуль при  $t = \pm 1$ . Действительная часть  $z(t)$  при этих значениях параметра равна 0 и 4.

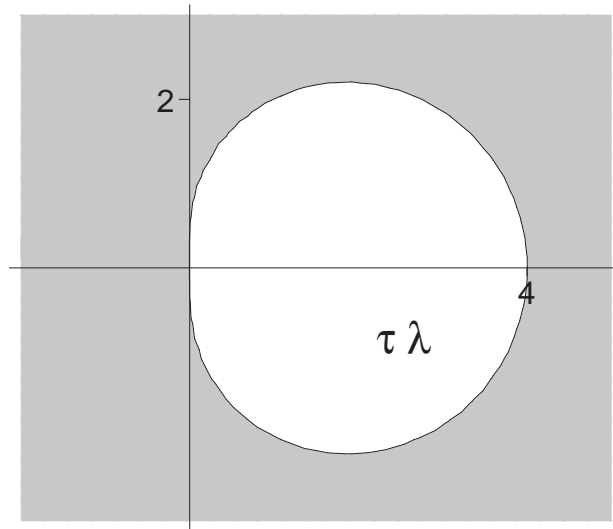


Рис. 5

Исследования показывают, что

$$\begin{aligned}
 \max_{[-1,1]} \operatorname{Im} z(t) &= \operatorname{Im} z \left( \frac{1 - \sqrt{3}}{2} \right) = \frac{(3 + \sqrt{3})\sqrt{3}}{2\sqrt{2}} \approx 2.20, \\
 \operatorname{Re} z \left( \frac{1 - \sqrt{3}}{2} \right) &= \frac{2 + \sqrt{3}}{2} \approx 1.86.
 \end{aligned}$$

Из (17.20) находим, что при

$$|q| \rightarrow \infty, \quad z \rightarrow \frac{3}{2} \in G,$$

и, следовательно, внутренность области — область неустойчивости. Тем самым, вне  $G$  (Рис. 5)  $|q| < 1$ , и метод абсолютно устойчив, а, следовательно, и  $A$ -устойчив. Этот метод второго порядка точности.

**Пример 8°.** Трехшаговая формула дифференцирования назад. (Упражнение 16.3)

$$\frac{11}{6}u_{n+1} - 3u_n + \frac{3}{2}u_{n-1} - \frac{1}{3}u_{n-2} = \tau\lambda u_{n+1}. \quad (17.22)$$

Характеристическое уравнение этого разностного уравнения имеет вид

$$\frac{11}{6}q^3 - 3q^2 + \frac{3}{2}q - \frac{1}{3} = \tau\lambda q^3. \quad (17.23)$$

Снова положим  $|q| = 1$ , т.е.  $q = e^{-i\varphi}$  и  $\tau\lambda = z$ . Тогда

$$z = \frac{11}{6} - 3e^{i\varphi} + \frac{3}{2}e^{2i\varphi} - \frac{1}{3}e^{3i\varphi}.$$

Обозначая  $\cos \varphi = t$ , после простых вычислений находим, что

$$z = -\frac{1}{3}(t-1)^2(4t-1) \pm \frac{i}{3}\sqrt{1-t^2}(4t^2-9t+8).$$

При  $t = \pm 1$   $\operatorname{Im} z = 0$ , а  $\operatorname{Re} z = 0$  или  $20/3$ . Исследования показывают, что  $\operatorname{Re} z$  как функция  $t$  принимает экстремальные значения при  $t = 1/2$  и  $t = 1$ . Значение  $t = 1$  мы уже рассмотрели, а

$$\operatorname{Re} z(1/2) = \min \operatorname{Re}(t) = -1/12, \quad \operatorname{Im} z(1/2) = \pm 3\sqrt{3}/4 \approx \pm 1.30$$

и, следовательно, часть границы области устойчивости расположена в левой полуплоскости. Как легко видеть, мнимую ось граница устойчивости пересекает при  $t = 1/4$  и

$$\operatorname{Im} z(1/4) = \pm\sqrt{15}/2 \approx \pm 1.94.$$

Экстремальные значения  $\operatorname{Im} z(t)$  принимает в точке

$$t^* = -\frac{1}{2} \left[ (2 + \sqrt{3})^{1/3} + (2 + \sqrt{3})^{-1/3} - 1 \right] \approx -0.60,$$

причем

$$\operatorname{Im} z(t^*) \approx \pm 3.96, \quad \operatorname{Re} z(t^*) \approx 2.89.$$

Область устойчивости изображена на рис. 6.

**Определение 17.8.** Линейный многошаговый метод называется  $A(\alpha)$ -устойчивым, если его область абсолютной устойчивости содержит угол

$$|\arg(-\tau\lambda)| < \alpha.$$

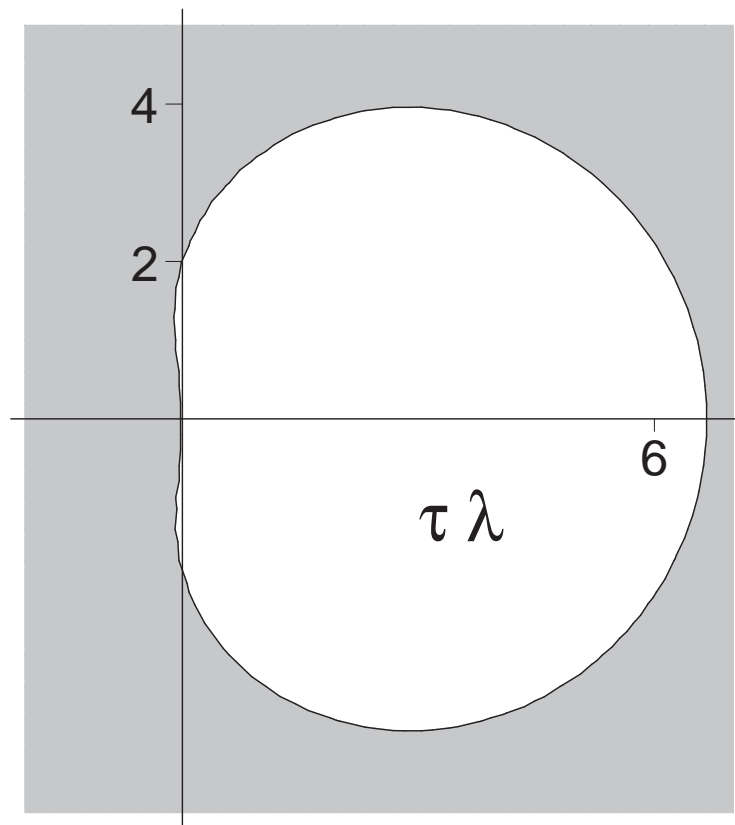


Рис. 6.

**Замечание 17.3.**  $A(\pi/2)$ - и  $A$ -устойчивости совпадают.

**Теорема 17.5.** Существуют многошаговые методы 3-го и 4-го порядков точности  $A(\alpha)$ -устойчивые при любых  $\alpha < \pi/2$ .

**Теорема 17.6.** Явные линейные многошаговые методы не являются  $A(\alpha)$ -устойчивыми ни при каких  $\alpha$ .

**Теорема 17.7.** Методы дифференцирования назад при  $k \leq 6$  являются  $A(\alpha)$ -устойчивыми при соответствующих значениях  $\alpha \neq 0$ .

**Упражнение 17.3.** Исследовать область абсолютной устойчивости двухшагового неявного метода Адамса.

## 17.4 Устойчивость методов Рунге-Кутты

Как было уже отмечено, методы (все) Рунге-Кутты являются нуль-устойчивыми. Исследуем области абсолютной устойчивости некоторых из этих

методов. Рассмотрим явный трехэтапный метод третьего порядка, задаваемый таблицей (15.48), которая имеет вид

$$\begin{array}{c|ccc} 1/2 & 1/2 & & \\ 1 & -1 & 2 & \\ \hline & 1/6 & 2/3 & 1/6 \end{array}$$

Применительно к уравнению (16.25) этот метод задается следующими соотношениями

$$\begin{aligned} Y_1 &= u_n, \\ Y_2 &= u_n + \frac{\tau\lambda}{2}Y_1, \\ Y_3 &= u_n - \tau\lambda Y_1 + 2\tau\lambda Y_2, \\ u_{n+1} &= u_n + \tau\lambda \left( \frac{1}{6}Y_1 + \frac{2}{3}Y_2 + \frac{1}{6}Y_3 \right). \end{aligned}$$

Исключая из этих соотношений промежуточные величины  $Y_1$ ,  $Y_2$  и  $Y_3$ , будем иметь

$$\begin{aligned} Y_2 &= \left( 1 + \frac{\tau\lambda}{2} \right) u_n, \\ Y_3 &= \left[ 1 - \tau\lambda + 2\tau\lambda \left( 1 + \frac{\tau\lambda}{2} \right) \right] u_n = (1 + \tau\lambda + \tau^2\lambda^2)u_n, \\ u_{n+1} &= \left\{ 1 + \tau\lambda \left[ \frac{1}{6} + \frac{2}{3} \left( 1 + \frac{\tau\lambda}{2} \right) + \frac{1}{6}(1 + \tau\lambda + \tau^2\lambda^2) \right] \right\} u_n = \\ &= \left( 1 + \tau\lambda + \frac{\tau^2\lambda^2}{2} + \frac{\tau^3\lambda^3}{6} \right) u_n. \end{aligned}$$

Это есть линейное разностное уравнение первого порядка, единственный корень характеристического уравнения которого равен

$$q = 1 + \tau\lambda + \frac{\tau^2\lambda^2}{2} + \frac{\tau^3\lambda^3}{6} = e^{\tau\lambda} + O(\tau^4\lambda^4).$$

Обозначим  $\tau\lambda$  через  $z$ . Тогда

$$q = 1 + z + \frac{z^2}{2} + \frac{z^3}{6}.$$

Этот корень характеристического уравнения есть многочлен третьей степени от  $z$  и в левой полуплоскости  $\operatorname{Re} z < 0$  ограниченным быть не может. Метод не является  $A(\alpha)$ -устойчивым ни при каком  $\alpha$ .

Рассмотрим теперь двухэтапный однократно неявный метод третьего порядка, задаваемый таблицей (15.36), которая имеет вид

$$\frac{\begin{array}{c|cc} \gamma & \gamma & 0 \\ 1-\gamma & 1-2\gamma & \gamma \end{array}}{\begin{array}{c|cc} \hline & 1/2 & 1/2 \end{array}} \quad \gamma = \frac{3 \pm \sqrt{3}}{6}. \quad (17.24)$$

Применительно к уравнению (16.25) этот метод записывается следующим образом

$$\begin{aligned} Y_1 &= u_n + \gamma\tau\lambda Y_1, \\ Y_2 &= u_n + \tau\lambda(1-2\gamma)Y_1 + \tau\lambda\gamma Y_2, \\ u_{n+1} &= u_n + \frac{\tau\lambda}{2}(Y_1 + Y_2). \end{aligned}$$

Как и в предыдущем примере, положим  $\tau\lambda = z$  и исключим  $Y_1$  и  $Y_2$ . Решая систему линейных алгебраических уравнений второго порядка относительно  $Y_1$  и  $Y_2$  (первые два уравнения) и подставляя результат в третье уравнение, находим, что

$$\begin{aligned} Y_1 &= \frac{1}{1-\gamma z} u_n, \quad Y_2 = \frac{1+(1-3\gamma)z}{(1-\gamma z)^2} u_n, \\ u_{n+1} &= \left[ 1 + \frac{z}{2} \left( \frac{1}{1-\gamma z} + \frac{1+(1-3\gamma)z}{(1-\gamma z)^2} \right) \right] u_n. \end{aligned}$$

Отсюда следует, что единственным корнем характеристического уравнения является

$$\begin{aligned} q &= \frac{1-2\gamma z + \gamma^2 z^2 + z/2 - \gamma z^2/2 + z/2 + (1-3\gamma)z^2/2}{(1-\gamma z)^2} = \\ &= \frac{1+(1-2\gamma)z + (\gamma^2 - 2\gamma + 1/2)z^2}{1-2\gamma z + \gamma^2 z^2} = \frac{P(z)}{Q(z)}. \end{aligned}$$

Этот корень является дробно-рациональной функцией, полюсом второго порядка которой является точка  $z = \gamma^{-1} = (3 \mp \sqrt{3})/6$ , расположенная в правой полуплоскости. В левой полуплоскости эта функция аналитична и, следовательно, максимум ее модуля здесь не превосходит максимума модуля на границе, т.е. при  $z = iy$ . Оценим ее модуль на мнимой оси. Имеем

$$\begin{aligned} |P(iy)|^2 &= [1 - (\gamma^2 - 2\gamma + 1/2)y^2]^2 + (1-2\gamma)^2 y^2 = \\ &= 1 - 2(\gamma^2 - 2\gamma + 1/2)y^2 + (\gamma^2 - 2\gamma + 1/2)^2 y^4 + (1-4\gamma + 4\gamma^2)y^2 = \\ &= 1 + 2\gamma^2 y^2 + (\gamma^2 - 2\gamma + 1/2)^2 y^4 \end{aligned}$$



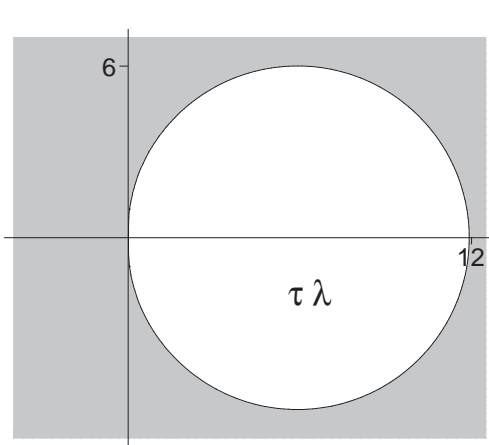


Рис. 7

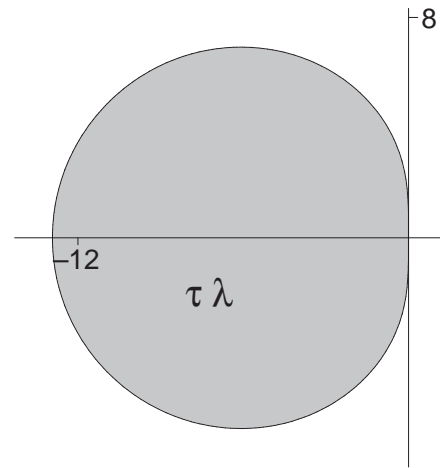


Рис. 8

и

$$|Q(iy)|^2 = (1 - \gamma^2 y^2)^2 + 4\gamma^2 y^2 = 1 + 2\gamma^2 y^2 + \gamma^4 y^4.$$

Отсюда

$$\left| \frac{P(iy)}{Q(iy)} \right|^2 = \frac{1 + 2\gamma^2 y^2 + (\gamma^2 - 2\gamma + 1/2)^2 y^4}{1 + 2\gamma^2 y^2 + \gamma^4 y^4}. \quad (17.25)$$

Эта функция не будет превосходить единицу, если ее числитель не больше знаменателя, т.е. если разность между числителем и знаменателем неположительна. Из (17.25) находим, что

$$|P(iy)|^2 - |Q(iy)|^2 = (1 - 4\gamma)(\gamma - 1/2)^2 y^4$$

и, следовательно, метод (17.24) является  $A$ -устойчивым только при  $\gamma \geq 1/4$ . Этому условию отвечает  $\gamma = (3 + \sqrt{3})/6$  из (17.24), в то время как при другом значении  $\gamma$  из (17.24) метод  $A$ -устойчивым не будет. Области абсолютной устойчивости этих методов изображены на рис. 7 и 8, соответственно.

Тем самым, один из методов (15.36), именно, отвечающий  $\gamma = (3 + \sqrt{3})/6$ , является  $A$ -устойчивым, в то время как второй таким свойством не обладает и даже не является  $A(\alpha)$ -устойчивым.

**Упражнение 17.4.** Доказать, что неявный двухэтапный метод Рунге-Кутты четвертого порядка (оптимальный двухэтапный метод) (15.38) является  $A(\alpha)$ -устойчивым.

**Упражнение 17.5.** Исследовать область абсолютной устойчивости метода (17.24) при  $\gamma = 1/4$

V

## Двухточечные краевые задачи

## Элементы теории разностных схем

### 18.1 Введение

Простейшим содержательным примером краевой задачи для обыкновенного дифференциального уравнения является следующий:

$$-u''(x) = f(x), \quad 0 < x < 1, \quad (18.1)$$

$$u(0) = g_0, \quad u(1) = g_1. \quad (18.2)$$

У краевой задачи, в отличие от задачи Коши, дополнительные условия, выделяющие единственное решение уравнения (18.1), задаются не в одной точке, а в нескольких (обычно в двух), и называются краевыми (или граничными) условиями. Это вносит дополнительные трудности в процесс решения задачи.

Мы будем изучать разностные методы решения краевых задач. Для этого на отрезке  $[0, 1]$  введем сетку

$$\bar{\omega} := \{x = x_i = ih \mid i = 0, \dots, N\}.$$

Точки  $x_i$  будем называть узлами сетки, а число  $h = 1/N$  — ее шагом. Введенная сетка является равномерной. Если бы расстояния между узлами менялось при переходе от одного узла к другому, то сетка была бы неравномерной.

Суть разностных методов решения краевых задач для дифференциальных уравнений состоит в том, что производные, входящие в дифференциальное уравнение и граничные условия, заменяются подходящими разностными отношениями. В результате краевая задача заменяется (аппроксимируется) системой алгебраических (линейных, если исходная

задача была линейной) уравнений, решение которой и принимается за приближенное решение краевой задачи.

Напомним простейшие аппроксимации первой и второй производных

$$\frac{u(x_i) - u(x_{i-1}))}{h} = u'(x_i) + O(h), \quad (18.3)$$

$$\frac{u(x_{i+1}) - u(x_i)}{h} = u'(x_i) + O(h), \quad (18.4)$$

$$\frac{u(x_{i+1}) - u(x_{i-1}))}{2h} = u'(x_i) + O(h^2), \quad (18.5)$$

$$\frac{-u(x_{i+2}) + 4u(x_{i+1}) - 3u(x_i)}{2h} = u'(x_i) + O(h^2), \quad (18.6)$$

$$\frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1}))}{h^2} = u''(x_i) + O(h^2). \quad (18.7)$$

Для справедливости соотношений (18.3) и (18.4) достаточно, чтобы  $u(x) \in C^2$ , для справедливости (18.5) и (18.6) —  $u(x) \in C^3$ , для справедливости (18.7) —  $u(x) \in C^4$ . В этом можно убедиться путем разложения левых частей (18.3)-(18.7) в точке  $x = x_i$  по формуле Тейлора.

**Упражнение 18.1.** Убедиться в справедливости (18.3)-(18.7).

**Замечание 18.1.** Если функцию  $u(x)$  заменить интерполяционным многочленом Лагранжа первой степени по узлам  $x_{i-1}$  и  $x_i$  или  $x_i$  и  $x_{i+1}$ , а затем его продифференцировать, то получим левые части соотношений (18.3), (18.4). Заменяя  $u(x)$  интерполяционным многочленом второй степени по узлам  $x_{i-1}, x_i, x_{i+1}$  или  $x_i, x_{i+1}, x_{i+2}$ , дифференцируя полученный интерполянт и полагая  $x = x_i$ , получим левые части (18.5) и (18.6), соответственно.

Воспользуемся соотношением (18.7) для замены второй производной в (18.1) разностным отношением. В узлах сетки будем иметь

$$-\frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1}))}{h^2} \approx f(x_i), \quad x_i = h, 2h, \dots, 1 - h.$$

Превратим приближенные равенства в точные путем замены точного решения  $u(x_i)$  в узле  $x_i$  на приближенное  $u_i^h$ :

$$-\frac{u_{i+1}^h - 2u_i^h + u_{i-1}^h}{h^2} = f_i, \quad i = 1, 2, \dots, N - 1. \quad (18.8)$$

Соотношения (18.8) представляют собой систему  $(N - 1)$  линейных алгебраических уравнений с  $N + 1$  неизвестными  $u_0^h, u_1^h, \dots, u_N^h$ . Система (18.8) недоопределена (как и следовало ожидать). Воспользуемся граничными условиями (18.2) и положим

$$u_0^h = g_0, \quad u_N^h = g_1. \quad (18.9)$$

Решение системы (18.8), (18.9), если оно существует, будем называть приближенным решением задачи (18.1), (18.2).

## 18.2 Основные понятия теории разностных схем

Обозначим дифференциальное выражение, стоящее в левой части (18.1), через  $Lu$ . Тогда дифференциальное уравнение (18.1) примет вид

$$Lu = f(x), \quad 0 < x < 1. \quad (18.10)$$

Граничные условия (18.2) запишем в виде

$$lu = g. \quad (18.11)$$

Аналогично, разностное выражение, стоящее в левой части (18.8), обозначим через  $L^h u^h$ . Тогда из (18.8) будем иметь

$$L^h u_i^h = f_i^h, \quad i = 1, 2, \dots, N - 1, \quad (18.12)$$

где  $f_i^h = f_i$ . Граничные условия (18.9) запишем в виде, аналогичном (18.11)

$$l^h u^h = g^h. \quad (18.13)$$

**Определение 18.1.** Сеточная функция

$$\Psi_v(x) := L^h v - Lv, \quad x \in \omega, \quad (18.14)$$

определенная на сетке  $\omega$ , где  $v$  — достаточно гладкая функция, заданная на  $[0, 1]$ , называется погрешностью аппроксимации дифференциального выражения  $Lv$  разностным выражением  $L^h v$ .

**Определение 18.2.** Разностное выражение  $L^h v$  аппроксимирует дифференциальное выражение  $Lv$ , если погрешность аппроксимации  $\Psi_v \rightarrow 0$  (в каком-нибудь смысле) при  $h \rightarrow 0$ .

