

С.П. Шарый

Курс  
ВЫЧИСЛИТЕЛЬНЫХ  
МЕТОДОВ

# Курс ВЫЧИСЛИТЕЛЬНЫХ МЕТОДОВ

С. П. ШАРЫЙ

Федеральный исследовательский центр  
информационных и вычислительных технологий  
Новосибирский государственный университет

Новосибирск – 2025

УДК 519.6

**Шарый С.П.** Курс вычислительных методов. – Москва–Ижевск:  
Институт компьютерных исследований, 2025. – 834 с.

Книга является систематическим учебником по курсу вычислительных методов и написана на основе лекций, читаемых автором на механико-математическом факультете Новосибирского государственного университета. Помимо традиционных разделов вычислительной математики, с которых начинается эта дисциплина, в книге широко изложены также методы интервального анализа и ряд результатов нелинейного анализа, нашедшие успешные применения в современных численных методах.

# Оглавление

<b>Предисловие</b>	<b>10</b>
<b>Глава 1. Общие вопросы вычислений</b>	<b>11</b>
1.1 Предмет вычислительной математики . . . . .	11
1.2 Погрешности приближённых величин . . . . .	14
1.3 Погрешности и вычисления . . . . .	18
1.4 Компьютерная арифметика . . . . .	26
1.5 Интервальная арифметика . . . . .	32
1.6 Интервальные расширения функций . . . . .	40
1.7 Обусловленность математических задач . . . . .	45
1.8 Устойчивость алгоритмов . . . . .	48
1.9 Элементы конструктивной математики . . . . .	53
1.10 Сложность задач и трудоёмкость алгоритмов . . . . .	56
1.11 Доказательные вычисления на ЭВМ . . . . .	60
Литература к главе 1 . . . . .	64
<b>Глава 2. Численные методы анализа</b>	<b>67</b>
2.1 Введение . . . . .	67
2.2 Интерполирование функций . . . . .	72
2.2а Постановка задачи и её свойства . . . . .	72
2.2б Алгебраическая интерполяция . . . . .	77
2.2в Интерполяционный полином Лагранжа . . . . .	80
2.2г Разделённые разности и их свойства . . . . .	84
2.2д Интерполяционный полином Ньютона . . . . .	91
2.2е Погрешность алгебраической интерполяции . . . . .	96
2.2ж Тригонометрическая интерполяция . . . . .	102
2.3 Полиномы Чебышёва . . . . .	108

2.3а	Определение и основные свойства . . . . .	108
2.3б	Применения полиномов Чебышёва . . . . .	114
2.3в	Обусловленность алгебраической интерполяции .	118
2.4	Интерполяция с кратными узлами . . . . .	121
2.5	Общие факты интерполяции . . . . .	127
2.5а	Интерполяционный процесс . . . . .	127
2.5б	Сводка результатов и обсуждение . . . . .	129
2.6	Сплайны . . . . .	137
2.6а	Элементы теории . . . . .	137
2.6б	Интерполяционные кубические сплайны . . . . .	142
2.6в	Кубические сплайны (продолжение) . . . . .	147
2.6г	Погрешность интерполирования сплайнами . . . .	150
2.6д	Экстремальное свойство кубических сплайнов .	152
2.7	Нелинейные методы интерполяции . . . . .	154
2.8	Численное дифференцирование . . . . .	160
2.8а	Интерполяционный подход . . . . .	161
2.8б	Оценка погрешности дифференцирования . . . .	166
2.8в	Порядок точности формул и методов . . . . .	168
2.8г	Метод неопределённых коэффициентов . . . . .	173
2.8д	Полная погрешность дифференцирования . . . .	176
2.9	Алгоритмическое дифференцирование . . . . .	180
2.10	Приближение функций . . . . .	185
2.10а	Обсуждение постановки задачи . . . . .	185
2.10б	Существование наилучшего приближения . . . .	189
2.10в	Единственность наилучшего приближения . . . .	193
2.10г	Квадратичные приближения . . . . .	198
2.11	Метод наименьших квадратов . . . . .	203
2.11а	Приближение в евклидовом подпространстве . .	203
2.11б	Приближение из линейной оболочки векторов .	209
2.11в	Геометрия квадратичного приближения . . . . .	212
2.11г	Псевдорешения систем линейных уравнений .	216
2.11д	Приложения к анализу данных . . . . .	220
2.11е	Среднеквадратичное приближение функций .	225
2.11ж	Базисы для среднеквадратичных приближений .	230
2.12	Полиномы Лежандра . . . . .	236
2.12а	Мотивация и определение . . . . .	236
2.12б	Формула Родрига . . . . .	238
2.12в	Основные свойства полиномов Лежандра . . .	243
2.13	Численное интегрирование . . . . .	248

