Б. П. ДЕМИДОВИЧ и И. А. МАРОН

# Основы вычислительной математики

Б. П. ДЕМИДОВИЧ и И. А. МАРОН

# ОСНОВЫ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ

ИЗДАНИЕ ТРЕТЬЕ,
ИСПРАВЛЕННОЕ

*Допущено Министерством
высшего и среднего специального образования РСФСР
в качестве учебного пособия
для высших технических учебных заведений*

ИЗДАТЕЛЬСТВО «НАУКА»
ГЛАВНАЯ РЕДАКЦИЯ
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ
МОСКВА 1966

518
Д 30
УДК 518. 0

## АННОТАЦИЯ

Книга посвящена изложению важнейших методов и приемов вычислительной математики на базе общего втузовского курса высшей математики. Основная часть книги является учебным пособием по курсу приближенных вычислений для втузов. Книга может быть полезна также для лиц, работающих в области прикладной математики.

*Борис Павлович Демидович* и *Исаак Абрамович Марон*
Основы вычислительной математики

М., 1966 г. 664 стр. с илл.

Редакторы *М. М. Горячая* и *В. М. Гринберг*

Техн. редактор *К. Ф. Брудко* Корректор *С. Д. Кайсер*

Сдано в набор 30/III 1966 г. Подписано к печати 28/VI 1966 г. Бумага 60×90/16
Физ. печ. л. 41,5 Условн. печ. л. 41,5 Уч.-изд. л. 41,06
Тираж 75 000 экз. Т-08270. Цена книги 1 р. 54 к. Заказ № 339.

Издательство «Наука»
Главная редакция
Физико-математической литературы
Москва, В-71, Ленинский проспект, 15

Первая Образцовая типография имени А. А. Жданова
Главполиграфпрома Комитета по печати при Совете Министров СССР
Москва, Ж-54, Валовая, 28

2-2-3
4-66

# ОГЛАВЛЕНИЕ

Предисловие к первому изданию . . . . . 9
Предисловие ко второму изданию . . . . . 12
Предисловие к третьему изданию . . . . . 12

Введение. Общие правила вычислительной работы . . . . . 13

Глава I. **Приближенные числа** . . . . . 17

§ 1. Абсолютная и относительная погрешности . . . . . 17
§ 2. Основные источники погрешностей . . . . . 20
§ 3. Десятичная запись приближенных чисел. Значащая цифра. Число верных знаков . . . . . 21
§ 4. Округление чисел . . . . . 24
§ 5. Связь относительной погрешности приближенного числа с количеством верных знаков этого числа . . . . . 25
§ 6. Таблицы для определения предельной относительной погрешности по числу верных знаков и наоборот . . . . . 28
§ 7. Погрешность суммы . . . . . 31
§ 8. Погрешность разности . . . . . 33
§ 9. Погрешность произведения . . . . . 35
§ 10. Число верных знаков произведения . . . . . 37
§ 11. Погрешность частного . . . . . 38
§ 12. Число верных знаков частного . . . . . 39
§ 13. Относительная погрешность степени . . . . . 39
§ 14. Относительная погрешность корня . . . . . 39
§ 15. Вычисления без точного учета погрешностей . . . . . 40
§ 16. Общая формула для погрешности . . . . . 41
§ 17. Обратная задача теории погрешностей . . . . . 43
§ 18. Точность определения аргумента для функции, заданной таблицей . . . . . 46
§ 19. Способ границ . . . . . 48
§ 20*. Понятие о вероятностной оценке погрешности . . . . . 51
Литература к первой главе . . . . . 52

Глава II. **Некоторые сведения из теории цепных дробей** . . . . . 53

§ 1. Определение цепной дроби . . . . . 53
§ 2. Обращение цепной дроби в обыкновенную и обратно . . . . . 54
§ 3. Подходящие дроби . . . . . 56
§ 4. Бесконечные цепные дроби . . . . . 64
§ 5. Разложение функций в цепные дроби . . . . . 70
Литература ко второй главе . . . . . 73

Глава III. **Вычисление значений функций** . . . . . 74

§ 1. Вычисление значений полинома. Схема Горнера . . . . . 74
§ 2. Обобщенная схема Горнера . . . . . 77
§ 3. Вычисление значений рациональных дробей . . . . . 79

§ 4. Приближенное нахождение сумм числовых рядов . . . . . 80
§ 5. Вычисление значений аналитической функции . . . . . 86
§ 6. Вычисление значений показательной функции . . . . . 88
§ 7. Вычисление значений логарифмической функции . . . . . 92
§ 8. Вычисление значений тригонометрических функций . . . . . 95
§ 9. Вычисление значений гиперболических функций . . . . . 98
§ 10. Применение метода итерации для приближенного вычисления значений функции . . . . . 100
§ 11. Вычисление обратной величины . . . . . 101
§ 12. Вычисление квадратного корня . . . . . 104
§ 13. Вычисление обратной величины квадратного корня . . . . . 108
§ 14. Вычисление кубического корня . . . . . 108
Литература к третьей главе . . . . . 111

Глава IV. **Приближенное решение алгебраических и трансцендентных уравнений** . . . . . 112

§ 1. Отделение корней . . . . . 112
§ 2. Графическое решение уравнений . . . . . 116
§ 3. Метод половинного деления . . . . . 118
§ 4. Способ пропорциональных частей (метод хорд) . . . . . 119
§ 5. Метод Ньютона (метод касательных) . . . . . 123
§ 6. Видоизмененный метод Ньютона . . . . . 131
§ 7. Комбинированный метод . . . . . 132
§ 8. Метод итерации . . . . . 135
§ 9. Метод итерации для системы двух уравнений . . . . . 148
§ 10. Метод Ньютона для системы двух уравнений . . . . . 152
§ 11. Метод Ньютона для случая комплексных корней . . . . . 153
Литература к четвертой главе . . . . . 157

Глава V. **Специальные приемы для приближенного решения алгебраических уравнений** . . . . . 158

§ 1. Общие свойства алгебраических уравнений . . . . . 158
§ 2. Границы действительных корней алгебраических уравнений . 163
§ 3. Метод знакопеременных сумм . . . . . 165
§ 4. Метод Ньютона . . . . . 167
§ 5. Число действительных корней полинома . . . . . 169
§ 6. Теорема Бюдана—Фурье . . . . . 171
§ 7. Идея метода Лобачевского—Греффе . . . . . 176
§ 8. Процесс квадрирования корней . . . . . 178
§ 9. Метод Лобачевского—Греффе для случая действительных различных корней . . . . . 180
§ 10. Метод Лобачевского—Греффе для случая комплексных корней 183
§ 11. Случай пары комплексных корней . . . . . 186
§ 12. Случай двух пар комплексных корней . . . . . 190
§ 13. Метод Бернулли . . . . . 195
Литература к пятой главе . . . . . 198

Глава VI. **Улучшение сходимости рядов** . . . . . 199

§ 1. Улучшение сходимости числовых рядов . . . . . 199
§ 2. Улучшение сходимости степенных рядов методом Эйлера—Абеля . . . . . 205
§ 3. Оценки коэффициентов Фурье . . . . . 210
§ 4. Улучшение сходимости тригонометрических рядов Фурье методом А. Н. Крылова . . . . . 213
§ 5. Приближенное суммирование тригонометрических рядов . . . 222
Литература к шестой главе . . . . . 224

Глава VII. **Алгебра матриц** . . . . . 225

§ 1. Основные определения . . . . . 225
§ 2. Действия с матрицами . . . . . 226
§ 3. Транспонированная матрица . . . . . 230
§ 4. Обратная матрица . . . . . 231
§ 5. Степени матрицы . . . . . 236
§ 6. Рациональные функции матрицы . . . . . 237
§ 7. Абсолютная величина и норма матрицы . . . . . 238
§ 8. Ранг матрицы . . . . . 244
§ 9. Предел матрицы . . . . . 245
§ 10. Матричные ряды . . . . . 247
§ 11. Клеточные матрицы . . . . . 252
§ 12. Обращение матриц при помощи разбиения на клетки . . . . 255
§ 13. Треугольные матрицы . . . . . 260
§ 14. Элементарные преобразования матриц . . . . . 263
§ 15. Вычисление определителей . . . . . 264
Литература к седьмой главе . . . . . 267

Глава VIII. **Решение систем линейных уравнений** . . . . . 268

§ 1. Общая характеристика методов решения систем линейных уравнений . . . . . 268
§ 2. Решение систем с помощью обратной матрицы. Формулы Крамера . . . . . 268
§ 3. Метод Гаусса . . . . . 272
§ 4. Уточнение корней . . . . . 279
§ 5. Метод главных элементов . . . . . 281
§ 6. Применение метода Гаусса для вычисления определителей . . 283
§ 7. Вычисление обратной матрицы методом Гаусса . . . . . 285
§ 8. Метод квадратных корней . . . . . 287
§ 9. Схема Халецкого . . . . . 290
§ 10. Метод итерации . . . . . 294
§ 11. Приведение линейной системы к виду, удобному для итерации 301
§ 12. Метод Зейделя . . . . . 303
§ 13. Случай нормальной системы . . . . . 305
§ 14. Метод релаксации . . . . . 307
§ 15. Исправление элементов приближенной обратной матрицы . . . 310
Литература к восьмой главе . . . . . 314

Глава IX*. **Сходимость итерационных процессов для систем линейных уравнений** . . . . . 315

§ 1. Достаточные условия сходимости процесса итерации . . . . . 315
§ 2. Оценка погрешности приближений процесса итерации . . . . 317
§ 3. Первое достаточное условие сходимости процесса Зейделя . . 320
§ 4. Оценка погрешности приближений процесса Зейделя по $m$-норме . . . . . 322
§ 5. Второе достаточное условие сходимости процесса Зейделя . . 323
§ 6. Оценка погрешности приближений процесса Зейделя по $l$-норме . . . . . 325
§ 7. Третье достаточное условие сходимости процесса Зейделя . . 326
Литература к девятой главе . . . . . 328

Глава X. **Основные сведения из теории линейных векторных пространств** . . . . . 329

§ 1. Понятие линейного векторного пространства . . . . . 329
§ 2. Линейная зависимость векторов . . . . . 330

§ 3. Скалярное произведение векторов . . . . . 335
§ 4. Ортогональные системы векторов . . . . . 338
§ 5. Преобразования координат вектора при изменениях базиса . . 340
§ 6. Ортогональные матрицы . . . . . 342
§ 7. Ортогонализация матриц . . . . . 343
§ 8. Применение методов ортогонализации к решению систем линейных уравнений . . . . . 351
§ 9. Пространство решений однородной системы . . . . . 356
§ 10. Линейные преобразования переменных . . . . . 359
§ 11. Обратное преобразование . . . . . 365
§ 12. Собственные векторы и собственные значения матрицы . . . . 367
§ 13. Подобные матрицы . . . . . 372
§ 14. Билинейная форма матрицы . . . . . 375
§ 15. Свойства симметрических матриц . . . . . 376
§ 16*. Свойства матриц с действительными элементами . . . . . 381
Литература к десятой главе . . . . . 385

Глава XI*. **Дополнительные сведения о сходимости итерационных процессов для систем линейных уравнений** . . . . . 386

§ 1. Сходимость матричных степенных рядов . . . . . 386
§ 2. Тождество Гамильтона—Кели . . . . . 389
§ 3. Необходимые и достаточные условия сходимости процесса итерации для системы линейных уравнений . . . . . 390
§ 4. Необходимые и достаточные условия сходимости процесса Зейделя для системы линейных уравнений . . . . . 392
§ 5. Сходимость процесса Зейделя для нормальной системы . . . 395
§ 6. Способы эффективной проверки условий сходимости . . . . . 397
Литература к одиннадцатой главе . . . . . 401

Глава XII. **Нахождение собственных значений и собственных векторов матрицы** . . . . . 402

§ 1. Вводные замечания . . . . . 402
§ 2. Развертывание вековых определителей . . . . . 402
§ 3. Метод А. М. Данилевского . . . . . 404
§ 4. Исключительные случаи в методе А. М. Данилевского . . . 410
§ 5. Вычисление собственных векторов по методу А. М. Данилевского . . . . . 411
§ 6. Метод А. Н. Крылова . . . . . 412
§ 7. Вычисление собственных векторов по методу А. Н. Крылова 416
§ 8. Метод Леверрье . . . . . 417
§ 9. Понятие о методе неопределенных коэффицентов . . . . . 419
§ 10. Сравнение различных методов развертывания векового определителя . . . . . 421
§ 11. Нахождение наибольшего по модулю собственного значения матрицы и соответствующего собственного вектора . . . . . 421
§ 12. Метод скалярных произведений для нахождения первого собственного значения действительной матрицы . . . . . 428
§ 13. Нахождение второго собственного значения матрицы и второго собственного вектора . . . . . 431
§ 14. Метод исчерпывания . . . . . 434
§ 15. Нахождение собственных элементов положительно определенной симметрической матрицы . . . . . 437
§ 16. Использование коэффициентов характеристического полинома матрицы для ее обращения . . . . . 442
§ 17. Метод Л. А. Люстерника улучшения сходимости процесса итерации для решения системы линейных уравнений . . . . . 444
Литература к двенадцатой главе . . . . . 449

Глава XIII. **Приближенное решение систем нелинейных уравнений** . . . . . 450

§ 1. Метод Ньютона . . . . . 450
§ 2. Общие замечания о сходимости процесса Ньютона . . . . . 456
§ 3*. Существование корней системы и сходимость процесса Ньютона . . . . . 460
§ 4*. Быстрота сходимости процесса Ньютона . . . . . 465
§ 5*. Единственность решения . . . . . 466
§ 6*. Устойчивость сходимости процесса Ньютона при варьировании начального приближения . . . . . 469
§ 7. Модифицированный метод Ньютона . . . . . 471
§ 8. Метод итерации . . . . . 474
9*. Понятие о сжимающем отображении . . . . . 477
§ 10*. Первое достаточное условие сходимости процесса итерации . . . . . 481
§ 11*. Второе достаточное условие сходимости процесса итерации 483
§ 12. Метод скорейшего спуска (метод градиента) . . . . . 485
§ 13. Метод скорейшего спуска для случая системы линейных уравнений . . . . . 490
§ 14*. Метод степенных рядов . . . . . 494
Литература к тринадцатой главе . . . . . 496

Глава XIV. **Интерполирование функций** . . . . . 497

§ 1. Конечные разности различных порядков . . . . . 497
§ 2. Таблица разностей . . . . . 500
§ 3. Обобщенная степень . . . . . 505
§ 4. Постановка задачи интерполирования . . . . . 507
§ 5. Первая интерполяционная формула Ньютона . . . . . 508
§ 6. Вторая интерполяционная формула Ньютона . . . . . 514
§ 7. Таблица центральных разностей . . . . . 518
§ 8. Интерполяционные формулы Гаусса . . . . . 519
§ 9. Интерполяционная формула Стирлинга . . . . . 521
§ 10. Интерполяционная формула Бесселя . . . . . 521
§ 11. Общая характеристика интерполяционных формул с постоянным шагом . . . . . 524
§ 12. Интерполяционная формула Лагранжа . . . . . 527
§ 13*. Вычисление лагранжевых коэффициентов . . . . . 531
§ 14. Оценка погрешности интерполяционной формулы Лагранжа 535
§ 15. Оценки погрешностей интерполяционных формул Ньютона 537
§ 16. Оценки погрешностей центральных интерполяционных формул 539
§ 17. О наилучшем выборе узлов интерполирования . . . . . 540
§ 18. Разделенные разности . . . . . 542
§ 19. Интерполяционная формула Ньютона для неравноотстоящих значений аргумента . . . . . 544
§ 20. Обратное интерполирование для случая равноотстоящих узлов 547
§ 21. Обратное интерполирование для случая неравноотстоящих узлов . . . . . 550
§ 22. Нахождение корней уравнения методом обратного интерполирования . . . . . 551
§ 23. Метод интерполяции для развертывания векового определителя . . . . . 553
§ 24*. Интерполирование функций двух переменных . . . . . 555
§ 25*. Двойные разности высших порядков . . . . . 557
§ 26*. Интерполяционная формула Ньютона для функции двух переменных . . . . . 558
Литература к четырнадцатой главе . . . . . 561

Глава XV. **Приближенное дифференцирование** . . . . . 562

§ 1. Постановка вопроса . . . . . 562
§ 2. Формулы приближенного дифференцирования, основанные на первой интерполяционной формуле Ньютона . . . . . 563
§ 3. Формулы приближенного дифференцирования, основанные на формуле Стирлинга . . . . . 567
§ 4. Формулы численного дифференцирования для равноотстоящих точек, выраженные через значения функции в этих точках . . 571
§ 5. Графическое дифференцирование . . . . . 574
§ 6*. Понятие о приближенном вычислении частных производных 575
Литература к пятнадцатой главе . . . . . 576

Глава XVI. **Приближенное интегрирование функций** . . . . . 577
§ 1. Общие замечания . . . . . 577
§ 2. Квадратурные формулы Ньютона—Котеса . . . . . 580
§ 3. Формула трапеций и ее остаточный член . . . . . 582
§ 4. Формула Симпсона и ее остаточный член . . . . . 583
§ 5. Формулы Ньютона—Котеса высших порядков . . . . . 586
§ 6. Общая формула трапеций (правило трапеций) . . . . . 588
§ 7. Общая формула Симпсона (параболическая формула) . . . . 589
§ 8. Понятие о квадратурной формуле Чебышева . . . . . 593
§ 9. Квадратурная формула Гаусса . . . . . 597
§ 10. Некоторые замечания о точности квадратурных формул . . . 604
§ 11*. Экстраполяция по Ричардсону . . . . . 607
§ 12*. Числа Бернулли . . . . . 611
§ 13*. Формула Эйлера—Маклорена . . . . . 613
§ 14. Приближенное вычисление несобственных интегралов . . . . 618
§ 15. Метод Л. В. Канторовича выделения особенностей . . . . . 621
§ 16. Графическое интегрирование . . . . . 624
§ 17*. Понятие о кубатурных формулах . . . . . 627
§ 18*. Кубатурная формула типа Симпсона . . . . . 629
Литература к шестнадцатой главе . . . . . 633

Глава XVII. **Метод Монте-Карло** . . . . . 634

§ 1. Идея метода Монте-Карло . . . . . 634
§ 2. Случайные числа . . . . . 635
§ 3. Способы получения случайных чисел . . . . . 638
§ 4. Вычисление кратных интегралов методом Монте-Карло . . . . 641
§ 5*. Решение систем линейных алгебраических уравнений методом Монте-Карло . . . . . 650
Литература к семнадцатой главе . . . . . 658

Предметный указатель . . . . . 659

## ПРЕДИСЛОВИЕ К ПЕРВОМУ ИЗДАНИЮ

Бурное развитие новейшей техники и все большее внедрение современных разделов математики в инженерные исследования неизмеримо повысили требования к математической подготовке инженеров и научных работников, занимающихся прикладными вопросами.

Математическое образование инженера-исследователя в настоящее время не может ограничиться традиционными разделами так называемого «классического анализа», сложившегося, в основных своих направлениях, к началу нашего века. От инженера, работающего в научно-исследовательском институте, требуется теперь знание многих разделов современной математики и в первую очередь основательное владение методами и приемами вычислительной математики, так как решение почти каждой инженерной задачи должно быть доведено до численного результата.

Вычислительная техника наших дней представляет новые мощные средства для фактического выполнения счетной работы. Благодаря этому во многих случаях стало возможным отказаться от приближенной трактовки прикладных вопросов и перейти к решению задач в точной постановке. Это предполагает использование более глубоких специальных разделов математики (нелинейные дифференциальные уравнения, функциональный анализ, теоретико-вероятностные методы и др.).

Разумное использование современной вычислительной техники не мыслимо без умелого применения методов приближенного и численного анализа. Этим и объясняется чрезвычайно возросший как у нас, так и за рубежом интерес к методам вычислительной математики.

В нашей стране было издано несколько оригинальных и переводных книг, посвященных приближенным и численным методам.

Однако это не удовлетворяет в полной мере потребности читателей, так как многие из этих книг стали библиографической редкостью, а часть из них устарела или носит слишком специальный характер.

Основное назначение настоящей книги — дать в известной мере систематическое и современное изложение важнейших методов и приемов вычислительной математики на базе общего втузовского

курса высшей математики. Книга составлена так, что основная часть ее представляет собой учебное пособие по первому концентру приближенных вычислений для высших технических учебных заведений.

Многие институты нашей страны приступили к подготовке специалистов для работы в вычислительных центрах. Большие разделы приближенного и численного анализа включены в программу аспирантской подготовки по ряду специальностей и в программы различных курсов усовершенствования инженеров. Поэтому в книгу включен дополнительный материал, выходящий за рамки обычного втузовского курса. Это обстоятельство не затруднит пользование книгой: читатель без ущерба для понимания выберет нужные ему разделы и опустит лишние. Для удобства пользования книгой главы и параграфы, необязательные при первом чтении, отмечены звездочкой.

В книге широко используются основы матричного исчисления. Понятие вектора, матрицы, обратной матрицы, собственного значения и собственного вектора матрицы и т. п. являются рабочими. Применение матриц дает ряд преимуществ при изложении, так как, пользуясь ими, легче удается выяснить закономерность многих расчетов. Особенно выигрышным в этом смысле является проведение доказательств теорем сходимости различных численных процессов. Кроме того, современные быстродействующие вычислительные машины легко осуществляют основные матричные операции.

Для полного понимания содержания книги от читателя требуется известный минимум сведений по линейной алгебре и теории линейных векторных пространств. Чтобы облегчить усвоение этого минимума и избежать отсылки к многим источникам, в книге приведен весь необходимый дополнительный материал. Соответствующие главы независимы от основного текста и могут быть опущены подготовленным читателем.

Вкратце остановимся на содержании книги. Книга в основном посвящена следующим вопросам: действия с приближенными числами, вычисление значений функций при помощи рядов и итеративных процессов, приближенное и численное решение алгебраических и трансцендентных уравнений, вычислительные методы линейной алгебры, интерполирование функций, численное дифференцирование и интегрирование функций, метод Монте-Карло.

Большое внимание обращено на удобные способы оценки погрешностей. Почти для всех процессов даются доказательства теорем сходимости, причем изложение построено так, что при желании можно их опустить и ограничиться лишь технической стороной дела. В отдельных случаях, в целях наглядности изложения и устранения излишней громоздкости, вычислительные приемы сообщаются рецептурно.

Основные методы доведены до численных приложений — даны расчетные схемы и приведены числовые примеры с подробным ходом решения. В целях лучшего понимания сути дела большинство

приведенных примеров рассматривается в упрощенной трактовке и носит иллюстративный характер. Использованная и дополнительная литература указана по главам.

Настоящая книга излагает избранные методы вычислительной математики, и в нее не включен материал, связанный с эмпирическими формулами, квадратичным аппроксимированием функций, приближенным решением дифференциальных уравнений и др. Авторы намерены посвятить этим вопросам отдельную книгу.

В книгу также не включены сведения о программировании и технике решения математических задач на счетных машинах; по этому вопросу следует обратиться к специальным руководствам.

Авторы приносят благодарность коллективу кафедры высшей математики Артиллерийской инженерной академии им. Ф. Э. Дзержинского, принимавшему участие в обсуждении рукописи книги. Особую признательность выражаем Л. А. Люстернику, Г. П. Толстову и Н. П. Бусленко, сделавшим ряд замечаний общего характера, Э. З. Шуваловой, представившей некоторые письменные материалы, Д. М. Гробману за ценные практические советы и А. А. Юшкевичу, прорецензировавшему главу XVII.

Авторы благодарны также проф. Х. Л. Смолицкому и доц. С. В. Фролову и Р. Я. Шостаку, рецензии которых позволили улучшить качество рукописи.

Считаем своим долгом отметить компетентную работу редактора Г. И. Бирюк.

Москва, 1959 г. *Авторы*

## ПРЕДИСЛОВИЕ КО ВТОРОМУ ИЗДАНИЮ

Второе издание книги печатается с незначительными изменениями по сравнению с первым. Внесены исправления замеченных ошибок. В конце некоторых глав добавлены отдельные фрагменты. Так, в введении даны дополнительные указания к общим правилам вычислительной работы. Сделаны также некоторые замечания к методу Бернулли и др.

В 1962 г. Физматгизом была выпущена книга Б. П. Демидович, И. А. Марон, Э. З. Шувалова «Численные методы анализа», посвященная приближению функций и дифференциальным уравнениям и являющаяся естественным продолжением настоящей книги. Обе эти книги вместе составляют двухтомное учебное пособие по вычислительной математике, охватывающее основные методы численного решения важнейших математических задач. Это учебное пособие содержит материал, достаточный для большого курса приближенных вычислений во втузах, и может быть использовано также студентами физико-математических факультетов, специализирующихся в области вычислительной математики.

Москва, 1963 г. *Авторы*

## ПРЕДИСЛОВИЕ К ТРЕТЬЕМУ ИЗДАНИЮ

Третье издание отличается от второго лишь тем, что исправлены замеченные опечатки.

Авторы считают приятным долгом выразить свою благодарность Л. Я. Цлафу за ряд ценных указаний.

Москва, 1965 г. *Авторы*

# ВВЕДЕНИЕ

## Общие правила вычислительной работы

При выполнении массовых вычислений важно придерживаться определенных простых правил, выработанных практикой, соблюдение которых экономит труд вычислителя и позволяет рационально использовать имеющуюся вычислительную технику и вспомогательные средства.

Прежде всего вычислитель должен разработать подробную *вычислительную схему,* точно указывающую порядок действий и дающую возможность получить искомый результат наиболее простым и быстрым путем. Это особенно необходимо при однотипных вычислениях, так как такая схема, автоматизируя вычисления, позволяет выполнять их более быстро и надежно, что с пользой окупает время, затраченное на составление схемы. Кроме того, имея детальную вычислительную схему для решения задачи, можно использовать труд менее квалифицированных вычислителей.

Составление вычислительной схемы проиллюстрируем на следующем примере. Пусть требуется вычислить значения данной аналитической функции

$$y = f(x)$$

для заданных значений аргумента $x = x_1, x_2, \ldots, x_n$. Если число этих значений велико, то неразумно вычислять отдельно сначала значение $f(x_1)$, затем значение $f(x_2)$ и т. д., каждый раз выполняя всю совокупность операций, указанных символом $f$. Гораздо целесообразнее, расчленив функцию $f$ на элементарные операции (*действия*)

$$f(x) = f_m(\ldots(f_2(f_1(x)))\ldots),$$

вычисления производить однотипными операциями:

$$\begin{array}{ll} u_i = f_1(x_i) & (i = 1, 2, \ldots, n); \\ v_i = f_2(u_i) & (i = 1, 2, \ldots, n); \\ \ldots\ldots\ldots\ldots & \\ y = f_m(w_i) & (i = 1, 2, \ldots, n), \end{array}$$

выполняя одну и ту же операцию $f_j$ ($j=1, 2, \ldots, m$) для всех рассматриваемых значений аргумента. При этом широко могут быть использованы соответствующие таблицы функций и специализированные счетные машины. Запись результатов вычислений следует производить на особых вычислительных *бланках* или *формулярах*, представляющих собой специальным образом разграфленные и размеченные листы бумаги (применительно к выбранной вычислительной схеме!). На этих бланках, в строго определенных местах, заносятся промежуточные результаты по мере их получения, а также окончательные результаты.

Вычислительные бланки обычно строятся таким образом, чтобы результаты каждой серии однотипных операций заносились в один столбец или в одну строку, причем расположение записей промежуточных результатов должно быть удобным для производства последующих вычислений.

Например, для составления таблицы значений функции

$$y=\frac{e^x+\cos x}{1+x^2}+\sqrt{1+\sin^2 x} \qquad (1)$$

можно рекомендовать вычислительный бланк, приведенный в таблице 1.

Таблица 1

**Вычислительный бланк для функции (1)**

| $x$ | $x^2$ (1)$^2$ | $e^x$ | $\sin x$ | $\cos x$ | $e^x+\cos x$ (3)+(5) | $1+x^2$ (1)+(2) | $\frac{e^x+\cos x}{1+x^2}$ (6):(7) | $\sin^2 x$ (4)$^2$ | $1+\sin^2 x$ (1)+(9) | $\sqrt{1+\sin^2 x}$ $\sqrt{(10)}$ | $y$ (8)+ +(11) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| | | | | | | | | | | | |

Вычисления ведутся по столбцам, причем характер выполняемых однотипных операций ясен из самого вычислительного бланка.

Сначала в столбец (1) записываются данные значения аргумента $x$. Затем все числа столбца (1) возводятся в квадрат и заносятся в столбец (2). Далее по таблицам определяются для каждого числа столбца (1) последовательно значения $e^x$, $\sin x$, $\cos x$ и заполняются соответственно столбцы (3), (4), (5).

В дальнейших столбцах указаны результаты промежуточных операций. Например, столбец (6) содержит значения сумм $e^x+\cos x$ (схематически (3)+(5)) и т. д. В последнем столбце (12) приводятся

значения искомой функции $y$. При правильно составленном бланке вычислитель в процессе вычисления уже фактически не пользуется формулой, по которой ведется расчет, его внимание сосредоточено исключительно на последовательном заполнении столбцов.

Заметим, что расчетная схема и форма бланка существенно зависят от используемой техники вычислений и вспомогательных таблиц. Так, например, в некоторых случаях отдельные промежуточные результаты хранятся в памяти машины и в бланк не заносятся. Иногда стандартные совокупности операций удобно рассматривать как отдельное действие. Например, при использовании логарифмической линейки численное значение выражения вида

$$\frac{ab}{c}$$

можно вычислять сразу, не фиксируя промежуточный результат, и поэтому нет необходимости расчленять его на простейшие операции умножения и деления. Аналогично при работе на электрических счетных машинах процесс отыскания суммы парных произведений

$$\sum_{k=1}^{n} a_k b_k$$

является единым действием. Во многих случаях выгодно преобразовывать данные выражения к специальному искусственному виду (например, заменять деление умножением на обратную величину, или приводить выражение к виду, удобному для логарифмирования, и т. п.).

Второе, на что нужно обратить серьезное внимание, — это *контроль вычислений*. Без проверки вычисление не может считаться законченным. Контроль разделяется на *текущий* и *заключительный*. При текущем контроле, производя добавочные действия, мы с большей или меньшей степенью достоверности убеждаемся, что полученные промежуточные результаты правильны. В противном случае производится пересчет соответствующего этапа. При заключительном контроле проверяется лишь окончательный результат. Например, если вычисляется корень уравнения, то найденное значение можно подставить в уравнение и таким образом узнать, правильно или нет решена задача. По здравому смыслу ясно, что если вычисление очень большое, то рискованно ставить под угрозу всю вычислительную работу, дожидаясь проверки окончательного результата. Поэтому целесообразно проверять правильность расчетов по этапам. В ответственных случаях расчеты контролируются путем независимого выполнения расчетов двумя различными вычислителями, или же задача решается одним и тем же вычислителем двумя различными способами.

Третий важный момент — *оценка точности*. В большинстве случаев вычисления производятся с приближенными числами и притом приближенно. Поэтому даже для точного метода решения задачи на каждом этапе вычислений возникают *погрешности действий* и

*погрешности округлений*. Если сам метод — приближенный, то к этим двум погрешностям присоединяется *погрешность метода*. При неблагоприятных обстоятельствах суммарная погрешность может быть столь велика, что полученный результат будет иметь лишь иллюзорное значение. В соответствующих главах книги указаны методы оценки погрешностей для основных вычислений.

В вычислительном бланке полезно предусмотреть столбцы для табличных разностей (см. гл. XIV, § 2), которые можно использовать для контроля вычислений. А именно, если правильность таблицы разностей нарушается на отдельном участке, то следует пересчитать соответствующие элементы таблицы (либо выявить причину нарушения).

Нужно обратить внимание также на *аккуратность* и *четкость* записи в вычислительных бланках. Практика показывает, что нечеткая запись цифр часто приводит к ошибкам и может погубить хорошо организованное вычисление. Особенно опасны ошибки в записи чисел, содержащих большое число нулей. Такие числа следует записывать в нормальной форме, выделяя целую степень десяти, например

$$0{,}00000345 = 3{,}45 \cdot 10^{-6}$$

и т. п.

Дальнейшая часть книги посвящена главным образом **методам вычислений**. Приводимые числовые примеры во многих случаях упрощены, причем промежуточные выкладки часто опускаются.

ГЛАВА I

# ПРИБЛИЖЕННЫЕ ЧИСЛА

## § 1. Абсолютная и относительная погрешности

*Приближенным числом* $a$ называется число, незначительно отличающееся от точного $A$ и заменяющее последнее в вычислениях. Если известно, что $a < A$, то $a$ называется приближенным значением числа $A$ *по недостатку*; если же $a > A$, то — *по избытку*. Например, для $\sqrt{2}$ число 1,41 будет приближенным значением по недостатку, а 1,42 — по избытку, так как $1{,}41 < \sqrt{2} < 1{,}42$. Если $a$ есть приближенное значение числа $A$, то пишут $a \approx A$.

Под *ошибкой* или *погрешностью* $\Delta a$ приближенного числа $a$ обычно понимается разность между соответствующим точным числом $A$ и данным приближенным, т. е.

$$\Delta a = A - a^{*)}.$$

Если $A > a$, то ошибка положительна: $\Delta a > 0$; если же $A < a$, то ошибка отрицательна: $\Delta a < 0$. Чтобы получить точное число $A$, нужно к приближенному числу $a$ прибавить его ошибку $\Delta a$, т. е.

$$A = a + \Delta a.$$

Таким образом, точное число можно рассматривать как приближенное с ошибкой, равной нулю.

Во многих случаях знак ошибки неизвестен. Тогда целесообразно пользоваться *абсолютной погрешностью приближенного числа*

$$\Delta = |\Delta a|.$$

Определение 1. *Абсолютной погрешностью* $\Delta$ приближенного числа $a$ называется абсолютная величина разности между соответствующим точным числом $A$ и числом $a$, т. е.

$$\Delta = |A - a|. \tag{1}$$

Здесь следует различать два случая:

1) число $A$ нам известно, тогда абсолютная погрешность $\Delta$ легко определяется по формуле (1):

---

*) Иногда ошибкой называют разность $a - A$

2) число $A$ нам не известно, что практически бывает чаще всего, и, следовательно, мы не можем определить и абсолютную погрешность $\Delta$ по формуле (1).

В этом случае полезно вместо неизвестной теоретической абсолютной погрешности $\Delta$ ввести ее оценку сверху, так называемую *предельную абсолютную погрешность*.

Определение 2. Под *предельной абсолютной погрешностью* приближенного числа понимается всякое число, не меньшее абсолютной погрешности этого числа.

Таким образом, если $\Delta_a$ — предельная абсолютная погрешность приближенного числа $a$, заменяющего точное $A$, то

$$\Delta = |A - a| \leqslant \Delta_a. \tag{2}$$

Отсюда следует, что точное число $A$ заключено в границах

$$a - \Delta_a \leqslant A \leqslant a + \Delta_a. \tag{3}$$

Следовательно, $a - \Delta_a$ есть приближение числа $A$ по недостатку, а $a + \Delta_a$ — приближение числа $A$ по избытку.

В этом случае для краткости пользуются записью

$$A = a \pm \Delta_a.$$

Пример 1. Определить предельную абсолютную погрешность числа $a = 3{,}14$, заменяющего число $\pi$.

Решение. Так как имеет место неравенство

$$3{,}14 < \pi < 3{,}15, \quad \text{то} \quad |a - \pi| < 0{,}01$$

и, следовательно, можно принять $\Delta_a = 0{,}01$.

Если учесть, что

$$3{,}14 < \pi < 3{,}142,$$

то будем иметь лучшую оценку: $\Delta_a = 0{,}002$.

Заметим, что сформулированное выше понятие предельной абсолютной погрешности является весьма широким, а именно: *под предельной абсолютной погрешностью приближенного числа $a$ понимается любой представитель бесконечного множества неотрицательных чисел $\Delta_a$, удовлетворяющих неравенству* (2). Отсюда логически вытекает, что всякое число, большее предельной абсолютной погрешности данного приближенного числа, также может быть названо предельной абсолютной погрешностью этого числа. Практически удобно в качестве $\Delta_a$ выбирать возможно меньшее при данных обстоятельствах число, удовлетворяющее неравенству (2).

В записи приближенного числа, полученного в результате измерения, обычно отмечают его предельную абсолютную погрешность. Например, если длина отрезка $l = 214$ *см* с точностью до 0,5 *см*, то пишут $l = 214$ *см* $\pm$ 0,5 *см*. Здесь предельная абсолютная погрешность $\Delta_l = 0{,}5$ *см*, а точная величина длины $l$ отрезка заключена в границах 213,5 *см* $\leqslant l \leqslant$ 214,5 *см*.

Абсолютная погрешность (или предельная абсолютная погрешность) не достаточна для характеристики точности измерения или вычисления. Так, например, если при измерении длин двух стержней получены результаты $l_1 = 100{,}8$ *см* $\pm\, 0{,}1$ *см* и $l_2 = 5{,}2$ *см* $\pm\, 0{,}1$ *см*, то, несмотря на совпадение предельных абсолютных погрешностей, качество первого измерения выше, чем второго. Для точности данных измерений существенна абсолютная погрешность, приходящаяся на единицу длины, которая носит название *относительной погрешности*.

Определение 3. *Относительной погрешностью* $\delta$ приближенного числа $a$ называется отношение абсолютной погрешности $\Delta$ этого числа к модулю соответствующего точного числа $A\,(A \neq 0)$, т. е.

$$\delta = \frac{\Delta}{|A|}. \tag{4}$$

Отсюда $\Delta = |A|\,\delta$.

Так же как и для абсолютной погрешности, введем понятие *предельной относительной погрешности*.

Определение 4. *Предельной относительной погрешностью* $\delta_a$ данного приближенного числа $a$ называется всякое число, не меньшее относительной погрешности этого числа. По определению имеем:

$$\delta \leqslant \delta_a, \tag{5}$$

т. е. $\frac{\Delta}{|A|} \leqslant \delta_a$, отсюда $\Delta \leqslant |A|\,\delta_a$.

Таким образом, за предельную абсолютную погрешность числа $a$ можно принять:

$$\Delta_a = |A|\,\delta_a. \tag{6}$$

Так как на практике $A \approx a$, то вместо формулы (6) часто пользуются формулой

$$\Delta_a = |a|\,\delta_a. \tag{6'}$$

Отсюда, зная предельную относительную погрешность $\delta_a$, получают границы для точного числа. То обстоятельство, что точное число лежит между $a\,(1-\delta_a)$ и $a\,(1+\delta_a)$, условно записывают так:

$$A = a\,(1 \pm \delta_a).$$

Пусть $a$ — приближенное число, заменяющее точное $A$, и $\Delta_a$ — предельная абсолютная погрешность числа $a$. Положим для определенности, что $A > 0$, $a > 0$ и $\Delta_a < a$. Тогда

$$\delta = \frac{\Delta}{A} \leqslant \frac{\Delta_a}{a - \Delta_a}.$$

Следовательно, в качестве предельной относительной погрешности числа $a$ можно принять число

$$\delta_a = \frac{\Delta_a}{a - \Delta_a}.$$

Аналогично получаем $\Delta = A\delta \leqslant (a+\Delta)\,\delta_a$; отсюда

$$\Delta_a = \frac{a\delta_a}{1-\delta_a}.$$

Если, как обычно бывает, $\Delta_a \ll a$ и $\delta_a \ll 1$ (знак $\ll$ обозначает «значительно меньше»), то приближенно можно принять:

$$\delta_a \approx \frac{\Delta_a}{a}$$

и

$$\Delta_a \approx a\delta_a.$$

Пример 2. Вес 1 *дм*³ воды при 0°C $p = 999{,}847\ Г \pm 0{,}001\ Г$. Определить предельную относительную погрешность результата взвешивания.

Решение. Очевидно, что $\Delta_p = 0{,}001\ Г$ и $p \leqslant 999{,}846\ Г$. Следовательно,

$$\delta_p = \frac{0{,}001}{999{,}846} \approx 10^{-4}\%.$$

Пример 3. При определении газовой постоянной для воздуха получили $R = 29{,}25$. Зная, что относительная погрешность этого значения равна $1^0/_{00}$, найти пределы, в которых заключается $R$.

Решение. Имеем $\delta_R = 0{,}001$, тогда $\Delta_R = R\delta_R \approx 0{,}03$. Следовательно, $29{,}22 \leqslant R \leqslant 29{,}28$.

## § 2. Основные источники погрешностей

Погрешности, встречающиеся в математических задачах, могут быть в основном разбиты на пять групп.

1. Погрешности, связанные с самой постановкой математической задачи. Математические формулировки редко точно отображают реальные явления: обычно они дают лишь более или менее идеализированные модели. Как правило, при изучении тех или иных явлений природы мы вынуждены принять некоторые, упрощающие задачу, условия, что вызывает ряд погрешностей *(погрешности задачи)*.

Иногда бывает и так, что решить задачу в точной постановке трудно или даже невозможно. Тогда ее заменяют близкой по результатам приближенной задачей. При этом возникает погрешность, которую можно назвать *погрешностью метода*.

2. Погрешности, связанные с наличием бесконечных процессов в математическом анализе. Функции, фигурирующие в математических формулах, часто задаются в виде бесконечных последовательностей или рядов $\left(\text{например, } \sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} + \ldots\right)$. Более того, многие математические уравнения можно решить, лишь описав бесконечные процессы, пределы которых и являются искомыми решениями.

Так как бесконечный процесс, вообще говоря, не может быть завершен в конечное число шагов, то мы вынуждены остановиться на некотором члене последовательности, считая его приближением к искомому решению. Понятно, что такой обрыв процесса вызывает погрешность, называемую обычно *остаточной погрешностью*.

3. Погрешности, связанные с наличием в математических формулах числовых параметров, значения которых могут быть определены лишь приближенно. Таковы, например, все физические константы. Условно назовем эту погрешность *начальной*.

4. Погрешности, связанные с системой счисления. При изображении даже рациональных чисел в десятичной системе или другой позиционной системе справа от запятой может быть бесконечное число цифр (например, может получиться бесконечная десятичная периодическая дробь). При вычислениях, очевидно, можно использовать лишь конечное число этих цифр. Так возникает *погрешность округления*. Например, полагая $\frac{1}{3}=0{,}333$, получаем погрешность $\Delta=4\cdot10^{-4}$. Приходится так же округлять и конечные числа, имеющие большое количество знаков.

5. Погрешности, связанные с действиями над приближенными числами *(погрешности действий)*. Понятно, что, производя вычисления с приближенными числами, погрешности исходных данных в какой-то мере мы переносим в результат вычислений. В этом отношении погрешности действий являются неустранимыми.

Само собой разумеется, что при решении конкретной задачи те или иные погрешности иногда отсутствуют, или влияние их ничтожно. Но, вообще говоря, для полного анализа погрешностей следует учитывать все их виды. В дальнейшем мы ограничимся в основном исчислением погрешностей действий и погрешностей методов.

## § 3. Десятичная запись приближенных чисел. Значащая цифра. Число верных знаков

Известно, что всякое положительное число $a$ может быть представлено в виде конечной или бесконечной десятичной дроби

$$a=\alpha_m 10^m+\alpha_{m-1}10^{m-1}+\alpha_{m-2}10^{m-2}+\ldots \\ \ldots+\alpha_{m-n+1}10^{m-n+1}+\ldots, \qquad (1)$$

где $\alpha_i$ — цифры числа $a$ ($\alpha_i=0, 1, 2, \ldots, 9$), причем старшая цифра $\alpha_m\neq0$, а $m$ — некоторое целое число (старший десятичный разряд числа $a$). Например,

$$3141{,}59\ldots=3\cdot10^3+1\cdot10^2+4\cdot10^1+1\cdot10^0+ \\ +5\cdot10^{-1}+9\cdot10^{-2}+\ldots$$

Каждая единица, стоящая на определенном месте в числе $a$, написанном в виде десятичной дроби (1), имеет свое значение. Единица, стоящая на первом месте, равна $10^m$, на втором — $10^{m-1}$, на $n$-м — $10^{m-n+1}$ и т. д.

На практике преимущественно приходится иметь дело с приближенными числами, представляющими собой конечные десятичные дроби

$$b=\beta_m 10^m+\beta_{m-1} 10^{m-1}+\ldots+\beta_{m-n+1} 10^{m-n+1} \quad (\beta_m \neq 0). \qquad (2)$$

Все сохраняемые десятичные знаки $\beta_i$ $(i=m,\ m-1,\ \ldots,\ m-n+1)$ называются *значащими цифрами* приближенного числа $b$, причем возможно, что некоторые из них равны нулю (за исключением $\beta_m$). При позиционном изображении числа $b$ в десятичной системе счисления иногда приходится вводить лишние нули в начале или в конце числа. Например,

$$b=7\cdot 10^{-3}+0\cdot 10^{-4}+1\cdot 10^{-5}+0\cdot 10^{-6}=\underline{0,00}70\underline{10},$$

или

$$b=2\cdot 10^9+0\cdot 10^8+0\cdot 10^7+3\cdot 10^6+0\cdot 10^5=2\,003\,\underline{000\,000}.$$

Такие нули (в приведенных примерах они подчеркнуты) не считаются значащими цифрами.

Определение 1. *Значащей цифрой* приближенного числа называются всякая цифра в его десятичном изображении, отличная от нуля, и нуль, если он содержится между значащими цифрами или является представителем сохраненного десятичного разряда. Все остальные нули, входящие в состав приближенного числа и служащие лишь для обозначения десятичных разрядов его, не причисляются к значащим цифрам.

Например, в числе 0,002 080 первые три нуля не являются значащими цифрами, так как они служат только для установления десятичных разрядов других цифр. Остальные два нуля являются значащими цифрами, так как первый из них находится между значащими цифрами 2 и 8, а второй, как это отражено в записи, указывает, что в приближенном числе сохранен десятичный разряд $10^{-6}$. В случае, если в данном числе 0,002 080 последняя цифра не является значащей, то это число должно быть записано в виде 0,002 08. С этой точки зрения числа 0,002 080 и 0,002 08 не равноценны, так как первое из них содержит четыре значащих цифры, а второе — лишь три значащих цифры.

При написании больших чисел нули справа могут служить как для обозначения значащих цифр, так и для определения разрядов остальных цифр. Поэтому при обычной записи чисел могут возникнуть неясности. Например, рассматривая число 689 000, мы не имеем возможности по его виду судить о том, сколько в нем значащих цифр, хотя можно утверждать, что их не меньше трех. Этой неопределенности можно избежать, выявив десятичный порядок числа

и записав его в виде $6{,}89\cdot 10^5$, если оно имеет три значащих цифры; или $6{,}8900\cdot 10^5$, если число имеет пять значащих цифр, и т. п. Вообще, такого рода запись удобна для чисел, содержащих большое количество незначащих нулей, например $0{,}000\,000\,120 = 1{,}20\cdot 10^{-7}$ и т. п.

Введем понятие о *верных десятичных знаках приближенного числа*.

Определение 2. Говорят, что $n$ первых значащих цифр (десятичных знаков) приближенного числа являются *верными*, если абсолютная погрешность этого числа не превышает половины единицы разряда, выражаемого $n$-й значащей цифрой, считая слева направо.

Таким образом, если для приближенного числа $a$ (1), заменяющего точное число $A$, известно, что

$$\Delta = |A - a| \leqslant \frac{1}{2}\cdot 10^{m-n+1},$$

то, по определению, первые $n$ цифр $\alpha_m, \alpha_{m-1}, \ldots, \alpha_{m-n+1}$ этого числа являются верными.

Например, для точного числа $A = 35{,}97$ число $a = 36{,}00$ является приближением с тремя верными знаками, так как $|A - a| = 0{,}03 < \frac{1}{2}\cdot 0{,}1$.

Заметим, что в математических таблицах все помещенные значащие цифры являются верными. Так, например, в пятизначных таблицах логарифмов гарантировано, что абсолютная погрешность мантиссы не превосходит $\frac{1}{2}\cdot 10^{-5}$ и т. п.

Термин «$n$ верных знаков» не следует понимать буквально, т. е. так, что в данном приближенном числе $a$, имеющем $n$ верных знаков, $n$ первых значащих цифр его совпадают с соответствующими цифрами точного числа $A$. Например, приближенное число $a = 9{,}995$, заменяющее точное $A = 10$, имеет три верных знака, причем все цифры этих чисел различны. Однако во многих случаях дело обстоит именно так, что верные знаки приближенного числа одинаковы с соответствующими цифрами точного числа.

Замечание. В некоторых случаях удобно говорить, что число $a$ является приближением точного числа $A$ с $n$ *верными знаками* в широком смысле, понимая под этим, что абсолютная погрешность $\Delta = |A - a|$ не превышает единицы десятичного разряда, выражаемого $n$-й значащей цифрой приближенного числа.

Например, для точного числа $A = 412{,}3567$ число $a = 412{,}356$ является приближением с шестью верными знаками в широком смысле, так как $\Delta = 0{,}0007 < 1\cdot 10^{-3}$.

В дальнейшем верные знаки приближенного числа мы будем понимать в смысле определения 2 (т. е. *в узком смысле*), если явно не оговорено противное.

## § 4. Округление чисел

Рассмотрим некоторое приближенное или точное число $a$, записанное в десятичной нумерации. Часто бывает надобность в *округлении* этого числа, т. е. в замене его числом $a_1$ с меньшим количеством значащих цифр. Число $a_1$ выбирают так, чтобы *погрешность округления* $|a_1 - a|$ была минимальной.

**Правило округления (по дополнению).** Чтобы округлить число до $n$ значащих цифр, отбрасывают все цифры его, стоящие справа от $n$-й значащей цифры, или, если это нужно для сохранения разрядов, заменяют их нулями. При этом:

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

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

3) если первая из отброшенных цифр равна 5 и среди остальных отброшенных цифр имеются ненулевые, то последняя оставшаяся цифра увеличивается на единицу;

3а) если же первая из отброшенных цифр равна 5 и все остальные отброшенные цифры являются нулями, то последняя оставшаяся цифра сохраняется неизменной, если она четная, и увеличивается на единицу, если она нечетная (правило четной цифры).

Иными словами, если при округлении числа отбрасывается меньше половины единицы последнего сохраняемого десятичного разряда, то цифры всех сохраненных разрядов остаются неизменными; если же отброшенная часть числа составляет больше половины единицы последнего сохраненного десятичного разряда, то цифра этого разряда увеличивается на единицу. В исключительном случае, когда отброшенная часть в точности равна половине единицы последнего сохраненного десятичного разряда, то для компенсации знаков ошибок округления используется правило четной цифры.

Очевидно, что при применении правила округления погрешность округления не превосходит $\frac{1}{2}$ единицы десятичного разряда, определяемого последней оставленной значащей цифрой.

Пример 1. Округляя число

$$\pi = 3{,}14159\,26535\ldots$$

до пяти, четырех и трех значащих цифр, получим приближенные числа 3,1416; 3,142; 3,14 с абсолютными погрешностями, меньшими $\frac{1}{2}\cdot 10^{-4}$; $\frac{1}{2}\cdot 10^{-3}$ и $\frac{1}{2}\cdot 10^{-2}$.

Пример 2. Округляя число 1,2500 до двух значащих цифр, получим приближенное число 1,2 с абсолютной погрешностью, равной $\frac{1}{2}\cdot 10^{-1} = 0{,}05$.

Точность приближенного числа зависит не от количества значащих цифр, а от количества верных значащих цифр [1], [2]. В тех случаях, когда приближенное число содержит излишнее количество неверных значащих цифр, прибегают к округлению. Обычно руководствуются следующим практическим правилом: *при выполнении приближенных вычислений число значащих цифр промежуточных результатов не должно превышать числа верных цифр более чем на одну или две единицы.* Окончательный результат может содержать не более чем одну излишнюю значащую цифру, по сравнению с верными. Если при этом абсолютная погрешность результата не превышает двух единиц последнего сохраненного десятичного разряда, то излишняя цифра называется *сомнительной.*

Приведенное правило позволяет без ущерба точности вычислений избегать написания лишних цифр и значительно экономит время вычислений. Сохранение запасных знаков имеет тот смысл, что обычно оценка погрешностей результатов производится для наихудших вариантов, и фактическая погрешность может оказаться значительно меньше максимальной теоретической. Таким образом, во многих случаях те значащие цифры, которые считаются неверными, на самом деле являются верными.

Приходится также округлять точные числа, содержащие слишком много или бесконечное количество значащих цифр, сообразуясь с общей точностью вычислений.

Заметим, что если точное число $A$ округлить по правилу дополнения до $n$ значащих цифр, то полученное таким образом приближенное число $a$ будет иметь $n$ верных цифр (в узком смысле).

Если же приближенное число $a$, имеющее $n$ верных цифр, округлить до $n$ значащих цифр, то полученное новое приближенное число $a_1$, вообще говоря, будет иметь $n$ верных цифр в широком смысле. Действительно, в силу неравенства

$$|A-a_1| \leqslant |A-a|+|a-a_1|$$

предельная абсолютная погрешность числа $a_1$ складывается из абсолютной погрешности числа $a$ и погрешности округления.

## § 5. Связь относительной погрешности приближенного числа с количеством верных знаков этого числа

Докажем теорему, которая связывает величину относительной погрешности приближенного числа с количеством верных знаков этого числа [3], [4].

Теорема. *Если положительное приближенное число $a$ имеет $n$ верных десятичных знаков в узком смысле, то относительная погрешность $\delta$ этого числа не превосходит $\left(\frac{1}{10}\right)^{n-1}$, деленную на первую*

*значащую цифру данного числа, т. е.*

$$\delta \leqslant \frac{1}{\alpha_m}\left(\frac{1}{10}\right)^{n-1},$$

*где* $\alpha_m$ — *первая значащая цифра числа* $a$.

Доказательство. Пусть число

$$a = \alpha_m 10^m + \alpha_{m-1} 10^{m-1} + \ldots + \alpha_{m-n+1} 10^{m-n+1} + \ldots \quad (\alpha_m \geqslant 1)$$

является приближенным значением точного числа $A$ и имеет $n$ верных знаков. Тогда по определению имеем:

$$\Delta = |A - a| \leqslant \frac{1}{2} \cdot 10^{m-n+1};$$

отсюда

$$A \geqslant a - \frac{1}{2} \cdot 10^{m-n+1}.$$

Последнее неравенство еще более усилится, если число $a$ заменим заведомо меньшим числом $\alpha_m 10^m$,

$$A \geqslant \alpha_m 10^m - \frac{1}{2} \cdot 10^{m-n+1} = \frac{1}{2} \cdot 10^m \left(2\alpha_m - \frac{1}{10^{n-1}}\right). \qquad (1)$$

Правая часть неравенства (1) достигает наименьшего значения при $n = 1$. Поэтому

$$A \geqslant \frac{1}{2} \cdot 10^m (2\alpha_m - 1), \qquad (2)$$

или, так как

$$2\alpha_m - 1 = \alpha_m + (\alpha_m - 1) \geqslant \alpha_m,$$

то

$$A \geqslant \frac{1}{2} \alpha_m 10^m.$$

Следовательно,

$$\delta = \frac{\Delta}{A} \leqslant \frac{\frac{1}{2} 10^{m-n+1}}{\frac{1}{2} \alpha_m 10^m} = \frac{1}{\alpha_m}\left(\frac{1}{10}\right)^{n-1}.$$

Итак,

$$\delta \leqslant \frac{1}{\alpha_m}\left(\frac{1}{10}\right)^{n-1}. \qquad (3)$$

Теорема доказана.

Замечание 1. Пользуясь неравенством (2), можно получить более точную оценку относительной погрешности $\delta$.

Следствие 1. За предельную относительную погрешность числа $a$ можно принять:

$$\delta_a = \frac{1}{\alpha_m}\left(\frac{1}{10}\right)^{n-1}, \tag{4}$$

где $\alpha_m$ — первая значащая цифра числа $a$.

Следствие 2. Если число $a$ имеет больше двух верных знаков, т. е. $n \geqslant 2$, то практически справедлива формула

$$\delta_a = \frac{1}{2\alpha_m}\left(\frac{1}{10}\right)^{n-1}. \tag{5}$$

Действительно, при $n \geqslant 2$ числом $\frac{1}{10^{n-1}}$ в неравенстве (1) можно пренебречь. Тогда

$$A \geqslant \frac{1}{2} \cdot 10^m \cdot 2\alpha_m = \alpha_m 10^m;$$

отсюда

$$\delta = \frac{\Delta}{A} \leqslant \frac{\frac{1}{2} \cdot 10^{m-n+1}}{\alpha_m 10^m} = \frac{1}{2\alpha_m}\left(\frac{1}{10}\right)^{n-1}.$$

Следовательно,

$$\delta_a = \frac{1}{2\alpha_m}\left(\frac{1}{10}\right)^{n-1}.$$

Замечание 2. Если приближенное число $a$ имеет $n$ верных десятичных знаков в широком смысле, то оценки (4) и (5) следует увеличить в два раза.

Пример 1. Какова предельная относительная погрешность, если вместо числа $\pi$ взять число $a = 3{,}14$?

Решение. В нашем случае $\alpha_m = 3$ и $n = 3$. Следовательно,

$$\delta_a = \frac{1}{2 \cdot 3}\left(\frac{1}{10}\right)^{3-1} = \frac{1}{600} = \frac{1}{6}\%.$$

Пример 2. Со сколькими десятичными знаками надо взять $\sqrt{20}$, чтобы погрешность не превышала 0,1%?

Решение. Так как первая цифра 4, то $\alpha_m = 4$, причем $\delta = 0{,}001$. Имеем $\frac{1}{4 \cdot 10^{n-1}} \leqslant 0{,}001$, отсюда $10^{n-1} \geqslant 250$ и $n \geqslant 4$.

Приведенная теорема дает возможность по числу верных знаков приближенного числа

$$a = \alpha_m \cdot 10^m + \alpha_{m-1} 10^{m-1} + \ldots \tag{6}$$

определить его относительную погрешность $\delta$.

Для решения обратной задачи — определения количества $n$ верных знаков числа (6), если известна его относительная погрешность $\delta$, обычно пользуются приближенной формулой

$$\delta = \frac{\Delta}{a} \ (a > 0),$$

где $\Delta$ — абсолютная погрешность числа $a$. Отсюда

$$\Delta = a\delta. \tag{7}$$

Учитывая старший десятичный разряд числа $\Delta$, легко установить количество верных знаков данного приближенного числа $a$. В частности, если

$$\delta \leqslant \frac{1}{10^n},$$

то из формул (6) и (7) имеем:

$$\Delta \leqslant (\alpha_m + 1)\, 10^m \cdot 10^{-n} \leqslant 10^{m-n+1},$$

т. е. число $a$ заведомо имеет $n$ верных десятичных знаков в широком смысле. Аналогично, если

$$\delta \leqslant \frac{1}{2 \cdot 10^n},$$

то число $a$ имеет $n$ верных знаков в узком смысле.

Пример 3. Приближенное число $a = 24\,253$ имеет относительную точность $1\,\%$. Сколько в нем верных знаков?

Решение. Имеем:

$$\Delta = 24\,253 \cdot 0{,}01 \approx 243 = 2{,}43 \cdot 10^2.$$

Следовательно, число $a$ имеет верными лишь первые две цифры ($n = 2$); цифра сотен является сомнительной. Согласно приведенному выше правилу число $a$ предпочтительнее записать в виде $a = 2{,}4\underline{3} \cdot 10^4$.

Замечание. Указанный способ определения числа верных знаков является приближенным. При точном подсчете верных цифр числа $a$ следует исходить из неравенств

$$\delta \geqslant \frac{\Delta}{a + \Delta}$$

и

$$\Delta \leqslant \frac{a\delta}{1 - \delta} \quad (0 \leqslant \delta < 1).$$

## § 6. Таблицы для определения предельной относительной погрешности по числу верных знаков и наоборот

Если приближенное число написано с указанными верными десятичными знаками, то можно легко подсчитать его предельную относительную погрешность. Практически с таким подсчетом приходится сталкиваться часто и поэтому желательно рационализиро-

вать эту операцию. Таблица 2 [5] указывает относительную погрешность в процентах приближенного числа в зависимости от количества верных в широком смысле десятичных знаков его и от первых двух значащих цифр числа, считая слева направо.

Таблица 2

**Относительная погрешность (в %) чисел с $n$ верными знаками**

| Первые две значащие цифры | $n$ | | |
|---|---|---|---|
| | 2 | 3 | 4 |
| 10—11 | 10 | 1 | 0,1 |
| 12—13 | 8,3 | 0,83 | 0,083 |
| 14, ..., 16 | 7,1 | 0,71 | 0,071 |
| 17, ..., 19 | 5,9 | 0,59 | 0,059 |
| 20, ..., 22 | 5 | 0,5 | 0,05 |
| 23, ..., 25 | 4,3 | 0,43 | 0,043 |
| 26, ..., 29 | 3,8 | 0,38 | 0,038 |
| 30, ..., 34 | 3,3 | 0,33 | 0,033 |
| 35, ..., 39 | 2,9 | 0,29 | 0,029 |
| 40, ..., 44 | 2,5 | 0,25 | 0,025 |
| 45, ..., 49 | 2,2 | 0,22 | 0,022 |
| 50, ..., 59 | 2 | 0,2 | 0,02 |
| 60, ..., 69 | 1,7 | 0,17 | 0,017 |
| 70, ..., 79 | 1,4 | 0,14 | 0,014 |
| 80, ..., 89 | 1,2 | 0,12 | 0,012 |
| 90, ..., 99 | 1,1 | 0,11 | 0,011 |

Пусть, например, имеем приближенное число 0,00354 с тремя верными десятичными знаками. Так как здесь $n=3$ и число 35 содержится в промежутке 35, ..., 39, то по таблице 2 находим $\delta=0{,}29$ %.

Если известна только первая цифра числа, например 4, то берем, конечно, большее из чисел 2,5 и 2,2, соответствующих возможным вариантам 40, ..., 44 и 45, ..., 49 (при $n=2$). Если и первая цифра неизвестна, то берем числа из первой строки (10 %; 1 %; 0,1 %), как наибольшие. Из этой таблицы мы видим, что три верных знака обеспечивают относительную точность (не менее 1 %), достаточную для большинства технических расчетов. Заметим, что если приближенное число имеет два, три или четыре верных знака в узком смысле, то все числа таблицы нужно уменьшить вдвое.

В таблице 3 [5] приведены верхние границы для относительных погрешностей (в процентах), обеспечивающих данному приближенному значению то или другое число верных знаков в широком смысле в зависимости от его первых двух цифр.

Таблица 3

**Число верных знаков приближенного числа в зависимости от предельной относительной погрешности (в %)**

| Первые две значащие цифры | $n$ | | |
|---|---|---|---|
| | 2 | 3 | 4 |
| 10—11 | 4,2 | 0,42 | 0,042 |
| 12—13 | 3,6 | 0,36 | 0,036 |
| 14, ..., 16 | 2,9 | 0,29 | 0,029 |
| 17, ..., 19 | 2,5 | 0,25 | 0,025 |
| 20, ..., 22 | 2,2 | 0,22 | 0,022 |
| 23, ..., 25 | 1,9 | 0,19 | 0,019 |
| 26, ..., 29 | 1,7 | 0,17 | 0,017 |
| 30, ..., 34 | 1,4 | 0,14 | 0,014 |
| 35, ..., 39 | 1,2 | 0,12 | 0,012 |
| 40, ..., 44 | 1,1 | 0,11 | 0,011 |
| 45, ..., 49 | 1 | 0,1 | 0,01 |
| 50, ..., 54 | 0,9 | 0,09 | 0,009 |
| 55, ..., 59 | 0,8 | 0,08 | 0,008 |
| 60, ..., 69 | 0,7 | 0,07 | 0,007 |
| 70, ..., 79 | 0,6 | 0,06 | 0,006 |
| 80, ..., 99 | 0,5 | 0,05 | 0,005 |

Покажем на примере, как надо пользоваться таблицей 3. Пусть, например, дано приближенное число $a=5{,}297$ с относительной погрешностью $\delta=0{,}5\ \%$. Здесь первые две значащие цифры 5 и 2; число, образованное этими цифрами, содержится между 50 и 54, причем последним, в зависимости от числа верхних знаков, соответствуют относительные погрешности 0,9 %; 0,09 %; 0,009 % и т. д. Так как $\delta=0{,}5\ \% < 0{,}9\ \%$ и относительная погрешность числа не зависит от того, какие десятичные разряды выражают цифры этого числа, то число $a=5{,}297$ имеет два верных десятичных знака в широком смысле.

Примеры. 1. Полагая $\pi=3{,}142$; $\sqrt{7}=2{,}65$; $e=2{,}718$; $\lg 5=0{,}699$; $\sin 1^\circ=0{,}0174$, по таблице 2 находим, что соответствующие относительные погрешности следующие: $\delta=0{,}033\%$; $\delta=0{,}19\%$; $\delta=0{,}019\%$; $\delta=0{,}17\%$; $\delta=0{,}59\%$.

2. По прогибу стального стержня вычислен модуль Юнга $E=2212\ldots\ T/см^2$ с точностью до 2%. Сколько верных знаков в найденном значении? По таблице 3 находим $n=2$. Следовательно, $E=22\cdot 10^2\ T/см^2$.

3. Для взрывчатой смеси в газомоторе вычислена газовая постоянная $R=31{,}5\ldots$ с относительной погрешностью $\delta=1\%$. Определить число верных знаков. По таблице 3 находим $n=2$. Значит, $R=32$.

## § 7. Погрешность суммы

Теорема 1. *Абсолютная погрешность алгебраической суммы нескольких приближенных чисел не превышает суммы абсолютных погрешностей этих чисел.*

Доказательство. Пусть $x_1, x_2, \ldots, x_n$ — данные приближенные числа. Рассмотрим их алгебраическую сумму

$$u = \pm x_1 \pm x_2 \pm \ldots \pm x_n.$$

Очевидно, что

$$\Delta u = \pm \Delta x_1 \pm \Delta x_2 \pm \ldots \pm \Delta x_n$$

и, следовательно,

$$|\Delta u| \leqslant |\Delta x_1| + |\Delta x_2| + \ldots + |\Delta x_n|. \tag{1}$$

Следствие. За предельную абсолютную погрешность алгебраической суммы можно принять сумму предельных абсолютных погрешностей слагаемых

$$\Delta_u = \Delta_{x_1} + \Delta_{x_2} + \ldots + \Delta_{x_n}. \tag{2}$$

Из формулы (2) следует, что предельная абсолютная погрешность суммы не может быть меньше предельной абсолютной погрешности наименее точного (в смысле абсолютной погрешности) из слагаемых, т. е. слагаемого, имеющего максимальную абсолютную погрешность. Следовательно, с какой бы степенью точности ни были определены остальные слагаемые, мы не можем за их счет увеличить точность суммы. Поэтому не имеет смысла сохранять излишние знаки и в более точных слагаемых. Отсюда вытекает следующее, обычно применяемое, практическое правило для сложения приближенных чисел.

Правило. Чтобы сложить числа различной абсолютной точности, следует:

1) выделить числа, десятичная запись которых обрывается ранее других, и оставить их без изменения;

2) остальные числа округлить по образцу выделенных, сохраняя один или два запасных десятичных знака;

3) произвести сложение данных чисел, учитывая все сохраненные знаки;

4) полученный результат округлить на один знак.

При округлении по правилу дополнения слагаемых суммы

$$u = x_1 + x_2 + \ldots + x_n$$

до $m$-го десятичного разряда погрешность округления суммы в самом неблагоприятном случае не превышает величины

$$\Delta_{\text{окр}} \leqslant n \cdot \frac{1}{2} \cdot 10^m. \tag{3}$$

Можно получить более точный расчет погрешности округления суммы, если учесть знаки ошибок округления слагаемых.

Пример. Найти сумму приближенных чисел: 0,348; 0,1834; 345,4; 235,2; 11,75; 9,27; 0,0849; 0,0214; 0,000354, каждое из которых имеет все верные значащие цифры (в широком смысле).

Решение. Выделяем числа наименьшей точности 345,4 и 235,2, абсолютная погрешность которых может достигать 0,1. Округляя остальные числа с точностью до 0,01, получим:

$$\begin{array}{r} 345{,}4 \\ 235{,}2 \\ 11{,}75 \\ 9{,}27 \\ 0{,}35 \\ 0{,}18 \\ 0{,}08 \\ 0{,}02 \\ 0{,}00 \\ \hline 602{,}25 \end{array}$$

Округляя результат до 0,1 по правилу четной цифры, получим приближенное значение суммы 602,2.

Полная погрешность $\Delta$ результата складывается из трех слагаемых:

1) суммы предельных погрешностей исходных данных

$$\Delta_1 = 10^{-3} + 10^{-4} + 10^{-1} + 10^{-1} + 10^{-2} + 10^{-2} + 10^{-4} + \\ + 10^{-4} + 10^{-6} = 0{,}221\,301 < 0{,}222;$$

2) абсолютной величины суммы ошибок (с учетом их знаков) округления слагаемых

$$\Delta_2 = |-0{,}002 + 0{,}0034 + 0{,}0049 + 0{,}0014 + 0{,}000\,354| = \\ = 0{,}008054 < 0{,}009;$$

3) заключительной погрешности округления результата

$$\Delta_3 = 0{,}050.$$

Следовательно,

$$\Delta = \Delta_1 + \Delta_2 + \Delta_3 \leqslant 0{,}222 + 0{,}009 + 0{,}050 = 0{,}281 < 0{,}3;$$

и, таким образом, искомая сумма есть $602{,}2 \pm 0{,}3$.

Теорема 2. *Если слагаемые — одного и того же знака, то предельная относительная погрешность их суммы не превышает наибольшей из предельных относительных погрешностей слагаемых.*

Доказательство. Пусть $u = x_1 + x_2 + \ldots + x_n$, где, для определенности, $x_i > 0$ $(i = 1, 2, \ldots, n)$.

Обозначим через $A_i$ ($A_i > 0$; $i = 1, 2, \ldots, n$) точные величины слагаемых $x_i$, а через $A = A_1 + A_2 + \ldots + A_n$ — точное значение суммы $u$. Тогда за предельную относительную погрешность суммы можно принять:

$$\delta_u = \frac{\Delta_u}{A} = \frac{\Delta_{x_1} + \Delta_{x_2} + \ldots + \Delta_{x_n}}{A_1 + A_2 + \ldots + A_n}. \tag{4}$$

Так как

$$\delta_{x_i} = \frac{\Delta_{x_i}}{A_i} \quad (i = 1, 2, \ldots, n),$$

то

$$\Delta_{x_i} = A_i \delta_{x_i}. \tag{4'}$$

Подставляя это выражение в формулу (4), получим:

$$\delta_u = \frac{A_1\delta_{x_1} + A_2\delta_{x_2} + \ldots + A_n\delta_{x_n}}{A_1 + A_2 + \ldots + A_n}.$$

Пусть $\overline{\delta}$ является наибольшей из относительных погрешностей $\delta_{x_i}$, т. е. $\overline{\delta}_{x_i} \leqslant \overline{\delta}$. Тогда

$$\delta_u \leqslant \frac{\overline{\delta}(A_1 + A_2 + \ldots + A_n)}{A_1 + A_2 + \ldots + A_n} = \overline{\delta}.$$

Следовательно, $\delta_u \leqslant \overline{\delta}$, т. е.

$$\delta_u \leqslant \max(\delta_{x_1}, \delta_{x_2}, \ldots, \delta_{x_n}).$$

## § 8. Погрешность разности

Рассмотрим разность двух приближенных чисел $u = x_1 - x_2$.

По формуле (2) § 7 предельная абсолютная погрешность $\Delta_u$ разности

$$\Delta_u = \Delta_{x_1} + \Delta_{x_2},$$

т. е. *предельная абсолютная погрешность разности равна сумме предельных абсолютных погрешностей уменьшаемого и вычитаемого.*

Отсюда предельная относительная погрешность разности

$$\delta_u = \frac{\Delta_{x_1} + \Delta_{x_2}}{A}, \tag{1}$$

где $A$ — точное значение абсолютной величины разности чисел $x_1$ и $x_2$.

Замечание о потере точности при вычитании близких чисел. Если приближенные числа $x_1$ и $x_2$ достаточно близки друг к другу и имеют малые абсолютные погрешности, то число $A$ мало. Из формулы (1) вытекает, что предельная относительная погрешность в этом случае может быть весьма большой, в то

время как относительные погрешности уменьшаемого и вычитаемого остаются малыми, т. е. здесь происходит *потеря точности*.

Вычислим, например, разность двух чисел: $x_1 = 47{,}132$ и $x_2 = 47{,}111$, каждое из которых имеет пять верных знаков. Вычитая, получим $u = 47{,}132 - 47{,}111 = 0{,}021$.

Таким образом, разность $u$ имеет лишь две значащие цифры, из которых последняя сомнительна, так как предельная абсолютная погрешность разности

$$\Delta_u = 0{,}0005 + 0{,}0005 = 0{,}001.$$

Предельные относительные погрешности вычитаемого, уменьшаемого и разности соответственно

$$\delta_{x_1} = \frac{0{,}0005}{47{,}132} \approx 0{,}00001;$$

$$\delta_{x_2} = \frac{0{,}0005}{47{,}111} \approx 0{,}00001;$$

$$\delta_u = \frac{0{,}001}{0{,}021} \approx 0{,}05.$$

Предельная относительная погрешность разности здесь примерно в 5000 раз больше предельных относительных погрешностей исходных данных.

Поэтому при приближенных вычислениях полезно преобразовывать выражения, вычисление числовых значений которых приводит к вычитанию близких чисел.

Пример. Найти разность

$$u = \sqrt{2{,}01} - \sqrt{2} \tag{2}$$

с тремя верными знаками.

Решение. Так как

$$\sqrt{2{,}01} = 1{,}4177\,4469\ldots$$

и

$$\sqrt{2} = 1{,}4142\,1356\ldots,$$

то искомый результат есть

$$u = 0{,}00353 = 3{,}53 \cdot 10^{-3}.$$

Этот результат можно получить, если записать выражение (2) в виде

$$u = \frac{0{,}01}{\sqrt{2{,}01} + \sqrt{2}}$$

и взять корни $\sqrt{2{,}01}$ и $\sqrt{2}$ лишь с тремя верными знаками. Действительно,

$$u = \frac{0{,}01}{1{,}42 + 1{,}41} = \frac{0{,}01}{2{,}83} = 10^{-2} \cdot 3{,}53 \cdot 10^{-1} = 3{,}53 \cdot 10^{-3}.$$

Исходя из вышесказанного, получаем следующее практическое правило: при приближенных вычислениях следует по возможности избегать вычитания двух почти равных приближенных чисел; если же в силу необходимости приходится вычитать такие числа, то следует уменьшаемое и вычитаемое брать с достаточным числом запасных верных знаков (если такая возможность имеется). Например, если известно, что при вычитании чисел $x_1$ и $x_2$ первые $m$ значащих цифр их пропадут, а результат необходимо иметь с $n$ верными значащими цифрами, то следует взять $x_1$ и $x_2$ с $m+n$ верными значащими цифрами.

## § 9. Погрешность произведения

Теорема. *Относительная погрешность произведения нескольких приближенных чисел, отличных от нуля, не превышает суммы относительных погрешностей этих чисел.*

Доказательство. Пусть $u = x_1 x_2 \ldots x_n$.

Предполагая для простоты, что приближенные числа $x_1, x_2, \ldots, x_n$ положительны, будем иметь:

$$\ln u = \ln x_1 + \ln x_2 + \ldots + \ln x_n.$$

Отсюда, используя приближенную формулу $\Delta \ln x \approx d \ln x = \frac{\Delta x}{x}$, находим:

$$\frac{\Delta u}{u} = \frac{\Delta x_1}{x_1} + \frac{\Delta x_2}{x_2} + \ldots + \frac{\Delta x_n}{x_n}.$$

Оценивая последнее выражение по абсолютной величине, получим:

$$\left|\frac{\Delta u}{u}\right| \leqslant \left|\frac{\Delta x_1}{x_1}\right| + \left|\frac{\Delta x_2}{x_2}\right| + \ldots + \left|\frac{\Delta x_n}{x_n}\right|.$$

Если $A_i$ $(i = 1, 2, \ldots, n)$ — точные значения сомножителей $x_i$ и $|\Delta x_i|$, как это бывает обычно, малы по сравнению с $x_i$, то приближенно можно положить:

$$\left|\frac{\Delta x_i}{x_i}\right| \approx \left|\frac{\Delta x_i}{A_i}\right| = \delta_i$$

и

$$\left|\frac{\Delta u}{u}\right| = \delta,$$

где $\delta_i$ — относительные погрешности сомножителей $x_i$ $(i = 1, 2, \ldots, n)$ и $\delta$ — относительная погрешность произведения.

Следовательно,

$$\delta \leqslant \delta_1 + \delta_2 + \ldots + \delta_n. \qquad (1)$$

Формула (1), очевидно, остается верной также, если сомножители $x_i$ $(i = 1, 2, \ldots, n)$ имеют различные знаки.

Следствие. Предельная относительная погрешность произведения равна сумме предельных относительных погрешностей сомножителей, т. е.

$$\delta_u = \delta_{x_1} + \delta_{x_2} + \ldots + \delta_{x_n}. \qquad (2)$$

Если все множители произведения $u$ весьма точны, за исключением одного, то из формулы (2) следует, что предельная относительная погрешность произведения в этом случае будет практически совпадать с предельной относительной погрешностью множителя, обладающего наименьшей точностью. В частном случае, если приближенным является лишь множитель $x_1$, то имеем просто

$$\delta_u = \delta_{x_1}.$$

Зная предельную относительную погрешность $\delta_u$ произведения $u$, можно определить его предельную абсолютную погрешность $\Delta_u$ по формуле

$$\Delta_u = |u|\,\delta_u.$$

Пример 1. Определить произведение $u$ приближенных чисел $x_1 = 12{,}2$ и $x_2 = 73{,}56$ и число верных знаков в нем, если все написанные цифры сомножителей верные.

Решение. Имеем $\Delta_{x_1} = 0{,}05$ и $\Delta_{x_2} = 0{,}005$. Отсюда

$$\delta_u = \frac{0{,}05}{12{,}2} + \frac{0{,}005}{73{,}56} = 0{,}0042.$$

Так как произведение $u = 897{,}432$, то $\Delta_u = u\delta_u = 897 \cdot 0{,}004 = 3{,}6$ (приблизительно).

Отсюда $u$ имеет лишь два верных знака и результат следует записать так:

$$u = 897 \pm 4.$$

Отметим частный случай

$$u = kx,$$

где $k$ — точный множитель, отличный от нуля. Имеем:

$$\delta_u = \delta_x$$

и

$$\Delta_u = |k|\,\Delta_x,$$

т. е. *при умножении приближенного числа на точный множитель $k$ относительная предельная погрешность не изменяется, а абсолютная предельная погрешность увеличивается в $|k|$ раз.*

Пример 2. При наведении ракеты на цель предельная угловая ошибка $\varepsilon = 1'$. Каково возможное отклонение $\Delta_u$ ракеты от цели на дальности $x = 2000$ *км* при отсутствии корректирования ошибки?

Решение. Здесь

$$\Delta_u = \frac{\pi}{180 \cdot 60} \cdot 2000 \text{ км} \approx 580 \text{ м}.$$

Очевидно, что относительная погрешность произведения не может быть меньше, чем относительная погрешность наименее точного из сомножителей. Поэтому здесь, как и в случае сложения, не имеет смысла сохранять в более точных сомножителях излишнее количество значащих цифр.

Полезно руководствоваться следующим правилом: чтобы найти произведение нескольких приближенных чисел с различным числом верных значащих цифр, достаточно:

1) округлить их так, чтобы каждое из них содержало на одну (или две) значащую цифру больше, чем число верных цифр в наименее точном из сомножителей;

2) в результате умножения сохранить столько значащих цифр, сколько верных цифр имеется в наименее точном из сомножителей (или удержать еще одну запасную цифру).

Пример 3. Найти произведение приближенных чисел $x_1 = 2{,}5$ и $x_2 = 72{,}397$, верных в написанных знаках.

Решение. Применяя правило, после округления имеем $x_1 = 2{,}5$ и $x_2 = 72{,}4$. Отсюда $x_1 x_2 = 2{,}5 \cdot 72{,}4 = 181 \approx 1{,}8 \cdot 10^2$.

## § 10. Число верных знаков произведения

Пусть имеем произведение $n$ сомножителей ($n \leqslant 10$) $u = x_1 x_2 \ldots x_n$ каждый из которых имеет по крайней мере $m$ ($m > 1$) верных цифр. Пусть, далее, $\alpha_1, \alpha_2, \ldots, \alpha_n$ — первые значащие цифры в десятичной записи множителей:

$$x_i = \alpha_i 10^{p_i} + \beta_i 10^{p_i - 1} + \ldots \quad (i = 1, 2, 3, \ldots, n).$$

Тогда по формуле (5) § 5 будем иметь:

$$\delta_{x_i} = \frac{1}{2\alpha_i} \left(\frac{1}{10}\right)^{m-1} \quad (i = 1, 2, \ldots, n)$$

и, следовательно,

$$\delta_u = \frac{1}{2} \left(\frac{1}{\alpha_1} + \frac{1}{\alpha_2} + \ldots + \frac{1}{\alpha_n}\right) \left(\frac{1}{10}\right)^{m-1}. \tag{1}$$

Так как $\frac{1}{\alpha_1} + \frac{1}{\alpha_2} + \ldots + \frac{1}{\alpha_n} \leqslant 10$, то $\delta_u \leqslant \frac{1}{2} \left(\frac{1}{10}\right)^{m-2}$.

Следовательно, в самом неблагоприятном случае произведение $u$ имеет $m - 2$ верных знака.

Правило. Если все сомножители имеют $m$ верных десятичных знаков и число их не больше 10, то число верных (в широком смысле) знаков произведения на одну или на две единицы меньше $m$.

Следовательно, если нужно обеспечить в произведении $m$ верных десятичных знаков, то сомножители следует брать с одним или двумя запасными знаками.

Если сомножители обладают различной точностью, то под $m$ следует понимать число верных знаков в наименее точном из сомножителей. Таким образом, *число верных знаков произведения небольшого числа сомножителей (порядка десяти) может быть на одну или две единицы меньше числа верных знаков в наименее точном из этих сомножителей.*

Пример 1. Определить относительную погрешность и количество верных цифр произведения $u = 93{,}87 \cdot 9{,}236$.

Решение. По формуле (1) имеем:

$$\delta_u = \frac{1}{2}\left(\frac{1}{9} + \frac{1}{9}\right)\frac{1}{10^3} = \frac{1}{9} \cdot 10^{-3} < \frac{1}{2} \cdot 10^{-3}.$$

Следовательно, произведение $u$ имеет по меньшей мере три верные цифры (см. § 5).

Пример 2. Определить относительную погрешность и число верных цифр произведения $u = 17{,}63 \cdot 14{,}285$.

Решение.

$$\delta_u = \frac{1}{2}\left(\frac{1}{1} + \frac{1}{1}\right)\frac{1}{10^3} = 1 \cdot 10^{-3}.$$

Следовательно, в произведении будут по крайней мере три верные цифры (в широком смысле).

## § 11. Погрешность частного

Если $u = \frac{x}{y}$, то $\ln u = \ln x - \ln y$

и

$$\frac{\Delta u}{u} = \frac{\Delta x}{x} - \frac{\Delta y}{y}.$$

Отсюда

$$\left|\frac{\Delta u}{u}\right| \leqslant \left|\frac{\Delta x}{x}\right| + \left|\frac{\Delta y}{y}\right|.$$

Из последней формулы вытекает, что теорема § 9 верна и для частного.

Теорема. *Относительная погрешность частного не превышает суммы относительных погрешностей делимого и делителя.*

Следствие. Если $u = \frac{x}{y}$, то $\delta_u = \delta_x + \delta_y$.

Пример. Найти число верных знаков частного $u = 25{,}7 : 3{,}6$, если все написанные знаки делимого и делителя верны.

Решение. Имеем:

$$\delta_u = \frac{0{,}05}{25{,}7} + \frac{0{,}05}{3{,}6} = 0{,}002 + 0{,}014 = 0{,}016.$$

Так как $u=7{,}14$, то $\Delta_u=0{,}016\cdot 7{,}14=0{,}11$. Поэтому частное $u$ имеет два верных знака в широком смысле, т. е. $u=7{,}1$ или, более точно,

$$u=7{,}14\pm 0{,}11.$$

## § 12. Число верных знаков частного

Пусть делимое $x$ и делитель $y$ имеют по меньшей мере $m$ верных цифр. Если $\alpha$ и $\beta$ — их первые значащие цифры, то за предельную относительную погрешность частного $u$ может быть принята величина

$$\delta_u=\frac{1}{2}\left(\frac{1}{\alpha}+\frac{1}{\beta}\right)\left(\frac{1}{10}\right)^{m-1}.$$

Отсюда получаем правило: 1) если $\alpha\geqslant 2$ и $\beta\geqslant 2$, то частное $u$ имеет по меньшей мере $m-1$ верных знаков; 2) если $\alpha=1$ или $\beta=1$, то частное $u$ заведомо имеет $m-2$ верных знака.

## § 13. Относительная погрешность степени

Пусть $u=x^m$ ($m$ — натуральное число), тогда $\ln u=m\ln x$ и, следовательно,

$$\left|\frac{\Delta u}{u}\right|=m\left|\frac{\Delta x}{x}\right|.$$

Отсюда

$$\delta_u=m\delta_x, \tag{1}$$

*т. е. предельная относительная погрешность $m$-й степени числа в $m$ раз больше предельной относительной погрешности самого числа.*

## § 14. Относительная погрешность корня

Пусть теперь $u=\sqrt[m]{x}$, тогда $u^m=x$. Отсюда

$$\delta_u=\frac{1}{m}\,\delta_x, \tag{1}$$

*т. е. предельная относительная погрешность корня $m$-й степени в $m$ раз меньше предельной относительной погрешности подкоренного числа.*

Пример. Определить, с какой относительной погрешностью и со сколькими верными цифрами можно найти сторону $a$ квадрата, если его площадь $s=12{,}34$ (с точностью до 0,01).

Решение. Имеем $a=\sqrt{s}=3{,}5128\ldots$ Так как

$$\delta_s=\frac{0{,}01}{12{,}33}\approx 0{,}0008,$$

то $\delta_a = \frac{1}{2}\delta_s = 0{,}0004$. Поэтому

$$\Delta_a = 3{,}5128 \cdot 0{,}0004 = 1{,}4 \cdot 10^{-3}.$$

Отсюда число $a$ будет иметь примерно четыре верных знака (в широком смысле) и, следовательно, $a = 3{,}513$.

## § 15. Вычисления без точного учета погрешностей

В предыдущих параграфах мы указали способы оценки предельной абсолютной погрешности действий. При этом предполагалось, что абсолютные погрешности компонент усиливают друг друга, что практически бывает сравнительно редко.

При массовых вычислениях, когда не учитывают погрешность каждого отдельного результата, рекомендуется пользоваться следующими правилами подсчета цифр [6].

1. При сложении и вычитании приближенных чисел младший сохраненный десятичный разряд результата должен являться наибольшим среди десятичных разрядов, выражаемых последними верными значащими цифрами исходных данных.

2. При умножении и делении приближенных чисел в результате следует сохранять столько значащих цифр, сколько их имеет приближенное данное с наименьшим числом верных значащих цифр.

3. При возведении в квадрат или куб приближенного числа в результате нужно сохранять столько значащих цифр, сколько верных значащих цифр имеет основание степени.

4. При извлечении квадратного и кубического корней из приближенного числа в результате следует брать столько значащих цифр, сколько верных цифр имеет подкоренное число.

5. Во всех промежуточных результатах следует сохранять на одну цифру больше, чем рекомендуют предыдущие правила. В окончательном результате эта «запасная цифра» отбрасывается.

6. При вычислениях с помощью логарифмов рекомендуется подсчитать число верных значащих цифр в приближенном числе, имеющем наименьшее число верных значащих цифр, и воспользоваться таблицей логарифмов с числом десятичных знаков, на единицу бо́льшим. В окончательном результате последняя значащая цифра отбрасывается.

7. Если данные можно брать с произвольной точностью, то для получения результата с $k$ верными цифрами исходные данные следует брать с таким числом цифр, которые согласно предыдущим правилам обеспечивают $k+1$ верную цифру в результате.

Если некоторые данные имеют излишние младшие десятичные разряды (при сложении и вычитании) или больше значащих цифр, чем другие (при умножении, делении, возведении в степень, извлечении корня), то их предварительно нужно округлить, сохраняя одну запасную цифру.

## § 16. Общая формула для погрешности

Основная задача теории погрешности заключается в следующем: известны погрешности некоторой системы величин, требуется определить погрешность данной функции от этих величин.

Пусть задана дифференцируемая функция

$$u = f(x_1, x_2, \ldots, x_n)$$

и пусть $|\Delta x_i|$ $(i = 1, 2, \ldots, n)$ — абсолютные погрешности аргументов функции. Тогда абсолютная погрешность функции

$$|\Delta u| = |f(x_1 + \Delta x_1, x_2 + \Delta x_2, \ldots, x_n + \Delta x_n) - f(x_1, x_2, \ldots, x_n)|.$$

Обычно на практике $|\Delta x_i|$ — малые величины, произведениями, квадратами и высшими степенями которых можно пренебречь. Поэтому можно положить:

$$|\Delta u| \approx |df(x_1, x_2, \ldots, x_n)| = \left| \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \Delta x_i \right| \leqslant \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i} \right| |\Delta x_i|.$$

Итак,

$$|\Delta u| \leqslant \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i} \right| |\Delta x_i|. \tag{1}$$

Отсюда, обозначая через $\Delta_{x_i}$ $(i = 1, 2, \ldots, n)$ предельные абсолютные погрешности аргументов $x_i$ и через $\Delta_u$ — предельную погрешность функции $u$, для малых $\Delta x_i$ получим:

$$\Delta_u = \sum_{i=1}^{n} \left| \frac{\partial u}{\partial x_i} \right| \Delta_{x_i}. \tag{2}$$

Разделив обе части неравенства (1) на $u$, будем иметь оценку для относительной погрешности функции $u$

$$\delta \leqslant \sum_{i=1}^{n} \left| \frac{\frac{\partial f}{\partial x_i}}{u} \right| |\Delta x_i| = \sum_{i=1}^{n} \left| \frac{\partial}{\partial x_i} \ln f(x_1, \ldots, x_n) \right| |\Delta x_i|. \tag{3}$$

Следовательно, за предельную относительную погрешность функции $u$ можно принять:

$$\delta_u = \sum_{i=1}^{n} \left| \frac{\partial}{\partial x_i} \ln u \right| \Delta_{x_i}. \tag{4}$$

Пример 1. Найти предельные абсолютную и относительную погрешности объема шара $V = \frac{1}{6}\pi d^3$, если диаметр $d = 3{,}7$ *см* $\pm$ $\pm\, 0{,}05$ *см*, а $\pi \approx 3{,}14$.

Решение. Рассматривая $\pi$ и $d$ как переменные величины, вычисляем частные производные

$$\frac{\partial V}{\partial \pi} = \frac{1}{6} d^3 = 8{,}44;$$

$$\frac{\partial V}{\partial d} = \frac{1}{2} \pi d^2 = 21{,}5.$$

В силу формулы (2) предельная абсолютная погрешность объема

$$\Delta V = \left|\frac{\partial V}{\partial \pi}\right| |\Delta \pi| + \left|\frac{\partial V}{\partial d}\right| |\Delta d| = 8{,}44 \cdot 0{,}0016 + 21{,}5 \cdot 0{,}05 =$$
$$= 0{,}013 + 1{,}075 = 1{,}088 \ см^3 \approx 1{,}1 \ см^3.$$

Поэтому

$$V = \frac{1}{6} \pi d^3 \approx 27{,}4 \ см^3 \pm 1{,}1 \ см^3. \qquad (5)$$

Отсюда предельная относительная погрешность объема

$$\delta_V = \frac{1{,}088 \ см^3}{27{,}4 \ см^3} = 0{,}0397 \approx 4\%.$$

Пример 2. Для определения модуля Юнга $E$ по прогибу стержня прямоугольного сечения применяется формула

$$E = \frac{1}{4} \cdot \frac{l^3 p}{a^3 b s},$$

где $l$ — длина стержня, $a$ и $b$ — измерения поперечного сечения стержня, $s$ — стрела прогиба, $p$ — нагрузка.

Вычислить предельную относительную погрешность при определении модуля Юнга $E$, если $p = 20$ *кГ*; $\delta_p = 0{,}1\%$; $a = 3$ *мм*; $\delta_a = 1\%$; $b = 44$ *мм*; $\delta_b = 1\%$; $l = 50$ *см*; $\delta_l = 1\%$; $s = 2{,}5$ *см*; $\delta_s = 1\%$.

Решение. $\ln E = 3 \ln l + \ln p - 3 \ln a - \ln b - \ln s - \ln 4$.

Отсюда, заменяя приращения дифференциалами, будем иметь:

$$\frac{\Delta E}{E} = 3 \frac{\Delta l}{l} + \frac{\Delta p}{p} - 3 \frac{\Delta a}{a} - \frac{\Delta b}{b} - \frac{\Delta s}{s}.$$

Следовательно,

$$\delta_E = 3\delta_l + \delta_p + 3\delta_a + \delta_b + \delta_s = 3 \cdot 0{,}01 + 0{,}001 + 3 \cdot 0{,}01 +$$
$$+ 0{,}01 + 0{,}01 = 0{,}081.$$

Таким образом, предельная относительная погрешность составляет 0,081, т. е. примерно 8% от измеряемой величины.

Произведя численные расчеты, имеем:

$$E = (2{,}10 \pm 0{,}17) \cdot 10^6 \frac{кГ}{см^2}.$$

## § 17. Обратная задача теории погрешностей

На практике важна также обратная задача: каковы должны быть абсолютные погрешности аргументов функции, чтобы абсолютная погрешность функции не превышала заданной величины.

Эта задача математически неопределенна, так как заданную предельную погрешность $\Delta_u$ функции $u=f(x_1, x_2, \ldots, x_n)$ можно обеспечить, устанавливая по-разному предельные абсолютные погрешности $\Delta_{x_i}$ ее аргументов.

Простейшее решение обратной задачи дается так называемым *принципом равных влияний*. Согласно этому принципу предполагается, что все частные дифференциалы

$$\frac{\partial f}{\partial x_i}\Delta x_i \quad (i=1, 2, \ldots, n)$$

одинаково влияют на образование общей абсолютной погрешности $\Delta_u$ функции $u=f(x_1, x_2, \ldots, x_n)$.

Пусть величина предельной абсолютной погрешности $\Delta_u$ задана. Тогда на основании формулы (2) § 16

$$\Delta_u=\sum_{i=1}\left|\frac{\partial u}{\partial x_i}\right|\Delta_{x_i}. \tag{1}$$

Предполагая, что все слагаемые равны между собой, будем иметь

$$\left|\frac{\partial u}{\partial x_1}\right|\Delta_{x_1}=\left|\frac{\partial u}{\partial x_2}\right|\Delta_{x_2}=\ldots=\left|\frac{\partial u}{\partial x_n}\right|\Delta_{x_n}=\frac{\Delta_u}{n}.$$

Отсюда

$$\Delta_{x_i}=\frac{\Delta_u}{n\left|\frac{\partial u}{\partial x_i}\right|} \quad (i=1, 2, \ldots, n). \tag{2}$$

Пример 1. Радиус основания цилиндра $R\approx 2$ *м*; высота цилиндра $H\approx 3$ *м*. С какими абсолютными погрешностями нужно определить $R$ и $H$, чтобы его объем $V$ можно было вычислить с точностью до 0,1 $м^3$?

Решение. Имеем $V=\pi R^2H$ и $\Delta_V=0{,}1$ $м^3$.

Полагая $R=2$ *м*; $H=3$ *м*; $\pi=3{,}14$; приближенно получим:

$$\frac{\partial V}{\partial \pi}=R^2H=12;$$

$$\frac{\partial V}{\partial R}=2\pi RH=37{,}7;$$

$$\frac{\partial V}{\partial H}=\pi R^2=12{,}6.$$

Отсюда, так как $n=3$, то на основании формулы (2) будем иметь:

$$\Delta_\pi = \frac{0,1}{3\cdot 12} < 0,003;$$

$$\Delta_R = \frac{0,1}{3\cdot 37,7} < 0,001;$$

$$\Delta_H = \frac{0,1}{3\cdot 12,6} < 0,003.$$

Пример 2. Требуется найти значение функции

$$u = 6x^2 (\lg x - \sin 2y)$$

с точностью до двух десятичных знаков (после запятой), причем приближенные значения $x$ и $y$ равны соответственно 15,2 и 57°. Найти допустимую абсолютную погрешность этих величин.

Решение. Здесь

$$u = 6x^2 (\lg x - \sin 2y) = 6\,(15,2)^2 (\lg 15,2 - \sin 114°) = 371,9;$$

$$\frac{\partial u}{\partial x} = 12x (\lg x - \sin 2y) + 6xM = 88,54,$$

где $M = 0,43429$ — модуль перехода;

$$\frac{\partial u}{\partial y} = -12x^2 \cos 2y = 1127,7.$$

Для того чтобы результат был верен до двух десятичных знаков, нужно выполнение равенства $\Delta_u = 0,005$. Тогда по принципу равных влияний имеем:

$$\Delta_x = \frac{\Delta_u}{2\left|\frac{\partial u}{\partial x}\right|} = \frac{0,005}{2\cdot 88,54} = 0,000028;$$

$$\Delta_y = \frac{\Delta_u}{2\left|\frac{\partial u}{\partial y}\right|} = \frac{0,005}{2\cdot 1127,7} = 0,0000022 \text{ рад} = 0'',45.$$

Нередко при решении обратной задачи по принципу равных влияний мы можем столкнуться с таким случаем, когда найденные по формуле (2) предельные абсолютные погрешности отдельных независимых переменных окажутся настолько малыми, что добиться соответствующей точности при измерении этих величин практически невозможно. В таких случаях следует отступить от принципа равных влияний и за счет разумного уменьшения погрешностей одной части переменных добиться увеличения погрешностей другой части переменных.

Пример 3. С какой точностью надо измерить радиус круга $R = 30,5$ *см* и со сколькими знаками взять $\pi$, чтобы площадь круга была известна с точностью до 0,1%?

Решение. Имеем $s=\pi R^2$ и $\ln s=\ln\pi+2\ln R$. Отсюда

$$\frac{\Delta_s}{s}=\frac{\Delta_\pi}{\pi}+\frac{2\Delta_R}{R}=0{,}001.$$

По принципу равных влияний следует положить:

$$\frac{\Delta_\pi}{\pi}=0{,}0005;\quad \frac{2\Delta_R}{R}=0{,}0005.$$

Отсюда $\Delta_\pi\leqslant 0{,}0016$ и $\Delta_R\leqslant 0{,}00025R=0{,}0076$ *см*.

Таким образом, следовало бы взять $\pi=3{,}14$ и измерять $R$ с точностью до тысячных долей сантиметра. Ясно, что такая точность измерения практически трудноосуществима. Поэтому выгоднее поступить следующим образом: взять $\pi=3{,}142$; отсюда $\frac{\Delta_\pi}{\pi}=0{,}00013$; тогда $\frac{2\Delta_R}{R}=0{,}001-0{,}00013=0{,}0087$ и $\Delta_R\leqslant 0{,}13$ *см*. Такая точность достигается сравнительно легко.

Иногда допускают, что предельная абсолютная погрешность всех аргументов $x_i$ $(i=1,\ 2,\ \ldots,\ n)$ одна и та же. Тогда, полагая

$$\Delta_{x_1}=\Delta_{x_2}=\ldots=\Delta_{x_n},$$

из формулы (1) будем иметь:

$$\Delta_{x_i}=\frac{\Delta_u}{\sum\limits_{i=1}^{n}\left|\frac{\partial u}{\partial x_i}\right|}\quad (i=1,\ 2,\ \ldots,\ n).$$

Наконец, можно предположить, что точность измерения всех аргументов $x_i$ $(i=1,\ 2,\ \ldots,\ n)$ одинакова, т. е. предельные относительные погрешности $\delta_{x_i}$ $(i=1,\ 2,\ \ldots,\ n)$ аргументов равны между собой:

$$\delta_{x_1}=\delta_{x_2}=\ldots=\delta_{x_n}.$$

Отсюда получим:

$$\frac{\Delta_{x_1}}{|x_1|}=\frac{\Delta_{x_2}}{|x_2|}=\ldots=\frac{\Delta_{x_n}}{|x_n|}=k,$$

где $k$ — общее значение отношений.

Следовательно,

$$\Delta_{x_i}=k\,|x_i|\qquad (i=1,\ 2,\ \ldots,\ n).$$

Подставляя эти значения в формулу (1), находим:

$$\Delta_u=k\sum_{i=1}^{n}\left|x_i\frac{\partial u}{\partial x_i}\right|$$

и

$$k = \frac{\Delta_u}{\sum_{i=1}^{n} \left| x_i \frac{\partial u}{\partial x_i} \right|}.$$

Таким образом, окончательно имеем:

$$\Delta_{x_i} = \frac{|x_i| \Delta_u}{\sum_{j=1}^{n} \left| x_j \frac{\partial u}{\partial x_j} \right|} \quad (i = 1, 2, \ldots, n).$$

Можно также использовать и другие варианты.

Аналогично решается вторая обратная задача теории погрешности, когда задана предельная относительная погрешность функции и ищутся предельные абсолютные или относительные погрешности аргумента.

Иногда в самой постановке задачи имеются условия, не позволяющие использовать принцип равных влияний.

Пример 4. Стороны прямоугольника $a \approx 5$ *м* и $b \approx 200$ *м*. Какова допустимая предельная абсолютная погрешность при измерении этих сторон, одинаковая для обеих сторон, чтобы площадь $S$ прямоугольника можно было определить с предельной абсолютной погрешностью $\Delta_S = 1$ *м*$^2$?

Решение. Так как

$$S = ab,$$

то

$$\Delta S \approx b \Delta a + a \Delta b$$

и

$$\Delta_S = b\Delta_a + a\Delta_b.$$

Согласно условию задачи

$$\Delta_a = \Delta_b,$$

поэтому

$$\Delta_a = \frac{\Delta_S}{a+b} = \frac{1}{205} \approx 0{,}005 \text{ м} = 5 \text{ мм}.$$

## § 18. Точность определения аргумента для функции, заданной таблицей

В вычислительной практике часто возникает необходимость определить аргумент по значению функции, заданной таблицей. Например, постоянно встречается необходимость определить число по его табличному логарифму или угол по табличному значению какой-либо тригонометрической функции и т. п. Понятно, что погрешность функции вызывает погрешность в определении аргумента.

Пусть имеем таблицу с одним входом для функции $y=f(x)$.

Если функция $f(x)$ дифференцируема, то для достаточно малых значений $|\Delta x|$ имеем:

$$|\Delta y|=|f'(x)||\Delta x|.$$

Отсюда

$$|\Delta x|=\frac{|\Delta y|}{|f'(x)|}, \tag{1}$$

или

$$\Delta_x=\frac{1}{|y'|}\Delta_y.$$

Применим формулу (1) к некоторым наиболее распространенным табулированным функциям.

## А. Логарифмы

Пусть $y=\ln x$, тогда $y'=\frac{1}{x}$.

Отсюда

$$\Delta_x=x\Delta_y. \tag{2}$$

Если же $y=\lg x$, то $y'=\frac{M}{x}$, где $M=0{,}43429$;

$$\Delta_x=\frac{1}{M}x\Delta_y=2{,}30\,x\Delta_y. \tag{2'}$$

Отсюда, в частности, получаем $\delta_x=2{,}30\Delta_y$, т. е. предельная относительная погрешность числа в таблице десятичных логарифмов равна примерно $2\frac{1}{2}$-кратной предельной абсолютной погрешности логарифма этого числа.

## Б. Тригонометрические функции

1. Если $y=\sin x\left(0<x<\frac{\pi}{2}\right)$, то $y'=\cos x$ и, следовательно,

$$\Delta_x=\Delta_y\sec x \text{ рад}. \tag{3}$$

2. Для функции

$$y=\operatorname{tg} x\left(0<x<\frac{\pi}{2}\right)$$

имеем

$$y'=\sec^2 x$$

и

$$\Delta_x=\Delta_y\cos^2 x \text{ рад}. \tag{4}$$

3. Если $y=\lg(\sin x)\left(0<x<\frac{\pi}{2}\right)$, то

$$y'=M\operatorname{ctg} x \text{ и } \Delta_x=2{,}30\operatorname{tg} x\Delta_y \text{ рад}. \tag{5}$$

4. Положим $y = \lg(\operatorname{tg} x)\left(0 < x < \frac{\pi}{2}\right)$, тогда

$$y' = \frac{2M}{\sin 2x} \text{ и } \Delta_x = 1{,}15 \sin 2x \Delta_y \text{ рад.} \tag{6}$$

Так как, очевидно, $\frac{\sin 2x}{2} < \operatorname{tg} x$ при $0 < x < \frac{\pi}{2}$, то из формул (5) и (6) следует, что угол $x$ по таблице логарифмов тангенсов определяется точнее, чем по таблице логарифмов синусов.

### В. Показательная функция

Если $y = e^x$, то $y' = e^x$ и

$$\Delta_x = \frac{\Delta_y}{e^x} \tag{7}$$

или

$$\Delta_x = \frac{\Delta_y}{y}.$$

Пример 1. С какой точностью можно определить число $x \approx 5000$, пользуясь четырехзначной таблицей десятичных логарифмов?

Решение. По формуле (2′) получаем:

$$\Delta_x = 2{,}30 \cdot 5000 \cdot \frac{1}{2} \cdot 10^{-4} \approx 0{,}6,$$

т. е. число $x$ имеет примерно четыре верные цифры.

Пример 2. Найти погрешность в определении угла $x \approx 60°$:

а) по пятизначной таблице логарифмов синусов,

б) по пятизначной таблице логарифмов тангенсов.

Решение. Для первого случая по формуле (5) имеем:

$$\Delta_x = 2{,}30 \cdot \sqrt{3} \cdot \frac{1}{2} \cdot 10^{-5} \text{ рад} = 0{,}00002 \text{ рад} \approx 4''.$$

Во втором случае по формуле (6) получаем:

$$\Delta_x = 1.15 \cdot \sqrt{3} \cdot \frac{1}{2} \cdot 10^{-5} \text{ рад} \approx 0{,}000005 \text{ рад} \approx 1'',$$

т. е. погрешность в четыре раза меньше.

## § 19. Способ границ

Обычно применяемая оценка погрешности функции (§ 16, формула (2)) является приближенной, так как эта оценка основана на пренебрежении произведениями ошибок. В некоторых случаях требуется иметь точные границы для искомого значения функции если известны границы изменения ее аргументов. Проще всего,

этого можно добиться, используя *способ двойных вычислений*, иначе называемый *способом границ*.

Пусть

$$u=f(x_1,\ x_2,\ \ldots,\ x_n)$$

— непрерывно дифференцируемая функция, монотонная по каждому аргументу $x_i$ $(i=1,\ 2,\ \ldots,\ n)$. Для этого достаточно предположить, что производные $\frac{\partial f}{\partial x_i}$ $(i=1,\ 2,\ \ldots,\ n)$ сохраняют постоянный знак в рассматриваемой области $\omega$ изменения аргументов. Допустим, что

$$\underline{x}_i < x_i < \overline{x}_i \quad (i=1,\ 2,\ \ldots,\ n), \tag{1}$$

причем параллелепипед (1) целиком принадлежит области $\omega$.

Положим, что $\tilde{x}_i=\underline{x}_i$, $\hat{x}_i=\overline{x}_i$, если функция $f$ — возрастающая по переменному $x_i$, и $\tilde{x}_i=\overline{x}_i$, $\hat{x}_i=\underline{x}_i$, если функция $f$ — убывающая по переменному $x_i$.

Тогда, очевидно,

$$\underline{u} < u < \overline{u}, \tag{2}$$

где

$$\underline{u}=f(\tilde{x}_1,\ \tilde{x}_2,\ \ldots,\ \tilde{x}_n)$$

и

$$\overline{u}=f(\hat{x}_1,\ \hat{x}_2,\ \ldots,\ \hat{x}_n).$$

Заметим, что переменные $\tilde{x}_i$ $(i=1,\ 2,\ \ldots,\ n)$ и результат действий $f$ над ними можно округлять лишь в сторону уменьшения величины $\underline{u}$, а переменные $\hat{x}_i$ $(i=1,\ 2,\ \ldots,\ n)$ и результат действий $f$ над ними можно округлять лишь в сторону увеличения величины $\overline{u}$. При этих обстоятельствах будет гарантировано строгое выполнение неравенства (2). В частном случае, если функция $f$ — монотонно возрастающая по каждому аргументу $x_i$ $(i=1,\ 2,\ \ldots,\ n)$, то имеем просто

$$f(\underline{x}_1,\ \underline{x}_2,\ \ldots,\ \underline{x}_n) < u < f(\overline{x}_1,\ \overline{x}_2,\ \ldots,\ \overline{x}_n). \tag{3}$$

Пример. Алюминиевый цилиндр с диаметром основания $d=2$ *см* $\pm 0{,}01$ *см* и высотой $h=11$ *см* $\pm 0{,}02$ *см* весит $p=93{,}4$ *г* $\pm$ $\pm 0{,}001$ *г*. Определить удельный вес $\gamma$ алюминия и оценить его предельную абсолютную погрешность.

Решение. Объем цилиндра равен

$$v=\frac{\pi d^2}{4}h;$$

отсюда

$$\gamma=\frac{p}{v}=\frac{4p}{\pi d^2 h}. \tag{4}$$

Из формулы (4) вытекает, что в области $p>0$, $d>0$, $h>0$ функция $\gamma$ — возрастающая по аргументу $p$ и убывающая по аргументам $d$ и $h$. Согласно условию задачи имеем:

$$1{,}99\ см \leqslant d \leqslant 2{,}01\ см;$$
$$10{,}98\ см \leqslant h \leqslant 11{,}02\ см;$$
$$93{,}399\,Г \leqslant p \leqslant 93{,}401\,Г.$$

Кроме того,

$$3{,}14159 < \pi < 3{,}1416.$$

Поэтому

$$\underline{\gamma} = \frac{4\cdot 93{,}399}{3{,}1416\cdot 2{,}01^2\cdot 11{,}02} = 2{,}671\frac{Г}{см^3}$$

(с недостатком) и

$$\overline{\gamma} = \frac{4\cdot 93{,}401}{3{,}14159\cdot 1{,}99^2\cdot 10{,}98} = 2{,}735\frac{Г}{см^3}$$

(с избытком). Взяв среднее арифметическое, получим:

$$\gamma = 2{,}703\frac{Г}{см^3} \pm 0{,}027\frac{Г}{см^3}, \tag{5}$$

или после округления

$$\gamma = 2{,}70\frac{Г}{см^3} \pm 0{,}03\frac{Г}{см^3}.$$

Для сравнения приведем приближенную оценку погрешности. Используя средние значения аргументов, получим:

$$\gamma = \frac{4\cdot 93{,}4}{3{,}1416\cdot 2^2\cdot 11} = 2{,}703\frac{Г}{см^3}.$$

Логарифмируя формулу (4), имеем:

$$\ln\gamma = \ln 4 + \ln p - \ln\pi - 2\ln d - \ln h;$$

отсюда, взяв полный дифференциал, получим:

$$\frac{\Delta\gamma}{\gamma} = \frac{\Delta p}{p} - \frac{\Delta\pi}{\pi} - \frac{2\,\Delta d}{d} - \frac{\Delta h}{h}.$$

Следовательно,

$$\delta_\gamma = \delta_p + \delta_\pi + 2\delta_d + \delta_h = \frac{0{,}001}{93{,}4} + \frac{0{,}00001}{3{,}1416} + \frac{2\cdot 0{,}01}{2} + \frac{0{,}02}{11} =$$
$$= 1{,}07\cdot 10^{-5} + 3{,}18\cdot 10^{-6} + 10^{-2} + 1{,}82\cdot 10^{-3} = 1{,}183\cdot 10^{-2}.$$

Далее, находим:

$$\Delta_\gamma = \delta_\gamma\cdot\gamma = 1{,}183\cdot 10^{-2}\cdot 2{,}703 = 3{,}2\cdot 10^{-2}\frac{Г}{см^3}.$$

Таким образом, приближенно имеем:

$$\gamma = 2{,}703\,\frac{г}{см^3} \pm 0{,}032\,\frac{г}{см^3},$$

что очень близко совпадает с точной оценкой (5).

## § 20*. Понятие о вероятностной оценке погрешности

Пусть имеем сумму $n$ слагаемых

$$u = x_1 + x_2 + \ldots + x_n.$$

Тогда предельная абсолютная погрешность суммы, как известно, равна

$$\Delta_u = \Delta_{x_1} + \Delta_{x_2} + \ldots + \Delta_{x_n}. \qquad (1)$$

Отсюда в случае, когда предельные абсолютные погрешности слагаемых одинаковы,

$$\Delta_{x_1} = \Delta_{x_2} = \ldots = \Delta_{x_n} = \Delta,$$

будем иметь:

$$\Delta_u = n\Delta. \qquad (1')$$

Формула (1) дает максимальное возможное значение абсолютной погрешности суммы. Эта предельная погрешность достигается лишь тогда, когда ошибки всех слагаемых: 1) наибольшие из возможных и 2) имеют одинаковые знаки. При большом количестве слагаемых такое неблагоприятное стечение обстоятельств является маловероятным. Фактически ошибки отдельных слагаемых, как правило, имеют различные знаки и, следовательно, частично компенсируют друг друга. Поэтому наряду с теоретической предельной погрешностью суммы $\Delta_u$ вводят *практическую предельную погрешность* $\Delta_u^*$, реализуемую с некоторой мерой достоверности.

Ограничимся рассмотрением простейшего случая. Пусть абсолютные погрешности $\Delta x_i$ ($i = 1, 2, \ldots, n$) слагаемых суммы (1) независимы и подчиняются нормальному закону с одной и той же мерой точности. Положим, что с вероятностью, превышающей число $\gamma$, абсолютные погрешности слагаемых не превышают числа $\Delta$, т. е.

$$\mathrm{P}(|\Delta x_i| \leqslant \Delta) > \gamma.$$

При этом условии в теории вероятностей доказывается, что с той же мерой достоверности абсолютная погрешность суммы $u$ будет удовлетворять неравенству $|\Delta u| \leqslant \Delta\sqrt{n}$, где $n$ — число слагаемых.

Таким образом, за предельную абсолютную погрешность суммы можно принять число

$$\Delta_u^* = \Delta\sqrt{n}. \qquad (2)$$

Например, складывая 100 чисел с абсолютной погрешностью 0,1, мы получим теоретическую предельную ошибку суммы $\Delta_u = 0{,}1 \cdot 100 = 10$. Фактически же можно ожидать, что эта ошибка не превзойдет величины $0{,}1 \cdot 10 = 1$.

В частности, рассмотрим среднее арифметическое $n$ чисел

$$\xi = \frac{1}{n}(x_1 + x_2 + \ldots + x_n).$$

Согласно строгой теории предельная абсолютная ошибка

$$\Delta_\xi = \frac{1}{n} \cdot n\Delta = \Delta;$$

тогда как с бо́льшей степенью достоверности можно утверждать, что практически

$$\Delta_\xi^* = \frac{\Delta\sqrt{n}}{n} = \frac{\Delta}{\sqrt{n}},$$

т. е. практически достоверно, *что среднее арифметическое приближенных чисел имеет повышенную точность по сравнению с этими числами, причем*

$$\Delta_\xi^* \to 0 \text{ при } n \to \infty.$$

Аналогично для случая умножения $n$ сомножителей с одинаковой относительной предельной погрешностью $\delta$ можно доказать, что практическая предельная относительная погрешность произведения определяется формулой

$$\delta_u^* = \delta\sqrt{n}. \tag{3}$$

## Литература к первой главе

1. А. Н. Крылов, Лекции о приближенных вычислениях, Изд. 2, АН СССР, Л., 1933, гл. I
2. Д. А. Вентцель, Е. С. Вентцель, Элементы теории приближенных вычислений, Изд. ВВИА им. Н. Е. Жуковского, М., 1949, гл. I.
3. Дж. Скарборо, Численные методы математического анализа, ГТТИ, 1934, гл. I.
4. Я. С. Безикович, Приближенные вычисления, Гостехиздат, 1949, гл. I и II.
5. Г. М. Фихтенгольц, Математика для инженеров, ГТТИ, 1933, ч. 1, гл. I.
6. В. М. Брадис, Устный и письменный счет. Вспомогательные средства вычислений. Энциклопедия элементарной математики, кн. 1, Учпедгиз, 1951.

## ГЛАВА II

# НЕКОТОРЫЕ СВЕДЕНИЯ ИЗ ТЕОРИИ ЦЕПНЫХ ДРОБЕЙ

### § 1. Определение цепной дроби

Выражение вида

$$a_0+\cfrac{b_1}{a_1+\cfrac{b_2}{a_2+\cfrac{b_3}{a_3+\dots}}}=\left[a_0;\ \frac{b_1}{a_1},\ \frac{b_2}{a_2},\ \frac{b_3}{a_3},\ \dots\right] \tag{1}$$

называется *цепной* или *непрерывной* дробью. Для цепной дроби (1) употребляется также сокращенная запись

$$a_0+\frac{b_1|}{|a_1}+\frac{b_2|}{|a_2}+\dots$$

В общем случае *элементы* цепной дроби $a_0$, $a_k$, $b_k$ $(k=1, 2, \dots)$ — вещественные или комплексные числа, или функции одной или нескольких переменных. Дроби $a_0=\frac{a_0}{1}$, $\frac{b_k}{a_k}$ $(k=1,\ 2,\ \dots)$ называются *звеньями* цепной дроби (1) (соответственно нулевым, первым и т. д.), а числа или функции $a_k$ и $b_k$ $(k\geqslant 1)$ — *членами* $k$-го звена (частными знаменателями или числителями). Мы будем предполагать, что $a_k\neq 0$. Заметим, что в сокращенной записи (1) звенья $\frac{b_k}{a_k}$ сокращать нельзя.

Если цепная дробь (1) содержит конечное число звеньев (например, $n$, не считая нулевого), то она называется *конечной* или *$n$-звенной* и сокращенно обозначается следующим образом:

$$\left[a_0;\ \frac{b_1}{a_1},\ \frac{b_2}{a_2},\ \dots,\ \frac{b_n}{a_n}\right]=\left[a_0;\ \frac{b_k}{a_k}\right]_1^n. \tag{2}$$

Конечная цепная дробь отождествляется с соответствующей обыкновенной, полученной в результате выполнения указанных действий. Цепная дробь (1), имеющая бесконечное множество звеньев,

называется *бесконечной*, причем употребляется обозначение

$$\left[a_0;\ \frac{b_k}{a_k}\right]_1^{\infty}. \tag{3}$$

Цепная дробь

$$a_0+\cfrac{1}{a_1+\cfrac{1}{a_2+\ddots}}=\left[a_0;\ \frac{1}{a_1},\ \frac{1}{a_2},\ \ldots\right], \tag{4}$$

у которой все частные числители равны 1, называется *обыкновенной* или стандартной *цепной дробью*. *Знаменатели звеньев* называются *неполными частными*. Заметим, что в теории чисел неполными частными обычно являются натуральные числа, т. е. целые положительные.

## § 2. Обращение цепной дроби в обыкновенную и обратно

Всякую конечную цепную дробь можно обратить в обыкновенную. Для этого достаточно произвести все действия, указанные в изображении цепной дроби.

Пример 1. Обратить цепную дробь

$$\left[3;\ \frac{1}{3},\ \frac{1}{1},\ \frac{1}{4}\right]=3+\cfrac{1}{3+\cfrac{1}{1+\cfrac{1}{4}}}$$

в обыкновенную.

Решение. Последовательно выполняя указанные действия, находим:

$$1)\ 1+\frac{1}{4}=\frac{5}{4};\quad 4)\ 1:\frac{19}{5}=\frac{5}{19};$$

$$2)\ 1:\frac{5}{4}=\frac{4}{5};\quad 5)\ 3+\frac{5}{19}=\frac{62}{19}.$$

$$3)\ 3+\frac{4}{5}=\frac{19}{5};$$

Следовательно,

$$\left[3;\ \frac{1}{3},\ \frac{1}{1},\ \frac{1}{4}\right]=3\frac{5}{19}.$$

Обратно всякое положительное рациональное число можно обратить в цепную дробь с натуральными элементами. Пусть, например, дана дробь $\frac{p}{q}$. Исключив из нее целую часть $a_0$, будем иметь:

$$\frac{p}{q}=a_0+\frac{r_0}{q},$$

где $r_0$ — остаток $\left(\text{если } \frac{p}{q} \text{ — правильная дробь, то } a_0=0 \text{ и } r_0=p\right)$.

Разделив теперь числитель и знаменатель дроби $\frac{r_0}{q}$ на $r_0$, получим:

$$\frac{r_0}{q}=\frac{1}{q:r_0}=\frac{1}{a_1+\frac{r_1}{r_0}},$$

где $a_1$ — целое частное, $r_1$ — остаток от деления $q$ на $r_0$.

Разделив числитель и знаменатель дроби $\frac{r_1}{r_0}$ на $r_1$, получим:

$$\frac{r_1}{r_0}=\frac{1}{r_0:r_1}=\frac{1}{a_2+\frac{r_2}{r_1}},$$

где $a_2$ — целое частное, $r_2$ — остаток от деления $r_0$ на $r_1$. Аналогично можно продолжить процесс дальше.

Так как $q>r_0>r_1>r_2>r_3>\ldots$ и $r_i$ $(i=0, 1, 2, \ldots)$ — целые положительные числа, то в конце концов мы будем иметь остаток $r_n=0$, т. е.

$$\frac{r_{n-1}}{r_{n-2}}=\frac{1}{a_n+0}.$$

Подставляя выражение дробей $\frac{r_i}{r_{i-1}}$, будем иметь:

$$\frac{p}{q}=a_0+\frac{r_0}{q}=a_0+\frac{1}{a_1+\frac{r_1}{r_0}}=a_0+\frac{1}{a_1+\frac{1}{a_2+\frac{r_2}{r_1}}}=$$

$$=a_0+\cfrac{1}{a_1+\cfrac{1}{a_2+\cdots+\cfrac{1}{a_n}}}.$$

Пример 2. Обратить $\frac{62}{19}$ в цепную дробь.

Решение. Последовательно имеем:

$$\frac{62}{19}=3+\frac{5}{19}=3+\frac{1}{\frac{19}{5}}=3+\frac{1}{3+\frac{4}{5}}=$$

$$=3+\frac{1}{3+\frac{1}{\frac{5}{4}}}=3+\frac{1}{3+\frac{1}{1+\frac{1}{4}}}.$$

Таким образом, $\frac{62}{19}=\left[3;\ \frac{1}{3},\ \frac{1}{1},\ \frac{1}{4}\right]$.

Аналогично преобразуются общие цепные дроби.

Пример 3. Обратить цепную дробь

$$\left[1;\ \frac{-x^2}{1},\ \frac{-x^2}{3},\ \frac{-x^2}{5}\right]=1-\cfrac{x^2}{1-\cfrac{x^2}{3-\cfrac{x^2}{5}}}$$

в обыкновенную.

Решение. Имеем:

1) $1-\cfrac{x^2}{3-\cfrac{x^2}{5}}=1-\frac{5x^2}{15-x^2}=\frac{15-6x^2}{15-x^2}$;

2) $1-\cfrac{x^2}{\cfrac{15-6x^2}{15-x^2}}=1-\frac{15x^2-x^4}{15-6x^2}=\frac{15-21x^2+x^4}{15-6x^2}$.

Таким образом,

$$\left[1;\ \frac{-x^2}{1},\ \frac{-x^2}{3},\ \frac{-x^2}{5}\right]=\frac{15-21x^2+x^4}{15-6x^2}.$$

## § 3. Подходящие дроби

Пусть дана конечная или бесконечная цепная дробь

$$\left[a_0;\ \frac{b_k}{a_k}\right]_1^n. \tag{1}$$

Обыкновенную дробь

$$\frac{P_k}{Q_k}\equiv\left[a_0;\ \frac{b_1}{a_1},\ \ldots,\ \frac{b_k}{a_k}\right]$$

($k=1,\ 2,\ \ldots$), где $k\leqslant n$, называют $k$-й *подходящей дробью* цепной дроби (1). Следуя Эйлеру, обычно полагают:

$$\frac{P_0}{Q_0}=\frac{a_0}{1};\qquad \frac{P_{-1}}{Q_{-1}}=\frac{1}{0},$$

причем для определенности считают, что

$$P_0=a_0,\qquad Q_0=1 \tag{2}$$

и

$$P_{-1}=1,\qquad Q_{-1}=0. \tag{2'}$$

При работе на электронной цифровой машине подходящие цепные дроби

$$\frac{P_n}{Q_n}=a_0+\cfrac{b_1}{a_1+\cfrac{b_2}{a_2+\cdots+\cfrac{b_n}{a_n}}}$$

удобно находить с помощью *схемы Горнера* (см. гл. III) *для деления*

$$c_1 = \frac{b_n}{a_n}, \qquad d_1 = a_{n-1} + c_1;$$
$$c_2 = \frac{b_{n-1}}{d_1}, \qquad d_2 = a_{n-2} + c_2;$$
$$\cdots\cdots\cdots\cdots$$
$$c_k = \frac{b_{n-k+1}}{d_{k-1}}, \qquad d_k = a_{n-k} + c_k;$$
$$\cdots\cdots\cdots\cdots$$
$$c_n = \frac{b_1}{d_{n-1}}, \qquad d_n = a_0 + c_n = \frac{P_n}{Q_n}.$$

Указанная последовательность действий легко программируется.

Теорема 1. (Закон составления подходящих дробей). *Числа* $P_k$, $Q_k$ $(k = -1, 0, 1, 2, \ldots)$, *определяемые из соотношений*

$$P_k = a_k P_{k-1} + b_k P_{k-2}, \tag{3}$$
$$Q_k = a_k Q_{k-1} + b_k Q_{k-2}, \tag{3'}$$

*где*

$$P_{-1} = 1, \quad Q_{-1} = 0; \quad P_0 = a_0, \quad Q_0 = 1, \tag{4}$$

*являются соответственно числителями и знаменателями подходящих дробей* $\frac{P_k}{Q_k}$ *цепной дроби* (1) *).

Доказательство. Пусть $R_k$ $(k = 1, 2, \ldots)$ — последовательные подходящие дроби цепной дроби (1). Нужно доказать, что

$$R_k = \frac{P_k}{Q_k} \qquad (k = 1, 2, \ldots).$$

Доказательство будем проводить методом математической индукции.

При $k = 1$ для подходящей дроби $R_1$ имеем:

$$R_1 = a_0 + \frac{b_1}{a_1} = \frac{a_0 a_1 + b_1}{a_1}.$$

С другой стороны, из соотношений (3) и (3′), учитывая (4), находим:

$$P_1 = a_1 a_0 + b_1,$$
$$Q_1 = a_1 \cdot 1 + b_1 \cdot 0 = a_1.$$

Следовательно, $R_1 = \frac{P_1}{Q_1}$ и для $k = 1$ утверждение теоремы справедливо.

---

*) Такие подходящие дроби будем называть *каноническими*.

Пусть теперь теорема верна для всех натуральных чисел, не превышающих $k$. Покажем, что теорема справедлива также для очередного натурального числа $k+1$. Из соотношений (3) и (3′) получаем:

$$P_{k+1}=a_{k+1}P_k+b_{k+1}P_{k-1},$$

$$Q_{k+1}=a_{k+1}Q_k+b_{k+1}Q_{k-1}.$$

Согласно индукционному предположению имеем:

$$R_k=\frac{P_k}{Q_k}=\frac{a_kP_{k-1}+b_kP_{k-2}}{a_kQ_{k-1}+b_kQ_{k-2}}.$$

По способу составления цепной дроби (1), подходящая дробь $R_{k+1}$ получается из подходящей дроби $R_k$ путем замены члена $a_k$ на сумму $a_k+\frac{b_{k+1}}{a_{k+1}}$. Поэтому

$$R_{k+1}=\frac{\left(a_k+\frac{b_{k+1}}{a_{k+1}}\right)P_{k-1}+b_kP_{k-2}}{\left(a_k+\frac{b_{k+1}}{a_{k+1}}\right)Q_{k-1}+b_kQ_{k-2}}=$$

$$=\frac{a_{k+1}(a_kP_{k-1}+b_kP_{k-2})+b_{k+1}P_{k-1}}{a_{k+1}(a_kQ_{k-1}+b_kQ_{k-2})+b_{k+1}Q_{k-1}}=\frac{a_{k+1}P_k+b_{k+1}P_{k-1}}{a_{k+1}Q_k+b_{k+1}Q_{k-1}}=\frac{P_{k+1}}{Q_{k+1}},$$

что и требовалось доказать.

Замечание. Так как члены подходящих дробей определяются неоднозначно, то в общем случае нельзя утверждать, что числитель и знаменатель подходящих дробей неканонического вида удовлетворяют уравнениям (3) и (3′). В дальнейшем мы будем предполагать, что рассматриваемые подходящие дроби являются каноническими.

Следствие. Для обыкновенной цепной дроби

$$a_0+\cfrac{1}{a_1+\cfrac{1}{a_2+\cdots}}$$

числители и знаменатели ее подходящих дробей $\frac{p_k}{q_k}$ $(k=1,\ 2,\ \ldots)$ могут быть определены из соотношений

$$\left.\begin{aligned}p_k&=a_kp_{k-1}+p_{k-2},\\ q_k&=a_kq_{k-1}+q_{k-2},\end{aligned}\right\}\qquad(3'')$$

где положено $p_0=a_0$, $p_{-1}=1$ и $q_0=1$, $q_{-1}=0$.

Замечание. Для нахождения по формулам (3) и (3′) членов последовательных подходящих дробей удобно применять следующую схему:

| $k$ | $-1$ | $0$ | $1$ | $2$ | $3$ | ... |
|---|---|---|---|---|---|---|
| $b_k$ | | $1$ | $b_1$ | $b_2$ | $b_3$ | ... |
| $a_k$ | | $a_0$ | $a_1$ | $a_2$ | $a_3$ | ... |
| $P_k$ | $1$ | $a_0$ | $P_1$ | $P_2$ | $P_3$ | ... |
| $Q_k$ | $0$ | $1$ | $Q_1$ | $Q_2$ | $Q_3$ | ... |

Для обыкновенной цепной дроби, где $b_k=1$ ($k=1, 2, \ldots$) и имеют место формулы (3″), в схеме опускают строку $b_k$.

Пример 1. Для цепной дроби

$$\frac{163}{59}=2+\cfrac{1}{1+\cfrac{1}{3+\cfrac{1}{4+\cfrac{1}{1+\cfrac{1}{2}}}}}$$

вычислить все подходящие дроби.

Решение. Используя схему, получим:

| $a_k$ | | $2$ | $1$ | $3$ | $4$ | $1$ | $2$ |
|---|---|---|---|---|---|---|---|
| $p_k$ | $p_{-1}=1$ | $2$ | $3$ | $11$ | $47$ | $58$ | $163$ |
| $q_k$ | $q_{-1}=0$ | $1$ | $1$ | $4$ | $17$ | $21$ | $59$ |

Таким образом,

$$\frac{p_0}{q_0}=\frac{2}{1};\ \frac{p_1}{q_1}=\frac{3}{1};\ \frac{p_2}{q_2}=\frac{11}{4};$$
$$\frac{p_3}{q_3}=\frac{47}{17};\ \frac{p_4}{q_4}=\frac{58}{21};\ \frac{p_5}{q_5}=\frac{163}{59}.$$

Пример 2. Найти все подходящие дроби для общей цепной дроби

$$\left[0;\ \frac{1}{2},\ \frac{3}{4},\ \frac{5}{8},\ \frac{7}{16}\right].$$

Решение. Применяя указанную выше схему, имеем:

| $k$ | $-1$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|
| $b_k$ | | 1 | 1 | 3 | 5 | 7 |
| $a_k$ | | 0 | 2 | 4 | 8 | 16 |
| $P_k$ | 1 | 0 | 1 | 4 | 37 | 620 |
| $Q_k$ | 0 | 1 | 2 | 11 | 98 | 1645 |

Отсюда

$$\frac{P_0}{Q_0}=\frac{0}{1};\quad \frac{P_1}{Q_1}=\frac{1}{2};\quad \frac{P_2}{Q_2}=\frac{4}{11};\quad \frac{P_3}{Q_3}=\frac{37}{98};\quad \frac{P_4}{Q_4}=\frac{620}{1645}.$$

Теорема 2. *Для двух соседних подходящих дробей* $\frac{P_{k-1}}{Q_{k-1}}$ *и* $\frac{P_k}{Q_k}$ *цепной дроби (1) справедлива формула*

$$\frac{P_k}{Q_k}-\frac{P_{k-1}}{Q_{k-1}}=(-1)^{k-1}\frac{b_1b_2\ldots b_k}{Q_{k-1}Q_k}\qquad (k\geqslant 1). \tag{4'}$$

Доказательство. Имеем:

$$\frac{P_k}{Q_k}-\frac{P_{k-1}}{Q_{k-1}}=\frac{\Delta_k}{Q_{k-1}Q_k}, \tag{5}$$

где

$$\Delta_k=\begin{vmatrix} P_k & P_{k-1}\\ Q_k & Q_{k-1}\end{vmatrix}.$$

Используя соотношения (3) и (3′), в силу известных свойств определителя получаем:

$$\Delta_k=\begin{vmatrix} a_kP_{k-1}+b_kP_{k-2} & P_{k-1}\\ a_kQ_{k-1}+b_kQ_{k-2} & Q_{k-1}\end{vmatrix}=b_k\begin{vmatrix} P_{k-2} & P_{k-1}\\ Q_{k-2} & Q_{k-1}\end{vmatrix}=-b_k\Delta_{k-1}.$$

Отсюда последовательно будем иметь:

$$\Delta_k=(-b_k)(-b_{k-1})\ldots(-b_1)\,\Delta_0=(-1)^k b_1b_2\ldots b_k\Delta_0,$$

где

$$\Delta_0=\begin{vmatrix} P_0 & P_{-1}\\ Q_0 & Q_{-1}\end{vmatrix}=\begin{vmatrix} a_0 & 1\\ 1 & 0\end{vmatrix}=-1.$$

Таким образом,

$$\Delta_k=(-1)^{k-1}b_1b_2\ldots b_k,$$

и, следовательно, на основании формулы (5) выводим:

$$\frac{P_k}{Q_k}-\frac{P_{k-1}}{Q_{k-1}}=(-1)^{k-1}\frac{b_1b_2\ldots b_k}{Q_{k-1}Q_k}.$$

Следствие 1. Если $\frac{P_{k-1}}{Q_{k-1}}$ и $\frac{P_k}{Q_k}$ $(k \geqslant 1)$ — две соседние подходящие дроби цепной дроби (1), то

$$\Delta_k = P_k Q_{k-1} - P_{k-1} Q_k = (-1)^{k-1} b_1 b_2 \ldots b_k.$$

Следствие 2. Для соседних подходящих дробей $\frac{p_{k-1}}{q_{k-1}}$, $\frac{p_k}{q_k}$ $(k \geqslant 1)$ обыкновенной цепной дроби справедливо равенство

$$\frac{p_k}{q_k} - \frac{p_{k-1}}{q_{k-1}} = \frac{(-1)^{k-1}}{q_{k-1} q_k}. \tag{4''}$$

Теорема 3. *Для двух одинаковой четности соседних подходящих дробей* $\frac{P_{k-2}}{Q_{k-2}}$ *и* $\frac{P_k}{Q_k}$ $(k \geqslant 2)$ *цепной дроби* (1) *справедливо соотношение*

$$\frac{P_k}{Q_k} - \frac{P_{k-2}}{Q_{k-2}} = (-1)^k \frac{b_1 b_2 \ldots b_{k-1} a_k}{Q_{k-2} Q_k}. \tag{6}$$

Доказательство. Имеем:

$$\frac{P_k}{Q_k} - \frac{P_{k-2}}{Q_{k-2}} = \frac{D_k}{Q_{k-2} Q_k}, \tag{7}$$

где

$$D_k = \begin{vmatrix} P_k & P_{k-2} \\ Q_k & Q_{k-2} \end{vmatrix}.$$

Отсюда на основании закона составления подходящих дробей и элементарных свойств определителя получаем:

$$D_k = \begin{vmatrix} a_k P_{k-1} + b_k P_{k-2} & P_{k-2} \\ a_k Q_{k-1} + b_k Q_{k-2} & Q_{k-2} \end{vmatrix} = a_k \begin{vmatrix} P_{k-1} & P_{k-2} \\ Q_{k-1} & Q_{k-2} \end{vmatrix} = a_k \Delta_{k-1},$$

где $\Delta_k$ — определитель, рассмотренный в теореме 2. Согласно следствию 1 теоремы 1 имеем:

$$\Delta_{k-1} = (-1)^k b_1 b_2 \ldots b_{k-1},$$

откуда

$$D_k = (-1)^k b_1 b_2 \ldots b_{k-1} a_k.$$

Следовательно, используя соотношение (7), получаем формулу (6).

Следствие. Если $\frac{p_{k-2}}{q_{k-2}}$ и $\frac{p_k}{q_k}$ — две соседние подходящие дроби одинаковой четности для обыкновенной цепной дроби

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cdots}},$$

то имеет место соотношение

$$\frac{p_k}{q_k} - \frac{p_{k-2}}{q_{k-2}} = (-1)^k \frac{a_k}{q_{k-2} q_k}. \tag{6'}$$

Теорема 4. *Если все элементы конечной цепной дроби положительны, то ее подходящие дроби четного порядка образуют монотонно возрастающую последовательность, а подходящие дроби нечетного порядка образуют монотонно убывающую последовательность. При этом каждая подходящая дробь четного порядка меньше любой подходящей дроби нечетного порядка. Само же число $\alpha$, выражаемое цепной дробью, содержится между двумя соседними подходящими дробями.*

Доказательство. Пусть имеем цепную дробь

$$\alpha = \left[a_0;\ \frac{b_1}{a_1},\ \frac{b_2}{a_2},\ \ldots,\ \frac{b_n}{a_n}\right] \tag{8}$$

с положительными элементами $a_k$ и $b_k$ и пусть $\frac{P_k}{Q_k}$ $(k=0, 1, \ldots, n)$ — ее последовательные канонические подходящие дроби. Очевидно, что $P_k > 0$ и $Q_k > 0$.

Рассмотрим два случая.

1. Пусть $k=2m$ — четное число. Тогда из соотношения (6), учитывая, что $a_k > 0$ и $b_i > 0$ $(i=1, \ldots, k)$, получаем:

$$\frac{P_{2m}}{Q_{2m}} - \frac{P_{2m-2}}{Q_{2m-2}} > 0.$$

Следовательно,

$$\frac{P_{2m-2}}{Q_{2m-2}} < \frac{P_{2m}}{Q_{2m}} \qquad (m=1,\ 2,\ \ldots)$$

или

$$\frac{P_0}{Q_0} < \frac{P_2}{Q_2} < \frac{P_4}{Q_4} < \ldots \tag{9}$$

2. Пусть $k=2m+1$ — нечетное число. Следовательно, $k-1$ будет четным числом. Тогда из того же соотношения (6) будем иметь:

$$\frac{P_{2m-1}}{Q_{2m-1}} > \frac{P_{2m+1}}{Q_{2m+1}}$$

или

$$\frac{P_1}{Q_1} > \frac{P_3}{Q_3} > \frac{P_5}{Q_5} > \ldots \tag{10}$$

Таким образом, доказано, что четные подходящие дроби образуют монотонно возрастающую последовательность, а нечетные — монотонно убывающую (рис. 1).

Далее, если в соотношении (4′) положить $k=2m$, то получим:

$$\frac{P_{2m-1}}{Q_{2m-1}} > \frac{P_{2m}}{Q_{2m}}, \tag{11}$$

т. е. всякая подходящая дробь нечетного порядка больше соседней подходящей дроби четного порядка. Отсюда заключаем, что любая подходящая дробь нечетного порядка больше любой подходящей

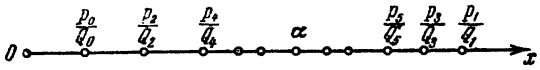


Рис. 1.

дроби четного порядка. Действительно, пусть $\frac{P_{2s-1}}{Q_{2s-1}}$ — какая-нибудь нечетная подходящая дробь. Если $s \leqslant m$, то

$$\frac{P_{2s-1}}{Q_{2s-1}} \geqslant \frac{P_{2m-1}}{Q_{2m-1}} > \frac{P_{2m}}{Q_{2m}};$$

если же $s > m$, то

$$\frac{P_{2s-1}}{Q_{2s-1}} > \frac{P_{2s}}{Q_{2s}} > \frac{P_{2m}}{Q_{2m}}.$$

Следовательно, при любых $s$ и $m$ имеем:

$$\frac{P_{2s-1}}{Q_{2s-1}} > \frac{P_{2m}}{Q_{2m}}. \tag{12}$$

Наконец, из способа образования цепной дроби

$$\alpha = a_0 + \cfrac{b_1}{a_1 + \cfrac{b_2}{a_2 + \ddots}}$$

имеем очевидные соотношения

$$\alpha > \frac{P_0}{Q_0}, \quad \alpha < \frac{P_1}{Q_1}, \quad \alpha > \frac{P_2}{Q_2}$$

и т. д. Следовательно,

$$\frac{P_k}{Q_k} < \alpha < \frac{P_{k+1}}{Q_{k+1}}, \tag{13}$$

если $k$ — четное число, и

$$\frac{P_k}{Q_k} > \alpha > \frac{P_{k+1}}{Q_{k+1}}, \tag{13'}$$

если $k$ — нечетное число. Очевидно, что для последней подходящей дроби вместо строгих неравенств (13) и (13′) мы будем иметь справа равенство.

Следствие 1. Если элементы цепной дроби (8) положительны и $\frac{P_k}{Q_k}$ — ее подходящие дроби, то справедлива оценка

$$\left|\alpha - \frac{P_k}{Q_k}\right| \leqslant \frac{b_1 b_2 \ldots b_{k+1}}{Q_k Q_{k+1}}.$$

Действительно, так как согласно доказанному имеем:

$$\left|\alpha-\frac{P_k}{Q_k}\right|\leqslant\left|\frac{P_{k+1}}{Q_{k+1}}-\frac{P_k}{Q_k}\right|,$$

то на основании формулы (4′) получим оценку (14).

Следствие 2. Если цепная дробь $\alpha$ с положительными элементами — обыкновенная и $\frac{p_k}{q_k}$ — ее последовательные подходящие дроби, то

$$\left|\alpha-\frac{p_k}{q_k}\right|\leqslant\frac{1}{q_kq_{k+1}}.$$

Замечание. Если элементы обыкновенной цепной дроби — натуральные, то можно показать [1], что подходящая дробь $\frac{p_k}{q_k}$ является *наилучшим приближением* числа $\alpha$, т. е. все остальные дроби $\frac{p}{q}$ со знаменателем $q\leqslant q_k$ отклоняются от числа $\alpha$ больше, чем дробь $\frac{p_k}{q_k}$.

Пример 3. Для дроби $\frac{163}{59}$ предпоследней подходящей дробью являлась $\frac{58}{21}$ (см. пример 1). Поэтому

$$\left|\frac{163}{59}-\frac{58}{21}\right|\leqslant\frac{1}{59\cdot21}<0{,}001.$$

## § 4. Бесконечные цепные дроби

Пусть

$$\left[a_0;\ \frac{b_1}{a_1},\ \frac{b_2}{a_2},\ \ldots\right]=a_0+\cfrac{b_1}{a_1+\cfrac{b_2}{a_2+\cdots}} \tag{1}$$

— бесконечная цепная дробь. Рассмотрим ее отрезок, т.е. конечную цепную дробь

$$\left[a_0;\ \frac{b_1}{a_1},\ \frac{b_2}{a_2},\ \ldots,\ \frac{b_n}{a_n}\right]=\frac{P_n}{Q_n}\quad(n=1,\ 2,\ 3,\ \ldots). \tag{2}$$

Определение. Бесконечная цепная дробь (1) называется сходящейся, если существует конечный предел

$$\alpha=\lim_{n\to\infty}\frac{P_n}{Q_n}, \tag{3}$$

причем число $\alpha$ принимается за значение этой дроби. Если же предел (3) не существует, то цепная дробь (1) называется *расходящейся* и ей не приписывается никакого числового значения.

Согласно *критерию Коши* [3] для сходимости последовательности $\frac{P_n}{Q_n}$ $(n=1, 2, 3, \ldots)$ необходимо и достаточно, чтобы для каждого $\varepsilon > 0$ существовало число $N=N(\varepsilon)$ такое, что

$$\left|\frac{P_{n+m}}{Q_{n+m}}-\frac{P_n}{Q_n}\right|<\varepsilon$$

при $n > N$ и любом $m > 0$.

Если $Q_k \neq 0$, то, очевидно, имеем:

$$\frac{P_n}{Q_n}=\frac{P_0}{Q_0}+\sum_{k=1}^{n}\left(\frac{P_k}{Q_k}-\frac{P_{k-1}}{Q_{k-1}}\right). \tag{4}$$

Отсюда

$$\lim_{n\to\infty}\frac{P_n}{Q_n}=\frac{P_0}{Q_0}+\sum_{k=1}^{\infty}\left(\frac{P_k}{Q_k}-\frac{P_{k-1}}{Q_{k-1}}\right)=\frac{P_0}{Q_0}+\sum_{k=1}^{\infty}(-1)^{k-1}\frac{b_1b_2\ldots b_k}{Q_{k-1}Q_k}, \tag{4'}$$

т. е. сходимость цепной дроби (1) эквивалентна сходимости ряда (4′). Если цепная дробь (1) сходится:

$$\alpha=\lim_{n\to\infty}\frac{P_n}{Q_n},$$

то в силу формул (4) и (4′) имеем оценку

$$\left|\alpha-\frac{P_n}{Q_n}\right|\leqslant\sum_{k=n+1}^{\infty}\left|\frac{P_k}{Q_k}-\frac{P_{k-1}}{Q_{k-1}}\right|\leqslant\sum_{k=n+1}^{\infty}\left|\frac{b_1b_2\ldots b_k}{Q_{k-1}Q_k}\right|.$$

Теорема 1. *Если все элементы* $a_k$, $b_k$ $(k=0, 1, \ldots)$ *цепной дроби* (1) *положительны, причем*

$$b_k\leqslant a_k \text{ и } a_k\geqslant d>0 \quad (k=1, 2, \ldots), \tag{5}$$

*то цепная дробь* (1) — *сходящаяся.*

Доказательство. При доказательстве первой части теоремы 4 предыдущего параграфа не было использовано свойство конечности цепной дроби. Поэтому, повторяя это доказательство, устанавливаем, что если элементы цепной дроби (1) положительны, то ее четные подходящие дроби $\frac{P_{2k}}{Q_{2k}}$ $(k=0, 1, 2, \ldots)$ образуют монотонно возрастающую последовательность, ограниченную сверху (например, числом $\frac{P_1}{Q_1}$). Отсюда в силу известной теоремы заключаем, что существует

$$\lim_{k\to\infty}\frac{P_{2k}}{Q_{2k}}=\alpha.$$

Аналогично в условиях нашей теоремы нечетные подходящие дроби $\frac{P_{2k+1}}{Q_{2k+1}}$ $(k=0, 1, 2, \ldots)$ цепной дроби (1) образуют монотонно убывающую последовательность, ограниченную снизу $\left(\text{например, числом } \frac{P_0}{Q_0}\right)$. Следовательно, существует также

$$\lim_{k\to\infty}\frac{P_{2k+1}}{Q_{2k+1}}=\beta,$$

причем $\beta\geqslant\alpha$. Кроме того, для любого $k\geqslant 0$ имеем:

$$\frac{P_{2k}}{Q_{2k}}<\alpha\leqslant\beta<\frac{P_{2k+1}}{Q_{2k+1}};$$

поэтому, используя теорему 2 § 3, получим:

$$0\leqslant\beta-\alpha<\frac{P_{2k+1}}{Q_{2k+1}}-\frac{P_{2k}}{Q_{2k}}=\frac{b_1b_2\ldots b_{2k+1}}{Q_{2k}Q_{2k+1}}=\eta_k. \tag{6}$$

Покажем, что $\eta_k\to 0$ при $k\to\infty$. В самом деле, на основании закона составления подходящих дробей при $k\geqslant 2$ имеем:

$$Q_k=a_kQ_{k-1}+b_kQ_{k-2}$$

и

$$Q_{k-1}=a_{k-1}Q_{k-2}+b_{k-1}Q_{k-3}.$$

Отсюда в силу условия (5) теоремы выводим:

$$Q_k\geqslant b_k(Q_{k-1}+Q_{k-2})$$

и

$$Q_{k-1}\geqslant dQ_{k-2}.$$

Следовательно,

$$Q_k\geqslant b_k(1+d)Q_{k-2}. \tag{7}$$

Из неравенства (7) последовательно получаем:

$$Q_{2k}\geqslant b_{2k}(1+d)Q_{2k-2}\geqslant\ldots$$
$$\ldots\geqslant b_{2k}b_{2k-2}\ldots b_2(1+d)^kQ_0=b_2b_4\ldots b_{2k}(1+d)^k \tag{8}$$

и

$$Q_{2k+1}\geqslant b_{2k+1}(1+d)Q_{2k-1}\geqslant\ldots$$
$$\ldots\geqslant b_{2k+1}\ldots b_3(1+d)^kQ_1\geqslant b_1b_3\ldots b_{2k+1}(1+d)^k, \tag{9}$$

так как $Q_1=a_1\geqslant b_1$. Перемножая неравенства (8) и (9), находим:

$$Q_{2k}Q_{2k+1}\geqslant b_1b_2\ldots b_{2k+1}(1+d)^{2k} \tag{10}$$

и, значит,

$$\eta_k=\frac{b_1b_2\ldots b_{2k+1}}{Q_{2k}Q_{2k+1}}\leqslant\frac{1}{(1+d)^{2k}}.$$

Таким образом, $\eta_k\to 0$ при $k\to\infty$.

Поэтому, переходя к пределу при $k \longrightarrow \infty$ в неравенстве (6), будем иметь $0 \leqslant \beta - \alpha \leqslant 0$, т. е.

$$\alpha = \beta = \lim_{n \to \infty} \frac{P_n}{Q_n},$$

и, следовательно, цепная дробь (1) сходится.

Замечание. Для сходящейся дроби (1) с положительными элементами ее значение $\alpha$ заключено между двумя последовательными подходящими дробями $\frac{P_{n-1}}{Q_{n-1}}$ и $\frac{P_n}{Q_n}$. Следовательно,

$$\left|\alpha - \frac{P_n}{Q_n}\right| \leqslant \left|\frac{P_n}{Q_n} - \frac{P_{n-1}}{Q_{n-1}}\right| = \frac{b_1 b_2 \ldots b_n}{Q_{n-1} Q_n}.$$

Следствие. Обыкновенная цепная дробь с натуральными элементами всегда сходится.

Можно доказать [1] также следующую теорему.

Теорема 2. *Каждое положительное число $\alpha$ можно разложить в обыкновенную сходящуюся цепную дробь с натуральными элементами, причем это разложение единственно.* Полученная цепная дробь конечна, если $\alpha$ — рациональное число, и бесконечна, если $\alpha$ — иррациональное число.

Пример. Разложить в цепную дробь число $\sqrt{41}$ и найти его приближенное значение.

Решение. Так как наибольшее целое число, заключающееся в $\sqrt{41}$, есть 6, то имеем:

$$\sqrt{41} = 6 + \frac{1}{a_1}. \tag{11}$$

Отсюда

$$a_1 = \frac{1}{\sqrt{41} - 6} = \frac{6 + \sqrt{41}}{5}.$$

Наибольшее целое число, заключающееся в $a_1$, есть 2, поэтому

$$a_1 = 2 + \frac{1}{a_2}. \tag{12}$$

Отсюда

$$a_2 = \frac{1}{a_1 - 2} = \frac{5}{\sqrt{41} - 4} = \frac{4 + \sqrt{41}}{5} = 2 + \frac{1}{a_3}. \tag{13}$$

Аналогично

$$a_3 = \frac{1}{a_2 - 2} = \frac{5}{\sqrt{41} - 6} = 6 + \sqrt{41} = 12 + \frac{1}{a_4}; \tag{14}$$

$$a_4 = \frac{1}{a_3 - 12} = \frac{1}{\sqrt{41} - 6} = \frac{6 + \sqrt{41}}{5} = 2 + \frac{1}{a_5}. \tag{15}$$

Мы замечаем, что $a_4 = a_1$, поэтому элементы цепной дроби будут повторяться, т. е. цепная дробь получится периодической. Производя последовательную подстановку в равенство (11) выражений (12), (13), (14), (15) и т. д., получим:

$$\sqrt{41} = 6 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{12 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{12 + \ddots}}}}}}$$

Таким образом, иррациональное число $\sqrt{41}$ выразилось бесконечной периодической цепной дробью:

$$\sqrt{41} = \left(6;\ \frac{1}{2},\ \frac{1}{2},\ \frac{1}{12},\ \frac{1}{2},\ \frac{1}{2},\ \frac{1}{12},\ \frac{1}{2},\ \frac{1}{2},\ \frac{1}{12},\ \ldots\right).$$

Подходящие дроби $\frac{p_k}{q_k}$ $(k = 0, 1, 2, \ldots)$ находим, пользуясь следующей схемой:

| $a_k$ | — | 6 | 2 | 2 | 12 | 2 | 2 | 12 |
|---|---|---|---|---|---|---|---|---|
| $p_k$ | $p_{-1} = 1$ | $p_0 = 6$ | 13 | 32 | 397 | 826 | 2049 | ... |
| $q_k$ | $q_{-1} = 0$ | $q_0 = 1$ | 2 | 5 | 62 | 129 | 320 | ... |

Ограничиваясь, например, пятой подходящей дробью, мы будем иметь приближенное значение $\sqrt{41}$ по избытку: $\sqrt{41} = \frac{2049}{320} = 6{,}403125$ с абсолютной погрешностью меньшей, чем

$$\Delta = \frac{1}{320\,(2 \cdot 320 + 129)} = \frac{1}{320 \cdot 769} < 5 \cdot 10^{-6}.$$

Теорема 3 (Принсгейма). *Если для бесконечной цепной дроби*

$$\left[0;\ \frac{b_1}{a_1},\ \frac{b_2}{a_2},\ \ldots,\ \frac{b_n}{a_n},\ \ldots\right] \tag{16}$$

*выполнены неравенства*

$$|b_n| + 1 \leqslant |a_n| \qquad (n = 1, 2, \ldots), \tag{17}$$

*то эта дробь — сходящаяся, причем абсолютное значение ее не превышает единицы* [4].

Доказательство. Пусть $\frac{P_k}{Q_k}$ $(k=1, 2, \ldots)$ — подходящие дроби цепной дроби (16).

Так как

$$Q_k = a_k Q_{k-1} + b_k Q_{k-2} \quad (k=1, 2, \ldots),$$

то

$$|Q_k| \geqslant |a_k||Q_{k-1}| - |b_k||Q_{k-2}|.$$

Отсюда, используя неравенство (17), получаем:

$$|Q_k| \geqslant (|b_k|+1)|Q_{k-1}| - |b_k||Q_{k-2}|,$$

или

$$|Q_k| - |Q_{k-1}| \geqslant |b_k|(|Q_{k-1}| - |Q_{k-2}|). \tag{18}$$

Последовательно применяя неравенство (18) и учитывая, что $Q_0 = 1$ и $Q_{-1} = 0$, будем иметь:

$$|Q_k| - |Q_{k-1}| \geqslant |b_k||b_{k-1}| \ldots |b_1|. \tag{19}$$

Из неравенства (19) вытекает, что $|Q_k|$ монотонно возрастает при возрастании $k$, причем $|Q_k| \geqslant |Q_0| = 1$.

Сходимость цепной дроби (16) эквивалентна сходимости ряда

$$\frac{P_0}{Q_0} + \sum_{k=1}^{\infty} \left(\frac{P_k}{Q_k} - \frac{P_{k-1}}{Q_{k-1}}\right) = \sum_{k=1}^{\infty} \frac{(-1)^{k-1} b_1 b_2 \ldots b_k}{Q_{k-1} Q_k}. \tag{20}$$

Рассмотрим ряд модулей

$$\sum_{k=1}^{\infty} \frac{|b_1||b_2| \ldots |b_k|}{|Q_{k-1}||Q_k|}. \tag{21}$$

На основании неравенства (19) имеем:

$$\sum_{k=1}^{n} \frac{|b_1||b_2| \ldots |b_k|}{|Q_{k-1}||Q_k|} \leqslant \sum_{k=1}^{n} \frac{|Q_k| - |Q_{k-1}|}{|Q_{k-1}||Q_k|} =$$

$$= \sum_{k=1}^{n} \left(\frac{1}{|Q_{k-1}|} - \frac{1}{|Q_k|}\right) = \frac{1}{|Q_0|} - \frac{1}{|Q_n|} < \frac{1}{|Q_0|} = 1 \qquad (n=1, 2, \ldots).$$

Таким образом, частные суммы ряда (21) ограничены и, следовательно, этот ряд сходится, причем

$$\sum_{k=1}^{\infty} \frac{|b_1||b_2| \ldots |b_k|}{|Q_{k-1}||Q_k|} \leqslant 1. \tag{22}$$

Но тогда, в силу признака сравнения, также сходится, и притом абсолютно, ряд (20), т. е. существует

$$\lim_{n\to\infty} \frac{P_n}{Q_n} = \sum_{k=1}^{\infty} \left(\frac{P_k}{Q_k} - \frac{P_{k-1}}{Q_{k-1}}\right) = \alpha.$$

Кроме того, учитывая неравенство (22), имеем:

$$|\alpha| \leqslant 1.$$

Замечание 1. Для сходимости цепной дроби (16) достаточно, чтобы неравенство (17) имело место при $n \geqslant m$, причем $Q_k \neq 0$ при $k \leqslant m$.

Замечание 2. В условиях теоремы 3 для значения цепной дроби $\alpha$ имеем следующую оценку:

$$\left|\alpha - \frac{P_n}{Q_n}\right| \leqslant \sum_{k=n+1}^{\infty} \left|\frac{P_k}{Q_k} - \frac{P_{k-1}}{Q_{k-1}}\right| =$$

$$= \sum_{k=n+1}^{\infty} \frac{|b_1||b_2| \ldots |b_k|}{|Q_{k-1}||Q_k|} \leqslant \sum_{k=n+1}^{\infty} \frac{|Q_k| - |Q_{k-1}|}{|Q_{k-1}||Q_k|} =$$

$$= \sum_{k=n+1}^{\infty} \left(\frac{1}{|Q_{k-1}|} - \frac{1}{|Q_k|}\right) = \frac{1}{|Q_n|} - \lim_{k \to \infty} \frac{1}{|Q_k|}.$$

В частности, если $|Q_k| \to +\infty$ при $k \to \infty$, то

$$\left|\alpha - \frac{P_n}{Q_n}\right| \leqslant \frac{1}{|Q_n|}.$$

## § 5. Разложение функций в цепные дроби

Цепные дроби являются удобным аппаратом для представления и вычисления функций. Подробности по этому вопросу изложены в специальной литературе (см., например, [2]), а мы ограничимся лишь рассмотрением отдельных примеров.

Заметим, что если функция $f(x)$ с помощью какого-нибудь приема разлагается в бесконечную цепную дробь, то в общем случае нужно доказать сходимость этой дроби и убедиться, что предельное значение цепной дроби равно функции $f(x)$.

### А. Разложение рациональной функции в цепную дробь

Если

$$f(x) = \frac{c_{10} + c_{11}x + c_{12}x^2 + \ldots}{c_{00} + c_{01}x + c_{02}x^2 + \ldots},$$

то в общем случае, производя элементарные преобразования, будем иметь:

$$f(x) = \cfrac{1}{\cfrac{c_{00}}{c_{10}} + \cfrac{c_{00} + c_{01}x + c_{02}x^2 + \ldots}{c_{10} + c_{11}x + c_{12}x^2 + \ldots} - \cfrac{c_{00}}{c_{10}}} = \frac{c_{10}}{c_{00} + xf_1(x)},$$

где

$$f_1(x)=\frac{c_{20}+c_{21}x+c_{22}x^2+\ldots}{c_{10}+c_{11}x+c_{12}x^2+\ldots}$$

и

$$c_{2k}=c_{10}c_{0,\,k+1}-c_{00}c_{1,\,k+1}\quad (k=0,\ 1,\ \ldots).$$

Аналогично

$$f_1(x)=\frac{c_{20}}{c_{10}+xf_2(x)},$$

где

$$f_2(x)=\frac{c_{30}+c_{31}x+c_{32}x^2+\ldots}{c_{20}+c_{21}x+c_{22}x^2+\ldots}$$

и

$$c_{3k}=c_{20}c_{1,\,k+1}-c_{10}c_{2,\,k+1}\quad (k=0,\ 1,\ \ldots)$$

и т. д.

Таким образом

$$f(x)=\cfrac{c_{10}}{c_{00}+\cfrac{c_{20}x}{c_{10}+\cfrac{c_{30}x}{c_{20}+\cdot}}}=\left[0;\ \frac{c_{10}}{c_{00}},\ \frac{c_{20}x}{c_{10}},\ \frac{c_{30}x}{c_{20}},\ \ldots,\ \frac{c_{n0}x}{c_{n-1,\,0}}\right],\qquad (1)$$

причем легко убедиться, что цепная дробь (1) получится конечной.

Коэффициенты разложений $c_{jk}$ удобно последовательно вычислять по формуле

$$c_{jk}=-\begin{vmatrix} c_{j-2,\,0} & c_{j-2,\,k+1}\\ c_{j-1,\,0} & c_{j-1,\,k+1}\end{vmatrix},$$

где $j\geqslant 2$.

Заметим, что в некоторых случаях коэффициенты $c_{jk}$ могут оказаться равными нулю. Тогда в разложение (1) нужно внести соответствующие изменения [2].

Пример 1. Разложить в цепную дробь функцию

$$f(x)=\frac{1-x}{1-5x+6x^2}.$$

Решение. Выписываем коэффициенты $c_{jk}$ в следующую схему:

| $j$ \ $k$ | 0 | 1 | 2 |
|---|---|---|---|
| 0 | 1 | −5 | 6 |
| 1 | 1 | −1 | 0 |
| 2 | −4 | 6 | 0 |
| 3 | −2 | 0 | 0 |
| 4 | −12 | 0 | 0 |

Следовательно,

$$\frac{1-x}{1-5x+6x^2}=\left[0;\ \frac{1}{1},\ \frac{-4x}{1},\ \frac{-2x}{-4},\ \frac{-12x}{-2}\right]=\cfrac{1}{1-\cfrac{4x}{1-\cfrac{2x}{-4+6x}}}.$$

### Б. Разложение $e^x$ в цепную дробь

Для $e^x$ Эйлер получил разложение [2]

$$e^x=\left[0;\ \frac{1}{1},\ \frac{-2x}{2+x},\ \frac{x^2}{6},\ \frac{x^2}{10},\ \ldots,\ \frac{x^2}{4n+2},\ \ldots\right], \tag{2}$$

сходящееся для любого значения $x$, действительного или комплексного [2].

Отсюда получаем подходящие дроби:

$$\frac{P_1}{Q_1}=\frac{1}{1};$$
$$\frac{P_2}{Q_2}=\frac{2+x}{2-x};$$
$$\frac{P_3}{Q_3}=\frac{12+6x+x^2}{12-6x+x^2};$$
$$\frac{P_4}{Q_4}=\frac{120+60x+12x^2+x^3}{120-60x+12x^2-x^3}$$

и т. д.

Полагая, в частности, $x=1$ и ограничиваясь четвертой подходящей дробью, имеем

$$e\approx\frac{193}{71}=2{,}7183\ \ldots$$

Для достижения той же точности нужно в разложении Маклорена

$$e=2+\frac{1}{2!}+\frac{1}{3!}+\ldots$$

взять не менее восьми членов.

### В. Разложение $\operatorname{tg} x$ в цепную дробь

Для $\operatorname{tg} x$ Ламбертом получено разложение

$$\operatorname{tg} x=\left[0;\ \frac{x}{1},\ \frac{-x^2}{3},\ \frac{-x^2}{5},\ \ldots,\ \frac{-x^2}{2n+1},\ \ldots\right], \tag{3}$$

сходящееся во всех точках непрерывности функции.

Пример 2. Найти приближенно $\operatorname{tg} 1$.

Решение. Полагая в формуле (3) $x=1$, будем иметь:

$$\operatorname{tg} 1=\left[0;\ \frac{1}{1},\ \frac{-1}{3},\ \frac{-1}{5},\ \ldots\right].$$

На основании формулы (3) из § 3 составим следующую схему для членов подходящих дробей:

| $k$ | $-1$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|
| $b_k$ | | 1 | 1 | $-1$ | $-1$ | 1 |
| $a_k$ | | 0 | 1 | 3 | 5 | 7 |
| $P_k$ | 1 | 0 | 1 | 3 | 14 | 95 |
| $Q_k$ | 0 | 1 | 1 | 2 | 9 | 61 |

Ограничиваясь четвертой подходящей дробью, будем иметь:

$$\operatorname{tg} 1 \approx \frac{95}{61} = 1{,}557377$$

(по таблицам $\operatorname{tg} 1 = 1{,}557396$).

## Литература ко второй главе

1. А. Я. Хинчин, Цепные дроби, Гостехиздат, 1949, гл. I.
2. А. Н. Хованский, Приложение цепных дробей и их обобщений к вопросам приближенного анализа, Гостехиздат, 1956, гл. I и II.
3. Г. М. Фихтенгольц, Основы математического анализа, Гостехиздат, 1955, т. 1, гл. III.
4. O. Perron, Die Lehre von den Kettenbrüchen, Teubner, 1913, гл. VII.

## ГЛАВА III

# ВЫЧИСЛЕНИЕ ЗНАЧЕНИЙ ФУНКЦИЙ

При вычислении с помощью счетных машин значений функций, заданных формулами, далеко не безразлично, в каком виде записана соответствующая формула. Математически эквивалентные выражения часто оказываются неравноценными с точки зрения приближенных вычислений. Поэтому возникает практически важная задача о нахождении для элементарных функций наиболее удобных аналитических выражений. Вычисление значений функций обычно сводится к последовательности элементарных арифметических действий. Учитывая ограниченность объема памяти машины, желательно эти операции разбивать на повторяющиеся циклы. Ниже мы рассмотрим некоторые типовые приемы вычислений.

## § 1. Вычисление значений полинома. Схема Гориера

Пусть дан полином $n$-й степени

$$P(x) = a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n \tag{1}$$

с действительными коэффициентами $a_k$ $(k = 0, 1, \ldots, n)$. Положим, что требуется найти значение этого полинома при $x = \xi$:

$$P(\xi) = a_0\xi^n + a_1\xi^{n-1} + \ldots + a_{n-1}\xi + a_n. \tag{2}$$

Вычисление числа $P(\xi)$ удобнее всего производить следующим образом. Представим формулу (2) в виде

$$P(\xi) = (\ldots(((a_0\xi + a_1)\xi + a_2)\xi + a_3)\xi + \ldots + a_{n-1})\xi + a_n).$$

Отсюда, последовательно вычисляя числа

$$\left.\begin{aligned} b_0 &= a_0, \\ b_1 &= a_1 + b_0\xi, \\ b_2 &= a_2 + b_1\xi, \\ b_3 &= a_3 + b_2\xi, \\ &\ldots\ldots \\ b_n &= a_n + b_{n-1}\xi, \end{aligned}\right\} \tag{3}$$

находим $b_n = P(\xi)$.

Покажем, что числа $b_0 = a_0$, $b_1, \ldots, b_{n-1}$ являются коэффициентами полинома $Q(x)$, полученного в качестве частного при делении данного полинома $P(x)$ на двучлен $x - \xi$. В самом деле, пусть

$$Q(x) = \beta_0 x^{n-1} + \beta_1 x^{n-2} + \ldots + \beta_{n-1} \tag{4}$$

и

$$P(x) = Q(x)(x - \xi) + \beta_n, \tag{5}$$

причем на основании теоремы Безу остаток от деления $\beta_n = P(\xi)$. Из формул (4) и (5) получим:

$$P(x) = (\beta_0 x^{n-1} + \beta_1 x^{n-2} + \ldots + \beta_{n-1})(x - \xi) + \beta_n,$$

или, раскрыв скобки и сделав приведение подобных членов, будем иметь:

$$P(x) = \beta_0 x^n + (\beta_1 - \beta_0 \xi) x^{n-1} + (\beta_2 - \beta_1 \xi) x^{n-2} + \ldots$$
$$\ldots + (\beta_{n-1} - \beta_{n-2}\xi) x + (\beta_n - \beta_{n-1}\xi).$$

Сравнивая коэффициенты при одинаковых степенях переменной $x$ в левой и правой частях последнего равенства, получим:

$$\begin{aligned} \beta_0 &= a_0, \\ \beta_1 - \beta_0\xi &= a_1, \\ \beta_2 - \beta_1\xi &= a_2, \\ &\cdots\cdots \\ \beta_{n-1} - \beta_{n-2}\xi &= a_{n-1}, \\ \beta_n - \beta_{n-1}\xi &= a_n. \end{aligned}$$

Отсюда

$$\begin{aligned} \beta_0 &= a_0 = b_0, \\ \beta_1 &= a_1 + \beta_0\xi = b_1, \\ \beta_2 &= a_2 + \beta_1\xi = b_2, \\ &\cdots\cdots\cdots \\ \beta_{n-1} &= a_{n-1} + \beta_{n-2}\xi = b_{n-1}, \\ \beta_n &= a_n + \beta_{n-1}\xi = b_n, \end{aligned}$$

что и требовалось доказать.

Таким образом, формулы (3) позволяют, не производя деления, определять коэффициенты частного $Q(x)$, а также остаток $P(\xi)$. Практически вычисления осуществляются по следующей схеме, называемой *схемой Горнера*:

$$\begin{array}{rccccc|l} & a_0 & a_1 & a_2 & \ldots & a_n & \xi \\ + & & b_0\xi & b_1\xi & \ldots & b_{n-1}\xi & \\ \hline & b_0 & b_1 & b_2 & \ldots & \underline{b_n = P(\xi)} & \end{array}$$

Пример 1. Вычислить значение полинома

$$P(x)=3x^3+2x^2-5x+7 \text{ при } x=3.$$

Решение. Составляем схему Горнера:

$$\begin{array}{r|l} & 3 \quad 2 \quad -5 \quad 7 \qquad\qquad \underline{|3} \\ + & \phantom{3\ } 9 \quad 33 \quad 84 \\ \hline & 3 \quad 11 \quad 28 \quad \underline{91=P(3)} \end{array}$$

Замечание. Пользуясь схемой Горнера, можно получить границы действительных корней данного полинома $P(x)$.

Положим, что при $x=\beta$ $(\beta>0)$ все коэффициенты $b_i$ в схеме Горнера неотрицательны, причем первый коэффициент положителен, т. е.

$$b_0=a_0>0 \text{ и } b_i\geqslant 0 \qquad (i=1,\ 2,\ \ldots,\ n). \tag{6}$$

Тогда можно утверждать, что все действительные корни $x_k$ $(k=1,\ 2,\ \ldots,\ m;\ m\leqslant n)$ полинома $P(x)$ расположены не правее $\beta$, т. е. $x_k\leqslant\beta$ $(k=1,\ 2,\ \ldots,\ m)$ (рис. 2).

В самом деле, так как

$$P(x)=(b_0x^{n-1}+\ldots+b_{n-1})(x-\beta)+b_n,$$

то при любом $x>\beta$ в силу условия (6) будем иметь $P(x)>0$, т. е. любое число, большее $\beta$, заведомо не является корнем полинома $P(x)$. Таким образом, имеем верхнюю оценку для действительных корней $x_k$ полинома.

Для получения нижней оценки корней $x_k$ составляем полином

$$(-1)^n P(-x)=a_0x^n-a_1x^{n-1}+\ldots+(-1)^n a_n.$$

Для этого нового полинома находим такое число $x=\alpha$ $(\alpha>0)$, чтобы все коэффициенты в соответствующей схеме Горнера были неотрицательны, за исключением первого, который, очевидно, будет положительным. Тогда согласно предыдущим рассуждениям для действительных корней полинома $(-1)^n P(-x)$, очевидно, равных $-x_k$ $(k=1,\ 2,\ \ldots,\ m)$, имеем неравенство $-x_k\leqslant\alpha$.

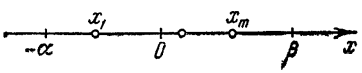

Рис. 2.

Следовательно, $x_k\geqslant-\alpha$ $(k=1,\ 2,\ \ldots,\ m)$. Таким образом, мы получили нижнюю границу $-\alpha$ действительных корней полинома $P(x)$. Отсюда следует, что все действительные корни полинома $P(x)$ расположены на отрезке $[-\alpha,\ \beta]$.

Пример 2. Найти границы действительных корней полинома

$$P(x)=x^4-2x^3+3x^2+4x-1.$$

Решение. Подсчитаем значение полинома $P(x)$, например, при $x=2$. Пользуясь схемой Горнера, получим:

$$\begin{array}{r|rrrrr|l} + & 1 & -2 & 3 & 4 & -1 & \underline{|2} \\ & & 2 & 0 & 6 & 20 & \\ \hline & 1 & 0 & 3 & 10 & \underline{19} & \end{array}$$

Так как все коэффициенты $b_i \geqslant 0$, то действительные корни $x_k$ полинома $P(x)$ (если они существуют) удовлетворяют неравенству $x_k < 2$. Верхняя граница действительных корней найдена. Перейдем к оценке нижней границы. Составим новый полином:

$$Q(x) = (-1)^4 P(-x) = x^4 + 2x^3 + 3x^2 - 4x - 1.$$

Подсчитывая значение полинома $Q(x)$, например, при $x=1$, имеем:

$$\begin{array}{r|rrrrr|l} + & 1 & 2 & 3 & -4 & -1 & \underline{|1} \\ & & 1 & 3 & 6 & 2 & \\ \hline & 1 & 3 & 6 & 2 & 1 & \end{array}$$

Все коэффициенты $b_i > 0$, значит, $-x_k < 1$, т. е. $x_k > -1$. Итак, все действительные корни данного полинома находятся внутри отрезка $[-1, 2]$.

## § 2. Обобщенная схема Горнера

Пусть в данном полиноме

$$P(x) = a_0 x^n + a_1 x^{n-1} + \ldots + a_n \tag{1}$$

по каким-то соображениям требуется произвести замену переменной $x$ по формуле

$$x = y + \xi, \tag{2}$$

где $\xi$ — фиксированное число и $y$ — новая переменная.

Подставив выражение (2) в полином (1) и произведя указанные действия, после приведения подобных членов получим новый полином относительно переменной $y$:

$$P(y+\xi) = A_0 y^n + A_1 y^{n-1} + \ldots + A_n. \tag{3}$$

Так как полином (3) можно рассматривать как разложение Тейлора по степеням $y$ функции $P(y+\xi)$, то коэффициенты $A_i$ $(i=0, 1, \ldots, n)$ могут быть вычислены по формуле

$$A_i = \frac{P^{(n-i)}(\xi)}{(n-i)!} \qquad (i=0, 1, \ldots, n).$$

Укажем более удобный на практике способ отыскания коэффициентов $A_i$ $(i=0, 1, 2, \ldots, n)$ с помощью схемы Горнера. Положим сначала $y=0$ в выражении (3). Тогда будем иметь $A_n = P(\xi)$.

Разделив полином (1) на двучлен $x-\xi$, получим:

$$P(x)=(x-\xi)P_1(x)+P(\xi), \tag{4}$$

где

$$P_1(x)=b_0x^{n-1}+b_1x^{n-2}+\ldots+b_{n-1}.$$

Далее, если в выражение (3) вместо $y$ подставить его значение $y=x-\xi$, то будем иметь:

$$P(x)=(x-\xi)\left[A_0(x-\xi)^{n-1}+A_1(x-\xi)^{n-2}+\ldots+A_{n-1}\right]+P(\xi). \tag{5}$$

Сопоставляя формулы (4) и (5), заключаем, что

$$P_1(x)=A_0(x-\xi)^{n-1}+A_1(x-\xi)^{n-2}+\ldots+A_{n-1}. \tag{6}$$

Отсюда

$$A_{n-1}=P_1(\xi). \tag{7}$$

Аналогично, разделив полином $P_1(x)$ на двучлен $x-\xi$, можем положить:

$$P_1(x)=(x-\xi)P_2(x)+P_1(\xi), \tag{8}$$

где $P_2(x)=c_0x^{n-2}+c_1x^{n-3}+\ldots+c_{n-2}$.

Кроме того, из формул (6) и (7) имеем:

$$P_1(x)=(x-\xi)\left[A_0(x-\xi)^{n-2}+A_1(x-\xi)^{n-3}+\ldots+A_{n-2}\right]+P_1(\xi). \tag{9}$$

Сопоставляя формулы (8) и (9), заключаем, что

$$P_2(x)=A_0(x-\xi)^{n-2}+A_1(x-\xi)^{n-3}+\ldots+A_{n-2}.$$

Отсюда $A_{n-2}=P_2(\xi)$.

Продолжая этот процесс далее, мы последовательно выразим все коэффициенты $A_i\,(i=0,\ 1,\ \ldots,\ n)$ через значения соответствующих полиномов $P_0(x)=P(x)$ и $P_1(x),\ \ldots,\ P_n(x)=a_0$ при $x=\xi$:

$$\begin{aligned}
A_n &= P\ (\xi);\\
A_{n-1} &= P_1(\xi);\\
A_{n-2} &= P_2(\xi);\\
&\cdots\cdots\\
A_0 &= P_n(\xi),
\end{aligned}$$

где полиномы $P_{k+1}(x)$, исходя из полиномов $P_k(x)$, строятся по формуле

$$P_k(x)=(x-\xi)P_{k+1}(x)+P_k(\xi) \qquad (k=0,\ 1,\ \ldots,\ n).$$

Для вычисления значений $P(\xi)$, $P_1(\xi)$, $P_2(\xi)$, ... пользуемся *обобщенной схемой Горнера*:

$$
\begin{array}{cccccc|l}
a_0 & a_1 & a_2 & a_3 & \dots\ a_{n-1} & a_n & \xi \\
 & b_0\xi & b_1\xi & b_2\xi & \dots\ b_{n-2}\xi & b_{n-1}\xi & \\
\hline
b_0 & b_1 & b_2 & b_3 & \dots\ b_{n-1} & \underline{b_n = P(\xi)} &
\end{array}
$$

$$
\begin{array}{cccccc}
 & c_0\xi & c_1\xi & \dots & & c_{n-2}\xi \\
\hline
c_0 & c_1 & c_2 & \dots & & \underline{c_{n-1} = P_1(\xi)} \\
\dots & \dots & \dots & \dots & \dots & \dots
\end{array}
$$

Пример. Чтобы уничтожить в полиноме

$$P(x) = x^4 - 8x^3 + 5x^2 + 2x - 7$$

член, содержащий третью степень неизвестной, полагают $x = y + 2$. Найти преобразованный полином.

Решение. Применяем обобщенную схему Горнера:

$$
\begin{array}{ccccc|l}
1 & -8 & 5 & 2 & -7 & \xi = 2 \\
 & 2 & -12 & -14 & -24 & \\
\hline
1 & -6 & -7 & -12 & \underline{-31} & \\
 & 2 & -8 & -30 & & \\
\hline
1 & -4 & -15 & \underline{-42} & & \\
 & 2 & -4 & & & \\
\hline
1 & -2 & \underline{-19} & & & \\
 & 2 & & & & \\
\hline
1 & \underline{0} & & & & \\
\hline
\underline{1} & & & & &
\end{array}
$$

Следовательно,

$$P(y+2) = y^4 - 19y^2 - 42y - 31.$$

## § 3. Вычисление значений рациональных дробей

Всякую *рациональную дробь* $R(x)$ можно представить в виде отношения двух полиномов, т. е.

$$R(x) = \frac{P(x)}{Q(x)}, \tag{1}$$

где

$$P(x) = a_0x^n + a_1x^{n-1} + \dots + a_n,$$

$$Q(x) = b_0x^m + b_1x^{m-1} + \dots + b_m.$$

Пусть нам требуется определить значение $R(x)$ в точке $x=\xi$:

$$R(\xi)=\frac{P(\xi)}{Q(\xi)}. \tag{2}$$

Числитель $P(\xi)$ и знаменатель $Q(\xi)$ дроби (2) можно найти, пользуясь схемой Горнера. Отсюда получаем простой способ вычисления числа $R(\xi)$.

Другой способ вычисления состоит в преобразовании функции $R(x)$ в цепную дробь. Для этого можно воспользоваться способом, указанным в главе II, § 3.

## § 4. Приближенное нахождение сумм числовых рядов

Определение. Числовой ряд

$$a_1+a_2+\ldots+a_n+\ldots \tag{1}$$

называется *сходящимся*, если существует предел последовательности его частных сумм

$$S=\lim_{n\to\infty} S_n, \tag{2}$$

где

$$S_n=a_1+a_2+\ldots+a_n.$$

Число $S$ называется *суммой ряда*.

Таким образом, сходимость ряда (1) эквивалентна сходимости последовательности его частных сумм. Согласно *критерию Коши* [1] эта последовательность сходится тогда и только тогда, когда для любого $\varepsilon>0$ существует $N=N(\varepsilon)$ такое, что

$$|S_{n+p}-S_n|<\varepsilon$$

при $n>N$ и произвольном $p>0$.

Из формулы (2) получаем:

$$S=S_n+R_n, \tag{3}$$

где $R_n$ — *остаток ряда*, причем $R_n\to 0$ при $n\to\infty$.

Для нахождения суммы $S$ сходящегося ряда (1) с заданной точностью $\varepsilon$ нужно выбрать число слагаемых $n$ столь большим, чтобы имело место неравенство

$$|R_n|<\varepsilon. \tag{4}$$

Тогда частная сумма $S_n$ приближенно может быть принята за точную сумму $S$ ряда (1).

Заметим, что на практике слагаемые $a_1$, $a_2$, ... также определяются приближенно. Кроме того, сумма $S_n$ обычно округляется до заданного числа десятичных знаков. Чтобы учесть все эти погрешности и обеспечить нужную точность, можно применить следующую методику: выберем, в общем случае, три положительных числа $\varepsilon_1$, $\varepsilon_2$ и $\varepsilon_3$ такие, что

$$\varepsilon_1+\varepsilon_2+\varepsilon_3=\varepsilon.$$

Возьмем число $n$ членов ряда столь большим, чтобы *остаточная погрешность* $|R_n|$ удовлетворяла неравенству

$$|R_n| \leqslant \varepsilon_1. \tag{5}$$

Каждое из слагаемых $a_k$ $(k=1, 2, \ldots, n)$ вычислим с предельной абсолютной погрешностью, не превышающей $\frac{\varepsilon_2}{n}$, и пусть $\overline{a}_k$ $(k=1, 2, \ldots, n)$ — соответствующие приближенные значения членов ряда (1), т. е.

$$|\overline{a}_k - a_k| \leqslant \frac{\varepsilon_2}{n}.$$

Тогда для суммы

$$\overline{S}_n = \sum_{k=1}^{n} \overline{a}_k$$

*погрешность действий* (суммирования) удовлетворяет неравенству

$$|S_n - \overline{S}_n| \leqslant \varepsilon_2. \tag{6}$$

Наконец, полученный приближенный результат $\overline{S}_n$ округлим до более простого числа $\overline{\overline{S}}_n$ с таким расчетом, чтобы *погрешность округления* была

$$|\overline{S}_n - \overline{\overline{S}}_n| \leqslant \varepsilon_3. \tag{7}$$

В таком случае число $\overline{\overline{S}}_n$ является приближенным значением суммы $S$ ряда (1) с заданной точностью $\varepsilon$. Действительно, из неравенств (5)—(7) имеем:

$$|S - \overline{\overline{S}}_n| \leqslant |S - S_n| + |S_n - \overline{S}_n| + |\overline{S}_n - \overline{\overline{S}}_n| \leqslant \varepsilon_1 + \varepsilon_2 + \varepsilon_3 = \varepsilon.$$

Разбиение числа $\varepsilon$ на положительные слагаемые $\varepsilon_1$, $\varepsilon_2$ и $\varepsilon_3$ следует производить, сообразуясь с потребным объемом работы для получения искомого результата. Если $\varepsilon = 10^{-m}$ и результат должен содержать $m$ верных десятичных знаков после запятой, то чаще всего принимают:

$$\varepsilon_1 = \frac{\varepsilon}{4}, \quad \varepsilon_2 = \frac{\varepsilon}{4}, \quad \varepsilon_3 = \frac{\varepsilon}{2}.$$

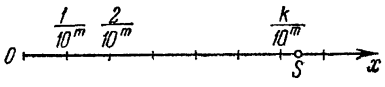


Рис. 3.

Если заключительное округление отсутствует, то обычно полагают:

$$\varepsilon_1 = \frac{\varepsilon}{2}, \quad \varepsilon_2 = \frac{\varepsilon}{2}, \quad \varepsilon_3 = 0.$$

Задача усложняется, если нужно найти сумму ряда с точностью до $m$ верных десятичных знаков после запятой в узком смысле слова. Геометрически это значит, что требуется найти элемент множества $\frac{k}{10^m}$ ($k$ — целое), ближайший к числу $S$ (рис. 3).

Пусть для определенности сумма $S$ положительна и

$$\tilde{S}=p_0+\frac{p_1}{10}+\ldots+\frac{p_m}{10^m}+\ldots+\frac{p_n}{10^n}$$

($p_k$ — целые неотрицательные числа, $n\geqslant m$) — рациональное приближение такое, что

$$|S-\tilde{S}|\leqslant\frac{1}{10^{m+1}}.$$

Допустим, что

$$p_{m+1}\neq 4 \quad \text{и} \quad p_{m+1}\neq 5.$$

Тогда, округляя число $\tilde{S}$, получим искомый результат:

$$\sigma=p_0+\frac{p_1}{10}+\ldots+\frac{p_m}{10^m}, \quad \text{если } p_{m+1}\leqslant 3, \tag{8}$$

или

$$\sigma=p_0+\frac{p_1}{10}+\ldots+\frac{p_m+1}{10^m}, \quad \text{если } p_{m+1}\geqslant 6. \tag{8'}$$

Действительно, в первом случае, при округлении по недостатку мы имеем:

$$0\leqslant\tilde{S}-\sigma=\frac{p_{m+1}}{10^{m+1}}+\frac{p_{m+2}}{10^{m+2}}+\ldots+\frac{p_n}{10^n}\leqslant$$
$$\leqslant\frac{3}{10^{m+1}}+\frac{9}{10^{m+2}}+\ldots+\frac{9}{10^n}<\frac{4}{10^{m+1}}.$$

Во втором случае, при округлении по избытку, получим:

$$0\leqslant\sigma-\tilde{S}=\frac{1}{10^m}-\frac{p_{m+1}}{10^{m+1}}-\ldots-\frac{p_n}{10^n}\leqslant\frac{1}{10^m}-\frac{6}{10^{m+1}}=\frac{4}{10^{m+1}}.$$

Поэтому в обоих случаях имеем:

$$|\tilde{S}-\sigma|\leqslant\frac{4}{10^{m+1}}$$

и, следовательно,

$$|S-\sigma|\leqslant|S-\tilde{S}|+|\tilde{S}-\sigma|\leqslant\frac{1}{10^{m+1}}+\frac{4}{10^{m+1}}=\frac{1}{2}\cdot 10^{-m}.$$

Таким образом,

$$S=\sigma\pm\frac{1}{2}\cdot 10^{-m}.$$

Если $p_{m+1}=4$ или $p_{m+1}=5$, то следует увеличить точность вычислений приближенной суммы $\tilde{S}$, привлекая очередной десятичный разряд.

В частном случае, если $p_{m+1}=4$ и известно, что

$$S<\tilde{S},$$

то $\sigma(8)$ есть приближенное значение суммы $S$ с точностью до $\frac{1}{2}\cdot 10^{-m}$ (по недостатку).

Аналогично, если $p_{m+1}=5$ и

$$S>\tilde{S},$$

то $\sigma(8')$ есть приближенное значение суммы $S$ с точностью до $\frac{1}{2}\cdot 10^{-m}$ (по избытку).

Для оценки остатка ряда (1)

$$R_n=a_{n+1}+a_{n+2}+\ldots$$

полезны следующие теоремы, которые мы приводим без доказательств [1].

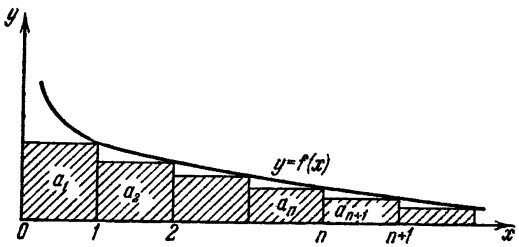


Рис. 4.

Теорема 1. *Если члены ряда* (1) *представляют собой соответствующие значения положительной монотонно убывающей функции* $f(x)$, *т. е.*

$$a_n=f(n) \qquad (n=1,\ 2,\ \ldots), \tag{9}$$

*то* (рис. 4)

$$\int_{n+1}^{\infty} f(x)\,dx<R_n<\int_{n}^{\infty} f(x)\,dx.$$

Теорема 2. *Если ряд* (1) — *знакочередующийся:*

$$a_1>0, \qquad a_2<0, \qquad a_3>0, \ \ldots$$

*и модули его членов монотонно убывают, то*

$$|R_n| \leqslant |a_{n+1}|$$

*и*

$$\operatorname{sgn} R_n = \operatorname{sgn} a_{n+1}\ ^{*)}.$$

Пример. Найти сумму ряда

$$S = \frac{1}{1^3} + \frac{1}{2^3} + \frac{1}{3^3} + \ldots + \frac{1}{n^3} + \ldots \tag{10}$$

с точностью до 0,001.

Решение. Примем остаточную погрешность

$$\varepsilon_1 = \frac{1}{4} \cdot 10^{-3} = \frac{1}{4000}.$$

Члены ряда (10) представляют собой соответствующие значения монотонно убывающей функции

$$f(x) = \frac{1}{x^3}.$$

Поэтому для $n$-го остатка ряда

$$R_n = \sum_{k=n+1}^{\infty} \frac{1}{k^3}$$

имеем оценку

$$R_n \leqslant \int_n^{\infty} \frac{dx}{x^3} = \frac{1}{2n^2}.$$

Решая неравенство

$$\frac{1}{2n^2} \leqslant \frac{1}{4000},$$

получим:

$$n \geqslant \sqrt{2000} \approx 44{,}7.$$

Примем $n = 45$.

Предельную погрешность суммирования выберем равной

$$\varepsilon_2 = \frac{1}{4} \cdot 10^{-3};$$

отсюда допустимая предельная абсолютная погрешность слагаемых частной суммы $S_{45}$ ряда (10) есть:

$$\frac{\varepsilon_2}{n} \leqslant \frac{\frac{1}{4} \cdot 10^{-3}}{45} = \frac{5}{9} \cdot 10^{-5}.$$

---

*) $\operatorname{sgn} R_n$ обозначает *знак* числа $R_n$, т. е. $\operatorname{sgn} R_n = +1$, если $R_n > 0$; $\operatorname{sgn} R_n = -1$, если $R_n < 0$; $\operatorname{sgn} R_n = 0$, если $R_n = 0$.

Положим

$$\frac{\varepsilon_2}{n} = \frac{1}{2} \cdot 10^{-5},$$

т. е. члены ряда (10) будем вычислять с пятью верными, в узком смысле, десятичными знаками после запятой. Ниже выписаны соответствующие значения слагаемых и результаты частичного суммирования:

| | | |
|---|---|---|
| 1,00000 | 0,00024 | 0,00003 |
| 0,12500 | 0,00020 | 0,00003 |
| 0,03704 | 0,00017 | 0,00003 |
| 0,01562 | 0,00014 | 0,00003 |
| 0,00800 | 0,00012 | 0,00002 |
| 0,00463 | 0,00011 | 0,00002 |
| 0,00292 | 0,00009 | 0,00002 |
| 0,00195 | 0,00008 | 0,00002 |
| 0,00137 | 0,00007 | 0,00002 |
| 0,00100 | 0,00006 | 0,00002 |
| 0,00075 | 0,00006 | 0,00001 |
| 0,00058 | 0,00005 | 0,00001 |
| 0,00046 | 0,00004 | 0,00001 |
| 0,00036 | 0,00004 | 0,00001 |
| 0,00030 | 0,00004 | 0,00001 |
| 1,19998 | 0,00151 | 0,00029 |

Следовательно,

$$S_{45} = 1{,}19998 + 0{,}00151 + 0{,}00029 = 1{,}20178.$$

Округляя это значение до тысячных, получим приближенное значение суммы

$$S \approx 1{,}202.$$

Так как погрешность округления

$$\varepsilon_3 = 0{,}00022 < \frac{1}{4} \cdot 10^{-3},$$

то суммарная погрешность найденного результата не превышает

$$\varepsilon < \frac{1}{4} \cdot 10^{-3} + \frac{1}{4} \cdot 10^{-3} + \frac{1}{4} \cdot 10^{-3} < \frac{3}{4} \cdot 10^{-3}.$$

Таким образом,

$$S = 1{,}202 \pm 0{,}001.$$

Оценку можно провести точнее, если учесть знаки округлений. Для сравнения приводим значение суммы $S$ с точностью до $\frac{1}{2} \cdot 10^{-6}$ [2]:

$$S = 1{,}202057.$$

Замечание. Так как расчет суммарной погрешности представляет собой весьма трудоемкую операцию, то на практике для обеспечения заданной точности $\varepsilon = 10^{-m}$ все промежуточные вычисления производят с одним или двумя запасными знаками. При этом не вполне строго предполагают, что допущенные погрешности не повлияют на десятичные знаки $m$-го разряда искомого результата.

Заметим, что при решении этого примера приходилось находить сумму сравнительно большого числа слагаемых. На практике данный ряд стараются предварительно преобразовать так, чтобы для достижения искомого результата можно было ограничиться малым количеством членов. Такого рода преобразование называется *улучшением сходимости ряда* и во многих случаях значительно экономит время вычислений. Этот вопрос рассмотрен в главе VI.

## § 5. Вычисление значений аналитической функции

Действительная функция $f(x)$ называется аналитической в точке $\xi$, если в некоторой окрестности $|x-\xi|<R$ этой точки функция разлагается в степенной ряд *(ряд Тейлора)*:

$$f(x)=f(\xi)+f'(\xi)(x-\xi)+\frac{f''(\xi)}{2!}(x-\xi)^2+\ldots$$
$$\ldots+\frac{f^{(n)}(\xi)}{n!}(x-\xi)^n+\ldots \qquad (1)$$

При $\xi=0$ получаем *ряд Маклорена*

$$f(x)=f(0)+f'(0)\,x+\frac{f''(0)}{2!}x^2+\ldots+\frac{f^{(n)}(0)}{n!}x^n+\ldots \qquad (2)$$

Разность

$$R_n(x)=f(x)-\sum_{k=0}^{n}\frac{f^{(k)}(\xi)}{k!}(x-\xi)^k$$

называется *остаточным членом* и представляет собой ошибку при замене функции $f(x)$ *полиномом Тейлора*

$$P_n(x)=\sum_{k=0}^{n}\frac{f^{(k)}(\xi)}{k!}(x-\xi)^k.$$

Как известно [1],

$$R_n(x)=\frac{f^{(n+1)}(\xi+\theta(x-\xi))}{(n+1)!}(x-\xi)^{n+1}, \qquad (3)$$

где $0<\theta<1$. В частности, для ряда Маклорена (2) имеем [1]:

$$R_n(x)=\frac{f^{(n+1)}(\theta x)}{(n+1)!}x^{n+1}, \qquad (4)$$

где $0<\theta<1$. Имеются также другие формы остаточных членов.

Разложение функции в ряд Тейлора во многих случаях является удобным способом вычисления значений этой функции.

Если $f(\xi)$ известно и требуется найти значение $f(\xi+h)$, где $h$ — «малая поправка», то формулу (1) выгодно записывать в виде

$$f(\xi+h)=f(\xi)+f'(\xi)h+\frac{f''(\xi)}{2!}h^2+\ldots+\frac{f^{(n)}(\xi)}{n!}h^n+R_n(h), \qquad (5)$$

где

$$R_n(h)=\frac{f^{(n+1)}(\xi+\theta h)}{(n+1)!}h^{n+1} \qquad (0<\theta<1).$$

Пример. Приближенно найти $\sqrt{10}$.