**Определение 18.3.** Сеточная функция

$$z = u^h - u, \quad x \in \bar{\omega}, \quad (18.15)$$

где  $u^h$  — решение задачи (18.12), (18.13), а  $u$  — решение задачи (18.10), (18.11), называется погрешностью решения.

Сформулируем задачу для погрешности решения  $z$ . Подставим в (18.12), (18.13)  $u^h$ , выражаемое из (18.15) через  $z$  и  $u$ :  $u^h = z + u$ . Будем иметь

$$L^h z = f^h - L^h u, \quad l^h z = g^h - l^h u. \quad (18.16)$$

**Определение 18.4.** Функция

$$\Psi = f^h - L^h u, \quad x \in \omega, \quad (18.17)$$

являющаяся правой частью уравнения для погрешности решения (18.16), называется погрешностью аппроксимации уравнения (18.10) уравнением (18.12).

**Определение 18.5.** Функция

$$\psi = g^h - l^h u, \quad (18.18)$$

являющаяся правой частью в граничных условиях для погрешности решения (18.16), называется погрешностью аппроксимации граничных условий (18.11) граничными условиями (18.13).

**Замечание 18.2.** Так как в силу (18.10)  $Lu - f = 0$ , то, добавляя этот нуль к представлению погрешности аппроксимации (18.17), будем иметь

$$\Psi = f^h - L^h u = f^h - f - (L^h u - Lu) = (f^h - f) - \Psi_u, \quad (18.19)$$

где  $\Psi_u$  определяется соотношением (18.14). Тем самым, погрешность аппроксимации уравнения представляет собой разность между погрешностью аппроксимации правой части и погрешностью аппроксимации дифференциального выражения. Аналогичные представления имеют место и для погрешности аппроксимации граничных условий:

$$\psi = g^h - l^h u = g^h - g - (l^h u - lu) = (g^h - g) - \psi_u. \quad (18.20)$$

**Определение 18.6.** Задача (18.12), (18.13) аппроксимирует задачу (18.10), (18.11), если  $\Psi$  и  $\psi$  стремятся к нулю при  $h \rightarrow 0$  вместе с  $\Psi_u$  и  $\psi_u$ .

**Определение 18.7.** Решение задачи (18.12), (18.13) сходится к решению задачи (18.10), (18.11), если  $z \rightarrow 0$  (в каком-либо смысле) при  $h \rightarrow 0$ .

**Определение 18.8.** Задача (18.12), (18.13) аппроксимирует задачу (18.10), (18.11) с погрешностью порядка  $n > 0$ , если

$$\|\Psi_u\|_{(1)} = o(1), \quad \|\psi_u\|_{(2)} = o(1), \quad \|\Psi\|_{(1)} = O(h^n), \quad \|\psi\|_{(2)} = O(h^n)$$

**Определение 18.9.** Решение задачи (18.12), (18.13) сходится к решению задачи (18.10), (18.11) со скоростью  $O(h^n)$ , если

$$\|z\|_{(3)} = O(h^n).$$

Проиллюстрируем введенные понятия на примере задачи (18.1), (18.2). Так как в данном случае  $L = -d^2v/dx^2$ , а

$$L^h v = -\frac{v(x_{i+1}) - 2v(x_i) + v(x_{i-1}))}{h^2},$$

то, в силу (18.7),

$$\Psi_v = O(h^2),$$

т.е. дифференциальное выражение  $v''$  аппроксимируется разностным выражением  $(v_{i+1} - 2v_i + v_{i-1})/h^2$  на функциях  $v(x) \in C^4$  с погрешностью  $O(h^2)$ .

Далее, так как  $f_i^h = f(x_i)$ , то с учетом (18.19) заключаем, что дифференциальное уравнение (18.1) аппроксимируется разностным уравнением (18.8) с погрешностью  $O(h^2)$ , если  $u(x) \in C^4[0, l]$ .

Наконец,

$$\begin{aligned} lu &= \{u(0), u(1)\}, \\ l^h u &= \{u_0, u_N\}, \\ g &= \{g_0, g_1\} = g^h, \end{aligned}$$

так что

$$\psi = g^h - l^h u = 0.$$

Итак, задача (18.8), (18.9) аппроксимирует задачу (18.1), (18.2) (при  $u(x) \in C^4[0, l]$ ) с погрешностью  $O(h^2)$ .

Очевидно, что, если вместо уравнения (18.1) рассмотреть уравнение

$$L_1 u := -u''(x) + q(x)u(x) = f(x), \quad x \in (0, 1) \quad (18.21)$$

и аппроксимировать его разностным уравнением

$$L_1^h u^h := -\frac{u_{i+1}^h - 2u_i^h + u_{i-1}^h}{h^2} + q(x_i)u_i^h = f(x_i), \quad i = 1, 2, \dots, N-1, \quad (18.22)$$

то задача (18.22), (18.9) будет аппроксимировать задачу (18.21), (18.2) тоже с погрешностью  $O(h^2)$ .

### 18.3 Разрешимость и сходимость

Исследуем вопрос о сходимости решения разностной задачи к решению задачи дифференциальной. Для уравнения (18.21) это сделать несколько проще, чем для уравнения (18.1). Поэтому к нему мы в первую очередь и обратимся. Но сначала установим существование и единственность решения задачи (18.22), (18.9).

**Теорема 18.1.** *Если*

$$q(x) \geq c_1 > 0, \quad 0 < x < 1, \quad (18.23)$$

*то решение задачи (18.22), (18.9) существует, единственно, и для него справедлива априорная оценка*

$$\|u^h\|_{L_\infty^h} := \max_i |u_i^h| \leq |g_0| + |g_1| + \frac{1}{c_1} \|f\|_{L_\infty^h}. \quad (18.24)$$

**Доказательство.** Задача (18.22), (18.9) представляет собой систему линейных алгебраических уравнений с квадратной матрицей порядка  $(N+1)$ . Поэтому всегда существует такая правая часть  $[g_0, f_1, \dots, f_{N-1}, g_1]$  этой системы (берется первое уравнение из (18.9), затем последовательно все уравнения (18.22) и, наконец, второе уравнение (18.9)), что решение  $u^h$  существует. Например, возьмем произвольный набор чисел  $u_0^h, u_1^h, \dots, u_N^h$  и подставим его в левые части (18.22), (18.9). Этим мы определим правые части (18.22), (18.9), при которых решение заведомо существует.

Получим априорную оценку этого решения. Пусть

$$\|u^h\|_{L_\infty^h} = |u_{i_0}^h|.$$

Если  $i_0 = 0$  или  $i_0 = N$ , то в силу краевых условий (18.9)

$$\max_i |u_i^h| \leq \max\{|g_0|, |g_1|\} \leq |g_0| + |g_1|, \quad (18.25)$$



что согласуется с (18.24). В противном случае максимум модуля достигается во внутреннем узле  $x_{i_0} \in \omega$ . Запишем уравнение (18.22) в этом узле

$$-\frac{u_{i_0-1}^h - 2u_{i_0}^h + u_{i_0+1}^h}{h^2} + q_{i_0}u_{i_0}^h = f_{i_0}.$$

Если  $u_{i_0}^h \geq 0$ , то

$$-[(\underbrace{u_{i_0-1}^h - u_{i_0}^h}_{\wedge 0}) + (\underbrace{u_{i_0+1}^h - u_{i_0}^h}_{\wedge 0})] \geq 0$$

и, следовательно,

$$q_{i_0}u_{i_0}^h \leq f_{i_0}.$$

Отсюда с учетом (18.23)

$$0 \leq u_{i_0}^h \leq \frac{f_{i_0}}{q_{i_0}} \leq \frac{1}{c_1} \|f\|_{L_\infty^h}. \quad (18.26)$$

Если же  $u_{i_0}^h < 0$ , то оно может быть только минимальным значением, и, следовательно,

$$-[(\underbrace{u_{i_0-1}^h - u_{i_0}^h}_{\vee 0}) + (\underbrace{u_{i_0+1}^h - u_{i_0}^h}_{\vee 0})] \leq 0. \quad (18.27)$$

Поэтому

$$q_{i_0}u_{i_0}^h \geq f_{i_0}.$$

Отсюда

$$-q_{i_0}|u_{i_0}^h| \geq f_{i_0}$$

и снова

$$|u_{i_0}^h| \leq -\frac{f_{i_0}}{q_{i_0}} \leq \frac{1}{c_1} \|f\|_{L_\infty^h}. \quad (18.28)$$

Собирая оценки (18.25), (18.26), (18.28), приходим к (18.24). Априорная оценка получена.

Докажем теперь, что решение единственно. Допустим противное, т.е. допустим существование двух решений  $u_{(1)}^h$  и  $u_{(2)}^h$ . Очевидно, что их разность  $z = u_{(1)}^h - u_{(2)}^h$  удовлетворяет однородному уравнению (18.22) и однородным граничным условиям (18.9). В силу априорной оценки (18.24)

$$\max_i |z_i| \leq 0.$$

Следовательно,  $z_i \equiv 0$ , что противоречит предположению. Мы доказали, что однородная система (18.22), (18.9) имеет лишь тривиальное решение.

Следовательно, матрица этой системы невырождена, и задача (18.22), (18.9) имеет единственное решение при любых  $g_0$ ,  $g_1$  и  $f_i$ . Теорема доказана.

**Теорема 18.2.** *Если выполнено условие (18.23), и решение  $u(x)$  задачи (18.21), (18.2) принадлежит  $C^4[0, 1]$ , то решение  $u^h$  задачи (18.22), (18.9) сходится к решению задачи (18.21), (18.2) со скоростью  $O(h^2)$ , т.е.*

$$|u(x_i) - u_i^h| = O(h^2).$$

**Доказательство.** Напишем задачу для погрешности решения  $z_i = u_i^h - u(x_i)$ . Будем иметь

$$-\frac{z_{i+1} - 2z_i + z_{i-1}}{h^2} + q_i z_i = \Psi_i, \quad z_0 = z_N = 0. \quad (18.29)$$

К задаче (18.29) применим теорему 18.1, в силу которой

$$\max_i |z_i| \leq \frac{1}{c_1} \max_i |\Psi_i|.$$

Но в силу вышедоказанного  $\Psi_i = O(h^2)$ , что и доказывает теорему.

**Замечание 18.3.** Более детальный анализ показывает, что

$$\max_i |u_i^h - u(x_i)| \leq \frac{1}{c_1} \max_{x \in [0,1]} |u^{IV}(x)| \frac{h^2}{12}.$$

**Теорема 18.3** (О монотонности). *Пусть  $L_1^h$  определяется (18.22). Тогда, если выполнено условие*

$$q_i \geq 0, \quad i = 1, 2, \dots, N-1, \quad (18.30)$$

*а сеточная функция  $U_i$ ,  $i = 0, 1, \dots, N$  такова, что*

$$U_0 \geq 0, \quad U_N \geq 0 \quad (18.31)$$

*и*

$$L_1^h U_i \geq 0, \quad i = 1, 2, \dots, N-1, \quad (18.32)$$

*то*

$$U_i \geq 0, \quad i = 1, 2, \dots, N-1. \quad (18.33)$$

**Доказательство.** Допустим противное, т.е. допустим, что функция  $U_i$  может принимать отрицательные значения. Тогда существует такой узел  $x_{i_0}$ ,  $i_0 \in \{1, 2, \dots, N-1\}$ , что

$$\min_i U_i = U_{i_0} < 0 \quad (18.34)$$

и с учетом (18.27) выражение  $-(U_{i_0-1} - 2U_{i_0} + U_{i_0+1})$  либо строго меньше нуля, либо нулю равняется. Исследуем обе эти возможности. Если  $-(U_{i_0-1} - 2U_{i_0} + U_{i_0+1}) < 0$ , то с учетом (18.30) и (18.34)

$$L_1^h U_{i_0} = -\frac{U_{i_0-1} - 2U_{i_0} + U_{i_0+1}}{h^2} + q_{i_0} U_{i_0} < 0,$$

и мы пришли к противоречию с (18.32). Если  $(U_{i_0-1} - 2U_{i_0} + U_{i_0+1}) = 0$ , а  $q_{i_0} \neq 0$ , мы снова получаем противоречие. Для выхода из этих противоречий мы должны предположить, что  $q_{i_0} = 0$  и  $(U_{i_0-1} - 2U_{i_0} + U_{i_0+1}) = 0$ . Но в силу (18.27), (18.34) это означает, что  $U_{i_0-1} = U_{i_0} = U_{i_0+1} < 0$ , и в качестве  $i_0$  из (18.34) можно взять также  $(i_0 - 1)$  или  $(i_0 + 1)$ . Делая этот выбор, мы теми же рассуждениями приходим к утверждению, что и  $U_{i_0-2} = U_{i_0}$  (или  $U_{i_0+2} = U_{i_0}$ ). И т.д. Поскольку в силу (18.31), (18.34) функция  $U_i$ ,  $i = \overline{0, N}$  не является постоянной, то существует такой узел  $x_{i_1}$ ,  $i_1 \in \{1, 2, \dots, N-1\}$ , что  $U_{i_1} = U_{i_0}$ , а  $U_{i_1-1}$  или  $U_{i_1+1}$  больше  $U_{i_1}$ . В этом узле  $-(U_{i_1-1} - 2U_{i_1} + U_{i_1+1}) < 0$ , и мы вернулись к уже рассмотренному случаю, который привел нас к противоречию с (18.32). Все противоречия снимаются, если мы откажемся от предположения, что  $U_i$  может принимать отрицательные значения. Теорема доказана.

**Определение 18.10.** Матрица  $A$  называется обратно монотонной, если для любого вектора  $x$  из условия  $Ax \geq 0$  следует  $x \geq 0$ .

**Упражнение 18.2.** Доказать, что обратно монотонная матрица невырождена, а обратная к ней имеет только неотрицательные элементы.

**Теорема 18.4** (Принцип сравнения). Пусть  $u_i^h$  — решение задачи (18.22), (18.9), а  $U_i$  — решение следующей задачи:

$$L_1^h U_i = F_i, \quad i = 1, 2, \dots, N-1, \quad U_0 = G_0, \quad U_N = G_1.$$

Пусть

$$|f_i| \leq F_i, \quad |g_0| \leq G_0, \quad |g_1| \leq G_1. \quad (18.35)$$

Тогда, если выполнено условие (18.30), то

$$|u_i^h| \leq U_i, \quad i = 1, 2, \dots, N-1. \quad (18.36)$$

**Доказательство.** Легко видеть, что функция  $(U_i - u_i^h)$  является решением задачи

$$\begin{aligned} L_1^h(U - u^h)_i &= F_i - f_i, \quad i = 1, 2, \dots, N-1, \\ U_0 - u_0^h &= G_0 - g_0, \quad U_N - u_N^h = G_1 - g_1. \end{aligned}$$

В силу (18.35) и теоремы 18.3 заключаем, что  $U_i - u_i^h \geq 0$ . Из аналогичных соображений находим, что и  $U_i + u_i^h \geq 0$ . Тем самым,  $-U_i \leq u_i^h \leq U_i$ , и теорема доказана.

**Определение 18.11.** Функция  $U_i$  из (18.36) называется *барьером*.

**Теорема 18.5.** Для решения задачи (18.22), (18.9) при выполнении условия (18.30) справедлива априорная оценка

$$\|u^h\|_{L_\infty^h} \leq |g_0| + |g_1| + \frac{1}{8} \|f\|_{L_\infty^h}.$$

**Доказательство.** Введем в рассмотрение функцию

$$U_i = |g_0|(1 - x_i) + x_i|g_1| + c x_i(1 - x_i) \geq 0, \quad (18.37)$$

где  $c > 0$  — некоторая постоянная. Очевидно, что  $U_0 = |u_0^h|$ ,  $U_N = |u_N^h|$ . Легко проверить, что

$$L_1^h U_i = 2c + q_i U_i =: F_i \geq 2c.$$

Пусть  $c = 1/2 \max_i |f_i|$ . Тогда  $|f_i| \leq F_i$ , и мы находимся в условиях теоремы 18.4, т.е.  $|u_i^h| \leq U_i$ . Но

$$\max_i U_i \leq |g_0| + |g_1| + c/4.$$

Теорема доказана.

**Упражнение 18.3.** Сформулировать и доказать теорему о скорости сходимости разностной задачи (18.22), (18.9).

## 18.4 Метод баланса (конечных объемов)

Рассмотрим общее самосопряженное уравнение второго порядка

$$L_2 u := -\frac{d}{dx} \left( p(x) \frac{du}{dx} \right) + q(x)u = f(x), \quad 0 < x < 1 \quad (18.38)$$

и изучим вопрос о его аппроксимации. На первый взгляд кажется вполне естественным раздифференцировать первое слагаемое левой части (18.38)

$$-p(x)\frac{d^2u}{dx^2} - p'(x)\frac{du}{dx} + q(x)u = f(x) \quad (18.39)$$

и в этом виде заменить  $d^2u/dx^2$  и  $du/dx$  соответствующими разностными отношениями. Но так поступать плохо в силу целого ряда причин. В частности, уравнение (18.38) является формально самосопряженным по Лагранжу (симметричным, т.е. если  $u(x)$  и  $v(x)$  обращаются в нуль при  $x = 0$  и  $x = 1$ , то  $\int_0^1 vL_2u dx = \int_0^1 uL_2v dx$ ). Сравнить с симметричной матрицей  $A = A^T : (Ax, y) = (x, Ay)$ . Если же аппроксимировать (18.39), которое эквивалентно (18.38) при гладкой  $p(x)$ , то аппроксимация, вообще говоря, симметричной не будет. Уравнение (18.38) нужно аппроксимировать сразу в исходном виде. Как и раньше, будем предполагать, что на  $[0, 1]$  задана равномерная сетка узлов  $x_i = ih$ ,  $h = 1/N$ ,  $i = 0, 1, \dots, N$ , которую будем называть основной. Помимо основной сетки введем на  $[0, 1]$  так называемую "сдвинутую" сетку с узлами  $x_{i+1/2} = x_i + h/2$ ,  $i = 0, 1, \dots, N - 1$ .

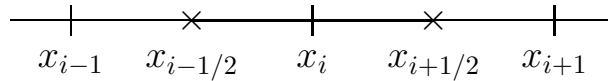


Рис. 1

Будем строить аппроксимацию (18.38) при помощи метода баланса (метода конечных объемов). Интегрируя уравнение (18.38) по отрезку  $(x_{i-1/2}, x_{i+1/2})$ , будем иметь

$$-p(x_{i+1/2})u'(x_{i+1/2}) + p(x_{i-1/2})u'(x_{i-1/2}) + \int_{x_{i-1/2}}^{x_{i+1/2}} [q(x)u(x) - f(x)] dx = 0. \quad (18.40)$$

Заменим в (18.40) интеграл квадратурной формулой прямоугольников, а производные — соответствующими разностными отношениями. Именно

$$\int_{x_{i-1/2}}^{x_{i+1/2}} [q(x)u - f(x)] dx \approx q_i u_i h - f_i h, \quad (18.41)$$

$$u'_{i+1/2} \approx \frac{u_{i+1} - u_i}{h}, \quad u'_{i-1/2} \approx \frac{u_i - u_{i-1}}{h}.$$

Подставляя (18.41) в (18.40), получим приближенное равенство. Заменяя приближенное равенство на точное, получим уравнение для приближенного решения. После деления на  $h$  оно примет вид:

$$-\frac{1}{h} \left[ p_{i+1/2} \frac{u_{i+1}^h - u_i^h}{h} - p_{i-1/2} \frac{u_i^h - u_{i-1}^h}{h} \right] + q_i u_i^h = f_i, \quad i = 1, 2, \dots, N-1. \quad (18.42)$$

Введем следующие обозначения

$$u_x := u_{x,i} := \frac{u_{i+1} - u_i}{h} \text{ — правое разностное отношение,}$$

$$u_{\bar{x}} := u_{\bar{x},i} := \frac{u_i - u_{i-1}}{h} \text{ — левое разностное отношение.}$$

Очевидно, что  $v_{x,i} \equiv v_{\bar{x},i+1}$ . Далее,

$$\begin{aligned} \frac{v_{i+1} - 2v_i + v_{i-1}}{h^2} &= \frac{1}{h} \left[ \frac{v_{i+1} - v_i}{h} - \frac{v_i - v_{i-1}}{h} \right] = \\ &= \frac{1}{h} (v_{x,i} - v_{\bar{x},i}) = \frac{1}{h} (v_{\bar{x},i+1} - v_{\bar{x},i}) = (v_{\bar{x}})_{x,i} = v_{\bar{x}x,i} =: v_{\bar{x}x}. \end{aligned}$$

Используя введенные обозначения, уравнение (18.42) можно переписать так:

$$L_2^h u^h := - (p^h u_{\bar{x}}^h)_{x,i} + q_i^h u_i^h = f_i^h, \quad i = 1, 2, \dots, N-1, \quad (18.43)$$

где

$$p^h := p_i^h := p(x_i - h/2), \quad q^h := q_i^h := q(x_i), \quad f^h := f_i^h := f(x_i). \quad (18.44)$$

## 18.5 Аппроксимация граничных условий

Применим теперь метод баланса для построения аппроксимации граничного условия, содержащего производную. Пусть для уравнения (18.38) в точке  $x = 0$  (граничной точке) задано граничное условие

$$\alpha \frac{du}{dx}(0) + \beta u(0) = \gamma. \quad (18.45)$$

Граничное условие (18.45) содержит в себе все основные граничные условия для уравнения (18.38): именно, граничные условия первого рода ( $\alpha = 0$ ), второго рода ( $\beta = 0$ ) и третьего рода. Нас будут интересовать граничные условия второго и третьего рода, т.е. условия, содержащие производную. Простейшая аппроксимация условия (18.45) имеет вид

$$\alpha \frac{u_1^h - u_0^h}{h} + \beta u_0^h = \gamma. \quad (18.46)$$

**Упражнение 18.4.** Доказать, что погрешность аппроксимации граничного условия (18.45) граничным условием (18.46) при  $\alpha \neq 0$  есть  $O(h)$ .