2.13а	Постановка и обсуждение задачи . . . . .	248
2.13б	Простейшие квадратурные формулы . . . . .	252
2.13в	Квадратурная формула Симпсона . . . . .	258
2.13г	Погрешность формулы Симпсона . . . . .	261
2.13д	Интерполяционные квадратурные формулы . . .	265
2.13е	Дальнейшие формулы Ньютона–Котеса . . . . .	268
2.14	Составные квадратурные формулы . . . . .	273
2.14а	Общая идея и её обоснование . . . . .	273
2.14б	Конкретные составные квадратурные формулы .	276
2.15	Квадратуры наивысшей степени точности . . . . .	279
2.15а	Задача оптимизации квадратурных формул . . .	279
2.15б	Простейшие квадратуры Гаусса . . . . .	282
2.15в	Выбор узлов для квадратурных формул Гаусса .	285
2.15г	Практическое применение формул Гаусса . . .	288
2.15д	Погрешность квадратур Гаусса . . . . .	292
2.15е	Теорема И.П. Мысовских . . . . .	296
2.16	Метод неопределённых коэффициентов . . . . .	298
2.17	Сходимость квадратур . . . . .	302
2.18	Вычисление интегралов методом Монте-Карло . . . . .	307
2.19	Правило Рунге для оценки погрешности . . . . .	313
	Литература к главе 2 . . . . .	316
<b>Глава 3.</b>	<b>Численные методы линейной алгебры</b>	<b>322</b>
3.1	Задачи вычислительной линейной алгебры . . . . .	322
3.2	Теоретическое введение . . . . .	325
3.2а	Необходимые сведения из линейной алгебры . . .	325
3.2б	Основные понятия теории матриц . . . . .	330
3.2в	Собственные числа и собственные векторы . . . .	343
3.2г	Разложения матриц, использующие их спектр . .	348
3.2д	Сингулярные числа и сингулярные векторы . . .	351
3.2е	Свойства сингулярных чисел и векторов . . . .	355
3.2ж	Сингулярное разложение матриц . . . . .	361
3.2з	Системы линейных алгебраических уравнений .	366
3.3	Нормы векторов и матриц . . . . .	368
3.3а	Векторные нормы . . . . .	368
3.3б	Топология на векторных пространствах . . . . .	373
3.3в	Эквивалентные векторные нормы . . . . .	377
3.3г	Покомпонентная сходимость . . . . .	380
3.3д	Матричные нормы . . . . .	382

3.3е	Подчинённые матричные нормы . . . . .	387
3.3ж	Топология на множествах матриц . . . . .	392
3.3з	Энергетическая норма . . . . .	396
3.3и	Спектральный радиус . . . . .	400
3.3к	Матричный ряд Неймана . . . . .	405
3.4	Обусловленность систем линейных уравнений . . . . .	408
3.4а	Число обусловленности матриц . . . . .	408
3.4б	Хорошо и плохо обусловленные матрицы . . . . .	413
3.4в	Матрицы с диагональным преобладанием . . . . .	418
3.4г	Практическое применение числа обусловленности	423
3.5	Приложения сингулярного разложения . . . . .	427
3.5а	Исследование неособенности и ранга матриц . . . . .	427
3.5б	Решение систем линейных уравнений . . . . .	431
3.5в	Малоранговые приближения матрицы . . . . .	433
3.5г	Метод главных компонент . . . . .	435
3.6	Прямые методы решения линейных систем . . . . .	437
3.6а	Основные понятия . . . . .	437
3.6б	Решение треугольных и трапециевидных систем . . . . .	441
3.6в	Метод Гаусса для решения линейных систем . . . . .	445
3.6г	Матричная интерпретация метода Гаусса . . . . .	449
3.6д	Метод Гаусса с выбором ведущего элемента . . . . .	452
3.6е	Алгоритмы Дулитла и Кроута . . . . .	457
3.6ж	Существование LU-разложения . . . . .	461
3.6з	Разложение Холесского . . . . .	463
3.6и	Метод Холесского . . . . .	466
3.7	Методы на основе ортогональных преобразований . . . . .	471
3.7а	Обусловленность и матричные преобразования . . . . .	471
3.7б	Ортогональность и матричные вычисления . . . . .	475
3.7в	QR-разложение матриц . . . . .	477
3.7г	Матрицы вращения и метод вращений . . . . .	480
3.7д	Ортогональные матрицы отражения . . . . .	484
3.7е	Метод отражений Хаусхольдера . . . . .	488
3.8	Процессы ортогонализации . . . . .	494
3.9	Метод прогонки . . . . .	501
3.10	Стационарные итерационные методы . . . . .	507
3.10а	Краткая теория . . . . .	507
3.10б	Сходимость стационарных одношаговых методов	511
3.10в	Подготовка системы к итерационному процессу .	518
3.10г	Метод Ричардсона и его оптимизация . . . . .	522

3.10д Итерационный метод Якоби . . . . .	527
3.10е Итерационный метод Гаусса–Зейделя . . . . .	532
3.10ж Методы релаксации . . . . .	538
3.11 Нестационарные итерационные методы . . . . .	545
3.11а Теоретическое введение . . . . .	545
3.11б Метод спуска для минимизации функций . . . . .	555
3.11в Наискорейший градиентный спуск . . . . .	558
3.11г Метод минимальных невязок . . . . .	563
3.11д Метод сопряжённых градиентов . . . . .	566
3.11е Сходимость метода сопряжённых градиентов . . . . .	571
3.11ж Другой подход к методу сопряжённых градиентов	580
3.12 Методы установления . . . . .	581
3.13 Теория А.А. Самарского . . . . .	584
3.14 Вычисление определителей и обратных матриц . . . . .	588
3.15 Оценка погрешности приближённого решения . . . . .	592
3.16 Линейная задача наименьших квадратов . . . . .	601
3.17 Матричная проблема собственных значений . . . . .	608
3.17а Обсуждение постановки задачи . . . . .	608
3.17б Матрицы простой структуры . . . . .	613
3.17в Обусловленность проблемы собственных значений	617
3.17г Коэффициенты перекоса матрицы . . . . .	623
3.17д Круги Гершгорина . . . . .	626
3.17е Отношение Рэлея . . . . .	629
3.17ж Предварительное упрощение матрицы . . . . .	634
3.18 Несимметрическая проблема собственных значений . . . . .	639
3.18а Степенной метод . . . . .	640
3.18б Обратные степенные итерации . . . . .	649
3.18в Сдвиги, исчерпывание и понижение порядка . . . . .	651
3.18г Базовый QR-алгоритм . . . . .	657
3.18д Модификации QR-алгоритма . . . . .	662
3.19 Симметрическая проблема собственных значений . . . . .	665
3.19а Симметрический QR-алгоритм . . . . .	666
3.19б Метод Якоби . . . . .	669
3.19в Итерации с отношением Рэлея . . . . .	677
3.19г Трёхдиагональные матрицы . . . . .	680
3.19д Численные методы сингулярного разложения . . . . .	683
Литература к главе 3 . . . . .	685