Решение. Имеем:

$$\sqrt{10}=\sqrt{3^2+1}=3\left(1+\frac{1}{9}\right)^{\frac{1}{2}}. \qquad (6)$$

Полагая

$$f(x)=(1+x)^{\frac{1}{2}},$$

последовательно получим:

$$f'(x)=\frac{1}{2}(1+x)^{-\frac{1}{2}},$$

$$f''(x)=-\frac{1}{4}(1+x)^{-\frac{3}{2}},$$

$$f'''(x)=\frac{3}{8}(1+x)^{-\frac{5}{2}},$$

$$f^{\mathrm{IV}}(x)=-\frac{15}{16}(1+x)^{-\frac{7}{2}}.$$

Отсюда, приняв $\xi=0$, $h=\frac{1}{9}$ и учитывая, что

$$f(0)=1, \quad f'(0)=\frac{1}{2}, \quad f''(0)=-\frac{1}{4}, \quad f'''(0)=\frac{3}{8},$$

в силу формулы (5) находим:

$$\left(1+\frac{1}{9}\right)^{\frac{1}{2}}=1+\frac{1}{2}\cdot\frac{1}{9}-\frac{1}{8}\cdot\frac{1}{81}+\frac{1}{16}\cdot\frac{1}{729}+R_3=$$
$$=1+0{,}05556-0{,}00154+0{,}00009+R_3=1{,}05411+R_3, \qquad (7)$$

где

$$R_3=-\frac{1}{24}\cdot\frac{15}{16}\cdot\left(1+\frac{\theta}{9}\right)^{-\frac{7}{2}}\cdot\frac{1}{6561}=$$
$$=-\frac{10}{1\,680\,616}\cdot\left(1+\frac{\theta}{9}\right)^{-\frac{7}{2}} \qquad (0<\theta<1).$$

Очевидно,

$$|R_3| < \frac{10}{1\,680\,616} < 6 \cdot 10^{-6}.$$

Из формул (6) и (7) получаем:

$$\sqrt{10} = 3{,}16233 + E, \tag{8}$$

где

$$|E| < 3 \cdot \frac{1}{2} \cdot 10^{-5} + 3 \cdot 6 \cdot 10^{-6} = 3{,}3 \cdot 10^{-5}.$$

Округляя найденное значение до четырех знаков после запятой, окончательно будем иметь:

$$\sqrt{10} = 3{,}1623 \pm 6 \cdot 10^{-5}.$$

Для сравнения приводим табличное значение

$$\sqrt{10} = 3{,}1622777 \ldots$$

## § 6. Вычисление значений показательной функции

Для экспоненциальной функции $e^x$ справедливо разложение [1]

$$e^x = 1 + x + \frac{x^2}{2!} + \ldots + \frac{x^n}{n!} + \ldots, \tag{1}$$

интервал сходимости которого $-\infty < x < +\infty$. Остаточный член ряда (1) имеет вид

$$R_n(x) = \frac{e^{\theta x}}{(n+1)!} x^{n+1} \qquad (0 < \theta < 1). \tag{2}$$

При больших по модулю значениях $x$ ряд (1) мало пригоден для вычислений. Поэтому обычно поступают следующим образом: пусть

$$x = E(x) + q,$$

где $E(x)$ — целая часть числа $x$ и $0 \leqslant q < 1$ — дробная часть его. Имеем:

$$e^x = e^{E(x)} \cdot e^q. \tag{3}$$

Первый множитель произведения (3) может быть найден с помощью умножения:

$$e^{E(x)} = \overbrace{ee \ldots e}^{E(x)\text{ раз}}, \qquad \text{если } E(x) \geqslant 0,$$

или

$$e^{E(x)} = \overbrace{\frac{1}{e} \cdot \frac{1}{e} \ldots \frac{1}{e}}^{-E(x)\text{ раз}}, \qquad \text{если } E(x) < 0,$$

где

$$e = 2{,}71828\ 18284\ 59045\ldots$$

и

$$\frac{1}{e} = 0{,}36787\ 94411\ 71442\ldots,$$

причем $e$ или $\frac{1}{e}$, для обеспечения заданной точности, следует взять с достаточно большим числом десятичных знаков (в настоящее время число $e$ определено свыше чем с 250 десятичными знаками).

Что касается второго множителя $e^q$ произведения (3), то для вычисления его пользуются приведенным выше разложением

$$e^q = \sum_{n=0}^{\infty} \frac{q^n}{n!}, \tag{4}$$

которое при $0 \leqslant q < 1$ образует быстро сходящийся ряд, так как на основании формулы (2) для остаточного члена $R_n(q)$ имеем оценку

$$0 \leqslant R_n(q) < \frac{3}{(n+1)!} q^{n+1}.$$

Выведем более точную формулу для оценки остатка $R_n(q)$ при $0 < q < 1$. Имеем:

$$R_n(q) = \frac{q^{n+1}}{(n+1)!} + \frac{q^{n+2}}{(n+2)!} + \frac{q^{n+3}}{(n+3)!} + \ldots =$$
$$= \frac{q^{n+1}}{(n+1)!}\left[1 + \frac{q}{n+2} + \frac{q^2}{(n+2)(n+3)} + \ldots\right] <$$
$$< \frac{q^{n+1}}{(n+1)!}\left[1 + \frac{q}{n+2} + \left(\frac{q}{n+2}\right)^2 + \ldots\right].$$

Отсюда, суммируя геометрическую прогрессию, стоящую в квадратных скобках, получим:

$$R_n(q) < \frac{q^{n+1}}{(n+1)!} \cdot \frac{1}{1 - \frac{q}{n+2}}; \tag{5}$$

или при $0 < q < 1$, учитывая, что

$$\frac{n+2}{n+1} < \frac{n+1}{n},$$

окончательно будем иметь:

$$0 < R_n(q) < \frac{q^{n+1}}{n!\,n},$$

т. е.

$$0 < R_n(q) < u_n \cdot \frac{q}{n}, \tag{6}$$

где $u_n = \frac{q^n}{n!}$ — последний сохраненный член.

Если остаточная погрешность $\varepsilon$ задана, то необходимое число членов $n$ можно найти подбором, решая неравенство

$$\frac{q^{n+1}}{n!\,n} < \varepsilon.$$

Приближенное вычисление $e^x$ по формуле (1) для малых $x$ удобно производить по схеме

$$e^x = u_0 + u_1 + u_2 + \ldots + u_n + R_n(x), \tag{7}$$

где

$$u_0 = 1, \quad u_k = \frac{xu_{k-1}}{k} \qquad (k = 1, 2, \ldots, n). \tag{8}$$

На счетных машинах вычисления удобно вести по схеме

$$u_k = \frac{x}{k} u_{k-1},$$

$$s_k = s_{k-1} + u_k \qquad (k = 0, 1, 2, \ldots, n),$$

где $u_0 = 1$, $s_{-1} = 0$, $s_0 = 1$. Число $s_n = \sum_{k=0}^{n} \frac{x^k}{k!}$ приближенно дает искомый результат $e^x$.

Если $\varepsilon$ — заданная допустимая остаточная погрешность и $n \geqslant 2|x| > 0$, то процесс суммирования следует прекратить, как только будет выполнено неравенство

$$|R_n(x)| \leqslant R_n(|x|) < \frac{|x|^{n+1}}{(n+1)!} \cdot \frac{1}{1 - \frac{|x|}{n+2}} <$$

$$< \frac{2|x|^{n+1}}{(n+1)!} = \frac{2|x|}{n+1} \cdot \frac{|x|^n}{n!} < |u_n| \leqslant \varepsilon,$$

т. е. если

$$|u_n(x)| \leqslant \varepsilon. \tag{9}$$

Иными словами, процесс суммирования прекращается, если последний выработанный член $u_n$ по модулю не превышает $\varepsilon$, при этом

$$|R_n(x)| < |u_n|.$$

Для расчета суммарной погрешности следует пользоваться общей схемой (§ 4).

Пример 1. Найти $\sqrt{e}$ с точностью до $10^{-5}$.

Решение. Принимаем остаточную погрешность

$$\varepsilon_1 = \frac{1}{4} \cdot 10^{-5} = 2{,}5 \cdot 10^{-6}.$$

Так как тогда число слагаемых в сумме (7), по грубой прикидке, порядка десяти, то слагаемые подсчитываем с точностью до $\frac{1}{2}\cdot 10^{-7}$, т. е. с двумя запасными десятичными знаками.

Полагая

$$u_0 = 1, \qquad u_k = \frac{u_{k-1}}{2k} \qquad (k = 1,\ 2,\ \ldots),$$

последовательно имеем:

$$\left.\begin{aligned}
u_0 &= 1 \\
u_1 &= \frac{1}{2} = 0{,}5000000 \\
u_2 &= \frac{u_1}{4} = 0{,}1250000 \\
u_3 &= \frac{u_2}{6} = 0{,}0208333 \\
u_4 &= \frac{u_3}{8} = 0{,}0026042 \\
u_5 &= \frac{u_4}{10} = 0{,}0002604 \\
u_6 &= \frac{u_5}{12} = 0{,}0000217 \\
u_7 &= \frac{u_6}{14} = 0{,}0000016 \\
\hline
&\phantom{=}\ 1{,}6487212
\end{aligned}\right\}$$

Округляя сумму до пяти десятичных знаков после запятой, получим:

$$\sqrt{e} = 1{,}64872, \tag{10}$$

с суммарной погрешностью

$$\varepsilon < 1{,}6\cdot 10^{-6} + 5\cdot\frac{1}{2}\cdot 10^{-7} + 1{,}2\cdot 10^{-6} = 3{,}05\cdot 10^{-6} < \frac{1}{2}\cdot 10^{-5},$$

т. е. все знаки результата (10) верные в узком смысле.

Для вычисления $e^x$ можно использовать также разложение в цепную дробь [4]

$$e^x = \left[0;\ \frac{1}{1},\ \frac{-2x}{2+x},\ \frac{x^2}{6},\ \frac{x^2}{10},\ \ldots,\ \frac{x^2}{4n+2},\ \ldots\right], \tag{11}$$

сходящуюся для любого значения $x$ (действительного или комплексного).

Пример 2. Найти $\sqrt{e}$, пользуясь формулой (11).

Решение. Полагая в формуле (11) $x=\frac{1}{2}$, составляем таблицу подходящих дробей для соответствующей цепной дроби.

| $k$ | $-1$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|
| $b_k$ | | 0 | 1 | $-1$ | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{4}$ |
| $a_k$ | 1 | 1 | 1 | $\frac{5}{2}$ | 6 | 10 | 14 |
| $P_k$ | 1 | 0 | 1 | $\frac{5}{2}$ | $\frac{61}{4}$ | $\frac{1\,225}{8}$ | $\frac{34\,361}{16}$ |
| $Q_k$ | 0 | 1 | 1 | $\frac{3}{2}$ | $\frac{37}{4}$ | $\frac{743}{8}$ | $\frac{20\,841}{16}$ |

Останавливаясь на 5-й подходящей дроби, имеем:

$$\sqrt{e} \approx \frac{P_5}{Q_5}=\frac{34361}{16}:\frac{20841}{16}=\frac{34361}{20841}=1{,}648721$$

с точностью до $\frac{1}{2}\cdot 10^{-6}$.

Для вычисления значений общей показательной функции $a^x$ $(a>0)$ можно использовать формулу

$$a^x=1+\ln a\cdot x+\frac{\ln^2 a}{2!}x^2+\frac{\ln^3 a}{3!}x^3+\ldots$$

## § 7. Вычисление значений логарифмической функции

Для натуральных логарифмов чисел, близких к единице, справедливо разложение [1]

$$\ln(1+x)=x-\frac{x^2}{2}+\frac{x^3}{3}-\frac{x^4}{4}+\ldots$$
$$\ldots+(-1)^{n-1}\frac{x^n}{n}+\ldots \qquad (-1<x\leqslant 1). \tag{1}$$

Формула (1) малопригодна для вычислений, так как диапазон чисел $0<1+x\leqslant 2$ невелик и, кроме того, при $|x|$, близком к единице, ряд (1) сходится медленно.

Введем более удобную формулу для вычислений натуральных логарифмов чисел. Заменяя $x$ в формуле (1) на $-x$, будем иметь:

$$\ln(1-x)=-x-\frac{x^2}{2}-\frac{x^3}{3}-\frac{x^4}{4}-\ldots-\frac{x^n}{n}-\ldots \tag{2}$$

Вычитая почленно из формулы (1) формулу (2), находим:

$$\ln\frac{1-x}{1+x}=-2\left(x+\frac{x^3}{3}+\frac{x^5}{5}+\ldots\right).$$

Полагая

$$\frac{1-x}{1+x}=z,$$

получим:

$$x=\frac{1-z}{1+z}$$

и, следовательно,

$$\ln z=-2\left[\frac{1-z}{1+z}+\frac{1}{3}\left(\frac{1-z}{1+z}\right)^3+\frac{1}{5}\left(\frac{1-z}{1+z}\right)^5+\ldots\right] \tag{3}$$

при $0<z<+\infty$.

Пусть $x$ — положительное число. Представим его в виде

$$x=2^m\cdot z,$$

где $m$ — целое число и $\frac{1}{2}\leqslant z<1$. Тогда, полагая

$$\frac{1-z}{1+z}=\xi,$$

где

$$0<\xi\leqslant\frac{1-\frac{1}{2}}{1+\frac{1}{2}}=\frac{1}{3},$$

и применяя формулу (3), будем иметь:

$$\ln x=\ln 2^m z=m\ln 2+\ln z=$$
$$=m\ln 2-2\left(\xi+\frac{\xi^3}{3}+\ldots+\frac{\xi^{2n-1}}{2n-1}\right)-R_n,$$

где

$$R_n=2\left(\frac{\xi^{2n+1}}{2n+1}+\frac{\xi^{2n+3}}{2n+3}+\frac{\xi^{2n+5}}{2n+5}+\ldots\right)<$$
$$<2\cdot\frac{\xi^{2n+1}}{2n+1}(1+\xi^2+\xi^4+\ldots)<\frac{2}{1-\xi^2}\cdot\frac{\xi^{2n+1}}{2n+1}.$$

При $0<\xi\leqslant\frac{1}{3}$ имеем:

$$\frac{2}{1-\xi^2}\leqslant\frac{9}{4},$$

и поэтому

$$0<R_n<\frac{9}{4}\cdot\frac{\xi^{2n+1}}{2n+1} \tag{4}$$

или, более грубо,

$$0<R_n<\frac{1}{4(2n+1)}\cdot\left(\frac{1}{3}\right)^{2n-1}.$$

Введя обозначение

$$u_k = \frac{\xi^{2k-1}}{2k-1} \qquad (k = 1,\ 2,\ \ldots),$$

получим:

$$\ln x = m \ln 2 - 2\,(u_1 + u_2 + \ldots + u_n) - R_n, \tag{5}$$

где

$$\ln 2 = 0{,}69314718\ \ldots$$

Процесс суммирования прекращается, как только

$$u_n < 4\varepsilon,$$

где $\varepsilon$ — допустимая остаточная погрешность, так как тогда в силу формулы (4) имеем:

$$R_n < \frac{9}{4}\,\xi^2 \cdot \frac{\xi^{2n-1}}{2n-1} \leqslant \frac{1}{4}\,u_n < \varepsilon.$$

Предельную погрешность суммы $\sum\limits_{k=1}^{n} u_k$ можно оценить, задавшись определенным количеством десятичных знаков слагаемых и установив на основании формулы (4) примерное число слагаемых $n$.

Пример. Найти $\ln 3$ с точностью до $10^{-5}$.

Решение. Вычисления будем производить с двумя запасными знаками. Положим

$$3 = 2^2 \cdot \frac{3}{4} = 2^2 \cdot 0{,}75.$$

Отсюда $z = 0{,}75$ и

$$\xi = \frac{1-z}{1+z} = \frac{0{,}25}{1{,}75} = \frac{1}{7} = 0{,}1428571.$$

Имеем:

$$\left.\begin{aligned} u_1 &= \xi = 0{,}1428571 \\ u_2 &= \frac{\xi^3}{3} = 0{,}0009718 \\ u_3 &= \frac{\xi^5}{5} = 0{,}0000119 \\ u_4 &= \frac{\xi^7}{7} = 0{,}0000002 \end{aligned}\right\}$$

$$\overline{\phantom{0{,}1438410}}$$

$$0{,}1438410$$

Используя формулу (5), получаем:

$$\ln 3 = 2 \cdot 0{,}69314718 - 2 \cdot 0{,}1438410 = 1{,}09861.$$

Замечание. Можно также вычислять натуральные логарифмы чисел, исходя из представления

$$x = e^p z,$$

где $p$ — целое число и $\frac{1}{e} < z \leqslant 1$ (см. [5]).

Для вычисления десятичных логарифмов используется формула

$$\lg x = M \ln x,$$

где

$$M = \lg e = 0{,}43429\,44819\,03252\ \ldots$$

## § 8. Вычисление значений тригонометрических функций

А. Вычисление значений синуса и косинуса

С помощью формул приведения аргумент $x$ можно заключить в отрезок $0 \leqslant x \leqslant \frac{\pi}{2}$. Если $0 \leqslant x \leqslant \frac{\pi}{4}$, то имеем:

$$\sin x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!}, \tag{1}$$

если же $\frac{\pi}{4} \leqslant x \leqslant \frac{\pi}{2}$, то полагают

$$\sin x = \cos z = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{(2n)!}, \tag{2}$$

где $z = \frac{\pi}{2} - x$ и $0 \leqslant z \leqslant \frac{\pi}{4}$.

Сумму ряда (1) удобно вычислять процессом суммирования

$$\sin x = u_1 + u_2 + \ldots + u_n + R_n, \tag{3}$$

где слагаемые $u_k$ $(k = 1, 2, \ldots, n)$ последовательно находятся с помощью рекуррентного соотношения

$$u_1 = x, \quad u_{k+1} = -\frac{x^2}{2k(2k+1)} u_k \quad (k = 1, 2, \ldots, n-1).$$

Так как ряд (1) знакочередующийся с монотонно убывающими, по модулю, членами, то для остаточного члена $R_n$ справедлива оценка

$$|R_n| \leqslant \frac{x^{2n+1}}{(2n+1)!} = |u_{n+1}|$$

и

$$\operatorname{sgn} R_n = \operatorname{sgn} u_{n+1}.$$

Поэтому процесс суммирования можно прекратить, как только будет обнаружено, что

$$|u_n| \leqslant \varepsilon,$$

где $\varepsilon$ — заданная остаточная погрешность.

Аналогично

$$\cos z = v_1 + v_2 + \ldots + v_n + R_n,$$

где

$$v_1 = 1, \quad v_{k+1} = -\frac{x^2}{(2k-1)\,2k} v_k \quad (k = 1, 2, \ldots, n-1)$$

и

$$|R_n| \leqslant \frac{z^{2n}}{(2n)!} = |v_{n+1}|, \quad \operatorname{sgn} R_n = \operatorname{sgn} v_{n+1}.$$

Пример. Найти $\sin 20°30'$ с точностью до $10^{-5}$.

Решение. Имеем:

$$x = \operatorname{arc} 20°30' = \frac{\pi}{9} + \frac{\pi}{360} = 0{,}349066 + 0{,}008727 = 0{,}357793.$$

Применяя формулу (3), получим:

$$\left.\begin{aligned} u_1 &= x = \phantom{-}0{,}357793 \\ u_2 &= \frac{x^2 u_1}{2\cdot 3} = -0{,}007634 \\ u_3 &= \frac{x^2 u_2}{4\cdot 5} = +0{,}000049 \\ u_4 &= \frac{x^2 u_3}{6\cdot 7} = -0{,}000000 \end{aligned}\right\}$$

$$-0{,}350208$$

Отсюда

$$\sin 20°30' = 0{,}35021.$$

Аналогично вычисляются значения $\cos x$.

## Б. Вычисление тангенса

Можно считать, что $0 \leqslant x \leqslant \frac{\pi}{4}$. Для $\operatorname{tg} x$ при $|x| < \frac{\pi}{2}$ справедливо разложение [6]

$$\operatorname{tg} x = x + \frac{x^3}{3} + \frac{2x^5}{15} + \frac{17x^7}{315} + \frac{62x^9}{2835} + \ldots$$

Коэффициенты разложения выражаются через *числа Бернулли* (см. гл. XVI, § 12).

Значение тангенса просто вычисляется с помощью цепных дробей. Полагая

$$\operatorname{tg} x = \frac{x}{y},$$

в силу формулы Ламберта (гл. II, § 6) будем иметь:

$$y = \left[1;\ \frac{-x^2}{3},\ \frac{-x^2}{5},\ \ldots,\ \frac{-x^2}{2n+1},\ \ldots\right].$$

Для вычисления $y$ с точностью до $10^{-10}$ достаточно принять $n=7$. Тогда

$$y=1-x^2:(3-x^2:(5-x^2:(7-x^2:(9-\\ -x^2:(11-x^2:(13-x^2:15)))))). \quad (4)$$

Вычисление $y$ обычно производится с помощью схемы Горнера для деления (начиная с конца):

$$\begin{aligned} y_1 &= 13-x^2:15,\\ y_2 &= 11-x^2:y_1,\\ y_3 &= 9-x^2:y_2,\\ y_4 &= 7-x^2:y_3,\\ y_5 &= 5-x^2:y_4,\\ y_6 &= 3-x^2:y_5,\\ y=y_7 &= 1-x^2:y_6. \end{aligned}$$

Отсюда $\operatorname{tg} x=\frac{x}{y}$.

Пример. Найти $\operatorname{tg} 40°$.

Решение. Имеем:

$$x=\operatorname{arc} 40°=0{,}698132$$

и

$$x^2=0{,}487388.$$

Отсюда

$$\begin{aligned} y_1 &= 13-\frac{0{,}487388}{15}=12{,}967508;\\ y_2 &= 11-\frac{0{,}487388}{12{,}967508}=10{,}962413;\\ y_3 &= 9-\frac{0{,}487388}{10{,}962413}=8{,}955540;\\ y_4 &= 7-\frac{0{,}487388}{8{,}955540}=6{,}955577;\\ y_5 &= 5-\frac{0{,}487388}{6{,}955577}=4{,}929928;\\ y_6 &= 3-\frac{0{,}487388}{4{,}929928}=2{,}901137;\\ y=y_7 &= 1-\frac{0{,}487388}{2{,}901137}=0{,}832001 \end{aligned}$$

и, следовательно,

$$\operatorname{tg} 40°=\frac{0{,}698132}{0{,}832001}=0{,}839100.$$

В полученном результате все знаки верные.

## § 9. Вычисление значений гиперболических функций

А. Вычисление значений гиперболического синуса

Как известно,

$$\operatorname{sh} x = \frac{e^x - e^{-x}}{2},$$

причем

$$\operatorname{sh}(-x) = -\operatorname{sh} x.$$

Для гиперболического синуса справедливо разложение

$$\operatorname{sh} x = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \ldots \qquad (-\infty < x < +\infty).$$

Предполагая, что $x > 0$, вычисления удобно производить процессом суммирования

$$\operatorname{sh} x = u_1 + u_2 + \ldots + u_n + R_n,$$

где

$$u_1 = x, \quad u_{k+1} = \frac{x^2}{2k(2k+1)} u_k \qquad (k = 1, 2, \ldots, n-1)$$

и $R_n$ — остаточный член. При $n \geqslant x > 0$ имеем:

$$R_n = \frac{x^{2n+1}}{(2n+1)!} + \frac{x^{2n+3}}{(2n+3)!} + \frac{x^{2n+5}}{(2n+5)!} + \ldots <$$

$$< \frac{x^{2n+1}}{(2n+1)!}\left[1 + \frac{x^2}{(2n+2)(2n+3)} + \frac{x^4}{(2n+2)^2(2n+3)^2} + \ldots\right] <$$

$$< \frac{x^{2n+1}}{(2n+1)!} \cdot \frac{1}{1 - \frac{x^2}{(2n+2)(2n+3)}} < \frac{4}{3}\frac{x^{2n+1}}{(2n+1)!} = \frac{4}{3} u_{n+1}.$$

Так как, очевидно,

$$u_{n+1} = \frac{x^2}{2n(2n+1)} u_n < \frac{1}{4} u_n,$$

то

$$R_n < \frac{1}{3} u_n.$$

Б. Вычисление значений гиперболического косинуса

Как известно,

$$\operatorname{ch} x = \frac{e^x + e^{-x}}{2},$$

причем

$$\operatorname{ch}(-x) = \operatorname{ch} x.$$

Для гиперболического косинуса справедливо разложение

$$\operatorname{ch} x = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \ldots \qquad (-\infty < x < +\infty).$$

Вычисления удобно производить процессом суммирования

$$\operatorname{ch} x = v_1 + v_2 + \ldots + v_n + R_n,$$

где

$$v_1 = 1, \quad v_{k+1} = \frac{x^2}{(2k-1)\,2k}\, v_k \qquad (k = 1,\ 2,\ \ldots,\ n-1)$$

и $R_n$ — остаточный член. При $n \geqslant |x| > 0$ имеем:

$$R_n = \frac{x^{2n}}{(2n)!} + \frac{x^{2n+2}}{(2n+2)!} + \frac{x^{2n+4}}{(2n+4)!} + \ldots <$$

$$< \frac{x^{2n}}{(2n)!}\left[1 + \frac{x^2}{(2n+1)(2n+2)} + \frac{x^4}{(2n+1)^2(2n+2)^2} + \ldots\right] <$$

$$< \frac{x^{2n}}{(2n)!} \cdot \frac{1}{1 - \dfrac{x^2}{(2n+1)(2n+2)}} < \frac{4}{3} \cdot \frac{x^{2n}}{(2n)!} = \frac{4}{3}\, v_{n+1}.$$

Так как при $n \geqslant 1$ справедливо неравенство

$$v_{n+1} = \frac{x^2}{(2n-1)\,2n}\, v_n \leqslant \frac{1}{2}\, v_n,$$

то

$$R_n < \frac{2}{3}\, v_n.$$

### В. Вычисление гиперболического тангенса

Имеем:

$$\operatorname{th} x = \frac{\operatorname{sh} x}{\operatorname{ch} x} = \frac{e^x - e^{-x}}{e^x + e^{-x}},$$

где

$$\operatorname{th}(-x) = -\operatorname{th} x.$$

При малых $|x|$ для вычисления значений гиперболического тангенса можно использовать разложение

$$\operatorname{th} x = x - \frac{x^3}{3} + \frac{2x^5}{15} - \frac{17x^7}{315} + \frac{62x^9}{2835} + \ldots \left(|x| < \frac{\pi}{2}\right).$$

При любом $x$ значение гиперболического тангенса выражается цепной дробью

$$\operatorname{th} x = \left[0;\ \frac{x}{1},\ \frac{x^2}{3},\ \frac{x^2}{5},\ \ldots,\ \frac{x^2}{2n-1},\ \ldots\right],$$

причем так как при $x > 0$ элементы этой дроби положительны, то $\operatorname{th} x$ при $x > 0$ заключен между двумя соседними подходящими дробями.

Если $x > 0$ велико, то $\operatorname{th} x$ удобно вычислять, применяя формулу

$$\operatorname{th} x = 1 - \frac{2}{e^{2x} + 1}.$$

## § 10. Применение метода итерации для приближенного вычисления значений функции

Пусть требуется вычислить значение непрерывной функции

$$y = f(x) \tag{1}$$

для заданного значения аргумента $x$. Если функция (1) достаточно сложна и нужно подсчитать большое количество ее значений, то вычисления обычно производятся на счетных машинах. Может случиться, что в силу конструктивных особенностей машины непосредственное вычисление значений функции по формуле (1) будет затруднительным. При этом самые простые действия могут оказаться «сложными» и даже невыполнимыми. Так, например, существуют счетные машины «без деления». Тогда во многих случаях оказывается полезным следующий прием. Запишем функцию (1) в неявном виде

$$F(x, y) = 0. \tag{2}$$

Предположим, что $F(x, y)$ непрерывна и имеет непрерывную частную производную $F'_y(x, y) \neq 0$.

Пусть $y_n$—приближенное значение $y$. Применяя теорему Лагранжа, будем иметь:

$$F(x, y_n) = F(x, y_n) - F(x, y) = (y_n - y) F'_y(x, \bar{y}_n),$$

где $\bar{y}_n$ — некоторое промежуточное значение между $y_n$ и $y$. Отсюда

$$y = y_n - \frac{F(x, y_n)}{F'_y(x, \bar{y}_n)}. \tag{3}$$

Значение $\bar{y}_n$ нам не известно. Полагая $\bar{y}_n \approx y_n$, для вычисления значения $y$ получим *итеративный процесс* [7]

$$y_{n+1} = y_n - \frac{F(x, y_n)}{F'_y(x, y_n)} \qquad (n = 0, 1, 2, \ldots). \tag{4}$$

Формула (3) имеет простой геометрический смысл. Зафиксируем значение $x$ и рассмотрим график функции

$$z = F(x, y). \tag{4'}$$

Из формулы (4) вытекает, что наш процесс представляет собой метод Ньютона (см. гл. IV, § 5), примененный к функции (4), т. е. последовательные приближения $y_{n+1}$ получаются как абсциссы точки пересечения с осью $Oy$ касательной к кривой (4), проведенной при $y = y_n$ $(n = 0, 1, 2, \ldots)$ (рис. 5). Сходимость процесса будет обеспечена, если

$$F'_y(x, y) \quad \text{и} \quad F''_{yy}(x, y)$$

сохраняют постоянные знаки в рассматриваемом интервале, содержащем корень $y$.

Начальное значение $y_0$, вообще говоря, произвольно и выбирается по возможности близким к искомому значению $y$. Процесс итерации продолжается до тех пор, пока в пределах заданной точности $\varepsilon$ два последовательных значения $y_n$ и $y_{n-1}$ не совпадут между собой: $|y_{n-1}-y_n|<\varepsilon$. При этом, строго говоря, не гарантируется, что

$$|y-y_n|<\varepsilon, \qquad (5)$$

поэтому в каждом конкретном случае требуется дополнительное исследование.

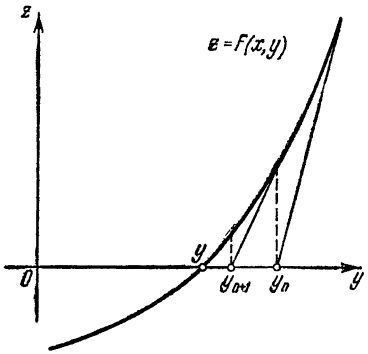


Рис. 5.

Достоинством итеративных процессов является однообразие операций и вследствие этого сравнительно легкая программируемость.

Заметим, что представление $F(x, y)=0$ для заданной функции (1) можно реализовать бесчисленным множеством способов. Этим следует воспользоваться, чтобы получить быстро сходящийся итерационный процесс. В следующих параграфах приведены типы основных процессов.

## § 11. Вычисление обратной величины

Пусть $y=\frac{1}{x}$.

Для определенности будем считать, что $x>0$. Положим

$$F(x, y)\equiv x-\frac{1}{y}=0,$$

тогда

$$F'_y(x, y)=\frac{1}{y^2}.$$

Применяя формулу (4) § 10, будем иметь:

$$y_{n+1}=y_n-\frac{x-\frac{1}{y_n}}{\frac{1}{y_n^2}}$$

или

$$y_{n+1}=y_n(2-xy_n) \qquad (n=0,\ 1,\ 2,\ \ldots), \qquad (1)$$

т. е. мы получили итерационный процесс, не содержащий деления. Начальное значение $y_0$ выбирается следующим образом. Пусть

аргумент $x$ записан в двоичной системе:

$$x = 2^m x_1, \text{ где } m \text{ — целое число и } \frac{1}{2} \leqslant x_1 < 1.$$

Тогда полагают

$$y_0 = 2^{-m}. \tag{2}$$

Выясним условия сходимости процесса (1). Из формулы (1) имеем:

$$\frac{1}{x} - y_n = \frac{1}{x} - 2y_{n-1} + xy_{n-1}^2 = x\left(\frac{1}{x} - y_{n-1}\right)^2; \tag{3}$$

отсюда

$$\frac{1}{x} - y_n = x^{2^n-1}\left(\frac{1}{x} - y_0\right)^{2^n} = \frac{1}{x}(1 - xy_0)^{2^n}. \tag{4}$$

Для сходимости процесса (4) необходимо и достаточно, чтобы было выполнено неравенство

$$|1 - xy_0| < 1$$

или

$$-1 < 1 - xy_0 < 1.$$

Таким образом, окончательно получаем следующий результат: если

$$0 < xy_0 < 2, \tag{5}$$

то

$$\lim_{n \to \infty} y_n = \frac{1}{x}.$$

Заметим, что при нашем выборе $y_0$ (2) имеем:

$$xy_0 = 2^m x_1 \cdot 2^{-m} = x_1;$$

поэтому

$$\frac{1}{2} \leqslant xy_0 < 1 \tag{6}$$

и, значит, условие (5) выполнено. Кроме того, из формулы (3) выводим:

$$\left|\frac{1}{x} - y_n\right| \leqslant \frac{1}{x}\left(\frac{1}{2}\right)^{2^n} \leqslant 2y_0\left(\frac{1}{2}\right)^{2^n},$$

т. е. сходимость итерационного процесса чрезвычайно быстрая.

Выведем другую оценку ошибки значения $y_n$, иногда практически более удобную. Прежде всего, заметим, что последовательные приближения $y_0, y_1, y_2 \ldots$ в данном случае получаются по методу Ньютона, примененному к гиперболе

$$z = x - \frac{1}{y} \quad (x = \text{const})$$

(рис. 6). Из неравенства (6) и формулы (3) следует, что

$$0 < y_n < \frac{1}{x} \quad (n = 0,\ 1,\ 2,\ \ldots).$$

Кроме того, так как

$$y_n - y_{n-1} = y_{n-1}(1 - xy_{n-1}) = xy_{n-1}\left(\frac{1}{x} - y_{n-1}\right) \geqslant 0, \tag{7}$$

то последовательные приближения $y_n$ монотонно возрастают:

$$y_0 \leqslant y_1 \leqslant y_2 \leqslant \ldots$$

Из формулы (7) имеем:

$$\frac{1}{x} - y_{n-1} = \frac{1}{xy_{n-1}}(y_n - y_{n-1}),$$

или так как

$$xy_{n-1} \geqslant xy_0 \geqslant \frac{1}{2},$$

то

$$\frac{1}{x} - y_{n-1} \leqslant 2(y_n - y_{n-1}).$$

Отсюда

$$\frac{1}{x} - y_n \leqslant y_n - y_{n-1}.$$

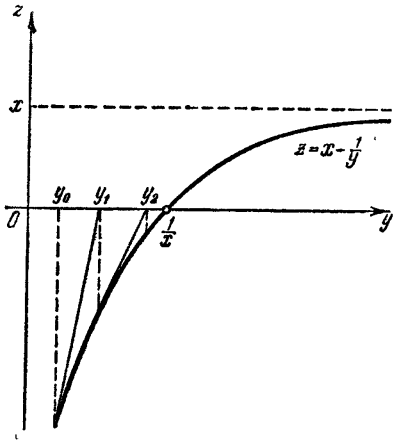


Рис 6.

Таким образом, если будет обнаружено, что $y_n - y_{n-1} < \varepsilon$, то и истинная погрешность

$$0 < \frac{1}{x} - y_n < \varepsilon.$$

Пример. С помощью (1) найти значение функции $y = \frac{1}{x}$ при $x = 3$.

Решение. Здесь $x = 2^2 \cdot \frac{3}{4}$. Полагаем $y_0 = \frac{1}{4}$, тогда

$$y_1 = \frac{1}{4}\left(2 - \frac{3}{4}\right) = \frac{5}{16} = 0{,}312;$$

$$y_2 = 0{,}312\,(2 - 3 \cdot 0{,}312) = 0{,}332 \text{ и т. д.}$$

Процесс итерации быстро сходится.

## § 12. Вычисление квадратного корня

Пусть

$$y=\sqrt{x} \qquad (x>0). \tag{1}$$

Положим

$$F(x,\ y)\equiv y^2-x=0,$$

тогда

$$F'_y(x,\ y)=2y.$$

Применяя формулу (4) § 10, имеем:

$$y_{n+1}=y_n-\frac{y_n^2-x}{2y_n}$$

или

$$y_{n+1}=\frac{1}{2}\left(y_n+\frac{x}{y_n}\right) \tag{2}$$

$(n=0,\ 1,\ 2,\ \ldots)$ — *процесс Герона.*

Последовательные приближения $y_0,\ y_1,\ y_2,\ \ldots$, очевидно, получаются по методу Ньютона, примененному к параболе

$$z=y^2-x \quad (x=\text{const})$$

(рис. 7).

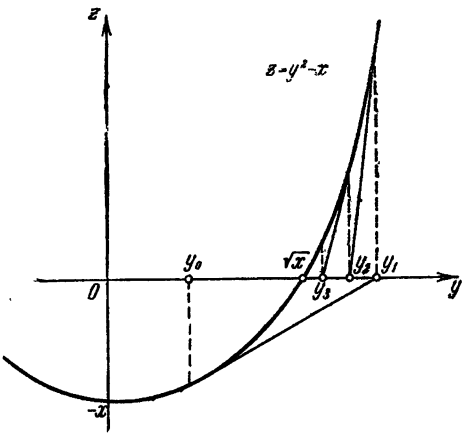


Рис. 7.

Заметим, что если за $y_0$ принять табличное значение, дающее $\sqrt{x}$ с относительной погрешностью $|\delta|$, то $y_1$, определенное по формуле (2), даст значение $\sqrt{x}$ приблизительно с относительной погрешностью $\frac{1}{2}\delta^2$.

Действительно, полагая

$$y_0=\sqrt{x}(1+\delta)$$

и пренебрегая степенями $\delta$, выше 3-й, будем иметь:

$$y_1=\frac{1}{2}\left(y_0+\frac{x}{y_0}\right)=\frac{1}{2}\left[\sqrt{x}(1+\delta)+\sqrt{x}(1+\delta)^{-1}\right]=$$
$$=\frac{1}{2}\sqrt{x}(1+\delta+1-\delta+\delta^2)=\sqrt{x}\left(1+\frac{\delta^2}{2}\right).$$

Отсюда получаем важный вывод: *при применении процесса Герона число верных цифр примерно удваивается на каждом этапе по сравнению с первоначальным количеством.*

Пример 1. Для $y=\sqrt{2}$ приближенно имеем:

$$y_0=1{,}4.$$

Уточняя это значение, получаем:

$$y_1=\frac{1}{2}\left(1{,}4+\frac{2}{1{,}4}\right)=0{,}7+0{,}714=1{,}414.$$

Еще раз повторяя процесс, будем иметь:

$$y_2=\frac{1}{2}\left(1{,}414+\frac{2}{1{,}414}\right)=0{,}707+0{,}7072136=1{,}4142136,$$

причем восемь или семь десятичных знаков являются верными. Действительно,

$$\sqrt{2}=1{,}41421356\ldots$$

Выясним условия сходимости процесса Герона. Из формулы (2), заменяя $n+1$ на $n$, при $y_0\neq 0$ имеем:

$$y_n-\sqrt{x}=\frac{1}{2y_{n-1}}(y_{n-1}-\sqrt{x})^2$$

и

$$y_n+\sqrt{x}=\frac{1}{2y_{n-1}}(y_{n-1}+\sqrt{x})^2.$$

Отсюда

$$\frac{y_n-\sqrt{x}}{y_n+\sqrt{x}}=\left(\frac{y_{n-1}-\sqrt{x}}{y_{n-1}+\sqrt{x}}\right)^2. \tag{3}$$

Следовательно,

$$\frac{y_n-\sqrt{x}}{y_n+\sqrt{x}}=\left(\frac{y_0-\sqrt{x}}{y_0+\sqrt{x}}\right)^{2^n}$$

и

$$y_n-\sqrt{x}=2\sqrt{x}\cdot\frac{q^{2^n}}{1-q^{2^n}}, \tag{4}$$

где

$$q=\frac{y_0-\sqrt{x}}{y_0+\sqrt{x}}. \tag{5}$$

Из формулы (4) вытекает, что процесс Герона сходится при

$$|q|<1,$$

т. е. если

$$y_0>0.$$

В этом случае, очевидно, имеем:

$$\lim_{n\to\infty} y_n=\sqrt{x},$$

причем

$$y_n\geqslant\sqrt{x}\qquad(n=1,\ 2,\ \ldots).$$

Заметим, что

$$y_{n-1}-y_n=y_{n-1}-\frac{1}{2}\left(y_{n-1}+\frac{x}{y_{n-1}}\right)=\frac{y_{n-1}^2-x}{2y_{n-1}}>0, \qquad (6)$$

поэтому приближения $y_n$ при $n\geqslant 1$ образуют монотонно убывающую последовательность

$$y_1\geqslant y_2\geqslant\ldots\geqslant y_{n-1}\geqslant y_n\geqslant\ldots\geqslant\sqrt{x}\,^{*)}.$$

При работе на счетной машине число $x$ удобно записывать в двоичной системе

$$x=2^m x_1, \qquad \text{где } m\text{—целое число и } \frac{1}{2}\leqslant x_1<1.$$

Тогда за нулевое приближение обычно принимают:

$$y_0=2^{E\left(\frac{m}{2}\right)}, \qquad (7)$$

где $E\left(\frac{m}{2}\right)$—целая часть числа $\frac{m}{2}$.

Пример 2. Найти $\sqrt{5}$.

Решение. Здесь $x=5=2^3\cdot\frac{5}{8}$. Поэтому

$$y_0=2^{E\left(\frac{3}{2}\right)}=2.$$

По формуле (2) последовательно находим:

$$y_1=\frac{1}{2}\left(2+\frac{5}{2}\right)=2{,}25,$$

$$y_2=\frac{1}{2}\left(2{,}25+\frac{5}{2{,}25}\right)=\frac{1}{2}(2{,}25+2{,}2222)=2{,}2361$$

и т. д. По таблицам квадратных корней имеем:

$$\sqrt{5}=2{,}236068\ldots$$

Оценим величину $|q|$, выражаемую формулой (5), исходя из значения $y_0$, определяемого формулой (7).

Если $m=2p$ есть число четное, то имеем:

$$y_0=2^{E\left(\frac{m}{2}\right)}=2^p>\sqrt{x}$$

и, следовательно,

$$|q|=\frac{y_0-\sqrt{x}}{y_0+\sqrt{x}}=\frac{2^p-2^p\sqrt{x_1}}{2^p+2^p\sqrt{x_1}}=\frac{1-\sqrt{x_1}}{1+\sqrt{x_1}}\leqslant\frac{1-\sqrt{\frac{1}{2}}}{1+\sqrt{\frac{1}{2}}}=(\sqrt{2}-1)^2.$$

---

*) Знак равенства может иметь место только при $y_0=\sqrt{x}$.

Аналогично, если $m=2p+1$ есть число нечетное, то

$$y_0=2^{E\left(\frac{m}{2}\right)}=2^p\leqslant\sqrt{x}.$$

Поэтому

$$|q|=\frac{\sqrt{x}-y_0}{\sqrt{x}+y_0}=\frac{2^p\sqrt{2x_1}-2^p}{2^p\sqrt{2x_1}+2^p}=$$

$$=\frac{\sqrt{2x_1}-1}{\sqrt{2x_1}+1}=1-\frac{2}{\sqrt{2x_1}+1}<1-\frac{2}{\sqrt{2}+1}=(\sqrt{2}-1)^2.$$

Таким образом, всегда имеем:

$$|q|\leqslant(\sqrt{2}-1)^2=0{,}1716\ldots<\frac{1}{5}.$$

Отсюда на основании формулы (4) получим:

$$0\leqslant y_n-\sqrt{x}<2\sqrt{x}\cdot\frac{\left(\frac{1}{5}\right)^{2^n}}{1-\left(\frac{1}{5}\right)^{2^n}}\leqslant\frac{25}{12}y_1\left(\frac{1}{5}\right)^{2^n}\quad\text{при}\quad n\geqslant1,$$

где

$$y_1=\frac{1}{2}\left(y_0+\frac{x}{y_0}\right)\leqslant\frac{3}{2}y_0.$$

Отсюда

$$0\leqslant y_n-\sqrt{x}<\frac{25}{8}y_0\left(\frac{1}{5}\right)^{2^n}. \tag{8}$$

Из формулы (8) легко можно определить число итераций $n=n(x)$, достаточное для обеспечения заданной точности.

Приведем еще одну формулу для оценки погрешности значения $y_n$ $(n\geqslant2)$. Так как

$$y_{n-1}\geqslant\sqrt{x}\ \text{и}\ \frac{x}{y_{n-1}}\leqslant\sqrt{x},$$

то, учитывая формулу (6), имеем:

$$y_{n-1}-\sqrt{x}\leqslant y_{n-1}-\frac{x}{y_{n-1}}=\frac{y_{n-1}^2-x}{y_{n-1}}=2(y_{n-1}-y_n).$$

Следовательно,

$$0\leqslant y_n-\sqrt{x}\leqslant y_{n-1}-y_n. \tag{9}$$

Таким образом, если $0\leqslant y_{n-1}-y_n<\varepsilon$ $(n\geqslant2)$, то гарантировано, что $0\leqslant y_n-\sqrt{x}<\varepsilon$.

Укажем еще один способ вычисления квадратного корня, оказывающийся иногда полезным. Заменим функцию (1) эквивалентным соотношением

$$F(x,y)\equiv\frac{x}{y^2}-1=0.$$

Тогда

$$F'_y(x, y) = -\frac{2x}{y^3}.$$

Применяя формулу (4) § 10, получим:

$$y_{n+1} = y_n + \frac{\frac{x}{y_n^2} - 1}{\frac{2x}{y_n^3}}$$

или

$$y_{n+1} = \frac{y_n}{2}\left(3 - \frac{y_n^2}{x}\right) \quad (n = 0, 1, 2, \ldots). \tag{10}$$

Выяснение условий сходимости итеративного процесса (10) и оценку погрешности мы оставляем без рассмотрения.

## § 13. Вычисление обратной величины квадратного корня

Положим

$$y = \frac{1}{\sqrt{x}} \quad (x > 0).$$

Записав функцию в виде

$$y = \sqrt{\frac{1}{x}},$$

из формулы (10) предыдущего параграфа получим итеративный процесс «без деления»

$$y_{n+1} = \frac{y_n}{2}(3 - xy_n^2) \quad (n = 0, 1, 2, \ldots). \tag{1}$$

Если $x = 2^m x_1$, где $\frac{1}{2} \leqslant x_1 < 1$, то за $y_0$ выбирается значение

$$y_0 = 2^{-E\left(\frac{m}{2}\right)}.$$

Заметим, что, пользуясь очевидным равенством

$$\sqrt{x} = x\sqrt{\frac{1}{x}},$$

в силу формулы (1) извлечение квадратного корня из числа можно производить также «без деления».

## § 14. Вычисление кубического корня

Если

$$y = \sqrt[3]{x} \quad (x > 0), \tag{1}$$

то, положив

$$F(x, y) \equiv y^3 - x = 0,$$

будем иметь:

$$F'_y(x, y) = 3y^2.$$

Отсюда, применяя формулу (4) § 10, получаем:

$$y_{n+1} = y_n - \frac{y_n^3 - x}{3y_n^2} \tag{2}$$

или

$$y_{n+1} = \frac{1}{3}\left(2y_n + \frac{x}{y_n^2}\right). \tag{3}$$

Геометрически процесс (3) представляет собой метод Ньютона, примененный к кубической параболе

$$z = y^3 - x \quad (x = \text{const})$$

(рис. 8). Процесс (3) сходится при $y_0 > 0$.

Если в качестве начального приближения $y_0$ взять табличное значение $\sqrt[3]{x}$, имеющее относительную погрешность $|\delta|$, т. е. положить

$$y_0 = \sqrt[3]{x}(1+\delta),$$

то значение $y_1$, найденное из формулы (3), даст $\sqrt[3]{x}$ с относительной погрешностью $\delta^2$. Действительно, применяя формулу (3), имеем:

$$y_1 = \frac{1}{3}\left(2y_0 + \frac{x}{y_0^2}\right) = \frac{1}{3}\left[2\sqrt[3]{x}(1+\delta) + \sqrt[3]{x}(1+\delta)^{-2}\right] = \frac{1}{3}\sqrt[3]{x}(2+2\delta+1-2\delta+3\delta^2) = \sqrt[3]{x}(1+\delta^2).$$

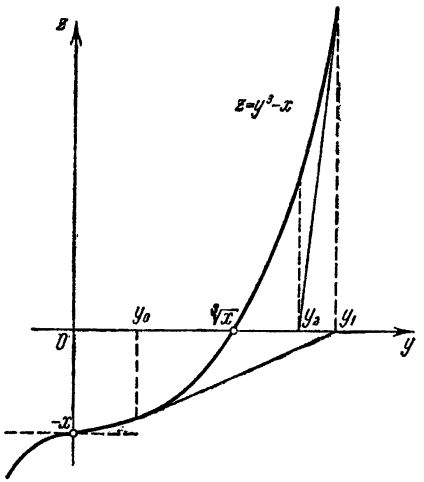


Рис. 8.

Отсюда, в частности, заключаем, что если $y_0$ имеет $p$ верных знаков в узком смысле, то $y_1$ будет иметь примерно $2p$ или $2p-1$ верных знаков в широком смысле (ср. § 12).

Пример. По трехзначным таблицам имеем:

$$\sqrt[3]{10} = 2{,}154,$$

где все знаки верные.

Применяя формулу (3), получаем:

$$\sqrt[3]{10}=\frac{1}{3}\left(2\cdot 2{,}154+\frac{10}{2{,}154^2}\right)=\frac{1}{3}(2\cdot 2{,}154+2{,}155304)=2{,}154435.$$

Для сравнения приводим значение из таблиц Барлоу

$$\sqrt[3]{10}=2{,}1544347\ldots$$

Если $x=2^m x_1$, где $m$ — целое число и $\frac{1}{2}\leqslant x_1<1$, то за начальное значение $y_0$ обычно выбирают

$$y_0=2^{E\left(\frac{m}{3}\right)}>0. \tag{4}$$

Так как

$$y_n-\sqrt[3]{x}=\frac{1}{3}\left(2y_{n-1}+\frac{x}{y_{n-1}^2}-3\sqrt[3]{x}\right)=$$

$$=\frac{1}{3y_{n-1}^2}(y_{n-1}-\sqrt[3]{x})^2(2y_{n-1}+\sqrt[3]{x})>0,$$

то

$$y_n\geqslant\sqrt[3]{x}\ \text{при}\ n\geqslant 1. \tag{5}$$

Кроме того, из формулы (2), заменяя $n+1$ на $n$, имеем:

$$y_{n-1}-y_n=\frac{y_{n-1}^3-x}{3y_{n-1}^2}; \tag{6}$$

поэтому

$$y_1\geqslant y_2\geqslant\ldots\geqslant y_{n-1}\geqslant y_n\geqslant\ldots\geqslant\sqrt[3]{x}. \tag{7}$$

Отсюда вытекает, что существует

$$\lim_{n\to\infty} y_n=y>0.$$

Переходя к пределу при $n\longrightarrow\infty$ в равенстве (3), будем иметь:

$$y=\frac{1}{3}\left(2y+\frac{x}{y^2}\right),$$

т. е. $y^3=x$ и, следовательно, $y=\sqrt[3]{x}$. Таким образом,

$$\lim_{n\to\infty} y_n=\sqrt[3]{x}.$$

Если начальное приближение $y_0$ выбирается на основании формулы (4), то можно доказать, что

$$0\leqslant y_n-\sqrt[3]{x}\leqslant\frac{3}{2}(y_{n-1}-y_n)$$

при $n\geqslant 2$.

## Литература к третьей главе

1. В. И. Смирнов, Курс высшей математики, т. I, изд. 17, Гостехиздат, М., 1957, гл. IV.
2. А. Марков, Исчисление конечных разностей, изд. 2, Матезис, 1911, гл. III.
3. Г. П. Толстов, Курс математического анализа, т. II, Гостехиздат, М., 1957, гл. XXIV.
4. А. Н. Хованский, Приложение цепных дробей и их обобщений к вопросам приближенного анализа, Гостехиздат, 1956, гл. II.
5. Б. М. Каган и Т. М. Тер-Микаэлян, Решение инженерных задач на автоматических цифровых вычислительных машинах, Госэнергоиздат, М.—Л., 1958, гл. III.
6. Г. М. Фихтенгольц, Курс дифференциального и интегрального исчисления, ОГИЗ, М.—Л., 1948, т. II, гл. XII.
7. Л. А. Люстерник, А. А. Абрамов, В. И. Шестаков, М. Р. Шура-Бура, Решение математических задач на автоматических цифровых машинах, Изд. АН СССР, 1952.

ГЛАВА IV

# ПРИБЛИЖЕННОЕ РЕШЕНИЕ АЛГЕБРАИЧЕСКИХ И ТРАНСЦЕНДЕНТНЫХ УРАВНЕНИЙ

## § 1. Отделение корней

Если уравнение алгебраическое или трансцендентное достаточно сложно, то его корни сравнительно редко удается найти точно. Кроме того, в некоторых случаях уравнение содержит коэффициенты, известные лишь приблизительно, и, следовательно, сама задача о точном определении корней уравнения теряет смысл. Поэтому важное значение приобретают способы приближенного нахождения корней уравнения и оценки степени их точности.

Пусть дано уравнение

$$f(x)=0, \tag{1}$$

где функция $f(x)$ определена и непрерывна в некотором конечном или бесконечном интервале $a<x<b$.

В дальнейшем в некоторых случаях нам понадобится существование и непрерывность первой производной $f'(x)$ или даже второй производной $f''(x)$, что будет оговорено в соответствующих местах.

Всякое значение $\xi$, обращающее функцию $f(x)$ в нуль, т. е. такое, что

$$f(\xi)=0,$$

называется *корнем уравнения* (1) или *нулем* функции $f(x)$.

Мы будем предполагать, что уравнение (1) имеет лишь *изолированные корни*, т. е. для каждого корня уравнения (1) существует окрестность, не содержащая других корней этого уравнения.

Приближенное нахождение изолированных действительных корней уравнения (1) обычно складывается из двух этапов:

1) отделение корней, т. е. установление возможно тесных промежутков $[\alpha, \beta]$, в которых содержится один и только один корень уравнения (1);

2) уточнение приближенных корней, т. е. доведение их до заданной степени точности.

Для отделения корней полезна известная теорема из математического анализа ([5], гл. IV).

Теорема 1. *Если непрерывная функция $f(x)$ принимает значения разных знаков на концах отрезка $[\alpha, \beta]$, т. е. $f(\alpha)f(\beta)<0$, то внутри этого отрезка содержится по меньшей мере один корень уравнения $f(x)=0$, т. е. найдется хотя бы одно число $\xi \in (\alpha, \beta)$*) такое, что $f(\xi)=0$* (рис. 9).

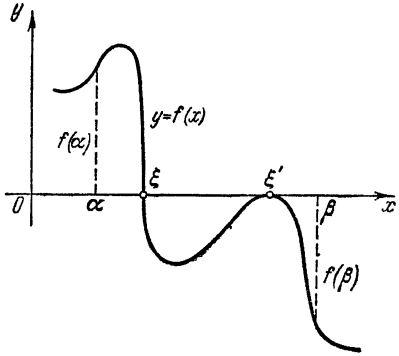

Рис. 9.

Корень $\xi$ заведомо будет единственным, если производная $f'(x)$ существует и сохраняет постоянный знак внутри интервала $(\alpha, \beta)$, т. е. если $f'(x)>0$ (или $f'(x)<0$) при $\alpha<x<\beta$ (рис. 10).

Процесс отделения корней начинается с установления знаков функции $f(x)$ в граничных точках $x=a$ и $x=b$ области ее существования.

Затем определяются знаки функции $f(x)$ в ряде промежуточных точек $x=\alpha_1, \alpha_2, \ldots$, выбор которых учитывает особенности функции $f(x)$. Если окажется, что $f(\alpha_k)f(\alpha_{k+1})<0$, то в силу теоремы 1 в интервале $(\alpha_k, \alpha_{k+1})$ имеется корень уравнения $f(x)=0$. Нужно тем или иным способом убедиться, является ли этот корень единственным. Для отделения корней практически часто бывает достаточно провести процесс половинного деления, приближенно деля данный интервал $(\alpha, \beta)$ на две, четыре, восемь и т. д. равных частей (до некоторого шага) и определяя знаки функции $f(x)$ в точках делений. Полезно помнить, что алгебраическое уравнение $n$-й степени

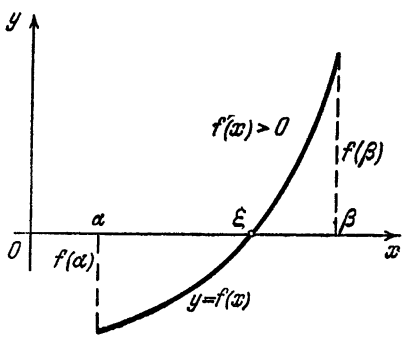

Рис. 10.

$$a_0x^n+a_1x^{n-1}+\ldots+a_n=0 \qquad (a_0\neq 0)$$

имеет не более $n$ действительных корней. Поэтому если для такого уравнения мы получили $n+1$ перемену знаков, то все корни его отделены.

Пример 1. Отделить корни уравнения

$$f(x)\equiv x^3-6x+2=0. \tag{2}$$

*) Запись $\xi \in (\alpha, \beta)$ обозначает, что точка $\xi$ принадлежит интервалу $(\alpha, \beta)$.

Решение. Составляем приблизительную схему:

| $x$ | $f(x)$ | $x$ | $f(x)$ |
|---|---|---|---|
| $-\infty$ | $-$ | 1 | $-$ |
| $-3$ | $-$ | 3 | $+$ |
| $-1$ | $+$ | $+\infty$ | $+$ |
| 0 | $+$ | | |

Следовательно, уравнение (2) имеет три действительных корня, лежащих в интервалах $(-3, -1)$, $(0, 1)$ и $(1, 3)$.

Если существует непрерывная производная $f'(x)$ и корни уравнения

$$f'(x) = 0$$

легко вычисляются, то процесс отделения корней уравнения (1) можно упорядочить. Для этого, очевидно, достаточно подсчитать лишь знаки функции $f(x)$ в точках нулей ее производной и в граничных точках $x = a$ и $x = b$.

Пример 2. Отделить корни уравнения

$$f(x) \equiv x^4 - 4x - 1 = 0. \tag{3}$$

Решение. Здесь $f'(x) = 4(x^3 - 1)$, поэтому $f'(x) = 0$ при $x = 1$.

Имеем $f(-\infty) > 0\,(+)$; $f(1) < 0\,(-)$; $f(+\infty) > 0\,(+)$. Следовательно, уравнение (3) имеет только два действительных корня, из которых один лежит в интервале $(-\infty, 1)$, а другой — в интервале $(1, +\infty)$.

Пример 3. Определить число действительных корней уравнения

$$f(x) \equiv x + e^x = 0. \tag{4}$$

Решение. Так как $f'(x) = 1 + e^x > 0$ и $f(-\infty) = -\infty$, $f(+\infty) = +\infty$, то уравнение (4) имеет только один действительный корень.

Дадим теперь оценку погрешности приближенного корня.

Теорема 2. *Пусть $\xi$ — точный, а $\bar{x}$ — приближенный корни уравнения $f(x) = 0$, находящиеся на одном и том же отрезке $[\alpha, \beta]$, причем $|f'(x)| \geqslant m_1 > 0$ при $\alpha \leqslant x \leqslant \beta$**).

*В таком случае справедлива оценка*

$$|\bar{x} - \xi| \leqslant \frac{|f(\bar{x})|}{m_1}. \tag{5}$$

---

*) В частности, за $m_1$ можно взять наименьшее значение $|f'(x)|$ при $\alpha \leqslant x \leqslant \beta$.

Доказательство. Применяя теорему Лагранжа, будем иметь:

$$f(\overline{x})-f(\xi)=(\overline{x}-\xi)f'(c),$$

где $c$ — промежуточное значение между $\overline{x}$ и $\xi$, т. е. $c\in(\alpha,\ \beta)$.

Отсюда, так как $f(\xi)=0$ и $|f'(c)|\geqslant m_1$, получим:

$$|f(\overline{x})-f(\xi)|=|f(\overline{x})|\geqslant m_1|\overline{x}-\xi|.$$

Следовательно,

$$|\overline{x}-\xi|\leqslant\frac{|f(\overline{x})|}{m_1}.$$

Замечание. Формула (5) может дать грубые результаты, и ее не всегда удобно применять. Поэтому на практике тем или иным способом сужают общий интервал $(\alpha,\ \beta)$, содержащий корень $\xi$ и его приближенное значение $\overline{x}$, и полагают $|\overline{x}-\xi|\leqslant\beta-\alpha$.

Пример 4. Приближенным корнем уравнения $f(x)\equiv x^4-x-1=0$ является $\overline{x}=1{,}22$. Оценить абсолютную погрешность этого корня.

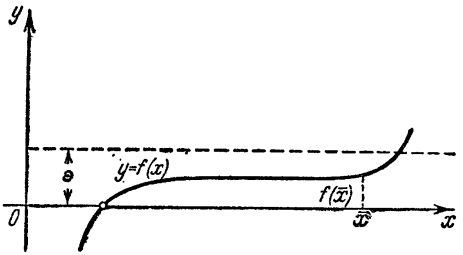


Рис. 11.

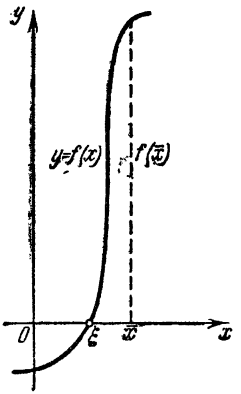


Рис. 12.

Решение. Имеем $f(\overline{x})=2{,}2153-1{,}22-1=-0{,}0047$.

Так как при $\overline{\overline{x}}=1{,}23$ получаем

$$f(\overline{\overline{x}})=2{,}2888-1{,}23-1=+0{,}0588,$$

то точный корень $\xi$ содержится в интервале $(1{,}22;\ 1{,}23)$. Производная $f'(x)=3x^3-1$ монотонно возрастает. Поэтому ее наименьшим значением в данном интервале является:

$$m_1=3\cdot1{,}22^3-1=3\cdot1{,}816-1=4{,}448.$$

Отсюда по формуле (5) получим:

$$|\overline{x}-\xi|\leqslant\frac{0{,}0047}{4{,}448}\approx0{,}001.$$

Замечание. Иногда на практике точность приближенного корня $\bar{x}$ оценивают по тому, насколько хорошо он удовлетворяет данному уравнению $f(x)=0$, т. е. если число $|f(\bar{x})|$ малое, то считают, что $\bar{x}$ является хорошим приближением точного корня $\xi$; если же $|f(\bar{x})|$ велико, то $\bar{x}$ полагают грубым значением точного корня $\xi$. Такой подход, как показывают рис. 11 и 12, является неправильным. Не следует также забывать, что если уравнение $f(x)=0$ умножить на произвольное число $N\neq 0$, то получается равносильное уравнение $Nf(x)=0$, причем число $|Nf(\bar{x})|$ можно сделать сколь угодно большим или сколь угодно малым за счет выбора множителя $N$.

## § 2. Графическое решение уравнений

Действительные корни уравнения

$$f(x)=0 \tag{1}$$

приближенно можно определить как абсциссы точек пересечения графика функции $y=f(x)$ с осью $Ox$ (рис. 9). Если уравнение (1) не имеет близких между собой корней, то этим способом его корни легко отделяются. На практике часто бывает выгодно уравнение (1) заменить равносильным ему уравнением*)

$$\varphi(x)=\psi(x), \tag{2}$$

где функции $\varphi(x)$ и $\psi(x)$ — более простые, чем функция $f(x)$. Тогда, построив графики функций $y=\varphi(x)$ и $y=\psi(x)$, искомые корни получим как абсциссы точек пересечения этих графиков.

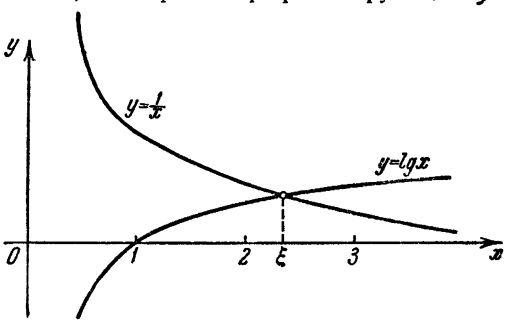


Рис. 13.

Пример 1. Графически решить уравнение

$$x\lg x=1. \tag{3}$$

Решение. Запишем уравнение (3) в виде равенства

$$\lg x=\frac{1}{x}.$$

Отсюда ясно, что корни уравнения (3) могут быть найдены как абсциссы точек пересечения логарифмической кривой $y=\lg x$ и гиперболы $y=\frac{1}{x}$. Построив эти кривые (рис. 13) на координатной

*) Два уравнения называются равносильными, если они имеют одинаковые корни.

бумаге, приближенно найдем единственный корень $\xi \approx 2{,}5$ уравнения (3).

Нахождение корней уравнения (2) упрощается, если одна из функций $\varphi(x)$ или $\psi(x)$ линейная, т. е., например, $\varphi(x) = ax + b$. В этом случае корни уравнения (2) находятся как абсциссы точек пересечения кривой $y = \psi(x)$ и прямой $y = ax + b$. Особенно выгодным оказывается этот прием при решении ряда однотипных уравнений, отличающихся только коэффициентами $a$ и $b$ линейной функции. Здесь графическое построение сводится к нахождению точек пересечения фиксированного графика $y = \psi(x)$ различными прямыми. К указанному типу, очевидно, относятся трехчленные уравнения

$$x^n + ax + b = 0.$$

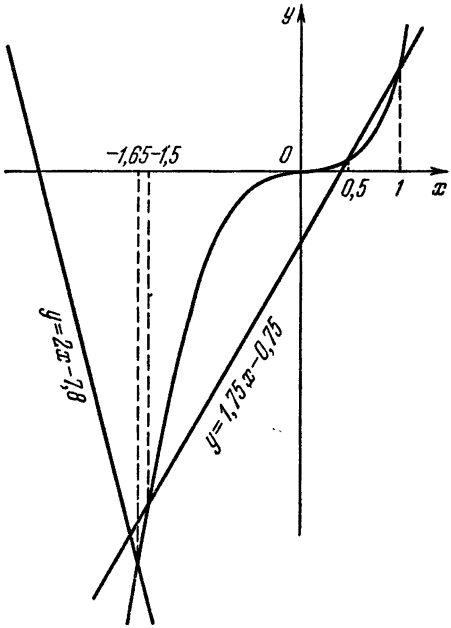


Рис. 14.

Пример 2. Решить кубические уравнения

$$x^3 - 1{,}75x + 0{,}75 = 0$$

и

$$x^3 + 2x + 7{,}8 = 0.$$

Решение. Построим кубическую параболу $y = x^3$. Искомые корни находятся как абсциссы точек пересечения этой параболы прямыми (рис. 14) $y = 1{,}75x - 0{,}75$ и $y = -2x - 7{,}8$. По чертежу ясно, что первое уравнение имеет три действительных корня: $x_1 = -1{,}5$; $x_2 = 0{,}5$; $x_3 = 1$, а второе уравнение — лишь один действительный корень $x_1 = -1{,}65$.

Отметим, что хотя графические методы решения уравнений весьма удобны и сравнительно просты, но они, как правило, применимы лишь для грубого определения корней. Особенно неблагоприятным в смысле потери точности является случай, когда линии пересекаются под очень острым углом и практически сливаются по некоторой дуге.

Разновидностью графических методов решения уравнений являются *номографические методы*, для ознакомления с которыми следует обратиться к специальным руководствам.

## § 3. Метод половинного деления

Пусть дано уравнение

$$f(x)=0, \tag{1}$$

где функция $f(x)$ непрерывна на $[a, b]$ и $f(a)f(b)<0$.

Для нахождения корня уравнения (1), принадлежащего отрезку $[a, b]$, делим этот отрезок пополам. Если $f\left(\frac{a+b}{2}\right)=0$, то $\xi=\frac{a+b}{2}$ является корнем уравнения. Если $f\left(\frac{a+b}{2}\right)\neq 0$, то выбираем ту из половин $\left[a, \frac{a+b}{2}\right]$ или $\left[\frac{a+b}{2}, b\right]$, на концах которой функция $f(x)$ имеет противоположные знаки. Новый суженный отрезок $[a_1, b_1]$ снова делим пополам и проводим то же рассмотрение и т. д. В результате получаем на каком-то этапе или точный корень уравнения (1), или же бесконечную последовательность вложенных друг в друга отрезков $[a_1, b_1]$, $[a_2, b_2]$, ..., $[a_n, b_n]$, ... таких, что

$$f(a_n)f(b_n)<0 \quad (n=1, 2, \ldots) \tag{2}$$

и

$$b_n-a_n=\frac{1}{2^n}(b-a). \tag{3}$$

Так как левые концы $a_1, a_2, \ldots, a_n, \ldots$ образуют монотонную неубывающую ограниченную последовательность, а правые концы $b_1, b_2, \ldots, b_n, \ldots$ — монотонную невозрастающую ограниченную последовательность, то в силу равенства (3) существует общий предел

$$\xi=\lim_{n\to\infty} a_n=\lim_{n\to\infty} b_n.$$

Переходя к пределу при $n\to\infty$ в неравенстве (2), в силу непрерывности функции $f(x)$ получим $[f(\xi)]^2\leqslant 0$. Отсюда $f(\xi)=0$, т. е. $\xi$ является корнем уравнения (1), причем, очевидно,

$$0\leqslant \xi-a_n\leqslant \frac{1}{2^n}(b-a). \tag{4}$$

Если корни уравнения (1) не отделены на отрезке $[a, b]$, то таким способом можно найти один из корней уравнения (1).

Метод половинного деления практически удобно применять для грубого нахождения корня данного уравнения, так как при увеличении точности значительно возрастает объем вычислительной работы.

Заметим, что метод половинного деления легко реализуется на электронных счетных машинах. Программа вычисления составляется так, чтобы машина находила значение правой части уравнения (1) в середине каждого из отрезков $[a_n, b_n]$ $(n=1, 2, \ldots)$ и выбирала соответствующую половину его.

Пример. Методом половинного деления уточнить корень уравнения

$$f(x) \equiv x^4 + 2x^3 - x - 1 = 0,$$

лежащий на отрезке [0, 1].

Решение. Последовательно имеем:

$$\begin{aligned}
f(0) &= -1; \; f(1) = 1;\\
f(0{,}5) &= 0{,}06 + 0{,}25 - 0{,}5 - 1 = -1{,}19;\\
f(0{,}75) &= 0{,}32 + 0{,}84 - 0{,}75 - 1 = -0{,}59;\\
f(0{,}875) &= 0{,}59 + 1{,}34 - 0{,}88 - 1 = +0{,}05;\\
f(0{,}8125) &= 0{,}436 + 1{,}072 - 0{,}812 - 1 = -0{,}304;\\
f(0{,}8438) &= 0{,}507 + 1{,}202 - 0{,}844 - 1 = -0{,}135;\\
f(0{,}8594) &= 0{,}546 + 1{,}270 - 0{,}859 - 1 = -0{,}043 \text{ и т. д.}
\end{aligned}$$

Можно принять

$$\xi = \frac{1}{2}(0{,}859 + 0{,}875) = 0{,}867.$$

## § 4. Способ пропорциональных частей (метод хорд)

Укажем (в предположениях § 3) более быстрый способ нахождения корня $\xi$ уравнения $f(x) = 0$, лежащего на заданном отрезке $[a, b]$ таком, что $f(a)f(b) < 0$.

Пусть для определенности $f(a) < 0$ и $f(b) > 0$. Тогда, вместо того чтобы делить отрезок $[a, b]$ пополам, более естественно разделить его в отношении $-f(a):f(b)$. Это дает нам приближенное значение корня

$$x_1 = a + h_1, \qquad (1)$$

где

$$h_1 = \frac{-f(a)}{-f(a) + f(b)}(b - a) = -\frac{f(a)}{f(b) - f(a)}(b - a). \qquad (2)$$

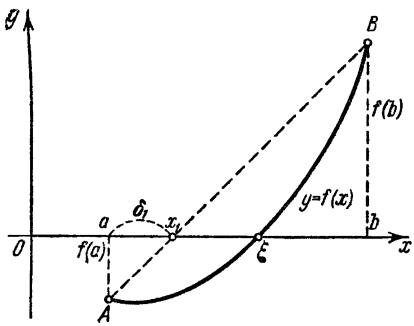

Рис. 15.

Далее, применяя этот прием к тому из отрезков $[a, x_1]$ или $[x_1, b]$, на концах которого функция $f(x)$ имеет противоположные знаки, получим второе приближение корня $x_2$ и т. д.

Геометрически способ пропорциональных частей эквивалентен замене кривой $y = f(x)$ хордой, проходящей через точки $A[a, f(a)]$ и $B[b, f(b)]$ (рис. 15). В самом деле, уравнение хорды $AB$ есть

$$\frac{x - a}{b - a} = \frac{y - f(a)}{f(b) - f(a)}.$$

Отсюда, полагая $x = x_1$ и $y = 0$, получим:

$$x_1 = a - \frac{f(a)}{f(b) - f(a)}(b - a). \qquad (1')$$

Формула (1′) полностью эквивалентна формулам (1) и (2).

Для доказательства сходимости процесса предположим, что корень отделен и вторая производная $f''(x)$ сохраняет постоянный знак на отрезке $[a, b]$.

Пусть для определенности $f''(x) > 0$ при $a \leqslant x \leqslant b$ (случай $f''(x) < 0$ сводится к нашему, если записать уравнение в виде

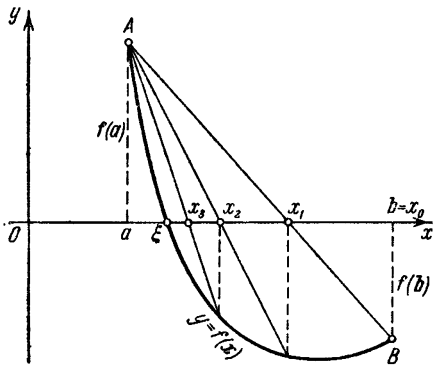


Рис. 16.

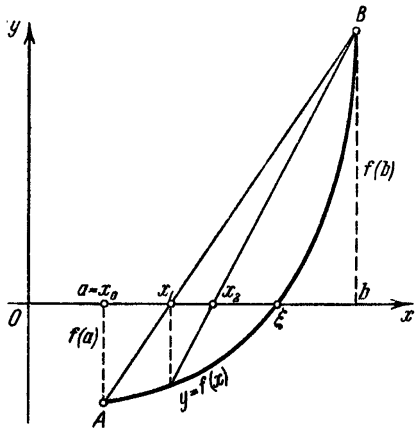


Рис. 17.

$-f(x) = 0$). Тогда кривая $y = f(x)$ будет выпукла вниз и, следовательно, расположена ниже своей хорды $AB$. Возможны два случая: 1) $f(a) > 0$ (рис. 16) и 2) $f(a) < 0$ (рис. 17).

В первом случае конец $a$ неподвижен и последовательные приближения: $x_0 = b$;

$$x_{n+1} = x_n - \frac{f(x_n)}{f(x_n) - f(a)}(x_n - a) \quad (n = 0, 1, 2, \ldots) \qquad (3)$$

образуют ограниченную монотонно убывающую последовательность, причем

$$a < \xi < \ldots < x_{n+1} < x_n < \ldots < x_1 < x_0.$$

Во втором случае неподвижен конец $b$, а последовательные приближения: $x_0 = a$;

$$x_{n+1} = x_n - \frac{f(x_n)}{f(b) - f(x_n)}(b - x_n) \qquad (4)$$

образуют ограниченную монотонно возрастающую последовательность, причем

$$x_0 < x_1 < x_2 < \ldots < x_n < x_{n+1} < \ldots < \xi < b.$$

Обобщая эти результаты, заключаем: 1) неподвижен тот конец, для которого знак функции $f(x)$ совпадает со знаком ее второй производной $f''(x)$; 2) последовательные приближения $x_n$ лежат по ту сторону корня $\xi$, где функция $f(x)$ имеет знак, противоположный знаку ее второй производной $f''(x)$. В обоих случаях каждое следующее приближение $x_{n+1}$ ближе к корню $\xi$, чем предшествующее $x_n$. Пусть

$$\overline{\xi}=\lim_{n\to\infty} x_n \qquad (a<\overline{\xi}<b)$$

(предел существует, так как последовательность $\{x_n\}$ ограничена и монотонна). Переходя к пределу в равенстве (3), для первого случая будем иметь:

$$\overline{\xi}=\overline{\xi}-\frac{f(\overline{\xi})}{f(\overline{\xi})-f(a)}(\overline{\xi}-a);$$

отсюда $f(\overline{\xi})=0$. Так как по предположению уравнение $f(x)=0$ имеет единственный корень $\xi$ на интервале $(a, b)$, то, следовательно, $\overline{\xi}=\xi$, что и требовалось доказать.

Совершенно так же переходом к пределу в равенстве (4) доказывается, что $\overline{\xi}=\xi$ для второго случая.

Для оценки точности приближения можно воспользоваться формулой (5) § 1

$$|x_n-\xi|\leqslant\frac{|f(x_n)|}{m_1},$$

где $|f'(x)|\geqslant m_1$ при $a\leqslant x\leqslant b$.

Приведем еще формулу, позволяющую оценивать абсолютную погрешность приближенного значения $x_n$, если известны два последовательных приближения $x_{n-1}$ и $x_n$.

Будем предполагать, что производная $f'(x)$ непрерывна на отрезке $[a, b]$, содержащем все приближения, и сохраняет постоянный знак, причем

$$0<m_1\leqslant|f'(x)|\leqslant M_1<+\infty. \tag{5}$$

Примем для определенности, что последовательные приближения $x_n$ точного корня $\xi$ вырабатываются по формуле (3) (рассмотрение формулы (4) аналогично)

$$x_n=x_{n-1}-\frac{f(x_{n-1})}{f(x_{n-1})-f(a)}(x_{n-1}-a)$$

($n=1, 2, \ldots$), где конец $a$ является неподвижным. Отсюда, учитывая, что $f(\xi)=0$, будем иметь:

$$f(\xi)-f(x_{n-1})=\frac{f(x_{n-1})-f(a)}{x_{n-1}-a}(x_n-x_{n-1}).$$

Применяя теорему Лагранжа о конечном приращении функции, получим:

$$(\xi - x_{n-1}) f'(\xi_{n-1}) = (x_n - x_{n-1}) f'(\bar{x}_{n-1}),$$

где $\xi_{n-1} \in (x_{n-1}, \xi)$ и $\bar{x}_{n-1} \in (a, x_{n-1})$. Следовательно,

$$|\xi - x_n| = \frac{|f'(\bar{x}_{n-1}) - f'(\xi_{n-1})|}{|f'(\xi_{n-1})|} |x_n - x_{n-1}|. \tag{6}$$

Так как $f'(x)$ сохраняет постоянный знак на отрезке $[a, b]$, причем $\bar{x}_{n-1} \in [a, b]$ и $\xi_{n-1} \in [a, b]$, то, очевидно, имеем:

$$|f'(\bar{x}_{n-1}) - f'(\xi_{n-1})| \leqslant M_1 - m_1.$$

Поэтому из формулы (6) выводим:

$$|\xi - x_n| \leqslant \frac{M_1 - m_1}{m_1} |x_n - x_{n-1}|, \tag{7}$$

где за $m_1$ и $M_1$ могут быть взяты соответственно наименьшее и наибольшее значения модуля производной $f'(x)$ на отрезке $[a, b]$. Если отрезок $[a, b]$ столь узок, что имеет место неравенство

$$M_1 \leqslant 2m_1,$$

то из формулы (7) получаем:

$$|\xi - x_n| \leqslant |x_n - x_{n-1}|.$$

Таким образом, в этом случае, как только будет обнаружено, что

$$|x_n - x_{n-1}| < \varepsilon,$$

где $\varepsilon$ — заданная предельная абсолютная погрешность, то гарантировано, что

$$|\xi - x_n| < \varepsilon.$$

Пример. Найти положительный корень уравнения

$$f(x) \equiv x^3 - 0{,}2x^2 - 0{,}2x - 1{,}2 = 0$$

с точностью до 0,002.

Решение. Прежде всего отделяем корень. Так как

$$f(1) = -0{,}6 < 0 \text{ и } f(2) = 5{,}6 > 0,$$

то искомый корень $\xi$ лежит в интервале (1, 2). Полученный интервал велик, поэтому разделим его пополам. Так как

$$f(1{,}5) = 1{,}425, \text{ то } 1 < \xi < 1{,}5.$$

Последовательно применяя формулы (1) и (2), будем иметь:

$$x_1 = 1 + \frac{0,6}{1,425+0,6}(1,5-1) = 1+0,15 = 1,15;$$

$$f(x_1) = -0,173;$$

$$x_2 = 1,15 + \frac{0,173}{1,425+0,073}(1,5-1,15) = 1,15+0,040 = 1,190;$$

$$f(x_2) = -0,036;$$

$$x_3 = 1,190 + \frac{0,036}{1,425+0,036}(1,5-1,190) = 1,190+0,008 = 1,198;$$

$$f(x_3) = -0,0072.$$

Так как $f'(x) = 3x^2 - 0,4x - 0,2$ и при $x_3 < x < 1,5$ имеем

$$f'(x) \geqslant 3\cdot 1,198^2 - 0,4\cdot 1,5 - 0,2 = 3\cdot 1,43 - 0,8 = 3,49,$$

то можно принять:

$$0 < \xi - x_3 < \frac{0,0072}{3,49} \approx 0,002.$$

Таким образом, $\xi = 1,198 + 0,002\theta$, где $0 < \theta \leqslant 1$.
Заметим, что точный корень уравнения (5) есть $\xi = 1,2$.

## § 5. Метод Ньютона (метод касательных)

Пусть корень $\xi$ уравнения

$$f(x) = 0 \tag{1}$$

отделен на отрезке $[a, b]$, причем $f'(x)$ и $f''(x)$ непрерывны и сохраняют определенные знаки при $a \leqslant x \leqslant b$. Найдя какое-нибудь $n$-е приближенное значение корня $x_n \approx \xi$ $(a \leqslant x_n \leqslant b)$, мы можем уточнить его по *методу Ньютона* следующим образом. Положим

$$\xi = x_n + h_n, \tag{2}$$

где $h_n$ считаем малой величиной. Отсюда, применяя формулу Тейлора, получим:

$$0 = f(x_n + h_n) \approx f(x_n) + h_n f'(x_n).$$

Следовательно,

$$h_n = -\frac{f(x_n)}{f'(x_n)}.$$

Внеся эту поправку в формулу (2), найдем следующее (по порядку) приближение корня

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad (n = 0, 1, 2, \ldots). \tag{3}$$

Геометрически метод Ньютона эквивалентен замене небольшой дуги кривой $y=f(x)$ касательной, проведенной в некоторой точке кривой. В самом деле, положим для определенности, что $f''(x)>0$ при $a\leqslant x\leqslant b$ и $f(b)>0$ (рис. 18).

Выберем, например, $x_0=b$, для которого $f(x_0)f''(x_0)>0$. Проведем касательную к кривой $y=f(x)$ в точке $B_0[x_0, f(x_0)]$.

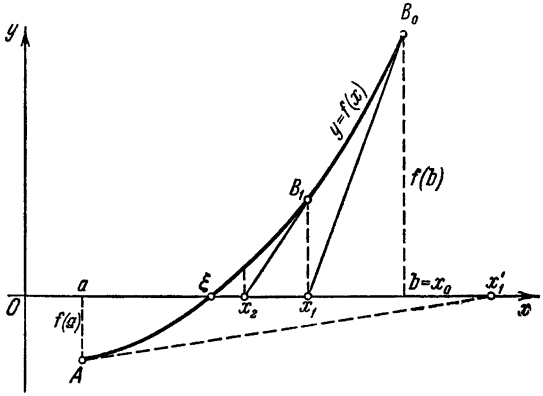


Рис. 18.

В качестве первого приближения $x_1$ корня $\xi$ возьмем абсциссу точки пересечения этой касательной с осью $Ox$. Через точку $B_1[x_1, f(x_1)]$ снова проведем касательную, абсцисса точки пересечения которой даст нам второе приближение $x_2$ корня $\xi$ и т. д. (рис. 18). Очевидно, что уравнение касательной в точке $B_n[x_n, f(x_n)]$ $(n=0, 1, 2, \ldots)$ есть

$$y-f(x_n)=f'(x_n)(x-x_n).$$

Полагая $y=0$, $x=x_{n+1}$, получим формулу (3)

$$x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)}.$$

Заметим, что если в нашем случае положить $x_0=a$ и, следовательно, $f(x_0)f''(x_0)<0$, то, проведя касательную к кривой $y=f(x)$ в точке $A[a, f(a)]$, мы получили бы точку $x_1$ (рис. 18), лежащую вне отрезка $[a, b]$, т. е. при этом выборе начального значения метод Ньютона оказывается непрактичным. Таким образом, в данном случае «хорошим» начальным приближением $x_0$ является то, для которого выполнено неравенство

$$f(x_0)f''(x_0)>0. \tag{4}$$

Докажем, что это правило является общим.

Теорема. *Если $f(a)f(b)<0$, причем $f'(x)$ и $f''(x)$ отличны от нуля и сохраняют определенные знаки при $a\leqslant x\leqslant b$, то,*

*исходя из начального приближения* $x_0 \in [a, b]$, *удовлетворяющего неравенству* (4), *можно вычислить методом Ньютона* (*формула* (3)) *единственный корень* $\xi$ *уравнения* (1) *с любой степенью точности.*

Доказательство. Пусть, например, $f(a) < 0$, $f(b) > 0$, $f'(x) > 0$, $f''(x) > 0$ при $a \leqslant x \leqslant b$ (остальные случаи рассматриваются аналогично). Согласно неравенству (4) имеем $f(x_0) > 0$ (например, можно принять $x_0 = b$).

Методом математической индукции докажем, что все приближения $x_n > \xi$ $(n = 0, 1, 2, \ldots)$ и, следовательно, $f(x_n) > 0$. В самом деле, прежде всего, $x_0 > \xi$.

Пусть теперь $x_n > \xi$. Положим

$$\xi = x_n + (\xi - x_n).$$

Применяя формулу Тейлора, получим:

$$0 = f(\xi) = f(x_n) + f'(x_n)(\xi - x_n) + \frac{1}{2} f''(c_n)(\xi - x_n)^2, \tag{5}$$

где $\xi < c_n < x_n$.

Так как $f''(x) > 0$, то имеем:

$$f(x_n) + f'(x_n)(\xi - x_n) < 0$$

и, следовательно,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} > \xi,$$

что и требовалось доказать.

Из формулы (3), учитывая знаки $f(x_n)$ и $f'(x_n)$, имеем $x_{n+1} < x_n$ $(n = 0, 1, \ldots)$, т. е. последовательные приближения $x_0, x_1, \ldots, x_n, \ldots$ образуют ограниченную монотонно убывающую последовательность. Следовательно, существует $\bar{\xi} = \lim\limits_{n \to \infty} x_n$.

Переходя к пределу в равенстве (3), будем иметь:

$$\bar{\xi} = \bar{\xi} - \frac{f(\bar{\xi})}{f'(\bar{\xi})},$$

т. е. $f(\bar{\xi}) = 0$. Отсюда $\bar{\xi} = \xi$, что и требовалось доказать.

Поэтому, применяя метод Ньютона, следует руководствоваться следующим правилом: *в качестве исходной точки* $x_0$ *выбирается тот конец интервала* $(a, b)$, *которому отвечает ордината того же знака, что и знак* $f''(x)$.

Замечание 1. Если: 1) функция $f(x)$ определена и непрерывна при $-\infty < x < +\infty$; 2) $f(a)f(b) < 0$; 3) $f'(x) \neq 0$ при $a \leqslant x \leqslant b$; 4) $f''(x)$ существует всюду и сохраняет постоянный знак, то при применении метода Ньютона для нахождения корня уравнения $f(x) = 0$, лежащего в интервале $(a, b)$, за начальное

приближение $x_0$ можно принять любое значение $c \in [a, b]$. В частности, можно взять $x_0 = a$ или $x_0 = b$.

Действительно, пусть, например, $f'(x) > 0$ при $a \leqslant x \leqslant b$, $f''(x) > 0$ и $x_0 = c$, где $a \leqslant c \leqslant b$.

Если $f(c) = 0$, то корень $\xi = c$ и задача, таким образом, решена.

Если $f(c) > 0$, то справедливо приведенное выше рассуждение и процесс Ньютона с начальным значением $c$ сходится к корню $\xi \in (a, b)$.

Наконец, если $f(c) < 0$, то находим значение

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = c - \frac{f(c)}{f'(c)} > c.$$

Применяя формулу Тейлора, будем иметь:

$$f(x_1) = f(c) - \frac{f(c)}{f'(c)} f'(c) + \frac{1}{2}\left[\frac{f(c)}{f'(c)}\right]^2 f''(\bar{c}) = \frac{1}{2}\left[\frac{f(c)}{f'(c)}\right]^2 f''(\bar{c}) > 0,$$

где $\bar{c}$ — некоторое промежуточное значение между $c$ и $x_1$.

Таким образом,

$$f(x_1) f''(x_1) > 0.$$

Кроме того, из условия $f''(x) > 0$ вытекает, что $f'(x)$ — возрастающая функция и, значит, $f'(x) > f'(a) > 0$ при $x > a$. Следовательно, $x_1$ можно принять за начальное значение для процесса Ньютона, сходящегося к некоторому корню $\bar{\xi}$ функции $f(x)$ такому, что $\bar{\xi} > c \geqslant a$. Так как в силу положительности производной $f'(x)$ при $x > a$ функция $f(x)$ имеет единственный корень на интервале $(a, +\infty)$, то

$$\bar{\xi} = \xi \in (a, b).$$

Аналогичное рассмотрение можно провести для других комбинаций знаков производных $f'(x)$ и $f''(x)$.

Замечание 2. Из формулы (3) видно, что чем больше численное значение производной $f'(x)$ в окрестности данного корня, тем меньше поправка, которую нужно прибавить к $n$-му приближению, чтобы получить $(n+1)$-е приближение. Поэтому метод Ньютона особенно удобно применять тогда, когда в окрестности данного корня график функции имеет большую крутизну. Но если численное значение производной $f'(x)$ близ корня мало, то поправки будут велики, и вычисление корня по этому методу может оказаться очень долгим, а иногда и вовсе невозможным. Следовательно, если кривая $y = f(x)$ вблизи точки пересечения с осью $Ox$ почти горизонтальна, то применять метод Ньютона для решения уравнения $f(x) = 0$ не рекомендуется.

Для оценки погрешности $n$-го приближения $x_n$ можно воспользоваться общей формулой (5) § 1

$$|\xi - x_n| \leqslant \frac{|f(x_n)|}{m_1}, \tag{6}$$

где $m_1$ — наименьшее значение $|f'(x)|$ на отрезке $[a, b]$.

Выведем еще одну формулу для оценки точности приближения $x_n$. Применяя формулу Тейлора, имеем:

$$f(x_n) = f[x_{n-1} + (x_n - x_{n-1})] =$$
$$= f(x_{n-1}) + f'(x_{n-1})(x_n - x_{n-1}) + \frac{1}{2} f''(\xi_{n-1})(x_n - x_{n-1})^2, \tag{7}$$

где $\xi_{n-1} \in (x_{n-1}, x_n)$. Так как в силу определения приближения $x_n$ имеем

$$f(x_{n-1}) + f'(x_{n-1})(x_n - x_{n-1}) = 0,$$

то из (7) находим:

$$|f(x_n)| \leqslant \frac{1}{2} M_2 (x_n - x_{n-1})^2,$$

где $M_2$ — наибольшее значение $|f''(x)|$ на отрезке $[a, b]$. Следовательно, на основании формулы (6) окончательно получаем:

$$|\xi - x_n| \leqslant \frac{M_2}{2m_1}(x_n - x_{n-1})^2. \tag{8}$$

Если процесс Ньютона сходится, то $x_n - x_{n-1} \to 0$ при $n \to \infty$. Поэтому при $n \geqslant N$ имеем:

$$|\xi - x_n| \leqslant |x_n - x_{n-1}|,$$

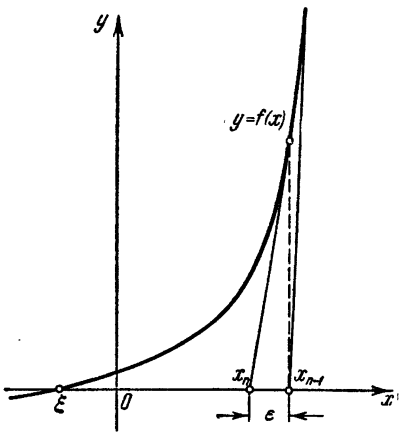


Рис. 19.

т. е. «*установившиеся*» начальные десятичные знаки приближений $x_{n-1}$ и $x_n$, начиная с некоторого приближения, являются верными.

Заметим, что в общем случае совпадение с точностью до $\varepsilon$ двух последовательных приближений $x_{n-1}$ и $x_n$ вовсе не гарантирует, что с той же точностью совпадает значение $x_n$ и точный корень $\xi$ (рис. 19).

Установим также формулу, связывающую абсолютные погрешности двух последовательных приближений $x_n$ и $x_{n+1}$. Из формулы (5) получаем:

$$\xi = x_n - \frac{f(x_n)}{f'(x_n)} - \frac{1}{2} \cdot \frac{f''(c_n)}{f'(x_n)}(\xi - x_n)^2,$$

где $c_n \in (x_n, \xi)$. Отсюда, учитывая формулу (3), будем иметь:

$$\xi - x_{n+1} = -\frac{1}{2} \cdot \frac{f''(c_n)}{f'(x_n)}(\xi - x_n)^2$$

и, следовательно,

$$|\xi - x_{n+1}| \leqslant \frac{M_2}{2m_1}(\xi - x_n)^2. \tag{9}$$

Формула (9) обеспечивает быструю сходимость процесса Ньютона, если начальное приближение $x_0$ таково, что

$$\frac{M_2}{2m_1}|\xi - x_0| \leqslant q < 1.$$

В частности, если

$$\mu = \frac{M_2}{2m_1} \leqslant 1 \text{ и } |\xi - x_n| < 10^{-m},$$

то из формулы (9) получаем:

$$|\xi - x_{n+1}| < 10^{-2m},$$

т. е. в этом случае, если приближение $x_n$ имело $m$ верных десятичных знаков после запятой, то следующее приближение $x_{n+1}$ будет иметь по меньшей мере $2m$ верных знаков; иными словами, если $\mu \leqslant 1$, то с помощью метода Ньютона число верных знаков после запятой искомого корня $\xi$ удваивается на каждом шаге.

Пример 1. Вычислить методом Ньютона отрицательный корень уравнения $f(x) \equiv x^4 - 3x^2 + 75x - 10\,000 = 0$ с пятью верными знаками.

Решение. Полагая в левой части уравнения $x = 0, -10, -100, \ldots$, получим $f(0) = -10\,000$, $f(-10) = -1050$, $f(-100) \approx +10^8$.

Следовательно, искомый корень $\xi$ находится в интервале $-100 < \xi < -10$. Сузим найденный интервал. Так как $f(-11) = 3453$, то $-11 < \xi < -10$. В этом последнем интервале $f'(x) < 0$ и $f''(x) > 0$. Так как $f(-11) > 0$ и $f''(-11) > 0$, то можем принять за начальное приближение $x_0 = -11$. Последовательные приближения $x_n$ $(n = 1, 2, \ldots)$ вычисляем по следующей схеме:

| $n$ | $x_n$ | $f(x_n)$ | $f'(x_n)$ | $h_n = -\frac{f(x_n)}{f'(x_n)}$ |
|---|---|---|---|---|
| 0 | $-11$ | 3453 | $-5183$ | 0,7 |
| 1 | $-10,3$ | 134,3 | $-4234$ | 0,03 |
| 2 | $-10,27$ | 37,8 | $-4196$ | 0,009 |
| 3 | $-10,261$ | 0,2 | — | — |

Останавливаясь на $n = 3$, проверяем знак значения $f(x_n + 0,001) = f(-10,260)$. Так как $f(-10,260) < 0$, то $-10,261 < \xi < -10,260$, и любое из этих чисел дает искомое приближение.

Пример 2. Найти по методу Ньютона наименьший положительный корень уравнения $\operatorname{tg} x = x$ с точностью до 0,0001.

Решение. Построив графики кривых $y = \operatorname{tg} x$ и $y = x$ (рис. 20), заключаем, что искомый корень $\xi$ находится в интервале $\pi < \xi < \frac{3\pi}{2}$. Переписав уравнение в виде

$$f(x) \equiv \sin x - x \cos x = 0,$$

будем иметь:

$$f'(x) = x \sin x;$$

$$f''(x) = \sin x + x \cos x.$$

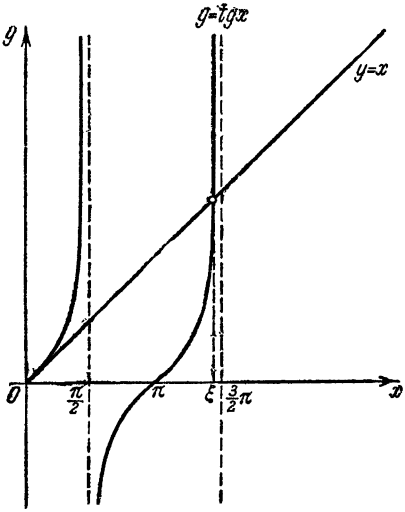


Рис. 20.

Отсюда $f'(x) < 0$ и $f''(x) < 0$ при $\pi < x < \frac{3\pi}{2}$. Так как $f\left(\frac{3\pi}{2}\right) = -1$, то за начальное приближение можно принять $x_0 = \frac{3\pi}{2}$. Вычисления производим по следующей схеме:

| $n$ | $x_n$ | $f(x_n)$ | $f'(x_n)$ | $h_n = -\frac{f(x_n)}{f'(x_n)}$ |
|---|---|---|---|---|
| 0 | $\frac{3\pi}{2} = 4{,}71239$ (270°) | $-1$ | $-4{,}712$ | $-0{,}212$ ($\approx -12°10'$) |
| 1 | 4,50004 (257°50′) | $-0{,}0291$ | $-4{,}399$ | $-0{,}0066$ ($\approx -22'44''$) |
| 2 | 4,49343 (257°27′16″) | $-0{,}00003$ | — | — |

Для оценки погрешности приближенного значения $x_n$ заметим, что последовательные приближения $x_n$ ($n = 0, 1, 2, \ldots$) в силу отрицательности второй производной $f''(x)$ монотонно убывают, причем $f(x_n) < 0$. Поэтому можно принять $\bar{x}_n < \xi < x_n$, где $\bar{x}_n$ — значение из интервала $\left(\pi, \frac{3\pi}{2}\right)$ такое, что $f(\bar{x}_n) > 0$. Значение $\bar{x}_n$ легко найти подбором*). Так, например, при $n = 2$,

*) Конечно, можно было бы взять $\bar{x}_n = \pi$, но это невыгодно, так как $f'(\pi) = 0$.

полагая приближенно

$$\bar{x}_2 = 4{,}49340 = \text{arc } 257°27'12'',$$

будем иметь:

$$f(\bar{x}_2) = \sin 257°27'12'' - 4{,}49340 \cdot \cos 257°27'12'' =$$
$$= -0{,}97612 + 4{,}49340 \cdot 0{,}21724 =$$
$$= -0{,}97612 + 0{,}97614 = +0{,}00002.$$

Следовательно, $\bar{x}_2$ выбрано правильно и

$$4{,}49340 < \xi < 4{,}49343.$$

Можно положить

$$\xi = 4{,}4934,$$

где все знаки верные.

Оценку погрешности значения $x_2$ нетрудно провести более точно. Так как при $x \in [\bar{x}_2, x_2]$ производная $f'(x)$ убывает и $f'(x) < 0$, то

$$m_1 = \min |f'(x)| = |f'(\bar{x}_2)|.$$

Отсюда

$$m_1 = 4{,}49340 \cdot 0{,}97612 > 4$$

и, следовательно,

$$|\xi - x_2| \leqslant \frac{|f(x_2)|}{4} = \frac{0{,}00003}{4} < 10^{-5}.$$

Таким образом,

$$\xi = 4{,}49343 - 0{,}00001\ \theta,$$

где $0 < \theta < 1$.

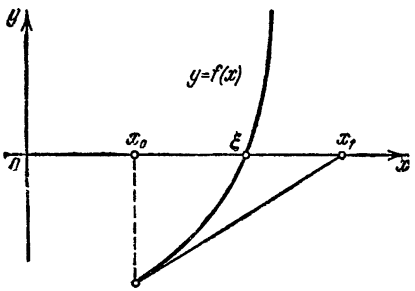


Рис. 21.

Пример 3. Рассмотрим уравнение

$$f(x) = 0, \qquad (10)$$

где $f''(x)$ непрерывна и сохраняет постоянный знак при $-\infty < x < +\infty$. В силу теоремы Ролля уравнение (10) не может иметь более двух действительных корней. Отметим два важных для практики случая.

I. Пусть

$$f(x_0) f'(x_0) < 0, \qquad f(x_0) f''(x) < 0$$

(рис. 21).

Тогда уравнение (10) имеет единственный корень $\xi$ в интервале $(x_0, x_1)$, где

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Корень $\xi$ может быть вычислен с заданной точностью методом Ньютона.

II. Пусть

$$f'(x_0)=0, \qquad f(x_0)f''(x)<0.$$

Тогда уравнение (10) имеет два корня $\xi$ и $\xi'$ в интервале $(-\infty, +\infty)$ (рис. 22).

Преобразуя левую часть уравнения (10) по формуле Тейлора, приближенно имеем:

$$f(x_0)+f'(x_0)(x-x_0)+\frac{1}{2}f''(x_0)(x-x_0)^2=0$$

или

$$f(x_0)+\frac{1}{2}f''(x_0)(x-x_0)^2=0.$$

Отсюда для корней $\xi$ и $\xi'$ получаем начальные приближения

$$x_1=x_0-\sqrt{-\frac{2f(x_0)}{f''(x_0)}}$$

и

$$x_1'=x_0+\sqrt{-\frac{2f(x_0)}{f''(x_0)}},$$

представляющие собой абсциссы точек пересечения параболы

$$Y=f(x_0)+\frac{1}{2}f''(x_0)(x-x_0)^2$$

с осью $Ox$ (рис. 23). Дальнейшие уточнения корней могут быть произведены обычным методом Ньютона.

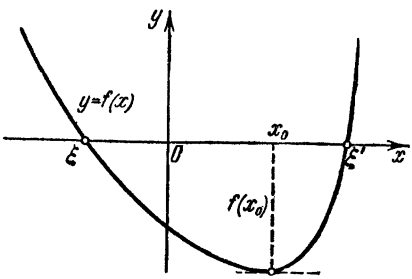


Рис. 22.

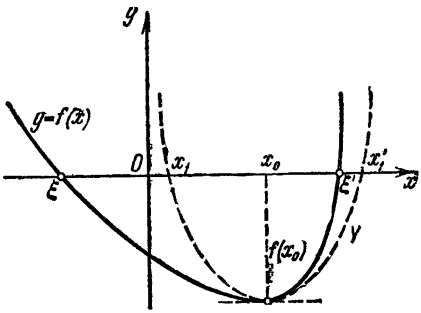


Рис. 23.

Утверждения I и II геометрически очевидны. Строгое доказательство предоставляем провести читателю.

## § 6. Видоизмененный метод Ньютона

Если производная $f'(x)$ мало изменяется на отрезке $[a, b]$, то в формуле (3) предыдущего параграфа можно положить:

$$f'(x_n)\approx f'(x_0). \tag{1}$$

Отсюда для корня $\xi$ уравнения $f(x)=0$ получаем последовательные приближения

$$x_{n+1}=x_n-\frac{f(x_n)}{f'(x_0)} \qquad (n=0,\ 1,\ \ldots). \tag{2}$$

Геометрически этот способ означает, что мы заменяем касательные в точках $B_n[x_n,\ f(x_n)]$ прямыми, параллельными касательной к кривой $y=f(x)$, в ее фиксированной точке $B_0[x_0,\ f(x_0)]$ (рис. 24).

Формула (1) избавляет нас от необходимости вычислять каждый раз значения производной $f'(x_n)$; поэтому эта формула весьма

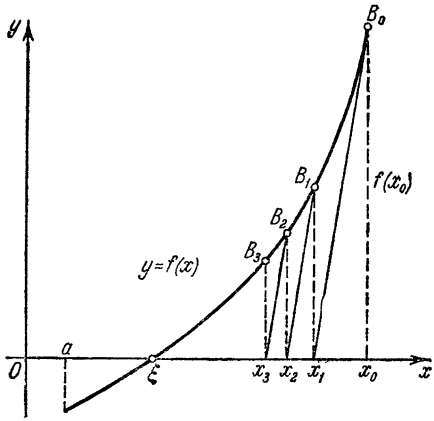


Рис. 24.

полезна, если $f'(x_n)$ сложна. Можно доказать, что в предположении постоянства знаков производных $f'(x)$ и $f''(x)$ последовательные приближения (2) дают сходящийся процесс.

## § 7. Комбинированный метод

Пусть $f(a)f(b)<0$, а $f'(x)$ и $f''(x)$ сохраняют постоянные знаки на отрезке $[a,\ b]$. Соединяя способ пропорциональных частей и метод Ньютона, получаем метод, на каждом этапе которого находим значения по недостатку и значения по избытку точного корня $\xi$ уравнения $f(x)=0$.

Отсюда, в частности, вытекает, что цифры, общие для $x_n$ и $\bar{x}_n$, обязательно принадлежат точному корню $\xi$. Теоретически здесь возможны четыре случая:

1) $f'(x)>0$; $f''(x)>0$ (рис. 25);
2) $f'(x)>0$; $f''(x)<0$ (рис. 26);
3) $f'(x)<0$; $f''(x)>0$ (рис. 27);
4) $f'(x)<0$; $f''(x)<0$ (рис. 28).

Мы ограничимся разбором первого случая. Остальные случаи изучаются аналогично, причем характер вычислений легко понять из

соответствующих чертежей. Заметим, что эти случаи можно свести к первому, если заменить рассматриваемое уравнение $f(x)=0$

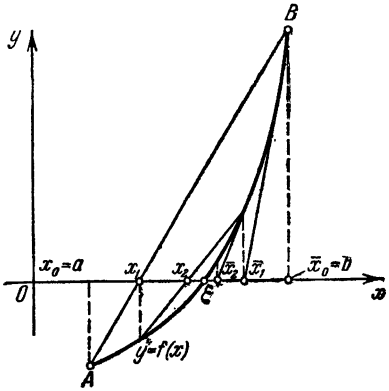


Рис. 25.

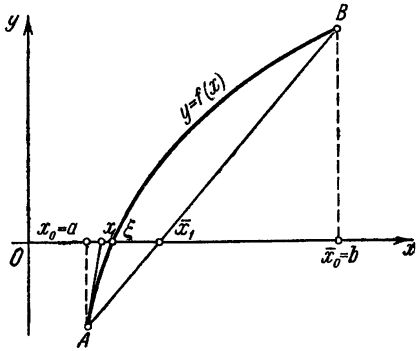


Рис. 26.

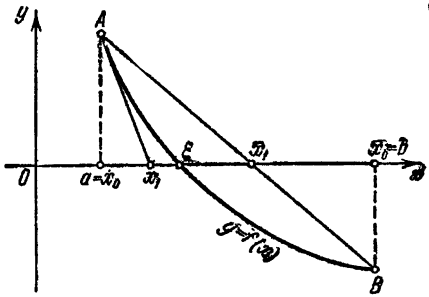


Рис. 27.

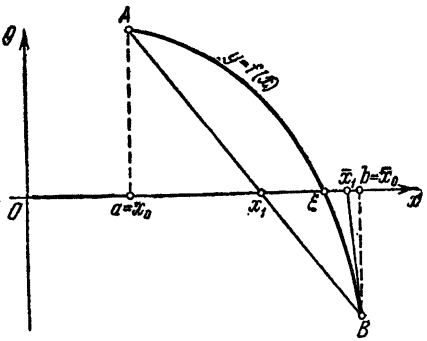


Рис. 28.

равносильными ему уравнениями: $-f(x)=0$ или $\pm f(-z)=0$, где $z=-x$.

Итак, пусть $f'(x)>0$ и $f''(x)>0$ при $a\leqslant x\leqslant b$. Полагаем $x_0=a$; $\bar{x}_0=b$ и

$$x_{n+1}=x_n-\frac{f(x_n)}{f(\bar{x}_n)-f(x_n)}(\bar{x}_n-x_n)\ ^{*)}; \tag{1}$$

$$\bar{x}_{n+1}=\bar{x}_n-\frac{f(\bar{x}_n)}{f'(\bar{x}_n)}\qquad (n=0,\ 1,\ 2,\ \ldots)\ ^{*)}. \tag{1'}$$

Из доказанного выше (§§ 5 и 6) следует, что

$$x_n<\xi<\bar{x}_n$$

*) На каждом шаге метод хорд применяется к новому отрезку $[x_n, \bar{x}_n]$.

и

$$0 < \xi - x_n < \bar{x}_n - x_n. \tag{2}$$

Если допустимая абсолютная погрешность приближенного корня $x_n$ задана заранее и равна $\varepsilon$, то процесс сближения прекращается в тот момент, когда будет обнаружено, что $\bar{x}_n - x_n < \varepsilon$. По окончании процесса за значение корня $\xi$ лучше всего взять среднее арифметическое полученных последних значений:

$$\bar{\xi} = \frac{1}{2}(x_n + \bar{x}_n).$$

Пример. Вычислить с точностью до 0,0005 единственный положительный корень уравнения

$$f(x) \equiv x^5 - x - 0{,}2 = 0.$$

Решение. Так как $f(1) < 0$ и $f(1{,}1) > 0$, то корень содержится в интервале (1; 1,1). Имеем:

$$f'(x) = 5x^4 - 1 \quad \text{и} \quad f''(x) = 20x^3.$$

В выбранном нами интервале $f'(x) > 0$; $f''(x) > 0$, т. е. знаки производных сохраняются.

Применим комбинированный метод, полагая $x_0 = 1$ и $\bar{x}_0 = 1{,}1$. Так как

$$f(x_0) = f(1) = -0{,}2; \qquad f(\bar{x}_0) = f(1{,}1) = 0{,}3105;$$
$$f'(\bar{x}_0) = f'(1{,}1) = 6{,}3205,$$

то формулы (1) и (1′) дают:

$$x_1 = 1 + \frac{0{,}1 \cdot 0{,}2}{0{,}51051} \approx 1{,}039; \qquad \bar{x}_1 = 1{,}1 - \frac{0{,}31051}{6{,}3205} \approx 1{,}051.$$

Ввиду того, что $\bar{x}_1 - x_1 = 0{,}012$, то точность недостаточная. Находим следующую пару приближений:

$$x_2 = 1{,}039 + \frac{0{,}012 \cdot 0{,}0282}{0{,}0595} \approx 1{,}04469; \qquad \bar{x}_2 = 1{,}051 - \frac{0{,}0313}{5{,}1005} \approx 1{,}04487.$$

Здесь $\bar{x}_2 - x_2 = 0{,}00018$, т. е. нужная степень точности достигнута. Можно положить

$$\bar{\xi} = \frac{1}{2}(1{,}04469 + 1{,}04487) = 1{,}04478 \approx 1{,}045$$

с абсолютной погрешностью, меньшей

$$\frac{1}{2} \cdot 0{,}00018 + 0{,}00022 = 0{,}00031 < \frac{1}{2} \cdot 10^{-3}.$$

## § 8. Метод итерации

Одним из наиболее важных способов численного решения уравнений является *метод итерации* *). Сущность этого метода заключается в следующем. Пусть дано уравнение

$$f(x)=0, \tag{15}$$

где $f(x)$ — непрерывная функция, и требуется определить его вещественные корни. Заменим уравнение (1) равносильным уравнением

$$x=\varphi(x). \tag{2}$$

Выберем каким-либо способом грубо приближенное значение корня $x_0$ и подставим его в правую часть уравнения (2). Тогда получим некоторое число

$$x_1=\varphi(x_0). \tag{3}$$

Подставляя теперь в правую часть равенства (3) вместо $x_0$ число $x_1$, получим новое число $x_2=\varphi(x_1)$. Повторяя этот процесс, будем иметь последовательность чисел

$$x_n=\varphi(x_{n-1}) \qquad (n=1,\ 2,\ \ldots). \tag{4}$$

Если эта последовательность — сходящаяся, т. е. существует предел $\xi=\lim\limits_{n\to\infty} x_n$, то, переходя к пределу в равенстве (4) и предполагая функцию $\varphi(x)$ непрерывной, найдем:

$$\lim_{n\to\infty} x_n=\varphi(\lim_{n\to\infty} x_{n-1})$$

или

$$\xi=\varphi(\xi). \tag{5}$$

Таким образом, предел $\xi$ является корнем уравнения (2) и может быть вычислен по формуле (4) с любой степенью точности.

Геометрически способ итерации может быть пояснен следующим образом. Построим на плоскости $xOy$ графики функций $y=x$ и $y=\varphi(x)$. Каждый действительный корень $\xi$ уравнения (2) является абсциссой точки пересечения $M$ кривой $y=\varphi(x)$ с прямой $y=x$ (рис. 29).

Отправляясь от некоторой точки $A_0[x_0;\ \varphi(x_0)]$, строим ломаную линию $A_0B_1A_1B_2A_2\ldots$ («лестница»), звенья которой попеременно параллельны оси $Ox$ и оси $Oy$, вершины $A_0$, $A_1$, $A_2$, ... лежат на кривой $y=\varphi(x)$, а вершины $B_1$, $B_2$, $B_3$, ... — на прямой $y=x$. Общие абсциссы точек $A_1$ и $B_1$, $A_2$ и $B_2$, ..., очевидно, представляют собой соответственно последовательные приближения $x_1$, $x_2$, ... корня $\xi$.

---

*) Часто метод итерации называют *методом последовательных приближений*.

Возможен также (рис. 30) другой вид ломаной $A_0B_1A_1B_2A_2\ldots$ («спираль»). Легко сообразить, что решение в виде «лестницы» получается, если производная $\varphi'(x)$ положительна, а решение в виде «спирали», если $\varphi'(x)$ отрицательна.

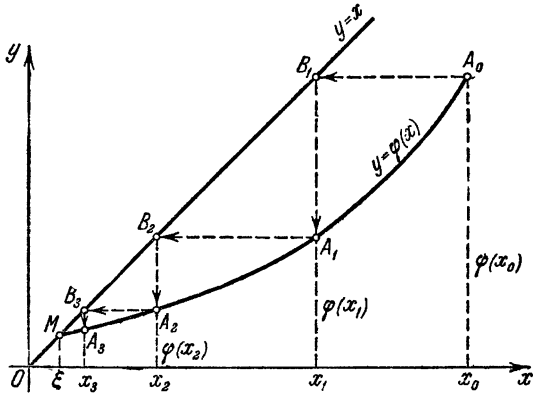


Рис. 29.

На рис. 29 кривая $y=\varphi(x)$ в окрестности корня $\xi$ — пологая, т. е. $|\varphi'(x)|<1$, и процесс итерации сходится. Однако, если рас-

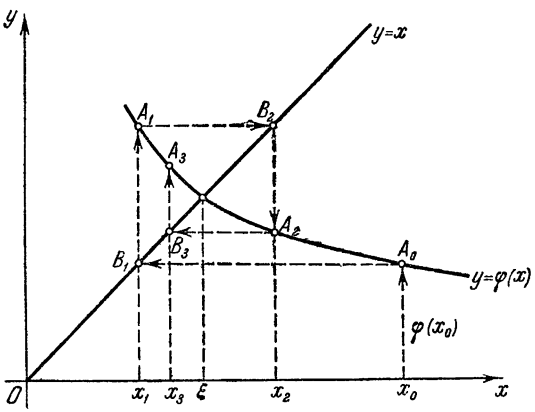


Рис. 30.

смотреть случай, где $|\varphi'(x)|>1$, то процесс итерации может быть расходящимся (рис. 31). Поэтому для практического применения метода итерации нужно выяснить достаточные условия сходимости итерационного процесса.

Теорема 1. *Пусть функция $\varphi(x)$ определена и дифференцируема на отрезке $[a, b]$, причем все ее значения $\varphi(x)\in[a, b]$.*

*Тогда, если существует правильная дробь $q$ такая*), что*

$$|\varphi'(x)| \leqslant q < 1 \tag{6}$$

*при $a < x < b$, то: 1) процесс итерации*

$$x_n = \varphi(x_{n-1}) \qquad (n = 1, 2, \ldots) \tag{7}$$

*сходится независимо от начального значения $x_0 \in [a, b]$; 2) предельное значение*

$$\xi = \lim_{n \to \infty} x_n$$

*является единственным корнем уравнения*

$$x = \varphi(x) \tag{8}$$

*на отрезке $[a, b]$.*

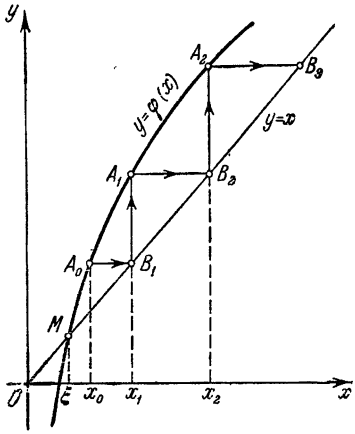


Рис. 31.

Доказательство. Рассмотрим два последовательных приближения

$$x_n = \varphi(x_{n-1}) \quad \text{и} \quad x_{n+1} = \varphi(x_n)$$

(которые в силу условий теоремы заведомо имеют смысл). Отсюда

$$x_{n+1} - x_n = \varphi(x_n) - \varphi(x_{n-1}).$$

Применяя теорему Лагранжа, будем иметь:

$$x_{n+1} - x_n = (x_n - x_{n-1})\,\varphi'(\bar{x}_n),$$

где $\bar{x}_n \in (x_{n-1}, x_n)$. Следовательно, на основании условия (6) получим:

$$|x_{n+1} - x_n| \leqslant q\,|x_n - x_{n-1}|. \tag{9}$$

Отсюда, давая значения $n = 1, 2, 3, \ldots$, последовательно выводим:

$$|x_2 - x_1| \leqslant q\,|x_1 - x_0|;$$
$$|x_3 - x_2| \leqslant q\,|x_2 - x_1| \leqslant q^2\,|x_1 - x_0|;$$
$$\cdots\cdots\cdots\cdots\cdots\cdots$$
$$|x_{n+1} - x_n| \leqslant q^n\,|x_1 - x_0|. \tag{10}$$

Рассмотрим ряд

$$x_0 + (x_1 - x_0) + (x_2 - x_1) + \ldots + (x_n - x_{n-1}) + \ldots, \tag{11}$$

для которого наши последовательные приближения $x_n$ являются $(n+1)$-ми частными суммами, т. е.

$$x_n = S_{n+1}.$$

---

*) За число $q$ можно принять наименьшее значение или нижнюю грань модуля производной $|\varphi'(x)|$ при $a \leqslant x \leqslant b$.

В силу неравенства (10) члены ряда (11) по абсолютной величине меньше соответствующих членов геометрической прогрессии со знаменателем $q < 1$, поэтому ряд (11) сходится и притом абсолютно. Следовательно, существует

$$\lim_{n \to \infty} S_{n+1} = \lim_{n \to \infty} x_n = \xi,$$

причем, очевидно, $\xi \in [a, b]$.

Переходя к пределу в равенстве (7), в силу непрерывности функции $\varphi(x)$ получаем:

$$\xi = \varphi(\xi). \tag{12}$$

Таким образом, $\xi$ есть корень уравнения (8). Другого корня на отрезке $[a, b]$ уравнение (8) не имеет. Действительно, если

$$\bar{\xi} = \varphi(\bar{\xi}), \tag{13}$$

то из равенств (12) и (13) получим:

$$\bar{\xi} - \xi = \varphi(\bar{\xi}) - \varphi(\xi)$$

и, следовательно,

$$(\bar{\xi} - \xi)[1 - \varphi'(c)] = 0, \tag{14}$$

где $c \in [\xi, \bar{\xi}]$. Так как выражение в квадратной скобке в равенстве (14) не равно нулю, то $\xi = \bar{\xi}$, т. е. корень $\xi$ — единственный.

Замечание 1. Теорема остается верной, если функция $\varphi(x)$ определена и дифференцируема в бесконечном интервале $-\infty < x < +\infty$, причем при $x \in (-\infty, +\infty)$ выполнено неравенство (6).

Замечание 2. В условиях теоремы 1 метод итерации сходится *при любом выборе начального значения* $x_0$ из $[a, b]$. Благодаря этому он является *самоисправляющимся*, т. е. отдельная ошибка в вычислениях, не выводящая за пределы отрезка $[a, b]$, не повлияет на конечный результат, так как ошибочное значение можно рассматривать как новое начальное значение $x_0$. Возможно, возрастет лишь объем работы. Свойство самоисправления делает метод итерации одним из надежнейших методов вычислений. Само собой разумеется, что систематические ошибки при применении этого метода могут помешать получению нужного результата.

**Оценка приближения.** Из формулы (10) имеем:

$$|x_{n+p} - x_n| \leqslant |x_{n+p} - x_{n+p-1}| + |x_{n+p-1} - x_{n+p-2}| + \ldots$$
$$\ldots + |x_{n+1} - x_n| \leqslant q^{n+p-1}|x_1 - x_0| + q^{n+p-2}|x_1 - x_0| + \ldots$$
$$\ldots + q^n|x_1 - x_0| = q^n|x_1 - x_0|(1 + q + q^2 + \ldots + q^{p-1})$$

Просуммировав геометрическую прогрессию, получим:

$$|x_{n+p} - x_n| \leqslant q^n|x_1 - x_0|\frac{1 - q^p}{1 - q} < \frac{q^n}{1 - q}|x_1 - x_0|.$$

Устремляя число $p$ к бесконечности и учитывая, что $\lim\limits_{p\to\infty} x_{n+p}=\xi$, находим окончательно:

$$|\xi-x_n|\leqslant\frac{q^n}{1-q}|x_1-x_0|. \qquad (15)$$

Отсюда ясно, что сходимость процесса итерации будет тем быстрее, чем меньше число $q$.

Для оценки приближения можно также дать другую формулу, полезную в некоторых случаях. Пусть

$$f(x)=x-\varphi(x).$$

Очевидно, что $f'(x)=1-\varphi'(x)\geqslant 1-q$. Отсюда, учитывая, что $f(\xi)=0$, получим:

$$|x_n-\varphi(x_n)|=|f(x_n)-f(\xi)|=|x_n-\xi||f'(\bar{x}_n)|\geqslant(1-q)|x_n-\xi|,$$

где $\bar{x}_n\in(x_n,\xi)$, и, следовательно,

$$|x_n-\xi|\leqslant\frac{|x_n-\varphi(x_n)|}{1-q}, \qquad (16)$$

т. е.

$$|\xi-x_n|\leqslant\frac{|x_{n+1}-x_n|}{1-q}. \qquad (16')$$

Используя формулу (9), имеем также:

$$|\xi-x_n|\leqslant\frac{q}{1-q}|x_n-x_{n-1}|; \qquad (16'')$$

откуда, в частности, следует, что если $q\leqslant\frac{1}{2}$, то

$$|\xi-x_n|\leqslant|x_n-x_{n-1}|,$$

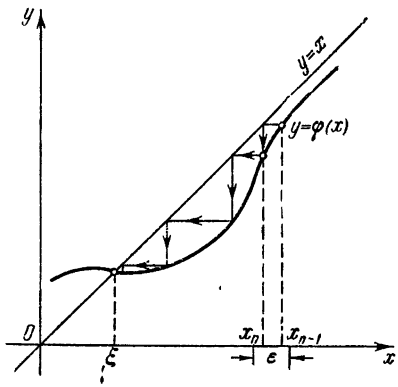

Рис. 32.

т. е. в этом случае из неравенства $|x_n-x_{n-1}|<\varepsilon$ вытекает неравенство

$$|\xi-x_n|<\varepsilon.$$

Замечание. Существует распространенное мнение, что если при применении метода итерации два последовательных приближения $x_{n-1}$ и $x_n$ совпадают между собой с заданной точностью $\varepsilon$ (например, для этих приближений установились $m$ первых десятичных знаков), то с той же точностью справедливо равенство $\xi\approx x_n$ (т. е., в частности, в приведенном примере $m$ знаков приближенного числа $x_n$ являются верными!). В общем случае, как наглядно показывает рис. 32, это утверждение ошибочно. Более того, легко показать, что если $\varphi'(x)$ близка к 1, то величина $|\xi-x_n|$ может быть большой, хотя величина $|x_n-x_{n-1}|$ весьма мала.

Формула (16″) дает возможность оценить погрешность приближенного значения $x_n$ по расхождению двух последовательных приближений $x_{n-1}$ и $x_n$.

Процесс итерации следует продолжать до тех пор, пока для двух последовательных приближений $x_{n-1}$ и $x_n$ не будет обеспечено выполнение неравенства

$$|x_n - x_{n-1}| \leqslant \frac{1-q}{q}\varepsilon,$$

где $\varepsilon$ — заданная предельная абсолютная погрешность корня $\xi$ и $|\varphi'(x)| \leqslant q$. Тогда в силу формулы (16″) будет иметь место неравенство

$$|\xi - x_n| \leqslant \varepsilon,$$

т. е.

$$\xi = x_n \pm \varepsilon.$$

Заметим, что если

$$x_n = \varphi(x_{n-1})$$

и

$$\xi = \varphi(\xi),$$

то

$$|\xi - x_n| = |\varphi(\xi) - \varphi(x_{n-1})| = \\ = |\xi - x_{n-1}||\varphi'(\bar{x}_{n-1})| \leqslant q|\xi - x_{n-1}| \qquad (\bar{x}_{n-1} \in (x_{n-1}, \xi)),$$

т. е.

$$|\xi - x_n| \leqslant |\xi - x_{n-1}|.$$

Таким образом, при сходящемся итеративном процессе погрешность $|\xi - x_n|$ стремится к нулю монотонно, т. е. каждое следующее значение $x_n$ является более точным, чем предшествующее значение $x_{n-1}$. Конечно, при всех этих выводах игнорируются погрешности округлений, т. е. предполагается, что последовательные приближения находятся точно.

На практике обычно бывает так, что грубым приемом устанавливается существование корня $\xi$ уравнения (2) и методом итерации требуется получить достаточно точное приближенное значение корня, причем неравенство (6) выполняется лишь в некоторой окрестности $(a, b)$ этого корня. Здесь при неудачном выборе начального значения $x_0$ последовательные приближения $x_n = \varphi(x_{n-1})$ $(n = 1, 2, \ldots)$ могут покинуть интервал $(a, b)$ или даже потерять смысл. Поэтому полезна другая формулировка теоремы 1.

Теорема 2. *Пусть функция $\varphi(x)$ определена и дифференцируема на некотором отрезке $[a, b]$, причем уравнение*

$$x = \varphi(x) \tag{17}$$

*имеет корень* $\xi$, *лежащий в более узком отрезке* $[\alpha, \beta]$, *где* $\alpha = a + \frac{1}{3}(b-a)$ *и* $\beta = b - \frac{1}{3}(b-a)$ (рис. 33).

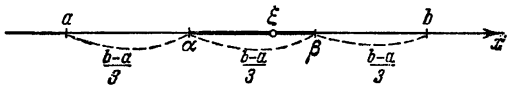


Рис. 33.

*Тогда, если*: а) $|\varphi'(x)| \leqslant q < 1$ *при* $a < x < b$; б) *начальное приближение* $x_0 \in [\alpha, \beta]$, *то*:

1) *все последовательные приближения содержатся в интервале* $(a, b)$:

$$x_n = \varphi(x_{n-1}) \in (a, b) \quad (n = 1, 2, \ldots),$$

2) *процесс последовательных приближений — сходящийся, т. е. существует*

$$\lim_{n\to\infty} x_n = \xi,$$

*причем* $\xi$ — *единственный корень на отрезке* $[a, b]$ *уравнения* (17), *и*

3) *справедлива оценка* (15).

Доказательство. 1) Действительно, пусть

$$x_0 \in [\alpha, \beta].$$

Тогда равенство

$$x_1 = \varphi(x_0),$$

очевидно, имеет смысл. Используя равенство

$$\xi = \varphi(\xi),$$

на основании теоремы Лагранжа получаем:

$$|x_1 - \xi| = |\varphi(x_0) - \varphi(\xi)| = |x_0 - \xi|\,|\varphi'(\bar{x}_0)| \leqslant q(\beta - \alpha) < \frac{b-a}{3};$$

отсюда

$$x_1 \in (a, b).$$

Вообще, если $x_{n-1} \in (a, b)$ $(n = 1, 2, \ldots)$ и $|x_{n-1} - \xi| < \frac{b-a}{3}$, то

$$x_n = \varphi(x_{n-1})$$

имеет смысл и

$$|x_n - \xi| = |\varphi(x_{n-1}) - \varphi(\xi)| =$$
$$= |x_{n-1} - \xi|\,|\varphi'(\bar{x}_{n-1})| \leqslant q\,|x_{n-1} - \xi| < \frac{b-a}{3}.$$

Следовательно, $x_n \in (a, b)$, где $n = 1, 2, 3, \ldots$

Что касается утверждений 2) и 3), то доказательство их вполне аналогично доказательству теоремы 1.

Замечание. Пусть в некоторой окрестности $(a, b)$ корня $\xi$ уравнения (17) производная $\varphi'(x)$ сохраняет постоянный знак и выполнено неравенство

$$|\varphi'(x)| \leqslant q < 1.$$

*Тогда, если производная $\varphi'(x)$ положительна, то последовательные приближения*

$$x_n = \varphi(x_{n-1}) \quad (n = 1, 2, \ldots), \quad x_0 \in (a, b)$$

*сходятся к корню $\xi$ монотонно.*

*Если же производная $\varphi'(x)$ отрицательна, то последовательные приближения колеблются около корня $\xi$.*

1) В самом деле, пусть $0 \leqslant \varphi'(x) \leqslant q < 1$ и, например,

$$x_0 < \xi.$$

Тогда

$$x_1 - \xi = \varphi(x_0) - \varphi(\xi) = (x_0 - \xi)\varphi'(\xi_1) < 0,$$

где $\xi_1 \in (x_0, \xi)$, причем

$$|x_1 - \xi| \leqslant q|x_0 - \xi| < |x_0 - \xi|.$$

Следовательно,

$$x_0 < x_1 < \xi.$$

Применяя метод математической индукции, получаем:

$$x_0 < x_1 < x_2 < \ldots < \xi$$

(рис. 34а).

Аналогичный результат получается при $x_0 > \xi$.

Таким образом, в случае положительной производной $\varphi'(x)$ достаточно выбрать лишь начальное приближение $x_0$, принадлежащее

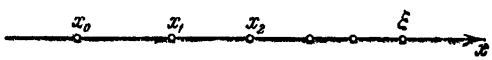


Рис. 34а.

окрестности $(a, b)$ интересующего нас корня $\xi$; все остальные приближения $x_n$ $(n = 1, 2, \ldots)$ автоматически будут содержаться в этой окрестности и с увеличением номера $n$ монотонно будут стремиться к корню $\xi$.

2) Пусть $-1 < -q \leqslant \varphi'(x) \leqslant 0$ и, например, $x_0 < \xi$, причем $x_1 = \varphi(x_0) \in (a, b)$.

Имеем:

$$x_1 - \xi = \varphi(x_0) - \varphi(\xi) = (x_0 - \xi)\varphi'(\xi_1) > 0,$$

т. е. $x_1 > \xi$ и $|x_1 - \xi| < |x_0 - \xi|$.

Повторяя эти рассуждения для приближений $x_1$, $x_2$, ..., получаем:

$$x_0 < x_2 < \ldots < \xi < \ldots < x_3 < x_1,$$

т. е. последовательные приближения будут то меньше, то больше корня $\xi$ (рис. 34б).

Таким образом, в случае отрицательной производной $\varphi'(x)$, если два приближения $x_0$ и $x_1$ принадлежат окрестности $(a, b)$ корня $\xi$,

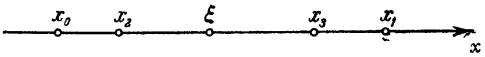


Рис. 34б.

то все остальные приближения $x_n$ $(n = 2, 3, \ldots)$ также принадлежат этой окрестности, причем последовательность $\{x_n\}$ «обертывает» корень $\xi$.

Заметим, что, очевидно,

$$|\xi - x_n| \leqslant |x_n - x_{n-1}|,$$

т. е. в этом случае установившиеся знаки приближения $x_n$ обязательно принадлежат точному корню $\xi$.

Пример 1. Найти действительные корни уравнения $x - \sin x = 0{,}25$ с точностью до трех значащих цифр.

Решение. Представим данное уравнение в виде

$$x = \sin x + 0{,}25.$$

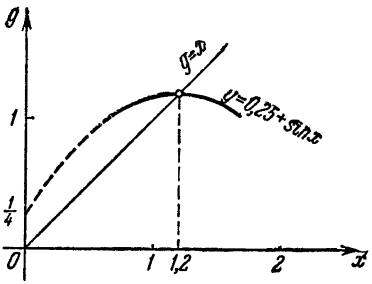


Рис. 35.

Графическим способом устанавливаем, что уравнение имеет в отрезке $[1{,}1;\ 1{,}3]$ один вещественный корень $\xi$, приближенно равный $x_0 = 1{,}2$ (рис. 35).

Придерживаясь обозначений теоремы 2, примем:

$$\alpha = 1{,}1 \quad \text{и} \quad \beta = 1{,}3;$$

отсюда

$$a = \alpha - (\beta - \alpha) = 0{,}9 \approx \text{arc } 52^\circ$$

и

$$b = \beta + (\beta - \alpha) = 1{,}5 \approx \text{arc } 86^\circ.$$

Так как

$$\varphi(x) = \sin x + 0{,}25$$

и

$$\varphi'(x) = \cos x,$$

то при $0{,}9 < x < 1{,}5$ имеем:

$$|\varphi'(x)| \leqslant \cos 52^\circ \approx 0{,}62 = q.$$

Если мы выберем $x_0 \in (1{,}1;\ 1{,}3)$, то все условия теоремы 2 будут полностью соблюдены и, следовательно, гарантировано, что последовательные приближения

$$x_n = \sin x_{n-1} + 0{,}25 \quad (n = 1,\ 2,\ \ldots)$$

1) содержатся в интервале (0,9; 1,5) и 2) $x_n \to \xi$ при $n \to \infty$.

Выбирая $x_0 = 1{,}2$ и задаваясь, согласно условию задачи, предельной абсолютной погрешностью

$$\varepsilon = \frac{1}{2} \cdot 10^{-2},$$

строим последовательные приближения $x_n$ $(n = 1,\ 2,\ \ldots)$ до тех пор, пока два соседних приближения $x_{n-1}$ и $x_n$ не совпадут друг с другом в пределах точности, равной

$$\frac{1-q}{q}\varepsilon = 0{,}51 \cdot \frac{1}{2} \cdot 10^{-2} \approx 0{,}0025.$$

Имеем:

$$x_1 = \sin 1{,}2 + 0{,}25 = 0{,}932 + 0{,}25 = 1{,}182;$$
$$x_2 = \sin 1{,}182 + 0{,}25 = 0{,}925 + 0{,}25 = 1{,}175;$$
$$x_3 = \sin 1{,}175 + 0{,}25 = 0{,}923 + 0{,}25 = 1{,}173;$$
$$x_4 = \sin 1{,}173 + 0{,}25 = 0{,}922 + 0{,}25 = 1{,}172;$$
$$x_5 = \sin 1{,}172 + 0{,}25 = 0{,}922 + 0{,}25 = 1{,}172.$$

Четвертое и пятое приближения совпали с точностью до четырех значащих цифр. Поэтому (см. (16″))

$$|x_5 - \xi| \leqslant \frac{0{,}62 \cdot 0{,}001}{1 - 0{,}62} = 0{,}0016.$$

Так как предельная абсолютная погрешность приближенного корня $x_5$, включая погрешность округления, не превышает

$$E = 0{,}0016 + 0{,}002 < \frac{1}{2} \cdot 10^{-2},$$

то можно принять:

$$\xi = 1{,}17 \pm 0{,}005.$$

Замечание. Данное уравнение

$$f(x) = 0 \tag{18}$$

можно записать в виде равенства

$$x = \varphi(x), \tag{18'}$$

выбирая различным образом функцию $\varphi(x)$.

Способ записи (18′) отнюдь не безразличен: в одних случаях $|\varphi'(x)|$ окажется малой в окрестности искомого корня $\xi$, в других —

большой. Для метода итераций выгодно то представление (18′), при котором выполнено неравенство

$$|\varphi'(x)| \leqslant q < 1, \tag{19}$$

причем, чем меньше число $q$, тем быстрее, вообще говоря, последовательные приближения сходятся к корню $\xi$.

Укажем один достаточно общий прием приведения уравнения (18) к виду (18′), для которого обеспечено выполнение неравенства (19). Пусть искомый корень $\xi$ уравнения лежит на отрезке $[a, b]$, причем

$$0 < m_1 \leqslant f'(x) \leqslant M_1 \tag{20}$$

при $a \leqslant x \leqslant b$ *). В частности, за $m_1$ можно взять наименьшее значение производной $f'(x)$ на отрезке $[a, b]$, которое должно быть положительным, а за $M_1$ — наибольшее значение $f'(x)$ на отрезке $[a, b]$. Заменим уравнение (18) эквивалентным ему уравнением

$$x = x - \lambda f(x) \quad (\lambda > 0).$$

Можно положить $\varphi(x) = x - \lambda f(x)$.

Подберем параметр $\lambda$ таким образом, чтобы в данной окрестности $[a, b]$ корня $\xi$ было выполнено неравенство

$$0 \leqslant \varphi'(x) = 1 - \lambda f'(x) \leqslant q < 1. \tag{21}$$

Отсюда на основании выражения (20) получаем:

$$0 \leqslant 1 - \lambda M_1 \leqslant 1 - \lambda m_1 \leqslant q.$$

Следовательно, можно выбрать:

$$\lambda = \frac{1}{M_1}$$

и

$$q = 1 - \frac{m_1}{M_1} < 1.$$

Таким образом, неравенство (21) выполнено.

Пример 2. Найти наибольший положительный корень $\xi$ уравнения

$$x^3 + x = 1000 \tag{22}$$

с точностью до $10^{-4}$.

Решение. Грубой прикидкой получаем приближенное значение корня $x_0 = 10$, причем, очевидно, $\xi < x_0$.

---

*) Если производная $f'(x)$ отрицательна, то вместо уравнения $f(x) = 0$ рассматриваем уравнение $-f(x) = 0$.

Уравнение (22) можно записать в виде

$$x = 1000 - x^3, \tag{22'}$$

или

$$x = \frac{1000}{x^2} - \frac{1}{x}, \tag{22''}$$

или

$$x = \sqrt[3]{1000 - x}, \tag{22'''}$$

и т. п. Наиболее выгодным из приведенных вариантов оказывается вариант (22'''), так как, взяв за основной интервал (9,10) и положив

$$\varphi(x) = \sqrt[3]{1000 - x},$$

будем иметь

$$\varphi'(x) = \frac{-1}{3\sqrt[3]{(1000-x)^2}}.$$

Отсюда

$$|\varphi'(x)| \leqslant \frac{1}{3\sqrt[3]{990^2}} \approx \frac{1}{300} = q.$$

Таблица 4

**Значения последовательных приближений $x_n$ и $y_n$**

| $n$ | $x_n$ | $y_n$ |
|---|---|---|
| 0 | 10 | 990 |
| 1 | 9,96655 | 990,03345 |
| 2 | 9,96666 | 990,03334 |
| 3 | 9,96667 | |

Вычисляем последовательные приближения $x_n$ с одним запасным знаком по формулам

$$y_n = 1000 - x_n;$$

$$x_{n+1} = \sqrt[3]{y_n} \quad (n = 0,\ 1,\ 2,\ \ldots).$$

Найденные значения помещены в таблице 4.

Так как $1 - q \approx 1$, то с точностью до $10^{-4}$ можно положить $\xi = 9{,}9667$.

Метод итерации можно применять также для вычисления корней уравнений, заданных в виде степенных рядов.

Пример 3. Найти действительный корень уравнения [2]

$$x - \frac{x^3}{3} + \frac{x^5}{10} - \frac{x^7}{42} + \frac{x^9}{216} - \frac{x^{11}}{1320} + \ldots$$
$$\ldots + (-1)^{n-1}\frac{x^{2n-1}}{(n-1)!\,(2n-1)} + \ldots = 0{,}4431135.$$

Решение. Имеем $x = \varphi(x)$, где

$$\varphi(x) = 0{,}4431135 + \frac{x^3}{3} - \frac{x^5}{10} + \frac{x^7}{42} - \frac{x^9}{216} + \frac{x^{11}}{1320} - \ldots$$

Отбрасывая все степени $x$ выше первой, определяем приближенное значение корня $x_0 = 0{,}44$. Далее,

$$x_1 = \varphi(0{,}44) \approx 0{,}47;$$
$$x_2 = \varphi(0{,}47) \approx 0{,}476;$$
$$x_3 = \varphi(0{,}476) \approx 0{,}4767;$$
$$x_4 = \varphi(0{,}4767) \approx 0{,}47689;$$
$$x_5 = \varphi(0{,}47689) \approx 0{,}476927;$$
$$x_6 = \varphi(0{,}476927) \approx 0{,}476934;$$
$$x_7 = \varphi(0{,}476934) \approx 0{,}476936.$$

Следовательно, $\xi = 0{,}47693$.

Укажем еще один прием улучшения сходимости процесса итерации, который может оказаться полезным в некоторых случаях [7].

Пусть имеем уравнение

$$x = \varphi(x)$$

такое, что в окрестности искомого корня $\xi$ выполнено неравенство

$$|\varphi'(x)| \geqslant k > 1.$$

Тогда процесс итерации для этого уравнения расходится. Однако, если данное уравнение заменить эквивалентным уравнением

$$x = \psi(x),$$

где $\psi(x) = \varphi^{-1}(x)$ — обратная функция, то мы получим уравнение, для которого процесс итерации сходится, так как

$$|\psi'(x)| = \left|\frac{1}{\varphi'(\psi(x))}\right| \leqslant \frac{1}{k} = q < 1.$$

Пример 4. Уравнение

$$f(x) \equiv x^3 - x - 1 = 0 \tag{23}$$

имеет корень $\xi \in (1,2)$, так как $f(1) = -1 < 0$ и $f(2) = 5 > 0$.

Уравнение (23) можно записать в виде

$$x = x^3 - 1. \tag{24}$$

Здесь

$$\varphi(x) = x^3 - 1 \quad \text{и} \quad \varphi'(x) = 3x^2;$$

поэтому

$$\varphi'(x) \geqslant 3 \quad \text{при} \quad 1 \leqslant x \leqslant 2$$

и, следовательно, условия сходимости процесса итерации не выполнены.

Если записать уравнение (23) в виде

$$x=\sqrt[3]{x+1}, \tag{25}$$

то будем иметь:

$$\psi(x)=\sqrt[3]{x+1} \text{ и } \psi'(x)=\frac{1}{3\sqrt[3]{(x+1)^2}}.$$

Отсюда $0<\psi'(x)<\frac{1}{3\sqrt[3]{4}}<\frac{1}{4}$ при $1\leqslant x\leqslant 2$ и, значит, процесс итерации для уравнения (25) быстро сойдется.

## § 9. Метод итерации для системы двух уравнений

Пусть даны два уравнения с двумя неизвестными

$$\left.\begin{aligned} F_1(x, y)=0, \\ F_2(x, y)=0, \end{aligned}\right\} \tag{1}$$

действительные корни которых требуется найти с заданной степенью точности.

Мы предположим, что система (1) допускает лишь изолированные корни. Число этих корней и их грубо приближенные значения можно установить, построив кривые $F_1(x, y)=0$; $F_2(x, y)=0$ и определив координаты их точек пересечения.

Пусть $x=x_0$; $y=y_0$ — приближенные значения корней системы (1), полученные графически или каким-нибудь другим способом (например, грубой прикидкой).

Дадим итерационный процесс, позволяющий при известных условиях уточнить данные приближенные значения корней. Для этого представим систему (1) в виде

$$\left.\begin{aligned} x=\varphi_1(x, y), \\ y=\varphi_2(x, y) \end{aligned}\right\} \tag{2}$$

и построим последовательные приближения по следующим формулам:

$$\begin{aligned} x_1=\varphi_1(x_0, y_0);\ y_1=\varphi_2(x_0, y_0); \\ x_2=\varphi_1(x_1, y_1);\ y_2=\varphi_2(x_1, y_1); \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ x_{n+1}=\varphi_1(x_n, y_n);\ y_{n+1}=\varphi_2(x_n, y_n); \\ \cdots\cdots\cdots\cdots\cdots\cdots \end{aligned} \tag{3}$$

Если итерационный процесс (3) сходится, т. е. существуют пределы

$$\xi=\lim_{n\to\infty} x_n \text{ и } \eta=\lim_{n\to\infty} y_n,$$

то, предполагая функции $\varphi_1(x, y)$ и $\varphi_2(x, y)$ непрерывными и переходя к пределу в равенстве (3) общего вида, получим:

$$\lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} \varphi_1(x_n, y_n),$$

$$\lim_{n\to\infty} y_{n+1} = \lim_{n\to\infty} \varphi_2(x_n, y_n).$$

Отсюда

$$\xi = \varphi_1(\xi, \eta); \quad \eta = \varphi_2(\xi, \eta),$$

т. е. предельные значения $\xi$ и $\eta$ являются корнями системы (2), а следовательно, и системы (1). Поэтому, взяв достаточно большое число итераций (3), мы получим числа $x_n$ и $y_n$, которые будут отличаться от точных корней $x=\xi$ и $y=\eta$ системы (1) сколь угодно мало. Поставленная задача, таким образом, окажется решенной. Если итерационный процесс (3) расходится, то им пользоваться нельзя.

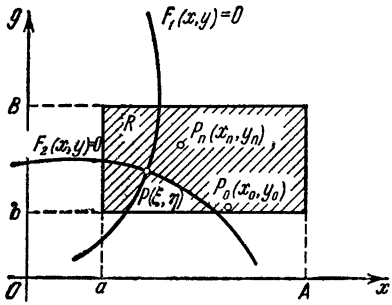

Рис. 36.

Теорема. *Пусть в некоторой замкнутой окрестности $R\{a \leqslant x \leqslant A;\ b \leqslant y \leqslant B\}$ (рис. 36) имеется одна и только одна пара корней $x=\xi$ и $y=\eta$ системы (2). Если: 1) функции $\varphi_1(x, y)$ и $\varphi_2(x, y)$ определены и непрерывно дифференцируемы в $R$; 2) начальные приближения $x_0$, $y_0$ и все последующие приближения $x_n$, $y_n$ $(n=1, 2, \ldots)$ принадлежат $R$; 3) в $R$ выполнены неравенства*

$$\left|\frac{\partial\varphi_1}{\partial x}\right| + \left|\frac{\partial\varphi_2}{\partial x}\right| \leqslant q_1 < 1,$$

$$\left|\frac{\partial\varphi_1}{\partial y}\right| + \left|\frac{\partial\varphi_2}{\partial y}\right| \leqslant q_2 < 1,$$

*то процесс последовательных приближений (3) сходится к корням $x=\xi$ и $y=\eta$ системы (2), т. е.*

$$\lim_{n\to\infty} x_n = \xi \ \text{ и } \ \lim_{n\to\infty} y_n = \eta.$$

Замечание. Теорема остается верной, если условие 3) заменить условием 3′)

$$\left|\frac{\partial\varphi_1}{\partial x}\right| + \left|\frac{\partial\varphi_1}{\partial y}\right| \leqslant q_1 < 1,$$

$$\left|\frac{\partial\varphi_2}{\partial x}\right| + \left|\frac{\partial\varphi_2}{\partial y}\right| \leqslant q_2 < 1.$$

Примерное доказательство теоремы см. в [2]. Более общая теорема приведена в главе XIII, §§ 10—11.

Пример. Для системы [2]

$$\left.\begin{aligned} f_1(x, y) &\equiv 2x^2 - xy - 5x + 1 = 0, \\ f_2(x, y) &\equiv x + 3\lg x - y^2 = 0 \end{aligned}\right\}$$

найти положительные корни с четырьмя значащими цифрами.

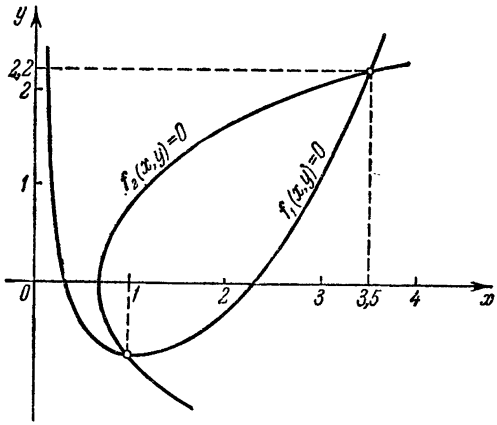


Рис. 37.

Решение. Строим графики функций $f_1(x, y) = 0$ и $f_2(x, y) = 0$ (рис. 37). Приближенные значения интересующих нас корней есть

$$x_0 = 3{,}5; \quad y_0 = 2{,}2.$$

Для применения метода итерации запишем нашу систему в таком виде:

$$x = \sqrt{\frac{x(y+5)-1}{2}} \equiv \varphi_1(x, y);$$

$$y = \sqrt{x + 3\lg x} \equiv \varphi_2(x, y).$$

Найдем частные производные

$$\frac{\partial\varphi_1}{\partial x} = \frac{y+5}{4\sqrt{\frac{x(y+5)-1}{2}}}, \quad \frac{\partial\varphi_2}{\partial x} = \frac{1+\frac{3M}{x}}{2\sqrt{x+3\lg x}},$$

где $M = 0{,}43429$,

$$\frac{\partial\varphi_1}{\partial y} = \frac{x}{4\sqrt{\frac{x(y+5)-1}{2}}}, \quad \frac{\partial\varphi_2}{\partial y} = 0.$$

Ограничиваясь окрестностью

$$R\{|x - 3{,}5| \leqslant 0{,}1; \quad |y - 2{,}2| \leqslant 0{,}1\},$$

будем иметь:

$$\left|\frac{\partial\varphi_1}{\partial x}\right| \leqslant \frac{2,3+5}{4\sqrt{\frac{3,4(2,1+5)-1}{2}}} < 0,54;$$

$$\left|\frac{\partial\varphi_1}{\partial y}\right| \leqslant \frac{3,6}{4\sqrt{\frac{3,4(2,1+5)-1}{2}}} < 0,27;$$

$$\left|\frac{\partial\varphi_2}{\partial x}\right| \leqslant \frac{1+\frac{3\cdot 0,43}{3,4}}{2\sqrt{3,4+2\lg 3,4}} < 0,42;$$

$$\left|\frac{\partial\varphi_2}{\partial y}\right| = 0.$$

Отсюда

$$\left|\frac{\partial\varphi_1}{\partial x}\right| + \left|\frac{\partial\varphi_2}{\partial x}\right| < 0,54+0,42 = 0,96 < 1; \tag{4}$$

$$\left|\frac{\partial\varphi_1}{\partial y}\right| + \left|\frac{\partial\varphi_2}{\partial y}\right| < 0,27+0 = 0,27 < 1. \tag{5}$$

Следовательно, если последовательные приближения $(x_n, y_n)$ не покинут области $R$ (что легко обнаружить в процессе вычислений), то итерационный процесс будет сходящимся.

Относительная близость суммы (4) к единице дает основания предполагать, что итерационный процесс в данном случае будет сходиться сравнительно медленно. Приступаем к вычислению последовательных приближений по формулам

$$x_{n+1} = \sqrt{\frac{x_n(y_n+5)-1}{2}};$$

$$y_{n+1} = \sqrt{x_n + 3\lg x_n} \quad (n=0,\ 1,\ 2,\ \ldots).$$

Соответствующие значения последовательных приближений помещены в таблице 5.

Таблица 5

**Значения последовательных приближений $x_n$ и $y_n$**

| $n$ | $x_n$ | $y_n$ |
|---|---|---|
| 0 | 3,5 | 2,2 |
| 1 | 3,479 | 2,259 |
| 2 | 3,481 | 2,260 |
| 3 | 3,484 | 2,261 |
| 4 | 3,486 | 2,261 |
| 5 | 3,487 | 2,262 |
| 6 | 3,487 | 2,262 |

Таким образом, можно принять $\xi = 3,487$; $\eta = 2,262$.

Замечание. Вместо рассмотренного процесса последовательных приближений (3) иногда удобнее пользоваться «*процессом Зейделя*»:

$$x_{n+1} = \varphi_1(x_n, y_n);$$

$$y_{n+1} = \varphi_2(x_{n+1}, y_n) \qquad (n=0,\ 1,\ 2,\ \ldots).$$

Метод итерации для общих систем рассмотрен в главе XIII (§§ 8—11).

## § 10. Метод Ньютона для системы двух уравнений

Пусть $x_n$, $y_n$ — приближенные корни системы уравнений

$$F(x, y)=0; \quad G(x, y)=0, \tag{1}$$

где $F$ и $G$ — непрерывно дифференцируемые функции. Полагая

$$x=x_n+h_n; \quad y=y_n+k_n,$$

получим:

$$\left.\begin{aligned} F(x_n+h_n;\ y_n+k_n)=0, \\ G(x_n+h_n;\ y_n+k_n)=0. \end{aligned}\right\} \tag{2}$$

Отсюда, применяя формулу Тейлора и ограничиваясь линейными членами относительно $h_n$ и $k_n$, будем иметь:

$$\left.\begin{aligned} F(x_n, y_n)+h_nF'_x(x_n, y_n)+k_nF'_y(x_n, y_n)=0, \\ G(x_n, y_n)+h_nG'_x(x_n, y_n)+k_nG'_y(x_n, y_n)=0. \end{aligned}\right\} \tag{3}$$

Если якобиан

$$J(x_n, y_n)=\begin{vmatrix} F'_x(x_n, y_n) & F'_y(x_n, y_n) \\ G'_x(x_n, y_n) & G'_y(x_n, y_n) \end{vmatrix} \neq 0,$$

то из системы (3) находим:

$$h_n=-\frac{1}{J(x_n, y_n)}\begin{vmatrix} F(x_n, y_n) & F'_y(x_n, y_n) \\ G(x_n, y_n) & G'_y(x_n, y_n) \end{vmatrix}, \tag{4}$$

$$k_n=-\frac{1}{J(x_n, y_n)}\begin{vmatrix} F'_x(x_n, y_n) & F(x_n, y_n) \\ G'_x(x_n, y_n) & G(x_n, y_n) \end{vmatrix}. \tag{5}$$

Следовательно, можно положить:

$$x_{n+1}=x_n-\frac{1}{J(x_n, y_n)}\begin{vmatrix} F(x_n, y_n) & F'_y(x_n, y_n) \\ G(x_n, y_n) & G'_y(x_n, y_n) \end{vmatrix}, \tag{6}$$

$$y_{n+1}=y_n-\frac{1}{J(x_n, y_n)}\begin{vmatrix} F'_x(x_n, y_n) & F(x_n, y_n) \\ G'_x(x_n, y_n) & G(x_n, y_n) \end{vmatrix} \tag{6'}$$

$$(n=0,\ 1,\ 2,\ \ldots).$$

Исходные приближения $x_0$, $y_0$ определяются грубо приближенно.

Пример. Найти вещественные корни системы

$$\begin{aligned} F(x, y)\equiv 2x^3-y^2-1=0; \\ G(x, y)\equiv xy^3-y-4=0. \end{aligned} \tag{1}$$

Решение. Графическим путем найдем грубо приближенные значения корней:

$$x_0=1{,}2; \quad y_0=1{,}7.$$

Подставив в систему (1), получим:

$$F(1{,}2;\ 1{,}7) = -0{,}434;$$
$$G(1{,}2;\ 1{,}7) = 0{,}1956.$$

Вычислим якобиан

$$J(x,\ y) = \begin{vmatrix} 6x^2 & -2y \\ y^3 & 3xy^2 - 1 \end{vmatrix};$$

отсюда

$$J = \begin{vmatrix} 8{,}64 & -3{,}40 \\ 4{,}91 & 9{,}40 \end{vmatrix} = 97{,}910.$$

По формуле (4) вычисляем $h_0$:

$$h_0 = -\frac{1}{97{,}910}\begin{vmatrix} -0{,}434 & -3{,}40 \\ 0{,}1956 & 9{,}40 \end{vmatrix} = \frac{3{,}389}{97{,}910} = 0{,}0349;$$

отсюда по формуле (6) находим:

$$x_1 = 1{,}2 + 0{,}0349 = 1{,}2349.$$

По формуле (5) вычисляем $k_0$:

$$k_0 = -\frac{1}{97{,}910}\begin{vmatrix} 8{,}64 & -0{,}434 \\ 4{,}91 & 0{,}1956 \end{vmatrix} = -0{,}0390;$$

отсюда по формуле (6) находим:

$$y_1 = 1{,}7 - 0{,}0390 = 1{,}6610.$$

Повторяя этот процесс с полученными значениями корней, получим:

$$x_2 = 1{,}2343;\quad y_2 = 1{,}6615 \text{ и т. д.}$$

Метод Ньютона для общих систем рассмотрен в главе XIII (§§ 1—7).

## § 11. Метод Ньютона для случая комплексных корней

На практике (например, при решении линейных дифференциальных уравнений) может встретиться надобность в уточнении комплексных корней данного уравнения

$$f(z) = 0. \tag{1}$$

Для этой цели иногда можно использовать метод, аналогичный методу Ньютона.

Допустим, что $f(z)$ $(z = x + iy,\ i^2 = -1)$ — аналитическая функция в некоторой выпуклой *) окрестности $U$ ее простого изолированного нуля

$$\zeta = \xi + i\eta \qquad (f(\zeta) = 0,\ f'(\zeta) \neq 0),$$

*) То есть любые две точки, принадлежащие окрестности $U$, являются концами отрезка, также принадлежащего $U$.

который, вообще говоря, является комплексным. Пусть $z_n$ — приближенное значение корня, принадлежащее окрестности $U$, и

$$z_{n+1} = z_n + \Delta z_n$$

— уточненное значение корня. Применяя разложение в ряд Тейлора в точке $z_n$ и считая, что $f(z_{n+1}) \approx 0$ с точностью до $\Delta z_n^2$, будем иметь:

$$f(z_{n+1}) \approx f(z_n) + \Delta z_n f'(z_n) = 0;$$

отсюда

$$\Delta z_n = -\frac{f(z_n)}{f'(z_n)}. \tag{2}$$

Таким образом, отправляясь от какого-нибудь значения $z_0$, шаг за шагом можно получать дальнейшие приближения корня по формуле

$$z_{n+1} = z_n - \frac{f(z_n)}{f'(z_n)} \qquad (n = 0,\ 1,\ 2,\ \ldots). \tag{3}$$

Если $z_n \in U$ $(n = 1,\ 2,\ \ldots)$ и последовательность $\{z_n\}$ сходится, то предел

$$\zeta = \lim_{n \to \infty} z_n$$

является корнем уравнения (1). Действительно, переходя к пределу при $n \to \infty$ в равенстве (3), будем иметь:

$$\lim_{n \to \infty} z_{n+1} = \lim_{n \to \infty} z_n - \frac{\lim\limits_{n \to \infty} f(z_n)}{\lim\limits_{n \to \infty} f'(z_n)}$$

или

$$\zeta = \zeta - \frac{f(\zeta)}{f'(\zeta)}.$$

Следовательно,

$$f(\zeta) = 0.$$

Для оценки погрешности приближенного значения $z_n$ предположим, что

$$|f'(z)| \geqslant m_1 > 0 \qquad \text{при } z \in U.$$

Тогда для данной функции

$$w = f(z)$$

в достаточно малой $R$-окрестности корня $\zeta$ существует однозначная обратная функция

$$z = f^{-1}(w),$$

определенная в некоторой окрестности $|w| < \rho$, производная которой, как известно, есть

$$\frac{dz}{dw} = \frac{1}{f'(z)}. \tag{4}$$

Предполагая, что $|f(z_n)| < \rho$, имеем:

$$z_n - \zeta = f^{-1}(f(z_n)) - f^{-1}(f(\zeta)) = \int_{f(\zeta)}^{f(z_n)} \frac{d}{dt}[f^{-1}(t)]\,dt = \int_0^{f(z_n)} \frac{dt}{f'(f^{-1}(t))}, \quad (5)$$

где $t$ — текущая точка, пробегающая прямолинейный отрезок между точками $f(\zeta)=0$ и $f(z_n)$ (рис. 38). Так как $|t| < \rho$, то $|f^{-1}(t)| < R$ и, следовательно,

$$|f'(f^{-1}(t))| \geqslant m_1.$$

Отсюда на основании формулы (5) будем иметь:

$$|z_n - \zeta| \leqslant \int_0^{f(z_n)} \frac{|dt|}{|f'(f^{-1}(t))|} \leqslant \frac{|f(z_n)|}{m_1}. \quad (6)$$

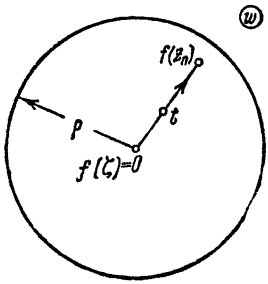

Рис. 38.

Приведем без доказательства достаточные условия существования корня уравнения (1), вытекающие из теоремы Островского [8], [9].

Теорема. *Если функция $f(z)$ аналитическая в замкнутой $R$-окрестности точки $z_0$, причем выполнены неравенства:*

1) $\left|\dfrac{1}{f'(z_0)}\right| \leqslant A_0$;

2) $\left|\dfrac{f(z_0)}{f'(z_0)}\right| \leqslant B_0 \leqslant \dfrac{R}{2}$;

3) $|f''(z)| \leqslant C$ при $|z - z_0| < R$;

4) $2A_0B_0C = \mu_0 \leqslant 1$,

*то уравнение* (1) *имеет единственный корень $\zeta$ в области $|z - z_0| \leqslant R$ и процесс Ньютона* (3), *определяемый начальным приближением $z_0$, сходится к этому корню, т. е.*

$$\zeta = \lim_{n\to\infty} z_n.$$

*Быстрота сходимости процесса характеризуется оценкой*

$$|\zeta - z_n| \leqslant B_0 \left(\frac{1}{2}\right)^{n-1} \mu_0^{2^n-1}. \quad (7)$$

Пример. Приближенно найти наименьшие по модулю корни уравнения

$$f(z) \equiv e^z - 0{,}2z + 1 = 0. \quad (8)$$

Решение. Здесь

$$f'(z) = e^z - 0{,}2.$$

Так как $f'(z) = 0$ при $\tilde{z} = \ln 0{,}2 \approx -1{,}79$ и

$$f(-\infty) = +\infty, \quad f(\tilde{z}) > 0, \quad f(+\infty) = +\infty,$$

то уравнение (8) действительных корней не имеет.

За начальное приближение искомого корня $\zeta$ примем наименьший по модулю корень $z_0$ уравнения

$$e^z + 1 = 0;$$

отсюда можно положить:

$$z_0 = \pi i.$$

Дальнейшие приближения $z_n$ ($n = 1, 2, 3, \ldots$) корня $\zeta$ последовательно определяем, применяя формулу (3):

$$z_1 = z_0 - \frac{f(z_0)}{f'(z_0)} = \pi i - \frac{0{,}2\pi i}{1{,}2} = \frac{5}{6}\pi i = 2{,}618i;$$

$$z_2 = z_1 - \frac{f(z_1)}{f'(z_1)} = \frac{5\pi i}{6} - \frac{0{,}132 - 0{,}024i}{-1{,}868 + 0{,}5i} = 0{,}069 + 2{,}624i \quad \text{и т. д.}$$

Результаты вычислений с точностью до 0,001 приведены в таблице 6.

Таблица 6

**Уточнение комплексных корней по методу Ньютона**

| $n$ | $z_n$ | $e^{z_n}$ | $f(z_n)$ | $f'(z_n)$ | $\Delta z_n = -\frac{f(z_n)}{f'(z_n)}$ |
|---|---|---|---|---|---|
| 0 | $3{,}142i$ | $-1$ | $-0{,}628i$ | $-1{,}2$ | $-0{,}524i$ |
| 1 | $2{,}618i$ | $-0{,}868+0{,}5i$ | $0{,}132-0{,}024i$ | $-1{,}068+0{,}5i$ | $0{,}153+0{,}040i$ |
| 2 | $0{,}153+2{,}658i$ | $-1{,}030+0{,}541i$ | $-0{,}061+0{,}009i$ | $-1{,}230+0{,}541i$ | $-0{,}044-0{,}012i$ |
| 3 | $0{,}109+2{,}646i$ | $-0{,}978+0{,}535i$ | $0+0{,}006i$ | $-1{,}178+0{,}535i$ | $-0{,}002+0{,}004i$ |
| 4 | $0{,}107+2{,}650i$ | $-0{,}981+0{,}525i$ | $-0{,}002-0{,}005i$ | $-1{,}181+0{,}525i$ | $-0{,}000-0{,}004i$ |
| 5 | $0{,}107+2{,}646i$ | $-0{,}977+0{,}534i$ | $+0{,}002+0{,}004i$ | $-1{,}177+0{,}534i$ | |

Для вычисления $e^z$ при $z = x + iy$ использовалась известная формула

$$e^z = e^x(\cos y + i \sin y).$$

Полагая

$$\zeta \approx z_5 = 0{,}107 + 2{,}646i,$$

будем иметь:

$$f(z_5) = 0{,}002 + 0{,}004i.$$

Приближенно считая

$$m_1 = |f'(z_5)| \approx 1{,}3,$$

на основании формулы (6) получаем погрешность

$$|\zeta - z_5| \approx \frac{|f(z_5)|}{m_1} = \frac{0{,}001 \cdot \sqrt{20}}{1{,}3} \approx 0{,}004.$$

Ввиду того, что левая часть уравнения (8) при вещественных $z$ принимает вещественные значения, то это уравнение имеет также сопряженный корень

$$\overline{\zeta} \approx 0{,}107 - 2{,}646i,$$

равный по модулю корню $\zeta$. Действительно, имеем:

$$f(\overline{\zeta}) = \overline{f(\zeta)} = 0.$$

Замечание. Другой способ решения уравнения (1) — это сведение его к системе двух действительных уравнений. Полагая

$$z = x + iy$$

в уравнении (1) и выделяя действительную и мнимую части функции $f(z)$, будем иметь:

$$f(z) \equiv u(x, y) + iv(x, y) = 0,$$

где $u$ и $v$ — действительные функции. Отсюда получаем, что уравнение (1) эквивалентно системе

$$\left.\begin{aligned} u(x, y) &= 0, \\ v(x, y) &= 0. \end{aligned}\right\} \qquad (9)$$

Уточнение корней системы вида (9) рассмотрено в §§ 9 и 10. Заметим, что этот новый способ годится также и в случае неаналитичности функции $f(z)$.

### Литература к четвертой главе


1. Я. С. Безикович, Приближенные вычисления, Гостехиздат, изд. 6, 1949, гл. VI.
2. Дж. Скарборо, Численные методы математического анализа, ГТТИ, 1934, гл. IX, X.
3. Э. Уиттекер и Г. Робинсон, Математическая обработка результатов наблюдений, ОНТИ, 1935, гл. VI.
4. Г. М. Фихтенгольц, Курс дифференциального и интегрального исчисления, т. I, Гостехиздат, 1957, гл. IV.
5. Г. П. Толстов, Курс математического анализа, т. I, Гостехиздат, 1954, гл. VII.
6. А. О. Гельфонд, Исчисление конечных разностей, Гостехиздат, 1952, гл. V.
7. Д. А. Вентцель, Е. С. Вентцель, Элементы теории приближенных вычислений, Изд. ВВИА им. Жуковского, 1949, гл. 3, § 4.
8. A. Ostrowski, Матем. сборник 2 (1937).
9. Л. В. Канторович, О методе Ньютона, Труды матем. ин-та им. В. А. Стеклова **XXVIII** (1949), 104—144.

ГЛАВА V

# СПЕЦИАЛЬНЫЕ ПРИЕМЫ ДЛЯ ПРИБЛИЖЕННОГО РЕШЕНИЯ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

## § 1. Общие свойства алгебраических уравнений

Рассмотрим *алгебраическое уравнение* $n$-й степени ($n \geqslant 1$)

$$P(x) \equiv a_0 x^n + a_1 x^{n-1} + \ldots + a_n = 0, \tag{1}$$

где коэффициенты $a_0$, $a_1$, ..., $a_n$ — действительные числа, причем

$$a_0 \neq 0.$$

В общем случае переменную $x$ будем считать комплексной.

Основная теорема алгебры. *Алгебраическое уравнение $n$-й степени* (1) (*а следовательно, и полином* $P(x)$) *имеет ровно $n$ корней, действительных или комплексных, при условии, что каждый корень считается столько раз, какова его кратность* [1], [2].

При этом говорят, что корень $\xi$ уравнения (1) имеет кратность $s$ (т. е. $\xi$ есть *$s$-кратный корень*), если

$$P(\xi) = P'(\xi) = \ldots = P^{(s-1)}(\xi) = 0,$$
$$P^{(s)}(\xi) \neq 0. \tag{2}$$

Комплексные корни уравнения (1) обладают свойством *парной сопряженности.*

Теорема 1. *Если коэффициенты алгебраического уравнения* (1) — *действительные, то комплексные корни этого уравнения попарно комплексно-сопряженные, т. е. если* $\xi = \alpha + i\beta$ ($\alpha$, $\beta$ — действительные) *есть корень уравнения* (1), *кратности $s$, то число* $\bar{\xi} = \alpha - i\beta$ *также является корнем этого уравнения и имеет ту же кратность $s$.*

Отметим, что модули этих корней одинаковы:

$$|\xi| = |\bar{\xi}| = \sqrt{\alpha^2 + \beta^2}.$$

Следствие. *Алгебраическое уравнение нечетной степени с действительными коэффициентами имеет по меньшей мере один действительный корень.*

Нетрудно дать грубую оценку модулей корней уравнения (1).

Теорема 2. *Пусть*

$$A = \max\{|a_1|, |a_2|, \ldots, |a_n|\},$$

*где $a_k$ — коэффициенты уравнения* (1).

*Тогда модули всех корней $x_k$ $(k = 1, \ldots, n)$ уравнения* (1) *удовлетворяют неравенству*

$$|x_k| < 1 + \frac{A}{|a_0|}, \tag{3}$$

*т. е. корни этого уравнения на комплексной плоскости $\xi O \eta$ $(x = \xi + i\eta)$ расположены внутри круга*

$$|x| < 1 + \frac{A}{|a_0|} = R$$

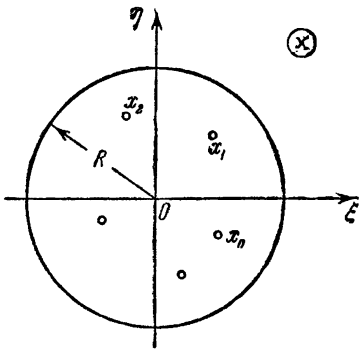


Рис. 39.

(рис. 39).

Доказательство. Полагая $|x| > 1$, из формулы (1) имеем:

$$|P(x)| \geqslant |a_0 x^n| - (|a_1 x^{n-1}| + |a_2 x^{n-2}| + \ldots + |a_n|) \geqslant$$
$$\geqslant |a_0||x|^n - A(|x|^{n-1} + |x|^{n-2} + \ldots + 1) =$$
$$= |a_0||x|^n - A\frac{|x|^n - 1}{|x| - 1} > \left(|a_0| - \frac{A}{|x| - 1}\right)|x|^n.$$

Отсюда, если

$$|a_0| - \frac{A}{|x| - 1} \geqslant 0,$$

т. е. если

$$|x| \geqslant 1 + \frac{A}{|a_0|}, \tag{4}$$

получаем, что

$$|P(x)| > 0.$$

Таким образом, значения $x$, удовлетворяющие неравенству (4), заведомо не являются корнями уравнения (1). Следовательно, все корни $x_k$ уравнения (1) удовлетворяют противоположному неравенству

$$|x_k| < 1 + \frac{A}{|a_0|}.$$

Следствие. *Пусть $a_n \neq 0$ и*

$$B = \max\{|a_0|, |a_1|, \ldots, |a_{n-1}|\}.$$

*Тогда все корни* $x_k$ $(k=1, 2, \ldots, n)$ *уравнения* (1) *удовлетворяют неравенству*

$$|x_k| > \frac{1}{1+\frac{B}{|a_n|}} = r, \tag{5}$$

*т. е. корни уравнения* (1) *расположены в круговом кольце*

$$r < |x| < R$$

(рис. 40).

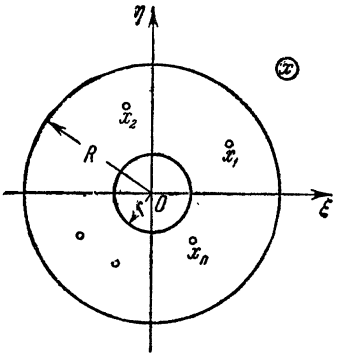


Рис. 40.

В самом деле, полагая

$$x = \frac{1}{y},$$

будем иметь:

$$P(x) = \frac{1}{y^n} Q(y),$$

где

$$Q(y) = a_n y^n + a_{n-1} y^{n-1} + \ldots + a_0.$$

Корни $y_k = \frac{1}{x_k}$ $(k=1, \ldots, n)$ полинома $Q(y)$ в силу нашей теоремы удовлетворяют неравенству

$$|y_k| = \frac{1}{|x_k|} < 1 + \frac{B}{|a_n|},$$

откуда

$$|x_k| > \frac{1}{1+\frac{B}{|a_n|}} = r \qquad (k=1, \ldots, n).$$

Замечание. Числа $r$ и $R$ являются соответственно *нижней* и *верхней* границами положительных корней уравнения (1).

Аналогично числа $-R$ и $-r$ служат соответственно нижней и верхней границами отрицательных корней уравнения (1).

Если

$$x_1, x_2, \ldots, x_n$$

— корни уравнения (1), то для левой части его справедливо разложение

$$P(x) = a_0 (x - x_1)(x - x_2) \ldots (x - x_n). \tag{6}$$

Отсюда, производя перемножение биномов в формуле (6) и приравнивая коэффициенты при одинаковых степенях $x$ в левой и

правой частях равенства (6), получим соотношения между корнями и коэффициентами алгебраического уравнения:

$$\left.\begin{aligned} &x_1+x_2+\ldots+x_n=-\frac{a_1}{a_0}, \\ &x_1x_2+x_1x_3+\ldots+x_{n-1}x_n=\frac{a_2}{a_0}, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ &x_1x_2\ldots x_n=(-1)^n\frac{a_n}{a_0}. \end{aligned}\right\} \tag{7}$$

Левые части равенств (7) представляют собой суммы сочетаний корней уравнения (1) по одному, по два и т. д.

Пример 1. Корни $x_1$, $x_2$, $x_3$ кубического уравнения

$$x^3+px^2+qx+r=0$$

удовлетворяют условиям:

$$\left.\begin{aligned} &x_1+x_2+x_3=-p, \\ &x_1x_2+x_1x_3+x_2x_3=q, \\ &x_1x_2x_3=-r. \end{aligned}\right\}$$

Если учитывать кратности корней, то разложение (6) принимает вид

$$P(x)=a_0(x-x_1)^{\alpha_1}(x-x_2)^{\alpha_2}\ldots(x-x_m)^{\alpha_m},$$

где $x_1, x_2, \ldots, x_m$ $(m\leqslant n)$ — различные корни уравнения (1) и $\alpha_1, \alpha_2, \ldots, \alpha_m$ — их кратности, причем

$$\alpha_1+\alpha_2+\ldots+\alpha_m=n.$$

Производная $P'(x)$ выражается следующим образом:

$$P'(x)=a_0(x-x_1)^{\alpha_1-1}(x-x_2)^{\alpha_2-1}\ldots(x-x_m)^{\alpha_m-1}Q(x),$$

где $Q(x)$ — полином такой, что

$$Q(x_k)\neq 0 \quad \text{при} \quad k=1,\ 2,\ \ldots,\ m.$$

Поэтому полином

$$R(x)=a_0(x-x_1)^{\alpha_1-1}(x-x_2)^{\alpha_2-1}\ldots(x-x_m)^{\alpha_m-1}$$

является наибольшим общим делителем полинома $P(x)$ и его производной $P'(x)$. Как известно, полином $R(x)$ может быть найден с помощью алгорифма Евклида [1]. Составляя частное

$$f(x)=\frac{P(x)}{R(x)},$$

получим полином

$$f(x)=A_0x^m+A_1x^{m-1}+\ldots+A_m \tag{8}$$

с действительными коэффициентами $A_0 = a_0$, $A_1$, ..., $A_m$, корни которого $x_1$, $x_2$, ..., $x_m$ различны.

Таким образом, решение алгебраического уравнения с кратными корнями сводится к решению алгебраического уравнения более низкой степени с различными корнями.

Полное число корней $x_1$, $x_2$, ..., $x_N$ уравнения

$$P(x) = 0,$$

расположенных на комплексной плоскости внутри простого замкнутого контура $\Gamma$ (рис. 41), можно определить на основании *принципа аргумента* [4], который состоит в следующем: *если полином $P(x)$ не имеет корней на замкнутом контуре $\Gamma$, то число корней $N$ этого полинома внутри контура $\Gamma$ в точности равно изменению* $\operatorname{Arg} P(x)$ *при положительном обходе контура $\Gamma$, деленному на $2\pi$*, т. е.

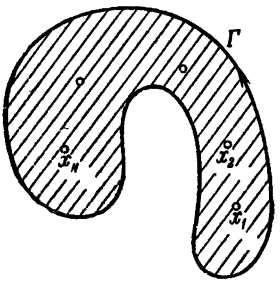

Рис. 41.

$$N = \frac{1}{2\pi} \Delta_\Gamma \operatorname{Arg} P(x),$$

причем каждый корень считается столько раз, какова его кратность.

Если уравнение контура $\Gamma$ есть

$$x = \xi(t) + i\eta(t) \qquad (0 \leqslant t \leqslant T)$$

($t$ — параметр), то для определения числа $N$ на плоскости $XOY$ строят кривую

$$X = X(t), \quad Y = Y(t) \qquad (0 \leqslant t \leqslant T), \tag{K}$$

где

$$P(x) = P(\xi(t) + i\eta(t)) = X(t) + iY(t)$$

($X(t)$, $Y(t)$ — действительные функции), и подсчитывают, сколько оборотов $N$ кривая $K$ делает вокруг начала координат.

Пример 2. Определить число корней уравнения

$$P(x) \equiv x^3 - 3x + 1 = 0, \tag{9}$$

содержащихся внутри круга $|x| < 2$.

Решение. Полагая

$$x = 2(\cos t + i \sin t),$$

будем иметь:

$$P(x) = 8(\cos t + i \sin t)^3 - 6(\cos t + i \sin t) + 1 =$$
$$= (8 \cos 3t - 6 \cos t + 1) + i(8 \sin 3t - 6 \sin t).$$

Отсюда

$$\left.\begin{aligned} X &= 8 \cos 3t - 6 \cos t + 1, \\ Y &= 8 \sin 3t - 6 \sin t. \end{aligned}\right\} \tag{K}$$

Таблица 7

| $t$ | 0 | $\pm\frac{\pi}{6}$ | $\pm\frac{\pi}{3}$ | $\pm\frac{\pi}{2}$ | $\pm\frac{2\pi}{3}$ | $\pm\frac{5\pi}{6}$ | $\pm\pi$ |
|---|---|---|---|---|---|---|---|
| $X$ | 3 | —4,22 | —10 | 1 | 15 | 6,22 | —1 |
| $Y$ | 0 | $\pm 5$ | $\pm 5{,}22$ | $\mp 14$ | $\mp 5{,}22$ | $\pm 5$ | 0 |

Построив по точкам кривую $K$ (см. таблицу 7), легко убедиться, что кривая три раза окружает начало координат (рис. 42). Поэтому $N=3$ и, следовательно, уравнение (9) имеет внутри круга $|x|<2$ три корня.

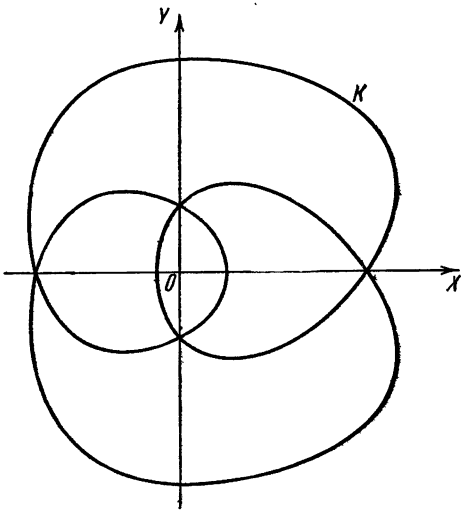


Рис. 42.

## § 2. Границы действительных корней алгебраических уравнений

В этом параграфе мы будем рассматривать полиномы вида

$$P(x)=a_0x^n+a_1x^{n-1}+\ldots+a_n \qquad (1)$$

с действительными коэффициентами $a_0, a_1, \ldots, a_n$, где $a_0\neq 0$. Нашей целью является установление границ, по возможности тесных для положительных и отрицательных корней $x_1, x_2, \ldots, x_m$ $(1\leqslant m\leqslant n)$ уравнения

$$P(x)=0, \qquad (2)$$

причем вопрос о существовании этих корней здесь не затрагивается. Заметим, что можно ограничиться нахождением верхней границы $R$ лишь положительных корней уравнений вида (2). В самом деле, наряду с уравнением (2) рассмотрим вспомогательные алгебраические уравнения

$$P_1(x)\equiv x^nP\left(\frac{1}{x}\right)=0,$$
$$P_2(x)\equiv P(-x)=0,$$
$$P_3(x)\equiv x^nP\left(-\frac{1}{x}\right)=0,$$

и пусть верхние границы их положительных корней соответственно есть $R_1$, $R_2$ и $R_3$. Тогда число $\frac{1}{R_1}$, очевидно, есть нижняя граница положительных корней уравнения (2), т. е. все положительные корни $x^+$ этого уравнения, если они существуют, удовлетворяют неравенству

$$\frac{1}{R_1} \leqslant x^+ \leqslant R.$$

Аналогично числа $-R_2$ и $-\frac{1}{R_3}$ являются соответственно нижней и верхней границами отрицательных корней уравнения (2), т. е. все отрицательные корни $x^-$ этого уравнения, если таковые имеются, удовлетворяют неравенству

$$-R_2 \leqslant x^- \leqslant -\frac{1}{R_3}.$$

Укажем некоторые простые приемы нахождения верхней границы $R$ положительных корней уравнения (2), причем некоторые из них приведем без доказательства.

Теорема Лагранжа. *Пусть $a_0 > 0$ и $a_k (k \geqslant 1)$ — первый из отрицательных коэффициентов*) полинома $P(x)$. Тогда за верхнюю границу положительных корней уравнения* (2) *может быть принято число*

$$R = 1 + \sqrt[k]{\frac{B}{a_0}}, \tag{3}$$

*где $B$ — наибольшая из абсолютных величин отрицательных коэффициентов полинома $P(x)$.*

Доказательство. Положим $x > 1$. Если в полиноме $P(x)$ каждый из неотрицательных коэффициентов $a_1, \ldots, a_{k-1}$ заменить нулем, а каждый из остальных коэффициентов $a_k, a_{k+1}, \ldots, a_n$ заменить отрицательным числом $-B$, то от этого полином (1) может лишь уменьшить свое значение и мы будем иметь неравенство

$$P(x) \geqslant a_0 x^n - B(x^{n-k} + x^{n-k-1} + \ldots + 1) = \\ = a_0 x^n - B\frac{x^{n-k+1}-1}{x-1}.$$

Отсюда при $x > 1$ получим:

$$P(x) > a_0 x^n - \frac{B}{x-1} x^{n-k+1} = \frac{x^{n-k+1}}{x-1}[a_0 x^{k-1}(x-1) - B] > \\ > \frac{x^{n-k+1}}{x-1}[a_0 (x-1)^k - B].$$

---

*) Если такого коэффициента нет, т. е. все коэффициенты полинома $P(x)$ неотрицательны, то полином $P(x)$ не имеет положительных корней.

Следовательно, при

$$x \geqslant 1 + \sqrt[k]{\frac{B}{a_0}} = R$$

будем иметь:

$$P(x) > 0,$$

т. е. все положительные корни $x^+$ уравнения (2) удовлетворяют неравенству

$$x^+ < R.$$

## § 3. Метод знакопеременных сумм

Идея метода Лагранжа может быть обобщена следующим образом: пусть полином $P(x)$ расположен по убывающим степеням переменной $x$, причем его старший коэффициент $a_0 > 0$. Представим $P(x)$ в виде знакопеременной суммы

$$P(x) = Q_1(x) - Q_2(x) + Q_3(x) - Q_4(x) + \ldots + Q_{2m-1}(x) - Q_{2m}(x),$$

где $Q_1(x)$ — сумма последовательных членов полинома $P(x)$ с положительными коэффициентами, начиная с $a_0x^n$, $-Q_2(x)$ — сумма последовательных членов полинома $P(x)$ с отрицательными коэффициентами, непосредственно примыкающих к членам первой суммы, и т. д., причем последнее слагаемое $-Q_{2m}(x)$ или состоит из членов с отрицательными коэффициентами, или тождественно равно нулю.

Обозначим через $c_j$ $(j = 1, 2, \ldots, m)$ положительные числа такие, что

$$Q_{2j-1}(c_j) - Q_{2j}(c_j) \geqslant 0 \tag{1}$$

$(j = 1, 2, \ldots, m)$. Тогда за верхнюю границу положительных корней уравнения (2) § 2 можно принять число

$$R = \max(c_1, c_2, \ldots, c_m). \tag{2}$$

В самом деле, положим:

$$Q_{2j-1}(x) - Q_{2j}(x) = b_1^{(j)}x^{n_j} + b_2^{(j)}x^{n_j-1} + \ldots + b_p^{(j)}x^{n_j-p+1} - \\ - b_{p+1}^{(j)}x^{n_j-p} - b_{p+2}^{(j)}x^{n_j-p-1} - \ldots - b_{p+q}^{(j)}x^{n_j-p-q+1},$$

где

$$b_s^{(j)} \geqslant 0 \qquad (s = 1, 2, \ldots, p+q),$$

причем $b_1^{(j)} > 0$ $(j = 1, 2, \ldots, m)$.

Полагая $x > 0$, имеем:

$$Q_{2j-1}(x) - Q_{2j}(x) = x^{n_j-p+1}\left[(b_1^{(j)}x^{p-1} + b_2^{(j)}x^{p-2} + \ldots + b_p^{(j)}) - \right.$$
$$\left. - \left(\frac{b_{p+1}^{(j)}}{x} + \frac{b_{p+2}^{(j)}}{x^2} + \ldots + \frac{b_{p+q}^{(j)}}{x^q}\right)\right]. \quad (3)$$

Из формулы (3) ясно, что функции $Q_{2j-1}(x) - Q_{2j}(x)$ $(j = 1, 2, \ldots, m)$ возрастают при возрастании $x$. Следовательно, при $x > c_j > 0$ имеем:

$$Q_{2j-1}(x) - Q_{2j}(x) > Q_{2j-1}(c_j) - Q_{2j}(c_j) \geqslant 0.$$

Отсюда при $x > R$ получаем:

$$P(x) = \sum_{j=1}^{m} [Q_{2j-1}(x) - Q_{2j}(x)] > 0,$$

т. е. все положительные корни $x^+$ уравнения (2) § 2 удовлетворяют условию

$$x^+ \leqslant R.$$

Пример. Определить границы действительных корней уравнения

$$2x^5 - 100x^2 + 2x - 1 = 0. \quad (4)$$

Решение. Здесь $a_0 = 2$ и $A = \max(100, 2, 1) = 100$. Поэтому верхняя граница $R$ положительных корней уравнения (4), согласно теореме 2 из § 1, есть

$$R = 1 + \frac{A}{a_0} = 1 + \frac{100}{2} = 51.$$

Применяя теорему Лагранжа, учитывая, что

$$a_k = a_3 = -100 \quad \text{и} \quad B = \max(100, 1) = 100,$$

будем иметь значительно лучшую оценку для верхней границы положительных корней

$$R = 1 + \sqrt[3]{\frac{100}{2}} = 1 + \sqrt[3]{50} \approx 4{,}7.$$

Наконец, применяя метод знакопеременных сумм, находим:

$$2x^5 - 100x^2 = 2x^2(x^3 - 50) > 0$$

при $x > \sqrt[3]{50}$ (например, при $x > 3{,}7$) и

$$2x - 1 = 2\left(x - \frac{1}{2}\right) > 0 \quad \text{при} \quad x > 0{,}5.$$

Следовательно, можно принять

$$R = \max(3{,}7;\ 0{,}5) = 3{,}7.$$

Для определения нижней границы $r$ положительных корней уравнения (4) положим:

$$x = \frac{1}{y}.$$

Тогда уравнение (4) примет вид

$$y^5 - 2y^4 + 100y^3 - 2 = 0.$$

Последовательно получаем:

$$y^5 - 2y^4 = y^4(y-2) > 0 \quad \text{при} \quad y > 2$$

и

$$100y^3 - 2 = 100(y^3 - 0{,}02) > 0 \quad \text{при} \quad y > 0{,}3.$$

Следовательно,

$$R_1 = \max(2;\ 0{,}3) = 2$$

и

$$r = \frac{1}{R_1} = 0{,}5.$$

Для нахождения границы отрицательных корней в уравнении (4) положим:

$$x = -z.$$

Отсюда

$$2z^5 + 10z^2 + 2z + 1 = 0. \tag{4'}$$

Так как коэффициенты уравнения (4′) положительны или равны нулю, то это уравнение не имеет положительных корней, а следовательно, данное уравнение (4) не имеет отрицательных корней.

## § 4. Метод Ньютона

Теорема Ньютона. *Если при $x = c > 0$ полином $P(x)$ и все его производные $P'(x)$, $P''(x)$, ..., $P^{(n)}(x)$ неотрицательны:*

$$P^{(k)}(c) \geqslant 0 \qquad (k = 0,\ 1,\ 2,\ \ldots,\ n), \tag{1}$$

*причем $P^{(n)}(c) = n!\,a_0 > 0$, то $R = c$ может быть принято за верхнюю границу положительных корней уравнения*

$$P(x) = 0. \tag{2}$$

Доказательство. При $x > c$, учитывая неравенства (1), на основании формулы Тейлора имеем:

$$P(x) = P(c) + P'(c)(x-c) + \ldots + \frac{P^{(n)}(c)}{n!}(x-c)^n > 0.$$

Следовательно, все положительные корни $x^+$ уравнения (2) удовлетворяют неравенству

$$x^+ \leqslant c.$$

Замечание. При практическом применении теоремы Ньютона методом проб (используя, например, схему Горнера) отыскивают монотонно возрастающую последовательность положительных чисел

$$0 < c_1 \leqslant c_2 \leqslant \ldots \leqslant c_{n-1} \leqslant c_n,$$

для которых справедливы неравенства

$$\begin{aligned} P^{(n-1)}(c_1) &\geqslant 0, \\ P^{(n-2)}(c_2) &\geqslant 0, \\ &\ldots\ldots \\ P'(c_{n-1}) &\geqslant 0, \\ P(c_n) &\geqslant 0. \end{aligned}$$

Такие числа заведомо существуют, так как для $a_0 > 0$ имеем:

$$P^{(m)}(x) \longrightarrow +\infty \qquad (m = 0,\ 1,\ 2,\ \ldots,\ n-1)$$

при $x \longrightarrow +\infty$. Окончательно можно принять $c = c_n$.

Действительно, так как

$$P^{(n)}(x) = n!\, a_0 > 0,$$

то функция $P^{(n-1)}(x)$ — возрастающая и, следовательно, при $x > c_1$ мы будем иметь:

$$P^{(n-1)}(x) > P^{(n-1)}(c_1) \geqslant 0.$$

Из последнего неравенства вытекает, что функция $P^{(n-2)}(x)$ — возрастающая в промежутке $[c_1,\ +\infty)$, и поэтому при $x > c_2 \geqslant c_1$ получаем:

$$P^{(n-2)}(x) > P^{(n-2)}(c_2) \geqslant 0.$$

Проводя последовательно это рассуждение, мы, наконец, убедимся, что $P(x)$ — возрастающая функция в промежутке $[c_{n-1},\ +\infty)$ и, следовательно, при $x > c_n \geqslant c_{n-1}$ имеем:

$$P(x) > P(c_n) \geqslant 0.$$

Значит, $x^+ \leqslant c_n$.

Пример. Рассмотрим приведенное в примере § 3 уравнение

$$P(x) = 2x^5 - 100x^2 + 2x - 1 = 0.$$

Здесь

$$\begin{aligned} P'(x) &= 10x^4 - 200x + 2, \\ P''(x) &= 40x^3 - 200, \\ P'''(x) &= 120x^2, \\ P^{\mathrm{IV}}(x) &= 240x, \\ P^{\mathrm{V}}(x) &= 240. \end{aligned}$$

Очевидно, $P'''(x) > 0$, $P^{IV}(x) > 0$, $P^{V}(x) > 0$ при $x > 0$. Имеем:

$$P''(x) = 40(x^3 - 5) > 0 \text{ при } x \geqslant 2.$$

Примем $c_1 = c_2 = c_3 = 2$. Так как

$$P'(2) = 10 \cdot 16 - 200 \cdot 2 + 2 < 0,$$

то определяем знак числа

$$P'(3) = 10 \cdot 81 - 200 \cdot 3 + 2 > 0.$$

Можно принять $c_4 = 3$. Далее, имеем:

$$P(3) = 2 \cdot 243 - 100 \cdot 9 + 2 \cdot 3 - 1 < 0;$$

поэтому вычисляем:

$$P(4) = 2 \cdot 1024 - 100 \cdot 16 + 2 \cdot 4 - 1 > 0.$$

Значит, $c_5 = 4$. Итак, верхняя граница положительных корней данного уравнения есть

$$R = 4.$$

Оценка по методу Ньютона получилась точнее, чем приведенная выше оценка по методу Лагранжа, но менее точная, чем оценка по способу знакопеременных сумм (см. пример § 3).

## § 5. Число действительных корней полинома

После того как установлены границы положительных и отрицательных корней алгебраического уравнения

$$P(x) = 0, \tag{1}$$

где $P(x)$ — данный полином, возникает вопрос о числе действительных корней данного уравнения на некотором известном интервале $(a, b)$.

Общую ориентировку о числе действительных корней уравнения (1) на интервале $(a, b)$ дает график функции $y = P(x)$ (рис. 43), где корнями $x_1, x_2, x_3$ являются абсциссы точек пересечения графика с осью $Ox$.

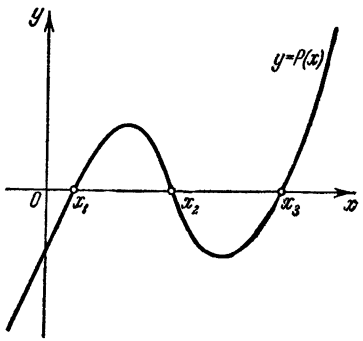

Рис. 43.

Отметим простые особенности целого полинома.

1) Если $P(a)P(b) < 0$, то на интервале $(a, b)$ имеется нечетное число корней полинома $P(x)$ с учетом их кратностей.

2) Если $P(a)P(b) > 0$, то на интервале $(a, b)$ или не имеется корней полинома $P(x)$, или таких корней существует четное число.

Полностью вопрос о числе действительных корней алгебраического уравнения на данном промежутке решается *методом Штурма* [1], [2].

Предварительно введем понятие о числе перемен знаков в числовой системе.

Определение. *Пусть дана упорядоченная конечная система действительных чисел, отличных от нуля:*

$$c_1,\ c_2,\ \ldots,\ c_n \qquad (n \geqslant 2). \tag{2}$$

*Говорят, что для пары рядом стоящих элементов $c_k$, $c_{k+1}$ системы* (2) *имеется изменение знака, если эти элементы обладают противоположными знаками, т. е.*

$$c_k c_{k+1} < 0,$$

*и нет изменения знака, если знаки их одинаковы, т. е.*

$$c_k c_{k+1} > 0.$$

Общее число изменений знаков всех пар соседних элементов $c_k$, $c_{k+1}$ ($k=1,\ 2,\ \ldots,\ n-1$) системы (2) называется *числом перемен знаков* в системе (2).