Мы не будем заниматься этой аппроксимацией из-за того, что она имеет большую погрешность. Построим другую аппроксимацию условия (18.45), но прежде его несколько преобразуем. По предположению  $\alpha \neq 0$ , и на этот коэффициент условие (18.45) можно разделить. Коэффициент  $p(x)$  уравнения (18.38) будем предполагать строго положительным

$$p(x) \geq c_0 > 0, \quad (18.47)$$

и домножение (18.45) на  $-p(0)$  приведет к эквивалентному уравнению. Будем вместо (18.45) рассматривать граничное условие

$$-p(0)\frac{du}{dx}(0) + \varkappa_0 u(0) = g_0, \quad (18.48)$$

которое при  $\alpha = -p(0) \neq 0$ ,  $\beta = \varkappa_0$  и  $\gamma = g_0$  совпадает с (18.45). Комбинация  $p(0)u'(0)$  в (18.48) хороша уже тем, что величина  $-p(x)u'(x)$  имеет смысл потока и фигурирует в самом уравнении (18.38). Знак минус перед производной должен свидетельствовать о том, что производная берется по "внешней нормали": производная  $u'(0)$  вычислена по направлению внутрь отрезка  $[0, 1]$ , а производная  $-u'(0)$  — по направлению, выходящему из отрезка.

Чтобы построить аппроксимацию (18.48), проинтегрируем уравнение (18.38) по отрезку  $(0, h/2)$ . Будем иметь

$$-p(h/2)\frac{du}{dx}(h/2) + p(0)\frac{du}{dx}(0) + \int_0^{h/2} [q(x)u(x) - f(x)] dx = 0. \quad (18.49)$$

Затем выразим  $p(0)u'(0)$  из (18.48)

$$p(0)\frac{du}{dx}(0) = \varkappa_0 u(0) - g_0, \quad (18.50)$$

аппроксимируем производную

$$\frac{du}{dx}(h/2) \approx \frac{u_1 - u_0}{h} \quad (18.51)$$

и аппроксимируем интеграл в (18.49) квадратурной формулой "левых прямоугольников"

$$\int_0^{h/2} [q(x)u(x) - f(x)] dx \approx [q(0)u(0) - f(0)]\frac{h}{2}. \quad (18.52)$$

Подставляя (18.50)-(18.52) в (18.49), получим приближенное равенство, которое превратим в точное путем замены точного решения  $u(x)$  на приближенное  $u^h(x)$ . В результате будем иметь

$$-p_{1/2} \frac{u_1^h - u_0^h}{h} + \left( \varkappa_0 + \frac{h}{2} q_0 \right) u_0^h = g_0 + \frac{h}{2} f_0$$

или, принимая обозначения (18.44),

$$-p_1^h u_{\bar{x},1}^h + \left( \varkappa_0 + \frac{h}{2} q_0^h \right) u_0^h = g_0 + \frac{h}{2} f_0^h. \quad (18.53)$$

Соотношение (18.53) представляет собой искомую аппроксимацию.

## 18.6 Исследование погрешности аппроксимации

Исследуем погрешность аппроксимации разностной схемы (18.43). Исследуем даже более общую схему. Пусть разностная схема имеет вид

$$-\frac{1}{h} [b_i u_{x,i}^h - a_i u_{\bar{x},i}^h] + q_i^h u_i^h = f_i^h. \quad (18.54)$$

Погрешность аппроксимации этой схемы есть

$$\begin{aligned} \Psi_i &= f_i^h + \frac{1}{h} [b_i u_{x,i} - a_i u_{\bar{x},i}] - q_i^h u_i = \\ &= [f_i^h - f(x_i)] - [q_i^h - q(x_i)] u_i + \frac{1}{h} [b_i u_{x,i} - a_i u_{\bar{x},i}] - (pu')'_i. \end{aligned} \quad (18.55)$$

При  $u(x) \in C^4[0, 1]$  имеют место следующие разложения

$$\begin{aligned} u_{x,i} &= u'_i + \frac{h}{2} u''_i + \frac{h^2}{6} u'''_i + O(h^3), \\ u_{\bar{x},i} &= u'_i - \frac{h}{2} u''_i + \frac{h^2}{6} u'''_i + O(h^3). \end{aligned}$$

Подставляя эти соотношения в (18.55), будем иметь

$$\begin{aligned} \Psi_i &= \frac{1}{h} \left[ b_i \left( u'_i + \frac{h}{2} u''_i + \frac{h^2}{6} u'''_i + O(h^3) \right) - a_i \left( u'_i - \frac{h}{2} u''_i + \frac{h^2}{6} u'''_i + O(h^3) \right) \right] - \\ &- (p'u' + pu'') - [q_i^h - q(x_i)] u_i + [f_i^h - f(x_i)] = \\ &= \left[ \frac{b_i - a_i}{h} - p'_i \right] u'_i + \left[ \frac{b_i + a_i}{2} - p_i \right] u''_i + h \frac{b_i - a_i}{6} u'''_i + O(h^2) - \\ &- (q_i^h - q_i) u_i + (f_i^h - f_i). \end{aligned}$$



Отсюда находим, что для аппроксимации  $O(h^2)$  необходимо и достаточно выполнения условий

$$\begin{aligned} 1^\circ. \quad & \frac{b_i - a_i}{h} - p'_i = O(h^2), \\ 2^\circ. \quad & \frac{b_i + a_i}{2} - p_i = O(h^2), \\ 3^\circ. \quad & q_i^h - q_i = O(h^2), \\ 4^\circ. \quad & f_i^h - f_i = O(h^2). \end{aligned} \tag{18.56}$$

Для схемы (18.43), (18.44) условия (18.56<sub>3</sub>) и (18.56<sub>4</sub>) очевидны. Обратимся к (18.56<sub>1</sub>) и (18.56<sub>2</sub>). Имеем

$$\begin{aligned} b_i = p_{i+1/2} &= p_i + \frac{h}{2}p'_i + \frac{h^2}{8}p''_i + O(h^3), \\ a_i = p_{i-1/2} &= p_i - \frac{h}{2}p'_i + \frac{h^2}{8}p''_i + O(h^3). \end{aligned}$$

Отсюда

$$\frac{b_i - a_i}{h} = p'_i + O(h^2), \quad \frac{b_i + a_i}{2} = p_i + O(h^2).$$

**Теорема 18.6.** *Если решение уравнения (18.38) обладает четвертыми непрерывными производными, то разностная схема (18.43), (18.44) имеет погрешность аппроксимации  $O(h^2)$ .*

**Упражнение 18.5.** Доказать, что разностная схема (18.43) при  $b_i = a_{i+1}$  и

$$\text{а) } a_i = \frac{p_i + p_{i-1}}{2}, \quad q_i^h = q_i, \quad f_i^h = f_i, \tag{18.57}$$

$$\text{б) } a_i = \frac{1}{h} \int_{x_{i-1}}^{x_i} p(x) dx, \quad q_i^h = \frac{1}{h} \int_{x_{i-1}}^{x_{i+1}} q(x)(1 - |x - x_i|) dx, \tag{18.58}$$

$$f_i^h = \frac{1}{h} \int_{x_{i-1}}^{x_{i+1}} f(x)(1 - |x - x_i|) dx$$

имеет погрешность аппроксимации  $O(h^2)$ .

Исследуем погрешность аппроксимации  $\psi_0$  граничного условия (18.53).

Имеем

$$\begin{aligned}\psi_0 &:= g_0 + \frac{h}{2}f_0 + p_{1/2}\frac{u_1 - u_0}{h} - (\varkappa_0 + \frac{h}{2}q_0)u_0 = \\ &= g_0 + \frac{h}{2}f_0 + \left(p_0 + \frac{h}{2}p'_0 + O(h^2)\right) \left(u'_0 + \frac{h}{2}u''_0 + O(h^2)\right) - (\varkappa_0 + \frac{h}{2}q_0)u_0 = \\ &= (p_0u'_0 - \varkappa_0u_0 + g_0) + \frac{h}{2}(p_0u''_0 + p'_0u'_0 - q_0u_0 + f_0) + O(h^2).\end{aligned}$$

Первая скобка в этом представлении равна нулю в силу (18.48), а вторая — в силу уравнения (18.38), продолженного по непрерывности с  $(0, 1)$  на  $[0, 1]$ . Тем самым, погрешность аппроксимации граничного условия (18.53) на решении уравнения (18.38) есть  $O(h^2)$ .

**Упражнение 18.6.** Методом баланса построить аппроксимацию граничного условия

$$p(1)\frac{du}{dx}(1) + \varkappa_1u(1) = g_1 \quad (18.59)$$

и исследовать погрешность полученной аппроксимации.

**Теорема 18.7.** Пусть выполнены условия

$$p_i^h \geq c_0 > 0, \quad q_i^h \geq c_1 > 0, \quad \varkappa_0 > 0. \quad (18.60)$$

Тогда существует единственное решение задачи (18.43), (18.53), (18.61)

$$u_N^h = g_1, \quad (18.61)$$

и для него справедлива априорная оценка

$$\max_i |u_i^h| \leq \frac{|g_0|}{\varkappa_0} + |g_1| + \max_i \frac{|f_i|}{c_1}. \quad (18.62)$$

**Упражнение 18.7.** Доказать теорему 18.7.

**Теорема 18.8.** Если выполнены условия (18.60), и решение задачи (18.38), (18.48), (18.63)  $u(x) \in C^4[0, 1]$ ,

$$u(1) = g_1, \quad (18.63)$$

то решение  $u^h$  задачи (18.43), (18.44), (18.53), (18.61) сходится к решению задачи (18.38), (18.48), (18.63) со скоростью  $O(h^2)$  равномерно по  $x_1 \in \omega$ , т.е.

$$\max_i |u(x_i) - u_i^h| = O(h^2).$$

**Упражнение 18.8.** Доказать теорему 18.8.

## 18.7 Уравнения с разрывными коэффициентами

Вновь обратимся к уравнению (18.38), но теперь разрешим коэффициентам этого уравнения и его правой части иметь разрывы первого рода в некотором конечном числе точек интервала  $(0, 1)$ . Достаточно рассмотреть случай, когда такая точка одна. Пусть это точка  $\xi \in (0, 1)$  и, например,  $p(\xi + 0) \neq p(\xi - 0)$ . Для того, чтобы уравнение (18.38) имело смысл, необходимо, чтобы в точке  $x = \xi$  функция  $u(x)$  была непрерывна, т.е.

$$u(x + 0) = u(x - 0).$$

Это условие принято записывать в виде

$$[u] \Big|_{x=\xi} = 0, \tag{18.64}$$

где квадратные скобки обозначают скачок функции, т.е.

$$[v] \Big|_{x=\xi} := v(\xi + 0) - v(\xi - 0).$$

Равенство нулю этого скачка как раз и говорит о том, что функция непрерывна в этой точке.

Для того, чтобы можно было выполнить второе дифференцирование в (18.38), необходима не непрерывность производной функции  $u(x)$ , а непрерывность всего выражения, стоящего под знаком внешней производной, т.е. потока

$$w(x) := -p(x)u'(x).$$

Тем самым, вторым условием должно быть условие

$$[pu'] \Big|_{x=\xi} = 0. \tag{18.65}$$

Условия (18.64), (18.65) обычно называют условиями сопряжения. Разрывы в коэффициенте  $q(x)$  и правой части  $f(x)$  никак не влияют на вид условий сопряжения.

Если коэффициенты уравнения (18.38) имеют точки разрыва, то при аппроксимации этого уравнения желательно указанные точки включать в число узловых точек сетки. В противном случае, вообще говоря, погрешность аппроксимации станет слишком большой, и точность разностной схемы уменьшится по сравнению с гладким случаем. Это, правда, почти всегда приводит к неравномерной (кусочно равномерной) сетке, но этой проблемы мы коснемся чуть позже.

Итак, пусть коэффициенты уравнения (18.38) и его правая часть имеют разрыв в точке  $\xi \in (0, 1)$ . Будем считать, что эта точка включена в число узлов сетки, и пусть, для простоты, сетка осталась равномерной. Так будет, например, если  $\xi = 1/2$ , а сетка на  $[0, 1]$  имеет шаг  $h = 1/2N$ . Пусть номер узла, отвечающего  $\xi$ , есть  $j$ . В силу непрерывности искомого решения в это точке (18.64) на сетке значению  $u(\xi)$  будет отвечать просто  $u_j^h$ . Построим аппроксимацию второго условия сопряжения (18.65). Для этого в очередной раз воспользуемся методом баланса. Проинтегрируем уравнение (18.38) по отрезку  $[\xi - h/2, \xi + h/2]$

$$-\int_{\xi-h/2}^{\xi+h/2} (pu')' dx = \int_{\xi-h/2}^{\xi+h/2} [f(x) - q(x)u(x)] dx.$$

Теперь вычисление левого интеграла следует проводить более осторожно: вычислять его как сумму двух интегралов по  $[\xi - h/2, \xi]$  и по  $[\xi, \xi + h/2]$ . Будем иметь

$$\begin{aligned} & -p\left(\xi + \frac{h}{2}\right)u'\left(\xi + \frac{h}{2}\right) + p(\xi + 0)u'(\xi + 0) - p(\xi - 0)u'(\xi - 0) + \\ & + p\left(\xi - \frac{h}{2}\right)u'\left(\xi - \frac{h}{2}\right) = \int_{\xi-h/2}^{\xi+h/2} [f(x) - q(x)u(x)] dx. \end{aligned} \quad (18.66)$$

В силу (18.65) второе и третье слагаемые левой части взаимно уничтожаются, и мы получаем в точности то же самое соотношение (18.40), что и в гладком случае. Мы несколько перестраховались, но это лучше, чем совершить ошибку. Дальнейшие наши действия напоминают действия в п. 18.4. Есть, правда, и отличия: заменять интеграл в правой части (18.66) квадратурной формулой прямоугольников, вообще говоря, нельзя, ибо в точке  $\xi$  функции  $f(x)$  и  $q(x)$  могут иметь разрывы. Эту трудность легко преодолеть, если в точке  $x = \xi$  положить

$$f(x) = [f(\xi + 0) + f(\xi - 0)]/2, \quad q(x) = [q(\xi + 0) + q(\xi - 0)]/2$$

и формально воспользоваться формулой прямоугольников. Разумеется, аппроксимация получится хуже из-за меньшей гладкости подынтегральной функции.

Итак, аппроксимация условий сопряжения принимает вид

$$-p_{j+1}^h u_{x,j}^h + p_j^h u_{\bar{x},j}^h = h (q_j^h u_j^h - f_j^h), \quad (18.67)$$

где, например,

$$q_j^h = \frac{q(\xi + 0) + q(\xi - 0)}{2}, \quad f_j^h = \frac{f(\xi + 0) + f(\xi - 0)}{2}. \quad (18.68)$$

Поделив это соотношение на  $h$ , получим формальную аппроксимацию уравнения (18.38) в точке разрыва коэффициентов

$$-(p^h u_{\bar{x}}^h)_{x,j} + q_j^h u_j^h = f_j^h. \quad (18.69)$$

Исследуем погрешность аппроксимации этого уравнения. Для простоты будем предполагать, что  $p(x)$  есть кусочно-постоянная функция

$$p(x) = \begin{cases} p_1, & 0 < x < \xi, \\ p_2, & \xi < x < 1. \end{cases} \quad (18.70)$$

Принимая во внимание (18.68), (18.70), будем иметь

$$\begin{aligned} \Psi_j &= \frac{f(\xi+0) + f(\xi-0)}{2} + \frac{1}{h} [p_2 u_{x,j} - p_1 u_{\bar{x},j}] - \frac{q(\xi+0) + q(\xi-0)}{2} u(\xi) = \\ &= \frac{f(\xi+0) - q(\xi+0)u(\xi) + f(\xi-0) - q(\xi-0)u(\xi)}{2} + \\ &+ \frac{1}{h} \left\{ p_2 \left[ u'(\xi+0) + \frac{h}{2} u''(\xi+0) + \frac{h^2}{6} u'''(\xi+0) \right] - \right. \\ &\left. - p_1 \left[ u'(\xi-0) - \frac{h}{2} u''(\xi-0) + \frac{h^2}{6} u'''(\xi-0) \right] + O(h^3) \right\}. \end{aligned}$$

Члены с  $O(h^{-1})$  взаимно уничтожаются в силу условия сопряжения (18.65), а члены  $O(1)$  — в силу условия сопряжения (18.64) и уравнения (18.38) слева и справа от точки разрыва  $\xi$ . Тем самым,

$$\Psi_j = \frac{h}{6} [pu''']_{x=\xi} + O(h^2).$$

Так как, вообще говоря,  $[pu''']_{x=\xi} \neq 0$ , то

$$\Psi_j = O(h). \quad (18.71)$$

**Замечание 18.4.** Вовсе не было большой необходимости рассматривать уравнение (18.69) как аппроксимацию (18.38) в точке разрыва (с погрешностью  $O(h)$ ). Достаточно было бы рассматривать эквивалентное соотношение (18.67) как аппроксимацию условия сопряжения (18.65) (с погрешностью  $O(h^2)$ ). Правда, при этом нам пришлось бы строить соответствующую априорную оценку решения новой сеточной задачи, включающей в свою формулировку условие (18.67). В трактовке (18.69) как аппроксимации уравнения (18.38) новую теорию строить не надо.

Исследуем влияние увеличения погрешности аппроксимации уравнения в одном узле до  $O(h)$  (см. (18.71)). Представим погрешность аппроксимации  $\Psi_i$ , например, задачи (18.43), (18.9) в виде

$$\Psi_i = \overset{\circ}{\Psi}_i + \tilde{\Psi}_i,$$

где

$$\overset{\circ}{\Psi}_i = \begin{cases} \Psi_i, & i \neq j, \\ 0, & i = j, \end{cases} \quad \tilde{\Psi}_i = \begin{cases} 0, & i \neq j, \\ \Psi_j, & i = j. \end{cases}$$

В силу линейности задачи погрешность решения  $z_i$  примет вид

$$z_i = \overset{\circ}{z}_i + \tilde{z}_i,$$

где

$$L_2^h \overset{\circ}{z}_i = \overset{\circ}{\Psi}_i, \quad L_2^h \tilde{z}_i = \tilde{\Psi}_i.$$

В силу теоремы 18.8 справедливо равенство  $\overset{\circ}{z}_i = O(h^2)$ , ибо  $\overset{\circ}{\Psi}_i = O(h^2)$ . С  $\tilde{z}_i$  нужно разбираться отдельно. Для этого воспользуемся принципом сравнения (теорема 18.4). Снова для простоты предположим, что коэффициент  $p(x)$  кусочно постоянный (18.70). Возьмем в качестве барьера кусочно линейную, непрерывную функцию

$$U_i = \begin{cases} Ax_i, & x_i \leq x_j = \xi, \\ A\xi, & x_i > \xi \end{cases} \quad (18.72)$$

Поскольку

$$L_2^h U_i = \begin{cases} -p_1 U_{\bar{x},i} + q_i U_i, & i < j, \\ -\frac{1}{h} (p_2 U_{x,j} - p_1 U_{\bar{x},j}) + q_j^h U_j, & i = j, \\ -p_2 U_{\bar{x},i} + q_i U_i, & i > j, \end{cases}$$

то

$$L_2^h U_i = \begin{cases} q_i U_i, & i \neq j, \\ \frac{Ap_1}{h} + q_j^h A\xi, & i = j. \end{cases}$$

Если  $|\Psi_j| \leq ch$ , то, положив  $A = ch^2/p_1$ , находим, что

$$L_2^h U_i =: F_i \geq |\tilde{\Psi}_i|$$

и, следовательно,

$$|\tilde{z}_i| = O(h^2).$$

Тем самым, показано, что, несмотря на то, что погрешность аппроксимации в одной точке (в нескольких точках, число которых не зависит от  $N$ ) есть  $O(h)$ , погрешность решения  $z_i = \overset{\circ}{z}_i + \tilde{z}_i$  остается величиной  $O(h^2)$  (как в гладком случае).

**Замечание 18.5.** Такая же ситуация с погрешностью аппроксимации возникает даже в том случае, когда коэффициенты уравнения непрерывны, однако сетка является кусочно-равномерной.

**Замечание 18.6.** Если бы коэффициент  $p(x)$  в уравнении (18.38) оставался переменным (а не кусочно-постоянным), то в качестве барьера нужно было бы брать несколько более сложную функцию.

**Упражнение 18.9.** Построить барьер, когда  $p_i^h$  не есть кусочно-постоянная функция.

## 18.8 Неравномерная сетка

При рассмотрении уравнения с разрывными коэффициентами мы столкнулись с тем обстоятельством, что для приемлемой аппроксимации дифференциального уравнения желательно использовать сетку, которая не является равномерной. Для этих целей достаточно использовать кусочно-равномерную сетку, но мы сейчас исследуем более общий случай, когда различными могут быть все шаги. Пусть

$$\widehat{\omega} = \{x_i \mid x_0 = 0 < x_1 < x_2 < \dots < x_{N-1} < x_N = 1\} \quad (18.73)$$

— произвольная *неравномерная сетка* на  $[0, 1]$ . Будем обозначать

$$h_i = x_i - x_{i-1}, \quad \bar{h}_i = \frac{h_i + h_{i+1}}{2}.$$

На сетке (18.73) для уравнения (18.38) методом баланса получим следующую аппроксимацию

$$\begin{aligned} & -\frac{1}{\bar{h}_i} \left[ p \left( x_i + \frac{h_{i+1}}{2} \right) \frac{u_{i+1}^h - u_i^h}{h_{i+1}} - p \left( x_i - \frac{h_{i+1}}{2} \right) \frac{u_i^h - u_{i-1}^h}{h_i} \right] + \\ & + q(x_i) u_i^h = f(x_i), \quad i = 1, \dots, N-1. \end{aligned} \quad (18.74)$$

Введем новое обозначение

$$(w_{i+1} - w_i) / \bar{h}_i =: w_{\hat{x}, i}.$$

С учетом этого обозначения уравнения (18.74) переписываются в виде

$$-(p^h u_{\bar{x}}^h)_{\hat{x}_i} + q_i^h u_i^h = f_i^h.$$

Если сетка (18.73) является произвольной, аппроксимация (18.74) имеет погрешность только  $O(h)$ , где  $h = \max h_i$ . Покажем это.