<b>Глава 4. Решение нелинейных уравнений и их систем</b>	<b>693</b>
4.1 Обзор постановок задачи . . . . .	693
4.2 Вычислительно-корректные задачи . . . . .	699
4.2а Предварительные сведения и определения . . . . .	699
4.2б Решение уравнений вычислительно некорректно .	703
4.2в $\varepsilon$ -решения уравнений . . . . .	704
4.2г Недостаточность $\varepsilon$ -решений . . . . .	707
4.3 Существование решений уравнений и систем уравнений .	709
4.3а Векторные поля . . . . .	710
4.3б Вращение векторных полей . . . . .	712
4.3в Индексы особых точек . . . . .	716
4.3г Устойчивость особых точек . . . . .	717
4.3д Вычислительно-корректная постановка . . . . .	719
4.3е Теоремы о сжимающих отображениях . . . . .	721
4.4 Классические методы решения уравнений . . . . .	729
4.4а Предварительная локализация решений . . . . .	731
4.4б Метод половинного деления (бисекции) . . . . .	735
4.4в Метод простой итерации . . . . .	739
4.4г Интерполяционные методы . . . . .	743
4.4д Метод Ньютона и его модификации . . . . .	748
4.4е Методы Чебышёва . . . . .	753
4.4ж Оценка погрешности приближённого решения .	756
4.5 Классические методы решения систем уравнений . . . . .	759
4.5а Метод простой итерации . . . . .	759
4.5б Метод Ньютона и его модификации . . . . .	762
4.6 Интервальные системы линейных уравнений . . . . .	765
4.6а Интервальные уравнения и их решения . . . . .	765
4.6б Численные методы для интервальных систем . .	771
4.7 Интервальные методы решения уравнений . . . . .	779
4.7а Основы интервальной техники . . . . .	779
4.7б Одномерный интервальный метод Ньютона . .	783
4.7в Многомерный интервальный метод Ньютона .	789
4.7г Метод Кравчика . . . . .	793
4.8 Глобальное решение уравнений и систем уравнений .	798
Литература к главе 4 . . . . .	805
<b>Обозначения</b>	<b>809</b>
<b>Краткий биографический словарь</b>	<b>813</b>

<i>Оглавление</i>	9
<b>Предметный указатель</b>	<b>823</b>

# Предисловие

Эта книга написана на основе курса лекций по вычислительным методам, которые читаются автором на механико-математическом факультете Новосибирского государственного университета. Её содержание в основной своей части традиционно и повторяет на современном уровне тематику, заданную в классических учебниках по этому предмету. Условно материал книги можно было бы назвать «вычислительные методы-1», поскольку в стандарте университетского образования существуют и следующие части этого большого курса, посвящённые численному решению дифференциальных уравнений (как обыкновенных, так и в частных производных), интегральных уравнений и др. Для понимания текста книги достаточно математической подготовки в объёме обычных курсов математического анализа и алгебры.

Вместе с тем книга имеет ряд особенностей. Во-первых, в ней широко представлены элементы интервального анализа и современные интервальные методы для решения традиционных задач вычислительной математики. Во-вторых, автор счёл уместным поместить в книгу краткие сведения из нелинейного анализа (очерк теории вращения векторных полей), которые необходимы при тщательном исследовании решений систем нелинейных уравнений. Книга имеет значительный объем, который превосходит возможности лекционного курса, но это сделано намеренно, чтобы учебник мог также служить «книгой для чтения» по предмету. С другой стороны, эта книга, конечно, не охватывает многие важные и интересные вопросы вычислительной математики даже в пределах затронутых тем, не претендую быть справочником.

Автор благодарен свой жене Ирине за любовь, моральную поддержку и неоценимую помошь в работе.

## Глава 1

# Общие вопросы вычислений

## 1.1 Предмет вычислительной математики

Курс методов вычислений является введением в обширную математическую дисциплину — вычислительную математику, которую можно неформально определить как «математику вычислений» или «математику, возникающую в связи с разнообразными процессами вычислений». При этом под «вычислениями» понимается не только получение числового ответа к задаче, т. е. доведение результата решения «до числа», но и нахождение конструктивных представлений или приближений для различных математических объектов. С 70-х годов XX века, когда качественно нового уровня достигли развитие вычислительных машин и их применение во всех сферах жизни общества, можно встретить расширительное толкование содержания вычислительной математики как «раздела математики, включающего круг вопросов, связанных с использованием ЭВМ» (определение А.Н. Тихонова).

Иногда в связи с вычислительной математикой и методами вычислений используют термин «численный анализ», возникший в США в конце 40-х годов XX века. Он более узок по содержанию, так как во главу угла ставит расчёты числового характера, а аналитические или символьные вычисления, без которых в настоящее время невозможно представить вычислительную математику и её приложения, отодвигает на второй план.

В действительности вычислительная математика — одна из самых древних ветвей математики, богатая своими собственными идеями и методами, и её положение на общем дереве математических наук замечательно своей тесной связью с практикой. В конце XIX – начале XX веков в связи с общим бурным развитием науки в Новом времени вычислительная математика выделилась в самостоятельное научное направление.