Для данного полинома $P(x)$ составим *систему Штурма*

$$P(x),\ P_1(x),\ \ldots,\ P_2(x),\ \ldots,\ P_m(x), \tag{3}$$

где $P_1(x)=P'(x)$, $P_2(x)$ — взятый с обратным знаком остаток при делении полинома $P(x)$ на $P_1(x)$, $P_3(x)$ — взятый с обратным знаком остаток при делении полинома $P_1(x)$ на $P_2(x)$ и т. д. Полиномы $P_k(x)$ ($k=2,\ \ldots,\ m$) могут быть найдены с помощью несущественно видоизмененного алгорифма Евклида, причем, если полином $P(x)$ не имеет кратных корней, то последний элемент $P_m(x)$ системы Штурма есть отличное от нуля действительное число. Заметим, что элементы системы Штурма можно вычислять с точностью до положительного числового множителя.

Обозначим через $N(c)$ число перемен знаков в системе Штурма при $x=c$, при условии, что нулевые элементы этой системы вычеркнуты.

Теорема Штурма. *Если полином $P(x)$ не имеет кратных корней и $P(a)\neq 0$, $P(b)\neq 0$, то число его действительных корней $N(a,\ b)$ на интервале $a<x<b$ в точности равно числу потерянных перемен знаков в системе Штурма полинома $P(x)$ при переходе от $x=a$ до $x=b$, т. е.*

$$N(a,\ b)=N(a)-N(b). \tag{4}$$

Следствие 1. Если $P(0)\neq 0$, то число $N_+$ положительных и число $N_-$ отрицательных корней полинома $P(x)$ соответственно равны

$$N_+=N(0)-N(+\infty)$$

и

$$N_-=N(-\infty)-N(0).$$

Следствие 2. Для того чтобы все корни полинома $P(x)$ степени $n$, не имеющего кратных корней, были действительны, необходимо и достаточно, чтобы выполнялось условие

$$N(-\infty)-N(+\infty)=n.$$

Таким образом, если

$$P(x)=a_0x^n+a_1x^{n-1}+\ldots+a_n,$$

где $a_0>0$, то все корни уравнения $P(x)=0$ будут действительны тогда и только тогда [1], когда: 1) система Штурма имеет максимальное число элементов $n+1$, т. е. $m=n$, и 2) выполнены неравенства $P_k(+\infty)>0$ $(k=1, 2, \ldots, n)$, т. е. старшие коэффициенты всех функций Штурма $P_k(x)$ должны быть положительны.

Пример. Определить число положительных и число отрицательных корней уравнения

$$x^4-4x+1=0. \tag{5}$$

Решение. Система Штурма имеет вид

$$\begin{aligned} P(x)&=x^4-4x+1,\\ P_1(x)&=x^3-1,\\ P_2(x)&=3x-1,\\ P_3(x)&=1; \end{aligned}$$

отсюда

$$N(-\infty)=2,\quad N(0)=2,\quad N(+\infty)=0.$$

Следовательно, уравнение (5) имеет:

$$N_+=2-0=2$$

положительных корней и

$$N_-=2-2=0$$

отрицательных корней. Поэтому два корня уравнения (5) — комплексные.

С помощью системы Штурма можно отделять корни алгебраических уравнений, разбивая интервал $(a, b)$, содержащий все действительные корни уравнения, на конечное число частичных интервалов $(\alpha, \beta)$ таких, что

$$N(\alpha)-N(\beta)=1.$$

## § 6. Теорема Бюдана — Фурье

Так как построение системы Штурма, вообще говоря, требует громоздких вычислений, то на практике ограничиваются более простыми частными приемами подсчета числа действительных корней алгебраических уравнений.

Уточним подсчет числа перемен знаков в числовой системе.

Определение. *Пусть дана конечная упорядоченная система действительных чисел*

$$c_1,\ c_2,\ \ldots,\ c_n, \tag{1}$$

*где* $c_1 \neq 0$ *и* $c_n \neq 0$.

*С одной стороны, назовем нижним числом перемен знаков* $\underline{N}$ *системы* (1) *число перемен знаков в ее соответствующей подсистеме, не содержащей нулевых элементов.*

*С другой стороны, назовем верхним числом перемен знаков* $\overline{N}$ *системы* (1) *число перемен знаков в преобразованной системе* (1), *где нулевые элементы*

$$c_k = c_{k+1} = \ldots = c_{k+l-1} = 0$$

($c_{k-1} \neq 0$, $c_{k+l} \neq 0$) *заменены элементами* $\tilde{c}_{k+i}$ ($i = 0, 1, 2, \ldots, l-1$) *такими, что*

$$\operatorname{sgn} \tilde{c}_{k+i} = (-1)^{l-i} \operatorname{sgn} c_{k+l}. \tag{2}$$

Очевидно, что если система (1) не имеет нулевых элементов, то число $N$ перемен знаков в этой системе по смыслу совпадает с ее нижним $\underline{N}$ и верхним $\overline{N}$ числами перемен знаков:

$$N = \underline{N} = \overline{N};$$

вообще же говоря, $\overline{N} \geqslant \underline{N}$.

Пример 1. Определить нижнее число и верхнее число перемен знаков в системе

$$1,\ 0,\ 0,\ -3,\ 1.$$

Решение. Игнорируя нули, получаем:

$$\underline{N} = 2.$$

Для подсчета $\overline{N}$, согласно формуле (2), составляем систему

$$1,\ -\varepsilon,\ \varepsilon,\ -3,\ 1,$$

где $\varepsilon > 0$. Отсюда

$$\overline{N} = 4.$$

Теорема Бюдана—Фурье. *Если числа* $a$ *и* $b$ ($a < b$) *не являются корнями полинома* $P(x)$ *степени* $n$, *то число* $N(a, b)$ *действительных корней уравнения*

$$P(x) = 0, \tag{3}$$

*содержащихся между* $a$ и $b$, *равно минимальному числу* $\Delta N$ *перемен знаков, потерянных в системе последовательных*

*производных*

$$P(x),\ P'(x),\ \ldots,\ P^{(n-1)}(x),\ P^{(n)}(x) \tag{4}$$

*при переходе от* $x=a$ *к* $x=b$, *или меньше числа* $\Delta N$ *на четное число, т. е.*

$$N(a,\ b)=\Delta N-2k,$$

где

$$\Delta N=\underline{N}(a)-\overline{N}(b)$$

и $\underline{N}(a)$ — нижнее число перемен знаков в системе (4) при $x=a$, $\overline{N}(b)$ — верхнее число перемен знаков в этой системе при $x=b$ $\left(k=0,\ 1,\ \ldots,\ E\left(\frac{\Delta N}{2}\right)\right)$ (см. [1]).

Предполагается, что каждый корень уравнения (3) считается столько раз, какова его кратность. Если производные $P^{(k)}(x)$ $(k=1,\ 2,\ \ldots,\ n)$ не обращаются в нуль при $x=a$ и $x=b$, то подсчет знаков упрощается, а именно:

$$\Delta N=N(a)-N(b).$$

Следствие 1. Если $\Delta N=0$, то между $a$ и $b$ нет действительных корней уравнения (3).

Следствие 2. Если $\Delta N=1$, то между $a$ и $b$ имеется ровно один действительный корень уравнения (3).

Замечание. Для подсчета числа потерянных знаков $\Delta N$ в системе (4), пользуясь схемой Горнера, составляем два разложения:

$$P(a+h)=\alpha_0+\alpha_1 h+\alpha_2 h^2+\ldots+\alpha_n h^n \tag{5}$$

и

$$P(b+h)=\beta_0+\beta_1 h+\beta_2 h^2+\ldots+\beta_n h^n. \tag{6}$$

Пусть $\underline{N}(a)$ — нижнее число перемен знаков коэффициентов разложения (5) и соответственно $\overline{N}(b)$ — верхнее число перемен знаков коэффициентов разложения (6). Так как

$$\alpha_k=\frac{P^{(k)}(a)}{k!},\quad \beta_k=\frac{P^{(k)}(b)}{k!}\quad (k=0,\ 1,\ 2,\ \ldots,\ n),$$

то знаки чисел $\alpha_k$ и $\beta_k$ совпадают со знаками системы (4) при $x=a$ и $x=b$. Поэтому

$$\Delta N=\underline{N}(a)-\overline{N}(b).$$

Пример 2. Определить число действительных корней уравнения

$$P(x)\equiv x^3-x^2+2x-3=0 \tag{7}$$

в интервале (0, 2).

Решение. Здесь $N(0)$, очевидно, есть число перемен знаков в системе чисел

$$-3,\ 2,\ -1,\ 1,$$

т. е.

$$N(0)=3.$$

Разложение $P(2+h)$ получается с помощью применения схемы Горнера

| | | | | |
|---|---|---|---|---|
| 1 | −1 | 2 | +3 | \| 2 |
| | 2 | 2 | 8 | |
| 1 | 1 | 4 | $\boxed{5}$ | |
| | 2 | 6 | | |
| 1 | 3 | $\boxed{10}$ | | |
| | 2 | | | |
| 1 | $\boxed{5}$ | | | |
| $\boxed{1}$ | | | | |

Следовательно, $N(2)$ есть число перемен знаков в системе чисел

$$5,\quad 10,\quad 5,\quad 1,$$

т. е. $N(2)=0$.

Отсюда

$$\Delta N=N(0)-N(2)=3.$$

Таким образом, уравнение (7) имеет в интервале (0, 2) три или один действительный корень.

Теорема Декарта. *Число положительных корней алгебраического уравнения*

$$P(x)\equiv a_0x^n+a_1x^{n-1}+\ldots+a_n=0 \qquad (a_0\neq 0) \tag{8}$$

*с учетом их кратностей равно числу перемен знаков в системе коэффициентов*

$$a_0,\ a_1,\ a_2,\ \ldots,\ a_n \tag{9}$$

(где коэффициенты, равные нулю, не учитываются), *или меньше этого числа на четное число.*

Теорема Декарта представляет собой применение теоремы Бюдана — Фурье к интервалу $(0,\ +\infty)$. В самом деле, так как

$$P^{(k)}(0)=k!\,a_{n-k} \qquad (k=0,\ 1,\ \ldots,\ n),$$

то система (9), с точностью до положительных множителей, есть

совокупность производных $P^{(k)}(0)$ $(k=0, 1, 2, \ldots, n)$, записанная так, что порядки их убывают. Поэтому число перемен знаков в системе (9) равно $\underline{N}(0)$, причем коэффициенты, равные нулю, не учитываются. С другой стороны, производные $P^{(k)}(+\infty)$ $(k=0, 1, 2, \ldots, n)$, очевидно, имеют один и тот же знак и, следовательно, $\overline{N}(+\infty)=0$. Поэтому имеем:

$$\Delta N=\underline{N}(0)-\overline{N}(+\infty)=\underline{N}(0),$$

причем на основании теоремы Бюдана — Фурье число положительных корней уравнения (8) или равно $\Delta N$, или меньше $\Delta N$ на четное число.

Следствие. Если коэффициенты уравнения (8) отличны от нуля, то число отрицательных корней уравнения (8), с учетом их кратностей, равно числу постоянств знака в системе (9) его коэффициентов или меньше этого числа на четное число.

Доказательство этого утверждения непосредственно следует из применения теоремы Декарта к полиному $P(-x)$.

Укажем еще необходимый признак вещественности всех корней полинома.

Теорема Гюа. *Если уравнение*

$$a_0x^n+a_1x^{n-1}+a_2x^{n-2}+\ldots+a_n=0 \tag{10}$$

*имеет действительные коэффициенты и все корни его действительны, то квадрат каждого некрайнего коэффициента этого уравнения больше произведения двух его соседних коэффициентов, т. е. выполнены неравенства*

$$a_k^2>a_{k-1}a_{k+1} \quad (k=1, 2, \ldots, n-1).$$

Следствие. Если при каком-нибудь $k$ выполнено неравенство

$$a_k^2\leqslant a_{k-1}a_{k+1},$$

то уравнение (10) имеет по меньшей мере одну пару комплексных корней.

Пример 3. Определить состав корней уравнения

$$x^4+8x^3-12x^2+104x-20=0. \tag{11}$$

Решение. Так как

$$(-12)^2<8\cdot 104,$$

то уравнение (11) имеет комплексные корни и, следовательно, число вещественных корней этого уравнения не больше двух. В ряде коэффициентов уравнения (11) имеется $\Delta N=3$ перемен знаков и

$\Delta P = 1$ постоянств знаков. Отсюда на основании теоремы Декарта и следствия к ней, учитывая наличие комплексных корней, делаем вывод: уравнение (11) имеет один положительный корень, один отрицательный корень и пару комплексных корней.

## § 7. Идея метода Лобачевского — Греффе

Рассмотрим алгебраическое уравнение $n$-й степени

$$a_0x^n + a_1x^{n-1} + \ldots + a_n = 0, \tag{1}$$

где $a_0 \neq 0$. Предположим, что корни $x_1, x_2, \ldots, x_n$ уравнения (1) таковы, что

$$|x_1| \gg |x_2| \gg |x_3| \gg \ldots \gg |x_n|, \tag{2}$$

т. е. корни различны по модулю, причем модуль каждого предыдущего корня значительно больше модуля последующего*). Иными словами, мы предполагаем, что отношение любых двух соседних корней, считая в порядке убывания их номеров, есть величина, малая по модулю, т. е.

$$\left.\begin{aligned} x_2 &= \varepsilon_1 x_1, \\ x_3 &= \varepsilon_2 x_2, \\ &\ldots\ldots \\ x_n &= \varepsilon_{n-1} x_{n-1}, \end{aligned}\right\} \tag{3}$$

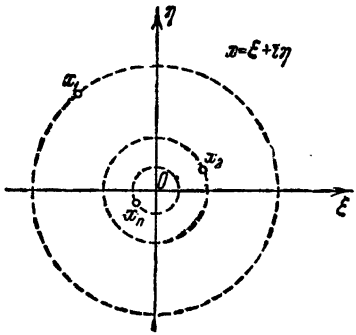


Рис. 44.

где $|\varepsilon_k| < \varepsilon$ и $\varepsilon$ — малая величина. Такие корни для краткости будем называть *отделенными* (рис. 44).

Воспользуемся теперь соотношениями между корнями и коэффициентами уравнения (1) (§ 1)

$$\left.\begin{aligned} x_1 + x_2 + \ldots + x_n &= -\frac{a_1}{a_0}, \\ x_1x_2 + x_1x_3 + \ldots + x_{n-1}x_n &= \frac{a_2}{a_0}, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ x_1x_2\ldots x_n &= (-1)^n \frac{a_n}{a_0}. \end{aligned}\right\}$$

---

*) Если коэффициенты уравнения (1) действительны, то из условия (2) следует, что все корни уравнения (1) действительны.

Отсюда в силу допущений (3) мы получаем:

$$\left.\begin{aligned} x_1(1+E_1) &= -\frac{a_1}{a_0}, \\ x_1x_2(1+E_2) &= \frac{a_2}{a_0}, \\ &\dots\dots\dots\dots \\ x_1x_2\dots x_n(1+E_n) &= (-1)^n\frac{a_n}{a_0}, \end{aligned}\right\} \tag{4}$$

где $E_1, E_2, \dots, E_n$ — малые по модулю величины по сравнению с единицей. Пренебрегая в равенствах (4) величинами $E_k$ $(k=1, 2, \dots, n)$, будем иметь приближенные соотношения

$$\left.\begin{aligned} x_1 &= -\frac{a_1}{a_0}, \\ x_1x_2 &= \frac{a_2}{a_0}, \\ &\dots\dots \\ x_1x_2\dots x_n &= (-1)^n\frac{a_n}{a_0}. \end{aligned}\right\} \tag{5}$$

Отсюда находим искомые корни

$$\left.\begin{aligned} x_1 &= -\frac{a_1}{a_0}, \\ x_2 &= -\frac{a_2}{a_1}, \\ &\dots\dots \\ x_n &= -\frac{a_n}{a_{n-1}}. \end{aligned}\right\} \tag{6}$$

Иными словами, если корни уравнения (1) отделены, то они приближенно определяются из цепи линейных уравнений

$$\begin{aligned} a_0x_1+a_1&=0, \\ a_1x_2+a_2&=0, \\ &\dots\dots\dots \\ a_{n-1}x_n+a_n&=0; \end{aligned}$$

причем точность этих корней зависит от того, насколько малы по модулю величины $\varepsilon_k$ в соотношениях (3).

Чтобы добиться отделения корней, исходя из уравнения (1), составляют преобразованное уравнение

$$a_0^{(m)}y^n+a_1^{(m)}y^{n-1}+\dots+a_n^{(m)}=0, \tag{7}$$

корнями которого $y_1, y_2, \dots, y_n$ являются $m$-е степени корней $x_1, x_2, \dots, x_n$ уравнения (1), т. е.

$$y_k=x_k^m \quad (k=1,\ 2,\ \dots,\ n). \tag{8}$$

Если корни уравнения (1), которые мы считаем расположенными в порядке убывания модулей, являются различными по модулю, то корни уравнения (7) при достаточно большой степени $m$ будут отделенными, так как

$$\frac{y_k}{y_{k-1}} = \left(\frac{x_k}{x_{k-1}}\right)^m \longrightarrow 0 \text{ при } m \longrightarrow \infty.$$

Например, пусть

$$x_1 = 2;\ x_2 = 1{,}5;\ x_3 = 1.$$

При $m = 100$ будем иметь:

$$y_1 = 1{,}27 \cdot 10^{30};\ y_2 = 4{,}06 \cdot 10^{17};\ y_3 = 1$$

и, следовательно,

$$\frac{y_2}{y_1} = 3{,}2 \cdot 10^{-13}; \qquad \frac{y_3}{y_2} = 2{,}5 \cdot 10^{-18}.$$

Обычно в качестве показателя $m$ берут степень числа 2, т. е. полагают $m = 2^p$, где $p$ — натуральное число, а само преобразование производят в $p$ приемов, каждый раз составляя уравнение, корнями которого являются квадраты корней предшествующего уравнения.

Приближенно вычислив корни $y_k$ ($k = 1, 2, \ldots, n$), из формул (8) можно определить и корни исходного уравнения (1). Точность вычислений зависит от того, насколько малым является отношение модулей соседних корней преобразованного уравнения.

Идея этого метода вычисления корней принадлежит Лобачевскому, практически удобная схема вычислений была предложена Греффе.

Достоинством метода Лобачевского — Греффе является то, что при применении этого метода нет необходимости изолировать корни. Нужно лишь избавиться от кратных корней с помощью приема, указанного в § 1. Само вычисление корней ведется однообразным регулярным способом. Как мы увидим далее, метод годится также и для нахождения комплексных корней. Неудобство метода состоит в необходимости оперирования с большими числами. Кроме того, отсутствует достаточно надежный контроль вычислений и затруднена оценка точности полученного результата.

Заметим, что если корни уравнения (1) различны, но модули некоторых из них близки между собой, то сходимость метода Лобачевского — Греффе весьма медленная. В этом случае такие корни следует рассматривать как равные по модулю и применять специальные приемы вычисления.

## § 8. Процесс квадрирования корней

Покажем теперь, как можно просто составить уравнение, корнями которого являются квадраты корней данного алгебраического уравнения, взятые со знаком минус. Последнее обстоятельство вызывается соображениями удобства, чтобы по возможности избежать появления

отрицательных коэффициентов. Процесс перехода от корней $x_k$ $(k=1, 2, \ldots, n)$ к корням

$$y_k = -x_k^2 \tag{1}$$

для краткости будем называть *квадрированием корней.*

Пусть

$$P(x) \equiv a_0x^n + a_1x^{n-1} + \ldots + a_n = 0$$

— данное уравнение, где $a_0 \neq 0$.

Обозначая через $x_1, x_2, \ldots, x_n$ корни этого уравнения, будем иметь:

$$P(x) = a_0(x-x_1)(x-x_2)\ldots(x-x_n).$$

Отсюда

$$P(-x) = (-1)^n a_0(x+x_1)(x+x_2)\ldots(x+x_n).$$

Следовательно,

$$P(x)P(-x) = (-1)^n a_0^2(x^2-x_1^2)(x^2-x_2^2)\ldots(x^2-x_n^2). \tag{2}$$

Полагая

$$y = -x^2,$$

в силу формулы (2) получим полином

$$Q(y) = P(x)P(-x),$$

корнями которого являются числа

$$y_k = -x_k^2 \quad (k=1, 2, \ldots, n).$$

Так как

$$P(-x) = (-1)^n[a_0x^n - a_1x^{n-1} + a_2x^{n-2} - \ldots + (-1)^n a_n],$$

то, производя перемножение полиномов $P(x)$ и $P(-x)$, будем иметь:

$$P(x)P(-x) = (-1)^n[a_0^2x^{2n} - (a_1^2 - 2a_0a_2)x^{2n-2} + \\ + (a_2^2 - 2a_1a_3 + 2a_0a_4)x^{2n-4} - \ldots + (-1)^n a_n^2].$$

Следовательно, интересующее нас уравнение есть

$$Q(y) \equiv A_0y^n + A_1y^{n-1} + A_2y^{n-2} + \ldots + A_n = 0,$$

где

$$A_0 = a_0^2,$$
$$A_1 = a_1^2 - 2a_0a_2,$$
$$A_2 = a_2^2 - 2a_1a_3 + 2a_0a_4,$$
$$\ldots\ldots\ldots\ldots$$
$$A_n = a_n^2.$$

Короче можно записать:

$$A_k = a_k^2 + 2\sum_{s=1}^{k}(-1)^s a_{k-s}a_{k+s} \quad (k=0,\ 1,\ 2,\ \ldots,\ n),$$

где предполагается $a_s = 0$ при $s < 0$ и $s > n$.

Правило. *При квадрировании корней каждый коэффициент преобразованного уравнения равен квадрату прежнего коэффициента, минус удвоенное произведение соседних с ним коэффициентов, плюс удвоенное произведение следующих в порядке близости коэффициентов и т. д.*, причем если нужный коэффициент отсутствует, то он считается равным нулю.

## § 9. Метод Лобачевского — Греффе для случая действительных различных корней

Пусть корни $x_1, x_2, \ldots, x_n$ уравнения $n$-й степени с действительными коэффициентами

$$a_0x^n + a_1x^{n-1} + \ldots + a_n = 0 \tag{1}$$

действительны и различны по модулю. Расположим их в порядке убывания модулей:

$$|x_1| > |x_2| > \ldots > |x_n|.$$

Последовательно применяя процесс квадрирования корней, составляем уравнение

$$b_0y^n + b_1y^{n-1} + \ldots + b_n = 0, \tag{2}$$

корнями которого служат числа

$$y_k = -x_k^{2^p} \quad (k=1,\ 2,\ \ldots,\ n). \tag{3}$$

Если $p$ достаточно велико, то корни $y_1, y_2, \ldots, y_n$ являются отделенными и на основании результатов § 7 могут быть определены из цепи линейных уравнений

$$\begin{gathered} b_0y_1 + b_1 = 0, \\ b_1y_2 + b_2 = 0, \\ \cdots\cdots\cdots \\ b_{n-1}y_n + b_n = 0. \end{gathered}$$

Отсюда получаем:

$$x_k = \pm\sqrt[2^p]{-y_k} = \sqrt[2^p]{\frac{b_k}{b_{k-1}}} \quad (k=1,\ 2,\ \ldots,\ n); \tag{4}$$

знаки корней $x_k$ определяются грубой прикидкой, при подстановке в данное уравнение, или на основании соотношений между корнями и коэффициентами уравнений. Процесс квадрирования корней обычно

продолжается до тех пор, пока удвоенные произведения не перестанут влиять на первые главные члены коэффициентов преобразованного уравнения.

Правило. *Процесс квадрирования корней следует прекратить, если коэффициенты некоторого преобразованного уравнения в пределах точности вычислений равны квадратам соответствующих коэффициентов последующего преобразованного уравнения за счет отсутствия удвоенных произведений.*

Действительно, если преобразованное уравнение, соответствующее степени $2^{p+1}$, имеет вид

$$c_0z^n + c_1z^{n-1} + \ldots + c_n = 0$$

и выполнены соотношения

$$c_k = b_k^2 \quad (k = 0,\ 1,\ 2,\ \ldots,\ n),$$

то, очевидно, получаем:

$$|x_k| = \sqrt[2^{p+1}]{\frac{c_k}{c_{k-1}}} = \sqrt[2^p]{\frac{b_k}{b_{k-1}}}.$$

Таким образом, при этих обстоятельствах мы не сможем увеличить точность вычисления корней.

Так как при применении метода Лобачевского — Греффе коэффициенты преобразованных уравнений, вообще говоря, быстро растут, то полезно выделять порядки их, записывая коэффициенты в стандартной форме $\alpha \cdot 10^m$, где $|\alpha| < 10$ и $m$ — целое число. При вычислениях, требующих большой точности, выгодно пользоваться логарифмами (см. [5]).

Пример. Методом Лобачевского — Греффе найти корни уравнения

$$x^3 - 3x + 1 = 0. \tag{5}$$

Решение. Результаты вычислений с четырьмя значащими цифрами помещены в таблице 8.

Останавливаясь на 64-й степени корней, будем иметь:

$$-x_1^{64} + 3{,}445 \cdot 10^{17} = 0,$$
$$-3{,}445 \cdot 10^{17} \cdot x_2^{64} + 2{,}486 \cdot 10^{29} = 0,$$
$$-2{,}486 \cdot 10^{29} \cdot x_3^{64} + 1 = 0.$$

Отсюда

$$x_1 = \pm \sqrt[64]{3{,}445 \cdot 10^{17}},$$
$$x_2 = \pm \sqrt[64]{\frac{2{,}486}{3{,}445} \cdot 10^{12}},$$
$$x_3 = \pm \sqrt[64]{\frac{1}{2{,}486} \cdot 10^{-29}}.$$

Таблица 8

**Вычисление действительных корней методом Лобачевского — Греффе**

| Степени | $x^3$ | $x^2$ | $x$ | $x^0$ |
|---|---|---|---|---|
| 1 | 1 | 0 | $-3$ | 1 |
| | | 0 | 9 | |
| | | 6 | 0 | |
| 2 | 1 | 6 | 9 | 1 |
| | | 36 | 81 | |
| | | $-18$ | $-12$ | |
| 4 | 1 | 18 | 69 | 1 |
| | | $3{,}24\cdot10^2$ | $4{,}761\cdot10^3$ | |
| | | $-1{,}38\cdot10^2$ | $-0{,}036\cdot10^3$ | |
| 8 | 1 | $1{,}86\cdot10^2$ | $4{,}725\cdot10^3$ | 1 |
| | | $3{,}460\cdot10^4$ | $2{,}233\cdot10^7$ | |
| | | $-0{,}945\cdot10^4$ | 0 | |
| 16 | 1 | $2{,}515\cdot10^4$ | $2{,}233\cdot10^7$ | 1 |
| | | $6{,}325\cdot10^8$ | $4{,}986\cdot10^{14}$ | |
| | | $-0{,}447\cdot10^8$ | 0 | |
| 32 | 1 | $5{,}878\cdot10^8$ | $4{,}986\cdot10^{14}$ | 1 |
| | | $3{,}455\cdot10^{17}$ | $2{,}486\cdot10^{29}$ | |
| | | $-0{,}010\cdot10^{17}$ | 0 | |
| 64 | 1 | $3{,}445\cdot10^{17}$ | $2{,}486\cdot10^{29}$ | 1 |
| | | $1{,}187\cdot10^{35}$ | $6{,}180\cdot10^{58}$ | |
| | | 0 | 0 | |
| 128 | 1 | $1{,}187\cdot10^{35}$ | $6{,}180\cdot10^{58}$ | 1 |

Логарифмируя, получим:

$$\lg|x_1| = \frac{1}{64}\cdot 17{,}53719 = 0{,}27402;$$

$$\lg|x_2| = \frac{1}{64}\cdot 11{,}85831 = 0{,}18528;$$

$$\lg|x_3| = \frac{1}{64}\cdot(-29{,}39550) = \bar{1}{,}54070,$$

и, следовательно,

$$x_1 = \pm 1{,}879;$$

$$x_2 = \pm 1{,}532;$$

$$x_3 = \pm 0{,}347.$$

Для определения знаков корней заметим, что согласно правилу Декарта уравнение (5) имеет один отрицательный корень и два

положительных корня *), причем

$$x_1 + x_2 + x_3 = 0. \tag{6}$$

Поэтому наибольшим по модулю должен быть отрицательный корень, и мы окончательно имеем:

$$x_1 = -1{,}879;$$
$$x_2 = 1{,}532;$$
$$x_3 = 0{,}347;$$

причем соотношение (6) выполнено в пределах заданной точности. Для сравнения приводим точные значения корней, полученные по формуле Кардана:

$$x_1 = 2\cos 160^\circ = -1{,}87938;$$
$$x_2 = 2\cos 40^\circ = 1{,}53208;$$
$$x_3 = 2\cos 80^\circ = 0{,}34730.$$

Заметим, что в нашем случае вычисление корней несколько упростилось благодаря тому, что крайние коэффициенты уравнения равны 1. Вообще, при применении метода Лобачевского — Греффе рекомендуется предварительно преобразовывать уравнение так, чтобы старший коэффициент уравнения был равен 1, а свободный член уравнения был равен $\pm 1$ (см. [5]).

## § 10. Метод Лобачевского — Греффе для случая комплексных корней

Обобщим теперь понятие отделения корней.

Пусть корни $x_1, x_2, \ldots, x_n$ уравнения

$$a_0x^n + a_1x^{n-1} + \ldots + a_n = 0 \tag{1}$$

удовлетворяют условиям

$$|x_1| \geqslant |x_2| \geqslant \ldots \geqslant |x_m| \gg |x_{m+1}| \geqslant |x_{m+2}| \geqslant \ldots \geqslant |x_n|. \tag{2}$$

Иными словами, предполагается, что корни уравнения (1) можно разбить на две *категории* (группы):

$$x_1, x_2, \ldots, x_m \quad (m < n)$$

и

$$x_{m+1}, x_{m+2}, \ldots, x_n,$$

*) Мы учитываем, что уравнение $P(x) \equiv x^3 - 3x + 1 = 0$ имеет положительные корни, так как $P(0) > 0$ и $P(1) < 0$.

так, что модули корней первой категории весьма велики по сравнению с модулями корней второй категории (см. рис. 45, где корни содержатся в заштрихованных областях, а внутренность незаштрихованного кругового кольца свободна от корней — «зона пустыни»).

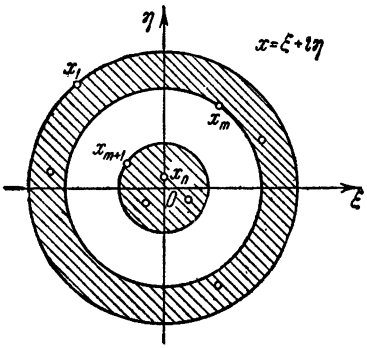


Рис. 45.

Выпишем первые $m$ соотношений между корнями и коэффициентами уравнения (1):

$$x_1+x_2+\ldots+x_m+$$
$$+(x_{m+1}+\ldots+x_n)=-\frac{a_1}{a_0},$$
$$x_1x_2+x_1x_3+\ldots+x_{m-1}x_m+$$
$$+(x_mx_{m+1}+\ldots+x_{n-1}x_n)=\frac{a_2}{a_0},$$
$$\cdots\cdots\cdots\cdots\cdots\cdots$$
$$x_1x_2\ldots x_m+(x_1x_2\ldots x_{m-1}x_{m+1}+\ldots$$
$$\ldots+x_{n-m+1}x_{n-m+2}\ldots x_n)=(-1)^m\frac{a_m}{a_0}.$$

Пренебрегая в последних равенствах взятыми в скобки относительно малыми по модулю членами, получим приближенные соотношения

$$\left.\begin{aligned} x_1+x_2+\ldots+x_m&=-\frac{a_1}{a_0},\\ x_1x_2+x_1x_3+\ldots+x_{m-1}x_m&=\frac{a_2}{a_0},\\ \cdots\cdots\cdots\cdots\cdots\cdots\\ x_1x_2\ldots x_m&=(-1)^m\frac{a_m}{a_0}.\end{aligned}\right\}\qquad(3)$$

Отсюда следует, что корни $x_1, x_2, \ldots, x_m$ первой категории (с большими модулями) приближенно являются корнями уравнения

$$a_0x^m+a_1x^{m-1}+\ldots+a_m=0. \qquad (4)$$

Из оставшихся неиспользованных $n-m$ соотношений между корнями и коэффициентами уравнения (1) получаем:

$$x_1x_2\ldots x_m(x_{m+1}+x_{m+2}+\ldots+x_n)+$$
$$+x_2x_3\ldots x_{m+1}x_{m+2}+\ldots+x_{n-m}x_{n-m+1}\ldots x_n=$$
$$=(-1)^{m+1}\frac{a_{m+1}}{a_0},$$
$$x_1x_2\ldots x_m(x_{m+1}x_{m+2}+\ldots+x_{n-1}x_n)+x_2x_3\ldots x_{m+1}x_{m+2}x_{m+3}+\ldots$$
$$\ldots+x_{n-m-1}\ldots x_n=(-1)^{m+2}\frac{a_{m+2}}{a_0},$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$x_1x_2\ldots x_mx_{m+1}\ldots x_n=(-1)^n\frac{a_n}{a_0}.$$

Отбрасывая в последних равенствах относительно малые по модулю члены, будем иметь приближенные соотношения

$$x_1x_2\ldots x_m(x_{m+1}+x_{m+2}+\ldots+x_n)=(-1)^{m+1}\frac{a_{m+1}}{a_0},$$

$$x_1x_2\ldots x_m(x_{m+1}x_{m+2}+\ldots+x_{n-1}x_n)=(-1)^{m+2}\frac{a_{m+2}}{a_0},$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$x_1x_2\ldots x_mx_{m+1}\ldots x_n=(-1)^n\frac{a_n}{a_0}.$$

Отсюда, используя последнее соотношение в формулах (3), находим:

$$\left.\begin{aligned} x_{m+2}+x_{m+3}+\ldots+x_n&=-\frac{a_{m+1}}{a_m},\\ x_{m+1}x_{m+2}+\ldots+x_{n-1}x_n&=\frac{a_{m+2}}{a_m},\\ &\ldots\ldots\ldots\ldots\\ x_{m+1}x_{m+2}\ldots x_n&=(-1)^{n-m}\frac{a_n}{a_m}.\end{aligned}\right\}\qquad(5)$$

Следовательно, корни $x_{m+1}$, $x_{m+2}$, ..., $x_n$ второй категории (с малыми модулями) приближенно являются корнями уравнения

$$a_mx^{n-m}+a_{m-1}x^{n-m-1}+\ldots+a_n=0.\qquad(6)$$

Таким образом, в наших условиях уравнение (1) распадается на два уравнения более низких степеней, каждое из которых приближенно определяет корни, принадлежащие одной из категорий.

Рассуждая по аналогии, заключаем, что если корни уравнения (1) могут быть разбиты на $p$ категорий

$$x_1,\ x_2,\ \ldots,\ x_{m_1};$$

$$x_{m_1+1},\ x_{m_1+2},\ \ldots,\ x_{m_2};$$

$$\ldots\ldots\ldots\ldots$$

$$x_{m_{p-1}+1},\ x_{m_{p-1}+2},\ \ldots,\ x_{m_p}$$

$$(m_1+m_2+\ldots+m_p=n),$$

так что выполнено условие

$$|x_1|\geqslant|x_2|\geqslant\ldots\geqslant|x_{m_1}|\gg|x_{m_1+1}|\geqslant|x_{m_1+2}|\geqslant\ldots\geqslant|x_{m_2}|\gg$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\gg|x_{m_{p-1}+1}|\geqslant|x_{m_{p-1}+2}|\geqslant\ldots\geqslant|x_{m_p}|,$$

т. е. модули корней, принадлежащих более низким категориям, значительно превышают модули корней, принадлежащих более высоким категориям (можно сказать, что эти корни являются *отделенными*

*в групповом смысле*), то корни каждой категории приближенно могут быть определены из соответствующих уравнений

$$\left.\begin{array}{l} a_0x^{m_1}+a_1x^{m_1-1}+\ldots+a_{m_1}=0, \\ a_{m_1}x^{m_2}+a_{m_1+1}x^{m_2-1}+\ldots+a_{m_1+m_2}=0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{m_1+m_2+\ldots+m_{p-1}}x^{m_p}+a_{m_1+m_2+\ldots+m_{p-1}+1}x^{m_p-1}+\ldots \\ \qquad\qquad\qquad\qquad\qquad\qquad \ldots+a_{m_1+m_2+\ldots+m_p}=0, \end{array}\right\} \tag{7}$$

степени которых соответственно $m_1$, $m_2$, ..., $m_p$. В частности, если корни уравнения (1) полностью отделены, то уравнения (7) — линейные; паре комплексных корней при отсутствии других корней того же модуля будет соответствовать квадратное уравнение в совокупности уравнений (7) и т. д.

Мы ограничимся здесь рассмотрением простейших случаев, когда уравнение (1), коэффициенты которого считаются действительными, имеет одну пару комплексных корней или же две пары комплексных корней с различными модулями, причем модули действительных корней различны и отличны от модулей комплексных корней. С более общими случаями можно познакомиться в книгах Крылова [5] и Скарборо [6].

## § 11. Случай пары комплексных корней

Пусть

$$\left.\begin{array}{l} x_m=u+iv, \\ x_{m+1}=u-iv \end{array}\right\} \tag{1}$$

($u$ и $v$ действительны, $v\neq 0$) — комплексные корни уравнения (1) § 10, причем все остальные корни $x_k$ ($k\neq m$, $k\neq m+1$) этого уравнения действительны и удовлетворяют условию

$$|x_1|>|x_2|>\ldots>|x_m|=|x_{m+1}|>\ldots>|x_n|. \tag{2}$$

Применяя процесс квадрирования корней, составим уравнение

$$b_0y^n+b_1y^{n-1}+\ldots+b_n=0,$$

корнями которого являются:

$$y_k=-x_k^{2^p} \qquad (k=1,\ 2,\ \ldots,\ n).$$

При достаточно большом натуральном $p$ действительные корни $y_1$, ..., $y_{m-1}$, $y_{m+2}$, ..., $y_n$ будут разделены с большой степенью

точности и могут быть определены из линейных уравнений

$$\begin{gathered} b_0 y_1 + b_1 = 0, \\ \cdots\cdots\cdots\cdots \\ b_{m-2} y_{m-1} + b_{m-1} = 0, \\ b_{m+1} y_{m+2} + b_{m+2} = 0, \\ \cdots\cdots\cdots\cdots \\ b_{n-1} y_n + b_n = 0. \end{gathered}$$

Отсюда получаем:

$$x_k = \pm \sqrt[2^p]{\frac{b_k}{b_{k-1}}} \quad (k \neq m, \ k \neq m+1).$$

Процесс квадрирования корней прекращается, когда в коэффициентах $b_1, \ldots, b_{m-1}, b_{m+1}, \ldots, b_n$ при следующем шаге пропадают, в пределах заданной степени точности, удвоенные произведения. Что касается коэффициента $b_m$, то в его состав, вообще говоря, не входят исчезающие удвоенные произведения. Может даже случиться, что эти произведения превалируют над квадратом и коэффициент $b_m$ имеет меняющийся знак. Это обстоятельство служит характерным признаком наличия комплексных корней или корней с равными модулями в уравнении (1) § 10, причем ведущий себя необычным образом коэффициент $b_m$ указывает м е с т о таких корней в ряде модулей (2).

Заметим, что рассматриваемое уравнение заведомо имеет комплексные корни, если коэффициент $b_m$ меняет знак; так, при наличии лишь действительных корней все коэффициенты преобразованных уравнений, очевидно, неотрицательны.

Согласно общей теории корни $y_m$ и $y_{m+1}$, соответствующие комплексным корням $x_m$ и $x_{m+1}$, приближенно удовлетворяют квадратному уравнению

$$b_{m-1} y^2 + b_m y + b_{m+1} = 0.$$

Обратим внимание, что коэффициент $b_m$ является с р е д н и м. Так как

$$x_k x_{k+1} = u^2 + v^2 = r^2,$$

где

$$r = |x_k| = |x_{k+1}|$$

— общий модуль комплексных корней, и

$$y_m y_{m+1} = x_k^{2^p} \cdot x_{k+1}^{2^p} = (x_k x_{k+1})^{2^p} = (r^2)^{2^p},$$

то по свойству корней квадратного уравнения имеем:

$$(r^2)^{2^p} = \frac{b_{m+1}}{b_{m-1}}.$$

Отсюда определяем квадрат модуля комплексных корней

$$r^2 = \sqrt[2^p]{\frac{b_{m+1}}{b_{m-1}}}. \tag{3}$$

Действительную часть $u$ комплексных корней проще всего найти, воспользовавшись соотношением

$$x_1+x_2+\ldots+x_{m-1}+(x_m+x_{m+1})+x_{m+2}+\ldots+x_n=-\frac{a_1}{a_0}.$$

Отсюда

$$2u=x_m+x_{m+1}=-\frac{a_1}{a_0}-\sum_{\substack{k\neq m\\ k\neq m+1}}'' x_k$$

и, следовательно,

$$u=-\frac{a_1}{2a_0}-\frac{1}{2}\sum_{\substack{k\neq m\\ k\neq m+1}}'' x_k. \tag{4}$$

Зная в силу формулы (3) общий модуль $r$ комплексных корней, находим коэффициент $v$ их мнимой части

$$v=\sqrt{r^2-u^2}. \tag{5}$$

По формулам (4) и (5) определяем искомые комплексные корни

$$x_{m,\,m+1}=u\pm iv.$$

Можно также искать комплексные корни в тригонометрическом виде

$$x_{m,\,m+1}=r(\cos\varphi\pm i\sin\varphi).$$

Пример. Найти корни уравнения [7]

$$x^4+x^3-10x^2-34x-26=0. \tag{6}$$

Решение. Результаты вычислений с четырьмя значащими цифрами приведены в таблице 9.

Из таблицы 9 видно, что в пятом преобразованном уравнении (степень корней $2^5=32$) действительные корни $x_1$ и $x_4$ (считая в порядке убывания модулей) являются отделенными. Эти корни можно найти из двучленных уравнений

$$-x_1^{32}+2{,}005\cdot10^{19}=0,$$

$$-2{,}704\cdot10^{43}x_4^{32}+1{,}901\cdot10^{45}=0.$$

Отсюда

$$x_1=\pm\sqrt[32]{2{,}005\cdot10^{19}},\qquad x_4=\pm\sqrt[32]{\frac{1{,}901}{2{,}704}\cdot10^2}.$$

Логарифмируя, будем иметь:

$$\lg|x_1|=\frac{1}{32}\cdot19{,}30211=0{,}60319;$$

$$\lg|x_4|=\frac{1}{32}\cdot(2{,}27898-0{,}43201)=0{,}05772.$$

Таблица 9

**Вычисление комплексных корней методом Лобачевского — Греффе**

| Степени | $x^4$ | $x^3$ | $x^2$ | $x$ | $x^0$ |
|---|---|---|---|---|---|
| 1 | 1 | 1<br>1<br>20 | —10<br>100<br>68<br>—52 | —34<br>1156<br>—520 | —26 |
| 2 | 1 | 21<br>441<br>—239 | 116<br>$1,346\cdot10^4$<br>$-2,671\cdot10^4$<br>$0,135\cdot10^4$ | 636<br>$4,045\cdot10^5$<br>$-1,568\cdot10^5$ | 676 |
| 4 | 1 | 209<br>$4,368\cdot10^4$<br>$2,380\cdot10^4$ | $-1,190\cdot10^4$<br>$1,416\cdot10^8$<br>$-1,035\cdot10^8$<br>$0,009\cdot10^8$ | $2,477\cdot10^5$<br>$6,135\cdot10^{10}$<br>$1,088\cdot10^{10}$ | $4,570\cdot10^5$ |
| 8 | 1 | $6,748\cdot10^4$<br>$4,554\cdot10^9$<br>$-0,078\cdot10^9$ | $3,90\cdot10^7$<br>$1,521\cdot10^{15}$<br>$-9,748\cdot10^{15}$<br>0 | $7,223\cdot10^{10}$<br>$5,216\cdot10^{21}$<br>$-0,016\cdot10^{21}$ | $2,088\cdot10^4$ |
| 16 | 1 | $4,476\cdot10^9$<br>$2,003\cdot10^{19}$<br>$0,002\cdot10^{19}$ | $-8,227\cdot10^{15}$<br>$6,768\cdot10^{31}$<br>$-4,655\cdot10^{31}$<br>0 | $5,200\cdot10^{21}$<br>$2,704\cdot10^{43}$<br>0 | $4,360\cdot10^{22}$ |
| 32 | 1 | $2,005\cdot10^{19}$<br>$4,020\cdot10^{38}$<br>0 | $2,113\cdot10^{31}$<br>$4,465\cdot10^{62}$<br>$-1,084\cdot10^{63}$<br>0 | $2,704\cdot10^{43}$<br>$7,312\cdot10^{86}$<br>0 | $1,901\cdot10^{45}$ |
| 64 | 1 | $4,020\cdot10^{38}$ | $-6,38\cdot10^{62}$ | $7,312\cdot10^{86}$ | $3,614\cdot10^{90}$ |

Следовательно,

$$x_1 = \pm 4,010; \qquad x_4 = \pm 1,142.$$

Грубой прикидкой убеждаемся, что корень $x_1$ — положительный, а корень $x_4$ — отрицательный. Таким образом, окончательно имеем:

$$x_1 = 4,010; \quad x_4 = -1,142.$$

Так как преобразованный коэффициент при $x^2$ меняет знак, то данное уравнение имеет комплексные корни $x = x_2$ и $x = x_3$, которые определяются из трехчленного уравнения

$$2,005\cdot10^{19}y^2 + 2,113\cdot10^{31}y + 2,704\cdot10^{43} = 0,$$

где

$$y = -x^{32}.$$

Согласно общей теории модуль корней

$$r = |x_2| = |x_3|$$

находится по формуле (3)

$$r^2 = \sqrt[32]{\frac{2,704}{2,005} \cdot 10^{24}}.$$

Отсюда

$$\lg r^2 = \frac{1}{32} \cdot (24,43201 - 0,30211) = 0,75406$$

и, следовательно,

$$r^2 = 5,6763.$$

Полагая

$$x_2 = u + iv, \quad x_3 = u - iv,$$

из соотношения

$$x_1 + x_2 + x_3 + x_4 = -1$$

получим:

$$u = \frac{1}{2}(-1 - 4,010 + 1,142) = -1,934.$$

Коэффициент мнимой части $v$ определяется по формуле

$$v = \sqrt{r^2 - u^2} = \sqrt{5,6763 - 3,7404} = \sqrt{1,9359} = 1,395.$$

Следовательно,

$$x_{2,3} = -1,934 \pm 1,395i.$$

Заметим, что корни $x_2$ и $x_3$ можно также определить из соотношений между корнями и коэффициентами уравнения (6), а именно, имеем:

$$x_1 + x_2 + x_3 + x_4 = -1,$$
$$x_1 x_2 x_3 x_4 = -26.$$

Отсюда, использовав найденные выше значения $x_1$ и $x_4$, получим:

$$x_2 + x_3 = -3,869;$$
$$x_2 x_3 = 5,677.$$

Поэтому $x_2$ и $x_3$ можно найти как корни квадратного уравнения

$$x^2 + 3,869x + 5,677 = 0,$$

решив которое, будем иметь:

$$x_{2,3} = -1,934 \pm 1,391i.$$

## § 12. Случай двух пар комплексных корней

Пусть уравнение (1) § 10 допускает две пары комплексных корней:

$$x_k = u_1 + iv_1, \quad x_{k+1} = u_1 - iv_1$$

и

$$x_m = u_2 + iv_2, \quad x_{m+1} = u_2 - iv_2$$

с различными модулями ($u_1$, $v_1$, $u_2$, $v_2$ действительны и $v_1 \neq 0$, $v_2 \neq 0$), причем все остальные корни $x_j$ ($j \neq k$, $j \neq k+1$, $j \neq m$, $j \neq m+1$) этого уравнения действительны, различны по абсолютной величине, не равны нулю*) и отличны по модулю от комплексных корней, т. е.

$$|x_1| > |x_2| > \ldots > |x_{k-1}| > |x_k| = |x_{k+1}| > \ldots > |x_m| = \\ = |x_{m+1}| > \ldots > |x_n| > 0. \quad (1)$$

Как обычно, производя квадрирование корней рассматриваемого уравнения до некоторой степени $2^p$, получим преобразованное уравнение

$$b_0 y^n + b_1 y^{n-1} + \ldots + b_n = 0,$$

корнями которого являются числа

$$y_j = -x_j^{2^p} \quad (j = 1, 2, \ldots, n).$$

При достаточно большом натуральном $p$ обнаружится, что при переходе к степени $2^{p+1}$ некоторые коэффициенты $c_j$ нового преобразованного уравнения

$$c_0 z^n + c_1 z^{n-1} + \ldots + c_n = 0$$

будут представлять собой, в пределах заданной точности, квадраты соответствующих коэффициентов $b_j$ предшествующего преобразованного уравнения. При нашем предположении (1), мы в конце концов получим:

$$c_j = b_j^2 \text{ при } j = 0, 1, 2, \ldots, k-1, k+1, \ldots, m-1, m+1, \ldots, n$$

и

$$c_k \neq b_k^2 \text{ и } c_m \neq b_m^2.$$

Это обстоятельство позволяет установить место комплексных корней. Заметим, что достаточным признаком наличия двух пар комплексных корней уравнения (1) § 10 служит изменение знаков коэффициентов $b_k$ и $b_m$ для различных показателей $2^p$.

Действительные корни $x_j$ рассматриваемого уравнения определяются из двучленных уравнений

$$-b_{j-1} x_j^{2^p} + b_j = 0.$$

Отсюда

$$x_j = \pm \sqrt[2^p]{\frac{b_j}{b_{j-1}}} \quad (j \neq k,\ j \neq k+1,\ j \neq m,\ j \neq m+1).$$

---

*) Нулевые корни могут быть выделены предварительно.

Комплексные корни $x_k$, $x_{k+1}$ и $x_m$, $x_{m+1}$ соответственно определяются из трехчленных уравнений

$$b_{k-1}x^{2^{p+1}} - b_k x^{2^p} + b_{k+1} = 0 \tag{2'}$$

и

$$b_{m-1}x^{2^{p+1}} - b_m x^{2^p} + b_{m+1} = 0. \tag{2''}$$

Введем обозначения:

$$r_1 = |x_k| = |x_{k+1}|$$

и

$$r_2 = |x_m| = |x_{m+1}|.$$

Учитывая, что

$$r_1^2 = x_k x_{k+1}$$

и

$$r_2^2 = x_m x_{m+1},$$

из уравнений (2′) и (2″) можно вычислить квадраты модулей комплексных корней

$$r_1^2 = \sqrt[2^p]{\frac{b_{k+1}}{b_{k-1}}} \quad \text{и} \quad r_2^2 = \sqrt[2^p]{\frac{b_{m+1}}{b_{m-1}}}.$$

Для определения действительных частей $u_1$ и $u_2$ комплексных корней воспользуемся соотношениями между корнями и коэффициентами уравнения (1) § 10, а именно, имеем:

$$x_2x_3 \ldots x_n + x_1x_3 \ldots x_n + \ldots + x_1x_2 \ldots x_{n-1} = (-1)^{n-1}\frac{a_{n-1}}{a_0}$$

и

$$x_1x_2 \ldots x_n = (-1)^n \frac{a_n}{a_0}.$$

Разделив первое равенство на второе, получим:

$$\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n} = -\frac{a_{n-1}}{a_n}.$$

Кроме того,

$$x_1 + x_2 + \ldots + x_n = -\frac{a_1}{a_0}.$$

Отсюда, учитывая соотношения

$$x_k + x_{k+1} + x_m + x_{m+1} = 2u_1 + 2u_2$$

и

$$\frac{1}{x_k} + \frac{1}{x_{k+1}} + \frac{1}{x_m} + \frac{1}{x_{m+1}} = \frac{2u_1}{r_1^2} + \frac{2u_2}{r_2^2},$$

имеем следующую линейную систему уравнений:

$$\left.\begin{aligned} u_1+u_2 &= -\frac{a_1}{2a_0}-\frac{1}{2}\sigma, \\ \frac{u_1}{r_1^2}+\frac{u_2}{r_2^2} &= -\frac{a_{n-1}}{2a_n}-\frac{1}{2}\sigma', \end{aligned}\right\} \tag{3}$$

где $\sigma$ — сумма действительных корней и $\sigma'$ — сумма обратных величин действительных корней, т. е.

$$\sigma=\sum_{j\neq k,\ k+1,\ m,\ m+1} x_j$$

и

$$\sigma'=\sum_{j\neq k,\ k+1,\ m,\ m+1} \frac{1}{x_j}.$$

Найдя из системы (3) $u_1$ и $u_2$, определяем коэффициенты $v_1$ и $v_2$ мнимых частей комплексных корней по формулам

$$v_1=\sqrt{r_1^2-u_1^2},\quad v_2=\sqrt{r_2^2-u_2^2}.$$

Таким образом, окончательно находим:

$$x_{k,\ k+1}=u_1\pm iv_1$$

и

$$x_{m,\ m+1}=u_2\pm iv_2.$$

Пример. Методом Лобачевского — Греффе решить уравнение [7]

$$x^4+4x^2-3x+3=0. \tag{4}$$

Решение. Применяя процесс квадрирования корней до 16-й степени и производя вычисления с точностью до четырех значащих цифр, получим результаты, приведенные в таблице 10.

Легко видеть, что при следующем преобразовании средний коэффициент будет равен квадрату своего прежнего значения. Поэтому мы прекращаем процесс квадрирования корней. Так как для 16-й степени среди коэффициентов преобразованного уравнения имеются два отрицательных коэффициента, то уравнение (4) допускает две пары комплексных корней:

$$x_{1,2}=u_1\pm iv_1$$

и

$$x_{3,4}=u_2\pm iv_2,$$

которые соответственно удовлетворяют трехчленным уравнениям

$$x^{32}+1{,}359\cdot 10^5x^{16}+5{,}720\cdot 10^9=0$$

и

$$5{,}720\cdot 10^9\cdot x^{32}+8{,}184\cdot 10^8\cdot x^{16}+4{,}305\cdot 10^7=0$$

Таблица 10

**Вычисление двух пар комплексных корней методом Лобачевского — Греффе**

| Степени | $x^4$ | $x^3$ | $x^2$ | $x$ | $x^0$ |
|---|---|---|---|---|---|
| 1 | 1 | 0 | 4 | −3 | 3 |
| | | 0 | 16 | 9 | |
| | | −8 | 0 | −24 | |
| | | | 6 | | |
| 2 | 1 | −8 | 22 | −15 | 9 |
| | | 64 | 484 | 225 | |
| | | −44 | −240 | −396 | |
| | | | 18 | | |
| 4 | 1 | 20 | 262 | −171 | 81 |
| | | $4\cdot10^2$ | $6,864\cdot10^4$ | $2,924\cdot10^4$ | |
| | | $-5,24\cdot10^2$ | $0,684\cdot10^4$ | $-4,244\cdot10^4$ | |
| | | | $0,016\cdot10^4$ | | |
| 8 | 1 | $-1,24\cdot10^2$ | $7,564\cdot10^4$ | $-1,320\cdot10^4$ | $6,561\cdot10^3$ |
| | | $1,538\cdot10^4$ | $5,723\cdot10^9$ | $1,743\cdot10^8$ | |
| | | $-15,128\cdot10^4$ | $-0,003\cdot10^9$ | $-9,927\cdot10^8$ | |
| | | | 0 | | |
| 16 | 1 | $-1,359\cdot10^5$ | $5,720\cdot10^9$ | $-8,184\cdot10^8$ | $4,305\cdot10^7$ |

Отсюда определяем квадраты модулей этих корней:

$$r_1^2=\sqrt[16]{5,720\cdot10^9}=4,072$$

и

$$r_2^2=\sqrt[16]{\frac{4,305}{5,720}\cdot10^{-2}}=0,7367.$$

Так как

$$\frac{1}{r_1^2}=0,2456;\quad \frac{1}{r_2^2}=1,3574,$$

то на основании системы (3) для нахождения действительных частей $u_1$ и $u_2$ корней имеем систему

$$u_1+u_2=0,$$
$$0,2456u_1+1,3574u_2=0,5.$$

Отсюда

$$u_1=-0,4497;$$
$$u_2=0,4497.$$

Используя найденные квадраты модулей $r_1^2$ и $r_2^2$ корней, определяем коэффициенты $v_1$ и $v_2$ мнимых частей корней:

$$v_1 = \sqrt{r_1^2 - u_1^2} = 1{,}967;$$

$$v_2 = \sqrt{r_2^2 - u_2^2} = 0{,}731.$$

Таким образом, корни уравнения (4) имеют вид

$$x_{1,2} = -0{,}450 \pm 1{,}967i$$

и

$$x_{3,4} = 0{,}450 \pm 0{,}731i.$$

## § 13. Метод Бернулли

Пусть дано алгебраическое уравнение

$$a_0x^n + a_1x^{n-1} + \ldots + a_n = 0 \qquad (a_0 \neq 0), \tag{1}$$

корни которого $x_1, x_2, \ldots, x_n$ различны.

На основе коэффициентов $a_k$ $(k = 0, 1, \ldots, n)$ построим так называемое *конечно-разностное уравнение*

$$a_0y_{n+i} + a_1y_{n+i-1} + \ldots + a_ny_i = 0 \qquad (i = 0, 1, 2, \ldots), \tag{2}$$

которое представляет собой рекуррентное соотношение, связывающее любые, следующие друг за другом, $n+1$ членов бесконечной последовательности

$$y_0, y_1, y_2, \ldots, y_i, \ldots \tag{3}$$

Последовательность (3) $y_i = f(i)$ $(i = 0, 1, 2, \ldots)$, члены которой удовлетворяют конечно-разностному уравнению (2), называется *решением* этого уравнения. Для построения решения $y_i$ достаточно задать $n$ его *начальных значений* $y_0, y_1, \ldots, y_{n-1}$; остальные члены $y_n$, $y_{n+1}, \ldots$ шаг за шагом можно определить из уравнения (2).

В теории конечных разностей [8] доказывается, что если корни $x_1$, $x_2, \ldots, x_n$ алгебраического уравнения (1) различны, то любое решение конечно-разностного уравнения (2) имеет вид

$$y_i = C_1x_1^i + C_2x_2^i + \ldots + C_nx_n^i \qquad (i = 0, 1, 2, \ldots), \tag{4}$$

где $C_1, C_2, \ldots, C_n$ — произвольные постоянные. Таким образом, уравнение (1) по отношению к уравнению (2) является *характеристическим*. Постоянные $C_1, C_2, \ldots, C_n$ могут быть определены из *начальных условий*:

$$\left.\begin{aligned} y_0 &= C_1 + C_2 + \ldots + C_n, \\ y_1 &= C_1x_1 + C_2x_2 + \ldots + C_nx_n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_{n-1} &= C_1x_1^{n-1} + C_2x_2^{n-1} + \ldots + C_nx_n^{n-1}. \end{aligned}\right\} \tag{5}$$

Теорема. *Пусть алгебраическое уравнение* (1) *имеет единственный наибольший по модулю корень* $x_1$. *Тогда отношение двух последовательных членов* $y_{i+1}$ *и* $y_i$ *решения конечно-разностного уравнения* (2) *стремится, вообще говоря, к пределу, равному* $x_1$, *т. е.*

$$\lim_{i\to\infty}\frac{y_{i+1}}{y_i}=x_1. \tag{6}$$

Доказательство. Пусть

$$|x_1|>|x_2|\geqslant\ldots\geqslant|x_n|. \tag{7}$$

Предполагая, что корни $x_k$ ($k=1, 2, \ldots, n$) различны, из формулы (4) имеем:

$$y_i=x_1^i\left[C_1+C_2\left(\frac{x_2}{x_1}\right)^i+\ldots+C_n\left(\frac{x_n}{x_1}\right)^i\right]$$

и

$$y_{i+1}=x_1^{i+1}\left[C_1+C_2\left(\frac{x_2}{x_1}\right)^{i+1}+\ldots+C_n\left(\frac{x_n}{x_1}\right)^{i+1}\right].$$

Отсюда

$$\frac{y_{i+1}}{y_i}=x_1\cdot\frac{C_1+C_2\left(\frac{x_2}{x_1}\right)^{i+1}+\ldots+C_n\left(\frac{x_n}{x_1}\right)^{i+1}}{C_1+C_2\left(\frac{x_2}{x_1}\right)^{i}+\ldots+C_n\left(\frac{x_n}{x_1}\right)^{i}}. \tag{8}$$

Если $C_1\neq 0$, то, переходя к пределу в формуле (8) при $i\to\infty$ и учитывая, что в силу неравенств (7) имеют место предельные соотношения

$$\left(\frac{x_2}{x_1}\right)^i\to 0,\ \ldots,\ \left(\frac{x_n}{x_1}\right)^i\to 0,$$

будем иметь:

$$\lim_{i\to\infty}\frac{y_{i+1}}{y_i}=x_1.$$

Замечание 1. Если при неудачном выборе решения окажется, что $C_1=0$, а $C_2\neq 0$, то предел (6) будет равен следующему наибольшему по модулю корню уравнения (1).

Замечание 2. Если для решения $y_i$ отношение $\frac{y_{i+1}}{y_i}$ колеблется, не стремясь к пределу, то возникает подозрение, что уравнение (1) имеет комплексные корни, наибольшие по модулю.

Замечание 3. Производя в уравнении (1) замену переменной

$$x=\frac{1}{z},$$

методом Бернулли можно найти отличный от нуля наименьший по модулю корень уравнения (1).

Таким образом, для приближенного нахождения наибольшего по модулю корня $x_1$ уравнения (1) можно пользоваться формулой

$$x_1 \approx \frac{y_i}{y_{i-1}},$$

где $i$ достаточно велико.

Для практического применения метода Бернулли нужно задать произвольные числа $y_0, y_1, \ldots, y_{n-1}$, а затем, пользуясь формулой

$$y_{n+i} = -\frac{1}{a_0}(a_n y_i + a_{n-1} y_{i-1} + \ldots + a_1 y_{n+i-1}) \qquad (i = 0, 1, 2, \ldots),$$

вычислить последовательность чисел $y_n, y_{n+1}, y_{n+2}, \ldots$ и отношения $\frac{y_n}{y_{n-1}}, \frac{y_{n+1}}{y_n}, \frac{y_{n+2}}{y_{n+1}}, \ldots$ Если отношение $\frac{y_{n+i}}{y_{n+i-1}}$ при возрастающем $i$ обнаруживает тенденцию приближения к некоторому числу $\xi$, то последнее принимается за наибольший по модулю корень $x_1$ уравнения (1). В противном случае весьма вероятно, что уравнение (1) имеет несколько корней, наибольших по модулю, или, что менее вероятно, что для начальной системы чисел $y_0, y_1, \ldots, y_{n-1}$ коэффициент $C_1 = 0$.

Если известно грубое значение $\alpha$ наибольшего по модулю корня $x_1$, то для ускорения сходимости процесса выгодно положить:

$$y_0 = 1, \; y_1 = \alpha, \; \ldots, \; y_{n-1} = \alpha^{n-1}.$$

Заметим, что метод Бернулли сводится к повторению однообразных операций и поэтому удобен для реализации на счетных машинах.

Начальные значения $y_i$ $(i = 0, 1, \ldots, n-1)$, вообще говоря, могут быть взяты произвольно. Обычно берут $y_0 = y_1 = \ldots = y_{n-2} = 0$; $y_{n-1} = 1$. Хильдебрант [9] предложил выбирать $y_i$ так, чтобы все коэффициенты $C_i$ в формуле (4) были равны 1. В этом случае, при наличии единственного наибольшего по модулю корня уравнения (1), процесс $\frac{y_i}{y_{i-1}}$ заведомо сходится при $i \to \infty$.

Метод Бернулли может быть применен также для вычисления комплексных корней уравнения (1) [10].

Пример. Найти наибольший по модулю корень $x_1$ уравнения

$$x^5 + 5x^4 - 5 = 0.$$

Решение. Соответствующее конечно-разностное уравнение имеет вид

$$y_{i+5} = 5(y_i - y_{i+4}) \qquad (i = 0, 1, 2, \ldots). \tag{9}$$

Произвольно берем значения

$$y_0 = 0, \quad y_1 = 0, \quad y_2 = 0, \quad y_3 = 0, \quad y_4 = 1.$$

По формуле (9) подсчитываем значения $y_i$ при $i \geqslant 5$. Найденные значения приведены в таблице 11.

Таблица 11

**Нахождение корней алгебраического уравнения методом Бернулли**

| $i$ | $y_i$ | $\frac{y_i}{y_{i-1}}$ | $i$ | $y_i$ | $\frac{y_i}{y_{i-1}}$ |
|---|---|---|---|---|---|
| 5 | —5 | —5 | 10 | 15 575 | —4,992 |
| 6 | 25 | —5 | 11 | —77 750 | —4,928 |
| 7 | —125 | —5 | 12 | 388 125 | —4,99196 |
| 8 | 625 | —5 | 13 | —1 937 500 | —4,991948 |
| 9 | —3120 | —4,992 | | | |

Остановившись на $y_{13}$, будем иметь:

$$x_1 \approx \frac{y_{13}}{y_{12}} = -\frac{1937500}{388125} = -4{,}991948.$$

Отсюда, учитывая $y_{12}$, приближенно можно положить:

$$x_1 = -4{,}99195.$$

В заключение отметим, что в последнее время появились новые методы решения алгебраических уравнений, обладающие удобными вычислительными схемами (метод Лина, метод Н. В. Палувера и др.) [10].

### Литература к пятой главе

1. А. Г. Курош, Курс высшей алгебры, Гостехиздат, М.—Л., 1946, гл. 7 и 8.
2. Г. М. Шапиро, Высшая алгебра, Изд. 4, ГУПИ, М., 1938, гл. III и VI.
3. Д. Граве, Элементы высшей алгебры, Киев, 1914, гл. X.
4. Б. А. Фукс и Б. В. Шабат, Функции комплексного переменного, Гостехиздат, М.—Л., 1949, гл. VII.
5. А. Н. Крылов, Лекции о приближенных вычислениях, Изд. 2, Изд-во АН СССР, 1933, гл. II.
6. Дж. Скарборо, Численные методы математического анализа, ГТТИ, М.—Л., 1934, гл. X.
7. Б. К. Млодзеевский, Решение численных уравнений, ГИЗ, М., 1924, гл. IV.
8. А. О. Гельфонд, Исчисление конечных разностей, Гостехиздат, 1952, гл. V.
9. F. B. Hildebrand, Introduction to numerical analysis, New York—Toronto—London, 1956.
10. В Л. Загускин, Справочник по численным методам решения алгебраических и трансцендентных уравнений, Физматгиз, 1960.

## ГЛАВА VI

# УЛУЧШЕНИЕ СХОДИМОСТИ РЯДОВ

### § 1. Улучшение сходимости числовых рядов

Говорят, что ряд

$$a_1 + a_2 + \ldots + a_n + \ldots \tag{1}$$

*сходится медленно*, если нужно взять весьма большое число членов ряда, чтобы получить его сумму с заданной степенью точности. Например, пусть требуется найти сумму ряда

$$S = \frac{1}{1^2} + \frac{1}{2^2} + \ldots + \frac{1}{n^2} + \ldots \tag{2}$$

с точностью до $10^{-6}$. Для $n$-го остатка ряда имеем оценку

$$R_n < \int_n^\infty \frac{dx}{x^2} = \frac{1}{n}.$$

Следовательно, наша точность будет гарантирована, если мы возьмем сумму 1 000 000 членов ряда, что практически невозможно. Поэтому ряд (2) при решении поставленной задачи следует рассматривать как медленно сходящийся ряд.

Таким образом, непосредственное нахождение суммы медленно сходящегося ряда с заданной точностью $\varepsilon$, вообще говоря, затруднительно или даже практически невыполнимо. Поэтому важное значение приобретают преобразования рядов, улучшающие их сходимость. Мы здесь ознакомимся с *преобразованием Куммера* [3], [4], в ряде случаев полезным для нашей цели.

Пусть ряд (1) сходится и сумма его равна $A$. Подберем вспомогательный сходящийся ряд

$$b_1 + b_2 + \ldots + b_n + \ldots \qquad (b_n \neq 0), \tag{2'}$$

сумма которого $B$ известна, причем такой, что существует

$$\lim_{n\to\infty}\frac{a_n}{b_n}=q\neq 0. \tag{3}$$

Тогда имеем очевидное равенство

$$\sum_{n=1}^{\infty}a_n=q\sum_{n=1}^{\infty}b_n+\sum_{n=1}^{\infty}(a_n-qb_n)$$

или

$$A=qB+\sum_{n=1}^{\infty}(a_n-qb_n). \tag{4}$$

В частности, если $a_n\sim b_n$, то $q=1$ и мы имеем:

$$A=B+\sum_{n=1}^{\infty}(a_n-b_n). \tag{4'}$$

Следовательно, нахождение суммы ряда (1) в общем случае заменяется нахождением суммы ряда

$$\sum_{n=1}^{\infty}(a_n-qb_n). \tag{5}$$

Остаток ряда (5) $\overline{R}_N$ может быть записан в следующем виде:

$$\overline{R}_N=\sum_{n=N+1}^{\infty}(a_n-qb_n)=\sum_{n=N+1}^{\infty}\left(1-q\frac{b_n}{a_n}\right)a_n=\sum_{n=N+1}^{\infty}\varepsilon_n a_n,$$

где $\varepsilon_n=1-q\frac{b_n}{a_n}\to 0$ при $n\to\infty$.

Поэтому, вообще говоря, ряд (5) сходится быстрее исходного ряда (1). Основная трудность применения преобразования Куммера состоит в удачном подборе вспомогательного ряда (2').

Покажем применение этого преобразования для знакоположительного ряда (1), члены которого $a_n$ есть рациональные функции целочисленной переменной $n$, т. е.

$$a_n=\frac{\alpha_0 n^p+\alpha_1 n^{p-1}+\ldots+\alpha_p}{\beta_0 n^q+\beta_1 n^{q-1}+\ldots+\beta_q}\qquad (n=1,\ 2,\ \ldots), \tag{6}$$

где $p$ и $q$—целые неотрицательные числа и $\alpha_0>0$, $\beta_0>0$. Для сходимости ряда с общим членом (6) необходимо и достаточно, чтобы имело место неравенство

$$q\geqslant p+2.$$

В этом случае

$$a_n = O\left(\frac{1}{n^2}\right)^{*)}$$

(по меньшей мере!).

Рассмотрим вспомогательные ряды

$$S^{(m)} = \sum_{n=1}^{\infty} \frac{1}{n(n+1)\ldots(n+m)} \qquad (m = 1,\ 2,\ \ldots). \tag{7}$$

Так как

$$\frac{1}{n(n+1)\ldots(n+m)} = \\ = \frac{1}{m}\left[\frac{1}{n(n+1)\ldots(n+m-1)} - \frac{1}{(n+1)(n+2)\ldots(n+m)}\right],$$

то

$$S_N^{(m)} = \sum_{n=1}^{N} \frac{1}{n(n+1)\ldots(n+m)} = \\ = \frac{1}{m}\left[\frac{1}{1\cdot 2\ldots m} - \frac{1}{(N+1)(N+2)\ldots(N+m)}\right].$$

Следовательно,

$$S^{(m)} = \lim_{N\to\infty} S_N^{(m)} = \frac{1}{mm!}. \tag{8}$$

Используя идею Стирлинга, общий член ряда, определяемый формулой (6), представим в виде конечной суммы обратных факториалов

$$a_n = \frac{A_1}{n(n+1)} + \frac{A_2}{n(n+1)(n+2)} + \ldots + \frac{A_m}{n(n+1)\ldots(n+m)} + a_n^{(m)},$$

где $A_1, A_2, \ldots, A_m$ — неопределенные коэффициенты и $a_n^{(m)}$ — остаточный член. Подберем коэффициенты $A_i$ $(i = 1,\ 2,\ \ldots,\ m)$ так, чтобы

$$a_n^{(m)} = O\left(\frac{1}{n^{2+m}}\right).$$

---

*) Говорят, что $a_n$ есть бесконечно малая порядка не меньше $m$ относительно $\frac{1}{n}$:

$$a_n = O\left(\frac{1}{n^m}\right),$$

если

$$\lim_{n\to\infty} \frac{a_n}{\left(\frac{1}{n}\right)^m} = c \neq \infty$$

Если при этом $c \neq 0$, то $a_n$ — бесконечно малая точно порядка $m$ относительно $\frac{1}{n}$.

Для этого достаточно коэффициенты $A_i$ последовательно определять по формулам

$$\left.\begin{aligned}
A_1 &= \lim_{n\to\infty} n(n+1)a_n,\\
A_2 &= \lim_{n\to\infty}\left[a_n - \frac{A_1}{n(n+1)}\right] n(n+1)(n+2),\\
&\dots\dots\dots\dots\dots\dots\dots\dots\\
A_m &= \lim_{n\to\infty}\left[a_n - \sum_{i=1}^{m-1}\frac{A_i}{n(n+1)\dots(n+i)}\right] n(n+1)\dots(n+m).
\end{aligned}\right\} \tag{9}$$

Согласно общей схеме за вспомогательный ряд (2) примем:

$$B=\sum_{n=1}^{\infty} b_n = \sum_{n=1}^{\infty}\left[\frac{A_1}{n(n+1)} + \frac{A_2}{n(n+1)(n+2)} + \dots + \frac{A_m}{n(n+1)\dots(n+m)}\right] =$$

$$= A_1 S^{(1)} + A_2 S^{(2)} + \dots + A_m S^{(m)} = \frac{A_1}{1\cdot 1!} + \frac{A_2}{2\cdot 2!} + \dots + \frac{A_m}{m\cdot m!}, \tag{10}$$

причем очевидно, что

$$\lim_{n\to\infty}\frac{a_n}{b_n} = 1$$

и

$$S=\sum_{n=1}^{\infty} a_n = B + \sum_{n=1}^{\infty} a_n^{(m)}. \tag{11}$$

Так как быстрая сходимость дополнительного ряда $\sum\limits_{n=1}^{\infty} a_n^{(m)}$ обнаруживается, вообще говоря, лишь при достаточно большом $n$, то практически удобно выполнять указанное преобразование, начиная лишь с некоторого члена ряда $a_{p+1}$. Полагая

$$S=\sum_{n=1}^{p} a_n + \sum_{n=p+1}^{\infty} a_n = S_p + \sum_{n=p+1}^{\infty} a_n,$$

будем иметь:

$$S = S_p + \sum_{n=p+1}^{\infty}\left[\frac{A_1}{n(n+1)} + \frac{A_2}{n(n+1)(n+2)} + \dots\right.$$

$$\left.\dots + \frac{A_m}{n(n+1)\dots(n+m)} + a_n^{(m)}\right] = S_p + A_1\sum_{n=p+1}^{\infty}\left(\frac{1}{n} - \frac{1}{n+1}\right) +$$

$$+ \frac{A_2}{2}\sum_{n=p+1}^{\infty}\left[\frac{1}{n(n+1)} - \frac{1}{(n+1)(n+2)}\right] + \dots$$

$$\ldots+\frac{A_m}{m}\sum_{n=p+1}^{\infty}\left[\frac{1}{n(n+1)\ldots(n+m-1)}-\frac{1}{(n+1)\ldots(n+m)}\right]+\sum_{n=p+1}^{\infty}a_n^{(m)}=$$

$$=S_p+A_1\cdot\frac{1}{p+1}+\frac{A_2}{2}\cdot\frac{1}{(p+1)(p+2)}+\ldots$$

$$\ldots+\frac{A_m}{m}\cdot\frac{1}{(p+1)\ldots(p+m)}+\sum_{n=p+1}^{\infty}a_n^{(m)}.$$

В частности, при $m\to\infty$, учитывая что $a_n^{(m)}\rightrightarrows 0$, получим *разложение Стирлинга*

$$\sum_{n=1}^{\infty}a_n=\sum_{n=1}^{p}a_n+A_1\cdot\frac{1}{p+1}+\frac{A_2}{2}\cdot\frac{1}{(p+1)(p+2)}+\ldots$$

$$\ldots+\frac{A_m}{m}\cdot\frac{1}{(p+1)(p+2)\ldots(p+m)}+\ldots$$

Пример. Найти сумму ряда

$$S=\sum_{n=1}^{\infty}\frac{1}{n^2+1} \tag{12}$$

с точностью до 0,001.

Решение. Положим:

$$\frac{1}{n^2+1}=\frac{A_1}{n(n+1)}+\frac{A_2}{n(n+1)(n+2)}+a_n^{(2)}.$$

Имеем:

$$A_1=\lim_{n\to\infty}\frac{n(n+1)}{n^2+1}=1;$$

$$A_2=\lim_{n\to\infty}\left[\frac{1}{n^2+1}-\frac{1}{n(n+1)}\right]n(n+1)(n+2)=\lim_{n\to\infty}\frac{(n-1)(n+2)}{n^2+1}=1.$$

Следовательно,

$$a_n^{(2)}=\frac{1}{n^2+1}-\frac{1}{n(n+1)}-\frac{1}{n(n+1)(n+2)}=$$

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

На основании формул (10) и (11) получим:

$$S=\frac{1}{1\cdot 1!}+\frac{1}{2\cdot 2!}+\sum_{n=1}^{\infty}\frac{n-3}{n(n+1)(n+2)(n^2+1)}. \tag{12'}$$

Так как при $n\geqslant 3$ имеем

$$\frac{n-3}{n(n+1)(n+2)(n^2+1)}\leqslant\frac{1}{n^4},$$

то

$$\rho_N = \sum_{n=N+1}^{\infty} \frac{n-3}{n(n+1)(n+2)(n^2+1)} < \int_N^{\infty} \frac{dx}{x^4} = \frac{1}{3N^3} < \frac{1}{2} \cdot 0{,}001.$$

Отсюда следует, что число слагаемых в сумме (12′) можно взять $N=10$, причем эти слагаемые нужно вычислять с четырьмя десятичными знаками в узком смысле. Таким образом, имеем:

$$S \approx 1{,}25 + (-0{,}1667) + (-0{,}0083) + 0 + 0{,}0005 + 0{,}0004 + \\ + 0{,}0002 + 0{,}0002 + 0{,}0001 + 0{,}0001 + 0{,}0001 = 1{,}0766,$$

причем, приняв во внимание, что сумма четырех первых слагаемых точная, для абсолютной погрешности результата получим оценку

$$\Delta < \frac{1}{3} \cdot 10^{-3} + 7 \cdot \frac{1}{2} \cdot 10^{-4} < 0{,}7 \cdot 10^{-3}.$$

Отсюда, округляя, находим значение

$$S \approx 1{,}077$$

с предельной абсолютной погрешностью

$$\overline{\Delta} = 0{,}7 \cdot 10^{-3} + 0{,}4 \cdot 10^{-3} = 1{,}1 \cdot 10^{-3}.$$

Заметим, что для остатка данного ряда (12) имеем оценку

$$R_N < \int_N^{\infty} \frac{dx}{x^2+1} < \int_N^{\infty} \frac{dx}{x^2} = \frac{1}{N} \leqslant \frac{1}{2} \cdot 0{,}001.$$

Отсюда $N \geqslant 2000$, т. е. без преобразования для достижения той же точности нужно взять примерно 2000 членов ряда.

Замечание. Для приближенного вычисления суммы ряда (1) с общим членом (6) можно использовать также ряды

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}; \quad \sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{\pi^4}{90}; \quad \sum_{n=1}^{\infty} \frac{1}{n^6} = \frac{\pi^6}{945} \text{ и т. д.}$$

Вообще говоря,

$$\sum_{n=1}^{\infty} \frac{1}{n^{2p}} = \frac{(-1)^{p-1}}{2} \cdot \frac{B_{2p}(2\pi)^{2p}}{(2p)!},$$

где $B_n$ $(n=1, 2, \ldots)$ — числа Бернулли [5], [6], определяемые символической формулой

$$(B+1)^n - B^n = 0,$$

в которой после развертывания по биному Ньютона полагаем $B^n = B_n$. В частности, имеем:

$$B_2 = \frac{1}{6}; \quad B_4 = -\frac{1}{30}; \quad B_6 = \frac{1}{42}; \quad B_8 = -\frac{1}{30}; \quad B_{10} = \frac{5}{66}$$

(см. гл. XVI, § 11).

## § 2. Улучшение сходимости степенных рядов методом Эйлера — Абеля

Рассмотрим сходящийся степенной ряд

$$f(x) = \sum_{n=0}^{\infty} a_n x^n, \tag{1}$$

где $f(x)$ — сумма ряда.

Пусть радиус сходимости $R$ ряда (1) конечен и отличен от нуля. Не нарушая общности рассуждения, можно считать, что $R = 1$ *).

Запишем ряд (1) в следующем виде:

$$f(x) = a_0 + x\varphi(x), \tag{2}$$

где

$$\varphi(x) = \sum_{n=1}^{\infty} a_n x^{n-1} = \sum_{n=0}^{\infty} a_{n+1} x^n. \tag{3}$$

Умножая обе части равенства (3) на бином $1 - x$, получим:

$$(1-x)\varphi(x) = \sum_{n=0}^{\infty} a_{n+1} x^n - \sum_{n=0}^{\infty} a_{n+1} x^{n+1}. \tag{4}$$

Полагая во второй сумме $n + 1 = m$ и учитывая, что сумма не зависит от обозначения индекса суммирования, будем иметь:

$$\sum_{n=0}^{\infty} a_{n+1} x^{n+1} = \sum_{m=1}^{\infty} a_m x^m = \sum_{n=1}^{\infty} a_n x^n.$$

Поэтому

$$(1-x)\varphi(x) = \sum_{n=0}^{\infty} a_{n+1} x^n - \sum_{n=1}^{\infty} a_n x^n =$$

$$= a_0 + \sum_{n=0}^{\infty} (a_{n+1} - a_n) x^n = a_0 + \sum_{n=0}^{\infty} \Delta a_n x^n,$$

где

$$\Delta a_n = a_{n+1} - a_n \qquad (n = 0, 1, 2, \ldots)$$

---

*) В самом деле, если $0 < R < \infty$ и $R \neq 1$, то, полагая $t = \frac{x}{R}$, получим степенной ряд относительно переменной $t$ с радиусом сходимости $\rho = 1$.

— *конечные разности первого порядка* коэффициентов $a_n$ (подробнее о конечных разностях см. гл. XIV, § 1). Следовательно, из формул (3) и (4) выводим:

$$\varphi(x)=\sum_{n=0}^{\infty} a_{n+1}x^n=\frac{a_0}{1-x}+\frac{1}{1-x}\sum_{n=0}^{\infty}\Delta a_n x^n$$

и, значит,

$$f(x)=a_0+\frac{a_0x}{1-x}+\frac{x}{1-x}\sum_{n=0}^{\infty}\Delta a_n x^n=$$

$$=\frac{a_0}{1-x}+\frac{x}{1-x}\sum_{n=0}^{\infty}\Delta a_n x^n,$$

т. е.

$$\sum_{n=0}^{\infty} a_n x^n=\frac{a_0}{1-x}+\frac{x}{1-x}\sum_{n=0}^{\infty}\Delta a_n x^n. \tag{5}$$

Указанное преобразование степенного ряда называется *преобразованием Эйлера—Абеля.* Аналогично, применив преобразование Эйлера—Абеля к степенному ряду $\sum\limits_{n=0}^{\infty}\Delta a_n x^n$, находим:

$$\sum_{n=0}^{\infty}\Delta a_n x^n=\frac{\Delta a_0}{1-x}+\frac{x}{1-x}\sum_{n=0}^{\infty}\Delta^2 a_n x^n,$$

где

$$\Delta^2 a_n=\Delta(\Delta a_n)=\Delta a_{n+1}-\Delta a_n$$

— *конечные разности второго порядка* коэффициентов $a_n$. Отсюда на основании формулы (5) получаем:

$$\sum_{n=0}^{\infty} a_n x^n=\frac{a_0}{1-x}+\frac{x}{1-x}\left(\frac{\Delta a_0}{1-x}+\frac{x}{1-x}\sum_{n=0}^{\infty}\Delta^2 a_n x^n\right)=$$

$$=\frac{a_0}{1-x}+\frac{x\,\Delta a_0}{(1-x)^2}+\left(\frac{x}{1-x}\right)^2\sum_{n=0}^{\infty}\Delta^2 a_n x^n.$$

Повторяя последовательно $p$ раз преобразование Эйлера—Абеля, будем иметь:

$$\sum_{n=0}^{\infty} a_n x^n=\frac{a_0}{1-x}+\frac{x\,\Delta a_0}{(1-x)^2}+\ldots+\frac{x^{p-1}\Delta^{p-1}a_0}{(1-x)^p}+$$

$$+\left(\frac{x}{1-x}\right)^p\sum_{n=0}^{\infty}\Delta^p a_n x^n,$$

где

$$\Delta^p a_n = \Delta^{p-1} a_{n+1} - \Delta^{p-1} a_n \qquad (n=0,\ 1,\ 2,\ \ldots)$$

— *конечные разности p-го порядка* коэффициентов $a_n$, а $\Delta^k a_0$ ($k=0$, 1, 2, ...) — последовательные конечные разности коэффициентов $a_n$ при $n=0$. Таким образом,

$$f(x) = \sum_{k=0}^{p-1} \Delta^k a_0 \frac{x^k}{(1-x)^{k+1}} + \left(\frac{x}{1-x}\right)^p \sum_{n=0}^{\infty} \Delta^p a_n x^n, \tag{6}$$

где положено $\Delta^0 a_0 = a_0$. Формулу (6) выгодно применять тогда, когда конечные разности $\Delta^p a_n$ при $n \to \infty$ имеют более высокий порядок убывания, чем коэффициенты $a_n$. Это обстоятельство встречается нередко. Например, если $a_n = \frac{1}{n}$, то получим:

$$\Delta a_n = \frac{1}{n+1} - \frac{1}{n} = -\frac{1}{n(n+1)},$$

т. е. здесь $\Delta a_n$ при $n \to \infty$ убывают быстрее, чем $a_n$.

В частности, если $a_n = P(n)$, где $P(n)$ — целый полином степени $p-1$, то формула (6) дает в конечном виде сумму ряда

$$\sum_{n=0}^{\infty} P(n)\, x^n = \sum_{k=0}^{p-1} \Delta^k P(0) \frac{x^k}{(1-x)^{k+1}} \qquad (|x| < 1), \tag{7}$$

так как $\Delta^p P(n) = 0$.

Формула (6) теряет смысл при $x=1$. Применительно к этому случаю можно видоизменить преобразование Эйлера — Абеля. Полагая $x = -t$, будем иметь:

$$f(x) = \sum_{n=0}^{\infty} a_n (-t)^n = \sum_{n=0}^{\infty} (-1)^n a_n t^n =$$
$$= \sum_{k=0}^{p-1} \Delta^k [(-1)^n a_n]_{n=0} \frac{t^k}{(1-t)^{k+1}} + \left(\frac{t}{1-t}\right)^p \sum_{n=0}^{\infty} \Delta^p [(-1)^n a_n]\, t^n.$$

Возвращаясь к прежней переменной, получим:

$$f(x) = \sum_{k=0}^{p-1} (-1)^k \Delta^k [(-1)^n a_n]_{n=0} \frac{x^k}{(1+x)^k} +$$
$$+ \left(\frac{x}{1+x}\right)^p \sum_{n=0}^{\infty} (-1)^{n+p} \Delta^p [(-1)^n a_n]\, x^n. \tag{8}$$

Формула (8) имеет смысл и при $x=1$.

Пример 1. Найти с точностью до 0,001 сумму ряда

$$f(x)=\sum_{n=0}^{\infty}\frac{x^n}{(n+1)(n+2)} \tag{9}$$

при $x=-1$.

Решение. Применим преобразование Эйлера два раза ($p=2$). Имеем:

$$a_n=\frac{1}{(n+1)(n+2)};$$

$$\Delta a_n=a_{n+1}-a_n=\frac{1}{(n+2)(n+3)}-\frac{1}{(n+1)(n+2)}=$$
$$=-\frac{2}{(n+1)(n+2)(n+3)};$$

$$\Delta^2 a_n=\Delta a_{n+1}-\Delta a_n=-\frac{2}{(n+2)(n+3)(n+4)}+$$
$$+\frac{2}{(n+1)(n+2)(n+3)}=\frac{6}{(n+1)(n+2)(n+3)(n+4)}.$$

Следовательно,

$$a_0=\frac{1}{1\cdot 2};\quad \Delta a_0=-\frac{2}{1\cdot 2\cdot 3}.$$

Отсюда на основании формулы (6) получаем:

$$f(-1)=\frac{1}{1\cdot 2}\cdot\frac{1}{2}+\frac{2}{1\cdot 2\cdot 3}\cdot\frac{1}{4}+$$
$$+\left(-\frac{1}{2}\right)^2\sum_{n=0}^{\infty}\frac{6}{(n+1)(n+2)(n+3)(n+4)}(-1)^n=$$
$$=\frac{1}{4}+\frac{1}{12}+\frac{3}{2}\cdot\frac{1}{24}-\frac{3}{2}\cdot\frac{1}{120}+\frac{3}{2}\cdot\frac{1}{360}-\frac{3}{2}\cdot\frac{1}{840}+$$
$$+\frac{3}{2}\cdot\frac{1}{1680}-\frac{3}{2}\cdot\frac{1}{3024}+\frac{3}{2}\cdot\frac{1}{5040}-\ldots \tag{10}$$

Ряд (10) — знакочередующийся с монотонно убывающими по модулю членами. Поэтому, если мы остановимся на члене

$$\frac{3}{2}\cdot\frac{1}{3024}=\frac{1}{2016},$$

то остаток ряда $R$ по модулю не будет превышать первого отброшенного члена:

$$|R|<\frac{3}{2}\cdot\frac{1}{5040}=\frac{1}{3360}<3\cdot 10^{-4}.$$

Таким образом, беря два запасных знака, имеем:

$$f(-1)=0{,}25000+0{,}08333+0{,}06250-0{,}01250+0{,}00417-$$
$$-0{,}00179+0{,}00089-0{,}00050=0{,}38610$$

с абсолютной погрешностью

$$\Delta<5\cdot\frac{1}{2}\cdot10^{-5}+3\cdot10^{-4}<4\cdot10^{-4}.$$

Округляя найденное число до трех знаков, получим приближенное значение $f(-1)=0{,}386$ с предельной абсолютной погрешностью

$$\Delta<4\cdot10^{-4}+1\cdot10^{-4}=\frac{1}{2}\cdot10^{-3}.$$

Точное значение суммы есть:

$$f(-1)=2\ln2-1=0{,}38630\ldots$$

Заметим, что если непосредственно вычислять число $f(-1)$, пользуясь рядом (9), то для достижения указанной точности пришлось бы взять примерно сорок пять членов этого ряда.

Пример 2. Найти сумму ряда

$$S(x)=\sum_{n=0}^{\infty}(n^2+n+1)\,x^n.$$

Решение. Имеем:

$$P(n)=n^2+n+1.$$

Составим таблицу 12.

Таблица 12

**Таблица конечных разностей**

| $n$ | $P(n)$ | $\Delta P(n)$ | $\Delta^2P(n)$ |
|---|---|---|---|
| 0 | 1 | 2 | 2 |
| 1 | 3 | 4 | |
| 2 | 7 | | |

По формуле (7) получаем:

$$S(x)=\frac{1}{1-x}+\frac{2x}{(1-x)^2}+\frac{2x^2}{(1-x)^3}$$

при $|x|<1$.

## § 3. Оценки коэффициентов Фурье

Тригонометрическим рядом Фурье данной функции $f(x)$ $(-\pi < x < \pi)$ *) называется ряд

$$\frac{a_0}{2}+\sum_{n=1}^{\infty}(a_n\cos nx+b_n\sin nx), \tag{1}$$

коэффициенты которого $a_n$, $b_n$ (*коэффициенты Фурье* функции $f(x)$) вычисляются по формулам

$$a_n=\frac{1}{\pi}\int_{-\pi}^{\pi}f(x)\cos nx\,dx \qquad (n=0,\ 1,\ \ldots), \tag{2}$$

$$b_n=\frac{1}{\pi}\int_{-\pi}^{\pi}f(x)\sin nx\,dx \qquad (n=1,\ 2,\ \ldots). \tag{2'}$$

Достаточным условием существования ряда Фурье функции $f(x)$ является интегрируемость этой функции на отрезке $[-\pi, \pi]$. В этом случае коэффициенты Фурье (2) и (2′) имеют определенные конечные значения.

Может оказаться, что полученный ряд Фурье расходится или сходится к другой функции. Приведем без доказательства [1], [7] условия, при выполнении которых тригонометрический ряд Фурье сходится к функции $f(x)$ во всех точках непрерывности последней.

Теорема сходимости. *Если функция $f(x)$ кусочно непрерывна и кусочно дифференцируема на отрезке $[-\pi, \pi]$, то ее ряд Фурье сходится на всей числовой оси и сумма его $S(x)$ есть периодическая функция с периодом $2\pi$, равная*

$$S(x_0)=\frac{f(x_0-0)+f(x_0+0)}{2} \tag{3}$$

*в любой точке $x_0\in(-\pi, \pi)$ и $S(\pm\pi)=2^{-1}[f(-\pi+0)+f(\pi-0)]$.*

В частности, $S(x_0)=f(x_0)$, если в точке $x=x_0$ функция непрерывна, т. е. если $f(x_0-0)=f(x_0+0)=f(x_0)$.

Если, сверх того, функция $f(x)$ — периодическая с периодом $2\pi$, то ее ряд Фурье сходится для каждого значения $x_0$ и имеет сумму (3).

При выполнении условий теоремы сходимости очевидно, что $a_n\to 0$ и $b_n\to 0$ при $n\to\infty$. Дадим более точные оценки коэффи-

---

*) Для простоты формулировок мы рассматриваем функцию, определенную на отрезке $[-\pi, \pi]$. Общий случай функции $\varphi(t)$, заданной на отрезке $[a, b]$, может быть сведен к нашему при помощи линейной замены

$$t=\frac{b+a}{2}+\frac{b-a}{2\pi}x.$$

циентов Фурье, накладывая известные ограничения на поведение функции $f(x)$.

Определение. *Говорят, что функция $f(x)$, заданная на отрезке $[-\pi, \pi]$, принадлежит классу периодичности $\tilde{C}^{(m)}$, если:*

1) *$f(x)$ непрерывна на отрезке $[-\pi, \pi]$ вместе со своими производными до $m$-го порядка включительно;*

2) *$f^{(k)}(-\pi+0)=f^{(k)}(\pi-0)$ для $k=0, 1, 2, \ldots, m$, т. е. на концах отрезка $[-\pi, \pi]$ должны совпадать значения функции $f(x)$ и ее $m$ первых производных.*

Из условий 1) и 2) следует, что периодическое продолжение функции $f(x)$ принадлежит классу $C^{(m)}(-\infty, +\infty)$.

Лемма. *Если функция $f(x)$ принадлежит классу периодичности $\tilde{C}^{(m)}$ на отрезке $[-\pi, \pi]$ (короче, $f(x) \in \tilde{C}^{(m)}[-\pi, \pi]$), то ее коэффициенты Фурье $a_n$ и $b_n$ есть бесконечно малые при $n \to \infty$ порядка выше $m$ относительно $\frac{1}{n}$, т. е.*

$$a_n = o\left(\frac{1}{n^m}\right); \quad b_n = o\left(\frac{1}{n^m}\right)^{*)}.$$

Доказательство. Проинтегрируем по частям $m$ раз правые члены следующих равенств:

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos nx\,dx \qquad (n=0, 1, \ldots), \tag{4}$$

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin nx\,dx \qquad (n=1, 2, \ldots). \tag{4'}$$

Полагая $u=f(x)$ и $dv=\cos nx\,dx$, находим $du=f'(x)\,dx$ и $v=\frac{1}{n}\sin nx$. Следовательно, по формуле интегрирования по частям имеем:

$$a_n = \frac{1}{\pi}\left[\frac{1}{n}f(x)\sin nx\right]_{-\pi}^{\pi} - \frac{1}{\pi n}\int_{-\pi}^{\pi} f'(x)\sin nx\,dx =$$

$$= \frac{1}{\pi n}\int_{-\pi}^{\pi} f'(x)\cos\left(\frac{\pi}{2}+nx\right)dx.$$

---

*) Запись $a_n = o\left(\frac{1}{n^m}\right)$ обозначает, что $\lim\limits_{n\to\infty}\left(\frac{a_n}{\frac{1}{n^m}}\right)=0$.

Применяя еще раз интегрирование по частям и учитывая, что $f'(-\pi)=f'(\pi)$, получим:

$$a_n=\frac{1}{\pi n}\left\{\left[\frac{1}{n}f'(x)\sin\left(\frac{\pi}{2}+nx\right)\Big|_{-\pi}^{\pi}+\right.\right.$$

$$\left.\left.+\frac{1}{n}\int_{-\pi}^{\pi}f''(x)\cos\left(\frac{\pi}{2}\cdot 2+nx\right)dx\right]\right\}=$$

$$=\frac{1}{\pi n^2}\int_{-\pi}^{\pi}f''(x)\cos\left(\frac{\pi}{2}\cdot 2+nx\right)dx$$

и т. д.

После $m$-кратного интегрирования по частям в формулах (4) и (4′) будем иметь:

$$a_n=\frac{1}{\pi n^m}\int_{-\pi}^{\pi}f^{(m)}(x)\cos\left(\frac{\pi}{2}\cdot m+nx\right)dx.$$

Аналогично

$$b_n=\frac{1}{\pi n^m}\int_{-\pi}^{\pi}f^{(m)}(x)\sin\left(\frac{\pi}{2}\cdot m+nx\right)dx.$$

Интегралы

$$\varepsilon_n=\frac{1}{\pi}\int_{-\pi}^{\pi}f^{(m)}(x)\cos\left(\frac{\pi}{2}\cdot m+nx\right)dx$$

и

$$\varepsilon_n'=\frac{1}{\pi}\int_{-\pi}^{\pi}f^{(m)}(x)\sin\left(\frac{\pi}{2}\cdot m+nx\right)dx$$

с точностью до знака являются коэффициентами Фурье непрерывной по условию функции $f^{(m)}(x)$. Как известно, коэффициенты Фурье непрерывной функции независимо от того, сходится или нет ее ряд Фурье, стремятся к нулю при неограниченном возрастании их номера *). Поэтому

$$\varepsilon_n\to 0 \quad \text{и} \quad \varepsilon_n'\to 0 \quad \text{при} \quad n\to\infty.$$

---

*) Это следует из того, что для любой кусочно непрерывной функции $f(x)$ с коэффициентами Фурье $a_n$ и $b_n$ ($n=0, 1, 2, \ldots$) имеет место *неравенство Бесселя* [7]

$$\frac{a_0^2}{2}+\sum_{n=1}^{\infty}(a_n^2+b_n^2)\leqslant\frac{1}{\pi}\int_{-\pi}^{\pi}f^2(x)\,dx.$$

Следовательно, ряд $\sum_{n=1}^{\infty}(a_n^2+b_n^2)$ сходится и $a_n\to 0$; $b_n\to 0$ при $n\to\infty$.

Но так как

$$a_n = \frac{\varepsilon_n}{n^m} \quad \text{и} \quad b_n = \frac{\varepsilon_n'}{n^m},$$

то коэффициенты Фурье $a_n$ и $b_n$ функции $f(x)$ являются бесконечно малыми более высокого порядка по сравнению с $\frac{1}{n^m}$:

$$a_n = o\left(\frac{1}{n^m}\right), \quad b_n = o\left(\frac{1}{n^m}\right).$$

Этот результат был положен А. Н. Крыловым в основу метода улучшения сходимости рядов Фурье.

Замечание. Если $f^{(m)}(x)$ удовлетворяет условиям теоремы сходимости, то легко доказать, что

$$\varepsilon_n = O\left(\frac{1}{n}\right) \quad \text{и} \quad \varepsilon_n' = O\left(\frac{1}{n}\right).$$

В этом случае для коэффициентов Фурье функции $f(x)$ получается лучшая оценка:

$$a_n = O\left(\frac{1}{n^{m+1}}\right) \quad \text{и} \quad b_n = O\left(\frac{1}{n^{m+1}}\right).$$

## § 4. Улучшение сходимости тригонометрических рядов Фурье методом А. Н. Крылова

Пусть функция $f(x)$ на отрезке $[-\pi, \pi]$ кусочно непрерывна и имеет кусочно непрерывные производные $f^{(i)}(x)$ $(i = 1, 2, \ldots, m)$ до $m$-го порядка включительно. Тогда в силу теоремы сходимости (§ 3) функцию $f(x)$ во всех ее точках непрерывности можно представить тригонометрическим рядом Фурье

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx), \tag{1}$$

где $a_n$ и $b_n$ — коэффициенты Фурье, определяемые формулами (2) и (2′) из § 3. В общем случае коэффициенты $a_n$ и $b_n$ ряда (1) медленно стремятся к нулю, и практически пользоваться этим рядом затруднительно, а тем более недопустимо почленно дифференцировать ряд (1), что бывает нужно при решении некоторых задач, в частности при применении метода Фурье.

Идея метода А. Н. Крылова [8] состоит в том, что из функции $f(x)$ выделяется элементарная функция $g(x)$ (обычно представляющая собой кусочно полиномиальную функцию), имеющая те же разрывы, что и функция $f(x)$, причем ее производные $g^{(i)}(x)$ $(i = 1,$

2, ..., $m$) до $m$-го порядка включительно обладают такими же разрывами, как и соответствующие производные $f^{(i)}(x)$ данной функции $f(x)$, и, сверх того, такая, что

$$f^{(i)}(-\pi+0)-g^{(i)}(-\pi+0)= \\ =f^{(i)}(\pi-0)-g^{(i)}(\pi-0) \qquad (i=0,\ 1,\ 2,\ \ldots,\ m).$$

В таком случае разность

$$\varphi(x)=f(x)-g(x)$$

будет принадлежать классу периодичности $\tilde{C}^{(m)}$.

Обозначая через $\alpha_n$ и $\beta_n$ ($n=0,\ 1,\ 2,\ \ldots$) коэффициенты Фурье функции $\varphi(x)$, получим:

$$f(x)=g(x)+\left[\frac{\alpha_0}{2}+\sum_{n=1}^{\infty}(\alpha_n\cos nx+\beta_n\sin nx)\right], \tag{2}$$

где $\alpha_n$ и $\beta_n$ — бесконечно малые при $n\to\infty$ порядка выше, чем $m$, относительно $\frac{1}{n}$, т. е. ряд (2) будет, вообще говоря, быстро сходящимся. Этот ряд можно дифференцировать почленно по меньшей мере $m-2$ раза.

Покажем, как по заданной функции $f(x)$ практически строится вспомогательная функция $g(x)$ [9]. Для этого рекуррентным спосо-

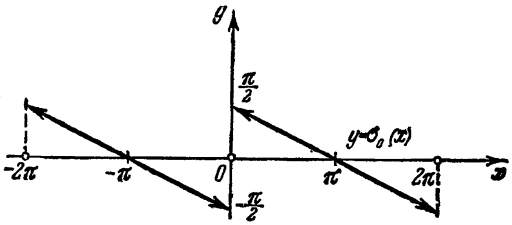


Рис. 46.

бом построим на отрезке $[-2\pi,\ 2\pi]$ последовательность функций $\sigma_0(x)$, $\sigma_1(x)$, ..., $\sigma_m(x)$, обладающих следующим свойством:

$$\sigma_k^{(k)}(+0)-\sigma_k^{(k)}(-0)=\pi \tag{3}$$

($k=0,\ 1,\ 2,\ \ldots,\ m$), причем таких, что производные $\sigma_k^{(j)}(x)$ ($j=0,\ 1,\ \ldots,\ k-1$) непрерывны на отрезке $[-2\pi,\ 2\pi]$.

Функцию $\sigma_0(x)$ определим следующим образом:

$$\sigma_0(x)=\begin{cases}\dfrac{-\pi-x}{2} & \text{при} \quad -2\pi<x<0,\\[2mm] \dfrac{\pi-x}{2} & \text{при} \quad 0<x<2\pi,\\[2mm] 0 & \text{при} \quad x=-2\pi,\ 0,\ 2\pi.\end{cases} \tag{4}$$

Ее график изображен на рис. 46. Эта функция нечетная, поэтому ее ряд Фурье содержит только синусы кратных дуг:

$$\sigma_0(x) = \sum_{n=1}^{\infty} b_n \sin nx,$$

где

$$b_n = \frac{2}{\pi}\int_0^{\pi} \frac{\pi - x}{2} \sin nx\, dx =$$

$$= \frac{2}{\pi}\left(-\frac{\pi - x}{2}\cdot\frac{\cos nx}{n}\Big|_0^{\pi} - \frac{1}{2n}\int_0^{\pi}\cos nx\, dx\right) =$$

$$= \frac{2}{\pi}\left(\frac{\pi}{2n} - \frac{1}{2n^2}\sin nx\Big|_0^{\pi}\right) = \frac{1}{n}.$$

Следовательно,

$$\sigma_0(x) = \frac{\sin x}{1} + \frac{\sin 2x}{2} + \ldots + \frac{\sin nx}{n} + \ldots \tag{5}$$

Очевидно, что функция $\sigma_0(x)$ имеет разрыв в точке $x = 0$ со скачком, равным $\pi$:

$$\sigma_0(+0) - \sigma_0(-0) = \frac{\pi}{2} - \left(-\frac{\pi}{2}\right) = \pi.$$

Поэтому функция

$$\psi(x) = \sigma_0(x - x_0) \qquad (-\pi \leqslant x \leqslant \pi;\ -\pi \leqslant x_0 \leqslant \pi)$$

имеет в точке $x_0$ такой же скачок, как и функция $\sigma_0(x)$:

$$\psi(x_0 + 0) - \psi(x_0 - 0) = \pi,$$

причем точка разрыва — единственная на отрезке $[-\pi, \pi]$.

Функцию $\sigma_1(x)$ определим формулой

$$\sigma_1(x) = c_1 + \int_0^x \sigma_0(x)\, dx, \tag{6}$$

где $c_1$ — некоторая постоянная.

Интегрируя почленно ряд (5), получим:

$$\sigma_1(x) = c_1 + \int_0^x \sum_{n=1}^{\infty} \frac{\sin nx}{n}\, dx = c_1 + \sum_{n=1}^{\infty}\frac{1}{n^2} - \sum_{n=1}^{\infty}\frac{\cos nx}{n^2}. \tag{7}$$

Постоянную $c_1$ подберем так, чтобы свободный член ряда (7) был равен нулю

$$c_1 + \sum_{n=1}^{\infty}\frac{1}{n^2} = 0.$$

Отсюда

$$c_1 = -\sum_{n=1}^{\infty} \frac{1}{n^2}.$$

Ряд $\sum_{n=1}^{\infty} \frac{1}{n^2}$, очевидно, является свободным членом ряда Фурье функции $\int_0^x \sigma_0(x)\,dx$. Отсюда, пользуясь формулой (4), имеем:

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{1}{\pi} \int_0^{\pi} dx \int_0^x \sigma_0(x)\,dx = \frac{1}{\pi} \int_0^{\pi} \left[\frac{\pi^2}{4} - \frac{(\pi - x)^2}{4}\right] dx =$$

$$= \frac{1}{\pi}\left(\frac{\pi^3}{4} - \frac{\pi^3}{12}\right) = \frac{\pi^2}{6}.$$

Поэтому

$$c_1 = -\frac{\pi^2}{6}.$$

Следовательно,

$$\sigma_1(x) = -\sum_{n=1}^{\infty} \frac{\cos nx}{n^2}, \tag{8}$$

причем

$$\sigma_1(x) = \begin{cases} \int_0^x \frac{\pi - x}{2}\,dx - \frac{\pi^2}{6} = \frac{\pi^2}{12} - \frac{(\pi - x)^2}{4} & \text{при} \quad 0 \leqslant x \leqslant 2\pi, \\ -\int_0^x \frac{\pi + x}{2}\,dx - \frac{\pi^2}{6} = \frac{\pi^2}{12} - \frac{(\pi + x)^2}{4} & \text{при} \quad -2\pi \leqslant x \leqslant 0. \end{cases}$$

График функции $\sigma_1(x)$ изображен на рис. 47. Функция $\sigma_1(x)$ непрерывна на отрезке $[-2\pi, 2\pi]$, но ее производная $\sigma_1'(x) = \sigma_0(x)$

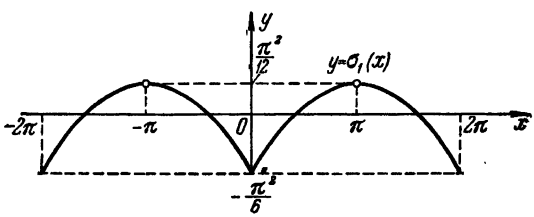


Рис 47.

имеет разрыв в точке $x = 0$, причем

$$\sigma_1'(+0) - \sigma_0'(-0) = \pi.$$

Таким же образом определяются функции

$$\sigma_2(x) = \int_0^x \sigma_1(x)\,dx + c_2;$$

$$\sigma_3(x) = \int_0^x \sigma_2(x)\,dx + c_3;$$

$$\cdots\cdots\cdots\cdots$$

$$\sigma_m(x) = \int_0^x \sigma_{m-1}(x)\,dx + c_m,$$

где произвольные постоянные $c_1, c_2, \ldots, c_m$ подбираются так, чтобы свободный член соответствующего ряда Фурье был равен нулю, т. е. постоянные $c_k$ $(k=1, 2, \ldots, m)$ последовательно находятся из условий

$$\int_0^\pi \left[c_k + \int_0^x \sigma_{k-1}(x)\,dx\right] dx = 0.$$

Функции $\sigma_k(x)$ $(k=1, 2, \ldots, m)$ и все производные до $(k-1)$-го порядка включительно непрерывны на отрезке $[-2\pi, 2\pi]$. При этом, так как $\sigma_k^{(k)}(x) = \sigma_0(x)$, то

$$\sigma_k^{(k)}(+0) - \sigma_k^{(k)}(-0) = \pi \quad (k=1, 2, \ldots, m),$$

т. е. производная $k$-го порядка функции $\sigma_k(x)$ имеет разрыв при $x=0$ со скачком $\pi$. Отсюда следует, что функция $\psi_k(x) = \sigma_k(x-x_0)$ $(-\pi \leqslant x \leqslant \pi)$, полученная сдвигом функции $\sigma_k(x)$, имеет разрыв лишь $k$-й производной в точке $x=x_0$:

$$\psi_k^{(k)}(x_0+0) - \psi_k^{(k)}(x_0-0) = \pi.$$

Пусть теперь

$x_1^{(0)}, x_2^{(0)}, \ldots, x_{k_0}^{(0)}$ — точки разрыва $f(x)$;
$x_1^{(1)}, x_2^{(1)}, \ldots, x_{k_1}^{(1)}$ — точки разрыва $f'(x)$;
. . . . . .
$x_1^{(m)}, x_2^{(m)}, \ldots, x_{k_m}^{(m)}$ — точки разрыва $f^{(m)}(x)$,

причем некоторые из этих точек могут повторяться.

Для соответствующих скачков функции и ее производных введем обозначения

$$f^{(l)}(x_j^{(l)}+0) - f^{(l)}(x_j^{(l)}-0) = h_j^{(l)}$$
$$(l=0, 1, \ldots, m;\ j=1, 2, \ldots, k_l).$$

Определим функцию $g(x)$ (*функция скачков*) формулой

$$g(x)=\sum_{s=1}^{s=k_0}\frac{h_s^{(0)}}{\pi}\sigma_0(x-x_s^{(0)})+\sum_{s=1}^{s=k_1}\frac{h_s^{(1)}}{\pi}\sigma_1(x-x_s^{(1)})+\ \ldots$$
$$\ldots+\sum_{s=1}^{s=k_m}\frac{h_s^{(m)}}{\pi}\sigma_m(x-x_s^{(m)}). \qquad (9)$$

Функция $g(x)$ обладает следующими свойствами:

1) в точках $x_1^{(0)}, x_2^{(0)}, \ldots, x_{k_0}^{(0)}$ функция $g(x)$ имеет разрывы, причем скачки ее в этих точках равны скачкам функции $f(x)$ в соответствующих точках:

$$g(x_j^{(0)}+0)-g(x_j^{(0)}-0)=\frac{h_j^{(0)}}{\pi}[\sigma_0(x_j-x_j+0)-\sigma_0(x_j-x_j-0)]=$$
$$=\frac{h_j^{(0)}}{\pi}\pi=h_j^{(0)};$$

2) производная $g^{(l)}(x)$ $(l=1, 2, \ldots, m)$ разрывна в точках $x_1^{(l)}, x_2^{(l)}, \ldots, x_{k_l}^{(l)}$, причем

$$g^{(l)}(x_j^{(l)}+0)-g^{(l)}(x_j^{(l)}-0)=$$
$$=\frac{h_j^{(l)}}{\pi}[\sigma_l(x_j^{(l)}-x_j^{(l)}+0)-\sigma_l(x_j^{(l)}-x_j^{(l)}-0)]=\frac{h_j^{(l)}}{\pi}\pi=h_j^{(l)},$$

т. е.

$$g^{(l)}(x_j+0)-g^{(l)}(x_j-0)=f^{(l)}(x_j+0)-f^{(l)}(x_j-0);$$

3) при $x\neq x_j^{(l)}$ функция $g(x)$ имеет непрерывные производные всех порядков.

Пусть

$$\varphi(x)=f(x)-g(x). \qquad (10)$$

В силу первого и второго свойств следует, что

$$\varphi^{(l)}(x_j^{(l)}+0)-\varphi^{(l)}(x_j^{(l)}-0)=0 \qquad (l=0, 1, 2, \ldots, m),$$

т. е.

$$\varphi(x)\in\tilde{C}^{(m)}[-\pi, \pi].$$

Таким образом, для разложения функции $f(x)$ можно воспользоваться быстро сходящимся рядом Фурье (2). Заметим, что,

пользуясь разложениями

$$\sigma_0(x-x_s^{(0)})=\sum_{n=1}^{\infty}\frac{\sin n(x-x_s^{(0)})}{n};$$

$$\sigma_1(x-x_s^{(1)})=-\sum_{n=1}^{\infty}\frac{\cos n\left(x-x_s^{(1)}\right)}{n^2};$$

$$\sigma_2(x-x_s^{(2)})=-\sum_{n=1}^{\infty}\frac{\sin n\left(x-x_s^{(2)}\right)}{n^3};$$

$$\cdots\cdots\cdots\cdots\cdots\cdots,$$

легко написать разложение в ряд Фурье функции $g(x)$. В итоге мы получим, что ряд Фурье функции $f(x)$ состоит: а) из медленно сходящейся части, элементарно суммирующейся к функции $g(x)$, и б) из быстро сходящегося остатка, представляющего ряд Фурье функции $\varphi(x)\in\tilde{C}^{(m)}[-\pi,\pi]$.

Замечание. Если на концах отрезка $[-\pi,\pi]$ предельные значения функции $f(x)$ или ее производных $f'(x),\ \ldots,\ f^{(k)}(x)$ $(k\leqslant m)$ не совпадают между собой, т. е.

$$f^{(l)}(-\pi+0)\neq f^{(l)}(\pi-0)\qquad(l=0,\ 1,\ 2,\ \ldots,\ k),$$

то точки $x=-\pi$ и $x=\pi$ следует считать точками разрыва функции $f(x)$ или соответственно производных $f^{(l)}(x)$.

Предполагая, что функция $f(x)$ периодически продолжена за пределы отрезка $[-\pi,\pi]$ с периодом $2\pi$, получаем, что скачок производных в точках $x=-\pi$ и $x=\pi$ один и тот же и равен

$$h^{(l)}=f^{(l)}(-\pi+0)-f^{(l)}(\pi-0).$$

В силу периодичности функции $\sigma_l(x)$ имеем:

$$\sigma_l(x+\pi)=\sigma_l(x-\pi),$$

причем функция $\sigma_l^{(l)}(x+\pi)$ на отрезке $[-\pi,\pi]$ допускает две точки разрыва ($x=-\pi$ и $x=\pi$) с одним и тем же скачком, равным $\pi$. Поэтому в формулу (9) нужно включить только одну концевую точку, например $x=-\pi$. В самом деле, на основании формулы (9) скачок производной $g^{(l)}(x)$ в точке $x=-\pi$ равен

$$g^{(l)}(-\pi+0)-g^{(l)}(-\pi-0)=\frac{h^{(l)}}{\pi}\left[\sigma^{(l)}(+0)-\sigma^{(l)}(-0)\right]=h^{(l)}.$$

В силу периодичности $g^{(l)}(x)$ эта производная имеет тот же скачок и при $x=\pi$. Следовательно, при составлении разности

$$f(x)-g(x)=\varphi(x),$$

где учтена лишь точка $x=-\pi$, снимается разрыв $l$-й производной функции $\varphi(x)$ как в точке $x=-\pi$, так и в точке $x=\pi$.

Пример. Методом А. Н. Крылова улучшить сходимость ряда Фурье функции (рис. 48а)

$$f(x)=\begin{cases} x^2+1 & \text{при} \quad -\pi<x<0, \\ x^2 & \text{при} \quad 0<x<\pi. \end{cases}$$

Решение. На отрезке $[-\pi, \pi]$ функция $f(x)$ в силу замечания имеет точки разрыва: $x_1=-\pi$; $x_2=0$; $x_3=\pi$. Вычисляя коэффициенты Фурье, получим:

$$a_0=1+\frac{2\pi^2}{3}; \qquad a_n=\frac{4}{n^2}(-1)^n; \qquad b_n=\begin{cases} -\dfrac{2}{\pi n} & \text{при } n \text{ нечетном;} \\ 0 & \text{при } n \text{ четном.} \end{cases}$$

Следовательно, ряд Фурье функции $f(x)$ имеет вид

$$f(x)=\frac{1}{2}+\frac{\pi^2}{3}+4\sum_{n=1}^{\infty}\frac{(-1)^n}{n^2}\cos nx-\frac{2}{\pi}\sum_{k=0}^{\infty}\frac{1}{2k+1}\sin(2k+1)x. \qquad (11)$$

Сходимость ряда (11) — плохая, так как коэффициенты $b_n=O\left(\frac{1}{n}\right)$ убывают медленно. Из функции $f(x)$ выделим функцию скачков $g(x)$ так, чтобы $\varphi(x)=[f(x)-g(x)]\in\tilde{C}^{(m)}[-\pi, \pi]$.

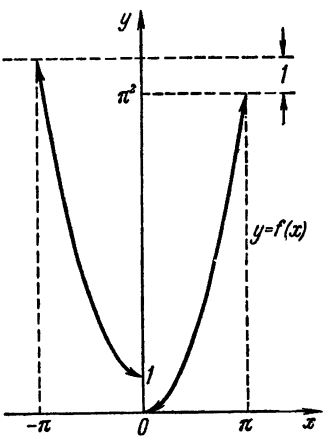


Рис. 48а.

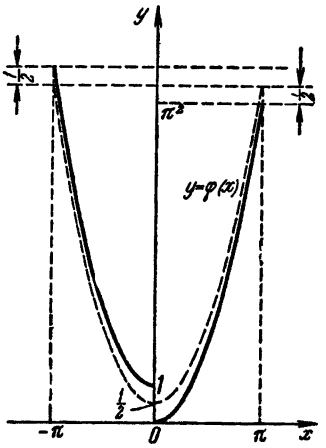


Рис. 48б.

Подсчитаем скачки $h_j^{(0)}$ нулевого рода в точках $x_j$ $(j=1, 2, 3)$:

$$h_1^{(0)}=f(-\pi+0)-f(\pi-0)=(\pi^2+1)-\pi^2=1;$$
$$h_2^{(0)}=f(+0)-f(-0)=0-1=-1;$$
$$h_3^{(0)}=h_1^{(0)}=1.$$

На основании формулы (9), учитывая замечание, получаем:

$$g(x)=\frac{1}{\pi}\cdot\sigma_0(x+\pi)-\frac{1}{\pi}\cdot\sigma_0(x)$$

или

$$g(x)=\frac{1}{\pi}\cdot\frac{\pi-(x+\pi)}{2}+\frac{1}{\pi}\cdot\frac{\pi+x}{2}=\frac{1}{2}$$

при $-\pi<x<0$ и

$$g(x)=\frac{1}{\pi}\cdot\frac{\pi-(x+\pi)}{2}-\frac{1}{\pi}\cdot\frac{\pi-x}{2}=-\frac{1}{2}$$

при $0<x<\pi$.

Вычитая из функции $f(x)$ функцию скачков $g(x)$, получаем функцию

$$\varphi(x)=x^2+\frac{1}{2},$$

непрерывную на отрезке $[-\pi, \pi]$ (рис. 48б). Так как

$$\sigma_0(x)=\sum_{n=1}^{\infty}\frac{\sin nx}{n}$$

и

$$\sigma_0(x+\pi)=\sum_{n=1}^{\infty}\frac{\sin n(x+\pi)}{n}=\sum_{n=1}^{\infty}\frac{(-1)^n}{n}\sin nx,$$

то

$$g(x)=\frac{1}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^n}{n}\sin nx-\frac{1}{\pi}\sum_{n=1}^{\infty}\frac{1}{n}\sin nx=$$

$$=\frac{1}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^n-1}{n}\sin nx=-\frac{2}{\pi}\sum_{k=0}^{\infty}\frac{\sin(2k+1)x}{2k+1}.$$

Следовательно,

$$f(x)=g(x)+\frac{1}{2}+\frac{\pi^2}{3}+4\sum_{n=1}^{\infty}\frac{(-1)^n}{n^2}\cos nx,$$

причем коэффициенты преобразованного ряда Фурье имеют порядок убывания $O\left(\frac{1}{n^2}\right)$.

Заметим, что если из функции $f(x)$ выделить функцию скачков $g(x)$ с точностью до разрывов производной, то остаток

будет тождественно равен нулю, т. е. мы получим точную сумму ряда (10).

Замечание. Метод А. Н. Крылова применим также к рядам Фурье периода $T=2l$. В самом деле, пусть функция $f(x)$ задана в основной области $a-l<x<a+l$. Произведя линейное преобразование

$$x=a+\frac{l}{\pi}t,$$

получим функцию $F(t)=f\left(a+\frac{l}{\pi}t\right)$ периода $2\pi$, определенную в стандартной области $-\pi<t<\pi$.

## § 5. Приближенное суммирование тригонометрических рядов

Пусть мы имеем сходящийся тригонометрический ряд

$$\sum_{n=0}^{\infty}(a_n\cos nx+b_n\sin nx)=S(x), \tag{1}$$

сумма которого $S(x)$ неизвестна. Требуется приближенно с некоторой наперед заданной точностью вычислить эту сумму.

Очевидно, что чем быстрее стремятся к нулю коэффициенты $a_n$ и $b_n$ ряда (1), тем меньше его членов нужно взять, чтобы обеспечить заданную точность. Поэтому, прежде чем приступить к подсчету суммы, желательно улучшить сходимость данного ряда. Для этого обычно пользуются следующим приемом: из данного ряда выделяют некоторый тригонометрический ряд, сумма которого $g(x)$ известна, так, чтобы оставшийся ряд

$$\sum_{n=0}^{\infty}(\alpha_n\cos nx+\beta_n\sin nx) \tag{2}$$

имел более быструю сходимость, чем исходный.

Если

$$g(x)=\sum_{n=0}^{\infty}(\bar{a}_n\cos nx+\bar{b}_n\sin nx),$$

то

$$S(x)=g(x)+\sum_{n=0}^{\infty}(\alpha_n\cos nx+\beta_n\sin nx), \tag{3}$$

где

$$\alpha_n=a_n-a_n \qquad (n=0,\ 1,\ 2,\ \ldots).$$

В простейших случаях для построения функции $g(x)$ можно использовать рассмотренные выше разложения:

$$\sum_{n=1}^{\infty}\frac{\sin nx}{n}=\sigma_0(x)=\frac{\pi-x}{2}\qquad(0<x<2\pi);$$

$$\sum_{n=1}^{\infty}\frac{\cos nx}{n^2}=-\sigma_1(x)=\frac{(\pi-x)^2}{4}-\frac{\pi^2}{12}\qquad(0\leqslant x\leqslant 2\pi);$$

$$\sum_{n=1}^{\infty}\frac{\sin nx}{n^3}=-\sigma_2(x)=\frac{2\pi^2x-3\pi x^2+x^3}{12}\qquad(0\leqslant x\leqslant 2\pi);$$

. . . . . .

Иногда полезны также разложения [7]

$$\sum_{n=1}^{\infty}\frac{\cos nx}{n}=-\ln\left(2\sin\frac{x}{2}\right)\qquad(0<x<2\pi);$$

$$\sum_{n=1}^{\infty}\frac{\sin nx}{n^2}=-\int_0^x\ln\left(2\sin\frac{x}{2}\right)dx\qquad(0\leqslant x\leqslant 2\pi);$$

$$\sum_{n=1}^{\infty}\frac{\cos nx}{n^3}=\int_0^x dx\int_0^x\ln\left(2\sin\frac{x}{2}\right)dx+\sum_{n=1}^{\infty}\frac{1}{n^3}\qquad(0\leqslant x\leqslant 2\pi),$$

где $\sum\limits_{n=1}^{\infty}\frac{1}{n^3}=1{,}202056903\ldots$

Пример. Найти сумму ряда

$$S(x)=\sum_{n=1}^{\infty}\frac{n}{n^2+1}\sin nx$$

с точностью до 0,001.

Решение. Коэффициенты ряда $b_n=\frac{n}{n^2+1}$ имеют порядок убывания $O\left(\frac{1}{n}\right)$, так как $\lim\limits_{n\to\infty}\left(b_n:\frac{1}{n}\right)=1$. Улучшим сходимость заданного ряда. Очевидно, что

$$\frac{n}{n^2+1}=\frac{n}{n^2}\left(\frac{1}{1+\frac{1}{n^2}}\right)=\frac{1}{n}\left(1-\frac{1}{n^2}+\frac{1}{n^4}-\ldots\right)=\frac{1}{n}-\frac{1}{n^3}+\gamma_n,$$

где

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

Тогда

$$\sum_{n=1}^{\infty}\frac{n}{n^2+1}\sin nx=\sum_{n=1}^{\infty}\frac{\sin nx}{n}-\sum_{n=1}^{\infty}\frac{\sin nx}{n^3}+\sum_{n=1}^{\infty}\gamma_n\sin nx.$$

Но

$$\sum_{n=1}^{\infty}\frac{\sin nx}{n}=\sigma_0(x) \quad \text{и} \quad \sum_{n=1}^{\infty}\frac{\sin nx}{n^3}=-\sigma_2(x).$$

Таким образом,

$$S(x)=\sigma_0(x)+\sigma_2(x)+\sum_{n=1}^{\infty}\gamma_n\sin nx,$$

где $\gamma_n=\frac{1}{n^3(n^2+1)}=O\left(\frac{1}{n^5}\right)$.

Пусть $N$ — число членов ряда $\sum_{n=1}^{\infty}\gamma_n\sin nx$, которые нужно взять, чтобы его остаток $R_N$ удовлетворял неравенству

$$|R_N|=\left|\sum_{N+1}^{\infty}\gamma_n\sin nx\right|<0{,}001.$$

Найдем число $N$. Имеем:

$$\left|\sum_{N+1}^{\infty}\frac{1}{n^3(n^2+1)}\sin nx\right|<\sum_{N+1}^{\infty}\frac{1}{n^5}<\int_N^{\infty}\frac{dx}{x^5}=\frac{1}{4N^4}.$$

Решая неравенство $\frac{1}{4N^4}<0{,}001$, получим, что достаточно взять $N=5$. Следовательно, с заданной точностью имеем:

$$S(x)=\frac{\pi-x}{2}-\frac{2\pi^2x-3\pi x^2+x^3}{12}+\sum_{n=1}^{5}\frac{\sin nx}{n^3(n^2+1)} \qquad (0<x<\pi).$$

## Литература к шестой главе


1. Г. М. Фихтенгольц, Основы математического анализа, т. II, Гостехиздат, 1956, гл. XV и XXIV.
2. А. Марков, Исчисление конечных разностей, Изд. 2, Матезис, 1910, гл. II.
3. Г. Салехов, Вычисление рядов, Гостехиздат, 1955, гл. I и III.
4. Я. С. Безикович, Исчисление конечных разностей, ЛГУ, 1939, гл. IX.
5. А. О. Гельфонд, Исчисление конечных разностей, Гостехиздат, 1952, гл. IV.
6. Валле-Пуссен, Курс анализа бесконечно малых, т. II, ГТТИ, 1933.
7. Г. П. Толстов, Ряды Фурье, Гостехиздат, 1951, гл. I—V
8. А. Н. Крылов, Лекции о приближенных вычислениях, Изд. 6, Гостехиздат, 1954, гл. V.
9. Л. В. Канторович, В. И. Крылов, Приближенные методы высшего анализа, Изд. 3, Гостехиздат, 1949, гл. I.

## ГЛАВА VII

# АЛГЕБРА МАТРИЦ

## § 1. Основные определения

Система $mn$ чисел (действительных или комплексных), расположенных в прямоугольной таблице из $m$ строк и $n$ столбцов,

$$A=\begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{m1} & a_{m2} & a_{m3} & & a_{mn} \end{bmatrix}, \tag{1}$$

называется *матрицей* (числовой). Строки и столбцы таблицы (1) называются *рядами матрицы*.

Числа $a_{ij}$ ($i=1, 2, \dots, m$; $j=1, 2, \dots, n$), составляющие данную матрицу, называются ее *элементами*. Здесь первый индекс $i$ обозначает номер строки элемента, а второй $j$ — номер его столбца.

Для матрицы (1) часто употребляется сокращенная запись

$$A=[a_{ij}] \qquad (i=1, 2, \dots, m;\ j=1, 2, \dots, n)$$

или

$$A=[a_{ij}]_{m,n},$$

причем говорят, что матрица $A$ имеет тип $m\times n$.

Если $m=n$, то матрица называется *квадратной порядка* $n$. Если же $m\neq n$, то матрица называется *прямоугольной*. В частности, матрица типа $1\times n$ называется *вектором-строкой*, а матрица типа $m\times 1$ — *вектором-столбцом*. Число (скаляр) можно рассматривать как матрицу типа $1\times 1$. Квадратная матрица вида

$$A=\begin{bmatrix} \alpha_1 & 0 & 0 & \dots & 0 \\ 0 & \alpha_2 & 0 & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \dots & \alpha_n \end{bmatrix} \tag{2}$$

называется *диагональной* и обозначается кратко так: $[\alpha_1, \alpha_2, \dots, \alpha_n]$.

В случае, если $\alpha_i = 1$ $(i = 1, 2, \ldots, n)$, то матрица (2) называется *единичной* и обозначается обычно буквой $E$, т. е.

$$E = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix}.$$

Введя *символ Кронекера*

$$\delta_{ij} = \begin{cases} 0, & \text{если} \quad i \neq j; \\ 1, & \text{если} \quad i = j, \end{cases}$$

можно записать

$$E = [\delta_{ij}].$$

Матрица, все элементы которой равны нулю, называется *нулевой* и обозначается через 0. Если желают указать еще число строк и столбцов нулевой матрицы, то употребляют обозначение $0_{mn}$.

С квадратной матрицей $A = [a_{ij}]_{n,n}$ связан *определитель (детерминант)*

$$\det A = \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{vmatrix}.$$

Не следует отождествлять эти два понятия: матрица представляет собой упорядоченную систему чисел, записанную в виде прямоугольной таблицы, а ее определитель $\det A$ есть число, определяемое по известным правилам, а именно:

$$\det A = \sum_{(\alpha_1, \alpha_2, \ldots, \alpha_n)} (-1)^{\varkappa} a_{1\alpha_1} a_{2\alpha_2} \ldots a_{n\alpha_n}, \tag{3}$$

где сумма (3) распространена на всевозможные перестановки $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ элементов $1, 2, \ldots, n$ и, следовательно, содержит $n!$ слагаемых, причем $\varkappa = 0$, если перестановка четная, и $\varkappa = 1$, если перестановка нечетная.

## § 2. Действия с матрицами

### А. Равенство матриц

Две матрицы $A = [a_{ij}]$ и $B = [b_{ij}]$ считаются равными: $A = B$, если они одного и того же типа, т. е. имеют одинаковое число строк и столбцов, и соответствующие элементы их равны, т. е.

$$a_{ij} = b_{ij}.$$

### Б. Сумма и разность матриц

*Суммой* двух матриц $A=[a_{ij}]$ и $B=[b_{ij}]$ одинакового типа называется матрица $C=[c_{ij}]$ того же типа, элементы которой $c_{ij}$ равны суммам соответствующих элементов $a_{ij}$ и $b_{ij}$ матриц $A$ и $B$, т. е. $c_{ij}=a_{ij}+b_{ij}$. Таким образом,

$$A+B=\begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \dots & a_{1n}+b_{1n} \\ a_{21}+b_{21} & a_{22}+b_{22} & \dots & a_{2n}+b_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1}+b_{m1} & a_{m2}+b_{m2} & \dots & a_{mn}+b_{mn} \end{bmatrix}.$$

Из определения суммы матриц непосредственно вытекают следующие ее свойства:

1) $A+(B+C)=(A+B)+C$;
2) $A+B=B+A$;
3) $A+0=A$.

Аналогично определяется *разность матриц*

$$A-B=\begin{bmatrix} a_{11}-b_{11} & a_{12}-b_{12} & \dots & a_{1n}-b_{1n} \\ a_{21}-b_{21} & a_{22}-b_{22} & \dots & a_{2n}-b_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1}-b_{m1} & a_{m2}-b_{m2} & \dots & a_{mn}-b_{mn} \end{bmatrix}.$$

### В. Умножение матрицы на число

*Произведением матрицы* $A=[a_{ij}]$ на число $\alpha$ (или произведением числа $\alpha$ на матрицу $A$) называется *матрица*, элементы которой получены умножением всех элементов матрицы $A$ на число $\alpha$, т. е.

$$A\alpha=\alpha A=\begin{bmatrix} \alpha a_{11} & \alpha a_{12} & \dots & \alpha a_{1n} \\ \alpha a_{21} & \alpha a_{22} & \dots & \alpha a_{2n} \\ \dots & \dots & \dots & \dots \\ \alpha a_{m1} & \alpha a_{m2} & \dots & \alpha a_{mn} \end{bmatrix}.$$

Из определения произведения числа на матрицу непосредственно вытекают следующие его свойства:

1) $1A=A$;
2) $0A=0$;
3) $\alpha(\beta A)=(\alpha\beta)A$;
4) $(\alpha+\beta)A=\alpha A+\beta A$;
5) $\alpha(A+B)=\alpha A+\alpha B$

(здесь $A$ и $B$ — матрицы; $\alpha$ и $\beta$ — числа).

Заметим, что если матрица $A$ — квадратная порядка $n$, то

$$\det \alpha A=\alpha^n \det A.$$

Матрица

$$-A=(-1)A$$

называется *противоположной*. Нетрудно видеть, что если матрицы $A$ и $B$ одинаковых типов, то

$$A-B=A+(-B).$$

## Г. Умножение матриц

Пусть

$$A=\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

и

$$B=\begin{bmatrix} b_{11} & b_{12} & \dots & b_{1q} \\ b_{21} & b_{22} & \dots & b_{2q} \\ \dots & \dots & \dots & \dots \\ b_{p1} & b_{p2} & \dots & b_{pq} \end{bmatrix}$$

— матрицы типов соответственно $m\times n$ и $p\times q$. Если число столбцов матрицы $A$ равно числу строк матрицы $B$, т. е.

$$n=p, \tag{1}$$

то для этих матриц определена матрица $C$ типа $m\times q$, называемая их *произведением*:

$$C=\begin{bmatrix} c_{11} & c_{12} & \dots & c_{1q} \\ c_{21} & c_{22} & \dots & c_{2q} \\ \dots & \dots & \dots & \dots \\ c_{m1} & c_{m2} & \dots & c_{mq} \end{bmatrix},$$

где

$$c_{ij}=a_{i1}b_{1j}+a_{i2}b_{2j}+\dots+a_{in}b_{nj}\ (i=1,\ 2,\ \dots,\ m;\ j=1,\ 2,\ \dots,\ q).$$

Из определения вытекает следующее правило умножения матриц: *чтобы получить элемент, стоящий в $i$-й строке и $j$-м столбце произведения двух матриц, нужно элементы $i$-й строки первой матрицы умножить на соответствующие элементы $j$-го столбца второй и полученные произведения сложить.*

Произведение $AB$ имеет смысл тогда и только тогда, когда матрица $A$ содержит в строках столько элементов, сколько элементов имеется в столбцах матрицы $B$. В частности, можно перемножать квадратные матрицы лишь одинакового порядка.

Пример 1.

$$A=\begin{bmatrix}3 & 2 & 8 & 1\\ 1 & -4 & 0 & 3\end{bmatrix};$$

$$B=\begin{bmatrix}2 & -1\\ 1 & -3\\ 0 & 1\\ 3 & 1\end{bmatrix}.$$

$$AB=\begin{bmatrix}3\cdot 2+2\cdot 1+8\cdot 0+1\cdot 3 & 3\cdot(-1)+2\cdot(-3)+8\cdot 1+1\cdot 1\\ 1\cdot 2+(-4)\cdot 1+0\cdot 0+3\cdot 3 & 1\cdot(-1)+(-4)\cdot(-3)+0\cdot 1+3\cdot 1\end{bmatrix}=$$

$$=\begin{bmatrix}11 & 0\\ 7 & 14\end{bmatrix}.$$

Пример 2.

$$\begin{bmatrix}1 & 2 & 3\\ 4 & 5 & 6\\ 7 & 8 & 9\end{bmatrix}\cdot\begin{bmatrix}1\\ 2\\ 3\end{bmatrix}=\begin{bmatrix}1\cdot 1+2\cdot 2+3\cdot 3\\ 4\cdot 1+5\cdot 2+6\cdot 3\\ 7\cdot 1+8\cdot 2+9\cdot 3\end{bmatrix}=\begin{bmatrix}14\\ 32\\ 50\end{bmatrix}.$$

Матричное произведение обладает следующими свойствами:

1) $A(BC)=(AB)C$; 3) $(A+B)C=AC+BC$;
2) $\alpha(AB)=(\alpha A)B$; 4) $C(A+B)=CA+CB$

($A$, $B$ и $C$ — матрицы; $\alpha$ — число).

Равенства 1) — 4) понимаются в том смысле, что если одна из их частей существует, то другая часть также существует и они равны между собой.

Произведение двух матриц не обладает переместительным свойством, т. е., вообще говоря, $AB\neq BA$, в чем можно убедиться на примерах.

Пример 3.

$$A=\begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix};\qquad B=\begin{bmatrix}5 & 6\\ 7 & 8\end{bmatrix}.$$

Тогда

$$AB=\begin{bmatrix}19 & 22\\ 43 & 50\end{bmatrix},\quad BA=\begin{bmatrix}23 & 34\\ 31 & 46\end{bmatrix},$$

т. е. здесь $AB\neq BA$.

Более того, может даже случиться, что произведение двух матриц, взятых в одном порядке, будет иметь смысл, а произведение тех же матриц, взятых в противоположном порядке, смысла иметь не будет.

Так, например, если

$$A=\begin{bmatrix}1 & 2 & 3\\ 4 & 5 & 6\end{bmatrix};\qquad B=\begin{bmatrix}3 & 2 & 1\\ 2 & 1 & 3\\ 4 & 3 & 0\end{bmatrix},$$

то

$$AB=\begin{bmatrix}19 & 13 & 7\\ 46 & 31 & 19\end{bmatrix},\ \text{а } BA \text{ не существует.}$$

В тех частных случаях, когда $AB=BA$, матрицы $A$ и $B$ называются *перестановочными* (коммутативными). Так, например, как нетрудно убедиться, единичная матрица $E$ перестановочна с любой квадратной матрицей $A$ того же порядка, причем

$$AE=EA=A.$$

Таким образом, единичная матрица $E$ играет *роль единицы* при умножении.

Если $A$ и $B$ — квадратные матрицы одного и того же порядка, то

$$\det(AB)=\det(BA)=\det A\cdot\det B.$$

Эта формула вытекает из правила перемножения определителей.

Например, для матриц, приведенных в примере 3, имеем:

$$\begin{vmatrix}19 & 22\\ 43 & 50\end{vmatrix}=\begin{vmatrix}1 & 2\\ 3 & 4\end{vmatrix}\begin{vmatrix}5 & 6\\ 7 & 8\end{vmatrix}$$

и

$$\begin{vmatrix}23 & 34\\ 31 & 46\end{vmatrix}=\begin{vmatrix}1 & 2\\ 3 & 4\end{vmatrix}\begin{vmatrix}5 & 6\\ 7 & 8\end{vmatrix}.$$

## § 3. Транспонированная матрица

Заменив в матрице

$$A=\begin{bmatrix}a_{11} & a_{12} & \dots & a_{1n}\\ a_{21} & a_{22} & \dots & a_{2n}\\ \cdot & \cdot & \cdot & \cdot\\ a_{m1} & a_{m2} & \dots & a_{mn}\end{bmatrix}$$

типа $m\times n$ строки соответственно столбцами, получим так называемую *транспонированную матрицу*

$$A'=A^T=\begin{bmatrix}a_{11} & a_{21} & \dots & a_{m1}\\ a_{12} & a_{22} & \dots & a_{m2}\\ \cdot & \cdot & \cdot & \cdot\\ a_{1n} & a_{2n} & \dots & a_{mn}\end{bmatrix},$$

типа $n\times m$. В частности, для вектора-строки

$$\boldsymbol{a}=[a_1\ a_2\ \dots\ a_n]$$

транспонированной матрицей является вектор-столбец

$$\boldsymbol{a}'=\begin{bmatrix}a_1\\ a_2\\ \cdot\\ \cdot\\ \cdot\\ a_n\end{bmatrix}.$$

Транспонированная матрица обладает следующими свойствами:

1) дважды транспонированная матрица совпадает с исходной:

$$A'' = (A')' = A;$$

2) транспонированная матрица суммы равна сумме транспонированных матриц слагаемых, т. е.

$$(A+B)' = A' + B';$$

3) транспонированная матрица произведения равна произведению транспонированных матриц сомножителей, взятому в обратном порядке, т. е.

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

Действительно, элемент $i$-й строки и $j$-го столбца матрицы $(AB)'$ равен элементу $j$-й строки и $i$-го столбца матрицы $AB$, т. е.

$$a_{j1}b_{1i} + a_{j2}b_{2i} + \ldots + a_{jn}b_{ni}.$$

Последнее выражение, очевидно, представляет собой сумму произведений элементов $i$-й строки матрицы $B'$ на соответствующие элементы $j$-го столбца матрицы $A'$, т. е. равно общему элементу матрицы $B'A'$.

Если матрица $A$ — квадратная, то, очевидно,

$$\det A' = \det A.$$

Матрица $A = [a_{ij}]$ называется *симметрической*, если она совпадает со своей транспонированной, т. е. если

$$A' = A. \tag{1}$$

Из равенства (1) вытекает, что: 1) симметрическая матрица — квадратная ($m = n$) и 2) элементы ее, симметричные относительно главной диагонали, равны между собой, т. е.

$$a_{ji} = a_{ij}.$$

Произведение

$$C = AA',$$

очевидно, представляет собой симметрическую матрицу, так как

$$C' = (AA')' = (A')'A' = AA' = C.$$

Например,

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} = \begin{bmatrix} 1^2+2^2+3^2 & 1\cdot4+2\cdot5+3\cdot6 \\ 4\cdot1+5\cdot2+6\cdot3 & 4^2+5^2+6^2 \end{bmatrix} = \begin{bmatrix} 14 & 32 \\ 32 & 77 \end{bmatrix}.$$

## § 4. Обратная матрица

Определение 1. *Обратной матрицей* по отношению к данной называется матрица, которая, будучи умноженной как справа, так и слева на данную матрицу, дает единичную матрицу.

Для матрицы $A$ обозначим обратную ей матрицу через $A^{-1}$. Тогда по определению имеем:

$$AA^{-1}=A^{-1}A=E, \tag{1}$$

где $E$ — единичная матрица.

Нахождение обратной матрицы для данной называется *обращением* данной матрицы.

Определение 2. Квадратная матрица называется *неособенной*, если определитель ее отличен от нуля.

В противном случае матрица называется *особенной*, или *сингулярной*.

Теорема. *Всякая неособенная матрица имеет обратную матрицу.*

Доказательство. Пусть дана неособенная матрица $n$-го порядка

$$A=\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix},$$

где $\det A=\Delta\neq 0$.

Составим для матрицы $A$ так называемую *присоединенную* (или *союзную*) матрицу

$$\tilde{A}=\begin{bmatrix} A_{11} & A_{21} & \dots & A_{n1} \\ A_{12} & A_{22} & \dots & A_{n2} \\ \dots & \dots & \dots & \dots \\ A_{1n} & A_{2n} & \dots & A_{nn} \end{bmatrix}, \tag{2}$$

где $A_{ij}$ — алгебраические дополнения (миноры со знаками) соответствующих элементов $a_{ij}$ ($i$, $j=1, 2, \dots, n$).

Заметим, что алгебраические дополнения элементов строк помещаются в соответствующих столбцах, т. е. произведена операция транспонирования.

Разделим все элементы последней матрицы на величину определителя матрицы $A$, т. е. на $\Delta$:

$$A^*=\begin{bmatrix} \frac{A_{11}}{\Delta} & \frac{A_{21}}{\Delta} & \dots & \frac{A_{n1}}{\Delta} \\ \frac{A_{12}}{\Delta} & \frac{A_{22}}{\Delta} & \dots & \frac{A_{n2}}{\Delta} \\ \dots & \dots & \dots & \dots \\ \frac{A_{1n}}{\Delta} & \frac{A_{2n}}{\Delta} & \dots & \frac{A_{nn}}{\Delta} \end{bmatrix}. \tag{3}$$

Докажем, что матрица $A^*$ есть искомая обратная матрица: $A^*=A^{-1}$.

Как известно, 1) сумма произведений элементов некоторого ряда (строки или столбца) определителя на алгебраические дополнения

этих элементов равна определителю и 2) сумма произведений элементов некоторого ряда определителя на алгебраические дополнения соответствующих элементов параллельного ряда равна нулю, т. е.

$$\sum_{k=1}^{n} a_{ik}A_{jk} = \delta_{ij}\Delta \tag{4}$$

и

$$\sum_{k=1}^{n} a_{ki}A_{kj} = \delta_{ij}\Delta, \tag{4'}$$

где

$$\delta_{ij} = \begin{cases} 1 & \text{при} \quad i=j, \\ 0 & \text{при} \quad i \neq j. \end{cases}$$

На основании этих свойств, составляя произведение $AA^*$, будем иметь:

$$AA^* = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} \frac{A_{11}}{\Delta} & \frac{A_{21}}{\Delta} & \dots & \frac{A_{n1}}{\Delta} \\ \frac{A_{12}}{\Delta} & \frac{A_{22}}{\Delta} & \dots & \frac{A_{n2}}{\Delta} \\ \dots & \dots & \dots & \dots \\ \frac{A_{1n}}{\Delta} & \frac{A_{2n}}{\Delta} & \dots & \frac{A_{nn}}{\Delta} \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix} = E. \tag{5}$$

Итак, $AA^* = E$.

Формулу (5) можно вывести короче, если воспользоваться сокращенными обозначениями

$$A = [a_{ij}] \text{ и } A^* = \left[\frac{A_{ji}}{\Delta}\right].$$

Учитывая соотношение (4), получим:

$$AA^* = \left[\sum_{k=1}^{n} a_{ik}\frac{A_{jk}}{\Delta}\right] = [\delta_{ij}] = E.$$

Аналогично можно удостовериться, что $A^*A = E$.

Следовательно, $A^* = A^{-1}$, т. е.

$$A^{-1} = \frac{1}{\Delta}[A_{ji}], \tag{6}$$

где

$$\Delta = \det A.$$

Замечание 1. Для данной матрицы $A$ ее обратная матрица $A^{-1}$ единственна. Более того, всякая правая обратная (левая обратная) матрица матрицы $A$ совпадает с ее обратной матрицей $A^{-1}$ (если последняя существует).

Действительно, если

$$AB=E,$$

то, умножая это равенство слева на $A^{-1}$, получим:

$$A^{-1}AB=A^{-1}E$$

или

$$B=A^{-1}.$$

Аналогично доказывается, что если

$$CA=E,$$

то $C=A^{-1}$.

Поэтому при проверке соотношения (1) достаточно ограничиться лишь одним равенством.

Замечание 2. Особенная квадратная матрица обратной не имеет. Действительно, так как матрица $A$ — особенная, то

$$\det A=0.$$

Из равенства (1) имеем:

$$\det A\cdot\det A^{-1}=\det E=1,$$

т. е.

$$0=1?!,$$

что невозможно. Утверждение доказано.

Пример. Для матрицы

$$A=\begin{bmatrix} 1 & 2 & 3 \\ -2 & -4 & -5 \\ 3 & 5 & 6 \end{bmatrix}$$

найти обратную матрицу.

Решение. Так как определитель

$$\Delta=\begin{vmatrix} 1 & 2 & 3 \\ -2 & -4 & -5 \\ 3 & 5 & 6 \end{vmatrix}=\begin{vmatrix} 1 & 2 & 3 \\ 0 & 0 & 1 \\ 0 & -1 & -3 \end{vmatrix}=1\neq 0.$$

то матрица $A$ неособенная.

Составим присоединенную матрицу

$$\tilde{A}=\begin{vmatrix} 1 & 3 & 2 \\ -3 & -3 & -1 \\ 2 & 1 & 0 \end{vmatrix}.$$

Разделим все элементы матрицы $\tilde{A}$ на $\Delta = 1$ и получим:

$$A^{-1} = \begin{bmatrix} 1 & 3 & 2 \\ -3 & -3 & -1 \\ 2 & 1 & 0 \end{bmatrix}.$$

Рекомендуется проверить, что действительно

$$AA^{-1} = E.$$

Укажем некоторые основные свойства обратной матрицы.

1. *Определитель обратной матрицы равен обратной величине определителя исходной матрицы.* Действительно, пусть

$$A^{-1}A = E.$$

Учитывая, что определитель произведения двух квадратных матриц равен произведению определителей этих матриц, получим:

$$\det A^{-1} \det A = \det E = 1.$$

Следовательно,

$$\det A^{-1} = \frac{1}{\det A}.$$

2. *Обратная матрица произведения квадратных матриц равна произведению обратных матриц сомножителей, взятому в обратном порядке, т. е.*

$$(AB)^{-1} = B^{-1}A^{-1}.$$

В самом деле,

$$AB(B^{-1}A^{-1}) = A(BB^{-1})A^{-1} = AEA^{-1} = AA^{-1} = E$$

и

$$(B^{-1}A^{-1})AB = B^{-1}(A^{-1}A)B = B^{-1}EB = B^{-1}B = E.$$

Значит, $B^{-1}A^{-1}$ есть обратная матрица для $AB$.

В более общем случае

$$(A_1A_2\ldots A_p)^{-1} = A_p^{-1}A_{p-1}^{-1}\ldots A_1^{-1}.$$

3. *Транспонированная обратная матрица равна обратной от транспонированной данной матрицы:*

$$(A^{-1})' = (A')^{-1}.$$

Действительно, транспонируя основное соотношение $A^{-1}A = E$, получим:

$$(A^{-1}A)' = A'(A^{-1})' = E' = E.$$

Отсюда, умножая последнее равенство слева на матрицу $(A')^{-1}$, будем иметь:

$$(A')^{-1}A'\,(A^{-1})' = (A')^{-1}E$$

или

$$(A^{-1})' = (A')^{-1},$$

что и требовалось доказать.

Замечание. С помощью обратной матрицы легко решаются матричные уравнения

$$AX = B \quad \text{и} \quad YA = B.$$

Действительно, если $\det A \neq 0$, то

$$X = A^{-1}B \quad \text{и} \quad Y = BA^{-1}.$$

## § 5. Степени матрицы

Пусть $A$ — квадратная матрица. Если $p$ — натуральное число, то полагают:

$$\underbrace{AA\ldots A}_{p\ \text{раз}} = A^p.$$

Дополнительно уславливаются, что $A^0 = E$, где $E$ — единичная матрица. Если матрица $A$ неособенная, то можно ввести отрицательную степень, определив ее соотношением

$$A^{-p} = (A^{-1})^p.$$

Для степеней матрицы с целыми показателями справедливы обычные правила:

1) $A^pA^q = A^{p+q}$;
2) $(A^p)^q = A^{pq}$.

Неквадратную матрицу, очевидно, в степень возводить нельзя.

Пример 1. Пусть

$$A = \begin{bmatrix} \alpha_1 & 0 & \ldots & 0 \\ 0 & \alpha_2 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & \alpha_n \end{bmatrix}.$$

Тогда

$$A^p = \begin{bmatrix} \alpha_1^p & 0 & \ldots & 0 \\ 0 & \alpha_2^p & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & \alpha_n^p \end{bmatrix}.$$

Пример 2. Найти

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}^2.$$

Решение. Имеем:

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}^2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Если $A$ и $B$ — квадратные матрицы одного и того же порядка, причем $AB=BA$, то справедлива формула бинома Ньютона

$$(A+B)^p = \sum_{k=0}^{p} C_p^k A^k B^{p-k}.$$

## § 6. Рациональные функции матрицы

Пусть

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ x_{n1} & x_{n2} & \dots & x_{nn} \end{bmatrix}$$

— произвольная квадратная матрица порядка $n$. По аналогии с формулами элементарной алгебры определяются целые рациональные функции матрицы $X$:

$$P(X) = A_0X^m + A_1X^{m-1} + \dots + A_mE \quad (\textit{правый полином});$$

$$\tilde{P}(X) = X^mA_0 + X^{m-1}A_1 + \dots + EA_m \quad (\textit{левый полином}),$$

где $A_\nu$ ($\nu = 0, 1, \dots, m$) — матрицы типа $m \times n$ или соответственно типа $n \times m$ и $E$ — единичная матрица порядка $n$.

Вообще говоря, $P(X) \neq \tilde{P}(X)$.

Можно ввести также *дробные рациональные функции* матрицы $X$, определив их формулами

$$R_1(X) = P(X)\,[Q(X)]^{-1}$$

и

$$R_2(X) = [Q(X)]^{-1}\,P(X),$$

где $P(X)$ и $Q(X)$ — матричные полиномы и $\det[Q(X)] \neq 0$.

Пример. Пусть

$$P(X)=X^2+\begin{bmatrix}1 & -1\\ 1 & 1\end{bmatrix}X-\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix},$$

где $X$ — переменная матрица второго порядка. Найти $P\left(\begin{bmatrix}0 & 0\\ 1 & 0\end{bmatrix}\right)$.

Решение. Имеем:

$$P\left(\begin{bmatrix}0 & 0\\ 1 & 0\end{bmatrix}\right)=\begin{bmatrix}0 & 0\\ 1 & 0\end{bmatrix}^2+\begin{bmatrix}1 & -1\\ 1 & 1\end{bmatrix}\begin{bmatrix}0 & 0\\ 1 & 0\end{bmatrix}-\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}=$$

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

## § 7. Абсолютная величина и норма матрицы

Неравенство

$$A\leqslant B \tag{1}$$

между матрицами $A=[a_{ij}]$ и $B=[b_{ij}]$ одинаковых типов обозначает, что

$$a_{ij}\leqslant b_{ij}. \tag{2}$$

В этом смысле не всякие две матрицы сравнимы между собой.

Под *абсолютной величиной* (*модулем*) матрицы $A=[a_{ij}]$ будем понимать матрицу

$$|A|=[|a_{ij}|],$$

где $|a_{ij}|$ — модули элементов матрицы $A$.

Если $A$ и $B$ — матрицы, для которых операции $A+B$ и $AB$ имеют смысл, то:

а) $|A+B|\leqslant|A|+|B|$;

б) $|AB|\leqslant|A|\cdot|B|$;

в) $|\alpha A|=|\alpha|\,|A|$

($\alpha$ — число).

В частности, получаем:

$$|A^p|\leqslant|A|^p$$

($p$ — натуральное число).

Под *нормой матрицы* $A=[a_{ij}]$ понимается действительное число $\|A\|$, удовлетворяющее условиям:

а) $\|A\|\geqslant 0$, причем $\|A\|=0$ тогда и только тогда, когда $A=0$;

б) $\|\alpha A\|=|\alpha|\,\|A\|$ ($\alpha$ — число) и, в частности, $\|-A\|=\|A\|$;

в) $\|A+B\|\leqslant\|A\|+\|B\|$;

г) $\|AB\|\leqslant\|A\|\cdot\|B\|$

($A$ и $B$ — матрицы, для которых соответствующие операции имеют смысл). В частности, для квадратной матрицы имеем:

$$\|A^p\| \leqslant \|A\|^p,$$

где $p$ — натуральное число.

Отметим еще одно важное неравенство между нормами матриц $A$ и $B$ одинакового типа. Применяя условие в), будем иметь:

$$\|B\| = \|A + (B - A)\| \leqslant \|A\| + \|B - A\|.$$

Отсюда

$$\|A - B\| = \|B - A\| \geqslant \|B\| - \|A\|.$$

Аналогично

$$\|A - B\| \geqslant \|A\| - \|B\|.$$

Следовательно,

$$\|A - B\| \geqslant \big| \|B\| - \|A\| \big|.$$

Назовем норму *канонической*, если дополнительно выполнены условия:

д) если $A = [a_{ij}]$, то

$$|a_{ij}| \leqslant \|A\|,$$

причем для скалярной матрицы $A = [a_{11}]$ имеем $\|A\| = |a_{11}|$;

е) из неравенства $|A| \leqslant |B|$ ($A$ и $B$ — матрицы) следует неравенство

$$\|A\| \leqslant \|B\|.$$

В частности, $\|A\| = \| \, |A| \, \|$.

В дальнейшем для матрицы $A = [a_{ij}]$ произвольного типа мы будем рассматривать главным образом три легко вычисляемые нормы:

1) $\|A\|_m = \max\limits_i \sum\limits_j |a_{ij}|$ (*m-норма*);

2) $\|A\|_l = \max\limits_j \sum\limits_i |a_{ij}|$ (*l-норма*);

3) $\|A\|_k = \sqrt{\sum\limits_{i,j} |a_{ij}|^2}$ (*k-норма*).

Пример. Пусть

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}.$$

Имеем:

$$\|A\|_m = \max(1+2+3,\ 4+5+6,\ 7+8+9) = \max\ (6,\ 15,\ 24) = 24;$$
$$\|A\|_l = \max(1+4+7,\ 2+5+8,\ 3+6+9) = \max\ (12,\ 15,\ 18) = 18;$$
$$\|A\|_k = \sqrt{1^2+2^2+3^2+4^2+5^2+6^2+7^2+8^2+9^2} =$$
$$= \sqrt{1+4+9+16+25+36+49+64+81} = \sqrt{285} \approx 16{,}9.$$

В частности, для вектора

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}$$

эти нормы имеют следующие значения:

$$\|x\|_m = \max_i |x_i|;$$
$$\|x\|_l = |x_1| + |x_2| + \ldots + |x_n|;$$
$$\|x\|_k = |x| = \sqrt{|x_1|^2 + |x_2|^2 + \ldots + |x_n|^2}$$

(*абсолютная величина вектора*). Если компоненты вектора действительны, то имеем просто

$$\|x\|_k = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}.$$

Для норм $\|A\|_m$, $\|A\|_l$ и $\|A\|_k$ проверим выполнение условий а) — г). Непосредственно очевидно, что условия а) и б) выполняются. Удостоверимся, что для этих норм выполнено условие в). Пусть $A = [a_{ij}]$ и $B = [b_{ij}]$, причем матрицы $A$ и $B$ — одинаковых типов. Имеем:

$$\|A+B\|_m = \max_i \sum_j |a_{ij} + b_{ij}| \leqslant \max_i \left\{ \sum_j |a_{ij}| + \sum_j |b_{ij}| \right\} \leqslant$$
$$\leqslant \max_i \sum_j |a_{ij}| + \max_i \sum_j |b_{ij}| = \|A\|_m + \|B\|_m.$$

Аналогично

$$\|A+B\|_l \leqslant \|A\|_l + \|B\|_l.$$

Далее,

$$\|A+B\|_k = \sqrt{\sum_{i,j} |a_{ij} + b_{ij}|^2} \leqslant$$
$$\leqslant \sqrt{\sum_{i,j} |a_{ij}|^2 + \sum_{i,j} |b_{ij}|^2 + 2\sum_{i,j} |a_{ij}||b_{ij}|}.$$

Применяя известное неравенство Коши *)

$$\sum_{i,j} |a_{ij}||b_{ij}| \leqslant \sqrt{\sum_{i,j} |a_{ij}|^2} \cdot \sqrt{\sum_{i,j} |b_{ij}|^2},$$

---

*) Приведем доказательство *неравенства Коши*

$$\left| \sum_{s=1}^{n} a_s b_s \right|^2 \leqslant \sum_{s=1}^{n} |a_s|^2 \cdot \sum_{s=1}^{n} |b_s|^2,$$

будем иметь:

$$\|A+B\|_k \leqslant \sqrt{\sum_{i,j} |a_{ij}|^2} + \sqrt{\sum_{i,j} |b_{ij}|^2} = \|A\|_k + \|B\|_k.$$

Таким образом, для всех трех норм условие в) выполнено.

Проверим теперь выполнение условия г). Пусть матрица $A=[a_{ij}]$ типа $m'\times n'$, а матрица $B=[b_{ij}]$ типа $m''\times n''$. Для возможности перемножения первой матрицы на вторую необходимо, чтобы $m''=n'$, причем матрица $AB$ будет иметь тип $m'\times n''$.

Имеем:

$$\begin{aligned}
\|AB\|_m &= \max_i \sum_{j=1}^{n''} \left| \sum_{s=1}^{n'} a_{is}a_{sj} \right| \leqslant \\
&\leqslant \max_i \left\{ \sum_{j=1}^{n''} \sum_{s=1}^{n'} |a_{is}||b_{sj}| \right\} = \\
&= \max_i \left\{ \sum_{s=1}^{n'} |a_{is}| \sum_{j=1}^{n''} |b_{sj}| \right\} \leqslant \\
&\leqslant \max_i \left\{ \sum_{s=1}^{n'} |a_{is}| \cdot \|B\|_m \right\} = \\
&= \max_i \left\{ \sum_{s=1}^{n'} |a_{is}| \right\} \cdot \|B\|_m = \|A\|_m \cdot \|B\|_m.
\end{aligned}$$

---

где $a_s$ и $b_s$ $(s=1, 2, \ldots, n)$—произвольные комплексные числа. Пусть $\lambda$—действительная переменная. Рассмотрим очевидное неравенство

$$\sum_{s=1}^{n} |a_s\lambda + b_s e^{i\varphi_s}|^2 \geqslant 0, \qquad (*)$$

где $\varphi_s$—некоторые действительные числа. Обозначая через $\bar{a}_s$ и $\bar{b}_s$ числа, сопряженные с $a_s$ и $b_s$, будем иметь:

$$\begin{aligned}
|a_s\lambda + b_s e^{i\varphi_s}|^2 &= (a_s\lambda + b_s e^{i\varphi_s})(\bar{a}_s\lambda + \bar{b}_s e^{-i\varphi_s}) = \\
&= a_s\bar{a}_s\lambda^2 + (a_s\bar{b}_s e^{-i\varphi_s} + \bar{a}_s b_s e^{i\varphi_s})\lambda + b_s\bar{b}_s = |a_s|^2\lambda^2 + 2\operatorname{Re}(a_s\bar{b}_s e^{-i\varphi_s})\lambda + |b_s|^2.
\end{aligned}$$

Отсюда неравенство $(*)$ примет вид

$$\lambda^2 \sum_{s=1}^{n} |a_s|^2 + 2\lambda \sum_{s=1}^{n} \operatorname{Re}(a_s\bar{b}_s e^{-i\varphi_s}) + \sum_{s=1}^{n} |b_s|^2 \geqslant 0.$$

Если положить

$$\varphi_s = \arg(a_s\bar{b}_s),$$

то

$$\begin{aligned}
\operatorname{Re}(a_s\bar{b}_s e^{-i\varphi_s}) &= \operatorname{Re}\left\{|a_s\bar{b}_s| e^{i\arg(a_s\bar{b}_s)} \cdot e^{-i\arg(a_s\bar{b}_s)}\right\} = \\
&= \operatorname{Re}\{|a_s\bar{b}_s|\} = |a_s\bar{b}_s| = |a_s b_s|
\end{aligned}$$

Аналогично

$$\|AB\|_l = \max_j \sum_{i=1}^{m'} \left| \sum_{s=1}^{n'} a_{is} b_{sj} \right| \leqslant$$

$$\leqslant \max_j \left\{ \sum_{i=1}^{m'} \sum_{s=1}^{n'} |a_{is}| \, |b_{sj}| \right\} =$$

$$= \max_j \left\{ \sum_{s=1}^{n'} |b_{sj}| \sum_{i=1}^{m'} |a_{is}| \right\} \leqslant$$

$$\leqslant \max_j \left\{ \sum_{s=1}^{n'} |b_{sj}| \cdot \|A\|_l \right\} =$$

$$= \|A\|_l \cdot \max_j \sum_{s=1}^{n'} |b_{sj}| = \|A\|_l \cdot \|B\|_l.$$

Далее,

$$\|AB\|_k = \sqrt{\sum_{i=1}^{m'} \sum_{j=1}^{n''} \left| \sum_{s=1}^{n'} a_{is} b_{sj} \right|^2} \leqslant \sqrt{\sum_{i=1}^{m'} \sum_{j=1}^{n''} \left\{ \sum_{s=1}^{n'} |a_{is}| \, |b_{sj}| \right\}^2}.$$

---

и, следовательно,

$$\lambda^2 \sum_{s=1}^{n} |a_s|^2 + 2\lambda \sum_{s=1}^{n} |a_s b_s| + \sum_{s=1}^{n} |b_s|^2 \geqslant 0.$$

Так как левая часть последнего неравенства в силу исходного неравенства (*) неотрицательна при любых вещественных $\lambda$, то соответствующее квадратное уравнение не может иметь различных действительных корней. Поэтому дискриминант уравнения

$$\left\{ \sum_{s=1}^{n} |a_s b_s| \right\}^2 - \sum_{s=1}^{n} |a_s|^2 \cdot \sum_{s=1}^{n} |b_s|^2 \leqslant 0,$$

т. е.

$$\left\{ \sum_{s=1}^{n} |a_s b_s| \right\}^2 \leqslant \sum_{s=1}^{n} |a_s|^2 \cdot \sum_{s=1}^{n} |b_s|^2.$$

Отсюда и подавно

$$\left| \sum_{s=1}^{n} a_s b_s \right|^2 \leqslant \left\{ \sum_{s=1}^{n} |a_s b_s| \right\}^2 \leqslant \sum_{s=1}^{n} |a_s|^2 \cdot \sum_{s=1}^{n} |b_s|^2.$$

Если числа $a_s$ и $b_s$ действительны, то получаем просто

$$\left( \sum_{s=1}^{n} a_s b_s \right)^2 \leqslant \sum_{s=1}^{n} a_s^2 \cdot \sum_{s=1}^{n} b_s^2.$$

Применяя неравенство Коши и учитывая, что $m''=n'$, будем иметь

$$\|AB\|_k \leqslant \sqrt{\sum_{i=1}^{m'}\sum_{j=1}^{n''}\left\{\sum_{s=1}^{n'}|a_{is}|^2\cdot\sum_{t=1}^{m''}|b_{tj}|^2\right\}} =$$

$$=\sqrt{\sum_{i=1}^{m'}\sum_{s=1}^{n'}|a_{is}|^2\cdot\sum_{t=1}^{m''}\sum_{j=1}^{n''}|b_{tj}|^2}=\sqrt{\|A\|_k^2\cdot\|B\|_k^2}=\|A\|_k\cdot\|B\|_k.$$

Следовательно, для рассматриваемых норм условие г) выполнено.

Покажем, что нормы $\|A\|_m$, $\|A\|_l$ и $\|A\|_k$ — канонические.

Если $a_{pq}$ — наибольший по модулю элемент матрицы $A=[a_{ij}]$ типа $m'\times n'$, то, очевидно, имеем:

$$\|a\|_m \geqslant |a_{p1}|+\ldots+|a_{pq}|+\ldots+|a_{pn'}|\geqslant|a_{pq}|;$$

$$\|A\|_l \geqslant |a_{1q}|+\ldots+|a_{pq}|+\ldots+|a_{m'q}|\geqslant|a_{pq}|$$

и

$$\|A\|_k=\sqrt{\sum_{i=1}^{m'}\sum_{j=1}^{n'}|a_{ij}|^2}\geqslant|a_{pq}|.$$

Таким образом,

$$|a_{ij}|\leqslant|a_{pq}|\leqslant\|A\|_s \quad (s=m,\ l,\ k).$$

Кроме того, если $A=[a_{11}]$, то

$$\|A\|_m=\|A\|_l=\|A\|_k=|a_{11}|.$$

Далее, если $|A|\leqslant|B|$, где $A=[a_{ij}]$ и $B=[b_{ij}]$, то $|a_{ij}|\leqslant|b_{ij}|$. Из определения норм $\|A\|_m$, $\|A\|_l$ и $\|A\|_k$ очевидно, что имеют место неравенства

$$\|A\|_s\leqslant\|B\|_s \quad (s=m,\ l,\ k).$$

Кроме того, для любой из этих норм имеем:

$$\|A\|_s=\||A|\|_s \quad (s=m,\ l,\ k).$$

Таким образом, условие е) также выполнено.

Следовательно, доказано, что нормы $\|A\|_m$, $\|A\|_l$ и $\|A\|_k$ — канонические.

Отметим, что если матрица $E$ — единичная, порядка $n$, то

$$\|E\|_m=\|E\|_l=1$$

и

$$\|E\|_k=\sqrt{n}.$$

## § 8. Ранг матрицы

Пусть дана прямоугольная матрица

$$A=\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}.$$

Если в этой матрице выбрать произвольным образом $k$ строк и $k$ столбцов, где $k \leqslant \min(m, n)$, то элементы, стоящие на пересечении этих строк и столбцов, образуют квадратную матрицу порядка $k$. Определитель этой последней матрицы называется *минором* $k$-го порядка матрицы $A$.

Определение. *Рангом матрицы называется максимальный порядок минора матрицы, отличного от нуля.* Иными словами, матрица $A$ имеет ранг $r$, если:

1) найдется по меньшей мере один ее минор $r$-го порядка, отличный от нуля;

2) все миноры матрицы $A$ порядка $r+1$ и выше равны нулю.

Ранг нулевой матрицы, т. е. матрицы, состоящей из нулей, считается равным нулю. Разность между наименьшим из чисел $m$ и $n$ и рангом матрицы называется *дефектом* матрицы. Если дефект равен нулю, то ранг матрицы — наибольший из возможных для данного типа.

При нахождении ранга матрицы полезно придерживаться следующего правила:

1) переходить от миноров меньших порядков (начиная с миноров первого порядка, т. е. элементов матрицы) к минорам бóльших порядков;

2) пусть найден минор $D$ $r$-го порядка, отличный от нуля, тогда нужно вычислить лишь миноры $(r+1)$-го порядка, окаймляющие минор $D$. Если все эти миноры равны нулю, то ранг матрицы равен $r$; если же хотя бы один из них отличен от нуля, то эту операцию нужно применить к нему, причем в этом случае ранг матрицы заведомо больше $r$.

Пример. Найти ранг матрицы

$$\begin{bmatrix} 2 & -4 & 3 & 1 & 0 \\ 1 & -2 & 1 & -4 & 2 \\ 0 & 1 & -1 & 3 & 1 \\ 4 & -7 & 4 & -4 & 5 \end{bmatrix}.$$

Решение. Минор второго порядка, стоящий в левом верхнем углу этой матрицы, равен нулю. Однако в матрице содержатся и отличные от нуля миноры второго порядка, например

$$D=\begin{vmatrix} -4 & 3 \\ -2 & 1 \end{vmatrix} \neq 0,$$

причем окаймляющий его минор третьего порядка

$$D' = \begin{vmatrix} 2 & -4 & 3 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{vmatrix} = 1$$

и оба минора четвертого порядка, окаймляющие минор $D'$, равны нулю:

$$\begin{vmatrix} 2 & -4 & 3 & 1 \\ 1 & -2 & 1 & -4 \\ 0 & 1 & -1 & 3 \\ 4 & -7 & 4 & -4 \end{vmatrix} = 0; \quad \begin{vmatrix} 2 & -4 & 3 & 0 \\ 1 & -2 & 1 & 2 \\ 0 & 1 & -1 & 1 \\ 4 & -7 & 4 & 5 \end{vmatrix} = 0.$$

Таким образом, ранг матрицы равен трем, а дефект $4-3=1$.

## § 9. Предел матрицы

Пусть имеется последовательность матриц

$$A_k = [a_{ij}^{(k)}] \quad (k = 1, 2, \ldots) \tag{1}$$

одного и того же типа $m \times n$ $(i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n)$. Под *пределом* последовательности матриц $A_k$ понимается матрица

$$A = \lim_{k \to \infty} A_k = \left[\lim_{k \to \infty} a_{ij}^{(k)}\right]. \tag{2}$$

Последовательность матриц, имеющая предел, называется *сходящейся*.

Лемма 1. *Для сходимости последовательности матриц $A_k$ $(k = 1, 2, \ldots)$ к матрице $A$ необходимо и достаточно, чтобы*

$$\| A - A_k \| \longrightarrow 0 \quad \text{при} \quad k \longrightarrow \infty, \tag{3}$$

*где $\| A \|$ — любая каноническая норма матрицы $A$. При этом*

$$\lim_{k \to \infty} \| A_k \| = \| A \|.$$

Действительно, если

$$A_k \longrightarrow A = [a_{ij}],$$

то

$$|a_{ij} - a_{ij}^{(k)}| < \varepsilon \quad \text{при} \quad k > N(\varepsilon).$$

Отсюда

$$| A - A_k | < \varepsilon I,$$

где $I$ — матрица типа $m \times n$, все элементы которой равны единице. В силу свойств нормы имеем:

$$\| A - A_k \| \leqslant \varepsilon \| I \| \quad \text{при} \quad k > N(\varepsilon),$$

следовательно,

$$\lim_{k \to \infty} \| A - A_k \| = 0. \tag{4}$$

Обратно, пусть выполнено условие (3). Тогда при $k > N(\varepsilon)$ имеем:

$$|a_{ij} - a_{ij}^{(k)}| \leqslant \|A - A_k\| < \varepsilon$$

и, следовательно,

$$\lim_{k \to \infty} a_{ij}^{(k)} = a_{ij},$$

т. е.

$$\lim_{k \to \infty} A_k = A.$$

Кроме того, если $A_k \longrightarrow A$, то имеем:

$$|\,\|A\| - \|A_k\|\,| \leqslant \|A - A_k\| \longrightarrow 0 \quad \text{при} \quad k \longrightarrow \infty.$$

Поэтому

$$\lim_{k \to \infty} \|A_k\| = \|A\|.$$

Следствие. Последовательность $A_k \longrightarrow 0$ при $k \longrightarrow \infty$ тогда и только тогда, когда

$$\lim_{k \to \infty} \|A_k\| = 0,$$

где $\|A_k\|$ — какая-нибудь каноническая норма.

Легко убедиться, что если

$$\lim_{k \to \infty} A_k = A \quad \text{и} \quad \lim_{k \to \infty} B_k = B,$$

то:

а) $\lim\limits_{k \to \infty} (A_k \pm B_k) = A \pm B$,

б) $\lim\limits_{k \to \infty} (A_k B_k) = AB$,

в) $\lim\limits_{k \to \infty} A_k^{-1} = A^{-1}$ $(\det A \neq 0)$,

в предположении, что соответствующие операции имеют смысл. В частности, если $C$ — постоянная матрица такая, что возможны перемножения $CA_k$ и $A_k C$ $(k = 1, 2, \ldots)$, то

$$\lim_{k \to \infty} CA_k = CA$$

и

$$\lim_{k \to \infty} A_k C = AC.$$

Лемма 2. *Для сходимости последовательности матриц $A_k$ $(k = 1, 2, \ldots)$ необходимо и достаточно, чтобы был выполнен обобщенный критерий Коши,* а именно: для всякого $\varepsilon > 0$ должен существовать такой номер $N = N(\varepsilon)$, что при $k > N$, $p > 0$

$$\|A_{k+p} - A_k\| < \varepsilon, \tag{5}$$

где $\|\ \|$ — любая каноническая норма.

Действительно, если справедливо неравенство (5), то для каждого элемента $a_{ij}^{(k)}$ матрицы $A_k$ будет выполнен критерий Коши (см. гл. III, § 4) и, следовательно, существует

$$\lim_{k\to\infty} A_k = \left[\lim_{k\to\infty} a_{ij}^{(k)}\right].$$

Обратно, если существует

$$A = \lim_{k\to\infty} A_k,$$

то в силу леммы 1

$$\| A - A_k \| \to 0 \quad \text{при} \quad k \to \infty$$

и, значит, будет иметь место неравенство (5).

## § 10. Матричные ряды

Пользуясь понятием предела матрицы, можно ввести в рассмотрение *матричные ряды*

$$\sum_{k=1}^{\infty} A_k = \lim_{N\to\infty} \sum_{k=1}^{N} A_k, \tag{1}$$

где $A_k$ — матрицы одного и того же типа.

Если предел (1) существует, то матричный ряд называется *сходящимся*, и матрица, полученная в пределе, называется *суммой* этого ряда. Если предела (1) не существует, то матричный ряд называется *расходящимся* и ему не приписывается никакой суммы.

**Необходимое условие сходимости матричного ряда.**

Теорема 1. *Если матричный ряд* (1) *сходится, то*

$$\lim_{k\to\infty} A_k = 0.$$

Доказательство. Пусть

$$S_k = \sum_{j=1}^{k} A_j.$$

Если ряд (1) сходится, то существует конечный предел

$$S = \lim_{k\to\infty} S_k.$$

Имеем:

$$A_k = S_k - S_{k-1},$$

откуда

$$\lim_{k\to\infty} A_k = \lim_{k\to\infty} S_k - \lim_{k\to\infty} S_{k-1} = S - S = 0.$$

Матричный ряд (1) называется *абсолютно сходящимся*, если сходится ряд

$$\sum_{k=1}^{\infty} |A_k|. \tag{2}$$

Теорема 2. *Абсолютно сходящийся матричный ряд есть ряд сходящийся.*

Доказательство. Пусть

$$A_k = \left[a_{ij}^{(k)}\right] \quad (k = 1, 2, \ldots).$$

Тогда

$$\sum_{k=1}^{\infty} |A_k| = \left[\sum_{k=1}^{\infty} |a_{ij}^{(k)}|\right].$$

Так как матричный ряд (2) сходится, то по определению каждый из числовых рядов $\sum_{k=1}^{\infty} |a_{ij}^{(k)}|$ $(i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n)$ является сходящимся. Отсюда в силу известной теоремы из теории рядов сходятся также, и притом абсолютно, все ряды $\sum_{k=1}^{\infty} a_{ij}^{(k)}$ $(i = 1, \ldots, m;$ $j = 1, \ldots, n)$, т. е. существует предел

$$S = \lim_{N \to \infty} S_N = \lim_{N \to \infty} \sum_{k=1}^{N} A_k$$

и, значит, матричный ряд (1) сходится.

Для грубого анализа сходимости матричного ряда (1) можно пользоваться приведенным ниже достаточным условием.

Теорема 3. *Если $\|A\|$ — любая каноническая норма и числовой ряд*

$$\sum_{k=1}^{\infty} \|A_k\| \tag{3}$$

*сходится, то матричный ряд* (1) *также сходится и притом абсолютно.*

Доказательство. Пусть

$$A_k = \left[a_{ij}^{(k)}\right] \quad (k = 1, 2, \ldots).$$

Рассмотрим числовые ряды

$$\sum_{k=1}^{\infty} a_{ij}^{(k)} \tag{4}$$

$(i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n)$. Так как

$$|a_{ij}^{(k)}| \leqslant \|A_k\|,$$

то каждый из рядов (4) сходится и притом абсолютно. Следовательно, матричный ряд

$$\sum_{k=1}^{\infty} A_k = \left[\sum_{k=1}^{\infty} a_{ij}^{(k)}\right]$$

в силу определения также сходится, причем сходимость — абсолютная.

В приложениях важное значение имеют *матричные степенные ряды:* правые

$$\sum_{k=0}^{\infty} A_k X^k \tag{5}$$

и левые

$$\sum_{k=0}^{\infty} X^k A_k, \tag{5'}$$

где $X$ — квадратная матрица порядка $n$. В первом случае $A_k$ — матрицы типа $m \times n$ или числа (например, $A_k$ могут представлять собой векторы-строки); во втором случае $A_k$ — матрицы типа $n \times m$ или числа (например, $A_k$ могут быть векторами-столбцами).

Теорема 4. *Если $r$ — радиус сходимости скалярного степенного ряда*

$$\sum_{k=0}^{\infty} \| A_k \| x^k, \tag{6}$$

*где $\| A_k \|$ $(k = 0, 1, 2, \ldots)$ — какая-нибудь каноническая норма, то матричные степенные ряды* (5) *и* (5′) *заведомо сходятся при*

$$\| X \| < r. \tag{7}$$

В частности, матричный степенной ряд

$$\sum_{k=0}^{\infty} a_k X^k$$

с числовыми коэффициентами $a_k$ $(k = 0, 1, 2, \ldots)$ сходится при

$$\| X \| < r,$$

где $r$ — радиус сходимости степенного ряда

$$\sum_{k=0}^{\infty} | a_k | x^k.$$

Доказательство. Так как

$$\| A_k X^k \| \leqslant \| A_k \| \, \| X \|^k,$$

то при выполнении неравенства (7) ряд

$$\sum_{k=0}^{\infty} \| A_k X^k \|$$

сходится. Следовательно, в силу теоремы 3 степенной ряд (5) также сходится.

Аналогичное рассуждение справедливо для ряда (5′).

Второе утверждение теоремы следует из того, что если $a_k$ — число, то

$$\| a_k \| = | a_k |.$$

Теорема 5. *Геометрические прогрессии*

$$A+AX+AX^2+\ldots+AX^k+\ldots \tag{8}$$

*и*

$$A+XA+X^2A+\ldots+X^kA+\ldots, \tag{8'}$$

*где $X$ — квадратная матрица, сходятся, если*

$$\|X\|<1. \tag{9}$$

*При этом*

$$\sum_{k=0}^{\infty} AX^k = A(E-X)^{-1}$$

*и*

$$\sum_{k=0}^{\infty} X^kA = (E-X)^{-1}A.$$

Действительно, в силу теоремы 4 при наличии условия (9) геометрическая прогрессия (8) сходится, т. е. существует конечная матрица

$$S=\sum_{k=0}^{\infty} AX^k.$$

Рассмотрим тождество

$$A(E+X+X^2+\ldots+X^k)(E-X)=A(E-X^{k+1}). \tag{10}$$

Переходя к пределу при $k\to\infty$ в равенстве (10) и учитывая, что в силу условия (9)

$$X^{k+1}\to 0 \quad \text{при} \quad k\to\infty,$$

будем иметь:

$$S(E-X)=AE=A. \tag{11}$$

В частности, полагая $A=E$ в равенстве (11), получим:

$$S_1(E-X)=E,$$

где

$$S_1=\sum_{k=0}^{\infty} X^k.$$

Отсюда

$$\det S_1\cdot\det(E-X)=\det E=1.$$

Так как $\det S_1$ конечен, то

$$\det(E-X)\neq 0$$

и, следовательно, матрица $E-X$ — неособенная, т. е. существует $(E-X)^{-1}$.

Умножая обе части равенства (11) справа на $(E-X)^{-1}$, получим окончательно:

$$S=\sum_{k=0}^{\infty} AX^k=A(E-X)^{-1}.$$

Аналогично доказывается, что

$$\sum_{k=0}^{\infty} X^kA=(E-X)^{-1}A$$

при

$$\|X\|<1.$$

Следствие. Если $\|X\|<1$, то существует обратная матрица

$$(E-X)^{-1}=\sum_{k=0}^{\infty} X^k.$$

Сверх того, если $\|E\|=1$, то

$$\|(E-X)^{-1}\|\leqslant\sum_{k=0}^{\infty}\|X\|^k=\frac{1}{1-\|X\|}.$$

Замечание. Если $\|X\|<1$, то нетрудно оценить норму остатка матричного ряда (8).

Имеем:

$$R_k\equiv\|A(E-X)^{-1}-A(E+X+X^2+\ldots+X^k)\|\leqslant\|A\|\,\|X^{k+1}+$$
$$+X^{k+2}+\ldots\|\leqslant\|A\|(\|X\|^{k+1}+\|X\|^{k+2}+\ldots)=\frac{\|A\|\,\|X\|^{k+1}}{1-\|X\|}.$$

Аналогично для ряда (8′) имеем:

$$R_k'=\|(E-X)^{-1}A-(E+X+X^2+\ldots+X^k)A\|\leqslant\frac{\|A\|\,\|X\|^{k+1}}{1-\|X\|}.$$

Матричные ряды дают возможность определить *трансцендентные функции матрицы*. Например, полагают

$$e^X=\sum_{n=0}^{\infty}\frac{X^n}{n!},\tag{12}$$

причем можно доказать, что ряд (12) сходится для любой квадратной матрицы $X$.

## § 11. Клеточные матрицы

Пусть дана некоторая матрица $A$. Разобьем ее на матрицы низших порядков (*клетки* или *блоки*) с помощью горизонтальных и вертикальных перегородок, идущих вдоль всей матрицы. Например,

$$A=\left[\begin{array}{cc|c} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ \hline a_{31} & a_{32} & a_{33} \end{array}\right],$$

где клетками являются матрицы

$$P=\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}; \quad Q=\begin{bmatrix} a_{13} \\ a_{23} \end{bmatrix}; \quad R=[a_{31} \;\; a_{32}]; \quad S=[a_{33}].$$

Тогда матрицу $A$ можно рассматривать как сложную матрицу, элементами которой служат клетки:

$$A=\begin{bmatrix} P & Q \\ R & S \end{bmatrix}.$$

Матрица, разбитая на клетки, называется *клеточной* или *блочной*. Понятно, что разбиение матрицы на клетки может быть осуществлено различными способами. Частным случаем клеточных матриц являются *квазидиагональные* матрицы

$$A=\left[\begin{array}{c|c|c|c} A_1 & & & \\ \hline & \ddots & & \\ \hline & & & \\ \hline & & & A_s \end{array}\right],$$

где клетки $A_i$ $(i=1, \ldots, s)$ есть квадратные матрицы, вообще говоря, различных порядков, а вне клеток стоят нули. Отметим, что

$$\det A=\det A_1 \ldots \det A_s.$$

Другой важный частный случай клеточных матриц представляют *окаймленные матрицы*

$$A_n=\left[\begin{array}{c|c} A_{n-1} & U_n \\ \hline V_n & a_{nn} \end{array}\right],$$

где

$$A_{n-1}=\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1,\,n-1} \\ a_{21} & a_{22} & \ldots & a_{2,\,n-1} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n-1,\,1} & a_{n-1,\,2} & \ldots & a_{n-1,\,n-1} \end{bmatrix}$$

— матрица порядка $n-1$;

$$U_n = \begin{bmatrix} a_{1,\,n} \\ a_{2,\,n} \\ \cdots \\ a_{n-1,\,n} \end{bmatrix} \text{— матрица-столбец;}$$

$V_n = [a_{n,1}\ a_{n,2} \ldots a_{n,\,n-1}]$ — матрица-строка и $a_{nn}$ — число.

Клеточные матрицы одинакового типа и с одинаковым разбиением условимся называть *конформными*. Удобство клеточных матриц состоит в том, что действия над ними совершаются формально по тем же правилам, что и над обыкновенными матрицами.

### А. Сложение и вычитание клеточных матриц

Если клеточные матрицы

$$A = \begin{bmatrix} A_{11} & A_{12} \ldots A_{1q} \\ \cdots & \cdots \\ A_{p1} & A_{p2} \ldots A_{pq} \end{bmatrix} \tag{1}$$

и

$$B = \begin{bmatrix} B_{11} & B_{12} \ldots B_{1s} \\ \cdots & \cdots \\ B_{r1} & B_{r2} \ldots B_{rs} \end{bmatrix} \tag{2}$$

конформны, т. е. $p=r$; $q=s$ и клетки $A_{ij}$ и $B_{ij}$ имеют одинаковый тип, то

$$A+B = \begin{bmatrix} A_{11}+B_{11} & A_{12}+B_{12} \ldots A_{1q}+B_{1q} \\ \cdots & \cdots \\ A_{p1}+B_{p1} & A_{p2}+B_{p2} \ldots A_{pq}+B_{pq} \end{bmatrix}.$$

В самом деле, чтобы сложить матрицы $A$ и $B$, надо сложить соответствующие элементы их, но очевидно, что то же самое получится, если мы сложим соответствующие клетки этих матриц.

Аналогично производится вычитание клеточных матриц.

Если $A$ — клеточная матрица (1) и $\alpha$ — число, то имеем:

$$\alpha A = \begin{bmatrix} \alpha A_{11} & \alpha A_{12} \ldots \alpha A_{1q} \\ \cdots & \cdots \\ \alpha A_{p1} & \alpha A_{p2} \ldots \alpha A_{pq} \end{bmatrix}.$$

### Б. Умножение клеточных матриц

Пусть клеточные матрицы $A$ и $B$ имеют структуру соответственно (1) и (2), причем $q=r$.

Предположим, что все клетки $A_{ij}$ и $B_{jk}$ ($i=1, 2, \ldots, p$; $j=1, 2, \ldots, q$; $k=1, 2, \ldots, s$) таковы, что число столбцов клетки $A_{ij}$ равняется числу строк клетки $B_{jk}$. В частном случае, если все клетки $A_{ij}$ и $B_{ij}$ — квадратные и имеют один и тот же порядок,

то это предположение заведомо выполняется. Тогда можно доказать, что произведение матриц $A$ и $B$ есть клеточная матрица

$$C=\begin{bmatrix} C_{11} & C_{12} \dots C_{1s} \\ C_{21} & C_{22} \dots C_{2s} \\ \cdot & \cdot \; \cdot \; \cdot \; \cdot \; \cdot \\ C_{p1} & C_{p2} \dots C_{ps} \end{bmatrix},$$

где $C_{ik}=A_{i1}B_{1k}+A_{i2}B_{2k}+\dots+A_{iq}B_{qk}$ $(i=1, 2, \dots, p;\ k=1, 2, \dots, s)$, т. е. матрицы $A$ и $B$ перемножаются так, как будто на месте клеток находятся числа [2].

Пример. Перемножая клеточные матрицы

$$A=\left[\begin{array}{c|c|c} & \leftarrow 2 \rightarrow & \leftarrow 1 \rightarrow \\ \hline \uparrow\ 2\ \downarrow & P & Q \end{array}\right]$$

и

$$B=\left[\begin{array}{c|c|c} & \leftarrow 1 \rightarrow & \leftarrow 2 \rightarrow \\ \hline \uparrow\ 2\ \downarrow & R & S \\ \hline \uparrow\ 1\ \downarrow & T & U \end{array}\right],$$

получим матрицу вида

$$AB=\left[\begin{array}{c|c|c} & \leftarrow 1 \rightarrow & \leftarrow 2 \rightarrow \\ \hline \uparrow\ 2\ \downarrow & PR+QT & PS+QU \end{array}\right].$$

Особенно просто производится сложение и умножение квазидиагональных матриц. Если

$$A=\begin{bmatrix} A_1 & & \\ & \ddots & \\ & & A_s \end{bmatrix}, \quad B=\begin{bmatrix} B_1 & & \\ & \ddots & \\ & & B_s \end{bmatrix}$$

и порядки матриц $A_i$, $B_i$ $(i=1, 2, \ldots, s)$ одинаковы, то, очевидно, имеем:

$$A+B=\begin{bmatrix} \boxed{A_1+B_1} & & \\ & \boxed{\ddots} & \\ & & \boxed{A_s+B_s} \end{bmatrix}$$

и

$$AB=\begin{bmatrix} \boxed{A_1B_1} & & \\ & \boxed{\ddots} & \\ & & \boxed{A_sB_s} \end{bmatrix}.$$

## § 12. Обращение матриц при помощи разбиения на клетки

Пусть для данной неособенной числовой матрицы $A$ требуется найти обратную матрицу $A^{-1}$. Разобьем матрицу $A$ на четыре клетки:

$$A=\begin{bmatrix} \alpha_{11}(r, r) & \alpha_{12}(r, s) \\ \alpha_{21}(s, r) & \alpha_{22}(s, s) \end{bmatrix}.$$

Здесь в скобках указаны порядки соответствующих клеток, причем $r+s=n$, где $n$ — порядок матрицы $A$. Будем искать обратную матрицу $A^{-1}$ также в виде четырехклеточной матрицы

$$A^{-1}=\begin{bmatrix} \beta_{11}(r, r) & \beta_{12}(r, s) \\ \beta_{21}(s, r) & \beta_{22}(s, s) \end{bmatrix}.$$

Тогда, так как $A^{-1}A=E$, то, перемножая эти матрицы, получим четыре матричных уравнения

$$\left.\begin{aligned} &\beta_{11}\alpha_{11}+\beta_{12}\alpha_{21}=E_r, \\ &\beta_{11}\alpha_{12}+\beta_{12}\alpha_{22}=0, \\ &\beta_{21}\alpha_{11}+\beta_{22}\alpha_{21}=0, \\ &\beta_{21}\alpha_{12}+\beta_{22}\alpha_{22}=E_s, \end{aligned}\right\} \qquad (1)$$

где $E_r$ и $E_s$ — единичные матрицы соответствующих порядков. Решив эту систему, определим клетки матрицы $A^{-1}$. Для решения системы (1) используем способ исключения неизвестных. Умножая справа первое уравнение системы (1) на $\alpha_{11}^{-1}\alpha_{12}$ и вычитая из результата умножения второе уравнение этой системы, получим:

$$\beta_{12}(\alpha_{21}\alpha_{11}^{-1}\alpha_{12}-\alpha_{22})=\alpha_{11}^{-1}\alpha_{12}.$$

Отсюда находим:

$$\beta_{12} = -\alpha_{11}^{-1}\alpha_{12}\left(\alpha_{22} - \alpha_{21}\alpha_{11}^{-1}\alpha_{12}\right)^{-1}$$

и

$$\beta_{11} = \alpha_{11}^{-1} - \beta_{12}\alpha_{21}\alpha_{11}^{-1}.$$

Аналогично из третьего и четвертого уравнений системы (1) будем иметь:

$$\beta_{22} = \left(\alpha_{22} - \alpha_{21}\alpha_{11}^{-1}\alpha_{12}\right)^{-1}$$

и

$$\beta_{21} = -\beta_{22}\alpha_{21}\alpha_{11}^{-1}.$$

Здесь, конечно, предполагается, что соответствующие операции имеют смысл. Введем в рассмотрение матрицы

$$\left.\begin{aligned} &X = \alpha_{11}^{-1}\alpha_{12}, \quad Y = \alpha_{21}\alpha_{11}^{-1}, \\ &\theta = \alpha_{22} - \alpha_{21}X = \alpha_{22} - Y\alpha_{12}. \end{aligned}\right\} \tag{2}$$

Тогда формулы для клеток $\beta_{ij}$ $(i,\ j = 1,\ 2)$ можно записать проще:

$$\beta_{11} = \alpha_{11}^{-1} + X\theta^{-1}Y,$$

$$\beta_{12} = -X\theta^{-1},$$

$$\beta_{21} = -\theta^{-1}Y, \quad \beta_{22} = \theta^{-1}.$$

Формулы (1) определяют клетки матрицы $A^{-1}$ при условии, что $\alpha_{11}^{-1}$ и $\theta^{-1}$ существуют. Вычисления удобно расположить в виде следующей схемы [4]:

| | $\alpha_{21}$ | $\alpha_{22}$ |
|---|---|---|
| $X = \alpha_{11}^{-1}\alpha_{12}$ | $\alpha_{11}^{-1}$ | $\alpha_{12}$ |
| $\theta^{-1}$ | $Y = \alpha_{21}\alpha_{11}^{-1}$ | $\theta = \alpha_{22} - Y\alpha_{12}$ |

и

$$A^{-1} = \left[\begin{array}{c|c} \alpha_{11}^{-1} + X\theta^{-1}Y & -X\theta^{-1} \\ \hline -\theta^{-1}Y & \theta^{-1} \end{array}\right].$$

Этот метод полезно применять, если матрица $\alpha_{11}$ легко обратима.