Снова для простоты будем предполагать, что  $p(x) \equiv 1$ . Тогда погрешность аппроксимации примет вид

$$\begin{aligned} \Psi_i &= f_i - q_i u(x_i) + \frac{1}{\bar{h}_i} \left[ \frac{u(x_{i+1}) - u(x_i)}{h_{i+1}} - \frac{u(x_i) - u(x_{i-1}))}{h_i} \right] = \\ &= f_i - q_i u_i + \frac{2}{h_i + h_{i+1}} \left[ u_i' + \frac{h_{i+1}}{2} u_i'' + \frac{h_{i+1}^2}{6} u_i''' + \frac{h_{i+1}^3}{24} u^{\text{IV}}(\xi_i) - \right. \\ &\quad \left. - \left( u_i' - \frac{h_i}{2} u_i'' + \frac{h_i^2}{6} u_i''' - \frac{h_i^3}{24} u^{\text{IV}}(\eta_i) \right) \right] = \\ &= (f - qu + u'')_i + \frac{h_{i+1}^2 - h_i^2}{3(h_{i+1} + h_i)} u_i''' + \frac{h_{i+1}^3 u^{\text{IV}}(\xi_i) + h_i^3 u^{\text{IV}}(\eta_i)}{12(h_{i+1} + h_i)} = \\ &= \frac{h_{i+1} - h_i}{3} u_i''' + \frac{h_i^2 + h_{i+1}^2 - h_i h_{i+1}}{12} u^{\text{IV}}(\zeta_i). \end{aligned} \quad (18.75)$$

Если

$$h_{i+1} - h_i = O(h_i^2),$$

то сетка называется квазиравномерной, и в этом случае

$$\Psi_i = O(h^2).$$

Если же

$$h_{i+1} - h_i \neq O(h_i^2),$$

то

$$\Psi_i = O(h).$$

**Замечание 18.7.** Квазиравномерные сетки образуются, например, в результате следующих построений. Пусть  $\varphi(t) \in C^2[0, 1]$ , причем  $\varphi' > 0$  на  $[0, 1]$  и  $\varphi(0) = 0$ ,  $\varphi(1) = 1$ . Тогда  $x = \varphi(t)$  есть взаимно однозначное отображение отрезка  $[0, 1]$  на себя. Введем на  $[0, 1]$  равномерную с шагом  $\tau = 1/N$  сетку узлов  $t = i/N$ . Этой сетке будет соответствовать неравномерная сетка с узлами  $x_i = \varphi(t_i)$ . Сетка  $x_i$  будет квазиравномерной, ибо

$$\begin{aligned} h_{i+1} - h_i &= (x_{i+1} - x_i) - (x_i - x_{i-1}) = \varphi_{i+1} - 2\varphi_i + \varphi_{i-1} = \\ &= \int_{t_{i-1}}^{t_{i+1}} (\tau - |t - t_i|) \varphi''(t) dt = N^{-2} \varphi''(\tilde{t}_i), \quad \tilde{t}_i \in (t_{i-1}, t_{i+1}). \end{aligned}$$



Какова же точность разностной схемы (18.74) на неравномерной сетке? Если сетка квазиравномерная, то так же, как в теореме 18.2 или 18.8 с использованием соответствующей априорной оценки (см., например, теорему 18.1 или теорему 18.7) доказывается, что точность разностной схемы (18.74) (при соответствующей гладкости решения) будет  $O(h^2)$ . Если же сетка (18.73) квазиравномерной не является, то дело с ответом на поставленный вопрос обстоит несколько сложнее. Разумеется, сходимость со скоростью  $O(h)$  здесь имеет место, и доказывается это уже известным нам способом, но это не вся правда. Более тщательный анализ показывает, что и в этом случае сходимость будет  $O(h^2)$ . Чтобы убедиться в этом, нужно несколько преобразовать погрешность аппроксимации (18.75) и вывести новую априорную оценку. Легко видеть, что прием, использованный при оценке скорости сходимости в случае разрывных коэффициентов, здесь не проходит, ибо погрешность аппроксимации имеет только первый порядок малости в слишком большом числе узлов.

Преобразуем первое слагаемое погрешности аппроксимации (18.75). Имеем

$$\frac{h_{i+1} - h_i}{3} u_i''' = \frac{h_{i+1}^2 - h_i^2}{6\bar{h}_i} u_i''' = \frac{1}{6} h_{\hat{x},i}^2 u_i''''.$$

С другой стороны,

$$\begin{aligned} (h^2 u''')_{\hat{x},i} &= \frac{h_{i+1}^2 u_{i+1}'''' - h_i^2 u_i'''' + h_{i+1}^2 u_i'''' - h_{i+1}^2 u_i''''}{\bar{h}_i} = \\ &= h_{\hat{x},i}^2 u_i'''' + \frac{h_{i+1}^2}{\bar{h}_i} (u_{i+1}'''' - u_i''') = h_{\hat{x},i}^2 u_i'''' + \frac{h_{i+1}^3}{\bar{h}_i} \tilde{u}_i^{\text{IV}}. \end{aligned}$$

Поэтому

$$\frac{h_{i+1} - h_i}{3} u_i''' = \frac{1}{6} (h^2 u''')_{\hat{x},i} - \frac{h_{i+1}^3}{\bar{h}_i} \tilde{u}_i^{\text{IV}}.$$

Используя это представление в (18.75), будем иметь

$$\Psi_i = \overset{\circ}{\psi}_{\hat{x},i} + \tilde{\psi}_i, \quad (18.76)$$

где

$$\overset{\circ}{\psi}_i = \frac{h^2 u_i''''}{6} = O(h^2), \quad (18.77)$$

$$\tilde{\psi}_i = \frac{h_i^2 + h_{i+1}^2 - h_i h_{i+1}}{12} u_i^{\text{IV}}(\zeta_i) - \frac{h_{i+1}^3}{\bar{h}_i} \tilde{u}_i^{\text{IV}} = O(h^2). \quad (18.78)$$

Итак, на произвольной неравномерной сетке (18.73) погрешность аппроксимации (18.75) представляет собой сумму двух сеточных функций, одна из которых  $O(h^2)$ , а другая является разностным отношением функции  $O(h^2)$ .

## 18.9 Априорные оценки и оценка точности

Теперь по поводу априорной оценки. Чтобы не погрязнуть в технических вопросах, нужную априорную оценку мы получим не для решения сеточной задачи, а для решения задачи дифференциальной. Для сеточной задачи все делается точно так же, но с бóльшими техническими трудностями.

Рассмотрим простейшую задачу

$$-u''(x) = f(x), \quad 0 < x < 1, \quad u(0) = u(1) = 0. \quad (18.79)$$

Для этой задачи имеет место принцип сравнения (см. теорему 18.4). Используя барьерную функцию

$$U(x) = x(1-x) \frac{1}{2} \max_{0 < x < 1} |f(x)|,$$

найдем, что

$$|u(x)| \leq U(x) \leq \max_{0 < x < 1} |f(x)|,$$

или

$$\|u\|_{L_\infty} \leq \frac{1}{8} \|f\|_{L_\infty}. \quad (18.80)$$

Эта априорная оценка аналогична оценке из теоремы 18.5.

Построим функцию Грина задачи (18.79). Для этого найдем сначала частное решение уравнения (18.79). Интегрируя уравнение (18.79) дважды, будем иметь

$$\bar{u}(x) = - \int_0^x d\xi \int_0^\xi f(\eta) d\eta.$$

Общее решение этого уравнения очевидно есть

$$u(x) = c_1 x + c_2(1-x) - \int_0^x d\xi \int_0^\xi f(\eta) d\eta.$$

Выбирая постоянные  $c_1$  и  $c_2$  так, чтобы выполнялись граничные условия (18.79), найдем решение задачи (18.79)

$$u(x) = x \int_0^1 d\xi \int_0^\xi f(\eta) d\eta - \int_0^x d\xi \int_0^\xi f(\eta) d\eta.$$

Преобразуем двойные интегралы к одинарным. Интегрируя по частям, будем иметь

$$u(x) = x \int_0^1 (1 - \xi) f(\xi) d\xi - \int_0^x (x - \xi) f(\xi) d\xi.$$

Разбивая первый из интегралов на сумму интегралов по  $(0, x)$  и  $(x, 1)$ , окончательно найдем, что

$$u(x) = \int_0^1 [x(1 - \xi) - (x - \xi)] f(\xi) d\xi + \int_x^1 x(1 - \xi) f(\xi) d\xi.$$

Пусть

$$G(x, \xi) = \begin{cases} x(1 - \xi), & x < \xi, \\ \xi(1 - x), & x > \xi. \end{cases} \quad (18.81)$$

Тогда

$$u(x) = \int_0^1 G(x, \xi) f(\xi) d\xi, \quad (18.82)$$

т.е. (18.81) есть *функция Грина* задачи (18.79).

Воспользуемся представлением (18.82) для получения априорных оценок решения задачи (18.79). Из (18.82) вытекает, что

$$\|u\|_{L_\infty} \leq \max_{0 \leq x, \xi \leq 1} |G(x, \xi)| \|f\|_{L_\infty}.$$

Поскольку

$$0 \leq G(x, \xi) \leq G(x, x) = x(1 - x) \leq 1/4,$$

то

$$\|u\|_{L_\infty} \leq \frac{1}{4} \|f\|_{L_\infty}.$$

Эта оценка (по коэффициенту) несколько слабее оценки (18.80), однако и оценка (18.80) может быть получена из (18.82). Именно

$$\|u\|_{L_\infty} \leq \|f\|_{L_\infty} \max_{0 < x < 1} \int_0^1 G(x, \xi) d\xi.$$

Но

$$\int_0^1 G(x, \xi) d\xi = \frac{x(1 - x)}{2} \leq \frac{1}{8},$$

Откуда и вытекает оценка (18.80).

Но не ради этого мы обратились к функции Грина и представлению решения в виде (18.82). Из (18.82) вытекает, например, такая новая оценка

$$\|u\|_{L_\infty} \leq \max_{0 < x, \xi < 1} G(x, \xi) \|f\|_{L_1} \leq \frac{1}{4} \|f\|_{L_1}. \quad (18.83)$$

В этой оценке для решения и правой части уравнения задействованы разные нормы, причем для решения использована более сильная норма, чем для правой части. Сеточный аналог этой оценки позволяет выявить правильную оценку скорости сходимости разностной схемы в рассмотренном ранее случае разрывных коэффициентов без построения специального барьера, как мы это делали в п. 18.7.

Но это еще не все. Преобразуем (18.82) интегрированием по частям

$$u(x) = - \int_0^1 G(x, \xi) d \left[ \int_\xi^1 f(\eta) d\eta - C \right] = \int_0^1 \frac{\partial G}{\partial \xi}(x, \xi) \left[ \int_\xi^1 f(\eta) d\eta - C \right] d\xi.$$

Отсюда находим, что

$$\|u\|_{L_\infty} \leq \min_C \max_{0 < \xi < 1} \left| \int_\xi^1 f(\eta) d\eta - C \right| \max_{0 < x < 1} \int_0^1 \left| \frac{\partial G}{\partial \xi}(x, \xi) \right| d\xi$$

Поскольку

$$\frac{\partial G}{\partial \xi} = \begin{cases} -x, & x < \xi \\ 1-x, & x > \xi, \end{cases}$$

то

$$\int_0^1 \left| \frac{\partial G}{\partial \xi}(x, \xi) \right| d\xi = 2x(1-x) \leq 1/2$$

и, следовательно,

$$\|u\|_{L_\infty} \leq \frac{1}{2} \min_C \max_{0 < \xi < 1} \left| \int_\xi^1 f(\eta) d\eta - C \right| =: \frac{1}{2} \|f\|_{W_\infty^{-1}}. \quad (18.84)$$

Ради этой оценки мы и ввели в рассмотрение функцию Грина.

Пусть

$$f(x) = \overset{\circ}{f}'(x) + \tilde{f}(x).$$

Тогда и решение задачи (18.79) можно представить в виде

$$u(x) = \overset{\circ}{u}(x) + \tilde{u}(x).$$

где  $\overset{\circ}{u}(x)$  удовлетворяет уравнению (18.79) с правой частью  $\overset{\circ}{f}'(x)$ , а  $\tilde{u}(x)$  — с правой частью  $\tilde{f}(x)$ . Для оценки  $\overset{\circ}{u}(x)$  воспользуемся соотношением

(18.84), положив  $C = \overset{\circ}{f}(1)$ , а для оценки  $\tilde{u}(x)$  — соотношением (18.80). В результате получим

$$\|u\|_{L_\infty} \leq \|\overset{\circ}{u}\|_{L_\infty} + \|\tilde{u}\|_{L_\infty} \leq \frac{1}{2} \|\overset{\circ}{f}\|_{L_\infty} + \frac{1}{8} \|\tilde{f}\|_{L_\infty}. \quad (18.85)$$

Если, например,

$$f(x) = k\pi \cos k\pi x + 1, \quad k \gg 1,$$

то, в силу оценки (18.80)

$$\|u\|_{L_\infty} = O(k),$$

а, полагая

$$\overset{\circ}{f}(x) = \sin k\pi x$$

и используя оценку (18.80), найдем, что

$$\|u\|_{L_\infty} = O(1).$$

**Замечание 18.8.** В общем случае уравнения (18.38) с переменными коэффициентами функция Грина в явном виде не выписывается, однако ограничена и имеет ограниченные производные как по  $x$ , так и по  $\xi$ . Так что оценка (18.84) имеет место и в общем случае.

Применим полученные результаты к анализу скорости сходимости разностных схем на неравномерной сетке. Пусть (18.74) есть аппроксимация уравнения (18.79), т.е.

$$-u_{\hat{x}\hat{x},i}^h = f_i, \quad i = 1, 2, \dots, N-1, \quad u_0^h = u_N^h = 0.$$

Легко проверить, что функция Грина этой задачи имеет вид

$$G^h(x_i, \xi_j) = \begin{cases} x_i(1 - \xi_j), & x_i \leq \xi_j, \\ \xi_j(1 - x_i), & x_i \geq \xi_j, \end{cases} \quad x_i, \xi_j \in \hat{\omega},$$

а для ее решения справедливо представление

$$u^h(x_i) = \sum_{j=1}^{N-1} G^h(x_i, \xi_j) f(\xi_j) \hat{h}_j, \quad (18.86)$$

Поскольку

$$-\left( \sum_{l=j}^{N-1} f_l \hat{h}_l \right)_{\hat{x},j} = f_j = f(\xi_j),$$

а при  $v_0 = 0$

$$\begin{aligned} \sum_{j=1}^{N-1} v_j w_{\hat{x},j} \bar{h}_j &= -v_1(w_2 - w_1) + v_2(w_3 - w_2) + \cdots + v_{N-1}(w_N - w_{N-1}) = \\ &= -w_1 v_1 - w_2(v_2 - v_1) - \cdots - w_N(v_N - v_{N-1}) = \sum_{j=1}^N v_{\bar{\xi},j} w_j \bar{h}_j, \end{aligned}$$

то представлению (18.86) можно придать вид

$$u^h(x_i) = \sum_{i=1}^N G_{\bar{\xi}}^h(x_i, \xi_j) \left( \sum_{l=j}^{N-1} f_l \bar{h}_l - C \right) h_j,$$

откуда и вытекает аналогичная (18.84) оценка

$$\|u^h\|_{L_\infty^h} \leq \frac{1}{2} \min_C \max_j \left| \sum_{l=j}^{N-1} f_l \bar{h}_l - C \right|.$$

Если

$$f_j = \overset{\circ}{f}_{\xi_j} + \tilde{f}_j, \quad (18.87)$$

то, как и в континуальном случае, находим, что

$$\|u^h\|_{L_\infty^h} \leq \frac{1}{2} \|\overset{\circ}{f}\|_{L_\infty^h} + \frac{1}{8} \|\tilde{f}\|_{L_\infty^h}. \quad (18.88)$$

Представление погрешности аппроксимации на неравномерной сетке (18.76) полностью совпадает с представлением правой части (18.87). Поэтому для погрешности решения  $z_i = u_i^h - u(x_i)$  справедлива оценка

$$\|z^h\|_{L_\infty^h} \leq \frac{1}{2} \|\overset{\circ}{\psi}\|_{L_\infty^h} + \frac{1}{8} \|\tilde{\psi}\|_{L_\infty^h},$$

вытекающая из (18.88). Если теперь принять во внимание соотношения (18.77), (18.78), то получим следующую оценку погрешности решения

$$\|z^h\|_{L_\infty^h} = O(h^2).$$

Итак, погрешность решения разностной схемы на неравномерной сетке, удовлетворяющей условию  $h_i \leq h$ , есть  $O(h^2)$ .

## 18.10 Аппроксимация производной

Укажем еще на одну полезную оценку, которая следует из представления (18.82). Дифференцируя  $u(x)$  из (18.82) и оценивая производную функции Грина, найдем, что

$$\|u'\|_{L_\infty} \leq \frac{1}{2} \|f\|_{L_\infty}.$$

Аналогичная оценка для сеточной задачи имеет вид

$$\|u_{\bar{x}}^h\|_{L_\infty^h} \leq \frac{1}{2} \|f^h\|_{L_\infty^h}. \quad (18.89)$$

Воспользуемся этой оценкой для анализа точности приближения первой производной решения дифференциальной задачи разностным отношением приближенного решения, вычисленного на равномерной сетке. В силу (18.89) для погрешности решения справедлива оценка

$$\|z_{\bar{x}}\|_{L_\infty^h} \leq \frac{1}{2} \|\Psi\|_{L_\infty^h}.$$

а, поскольку сетка предполагается равномерной, то  $\Psi = O(h^2)$  и, следовательно,

$$z_{\bar{x},i} = O(h^2). \quad (18.90)$$

Будем аппроксимировать производную решения задачи (18.79) при помощи двух соседних значений приближенного решения  $u_i^h$  и  $u_{i-1}^h$ . Именно, в качестве аппроксимирующего значения возьмем  $u_{\bar{x},i}^h$ . По определению,  $u_i^h = u(x_i) + z_i$  и, следовательно

$$u_{\bar{x},i}^h = u_{\bar{x},i} + z_{\bar{x},i}.$$

Поскольку

$$u_{\bar{x},i} = u'(x_{i-1/2}) + O(h^2),$$

то, с учетом (18.90),

$$u_{\bar{x},i}^h - u'(x_{i-1/2}) = O(h^2),$$

т.е. точности в определении и решения и производной имеют один и тот же порядок малости, что в корне отличается от ситуации, описанной в § 15. В данном случае о погрешности приближенного решения у нас больше информации, чем было раньше.

**Упражнение 18.10.** Выяснить, какую точность при приближении производной дает в рассматриваемом случае формула  $(u_{i+1}^h - u_{i-1}^h)/2h$ .

**Упражнение 18.11.** Что можно сказать о точности  $u_{\bar{x},i}^h$ ?

## 18.11 Уравнение конвекции-диффузии

Добавим к уравнению (18.38) еще один член — первую производную искомого решения, умноженную на некоторый коэффициент

$$-(pu')' - r(x)u' + q(x)u = f. \quad (18.91)$$

Как аппроксимировать первый и последний члены левой части (18.91), мы знаем. Осталось построить аппроксимацию второго члена. С точки зрения наилучшего порядка аппроксимации следует положить

$$u'_i \approx \frac{u_{i+1} - u_{i-1}}{2h} =: u_{\bar{x},i}. \quad (18.92)$$

Тогда аппроксимация уравнения (18.91) примет вид

$$-(p_{i-1/2}u_{\bar{x}}^h)_{x,i} - r_i u_{\bar{x},i}^h + q_i u_i^h = f_i, \quad i = 1, \dots, N-1. \quad (18.93)$$

**Теорема 18.9.** *Если  $u \in C^4[0, 1]$ , то погрешность аппроксимации разностной схемы (18.93)  $\Psi = O(h^2)$ .*

Дополним уравнение (18.91) граничными условиями. Пусть, например,

$$u(0) = g_0, \quad u(1) = g_1. \quad (18.94)$$

Тогда разностные уравнения (18.93) нужно дополнить граничными условиями

$$u_0^h = g_0, \quad u_N^h = g_1. \quad (18.95)$$

Имеет место

**Теорема 18.10.** *Если коэффициенты уравнения (18.91) удовлетворяют условиям (18.23), (18.47), а сетка такова, что*

$$\max_i \frac{|r(x_i)|h}{2c_0} \leq 1, \quad (18.96)$$

*то задача (18.93), (18.95) имеет единственное решение, и для него справедлива априорная оценка*

$$\max_i |u_i^h| \leq |g_0| + |g_1| + \max_i \frac{|f_i|}{c_1}. \quad (18.97)$$



**Доказательство.** Представим

$$u_x^h = \frac{u_{i+1}^h - u_{i-1}^h}{2h} = \frac{u_{i+1}^h - u_i^h + u_i^h - u_{i-1}^h}{2h} = \frac{1}{2}u_x + \frac{1}{2}u_{\bar{x}}.$$

Подставим это представление в (18.93)

$$-\frac{1}{h} (p_{i+1/2}u_{x,i}^h - p_{i-1/2}u_{\bar{x},i}^h) - \frac{r_i}{2} (u_{x,i}^h + u_{\bar{x},i}^h) + q_i u_i^h = f_i.$$

Отсюда

$$-\left(\frac{p_{i+1/2}}{h} + \frac{r_i}{2}\right) \frac{u_{i+1}^h - u_i^h}{h} + \left(\frac{p_{i-1/2}}{h} - \frac{r_i}{2}\right) \frac{u_i^h - u_{i-1}^h}{h} + q_i u_i^h = f_i.$$

При выполнении условий (18.96) выражения в скобках неотрицательные. Этому замечания достаточно для того, чтобы завершить доказательство этой теоремы, используя те же самые рассуждения, что и при доказательстве теорем 18.7 и 18.1.