Развитие вычислительной математики в различные исторические периоды имело свои особенности и акценты. Начиная с античности (вспомним Евклида и Архимеда) и вплоть до настоящего времени вычисления гармонично входили в сферу научных интересов крупнейших математиков — И. Кеплера, И. Ньютона, Л. Эйлера, Ж.-Л. Лагранжа, К.Ф. Гаусса, Н.И. Лобачевского, К.Г. Якоби, П.Л. Чебышёва, А.Н. Тихонова, С.Л. Соболева и многих других, чьи имена остались в названиях популярных численных методов и важнейших результатов вычислительной математики. Дальнейшее развитие и дифференциация математики, дробление её на ветви и отдельные дисциплины, привели к большей специализации, в частности, в вычислительной математике. Её типичные задачи и общее состояние на начало XX века можно увидеть в первом в мире систематическом учебнике методов вычислений — «Лекциях о приближённых вычислениях» акад. А.Н. Крылова [9]. Они были впервые изданы в 1906 году и вплоть до середины XX века выдержали семь изданий.

В XX веке, и особенно в его второй половине, на первый план выдвинулись разработка и применение конкретных практических алгоритмов для решения сложных задач математического моделирования (в основном вычислительной физики, механики и управления). Необходимо было строить летательные аппараты, суда, автомобили и другие сложные технические устройства, управлять ими, запускать и наводить ракеты, выходить в космос и т. п.

Глубокое и сильное влияние на математику, в частности вычислительную, оказали формирование в первой половине XX века понятия алгоритма и далее развернувшиеся его исследования. Напомним, что *алгоритм* — это конечная совокупность инструкций, однозначно определяющая содержание и порядок выполнения действий исполнителя для решения некоторой задачи или класса задач. С другой стороны, на развитие вычислительной математики очень большое влияние оказывали конкретные способы вычислений и вычислительные устройства, которые возникали по ходу развития технологий и применялись для

решения задач практики. В частности, огромный по своим последствиям импульс был придан вычислительной математике в середине XX века в связи с появлением и распространением электронных цифровых вычислительных машин, кратко называемых ныне «компьютерами».<sup>1</sup>

В связи с процессами вычислений нас интересуют в основном три типа задач:

- Как найти числовые значения интересующих практику величин, которые определяются какими-то математическими конструкциями или операциями? К примеру, как с помощью конечных вычислительных процедур найти производную, интеграл, решение уравнения?

Как конструктивно найти (вычислить) тот или иной математический объект или его приближение? В частности, как построить решение дифференциального или интегрального уравнения и т. п.?

- Каково количество элементарных операций, необходимое для нахождения решений тех или иных задач? Иными словами: какова трудоёмкость задачи? Может ли она быть уменьшена и как именно? Сюда же относятся затраты других ресурсов вычислительного процесса — памяти, обращений к внешним устройствам и т. п.
- Если алгоритм для решения задачи уже известен, то как наилучшим образом организовать вычисления по этому алгоритму на том или ином конкретном вычислительном устройстве? Это относится, например, к ускорению выполнения алгоритма. Часто требуется уменьшить погрешности вычислений по данному алгоритму, сократить затраты памяти и т. п.

Вопросы из последнего пункта стали особенно актуальными в связи с развитием различных архитектур электронных вычислительных машин, в частности, как результат вхождения в нашу повседневную жизнь многопроцессорных и параллельных компьютеров. Даже на бытовом уровне эти способы организации вычислений повсеместно реа-

---

<sup>1</sup>Строго говоря, термин «компьютер» является более широким по значению, и электронная цифровая вычислительная машина — одна из его разновидностей. Вообще, компьютеры могут быть не только электронными, но и механическими, оптическими, биологическими, квантовыми и т. п.

лизуются сейчас в виде многоядерных процессоров для персональных компьютеров и смартфонов.

Ясно, что все три отмеченных выше типа вопросов тесно связаны между собой. К примеру, если нам удаётся построить алгоритм для решения какой-либо задачи, то, оценив сложность его исполнения, мы тем самым предъявляем и верхнюю оценку трудоёмкости решения этой задачи.

Исторически сложилось, что исследования по второму пункту относятся главным образом к другим разделам математики — к различным теориям вычислительной сложности и к теории алгоритмов, которая в 30-е годы XX века вычленилась из абстрактной математической логики. Но традиционная вычислительная математика, предметом которой считается построение и исследование конкретных численных методов, также немало способствует прогрессу в этой области.

Аналогично, исторические и организационные причины привели к тому, что вычислительные методы для решения некоторых задач порой относятся к своим специфичным математическим дисциплинам. Например, численные методы для отыскания экстремумов различных функций являются предметом вычислительной оптимизации, теории принятия решений, исследования операций и даже теории систем и теории управления. В существующей классификации математических дисциплин [34] это отдельные разделы, не включённые в вычислительную математику.

## 1.2 Погрешности приближённых величин

Общеизвестно, что в практических задачах числовые данные почти всегда не вполне точны и содержат некоторые погрешности и неточности. Если эти данные являются, к примеру, результатами измерений непрерывно изменяющихся величин, то за редким исключением они не могут быть произведены абсолютно точно. То же самое относится к результатам большинства вычислений с вещественным типом данных (как на ЭВМ, так и вручную). Наконец, погрешность численного решения задачи может быть вызвана иррациональностью каких-то величин, которые используются при её математической формализации. Например, такова ситуация с вычислением длины окружности радиуса  $r$ , равной  $L = 2\pi r$ , или площади этого круга  $S = \pi r^2$ . Здесь  $\pi = 3.14159265\dots$  — число иррациональное.