Пример 1. Обратить матрицу

$$\begin{bmatrix} 1 & 0 & 3 & -4 \\ 0 & 1 & 5 & 6 \\ -3 & 4 & 0 & 2 \\ -5 & -6 & 2 & 0 \end{bmatrix}.$$

Решение. Положим

$$\alpha_{11}=\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \qquad \alpha_{12}=\begin{bmatrix} 3 & -4 \\ 5 & 6 \end{bmatrix};$$

$$\alpha_{21}=\begin{bmatrix} -3 & 4 \\ -5 & -6 \end{bmatrix}; \quad \alpha_{22}=\begin{bmatrix} 0 & 2 \\ 2 & 0 \end{bmatrix}.$$

Применяя приведенную выше схему, будем иметь:

$$\begin{array}{c|c|cc|cc|}
 & & -3 & 4 & 0 & 2 \\
 & & -5 & -6 & 2 & 0 \\
\hline
X & \begin{matrix} 3 & -4 \\ 5 & 6 \end{matrix} & \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix} & & \begin{matrix} 3 & -4 \\ 5 & 6 \end{matrix} & \\
\hline
\theta^{-1}\dfrac{1}{1422} & \begin{matrix} 16 & 34 \\ -47 & -11 \end{matrix} & \begin{matrix} -3 & 4 \\ -5 & -6 \end{matrix} & & \begin{matrix} -11 & -34 \\ 47 & 16 \end{matrix} & \\
\hline
 & & Y & & \theta &
\end{array}$$

Отсюда

$$X\theta^{-1}=\frac{1}{1422}\begin{bmatrix} 3 & -4 \\ 5 & 6 \end{bmatrix}\begin{bmatrix} 16 & 34 \\ -47 & -11 \end{bmatrix}=\frac{1}{1422}\begin{bmatrix} 236 & 146 \\ -202 & 104 \end{bmatrix},$$

$$\theta^{-1}Y=\frac{1}{1422}\begin{bmatrix} 16 & 34 \\ -47 & -11 \end{bmatrix}\begin{bmatrix} -3 & 4 \\ -5 & -6 \end{bmatrix}=\frac{1}{1422}\begin{bmatrix} -218 & -140 \\ 196 & -122 \end{bmatrix},$$

$$X\theta^{-1}Y=\frac{1}{1422}\begin{bmatrix} 236 & 146 \\ -202 & 104 \end{bmatrix}\begin{bmatrix} -3 & 4 \\ -5 & -6 \end{bmatrix}=\frac{1}{1422}\begin{bmatrix} -1438 & 68 \\ 86 & -1432 \end{bmatrix}.$$

Для контроля произведение $X\theta^{-1}Y$ вычисляем двумя способами:

$$X\theta^{-1}Y=(X\theta^{-1})Y \quad \text{и} \quad Y\theta^{-1}Y=X(\theta^{-1}Y).$$

По общей схеме имеем:

$$A^{-1}=\frac{1}{1422}\left[\begin{array}{cc|cc} -16 & 68 & -236 & -146 \\ 86 & -10 & 202 & -104 \\ \hline 218 & 140 & 16 & 34 \\ -196 & 122 & -47 & -11 \end{array}\right].$$

Частным случаем изложенного метода является так называемый *метод окаймления*. Сущность его заключается в следующем. Пусть дана матрица

$$A=\begin{bmatrix} a_{11}\ldots a_{1n} \\ \ldots\ldots \\ a_{n1}\ldots a_{nn} \end{bmatrix}.$$

Образуем последовательность матриц

$$S_1=[a_{11}];$$

$$S_2=\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix};$$

$$S_3=\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}=\left[\begin{array}{cc|c} \multicolumn{2}{c|}{S_2} & \begin{matrix} a_{13} \\ a_{23} \end{matrix} \\ \hline a_{31} & a_{32} & a_{33} \end{array}\right];$$

$$S_4=\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}=\left[\begin{array}{ccc|c} \multicolumn{3}{c|}{S_3} & \begin{matrix} a_{14} \\ a_{24} \\ a_{34} \end{matrix} \\ \hline a_{41} & a_{42} & a_{43} & a_{44} \end{array}\right]$$

и т. д. Каждая следующая матрица получена из предыдущей при помощи окаймления. Обратная ко второй из этих матриц $S_2^{-1}$ находится непосредственно:

$$S_2^{-1}=\begin{bmatrix} \frac{a_{22}}{\Delta} & -\frac{a_{12}}{\Delta} \\ -\frac{a_{21}}{\Delta} & \frac{a_{11}}{\Delta} \end{bmatrix},$$

где

$$\Delta=a_{11}a_{22}-a_{12}a_{21}.$$

При помощи матрицы $S_2^{-1}$, применив к $S_3$ приведенную выше схему вычисления, можно получить $S_3^{-1}$, а затем при помощи $S_3^{-1}$ аналогично получить $S_4^{-1}$ и, наконец, $S_n^{-1}=A^{-1}$.

Метод окаймления становится непригодным, если одна из промежуточных матриц $S_i$ является особенной. Положение может быть исправлено с помощью перестановки строк матрицы [5].

Пример 2. Найти обратную матрицу для матрицы

$$A=\begin{bmatrix} 1 & 4 & 1 & 3 \\ 0 & -1 & 3 & -1 \\ 3 & 1 & 0 & 2 \\ 1 & -2 & 5 & 1 \end{bmatrix}.$$

Решение. Здесь

$$S_2=\begin{bmatrix} 1 & 4 \\ 0 & -1 \end{bmatrix}=S_2^{-1}.$$

Схема вычисления $S_3^{-1}$ имеет следующий вид:

$$
\begin{array}{c|c|cc|c|}
 & & 3 & 1 & 0 \\
\hline
X & \begin{matrix}13\\-3\end{matrix} & \begin{matrix}1\\0\end{matrix} & \begin{matrix}4\\-1\end{matrix} & \begin{matrix}1\\3\end{matrix} \\
\hline
\theta^{-1} & -\frac{1}{36} & 3 & 11 & -36 \\
\hline
 & & Y & & 0
\end{array}
$$

$$
X\theta^{-1}Y = \begin{bmatrix} -\frac{13}{12} & -\frac{143}{36} \\ \frac{1}{4} & \frac{11}{12} \end{bmatrix}.
$$

Следовательно,

$$
S_3^{-1} = \left[\begin{array}{cc|c} -\frac{1}{12} & \frac{1}{36} & \frac{13}{36} \\ \frac{1}{4} & -\frac{1}{12} & -\frac{1}{12} \\ \hline \frac{1}{12} & \frac{11}{36} & -\frac{1}{36} \end{array}\right].
$$

Для вычисления $S_4^{-1}$ служит следующая схема:

$$
\begin{array}{c|c|ccc|c|}
 & & 1 & -2 & 5 & 1 \\
\hline
X & \begin{matrix}\frac{4}{9}\\ \frac{2}{3}\\ -\frac{1}{9}\end{matrix} & \begin{matrix}-\frac{1}{12}\\ \frac{1}{4}\\ \frac{1}{12}\end{matrix} & \begin{matrix}\frac{1}{36}\\ -\frac{1}{12}\\ \frac{11}{36}\end{matrix} & \begin{matrix}\frac{13}{36}\\ -\frac{1}{12}\\ \frac{1}{36}\end{matrix} & \begin{matrix}3\\ -1\\ 2\end{matrix} \\
\hline
\theta^{-1} & \frac{9}{22} & -\frac{1}{6} & \frac{31}{18} & \frac{7}{18} & \frac{22}{9} \\
\hline
 & & & Y & & \theta
\end{array}
$$

$$
X\theta^{-1}Y = \begin{bmatrix} -\frac{1}{33} & \frac{31}{99} & \frac{7}{99} \\ -\frac{1}{22} & \frac{31}{66} & \frac{7}{66} \\ \frac{1}{132} & -\frac{31}{396} & -\frac{7}{396} \end{bmatrix}.
$$

Следовательно,

$$S_4^{-1}=A^{-1}=\left[\begin{array}{cccc|c} -\frac{5}{44} & \frac{15}{44} & \frac{19}{44} & & -\frac{2}{11} \\ \frac{9}{44} & \frac{17}{44} & \frac{1}{44} & & -\frac{3}{11} \\ \frac{4}{44} & \frac{10}{44} & -\frac{2}{44} & & \frac{1}{22} \\ \hline \frac{3}{44} & -\frac{31}{44} & -\frac{7}{44} & & \frac{9}{22} \end{array}\right]=$$

$$=\frac{1}{44}\begin{bmatrix} -5 & 15 & 19 & -8 \\ 9 & 17 & 1 & -12 \\ 4 & 10 & -2 & 2 \\ 3 & -31 & -7 & 18 \end{bmatrix}.$$

## § 13. Треугольные матрицы

Определение. Квадратная матрица называется *треугольной*, если элементы, стоящие выше (ниже) главной диагонали, равны нулю. Например,

$$T=\begin{bmatrix} t_{11} & t_{12} & \dots & t_{1n} \\ 0 & t_{22} & \dots & t_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & t_{nn} \end{bmatrix},$$

где $t_{ij}=0$ при $i>j$, есть верхняя треугольная матрица. Аналогично

$$T_1=\begin{bmatrix} t_{11} & 0 & \dots & 0 \\ t_{21} & t_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ t_{n1} & t_{n2} & \dots & t_{nn} \end{bmatrix},$$

где $t_{ij}=0$ для $j>i$, есть нижняя треугольная матрица.

Диагональная матрица является частным случаем как верхней, так и нижней треугольной матрицы. Определитель треугольной матрицы равен произведению ее диагональных элементов, а именно: если $T=[t_{ij}]$ — треугольная матрица, то очевидно, что $\det T=$ $=t_{11}t_{22}\dots t_{nn}$. Поэтому треугольная матрица является неособенной только тогда, когда все ее диагональные элементы отличны от нуля.

Можно доказать, что: 1) сумма и произведение треугольных матриц одинакового типа и одной и той же структуры, т. е. одновременно только верхних или только нижних, есть также треугольные матрицы того же типа и общей структуры; 2) обратная матрица неособенной треугольной матрицы есть также треугольная матрица того же типа и структуры. Пользуясь последним обстоятельством мы легко можем обращать треугольную матрицу.

Пример 1. Обратить матрицу

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

Решение. Положим

$$A^{-1}=\begin{bmatrix}t_{11} & 0 & 0\\ t_{21} & t_{22} & 0\\ t_{31} & t_{32} & t_{33}\end{bmatrix}.$$

Перемножая матрицы $A$ и $A^{-1}$, будем иметь:

$$\left.\begin{aligned} t_{11}&=1, & t_{11}+2t_{21}+3t_{31}&=0,\\ t_{11}+2t_{21}&=0, & 2t_{22}+3t_{32}&=0,\\ 2t_{22}&=1, & 3t_{33}&=1.\end{aligned}\right\}$$

Отсюда последовательно находим:

$$t_{11}=1;\quad t_{21}=-\frac{1}{2};\quad t_{22}=\frac{1}{2};$$
$$t_{31}=0;\quad t_{32}=-\frac{1}{3};\quad t_{33}=\frac{1}{3}.$$

Следовательно,

$$A^{-1}=\begin{bmatrix}1 & 0 & 0\\ -\frac{1}{2} & \frac{1}{2} & 0\\ 0 & -\frac{1}{3} & \frac{1}{3}\end{bmatrix}.$$

Имеет место важная теорема [3].

Теорема. *Всякую квадратную матрицу*

$$A=\begin{bmatrix}a_{11} & a_{12} & \dots & a_{1n}\\ a_{21} & a_{22} & \dots & a_{2n}\\ \dots & \dots & \dots & \dots\\ a_{n1} & a_{n2} & \dots & a_{nn}\end{bmatrix},$$

*имеющую отличные от нуля главные диагональные миноры*

$$\Delta_1=a_{11}\neq 0;\qquad \Delta_2=\begin{vmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{vmatrix}\neq 0;\ \dots;\quad \Delta_n=|A|\neq 0,$$

*можно представить в виде произведения двух треугольных матриц различных структур (нижней и верхней), причем это разложение будет единственным, если заранее зафиксировать диагональные элементы одной из треугольных матриц (например, положить их равными* 1).

Не приводя доказательства теоремы, ограничимся указанием способа отыскания элементов искомых треугольных матриц. Пусть

$$A = T_1 T_2, \tag{1}$$

где

$$T_1 = [b_{ij}], \quad b_{ij} = 0 \quad \text{для} \quad j > i, \tag{2}$$

есть нижняя треугольная матрица порядка $n$;

$$T_2 = [c_{ij}], \quad c_{ij} = 0 \quad \text{для} \quad i > j, \tag{3}$$

есть верхняя треугольная матрица порядка $n$. Перемножая эти матрицы, в силу формулы (1) получим:

$$\sum_{k=1}^{n} b_{ik} c_{kj} = a_{ij} \qquad (i,\ j = 1,\ 2,\ \ldots,\ n). \tag{4}$$

Система (4) в силу условий (2) и (3) принимает вид

$$\sum_{k=1}^{j} b_{ik} c_{kj} = a_{ij} \quad \text{при} \quad i \geqslant j \quad (j = 1,\ 2,\ \ldots,\ n) \tag{4'}$$

и

$$\sum_{k=1}^{i} b_{ik} c_{kj} = a_{ij} \quad \text{при} \quad i < j \quad (i = 1,\ 2,\ \ldots,\ n-1). \tag{4''}$$

Системы (4′) и (4″) в силу их особой структуры легко решаются с точностью до диагональных элементов $b_{ii}$ и $c_{ii}$. Для определенности можно положить $c_{ii} = 1$ $(i = 1,\ 2,\ \ldots,\ n)$.

Пример 2. Представить матрицу

$$A = \begin{bmatrix} 1 & -1 & 2 \\ -1 & 5 & 4 \\ 2 & 4 & 14 \end{bmatrix}$$

в виде произведения двух треугольных матриц $T_1$ и $T_2$.

Решение. $A = T_1 T_2$. Будем искать $T_1$ и $T_2$ в виде

$$T_1 = \begin{bmatrix} t_{11} & 0 & 0 \\ t_{21} & t_{22} & 0 \\ t_{31} & t_{32} & t_{33} \end{bmatrix} \text{ и } T_2 = \begin{bmatrix} 1 & r_{12} & r_{13} \\ 0 & 1 & r_{23} \\ 0 & 0 & 1 \end{bmatrix}.$$

Имеем:

$$\begin{bmatrix} 1 & -1 & 2 \\ -1 & 5 & 4 \\ 2 & 4 & 14 \end{bmatrix} = \begin{bmatrix} t_{11} & t_{11}r_{12} & t_{11}r_{13} \\ t_{21} & t_{21}r_{12} + t_{22} & t_{21}r_{13} + t_{22}r_{23} \\ t_{31} & t_{31}r_{12} + t_{32} & t_{31}r_{13} + t_{32}r_{23} + t_{33} \end{bmatrix};$$

откуда

$$\begin{array}{lll} t_{11} = 1; & t_{11}r_{12} = -1; & t_{11}r_{13} = 2; \\ t_{21} = -1; & t_{21}r_{12} + t_{22} = 5; & t_{21}r_{13} + t_{22}r_{23} = 4; \\ t_{31} = 2; & t_{31}r_{12} + t_{32} = 4; & t_{31}r_{13} + t_{32}r_{23} + t_{33} = 14. \end{array}$$

Решив систему, получим:

$$t_{11}=1; \quad t_{21}=-1; \quad t_{31}=2;$$
$$t_{22}=4; \quad t_{32}=6; \quad t_{33}=1;$$
$$r_{12}=-1; \quad r_{13}=2; \quad r_{23}=\frac{3}{2}.$$

Таким образом,

$$T_1=\begin{bmatrix} 1 & 0 & 0 \\ -1 & 4 & 0 \\ 2 & 6 & 1 \end{bmatrix}$$

и

$$T_2=\begin{bmatrix} 1 & -1 & 2 \\ 0 & 1 & \frac{3}{2} \\ 0 & 0 & 1 \end{bmatrix}.$$

Пользуясь представлением квадратной матрицы $A$ ($\det A \neq 0$) в виде произведения двух треугольных матриц, можно указать еще один способ вычисления обратной матрицы $A^{-1}$, а именно, если

$$A=T_1T_2,$$

то

$$A^{-1}=T_2^{-1}T_1^{-1}.$$

Обратные матрицы для треугольных, как мы видели выше, находятся сравнительно просто.

## § 14. Элементарные преобразования матриц

Следующие преобразования матриц носят названия *элементарных*:

1) перестановка двух строк или столбцов;

2) умножение всех элементов какой-либо строки (столбца) на одно и то же число, отличное от нуля;

3) прибавление к элементам какой-либо строки (столбца) соответствующих элементов другой строки (столбца), умноженных на одно и то же число.

Две матрицы называются *эквивалентными*, если одна получается из другой с помощью конечного числа элементарных преобразований. Такие матрицы не являются, вообще говоря, равными, но, как можно доказать, имеют один и тот же ранг [6].

Легко убедиться, что каждое элементарное преобразование квадратной матрицы $A$ равносильно умножению последней на некоторую неособенную матрицу. При этом, если преобразование производится над строками (столбцами) матрицы $A$, то множитель должен быть

левым (правым) и представлять собой результат применения соответствующего элементарного преобразования к единичной матрице [6]. Например, переставляя в матрице

$$A=\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

вторую и третью строки, будем иметь эквивалентную матрицу

$$\tilde{A}=\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{31} & a_{32} & a_{33} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}.$$

Та же матрица $\tilde{A}$ получится, если в единичной матрице

$$E=\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

переставить вторую и третью строки и полученную матрицу

$$\tilde{E}=\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

умножить слева на матрицу $A$, т. е. $\tilde{A}=\tilde{E}A$.

Аналогичным способом производят и другие элементарные преобразования. Заметим, что если в равенстве $AA^{-1}=E$ совершать одинаковые преобразования строк матриц $A$ и $E$ до тех пор, пока матрица $A$ не превратится в единичную, то будем иметь $\tilde{E}AA^{-1}=\tilde{E}$, где $\tilde{E}$—преобразованная единичная матрица. Отсюда, так как $\tilde{E}A=E$, получим $A^{-1}=\tilde{E}$, т. е. обратная матрица $A^{-1}$ представляет собой преобразованную единичную матрицу. На этом основан способ вычисления обратной матрицы при помощи преобразования строк [4].

## § 15. Вычисление определителей

Элементарные преобразования матрицы дают наиболее удобный способ вычисления определителя этой матрицы. Пусть, например,

$$\Delta_n=\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}. \qquad (1)$$

Предполагая, что $a_{11} \neq 0$, будем иметь:

$$\Delta_n = a_{11} \begin{bmatrix} 1 & a_{12} & \dots & a_{1n} \\ \frac{a_{21}}{a_{11}} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \frac{a_{n1}}{a_{11}} & a_{n2} & \dots & a_{nn} \end{bmatrix}.$$

Отсюда, вычитая из элементов $a_{ij}$, принадлежащих $j$-му столбцу ($j \geqslant 2$), соответствующие элементы первого столбца, умноженные на $a_{1j}$, получим:

$$\Delta_n = a_{11} \begin{bmatrix} 1 & 0 & \dots & 0 \\ \frac{a_{21}}{a_{11}} & a_{22}^{(1)} & \dots & a_{2n}^{(1)} \\ \cdot & \cdot & \cdot & \cdot \\ \frac{a_{n1}}{a_{11}} & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} \end{bmatrix} = a_{11}\Delta_{n-1},$$

где

$$\Delta_{n-1} = \begin{vmatrix} a_{22}^{(1)} & a_{23}^{(1)} & \dots & a_{2n}^{(1)} \\ a_{32}^{(1)} & a_{33}^{(1)} & \dots & a_{3n}^{(1)} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n2}^{(1)} & a_{n3}^{(1)} & \dots & a_{nn}^{(1)} \end{vmatrix} \tag{2}$$

и

$$a_{ij}^{(1)} = a_{ij} - \frac{a_{i1}a_{1j}}{a_{11}} \quad (i,\ j = 2,\ 3,\ \dots,\ n).$$

К определителю $\Delta_{n-1}$ применяем тот же прием. Если все элементы

$$a_{ii}^{(i-1)} \neq 0 \quad (i = 1,\ 2,\ \dots,\ n),$$

то окончательно находим:

$$\Delta_n = a_{11}a_{22}^{(1)} \dots a_{nn}^{(n-1)}. \tag{3}$$

Если в каком-нибудь промежуточном определителе $\Delta_{n-k}$ левый верхний элемент $a_{k+1,\,k+1}^{(k)} = 0$, то следует переставить строки или столбцы определителя $\Delta_{n-k}$ так, чтобы нужный нам элемент был отличен от нуля (это возможно всегда, если определитель $\Delta \neq 0$). Конечно, при этом следует учесть изменение знака определителя $\Delta_{n-k}$. Можно дать более общее правило. Пусть определитель $\tilde{\Delta}_n = \det[\alpha_{ij}]$

преобразован так, что $\alpha_{pq}=1$ ($\alpha_{pq}$ — главный элемент), т. е.

$$\tilde{\Delta}_n=\begin{vmatrix} \alpha_{11} & \dots & \alpha_{1q} & \dots & \alpha_{1j} & \dots & \alpha_{1n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \alpha_{i1} & \dots & \boxed{\alpha_{iq}} & \dots & \alpha_{ij} & \dots & \alpha_{in} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \alpha_{p1} & \dots & 1 & \dots & \boxed{\alpha_{pj}} & \dots & \alpha_{pn} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \alpha_{n1} & \dots & \alpha_{nq} & \dots & \alpha_{nj} & \dots & \alpha_{nn} \end{vmatrix}.$$

Тогда

$$\tilde{\Delta}_n=(-1)^{p+q}\,\tilde{\Delta}_{n-1},$$

где $\tilde{\Delta}_{n-1}=\det\left[\alpha_{ij}^{(1)}\right]$ есть определитель $(n-1)$-го порядка, получающийся из $\Delta_n$ путем выбрасывания $p$-й строки и $q$-го столбца с последующим преобразованием элементов по формуле

$$\alpha_{ij}^{(1)}=\alpha_{ij}-\alpha_{iq}\alpha_{pj},$$

т. е. *каждый элемент $\alpha_{ij}^{(1)}$ определителя $\tilde{\Delta}_{n-1}$ равен соответствующему элементу $\alpha_{ij}$ определителя $\tilde{\Delta}_n$, уменьшенному на произведение его «проекций» $\alpha_{iq}$ и $\alpha_{pj}$ на отброшенные столбец и строку исходного определителя.* Доказательство этого положения легко вытекает из общих свойств определителей [7].

Пример. Вычислить

$$\Delta_5=\begin{vmatrix} 3 & 1 & -1 & 2 & \boxed{1} \\ -2 & 3 & 1 & 4 & 3 \\ 1 & 4 & 2 & 3 & 1 \\ 5 & -2 & -3 & 5 & -1 \\ -1 & 1 & 2 & 3 & 2 \end{vmatrix}.$$

Решение. Принимая за главный элемент $a_{15}=1$, будем иметь:

$$\Delta_5=(-1)^{1+5}\begin{vmatrix} -2-3\cdot 3 & 3-1\cdot 3 & 1-(-1)\cdot 3 & 4-2\cdot 3 \\ 1-3\cdot 1 & 4-1\cdot 1 & 2-(-1)\cdot 1 & 3-2\cdot 1 \\ 5-3\cdot(-1) & -2-1\cdot(-1) & -3-(-1)(-1) & 5-2\cdot(-1) \\ -1-3\cdot 2 & 1-1\cdot 2 & 2-(-1)2 & 3-2\cdot 2 \end{vmatrix}=$$

$$=\begin{vmatrix} -11 & 0 & 4 & -2 \\ -2 & 3 & 3 & \boxed{1} \\ 8 & -1 & -4 & 7 \\ -7 & -1 & 4 & -1 \end{vmatrix}.$$

Далее, принимая за главный элемент $a_{24}=1$ и применяя аналогичное преобразование, получим:

$$\Delta_4=(-1)^6\begin{vmatrix}-15 & 6 & 10\\ 22 & -22 & -25\\ -9 & 2 & 7\end{vmatrix}=2\begin{vmatrix}-15 & 3 & 10\\ 22 & -11 & -25\\ -9 & \boxed{1} & 7\end{vmatrix}=$$

$$=2\cdot(-1)^{3+2}\begin{vmatrix}12 & -11\\ -77 & 52\end{vmatrix}=446.$$

Заметим, что число умножений и делений, нужных для вычисления определителя $n$-го порядка, равно [8]

$$\frac{n-1}{3}(n^2+n+3)$$

## Литература к седьмой главе


1. О. Шрейер и Е. Шпернер, Теория матриц, ОНТИ, 1936, § 1, 2.
2. А. Мальцев, Основы линейной алгебры, Изд. 2, Гостехиздат, 1956.
3. В. Н. Фаддеева, Вычислительные методы линейной алгебры, Гостехиздат, 1950.
4. Р. Фрезер, В. Дункан, А. Коллар, Теория матриц и ее приложения к дифференциальным уравнениям и динамике, ИЛ, 1950.
5. Б. В. Булгаков, Колебания, Гостехиздат, 1954, гл. I.
6. Е. С. Ляпин, Курс высшей алгебры, Учпедгиз, 1953, гл. IX.
7. Э. Уиттекер и Г. Робинсон, Математическая обработка результатов наблюдений, ОНТИ, 1935, гл. V.
8. Д. К. Фаддеев и В. Н. Фаддеева, Вычислительные методы линейной алгебры, Физматгиз, 1960, гл. II.

ГЛАВА VIII

# РЕШЕНИЕ СИСТЕМ ЛИНЕЙНЫХ УРАВНЕНИЙ

## § 1. Общая характеристика методов решения систем линейных уравнений

Способы решения систем линейных уравнений в основном разделяются на две группы: 1) *точные методы*, представляющие собой конечные алгорифмы для вычисления корней системы (таковы, например, правило Крамера, метод Гаусса, метод главных элементов, метод квадратных корней и др.), и 2) *итерационные методы*, позволяющие получать корни системы с заданной точностью путем сходящихся бесконечных процессов (к числу их относятся *метод итерации, метод Зейделя, метод релаксации* и др.).

Вследствие неизбежных округлений результаты даже точных методов являются приближенными, причем оценка погрешностей корней в общем случае затруднительна. При использовании итерационных процессов, сверх того, добавляется погрешность метода.

Заметим, что эффективное применение итерационных методов существенно зависит от удачного выбора начального приближения и быстроты сходимости процесса.

## § 2. Решение систем с помощью обратной матрицы. Формулы Крамера

Пусть дана система $n$ линейных уравнений с $n$ неизвестными

$$\left.\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n. \end{array}\right\} \qquad (1)$$

Обозначим через

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} \qquad (2)$$

матрицу из коэффициентов системы (1), через

$$\boldsymbol{b}=\begin{bmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{bmatrix} \tag{3}$$

— столбец ее свободных членов и через

$$\boldsymbol{x}=\begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \tag{4}$$

— столбец из неизвестных (искомый вектор). Тогда система (1) кратко может быть записана в виде матричного уравнения

$$A\boldsymbol{x}=\boldsymbol{b}. \tag{5}$$

Совокупность чисел $x_1, x_2, \ldots, x_n$ (или, короче, вектор $\boldsymbol{x}$), обращающих систему (1) в тождество, называется *решением* этой системы, а сами числа $x_i$ — ее *корнями*.

Если матрица $A$ — неособенная, т. е.

$$\det A=\begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{vmatrix}=\Delta\neq 0, \tag{6}$$

то система (1), или эквивалентное ей матричное уравнение (5), имеет единственное решение.

В самом деле, при условии $\det A\neq 0$ существует обратная матрица $A^{-1}$. Умножая обе части уравнения (5) слева на матрицу $A^{-1}$, получим

$$A^{-1}A\boldsymbol{x}=A^{-1}\boldsymbol{b}$$

или

$$\boldsymbol{x}=A^{-1}\boldsymbol{b}. \tag{7}$$

Формула (7), очевидно, дает решение уравнения (5), причем так как каждое решение имеет вид (7), то решение единственно.

Пример 1. Решить систему уравнений

$$\left.\begin{aligned} 3x_1-x_2&=5, \\ -2x_1+x_2+x_3&=0, \\ 2x_1-x_2+4x_3&=15. \end{aligned}\right\}$$

Решение. Запишем систему в матричной форме

$$\begin{bmatrix} 3 & -1 & 0 \\ -2 & 1 & 1 \\ 2 & -1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 5 \\ 0 \\ 15 \end{bmatrix}.$$

Определитель матрицы $A$ данной системы

$$\det A = \begin{vmatrix} 3 & -1 & 0 \\ -2 & 1 & 1 \\ 2 & -1 & 4 \end{vmatrix} = 5 \neq 0.$$

Вычисляя обратную матрицу $A^{-1}$, получим:

$$A^{-1} = \begin{bmatrix} 1 & \frac{4}{5} & -\frac{1}{5} \\ 2 & \frac{12}{5} & -\frac{3}{5} \\ 0 & \frac{1}{5} & \frac{1}{5} \end{bmatrix}.$$

Отсюда

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 & \frac{4}{5} & -\frac{1}{5} \\ 2 & \frac{12}{5} & -\frac{3}{5} \\ 0 & \frac{1}{5} & \frac{1}{5} \end{bmatrix} \begin{bmatrix} 5 \\ 0 \\ 15 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}.$$

Значит, $x_1 = 2$; $x_2 = 1$; $x_3 = 3$.

Для матрицы $A$ порядка $n > 4$ непосредственное нахождение обратной матрицы $A^{-1}$ требует много времени. Поэтому формула (7) редко употребляется на практике.

Пользуясь формулой (7), легко получить формулы для неизвестных системы (1). Как известно (гл. VII, § 4),

$$A^{-1} = \frac{1}{\Delta} \tilde{A},$$

где

$$\tilde{A} = \begin{bmatrix} A_{11} & A_{21} & \dots & A_{n1} \\ \cdot & \cdot & \cdot & \cdot \\ A_{1n} & A_{2n} & \dots & A_{nn} \end{bmatrix}$$

— матрица, союзная с $A$ ($A_{ij}$ — алгебраические дополнения элементов $a_{ij}$). Поэтому

$$\boldsymbol{x} = \frac{1}{\Delta} \tilde{A} \boldsymbol{b}$$

или

$$\begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = \frac{1}{\Delta} \begin{bmatrix} \Delta_1 \\ \Delta_2 \\ \cdot \\ \cdot \\ \cdot \\ \Delta_n \end{bmatrix}, \tag{8}$$

где

$$\Delta_i = \sum_{j=1}^{n} A_{ji} b_j = \begin{vmatrix} a_{11} & \dots & a_{1,\,i-1} & b_1 & a_{1,\,i+1} & \dots & a_{1n} \\ a_{21} & \dots & a_{2,\,i-1} & b_2 & a_{2,\,i+1} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & \dots & a_{n,\,i-1} & b_n & a_{n,\,i+1} & \dots & a_{nn} \end{vmatrix}$$

— определители, получающиеся из определителя $\Delta$ [формула (6)] путем замены его $i$-го столбца столбцом свободных членов системы (1). Из равенства (8) получаем *формулы Крамера*

$$x_1 = \frac{\Delta_1}{\Delta}, \quad x_2 = \frac{\Delta_2}{\Delta}, \quad \dots, \quad x_n = \frac{\Delta_n}{\Delta}. \tag{9}$$

Следовательно, если определитель системы (1) $\Delta \neq 0$, то система имеет единственное решение $\boldsymbol{x}$, определяемое матричной формулой (7) или эквивалентными ей скалярными формулами (9).

Пример 2. Решить систему линейных уравнений

$$\left.\begin{aligned} 2x_1 + x_2 - 5x_3 + x_4 &= 8, \\ x_1 - 3x_2 - 6x_4 &= 9, \\ 2x_2 - x_3 + 2x_4 &= -5, \\ x_1 + 4x_2 - 7x_3 + 6x_4 &= 0. \end{aligned}\right\}$$

Решение. Определитель этой системы

$$\Delta = \begin{vmatrix} 2 & 1 & -5 & 1 \\ 1 & -3 & 0 & -6 \\ 0 & 2 & -1 & 2 \\ 1 & 4 & -7 & 6 \end{vmatrix} = 27 \neq 0.$$

Вычисляя дополнительные определители, получим:

$$\Delta_1 = \begin{vmatrix} 8 & 1 & -5 & 1 \\ 9 & -3 & 0 & -6 \\ -5 & 2 & -1 & 2 \\ 0 & 4 & -7 & 6 \end{vmatrix} = 81;$$

$$\Delta_2 = \begin{vmatrix} 2 & 8 & -5 & 1 \\ 1 & 9 & 0 & -6 \\ 0 & -5 & -1 & 2 \\ 1 & 0 & -7 & 6 \end{vmatrix} = -108;$$

$$\Delta_3 = \begin{vmatrix} 2 & 1 & 8 & 1 \\ 1 & -3 & 9 & -6 \\ 0 & 2 & -5 & 2 \\ 1 & 4 & 0 & 6 \end{vmatrix} = -27;$$

$$\Delta_4 = \begin{vmatrix} 2 & 1 & -5 & 8 \\ 1 & -3 & 0 & 9 \\ 0 & 2 & -1 & -5 \\ 1 & 4 & -7 & 0 \end{vmatrix} = 27.$$

Отсюда

$$x_1 = \frac{\Delta_1}{\Delta} = \frac{81}{27} = 3;$$

$$x_2 = \frac{\Delta_2}{\Delta} = -\frac{108}{27} = -4;$$

$$x_3 = \frac{\Delta_3}{\Delta} = -\frac{27}{27} = -1;$$

$$x_4 = \frac{\Delta_4}{\Delta} = \frac{27}{27} = 1.$$

Таким образом, решение линейной системы (1) с $n$ неизвестными сводится к вычислению $(n+1)$-го определителя порядка $n$. Если число $n$ велико, то вычисление определителей является трудоемкой операцией. Поэтому разработаны прямые приемы нахождения корней линейной системы.

## § 3. Метод Гаусса

Наиболее распространенным приемом решения систем линейных уравнений является алгорифм последовательного исключения неизвестных. Этот метод носит название *метода Гаусса*. Для простоты рассуждений ограничимся рассмотрением системы четырех уравнений с четырьмя неизвестными

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 &= a_{15}, \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 &= a_{25}, \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4 &= a_{35}, \\ a_{41}x_1 + a_{42}x_2 + a_{43}x_3 + a_{44}x_4 &= a_{45}. \end{aligned}\right\} \qquad (1)$$

Пусть $a_{11} \neq 0$ (ведущий элемент). Разделив коэффициенты первого уравнения системы (1) на $a_{11}$, получим:

$$x_1 + b_{12}x_2 + b_{13}x_3 + b_{14}x_4 = b_{15}, \qquad (2)$$

где

$$b_{1j} = \frac{a_{1j}}{a_{11}} \qquad (j > 1).$$

Пользуясь уравнением (2), легко исключить из системы (1) неизвестную $x_1$. Для этого достаточно из второго уравнения системы (1) вычесть уравнение (2), умноженное на $a_{21}$, из третьего уравнения системы (1) вычесть уравнение (2), умноженное на $a_{31}$, и т. д. В результате получим систему из трех уравнений

$$\left.\begin{aligned} a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + a_{24}^{(1)}x_4 &= a_{25}^{(1)}, \\ a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 + a_{34}^{(1)}x_4 &= a_{35}^{(1)}, \\ a_{42}^{(1)}x_2 + a_{43}^{(1)}x_3 + a_{44}^{(1)}x_4 &= a_{45}^{(1)}, \end{aligned}\right\} \qquad (1')$$

где коэффициенты $a_{ij}^{(1)}$ $(i,\ j \geqslant 2)$ вычисляются по формуле

$$a_{ij}^{(1)} = a_{ij} - a_{i1}b_{1j} \qquad (i,\ j \geqslant 2).$$

Разделив, далее, коэффициенты первого уравнения системы (1′) на «ведущий элемент» $a_{22}^{(1)}$, получим уравнение

$$x_2 + b_{23}^{(1)}x_3 + b_{24}^{(1)}x_4 = b_{25}^{(1)}, \qquad (2')$$

где

$$b_{2j}^{(1)} = \frac{a_{2j}^{(1)}}{a_{22}^{(1)}} \qquad (j > 2).$$

Исключая теперь $x_2$ таким же способом, каким мы исключили $x_1$, придем к следующей системе уравнений:

$$\left.\begin{aligned} a_{33}^{(2)}x_3 + a_{34}^{(2)}x_4 &= a_{35}^{(2)}, \\ a_{43}^{(2)}x_3 + a_{44}^{(2)}x_4 &= a_{45}^{(2)}, \end{aligned}\right\} \qquad (1'')$$

где

$$a_{ij}^{(2)} = a_{ij}^{(1)} - a_{i2}^{(1)}\, b_{2j}^{(1)} \qquad (i,\ j \geqslant 3).$$

Разделив коэффициенты первого уравнения системы (1″) на «ведущий элемент» $a_{33}^{(2)}$, получим:

$$x_3 + b_{34}^{(2)}x_4 = b_{35}^{(2)}, \qquad (2'')$$

где

$$b_{3j}^{(2)} = \frac{a_{3j}^{(2)}}{a_{33}^{(2)}} \qquad (j > 3).$$

Исключив теперь $x_3$ аналогичным путем из системы (1″), будем иметь:

$$a_{44}^{(3)}x_4 = a_{45}^{(3)}, \qquad (1''')$$

где

$$a_{ij}^{(3)} = a_{ij}^{(2)} - a_{i3}^{(2)}\, b_{3j}^{(2)} \qquad (i,\ j \geqslant 4).$$

Отсюда

$$x_4 = \frac{a_{45}^{(3)}}{a_{44}^{(3)}} = b_{45}^{(3)}. \qquad (2''')$$

Остальные неизвестные последовательно определяются из уравнений (2″), (2′) и (2):

$$x_3 = b_{35}^{(2)} - b_{34}^{(2)} x_4,$$

$$x_2 = b_{25}^{(1)} - b_{24}^{(1)} x_4 - b_{23}^{(1)} x_3,$$

$$x_1 = b_{15} - b_{14} x_4 - b_{13} x_3 - b_{12} x_2.$$

Таким образом, процесс решения линейной системы по методу Гаусса сводится к построению эквивалентной системы (2), (2′), (2″), (2‴), имеющей треугольную матрицу. Необходимым и достаточным условием применимости метода является неравенство нулю всех «ведущих элементов». Вычисления удобно поместить в таблицу 13. Приведенная в ней схема называется *схемой единственного деления*. Процесс нахождения коэффициентов $b_{ij}^{(i-1)}$ треугольной системы обычно называется *прямым ходом*, процесс получения значений неизвестных — *обратным ходом*.

Прямой ход начинается с выписывания коэффициентов системы, включая свободные члены (раздел $A$). Последняя строка раздела $A$ схемы представляет собой результат деления первой строки раздела на «ведущий элемент» $a_{11}$. Элементы $a_{ij}^{(1)}$ ($i$, $j \geqslant 2$) следующего раздела схемы (раздел $A_1$) равны соответствующим элементам $a_{ij}$ предшествующего раздела без произведения их «проекций» на ряды раздела $A$, содержащие элемент 1 (т. е. на первый столбец и последнюю строку).

Последняя строка раздела $A_1$ находится путем деления первой строки раздела на «ведущий элемент» $a_{22}^{(1)}$. Аналогично строятся следующие разделы. Прямой ход заканчивается, когда мы дойдем до раздела, состоящего из одной строки, не считая преобразованной (раздел $A_3$ в нашем частном случае).

При обратном ходе используются лишь строки разделов $A_i$, содержащие единицы (*отмеченные строки*), начиная с последней. Элемент $b_{45}^{(3)}$ из раздела $A_3$, стоящий в столбце свободных членов отмеченной строки раздела, дает значение $x_4$. Далее, все остальные неизвестные $x_i$ ($i = 3, 2, 1$) шаг за шагом находятся с помощью вычитания из свободного члена отмеченной строки суммы произведений ее коэффициентов на соответствующие значения ранее найденных неизвестных. Значения неизвестных последовательно выписываются в последний раздел $B$. Расставленные там единицы помогают находить для $x_i$ соответствующие коэффициенты в отмеченных строках.

Таблица 13

**Схема единственного деления**

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | Свободные члены | $\Sigma$ | Разделы схемы |
|---|---|---|---|---|---|---|
| $a_{11}$<br>$a_{21}$<br>$a_{31}$<br>$a_{41}$<br>1 | $a_{12}$<br>$a_{22}$<br>$a_{32}$<br>$a_{42}$<br>$b_{12}$ | $a_{13}$<br>$a_{23}$<br>$a_{33}$<br>$a_{43}$<br>$b_{13}$ | $a_{14}$<br>$a_{24}$<br>$a_{34}$<br>$a_{44}$<br>$b_{14}$ | $a_{15}$<br>$a_{25}$<br>$a_{35}$<br>$a_{45}$<br>$b_{15}$ | $a_{16}$<br>$a_{26}$<br>$a_{36}$<br>$a_{46}$<br>$b_{16}$ | $A$ |
| | $a_{22}^{(1)}$<br>$a_{32}^{(1)}$<br>$a_{42}^{(1)}$<br>1 | $a_{23}^{(1)}$<br>$a_{33}^{(1)}$<br>$a_{43}^{(1)}$<br>$b_{23}^{(1)}$ | $a_{24}^{(1)}$<br>$a_{34}^{(1)}$<br>$a_{44}^{(1)}$<br>$b_{24}^{(1)}$ | $a_{25}^{(1)}$<br>$a_{35}^{(1)}$<br>$a_{45}^{(1)}$<br>$b_{25}^{(1)}$ | $a_{26}^{(1)}$<br>$a_{36}^{(1)}$<br>$a_{46}^{(1)}$<br>$b_{26}^{(1)}$ | $A_1$ |
| | | $a_{33}^{(2)}$<br>$a_{43}^{(2)}$<br>1 | $a_{34}^{(2)}$<br>$a_{44}^{(2)}$<br>$b_{34}^{(2)}$ | $a_{35}^{(2)}$<br>$a_{45}^{(2)}$<br>$b_{35}^{(2)}$ | $a_{36}^{(2)}$<br>$a_{46}^{(2)}$<br>$b_{36}^{(2)}$ | $A_2$ |
| | | | $a_{44}^{(3)}$<br>1 | $a_{45}^{(3)}$<br>$b_{45}^{(3)}$<br>$(x_4)$ | $a_{46}^{(3)}$<br>$b_{46}^{(3)}$<br>$(\overline{x_4})$ | $A_3$ |
| 1 | 1 | 1 | 1 | $x_4$<br>$x_3$<br>$x_2$<br>$x_1$ | $\overline{x_4}$<br>$\overline{x_3}$<br>$\overline{x_2}$<br>$\overline{x_1}$ | $B$ |

Для контроля вычислений используются так называемые «контрольные суммы»

$$a_{i6} = \sum_{j=1}^{5} a_{ij} \qquad (i = 1,\ 2,\ \ldots,\ 5), \tag{3}$$

помещенные в столбце $\sum$ и представляющие собой сумму элементов строк матрицы исходной системы (1), включая свободные члены.

Если $a_{i6}$ принять за новые свободные члены в системе (1), то преобразованная линейная система

$$\sum_{j=1}^{4} a_{ij}\bar{x}_j = a_{i6} \qquad (i=1,\ 2,\ 3,\ 4) \tag{4}$$

будет иметь неизвестные $\bar{x}_j$, связанные с прежними неизвестными $x_j$ соотношениями

$$\bar{x}_j = x_j + 1 \qquad (j=1,\ 2,\ 3,\ 4). \tag{5}$$

В самом деле, подставляя формулы (5) в уравнение (4), в силу системы (1) и формул (3) получим тождество

$$\sum_{j=1}^{4} a_{ij}x_j + \sum_{j=1}^{4} a_{ij} = \sum_{j=1}^{5} a_{ij} \equiv a_{i6} \qquad (j=1,\ 2,\ 3,\ 4).$$

Вообще, если над контрольными суммами в каждой строке проделывать те же операции, что и над остальными элементами этой строки, то при отсутствии ошибок в вычислениях элементы столбца $\sum$ равны суммам элементов соответствующих преобразованных строк. Это обстоятельство служит контролем прямого хода. Обратный ход контролируется нахождением чисел $\bar{x}_j$, которые должны совпадать с числами $x_j + 1$.

Пример. Решить систему

$$\left.\begin{aligned} 7{,}9x_1 + 5{,}6x_2 + 5{,}7x_3 - 7{,}2x_4 &= 6{,}68;\\ 8{,}5x_1 - 4{,}8x_2 + 0{,}8x_3 + 3{,}5x_4 &= 9{,}95;\\ 4{,}3x_1 + 4{,}2x_2 - 3{,}2x_3 + 9{,}3x_4 &= 8{,}6;\\ 3{,}2x_1 - 1{,}4x_2 - 8{,}9x_3 + 3{,}3x_4 &= 1. \end{aligned}\right\} \tag{6}$$

Решение. В раздел $A$ таблицы 14 впишем матрицу коэффициентов системы, ее свободные члены и контрольные суммы. Далее, заполняем последнюю (пятую) строку раздела $A$, деля первую строку на 7,9 (на $a_{11}$).

Переходим к заполнению раздела $A_1$ таблицы. Взяв любой элемент раздела $A$ (не находящийся в первой строке), вычитаем из него произведение первого элемента его строки на последний элемент столбца, к которому он принадлежит, и записываем на соответствующее место в разделе $A_1$ схемы. Например, выбрав $a_{43} = -8{,}9$, будем иметь:

$$a_{43}^{(1)} = a_{43} - a_{41}b_{13} = -8{,}9 - 3{,}2 \cdot 0{,}72152 = -11{,}20886.$$

Чтобы получить последнюю строку раздела $A_1$, делим все члены первой строки этого раздела на $a_{22}^{(1)} = -10{,}82531$. Например,

$$b_{23}^{(1)} = \frac{a_{23}^{(1)}}{a_{22}^{(1)}} = \frac{-5{,}33292}{-10{,}82531} = 0{,}49263.$$

Таблица 14

**Решение системы по схеме единственного деления**

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | Свободные разделы | $\Sigma$ | Разделы схемы |
|---|---|---|---|---|---|---|
| 7,9 | 5,6 | 5,7 | —7,2 | 6,68 | 18,68 | $A$ |
| 8,5 | —4,8 | 0,8 | 3,5 | 9,95 | 17,95 | |
| 4,3 | 4,2 | —3,2 | 9,3 | 8,6 | 23,2 | |
| 3,2 | —1,4 | —8,9 | 3,3 | 1 | —2,8 | |
| 1 | 0,70886 | 0,72152 | —0,91139 | 0,84557 | 2,36456 | |
| | —10,82531 | —5,33292 | 11,24682 | 2,76265 | —2,14876 | $A_1$ |
| | 1,15190 | —6,30254 | 13,21898 | 4,96405 | 13,03239 | |
| | —3,66835 | —11,20886 | 6,21645 | —1,70582 | —10,36658 | |
| | 1 | 0,49263 | —1,03894 | —0,25520 | 0,19849 | |
| | | —6,87000 | 14,41573 | 5,25801 | 12,80374 | $A_2$ |
| | | —9,40172 | 2,40525 | —2,64198 | —9,63845 | |
| | | 1 | —2,09836 | —0,76536 | —1,86372 | |
| | | | —17,32294 | —9,83768 | —27,16062 | $A_3$ |
| | | | 1 | 0,56790 | 1,56790 | |
| | | | 1 | 0,56790 | 1,56790 | $B$ |
| | | 1 | | 0,42630 | 1,42630 | |
| | 1 | | | 0,12480 | 1,12480 | |
| 1 | | | | 0,96710 | 1,96710 | |

Аналогичным путем заполняются остальные разделы таблицы. Например,

$$a_{44}^{(2)} = a_{44}^{(1)} - a_{42}^{(1)} b_{24}^{(1)} = \\ = 6{,}21645 - (-3{,}66835)\cdot(-1{,}03894) = 2{,}40525.$$

Для нахождения неизвестных используем строки, содержащие единицы, начиная с последней (*отмеченные строки*). Неизвестное $x_4$ представляет собой свободный член последней строки раздела $A_3$:

$$x_4 = b_{45}^{(3)} = 0{,}56790.$$

Значения остальных неизвестных $x_3$, $x_2$, $x_1$ получаются последовательно в результате вычитания из свободных членов отмеченных

строк суммы произведений соответствующих коэффициентов $b_{ij}^{(\cdot)}$ на ранее найденные значения неизвестных.

Имеем:

$$x_3 = b_{35}^{(2)} - b_{34}^{(2)} x_4 = -0{,}76536 - (-2{,}09836) \cdot 0{,}56790 = 0{,}42630;$$

$$x_2 = b_{25}^{(1)} - b_{24}^{(1)} x_4 - b_{23}^{(1)} x_3 =$$
$$= -0{,}25520 - (-1{,}03894) \cdot 0{,}56790 - 0{,}49263 \cdot 0{,}42630 = 0{,}12480;$$

$$x_1 = b_{15} - b_{14} x_4 - b_{13} x_3 - b_{12} x_2 = 0{,}84557 - (-0{,}91139) \times$$
$$\times 0{,}56790 - 0{,}72152 \cdot 0{,}42630 - 0{,}70886 \cdot 0{,}12480 = 0{,}96710.$$

Итак,

$$x_1 = 0{,}96710; \quad x_2 = 0{,}12480; \quad x_3 = 0{,}42630; \quad x_4 = 0{,}56790.$$

Текущий контроль вычислений осуществляется с помощью столбца $\sum$, над которым производятся те же действия, что и над остальными столбцами.

В результате: 1) сумма элементов каждой строки схемы (не принадлежащих столбцу $\sum$) должна быть равна элементу этой строки из столбца $\sum$; 2) корни $\bar{x}_i$, соответствующие столбцу $\sum$, должны быть на единицу больше соответствующих корней системы.

Кстати, если учесть единицы, написанные в разделе $B$, то опять получится, что и в этом разделе элементы столбца $\sum$ являются суммами элементов отвечающих им строк. В нашем случае первое и второе условия выполняются с точностью до единицы последнего разряда. Следовательно, почти достоверно, что вычисления выполнены правильно.

Заметим, что если матрица системы — симметрическая, то соответствующие части разделов $A$, $A_1$, $A_2$, ... схемы единственного деления также получаются симметрическими. Это обстоятельство можно использовать для упрощения таблицы.

Нетрудно оценить число $N$ арифметических действий, необходимых для решения линейной системы с $n$ неизвестными методом Гаусса [5] (не учитывая контроля).

Для прямого хода требуется следующее число умножений и делений:

$$n(n+1) + (n-1)n + \ldots + 1 \cdot 2 = (1^2 + 2^2 + \ldots + n^2) +$$
$$+ (1 + 2 + \ldots + n) = \frac{n(n+1)(n+2)}{3}$$

и столько же вычитаний. Для обратного хода потребуется $\frac{n(n-1)}{2}$ умножений и делений и такое же количество вычитаний. Следова-

тельно, общее количество арифметических действий в методе Гаусса есть

$$N=\frac{2n(n+1)(n+2)}{3}+n(n-1)<n^3$$

при $n>7$.

Таким образом, время, необходимое для решения линейной системы методом Гаусса, примерно пропорционально кубу числа неизвестных. Например, для решения методом Гаусса системы 100 линейных уравнений со 100 неизвестными на быстродействующей машине, выполняющей $10^4$ операций в секунду, потребуется время

$$T=10^6\cdot 10^{-4}=100 \text{ сек.}$$

Фактическое машинное время будет значительно больше ввиду наличия в программе, кроме арифметических действий, других операций (переадресация, логические операции, пересылки, формирование и др.).

## § 4. Уточнение корней

Полученные методом Гаусса приближенные значения корней можно уточнить. Покажем, как это делается, если поправки корней малы по абсолютной величине.

Пусть для системы

$$Ax=b$$

найдено приближенное решение $x_0$. Полагая

$$x=x_0+\delta,$$

для поправки $\delta=\begin{bmatrix}\delta_1\\ \cdot\\ \cdot\\ \cdot\\ \delta_n\end{bmatrix}$ корня $x_0$ будем иметь уравнение

$$A(x_0+\delta)=b$$

или

$$A\delta=\varepsilon,$$

где $\varepsilon=b-Ax_0$ — *невязка для приближенного решения* $x_0$. Таким образом, чтобы найти $\delta$, нужно решить линейную систему с прежней матрицей $A$ и новым свободным членом $\varepsilon=\begin{bmatrix}\varepsilon_1\\ \cdot\\ \cdot\\ \cdot\\ \varepsilon_n\end{bmatrix}$. Для этого достаточно к основной схеме вычислений присоединить столбец $\varepsilon$ свободных членов и преобразовать его по общим правилам. Поправки $\delta_1$, $\delta_2$, ..., $\delta_n$, как обычно, определяются из отмеченных строк, причем коэффициенты при этих неизвестных поправках уже имеются готовыми

в таблице. Заметим, что преобразованные коэффициенты матрицы $A$ можно не уточнять, так как при малых невязках соответствующие ошибки будут обладать более высоким порядком малости.

Пример. Решить методом Гаусса с тремя знаками (например, на счетной линейке или вручную) систему

$$\left.\begin{aligned} 6x_1 - x_2 - x_3 &= 11{,}33; \\ -x_1 + 6x_2 - x_3 &= 32; \\ -x_1 - x_2 + 6x_3 &= 42. \end{aligned}\right\} \qquad (1)$$

Используя полученные значения как начальные приближения, уточнить корни до $10^{-4}$.

Решение. Применяем обычную схему единственного деления (таблица 15), выполняя все действия с тремя значащими цифрами.

Таблица 15

**Уточнение корней, вычисленных методом Гаусса**

| $x_1$ | $x_2$ | $x_3$ | Свободные члены | $\Sigma$ | Невязка $\varepsilon$ |
|---|---|---|---|---|---|
| 6 | —1 | —1 | 11,33 | 15,33 | —0,02 |
| —1 | 6 | —1 | 32 | 36 | 0 |
| —1 | —1 | 6 | 42 | 46 | —0,01 |
| 1 | —0,167 | —0,167 | 1,89 | 2,56 | —0,0033 |
| | 5,83 | —1,17 | 33,9 | 38,6 | —0,0033 |
| | —1,17 | 5,83 | 43,9 | 48,6 | —0,0133 |
| | 1 | —0,200 | 5,80 | 6,60 | —0,0006 |
| | | 5,60 | 50,7 | 56,3 | —0,0140 |
| | | 1 | 9,05<br>9,0475 | 10,05 | —0,0025 |
| | 1 | | 7,62<br>7,6189 | 8,62 | —0,0011 |
| 1 | | | 4,67<br>4,6661 | | —0,0039 |

Имеем приближенные значения корней:

$$x_1^{(0)} = 4{,}67; \quad x_2^{(0)} = 7{,}62; \quad x_3^{(0)} = 9{,}05.$$

Подставляя эти числа в данную систему (1), вычисляем соответствующие невязки (т. е. разности между правой и левой частями системы (1))

$$\varepsilon_1^{(0)} = -0{,}02; \quad \varepsilon_2^{(0)} = 0; \quad \varepsilon_3^{(0)} = -0{,}01.$$

Используя эти значения в качестве свободных членов (таблица 15), получаем поправки корней

$$\delta_1^{(0)} = -0{,}0039; \quad \delta_2^{(0)} = -0{,}0011; \quad \delta_3^{(0)} = -0{,}0025.$$

Отсюда находим уточненные значения корней

$$x_1 = 4{,}6661; \quad x_2 = 7{,}6189; \quad x_3 = 9{,}0475,$$

причем невязки равны

$$\delta_1 = -2 \cdot 10^{-4}, \quad \delta_2 = 2 \cdot 10^{-4}, \quad \delta_3 = 0.$$

Иногда требуется определить возможную ошибку $\Delta \boldsymbol{x}$ корня $\boldsymbol{x}$ линейной системы по известным малым ошибкам $\Delta A$ и $\Delta \boldsymbol{b}$ матрицы $A$ системы и ее свободного члена $\boldsymbol{b}$.

Имеем:

$$A\boldsymbol{x} = \boldsymbol{b} \tag{2}$$

и

$$(A + \Delta A)(\boldsymbol{x} + \Delta \boldsymbol{x}) = \boldsymbol{b} + \Delta \boldsymbol{b}. \tag{3}$$

Отсюда, пренебрегая малым членом $\Delta A \cdot \Delta \boldsymbol{x}$, получим:

$$A\boldsymbol{x} + A\Delta \boldsymbol{x} + \Delta A \boldsymbol{x} = \boldsymbol{b} + \Delta \boldsymbol{b}$$

или

$$A\Delta \boldsymbol{x} = \Delta \boldsymbol{b} - \Delta A \boldsymbol{x}. \tag{4}$$

Таким образом, для приближенного нахождения $\Delta \boldsymbol{x}$ можно использовать схему Гаусса для основной системы (2), пополнив эту схему новым столбцом свободных членов $\Delta \boldsymbol{b} - \Delta A \boldsymbol{x}$.

## § 5. Метод главных элементов

Пусть дана линейная система

$$\left.\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = a_{1,\ n+1}, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = a_{2,\ n+1}, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = a_{n,\ n+1}. \end{array}\right\} \tag{1}$$

Рассмотрим расширенную прямоугольную матрицу, состоящую из коэффициентов системы и ее свободных членов,

$$M=\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1j} & \dots & a_{1q} & \dots & a_{1n} & a_{1,\,n+1} \\ a_{21} & a_{22} & \dots & a_{2j} & \dots & a_{2q} & \dots & a_{2n} & a_{2,\,n+1} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{i1} & a_{i2} & \dots & a_{ij} & \dots & a_{iq} & \dots & a_{in} & a_{i,\,n+1} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{p1} & a_{p2} & \dots & a_{pj} & \dots & \boxed{a_{pq}} & \dots & a_{pn} & a_{p,\,n+1} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nj} & \dots & a_{nq} & \dots & a_{nn} & a_{n,\,n+1} \end{bmatrix}.$$

Выберем ненулевой, как правило, наибольший по модулю, не принадлежащий к столбцу свободных членов ($q \neq n+1$) элемент $a_{pq}$ матрицы $M$, который называется *главным элементом*, и вычислим множители

$$m_i = -\frac{a_{iq}}{a_{pq}}$$

для всех $i \neq p$.

Строка с номером $p$ матрицы $M$, содержащая главный элемент, называется *главной строкой*. Далее, произведем следующую операцию: к каждой неглавной строке прибавим главную строку, умноженную на соответствующий множитель $m_i$ для этой строки. В результате мы получим новую матрицу, у которой $q$-й столбец состоит из нулей. Отбрасывая этот столбец и главную $p$-ю строку, получим новую матрицу $M^{(1)}$ с меньшим на единицу числом строк и столбцов.

Над матрицей $M^{(1)}$ повторяем те же операции, после чего получаем матрицу $M^{(2)}$, и т. д. Таким образом, мы построим последовательность матриц

$$M,\ M^{(1)},\ \dots,\ M^{(n-1)},$$

последняя из которых представляет двучленную матрицу-строку; ее также считаем главной строкой.

Для определения неизвестных $x_i$ объединяем в систему все главные строки, начиная с последней, входящей в матрицу $M^{(n-1)}$.

После надлежащего изменения нумерации неизвестных получается система с треугольной матрицей, из которой легко шаг за шагом найти неизвестные данной системы (1). Метод главных элементов всегда применим, если определитель системы

$$\det A = \begin{vmatrix} a_{11} & \dots & a_{1n} \\ \cdot & \cdot & \cdot \\ a_{n1} & \dots & a_{nn} \end{vmatrix} \neq 0.$$

Заметим, что метод Гаусса является частным случаем метода главных элементов, а схема метода Гаусса получается, если за главный элемент всегда выбирать левый верхний элемент соответствующей матрицы.

## § 6. Применение метода Гаусса для вычисления определителей

Пусть

$$A=\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \tag{1}$$

и

$$\Delta=\det A. \tag{2}$$

Рассмотрим линейную систему

$$Ax=\mathbf{0}. \tag{3}$$

При решении системы (3) по методу Гаусса мы заменяли матрицу $A$ треугольной матрицей $B$, состоящей из элементов отмеченных строк,

$$B=\begin{bmatrix} 1 & b_{12} & b_{13} & \dots & b_{1n} \\ 0 & 1 & b_{23}^{(1)} & \dots & b_{2n}^{(1)} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}.$$

В результате получалась эквивалентная система

$$Bx=\mathbf{0}. \tag{4}$$

Элементы матрицы $B$ последовательно получались из элементов матрицы $A$ и дальнейших вспомогательных матриц $A_1, A_2, \dots, A_{n-1}$ с помощью следующих элементарных преобразований:

1) деления на «ведущие» элементы $a_{11}, a_{22}^{(1)}, \dots, a_{nn}^{(n-1)}$, которые предполагались отличными от нуля, и

2) вычитания из строк матрицы $A$ и промежуточных матриц $A_i$ ($i=1, 2, \dots, n-1$) чисел, пропорциональных элементам соответствующих ведущих строк. При первой операции определитель матрицы также делится на соответствующий «ведущий» элемент, при второй — определитель матрицы остается неизменным. Поэтому

$$\det B=1=\frac{\det A}{a_{11}a_{22}^{(1)}\ \dots\ a_{nn}^{(n-1)}}.$$

Следовательно,

$$\Delta=\det A=a_{11}a_{22}^{(1)}\ \dots\ a_{nn}^{(n-1)}, \tag{5}$$

т. е. *определитель равен произведению «ведущих» элементов для соответствующей схемы Гаусса.* Отсюда заключаем, что приведенная

нами в § 3 схема единственного деления может быть использована для вычисления определителей, причем столбец свободных членов тогда становится излишним.

Заметим, что если для какого-нибудь шага элемент $a_{ii}^{(i-1)}=0$ или близок к нулю (что влечет за собой уменьшение точности вычислений), то следует соответствующим образом изменить порядок строк и столбцов матрицы.

Пример. Вычислить определитель

$$\Delta=\begin{vmatrix} 7{,}4 & 2{,}2 & -3{,}1 & 0{,}7 \\ 1{,}6 & 4{,}8 & -8{,}5 & 4{,}5 \\ 4{,}7 & 7{,}0 & -6{,}0 & 6{,}6 \\ 5{,}9 & 2{,}7 & 4{,}9 & -5{,}3 \end{vmatrix}.$$

Решение. Используя элементы определителя $\Delta$, составляем схему единственного деления (таблица 16).

Таблица 16

**Вычисление определителя методом Гаусса**

| 1-й столбец | 2-й столбец | 3-й столбец | 4-й столбец | $\Sigma$ | |
|---|---|---|---|---|---|
| $\boxed{7{,}4}$ | 2,2 | —3,1 | 0,7 | 7,2 | |
| 1,6 | 4,8 | —8,5 | 4,5 | 2,4 | |
| 4,7 | 7,0 | —6,0 | 6,6 | 12,3 | $A$ |
| 5,9 | 2,7 | 4,9 | —5,3 | 8,2 | |
| 1 | 0,29729 | —0,41891 | 0,09459 | 0,97297 | |
| | $\boxed{4{,}32434}$ | —7,82974 | 4,34866 | 0,84326 | |
| | 5,60274 | —4,03112 | 6,15543 | 7,72705 | |
| | 0,94599 | 7,37157 | —5,85808 | 2,45948 | $A_1$ |
| | 1 | —1,81062 | 1,00562 | 0,19500 | |
| | | $\boxed{6{,}11331}$ | 0,52120 | 6,63451 | |
| | | 9,08440 | —6,80939 | 2,27501 | $A_2$ |
| | | 1 | 0,08526 | 1,08526 | |
| | | | $\boxed{-7{,}58393}$ | —7,58393 | $A_3$ |
| | | | | $\Delta=-1483{,}61867$ | |

Перемножая «ведущие» элементы (заключенные в рамки), получим:

$$\Delta = 7{,}4 \cdot 4{,}32434 \cdot 6{,}11331 \cdot (-7{,}58393) = -1483{,}61867.$$

Обратим внимание на следующее обстоятельство. Для того чтобы решить систему $n$ линейных уравнений с $n$ неизвестными по формулам Крамера, нужно вычислить $n+1$ определителей $n$-го порядка. Между тем для вычисления одного определителя $n$-го порядка по схеме единственного деления требуется почти такой же объем работы, как и для полного решения системы уравнений. Поэтому пользоваться формулами Крамера для численного решения линейной системы при $n > 3$, вообще говоря, нецелесообразно.

## § 7. Вычисление обратной матрицы методом Гаусса

Пусть дана неособенная матрица

$$A = [a_{ij}] \quad (i,\ j = 1,\ 2,\ \ldots,\ n). \tag{1}$$

Для нахождения ее обратной матрицы

$$A^{-1} = [x_{ij}] \tag{2}$$

используем основное соотношение

$$AA^{-1} = E, \tag{3}$$

где $E$ — единичная матрица.

Перемножая матрицы $A$ и $A^{-1}$, будем иметь $n$ систем уравнений относительно $n^2$ неизвестных $x_{ij}$

$$\sum_{k=1}^{n} a_{ik} x_{kj} = \delta_{ij} \quad (i,\ j = 1,\ 2,\ \ldots,\ n),$$

где

$$\delta_{ij} = \begin{cases} 1, & \text{когда } i = j, \\ 0, & \text{когда } i \neq j. \end{cases}$$

Полученные $n$ систем линейных уравнений для $j = 1,\ 2,\ \ldots,\ n$, имеющих одну и ту же матрицу $A$ и различные свободные члены, одновременно можно решить методом Гаусса.

Пример. Найти обратную матрицу $A^{-1}$ для матрицы

$$A = \begin{bmatrix} 1{,}8 & -3{,}8 & 0{,}7 & -3{,}7 \\ 0{,}7 & 2{,}1 & -2{,}6 & -2{,}8 \\ 7{,}3 & 8{,}1 & 1{,}7 & -4{,}9 \\ 1{,}9 & -4{,}3 & -4{,}9 & -4{,}7 \end{bmatrix}.$$

Решение. Составим схему с единственным делением. При этом будем иметь четыре столбца свободных членов (таблица 17). Заметим, что элементы строк обратной матрицы получаются в обратном порядке.

Таблица 17

**Вычисление обратной матрицы методом Гаусса**

| $x_{1j}$ | $x_{2j}$ | $x_{3j}$ | $x_{4j}$ | $j=1$ | $j=2$ | $j=3$ | $j=4$ | $\Sigma$ |
|---|---|---|---|---|---|---|---|---|
| 1,8 | —3,8 | 0,7 | —3,7 | 1 | 0 | 0 | 0 | —4,0 |
| 0,7 | 2,1 | —2,6 | —2,8 | 0 | 1 | 0 | 0 | —1,6 |
| 7,3 | 8,1 | 1,7 | —4,9 | 0 | 0 | 1 | 0 | 13,2 |
| 1,9 | —4,3 | —4,9 | —4,7 | 0 | 0 | 0 | 1 | —11,0 |
| 1 | —2,11111 | 0,38889 | —2,05556 | 0,55556 | 0 | 0 | 0 | —2,22223 |
| | 3,57778 | —2,87222 | —1,36111 | —0,38885 | 1 | 0 | 0 | —0,04440 |
| | 23,51110 | —1,13890 | 10,10559 | —4,05551 | 0 | 1 | 0 | 29,42228 |
| | —0,28889 | —5,63889 | —0,79444 | —1,05554 | 0 | 0 | 1 | —6,77776 |
| | 1 | —0,80279 | —0,38043 | —0,10868 | 0,27950 | 0 | 0 | —0,01241 |
| | | 17,73557 | 19,04992 | —1,50032 | —6,57135 | 1 | 0 | 29,71405 |
| | | —5,87081 | —0,90434 | —1,08694 | 0,08074 | 0 | 1 | —6,78134 |
| | | 1 | 1,07411 | —0,08459 | —0,37108 | 0,05638 | 0 | 1,67539 |
| | | | 5,40155 | —1,58355 | —2,09780 | 0,33100 | 1 | 3,05456 |
| | | | 1 | —0,29316 | —0,38837 | 0,06128 | 0,18513 | 0,56540 |
| | | 1 | | 0,23030 | 0,04607 | —0,00944 | —0,19885 | 1,06809 |
| | 1 | | | —0,03533 | 0,16873 | 0,01573 | —0,08920 | 1,06013 |
| 1 | | | | —0,21121 | —0,46003 | 0,16284 | 0,26956 | 0,76266 |

На основании результатов таблицы 17 получаем:

$$A^{-1}=\begin{bmatrix} -0{,}21121 & -0{,}46003 & 0{,}16284 & 0{,}26956 \\ -0{,}03533 & 0{,}16873 & 0{,}01573 & -0{,}08920 \\ 0{,}23030 & 0{,}04607 & -0{,}00944 & -0{,}19885 \\ -0{,}29316 & -0{,}38837 & 0{,}06128 & 0{,}18513 \end{bmatrix}.$$

Для проверки составим произведение

$$AA^{-1}=\begin{bmatrix}1{,}8 & -3{,}8 & 0{,}7 & -3{,}7\\ 0{,}7 & 2{,}1 & -2{,}6 & -2{,}8\\ 7{,}3 & 8{,}1 & 1{,}7 & -4{,}9\\ 1{,}9 & -4{,}3 & -4{,}9 & -4{,}7\end{bmatrix}\begin{bmatrix}-0{,}21121 & -0{,}46003 & 0{,}16284 & 0{,}26956\\ -0{,}03533 & 0{,}16873 & 0{,}01573 & -0{,}08920\\ 0{,}23030 & 0{,}04607 & -0{,}00944 & -0{,}19885\\ -0{,}29316 & -0{,}38837 & 0{,}06128 & 0{,}18513\end{bmatrix}=$$

$$=\begin{bmatrix}0{,}99997 & 0{,}00000 & -0{,}00001 & 0{,}00000\\ -0{,}00025 & 0{,}99997 & -0{,}00002 & -0{,}00039\\ -0{,}00808 & -0{,}01017 & 0{,}99982 & 0{,}00009\\ 0{,}00000 & 0{,}00000 & 0{,}00000 & 1{,}00048\end{bmatrix}=$$

$$=E-10^{-3}\cdot\begin{bmatrix}0{,}03 & 0{,}00 & 0{,}01 & 0{,}00\\ 0{,}25 & 0{,}03 & 0{,}02 & 0{,}39\\ 8{,}08 & 10{,}17 & 0{,}18 & -0{,}09\\ 0{,}00 & 0{,}00 & 0{,}00 & -0{,}48\end{bmatrix}.$$

Мы видим, что благодаря округлению обратная матрица получилась не вполне точной. Ниже мы укажем (см. § 15) метод исправления элементов приближенной обратной матрицы.

## § 8. Метод квадратных корней

Пусть дана линейная система

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

где $A=[a_{ij}]$ — симметрическая матрица, т. е. $A'=[a_{ji}]=A$. Тогда матрицу $A$ можно представить в виде произведения двух транспонированных между собой треугольных матриц

$$A=T'T, \tag{2}$$

где

$$T=\begin{bmatrix}t_{11} & t_{12} & \dots & t_{1n}\\ 0 & t_{22} & \dots & t_{2n}\\ \cdot & \cdot & \cdot & \cdot\\ 0 & 0 & \dots & t_{nn}\end{bmatrix} \quad \text{и} \quad T'=\begin{bmatrix}t_{11} & 0 & \dots & 0\\ t_{12} & t_{22} & \dots & 0\\ \cdot & \cdot & \cdot & \cdot\\ t_{1n} & t_{2n} & \dots & t_{nn}\end{bmatrix}.$$

Производя перемножение матриц $T'$ и $T$, для определения элементов $t_{ij}$ матрицы $T$ получим следующие уравнения:

$$\left.\begin{aligned} t_{1i}t_{1j}+t_{2i}t_{2j}+\dots+t_{ii}t_{ij}&=a_{ij} \qquad (i<j),\\ t_{1i}^2+t_{2i}^2+\dots+t_{ii}^2&=a_{ii}.\end{aligned}\right\}$$

Отсюда последовательно находим:

$$\left.\begin{aligned}
&t_{11}=\sqrt{a_{11}}, \quad t_{1j}=\frac{a_{1j}}{t_{11}} \qquad (j>1),\\
&t_{ii}=\sqrt{a_{ii}-\sum_{k=1}^{i-1} t_{ki}^2} \qquad (1<i\leqslant n),\\
&t_{ij}=\frac{a_{ij}-\sum_{k=1}^{i-1} t_{ki}t_{kj}}{t_{ii}} \qquad (i<j),\\
&t_{ij}=0 \quad \text{при} \quad i>j.
\end{aligned}\right\} \tag{3}$$

Система (1) имеет определенное единственное решение, если $t_{ii}\neq 0$ $(i=1, 2, \ldots, n)$, так как тогда

$$\det A=\det T'\cdot\det T=(\det T)^2=(t_{11}t_{22}\ldots t_{nn})^2\neq 0.$$

Коэффициенты матрицы $T$ будут действительны, если $t_{ii}^2>0$. В дальнейшем мы, вообще говоря, не будем предполагать это последнее условие выполненным.

При наличии соотношения (2) уравнение (1) эквивалентно двум уравнениям:

$$T'y=b \quad \text{и} \quad Tx=y,$$

или в раскрытом виде

$$\left.\begin{aligned}
&t_{11}y_1=b_1,\\
&t_{12}y_1+t_{22}y_2=b_2,\\
&\ldots\ldots\ldots\ldots\\
&t_{1n}y_1+t_{2n}y_2+\ldots+t_{nn}y_n=b_n
\end{aligned}\right\} \tag{4}$$

и

$$\left.\begin{aligned}
&t_{11}x_1+t_{12}x_2+\ldots+t_{1n}x_n=y_1,\\
&\qquad\quad t_{22}x_2+\ldots+t_{2n}x_n=y_2,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\\
&\qquad\qquad\qquad\qquad t_{nn}x_n=y_n.
\end{aligned}\right\} \tag{5}$$

Отсюда последовательно находим:

$$\left.\begin{aligned}
&y_1=\frac{b_1}{t_{11}},\\
&y_i=\frac{b_i-\sum_{k=1}^{i-1} t_{ki}y_k}{t_{ii}} \qquad (i>1)
\end{aligned}\right\} \tag{6}$$

и

$$\left.\begin{aligned} x_n &= \frac{y_n}{t_{nn}}, \\ x_i &= \frac{y_i - \sum\limits_{k=i+1}^{n} t_{ik}x_k}{t_{ii}} \qquad (i < n). \end{aligned}\right\} \tag{7}$$

Изложенный способ решения линейной системы носит название *метода квадратных корней*. Так как матрица $A$ — симметрическая, а матрица $T$ — верхняя треугольная, то в вычислительной схеме можно записывать только $\frac{n}{2}(n+1)$ верхних коэффициентов $a_{ij}$ и $t_{ij}$ $(i \geqslant j)$. При вычислениях применяется обычный контроль с помощью сумм, причем при составлении суммы учитываются все коэффициенты соответствующей строки.

Заметим, что если для некоторой $s$-й строки имеем $t_{ss}^2 < 0$, то соответствующие элементы $t_{sj}$ будут мнимыми. Метод формально применим и в этом случае.

При практическом применении метода квадратных корней *прямым ходом* с помощью формул (3) и (6) последовательно вычисляются коэффициенты $t_{ij}$ и $y_i$ $(i = 1, 2, \ldots, n)$, а затем *обратным ходом* с помощью формулы (7) находятся неизвестные $x_i$ $(i = n, n-1, \ldots, 1)$.

Пример. Методом квадратных корней решить систему уравнений

$$\left.\begin{aligned} x_1 + 3x_2 - 2x_3 \qquad\quad - 2x_5 &= 0{,}5; \\ 3x_1 + 4x_2 - 5x_3 + x_4 - 3x_5 &= 5{,}4; \\ -2x_1 - 5x_2 + 3x_3 - 2x_4 + 2x_5 &= 5{,}0; \\ x_2 - 2x_3 + 5x_4 + 3x_5 &= 7{,}5; \\ -2x_1 - 3x_2 + 2x_3 + 3x_4 + 4x_5 &= 3{,}3. \end{aligned}\right\}$$

Решение. Записываем коэффициенты $a_{ij}$ и свободные члены $b_i$ данной системы в начальный раздел $A$ таблицы (таблица 18) и подсчитываем столбец $\sum$. Применяя формулы (3) и (6), последовательно переходя от строки к строке, вычисляем коэффициенты $t_{ij}$ и новые свободные члены $y_i$ и, таким образом, заполняем раздел $Б$ таблицы.

Например,

$$t_{35} = \frac{a_{35} - t_{13}t_{15} - t_{23}t_{25}}{t_{33}} = \frac{2-(-2)(-2)-(-0{,}4472i)(-1{,}3416i)}{0{,}8944i} = 1{,}5653i.$$

Для контроля подсчитываем столбец $\sum$. На основании формул (7) находим значения неизвестных $x_i$ и контрольные величины

Таблица 18

**Решение линейной системы методом квадратных корней**

| $a_{i1}$ | $a_{i2}$ | $a_{i3}$ | $a_{i4}$ | $a_{i5}$ | $b_i$ | $\Sigma$ | Разделы схемы |
|---|---|---|---|---|---|---|---|
| 1 | 3 | −2 | 0 | −2 | 0,5 | 0,5 | |
| 3 | 4 | −5 | 1 | −3 | 5,4 | 5,4 | |
| −2 | −5 | 3 | −2 | 2 | 5,0 | 1,0 | *А* |
| 0 | 1 | −2 | 5 | 3 | 7,5 | 14,5 | |
| −2 | −3 | 2 | 3 | 4 | 3,3 | 7,3 | |
| $t_{i1}$ | $t_{i2}$ | $t_{i3}$ | $t_{i4}$ | $t_{i5}$ | $y_i$ | $\Sigma$ | |
| 1 | 3 | −2 | 0 | −2 | 0,5 | 0,5 | |
| | 2,2361$i$ | −0,4472$i$ | −0,4472$i$ | −1,3416$i$ | −1,7471$i$ | −1,7471$i$ | |
| | | 0,8944$i$ | 2,0125$i$ | 1,5653$i$ | −7,5803$i$ | −3,1081$i$ | *Б* |
| | | | 3,0414 | 2,2194 | −2,2928 | 2,9679 | |
| | | | | 0,8221$i$ | 0,1643$i$ | 0,9859$i$ | |
| −6,0978 | −2,2016 | −6,8011 | −0,8996 | 0,1998 | | $\frac{x_i}{\bar{x}_i}$ | *В* |
| −5,0973 | −1,2017 | −5,8004 | 0,1007 | 1,1992 | | | |

$\bar{x}_i = x_i + 1$, помещая их в разделе *В*. Например,

$$x_3 = \frac{y_3 - t_{35}x_5 - t_{34}x_4}{t_{33}} = \frac{-7{,}5803i - 1{,}5652i \cdot 0{,}1998 - 2{,}0125i \cdot (-0{,}8996)}{0{,}8944i} = -6{,}8011.$$

## § 9. Схема Халецкого

Для удобства рассуждений систему линейных уравнений запишем в матричном виде

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

где $A = [a_{ij}]$ — квадратная матрица порядка $n$ и

$$x = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}, \qquad b = \begin{bmatrix} a_{1,\,n+1} \\ \cdot\ \cdot\ \cdot \\ a_{n,\,n+1} \end{bmatrix}$$

— векторы-столбцы. Представим матрицу $A$ в виде произведения нижней треугольной матрицы $B = [b_{ij}]$ и верхней треугольной матрицы $C = [c_{ij}]$ с единичной диагональю, т. е.

$$A = BC, \tag{2}$$

где

$$B=\begin{bmatrix} b_{11} & 0 & \dots & 0 \\ b_{21} & b_{22} & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ b_{n1} & b_{n2} & \dots & b_{nn} \end{bmatrix} \quad \text{и} \quad C=\begin{bmatrix} 1 & c_{12} & \dots & c_{1n} \\ 0 & 1 & \dots & c_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & 1 \end{bmatrix}.$$

Тогда элементы $b_{ij}$ и $c_{ij}$ определяются по формулам

$$\left.\begin{aligned} b_{i1} &= a_{i1}, \\ b_{ij} &= a_{ij} - \sum_{k=1}^{j-1} b_{ik}c_{kj} \qquad (i \geqslant j > 1) \end{aligned}\right\} \tag{3}$$

и

$$\left.\begin{aligned} c_{1j} &= \frac{a_{1j}}{b_{11}}, \\ c_{ij} &= \frac{1}{b_{ii}}\left(a_{ij} - \sum_{k=1}^{i-1} b_{ik}c_{kj}\right) \qquad (1 < i < j). \end{aligned}\right\} \tag{4}$$

Отсюда искомый вектор $\boldsymbol{x}$ может быть вычислен из цепи уравнений

$$B\boldsymbol{y} = \boldsymbol{b}, \qquad C\boldsymbol{x} = \boldsymbol{y}. \tag{5}$$

Так как матрицы $B$ и $C$ — треугольные, то системы (5) легко решаются, а именно:

$$\left.\begin{aligned} y_1 &= \frac{a_{1,\,n+1}}{b_{11}}, \\ y_i &= \frac{1}{b_{ii}}\left(a_{i,\,n+1} - \sum_{k=1}^{i-1} b_{ik}y_k\right) \qquad (i > 1) \end{aligned}\right\} \tag{6}$$

и

$$\left.\begin{aligned} x_n &= y_n, \\ x_i &= y_i - \sum_{k=i+1}^{n} c_{ik}x_k \qquad (i < n). \end{aligned}\right\} \tag{7}$$

Из формул (6) видно, что числа $y_i$ выгодно вычислять вместе с коэффициентами $c_{ij}$. Этот метод получил название *схемы Халецкого*. В схеме применяется обычный контроль с помощью сумм.

Заметим, что если матрица $A$ — симметрическая, т. е. $a_{ij} = a_{ji}$, то

$$c_{ij} = \frac{b_{ji}}{b_{ii}} \qquad (i < j).$$

Схема Халецкого удобна для работы на вычислительных машинах, так как в этом случае операции «накопления» (3) и (4) можно проводить без записи промежуточных результатов.

Пример. Решить систему

$$\left.\begin{aligned} 3x_1+x_2-x_3+2x_4&=6;\\ -5x_1+x_2+3x_3-4x_4&=-12;\\ 2x_1+x_3-x_4&=1;\\ x_1-5x_2+3x_3-3x_4&=3. \end{aligned}\right\}$$

Решение (см. таблицу 19).

В первый раздел таблицы 19 впишем матрицу коэффициентов системы, ее свободные члены и контрольные суммы.

Далее, так как $b_{i1}=a_{i1}$ $(i=1, 2, 3, 4)$, то первый столбец из раздела I переносится в первый столбец раздела II.

Чтобы получить первую строку раздела II, делим все элементы первой строки раздела I на элемент $a_{11}=b_{11}$, в нашем случае на 3. Имеем:

$$c_{12}=\frac{1}{3}=0,(3);$$

$$c_{13}=-\frac{1}{3}=-0,(3);$$

$$c_{14}=\frac{2}{3}=0,(6);$$

$$c_{15}=\frac{6}{3}=2;$$

$$c_{16}=\frac{11}{3}=2,(6).$$

Переходим к заполнению второго столбца раздела II, начиная со второй строки. Пользуясь формулами (3), определяем $b_{j2}$:

$$b_{22}=a_{22}-b_{21}c_{12}=1-\left(-5\cdot\frac{1}{3}\right)=\frac{8}{3}=2,66\,(6);$$

$$b_{32}=a_{32}-b_{31}c_{12}=0-2\cdot\frac{1}{3}=-\frac{2}{3}=0,(6);$$

$$b_{42}=a_{42}-b_{41}c_{12}=-5-1\cdot\frac{1}{3}=-5\frac{1}{3}=-5,(3).$$

Далее, определяя $c_{2j}$ $(j=3, 4, 5, 6)$ по формулам (4), заполняем вторую строку раздела II:

$$c_{23}=\frac{1}{b_{22}}(a_{23}-b_{21}c_{13})=\frac{3}{8}\left[3-(-5)\cdot\left(-\frac{1}{3}\right)\right]=\frac{1}{2};$$

$$c_{24}=\frac{1}{b_{22}}(a_{24}-b_{21}c_{14})=\frac{3}{8}\left[(-4)-(-5)\cdot\frac{2}{3}\right]=-\frac{1}{4};$$

$$c_{25}=\frac{1}{b_{22}}(a_{25}-b_{21}c_{15})=\frac{3}{8}\left[(-12)-(-5)\cdot 2\right]=-\frac{3}{4};$$

$$c_{26}=\frac{1}{b_{22}}(a_{26}-b_{21}c_{16})=\frac{3}{8}\left[(-17)-(-5)\cdot\frac{11}{3}\right]=\frac{1}{2}.$$

Таблица 19

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | | $\Sigma$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ | | $\Sigma$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| I | $a_{11}$ | $a_{12}$ | $a_{13}$ | $a_{14}$ | $a_{15}$ | $a_{16}$ | 3 | 1 | —1 | 2 | 6 | 11 |
| | $a_{21}$ | $a_{22}$ | $a_{23}$ | $a_{24}$ | $a_{25}$ | $a_{26}$ | —5 | 1 | 3 | —4 | —12 | —17 |
| | $a_{31}$ | $a_{32}$ | $a_{33}$ | $a_{34}$ | $a_{35}$ | $a_{36}$ | 2 | 0 | 1 | —1 | 1 | 3 |
| | $a_{41}$ | $a_{42}$ | $a_{43}$ | $a_{44}$ | $a_{45}$ | $a_{46}$ | 1 | —5 | 3 | —3 | 3 | —1 |
| II | $b_{11}$ \| 1 | $c_{12}$ | $c_{13}$ | $c_{14}$ | $c_{15}$ | $c_{16}$ | 3 \| 1 | 0,333333 | —0,333333 | 0,666667 | 2 | 3,666667 |
| | $b_{21}$ | $b_{22}$ \| 1 | $c_{23}$ | $c_{24}$ | $c_{25}$ | $c_{26}$ | —5 | 2,666667 \| 1 | 0,5 | —0,25 | —0,75 | 0,5 |
| | $b_{31}$ | $b_{32}$ | $b_{33}$ \| 1 | $c_{34}$ | $c_{35}$ | $c_{36}$ | 2 | —0,666667 | 2 \| 1 | —1,25 | —1,75 | —2 |
| | $b_{41}$ | $b_{42}$ | $b_{43}$ | $b_{44}$ \| 1 | $c_{45}$ | $c_{46}$ | 1 | —5,333333 | 6 | 2,5 \| 1 | 3 | 4 |
| III | | | | | $y_1$ | $x_1$ | | | | | 2 | 1 |
| | | | | | $y_2$ | $x_2$ | | | | | —0,75 | —1 |
| | | | | | $y_3$ | $x_3$ | | | | | —1,75 | 2 |
| | | | | | $y_4$ | $x_4$ | | | | | 3 | 3 |

Затем переходим к третьему столбцу, вычисляя его элементы $b_{33}$ и $b_{34}$ по формулам (3) и т. д., пока не будет заполнена вся таблица раздела II. Таким образом, заполнение раздела II происходит способом «елочки»: столбец — строка, столбец — строка и т. д.

В разделе III, пользуясь формулами (6) и (7), определяем $y_i$ и $x_i$ $(i=1, 2, 3, 4)$.

Текущий контроль осуществляется с помощью столбца $\sum$, над которым производятся те же действия, что и над столбцом свободных членов.

## § 10. Метод итерации

При большом числе неизвестных линейной системы схема метода Гаусса, дающая точное решение, становится весьма сложной. В этих условиях для нахождения корней системы иногда удобнее пользоваться приближенными численными методами. Изложим здесь один из этих методов — *метод итерации*.

Пусть дана линейная система

$$\left.\begin{array}{l} a_{11}x_1+a_{12}x_2+\ \ldots\ +a_{1n}x_n=b_1, \\ a_{21}x_1+a_{22}x_2+\ \ldots\ +a_{2n}x_n=b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1+a_{n2}x_2+\ \ldots\ +a_{nn}x_n=b_n. \end{array}\right\} \qquad (1)$$

Введя в рассмотрение матрицы

$$\boldsymbol{A}=\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}, \qquad \boldsymbol{x}=\begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}, \qquad \boldsymbol{b}=\begin{bmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{bmatrix},$$

систему (1) коротко можно записать в виде матричного уравнения

$$A\boldsymbol{x}=\boldsymbol{b}. \qquad (1')$$

Предполагая, что диагональные коэффициенты

$$a_{ii}\neq 0 \qquad (i=1,\ 2,\ \ldots,\ n),$$

разрешим первое уравнение системы (1) относительно $x_1$, второе — относительно $x_2$ и т. д. Тогда получим эквивалентную систему

$$\left.\begin{array}{l} x_1=\beta_1+\alpha_{12}x_2+\alpha_{13}x_3+\ \ldots\ +\alpha_{1n}x_n, \\ x_2=\beta_2+\alpha_{21}x_1+\alpha_{23}x_3+\ \ldots\ +\alpha_{2n}x_n, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n=\beta_n+\alpha_{n1}x_1+\alpha_{n2}x_2+\ \ldots\ +\alpha_{n,\,n-1}x_{n-1}, \end{array}\right\} \qquad (2)$$

где

$$\beta_i = \frac{b_i}{a_{ii}}; \qquad \alpha_{ij} = -\frac{a_{ij}}{a_{ii}} \qquad \text{при } i \neq j$$

и $\alpha_{ij} = 0$ при $i = j$ $(i, j = 1, 2, \ldots, n)$.

Введя матрицы

$$\alpha = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \alpha_{n1} & \alpha_{n2} & \ldots & \alpha_{nn} \end{bmatrix} \quad \text{и} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \beta_n \end{bmatrix},$$

систему (2) можем записать в матричной форме

$$x = \beta + \alpha x. \tag{2'}$$

Систему (2) будем решать методом последовательных приближений. За нулевое приближение принимаем, например, столбец свободных членов $x^{(0)} = \beta$.

Далее, последовательно строим матрицы-столбцы

$$x^{(1)} = \beta + \alpha x^{(0)}$$

*(первое приближение)*,

$$x^{(2)} = \beta + \alpha x^{(1)}$$

*(второе приближение)* и т. д.

Вообще говоря, любое $(k+1)$-е приближение вычисляют по формуле

$$x^{(k+1)} = \beta + \alpha x^{(k)} \qquad (k = 0, 1, 2, \ldots). \tag{3}$$

Если последовательность приближений $x^{(0)}, x^{(1)}, \ldots, x^{(k)}, \ldots$ имеет предел

$$x = \lim_{k \to \infty} x^{(k)},$$

то этот предел является решением системы (2). В самом деле, переходя к пределу в равенстве (3), будем иметь:

$$\lim_{k \to \infty} x^{(k+1)} = \beta + \alpha \lim_{k \to \infty} x^{(k)}$$

или

$$x = \beta + \alpha x,$$

т. е. предельный вектор $x$ является решением системы (2′), а следовательно, и системы (1).

Напишем формулы приближений в развернутом виде:

$$\left.\begin{aligned} & x_i^{(0)} = \beta_i, \\ & x_i^{(k+1)} = \beta_i + \sum_{j=1}^{n} \alpha_{ij} x_j^{(k)} \\ & (\alpha_{ii} = 0;\ i = 1, \ldots, n;\ k = 0, 1, 2, \ldots). \end{aligned}\right\} \tag{3'}$$

Заметим, что иногда выгоднее приводить систему (1) к виду (2) так, чтобы коэффициенты $\alpha_{ii}$ не были равны нулю. Например, уравнение

$$1{,}02x_1 - 0{,}15x_2 = 2{,}7$$

для применения метода последовательных приближений естественно записать в виде

$$x_1 = 2{,}7 - 0{,}02x_1 + 0{,}15x_2.$$

Вообще, имея систему

$$\sum_{j=1}^{n} a_{ij}x_j = b_i \qquad (i = 1,\ 2,\ \ldots,\ n),$$

можно положить:

$$a_{ii} = a_{ii}^{(1)} + a_{ii}^{(2)},$$

где $a_{ii}^{(1)} \neq 0$. Тогда данная система эквивалентна приведенной системе

$$x_i = \beta_i + \sum_{j=1}^{n} \alpha_{ij}x_j \qquad (i = 1,\ 2,\ \ldots,\ n),$$

где

$$\beta_i = \frac{b_i}{a_{ii}^{(1)}}, \qquad \alpha_{ii} = -\frac{a_{ii}^{(2)}}{a_{ii}^{(1)}}, \qquad \alpha_{ij} = -\frac{a_{ij}}{a_{ii}^{(1)}} \quad \text{при } i \neq j.$$

Поэтому при дальнейших рассуждениях мы не будем, вообще говоря, предполагать, что $\alpha_{ii} = 0$.

Метод последовательных приближений, определяемых формулой (3) или (3′), носит название *метода итерации*. Процесс итерации (3) хорошо сходится, т. е. число приближений, необходимых для получения корней системы (1) с заданной точностью, невелико, если элементы матрицы $\alpha$ малы по абсолютной величине. Иными словами, для успешного применения процесса итерации модули диагональных коэффициентов системы (1) должны быть велики по сравнению с модулями недиагональных коэффициентов этой системы (свободные члены при этом роли не играют).

Пример 1. Решить систему

$$\left.\begin{aligned} 4x_1 + 0{,}24x_2 - 0{,}08x_3 &= 8, \\ 0{,}09x_1 + 3x_2 - 0{,}15x_3 &= 9, \\ 0{,}04x_1 - 0{,}08x_2 + 4x_3 &= 20 \end{aligned}\right\} \qquad (4)$$

методом итерации.

Решение. Здесь диагональные коэффициенты 4; 3; 4 системы значительно преобладают над остальными коэффициентами при неизвестных. Приведем эту систему к нормальному виду (2)

$$\left.\begin{aligned} x_1 &= 2 - 0{,}06x_2 + 0{,}02x_3, \\ x_2 &= 3 - 0{,}03x_1 + 0{,}05x_3, \\ x_3 &= 5 - 0{,}01x_1 + 0{,}02x_2. \end{aligned}\right\} \qquad (5)$$

В матричной форме систему (5) можно записать так:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 5 \end{bmatrix} + \begin{bmatrix} 0 & -0{,}06 & 0{,}02 \\ -0{,}03 & 0 & 0{,}05 \\ -0{,}01 & 0{,}02 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

За нулевые приближения корней системы (4) принимаем:

$$x_1^{(0)} = 2; \quad x_2^{(0)} = 3; \quad x_3^{(0)} = 5.$$

Подставляя эти значения в правые части уравнений (5), получим первые приближения корней:

$$x_1^{(1)} = 2 - 0{,}06 \cdot 3 + 0{,}02 \cdot 5 = 1{,}92;$$
$$x_2^{(1)} = 3 - 0{,}03 \cdot 2 + 0{,}05 \cdot 5 = 3{,}19;$$
$$x_3^{(1)} = 5 - 0{,}01 \cdot 2 + 0{,}02 \cdot 3 = 5{,}04.$$

Далее, подставляя эти найденные приближения в формулу (5) получим вторые приближения корней:

$$x_1^{(2)} = 1{,}9094; \quad x_2^{(2)} = 3{,}1944;$$
$$x_3^{(2)} = 5{,}0446.$$

После новой подстановки будем иметь третьи приближения корней:

$$x_1^{(3)} = 1{,}90923; \quad x_2^{(3)} = 3{,}19495;$$
$$x_3^{(3)} = 5{,}04485 \text{ и т. д.}$$

Результаты вычисления помещены в таблице 20.

Таблица 20

**Вычисление решения линейной системы методом итерации**

| $k$ | $x_1^{(k)}$ | $x_2^{(k)}$ | $x_3^{(k)}$ |
|---|---|---|---|
| 0 | 2 | 2 | 5 |
| 1 | 1,92 | 3,19 | 5,04 |
| 2 | 1,9094 | 3,1944 | 5,0446 |
| 3 | 1,90923 | 3,19495 | 5,04485 |

Замечание. При применении метода итерации (формула (3)) нет необходимости за нулевое приближение $x^{(0)}$ принимать столбец свободных членов. Как будет доказано ниже, сходимость процесса итерации зависит только от свойств матрицы $\alpha$, причем при выполнении известных условий, если этот процесс сходится при какомнибудь выборе исходного начального приближения, то он будет сходиться к тому же предельному вектору и при любом другом выборе этого начального приближения. Поэтому начальный вектор $x^{(0)}$ в процессе итерации может быть взят произвольным.

Целесообразно за компоненты начального вектора выбирать приближенные значения корней системы, находимые грубой прикидкой.

Сходящийся процесс итерации обладает важным свойством самоисправляемости, т. е. отдельная ошибка в вычислениях не отразится на окончательном результате, так как ошибочное приближение можно рассматривать как новый начальный вектор.

Отметим, что иногда бывает удобнее подсчитывать не сами приближения, а их разности. Введя обозначения

$$\Delta^{(k)} = x^{(k)} - x^{(k-1)} \qquad (k = 0,\ 1,\ 2,\ \ldots),$$

из формулы (3) имеем:

$$x^{(k+1)} = \beta + \alpha x^{(k)} \tag{6}$$

и

$$x^{(k)} = \beta + \alpha x^{(k-1)}. \tag{7}$$

Отсюда, вычитая из равенства (6) равенство (7), получим:

$$\Delta^{(k+1)} = \alpha\left(x^{(k)} - x^{(k-1)}\right) = \alpha\Delta^{(k)},$$

т. е.

$$\Delta^{(k+1)} = \alpha\Delta^{(k)} \qquad (k = 1,\ 2,\ \ldots). \tag{8}$$

За нулевое приближение принимаем:

$$\Delta^{(0)} = x^{(0)}, \tag{9}$$

тогда $m$-е приближение есть

$$x^{(m)} = \sum_{k=0}^{m} \Delta^{(k)}. \tag{10}$$

Если, как обычно, положить $\Delta^{(0)} = x^{(0)} = \beta$, то равенство (8) будет выполнено и при $k = 0$. В противном случае равенство (8) при $k = 0$ не имеет места. Отсюда вытекает такая методика вычисления этого варианта итерации:

1) если $\Delta^{(0)} = x^{(0)} = \beta$, то

$$\Delta^{(k)} = \alpha\Delta^{(k-1)} = \alpha^k\beta \qquad (k = 0,\ 1,\ 2,\ \ldots)$$

и

$$x^{(k)} = \sum_{s=0}^{k} \Delta^{(s)} = \sum_{s=0}^{k} \alpha^s\beta;$$

2) если же $\Delta^{(0)} = x^{(0)} \neq \beta$, то находим

$$\Delta^{(1)} = x^{(1)} - x^{(0)} = \alpha x^{(0)} + \beta - x^{(0)}$$

и полагаем

$$\Delta^{(k)} = \alpha\Delta^{(k-1)} = \alpha^{k-1}\Delta^{(1)} \qquad (k = 1,\ 2,\ 3,\ \ldots).$$

Следовательно,

$$x^{(k)} = \sum_{s=0}^{k} \Delta^{(s)} = x^{(0)} + \sum_{s=1}^{k} \alpha^{s-1}\Delta^{(1)}.$$

Пример 2. Решить систему

$$\left.\begin{aligned} 2x_1 - x_2 + x_3 &= -3, \\ 3x_1 + 5x_2 - 2x_3 &= 1, \\ x_1 - 4x_2 + 10x_3 &= 0. \end{aligned}\right\} \qquad (11)$$

Решение. Приведем систему (11) к виду (2):

$$\begin{aligned} x_1 &= -1{,}5 + 0{,}5x_2 - 0{,}5x_3; \\ x_2 &= 0{,}2 - 0{,}6x_1 + 0{,}4x_3; \\ x_3 &= -0{,}1x_1 + 0{,}4x_2. \end{aligned}$$

Здесь

$$\alpha = \begin{bmatrix} 0 & 0{,}5 & -0{,}5 \\ -0{,}6 & 0 & 0{,}4 \\ -0{,}1 & 0{,}4 & 0 \end{bmatrix}$$

и

$$\beta = \begin{bmatrix} -1{,}5 \\ 0{,}2 \\ 0 \end{bmatrix}.$$

Пользуясь формулами (8) и (9), получим:

$$\Delta^{(0)} = \beta = \begin{bmatrix} -1{,}5 \\ 0{,}2 \\ 0 \end{bmatrix};$$

$$\Delta^{(1)} = \alpha\Delta^{(0)} = \begin{bmatrix} 0 & 0{,}5 & -0{,}5 \\ -0{,}6 & 0 & 0{,}4 \\ -0{,}1 & 0{,}4 & 0 \end{bmatrix} \begin{bmatrix} -1{,}5 \\ 0{,}2 \\ 0 \end{bmatrix} = \begin{bmatrix} 0{,}1 \\ 0{,}9 \\ 0{,}23 \end{bmatrix};$$

$$\Delta^{(2)} = \alpha\Delta^{(1)} = \begin{bmatrix} 0 & 0{,}5 & -0{,}5 \\ -0{,}6 & 0 & 0{,}4 \\ -0{,}1 & 0{,}4 & 0 \end{bmatrix} \begin{bmatrix} 0{,}1 \\ 0{,}9 \\ 0{,}23 \end{bmatrix} = \begin{bmatrix} 0{,}335 \\ 0{,}032 \\ 0{,}350 \end{bmatrix}$$

и т. д. Результаты записываем в таблицу 21.

Таким образом, приближенные значения корней есть

$$x_1 = -1{,}235; \quad x_2 = 1{,}089; \quad x_3 = 0{,}560.$$

Недостатком этого варианта метода итерации является систематическое накопление ошибок при увеличении числа слагаемых, в результате чего могут возникнуть значительные погрешности искомых корней. Кроме того, ошибка, допущенная в вычислениях, повлияет на окончательный результат. Поэтому надежнее пользоваться первым вариантом метода итерации.

Таблица 21

**Вычисление решения линейной системы видоизмененным методом итерации**
(метод накопления)

| $k$ | $\Delta^{(k)}x_1$ | $\Delta^{(k)}x_2$ | $\Delta^{(k)}x_3$ |
|---|---|---|---|
| 0 | —1,500 | 0,200 | 0,000 |
| 1 | 0,100 | 0,900 | 0,230 |
| 2 | 0,335 | 0,032 | 0,350 |
| 3 | —0,159 | —0,061 | —0,021 |
| 4 | —0,020 | 0,011 | —0,008 |
| 5 | 0,010 | 0,009 | 0,006 |
| 6 | 0,002 | —0,004 | 0,003 |
| 7 | —0,004 | 0,000 | —0,001 |
| 8 | 0,000 | 0,002 | 0,000 |
| 9 | 0,001 | 0,000 | 0,001 |
| Σ | —1,235 | 1,089 | 0,560 |

**Замечания о точности расчета.** Если все коэффициенты и свободные члены данной системы являются точными числами, то решение ее методом последовательных приближений может быть получено с любым заранее заданным числом $m$ верных десятичных знаков. При этом в значениях последовательных приближений следует удерживать $m+1$ десятичных знаков и последовательные приближения вычислять до их совпадения, после чего нужно округлить результат на один знак. Если коэффициенты и свободные члены данной системы являются приближенными числами, написанными с $p$ знаками, то решение этой системы производится, как в случае точных чисел, с точностью до $m=p$ знаков.

Приведем без доказательства достаточное условие сходимости процесса итерации (доказательство см. гл. IX, § 1).

Теорема. *Если для приведенной системы* (2) *выполнено по меньшей мере одно из условий*

1) $$\sum_{j=1}^{n} |\alpha_{ij}| < 1 \qquad (i=1,\ 2,\ \ldots,\ n)$$

*или*

2) $$\sum_{i=1}^{n} |\alpha_{ij}| < 1 \qquad (j=1,\ 2,\ \ldots,\ n),$$

*то процесс итерации* (3) *сходится к единственному решению этой системы, независимо от выбора начального приближения.*

Следствие. Для системы

$$\sum_{j=1}^{n} a_{ij}x_j = b_i \qquad (i=1,\ 2,\ \ldots,\ n)$$

метод итерации сходится, если выполнены неравенства

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}| \quad (i = 1, 2, \ldots, n),$$

т. е. если модули диагональных коэффициентов для каждого уравнения системы больше суммы модулей всех остальных коэффициентов (не считая свободных членов).

## § 11. Приведение линейной системы к виду, удобному для итерации

Теорема сходимости (§ 10) накладывает жесткие условия на коэффициенты данной линейной системы

$$Ax = b. \tag{1}$$

Однако, если $\det A \neq 0$, то с помощью линейного комбинирования уравнений системы (15) последнюю всегда можно заменить эквивалентной системой

$$x = \beta + \alpha x, \tag{2}$$

такой, что условия теоремы сходимости будут выполнены.

В самом деле, умножим уравнение (1) на матрицу $D = A^{-1} - \varepsilon$, где $\varepsilon = [\varepsilon_{ij}]$ — матрица с малыми по модулю элементами. Тогда будем иметь:

$$(A^{-1} - \varepsilon) Ax = Db$$

или

$$x = \beta + \alpha x, \tag{3}$$

где $\alpha = \varepsilon A$ и $\beta = Db$. Если $|\varepsilon_{ij}|$ достаточно малы, то очевидно, что система (3) удовлетворяет условиям теоремы сходимости.

Умножение на матрицу $D$ эквивалентно совокупности элементарных преобразований над уравнениями системы. Задача заключается в том, чтобы прийти к стандартному виду (3) с наименьшей затратой труда.

Практически поступают следующим образом. Из заданной системы выделяют уравнения с коэффициентами, модули которых больше суммы модулей остальных коэффициентов уравнения. Каждое выделенное уравнение выписывают в такую строку новой системы, чтобы наибольший по модулю коэффициент оказался диагональным.

Из оставшихся неиспользованных и выделенных уравнений системы составляют линейно независимые между собой линейные комбинации с таким расчетом, чтобы был соблюден указанный выше принцип комплектования новой системы и все свободные строки оказались заполненными. При этом нужно позаботиться, чтобы каждое неиспользованное ранее уравнение попало хотя бы в одну линейную комбинацию, являющуюся уравнением новой системы. Поясним все сказанное на примере.

Пример. Систему

$$\left.\begin{array}{ll}
(A) & 2x_1+3x_2-4x_3+x_4-3=0,\\
(Б) & x_1-2x_2-5x_3+x_4-2=0,\\
(B) & 5x_1-3x_2+x_3-4x_4-1=0,\\
(Г) & 10x_1+2x_2-x_3+2x_4+4=0
\end{array}\right\}$$

привести к виду, годному для применения метода итерации.

Решение. В уравнении (Б) коэффициент при $x_3$ по модулю больше суммы модулей остальных коэффициентов, поэтому можно принять это уравнение за третье уравнение новой системы. Коэффициент при $x_1$ в уравнении (Г) также больше суммы модулей остальных коэффициентов уравнения (Г), поэтому можно принять это уравнение за первое уравнение новой системы. Таким образом, новая система имеет следующий вид:

$$\left.\begin{array}{ll}
(\mathrm{I}) & 10x_1+2x_2-x_3+2x_4+4=0,\\
(\mathrm{II}) & \dots\dots\dots\dots\dots\dots\\
(\mathrm{III}) & x_1-2x_2-5x_3+x_4-2=0,\\
(\mathrm{IV}) & \dots\dots\dots\dots\dots\dots
\end{array}\right\}$$

Анализируя данную систему, легко заметить, что для получения уравнения (II) с максимальным по модулю коэффициентом при $x_2$ достаточно составить разность (А)—(В):

$$(\mathrm{II})\qquad x_1+5x_2+x_3+0x_4-1=0.$$

Теперь в новую систему вошли уравнения (А), (Б) и (Г), поэтому в уравнение (IV) обязательно должно войти уравнение (В) данной системы. Подбором убеждаемся, что за уравнение (IV) можно взять линейную комбинацию 2 (А) — (Б) + 2 (В) — (Г), т. е.

$$(\mathrm{IV})\qquad 3x_1+0x_2+0x_3-9x_4-10=0.$$

В итоге получим преобразованную систему уравнений I—IV, эквивалентную исходной и удовлетворяющую условиям сходимости процесса итерации. Разрешив эту систему относительно диагональных неизвестных, будем иметь систему

$$\left.\begin{array}{l}
x_1=0x_1-0{,}2x_2+0{,}1x_3-0{,}2x_4-0{,}4;\\
x_2=0{,}2x_1+0x_2-0{,}2x_3+0x_4+0{,}2;\\
x_3=0{,}2x_1-0{,}4x_2+0x_3+0{,}2x_4-0{,}4;\\
x_4=0{,}333x_1+0x_2+0x_3+0x_4-1{,}111,
\end{array}\right\}$$

к которой можно применить метод итерации.

## § 12. Метод Зейделя

Метод Зейделя представляет собой некоторую модификацию метода итерации. Основная его идея заключается в том, что при вычислении $(k+1)$-го приближения неизвестной $x_i$ учитываются уже вычисленные ранее $(k+1)$-е приближения неизвестных $x_1, x_2, \ldots, x_{i-1}$.

Пусть дана приведенная линейная система

$$x_i = \beta_i + \sum_{j=1}^{n} \alpha_{ij} x_j \qquad (i = 1, 2, \ldots, n).$$

Выберем произвольно начальные приближения корней

$$x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)},$$

стараясь, конечно, чтобы они в какой-то мере соответствовали искомым неизвестным

$$x_1, x_2, \ldots, x_n.$$

Далее, предполагая, что $k$-е приближения $x_i^{(k)}$ корней известны, согласно Зейделю будем строить $(k+1)$-е приближения корней по следующим формулам:

$$x_1^{(k+1)} = \beta_1 + \sum_{j=1}^{n} \alpha_{1j} x_j^{(k)};$$

$$x_2^{(k+1)} = \beta_2 + \alpha_{21} x_1^{(k+1)} + \sum_{j=2}^{n} \alpha_{2j} x_j^{(k)};$$

$$\cdots\cdots\cdots\cdots\cdots\cdots$$

$$x_i^{(k+1)} = \beta_i + \sum_{j=1}^{i-1} \alpha_{ij} x_j^{(k+1)} + \sum_{j=i}^{n} \alpha_{ij} x_j^{(k)};$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$x_n^{(k+1)} = \beta_n + \sum_{j=1}^{n-1} \alpha_{nj} x_j^{(k+1)} + \alpha_{nn} x_n^{(k)} \qquad (k = 0, 1, 2, \ldots).$$

Заметим, что указанная выше (§ 10) теорема сходимости для простой итерации остается верной для итерации по методу Зейделя (см. гл. IX, §§ 3—7).

Обычно метод Зейделя дает лучшую сходимость, чем метод простой итерации, но, вообще говоря, он приводит к более громоздким вычислениям. Процесс Зейделя может сходиться даже в том случае, если расходится процесс итерации. Однако это бывает не всегда. Возможны случаи, когда процесс Зейделя сходится медленнее процесса итерации. Более того, могут быть случаи, когда процесс итерации сходится, а процесс Зейделя расходится [1] (см. гл. XI, § 6).

Пример. Методом Зейделя решить систему уравнений

$$\left.\begin{aligned} 10x_1+x_2+x_3&=12,\\ 2x_1+10x_2+x_3&=13,\\ 2x_1+2x_2+10x_3&=14. \end{aligned}\right\}$$

Решение. Приведем эту систему к виду, удобному для итерации,

$$\left.\begin{aligned} x_1&=1{,}2-0{,}1x_2-0{,}1x_3;\\ x_2&=1{,}3-0{,}2x_1-0{,}1x_3;\\ x_3&=1{,}4-0{,}2x_1-0{,}2x_2. \end{aligned}\right\}$$

В качестве нулевых приближений корней возьмем:

$$x_1^{(0)}=1{,}2;\quad x_2^{(0)}=0;\quad x_3^{(0)}=0.$$

Применяя процесс Зейделя, последовательно получим:

$$\left.\begin{aligned} x_1^{(1)}&=1{,}2-0{,}1\cdot 0-0{,}1\cdot 0=1{,}2;\\ x_2^{(1)}&=1{,}3-0{,}2\cdot 1{,}2-0{,}1\cdot 0=1{,}06;\\ x_3^{(1)}&=1{,}4-0{,}2\cdot 1{,}2-0{,}2\cdot 1{,}06=0{,}948; \end{aligned}\right\}\qquad \text{(I)}$$

$$\left.\begin{aligned} x_1^{(2)}&=1{,}2-0{,}1\cdot 1{,}06-0{,}1\cdot 0{,}948=0{,}9992;\\ x_2^{(2)}&=1{,}3-0{,}2\cdot 0{,}9992-0{,}1\cdot 0{,}948=1{,}00536;\\ x_3^{(2)}&=1{,}4-0{,}2\cdot 0{,}9992-0{,}2\cdot 1{,}00536=0{,}999098 \text{ и т. д.} \end{aligned}\right\}\qquad \text{(II)}$$

Результаты вычислений с точностью до четырех знаков помещены в таблице 22.

Таблица 22

**Нахождение корней линейной системы методом Зейделя**

| $k$ | $x_1^{(k)}$ | $x_2^{(k)}$ | $x_3^{(k)}$ |
|---|---|---|---|
| 0 | 1,2000 | 0,0000 | 0,0000 |
| 1 | 1,2000 | 1,0600 | 0,9480 |
| 2 | 0,9992 | 1,0054 | 0,9991 |
| 3 | 0,9996 | 1,0001 | 1,0001 |
| 4 | 1,0000 | 1,0000 | 1,0000 |
| 5 | 1,0000 | 1,0000 | 1,0000 |

Точные значения корней: $x_1=1$; $x_2=1$; $x_3=1$.

## § 13. Случай нормальной системы

Определение 1. Целый однородный полином второй степени от $n$ переменных называется *квадратичной формой* этих переменных. В общем случае квадратичная форма имеет вид

$$u(x_1, x_2, \ldots, x_n) = a_{11}x_1^2 + a_{22}x_2^2 + \ldots + a_{nn}x_n^2 + \\ + 2a_{12}x_1x_2 + 2a_{13}x_1x_3 + \ldots + 2a_{n-1,n}x_{n-1}x_n, \quad (1)$$

где $a_{ij}$ $(i, j = 1, 2, \ldots, n)$ — постоянные числа, причем для удобства соответствующие коэффициенты при $i \neq j$ взяты в четной форме $2a_{ij}$. Приравняв $u$ постоянной $c$, получим уравнение центральной поверхности второго порядка

$$u(x_1, x_2, \ldots, x_n) = c$$

в $n$-мерном пространстве.

Если положить

$$a_{ij} = a_{ji}, \quad (2)$$

т. е. $2a_{ij} = a_{ij} + a_{ji}$, то формулу (1) короче можно записать следующим образом:

$$u(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}x_ix_j. \quad (1')$$

Матрица

$$A = [a_{ij}] \quad (3)$$

носит название *матрицы квадратичной формы* (1′). В силу условия (2) матрица $A$ будет симметрической, т. е. совпадет со своей транспонированной матрицей. Наоборот, для всякой симметрической матрицы $A = [a_{ij}]$ можно построить соответствующую квадратичную форму (1′).

Определение 2. Квадратичная форма (1) называется *положительно (отрицательно) определенной*, если она принимает положительные (отрицательные) значения, обращаясь в нуль лишь при

$$x_1 = x_2 = \ldots = x_n = 0.$$

Если $u(x_1, x_2, \ldots, x_n)$ — положительно определенная квадратичная форма, то уравнение

$$u(x_1, x_2, \ldots, x_n) = c \qquad (c > 0)$$

представляет собой уравнение эллипсоида. Заметим, что в этом случае

$$a_{ii} > 0 \qquad (i = 1, 2, \ldots, n),$$

так как

$$a_{11} = u(1, 0, \ldots, 0) > 0,$$
$$a_{22} = u(0, 1, \ldots, 0) > 0,$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$a_{nn} = u(0, 0, \ldots, 1) > 0.$$

Определение 3. Назовем линейную систему

$$\sum_{j=1}^{n} a_{ij}x_j = b_i \qquad (i = 1,\ 2,\ \ldots,\ n) \tag{4}$$

*нормальной*, если: 1) матрица коэффициентов $A = [a_{ij}]$ — симметрическая, т. е. $a_{ij} = a_{ji}$, 2) соответствующая квадратичная форма

$$u = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}x_i x_j \text{ — положительно определенная.}$$

Нормальные системы встречаются при решении многих вопросов, например, в способе наименьших квадратов, при нахождении направлений главных осей эллипсоида и т. д.

Нормальную систему (4) приведем обычным способом к специальному виду

$$x_i = \sum_{j \neq i} \alpha_{ij}x_j + \beta_i, \tag{4'}$$

где

$$\alpha_{ij} = -\frac{a_{ij}}{a_{ii}} \qquad (j \neq i) \quad \text{и} \quad \beta_i = \frac{b_i}{a_{ii}}.$$

Теорема 1. *Если линейная система* (4) — *нормальная, то процесс Зейделя для эквивалентной ей приведенной системы* (4′) *всегда сходится.*

Доказательство см. главу XI, § 5, а также [2].

Способ приведения линейной системы к нормальному виду указывается следующей теоремой.

Теорема 2. *Если обе части линейной системы*

$$A\boldsymbol{x} = \boldsymbol{b} \tag{5}$$

*с неособенной матрицей* $A = [a_{ij}]$ *умножить слева на транспонированную матрицу* $A' = [a_{ji}]$, *то полученная новая система*

$$A'A\boldsymbol{x} = A'\boldsymbol{b} \tag{6}$$

*будет нормальной.*

Докажем сначала, что матрица $A'A$ есть симметрическая матрица. В самом деле, имеем:

$$(A'A)' = A'A'' = A'A.$$

Теперь докажем, что квадратичная форма, соответствующая матрице $A'A$, — положительно определенная. Составим квадратичную форму с матрицей $A'A$:

$$u(x_1,\ x_2,\ \ldots,\ x_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} a_{ki}a_{kj}x_i x_j.$$

Изменяя порядок суммирования, получим:

$$u=\sum_{k=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ki}x_ia_{kj}x_j=\sum_{k=1}^{n}\left(\sum_{i=1}^{n}a_{ki}x_i\cdot\sum_{j=1}^{n}a_{kj}x_j\right).$$

Так как значение суммы не зависит от обозначения индекса суммирования, то

$$u=\sum_{k=1}^{n}\left(\sum_{i=1}^{n}a_{ki}x_i\right)^2\geqslant 0.$$

Согласно предположению $\det A=\det[a_{ij}]\neq 0$. Поэтому однородная система

$$\sum_{i=1}^{n}a_{ki}x_i=0 \qquad (k=1,\ 2,\ \ldots,\ n)$$

имеет лишь нулевые решения. Следовательно,

$$u(x_1,\ x_2,\ \ldots,\ x_n)>0 \quad \text{при } |x_1|+|x_2|+\ldots+|x_n|\neq 0.$$

Теорема доказана.

## § 14. Метод релаксации

Пусть имеем систему линейных уравнений

$$\left.\begin{array}{l} a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n=b_1,\\ a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_n=b_2,\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ a_{n1}x_1+a_{n2}x_2+\ldots+a_{nn}x_n=b_n. \end{array}\right\} \qquad (1)$$

Преобразуем эту систему следующим образом: перенесем свободные члены налево и разделим первое уравнение на $-a_{11}$, второе на $-a_{22}$ и т. д. Тогда получим систему, *приготовленную к релаксации*,

$$\left.\begin{array}{l} -x_1+b_{12}x_2+\ldots+b_{1n}x_n+c_1=0,\\ b_{21}x_1-\quad x_2+\ldots+b_{2n}x_n+c_2=0,\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ b_{n1}x_1+b_{n2}x_2+\ldots-\quad x_n+c_n=0, \end{array}\right\} \qquad (2)$$

где

$$b_{ij}=-\frac{a_{ij}}{a_{ii}} \qquad (i\neq j) \quad \text{и} \quad c_i=\frac{b_i}{a_{ii}}.$$

Пусть $x^{(0)}=(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})$ — начальное приближение решения системы (2). Подставляя эти значения в систему (2), получим *невязки*

$$\left.\begin{aligned} R_1^{(0)} &= c_1 - x_1^{(0)} + \sum_{j=2}^{n} b_{1j}x_j^{(0)}, \\ R_2^{(0)} &= c_2 - x_2^{(0)} + \sum_{\substack{j=1\\ j\neq 2}}^{n}{}' b_{2j}x_j^{(0)}, \\ &\cdots\cdots\cdots\cdots\cdots \\ R_n^{(0)} &= c_n - x_n^{(0)} + \sum_{j=1}^{n-1} b_{nj}x_j^{(0)}. \end{aligned}\right\} \qquad (3)$$

Если одной из неизвестных $x_s^{(0)}$ дать приращение $\delta x_s^{(0)}$, то соответствующая невязка $R_s^{(0)}$ уменьшится на величину $\delta x_s^{(0)}$, а все остальные невязки $R_i^{(0)}$ $(i\neq s)$ увеличатся на величину $b_{is}\delta x_s^{(0)}$. Таким образом, чтобы обратить очередную невязку $R_s^{(1)}$ в нуль, достаточно величине $x_s^{(0)}$ дать приращение

$$\delta x_s^{(0)} = R_s^{(0)},$$

и мы будем иметь:

$$R_s^{(1)} = 0$$

и

$$R_i^{(1)} = R_i^{(0)} + b_{is}\delta x_s^{(0)} \quad \text{при } i\neq s.$$

*Метод релаксации* (по-русски: *метод ослабления*) [3], [4] в его простейшей форме заключается в том, что на каждом шаге обращают в нуль максимальную по модулю невязку путем изменения значения соответствующей компоненты приближения. Процесс заканчивается, когда все невязки последней преобразованной системы будут равны нулю с заданной точностью. Вопрос о сходимости этого процесса мы оставляем без рассмотрения [4].

Пример. Методом релаксации решить систему [3]

$$\left.\begin{aligned} 10x_1 - 2x_2 - 2x_3 &= 6, \\ -x_1 + 10x_2 - 2x_3 &= 7, \\ -x_1 - x_2 + 10x_3 &= 8, \end{aligned}\right\} \qquad (4)$$

производя вычисления с двумя десятичными знаками.

Решение. Приводим систему (4) к виду, удобному для релаксации

$$\left.\begin{aligned} -x_1 + 0{,}2x_2 + 0{,}2x_3 + 0{,}6 &= 0, \\ -x_2 + 0{,}1x_1 + 0{,}2x_3 + 0{,}7 &= 0, \\ -x_3 + 0{,}1x_1 + 0{,}1x_2 + 0{,}8 &= 0. \end{aligned}\right\}$$

Выбирая в качестве начальных приближений корней нулевые значения

$$x_1^{(0)} = x_2^{(0)} = x_3^{(0)} = 0,$$

находим соответствующие им невязки

$$R_1^{(0)} = 0{,}60; \quad R_2^{(0)} = 0{,}70; \quad R_3^{(0)} = 0{,}80.$$

Согласно общей теории полагаем:

$$\delta x_3^{(0)} = 0{,}80.$$

Отсюда получаем невязки

$$R_1^{(1)} = R_1^{(0)} + 0{,}2 \cdot 0{,}8 = 0{,}60 + 0{,}16 = 0{,}76;$$
$$R_2^{(1)} = R_2^{(0)} + 0{,}2 \cdot 0{,}8 = 0{,}70 + 0{,}16 = 0{,}86;$$
$$R_3^{(1)} = R_1^{(0)} - R_2^{(0)} = 0.$$

Таблица 23

**Решение линейной системы методом релаксации**

| | $x_1$ | $R_1$ | $x_2$ | $R_2$ | $x_3$ | $R_3$ |
|---|---|---|---|---|---|---|
| | 0 | 0,60 | 0 | 0,70 | 0 | 0,80 |
| | | 0,16 | | 0,16 | 0,80 | −0,80 |
| | | 0,76 | | 0,86 | | 0 |
| | | 0,17 | 0,86 | −0,86 | | 0,09 |
| | | 0,93 | | 0 | | 0,09 |
| | 0,93 | −0,93 | | 0,09 | | 0,09 |
| | | 0 | | 0,09 | | 0,18 |
| | | 0,04 | | 0,04 | 0,18 | −0,18 |
| | | 0,04 | | 0,13 | | 0 |
| | | 0,03 | 0,13 | −0,13 | | 0,01 |
| | | 0,07 | | 0 | | 0,01 |
| | 0,07 | −0,07 | | 0,01 | | 0,01 |
| | | 0 | | 0,01 | | 0,02 |
| | | 0 | | 0 | 0,02 | −0,02 |
| | | 0 | | 0,01 | | 0 |
| | | 0 | 0,01 | −0,01 | | 0 |
| | | 0 | | 0 | | 0 |
| $\Sigma$ | 1,00 | | 1,00 | | 1,00 | |

Далее, полагаем

$$\delta x_2^{(1)} = 0{,}86$$

и т. д. Соответствующие результаты вычислений приведены в таблице 23.

Суммируя все приращения $\delta_i^{(k)}$ $(i = 1, 2, 3;\ k = 0, 1, \ldots)$, получим значения корней

$$x_1 = 0 + 0{,}93 + 0{,}07 = 1{,}00;$$
$$x_2 = 0 + 0{,}86 + 0{,}13 + 0{,}01 = 1{,}00;$$
$$x_3 = 0 + 0{,}80 + 0{,}18 + 0{,}02 = 1{,}00.$$

Для контроля подставляем найденные значения корней в исходные уравнения; в данном случае система (4) решена точно.

## § 15. Исправление элементов приближенной обратной матрицы

Пусть имеем неособенную матрицу $A$ и требуется найти обратную матрицу $A^{-1}$. Положим, что мы получили приближенное значение обратной матрицы $D_0 \approx A^{-1}$. Тогда для улучшения точности можно воспользоваться методом последовательных приближений в специальной форме. В качестве предварительной меры погрешности используем разность

$$F_0 = E - AD_0.$$

Если $F_0 = 0$, то очевидно, что $D_0 = A^{-1}$, поэтому, если модули элементов матрицы $F_0$ малы, то матрицы $A^{-1}$ и $D_0$ близки между собой. Будем строить последовательные приближения по формуле

$$D_k = D_{k-1} + D_{k-1}F_{k-1} \qquad (k = 1, 2, 3, \ldots), \tag{1}$$

причем соответствующая погрешность есть

$$F_k = E - AD_k.$$

Оценим быстроту сходимости последовательных приближений. Имеем:

$$F_1 = E - AD_1 = E - A(D_0 + D_0F_0) = E - AD_0(E + F_0) =$$
$$= E - (E - F_0)(E + F_0) = E - (E - F_0^2) = F_0^2.$$

Аналогично

$$F_2 = F_1^2 = F_0^4$$

и, вообще,

$$F_k = F_0^{2^k} \qquad (k = 1, 2, 3, \ldots). \tag{2}$$

Докажем, что если

$$\|F_0\| \leqslant q < 1, \tag{3}$$

где $\|F_0\|$ — какая-нибудь каноническая норма матрицы $F_0$ (гл. VII, § 7), то процесс итерации (1) сходится, т. е.

$$\lim_{k\to\infty} D_k = A^{-1}.$$

Действительно, из формулы (2) имеем:

$$\|F_k\| \leqslant \|F_0\|^{2^k} \leqslant q^{2^k}.$$

Поэтому

$$\lim_{k\to\infty} \|F_k\| = 0$$

и, следовательно,

$$\lim_{k\to\infty} F_k = \lim_{k\to\infty} (E - AD_k) = 0$$

или

$$E - A \lim_{k\to\infty} D_k = 0,$$

т. е.

$$\lim_{k\to\infty} D_k = A^{-1}E = A^{-1}.$$

Таким образом, утверждение доказано.

В частности, используя $m$-норму (гл. VII, § 7), получаем, что если элементы матрицы $F_0 = [f_{ij}]$ удовлетворяют неравенству

$$|f_{ij}| \leqslant \frac{q}{n},$$

где $n$ — порядок матрицы и $0 \leqslant q < 1$, то процесс итерации (1) заведомо сходится.

Предполагая неравенство (3) выполненным, оценим погрешность

$$R_k = \|A^{-1} - D_k\| \leqslant \|A^{-1}\| \|E - AD_k\| = \|A^{-1}\| \|F_k\| \leqslant \|A^{-1}\| q^{2^k}.$$

Так как

$$AD_0 = E - F_0,$$

то

$$A^{-1} = D_0 (E - F_0)^{-1} = D_0 (E + F_0 + F_0^2 + \ldots).$$

Отсюда

$$\|A^{-1}\| \leqslant \|D_0\| \{\|E\| + q + q^2 + \ldots\} = \|D_0\| \left\{\|E\| + \frac{q}{1-q}\right\}.$$

Для $m$-нормы или $l$-нормы имеем $\|E\| = 1$, и поэтому

$$\|A^{-1}\| < \frac{\|D_0\|}{1-q}.$$

Таким образом,

$$\|A^{-1}-D_k\| \leqslant \frac{\|D_0\|}{1-q}\|F_k\| \tag{4}$$

или

$$\|A^{-1}-D_k\| \leqslant \frac{\|D_0\|}{1-q} q^{2^k}, \tag{5}$$

где норма понимается в смысле $m$-нормы или $l$-нормы. Из формулы (4) следует, что сходимость процесса (1) при $q \ll 1$ очень быстрая.

На практике процесс уточнения элементов обратной матрицы прекращают, когда обеспечено неравенство

$$\|D_k - D_{k-1}\| \leqslant \varepsilon,$$

где $\varepsilon$ — заданная точность.

Пример. Исправить элементы приближенной обратной матрицы, полученной в примере § 7 на стр. 286.

Решение. Для матрицы

$$A = \begin{bmatrix} 1,8 & -3,8 & 0,7 & -3,7 \\ 0,7 & 2,1 & -2,6 & -2,8 \\ 7,3 & 8,1 & 1,7 & -4,9 \\ 1,9 & -4,3 & -4,9 & -4,7 \end{bmatrix}$$

методом Гаусса получена приближенная обратная матрица

$$D_0 = \begin{bmatrix} -0,21121 & -0,46003 & 0,16284 & 0,26956 \\ -0,03533 & 0,16873 & 0,01573 & -0,08920 \\ 0,23030 & 0,04607 & -0,00944 & -0,19885 \\ -0,29316 & -0,38837 & -0,06128 & 0,18513 \end{bmatrix},$$

причем

$$AD_0 = E - 10^{-3} \cdot \begin{bmatrix} 0,03 & 0,00 & 0,01 & 0,00 \\ 0,25 & 0,03 & 0,02 & 0,39 \\ 8,08 & 10,17 & 0,18 & -0,09 \\ 0,00 & 0,00 & 0,00 & -0,48 \end{bmatrix}.$$

Отсюда

$$F_0 = E - AD_0 = 10^{-3} \cdot \begin{bmatrix} 0,03 & 0,00 & 0,01 & 0,00 \\ 0,25 & 0,03 & 0,02 & 0,39 \\ 8,08 & 10,17 & 0,18 & -0,09 \\ 0,00 & 0,00 & 0,00 & -0,48 \end{bmatrix}.$$

Для дальнейшего уточнения элементов матрицы $D_0$ воспользуемся итерационным процессом

$$D_{k+1} = D_k + D_k F_k, \quad F_k = E - AD_k \quad (k = 0, 1, 2, \ldots).$$

Так как

$$q = \|F_0\|_l = 10^{-3}\cdot(0{,}03 + 10{,}17) = 1{,}02\cdot 10^{-2} \ll 1,$$

то процесс итерации быстро сходится.

Имеем:

$$D_0F_0 = \begin{bmatrix} -0{,}21121 & -0{,}46003 & 0{,}16284 & 0{,}26956 \\ -0{,}03533 & 0{,}16873 & 0{,}01573 & -0{,}08920 \\ 0{,}23030 & 0{,}04607 & -0{,}00944 & -0{,}19885 \\ -0{,}29316 & -0{,}38837 & -0{,}06128 & 0{,}18513 \end{bmatrix} \times$$

$$\times 10^{-3}\cdot\begin{bmatrix} 0{,}03 & 0{,}00 & 0{,}01 & 0{,}00 \\ 0{,}25 & 0{,}03 & 0{,}02 & 0{,}39 \\ 8{,}08 & 10{,}17 & 0{,}18 & -0{,}09 \\ 0{,}00 & 0{,}00 & 0{,}00 & -0{,}48 \end{bmatrix} =$$

$$= 10^{-3}\cdot\begin{bmatrix} 1{,}19 & 1{,}64 & 0{,}02 & -0{,}32 \\ 0{,}17 & 0{,}16 & 0{,}01 & 0{,}11 \\ -0{,}06 & -0{,}09 & 0{,}00 & 0{,}11 \\ 0{,}39 & 0{,}61 & 0{,}00 & -0{,}24 \end{bmatrix}.$$

Отсюда

$$D_1 = D_0 + D_0F_0 = \begin{bmatrix} -0{,}21121 & -0{,}46003 & 0{,}16284 & 0{,}26956 \\ -0{,}03533 & 0{,}16873 & 0{,}01573 & -0{,}08920 \\ 0{,}23030 & 0{,}04607 & -0{,}00944 & -0{,}19885 \\ -0{,}29316 & -0{,}38837 & 0{,}06128 & 0{,}18513 \end{bmatrix} +$$

$$+ 10^{-3}\cdot\begin{bmatrix} 1{,}19 & 1{,}64 & 0{,}02 & -0{,}32 \\ 0{,}17 & 0{,}16 & 0{,}01 & 0{,}11 \\ -0{,}06 & -0{,}09 & 0{,}00 & 0{,}11 \\ 0{,}39 & 0{,}61 & 0{,}00 & -0{,}24 \end{bmatrix} =$$

$$= \begin{bmatrix} -0{,}21002 & -0{,}45839 & 0{,}16286 & 0{,}26924 \\ -0{,}03516 & 0{,}16889 & 0{,}01574 & -0{,}08909 \\ 0{,}23024 & 0{,}04598 & -0{,}00944 & -0{,}19874 \\ -0{,}29277 & -0{,}38776 & 0{,}06128 & 0{,}18489 \end{bmatrix}.$$

Можно считать

$$A^{-1} \approx D_1,$$

так как

$$AD_1 = E - 10^{-5}\cdot\begin{bmatrix} 2 & -2 & 1 & 3 \\ 0 & 2 & -1 & 0 \\ 3 & 4 & -5 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

и

$$F_1 = E - AD_1 = 10^{-5} \cdot \begin{bmatrix} 2 & -2 & 1 & 3 \\ 0 & 2 & -1 & 0 \\ 3 & 4 & -5 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}.$$

На основании формулы (4) для погрешности имеем оценку

$$\|A^{-1} - D_1\|_l \leqslant \frac{\|D_0\|_l}{1-q} \|F_1\|_l.$$

Так как

$$\|D_0\|_l = 0{,}46003 + 0{,}16873 + 0{,}04607 + 0{,}38837 < 1{,}07$$

и

$$\|F_1\|_l = 10^{-5} \cdot (2+2+4) = 8 \cdot 10^{-5},$$

то окончательно имеем:

$$\|A^{-1} - D_1\|_l \leqslant \frac{1{,}07}{1 - 1{,}02 \cdot 10^{-2}} \cdot 8 \cdot 10^{-5} < 9 \cdot 10^{-5}.$$

Замечание. Подбор приближенной обратной матрицы может быть осуществлен различными способами. В частности, используется способ обращения матриц, указанный в главе VII, § 12.

В заключение главы отметим следующее. В настоящее время разработаны многие другие методы решения системы линейных алгебраических уравнений (Метод Перселла, эскалаторный метод [6], метод Ричардсона [7] и др.)

### Литература к восьмой главе

1. В. Н. Фаддеева, Вычислительные методы линейной алгебры, Гостехиздат, 1950, гл. II.
2. Дж. Скарборо, Численные методы математического анализа, ГТТИ, 1934, дополнение I.
3. М. Дж. Сальвадори, Численные методы в технике, ИЛ, М., 1955, гл. 1, § 10.
4. Современная математика для инженеров, под ред. Э. Ф. Беккенбаха, ИЛ, М., 1958, гл. 15.
5. Х. Л. Смолицкий, Вычислительная математика (конспект лекций), ЛКВВИА им. Можайского, Л., 1960.
6. Д. К. Фаддеев и В. Н. Фаддеева, Вычислительные методы линейной алгебры, Физматгиз, 1960, гл. II.
7. И. С. Березин и Н. П. Жидков, Методы вычислений, Физматгиз, 1959, гл. VI.

ГЛАВА IX*

# СХОДИМОСТЬ ИТЕРАЦИОННЫХ ПРОЦЕССОВ ДЛЯ СИСТЕМ ЛИНЕЙНЫХ УРАВНЕНИЙ

## § 1. Достаточные условия сходимости процесса итерации

Пусть мы имеем приведенную линейную систему

$$x = \alpha x + \beta, \tag{1}$$

$$\alpha = [\alpha_{ij}], \qquad \beta = \begin{bmatrix} \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_n \end{bmatrix}$$

— заданные матрица и вектор и $x = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}$ — искомый вектор.

Теорема. *Процесс итерации для приведенной линейной системы* (1) *сходится к единственному ее решению, если какая-нибудь каноническая норма матрицы* $\alpha$ *меньше единицы, т. е. для итерационного процесса*

$$x^{(k)} = \beta + \alpha x^{(k-1)} \qquad (k = 1, 2, \ldots)$$

($x^{(0)}$ *произвольно*) *достаточное условие сходимости есть*

$$\|\alpha\| < 1. \tag{2}$$

Доказательство. Отправляясь от произвольного вектора $x^{(0)}$, строим последовательность приближений

$$x^{(1)} = \beta + \alpha x^{(0)},$$
$$x^{(2)} = \beta + \alpha x^{(1)},$$
$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$
$$x^{(k)} = \beta + \alpha x^{(k-1)}.$$

Отсюда

$$x^{(k)} = (E + \alpha + \alpha^2 + \ldots + \alpha^{k-1})\,\beta + \alpha^k x^{(0)}. \tag{3}$$

Так как при $\|\alpha\| < 1$ имеем $\|\alpha^k\| \to 0$ при $k \to \infty$, то (см. гл. VII, § 10)

$$\lim_{k\to\infty} \alpha^k = 0$$

и

$$\lim_{k\to\infty} (E+\alpha+\alpha^2+\ldots+\alpha^{k-1}) = \sum_{k=0}^{\infty} \alpha^k = (E-\alpha)^{-1}.$$

Поэтому, переходя к пределу при $k \to \infty$ в равенстве (3), получим:

$$x = \lim_{k\to\infty} x^{(k)} = (E-\alpha)^{-1}\beta. \tag{4}$$

Этим доказана сходимость итеративного процесса. Кроме того, из равенства (4) имеем:

$$(E-\alpha)\,x = \beta$$

или

$$x = \alpha x + \beta,$$

т. е. предельный вектор $x$ является решением системы (1). Так как матрица системы (1) $E-\alpha$ — неособенная, то решение $x$ единственно.

Следствие 1. Процесс итерации для системы (1) сходится, если:

а) $$\|\alpha\|_m = \max_i \sum_{j=1}^{n} |\alpha_{ij}| < 1;$$

или

б) $$\|\alpha\|_l = \max_j \sum_{i=1}^{n} |\alpha_{ij}| < 1;$$

или

в) $$\|\alpha\|_k = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} |\alpha_{ij}|^2} < 1.$$

В частности, процесс итерации заведомо сходится, если элементы матрицы $\alpha$ удовлетворяют неравенству

$$|\alpha_{ij}| < \frac{1}{n},$$

где $n$ — число неизвестных в системе (1).

Действительно, а), б) и в) являются простейшими каноническими нормами матрицы $\alpha$.

Следствие 2. Для системы

$$\sum_{j=1}^{n} a_{ij}x_j = b_i \qquad (i = 1,\ 2,\ \ldots,\ n) \tag{5}$$

процесс итерации сходится, если выполнены неравенства:

а′) $$|a_{ii}| > \sum_{j=1}^{n}{}' |a_{ij}| \qquad (i=1,\ 2,\ \ldots,\ n)$$

или

б′) $$|a_{jj}| > \sum_{i=1}^{n}{}' |a_{ij}| \qquad (j=1,\ 2,\ \ldots,\ n),$$

где штрих у знака суммы означает, что при суммировании пропускаются значения $i=j$, т. е. сходимость имеет место, если модули диагональных элементов матрицы $A=[a_{ij}]$ системы (1) или для каждой строки превышают сумму модулей недиагональных элементов этой строки, или же для каждого столбца превышают сумму модулей недиагональных элементов этого столбца.

В самом деле, при наличии неравенства а′), очевидно, выполнено соответствующее неравенство а) следствия 1.

Для доказательства второго утверждения в системе (5) положим:

$$x_i = \frac{z_i}{a_{ii}} \qquad (i=1,\ 2,\ \ldots,\ n),$$

где $z_i$ — новые неизвестные. Тогда получим систему

$$\sum_{j=1}^{n} \frac{a_{ij}}{a_{jj}} z_j = b_i \qquad (i=1,\ 2,\ \ldots,\ n), \tag{5}$$

для которой процесс итерации сходится или расходится одновременно с процессом итерации для исходной системы (5). Приведя обычным способом систему (5′) к специальному виду (1) и используя условие б) следствия 1, получим достаточное условие сходимости процесса итерации для системы (5):

$$\sum_{i=1}^{n}{}' \left|\frac{a_{ij}}{a_{jj}}\right| < 1 \qquad (j=1,\ 2,\ \ldots,\ n)$$

или

$$|a_{jj}| > \sum_{i=1}^{n}{}' |a_{ij}| \qquad (j=1,\ 2,\ \ldots,\ n).$$

## § 2. Оценка погрешности приближений процесса итерации

Пусть $x^{(k-1)}$ и $x^{(k)}$ $(k \geqslant 1)$ — два последовательных приближения решения линейной системы $x=\alpha x+\beta$. При $p \geqslant 1$ имеем:

$$\|x^{(k+p)} - x^{(k)}\| \leqslant \|x^{(k+1)} - x^{(k)}\| + \\ + \|x^{(k+2)} - x^{(k+1)}\| + \ldots + \|x^{(k+p)} - x^{(k+p-1)}\|. \tag{1}$$

Так как

$$x^{(m+1)} = \alpha x^{(m)} + \beta$$

и

$$x^{(m)} = \alpha x^{(m-1)} + \beta,$$

то

$$x^{(m+1)} - x^{(m)} = \alpha (x^{(m)} - x^{(m-1)})$$

и, следовательно,

$$\| x^{(m+1)} - x^{(m)} \| \leqslant \| \alpha \| \, \| x^{(m)} - x^{(m-1)} \| \leqslant$$
$$\leqslant \| \alpha \|^{m-k} \| x^{(k+1)} - x^{(k)} \| \quad \text{при} \quad m > k \geqslant 1.$$

Поэтому из формулы (1) получаем:

$$\| x^{(p+k)} - x^{(k)} \| \leqslant \| x^{(k+1)} - x^{(k)} \| +$$
$$+ \| \alpha \| \, \| x^{(k+1)} - x^{(k)} \| + \ldots + \| \alpha \|^{p-1} \| x^{(k+1)} - x^{(k)} \| \leqslant$$
$$\leqslant \frac{1}{1 - \| \alpha \|} \| x^{(k+1)} - x^{(k)} \|.$$

Переходя в последнем неравенстве к пределу при $p \longrightarrow \infty$, получим окончательно:

$$\| x - x^{(k)} \| \leqslant \frac{\| x^{(k+1)} - x^{(k)} \|}{1 - \| \alpha \|} \tag{2}$$

при $k \geqslant 1$, или

$$\| x - x^{(k)} \| \leqslant \frac{\| \alpha \|}{1 - \| \alpha \|} \| x^{(k)} - x^{(k-1)} \|.$$

Если

$$\| \alpha \| \leqslant \frac{1}{2},$$

то предыдущая формула принимает вид

$$\| x - x^{(k)} \| \leqslant \| x^{(k)} - x^{(k-1)} \|,$$

т. е. в этом случае, если

$$\| x^{(k)} - x^{(k-1)} \| < \varepsilon,$$

то и

$$\| x - x^{(k)} \| < \varepsilon.$$

В общем случае, если в процессе вычислений будет обнаружено, что

$$\| x^{(k)} - x^{(k-1)} \| \leqslant \frac{1-q}{q} \varepsilon,$$

где $q = \| \alpha \| < 1$, то

$$\| x - x^{(k)} \| \leqslant \varepsilon$$

и, таким образом,

$$| x_i - x_i^{(k)} | \leqslant \varepsilon \quad (i = 1, 2, \ldots, n).$$

При этом предполагается, конечно, что последовательные приближения $x^{(j)}$ ($j=0, 1, \ldots, k$) вычисляются точно, т. е. в них полностью отсутствуют погрешности округлений.

Из формулы (2), используя полученные выше оценки для нормы разности двух последовательных приближений, будем иметь:

$$\| x - x^{(k)} \| \leqslant \frac{\| \alpha \|^k}{1 - \| \alpha \|} \| x^{(1)} - x^{(0)} \|.$$

В частности, если выбрать

$$x^{(0)} = \beta,$$

то

$$x^{(1)} = \alpha\beta + \beta$$

и

$$\| x^{(1)} - x^{(0)} \| = \| \alpha\beta \| \leqslant \| \alpha \| \| \beta \|.$$

Следовательно,

$$\| x - x^{(k)} \| \leqslant \frac{\| \alpha \|^{k+1}}{1 - \| \alpha \|} \| \beta \|. \tag{2'}$$

Пример. Показать, что для системы

$$\left.\begin{aligned} 10x_1 - x_2 + 2x_3 - 3x_4 &= 0, \\ x_1 + 10x_2 - x_3 + 2x_4 &= 5, \\ 2x_1 + 3x_2 + 20x_3 - x_4 &= -10, \\ 3x_1 + 2x_2 + x_3 + 20x_4 &= 15 \end{aligned}\right\} \tag{3}$$

процесс итерации сходится. Сколько итераций следует выполнить, чтобы найти корни системы (3) с точностью до $10^{-4}$?

Решение. Приводя систему (3) к специальному виду, получим:

$$\left.\begin{aligned} x_1 &= 0{,}1x_2 - 0{,}2x_3 + 0{,}3x_4; \\ x_2 &= -0{,}1x_1 + 0{,}1x_3 - 0{,}2x_4 + 0{,}5; \\ x_3 &= -0{,}1x_1 - 0{,}15x_2 + 0{,}05x_4 - 0{,}5; \\ x_4 &= -0{,}15x_1 - 0{,}1x_2 - 0{,}05x_3 + 0{,}75. \end{aligned}\right\} \tag{3'}$$

Отсюда матрица системы

$$\alpha = \begin{bmatrix} 0 & 0{,}1 & -0{,}2 & 0{,}3 \\ -0{,}1 & 0 & 0{,}1 & -0{,}2 \\ -0{,}1 & -0{,}15 & 0 & 0{,}05 \\ -0{,}15 & -0{,}1 & -0{,}05 & 0 \end{bmatrix}.$$

Используя, например, норму $\| \alpha \|_l$, получим:

$$\| \alpha \|_l = \max(0{,}35;\ 0{,}35;\ 0{,}35;\ 0{,}55) = 0{,}55 < 1.$$

Следовательно, процесс итерации для системы (3′) сходится.

За начальное приближение корня $x$ примем:

$$x^{(0)} = \beta = \begin{bmatrix} 0, \\ 0,5 \\ -0,5 \\ 0,75 \end{bmatrix}.$$

Отсюда

$$\|\beta\|_l = 0 + 0,5 + 0,5 + 0,75 = 1,75.$$

Пусть $k$ — число итераций, необходимое для достижения заданной точности. Применяя формулу (2′), будем иметь:

$$\|x - x^{(k)}\| \leqslant \frac{\|\alpha\|_l^{k+1} \|\beta\|_l}{1 - \|\alpha\|_l} = \frac{0,55^{k+1} \cdot 1,75}{0,45} < 10^{-4}.$$

Отсюда

$$0,55^{k+1} < \frac{45}{175} \cdot 10^{-4}$$

и

$$(k+1) \lg 0,55 < \lg 45 - \lg 175 - 4,$$

т. е.

$$-(k+1) \cdot 0,25964 < 1,65321 - 2,24304 - 4 = -4,58983.$$

Следовательно,

$$k + 1 > \frac{4,58983}{0,25964} \approx 17,7$$

и

$$k > 16,7.$$

Можно принять $k = 17$.

Следует отметить, что теоретическая оценка числа итераций, необходимых для обеспечения заданной точности, практически оказывается весьма завышенной.

## § 3. Первое достаточное условие сходимости процесса Зейделя

Теорема. *Если для линейной системы*

$$x = \alpha x + \beta \tag{1}$$

*выполнено условие*

$$\|\alpha\|_m < 1, \tag{2}$$

*где*

$$\|\alpha\|_m = \max_i \sum_{j=1}^{n} |\alpha_{ij}|,$$

*то процесс Зейделя для системы* (1) *сходится к единственному ее решению при любом выборе начального вектора* $x^{(0)}$.

Доказательство. Пусть $x^{(k)}=\{x_1^{(k)}, \ldots, x_n^{(k)}\}$ — $k$-е приближение процесса Зейделя. Имеем:

$$x_i^{(k)}=\sum_{j=1}^{i-1}\alpha_{ij}x_j^{(k)}+\sum_{j=i}^{n}\alpha_{ij}x_j^{(k-1)}+\beta_i \tag{3}$$
$$(i=1, 2, \ldots, n;\ k=1, 2, \ldots).$$

При выполнении условия (2) система (1) допускает единственное решение $x=\{x_1, \ldots, x_n\}$, которое может быть найдено, например, методом итерации. Имеем:

$$x_i=\sum_{j=1}^{n}\alpha_{ij}x_j+\beta_i \qquad (i=1, 2, \ldots). \tag{4}$$

Вычитая из равенства (4) равенство (3), получим:

$$x_i-x_i^{(k)}=\sum_{j=1}^{i-1}\alpha_{ij}\left(x_j-x_j^{(k)}\right)+\sum_{j=i}^{n}\alpha_{ij}\left(x_j-x_j^{(k-1)}\right);$$

отсюда

$$\left|x_i-x_i^{(k)}\right|\leqslant\sum_{j=1}^{i-1}|\alpha_{ij}|\left|x_j-x_j^{(k)}\right|+\sum_{j=i}^{n}|\alpha_{ij}|\left|x_j-x_j^{(k-1)}\right| \tag{5}$$
$$(i=1, 2, \ldots, n).$$

Согласно смыслу принятой нормы

$$\|x-x^{(k)}\|_m=\max_i\left|x_i-x_i^{(k)}\right|,$$

поэтому

$$\left|x_j-x_j^{(k)}\right|\leqslant\|x-x^{(k)}\|_m$$

($j=1, 2, \ldots, n$). Следовательно, из неравенства (5) выводим:

$$\left|x_i-x_i^{(k)}\right|\leqslant p_i\|x-x^{(k)}\|_m+q_i\|x-x^{(k-1)}\|_m, \tag{6}$$

где

$$p_i=\sum_{j=1}^{i-1}|\alpha_{ij}| \quad \text{и} \quad q_i=\sum_{j=i}^{n}|\alpha_{ij}|.$$

Пусть $s=s(k)$ есть то значение индекса $i$, для которого

$$\left|x_s-x_s^{(k)}\right|=\max_i\left|x_i-x_i^{(k)}\right|=\|x-x^{(k)}\|_m.$$

Полагая $i=s$ в неравенстве (6), получим:

$$\|x-x^{(k)}\|_m\leqslant p_s\|x-x^{(k)}\|_m+q_s\|x-x^{(k-1)}\|_m$$

или

$$\|x-x^{(k)}\|_m\leqslant\frac{q_s}{1-p_s}\|x-x^{(k-1)}\|_m.$$

Отсюда

$$\|x-x^{(k)}\|_m\leqslant\mu\|x-x^{(k-1)}\|_m, \tag{7}$$

где

$$\mu = \max_i \frac{q_i}{1-p_i}. \tag{8}$$

Покажем, что

$$\mu \leqslant \|\alpha\|_m < 1.$$

Действительно, так как

$$p_i + q_i = \sum_{j=1}^{n} |\alpha_{ij}| \leqslant \|\alpha\|_m < 1,$$

то

$$q_i \leqslant \|\alpha\|_m - p_i$$

и, следовательно,

$$\frac{q_i}{1-p_i} \leqslant \frac{\|\alpha\|_m - p_i}{1-p_i} \leqslant \frac{\|\alpha\|_m - p_i \|\alpha\|_m}{1-p_i} = \|\alpha\|_m.$$

Поэтому

$$\mu = \|\alpha\|_m < 1.$$

Из неравенства (7) вытекает, что

$$\|x - x^{(k)}\|_m \leqslant \mu^k \|x - x^{(0)}\|_m$$

и, следовательно,

$$\lim_{k\to\infty} x^{(k)} = x,$$

тем самым сходимость процесса Зейделя к искомому решению доказана.

Замечание. Так как для метода итерации мы имеем

$$\|x - x^{(k)}\| \leqslant \|\alpha\|_m \|x - x^{(k-1)}\|,$$

а для метода Зейделя получим

$$\|x - x^{(k)}\| \leqslant \mu \|x - x^{(k-1)}\|,$$

где $\mu \leqslant \|\alpha\|_m$, то в условиях теоремы 1 сходимость процесса Зейделя в общем несколько лучше, чем сходимость процесса простой итерации. Из формулы (8) следует, что в этом случае, при применении метода Зейделя, систему (1) выгодно располагать так, чтобы первое уравнение системы имело наименьшую сумму модулей коэффициентов

$$q_1 = \sum_{j=1}^{n} |\alpha_{1j}|.$$

## § 4. Оценка погрешности приближений процесса Зейделя по $m$-норме

Пусть $x^{(k)}$ и $x^{(k+1)}$ — две последовательные итерации процесса Зейделя. Применяя к этим итерациям преобразования, использованные при доказательстве теоремы § 3, получим неравенство, аналогичное неравенству (7) из § 3:

$$\|x^{(k+1)} - x^{(k)}\|_m \leqslant \mu \|x^{(k)} - x^{(k-1)}\|_m.$$

Отсюда

$$\|x^{(k+p)}-x^{(k)}\|_m \leqslant \|x^{(k+p)}-x^{(k+p-1)}\|_m + \\ + \|x^{(k+p-1)}-x^{(k+p-2)}\|_m + \ldots + \|x^{(k+1)}-x^{(k)}\|_m \leqslant \\ \leqslant \mu^p \|x^{(k)}-x^{(k-1)}\|_m + \mu^{p-1}\|x^{(k)}-x^{(k-1)}\|_m + \ldots \\ \ldots + \mu \|x^{(k)}-x^{(k-1)}\|_m \leqslant \frac{\mu}{1-\mu}\|x^{(k)}-x^{(k-1)}\|_m.$$

При $p \to \infty$ будем иметь:

$$\lim_{p\to\infty} x^{(k+p)} = x$$

и, следовательно,

$$\|x-x^{(k)}\|_m \leqslant \frac{\mu}{1-\mu}\|x^{(k)}-x^{(k-1)}\|_m,$$

где

$$\mu = \max_i \frac{\sum\limits_{j=i}^{n}|\alpha_{ij}|}{1-\sum\limits_{j=1}^{i-1}|\alpha_{ij}|} \leqslant \|\alpha\|_m.$$

В частности, из полученного неравенства выводим:

$$\|x-x^{(k)}\|_m \leqslant \frac{\mu^k}{1-\mu}\|x^{(1)}-x^{(0)}\|_m,$$

т. е.

$$|x_i-x_i^{(k)}| \leqslant \frac{\mu^k}{1-\mu}\max_j |x_j^{(1)}-x_j^{(0)}| \qquad (i=1,\ 2,\ \ldots,\ n).$$

## § 5. Второе достаточное условие сходимости процесса Зейделя

Теорема. *Если для линейной системы*

$$x = \alpha x + \beta \tag{1}$$

*выполнено условие*

$$\|\alpha\|_l < 1,$$

*где*

$$\|\alpha\|_l = \max_j \sum_{i=1}^{n} |\alpha_{ij}|,$$

*то процесс Зейделя сходится к единственному решению системы* (1) *при любом выборе начального вектора.*

Доказательство. Пусть

$$x_i^{(k)} = \sum_{j=1}^{i-1}\alpha_{ij}x_j^{(k)} + \sum_{j=i}^{n}\alpha_{ij}x_j^{(k-1)} + \beta_i \quad (i=1,\ 2,\ldots,\ n;\ k=1,\ 2,\ldots). \tag{2}$$

Для точного решения $x=\{x_1, x_2, \ldots, x_n\}$, которое существует и единственно, имеем:

$$x_i=\sum_{j=1}^{i-1}\alpha_{ij}x_j+\sum_{j=i}^{n}\alpha_{ij}x_j+\beta_i. \tag{3}$$

Вычитая из равенств (3) соответствующие равенства (2), получим:

$$x_i-x_i^{(k)}=\sum_{j=1}^{i-1}\alpha_{ij}\left(x_j-x_j^{(k)}\right)+\sum_{j=i}^{n}\alpha_{ij}\left(x_j-x_j^{(k-1)}\right).$$

Отсюда

$$\left|x_i-x_i^{(k)}\right|\leqslant\sum_{j=1}^{i-1}|\alpha_{ij}|\left|x_j-x_j^{(k)}\right|+\sum_{j=i}^{n}|\alpha_{ij}|\left|x_j-x_j^{(k-1)}\right|$$
$$(i=1, 2, \ldots, n).$$

Просуммировав последние неравенства, будем иметь:

$$\sum_{i=1}^{n}\left|x_i-x_i^{(k)}\right|\leqslant\sum_{i=1}^{n}\sum_{j=1}^{i-1}|\alpha_{ij}|\left|x_j-x_j^{(k)}\right|+\sum_{i=1}^{n}\sum_{j=i}^{n}|\alpha_{ij}|\left|x_j-x_j^{(k-1)}\right|,$$

или, меняя порядок суммирования, получим:

$$\sum_{i=1}^{n}\left|x_i-x_i^{(k)}\right|\leqslant\sum_{j=1}^{n-1}\left|x_j-x_j^{(k)}\right|\sum_{i=j+1}^{n}|\alpha_{ij}|+\sum_{j=1}^{n}\left|x_j-x_j^{(k-1)}\right|\sum_{i=1}^{j}|\alpha_{ij}|. \tag{4}$$

Положим

$$s_j=\sum_{i=j+1}^{n}|\alpha_{ij}|, \quad t_j=\sum_{i=1}^{j}|\alpha_{ij}| \quad (j=1, 2, \ldots, n-1)$$

и

$$s_n=0, \quad t_n=\sum_{i=1}^{n}|\alpha_{ij}|.$$

Очевидно,

$$s_j+t_j=\sum_{i=1}^{n}|\alpha_{ij}|\leqslant\|\alpha\|_l<1;$$

отсюда

$$s_j<1.$$

Неравенство (4) принимает вид

$$\sum_{i=1}^{n}\left|x_i-x_i^{(k)}\right|\leqslant\sum_{j=1}^{n}s_j\left|x_j-x_j^{(k)}\right|+\sum_{j=1}^{n}t_j\left|x_j-x_j^{(k-1)}\right|$$

или

$$\sum_{j=1}^{n}(1-s_j)\left|x_j-x_j^{(k)}\right|\leqslant\sum_{j=1}^{n}t_j\left|x_j-x_j^{(k-1)}\right|.$$

Так как

$$t_j \leqslant \|\alpha\|_l - s_j \leqslant \|\alpha\|_l - s_j\|\alpha\|_l = \|\alpha\|_l(1-s_j), \tag{5}$$

то далее имеем:

$$\sum_{j=1}^{n}(1-s_j)\left|x_j - x_j^{(k)}\right| \leqslant \|\alpha\|_l \sum_{i=1}^{n}(1-s_j)\left|x_j - x_j^{(k-1)}\right| \leqslant$$

$$\leqslant \|\alpha\|_l^k \sum_{i=1}^{n}(1-s_j)\left|x_j - x_j^{(0)}\right|. \tag{6}$$

Отсюда, переходя к пределу при $k \to \infty$ и учитывая, что $\|\alpha\|_l < 1$, получим:

$$\lim_{k\to\infty}\sum_{j=1}^{n}(1-s_j)\left|x_j - x_j^{(k)}\right| = 0.$$

Следовательно,

$$\lim_{k\to\infty} x_j^{(k)} = x_j \quad (j=1,\ 2,\ \ldots,\ n),$$

что и требовалось доказать.

## § 6. Оценка погрешности приближений процесса Зейделя по $l$-норме

Пусть

$$\sigma_{k+1} = \sum_{j=1}^{n}(1-s_j)\left|x_j^{(k+1)} - x_j^{(k)}\right| \quad (k=0,\ 1,\ 2,\ \ldots).$$

Использовав преобразования, аналогичные примененным при доказательстве теоремы предыдущего параграфа, для двух последовательных итераций $x_j^{(k)}$ и $x_j^{(k+1)}$ получим неравенство ((6) § 5)

$$\sigma_{k+1} \leqslant \rho\sigma_k, \tag{1}$$

где в силу неравенства (5) из § 5

$$\rho = \max_j \frac{t_j}{1-s_j} \leqslant \|\alpha\|_l.$$

Отсюда

$$\sigma_{k+p} \leqslant \rho^p\sigma_k \quad (p=1,\ 2,\ \ldots).$$

Далее имеем:

$$\sum_{j=1}^{n}(1-s_j)\left|x_j^{(k+p)} - x_j^{(k)}\right| \leqslant \sigma_{k+p} + \sigma_{k+p-1} + \ldots + \sigma_{k+1} \leqslant$$

$$\leqslant \rho^p\sigma_k + \rho^{p-1}\sigma_k + \ldots + \rho\sigma_k \leqslant \frac{\rho\sigma_k}{1-\rho}.$$

Отсюда при $p \to \infty$ получим:

$$\sum_{j=1}^{n} (1-s_j) \left| x_j - x_j^{(k)} \right| \leqslant \frac{\rho\sigma_k}{1-\rho}$$

или

$$\sum_{j=1}^{n} \left| x_j - x_j^{(k)} \right| \leqslant \frac{\rho}{(1-s)(1-\rho)} \sum_{j=1}^{n} \left| x_j^{(k)} - x_j^{(k-1)} \right|,$$

где

$$s = \max_j s_j = \max_j \sum_{i=j+1}^{n} |\alpha_{ij}|.$$

Так как из формулы (1) вытекает, что

$$\sigma_k \leqslant \rho^{k-1}\sigma_1,$$

то справедлива также оценка

$$\left\| x_j - x_j^{(k)} \right\|_l = \sum_{j=1}^{n} \left| x_j - x_j^{(k)} \right| \leqslant \frac{\rho^k}{(1-s)(1-\rho)} \sigma_1 \leqslant$$

$$\leqslant \frac{\rho^k}{(1-s)(1-\rho)} \sum_{j=1}^{n} \left| x_j^{(1)} - x_j^{(0)} \right|.$$

## § 7. Третье достаточное условие сходимости процесса Зейделя

Теорема. *Если для линейной системы*

$$x = \alpha x + \beta \tag{1}$$

*выполнено условие*

$$\|\alpha\|_k < 1,$$

*где*

$$\|\alpha\|_k = \sqrt{\sum_{i,j} |\alpha_{ij}|^2},$$

*то процесс Зейделя для системы* (1) *сходится к единственному ее решению при любом выборе начального вектора.*

Доказательство. Пусть

$$x = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \quad \text{и} \quad x^{(p)} = \begin{bmatrix} x_1^{(p)} \\ \cdot \\ \cdot \\ \cdot \\ x_n^{(p)} \end{bmatrix}$$

— соответственно точное решение системы (1) и $p$-е приближение ($p=0, 1, 2, \ldots$) процесса Зейделя для этой системы. Имеем:

$$x_i=\sum_{j=1}^{i-1}\alpha_{ij}x_j+\sum_{j=i}^{n}\alpha_{ij}x_j+\beta_i$$

и

$$x_i^{(p)}=\sum_{j=1}^{i-1}\alpha_{ij}x_j^{(p)}+\sum_{j=i}^{n}\alpha_{ij}x_j^{(p-1)}+\beta_i$$

($i=1, 2, \ldots, n$). Отсюда

$$x_i-x_i^{(p)}=\sum_{i=1}^{i-1}\alpha_{ij}\left(x_j-x_j^{(p)}\right)+\sum_{j=i}^{n}\alpha_{ij}\left(x_j-x_j^{(p-1)}\right)$$

и, следовательно,

$$|x_i-x_i^{(p)}|^2\leqslant\left\{\sum_{j=1}^{i-1}|\alpha_{ij}||x_j-x_j^{(p)}|+\sum_{j=i}^{n}|\alpha_{ij}||x_j-x_j^{(p-1)}|\right\}^2.$$

Применяя неравенство Коши (гл. VII, § 7) к сумме всех слагаемых, стоящих в фигурной скобке, будем иметь:

$$|x_i-x_i^{(p)}|^2\leqslant s_i\left\{\sum_{j=1}^{i-1}|x_j-x_j^{(p)}|^2+\sum_{j=i}^{n}|x_j-x_j^{(p-1)}|^2\right\}, \qquad (2)$$

где

$$s_i=\sum_{j=1}^{n}|\alpha_{ij}|^2 \quad (i=1, 2, \ldots, n).$$

Суммируя неравенства (2) по $i$ от 1 до $n$, получим:

$$\sum_{i=1}^{n}|x_i-x_i^{(p)}|^2\leqslant\sum_{i=1}^{n}\sum_{j=1}^{i-1}s_i|x_j-x_j^{(p)}|^2+\sum_{i=1}^{n}\sum_{j=i}^{n}s_i|x_j-x_j^{(p-1)}|^2.$$

Изменяя индекс суммирования в левой части и порядок суммирования в правой части последнего неравенства, будем иметь:

$$\sum_{j=1}^{n}|x_j-x_j^{(p)}|^2\leqslant\sum_{j=1}^{n-1}|x_j-x_j^{(p)}|^2\sum_{i=j+1}^{n}s_i+\sum_{j=1}^{n}|x_j-x_j^{(p-1)}|^2\sum_{i=1}^{j}s_i. \qquad (3)$$

Пусть

$$S_j=\sum_{i=j+1}^{n}s_i, \quad T_j=\sum_{i=1}^{j}s_i \quad (j=1, 2, \ldots, n-1)$$

и

$$S_n=0, \quad T_n=\sum_{i=1}^{n}s_i.$$

Очевидно,

$$S_j+T_j=\sum_{i=1}^{n} s_i=\sum_{i=1}^{n}\sum_{j=1}^{n}|\alpha_{ij}|^2=\|\alpha\|_k^2<1 \quad (j=1,\ 2,\ \ldots,\ n). \quad (4)$$

Пользуясь этими обозначениями, неравенству (3) можно придать вид

$$\sum_{j=1}^{n}|x_j-x_j^{(p)}|^2\leqslant\sum_{j=1}^{n}S_j|x_j-x_j^{(p)}|^2+\sum_{j=1}^{n}T_j|x_j-x_j^{(p-1)}|^2$$

или

$$\sum_{j=1}^{n}(1-S_j)|x_j-x_j^{(p)}|^2\leqslant\sum_{j=1}^{n}T_j|x_j-x_j^{(p-1)}|^2.$$

На основании формулы (4) получаем:

$$T_j=\|\alpha\|_k^2-S_j\leqslant\|\alpha\|_k^2-\|\alpha\|_k^2S_j=\|\alpha\|_k^2(1-S_j).$$

Поэтому

$$\sum_{j=1}^{n}(1-S_j)|x_j-x_j^{(p)}|^2\leqslant\|\alpha\|_k^2\sum_{j=1}^{n}(1-S_j)|x_j-x_j^{(p-1)}|^2. \quad (5)$$

Из неравенства (5) при $p>1$ последовательно выводим:

$$\sum_{j=1}^{n}(1-S_j)|x_j-x_j^{(p)}|^2\leqslant(\|\alpha\|_k^2)^p\sum_{j=1}^{n}(1-S_j)|x_j-x_j^{(0)}|^2.$$

Так как $\|\alpha\|_k<1$, то отсюда будем иметь:

$$\lim_{p\to\infty}\sum_{j=1}^{n}(1-S_j)|x_j-x_j^{(p)}|^2=0,$$

и, следовательно, учитывая, что $0\leqslant S_j<1$ $(j=1,\ 2,\ \ldots,\ n)$, получим:

$$\lim_{p\to\infty}x_j^{(p)}=x_j \quad (j=1,\ 2,\ \ldots,\ n),$$

что и требовалось доказать.

Замечание. Погрешность итераций $x^{(p)}$ $(p=1,\ 2,\ \ldots)$ оценивается аналогично тому, как это было сделано в § 6.

### Литература к девятой главе

1. В. Н. Фаддеева. Вычислительные методы линейной алгебры, Гостехиздат, М.—Л., 1950, гл. II, § 17 и 19.

ГЛАВА X

# ОСНОВНЫЕ СВЕДЕНИЯ ИЗ ТЕОРИИ ЛИНЕЙНЫХ ВЕКТОРНЫХ ПРОСТРАНСТВ

## § 1. Понятие линейного векторного пространства

Определение. Упорядоченная совокупность $n$ чисел $x=(x_1, x_2, \ldots, x_n)$, вообще говоря, комплексных, называется *точкой* или *вектором* $n$-мерного пространства, а числа $x_1, x_2, \ldots, x_n$ называются *координатами* вектора $x$ [1], [2], [3]. В качестве примеров векторов укажем следующие:

1) свободные векторы на плоскости или в трехмерном пространстве будут соответственно двумерными или трехмерными векторами в смысле данного выше определения;

2) всякое решение любой системы линейных уравнений с $n$ неизвестными будет $n$-мерным вектором;

3) если дана матрица из $n$ строк и $m$ столбцов, то ее строки будут $m$-мерными векторами, столбцы — $n$-мерными векторами.

Два вектора $x=(x_1, x_2, \ldots, x_n)$ и $y=(y_1, y_2, \ldots, y_n)$ считаются равными тогда и только тогда, когда совпадают их координаты, стоящие на одинаковых местах, т. е. если $x_i=y_i$ при $i=1, 2, \ldots, n$.

Обозначим вектор $(0, 0, \ldots, 0)$ через $\mathbf{0}$ и назовем его *нулевым вектором*.

*Суммой векторов* $x=(x_1, x_2, \ldots, x_n)$, $y=(y_1, y_2, \ldots, y_n)$ называется вектор

$$x+y=(x_1+y_1;\ x_2+y_2;\ \ldots;\ x_n+y_n),$$

координаты которого суть суммы соответствующих координат слагаемых векторов. Сложение векторов подчиняется перестановочному и сочетательному законам:

1) $$x+y=y+x;$$

2) $$(x+y)+z=x+(y+z).$$

Аналогично определяется разность векторов $x$ и $y$. Вектор $-x$, удовлетворяющий условию $(-x)+x=\mathbf{0}$, называется противоположным вектору $x$. Легко показать, что

$$x-y=x+(-y).$$

*Произведением вектора* $x=(x_1, x_2, \ldots, x_n)$ на число $k$ называется вектор

$$kx=(kx_1, kx_2, \ldots, kx_n).$$

Из этого определения вытекают следующие свойства произведения вектора на число:

1) $$k(x \pm y)=kx \pm ky;$$

2) $$(k \pm l)x=kx \pm lx;$$

3) $$k(lx)=(kl)x;$$

4) $$0x=0;$$

5) $$1x=x;$$

6) $$(-1)x=-x,$$

где $k$ и $l$ — произвольные числа, а $x$ и $y$ — векторы.

Для векторов $x$ и $y$ естественно определяется *линейная комбинация*

$$\alpha x+\beta y$$

($\alpha$, $\beta$ — числа), как вектор с координатами $\alpha x_j+\beta y_j$ $(j=1, 2, \ldots, n)$.

Любая совокупность $n$-мерных векторов, рассматриваемая с установленными в ней операциями сложения векторов и умножения вектора на число, не выводящими за пределы этой совокупности, называется *линейным векторным пространством*. В частности, совокупность всех $n$-мерных векторов образует $n$-мерное векторное пространство $E_n$.

## § 2. Линейная зависимость векторов

Определение 1. Векторы $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ пространства $E_n$ называются *линейно зависимыми*, если существуют числа $c_1, c_2, \ldots, c_m$, не все равные нулю и такие, что

$$c_1x^{(1)}+c_2x^{(2)}+\ldots+c_mx^{(m)}=0. \tag{1}$$

Пусть, например, $c_m \neq 0$. Тогда из равенства (1) будем иметь:

$$x^{(m)}=\gamma_1x^{(1)}+\gamma_2x^{(2)}+\ldots+\gamma_{m-1}x^{(m-1)},$$

где

$$\gamma_j=-\frac{c_j}{c_m} \quad (j=1, 2, \ldots, m-1).$$

Таким образом, *данные векторы линейно зависимы тогда и только тогда, когда один из них является линейной комбинацией остальных.*

Если же равенство (1) возможно в единственном случае, когда $c_1=c_2=\ldots=c_m=0$, то векторы $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ называются

*линейно независимыми*, т. е. *векторы линейно независимы тогда и только тогда, когда никакая линейная комбинация их, не все коэффициенты которой равны нулю, не является нуль-вектором.* Заметим, что среди линейно независимых векторов, очевидно, не должно быть нуль-вектора.

Пример 1. Для случая трехмерного векторного пространства $E_3$ линейная зависимость двух векторов $\boldsymbol{x}$ и $\boldsymbol{y}$ означает, что они параллельны некоторой прямой, а линейная зависимость трех векторов $\boldsymbol{x}$, $\boldsymbol{y}$ и $\boldsymbol{z}$ — что они параллельны некоторой плоскости.

Отметим, что если часть векторов линейно зависима, то вся совокупность векторов также линейно зависима.

Пусть имеется совокупность векторов

$$\boldsymbol{x}^{(j)} = \left(x_1^{(j)},\ x_2^{(j)},\ \ldots,\ x_n^{(j)}\right) \quad (j = 1,\ 2,\ \ldots,\ m).$$

Тогда для определения постоянных $c_k\,(k = 1,\ 2,\ \ldots,\ m)$ в силу равенства (1) получаем систему

$$\left.\begin{array}{l} c_1x_1^{(1)} + c_2x_1^{(2)} + \ldots + c_mx_1^{(m)} = 0, \\ c_1x_2^{(1)} + c_2x_2^{(2)} + \ldots + c_mx_2^{(m)} = 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ c_1x_n^{(1)} + c_2x_n^{(2)} + \ldots + c_mx_n^{(m)} = 0. \end{array}\right\} \qquad (2)$$

Если эта система имеет ненулевые решения, то данные векторы линейно зависимы. В противном случае они линейно независимы.

Рассмотрим матрицу координат

$$X = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \ldots & x_1^{(m)} \\ x_2^{(1)} & x_2^{(2)} & \ldots & x_2^{(m)} \\ \ldots & \ldots & \ldots & \ldots \\ x_n^{(1)} & x_n^{(2)} & \ldots & x_n^{(m)} \end{bmatrix}.$$

Пусть $r$ — ранг этой матрицы. В алгебре доказывается [2], что система (2) имеет ненулевые решения тогда и только тогда, когда $r < m$. Следовательно, векторы $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(m)}$ линейно зависимы, если $r < m$, и линейно независимы, если $r = m$ (больше $m$ ранг $r$, очевидно, быть не может).

Отсюда вытекает, что ранг матрицы $X$ дает нам максимальное число линейно независимых векторов, содержащихся в данной совокупности векторов.

Таким образом, если ранг матрицы $X$ равен $r$, то среди векторов-столбцов $\boldsymbol{x}^{(j)}$ $(j = 1, 2, \ldots, m)$: 1) найдется $r$ линейно независимых и 2) каждые $r+1$ векторов $(r+1 \leqslant m)$ из этой совокупности линейно зависимы. То же самое справедливо относительно векторов-строк $\left(x_i^{(1)}, \ldots, x_i^{(m)}\right)$ $(i = 1,\ 2,\ \ldots,\ n)$ матрицы $X$.

Пример 2. Исследовать на линейную зависимость систему векторов:

$$\begin{aligned} x^{(1)} &= (1, \ -1, \ 1, \ -1, \ 1); \\ x^{(2)} &= (1, \ 0, \ 2, \ 0, \ 1); \\ x^{(3)} &= (1, \ -5, \ -1, \ 2, \ -1); \\ x^{(4)} &= (3, \ -6, \ 2, \ 1, \ 1). \end{aligned}$$

Решение. Составляем матрицу координат

$$X = \begin{bmatrix} 1 & 1 & 1 & 3 \\ -1 & 0 & -5 & -6 \\ 1 & 2 & -1 & 2 \\ -1 & 0 & 2 & 1 \\ 1 & 1 & -1 & 1 \end{bmatrix}.$$

Для определения ранга $r$ матрицы $X$ проведем некоторые элементарные преобразования, а именно, вычитая из четвертого столбца матрицы сумму трех первых, получаем:

$$X \sim \begin{bmatrix} 1 & 1 & 1 & 0 \\ -1 & 0 & -5 & 0 \\ 1 & 2 & -1 & 0 \\ -1 & 0 & 2 & 0 \\ 1 & 1 & -1 & 0 \end{bmatrix}.$$

Отсюда заключаем, что все определители четвертого порядка матрицы $X$ равны нулю. Очевидно, что имеются миноры третьего порядка матрицы $X$, отличные от нуля. Следовательно, $r = 3$, и так как ранг матрицы меньше числа векторов, то векторы $x^{(1)}$, $x^{(2)}$, $x^{(3)}$, $x^{(4)}$ линейно зависимы. В данном случае это ясно, так как

$$x^{(1)} + x^{(2)} + x^{(3)} - x^{(4)} = 0.$$

Теорема 1. *Максимальное число линейно независимых векторов $n$-мерного пространства $E_n$ в точности равно размерности этого пространства.*

Доказательство. Прежде всего, в пространстве $E_n$ имеются системы из $n$ линейно независимых векторов. Такова, например, совокупность $n$ единичных векторов (*ортов*):

$$\begin{aligned} e_1 &= (1, \ 0, \ 0, \ \ldots, \ 0); \\ e_2 &= (0, \ 1, \ 0, \ \ldots, \ 0); \\ &\ldots\ldots\ldots\ldots \\ e_n &= (0, \ 0, \ 0, \ \ldots, \ 1). \end{aligned}$$

Так как если

$$c_1 e_1 + c_2 e_2 + \ldots + c_n e_n = (c_1, \ c_2, \ \ldots, \ c_n) = 0,$$

то очевидно, что $c_1 = c_2 = \ldots = c_n = 0$.

Покажем, что если число векторов $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ больше $n$ $(m > n)$, то они обязательно линейно зависимы. Действительно, матрица координат этих векторов имеет тип $n \times m$ и, следовательно, ранг ее $r \leqslant \min(n, m) = n < m$. Отсюда следует, что эти векторы линейно зависимы.

Определение 2. Любая совокупность $n$ линейно независимых векторов $n$-мерного пространства называется *базисом* этого пространства.

Теорема 2. *Каждый вектор $n$-мерного пространства $E_n$ может быть представлен, и притом единственным образом, в виде линейной комбинации векторов базиса.*

Доказательство. Пусть $x \in E_n$ и $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ — базис пространства $E_n$. В силу теоремы 1 векторы $x, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ линейно зависимы, т. е.

$$c_0 x + c_1 \varepsilon_1 + c_2 \varepsilon_2 + \ldots + c_n \varepsilon_n = 0, \tag{3}$$

где некоторый коэффициент $c_j \neq 0$ $(0 \leqslant j \leqslant n)$.

В равенстве (3) коэффициент $c_0 \neq 0$, так как в противном случае мы бы имели

$$c_1 \varepsilon_1 + c_2 \varepsilon_2 + \ldots + c_n \varepsilon_n = 0,$$

где $c_j \neq 0$ $(j \geqslant 1)$, что противоречит линейной независимости векторов $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$. Следовательно, мы можем разрешить равенство (3) относительно $x$:

$$x = \xi_1 \varepsilon_1 + \xi_2 \varepsilon_2 + \ldots + \xi_n \varepsilon_n, \tag{4}$$

где

$$\xi_1 = -\frac{c_1}{c_0}, \quad \xi_2 = -\frac{c_2}{c_0}, \quad \ldots, \quad \xi_n = -\frac{c_n}{c_0}.$$

Таким образом, любой вектор $x$ пространства $E_n$ есть линейная комбинация векторов базиса. Разложение (4) единственно. В самом деле, если имеется другое разложение

$$x = \xi_1' \varepsilon_1 + \xi_2' \varepsilon_2 + \ldots + \xi_n' \varepsilon_n, \tag{4'}$$

отличное от первого, то, вычитая из равенства (4) равенство (4′), получим:

$$0 = (\xi_1 - \xi_1') \varepsilon_1 + (\xi_2 - \xi_2') \varepsilon_2 + \ldots + (\xi_n - \xi_n') \varepsilon_n, \tag{5}$$

где по меньшей мере один из коэффициентов $\xi_j - \xi_j' \neq 0$. Равенство (5) невозможно, так как векторы базиса линейно независимы. Следовательно, существует только одно разложение вида (4).

**Геометрическая иллюстрация.** Для случая трехмерного пространства формула (4) эквивалентна разложению вектора $x$ по направлениям трех данных векторов $\varepsilon_1$, $\varepsilon_2$ и $\varepsilon_3$ общего положения (рис. 49).

Определение 3. Если $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ есть базис $n$-мерного пространства и

$$x = \xi_1\varepsilon_1 + \xi_2\varepsilon_2 + \ldots + \xi_n\varepsilon_n,$$

то числа $\xi_1, \xi_2, \ldots, \xi_n$ называют *координатами* вектора $x$ в данном базисе $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$. Заметим, что координаты вектора

$$x = (x_1, x_2, \ldots, x_n)$$

есть координаты его в базисе ортов

$$e_j = (\delta_{1j}, \delta_{2j}, \ldots, \delta_{nj})$$
$$(j = 1, 2, \ldots, n),$$

где $\delta_{nj}$ — символ Кронекера. Следовательно, имеем основное разложение

$$x = x_1e_1 + x_2e_2 + \ldots + x_ne_n. \quad (6)$$

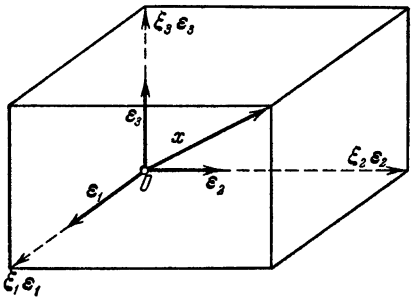


Рис. 49.

Базис ортов $e_j$ $(j = 1, 2, \ldots, n)$ будем называть *исходным базисом* пространства.

Определение 4. Совокупность $E_k$ векторов из $n$-мерного пространства $E_n$ называется *линейным подпространством* пространства $E_n$, если выполнены следующие условия:

1) из $x \in E_k$ и $y \in E_k$ следует $x + y \in E_k$;

2) из $x \in E_k$ следует $\alpha x \in E_k$, где $\alpha$ — любое число. В частности, $0 \in E_k$.

Следовательно $E_k$ можно также считать векторным пространством. Максимальное число $k$ линейно независимых векторов в пространстве $E_k$ называется *размерностью* этого подпространства.

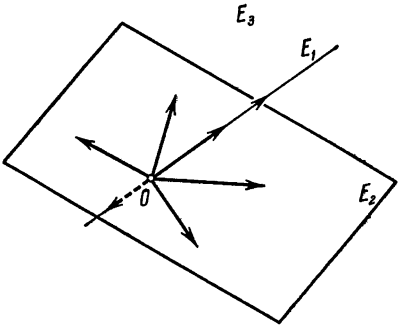


Рис. 50.

Из теоремы 1 следует, что $k \leqslant n$. Таким образом, в пространстве $E_n$ могут быть подпространства: $E_1$ — одного измерения, $E_2$ — двух измерений и т. д. до $E_n$ — $n$ измерений (само пространство). Нулевой вектор $0$ можно рассматривать как пространство нулевого измерения.

Пример 3. В обычном трехмерном пространстве $E_3$ подпространство $E_1$ одного измерения является прямой; подпространство $E_2$ двух измерений — плоскостью (рис. 50).

Теорема 3. *Если $z_1, z_2, \ldots, z_k$ — векторы $n$-мерного пространства $E_n$, то полная совокупность векторов*

$$x = a_1 z_1 + a_2 z_2 + \ldots + a_k z_k, \tag{7}$$

*где $a_j$ $(j = 1, 2, \ldots, k)$ — произвольные числа, представляет собой подпространство пространства $E_n$, причем, если векторы $z_1, z_2, \ldots, z_k$ $(k \leqslant n)$ линейно независимы, то размерность этого подпространства равна $k$.*

*Обратно, всякое подпространство $E_k$ пространства $E_n$ совпадает с совокупностью всех линейных комбинаций линейно независимых векторов $z_1, z_2, \ldots, z_k$ этого подпространства (базисные векторы).*

Доказательство. Справедливость первого утверждения теоремы проверяется непосредственно.

Докажем второе утверждение теоремы. Пусть $x \in E_k$ и $x$ не является линейной комбинацией базисных векторов $z_1, z_2, \ldots, z_k$. Тогда очевидно, что векторы $x, z_1, z_2, \ldots, z_k$ линейно независимы, и, таким образом, в пространстве $E_k$ имеется $k+1$ линейно независимых векторов. Но последнее обстоятельство невозможно, так как согласно предположению максимальное число линейно независимых векторов пространства $E_k$ равно $k$.

Следовательно, при каком-нибудь выборе чисел $a_1, a_2, \ldots, a_k$ имеем:

$$x = a_1 z_1 + a_2 z_2 + \ldots + a_k z_k,$$

что и требовалось доказать.

Следствие. Совокупность векторов $x$, определяемых формулой (7), представляет собой наименьшее линейное пространство, содержащее векторы $z_1, z_2, \ldots, z_k$ (так называемое *пространство, порожденное векторами* $z_1, z_2, \ldots, z_k$, или пространство, *натянутое на векторы* $z_1, z_2, \ldots, z_k$).

## § 3. Скалярное произведение векторов

Пусть в $n$-мерном пространстве $E_n$ имеем векторы

$$x = (x_1, x_2, \ldots, x_n) \text{ и } y = (y_1, y_2, \ldots, y_n).$$

Будем считать, что координаты векторов — комплексные числа:

$$x_j = \xi_j + i\xi_j'; \quad y_j = \eta_j + i\eta_j',$$

где $i = \sqrt{-1}$; $j = 1, 2, \ldots, n$.

Введем сопряженные величины

$$x_j^* = \xi_j - i\xi_j'; \qquad y_j^* = \eta_j - i\eta_j'.$$

Тогда очевидно, что

$$x_j x_j^* = |x_j|^2.$$

Под *скалярным произведением* двух векторов понимается число, равное

$$(x, y) = \sum_{j=1}^{n} x_j y_j^*. \tag{1}$$

Скалярное произведение обладает следующими свойствами.

**1. Свойство положительной определенности.** Скалярное произведение вектора самого на себя есть неотрицательное число, которое равно нулю тогда и только тогда, когда вектор равен нулю. В самом деле, из формулы (1) имеем:

$$(x, x) = \sum_{j=1}^{n} x_j x_j^* = \sum_{j=1}^{n} |x_j|^2 \geqslant 0.$$

Очевидно, что $(\mathbf{0}, \mathbf{0}) = 0$. Наоборот, если $(x, x) = 0$, то $x_j = 0$ $(j = 1, 2, \ldots, n)$ и, следовательно, $x = \mathbf{0}$.

**2. Эрмитова симметрия.** При перестановке двух множителей скалярное произведение заменяется сопряженным. Действительно, пользуясь теоремами о сопряженной величине суммы и сопряженной величине произведения*), имеем:

$$(y, x) = \sum_{j=1}^{n} y_j x_j^* = \sum_{j=1}^{n} x_j^* y_j = \left(\sum_{j=1}^{n} x_j y_j^*\right)^* = (x, y)^*.$$

Следовательно,

$$(y, x) = (x, y)^*. \tag{2}$$

**3. Скалярный множитель,** стоящий на первом месте, можно выносить за знак скалярного произведения, т. е.

$$(\alpha x, y) = \alpha (x, y). \tag{3}$$

Доказательство этого свойства непосредственно вытекает из формулы (1).

Следствие. Скалярный множитель, стоящий на втором месте, можно выносить за знак скалярного произведения, заменяя его сопряженным. Имеем:

$$(x, \alpha y) = (\alpha y, x)^* = [\alpha (y, x)]^* = \alpha^* (y, x)^* = \alpha^* (x, y).$$

Итак,

$$(x, \alpha y) = \alpha^* (x, y).$$

**4. Свойство дистрибутивности.** Если первый или второй векторы представляют собой сумму двух векторов, то скалярное произведение

---

*) Мы здесь воспользовались следующими теоремами:

а) сопряженная величина суммы равна сумме сопряженных величин слагаемых;

б) сопряженная величина произведения равна произведению сопряженных величин сомножителей.

этого вектора есть сумма соответствующих скалярных произведений слагаемых этого вектора. В самом деле, пусть

$$x = x^{(1)} + x^{(2)},$$

где $x^{(k)} = (x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)})$ $(k = 1, 2)$.

Исходя из определения суммы векторов, согласно формуле (1) получаем:

$$(x^{(1)} + x^{(2)}, y) = \sum_{j=1}^{n} (x_j^{(1)} + x_j^{(2)}) y_j^* =$$

$$= \sum_{j=1}^{n} x_j^{(1)} y_j^* + \sum_{j=1}^{n} x_j^{(2)} y_j^* = (x^{(1)}, y) + (x^{(2)}, y),$$

т. е.

$$(x^{(1)} + x^{(2)}, y) = (x^{(1)}, y) + (x^{(2)}, y). \tag{4}$$

Далее,

$$(x, y^{(1)} + y^{(2)}) = (y^{(1)} + y^{(2)}, x)^* = (y^{(1)}, x)^* + (y^{(2)}, x)^* =$$

$$= (x, y^{(1)}) + (x, y^{(2)}). \tag{5}$$

Формулы (4) и (5) легко распространяются на любое конечное число векторов, а именно:

$$\left(\sum_{j=1}^{m} x^{(j)}, \sum_{k=1}^{l} y^{(k)}\right) = \sum_{j=1}^{m} \sum_{k=1}^{l} (x^{(j)}, y^{(k)}).$$

Кроме введенного $n$-мерного *комплексного пространства*, полезно рассматривать *$n$-мерное вещественное* (*действительное*) пространство, т. е. совокупность векторов с вещественными координатами.

В действительном $n$-мерном пространстве скалярное произведение равно сумме произведений соответствующих координат векторов

$$(x, y) = \sum_{j=1}^{n} x_j y_j. \tag{1'}$$

Рассмотренные выше свойства скалярного произведения формулируются так:

1) $(x, x) \geqslant 0$, причем если $(x, x) = 0$, то $x = 0$;
2) $(x, y) = (y, x)$;
3) $(\alpha x, y) = (x, \alpha y) = \alpha\,(x, y)$ ($\alpha$ действительно);
4) $(x + y, z) = (x, z) + (y, z)$;
   $(x, y + z) = (x, y) + (x, z)$.

С помощью скалярного произведения можно дать определение и основных метрических понятий в $n$-мерном пространстве: длины вектора и угла между парой векторов.

1. **Длина вектора.** Длиной вектора в $n$-мерном пространстве называется неотрицательное число

$$|x| = +\sqrt{(x, x)}.$$

Очевидно, что это определение согласуется с понятием длины вектора в трехмерном пространстве.

2. **Угол между векторами.** Углом $\varphi$ между парой векторов $x$ и $y$ называется тот угол (в пределах от 0 до 180°), для которого

$$\cos\varphi = \frac{(x, y)}{|x||y|}.$$

Для векторов в трехмерном пространстве это определение согласуется с обычным выражением угла между векторами через скалярное произведение. Можно доказать, что справедливо неравенство [1]

$$|(x, y)| \leqslant |x||y|.$$

Поэтому угол между векторами в вещественном пространстве будет действителен.

## § 4. Ортогональные системы векторов

Определение 1. Два вектора $x$ и $y$ пространства $E_n$ называются *ортогональными*, если их скалярное произведение равно нулю, т. е.

$$(x, y) = 0. \tag{1}$$

Если векторы не нулевые, то ортогональность означает, что угол между ними равен $\frac{\pi}{2}$. Нулевой вектор, очевидно, ортогонален любому вектору пространства.

Таким образом, ортогональность есть обобщенное свойство перпендикулярности.

Определение 2. Система векторов $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ называется *ортогональной*, если любые два вектора системы ортогональны друг другу, т. е.

$$(x^{(j)}, x^{(k)}) = 0 \quad \text{при} \quad j \neq k.$$

Заметим, что если вектор $x^{(1)}$ ортогонален векторам $x^{(2)}, \ldots, x^{(m)}$, то этот вектор ортогонален также любой линейной комбинации последних векторов, иными словами, вектор $x^{(1)}$ ортогонален пространству, натянутому на векторы $x^{(2)}, \ldots, x^{(m)}$. Действительно, если

$$(x^{(1)}, x^{(k)}) = 0 \quad \text{при} \quad k = 2, \ldots, m,$$

то имеем:

$$\left(x^{(1)}, \sum_{k=2}^{m} c_k x^{(k)}\right) = \sum_{k=2}^{m} c_k^* (x^{(1)}, x^{(k)}) = 0,$$

где $c_2, \ldots, c_m$ — произвольные постоянные.

Теорема. *Ненулевые попарно ортогональные векторы* $x^{(1)}$, $x^{(2)}, \ldots, x^{(m)}$ *линейно независимы.*

Доказательство. В самом деле, пусть

$$c_1 x^{(1)} + c_2 x^{(2)} + \ldots + c_m x^{(m)} = 0. \qquad (2)$$

Умножая скалярно обе части равенства (2) на $x^{(1)}$, получим:

$$c_1^* (x^{(1)}, x^{(1)}) + c_2^* (x^{(1)}, x^{(2)}) + \ldots + c_m^* (x^{(1)}, x^{(m)}) = 0,$$

или, так как

$(x^{(1)}, x^{(1)}) \neq 0$ и $(x^{(1)}, x^{(j)}) = 0$ при $j \neq 1$, то $c_1^* = 0$ и $c_1 = 0$.

Совершенно так же доказываем, что $c_2 = 0, \ldots, c_m = 0$. Следовательно, векторы $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ линейно независимы.

Следствие. В $n$-мерном пространстве $E_n$ ортогональная система содержит не более $n$ векторов.

Определение 3. Базис $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ пространства $E_n$ называется *ортогональным,* если базисные векторы попарно ортогональны, т. е.

$$(\varepsilon_j, \varepsilon_k) = 0 \quad \text{при} \quad j \neq k \quad (j, k = 1, 2, \ldots, n).$$

Если, сверх того, векторы $\varepsilon_j$ $(j = 1, 2, \ldots, n)$ — единичные, то ортогональный базис называется *нормированным* (короче, *ортонормированным*). В этом случае имеем:

$$(\varepsilon_j, \varepsilon_k) = \delta_{jk},$$

где $\delta_{jk}$ — символ Кронекера.

Простейший ортонормированный базис пространства $E_n$, как нетрудно убедиться, представляет собой систему ортов

$$\begin{aligned} e_1 &= (1, 0, 0, \ldots, 0), \\ e_2 &= (0, 1, 0, \ldots, 0), \\ &\ldots\ldots\ldots\ldots \\ e_n &= (0, 0, 0, \ldots, 1), \end{aligned}$$

образующих исходный базис.

Ортогональный базис $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ всегда можно нормировать, разделив каждый из векторов $\varepsilon_j$ на его длину. Полученные новые векторы

$$\tilde{\varepsilon}_j^{(0)} = \frac{\varepsilon_j}{\sqrt{(\varepsilon_j, \varepsilon_j)}} \qquad (j = 1, 2, \ldots, n)$$

образуют ортонормированный базис.