**Упражнение 18.12.** Завершить доказательство теоремы 18.10.

**Упражнение 18.13.** Сформулировать и доказать теорему о сходимости разностной задачи (18.93), (18.95).

## Сингулярно возмущенные уравнения

### 19.1 Осцилляции решения и сингулярно возмущенные уравнения

При исследовании разрешимости и сходимости разностной схемы (18.93) для уравнения конвекции-диффузии (18.91) мы ввели ограничение (18.96) на шаг сетки. Это ограничение в ряде случаев оказывается излишне обременительным, и тогда от аппроксимации (18.92) первой производной приходится отказываться. Обсудим этот вопрос на примере простейшего однородного уравнения с постоянными коэффициентами

$$\frac{d^2 u}{dx^2} + r \frac{du}{dx} = 0, \quad r = \text{const.} \quad (19.1)$$

Наряду с аппроксимацией (18.92) производной  $u'$  рассмотрим также ее аппроксимации односторонними разностными отношениями  $u_x$  и  $u_{\bar{x}}$ . Разумеется, порядок погрешности аппроксимации в этих случаях будет ниже. Будем рассматривать одновременно все три из указанных аппроксимаций  $u'$ . Для этого в разностное уравнение введем параметр  $\sigma$

$$u_{\bar{x}\bar{x}}^h + r [\sigma u_x^h + (1 - \sigma) u_{\bar{x}}^h] = 0. \quad (19.2)$$

При  $\sigma = 1/2$  имеем центральное разностное отношение  $u_{\circ}^h = (u_x + u_{\bar{x}})/2$ , при  $\sigma = 1$  — правое разностное отношение  $u_x$ , а при  $\sigma = 0$  — левое разностное отношение  $u_{\bar{x}}$ . Перепишем (19.2) в поточечном виде

$$\left( \frac{1}{h^2} + \frac{\sigma r}{h} \right) u_{i+1}^h - \left( \frac{2}{h^2} + \frac{(2\sigma - 1)r}{h} \right) u_i^h + \left( \frac{1}{h^2} + \frac{(\sigma - 1)r}{h} \right) u_{i-1}^h = 0.$$

Это есть разностное уравнение с постоянными коэффициентами. Его характеристическое уравнение имеет вид

$$\left(\frac{1}{h} + \sigma r\right) q^2 - \left(\frac{2}{h} + (2\sigma - 1)r\right) q + \left(\frac{1}{h} + (\sigma - 1)r\right) = 0. \quad (19.3)$$

Поскольку сумма коэффициентов уравнения (19.3) равна нулю, то среди его корней есть корень  $q_1 = 1$ . Второй корень

$$q_2 = q = \frac{1 + (\sigma - 1)rh}{1 + \sigma rh}. \quad (19.4)$$

Проведем качественное сравнение решений дифференциального уравнения (19.1) и разностного уравнения (19.2). Для этого предположим, что

$$r > 0 \quad (19.5)$$

и поставим задачу для (19.1) на положительной полуоси  $Ox$

$$u(0) = 1, \quad u(\infty) = 0. \quad (19.6)$$

Очевидно, что решение задачи (19.1), (19.5), (19.6) имеет вид

$$u(x) = e^{-rx}. \quad (19.7)$$

Функция (19.7) положительна и монотонно убывает при  $x \rightarrow \infty$ .

Будем искать решение разностного уравнения (19.2), удовлетворяющее условиям

$$u_0^h = 1, \quad \lim_{i \rightarrow \infty} u_i^h = 0. \quad (19.8)$$

Общее решение уравнения (19.2) в силу вышесказанного есть

$$u_i^h = c_1 + c_2 q^i. \quad (19.9)$$

Для того, чтобы это решение на бесконечности было хотя бы ограниченным, нужно потребовать, чтобы (см. (19.4))

$$|q| \leq 1, \quad \text{т.е.} \quad -1 \leq \frac{1 + (\sigma - 1)rh}{1 + \sigma rh} \leq 1. \quad (19.10)$$

Пусть  $\sigma \geq 0$ . Тогда знаменатель в (19.10) положителен, правая часть неравенства имеет место всегда, и поэтому остается только ограничение

$$-1 - \sigma rh \leq 1 + (\sigma - 1)rh,$$

т.е.

$$2 + (2\sigma - 1)rh \geq 0.$$

Если  $\sigma = 1$  или  $\sigma = 1/2$ , то это условие выполнено со знаком строгого неравенства, и решение задачи (19.2), (19.8) при этих значениях  $\sigma$  имеет вид

$$u_i^h = q^i, \quad i \in \mathbb{N}. \quad (19.11)$$

Наложим более сильное условие на сеточное решение. Потребуем, чтобы оно было монотонным как и решение (19.7) дифференциальной задачи. Решение (19.11) будет монотонным тогда и только тогда, когда  $q \geq 0$ , т.е. если

$$1 + (\sigma - 1)rh \geq 0. \quad (19.12)$$

При  $\sigma = 1$  это условие выполнено, а при  $\sigma = 1/2$  требуется, чтобы

$$rh \leq 2 \quad (19.13)$$

(сравнить с (18.96)).

Итак, если  $\sigma = 1$ , погрешность аппроксимации уравнения (19.2) есть  $O(h)$ , но решение (19.11) задачи (19.2), (19.8) монотонно при любых  $h$ . Если  $\sigma = 1/2$ , то погрешность аппроксимации есть  $O(h^2)$ , но решение (19.11) монотонно только при выполнении условия (19.13). В противном случае решение (19.11) будет колебаться (см. рис. 1), меняя знак при переходе от одного узла к другому.

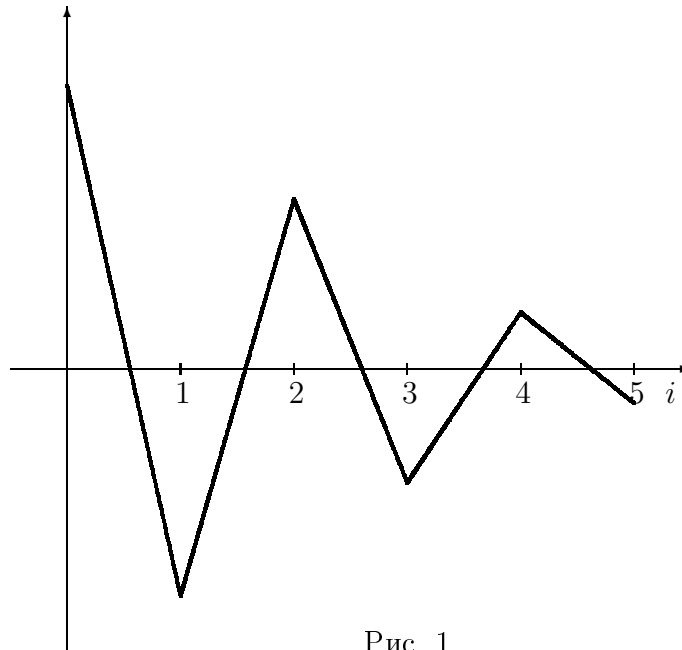


Рис. 1

Именно эти осцилляции решения разностной схемы (19.2) при  $\sigma = 1/2$  и не любят прикладники.

**Замечание 19.1.** Проведенный анализ показал принципиальное различие между схемами (19.2) при  $\sigma = 1$  и при  $\sigma = 0$ , хотя обе эти схемы имеют погрешность  $O(h)$  и в этом смысле близки. Причина различия состоит в знаке коэффициента  $r$ . Если бы он был отрицательным, то схемы с  $\sigma = 1$  и  $\sigma = 0$  поменялись бы ролями.

Казалось бы, ограничение (19.13) не является слишком обременительным, чтобы всегда требовать его выполнения. Для обычных задач это так. Но есть важный класс так называемых сингулярно возмущенных уравнений, когда ограничение (19.13) оказывается весьма обременительным. Простейшим примером является уравнение

$$\varepsilon u'' + u' = 0. \quad (19.14)$$

Здесь  $\varepsilon \in (0, 1]$  — малый параметр. При  $\varepsilon \rightarrow 0$  дифференциальное уравнение второго порядка (19.14) переходит в уравнение первого порядка, для которого одно из двух граничных условий, выделяющих единственное решение уравнения (19.14), становится лишним. Это и является причиной непростого поведения решения соответствующей задачи для уравнения (19.14) при малых  $\varepsilon$ . Если для уравнения (19.14) поставить граничные условия

$$u(0) = 0, \quad u(1) = 1, \quad (19.15)$$

то решением этой задачи будет функция

$$u(x) = \frac{1 - e^{-x/\varepsilon}}{1 - e^{-1/\varepsilon}} = 1 - \frac{e^{-x/\varepsilon} - e^{-1/\varepsilon}}{1 - e^{-1/\varepsilon}}, \quad (19.16)$$

являющаяся суммой гладкой, медленно меняющейся функции  $u_0(x) := 1$  и быстро меняющейся функции  $u_1(x) := (e^{-x/\varepsilon} - e^{-1/\varepsilon})/(1 - e^{-1/\varepsilon})$ .

Поскольку уравнения (19.1) и (19.14) переходят одно в другое при  $r = 1/\varepsilon$ , то условие (19.13) применительно к разностной схеме (19.2) для уравнения (19.14) примет вид

$$h \leq 2\varepsilon. \quad (19.17)$$

Но в (19.14) параметр  $\varepsilon$  может принимать значения  $10^{-2}$ ,  $10^{-4}$  или даже  $10^{-8}$ , и ограничение (19.17) становится слишком обременительным. В этой ситуации следует либо ограничиться схемой с  $\sigma = 1$ , которая не накладывает никаких ограничений на шаг сетки с точки зрения осцилирования решения, и довольствоваться погрешностью аппроксимации  $O(h)$ , либо пытаться строить другие схемы, которые имеют погрешность  $O(h^2)$  и не требуют ограничения типа (19.17).

## 19.2 Четырехточечная схема

Построим другую аппроксимацию уравнения (19.1). Будем аппроксимировать в (19.1) второе слагаемое при помощи соотношения

$$u'(x_i) \approx \frac{-u_{i+2} + 4u_{i+1} - 3u_i}{2h},$$

погрешность аппроксимации которого есть  $O(h^2)$ . Используя эту аппроксимацию, вместо (19.2) будем иметь

$$\frac{u_{i+1}^h - 2u_i^h + u_{i-1}^h}{h^2} + r \frac{-u_{i+2}^h + 4u_{i+1}^h - 3u_i^h}{2h} = 0. \quad (19.18)$$

Напишем характеристическое уравнение этого разностного уравнения с постоянными коэффициентами

$$\frac{q^2 - 2q + 1}{h^2} + r \frac{-q^3 + 4q^2 - 3q}{2h} = 0.$$

Обозначим  $rh/2 = \xi$  и перепишем характеристическое уравнение в виде

$$-\xi q^3 + (1 + 4\xi)q^2 - (2 + 3\xi)q + 1 = 0.$$

Сумма коэффициентов этого уравнения равна нулю, и следовательно,  $q = 1$  есть корень этого уравнения. После деления многочлена из левой части на  $(q - 1)$  получим уравнение

$$-\xi q^2 + (1 + 3\xi)q - 1 = 0,$$

корнями которого являются числа

$$q_{2,3} = \frac{1 + 3\xi \pm \sqrt{1 + 6\xi + 9\xi^2 - 4\xi}}{2\xi}.$$

Очевидно, что оба эти корня положительны при любых положительных  $\xi$ . Поскольку общее решение уравнения (19.18) имеет вид

$$u_i^h = c_1 + c_2 q_2^i + c_3 q_3^i,$$

то осцилляции этого решения будут отсутствовать на любой сетке, т.е. при любых  $h$ .

Аппроксимирующим экспоненту  $e^{-rh}$  будет корень

$$\begin{aligned} q_2 &= \frac{1}{2\xi} \left[ 1 + 3\xi - \sqrt{1 + 2\xi + 9\xi^2} \right] = \\ &= \frac{1}{2\xi} \left\{ 1 + 3\xi - \left[ 1 + \frac{2\xi + 9\xi^2}{2} - \frac{(2\xi + 9\xi^2)^2}{8} + \frac{8}{16}\xi^3 + O(\xi^4) \right] \right\} = \\ &= 1 - 2\xi + 2\xi^2 + O(\xi^3) = 1 - rh + \frac{r^2 h^2}{2} + O(h^3) = e^{-rh} + O(h^3) \end{aligned}$$

**Замечание 19.2.** Поскольку уравнение (19.18) нельзя написать для  $i = N - 1$ , то в этом узле должна быть написана другая аппроксимация уравнения (19.1), например, (19.2) при  $i = N - 1$  с любым  $\sigma$  (либо  $\sigma = 1/2$ , либо  $\sigma = 1$ ).

**Замечание 19.3.** Мы рассмотрели случай  $r > 0$ . Если  $r < 0$ , то, сделав в (19.1) замену независимой переменной  $1 - x = t$ , придем к уравнению

$$\frac{d^2u}{dt^2} - r \frac{du}{dt} = u'' + |r|u' = 0.$$

Отсюда следует, что при  $r < 0$  в исходных переменных нужно использовать аппроксимацию, зеркальную к той, которая используется при  $r > 0$ . Именно, вместо разности вперед  $(u_{i+1} - u_i)/h$  — разность назад  $(u_i - u_{i-1})/h$ , а вместо  $(-u_{i+2} + 4u_{i+1} - 3u_i)/2h$  — аппроксимацию

$$u_{\bar{x},i} + \frac{h}{2}u_{\bar{x}x,i} = \frac{u_{i-2} - 4u_{i-1} + 3u_i}{2h}.$$

### 19.3 Монотонная схема Самарского

Другой подход к построению монотонной разностной схемы формально второго порядка аппроксимации был предложен А.А. Самарским. Суть его рассуждений состоит в следующем. Пусть рассматривается уравнение (18.91), коэффициенты которого, для простоты, будем считать постоянными. Напишем для этого уравнения аппроксимацию с центральным разностным отношением

$$pu_{\bar{x}x,i}^h + ru_{x,i}^h = -f_i. \quad (19.19)$$

Поскольку

$$u_{\bar{x}}^h = \frac{1}{2}u_x^h + \frac{1}{2}u_{\bar{x}}^h = u_x^h - \frac{1}{2}(u_x^h - u_{\bar{x}}^h) = u_x^h - \frac{h}{2}u_{\bar{x}x}^h.$$

то уравнение (19.19) можно переписать в виде

$$p \left( 1 - \frac{rh}{2p} \right) u_{\bar{x}x,i}^h + ru_{x,i}^h = -f_i. \quad (19.20)$$

Замечая, что

$$\frac{1}{1 + \frac{rh}{2p}} = 1 - \frac{rh}{2p} + O\left(\left(\frac{rh}{2p}\right)^2\right),$$

напишем вместо (19.20) новое уравнение

$$\frac{p}{1 + \frac{rh}{2p}} u_{\bar{x}x,i}^h + r u_{x,i}^h = -f_i. \quad (19.21)$$

Разностная схема (19.21) называется монотонной схемой Самарского.

Характеристическое уравнение для (19.21) имеет вид

$$\left( p + rh + \frac{r^2 h^2}{2p} \right) q^2 - \left( 2p + rh + \frac{r^2 h^2}{2p} \right) q + p = 0.$$

Легко проверить, что один из корней этого уравнения равен 1, а второй

$$q = \left( 1 + \frac{rh}{p} + \frac{r^2 h^2}{2p^2} \right)^{-1}.$$

т.е. оба корня положительны, и решение однородного уравнения (19.21) колебаний иметь не может. В этом смысле решения (19.21) ведут себя так же, как решения (19.2) при  $\sigma = 1$ . Напомним, что схема (19.2) при  $\sigma = 1$  имеет первый порядок аппроксимации, а схема (19.21) позиционируется как схема второго порядка.

Исследуем более подробно погрешность аппроксимации схемы (19.21)

$$\begin{aligned} \psi &= f_i - \frac{p}{1 + \frac{rh}{2p}} u_{\bar{x}x,i} - r u_{x,i} = \\ &= f_i - \frac{p}{1 + \frac{rh}{2p}} \left( u_i'' + \frac{h^2}{12} \tilde{u}_i^{IV} \right) - r \left( u_i' + \frac{h}{2} u_i'' + \frac{h^2}{6} \tilde{u}_i''' \right). \end{aligned}$$

Будем предполагать, что  $u(x)$  — гладкое решение уравнения (18.91), т.е. все используемые в настоящих рассуждениях производные этого решения ограничены. Тогда

$$\begin{aligned} \psi_i &= f_i - \frac{p}{1 + \frac{rh}{2p}} u_i'' - r \left( u_i' - \frac{h}{2} u_i'' \right) + O(h^2) = \\ &= \frac{(1 + rh/(2p)) f_i - p u_i' - (1 + rh/(2p)) r u_i' - (1 + rh/(2p)) \frac{rh}{2p} p u_i''}{1 + rh/(2p)} + O(h^2) = \\ &= f_i - r u_i' - u_i'' - \frac{r^2 h^2}{2(2p + rh)} u_i'' + O(h^2) = -\frac{r^2 h^2}{2(2p + rh)} u_i'' + O(h^2). \end{aligned}$$

Отсюда следует, что, если  $p$  не мал, то погрешность монотонной схемы Самарского есть  $O(h^2)$ , как и у схемы с центральной разностью. Последняя при конечных  $p$  также является монотонной. Если  $p$  становится малым, когда у схемы с центральной разностью возникают проблемы, порядок погрешности аппроксимации монотонной схемы Самарского становится первым.



## 19.4 О равномерной по $\varepsilon$ сходимости

Исследования показывают, что какой бы метод аппроксимации уравнения (19.1) из числа рассмотренных выше мы ни избрали, в любом случае при фиксированном  $N$  и  $\varepsilon \rightarrow 0$  найдутся такие узлы равномерной сетки, в которых погрешность решения будет  $O(1)$ . Чтобы отметить этот факт, говорят, что разностная схема не обладает свойством *равномерной по малому параметру сходимости*.

Один из путей обеспечения равномерной по малому параметру сходимости — использование сгущающихся сеток. Одна из простейших сеток, называемая *сеткой Шишкина*, имеет вид (см. рис. 2)

$$\begin{aligned} \bar{\Omega} = \{x_i \mid x_i = ih, i = 0, 1, \dots, N/2, x_i = x_{N/2} + (i - N/2)H, \\ i = N/2 + 1, \dots, N, \quad h = \delta/(N/2), \quad H = (1 - \delta)/(N/2), \\ \delta = \min \{c\varepsilon \ln N, 1/2\}\}, \end{aligned}$$

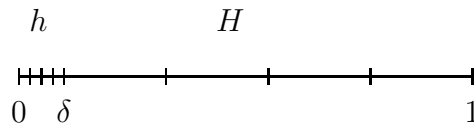


Рис. 2

или

$$x_i = x(t_i), \quad \text{где } t_i = i/N, \quad \text{а}$$

$$x(t) = \begin{cases} 2\delta t, & 0 \leq t \leq 1/2, \\ 1 - 2(1 - \delta)(1 - t), & 1/2 < t \leq 1 \end{cases}$$

есть кусочно-линейное непрерывное отображение отрезка  $[0, 1]$  на себя. Эта сетка является кусочно равномерной с шагом  $h \ll H$  на отрезке  $[0, \delta]$  и с шагом  $H$  на отрезке  $[\delta, 1]$ .

Равномерная по малому параметру точность разностной схемы определяется погрешностью аппроксимации разностной схемы и величиной параметра  $c$ , который должен быть выбран таким, чтобы на длине  $\delta$  быстро меняющаяся составляющая точного решения успела принять столь малое значение, которое уже не влияет на погрешность приближенного решения.