*Погрешностью* приближённого значения  $\tilde{x}$  какой-либо величины называют разность между  $\tilde{x}$  и точным значением  $x^*$  этой величины, т. е.  $\tilde{x} - x^*$ .<sup>2</sup> На практике точное значение  $x^*$  интересующей нас величины (которое называют также «истинным значением»), как правило, неизвестно, что имеет важные методические следствия.

Во-первых, чаще всего неизвестен и знак погрешности, так что более удобно оперировать *абсолютной погрешностью*  $\tilde{\Delta}$  приближённой величины, которая определяется как

$$\tilde{\Delta} = |\tilde{x} - x^*|, \quad (1.1)$$

т. е. как модуль погрешности.<sup>3</sup>

Во-вторых, вместо точной абсолютной погрешности приходится довольствоваться её приближёнными верхними оценками. Их обычно находят, анализируя способ получения приближённого значения, т. е. использованной технологии измерений, свойств прибора, численного метода и т. п. Наилучшую возможную в данных условиях оценку сверху для абсолютной погрешности называют *пределной* (или *граничной*) *абсолютной погрешностью*. В самом этом термине содержится желание иметь эту величину как можно более точной, т. е. как можно меньшей.

Таким образом, если  $\Delta$  — предельная абсолютная погрешность приближения  $\tilde{x}$  для точного значения  $x^*$ , то

$$\tilde{\Delta} = |\tilde{x} - x^*| \leq \Delta,$$

и потому

$$\tilde{x} - \Delta \leq x^* \leq \tilde{x} + \Delta.$$

Это двустороннее неравенство часто выражают следующей краткой условной записью:

$$x^* = \tilde{x} \pm \Delta.$$

Фактически вместо точного числа мы имеем здесь целый диапазон значений — числовой интервал  $[\tilde{x} - \Delta, \tilde{x} + \Delta]$  возможных представителей для точного значения рассматриваемой величины.

<sup>2</sup> В обыденной речи наряду с термином «погрешность» используется также слово «ошибка», но в современной метрологии так называют значения величины, которые имеют с ней мало общего, отбраковываются и не идут в дальнейшую обработку.

<sup>3</sup> Нередко термин «абсолютная погрешность» используется в другом смысле и обозначает величину, которую мы называли просто «погрешностью».

На практике указание одной только абсолютной погрешности недостаточно для характеристики качества рассматриваемого приближения. К примеру, для  $x^* = 10$  абсолютная погрешность, равная единице, соответствует довольно грубому приближению, тогда как для  $x^* = 10\,000$  та же погрешность обеспечивается, как правило, лишь весьма тщательным и высокоточным измерением. Более полное понятие о качестве приближения даёт *относительная погрешность* приближения, которая определяется как отношение абсолютной погрешности к модулю значения величины:

$$\tilde{\delta} = \frac{\tilde{\Delta}}{|x^*|}. \quad (1.2)$$

Ясно, что она имеет смысл для ненулевых величин.

В условиях недоступности точного (истинного) значения  $x^*$  относительную погрешность обычно полагают равной

$$\tilde{\delta} = \frac{\tilde{\Delta}}{|\tilde{x}|}, \quad (1.3)$$

где  $\tilde{x}$  — рассматриваемое приближение к  $x^*$ . При близости  $x^*$  и  $\tilde{x}$  такая замена почти не отражается на значении и содержательном смысле относительной погрешности. Ниже мы в равной мере будем пользоваться обеими формулами (1.2) и (1.3).

Относительная погрешность — безразмерная величина, и часто её выражают в процентах. О практичесности и применимости относительной погрешности можно сказать то же самое, что и по поводу абсолютной погрешности, на которую она опирается: точное значение  $\tilde{\delta}$  часто неизвестно ввиду недоступности  $x^*$  и  $\tilde{\Delta}$  на практике. Как следствие, оперируют верхними оценками для  $\tilde{\delta}$ . *Предельной относительной погрешностью* некоторого приближённого значения называют число  $\delta$ , в данных условиях наилучшим образом оценивающее сверху его относительную погрешность. Таким образом, если  $\Delta$  — предельная абсолютная погрешность значения  $\tilde{x}$  для величины с точным значением  $x^*$ , то

$$\tilde{\delta} = \frac{|\tilde{x} - x^*|}{|\tilde{x}|} \leq \delta = \frac{\Delta}{|\tilde{x}|}. \quad (1.4)$$

Отсюда следует двустороннее неравенство

$$\tilde{x} - \delta |\tilde{x}| \leq x^* \leq \tilde{x} + \delta |\tilde{x}|.$$

Стоит отметить логарифмическую относительную погрешность, рассмотренную в работе [18] и определяемую как

$$\tilde{\delta}_{\log} := \left| \ln \frac{\tilde{x}}{x^*} \right|,$$

т. е. как модуль натурального логарифма отношения приближённого значения величины к точному. Она обладает большей гибкостью в сравнении с классическими определениями (1.2) и (1.3), хотя на практике её оценивание нередко менее удобно. Логарифмическая относительная погрешность, в частности, симметрична относительно своих аргументов. В то же время при близких  $\tilde{x}$  и  $x^*$  справедливо  $\tilde{\delta} \approx \tilde{\delta}_{\log}$ , так как

$$\frac{|\tilde{x} - x^*|}{|x^*|} = \left| \frac{\tilde{x}}{x^*} - 1 \right| \approx \left| \ln \frac{\tilde{x}}{x^*} \right|.$$

Говоря про абсолютную или относительную погрешность, обычно опускают прилагательное «предельная», поскольку именно предельные (границочные) погрешности являются реально доступными нам величинами, с которыми и работают на практике.<sup>4</sup> При этом относительная погрешность является наиболее адекватным эквивалентом популярного, но нестрогого понятия «точности» приближённой величины. Фактически относительная погрешность показывает, сколько в долях единицы приближённой величине «не хватает» до истинного значения и насколько мы можем в ней сомневаться. Относительная погрешность, большая 100 %, означает, что даже знак числа не известен надёжно, и это нередко происходит при работе с малыми приближёнными величинами.