Выразим координаты вектора $x$ в ортонормированном базисе $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$. Если

$$x = \xi_1\varepsilon_1 + \xi_2\varepsilon_2 + \ldots + \xi_n\varepsilon_n, \tag{3}$$

то, умножая скалярно равенство (3) справа на $\varepsilon_j$, получим:

$$\xi_j = (x, \varepsilon_j) \quad (j = 1, 2, \ldots, n). \tag{4}$$

По аналогии с векторной алгеброй можно сказать, что *координаты вектора в ортонормированном базисе равны проекциям вектора на соответствующие векторы базиса.*

Возводя равенство (3) в квадрат, будем иметь:

$$(x, x) = \left(\sum_{j=1}^{n} \xi_j\varepsilon_j, \sum_{k=1}^{n} \xi_k\varepsilon_k\right) = \\ = \sum_{j=1}^{n}\sum_{k=1}^{n} \xi_j\xi_k^*(\varepsilon_j, \varepsilon_k) = \sum_{j=1}^{n} \xi_j\xi_j^* = \sum_{j=1}^{n} |\xi_j|^2, \tag{5}$$

т. е. *квадрат длины вектора равен сумме квадратов модулей его проекций на базисные ортонормированные векторы* (*аналог теоремы Пифагора*). В частности, если пространство $E_n$ — действительное, то формулу (5) можно записать без модуля:

$$(x, x) = \sum_{j=1}^{n} (\xi_j)^2. \tag{5'}$$

## § 5. Преобразования координат вектора при изменениях базиса

Пусть $e_1, e_2, \ldots, e_n$ и $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ — два базиса одного и того же линейного пространства $E_n$. Каждый вектор нового (второго) базиса $\varepsilon_j$ имеет в старом (первом) базисе $e_j$ некоторые координаты $s_{1j}, s_{2j}, \ldots, s_{nj}$*), т. е.

$$\varepsilon_j = s_{1j}e_1 + s_{2j}e_2 + \ldots + s_{nj}e_n \quad (j = 1, 2, \ldots, n). \tag{1}$$

Неособенная матрица $S = [s_{ij}]$ называется *матрицей перехода* от старого базиса к новому **). Эта матрица является транспонированной по отношению к матрице, задающей преобразование базиса. Пусть $x$ — данный вектор. Обозначим через $x_i$ координаты этого вектора в старом базисе и через $\xi_i$ — его координаты в новом базисе.

---

*) При обозначении координат на первом месте указывается номер соответствующего старого базисного вектора, а на втором — нового базисного вектора.

**) Определитель $\det S \neq 0$, так как в противном случае векторы $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ были бы линейно зависимы.

Очевидно, что

$$x=\sum_{i=1}^{n} x_i e_i=\sum_{j=1}^{n} \xi_j \varepsilon_j.$$

Отсюда, подставляя во вторую сумму выражение (1) для $\varepsilon_j$, получим:

$$x=\sum_{i=1}^{n} x_i e_i=\sum_{j=1}^{n} \xi_j \sum_{i=1}^{n} s_{ij} e_i=\sum_{i=1}^{n} e_i \sum_{j=1}^{n} s_{ij}\xi_j.$$

Следовательно, в силу линейной независимости векторов $e_1$, $e_2$, ..., $e_n$ находим:

$$x_i=\sum_{j=1}^{n} s_{ij}\xi_j \quad (i=1,\ 2,\ \ldots,\ n). \tag{2}$$

Если обозначить

$$x=\begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \quad \text{и} \quad \xi=\begin{bmatrix} \xi_1 \\ \cdot \\ \cdot \\ \cdot \\ \xi_n \end{bmatrix}^{*)},$$

то соотношение (2) можно переписать в следующем матричном виде:

$$x=S\xi, \tag{3}$$

т. е. *вектор в старых координатах (базисе) равен матрице перехода $S$* (или транспонированной матрице, задающей новый базис), *умноженной на вектор в новых координатах.*

Из формулы (3) получаем:

$$\xi=S^{-1}x. \tag{4}$$

Отметим один важный частный случай, аналогичный преобразованию прямоугольных координат. Пусть старый базис $e_1$, $e_2$, ..., $e_n$ и новый базис $\varepsilon_1$, $\varepsilon_2$, ..., $\varepsilon_n$ действительны и ортонормированы, т. е.

$$(e_i,\ e_j)=\delta_{ij} \tag{5}$$

и

$$(\varepsilon_i,\ \varepsilon_j)=\delta_{ij}, \tag{5'}$$

где $\delta_{ij}$ — символ Кронекера.

Тогда из формулы (1) вытекает:

$$s_{ij}=(\varepsilon_j,\ e_i) \quad (i,\ j=1,\ 2,\ \ldots,\ n), \tag{6}$$

---

*) Иными словами, мы рассматриваем вектор $x$ в новых координатах как преобразованный вектор, отнесенный к старому базису.

т. е. элементы матрицы перехода $S$ являются *направляющими косинусами* и могут быть заданы таблицей 24.

Таблица 24

**Косинусы углов между ортами двух базисов**

| Орты новой системы | Орты старой системы | | | |
|---|---|---|---|---|
| | $e_1$ | $e_2$ | ... | $e_n$ |
| $\varepsilon_1$ | $s_{11}$ | $s_{21}$ | ... | $s_{n1}$ |
| $\varepsilon_2$ | $s_{12}$ | $s_{22}$ | ... | $s_{n2}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | ... | $\vdots$ |
| $\varepsilon_n$ | $s_{1n}$ | $s_{2n}$ | ... | $s_{nn}$ |

Подставляя выражение (1) в формулу (5′), в силу формул (5) получим:

$$(\varepsilon_j, \varepsilon_k) = \left(\sum_{i=1}^{n} s_{ij}e_i, \sum_{i=1}^{n} s_{ik}e_i\right) = \sum_{i=1}^{n} s_{ij}s_{ik} = \delta_{jk},$$

т. е. 1) *суммы парных произведений соответствующих направляющих косинусов различных координатных осей новой ортонормированной системы равны нулю* и 2) *сумма квадратов направляющих косинусов для каждой новой координатной оси равна единице*. Отсюда

$$S'S = E, \tag{7}$$

т. е. *матрица перехода от одного ортонормированного базиса к другому* ортогональна (подробнее об ортогональных матрицах см. § 6).

## § 6. Ортогональные матрицы

Определение. Действительная матрица $A$ называется *ортогональной*, если ее транспонированная матрица $A'$ совпадает с обратной $A^{-1}$, т. е.

$$A' = A^{-1} \tag{1}$$

или

$$AA' = A'A = E. \tag{2}$$

Ортогональная матрица имеет следующие свойства.

1. Строки (столбцы) ортогональной матрицы попарно ортогональны.

Действительно, если $A=[a_{ij}]$, то из равенства (2) имеем:

$$\sum_{k=1}^{n} a_{ik}a_{jk}=0 \quad \text{при} \quad i\neq j$$

и

$$\sum_{k=1}^{n} a_{ki}a_{kj}=0 \quad \text{при} \quad i\neq j.$$

2. Сумма квадратов элементов каждой строки (столбца) ортогональной матрицы равна 1.

Из равенства (2) при $i=j$ получаем:

$$\sum_{k=1}^{n} a_{ik}^2=\sum_{k=1}^{n} a_{ki}^2=1.$$

3. Определитель ортогональной матрицы равен $\pm 1$.

В самом деле, на основании равенства (2) имеем:

$$\det A \det A' = \det E.$$

Отсюда, так как $\det A' = \det A$ и $\det E = 1$, то

$$(\det A)^2 = 1$$

и, следовательно,

$$\det A = \pm 1.$$

4. Транспонированная и обратная матрицы ортогональной матрицы суть также ортогональные матрицы. Это свойство непосредственно вытекает из формул (1) и (2).

## § 7. Ортогонализация матриц

Пусть имеем матрицу с действительными элементами

$$A=\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}.$$

Столбцы матрицы $A$ будем рассматривать как векторы

$$\boldsymbol{a}^{(j)}=\begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{nj} \end{bmatrix} \quad (j=1,\ 2,\ \dots,\ n).$$

Следовательно, эту матрицу можно записать в таком виде:

$$A=\left[\boxed{\boldsymbol{a}^{(1)} \mid \dots \mid \boldsymbol{a}^{(n)}}\right].$$

Теорема 1. *Всякую действительную неособенную матрицу $A$ можно представить в виде произведения матрицы с ортогональными столбцами на верхнюю треугольную матрицу, т. е.*

$$A = RT,$$

*где $R$ — матрица с ортогональными столбцами и $T$ — верхняя треугольная матрица с единичной диагональю.*

Доказательство. Для простоты доказательство теоремы проведем для случая, когда порядок матрицы $n=3$. Однако рассуждения будут иметь общий характер. Пусть

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}.$$

Запишем эту матрицу в виде

$$A = [\boldsymbol{a}^{(1)}\ \boldsymbol{a}^{(2)}\ \boldsymbol{a}^{(3)}],$$

где $\boldsymbol{a}^{(j)} = \begin{bmatrix} a_{1j} \\ a_{2j} \\ a_{3j} \end{bmatrix}$ — векторы-столбцы.

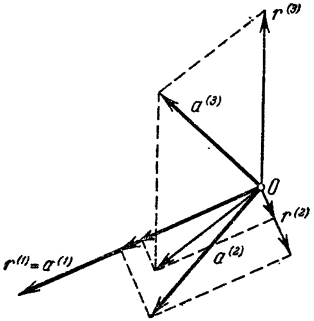


Рис. 51.

Так как матрица $A$ — неособенная, то векторы $\boldsymbol{a}^{(1)}$, $\boldsymbol{a}^{(2)}$, $\boldsymbol{a}^{(3)}$ линейно независимы.

Действительно, если бы эти векторы были линейно зависимы, то в $\det A$ один из столбцов являлся бы линейной комбинацией двух других и, следовательно, $\det A = 0$, что невозможно.

Будем искать матрицу $R$ также в виде

$$R = [\boldsymbol{r}^{(1)}\ \boldsymbol{r}^{(2)}\ \boldsymbol{r}^{(3)}],$$

где $\boldsymbol{r}^{(j)}$ $(j=1,\ 2,\ 3)$ — искомые ортогональные столбцы.

Положим

$$\boldsymbol{r}^{(1)} = \boldsymbol{a}^{(1)}. \tag{1}$$

Далее, вектор $\boldsymbol{a}^{(2)}$ раскладываем на составляющие $t_{12}\boldsymbol{r}^{(1)}$ и $\boldsymbol{r}^{(2)}$, из которых первая направлена по вектору $\boldsymbol{r}^{(1)}$, а вторая перпендикулярна (ортогональна) к нему (рис. 51), т. е.

$$\boldsymbol{a}^{(2)} = t_{12}\boldsymbol{r}^{(1)} + \boldsymbol{r}^{(2)}, \tag{2}$$

где

$$(\boldsymbol{r}^{(1)},\ \boldsymbol{r}^{(2)}) = 0. \tag{2'}$$

Аналогично вектор $\boldsymbol{a}^{(3)}$ раскладываем на три составляющие $t_{13}\boldsymbol{r}^{(1)}$, $t_{23}\boldsymbol{r}^{(2)}$ и $\boldsymbol{r}^{(3)}$, из которых первые две направлены соответственно по

векторам $r^{(1)}$ и $r^{(2)}$, а последняя перпендикулярна как к вектору $r^{(1)}$, так и к вектору $r^{(2)}$ (рис. 51), т. е.

$$a^{(3)} = t_{13}r^{(1)} + t_{23}r^{(2)} + r^{(3)}, \tag{3}$$

где

$$(r^{(1)}, r^{(3)}) = 0 \text{ и } (r^{(2)}, r^{(3)}) = 0. \tag{3'}$$

Из построения ясно, что векторы $r^{(1)}$, $r^{(2)}$ и $r^{(3)}$ будут взаимно перпендикулярны. Определим из системы (2) и (3) как векторы $r^{(2)}$ и $r^{(3)}$, так и коэффициенты $t_{ij}$. Умножая скалярно обе части уравнения (2) на $r^{(1)} = a^{(1)}$, в силу условия ортогональности (2′) получим:

$$(a^{(2)}, r^{(1)}) = t_{12}(r^{(1)}, r^{(1)}),$$

причем

$$(r^{(1)}, r^{(1)}) \neq 0.$$

Следовательно,

$$t_{12} = \frac{(a^{(2)}, r^{(1)})}{(r^{(1)}, r^{(1)})}$$

и

$$r^{(2)} = a^{(2)} - t_{12}r^{(1)}.$$

Заметим, что в силу неособенности матрицы $A$ вектор $r^{(1)} = a^{(1)} \neq 0$ и поэтому $(r^{(1)}, r^{(1)}) \neq 0$. Кроме того, $r^{(2)} \neq 0$, так как в противном случае векторы $a^{(1)}$ и $a^{(2)}$ были бы линейно зависимы.

Аналогично, умножая скалярно обе части уравнения (3) последовательно на $r^{(1)}$ и $r^{(2)}$, в силу условий ортогональности (2′) и (3′) получим:

$$(a^{(3)}, r^{(1)}) = t_{13}(r^{(1)}, r^{(1)});$$
$$(a^{(3)}, r^{(2)}) = t_{23}(r^{(2)}, r^{(2)}).$$

Отсюда, учитывая, что $(r^{(1)}, r^{(1)}) \neq 0$ и $(r^{(2)}, r^{(2)}) \neq 0$, будем иметь:

$$t_{13} = \frac{(a^{(3)}, r^{(1)})}{(r^{(1)}, r^{(1)})}, \quad t_{23} = \frac{(a^{(3)}, r^{(2)})}{(r^{(2)}, r^{(2)})}$$

и

$$r^{(3)} = a^{(3)} - t_{13}r^{(1)} - t_{23}r^{(2)}.$$

Легко проверить, что так построенные векторы $r^{(1)}$, $r^{(2)}$ и $r^{(3)}$ попарно ортогональны. Таким образом, окончательно имеем:

$$\left.\begin{aligned} a^{(1)} &= r^{(1)}, \\ a^{(2)} &= t_{12}r^{(1)} + r^{(2)}, \\ a^{(3)} &= t_{13}r^{(1)} + t_{23}r^{(2)} + r^{(3)}, \end{aligned}\right\} \tag{4}$$

где

$$t_{ij} = \frac{(a^{(j)}, r^{(i)})}{(r^{(i)}, r^{(i)})} \quad (i < j)$$

и

$$(r^{(i)}, r^{(j)}) = 0 \text{ при } i \neq j.$$

Очевидно, что система (4) эквивалентна матричному уравнению

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \cdot \begin{bmatrix} 1 & t_{12} & t_{13} \\ 0 & 1 & t_{23} \\ 0 & 0 & 1 \end{bmatrix}$$

или

$$A = RT, \tag{5}$$

где $R = [r_{ij}]$ — матрица с ортогональными столбцами, а $T = [t_{ij}]$ — верхняя треугольная матрица с единичной диагональю.

Пример. Ортогонализировать столбцы матрицы

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

Решение. Положим

$$r^{(1)} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} = a^{(1)}.$$

Тогда

$$t_{12} = \frac{(a^{(2)}, r^{(1)})}{(r^{(1)}, r^{(1)})} = \frac{1\cdot 0 + 2\cdot 1 + 0\cdot 2}{0^2 + 1^2 + 2^2} = 0{,}4.$$

Теперь найдем

$$r^{(2)} = a^{(2)} - t_{12} r^{(1)} = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} - 0{,}4 \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1{,}6 \\ -0{,}8 \end{bmatrix}.$$

Для определения $r^{(3)}$ вычислим $t_{13}$ и $t_{23}$. Имеем:

$$t_{13} = \frac{(a^{(3)}, r^{(1)})}{(r^{(1)}, r^{(1)})} = \frac{2\cdot 0 + 0\cdot 1 + 1\cdot 2}{5} = \frac{2}{5} = 0{,}4;$$

$$t_{23} = \frac{(a^{(3)}, r^{(2)})}{(r^{(2)}, r^{(2)})} = \frac{2\cdot 1 + 0\cdot 1{,}6 + 1\cdot(-0{,}8)}{1^2 + 1{,}6^2 + 0{,}8^2} = \frac{1{,}2}{4{,}2} \approx 0{,}3.$$

Отсюда

$$r^{(3)} = a^{(3)} - t_{13} r^{(1)} - t_{23} r^{(2)} = \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} - 0{,}4 \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} - 0{,}3 \begin{bmatrix} 1 \\ 1{,}6 \\ -0{,}8 \end{bmatrix} = \begin{bmatrix} 1{,}70 \\ -0{,}88 \\ 0{,}44 \end{bmatrix}.$$

Итак,

$$A = \begin{bmatrix} 0 & 1 & 1{,}7 \\ 1 & 1{,}6 & -0{,}88 \\ 2 & -0{,}8 & 0{,}44 \end{bmatrix} \begin{bmatrix} 1 & 0{,}4 & 0{,}4 \\ 0 & 1 & 0{,}3 \\ 0 & 0 & 1 \end{bmatrix},$$

причем векторы

$$r^{(1)}=\begin{bmatrix}0\\1\\2\end{bmatrix};\quad r^{(2)}=\begin{bmatrix}1\\1,6\\-0,8\end{bmatrix};\quad r^{(3)}=\begin{bmatrix}1,7\\-0,88\\0,44\end{bmatrix}$$

попарно ортогональны, в чем можно убедиться непосредственной проверкой.

В некоторых случаях выгоднее ортогонализировать не столбцы, а строки матрицы, рассматривая их как соответствующие векторы.

Пусть $A'$ — транспонированная матрица для данной матрицы $A$ — приведена к виду

$$A'=RT, \tag{6}$$

где $R$ — матрица с ортогональными столбцами и $T$ — верхняя треугольная матрица с единичной диагональю. Транспонируя равенство (6), получим:

$$A=T'R', \tag{7}$$

где $T'$ — нижняя треугольная матрица и $R'$ — матрица с ортогональными строками. Таким образом, указанный выше прием ортогонализации столбцов матрицы годится также и для ортогонализации строк, и мы имеем теорему.

Теорема 2. *Всякую действительную неособенную матрицу можно представить в виде произведения нижней треугольной матрицы с единичной диагональю и матрицы с ортогональными строками.*

Укажем еще один прием ортогонализации строк матрицы, иногда практически более удобный [5]. Пусть дана действительная неособенная матрица

$$A=\begin{bmatrix}a_{11} & a_{12} & \dots & a_{1n}\\ a_{21} & a_{22} & \dots & a_{2n}\\ \dots & \dots & \dots & \dots\\ a_{n1} & a_{n2} & \dots & a_{nn}\end{bmatrix}.$$

Из каждой $i$-й строки матрицы $A$, начиная со второй, вычтем ее первую строку, умноженную на некоторое число $\lambda_{i1}$ ($i=2, \dots, n$), зависящее от номера строки. В результате будем иметь преобразованную матрицу

$$A^{(1)}=\begin{bmatrix}a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)}\\ a_{21}^{(1)} & a_{22}^{(1)} & \dots & a_{2n}^{(1)}\\ \dots & \dots & \dots & \dots\\ a_{n1}^{(1)} & a_{n2}^{(1)} & \dots & a_{nn}^{(1)}\end{bmatrix},$$

где $a_{ij}^{(1)}=a_{ij}$ при $i=1$ и $a_{ij}^{(1)}=a_{ij}-\lambda_{i1}a_{1j}$ при $i\geqslant 2$.

Подберем множители $\lambda_{i1}$ так, чтобы первая строка матрицы $A^{(1)}$ была ортогональна ко всем остальным строкам этой матрицы. Имеем:

$$\sum_{j=1}^{n} a_{1j}^{(1)} a_{ij}^{(1)} = \sum_{j=1}^{n} a_{1j}(a_{ij} - \lambda_{i1} a_{1j}) = \sum_{j=1}^{n} a_{1j} a_{ij} - \lambda_{i1} \sum_{j=1}^{n} a_{1j}^2 = 0.$$

Отсюда

$$\lambda_{i1} = \frac{\sum_{j=1}^{n} a_{1j} a_{ij}}{\sum_{j=1}^{n} a_{1j}^2} \quad (i = 2, \ldots, n).$$

Над матрицей $A^{(1)}$ проделываем аналогичную операцию, а именно: первые две строки ее оставляем неизменными, а из каждой $i$-й строки, где $i \geqslant 3$, вычитаем вторую строку матрицы $A^{(1)}$, умноженную на число $\lambda_{i2}$ $(i = 3, \ldots, n)$. Получаем новую матрицу

$$A^{(2)} = \begin{bmatrix} a_{11}^{(2)} & a_{12}^{(2)} & \ldots & a_{1n}^{(2)} \\ a_{21}^{(2)} & a_{22}^{(2)} & \ldots & a_{2n}^{(2)} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1}^{(2)} & a_{n2}^{(2)} & \ldots & a_{nn}^{(2)} \end{bmatrix},$$

где $a_{ij}^{(2)} = a_{ij}^{(1)}$ при $i = 1, 2$ и $a_{ij}^{(2)} = a_{ij}^{(1)} - \lambda_{i2} a_{2j}^{(1)}$ при $i \geqslant 3$.

Так как первая строка матрицы $A^{(2)}$ совпадает с первой строкой матрицы $A^{(1)}$ и все остальные строки матрицы $A^{(2)}$ представляют собой линейные комбинации строк матрицы $A^{(1)}$, ортогональных первой строке матрицы $A^{(1)}$, то строки матрицы $A^{(2)}$ будут также ортогональными к ее первой строке. Выберем множители $\lambda_{i2}$ так, чтобы строки матрицы $A^{(2)}$, начиная с третьей, были ортогональны ко второй ее строке. Получаем:

$$\sum_{j=1}^{n} a_{2j}^{(2)} a_{ij}^{(2)} = \sum_{j=1}^{n} a_{2j}^{(1)} (a_{ij}^{(1)} - \lambda_{i2} a_{2j}^{(1)}) = \sum_{j=1}^{n} a_{2j}^{(1)} a_{ij}^{(1)} - \lambda_{i2} \sum_{j=1}^{n} [a_{2j}^{(1)}]^2 = 0.$$

Отсюда

$$\lambda_{i2} = \frac{\sum_{j=1}^{n} a_{2j}^{(1)} a_{ij}^{(1)}}{\sum_{j=1}^{n} [a_{2j}^{(1)}]^2} \quad (i = 3, \ldots, n). \qquad \text{(A)}$$

Указанный выше процесс продолжаем до тех пор, пока не получится матрица

$$A^{(n-1)}=\begin{bmatrix} a_{11}^{(n-1)} & a_{12}^{(n-1)} & \dots & a_{1n}^{(n-1)} \\ a_{21}^{(n-1)} & a_{22}^{(n-1)} & \dots & a_{2n}^{(n-1)} \\ \dots & \dots & \dots & \dots \\ a_{n1}^{(n-1)} & a_{n2}^{(n-1)} & \dots & a_{nn}^{(n-1)} \end{bmatrix},$$

все строки которой попарно ортогональны:

$$\sum_{j=1}^{n} a_{kj}^{(n-1)} a_{ij}^{(n-1)}=0 \text{ при } k\neq i.$$

Матрица $A^{(n-1)}=\tilde{R}$ с ортогональными строками получилась из данной матрицы $A$ в результате цепи элементарных преобразований. Поэтому справедливо равенство

$$\tilde{R}=\Lambda A, \tag{8}$$

где $\Lambda$ — неособенная матрица, которая в нашем случае является нижней треугольной матрицей.

Матрицу $\Lambda$ легко восстановить, проделав над единичной матрицей $E$ все элементарные преобразования, совершенные над матрицей $A$. Из формулы (8) имеем окончательно:

$$A=\tilde{T}\tilde{R},$$

где $\tilde{T}=\Lambda^{-1}$ — нижняя треугольная матрица.

Укажем некоторые свойства матриц с ортогональными рядами.

Лемма. *Если столбцы действительной матрицы составляют ортогональную систему векторов, то произведение транспонированной матрицы на саму матрицу равно диагональной матрице.*

Доказательство. Пусть $A=[a_{ij}]$ — данная матрица. Требуется доказать, что $A'A=D$, где $A'=[a_{ji}]$ — транспонированная матрица $A$ и

$$D=\begin{bmatrix} d_{11} & 0 & \dots & 0 \\ 0 & d_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & d_{nn} \end{bmatrix} \text{ — диагональная матрица.}$$

Полагая $D=[d_{ij}]$, согласно правилу умножения матриц имеем:

$$d_{ij}=\sum_{k=1}^{n} a_{ki}a_{kj}.$$

Отсюда, так как $a_{ki}$ — координаты $i$-го вектора $a^{(i)}$ и $a_{kj}$ — координаты $j$-го вектора $a^{(j)}$, получаем:

$$d_{ij}=\sum_{k=1}^{n} a_{ki}a_{kj}=(a^{(i)},\ a^{(j)})=0, \quad \text{если } i\neq j.$$

Следовательно, $D=[d_{ij}]$ — диагональная матрица.

Следствие. Произведение действительной матрицы с ортогональными строками на транспонированную матрицу равно диагональной матрице, т. е. $AA'=D$.

Теорема 3. *Всякая неособенная действительная матрица $A$ с ортогональными столбцами представляет собой ортогональную матрицу, умноженную справа на диагональную матрицу.*

Доказательство. В силу леммы имеем:

$$A'A=D, \tag{9}$$

где $D=[d_{ij}]$ — диагональная матрица. Если $A=[a_{ij}]$, то, очевидно,

$$d_{ii}=\sum_{k=1}^{n} a_{ki}^2>0.$$

Пусть

$$\rho_i=\sqrt{d_{ii}}>0 \quad (i=1,\ 2,\ \ldots,\ n)$$

и

$$d=\begin{bmatrix} \rho_1 & 0 & \ldots & 0 \\ 0 & \rho_2 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & \rho_n \end{bmatrix}.$$

Очевидно, что $D=d^2$. Из формулы (9) имеем $A'A=d^2$, откуда

$$d^{-1}A'Ad^{-1}=E.$$

Так как

$$(d^{-1})'=d^{-1}, \text{ то } (Ad^{-1})'(Ad^{-1})=E.$$

Следовательно, матрица $Ad^{-1}=U$ ортогональна и, значит,

$$A=Ud, \tag{10}$$

что и требовалось доказать.

Следствие. Неособенную действительную матрицу с ортогональными строками можно представить в виде произведения диагональной матрицы на ортогональную матрицу.

Действительно, пусть $A$ — матрица с ортогональными строками; тогда $A'$ — матрица с ортогональными столбцами. В силу формулы (10) имеем $A'=Ud$, где $U$ — ортогональная матрица и $d$ — диагональная матрица, которая может быть определена из соотношения

$$AA'=d^2.$$

Отсюда получаем:

$$A=(A')'=d'U'=dU',$$

где $U'$ — также ортогональная матрица.

Замечание. Чтобы данную действительную неособенную матрицу $A$ с ортогональными столбцами (строками) преобразовать в ортогональную, достаточно нормировать ее столбцы (или соответственно строки), т. е. элемент каждого столбца (строки) разделить на корень квадратный из суммы квадратов элементов этого столбца (строки). Например, если $A=[a_{ij}]$ есть матрица с ортогональными столбцами, то матрица

$$\tilde{A}=[\tilde{a}_{ij}],$$

где $\tilde{a}_{ij}=\dfrac{a_{ij}}{\sqrt{\sum\limits_{k=1}^{n} a_{kj}^2}}$ $(i,\ j=1,\ 2,\ \ldots,\ n)$, есть ортогональная матрица.

## § 8. Применение методов ортогонализации к решению систем линейных уравнений

### А. Первый способ (ортогонализация столбцов)

Пусть имеем систему линейных уравнений

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

с действительной неособенной матрицей $A$. Ортогонализируя столбцы матрицы $A$, получим матрицу $R$, причем $A=RT$, где $T$ — верхняя треугольная матрица. Имеем:

$$RTx=b. \tag{2}$$

Умножая слева на $R'$ обе части равенства (2), получим:

$$R'RTx=R'b. \tag{3}$$

Но, как известно, $R'R=D$, где $D$ — диагональная матрица. Вводя обозначение $R'b=\beta$, будем иметь:

$$DTx=\beta,$$

откуда

$$x=(DT)^{-1}\beta=T^{-1}D^{-1}\beta. \tag{4}$$

Матрица $D^{-1}$, обратная диагональной, находится легко, а именно, если

$$D=\begin{bmatrix} d_{11} & 0 & \ldots & 0 \\ 0 & d_{22} & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & d_{nn} \end{bmatrix},$$

то

$$D^{-1}=\begin{bmatrix} d_{11}^{-1} & 0 & \ldots & 0 \\ 0 & d_{22}^{-1} & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & d_{nn}^{-1} \end{bmatrix}.$$

Относительно просто находится также обратная матрица $T^{-1}$ треугольной матрицы $T$.

Пример 1. Методом ортогонализации столбцов решить систему уравнений

$$\left.\begin{aligned} 0{,}4x_1+0{,}3x_2-0{,}2x_3&=2;\\ 0{,}6x_1-0{,}5x_2+0{,}3x_3&=2{,}5;\\ 0{,}3x_1+0{,}2x_2+0{,}5x_3&=11.\end{aligned}\right\}$$

Решение. Представим матрицу $A$ данной системы в виде произведения матрицы $R$ с ортогональными столбцами на треугольную матрицу с единичной диагональю:

$$A=RT=\begin{bmatrix} r_{11} & r_{12} & r_{13}\\ r_{21} & r_{22} & r_{23}\\ r_{31} & r_{32} & r_{33}\end{bmatrix}\begin{bmatrix} 1 & \lambda_{12} & \lambda_{13}\\ 0 & 1 & \lambda_{23}\\ 0 & 0 & 1\end{bmatrix}.$$

Полагаем:

$$\boldsymbol{r}^{(1)}=\boldsymbol{a}^{(1)};\quad \boldsymbol{r}^{(2)}=\boldsymbol{a}^{(2)}-\lambda_{12}\boldsymbol{r}^{(1)};\quad \boldsymbol{r}^{(3)}=\boldsymbol{a}^{(3)}-\lambda_{13}\boldsymbol{r}^{(1)}-\lambda_{23}\boldsymbol{r}^{(2)}.$$

Имеем:

$$\boldsymbol{r}^{(1)}=\begin{bmatrix}0{,}4\\0{,}6\\0{,}3\end{bmatrix}.$$

По формулам (4) предыдущего параграфа находим:

$$\lambda_{12}=\frac{(\boldsymbol{a}^{(2)},\boldsymbol{r}^{(1)})}{(\boldsymbol{r}^{(1)},\boldsymbol{r}^{(1)})}=\frac{0{,}12-0{,}3+0{,}06}{0{,}16+0{,}36+0{,}09}=-\frac{0{,}12}{0{,}61}=-0{,}1967;$$

$$\boldsymbol{r}^{(2)}=\begin{bmatrix}0{,}3\\-0{,}5\\0{,}2\end{bmatrix}+0{,}1967\begin{bmatrix}0{,}4\\0{,}6\\0{,}3\end{bmatrix}=\begin{bmatrix}0{,}3787\\-0{,}3820\\0{,}2590\end{bmatrix}.$$

Контроль:

$$(\boldsymbol{r}^{(1)},\boldsymbol{r}^{(2)})=\begin{bmatrix}0{,}4\\0{,}6\\0{,}3\end{bmatrix}'\begin{bmatrix}0{,}3787\\-0{,}3820\\0{,}2590\end{bmatrix}=\left.\begin{matrix}0{,}1515\\-0{,}2292\\0{,}0777\end{matrix}\right\}=0;$$

$$\lambda_{13}=\frac{(\boldsymbol{a}^{(3)},\boldsymbol{r}^{(1)})}{(\boldsymbol{r}^{(1)},\boldsymbol{r}^{(1)})}=\frac{-0{,}08+0{,}18+0{,}15}{0{,}61}=\frac{0{,}25}{0{,}61}=0{,}4098;$$

$$\lambda_{23}=\frac{(\boldsymbol{a}^{(3)},\boldsymbol{r}^{(2)})}{(\boldsymbol{r}^{(2)},\boldsymbol{r}^{(2)})}=-\frac{0{,}07574-0{,}11460+0{,}12950}{0{,}35}=-0{,}1714;$$

$$\boldsymbol{r}^{(3)}=\begin{bmatrix}-0{,}2\\0{,}3\\0{,}5\end{bmatrix}-0{,}4098\begin{bmatrix}0{,}4\\0{,}6\\0{,}3\end{bmatrix}+0{,}1714\begin{bmatrix}0{,}3787\\-0{,}3820\\0{,}2590\end{bmatrix}=\begin{bmatrix}-0{,}2990\\-0{,}0114\\0{,}4215\end{bmatrix}.$$

Контроль:

$$(\boldsymbol{r}^{(1)},\boldsymbol{r}^{(3)})=(\boldsymbol{r}^{(2)},\boldsymbol{r}^{(3)})=0.$$

Таким образом,

$$A=\underbrace{\begin{bmatrix}0{,}4 & 0{,}3787 & -0{,}2990\\ 0{,}6 & -0{,}3820 & -0{,}0114\\ 0{,}3 & 0{,}2590 & 0{,}4215\end{bmatrix}}_{R}\underbrace{\begin{bmatrix}1 & -0{,}1967 & 0{,}4098\\ 0 & 1 & -0{,}1714\\ 0 & 0 & 1\end{bmatrix}}_{T}.$$

По формуле (4) имеем:

$$\boldsymbol{x}=T^{-1}D^{-1}R'\boldsymbol{b},$$

где $D=R'R$ — диагональная матрица и

$$\boldsymbol{b}=\begin{bmatrix}2\\ 2{,}5\\ 11\end{bmatrix}.$$

Для матрицы $D$ и ее обратной матрицы $D^{-1}$ получаем такие значения:

$$D=\begin{bmatrix}0{,}61 & 0 & 0\\ 0 & 0{,}35 & 0\\ 0 & 0 & 0{,}2672\end{bmatrix} \quad \text{и} \quad D^{-1}=\begin{bmatrix}1{,}64 & 0 & 0\\ 0 & 2{,}81 & 0\\ 0 & 0 & 3{,}75\end{bmatrix}.$$

Далее,

$$R'\boldsymbol{b}=\begin{bmatrix}0{,}4 & 0{,}6 & 0{,}3\\ 0{,}3787 & -0{,}3820 & 0{,}2590\\ -0{,}2990 & -0{,}0114 & 0{,}4215\end{bmatrix}\begin{bmatrix}2\\ 2{,}5\\ 11\end{bmatrix}=\begin{bmatrix}5{,}6\\ 2{,}67\\ 4{,}08\end{bmatrix}.$$

Наконец, обычным приемом подсчитываем:

$$T^{-1}=\begin{bmatrix}1 & 0{,}1967 & -0{,}3761\\ 0 & 1 & 0{,}1714\\ 0 & 0 & 1\end{bmatrix}.$$

В итоге получим:

$$\boldsymbol{x}=\begin{bmatrix}1 & 0{,}1967 & -0{,}3761\\ 0 & 1 & 0{,}1714\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}1{,}64 & 0 & 0\\ 0 & 2{,}81 & 0\\ 0 & 0 & 3{,}75\end{bmatrix}\begin{bmatrix}5{,}6\\ 2{,}67\\ 4{,}08\end{bmatrix}=\begin{bmatrix}5{,}0238\\ 10{,}0475\\ 15{,}0087\end{bmatrix}.$$

Следовательно,

$$x_1=5{,}0238; \qquad x_2=10{,}0475; \qquad x_3=15{,}0087;$$

точные значения корней: $x_1=5$; $x_2=10$; $x_3=15$.

### Б. Второй способ (ортогонализация строк)

Пусть дана система

$$A\boldsymbol{x}=\boldsymbol{b}, \tag{5}$$

где $\det A\neq 0$.

Преобразуем строки системы (5) с помощью приема, указанного в предыдущем параграфе, так, чтобы матрица $A$ перешла в матрицу $R$

с ортогональными строками. При этом вектор $b$ перейдет в какой-то вектор $\beta$. В результате получим эквивалентную систему

$$Rx=\beta. \tag{6}$$

Следовательно,

$$x=R^{-1}\beta. \tag{7}$$

Как известно, $RR'=D=d^2$, где $d$ — диагональная матрица, и $R=dU$, где $U$ — ортогональная матрица. Поэтому

$$R^{-1}=(dU)^{-1}=U^{-1}d^{-1}=U'd'd^{-2}=(dU)'d^{-2}=R'd^{-2}=R'D^{-1}.$$

Таким образом, на основании формулы (7) окончательно имеем:

$$x=R'D^{-1}\beta, \tag{8}$$

где

$$D=RR'. \tag{9}$$

Используя формулу (8), можно избежать наиболее трудоемкого процесса нахождения обратной матрицы для недиагональной матрицы. Наличие матрицы $D^{-1}$ не вносит усложнения, так как $D$ — диагональная матрица. Формула (9), нужная по существу, может быть использована также и для контроля.

Пример 2. Методом ортогонализации строк решить систему уравнений

$$\left.\begin{aligned}3{,}00x_1+0{,}15x_2-0{,}09x_3&=6{,}00;\\0{,}08x_1+4{,}00x_2-0{,}16x_3&=12{,}00;\\0{,}05x_1+0{,}30x_2+5{,}00x_3&=20{,}00.\end{aligned}\right\} \tag{I}$$

Решение. По формулам предыдущего параграфа определяем множители:

$$\lambda_{21}=\frac{3{,}00\cdot0{,}08+0{,}15\cdot4{,}00+(-0{,}09)\cdot(-0{,}16)}{3{,}00^2+0{,}15^2+0{,}09^2}=\frac{0{,}8544}{9{,}0306}=0{,}0946;$$

$$\lambda_{31}=\frac{3{,}00\cdot0{,}05+0{,}15\cdot0{,}30-0{,}09\cdot5{,}00}{3{,}00^2+0{,}15^2+0{,}09^2}=-\frac{0{,}2550}{9{,}0306}=-0{,}0282.$$

Сохраняя первое уравнение системы (I), из каждого следующего вычитаем первое, умноженное на соответствующие множители $\lambda_{i1}$ ($i=2,\ 3$):

$$\left.\begin{aligned}3{,}00x_1+0{,}15x_2-0{,}09x_3&=6{,}00;\\-0{,}2038x_1+3{,}9858x_2-0{,}1685x_3&=11{,}4324;\\0{,}1346x_1+0{,}3042x_2+4{,}9975x_3&=20{,}1692.\end{aligned}\right\} \tag{II}$$

Для системы (II) определяем множитель

$$\lambda_{32}=\frac{-0{,}2038\cdot0{,}1346+3{,}9858\cdot0{,}3042-0{,}1685\cdot4{,}9975}{0{,}2038^2+3{,}9858^2+0{,}1685^2}=\frac{0{,}3430}{15{,}9565}=0{,}0215.$$

Сохраняя два первых уравнения системы (II), из ее третьего уравнения вычитаем второе, умноженное на множитель $\lambda_{32}$:

$$\left.\begin{aligned} 3{,}00x_1+0{,}15x_2-0{,}09x_3&=6{,}00;\\ -0{,}2038x_1+3{,}9858x_2-0{,}1685x_3&=11{,}4324;\\ 0{,}1390x_1+0{,}2185x_2+5{,}0011x_3&=19{,}9234.\end{aligned}\right\}\qquad\text{(III)}$$

Матрица

$$R=\begin{bmatrix} 3{,}00 & 0{,}15 & -0{,}09\\ -0{,}2038 & 3{,}9858 & -0{,}1685\\ 0{,}1390 & 0{,}2185 & 5{,}0011\end{bmatrix}$$

имеет ортогональные строки. Для контроля составляем матрицу

$$D=RR'=\begin{bmatrix} 9{,}0306 & 0{,}0017 & -0{,}0002\\ 0{,}0017 & 15{,}9565 & -0{,}0018\\ -0{,}0002 & -0{,}0018 & 25{,}0780\end{bmatrix}\approx\begin{bmatrix} 9{,}0306 & 0 & 0\\ 0 & 15{,}9565 & 0\\ 0 & 0 & 25{,}0780\end{bmatrix}.$$

Применяя формулу (8), получим:

$$x=R'D^{-1}\beta=\begin{bmatrix} 3{,}00 & -0{,}2038 & 0{,}1390\\ 0{,}15 & 3{,}9858 & 0{,}2185\\ -0{,}09 & -0{,}1685 & 5{,}0011\end{bmatrix}\times$$

$$\times\begin{bmatrix} 0{,}1107 & 0 & 0\\ 0 & 0{,}0626 & 0\\ 0 & 0 & 0{,}0399\end{bmatrix}\begin{bmatrix} 6{,}00\\ 11{,}4324\\ 19{,}9234\end{bmatrix}=\begin{bmatrix} 1{,}957\\ 3{,}126\\ 3{,}803\end{bmatrix}.$$

Следовательно,

$$x_1=1{,}957;\qquad x_2=3{,}126;\qquad x_3=3{,}803.$$

В. Третий способ (метод ортогональных матриц)

Пусть линейная система приведена к виду

$$Rx=\beta,\qquad(10)$$

где $R=[r_{ij}]$ — неособенная матрица с ортогональными строками и

$$\beta=\begin{bmatrix}\beta_i\\ \cdot\\ \cdot\\ \cdot\\ \beta_n\end{bmatrix}\text{ — вектор свободных членов.}$$

Умножая каждое уравнение системы (10) на нормирующий множитель

$$\mu_i=\frac{1}{\sqrt{\sum\limits_{j=1}^{n} r_{ij}^2}}\qquad(i=1,\ 2,\ \ldots,\ n),$$

получим систему

$$\tilde{R}x = \tilde{\beta}, \tag{11}$$

где $\tilde{R} = [\mu_i r_{ij}]$ — ортогональная матрица и

$$\tilde{\beta} = \begin{bmatrix} \mu_1\beta_1 \\ \mu_2\beta_2 \\ \cdots \\ \mu_n\beta_n \end{bmatrix}$$ — новый вектор свободных членов.

Из уравнения (11) будем иметь:

$$x = \tilde{R}^{-1}\tilde{\beta} = \tilde{R}'\tilde{\beta}. \tag{12}$$

## § 9. Пространство решений однородной системы

Рассмотрим однородную линейную систему

$$\left.\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = 0; \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = 0; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = 0 \end{array}\right\} \tag{1}$$

или, короче,

$$Ax = 0, \tag{1'}$$

где $A = [a_{ij}]$ и $x = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}$ — искомый вектор.

Если $\det A \neq 0$, то в силу формул Крамера система (1) имеет единственное нулевое решение $x = 0$.

Пусть $\det A = 0$. В этом случае система (1) имеет бесконечное множество решений (в том числе и ненулевые).

Из формулы (1′) вытекает: 1) если $x$ — решение уравнения (1′), то $cx$, где $c$ — произвольная постоянная, есть также решение этого уравнения; 2) если $x^{(1)}$ и $x^{(2)}$ — решения уравнения (1′), то сумма $x^{(1)} + x^{(2)}$ есть также решение этого уравнения.

Действительно, если

$$Ax = 0,$$

то

$$A(cx) = cAx = 0.$$

Следовательно, $cx$ — решение однородной системы (1).

Аналогично, если

$$Ax^{(1)} = 0 \quad \text{и} \quad Ax^{(2)} = 0,$$

то

$$A(x^{(1)} + x^{(2)}) = Ax^{(1)} + Ax^{(2)} = 0 + 0 = 0,$$

т. е. $x^{(1)} + x^{(2)}$ — также решение однородной системы (1).

Отсюда получаем, что *любая линейная комбинация решений однородной системы* (1) *есть также решение этой системы.* Следовательно, совокупность всех решений однородной системы (1) образует векторное пространство, которое называется *пространством решений.* Справедлива теорема.

Теорема. *Если $n$ — число неизвестных однородной системы* (1) *и $r$ есть ранг ее матрицы $A$, то размерность пространства решений равна $k=n-r$.*

Базис пространства решений называется *фундаментальной системой* решений. Если для системы (1) известна фундаментальная система решений

$$\begin{aligned} x^{(1)} &= (x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}), \\ x^{(2)} &= (x_1^{(2)}, x_2^{(2)}, \ldots, x_n^{(2)}), \\ &\ldots\ldots\ldots\ldots \\ x^{(k)} &= (x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}), \end{aligned}$$

то все ее решения содержатся в формуле

$$x = c_1 x^{(1)} + c_2 x^{(2)} + \ldots + c_k x^{(k)} \tag{2}$$

или, более подробно,

$$\left.\begin{aligned} x_1 &= c_1 x_1^{(1)} + c_2 x_1^{(2)} + \ldots + c_k x_1^{(k)}, \\ x_2 &= c_1 x_2^{(1)} + c_2 x_2^{(2)} + \ldots + c_k x_2^{(k)}, \\ &\ldots\ldots\ldots\ldots \\ x_n &= c_1 x_n^{(1)} + c_2 x_n^{(2)} + \ldots + c_k x_n^{(k)}, \end{aligned}\right\}$$

где $c_1, c_2, \ldots, c_k$ — произвольные постоянные числа.

Для нахождения фундаментальной системы решений в матрице $A$ выделяют минор $r$-го порядка $\delta_r$, отличный от нуля. Пусть

$$\delta_r = \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1r} \\ a_{21} & a_{22} & \ldots & a_{2r} \\ \ldots & \ldots & \ldots & \ldots \\ a_{r1} & a_{r2} & \ldots & a_{rr} \end{vmatrix} \neq 0.$$

Этого всегда можно добиться путем перестановки уравнений системы (1) и изменения нумерации ее неизвестных. Тогда легко доказать, что уравнения системы (1), начиная с $(r+1)$-го, являются следствиями первых $r$ уравнений этой системы, т. е. они удовлетворяются, если будут удовлетворены $r$ первых уравнений системы (1). Поэтому достаточно рассмотреть подсистему

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1r}x_r &= -a_{1,r+1}x_{r+1} - \ldots - a_{1n}x_n, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2r}x_r &= -a_{2,r+1}x_{r+1} - \ldots - a_{2n}x_n, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{r1}x_1 + a_{r2}x_2 + \ldots + a_{rr}x_r &= -a_{r,r+1}x_{r+1} - \ldots - a_{rn}x_n, \end{aligned}\right\} \tag{3}$$

определитель которой $\delta_r$ отличен от нуля.

В системе (3) значения неизвестных

$$x_{r+1}=c_1,\quad x_{r+2}=c_2,\ \ldots,\ x_n=c_{n-r}=c_k$$

можно считать произвольными. Разрешив систему (3) относительно неизвестных $x_1, x_2, \ldots, x_r$, получим:

$$\left.\begin{aligned} x_1&=\alpha_{11}c_1+\alpha_{12}c_2+\ldots+\alpha_{1k}c_k,\\ x_2&=\alpha_{21}c_1+\alpha_{22}c_2+\ldots+\alpha_{2k}c_k,\\ &\ldots\ldots\ldots\ldots\ldots\ldots\\ x_r&=\alpha_{r1}c_1+\alpha_{r2}c_2+\ldots+\alpha_{rk}c_k,\end{aligned}\right\}\tag{4}$$

где $\alpha_{ij}$ $(i=1,\ 2,\ \ldots,\ r;\ j=1,\ 2,\ \ldots,\ k)$ — вполне определенные постоянные.

Кроме того,

$$\left.\begin{aligned} x_{r+1}&=c_1,\\ x_{r+2}&=\quad c_2,\\ &\ldots\ldots\ldots\\ x_n&=\qquad c_k.\end{aligned}\right\}\tag{4'}$$

Формулы (4) и (4′) дают полную систему решений системы (1). За фундаментальную систему решений можно принять:

$$x^{(1)}=\begin{bmatrix}\alpha_{11}\\ \cdot\\ \cdot\\ \cdot\\ \alpha_{r1}\\ 1\\ 0\\ 0\\ \vdots\\ 0\end{bmatrix},\quad x^{(2)}=\begin{bmatrix}\alpha_{12}\\ \cdot\\ \cdot\\ \cdot\\ \alpha_{r2}\\ 0\\ 1\\ 0\\ \vdots\\ 0\end{bmatrix},\ \ldots,\quad x^{(k)}=\begin{bmatrix}\alpha_{1k}\\ \cdot\\ \\ \alpha_{rk}\\ 0\\ 0\\ 0\\ \vdots\\ 1\end{bmatrix}$$

Эти последние решения могут быть найдены непосредственно из системы (3), если в ней последовательно полагать:

$$x_{r+1}=1,\ x_{r+2}=\ldots=x_n=0;$$
$$x_{r+1}=0,\ x_{r+2}=1,\ x_{r+3}=\ldots=x_n=0;$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$x_{r+1}=\ldots=x_{n-1}=0,\ x_n=1.$$

Пример. Найти фундаментальную систему решений системы однородных уравнений

$$\left.\begin{aligned} x_1-\ x_2+5x_3-\ x_4&=0,\\ x_1+\ x_2-2x_3+3x_4&=0,\\ 3x_1-\ x_2+8x_3+\ x_4&=0,\\ x_1+3x_2-9x_3+7x_4&=0.\end{aligned}\right\}\tag{5}$$

Решение. Ранг матрицы системы (5) $r=2$, причем

$$\delta=\begin{vmatrix}1 & -1\\ 1 & 1\end{vmatrix}=2\neq 0.$$

Поэтому два последних уравнения системы (5) являются следствием двух первых. Решаем подсистему

$$\left.\begin{aligned}x_1-x_2&=-5x_3+x_4,\\ x_1+x_2&=2x_3-3x_4.\end{aligned}\right\}$$

Полагая сперва $x_3=1$; $x_4=0$, а затем $x_3=0$ и $x_4=1$, получим два линейно независимых решения

$$x^{(1)}=\left(-\frac{3}{2},\ \frac{7}{2},\ 1,\ 0\right);$$

$$x^{(2)}=(-1,\ -2,\ 0,\ 1),$$

которые образуют фундаментальную систему решений системы (5).

Векторы $x^{(1)}$ и $x^{(2)}$ образуют базис пространства решений данной системы, и все ее решения определяются формулами

$$\left.\begin{aligned}x_1&=-3c_1-c_2,\\ x_2&=7c_1-2c_2,\\ x_3&=2c_1,\\ x_4&=c_2,\end{aligned}\right\}$$

где $c_1$ и $c_2$ — произвольные постоянные (здесь для удобства первая постоянная взята в виде $2c_1$).

## § 10. Линейные преобразования переменных

Пусть $x_1, x_2, \ldots, x_n$ — некоторая система переменных величин, а $y_1, y_2, \ldots, y_n$ — другая система переменных величин, связанная с первой следующими соотношениями:

$$\left.\begin{aligned}y_1&=f_1(x_1, x_2, \ldots, x_n),\\ y_2&=f_2(x_1, x_2, \ldots, x_n),\\ &\ldots\ldots\ldots\ldots\\ y_n&=f_n(x_1, x_2, \ldots, x_n),\end{aligned}\right\}\qquad(1)$$

где $f_1, f_2, \ldots, f_n$ — заданные функции.

Переход от системы $x_1, x_2, \ldots, x_n$ к системе $y_1, y_2, \ldots, y_n$ будем называть *преобразованием*.

Определение. Преобразование (1) называется *линейным*, если новые переменные $y_1, y_2, \ldots, y_n$ являются линейными однородными

функциями старых переменных $x_1, x_2, \ldots, x_n$, т. е.

$$\left.\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n, \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n &= a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n, \end{aligned}\right\} \tag{2}$$

где $a_{ij}$ $(i=1, 2, \ldots, n;\ j=1, 2, \ldots, n)$ — постоянные.

Линейное преобразование (2) однозначно определяется своей квадратной матрицей коэффициентов (*матрицей преобразования*)

$$A = \begin{bmatrix} a_{11} & a_{12} \ldots a_{1n} \\ a_{21} & a_{22} \ldots a_{2n} \\ \ldots & \ldots\ldots \\ a_{n1} & a_{n2} \ldots a_{nn} \end{bmatrix}.$$

Обратно, имея квадратную матрицу $A = [a_{ij}]$, можно построить линейное преобразование, для которого эта матрица служит матрицей коэффициентов. Таким образом, между линейными преобразованиями и квадратными матрицами существует *взаимнооднозначное* соответствие.

Системы переменных $x_1, x_2, \ldots, x_n$ и $y_1, y_2, \ldots, y_n$ можно рассматривать как векторы-столбцы

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \quad \text{и} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}$$

одного и того же векторного пространства $E_n$. Поэтому преобразование (2) отображает пространство $E_n$ на самого себя или на свою правильную часть.

Пример 1. Преобразование

$$\left.\begin{aligned} y_1 &= x_1 + x_2, \\ y_2 &= x_1 + x_2 \end{aligned}\right\}$$

переводит пространство $E_2$ в подпространство $y_1 = y_2$ размерности 1.

Соотношения (2) эквивалентны одному матричному соотношению

$$y = Ax. \tag{3}$$

Действительно, по правилу умножения матриц мы имеем:

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \\ \ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n \end{bmatrix}.$$

Отсюда на основании понятия о равенстве матриц получаем формулы (2). Согласно формуле (3) матрицу $A$ можно рассматривать как *оператор* линейного преобразования.

Линейное преобразование, как легко убедиться непосредственно, обладает двумя основными свойствами:

1) постоянный множитель можно выносить за знак оператора линейного преобразования, т. е.

$$A(\alpha x) = \alpha A x;$$

2) оператор линейного преобразования от суммы нескольких векторов равен сумме операторов от этих векторов, т. е.

$$A(x+z) = Ax + Az.$$

Как следствие имеем:

$$A(\alpha x + \beta z) = \alpha A x + \beta A z$$

($x$, $z$ — векторы, $\alpha$ и $\beta$ — скаляры).

Пример 2. Пусть на плоскости $Ox_1x_2$ каждому вектору

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

ставится в соответствие вектор

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix},$$

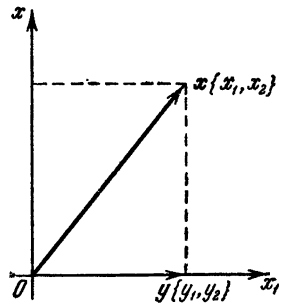


Рис. 52.

являющийся проекцией вектора $x$ на ось $Ox_1$ (*преобразование проектирования*) (рис. 52). Показать, что данное преобразование — линейное, и найти матрицу преобразования.

Решение. Очевидно, имеем:

$$\left.\begin{aligned} y_1 &= x_1, \\ y_2 &= 0, \end{aligned}\right\}$$

и, следовательно, преобразование проектирования является линейным. Матрица преобразования есть

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

Выясним смысл элементов $a_{ij}$ матрицы преобразования $A$. Рассмотрим единичные векторы (орты), направленные по осям координат $Ox_1$, $Ox_2$, ..., $Ox_n$:

$$e_1 = \begin{bmatrix} 1 \\ 0 \\ \cdot \\ \cdot \\ 0 \end{bmatrix}, \quad e_2 = \begin{bmatrix} 0 \\ 1 \\ \cdot \\ \cdot \\ 0 \end{bmatrix}, \quad \ldots, \quad e_n = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ 1 \end{bmatrix}.$$

Применяя преобразование $A$ к $e_j$, будем иметь:

$$Ae_j = \begin{bmatrix} a_{11} & a_{12} \ldots a_{1n} \\ a_{21} & a_{22} \ldots a_{2n} \\ \cdot & \cdot \quad \cdot \quad \cdot \\ a_{n1} & a_{n2} \ldots a_{nn} \end{bmatrix} \times \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \cdot \\ \cdot \\ \cdot \\ a_{nj} \end{bmatrix} \quad (j = 1, 2, \ldots, n).$$

Таким образом, $a_{ij}$ представляет собой $i$-ю координату преобразованного $j$-го единичного вектора.

Пример 3. Пусть на плоскости $Ox_1x_2$ каждый радиус-вектор $x$ заменяется той же длины радиусом-вектором $y$, повернутым относительно первого на угол $\alpha$ (*преобразование вращения*) (рис. 53).

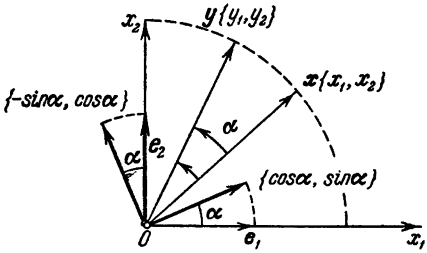


Рис. 53.

Показать, что данное преобразование—линейное, и найти матрицу преобразования.

Решение. С вектором $y$ свяжем координатную систему $Oy_1y_2$, которая повернута относительно координатной системы $Ox_1x_2$ на угол $\alpha$. Так как координаты вектора $y$ в системе $Oy_1y_2$, очевидно, есть $x_1$ и $x_2$, то координаты этого вектора в старой системе $Ox_1x_2$ по известным формулам аналитической геометрии выразятся следующим образом:

$$\left.\begin{aligned} y_1 &= x_1 \cos\alpha - x_2 \sin\alpha, \\ y_2 &= x_1 \sin\alpha + x_2 \cos\alpha. \end{aligned}\right\} \tag{4}$$

Таким образом, преобразование вращения есть линейное, и его матрица имеет вид

$$A = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix}.$$

Можно иначе решить задачу: в результате преобразования единичный вектор $e_1$, очевидно, переходит в вектор $\{\cos\alpha, \sin\alpha\}$, а единичный вектор $e_2$—в вектор $\{-\sin\alpha, \cos\alpha\}$. Следовательно, согласно сказанному выше матрица преобразования есть

$$A = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix},$$

и мы снова приходим к формулам (4).

Определим следующие действия над линейными преобразованиями $A$ и $B$: 1) сложение $A+B$; 2) умножение на скаляр $\alpha A$; 3) умножение (композиция) $BA$ с помощью естественных соглашений:

1) $$(A+B)\,\boldsymbol{x}=A\boldsymbol{x}+B\boldsymbol{x}; \tag{5}$$

2) $$(\alpha A)\,\boldsymbol{x}=\alpha\,(A\boldsymbol{x}); \tag{5'}$$

3) $$(BA)\,\boldsymbol{x}=B\,(A\boldsymbol{x}), \tag{5''}$$

где $\boldsymbol{x}$ — вектор.

Легко проверить, что если $A$ и $B$ — линейные преобразования, то результирующие преобразования $A+B$, $\alpha A$ и $BA$ также линейные.

Имеет место следующий принцип: *каждому действию* 1) — 3) *над линейными преобразованиями соответствует такое же действие над матрицами этих линейных преобразований*, т. е. формулы (5), (5′), (5″) можно понимать в буквальном смысле.

Для доказательства ограничимся третьим случаем (5″) [формулы (5) и (5′) проверяются более просто].

Пусть нам требуется последовательно выполнить два линейных преобразования:

$$\left.\begin{aligned} y_1&=a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n,\\ y_2&=a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_n,\\ &\ldots\ldots\ldots\ldots\ldots\ldots\\ y_n&=a_{n1}x_1+a_{n2}x_2+\ldots+a_{nn}x_n \end{aligned}\right\} \tag{6}$$

с матрицей $A=[a_{ij}]$ и

$$\left.\begin{aligned} z_1&=b_{11}y_1+b_{12}y_2+\ldots+b_{1n}y_n,\\ z_2&=b_{21}y_1+b_{22}y_2+\ldots+b_{2n}y_n,\\ &\ldots\ldots\ldots\ldots\ldots\ldots\\ z_n&=b_{n1}y_1+b_{n2}y_2+\ldots+b_{nn}y_n \end{aligned}\right\} \tag{7}$$

с матрицей $B=[b_{ij}]$.

Обозначим через $C=[c_{ij}]$ матрицу композиции этих преобразований в указанном порядке, т. е. переход от переменных $x_1, x_2, \ldots, x_n$ к переменным $z_1, z_2, \ldots, z_n$. Коротко записывая преобразования (6) и (7) в виде

$$y_k=\sum_{j=1}^{n} a_{kj}x_j \qquad (k=1,\ 2,\ \ldots,\ n). \tag{6'}$$

$$z_i=\sum_{k=1}^{n} b_{ik}y_k \quad (i=1,\ 2,\ \ldots,\ n) \tag{7}$$

и подставляя формулу (6′) в формулу (7′), получим:

$$z_i=\sum_{k=1}^{n} b_{ik}\left(\sum_{j=1}^{n} a_{kj}x_j\right)=\sum_{j=1}^{n} x_j \sum_{k=1}^{n} b_{ik}a_{kj}. \tag{8}$$

Таким образом, коэффициент при $x_j$ в выражении для $z_i$, т. е. элемент $c_{ij}$ матрицы $C$, имеет вид

$$c_{ij}=\sum_{k=1}^{n} b_{ik}a_{kj}=b_{i1}a_{1j}+b_{i2}a_{2j}+\ldots+b_{in}a_{nj}.$$

Мы видим, что элемент матрицы $C$, стоящий в $i$-й строке и $j$-м столбце, равен сумме произведений соответствующих элементов $i$-й строки матрицы $B$ и $j$-го столбца матрицы $A$, т. е. совпадает с соответствующим элементом произведения матрицы $B$ на матрицу $A$. Следовательно, $C=BA$.

Пользуясь матричными обозначениями, доказательство можно провести значительно проще. Пусть

$$x=\begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}; \quad y=\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \quad \text{и} \quad z=\begin{bmatrix} z_1 \\ z_2 \\ \cdot \\ \cdot \\ \cdot \\ z_n \end{bmatrix}$$

— соответствующие векторы. Из формул (6) и (7) имеем:

$$y=Ax \quad \text{и} \quad z=By.$$

Отсюда

$$z=B(Ax)=(BA)x.$$

Следовательно, матрица результирующего преобразования $C=BA$.

Пример 4. Найти результат последовательного выполнения линейных преобразований:

$$\begin{aligned} y_1&=5x_1-x_2+3x_3; \\ y_2&=x_1-2x_2; \\ y_3&=7x_2-x_3 \end{aligned}$$

и

$$\begin{aligned} z_1&=2y_1+y_3; \\ z_2&=y_2-5y_3; \\ z_3&=2y_2. \end{aligned}$$

Решение. Напишем матрицы преобразований

$$A=\begin{bmatrix} 5 & -1 & 3 \\ 1 & -2 & 0 \\ 0 & 7 & -1 \end{bmatrix} \quad \text{и} \quad B=\begin{bmatrix} 2 & 0 & 1 \\ 0 & 1 & -5 \\ 0 & 2 & 0 \end{bmatrix}.$$

Перемножая эти матрицы, получим матрицу результирующего преобразования

$$C=BA=\begin{bmatrix}2 & 0 & 1\\ 0 & 1 & -5\\ 0 & 2 & 0\end{bmatrix}\begin{bmatrix}5 & -1 & 3\\ 1 & -2 & 0\\ 0 & 7 & -1\end{bmatrix}=\begin{bmatrix}10 & 5 & 5\\ 1 & -37 & 5\\ 2 & -4 & 0\end{bmatrix}.$$

Следовательно, искомое линейное преобразование имеет следующий вид:

$$z_1=10x_1+5x_2-5x_3;$$
$$z_2=x_1-37x_2+5x_3;$$
$$z_3=2x_1-4x_2.$$

## § 11. Обратное преобразование

Пусть данное линейное преобразование

$$\left.\begin{aligned} y_1&=a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n,\\ y_2&=a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_n,\\ &\ldots\ldots\ldots\ldots\ldots\ldots\\ y_n&=a_{n1}x_1+a_{n2}x_2+\ldots+a_{nn}x_n \end{aligned}\right\} \qquad (1)$$

переводит систему переменных $x_1, x_2, \ldots, x_n$ в систему переменных $y_1, y_2, \ldots, y_n$.

Определение. Преобразование, переводящее систему переменных $y_1, y_2, \ldots, y_n$ в систему переменных $x_1, x_2, \ldots, x_n$, называется *обратным* по отношению к преобразованию (1).

Формулы обратного преобразования мы получим, разрешив систему (1) относительно переменных $x_1, x_2, \ldots, x_n$.

Пусть $A_{ij}$ $(i, j=1, 2, \ldots, n)$ — алгебраические дополнения элементов $a_{ij}$ матрицы $A=[a_{ij}]$ и

$$\det A=\det[a_{ij}]=\Delta\neq 0.$$

Умножая уравнения системы (1) соответственно на $A_{11}, A_{21}, \ldots, A_{n1}$ и складывая, в силу известной формулы будем иметь:

$$A_{11}y_1+A_{21}y_2+\ldots+A_{n1}y_n=\Delta x_1.$$

Аналогично выводим:

$$A_{12}y_1+A_{22}y_2+\ldots+A_{n2}y_n=\Delta x_2;$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$A_{1n}y_1+A_{2n}y_2+\ldots+A_{nn}y_n=\Delta x_n.$$

Отсюда

$$\left.\begin{aligned} x_1 &= \frac{A_{11}}{\Delta} y_1 + \frac{A_{21}}{\Delta} y_2 + \ldots + \frac{A_{n1}}{\Delta} y_n, \\ x_2 &= \frac{A_{12}}{\Delta} y_1 + \frac{A_{22}}{\Delta} y_2 + \ldots + \frac{A_{n2}}{\Delta} y_n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n &= \frac{A_{1n}}{\Delta} y_1 + \frac{A_{2n}}{\Delta} y_2 + \ldots + \frac{A_{nn}}{\Delta} y_n. \end{aligned}\right\} \tag{2}$$

Таким образом, обратное преобразование линейного преобразования также является линейным (если оно существует).

Теорема. *Линейное преобразование имеет однозначное обратное преобразование тогда и только тогда, когда матрица данного преобразования — неособенная. Обратное преобразование линейного преобразования есть линейное, и его матрица является обратной по отношению к матрице исходного преобразования.*

Доказательство. Если $A=[a_{ij}]$ — матрица преобразования (1) и $\Delta=\det A\neq 0$, то обратное преобразование существует и определяется формулами (2). Матрица обратного преобразования, очевидно, равна

$$\left(\frac{A_{ji}}{\Delta}\right)=A^{-1}.$$

Если $\Delta=0$, то, как известно из алгебры, уравнения (1) нельзя однозначно разрешить относительно переменных $x_1, x_2, \ldots, x_n$. Следовательно, однозначного обратного преобразования не существует, причем обязательно найдутся значения переменных $y_1, y_2, \ldots, y_n$, для которых не существует соответствующих значений переменных $x_1, x_2, \ldots, x_n$. Линейное преобразование в этом случае называется *вырожденным*.

Замечание 1. Запишем преобразование (1) в матричной форме

$$y=Ax, \tag{3}$$

где $A=[a_{ij}]$ — матрица преобразования; $x$ и $y$ — векторы-столбцы.

Если преобразование $A$ — *невырожденное* ($\det A\neq 0$), то существует обратное преобразование

$$x=A^{-1}y, \tag{4}$$

и каждому вектору $x$ из $n$-мерного пространства $Ox_1x_2\ldots x_n$, в силу формулы (3), соответствует один и только один вектор $y$ этого пространства, т. е. формула (3) преобразует пространство $Ox_1x_2\ldots x_n$ в самого себя.

Если преобразование $A$ — *вырожденное* ($\det A=0$), то формула (3) преобразует пространство $Ox_1x_2\ldots x_n$ в подпространство низшего числа измерений.

Пример. Рассмотрим преобразование проектирования (§ 10, пример 2), определяемое матрицей

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

Здесь матрица $A$ — особенная и преобразование $y=Ax$ переводит пространство $Ox_1x_2$ в координатную ось $Ox_1$.

Замечание 2. Будем понимать под $Ex$ тождественное преобразование, оставляющее вектор $x$ неизменным.

Так как из соотношений

$$y=Ax \text{ и } x=A^{-1}y$$

следует

$$y=AA^{-1}y \text{ и } x=A^{-1}Ax,$$

то

$$AA^{-1}=A^{-1}A=E.$$

## § 12. Собственные векторы и собственные значения матрицы

Пусть дана квадратная матрица $A=[a_{ij}]$. Рассмотрим линейное преобразование

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

где $x$ и $y$ — $n$-мерные векторы (столбцовые матрицы) некоторого, вообще говоря, комплексного $n$-мерного пространства.

Определение 1. Ненулевой вектор называется *собственным вектором* данной матрицы (или определяемого ею линейного преобразования), если в результате соответствующего линейного преобразования этот вектор переходит в коллинеарный ему, т. е. если преобразованный вектор отличается от исходного только скалярным множителем.

Иначе говоря, вектор $x\neq 0$ называется *собственным* вектором матрицы $A$, если эта матрица переводит вектор $x$ в вектор

$$Ax=\lambda x. \tag{2}$$

Число $\lambda$, стоящее в равенстве (2), называется *собственным значением*, или *характеристическим числом*, матрицы $A$, соответствующим данному собственному вектору $x$.

Пример 1. Рассмотрим преобразование проектирования в двумерном пространстве $Ox_1x_2$, определяемое матрицей

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

Здесь собственными векторами являются: 1) ненулевые векторы $x$, направленные по оси $Ox_1$, с собственным значением $\lambda_1 = 1$ и 2) ненулевые векторы $y$, направленные по оси $Ox_2$, с собственным значением $\lambda_2 = 0$ (рис. 54).

Теорема 1. *В комплексном векторном пространстве каждое линейное преобразование (матрица) имеет по меньшей мере один действительный или комплексный собственный вектор.*

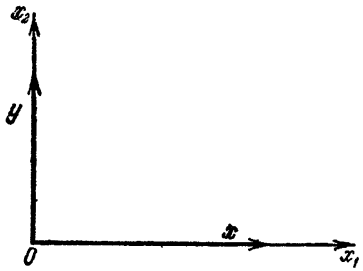


Рис. 54.

Доказательство. Пусть $A$ — матрица линейного преобразования. Собственные векторы матрицы $A$ являются ненулевыми решениями матричного уравнения

$$Ax = \lambda x$$

или

$$(A - \lambda E)\, x = 0, \tag{3}$$

где матрица $A - \lambda E$ называется *характеристической матрицей*. Уравнение (3) представляет собой линейную однородную систему, которая имеет ненулевые решения тогда и только тогда, когда определитель системы равен нулю, т. е. должно выполняться условие

$$\det (A - \lambda E) = 0. \tag{4}$$

Определитель (4) называется *характеристическим* (*вековым*) определителем матрицы $A$, а уравнение (4) называется *характеристическим* (*вековым*) уравнением матрицы $A$. В развернутом виде характеристическое уравнение (4) запишется следующим образом:

$$\begin{vmatrix} a_{11}-\lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22}-\lambda & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn}-\lambda \end{vmatrix} = 0 \tag{4'}$$

или

$$\lambda^n - \sigma_1 \lambda^{n-1} + \sigma_2 \lambda^{n-2} - \dots + (-1)^{n-1} \sigma_{n-1} \lambda + (-1)^n \sigma_n = 0. \tag{5}$$

Полином, стоящий в левой части уравнения (5), называется *характеристическим полиномом* матрицы $A$. Коэффициенты его $\sigma_i$ $(i = 1, 2, \dots, n)$ определяются по следующим правилам. Коэффициент $\sigma_1$ равен сумме диагональных элементов матрицы $A$, т. е. $\sigma_1 = \sum_{i=1}^{n} a_{ii}$. Это число называется *следом* матрицы $A$ и обозначается так: $\sigma_1 = \operatorname{Sp} A$. Коэффициент $\sigma_2$ есть сумма всех диагональных миноров второго порядка матрицы $A$. Вообще, коэффициент $\sigma_k$

есть сумма всех диагональных миноров $k$-го порядка матрицы $A$. Наконец, свободный член $\sigma_n$ равен определителю матрицы $A$:

$$\sigma_n = \det A.$$

Характеристическое уравнение (5) есть алгебраическое уравнение $n$-й степени относительно $\lambda$ и, следовательно, как доказывается в алгебре, имеет по меньшей мере один действительный или комплексный корень. Пусть $\lambda_1, \lambda_2, \ldots, \lambda_m$ $(m \leqslant n)$ — различные корни уравнения (5). Эти корни называются *собственными значениями*, или *характеристическими числами*, матрицы $A$, а совокупность *всех* собственных значений называется *спектром* матрицы $A$. Возьмем какой-нибудь корень $\lambda = \lambda_j$ и подставим его в уравнение (4). Тогда будем иметь $(A - \lambda_j E)\,x = 0$ или, в развернутом виде,

$$\left.\begin{array}{l} (a_{11} - \lambda_j)\,x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = 0, \\ a_{21}x_1 + (a_{22} - \lambda_j)\,x_2 + \ldots + a_{2n}x_n = 0, \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + (a_{nn} - \lambda_j)x_n = 0. \end{array}\right\} \qquad (6)$$

Так как определитель системы (6) $\det(A - \lambda_j E) = 0$, то эта система заведомо имеет ненулевые решения, которые и являются собственными векторами матрицы $A$, соответствующими собственному значению $\lambda_j$. Если ранг матрицы $A - \lambda_j E$ равен $r$ $(r < n)$, то существует $k = n - r$ линейно независимых собственных векторов

$$x^{(1j)},\ x^{(2j)},\ \ldots,\ x^{(kj)},$$

отвечающих корню $\lambda_j$. Теорема доказана.

Замечание. Можно доказать, что число линейно независимых собственных векторов, отвечающих одному и тому же корню характеристического уравнения, не превышает кратности этого корня. Отсюда, в частности, следует, что если корни характеристического уравнения (5) различны, то каждому собственному значению соответствует с точностью до коэффициента пропорциональности один и только один собственный вектор.

Пример 2. Найти собственные значения и собственные векторы матрицы

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}.$$

Решение. Составим характеристическое уравнение матрицы $A$:

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

Отсюда $(\lambda-1)^2(4-\lambda)=0$ и $\lambda_1=\lambda_2=1$; $\lambda_3=4$.

Возьмем $\lambda_1=1$ и подставим в уравнение

$$(A-\lambda_j E)\,\boldsymbol{x}=\mathbf{0}. \tag{7}$$

Имеем:

$$\begin{bmatrix}1&1&1\\1&1&1\\1&1&1\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}=\mathbf{0}$$

или

$$\left.\begin{aligned}x_1+x_2+x_3&=0,\\x_1+x_2+x_3&=0,\\x_1+x_2+x_3&=0.\end{aligned}\right\} \tag{8}$$

Так как ранг матрицы системы (8) $r=1$, то два ее уравнения являются следствием третьего (что, впрочем, очевидно). Поэтому достаточно решить уравнение

$$x_1+x_2+x_3=0.$$

Положив $x_1=c_1$; $x_2=c_2$, получим:

$$x_3=-(c_1+c_2),$$

где $c_1$ и $c_2$ — любые числа, не равные нулю одновременно.

В частности, выбирая сперва $c_1=1$; $c_2=0$, а затем $c_1=0$; $c_2=1$, будем иметь простейшую фундаментальную систему решений, состоящую из двух линейно независимых собственных векторов матрицы $A$:

$$\boldsymbol{x}^{(1)}=\begin{bmatrix}1\\0\\-1\end{bmatrix}\quad\text{и}\quad\boldsymbol{x}^{(2)}=\begin{bmatrix}0\\1\\-1\end{bmatrix}.$$

Все остальные собственные векторы матрицы $A$, соответствующие характеристическому числу $\lambda_1=1$, являются линейной комбинацией этих базисных векторов и заполняют плоскость, натянутую на векторы $\boldsymbol{x}^{(1)}$ и $\boldsymbol{x}^{(2)}$ (исключая начало координат).

Возьмем теперь $\lambda_3=4$. Подставляя это значение в уравнение (7), получим:

$$\begin{bmatrix}-2&1&1\\1&-2&1\\1&1&-2\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}=\mathbf{0}$$

или

$$\left.\begin{aligned}-2x_1+x_2+x_3&=0,\\x_1-2x_2+x_3&=0,\\x_1+x_2-2x_3&=0.\end{aligned}\right\} \tag{9}$$

Ранг матрицы системы (9) $r=2$, причем левый верхний минор

$$\delta=\begin{vmatrix}-2 & 1\\ 1 & -2\end{vmatrix}\neq 0.$$

Следовательно, третье уравнение системы есть следствие двух первых, поэтому можно ограничиться системой из первых двух уравнений:

$$\left.\begin{aligned}-2x_1+x_2+x_3=0,\\ x_1-2x_2+x_3=0.\end{aligned}\right\}$$

Отсюда

$$\frac{x_1}{\begin{vmatrix}1 & 1\\ -2 & 1\end{vmatrix}}=\frac{x_2}{-\begin{vmatrix}-2 & 1\\ 1 & 1\end{vmatrix}}=\frac{x_3}{\begin{vmatrix}-2 & 1\\ 1 & -2\end{vmatrix}}$$

или

$$\frac{x_1}{3}=\frac{x_2}{3}=\frac{x_3}{3},\ \text{т. е.}\quad x_1=x_2=x_3=c,$$

где $c$ — постоянная, отличная от нуля.

Положив $c=1$, получим простейшее решение, реализующее собственный вектор матрицы $A$:

$$x^{(3)}=\begin{bmatrix}1\\1\\1\end{bmatrix}.$$

Определение 2. Линейное подпространство $E_k$ $(k\leqslant n)$ называется *инвариантным* относительно данного линейного преобразования

$$y=Ax,$$

если каждый преобразованный вектор этого подпространства также принадлежит ему, т. е. из $x\in E_k$ следует $Ax\in E_k$.

Очевидно, что доказательство теоремы 1 полностью остается в силе, если рассматривать линейное преобразование, определяемое матрицей $A$, в некотором инвариантном пространстве.

Теорема 1′. *Каждое линейное преобразование, определенное на инвариантном подпространстве комплексного векторного пространства, имеет по меньшей мере один собственный вектор.*

Отметим еще одно важное свойство собственных векторов.

Теорема 2. *Собственные векторы матрицы, соответствующие попарно различным между собой собственным значениям, линейно независимы.*

Доказательство. Пусть $A$ — данная матрица и

$$Ax^{(j)}=\lambda_j x^{(j)}\qquad (j=1,\ 2,\ \ldots,\ m),\tag{10}$$

где

$$x^{(j)}\neq 0\quad\text{и}\quad \lambda_j\neq\lambda_k\quad\text{при}\quad j\neq k.$$

Допустим, что

$$c_1x^{(1)}+c_2x^{(2)}+\ldots+c_mx^{(m)}=\mathbf{0}, \tag{11}$$

где $|c_1|+|c_2|+\ldots+|c_m|\neq 0$.

Пусть для определенности $c_1\neq 0$. Применяя к равенству (11) преобразование $A$, в силу формул (10) будем иметь:

$$\lambda_1c_1x^{(1)}+\lambda_2c_2x^{(2)}+\ldots+\lambda_mc_mx^{(m)}=\mathbf{0}. \tag{12}$$

Отсюда, умножая равенство (11) на $\lambda_m$ и вычитая из полученного равенства равенство (12), находим:

$$(\lambda_m-\lambda_1)\,c_1x^{(1)}+(\lambda_m-\lambda_2)\,c_2x^{(2)}+\ldots$$
$$\ldots+(\lambda_m-\lambda_{m-1})\,c_{m-1}x^{(m-1)}=\mathbf{0}. \tag{13}$$

Далее, из равенства (13) аналогичным приемом можно исключить вектор $x^{(m-1)}$ и т. д. В результате, исключая векторы

$$x^{(m)},\ x^{(m-1)},\ \ldots,\ x^{(2)},$$

получим:

$$(\lambda_m-\lambda_1)(\lambda_{m-1}-\lambda_1)\ \ldots\ (\lambda_2-\lambda_1)\,c_1x^{(1)}=\mathbf{0}. \tag{14}$$

Но последнее равенство невозможно, так как ни один из сомножителей его левой части не равен нулю. Следовательно, наше допущение ложно и собственные векторы $x^{(1)}$, $x^{(2)}$, ..., $x^{(m)}$ линейно независимы.

Следствие. Если все собственные значения матрицы $A$ порядка $n$ попарно различны, то отвечающие им собственные векторы этой матрицы в числе, равном $n$*), образуют базис соответствующего $n$-мерного пространства.

## § 13. Подобные матрицы

Определение. Две матрицы, соответствующие одному и тому же линейному преобразованию в различных базисах, называются *подобными*.

Если матрица $A$ подобна матрице $B$, то пишут $A\backsim B$. Выведем условие подобия двух матриц. Пусть матрица $A$ в некотором базисе реализует линейное преобразование

$$y=Ax. \tag{1}$$

В новом базисе (в новых координатах) это же линейное преобразование будет описываться другой матрицей $B$:

$$\eta=B\xi, \tag{2}$$

где

$$B\backsim A.$$

*) Предполагается, что для каждого собственного значения берется один собственный вектор.

Обозначим через $S$ матрицу перехода от новой системы к старой, т. е. пусть

$$x = S\xi, \qquad y = S\eta, \tag{3}$$

где

$$\det S \neq 0.$$

Подставив формулы (3) в уравнение (1), получим:

$$S\eta = AS\xi.$$

Отсюда, умножая слева последнее равенство на обратную матрицу $S^{-1}$, будем иметь:

$$\eta = S^{-1}AS\xi. \tag{4}$$

Сравнивая формулы (4) и (2), получим:

$$B = S^{-1}AS. \tag{5}$$

Относительно матриц $A$ и $B$, связанных соотношением (5), говорят, что матрица $B$ получается из матрицы $A$ *путем преобразования с помощью матрицы* $S$. Таким образом, заключаем, что *две матрицы подобны тогда и только тогда, когда одна получается из другой путем преобразования с помощью некоторой неособенной матрицы.*

Из равенства (5) выводим $A = SBS^{-1}$, т. е. *если матрица $B$ подобна матрице $A$, то и, наоборот, матрица $A$ подобна матрице $B$.* Отметим некоторые свойства преобразования с помощью матрицы $S$.

1. *Преобразование суммы равно сумме преобразований:*

$$S^{-1}(A+B)S = S^{-1}AS + S^{-1}BS.$$

2. *Преобразование произведения равно произведению преобразований сомножителей:*

$$S^{-1}(AB)S = S^{-1}AS \cdot S^{-1}BS.$$

3. *Преобразование обратной матрицы равно обратной матрице от преобразованной:*

$$S^{-1}A^{-1}S = (S^{-1}AS)^{-1}.$$

4. *Преобразование целой степени (положительной или отрицательной) равно той же степени преобразования:*

$$S^{-1}A^nS = (S^{-1}AS)^n.$$

Теорема 1. *Подобные матрицы имеют одинаковые характеристические полиномы.*

Доказательство. Пусть $B \backsim A$. Требуется доказать, что

$$\det(A-\lambda E)=\det(B-\lambda E).$$

Так как

$$B=S^{-1}AS \quad (\det S \neq 0),$$

то

$$\det(B-\lambda E)=\det[S^{-1}(A-\lambda E)S]=$$
$$=\det S^{-1}\det(A-\lambda E)\det S=\det(A-\lambda E)\ ^{*}).$$

Итак,

$$\det(B-\lambda E)=\det(A-\lambda E).$$

Следствие 1. Подобные матрицы имеют одинаковые следы и одинаковые собственные числа (включая их кратности).

Следствие 2. Свойство вектора быть собственным для данного линейного преобразования не зависит от выбора базиса.

В самом деле, пусть

$$Ax=\lambda x \qquad (x \neq 0).$$

Если в новом базисе вектор $x$ эквивалентен вектору $\xi$, то имеем:

$$x=S\xi,$$

где $S$ — матрица перехода.

Отсюда $AS\xi=\lambda S\xi$ и, следовательно, $S^{-1}AS\xi=\lambda\xi$, т. е. $\xi$ есть собственный вектор для матрицы $B=S^{-1}AS \backsim A$, описывающей в новом базисе наше линейное преобразование.

Замечание. Так как характеристический полином, собственные значения и собственные векторы одинаковы для всех матриц, реализующих данное линейное преобразование, то они называются соответственно *характеристическим полиномом, собственными значениями и собственными векторами самого линейного преобразования.*

Теорема 2. *Если данная квадратная матрица порядка $n$ имеет $n$ линейно независимых собственных векторов, то, приняв последние за базисные, получим диагональную матрицу, подобную данной.*

Доказательство. Пусть имеем квадратную матрицу $A$. Из ее собственных векторов $e_1, e_2, \ldots, e_n$ образуем базис. Так как векторы $e_j$ — собственные, то

$$Ae_j=\lambda_j e_j \qquad (j=1, 2, \ldots, n).$$

---

*) Мы здесь пользовались известными теоремами (см. гл. VII, § 2 и § 4): 1) определитель произведения двух квадратных матриц одинакового порядка равен произведению определителей этих матриц, 2) определитель обратной матрицы равен обратной величине определителя исходной матрицы.

Рассмотрим любой вектор $x$ нашего пространства. Раскладывая его по базисным векторам $e_j$ ($j=1, 2, \ldots, n$), будем иметь:

$$x=\sum_{j=1}^{\infty} x_j e_j,$$

где $x_j$ — координаты вектора $x$ в данном базисе.

Применяя преобразование $A$ к вектору $x$, получим новый вектор

$$y=Ax=A\sum_{j=1}^{n} x_j e_j$$

или, так как преобразование $A$ линейное,

$$y=\sum_{j=1}^{n} x_j A e_j=\sum_{j=1}^{n} x_j \lambda_j e_j.$$

Отсюда видно, что координаты вектора $y$ в данном базисе есть

$$y_j=\lambda_j x_j \qquad (j=1, 2, \ldots, n)$$

или

$$y_j=\sum_{k=1}^{n} \delta_{jk}\lambda_j x_k,$$

где $\delta_{jk}$ — символ Кронекера.

Следовательно, в новом базисе матрица преобразования есть диагональная матрица

$$\Lambda=(\delta_{jk}\lambda_j)$$

или, в развернутом виде,

$$\Lambda=\begin{bmatrix} \lambda_1 & 0 & 0 & \ldots & 0 \\ 0 & \lambda_2 & 0 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & \lambda_n \end{bmatrix}.$$

Следствие. Всякую квадратную матрицу, собственные значения которой попарно различны, путем преобразования подобия можно привести к диагональному виду.

Этот результат вытекает непосредственно из теоремы 2 предыдущего параграфа.

## § 14. Билинейная форма матрицы

Пусть $A=[a_{jk}]$ — действительная квадратная матрица и $x, y$ — векторы $n$-мерного комплексного пространства. Составим скалярное произведение

$$(Ax, y)=\sum_{j=1}^{n} (Ax)_j y_j^*=\sum_{j=1}^{n}\sum_{k=1}^{n} a_{jk}x_k y_j^*. \tag{1}$$

Выражение (1) называется *билинейной формой* матрицы $A$.

Выведем одно важное свойство билинейной формы. Сумма (1), очевидно, будет иметь прежнее значение, если изменить порядок суммирования и одновременно взаимно поменять обозначения индексов суммирования. Поэтому получим:

$$(Ax,\ y) = \sum_{j=1}^{n} \sum_{k=1}^{n} a_{kj} x_j y_k^*.$$

Запишем полученную сумму в виде скалярного произведения

$$(Ax,\ y) = \sum_{j=1}^{n} \sum_{k=1}^{n} a_{kj} x_j y_k^* = \left( \sum_{j=1}^{n} \sum_{k=1}^{n} a_{kj} y_k x_j^* \right)^* = (A'y,\ x)^* = (x,\ A'y).$$

Таким образом,

$$(Ax,\ y) = (x,\ A'y), \tag{2}$$

т. е. *в скалярном произведении* (1) *действительную матрицу* $A$ *можно переставлять с первого места на второе, заменяя ее транспонированной.*

Следствие. Если матрица $A$ — действительная и симметрическая ($A' = A$), то

$$(Ax,\ y) = (x,\ Ay), \tag{3}$$

т. е. *в скалярном произведении действительную симметрическую матрицу можно переставлять с первого места на второе.*

## § 15. Свойства симметрических матриц

Теорема 1. *Все собственные значения симметрической матрицы с действительными элементами — действительные.*

Доказательство. Пусть $\lambda$ — собственное значение матрицы $A$ и $x$ — соответствующий собственный вектор, т. е.

$$Ax = \lambda x \qquad (x \neq 0). \tag{1}$$

Так как $A' = A$, то

$$(Ax,\ x) = (x,\ Ax)$$

или в силу равенства (1)

$$(\lambda x,\ x) = (x,\ \lambda x).$$

Отсюда

$$\lambda (x,\ x) = \lambda^* (x,\ x).$$

Собственный вектор по определению — ненулевой, поэтому

$$(x,\ x) \neq 0$$

и, следовательно, $\lambda = \lambda^*$, т. е. $\lambda$ — действительное число.

Следствие. Характеристическое уравнение для действительной симметрической матрицы имеет только действительные корни.

Теорема 2. *Собственные векторы действительной симметрической матрицы, соответствующие различным собственным значениям, ортогональны между собой.*

Доказательство. Пусть $A$ — действительная симметрическая матрица. Рассмотрим два собственных вектора $x^{(i)}$ и $x^{(j)}$, соответствующие собственным числам $\lambda_i$ и $\lambda_j$ $(\lambda_i \neq \lambda_j)$. Имеем:

$$Ax^{(i)} = \lambda_i x^{(i)} \tag{2}$$

и

$$Ax^{(j)} = \lambda_j x^{(j)}. \tag{3}$$

Составим скалярное произведение

$$(Ax^{(i)},\ x^{(j)}) = (x^{(i)},\ Ax^{(j)}).$$

Отсюда в силу равенств (2) и (3) получаем:

$$(\lambda_i x^{(i)},\ x^{(j)}) = (x^{(i)},\ \lambda_j x^{(j)})$$

и

$$\lambda_i (x^{(i)},\ x^{(j)}) = \lambda_j^* (x^{(i)},\ x^{(j)}). \tag{4}$$

Так как на основании теоремы 1 собственное число $\lambda_j$ — действительное, то $\lambda_j^* = \lambda_j$.

Следовательно, из формулы (4) имеем:

$$(\lambda_i - \lambda_j)(x^{(i)},\ x^{(j)}) = 0.$$

Но

$$\lambda_i - \lambda_j \neq 0,$$

поэтому

$$(x^{(i)},\ x^{(j)}) = 0,$$

т. е. собственные векторы $x^{(i)}$ и $x^{(j)}$ ортогональны между собой.

Замечание. Собственные векторы симметрической матрицы с действительными элементами можно полагать действительными.

Теорема 3. *Всякую действительную симметрическую матрицу при помощи преобразования подобия можно привести к диагональному виду.*

Доказательство. Ограничимся при доказательстве для наглядности случаем трехмерного пространства $E_3$.

Пусть дана симметрическая матрица $A$ третьего порядка. Как известно, всякая матрица имеет по меньшей мере один собственный вектор (§ 12, теорема 1). Обозначим через $e_1$ собственный вектор матрицы $A$. Так как эта матрица — симметрическая, то вектор $e_1$ можно выбрать действительным.

Рассмотрим все векторы $x$, ортогональные к вектору $e_1$, т. е. такие, что

$$(x,\ e_1) = 0. \tag{5}$$

Покажем, что эти векторы образуют инвариантное подпространство $E_2$ относительно преобразования $A$ (рис. 55).

В самом деле, прежде всего, если $x \in E_2$ и $y \in E_2$, т. е.

$$(x, e_1) = (y, e_1) = 0,$$

то для любых чисел $\alpha$ и $\beta$ имеем:

$$(\alpha x + \beta y, e_1) = \alpha (x, e_1) + \beta (y, e_1) = 0$$

и, следовательно,

$$\alpha x + \beta y \in E_2.$$

Таким образом, множество векторов, удовлетворяющих условию (5), образует линейное пространство, причем легко убедиться, что измерение этого пространства равно двум.

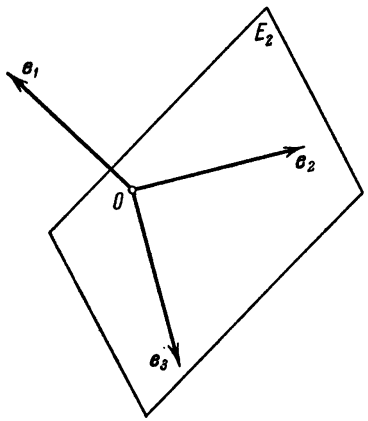

Рис. 55.

Пусть теперь $x \in E_2$. Рассмотрим скалярное произведение

$$(Ax, e_1) = (x, Ae_1) = (x, \lambda_1 e_1) = \\ = \lambda_1 (x, e_1) = 0,$$

т. е.

$$Ax \in E_2.$$

В силу теоремы 1′ (§ 12) в двумерном пространстве $E_2$ также существует собственный вектор $e_2$ матрицы $A$. Будем теперь рассматривать векторы $x$, ортогональные как к вектору $e_1$, так и к вектору $e_2$, т. е. такие, что

$$(x, e_1) = (x, e_2) = 0.$$

Аналогично предыдущему доказывается, что эти векторы образуют одномерное пространство $E_1$, инвариантное относительно преобразования $A$. В пространстве $E_1$ снова имеется собственный вектор $e_3$ матрицы $A$. Векторы $e_1$, $e_2$, $e_3$, будучи попарно ортогональными, линейно независимы между собой. Таким образом, мы построим ортогональный базис пространства $E_3$, состоящий из собственных векторов матрицы $A$.

Обозначим через $\lambda_j$ собственные значения, соответствующие собственным векторам $e_j$. В силу теоремы 2 из § 13 матрица $\Lambda$ данного линейного преобразования в собственном базисе $e_1$, $e_2$, $e_3$ будет диагональной, причем в нашем случае

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}.$$

Аналогично доказывается теорема и в общем случае.

Следствие 1. Для всякого линейного преобразования с действительной симметрической матрицей существует ортогональный базис, состоящий из действительных собственных векторов данной матрицы, в котором матрица преобразования — диагональная.

Следствие 2. Если матрица — симметрическая, то каждому собственному значению ее соответствует столько линейно независимых собственных векторов, какова кратность этого собственного значения.

Теорема 4 (экстремальное свойство собственных значений). *Пусть $A$ — действительная симметрическая матрица и*

$$\lambda = \min(\lambda_1, \lambda_2, \ldots, \lambda_n),$$

$$\Lambda = \max(\lambda_1, \lambda_2, \ldots, \lambda_n),$$

*где $\lambda_1, \lambda_2, \ldots, \lambda_n$ — все собственные значения матрицы $A$.*

*Тогда для любого вектора $\boldsymbol{x}$ справедливо неравенство*

$$\lambda(\boldsymbol{x}, \boldsymbol{x}) \leqslant (A\boldsymbol{x}, \boldsymbol{x}) \leqslant \Lambda(\boldsymbol{x}, \boldsymbol{x}). \tag{6}$$

Доказательство. В силу следствия 1 к теореме 3 существует система собственных векторов $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ матрицы $A$:

$$A\boldsymbol{e}_j = \lambda_j \boldsymbol{e}_j \qquad (j = 1, 2, \ldots, n),$$

образующих нормированный ортогональный базис пространства $E_n$. Тогда любой вектор $\boldsymbol{x}$ можно представить в виде

$$\boldsymbol{x} = x_1\boldsymbol{e}_1 + x_2\boldsymbol{e}_2 + \ldots + x_n\boldsymbol{e}_n,$$

где $x_1, x_2, \ldots, x_n$ — координаты вектора $\boldsymbol{x}$ в данном базисе. Отсюда

$$A\boldsymbol{x} = x_1A\boldsymbol{e}_1 + x_2A\boldsymbol{e}_2 + \ldots + x_nA\boldsymbol{e}_n = \lambda_1x_1\boldsymbol{e}_1 + \lambda_2x_2\boldsymbol{e}_2 + \ldots + \lambda_nx_n\boldsymbol{e}_n.$$

Учитывая ортогональность векторов базиса, будем иметь:

$$(A\boldsymbol{x}, \boldsymbol{x}) = \left(\sum_{j=1}^{n} \lambda_j x_j \boldsymbol{e}_j, \sum_{k=1}^{n} x_k \boldsymbol{e}_k\right) = \sum_{j=1}^{n}\sum_{k=1}^{n} \lambda_j x_j x_k^* (\boldsymbol{e}_j, \boldsymbol{e}_k) =$$

$$= \sum_{j=1}^{n}\sum_{k=1}^{n} \lambda_j x_j x_k^* \delta_{jk} = \sum_{j=1}^{n} \lambda_j |x_j|^2,$$

т. е.

$$(A\boldsymbol{x}, \boldsymbol{x}) = \sum_{j=1}^{n} \lambda_j |x_j|^2. \tag{7}$$

Заменяя в равенстве (7) $\lambda_j$ на наименьшее значение $\lambda$, получим:

$$(A\boldsymbol{x}, \boldsymbol{x}) \geqslant \lambda \sum_{j=1}^{n} |x_j|^2 = \lambda(\boldsymbol{x}, \boldsymbol{x}).$$

Аналогично, подставляя в равенство (7) вместо $\lambda_j$ наибольшее значение $\Lambda$, находим:

$$(Ax, x) \leqslant \Lambda \sum_{j=1}^{n} |x_j|^2 = \Lambda (x, x).$$

Таким образом, неравенство (6) доказано.

Следствие. Минимальное собственное значение $\lambda$ и максимальное собственное значение $\Lambda$ симметрической действительной матрицы $A$ являются соответственно наименьшим и наибольшим значениями квадратичной формы

$$u = (Ax, x)$$

на единичной сфере $(x, x) = 1$.

Действительно, полагая $(x, x) = 1$ в неравенстве (6), будем иметь:

$$\lambda \leqslant (Ax, x) \leqslant \Lambda.$$

Кроме того, если $Ax = \lambda x$, то

$$(Ax, x) = (\lambda x, x) = \lambda;$$

аналогично, если $Ax = \Lambda x$, то

$$(Ax, x) = (\Lambda x, x) = \Lambda.$$

Таким образом,

$$\lambda = \min (Ax, x) \quad \text{при} \quad (x, x) = 1$$

и

$$\Lambda = \max (Ax, x) \quad \text{при} \quad (x, x) = 1.$$

Действительную симметрическую матрицу $A = [a_{ij}]$ назовем *положительно определенной,* если соответствующая квадратичная форма

$$u = (Ax, x) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j^*$$

является положительно определенной (гл. VIII, § 13), т. е. для любого вектора $x \neq 0$ имеем:

$$(Ax, x) > 0.$$

Теорема 5. *Действительная симметрическая матрица является положительно определенной тогда и только тогда, когда все собственные значения ее положительны.*

Доказательство. Если $A$ — действительная симметрическая матрица и ее собственные значения $\lambda_j$ таковы, что $\lambda_j > 0$ ($j = 1, 2, \ldots, n$), то на основании формулы (7) из доказательства предыдущей теоремы имеем:

$$(Ax, x) = \sum_{j=1}^{n} \lambda_j |x_j|^2,$$

где $\boldsymbol{x}=(x_1, x_2, \ldots, x_n)$. Отсюда при $\boldsymbol{x}\neq\boldsymbol{0}$ получим:

$$(A\boldsymbol{x}, \boldsymbol{x})>0,$$

т. е. матрица $A$ — положительно определенная.

Обратно, пусть $A$ — действительная симметрическая положительно определенная матрица.

В силу теоремы 1 все ее собственные значения $\lambda_1, \lambda_2, \ldots, \lambda_n$ действительны, причем

$$\lambda=\min(\lambda_1, \lambda_2, \ldots, \lambda_n),$$

является наименьшим значением квадратичной формы $u=(A\boldsymbol{x}, \boldsymbol{x})$ на сфере $(\boldsymbol{x}, \boldsymbol{x})=1$. Так как сфера $(\boldsymbol{x}, \boldsymbol{x})=1$ — компактное и ограниченное множество и квадратичная форма $u$ непрерывна и положительна на этой сфере, то по теореме Вейерштрасса наименьшее значение $u$ существует и также положительно, т. е.

$$\lambda>0.$$

Отсюда и подавно

$$\lambda_j>0 \quad \text{при} \quad j=1, 2, \ldots, n.$$

Приведем без доказательства условия положительной определенности действительной матрицы [2].

Теорема 6. *Для положительной определенности действительной матрицы $A=[a_{ij}]$, где $a_{ij}=a_{ji}$, необходимо и достаточно, чтобы были выполнены условия Сильвестра:*

$$\Delta_1=a_{11}>0; \qquad \Delta_2=\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}>0; \ \ldots;$$

$$\Delta_n=\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}>0,$$

т. е. действительная симметрическая матрица $A$ положительно определенная тогда и только тогда, если все главные диагональные миноры ее определителя $\det A$ строго положительны.

## § 16*. Свойства матриц с действительными элементами

В дальнейшем мы, как правило, будем рассматривать матрицы $A=[a_{ij}]$, элементы которых $a_{ij}$ действительны; такие матрицы называются *действительными* или *вещественными*.

Пусть $A=[a_{ij}]$ — действительная квадратная матрица порядка $n$. Так как ее характеристическое уравнение

$$\det(A-\lambda E)=0$$

есть полином с действительными коэффициентами, то корни $\lambda_1, \lambda_2, \ldots, \lambda_n$ характеристического уравнения, представляющие собой собственные значения матрицы $A$, в случае их комплексности попарно сопряжены (гл. V, § 1), т. е. если $\lambda_s$ есть собственное значение матрицы $A$, то сопряженное число $\lambda_s^*$ также является собственным значением матрицы $A$ и имеет ту же кратность.

Действительная матрица может не иметь действительных собственных значений. Однако в одном важном случае, когда элементы матрицы положительны, гарантируется существование хотя бы одного действительного собственного значения [6].

Теорема Перрона. *Если все элементы квадратной матрицы положительны, то наибольшее по модулю собственное значение ее также положительно и является простым корнем характеристического уравнения матрицы, причем ему соответствует собственный вектор с положительными координатами.*

Собственные векторы действительной матрицы $A$ с различными собственными значениями в общем случае комплексные и не обладают свойством ортогональности. Однако, привлекая собственные векторы транспонированной матрицы $A'$, можно получить так называемые *соотношения биортогональности*, которые для случая симметрической матрицы эквивалентны обычным соотношениям ортогональности.

Теорема 1. *Если матрица $A$ — действительная и собственные значения ее попарно различны, то существуют два базиса $\{x_j\}$ и $\{x'_j\}$ пространства $E_n$, состоящих соответственно из собственных векторов матрицы $A$ и собственных векторов транспонированной матрицы $A'$, удовлетворяющих следующим условиям биортонормировки:*

$$(x_j, x'_k) = \begin{cases} 0 & \text{при} \quad j \neq k, \\ 1 & \text{при} \quad j = k. \end{cases}$$

Доказательство. Пусть $\lambda_1, \lambda_2, \ldots, \lambda_n$ — собственные значения матрицы $A$. Так как матрица $A$ — действительная, то, как мы знаем, собственные значения ее — попарно сопряженные, т. е. наряду с собственным значением $\lambda_j$ сопряженное число $\lambda_j^*$ также является собственным значением матрицы $A$. Обозначим через $x_j$ $(j = 1, 2, \ldots, n)$ соответствующие собственные векторы матрицы $A$, т. е.

$$Ax_j = \lambda_j x_j \qquad (j = 1, 2, \ldots, n). \tag{1}$$

Векторы $\{x_j\}$ образуют базис пространства $E_n$ (§ 12, теорема 2, следствие).

Так как определитель не изменяет своего значения при замене строк столбцами, то

$$\det(A' - \lambda E) \equiv \det(A - \lambda E)$$

и, следовательно, транспонированная матрица $A'$ имеет те же собственные значения $\lambda_j$, что и матрица $A$. Пусть $x'_j$ $(j=1, 2 \ldots, n)$ — собственные векторы матрицы $A'$, соответствующие сопряженным собственным значениям $\lambda_j^*$, т. е.

$$A'x'_j = \lambda_j^* x'_j \qquad (j=1, 2, \ldots, n). \tag{2}$$

Векторы $\{x'_j\}$ также образуют базис пространства $E_n$.

Базисы $\{x_j\}$ и $\{x'_j\}$ *биортогональны*, а именно:

$$(x_j, x'_k) = 0 \quad \text{при} \quad j \neq k. \tag{3}$$

Действительно, с одной стороны, имеем:

$$(Ax_j, x'_k) = (\lambda_j x_j, x'_k) = \lambda_j (x_j, x'_k). \tag{4}$$

С другой стороны, учитывая вещественность матрицы $A$, получаем:

$$(Ax_j, x'_k) = (x_j, A'x'_k) = (x_j, \lambda_k^* x'_k) = \lambda_k (x_j, x'_k). \tag{5}$$

Из равенств (4) и (5) выводим:

$$\lambda_j (x_j, x'_k) = \lambda_k (x_j, x'_k). \tag{6}$$

Так как $\lambda_j \neq \lambda_k$ при $j \neq k$, то из равенства (6) вытекает равенство (3). Покажем, что векторы $\{x_j\}$ и $\{x'_j\}$ можно нормировать так, чтобы

$$(x_j, x'_j) = 1 \qquad (j=1, 2, \ldots, n). \tag{7}$$

В самом деле, разлагая вектор $x_j$ по векторам базиса $\{x'_1, x'_2, \ldots, x_n\}$, будем иметь:

$$x_j = c_1 x'_1 + \ldots + c_j x'_j + \ldots + c_n x'_n.$$

Отсюда, учитывая условие биортогональности (3), получим:

$$(x_j, x_j) = c_1^* (x_j, x'_1) + \ldots + c_j^* (x_j, x'_j) + \ldots$$
$$\ldots + c_n^* (x_j, x'_n) = c_j^* (x_j, x'_j) > 0;$$

поэтому

$$(x_j, x'_j) = \alpha_j \neq 0.$$

Взяв вместо векторов $x'_1, \ldots, x'_n$ векторы $\frac{1}{\alpha_1^*} x'_1, \ldots, \frac{1}{\alpha_n^*} x'_n$, получим требуемую нормировку (7), так как

$$\left(x_j, \frac{1}{\alpha_j^*} x'_j\right) = \frac{1}{\alpha_j} (x_j, x'_j) = \frac{1}{\alpha_j} \cdot \alpha_j = 1 \qquad (j=1, 2, \ldots, n).$$

Таким образом, если собственные значения действительной матрицы $A$ различны, то для собственного базиса $\{x_j\}$ матрицы $A$

всегда можно найти собственный базис $\{x'_j\}$ транспонированной матрицы $A'$ такой, что

$$(x_j, x'_k) = \delta_{jk}, \tag{8}$$

где $\delta_{jk}$ — символ Кронекера.

Следствие. Если матрица $A$ — действительная и симметрическая ($A' = A$), то можно положить: $x'_j = x_j$ ($j = 1, 2, \ldots, n$), где $x_j$ — нормированные собственные векторы матрицы $A$ (см. § 15). Тогда

$$(x_j, x_k) = \delta_{jk}.$$

Выведем еще так называемое *билинейное разложение матрицы A*.

Теорема 2. *Пусть $A$ — квадратная действительная матрица и*

$$X_j = \begin{bmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{bmatrix}$$

($j = 1, 2, \ldots, n$) — *ее собственные векторы, рассматриваемые как матрицы-столбцы, и*

$$X'_k = [x'_{1k} \ \ldots \ x'_{nk}]$$

($k = 1, 2, \ldots, n$) — *соответствующие*) собственные векторы транспонированной матрицы $A'$, рассматриваемые как матрицы-строки, причем выполнены условия биортонормировки* (8):

$$(X_j, X'^*_k) = X'_k X_j = \delta_{jk}. \tag{9}$$

*Тогда имеет место соотношение*

$$A = \lambda_1 X_1 X'_1 + \lambda_2 X_2 X'_2 + \ldots + \lambda_n X_n X'_n, \tag{10}$$

*где $\lambda_1, \lambda_2, \ldots, \lambda_n$ — собственные значения матрицы $A$.*

Доказательство. Рассмотрим матрицы

$$X = \begin{bmatrix} x_{11} & \cdot & x_{1n} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdot & x_{nn} \end{bmatrix} \quad \text{и} \quad X' = \begin{bmatrix} x'_{11} & \ldots & x'_{n1} \\ \cdot & \cdot & \cdot \\ x'_{1n} & \ldots & x'_{nn} \end{bmatrix},$$

состоящие соответственно из столбцов $X_j$ ($j = 1, \ldots, n$) и строк $X'_k$ ($k = 1, \ldots, n$).

В силу равенства (9) будем иметь:

$$X'X = \left[\sum_{k=1}^{n} x'_{ki} x_{kj}\right] = [X_j X'_i] = [\delta_{ji}] = E, \tag{11}$$

где $E$ — единичная матрица. Так как матрица $X$ состоит из линейно

*) То есть отвечающие тем же собственным значениям матриц $A$ и $A'$.

независимых столбцов, то она — неособенная, т. е. $\det X \neq 0$ и, следовательно, существует обратная матрица $X^{-1}$. На основании равенства (11) (см. гл. VII, § 4, теорема, замечание 1) имеем:

$$X^{-1} = X'.$$

Отсюда вытекает, что

$$XX' = E,$$

и, таким образом, мы получаем *вторые соотношения биортогональности* [7]

$$\sum_{k=1}^{n} x_{ik} x'_{jk} = \delta_{ij}. \tag{12}$$

Используя эти соотношения, имеем:

$$X_1 X'_1 + X_2 X'_2 + \ldots + X_n X'_n = [x_{i1} x'_{j1}] + [x_{i2} x'_{j2}] + \ldots + [x_{in} x'_{jn}] = \\ = \left[\sum_{k=1}^{n} x_{ik} x'_{jk}\right] = [\delta_{ij}] = E,$$

т. е.

$$E = X_1 X'_1 + X_2 X'_2 + \ldots + X_n X'_n.$$

Умножая это равенство слева на матрицу $A$ и учитывая, что

$$AX_j = \lambda_j X_j \quad (j = 1, 2, \ldots, n),$$

очевидно, получим равенство (10).

Обратим внимание на то, что в формуле (10) $X_j$ и $X'_j$ $(j = 1, 2, \ldots, n)$ — собственные векторы матриц $A$ и $A'$, соответствующие одному и тому же собственному значению $\lambda_j$, вопреки обозначениям теоремы 1, где $x_j$ и $x'_j$ — собственные векторы матриц $A$ и $A'$, соответствующие комплексно сопряженным собственным значениям $\lambda_j$ и $\lambda_j^*$.

## Литература к десятой главе


1. Г. Е. Шилов, Введение в теорию линейных пространств, Гостехиздат, М—Л., 1952, гл. I—IX.
2. И. М. Гельфанд, Лекции по линейной алгебре, Изд. 2., Гостехиздат, М.—Л., 1951, гл. I—II.
3. А. И. Мальцев, Основы линейной алгебры, Гостехиздат, М.—Л., 1948, гл. I—III.
4. А. С. Хаусхолдер, Основы численного анализа, ИЛ, 1956, гл. II.
5. Ю. А. Шрейдер, Решение систем линейных алгебраических уравнений, Докл. АН СССР, 5 (1951).
6. Ф. Р. Гантмахер, Теория матриц, Гостехиздат, М., 1953, гл. VIII.
7. В. Н. Фаддеева, Вычислительные методы линейной алгебры, Гостехиздат, М.—Л., 1950, гл. I.

ГЛАВА XI*

# ДОПОЛНИТЕЛЬНЫЕ СВЕДЕНИЯ О СХОДИМОСТИ ИТЕРАЦИОННЫХ ПРОЦЕССОВ ДЛЯ СИСТЕМ ЛИНЕЙНЫХ УРАВНЕНИЙ

## § 1. Сходимость матричных степенных рядов

Теорема 1. *Матричный степенной ряд*

$$\sum_{k=0}^{\infty} a_k X^k \tag{1}$$

*с числовыми коэффициентами* $a_k$ *сходится, если все собственные значения* $\lambda_1, \lambda_2, \ldots, \lambda_n$ *матрицы* $X$ *расположены з замкнутом круге сходимости* $|x| \leqslant R$ (рис. 56) *скалярного степенного ряда*

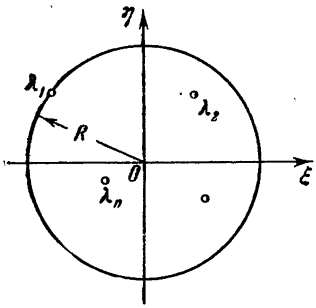


Рис. 56.

$$\sum_{k=0}^{\infty} a_k x^k \tag{2}$$

($x = \xi + i\eta$), *причем собственные значения, лежащие на окружности круга сходимости, простые и являются точками сходимости ряда* (2).

*Ряд* (1) *расходится, если хотя бы одно собственное значение матрицы* $X$ *находится вне замкнутого круга сходимости ряда* (2), *или имеется собственное значение матрицы* $X$, *лежащее на окружности круга сходимости, для которого ряд* (2) *расходится.*

Доказательство. 1) Пусть матрица $X$ такова, что

$$|\lambda_1| \leqslant R, \ \ldots, \ |\lambda_n| \leqslant R.$$

Предположим для простоты, что собственные значения $\lambda_j$ ($j = 1, 2, \ldots, n$) матрицы $X$ — простые. Тогда матрица $X$ с помощью неособенной матрицы $S$ может быть приведена к диагональному виду

$$X = S^{-1}[\lambda_1, \ \ldots, \ \lambda_n]S.$$

Введем обозначения

$$F_m(X)=\sum_{k=0}^{m} a_k X^k, \qquad f_m(x)=\sum_{k=0}^{m} a_k x^k$$

и

$$f(x)=\lim_{m\to\infty} f_m(x)=\sum_{k=0}^{\infty} a_k x^k.$$

Имеем:

$$F_m(X)=\sum_{k=0}^{m} a_k \{S^{-1}[\lambda_1, \ldots, \lambda_n]S\}^k = S^{-1}\left\{\sum_{k=0}^{m} a_k [\lambda_1^k, \ldots, \lambda_n^k]\right\}S=$$
$$=S^{-1}[f_m(\lambda_1), \ldots, f_n(\lambda_n)]S. \qquad (3)$$

Так как числа $\lambda_j$ расположены внутри круга сходимости степенного ряда (2) или совпадают с точками сходимости этого ряда, принадлежащими окружности круга сходимости, то существуют конечные пределы

$$f(\lambda_j)=\lim_{m\to\infty} f_m(\lambda_j) \quad (j=1, 2, \ldots, n).$$

Поэтому, переходя к пределу при $m\to\infty$ в формуле (3), получаем:

$$F(X)=\lim_{m\to\infty} F_m(X)=S^{-1}[f(\lambda_1), \ldots, f(\lambda_n)]S, \qquad (4)$$

т. е. матричный ряд (1) в точке $X$ сходится.

Можно доказать также, что утверждение теоремы верно и для случая кратных собственных значений $\lambda_j$, на чем мы останавливаться не будем [1].

2) Пусть, например, хотя бы одно собственное значение $\lambda_1$ матрицы $X$ таково, что

$$|\lambda_1|>R.$$

Так как $\lambda_1$ лежит вне круга сходимости степенного ряда (2), то $f_m(\lambda_1)$ при $m\to\infty$ не имеет предела. Из формулы (3) следует, что $F_m(X)$ при $m\to\infty$ также не имеет предела, т. е. ряд (1) в точке $X$ расходится.

Аналогичный результат получается, если $|\lambda_1|=R$ и ряд $\sum_{k=0}^{\infty} a_k\lambda_1^k$ расходится.

Замечание. Из формулы (4) следует, что если $\lambda_1, \lambda_2, \ldots, \lambda_n$ есть простые собственные значения матрицы $X$, то $f(\lambda_1), f(\lambda_2), \ldots$ $\ldots, f(\lambda_n)$, где

$$f(x)=\sum_{k=0}^{\infty} a_k x^k,$$

являются собственными значениями функции

$$F(X)=\sum_{k=0}^{\infty} a_k X^k.$$

В частности, собственными значениями матрицы $X^k$ являются числа $\lambda_1^k, \ldots, \lambda_n^k$.

Теорема 2. *Матричная геометрическая прогрессия*

$$E+X+X^2+\ldots+X^k+\ldots, \tag{5}$$

*где $X$ — квадратная матрица порядка $n$, сходится тогда и только тогда, когда все собственные значения*

$$\lambda_j=\lambda_j(X) \quad (j=1, 2, \ldots, n)$$

*матрицы $X$ расположены внутри единичного круга*

$$|\lambda_j|<1 \quad (j=1, 2, \ldots, n), \tag{6}$$

*причем если ряд* (5) *расходится, то $X^k \not\to 0$ при $k \to \infty$.*

Доказательство. Действительно, так как для соответствующего степенного ряда

$$\sum_{k=0}^{\infty} x^k \tag{7}$$

радиус сходимости $R=1$, причем при $|x|=1$ ряд (7) расходится, то в силу теоремы 1 геометрическая прогрессия (5) сходится лишь при выполнении условий (6).

Если ряд (5) расходится, то

$$|\lambda_j| \geqslant 1 \quad (j=1, 2, \ldots, n).$$

Отсюда, предполагая для простоты, что собственные значения $\lambda_1, \ldots, \lambda_n$ различны, будем иметь:

$$X=S^{-1}[\lambda_1, \ldots, \lambda_n]S,$$

где $S$ — неособенная матрица. Поэтому

$$X^k=S^{-1}[\lambda_1^k, \ldots, \lambda_n^k]S,$$

и, следовательно, $X^k \not\to 0$ при $k \to \infty$. Последнее утверждение остается справедливым и при наличии кратных значений $\lambda_j$, на чем мы останавливаться не будем.

Теорема 3. *Модуль каждого собственного значения $\lambda_1, \ldots, \lambda_n$ квадратной матрицы $X$ не превосходит любой ее канонической нормы,* т. е.

$$|\lambda_j| \leqslant \|X\| \quad (j=1, 2, \ldots, n).$$

Доказательство. Положим

$$\|X\| = \rho$$

и рассмотрим матрицу

$$Y = \frac{1}{\rho+\varepsilon} X, \tag{8}$$

где $\varepsilon > 0$. Очевидно,

$$\|Y\| = \frac{1}{\rho+\varepsilon}\|X\| = \frac{\rho}{\rho+\varepsilon} < 1.$$

Следовательно (гл. VII, § 10, теорема 5) ряд

$$E + Y + Y^2 + \ldots + Y^k + \ldots$$

сходится.

Отсюда в силу теоремы 2 заключаем, что собственные значения $\mu_1, \ldots, \mu_n$ матрицы $Y$ удовлетворяют неравенствам

$$|\mu_j| < 1 \quad (j = 1, 2, \ldots, n).$$

Но из формулы (8) вытекает, что

$$\mu_j = \frac{1}{\rho+\varepsilon}\lambda_j \quad (j = 1, 2, \ldots, n).$$

Следовательно,

$$|\lambda_j| < \rho + \varepsilon \qquad (j = 1, 2, \ldots, n)$$

или, ввиду произвольности числа $\varepsilon$,

$$|\lambda_j| \leqslant \rho = \|X\| \quad (j = 1, 2, \ldots, n),$$

что и требовалось доказать.

## § 2. Тождество Гамильтона — Кели

Теорема. *Всякая квадратная матрица $X$ является корнем своего характеристического полинома, т. е. если*

$$\psi(\lambda) = \lambda^n + p_1\lambda^{n-1} + \ldots + p_n,$$

*где* $\psi(\lambda) = \det(\lambda E - X)$, *то*

$$\psi(X) = X^n + p_1 X^{n-1} + \ldots + p_n E \equiv 0.$$

Доказательство. Пусть все собственные значения $\lambda_1, \lambda_2, \ldots, \lambda_n$ матрицы $X$, т. е. корни характеристического уравнения $\psi(\lambda) = 0$, различны. Тогда матрицу $X$ с помощью некоторой неособенной матрицы $S$ можно привести к диагональному виду

$$X = S^{-1}[\lambda_1, \lambda_2, \ldots, \lambda_n]S.$$

Так как $\psi(X)$ представляет собой частный случай матричного степенного ряда, то на основании формулы (4) из § 1 имеем:

$$\psi(X)=S^{-1}[\psi(\lambda_1),\ \psi(\lambda_2),\ \ldots,\ \psi(\lambda_n)]\,S.$$

Но, очевидно,

$$\psi(\lambda_j)=0 \qquad (j=1,\ 2,\ \ldots,\ n).$$

Поэтому

$$\psi(X)=S^{-1}[0,\ 0,\ \ldots,\ 0]\,S=0.$$

Если характеристическое уравнение $\psi(\lambda)=0$ имеет кратные корни, то их можно рассматривать как пределы несовпадающих корней при бесконечно малых возмущениях коэффициентов уравнения [1]. В результате теорема обобщается и на этот случай.

## § 3. Необходимые и достаточные условия сходимости процесса итерации для системы линейных уравнений

Используя собственные значения матрицы $\alpha=[\alpha_{ij}]$, можно дать необходимые и достаточные условия сходимости процесса итерации для линейной системы

$$\boldsymbol{x}=\alpha\boldsymbol{x}+\boldsymbol{\beta}, \tag{1}$$

где

$$\boldsymbol{\beta}=\begin{bmatrix}\beta_1\\ \vdots\\ \beta_n\end{bmatrix} \quad \text{и} \quad \boldsymbol{x}=\begin{bmatrix}x_1\\ \vdots\\ x_n\end{bmatrix}.$$

Теорема. *Для сходимости процесса итерации*

$$\boldsymbol{x}^{(k)}=\alpha\boldsymbol{x}^{(k-1)}+\boldsymbol{\beta} \tag{2}$$
$$(k=1,\ 2,\ \ldots)$$

*при любом выборе начального вектора* $\boldsymbol{x}^{(0)}$ *и любом свободном члене* $\boldsymbol{\beta}$ *необходимо и достаточно, чтобы собственные значения матрицы* $\alpha$, *т. е. корни характеристического уравнения*

$$\det(\alpha-\lambda E)=0,$$

*были по модулю меньше единицы.*

Доказательство. Из формулы (2) получаем:

$$\boldsymbol{x}^{(k)}=(E+\alpha+\alpha^2+\ldots+\alpha^{k-1})\,\boldsymbol{\beta}+\alpha^k\boldsymbol{x}^{(0)}. \tag{3}$$

Отсюда следует, что сходимость процесса итерации (2) при произвольных $\boldsymbol{\beta}$ и $\boldsymbol{x}^{(0)}$ эквивалентна сходимости матричной геометрической прогрессии

$$E+\alpha+\alpha^2+\ldots=\sum_{k=0}^{\infty}\alpha^k. \tag{4}$$

В силу теоремы 2 из § 1 геометрическая прогрессия (4) сходится, если все собственные значения $\lambda_j$ $(j=1, 2, \ldots, n)$ матрицы $\alpha$ удовлетворяют неравенствам

$$|\lambda_j| < 1 \qquad (j=1, 2, \ldots, n). \tag{5}$$

Так как при этом $\alpha^k \to 0$ при $k \to \infty$, то из формулы (3) вытекает, что процесс итерации сходится при любых $\boldsymbol{\beta}$ и $\boldsymbol{x}^{(0)}$, т. е. существует

$$\lim_{k \to \infty} \boldsymbol{x}^{(k)} = \boldsymbol{x},$$

где $\boldsymbol{x}$ —, очевидно, решение системы (1).

Если неравенства (5) не выполнены, то ряд (4) расходится. В таком случае при некотором выборе начального вектора $\boldsymbol{x}^{(0)}$ процесс итерации, очевидно, также расходится.

Таким образом, для сходимости процесса итерации (2) необходимо и достаточно, чтобы все корни $\lambda_1, \lambda_2, \ldots, \lambda_n$ характеристического уравнения

$$\begin{vmatrix} \alpha_{11}-\lambda & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22}-\lambda & \ldots & \alpha_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \alpha_{n1} & \alpha_{n2} & \ldots & \alpha_{nn}-\lambda \end{vmatrix} = 0$$

удовлетворяли условиям: $|\lambda_j| < 1$ $(j=1, 2, \ldots, n)$.

Следствие. Для сходимости процесса итерации (2) достаточно, чтобы

$$\|\alpha\| < 1, \tag{6}$$

какова бы ни была каноническая норма (ср. гл. IX, § 1).

Действительно, в силу теоремы 3 из § 1, на основании неравенства (6) получаем неравенства (5).

Замечание. Рассмотрим линейную систему

$$A\boldsymbol{x} = \boldsymbol{b}, \tag{7}$$

где $A = [a_{ij}]$ и $\boldsymbol{b} = [b_1 \ \ldots \ b_n]$ — вектор-столбец.

Пусть

$$D = \begin{bmatrix} a_{11} & 0 & \ldots & 0 \\ 0 & a_{22} & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & a_{nn} \end{bmatrix} \neq 0.$$

Для приведения системы (7) к специальному виду (1) обычно полагают:

$$A = D + (A - D).$$

Отсюда

$$D\boldsymbol{x} = \boldsymbol{b} - (A - D)\,\boldsymbol{x},$$

и так как $\det D = a_{11}a_{22} \ldots a_{nn} \neq 0$, то

$$\boldsymbol{x} = D^{-1}\boldsymbol{b} + D^{-1}(D - A)\boldsymbol{x}.$$

Можно принять

$$\alpha = D^{-1}(D - A).$$

Таким образом, для сходимости обычного процесса итерации для линейной системы (7) при любом свободном члене $\boldsymbol{b}$ и любом начальном векторе $\boldsymbol{x}^{(0)}$ необходимо и достаточно, чтобы все корни $\lambda_1$, $\lambda_2$, ..., $\lambda_n$ характеристического уравнения

$$\det[D^{-1}(D - A) - \lambda E] = 0 \tag{8}$$

были по модулю меньше единицы. Пользуясь теоремой об определителе произведения двух матриц, уравнение (8) можно преобразовать следующим образом:

$$\det D^{-1} \det[(D - A) - \lambda D] = 0$$

или

$$\det[\lambda D + (A - D)] = 0,$$

т. е.

$$\begin{vmatrix} a_{11}\lambda & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22}\lambda & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn}\lambda \end{vmatrix} = 0.$$

## § 4. Необходимые и достаточные условия сходимости процесса Зейделя для системы линейных уравнений

Для линейной системы

$$\boldsymbol{x} = \alpha\boldsymbol{x} + \boldsymbol{\beta}, \tag{1}$$

где $\alpha = [\alpha_{ij}]$ и $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_n \end{bmatrix}$, рассмотрим процесс Зейделя

$$x_i^{(k)} = \sum_{j=1}^{i-1} \alpha_{ij}x_j^{(k)} + \sum_{j=i}^{n} \alpha_{ij}x_j^{(k-1)} + \beta_i \quad (i = 1, 2, \ldots, n;\ k = 1, 2, \ldots)$$

при произвольном начальном векторе

$$\boldsymbol{x}^{(0)} = \begin{bmatrix} x_1^{(0)} \\ \cdot \\ \cdot \\ \cdot \\ x_n^{(0)} \end{bmatrix}.$$

Положим

$$\alpha = B + C,$$

где

$$B = \begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ \alpha_{21} & 0 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \alpha_{n1} & \alpha_{n2} & \dots & \alpha_{n,\,n-1} & 0 \end{bmatrix}, \quad C = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ 0 & \alpha_{22} & \dots & \alpha_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & \alpha_{nn} \end{bmatrix}.$$

Тогда процесс Зейделя в матричном виде можно записать следующим образом:

$$x^{(k)} = Bx^{(k)} + Cx^{(k-1)} + \beta \qquad (k = 1,\ 2,\ \dots). \tag{2}$$

Теорема. *Для сходимости процесса Зейделя* (2) *для системы* (1) *при любом выборе свободного члена* $\beta$ *и начального вектора* $x^{(0)}$ *необходимо и достаточно, чтобы все корни* $\lambda_1, \dots, \lambda_n$ *уравнения*

$$\det[C - (E - B)\lambda] \equiv \begin{vmatrix} \alpha_{11} - \lambda & \alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21}\lambda & \alpha_{22} - \lambda & \dots & \alpha_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \alpha_{n1}\lambda & \alpha_{n2}\lambda & \dots & \alpha_{nn} - \lambda \end{vmatrix} = 0 \tag{3}$$

*были по модулю меньше единицы.*

Доказательство. Из формулы (2) вытекает:

$$(E - B)\, x^{(k)} = Cx^{(k-1)} + \beta. \tag{4}$$

Матрица $E - B$ — неособенная, так как

$$\det(E - B) = 1.$$

Поэтому равенство (4) можно записать в виде

$$x^{(k)} = (E - B)^{-1} Cx^{(k-1)} + (E - B)^{-1}\beta. \tag{5}$$

Отсюда ясно, что процесс Зейделя эквивалентен процессу простой итерации, примененному к линейной системе

$$x = (E - B)^{-1} Cx + (E - B)^{-1}\beta.$$

В силу теоремы предыдущего параграфа, для сходимости итерационного процесса (5) необходимо и достаточно, чтобы корни $\lambda_1, \dots, \lambda_n$ характеристического уравнения

$$\det[(E - B)^{-1} C - \lambda E] = 0 \tag{6}$$

удовлетворяли условиям

$$|\lambda_j| < 1 \qquad (j = 1,\ 2,\ \dots,\ n).$$

Уравнение (6), очевидно, равносильно уравнению (3).

Замечание. Пусть

$$Ax = b. \tag{7}$$

Положим

$$D = \begin{bmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_{nn} \end{bmatrix} \neq 0.$$

Для применения метода Зейделя систему (7) обычно записывают в форме

$$Dx = (D - A)x + b$$

или

$$x = D^{-1}(D - A)x + D^{-1}b. \tag{8}$$

Положим

$$A - D = B_1 + C_1,$$

где

$$B_1 = \begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ a_{21} & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{n,\,n-1} & 0 \end{bmatrix}$$

и

$$C_1 = \begin{bmatrix} 0 & a_{12} & \dots & a_{1,\,n-1} & a_{1n} \\ 0 & 0 & \dots & a_{2,\,n-1} & a_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix}.$$

Тогда

$$D^{-1}(D - A) = B + C,$$

где

$$B = -D^{-1}B_1 \quad \text{и} \quad C = -D^{-1}C_1,$$

причем треугольные матрицы $B$ и $C$ реализуют разбиение матрицы системы (8), необходимое для применения процесса Зейделя. На основании формулы (3) сходимость процесса Зейделя для системы (7) зависит от свойств корней уравнения

$$\det[-D^{-1}C_1 - (E + D^{-1}B_1)\lambda] = 0. \tag{9}$$

Уравнение (9) можно заменить эквивалентным уравнением

$$\det[(D + B_1)\lambda + C_1] = 0$$

или

$$\begin{vmatrix} a_{11}\lambda & a_{12} & \dots & a_{1n} \\ a_{21}\lambda & a_{22}\lambda & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1}\lambda & a_{n2}\lambda & \dots & a_{nn}\lambda \end{vmatrix} = 0. \tag{10}$$

Таким образом, для сходимости процесса Зейделя для системы (7) при любом свободном члене $\boldsymbol{b}$ и любом начальном приближении $\boldsymbol{x}^{(0)}$ необходимо и достаточно, чтобы все корни $\lambda_j$ уравнения (10) удовлетворяли условиям

$$|\lambda_j| < 1 \qquad (j = 1, 2, \ldots, n).$$

## § 5. Сходимость процесса Зейделя для нормальной системы

Теорема. *Для нормальной системы обычный процесс Зейделя сходится при любом выборе начального вектора.*

Доказательство. Пусть линейная система

$$A\boldsymbol{x} = \boldsymbol{b} \tag{1}$$

— нормальная, т. е. матрица $A = [a_{ij}]$ — симметрическая и положительно определенная.

Положим

$$A = D + V + V^*,$$

где

$$D = \begin{bmatrix} a_{11} & 0 & \ldots & 0 \\ 0 & a_{22} & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & a_{nn} \end{bmatrix}$$

— диагональная матрица,

$$V = \begin{bmatrix} 0 & 0 & \ldots & 0 \\ a_{21} & 0 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & 0 \end{bmatrix}$$

— нижняя треугольная матрица,

$$V^* = \begin{bmatrix} 0 & a_{12} & \ldots & a_{1n} \\ 0 & 0 & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & 0 \end{bmatrix}$$

— верхняя треугольная матрица, являющаяся ввиду симметричности матрицы $A$ транспонированной по отношению к матрице $V$. Тогда имеем:

$$(D + V + V^*)\,\boldsymbol{x} = \boldsymbol{b}.$$

Отсюда

$$D\boldsymbol{x} = \boldsymbol{b} - (V + V^*)\,\boldsymbol{x}$$

и, следовательно,

$$\boldsymbol{x} = D^{-1}\boldsymbol{b} - D^{-1}(V + V^*)\,\boldsymbol{x}, \tag{2}$$

где

$$D^{-1}=\begin{bmatrix} \frac{1}{a_{11}} & 0 & \dots & 0 \\ 0 & \frac{1}{a_{22}} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{1}{a_{nn}} \end{bmatrix}.$$

Согласно предыдущему процесс Зейделя для системы (1) или эквивалентной ей системы (2) строится следующим образом:

$$x^{(k)}=D^{-1}b+Bx^{(k)}+Cx^{(k-1)} \qquad (k=1,\ 2,\ \dots), \tag{3}$$

где

$$B=-D^{-1}V \quad \text{и} \quad C=-D^{-1}V^*.$$

Для сходимости процесса в силу теоремы предыдущего параграфа необходимо и достаточно, чтобы все собственные значения $\lambda$ матрицы

$$M=(E-B)^{-1}C$$

были по модулю меньше единицы.

В нашем случае имеем:

$$M=-(E+D^{-1}V)^{-1}D^{-1}V^*=-[D^{-1}(D+V)]^{-1}D^{-1}V^*=$$
$$=-(D+V)^{-1}DD^{-1}V^*=-(D+V)^{-1}V^*.$$

Пусть $e$ — единичный собственный вектор матрицы $M$, соответствующий собственному значению $\lambda$, т. е.

$$(D+V)^{-1}V^*e=-\lambda e$$

или

$$V^*e=-\lambda(D+V)e.$$

Отсюда

$$(V^*e,\ e)=-\lambda[(D+V)e,\ e]$$

и, следовательно,

$$\lambda=-\frac{(V^*e,\ e)}{(De,\ e)+(Ve,\ e)}.$$

Введем обозначения

$$(De,\ e)=\sum_{j=1}^{n}a_{jj}|e_j|^2=\sigma>0$$

и

$$(Ve,\ e)=\alpha+i\beta$$

($\alpha$ и $\beta$ действительны и $i=\sqrt{-1}$).

Ввиду того, что матрица $V^*$ является транспонированной по отношению к матрице $V$, получаем:

$$(V^*e,\ e)=(e,\ Ve)=(Ve,\ e)^*=\alpha-i\beta.$$

Поэтому

$$\lambda=-\frac{\alpha-i\beta}{(\sigma+\alpha)+i\beta}$$

и, следовательно,

$$|\lambda|=\frac{\sqrt{\alpha^2+\beta^2}}{\sqrt{(\sigma+\alpha)^2+\beta^2}}. \tag{4}$$

Используя положительную определенность матрицы $A$, будем иметь:

$$(Ae,\ e)=(De,\ e)+(Ve,\ e)+(V^*e,\ e)=$$
$$=\sigma+(\alpha+i\beta)+(\alpha-i\beta)=\sigma+2\alpha>0,$$

т. е.

$$\sigma+\alpha>-\alpha. \tag{5}$$

Далее, учитывая положительность числа $\sigma$, очевидно, имеем:

$$\sigma+\alpha>\alpha.$$

Таким образом, всегда выполнено неравенство

$$\sigma+\alpha>|\alpha|. \tag{6}$$

Отсюда для членов дроби (4) будем иметь:

$$\sqrt{(\sigma+\alpha)^2+\beta^2}>\sqrt{\alpha^2+\beta^2}\geqslant 0,$$

т. е.

$$|\lambda|<1,$$

что и требовалось доказать.

## § 6. Способы эффективной проверки условий сходимости

Для проверки условий теорем сходимости итерационных процессов нужно иметь эффективные критерии, позволяющие определять, удовлетворяют ли корни $\lambda_1, \lambda_2, \ldots, \lambda_n$ данного алгебраического полинома

$$f(\lambda)=p_0\lambda^n+p_1\lambda^{n-1}+\ldots+p_n \tag{1}$$

требованию

$$|\lambda_j|<1 \qquad (j=1,\ 2,\ \ldots,\ n) \tag{2}$$

или не удовлетворяют ему. Этот вопрос можно просто разрешить, воспользовавшись известными *условиями Гурвица* [2].

Теорема Гурвица. *Пусть коэффициенты* $p_k$ $(k=0, 1, \ldots, n)$ *полинома* (1) *действительны, причем*

$$p_0 > 0$$

*и*

$$M=\begin{bmatrix} p_1 & p_0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ p_3 & p_2 & p_1 & p_0 & \ldots & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \ldots & p_n & p_{n-1} & p_{n-2} \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & p_n \end{bmatrix}$$

*— матрица n-го порядка, строки которой представляют собой расширенные последовательности коэффициентов полинома* (1)

$$p_{2m-1},\ p_{2m-2},\ \ldots,\ p_{2m-n},$$

*где положено* $p_k=0$ *при* $k<0$ *и* $k>n$. *Тогда все корни* $\lambda_1$, $\lambda_2$, $\ldots$, $\lambda_n$ *полинома* (1) *имеют отрицательные вещественные части*

$$\operatorname{Re}\lambda_j < 0 \qquad (j=1,\ 2,\ \ldots,\ n)$$

(*т. е. расположены в левой полуплоскости комплексной плоскости* $\lambda=\alpha+i\beta$) *в том и только том случае, если главные диагональные миноры матрицы M положительны, т. е.*

$$\left.\begin{aligned} &\Delta_1 = p_1 > 0, \\ &\Delta_2 = \begin{vmatrix} p_1 & p_0 \\ p_3 & p_2 \end{vmatrix} > 0, \\ &\cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ &\Delta_n = p_n\Delta_{n-1} > 0. \end{aligned}\right\} \tag{3}$$

Пример 1. Для квадратного трехчлена

$$p_0\lambda^2+p_1\lambda+p_2$$

условия Гурвица таковы:

$$p_0>0,\ p_1>0,\ p_2>0.$$

Нас интересует вопрос, когда корни полинома (1) удовлетворяют условию (2), т. е. лежат на комплексной плоскости $\lambda$ внутри единичного круга

$$|\lambda|<1.$$

С помощью дробно-линейной функции

$$\lambda=\frac{\mu+1}{\mu-1}$$

внутренность единичного круга $|\lambda| < 1$ преобразуется в левую полуплоскость $\operatorname{Re} \mu < 0$. При этом наш полином (1) принимает вид

$$f\left(\frac{\mu+1}{\mu-1}\right) = p_0\left(\frac{\mu+1}{\mu-1}\right)^n + p_1\left(\frac{\mu+1}{\mu-1}\right)^{n-1} + \ldots + p_n =$$
$$= \frac{1}{(\mu-1)^n}\left[p_0(\mu+1)^n + p_1(\mu+1)^{n-1}(\mu-1) + \ldots + p_n(\mu-1)^n\right].$$

Следовательно, корни полинома (1) тогда и только тогда расположены внутри единичного круга, когда вспомогательный полином

$$F(\mu) = \pm\left[p_0(\mu+1)^n + p_1(\mu+1)^{n-1}(\mu-1) + \ldots + p_n(\mu-1)^n\right]$$

удовлетворяет условиям Гурвица (3), причем знак выбирается так, чтобы старший коэффициент

$$\pm(p_0 + p_1 + \ldots + p_n) > 0.$$

Пример 2. Рассмотрим квадратный трехчлен

$$f(\lambda) = \lambda^2 + p\lambda + q, \tag{4}$$

где $p$ и $q$ действительны. Вспомогательный полином имеет вид

$$F(\mu) = \pm\left[(\mu+1)^2 + p(\mu+1)(\mu-1) + q(\mu-1)^2\right] =$$
$$= \pm\left[(1+p+q)\mu^2 + 2(1-q)\mu + (1-p+q)\right].$$

Из условий Гурвица получаем:

$$\left.\begin{aligned} &\pm(1+p+q) > 0, \\ &\pm(1-q) > 0, \\ &\pm(1-p+q) > 0. \end{aligned}\right\}$$

Рассмотрим случаи:

а) $q < 1$, тогда $q > -p-1$ и $q > p-1$;

б) $q > 1$, тогда $q < -p-1$ и $q < p-1$, что невозможно.

Следовательно, уравнение (4) имеет корни $\lambda_1$, $\lambda_2$, по модулю меньшие единицы, тогда и только тогда, когда

$$|p| < 1+q, \quad |q| < 1. \tag{5}$$

Так как при $n=2$ характеристическое уравнение матрицы $\alpha$ имеет вид

$$\begin{vmatrix} \alpha_{11}-\lambda & \alpha_{12} \\ \alpha_{21} & \alpha_{22}-\lambda \end{vmatrix} = 0,$$

или

$$\lambda^2 - (\alpha_{11}+\alpha_{22})\lambda + \det\alpha = 0,$$

то для сходимости соответствующего процесса итерации для системы двух уравнений необходимо, чтобы

$$|\det\alpha| < 1.$$

Области сходимости процесса обычной итерации и процесса Зейделя, вообще говоря, перекрываются. Можно привести примеры линейных систем, для которых процесс обычной итерации сходится, а процесс Зейделя расходится, и обратно [3].

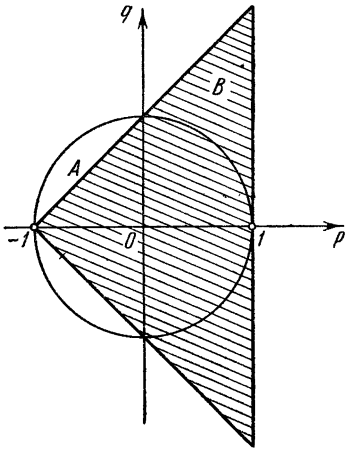


Рис. 57.

Пример 3. Рассмотрим линейную систему

$$x = \alpha x + \beta \tag{6}$$

с кососимметрической матрицей

$$\alpha = \begin{bmatrix} p & q \\ -q & p \end{bmatrix}$$

($p$ и $q$ действительны).

Характеристическое уравнение матрицы $\alpha$ имеет вид

$$\begin{vmatrix} p-\lambda & q \\ -q & p-\lambda \end{vmatrix} = 0$$

или

$$(\lambda - p)^2 + q^2 = 0.$$

Отсюда

$$\lambda_{1,2} = p \pm iq.$$

Для сходимости метода обычной итерации необходимо и достаточно, чтобы

$$|\lambda_{1,2}| = \sqrt{p^2+q^2} < 1,$$

т. е.

$$p^2 + q^2 < 1$$

(область $A$ на рис. 57).

Для метода Зейделя уравнение, определяющее сходимость, имеет вид

$$\begin{vmatrix} p-\lambda & q \\ -q\lambda & p-\lambda \end{vmatrix} = 0$$

или

$$\lambda^2 - (2p - q^2)\lambda + p^2 = 0. \tag{7}$$

На основании результатов примера 2 для того, чтобы корни $\lambda_1$ и $\lambda_2$ уравнения (7) удовлетворяли условиям

$$|\lambda_1| < 1, \qquad |\lambda_2| < 1,$$

необходимо и достаточно выполнение неравенств

$$|2p - q^2| < 1 + p^2, \qquad p^2 < 1,$$

откуда

$$|p| < 1, \qquad |q| < 1 + p$$

(область $B$ на рис. 57). Так как области $A$ и $B$ частично перекрываются, то отсюда следует, что для системы (6) можно выбрать коэффициенты $p$ и $q$, во-первых, так, чтобы метод итерации сходился, а метод Зейделя расходился (например, $p=-0{,}5$; $q=0{,}6$), и, во-вторых, так, чтобы, наоборот, метод Зейделя сходился, а метод итерации расходился (например, $p=0{,}5$; $q=1$).

## Литература к одиннадцатой главе

1. В. И. Смирнов, Курс высшей математики, т. 3, ГТТИ, М.—Л., 1933, гл. VII.
2. А. Г. Курош, Курс высшей алгебры, Гостехиздат, М. —Л., 1946, гл. VIII.
3. В. Н. Фаддеева, Вычислительные методы линейной алгебры, Гостехиздат, М. — Л., 1950, гл. II.

ГЛАВА XII

# НАХОЖДЕНИЕ СОБСТВЕННЫХ ЗНАЧЕНИЙ И СОБСТВЕННЫХ ВЕКТОРОВ МАТРИЦЫ

## § 1. Вводные замечания

При решении теоретических и практических задач часто возникает надобность определить собственные значения данной матрицы $A$, т. е. вычислить корни ее *векового* (*характеристического*) уравнения

$$\det(A-\lambda E)=0, \tag{1}$$

а также найти соответствующие собственные векторы матрицы $A$. Вторая задача является более простой, так как если корни характеристического уравнения известны, то нахождение собственных векторов сводится к отысканию ненулевых решений некоторых однородных линейных систем. Поэтому мы в первую очередь будем заниматься первой задачей—вычислением корней характеристического уравнения (1).

Здесь в основном применяются два приема: 1) развертывание векового определителя в полином $n$-й степени

$$D(\lambda)=\det(A-\lambda E)$$

с последующим решением уравнения $D(\lambda)=0$ одним из известных приближенных, вообще говоря, способов (например, методом Лобачевского—Греффе, гл. V, §§ 7—12) и 2) приближенное определение корней характеристического уравнения (чаще всего наибольших по модулю) методом итерации, без предварительного развертывания векового определителя.

В этой главе будут изложены основные методы решения поставленной общей задачи, причем мы начнем с *развертывания вековых определителей*.

## § 2. Развертывание вековых определителей

Как известно, *вековым определителем* матрицы $A=[a_{ij}]$ называется определитель вида

$$D(\lambda)=\det(A-\lambda E)=\begin{vmatrix} a_{11}-\lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22}-\lambda & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn}-\lambda \end{vmatrix}. \tag{1}$$

Приравнивая этот определитель нулю, получаем *характеристическое уравнение*

$$D(\lambda)=0.$$

Если требуется найти все корни характеристического уравнения, то целесообразно предварительно вычислить определитель (1).

Развертывая определитель (1), получаем полином $n$-й степени

$$D(\lambda)=(-1)^n[\lambda^n-\sigma_1\lambda^{n-1}+\sigma_2\lambda^{n-2}-\ldots+(-1)^n\sigma_n], \qquad (2)$$

где

$$\sigma_1=\sum_{\alpha=1}^{n} a_{\alpha\alpha}$$

есть сумма всех диагональных миноров первого порядка матрицы $A$.

$$\sigma_2=\sum_{\alpha<\beta}\begin{vmatrix} a_{\alpha\alpha} & a_{\alpha\beta} \\ a_{\beta\alpha} & a_{\beta\beta}\end{vmatrix}$$

есть сумма всех диагональных миноров второго порядка матрицы $A$;

$$\sigma_3=\sum_{\alpha<\beta<\gamma}\begin{vmatrix} a_{\alpha\alpha} & a_{\alpha\beta} & a_{\alpha\gamma} \\ a_{\beta\alpha} & a_{\beta\beta} & a_{\beta\gamma} \\ a_{\gamma\alpha} & a_{\gamma\beta} & a_{\gamma\gamma}\end{vmatrix}$$

— сумма всех диагональных миноров третьего порядка матрицы $A$ и т. д. Наконец,

$$\sigma_n=\det A.$$

Легко убедиться, что число диагональных миноров $k$-го порядка матрицы $A$ равно

$$C_n^k=\frac{n(n-1)\ldots(n-k+1)}{k!} \qquad (k=1,\ 2,\ \ldots,\ n).$$

Отсюда получаем, что непосредственное вычисление коэффициентов характеристического полинома (2) эквивалентно вычислению

$$C_n^1+C_n^2+\ldots+C_n^n=2^n-1$$

определителей различных порядков. Последняя задача, вообще говоря, технически трудно осуществима для сколько-нибудь больших значений $n$. Поэтому созданы специальные методы развертывания вековых определителей (методы А. Н. Крылова, А. М. Данилевского, Леверье, метод неопределенных коэффициентов, метод интерполяции и др.) (см. [1]). В следующих параграфах мы изложим некоторые из этих методов.

## § 3. Метод А. М. Данилевского

Сущность метода А. М. Данилевского [1] заключается в приведении векового определителя к так называемому *нормальному виду Фробениуса*

$$D(\lambda)=\begin{vmatrix} p_1-\lambda & p_2 & p_3 & \dots & p_n \\ 1 & -\lambda & 0 & \dots & 0 \\ 0 & 1 & -\lambda & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & -\lambda \end{vmatrix}. \tag{1}$$

Если нам удалось записать вековой определитель в форме (1), то, разлагая его по элементам первой строки, будем иметь:

$$D(\lambda)=(p_1-\lambda)(-\lambda)^{n-1}-p_2(-\lambda)^{n-2}+p_3(-\lambda)^{n-3}-\dots+(-1)^{n-1}p_n$$

или

$$D(\lambda)=(-1)^n(\lambda^n-p_1\lambda^{n-1}-p_2\lambda^{n-2}-p_3\lambda^{n-3}-\dots-p_n). \tag{2}$$

Таким образом, развертывание векового определителя, записанного в нормальной форме (1), не представляет затруднений. Обозначим через

$$A=\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

данную матрицу, а через

$$P=\begin{bmatrix} p_1 & p_2 & \dots & p_{n-1} & p_n \\ 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}$$

— подобную ей *матрицу Фробениуса*, т. е.

$$P=S^{-1}AS,$$

где $S$ — неособенная матрица.

Так как подобные матрицы обладают одинаковыми характеристическими полиномами, то имеем:

$$\det(A-\lambda E)=\det(P-\lambda E). \tag{3}$$

Поэтому для обоснования метода достаточно показать, каким образом, исходя из матрицы $A$, строится матрица $P$. Согласно методу А. М. Данилевского, переход от матрицы $A$ к подобной ей матрице $P$ осуществляется с помощью $n-1$ преобразований подобия, последовательно преобразующих строки матрицы $A$, начиная с последней, в соответствующие строки матрицы $P$.

Покажем начало процесса. Нам нужно строку

$$a_{n1}\ a_{n2}\ \dots\ a_{n,\,n-1}\ a_{nn}$$

перевести в строку 0 0 ... 1 0. Предполагая, что $a_{n,\,n-1} \neq 0$, разделим все элементы $(n-1)$-го столбца матрицы $A$ на $a_{n,\,n-1}$. Тогда ее $n$-я строка примет вид

$$a_{n1}\, a_{n2} \ldots 1\; a_{nn}.$$

Затем вычтем $(n-1)$-й столбец преобразованной матрицы, умноженный соответственно на числа $a_{n1}, a_{n2}, \ldots, a_{nn}$, из всех остальных ее столбцов.

В результате получим матрицу, последняя строка которой имеет желаемый вид 0 0 ... 1 0. Указанные операции являются элементарными преобразованиями, производимыми над столбцами матрицы $A$. Произведя эти же преобразования над единичной матрицей, получим матрицу

$$M_{n-1} = \begin{bmatrix} 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ m_{n-1,\,1} & m_{n-1,\,2} & \ldots & m_{n-1,\,n-1} & m_{n-1,\,n} \\ 0 & 0 & \ldots & 0 & 1 \end{bmatrix},$$

где

$$m_{n-1,\,i} = -\frac{a_{ni}}{a_{n,\,n-1}} \quad \text{при} \quad i \neq n-1 \tag{4}$$

и

$$m_{n-1,\,n-1} = \frac{1}{a_{n,\,n-1}}. \tag{4'}$$

Отсюда заключаем (см. гл. VII, § 14), что произведенные операции равносильны умножению справа матрицы $M_{n-1}$ на матрицу $A$, т. е. после указанных преобразований получим матрицу

$$AM_{n-1} = B = \begin{bmatrix} b_{11} & b_{12} & \ldots & b_{1,\,n-1} & b_{1,\,n} \\ b_{21} & b_{22} & \ldots & b_{2,\,n-1} & b_{2,\,n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ b_{n-1,\,1} & b_{n-1,\,2} & \ldots & b_{n-1,\,n-1} & b_{n-1,\,n} \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix}. \tag{5}$$

Используя правило умножения матриц, находим, что элементы матрицы $B$ вычисляются по следующим формулам:

$$b_{ij} = a_{ij} + a_{i,\,n-1} m_{n-1,\,j} \quad \text{при} \quad 1 \leqslant i \leqslant n;\; j \neq n-1; \tag{6}$$

$$b_{i,\,n-1} = a_{i,\,n-1} m_{n-1,\,n-1} \quad \text{при} \quad 1 \leqslant i \leqslant n. \tag{6'}$$

Однако построенная матрица $B = AM_{n-1}$ не будет подобна матрице $A$. Чтобы иметь преобразование подобия, нужно обратную матрицу $M_{n-1}^{-1}$ слева умножить на матрицу $B$:

$$M_{n-1}^{-1} A M_{n-1} = M_{n-1}^{-1} B.$$

Непосредственной проверкой легко убедиться, что обратная матрица $M_{n-1}^{-1}$ имеет вид

$$M_{n-1}^{-1}=\begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ a_{n1} & a_{n2} & \dots & a_{n,\,n-1} & a_{nn} \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix}. \tag{7}$$

Пусть

$$M_{n-1}^{-1}AM_{n-1}=C.$$

Следовательно,

$$C=M_{n-1}^{-1}B. \tag{8}$$

Так как, очевидно, умножение слева матрицы $M_{n-1}^{-1}$ на матрицу $B$ не изменяет преобразованной строки последней, то матрица $C$ имеет вид

$$C=\begin{bmatrix} c_{11} & c_{12} & \dots & c_{1,\,n-1} & c_{1n} \\ c_{21} & c_{22} & \dots & c_{2,\,n-1} & c_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ c_{n-1,\,1} & c_{n-1,\,2} & \dots & c_{n-1,\,n-1} & c_{n-1,\,n} \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}. \tag{9}$$

Перемножая матрицы $M_{n-1}^{-1}$ (7) и $B$ (5), будем иметь:

$$c_{ij}=b_{ij} \quad \text{при} \quad 1\leqslant i\leqslant n-2 \tag{10}$$

и

$$c_{n-1,\,j}=\sum_{k=1}^{n} a_{nk}b_{kj} \quad \text{при} \quad 1\leqslant j\leqslant n. \tag{10'}$$

Таким образом, умножение $M_{n-1}^{-1}$ на матрицу $B$ меняет лишь $(n-1)$-ю строку матрицы $B$. Элементы этой строки находятся по формулам (10) и (10′). Полученная матрица $C$ подобна матрице $A$ и имеет одну приведенную строку. Этим заканчивается первый этап процесса.

Далее, если $c_{n-1,\,n-2}\neq 0$, то над матрицей $C$ можно повторить аналогичные операции, взяв за основную $(n-2)$-ю ее строку. В результате получим матрицу

$$D=M_{n-2}^{-1}CM_{n-2}$$

с двумя приведенными строками. Над последней матрицей проделываем те же операции. Продолжая этот процесс, мы, наконец, получим матрицу Фробениуса

$$P=M_1^{-1}\ \dots\ M_{n-2}^{-1}M_{n-1}^{-1}AM_{n-1}M_{n-2}\ \dots\ M_1,$$

если, конечно, все $n-1$ промежуточных преобразований возможны.

Весь процесс может быть оформлен в удобную вычислительную схему, составление которой покажем на следующем примере.

Пример. Привести к виду Фробениуса матрицу

$$A=\begin{bmatrix}1 & 2 & 3 & 4\\ 2 & 1 & 2 & 3\\ 3 & 2 & 1 & 2\\ 4 & 3 & 2 & 1\end{bmatrix}.$$

Решение. Вычисления располагаем в таблицу 25.

В строках 1—4 таблицы 25 помещаем элементы $a_{ij}$ ($i$, $j=1$, 2, 3, 4) данной матрицы и контрольные суммы $a_{i5}=\sum_{j=1}^{4} a_{ij}$ ($i=1$, 2, 3, 4) ($\Sigma$). Отмечаем элемент $a_{43}=2$, принадлежащий третьему столбцу (*отмеченный столбец*). В строке I записываем элементы третьей строки матрицы $M_{n-1}=M_3$, вычисляемые по формулам (4) и (4′):

$$m_{31}=-\frac{a_{41}}{a_{43}}=-\frac{4}{2}=-2;$$
$$m_{32}=-\frac{a_{42}}{a_{43}}=-\frac{3}{2}=-1{,}5;$$
$$m_{33}=\frac{1}{a_{43}}=\frac{1}{2}=0{,}5;$$
$$m_{34}=-\frac{a_{44}}{a_{43}}=-\frac{1}{2}=-0{,}5.$$

Сюда же (строка I таблицы 25) помещаем элемент

$$m_{35}=-\frac{a_{45}}{a_{43}}=-\frac{10}{2}=-5,$$

получаемый аналогичным приемом из контрольного столбца $\Sigma$. Число —5 должно совпасть с суммой элементов строки I, не входящих в контрольный столбец (после замены элемента $m_{33}$ на —1). Для удобства число —1 записываем рядом с элементом $m_{33}$, отделяя от последнего чертой.

В строках 5—8 в графе $M^{-1}$ выписываем третью строку матрицы $M^{-1}$, которая в силу формулы (7) совпадает с четвертой строкой исходной матрицы $A$. В строках 5—8 в соответствующих столбцах выписываем элементы матрицы

$$B=AM_3,$$

вычисляемые по двучленным формулам (6) для неотмеченных столбцов и по одночленной формуле (6′) для отмеченного столбца. Например, для первого столбца имеем:

$$b_{11}=1+3(-2)=-5;$$
$$b_{21}=2+2(-2)=-2;$$
$$b_{31}=3+1(-2)=1;$$
$$b_{41}=4+2(-2)=0$$

и т. д.

Таблица 25

**Вычислительная схема метода А. М. Данилевского**

| Номер строки | $M^{-1}$ | Столбцы матрицы | | | | Σ | Σ′ |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | | |
| 1 | | 1 | 2 | 3 | 4 | 10 | |
| 2 | | 2 | 1 | 2 | 3 | 8 | |
| 3 | | 3 | 2 | 1 | 2 | 8 | |
| 4 | | 4 | 3 | $\boxed{2}$ | 1 | 10 | |
| I | $M_3^{-1}$ \| $M_3$ | —2 | —1,5 | 0,5 \| —1 | —0,5 | —5 | |
| 5 | $A$ | —5 | —2,5 | 1,5 | 2,5 | —3,5 | —5 |
| 6 | 3 | 2 | —2 | 1 | 2 | —1 | —2 |
| $\boxed{7}$ | 2 | 1 | 0,5 | 0,5 | 1,5 | 3,5 | 3 |
| 8 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| $\boxed{7'}$ | | —24 | $\boxed{-15}$ | 11 | 19 | —9 | |
| II | $M_2^{-1}$ \| $M_2$ | —1,600 | —0,067 \| —1 | 0,733 | 1,267 | —0,600 | |
| 9 | —24 | —1 | 0,167 | —0,333 | —0,667 | —1,833 | —2 |
| $\boxed{10}$ | —15 | 1,2 | 0,133 | —0,467 | —0,533 | 0,333 | 0,2 |
| 11 | 11 | 0 | 1 | 0 | 0 | 1 | 0 |
| 12 | 19 | 0 | 0 | 1 | 0 | 1 | 1 |
| $\boxed{10'}$ | | $\boxed{6}$ | 5 | 34 | 24 | 69 | |
| III | $M_1^{-1}$ \| $M_1$ | 0,167 \| —1 | —0,833 | —5,667 | —4,000 | —11,500 | |
| $\boxed{13}$ | 6 | —0,167 | 1 | 5,333 | 3,333 | 9,500 | 9,667 |
| 14 | 5 | 1 | 0 | 0 | 0 | 1 | 0 |
| 15 | 34 | 0 | 1 | 0 | 0 | 1 | 1 |
| 16 | 24 | 0 | 0 | 1 | 0 | 1 | 1 |
| $\boxed{13'}$ | | 4 | 40 | 56 | 20 | 120 | |

Преобразованные элементы третьего (отмеченного) столбца получаются с помощью умножения исходных элементов на $m_{33}=0{,}5$. Например,

$$b_{13}=3\cdot 0{,}5=1{,}5;$$
$$b_{23}=2\cdot 0{,}5=1;$$
$$b_{33}=1\cdot 0{,}5=0{,}5;$$
$$b_{43}=2\cdot 0{,}5=1.$$

Заметим, что последняя строка матрицы $B$ должна иметь вид

$$0\ \ 0\ \ 1\ \ 0.$$

Для контроля пополняем матрицу $B$ преобразованными по аналогичным двучленным формулам с $m_{35}=-5$ соответствующими элементами столбца $\Sigma$. Например,

$$b_{16}=10+3\cdot(-5)=-5;$$
$$b_{26}=8+2\cdot(-5)=-2;$$
$$b_{36}=8+1\cdot(-5)=3;$$
$$b_{46}=10+2\cdot(-5)=0.$$

Полученные результаты записываем в столбце $\Sigma'$ в соответствующих строках. Прибавив к ним элементы третьего столбца, будем иметь контрольные суммы

$$b_{i5}=\sum_{j=1}^{4} b_{ij} \qquad (i=1,\ 2,\ 3,\ 4)$$

для строк 5—8 (столбец $\Sigma$).

Преобразование $M_3^{-1}$, произведенное над матрицей $B$ и дающее матрицу $C=M_3^{-1}B$, изменяет лишь третью строку матрицы $B$, т. е. седьмую строку таблицы. Элементы этой преобразованной строки $7'$ получаются по формуле (10), т. е. представляют собой суммы парных произведений элементов столбца $M^{-1}$, находящихся в строках 5—8, на соответствующие элементы каждого из столбцов матрицы $B$. Например,

$$c_{31}=4(-5)+3(-2)+2\cdot 1=-24$$

и т. д.

Те же преобразования производим над столбцом $\Sigma$:

$$c_{35}=4(-3{,}5)+3(-1)+2\cdot 3{,}5+1\cdot 1=-9.$$

В результате получаем матрицу $C$, состоящую из строк 5, 6, $7'$, 8 с контрольными суммами $\Sigma$, причем матрица $C$ подобна матрице $A$ и имеет одну приведенную строку 8. Этим заканчивается построение первого подобного преобразования $C=M_3^{-1}AM_3$.

Далее, приняв матрицу $C$ за исходную и выделив элемент $c_{22} = -15$ (второй столбец), продолжаем процесс аналогичным образом. В результате получаем матрицу $D = M_2^{-1}CM_2$, элементы которой расположены в строках 9, 10′, 11, 12, содержащую две приведенные строки. Наконец, отправляясь от элемента $d_{21} = 6$ (первый столбец) и преобразуя матрицу $D$ в подобную ей, получаем искомую матрицу Фробениуса $P$, элементы которой записаны в строках 13′, 14, 15, 16. На каждом этапе процесса контроль осуществляется с помощью столбцов $\Sigma$ и $\Sigma'$.

Таким образом, матрица Фробениуса будет:

$$P = \begin{bmatrix} 4 & 40 & 56 & 20 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

Отсюда вековой определитель, приведенный к нормальному виду Фробениуса, запишется так:

$$D(\lambda) = \begin{vmatrix} 4-\lambda & 40 & 56 & 20 \\ 1 & -\lambda & 0 & 0 \\ 0 & 1 & -\lambda & 0 \\ 0 & 0 & 1 & -\lambda \end{vmatrix}$$

или

$$D(\lambda) = \lambda^4 - 4\lambda^3 - 40\lambda^2 - 56\lambda - 20.$$

## § 4. Исключительные случаи в методе А. М. Данилевского

Процесс А. М. Данилевского происходит без всяких осложнений, если все выделяемые элементы отличны от нуля. Мы остановимся сейчас на исключительных случаях, когда это требование нарушается.

Допустим, что при преобразовании матрицы $A$ в матрицу Фробениуса $P$ мы после нескольких шагов пришли к матрице вида

$$D = \begin{bmatrix} d_{11} & d_{12} & \dots & d_{1k} & \dots & d_{1,\,n-1} & d_{1n} \\ d_{21} & d_{22} & \dots & d_{2k} & \dots & d_{2,\,n-1} & d_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ d_{k1} & d_{k2} & \dots & d_{kk} & \dots & d_{k\,n-1} & d_{kn} \\ 0 & 0 & \dots & 1 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & 0 & \dots & 1 & 0 \end{bmatrix},$$

причем оказалось, что $d_{k,\,k-1} = 0$.

Тогда продолжать преобразование по методу А. М. Данилевского нельзя. Здесь возможны два случая.

1. Пусть какой-то элемент матрицы $D$, стоящий левее нулевого элемента $d_{k,k-1}$, отличен от нуля, т. е. $d_{k,l} \neq 0$, где $l < k-1$. Тогда этот элемент выдвигаем на место нулевого элемента $d_{k,k-1}$, т. е. переставляем $(k-1)$-й и $l$-й столбцы матрицы $D$ и одновременно переставляем ее $(k-1)$-ю и $l$-ю строки. Можно доказать, что полученная новая матрица $D'$ будет подобна прежней. К новой матрице применяем метод А. М. Данилевского.

2. Пусть $d_{kl} = 0$ $(l = 1, 2, \ldots, k-1)$, тогда $D$ имеет вид

$$D = \left[\begin{array}{ccc|ccc} & (D_1) & & & (L) & \\ c_{11} & c_{12} & \ldots\ c_{1,k-1} & c_{1k} & \ldots\ c_{1,n-1} & c_{1n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ c_{k-1,1} & c_{k-1,2} & \ldots\ c_{k-1,k-1} & c_{k-1,k} & \ldots\ c_{k-1,n-1} & c_{k-1,n} \\ \hline 0 & 0 & \ldots\ 0 & c_{kk} & \ldots\ c_{k,n-1} & c_{kn} \\ 0 & 0 & \ldots\ 0 & 1 & \ldots\ 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \ldots\ 0 & 0 & \ldots\ 1 & 0 \\ & (0) & & & (D_2) & \end{array}\right] =$$

$$= \left[\begin{array}{c|c} D_1 & L \\ \hline 0 & D_2 \end{array}\right].$$

В таком случае вековой определитель $\det(D - \lambda E)$ распадается на два определителя

$$\det(D - \lambda E) = \det(D_1 - \lambda E)\det(D_2 - \lambda E).$$

При этом матрица $D_2$ уже приведена к канонической форме Фробениуса и поэтому $\det(D_2 - \lambda E)$ вычисляется сразу. Остается применить метод А. М. Данилевского к матрице $D_1$.

## § 5. Вычисление собственных векторов по методу А. М. Данилевского

Метод А. М. Данилевского дает возможность определять собственные векторы данной матрицы $A$, если известны ее собственные значения. Пусть $\lambda$ — собственное значение матрицы $A$, а следовательно, и собственное значение подобной ей матрицы Фробениуса $P$.

Найдем собственный вектор $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ матрицы $P$, соответствующий данному значению $\lambda$: $P\boldsymbol{y} = \lambda\boldsymbol{y}$. Отсюда $(P - \lambda E)\boldsymbol{y} = \boldsymbol{0}$ или

$$\begin{bmatrix} p_1 - \lambda & p_2 & p_3 & \ldots & p_n \\ 1 & -\lambda & 0 & \ldots & 0 \\ 0 & 1 & -\lambda & \ldots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \ldots & -\lambda \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \boldsymbol{0}.$$

Перемножая матрицы, получим систему для определения координат $y_1, y_2, \ldots, y_n$ собственного вектора $y$:

$$\left.\begin{aligned} (p_1-\lambda)y_1+p_2y_2+\ldots+p_ny_n&=0,\\ y_1-\lambda y_2 &=0,\\ y_2-\lambda y_3 &=0,\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ y_{n-1}-\lambda y_n&=0. \end{aligned}\right\} \qquad (1)$$

Система (1) — однородная. С точностью до коэффициента пропорциональности решения ее могут быть найдены следующим образом. Положим $y_n=1$. Тогда последовательно получим:

$$\left.\begin{aligned} y_{n-1}&=\lambda,\\ y_{n-2}&=\lambda^2,\\ \ldots\ldots\ldots\\ y_1&=\lambda^{n-1}. \end{aligned}\right\} \qquad (2)$$

Таким образом, искомый собственный вектор есть

$$y=\begin{bmatrix}\lambda^{n-1}\\ \lambda^{n-2}\\ \cdot\\ \cdot\\ \cdot\\ 1\end{bmatrix}.$$

Обозначим теперь через $x$ собственный вектор матрицы $A$, соответствующий значению $\lambda$. Тогда, очевидно, имеем:

$$x=M_{n-1}M_{n-2}\ldots M_2M_1y.$$

Преобразование $M_1$, совершенное над $y$, дает:

$$M_1y=\begin{bmatrix} m_{11} & m_{12} & \ldots & m_{1n}\\ 0 & 1 & \ldots & 0\\ \ldots & \ldots & \ldots & \ldots\\ 0 & 0 & \ldots & 1\end{bmatrix}\begin{bmatrix}y_1\\ y_2\\ \cdot\\ \cdot\\ \cdot\\ y_n\end{bmatrix}=\begin{bmatrix}\sum\limits_{k=1}^{n} m_{1k}y_k\\ y_2\\ \cdot\\ \cdot\\ \cdot\\ y_n\end{bmatrix}=\begin{bmatrix}\sum\limits_{k=1}^{n} m_{1k}y_k\\ \lambda^{n-2}\\ \cdot\\ \cdot\\ \cdot\\ 1\end{bmatrix}.$$

Таким образом, преобразование $M_1$ изменяет лишь первую координату вектора. Аналогично преобразование $M_2$ изменит лишь вторую координату вектора $M_1y$ и т. д. Повторив этот процесс $n-1$ раз, получим искомый собственный вектор $x$ матрицы $A$.

## § 6. Метод А. Н. Крылова

Приведем метод развертывания векового определителя, принадлежащий А. Н. Крылову [2] и основанный на существенно иной идее, чем метод А. М. Данилевского.

Пусть

$$D(\lambda) \equiv \det(\lambda E - A) = \lambda^n + p_1\lambda^{n-1} + \ldots + p_n \tag{1}$$

— характеристический полином (с точностью до знака) матрицы $A$. Согласно тождеству Гамильтона—Кели (гл. XI, § 2), матрица $A$ обращает в нуль свой характеристический полином; поэтому

$$A^n + p_1 A^{n-1} + \ldots + p_n E = 0. \tag{2}$$

Возьмем теперь произвольный ненулевой вектор

$$\mathbf{y}^{(0)} = \begin{bmatrix} y_1^{(0)} \\ \cdot \\ \cdot \\ \cdot \\ y_n^{(0)} \end{bmatrix}.$$

Умножая обе части равенства (2) справа на $\mathbf{y}^{(0)}$, получим:

$$A^n\mathbf{y}^{(0)} + p_1 A^{n-1}\mathbf{y}^{(0)} + \ldots + p_n\mathbf{y}^{(0)} = \mathbf{0}. \tag{3}$$

Положим:

$$A^k\mathbf{y}^{(0)} = \mathbf{y}^{(k)} \qquad (k = 1, 2, \ldots, n); \tag{4}$$

тогда равенство (3) приобретает вид

$$\mathbf{y}^{(n)} + p_1\mathbf{y}^{(n-1)} + \ldots + p_n\mathbf{y}^{(0)} = \mathbf{0} \tag{5}$$

или

$$\begin{bmatrix} y_1^{(n-1)} & y_1^{(n-2)} & \cdot & y_1^{(0)} \\ y_2^{(n-1)} & y_2^{(n-2)} & \cdot & y_2^{(0)} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ y_n^{(n-1)} & y_n^{(n-2)} & \cdot & y_n^{(0)} \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \cdot \\ \cdot \\ \cdot \\ p_n \end{bmatrix} = -\begin{bmatrix} y_1^{(n)} \\ y_2^{(n)} \\ \cdot \\ \cdot \\ \cdot \\ y_n^{(n)} \end{bmatrix}, \tag{5'}$$

где

$$\mathbf{y}^{(k)} = \begin{bmatrix} y_1^{(k)} \\ y_2^{(k)} \\ \cdot \\ \cdot \\ \cdot \\ y_n^{(k)} \end{bmatrix} \qquad (k = 0, 1, 2, \ldots, n).$$

Следовательно, векторное равенство (5) эквивалентно системе уравнений

$$p_1 y_j^{(n-1)} + p_2 y_j^{(n-2)} + \ldots + p_n y_j^{(0)} = -y_j^{(n)} \qquad (j = 1, 2, \ldots, n), \tag{6}$$

из которой, вообще говоря, можно определить неизвестные коэффициенты $p_1, p_2, \ldots, p_n$.

Так как на основании формулы (4)

$$y^{(k)} = Ay^{(k-1)}$$

($k = 1, 2, \ldots, n$), то координаты $y_1^{(k)}, y_2^{(k)}, \ldots, y_n^{(k)}$ вектора $y^{(k)}$ последовательно вычисляются по формулам

$$\left.\begin{aligned} y_i^{(1)} &= \sum_{j=1}^{n} a_{ij} y_j^{(0)}, \\ y_i^{(2)} &= \sum_{j=1}^{n} a_{ij} y_j^{(1)}, \\ &\cdots\cdots\cdots \\ y_i^{(n)} &= \sum_{j=1}^{n} a_{ij} y_j^{(n-1)} \qquad (i = 1, 2, \ldots, n). \end{aligned}\right\} \tag{7}$$

Таким образом, определение коэффициентов $p_j$ характеристического полинома (1) методом А. Н. Крылова сводится к решению линейной системы уравнений (6), коэффициенты которой вычисляются по формулам (7), причем координаты начального вектора

$$y^{(0)} = \begin{bmatrix} y_1^{(0)} \\ \cdot \\ \cdot \\ \cdot \\ y_n^{(0)} \end{bmatrix}$$

произвольны. Если система (6) имеет единственное решение, то ее корни $p_1, p_2, \ldots, p_n$ являются коэффициентами характеристического полинома (1). Это решение может быть найдено, например, методом Гаусса (гл. VIII, § 3). Если система (6) не имеет единственного решения, то задача усложняется [1]. В этом случае рекомендуется изменить начальный вектор.

Пример. Методом А. Н. Крылова найти характеристический полином матрицы (см. § 3)

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix}.$$

Решение. Выберем начальный вектор

$$y^{(0)} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Пользуясь формулами (7), определим координаты векторов

$$y^{(k)} = A^k y^{(0)} \qquad (k = 1, 2, 3, 4).$$

Имеем:

$$y^{(1)} = Ay^{(0)} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix};$$

$$y^{(2)} = Ay^{(1)} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 30 \\ 22 \\ 18 \\ 20 \end{bmatrix};$$

$$y^{(3)} = Ay^{(2)} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} 30 \\ 22 \\ 18 \\ 20 \end{bmatrix} = \begin{bmatrix} 208 \\ 178 \\ 192 \\ 242 \end{bmatrix};$$

$$y^{(4)} = Ay^{(3)} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} 208 \\ 178 \\ 192 \\ 242 \end{bmatrix} = \begin{bmatrix} 2108 \\ 1704 \\ 1656 \\ 1992 \end{bmatrix}.$$

Составим систему (6):

$$\begin{bmatrix} y_1^{(3)} & y_1^{(2)} & y_1^{(1)} & y_1^{(0)} \\ y_2^{(3)} & y_2^{(2)} & y_2^{(1)} & y_2^{(0)} \\ y_3^{(3)} & y_3^{(2)} & y_3^{(1)} & y_3^{(0)} \\ y_4^{(3)} & y_4^{(2)} & y_4^{(1)} & y_4^{(0)} \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix} = - \begin{bmatrix} y_1^{(4)} \\ y_2^{(4)} \\ y_3^{(4)} \\ y_4^{(4)} \end{bmatrix},$$

которая в нашем случае имеет вид

$$\begin{bmatrix} 208 & 30 & 1 & 1 \\ 178 & 22 & 2 & 0 \\ 192 & 18 & 3 & 0 \\ 242 & 20 & 4 & 0 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix} = - \begin{bmatrix} 2108 \\ 1704 \\ 1656 \\ 1992 \end{bmatrix}.$$

Отсюда

$$\left.\begin{aligned} 208p_1 + 30p_2 + p_3 + p_4 &= -2108, \\ 178p_1 + 22p_2 + 2p_3 &= -1704, \\ 192p_1 + 18p_2 + 3p_3 &= -1656, \\ 242p_1 + 20p_2 + 4p_3 &= -1992. \end{aligned}\right\}$$

Решив эту систему, получим:

$$p_1 = -4; \quad p_2 = -40; \quad p_3 = -56; \quad p_4 = -20.$$

Следовательно,

$$\det(\lambda E - A) = \lambda^4 - 4\lambda^3 - 40\lambda^2 - 56\lambda - 20,$$

что совпадает с результатом, найденным по методу А. М. Данилевского (§ 3).

## § 7. Вычисление собственных векторов по методу А. Н. Крылова

Метод А. Н. Крылова дает возможность просто найти соответствующие собственные векторы [1].

Для простоты ограничимся случаем, когда характеристический полином

$$D(\lambda) = \lambda^n + p_1\lambda^{n-1} + \ldots + p_n \tag{1}$$

имеет различные корни $\lambda_1, \lambda_2, \ldots, \lambda_n$. Предположим, что коэффициенты полинома (1) и его корни определены. Требуется найти собственные векторы $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$, отвечающие соответственно собственным значениям $\lambda_1, \lambda_2, \ldots, \lambda_n$.

Пусть $y^{(0)}, y^{(1)} = Ay^{(0)}, \ldots, y^{(n-1)} = A^{n-1}y^{(0)}$ — векторы, использованные в методе А. Н. Крылова для нахождения коэффициентов $p_i$ $(i = 1, 2, \ldots, n)$. Разлагая вектор $y^{(0)}$ по собственным векторам $x^{(i)}$ $(i = 1, 2, \ldots, n)$, будем иметь:

$$y^{(0)} = c_1x^{(1)} + c_2x^{(2)} + \ldots + c_nx^{(n)}, \tag{2}$$

где $c_i$ $(i = 1, 2, \ldots, n)$ — некоторые числовые коэффициенты. Отсюда, учитывая, что

$$\begin{aligned} Ax^{(i)} &= \lambda_i x^{(i)}, \\ A^2x^{(i)} &= \lambda_i^2 x^{(i)}, \\ &\ldots\ldots\ldots \end{aligned} \qquad (i = 1, 2, \ldots, n),$$

получим:

$$\begin{aligned} y^{(1)} &= c_1\lambda_1x^{(1)} + c_2\lambda_2x^{(2)} + \ldots + c_n\lambda_nx^{(n)}, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ y^{(n-1)} &= c_1\lambda_1^{n-1}x^{(1)} + c_2\lambda_2^{n-1}x^{(2)} + \ldots + c_n\lambda_n^{n-1}x^{(n)}. \end{aligned} \tag{3}$$

Пусть

$$\varphi_i(\lambda) = \lambda^{n-1} + q_{1i}\lambda^{n-2} + \ldots + q_{n-1,i} \tag{4}$$

$(i = 1, 2, \ldots, n)$ — произвольная система полиномов. Составляя линейную комбинацию векторов $y^{(n-1)}, y^{(n-2)}, \ldots, y^{(0)}$ с коэффициентами $1, q_{1-i}, \ldots, q_{n-1,i}$, в силу соотношений (2) и (3) находим:

$$\begin{aligned} &y^{(n-1)} + q_{1i}y^{(n-2)} + \ldots + q_{n-1,i}y^{(0)} = \\ &\qquad = c_1\varphi_i(\lambda_1)x^{(1)} + c_2\varphi_i(\lambda_2)x^{(2)} + \ldots + c_n\varphi_i(\lambda_n)x^{(n)}. \end{aligned} \tag{5}$$

Если положить

$$\varphi_i(\lambda)=\frac{D(\lambda)}{\lambda-\lambda_i} \qquad (i=1,\ 2,\ \ldots,\ n), \tag{6}$$

то, очевидно,

$$\varphi_i(\lambda_j)=0 \quad \text{при } i\neq j$$

и

$$\varphi_i(\lambda_i)=D'(\lambda_i)\neq 0.$$

Формула (5) при этом принимает вид

$$c_i\varphi_i(\lambda_i)\,x^{(i)}=y^{(n-1)}+q_{1i}y^{(n-2)}+\ldots+q_{n-1,\,i}y^{(0)}\ (i=1,\ 2,\ \ldots,\ n). \tag{7}$$

Таким образом, если $c_i\neq 0$, то полученная линейная комбинация векторов $y^{(n-1)}, y^{(n-2)}, \ldots, y^{(0)}$ дает собственный вектор $x^{(i)}$ с точностью до числового множителя. Коэффициенты $q_{j,\,i}$ $(j=1,\ 2,\ \ldots,\ n-1)$ могут быть легко определены по *схеме Горнера*

$$\left.\begin{aligned} q_{0i}&=1,\\ q_{ji}&=\lambda_i q_{j-1,\,i}+p_j.\end{aligned}\right\}$$

## § 8. Метод Леверрье

Этот метод раскрытия векового определителя основан на формулах Ньютона [3] для сумм степеней корней алгебраического уравнения.

Пусть

$$\det(\lambda E-A)=\lambda^n+p_1\lambda^{n-1}+\ldots+p_n \tag{1}$$

— характеристический полином данной матрицы $A=[a_{ij}]$ и $\lambda_1,\ \lambda_2,\ \ldots,\ \lambda_n$ — полная совокупность его корней, где каждый корень повторяется столько раз, какова его кратность.

Положим

$$s_k=\lambda_1^k+\lambda_2^k+\ldots+\lambda_n^k \qquad (k=0,\ 1,\ 2,\ \ldots,\ n).$$

Тогда при $k\leqslant n$ справедливы *формулы Ньютона* [3]

$$s_k+p_1s_{k-1}+\ldots+p_{k-1}s_1=-kp_k \qquad (k=1,\ 2,\ \ldots,\ n). \tag{2}$$

Отсюда

$$\left.\begin{aligned} p_1&=-s_1,\\ p_2&=-\frac{1}{2}(s_2+p_1s_1),\\ &\ldots\ldots\ldots\ldots\ldots\ldots\\ p_n&=-\frac{1}{n}(s_n+p_1s_{n-1}+\ldots+p_{n-1}s_1).\end{aligned}\right\} \tag{3}$$

Если суммы $s_1,\ s_2,\ \ldots,\ s_n$ известны, то с помощью формул (3) можно шаг за шагом определить коэффициенты $p_1,\ p_2,\ \ldots,\ p_n$ характеристического полинома (1).

Суммы $s_1, s_2, \ldots, s_n$ вычисляются следующим образом: для $s_1$ имеем (гл. X, § 12):

$$s_1 = \lambda_1 + \lambda_2 + \ldots + \lambda_n = \operatorname{Sp} A,$$

т. е.

$$s_1 = \sum_{i=1}^{n} a_{ii}. \tag{4}$$

Далее, как известно (гл. XI, § 1), $\lambda_1^k, \lambda_2^k, \ldots, \lambda_n^k$ являются собственными значениями матрицы $A^k$. Поэтому

$$s_k = \lambda_1^k + \lambda_2^k + \ldots + \lambda_n^k = \operatorname{Sp} A^k,$$

т. е. если

$$A^k = \left[a_{ij}^{(k)}\right],$$

то

$$s_k = \sum_{i=1}^{n} a_{ii}^{(k)}. \tag{5}$$

Степени $A^k = A^{k-1}A$ находятся непосредственным перемножением.

Таким образом, схема раскрытия векового определителя по методу Леверрье весьма простая, а именно: сначала вычисляются $A^k$ ($k = 1, 2, \ldots, n$) — степени данной матрицы $A$, затем находятся соответствующие $s_k$ — суммы элементов главных диагоналей матриц $A^k$ и, наконец, по формулам (3) определяются искомые коэффициенты $p_i$ ($i = 1, 2, \ldots, n$).

Метод Леверрье весьма трудоемок, так как приходится подсчитывать высокие степени данной матрицы. Достоинство его — несложная схема вычислений и отсутствие исключительных случаев.

Пример. Методом Леверрье развернуть характеристический определитель матрицы (см. § 3)

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix}.$$

Решение. Образуем степени $A^k$ ($k = 2, 3, 4$) матрицы $A$. Имеем:

$$A^2 = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 30 & 22 & 18 & 20 \\ 22 & 18 & 16 & 18 \\ 18 & 16 & 18 & 22 \\ 20 & 18 & 22 & 30 \end{bmatrix};$$

$$A^3=\begin{bmatrix}30 & 22 & 18 & 20\\ 20 & 18 & 16 & 18\\ 18 & 16 & 18 & 22\\ 20 & 18 & 22 & 30\end{bmatrix}\begin{bmatrix}1 & 2 & 3 & 4\\ 2 & 1 & 2 & 3\\ 3 & 2 & 1 & 2\\ 4 & 3 & 2 & 1\end{bmatrix}=\begin{bmatrix}208 & 178 & 192 & 242\\ 178 & 148 & 154 & 192\\ 192 & 154 & 148 & 178\\ 242 & 192 & 178 & 208\end{bmatrix};$$

$$A^4=\begin{bmatrix}208 & 178 & 192 & 242\\ 178 & 148 & 154 & 192\\ 192 & 154 & 148 & 178\\ 242 & 192 & 178 & 208\end{bmatrix}\begin{bmatrix}1 & 2 & 3 & 4\\ 2 & 1 & 2 & 3\\ 3 & 2 & 1 & 2\\ 4 & 3 & 2 & 1\end{bmatrix}=\begin{bmatrix}2108 & 1704 & 1656 & 1992\\ 1704 & 1388 & 1368 & 1656\\ 1656 & 1368 & 1388 & 1704\\ 1992 & 1656 & 1704 & 2108\end{bmatrix}.$$

Заметим, что не было необходимости вычислять $A^4$ полностью, достаточно было найти лишь главные диагональные элементы этой матрицы.

Отсюда

$$\begin{aligned}
s_1&=\operatorname{Sp}A=1+1+1+1=4;\\
s_2&=\operatorname{Sp}A^2=30+18+18+30=96;\\
s_3&=\operatorname{Sp}A^3=208+148+148+208=712;\\
s_4&=\operatorname{Sp}A^4=2108+1388+1388+2108=6992.
\end{aligned}$$

Следовательно, по формулам (3) будем иметь:

$$\begin{aligned}
p_1&=-s_1=-4;\\
p_2&=-\frac{1}{2}(s_2+p_1s_1)=-\frac{1}{2}(96-4\cdot 4)=-40;\\
p_3&=-\frac{1}{3}(s_3+p_1s_2+p_2s_1)=-\frac{1}{3}(712-4\cdot 96-40\cdot 4)=-56;\\
p_4&=-\frac{1}{4}(s_4+p_1s_3+p_2s_2+p_3s_1)=\\
&=-\frac{1}{4}(6992-4\cdot 712-40\cdot 96-56\cdot 4)=-20.
\end{aligned}$$

Таким образом, мы получаем уже известный результат (см. § 3):

$$\begin{vmatrix}\lambda-1 & -2 & -3 & -4\\ -2 & \lambda-1 & -2 & -3\\ -3 & -2 & \lambda-1 & -2\\ -4 & -3 & -2 & \lambda-1\end{vmatrix}=\lambda^4-4\lambda^3-40\lambda^2-56\lambda-20.$$

## § 9. Понятие о методе неопределенных коэффициентов

Развертывание векового определителя можно также осуществить при помощи нахождения достаточно большого количества его числовых значений.

Пусть

$$D(\lambda)=\lambda^n+p_1\lambda^{n-1}+\ldots+p_n \qquad (1)$$

есть вековой определитель матрицы $A$, т. е.

$$D(\lambda)=\det(\lambda E-A).$$

Если в равенстве (1) последовательно положить $\lambda=0, 1, 2, \ldots, n-1$, то для коэффициентов $p_i$ $(i=1, 2, \ldots, n)$ получим систему линейных уравнений

$$\left.\begin{aligned} p_n &= D(0),\\ 1^n+p_1\cdot 1^{n-1}+\ldots+p_n &= D(1),\\ 2^n+p_1\cdot 2^{n-1}+\ldots+p_n &= D(2),\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ (n-1)^n+p_1(n-1)^{n-1}+\ldots+p_n &= D(n-1). \end{aligned}\right\} \tag{2}$$

Отсюда

$$\left.\begin{aligned} &p_1+p_2+\ldots+p_{n-1}=D(1)-D(0)-1,\\ &2^{n-1}p_1+2^{n-2}p_2+\ldots+2p_{n-1}=D(2)-D(0)-2^n,\\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ &(n-1)^{n-1}p_1+(n-1)^{n-2}p_2+\ldots+(n-1)p_{n-1}=\\ &\qquad\qquad =D(n-1)-D(0)-(n-1) \end{aligned}\right\} \tag{3}$$

и

$$p_n=D(0)=\det(-A).$$

Из системы (3) можно определить коэффициенты $p_i$ $(i=1, 2, \ldots, n)$ характеристического полинома (1).

Вводя матрицу

$$C_n=\begin{bmatrix} 1 & 1 & \ldots & 1\\ 2^{n-1} & 2^{n-2} & \ldots & 2\\ \ldots & \ldots & \ldots & \ldots\\ (n-1)^{n-1} & (n-1)^{n-2} & \ldots & n-1 \end{bmatrix}$$

и векторы

$$\boldsymbol{D}=\begin{bmatrix} D(1)-D(0)-1^n\\ D(2)-D(0)-2^n\\ \ldots\ldots\ldots\ldots\\ D(n-1)-D(0)-(n-1)^n \end{bmatrix},\quad \boldsymbol{P}=\begin{bmatrix} p_1\\ p_2\\ \cdot\\ \cdot\\ \cdot\\ p_{n-1} \end{bmatrix},$$

систему (3) можно записать в виде матричного уравнения

$$C_n\boldsymbol{P}=\boldsymbol{D}; \tag{4}$$

отсюда

$$\boldsymbol{P}=C_n^{-1}\boldsymbol{D}. \tag{5}$$

Заметим, что обратная матрица $C_n^{-1}$ зависит только от порядка $n$ векового определителя и может быть найдена заранее, если приходится иметь дело с массовым раскрытием вековых определителей одного и того же порядка.

Таким образом, применение этого метода сводится к вычислению числовых определителей

$$D(k) = \det(kE - A) \qquad (k = 0,\ 1,\ 2,\ \ldots,\ n-1)$$

и нахождению решения стандартной линейной системы (4).

## § 10. Сравнение различных методов развертывания векового определителя

Об относительной эффективности различных методов развертывания векового определителя можно судить по приведенной ниже таблице 26 [4], в которой указаны количества действий, требуемых каждым из рассмотренных методов, в зависимости от порядка определителя.

Таблица 26

**Количество действий, используемых различными методами развертывания векового определителя, в зависимости от порядка его**

| Метод | Порядок | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 3 | | 4 | | 5 | | 7 | | 9 | |
| | Умножений—делений У—Д | Сложений—вычитаний С—В | У—Д | С—В | У—Д | С—В | У—Д | С—В | У—Д | С—В |
| Непосредственное развертывание | 12 | 10 | 60 | 46 | 320 | 238 | 13 692 | 10 078 | 986 400 | 725 758 |
| Данилевского | 14 | 12 | 42 | 36 | 92 | 80 | 282 | 252 | 632 | 576 |
| Крылова | 67 | 38 | 179 | 118 | 389 | 280 | 1 287 | 1 022 | 3 209 | 2 688 |
| Леверрье | 41 | 27 | 153 | 114 | 414 | 330 | 1 791 | 1 533 | 5 228 | 4 644 |
| Неопределенных коэффициентов | 67 | 41 | 171 | 116 | 364 | 265 | 1 189 | 945 | 2 966 | 2 481 |
| Интерполирования*) | 46 | 38 | 125 | 102 | 279 | 230 | 972 | 826 | 2 525 | 2 202 |

*) См. гл. XIV, § 23.

Из этой таблицы видно, что для развертывания вековых определителей порядка выше пятого наиболее выгодным, с точки зрения количества действий, является метод А. М. Данилевского.

## § 11. Нахождение наибольшего по модулю собственного значения матрицы и соответствующего собственного вектора

Пусть имеем характеристическое уравнение

$$\det(A - \lambda E) = 0.$$

Корни этого уравнения $\lambda_1, \lambda_2, \ldots, \lambda_n$ являются собственными значениями матрицы $A$. Пусть им соответствуют линейно независимые

собственные векторы $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$. Укажем некоторые итерационные методы вычисления наибольшего по модулю собственного значения матрицы $A$, не требующие раскрытия ее векового определителя.

Случай 1. Среди собственных значений матрицы $A$ есть одно, наибольшее по модулю. Для определенности предположим, что

$$|\lambda_1| > |\lambda_2| \geqslant |\lambda_3| \geqslant \ldots \geqslant |\lambda_n|, \tag{1}$$

так что наибольшим по модулю является первое собственное значение. Очевидно, для действительной матрицы наибольшее по модулю собственное значение $\lambda_1$ действительно. Заметим, что такой случай имеет место, если матрица $A$ — действительная и элементы ее положительны (гл. X, § 16, теорема Перрона).

Укажем приближенный способ вычисления корня $\lambda_1$. Возьмем произвольный вектор $y$ и разложим его по собственным векторам $x^{(j)}$ матрицы $A$:

$$y = \sum_{j=1}^{n} c_j x^{(j)},$$

где $c_j$ $(j = 1, 2, \ldots, n)$ — постоянные коэффициенты. Произведя преобразование $A$ над вектором $y$, будем иметь:

$$Ay = \sum_{j=1}^{n} c_j A x^{(j)}.$$

Отсюда, так как $x^{(j)}$ есть собственный вектор преобразования $A$, т. е. $Ax^{(j)} = \lambda_j x^{(j)}$, получим:

$$Ay = \sum_{j=1}^{n} c_j \lambda_j x^{(j)};$$

$Ay$ назовем *итерацией* вектора $y$.

Последовательно образуя итерации $Ay, A^2y, \ldots, A^my$, находим:

$$A^m y = \sum_{j=1}^{n} c_j \lambda_j^m x^{(j)} \tag{2}$$

(*m-я итерация*).

Выберем в пространстве $E_n = \{y\}$ базис $e_1, e_2, \ldots, e_n$ (не обязательно единичный). Пусть

$$A^m y = y^{(m)} \qquad (m = 1, 2, 3, \ldots)$$

и

$$y^{(m)} = \begin{bmatrix} y_1^{(m)} \\ \cdot \\ \cdot \\ \cdot \\ y_n^{(m)} \end{bmatrix},$$

где $y_i^{(m)}$ $(i=1, 2, \ldots, n)$ — координаты вектора $y^{(m)}$ в выбранном базисе.

Разлагая собственные векторы $x^{(j)}$ по векторам базиса, будем иметь:

$$x^{(j)}=\sum_{i=1}^{n} x_{ij}e_i. \tag{3}$$

Отсюда, подставляя выражение (3) в формулу (2), получим:

$$y^{(m)}=\sum_{j=1}^{n} c_j\lambda_j^m \sum_{i=1}^{n} x_{ij}e_i$$

или, изменяя порядок суммирования,

$$y^{(m)}=\sum_{i=1}^{n} e_i \sum_{j=1}^{n} c_j x_{ij}\lambda_j^m. \tag{4}$$

Коэффициент при $e_i$ есть $i$-я координата вектора $y^{(m)}$. Следовательно, можно написать:

$$y_i^{(m)}=\sum_{j=1}^{n} c_j x_{ij}\lambda_j^m. \tag{4'}$$

Аналогично

$$y_i^{(m+1)}=\sum_{j=1}^{n} c_j x_{ij}\lambda_j^{m+1}. \tag{4''}$$

Разделив вторую сумму на первую, будем иметь

$$\frac{y_i^{(m+1)}}{y_i^{(m)}}=\frac{c_1x_{i1}\lambda_1^{m+1}+\ldots+c_nx_{in}\lambda_n^{m+1}}{c_1x_{i1}\lambda_1^m+\ldots+c_nx_{in}\lambda_n^m}. \tag{5}$$

Предположим, что $c_1\neq 0$ и $x_{i1}\neq 0$. Этого можно добиться, выбирая надлежащим образом исходный вектор $y$ и базис $(e_1, e_2, \ldots, e_n)$.

Преобразуем выражение (5) следующим образом:

$$\frac{y_i^{(m+1)}}{y_i^{(m)}}=\lambda_1\frac{1+\frac{c_2x_{i2}}{c_1x_{i1}}\left(\frac{\lambda_2}{\lambda_1}\right)^{m+1}+\ldots+\frac{c_nx_{in}}{c_1x_{i1}}\left(\frac{\lambda_n}{\lambda_1}\right)^{m+1}}{1+\frac{c_2x_{i2}}{c_1x_{i1}}\left(\frac{\lambda_2}{\lambda_1}\right)^{m}+\ldots+\frac{c_nx_{in}}{c_1x_{i1}}\left(\frac{\lambda_n}{\lambda_1}\right)^{m}}.$$

Отсюда, переходя к пределу при $m\to\infty$ и учитывая неравенства (1), получим:

$$\lim_{m\to\infty}\frac{y_i^{(m+1)}}{y_i^{(m)}}=\lambda_1 \tag{6}$$

$\left(\text{так как } \lim\limits_{m\to\infty}\left(\frac{\lambda_j}{\lambda_1}\right)^m=0 \text{ при } j>1\right)$ или приближенно

$$\lambda_1 \approx \frac{y_i^{(m+1)}}{y_i^{(m)}} \qquad (i=1,\ 2,\ \ldots,\ n), \tag{7}$$

точнее,

$$\lambda_1 = \frac{y_i^{(m+1)}}{y_i^{(m)}} + O\left(\left(\frac{\lambda_2}{\lambda_1}\right)^m\right).$$

Взяв достаточно большой номер итерации $m$, мы сможем с любой степенью точности определить по формуле (7) наибольший по модулю корень $\lambda_1$ характеристического уравнения данной матрицы $A$. Для нахождения этого корня может быть использована любая координата вектора $y^{(m)}$, в частности, можно взять среднее арифметическое соответствующих отношений.

Замечание 1. В исключительных случаях, при неудачном выборе начального вектора $y$, формула (6) может не дать нужного корня или даже вообще может не иметь смысла, т. е. предел отношения $\frac{y_i^{(m+1)}}{y_i^{(m)}}$ может не существовать. Последнее легко заметить по «прыгающим» значениям этого отношения. В таких случаях следует испробовать другой начальный вектор.