## Численные методы для задач с негладкими решениями

Рассмотрим следующее дифференциальное уравнение

$$-\frac{1}{x}(xu')' + \frac{\lambda^2}{x^2}u = 0, \quad 0 < x < 1. \quad (20.1)$$

Это уравнение не вкладывается в тот класс уравнений, который мы для себя выделили. Именно, коэффициент  $p(x) := x \geq 0$ , но не отрезан от нуля постоянной (на рассматриваемом отрезке). Поэтому для уравнения (20.1) в точке  $x = 0$  нельзя ставить произвольное граничное условие. В самом деле, будем искать решение уравнения (20.1) в виде

$$u(x) = x^\alpha.$$

Подставляя это выражение в (20.1), находим, что для удовлетворения уравнения требуется выполнение условия

$$\alpha^2 = \lambda^2,$$

т.е.  $\alpha = \pm\lambda$ . Тем самым, мы нашли два фундаментальных решения уравнения (20.1), и его общее решение есть

$$u(x) = c_1x^\lambda + c_2x^{-\lambda}. \quad (20.2)$$

Без ограничения общности можно считать, что  $\lambda > 0$ . Если нас интересует ограниченное решение (что естественно с точки зрения приложений), то  $c_2 = 0$  и

$$u(x) = c_1x^\lambda.$$

Отсюда находим, что единственным допустимым граничным условием из числа классических является условие

$$u(0) = 0. \quad (20.3)$$

(именно это условие и будет выделять из (20.2) ограниченное решение). При  $x = 1$  можно ставить любое граничное условие, например,

$$u(1) = 1. \quad (20.4)$$

Тогда решением задачи (20.1), (20.3), (20.4) будет функция

$$u(x) = x^\lambda. \quad (20.5)$$

Если  $0 < \lambda < 1$ , то уже первая производная интересующего нас решения не ограничена, не говоря уже о четвертой производной, которая фигурирует в погрешности аппроксимации. О хорошей сходимости численного решения на равномерной сетке говорить трудно. Выход из создавшегося положения можно найти на пути использования специальной сгущающейся к точке  $x = 0$  сетки. Как эту сетку построить? Пусть  $x = x(t)$  есть отображение отрезка  $[0, 1]$  на себя. Для  $t \in [0, 1]$  введем равномерную сетку с шагом  $\tau = 1/N$ . Тогда

$$x_i = x(t_i)$$

будет задавать узлы неравномерной сетки по  $x$ . На этой неравномерной сетке

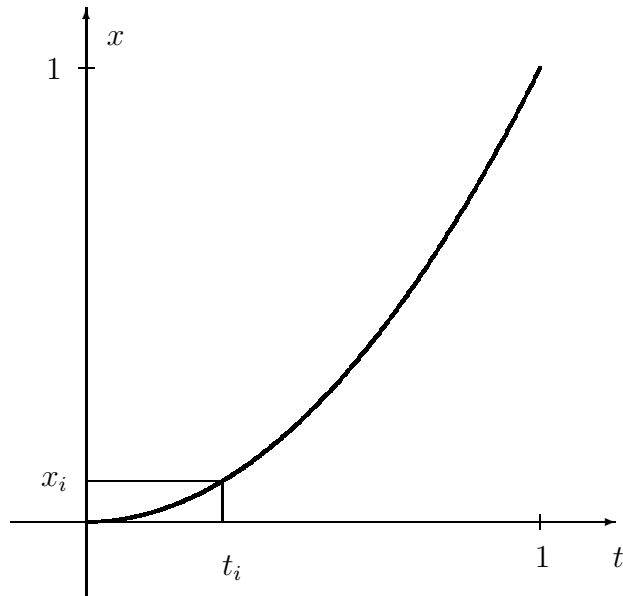


Рис. 1

и аппроксимируем уравнение (20.1). Пусть

$$h_i = x_i - x_{i-1}, \quad \bar{h}_i = (h_i + h_{i+1})/2.$$

Тогда, используя, например, метод баланса (см. § 19), для уравнения (20.1) получим следующую аппроксимацию

$$\frac{1}{x_i} \frac{1}{\bar{h}_i} \left( x_{i+1/2} \frac{u_{i+1}^h - u_i^h}{h_{i+1}} - x_{i-1/2} \frac{u_i^h - u_{i-1}^h}{h_i} \right) - \frac{\lambda^2}{x_i^2} u_i^h = 0. \quad (20.6)$$

Исследуем погрешность аппроксимации этой разностной схемы на неравномерной сетке. Используя формулу Тейлора, будем иметь

$$\begin{aligned} \Psi_i &= \frac{1}{x_i} \frac{1}{\bar{h}_i} [x_{i+1/2} u_{x,i} - x_{i-1/2} u_{\bar{x},i}] - \frac{1}{x_i} (xu')'_i = \\ &= \frac{1}{x_i} \frac{1}{\bar{h}_i} \left[ \left( x_i + \frac{h_{i+1}}{2} \right) \left( u'_i + \frac{h_{i+1}}{2} u''_i + \frac{h_{i+1}^2}{6} u'''_i + \frac{h_{i+1}^3}{24} \tilde{u}_i^{IV} \right) - \right. \\ &\quad \left. - \left( x_i - \frac{h_i}{2} \right) \left( u'_i - \frac{h_i}{2} u''_i + \frac{h_i^2}{6} u'''_i - \frac{h_i^3}{24} \tilde{u}_i^{IV} \right) \right] - \frac{1}{x_i} (xu')'_i = \\ &= \frac{1}{x_i} \frac{1}{\bar{h}_i} \left[ h_{i+1}^2 \left( \frac{x_i}{6} u'''_i + \frac{1}{4} u''_i \right) + h_{i+1}^3 \left( \frac{x_i}{24} \tilde{u}_i^{IV} + \frac{1}{12} u'''_i \right) + \frac{h_{i+1}^4}{48} \tilde{u}_i^{IV} - \right. \\ &\quad \left. - h_i^2 \left( \frac{x_i}{6} u'''_i + \frac{1}{4} u''_i \right) + h_i^3 \left( \frac{x_i}{24} \tilde{u}_i^{IV} + \frac{1}{12} u'''_i \right) - \frac{h_i^4}{48} \tilde{u}_i^{IV} \right] = \\ &= \frac{h_{i+1}^2 - h_i^2}{\bar{h}_i} \left( \frac{1}{6} u'''_i + \frac{1}{4} \frac{u''_i}{x_i} \right) + \frac{h_{i+1}^3}{\bar{h}_i} \left( \frac{1}{24} \tilde{u}_i^{IV} + \frac{1}{12} \frac{u'''_i}{x_i} \right) + \\ &\quad + \frac{h_i^3}{\bar{h}_i} \left( \frac{1}{24} \tilde{u}_i^{IV} + \frac{1}{12} \frac{u'''_i}{x_i} \right) + \frac{1}{48} \frac{h_{i+1}^4}{\bar{h}_i} \frac{\tilde{u}_i^{IV}}{x_i} - \frac{1}{48} \frac{h_i^4}{\bar{h}_i} \frac{\tilde{u}_i^{IV}}{x_i}. \end{aligned} \quad (20.7)$$

Подставим сюда истинное значение  $u(x)$  из (20.5) и оценим вклад в погрешность решения типичной составляющей погрешности аппроксимации

$$\overset{\circ}{\psi}_i = c(x_i) h_i^2 x_i^{\lambda-4}.$$

Эта составляющая представлена в погрешности аппроксимации (20.7) вторым и третьим слагаемыми. Составляющую погрешности решения, отвечающую  $\overset{\circ}{\psi}_i$ , обозначим через  $\overset{\circ}{z}_i$ . Для нее имеем уравнение

$$-\frac{1}{x_i} \frac{1}{\bar{h}_i} \left( x_{i+1/2} \frac{\overset{\circ}{z}_{i+1} - \overset{\circ}{z}_i}{h_{i+1}} - x_{i-1/2} \frac{\overset{\circ}{z}_i - \overset{\circ}{z}_{i-1}}{h_i} \right) + \frac{\lambda^2}{x_i^2} \overset{\circ}{z}_i = c(x_i) h_i^2 x_i^{\lambda-4}.$$

Как и при доказательстве теоремы 18.1 для максимума  $|\overset{\circ}{z}_i|$  получаем оценку

$$\max_i |\overset{\circ}{z}_i| \leq \max_i \frac{c(x_i) h_i^2 x_i^{\lambda-2}}{\lambda^2}. \quad (20.8)$$

Из этой оценки следует, что, если  $\lambda \geq 2$ , то никаких проблем нет, ибо в этом случае выражение, стоящее в правой части под знаком  $\max$ , имеет равномерную по  $x_i$  малость  $O(h_i^2)$ , и сетку можно брать равномерной. Если же  $\lambda < 2$ , то равномерной по  $x_i$  малости  $O(h_i^2)$  указанного выражения не гарантируется, если сетка не выбрана надлежащим образом. Поскольку  $c(x_i)$  из правой части (20.8) меняется мало, выберем сетку при  $\lambda < 2$  так, чтобы

$$h_i^2 x_i^{\lambda-2} \approx \text{const.}$$

Так как

$$h_i = x_i - x_{i-1} = N^{-1} x'(t_i^*), \quad (20.9)$$

то

$$h_i^2 x_i^{\lambda-2} = N^{-2} x'^2(t_i^*) x_i^{\lambda-2}.$$

Пусть

$$x'^2 x^{\lambda-2} = c,$$

где  $c$  — некоторая постоянная, или

$$x' x^{\lambda/2-1} = \sqrt{c} = c_1.$$

Интегрируя это уравнение, находим, что

$$x^{\lambda/2} = c_1 t + c_2,$$

или

$$x = (c_1 t + c_2)^{2/\lambda}.$$

Так как  $x(0) = 0$ , а  $x(1) = 1$ , то  $c_2 = 0$ , а  $c_1 = 1$ . Тем самым,

$$x = t^{2/\lambda}, \quad (20.10)$$

и, следовательно,

$$x_i = (i/N)^{2/\lambda}. \quad (20.11)$$

Если узлы сетки будут заданы по закону (20.11), то, в силу (20.9), (20.10) при  $\lambda < 2$

$$h_i = 2N^{-1} t_i^{2/\lambda-1} / \lambda,$$

и величины шагов сетки уменьшаются при приближении к границе  $x = 0$ , т.е. построенная сетка является сгущающейся в окрестности  $x = 0$ . Если  $i \sim N$ , то  $h_i \sim cN^{-1}$ , а если  $i = 1$ , то

$$h_1 = \frac{2}{\lambda} N^{-2/\lambda} \quad (\lambda < 2).$$

Принимая во внимание сказанное, а также (20.5), (20.10) и (20.8), легко проверить, что вклад последних четырех слагаемых погрешности аппроксимации (20.7) в погрешность решения является величиной  $O(N^{-2})$ .

Обратимся, например, к первому слагаемому правой части (20.7). Используя формулу Тейлора, находим, что

$$\frac{h_{i+1}^2 - h_i^2}{\bar{h}_i} = 2(h_{i+1} - h_i) = 2(x_{i+1} - 2x_i + x_{i-1}) = 2N^{-2}x''(t^*) = cN^{-2}x^{1-\lambda}.$$

Снова принимая во внимание (20.5), (20.10) и (20.8), заключаем, что вклад и первого слагаемого погрешности аппроксимации (20.7) в погрешность решения оценивается величиной  $O(N^{-2})$ .

# VI

## Численные методы для дифференциальных уравнений с частными производными

## Разностные методы для уравнения теплопроводности

Нестационарное уравнение теплопроводности является собой простейший пример параболического уравнения — уравнения с частными производными. Возьмем его в виде

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x, t), \quad 0 < x < 1, \quad 0 < t \leq T. \quad (21.1)$$

Чтобы выделить единственное решение уравнения (21.1), нужно задать дополнительные условия. Таковыми могут быть граничные условия, задаваемые при  $x = 0$  и  $x = 1$ , и начальное условие, задаваемое при  $t = 0$ . Пусть, например, граничные условия имеют вид

$$u(0, t) = u(1, t) = 0, \quad (21.2)$$

а начальное условие —

$$u(x, 0) = \varphi(x). \quad (21.3)$$

Как известно из курса методов математической физики, задача (21.1)–(21.3) поставлена корректно и при надлежащей гладкости  $f(x, t)$  и  $\varphi(x)$  имеет единственное решение.

Посмотрим на уравнение (21.1) с точки зрения краевых задач для обыкновенных дифференциальных уравнений. Для этого обозначим  $\partial u / \partial t = \dot{u}$  и перепишем (21.1) в виде

$$-\frac{\partial^2 u}{\partial x^2} = f(x, t) - \dot{u} \equiv \mathcal{F}(x, t). \quad (21.4)$$

Считая  $\mathcal{F}(x, t)$  в (21.4) заданной функцией, а  $t$  — параметром, мы можем условно рассматривать (21.4) как обыкновенное дифференциальное



уравнение, аппроксимацию которого мы строить умеем. На  $[0, 1]$  введем сетку

$$\bar{\omega}^h = \{x = x_i = ih \mid i = 0, \dots, N\}$$

с внутренними узлами

$$\omega^h = \{x_i \in \bar{\omega}^h \mid i = 1, \dots, N-1\}$$

и на этой сетке дифференциальное уравнение (21.4) аппроксимируем разностным уравнением

$$-u_{\bar{x}x,i}^h = \mathcal{F}^h(x_i, t), \quad x_i \in \omega^h. \quad (21.5)$$

Теперь нужно вспомнить (21.4), в силу которого

$$\mathcal{F}(x_i, t) = f(x_i, t) - \dot{u}(x_i, t).$$

Поэтому естественно положить

$$\mathcal{F}^h(x_i, t) = f^h(x_i, t) - \dot{u}_i^h.$$

Тогда (21.5) примет вид

$$-u_{\bar{x}x,i}^h = f_i^h(t) - \dot{u}_i^h, \quad x_i \in \omega^h. \quad (21.6)$$

Это соотношение представляет собой систему  $(N-1)$  обыкновенных дифференциальных уравнений первого порядка с  $(N+1)$  неизвестными  $u_i^h$ ,  $i = 0, \dots, N$ . Воспользуемся граничными условиями (21.2) и положим

$$u_0^h(t) = u_N^h(t) = 0. \quad (21.7)$$

После исключения этих неизвестных из (21.6) будем иметь систему  $(N-1)$  уравнений с  $(N-1)$  неизвестными.

Перепишем теперь (21.6) по-другому, поставив на первое место производную

$$\dot{u}_i^h = u_{\bar{x}x,i}^h + f_i^h(t), \quad i = 1, \dots, N-1, \quad (21.8)$$

и введем обозначения

$$\begin{aligned} U &= [u_1^h \dots u_{N-1}^h]^T, \\ \Lambda &= \frac{1}{h^2} \begin{bmatrix} -2 & 1 & 0 & \dots & 0 & 0 \\ 1 & -2 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & -2 \end{bmatrix}, \\ F &= [f_1^h \dots f_{N-1}^h]. \end{aligned} \quad (21.9)$$

Тогда система (21.8) с учетом (21.7) примет вид

$$\frac{dU}{dt} = \Lambda U + F. \quad (21.10)$$

Введем еще одно обозначение

$$\Phi = [\varphi_1 \dots \varphi_{N-1}]^T$$

и положим

$$U(0) = \Phi. \quad (21.11)$$

Соотношения (21.10), (21.11) представляют собой задачу Коши для системы обыкновенных дифференциальных уравнений первого порядка. Для приближенного решения этой задачи можно использовать уже изученные методы. Например, метод Эйлера, который приводит к соотношениям

$$\frac{U^{j+1} - U^j}{\tau} = \Lambda U^j + F^j, \quad U^0 = \Phi, \quad (21.12)$$

или неявный метод Эйлера

$$\frac{U^{j+1} - U^j}{\tau} = \Lambda U^{j+1} + F^{j+1}, \quad U^0 = \Phi, \quad (21.13)$$

а можно и метод трапеций

$$\frac{U^{j+1} - U^j}{\tau} = \frac{1}{2}\Lambda [U^{j+1} + U^j] + \frac{1}{2}(F^{j+1} + F^j), \quad U^0 = \Phi. \quad (21.14)$$

Мы не будем изучать общие методы решения задачи (21.10), (21.11), а ограничимся одношаговыми, как (21.12)-(21.14), которые в теории разностных схем для параболических уравнений принято называть двухслойными.

Изучение (21.12), (21.13) и (21.14) можно проводить одновременно, если записать их единым образом за счет введения параметра  $\sigma$ :

$$\frac{U^{j+1} - U^j}{\tau} = \sigma \Lambda U^{j+1} + (1 - \sigma) \Lambda U^j + \sigma F^{j+1} + (1 - \sigma) F^j. \quad (21.15)$$

Полагая здесь  $\sigma = 0, 1$  или  $1/2$ , получим (21.12), (21.13) или (21.14), соответственно.

Посмотрим теперь на (21.15) с точки зрения аппроксимации не задачи Коши для системы обыкновенных дифференциальных уравнений (21.10), (21.11), а с точки зрения аппроксимации задачи (21.1)-(21.3). В результате

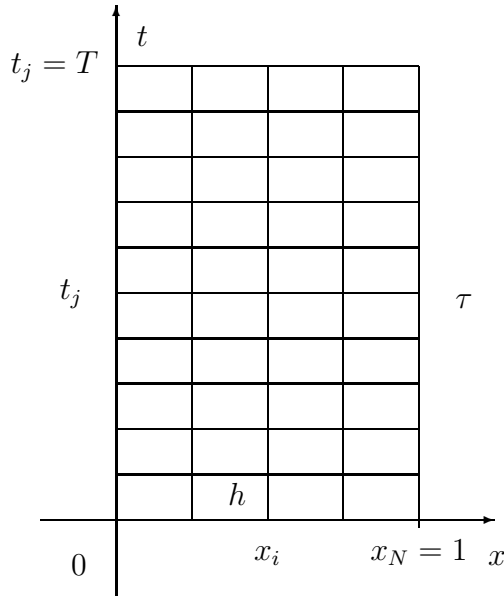


Рис. 1

двух шагов аппроксимации в области  $[0, 1] \times [0, T]$  образована сетка (см. рис. 1),

на которой дифференциальное уравнение (21.1) аппроксимировано системой разностных уравнений

$$\frac{u_i^{hj+1} - u_i^{hj}}{\tau} = \sigma u_{\bar{x}x,i}^{hj+1} + (1-\sigma)u_{\bar{x}x,i}^{hj} + f_i^{hj}, \quad i = 1, \dots, N-1, \quad j = 0, \dots, J-1, \quad (21.16)$$

а граничные (21.2) и начальное (21.3) условия — соотношениями

$$u_0^{hj} = 0, \quad u_N^{hj} = 0, \quad j = 1, \dots, J, \quad (21.17)$$

и

$$u_i^{h0} = \varphi_i, \quad i = 0, \dots, N, \quad (21.18)$$

соответственно. (На связи  $f_i^{hj}$  с  $f(x, t)$  мы не останавливаемся).

Введем дополнительные обозначения

$$u_i^j = u, \quad u_i^{j+1} = \hat{u}, \quad (\hat{u} - u)/\tau = u_t.$$

В новых обозначениях уравнения (21.16) примут вид

$$u_t^h = \sigma \hat{u}_{\bar{x}x}^h + (1-\sigma)u_{\bar{x}x}^h + f^h. \quad (21.19)$$

Погрешностью аппроксимации уравнения (21.1) уравнениями (21.19) будет сеточная функция

$$\Psi = f^h + \sigma \hat{u}_{\bar{x}\bar{x}} + (1 - \sigma)u_{\bar{x}\bar{x}} - u_t,$$

где  $u = u(x_i, t_j)$  — значения решения уравнения (21.1) в узлах  $(x_i, t_j)$ .

**Упражнение 21.1.** Доказать, что при надлежащей гладкости (какой?)

$$\Psi = \begin{cases} O(\tau + h^2) & \text{при } \sigma = 0, \sigma = 1, \\ O(\tau^2 + h^2) & \text{при } \sigma = 1/2. \end{cases}$$

**Замечание 21.1.** Для написания уравнений (21.19) при  $\sigma = 0$ ,  $\sigma = 1$  или  $\sigma = 1/2$  требуются следующие множества узлов

$$\begin{array}{cccc} j+1 & & \bullet & & \bullet & \bullet & \bullet & & \bullet & \bullet & \bullet \\ j & & \bullet & \bullet & \bullet & & \bullet & & \bullet & \bullet & \bullet \end{array}$$

соответственно, называемые шаблонами.

## 21.1 Устойчивость по начальным данным

Исследуем разностную схему (21.16)-(21.18) на предмет ее устойчивости по начальным данным. Для этого будем считать, что правая часть в уравнениях (21.16) равна нулю, т.е.

$$\frac{u_i^{hj+1} - u_i^{hj}}{\tau} = \sigma u_{\bar{x}\bar{x},i}^{hj+1} + (1 - \sigma)u_{\bar{x}\bar{x},i}^{hj}, \quad \begin{array}{l} i = 1, \dots, N-1, \\ j = 0, \dots, J-1. \end{array} \quad (21.20)$$

Чтобы исследовать вопрос об устойчивости, найдем решение задачи (21.20), (21.17), (21.18). Решение будем искать методом разделения переменных. Будем искать частные решения уравнений (21.20) в виде

$$u_i^j = X_i T_j.$$

Тогда

$$\frac{T_{j+1} - T_j}{\tau} X_i = (\sigma T_{j+1} + (1 - \sigma)T_j) X_{\bar{x}\bar{x},i}$$

или

$$\frac{(T_{j+1} - T_j)/\tau}{\sigma T_{j+1} + (1 - \sigma)T_j} = \frac{X_{\bar{x}\bar{x},i}}{X_i} = -\lambda^h, \quad (21.21)$$

где  $\lambda^h$  — постоянная. С учетом граничных условий (21.17) для  $X_i$  из (21.21) получим задачу

$$X_{\bar{x}x,i} + \lambda^h X_i = 0, \quad i = 1, 2, \dots, N-1, \quad X_0 = X_N = 0,$$

или, в развернутом виде,

$$-X_{i-1} + 2X_i - X_{i+1} = h^2 \lambda^h X_i, \quad i = 1, 2, \dots, N-1, \quad X_0 = X_N = 0. \quad (21.22)$$

Но эта задача совпадает с рассмотренной нами ранее задачей (10.33), если в последней под  $\lambda$  понимать  $h^2 \lambda^h$ . Поэтому, в силу (10.35)

$$X_i^{(k)} = \sqrt{2} \sin k\pi x_i, \quad i = 1, \dots, N-1 \quad (21.23)$$

суть собственные векторы задачи (21.22), которые ортогональны в смысле скалярного произведения

$$(u, v) = \sum_{i=1}^{N-1} u_i v_i h$$

и нормированы, т.е.  $\|X_i^{(k)}\|^2 = (X_i^{(k)}, X_i^{(k)}) = 1$ . В силу (10.42)

$$\lambda_k^h = \frac{4}{h^2} \sin^2 \frac{k\pi h}{2}, \quad k = 1, \dots, N-1 \quad (21.24)$$

— различные собственные значения этой задачи.