Значащими цифрами приближённого числа называются цифры из его представления в заданной системе счисления, начиная с первой слева, отличной от нуля, и все следующие за ней. Содержательное определение этого понятия состоит в том, что значащая цифра — это цифра из представления числа, которая «что-то значит» — даёт существенную информацию о его относительной погрешности. Например, в каждом из десятичных чисел 0.01234567, 1 234 567 и 1 234.567 значащими являются по 7 цифр, начиная с 1. Нули слева в записи числа, меньшего единицы, отвечают за масштаб этого числа, который не влияет на его относительную погрешность.

---

<sup>4</sup>Они могут быть, к примеру, даны в спецификациях используемых технических устройств, могут быть a priori оценены из каких-либо содержательных соображений или же найдены из опыта и т. п.

Полезно различать *верные* и *сомнительные* (ненадёжные) значащие цифры приближённого числа. Значащая цифра называется верной, если абсолютная погрешность числа не превосходит половины единицы разряда, который соответствует этой цифре. Очевидно, что в этом случае она наилучшим образом представляет рассматриваемое значение, и большего мы требовать от значащей цифры не можем. В действительности это условие является довольно жёстким (особенно для последних значащих цифр) и в реальной жизни может быть выполнено не всегда. Часто его заменяют более мягким и реалистичным условием, что абсолютная погрешность не превосходит единицы соответствующего разряда. Если сформулированные условия на значащую цифру не удовлетворены, то она называется *сомнительной*.

Отметим, что если какая-то значащая цифра верна, то ясно, что и предшествующие ей слева значащие цифры также являются верными, поскольку для них условие на величину погрешности также выполнено. По этой причине для того, чтобы охарактеризовать точность представления какого-либо приближённого числа часто говорят о количестве его верных значащих цифр. Соответственно, синонимом ухудшения (увеличения) погрешности приближённого числа является «потеря значащих цифр».

При записи приближённых чисел имеет смысл изображать их так, чтобы сама форма написания давала характеристику об их точности. Ясно, что нет большого смысла указывать много сомнительных (ненадёжных) цифр в представлении чисел. Обычно принимают за правило писать числа так, чтобы все их значащие цифры, кроме может быть последней, были верны, а последняя цифра была бы неточной не более чем на единицу. Например, согласно этому правилу число 1 234 000, у которого цифра 4 уже неточна и может быть равна 3, 4, 5 или 6, нужно записывать в виде  $1.23 \cdot 10^6$ .

### 1.3 Погрешности и вычисления

В этом параграфе мы исследуем вопрос о том, как изменяются абсолютные и относительные погрешности при выполнении арифметических операций с приближёнными числами.

Приближённое число с заданной абсолютной погрешностью — это фактически целый интервал значений. По этой причине для абсолютных погрешностей поставленный выше вопрос решается с помощью так

называемой интервальной арифметики, которая рассматривается далее в § 1.5. Здесь мы приводим несколько иное решение вопроса, иногда более удобное для теории или практического использования, но имеющее приближённый характер для умножения и деления.

**Предложение 1.3.1** *Абсолютная погрешность суммы или разности приближённых чисел равна сумме их абсолютных погрешностей.*

**Доказательство.** Если  $x_1^*, x_2^*$  — точные значения рассматриваемых чисел,  $\tilde{x}_1, \tilde{x}_2$  — их приближённые значения, а  $\Delta_1, \Delta_2$  — соответствующие предельные абсолютные погрешности, то

$$\tilde{x}_1 - \Delta_1 \leq x_1^* \leq \tilde{x}_1 + \Delta_1, \quad (1.5)$$

$$\tilde{x}_2 - \Delta_2 \leq x_2^* \leq \tilde{x}_2 + \Delta_2. \quad (1.6)$$

Складывая эти неравенства почленно, получим

$$(\tilde{x}_1 + \tilde{x}_2) - (\Delta_1 + \Delta_2) \leq x_1^* + x_2^* \leq (\tilde{x}_1 + \tilde{x}_2) + (\Delta_1 + \Delta_2).$$

Полученное соотношение означает, что величина  $\Delta_1 + \Delta_2$  является предельной абсолютной погрешностью суммы  $\tilde{x}_1 + \tilde{x}_2$ .

Умножая обе части неравенства (1.6) на  $(-1)$ , получим

$$-\tilde{x}_2 - \Delta_2 \leq -x_2^* \leq -\tilde{x}_2 + \Delta_2.$$

Складывая почленно с неравенством (1.5), получим

$$(\tilde{x}_1 - \tilde{x}_2) - (\Delta_1 + \Delta_2) \leq x_1^* - x_2^* \leq (\tilde{x}_1 - \tilde{x}_2) + (\Delta_1 + \Delta_2).$$

Отсюда видно, что величина  $\Delta_1 + \Delta_2$  является предельной абсолютной погрешностью разности  $\tilde{x}_1 - \tilde{x}_2$ . ■

Несмотря на простоту доказанного результата, он имеет важные практические следствия. Если при вычислении некоторой суммы (ряда и т. п.) погрешность какого-либо слагаемого окажется большой, то дальнейшая полная погрешность суммы уже не сможет стать меньше погрешности этого слагаемого, сколь бы точными ни были последующие слагаемые. Поэтому для получения точной суммы необходимо обеспечить точность всех её слагаемых.

Для умножения и деления формулы преобразования абсолютной погрешности более громоздки и менее точны.