Замечание 2. Для ускорения сходимости итерационного процесса (6) иногда выгодно составлять последовательность матриц

$$\begin{aligned} A^2 &= A\cdot A,\\ A^4 &= A^2\cdot A^2,\\ A^8 &= A^4\cdot A^4,\\ &\ldots\ldots\\ A^{2^k} &= A^{2^{k-1}}\cdot A^{2^{k-1}}. \end{aligned}$$

Отсюда находим:

$$y^{(m)} = A^m y$$

и

$$y^{(m+1)} = Ay^{(m)},$$

где $m=2^k$. Затем, как обычно, полагаем:

$$\lambda_1 \approx \frac{y_i^{(m+1)}}{y_i^{(m)}} \qquad (i=1,\ 2,\ \ldots,\ n).$$

Вектор $y^{(m)}=A^m y$ приближенно представляет собой собственный вектор матрицы $A$, соответствующий собственному значению $\lambda_1$.

В самом деле, из формулы (2) имеем:

$$A^m y = c_1 \lambda_1^m x^{(1)} + \sum_{j=2}^{n} c_j \lambda_j^m x^{(j)},$$

где $x^{(j)}$ $(j=1, 2, \ldots, n)$ — собственные векторы матрицы $A$.

Отсюда

$$A^m y = c_1 \lambda_1^m \left\{ x^{(1)} + \sum_{j=2}^{n} \frac{c_j}{c_1} \left(\frac{\lambda_j}{\lambda_1}\right)^m x^{(j)} \right\}.$$

Так как $\left(\frac{\lambda_j}{\lambda_1}\right)^m \longrightarrow 0$ при $m \longrightarrow \infty$ $(j > 1)$, то при достаточно большом $m$ с любой степенью точности будем иметь:

$$A^m y \approx c_1 \lambda_1^m x^{(1)},$$

т. е. $A^m y$ лишь числовым множителем отличается от собственного вектора $x^{(1)}$ и, следовательно, также является собственным вектором, соответствующим тому же самому собственному значению $\lambda_1$.

Пример. Найти наибольшее собственное значение матрицы

$$A = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix} \tag{8}$$

и соответствующий ему собственный вектор.

Решение. Выбираем начальный вектор

$$y = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Составляем таблицу 27.

Таблица 27

**Вычисление первого собственного значения**

| $y$ | $Ay$ | $A^2y$ | $A^3y$ | $A^4y$ | $A^5y$ | $A^6y$ | $A^7y$ | $A^8y$ | $A^9y$ | $A^{10}y$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 5 | 24 | 111 | 504 | 2268 | 10161 | 45433 | 202833 | 905238 | 4 038939 |
| 1 | 4 | 15 | 60 | 252 | 1089 | 4779 | 21141 | 93906 | 417987 | 1 862460 |
| 1 | 2 | 6 | 21 | 81 | 333 | 1422 | 6201 | 27342 | 121248 | 539235 |

Остановившись на итерациях $A^9y=y^{(9)}$ и $A^{10}y=y^{(10)}$, получаем значения

$$\frac{y_1^{(10)}}{y_1^{(9)}}=\frac{4038939}{905238}=4{,}462;$$

$$\frac{y_2^{(10)}}{y_2^{(9)}}=\frac{1862460}{417987}=4{,}456;$$

$$\frac{y_3^{(10)}}{y_3^{(9)}}=\frac{539235}{121248}=4{,}447.$$

Следовательно, приближенно можно принять:

$$\lambda_1=\frac{1}{3}(4{,}462+4{,}456+4{,}447)=4{,}455\approx 4{,}46.$$

В качестве первого собственного вектора матрицы $A$ можно взять:

$$A^{10}y=\begin{bmatrix}4038939\\1862460\\539235\end{bmatrix}.$$

Нормируя его, окончательно получим:

$$x^{(1)}=\begin{bmatrix}0{,}90\\0{,}42\\0{,}12\end{bmatrix}.$$

Случай 2. Наибольшее по модулю собственное значение матрицы $A$ является кратным.

Пусть

$$\lambda_1=\lambda_2=\ldots=\lambda_s$$

и

$$|\lambda_1|>|\lambda_k| \qquad \text{при} \qquad k>s.$$

Из формулы (5) имеем:

$$\frac{y_i^{(m+1)}}{y_i^{(m)}}=\frac{c_1x_{i1}\lambda_1^{m+1}+\ldots+c_sx_{is}\lambda_1^{m+1}+c_{s+1}x_{i,\ s+1}\lambda_{s+1}^{m+1}+\ldots+c_nx_{in}\lambda_n^{m+1}}{c_1x_{i1}\lambda_1^{m}+\ldots+c_sx_{is}\lambda_1^{m}+c_{s+1}x_{i,\ s+1}\lambda_{s+1}^{m}+\ldots+c_nx_{in}\lambda_n^{m}}=$$

$$=\lambda_1\frac{c_1x_{i1}+\ldots+c_sx_{is}+c_{s+1}x_{i,\ s+1}\left(\frac{\lambda_{s+1}}{\lambda_1}\right)^{m+1}+\ldots+c_nx_{in}\left(\frac{\lambda_n}{\lambda_1}\right)^{m+1}}{c_1x_{i1}+\ldots+c_sx_{is}+c_{s+1}x_{i,\ s+1}\left(\frac{\lambda_{s+1}}{\lambda_1}\right)^{m}+\ldots+c_nx_{in}\left(\frac{\lambda_n}{\lambda_1}\right)^{m}}.$$

Отсюда, если $c_1x_{i1}+\ldots+c_sx_{is}\neq 0$, учитывая, что

$$\left(\frac{\lambda_k}{\lambda_1}\right)^m\longrightarrow 0 \text{ при } m\to\infty \text{ и } k>s,$$

получим:

$$\lim_{m\to\infty}\frac{y_i^{(m+1)}}{y_i^{(m)}}=\lambda_1 \qquad (i=1,\ 2,\ \ldots,\ n)$$

или, точнее,

$$\lambda_1=\frac{y_i^{(m+1)}}{y_i^{(m)}}+O\left(\left(\frac{\lambda_{s+1}}{\lambda_1}\right)^m\right).$$

Таким образом, указанный выше способ вычисления $\lambda_1$ применим и в этом случае.

Как и прежде,

$$y^{(m)}=A^m y$$

представляет собой один из приближенных собственных векторов матрицы $A$, соответствующих значению $\lambda_1$. Изменяя начальный вектор $y$, мы, вообще говоря, получаем другой линейно независимый вектор матрицы $A$. Заметим, что в этом случае не гарантировано, что нашим приемом будет определена вся совокупность линейно независимых собственных векторов матрицы $A$ для значения $\lambda_1$.

Для случаев 1—2 можно указать более быстрый итерационный процесс нахождения наибольшего по модулю собственного значения $\lambda_1$ матрицы $A$, а именно: образуем последовательность матриц

$$A,\ A^2,\ A^4,\ A^8,\ \ldots,\ A^{2^k}.$$

Как известно (гл. X, § 12),

$$\sum_{i=1}^{n}\lambda_i=\operatorname{Sp}A;$$

аналогично

$$\sum_{i=1}^{n}\lambda_i^m=\operatorname{Sp}A^m,$$

где $m=2^k$. Ограничиваясь для простоты случаем 1, имеем:

$$\lambda_1^m+\lambda_2^m+\ldots+\lambda_n^m=\lambda_1^m\left[1+\left(\frac{\lambda_2}{\lambda_1}\right)^m+\ldots+\left(\frac{\lambda_n}{\lambda_1}\right)^m\right]=\operatorname{Sp}A^m;$$

отсюда

$$|\lambda_1|\left[1+\left(\frac{\lambda_2}{\lambda_1}\right)^m+\ldots+\left(\frac{\lambda_n}{\lambda_1}\right)^m\right]^{\frac{1}{m}}=\sqrt[m]{\operatorname{Sp}A^m}.$$

При $m\to\infty$ получаем:

$$|\lambda_1|=\lim_{m\to\infty}\sqrt[m]{\operatorname{Sp}A^m},$$

т. е.

$$|\lambda_1|\approx\sqrt[m]{\operatorname{Sp}A^m},$$

где $m$ достаточно велико.

Чтобы избежать извлечения корней высокой степени, можно найти

$$A^{m+1} = A^m A.$$

Тогда

$$\lambda_1^{m+1} + \lambda_2^{m+1} + \ldots + \lambda_n^{m+1} = \operatorname{Sp} A^{m+1}$$

и

$$\lambda_1^m + \lambda_2^m + \ldots + \lambda_n^m = \operatorname{Sp} A^m.$$

Отсюда, учитывая относительную малость $|\lambda_2|, \ldots, |\lambda_n|$ по сравнению с $|\lambda_1|$, получим:

$$\lambda_1 \approx \operatorname{Sp} A^{m+1} / \operatorname{Sp} A^m.$$

## § 12. Метод скалярных произведений для нахождения первого собственного значения действительной матрицы

Для отыскания первого собственного значения $\lambda_1$ действительной матрицы $A$ можно указать несколько иной итерационный процесс, являющийся иногда более выгодным. Метод основан на образовании скалярных произведений

$$(A^k y_0,\ A'^k y_0) \quad \text{и} \quad (A^{k-1} y_0,\ A'^k y_0)$$

($k = 1, 2, \ldots$), где $A'$ — матрица, транспонированная с матрицей $A$, и $y_0$ — выбранный каким-либо образом начальный вектор.

Переходим теперь к изложению самого метода.

Пусть $A$ — действительная матрица и $\lambda_1, \lambda_2, \ldots, \lambda_n$ — ее собственные значения, которые предполагаются различными, причем

$$|\lambda_1| > |\lambda_2| \geqslant \ldots \geqslant |\lambda_n|.$$

Возьмем некоторый ненулевой вектор $y_0$ и с помощью матрицы $A$ построим последовательность итераций

$$y_k = A y_{k-1} \quad (k = 1, 2, \ldots). \tag{1}$$

Для вектора $y_0$ образуем также с помощью транспонированной матрицы $A'$ вторую последовательность итераций

$$y'_k = A' y'_{k-1} \qquad (k = 1, 2, \ldots), \tag{2}$$

где $y'_0 = y_0$.

Согласно теореме 1 главы X, § 16 в пространстве $E_n$ выберем два собственных базиса $\{x_j\}$ и $\{x'_j\}$ соответственно для матриц $A$ и $A'$, удовлетворяющих условиям биортонормировки:

$$(x_i,\ x'_j) = \delta_{ij}, \tag{3}$$

где $Ax_i = \lambda_i x_i$ и $A' x'_j = \lambda_j^* x'_j$ ($i, j = 1, 2, \ldots, n$). Обозначим координаты вектора $y_0$ в базисе $\{x_j\}$ через $a_1, \ldots, a_n$, а в

базисе $\{x'_j\}$ — через $b_1, \ldots, b_n$, т. е.

$$y_0 = a_1 x_1 + \ldots + a_n x_n \text{ и } y_0 = b_1 x'_1 + \ldots + b_n x'_n.$$

Отсюда

$$y_k = A^k y_0 = \sum_{j=1}^{n} a_j \lambda_j^k x_j \tag{4}$$

и

$$y'_k = A'^k y_0 = \sum_{j=1}^{n} b_j \lambda_j^{*k} x'_j \qquad (k = 1, 2, \ldots). \tag{4'}$$

Составим скалярное произведение

$$(y_k, y'_k) = (A^k y_0, A'^k y_0) = (y_0, A'^{2k} y_0) = \left(\sum_{i=1}^{n} a_i x_i, \sum_{j=1}^{n} b_j \lambda_j^{*2k} x'_j\right).$$

Отсюда в силу условия ортонормирования находим:

$$(y_k, y'_k) = \sum_{j=1}^{n} a_j b_j^* \lambda_j^{2k} = a_1 b_1^* \lambda_1^{2k} + a_2 b_2^* \lambda_2^{2k} + \ldots + a_n b_n^* \lambda_n^{2k}. \tag{5}$$

Аналогично

$$(y_{k-1}, y'_k) = a_1 b_1^* \lambda_1^{2k-1} + a_2 b_2^* \lambda_2^{2k-1} + \ldots + a_n b_n^* \lambda_n^{2k-1}. \tag{6}$$

Следовательно, при $a_1 b_1^* \neq 0$ имеем:

$$\frac{(y_k, y'_k)}{(y_{k-1}, y'_k)} = \frac{a_1 b_1^* \lambda_1^{2k} + a_2 b_2^* \lambda_2^{2k} + \ldots + a_n b_n^* \lambda_n^{2k}}{a_1 b_1^* \lambda_1^{2k-1} + a_2 b_2^* \lambda_2^{2k-1} + \ldots + a_n b_n^* \lambda_n^{2k-1}} = \lambda_1 + O\left(\left(\frac{\lambda_2}{\lambda_1}\right)^{2k}\right).$$

Таким образом,

$$\lambda_1 \approx \frac{(y_k, y'_k)}{(y_{k-1}, y'_k)} = \frac{(A^k y_0, A'^k y_0)}{(A^{k-1} y_0, A'^k y_0)}, \tag{7}$$

Этот метод особенно удобен для симметрической матрицы $A$, так как тогда $A' = A$, и мы имеем просто

$$\lambda_1 \approx \frac{(A^k y_0, A^k y_0)}{(A^{k-1} y_0, A^k y_0)}, \tag{8}$$

и, следовательно, здесь нужно построить только одну последовательность $y_k = A^k y_0$ $(k = 1, 2, \ldots)$.

Пример. Методом скалярных произведений найти наибольшее собственное значение матрицы (§ 11)

$$A = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

Решение. Так как матрица $A$ — симметрическая, то достаточно построить лишь одну последовательность итераций $A^k y_0$ $(k = 1, 2, \ldots)$.

Выбирая за начальный вектор

$$y_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix},$$

можно использовать результаты таблицы 27. Например, при $k=5$ и $k=6$ имеем:

$$A^5 y_0 = \begin{bmatrix} 2\,268 \\ 1\,089 \\ 333 \end{bmatrix} \quad \text{и} \quad A^6 y_0 = \begin{bmatrix} 10\,161 \\ 4\,779 \\ 1\,422 \end{bmatrix}.$$

Отсюда

$$(A^5 y_0,\ A^6 y_0) = 2268 \cdot 10\,161 + 1089 \cdot 4779 + 333 \cdot 1422 = 28\,723\,005$$

и

$$(A^6 y_0,\ A^6 y_0) = 10\,161^2 + 4779^2 + 1422^2 = 128\,106\,846.$$

Следовательно,

$$\lambda_1 \approx \frac{(A^6 y_0,\ A^6 y_0)}{(A^5 y_0,\ A^6 y_0)} = \frac{128\,106\,846}{28\,723\,005} = 4{,}46,$$

что совпадает в написанных знаках со значением, найденным прежде с помощью $A^{10} y_0$ (см. § 11).

Замечание. Методы нахождения наибольшего по модулю корня характеристического уравнения (§ 11) можно использовать для нахождения наибольшего по модулю корня алгебраического уравнения

$$x^n + p_1 x^{n-1} + \ldots + p_n = 0. \tag{9}$$

Действительно, уравнение (9), как легко непосредственно проверить, является вековым для матрицы (ср. § 3, матрица Фробениуса)

$$P = \begin{bmatrix} -p_1 & -p_2 & \ldots & -p_{n-1} & -p_n \\ 1 & 0 & \ldots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix},$$

т. е. уравнение (9) эквивалентно уравнению

$$\det(xP - E) = 0.$$

Если уравнение (9) не имеет нулевых корней, то аналогичным способом может быть определен наименьший по модулю корень этого уравнения, а именно, при $p_n \neq 0$, полагая $\frac{1}{x} = y$, получим:

$$y^n + \frac{p_{n-1}}{p_n} y^n + \ldots + \frac{1}{p_n} = 0. \tag{10}$$

Обратная величина наибольшего по модулю корня уравнения (10), очевидно, даст нам наименьший по модулю корень уравнения (9).

## § 13. Нахождение второго собственного значения матрицы и второго собственного вектора

Пусть собственные значения $\lambda_j$ $(j=1, 2, \ldots, n)$ матрицы $A$ таковы, что

$$|\lambda_1| > |\lambda_2| > |\lambda_3| \geqslant \ldots \geqslant |\lambda_n|, \tag{1}$$

т. е. имеются два отличных друг от друга, наибольших по модулю собственных значения $\lambda_1$ и $\lambda_2$ матрицы $A$. В таком случае приемом, аналогичным разобранному выше (§ 11), можно приближенно найти второе собственное значение $\lambda_2$ и отвечающий ему собственный вектор $x^{(2)}$.

Из формулы (2) § 11 имеем:

$$A^m y = c_1\lambda_1^m x^{(1)} + c_2\lambda_2^m x^{(2)} + \ldots + c_n\lambda_n^m x^{(n)} \tag{2}$$

и

$$A^{m+1} y = c_1\lambda_1^{m+1} x^{(1)} + c_2\lambda_2^{m+1} x^{(2)} + \ldots + c_n\lambda_n^{m+1} x^{(n)}. \tag{3}$$

Исключим из формул (2) и (3) члены, содержащие $\lambda_1$. Для этого из равенства (3) вычтем равенство (2), умноженное на $\lambda_1$. В результате получим:

$$A^{m+1}y - \lambda_1 A^m y = c_2\lambda_2^m (\lambda_2 - \lambda_1) x^{(2)} + \ldots + c_n\lambda_n^m(\lambda_n - \lambda_1) x^{(n)}. \tag{4}$$

Для краткости введем обозначение

$$\Delta_\lambda A^m y = A^{m+1}y - \lambda A^m y, \tag{5}$$

причем выражение (5) будем называть $\lambda$-*разностью* от $A^m y$. Если $c_2 \neq 0$, то очевидно, что первое слагаемое в правой части равенства (4) является ее главным членом при $m \to \infty$, и мы имеем приближенное равенство

$$\Delta_{\lambda_1} A^m y \approx c_2\lambda_2^m (\lambda_2 - \lambda_1) x^{(2)}. \tag{6}$$

Отсюда

$$\Delta_{\lambda_1} A^{m-1} y \approx c_2\lambda_2^{m-1} (\lambda_2 - \lambda_1) x^{(2)} \tag{7}$$

Пусть

$$A^m y = y^{(m)} = \begin{bmatrix} y_1^{(m)} \\ y_2^{(m)} \\ \cdot \\ \cdot \\ \cdot \\ y_n^{(m)} \end{bmatrix}.$$

Из формул (6) и (7) выводим:

$$\lambda_2 \approx \frac{\Delta_{\lambda_1} y_i^{(m)}}{\Delta_{\lambda_1} y_i^{(m-1)}} = \frac{y_i^{(m+1)} - \lambda_1 y_i^{(m)}}{y_i^{(m)} - \lambda_1 y_i^{(m-1)}} \qquad (i = 1, 2, \ldots, n). \tag{8}$$

Пользуясь формулой (8), можно приближенно вычислить второе собственное значение $\lambda_2$. Заметим, что на практике ввиду потери точности при вычитании близких чисел иногда выгоднее номер итерации $k$ для определения $\lambda_2$ брать меньшим, чем номер итерации $m$ для определения $\lambda_1$, т. е. целесообразно полагать:

$$\lambda_2 \approx \frac{y_i^{(k+1)} - \lambda_1 y_i^{(k)}}{y_i^{(k)} - \lambda_1 y_i^{(k-1)}} \qquad (k < m), \tag{9}$$

где $k$ — наименьшее из чисел, при котором начинает сказываться преобладание $\lambda_2$ над следующими собственными значениями. Формула (9), вообще говоря, дает грубые значения для $\lambda_2$. Заметим, что если модули всех собственных значений различны между собой, то при помощи формул, аналогичных формуле (9), можно вычислить и остальные собственные значения данной матрицы. Однако результаты вычислений будут еще менее надежны.

Что касается собственного вектора $x^{(2)}$, то, как вытекает из формулы (6), можно положить:

$$x^{(2)} \approx \Delta_{\lambda_1} y^{(k)}. \tag{10}$$

Имеется распространение данного метода на случай кратных корней характеристического уравнения [1].

Пример. Определить дальнейшие собственные значения и собственные векторы матрицы (см. пример из § 11)

$$A = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

Решение. Для нахождения второго собственного значения примем $k = 8$. Имеем (см. таблицу 27):

| $A^7y$ | $A^8y$ | $A^9y$ |
|---|---|---|
| 45 433 | 202 833 | 905 238 |
| 21 141 | 93 906 | 417 987 |
| 6 201 | 27 342 | 121 248 |

Составляем $\lambda$-разности по формуле

$$\Delta_{\lambda_1} y_i^{(j)} = y_i^{(j+1)} - \lambda_1 y_i^{(j)} \qquad (i = 1, 2, 3),$$

где $y^{(j)} = A^j y$. Для каждого из столбцов принимается свое значение $\lambda_1$, а именно: $\lambda_1 = 4{,}462$; $\lambda_1 = 4{,}456$; $\lambda_1 = 4{,}447$ (таблица 28).

Таблица 28

**Вычисление второго собственного значения**

| $A^8y$ | $\lambda_1 A^7 y$ | $\Delta_{\lambda_1} A^7 y$ | $A^9 y$ | $\lambda_1 A^8 y$ | $\Delta_{\lambda_1} A^8 y$ |
|---|---|---|---|---|---|
| 202 833 | 202 722 | 111 | 905 238 | 905 041 | 197 |
| 93 906 | 94 204 | —298 | 417 987 | 418 445 | —458 |
| 27 342 | 27 576 | —234 | 121 248 | 121 590 | —342 |

Отсюда получаем:

$$\frac{\Delta_{\lambda_1} y_1^{(8)}}{\Delta_{\lambda_1} y_1^{(7)}} = \frac{197}{111} = 1{,}78; \quad \frac{\Delta_{\lambda_1} y_2^{(8)}}{\Delta_{\lambda_1} y_2^{(7)}} = \frac{-458}{-298} = 1{,}54; \quad \frac{\Delta_{\lambda_1} y_3^{(8)}}{\Delta_{\lambda_1} y_3^{(7)}} = \frac{-342}{-234} = 1{,}46.$$

Следовательно, приближенно можно принять:

$$\lambda_2 = \frac{1}{3}(1{,}78 + 1{,}54 + 1{,}46) \approx 1{,}59.$$

В качестве второго собственного вектора можно принять:

$$\Delta_{\lambda_1} A^8 y = \begin{bmatrix} 197 \\ -458 \\ -342 \end{bmatrix}.$$

Нормируя этот вектор, получим:

$$x^{(2)} = \begin{bmatrix} 0{,}33 \\ -0{,}76 \\ -0{,}56 \end{bmatrix}.$$

Так как матрица $A$ — симметрическая, то векторы $x^{(1)}$ (§ 11) и $x^{(2)}$ должны быть ортогональны между собой. Проверка дает:

$$(x^{(1)}, x^{(2)}) = 0{,}90 \cdot 0{,}33 + 0{,}42 \cdot (-0{,}76) + 0{,}12 \cdot (-0{,}56) = 0{,}09.$$

Отсюда $\measuredangle (x^{(1)}, x^{(2)}) = 85°$, что довольно неточно.

Третье собственное значение $\lambda_3$ находим по следу матрицы $A$:

$$\lambda_1 + \lambda_2 + \lambda_3 = \operatorname{Sp} A = 4 + 2 + 1 = 7.$$

Отсюда

$$\lambda_3 = 7 - 4{,}46 - 1{,}59 \approx 0{,}95.$$

Собственный вектор

$$x^{(3)} = \begin{bmatrix} x_1^{(3)} \\ x_2^{(3)} \\ x_3^{(3)} \end{bmatrix}$$

можно вычислить из условий ортогональности:

$$\left.\begin{aligned} 0{,}90x_1^{(3)} + 0{,}42x_2^{(3)} + 0{,}12x_3^{(3)} = 0, \\ 0{,}33x_1^{(3)} + (-0{,}76)\,x_2^{(3)} + (-0{,}56)\,x_3^{(3)} = 0. \end{aligned}\right\}$$

Отсюда

$$\frac{x_1^{(3)}}{\begin{vmatrix} 0{,}42 & 0{,}12 \\ -0{,}76 & -0{,}56 \end{vmatrix}} = \frac{x_2^{(3)}}{\begin{vmatrix} 0{,}12 & 0{,}90 \\ -0{,}56 & 0{,}33 \end{vmatrix}} = \frac{x_3^{(3)}}{\begin{vmatrix} 0{,}90 & 0{,}42 \\ 0{,}33 & -0{,}76 \end{vmatrix}}$$

или

$$\frac{x_1^{(3)}}{-0{,}144} = \frac{x_2^{(3)}}{0{,}539} = \frac{x_3^{(3)}}{-0{,}818}.$$

После нормировки окончательно получим:

$$x^{(3)} = \begin{bmatrix} -0{,}14 \\ 0{,}53 \\ -0{,}81 \end{bmatrix}.$$

## § 14. Метод исчерпывания

Для определения второго собственного значения матрицы и принадлежащего ему собственного вектора можно указать еще один способ, называемый *методом исчерпывания* [1].

Пусть матрица $A = [a_{ij}]$ действительна и обладает различными собственными значениями $\lambda_1, \lambda_2, \ldots, \lambda_n$, причем

$$|\lambda_1| > |\lambda_2| > |\lambda_3| \geqslant \ldots \geqslant |\lambda_n|.$$

Наряду с матрицей $A$ рассмотрим матрицу

$$A_1 = A - \lambda_1 X_1 X_1', \tag{1}$$

где $\lambda_1$ — первое собственное значение матрицы $A$,

$$X_1 = \begin{bmatrix} x_{11} \\ x_{21} \\ \cdot \\ \cdot \\ \cdot \\ x_{n1} \end{bmatrix}.$$

— соответствующий собственный вектор матрицы $A$, рассматриваемый как матрица-столбец, и

$$X_1' = [x_{11}' \; x_{21}' \; \ldots \; x_{n1}']$$

— отвечающий $\lambda_1$ собственный вектор транспонированной матрицы $A'$, рассматриваемый как матрица-строка, причем векторы $X_1$ и $X_1'$ нормированы так, что их скалярное произведение равно единице:

$$(X_1, X_1'^*) = X_1' X_1 = \sum_{j=1}^{n} x_{j1} x_{j1}' = 1. \tag{2}$$

Число $\lambda_1$ и векторы $X_1$ и $X_1'$ предполагаются известными.

В развернутом виде матрица $A_1$ записывается следующим образом:

$$A_1 = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} - \lambda_1 \begin{bmatrix} x_{11} \\ x_{21} \\ \cdot \\ \cdot \\ x_{n1} \end{bmatrix} [x_{11}' \; x_{21}' \; \ldots \; x_{n1}'] =$$

$$= \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} - \lambda_1 \begin{bmatrix} x_{11}x_{11}' & x_{11}x_{21}' & \ldots & x_{11}x_{n1}' \\ x_{21}x_{11}' & x_{21}x_{21}' & \ldots & x_{21}x_{n1}' \\ \cdot & \cdot & \cdot & \cdot \\ x_{n1}x_{11}' & x_{n1}x_{21}' & \ldots & x_{n1}x_{n1}' \end{bmatrix}. \tag{1'}$$

Докажем, что все собственные векторы $X_j$ $(j = 1, 2, \ldots, n)$ матрицы $A$ являются также собственными векторами матрицы $A_1$, причем соответствующие собственные значения сохраняются, за исключением $\lambda_1$, вместо которого появляется нулевое собственное значение.

В самом деле, используя свойство ассоциативности матричного произведения и условие нормировки (2), имеем:

$$A_1 X_1 = A X_1 - \lambda_1 (X_1 X_1') X_1 = \lambda_1 X_1 - \lambda_1 X_1 (X_1' X_1) = \lambda_1 X_1 - \lambda_1 X_1 = 0,$$

т. е.

$$A_1 X_1 = 0 X_1$$

и, следовательно, нуль является собственным значением матрицы $A_1$.

Далее, при $j > 1$, учитывая, что

$$(X_j, X_1'^*) = X_1' X_j = 0 \qquad (j = 2, \ldots, n)$$

(см. гл. X, § 16, теорема 1), получаем:

$$A_1 X_j = A X_j - \lambda_1 (X_1 X_1') X_j = \lambda_1 X_j - \lambda_1 X_1 (X_1' X_j) = \lambda_j X_j$$
$$(j = 2, \ldots, n).$$

Таким образом, для матрицы $A_1$ наибольшим по модулю собственным значением является $\lambda_2$. Поэтому для определения $\lambda_2$ и соответствующего собственного вектора $X_2$ можно воспользоваться указанными выше методами (§§ 11 и 12). Этот прием называется *методом исчерпывания*. Например, исходя из произвольного вектора $y_0$, можно вычислить $\lambda_2$ по формуле

$$\lambda_2 \approx \frac{(A_1^m y_0)_i}{(A_1^{m-1} y_0)_i} \quad (i = 1, 2, \ldots, n),$$

причем

$$X_2 \approx c A_1^m y_0 \quad (c \neq 0).$$

Покажем, что для нахождения итераций $A_1^m y_0$ $(m = 1, 2, \ldots)$ можно воспользоваться формулой

$$A_1^m y_0 = A^m y_0 - \lambda_1^m X_1 X_1' y_0, \tag{3}$$

позволяющей избежать непосредственного итерирования матрицы $A_1$.

Действительно, пусть собственные векторы $X_j$ и $X_j'$ $(j = 1, 2, \ldots, n)$ матрицы $A$ и транспонированной матрицы $A'$ удовлетворяют условиям биортонормировки (гл. X, § 16, теорема 2)

$$X_k' X_j = \delta_{jk},$$

где $\delta_{jk}$ — символ Кронекера. Тогда имеет место билинейное разложение матрицы $A$

$$A = \lambda_1 X_1 X_1' + \lambda_2 X_2 X_2' + \ldots + \lambda_n X_n X_n'. \tag{4}$$

Отсюда

$$A_1 = A - \lambda_1 X_1 X_1' = \lambda_2 X_2 X_2' + \ldots + \lambda_n X_n X_n'. \tag{5}$$

Так как

$$A^m X_j = \lambda_j^m X_j \quad (j = 1, 2, \ldots, n),$$

то, умножая равенство (4) слева на $A^{m-1}$, будем иметь:

$$A^m = A^m X_1 X_1' + A^m X_2 X_2' + \ldots + A^m X_n X_n' = \\ = \lambda_1^m X_1 X_1' + \lambda_2^m X_2 X_2' + \ldots + \lambda_n^m X_n X_n'. \tag{6}$$

Аналогично, учитывая, что

$$A_1^m X_1 = A_1^{m-1} (A_1 X_1) = 0$$

и

$$A_1^m X_j = \lambda_j^m X_j \quad (j = 2, 3, \ldots, n),$$

после умножения равенства (5) слева на $A_1^{m-1}$ получим:

$$A_1^m = A_1^m X_2 X_2' + \ldots + A_1^m X_n X_n' = \lambda_2^m X_2 X_2' + \ldots + \lambda_n^m X_n X_n'. \qquad (7)$$

Из формул (6) и (7) вытекает:

$$A_1^m = A^m - \lambda_1^m X_1 X_1',$$

что эквивалентно соотношению (3).

## § 15. Нахождение собственных элементов положительно определенной симметрической матрицы

Приведем итерационный метод одновременного нахождения собственных значений и собственных векторов положительно определенной матрицы [5].

Как известно (гл. X, § 15), если действительная матрица

$$A = [a_{ij}]$$

— симметрическая и положительно определенная, то:

1) корни $\lambda_1, \lambda_2, \ldots, \lambda_n$ ее характеристического уравнения

$$\begin{bmatrix} a_{11}-\lambda & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22}-\lambda & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn}-\lambda \end{bmatrix} = 0 \qquad (1)$$

действительны и положительны;

2) собственные векторы

$$x^{(j)} = \begin{bmatrix} x_1^{(j)} \\ \cdot \\ \cdot \\ \cdot \\ x_n^{(j)} \end{bmatrix} \qquad (j = 1, 2, \ldots, n)$$

могут быть взяты действительными и удовлетворяют условиям ортогональности

$$\sum_{i=1}^{n} x_i^{(j)} x_i^{(k)} = 0 \quad \text{при} \quad j \neq k. \qquad (2)$$

Напишем систему, служащую для определения собственного вектора $x^{(1)}$:

$$\left.\begin{aligned} &(a_{11}-\lambda_1)x_1^{(1)} + a_{12}x_2^{(1)} + \ldots + a_{1n}x_n^{(1)} = 0, \\ &a_{21}x_1^{(1)} + (a_{22}-\lambda_1)x_2^{(1)} + \ldots + a_{2n}x_n^{(1)} = 0, \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ &a_{n1}x_1^{(1)} + a_{n2}x_2^{(1)} + \ldots + (a_{nn}-\lambda_1)x_n^{(1)} = 0 \end{aligned}\right\}$$

или

$$\left.\begin{aligned}
x_1^{(1)} &= \frac{1}{\lambda_1}\left(a_{11}x_1^{(1)} + a_{12}x_2^{(1)} + \ldots + a_{1n}x_n^{(1)}\right),\\
x_2^{(1)} &= \frac{1}{\lambda_1}\left(a_{21}x_1^{(1)} + a_{22}x_2^{(1)} + \ldots + a_{2n}x_n^{(1)}\right),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
x_{n-1}^{(1)} &= \frac{1}{\lambda_1}\left(a_{n-1,1}x_1^{(1)} + a_{n-1,2}x_2^{(1)} + \ldots + a_{n-1,n}x_n^{(1)}\right),\\
\lambda_1 &= \frac{1}{x_n^{(1)}}\left(a_{n1}x_1^{(1)} + a_{n2}x_2^{(1)} + \ldots + a_{nn}x_n^{(1)}\right).
\end{aligned}\right\} \qquad (3)$$

Так как координаты собственных векторов определяются с точностью до множителя пропорциональности, то одна из них произвольна, например, за исключением особого случая, можно положить $x_n^{(1)} = 1$. Систему (3), вообще говоря, можно решить методом итерации [5], выбирая подходящие начальные значения $x_i^{(1,\,0)}$, $\lambda_1^{(0)}$ и полагая

$$x_i^{(1,\,k+1)} = \frac{1}{\lambda_1^{(k)}}\left(\sum_{j=1}^{n-1} a_{ij}x_j^{(1,\,k)} + a_{in}\right) \qquad (i = 1,\ 2,\ \ldots,\ n-1);$$

$$\lambda_1^{(k+1)} = \sum_{j=1}^{n-1} a_{nj}x_j^{(1,\,k+1)} + a_{nn} \qquad (k = 0,\ 1,\ 2,\ \ldots).$$

Можно также использовать процесс Зейделя. Таким образом, находятся первый корень характеристического уравнения (1)

$$\lambda_1 \approx \lambda_1^{(k)} \qquad (4)$$

и первый собственный вектор

$$x^{(1)} \approx \begin{bmatrix} x_1^{(1,\,k)} \\ \cdot \\ \cdot \\ \cdot \\ x_{n-1}^{(1,\,k)} \\ 1 \end{bmatrix}.$$

Для определения второго корня $\lambda_2$ уравнения (1) и второго собственного вектора $x^{(2)}$ напишем соответствующую систему уравнений:

$$\lambda_2 x_i^{(2)} = \sum_{j=1}^{n} a_{ij}x_j^{(2)} \qquad (i = 1,\ 2,\ \ldots,\ n). \qquad (5)$$

Из соотношения ортогональности

$$\sum_{j=1}^{n} x_j^{(1)}x_j^{(2)} = 0 \qquad (6)$$

исключим одно из неизвестных $x_j^{(2)}$, например $x_n^{(2)}$. Тогда система (5) заменится эквивалентной системой

$$\left.\begin{aligned} x_i^{(2)} &= \frac{1}{\lambda_2}\sum_{j=1}^{n-1} a_{ij}^{(2)} x_j^{(2)} \quad (i=1,\ 2,\ \ldots,\ n-2), \\ \lambda_2 &= \frac{1}{x_{n-1}^{(2)}}\sum_{j=1}^{n-1} a_{n-1,\,j}^{(2)} x_j^{(2)}. \end{aligned}\right\} \tag{7}$$

Полагая $x_{n-1}^{(2)}=1$, решаем систему (7) методом итерации. В результате будут найдены второй корень $\lambda_2$ характеристического уравнения (1) и собственный вектор $x^{(2)}$, причем $n$-я координата этого вектора определяется из условия ортогональности (6). Аналогично отыскиваются остальные корни $\lambda_j$ $(j=3,\ \ldots,\ n)$ уравнения (1) и соответствующие им собственные векторы $x^{(j)}$.

Мы не рассматриваем исключительные случаи, которые могут возникнуть при применении этого метода.

Пример. Для матрицы [5]

$$A=\begin{bmatrix} 4 & 2 & 2 \\ 2 & 5 & 1 \\ 2 & 1 & 6 \end{bmatrix}$$

найти корни $\lambda_j$ характеристического уравнения и собственные векторы $x^{(j)}$.

Решение. Матрица $A$ является симметрической и положительно определенной, так как

$$\Delta_1 = 4 > 0;$$
$$\Delta_2 = \begin{vmatrix} 4 & 2 \\ 2 & 5 \end{vmatrix} = 16 > 0;$$
$$\Delta_3 = \det A = 80 > 0.$$

Соответствующая система имеет вид

$$\left.\begin{aligned} \lambda_j x_1^{(j)} &= 4x_1^{(j)} + 2x_2^{(j)} + 2x_3^{(j)}, \\ \lambda_j x_2^{(j)} &= 2x_1^{(j)} + 5x_2^{(j)} + x_3^{(j)}, \\ \lambda_j x_3^{(j)} &= 2x_1^{(j)} + x_2^{(j)} + 6x_3^{(j)} \end{aligned}\right\} \quad (j=1,\ 2,\ 3). \tag{8}$$

Полагая $j=1$ и $x_3^{(1)}=1$, получим:

$$\left.\begin{aligned} x_1^{(1)} &= \frac{1}{\lambda_1}\left(4x_1^{(1)} + 2x_2^{(1)} + 2\right), \\ x_2^{(1)} &= \frac{1}{\lambda_1}\left(2x_1^{(1)} + 5x_2^{(1)} + 1\right), \\ \lambda_1 &= 2x_1^{(1)} + x_2^{(1)} + 6. \end{aligned}\right\} \tag{9}$$

Систему (9) решаем методом итерации, выбирая начальные значения

$$x_1^{(1,0)}=1 \text{ и } x_2^{(1,0)}=1.$$

Тогда из последнего уравнения системы (9) будем иметь $\lambda_1^{(0)}=9$. Результаты вычислений приведены в таблице 29.

Таблица 29

**Вычисление методом итерации собственных элементов матрицы, отвечающих первому корню характерического уравнения**

| $k$ | $x_1^{(1k)}$ | $x_2^{(1k)}$ | $x_3^{(1k)}$ | $\lambda_1^{(k)}$ |
|---|---|---|---|---|
| 0 | 1 | 1 | 1 | 9 |
| 1 | 0,89 | 0,89 | 1 | 8,67 |
| 2 | 0,85 | 0,83 | 1 | 8,53 |
| 3 | 0,83 | 0,80 | 1 | 8,46 |
| 4 | 0,81 | 0,78 | 1 | 8,40 |
| 5 | 0,805 | 0,770 | 1 | 8,38 |
| 6 | 0,806 | 0,771 | 1 | 8,383 |
| 7 | 0,807 | 0,771 | 1 | 8,385 |
| 8 | 0,8074 | 0,7715 | 1 | 8,3863 |
| 9 | 0,8076 | 0,7717 | 1 | 8,3869 |
| 10 | 0,8076 | 0,7719 | 1 | 8,3871 |
| 11 | 0,8077 | 0,7720 | 1 | 8,3874 |

Можно принять

$$\lambda_1=8{,}3874$$

и

$$x^{(1)}=\begin{bmatrix}0{,}8077\\0{,}7720\\1\end{bmatrix}.$$

Положим теперь в системе (8) $j=2$. Из условия ортогональности векторов $x^{(1)}$ и $x^{(2)}$ имеем:

$$0{,}8077\,x_1^{(2)}+0{,}7720\,x_2^{(2)}+x_3^{(2)}=0.$$

Отсюда

$$x_3^{(2)}=-0{,}8077\,x_1^{(2)}-0{,}7720\,x_2^{(2)}. \quad (10)$$

Подставляя это выражение в систему (8) и полагая $x_2^{(2)}=1$, получим:

$$\left.\begin{aligned} x_1^{(2)}&=\frac{1}{\lambda_2}(2{,}3846\,x_1^{(2)}+0{,}4560),\\ \lambda_2&=1{,}1923\,x_1^{(2)}+4{,}2280.\end{aligned}\right\} \quad (11)$$

Систему (11) решаем методом итерации, полагая:

$$x_1^{(2,0)}=1 \text{ и } \lambda_2^{(0)}=5{,}42.$$

Результаты вычислений даны в таблице 30.

Можно принять $\lambda_2=4{,}4867$ и $x_1^{(2)}=0{,}2170$; $x_2^{(2)}=1$.

Третья координата определяется из соотношений ортогональности (10):

$$x_3^{(2)}=-0{,}9473,$$

поэтому

$$x^{(2)}=\begin{bmatrix}0{,}2170\\1\\-0{,}9473\end{bmatrix}.$$

Таблица 30

**Вычисление методом итерации собственных элементов матрицы, отвечающих второму корню характеристического уравнения**

| $k$ | $x_1^{(2k)}$ | $x_2^{(2k)}$ | $\lambda_2^{(k)}$ | $k$ | $x_1^{(2k)}$ | $x_2^{(2k)}$ | $\lambda_2^{(k)}$ |
|---|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 5,42 | 6 | 0,223 | 1 | 4,494 |
| 1 | 0,52 | 1 | 4,85 | 7 | 0,220 | 1 | 4,490 |
| 2 | 0,35 | 1 | 4,64 | 8 | 0,218 | 1 | 4,488 |
| 3 | 0,28 | 1 | 4,56 | 9 | 0,2174 | 1 | 4,487 |
| 4 | 0,25 | 1 | 4,53 | 10 | 0,2171 | 1 | 4,4868 |
| 5 | 0,23 | 1 | 4,500 | 11 | 0,2170 | 1 | 4,4867 |

Третий собственный вектор $\boldsymbol{x}^{(3)}$ непосредственно определяется из двух соотношений ортогональности

$$\left.\begin{aligned} 0{,}8077\,x_1^{(3)} + 0{,}7720\,x_2^{(3)} + x_3^{(3)} = 0, \\ 0{,}2170\,x_1^{(3)} + x_2^{(3)} - 0{,}9473\,x_3^{(3)} = 0. \end{aligned}\right\}$$

Полагая $x_1^{(3)} = 1$, получим $x_2^{(3)} = -0{,}5673$; $x_3^{(3)} = -0{,}3698$. Следовательно,

$$\boldsymbol{x}^{(3)} = \begin{bmatrix} 1 \\ -0{,}5673 \\ -0{,}3698 \end{bmatrix}.$$

Из последнего уравнения системы (8) при $j = 3$ находим также

$$\lambda_3 = 2{,}1260.$$

Для контроля составим след матрицы A:

$$\operatorname{Sp} A = \lambda_1 + \lambda_2 + \lambda_3 = 8{,}3874 + 4{,}4867 + 2{,}1260 = 15{,}0001 \approx 4 + 5 + 6.$$

Заметим, что корни, получаемые процессом итерации, как правило, расположены в порядке убывания их модулей. Собственные векторы матрицы определяются с точностью до коэффициентов пропорциональности, поэтому все решения системы (8) таковы:

| $\lambda_j$ | $x_1^{(j)}$ | $x_2^{(j)}$ | $x_3^{(j)}$ |
|---|---|---|---|
| 8,3874 | $0{,}8077c_1$ | $0{,}7720c_1$ | $c_1$ |
| 4,4867 | $0{,}2170c_2$ | $c_2$ | $-0{,}9473c_2$ |
| 2,1260 | $c_3$ | $-0{,}5673c_3$ | $-0{,}3698c_3$ |

($c_1$, $c_2$, $c_3$—произвольные постоянные, отличные от нуля).

## § 16. Использование коэффициентов характеристического полинома матрицы для ее обращения

Выше были приведены способы развертывания векового определителя матрицы в полином (§§ 3—9). С помощью коэффициентов этого характеристического полинома и составления степеней $A$, $A^2, \ldots, A^{n-1}$ неособенной матрицы $A$ порядка $n$ сравнительно просто можно найти обратную матрицу $A^{-1}$. Особенно выгодным в этом отношении является метод Леверрье (§ 8).

Пусть имеем неособенную матрицу $A$ порядка $n$. Рассмотрим ее характеристический полином

$$\det(\lambda E - A) = \lambda^n + p_1\lambda^{n-1} + \ldots + p_{n-1}\lambda + p_n.$$

Согласно тождеству Гамильтона—Кели (гл. XI, § 2) имеем:

$$A^n + p_1 A^{n-1} + \ldots + p_{n-1}A + p_n E = 0. \tag{1}$$

Умножая матричное равенство (1) слева на $A^{-1}$, получим:

$$A^{n-1} + p_1 A^{n-2} + \ldots p_{n-1}E + p_n A^{-1} = 0. \tag{2}$$

Отсюда при $p_n \neq 0$ будем иметь:

$$A^{-1} = -\frac{1}{p_n}(A^{n-1} + p_1 A^{n-2} + \ldots + p_{n-1}E). \tag{3}$$

Таким образом, если известны коэффициенты характеристического полинома матрицы $A$ и составлены степени этой матрицы до $(n-1)$-й включительно, то обратная матрица $A^{-1}$ легко вычисляется по формуле (3).

Заметим, что если $p_n = 0$ и $p_{n-1} \neq 0$, то для получения формулы, содержащей $A^{-1}$, векторное равенство (1) нужно умножать слева на $A^{-2}$ и т. д.

Пример. Для матрицы (см. § 8, пример)

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix}$$

найти обратную матрицу $A^{-1}$.

Решение. Воспользуемся найденными прежде степенями матрицы $A$ (§ 8):

$$A^2 = \begin{bmatrix} 30 & 22 & 18 & 20 \\ 22 & 18 & 16 & 18 \\ 18 & 16 & 18 & 22 \\ 20 & 18 & 22 & 30 \end{bmatrix}$$

и

$$A^3 = \begin{bmatrix} 208 & 178 & 192 & 242 \\ 178 & 148 & 154 & 192 \\ 192 & 154 & 148 & 178 \\ 242 & 192 & 178 & 208 \end{bmatrix}.$$

Так как характеристический полином матрицы $A$ имеет вид

$$\det(\lambda A - E) = \lambda^4 - 4\lambda^3 - 40\lambda^2 - 56\lambda - 20,$$

то по формуле (3) получаем:

$$A^{-1} = -\frac{1}{-20}\left\{\begin{bmatrix} 208 & 178 & 192 & 242 \\ 178 & 148 & 154 & 192 \\ 192 & 154 & 148 & 178 \\ 242 & 192 & 178 & 208 \end{bmatrix} - \right.$$

$$\left. -4\begin{bmatrix} 30 & 22 & 18 & 20 \\ 22 & 18 & 16 & 18 \\ 18 & 16 & 18 & 22 \\ 20 & 18 & 22 & 30 \end{bmatrix} - 40\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix} - 56\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}\right\} =$$

$$= \frac{1}{10}\left\{\begin{bmatrix} 104 & 89 & 96 & 121 \\ 89 & 74 & 77 & 96 \\ 96 & 77 & 74 & 89 \\ 121 & 96 & 89 & 104 \end{bmatrix} - \begin{bmatrix} 60 & 44 & 36 & 40 \\ 44 & 36 & 32 & 36 \\ 36 & 32 & 36 & 44 \\ 40 & 36 & 44 & 60 \end{bmatrix} - \right.$$

$$\left. - \begin{bmatrix} 20 & 40 & 60 & 80 \\ 40 & 20 & 40 & 60 \\ 60 & 40 & 20 & 40 \\ 80 & 60 & 40 & 20 \end{bmatrix} - \begin{bmatrix} 28 & 0 & 0 & 0 \\ 0 & 28 & 0 & 0 \\ 0 & 0 & 28 & 0 \\ 0 & 0 & 0 & 28 \end{bmatrix}\right\} =$$

$$= \begin{bmatrix} -0{,}4 & 0{,}5 & 0 & 0{,}1 \\ 0{,}5 & -1 & 0{,}5 & 0 \\ 0 & 0{,}5 & -1 & 0{,}5 \\ 0{,}1 & 0 & 0{,}5 & -0{,}4 \end{bmatrix}.$$

Для контроля составляем произведение

$$AA^{-1} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} -0{,}4 & 0{,}5 & 0 & 0{,}1 \\ 0{,}5 & -1 & 0{,}5 & 0 \\ 0 & 0{,}5 & -1 & 0{,}5 \\ 0{,}1 & 0 & 0{,}5 & -0{,}4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = E.$$

## § 17. Метод Л. А. Люстерника улучшения сходимости процесса итерации для решения системы линейных уравнений

Пусть система линейных уравнений

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

приведена к виду, удобному для итерации,

$$x = \beta + \alpha x. \tag{1'}$$

Согласно методу итерации (гл. VIII, § 8), последовательные приближения решения $x$ системы (1′) определяются по формуле

$$x^{(m)} = \beta + \alpha x^{(m-1)} \qquad (m = 1,\ 2,\ \ldots), \tag{2}$$

где $x^{(0)}$ — произвольный начальный вектор.

Предположим, что собственные значения $\lambda_1, \lambda_2, \ldots, \lambda_n$ матрицы $\alpha$ различны, причем

$$|\lambda_1| > |\lambda_2| \geqslant \ldots \geqslant |\lambda_n|. \tag{3}$$

Процесс итерации (2) сходится, если

$$|\lambda_1| < 1.$$

Первое собственное значение $\lambda_1$ может быть приближенно определено с помощью указанных выше методов (§§ 11—12). Как доказал Л. А. Люстерник [6], используя число $\lambda_1$, можно существенно улучшить сходимость итерационного процесса (2) для решения системы (1′). Покажем, как это делается.

При достаточно большом $m$ приближенно можно положить:

$$x \approx x^{(m)}.$$

Оценим ошибку $x - x^{(m)}$. При условии сходимости процесса (2) имеем:

$$x = \lim_{m\to\infty} x^{(m)} = x^{(0)} + \sum_{k=1}^{m} (x^{(k)} - x^{(k-1)});$$

кроме того,

$$x^{(m)} = x^{(0)} + \sum_{k=1}^{m} (x^{(k)} - x^{(k-1)}).$$

Поэтому

$$x - x^{(m)} = \sum_{k=m+1}^{\infty} (x^{(k)} - x^{(k-1)}) = \\ = [x^{(m+1)} - x^{(m)}] + [x^{(m+2)} - x^{(m+1)}] + \ldots \tag{4}$$

Так как

$$x^{(k)} - x^{(k-1)} = [\beta + \alpha x^{(k-1)}] - [\beta + \alpha x^{(k-2)}] = \\ = \alpha (x^{(k-1)} - x^{(k-2)}) = \alpha^{k-1} (x^{(1)} - x^{(0)}) \text{ при } k = 1,\ 2,\ \ldots,$$

то

$$x - x^{(m)} = \alpha^m (x^{(1)} - x^{(0)}) + \alpha^{m+1} (x^{(1)} - x^{(0)}) + \ldots \tag{5}$$

Пусть $y_1, y_2, \ldots, y_n$ — собственные векторы матрицы $\alpha$, соответствующие собственным значениям $\lambda_1, \lambda_2, \ldots, \lambda_n$ и образующие базис пространства $E_n$. Разлагая вектор $x^{(1)} - x^{(0)}$ по векторам этого базиса, будем иметь:

$$x^{(1)} - x^{(0)} = c_1 y_1 + c_2 y_2 + \ldots + c_n y_n,$$

где $c_j$ $(j = 1, 2, \ldots, n)$ — некоторые определенные числа. Отсюда

$$x^{(k)} - x^{(k-1)} = \alpha^{k-1} (x^{(1)} - x^{(0)}) = \\ = c_1 \lambda_1^{k-1} y_1 + c_2 \lambda_2^{k-1} y_2 + \ldots + c_n \lambda_n^{k-1} y_n \tag{6}$$
$$(k = m + 1,\ m + 2,\ \ldots).$$

Следовательно, на основании формулы (5) находим:

$$x - x^{(m)} = c_1 \lambda_1^m (1 + \lambda_1 + \lambda_1^2 + \ldots) y_1 + c_2 \lambda_2^m (1 + \lambda_2 + \lambda_2^2 + \ldots) y_2 + \\ + \ldots + c_n \lambda_n^m (1 + \lambda_n + \lambda_n^2 + \ldots) y_n = \\ = \frac{c_1 \lambda_1^m}{1 - \lambda_1} y_1 + \frac{c_2 \lambda_2^m}{1 - \lambda_2} y_2 + \ldots + \frac{c_n \lambda_n^m}{1 - \lambda_n} y_n.$$

Отсюда, учитывая неравенства (3), получим:

$$x - x^{(m)} = \frac{c_1 \lambda_1^m}{1 - \lambda_1} y_1 + O(\lambda_2^m). \tag{7}$$

Кроме того, из формулы (6) при $k = m + 1$ выводим:

$$x^{(m+1)} - x^{(m)} = c_1 \lambda_1^m y_1 + O(\lambda_2^m). \tag{8}$$

Поэтому

$$x - x^{(m)} = \frac{x^{(m+1)} - x^{(m)}}{1 - \lambda_1} + O(\lambda_2^m).$$

Таким образом, окончательно имеем:

$$x \approx x^{(m)} + \frac{x^{(m+1)} - x^{(m)}}{1 - \lambda_1}. \tag{9}$$

Добавочный член $\frac{x^{(m+1)} - x^{(m)}}{1 - \lambda_1}$ заметно улучшает сходимость итерационного процесса (2).

Так как из формулы (8) вытекает, что

$$x^{(m+1)} - x^{(m)} = \lambda_1 (x^{(m)} - x^{(m-1)}) + O(\lambda_2^m), \tag{10}$$

то формулу (9) можно заменить следующей:

$$x \approx x^{(m)} + \frac{\lambda_1}{1 - \lambda_1} (x^{(m)} - x^{(m-1)}). \tag{11}$$

Формула (11) освобождает от необходимости вычислять следующее по порядку приближение.

На основании формулы (10) наибольшее собственное значение $\lambda_1$ может быть определено по формуле.

$$\lambda_1 \approx \frac{(x^{(m)}-x^{(m-1)})_i}{(x^{(m-1)}-x^{(m-2)})_i} \qquad (i=1,\ 2,\ \ldots,\ n).$$

В случае симметрической матрицы $\alpha$, пользуясь методом скалярных произведений, получаем более точную формулу:

$$\lambda_1 \approx \frac{(x^{(m)}-x^{(m-1)},\ x^{(m)}-x^{(m-1)})}{(x^{(m-1)}-x^{(m-2)},\ x^{(m)}-x^{(m-1)})}.$$

В частности, если

$$x^{(0)}=\beta,$$

то

$$x^{(m)}-x^{(m-1)}=\alpha^{m-1}(x^{(1)}-x^{(0)})=\alpha^m\beta$$

и

$$x^{(m)}=x^{(0)}+\sum_{k=1}^{m}\alpha^{k-1}(x^{(1)}-x^{(0)})=\sum_{k=0}^{m}\alpha^k\beta.$$

Поэтому

$$\lambda_1 \approx \frac{(\alpha^m\beta)_i}{(\alpha^{m-1}\beta)_i} \qquad (i=1,\ 2,\ \ldots,\ n), \tag{12}$$

где $(\alpha^m\beta)_i$ и $(\alpha^{m-1}\beta)_i$ — $i$-е координаты соответственно векторов $\alpha^m\beta$ и $\alpha^{m-1}\beta$. Аналогично, если матрица $\alpha$ — симметрическая, то

$$\lambda_1 \approx \frac{(\alpha^m\beta,\ \alpha^m\beta)}{(\alpha^{m-1}\beta,\ \alpha^m\beta)}. \tag{13}$$

Пример. Методом итерации решить систему [1]

$$\left.\begin{aligned}
0{,}78x_1-0{,}02x_2-0{,}12x_3-0{,}14x_4&=0{,}76;\\
-0{,}02x_1+0{,}86x_2-0{,}04x_3+0{,}06x_4&=0{,}08;\\
-0{,}12x_1-0{,}04x_2+0{,}72x_3-0{,}08x_4&=1{,}12;\\
-0{,}14x_1+0{,}06x_2-0{,}08x_3+0{,}74x_4&=0{,}68,
\end{aligned}\right\}$$

применив для уточнения корней способ Л. А. Люстерника.

Решение. Приводим систему к виду, удобному для применения метода итерации:

$$\left.\begin{aligned}
x_1&=0{,}22x_1+0{,}02x_2+0{,}12x_3+0{,}14x_4+0{,}76;\\
x_2&=0{,}02x_1+0{,}14x_2+0{,}04x_3-0{,}06x_4+0{,}08;\\
x_3&=0{,}12x_1+0{,}04x_2+0{,}28x_3+0{,}08x_4+1{,}12;\\
x_4&=0{,}14x_1-0{,}06x_2+0{,}08x_3+0{,}26x_4+0{,}68
\end{aligned}\right\} \tag{14}$$

или в матричной форме

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0{,}76 \\ 0{,}08 \\ 1{,}12 \\ 0{,}68 \end{bmatrix} + \begin{bmatrix} 0{,}22 & 0{,}02 & 0{,}12 & 0{,}14 \\ 0{,}02 & 0{,}14 & 0{,}04 & -0{,}06 \\ 0{,}12 & 0{,}04 & 0{,}28 & 0{,}08 \\ 0{,}14 & -0{,}06 & 0{,}08 & 0{,}26 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}. \tag{14'}$$

Отсюда

$$\alpha = \begin{bmatrix} 0{,}22 & 0{,}02 & 0{,}12 & 0{,}14 \\ 0{,}02 & 0{,}14 & 0{,}04 & -0{,}06 \\ 0{,}12 & 0{,}04 & 0{,}28 & 0{,}08 \\ 0{,}14 & -0{,}06 & 0{,}08 & 0{,}26 \end{bmatrix} \quad \text{и} \quad \beta = \begin{bmatrix} 0{,}76 \\ 0{,}08 \\ 1{,}12 \\ 0{,}68 \end{bmatrix}.$$

Так как

$$\|\alpha\|_m = \max(0{,}50;\ 0{,}26;\ 0{,}52;\ 0{,}54) = 0{,}54 < 1,$$

то процесс итерации для системы (14) сходится.

Используя в качестве начального вектора $x^{(0)}$ вектор $\beta$, для $m$-го приближения $x^{(m)}$ искомого решения

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

получим следующее выражение:

$$x^{(m)} = \sum_{k=0}^{m} \alpha^k \beta. \tag{15}$$

Таким образом, для вычисления $x^{(m)}$ нужно образовывать последовательные итерации вектора $\beta$ с помощью матрицы $\alpha$. Имеем:

$$\alpha\beta = \begin{bmatrix} 0{,}22 & 0{,}02 & 0{,}12 & 0{,}14 \\ 0{,}02 & 0{,}14 & 0{,}04 & -0{,}06 \\ 0{,}12 & 0{,}04 & 0{,}28 & 0{,}08 \\ 0{,}14 & -0{,}06 & 0{,}08 & 0{,}26 \end{bmatrix} \begin{bmatrix} 0{,}76 \\ 0{,}08 \\ 1{,}12 \\ 0{,}68 \end{bmatrix} = \begin{bmatrix} 0{,}3984 \\ 0{,}0304 \\ 0{,}4624 \\ 0{,}3680 \end{bmatrix};$$

$$\alpha^2\beta = \alpha \cdot \alpha\beta = \begin{bmatrix} 0{,}22 & 0{,}02 & 0{,}12 & 0{,}14 \\ 0{,}02 & 0{,}14 & 0{,}04 & -0{,}06 \\ 0{,}12 & 0{,}04 & 0{,}28 & 0{,}08 \\ 0{,}14 & -0{,}06 & 0{,}08 & 0{,}26 \end{bmatrix} \begin{bmatrix} 0{,}3984 \\ 0{,}0304 \\ 0{,}4624 \\ 0{,}3680 \end{bmatrix} = \begin{bmatrix} 0{,}195264 \\ 0{,}008640 \\ 0{,}207936 \\ 0{,}186624 \end{bmatrix}$$

и т. д.

Результаты соответствующих вычислений приведены в таблице 31.

Таблица 31

**Последовательные итерации вектора $\beta$ матрицей $\alpha$**

| $\beta$ | $\alpha\beta$ | $\alpha^2\beta$ | $\alpha^3\beta$ | $\alpha^4\beta$ |
|---|---|---|---|---|
| 0,76 | 0,3984 | 0,195264 | 0,09421056 | 0,04527913 |
| 0,08 | 0,0304 | 0,008640 | 0,00223488 | 0,00055572 |
| 1,12 | 0,4624 | 0,207936 | 0,09692928 | 0,04589292 |
| 0,68 | 0,3680 | 0,186624 | 0,09197568 | 0,04472340 |

| $\alpha^5\beta$ | $\alpha^6\beta$ | $\alpha^7\beta$ | $\alpha^8\beta$ | $x^{(8)}=\sum_{k=0}^{8}\alpha^k\beta$ |
|---|---|---|---|---|
| 0,02174095 | 0,01043649 | 0,00500961 | 0,00240463 | 1,532746 |
| 0,00013570 | 0,00003285 | 0,00000792 | 0,00000190 | 0,122009 |
| 0,02188361 | 0,01047017 | 0,00501763 | 0,00240654 | 1,972937 |
| 0,02160525 | 0,01040364 | 0,00500170 | 0,00240272 | 1,410737 |

В формуле (11) примем $m=8$. Так как матрица $\alpha$ — симметрическая, то для вычисления ее первого собственного значения $\lambda_1$ используем метод скалярных произведений. Имеем:

$$\lambda_1\approx\frac{(\alpha^8\beta,\ \alpha^8\beta)}{(\alpha^7\beta,\ \alpha^8\beta)}=$$

$$=\frac{240\,463^2+190^2+240\,654^2+240\,272^2}{500\,961\cdot240\,463+792\cdot190+501\,763\cdot240\,654+500\,170\cdot240\,272}=0{,}480000.$$

Отсюда, учитывая, что $x^{(8)}-x^{(7)}=\alpha^8\beta$, находим:

$$x\approx x^{(8)}+\lambda_1\cdot\frac{\alpha^8\beta}{1-\lambda_1}=$$

$$=\begin{bmatrix}1{,}532746\\0{,}122009\\1{,}972937\\1{,}410737\end{bmatrix}+\frac{12}{13}\begin{bmatrix}0{,}002405\\0{,}000002\\0{,}002406\\0{,}002403\end{bmatrix}=\begin{bmatrix}1{,}534965\\0{,}122011\\1{,}975159\\1{,}412955\end{bmatrix}.$$

Для сравнения приводим значения корней системы (11), полученные методом Гаусса [1]:

$$x_1=1{,}534965;\qquad x_2=0{,}122010$$
$$x_3=1{,}975166;\qquad x_4=1{,}412955.$$

Таким образом, если $x^{(8)}$ давало значения корней $x_i$ $(i=1, 2, 3, 4)$ примерно с точностью $1\cdot10^{-3}-2\cdot10^{-3}$, то после поправок

Л. А. Люстерника мы получаем эти корни приблизительно с точностью до $10^{-6}$.

Метод Л. А. Люстерника улучшения сходимости может быть применен также к процессу Зейделя. Как известно, процесс Зейделя для системы (2) представляет собой процесс итерации для равносильной системы

$$x = \beta_1 + \alpha_1 x,$$

где матрица $\alpha_1$ однозначно определяется через матрицу $\alpha$ (см. гл XI, § 3), а именно, если

$$\alpha = B + C,$$

где $B$ — нижняя треугольная матрица с нулевой диагональю и $C$ — верхняя треугольная матрица, то

$$\alpha_1 = (E - B)^{-1} C.$$

Поэтому, если $\xi^{(m)}$ $(m = 1, 2, \ldots)$ — последовательные приближения по Зейделю корня $x$ системы (2), то можно положить:

$$x \approx \xi^{(m)} + \frac{\xi^{(m+1)} - \xi^{(m)}}{1 - \mu_1},$$

где $\mu_1$ — наибольшее по модулю собственное значение матрицы $\alpha_1$.

Отметим, что существуют также и другие методы улучшения сходимости итерационных процессов для решения систем линейных уравнений, например, метод М. К. Гавурина [7], [8], метод А. А. Абрамова [9].

### Литература к двенадцатой главе


1. В. Н. Фаддеева, Вычислительные методы линейной алгебры, Гостехиздат, М., 1950, гл. III.
2. И. М. Гельфанд, Лекции по линейной алгебре, Изд. 2, Гостехиздат, М.—Л., 1951, добавл. I.
3. А. Г. Курош, Курс высшей алгебры, Гостехиздат, М.—Л., 1946, гл. VI.
4. Гарольд Вейленд, Представление векового уравнения в виде многочлена, УМН II, вып. 4 (20) (1947), 128—158.
5. В. Э. Милн, Численный анализ, ИЛ, 1951, гл. II.
6. Л. А. Люстерник, Труды Матем. ин-та им. В. А. Стеклова, 20 (1947), стр. 49.
7. М. К. Гавурин, Применение полиномов наилучшего приближения к улучшению сходимости итеративных процессов, УМН, 5 : 3 (37) (1950), 156—160.
8. И. С. Березин и Н. П. Жидков, Методы вычислений, Физматгиз, 1959, т. 2, гл. VIII.
9. Д. К. Фаддеев и В. Н. Фаддеева, Вычислительные методы линейной алгебры, Физматгиз, 1960, гл. IX.

## ГЛАВА XIII

# ПРИБЛИЖЕННОЕ РЕШЕНИЕ СИСТЕМ НЕЛИНЕЙНЫХ УРАВНЕНИЙ

## § 1. Метод Ньютона

Рассмотрим, вообще говоря, нелинейную систему уравнений

$$\left.\begin{array}{l} f_1(x_1, x_2, \ldots, x_n)=0, \\ f_2(x_1, x_2, \ldots, x_n)=0, \\ \ldots\ldots\ldots\ldots\ldots \\ f_n(x_1, x_2, \ldots, x_n)=0 \end{array}\right\} \qquad (1)$$

с действительными левыми частями.

Запишем короче систему (1). Совокупность аргументов $x_1$, $x_2$, ..., $x_n$ можно рассматривать как $n$-мерный вектор

$$\boldsymbol{x}=\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

Аналогично совокупность функций $f_1, f_2, \ldots, f_n$ представляет собой также $n$-мерный вектор (*вектор-функцию*)

$$\boldsymbol{f}=\begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix}.$$

Поэтому система (1) кратко записывается так:

$$\boldsymbol{f}(\boldsymbol{x})=\boldsymbol{0}. \qquad (1')$$

Для решения системы (1′) будем пользоваться методом последовательных приближений.

Предположим, что найдено $p$-е приближение

$$\boldsymbol{x}^{(p)}=\left(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)}\right)$$

одного из изолированных корней $x=(x_1, x_2, \ldots, x_n)$ векторного уравнения (1′). Тогда точный корень уравнения (1′) можно представить в виде

$$x = x^{(p)} + \varepsilon^{(p)}, \tag{2}$$

где $\varepsilon^{(p)} = (\varepsilon_1^{(p)}, \varepsilon_2^{(p)}, \ldots, \varepsilon_n^{(p)})$ — поправка (погрешность корня).

Подставляя выражение (2) в уравнение (1′), будем иметь:

$$f(x^{(p)} + \varepsilon^{(p)}) = 0. \tag{3}$$

Предполагая, что функция $f(x)$ непрерывно дифференцируема в некоторой выпуклой области, содержащей $x$ и $x^{(p)}$, разложим левую часть уравнения (3) по степеням малого вектора $\varepsilon^{(p)}$, ограничиваясь линейными членами,

$$f(x^{(p)} + \varepsilon^{(p)}) = f(x^{(p)}) + f'(x^{(p)})\,\varepsilon^{(p)} = 0 \tag{4}$$

или, в развернутом виде,

$$\left.\begin{aligned}
&f_1(x_1^{(p)} + \varepsilon_1^{(p)}, x_2^{(p)} + \varepsilon_2^{(p)}, \ldots, x_n^{(p)} + \varepsilon_n^{(p)}) = \\
&\quad = f_1(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)}) + f'_{1x_1}(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})\,\varepsilon_1^{(p)} + \\
&\qquad + f'_{1x_2}(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})\,\varepsilon_2^{(p)} + \ldots \\
&\qquad\qquad \ldots + f'_{1x_n}(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})\,\varepsilon_n^{(p)} = 0, \\
&f_2(x_1^{(p)} + \varepsilon_1^{(p)}, x_2^{(p)} + \varepsilon_2^{(p)}, \ldots, x_n^{(p)} + \varepsilon_n^{(p)}) = \\
&\quad = f_2(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)}) + f'_{2x_1}(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})\,\varepsilon_1^{(p)} + \\
&\qquad + f'_{2x_2}(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})\,\varepsilon_2^{(p)} + \ldots \\
&\qquad\qquad \ldots + f'_{2x_n}(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})\,\varepsilon_n^{(p)} = 0, \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
&f_n(x_1^{(p)} + \varepsilon_1^{(p)}, x_2^{(p)} + \varepsilon_2^{(p)}, \ldots, x_n^{(p)} + \varepsilon_n^{(p)}) = \\
&\quad = f_n(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)}) + f'_{nx_1}(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})\,\varepsilon_1^{(p)} + \\
&\qquad + f'_{nx_2}(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})\,\varepsilon_2^{(p)} + \ldots \\
&\qquad\qquad \ldots + f'_{nx_n}(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})\,\varepsilon_n^{(p)} = 0.
\end{aligned}\right\} \tag{4'}$$

Из формул (4) и (4′) вытекает, что под производной $f'(x)$ следует понимать *матрицу Якоби* системы функций $f_1, f_2, \ldots, f_n$ относительно переменных $x_1, x_2, \ldots, x_n$, т. е.

$$f'(x) = W(x) = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} & \cdots & \dfrac{\partial f_2}{\partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \dfrac{\partial f_n}{\partial x_1} & \dfrac{\partial f_n}{\partial x_2} & \cdots & \dfrac{\partial f_n}{\partial x_n} \end{bmatrix},$$

или в краткой записи

$$f'(x) = W(x) = \left[\frac{\partial f_i}{\partial x_j}\right] \qquad (i,\ j = 1,\ 2,\ \ldots,\ n).$$

Система (4′) представляет собой линейную систему относительно поправок $\varepsilon_i^{(p)}$ $(i = 1,\ 2,\ \ldots,\ n)$ с матрицей $W(x)$, поэтому формула (4) может быть записана в следующем виде:

$$f(x^{(p)}) + W(x^{(p)})\,\varepsilon^{(p)} = \mathbf{0}.$$

Отсюда, предполагая, что матрица $W(x^{(p)})$ — неособенная, получим:

$$\varepsilon^{(p)} = -W^{-1}(x^{(p)})\,f(x^{(p)}).$$

Следовательно,

$$x^{(p+1)} = x^{(p)} - W^{-1}(x^{(p)})\,f(x^{(p)}) \quad (p = 0,\ 1,\ 2,\ \ldots) \tag{5}$$

(*метод Ньютона*).

За нулевое приближение $x^{(0)}$ можно взять грубое значение искомого корня.

Пример 1. Приближенно найти положительные решения системы уравнений (ср. гл. IV, § 9)

$$\left.\begin{aligned} f_1(x_1,\ x_2) &\equiv x_1 + 3\lg x_1 - x_2^2 = 0, \\ f_2(x_1,\ x_2) &\equiv 2x_1^2 - x_1x_2 - 5x_1 + 1 = 0. \end{aligned}\right\} \tag{6}$$

Решение. Кривые, определяемые системой (6), пересекаются приблизительно в точках $M_1$ (1,4; — 1,5) и $M_2$ (3,4; 2,2). Исходя из начального приближения

$$x^{(0)} = \begin{bmatrix} 3{,}4 \\ 2{,}2 \end{bmatrix},$$

вычислим вторые приближения корней, производя вычисления с точностью до четырех десятичных знаков. Полагая

$$f(x) = \begin{bmatrix} f_1(x_1,\ x_2) \\ f_2(x_1,\ x_2) \end{bmatrix},$$

имеем:

$$f(x^{(0)}) = \begin{bmatrix} 3{,}4 + 3\lg 3{,}4 - 2{,}2^2 \\ 2\cdot 3{,}4^2 - 3{,}4\cdot 2{,}2 - 5\cdot 3{,}4 + 1 \end{bmatrix} = \begin{bmatrix} 0{,}1544 \\ -0{,}3600 \end{bmatrix}.$$

Составим теперь матрицу Якоби

$$W(x) = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} \\ \dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 1 + \dfrac{3M}{x_1} & -2x_2 \\ 4x_1 - x_2 - 5 & -x_1 \end{bmatrix},$$

где $M=0{,}43429$. Отсюда

$$W(x^{(0)})=\begin{bmatrix}1+\dfrac{3\cdot 0{,}43429}{3{,}4} & -2\cdot 2{,}2\\ 4\cdot 3{,}4-2{,}2-5 & -3{,}4\end{bmatrix}=\begin{bmatrix}1{,}3832 & -4{,}4\\ 6{,}4 & -3{,}4\end{bmatrix},$$

причем

$$\Delta=\det W(x^{(0)})=23{,}4571.$$

Следовательно, матрица $W(x^{(0)})$ — неособенная. Составим обратную ей матрицу

$$W^{-1}(x^{(0)})=\frac{1}{\Delta}\begin{bmatrix}-3{,}4 & 4{,}4\\ -6{,}4 & 1{,}3832\end{bmatrix}.$$

Используя формулу (5), получим:

$$x^{(1)}=\begin{bmatrix}3{,}4\\ 2{,}2\end{bmatrix}-\frac{1}{23{,}4571}\begin{bmatrix}-3{,}4 & 4{,}4\\ -6{,}4 & 1{,}3832\end{bmatrix}\begin{bmatrix}0{,}1544\\ -0{,}3600\end{bmatrix}=$$

$$=\begin{bmatrix}3{,}4\\ 2{,}2\end{bmatrix}-\frac{1}{23{,}4571}\begin{bmatrix}-2{,}10896\\ -1{,}48604\end{bmatrix}=\begin{bmatrix}3{,}4\\ 2{,}2\end{bmatrix}+\begin{bmatrix}0{,}0899\\ 0{,}0633\end{bmatrix}=\begin{bmatrix}3{,}4899\\ 2{,}2633\end{bmatrix}.$$

Аналогично находятся дальнейшие приближения. Результаты вычислений приведены в таблице 32.

Таблица 32

**Последовательные приближения корней системы (6)**

| $i$ | $x_1$ | $\varepsilon_1=\Delta x_1$ | $x_2$ | $\varepsilon_2=\Delta x_2$ |
|---|---|---|---|---|
| 0 | 3,4 | 0,0899 | 2,2 | 0,0633 |
| 1 | 3,4899 | −0,0008 | 2,2633 | −0,0012 |
| 2 | 3,4891 | −0,0016 | 2,2621 | −0,0005 |
| 3 | 3,4875 | | 2,2616 | |

Останавливаясь на приближении $x^{(3)}$, будем иметь:

$$x_1=3{,}4875;\qquad x_2=2{,}2616,$$

причем

$$f(x^{(3)})=\begin{bmatrix}0{,}0002\\ 0{,}0000\end{bmatrix}.$$

Пример 2. Методом Ньютона приближенно найти положительное решение системы уравнений

$$\left.\begin{aligned}x^2+y^2+z^2&=1,\\ 2x^2+y^2-4z&=0,\\ 3x^2-4y+z^2&=0,\end{aligned}\right\}$$

исходя из начального приближения

$$x_0=y_0=z_0=0{,}5.$$

Решение. Имеем:

$$f(x)=\begin{bmatrix} x^2+y^2+z^2-1 \\ 2x^2+y^2-4z \\ 3x^2-4y+z^2 \end{bmatrix}.$$

Отсюда

$$f(x^{(0)})=\begin{bmatrix} 0{,}25+0{,}25+0{,}25-1 \\ 0{,}50+0{,}25-2{,}00 \\ 0{,}75-2{,}00+0{,}25 \end{bmatrix}=\begin{bmatrix} -0{,}25 \\ -1{,}25 \\ -1{,}00 \end{bmatrix}.$$

Составим матрицу Якоби

$$W(x)=\begin{bmatrix} 2x & 2y & 2z \\ 4x & 2y & -4 \\ 6x & -4 & 2z \end{bmatrix}.$$

Имеем

$$W(x^{(0)})=\begin{bmatrix} 1 & 1 & 1 \\ 2 & 1 & -4 \\ 3 & -4 & 1 \end{bmatrix}$$

и

$$\det W(x^{(0)})=\begin{vmatrix} 1 & 1 & 1 \\ 2 & 1 & -4 \\ 3 & -4 & 1 \end{vmatrix}=-40.$$

Находим обратную матрицу

$$W^{-1}(x^{(0)})=-\frac{1}{40}\begin{bmatrix} -15 & -5 & -5 \\ -14 & -2 & 6 \\ -11 & 7 & -1 \end{bmatrix}=\begin{bmatrix} \frac{3}{8} & \frac{1}{8} & \frac{1}{8} \\ \frac{7}{20} & \frac{1}{20} & -\frac{3}{20} \\ \frac{11}{40} & -\frac{7}{40} & \frac{1}{40} \end{bmatrix}.$$

По формуле (5) получаем первое приближение

$$x^{(1)}=x^{(0)}-W^{-1}(x^{(0)})f(x^{(0)})=$$

$$=\begin{bmatrix} 0{,}5 \\ 0{,}5 \\ 0{,}5 \end{bmatrix}-\begin{bmatrix} \frac{3}{8} & \frac{1}{8} & \frac{1}{8} \\ \frac{7}{20} & \frac{1}{20} & \frac{-3}{20} \\ \frac{11}{40} & \frac{-7}{40} & \frac{1}{40} \end{bmatrix}\begin{bmatrix} -0{,}25 \\ -1{,}25 \\ -1{,}00 \end{bmatrix}=$$

$$=\begin{bmatrix} 0{,}5 \\ 0{,}5 \\ 0{,}5 \end{bmatrix}+\begin{bmatrix} 0{,}375 \\ 0 \\ -0{,}125 \end{bmatrix}=\begin{bmatrix} 0{,}875 \\ 0{,}500 \\ 0{,}375 \end{bmatrix}.$$

Далее вычисляем второе приближение $x^{(2)}$. Имеем:

$$f(x^{(1)})=\begin{bmatrix}0,875^2+0,500^2+0,375^2-1\\2\cdot0,875^2+0,500^2-4\cdot0,375\\3\cdot0,875^2-4\cdot0,500+0,375^2\end{bmatrix}=\begin{bmatrix}0,15625\\0,28125\\0,43750\end{bmatrix}$$

и

$$W(x^{(1)})=\begin{bmatrix}2\cdot0,875 & 2\cdot0,500 & 2\cdot0,375\\4\cdot0,875 & 2\cdot0,500 & -4\\6\cdot0,875 & -4 & 2\cdot0,375\end{bmatrix}=$$

$$=\begin{bmatrix}1,750 & 1 & 0,750\\3,500 & 1 & -4\\5,250 & -4 & 0,750\end{bmatrix}.$$

Отсюда

$$\det W(x^{(1)})=\begin{vmatrix}1,750 & 1 & 0,750\\3,500 & 1 & -4\\5,250 & -4 & 0,750\end{vmatrix}=$$

$$=\begin{vmatrix}1,750 & 1 & 0,750\\1,750 & 0 & -4,750\\12,250 & 0 & 3,750\end{vmatrix}=-64,75$$

и

$$W^{-1}(x^{(1)})=-\frac{1}{64,75}\begin{bmatrix}-15,25 & -3,75 & -4,75\\-23,625 & -2,6250 & 9,625\\-19,25 & 12,25 & -1,75\end{bmatrix}.$$

Используя формулу (5), получаем:

$$x^{(2)}=x^{(1)}-W^{-1}(x^{(1)})f(x^{(1)})=$$

$$=\begin{bmatrix}0,875\\0,500\\0,375\end{bmatrix}+\frac{1}{64,75}\begin{bmatrix}-15,25 & -3,75 & -4,75\\-23,625 & -2,6250 & 9,625\\-19,25 & 12,25 & -1,75\end{bmatrix}\begin{bmatrix}0,15625\\0,28125\\0,43750\end{bmatrix}=$$

$$=\begin{bmatrix}0,875\\0,500\\0,375\end{bmatrix}-\begin{bmatrix}0,08519\\0,00338\\0,00507\end{bmatrix}=\begin{bmatrix}0,78981\\0,49662\\0,36993\end{bmatrix}.$$

Аналогично находятся дальнейшие приближения:

$$x^{(3)}=\begin{bmatrix}0,78521\\0,49662\\0,36992\end{bmatrix},\quad f(x^{(3)})=\begin{bmatrix}0,00001\\0,00004\\0,00005\end{bmatrix}$$

и т. д.

Ограничиваясь третьим приближением, получим:

$$x=0,7852;\quad y=0,4966;\quad z=0,3699.$$

## § 2. Общие замечания о сходимости процесса Ньютона

В § 1 был дан формальный аспект метода Ньютона. Условия сходимости этого метода для системы исследованы Виллерсом, Стениным, Островским, Канторовичем и др. Ниже излагается частный случай теоремы Л. В. Канторовича (теорема 1) [1] о сходимости процесса Ньютона в функциональных пространствах применительно к конечным системам нелинейных уравнений, причем для простоты рассуждений используются более грубые оценки. Следуя Л. В. Канторовичу, устанавливаем также быстроту сходимости процесса Ньютона, единственность корня системы и устойчивость процесса по отношению к выбору начального приближения (теоремы 2—4). Как частный случай получается теорема Островского [2] о сходимости процесса Ньютона для уравнения с аналитической комплексной правой частью.

В дальнейшем совокупности функций будет удобно рассматривать как *вектор-функцию* или *матричную функцию*. Для облегчения изложения обобщим понятие производной на эти случаи.

Пусть $x=(x_1, \ldots, x_n)$ и

$$f(x)=\begin{bmatrix} f_1(x) \\ \cdot \\ \cdot \\ \cdot \\ f_n(x) \end{bmatrix},$$

где $f_i \in C^{(1)}$ $(i=1, 2, \ldots, n)$.

Определение 1. Под производной $f'(x)$ понимается матрица Якоби системы функций $f_i$ $(i=1,\ldots,n)$ по переменным $x_1, \ldots, x_n$, т. е.

$$f'(x)=\left[\frac{\partial f_i}{\partial x_j}\right]. \tag{1}$$

Матричную функцию

$$F(x)=\begin{bmatrix} f_{11}(x)\ldots f_{1r}(x) \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ f_{n1}(x)\ldots f_{nr}(x) \end{bmatrix}$$

можно рассматривать как совокупность $m$ вектор-функций

$$F_1(x)=\begin{bmatrix} f_{11}(x) \\ \cdot \\ \cdot \\ \cdot \\ f_{n1}(x) \end{bmatrix}, \ldots, F_r(x)=\begin{bmatrix} f_{1r}(x) \\ \cdot \\ \cdot \\ \cdot \\ f_{nr}(x) \end{bmatrix}.$$

Поэтому под производной $F'(x)$ естественно понимать совокупность

$$F'(x)=[F_1'(x) \ \ldots \ F_r'(x)],$$

где

$$F_k'(x)=\begin{bmatrix}\frac{\partial f_{1k}}{\partial x_1}\dots\frac{\partial f_{1k}}{\partial x_n}\\ \dots\dots\\ \frac{\partial f_{nk}}{\partial x_1}\dots\frac{\partial f_{nk}}{\partial x_n}\end{bmatrix}$$

— матрицы Якоби ($k=1, 2, \dots, r$).

Определение 2. Если $F(x)=[f_{ij}(x)]$ — функциональная матрица типа $n\times r$ и $f_{ij}(x)\in C^{(1)}$, то

$$F'(x)=[F'_k(x)], \tag{2}$$

где

$$F_k'(x)=\left[\frac{\partial f_{ik}}{\partial x_j}\right] \quad (i,\ j=1, 2, \dots, n;\ k=1, 2, \dots, r).$$

В частности, если вектор-функция $f(x)=[f_i(x)]$ такова, что $f_i(x)\in C^{(2)}$, то

$$f''(x)=[W_1(x)\dots W_n(x)],$$

где

$$W_k(x)=\left[\frac{\partial^2 f_i}{\partial x_k\partial x_j}\right] \quad (k=1, 2, \dots, n).$$

В этом параграфе для оценки матриц мы будем пользоваться $m$-нормой (гл. VII, § 7), причем значок $m$ для краткости опускается, а именно:

$$\|f(x)\|=\max_i|f_i(x)|;$$

$$\|f'(x)\|=\max_i\sum_{j=1}^n\left|\frac{\partial f_i(x)}{\partial x_j}\right|;$$

$$\|f''(x)\|=\max_k\|W_k(x)\|=\max_k\left\{\max_i\sum_{j=1}^n\left|\frac{\partial^2 f_i(x)}{\partial x_k\partial x_j}\right|\right\} \text{ и т. п.}$$

Аналогично

$$\|F(x)\|=\max_i\sum_{j=1}^r|f_{ij}(x)|$$

$$\|F'(x)\|=\max_{i,j}\sum_{k=1}^n\left|\frac{\partial f_{ij}(x)}{\partial x_k}\right| \text{ *)}.$$

---

*) Так как, очевидно, для любой конечной совокупности чисел $\{a_{ij}\}$ имеем:

$$\max_i\left(\max_j a_{ij}\right)=\max_{i,j}a_{ij}.$$

Предварительно выведем несколько оценок для $m$-норм разностей значений матричных функций, аналогичных формуле конечного приращения, которые окажутся полезными в дальнейшем (ср. [1]).

Лемма 1. *Если*

$$F(x)=[f_{ij}(x)] \qquad (n\times r),$$

*где $f_{ij}(x)$ непрерывны вместе со своими частными производными первого порядка в выпуклой области, содержащей точки $x$ и $x+\Delta x$, то*

$$\|F(x+\Delta x)-F(x)\|\leqslant r\|\Delta x\|\cdot\|F'(\xi)\|, \tag{3}$$

*где $\xi=x+\theta\Delta x$, $0<\theta<1$ и норма матриц понимается в смысле $m$-нормы.*

Доказательство. Применяя формулу Тейлора, получим:

$$F(x+\Delta x)-F(x)=[f_{ij}(x+\Delta x)-f_{ij}(x)]=\left[\sum_{k=1}^{n}\frac{\partial f_{ij}(\xi_{ij})}{\partial x_k}\Delta x_k\right],$$

где $\xi_{ij}=x+\theta_{ij}\Delta x$, $0<\theta_{ij}<1$; $i=1, 2, \ldots, n$; $j=1, 2, \ldots, r$. Отсюда, фиксируя $x$ и $x+\Delta x$, будем иметь:

$$\begin{aligned}\|F(x+\Delta x)-F(x)\|&=\\ &=\max_i\sum_{j=1}^{r}\left|\sum_{k=1}^{n}\frac{\partial f_{ij}(\xi_{ij})}{\partial x_k}\Delta x_k\right|\leqslant\\ &\leqslant\max_i\sum_{j=1}^{r}\sum_{k=1}^{n}\left|\frac{\partial f_{ij}(\xi_{ij})}{\partial x_k}\right|\,|\Delta x_k|\leqslant\\ &\leqslant\max_k|\Delta x_k|\cdot\sum_{j=1}^{r}\max_{i,j}\sum_{k=1}^{n}\left|\frac{\partial f_{ij}(\xi_{ij})}{\partial x_k}\right|=\\ &=r\|\Delta x\|\max_{i,j}\sum_{k=1}^{n}\left|\frac{\partial f_{ij}(\xi_{ij})}{\partial x_k}\right|.\end{aligned}$$

Так как число пар $(i, j)$ конечно, то найдется пара $(p, q)$ такая, что

$$\max_{i,j}\sum_{k=1}^{n}\left|\frac{\partial f_{ij}(\xi_{ij})}{\partial x_k}\right|=\sum_{k=1}^{n}\left|\frac{\partial f_{pq}(\xi_{pq})}{\partial x_k}\right|\leqslant\max_{i,j}\sum_{k=1}^{n}\left|\frac{\partial f_{ij}(\xi_{pq})}{\partial x_k}\right|=\|F'(\xi)\|,$$

где $\xi=\xi_{pq}$.

Таким образом,

$$\|F(x+\Delta x)-F(x)\|\leqslant r\|\Delta x\|\,\|F'(\xi)\|,$$

что и требовалось доказать.

Следствие 1. Если

$$f(x)=\begin{bmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{bmatrix},$$

то

$$\|f(x+\Delta x)-f(x)\| \leqslant \|\Delta x\|\cdot\|f'(\xi)\|,$$

где $\xi=x+\theta\Delta x$ и $0<\theta<1$.

Здесь $r=1$.

Следствие 2. При $f(x)\in C^{(2)}$ имеем:

$$\|f'(x+\Delta x)-f'(x)\| \leqslant n\|\Delta x\|\,\|f''(\xi)\|.$$

где $\xi=x+\theta\Delta x$ и $0<\theta<1$.

Лемма 2. *Если*

$$f(x)=\begin{bmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{bmatrix}\in C^{(2)}$$

*в выпуклой области, содержащей точки* $x$ *и* $x+\Delta x$, *то*

$$\|f(x+\Delta x)-f(x)-f'(x)\,\Delta x\| \leqslant \frac{1}{2}n\|\Delta x\|^2\cdot\|f''(\xi)\|, \tag{4}$$

*где* $\xi=x+\theta\Delta x$ *и* $0<\theta<1$.

Доказательство. Используя двучленную формулу Тейлора, получаем:

$$\begin{aligned}
&\|f(x+\Delta x)-f(x)-f'(x)\,\Delta x\| = \\
&\quad = \|[f_i(x+\Delta x)-f_i(x)-df_i(x_i)]\| = \\
&\quad = \frac{1}{2}\left\|\left[\sum_{j,\,k}\frac{\partial^2 f_i(\xi_i)}{\partial x_j\,\partial x_k}\,\Delta x_j\,\Delta x_k\right]\right\| \leqslant \\
&\quad \leqslant \frac{1}{2}\left\|\left[\sum_j |\Delta x_j|\sum_k\left|\frac{\partial^2 f_i(\xi_i)}{\partial x_j\,\partial x_k}\right|\,|\Delta x_k|\right]\right\| \leqslant \\
&\quad \leqslant \frac{1}{2}\max_j|\Delta x_j|\cdot\max_k|\Delta x_k|\cdot\left\|\left[\sum_j\sum_k\left|\frac{\partial^2 f_i(\xi_i)}{\partial x_j\,\partial x_k}\right|\right]\right\| = \\
&\quad = \frac{1}{2}\|\Delta x\|^2\left\|\left[\sum_j\sum_k\left|\frac{\partial^2 f_i(\xi_i)}{\partial x_j\,\partial x_k}\right|\right]\right\|,
\end{aligned} \tag{5}$$

где $\xi_i=x+\theta_i\Delta x$, $0<\theta_i<1$.

Так как

$$\sum_k \left| \frac{\partial^2 f_i(\xi_i)}{\partial x_j \partial x_k} \right| \leqslant \max_{i,j} \sum_k \left| \frac{\partial^2 f_i(\xi_i)}{\partial x_j \partial x_k} \right| =$$

$$= \sum_k \left| \frac{\partial^2 f_p(\xi_p)}{\partial x_q \partial x_k} \right| \leqslant \max_{i,j} \sum_k \left| \frac{\partial^2 f_i(\xi_p)}{\partial x_j \partial x_k} \right| = \|f''(\xi_p)\|,$$

то из неравенства (5), учитывая смысл нормы, получаем:

$$\|f(x+\Delta x) - f(x) - f'(x)\Delta x\| \leqslant \frac{1}{2}\|\Delta x\|^2 [\|f''(\xi)\|] =$$

$$= \frac{n}{2}\|\Delta x\|^2 \|f''(\xi)\|,$$

где $\xi = \xi_p = x + \theta\Delta x$ и $0 < \theta < 1$.

## § 3*. Существование корней системы и сходимость процесса Ньютона

Теорема 1. *Пусть дана нелинейная система алгебраических или трансцендентных уравнений с действительными коэффициентами*

$$f(x) = 0, \tag{1}$$

*где вектор-функция*

$$f(x) = \begin{bmatrix} f_1(x_1, \ldots, x_n) \\ \cdots\cdots \\ f_n(x_1, \ldots, x_n) \end{bmatrix}$$

*определена и непрерывна, вместе со своими частными производными первого и второго порядков, в некоторой области* $\omega$, *т. е.*

$$f(x) \in C^{(2)}(\omega).$$

*Положим, что* $x^{(0)}$ *есть точка, лежащая в* $\omega$ *вместе со своей замкнутой* $\mathcal{H}$*-окрестностью:*

$$\overline{U}_{\mathcal{H}}(x^{(0)}) = \{\|x - x^{(0)}\| \leqslant \mathcal{H}\} \subset \omega,$$

*где норма понимается в смысле m-нормы*)* (см. гл. VII, § 7), *причем выполнены следующие условия:*

---

*) То есть если $A = [a_{ij}]$, то

$$\|A\| = \|A\|_m = \max_i \sum_j |a_{ij}|.$$

1) *матрица Якоби* $W(x)=\left[\frac{\partial f_i}{\partial x_j}\right]$ *при* $x=x^{(0)}$ *имеет обратную* $\Gamma_0=W^{-1}(x^{(0)})$, *где*

$$\|\Gamma_0\|\leqslant A_0{}^{*});$$

2) $\|\Gamma_0 f(x^{(0)})\|\leqslant B_0\leqslant\frac{\mathscr{H}}{2}$;

3) $\sum\limits_{k=1}^{n}\left|\frac{\partial^2 f_i(x)}{\partial x_j\partial x_k}\right|\leqslant C$

*при* $i,\ j=1,\ 2,\ \ldots,\ n$ *и* $x\in\overline{U}_{\mathscr{H}}(x^{(0)})$;

4) *постоянные* $A_0$, $B_0$ *и* $C$ *удовлетворяют неравенству*

$$\mu_0=2nA_0B_0C\leqslant 1. \tag{2}$$

*Тогда процесс Ньютона*

$$x^{(p+1)}=x^{(p)}-W^{-1}(x^{(p)})f(x^{(p)}) \tag{3}$$

$(p=0,\ 1,\ 2,\ \ldots)$ *при начальном приближении* $x^{(0)}$ *сходится и предельный вектор*

$$x^*=\lim_{p\to\infty}x^{(p)}$$

*есть решение системы* (1) *такое, что*

$$\|x^*-x^{(0)}\|\leqslant 2B_0\leqslant\mathscr{H}.$$

Доказательство. Введем обозначения

$$h_p=\|x^{(p+1)}-x^{(p)}\|=\max_k\left|x_k^{(p+1)}-x_k^{(p)}\right|,$$

$$\Gamma_p=W^{-1}(x^{(p)})\quad(p=0,\ 1,\ 2,\ \ldots).$$

Из формулы (3) имеем:

$$h_p=\|\Gamma_p f(x^{(p)})\|.$$

Исходя из условий 1)—4), получим оценки для величин $\Gamma_p$ и $\Gamma_p f(x^{(p)})$.

Рассмотрим сначала случай $p=1$. Используя условие 2), имеем

$$h_0=\|x^{(1)}-x^{(0)}\|=\|W^{-1}(x^{(0)})f(x^{(0)})\|\leqslant B_0\leqslant\frac{\mathscr{H}}{2};$$

---

*) Иными словами, если $W(x^{(0)})=[a_{ij}]$, то $\Gamma_0=W^{-1}(x^{(0)})=\left[\frac{A_{ji}}{\Delta}\right]$, где $A_{ij}$ — алгебраические дополнения элементов $a_{ij}$ и $\Delta=\det[a_{ij}]$ и, следовательно,

$$\|\Gamma_0\|=\max_i\frac{1}{|\Delta|}\sum_{j=1}^{n}|A_{ji}|.$$

следовательно,

$$h_0 \leqslant B_0$$

и

$$\overline{U}_{\frac{\mathcal{H}}{2}}(x^{(1)}) \subset \overline{U}_{\mathcal{H}}(x^{(0)}).$$

Для оценки $\Gamma_1 = W^{-1}(x^{(1)})$, воспользовавшись соотношением $(AB)^{-1} = B^{-1}A^{-1}$, представим эту величину в виде

$$\Gamma_1 = [W(x^{(0)}) \cdot \Gamma_0 W(x^{(1)})]^{-1} = [\Gamma_0 W(x^{(1)})]^{-1} \cdot \Gamma_0. \tag{4}$$

Учитывая условие 1) теоремы, имеем:

$$\|E - \Gamma_0 W(x^{(1)})\| = \|\Gamma_0 [W(x^{(0)}) - W(x^{(1)})]\| \leqslant$$
$$\leqslant \|\Gamma_0\| \, \|W(x^{(0)}) - W(x^{(1)})\| \leqslant A_0 \|W(x^{(1)}) - W(x^{(0)})\|.$$

Так как из условия 3) вытекает, что

$$\|f''(x)\| = \max_{i,j} \sum_{k=1}^{n} \left| \frac{\partial^2 f_i(x)}{\partial x_j \partial x_k} \right| \leqslant C,$$

то в силу следствия 2 к лемме 1 получаем:

$$\|W(x^{(1)}) - W(x^{(0)})\| = \|f'(x^{(1)}) - f'(x^{(0)})\| \leqslant n \|x^{(1)} - x^{(0)}\| C \leqslant n B_0 C;$$

поэтому

$$\|E - \Gamma_0 W(x^{(1)})\| \leqslant n A_0 B_0 C = \frac{\mu_0}{2} \leqslant \frac{1}{2}.$$

Следовательно (гл. VII, § 10, теорема 5, следствие), существует обратная матрица

$$[\Gamma_0 W(x^{(1)})]^{-1} = \{E - (E - \Gamma_0 W(x^{(1)}))\}^{-1},$$

причем, так как $\|E\| = \|E\|_m = 1$, то

$$\|[\Gamma_0 W(x^{(1)})]^{-1}\| \leqslant \frac{1}{1 - \frac{\mu_0}{2}} \leqslant 2. \tag{5}$$

Теперь из формулы (4) выводим:

$$\|\Gamma_1\| \leqslant \|[\Gamma_0 W(x^{(1)})]^{-1}\| \, \|\Gamma\|_0 \| \leqslant 2A_0 = A_1. \tag{6}$$

Далее, из формулы (3) следует:

$$f(x^{(0)} + f')(x^{(0)})(x^{(1)} - x^{(0)}) = 0.$$

Отсюда на основании леммы 2 будем иметь:

$$\|f(x^{(1)})\| = \|f(x^{(1)}) - f(x^{(0)}) - f'(x^{(0)})(x^{(1)} - x^{(0)})\| \leqslant$$
$$\leqslant \frac{1}{2} n \|x^{(1)} - x^{(0)}\|^2 \|f''(\xi)\| \leqslant \frac{1}{2} n B_0^2 C,$$

где

$$\xi = x^{(0)} + \theta (x^{(1)} - x^{(0)}) \quad \text{и} \quad 0 < \theta < 1.$$

Поэтому, учитывая неравенство (6), получим:

$$\| \Gamma_1 f(x^{(1)}) \| \leqslant \| \Gamma_1 \| \, \| f(x^{(1)}) \| \leqslant$$
$$\leqslant 2A_0 \cdot \frac{1}{2} nB_0^2 C = nA_0 B_0^2 C = \frac{1}{2} \mu_0 B_0 = B_1. \qquad (7)$$

Итак, для точки $x^{(1)}$ мы имеем:

$$\overline{U}_{\frac{\mathcal{H}}{2}}(x^{(1)}) \subset \overline{U}_{\mathcal{H}}(x^{(0)}) \subset \omega$$

и, кроме того,

$$\| \Gamma_1 \| \leqslant A_1, \qquad h_1 = \| \Gamma_1 f(x^{(1)}) \| \leqslant B_1,$$

где

$$A_1 = 2A_0,$$
$$B_1 = \frac{1}{2} \mu_0 B_0 \leqslant \frac{\mathcal{H}}{4}.$$

Отсюда получаем:

$$\mu_1 = 2nA_1 B_1 C = 2n \cdot 2A_0 \cdot \frac{1}{2} \mu_0 B_0 C = \mu_0 \cdot 2nA_0 B_0 C = \mu_0^2 \leqslant 1. \qquad (8)$$

Следовательно, мы снова находимся в условиях теоремы с той только разницей, что вместо окрестности $\overline{U}_{\mathcal{H}}(x^{(0)})$ имеем окрестность $\overline{U}_{\frac{\mathcal{H}}{2}}(x^{(1)})$, вложенную в первую.

Повторяя аналогичные рассуждения, мы установим, что последовательные приближения $x^{(p)}$ $(p = 1, 2, \ldots)$ имеют смысл и таковы, что

$$\overline{U}_{\mathcal{H}}(x^{(0)}) \supset \overline{U}_{\frac{\mathcal{H}}{2}}(x^{(1)}) \supset \ldots \supset \overline{U}_{\frac{\mathcal{H}}{2^p}}(x^{(p)}) \supset \ldots,$$

причем

$$\| \Gamma_p \| = \| W^{-1}(x^{(p)}) \| \leqslant A_p,$$
$$\| \Gamma_p f(x^{(p)}) \| = \| x^{(p+1)} - x^{(p)} \| \leqslant B_p,$$

где постоянные $A_p$ и $B_p$ связаны между собой рекуррентными соотношениями

$$\left.\begin{aligned} A_p &= 2A_{p-1}, \\ B_p &= \frac{1}{2} \mu_{p-1} B_{p-1} \end{aligned}\right\} \qquad (9)$$

и

$$\mu_p = 2nA_p B_p C \quad (p = 1, 2, \ldots). \qquad (10)$$

Покажем, что для последовательности приближений $x^{(p)}$ ($p=0$, 1, 2, ...) выполнен критерий Коши (гл. VII, § 9). Действительно, при $q>0$ имеем:

$$x^{(p+q)} \in \overline{U}_{\frac{\mathcal{H}}{2^p}}(x^{(p)}).$$

Поэтому

$$\| x^{(p+q)} - x^{(p)} \| \leqslant \frac{\mathcal{H}}{2^p} < \varepsilon,$$

если $p>N$ и $q>0$, что эквивалентно критерию Коши. Отсюда следует, что существует

$$\lim_{p\to\infty} x^{(p)} = x^* \in \overline{U}_{\mathcal{H}}(x^{(0)}).$$

Убедимся теперь, что $x^*$ есть решение системы (1). Из соотношения (3) имеем:

$$f(x^{(p)}) + W(x^{(p)})(x^{(p+1)} - x^{(p)}) = \mathbf{0}.$$

Переходя в этом равенстве к пределу при $p\to\infty$ и учитывая, что при этом

$$x^{(p+1)} - x^{(p)} \to \mathbf{0},$$

а также, что $W(x^{(p)})$ непрерывна и ограничена в $\overline{U}_{\mathcal{H}}(x^{(0)})$, будем иметь:

$$\lim_{p\to\infty} f(x^{(p)}) = \mathbf{0}.$$

Отсюда в силу непрерывности функции $f(x)$ получим:

$$f\left(\lim_{p\to\infty} x^{(p)}\right) = f(x^*) = \mathbf{0},$$

т. е. $x^*$ есть решение системы (1). Кроме того,

$$\| x^* - x^{(0)} \| = \left\| \sum_{p=0}^{\infty} [x^{(p+1)} - x^{(p)}] \right\| \leqslant$$

$$\leqslant \sum_{p=0}^{\infty} \| x^{(p+1)} - x^{(p)} \| \leqslant \sum_{p=0}^{\infty} B_p \leqslant B_0 + \frac{B_0}{2} + \ldots = 2B_0 \leqslant \mathcal{H}.$$

Теорема доказана полностью.

Замечание 1. Если $f(x) \in C^{(2)}(\omega)$ и в области $\omega$ система (1) имеет простое решение $x^*$, т. е. такое, что

$$f(x^*) = \mathbf{0},\ f'(x^*) = W(x^*) \neq 0,$$

то для каждой точки $x^{(0)}$, достаточно близкой к $x^*$, условия теоремы 1, очевидно, будут выполнены.

Для проверки условия 2) полезно отметить, что $B_0$ дает оценку расхождения начального и первого приближений процесса Ньютона:

$$\| \Gamma_0 f(x^{(0)}) \| = \| x^{(1)} - x^{(0)} \| \leqslant B_0,$$

и поэтому это неравенство легко может быть проверено, как только будет найдено приближение $x^{(1)}$.

Замечание 2. Аналогичные формулировки для теоремы сходимости получаются, если вместо нормы $\|A\|_m$ использовать нормы $\|A\|_l$ или $\|A\|_k$.

## § 4*. Быстрота сходимости процесса Ньютона

Теорема 2. *Если выполнены условия* 1) — 4) *теоремы* 1 *из* § 3, *то для последовательных приближений* $x^{(p)}$ $(p=0, 1, 2, \ldots)$ *справедливо неравенство*

$$\| x^* - x^{(p)} \| \leqslant \left(\frac{1}{2}\right)^{p-1} \mu_0^{2^p-1} B_0,$$

*где* $x^*$ — *решение системы и* $\mu_0$ *определяется формулой* (2) *из* § 3.

Доказательство. Используя соотношения (9) и (10) из § 3, имеем:

$$\mu_p = 2nA_pB_pC = 2n \cdot 2A_{p-1} \cdot \frac{1}{2} \mu_{p-1}B_{p-1} \cdot C =$$
$$= \mu_{p-1} \cdot 2nA_{p-1}B_{p-1}C = \mu_{p-1}^2.$$

Отсюда получаем

$$\left.\begin{aligned} &\mu_1 = \mu_0^2, \\ &\mu_2 = \mu_1^2 = \mu_0^4, \\ &\cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ &\mu_p = \mu_0^{2^p}. \end{aligned}\right\} \qquad (1)$$

Далее,

$$B_p = \frac{1}{2} \mu_{p-1} B_{p-1} = \frac{1}{2} \mu_0^{2^{p-1}} B_{p-1}.$$

Поэтому

$$B_p = \frac{1}{2} \mu_0^{2^{p-1}} \cdot \frac{1}{2} \mu_0^{2^{p-2}} \ldots \frac{1}{2} \mu_0^{2^0} B_0 =$$
$$= \left(\frac{1}{2}\right)^p \cdot \mu_0^{2^{p-1}+2^{p-2}+\cdots+1} B_0 = \left(\frac{1}{2}\right)^p \mu_0^{2^p-1} B_0. \qquad (2)$$

Так как

$$\| x^{(p+1)} - x^{(p)} \| \leqslant B_p,$$

то при $q > 1$ имеем:

$$\begin{aligned}\| x^{(p+q)} - x^{(p)} \| &\leqslant \| x^{(p+1)} - x^{(p)} \| + \\ &+ \| x^{(p+2)} - x^{(p+1)} \| + \ldots + \| x^{(p+q)} - x^{(p+q-1)} \| \leqslant \\ &\leqslant B_p + B_{p+1} + \ldots + B_{p+q-1} = \\ &= \left(\frac{1}{2}\right)^p \mu_0^{2^p-1} B_0 + \left(\frac{1}{2}\right)^{p+1} \mu_0^{2^{p+1}-1} B_0 + \ldots + \left(\frac{1}{2}\right)^{p+q-1} \mu_0^{2^{p+q-1}-1} B_0 = \\ &= \left(\frac{1}{2}\right)^p \mu_0^{2^p-1} B_0 \left[1 + \frac{1}{2} \cdot \mu_0^{2^p} + \ldots + \left(\frac{1}{2}\right)^{q-1} \mu_0^{2^p(2^{q-1}-1)}\right].\end{aligned}$$

Отсюда, учитывая, что $\mu_0 \leqslant 1$, получаем:

$$\begin{aligned}\| x^{(p+q)} - x^{(p)} \| &\leqslant \\ &\leqslant \left(\frac{1}{2}\right)^p \mu_0^{2^p-1} B_0 \left[1 + \frac{1}{2} + \ldots + \left(\frac{1}{2}\right)^{q-1}\right] \leqslant \left(\frac{1}{2}\right)^{p-1} \mu_0^{2^p-1} B_0.\end{aligned}$$

Переходя к пределу при $q \to \infty$, окончательно находим:

$$\| x^* - x^{(p)} \| \leqslant \left(\frac{1}{2}\right)^{p-1} \mu_0^{2^p-1} B_0 \leqslant \left(\frac{1}{2}\right)^p \mu_0^{2^p-1} \mathscr{H},$$

где

$$\mu_0 = 2nA_0B_0C \leqslant 1.$$

Таким образом, при $\mu_0 < 1$ сходимость процесса Ньютона — сверхбыстрая. В частности, при $p = 0$ будем иметь:

$$\| x^* - x^{(0)} \| \leqslant 2B_0 \leqslant \mathscr{H}.$$

## § 5*. Единственность решения

Теорема 3. *При наличии условий* 1) — 4) *теоремы* 1 *из* § 3 *в области*

$$\| x - x^{(0)} \| \leqslant 2B_0 \tag{1}$$

*содержится единственное решение системы* (1) (§ 3).

Доказательство. Пусть, кроме решения $x^*$ системы (1) из § 3, определяемого процессом Ньютона, имеется другое решение $x^{**}$ этой системы такое, что

$$\| x^{**} - x^{(0)} \| \leqslant 2B_0. \tag{2}$$

Последовательные приближения $x^{(p)}$ ($p = 0, 1, 2, \ldots$) процесса Ньютона содержатся в окрестности (1) и удовлетворяют условию

$$f(x^{(p)}) + W_p (x^{(p+1)} - x^{(p)}) = 0,$$

где

$$W_p = W(x^{(p)}).$$

Отсюда, учитывая, что

$$f(x^{**})=0,$$

получаем:

$$W_p(x^{(p+1)}-x^{**})=f(x^{**})-f(x^{(p)})-W_p(x^{**}-x^{(p)})$$

и, следовательно,

$$x^{(p+1)}-x^{**}=\Gamma_p[f(x^{**})-f(x^{(p)})-W_p(x^{**}-x^{(p)})],$$

где

$$\Gamma_p=W_p^{-1}.$$

Производя оценку по норме, будем иметь:

$$\|x^{**}-x^{(p+1)}\|\leqslant\|\Gamma_p\|\,\|f(x^{**})-f(x^{(p)})-W_p(x^{**}-x^{(p)})\|.$$

Согласно обозначениям § 3 (см. теорему 1)

$$\|\Gamma_p\|\leqslant A_p.$$

Применяя лемму 2 из § 2, получим неравенство

$$\|f(x^{**})-f(x^{(p)})-W_p(x^{**}-x^{(p)})\|\leqslant\frac{1}{2}nC\|x^{**}-x^{(p)}\|^2,$$

где постоянная $C$ определена из условия 3) теоремы 1. Поэтому

$$\|x^{**}-x^{(p+1)}\|\leqslant\frac{1}{2}nA_pC\|x^{**}-x^{(p)}\|^2\quad(p=0,\ 1,\ 2,\ \ldots).\tag{3}$$

Полагая $p=0$ в неравенстве (3) и используя неравенство (2), получим:

$$\|x^{**}-x^{(1)}\|\leqslant\frac{1}{2}nA_0C\|x^{**}-x^{(0)}\|^2\leqslant 2nA_0B_0^2C,$$

или, вводя числа, определяемые соотношениями

$$\left.\begin{aligned}\mu_p&=2nA_pB_pC,\\ B_{p+1}&=\frac{1}{2}\mu_pB_p\end{aligned}\right\}\quad(p=0,\ 1,\ 2,\ \ldots),\tag{4}$$

находим

$$\|x^{**}-x^{(1)}\|\leqslant\mu_0B_0=2B_1.\tag{5}$$

Аналогично при $p=1$ из формул (3), (5) и (4) выводим:

$$\|x^{**}-x^{(2)}\|\leqslant\frac{1}{2}nA_1C\|x^{**}-x^{(1)}\|^2\leqslant 2nA_1B_1^2C=\mu_1B_1=2B_2.$$

Вообще,

$$\|x^{**}-x^{(p)}\|\leqslant 2B_p\quad(p=0,\ 1,\ 2,\ \ldots).\tag{6}$$

Так как на основании формулы (2) из § 4 величина $B_p\longrightarrow 0$ при $p\longrightarrow\infty$, то, переходя к пределу в неравенстве (6), будем иметь:

$$x^{**}=\lim_{p\to\infty}x^{(p)}=x^*,$$

т. е. решение системы (1) в области $\|x-x^{(0)}\|\leqslant 2B_0$ единственно.

Замечание. Если область $\overline{U}_{\mathcal{H}}(x^{(0)})$, такова, что

$$\frac{2}{\mu_0} B_0 \leqslant \mathcal{H},$$

то в расширенной области (1)

$$\| x - x^{(0)} \| \leqslant \frac{2}{\mu_0} B_0 \tag{7}$$

не имеется других решений системы (1), кроме $x^*$.

Действительно, предполагая, что в области (7) находится решение $x^{**}$ системы (1) (§ 3) и повторяя рассуждения теоремы, мы получим неравенство вида (3)

$$\| x^{**} - x^{(p+1)} \| \leqslant \frac{1}{2} n A_p C \| x^{**} - x^{(p)} \|^2,$$

где $x^{(p)}$ $(p=0, 1, 2, \ldots)$ — последовательные приближения процесса Ньютона с начальным приближением $x^{(0)}$. Отсюда, так как

$$\| x^{**} - x^{(0)} \| \leqslant \frac{2}{\mu_0} B_0,$$

то, используя числа $\mu_{p+1} = \mu_p^2$, последовательно имеем:

$$\| x^{**} - x^{(1)} \| \leqslant \frac{1}{2} n A_0 C \frac{4}{\mu_0^2} B_0^2 =$$

$$= 2nA_0B_0C \cdot \frac{1}{\mu_0^2} B_0 = \frac{1}{\mu_0} B_0 = \frac{2}{\mu_0^2} B_1 = \frac{2}{\mu_1} B_1,$$

$$\| x^{**} - x^{(2)} \| \leqslant \frac{1}{2} n A_1 C \cdot \frac{4}{\mu_1^2} B_1^2 = 2nA_1B_1C \cdot \frac{1}{2} \mu_1 B_1 \cdot \frac{2}{\mu_1^3} =$$

$$= \mu_1 \cdot B_2 \cdot \frac{2}{\mu_1^3} = \frac{2}{\mu_1^2} B_2 = \frac{2}{\mu_2} B_2$$

и т. д.

Вообще,

$$\| x^{**} - x^{(p)} \| \leqslant \frac{2}{\mu_p} B_p \qquad (p=0, 1, 2, \ldots).$$

Так как

$$B_p = \frac{1}{2} \mu_{p-1} B_{p-1}$$

и

$$\mu_p = \mu_{p-1}^2,$$

то

$$\frac{B_p}{\mu_p} = \frac{1}{2} \cdot \frac{B_{p-1}}{\mu_{p-1}} = \left(\frac{1}{2}\right)^p \cdot \frac{B_0}{\mu_0}. \tag{8}$$

Последнее соотношение можно также непосредственно получить из формул (1) и (2) (§ 4).

Таким образом,

$$\| x^{**} - x^{(p)} \| \leqslant \left(\frac{1}{2}\right)^{p-1} \frac{B_0}{\mu_0} \qquad (p = 0,\ 1,\ 2,\ \ldots).$$

Следовательно,

$$x^{**} = \lim_{p \to \infty} x^{(p)} = x^*,$$

что и требовалось доказать.

## § 6*. Устойчивость сходимости процесса Ньютона при варьировании начального приближения

Теорема 4. *Если выполнены условия* 1)—4) *теоремы* 1 *из* § 3 *и*

$$\frac{2}{\mu_0} B_0 \leqslant \mathcal{H},$$

*где* $\mu_0 = 2nA_0B_0C < 1$, *то процесс Ньютона сходится к единственному решению* $x^*$ *системы* (1) (§ 3) *в основной области* $\| x - x^{(0)} \| \leqslant 2B_0$ *при любом выборе начального приближения* $x'^{(0)}$ *из области*

$$\| x'^{(0)} - x^{(0)} \| \leqslant \frac{1-\mu_0}{2\mu_0} B_0. \qquad (1)$$

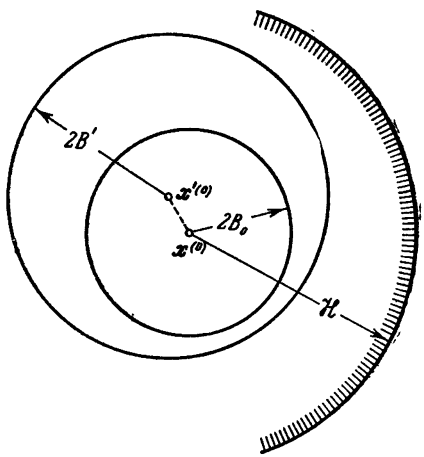

Рис. 58.

Доказательство. По аналогии с введенными выше обозначениями

$$W_0 = W(x^{(0)}) \text{ и } \Gamma_0 = W_0^{-1}$$

введем обозначения

$$W_0' = W(x'^{(0)}) \text{ и } \Gamma_0' = (W_0')^{-1}.$$

Покажем, что в точке $x'^{(0)}$ будут выполнены условия, аналогичные условиям 1)—4) теоремы 1.

Используя обозначения и метод доказательства теоремы 1, имеем:

$$\| E - \Gamma_0 W_0' \| = \| \Gamma_0 (W_0 - W_0') \| \leqslant$$
$$\leqslant \| \Gamma_0 \| \, \| W_0 - W_0' \| \leqslant A_0 nC \, \| x'^{(0)} - x^{(0)} \|.$$

Отсюда, учитывая неравенство (1), получим:

$$\| E - \Gamma_0 W_0' \| \leqslant A_0 nC \frac{1-\mu_0}{2\mu_0} B_0 = \frac{1-\mu_0}{4} \leqslant \frac{1}{4}.$$

Следовательно,

$$\|(\Gamma_0 W_0')^{-1}\| = \|[E-(E-\Gamma_0 W_0')]^{-1}\| \leqslant$$
$$\leqslant \frac{1}{1-\|E-\Gamma_0 W_0'\|} \leqslant \frac{1}{1-\frac{1-\mu_0}{4}} = \frac{4}{3+\mu_0}. \tag{2}$$

Поэтому существует

$$\Gamma_0' = (\Gamma_0 W_0')^{-1}\Gamma_0$$

и

$$\|\Gamma_0'\| \leqslant \|(\Gamma_0 W_0')^{-1}\| \, \|\Gamma_0\| \leqslant \frac{4A_0}{3+\mu_0} = A'. \tag{3}$$

Далее, выводим:

$$\|\Gamma_0 f(x'^{(0)})\| \leqslant \|\Gamma_0\| \, \|f(x'^{(0)}) - f(x^{(0)}) -$$
$$- W_0(x'^{(0)} - x^{(0)})\| + \|\Gamma_0 f(x^{(0)})\| + \|x'^{(0)} - x^{(0)}\| \leqslant$$
$$\leqslant \frac{1}{2} A_0 nC \|x'^{(0)} - x^{(0)}\|^2 + B_0 + \|x'^{(0)} - x^{(0)}\| \leqslant$$
$$\leqslant \frac{1}{4}\mu_0 B_0 \frac{1-2\mu_0+\mu_0^2}{4\mu_0^2} + B_0 + \frac{1-\mu_0}{2\mu_0} B_0 =$$
$$= \frac{1-2\mu_0+\mu_0^2+16\mu_0+8-8\mu_0}{16\mu_0} B_0 = \frac{(3+\mu_0)^2}{16\mu_0} B_0.$$

Отсюда, используя неравенство (2), имеем:

$$\|\Gamma_0' f(x'^{(0)})\| = \|(\Gamma_0 W_0')^{-1} \cdot \Gamma_0 f(x'^{(0)})\| \leqslant \|(\Gamma_0 W_0')^{-1}\| \cdot \|\Gamma_0 f(x'^{(0)})\| \leqslant$$
$$\leqslant \frac{4}{3+\mu_0} \cdot \frac{(3+\mu_0)^2}{16\mu_0} B_0 = \frac{3+\mu_0}{4\mu_0} B_0 = B'. \tag{4}$$

На основании неравенств (3 и (4) получаем:

$$\mu' = 2nA'B'C = 2n\frac{4A_0}{3+\mu_0} \cdot \frac{3+\mu_0}{4\mu_0} B_0 C = 2nA_0 B_0 C \frac{1}{\mu_0} = 1.$$

Кроме того,

$$2B' + \|x'^{(0)} - x^{(0)}\| \leqslant \frac{3+\mu_0}{2\mu_0} B_0 + \frac{1-\mu_0}{2\mu_0} B_0 = \frac{2B_0}{\mu_0} \leqslant \mathcal{H}$$

и, значит, и подавно

$$2B' \leqslant \frac{2B_0}{\mu_0} \leqslant \mathcal{H}.$$

Таким образом, в точке $x'^{(0)}$ полностью выполнены условия теоремы 1, причем

$$\bar{U}_{2B'}(x'^{(0)}) \subset \bar{U}_{\frac{2B_0}{\mu_0}}(x^{(0)}) \subset \bar{U}_{\mathcal{H}}(x^{(0)}) \tag{5}$$

(рис. 58).

Поэтому процесс Ньютона

$$x'^{(p+1)} = x'^{(p)} - \Gamma'_p f(x'^{(p)}),$$

где

$$\Gamma'_p = W^{-1}(x'^{(p)}) \quad (p = 0, 1, 2, \ldots),$$

сходится к некоторому решению $x'^*$ системы (1) § 3, лежащему в области $\overline{U}_{2B'}(x'^{(0)})$. На основании формулы (5)

$$x'^* \in \overline{U}_{\frac{2B_0}{\mu_0}}(x^{(0)}).$$

Но в силу замечания к теореме 3 предыдущего параграфа в области $\overline{U}_{\frac{2B_0}{\mu_0}}(x^{(0)})$ имеется единственное решение $x^*$ основной системы (1). Поэтому

$$x'^* = x^*$$

и

$$x^* = \lim_{p \to \infty} x'^{(p)},$$

что и требовалось доказать.

Замечание. Если $2B_0 < \mathcal{H}$ и $\mu_0 < 1$, то для начального приближения $x^{(0)}$ всегда имеется окрестность, любая точка которой может быть принята за начальное приближение процесса Ньютона, сходящегося к искомому решению $x^*$.

Действительно, пусть

$$2B_0 < 2qB_0 = \mathcal{H},$$

где $q > 1$. Полагая

$$\mu_0^* = \max\left(\mu_0, \frac{1}{q}\right),$$

получим, что в силу теорем 1 и 4 для любого начального приближения $x'^{(0)}$, удовлетворяющего условию

$$\| x'^{(0)} - x^{(0)} \| \leqslant \frac{1-\mu_0^*}{2\mu_0} B_0,$$

соответствующий процесс Ньютона будет сходиться к решению $x^*$ системы (1).

## § 7. Модифицированный метод Ньютона

При построении процесса Ньютона

$$x^{(p+1)} = x^{(p)} - W^{-1}(x^{(p)}) f(x^{(p)}) \quad (p = 0, 1, 2, \ldots) \tag{1}$$

существенным неудобством является необходимость для каждого шага наново вычислять обратную матрицу $W^{-1}(x^{(p)})$. Если матрица $W^{-1}(x)$

непрерывна в окрестности искомого решения $x^*$ и начальное приближение $x^{(0)}$ достаточно близко к $x^*$, то приближенно можно положить:

$$W^{-1}(x^{(p)}) \approx W^{-1}(x^{(0)}),$$

и мы, таким образом, приходим к *модифицированному процессу Ньютона*

$$\xi^{(p+1)} = \xi^{(p)} - W^{-1}(x^{(0)}) f(\xi^{(p)}) \tag{2}$$

($p = 0, 1, 2, \ldots$), где $\xi^{(0)} = x^{(0)}$. Заметим, что для процессов (1) и (2) первые приближения $x^{(1)}$ и $\xi^{(1)}$ совпадают между собой, т. е.

$$x^{(1)} = \xi^{(1)}.$$

Сходимость модифицированного процесса Ньютона (2) исследовалась Л. В. Канторовичем [1].

Теорема. *Если выполнены условия* 1)—4) *теоремы* 1 *из* § 3 *и*

$$\mu_0 = 2nA_0B_0C < 1,$$

*то модифицированный процесс Ньютона* (2), *определяемый начальным приближением* $\xi^{(0)} = x^{(0)}$, *сходится к решению* $x^*$ *системы*

$$f(x) = 0,$$

*причем*

$$\| x^* - \xi^{(p)} \| \leqslant \mu_0^p \| x^* - x^{(0)} \| \leqslant 2B_0\mu_0^p \quad (p = 0, 1, 2, \ldots), \tag{3}$$

*где норма понимается в смысле m-нормы.*

Доказательство. Рассмотрим вектор-функцию

$$F(x) = x - \Gamma_0 f(x) = [F_i(x)],$$

где $\Gamma_0 = W^{-1}(x^{(0)})$.

Очевидно,

$$F(\xi^{(p)}) = \xi^{(p)} - \Gamma_0 f(\xi^{(p)}) = \xi^{(p+1)} \quad (p = 0, 1, 2, \ldots). \tag{4}$$

Кроме того,

$$F'(x) = E - \Gamma_0 f'(x); \tag{5}$$

отсюда, в частности,

$$F'(x^{(0)}) = E - \Gamma_0 f'(x^{(0)}) = E - E = 0. \tag{6}$$

Методом математической индукции докажем, что все приближения $\xi^{(p)}$ ($p = 1, 2, \ldots$) содержатся в $2B_0$-окрестности точки $x^{(0)}$, т. е.

$$\| \xi^{(p)} - x^{(0)} \| < 2B_0. \tag{7}$$

Действительно, при $p = 1$ равенство (7) очевидно, так как в силу условия 2) теоремы имеем:

$$\| \xi^{(1)} - x^{(0)} \| = \| x^{(1)} - x^{(0)} \| \leqslant B_0.$$

Пусть теперь для некоторого $p$ выполнено неравенство (7). Тогда, используя лемму 2 (§ 2), имеем:

$$\begin{aligned}\| \xi^{(p+1)} - x^{(0)} \| &= \| F(\xi^{(p)}) - x^{(0)} \| = \| \xi^{(p)} - \Gamma_0 f(\xi^{(p)}) - x^{(0)} \| = \\ &= \| \Gamma_0 [f(\xi^{(p)}) - W(x^{(0)})(\xi^{(p)} - x^{(0)})] \| \leqslant \| \Gamma_0 f(x^{(0)}) \| + \\ &+ \| \Gamma_0 \{ f(\xi^{(p)}) - f(x^{(0)}) - W(x^{(0)})(\xi^{(p)} - x^{(0)}) \} \| \leqslant \\ &\leqslant B_0 + \frac{1}{2} A_0 n C \| \xi^{(p)} - x^{(0)} \|^2.\end{aligned}$$

Используя неравенство (7), находим:

$$\begin{aligned}\| \xi^{(p+1)} - x^{(0)} \| &< B_0 + \frac{1}{2} n A_0 C \cdot 4 B_0^2 = \\ &= B_0 + 2 n A_0 B_0 C \cdot B_0 = (1 + \mu_0) B_0 < 2 B_0,\end{aligned}$$

что и доказывает наше утверждение.

Так как условия теоремы 1 (§ 3) предполагаются выполненными, то система $f(x) = 0$ имеет корень $x^*$ такой, что $\| x^* - x^{(0)} \| \leqslant 2B_0$.

Рассмотрим разность $x^* - \xi^{(p)}$, где $p \geqslant 1$. Учитывая, что

$$F(x^*) \equiv x^* - \Gamma_0 f(x^*) = x^*$$

и используя лемму 1 (§ 2), имеем:

$$\| x^* - \xi^{(p)} \| = \| F(x^*) - F(\xi^{(p-1)}) \| \leqslant \| x^* - \xi^{(p-1)} \| \cdot \| F'(\theta) \|, \quad (8)$$

где $\theta$ — точка отрезка $[x^*, \xi^{(p-1)}]$.

Далее (см. § 2, лемма 1, следствие 2)

$$\| F'(\theta) \| = \| F'(\theta) - F'(x^{(0)}) \| \leqslant n \| \theta - x^{(0)} \| \max \| F''(\eta) \|, \quad (9)$$

где $\eta$ — точка отрезка $[\theta, x^{(0)}]$. Из формулы (5) имеем:

$$F'(x) = \left[ \delta_{ij} - \sum_{s=1}^{n} \gamma_{is} \frac{\partial f_s}{\partial x_j} \right],$$

где $\delta_{ij}$ — символ Кронекера и $\Gamma_0 = [\gamma_{ij}]$. Поэтому

$$\frac{\partial F_i}{\partial x_j} = \delta_{ij} - \sum_{s=1}^{n} \gamma_{is} \frac{\partial f_s}{\partial x_j}$$

и

$$\frac{\partial^2 F_i}{\partial x_j \partial x_k} = - \sum_{s=1}^{n} \gamma_{is} \frac{\partial^2 f_s}{\partial x_j \partial x_k}.$$

Следовательно,

$$\begin{aligned}\| F''(\eta) \| &= \max_{i,j} \sum_{k=1}^{n} \left| \frac{\partial^2 F_i(\eta)}{\partial x_j \partial x_k} \right| = \max_{i,j} \sum_{k=1}^{n} \left| \sum_{s=1}^{n} \gamma_{is} \frac{\partial^2 f_s(\eta)}{\partial x_j \partial x_k} \right| \leqslant \\ &\leqslant \max_{i,j} \sum_{s=1}^{n} | \gamma_{is} | \sum_{k=1}^{n} \left| \frac{\partial^2 f_s(\eta)}{\partial x_j \partial x_k} \right| \leqslant \max_{i,j} \sum_{s=1}^{n} | \gamma_{is} | C = C \| \Gamma_0 \| \leqslant A_0 C\end{aligned}$$

и, значит, на основании (9),

$$\| F'(\theta) \| \leqslant nA_0C \| \theta - x^{(0)} \|.$$

Так как точка $\theta$, очевидно, принадлежит $2B_0$-окрестности точки $x^{(0)}$, то

$$\| \theta - x^{(0)} \| \leqslant 2B_0$$

и, таким образом,

$$\| F'(\theta) \| \leqslant 2nA_0B_0C = \mu_0. \tag{10}$$

Учитывая неравенство (10), из неравенства (8) выводим:

$$\| x^* - \xi^{(p)} \| \leqslant \mu_0 \| x^* - \xi^{(p-1)} \|,$$

откуда

$$\| x^* - \xi^{(p)} \| \leqslant \mu_0^p \| x^* - \xi^{(0)} \| = \mu_0^p \| x^* - x^{(0)} \| \leqslant 2B\mu_0^p.$$

При $\mu_0 < 1$ из последнего неравенства вытекает, что

$$\lim_{p \to \infty} \xi^{(p)} = x^*.$$

Теорема доказана полностью.

## § 8. Метод итерации

Пусть дана система нелинейных уравнений специального вида

$$\left.\begin{aligned} x_1 &= \varphi_1(x_1, x_2, \ldots, x_n), \\ x_2 &= \varphi_2(x_1, x_2, \ldots, x_n), \\ &\ldots\ldots\ldots\ldots \\ x_n &= \varphi_n(x_1, x_2, \ldots, x_n), \end{aligned}\right\} \tag{1}$$

где функции $\varphi_1, \varphi_2, \ldots, \varphi_n$ действительны, определены и непрерывны в некоторой окрестности $\omega$ изолированного решения $(x_1^*, x_2^*, \ldots, x_n^*)$ этой системы.

Введя в рассмотрение векторы

$$x = (x_1, x_2, \ldots, x_n) \quad \text{и} \quad \varphi(x) = (\varphi_1(x), \varphi_2(x), \ldots, \varphi_n(x)),$$

систему (1) можно записать более кратко:

$$x = \varphi(x). \tag{2}$$

Для нахождения вектор-корня $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$ уравнения (2) часто удобно использовать *метод итерации*

$$x^{(p+1)} = \varphi(x^{(p)}) \qquad (p = 0, 1, 2, \ldots), \tag{3}$$

где начальное приближение $x^{(0)} \approx x^*$. Сходимость этого процесса будет исследована ниже. Заметим, что если процесс итерации (3) сходится, то предельное значение

$$\xi = \lim_{p \to \infty} x^{(p)} \tag{4}$$

обязательно является корнем уравнения (2). Действительно, предполагая, что соотношение (4) выполнено, и переходя к пределу в равенстве (3) при $p \to \infty$, в силу непрерывности функции $\varphi(x)$ будем иметь:

$$\lim_{p \to \infty} x^{(p+1)} = \varphi\left(\lim_{p \to \infty} x^{(p)}\right),$$

т. е.

$$\xi = \varphi(\xi).$$

Таким образом, $\xi$ есть корень векторного уравнения (2).

Если, сверх того, все приближения $x^{(p)}$ $(p = 0, 1, 2, \ldots)$ принадлежат области $\omega$ и $x^*$ — единственный корень системы (2) в $\omega$, то, очевидно,

$$\xi = x^*.$$

Метод итерации может быть применен также к общей системе

$$f(x) = 0, \tag{5}$$

где $f(x)$ — вектор-функция, определенная и непрерывная в окрестности $\omega$ изолированного вектор-корня $x^*$. Например, перепишем эту систему в следующем виде:

$$x = x + \Lambda f(x),$$

где $\Lambda$ — неособенная матрица. Введя обозначение

$$x + \Lambda f(x) = \varphi(x), \tag{6}$$

будем иметь:

$$x = \varphi(x). \tag{7}$$

К последнему уравнению легко применяется обычный метод итерации (3).

Если функция $f(x)$ имеет непрерывную производную $f'(x)$ в $\omega$, то из формулы (6) вытекает:

$$\varphi'(x) = E + \Lambda f'(x).$$

В следующих параграфах будет доказано, что процесс итерации для уравнения (7) быстро сходится, если $\varphi'(x)$ мала по норме. Учитывая это обстоятельство, выбираем матрицу $\Lambda$ так, чтобы

$$\varphi'(x^{(0)}) = E + \Lambda f'(x^{(0)}) = 0;$$

отсюда, если матрица $f'(x^{(0)})$ — неособенная, будем иметь:

$$\Lambda = -[f'(x^{(0)})]^{-1}.$$

Заметим, что это, в сущности, модифицированный процесс Ньютона, примененный к уравнению (5) (см. § 7).

В случае, если $\det f'(x^{(0)}) = 0$, то следует выбрать другое начальное приближение $x^{(0)}$.

Употребляются также иные способы замены системы (5) эквивалентной ей системой (7).

Пример. Методом итерации приближенно решить систему

$$\left.\begin{aligned} x_1^2+x_2^2&=1,\\ x_1^3-x_2&=0. \end{aligned}\right\} \tag{8}$$

Решение. Из графического построения видно (рис. 59), что система (8) имеет два решения, отличающихся только знаком. Ограничимся нахождением положительного решения. Из чертежа усматриваем, что за начальное приближение положительного решения системы (8) можно принять:

$$x^{(0)}=\begin{bmatrix}0,9\\0,5\end{bmatrix}.$$

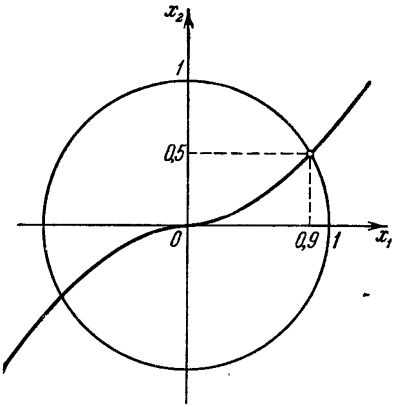


Рис. 59

Полагая

$$f(x)=\begin{bmatrix}x_1^2+x_2^2-1\\x_1^3-x_2\end{bmatrix},$$

будем иметь:

$$f'(x)=\begin{bmatrix}2x_1 & 2x_2\\3x_1^2 & -1\end{bmatrix}.$$

Отсюда

$$f'(x^{(0)})=\begin{bmatrix}1,8 & 1\\2,43 & -1\end{bmatrix}$$

и

$$\det f'(x^{(0)})=-1,8-2,43=-4,23.$$

Так как матрица $f'(x^{(0)})$ — неособенная, то существует обратная матрица

$$[f'(x^{(0)})]^{-1}=-\frac{1}{4,23}\begin{bmatrix}-1 & -1\\-2,43 & 1,8\end{bmatrix}.$$

Таким образом,

$$\Lambda=-[f'(x^{(0)})]^{-1}=\frac{1}{4,23}\begin{bmatrix}-1 & -1\\-2,43 & 1,8\end{bmatrix}.$$

Положим

$$\varphi(x)=x+\Lambda f(x)=\begin{bmatrix}x_1\\x_2\end{bmatrix}-\frac{1}{4,23}\begin{bmatrix}1 & 1\\2,43 & -1,8\end{bmatrix}\begin{bmatrix}x_1^2+x_2^2-1\\x_1^3-x_2\end{bmatrix}.$$

Тогда система (8) будет эквивалентна стандартному матричному уравнению

$$x=\varphi(x). \tag{9}$$

Пользуясь формулой (4), находим для решения системы (9) последовательные приближения:

$$x^{(1)}=\begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \end{bmatrix}-\frac{1}{4,23}\begin{bmatrix} 1 & 1 \\ 2,43 & 1,8 \end{bmatrix}\begin{bmatrix} x_1^{(0)^2}+x_2^{(0)^2}-1 \\ x_1^{(0)^3}-x_2^{(0)} \end{bmatrix}=$$

$$=\begin{bmatrix} 0,9 \\ 0,5 \end{bmatrix}-\frac{1}{4,23}\begin{bmatrix} 1 & 1 \\ 2,43 & -1,8 \end{bmatrix}\begin{bmatrix} 0,060 \\ 0,229 \end{bmatrix}=\begin{bmatrix} 0,9 \\ 0,5 \end{bmatrix}-\begin{bmatrix} 0,0683 \\ -0,0630 \end{bmatrix}=\begin{bmatrix} 0,8317 \\ 0,5630 \end{bmatrix};$$

$$x^{(2)}=\begin{bmatrix} 0,8317 \\ 0,5630 \end{bmatrix}-\frac{1}{4,23}\begin{bmatrix} 1 & 1 \\ 2,43 & -1,8 \end{bmatrix}\begin{bmatrix} 0,8317^2+0,5630^2-1 \\ 0,8317^3-0,5630 \end{bmatrix}=$$

$$=\begin{bmatrix} 0,8317 \\ 0,5630 \end{bmatrix}-\begin{bmatrix} -0,0049 \\ 0,0003 \end{bmatrix}=\begin{bmatrix} 0,8268 \\ 0,5633 \end{bmatrix};$$

$$x^{(3)}=\begin{bmatrix} 0,8268 \\ 0,5633 \end{bmatrix}-\begin{bmatrix} 0,0007 \\ 0,0002 \end{bmatrix}=\begin{bmatrix} 0,8261 \\ 0,5631 \end{bmatrix};$$

$$x^{(4)}=\begin{bmatrix} 0,8261 \\ 0,5631 \end{bmatrix}-\begin{bmatrix} 0,0000 \\ -0,0005 \end{bmatrix}=\begin{bmatrix} 0,8261 \\ 0,5636 \end{bmatrix}$$

и т. д.

Ограничиваясь четвертым приближением, имеем корни

$$x_1=0,8261;\quad x_2=0,5636,$$

причем

$$f(x)=\begin{bmatrix} 0,0000 \\ 0,0002 \end{bmatrix}$$

с точностью до $10^{-4}$.

## § 9*. Понятие о сжимающем отображении

Пусть дана нелинейная система

$$\left.\begin{aligned} y_1&=\varphi_1(x_1, x_2, \ldots, x_n), \\ y_2&=\varphi_2(x_1, x_2, \ldots, x_n), \\ &\ldots\ldots\ldots\ldots \\ y_n&=\varphi_n(x_1, x_2, \ldots, x_n), \end{aligned}\right\} \tag{1}$$

где функции $\varphi_1, \varphi_2, \ldots, \varphi_n$ определены и непрерывны в известной области $G$ действительного $n$-мерного пространства $E_n$ (гл. X, § 1), причем значения их $(y_1, y_2, \ldots, y_n)$ при $(x_1, x_2, \ldots, x_n) \in G$ заполняют некоторую область $G' \subset E_n$. Система (1) устанавливает *отображение* области $G$ на область $G'$. Введя векторы

$$x=\begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix},\quad y=\begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix},\quad \varphi=\begin{bmatrix} \varphi_1 \\ \cdot \\ \cdot \\ \cdot \\ \varphi_n \end{bmatrix},$$

систему (1) коротко можно записать следующим образом:

$$y=\varphi(x). \tag{1'}$$

В пространстве $E_n$ введем каноническую норму $\|x\|$, удовлетворяющую обычным условиям. Например, можно положить:

$$\|x\|_m = \max_i |x_i|,$$

или

$$\|x\|_l = \sum_i |x_i|,$$

или

$$\|x\|_k = \sqrt{\sum_i x_i^2}$$

и т. п.

Отображение (1) или (1′) называется *сжимающим* в области $G$, если существует правильная дробь $q$ такая, что для любых двух точек $x_1, x_2 \in G$ их образы $y_1 = \varphi(x_1)$ и $y_2 = \varphi(x_2)$ удовлетворяют условию

$$\|y_1 - y_2\| \leqslant q \|x_1 - x_2\|, \tag{2}$$

т. е.

$$\|\varphi(x_1) - \varphi(x_2)\| \leqslant q \|x_1 - x_2\| \qquad (0 \leqslant q < 1). \tag{2′}$$

Рассмотрим нелинейное векторное уравнение

$$x = \varphi(x), \tag{3}$$

эквивалентное нелинейной системе уравнений специального вида

$$\left.\begin{aligned} x_1 &= \varphi_1(x_1, x_2, \ldots, x_n), \\ x_2 &= \varphi_2(x_1, x_2, \ldots, x_n), \\ &\ldots\ldots\ldots\ldots \\ x_n &= \varphi_n(x_1, x_2, \ldots, x_n). \end{aligned}\right\} \tag{3′}$$

Решение $x^*$ этого уравнения, если оно существует, является *неподвижной точкой* преобразования (1). Для нахождения $x^*$ построим *итерационный процесс*

$$x^{(p)} = \varphi(x^{(p-1)}) \qquad (p = 1, 2, \ldots), \tag{4}$$

где $x^{(0)} \in G$.

Теорема 1. *Пусть область $G$ замкнута и отображение* (1) *является сжимающим в $G$, т. е. выполнено условие* (2). *Тогда, если для итерационного процесса* (4) *все последовательные приближения $x^{(p)} \in G$ ($p = 0, 1, 2, \ldots$), то* 1) *независимо от выбора начального приближения $x^{(0)}$ процесс* (4) *сходится, т. е. существует*

$$x^* = \lim_{p \to \infty} x^{(p)}; \tag{5}$$

2) *предельный вектор $x^*$ является единственным решением уравнения* (3) *в области $G$;* 3) *справедлива оценка*

$$\|x^* - x^{(p)}\| \leqslant \frac{q^p}{1-q} \|x^{(1)} - x^{(0)}\|. \tag{6}$$

Доказательство. 1) Для доказательства сходимости последовательности приближений $x^{(p)}$ $(p=0, 1, 2, \ldots)$ применим *критерий Коши* (см. гл. VII, § 9). Имеем:

$$\| x^{(p+k)} - x^{(p)} \| = \| (x^{(p+1)} - x^{(p)}) + (x^{(p+2)} - x^{(p+1)}) + \ldots + (x^{(p+k)} - x^{(p+k-1)}) \| \leqslant \| x^{(p+1)} - x^{(p)} \| + \| x^{(p+2)} - x^{(p+1)} \| + \ldots + \| x^{(p+k)} - x^{(p+k-1)} \|. \quad (7)$$

Используя соотношение (4) и «условие сжатости» (2′), последовательно получаем:

$$\| x^{(s+1)} - x^{(s)} \| = \| \varphi(x^{(s)}) - \varphi(x^{(s-1)}) \| \leqslant q \| x^{(s)} - x^{(s-1)} \| \leqslant q^2 \| x^{(s-1)} - x^{(s-2)} \| \leqslant q^s \| x^{(1)} - x^{(0)} \|, \quad (8)$$

где $s \geqslant 0$. Поэтому, усиливая правую часть неравенства (7), будем иметь:

$$\| x^{(p+k)} - x^{p} \| \leqslant q^p \| x^{(1)} - x^{(0)} \| + q^{p+1} \| x^{(1)} - x^{(0)} \| + \ldots + q^{p+k-1} \| x^{(1)} - x^{(0)} \|,$$

или, воспользовавшись формулой для суммы членов геометрической прогрессии, находим:

$$\| x^{(p+k)} - x^{(p)} \| \leqslant \frac{q^p - q^{p+k}}{1-q} \| x^{(1)} - x^{(0)} \| \leqslant \frac{q^p}{1-q} \| x^{(1)} - x^{(0)} \|. \quad (9)$$

Так как $0 \leqslant q < 1$ и, следовательно, $q^p \to 0$ при $p \to \infty$, то из формулы (9) вытекает, что для всякого $\varepsilon > 0$ существует $N = N(\varepsilon)$ такое, что при $p > N(\varepsilon)$ и $k > 0$ будет справедливо неравенство

$$\| x^{(p+k)} - x^{(p)} \| < \varepsilon,$$

т. е. для последовательности $x^{(p)}$ $(p=0, 1, 2, \ldots)$ выполнен критерий Коши. Поэтому существует:

$$x^* = \lim_{p \to \infty} x^{(p)},$$

причем $x^* \in G$ в силу замкнутости области $G$.

2) Вектор $x^*$ является решением уравнения (3), так как, переходя к пределу при $p \to \infty$ в равенстве (4) и учитывая непрерывность в $G$ вектор-функции $\varphi(x)$, будем иметь:

$$\lim_{p \to \infty} x^{(p)} = \varphi(\lim_{p \to \infty} x^{(p-1)}),$$

т. е.

$$x^* = \varphi(x^*). \quad (10)$$

Это решение единственно в $G$. Действительно, пусть $x^{*\prime}$ есть другое решение уравнения (3), т. е.

$$x^{*\prime} = \varphi(x^{*\prime}). \quad (11)$$

Вычитая из равенства (10) равенство (11), получим:

$$x^* - x^{*\prime} = \varphi(x^*) - \varphi(x^{*\prime}),$$

отсюда

$$\| x^* - x^{*\prime} \| = \| \varphi(x^*) - \varphi(x^{*\prime}) \| \leqslant q \| x^* - x^{*\prime} \|$$

или

$$(1-q) \| x^* - x^{*\prime} \| \leqslant 0. \tag{12}$$

Так как $1-q > 0$, то неравенство (12) может иметь место лишь при $\| x^* - x^{*\prime} \| = 0$, т. е. тогда, когда $x^* = x^{*\prime}$. Таким образом, другого решения уравнения (3) в области $G$ быть не может.

3) Переходя к пределу при $k \to \infty$ в неравенстве (9), получим оценку (6).

Теорема 1 доказана полностью.

Замечание 1. Если область $G$ совпадает со всем пространством $E_n$, то условие $x^{(p)} \in G$ $(p = 0, 1, 2, \ldots)$, очевидно, становится излишним.

Замечание 2. Используя неравенства

$$\| x^{(p+1)} - x^{(p)} \| \leqslant q \| x^{(p)} - x^{(p-1)} \|,$$
$$\| x^{(p+2)} - x^{(p+1)} \| \leqslant q^2 \| x^{(p)} - x^{(p-1)} \|,$$
$$\cdots\cdots\cdots\cdots\cdots\cdots$$

из формулы (7) будем иметь:

$$\| x^{(p+k)} - x^{(p)} \| \leqslant q \| x^{(p)} - x^{(p-1)} \| +$$
$$+ q^2 \| x^{(p)} - x^{(p-1)} \| + \ldots + q^k \| x^{(p)} - x^{(p-1)} \| \leqslant \frac{q}{1-q} \| x^{(p)} - x^{(p-1)} \|.$$

Отсюда при $k \to \infty$ получим:

$$\| x^* - x^{(p)} \| \leqslant \frac{q}{1-q} \| x^{(p)} - x^{(p-1)} \|. \tag{13}$$

В частности, если $0 \leqslant q \leqslant \frac{1}{2}$, то из формулы (13) следует, что при

$$\| x^{(p)} - x^{(p-1)} \| \leqslant \varepsilon$$

справедливо неравенство

$$\| x^* - x^{(p)} \| \leqslant \varepsilon.$$

В условии теоремы 1 требуется, чтобы все приближения $x^{(p)}$ принадлежали фиксированной области $G$. На практике это обстоятельство иногда затруднительно проверить. Поэтому мы приведем несколько видоизмененную формулировку теоремы 1.

Теорема 2. *Пусть отображение* (1) *является сжимающим в замкнутой области* $G^1$ *и* $g$ — *ограниченная область, лежащая в* $G$ *вместе со своей* $\rho$-*окрестностью* (в смысле введенной нормы), *где*

$$\rho \geqslant \frac{Dq}{1-q}, \tag{14}$$

*$D$ — диаметр области $g$ и $q$ — соответствующий коэффициент в неравенстве* (2). *Тогда, если*

$$x^{(0)} \in g \quad \text{и} \quad x^{(1)} = \varphi(x^{(0)}) \in g,$$

*то итерационный процесс* (4) *сходится и справедливы заключения теоремы* 1.

Доказательство. Очевидно, достаточно доказать, что все последовательные приближения содержатся в области $G$:

$$x^{(p)} \in G \qquad (p = 2, 3, \ldots). \tag{15}$$

Доказательство будем проводить методом математической индукции.

Прежде всего, по условию теоремы $x^{(0)} \in G$ и $x^{(1)} \in G$. Пусть теперь $x^{(s)} \in G$ $(s = 0, 1, \ldots, p)$. Так как

$$\| x^{(1)} - x^{(0)} \| \leqslant D,$$

то на основании неравенства (8) имеем:

$$\| x^{(s+1)} - x^{(s)} \| \leqslant q^s \| x^{(1)} - x^{(0)} \| \leqslant Dq^s \qquad (s = 0, 1, 2, \ldots, p).$$

Поэтому

$$\| x^{(p+1)} - x^{(1)} \| \leqslant \| x^{(2)} - x^{(1)} \| + \ldots + \| x^{(p+1)} - x^{(p)} \| \leqslant$$
$$\leqslant Dq + \ldots + Dq^p \leqslant \frac{Dq}{1-q} \leqslant \rho.$$

Следовательно, $x^{(p+1)} \in G$.

Таким образом, свойство (15) установлено и мы находимся в условиях теоремы (1). Теорема 2 доказана.

Замечание. Условия теоремы 2 равносильны следующим: пусть $x^{(0)} \in G$ и $x^{(1)} \in G$, причем $\| x^{(1)} - x^{(0)} \| \leqslant D$. Если расстояние точки $x^{(1)}$ до границы $\Gamma$ области $G$ не больше $\frac{Dq}{1-q}$, то итерационный процесс (4) сходится.

## § 10*. Первое достаточное условие сходимости процесса итерации

Рассмотрим действительную приведенную систему

$$\left.\begin{aligned} x_1 &= \varphi_1(x_1, x_2, \ldots, x_n), \\ x_2 &= \varphi_2(x_1, x_2, \ldots, x_n), \\ &\cdots\cdots\cdots\cdots \\ x_n &= \varphi_n(x_1, x_2, \ldots, x_n), \end{aligned}\right\} \tag{1}$$

или в матричном виде

$$x = \varphi(x), \tag{1'}$$

где

$$x = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}, \quad \varphi(x) = \begin{bmatrix} \varphi_1(x) \\ \cdot \\ \cdot \\ \cdot \\ \varphi_n(x) \end{bmatrix}.$$

Предполагается, что вектор-функция $\varphi(x)$ определена и непрерывна вместе со своей производной $\varphi'(x)=\left[\frac{\partial\varphi_i}{\partial x_j}\right]$ в выпуклой ограниченной замкнутой области $G\subset E_n$.

В этом параграфе мы будем пользоваться двумя нормами:

$$\|x\|_m=\max_i|x_i|$$

и

$$\|x\|_l=\sum_{i=1}^{n}|x_i|.$$

Относительно области $G$ введем нормы:

$$\|\varphi'(x)\|_{\mathrm{I}}=\max_{x\in G}\|\varphi'(x)\|_m \tag{2}$$

и

$$\|\varphi'(x)\|_{\mathrm{II}}=\max_{x\in G}\|\varphi'(x)\|_l, \tag{3}$$

где

$$\|\varphi'(x)\|_m=\max_i\sum_{j=1}^{n}\left|\frac{\partial\varphi_i(x)}{\partial x_j}\right| \tag{2'}$$

и

$$\|\varphi'(x)\|_l=\max_j\sum_{i=1}^{n}\left|\frac{\partial\varphi_i(x)}{\partial x_j}\right|. \tag{3'}$$

Теорема. *Пусть функции $\varphi(x)$ и $\varphi'(x)$ непрерывны в области $G$, причем в $G$ выполнено неравенство*

$$\|\varphi'(x)\|_{\mathrm{I}}\leqslant q<1, \tag{4}$$

*где $q$ — некоторая постоянная.*

*Если последовательные приближения*

$$x^{(p+1)}=\varphi(x^{(p)}) \tag{5}$$

*($p=0, 1, 2, \ldots$) не выходят из области $G$, то процесс итерации* (5) *сходится и предельный вектор*

$$x^*=\lim_{p\to\infty}x^{(p)}$$

*является в области $G$ единственным решением системы* (1).

Доказательство. В силу теоремы 1 предыдущего параграфа достаточно показать, что отображение

$$y=\varphi(x) \tag{6}$$

при наличии условия (4) является сжимающим в области $G$ в смысле $m$-нормы.

Пусть $x_1, x_2 \in G$ и $y_i = \varphi(x_i)$ $(i = 1, 2)$. На основании следствия 1 к лемме 1 из § 2 имеем:

$$\| y_1 - y_2 \|_m = \| \varphi(x_1) - \varphi(x_2) \|_m \leqslant$$
$$\leqslant \| x_1 - x_2 \|_m \| \varphi'(\xi) \|_m \leqslant \| x_1 - x_2 \|_m \| \varphi'(x) \|_1.$$

Отсюда

$$\| y_1 - y_2 \|_m \leqslant q \| x_1 - x_2 \|_m,$$

где $0 \leqslant q < 1$, что и требовалось доказать.

Следствие. Процесс итерации (5) сходится, если

$$\sum_{j=1}^{n} \left| \frac{\partial \varphi_i(x)}{\partial x_j} \right| \leqslant q_i < 1 \qquad (i = 1, 2, \ldots, n) \tag{7}$$

при $x \in G$.

Очевидно, из системы неравенств (7) вытекает условие (4) теоремы.

Замечание. На основании теоремы 1 из § 9 для приближения $x^{(p)}$ получаем следующую оценку:

$$\| x^* - x^{(p)} \|_m \leqslant \frac{q^p}{1-q} \| x^{(1)} - x^{(0)} \|_m \qquad (p = 0, 1, 2, \ldots),$$

где $x^{(1)} = \varphi(x^{(0)})$.

## § 11*. Второе достаточное условие сходимости процесса итерации

Прежде чем переходить к доказательству теоремы сходимости, использующей нормы $\|\varphi'(x)\|_{11}$, выведем предварительно одну оценку для разности значений вектор-функции, аналогичную теореме о среднем и представляющую также самостоятельный интерес.

Лемма. *Если вектор-функция*

$$f(x) = \begin{bmatrix} f_1(x) \\ \cdot \\ \cdot \\ \cdot \\ f_n(x) \end{bmatrix}$$

*непрерывна, вместе со своей производной $f'(x)$, в выпуклой области, содержащей точки $x$ и $x + \Delta x$, то*

$$\| f(x + \Delta x) - f(x) \|_l \leqslant \| \Delta x \|_l \cdot \| f'(\xi) \|_l, \tag{1}$$

*где $\xi = x + \theta \Delta x$ и $0 < \theta < 1$.*

Доказательство. Рассмотрим вспомогательную функцию

$$\Phi(t) = \sum_{i=1}^{n} \varepsilon_i [f_i(x + t\Delta x) - f_i(x)],$$

где $0 \leqslant t \leqslant 1$ — скалярный аргумент и $\varepsilon_i$ — система чисел, принимающих значения $-1, 0, 1$. Очевидно, $\Phi(0) = 0$. Применяя теорему

Лагранжа о конечном приращении функции, получим:

$$\sum_{i=1}^{n} \varepsilon_i [f_i(\boldsymbol{x}+\Delta\boldsymbol{x}) - f_i(\boldsymbol{x})] = \Phi(1) - \Phi(0) = \Phi'(\theta) = \\ = \sum_{i=1}^{n} \varepsilon_i \sum_{j=1}^{n} \frac{\partial f_i(\xi)}{\partial x_j} \Delta x_j,$$

где $\xi = \boldsymbol{x} + \theta\,\Delta\boldsymbol{x}$ и $0 < \theta < 1$.

Отсюда, учитывая, что $|\varepsilon_i| \leqslant 1$, будем иметь:

$$\sum_{i=1}^{n} \varepsilon_i [f_i(\boldsymbol{x}+\Delta\boldsymbol{x}) - f_i(\boldsymbol{x})] \leqslant \\ \leqslant \sum_{i=1}^{n} \sum_{j=1}^{n} \left| \frac{\partial f_i(\xi)}{\partial x_j} \right| \cdot |\Delta x_j| = \sum_{j=1}^{n} |\Delta x_j| \sum_{i=1}^{n} \left| \frac{\partial f_i(\xi)}{\partial x_j} \right|. \qquad (2)$$

Так как

$$\sum_{i=1}^{n} \left| \frac{\partial f_i(\xi)}{\partial x_j} \right| \leqslant \max_j \sum_{i=1}^{n} \left| \frac{\partial f_i(\xi)}{\partial x_j} \right| = \|\boldsymbol{f}'(\xi)\|_l,$$

то, усиливая неравенство (2), получаем:

$$\sum_{i=1}^{n} \varepsilon_i [f_i(\boldsymbol{x}+\Delta\boldsymbol{x}) - f_i(\boldsymbol{x})] \leqslant \sum_{j=1}^{n} |\Delta x_j| \, \|\boldsymbol{f}'(\xi)\|_l = \\ = \|\boldsymbol{f}'(\xi)\|_l \cdot \sum_{j=1}^{n} |\Delta x_j| = \|\boldsymbol{f}'(\xi)\|_l \cdot \|\Delta\boldsymbol{x}\|_l.$$

Полагая в последнем неравенстве

$$\varepsilon_i = \operatorname{sgn}[f_i(\boldsymbol{x}+\Delta\boldsymbol{x}) - f_i(\boldsymbol{x})] \qquad (i = 1, 2, \ldots, n),$$

окончательно находим:

$$\sum_{i=1}^{n} |f_i(\boldsymbol{x}+\Delta\boldsymbol{x}) - f_i(\boldsymbol{x})| \leqslant \|\boldsymbol{f}'(\xi)\|_l \, \|\Delta\boldsymbol{x}\|_l,$$

т. е.

$$\|\boldsymbol{f}(\boldsymbol{x}+\Delta\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x})\|_l \leqslant \|\Delta\boldsymbol{x}\|_l \, \|\boldsymbol{f}'(\xi)\|_l, \qquad (2')$$

что и требовалось доказать *).

---

*) Если непосредственно применить теорему о среднем к каждой компоненте вектора $\boldsymbol{f}(\boldsymbol{x}+\Delta\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x})$, то получается оценка, зависящая от значений производных $\frac{\partial f_i(\xi_i)}{\partial x_j}$ в различных точках $\xi_i$ $(i = 1, 2, \ldots, n)$ интервала $(\boldsymbol{x}, \boldsymbol{x}+\Delta\boldsymbol{x})$. Неравенство (2′) показывает, что можно ограничиться значениями производных $\frac{\partial f_i(\xi)}{\partial x_j}$ в одной и той же точке $\xi \in (\boldsymbol{x}, \boldsymbol{x}+\Delta\boldsymbol{x})$.

Теорема. *Пусть вектор-функция* $\varphi(x)$ *непрерывна, вместе со своей производной* $\varphi'(x)$, *в ограниченной выпуклой замкнутой области* $G$ *и*

$$\|\varphi'(x)\|_{II} \leqslant q < 1, \tag{3}$$

*где* $q$ — *постоянная. Если* $x^{(0)} \in G$ *и все последовательные приближения*

$$x^{(p+1)} = \varphi(x^{(p)}) \qquad (p = 0,\ 1,\ 2,\ \ldots) \tag{4}$$

*также содержатся в* $G$, *то процесс итерации* (4) *сходится к единственному решению уравнения*

$$x = \varphi(x) \tag{5}$$

*в области* $G$.

Доказательство. Докажем, что отображение $y = \varphi(x)$ является сжимающим в $G$ в смысле $l$-нормы.

Пусть $x_1,\ x_2 \in G$ и $y_i = \varphi(x_i)$ $(i = 1,\ 2)$. Используя лемму, имеем:

$$\|y_1 - y_2\|_l = \|\varphi(x_1) - \varphi(x_2)\|_l \leqslant \|x_1 - x_2\|_l \cdot \|\varphi'(\xi)\|_l, \tag{6}$$

где $\xi \in G$.

Так как

$$\|\varphi'(\xi)\|_l \leqslant \max_{x \in G} \|\varphi'(x)\|_l = \|\varphi'(x)\|_{II} \leqslant q,$$

то из неравенства (6) получим:

$$\|y_1 - y_2\|_l \leqslant q\,\|x_1 - x_2\|_l,$$

где $0 \leqslant q < 1$.

В силу теоремы из § 10 теорема является доказанной.

Следствие. Процесс итерации (4) сходится к единственному решению уравнения (5), если при $x \in G$ выполнены неравенства

$$\sum_{j=1}^{n} \left| \frac{\partial \varphi_j(x)}{\partial x_i} \right| \leqslant q_i < 1 \tag{7}$$

$(i = 1,\ 2,\ \ldots,\ n)$.

Замечание. На основании теоремы из § 10 для приближения $x^{(p)}$ имеем следующую оценку:

$$\|x^* - x^{(p)}\|_l \leqslant \frac{q^p}{1-q} \|x^{(1)} - x^{(0)}\|_l,$$

где $x^{(1)} = \varphi(x^{(0)})$.

## § 12*. Метод скорейшего спуска (метод градиента)

Пусть имеем систему уравнений

$$\left.\begin{aligned} f_1(x_1,\ x_2,\ \ldots,\ x_n) &= 0, \\ f_2(x_1,\ x_2,\ \ldots,\ x_n) &= 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ f_n(x_1,\ x_2,\ \ldots,\ x_n) &= 0 \end{aligned}\right\} \tag{1}$$

или в матричной форме

$$f(x)=0, \tag{2}$$

где $f=\begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix}$.

Предположим, что функции $f_i$ действительны и непрерывно дифференцируемы в их общей области определения. Рассмотрим функцию

$$U(x)=\sum_{i=1}^{n}[f_i(x)]^2=(f(x),\ f(x)). \tag{3}$$

Очевидно, что каждое решение системы (1) обращает в нуль функцию $U(x)$; наоборот, числа $x_1, x_2, \ldots, x_n$, для которых функция $U(x)$ равна нулю, являются корнями системы (1).

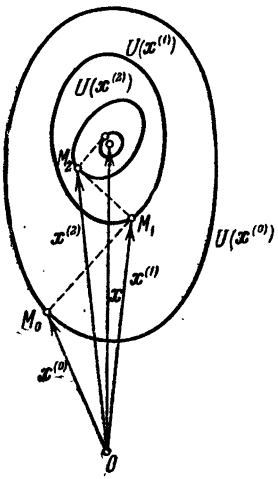


Рис. 60.

Будем предполагать, что система (1) имеет лишь изолированное решение, которое представляет собой точку строгого минимума функции $U(x)$. Таким образом, задача сводится к нахождению минимума функции $U(x)$ в $n$-мерном пространстве $E_n=\{x_1, x_2, \ldots, x_n\}$.

Пусть $x$ — вектор-корень системы (1) и $x^{(0)}$ — его нулевое приближение. Через точку $x^{(0)}$ проведем поверхность уровня функции $U(x)$. Если точка $x^{(0)}$ достаточно близка к корню $x$, то при наших предположениях поверхность уровня

$$U(x)=U(x^{(0)})$$

будет похожа на эллипсоид.

Из точки $x^{(0)}$ двигаемся по нормали к поверхности $U(x)=U(x^{(0)})$ до тех пор, пока эта нормаль не коснется в некоторой точке $x^{(1)}$ какой-то другой поверхности уровня (рис. 60)

$$U(x)=U(x^{(1)}).$$

Затем, отправляясь от точки $x^{(1)}$, снова двигаемся по нормали к поверхности уровня $U(x)=U(x^{(1)})$ до тех пор, пока эта нормаль не коснется в некоторой точке $x^{(2)}$ новой поверхности уровня $U(x)=U(x^{(2)})$, и т. д.

Так как $U(x^{(0)})>U(x^{(1)})>U(x^{(2)})>\ldots$, то, двигаясь по такому пути, мы быстро приближаемся к точке с наименьшим зна-

чением $U$ (дно «ямы»), которая соответствует искомому корню $x$ системы (1). Обозначим через

$$\nabla U(x)=\begin{bmatrix}\dfrac{\partial U}{\partial x_1}\\ \cdot\\ \cdot\\ \cdot\\ \dfrac{\partial U}{\partial x_n}\end{bmatrix}$$

градиент *) функции $U(x)$.

Из векторных треугольников $OM_0M_1$, $OM_1M_2$, ... заключаем, что

$$x^{(p+1)}=x^{(p)}-\lambda_p\nabla U(x^{(p)})\qquad(p=0,\ 1,\ 2,\ \ldots).$$

Остается определить множители $\lambda_p$. Для этого рассмотрим скалярную функцию

$$\Phi(\lambda)=U[x^{(p)}-\lambda\nabla U(x^{(p)})].$$

Функция $\Phi(\lambda)$ дает изменение уровня функции $U$ вдоль соответствующей нормали к поверхности уровня в точке $x^{(p)}$. Множитель $\lambda=\lambda_p$ надо выбрать таким образом, чтобы $\Phi(\lambda)$ имела минимум. Беря производную по $\lambda$ и приравнивая ее нулю, получаем уравнение

$$\Phi'(\lambda)=\frac{d}{d\lambda}U[x^{(p)}-\lambda\nabla U(x^{(p)})]=0. \tag{4}$$

Наименьший положительный корень уравнения (4) и даст нам значение $\lambda_p$. Уравнение (4), вообще говоря, нужно решать численно. Поэтому укажем метод приближенного нахождения чисел $\lambda_p$. Будем считать, что $\lambda$ — малая величина, квадратом и высшими степенями которой можно пренебречь. Имеем:

$$\Phi(\lambda)=\sum_{i=1}^{n}\{f_i[x^{(p)}-\lambda\nabla U(x^{(p)})]\}^2.$$

Разлагая функции $f_i$ по степеням $\lambda$ с точностью до линейных членов, получим:

$$\Phi(\lambda)=\sum_{i=1}^{n}\left[f_i(x^{(p)})-\lambda\frac{\partial f_i(x^{(p)})}{\partial x}\nabla U(x^{(p)})\right]^2,$$

---

*) Градиент функции $U(x)$ (обозначается $\operatorname{grad} U$ или $\nabla U$, где символ $\nabla$ называется *наблой*) есть вектор, приложенный в точке $x$, имеющий направление нормали $n$ к поверхности уровня функции в данной точке в сторону возрастания $U$ и длину, равную $\dfrac{\partial U}{\partial n}$.

Справедлива формула

$$\nabla U=\frac{\partial U}{\partial x_1}e_1+\frac{\partial U}{\partial x_2}e_2+\ldots+\frac{\partial U}{\partial x_n}e_n,$$

где $e_i$ $(i=1,\ 2,\ \ldots,\ n)$ — орты пространства $E_n$.

где

$$\frac{\partial f_i}{\partial x}=\left[\frac{\partial f_i}{\partial x_1}, \frac{\partial f_i}{\partial x_2}, \ldots, \frac{\partial f_i}{\partial x_n}\right].$$

Отсюда

$$\Phi'(\lambda)=-2\sum_{i=1}^{n}\left[f_i(x^{(p)})-\lambda\frac{\partial f_i(x^{(p)})}{\partial x}\nabla U(x^{(p)})\right]\times$$
$$\times\frac{\partial f_i(x^{(p)})}{\partial x}\nabla U(x^{(p)})=0.$$

Следовательно,

$$\lambda_p=\frac{\sum\limits_{i=1}^{n} f_i(x^{(p)})\frac{\partial f_i(x^{(p)})}{\partial x}\nabla U(x^{(p)})}{\sum\limits_{i=1}^{n}\left[\frac{\partial f_i(x^{(p)})}{\partial x}\nabla U(x^{(p)})\right]^2}=$$
$$=\frac{(f(x^{(p)}), W(x^{(p)})\nabla U(x^{(p)}))}{(W(x^{(p)})\nabla U(x^{(p)}), W(x^{(p)})\nabla U(x^{(p)}))},$$

где

$$W(x)=\frac{df}{dx}=\begin{bmatrix}\frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n}\\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n}\\ \cdots & \cdots & \cdots & \cdots\\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n}\end{bmatrix}$$

— матрица Якоби вектор-функции $f$.

Далее, имеем:

$$\frac{\partial U}{\partial x_j}=\frac{\partial}{\partial x_j}\left\{\sum_{i=1}^{n}[f_i(x)]^2\right\}=2\sum_{i=1}^{n}f_i(x)\frac{\partial f_i(x)}{\partial x_j}.$$

Отсюда

$$\nabla U(x)=2\begin{bmatrix}\sum\limits_{i=1}^{n}\frac{\partial f_i(x)}{\partial x_1}f_i(x)\\ \cdots\cdots\cdots\\ \sum\limits_{i=1}^{n}\frac{\partial f_i(x)}{\partial x_n}f_i(x)\end{bmatrix}=2W'(x)f(x),$$

где $W'(x)$ — транспонированная матрица Якоби.

Поэтому окончательно

$$\mu_p=2\lambda_p=\frac{(f^{(p)}, W_pW_p'f^{(p)})}{(W_pW_p'f^{(p)}, W_pW_p'f^{(p)})}, \tag{5}$$

где для краткости положено

$$f^{(p)} = f(x^{(p)}); \quad W_p = W(x^{(p)}),$$

причем

$$x^{(p+1)} = x^{(p)} - \mu_p W'_p f^{(p)} \qquad (p = 0,\ 1,\ 2,\ \ldots). \tag{6}$$

Если допустить, что функция $f(x)$ дважды непрерывно дифференцируема в окрестности искомого корня $x$, то можно получить более точные формулы для поправок $\Delta x^{(p)} = x^{(p+1)} - x^{(p)}$ (см. [7]).

Пример. Методом скорейшего спуска приближенно вычислить корни системы

$$\left.\begin{aligned} &x + x^2 - 2yz = 0{,}1;\\ &y - y^2 + 3xz = -0{,}2;\\ &z + z^2 + 2xy = 0{,}3, \end{aligned}\right\}$$

расположенные в окрестности начала координат.

Решение. Имеем:

$$x^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Здесь

$$f = \begin{bmatrix} x + x^2 - 2yz - 0{,}1 \\ y - y^2 + 3xz + 0{,}2 \\ z + z^2 + 2xy - 0{,}3 \end{bmatrix}$$

и

$$W = \begin{bmatrix} 1 + 2x & -2z & -2y \\ 3z & 1 - 2y & 3x \\ 2y & 2x & 1 + 2z \end{bmatrix}.$$

Подставляя нулевое приближение, будем иметь:

$$f^{(0)} = \begin{bmatrix} -0{,}1 \\ 0{,}2 \\ -0{,}3 \end{bmatrix} \quad \text{и} \quad W_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = E.$$

По формулам (5) и (6) получаем первое приближение

$$\mu_0 = \frac{(f^{(0)},\ f^{(0)})}{(f^{(0)},\ f^{(0)})} = 1$$

и

$$x^{(1)} = x^{(0)} - 1 \cdot E f^{(0)} = \begin{bmatrix} 0{,}1 \\ -0{,}2 \\ 0{,}3 \end{bmatrix}.$$

Аналогично находим второе приближение $x^{(1)}$. Имеем:

$$f^{(1)}=\begin{bmatrix}0,13\\0,05\\0,05\end{bmatrix};\quad W_1=\begin{bmatrix}1,2 & -0,6 & 0,4\\0,9 & 1,4 & 0,3\\-0,4 & 0,2 & 1,6\end{bmatrix}.$$

Отсюда

$$W_1'f^{(1)}=\begin{bmatrix}0,181\\0,002\\0,147\end{bmatrix}$$

и

$$W_1W_1'f^{(1)}=\begin{bmatrix}0,2748\\0,2098\\0,1632\end{bmatrix}.$$

Следовательно,

$$\mu_1=\frac{0,13\cdot 0,2748+0,05\cdot 0,2098+0,05\cdot 0,1632}{0,2748^2+0,2098^2+0,1632^2}=\frac{0,054374}{0,14619797}=0,3719$$

и

$$x^{(2)}=\begin{bmatrix}0,1\\-0,2\\0,3\end{bmatrix}-0,37119\cdot\begin{bmatrix}0,181\\0,002\\0,147\end{bmatrix}=\begin{bmatrix}0,0327\\-0,2007\\0,2453\end{bmatrix}.$$

Для контроля вычислим невязку

$$f^{(2)}=\begin{bmatrix}0,032\\-0,017\\-0,007\end{bmatrix}.$$

## § 13. Метод скорейшего спуска для случая системы линейных уравнений

Рассмотрим систему линейных уравнений

$$\left.\begin{aligned}f_1&\equiv\sum_{j=1}^{n}a_{1j}x_j-b_1=0,\\f_2&\equiv\sum_{j=1}^{n}a_{2j}x_j-b_2=0,\\&\dots\dots\dots\dots\\f_n&\equiv\sum_{j=1}^{n}a_{nj}x_j-b_n=0\end{aligned}\right\}\qquad(1)$$

с действительной матрицей $A=[a_{ij}]$ и столбцом свободных членов

$$b=\begin{bmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{bmatrix}.$$

Тогда

$$f=Ax-b$$

$$W=\frac{df}{dx}=\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}=A.$$

Следовательно,

$$x^{(p+1)}=x^{(p)}-\mu_p A' r_p, \tag{2}$$

где $r_p=Ax^{(p)}-b$ — невязка вектора $x^{(p)}$ и

$$\mu_p=\frac{(r_p,\ AA'r_p)}{(AA'r_p,\ AA'r_p)} \qquad (p=0,\ 1,\ 2,\ \dots) \tag{3}$$

(ср. [5], [6]).

Применение формул (2) и (3) приводит к громоздким вычислениям. Поэтому на практике часто вместо «скорейшего спуска» пользуются просто «спуском», добиваясь минимума функции

$$U=(Ax-b,\ Ax-b).$$

При этом число шагов процесса, обеспечивающих заданную точность корней системы (1), вообще говоря, возрастает; однако можно добиться того, чтобы вычисление каждого шага было более простым.

В общей постановке полагают:

$$x^{(p+1)}=x^{(p)}-\lambda_p y^{(p)} \qquad (p=0,\ 1,\ 2,\ \dots),$$

где $y^{(p)}$ — произвольный вектор, направленный наружу поверхности уровня $U=\text{const}$, проходящей через точку $x^{(p)}$, т. е.

$$(\operatorname{grad} U(x^{(p)}),\ y^{(p)})>0.$$

Имеем:

$$r_{p+1}=Ax^{(p+1)}-b=Ax^{(p)}-b-\lambda_p Ay^{(p)}=r_p-\lambda_p Ay^{(p)}.$$

Один из возможных путей определения скалярного множителя $\lambda_p$ исходит из требования [7]

$$(r_{p+1},\ y^{(p)})=(r_p,\ y^{(p)})-\lambda_p(Ay^{(p)},\ y^{(p)})=0.$$

Отсюда

$$\lambda_p=\frac{(r_p,\ y^{(p)})}{(Ay^{(p)},\ y^{(p)})}.$$

В зависимости от выбора вектора $y^{(p)}$ получаются те или иные расчетные схемы. В частности, если матрица $A = A'$ — положительно определенная (гл. X, § 15), то, полагая $y^{(p)} = r_p$, будем иметь:

$$x^{(p+1)} = x^{(p)} - \frac{(r_p, r_p)}{(Ar_p, r_p)} r_p$$

($p = 0, 1, 2, \ldots$), причем $(\operatorname{grad} U(x^{(p)}), y^{(p)}) = 2(Ar_p, r_p) > 0$ при $r_p \neq 0$.

Пример. Методом скорейшего спуска решить систему уравнений

$$\left.\begin{aligned} 8x_1 - x_2 - 2x_3 &= 2{,}3; \\ 10x_2 + x_3 + 2x_4 &= -0{,}5; \\ -x_1 + 6x_3 + 2x_4 &= -1{,}2; \\ 3x_1 - x_2 + 2x_3 + 12x_4 &= 3{,}7. \end{aligned}\right\} \qquad (4)$$

Решение. Так как в матрице системы преобладают диагональные элементы, то в качестве начального вектора $x^{(0)}$ примем вектор, координаты которого представляют собой округленные значения корней системы:

$$8x_1 = 2{,}3; \qquad 6x_3 = -1{,}2;$$
$$10x_2 = -0{,}5; \qquad 12x_4 = 3{,}7.$$

Отсюда, например,

$$x^{(0)} = \begin{bmatrix} 0{,}3 \\ -0{,}05 \\ -0{,}2 \\ 0{,}3 \end{bmatrix}.$$

Следовательно,

$$r_0 = Ax^{(0)} - b = \begin{bmatrix} 8 & -1 & -2 & 0 \\ 0 & 10 & 1 & 2 \\ -1 & 0 & 6 & 2 \\ 3 & -1 & 2 & 12 \end{bmatrix} \begin{bmatrix} 0{,}3 \\ -0{,}05 \\ -0{,}2 \\ 0{,}3 \end{bmatrix} - \begin{bmatrix} 2{,}3 \\ -0{,}5 \\ -1{,}2 \\ 3{,}7 \end{bmatrix} = \begin{bmatrix} 0{,}55 \\ 0{,}4 \\ 0{,}3 \\ 0{,}45 \end{bmatrix}.$$

Далее,

$$A'r_0 = \begin{bmatrix} 8 & 0 & -1 & 3 \\ -1 & 10 & 0 & -1 \\ -2 & 1 & 6 & 2 \\ 0 & 2 & 2 & 12 \end{bmatrix} \begin{bmatrix} 0{,}55 \\ 0{,}4 \\ 0{,}3 \\ 0{,}45 \end{bmatrix} = \begin{bmatrix} 5{,}45 \\ 3{,}0 \\ 2{,}0 \\ 6{,}8 \end{bmatrix}$$

и

$$AA'r_0 = \begin{bmatrix} 8 & -1 & -2 & 0 \\ 0 & 10 & 1 & 2 \\ -1 & 0 & 6 & 2 \\ 3 & -1 & 2 & 12 \end{bmatrix} \begin{bmatrix} 5{,}45 \\ 3{,}0 \\ 2{,}0 \\ 6{,}8 \end{bmatrix} = \begin{bmatrix} 36{,}6 \\ 45{,}6 \\ 20{,}15 \\ 98{,}95 \end{bmatrix}.$$

Применяя формулу (3), получаем:

$$\mu_0 = \frac{(r_0,\ AA'r_0)}{(AA'r_0,\ AA'r_0)} = \frac{0{,}55\cdot 36{,}6 + 0{,}4\cdot 45{,}6 + 0{,}3\cdot 20{,}15 + 0{,}45\cdot 98{,}95}{36{,}6^2 + 45{,}6^2 + 20{,}15^2 + 98{,}95^2} = \frac{88{,}9425}{13616{,}0452} = 0{,}006532.$$

Отсюда

$$x^{(1)} = x^{(0)} - \mu_0 A'r_0 = \begin{bmatrix} 0{,}3 \\ -0{,}05 \\ -0{,}2 \\ 0{,}3 \end{bmatrix} - 0{,}006\,532 \begin{bmatrix} 5{,}45 \\ 3{,}0 \\ 2{,}0 \\ 6{,}8 \end{bmatrix} = \begin{bmatrix} 0{,}2644 \\ -0{,}0696 \\ -0{,}2131 \\ 0{,}2556 \end{bmatrix},$$

причем

$$r^{(1)} = Ax^{(1)} - b = \begin{bmatrix} 0{,}3109 \\ 0{,}1020 \\ 0{,}1684 \\ -0{,}1966 \end{bmatrix}.$$

Аналогично находятся дальнейшие приближения и соответствующие невязки:

$$x^{(2)} = \begin{bmatrix} 0{,}2351 \\ -0{,}0849 \\ -0{,}2147 \\ 0{,}2863 \end{bmatrix}, \quad r_2 = \begin{bmatrix} 0{,}0956 \\ 0{,}0087 \\ 0{,}2493 \\ 0{,}0967 \end{bmatrix};$$

$$x^{(3)} = \begin{bmatrix} 0{,}2296 \\ -0{,}0842 \\ -0{,}2251 \\ 0{,}2748 \end{bmatrix}, \quad r_3 = \begin{bmatrix} 0{,}0712 \\ -0{,}0280 \\ 0{,}1692 \\ -0{,}0806 \end{bmatrix};$$

$$x^{(4)} = \begin{bmatrix} 0{,}2266 \\ -0{,}0792 \\ -0{,}2379 \\ 0{,}2875 \end{bmatrix}, \quad r_4 = \begin{bmatrix} 0{,}0680 \\ 0{,}0354 \\ 0{,}1211 \\ 0{,}0334 \end{bmatrix};$$

$$x^{(5)} = \begin{bmatrix} 0{,}2228 \\ -0{,}0810 \\ -0{,}2430 \\ 0{,}2823 \end{bmatrix}, \quad r_5 = \begin{bmatrix} 0{,}0493 \\ 0{,}0013 \\ 0{,}0839 \\ -0{,}0493 \end{bmatrix}$$

и т. д.

Заметим, что в данном случае процесс приближений сходится медленно: после пятого приближения мы еще далеки от точных корней системы (4), равных $x_1=0{,}2$; $x_2=-0{,}1$; $x_3=-0{,}3$; $x_4=0{,}3$.

## § 14*. Метод степенных рядов

Пусть дана нелинейная система

$$f_k(x_1, x_2, \ldots, x_n)=0 \tag{1}$$

($k=1, 2, \ldots, n$), где функции $f_k$ — аналитические в окрестности изолированного решения $x^*=(x_1^*, x_2^*, \ldots, x_n^*)$.

Рассмотрим более общую систему [8]

$$F_k(x_1, x_2, \ldots, x_n; \lambda)=0 \tag{2}$$

($k=1, 2, \ldots, n$), зависящую от действительного параметра $\lambda$ и такую, что при $\lambda=0$ система (2) решается непосредственно, а при $\lambda=1$ — тождественна системе (1), т. е.

$$F_k(x_1, x_2, \ldots, x_n; 1)\equiv f_k(x_1, x_2, \ldots, x_n)$$

($k=1, 2, \ldots, n$). Параметр $\lambda$ следует вводить так, чтобы зависимость функций $F_k$ от $\lambda$ была по возможности простой. Например, если $x^{(0)}=(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})$ — грубое приближение решения, то можно положить:

$$\sum_{j=1}^{n}(x_j-x_j^{(0)})\frac{\partial f_k(x^{(0)})}{\partial x_j}+\lambda\left[f_k(x)-\sum_{j=1}^{n}(x_j-x_j^{(0)})\frac{\partial f_k(x^{(0)})}{\partial x_j}\right]=0$$

($k=1, 2, \ldots, n$), где

$$x=(x_1, x_2, \ldots, x_n).$$

Мы будем предполагать, что $F_k$ — аналитические функции от $\lambda$ при $|\lambda|\leqslant 1$.

Пусть при $|\lambda|\leqslant 1$ система (2) имеет простое аналитическое решение $x_j(\lambda)$ ($j=1, 2, \ldots, n$), которое при $\lambda=1$ совпадает с $x_j^*$ ($j=1, 2, \ldots, n$). Положим

$$x_j(0)=x_j^{(0)} \qquad (j=1, 2, \ldots, n),$$

где $x_j^{(0)}$ ($j=1, 2, \ldots, n$) — известное решение системы (2) при $\lambda=0$. Разлагая функции $x_j(\lambda)$ в ряд Тейлора в точке $\lambda=0$, получим:

$$x_j(\lambda)=x_j(0)+\lambda x_j'(0)+\frac{\lambda^2}{2!}x_j''(0)+\ldots \qquad (j=1, 2, \ldots, n). \tag{3}$$

Для определения коэффициентов $x_j'(0)$ продифференцируем по параметру $\lambda$ равенство (2):

$$\sum_{j=1}^{n} \frac{\partial F_k}{\partial x_j} x_j'(\lambda) + \frac{\partial F_k}{\partial \lambda} = 0 \qquad (k = 1, 2, \ldots, n). \tag{4}$$

Полагая $x = x^{(0)}$ и $\lambda = 0$, будем иметь:

$$\sum_{j=1}^{n} \frac{\partial F_k(x^{(0)}; 0)}{\partial x_j} x_j'(0) = -\frac{\partial F_k(x^{(0)}; 0)}{\partial \lambda} \qquad (k = 1, 2, \ldots, n).$$

Отсюда, если

$$\det\left[\frac{\partial F_k(x^{(0)}; 0)}{\partial x_j}\right] \neq 0,$$

находим $x_j'(0)$.

Далее, дифференцируя по $\lambda$ равенство (4), получим:

$$\sum_{j=1}^{n} \frac{\partial F_k}{\partial x_j} x_j''(\lambda) + \sum_{j=1}^{n} \sum_{l=1}^{n} \frac{\partial^2 F_k}{\partial x_j \partial x_l} x_j'(\lambda) x_l'(\lambda) +$$

$$+ 2\sum_{j=1}^{n} \frac{\partial^2 F_k}{\partial x_j \partial \lambda} x_j'(\lambda) + \frac{\partial^2 F_k}{\partial \lambda^2} = 0.$$

Отсюда при $x = x^{(0)}$ и $\lambda = 0$ находим:

$$\sum_{j=1}^{n} \frac{\partial F_k(x^{(0)}; 0)}{\partial x_j} x_j''(0) = -\sum_{j=1}^{n} \sum_{l=1}^{n} \frac{\partial^2 F_k(x^{(0)}; 0)}{\partial x_j \partial x_l} x_j'(0) x_l'(0) -$$

$$-2\sum_{j=1}^{n} \frac{\partial^2 F_k(x^{(0)}; 0)}{\partial x_j \partial \lambda} x_j'(0) - \frac{\partial^2 F_k(x^{(0)}; 0)}{\partial \lambda^2} \qquad (k = 1, 2, \ldots, n). \tag{5}$$

Так как $x_j'(0)$ известны, то из системы (5) можно определить $x_j''(0)$. Аналогично вычисляются производные $x'''(0)$, $x^{\mathrm{IV}}(0)$, ...

Заметим, что матрица коэффициентов при старших производных оказывается все время одной и той же и равна матрице Якоби функций $F_1, F_2, \ldots, F_n$ относительно переменных $x_1, x_2, \ldots, x_n$ при $x_j = x_j^{(0)}$ $(j = 1, 2, \ldots, n)$ и $\lambda = 0$.

Предполагая, что ряды (3) сходятся при $\lambda = 1$, окончательно находим:

$$x_j^* = x_j(1) = x_j(0) + x_j'(0) + \frac{1}{2!} x_j''(0) + \ldots \qquad (j = 1, 2, \ldots, n). \tag{6}$$

Недостатком метода является сложность вычислений в общем случае производных высших порядков. Кроме того, сходимость ряда (6) может быть недостаточно быстрой.

При применении метода не обязательно предполагать аналитичность функций $x_j(\lambda)$ $(j=1, 2, \ldots, n)$, а именно: вместо ряда Тейлора можно воспользоваться формулой Тейлора, оборвав ряды $x_j(\lambda)$ на некоторой степени $\lambda^s$ и оценив их остатки по известным формулам (гл. III, § 4).

### Литература к тринадцатой главе

1. Л. В. Канторович, О методе Ньютона, Труды Матем. и-та им. Стеклова, XXVIII, М.—Л. (1949), 104—144.
2. A. Ostrowski, Сборник работ памяти Д. А. Граве, 1940, стр. 213.
3. Дж. Скарборо, Численные методы математического анализа, ГТТИ, М.—Л., 1934, гл. IX.
4. Д. А. Вентцель, Е. С. Вентцель, Элементы теории приближенных вычислений, Изд. ВВИА им. Жуковского, М., 1949, гл. III, § 8.
5. В. Э. Милн, Численное решение дифференциальных уравнений, ИЛ, 1955, гл. IX.
6. А. С. Хаусхолдер, Основы численного анализа, ИЛ, 1956, гл. III.
7. Э. Бут, Численные методы, Физматгиз, М., 1959.
8. Современная математика для инженеров, под редакцией Э. Ф. Беккенбаха, ИЛ, М., 1958, гл. XIV. Ч. Моррей, Нелинейные методы.

ГЛАВА XIV

# ИНТЕРПОЛИРОВАНИЕ ФУНКЦИЙ

## § 1. Конечные разности различных порядков

Пусть

$$y = f(x)$$

— заданная функция. Обозначим через $\Delta x = h$ фиксированную величину приращения аргумента (*шаг*). Тогда выражение

$$\Delta y \equiv \Delta f(x) = f(x + \Delta x) - f(x) \tag{1}$$

называется *первой конечной разностью* функции $y$. Аналогично определяются *конечные разности высших порядков*

$$\Delta^n y = \Delta(\Delta^{n-1} y) \qquad (n = 2,\ 3,\ \ldots).$$

Например,

$$\Delta^2 y = \Delta[f(x + \Delta x) - f(x)] = [f(x + 2\Delta x) - f(x + \Delta x)] -$$
$$- [f(x + \Delta x) - f(x)] = f(x + 2\Delta x) - 2f(x + \Delta x) + f(x).$$

Пример. Построить конечные разности для функции

$$P(x) = x^3,$$

считая шаг $\Delta x = 1$.

Решение. Имеем:

$$\Delta P(x) = (x + 1)^3 - x^3 = 3x^2 + 3x + 1,$$
$$\Delta^2 P(x) = [3(x + 1)^2 + 3(x + 1) + 1] - (3x^2 + 3x + 1) = 6x + 6,$$
$$\Delta^3 P(x) = [6(x + 1) + 6] - (6x + 6) = 6,$$
$$\Delta^n P(x) = 0 \quad \text{при} \quad n > 3.$$

Обратим внимание, что конечная разность третьего порядка функции $P(x)$ постоянна.

Вообще, справедливо утверждение: если

$$P_n(x) = a_0 x^n + a_1 x^{n-1} + \ \ldots \ + a_n$$

— полином $n$-й степени, то $\Delta^n P_n(x) = n! a_0 h^n = \text{const}$, где $\Delta x = h$.

Действительно, имеем:

$$\Delta P_n(x) = P_n(x+h) - P_n(x) = a_0[(x+h)^n - x^n] + \\ + a_1[(x+h)^{n-1} - x^{n-1}] + \ldots + a_{n-1}[(x+h) - h].$$

Раскрыв по биному Ньютона круглые скобки, легко убедиться, что $\Delta P_n(x)$ представляет собой полином $(n-1)$-й степени:

$$\Delta P_n(x) = b_0 x^{n-1} + b_1 x^{n-2} + \ \ldots \ + b_{n-1},$$

где

$$b_0 = nha_0.$$

Рассуждая аналогично, приходим к выводу, что вторая разность $\Delta^2 P_n(x)$ есть полином $(n-2)$-й степени:

$$\Delta^2 P_n(x) = c_0 x^{n-2} + c_1 x^{n-3} + \ \ldots \ + c_{n-2},$$

причем

$$c_0 = (n-1)\,hb_0 = n(n-1)\,h^2 a_0.$$

Проводя последовательно аналогичные рассуждения, мы в конце концов установим, что

$$\Delta^n P_n(x) = n! a_0 h^n = \text{const}.$$

Как следствие получаем:

$$\Delta^s P_n(x) = 0 \quad \text{при} \quad s > n.$$

Символ $\Delta$ (дельта) можно рассматривать как *оператор*, ставящий в соответствие функции $y = f(x)$ функцию $\Delta y = f(x+\Delta x) - f(x)$ ($\Delta x$ постоянно). Легко проверить основные свойства оператора $\Delta$:

1) $\Delta(u+v) = \Delta u + \Delta v$;

2) $\Delta(Cu) = C\,\Delta u$ ($C$ — постоянная);

3) $\Delta^m(\Delta^n y) = \Delta^{m+n} y$,

где $m$ и $n$ — целые неотрицательные числа, причем по определению полагают $\Delta^0 y = y$.

Из формулы (1) имеем:

$$f(x+\Delta x) = f(x) + \Delta f(x);$$

отсюда, рассматривая $\Delta$ как символический множитель, получим:

$$f(x+\Delta x) = (1+\Delta) f(x). \tag{2}$$

Последовательно применяя это соотношение $n$ раз, будем иметь:

$$f(x+n\Delta x) = (1+\Delta)^n f(x). \tag{3}$$

Воспользовавшись формулой бинома Ньютона*), окончательно выводим:

$$f(x+n\,\Delta x)=\sum_{m=0}^{n} C_n^m \Delta^m f(x), \tag{4}$$

где

$$C_n^m=\frac{n\,(n-1)\,\ldots\,[n-(m-1)]}{m!}$$

— число сочетаний из $n$ элементов по $m$.

Таким образом, с помощью формулы (4) последовательные значения функции $f(x)$ выражаются через ее конечные разности различных порядков.

Воспользовавшись тождеством

$$\Delta=(1+\Delta)-1 \tag{5}$$

и применяя бином Ньютона, получаем:

$$\Delta^n f(x)=[(1+\Delta)-1]^n f(x)=(1+\Delta)^n f(x)-C_n^1(1+\Delta)^{n-1}f(x)+$$
$$+C_n^2(1+\Delta)^{n-2}f(x)-\ \ldots\ +(-1)^n f(x).$$

Отсюда в силу формулы (3) будем иметь:

$$\Delta^n f(x)=f(x+n\,\Delta x)-C_n^1 f[x+(n-1)\,\Delta x]+$$
$$+C_n^2 f[x+(n-2)\,\Delta x]-\ldots+(-1)^n f(x). \tag{6}$$

Формула (6) дает выражение конечной разности $n$-го порядка функции $f(x)$ через последовательные значения этой функции.

Пусть функция $f(x)$ имеет непрерывную производную $f^{(n)}(x)$ на отрезке $[x,\ x+n\,\Delta x]$. Тогда справедлива важная формула

$$\Delta^n f(x)=(\Delta x)^n f^{(n)}(x+\theta n\,\Delta x), \tag{7}$$

где

$$0<\theta<1.$$

Формулу (7) проще всего доказать, используя метод математической индукции.

В самом деле, при $n=1$ мы получаем теорему Лагранжа о конечном приращении функции и, следовательно, формула (7) верна. Пусть теперь при $k<n$ имеем:

$$\Delta^k f(x)=(\Delta x)^k f^{(k)}(x+\theta' k\,\Delta x),$$

где

$$0<\theta'<1.$$

---

*) Законность применения формулы бинома Ньютона предоставляем обосновать читателю.

Тогда

$$\Delta^{k+1}f(x)=\Delta^k[f(x+\Delta x)-f(x)]=$$
$$=(\Delta x)^k[f^{(k)}(x+\Delta x+\theta' k\,\Delta x)-f^{(k)}(x+\theta' k\,\Delta x)].$$

Применяя теорему Лагранжа к получившемуся приращению производной $f^{(k)}(x)$, будем иметь:

$$\Delta^{k+1}f(x)=(\Delta x)^k\,\Delta x f^{(k+1)}(x+\theta' k\,\Delta x+\theta''\,\Delta x),$$

где $0<\theta''<1$. Полагая

$$\frac{\theta' k+\theta''}{k+1}=\theta, \tag{8}$$

окончательно получим:

$$\Delta^{k+1}f(x)=(\Delta x)^{k+1}f^{(k+1)}(x+\theta(k+1)\,\Delta x),$$

причем, очевидно,

$$0<\theta<1.$$

Таким образом, установлен переход от $k$ к $k+1$ и, следовательно формула (7) доказана.