Далее, из (21.21) находим, что

$$\frac{T_{j+1}^{(k)} - T_j^{(k)}}{\tau} + \lambda_k^h \left[ \sigma T_{j+1}^{(k)} + (1 - \sigma) T_j^{(k)} \right] = 0$$

или

$$(1 + \sigma \tau \lambda_k^h) T_{j+1}^{(k)} = (1 - (1 - \sigma) \tau \lambda_k^h) T_j^{(k)}.$$

Отсюда следует, что

$$T_{j+1}^{(k)} = q_k T_j^{(k)},$$

где

$$q_k = \frac{1 - (1 - \sigma) \tau \lambda_k^h}{1 + \sigma \tau \lambda_k^h} \quad (21.25)$$

и поэтому

$$T_j^{(k)} = c_k q_k^j. \quad (21.26)$$

Итак, мы нашли, что функции

$$u_i^{j(k)} = X_i^{(k)} T_j^{(k)}, \quad k = 1, \dots, N-1,$$

где  $X_i^{(k)}$  и  $T_j^{(k)}$  из (21.23) и (21.26), соответственно, являются частными решениями уравнений (21.20), удовлетворяющими граничным условиям (21.17). Построим линейную комбинацию этих решений

$$u_i^{hj} = \sum_{k=1}^{N-1} c_k X_i^{(k)} q_k^j. \quad (21.27)$$

Полагая здесь  $j = 0$ , получим

$$u_i^{h0} = \sum_{k=1}^{N-1} c_k X_i^{(k)},$$

а принимая во внимание (21.18), заключаем, что функция (21.27) будет удовлетворять начальным условиям (21.18), если

$$\sum_{k=1}^{N-1} c_k X_i^{(k)} = \varphi_i,$$

т.е. если постоянные  $c_k$  суть коэффициенты Фурье функции  $\varphi_i$  при разложении по ортонормированной системе  $X_i^{(k)}$

$$c_k = (\varphi, X_i^{(k)}) = \sum_{i=1}^{N-1} \varphi_i X_i^{(k)} h. \quad (21.28)$$

Итак, сеточная функция (21.27) с коэффициентами  $c_k$  из (21.28) удовлетворяет уравнениям (21.20), граничным условиям (21.17) и начальному условию (21.18), а поэтому является решением задачи (21.20), (21.17), (21.18).

Найдем оценку этого решения. Возводя левую и правую части (21.27) в квадрат и суммируя результат по  $i$  от 1 до  $N-1$ , с учетом ортогональности  $X_i^{(k)}$ , будем иметь

$$\begin{aligned} \|u^{hj}\|_{L_2^h}^2 &= \sum_{i=1}^{N-1} (u_i^{hj})^2 h = \sum_{i=1}^{N-1} h \sum_{k,l=1}^{N-1} c_k c_l X_i^{(k)} X_i^{(l)} q_k^j q_l^j = \\ &= \sum_{k,l=1}^{N-1} c_k c_l q_k^j q_l^j (X_i^{(k)}, X_i^{(l)}) = \sum_{k=1}^{N-1} c_k^2 q_k^{2j} \leq \max_k q_k^{2j} \sum_{k=1}^{N-1} c_k^2 = \max_k q_k^{2j} \|\varphi\|_{L_2^h}^2. \end{aligned}$$

Пусть

$$|q_k| \leq 1. \quad (21.29)$$

Тогда

$$\|u^{hj}\|_{L_2^h} \leq \|\varphi\|_{L_2^h}, \quad (21.30)$$

т.е.  $L_2^h$ -норма решения при любом  $j$  не превосходит  $L_2^h$ -нормы начального условия.

Выясним, когда выполняется условие (21.29). С учетом (21.24), (21.25) при  $\sigma \geq 0$  имеем

$$-\left(1 + \frac{4\tau}{h^2}\sigma \sin^2 \frac{k\pi h}{2}\right) \leq 1 - \frac{4\tau}{h^2}(1 - \sigma) \sin^2 \frac{k\pi h}{2}$$

или, после приведения подобных членов,

$$2 - \frac{4\tau}{h^2}(1 - 2\sigma) \sin^2 \frac{k\pi h}{2} \geq 0, \quad k = 1, \dots, N - 1.$$

Отсюда вытекает условие

$$(1 - 2\sigma) \leq \frac{h^2}{2\tau} \min_k \frac{1}{\sin^2 \frac{k\pi h}{2}}.$$

Поскольку  $\min_k \sin^{-2} \frac{k\pi h}{2} \geq 1$ , то (21.29) будет выполнено, если

$$(1 - 2\sigma) \leq \frac{h^2}{2\tau}$$

или, что эквивалентно,

$$\sigma \geq \frac{1}{2} - \frac{h^2}{4\tau}. \quad (21.31)$$

Итак, нами доказана

**Теорема 21.1.** *Если параметр  $\sigma$  схемы (21.20) удовлетворяет условию (21.31), то для решения задачи (21.20), (21.17), (21.18) справедлива априорная оценка*

$$\max_j \|u^{hj}\|_{L_2^h} \leq \|u^{h0}\|_{L_2^h}, \quad j = 1, 2, \dots, J. \quad (21.32)$$

**Определение 21.1.** Говорят, что разностная схема (21.20) устойчива по начальным данным, если для решения задачи (21.20), (21.17), (21.18) справедлива оценка

$$\|u^{hj}\|_{(1)} \leq M \|u^{h0}\|_{(2)},$$

где  $\|\cdot\|_{(1)}$  и  $\|\cdot\|_{(2)}$  — некоторые нормы, а  $M = \text{const} > 0$  не зависит от  $\tau$  и  $h$ .

**Следствие 1.** Теорема 21.1 утверждает устойчивость по начальным данным схемы (21.20) при выполнении условий (21.31), когда

$$\|\cdot\|_{(2)} = \|\cdot\|_{L_2^h}, \quad \|\cdot\|_{(1)} = \|\cdot\|_{L_\infty^h(0,T) \times L_2^h(0,1)}.$$

Обсудим условие (21.31). Если  $\sigma = 1$ , т.е. использован неявный метод Эйлера для системы, то (21.31) выполнено при любых  $\tau$  и  $h$ . То же самое имеет место и при  $\sigma = 1/2$  (схема трапеций). Если же  $\sigma = 0$ , то для выполнения (21.31) нужно, чтобы

$$\tau \leq h^2/2. \quad (21.33)$$

Про первые две схемы (при  $\sigma = 1$  и  $\sigma = 1/2$ ) говорят, что они безусловно устойчивы, а третья ( $\sigma = 0$ ) устойчива условно (для устойчивости шага по временной переменной и по пространственной связаны неравенством (21.33)).

Напомним, что все три схемы нуль-устойчивы по терминологии из обыкновенных дифференциальных уравнений, а первые две еще и А-устойчивы.

Отметим, что числа  $(-\lambda_k^h)$  являются собственными числами матрицы (21.9)

$$\begin{aligned} \frac{\max_k |-\lambda_k^h|}{\min_k |-\lambda_k^h|} &= \frac{\lambda_{N-1}^h}{\lambda_1^h} = \frac{\sin^2 \frac{(N-1)\pi h}{2}}{\sin^2 \frac{\pi h}{2}} = \\ &= \frac{\cos^2 \frac{\pi h}{2}}{\sin^2 \frac{\pi h}{2}} = \operatorname{ctg}^2 \frac{\pi h}{2} \gg 1 \quad \text{при } h \ll 1, \end{aligned}$$

т.е. система уравнений (21.8) жесткая.

## 21.2 Устойчивость по правой части

Обратимся теперь к неоднородному уравнению (21.19), а вместо (21.18) поставим однородные начальные условия

$$u_i^{h0} = 0, \quad i = 0, \dots, N. \quad (21.34)$$

**Теорема 21.2.** Если параметр  $\sigma$  схемы (21.19) удовлетворяет условию (21.31), то для решения задачи (21.19), (21.17), (21.34) справедлива априорная оценка

$$\max_j \|u^{hj}\|_{L_2^h} \leq T \max_j \|f^{hj}\|_{L_2^h}. \quad (21.35)$$



**Доказательство.** Разложим  $u_i^{hj}$  и  $f_i^{hj}$  при каждом  $j$  по собственным векторам задачи (21.22)

$$u_i^{hj} = \sum_{k=1}^{N-1} T_j^{(k)} X_i^{(k)}, \quad f_i^{hj} = \sum_{k=1}^{N-1} f_j^{(k)} X_i^{(k)}.$$

Подставляя эти разложения в (21.19) и принимая во внимание ортогональность  $X_i^{(k)}$ , получим

$$\frac{T_{j+1}^{(k)} - T_j^{(k)}}{\tau} + \lambda_k^h \left[ \sigma T_{j+1}^{(k)} + (1 - \sigma) T_j^{(k)} \right] = f_j^{(k)}.$$

Приводя подобные члены, найдем, что

$$[1 + \sigma \tau \lambda_k^h] T_{j+1}^{(k)} = [1 - (1 - \sigma) \tau \lambda_k^h] T_j^{(k)} + \tau f_j^{(k)},$$

а, разрешая относительно  $T_{j+1}^{(k)}$ , будем иметь

$$T_{j+1}^{(k)} = q_k T_j^{(k)} + \frac{\tau}{1 + \sigma \tau \lambda_k^h} f_j^{(k)}.$$

В силу (21.29)  $|q_k| \leq 1$ , а при  $\sigma \geq 0$  знаменатель  $(1 + \sigma \tau \lambda_k^h) \geq 1$  и поэтому

$$|T_{j+1}^{(k)}| \leq |T_j^{(k)}| |q_k| + \tau |f_j^{(k)}| \leq |T_j^{(k)}| + \tau |f_j^{(k)}|.$$

Далее

$$\|u^{hj+1}\|_{L_2^h} = \sqrt{\sum_{k=1}^{N-1} \left(T_{j+1}^{(k)}\right)^2} \leq \sqrt{\sum_{k=1}^{N-1} \left(|T_j^{(k)}| + \tau |f_j^{(k)}|\right)^2} \leq \|u^{hj}\|_{L_2^h} + \tau \|f^{hj}\|_{L_2^h}.$$

Суммируя это неравенство по  $j$  в нужных пределах, приходим к (21.35). Теорема доказана.

**Теорема 21.3** (сходимости). *Если выполнено условие (21.31), и решение задачи (21.1)-(21.3)  $u(x, t) \in C^4[0, 1] \times C^3[0, T]$ , то решение  $u^h$  задачи (21.16)-(21.18) при соответствующей  $f_i^{hj}$  сходится к решению  $u$  задачи (21.1)-(21.3) со скоростью  $O(h^2 + (\sigma - 1/2)\tau + \tau^2)$ .*

**Доказательство.** Пусть  $z_i^j = u_i^{hj} - u(x_i, t_j)$  — погрешность решения. Выражая  $u_i^{hj}$  через  $z_i^j$  и  $u(x_i, t_j)$  и подставляя результат в (21.16)-(21.18), для  $z_i^j$  получим задачу

$$\begin{aligned} \frac{z_i^{j+1} - z_i^j}{\tau} &= \sigma z_{\bar{x},i}^{j+1} + (1 - \sigma) z_{\bar{x},i}^j + \Psi_i^j, \\ z_0^j &= z_N^j = 0, \quad z_i^0 = 0. \end{aligned} \quad (21.36)$$

Для задачи (21.36) справедлива оценка, устанавливаемая теоремой 21.2, т.е.

$$\max_j \|z_i^j\|_{L_2^h} \leq T \max_j \|\Psi_i^j\|_{L_2^h}.$$

Используя теперь результаты упражнения 21.1, приходим к утверждению теоремы.

### 21.3 Устойчивость в смысле максимума модуля

**Теорема 21.4.** *Если выполнено условие*

$$\frac{2(1-\sigma)}{h^2}\tau \leq 1, \quad (21.37)$$

то для решения задачи (21.16)-(21.18) справедлива априорная оценка

$$\max_{ij} |u_i^{hj}| \leq \max_i |\varphi_i| + T \max_{ij} |f_i^{hj}|. \quad (21.38)$$

**Доказательство.** Перепишем уравнение (21.16) в поточечном виде

$$\frac{u_i^{hj+1} - u_i^{hj}}{\tau} = \sigma \frac{u_{i-1}^{hj+1} - 2u_i^{hj+1} + u_{i+1}^{hj+1}}{h^2} + (1-\sigma) \frac{u_{i-1}^{hj} - 2u_i^{hj} + u_{i+1}^{hj}}{h^2} + f_i^{hj}$$

и приведем подобные члены

$$\begin{aligned} & \left( \frac{1}{\tau} + \frac{2\sigma}{h^2} \right) u_i^{hj+1} = \\ & = \frac{\sigma}{h^2} u_{i-1}^{hj+1} + \frac{\sigma}{h^2} u_{i+1}^{hj+1} + \left( \frac{1}{\tau} - \frac{2(1-\sigma)}{h^2} \right) u_i^{hj} + \frac{1-\sigma}{h^2} u_{i-1}^{hj} + \frac{1-\sigma}{h^2} u_{i+1}^{hj} + f_i^{hj}. \end{aligned}$$

Возьмем модули левой и правой частей и оценим правую часть этого соотношения через максимальные значения модулей  $u_i^{hj}$ ,  $u_i^{hj+1}$  и  $f_i^{hj}$ .

Будем иметь

$$\begin{aligned} & \left( \frac{1}{\tau} + \frac{2\sigma}{h^2} \right) |u_i^{hj+1}| \leq \\ & \leq \frac{2\sigma}{h^2} \max_i |u_i^{hj+1}| + \left( \left| \frac{1}{\tau} - \frac{2(1-\sigma)}{h^2} \right| + \frac{2(1-\sigma)}{h^2} \right) \max_i |u_i^{hj}| + \max_i |f_i^{hj}|. \end{aligned}$$

Беря теперь максимум по  $i$  левой части и приводя подобные члены, после домножения на  $\tau$  получим:

$$\max_i |u_i^{hj+1}| \leq \left( \left| 1 - \frac{2(1-\sigma)\tau}{h^2} \right| + \frac{2(1-\sigma)\tau}{h^2} \right) \max_i |u_i^{hj}| + \tau \max_i |f_i^{hj}|.$$

Изучим коэффициент при  $\max_i |u_i^{hj}|$ :

$$\left| 1 - \frac{2(1-\sigma)\tau}{h^2} \right| + \frac{2(1-\sigma)\tau}{h^2} = \begin{cases} 1 & \text{при } \frac{2(1-\sigma)\tau}{h^2} \leq 1, \\ \frac{4(1-\sigma)\tau}{h^2} - 1 > 1 & \text{при } \frac{2(1-\sigma)\tau}{h^2} > 1 \end{cases}.$$

В силу условия (21.37) теоремы реализуется первая возможность, и следовательно

$$\max_i |u_i^{h,j+1}| \leq \max_i |u_i^{hj}| + \tau \max_i |f_i^{hj}|. \quad (21.39)$$

Пусть

$$\max_j \max_i |u_i^{hj}| = \max_i |u_i^{h,j_0}|.$$

Тогда

$$\max_{ij} |u_i^{hj}| = \max_i |u_i^{h,j_0}| \leq \max_i |u_i^{h,j_0-1}| + \tau \max_i |f_i^{h,j_0-1}|.$$

Прибавляя сюда все предыдущие неравенства (21.39) при  $j = j_0 - 2, \dots, j = 0$ , получим

$$\max_{ij} |u_i^{hj}| \leq \max_i |\varphi_i| + \sum_{j=1}^J \tau \max_i |f_i^{hj-1}| \leq \max_i |\varphi_i| + T \max_{ij} |f_i^{hj}|.$$

Теорема доказана.

**Следствие 2.** При  $\sigma = 1$  оценка (21.38) верна на любой сетке. При  $\sigma = 0$  (21.37) совпадает с (21.31), и для справедливости оценки (21.38) должно быть выполнено условие (21.33). Если же  $\sigma = 1/2$ , то оценка (21.38) верна при

$$\tau/h^2 \leq 1. \quad (21.40)$$

## 21.4 Сеточное преобразование Фурье

Пусть  $\{x_m\}$  — совокупность равноотстоящих узлов на оси  $Ox$ . Будем использовать обозначение

$$v(x_m) = v_m, \quad m \in \mathbb{Z}.$$

Будем предполагать, что  $v_m \in l_2$ , т.е.

$$\sum_{m \in \mathbb{Z}} |v_m|^2 < \infty. \quad (21.41)$$

**Определение 21.2.** Будем называть  $2\pi$ -периодическую функцию

$$(Fv_m)(\xi) = \sum_{m \in \mathbb{Z}} v_m e^{-im\xi} = \tilde{v}(\xi) \quad (21.42)$$

сеточным преобразованием Фурье.

**Определение 21.3.** Обратным сеточным преобразованием Фурье называется сеточная функция

$$(F^{-1}\tilde{v})_m = \frac{1}{2\pi} \int_0^{2\pi} \tilde{v}(\xi) e^{im\xi} d\xi = v_m. \quad (21.43)$$

**Замечание 21.2.** Соотношение (21.42) на самом деле представляет собой сумму ряда Фурье, коэффициентами которого являются значения рассматриваемой нами сеточной функции  $v_m$ . С этой точки зрения (21.43) есть формула для коэффициентов Фурье  $2\pi$ -периодической функции  $\tilde{v}(\xi)$ .

**Замечание 21.3.** Можно было бы называть сеточным преобразованием Фурье функцию

$$Fv_m = \sum_{m \in \mathbb{Z}} hv_m e^{-i(mh)\xi/h} \equiv \tilde{v}(\xi/h), \quad \xi/h = \xi', \quad (21.44)$$

где  $h$  — расстояние между соседними узлами  $h = x_m - x_{m-1}$ . Тогда обратное преобразование приняло бы вид

$$F^{-1}\tilde{v} = \frac{1}{2\pi h} \int_0^{2\pi} \tilde{v}(\xi') e^{i(mh)\xi/h} d\xi = \frac{1}{2\pi} \int_{-\pi/h}^{\pi/h} \tilde{v}(\xi') e^{i(mh)\xi'} d\xi'. \quad (21.45)$$

Устремляя в (21.44) и в (21.45)  $h$  к нулю, получим обычные прямое и обратное преобразование Фурье:

$$Fv(x) = \int_{-\infty}^{\infty} v(x) e^{-ix\xi} dx = \tilde{v}(\xi),$$

$$F^{-1}\tilde{v}(\xi) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{v}(\xi) e^{ix\xi} d\xi = v(x).$$

Для дальнейшего нам потребуется известное из теории рядов Фурье равенство Парсеваля

$$\int_0^{2\pi} |\tilde{v}|^2 d\xi =: \|\tilde{v}\|_{L_2(0,2\pi)}^2 = \frac{1}{2\pi} \|v_m\|_{l_2}^2 := \frac{1}{2\pi} \sum_{m \in \mathbb{Z}} |v_m|^2. \quad (21.46)$$

Пусть  $T$  есть оператор сдвига направо, т.е.

$$Tv_m = v_{m+1}.$$

Обратным к нему будет оператор сдвига налево

$$T^{-1}v_m = v_{m-1}.$$

Найдем сеточное преобразование Фурье этих операторов

$$F(Tv_m) = \sum_{m \in \mathbb{Z}} v_{m+1} e^{-im\xi} e^{-i\xi} e^{i\xi} = e^{i\xi} \tilde{v}(\xi).$$

Аналогично

$$F(T^{-1}v_m) = e^{-i\xi} \tilde{v}(\xi).$$

Теперь найдем преобразование Фурье разностных отношений. Имеем

$$Fv_{x,m} = \frac{1}{h} F(T - I)v_m = \frac{e^{i\xi} - 1}{h} \tilde{v}(\xi), \quad (21.47)$$

$$Fv_{\bar{x},m} = \frac{1}{h} (I - T^{-1})v_m = \frac{1 - e^{-i\xi}}{h} \tilde{v}(\xi), \quad (21.48)$$

$$\begin{aligned} Fv_{\bar{x}x,m} &= \frac{1}{h} F(v_{x,m} - v_{\bar{x},m}) = \frac{e^{i\xi} - 1 - 1 + e^{-i\xi}}{h^2} \tilde{v}(\xi) = \\ &= \frac{(e^{i\xi/2} - e^{-i\xi/2})^2}{h^2} \tilde{v}(\xi) = -\frac{4 \sin^2 \xi/2}{h^2} \tilde{v}(\xi). \end{aligned} \quad (21.49)$$

## 21.5 Устойчивость по начальным данным разностной схемы для уравнения теплопроводности

Рассмотрим разностную схему (21.19) при  $f^h \equiv 0$  на сетке, заданной на всей оси  $Ox$ , т.е. пусть

$$u_{t,m}^h = \sigma \hat{u}_{\bar{x}x,m}^h + (1 - \sigma) u_{\bar{x}x,m}^h, \quad m \in \mathbb{Z}. \quad (21.50)$$

**Теорема 21.5.** Если параметр  $\sigma$  схемы (21.50) положителен и удовлетворяет условию

$$\sigma \geq \frac{1}{2} - \frac{h^2}{4\tau} \quad (21.51)$$

то для решения (21.50) имеет место априорная оценка

$$\max_j \|u^{hj}\|_{L_2^h} \leq \|u^{h0}\|_{L_2^h}, \quad j = 1, 2, \dots \quad (21.52)$$

**Доказательство.** Сделаем в (21.50) сеточное преобразование Фурье

$$\tilde{u}_t + \frac{4 \sin^2 \xi/2}{h^2} (\sigma \hat{u} + (1 - \sigma) \tilde{u}) = 0.$$

Разрешая это обыкновенное разностное уравнение первого порядка относительно  $\hat{u}$ , получим

$$\hat{u} = q(\xi) \tilde{u}, \quad (21.53)$$

где

$$q(\xi) = \frac{1 - (1 - \sigma) \frac{4\tau}{h^2} \sin^2 \frac{\xi}{2}}{1 + \sigma \frac{4\tau}{h^2} \sin^2 \frac{\xi}{2}}. \quad (21.54)$$

Из (21.53)

$$\|\hat{u}\|_{L_2(0,2\pi)} = \|q(\xi) \tilde{u}\|_{L_2(0,2\pi)} \leq \max_{0 \leq \xi \leq 2\pi} |q(\xi)| \|\tilde{u}\|_{L_2(0,2\pi)}.$$

Отсюда следует, что  $L_2$ -норма образа Фурье решения не будет возрастать, если

$$|q(\xi)| \leq 1. \quad (21.55)$$

При этом

$$\|\hat{u}\|_{L_2(0,2\pi)} \leq \|\tilde{u}\|_{L_2(0,2\pi)} \leq \dots \leq \|\tilde{u}^0\|_{L_2(0,2\pi)}.$$

Принимая теперь во внимание равенство Парсевала (21.46), приходим к (21.52).