**Предложение 1.3.2** Если приближённые величины  $\tilde{x}_1, \tilde{x}_2$  имеют абсолютные погрешности  $\Delta_1$  и  $\Delta_2$ , то абсолютная погрешность произведения  $\tilde{x}_1\tilde{x}_2$  не превосходит  $|\tilde{x}_1|\Delta_2 + |\tilde{x}_2|\Delta_1 + \Delta_1\Delta_2$ .

Если  $\tilde{x}_2 \neq 0$  и известно, что точное значение этой величины  $x_2^*$  — ненулевое, а её относительная погрешность  $\delta_2 = \Delta_2/|x_2^*|$  меньше 1, то абсолютная погрешность частного  $\tilde{x}_1/\tilde{x}_2$  не превосходит

$$\frac{|\tilde{x}_1|\Delta_2 + |\tilde{x}_2|\Delta_1}{\tilde{x}_2^2} \cdot \frac{1}{1 - \delta_2}.$$

**Доказательство.** Имеем

$$\begin{aligned} |\tilde{x}_1\tilde{x}_2 - x_1^*x_2^*| &= |\tilde{x}_1\tilde{x}_2 - \tilde{x}_1x_2^* + \tilde{x}_1x_2^* - x_1^*x_2^*| \leq \\ &\leq |\tilde{x}_1(\tilde{x}_2 - x_2^*)| + |x_2^*(\tilde{x}_1 - x_1^*)| = \\ &= |\tilde{x}_1(\tilde{x}_2 - x_2^*)| + |(\tilde{x}_2 - (\tilde{x}_2 - x_2^*))(\tilde{x}_1 - x_1^*)| \leq \\ &\leq |\tilde{x}_1(\tilde{x}_2 - x_2^*)| + |\tilde{x}_2(\tilde{x}_1 - x_1^*)| + |(\tilde{x}_1 - x_1^*)(\tilde{x}_2 - x_2^*)| \leq \\ &\leq |\tilde{x}_1|\Delta_2 + |\tilde{x}_2|\Delta_1 + \Delta_1\Delta_2, \end{aligned}$$

что доказывает первое утверждение.

Доказательство второго утверждения:

$$\begin{aligned} \left| \frac{\tilde{x}_1}{\tilde{x}_2} - \frac{x_1^*}{x_2^*} \right| &= \left| \frac{\tilde{x}_1x_2^* - x_1^*\tilde{x}_2}{\tilde{x}_2x_2^*} \right| = \\ &= \left| \frac{\tilde{x}_1(\tilde{x}_2 - (\tilde{x}_2 - x_2^*)) - (\tilde{x}_1 - (\tilde{x}_1 - x_1^*))\tilde{x}_2}{\tilde{x}_2x_2^*} \right| = \\ &= \frac{|-\tilde{x}_1(\tilde{x}_2 - x_2^*) + (\tilde{x}_1 - x_1^*)\tilde{x}_2|}{|\tilde{x}_2x_2^*|} \leq \frac{|\tilde{x}_1|\Delta_2 + |\tilde{x}_2|\Delta_1}{|\tilde{x}_2x_2^*|} = \\ &= \frac{|\tilde{x}_1|\Delta_2 + |\tilde{x}_2|\Delta_1}{|\tilde{x}_2^2|} \cdot \frac{1}{|1 - (\tilde{x}_2 - x_2^*)/\tilde{x}_2|} = \\ &= \frac{|\tilde{x}_1|\Delta_2 + |\tilde{x}_2|\Delta_1}{\tilde{x}_2^2} \cdot \frac{1}{1 - \delta_2}. \end{aligned}$$
■

Интересно, что в формуле для абсолютной погрешности частного оказалось задействована относительная погрешность делителя. Полученные формулы выглядят довольно громоздко, и если нужны конкретные числовые значения погрешности, то их можно получить с

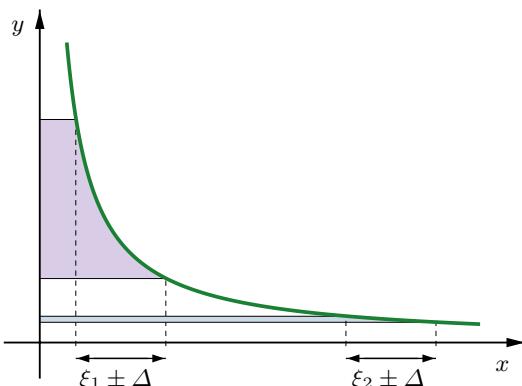


Рис. 1.1. Изменения частного больше при малых значениях делителя

помощью интервальной арифметики, рассматриваемой ниже в § 1.5. С другой стороны, формулы из предложения 1.3.2 позволяют проанализировать ситуацию с погрешностями «в целом».

Отметим, что погрешность деления выше в области малых значений делителя, когда  $\tilde{x}_2^2$  в знаменателе мало. Это иллюстрируется на рис. 1.1, где показана гипербола  $y = a/x$  и влияние изменения делителя  $x$  на результат частного при различных значениях делителя,  $\xi_1$  и  $\xi_2$ . Таким образом, для уменьшения погрешностей деления следует, при прочих равных условиях, выбирать делитель как можно большим.

Рассмотрим теперь эволюцию относительной погрешности в вычислениях.