Из формулы (7) имеем:

$$f^{(n)}(x+\theta n\,\Delta x)=\frac{\Delta^n f(x)}{(\Delta x)^n}.$$

Отсюда, переходя к пределу при $\Delta x\to 0$ и предполагая, что производная $f^{(n)}(x)$ непрерывна, получим:

$$f^{(n)}(x)=\lim_{\Delta x\to 0}\frac{\Delta^n f(x)}{(\Delta x)^n}. \tag{9}$$

Следовательно, при малых $\Delta x$ справедлива приближенная формула

$$f^{(n)}(x)\approx\frac{\Delta^n f(x)}{(\Delta x)^n}. \tag{10}$$

## § 2. Таблица разностей

Часто приходится рассматривать функции $y=f(x)$, заданные табличными значениями $y_i=f(x_i)$ для системы равноотстоящих точек $x_i$ $(i=0,\ 1,\ 2,\ \ldots)$, где

$$\Delta x_i=x_{i+1}-x_i=h=\text{const}.$$

Конечные разности последовательности $y_i$ естественно определяются соотношениями

$$\Delta y_i=y_{i+1}-y_i,$$
$$\Delta^2 y_i=\Delta(\Delta y_i)=\Delta y_{i+1}-\Delta y_i,$$
$$\cdots\cdots\cdots\cdots$$
$$\Delta^n y_i=\Delta(\Delta^{n-1}y_i)=\Delta^{n-1}y_{i+1}-\Delta^{n-1}y_i.$$

Из первого равенства имеем:

$$y_{i+1} = y_i + \Delta y_i = (1 + \Delta) y_i.$$

Отсюда последовательно выводим:

$$\begin{aligned} y_{i+2} &= (1 + \Delta) y_{i+1} = (1 + \Delta)^2 y_i, \\ y_{i+3} &= (1 + \Delta) y_{i+2} = (1 + \Delta)^3 y_i, \\ &\ldots\ldots\ldots\ldots \\ y_{i+n} &= (1 + \Delta)^n y_i. \end{aligned}$$

Использовав формулу бинома Ньютона, получим:

$$y_{i+n} = y_i + C_n^1 \Delta y_i + C_n^2 \Delta^2 y_i + \ldots + \Delta^n y_i.$$

Обратно, имеем:

$$\Delta^n y_i = [(1 + \Delta) - 1]^n y_i = (1 + \Delta)^n y_i - C_n^1 (1 + \Delta)^{n-1} y_i + C_n^2 (1 + \Delta)^{n-2} y_i - \ldots \ldots + (-1)^n y_i$$

или

$$\Delta^n y_i = y_{n+i} - C_n^1 y_{n+i-1} + C_n^2 y_{n+i-2} - \ldots + (-1)^n y_i.$$

Например,

$$\Delta^2 y_i = y_{i+2} - 2y_{i+1} + y_i,$$
$$\Delta^3 y_i = y_{i+3} - 3y_{i+2} + 3y_{i+1} - y_i$$

и т. д. Заметим, что для вычисления $n$-й разности $\Delta^n y_i$ нужно знать $n+1$ членов $y_i, y_{i+1}, \ldots, y_{i+n}$ данной последовательности.

Таблица 33

**Горизонтальная таблица разностей**

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ | $\Delta^3 y$ |
|---|---|---|---|---|
| $x_0$ | $y_0$ | $\Delta y_0$ | $\Delta^2 y_0$ | $\Delta^3 y_0$ |
| $x_1$ | $y_1$ | $\Delta y_1$ | $\Delta^2 y_1$ | $\Delta^3 y_1$ |
| $x_2$ | $y_2$ | $\Delta y_2$ | $\Delta^2 y_2$ | $\Delta^3 y_2$ |
| . . . | . . . | . . . | . . . | . . . |
| . . . | . . . | . . . | . . . | . . |

Таблица 34

**Диагональная таблица разностей**

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ | $\Delta^3 y$ |
|---|---|---|---|---|
| $x_0$ | $y_0$ | | | |
| | | $\Delta y_0$ | | |
| $x_1$ | $y_1$ | | $\Delta^2 y_0$ | |
| | | $\Delta y_1$ | | $\Delta^3 y_0$ |
| $x_2$ | $y_2$ | | $\Delta^2 y_1$ | |
| | | $\Delta y_2$ | | |
| $x_3$ | $y_3$ | | | |

Конечные разности различных порядков удобно располагать в форме таблиц двух видов: *горизонтальной таблицы разностей* (таблица 33) или *диагональной таблицы разностей* (таблица 34).

Пример 1. Составить горизонтальную таблицу разностей функции

$$y = 2x^3 - 2x^2 + 3x - 1 \qquad (1)$$

от начального значения $x_0 = 0$, приняв шаг $h = 1$.

Решение. Полагая $x_0=0$, $x_1=1$, $x_2=2$, находим соответствующие значения $y_0=-1$, $y_1=2$, $y_2=13$. Отсюда имеем:

$$\Delta y_0=y_1-y_0=3,$$
$$\Delta y_1=y_2-y_1=11,$$
$$\Delta^2 y_0=\Delta y_1-\Delta y_0=8.$$

Эти значения заносим в таблицу (таблица 35). Так как наша функция есть полином третьей степени, то третья разность ее постоянна (см. § 1) и равна

$$\Delta^3 y_i=2\cdot 3!=12.$$

Таблица 35

**Горизонтальная таблица разностей кубической функции (1)**

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ | $\Delta^3 y$ |
|---|---|---|---|---|
| 0 | −1 | 3 | 8 | 12 |
| 1 | 2 | 11 | 20 | 12 |
| 2 | 13 | 31 | 32 | 12 |
| 3 | 44 | 63 | 44 | 12 |
| 4 | 107 | 107 | 56 | 12 |
| 5 | 214 | 163 | 68 | 12 |
| . . . | . . . | . . . | . . . | . . . |

Поэтому дальнейшее заполнение таблицы 35 можно производить при помощи суммирования, используя формулы

$$\Delta^2 y_{i+1}=\Delta^2 y_i+12$$
$$(i=0,\ 1,\ 2,\ \ldots),$$
$$\Delta y_{i+1}=\Delta y_i+\Delta^2 y_i$$
$$(i=1,\ 2,\ \ldots),$$
$$y_{i+1}=y_i+\Delta y_i$$
$$(i=2,\ 3,\ \ldots).$$

Ступенчатой ломаной отмечены исходные данные для составления таблицы.

Замечание. При составлении таблицы разностей возможны случайные ошибки вычислителя. Посмотрим, как отразится на значениях разностей ошибка $\varepsilon$ в значении $y_n$. Составляя соответствующую диагональную таблицу разностей, получим таблицу 36.

Из таблицы 36 видно: 1) если $y_n$ содержит ошибку, то ошибочными являются также разности

$$\Delta y_{n-1},\qquad \Delta y_n;\qquad \Delta^2 y_{n-2},\qquad \Delta^2 y_{n-1},\ \Delta^2 y_n$$

и т. д.; 2) в $k$-е разности $\Delta^k y$ ошибки входят со знакопеременными биномиальными коэффициентами, а именно, ошибки соответственно имеют значения

$$C_k^0\varepsilon,\qquad -C_k^1\varepsilon,\qquad C_k^2\varepsilon,\ \ldots,\ (-1)^k C_k^k\varepsilon$$

и, следовательно, абсолютное значение максимальной ошибки $k$-й разности быстро растет вместе с номером разности; 3) для каждой конечной разности $\Delta^k y$ сумма ошибок с учетом их знаков равна

Таблица 36

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ | $\Delta^3 y$ | $\Delta^4 y$ |
|---|---|---|---|---|---|
| ..... | ..... | ..... | ..... | ..... | ..... |
| $x_{n-4}$ | $y_{n-4}$ | | | | |
| | | $\Delta y_{n-4}$ | | | |
| $x_{n-3}$ | $y_{n-3}$ | | $\Delta^2 y_{n-4}$ | | |
| | | $\Delta y_{n-3}$ | | $\Delta^3 y_{n-4}$ | |
| $x_{n-2}$ | $y_{n-2}$ | | $\Delta^2 y_{n-3}$ | | $\Delta^4 y_{n-4}+\varepsilon$ |
| | | $\Delta y_{n-2}$ | | $\Delta^3 y_{n-3}+\varepsilon$ | |
| $x_{n-1}$ | $y_{n-1}$ | | $\Delta^2$ $+\varepsilon$ | | $\Delta^4 y_{n-3}-4\varepsilon$ |
| | | $\Delta y_{n-1}+\varepsilon$ | | $\Delta^3 y_{n-2}-3\varepsilon$ | |
| $x_n$ | $y_n+\varepsilon$ | | $\Delta^2 y_{n-1}-2\varepsilon$ | | $\Delta^4 y_{n-2}+6\varepsilon$ |
| | | $\Delta y_n-\varepsilon$ | | $\Delta^3 y_{n-1}+3\varepsilon$ | |
| $x_{n+1}$ | $y_{n+1}$ | | $\Delta^2 y_n+\varepsilon$ | | $\Delta^4 y_{n-1}-4\varepsilon$ |
| | | $\Delta y_{n+1}$ | | $\Delta^3 y_n-\varepsilon$ | |
| $x_{n+2}$ | $y_{n+2}$ | | $\Delta^2 y_{n+1}$ | | $\Delta^4 y_n+\varepsilon$ |
| | | $\Delta y_{n+2}$ | | $\Delta^3 y_{n+1}$ | |
| $x_{n+3}$ | $y_{n+3}$ | | $\Delta^2 y_{n+2}$ | | |
| | | $\Delta y_{n+3}$ | | | |
| $x_{n+4}$ | $y_{n+4}$ | | | | |

нулю, а сумма абсолютных величин ошибок равна $|\varepsilon|\cdot 2^k$. Таким образом, даже незначительная ошибка в значении функции приводит к значительным ошибкам в ее разностях высокого порядка. Заметим, что максимальная ошибка разностей $\Delta^k y$ в случае диагональной таблицы разностей находится в той же горизонтальной строке, что и ошибочная табличная величина $y_n$, или же в верхней и нижней соседних строках.

Рассмотренный *закон распространения ε-ошибки* в таблице конечных разностей дает возможность в некоторых случаях установить

наличие и место ошибки, а также ее числовое значение, что позволяет исправить таблицу.

Таблицы разностей обычно составляются с точностью до некоторого фиксированного десятичного разряда. Если функция $y=f(x)$ имеет непрерывные производные до $m$-го порядка, то при достаточно малом шаге $h=\Delta x$ ее конечные разности до $m$-го порядка включительно изменяются плавно, причем разность $m$-го порядка почти постоянна в пределах данных десятичных разрядов. Нарушение последнего условия на каком-нибудь участке таблицы при отсутствии особенностей функции, вообще говоря, свидетельствует о наличии вычислительной ошибки.

Установив максимальное отклонение $m$-й разности от нормы, можно определить место этой ошибки в столбце значений функции $y$ в предположении, что: 1) эта ошибка — одиночная и заключается в неверном подсчете одного значения функции и 2) при вычислении конечных разностей новых ошибок не было. Если такая ошибка в таблице разностей обнаружена, то исправление ее может быть осуществлено с помощью значений разностей. Покажем, каким путем это достигается, причем для простоты ограничимся случаем постоянства разностей второго или третьего порядка.

Пусть ошибочное табличное значение есть $y_n+\varepsilon$, где индекс $n$ установлен, а величина ошибки $\varepsilon$ неизвестна.

Если третьи разности практически постоянны, то вторые разности образуют арифметическую прогрессию, и поэтому верное значение второй разности $\Delta^2 y_{n-1}$ будет равно среднему арифметическому трех смежных ошибочных разностей:

$$\Delta^2 y_{n-1}=\frac{1}{3}\left[(\Delta^2 y_{n-2}+\varepsilon)+(\Delta^2 y_{n-1}-2\varepsilon)+(\Delta^2 y_n+\varepsilon)\right],$$

так как члены, содержащие $\varepsilon$, сокращаются.

По найденному верному значению второй разности $\Delta^2 y_{n-1}$ можно найти величину ошибки $\varepsilon$, а именно: эта ошибка будет равна полуразности между исправленным и ошибочным значениями разности $\Delta^2 y_{n-1}$

$$\varepsilon=\frac{1}{2}\left[\Delta^2 y_{n-1}-(\Delta^2 y_{n-1}-2\varepsilon)\right].$$

Верное же значение самой функции $y_n$ найдем из тождества

$$y_n=(y_n+\varepsilon)-\varepsilon.$$

Для контроля следует снова вычислить конечные разности.

Пример 2. Исправить ошибку в таблице 37.

Решение. Здесь плавный ход вторых разностей максимально нарушается при $x=19$. Имеющаяся ошибка распространяется на

три строки, объединенные фигурной скобкой. Находим среднее арифметическое значение второй разности для средней из трех объединенных строк

$$\Delta^2 y_{n-1} = \frac{10^{-3}}{3}(-4+8-4) = 0.$$

Отсюда

$$\varepsilon = \frac{1}{2}[0-0{,}008] = -0{,}004.$$

Внося исправление в табличное значение $y$ для $x=19$, получим:

$$y_n = (y_n+\varepsilon)-\varepsilon = 16{,}792-(-0{,}004) = 16{,}796.$$

Таблица 37

**Таблица разностей, содержащая одиночную ошибку**

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ | Ошибка |
|---|---|---|---|---|
| 15 | 13,260 | | | |
| | | 884 | | |
| 16 | 14,144 | | 0 | |
| | | 884 | | |
| 17 | 15,028 | | 0 | |
| | | 884 | | $\varepsilon$ |
| 18 | 15,912 | | (−4) 0 | |
| | | 88(0)4 | | $-2\varepsilon$ |
| 19 | 16,79(2)6 | | (8) 0 | |
| | | 88(8)4 | | |
| 20 | 17,680 | | (−4) 0 | $\varepsilon$ |
| | | 884 | | |
| 21 | 18,564 | | 0 | |
| | | 884 | | |
| 22 | 19,448 | | 0 | |
| | | 884 | | |
| 23 | 20,332 | | | |

После исправления получаем таблицу с плавным изменением первых разностей и постоянной второй разностью (неверные цифры взяты в скобки). Отметим, что указанным методом можно исправить лишь отдельные вычислительные ошибки или описки. Для устранения же большого количества ошибок, могущих появиться по разным причинам, а также для уменьшения накопления ошибок, возникающих в результате неточностей самих вычислительных методов и округления промежуточных результатов до данного числа знаков, применяются специальные приемы «сглаживания» [1].

## § 3. Обобщенная степень

В дальнейшем нам понадобится понятие об *обобщенной степени* [1].

Определение. Обобщенной $n$-степенью числа $x$ называется произведение $n$ сомножителей, первый из которых равен $x$, а каждый следующий на $h$ меньше предыдущего:

$$x^{[n]} = x(x-h)(x-2h)\ \ldots\ [x-(n-1)h], \tag{1}$$

где $h$ — некоторое фиксированное постоянное число.

Показатель обобщенной степени обычно записывается в квадратных скобках. Полагают $x^{[0]}=1$.

При $h=0$ обобщенная степень (1) совпадает с обычной

$$x^{[n]}=x^n.$$

Вычислим конечные разности для обобщенной степени, полагая $\Delta x=h$. Для первой разности имеем:

$$\begin{aligned}\Delta x^{[n]}&=(x+h)^{[n]}-x^{[n]}=\\&=(x+h)\,x\ \ldots\ [x-(n-2)\,h]-x\,(x-h)\ \ldots\ [x-(n-1)\,h]=\\&=x\,(x-h)\ \ldots\ [x-(n-2)\,h]\cdot\{(x+h)-[x-(n-1)\,h]\}=\\&=x\,(x-h)\ \ldots\ [x-(n-2)\,h]\,nh=nhx^{[n-1]},\end{aligned}$$

т. е.

$$\Delta x^{[n]}=nhx^{[n-1]}. \tag{2}$$

Подсчитываем вторую разность:

$$\begin{aligned}\Delta^2x^{[n]}=\Delta\,(\Delta x^{[n]})=\Delta\,(nhx^{[n-1]})=nh\cdot(n-1)\,hx^{[n-2]}=\\=nh^2\,(n-1)\,x^{[n-2]}.\end{aligned}$$

Итак,

$$\Delta^2x^{[n]}=n\,(n-1)\,h^2x^{[n-2]}.$$

Методом математической индукции легко доказать общую формулу

$$\Delta^kx^{[n]}=n\,(n-1)\ \ldots\ [n-(k-1)]\,h^kx^{[n-k]},$$

где $k=1,\ 2,\ \ldots,\ n$.

Очевидно,

$$\Delta^kx^{[n]}=0 \text{ при } k>n.$$

Из формулы (2) вытекает также простая формула *конечного суммирования*. Пусть

$$x_0,\ x_1,\ x_2,\ \ldots$$

— равноотстоящие точки с шагом $h$

$$x_{i+1}-x_i=h \qquad (i=0,\ 1,\ 2\ \ldots).$$

Рассмотрим сумму

$$S_N=\sum_{i=0}^{N-1}x_i^{[n]}.$$

Так как в силу формулы (2) имеем:

$$x^{[n]}=\frac{\Delta x^{[n+1]}}{h(n+1)},$$

то

$$S_N = \frac{1}{h(n+1)} \sum_{i=0}^{N-1} \Delta x_i^{[n+1]} =$$

$$= \frac{1}{h(n+1)} \left\{ x_1^{[n+1]} - x_0^{[n+1]} + x_2^{[n+1]} - x_1^{[n+1]} + \ldots + x_N^{[n+1]} - x_{N-1}^{[n+1]} \right\} =$$

$$= \frac{1}{h(n+1)} \left( x_N^{[n+1]} - x_0^{[n+1]} \right).$$

Итак,

$$\sum_{i=0}^{N-1} x_i^{[n]} = \frac{x_N^{[n+1]} - x_0^{[n+1]}}{h(n+1)}. \tag{3}$$

Формула (3) аналогична формуле Ньютона — Лейбница для целой положительной степени.

## § 4. Постановка задачи интерполирования

Простейшая задача интерполирования [2] заключается в следующем. На отрезке $[a, b]$ заданы $n+1$ точки $x_0, x_1, \ldots, x_n$, которые

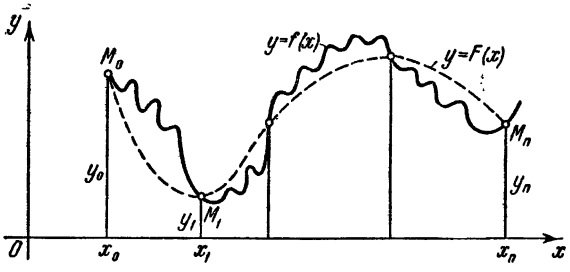


Рис. 61.

называются *узлами интерполяции*, и значения некоторой функции $f(x)$ в этих точках

$$f(x_0) = y_0, \quad f(x_1) = y_1, \quad \ldots, \quad f(x_n) = y_n. \tag{1}$$

Требуется построить функцию $F(x)$ (*интерполирующая функция*), принадлежащую известному классу и принимающую в узлах интерполяции те же значения, что и $f(x)$, т. е. такую, что

$$F(x_0) = y_0, \quad F(x_1) = y_1, \quad \ldots, \quad F(x_n) = y_n. \tag{2}$$

Геометрически это означает, что нужно найти кривую $y = F(x)$ некоторого определенного типа, проходящую через заданную систему точек $M_i(x_i, y_i)$ $(i = 0, 1, 2, \ldots)$ (рис. 61).

В такой общей постановке задача может иметь бесчисленное множество решений или совсем не иметь решений.

Однако эта задача становится однозначной, если вместо произвольной функции $F(x)$ искать полином $P_n(x)$ степени не выше $n$, удовлетворяющий условиям (2), т. е. такой, что

$$P_n(x_0)=y_0, \qquad P_n(x_1)=y_1, \ \ldots, \ P_n(x_n)=y_n.$$

Полученную интерполяционную формулу

$$y=F(x)$$

обычно используют для приближенного вычисления значений данной функции $f(x)$ для значений аргумента $x$, отличных от узлов интерполирования. Такая операция называется *интерполированием функции* $f(x)$. При этом различают *интерполирование в узком смысле*, когда $x \in [x_0, x_n]$, т. е. значение $x$ является промежуточным между $x_0$ и $x_n$, и *экстраполирование*, когда $x \overline{\in} [x_0, x_n]$. В дальнейшем под термином *интерполирование* мы будем понимать как первую, так и вторую операции.

## § 5. Первая интерполяционная формула Ньютона

Пусть для функции $y=f(x)$ заданы значения $y_i=f(x_i)$ для равноотстоящих значений назависимой переменной: $x_i=x_0+ih$ $(i=0, 1, 2, \ldots, n)$, где $h$ — *шаг интерполяции*. Требуется подобрать полином $P_n(x)$ степени не выше $n$, принимающий в точках $x_i$ значения

$$P_n(x_i)=y_i \quad (i=0, 1, \ldots, n). \tag{1}$$

Условия (1) эквивалентны тому, что

$$\Delta^m P_n(x_0)=\Delta^m y_0$$

при $m=0, 1, 2, \ldots, n$.

Следуя Ньютону, будем искать полином в виде

$$\begin{aligned} P_n(x)=a_0+a_1(x-x_0)+a_2(x-x_0)(x-x_1)+ \\ +a_3(x-x_0)(x-x_1)(x-x_2)+\ldots \\ \ldots+a_n(x-x_0)(x-x_1)\ \ldots\ (x-x_{n-1}). \end{aligned} \tag{2}$$

Пользуясь обобщенной степенью, выражение (1) запишем так:

$$\begin{aligned} P_n(x)=a_0+a_1(x-x_0)^{[1]}+a_2(x-x_0)^{[2]}+ \\ +a_3(x-x_0)^{[3]}+\ldots+a_n(x-x_0)^{[n]}. \end{aligned} \tag{2'}$$

Наша задача состоит в определении коэффициентов $a_i$ $(i=0, 1, 2, \ldots, n)$ полинома $P_n(x)$. Полагая $x=x_0$ в выражении (2′), получим:

$$P_n(x_0)=y_0=a_0.$$

Чтобы найти коэффициент $a_1$, составим первую конечную разность

$$\Delta P_n(x) = a_1 h + 2a_2 (x-x_0)^{[1]} h + \\ + 3a_3 (x-x_0)^{[2]} h + \ldots + n a_n (x-x_0)^{[n-1]} h.$$

Полагая в последнем выражении $x = x_0$, получим:

$$\Delta P_n(x_0) = \Delta y_0 = a_1 h;$$

откуда

$$a_1 = \frac{\Delta y_0}{1!\,h}.$$

Для определения коэффициента $a_2$ составим конечную разность второго порядка

$$\Delta^2 P_n(x) = 2!\,h^2 a_2 + 2\cdot 3h^2 a_3 (x-x_0)^{[1]} + \ldots \\ \ldots + (n-1)\,n h^2 a_n (x-x_0)^{[n-2]}.$$

Положив $x = x_0$, получим:

$$\Delta^2 P_n(x_0) = \Delta^2 y_0 = 2!\,h^2 a_2;$$

откуда

$$a_2 = \frac{\Delta^2 y_0}{2!\,h^2}.$$

Последовательно продолжая этот процесс, мы обнаружим, что

$$a_i = \frac{\Delta^i y_0}{i!\,h^i} \quad (i = 0,\ 1,\ 2,\ \ldots,\ n),$$

где положено

$$0! = 1 \quad \text{и} \quad \Delta^0 y = y.$$

Подставляя найденные значения коэффициентов $a_i$ в выражение (2′), получим *интерполяционный полином Ньютона*

$$P_n(x) = y_0 + \frac{\Delta y_0}{1!\,h}(x-x_0)^{[1]} + \frac{\Delta^2 y_0}{2!\,h^2}(x-x_0)^{[2]} + \ldots \\ \ldots + \frac{\Delta^n y_0}{n!\,h^n}(x-x_0)^{[n]}. \qquad (3)$$

Легко видеть, что полином (3) полностью удовлетворяет требованиям поставленной задачи. Действительно, во-первых, степень полинома $P_n(x)$ не выше $n$, во-вторых,

$$P_n(x_0) = y_0$$

и

$$P_n(x_k) = y_0 + \frac{\Delta y_0}{h}(x_k - x_0) + \frac{\Delta^2 y_0}{2!\,h^2}(x_k - x_0)(x_k - x_1) + \ldots \\ \ldots + \frac{\Delta^k y_0}{k!\,h^k}(x_k - x_0)(x_k - x_1)\ldots(x_k - x_{k-1}) = \\ = y_0 + k\,\Delta y_0 + \frac{k(k-1)}{2!}\Delta^2 y_0 + \ldots + \frac{k(k-1)\ldots 1}{k!}\Delta^k y_0 = \\ = (1+\Delta)^k y_0 = y_k \quad (k = 1,\ 2,\ \ldots,\ n).$$

Заметим, что при $h \to 0$ формула (3) превращается в полином Тейлора для функции $y$.

В самом деле,

$$\lim_{h \to 0} \frac{\Delta^k y_0}{h^k} = \left(\frac{d^k y}{dx^k}\right)_{x=x_0} = y^{(k)}(x_0).$$

Кроме того, очевидно,

$$\lim_{h \to 0} (x - x_0)^{[n]} = (x - x_0)^n.$$

Отсюда при $h \to 0$ формула (3) принимает вид полинома Тейлора:

$$P_n(x) = y(x_0) + y'(x_0)(x - x_0) + \ldots + \frac{y^{(n)}(x_0)}{n!}(x - x_0)^n.$$

Для практического использования интерполяционную формулу Ньютона (3) обычно записывают в несколько преобразованном виде. Для этого введем новую переменную $q$ по формуле

$$q = \frac{x - x_0}{h};$$

тогда

$$\frac{(x - x_0)^{[i]}}{h^i} = \frac{(x - x_0)}{h} \cdot \frac{(x - x_0 - h)}{h} \cdot \frac{(x - x_0 - 2h)}{h} \ldots$$

$$\ldots \frac{[x - x_0 - (i-1)h]}{h} = q(q-1)(q-2)\ldots(q-i+1)$$

$$(i = 1, 2, \ldots, n).$$

Подставляя эти выражения в формулу (3), получим:

$$P_n(x) = y_0 + q\,\Delta y_0 + \frac{q(q-1)}{2!}\Delta^2 y_0 + \ldots$$

$$\ldots + \frac{q(q-1)\ldots(q-n+1)}{n!}\Delta^n y_0, \qquad (4)$$

где $q = \frac{x - x_0}{h}$ представляет собой *число шагов*, необходимых для достижения точки $x$, исходя из точки $x_0$. Это и есть окончательный вид *первой интерполяционной формулы Ньютона*.

Формулу (4) выгодно использовать для интерполирования функции $y = f(x)$ в окрестности начального значения $x_0$, где $q$ мало по абсолютной величине.

Если в формуле (4) положить $n = 1$, то получим формулу *линейного интерполирования*:

$$P_1(x) = y_0 + q\,\Delta y_0.$$

При $n = 2$ будем иметь формулу *параболического или квадратичного интерполирования:*

$$P_2(x) = y_0 + q\,\Delta y_0 + \frac{q(q-1)}{2}\Delta^2 y_0.$$

Если дана неограниченная таблица значений функции $y$, то число $n$ в интерполяционной формуле (4) может быть любым. Практически в этом случае число $n$ выбирают так, чтобы разность $\Delta^n y_i$ была постоянной с заданной степенью точности. За начальное значение $x_0$ можно принимать любое табличное значение аргумента $x$.

Если таблица значений функции конечна, то число $n$ ограничено, а именно: $n$ не может быть больше числа значений функции $y$, уменьшенного на единицу.

Заметим, что при применении первой интерполяционной формулы Ньютона удобно пользоваться горизонтальной таблицей разностей, так как тогда нужные значения разностей функции находятся в соответствующей горизонтальной строке таблицы.

Пример 1. Приняв шаг $h=0{,}05$, построить на отрезке [3,5; 3,6] интерполяционный полином Ньютона для функции $y=e^x$, заданной таблицей

| $x$ | 3,50 | 3,55 | 3,60 | 3,65 | 3,70 |
|---|---|---|---|---|---|
| $y$ | 33,115 | 34,813 | 36,598 | 38,475 | 40,447 |

Решение. Составляем таблицу разностей (таблица 38). Заметим, что в столбцах разностей, следуя обычной практике, мы не указываем десятичные разряды (которые ясны из столбца значений функции). Так как разности третьего порядка практически постоянны, то в формуле (4) полагаем $n=3$. Приняв $x_0=3{,}50$, $y_0=33{,}115$, будем иметь:

$$P_3(x)=33{,}115+1{,}698q+$$
$$+0{,}087\frac{q(q-1)}{2}+$$
$$+0{,}005\frac{q(q-1)(q-2)}{6}$$

или

$$P_3(x)=33{,}115+1{,}698q+$$
$$+0{,}0435q(q-1)+$$
$$+0{,}00083q(q-1)(q-2),$$

где

$$q=\frac{x-3{,}50}{0{,}05}=20(x-3{,}5).$$

Таблица 38

**Таблица разностей функции $y=e^x$**

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ | $\Delta^3 y$ |
|---|---|---|---|---|
| 3,50 | 33,115 | 1698 | 87 | 5 |
| 3,55 | 34,813 | 1785 | 92 | 3 |
| 3,60 | 36,598 | 1877 | 95 | |
| 3,65 | 38,475 | 1972 | | |
| 3,70 | 40,447 | | | |

Пример 2. В таблице 39 приведены значения *интеграла вероятностей*

$$\Phi(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-x^2}dx.$$

Применяя первую интерполяционную формулу Ньютона, приближенно найти $\Phi(1{,}43)$.

Решение. Дополняем таблицу 39 конечными разностями функции $y$ до третьего порядка включительно.

За $x_0$ принимаем ближайшее табличное значение к искомому значению $x = 1{,}43$, т. е. полагаем $x_0 = 1{,}4$. Так как $h = 0{,}1$, то

$$q = \frac{1{,}43 - 1{,}4}{0{,}1} = 0{,}3.$$

Таблица 39

**Таблица разностей функции $y = \Phi(x)$**

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ | $\Delta^3 y$ |
|---|---|---|---|---|
| 1,0 | 0,8427 | 375 | —74 | 10 |
| 1,1 | 0,8802 | 301 | —64 | 10 |
| 1,2 | 0,9103 | 237 | —54 | 9 |
| 1,3 | 0,9340 | 183 | —45 | 9 |
| 1,4 | 0,9523 | 138 | —36 | 9 |
| 1,5 | 0,9661 | 102 | —27 | 5 |
| 1,6 | 0,9763 | 75 | —22 | 6 |
| 1,7 | 0,9838 | 53 | —16 | 4 |
| 1,8 | 0,9891 | 37 | —12 | |
| 1,9 | 0,9928 | 25 | | |
| 2,0 | 0,9953 | | | |

Подставляя в формулу (4), получим:

$$y \approx 0{,}9523 + 0{,}3 \cdot 0{,}0138 + \frac{0{,}3\,(0{,}3-1)}{2!}(-0{,}0036) + \frac{0{,}3\,(0{,}3-1)\,(0{,}3-2)}{3!} \cdot 0{,}0009 = 0{,}95686.$$

(Табличное значение: $\Phi(1{,}43) = 0{,}9569$, см. «Таблицы функций» Янке и Эмде.)

На практике часто встречается надобность для функции, заданной таблично, подобрать аналитическую формулу, представляющую с некоторой точностью данные табличные значения функции. Такая формула называется *эмпирической*, причем задача построения ее неоднозначна.

При построении эмпирической формулы следует учитывать общие свойства функции. Если из таблицы разностей будет обнаружено, что $n$-е разности функции для равноотстоящих значений аргумента постоянны, то в качестве эмпирической формулы можно взять соответствующую первую интерполяционную формулу Ньютона.

Пример 3. Построить эмпирическую формулу для функции $y$, заданной таблицей

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| $y$ | 5,2 | 8,0 | 10,4 | 12,4 | 14,0 | 15,2 |

Решение. Составляя таблицу разностей (таблица 40), убеждаемся, что вторая разность постоянна. Используя интерполяционную формулу Ньютона в форме (3) и учитывая, что $h = 1$, будем

иметь:

$$y = 5{,}2 + 2{,}8x - \frac{0{,}4}{2} x(x-1)$$

или

$$y = 5{,}2 + 3x - 0{,}2x^2.$$

Пример 4. Найти сумму квадратов

$$S_n = 1^2 + 2^2 + \ldots + n^2$$

натуральных чисел от 1 до $n$.

Решение. Очевидно, имеем:

$$\Delta S_n = S_{n+1} - S_n = (n+1)^2.$$

Отсюда

$$\Delta^2 S_n = 2n + 3, \qquad \Delta^3 S_n = 2$$

и, следовательно, $S_n$ можно искать в виде полинома третьей степени относительно $n$.

Для определения разностей

$$\Delta S_1, \quad \Delta^2 S_1$$

нужно вычислить три значения $S_1$, $S_2$ и $S_3$. Имеем:

$$S_1 = 1,$$
$$S_2 = S_1 + 2^2 = 1 + 4 = 5,$$
$$S_3 = S_2 + 3^2 = 5 + 9 = 14.$$

Отсюда

$$\Delta S_1 = 5 - 1 = 4,$$
$$\Delta S_2 = 14 - 5 = 9,$$
$$\Delta^2 S_1 = 9 - 4 = 5,$$

причем

$$\Delta^3 S_1 = 2.$$

Таблица 40

**Конечные разности функции $y$**

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ |
|---|---|---|---|
| 0 | 5,2 | 2,8 | −0,4 |
| 1 | 8,0 | 2,4 | −0,4 |
| 2 | 10,4 | 2,0 | −0,4 |
| 3 | 12,4 | 1,6 | −0,4 |
| 4 | 14,0 | 1,2 | |
| 5 | 15,2 | | |

Применяя первую интерполяционную формулу Ньютона и учитывая, что

$$q = \frac{n-1}{1} = n - 1,$$

получим:

$$S_n = 1 + 4(n-1) + \frac{5(n-1)(n-2)}{2} + \frac{2(n-1)(n-2)(n-3)}{6}$$

или

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

## § 6. Вторая интерполяционная формула Ньютона

Первая интерполяционная формула Ньютона практически неудобна для интерполирования функции вблизи конца таблицы. В этом случае обычно применяется *вторая интерполяционная формула Ньютона*. Выводом этой формулы мы и займемся.

Пусть имеем систему значений функции

$$y_i = y(x_i) \quad (i = 0,\ 1,\ 2,\ \ldots,\ n)$$

для равноотстоящих значений аргумента

$$x_i = x_0 + ih.$$

Построим интерполирующий полином следующего вида:

$$P_n(x) = a_0 + a_1(x - x_n) + a_2(x - x_n)(x - x_{n-1}) + \\ + a_3(x - x_n)(x - x_{n-1})(x - x_{n-2}) + \ldots \\ \ldots + a_n(x - x_n)(x - x_{n-1}) \ldots (x - x_1),$$

или, используя обобщенную степень, получаем:

$$P_n(x) = a_0 + a_1(x - x_n)^{[1]} + a_2(x - x_{n-1})^{[2]} + \\ + a_3(x - x_{n-2})^{[3]} + \ldots + a_n(x - x_1)^{[n]}. \qquad (1)$$

Наша задача состоит в определении коэффициентов $a_0$, $a_1$, $a_2$, $a_3$, ..., $a_n$ таким образом, чтобы были выполнены равенства

$$P_n(x_i) = y_i \qquad (i = 0,\ 1,\ 2,\ \ldots,\ n).$$

Для этого необходимо и достаточно, чтобы

$$\Delta^i P_n(x_{n-i}) = \Delta^i y_{n-i} \qquad (i = 0,\ 1,\ \ldots,\ n). \qquad (2)$$

Положим $x = x_n$ в формуле (1). Тогда будем иметь:

$$P_n(x_n) = y_n = a_0,$$

следовательно,

$$a_0 = y_n.$$

Далее, берем от левой и правой частей формулы (1) конечные разности первого порядка

$$\Delta P_n(x) = a_1 \cdot 1h + a_2 \cdot 2h(x - x_{n-1})^{[1]} + \\ + a_3 \cdot 3h(x - x_{n-2})^{[2]} + \ldots + a_n nh(x - x_1)^{[n-1]}.$$

Отсюда, полагая $x = x_{n-1}$ и учитывая соотношения (2), будем иметь:

$$\Delta P_n(x_{n-1}) = \Delta y_{n-1} = a_1 h.$$

Следовательно,

$$a_1 = \frac{\Delta y_{n-1}}{h}.$$

Аналогично составив вторую разность от $P_n(x)$, получим:

$$\Delta^2 P_n(x) = a_2 2!\, h^2 + a_3 3\cdot 2h^2 (x - x_{n-2})^{[1]} + \ldots$$
$$\ldots + a_n n(n-1)\, h^2 (x - x_1)^{[n-2]}.$$

Полагая $x = x_{n-2}$, находим:

$$\Delta^2 P_n(x_{n-2}) = \Delta^2 y_{n-2} = a_2 2!\, h^2$$

и, таким образом,

$$a_2 = \frac{\Delta^2 y_{n-2}}{2!\, h^2}.$$

Характер закономерности коэффициентов $a_i$ достаточно ясен. Применяя метод математической индукции, можно строго доказать, что

$$a_i = \frac{\Delta^i y_{n-i}}{i!\, h^i} \qquad (i = 0,\ 1,\ 2,\ \ldots,\ n). \tag{3}$$

Подставляя эти значения в формулу (1), будем иметь окончательно:

$$P_n(x) = y_n + \frac{\Delta y_{n-1}}{1!\, h}(x - x_n) + \frac{\Delta^2 y_{n-2}}{2!\, h^2}(x - x_n)(x - x_{n-1}) +$$
$$+ \frac{\Delta^3 y_{n-3}}{3!\, h^3}(x - x_n)(x - x_{n-1})(x - x_{n-2}) + \ldots$$
$$\ldots + \frac{\Delta^n y_0}{n!\, h^n}(x - x_n)\ldots(x - x_1). \tag{4}$$

Формула (4) носит название *второй интерполяционной формулы Ньютона.*

Введем более удобную запись формулы (4). Пусть

$$q = \frac{x - x_n}{h},$$

тогда

$$\frac{x - x_{n-1}}{h} = \frac{x - x_n + h}{h} = q + 1,$$
$$\frac{x - x_{n-2}}{h} = q + 2 \text{ и т. д.}$$

Подставив эти значения в формулу (4), получим:

$$P_n(x) = y_n + q\,\Delta y_{n-1} + \frac{q(q+1)}{2!}\Delta^2 y_{n-2} + \frac{q(q+1)(q+2)}{3!}\Delta^3 y_{n-3} + \ldots$$
$$\ldots + \frac{q(q+1)\ldots(q+n-1)}{n!}\Delta^n y_0. \tag{4'}$$

Это и есть обычный вид *второй интерполяционной формулы Ньютона.* Для приближенного вычисления значений функции $y$ полагают:

$$y = P_n(x).$$

Пример 1. Дана таблица значений $y = \lg x$ семизначных логарифмов

| $x$ | $y$ |
|---|---|
| 1000 | 3,0000000 |
| 1010 | 3,0043214 |
| 1020 | 3,0086002 |
| 1030 | 3,0128372 |
| 1040 | 3,0170333 |
| 1050 | 3,0211893 |

Найти $\lg 1044$.

Решение. Составляем таблицу разностей (таблица 41).

Таблица 41

**Конечные разности функции $y = \lg x$**

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ | $\Delta^3 y$ |
|---|---|---|---|---|
| 1000 | 3,0000000 | 43 214 | —426 | 2 |
| 1010 | 3,0043214 | 42 788 | —418 | 9 |
| 1020 | 3,0086002 | 42 370 | —409 | 8 |
| 1030 | 3,0128372 | 41 961 | —401 | |
| 1040 | 3,0170333 | 41 560 | | |
| 1050 | 3,0211893 | | | |

Примем

$$x_n = 1050,$$

тогда

$$q = \frac{x - x_n}{h} = \frac{1044 - 1050}{10} = -0{,}6.$$

Используя подчеркнутые разности, в силу формулы (4′) будем иметь:

$$\lg 1044 = 3{,}021\,1893 + (-0{,}6)\cdot 0{,}0041560 + \frac{(-0{,}6)\cdot(-0{,}6+1)}{2}\cdot 0{,}0000401 + \frac{(-0{,}6)\cdot(-0{,}6+1)\cdot(-0{,}6+2)}{6}\cdot 0{,}0000008 = 3{,}0187005.$$

В полученном результате все знаки верные.

Как первая, так и вторая интерполяционные формулы Ньютона могут быть использованы для экстраполирования функции, т. е. для нахождения значений функции $y$ для значений аргументов $x$, лежащих вне пределов таблицы. Если $x < x_0$ и $x$ близко к $x_0$, то

выгодно применять первую интерполяционную формулу Ньютона, причем тогда

$$q=\frac{x-x_0}{h}<0.$$

Если же $x>x_n$ и $x$ близко к $x_n$, то удобнее пользоваться второй интерполяционной формулой Ньютона, причем

$$q=\frac{x-x_n}{h}>0.$$

Таким образом, первая интерполяционная формула Ньютона обычно используется для *интерполирования вперед* и *экстраполирования назад*, а вторая интерполяционная формула Ньютона, наоборот,— для *интерполирования назад* и *экстраполирования вперед*.

Заметим, что операция экстраполирования, вообще говоря, менее точна, чем операция интерполирования в узком смысле слова.

Таблица 42

**Таблица разностей функции $y=\sin x$**

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ | $\Delta^3 y$ |
|---|---|---|---|---|
| 15° | 0,2588 | 832 | —26 | —6 |
| 20° | 0,3420 | 806 | —32 | —6 |
| 25° | 0,4226 | 774 | —38 | —6 |
| 30° | 0,5000 | 736 | —44 | —5 |
| 35° | 0,5736 | 692 | —49 | —5 |
| 40° | 0,6428 | 643 | —54 | —3 |
| 45° | 0,7071 | 589 | —57 | |
| 50° | 0,7660 | 532 | | |
| 55° | 0,8192 | | | |

Пример 2. Имея таблицу значений функции $y=\sin x$ в пределах от $x=15°$ до $x=55°$ с шагом $h=5°$ (таблица 42), найти $\sin 14°$ и $\sin 56°$.

Решение. Составим таблицу разностей (таблица 42). Мы видим, что третьи разности функции $y$ практически постоянны, и поэтому можно ограничиться ими.

Для вычисления $\sin 14°$ примем:

$$x_0=15° \quad \text{и} \quad x=14°;$$

отсюда

$$q=\frac{14°-15°}{5°}=-0{,}2.$$

Применяя первую интерполяционную формулу Ньютона и используя подчеркнутые разности, будем иметь:

$$\sin 14°=0{,}2588+(-0{,}2)\cdot 0{,}0832+\frac{(-0{,}2)(-1{,}2)}{2!}(-0{,}0026)+$$
$$+\frac{(-0{,}2)(-1{,}2)(-2{,}2)}{3!}(-0{,}0006)=0{,}2419.$$

По таблицам $\sin 14°=0{,}24192$.

Для отыскания $\sin 56°$ положим:

$$x_n=55° \quad \text{и} \quad x=56°;$$

отсюда

$$q=\frac{56°-55°}{5°}=0{,}2.$$

Применяя вторую интерполяционную формулу Ньютона и используя дважды подчеркнутые разности, будем иметь:

$$\sin 56^\circ = 0{,}8192 + 0{,}2\cdot 0{,}0532 + \frac{0{,}2\cdot 1{,}2}{2!}(-0{,}0057) + \\ + \frac{0{,}2\cdot 1{,}2\cdot 2{,}2}{3!}(-0{,}0003) = 0{,}8291.$$

По таблицам $\sin 56^\circ = 0{,}82904$.

## § 7. Таблица центральных разностей

При построении интерполяционных формул Ньютона используются лишь значения функции, лежащие по одну сторону от выбранного начального значения, т. е. эти формулы носят односторонний характер.

Таблица 43

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ | $\Delta^3 y$ | $\Delta^4 y$ | $\Delta^5 y$ | $\Delta^6 y$ |
|---|---|---|---|---|---|---|---|
| | | | | | | | |
| $x_{-4}$ | $y_{-4}$ | | | | | | |
| | | $\Delta y_{-4}$ | | | | | |
| $x_{-3}$ | $y_{-3}$ | | $\Delta^2 y_{-4}$ | | | | |
| | | $\Delta y_{-3}$ | | $\Delta^3 y_{-4}$ | | | |
| $x_{-2}$ | $y_{-2}$ | | $\Delta^2 y_{-3}$ | | $\Delta^4 y_{-4}$ | | |
| | | $\Delta y_{-2}$ | | $\Delta^3 y_{-3}$ | | $\Delta^5 y_{-4}$ | |
| $x_{-1}$ | $y_{-1}$ | | $\Delta^2 y_{-2}$ | | $\Delta^4 y_{-3}$ | | $\Delta^6 y_{-4}$ |
| | | $\Delta y_{-1}$ | | $\Delta^3 y_{-2}$ | | $\Delta^5 y_{-3}$ | |
| $x_0$ | $y_0$ | | $\Delta^2 y_{-1}$ | | $\Delta^4 y_{-2}$ | | $\Delta^6 y_{-3}$ |
| | | $\Delta y_0$ | | $\Delta^3 y_{-1}$ | | $\Delta^5 y_{-2}$ | |
| $x_1$ | $y_1$ | | $\Delta^2 y_0$ | | $\Delta^4 y_{-1}$ | | $\Delta^6 y_{-2}$ |
| | | $\Delta y_1$ | | $\Delta^3 y_0$ | | $\Delta^5 y_{-1}$ | |
| $x_2$ | $y_2$ | | $\Delta^2 y_1$ | | $\Delta^4 y_0$ | | |
| | | $\Delta y_2$ | | $\Delta^3 y_1$ | | | |
| $x_3$ | $y_3$ | | $\Delta^2 y_2$ | | | | |
| | | $\Delta y_3$ | | | | | |
| $x_4$ | $y_4$ | | | | | | |
| | | | | | | | |

Во многих случаях оказываются полезными интерполяционные формулы, содержащие как последующие, так и предшествующие значения функции по отношению к начальному значению ее. Наиболее употребительными из них являются те, которые содержат разности, расположенные в горизонтальной строке диагональной таблицы разностей данной функции, соответствующей начальным значениям $x_0$ и $y_0$, или в строках, непосредственно примыкающих к ней. Эти разности $\Delta y_{-1}$, $\Delta y_0$, $\Delta^2 y_{-1}$, ... называются *центральными разностями* (таблица 43), где

$$x_i = x_0 + ih \qquad (i = 0, \pm 1, \pm 2, \ldots), \qquad y_i = f(x_i),$$
$$\Delta y_i = y_{i+1} - y_i; \qquad \Delta^2 y_i = \Delta y_{i+1} - \Delta y_i \quad \text{и т. д.}$$

Соответствующие интерполяционные формулы носят название *интерполяционных формул с центральными разностями*. К числу их относятся формулы Гаусса, Стирлинга и Бесселя [3].

## § 8. Интерполяционные формулы Гаусса

Выведем сначала интерполяционные формулы Гаусса.

Пусть имеется $2n+1$ равноотстоящих узлов интерполирования

$$x_{-n},\ x_{-(n-1)},\ \ldots,\ x_{-1},\ x_0,\ x_1,\ \ldots,\ x_{n-1},\ x_n,$$

где

$$\Delta x_i = x_{i+1} - x_i = h = \text{const} \qquad (i = -n,\ -(n-1),\ \ldots,\ n-1),$$

и для функции $y = f(x)$ известны ее значения в этих узлах

$$y_i = f(x_i) \qquad (i = 0,\ \pm 1,\ \ldots,\ \pm n).$$

Требуется построить полином $P(x)$ степени не выше $2n$ такой, что

$$P(x_i) = y_i \quad \text{при } i = 0,\ \pm 1,\ \ldots,\ \pm n.$$

Из последнего условия вытекает, что

$$\Delta^k P(x_i) = \Delta^k y_i \tag{1}$$

для всех соответствующих значений $i$ и $k$.

Будем искать этот полином в виде

$$\begin{aligned} P(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \\ + a_3(x - x_{-1})(x - x_0)(x - x_1) + a_4(x - x_{-1})(x - x_0)(x - x_1) \times \\ \times (x - x_2) + a_5(x - x_{-2})(x - x_{-1})(x - x_0)(x - x_1)(x - x_2) + \ldots \\ \ldots + a_{2n-1}(x - x_{-(n-1)}) \ldots (x - x_{-1})(x - x_0)(x - x_1) \ldots \\ \ldots (x - x_{n-1}) + a_{2n}(x - x_{-(n-1)}) \ldots (x - x_{-1})(x - x_0)(x - x_1) \ldots \\ \ldots (x - x_{n-1})(x - x_n). \end{aligned} \tag{2}$$

Вводя обобщенные степени, получим:

$$P(x)=a_0+a_1(x-x_0)^{[1]}+a_2(x-x_0)^{[2]}+a_3(x-x_{-1})^{[3]}+ \\ +a_4(x-x_{-1})^{[4]}+\ldots+a_{2n-1}(x-x_{-(n-1)})^{[2n-1]}+ \\ +a_{2n}(x-x_{-(n-1)})^{[2n]}. \qquad (3)$$

Применяя для вычисления коэффициентов $a_i$ $(i=0,\ 1,\ \ldots,\ 2n)$ тот же способ, что и при выводе интерполяционных формул Ньютона, и учитывая формулу (1), последовательно находим:

$$a_0=y_0,\quad a_1=\frac{\Delta y_0}{1!\,h},\quad a_2=\frac{\Delta^2 y_{-1}}{2!\,h^2},\quad a_3=\frac{\Delta^3 y_{-1}}{3!\,h^3},$$

$$a_4=\frac{\Delta^4 y_{-2}}{4!\,h^4},\ \ldots,\ a_{2n-1}=\frac{\Delta^{2n-1}y_{-(n-1)}}{(2n-1)!\,h^{2n-1}},\quad a_{2n}=\frac{\Delta^{2n}y_{-n}}{(2n)!\,h^{2n}}.$$

Далее, введя переменную

$$q=\frac{x-x_0}{h}$$

и сделав соответствующую замену в формуле (3), получим *первую интерполяционную формулу Гаусса*

$$P(x)=y_0+q\,\Delta y_0+\frac{q(q-1)}{2!}\Delta^2 y_{-1}+ \\ +\frac{(q+1)\,q\,(q-1)}{3!}\Delta^3 y_{-1}+\frac{(q+1)\,q\,(q-1)\,(q-2)}{4!}\Delta^4 y_{-2}+ \\ +\frac{(q+2)\,(q+1)\,q\,(q-1)\,(q-2)}{5!}\Delta^5 y_{-2}+\ldots \\ \ldots+\frac{(q+n-1)\ldots(q-n+1)}{(2n-1)!}\Delta^{2n-1}y_{-(n-1)}+\frac{(q+n-1)\ldots(q-n)}{(2n)!}\Delta^{2n}y_{-n} \qquad (4)$$

или, короче,

$$P(x)=y_0+q\,\Delta y_0+\frac{q^{[2]}}{2!}\Delta^2 y_{-1}+\frac{(q+1)^{[3]}}{3!}\Delta^3 y_{-1}+ \\ +\frac{(q+1)^{[4]}}{4!}\Delta^4 y_{-2}+\ldots+\frac{(q+n-1)^{[2n-1]}}{(2n-1)!}\Delta^{2n-1}y_{-(n-1)}+ \\ +\frac{(q+n-1)^{[2n]}}{(2n)!}\Delta^{2n}y_{-n}, \qquad (4')$$

где $x=x_0+qh$ и $q^{[m]}=q\,(q-1)\ldots[q-(m-1)]$.

Первая интерполяционная формула Гаусса содержит центральные разности

$$\Delta y_0,\quad \Delta^2 y_{-1},\quad \Delta^3 y_{-1},\quad \Delta^4 y_{-2},\quad \Delta^5 y_{-2},\quad \Delta^6 y_{-3},\ \ldots$$

(см. таблицу 43, где эти разности образуют нижнюю ломаную строку по ходу стрелки). Аналогично можно получить *вторую интерполяционную формулу Гаусса*, содержащую центральные разности

$$\Delta y_{-1},\quad \Delta^2 y_{-1},\quad \Delta^3 y_{-2},\quad \Delta^4 y_{-2},\quad \Delta^5 y_{-3},\quad \Delta^6 y_{-3},\ \ldots$$

(в таблице 43 эти разности образуют верхнюю ломаную строку по ходу стрелки).

Вторая интерполяционная формула Гаусса имеет вид

$$P(x)=y_0+q\,\Delta y_{-1}+\frac{(q+1)\,q}{2!}\,\Delta^2 y_{-1}+\frac{(q+1)\,q\,(q-1)}{3!}\,\Delta^3 y_{-2}+$$
$$+\frac{(q+2)\,(q+1)\,q\,(q-1)}{4!}\,\Delta^4 y_{-2}+\ldots$$
$$\ldots+\frac{(q+n-1)\ldots(q-n+1)}{(2n-1)!}\,\Delta^{2n-1}y_{-n}+$$
$$+\frac{(q+n)\,(q+n-1)\ldots(q-n+1)}{(2n)!}\,\Delta^{2n}y_{-n} \qquad (5)$$

или, в сокращенных обозначениях,

$$P(x)=y_0+q\,\Delta y_{-1}+\frac{(q+1)^{[2]}}{2!}\,\Delta^2 y_{-1}+$$
$$+\frac{(q+1)^{[3]}}{3!}\,\Delta^3 y_{-2}+\frac{(q+2)^{[4]}}{4!}\,\Delta^4 y_{-2}+\ldots$$
$$\ldots+\frac{(q+n-1)^{[2n-1]}}{(2n-1)!}\,\Delta^{2n-1}y_{-n}+\frac{(q+n)^{[2n]}}{(2n)!}\,\Delta^{2n}y_{-n}, \qquad (5')$$

где

$$x=x_0+qh.$$

## § 9. Интерполяционная формула Стирлинга

Взяв среднее арифметическое первой и второй интерполяционных формул Гаусса (4) и (5) (§ 8), получим *формулу Стирлинга*

$$P(x)=y_0+q\cdot\frac{\Delta y_{-1}+\Delta y_0}{2}+\frac{q^2}{2}\,\Delta^2 y_{-1}+\frac{q\,(q^2-1^2)}{3!}\cdot\frac{\Delta^3 y_{-2}+\Delta^3 y_{-1}}{2}+$$
$$+\frac{q^2\,(q^2-1^2)}{4!}\,\Delta^4 y_{-2}+\frac{q\,(q^2-1^2)\,(q^2-2^2)}{5!}\cdot\frac{\Delta^5 y_{-3}+\Delta^5 y_{-2}}{2}+$$
$$+\frac{q^2\,(q^2-1^2)\,(q^2-2^2)}{6!}\,\Delta^6 y_{-3}+\ldots+$$
$$+\frac{q\,(q^2-1^2)\,(q^2-2^2)\,(q^2-3^2)\ldots[q^2-(n-1)^2]}{(2n-1)!}\times$$
$$\times\frac{\Delta^{2n-1}y_{-n}+\Delta^{2n-1}y_{-(n-1)}}{2}+\frac{q^2\,(q^2-1^2)\,(q^2-2^2)\ldots[q^2-(n-1)^2]}{(2n)!}\,\Delta^{2n}y_{-n},$$

где

$$q=\frac{x-x_0}{h}.$$

Легко видеть, что

$$P(x_i)=y_i \quad \text{при} \quad i=0,\ \pm 1,\ \ldots,\ \pm n.$$

## § 10. Интерполяционная формула Бесселя

Кроме формулы *Стирлинга*, часто употребляется *формула Бесселя*. Для вывода этой формулы воспользуемся второй интерполяционной формулой Гаусса (5) (см. § 8).

Возьмем $2n+2$ равноотстоящих узлов интерполирования

$$x_{-n},\ x_{-(n-1)},\ \ldots,\ x_0,\ \ldots,\ x_{n-1},\ x_n,\ x_{n+1}$$

с шагом $h$, и пусть

$$y_i = f(x_i) \qquad (i = -n,\ \ldots,\ n+1)$$

— заданные значения функции $y = f(x)$.

Если выбрать за начальные значения $x = x_0$ и $y = y_0$, то, используя узлы $x_k$ $(k = 0,\ \pm 1,\ \ldots,\ \pm n)$, будем иметь:

$$P(x) = y_0 + q\,\Delta y_{-1} + \frac{(q+1)\,q}{2!}\,\Delta^2 y_{-1} + \\ + \frac{(q+1)\,q\,(q-1)}{3!}\,\Delta^3 y_{-2} + \frac{(q+2)\,(q+1)\,q\,(q-1)}{4!}\,\Delta^4 y_{-2} + \ldots \\ \ldots + \frac{(q+n-1)\ldots(q-n+1)}{(2n-1)!}\,\Delta^{2n-1} y_{-n} + \\ + \frac{(q+n)\,(q+n-1)\ldots(q-n+1)}{(2n)!}\,\Delta^{2n} y_{-n}. \qquad (1)$$

Примем теперь за начальные значения $x = x_1$ и $y = y_1$ и используем узлы $x_{1+k}$ $(k = 0,\ \pm 1,\ \ldots,\ \pm n)$. Тогда

$$\frac{x - x_1}{h} = \frac{x - x_0 - h}{h} = q - 1,$$

причем соответственно индексы всех разностей в правой части формулы (1) возрастут на единицу. Заменив в правой части формулы (1) $q$ на $q-1$ и увеличив индексы всех разностей на 1, получим вспомогательную интерполяционную формулу

$$P(x) = y_1 + (q-1)\,\Delta y_0 + \frac{q(q-1)}{2!}\,\Delta^2 y_0 + \frac{q\,(q-1)\,(q-2)}{3!}\,\Delta^3 y_{-1} + \\ + \frac{(q+1)\,q\,(q-1)\,(q-2)}{4!}\,\Delta^4 y_{-1} + \frac{(q+1)\,q\,(q-1)\,(q-2)\,(q-3)}{5!}\,\Delta^5 y_{-2} + \ldots \\ \ldots + \frac{(q+n-2)\ldots(q-n)}{(2n-1)!}\,\Delta^{2n-1} y_{-(n-1)} + \\ + \frac{(q+n-1)\ldots(q-n)}{(2n)!}\,\Delta^{2n} y_{-(n-1)}. \qquad (2)$$

Взяв среднее арифметическое формул (1) и (2), после несложных преобразований получим *интерполяционную формулу Бесселя*

$$P(x) = \frac{y_0 + y_1}{2} + \left(q - \frac{1}{2}\right)\Delta y_0 + \frac{q\,(q-1)}{2} \cdot \frac{\Delta^2 y_{-1} + \Delta^2 y_0}{2} + \\ + \frac{\left(q - \frac{1}{2}\right) q\,(q-1)}{3!}\,\Delta^3 y_{-1} + \frac{q\,(q-1)\,(q+1)\,(q-2)}{4!} \cdot \frac{\Delta^4 y_{-2} + \Delta^4 y_{-1}}{2} + \\ + \frac{\left(q - \frac{1}{2}\right) q\,(q-1)\,(q+1)\,(q-2)}{5!}\,\Delta^5 y_{-2} +$$

$$+\frac{q(q-1)(q+1)(q-2)(q+2)(q-3)}{6!}\cdot\frac{\Delta^6 y_{-3}+\Delta^6 y_{-2}}{2}+\ldots$$

$$\ldots+\frac{q(q-1)(q+1)(q-2)(q+2)\ldots(q-n)(q+n-1)}{(2n)!}\cdot\frac{\Delta^{2n}y_{-n}+\Delta^{2n}y_{-n+1}}{2}+$$

$$+\frac{\left(q-\frac{1}{2}\right)q(q-1)(q+1)(q-2)(q+2)\ldots(q-n)(q+n-1)}{(2n+1)!}\Delta^{2n+1}y_{-n}, \quad (3)$$

где

$$q=\frac{x-x_0}{h}.$$

Интерполяционная формула Бесселя (3), как следует из способа получения ее, представляет собой полином, совпадающий с данной функцией $y=f(x)$ в $2n+2$ точках

$$x_{-n},\ x_{-(n-1)},\ \ldots,\ x_n,\ x_{n+1}.$$

В частном случае, при $n=1$, пренебрегая разностью $\Delta^3 y_{-1}$, имеем *формулу квадратичной интерполяции по Бесселю*

$$P(x)=\frac{y_0+y_0+\Delta y_0}{2}+\left(q-\frac{1}{2}\right)\Delta y_0+$$
$$+\frac{q(q-1)}{2}\cdot\frac{\Delta y_0-\Delta y_{-1}+\Delta y_1-\Delta y_0}{2}$$

или

$$P(x)=y_0+q\,\Delta y_0-q_1(\Delta y_1-\Delta y_{-1}),$$

где

$$q_1=\frac{q(1-q)}{4}.$$

В формуле Бесселя все члены, содержащие разности нечетного порядка, имеют множитель $q-\frac{1}{2}$; поэтому при $q=\frac{1}{2}$ формула (3) значительно упрощается:

$$P\left(\frac{x_0+x_1}{2}\right)=\frac{y_0+y_1}{2}-\frac{1}{8}\,\frac{\Delta^2 y_{-1}+\Delta^2 y_0}{2}+$$
$$+\frac{3}{128}\,\frac{\Delta^4 y_{-2}+\Delta^4 y_{-1}}{2}-\frac{5}{1024}\,\frac{\Delta^6 y_{-3}+\Delta^6 y_{-2}}{2}+\ldots$$
$$\ldots+(-1)^n\frac{[1\cdot3\cdot5\ldots(2n-1)]^2}{2^{2n}(2n)!}\,\frac{\Delta^{2n}y_{-n}+\Delta^{2n}y_{-n+1}}{2}.$$

Этот специальный случай формулы Бесселя называется *формулой интерполирования на середину*. Если в формуле Бесселя (3) сделать замену переменной по формуле $q-\frac{1}{2}=p$, то она приобретает более

симметричный вид

$$P(x)=\frac{y_0+y_1}{2!}+p\,\Delta y_0+\frac{\left(p^2-\frac{1}{4}\right)}{2}\cdot\frac{\Delta^2 y_{-1}+\Delta^2 y_0}{2}+$$
$$+\frac{p\left(p^2-\frac{1}{4}\right)}{3!}\Delta^3 y_{-1}+\frac{\left(p^2-\frac{1}{4}\right)\left(p^2-\frac{9}{4}\right)}{4!}\cdot\frac{\Delta^4 y_{-2}+\Delta^4 y_{-1}}{2}+$$
$$+\frac{p\left(p^2-\frac{1}{4}\right)\left(p^2-\frac{9}{4}\right)}{5!}\Delta^5 y_{-2}+\frac{\left(p^2-\frac{1}{4}\right)\left(p^2-\frac{9}{4}\right)\left(p^2-\frac{25}{4}\right)}{6!}\times$$
$$\times\frac{\Delta^6 y_{-3}+\Delta^6 y_{-2}}{2}+\ldots+\frac{\left(p^2-\frac{1}{4}\right)\left(p^2-\frac{9}{4}\right)\ldots\left[p^2-\frac{(2n-1)^2}{4}\right]}{(2n)!}\times$$
$$\times\frac{\Delta^{2n} y_{-n}+\Delta^{2n} y_{-n+1}}{2}+\frac{p\left(p^2-\frac{1}{4}\right)\left(p^2-\frac{9}{4}\right)\ldots\left[p^2-\frac{(2n-1)^2}{4}\right]}{(2n+1)!}\times$$
$$\times\Delta^{2n+1} y_{-n+1}, \qquad (3')$$

где $p=\frac{1}{h}\left(x-\frac{x_0+x_1}{2}\right)$.

## § 11. Общая характеристика интерполяционных формул с постоянным шагом

Давая общую характеристику интерполяционным формулам, отметим следующее: при построении интерполяционных формул Ньютона в качестве начального значения $x_0$ выбирается первый или последний узел интерполирования; для центральных же формул интерполирования начальный узел является средним. Приведенная ниже схема (таблица 44) показывает используемые разности в основных интерполяционных формулах, причем для удобства обозрения во второй интерполяционной формуле Ньютона изменена нумерация индексов.

Более детальное рассмотрение интерполяционных формул показывает, что при $|q|\leqslant 0{,}25$ целесообразно применять формулу Стирлинга, а при $0{,}25\leqslant q\leqslant 0{,}75$ — формулу Бесселя. Первую и вторую интерполяционные формулы Ньютона выгодно применять тогда, когда интерполирование производится в начале или соответственно в конце таблицы и нужных центральных разностей не хватает [4].

Пример 1. Значения интеграла вероятностей [3]

$$\Phi(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-x^2}\,dx$$

даны в таблице 45. Найти $\Phi(0{,}5437)$.

Таблица 44

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ | $\Delta^3 y$ | $\Delta^4 y$ | Примечание |
|---|---|---|---|---|---|---|
| | | | | | | 2-я формула Ньютона |
| $x_{-2}$ | $y_{-2}$ | | $\Delta^2 y_{-3}$ | | $\Delta^4 y_{-4}$ | |
| | | $\Delta y_{-2}$ | | $\Delta^3 y_{-3}$ | | |
| $x_{-1}$ | $y_{-1}$ | | $\Delta^2 y_{-2}$ | | $\Delta^4 y_{-3}$ | |
| | | $\Delta y_{-1}$ | | $\Delta^3 y_{-2}$ | | |
| $x_0$ | $y_0$ | | $\Delta^2 y_{-1}$ | | $\Delta^4 y_{-2}$ | Формула Стирлинга |
| | | $\Delta y_0$ | | $\Delta^3 y_{-1}$ | | Формула Бесселя |
| $x_1$ | $y_1$ | | $\Delta^2 y_0$ | | $\Delta^4 y_{-1}$ | |
| | | $\Delta y_1$ | | $\Delta^3 y_0$ | | |
| $x_2$ | $y_2$ | | $\Delta^2 y_1$ | | $\Delta^4 y_0$ | |
| | | $\Delta y_2$ | | $\Delta^3 y_1$ | | 1-я формула Ньютона |
| $x_3$ | $y_3$ | | $\Delta^2 y_2$ | | $\Delta^4 y_1$ | |
| | | | | | | |

Таблица 45

**Таблица разностей функции $y = \Phi(x)$**

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ | $\Delta^3 y$ |
|---|---|---|---|---|
| 0,51 | 0,5292437 | | | |
| | | 86 550 | | |
| 0,52 | 0,5378987 | | —896 | |
| | | 85 654 | | —7 |
| 0,53 | 0,5464641 | | —903 | |
| | | 84 751 | | —7 |
| 0,54 | 0,5549392 | | —910 | |
| | | 83 841 | | —7 |
| 0,55 | 0,5633233 | | —917 | |
| | | 82 924 | | —6 |
| 0,56 | 0,5716157 | | —923 | |
| | | 82 001 | | |
| 0,57 | 0,5798158 | | | |

Решение. Таблицу 45 дополняем конечными разностями данной функции $y=\Phi(x)$. Принимаем $x_0=0{,}54$ и $x=0{,}5437$, тогда

$$q=\frac{x-x_0}{h}=\frac{0{,}5437-0{,}54}{0{,}01}=0{,}37.$$

Так как $\frac{1}{4}<q<\frac{3}{4}$, то воспользуемся формулой Бесселя (3′). Имеем:

$$p=q-\frac{1}{2}=0{,}37-0{,}50=-0{,}13;$$

отсюда, используя подчеркнутые разности, получаем:

$$\Phi(0{,}5437)=\frac{0{,}5549392+0{,}5633233}{2}+$$
$$+(-0{,}13)\,0{,}0083841+\frac{0{,}0169-0{,}25}{2}\cdot\frac{-0{,}0000910-0{,}0000917}{2}+$$
$$+\frac{-0{,}13\,(0{,}0169-0{,}25)}{6}\cdot(-0{,}0000007)=$$
$$=0{,}55913125-0{,}00108993+0{,}00001065=0{,}5580520.$$

Пример 2. Имея таблицу 46 значений полного эллиптического интеграла

$$K(\alpha)=\int_0^{\frac{\pi}{2}}\frac{dx}{\sqrt{1-\sin^2\alpha\sin^2 x}},$$

найти $K(78°\,30')$.

Таблица 46

**Значения полного эллиптического интеграла $K(\alpha)$**

| $\alpha$ | $K(\alpha)$ | $\Delta K$ | $\Delta^2 K$ | $\Delta^3 K$ | $\Delta^4 K$ | $\Delta^5 K$ | $\Delta^6 K$ |
|---|---|---|---|---|---|---|---|
| 75° | 2,76806 | | | | | | |
| | | 6 461 | | | | | |
| 76° | 2,83267 | | 528 | | | | |
| | | 6 989 | | 84 | | | |
| 77° | 2,90256 | | 612 | | 19 | | |
| | | 7 601 | | 103 | | 13 | |
| 78° | 2,97857 | | 715 | | 32 | | —5 |
| | | 8 316 | | 135 | | 8 | |
| 79° | 3,06173 | | 850 | | 40 | | 18 |
| | | 9 166 | | 175 | | 26 | |
| 80° | 3,15339 | | 1 025 | | 66 | | —1 |
| | | 10 191 | | 241 | | 25 | |
| 81° | 3,25530 | | 1 266 | | 91 | | 43 |
| | | 11 457 | | 332 | | 68 | |
| 82° | 3,36987 | | 1 598 | | 159 | | |
| | | 13 055 | | 491 | | | |
| 83° | 3,50042 | | 2 089 | | | | |
| | | 15 144 | | | | | |
| 84° | 3,65186 | | | | | | |

Решение. Полагаем $x_0=78°$; $h=1°$; $x=78°30'$; отсюда $q=0{,}5$. Если воспользоваться формулой Бесселя для интерполирования на середину, то, ограничиваясь разностями пятого порядка, будем иметь:

$$K(78°30')=2{,}97857+0{,}5\cdot 8316\cdot 10^{-5}-0{,}125\cdot\frac{715+850}{2}\cdot 10^{-5}+$$
$$+0{,}023437\cdot\frac{32+40}{2}\cdot 10^{-5}=2{,}97857-0{,}04158-0{,}000978+$$
$$+0{,}000008=3{,}019180.$$

Для сравнения применим теперь формулу Стирлинга

$$K(78°30')=2{,}97857+0{,}5\,\frac{7601+8316}{2}\cdot 10^{-5}+$$
$$+0{,}125715\cdot 10^{-5}-0{,}0625\cdot\frac{103+135}{2}\cdot 10^{-5}-0{,}0078\cdot 32\cdot 10^{-5}+$$
$$+0{,}0117\cdot\frac{13+8}{2}\cdot 10^{-5}=2{,}97857+0{,}039792+$$
$$+0{,}000894-0{,}000074-0{,}000002+0{,}000001=3{,}019181.$$

## § 12. Интерполяционная формула Лагранжа

Выведенные нами в предыдущих параграфах интерполяционные формулы пригодны лишь в случае равноотстоящих узлов интерполирования. Для произвольно заданных узлов интерполирования пользуются более общей формулой, так называемой *интерполяционной формулой Лагранжа*.

Пусть на отрезке $[a, b]$ даны $n+1$ различных значений аргумента: $x_0, x_1, x_2, \ldots, x_n$ и известны для функции $y=f(x)$ соответствующие значения:

$$f(x_0)=y_0,$$
$$f(x_1)=y_1,\ \ldots,\ f(x_n)=y_n.$$

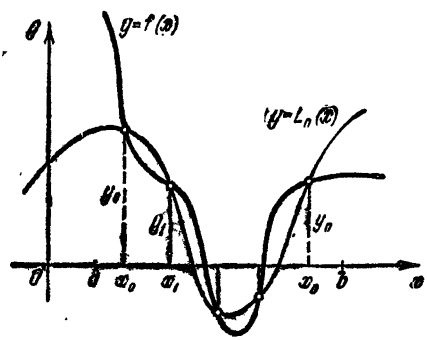

Рис. 62а.

Требуется построить полином $L_n(x)$ степени не выше $n$, имеющий в заданных узлах $x_0, x_1, \ldots, x_n$ те же значения, что и функция $f(x)$, т. е. такой, что

$$L_n(x_i)=y_i \qquad (i=0,\ 1,\ 2,\ \ldots,\ n)$$

(рис. 62а).

Решим сначала частную задачу: построим полином $p_i(x)$ такой, что

$$p_i(x_j)=0 \quad \text{при} \quad j\neq i \quad \text{и} \quad p_i(x_i)=1$$

(рис. 62б).

Короче эти условия можно записать следующим образом:

$$p_i(x_j)=\delta_{ij}=\begin{cases} 1, & \text{если } j=i; \\ 0, & \text{если } j\neq i, \end{cases} \tag{1}$$

где $\delta_{ij}$ — *символ Кронекера.*

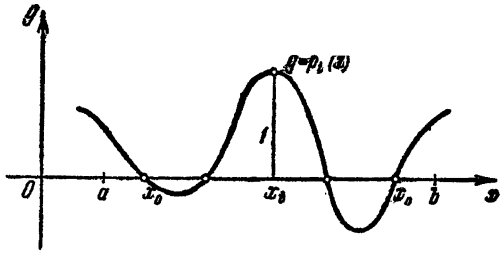


Рис. 62б.

Так как искомый полином обращается в нуль в $n$ точках $x_0$, $x_1$, ..., $x_{i-1}$, $x_{i+1}$, ..., $x_n$, то он имеет вид

$$p_i(x)=C_i(x-x_0)(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n), \tag{2}$$

где $C_i$ — постоянный коэффициент. Полагая $x=x_i$ в формуле (2) и учитывая, что $p_i(x_i)=1$, получим:

$$C_i(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)=1.$$

Отсюда

$$C_i=\frac{1}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)}.$$

Подставив это значение в формулу (2), будем иметь:

$$p_i(x)=\frac{(x-x_0)(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)}. \tag{3}$$

Теперь перейдем к решению общей задачи: к отысканию полинома $L_n(x)$, удовлетворяющего указанным выше условиям $L_n(x_i)=y_i$. Этот полином имеет следующий вид:

$$L_n(x)=\sum_{i=0}^{n} p_i(x)\,y_i. \tag{4}$$

В самом деле, во-первых, очевидно, степень построенного полинома $L_n(x)$ не выше $n$ и, во-вторых, в силу условия (1) имеем:

$$L_n(x_j)=\sum_{i=0}^{n} p_i(x_j)\,y_i=p_j(x_j)\,y_j=y_j \qquad (j=0,\ 1,\ \ldots,\ n).$$

Подставив в формулу (4) значение $p_i(x)$ из (3), получим:

$$L_n(x)=\sum_{i=0}^{n} y_i\,\frac{(x-x_0)(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)}. \qquad (5)$$

Это и есть *интерполяционная формула Лагранжа.*

Докажем *единственность* полинома Лагранжа.

Предположим противное.

Пусть $\tilde{L}_n(x)$ — полином, отличный от $L_n(x)$, степени не выше $n$ и такой, что

$$\tilde{L}_n(x_i)=y_i \qquad (i=0,\ 1,\ \ldots,\ n).$$

Тогда полином

$$Q_n(x)=\tilde{L}_n(x)-L_n(x),$$

степень которого, очевидно, не выше $n$, обращается в нуль в $n+1$ точках $x_0,\ x_1,\ x_2,\ \ldots,\ x_n$, т. е.

$$Q_n(x)\equiv 0.$$

Следовательно,

$$\tilde{L}_n(x)\equiv L_n(x).$$

Отсюда, в частности, следует, что если узлы интерполирования — равноотстоящие, то интерполяционный полином Лагранжа совпадает с соответствующим интерполяционным полиномом Ньютона.

Заметим, вообще, что все построенные выше интерполяционные формулы получаются из интерполяционной формулы Лагранжа при соответствующем выборе узлов.

Формуле (5) Лагранжа можно придать более сжатый вид. Для этого введем обозначение

$$\Pi_{n+1}(x)=(x-x_0)(x-x_1)\ldots(x-x_n). \qquad (6)$$

Дифференцируя по $x$ это произведение, получим:

$$\Pi'_{n+1}(x)=\sum_{j=0}^{n}(x-x_0)(x-x_1)\ldots(x-x_{j-1})(x-x_{j+1})\ldots(x-x_n).$$

Полагая $x=x_i$ $(i=0,\ 1,\ 2,\ \ldots,\ n)$, будем иметь:

$$\Pi'_{n+1}(x_i)=(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_0). \qquad (7)$$

Внося выражения (6) и (7) в формулу (5), получим:

$$L_n(x)=\Pi_{n+1}(x)\sum_{i=0}^{n}\frac{y_i}{\Pi'_{n+1}(x_i)\,(x-x_i)}. \qquad (5')$$

Следует отметить, что формула Лагранжа в отличие от предыдущих интерполяционных формул содержит явно $y_i$, что бывает иногда важно.

Рассмотрим два частных случая интерполяционного полинома Лагранжа.

При $n=1$ мы имеем две точки, и формула Лагранжа представляет в этом случае уравнение прямой $y=L_1(x)$, проходящей через две заданные точки:

$$y=\frac{x-b}{a-b}y_0+\frac{x-a}{b-a}y_1,$$

где $a$, $b$ — абсциссы этих точек.

При $n=2$ получим уравнение параболы $y=L_2(x)$, проходящей через три точки:

$$y=\frac{(x-b)(x-c)}{(a-b)(a-c)}y_0+\frac{(x-a)(x-c)}{(b-a)(b-c)}y_1+\frac{(x-a)(x-b)}{(c-a)(c-b)}y_2,$$

где $a$, $b$, $c$ — абсциссы данных точек.

Пример 1. Для функции $y=\sin\pi x$ построить интерполяционный полином Лагранжа, выбрав узлы

$$x_0=0,\quad x_1=\frac{1}{6},\quad x_2=\frac{1}{2}.$$

Решение. Вычисляем соответствующие значения функции:

$$y_0=0,\quad y_1=\sin\frac{\pi}{6}=\frac{1}{2},\quad y_2=\sin\frac{\pi}{2}=1.$$

Применяя формулу (5), получим:

$$L_2(x)=\frac{\left(x-\frac{1}{6}\right)\left(x-\frac{1}{2}\right)}{\left(-\frac{1}{6}\right)\left(-\frac{1}{2}\right)}\cdot 0+\frac{x\left(x-\frac{1}{2}\right)}{\frac{1}{6}\left(\frac{1}{6}-\frac{1}{2}\right)}\cdot\frac{1}{2}+\frac{x\left(x-\frac{1}{6}\right)}{\frac{1}{2}\left(\frac{1}{2}-\frac{1}{6}\right)}\cdot 1$$

или

$$L_2(x)=\frac{7}{2}x-3x^2.$$

Пример 2. Дана таблица значений функции $y=f(x)$ [3]:

| $x$ | $y$ |
|---|---|
| 321,0 | 2,50651 |
| 322,8 | 2,50893 |
| 324,2 | 2,51081 |
| 325,0 | 2,51188 |

Вычислить значение $f(323{,}5)$.

Решение. Положим $x=323{,}5$; $n=3$. Тогда по формуле (5) будем иметь:

$$f(323{,}5)=\frac{(323{,}5-322{,}8)(323{,}5-324{,}2)(323{,}5-325{,}0)}{(321-322{,}8)(321-324{,}2)(321-325)}\cdot 2{,}50651+$$
$$+\frac{(323{,}5-321)(323{,}5-324{,}2)(323{,}5-325)}{(322{,}8-321)(322{,}8-324{,}2)(322{,}8-325)}\cdot 2{,}50893+$$
$$+\frac{(323{,}5-321)(323{,}5-322{,}8)(323{,}5-325)}{(324{,}2-321)(324{,}2-322{,}8)(324{,}2-325)}\cdot 2{,}51081+$$
$$+\frac{(323{,}5-321)(323{,}5-322{,}8)(323{,}5-324{,}2)}{(325-321)(325-322{,}8)(325-324{,}2)}\cdot 2{,}51188=$$
$$=-0{,}07996+1{,}18794+1{,}83897-0{,}43708=2{,}50987.$$

## § 13*. Вычисление лагранжевых коэффициентов

Укажем схему, облегчающую вычисление коэффициентов при $y_i$ $(i=0,\ 1,\ 2,\ \ldots,\ n)$ в формуле Лагранжа, так называемых *лагранжевых коэффициентов*

$$L_i^{(n)}(x)=\frac{(x-x_0)(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)},\qquad(1)$$

или в более компактной записи

$$L_i^{(n)}(x)=\frac{\Pi_{n+1}(x)}{(x-x_i)\,\Pi'_{n+1}(x_i)},\qquad(2)$$

где

$$\Pi_{n+1}(x)=(x-x_0)\ldots(x-x_n).$$

Формула Лагранжа при этом имеет вид

$$L_n(x)=\sum_{i=0}^{n}L_i^{(n)}(x)\,y_i.$$

Отметим, что форма лагранжевых коэффициентов инвариантна относительно целой линейной подстановки $x=at+b$ ($a$, $b$ постоянны и $a\neq 0$). Действительно, положив в формуле (1)

$$x=at+b;\quad x_j=at_j+b\qquad(j=0,\ 1,\ \ldots,\ n),$$

после сокращения числителя и знаменателя на $a^n$, получим:

$$L_i^{(n)}(t)=\frac{(t-t_0)(t-t_1)\ldots(t-t_{i-1})(t-t_{i+1})\ldots(t-t_n)}{(t_i-t_0)(t_i-t_1)\ldots(t_i-t_{i-1})(t_i-t_{i+1})\ldots(t_i-t_n)}\qquad(3)$$

или

$$L_i^{(n)}=\frac{\Pi_{n+1}(t)}{(t-t_i)\,\Pi'_{n+1}(t_i)},\qquad(3')$$

где

$$\Pi_{n+1}(t)=(t-t_0)(t-t_1)\ldots(t-t_n),$$

что и требовалось доказать.

Для вычисления лагранжевых коэффициентов может быть использована приведенная ниже схема, особенно удобная при применении счетной машины. Сначала располагаем в таблицу разности следующим образом:

$$\begin{array}{cccc} \underline{x-x_0} & x_0-x_1 & x_0-x_2 \ldots & x_0-x_n \\ x_1-x_0 & \underline{x-x_1} & x_1-x_2 \ldots & x_1-x_n \\ x_2-x_0 & x_2-x_1 & \underline{x-x_2} \ldots & x_2-x_n \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ x_n-x_0 & x_n-x_1 & x_n-x_2 \ldots & \underline{x-x_n}. \end{array} \qquad (*)$$

Обозначим произведение элементов первой строки через $D_0$, второй — через $D_1$ и т. д. Произведение же элементов главной диагонали (в схеме эти элементы подчеркнуты), очевидно, будет $\Pi_{n+1}(x)$. Отсюда следует, что

$$L_i^{(n)}(x)=\frac{\Pi_{n+1}(x)}{D_i} \qquad (i=0,\ 1,\ \ldots,\ n). \tag{4}$$

Следовательно,

$$L_n(x)=\Pi_{n+1}(x)\sum_{i=0}^{n}\frac{y_i}{D_i}. \tag{5}$$

В случае равноотстоящих точек лагранжевы коэффициенты могут быть приведены к более простому виду.

В самом деле, полагая

$$x=x_0+th,$$

будем иметь:

$$t_0=0, \quad t_1=1, \ \ldots, \ t_n=n.$$

Отсюда

$$\Pi_{n+1}(t)=t(t-1)(t-2)\ldots(t-n)$$

и

$$\Pi'_{n+1}(i)=(-1)^{n-i}\,i!\,(n-i)!$$

Подставив эти выражения в формулу (3′), получим:

$$L_i^{(n)}(t)=\frac{1}{n!}\Pi_{n+1}(t)\cdot\frac{(-1)^{n-i}C_n^i}{t-i} \qquad (i=0,\ 1,\ \ldots,\ n), \tag{6}$$

где

$$C_n^i=\frac{n!}{i!\,(n-i)!}.$$

Отсюда

$$L_n(x)=\frac{1}{n!}\Pi_{n+1}(t)\sum_{i=0}^{n}(-1)^{n-i}\frac{C_n^i}{t-i}y_i, \tag{7}$$

где

$$t=\frac{x-x_0}{h}.$$

Задача интерполирования в случае постоянного шага $h$ облегчается еще тем, что имеются таблицы для лагранжевых коэффициентов (см. [5]), так что фактически все вычисления сводятся к умножению табличных коэффициентов на соответствующие значения функции $y_i$ и к суммированию.

Пример 1. Для функции $y=y(x)$ дана таблица значений

| $x$ | 0,05 | 0,15 | 0,20 | 0,25 | 0,35 | 0,40 | 0,50 | 0,55 |
|---|---|---|---|---|---|---|---|---|
| $y$ | 0,9512 | 0,8607 | 0,8187 | 0,7788 | 0,7047 | 0,6703 | 0,6065 | 0,5769 |
| $t$ | 1 | 3 | 4 | 5 | 7 | 8 | 10 | 11 |

Найти $y(0{,}45)$.

Решение. Для упрощения вычислений полагаем:

$$x=0{,}05t.$$

Тогда значения новой переменной $t$, соответствующие узлам интерполирования, будут 1, 3, 4, 5, 7, 8, 10, 11. Нам нужно найти значение $y$ при $x=0{,}45$, т. е. при $t=9$. Полагая $t=t_i$ ($i=0, 1, \ldots, 7$), вычисления располагаем по приведенной ниже схеме (таблица 47).

Таблица 47

**Схема вычисления лагранжевых коэффициентов**

| $i$ | $t_i-t_j$ ($j\neq i$) | | | | | | | | $D_i$ | $y_i$ | $\frac{y_i}{D_i}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | $\boxed{8}$ | $-2$ | $-3$ | $-4$ | $-6$ | $-7$ | $-9$ | $-10$ | $-725\,760$ | 0,9512 | $-0{,}0131\cdot10^{-4}$ |
| 1 | 2 | $\boxed{6}$ | $-1$ | $-2$ | $-4$ | $-5$ | $-7$ | $-8$ | 26 880 | 0,8607 | $0{,}3202\cdot10^{-4}$ |
| 2 | 3 | 1 | $\boxed{5}$ | $-1$ | $-3$ | $-4$ | $-6$ | $-7$ | $-7\,560$ | 0,8187 | $-1{,}0829\cdot10^{-4}$ |
| 3 | 4 | 2 | 1 | $\boxed{4}$ | $-2$ | $-3$ | $-5$ | $-6$ | 5 760 | 0,7788 | $1{,}3520\cdot10^{-4}$ |
| 4 | 6 | 4 | 3 | 2 | $\boxed{2}$ | $-1$ | $-3$ | $-4$ | $-3\,456$ | 0,7047 | $-2{,}0390\cdot10^{-4}$ |
| 5 | 7 | 5 | 4 | 3 | 1 | $\boxed{1}$ | $-2$ | $-3$ | 2 520 | 0,6703 | $2{,}6530\cdot10^{-4}$ |
| 6 | 9 | 7 | 6 | 5 | 3 | 2 | $\boxed{-1}$ | $-1$ | 11 340 | 0,6065 | $0{,}5348\cdot10^{-4}$ |
| 7 | 10 | 8 | 7 | 6 | 4 | 3 | 1 | $\boxed{-2}$ | $-80\,640$ | 0,5769 | $-0{,}0715\cdot10^{-4}$ |
| | $\Pi(9)=3840$ | | | | | | | | | | $S=1{,}6535\cdot10^{-4}$ |

Отсюда

$$y(0{,}45)=\Pi(9)\sum_{i=0}^{i=7}\frac{y_i}{D_i}=\Pi(9)\cdot S=3840\cdot 1{,}6535\cdot 10^{-4}=\underline{0{,}6349}.$$

Пример 2. Функция $y=\cos x$ задана таблицей [5]

| $x$ | 5,0 | 5,1 | 5,2 | 5,3 |
|---|---|---|---|---|
| $y$<br>$t$ | 0,283662185<br>0 | 0,377977743<br>1 | 0,468516671<br>2 | 0,554374336<br>3 |
| $x$ | 5,4 | 5,5 | 5,6 | 5,7 |
| $y$<br>$t$ | 0,634692876<br>4 | 0,708669774<br>5 | 0,775565879<br>6 | 0,834712785<br>7 |

Найти $\cos 5{,}347$.

Решение. Сделаем замену переменной по формуле

$$x=0{,}1t+5.$$

Тогда значения переменной $t$, соответствующие узлам интерполирования, будут 0, 1, 2, 3, 4, 5, 6, 7, а искомое значение $x=5{,}347$ перейдет в $t=3{,}47$. Учитывая, что узлы $t_i=i$ $(i=0, 1, \ldots, 7)$ — равноотстоящие, вычисления можно произвести по указанной выше схеме (таблица 48).

Таблица 48

**Схема вычисления лагранжевых коэффициентов для случая равноотстоящих узлов интерполирования**

| $i$ | $x_i$ | $y_i$ | $t-i$ | $(-1)^{7-i}C_7^i$ | $(-1)^{7-i}C_7^i\frac{y_i}{t-i}$ |
|---|---|---|---|---|---|
| 0 | 5,0 | 0,283662185 | 3,47 | —1 | —0,08174702 |
| 1 | 5,1 | 0,377977743 | 2,47 | 7 | 1,07119198 |
| 2 | 5,2 | 0,468516671 | 1,47 | —21 | —6,69309530 |
| 3 | 5,3 | 0,554374336 | 0,47 | 35 | 41,28319523 |
| 4 | 5,4 | 0,634692876 | —0,53 | —35 | 41,91368048 |
| 5 | 5,5 | 0,708669774 | —1,53 | 21 | —9,72684003 |
| 6 | 5,6 | 0,775565879 | —2,53 | —7 | 2,14583444 |
| 7 | 5,7 | 0,834712785 | —3,53 | 1 | —0,23646254 |
| $\Pi=42{,}8848749$ | | | | | $S=69{,}67575724$ |

Из таблицы 48 имеем:

$$\Pi(3{,}47)=\prod_{i=0}^{7}(3{,}47-i)=42{,}8848749$$

и

$$S=\sum_{i=0}^{7}(-1)^{7-i}C_7^i\frac{y_i}{3{,}47-i}=69{,}67575724.$$

На основании формулы (7) получаем:

$$\cos 5{,}347=\frac{1}{7!}\cdot\Pi(3{,}47)\cdot S=0{,}592864312.$$

## § 14. Оценка погрешности интерполяционной формулы Лагранжа

Для функции $y=f(x)$ мы построили в § 12 интерполяционный полином Лагранжа $L_n(x)$, принимающий в точках $x_0, x_1, \ldots, x_n$ заданные значения

$$y_0=f(x_0),\quad y_1=f(x_1),\quad \ldots,\quad y_n=f(x_n).$$

Возникает вопрос, насколько близко построенный полином приближается к функции $f(x)$ в других точках, т. е. как велик остаточный член

$$R_n(x)=f(x)-L_n(x).$$

Для определения этой степени приближения наложим на функцию $y=f(x)$ дополнительные ограничения. Именно, мы будем предполагать, что в рассматриваемой области $a\leqslant x\leqslant b$ изменения $x$, содержащей узлы интерполирования, функция $f(x)$ имеет все производные $f'(x), f''(x), \ldots, f^{(n+1)}(x)$ до $(n+1)$-го порядка включительно.

Введем вспомогательную функцию

$$u(x)=f(x)-L_n(x)-k\Pi_{n+1}(x), \tag{1}$$

где

$$\Pi_{n+1}(x)=(x-x_0)(x-x_1)\ldots(x-x_n)$$

и $k$ — постоянный коэффициент, который будет выбран ниже.

Функция $u(x)$, очевидно, имеет $n+1$ корень в точках

$$x_0,\ x_1,\ \ldots,\ x_n.$$

Подберем теперь коэффициент $k$ так, чтобы $u(x)$ имела $(n+2)$-й корень в любой, но фиксированной точке $\bar{x}$ отрезка $[a, b]$, не

совпадающей с узлами интерполирования (рис. 63). Для этого достаточно положить

$$f(\bar{x}) - L_n(\bar{x}) - k\Pi_{n+1}(\bar{x}) = 0.$$

Отсюда, так как $\Pi_{n+1}(\bar{x}) \neq 0$, то

$$k = \frac{f(\bar{x}) - L_n(\bar{x})}{\Pi_{n+1}(\bar{x})}. \tag{2}$$

При этом значении множителя $k$ функция $u(x)$ имеет $n+2$ корня на отрезке $[a, b]$ и будет обращаться в нуль на концах каждого из отрезков

$$[x_0, x_1], [x_1, x_2], \ldots, [x_i, \bar{x}], [\bar{x}, x_{i+1}], \ldots, [x_{n-1}, x_n].$$

Применяя теорему Ролля к каждому из этих отрезков, убеждаемся, что производная $u'(x)$ имеет не менее $n+1$ корня на отрезке $[a, b]$.

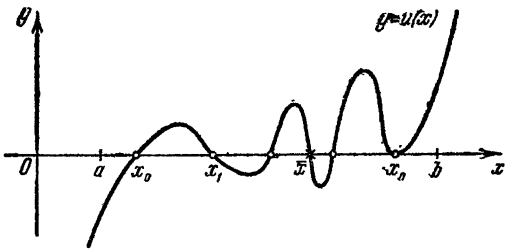


Рис. 63.

Применив теорему Ролля к производной $u'(x)$, мы убедимся, что вторая производная $u''(x)$ обращается в нуль не менее $n$ раз на отрезке $[a, b]$.

Продолжая эти рассуждения, придем к заключению, что на рассматриваемом отрезке $[a, b]$ производная $u^{(n+1)}(x)$ имеет хотя бы один нуль, который обозначим через $\xi$, т. е. $u^{(n+1)}(\xi) = 0$.

Из формулы (1), так как

$$L_n^{(n+1)}(x) = 0 \quad \text{и} \quad \Pi_{n+1}^{(n+1)}(x) = (n+1)!,$$

имеем:

$$u^{(n+1)}(x) = f^{(n+1)}(x) - k(n+1)!$$

При $x = \xi$ получаем:

$$0 = f^{(n+1)}(\xi) - k(n+1)!$$

Отсюда

$$k = \frac{f^{(n+1)}(\xi)}{(n+1)!}. \tag{3}$$

Сравнивая правые части формул (2) и (3), будем иметь:

$$\frac{f(\bar{x}) - L_n(\bar{x})}{\Pi_{n+1}(\bar{x})} = \frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

т. е.

$$f(\bar{x}) - L_n(\bar{x}) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\Pi_{n+1}(\bar{x}). \tag{4}$$

Так как $\bar{x}$ произвольно, то формулу (4) можно записать и так:

$$R_n(x) = f(x) - L_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\Pi_{n+1}(x), \tag{5}$$

где $\xi$ зависит от $x$ и лежит внутри отрезка $[a, b]$.

Отметим, что формула (5) справедлива для всех точек отрезка $[a, b]$, в том числе и для узлов интерполирования.

Обозначая через

$$M_{n+1} = \max_{a \leqslant x \leqslant b} |f^{(n+1)}(x)|,$$

мы получаем следующую оценку для абсолютной погрешности интерполяционной формулы Лагранжа:

$$|R_n(x)| = |f(x) - L_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!}|\Pi_{n+1}(x)|, \tag{6}$$

где

$$\Pi_{n+1}(x) = (x - x_0)(x - x_1)\ldots(x - x_n). \tag{6'}$$

Пример. С какой точностью можно вычислить $\sqrt{115}$ с помощью интерполяционной формулы Лагранжа для функции $y = \sqrt{x}$, выбрав узлы интерполирования $x_0 = 100$, $x_1 = 121$, $x_2 = 144$?

Решение. Имеем:

$$y' = \frac{1}{2}x^{-\frac{1}{2}}, \quad y'' = -\frac{1}{4}x^{-\frac{3}{2}}, \quad y''' = \frac{3}{8}x^{-\frac{5}{2}}.$$

Отсюда

$$M_3 = \max|y'''| = \frac{3}{8}\cdot\frac{1}{\sqrt{100^5}} = \frac{3}{8}\cdot 10^{-5} \quad \text{при} \quad 100 \leqslant x \leqslant 144.$$

На основании формулы (6) получаем:

$$|R_2| \leqslant \frac{3}{8}\cdot 10^{-5}\cdot\frac{1}{3!}|(115-100)(115-121)(115-144)| =$$

$$= \frac{1}{16}\cdot 10^{-5}\cdot 15\cdot 6\cdot 29 \approx 1{,}6\cdot 10^{-3}.$$

## § 15. Оценки погрешностей интерполяционных формул Ньютона

Если узлы интерполирования $x_0, x_1, \ldots, x_n$ — равноотстоящие, причем

$$x_{i+1} - x_i = h \qquad (i = 0, 1, 2, \ldots, n-1),$$

то, полагая

$$q = \frac{x - x_0}{h},$$

на основании формулы (5) из предыдущего параграфа получим *остаточный член первой интерполяционной формулы Ньютона*

$$R_n(x) = h^{n+1} \cdot \frac{q(q-1)\ \ldots\ (q-n)}{(n+1)!} f^{(n+1)}(\xi), \tag{1}$$

где $\xi$ — некоторое промежуточное значение между узлами интерполирования $x_0, x_1, \ldots, x_n$ и рассматриваемой точкой $x$. Заметим, что для случая интерполирования в узком смысле слова $\xi \in [x_0, x_n]$; при экстраполировании возможно, что $\xi \overline{\in} [x_0, x_n]$.

Аналогично, полагая в формуле (5) из § 14

$$q = \frac{x - x_n}{h},$$

получим *остаточный член второй интерполяционной формулы Ньютона*

$$R_n(x) = h^{n+1} \cdot \frac{q(q+1)\ \ldots\ (q+n)}{(n+1)!} f^{(n+1)}(\xi), \tag{2}$$

где $\xi$ — некоторое промежуточное значение между узлами интерполирования $x_0, x_1, \ldots, x_n$ и точкой $x$.

Обычно при практических вычислениях интерполяционная формула Ньютона обрывается на членах, содержащих такие разности, которые в пределах заданной точности можно считать постоянными.

Предполагая, что $\Delta^{n+1}y$ почти постоянны для функции $y = f(x)$ и $h$ достаточно мало, и учитывая, что

$$f^{(n+1)}(x) = \lim_{h \to 0} \frac{\Delta^{n+1}y}{h^{n+1}},$$

приближенно можно положить:

$$f^{(n+1)}(\xi) \approx \frac{\Delta^{n+1}y_0}{h^{n+1}}.$$

В этом случае остаточный член первой интерполяционной формулы Ньютона равен

$$R_n(x) \approx \frac{q(q-1)\ \ldots\ (q-n)}{(n+1)!} \Delta^{n+1}y_0.$$

В этих же условиях для остаточного члена второй интерполяционной формулы Ньютона получаем выражение

$$R_n(x) \approx \frac{q(q+1)\ \ldots\ (q+n)}{(n+1)!} \Delta^{n+1}y_n.$$

Пример 1. В пятизначных таблицах логарифмов даются логарифмы целых чисел от $x = 1000$ до $x = 10\,000$ с предельной абсолютной погрешностью, равной $\frac{1}{2} \cdot 10^{-5}$. Возможно ли линейное интерполирование с той же степенью точности?

Решение. Полагая

$$y = \lg x,$$

будем иметь:

$$y' = \frac{M}{x} \text{ и } y'' = -\frac{M}{x^2},$$

где $M = 0{,}43$. Отсюда

$$M_2 = \max |y''| < \frac{0{,}5}{10^6} = \frac{1}{2} \cdot 10^{-6}.$$

Из формулы (1) при $n = 2$ и $h = 1$ получаем оценку для погрешности линейного интерполирования:

$$|R_1(x)| \leqslant \frac{|q(q-1)|}{2!} M_2 \leqslant \frac{q(1-q)}{2} \cdot \frac{1}{2} \cdot 10^{-6}.$$

Так как при $0 \leqslant q \leqslant 1$ имеем

$$q(1-q) = \frac{1}{4} - \left(\frac{1}{2} - q\right)^2 \leqslant \frac{1}{4},$$

то окончательно получаем:

$$|R_1(x)| \leqslant \frac{\frac{1}{4}}{2} \cdot \frac{1}{2} \cdot 10^{-6} < 10^{-7}.$$

Следовательно, линейное интерполирование вполне допустимо.

Пример 2. Оценить погрешность, получающуюся при приближении функции $f(x) = \sin x$ интерполяционным полиномом пятой степени $P_5(x)$, совпадающим с данной функцией при значениях $x = 0°, 5°, 10°, 15°, 20°, 25°$.

Решение. Здесь $f^{(6)}(x) = -\sin x$; поэтому $|f^{(6)}(x)| \leqslant 1$. На основании формулы (1) имеем:

$$|\sin x - P_5(x)| \leqslant$$
$$\leqslant \frac{1}{6!} \left| x\left(x - \frac{\pi}{36}\right)\left(x - \frac{\pi}{18}\right)\left(x - \frac{\pi}{12}\right)\left(x - \frac{\pi}{9}\right)\left(x - \frac{5\pi}{36}\right) \right|.$$

Например, при $x = 12°\ 30' = \text{arc}\, 0{,}21816$ получим:

$$|\sin x - P_5(x)| < 2{,}2 \cdot 10^{-9}.$$

## § 16. Оценки погрешностей центральных интерполяционных формул

Приводим без доказательства остаточные члены для формул Стирлинга и Бесселя [3].

а) *Остаточный член интерполяционной формулы Стирлинга.* Если $2n$ — порядок максимальной используемой разности таблицы и $x \in [x_0 - nh,\ x_0 + nh]$, то

$$R_n(x) = \frac{h^{2n+1} f^{(2n+1)}(\xi)}{(2n+1)!} q(q^2 - 1^2)(q^2 - 2^2)(q^2 - 3^2) \ \ldots \ (q^2 - n^2),$$

где

$$q=\frac{x-x_0}{h} \text{ и } \xi\in[x_0-nh,\ x_0+nh].$$

Если же аналитическое выражение функции $f(x)$ неизвестно, то при $h$ малом полагают:

$$R_n(x)\approx\frac{\Delta^{2n+1}y_{-n-1}+\Delta^{2n+1}y_{-n}}{2\cdot(2n+1)!}q(q^2-1^2)(q^2-2^2)\ \ldots\ (q^2-n^2).$$

б) *Остаточный член интерполяционной формулы Бесселя.* Если $2n+1$ — порядок максимальной используемой разности таблицы и $x\in[x_0-nh,\ x_0+(n+1)h]$, то

$$R_n(x)=\frac{h^{2n+2}}{(2n+2)!}f^{(2n+2)}(\xi)\,q(q^2-1^2)(q^2-2^2)\ldots(q^2-n^2)[q-(n+1)],$$

где

$$q=\frac{x-x_0}{h} \quad \text{и} \quad \xi\in[x_0-nh,\ x_0+(n+1)h].$$

Если же функция $f(x)$ задана таблично и шаг $h$ мал, то принимают:

$$R_n(x)\approx\frac{\Delta^{2n+2}y_{-n-1}+\Delta^{2n+2}y_{-n}}{2(2n+2)!}q(q^2-1^2)(q^2-2^2)\ldots$$
$$\ldots\ (q^2-n^2)[q-(n+1)].$$

В частности, при $q=\frac{1}{2}$ получаем погрешность при *интерполировании на середину*

$$R_n=\frac{h^{2n+2}f^{(2n+2)}(\xi)}{(2n+2)!}(-1)^{n+1}\frac{[1\cdot3\cdot5\ \ldots\ (2n+1)]^2}{2^{2n+2}}$$

или

$$R_n\approx\frac{\Delta^{2n+2}y_{-n-1}+\Delta^{2n+2}y_{-n}}{2(2n+2)!}(-1)^{n+1}\frac{[1\cdot3\cdot5\ \ldots\ (2n+1)]^2}{2^{2n+2}}.$$

Если положить

$$q=p+\frac{1}{2},$$

то формула для остаточного члена формулы Бесселя принимает вид

$$R_n(x)=\frac{h^{2n+2}}{(2n+2)!}f^{(2n+2)}(\xi)\left(p^2-\frac{1}{4}\right)\left(p^2-\frac{9}{4}\right)\ \ldots\ \left[p^2-\frac{(2n+1)^2}{4}\right].$$

## § 17. О наилучшем выборе узлов интерполирования

Анализируя формулу (5) из § 14, мы видим, что погрешность $R_n(x)$ формулы Лагранжа представляет собой, с точностью до числовой постоянной, произведение двух множителей, из которых один, $f^{(n+1)}(\xi)$, зависит от свойств функции $f(x)$ и не поддается регулированию,

а величина другого, $\Pi_{n+1}(x)$, определяется исключительно выбором узлов интерполирования.

При неудачном расположении узлов интерполирования $x_i$ верхняя грань модуля погрешности $R_n(x)$ ((6) § 14) может быть весьма большой. Например, если мы сконцентрируем узлы $x_i$ вблизи одного конца отрезка $[a, b]$, то $R_n(x)$ при $l=b-a>1$, вообще говоря, будет велик в точках $x$, близких к другому концу отрезка. Поэтому возникает задача о наиболее рациональном выборе узлов интерполирования $x_i$ (при заданном числе узлов $n$) с тем, чтобы находящаяся в нашей власти часть погрешности — полином $\Pi_{n+1}(x)$ имел наименьшее максимальное значение по абсолютной величине на отрезке $[a, b]$, или, как коротко говорят, «наименее отклонялся от нуля на $[a, b]$». Эта задача была решена русским математиком П. Л. Чебышевым [2], [6], который доказал, что наилучший выбор в указанном смысле узлов интерполирования дается формулой

$$x_i=\frac{b+a}{2}+\frac{b-a}{2}\xi_i,$$

где

$$\xi_i=-\cos\frac{2i+1}{2n+2}\pi \qquad (i=0,\ 1,\ 2,\ \ldots,\ n)$$

— нули так называемого полинома Чебышева $T_{n+1}(x)$. В этом случае мы будем иметь:

$$|\Pi_{n+1}(x)|\leqslant 2\left(\frac{b-a}{4}\right)^{n+1}.$$

Интересно отметить, что эти узлы не являются равноотстоящими, а сгущаются около концов отрезка. Даже при таком подборе узлов в общем случае нельзя гарантировать, что абсолютная величина погрешности будет сколь угодно мала при достаточно большом $n$.

Сделаем общие замечания об определении погрешностей интерполяционных формул. Если максимальные разности практически постоянны, то результат интерполирования в узком смысле обыкновенно имеет столько верных десятичных знаков, сколько их есть в табличных данных, и поэтому оценка погрешностей не обязательна. При пользовании интерполяционной формулой Лагранжа нет возможности следить за ходом конечных разностей, и поэтому следует, если это возможно, оценивать остаточный член.

Если функция $f(x)$ задана таблично и аналитическое выражение ее неизвестно, то оценка погрешности интерполяционного полинома, строго говоря, является невозможной. Действительно, для данного полинома теоретически можно построить бесчисленное множество различных функций, совпадающих с этим полиномом в данной системе узлов. Таким образом, в промежуточных точках отклонение интерполяционного полинома от функции может быть каким угодно.

Однако, если природа функции такова, что график ее представляет собой плавную кривую, то приближенно погрешности интерполирующих полиномов с большой степенью уверенности можно определять на основании значений конечных разностей высших порядков по приведенным выше формулам.

## § 18. Разделенные разности

При построении таблицы разностей мы до сих пор предполагали, что значения аргумента функции — равноотстоящие, т. е. имеют *постоянный шаг*. Однако на практике встречаются также таблицы для неравноотстоящих значений аргумента, т. е. таблицы с переменным шагом. Например, такой характер часто имеют эмпирические данные. Для таблиц с переменным шагом понятие конечных разностей обобщается, а именно: вводятся так называемые *разделенные разности*.

Пусть функция $y=f(x)$ задана таблично и $x_0, x_1, x_2, \ldots$ — значения аргумента, а $y_0, y_1, y_2, \ldots$ — соответствующие значения функции, где разности

$$\Delta x_i = x_{i+1} - x_i \neq 0 \qquad (i=0, 1, \ldots)$$

не равны между собой.

Отношения

$$[x_i, x_{i+1}] = \frac{y_{i+1}-y_i}{x_{i+1}-x_i}$$

$(i=0, 1, 2, \ldots)$ называются *разделенными разностями первого порядка*. Например,

$$[x_0, x_1] = \frac{y_1-y_0}{x_1-x_0}; \qquad [x_1, x_2] = \frac{y_2-y_1}{x_2-x_1} \text{ и т. д.}$$

Аналогично определяются *разделенные разности второго порядка*

$$[x_i, x_{i+1}, x_{i+2}] = \frac{[x_{i+1}, x_{i+2}]-[x_i, x_{i+1}]}{x_{i+2}-x_i}$$

$(i=0, 1, 2, \ldots)$. Например,

$$[x_0, x_1, x_2] = \frac{[x_1, x_2]-[x_0, x_1]}{x_2-x_0}$$

и т. д.

Вообще, *разделенные разности $n$-го порядка* получаются из разделенных разностей $(n-1)$-го порядка с помощью рекуррентного соотношения

$$[x_i, x_{i+1}, \ldots, x_{i+n}] = \frac{[x_{i+1}, \ldots, x_{i+n}]-[x_i, \ldots, x_{i+n-1}]}{x_{i+n}-x_i} \qquad (1)$$

$(n=1, 2, \ldots;\ i=0, 1, 2, \ldots)$.

Заметим, что разделенные разности не меняются при перестановке элементов, т. е. представляют собой *симметрические функции* своих аргументов. Например,

$$[x_0,\ x_1]=\frac{y_1-y_0}{x_1-x_0}=\frac{y_0-y_1}{x_0-x_1}=[x_1,\ x_0] \text{ и т. д.}$$

Разделенные разности обычно располагаются в таблицу приведенного ниже вида (таблица 49).

Таблица 49

**Таблица разделенных разностей**

| $x$ | $y$ | Разделенные разности | | | |
|---|---|---|---|---|---|
| | | 1-го пор. | 2-го пор. | 3-го пор. | 4-го пор. |
| $x_0$ | $y_0$ | | | | |
| | | $[x_0, x_1]$ | | | |
| $x_1$ | $y_1$ | | $[x_0, x_1, x_2]$ | | |
| | | $[x_1, x_2]$ | | $[x_0, x_1, x_2, x_3]$ | |
| $x_2$ | $y_2$ | | $[x_1, x_2, x_3]$ | | $[x_0, x_1, x_2, x_3, x_4]$ |
| | | $[x_2, x_3]$ | | $[x_1, x_2, x_3, x_4]$ | |
| $x_3$ | $y_3$ | | $[x_2, x_3, x_4]$ | | |
| | | $[x_3, x_4]$ | | | |
| $x_4$ | $y_4$ | | | | |

Пример. Составить разделенные разности для функции, заданной следующей таблицей:

| $x$ | 0 | 0,2 | 0,3 | 0,4 | 0,7 | 0,9 |
|---|---|---|---|---|---|---|
| $y$ | 132,651 | 148,877 | 157,464 | 166,375 | 195,112 | 216,000 |

Решение. Последовательно применяя формулу (1), будем иметь

$$[x_0,\ x_1]=\frac{148{,}877-132{,}651}{0{,}2-0}=81{,}13;$$

$$[x_1,\ x_2]=\frac{157{,}464-148{,}877}{0{,}3-0{,}2}=85{,}87;$$

$$[x_0,\ x_1,\ x_2]=\frac{85{,}87-81{,}13}{0{,}3-0}=15{,}8$$

и т. д. Результаты вычислений приведены в таблице 50.

Таблица 50

**Разделенные разности функции $y$**

| $x$ | $y$ | 1-й пор. | 2-й пор. | 3-й пор. | 4-й пор. |
|---|---|---|---|---|---|
| 0 | 132,651 | | | | |
| | | 81,13 | | | |
| 0,2 | 140,877 | | 15,8 | | |
| | | 85,87 | | 1 | |
| 0,3 | 157,464 | | 16,2 | | 0 |
| | | 89,11 | | 1 | |
| 0,4 | 166,375 | | 16,7 | | 0 |
| | | 95,79 | | 1 | |
| 0,7 | 195,112 | | 17,3 | | |
| | | 104,44 | | | |
| 0,9 | 216,000 | | | | |

## § 19. Интерполяционная формула Ньютона для неравноотстоящих значений аргумента

Пользуясь понятием разделенных разностей, интерполяционную формулу Лагранжа можно представить в виде, аналогичном первой интерполяционной формуле Ньютона. Докажем предварительно одну лемму, представляющую также самостоятельный интерес.

Лемма. *Если $y=P(x)$ есть полином $n$-й степени, то его разделенная разность $(n+1)$-го порядка тождественно равна нулю, т. е.*

$$[x, x_0, x_1, \ldots, x_n]\equiv 0$$

*для любой системы различных между собой чисел $x, x_0, x_1, \ldots, x_n$.*

Действительно, если $P(x)$ — полином $n$-й степени, то

$$[x, x_0]=\frac{P(x)-P(x_0)}{x-x_0}\equiv P(x, x_0)$$

является полиномом $(n-1)$-й степени относительно $x$. Далее,

$$[x, x_0, x_1]=\frac{P(x, x_0)-P(x_0, x_1)}{x-x_1}\equiv P(x, x_0, x_1)$$

представляет собой полином $(n-2)$-й степени относительно $x$. В самом деле, функция $P(x, x_0)-P(x_0, x_1)=P(x, x_0)-P(x_1, x_0)$ имеет корень $x=x_1$ и, следовательно, на основании теоремы Безу полином $P(x, x_0)-P(x_0, x_1)$ без остатка делится на двучлен $x-x_1$. С помощью аналогичных рассуждений убеждаемся, что

$$[x, x_0, \ldots, x_{n-1}]\equiv P(x, x_0, \ldots, x_{n-1})$$

есть полином нулевой степени, т. е.

$$P(x, x_0, \ldots, x_{n-1})=C.$$

Отсюда

$$[x, x_0, \ldots, x_n] = \frac{C-C}{x-x_n} = 0.$$

Пусть теперь $P(x)$ — полином Лагранжа степени $n$ такой, что

$$P(x_i) = f(x_i) = y_i \tag{1}$$

($i = 0, 1, \ldots, n$), где $y = f(x)$ — данная функция. Обозначим через $P(x, x_0)$, $P(x, x_0, x_1)$, $\ldots$, $P(x, x_0, \ldots, x_n)$ последовательные разделенные разности полинома $P(x)$. Имеем:

$$\left.\begin{aligned} &P(x_0, x_1) = [x_0, x_1], \\ &P(x_0, x_1, x_2) = [x_0, x_1, x_2], \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ &P(x_0, x_1, \ldots, x_n) = [x_0, x_1, \ldots, x_n]; \end{aligned}\right\} \tag{2}$$

кроме того, на основании леммы

$$P(x, x_0, \ldots, x_n) = 0. \tag{3}$$

По определению получаем:

$$\frac{P(x) - P(x_0)}{x - x_0} = P(x, x_0); \tag{4}$$

отсюда

$$P(x) = P(x_0) + P(x, x_0)(x - x_0). \tag{5}$$

Из определения разделенных разностей следует, что

$$P(x, x_0, \ldots, x_m) = \frac{P(x, x_0, \ldots, x_{m-1}) - P(x_0, \ldots, x_m)}{x - x_m}.$$

Отсюда

$$P(x, x_0, \ldots, x_{m-1}) = P(x_0, \ldots, x_m) + \\ + (x - x_m) P(x, x_0, \ldots, x_m) \tag{6}$$

($m = 1, 2, \ldots, n$).

Используя формулу (6), из формулы (5) последовательно выводим:

$$\begin{aligned} P(x) &= P(x_0) + P(x, x_0)(x - x_0) = \\ &= P(x_0) + P(x_0, x_1)(x - x_0) + \\ &\quad + P(x, x_0, x_1)(x - x_0)(x - x_1) = \\ &= P(x_0) + P(x_0, x_1)(x - x_0) + \\ &\quad + P(x_0, x_1, x_2)(x - x_0)(x - x_1) + \ldots \\ &\ldots + P(x_0, x_1, \ldots, x_n)(x - x_0)(x - x_1) \ldots (x - x_{n-1}) + \\ &\quad + P(x, x_0, \ldots, x_n)(x - x_0)(x - x_1) \ldots (x - x_n), \end{aligned}$$

или, учитывая равенства (2) и (3), окончательно получаем *интерполяционную формулу Ньютона для неравноотстоящих значений аргумента*

$$P(x)=y_0+[x_0,\ x_1](x-x_0)+[x_0,\ x_1,\ x_2](x-x_0)(x-x_1)+\ldots$$
$$\ldots+[x_0,\ x_1,\ \ldots,\ x_n](x-x_0)(x-x_1)\ \ldots\ (x-x_{n-1}). \qquad (7)$$

Погрешность формулы (7), как обычно, равна

$$R(x)=f(x)-P(x)=\frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)(x-x_1)\ \ldots\ (x-x_n), \qquad (8)$$

где $\xi$ — промежуточное значение между точками $x_0,\ x_1,\ \ldots,\ x_n$ и $x$.

Пример. Составить интерполяционный полином для функции $y=f(x)$, заданной таблицей

| $x$ | 0 | 2,5069 | 5,0154 | 7,52270 |
|---|---|---|---|---|
| $y$ | 0,3989423 | 0,3988169 | 0,3984408 | 0,3978138 |

С помощью этого полинома найти $f(3{,}7608)$.

Решение. Находим разделенные разности функции $y$ (таблица 51).

Таблица 51

**Разделенные разности функции $y$**

| $x$ | $y$ | 1-й пор. | 2-й пор. | 3-й пор. |
|---|---|---|---|---|
| 0 | 0,3989423 | | | |
| | | —500 | | |
| 2,5069 | 0,3988169 | | —199 | |
| | | —1499 | | 0 |
| 5,0154 | 0,3984408 | | —199 | |
| | | —2496 | | |
| 7,5270 | 0,3978138 | | | |

Используя формулу (7), находим:

$$y=0{,}3989423-0{,}0000500x-0{,}0000199x(x-2{,}5069).$$

Отсюда

$$y(3{,}7608)=0{,}3989423-0{,}0000500\cdot 3{,}7608-$$
$$-0{,}0000199\cdot 3{,}7608\cdot(3{,}7608-2{,}5069)=0{,}3986604.$$

## § 20. Обратное интерполирование для случая равноотстоящих узлов

Пусть функция $y=f(x)$ задана таблично.

Задача *обратного интерполирования* заключается в том, чтобы по заданному значению функции $y$ определить соответствующее значение аргумента $x$.

Остановимся сначала на случае равноотстоящих узлов. Здесь обычно используется *метод последовательных приближений*.

Предположим, что функция $f(x)$ монотонна и данное значение $y$ содержится между $y_0=f(x_0)$ и $y_1=f(x_1)$.

Заменяя функцию $y$ первым интерполяционным полиномом Ньютона, будем иметь:

$$y=y_0+\frac{\Delta y_0}{1!}q+\frac{\Delta^2 y_0}{2!}q(q-1)+\ \ldots\ +\frac{\Delta^n y_0}{n!}q(q-1)\ \ldots\ (q-n+1);$$

отсюда $q=\varphi(q)$, где

$$\varphi(q)=\frac{y-y_0}{\Delta y_0}-\frac{\Delta^2 y_0}{2!\,\Delta y_0}q(q-1)-\ldots$$
$$\ldots-\frac{\Delta^n y_0}{n!\,\Delta y_0}q(q-1)\ \ldots\ (q-n+1).$$

За начальное приближение принимаем:

$$q_0=\frac{y-y_0}{\Delta y_0}.$$

Тогда, применяя метод итерации, получим:

$$q_m=\varphi(q_{m-1})\qquad (m=1,\ 2,\ \ldots). \tag{1}$$

Если $f(x)\in C^{(n+1)}[a,b]$, где отрезок $[a,b]$ содержит узлы интерполяции, и шаг $h$ достаточно мал, то этот процесс сходится, т. е.

$$\lim_{m\to\infty} q_m=q,$$

где $q$ — истинное решение.

На практике процесс итерации продолжают до тех пор, пока не установятся цифры, соответствующие требуемой точности, причем полагают $q\approx q_s$, где $q_s$ — последнее приближение.

Найдя $q$, определяем затем $x$ из формулы

$$\frac{x-x_0}{h}=q;$$

отсюда

$$x=x_0+qh.$$

Пример 1. Используя значения функции $y = \lg x$, данные в таблице

| $x$ | 20 | 25 | 30 |
|---|---|---|---|
| $y$ | 1,3010 | 1,3979 | 1,4771 |

найти значения $x$ такое, что $y = \lg x = 1{,}35$.

Решение. Составляем таблицу разностей.

Таблица 52

**Конечные разности функции $y$**

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ |
|---|---|---|---|
| 20 | 1,3010 | 969 | —177 |
| 25 | 1,3979 | 792 | |
| 30 | 1,4771 | | |

Принимая $y_0 = 1{,}3010$, будем иметь:

$$q_0 = \frac{y - y_0}{\Delta y_0} = \frac{1{,}35 - 1{,}3010}{0{,}0969} = \frac{490}{969} = 0{,}506.$$

Далее, удерживая три знака после запятой, последовательно получаем:

$$q_1 = 0{,}506 - \frac{177}{2 \cdot 969} \cdot 0{,}506\,(1 - 0{,}506) = 0{,}506 - 0{,}023 = 0{,}483;$$

$$q_2 = 0{,}506 - \frac{177}{2 \cdot 969} \cdot 0{,}483\,(1 - 0{,}483) = 0{,}506 - 0{,}023 = 0{,}483.$$

Принимаем

$$q = 0{,}483.$$

Отсюда

$$x + x_0 + qh = 20 + 0{,}483 \cdot 5 = 22{,}42.$$

По таблице антилогарифмов имеем $x = 22{,}39$. Значительное расхождение вычисленного значения и точного объясняется тем, что шаг $h = 5$ велик.

Мы применили метод итерации для решения задачи обратного интерполирования, пользуясь первой интерполяционной формулой Ньютона. Но совершенно аналогично можно применить этот способ и к другим интерполяционным формулам: ко второй формуле Ньютона, к формулам Стирлинга, Бесселя и др. Покажем это на следующем примере.

Пример 2. В таблице 53 приведены значения интеграла вероятностей [3]

$$y=\frac{2}{\sqrt{\pi}}\int_0^x e^{-x^2}dx.$$

При каком значении $x$ интеграл $y$ равен $\frac{1}{2}$?

Таблица 53

**Значение интеграла вероятностей**

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ | $\Delta^3 y$ | $\Delta^4 y$ |
|---|---|---|---|---|---|
| 0,45 | 0,4754818 | | | | |
| | | 91737 | | | |
| 0,46 | 0,4846555 | | —840 | | |
| | | 90897 | | —11 | |
| 0,47 | 0,4937452 | | —851 | | 1 |
| | | 90046 | | —10 | |
| 0,48 | 0,5027498 | | —861 | | 2 |
| | | 89185 | | — 8 | |
| 0,49 | 0,5116683 | | —869 | | |
| | | 88316 | | | |
| 0,50 | 0,5204999 | | | | |

Решение. Дополняем таблицу 53 конечными разностями функции $y$. Ближайшим табличным значением для аргумента $x$, соответствующим значению функции $y=\frac{1}{2}$, является $x_0=0{,}47$. Здесь удобно применить формулу Бесселя.

Имеем $x_0=0{,}47$; $h=0{,}01$; $y=0{,}5$.

Подставляя эти значения в формулу (8) из § 7 и используя соответствующие табличные данные, получим:

$$0{,}5=0{,}4982475+0{,}0090046\,p+\frac{p^2-0{,}25}{2}\left(\frac{-851-861}{2}\right)\cdot 10^{-7}+$$
$$+\frac{p(p^2-0{,}25)}{6}(-10)\cdot 10^{-7}. \qquad (2)$$

Отсюда, разделив обе части равенства (2) на 0,0090046 и изолировав член, содержащий $p$ в первой степени, будем иметь:

$$p=0{,}194623+4{,}753\cdot 10^{-3}(p^2-0{,}25)+1{,}85\cdot 10^{-5}p(p^2-0{,}25). \qquad (3)$$

В качестве первого приближения параметра $p$ возьмем:

$$p^{(1)}=0{,}194623.$$

Подставив $p^{(1)}$ в выражение (3), получим второе приближение:

$$p^{(2)} = 0{,}194623 + 4{,}753 \cdot 10^{-3}\,[(0{,}194623)^2 - 0{,}25] + \\ + 1{,}85 \cdot 10^{-5} \cdot 0{,}194623 \cdot [(0{,}194623)^2 - 0{,}25] = \\ = 0{,}194623 - 0{,}001008 - 0{,}000001 = 0{,}193614.$$

Аналогично, подставляя $p^{(2)}$ вместо $p$ в формулу (3), получим третье приближение:

$$p^{(3)} = 0{,}193612.$$

Так как имеем совпадение первых пяти знаков после запятой, то процесс итерации считается законченным.

Далее, последовательно находим:

$$q = p + \frac{1}{2} = 0{,}693612$$

и

$$x = x_0 + qh = 0{,}47 + 0{,}01 \cdot 0{,}693612 = 0{,}47693612.$$

Это значение верно до шестого десятичного знака.

## § 21. Обратное интерполирование для случая неравноотстоящих узлов

Задача обратного интерполирования функции для случая неравноотстоящих значений аргумента $x_0, x_1, \ldots, x_n$ непосредственно может быть решена с помощью интерполяционной формулы Лагранжа.

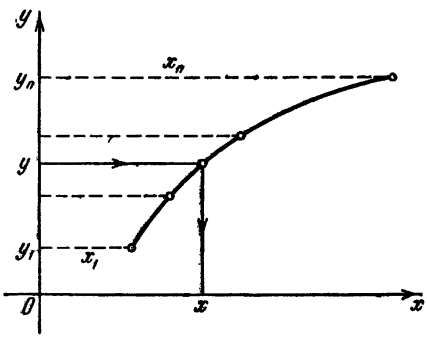


Рис. 64.

Для этого достаточно принять переменную $y$ за независимую и написать формулу, выражающую $x$ как функцию $y$ (рис. 64)

$$x = \sum_{i=0}^{n} \frac{(y-y_1)(y-y_2)\ldots(y-y_{i-1})(y-y_{i+1})\ldots(y-y_n)}{(y_i-y_1)(y_i-y_2)\ldots(y_i-y_{i-1})(y_i-y_{i+1})\ldots(y_i-y_n)}\, x_i \qquad (1)$$

где $y_i = f(x_i)$ $(i = 0, 1, \ldots, n)$. Можно также, считая $y$ аргументом, использовать интерполяционную формулу Ньютона для неравноотстоящих значений аргумента (см. § 19):

$$x = x_0 + [y_0, y_1](y - y_0) + [y_0, y_1, y_2](y - y_0)(y - y_1) + \ldots$$
$$\ldots + [y_0, y_1, \ldots, y_n](y - y_0)(y - y_1)\ldots(y - y_{n-1}), \qquad (2)$$

где $[y_0, y_1]$, $[y_0, y_1, y_2]$, $\ldots$, $[y_0, y_1, \ldots, y_n]$ — соответствующие разделенные разности.

Пример. Решить пример 2 из § 20 с помощью формулы Лагранжа для обратного интерполирования [3].

Решение. Ограничимся четырьмя значениями:

$$x_0 = 0{,}46;\ x_1 = 0{,}47;\ x_2 = 0{,}48;\ x_4 = 0{,}49.$$

Полагая

$$u = 10^7 y - \frac{1}{2} \cdot 10^7,$$

будем иметь следующую таблицу:

| $x$ | 0,46 | 0,47 | 0,48 | 0,49 |
|---|---|---|---|---|
| $u$ | —153445 | —62548 | 27498 | 116683 |

Данное значение $y = \frac{1}{2}$ соответствует $u = 0$. Применяя формулу (2), где $y$ заменено на $u$, получим:

$$x = \frac{62548 \cdot (-27498) \cdot (-116683)}{(-153445 + 62548) \cdot (-153445 - 27498) \cdot (-153445 - 116683)} \cdot 0{,}46 +$$
$$+ \frac{153445 \cdot (-27498)(-116683)}{(-62548 + 153445) \cdot (-62548 - 27498) \cdot (-62548 - 116683)} \cdot 0{,}47 +$$
$$+ \frac{153445 \cdot 62548 \cdot (-116683)}{(27498 + 153445) \cdot (27498 + 62548) \cdot (27498 - 116683)} \cdot 0{,}48 +$$
$$+ \frac{153445 \cdot 62548 \cdot (-27498)}{(116683 + 153445) \cdot (116683 + 62548) \cdot (116683 - 27498)} \cdot 0{,}49 =$$
$$= -0{,}020779 + 0{,}157737 + 0{,}369928 - 0{,}029950 = 0{,}476936.$$

## § 22. Нахождение корней уравнения методом обратного интерполирования

Отметим в заключение, что решение уравнения

$$f(x) = 0$$

можно свести к задаче обратного интерполирования. Для этого нужно составить таблицу значений функции $y = f(x)$ и построить соответствующую таблицу конечных разностей для значений $x$,

близких к корню. А затем применить приемы обратного интерполирования, отыскивая значение $x$, соответствующее $y=0$.

Пример. По данной таблице значений функции Бесселя $y=J_0(x)$

| $x$ | 2,4 | 2,5 | 2,6 |
|---|---|---|---|
| $y$ | 0,0025 | —0,0484 | —0,0968 |

найти с точностью до $10^{-3}$ корень уравнения $J_0(x)=0$, лежащий в интервале (2,4; 2,6).

Решение. Составляем таблицу разностей (таблица 54).

Таблица 54

**Конечные функции бесселевой функции $y=J_0(x)$**

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ |
|---|---|---|---|
| 2,4 | 0,0025 | —509 | 25 |
| 2,5 | —0,0484 | —484 | |
| 2,6 | —0,0968 | | |

Полагая $y=0$ и $x_0=2,4$; $y_0=0,0025$, на основании приведенной выше в § 20 формулы (1) получаем:

$$q_0=\frac{y-y_0}{\Delta y_0}=\frac{0,0025}{0,0509}=0,049;$$

$$q_1=q_0+\frac{\Delta^2 y_0}{2\Delta y_0}q_0(1-q_0)=$$

$$=0,049-\frac{25}{2\cdot 509}\cdot 0,049\cdot 0,951=0,049-0,001=0,048;$$

$$q_2=0,049-\frac{25}{2\cdot 509}\cdot 0,048\cdot 0,952=0,049-0,001=0,048.$$

Принимаем

$$q=0,048;$$

отсюда

$$x=x_0+qh=2,4+0,048\cdot 0,1=2,405.$$

По таблицам

$$x=2,4048.$$

## § 23. Метод интерполяции для развертывания векового определителя

Интерполирование функций может быть использовано для развертывания векового (характеристического) определителя (см. гл. XII)

$$D(\lambda) = \det(A - \lambda E),$$

где $A = [a_{ij}]$.

Выберем равноотстоящие узлы

$$\lambda_0 = 0, \ \lambda_1 = 1, \ \ldots, \ \lambda_n = n$$

и для определителя $D(\lambda)$ вычислим соответствующие значения

$$D(0) = D_0, \ D(1) = D_1, \ \ldots, \ D(n) = D_n.$$

Составляя горизонтальную таблицу разностей для последовательности чисел $D(0)$, $D(1)$, ..., $D(n)$, обычным приемом находим разности $\Delta^i D(0)$ $(i = 0, 1, \ldots, n)$. Отсюда, применяя первую интерполяционную формулу Ньютона, получим полиномиальное выражение для векового определителя

$$D(\lambda) = D(0) + \sum_{i=1}^{n} \frac{\Delta^i D(0)}{i!} \lambda(\lambda - 1) \ldots (\lambda - i + 1). \tag{1}$$

Если положить

$$\frac{\lambda(\lambda-1)\ldots(\lambda-i+1)}{i!} = \sum_{m=1}^{i} c_{mi}\lambda^m \quad (i = 1, 2, \ldots), \tag{2}$$

то после несложных преобразований получаем *формулу А. А. Маркова*

$$D(\lambda) = D(0) + \sum_{m=1}^{n} \lambda^m \sum_{i=m}^{n} c_{mi}\Delta^i D(0). \tag{3}$$

Для облегчения вычислений по формуле (2) составлены таблицы коэффициентов $c_{mi}$ [8].

В более общем случае, если в качестве узлов интерполирования взять числа $\lambda_i = a + ih$ $(i = 0, 1, \ldots, n)$, то формула (3) примет вид

$$D(\lambda) = D(a) + \sum_{m=1}^{n} (\lambda - a)^m \sum_{i=m}^{n} c_{mi} h^i \Delta^i D(a). \tag{4}$$

Хотя изложенный здесь *метод интерполяции* и требует трудоемких вычислений $n+1$ определителей $n$-го порядка, тем не менее этот метод удобен своей простой вычислительной схемой. Кроме того, он применим к развертыванию определителя более общего вида

$$F(\lambda) = \det[f_{ij}(\lambda)],$$

где $f_{ij}(\lambda)$ — целые полиномы от $\lambda$.

Пример. Пользуясь методом интерполяции, раскрыть характеристический определитель

$$D(\lambda)=\begin{vmatrix} 1-\lambda & 2 & 3 & 4 \\ 2 & 1-\lambda & 2 & 3 \\ 3 & 2 & 1-\lambda & 2 \\ 4 & 3 & 2 & 1-\lambda \end{vmatrix}$$

(ср. гл. XII, § 3, пример).

Решение. Последовательно вычислим $D(i)$ при $i=0, 1, 2, 3, 4$. Имеем:

$$D(0)=-20,\ D(1)=-119,\ D(2)=-308,$$
$$D(3)=-575,\ D(4)=-884.$$

Конечные разности $\Delta^i D(0)$ $(i=0, 1, 2, 3, 4)$ приведены в таблице 55.

Таблица 55

**Конечные разности чисел $D(\lambda)$**

| $\lambda$ | $D(\lambda)$ | $\Delta D(\lambda)$ | $\Delta^2 D(\lambda)$ | $\Delta^3 D(\lambda)$ | $\Delta^4 D(\lambda)$ |
|---|---|---|---|---|---|
| 0 | $-20$ | $-99$ | $-90$ | $12$ | $24$ |
| 1 | $-119$ | $-189$ | $-78$ | $36$ | |
| 2 | $-308$ | $-267$ | $-42$ | | |
| 3 | $-575$ | $-309$ | | | |
| 4 | $-884$ | | | | |

Так как

$$\frac{\lambda}{1!}=\lambda;$$
$$\frac{\lambda(\lambda-1)}{2!}=\frac{\lambda^2}{2}-\frac{\lambda}{2};$$
$$\frac{\lambda(\lambda-1)(\lambda-2)}{3!}=\frac{\lambda^3}{6}-\frac{\lambda^2}{2}+\frac{\lambda}{3};$$
$$\frac{\lambda(\lambda-1)(\lambda-2)(\lambda-3)}{4!}=\frac{\lambda^4}{24}-\frac{\lambda^3}{4}+\frac{11\lambda^2}{24}-\frac{\lambda}{4},$$

то из формулы (2) получаем:

$$c_{11}=1;$$
$$c_{22}=\frac{1}{2},\quad c_{12}=-\frac{1}{2};$$
$$c_{33}=\frac{1}{6},\quad c_{23}=-\frac{1}{2},\quad c_{13}=\frac{1}{3};$$
$$c_{41}=\frac{1}{24},\quad c_{34}=-\frac{1}{4},\quad c_{24}=\frac{11}{24},\quad c_{14}=-\frac{1}{4}.$$

Отсюда, применяя формулу Маркова (3), будем иметь:

$$D(\lambda)=D(0)+[c_{11}\Delta D(0)+c_{12}\Delta^2 D(0)+c_{13}\Delta^3 D(0)+c_{14}\Delta^4 D(0)]\lambda+$$
$$+[c_{22}\Delta^2 D(0)+c_{23}\Delta^3 D(0)+c_{24}\Delta^4 D(0)]\lambda^2+$$
$$+[c_{33}\Delta^3 D(0)+c_{34}\Delta^4 D(0)]\lambda^2+c_{44}\Delta^4 D(0)\lambda^4=$$
$$=-20+\left(-99\cdot 1+90\cdot\frac{1}{2}+12\cdot\frac{1}{3}-24\cdot\frac{1}{4}\right)\lambda+$$
$$+\left(-90\cdot\frac{1}{2}-12\cdot\frac{1}{2}+24\cdot\frac{11}{24}\right)\lambda^2+\left(12\cdot\frac{1}{6}-24\cdot\frac{1}{4}\right)\lambda^3+$$
$$+24\cdot\frac{1}{24}\lambda^4=-20-56\lambda-40\lambda^2-4\lambda^3+\lambda^4.$$

## § 24*. Интерполирование функций двух переменных

Пусть функция

$$z=f(x,\ y)$$

задана на системе равноотстоящих точек $(x_i,\ y_j)$ $(i,\ j=0,\ 1,\ 2,\ \ldots)$, где

$$x_i=x_0+ih,\quad y_i=y_0+jk,$$

причем

$$h=\Delta x_i=\text{const};\quad k=\Delta y_j=\text{const}.$$

Для краткости введем обозначения

$$z_{ij}=f(x_i,\ y_j).$$

Значения функции $z$ можно оформить в виде *таблицы с двумя входами* (таблица 56).

Таблица 56

**Значения функции двух переменных**

| $y \backslash x$ | $x_0$ | $x_1$ | $x_2$ | $\ldots$ |
|---|---|---|---|---|
| $y_0$ | $z_{00}$ | $z_{10}$ | $z_{20}$ | $\ldots$ |
| $y_1$ | $z_{01}$ | $z_{11}$ | $z_{21}$ | $\ldots$ |
| $y_2$ | $z_{02}$ | $z_{12}$ | $z_{22}$ | $\ldots$ |
| $\ldots$ | $\ldots$ | $\ldots$ | $\ldots$ | $\ldots$ |

Интерполирование функции двух переменных

$$z=f(x,\ y),$$

т. е. нахождение ее нетабличных значений, можно последовательно проводить по каждому переменному $x$ и $y$ в отдельности. Пусть, например, требуется найти значение

$$\bar{z}=f(\bar{x},\bar{y}).$$

Интерполируя надлежащим образом выбранные функции одной переменной $x$:

$$f_k(x) = f(x, y_k),$$

где $y_k \approx \bar{y}$, находим значения $f_k(\bar{x})$. Для этого используются соответствующие строки двойной таблицы. Рассматривая полученные значения $f_k(\bar{x}) = f(\bar{x}, y_k)$ как значения функции $f(\bar{x}, y)$ единственной переменной $y$, с помощью одной из интерполяционных формул находим искомое значение $f(\bar{x}, \bar{y}) = \bar{z}$.

Можно также производить интерполирование в обратном порядке.

Пример. Значения *функции последействия*

$$f(x, y) = \int_{-\infty}^{+\infty} e^{-y^2z^2 - z - xe^{-z}}\, dz$$

даются следующей таблицей (см. Янке и Эмде, «Таблицы функций»):

| $y$ \ $x$ | 0,4 | 0,7 | 1,0 |
|---|---|---|---|
| 0,00 | 2,500 | 1,429 | 1,000 |
| 0,05 | 2,487 | 1,419 | 0,995 |
| 0,10 | 2,456 | 1,400 | 0,981 |

Найти $f(0{,}5;\ 0{,}03)$.

Решение. Составляем таблицы 57а, 57б и 57в, используя строки данной двойной таблицы.

$y=0$ Таблица 57а

| $x$ | $f$ | $\Delta f$ | $\Delta^2 f$ |
|---|---|---|---|
| 0,4 | 2,500 | —1,071 | 0,642 |
| 0,7 | 1,429 | —0,429 | |
| 1,0 | 1,000 | | |

$y=0{,}05$ Таблица 57б

| $x$ | $f$ | $\Delta f$ | $\Delta^2 f$ |
|---|---|---|---|
| 0,4 | 2,487 | —1,068 | 0,644 |
| 0,7 | 1,419 | —0,424 | |
| 1,0 | 0,995 | | |

$y=0{,}10$ Таблица 57в

| $x$ | $f$ | $\Delta f$ | $\Delta^2 f$ |
|---|---|---|---|
| 0,4 | 2,456 | —1,056 | 0,637 |
| 0,7 | 1,400 | —0,419 | |
| 1,0 | 0,981 | | |

Так как для этих таблиц

$$h = 0{,}7 - 0{,}4 = 0{,}3,$$

то, принимая $x_0 = 0{,}4$, будем иметь:

$$q = \frac{x - x_0}{h} = \frac{0{,}5 - 0{,}4}{0{,}3} = \frac{1}{3}.$$

Отсюда, используя первую интерполяционную формулу Ньютона, последовательно получим:

$$f_0 = f(0{,}5;\ 0) = 2{,}500 - \frac{1}{3} \cdot 1{,}071 + \frac{\frac{1}{3}\left(-\frac{2}{3}\right)}{2} \cdot 0{,}642 = 2{,}072;$$

$$f_1 = f(0{,}5;\ 0{,}05) = 2{,}487 - \frac{1}{3} \cdot 1{,}068 - \frac{1}{9} \cdot 0{,}644 = 2{,}069;$$

$$f_2 = f(0{,}5;\ 0{,}10) = 2{,}456 - \frac{1}{3} \cdot 1{,}056 - \frac{1}{9} \cdot 0{,}637 = 2{,}033.$$

Составляем таблицу найденных значений (таблица 58).

Таблица 58

| $y$ | $f$ | $\Delta f$ | $\Delta^2 f$ |
|---|---|---|---|
| 0 | 2,072 | —0,003 | —0,033 |
| 0,05 | 2,069 | —0,036 | |
| 0,10 | 2,033 | | |

Принимая $k = 0{,}05 - 0 = 0{,}05$ и $y_0 = 0$, получим:

$$q' = \frac{0{,}03 - 0}{0{,}05} = \frac{3}{5}.$$

Отсюда

$$f(0{,}5;\ 0{,}03) = 2{,}072 - \frac{3}{5} \cdot 0{,}003 + \frac{\frac{3}{5} \cdot \left(-\frac{2}{5}\right)}{2} \cdot (-0{,}033) = 2{,}074.$$

## § 25*. Двойные разности высших порядков

Для функции $z = f(x, y)$, заданной двойной таблицей $\{z_{ij}\}$, можно определить *частные конечные разности*

$$\Delta_x z_{ij} = z_{i+1,j} - z_{ij} \quad \text{и} \quad \Delta_y z_{ij} = z_{i,j+1} - z_{ij}.$$

Повторно применяя эти операции, получим *двойные разности высших порядков*

$$\Delta^{m+n} z_{ij} = \Delta_{x^m y^n}^{m+n} z_{ij} = \Delta_{x^m}^{m} (\Delta_{y^n}^{n} z_{ij}) = \Delta_{y^n}^{n} (\Delta_{x^m}^{m} z_{ij}),$$

где положено $\Delta^{0+0}z_{ij}=z_{ij}$. Например,

$$\Delta^{1+2}z_{ij}=\Delta_x(\Delta^2_{yy}z_{ij})=\Delta_x(z_{i,j+2}-2z_{i,j+1}+z_{ij})=$$
$$=(z_{i+1,j+2}-2z_{i+1,j+1}+z_{i+1,j})-(z_{i,j+2}-2z_{i,j+1}+z_{ij}).$$

## § 26*. Интерполяционная формула Ньютона для функции двух переменных

Используя разности функции двух переменных $z=f(x, y)$, можно построить интерполяционный полином, аналогичный интерполяционному полиному Ньютона. Пусть $P(x, y)$ — целый полином такой, что

$$\Delta^{m+n}_{x^m y^n}P(x_0, y_0)=\Delta^{m+n}z_{00} \tag{1}$$

($m$, $n=0$, 1, 2, ...). Положим, что $P(x, y)$ разложен по обобщенным степеням разностей $x-x_0$ и $y-y_0$, т. е.

$$P(x, y)=c_{00}+c_{10}(x-x_0)+c_{01}(y-y_0)+c_{20}(x-x_0)(x-x_1)+$$
$$+c_{11}(x-x_0)(y-y_0)+c_{02}(y-y_0)(y-y_1)+\ldots \tag{2}$$

Полагая $x=x_0$ и $y=y_0$, в силу условия (1) будем иметь:

$$P(x_0, y_0)=z_{00}=c_{00}.$$

Составляя для полинома $P(x, y)$ конечные разности первого порядка, получим:

$$\Delta_x P(x, y)=c_{10}h+2c_{20}h(x-x_0)+c_{11}h(y-y_0)+\ldots$$

и

$$\Delta_y P(x, y)=c_{01}k+c_{11}k(x-x_0)+2c_{02}k(y-y_0)+\ldots$$

Отсюда, полагая $x=x_0$ и $y=y_0$, на основании условия (1) будем иметь:

$$\Delta_x P(x_0, y_0)=\Delta^{1+0}z_{00}=c_{10}h$$

и

$$\Delta_y P(x_0, y_0)=\Delta^{0+1}z_{00}=c_{01}k,$$

т. е.

$$c_{10}=\frac{\Delta^{1+0}z_{00}}{h},\quad c_{10}=\frac{\Delta^{0+1}z_{00}}{k}.$$

Далее, подсчитывая для полинома $P(x, y)$ конечные разности второго порядка, найдем:

$$\Delta_{xx}P(x, y)=2!c_{20}h^2+\ldots,$$
$$\Delta_{xy}P(x, y)=c_{11}hk+\ldots,$$
$$\Delta_{yy}P(x, y)=2!c_{02}k^2+\ldots$$

Отсюда при $x=x_0$ и $y=y_0$ получим:

$$\Delta_{xx}P(x_0, y_0)=\Delta^{2+0}z_{00}=2!c_{20}h^2,$$
$$\Delta_{xy}P(x_0, y_0)=\Delta^{1+1}z_{00}=c_{11}hk,$$
$$\Delta_{yy}P(x_0, y_0)=\Delta^{0+2}z_{00}=2!\,c_{02}k^2,$$

т. е.

$$c_{20}=\frac{1}{2!}\cdot\frac{\Delta^{2+0}z_{00}}{h^2},\quad c_{11}=\frac{\Delta^{1+1}z_{00}}{hk},\quad c_{02}=\frac{1}{2!}\cdot\frac{\Delta^{0+2}z_{00}}{k^2}.$$

Аналогично находятся дальнейшие коэффициенты разложения (2). Подставляя найденные значения коэффициентов в формулу (2), получим *интерполяционный полином для функции двух переменных*

$$P(x, y)=z_{00}+\left[\frac{\Delta^{1+0}z_{00}}{h}(x-x_0)+\frac{\Delta^{0+1}z_{00}}{k}(y-y_0)\right]+$$
$$+\frac{1}{2!}\left[\frac{\Delta^{2+0}z_{00}}{h^2}(x-x_0)^{[2]}+\right.$$
$$\left.+2\cdot\frac{\Delta^{1+1}z_{00}}{hk}(x-x_0)(y-y_0)+\frac{\Delta^{0+2}z_{00}}{k^2}(y-y_0)^{[2]}\right]+\ldots \qquad (3)$$

При интерполировании функции $f(x, y)$ полагают:

$$f(x, y)\approx P(x, y).$$

Для удобства вычислений обычно вводят переменные

$$\frac{x-x_0}{h}=p,\qquad \frac{y-y_0}{k}=q;$$

тогда

$$\frac{x-x_1}{h}=p-1,\quad \frac{y-y_1}{k}=q-1$$

и т. д. Отсюда формула (3) принимает вид

$$z\approx z_{00}+(p\Delta^{1+0}z_{00}+q\Delta^{0+1}z_{00})+$$
$$+\frac{1}{2!}[p(p-1)\Delta^{2+0}z_{00}+2pq\Delta^{1+1}z_{00}+q(q-1)\Delta^{0+2}z_{00}]+\ldots, \qquad (4)$$

где

$$x=x_0+ph,\quad y=y_0+qk.$$

Если положить $p=0$ или $q=0$, то формула (4) перейдет в соответствующую интерполяционную формулу Ньютона.

Пример. Пользуясь интерполяционной формулой (4), найти $f=f(0{,}5;\ 0{,}03)$ для функции $f(x, y)$, рассмотренной в примере из § 24.

Решение. Принимая $x_0=0{,}4$; $y_0=0$, составляем для функции $f$ таблицы конечных разностей первого порядка (таблицы 59а и 59б).

Таблица 59а

| | $\Delta^{1+0}f_{0j}$ | $\Delta^{1+0}f_{1j}$ |
|---|---|---|
| $j=0$ | —1,071 | —0,429 |
| $j=1$ | —1,068 | —0,424 |
| $j=2$ | —1,056 | —0,419 |

Таблица 59б

| | $i=0$ | $i=1$ | $i=2$ |
|---|---|---|---|
| $\Delta^{0+1}f_{i0}$ | —0,013 | —0,010 | —0,005 |
| $\Delta^{0+1}f_{i1}$ | —0,031 | —0,019 | —0,014 |

Отсюда находим конечные разности второго порядка

$$\Delta^{2+0}f_{00}=\Delta^{1+0}f_{10}-\Delta^{1+0}f_{00}=-0{,}429-(-1{,}071)=0{,}642;$$
$$\Delta^{1+1}f_{00}=\Delta^{1+0}f_{01}-\Delta^{1+0}f_{00}=-1{,}068-(-1{,}071)=0{,}003$$

или

$$\Delta^{1+1}f_{00}=\Delta^{0+1}f_{10}-\Delta^{0+1}f_{00}=-0{,}010-(-0{,}013)=0{,}003;$$
$$\Delta^{0+2}f_{00}=\Delta^{0+1}f_{01}-\Delta^{0+1}f_{00}=-0{,}031-(-0{,}013)=-0{,}018.$$

Так как

$$p=\frac{x-x_0}{h}=\frac{1}{3};\qquad q=\frac{y-y_0}{k}=\frac{3}{5},$$

то, применяя формулу (4), получим:

$$f=2{,}500+\frac{1}{3}\cdot(-1{,}071)+\frac{3}{5}\cdot(-0{,}013)+$$
$$+\frac{1}{2}\left[\frac{1}{3}\cdot\left(-\frac{2}{3}\right)\cdot 0{,}642+2\cdot\frac{1}{3}\cdot\frac{3}{5}\cdot 0{,}003+\right.$$
$$\left.+\frac{3}{5}\cdot\left(-\frac{2}{5}\right)(-0{,}018)\right]=2{,}500-0{,}357-0{,}0078-0{,}0713+$$
$$+0{,}0006+0{,}0021=2{,}067.$$

Сравнивая с ответом $f=2{,}074$, полученным первым способом, мы видим, что цифры тысячных являются ненадежными.

## Литература к четырнадцатой главе

1. Э. Уиттекер и Г. Робинсон, Математическая обработка результатов наблюдений, ГТТИ, М.—Л., 1933, гл. I.
2. В. Л. Гончаров, Теория интерполирования и приближения функций, ГТТИ, М.—Л., 1934, гл. I, §§ 18—21.
3. Дж. Скарборо, Численные методы математического анализа, ГТТИ, М.—Л., 1934, IV, разд. II.
4. В. М. Брадис, Теория и практика вычислений, Учпедгиз, М., 1935, гл. IX.
5. В. Э. Милн. Численный анализ, ИЛ, 1951, гл. III, VI.
6. Е. Я. Ремез, Общие вычислительные методы чебышевского приближения, Изд. АН УССР, 1957, ч. 1, гл. I.
7. Математический практикум на счетно-вычислительных приборах и инструментах. Под общей редакцией Н. А. Леднева, «Советская наука», М., 1959, гл. III.
8. В. Н. Фаддеева, Вычислительные методы линейной алгебры, Гостехиздат, М.—Л., 1950, гл. III, § 27.

## ГЛАВА XV

# ПРИБЛИЖЕННОЕ ДИФФЕРЕНЦИРОВАНИЕ

### § 1. Постановка вопроса

При решении практических задач часто нужно найти производные указанных порядков от функции $y=f(x)$, заданной таблично. Возможно также, что в силу сложности аналитического выражения функции $f(x)$ непосредственное дифференцирование ее затруднительно. В этих случаях обычно прибегают к *приближенному дифференцированию*.

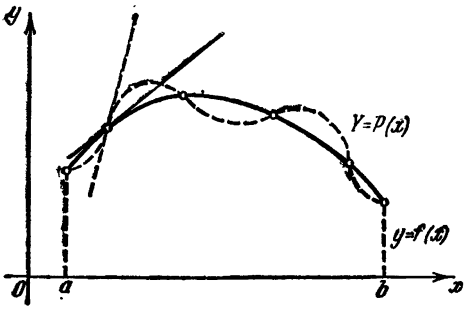


Рис. 65.

Для вывода формул приближенного дифференцирования заменяют данную функцию $f(x)$ на интересующем отрезке $[a, b]$ интерполирующей функцией $P(x)$ (чаще всего полиномом), а затем полагают:

$$f'(x)=P'(x) \qquad (1)$$

при

$$a \leqslant x \leqslant b.$$

Аналогично поступают при нахождении производных высших порядков функции $f(x)$.

Если для интерполирующей функции $P(x)$ известна погрешность

$$R(x)=f(x)-P(x),$$

то погрешность производной $P'(x)$ выражается формулой

$$r(x)=f'(x)-P'(x)=R'(x), \qquad (2)$$

т. е. *погрешность производной интерполирующей функции равна производной от погрешности этой функции*. То же самое справедливо и для производных высших порядков.

Следует отметить, что, вообще говоря, приближенное дифференцирование представляет собой операцию менее точную, чем интерполирование. Действительно, близость друг к другу ординат двух кривых

$$y=f(x) \quad \text{и} \quad Y=P(x)$$

на отрезке $[a, b]$ еще не гарантирует близости на этом отрезке их производных $f'(x)$ и $P'(x)$, т. е. малого расхождения угловых коэффициентов касательных к рассматриваемым кривым при одинаковых значениях аргумента (рис. 65).

## § 2. Формулы приближенного дифференцирования, основанные на первой интерполяционной формуле Ньютона

Пусть имеем функцию $y(x)$, заданную в равноотстоящих точках $x_i$ $(i=0, 1, 2, \ldots, n)$ отрезка $[a, b]$ с помощью значений $y_i=f(x_i)$. Для нахождения на $[a, b]$ производных $y'=f'(x)$, $y''=f''(x)$ и т. д.*) функцию $y$ приближенно заменим интерполяционным полиномом Ньютона, построенным для системы узлов $x_0, x_1, \ldots, x_k$ $(k \leqslant n)$.

Имеем:

$$y(x)=y_0+q\Delta y_0+\frac{q(q-1)}{2!}\Delta^2 y_0+\frac{q(q-1)(q-2)}{3!}\Delta^3 y_0+ \\ +\frac{q(q-1)(q-2)(q-3)}{4!}\Delta^4 y_0+\ldots, \qquad (1)$$

где

$$q=\frac{x-x_0}{h} \quad \text{и} \quad h=x_{i+1}-x_i \quad (i=0, 1, \ldots).$$

Производя перемножение биномов, получим:

$$y(x)=y_0+q\Delta y_0+\frac{q^2-q}{2}\Delta^2 y_0+\frac{q^3-3q^2+2q}{6}\Delta^3 y_0+ \\ +\frac{q^4-6q^3+11q^2-6q}{24}\Delta^4 y_0+\ldots \qquad (1')$$

Так как

$$\frac{dy}{dx}=\frac{dy}{dq}\cdot\frac{dq}{dx}=\frac{1}{h}\frac{dy}{dq},$$

то

$$y'(x)=\frac{1}{h}\left[\Delta y_0+\frac{2q-1}{2}\Delta^2 y_0+\frac{3q^2-6q+2}{6}\Delta^3 y_0+ \right. \\ \left. +\frac{2q^3-9q^2+11q-3}{12}\Delta^4 y_0+\ldots\right]. \qquad (2)$$

Аналогично, так как

$$y''(x)=\frac{d(y')}{dx}=\frac{d(y')}{dq}\cdot\frac{dq}{dx},$$

то

$$y''(x)=\frac{1}{h^2}\left[\Delta^2 y_0+(q-1)\Delta^3 y_0+\frac{6q^2-18q+11}{12}\Delta^4 y_0+\ldots\right]. \qquad (3)$$

---

*) Само собой разумеется, что заранее должно быть известно о существовании соответствующих производных функции $f(x)$, иначе выкладки носят иллюзорный характер.

Таким же способом в случае надобности можно вычислить и производные функции $y(x)$ любого порядка.

Заметим, что при нахождении производных $y'(x)$, $y''(x)$, ... в фиксированной точке $x$ в качестве $x_0$ следует выбирать ближайшее табличное значение аргумента.

Иногда требуется находить производные функции $y$ в основных табличных точках $x_i$. В этом случае формулы численного дифференцирования упрощаются. Так как каждое табличное значение можно считать за начальное, то положим $x=x_0$, $q=0$; тогда будем иметь:

$$y'(x_0)=\frac{1}{h}\left(\Delta y_0-\frac{\Delta^2 y_0}{2}+\frac{\Delta^3 y_0}{3}-\frac{\Delta^4 y_0}{4}+\frac{\Delta^5 y_0}{5}-\ldots\right) \tag{4}$$

и

$$y''(x_0)=\frac{1}{h^2}\left(\Delta^2 y_0-\Delta^3 y_0+\frac{11}{12}\Delta^4 y_0-\frac{5}{6}\Delta^5 y_0+\ldots\right). \tag{5}$$

Если $P_k(x)$ — интерполяционный полином Ньютона, содержащий разности $\Delta y_0$, $\Delta^2 y_0$, ..., $\Delta^k y_0$, и

$$R_k(x)=y(x)-P_k(x)$$

— соответствующая погрешность, то погрешность в определении производной есть

$$R'_k(x)=y'(x)-P'_k(x).$$

Как известно (гл. XIV, § 15),

$$R_k(x)=\frac{(x-x_0)(x-x_1)\ldots(x-x_k)}{(k+1)!}y^{(k+1)}(\xi)=$$
$$=h^{k+1}\frac{q(q-1)\ldots(q-k)}{(k+1)!}y^{(k+1)}(\xi),$$

где $\xi$ — некоторое промежуточное число между значениями $x_0$, $x_1$, ... ..., $x_k$, $x$. Поэтому, предполагая, что $y(x)\in C^{(k+2)}$, получим:

$$R'_k(x)=\frac{dR_k}{dq}\cdot\frac{dq}{dx}=\frac{h^k}{(k+1)!}\left\{y^{(k+1)}(\xi)\frac{d}{dq}[q(q-1)\ldots(q-k)]+\right.$$
$$\left.+q(q-1)\ldots(q-k)\frac{d}{dq}[y^{(k+1)}(\xi)]\right\}.$$

Отсюда при $x=x_0$ и, следовательно, при $q=0$ и учитывая, что $\frac{d}{dq}[q(q-1)\ldots(q-k)]_{q=0}=(-1)^k k!$, будем иметь:

$$R'_k(x_0)=(-1)^k\frac{h^k}{k+1}y^{(k+1)}(\xi). \tag{6}$$

Так как $y^{(k+1)}(\xi)$ во многих случаях трудно оценить, то при $h$ малом приближенно полагают:

$$y^{(k+1)}(\xi) \approx \frac{\Delta^{k+1} y_0}{h^{k+1}}$$

и, следовательно,

$$R'_k(x_0) \approx \frac{(-1)^k}{h} \frac{\Delta^{k+1} y_0}{k+1}. \tag{7}$$

Аналогично может быть найдена погрешность $R''_k(x_0)$ для второй производной $y''(x_0)$.

Пример 1. Найти $y'(50)$ функции $y = \lg x$, заданной таблично (таблица 60).

Таблица 60

**Значения функции $y = \lg x$**

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ | $\Delta^3 y$ |
|---|---|---|---|---|
| 50 | 1,6990 | 414 | —36 | 5 |
| 55 | 1,7404 | 378 | —31 | |
| 60 | 1,7782 | 347 | | |
| 65 | 1,8129 | | | |

Решение. Здесь $h = 5$. Дополняем таблицу 60 столбцами конечных разностей (десятичные разряды, как обычно, не указываются; они определяются десятичными разрядами значений функции).

Используя первую строчку таблицы, на основании формулы (4), с точностью до разностей третьего порядка, будем иметь:

$$y'(50) = \frac{1}{5}(0{,}0414 + 0{,}0018 + 0{,}0002) = 0{,}0087.$$

Для оценки точности найденного значения заметим, что так как табулированная выше функция есть $y = \lg x$, то

$$y'_x = \frac{M}{x} = \frac{0{,}43429}{x}.$$

Следовательно,

$$y'(50) = \frac{0{,}43429}{50} = 0{,}0087.$$

Таким образом, результаты совпадают с точностью до четвертого десятичного знака.

Пример 2. Путь $y=f(t)$, пройденный прямолинейно движущейся точкой за время $t$, дается следующей таблицей [1]:

| $i$ | Время $t_i$ в сек. | Путь $y(t_i)$ в *см* |
|---|---|---|
| 0 | 0,00 | 0,000 |
| 1 | 0,01 | 1,519 |
| 2 | 0,02 | 6,031 |
| 3 | 0,03 | 13,397 |
| 4 | 0,04 | 23,396 |
| 5 | 0,05 | 35,721 |
| 6 | 0,06 | 50,000 |
| 7 | 0,07 | 65,798 |
| 8 | 0,08 | 82,635 |
| 9 | 0,09 | 100,000 |

Используя конечные разности до пятого порядка включительно, приближенно найти скорость $V=\frac{dy}{dt}$ и ускорение $W=\frac{d^2y}{dt^2}$ точки для моментов $t=0$; 0,01; 0,02; 0,03; 0,04.

Решение. Составляем таблицу разностей (таблица 61).

Таблица 61

**Конечные разности функции $y=f(t)$**

| $i$ | $\Delta y_i$ | $\Delta^2 y_i$ | $\Delta^3 y_i$ | $\Delta^4 y_i$ | $\Delta^5 y_i$ |
|---|---|---|---|---|---|
| 0 | 1,519 | 2,993 | —0,139 | —0,082 | —0,004 |
| 1 | 4,512 | 2,854 | —0,221 | —0,086 | 0,021 |
| 2 | 7,366 | 2,633 | —0,307 | —0,065 | 0,002 |
| 3 | 9,999 | 2,326 | —0,372 | —0,063 | 0,018 |
| 4 | 12,325 | 1,954 | —0,435 | —0,045 | 0,014 |
| 5 | 14,279 | 1,519 | —0,480 | —0,031 | — |
| 6 | 15,798 | 1,039 | —0,511 | — | |
| 7 | 16,837 | 0,528 | — | | |
| 8 | 17,365 | — | | | |
| 9 | — | | | | |

Полагая $h=0{,}01$ и применяя формулы (4) и (5), получаем приближенные значения величины скорости $V$ (*см/сек*) и величины ускорения $W$ (*см/сек*$^2$). Например,

$$V(0)=100\,(1{,}519-1{,}496-0{,}046+0{,}020-0{,}001)=-0{,}4\ \textit{см/сек},$$
$$W(0)=10\,000\,(2{,}993+0{,}139-0{,}075+0{,}003)=30\,600\ \textit{см/сек}^2.$$

Соответствующие значения $V$ и $W$ помещены в таблице 62.

Таблица 62

**Значения скорости $V$ и ускорения $W$ для закона движения $y=f(t)$**

| $t$ | $V$ | $W$ | $\tilde{V}$ | $\tilde{W}$ |
|---|---|---|---|---|
| 0,00 | 0,4 | 30600 | 0,00 | 30462 |
| 0,01 | 303,6 | 29780 | 303,08 | 30001 |
| 0,02 | 596,3 | 28780 | 596,98 | 28625 |
| 0,03 | 873,2 | 26250 | 872,66 | 26381 |
| 0,04 | 1121,7 | 23360 | 1121,9 | 23340 |

Заметим, что табулированный закон движения дается формулой

$$y=100\left(1-\cos\frac{50\pi t}{9}\right).$$

Отсюда

$$V=\frac{dy}{dt}=\frac{5000\pi}{9}\sin\frac{50\pi t}{9}$$

и

$$W=\frac{d^2y}{dt^2}=\frac{250000\pi^2}{81}\cos\frac{50\pi t}{9}.$$

Для сравнения точные значения $\tilde{V}$ и $\tilde{W}$ приведены в правой половине таблицы 62.

Отметим, что можно вывести также формулы приближенного дифференцирования, исходя из второй интерполяционной формулы Ньютона.

## § 3. Формулы приближенного дифференцирования, основанные на формуле Стирлинга

Выведенные в § 2 формулы численного дифференцирования для функции $y$ в точке $x=x_0$ обладают тем недостатком, что они используют лишь односторонние значения функции при $x>x_0$. Относительно бóльшую точность имеют симметрические формулы дифференцирования, учитывающие значения данной функции $y$ как при $x>x_0$, так и при $x<x_0$. Эти формулы обычно называются *центральными формулами дифференцирования*. Мы ограничимся выводом одной из них, взяв за основу интерполяционную формулу Стирлинга.

Пусть $\ldots, x_{-3}, x_{-2}, x_{-1}, x_0, x_1, x_2, x_3, \ldots$ — система равноотстоящих точек с шагом $x_{i+1}-x_i=h$ и $y_i=f(x_i)$ — соответствующие значения данной функции $y=f(x)$. Полагая

$$q=\frac{x-x_0}{h}$$

и заменяя приближенно функцию $y$ интерполяционным полиномом Стирлинга, будем иметь:

$$y(x)=y_0+q\Delta y_{-\frac{1}{2}}+\frac{q^2}{2!}\Delta^2 y_{-1}+\frac{q(q^2-1)}{3!}\Delta^3 y_{-\frac{3}{2}}+$$
$$+\frac{q^2(q^2-1)}{4!}\Delta^4 y_{-2}+\frac{q(q^2-1)(q^2-2^2)}{5!}\Delta^5 y_{-\frac{5}{2}}+$$
$$+\frac{q^2(q^2-1)(q^2-2^2)}{6!}\Delta^6 y_{-3}+\ldots, \qquad (1)$$

где для краткости введены обозначения

$$\Delta y_{-\frac{1}{2}}=\frac{\Delta y_{-1}+\Delta y_0}{2},$$
$$\Delta^3 y_{-\frac{3}{2}}=\frac{\Delta^3 y_{-2}+\Delta^3 y_{-1}}{2},$$
$$\Delta^5 y_{-\frac{5}{2}}=\frac{\Delta^5 y_{-3}+\Delta^5 y_{-2}}{2}$$

и т. д.

Из формулы (1), учитывая, что

$$\frac{dq}{dx}=\frac{1}{h},$$

получаем:

$$y'(x)=\frac{1}{h}\left(\Delta y_{-\frac{1}{2}}+q\Delta^2 y_{-1}+\frac{3q^2-1}{6}\Delta^3 y_{-\frac{3}{2}}+\frac{2q^3-q}{12}\Delta^4 y_{-2}+\right.$$
$$\left.+\frac{5q^4-15q^2+4}{120}\Delta^5 y_{-\frac{5}{2}}+\frac{3q^5-10q^3+4q}{360}\Delta^6 y_{-3}+\ldots\right), \qquad (2)$$
$$y''(x)=\frac{1}{h^2}\left(\Delta^2 y_{-1}+q\Delta^3 y_{-\frac{3}{2}}+\frac{6q^2-1}{12}\Delta^4 y_{-2}+\right.$$
$$\left.+\frac{2q^3-3q}{12}\Delta^4 y_{-\frac{5}{2}}+\frac{15q^4-30q^2+4}{360}\Delta^6 y_{-3}+\ldots\right). \qquad (2')$$

В частности, полагая $q=0$, будем иметь:

$$y'(x_0)=\frac{1}{h}\left(\Delta y_{-\frac{1}{2}}-\frac{1}{6}\Delta^3 y_{-\frac{3}{2}}+\frac{1}{30}\Delta^5 y_{-\frac{5}{2}}+\ldots\right) \qquad (3)$$

и

$$y''(x_0)=\frac{1}{h^2}\left(\Delta^2 y_{-1}-\frac{1}{12}\Delta^4 y_{-2}+\frac{1}{90}\Delta^6 y_{-3}+\ldots\right). \qquad (3')$$

Пример 1. Найти $y'(1)$ и $y''(1)$ для функции $y=y(x)$, заданной таблицей 63.

Таблица 63

**Значения функции $y=y(x)$**

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ | $\Delta^3 y$ | $\Delta^4 y$ |
|---|---|---|---|---|---|
| 0,96 | 0,7825361 | | | | |
| | | —86029 | | | |
| 0,98 | 0,7739332 | | —1326 | | |
| | | —87355 | | 25 | |
| 1,00 | 0,7651977 | | —1301 | | 1 |
| | | —88656 | | 26 | |
| 1,02 | 0,7563321 | | —1275 | | |
| | | —89931 | | | |
| 1,04 | 0,7473390 | | | | |

Решение. Составляя разности функции $y$ (таблица 63) и используя подчеркнутые члены, на основании формулы (3) будем иметь:

$$y'(1)=\frac{1}{0{,}02}\left(-\frac{87\,355+88\,656}{2}\cdot 10^{-7}-\frac{1}{6}\cdot\frac{25+26}{2}\cdot 10^{-7}+\frac{1}{30}\cdot 1\cdot 10^{-7}\right)=$$
$$=-50\cdot(88\,005{,}5+4{,}2+0)\cdot 10^{-7}=-0{,}4400485.$$

Для проверки заметим, что табулированная функция есть функция Бесселя нулевого индекса $y=J_0(x)$.

Как известно,

$$J_0'(1)=-J_1(x)\,|_{x=1}=-0{,}4400506.$$

Аналогично, используя дважды подчеркнутые члены и применяя формулу (3′), будем иметь:

$$y''(1)=\frac{1}{0{,}02^2}\cdot\left(-1301\cdot 10^{-7}-\frac{1}{12}\cdot 1\cdot 10^{-7}\right)=$$
$$=-2500\cdot 1301\cdot 10^{-7}=-3{,}2525\cdot 10^{-1}=-0{,}325250.$$

Для сравнения приведем получающееся на основании соотношений между бесселевыми функциями точное значение

$$y''(1)=J_0''(1)=J_1(1)-J_0(1)=$$
$$=0{,}4400506-0{,}7651977=-0{,}325147.$$

Таким образом, численное нахождение второй производной есть операция, вообще говоря, менее надежная, чем первой.

Замечание. Иногда требуется найти экстремум дифференцируемой функции $y=y(x)$, заданной таблично. Для этого необходимо,

чтобы в точке экстремума $\tilde{x}$ было выполнено равенство $y'(\tilde{x})=0$. Приравняв нулю производную $y'(x)$ в формуле (2), методом последовательных приближений находим соответствующее значение $q$. Отсюда

$$\tilde{x}=x_0+qh,$$

причем значение $\tilde{y}$ вычисляется по формуле (1) или по какой-нибудь другой интерполяционной формуле. Найденное значение $\tilde{y}$ является экстремумом функций, если в окрестности точки $\tilde{x}$ вторая разность $\Delta^2 y$ сохраняет постоянный знак.

Пример 2. Найти нуль производной функции $y=J_1(x)$, заданной таблицей 64.

Таблица 64

**Значение функции $y=J_1(x)$**

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ | $\Delta^3 y$ |
|---|---|---|---|---|
| 1,80 | 0,5815170 | | | |
| | | 2561 | | |
| 1,82 | 0,5817731 | | —1643 | |
| | | 918 | | 2 |
| 1,84 | 0,5818649 | | —1641 | |
| | | —723 | | 4 |
| 1,86 | 0,5817926 | | —1637 | |
| | | —2360 | | 2 |
| 1,88 | 0,5815566 | | —1635 | |
| | | —3995 | | |
| 1,90 | 0,5811571 | | | |

Дополняем таблицу 64 конечными разностями функции $y$.

Решение. Принимаем $x_0=1{,}84$. Используя подчеркнутые разности, на основании формулы (2) будем иметь:

$$0=\frac{918-723}{2}+q(-1641)+\frac{3q^2-1}{6}\cdot\frac{2+4}{2}$$

или

$$0=97-1641q+\frac{3}{2}q^2.$$

Отсюда

$$q=\frac{97}{1641}+\frac{1}{1094}q^2. \tag{4}$$

Отбрасывая малый нелинейный член, получим первое приближение:

$$q^{(1)}=\frac{97}{1641}=5{,}911\cdot10^{-2}.$$

Уточняя это значение, из формулы (4) получим второе приближение:

$$q^{(2)}=q^{(1)}+\frac{1}{1094}[q^{(1)}]^2=5{,}911\cdot10^{-2}+\frac{1}{1094}\cdot3{,}494\cdot10^{-3}=$$
$$=5{,}911\cdot10^{-2}+3{,}2\cdot10^{-6}=5{,}911\cdot10^{-2}.$$

Следовательно, можно положить:

$$q=0{,}05911.$$

Отсюда

$$x=x_0+qh=1{,}84+0{,}05911\cdot0{,}02=1{,}8411822.$$

Таким образом,

$$J_1'(1{,}8411822)=0.$$

## § 4. Формулы численного дифференцирования для равноотстоящих точек, выраженные через значения функции в этих точках

Пусть точки $x_0, x_1, x_2, \ldots, x_n$ — равноотстоящие, т. е.

$$x_{i+1}-x_i=h \qquad (i=0,\ 1,\ 2,\ \ldots,\ n-1),$$

и пусть для функции $y=y(x)$ известны значения $y_i=y(x_i)$ $(i=0, 1, \ldots, n)$. Для данной системы узлов $x_i$ построим интерполяционный полином Лагранжа (гл. XIV, § 12)

$$L_n(x)=\sum_{i=0}^{n}\frac{\Pi_{n+1}(x)\,y_i}{(x-x_i)\,\Pi'_{n+1}(x_i)},$$

где

$$\Pi_{n+1}(x)=(x-x_0)(x-x_1)\ldots(x-x_n).$$

Тогда

$$L_n(x_i)=y_i \qquad (i=0,\ 1,\ \ldots,\ n).$$

Полагая

$$\frac{x-x_0}{h}=q,$$

получим:

$$\Pi_{n+1}(x)=h^{n+1}q(q-1)\ldots(q-n)=h^{n+1}q^{[n+1]}$$

и

$$\Pi'_{n+1}(x_i)=(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)=$$
$$=h^n i(i-1)\ldots1(-1)\ldots[-(n-i)]=(-1)^{n-i}h^n i!\,(n-i)! \qquad (1)$$

Следовательно, для полинома Лагранжа $L_n(x)$ имеем выражение

$$L_n(x)=\sum_{i=0}^{n}\frac{(-1)^{n-i}y_i}{i!\,(n-i)!}\cdot\frac{q^{[n+1]}}{q-i}. \qquad (2)$$

Отсюда, учитывая, что

$$\frac{dx}{dq}=h,$$

получаем:

$$y'(x)\approx L_n'(x)=\frac{1}{h}\sum_{i=0}^{n}\frac{(-1)^{n-i}y_i}{i!\,(n-i)!}\frac{d}{dq}\left\{\frac{q^{[n+1]}}{q-i}\right\}. \tag{3}$$

Аналогично могут быть найдены производные высших порядков данной функции $y(x)$. Для оценки погрешности

$$r_n(x)=y'(x)-L_n'(x)$$

воспользуемся известной формулой погрешности интерполяционной формулы (2) (гл. XIV, § 14)

$$R_n(x)=y(x)-L_n(x)=\frac{y^{(n+1)}(\xi)}{(n+1)!}\Pi_{n+1}(x), \tag{4}$$

где $\xi=\xi(x)$ — промежуточное значение между точками $x_0, x_1, \ldots, x_n$ и $x$.

Предполагая, что $y(x)\in C^{(n+2)}$, выводим:

$$r_n(x)=R_n'(x)=$$
$$=\frac{1}{(n+1)!}\left\{y^{(n+1)}(\xi)\,\Pi_{n+1}'(x)+\Pi_{n+1}(x)\frac{d}{dx}[y^{(n+1)}(\xi)]\right\}.$$

Отсюда, учитывая формулу (1), получаем погрешность производной в узлах

$$R_n'(x_i)=(-1)^{n-i}h^n\frac{i!\,(n-i)!}{(n+1)!}y^{(n+1)}(\xi), \tag{5}$$

где $\xi$ — промежуточное значение между $x_0, x_1, \ldots, x_n$.

I. Произведем расчет для $n=2$ (три точки). Из формулы (2) получаем:

$$L_2(x)=\frac{1}{2}y_0(q-1)(q-2)-y_1q(q-2)+\frac{1}{2}y_2q(q-1).$$

Отсюда, учитывая, что $\frac{dx}{dq}=h$, будем иметь:

$$y'(x)\approx L_2'(x)=\frac{1}{h}\left[\frac{1}{2}y_0(2q-3)-y_1(2q-2)+\frac{1}{2}y_2(2q-1)\right].$$

В частности, для производных

$$y'(x_i)=y_i' \quad (i=0,\ 1,\ 2)$$

получим следующие выражения:

$$y'_0 = \frac{1}{2h}(-3y_0 + 4y_1 - y_2);$$
$$y'_1 = \frac{1}{2h}(-y_0 + y_2);$$
$$y'_2 = \frac{1}{2h}(y_0 - 4y_1 + 3y_2)$$

с соответствующими погрешностями:

$$r_0 = \frac{1}{3}h^2 y'''(\xi_0);$$
$$r_1 = -\frac{1}{6}h^2 y'''(\xi_1);$$
$$r_2 = \frac{1}{3}h^2 y'''(\xi_2).$$

Приведем без доказательства формулы дифференцирования для четырех и пяти точек [3], справедливость которых читатель легко может проверить самостоятельно.

II. $n = 3$ (четыре точки).

$$y'_0 = \frac{1}{6h}(-11y_0 + 18y_1 - 9y_2 + 2y_3) - \frac{h^3}{4}y^{(4)}(\xi);$$
$$y'_1 = \frac{1}{6h}(-2y_0 - 3y_1 + 6y_2 - y_3) + \frac{h^3}{12}y^{(4)}(\xi);$$
$$y'_2 = \frac{1}{6h}(y_0 - 6y_1 + 3y_2 + 2y_3) - \frac{h^3}{12}y^{(4)}(\xi);$$
$$y'_3 = \frac{1}{6h}(-2y_0 + 9y_1 - 18y_2 + 11y_3) + \frac{h^3}{4}y^4(\xi).$$

III. $n = 4$ (пять точек).

$$y'_0 = \frac{1}{12h}(-25y_0 + 48y_1 - 36y_2 + 16y_3 - 3y_4) + \frac{h^4}{5}y^{(5)}(\xi),$$
$$y'_1 = \frac{1}{12h}(-3y_0 - 10y_1 + 18y_2 - 6y_3 + y_4) - \frac{h^4}{20}y^{(5)}(\xi),$$
$$y'_2 = \frac{1}{12h}(y_0 - 8y_1 + 8y_3 - y_4) + \frac{h^4}{30}y^{(5)}(\xi);$$
$$y'_3 = \frac{1}{12h}(-y_0 + 6y_1 - 18y_2 + 10y_3 + 3y_4) - \frac{h^4}{20}y^{(5)}(\xi);$$
$$y'_4 = \frac{1}{12h}(3y_0 - 16y_1 + 36y_2 - 48y_3 + 25y_4) + \frac{h^4}{4}y^5(\xi).$$

Рассмотрение формул I—III показывает, что если число точек нечетно и производная берется в средней точке, то соответствующая

формула численного дифференцирования выражается более просто и обладает повышенной точностью.

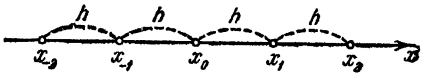


Рис. 66.

Ниже приводятся для случая $n=2$ и $n=4$ формулы таких *центральных производных* [3], причем для выявления симметрии изменена нумерация точек (рис. 66):

I. $n=2$.

$$y_0'=\frac{1}{2h}(y_2-y_{-1})-\frac{h^2}{6}y^{(3)}(\xi),$$

где $y_i=y(x_i)$ и $i=-1, 0, 1$;

II. $n=4$.

$$y_0'=\frac{2}{3h}(y_1-y_{-1})-\frac{1}{12h}(y_2-y_{-2})+\frac{h^4}{30}y^{(5)}(\xi),$$

где $y_i=y(x_i)$ и $i=-2, -1, 0, 1, 2$.

## § 5. Графическое дифференцирование

Задача графического дифференцирования заключается в построении по заданному графику функции $y=f(x)$ графика ее производной

$$Y=f'(x).$$

Пусть дан график функции $y=f(x)$ (рис. 67). Для построения в известном масштабе $l$ графика ее производной выбираем на данной кривой достаточно густую сеть точек 1, 2, 3, 4, 5, ..., включающую по возможности характерные для графика точки. В этих точках с возможной тщательностью строим касательные к графику функции, проводя их «на глаз». Далее, на оси $Ox$ выбираем точку $P(-l, 0)$ (полюс) и проводим параллельные соответствующим касательным прямые $P1'$, $P2'$, $P3'$, $P4'$, $P5'$, ... до пересечения их с осью $Oy$. Отрезки оси $Oy$: $01'$, $02'$, $03'$, $04'$, $05'$, ... представляют собой соответственно величины, пропорциональные значениям производной $y'=f'(x)$ в выбранных точках, т. е. являются ординатами графика производной. В самом деле, например, для точки 1 из рис. 67 имеем:

$$OA=l\operatorname{tg}\alpha_1=lf'(x_1).$$

Аналогичные результаты получаем для всех других точек. Поэтому точки пересечения $1''$, $2''$, $3''$, $4''$, $5''$, ... параллелей, проходящих через точки $1'$, $2'$, $3'$, $4'$, $5'$, ..., с соответствующими вертикалями, проходящими через точки касания 1, 2, 3, 4, 5, ..., принадлежат графику производной $y=lf'(x)$.

Соединяя точки 1″, 2″, 3″, 4″, 5″,... линией, характер которой учитывает положение промежуточных точек, мы приближенно получим график производной $y'$ в масштабе $l$. Если выбрать $l=1$, то график производной получится в натуральном масштабе.

Для увеличения точности графического построения рекомендуется сначала определять направление касательной, а затем намечать точку касания. Для этого график данной функции разбивают на небольшие участки, мало отличающиеся от прямолинейных. Рассмотрим один из таких участков $AB$ (рис. 68). Построим семейство хорд, параллельных секущей $AB$. Геометрическое место середин этих хорд представляет собой кривую $K$, пересекающую график функции в точке $C$, касательная в которой параллельна секущей $AB$. Таким приемом на каждом участке можно найти точку и соответствующее направление касательной. Дальнейшее построение выполняется указанным выше способом.

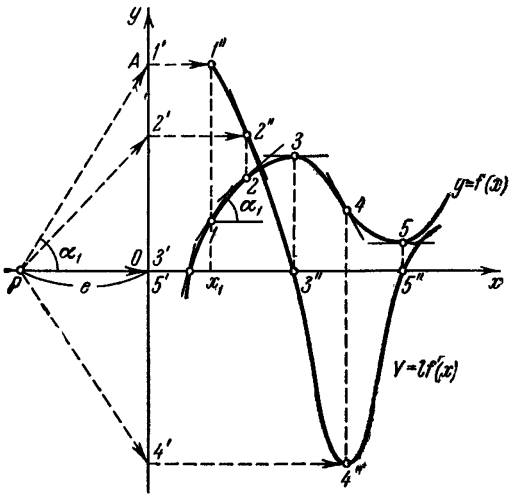


Рис. 67.

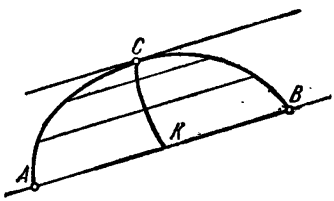


Рис. 68.

За указаниями более детального характера следует обратиться к специальной литературе (см., например, [5]).

## § 6*. Понятие о приближенном вычислении частных производных

Если функция $z=f(x, y)$ задана на прямоугольной сетке

$$x=x_0+ih; \quad y=y_0+jk$$

($i, j=0, 1, 2, \ldots$), то ее приближенно можно представить интерполяционной формулой (гл. XIV, § 26)

$$\begin{aligned} z = z_{00} &+ [p\Delta^{1+0} z_{00} + q\,\Delta^{0+1} z_{00}] + \\ &+ \frac{1}{2!}[p(p-1)\,\Delta^{2+0}\, z_{00} + 2pq\Delta^{1+1} z_{00} + q(q-1)\,\Delta^{0+2} z_{00}] + \\ &+ \frac{1}{3!}[p(p-1)(p-2)\,\Delta^{3+0} z_{00} + 3p(p-1)\,q\,\Delta^{2+1} z_{00} + \\ &+ 3pq(q-1)\,\Delta^{1+2} z_{00} + q(q-1)(q-2)\,\Delta^{0+3} z_{00}] + \ldots, \end{aligned} \tag{1}$$

где

$$p=\frac{x-x_0}{h}, \qquad q=\frac{y-y_0}{k}$$

и $\Delta^{m+n}z_{00}=\Delta_{x^m y^n}^{m+n}z\,(0,\ 0)$ — смешанные двойные разности.

Из формулы (1) легко находятся частные производные

$$\frac{\partial z}{\partial x}=\frac{\partial z}{\partial p}\cdot\frac{dp}{dx}=\frac{1}{h}\frac{\partial z}{\partial p},\quad \frac{\partial z}{\partial y}=\frac{\partial z}{\partial q}\cdot\frac{dq}{dy}=\frac{1}{k}\frac{\partial z}{\partial q}$$

и т. д.

### Литература к пятнадцатой главе

1. А. Н. Крылов, Лекции о приближенных вычислениях, Изд. 6, Гостехиздат, М., 1954, стр. 228.
2. Дж. Скарборо, Численные методы математического анализа, ГТТИ, М.—Л., 1934, гл. VII.
3. В. Э. Милн, Численный анализ, ИЛ, М., 1951, гл. IV.
4. Ш. Е. Микеладзе. Численные методы математического анализа, Гостехиздат, М., 1953, гл. XII.
5. К. Рунге, Графические методы математических вычислений, ГТТИ, М.—Л., 1932, гл. III, § 14.

## ГЛАВА XVI

# ПРИБЛИЖЕННОЕ ИНТЕГРИРОВАНИЕ ФУНКЦИЙ

## § 1. Общие замечания

Если функция $f(x)$ непрерывна на отрезке $[a, b]$ и известна ее первообразная $F(x)$, то определенный интеграл от этой функции в пределах от $a$ до $b$ может быть вычислен по *формуле Ньютона — Лейбница*

$$\int_a^b f(x)\,dx = F(b) - F(a), \tag{1}$$

где $F'(x) = f(x)$.

Однако во многих случаях первообразная функция $F(x)$ не может быть найдена с помощью элементарных средств или является слишком сложной; вследствие этого вычисление определенного интеграла по формуле (1) может быть затруднительным или даже практически невыполнимым.

Кроме того, на практике подынтегральная функция $f(x)$ часто задается таблично и тогда само понятие первообразной теряет смысл. Аналогичные вопросы возникают при вычислении кратных интегралов. Поэтому важное значение имеют приближенные и в первую очередь *численные методы* вычисления определенных интегралов.

Задача численного интегрирования функции заключается в вычислении значения определенного интеграла на основании ряда значений подынтегральной функции.

Численное вычисление однократного интеграла называется *механической квадратурой*, двойного — *механической кубатурой*. Соответствующие формулы мы будем называть *квадратурными* и *кубатурными* формулами.

Мы сначала остановимся на численном вычислении однократных интегралов. Обычный прием механической квадратуры состоит в том, что данную функцию $f(x)$ на рассматриваемом отрезке $[a, b]$ заменяют интерполирующей или аппроксимирующей функцией $\varphi(x)$ простого вида (например, полиномом), а затем приближенно

полагают:

$$\int_a^b f(x)\,dx = \int_a^b \varphi(x)\,dx. \tag{2}$$

Функция $\varphi(x)$ должна быть такова, чтобы интеграл $\int_a^b \varphi(x)\,dx$ вычислялся непосредственно.

Если функция $f(x)$ задана аналитически, то ставится вопрос об оценке погрешности формулы (2).

Рассмотрим более подробно применение для этой цели интерполяционного полинома Лагранжа (гл. XIV, § 12).

Пусть для функции $y=f(x)$ известны в $n+1$ точках $x_0$, $x_1$, $x_2$, ..., $x_n$ отрезка $[a, b]$ соответствующие значения

$$f(x_i)=y_i \quad (i=0,\ 1,\ 2,\ \ldots,\ n). \tag{3}$$

Требуется приближенно найти:

$$\int_a^b y\,dx = \int_a^b f(x)\,dx.$$

По заданным значениям $y_i$ построим полином Лагранжа

$$L_n(x)=\sum_{i=0}^{n} \frac{\Pi_{n+1}(x)}{(x-x_i)\,\Pi'_{n+1}(x_i)}\,y_i, \tag{4}$$

где

$$\Pi_{n+1}(x)=(x-x_0)(x-x_1)\ldots(x-x_n),$$

причем

$$L_n(x_i)=y_i \quad (i=0,\ 1,\ 2,\ \ldots,\ n).$$

Заменяя функцию $f(x)$ полиномом $L_n(x)$, получим равенство

$$\int_a^b f(x)\,dx = \int_a^b L_n(x)\,dx + R_n[f], \tag{5}$$

где $R_n[f]$ — ошибка квадратурной формулы (5) (*остаточный член*). Отсюда, воспользовавшись выражением (4), получаем приближенную квадратурную формулу

$$\int_a^b y\,dx = \sum_{i=0}^{n} A_i y_i, \tag{6}$$

где

$$A_i=\int_a^b \frac{\Pi_{n+1}(x)}{(x-x_i)\,\Pi'_{n+1}(x_i)}\,dx \quad (i=0,\ 1,\ 2,\ \ldots,\ n). \tag{7}$$

Если пределы интегрирования $a$ и $b$ являются узлами интерполирования, то квадратурная формула (6) называется «*замкнутого типа*», в противном случае — «*открытого типа*».

Для вычисления коэффициентов $A_i$ заметим, что

1) коэффициенты $A_i$ при данном расположении узлов не зависят от выбора функции $f(x)$;

2) для полинома степени $n$ формула (6) — точная, так как тогда $L_n(x) \equiv f(x)$; следовательно, в частности, формула (6) — точная при $y = x^k$ $(k = 0, 1, \ldots, n)$, т. е. $R_n[x^k] = 0$ при $k = 0, 1, \ldots, n$.

Полагая $y = x^k$ $(k = 0, 1, 2, \ldots, n)$ в формуле (6), получим линейную систему из $n+1$ уравнений

$$\left.\begin{aligned} I_0 &= \sum_{i=0}^{n} A_i, \\ I_1 &= \sum_{i=0}^{n} A_i x_i, \\ &\ldots\ldots\ldots\ldots \\ I_n &= \sum_{i=0}^{n} A_i x_i^n, \end{aligned}\right\} \tag{8}$$

где

$$I_k = \int_a^b x^k dx = \frac{b^{k+1} - a^{k+1}}{k+1} \quad (k = 0, 1, \ldots, n),$$

из которой можно определить коэффициенты $A_0, A_1, \ldots, A_n$ [1], [2]. Определитель системы (8) есть определитель Вандермонда

$$D = \prod_{i>j} (x_i - x_j) \neq 0.$$

Заметим, что при применении этого метода фактическое построение полинома Лагранжа $L_n(x)$ является излишним.

Простой метод подсчета погрешностей квадратурных формул разработан С. М. Никольским [3].

Пример. Вывести квадратурную формулу вида

$$\int_0^1 y\, dx = A_0 y\left(\frac{1}{4}\right) + A_1 y\left(\frac{1}{2}\right) + A_2 y\left(\frac{3}{4}\right). \tag{9}$$

Решение. Полагая в формуле (9)

$$y = x^k \quad (k = 0, 1, 2)$$

и учитывая, что

$$\int_0^1 dx = 1, \qquad \int_0^1 x\, dx = \frac{1}{2}, \qquad \int_0^1 x^2 dx = \frac{1}{3},$$

получим систему

$$\left.\begin{aligned} 1 &= A_0 + A_1 + A_2, \\ \frac{1}{2} &= \frac{1}{4} A_0 + \frac{1}{2} A_1 + \frac{3}{4} A_2, \\ \frac{1}{3} &= \frac{1}{16} A_0 + \frac{1}{4} A_1 + \frac{9}{16} A_2. \end{aligned}\right\}$$

Отсюда

$$A_0 = \frac{2}{3}, \quad A_1 = -\frac{1}{3}, \quad A_2 = \frac{2}{3}$$

и, следовательно,

$$\int_0^1 y dx = \frac{2}{3} y\left(\frac{1}{4}\right) - \frac{1}{3} y\left(\frac{1}{2}\right) + \frac{2}{3} y\left(\frac{3}{4}\right). \tag{10}$$

Квадратурная формула (10) открытого типа и является точной для всех полиномов степени не выше второй. Нетрудно убедиться, что формула (10) дает правильный результат и при $y = x^3$. Поэтому эта формула является точной также для полиномов третьей степени.

## § 2. Квадратурные формулы Ньютона — Котеса

Пусть для данной функции $y = f(x)$ требуется вычислить интеграл

$$\int_a^b y\, dx.$$

Выбрав шаг

$$h = \frac{b-a}{n},$$

разобьем отрезок $[a, b]$ с помощью равноотстоящих точек

$$x_0 = a, \quad x_i = x_0 + ih \quad (i = 1, 2, \ldots, n-1), \quad x_n = b$$

на $n$ равных частей, и пусть

$$y_i = f(x_i) \quad (i = 0, 1, 2, \ldots, n).$$

Заменяя функцию $y$ соответствующим интерполяционным полиномом Лагранжа $L_n(x)$, получим приближенную квадратурную формулу

$$\int_{x_0}^{x_n} y dx = \sum_{i=0}^{n} A_i y_i, \tag{1}$$

где $A_i$ — некоторые постоянные коэффициенты.

Выведем явные выражения для коэффициентов $A_i$ формулы (1). Как известно (гл. XIV, § 12),

$$L_n(x) = \sum_{i=0}^{n} p_i(x) y_i, \tag{2}$$

где

$$p_i(x) = \frac{(x-x_0)(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)}. \tag{3}$$

Введя обозначения

$$q = \frac{x-x_0}{h} \tag{4}$$

и

$$q^{[n+1]} = q(q-1)\ldots(q-n), \tag{5}$$

будем иметь (ср. гл. XV, § 4, формулу (2)):

$$L_n(x) = \sum_{i=0}^{n} \frac{(-1)^{n-i}}{i!\,(n-i)!} \cdot \frac{q^{[n+1]}}{q-i} y_i. \tag{6}$$

Заменяя в формуле (1) функцию $y$ полиномом $L_n(x)$, в силу формулы (6) получим:

$$A_i = \int_{x_0}^{x_n} \frac{(-1)^{n-i}}{i!\,(n-i)!} \cdot \frac{q^{[n+1]}}{q-i}\, dx$$

или, так как

$$q = \frac{x-x_0}{h}, \quad dq = \frac{dx}{h},$$

то, сделав замену переменных в определенном интеграле, будем иметь:

$$A_i = h\,\frac{(-1)^{n-i}}{i!\,(n-i)!} \int_0^n \frac{q^{[n+1]}}{q-i}\, dq \qquad (i = 0,\ 1,\ 2,\ \ldots,\ n).$$

Так как

$$h = \frac{b-a}{n},$$

то обычно полагают:

$$A_i = (b-a) H_i,$$

где

$$H_i = \frac{1}{n}\,\frac{(-1)^{n-i}}{i!\,(n-i)!} \int_0^n \frac{q^{[n+1]}}{q-i}\, dq \quad (i = 0,\ 1,\ 2,\ \ldots,\ n) \tag{7}$$

— постоянные, называемые *коэффициентами Котеса* (см., например, [1], [4]).

Квадратурная формула (1) при этом принимает вид

$$\int_a^b y\,dx = (b-a)\sum_{i=0}^{n} H_i y_i, \tag{8}$$

где

$$h = \frac{b-a}{n} \quad \text{и} \quad y_i = f_i(a+ih) \quad (i=0,\ 1,\ \ldots,\ n).$$

Нетрудно убедиться, что справедливы соотношения:

1) $\sum_{i=0}^{n} H_i = 1$; 2) $H_i = H_{n-i}$.

## § 3. Формула трапеций и ее остаточный член

Применяя формулу (7) предыдущего параграфа, при $n=1$ имеем:

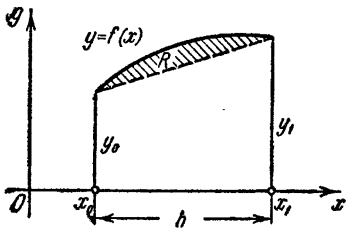


Рис. 69.

$$H_0 = -\int_0^1 \frac{q(q-1)}{q}\,dq = \frac{1}{2},$$

$$H_1 = \int_0^1 q\,dq = \frac{1}{2};$$

отсюда

$$\int_{x_0}^{x_1} y\,dx = \frac{h}{2}(y_0+y_1). \tag{1}$$

Мы получили известную *формулу трапеций* для приближенного вычисления определенного интеграла (рис. 69).