Покажем теперь, что (21.55) следует из (21.51). Так как  $\sigma \geq 0$ , то знаменатель в (21.54) положителен, и всегда  $q \leq 1$ . Осталось проверить условие  $q \geq -1$ , которое эквивалентно условию

$$2 - (1 - 2\sigma) \frac{4\tau}{h^2} \sin^2 \frac{\xi}{2} \geq 0$$

или

$$1 - 2\sigma \leq \frac{h^2}{2\tau \sin^2 \frac{\xi}{2}}.$$

Но это условие будет выполнено, если

$$1 - 2\sigma \leq \min_{\xi} \frac{h^2}{2\tau \sin^2 \frac{\xi}{2}} = \frac{h^2}{2\tau},$$

что эквивалентно (21.51). Теорема доказана.

**Упражнение 21.2.** Рассмотреть неоднородное уравнение и установить оценку решения через правую часть.

## Разностные схемы для уравнения колебаний струны

### 22.1 Аппроксимация

Рассмотрим другой пример уравнения с частными производными — уравнение колебаний струны

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 1, \quad 0 < t < T. \quad (22.1)$$

Это — гиперболическое уравнение. Корректной для него является смешанная задача, например,

$$u(0, t) = u(1, t) = 0, \quad u(x, 0) = \bar{u}(x), \quad \frac{\partial u}{\partial t}(x, 0) = \bar{\dot{u}}(x). \quad (22.2)$$

Граничные условия при  $x = 0$  и  $x = 1$  предполагаются однородными граничными условиями первого рода, а в качестве начальных функций взяты некоторые функции  $\bar{u}(x)$  и  $\bar{\dot{u}}(x)$ .

Как и при построении аппроксимации уравнения теплопроводности, аппроксимируем сначала производную по пространственной переменной  $x$ . В результате получим задачу

$$\begin{aligned} \ddot{u}_i^h(t) &= u_{\bar{x}\bar{x},i}^h(t), \quad i = 1, \dots, N-1, \\ u_0^h(t) &= u_N^h(t) = 0, \\ u^h(x_i, 0) &= \bar{u}(x_i), \quad \dot{u}^h(x_i, 0) = \bar{\dot{u}}(x_i), \end{aligned} \quad (22.3)$$

которая представляет собой задачу Коши для системы  $(N-1)$  дифференциальных уравнений второго порядка. Теперь произведем аппроксимацию по временной переменной: производную  $\ddot{u}(t)$  заменим вторым

разностным отношением

$$u_{\bar{t}t}(t_j) \equiv [u(t_{j+1}) - 2u(t_j) + u(t_{j-1})]/\tau^2,$$

а правую часть (22.3) — линейной комбинацией ее значений при  $t = t_{j-1}$ ,  $t = t_j$  и  $t = t_{j+1}$ . В результате будем иметь

$$u_{\bar{t}t,i}^h = \sigma \hat{u}_{\bar{x}x,i}^h + (1 - 2\sigma)u_{\bar{x}x,i}^h + \sigma \check{u}_{\bar{x}x,i}^h, \quad i = 1, \dots, N - 1, \quad (22.4)$$

где наряду с уже введенным ранее обозначением  $\hat{v}_i = v_i(t_{j+1})$  принято обозначение  $\check{v}_i = v_i(t_{j-1})$ . Правая часть (22.4) представляет собой не общую линейную комбинацию, а линейную комбинацию, симметричную относительно  $t_{j-1}$  и  $t_{j+1}$ .

К уравнениям (22.4) нужно добавить граничные и начальные условия, которые должны аппроксимировать соответственно условия (22.3)

$$u_0^{hj} = u_N^{hj} = 0, \quad j = 0, \dots, J, \quad (22.5)$$

$$u_i^{h0} = \bar{u}(x_i), \quad i = 1, \dots, N - 1. \quad (22.6)$$

Второе из начальных условий (22.3) содержит производную. Аппроксимируя ее по двум точкам, получим

$$u_{t,i}^{h0} = \bar{u}(x_i), \quad i = 1, \dots, N - 1. \quad (22.7)$$

**Теорема 22.1.** *Если решение  $u(x, t)$  уравнения (22.1) обладает непрерывными четвертыми производными, то погрешность аппроксимации разностной схемы (22.4) есть  $O(\tau^2 + h^2)$ .*

**Доказательство.**

$$\begin{aligned} \Psi_i^j &= \sigma \hat{u}_{\bar{x}x,i}^j + (1 - 2\sigma)u_{\bar{x}x,i}^j + \sigma \check{u}_{\bar{x}x,i}^j - u_{\bar{t}t,i}^j = \sigma \tau^2 u_{\bar{x}x\bar{t}t,i}^j + u_{\bar{x}x,i}^j - u_{\bar{t}t,i}^j = \\ &= \frac{\partial^2 u}{\partial x^2} + O(h^2) - \frac{\partial^2 u}{\partial t^2} + O(\tau^2) = O(\tau^2 + h^2). \end{aligned}$$

Теорема доказана.

Найдем погрешность аппроксимации начального условия (22.7)

$$\begin{aligned} \psi_i &= -u_{t,i}^0 + \bar{u}(x_i) = -\frac{\partial u}{\partial t}(x_i, 0) - \frac{\tau}{2} \frac{\partial^2 u}{\partial t^2}(x_i, 0) + O(\tau^2) + \bar{u}(x_i) = \\ &= -\frac{\tau}{2} \frac{\partial^2 u}{\partial t^2}(x_i, 0) + O(\tau^2). \end{aligned} \quad (22.8)$$



Погрешность аппроксимации начального условия (22.7) есть  $O(\tau)$ . Построим другую аппроксимацию с погрешностью не хуже  $O(\tau^2 + h^2)$ . Для этого преобразуем (22.8). В силу (22.1)  $\frac{\partial^2 u}{\partial t^2}(x, 0) = \frac{\partial^2 u}{\partial x^2}(x, 0)$ , и поэтому

$$\overset{\circ}{\psi}_i = -\frac{\tau}{2} \frac{\partial^2 u}{\partial x^2}(x_i, 0) + O(\tau^2).$$

Принимая теперь во внимание (22.2), будем иметь

$$\overset{\circ}{\psi}_i = -\frac{\tau}{2} \bar{u}''(x_i) + O(\tau^2).$$

Отсюда и из (22.8) следует, что если вместо  $\bar{u}(x_i)$  в (22.7) положить  $\bar{u}(x_i) + \frac{\tau}{2} \bar{u}''(x_i)$ , т.е. написать условие

$$u_{t,i}^{h0} = \bar{u}_i + \frac{\tau}{2} \bar{u}_i'', \quad (22.9)$$

то погрешность этой аппроксимации будет  $O(\tau^2)$ . Очевидно также, что если вместо  $\bar{u}_i''$  в (22.9) подставить  $\bar{u}_{\bar{x}\bar{x},i}$ , т.е. взять

$$u_{t,i}^{h0} = \bar{u}_i + \frac{\tau}{2} \bar{u}_{\bar{x}\bar{x},i}, \quad (22.10)$$

то погрешность этой аппроксимации будет  $O(\tau^2 + h^2)$ .

Аппроксимация (22.10) всем хороша за исключением одного но. Именно, для аппроксимации уравнения (22.1) мы использовали однопараметрическое семейство разностных схем, среди которых имеются как явная ( $\sigma = 0$ ), так и неявные ( $\sigma \neq 0$ ). Аппроксимация же (22.10) всегда явная. Внесем параметр и в начальное условие. Пусть

$$u_{t,i}^{h0} = \bar{u}_i + \frac{\tau}{2} [\gamma u_{\bar{x}\bar{x},i}^{h1} + (1 - \gamma) u_{\bar{x}\bar{x},i}^{h0}].$$

Ясно, что погрешность этой аппроксимации снова не хуже  $O(\tau^2 + h^2)$ . Наконец, согласуем параметр  $\gamma$  с параметром  $\sigma$ , полагая  $\gamma/2 = \sigma$ . Аппроксимация второго начального условия (22.2) примет следующий окончательный вид

$$u_{t,i}^{h0} = \tau \sigma u_{\bar{x}\bar{x},i}^{h1} + \tau \left( \frac{1}{2} - \sigma \right) u_{\bar{x}\bar{x},i}^{h0} + \bar{u}_i. \quad (22.11)$$

## 22.2 Устойчивость по начальным данным

Исследуем вопрос об устойчивости схемы (22.4) по начальным данным. Ограничимся изучением задачи Коши, т.е. будем предполагать, что уравнения (22.4) и начальные условия (22.6) и (22.11) заданы для всех  $i \in \mathbb{Z}$ .

Именно, будем рассматривать следующую задачу

$$u_{tt,i}^h = \sigma \hat{u}_{\bar{x}x,i}^h + (1 - 2\sigma)u_{\bar{x}x,i}^h + \sigma \check{u}_{\bar{x}x,i}^h, \quad i \in \mathbb{Z}, \quad (22.12)$$

$$u_i^{h0} = \bar{u}_i, \quad u_{t,i}^{h0} = \tau \sigma u_{\bar{x}x,i}^{h1} + \tau \left( \frac{1}{2} - \sigma \right) u_{\bar{x}x,i}^{h0} + \bar{u}_i, \quad i \in \mathbb{Z}. \quad (22.13)$$

**Теорема 22.2.** Если параметр  $\sigma$  задачи (22.12), (22.13) неотрицателен и удовлетворяет условию

$$\sigma \geq \frac{1}{4} - \frac{h^2}{4\tau^2}, \quad (22.14)$$

то для решения этой задачи справедлива априорная оценка

$$\|u^{hj}\|_{L_2^h} \leq \|\bar{u}\|_{L_2^h} + T \|\bar{u}\|_{L_2^h}. \quad (22.15)$$

**Доказательство.** Сделаем сеточное преобразование Фурье (22.12), (22.13). Для образа Фурье  $\tilde{u}^j(\xi)$  решения  $u_i^{hj}$  получим задачу

$$\begin{aligned} \tilde{u}_{\bar{t}\bar{t}} + \frac{4 \sin^2 \frac{\xi}{2}}{h^2} [\sigma \hat{\tilde{u}} + (1 - 2\sigma)\tilde{u} + \sigma \check{\tilde{u}}] &= 0, \\ \tilde{u}^0 = \tilde{\bar{u}}, \quad \tilde{u}_t^0 + \frac{4\tau \sin^2 \frac{\xi}{2}}{h^2} \left[ \sigma \tilde{u}^1 + \left( \frac{1}{2} - \sigma \right) \tilde{u}^0 \right] &= \tilde{\bar{u}}. \end{aligned} \quad (22.16)$$

Умножим теперь уравнение (22.16) на  $\tau^2$  и перепишем в поточечном виде

$$\tilde{u}^{j+1} - 2\tilde{u}^j + \tilde{u}^{j-1} + \frac{4\tau^2}{h^2} \sin^2 \frac{\xi}{2} [\sigma \tilde{u}^{j+1} + (1 - 2\sigma)\tilde{u}^j + \sigma \tilde{u}^{j-1}] = 0.$$

Введем обозначение

$$\frac{2\tau}{h} \sin \frac{\xi}{2} = \lambda \quad (22.17)$$

и напишем характеристическое уравнение разностного уравнения

$$(1 + \sigma\lambda^2)q^2 - 2 \left( 1 + \left( \sigma - \frac{1}{2} \right) \lambda^2 \right) q + (1 + \sigma\lambda^2) = 0$$

или

$$q^2 - 2 \frac{1 + (\sigma - 1/2)\lambda^2}{1 + \sigma\lambda^2} q + 1 = 0.$$

Отсюда находим корни

$$\begin{aligned} q_{1,2} &= \frac{1 + (\sigma - 1/2)\lambda^2}{1 + \sigma\lambda^2} \pm \\ &\pm \frac{\sqrt{[1 + (\sigma - 1/2)\lambda^2 + 1 + \sigma\lambda^2][1 + (\sigma - 1/2)\lambda^2 - 1 - \sigma\lambda^2]}}{1 + \sigma\lambda^2} = \\ &= \frac{1 + (\sigma - 1/2)\lambda^2 \pm \sqrt{[1 + (\sigma - 1/4)\lambda^2](-\lambda^2)}}{1 + \sigma\lambda^2}. \end{aligned} \quad (22.18)$$

Если

$$1 + \left(\sigma - \frac{1}{4}\right) \lambda^2 > 0,$$

то корни  $q_1$  и  $q_2$  будут комплексными и равными по модулю 1. Если же

$$1 + \left(\sigma - \frac{1}{4}\right) \lambda^2 = 0,$$

то

$$q_{1,2} = \frac{1 + (\sigma - 1/2)\lambda^2}{1 + \sigma\lambda^2} = \frac{-1/4\lambda^2}{1/4\lambda^2} = -1$$

и снова  $|q_{1,2}| = 1$ .

Выясним, когда

$$1 + \left(\sigma - \frac{1}{4}\right) \lambda^2 \geq 0,$$

или, что то же самое,

$$(1/4 - \sigma) \leq \frac{1}{\lambda^2}.$$

Это неравенство будет выполнено при всех  $\xi \in (0, 2\pi]$ , если (см. (22.17))

$$\frac{1}{4} - \sigma \leq \min_{\xi} \frac{1}{\lambda^2} = \frac{h^2}{4\tau^2}.$$

Но это условие совпадает с условием (22.14) теоремы 22.2, и, следовательно,  $|q_{1,2}| = 1$ .

Введем обозначение

$$q_{1,2} = e^{\pm i\varphi(\xi)} = \cos \varphi \pm i \sin \varphi,$$

где, согласно (22.18),

$$\cos \varphi = \frac{1 + (\sigma - 1/2)\lambda^2}{1 + \sigma\lambda^2}, \quad \sin \varphi = \frac{|\lambda| \sqrt{1 + (\sigma - 1/4)\lambda^2}}{1 + \sigma\lambda^2}, \quad (22.19)$$

и найдем решение задачи (22.16). Общее решение разностного уравнения (22.16) есть

$$\tilde{u}^j = c_1 \cos j\varphi + c_2 \sin j\varphi.$$

При  $j = 0$

$$\tilde{u}^0 = c_1 = \tilde{u},$$

а при  $j = 1$

$$\tilde{u}^1 = c_1 \cos \varphi + c_2 \sin \varphi.$$

Из вышесказанного следует, что

$$c_2 = \frac{\tilde{u}^1 - \tilde{\tilde{u}} \cos \varphi}{\sin \varphi}. \quad (22.20)$$

Далее, из второго начального условия (22.16)

$$\tilde{u}^1(1 + \sigma\lambda^2) = \left(1 + \left(\sigma - \frac{1}{2}\right)\lambda^2\right) \tilde{u}^0 + \tau\tilde{\tilde{u}},$$

и, следовательно,

$$\tilde{u}^1 = \frac{1 + (\sigma - 1/2)\lambda^2}{1 + \sigma\lambda^2} \tilde{u}^0 + \frac{\tau\tilde{\tilde{u}}}{1 + \sigma\lambda^2},$$

а с учетом (22.19)

$$\tilde{u}^1 = \cos \varphi \tilde{\tilde{u}} + \frac{\tau\tilde{\tilde{u}}}{1 + \sigma\lambda^2}.$$

Подставляя это значение  $\tilde{u}^1$  в (22.20), найдем, что

$$c_2 = \frac{\tau\tilde{\tilde{u}}}{(1 + \sigma\lambda^2) \sin \varphi}.$$

Окончательно для решения задачи (22.16) получаем представление

$$\tilde{u}^j = \tilde{\tilde{u}} \cos j\varphi + \tau \frac{\tilde{\tilde{u}}}{1 + \sigma\lambda^2} \frac{\sin j\varphi}{\sin \varphi}. \quad (22.21)$$

Чтобы оценить правую часть (22.21), нам потребуется

**Лемма 22.1.** При  $n \in \mathbb{N}$

$$|\sin(n\varphi)/\sin \varphi| \leq n.$$

**Доказательство.** Имеем

$$\begin{aligned} & \left| \frac{e^{in\varphi} - e^{-in\varphi}}{e^{i\varphi} - e^{-i\varphi}} \right| = \left| \frac{e^{2in\varphi} - 1}{e^{2i\varphi} - 1} \frac{e^{i\varphi}}{e^{in\varphi}} \right| = \left| \frac{e^{2in\varphi} - 1}{e^{2i\varphi} - 1} \right| \\ & = \left| e^{2i(n-1)\varphi} + e^{2i(n-2)\varphi} + \dots + 1 \right| \leq n. \end{aligned}$$

Лемма доказана.

Используя лемму 22.1, из (22.21) находим, что

$$|\tilde{u}^j| \leq |\tilde{\tilde{u}}| + \tau j |\tilde{\tilde{u}}| \leq |\tilde{\tilde{u}}| + T |\tilde{\tilde{u}}|.$$

Отсюда

$$\|\tilde{u}^j\|_{L_2(0,2\pi)} \leq \|\tilde{u}\|_{L_2(0,2\pi)} + T\|\tilde{\bar{u}}\|_{L_2(0,2\pi)},$$

а с учетом равенства Парсеваля

$$\|u^j\|_{L_2^h} \leq \|\bar{u}\|_{L_2^h} + T\|\bar{\bar{u}}\|_{L_2^h}.$$

Теорема доказана.

**Следствие 3.** При  $\sigma \geq 1/4$  условие (22.14) выполнено, и оценка (22.15) решения имеет место для любых  $h$  и  $\tau$ . При  $\sigma = 0$  (явная схема) условие (22.14) выполнено, если  $\tau \leq h$ , и поэтому оценка (22.15) имеет место только при указанном соотношении между  $\tau$  и  $h$ .

# Литература

1. А.А. Амосов, Ю.А. Дубинский, Н.В. Копченова. Вычислительные методы для инженеров. М.: Высшая школа. 1991.
2. Н.С.Бахвалов, Н.П.Жидков, Г.М.Кобельков. Численные методы. М.: Наука. 1989.
3. В.М.Вербжицкий. Основы численных методов. М.: Высшая школа. 2002.
4. Дж. Голуб, Ч. Ван Лоун. Матричные вычисления. М.: Мир. 1999.
5. Дж. Деммель. Вычислительная линейная алгебра. Теория и приложения. М.: Мир. 2001.
6. Х.Д.Икрамов. Численные методы линейной алгебры. М.: Знание. 1987 (Новое в жизни, науке, технике. Сер. Математика, кибернетика. N 4/1987).
7. Д. Каханер, К. Моулер, С. Нэш. Численные методы и программное обеспечение. М.: Мир. 2001.
8. А.А.Самарский, А.В.Гулин. Численные методы. М.: Наука. 1989.
9. А.А.Самарский, Николаев Е.С. Методы решения сеточных уравнений. М.: Наука. 1978.
10. Современные численные методы решения обыкновенных дифференциальных уравнений. Ред. Дж. Холл и Дж. Уатт. М.: Мир. 1979.
11. Е.Е. Тыртышников. Методы численного анализа. М.: Академия. 2007.
12. Д. Уоткинс. Основы матричных вычислений. М.: БИНОМ. Лаборатория знаний. 2006.
13. Э.Хайрер, С.Нёрсетт, Г.Ваннер. Решение обыкновенных дифференциальных уравнений. Нежесткие задачи. М.: Мир. 1990.

14. A. Quarteroni, R. Sacco, E. Saleri. Numerical Mathematics. Springer-Verlag. 2000.
15. L.N. Trefethen, D. Bau, III. Numerical linear algebra. Philadelphia SIAM. 1997.

# Предметный указатель

- Абсолютная устойчивость 246
- Априорные оценки 279
- Ведущая подматрица 7
- Верхняя форма Хессенберга 115
- Гибридные алгоритмы 187
- Жесткие задачи 242
- Итерации
  - обратные 110
  - с отношением Рэлея 110
- Каноническая форма
  - Жордана 102
  - Шура 102
  - — вещественная 102
- Метод
  - баланса 265
  - бисекции 170
  - блочного исключения 47
  - Бройдена 193
  - вращений Гивенса 21
  - Гаусса 5
  - — с выбором главного элемента 60
  - —, обратный ход 8
  - —, прямой ход 7
  - наискорейшего спуска 78
  - неопределенных коэффициентов 160
  - Ньютона 174, 191
  - отражений Хаусхолдера 27
  - последовательного исключения 5
  - прогонки 34
  - простых итераций 171
  - Рунге 200
  - Рунге-Кутты  $s$ -этапный 206
  - секущих 180, 192
  - сопряженных градиентов 84
  - степенной 104
  - трапеций 201
  - Хойна 218
  - Холецкого 15
  - Чебышевский итерационный 71
  - Эйлера 199
  - — неявный 200
- Методы
  - Адамса 216
  - Рунге-Кутты 204
  - — диагонально неявные 207
  - — однократно неявные 207
  - линейные многошаговые 226
- Матрица
  - вращения 23
  - отражения 27
  - Хаусхолдера 27
  - ленточная 35
  - трехдиагональная 36
  - перестановок 60
  - , число обусловленности 52
  - плохо обусловленная 53
- Матрицы подобные 101
- Многочлен
  - интерполяционный Лагранжа 162
  - — Ньютона 165
  - — Эрмита 165
  - Чебышева 73
- Многочлены
  - ортогональные 146



— Чебышева 152

— Лежандра 155

Множитель Холецкого 15

Монотонная схема Самарского 292

Невязка 79

Неполное разложение 97

Норма 51

— подчиненная 51

Область абсолютной устойчивости 246

Обратная подстановка 8

Отношение Релея 103

Сетка

— неравномерная 276

— квазиравномерная 277

— Шишкина 294

Сеточное преобразование Фурье 312

Таблица Бутчера 206

Уравнение

— сингулярно возмущенное 287

— разностное 133

— характеристическое 138

— — системы 142

Угловой минор 13

Формула

— центральных прямоугольников 199

— левых прямоугольников 199

— правых прямоугольников 199

— трапеций 201

Формулы дифференцирования назад 245, 246

Функция Грина 230, 231

$A$ -устойчивость 245

$A(\alpha)$ -устойчивость 250

$LU$  разложение 8

$LDL^T$ -разложением 19

$QR$ -алгоритм 114

$QR$ -факторизации 20