Остаточный член (ошибка) квадратурной формулы (1) равен

$$R = \int_{x_0}^{x_1} y\,dx - \frac{h}{2}(y_0+y_1).$$

Предполагая, что $y \in C^{(2)}[a, b]$, выведем простую формулу для остаточного члена. Будем рассматривать $R = R(h)$ как функцию шага $h$; тогда можно положить:

$$R(h) = \int_{x_0}^{x_0+h} y\,dx - \frac{h}{2}[y(x_0)+y(x_0+h)].$$

Дифференцируя эту формулу по $h$ последовательно два раза, получим:

$$R'(h) = y(x_0+h) - \frac{1}{2}[y(x_0)+y(x_0+h)] - \frac{h}{2}y'(x_0+h) =$$

$$= \frac{1}{2}[y(x_0+h) - y(x_0)] - \frac{h}{2}y'(x_0+h)$$

и

$$R''(h)=\frac{1}{2}y'(x_0+h)-\frac{1}{2}y'(x_0+h)-\frac{h}{2}y''(x_0+h)=-\frac{h}{2}y''(x_0+h),$$

причем

$$R(0)=0, \quad R'(0)=0.$$

Отсюда, интегрируя по $h$ и используя теорему о среднем, последовательно выводим:

$$R'(h)=R'(0)+\int_0^h R''(t)\,dt=-\frac{1}{2}\int_0^h ty''(x_0+t)\,dt=$$

$$=-\frac{1}{2}y''(\xi_1)\int_0^h t\,dt=-\frac{h^2}{4}y''(\xi_1),$$

где $\xi_1\in(x_0,\ x_0+h)$, и

$$R(h)=R(0)+\int_0^h R'(t)\,dt=-\frac{1}{4}\int_0^h t^2y''(\xi_1)\,dt=$$

$$=-\frac{1}{4}y''(\xi)\int_0^h t^2\,dt=-\frac{h^3}{12}y''(\xi),$$

где $\xi\in(x_0,\ x_0+h)$.

Таким образом, окончательно имеем:

$$R=-\frac{h^3}{12}y''(\xi), \tag{2}$$

где $\xi\in(x_0,\ x_1)$.

Отсюда, в частности, следует, что если $y''>0$, то формула (1) дает значение интеграла с *избытком*, если же $y''<0$ — то с *недостатком*.

## § 4. Формула Симпсона и ее остаточный член

Из формулы (7) § 2 при $n=2$ получаем:

$$H_0=\frac{1}{2}\cdot\frac{1}{2}\int_0^2(q-1)(q-2)\,dq=\frac{1}{4}\left(\frac{8}{3}-6+4\right)=\frac{1}{6},$$

$$H_1=-\frac{1}{2}\cdot\frac{1}{1}\int_0^2 q(q-2)\,dq=\frac{2}{3},$$

$$H_2=\frac{1}{2}\cdot\frac{1}{2}\int_0^2 q(q-1)\,dq=\frac{1}{6}.$$

Следовательно, так как $x_2 - x_0 = 2h$, имеем:

$$\int_{x_0}^{x_2} y\,dx = \frac{h}{3}(y_0 + 4y_1 + y_2). \tag{1}$$

Формула (1) носит название *формулы Симпсона*. Геометрически эта формула получается в результате замены данной кривой $y = f(x)$ параболой $y = L_2(x)$, проходящей через три точки $M_0(x_0, y_0)$, $M_1(x_1, y_1)$ и $M_2(x_2, y_2)$ (рис. 70).

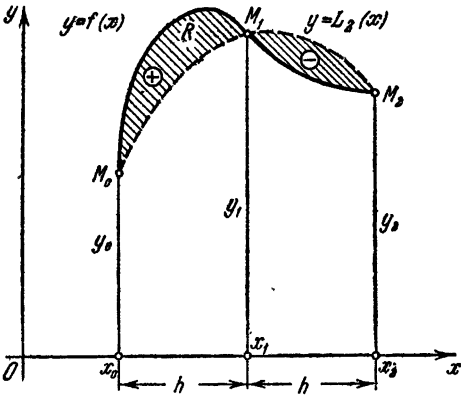

Рис. 70.

Остаточный член формулы Симпсона равен

$$R = \int_{x_0}^{x_2} y\,dx - \frac{h}{3}(y_0 + 4y_1 + y_2).$$

Предполагая, что $y \in C^{(4)}[a, b]$, аналогично тому как это делалось для формулы трапеций, выведем более простое выражение для $R$. Фиксируя среднюю точку $x_1$ и рассматривая $R = R(h)$ как функцию шага $h$ $(h \geqslant 0)$, будем иметь:

$$R(h) = \int_{x_1-h}^{x_1+h} y\,dx - \frac{h}{3}[y(x_1 - h) + 4y(x_1) + y(x_1 + h)].$$

Отсюда, дифференцируя функцию $R(h)$ по $h$ последовательно три раза, получим:

$$R'(h) = [y(x_1 + h) + y(x_1 - h)] - \frac{1}{3}[y(x_1 - h) + 4y(x_1) + y(x_1 + h)] -$$
$$- \frac{h}{3}[-y'(x_1 - h) + y'(x_1 + h)] = \frac{2}{3}[y(x_1 - h) + y(x_1 + h)] -$$
$$- \frac{4}{3}y(x_1) - \frac{h}{3}[-y'(x_1 - h) + y'(x_1 + h)];$$

$$R''(h) = \frac{2}{3}[-y'(x_1 - h) + y'(x_1 + h)] -$$
$$- \frac{1}{3}[-y'(x_1 - h) + y'(x_1 + h)] - \frac{h}{3}[y''(x_1 - h) + y''(x_1 + h)] =$$
$$= \frac{1}{3}[-y'(x_1 - h) + y'(x_1 + h)] - \frac{h}{3}[y''(x_1 - h) + y''(x_1 + h)];$$

$$R'''(h)=\frac{1}{3}[y''(x_1-h)+y''(x_1+h)]-$$
$$-\frac{1}{3}[y''(x_1-h)+y''(x_1+h)]-\frac{h}{3}[-y'''(x_1-h)+y'''(x_1+h)]=$$
$$=-\frac{h}{3}[y'''(x_1+h)-y'''(x_1-h)]=-\frac{2h^2}{3}y^{\mathrm{IV}}(\xi_3),$$

где $\xi_3\in(x_1-h,\ x_1+h)$.

Кроме того, имеем:

$$R(0)=0,\ R'(0)=0,\ R''(0)=0.$$

Последовательно интегрируя $R'''(h)$, используя теорему о среднем, находим:

$$R''(h)=R''(0)+\int_0^h R'''(t)\,dt=-\frac{2}{3}\int_0^h t^2y^{\mathrm{IV}}(\xi_3)\,dt=$$
$$=-\frac{2}{3}y^{\mathrm{IV}}(\xi_2)\int_0^h t^2\,dt=-\frac{2}{9}h^3y^{\mathrm{IV}}(\xi_2),$$

где $\xi_2\in(x_1-h,\ x_1+h)$;

$$R'(h)=R'(0)+\int_0^h R''(t)\,dt=-\frac{2}{9}\int_0^h t^3y^{\mathrm{IV}}(\xi_2)\,dt=$$
$$=-\frac{2}{9}y^{\mathrm{IV}}(\xi_1)\int_0^h t^3\,dt=-\frac{1}{18}h^4y^{\mathrm{IV}}(\xi_1),$$

где $\xi_1\in(x_1-h,\ x_1+h)$;

$$R(h)=R(0)+\int_0^h R'(t)\,dt=-\frac{1}{18}\int_0^h t^4y^{\mathrm{IV}}(\xi_1)\,dt=$$
$$=-\frac{1}{18}y^{\mathrm{IV}}(\xi)\int_0^h t^4\,dt=-\frac{h^5}{90}y^{\mathrm{IV}}(\xi),$$

где $\xi\in(x_1-h,\ x_1+h)$.

Таким образом, *остаточный член формулы Симпсона* равен

$$R=-\frac{h^5}{90}y^{\mathrm{IV}}(\xi),\tag{2}$$

где $\xi\in(x_0,\ x_2)$.

Следовательно, эта формула является *точной* для полиномов не только второй, но и третьей степени, т. е. формула Симпсона при относительно малом числе ординат обладает повышенной точностью.

## § 5. Формулы Ньютона — Котеса высших порядков

Производя соответствующие вычисления при $n=3$, получим из формулы (7) § 2 *квадратурную формулу Ньютона*

$$\int_{x_0}^{x_3} y\,dx = \frac{3h}{8}(y_0+3y_1+3y_2+y_3) \tag{1}$$

(*правило трех восьмых*).

Остаточный член формулы (1) равен [2]

$$R = -\frac{3h^5}{80} y^{\mathrm{IV}}(\xi),$$

где $\xi \in (x_0, x_3)$, т. е. при одинаковом шаге формула Ньютона, вообще говоря, менее точна, чем формула Симпсона.

Дальнейшие квадратурные формулы Ньютона — Котеса приведены в [1], [2]. Остаточные члены этих формул даны Стеффенсеном (см. [1], [5], [6]).

Заметим, что ошибка формулы Ньютона — Котеса с $n+1$ ординатами при достаточной гладкости функции $y=f(x)$ по меньшей мере имеет порядок [1], [6]

$$R = O\left[h^{2E\left(\frac{n}{2}\right)+3}\right],$$

где $E\left(\frac{n}{2}\right)$ — целая часть дроби $\frac{n}{2}$.

Отсюда видно, что в смысле порядка точности квадратурные формулы с нечетным числом ординат являются более выигрышными.

Таблица 65

**Коэффициенты Котеса**

| $n$ | $\hat{H}_0$ | $\hat{H}_1$ | $\hat{H}_2$ | $\hat{H}_3$ | $\hat{H}_4$ | $\hat{H}_5$ | $\hat{H}_6$ | $\hat{H}_7$ | $\hat{H}_8$ | Общий знаменатель $N$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | | | | | | | | 2 |
| 2 | 1 | 4 | 1 | | | | | | | 6 |
| 3 | 1 | 3 | 3 | 1 | | | | | | 8 |
| 4 | 7 | 32 | 12 | 32 | 7 | | | | | 90 |
| 5 | 19 | 75 | 50 | 50 | 75 | 19 | | | | 288 |
| 6 | 41 | 216 | 27 | 272 | 27 | 216 | 41 | | | 840 |
| 7 | 751 | 3577 | 1323 | 2989 | 2989 | 1323 | 3577 | 751 | | 17280 |
| 8 | 989 | 5888 | —928 | 10496 | —4540 | 10496 | —928 | 5888 | 989 | 28350 |

Приводим для справок таблицу коэффициентов Котеса (таблица 65). Для удобства записи коэффициенты Котеса для каждого $n$ представлены в виде дробей

$$H_i = \frac{\hat{H}_i}{N}$$

с общим знаменателем $N$. Для контроля заметим, что

$$\sum_{i=0}^{n} \hat{H}_i = N.$$

Следует обратить внимание на то, что коэффициенты Котеса при больших $n$ могут быть отрицательными (см., например, $n=8$).

Пример. Вычислить

$$I = \int_0^1 \frac{dx}{1+x},$$

применяя формулу Ньютона — Котеса с семью ординатами ($n=6$).

Решение. Полагая шаг

$$h = \frac{1-0}{6} = \frac{1}{6},$$

составляем приведенную ниже таблицу значений (таблица 66), где для удобства принято $\hat{H}_i = 840 H_i$.

Таблица 66

**Вычисление интеграла по формуле Ньютона — Котеса**

| $i$ | $x_i$ | $y_i$ | $\hat{H}_i$ | $\hat{H}_i y_i$ |
|---|---|---|---|---|
| 0 | 0 | 1 | 41 | 41 |
| 1 | $\frac{1}{6}$ | $\frac{6}{7}$ | 216 | 185,142857 |
| 2 | $\frac{1}{3}$ | $\frac{3}{4}$ | 27 | 20,25 |
| 3 | $\frac{1}{2}$ | $\frac{2}{3}$ | 272 | 181,333333 |
| 4 | $\frac{2}{3}$ | $\frac{3}{5}$ | 27 | 16,2 |
| 5 | $\frac{5}{6}$ | $\frac{6}{11}$ | 216 | 117,818182 |
| 6 | 1 | $\frac{1}{2}$ | 41 | 20,25 |
| $\Sigma$ | | | | 581,994372 |

Отсюда

$$I = \frac{1}{840} \cdot 581{,}994372 = 0{,}6933.$$

Точное значение

$$I = \ln 2 = 0{,}69315\ \ldots$$

Так как коэффициенты Котеса при большом числе ординат весьма сложны, то практически для приближенного вычисления определенных интегралов поступают следующим образом: разбивают промежуток интегрирования на достаточно большое число мелких промежутков и к каждому из них применяют квадратурную формулу Ньютона — Котеса с малым числом ординат (см., например, [7]). Получаются формулы более простой структуры, точность которых может быть произвольно высокой.

В следующих параграфах мы рассмотрим примеры таких формул.

## § 6. Общая формула трапеций (правило трапеций)

Для вычисления интеграла

$$\int_a^b y\,dx$$

разделим промежуток интегрирования $[a, b]$ на $n$ равных частей $[x_0, x_1]$, $[x_1, x_2]$, ..., $[x_{n-1}, x_n]$ и к каждому из них применим формулу трапеций (см. § 3 (1)). Полагая $h = \frac{b-a}{n}$ и обозначая через $y_i = f(x_i)$ $(i = 0, 1, \ldots, n)$ значения подынтегральной функции в точках $x_i$, будем иметь:

$$\int_a^b y\,dx = \frac{h}{2}(y_0 + y_1) + \frac{h}{2}(y_1 + y_2) + \ldots + \frac{h}{2}(y_{n-1} + y_n)$$

или

$$\int_a^b y\,dx = h\left(\frac{y_0}{2} + y_1 + y_2 + \ldots + y_{n-2} + y_{n-1} + \frac{y_n}{2}\right). \tag{1}$$

Геометрически формула (1) получается в результате замены графика подынтегральной функции $y = f(x)$ ломаной линией (рис. 71).

Если $y \in C^{(2)}[a, b]$, то *остаточный член* квадратурной формулы (1) в силу (2) из § 3 равен

$$R = \int_{x_0}^{x_n} y\,dx - \frac{h}{2}\sum_{i=1}^{n}(y_{i-1} + y_i) =$$

$$= \sum_{i=1}^{n}\left[\int_{x_{i-1}}^{x_i} y\,dx - \frac{h}{2}(y_{i-1} + y_i)\right] = -\frac{h^3}{12}\sum_{i=1}^{n} y''(\xi_i), \tag{2}$$

где $\xi_i \in (x_{i-1}, x_i)$.

Рассмотрим среднее арифметическое

$$\mu = \frac{1}{n} \sum_{i=1}^{n} y''(\xi_i). \tag{3}$$

Очевидно, $\mu$ заключается между наименьшим $m_2$ и наибольшим $M_2$ значениями второй производной $y''$ на отрезке $[a, b]$, т. е.

$$m_2 \leqslant \mu \leqslant M_2.$$

Так как $y''$ непрерывна на отрезке $[a, b]$, то в качестве своих

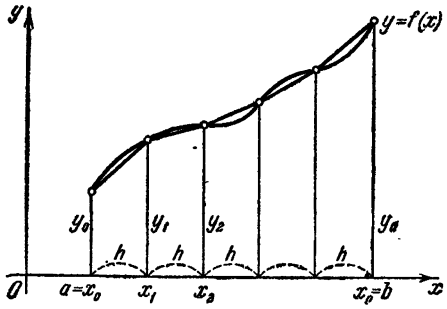


Рис. 71.

значений на $[a, b]$ она принимает все промежуточные числа между $m_2$ и $M_2$. Следовательно, найдется точка $\xi \in [a, b]$ такая, что

$$\mu = f''(\xi).$$

Из формул (2) и (3) имеем:

$$R = -\frac{nh^3}{12} y''(\xi) = -\frac{(b-a)h^2}{12} y''(\xi),$$

где $\xi \in [a, b]$.

## § 7. Общая формула Симпсона (параболическая формула)

Пусть $n = 2m$ есть четное число и $y_i = f(x_i)$ $(i = 0, 1, 2, \ldots, n)$ — значения функции $y = f(x)$ для равноотстоящих точек $a = x_0, x_1, \ldots$ $\ldots, x_n = b$ с шагом

$$h = \frac{b-a}{n} = \frac{b-a}{2m}.$$

Применяя формулу Симпсона (§ 4, (1)) к каждому удвоенному промежутку $[x_0, x_2]$, $[x_2, x_4]$, $\ldots$, $[x_{2m-2}, x_{2m}]$ длины $2h$ (рис. 72),

будем иметь:

$$\int_a^b y\,dx = \frac{h}{3}(y_0+4y_1+y_2)+\frac{h}{3}(y_2+4y_3+y_4)+\ldots$$
$$\ldots+\frac{h}{3}(y_{2m-2}+4y_{2m-1}+y_{2m}).$$

Отсюда получаем *общую формулу Симпсона*

$$\int_a^b y\,dx = \frac{h}{3}[(y_0+y_{2m})+4(y_1+y_3+\ldots+y_{2m-1})+$$
$$+2(y_2+y_4+\ldots+y_{2m-2})]. \qquad (1)$$

Введя обозначения

$$\sigma_1 = y_1+y_3+\ldots+y_{2m-1},$$
$$\sigma_2 = y_2+y_4+\ldots+y_{2m},$$

формулу (1) можно записать в более простом виде:

$$\int_a^b y\,dx = \frac{h}{3}[(y_0+y_n)+4\sigma_1+2\sigma_2]. \qquad (1')$$

Если $y \in C^{(4)}[a, b]$, то ошибка формулы Симпсона на каждом удвоенном промежутке $[x_{2k-2}, x_{2k}]$ $(k=1, 2, \ldots, m)$ на основании § 4, (2) дается формулой

$$r_k = -\frac{h^5}{90}y^{\mathrm{IV}}(\xi_k),$$

где $\xi_k \in (x_{2k-2}, x_{2k})$. Суммируя все эти ошибки, получим *остаточный член общей формулы Симпсона* в виде

$$R = -\frac{h^5}{90}\sum_{k=1}^{m} y^{\mathrm{IV}}(\xi_k).$$

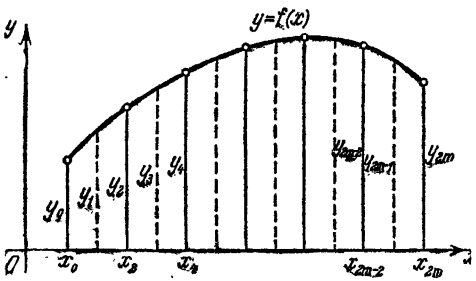

Рис. 72.

Так как $y^{\mathrm{IV}}(x)$ непрерывна на отрезке $[a, b]$, то найдется точка $\xi \in [a, b]$ такая, что

$$y^{\mathrm{IV}}(\xi) = \frac{1}{m}\sum_{k=1}^{m} y^{\mathrm{IV}}(\xi_k).$$

Поэтому будем иметь:

$$R = -\frac{mh^5}{90}y^{\mathrm{IV}}(\xi) = -\frac{(b-a)\,h^4}{180}y^{\mathrm{IV}}(\xi), \qquad (2)$$

где $\xi \in [a, b]$.

Если задана предельная допустимая погрешность $\varepsilon > 0$, то, обозначив

$$M_4 = \max |y^{\mathrm{IV}}(x)|,$$

будем иметь для определения шага $h$ неравенство

$$(b-a)\frac{h^4}{180}M_4 < \varepsilon;$$

отсюда

$$h < \sqrt[4]{\frac{180\varepsilon}{(b-a)M_4}},$$

т. е. $h$ имеет порядок $\sqrt[4]{\varepsilon}$.

Во многих случаях оценка погрешности квадратурной формулы Симпсона (1) по формуле (2) весьма затруднительна. Тогда обычно применяют двойной пересчет с шагами $h$ и $2h$ и считают, что совпадающие десятичные знаки принадлежат точному значению интеграла.

Можно указать еще один практически удобный способ подсчета ошибки квадратурной формулы Симпсона. Предполагая, что на отрезке $[a, b]$ производная $y^{\mathrm{IV}}(x)$ меняется медленно, в силу формулы (2) получаем приближенное выражение для искомой ошибки

$$R = Mh^4,$$

где коэффициент $M$ будем считать постоянным. Пусть $\Sigma_h$ и $\Sigma_H$ — приближенные значения интеграла

$$I = \int_a^b y\,dx,$$

полученные по формуле Симпсона соответственно с шагом $h$ и шагом $H = 2h$. Имеем:

$$I = \Sigma_h + Mh^4$$

и

$$I = \Sigma_H + M(2h)^4.$$

Отсюда

$$R = \frac{\Sigma_h - \Sigma_H}{15}.$$

За приближенное значение интеграла $I$ целесообразно принять исправленное значение

$$I = \Sigma_h + \frac{\Sigma_h - \Sigma_H}{15}.$$

Заметим, что если число делений $n$ кратно 4, то для вычисления суммы $\Sigma_H$ можно воспользоваться имеющимися табличными значениями, делая выборку значений через одно.

Пример. С помощью формулы Симпсона вычислить интеграл

$$I=\int_0^1 \frac{dx}{1+x},$$

приняв $n=10$.

Решение. Имеем $2m=10$. Отсюда

$$h=\frac{1-0}{10}=0{,}1.$$

Результаты вычислений приведены в таблице 67.

Таблица 67

**Вычисление интеграла по формуле Симпсона**

| $i$ | $x_i$ | $y_{2j-1}$ | $y_{2j}$ |
|---|---|---|---|
| 0 | 0 | | $\boxed{y_0=1{,}00000}$ |
| 1 | 0,1 | 0,90909 | |
| 2 | 0,2 | | 0,83333 |
| 3 | 0,3 | 0,76923 | |
| 4 | 0,4 | | 0,71429 |
| 5 | 0,5 | 0,66667 | |
| 6 | 0,6 | | 0,62500 |
| 7 | 0,7 | 0,58824 | |
| 8 | 0,8 | | 0,55556 |
| 9 | 0,9 | 0,52632 | |
| 10 | 1,0 | | $\boxed{0{,}50000=y_n}$ |
| $\Sigma$ | | $\boxed{3{,}45955}$ $(\sigma_1)$ | $\boxed{2{,}72818}$ $(\sigma_2)$ |

По формуле (1′) получаем:

$$I\approx\frac{h}{3}(y_0+y_n+4\sigma_1+2\sigma_2)=0{,}69315. \qquad (3)$$

Подсчитаем погрешность результата (3). Полная погрешность $R$ складывается из погрешности действий $R_1$ и остаточного члена $R_2$. Очевидно,

$$R_1=\sum_{i=0}^{n} A_i\varepsilon,$$

где $A_i$ — коэффициенты формулы Симпсона и $\varepsilon$ — максимальная ошибка округления значений подынтегральной функции.

В нашем случае

$$R_1=nh\varepsilon=(b-a)\,\varepsilon=1\cdot\frac{1}{2}\cdot 10^{-5}=0{,}5\cdot 10^{-5}.$$

Остаточный член оценим по формуле (2). Так как

$$y=\frac{1}{1+x}=(1+x)^{-1},$$

то

$$y^{\mathrm{IV}}=(-1)(-2)(-3)(-4)(1+x)^{-5}=\frac{24}{(1+x)^5}.$$

Отсюда

$$\max|y^{\mathrm{IV}}|=24 \quad \text{при} \quad 0\leqslant x\leqslant 1$$

и, следовательно,

$$|R_2|\leqslant 1\cdot\frac{(0,1)^4}{180}\cdot 24=1,3\cdot 10^{-5}.$$

Таким образом, предельная полная погрешность есть

$$R=0,5\cdot 10^{-5}+1,3\cdot 10^{-5}=1,8\cdot 10^{-5}<0,00002$$

и, значит,

$$I=0,69315\pm 0,00002.$$

## § 8. Понятие о квадратурной формуле Чебышева

Рассмотрим квадратурную формулу

$$\int_{-1}^{1} f(t)\,dt=\sum_{i=1}^{n} B_i f(t_i), \tag{1}$$

где $B_i$ — постоянные коэффициенты.

Чебышев предложил выбрать абсциссы $t_i$ таким образом, чтобы:

1) коэффициенты $B_i$ были равны между собой;

2) квадратурная формула (1) являлась точной для всех полиномов до степени $n$ включительно.

Покажем, как могут быть найдены в этом случае величины $B_i$ и $t_i$. Полагая

$$B_1=B_2=\ldots=B_n=B$$

и учитывая, что при $f(t)\equiv 1$ будем иметь

$$2=\sum_{i=1}^{n} B_i,$$

отсюда получаем:

$$B=\frac{2}{n}.$$

Следовательно, *квадратурная формула Чебышева* имеет вид

$$\int_{-1}^{1} f(t)\,dt=\frac{2}{n}\sum_{i=1}^{n} f(t_i). \tag{2}$$

Для определения абсцисс $t_i$ заметим, что формула (2), согласно условию 2), должна быть точной для функций вида

$$f(t)=t,\ t^2,\ \ldots,\ t^n.$$

Подставляя эти функции в формулу (2), получим систему уравнений

$$\left.\begin{aligned}
&t_1+t_2+\ldots+t_n=0,\\
&t_1^2+t_2^2+\ldots+t_n^2=\frac{n}{3},\\
&t_1^3+t_2^3+\ldots+t_n^3=0,\\
&t_1^4+t_2^4+\ldots+t_n^4=\frac{n}{5},\\
&\ldots\ldots\ldots\ldots\ldots\\
&t_1^n+t_2^n+\ldots+t_n^n=\frac{n\,[1-(-1)^{n+1}]}{2\,(n+1)},
\end{aligned}\right\}\qquad(3)$$

из которой могут быть определены неизвестные $t_i$ $(i=1,\ 2,\ \ldots,\ n)$. Чебышев показал, что решение системы (3) сводится к нахождению корней некоторого алгебраического уравнения степени $n$ [6], [8]. В таблице 68 приведены значения корней $t_i$ системы (3) для $n=2,\ 3,\ \ldots,\ 7$.

Таблица 68

**Значения абсцисс $t_i$ в формуле Чебышева**

| $n$ | $i$ | $t_i$ | $n$ | $i$ | $t_i$ |
|---|---|---|---|---|---|
| 2 | 1; 2 | ∓0,577350 | 6 | 1; 6 | ∓0,866247 |
| 3 | 1; 3 | ∓0,707107 | | 2; 5 | ∓0,422519 |
| | 2 | 0 | | 3; 4 | ∓0,266635 |
| 4 | 1; 4 | ∓0,794654 | 7 | 1; 7 | ∓0,883862 |
| | 2; 3 | ∓0,187592 | | 2; 6 | ∓0,529657 |
| 5 | 1; 5 | ∓0,832498 | | 3; 5 | ∓0,323912 |
| | 2; 4 | ∓0,374541 | | 4 | 0 |
| | 3 | 0 | | | |

Заметим, что система (3), как показал С. Н. Бернштейн, при $n=8$ и $n\geqslant 10$ не имеет действительных решений. В этом состоит принципиальный недостаток квадратурной формулы Чебышева.

Пример 1. Вывести формулу Чебышева с тремя ординатами $(n=3)$.

Решение. Для определения абсцисс $t_i$ $(i=1,\ 2,\ 3)$ имеем систему уравнений

$$\left.\begin{aligned}
&t_1+t_2+t_3=0,\\
&t_1^2+t_2^2+t_3^2=1,\\
&t_1^3+t_2^3+t_3^3=0.
\end{aligned}\right\}\qquad(4)$$

Рассмотрим симметрические функции корней

$$C_1 = t_1 + t_2 + t_3,$$
$$C_2 = t_1t_2 + t_1t_3 + t_2t_3,$$
$$C_3 = t_1t_2t_3.$$

Из системы (4) имеем:

$$C_1 = 0;$$
$$C_2 = \frac{1}{2}\left[(t_1+t_2+t_3)^2 - (t_1^2+t_2^2+t_3^2)\right] = \frac{1}{2}(0-1) = -\frac{1}{2};$$
$$C_3 = \frac{1}{6}\left[(t_1+t_2+t_3)^3 - 3(t_1+t_2+t_3)(t_1^2+t_2^2+t_3^2) + \right.$$
$$\left. + 2(t_1^3+t_2^3+t_3^3)\right] = \frac{1}{6}(0-0+0) = 0.$$

Отсюда заключаем, что $t_i$ есть корни вспомогательного уравнения

$$t^3 - C_1t^2 + C_2t - C_3 = 0$$

или

$$t^3 - \frac{1}{2}t = 0.$$

Следовательно, можно принять:

$$t_1 = -\frac{\sqrt{2}}{2}, \quad t_2 = 0, \quad t_3 = \frac{\sqrt{2}}{2}.$$

Таким образом, соответствующая формула Чебышева имеет вид

$$\int_{-1}^{1} f(t)\,dt = \frac{2}{3}\left[f\left(-\frac{1}{\sqrt{2}}\right) + f(0) + f\left(\frac{1}{\sqrt{2}}\right)\right].$$

Чтобы применить квадратурную формулу Чебышева к интегралу вида

$$\int_a^b f(x)\,dx,$$

следует преобразовать его с помощью подстановки

$$x = \frac{b+a}{2} + \frac{b-a}{2}t,$$

переводящей отрезок $a \leqslant x \leqslant b$ в отрезок $-1 \leqslant t \leqslant 1$. Применяя к преобразованному интегралу формулу Чебышева (2), будем иметь:

$$\int_a^b f(x)\,dx = \frac{b-a}{n}\sum_{i=1}^{n} f(x_i), \tag{5}$$

где

$$x_i = \frac{b+a}{2} + \frac{b-a}{2} t_i \tag{6}$$

и $t_i$ $(i=1, 2, \ldots, n)$ — корни системы (3) (помещены в таблице 68).

Квадратурная формула Чебышева употребляется главным образом в кораблестроении.

Пример 2. Вычислить интеграл

$$I = \int_0^1 \frac{x\,dx}{1+x}$$

по формуле Чебышева с пятью ординатами ($n=5$).

Решение. Введем обозначение

$$f(x) = \frac{x}{1+x},$$

имеем:

$$I = \frac{1}{5}[f(x_1) + f(x_2) + f(x_3) + f(x_4) + f(x_5)],$$

где в силу формулы (6)

$$x_1 = \frac{1}{2} + \frac{1}{2} t_1 = \frac{1}{2} + \frac{1}{2} \cdot (-0{,}83250) = 0{,}08375;$$

$$x_2 = \frac{1}{2} + \frac{1}{2} t_2 = \frac{1}{2} + \frac{1}{2} \cdot (-0{,}37454) = 0{,}31273;$$

$$x_3 = \frac{1}{2} + \frac{1}{2} t_3 = \frac{1}{2} + \frac{1}{2} \cdot 0 = 0{,}5;$$

$$x_4 = 1 - x_2 = 0{,}68727;$$

$$x_5 = 1 - x_1 = 0{,}91625.$$

Таблица 69

**Вычисление интеграла по формуле Чебышева**

| $i$ | $x_i$ | $y_i$ |
|---|---|---|
| 1 | 0,08375 | 0,0773 |
| 2 | 0,31273 | 0,2382 |
| 3 | 0,50000 | 0,3333 |
| 4 | 0,68727 | 0,4073 |
| 5 | 0,91625 | 0,4781 |
| Σ | | 1,5342 |

Соответствующие значения $y_i = f(x_i)$ $(i=1, 2, 3, 4, 5)$ подынтегральной функции помещены в таблице 69.

Отсюда

$$I = \frac{1}{5} \cdot 1{,}5342 = 0{,}3068.$$

Для сравнения приводим точное значение интеграла с шестью значащими цифрами

$$I = 0{,}306846\ldots$$

## § 9. Квадратурная формула Гаусса

В этом параграфе нам потребуются некоторые сведения о полиномах Лежандра. Полиномы вида

$$P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} [(x^2-1)^n] \qquad (n = 0, 1, 2, \ldots)$$

называются полиномами Лежандра.

Отметим важнейшие свойства полиномов Лежандра [1]:

1) $P_n(1) = 1$, $P_n(-1) = (-1)^n$ $(n = 0, 1, \ldots)$,

2) $\int\limits_{-1}^{1} P_n(x) \cdot Q_k(x)\, dx = 0$ $(k < n)$, где $Q_k(x)$ — любой полином степени $k$, меньшей $n$;

3) полином Лежандра $P_n(x)$ имеет $n$ различных и действительных корней, которые расположены на интервале $(-1, 1)$.

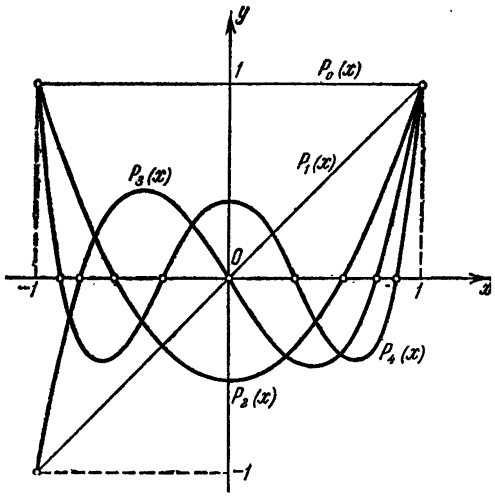


Рис. 73.

Ниже приводим первые пять полиномов Лежандра и их графики (рис. 73):

$$P_0(x) = 1,$$
$$P_1(x) = x,$$
$$P_2(x) = \frac{1}{2}(3x^2 - 1),$$
$$P_3(x) = \frac{1}{2}(5x^3 - 3x),$$
$$P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3).$$

Перейдем сейчас к выводу *квадратурной формулы Гаусса.*

Рассмотрим сначала функцию $y=f(t)$, заданную на стандартном промежутке $[-1;\ 1]$. Общий случай легко свести к нашему путем линейной замены независимой переменной.

Поставим задачу: как нужно подобрать точки $t_1, t_2, \ldots, t_n$ и коэффициенты $A_1, A_2, \ldots, A_n$, чтобы квадратурная формула

$$\int_{-1}^{1} f(t)\,dt = \sum_{i=1}^{n} A_i f(t_i) \tag{1}$$

была точной для всех полиномов $f(t)$ наивысшей возможной степени $N$.

Так как в нашем распоряжении имеется $2n$ постоянных $t_i$ и $A_i$ $(i=1, 2, \ldots, n)$, а полином степени $2n-1$ определяется $2n$ коэффициентами, то эта наивысшая степень в общем случае, очевидно, равна $N=2n-1$.

Для обеспечения равенства (1) необходимо и достаточно, чтобы оно было верным при

$$f(x)=1,\ t,\ t^2,\ \ldots,\ t^{2n-1}.$$

Действительно, полагая

$$\int_{-1}^{1} t^k\,dt = \sum_{i=1}^{n} A_i t_i^k \qquad (k=0,\ 1,\ 2,\ \ldots,\ 2n-1) \tag{2}$$

и

$$f(t)=\sum_{k=0}^{2n-1} C_k t^k,$$

будем иметь:

$$\int_{-1}^{1} f(t)\,dt = \sum_{k=0}^{2n-1} C_k \int_{-1}^{1} t^k\,dt = \sum_{k=0}^{2n-1} C_k \sum_{i=1}^{n} A_i t_i^k =$$
$$= \sum_{i=1}^{n} A_i \sum_{k=0}^{2n-1} C_k t_i^k = \sum_{i=1}^{n} A_i f(t_i).$$

Таким образом, учитывая соотношения:

$$\int_{-1}^{1} t^k\,dt = \frac{1-(-1)^{k+1}}{k+1} = \begin{cases} \dfrac{2}{k+1} & \text{при } k \text{ четном;} \\ 0 & \text{при } k \text{ нечетном,} \end{cases}$$

заключаем, что для решения поставленной задачи [2], [3], [6]

достаточно определить $t_i$ и $A_i$ из системы $2n$ уравнений

$$\left.\begin{aligned} &\sum_{i=1}^{n} A_i = 2, \\ &\sum_{i=1}^{n} A_i t_i = 0, \\ &\cdots\cdots\cdots \\ &\sum_{i=1}^{n} A_i t_i^{2n-2} = \frac{2}{2n-1}, \\ &\sum_{i=1}^{n} A_i t_i^{2n-1} = 0. \end{aligned}\right\} \tag{3}$$

Система (3) — нелинейная, и решение ее обычным путем представляет большие математические трудности. Однако здесь можно применить следующий искусственный прием.

Рассмотрим полиномы

$$f(t) = t^k P_n(t) \qquad (k = 0,\ 1,\ \ldots,\ n-1),$$

где $P_n(t)$ — полином Лежандра.

Так как степени этих полиномов не превышают $2n-1$, то на основании системы (3) для них должна быть справедлива формула (1) и

$$\int_{-1}^{1} t^k P_n(t)\,dt = \sum_{i=1}^{n} A_i t_i^k P_n(t_i) \qquad (k = 0,\ 1,\ \ldots,\ n-1). \tag{4}$$

С другой стороны, в силу свойства ортогональности полиномов Лежандра (свойство 2)) выполнены равенства

$$\int_{-1}^{1} t^k P_n(t)\,dt = 0 \quad \text{при} \quad k < n,$$

поэтому

$$\sum_{i=1}^{n} A_i t_i^k P_n(t_i) = 0 \qquad (k = 0,\ 1,\ \ldots,\ n-1). \tag{5}$$

Равенства (5) заведомо будут обеспечены при любых значениях $A_i$, если положить

$$P_n(t_i) = 0 \qquad (i = 1,\ 2,\ \ldots,\ n), \tag{6}$$

т. е. для достижения наивысшей точности квадратурной формулы (1) в качестве точек $t_i$ достаточно взять нули соответствующего полинома Лежандра. Как известно (свойство 3)), эти нули действительны, различны и расположены на интервале $(-1,\ 1)$. Зная абсциссы $t_i$,

легко можно найти из линейной системы первых $n$ уравнений системы (3) коэффициенты $A_i$ ($i=1, 2, \ldots, n$). Определитель этой подсистемы есть определитель Вандермонда

$$D=\prod_{i>j}(t_i-t_j)\neq 0$$

и, следовательно, $A_i$ определяются однозначно.

Формула (1), где $t_i$ — нули полинома Лежандра $P_n(t)$ и $A_i$ ($i=1, 2, \ldots, n$) определяются из системы (3), называется *квадратурной формулой Гаусса.*

Пример 1. Вывести квадратурную формулу Гаусса для случая трех ординат ($n=3$).

Решение. Полином Лежандра третьей степени есть

$$P_3(t)=\frac{1}{2}(5t^3-3t).$$

Приравнивая этот полином нулю, находим корни

$$t_1=-\sqrt{\frac{3}{5}}\approx -0{,}774597;$$
$$t_2=0;$$
$$t_3=\sqrt{\frac{3}{5}}\approx 0{,}774597.$$

Для определения коэффициентов $A_1$, $A_2$, $A_3$ в силу (3) имеем систему:

$$\left.\begin{aligned} A_1+A_2+A_3&=2,\\ -\sqrt{\frac{3}{5}}A_1+\sqrt{\frac{3}{5}}A_3&=0;\\ \frac{3}{5}A_1+\frac{3}{5}A_3&=\frac{2}{3}, \end{aligned}\right\}$$

отсюда

$$A_1=A_3=\frac{5}{9}, \qquad A_2=\frac{8}{9}.$$

Следовательно,

$$\int_{-1}^{1} f(t)\,dt=\frac{1}{9}\left[5f\left(-\sqrt{\frac{3}{5}}\right)+8f(0)+5f\left(\sqrt{\frac{3}{5}}\right)\right].$$

Для справок приводим (таблица 70) приближенные значения абсцисс $t_i$ и коэффициентов $A_i$ в квадратурной формуле Гаусса (1) для $n=1-8$ (см. [1], [4], [6]).

Неудобство применения квадратурной формулы Гаусса состоит в том, что абсциссы точек $t_i$ и коэффициенты $A_i$ — вообще говоря, иррациональные числа. Этот недостаток отчасти искупается ее высокой точностью при сравнительно малом числе ординат.

Рассмотрим теперь использование квадратурной формулы Гаусса для вычисления общего интеграла

$$\int_a^b f(x)\,dx.$$

Делая замену переменной

$$x=\frac{b+a}{2}+\frac{b-a}{2}t,$$

получим:

$$\int_a^b f(x)\,dx=\frac{b-a}{2}\int_{-1}^{1} f\left(\frac{b+a}{2}+\frac{b-a}{2}t\right)dt.$$

Применяя к последнему интегралу квадратурную формулу Гаусса (1) будем иметь:

$$\int_a^b f(x)\,dx=\frac{b-a}{2}\sum_{i=1}^{n} A_i f(x_i), \tag{7}$$

где

$$x_i=\frac{b+a}{2}+\frac{b-a}{2}t_i \qquad (i=1,\ 2,\ \ldots,\ n), \tag{8}$$

$t_i$ — нули полинома Лежандра $P_n(t)$, т. е.

$$P_n(t_i)=0.$$

Остаточный член формулы Гаусса (7) с $n$ узлами выражается следующим образом [1], [6]:

$$R_n=\frac{(b-a)^{2n+1}(n!)^4 f^{(2n)}(\xi)}{[(2n)!]^3(2n+1)};$$

отсюда получаем:

$$R_2=\frac{1}{135}\left(\frac{b-a}{2}\right)^5 f^{(4)}(\xi),$$

$$R_3=\frac{1}{15750}\left(\frac{b-a}{2}\right)^7 f^{(6)}(\xi),$$

$$R_4=\frac{1}{3472875}\left(\frac{b-a}{2}\right)^9 f^{(8)}(\xi),$$

$$R_5=\frac{1}{1237732650}\left(\frac{b-a}{2}\right)^{11} f^{(10)}(\xi),$$

$$R_6=\frac{1}{648984486150}\left(\frac{b-a}{2}\right)^{13} f^{(12)}(\xi) \text{ и т. д.}$$

Пример 2. Вычислить интеграл

$$I=\int_0^1 \sqrt{1+2x}\,dx,$$

применяя формулу Гаусса с тремя ординатами ($n=3$).

Таблица 70

**Элементы формулы Гаусса**

| $n$ | $i$ | $t_i$ | $A_i$ |
|---|---|---|---|
| 1 | 1 | 0 | 2 |
| 2 | 1; 2 | ∓ 0,57735027 | 1 |
| 3 | 1; 3 | ∓ 0,77459667 | $\frac{5}{9}=0,55555556$ |
| | 2 | 0 | $\frac{8}{9}=0,88888889$ |
| 4 | 1; 4 | ∓ 0,86113631 | 0,34785484 |
| | 2; 3 | ∓ 0,33998104 | 0,65214516 |
| 5 | 1; 5 | ∓ 0,90617985 | 0,23692688 |
| | 2; 4 | ∓ 0,53846931 | 0,47862868 |
| | 3 | 0 | 0,56888889 |
| 6 | 1; 6 | ∓ 0,93246951 | 0,17132450 |
| | 2; 5 | ∓ 0,66120939 | 0,36076158 |
| | 3; 4 | ∓ 0,23861919 | 0,46791394 |
| 7 | 1; 7 | ∓ 0,94910791 | 0,12948496 |
| | 2; 6 | ∓ 0,74153119 | 0,27970540 |
| | 3; 5 | ∓ 0,40584515 | 0,38183006 |
| | 4 | 0 | 0,41795918 |
| 8 | 1; 8 | ∓ 0,96028986 | 0,10122854 |
| | 2; 7 | ∓ 0,79666648 | 0,22238104 |
| | 3; 6 | ∓ 0,52553242 | 0,31370664 |
| | 4; 5 | ∓ 0,18343464 | 0,36268378 |

Решение. Имеем $a=0$ и $b=1$. В силу формулы (8) и таблицы 70 абсциссы точек с точностью до пяти значащих цифр будут

иметь следующие значения:

$$x_1 = \frac{1}{2} + \frac{1}{2} t_1 = 0{,}11270;$$

$$x_2 = \frac{1}{2} + \frac{1}{2} t_2 = 0{,}50000;$$

$$x_3 = \frac{1}{2} + \frac{1}{2} t_3 = 0{,}88730.$$

Соответствующие коэффициенты формулы (7) в нашем случае будут:

$$C_1 = \frac{b-a}{2} A_1 = \frac{1}{2} \cdot \frac{5}{9} = \frac{5}{18} = 0{,}27778;$$

$$C_2 = \frac{b-a}{2} A_2 = \frac{1}{2} \cdot \frac{8}{9} = \frac{4}{9} = 0{,}44444;$$

$$C_3 = \frac{b-a}{2} A_3 = \frac{1}{2} \cdot \frac{5}{9} = \frac{5}{18} = 0{,}27778.$$

Дальнейшие вычисления сведены в таблицу 71.

Таблица 71

**Схема вычисления интеграла по формуле Гаусса**

| $i$ | $x_i$ | $y_i$ | $C_i$ | $C_i y_i$ |
|---|---|---|---|---|
| 1 | 0,11270 | 1,10698 | 0,27778 | 0,30747 |
| 2 | 0,50000 | 1,41421 | 0,44444 | 0,62853 |
| 3 | 0,88730 | 1,66571 | 0,27778 | 0,46270 |
| $\Sigma$ | | | | 1,39870 |

Следовательно,

$$I = \sum_{i=1}^{3} C_i y_i = 1{,}39870.$$

Для оценки остаточной погрешности $R_3$ можно воспользоваться формулой

$$R_3 = \frac{1}{15750} \left(\frac{b-a}{2}\right)^7 f^{(6)}(\xi), \text{ где } \xi \in (a, b).$$

Полагая

$$f(x) = \sqrt{1+2x} = (1+2x)^{\frac{1}{2}},$$

имеем:

$$f^{(6)}(x) = \frac{1}{2}\left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right)\left(-\frac{5}{2}\right)\left(-\frac{7}{2}\right)\left(-\frac{9}{2}\right)(1+2x)^{-\frac{11}{2}} \cdot 2^6 =$$

$$= -945\,(1+2x)^{-\frac{11}{2}}.$$

Отсюда

$$\max|f^{(6)}(x)|=945 \quad \text{при} \quad 0\leqslant x\leqslant 1$$

и следовательно,

$$|R_3|\leqslant\frac{945}{15750}\left(\frac{1}{2}\right)^7\approx\frac{1}{2000}.$$

Заметим, что точное значение интеграла есть

$$I=\sqrt{3}-\frac{1}{3}\approx 1{,}39872.$$

## § 10. Некоторые замечания о точности квадратурных формул

Рассмотренные нами квадратурные формулы имеют следующую структуру:

$$\int_a^b f(x)\,dx=\sum_{i=1}^{n}A_i f(x_i)+R[f], \tag{1}$$

где $x_1, x_2, \ldots, x_n$ — данная *система узлов* из отрезка интегрирования $[a, b]$, $A_i$ — некоторые известные *постоянные коэффициенты* и $R[f]$ — остаточный член.

При одном и том же числе ординат точность различных квадратурных формул различна.

Пример. Сравнить точность различных квадратурных формул с тремя ординатами для интегралов

$$I=\int_{-1}^{1}\sqrt{2+x}\,dx=2\sqrt{3}-\frac{2}{3}=2{,}797435\ldots$$

Решение. Применяя формулу Симпсона, получим:

$$I\approx\frac{1}{3}\left[\sqrt{2-1}+4\sqrt{2+0}+\sqrt{2+1}\right]=\frac{1}{3}\cdot 8{,}428905=2{,}809635.$$

Формула Чебышева дает такой результат:

$$I\approx\frac{2}{3}\left[\sqrt{2-\frac{\sqrt{2}}{2}}+\sqrt{2+0}+\sqrt{2+\frac{\sqrt{2}}{2}}\right]=$$
$$=\frac{2}{3}\cdot 4{,}220097=2{,}813398.$$

Наконец, формула Гаусса приводит к следующему значению:

$$I\approx 0{,}555566\left(\sqrt{2-0{,}774597}+\sqrt{2+0{,}774597}\right)+$$
$$+0{,}888889\sqrt{2+0}=2{,}797460.$$

Таким образом, здесь формула Гаусса оказывается наиболее точной.

Мы ограничимся исследованием квадратурных формул *с равноотстоящими узлами*; к числу их относятся наиболее распространенные формулы: трапеций, Симпсона, Ньютона—Котеса. В этом случае точность квадратурной формулы в основном характеризуется *порядком* остаточного члена

$$R = O(h^m), \qquad (2)$$

где

$$h = \frac{b-a}{n}$$

— шаг ($n$ — число делений) и $m$ — натуральное число. Например, для формулы трапеций имеем (§ 3):

$$R[f] = -\frac{b-a}{12} h^2 f''(\xi),$$

поэтому $m = 2$; для формулы Симпсона (§ 4)

$$R[f] = -\frac{b-a}{180} h^4 f^{\mathrm{IV}}(\xi),$$

отсюда $m = 4$. Квадратурная формула считается тем точнее, чем больше число $m$; в этом смысле формула Симпсона является более точной, нежели формула трапеций. Качество формулы обнаруживается при достаточно малом шаге $h$.

Отсюда вовсе не следует, что в конкретных случаях более грубая квадратурная формула при одном и том же шаге не может дать лучших результатов, чем более точная. Например, для функции (рис. 74)

$$f(x) = -8 + 45x^2 - 25x^4$$

имеем:

$$I = \int_{-1}^{1} f(x)\,dx = 2(-8 + 15 - 5) = 4.$$

При $h = 1$ формула трапеций дает точное значение

$$I_1 = \frac{1}{2} f(-1) + f(0) + \frac{1}{2} f(1) = 6 - 8 + 6 = 4,$$

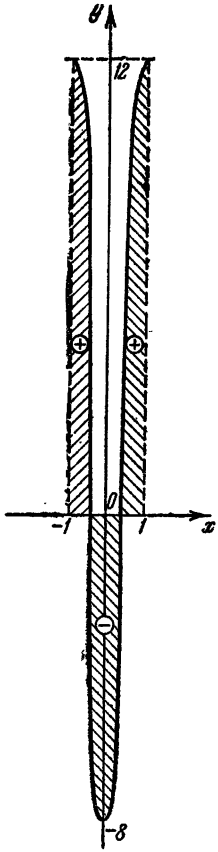

Рис. 74.

тогда как формула Симпсона при $h = 1$ не обеспечивает даже знака интеграла:

$$I_2 = \frac{1}{3}[f(-1) + 4f(0) + f(1)] = \frac{1}{3}(12 - 32 + 12) = -\frac{8}{3}.$$

Точность квадратурной формулы при фиксированном числе узлов существенно зависит от расположения этих узлов. При неудачном расположении узлов квадратурная формула может дать сильно

искаженные результаты. Например, для функции $y=f(x)$, изображенной на рис. 75, выбирая равноотстоящие узлы $a=x_0$, $x_1$, $x_2$, $x_3$, $x_4=b$ и пользуясь соответствующей формулой Котеса для пяти ординат, получим:

$$I=\int_a^b f(x)\,dx<0,$$

тогда как очевидно, что $I>0$.

Не представляет большого труда построить аналогичные примеры для любой квадратурной формулы с произвольным числом ординат.

Вообще, при наличии значительного числа нулей подынтегральной функции $f(x)$ или при большом числе экстремумов ее (т. е когда имеется много нулей производной $f'(x)$), благодаря неизбежным большим значениям старших производных точность квадратурных формул сильно понижается. Поэтому шаг $h$ следует выбирать так, чтобы он был намного меньше расстояний между соседними нулями функции $f(x)$ и ее производной $f'(x)$. Для этого рекомендуется основной отрезок интегрирования $[a, b]$ разбить на частичные отрезки $[\alpha, \beta]$, внутри каждого из которых функции $f(x)$ и $f'(x)$ сохраняют постоянный знак (если это возможно), и производить вычисление интеграла частями, выбирая для каждого частичного отрезка, вообще говоря, свой шаг. В более сложных случаях нужно учитывать также поведение производных высших порядков $f^{(n)}(x)$ $(n\geqslant 2)$. Для общей ориентировки полезно предварительно построить график подынтегральной функции $y=f(x)$. Если функция сильно колеблющаяся, то следует применять специальные приемы вычислений. Разработаны также общие приемы увеличения точности квадратурных формул [9].

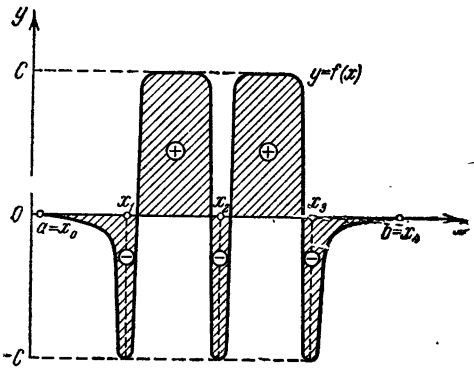

Рис. 75.

При нахождении *полной предельной погрешности* квадратурной формулы (1) следует учесть также *погрешность суммирования* $R_1$. Пусть слагаемые $f(x_i)$ $(i=1, 2, \ldots, n)$ вычислены с абсолютной погрешностью, не превышающей $\varepsilon$, а коэффициенты $A_i$ квадратурной формулы являются точными положительными постоянными. Тогда можно положить:

$$R_1\leqslant\sum_{i=1}^{n}A_i\varepsilon=\varepsilon\sum_{i=1}^{n}A_i. \qquad (3)$$

Так как квадратурная формула (1) верна для $f(x) \equiv 1$, то

$$\int_a^b dx = b - a = \sum_{i=1}^{n} A_i.$$

Поэтому из формулы (3) имеем:

$$R_1 \leqslant (b-a)\,\varepsilon. \tag{4}$$

Следовательно, полная предельная погрешность квадратурной формулы без учета заключительной погрешности округления равна

$$\tilde{R} = (b-a)\,\varepsilon + |R[f]|,$$

где $|R[f]|$ — *погрешность метода*, которая может быть определена указанным выше способом.

Заметим, что если подынтегральная функция $y = f(x)$ задана таблично значениями $y_i = f(x_i)$ $(i = 1, 2, \ldots, n)$, то, строго говоря, мы лишены возможности оценить точность квадратурной формулы (1). Дело в том, что через конечную систему точек $M_i(x_i, y_i)$ можно провести бесчисленное множество кривых $y = f(x)$ (рис. 76), ограничивающих на данном отрезке $[a, b]$ различные площади, т. е. интеграл

$$I = \int_a^b f(x)\,dx$$

априори может иметь совершенно произвольное значение (см. рис. 76). Пользование в этом случае квадратурными формулами допустимо лишь тогда, если нам в какой-то мере известны неиспользованные промежуточные значения подынтегральной функции и ее общие свойства, позволяющие судить о характере графика функции.

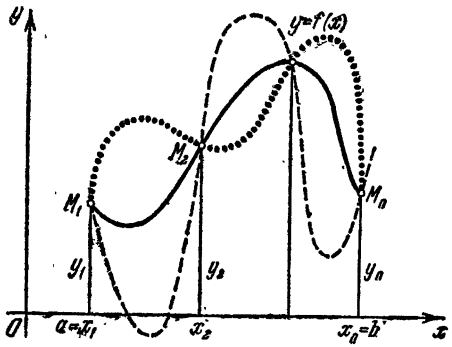

Рис. 76.

## § 11*. Экстраполяция по Ричардсону

Если для квадратурной формулы (1) § 10 известен порядок остаточного члена $R = R[f]$, то для определения величины $R$ можно использовать *метод двойного пересчета*. Пусть

$$R = O(h^m) \qquad (m \geqslant 1),$$

где

$$h=\frac{b-a}{n}$$

($n$ — число делений); тогда приближенно можно положить:

$$R=Mh^m, \tag{1}$$

где $M$ — некоторая величина, которую для данной подынтегральной функции $f(x)$ будем считать постоянной на промежутке интегрирования $[a, b]$. Выберем два различных шага

$$h_1=\frac{b-a}{n_1} \quad \text{и} \quad h_2=\frac{b-a}{n_2},$$

где $n_1$ и $n_2$ $(n_2>n_1)$ — количество частичных отрезков в первом и во втором случаях.

Обозначим через $I_{n_1}$ и $I_{n_2}$ соответствующие приближенные значения интеграла $I$. Из формулы (1) имеем:

$$R_{n_1}=I-I_{n_1}=M\left(\frac{b-a}{n_1}\right)^m \tag{2}$$

и

$$R_{n_2}=I-I_{n_2}=M\left(\frac{b-a}{n_2}\right)^m, \tag{2'}$$

где $R_{n_1}$ и $R_{n_2}$ — соответствующие остаточные члены. Отсюда

$$I_{n_2}-I_{n_1}=M(b-a)^m\left(\frac{1}{n_1^m}-\frac{1}{n_2^m}\right)$$

и, следовательно,

$$M=\frac{(n_1n_2)^m}{(b-a)^m}\cdot\frac{I_{n_2}-I_{n_1}}{n_2^m-n_1^m}.$$

На основании формулы (1) получаем выражение для остаточного члена

$$R=\left(\frac{n_1n_2}{n}\right)^m\cdot\frac{I_{n_2}-I_{n_1}}{n_2^m-n_1^m};$$

в частности при $h=h_2$, т. е. при $n=n_2$, имеем:

$$R_{n_2}=\frac{n_1^m}{n_2^m-n_1^m}(I_{n_2}-I_{n_1}). \tag{3}$$

Пользуясь поправкой (3), в силу формулы (2'), для интеграла $I$ получаем уточненное значение:

$$I_{n_1,\,n_2}=I_{n_2}+\frac{n_1^m}{n_2^m-n_1^m}(I_{n_2}-I_{n_1}). \tag{4}$$

Таблица 72а

**Экстраполирование для случая формулы трапеций**

| № п/п | | $I_2$ | $I_4$ | $I_{2,4}$ | $I$ |
|---|---|---|---|---|---|
| 1 | $I=\int_0^{\pi}\sin x\,dx$ | 1,571 | 1,896 | 2,004 | 2,000 |
| 2 | $I=\int_0^{2}e^{-x^2}dx$ | 0,877 | 0,881 | 0,8823 | 0,8821 |
| 3 | $I=\int_3^{7}x^2\ln x\,dx$ | 185,7090 | 179,5385 | 177,4819 | 177,4836 |
| 4 | $I=\int_0^{4}\frac{dx}{\sqrt{5-x^2}}$ | 0,9695 | 0,9389 | 0,9286 | 0,9267 |

| № п/п | $e_1=I-I_2$ | $e_2=I-I_4$ | $e_{1,2}=I-I_{2,4}$ |
|---|---|---|---|
| 1 | 0,429 | 0,104 | —0,004 |
| 2 | 0,0051 | 0,0011 | —0,0002 |
| 3 | —8,2254 | —2,0549 | 0,0017 |
| 4 | —0,0428 | —0,0122 | —0,0019 |

Этот прием называется *экстраполяцией по Ричардсону* [10]. Введя обозначение

$$\frac{n_2}{n_1}=\alpha \quad (\alpha>1),$$

будем иметь:

$$I_{n_1,\,n_2}=I_{n_2}+\beta\,(I_{n_2}-I_{n_1}), \tag{5}$$

где

$$\beta=\frac{1}{\alpha^m-1}. \tag{6}$$

Коэффициенты $\beta$ табулированы для различных значений $\alpha$ и $m$. Заметим, что для формулы трапеций $m=2$, а для формулы Симпсона $m=4$. Частный случай формулы (5) был приведен в § 7.

Покажем, что если $I_{n_1}\neq I_{n_2}$, то $I_{n_1,\,n_2}$ всегда лежит вне отрезка $[I_{n_1},\ I_{n_2}]$.

Действительно, если

$$I_{n_2}>I_{n_1},$$

Таблица 72б

**Экстраполирование для случая формулы Симпсона**

| № п/п | | $I_2$ | $I_4$ | $I_{2,4}$ | $I$ |
|---|---|---|---|---|---|
| 1 | $I=\int_0^{\pi}\sin x\,dx$ | 2,094 | 2,004 | 2,010 | 2,000 |
| 2 | $I=\int_0^{1}\frac{dx}{1+x^2}$ | 0,7833 | 0,7853 | 0,7855 | 0,7854 |
| 3 | $I=\int_3^{7}x^2\ln x\,dx$ | 177,454 | 177,481 | 177,483 | 177,4836 |
| 4 | $I=\int_0^{4}\frac{dx}{(25-x^2)^{3/2}}$ | 0,0577 | 0,0541 | 0,0538 | 0,0533 |

| № п/п | $e_1=I-I_1$ | $e_2=I-I_4$ | $e_{1,2}=I-I_{2,4}$ |
|---|---|---|---|
| 1 | —0,094 | —0,004 | —0,010 |
| 2 | 0,0021 | 0,0001 | —0,0001 |
| 3 | 0,0296 | 0,0026 | 0,0006 |
| 4 | —0,0044 | —0,0008 | —0,0005 |

то из формулы (5) имеем:

$$I_{n_1,n_2} > I_{n_2} = \max\{I_{n_1},\ I_{n_2}\}.$$

Если же

$$I_{n_2} < I_{n_1},$$

то из той же формулы (5) получаем:

$$I_{n_1,n_2} = I_{n_2} - \beta(I_{n_1} - I_{n_2}) < I_{n_2} = \min\{I_{n_1},\ I_{n_2}\}.$$

Таким образом,

$$I_{n_1,n_2} \overline{\in} [I_{n_1},\ I_{n_2}],$$

т. е. $I_{n_1,n_2}$ получается из $I_{n_1}$ и $I_{n_2}$ в результате операции экстраполирования. Этим и объясняется название способа.

Если $I_{n_1}=I_{n_2}$, то, очевидно,

$$I_{n_1,n_2} = I_{n_1} = I_{n_2}.$$

Можно показать, что при достаточной гладкости подынтегральной функции $f(x)$ порядок остаточного члена для $I_{n_1, n_2}$ равен по меньшей мере $m+1$.

Замечание. В таблицах 72а и 72б приведены примеры на экстраполирование по Ричардсону.

Из таблицы 72а и 72б видно, что для функций, не имеющих особенностей, экстраполирование, как правило, повышает точность вычислений.

Можно также вывести более точные формулы экстраполирования, использующие значения $I_{n_1}$, $I_{n_2}$ и $I_{n_3}$ искомого интеграла, соответствующие трем различным шагам

$$h_s = \frac{b-a}{n_s} \quad (s=1,\ 2,\ 3)$$

и учитывающие два первых члена разложения остаточного члена квадратурной формулы [10].

## § 12*. Числа Бернулли

Рассмотрим функцию

$$f(x) = \frac{x}{e^x - 1}. \tag{1}$$

Воспользовавшись известным разложением

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots,$$

можно записать:

$$f(x) = \frac{x}{\frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots} = \frac{1}{1 + \frac{x}{2!} + \frac{x^2}{3!} + \ldots}. \tag{2}$$

Отсюда ясно, что функция $f(x)$ в окрестности $x=0$ допускает разложение в степенной ряд, который для удобства дальнейших выкладок представим в следующем виде:

$$\frac{x}{e^x - 1} = \sum_{n=0}^{\infty} \frac{B_n}{n!} x^n, \tag{3}$$

где $B_0 = f(0) = 1$. Для определения остальных коэффициентов разложения $B_n$ $(n=1,\ 2,\ \ldots)$, носящих название чисел *Бернулли*, используем получаемое на основании формулы (2) тождество

$$\sum_{n=0}^{\infty} \frac{x^n}{(n+1)!} \cdot \sum_{n=0}^{\infty} \frac{B_n}{n!} x^n \equiv 1.$$

Производя перемножение степенных рядов и приравнивая нулю коэффициенты при положительных степенях переменной $x$, будем иметь бесконечную систему линейных уравнений

$$\frac{B_n}{n!}\cdot\frac{1}{1!}+\frac{B_{n-1}}{(n-1)!}\cdot\frac{1}{2!}+\ldots+\frac{B_0}{0!}\,\frac{1}{(n+1)!}=0 \quad (n=1,\ 2,\ 3,\ \ldots)$$

или, умножая на $(n+1)!$ и учитывая, что

$$\frac{(n+1)!}{(n-k)!\,(k+1)!}=C_{n+1}^{n-k} \quad (k=0,\ 1,\ \ldots,\ n+1),$$

получим:

$$C_{n+1}^{1}B_n+C_{n+1}^{2}B_{n-1}+\ldots+C_{n+1}^{n}B_1+1=0. \tag{4}$$

Если условно положить

$$B_k=B^k, \tag{5}$$

то формулу (4) кратко можно записать в следующем *символическом* виде:

$$(B+1)^{n+1}-B^{n+1}=0$$

или, заменяя $n+1$ на $n$,

$$(B+1)^n-B^n=0. \tag{6}$$

Полагая $n=2,\ 3,\ 4,\ \ldots$ в формуле (6), получим бесконечную систему уравнений

$$\left.\begin{aligned} 2B_1+1&=0,\\ 3B_2+3B_1+1&=0,\\ 4B_3+6B_2+4B_1+1&=0,\\ 5B_4+10B_3+10B_2+5B_1+1&=0,\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{aligned}\right\} \tag{7}$$

Отсюда последовательно находим:

$$B_1=-\frac{1}{2};\quad B_2=\frac{1}{6};\quad B_3=0;\quad B_4=-\frac{1}{30};\quad B_5=0;$$

$$B_6=\frac{1}{42};\quad B_7=0;\quad B_8=-\frac{1}{30};\quad B_9=0;\quad B_{10}=\frac{5}{66};$$

$$B_{11}=0;\quad B_{12}=-\frac{691}{2730};\quad B_{13}=0;\quad B_{14}=\frac{7}{6};\quad B_{15}=0;$$

$$B_{16}=-\frac{3617}{510};\quad B_{17}=0;\quad B_{18}=\frac{43867}{798};\quad B_{19}=0;\quad B_{20}=-\frac{174611}{330}$$

и т. д.

Таким образом, числа Бернулли могут быть шаг за шагом определены из символической формулы (6), причем после развертывания бинома по правилу Ньютона степени чисел $B$ должны быть заменены числами Бернулли с соответствующими индексами.

Функция (1) называется *производящей функцией* чисел Бернулли. Воспользовавшись обозначением (5), разложение (3) символически можно записать следующим образом:

$$\frac{x}{e^x-1}=e^{Bx}.$$

Из структуры системы (7) очевидно, что все числа Бернулли рациональны. Кроме того, обнаружилось, что числа Бернулли с нечетными индексами, кроме $B_1$, равны нулю. Докажем это свойство в общем виде. Учитывая, что

$$B_0=1 \quad \text{и} \quad B_1=-\frac{1}{2},$$

имеем:

$$\varphi(x)=\frac{x}{e^x-1}-B_1x=\frac{x}{e^x-1}+\frac{x}{2}=1+\sum_{n=2}^{\infty}\frac{B_n}{n!}x^n. \tag{8}$$

Очевидно,

$$\varphi(x)=\frac{x(e^x+1)}{2(e^x-1)}=\frac{x}{2}\cdot\frac{e^{\frac{x}{2}}+e^{-\frac{x}{2}}}{e^{\frac{x}{2}}-e^{-\frac{x}{2}}}=\frac{x}{2}\operatorname{cth}\frac{x}{2}$$

есть функция четная. Поэтому ее разложение (8) содержит лишь четные степени переменной $x$ и, следовательно,

$$B_n=0 \quad \text{при} \quad n=3,\ 5,\ 7,\ \ldots$$

Числа Бернулли находят применения во многих вопросах. В частности, они используются в важной формуле *Эйлера — Маклорена*, к выводу которой мы переходим.

## § 13*. Формула Эйлера — Маклорена

Пусть $y=f(x)$ — некоторая функция, определенная в области $x\geqslant x_0$. Рассмотрим оператор *конечной разности*

$$\Delta f(x)=f(x+h)-f(x),$$

где $h$ — фиксированная положительная величина. Под *обратным оператором* $\frac{1}{\Delta}$ от функции $f(x)$ естественно понимается функция $F(x)$, удовлетворяющая конечно-разностному уравнению

$$\Delta F(x)=f(x). \tag{1}$$

Таким образом, из уравнения (1) имеем:

$$F(x)=\frac{1}{\Delta}f(x). \tag{2}$$

Если функция $f(x)$ рассматривается на множестве равноотстоящих точек

$$x_0,\ x_1,\ x_2,\ \ldots,$$

где $\Delta x_i = x_{i+1} - x_i = h$ $(i = 0,\ 1,\ 2,\ \ldots)$, то обратный оператор $F(x_i) = \frac{1}{\Delta} f(x_i)$ легко построить. Действительно, составим конечную сумму

$$S(x_i) = \sum_{j=0}^{i-1} f(x_j) \quad (i = 1,\ 2,\ \ldots),$$

причем будем условно считать, что $S(x_0) = 0$. Очевидно, получаем:

$$\Delta S(x_i) = S(x_{i+1}) - S(x_i) = f(x_i). \tag{3}$$

С другой стороны, согласно уравнению (1) имеем:

$$\Delta F(x_i) = f(x_i). \tag{4}$$

Вычитая из равенства (4) равенство (3), находим:

$$\Delta\left[F(x_i) - S(x_i)\right] = 0$$

при $i = 0,\ 1,\ 2,\ \ldots$ Следовательно, разность $F(x_i) - S(x_i)$ не зависит от индекса $i$ и мы можем положить:

$$F(x_i) - S(x_i) = F(x_0) - S(x_0) = F(x_0);$$

отсюда

$$F(x_i) = F(x_0) + S(x_i),$$

где $F(x_0)$ — произвольная постоянная величина. Итак,

$$\frac{1}{\Delta} f(x_i) = F(x_0) + S(x_i), \tag{5}$$

т. е. *обратный оператор для конечной разности есть оператор конечного суммирования.*

Введем теперь *оператор дифференцирования*

$$Df(x) = \frac{df(x)}{dx}.$$

Под *обратным оператором* $\frac{1}{D}$ понимается операция интегрирования

$$\frac{1}{D} f(x) = \int_{x_0}^{x} f(x)\,dx.$$

Используя ряд Тейлора, находим:

$$\Delta f(x) = \sum_{k=1}^{\infty} \frac{h^k}{k!} D^k f(x) = \left\{\sum_{k=1}^{\infty} \frac{h^k D^k}{k!}\right\} f(x) = (e^{hD} - 1) f(x).$$

Следовательно,

$$\Delta = (e^{hD} - 1).$$

Отсюда для обратного оператора $\frac{1}{\Delta}$ получаем следующее выражение:

$$\frac{1}{\Delta} = \frac{1}{e^{hD} - 1}.$$

Умножая обе части последнего равенства на $hD$, имеем:

$$hD\frac{1}{\Delta} = \frac{hD}{e^{hD} - 1}.$$

В правой части получалась производящая функция для чисел Бернулли. Поэтому

$$hD\frac{1}{\Delta} = \sum_{k=0}^{\infty} \frac{B_k}{k!} h^k D$$

или более подробно

$$\frac{d}{dx}\left[\frac{1}{\Delta} f(x)\right] = \sum_{k=0}^{\infty} \frac{B_k}{k!} h^{k-1} D^k f(x). \tag{6}$$

Интегрируя равенство (6) в пределах от $x = x_0$ до $x = x_n$ и используя формулу (5), будем иметь:

$$\frac{1}{\Delta} f(x_n) - \frac{1}{\Delta} f(x_0) =$$
$$= \frac{1}{h}\int_{x_0}^{x_n} f(x)\,dx + \sum_{k=1}^{\infty} \frac{B_k}{k!} h^{k-1}\left[f^{(k-1)}(x_n) - f^{(k-1)}(x_0)\right],$$

или

$$F(x_0) + \sum_{j=0}^{n-1} f(x_j) - F(x_0) = \sum_{j=0}^{n-1} f(x_j) =$$
$$= \frac{1}{h}\int_{x_0}^{x_n} f(x)\,dx + \sum_{k=1}^{\infty} \frac{B_k}{k!} h^{k-1}\left[f^{(k-1)}(x_n) - f^{(k-1)}(x_0)\right].$$

Учитывая, что

$$B_1 = -\frac{1}{2} \quad \text{и} \quad B_{2k+1} = 0 \quad \text{при} \quad k = 1, 2, \ldots,$$

получаем *формулу Эйлера—Маклорена*

$$\int_{x_0}^{x_n} f(x)\,dx = h\left[\frac{1}{2}f(x_0)+f(x_1)+f(x_2)+\ldots+f(x_{n-1})+\frac{1}{2}f(x_n)\right]-$$

$$-\sum_{k=1}^{m}\frac{B_{2k}}{(2k)!}h^{2k}\left[f^{(2k-1)}(x_n)-f^{(2k-1)}(x_0)\right]+R_{2m}, \qquad (7)$$

где $R_{2m}$ — *остаточный член*. Запись формулы (7) в виде бесконечного ряда не всегда законна, так как ряд может расходиться. Подставляя значения чисел Бернулли, будем иметь:

$$\int_{x_0}^{x_n} y\,dx = h\left(\frac{1}{2}y_0+y_1+y_2+\ldots+y_{n-1}+\frac{1}{2}y_n\right)-\frac{h^2}{12}(y_n'-y_0')+$$

$$+\frac{h^4}{720}(y_n'''-y_0''')-\frac{h^6}{30240}(y_n^{\mathrm{V}}-y_0^{\mathrm{V}})+\ldots$$

$$\ldots-\frac{B_{2m}}{(2m)!}h^{2m}\left[f^{(2m-1)}(x_n)-f^{(2m-1)}(x_0)\right]+R_{2m}. \qquad (8)$$

Остаточный член формулы Эйлера—Маклорена имеет вид [6]

$$R_{2m} = -nh^{2m+3}\frac{B_{2m+2}}{(2m+2)!}f^{(2m+2)}(\xi),$$

где $\xi \in (x_0, x_n)$.

Формулу Эйлера—Маклорена (8) используют для приближенного вычисления определенных интегралов, а также для приближенного суммирования значений функций при равноотстоящих значениях аргумента. Действительно, из формулы (8) мы имеем:

$$\sum_{i=0}^{n} f(x_i) = \frac{1}{h}\int_{x_0}^{x_n} f(x)\,dx+\frac{f(x_0)+f(x_n)}{2}+$$

$$+\sum_{k=1}^{m}\frac{B_{2k}}{(2k)!}h^{2k-1}\left[f^{(2k-1)}(x_n)-f^{(2k-1)}(x_0)\right]+nh^{2m+2}\frac{B_{2m+2}}{(2m+2)!}f^{(2m+2)}(\xi). \qquad (9)$$

Пример 1. Используя формулу Эйлера—Маклорена, приближенно вычислить определенный интеграл

$$I = \int_{0,2}^{1}(\sin x-\ln x+e^x)\,dx.$$

Решение. Разделим отрезок [0, 2; 1], например, на восемь промежутков, беря $h=0{,}1$ и полагая

$$x_i = 0{,}2+i\cdot 0{,}1 \qquad (i=0,\ 1,\ \ldots,\ 8).$$

Результаты вычислений соответствующих значений функции $f(x) = \sin x - \ln x + e^x$ приведены в таблице 73.

Таблица 73

Значения функции $f(x) = \sin x - \ln x + e^x$

| $x$ | 0,2 | 0,3 | 0,4 | 0,5 | 0,6 |
|---|---|---|---|---|---|
| $f(x)$ | 3,02951 | 2,84936 | 2,79754 | 2,82130 | 2,89759 |
| $x$ | 0,7 | 0,8 | 0,9 | 1,0 | |
| $f(x)$ | 3,01435 | 3,16605 | 3,34830 | 3,55975 | |

Отсюда

$$\frac{1}{2}f(x_0) + f(x_1) + \ldots + f(x_7) + \frac{1}{2}f(x_8) = 24,1894.$$

Ограничиваясь производной пятого порядка, имеем:

$$f'(x) = \cos x - \frac{1}{x} + e^x,$$

$$f'''(x) = -\cos x - \frac{2}{x^3} + e^x,$$

$$f^{\mathrm{V}}(x) = \cos x - \frac{24}{x^5} + e^x.$$

Следовательно,

$$f'(0,2) = -2,7985; \qquad f'(1) = 2,2586;$$
$$f'''(0,2) = -249,7587; \qquad f'''(1) = 0,1780;$$
$$f^{\mathrm{V}}(0,2) = -74\,997,7985; \qquad f^{\mathrm{V}}(1) = -20,7415.$$

Подставив найденные значения в формулу (8), получим:

$$I = 24,1894 \cdot 0,1 - \frac{(0,1)^2}{12} \cdot (2,2586 + 2,7985) +$$
$$+ \frac{(0,1)^4}{720} \cdot (0,1780 + 249,7587) - \frac{(0,1)^6}{30\,240} \cdot (-20,7415 + 74\,997,7985) =$$
$$= 2,41894 - 0,00421 + 0,00004 = 2,41477.$$

Непосредственное интегрирование дает:

$$I = [-\cos x - x(\ln x - 1) + e^x]\Big|_{0,2}^{1} \approx 2,4148.$$

Пример 2. Найти сумму

$$\frac{1}{51^2} + \frac{1}{53^2} + \frac{1}{55^2} + \ldots + \frac{1}{99^2}.$$

Решение. В нашем случае

$$f(x)=\frac{1}{x^2};\quad h=2;\quad x_0=51;\quad x_n=99.$$

Находим производные нечетного порядка от функции $f(x)$:

$$f'(x)=-\frac{2}{x^3},$$

$$f'''(x)=-\frac{24}{x^5},$$

$$f^{\mathrm{V}}(x)=-\frac{720}{x^7},$$

$$f^{\mathrm{VII}}(x)=-\frac{40\,320}{x^9}\quad \text{и т. д.}$$

Подставляя в формулу (9) и ограничиваясь производной седьмого порядка, получим:

$$\sum_{x=51}^{x=99}\frac{1}{x^2}=\frac{1}{2}\int_{51}^{99}\frac{dx}{x^2}+\frac{1}{2}\left(\frac{1}{51^2}+\frac{1}{99^2}\right)+\frac{1}{3}\left(\frac{1}{51^3}-\frac{1}{99^3}\right)-$$

$$-\frac{4}{15}\left(\frac{1}{51^5}-\frac{1}{99^5}\right)+\frac{16}{21}\left(\frac{1}{51^7}-\frac{1}{99^7}\right)-\frac{64}{15}\left(\frac{1}{51^9}-\frac{1}{99^9}\right)=$$

$$=0{,}004\,753\,416+0{,}000\,243\,490+$$

$$+0{,}000\,002\,169-0{,}000\,000\,001=0{,}004\,999\,074.$$

Согласно формуле (9), где нужно положить $h=2$, $n=24$, $m=4$, ошибка полученного результата есть

$$R=24\cdot 2^{10}\cdot\frac{B_{10}}{8!}\cdot f^{(10)}(\xi)<24\cdot 2^{10}\cdot\frac{5}{66}\cdot\frac{1}{8!}\cdot\frac{11!}{50^{12}}<\frac{2}{25^{10}}\approx 10^{-14}.$$

## § 14. Приближенное вычисление несобственных интегралов

Интеграл

$$\int_a^b f(x)\,dx \tag{1}$$

называется *собственным*, если

1) промежуток интегрирования $[a, b]$ конечен;

2) подынтегральная функция $f(x)$ непрерывна на $[a, b]$.

В противном случае интеграл (1) называется *несобственным*.

Рассмотрим сначала приближенное вычисление несобственного интеграла

$$\int_a^{\infty} f(x)\,dx \tag{2}$$

с *бесконечным промежутком интегрирования*, где функция $f(x)$ непрерывна при $a \leqslant x < \infty$.

Интеграл (2) называется *сходящимся*, если существует конечный предел

$$\lim_{b \to \infty} \int_a^b f(x)\,dx, \tag{3}$$

и по определению полагают:

$$\int_a^\infty f(x)\,dx = \lim_{b \to \infty} \int_a^b f(x)\,dx. \tag{4}$$

Если предел (3) не существует, то интеграл (2) называется *расходящимся*, и такой интеграл считается лишенным смысла. Поэтому, прежде чем приступить к вычислению несобственного интеграла, нужно предварительно убедиться, пользуясь известными признаками сходимости [10], что этот интеграл сходится.

Чтобы вычислить сходящийся несобственный интеграл (2) с заданной точностью $\varepsilon$, представим его в виде

$$\int_a^\infty f(x)\,dx = \int_a^b f(x)\,dx + \int_b^\infty f(x)\,dx. \tag{5}$$

В силу сходимости интеграла число $b$ можно выбрать столь большим, чтобы имело место неравенство

$$\left| \int_b^\infty f(x)\,dx \right| < \frac{\varepsilon}{2}. \tag{6}$$

Собственный интеграл

$$\int_a^b f(x)\,dx$$

можно вычислить по одной из квадратурных формул. Пусть $S$ — приближенное значение этого интеграла с точностью до $\frac{\varepsilon}{2}$, т. е.

$$\left| \int_a^b f(x)\,dx - S \right| < \frac{\varepsilon}{2}. \tag{7}$$

Из формул (5), (6) и (7) имеем:

$$\left| \int_a^\infty f(x)\,dx - S \right| < \varepsilon,$$

т. е. поставленная задача будет решена.

Предположим теперь, что промежуток интегрирования $[a, b]$ конечен и подынтегральная функция $f(x)$ имеет конечное число точек разрыва на $[a, b]$. Так как в наших предположениях промежуток интегрирования можно разбить на частичные промежутки с единственной точкой разрыва подынтегральной функции, то достаточно разобрать лишь случай, когда на $[a, b]$ имеется единственная точка разрыва $c$ функции $f(x)$, причем второго рода*).

Если $c$ есть внутренняя точка отрезка $[a, b]$, то по определению полагают:

$$\int_a^b f(x)\,dx = \lim_{\substack{\delta_1 \to +0 \\ \delta_2 \to +0}} \left\{ \int_a^{c-\delta_1} f(x)\,dx + \int_{c+\delta_2}^b f(x)\,dx \right\}, \tag{8}$$

и в случае существования этого предела интеграл называют *сходящимся*, в противном случае — *расходящимся*.

Аналогично определяется сходимость несобственного интеграла (8), если точка разрыва $c$ подынтегральной функции $f(x)$ совпадает с одним из концов промежутка интегрирования $[a, b]$.

Для приближенного вычисления с заданной точностью $\varepsilon$ сходящегося несобственного интеграла (8), где точка разрыва $c \in (a, b)$, выбирают положительные числа $\delta_1$ и $\delta_2$ столь малыми, чтобы имело место неравенство

$$\left| \int_{c-\delta_1}^{c+\delta_2} f(x)\,dx \right| < \frac{\varepsilon}{2}.$$

Затем по известным квадратурным формулам приближенно вычисляют собственные интегралы

$$\int_a^{c-\delta_1} f(x)\,dx \quad \text{и} \quad \int_{c+\delta_2}^b f(x)\,dx. \tag{9}$$

---

*) Если $c$ — точка разрыва первого рода, т. е. существуют конечные односторонние пределы

$$f(c-0) = \lim_{x \to c,\ x < c} f(x) \quad \text{и} \quad f(c+0) = \lim_{x \to c,\ x > c} f(x),$$

то можно положить:

$$\int_a^b f(x)\,dx = \int_a^c f_1(x)\,dx + \int_c^b f_2(x)\,dx,$$

где

$$f_1(x) = \begin{cases} f(x), & \text{если } a \leqslant x < c; \\ f(c-0), & \text{если } x = c; \end{cases} \quad \text{и} \quad f_2(x) = \begin{cases} f(c+0), & \text{если } x = c; \\ f(x), & \text{если } c < x \leqslant b; \end{cases}$$

причем функции $f_1(x)$ и $f_2(x)$ являются непрерывными соответственно на отрезках $[a, c]$ и $[c, b]$. Таким образом, наш интеграл сводится к сумме двух собственных интегралов.

Очевидно, если $S_1$ и $S_2$ — приближенные значения интегралов (9) с точностью до $\frac{\varepsilon}{4}$, то

$$\int_a^b f(x)\,dx \approx S_1 + S_2$$

с точностью до $\varepsilon$. Если точка разрыва $c$ подынтегральной функции $f(x)$ является концевой для промежутка интегрирования $[a, b]$, то методика вычисления очевидным образом видоизменяется.

## § 15. Метод Л. В. Канторовича выделения особенностей

Часто для приближенного вычисления интеграла от разрывной функции полезным оказывается *метод Л. В. Канторовича выделения особенностей* [1], [6], [10]. Идея этого метода состоит в том, что из подынтегральной функции $f(x)$ выделяют некоторую функцию $g(x)$, имеющую те же особенности, что и функция $f(x)$, элементарно интегрирующуюся на данном промежутке $[a, b]$ и такую, чтобы разность $f(x) - g(x)$ была достаточно гладкой на отрезке $[a, b]$. Например,

$$f(x) - g(x) \in C^{(m)}[a, b], \text{ где } m \geqslant 1.$$

Тогда будем иметь:

$$\int_a^b f(x)\,dx = \int_a^b g(x)\,dx + \int_a^b [f(x) - g(x)]\,dx,$$

где первый интеграл берется непосредственно, а второй без труда вычисляется с помощью стандартных формул.

Мы рассмотрим применение этого метода для вычисления интеграла вида

$$\int_a^b \frac{\varphi(x)}{(x - x_0)^\alpha}\,dx, \tag{1}$$

где $x_0 \in [a, b]$, $0 < \alpha < 1$ и $\varphi(x)$ непрерывна на отрезке $[a, b]$.

Пусть $\varphi(x) \in C^{(m+1)}[a, b]$, т. е. $\varphi(x)$ имеет на отрезке $[a, b]$ непрерывные производные до $(m+1)$-го порядка включительно.

Используя формулу Тейлора, будем иметь:

$$\varphi(x) = \sum_{k=0}^{m} \frac{\varphi^{(k)}(x_0)}{k!}(x - x_0)^k + \psi(x), \tag{2}$$

где

$$\psi(x) = \varphi(x) - \sum_{k=0}^{m} \frac{\varphi^{(k)}(x_0)}{k!}(x - x_0)^k = \frac{\varphi^{(m+1)}(\xi)}{(m+1)!}(x - x_0)^{m+1} \tag{3}$$

$$(\xi \in (a, b)).$$

Отсюда для интеграла (1) получим:

$$\int_a^b \frac{\varphi(x)\,dx}{(x-x_0)^\alpha} = \sum_{k=0}^{m} \frac{\varphi^{(k)}(x_0)}{k!} \int_a^b (x-x_0)^{k-\alpha} dx + \int_a^b \frac{\psi(x)\,dx}{(x-x_0)^\alpha} =$$

$$= \sum_{k=0}^{m} \frac{\varphi^{(k)}(x_0)}{k!\,(k+1-\alpha)} \left[ (b-x_0)^{k+1-\alpha} - (a-x_0)^{k+1-\alpha} \right] + I, \qquad (4)$$

где

$$I = \int_a^b \frac{\psi(x)\,dx}{(x-x_0)^\alpha}. \qquad (5)$$

Из формулы (3) вытекает, что

$$\frac{\psi(x)}{(x-x_0)^\alpha} \in C^{(m)}[a, b]$$

(по меньшей мере!), следовательно, интеграл (5) является собственным и может быть вычислен с любой степенью точности по подходящей квадратурной формуле.

Метод Канторовича применим также к несобственным интегралам, подынтегральная функция которых имеет несколько точек разрыва рассмотренного типа. В этом случае для вычисления интеграла достаточно разбить промежуток интегрирования на части, содержащие лишь одну особую точку подынтегральной функции, и воспользоваться свойством аддитивности интеграла.

Пример 1. Приближенно вычислить несобственный интеграл [11]

$$I = \int_0^{\frac{1}{2}} \frac{dx}{\sqrt{x(1-x)}}.$$

Решение. Подынтегральная функция

$$f(x) = x^{-\frac{1}{2}}(1-x)^{-\frac{1}{2}}$$

имеет на отрезке $\left[0, \frac{1}{2}\right]$ единственную особую точку $x=0$. Разложим в ряд Тейлора по степеням $x$ функцию

$$\varphi(x) = (1-x)^{-\frac{1}{2}}$$

с точностью до $x^4$. Применяя бином Ньютона, будем иметь:

$$\varphi(x) = 1 + \frac{1}{2}x + \frac{3}{8}x^2 + \frac{5}{16}x^3 + \frac{35}{128}x^4.$$

Отсюда

$$I=\int_0^{\frac{1}{2}} x^{-\frac{1}{2}}dx+\frac{1}{2}\int_0^{\frac{1}{2}} x^{\frac{1}{2}}dx+$$

$$+\frac{3}{8}\int_0^{\frac{1}{2}} x^{\frac{3}{2}}dx+\frac{5}{16}\int_0^{\frac{1}{2}} x^{\frac{5}{2}}dx+\frac{35}{128}\int_0^{\frac{1}{2}} x^{\frac{7}{2}}dx+I_1=\frac{715801}{645120}\sqrt{2}+I_1=$$

$$=1{,}5691585+I_1, \qquad (6)$$

где

$$I_1=\int_0^{\frac{1}{2}} \frac{\psi(x)}{\sqrt{x}}dx \qquad (7)$$

и

$$\psi(x)=\frac{1}{\sqrt{1-x}}-\left(1+\frac{1}{2}x+\frac{3}{8}x^2+\frac{5}{16}x^3+\frac{35}{128}x^4\right); \qquad \psi(0)=0.$$

Собственный интеграл (7) вычисляем по формуле Симпсона,

Таблица 74

**Вычисление интеграла по формуле Симпсона**

| $i$ | $x_i$ | $y_{2j-1}$ | $y_{2j}$ |
|---|---|---|---|
| 0 | 0 | | 0,000000 |
| 1 | 0,05 | | |
| 2 | 0,10 | 0,000000 | 0,000009 |
| 3 | 0,15 | | |
| 4 | 0,20 | 0,000056 | 0,000216 |
| 5 | 0,25 | | |
| 6 | 0,30 | 0,000624 | 0,001508 |
| 7 | 0,35 | | |
| 8 | 0,40 | 0,003225 | 0,006316 |
| 9 | 0,45 | | |
| 10 | 0,50 | 0,011588 | 0,020239 |
| Σ | | 0,015493 | 0,008049 |

приняв $n=10$ и шаг $h=\frac{1}{20}=0{,}05$. Результаты вычислений с точностью до шести знаков приведены в таблице 74.

Отсюда

$$I_1 = \frac{1}{20\cdot 3}(0{,}020239 + 4\cdot 0{,}015493 + 2\cdot 0{,}008049) = \\ = \frac{1}{60}\cdot 0{,}098309 = 0{,}0016385.$$

Следовательно, в силу формулы (6) имеем:

$$I = \left.\begin{matrix} 1{,}5691585 \\ +\,0{,}0016385 \end{matrix}\right\} = 1{,}5707970.$$

Заметим, что интеграл $I$ находится элементарно и его точное значение есть

$$I = \frac{\pi}{2} = 1{,}5707963\ldots$$

Замечание. В некоторых случаях несобственный интеграл может быть преобразован в собственный с помощью замены переменной или интегрирования по частям.

Пример 2. Преобразовать в собственный интеграл

$$I = \int_1^\infty \frac{dx}{(1+x)\sqrt{x}}. \tag{8}$$

Решение. Полагая в интеграле (8) $x = \frac{1}{z}$, получим интеграл с конечными пределами

$$I = \int_0^1 \frac{dz}{(z+1)\sqrt{z}} = \int_0^1 \frac{dx}{(1+x)\sqrt{x}}, \tag{9}$$

но имеющий особенность при $x = 0$.

Производя надлежащее интегрирование по частям, будем иметь:

$$I = \int_0^1 \frac{1}{1+x}\, d(2\sqrt{x}) = \frac{2\sqrt{x}}{1+x}\Bigg|_0^1 + \int_0^1 2\sqrt{x}\,\frac{dx}{(1+x)^2} = 1 + 2\int_0^1 \frac{\sqrt{x}}{(1+x)^2}\,dx,$$

причем оставшийся интеграл есть собственный, и применение к нему квадратурных формул не встречает затруднений. Употребляются также другие приемы преобразования несобственных интегралов [6].

## § 16. Графическое интегрирование

Задача графического интегрирования состоит в следующем: по данному графику непрерывной функции $y = f(x)$ требуется построить график ее первообразной функции

$$F(x) = \int_a^x f(x)\,dx.$$

Иными словами, нужно построить такую кривую $y=F(x)$, ордината в каждой точке $x$ которой численно равна площади криволинейной трапеции с основанием $[a, x]$, ограниченной данной кривой $y=f(x)$.

Для приближенного построения графика первообразной функции $y=F(x)$ разбиваем площадь соответствующей криволинейной трапеции, ограниченной кривой $y=f(x)$, на узкие вертикальные полоски с помощью ординат, проведенных в точках $x_0, x_1, \ldots$ $(a=x_0<x_1<x_2<\ldots)$ (рис. 77). Каждую из таких полосок заменяем,

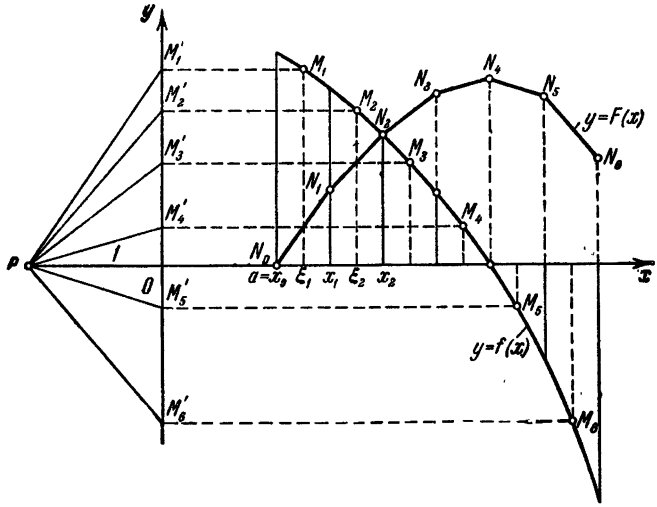


Рис. 77.

используя теорему о среднем, равновеликим (по возможности) прямоугольником с тем же основанием и высотой, равной $f(\xi_i)$, где $\xi_i$ $(i=1, 2, \ldots)$ — некоторая промежуточная точка $i$-го по счету отрезка $[x_{i-1}, x_i]$, т. е. полагаем:

$$\int_{x_{i-1}}^{x_i} f(x)\,dx = f(\xi_i)\ (x_i - x_{i-1}),$$

где

$$x_{i-1} \leqslant \xi_i \leqslant x_i \qquad (i=1, 2, \ldots).$$

Значения первообразной функции

$$F(x) = \int_{x_0}^{x} f(x)\,dx$$

в точках $x_i$ можно подсчитать *методом накопления*:

$F(x_0)=0$;

$$F(x_i)=\int_{x_0}^{x_i} f(x)\,dx=\int_{x_0}^{x_{i-1}} f(x)\,dx+\int_{x_{i-1}}^{x_i} f(x)\,dx=$$

$$=F(x_{i-1})+f(\xi_i)(x_i-x_{i-1}) \qquad (i=1,\ 2,\ \ldots). \qquad (1)$$

Пусть $M_1(\xi_1, f(\xi_1))$, $M_2(\xi_2, f(\xi_2))$, ... — соответствующие точки кривой $y=f(x)$. Проектируя их на ось $Oy$, получим точки $M_1'$, $M_2'$, ... (рис. 77).

Выберем теперь полюс $P$ с расстоянием $OP=1$ и проведем лучи $PM_1'$, $PM_2'$, ... Искомую линию $y=F(x)$ приближенно можно заменить ломаной $N_0N_1N_2N_3\ldots$ с вершинами $N_0(x_0,\ 0)$, $N_1(x_1,\ F(x_1))$, $N_2(x_2,\ F(x_2))$, ... Последовательные звенья этой ломаной будут параллельны соответствующим лучам, а именно: $N_0N_1 \parallel PM_1'$; $N_1N_2 \parallel PM_2'$; $N_2N_3 \parallel PM_3'$; ... В самом деле, угловой коэффициент звена $N_{i-1}N_i$ на основании формулы (1) равен

$$k=\frac{F(x_i)-F(x_{i-1})}{x_i-x_{i-1}}=f(\xi_i),$$

в силу же построения угловой коэффициент луча $OM_i'$ есть

$$k_i'=\frac{f(\xi_i)}{1}=f(\xi_i).$$

Следовательно,

$$N_{i-1}N_i \parallel OM_i' \qquad (i=1,\ 2,\ \ldots).$$

Таким образом, технически построение графика функции $y=F(x)$ может быть осуществлено так: из точки $N_0(x_0,\ 0)$ проводим прямую $N_0N_1$, параллельную лучу $OM_1'$, до пересечения в точке $N_1$ с вертикалью $x=x_1$; из точки $N_1$ проводим прямую $N_1N_2$, параллельную лучу $OM_2'$, до пересечения в точке $N_2$ с вертикалью $x=x_2$ и т. д.

Следует отметить, что при применении данного метода графического интегрирования точки $x_i$ $(i=0,\ 1\ \ldots)$ не обязательно брать равноотстоящими. Для увеличения точности построения рекомендуется характерные точки графика интегрируемой функции (нули, точки экстремума, точки перегиба) обязательно включать в состав точек $x_i$.

Графическое интегрирование обладает, вообще говоря, малой точностью. Поэтому этот прием полезно использовать тогда, когда требуется иметь общее представление об интеграле функции или когда подынтегральная функция задана графически и ее аналитическое выражение нам неизвестно.

## § 17*. Понятие о кубатурных формулах

*Кубатурные формулы* или, иначе, *формулы численных кубатур* предназначены для численного вычисления двойных интегралов [1].

Пусть функция $z=f(x, y)$ определена и непрерывна в некоторой ограниченной области $\sigma$ (рис. 78). В этой области $\sigma$ выбирается система точек (узлов) $M_i(x_i, y_i)$ $(i=1, 2, \ldots, N)$. Для вычисления двойного интеграла $\iint\limits_{(\sigma)} f(x, y)\,dx\,dy$ приближенно полагают:

$$\iint\limits_{(\sigma)} f(x, y)\,dx\,dy = \sum_{i=1}^{N} A_i f(x_i, y_i). \qquad (1)$$

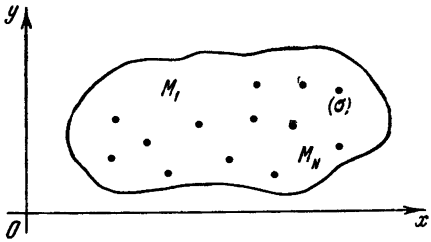

Рис. 78.

Чтобы найти коэффициенты $A_i$, потребуем точного выполнения кубатурной формулы (1) для всех полиномов

$$P_n(x, y) = \sum_{k+l\leqslant n} c_{kl} x^k y^l, \qquad (2)$$

степень которых не превышает заданного числа $n$. Для этого необходимо и достаточно, чтобы формула (1) была точной для произведения степеней

$$x^k y^l$$

$$(k, l=0, 1, 2, \ldots, n;\ k+l\leqslant n).$$

Полагая в (1) $f(x, y)=x^k y^l$, будем иметь:

$$I_{kl} = \iint\limits_{(\sigma)} x^k y^l\,dx\,dy = \sum_{i=1}^{N} A_i x_i^k y_i^l$$

$$(k, l=0, 1, 2, \ldots, n;\ k+l\leqslant n). \qquad (3)$$

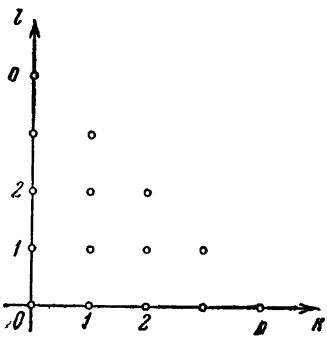

Рис. 79.

Таким образом, коэффициенты $A_i$ формулы (1), вообще говоря, могут быть определены из системы линейных уравнений (3).

Для того чтобы система (3) была определенной, необходимо, чтобы число неизвестных $N$ было равно числу уравнений. Отсюда, составляя «решетку показателей» (рис. 79), получаем:

$$N=(n+1)+n+\ldots+1=\frac{(n+1)(n+2)}{2}.$$

Остается открытым трудный вопрос о наиболее рациональном выборе узлов для данной области.

Можно указать еще один достаточно общий прием вычисления двойного интеграла. Пусть область интегрирования ограничена непрерывными однозначными кривыми

$$y=\varphi(x), \quad y=\psi(x) \quad (\varphi(x)\leqslant\varphi(x))$$

и двумя вертикалями $x=a$, $x=b$ (рис. 80).

Расставляя по известным правилам в двойном интеграле

$$I=\iint\limits_{(\sigma)} f(x, y)\,dx\,dy \tag{4}$$

пределы интегрирования, будем иметь:

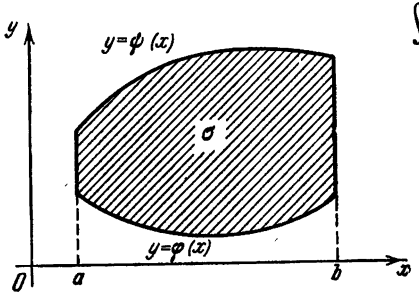


Рис. 80.

$$\iint\limits_{(\sigma)} f(x, y)\,dx\,dy=$$

$$=\int\limits_a^b dx\int\limits_{\varphi(x)}^{\psi(x)} f(x, y)\,dy.$$

Пусть

$$F(x)=\int\limits_{\varphi(x)}^{\psi(x)} f(x, y)\,dy. \tag{5}$$

Тогда

$$\iint\limits_{(\sigma)} f(x, y)\,dx\,dy=\int\limits_a^b F(x)\,dx. \tag{6}$$

Применяя к однократному интегралу, стоящему в правой части равенства (6), одну из квадратурных формул, получим:

$$\iint\limits_{(\sigma)} f(x, y)\,dx\,dy=\sum_{i=1}^{n} A_i F(x_i), \tag{7}$$

где $x_i\in[a, b]$ $(i=1, 2, \ldots, n)$ и $A_i$ — некоторые постоянные коэффициенты. В свою очередь значения

$$F(x_i)=\int\limits_{\varphi(x_i)}^{\psi(x_i)} f(x_i, y)\,dy$$

могут быть также найдены по некоторым формулам квадратур

$$F(x_i)=\sum_{j=1}^{m_i} B_{ij} f(x_i, y_j),$$

где $B_{ij}$ — соответствующие постоянные.

Из формулы (7) выводим:

$$\iint\limits_{(\sigma)} f(x,\ y)dx\,dy=\sum_{i=1}^{n}\sum_{j=1}^{m_i}A_iB_{ij}f(x_i,\ y_j), \tag{8}$$

где $A_i$ и $B_{ij}$ — известные постоянные.

Геометрически этот метод эквивалентен вычислению объма $I$, выражаемого интегралом (4) с помощью поперечных сечений.

Для кубатурных формул типа (8) сохраняют силу с соответствующими видоизменениями общие замечания, относящиеся к вычислению однократных интегралов (см. § 10).

## § 18*. Кубатурная формула типа Симпсона

Пусть сначала область интегрирования есть прямоугольник

$$R\{a\leqslant x\leqslant A;\ b\leqslant y\leqslant B\}$$

(рис. 81), стороны которого параллельны осям координат. Каждый из промежутков $[a,\ A]$ и $[b,\ B]$ разобьем пополам точками

$$x_0=a,\ x_1=a+h,\ x_2=a+2h=A$$

и соответственно

$$y_0=b,\ y_1=b+k,\ y_2=b+2k=B,$$

где

$$h=\frac{A-a}{2},\quad k=\frac{B-b}{2}.$$

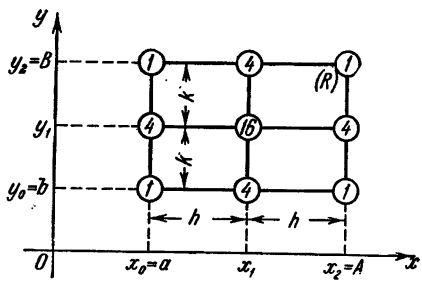


Рис. 81.

Всего, таким образом, получим девять точек $(x_i,\ y_j)$ $(i,\ j=0,\ 1,\ 2,\ \ldots,\ 9)$. Имеем:

$$\iint\limits_{(R)} f(x,\ y)\,dx\,dy=\int_a^A dx\int_b^B f(x,\ y)\,dy. \tag{1}$$

Отсюда, вычисляя внутренний интеграл по квадратурной формуле Симпсона, находим:

$$\iint f(x,\ y)\,dx\,dy=\int_a^A dx\cdot\frac{k}{3}[f(x,\ y_0)+4f(x,\ y_1)+f(x,\ y_2)]=$$

$$=\frac{k}{3}\left[\int_a^A f(x,\ y_0)\,dx+4\int_a^A f(x,\ y_1)\,dx+\int_a^A f(x,\ y_2)\,dx\right].$$

Применяя к каждому интегралу снова формулу Симпсона, получим:

$$\iint\limits_{(R)} f(x, y)\,dx\,dy = \frac{hk}{9}\{[f(x_0, y_0) + 4f(x_1, y_0) + f(x_2, y_0)] + \\ + 4[f(x_0, y_1) + 4f(x_1, y_1) + f(x_2, y_1)] + \\ + [f(x_0, y_2) + 4f(x_1, y_2) + f(x_2, y_2)]\}$$

или

$$\iint\limits_{(R)} f(x, y)\,dx\,dy = \frac{hk}{9}\{[f(x_0, y_0) + f(x_2, y_0) + f(x_0, y_2) + \\ + f(x_2, y_2)] + 4[f(x_1, y_0) + f(x_0, y_1) + \\ + f(x_2, y_1) + f(x_1, y_2)] + 16f(x_1, y_1)\}. \quad (2)$$

Формулу (2) будем называть *кубатурной формулой Симпсона.* Следовательно,

$$\iint\limits_{(R)} f(x, y)\,dx\,dy = \frac{hk}{9}(\sigma_0 + 4\sigma_1 + 16\sigma_2), \quad (2')$$

где $\sigma_0$ — сумма значений подынтегральной функции $f(x, y)$ в вершинах прямоугольника $R$, $\sigma_1$ — сумма значений $f(x, y)$ в серединах сторон прямоугольника $R$, $\sigma_2 = f(x_1, y_1)$ — значение функции $f(x, y)$ в центре прямоугольника $R$. Кратности этих значений обозначены на рис. 81.

Пример 1. Применяя кубатурную формулу Симпсона, вычислить двойной интеграл [7]

$$I = \int\limits_4^{4,4} \int\limits_2^{2,6} \frac{dx\,dy}{xy}.$$

Решение. Берем

$$h = \frac{4,4 - 4}{2} = 0,2 \quad \text{и} \quad k = \frac{2,6 - 2}{2} = 0,3.$$

Соответствующие значения подынтегральной функции $z = \frac{1}{xy}$ помещены в таблице 75.

Таблица 75

**Вычисление двойного интеграла по формуле Симпсона**

| $y_j$ \ $x_i$ | 4,0 | 4,2 | 4,4 |
|---|---|---|---|
| 2,0 | 0,125000 | 0,119048 | 0,113636 |
| 2,3 | 0,108696 | 0,103520 | 0,0988142 |
| 2,6 | 0,096154 | 0,0915751 | 0,0874126 |

Применяя кубатурную формулу (2), получим:

$$I=\frac{0,2\cdot 0,3}{9}[(0,125000+0,113636+0,096154+0,0874126)+$$
$$+4(0,119048+0,108696+0,0988142+0,0915751)+$$
$$+16\cdot 0,103520]=0,0250070.$$

Точное значение этого двойного интеграла будет:

$$\int_{4}^{4,4}\int_{2}^{2,6}\frac{dx\,dy}{xy}=\ln 1,3\cdot\ln 1,1=0,0953108\cdot 0,262364=0,0250061.$$

Следовательно, остаточная погрешность

$$\Delta=|0,025006-0,0250070|=0,0000009\approx 10^{-6}.$$

Если размеры прямоугольника $R\{a\leqslant x\leqslant A;\ b\leqslant y\leqslant B\}$ велики, то для увеличения точности кубатурной формулы (2) область $R$ разбивают на систему прямоугольников, к каждому из которых применяют кубатурную формулу Симпсона.

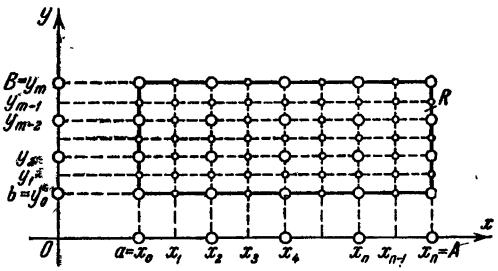

Рис. 82.

Положим, что стороны прямоугольника $R$ мы разделили соответственно на $n$ и $m$ равных частей; в результате получилась относительно крупная сеть $nm$ прямоугольников (на рис. 82 вершины этих прямоугольников отмечены более крупными кружками). Каждый из этих прямоугольников в свою очередь разделим на четыре равные части. Вершины этой последней мелкой сети прямоугольников примем за узлы $M_{ij}$ кубатурной формулы.

Пусть

$$h=\frac{A-a}{2n}$$

и

$$k=\frac{B-b}{2m}.$$

Тогда сеть узлов будет иметь следующие координаты:

$$x_i=x_0+ih \qquad (x_0=a;\ i=0,\ 1,\ 2,\ \ldots,\ 2n)$$

и

$$y_j=y_0+jk \qquad (y_0=b;\ j=0,\ 1,\ 2,\ \ldots,\ 2m).$$

Для сокращения введем обозначение

$$f(x_i, y_i) = f_{ij}.$$

Применяя формулу (2) к каждому из прямоугольников крупной сети, будем иметь (рис. 82):

$$\iint\limits_{(R)} f(x, y)\,dx\,dy = \frac{hk}{9}\sum_{i=0}^{n}\sum_{j=0}^{m}[(f_{2i,\,2j} + f_{2i+2,\,2j} + f_{2i+2,\,2j+2} + f_{2i,\,2j+2}) + 4(f_{2i+1,\,2j} + f_{2i+2,\,2j+1} + f_{2i+1,\,2j+2} + f_{2i,\,2j+1}) + 16 f_{2i+1,\,2j+1}].$$

Отсюда, делая приведение подобных членов, окончательно находим:

$$\iint\limits_{(R)} f(x, y)\,dx\,dy = \frac{hk}{9}\sum_{i=0}^{2n}\sum_{j=0}^{2m}\lambda_{ij} f_{ij}, \tag{3}$$

где коэффициенты $\lambda_{ij}$ являются соответствующими элементами матрицы

$$\Lambda = \begin{bmatrix} 1 & 4 & 2 & 4 & 2 & \dots & 4 & 2 & 4 & 1 \\ 4 & 16 & 8 & 16 & 8 & \dots & 16 & 8 & 16 & 4 \\ 2 & 8 & 4 & 8 & 4 & \dots & 8 & 4 & 8 & 2 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 2 & 8 & 4 & 8 & 4 & \dots & 8 & 4 & 8 & 2 \\ 4 & 16 & 8 & 16 & 8 & \dots & 16 & 8 & 16 & 4 \\ 1 & 4 & 2 & 4 & 2 & \dots & 4 & 2 & 4 & 1 \end{bmatrix}.$$

Если область интегрирования $\sigma$ — криволинейная, то строим

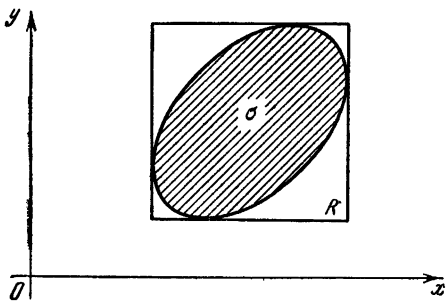


Рис. 83.

прямоугольник $R \supset \sigma$, стороны которого параллельны осям координат (рис. 83). Рассмотрим вспомогательную функцию

$$f^*(x, y) = \begin{cases} f(x, y), & \text{если } (x, y) \in \sigma; \\ 0, & \text{если } (x, y) \in R - \sigma. \end{cases}$$

В таком случае, очевидно, имеем:

$$\iint\limits_{(\sigma)} f(x, y)\,dx\,dy = \iint\limits_{(R)} f^*(x, y)\,dx\,dy.$$

Последний интеграл приближенно может быть вычислен по общей кубатурной формуле (3).

## Литература к шестнадцатой главе


1. Ш. Е. Микеладзе, Численные методы математического анализа, Гостехиздат, 1953, гл. XIII, XVIII.
2. В. Э. Милн, Численный анализ, ИЛ, 1951, гл. IV.
3. С. М. Никольский, Квадратурные формулы, Физматгиз, М., 1958.
4. А. Марков, Исчисление конечных разностей, изд. 2. Матезис, 1911, гл. V.
5. И. Ф. Стеффенсон, Теория интерполяции, М.—Л., 1935.
6. И. С. Березин и Н. П. Жидков, Методы вычислений, Физматгиз, М., 1959, т. I, гл. III.
7. Дж. Скарборо, Численные методы математического анализа, ГТТИ, 1934, гл, VII.
8. А. Н. Крылов, Лекции о приближенных вычислениях, изд. 2, ИАН СССР, 1933, гл. III.
9. В. И. Крылов, Приближенное вычисление интегралов, Физматгиз, М., 1959.
10. М. Дж. Сальвадори, Численные методы в технике, ИЛ, М., 1955.
11. Г. М. Фихтенгольц, Курс дифференциального и интегрального исчисления, 1948, Гостехиздат, т. 2, гл. IX, XIII.

## ГЛАВА XVII

# МЕТОД МОНТЕ-КАРЛО

## § 1. Идея метода Монте-Карло

Обычный путь решения задачи состоит в том, что указывается *алгоритм* (последовательность действий), с помощью которого искомая величина $f$ находится или точно, или с заданной погрешностью. А именно, если через $f_1, f_2, \ldots, f_n, \ldots$ обозначить соответствующие результаты последовательно накопляющихся действий, то

$$f = \lim_{n \to \infty} f_n, \tag{1}$$

причем в случае конечного числа операций процесс обрывается на некотором шаге. Здесь процесс вычислений является строго детерминированным: два различных вычислителя при отсутствии ошибок приходят к одному и тому же результату.

Однако встречаются задачи, где построение такого рода алгоритма практически невыполнимо или сам алгоритм оказывается чрезмерно сложным. В этих случаях часто прибегают к моделированию математической или физической сущности задачи и использованию законов больших чисел теории вероятностей. Оценки $f_1, f_2, \ldots, f_n, \ldots$ искомой величины $f$ получаются на основании статистической обработки материала, связанного с результатами некоторых многократных *случайных испытаний*. При этом требуется, чтобы случайная величина $f_n$ при $n \to \infty$ по вероятности сходилась к искомой величине $f$ [1], [2], т. е. для любого $\varepsilon > 0$ должно иметь место предельное соотношение

$$\lim_{n \to \infty} \boldsymbol{P}(|f - f_n| < \varepsilon) = 1, \tag{2}$$

где $\boldsymbol{P}$ обозначает соответствующую вероятность.

Выбор величины $f_n$ обусловливается конкретными особенностями задачи. Например, часто искомую величину $f$ трактуют как вероятность некоторого случайного события (или, более общо, как математическое ожидание некоторой случайной величины). Тогда частоту $f_n$ появления события при $n$ соответствующих случайных испытаниях (или соответственно эмпирическое среднее значений случайной величины) в ши-

рокихпредположенияхможно рассматривать какверoятностнуюоценку искомой величины. Возможны также и другие варианты. Заметим, что в этих случаях вычислительный процесс является недетерминированным, так как он определяется итогами случайных испытаний.

Способы решения задач, использующие случайные величины, получили общее название *метода Монте-Карло*. Более точно под *методом Монте-Карло* [3], [4], [5], [6] понимается совокупность приемов, позволяющих получать решения математических или физических задач при помощи многократных случайных испытаний. Оценки искомой величины выводятся статистическим путем и носят вероятностный характер. На практике случайные испытания заменяются результатами некоторых вычислений, производимых над *случайными числами* (см. § 2).

Эффективное применение метода Монте-Карло стало возможным после появления быстродействующих электронных машин, так как для получения достаточно точной оценки искомой величины требуются выполнение вычислений для весьма большого количества частных случаев и последующая статистическая обработка колоссального числового материала. Заметим, что при пользовании методом Монте-Карло нет необходимости знать точные соотношения между данными и искомыми величинами задачи, а достаточно лишь выявить тот комплекс условий, при наличии которого соответствующее явление имеет место. Это обстоятельство делает возможным использование метода Монте-Карло для решения логических задач.

Из математических задач, для которых разработано применение метода Монте-Карло, отметим следующие: решение систем линейных уравнений, обращение матриц, нахождение собственных значений и собственных векторов матрицы, вычисление кратных интегралов, решение задачи Дирихле, решение функциональных уравнений различных типов и др. Метод Монте-Карло успешно используется также для решения задач ядерной физики. Заметим, что для решения одной и той же конкретной задачи схема применения метода может быть существенно различной.

В этой главе будет рассмотрено вычисление кратных интегралов и решение систем линейных уравнений методом Монте-Карло. Что касается других указанных задач, то сведения о них можно почерпнуть в специальной литературе (см., например [3], библиография; а также [6]).

## § 2. Случайные числа

При практическом применении метода Монте-Карло случайные испытания обычно заменяют выборкой *случайных чисел*.

Определение 1. Величина называется *случайной*, если она принимает те или иные значения в зависимости от появления или непоявления некоторого случайного события.

Случайная величина $X$ задается законом распределения

$$P(X < x) = \Phi(x),$$

где $x$ — любое действительное число и $\Phi(x)$ — известная функция *(функция распределения)*. Значения случайной величины называются *случайными числами*.

Определение 2. Если случайная величина имеет заданный закон распределения [1], [2] (равномерный, нормальный и т. п.), то будем говорить, что соответствующие случайные числа *распределены по этому закону*.

Пусть числа $x_1, x_2, \ldots x_n, \ldots$ являются значениями одной и той же случайной величины $X$ при независимых между собой испытаниях с повторяющимися условиями. Тогда *последовательность случайных чисел*

$$\{x_n\} \tag{1}$$

будем называть *случайной*, с соответствующим законом распределения. В дальнейшем, как правило, мы будем рассматривать *равномерно распределенные* на единичном отрезке $0 \leqslant x \leqslant 1$ случайные последовательности (1). Если $(a, b)$ — любой промежуток*) из отрезка $[0, 1]$ и $\nu_n = \nu_n(a, b)$ — число элементов конечной подпоследовательности $x_1, x_2, \ldots, x_n$, принадлежащих промежутку $(a, b)$, то для равномерно распределенной последовательности (1) имеет место предельное соотношение

$$\lim_{n \to \infty} \frac{\nu_n(a, b)}{n} = b - a, \tag{2}$$

т. е. *предельная относительная частота равномерно распределенной на* $[0,1]$ *последовательности* $\{x_n\}$ *для каждого частичного промежутка* $(a, b)$ *равна длине этого промежутка*.

Если случайная последовательность $\{x_n\}$ равномерно распределена на отрезке $[0, 1]$, то линейное преобразование

$$y_n = A + (B - A)\, x_n \qquad (n = 1, 2, \ldots), \tag{3}$$

где $A$ и $B$ — данные числа, приводит к случайной последовательности $\{y_n\}$, равномерно распределенной на отрезке $[A, B]$.

Вообще, имея случайную последовательность $\{x_n\}$, равномерно распределенную на отрезке $[0, 1]$, можно построить случайную последовательность $\{y_n\}$ с заданным законом распределения $\Phi(y)$. А именно, пусть

$$\Phi(y) = \int_{-\infty}^{y} \varphi(t)\, dt$$

— соответствующая функция распределения**), где $\varphi(t)$ — *плотность вероятности*.

---

*) Концы $a$ и $b$ по договоренности могут как включаться в промежуток $(a, b)$, так и не включаться в него.

**) Если $y_n$ $(n = 1, 2, \ldots)$ содержатся в конечном отрезке $A \leqslant y \leqslant B$, то, как обычно, полагают $\varphi(y) = 0$ при $y \overline{\in} [A, B]$.

Для простоты будем предполагать, что функция

$$x = \Phi(y)$$

непрерывна и строго монотонна (рис. 84). Тогда для каждого $x_n$, определяя $y_n$ из уравнения

$$x_n = \Phi(y_n) \qquad (n = 1, 2, \ldots),$$

получим случайную последовательность $\{y_n\}$, имеющую заданный закон распределения $\Phi(y)$. А именно, по способу построения, для

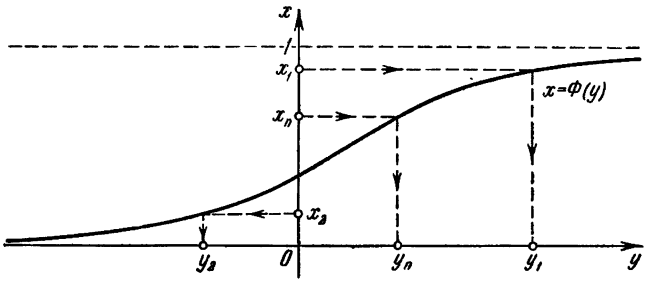


Рис. 84.

последовательности $\{y_n\}$ будет справедливо предельное соотношение

$$\lim_{n \to \infty} \frac{\tilde{\nu}_n(a, b)}{n} = \int_a^b \varphi(y)dy, \tag{4}$$

где $\tilde{\nu}_n(a, b)$ — число элементов конечной подпоследовательности $y_1, \ldots, y_n$, принадлежащих произвольному промежутку $(a, b)$.

В частности, полагая

$$\varphi(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}},$$

этим способом получим каноническую нормально распределенную (гауссову) случайную последовательность $\{y_n\}$, соответствующую случайной величине $Y$ с математическим ожиданием $MY = 0$ и дисперсией $DY = 1$. Линейное преобразование

$$z_n = \sigma y_n + c \qquad (n = 1, 2, \ldots)$$

дает нормально распределенную случайную последовательность $\{z_n\}$, соответствующую случайной величине $Z$, для которой математическое ожидание $MZ = c$ и дисперсия $DZ = \sigma^2$.

## § 3. Способы получения случайных чисел

Для выработки случайных чисел можно использовать результаты случайных физических процессов (например, бросание игральной кости, вращение рулетки, вспышки в счетчике Гейгера, шум при электрических передачах и т. п.). Имеются также готовые таблицы случайных чисел (см., например, [7], [8]).

Строго говоря, при пользовании механическими приспособлениями для получения случайных чисел нет абсолютной уверенности, что мы имеем дело со случайными событиями с заданным распределением вероятностей. Поэтому полученный материал обычно подвергается статистической «проверке на случайность». В этом смысле надежнее употреблять случайные числа из таблиц, где такая проверка уже проделана; однако использование таблиц случайных чисел для решения задач на электронных цифровых машинах часто связано с серьезными неудобствами [9].

Для решения задач методом Монте-Карло обычно требуется весьма большое количество случайных чисел. Эти числа практически наиболее удобно получаются с помощью *специальных датчиков* случайных чисел, подключаемых к машине. Действие этих датчиков регулируется случайными физическими процессами (например, радиоактивным распадом, шумами в электронных лампах и т. п.) [9].

Так как воспроизводство случайных чисел, отвечающих данной теоретической модели, есть процесс весьма тонкий и сложный, то на практике часто ограничиваются получением так называемых *псевдослучайных чисел*, в основных чертах похожих на соответствующие случайные. Источниками (датчиками) псевдослучайных чисел служат достаточно сложные математические алгорифмы. В дальнейшем под термином «случайное число» мы будем понимать как случайные, так и псевдослучайные числа, если различие между ними не имеет существенного значения.

Укажем некоторые простые приемы получения случайных чисел, в обобщенном смысле, равномерно распределенных на отрезке [0, 1]. Для простоты будем предполагать, что эти числа представляют собой правильные десятичные дроби с фиксированным количеством десятичных знаков после запятой, например $s$ (*s-разрядные* десятичные дроби), т. е. могут быть записаны в виде

$$x=\frac{\alpha_1}{10}+\frac{\alpha_2}{10^2}+\ldots+\frac{\alpha_s}{10^s}, \tag{1}$$

где $\alpha_i$ $(i=1, 2, \ldots, s)$ — цифры этого числа, принимающие значения 0, 1, 2, 3, 4, 5, 6, 7, 8, 9.

Для составления таблицы случайных чисел вида (1), равномерно распределенных на [0, 1], достаточно указать способы получения цифр $\alpha_i$ с соблюдением следующих условий:

а) $\alpha$ есть *случайная выборка* из системы чисел 0—9, причем все указанные значения равновероятны;

б) выбор предыдущей цифры $\alpha_i$ никоим образом не влияет на выбор последующей $\alpha_{i+1}$.

Для получения $s$-разрядного случайного числа такая выборка производится $s$ раз.

Система выбора, удовлетворяющая условиям а) и б), практически может быть реализована многими способами. Рассмотрим некоторые из них.

1. В урну опускают десять одинаковых перенумерованных шаров с номерами 0—9. Из урны последовательно извлекается шар и записывается его номер $\alpha$. После каждого извлечения шар возвращают в урну и перед каждым следующим тиражом все шары в урне перемешиваются.

2. Одновременно бросают две игральные кости. Если $n_1$ и $n_2$ — числа выпавших очков ($n_1, n_2 = 1, 2, 3, 4, 5, 6$) соответственно на первой и второй костях (кости должны быть различаемыми), то очередная цифра $\alpha$ случайного числа берется равной остатку от деления суммы $6(n_1-1)+n_2$ на 10, где $n_1 < 6$, т. е. $\alpha$ есть целое неотрицательное число, меньшее 10, удовлетворяющее сравнению*),

$$6(n_1-1)+n_2 \equiv \alpha \pmod{10}. \tag{2}$$

Если $n_1 = 6$, то кости перебрасываются. Из формулы (2) вытекает, что цифра $\alpha$ с равной вероятностью может принять любое значение от 0 до 9 (см. [7]).

3. Берется $s$-разрядное целое число. Число возводится в квадрат и выбираются средние $s$ его разрядов; затем процесс повторяется. Если $s$ достаточно велико, например $s \geqslant 10$, то выбираемые разряды могут быть приняты за наборы элементов $s$-разрядных псевдослучайных чисел [3].

Для получения последовательности псевдослучайных чисел можно также использовать умножение многозначного числа на постоянный множитель и извлечение средних разрядов или возведение многозначного числа в квадрат и приведение результата по модулю некоторого достаточно большого простого числа.

4. Псевдослучайная последовательность $\{x_n\}$ вырабатывается с помощью процесса [10]

$$x_n = 2^{-42} u_n,$$

где

$$u_0 = 1, \quad u_{n+1} \equiv 5^{17} u_n \pmod{2^{42}}.$$

5. Используется десятичное разложение положительного иррационального числа

$$\omega = \beta_0, \ \beta_1\beta_2 \ \ldots \beta_s \ \ldots = \beta_0 + (\omega),$$

где $\beta_0$ — целая часть числа $\omega$ и $(\omega)$ — его дробная часть.

---

*) Запись $a \equiv b \pmod{k}$ ($a$, $b$, $k$ — целые числа) обозначает, что разность $a-b$ делится без остатка на $k$.

Для получения случайной последовательности $\{x_n\}$ полагают:

$$x_n = (n\omega) \qquad (n = 1,\ 2, \ldots).$$

Если требуется случайная последовательность, состоящая из $s$-разрядных чисел, то в числах $(n\omega)$ ограничиваются соответствующими разрядами.

Для решения некоторых задач нужно иметь несколько случайных последовательностей

$$\{x_n^{(1)}\},\ \{x_n^{(2)}\},\ \ldots,\ \{x_n^{(m)}\}.$$

В этом случае выбирают $m$ линейно независимых положительных иррациональностей $\omega_1, \omega_2, \ldots, \omega_m$ и полагают:

$$x_n^{(k)} = (n\omega_k) \qquad (k = 1,\ 2,\ \ldots,\ m;\ n = 1,\ 2,\ \ldots).$$

Можно также взять одну равномерно распределенную случайную последовательность $\{x_n\}$ и произвести из нее $m$ выборок:

$$\begin{array}{l} \{x_1,\ x_{m+1},\ x_{2m+1},\ \ldots\}, \\ \{x_2,\ x_{m+2},\ \ldots\ x_{2m+2},\ \ldots\}, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \{x_m,\ x_{2m},\ x_{3m},\ \ldots\}, \end{array}$$

беря числа не подряд, а через $m$ элементов. Очевидно, таким образом, мы будем иметь $m$ равномерно распределенных подпоследовательностей.

С помощью этих и других методов составлены таблицы случайных чисел. В таких таблицах обычно даются случайные десятичные цифры; из них легко можно конструировать случайные числа определенного разряда. Для примера приводим часть одной из таблиц (см.[7]) пятиразрядных случайных чисел (таблица 76).

Таблица 76

**Случайные числа, равномерно распределенные на отрезке [0, 1]**

| | | | | |
|---|---|---|---|---|
| 0,57705 | 0,35483 | 0,11578 | 0,65339 | 0,66674 |
| 0,71618 | 0,09393 | 0,93045 | 0,93382 | 0,99279 |
| 0,73710 | 0,30304 | 0,93011 | 0,05758 | 0,24202 |
| 0,70131 | 0,55186 | 0,42844 | 0,00336 | 0,94010 |
| 0,16961 | 0,64003 | 0,52906 | 0,88222 | 0,60981 |
| 0,53324 | 0,20514 | 0,09461 | 0,98585 | 0,13094 |
| 0,43166 | 0,00188 | 0,99602 | 0,52103 | 0,35193 |
| 0,26275 | 0,55709 | 0,69962 | 0,91827 | 0,64560 |
| 0,05926 | 0,86977 | 0,31311 | 0,07069 | 0,64559 |
| 0,66289 | 0,31303 | 0,27004 | 0,13928 | 0,68008 |

## § 4. Вычисление кратных интегралов методом Монте-Карло

Пусть функция

$$y=f(x_1,\ x_2,\ \ldots,\ x_m)$$

непрерывна в ограниченной замкнутой области $S$ и требуется вычислить $m$-кратный интеграл

$$I=\iint\limits_{(S)}\ldots\int f(x_1,\ x_2,\ \ldots,\ x_m)\,dx_1\,dx_2\ \ldots\ dx_m. \qquad (1)$$

Геометрически число $I$ представляет собой $(m+1)$-мерный объем прямого цилиндроида*) в пространстве $Ox_1x_2\ \ldots\ x_my$, построенного на основании $S$ и ограниченного сверху данной поверхностью $y=f(\boldsymbol{x})$, где $\boldsymbol{x}=(x_1,\ x_2,\ \ldots,\ x_m)$ (рис. 85).

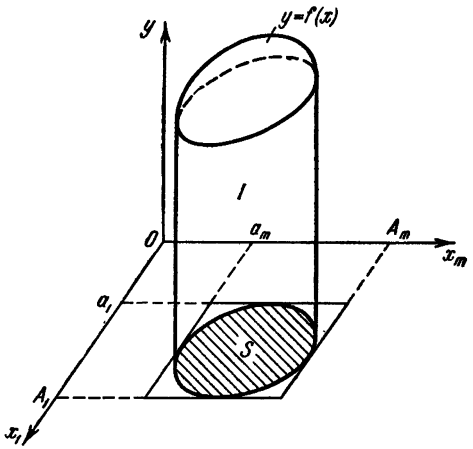


Рис. 85.

Преобразуем интеграл (1) так, чтобы новая область интегрирования целиком содержалась внутри единичного $m$-мерного куба. Пусть область $S$ расположена в $m$-мерном параллелепипеде

$$a_i\leqslant x_i\leqslant A_i \qquad (2)$$
$$(i=1,\ 2,\ \ldots,\ m).$$

Сделаем замену переменных

$$x_i=a_i+(A_i-a_i)\xi_i \qquad (3)$$
$$(i=1,\ 2,\ \ldots,\ m).$$

Тогда, очевидно, $m$-мерный параллелепипед (2) преобразуется в $m$-мерный единичный куб

$$0\leqslant\xi_i\leqslant 1 \qquad (i=1,\ 2,\ \ldots,\ m) \qquad (4)$$

и, следовательно, новая область интегрирования $\sigma$, которая находится по обычным правилам, будет целиком расположена внутри этого куба (рис. 86).

Вычисляя якобиан преобразования будем иметь:

$$\frac{D(x_1,\ x_2,\ \ldots,\ x_m)}{D(\xi_1,\ \xi_2,\ \ldots,\ \xi_m)}=\begin{vmatrix} A_1-a_1 & 0 & \ldots & 0\\ 0 & A_2-a_2 & \ldots & 0\\ \cdots & \cdots & \cdots & \cdots\\ 0 & 0 & \ldots & A_m-a_m\end{vmatrix}=$$
$$=(A_1-a_1)(A_2-a_2)\ \ldots\ (A_m-a_m).$$

*) Точнее, алгебраический объем, где предполагается, что части цилиндроида, расположенные выше гиперплоскости $Ox_1x_2,\ \ldots,\ x_m$, имеют положительную меру, а ниже — отрицательную.

Таким образом,

$$I = \iint\limits_{(\sigma)} \ldots \int F(\xi_1, \xi_2, \ldots, \xi_m)\, d\xi_1\, d\xi_2 \ldots d\xi_m, \tag{5}$$

где

$$F(\xi_1, \xi_2, \ldots, \xi_m) = (A_1 - a_1)(A_2 - a_2) \ldots (A_m - a_m) \times$$
$$\times f(a_1 + (A_1 - a_1)\xi_1,\ a_2 + (A_2 - a_2)\xi_2,\ \ldots,\ a_m + (A_m - a_m)\xi_m).$$

Введя обозначения

$$\xi = (\xi_1, \xi_2, \ldots, \xi_m)$$

и

$$d\sigma = d\xi_1 d\xi_2 \ldots d\xi_m,$$

запишем интеграл (5) короче в следующем виде:

$$I = \iint\limits_{(\sigma)} \ldots \int F(\xi)\, d\sigma. \tag{5'}$$

Мы укажем два способа вычисления интеграла (5′) методом случайных испытаний.

Первый способ. Выбираем $m$ равномерно распределенных на отрезке [0, 1] последовательностей случайных чисел:

$$\begin{array}{l} \xi_1^{(1)}, \xi_2^{(1)}, \ldots, \xi_n^{(1)}, \ldots; \\ \xi_1^{(2)}, \xi_2^{(2)}, \ldots, \xi_n^{(2)}, \ldots; \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \xi_1^{(m)}, \xi_2^{(m)}, \ldots, \xi_n^{(m)}, \ldots \end{array}$$

Точки $M_i(\xi_i^{(1)}, \xi_i^{(2)}, \ldots, \xi_i^{(m)})$ $(i = 1, 2, \ldots)$ можно рассматривать

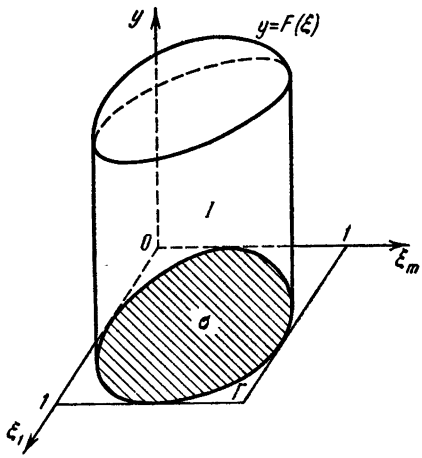


Рис. 86.

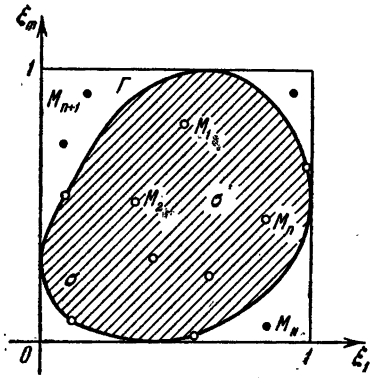


Рис. 87.

как случайные. Выбрав достаточно большое число $N$ точек $M_1, M_2, \ldots, M_N$, проверяем, какие из них принадлежат области $\sigma$

(*первая категория*) и какие не принадлежат ей (*вторая категория*). Пусть (рис. 87)

1) $$M_i \in \sigma \text{ при } i=1,\ 2,\ \ldots,\ n \tag{6}$$

и

2) $$M_i \overline{\in} \sigma \text{ при } i=n+1,\ n+2,\ \ldots,\ N \tag{6'}$$

(для удобства мы здесь изменяем нумерацию точек!). Заметим, что относительно границы Г области σ следует заранее договориться, причисляются ли граничные точки или часть их к области σ, или не причисляются к ней. В общем случае при гладкой границе Г это не имеет существенного значения; в отдельных случаях нужно решать вопрос с учетом конкретной обстановки.

Взяв достаточно большое число $n$ точек $M_i \in \sigma$, приближенно можно положить:

$$y_{\text{ср}} = \frac{1}{n} \sum_{i=1}^{n} F(M_i);$$

отсюда искомый интеграл выражается формулой

$$I = y_{\text{ср}}\sigma = \frac{\sigma}{n} \sum_{i=1}^{n} F(M_i), \tag{7}$$

где под σ понимается $m$-мерный объем области интегрирования σ. Если вычисление объема σ затруднительно, то можно принять:

$$\sigma \approx \frac{n}{N};$$

отсюда

$$I \approx \frac{1}{N} \sum_{i=1}^{n} F(M_i).$$

В частном случае, когда σ есть единичный куб ($\sigma = 1$), проверка становится излишней, т. е. $n = N$ и мы имеем просто

$$I = \frac{1}{N} \sum_{i=1}^{N} F(M_i).$$

Для проверки условий (6) и (6′) обычно исходят из аналитического задания границы Г области σ. В простейшем случае, если поверхность Г задана уравнением

$$\varphi(\xi) = 0, \tag{8}$$

где при $\varphi(\xi) < 0$ точка $\xi \in \sigma$ и при $\varphi(\xi) > 0$ точка $\xi \overline{\in} \sigma$, мы имеем: 1) если $\varphi(M_i) < 0$, то точка $M_i$ — первой категории и 2) если $\varphi(M_i) > 0$, то точка $M_i$ — второй категории. Точки $M_i$, для которых $\varphi(M_i) = 0$, причисляются к первой или второй категории по соглашению. Заметим, что уравнение (8) можно заменить любым равносильным ему уравнением, что иногда дает возможность значительно

облегчить выкладки. Так, например, неравенство для круга

$$x^2+y^2-x-y+\frac{1}{4}\leqslant 0$$

удобнее заменить эквивалентным

$$\left(x-\frac{1}{2}\right)^2+\left(y-\frac{1}{2}\right)^2\leqslant\frac{1}{4},$$

так как второе неравенство проверяется проще.

Если область $\sigma$ — стандартная и задана неравенствами:

$$\left.\begin{array}{l}\underline{\xi_1}\leqslant\xi_1\leqslant\overline{\xi_1},\\ \underline{\xi_2}(\xi_1)\leqslant\xi_2\leqslant\overline{\xi_2}(\xi_1),\\ \ldots\ldots\ldots\\ \underline{\xi_m}(\xi_1,\ \ldots,\ \xi_{m-1})\leqslant\xi_m\leqslant\overline{\xi_m}(\xi_1,\ \ldots,\ \xi_{m-1}),\end{array}\right\}\qquad(9)$$

то для определения принадлежности случайной точки $M$ ($\xi_1$, $\xi_2$, ..., $\xi_m$) к первой или второй категории проверяют выполнение этих неравенств.

Таблица 77

**Схема определения принадлежности случайной точки $M$ ($\xi_1$, ..., $\xi_m$) к стандартной области (9)**

| $\xi_1$ | $\underline{\xi_1}$ | $\overline{\xi_1}$ | $\varepsilon_1$ | $\xi_2$ | $\underline{\xi_2}$ | $\overline{\xi_2}$ | $\varepsilon_2$ |
|---|---|---|---|---|---|---|---|
| | | | | | | | |
| ... | $\xi_m$ | $\underline{\xi_m}$ | $\overline{\xi_m}$ | $\varepsilon_m$ | $\varepsilon$ | $y$ | |
| | | | | | | | |

Практически это удобно делать по схеме, приведенной в таблице 77. Здесь

$$\varepsilon_i=\begin{cases}1, & \text{если } \xi_i\in[\underline{\xi_i},\ \overline{\xi_i}]\\ 0, & \text{если } \xi_i\overline{\in}[\underline{\xi_i},\ \overline{\xi_i}]\end{cases},$$

($i=1,\ 2,\ \ldots,\ m$) и $\varepsilon=\varepsilon_1\varepsilon_2\ \ldots\varepsilon_m$. Очевидно,

если $\varepsilon=1$, то $M\in\sigma$;

если $\varepsilon=0$, то $M\overline{\in}\sigma$.

Заметим, что если $\varepsilon_j = 0$ $(j < m)$, то дальнейшие значения $\varepsilon_{j+1}, \ldots, \varepsilon_m$ можно не подсчитывать, так как они не повлияют на окончательный результат. Значение функции $y = F(M)$ подсчитывается только для тех точек $M$, для которых $\varepsilon = 1$. Затем для вычисления интеграла $I$ используется формула (7).

Пример. Методом Монте-Карло приближенно вычислить интеграл

$$I = \iint\limits_{(\sigma)} (x^2 + y^2)\, dx\, dy, \tag{10}$$

где область интегрирования $\sigma$ определяется следующими неравенствами:

$$\left.\begin{aligned} &\frac{1}{2} \leqslant x \leqslant 1, \\ &0 \leqslant y \leqslant 2x - 1 \end{aligned}\right\} \quad (\sigma)$$

(рис. 88).

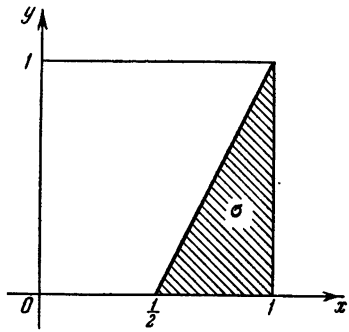

Рис. 88.

Решение. Интеграл (10) дан в приведенном виде, т. е. область интегрирования $\sigma$ расположена в единичном квадрате

$$0 \leqslant x \leqslant 1, \quad 0 \leqslant y \leqslant 1.$$

Для решения задачи воспользуемся таблицей случайных чисел (таблица 76), рассматривая каждую очередную пару чисел таблицы как соответствующие координаты $x$ и $y$ случайной точки $M(x, y)$. Так как вычисление носит иллюстративный характер, то ограничимся $N = 20$ случайными точками, причем координаты их для простоты округлим до трех десятичных знаков. Результаты вычислений приведены в таблице 78, где положено

$$\underline{x} = \frac{1}{2}, \qquad \overline{x} = 1;$$

$$\underline{y}(x) = 0, \qquad \overline{y}(x) = 2x - 1;$$

$$z = x^2 + y^2.$$

Отсюда

$$z_{\text{ср}} = \frac{1}{4} \cdot 3{,}837 = 0{,}96$$

и, следовательно, по формуле (7), учитывая, что $\sigma = \frac{1}{4}$, имеем:

$$I = z_{\text{ср}} \cdot \sigma = 0{,}96 \cdot \frac{1}{4} = 0{,}24. \tag{11}$$

Если приближенно принять

$$\sigma \approx \frac{n}{N} = \frac{4}{20} = \frac{1}{5},$$

то получим:

$$I \approx 0{,}96 \cdot \frac{1}{5} = 0{,}19.$$

Таблица 78

**Вычисление двойного интеграла (10) методом Монте-Карло**

| $x$ | $\underline{x}$ | $\overline{x}$ | $\varepsilon_1$ | $y$ | $\underline{y}(x)$ | $\overline{y}(x)$ | $\varepsilon_2$ | $\varepsilon$ | $z$ |
|---|---|---|---|---|---|---|---|---|---|
| 0,577 | 0,500 | 1,000 | 1 | 0,716 | 0 | 0,154 | 0 | 0 | |
| 0,737 | 0,500 | 1,000 | 1 | 0,701 | 0 | 0,474 | 0 | 0 | |
| 0,170 | 0,500 | 1,000 | 0 | 0,533 | | | | 0 | |
| 0,432 | 0,500 | 1,000 | 0 | 0,263 | | | | 0 | |
| 0,059 | 0,500 | 1,000 | 0 | 0,663 | | | | 0 | |
| 0,355 | 0,500 | 1,000 | 0 | 0,094 | | | | 0 | |
| 0,303 | 0,500 | 1,000 | 0 | 0,552 | | | | 0 | |
| 0,640 | 0,500 | 1,000 | 1 | 0,205 | 0 | 0,280 | 1 | 1 | 0,452 |
| 0,002 | 0,500 | 1,000 | 0 | 0,557 | | | | 0 | |
| 0,870 | 0,500 | 1,000 | 1 | 0,323 | 0 | 0,740 | 1 | 1 | 0,855 |
| 0,116 | 0,500 | 1,000 | 0 | 0,930 | | | | 0 | |
| 0,930 | 0,500 | 1,000 | 1 | 0,428 | 0 | 0,860 | 1 | 1 | 1,048 |
| 0,529 | 0,500 | 1,000 | 1 | 0,095 | 0 | 0,058 | 0 | 0 | |
| 0,996 | 0,500 | 1,000 | 1 | 0,700 | 0 | 0,992 | 1 | 1 | 1,482 |
| 0,313 | 0,500 | 1,000 | 0 | 0,270 | | | | 0 | |
| 0,653 | 0,500 | 1,000 | 1 | 0,934 | 0 | 0,306 | 0 | 0 | |
| 0,058 | 0,500 | 1,000 | 0 | 0,003 | | | | 0 | |
| 0,882 | 0,500 | 1,000 | 1 | 0,986 | 0 | 0,764 | 0 | 0 | |
| 0,521 | 0,500 | 1,000 | 1 | 0,918 | 0 | 0,042 | 0 | 0 | |
| 0,071 | 0,500 | 1,000 | 0 | 0,239 | | | | 0 | |
| $\Sigma$ | | | | | | | | 4 | 3,837 |

Заметим, что точное значение интеграла

$$I=\frac{7}{32}\approx 0{,}22;$$

поэтому относительная погрешность результата (11) равна

$$\delta=\frac{0{,}24-0{,}22}{0{,}22}\approx 9\%.$$

Конечно, тут число точек $N=20$ недостаточно, чтобы статистические закономерности могли проявиться в должной мере, но тем не менее для грубой ориентировки получен удовлетворительный результат.

Второй способ. Если функция $F(\xi)=F(\xi_1, \xi_2, \ldots, \xi_m)$ неотрицательна, то интеграл (5) можно рассматривать как объем тела $V$ в $(m+1)$-мерном пространстве $O\xi_1\xi_2\ldots\xi_m y$, т. е.

$$I=\iint\limits_{(V)}\ldots\int d\xi_1\, d\xi_2\ldots d\xi_m\, dy, \tag{12}$$

где область интегрирования $V$ определяется условиями

$$\xi=(\xi_1, \xi_2, \ldots, \xi_m)\in\sigma, \quad 0\leqslant y\leqslant F(\xi).$$

Пусть

$$0 \leqslant F(\xi) \leqslant B. \tag{13}$$

Введя в интеграле (12) новую переменную

$$\eta = \frac{1}{B} y, \tag{14}$$

получим:

$$I = B \iint \ldots \int\limits_{(v)} d\xi_1\, d\xi_2 \ldots d\xi_m\, d\eta,$$

где новая область $v$ есть цилиндроид пространства $O\xi_1\xi_2\ldots\xi_m\eta$, построенный на области $\sigma$ и ограниченный снизу гиперплоскостью $\eta = 0$ и сверху гиперповерхностью

$$\eta = \frac{1}{B} F(\xi)$$

(рис. 89). В силу неравенства (13) объем $v$ целиком лежит в $(m+1)$-мерном кубе

$$0 \leqslant \xi_i \leqslant 1$$
$$(i = 1, 2, \ldots, m),$$
$$0 \leqslant \eta < 1.$$

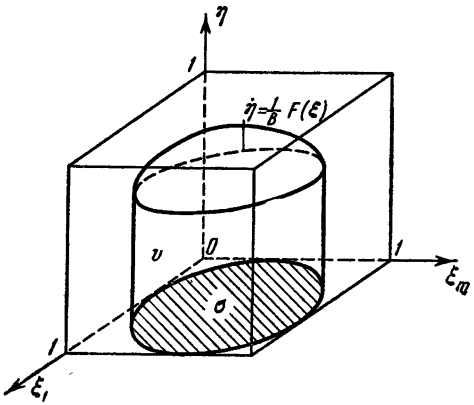


Рис. 89.

Возмьем теперь $m+1$ равномерно распределенных на [0, 1] случайных последовательностей

$$\{\xi_i^{(1)}\}, \{\xi_i^{(2)}\}, \ldots, \{\xi_i^{(m)}\}, \{\eta_i\},$$

соответствующие элементы которых будем рассматривать как координаты случайных точек

$$M_i\{\xi_i^{(1)}, \xi_i^{(2)}, \ldots, \xi_i^{(m)}, \eta_i\}, \qquad (i = 1, 2, \ldots)$$

пространства $O\xi_1\xi_2\ldots\xi_m\eta$. Если из общего числа $N$ случайных точек $n$ принадлежит объему $v$, а $N-n$ не принадлежит этому объему, то при достаточно большом числе $N$ приближенно полагают:

$$I \approx B \cdot \frac{n}{N}, \tag{15}$$

т. е.

$$I = B \cdot \boldsymbol{P}(M \in v),$$

где точка $M$ с одинаковой вероятностью может занимать положения $M_1, M_2, \ldots, M_N$. Выполнение соотношения

$$M \in v$$

проверяется аналогично тому, как было указано в первом способе. Заметим, что если $\sigma$ есть единичный куб $0 \leqslant \xi_i \leqslant 1$ $(i = 1, 2, \ldots, m)$, то для точки $M_i$ $(\xi_i^{(1)}, \ldots, \xi_i^{(m)}, \eta_i)$, все координаты которой предполагаются принадлежащими единичному отрезку [0, 1,] достаточно проверять лишь выполнение соотношения

$$\eta_i \leqslant \frac{1}{B} F(\xi_1^{(1)}, \xi_2^{(2)}, \ldots, \xi_m^{(m)}).$$

Рассмотрим теперь общий случай, когда функция

$$F(\xi) = F(\xi_1, \xi_2, \ldots, \xi_m)$$

— знакопеременная. Пусть

$$-b \leqslant F(\xi) \leqslant B, \tag{16}$$

где $b$ и $B$ — неотрицательные числа. Положим

$$F(\xi) = -b + (B+b)\tilde{F}(\xi),$$

тогда будем иметь:

$$\iint\limits_{(\sigma)} \ldots \int F(\xi)\, d\sigma = -b\sigma + (B+b) \iint\limits_{(\sigma)} \ldots \int \tilde{F}(\xi)\, d\sigma,$$

где функция $\eta = \tilde{F}(\xi)$ в силу неравенства (16) удовлетворяет неравенствам

$$0 \leqslant \tilde{F}(\xi) \leqslant 1.$$

Интеграл

$$\iint\limits_{(\sigma)} \ldots \int \tilde{F}(\xi)\, d\sigma = \iint\limits_{(\tilde{v})} \ldots \int d\sigma\, d\eta$$

может быть вычислен способом, указанным выше.

Для оценки приближенного равенства *)

$$I_0 = \iint\limits_{(v)} \ldots \int d\sigma\, d\eta = P(M \in v) \approx \frac{n}{N} \tag{17}$$

предположим сначала, что мы имеем дело с идеальными случайными равномерно распределенными последовательностями точек $M_i$ $(i = 1, 2, \ldots)$, координаты которых принадлежат единичному отрезку [0, 1].

На основании теоремы Бернулли, применяя неравенство Чебышева, будем иметь:

$$P\left(\left|\frac{n}{N} - I_0\right| < \varepsilon\right) \geqslant 1 - \frac{I_0(1-I_0)}{\varepsilon^2 N} \geqslant 1 - \frac{1}{4\varepsilon^2 N}. \tag{18}$$

---

*) Множитель $B$ не имеет существенного значения.

Задавшись для данного $\varepsilon$ гарантийной вероятностью

$$P\left(\left|\frac{n}{N}-I_0\right|<\varepsilon\right)\geqslant 1-\delta, \tag{19}$$

из неравенства (18) получаем, что условие (19) заведомо имеет место, если

$$\frac{1}{4\varepsilon^2 N}=\delta. \tag{20}$$

Отсюда выводим:

$$\varepsilon=\frac{1}{2\sqrt{\delta N}}. \tag{21}$$

Таким образом, точность оценки

$$I_0\approx\frac{n}{N}$$

при заданной предельной вероятности ее обратно пропорциональна корню квадратному из числа испытаний, т. е. $\varepsilon=O\left(\frac{1}{\sqrt{N}}\right)$. Это обстоятельство обусловливает сравнительно медленную сходимость метода Монте-Карло: например, чтобы уменьшить погрешность результата в 10 раз, число испытаний нужно увеличить в 100 раз! Если точность оценки $\varepsilon$ и гарантийная вероятность $1-\delta$ заданы, то из формулы (20) выводим необходимое число испытаний

$$N=\frac{1}{4\varepsilon^2\delta}. \tag{22}$$

Например, при $\varepsilon=0{,}001$ и $\delta=0{,}01$ имеем:

$$N=25\,000\,000.$$

Оценка (22) является завышенной и может быть значительно улучшена!

Отметим одно важное обстоятельство: число испытаний $N$ не зависит от размерности интеграла $I_0$ и поэтому метод Монте-Карло выгодно применять для вычисления кратных интегралов высоких размерностей, где применение обычных кубатурных формул встречает значительные затруднения. Например, для приближенного вычисления обычным путем 10-кратного интеграла, распространенного на единичный объем, при выборе шага $h=0{,}1$ понадобится сумма, содержащая примерно $10^{10}$ слагаемых!

При практическом применении метода Монте-Карло для вычисления кратных интегралов обычно пользуются $s$-разрядными равномерно распределенными случайными последовательностями. В этом случае дробь $\frac{n}{N}$, если $N$ велико, будет близка не к истинному объему $I_0$, а к некоторому фиктивному объему $I'_0$, приближенно

представляющему собой относительную меру числа точек $M$ с координатами вида

$$\xi_i = \frac{k_i}{10^s}, \qquad \eta = \frac{k}{10^s} \tag{23}$$

$$(i = 1, 2, \ldots, m; \quad k_i, \ k = 0, 1, 2, \ldots, 10^s),$$

попавших в объем $v$ (ср. § 3), причем $I_0'$, строго говоря, зависит от того, причисляются ли граничные точки к объему $v$ или нет. Полная погрешность результата оценивается следующим образом (см. [2]):

$$\left|\frac{n}{N} - I_0\right| \leqslant \left|I_1' - I_0\right| + \left|I_0' - \frac{n}{N}\right|. \tag{24}$$

Первое слагаемое $|I'_0 - I_0|$ правой части неравенства (24) представляет собой *обычную вычислительную погрешность,* получающуюся при замене интеграла $I_0$ интегральной суммой, соответствующей разбиению объема $v$ на элементарные кубические ячейки, вершины которых принадлежат сетке (23). Величину этой погрешности можно оценить с помощью неравенства

$$|I'_0 - I_0| \leqslant \bar{v} - \underline{v}, \tag{25}$$

где $\bar{v}$ — верхняя интегральная сумма (в нашем случае для интеграла (17) просто объем описанного ступенчатого тела) и $\underline{v}$ — нижняя интегральная сумма (т. е. объем вписанного ступенчатого тела). Величина погрешности $|I'_0 - I_0|$ существенно зависит от разрядности $s$ случайных чисел, и если граница тела $v$ кусочно гладкая, то эта погрешность при достаточно большом $s$ может быть сделана сколь угодно малой. Неудобство от увеличения разрядности состоит в том, что возрастает объем работы, так как вычисления приходится производить с дополнительными знаками. Второе слагаемое $\left|I'_0 - \frac{n}{N}\right|$ правой части неравенства (24) называется *погрешностью выборки* и может быть оценено вероятностным путем с помощью теоремы Бернулли, как указано выше.

## § 5*. Решение систем линейных алгебраических уравнений методом Монте-Карло

Рассмотрим линейную систему

$$\sum_{j=1}^{n} a_{ij} x_j = b_i \qquad (i = 1, \ldots, n). \tag{1}$$

Некоторым способом приведем систему (1) к специальному виду

$$x_i = \sum_{j=1}^{n} \alpha_{ij} x_j + \beta_i \qquad (i = 1, \ldots, n). \tag{2}$$

Введя матрицу $\alpha = [\alpha_{ij}]$ и векторы

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix},$$

систему (2) можно записать в матрично-векторной форме

$$x = \alpha x + \beta. \tag{2'}$$

Будем предпологать, что все собственные значения матрицы $\alpha$ по модулю меньше единицы. В частности, достаточно считать, что какая-нибудь каноническая норма матрицы $\alpha$ подчинена неравенству

$$\|\alpha\| < 1. \tag{3}$$

В этом случае система (2′) имеет единственное решение, которое может быть найдено методом итерации (гл. VIII, § 10).

Подберем систему множителей $v_{ij}$ таким образом, чтобы числа $p_{ij}$, определяемые уравнениями

$$\alpha_{ij} = p_{ij} v_{ij} \qquad (i,\ j = 1, \ldots, n), \tag{4}$$

удовлетворяли следующим условиям:

1) $p_{ij} \geqslant 0$, причем $p_{ij} > 0$ при $\alpha_{ij} \neq 0$;

2) $\sum\limits_{j=1}^{n} p_{ij} < 1 \quad (i = 1, \ldots, n)$.

Пусть

$$p_{i,\,n+1} = 1 - \sum_{j=1}^{n} p_{ij} \qquad (i = 1, \ldots, n).$$

Кроме того, условно положим

$$p_{n+1,\,j} = 0 \text{ при } j < n+1$$

и

$$p_{n+1,\,n+1} = 1.$$

Рассмотрим теперь некоторую блуждающую частицу, обладающую конечным числом возможных и несовместимых состояний

$$S_1,\ S_2, \ldots,\ S_n,\ S_{n+1}.$$

Эта частица такова, что с вероятностью $p_{ij}$ $(i,\ j = 1, \ldots, n+1)$ переходит из состояния $S_i$ в состояние $S_j$, независимо от прошлых состояний и с неопределенностью будущих. Состояние $S_{n+1} = \Gamma$ («граница» или «поглощающий экран») является особым и соответствует полной остановке частицы, так как в силу условия $p_{n+1,\,j} = 0$ $(j = 1, \ldots, n)$ переходы из состояния $S_{n+1}$ в состояние $S_j$ при

$j < n+1$ невозможны. Таким образом, процесс блуждания прекращается, как только частица первый раз попадает на границу $\Gamma$. Описанная смена состояний обычно называется *дискретной цепью Маркова**) с конечным числом состояний [2]. Числа $p_{ij}$ называются *переходными вероятностями*, а матрица

$$\Pi = \begin{bmatrix} p_{11} & \dots & p_{1n} & p_{1,\,n+1} \\ \cdot & \cdot & \cdot & \cdot \\ p_{n1} & \dots & p_{nn} & p_{n,\,n+1} \\ 0 & \dots & 0 & 1 \end{bmatrix}$$

представляет собой *матрицу перехода* состояний $\{S_i\}$ (*закон цепи*).

Пусть $S_i$ — некоторое фиксированное состояние, отличное от граничного ($i < n+1$). Рассмотрим случайное блуждание частицы, начинающееся в данном состоянии $S_i = S_{i_0}$ и после ряда промежуточных состояний $S_{i_1}, S_{i_2}, \dots, S_{i_m}$ заканчивающееся на границе $S_{i_{m+1}} = \Gamma$. Таким образом, $S_{i_m}$ $(m \geqslant 0)$ есть состояние частицы, непосредственно предшествующее выходу ее на границу. Совокупность состояний

$$T_i = \{S_{i_0}, S_{i_1}, \dots, S_{i_m}, S_{i_{m+1}}\} \tag{5}$$

для краткости будем называть *траекторией*. Пусть $X_i$ — случайная величина, зависящая от случайных траекторий $T_i$, начинающихся в состоянии $S_i$ (*функционал траектории* $T_i$), и принимающая для траектории (5) значение

$$\xi(T_i) = \beta_{i_0} + v_{i_0 i_1}\beta_{i_1} + v_{i_0 i_1} v_{i_1 i_2}\beta_{i_2} + \dots + v_{i_0 i_1} \dots v_{i_{m-1} i_m}\beta_{i_m}, \tag{6}$$

где $\beta_j$ $(j = i_0, i_1, \dots, i_m)$ — соответствующие свободные члены приведенной системы (2).

В частности, если $v_{ij} = 1$, то имеем просто

$$\xi(T_i) = \beta_{i_0} + \beta_{i_1} + \dots + \beta_{i_m}. \tag{6'}$$

По теореме умножения вероятностей траектория $T_i$, а следовательно, и значение $\xi(T_i)$, реализуется с вероятностью

$$P(T_i) = p_{i_0 i_1} p_{i_1 i_2} \dots p_{i_m i_{m+1}}, \tag{7}$$

где $i_0 = i$ и $i_{m+1} = n+1$.

Теорема. *Математические ожидания*

$$MX_i = x_i \qquad (i = 1, 2, \dots, n)$$

*являются корнями системы* (2).

---

*) Точнее, *простой однородной* [2].

Доказательство. Траектории $T_i$, начинающиеся из состояния $S_i$, в зависимости от первого шага можно разбить на $n+1$ категорий

$$\begin{aligned} T_{i_1} &= \{S_i,\ S_1,\ S_{i_2},\ \ldots\}; \\ T_{i_2} &= \{S_i,\ S_2,\ S_{i_2},\ \ldots\}; \\ &\ldots\ldots\ldots\ldots \\ T_{i_n} &= \{S_i,\ S_n,\ S_{i_2},\ \ldots\}; \\ T_{i,n+1} &= \{S_i,\ S_{n+1}\}, \end{aligned}$$

т. е. частица, начав блуждание из состояния $S_i$, при первом шаге может переходить или в состояние $S_1$, или в состояние $S_2$ и т. д., а затем через некоторое число шагов закончить блуждание на границе.

Если частица имеет траекторию

$$T_{ij} = \{S_i,\ S_j,\ S_{i_2},\ \ldots,\ S_{i_m},\ S_{i_{m+1}} = \Gamma\},$$

где $j \neq n+1$, то случайная величина $X_i$ в силу формулы (6) примет значение

$$\xi(T_{ij}) = \beta_i + v_{ij}\beta_j + v_{ij}v_{ji_2}\beta_{i_2} + \ldots + v_{ij}v_{ji_2}\ldots v_{i_{m-1}i_m}\beta_{i_m} =$$
$$= \beta_i + v_{ij}(\beta_j + v_{ji_2}\beta_{i_2} + \ldots + v_{ji_2}\ldots v_{i_{m-1}i_m}\beta_{i_m}) = \beta_i + v_{ij}\xi(T_j), \qquad (8)$$

где $T_j$ — некоторая траектория с начальным состоянием $S_j$.

В том случае, когда частица сразу попадает на границу $\Gamma$, т. е. траектория имеет вид $T_{i,\,n+1} = \{S_i,\ S_{n+1}\}$, то

$$\xi(T_{i,\,n+1}) = \beta_i. \qquad (8')$$

Вероятность того, что траектория $T_i$ есть траектория типа $T_{ij}$, очевидно, равна $p_{ij}$.

В силу определения математического ожидания имеем:

$$MX_i = \sum_{T_i} \xi(T_i)P(T_i) = \sum_j \sum_{T_{ij}} \xi(T_{ij})P(T_{ij}).$$

Если $j < n+1$, то траектория $T_{ij}$ состоит из отрезка $(S_i,\ S_j)$ и некоторой траектории $T_j$. Поэтому $P(T_{ij}) = p_{ij}P(T_j)$. При $j = n+1$ имеем:

$$\xi(T_{i,\,n+1}) = \beta_i \text{ и } P(T_{i,\,n+1}) = p_{i,\,n+1}.$$

Кроме того, так как каждой траектории $T_{ij}$ при $j < n+1$ однозначно соответствует траектория $T_j$ и обратно, то суммирование по траекториям $T_{ij}$ для $j = 1,\ 2,\ \ldots,\ n$ можно заменить суммированием по траекториям $T_j$.

Отсюда, учитывая формулу (8), получаем:

$$MX_i = \sum_{j=1}^{n} \sum_{T_j} [\beta_i + v_{ij}\xi(T_j)] \cdot p_{ij}P(T_j) + \beta_i p_{i,\,n+1}$$

или

$$MX_i = \sum_{j=1}^{n} p_{ij} v_{ij} \sum_{T_j} \xi(T_j) P(T_j) + \beta_i \left[ \sum_{j=1}^{n} p_{ij} \sum_{T_j} P(T_j) + p_{i,\, n+1} \right].$$

Но, очевидно,

$$\sum_{T_j} \xi(T_j) P(T_j) = MX_i \qquad (j = 1,\ 2,\ \ldots,\ n).$$

Кроме того, $\sum_{T_j} P(T_j) = 1$ и

$$\sum_{j=1}^{n} p_{ij} \sum_{T_j} P(T_j) + p_{i,\, n+1} = \sum_{j=1}^{n+1} p_{ij} = 1.$$

Следовательно,

$$MX_i = \sum_{j=1}^{n} \alpha_{ij} MX_j + \beta_i \qquad (i = 1,\ \ldots,\ n),$$

где $\alpha_{ij} = p_{ij} v_{ij}$.

Теорема доказана.

Замечание. При доказательстве теоремы предполагалось, что математические ожидания

$$x_i = MX_i \quad (i = 1,\ \ldots,\ n)$$

существуют. Можно доказать, что при выполнении условия (3) случайные величины $X_i$ обладают конечными математическими ожиданиями.

Из доказанной теоремы следует, что корни системы (2) можно рассматривать как математические ожидания случайных величин $X_1,\ \ldots,\ X_n$. Для экспериментального определения величины $x_i = MX_i$ организуют $N$ случайных блужданий, со случайными траекториями $T_i^{(k)}$ $(k = 1,\ \ldots,\ N)$ с начальным состоянием $S_i$ и каждый раз регистрируют значение $\xi(T_i^{(k)})$ случайной величины $X_i$. Предположим, что испытания независимы между собой и величина $X_i$ имеет ограниченную дисперсию. Тогда в силу теоремы Чебышева [1], [2] при $N$ достаточно большом с вероятностью, сколь угодно близкой к 1, будет справедливо неравенство

$$\left| x_i - \frac{1}{N} \sum_{k=1}^{N} \xi(T_i^{(k)}) \right| < \varepsilon,$$

где $\varepsilon$ — заданная предельная погрешность. Таким образом, корни системы (2) приближенно могут быть определены по формулам

$$x_i \approx \frac{1}{N} \sum_{k=1}^{N} \xi(T_i^{(k)}). \tag{9}$$

В частности, этим способом можно обращать матрицы вида

$$A = E - \alpha, \tag{10}$$

где $\|\alpha\| < 1$ и $E = [\delta_{ij}]$ — единичная матрица. Для этого заметим, что элементы обратной матрицы

$$A^{-1} = [x_{ij}]$$

являются корнями линейной системы

$$\sum_{k=1}^{n} (\delta_{ik} - \alpha_{ik}) x_{kj} = \delta_{ij} \qquad (i,\ j = 1,\ \ldots,\ n).$$

Отсюда получаем, что элементы каждого столбца

$$x_{1j},\ \ldots,\ x_{nj} \qquad (j = 1,\ \ldots,\ n)$$

матрицы $A^{-1}$ определяются из линейной подсистемы

$$x_{ij} = \sum_{k=1}^{n} \alpha_{ik} x_{kj} + \delta_{ij} \qquad (i = 1,\ \ldots,\ n). \tag{11}$$

На основании предыдущего, отправляясь из состояния $S_i = S_{i_0}$, при фиксированном $j$ получаем случайную величину $X_{ij}$ со значениями

$$\xi_j(T_i) = \delta_{i_0 j} + \delta_{i_1 j} v_{i_0 i_1} + \ldots + \delta_{i_m j} v_{i_0 i_1} \ldots v_{i_{m-1} i_m},$$

где $T_i = \{S_{i_0},\ S_{i_1},\ \ldots,\ S_{i_m}, S_{i_{m+1}} = \Gamma\}$ и числа $v_{ij}$ таковы, что $p_{ij}$, определяемые из уравнений $\alpha_{ij} = p_{ij} v_{ij}$, представляют собой вероятности перехода из состояния $S_i$ в состояние $S_j$. Математические ожидания $MX_{ij} = x_{ij}$ дадут искомые элементы матрицы $A^{-1}$.

Покажем теперь, как практически возможно организовать случайное блуждание частицы с заданными вероятностями перехода $p_{ij}$. Для простоты продположим, что $p_{ij}$ суть десятичные дроби с общим знаменателем $10^s$ ($s$ — натуральное число):

$$p_{i1} = \frac{t_{i1}}{10^s}, \qquad p_{i2} = \frac{t_{i2}}{10^s}, \qquad \ldots, \qquad p_{i,n+1} = \frac{t_{i,n+1}}{10^s},$$

где $t_{i1},\ t_{i2},\ \ldots,\ t_{i,n+1}$ — целые неотрицательные числа, причем

$$t_{i1} + t_{i2} + \ldots + t_{i,n+1} = 10^s \qquad (i = 1,\ 2,\ \ldots,\ n).$$

Рассмотрим частицу, имеющую начальное состояние $S_i$. Пусть $\{x\}$ есть $s$-разрядные числа, меньшие единицы, равномерно распределенные на отрезке [0, 1], например, элементы из соответствующей таблицы случайных чисел. Произведем розыгрыш случайного числа $x$. Если окажется, что выполнено неравенство

$$0 \leqslant x < \frac{t_{i1}}{10^s},$$

то будем считать, что частица переходит из состояния $S_i$ в состояние $S_1$. Далее, если

$$\frac{t_{i1}}{10^s} \leqslant x < \frac{t_{i1}+t_{i2}}{10^s},$$

то полагаем, что частица переходит из состояния $S_i$ в состояние $S_2$. Аналогично определяются остальные переходы. В частности, частица попадает на границу $S_{n+1}=\Gamma$, если случайное число $x$ таково, что

$$\frac{t_{i1}+\ldots+t_{in}}{10^s} \leqslant x < \frac{t_{i1}+\ldots+t_{in}+t_{i,\,n+1}}{10^s} = 1.$$

На основании данного соглашения ясно, что количества благоприятных случаев для переходов $S_i \longrightarrow S_j$ $(j=1,\ 2,\ \ldots,\ n+1)$ пропорциональны соответственно числам

$$t_{i1},\ t_{i2},\ \ldots,\ t_{i,\,n+1},$$

причем эти случаи равновероятны. Поэтому вероятности перехода

$$P(S_i \longrightarrow S_j) = \frac{t_{ij}}{10^s} = p_{ij} \qquad (i=1,\ \ldots,\ n;\ j=1,\ \ldots,\ n+1).$$

Выбирая последовательность случайных чисел и руководствуясь указанным выше правилом, получаем случайное блуждание частицы с фиксированным начальным состоянием и данными вероятностями перехода. Для достижения нужной точности корней (в вероятностном смысле) следует рассмотреть достаточно большое количество независимых блужданий.

Пример. Методом Монте-Карло решить систему уравнений

$$\left.\begin{aligned} x_1 &= 0{,}1x_1 + 0{,}2x_2 + 0{,}7; \\ x_2 &= 0{,}2x_1 - 0{,}3x_2 + 1{,}1. \end{aligned}\right\} \qquad (12)$$

Решение. Можно положить

$$v_{11}=1, \qquad v_{12}=1,$$
$$v_{21}=1, \qquad v_{22}=-1.$$

Отсюда матрица перехода есть

$$\Pi = \left[\begin{array}{cc|c} 0{,}1 & 0{,}2 & 0{,}7 \\ 0{,}2 & 0{,}3 & 0{,}5 \\ \hline 0 & 0 & 1 \end{array}\right],$$

где элементы первой строки представляют собой соответственно вероятности перехода из состояния $S_1$ в состояния $S_1$, $S_2$ и $S_3=\Gamma$, а элементы второй строки — из состояния $S_2$ в состояния $S_1$, $S_2$ и $S_3$, причем «кайма» соответствует границе $\Gamma$.

Так как элементы матрицы $\Pi$ кратны 0,1, то можно использовать одноразрядные случайные числа, цифры которых рекрутируются из

Таблица 79

**Нахождение неизвестного $x_1$ системы (12) методом Монте-Карло**

| № п/п | Случайное число $x$ | Траектория блуждания | Значение случайной величины $X_1$ |
|---|---|---|---|
| 1 | 0,5 | $S_1 \to \Gamma$ | 0,7 |
| 2 | 0,7 | $S_1 \to \Gamma$ | 0,7 |
| 3 | 0,7 | $S_1 \to \Gamma$ | 0,7 |
| 4 | 0,0<br>0,5 | $S_1 \to S_1 \to \Gamma$ | 0,7+0,7 |
| 5 | 0,7 | $S_1 \to \Gamma$ | 0,7 |
| 6 | 0,1<br>0,6 | $S_1 \to S_2 \to \Gamma$ | 0,7+1,1 |
| 7 | 0,1<br>0,8 | $S_1 \to S_2 \to \Gamma$ | 0,7+1,1 |
| 8 | 0,7 | $S_1 \to \Gamma$ | 0,7 |
| 9 | 0,3 | $S_1 \to \Gamma$ | 0,7 |
| 10 | 0,7 | $S_1 \to \Gamma$ | 0,7 |
| 11 | 0,1<br>0,0<br>0,7 | $S_1 \to S_2 \to$<br>$\to S_1 \to \Gamma$ | 0,7+1,1+0,7 |
| 12 | 0,0<br>0,1<br>0,3<br>0,1<br>0,1<br>0,6 | $S_1 \to S_1 \to S_2 \to$<br>$\to S_2 \to S_1 \to$<br>$\to S_2 \to \Gamma$ | 0,7+0,7+1,1 −<br>− 1,1 − 0,7 − 1,1 |
| 13 | 0,9 | $S_1 \to \Gamma$ | 0,7 |
| 14 | 0,6 | $S_1 \to \Gamma$ | 0,7 |
| 15 | 0,1<br>0,5 | $S_1 \to S_2 \to \Gamma$ | 0,7+1,1 |
| 16 | 0,3 | $S_1 \to \Gamma$ | 0,7 |
| 17 | 0,3 | $S_1 \to \Gamma$ | 0,7 |
| 18 | 0,2<br>0,4<br>0,4<br>0,3<br>0,1<br>0,6 | $S_1 \to S_2 \to S_2 \to$<br>$\to S_2 \to S_2 \to$<br>$\to S_1 \to \Gamma$ | 0,7+1,1 − 1,1 +<br>+ 1,1 − 1,1 − 0,7 |
| 19 | 0,6 | $S_1 \to \Gamma$ | 0,7 |
| 20 | 0,2<br>0,6 | $S_1 \to S_2 \to \Gamma$ | 0,7+1,1 |
| | | $\Sigma$ | $21 \cdot 0{,}7 + 4 \cdot 1{,}1$ |

какой-нибудь случайной последовательности, например являются элементами случайных чисел из таблицы 76 (§ 3).

Полученные результаты для 20 случайных блужданий с начальным состоянием $S_1$ приведены в таблице 79. Случайное число $x$ обеспечивало переходы состояний согласно следующей инструкции:

I. Для начального состояния $S_1$:

1) если $0 \leqslant x < 0{,}1$, то $S_1 \to S_1$;
2) » $0{,}1 \leqslant x < 0{,}3$, » $S_1 \to S_2$;
3) » $0{,}3 \leqslant x < 1$, » $S_1 \to \Gamma$;

II. Для начального состояния $S_2$:

1) если $0 \leqslant x < 0{,}2$, то $S_2 \to S_1$;
2) » $0{,}2 \leqslant x < 0{,}5$, » $S_2 \to S_2$;
3) » $0{,}5 \leqslant x < 1$, » $S_2 \to \Gamma$.

В последнем столбце таблицы 79 помещены значения случайной величины $X_1$, вычисленные по формуле (6). Отсюда

$$x_1 = MX_1 \approx \frac{1}{20}(20 \cdot 0{,}7 + 0{,}7 + 4 \cdot 1{,}1) = 0{,}96.$$

Аналогично вычисляется неизвестное $x_2$.

Заметим, что точные корни системы (12) суть $x_1 = 1$ и $x_2 = 1$.

Для решения алгебраических линейных уравнений по методу Монте-Карло используются также другие способы [11].

### Литература к семнадцатой главе

1. Е. С. Вентцель, Теория вероятностей, Физматгиз, М., 1958, гл. I—VI
2. Б. В. Гнеденко, Курс теории вероятностей, Гостехиздат. М.—Л., 1950, гл. I—VI.
3. А. С. Хаусхолдер, Основы численного анализа, ИЛ, М., 1956, гл. VIII.
4. В. Э. Милн. Численное решение дифференциальных уравнений. Приложение В, ИЛ, М., 1955.
5. Ю. А. Шрейдер, Метод статистических испытаний (Монте-Карло), Приборостроение, №7 (1956).
6. Современная математика для инженеров. Под ред. Э. Ф. Беккенбаха, ИЛ, М., 1958. Дж. В. Браун, Методы Монте-Карло.
7. Ф. М. Морз и Дж. Е. Кимбелл, Методы исследования операций, Сов. радио, М., 1956, гл. VI, § 4.
8. М. Кадыров, Таблицы случайных чисел, Изд. Среднеазиатского гос. ун-та, Ташкент, 1936.
9. А. И. Китов и Н. А. Криницкий, Электронные цифровые машины и программирование, Физматгиз, М., 1959, гл. VIII.
10. Дейвис, Рабинович, Опыты по вычислению кратных интегралов методом Монте-Карло. Реферативный журнал (математика) № 2 (1957), 1835.
11. Ю. А. Шрейдер, Решение систем линейных алгебраических уравнений по методу Монте-Карло. Вопросы теории математических машин, сб. I, Физматгиз, М., 1958.

# ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ


Абрамова А. А. метод 449
Абсолютная погрешность 17
Аналитическая функция 86

Базис пространства 333
— — ортонормированный 339
Бернулли метод 195
— числа 64
Бесселя неравенство 212
— формула интерполяционная 522, 523
Билинейная форма матрицы 376
Биортогональность 382
Бюдана—Фурье теорема 172

Вековое уравнение 368
Вековой определитель 403
Вектор матрицы собственный 367
— нулевой 329
— $n$-мерный 329
Вектор-столбец 225
Вектор-строка 225
Вектор-функция 450
Векторы линейно зависимые 330
Величина обратная 101
Верные десятичные знаки 23
Вероятностная оценка погрешности 51

Гавурина М. К. метод 449
Гамильтона—Кели тождество 389
Гаусса метод 268, 272
— формула интерполяционная 520
— — квадратурная 600
Герона процесс 104
Главная строка 282
Главный элемент 268, 282
Горизонтальная таблица разностей 501
Горнера схема 74
— — обобщенная 77
Границы действительных корней 165, 167
Графическое дифференцирование 574
— интегрирование 624
— решение уравнения 116
Гурвица теорема 398
— условия 397
Гюа теорема 175

Данилевского метод 403, 404, 410, 411, 421
Датчики случайных чисел 638
Двойной пересчет 607
Двойные разности высших порядков 557
Действительная матрица 381
Действительные корни уравнения 159, 163, 165, 167, 169, 180
Декарта теорема 174
Десятичные знаки верные 23
Детерминант матрицы 226, 264, 283, 380, 402, 403, 421
Дефект матрицы 244
Диагональная матрица 225
— таблица разностей 501
Дифференцирование графическое 574
— приближенное 562
— численное 571—573
Дроби подходящие 56, 57
— —, закон составления 57
Дробь рациональная 79
— цепная 53, 54
— — бесконечная 64 и д.
— — — расходящаяся 64
— — — сходящаяся 64
— —, звено 53
— — конечная 53

Единичная матрица 226
Единственность корня системы нелинейных уравнений 466
— — уравнения 113
— решения системы нелинейных уравнений 466

Задача теории погрешностей обратная 43
Закон распространения $\varepsilon$-ошибки в таблице конечных разностей 503
— цепи 652
Звено цепной дроби 53
Зейделя метод 268, 303
— процесс 151
Знаки верные 25
Значения собственные матрицы 421, 431
— — —, свойство экстремальное 379

Интеграл несобственный 618
— — расходящийся 619, 620
— — сходящийся 619, 620
— собственный 618
Интегралы кратные, метод Монте-Карло 641
Интегрирование 577
— графическое 624
Интерполирование в узком смысле 508
— вперед 517
— на середину 523
— назад 517
— квадратичное 510

Интерполирование линейное 510
— обратное для случая неравноотстоящих узлов 550
— — — — равноотстоящих узлов 547
— параболическое 510
— функций 507
— — двух переменных 555
Интерполяционная формула — см. *Формула интерполирования*
Итерация 100, 135, 268, 294

Каноническая норма матрицы 239
Канторовича Л. В. метод 621
Квадратичная форма 305
Квадратная матрица 225
Квадратный корень 104
Квадратура механическая 577
Квадратурная формула 578
Квадрирование корней 179, 180, 181
Квазидиагональная матрица 252
Клеточная матрица 252
Комбинация линейная векторов 330
Комбинированный метод 132
Комплексные корни 158
Конечно-разностное уравнение 195
Конечные разности 497
— — второго порядка 206
— — первого порядка 206
— — порядка *p* 207
Контроль вычислений заключительный 15
— — текущий 15
Конформные матрицы 253
Корень квадратный 104
— —, обратная величина 108
— кубический 108
— полинома действительный 169
— системы линейных уравнений 269
— уравнения 112
Корни линейной системы, уточнение 279
— уравнения в групповом смысле 186
— — действительные 159, 163, 165, 167, 169, 180
— — комплексные 158, 183, 186, 190
— — —, пара комплексных корней 186
— — —, случай двух пар 190
— — кратные 158
— —, отделение 112
— — отделенные 176
Косинус 95
— гиперболический 98
Косинусы направляющие 342
Котеса квадратурная формула 580
— коэффициенты — см. *Коэффициенты Котеса*
Коэффициенты Котеса 581, 586
— Лагранжа 531
— Фурье 210
Крамера правило 268
— формулы 271
Кратность корня 158
Кратные интегралы 641
— корни 158
Кронекера символ 226
Крылова А. Н. метод 213, 412, 421
Кубатура механическая 574
Кубический корень 108
Куммера преобразование 199

Лагранжа коэффициенты 531
— теорема 164
— формула интерполяционная 529, 535
Леверрье метод 417
Лежандра полиномы 597
Линейная зависимость векторов 330
— комбинация векторов 330
— система 306
Линейное векторное пространство 329
— подпространство 334
— преобразование 359
Лобачевского—Граффе метод 176
Логарифмическая функция 92
Люстерника метод 444

Маклорена ряд 86
Маркова формула 553
— цепь 652
Матриц равенство 226
— сумма и разность 227
— умножение 228
Матрица 225
—, величина (модуль) 238
— действительная 381
— диагональная 225
— единичная 226
— квадратичной формы 305
— квадратная 225
— квазидиагональная 252
— клеточная 252
— неособенная 232
—, норма 238, 239
— нулевая 226
— обратная 231
— —, уточнение 310
— окаймленная 252
— ортогональная 342
— особенная (сингулярная) 232
— перехода к новому базису 340
—, предел 245
— присоединенная 232
— противоположная 228
— прямоугольная 225
—, ранг 244
—, рациональная функция 237
— симметрическая 231, 376
— — положительно определенная 380
— сингулярная — см. *Матрица особенная*
—, степень 236
— транспонированная 230
— треугольная 260
—, умножение на число 227
— Фробениуса 404
— характеристическая 368
—, элементарное преобразование 263
— Якоби 451
Матрицы конформные 253
— подобные 372
— равные 226
— эквивалентные 263
Матричный ряд 247
Медленная сходимость ряда 199
Метод Абрамова А. А. 449
— Бернулли 195
— Гавурина М. К. 449
— Гаусса 268, 272, 283
— главных элементов 268, 281
— градиента — см. *Метод скорейшего спуска*
— границ 48, 49
— Данилевского развертывания векового определителя 403, 404, 410, 411, 421
— двойного пересчета 607
— Зейделя 268, 303
— знакопеременных сумм 165
— интерполяции развертывания векового определителя 403, 421, 553

Метод исчерпывания 434
— итерации 100, 135, 268, 294
— — для системы двух уравнений 148
— — — — нелинейных уравнений 474
— Канторовича Л. В. выделения особенностей 621
— касательных — см. *Метод Ньютона*
— квадратных корней 268, 289
— комбинированный 132
— Крылова А. Н. 213, 412, 421
— — развертывания векового определителя 403, 421
— —, собственные векторы матрицы 416
— — улучшения сходимости тригонометрических рядов Фурье 213
— Леверрье 417
— — развертывания векового определителя 403, 421
— Лобачевского—Греффе 176
— — — для случая действительных корней 180
— — — — — комплексных корней 183
— Люстерника улучшения сходимости процесса итерации для решения системы линейных уравнений 444
— Монте-Карло 635
— — вычисления кратных интегралов 641
— —, решение систем линейных уравнений 650
— неопределенных коэффициентов 419
— Ньютона 123 и д., 167
— — видоизмененный 131 и д.
— — для системы двух линейных уравнений 152
— — — случая комплексных корней 153
— — решения систем нелинейных уравнений 450, 452
— — — — — — модифицированный 472 и д.
— обратного интерполирования для решений уравнений 551
— окаймления 258
— ослабления — см. *Метод релаксации*
— половинного деления 118
— последовательных приближений — см. *Метод итерации*
— релаксации 268, 307
— Ричардсона 314
— скалярных произведений 428
— скорейшего спуска для решения системы линейных уравнений 490
— — — — — — нелинейных уравнений (метод градиента) 485
— степенных рядов для решения системы нелинейных уравнений 494
— Штурма 169
— хорд 119
— Эйлера—Абеля 205
— эскалаторный 314
Механическая квадратура 577
— кубатура 577
Минор матрицы 244
Модуль матрицы 238
Монте-Карло метод 635

Направляющие косинусы 342
Начальная погрешность 21
Невязка приближенного решения 279
Неособенная матрица 232
Неподвижная точка преобразования 478
Неполное частное цепной дроби 54
Несобственный интеграл 618
Норма матрицы 238
— — каноническая 239
Нормальные системы линейных уравнений 310
Нулевая матрица 226
Ньютона метод — см. *Метод Ньютона*
— теорема 167
— — интерполяционная 510, 515
— — квадратурная 580, 586

Обратная величина 101
— задача теории погрешностей 43
—матрица 231
Обратное интерполирование 547, 550
— преобразование 365
Обращение матрицы 232, 255, 442
— — методом Гаусса 285
Окаймленная матрица 252
Округление 24
Определитель 226, 264, 283
— вековой (характеристический) 380, 402
— —, развертывание 403, 421
Ортогонализация матриц 343
— столбцов 351
Ортогональная матрица 342
— система векторов 338
Основная теорема алгебры 158
Особенная матрица 232
Остаток ряда 80
Остаточная погрешность 21
Отделение корней уравнения 112 и д.
Относительная погрешность 49
Отображение сжимающееся 478
Отрицательно определенная квадратичная форма 305
Оценка погрешности вероятностная 51
— — процесса Зейделя 323, 325
— — — итераций 315
Оценки коэффициентов Фурье 211
Ошибка приближенного числа 17

Параболическая формула 589
Переход к новому базису 340
Переходные вероятности 652
Перрона теорема 382
Плотность вероятности 636
Погрешность абсолютная 17
— — разности 33
— — суммы 31
— действий 15
— задачи 20
— интерполяционных формул Ньютона 537
— — — Лагранжа 535
— квадратурных формул 604
— метода 16, 20
— начальная 21
— округления 15, 21, 24
— остаточная 21
— относительная 19, 25
— — корня 39
— — произведения 35
— — степени 39
— — частного 38
— предельная 19
— — практическая 51
— приближений итерационных 138, 317
— — процесса Зейделя 322, 325
— произведения 35
— разности 33
— суммы 31

Погрешность суммы относительная 28, 29, 32
— формулы квадратурной 606
— центральных интерполяционных формул 539
— частного 38
Подобие матриц 372
Подпространство линейное 334
Подходящие дроби 56, 57
— —, закон составления 57
Показательная функция 88
Полином 74
— интерполяционный Лежандра 597
— — Ньютона 509
— характеристический матрицы 368
Половинное деление 118
Положительно определенная квадратичная форма 305
— — матрица 380
Порядок матрицы 225
Правило Крамера 268
— трех восьмых 586
Предел последовательности матрицы 245
Предельная погрешность 19
Преобразование вращения 362
— Куммера 199
— линейное 359
— матриц 263
— обратное 365
— проектирования 361
— Эйлера—Абеля 206
Приближенное вычисление частных производных 576
— число 17
Принсгейма матрица 68
Принцип аргумента 162
— равных влияний 43
Присоединенная матрица 232
Произведение вектора на число 330
— векторов скалярное 336
— матриц 227, 228, 253
Пространство линейное векторное 330
— решение однородной системы 356
Противоположная матрица 228
Процесс Герона 104
— Зейделя 151
Прямоугольная матрица 225

Равенство матриц 226
Развертывание вековых определителей 402 и д.
Разделенные разности 542
— —, таблица 543
Разложение в цепную дробь рациональной функции 71
— — — — $e^x$ 72
— — — — $\operatorname{tg} x$ 72
— матрицы билинейное 384
Размерность пространства 332
Разности конечные 497
— — второго порядка 206
— — двойные высших порядков 557
— — первого порядка 206
— — разделенные 542
— — центральные, таблица 519
— — частные 557
— — $p$-го порядка 207
Разность матриц 227, 253
Ранг матрицы 244
Рациональная дробь 79
— функция матрицы 237
Релаксация 268, 307
Решение конечно-разностного уравнения 195
Решение уравнений графическое 116 и д.
Ричардсона метод 314
— экстраполяция 609
Ряд Маклорена 86
— матричный 225, 247, 249
— Тейлора 86
— тригонометрический 247
— числовой 80
— — сходящийся 80

Сжимающееся отображение 478
Символ Кронекера 229
Симметрическая матрица 231, 376
Симметрия эрмитова 336
Симпсона формула 584
— — общая 58
Сингулярная матрица 232
Синус 95
— гиперболический 98
Система векторов ортогональная 338
— двух уравнений 152
— линейная нормальная 306
— линейных уравнений, корни 269
— — —, метод Монте-Карло 650
— — —, методы точные 268
— — — нормальная 310
Скалярное произведение векторов 336
Скорость сходимости процесса Ньютона для системы 465
— — — — — уравнения 128
След матрицы 360
Случайная величина 635
Случайные числа 635
— —, способ получения 638
Собственное значение матрицы 367
— — — второе 431
— — — первое 421
Собственные векторы, метод Данилевского 411
— —, — Крылова 416
— значения матрицы 421, 431
— — —, свойство экстремальное 379
— элементы положительно определенной симметрической матрицы 437
Собственный вектор матрицы 367
— интеграл 618
Соотношения биортогональности 382
Спектр матрицы 369
Степенной ряд (ряд Тейлора) 86
— — матричный 386
Степень матрицы 236
— обобщенная 505
Стирлинга формула дифференцирования 567
— — интерполяционная 521
Строка главная 282
Сумма векторов 329
—матриц 227, 253
— матричного ряда 247
— числового ряда 80
Суммирование приближенное тригонометрических рядов 222
Схема вычислительная 13
— Горнера 74
— — обобщенная 77
— единственного деления 274
— — —, таблица 275
— Халецкого 291
Сходимость процесса Зейделя 320, 322, 326, 392
— — — для нормальной системы линейных уравнений 395
— — итерации 136 и д.

Сходимость процесса итерации для систем линейных уравнений 315, 390
— — — — — нелинейных уравнений 481, 483
— — Ньютона для систем нелинейных уравнений 456, 460, 465, 469
— — —, скорость 128, 465
— — —, устойчивость 469
— ряда матричного 247, 386
— — — абсолютная 247

Таблица конечных разностей функции $y=\lg x$ 516
— предельной относительной погрешности (определение по числу верных знаков) 29
— разностей горизонтальная 501
— — диагональная 501
— — разделенных 543
— — функции $y=e^x$ 511
— — — $y=\sin x$ 517
— распространения ε-ошибки в таблице конечных разностей 503
— центральных разностей 519
— числа верных знаков в зависимости от предельной относительной погрешности 30
Тангенс 96
— гиперболический 99
Тейлора ряд 86
Текущий контроль вычислений 115
Теорема Бюдана—Фурье 172
— Гурвица 398
— Гюа 175
— Декарта 174
— Ньютона 167
— основная алгебры 158
— Перрона 382
— Принсгейма 68
— Штурма 170
Тождество Гамильтона—Кели 389
Точка неподвижная преобразования 478
Транспонирование матрицы 230
Трансцендентные функции матрицы 217
Треугольная матрица 260
Тригонометрические ряды 222

Узлы интерполирования 507, 541
Улучшение сходимости ряда 86, 199
— — — степенного методом Эйлера—Абеля 205
— — — Фурье методом Крылова 213
Уравнение вековое 368
— конечно-разностное 195
— — ,решение 195
— матрицы характеристическое 195, 368
Условие сходимости процесса Зейделя второе 322
— — — — первое 320
— — — — по $l$-норме 326
Условия Гурвица 397
Устойчивость сходимости процесса Ньютона при варьировании начального приближения 469
Уточнение корней линейной системы 279
— обратной матрицы 310

Форма квадратичная 305
— матрицы билинейная 376
Формула интегрирования Эйлера—Маклорена 616
— интерполяции квадратичного 510
— — Бесселя 522
— — — квадратичная 523
— — — на середину 523
— — Гаусса вторая 520
— — — первая 521
— — Лагранжа 529
— — —, оценка погрешности 535
— — линейного 510
— — Ньютона вторая 515
— — — для значений аргумента неравноотстоящих 546
— — — для функции двух переменных 559
— — —, оценка погрешности 538
— — — первая 509, 510
— — — —, дифференцирование 563
— — — —, оценка погрешности 538
— — — с разделенными разностями 546
— — параболического 510
— — Стирлинга 521
— квадратурная 578
— — Гуасса 600
— — Котеса 580
— — Ньютона 580, 586
— — Чебышева 593
— кубатурная 627
— — типа Симпсона 629
— Маркова А. А. 553
— параболическая 589
— погрешности общая 41
— Симпсона 584
— — общая 589
— —, остаточный член 585
— Стирлинга, дифференцирование 567
— трапеций 582
— — общая 588
— —, остаточный член 588
Формулы дифференцирования центральные 567
— — численного 571
— интерполяционные с постоянным шагом 525
— — с центральными разностями 519
— Крамера 271
— Ньютона — Котеса 581
— — — высших порядков 586
Фробениуса матрица 404
— нормальный вид определителя 404
Функции матрицы рациональные 237
— — трансцендентные 251
Функция двух переменных 555
—, заданная таблицей 46
— интерполирующая 507
— матричная 456
— распределения 636
— $y=e^x$ 88
— $y=\lg x$ 92, 516
— $y=\sin x$ 517
Фурье коэффициенты 210

Халецкого схема 291
Характеристический определитель матрицы 380
— полином матрицы 368
Характеристическое уравнение 195

Центральные формулы дифференцирования 567

Цепная дробь 53, 54
— — бесконечная 64
— — — расходящаяся 64
— — — сходящаяся 64
— —, звено 53
— — конечная 53
Цепь Маркова дискретная 652
Цифра значащая 22
— сомнительная 25

Частное неполное цепной дроби 54
Частные производные 575
— разности 557
Чебышева формула квадратурная 593
Числа случайные 635, 638
Численное дифференцирование 571
Число Бернулли 64
— верных знаков произведения 37
— — — частного 39
— действительных корней полинома 169
— перемен знаков системы чисел верхнее 172
— — — — — нижнее 172
— приближенное 17
— —, погрешность 17
— характеристическое 367
Числовой ряд 80
Член $k$-го звена цепной дроби 53

Шаг интерполяции 508
Штурма метод 169
— теорема 170

Эйлера — Абеля метод 205
— — — преобразование 206
Эйлера—Маклорена формула интегрирования 616
Эквивалентность матриц 263
Экстраполяция Ричардсона 609
— — для случая формулы трапеций 609
— — формулы Симпсона 610
Экстраполирование 508
— вперед 517
— назад 517
Экстремальное свойство собственных значений матрицы 379
Элемент главный 268, 282
— матрицы 225
Элементы собственные положительно определенной симметрической матрицы 437
Эрмитова симметрия 336

Якоби матрица 451