**Предложение 1.3.3** *Если все слагаемые в сумме имеют одинаковый знак, то относительная погрешность суммы не превосходит наибольшей из относительных погрешностей слагаемых и не является меньшей, чем наименьшая из их относительных погрешностей.*

**Доказательство.** Пусть складываются две приближённые величины, значения которых равны  $\tilde{x}_1$  и  $\tilde{x}_2$ , а относительные погрешности суть  $\delta_1$  и  $\delta_2$ . Тогда их абсолютные погрешности —

$$\Delta_1 = \delta_1 |\tilde{x}_1| \quad \text{и} \quad \Delta_2 = \delta_2 |\tilde{x}_2|.$$

Если  $\delta = \max\{\delta_1, \delta_2\}$ , то  $\Delta_1 \leq \delta |\tilde{x}_1|$  и  $\Delta_2 \leq \delta |\tilde{x}_2|$ . Складывая эти

неравенства почленно, получим

$$\Delta_1 + \Delta_2 \leq \delta(|\tilde{x}_1| + |\tilde{x}_2|),$$

что при ненулевых  $\tilde{x}_1$  и  $\tilde{x}_2$  равносильно

$$\frac{\Delta_1 + \Delta_2}{|\tilde{x}_1| + |\tilde{x}_2|} \leq \delta.$$

В случае, когда слагаемые имеют один и тот же знак, справедливо  $|\tilde{x}_1| + |\tilde{x}_2| = |\tilde{x}_1 + \tilde{x}_2|$ , откуда

$$\frac{\Delta_1 + \Delta_2}{|\tilde{x}_1 + \tilde{x}_2|} \leq \delta.$$

Так как в числителе дроби из левой части стоит предельная абсолютная погрешность суммы, а в знаменателе — абсолютное значение суммы, то полученное неравенство завершает доказательство предложения.

Адаптация проведённых рассуждений для нижней границы относительной погрешности очевидна. ■

Ситуация с относительной погрешностью принципиально меняется, когда в сумме слагаемые имеют разный знак, т. е. наряду со сложением в ней также встречается вычитание. Если результат имеет меньшую абсолютную величину, чем сумма абсолютных величин операндов, то значение дроби (1.2) возрастёт, т. е. относительная погрешность станет больше. А если вычитаемые числа очень близки друг к другу, то знаменатель в (1.2) сделается очень маленьким и относительная погрешность результата вычитания может катастрофически возрасти.

**Пример 1.3.1** Рассмотрим вычитание чисел 1001 и 1000, каждое из которых является приближённым, с абсолютной погрешностью 0.1. Таким образом, относительные погрешности обоих чисел примерно равны 0.01 % (и это довольно высокая точность!).

Выполняя их вычитание, получим результат  $1001 - 1000 = 1$ , который имеет абсолютную погрешность  $0.1 + 0.1 = 0.2$ . Как следствие, относительная погрешность результата стала равной  $0.2/1$ , достигнув 20 %, т. е. увеличилась в 2000 (две тысячи) раз. ■

Отмеченное явление резкого увеличения относительной погрешности при вычитании называют *эффектом потери точности*. Часто используется также термин *эффект потери значащих цифр*, поскольку следствием является уменьшение количества верных значащих цифр в представлении результата.<sup>5</sup> Таким образом, при реализации вычислительных алгоритмов нужно стремиться избегать вычитания близких чисел, заменяя по возможности эту операцию на более безопасные. Например, преобразуя вычислительные формулы так, чтобы малые разности двух величин находились косвенным образом, без вычитания самих этих величин.

**Пример 1.3.2** Предположим, что требуется наиболее точно вычислять значение функции

$$f(t) = \sqrt{t+1} - \sqrt{t}$$

при различных положительных  $t$ . В исходном виде выражение для  $f(t)$  содержит разность чисел, которые по мере увеличения  $t$  становятся всё ближе друг к другу, так как производная квадратного корня уменьшается (рис. 1.5). За счёт этого относительная погрешность результата с ростом  $t$  также увеличивается.

Преобразуем выражение для  $f(t)$ :

$$\begin{aligned} f(t) &= \sqrt{t+1} - \sqrt{t} = \frac{(\sqrt{t+1} - \sqrt{t})(\sqrt{t+1} + \sqrt{t})}{\sqrt{t+1} + \sqrt{t}} = \\ &= \frac{(t+1) - t}{\sqrt{t+1} + \sqrt{t}} = \frac{1}{\sqrt{t+1} + \sqrt{t}}. \end{aligned}$$

В модифицированном выражении уже нет вычитаний, так что условия для проявления эффекта потери точности устранины. Результаты расчётов в арифметике двойной точности по различным выражениям

---

<sup>5</sup> Соответствующий английский термин — loss of significance.

в самом деле демонстрируют заметное различие при больших  $t$ :

$t$	Исходное $f(t)$	Модифицированное $f(t)$
$10^{12}$	$5.00004 \cdot 10^{-7}$	$5.00000 \cdot 10^{-7}$
$10^{13}$	$1.57859 \cdot 10^{-7}$	$1.58114 \cdot 10^{-7}$
$10^{14}$	$5.02914 \cdot 10^{-8}$	$5.00000 \cdot 10^{-8}$
$10^{15}$	$1.86265 \cdot 10^{-8}$	$1.58114 \cdot 10^{-8}$

При  $t = 10^{16}$  и больших значениях исходное выражение для  $f(t)$  за- нуляется, так как значения  $\sqrt{t+1}$  и  $\sqrt{t}$  становятся настолько близкими, что их вычисленные на компьютере значения уже совпадают (детали явления читатель может уяснить из следующего § 1.4). Это не позволяет находить с помощью исходного выражения функции сколько-нибудь адекватную оценку её значений при  $t \geq 10^{16}$ . Но модифицированное выражение продолжает работать и в этом случае. ■

Следующие два результата являются аналогами предложения 1.3.1 для относительных погрешностей операндов.

**Предложение 1.3.4** *Если погрешности приближённых чисел малы, то относительная погрешность их произведения приближённо равна (с точностью до членов более высокого порядка малости) сумме относительных погрешностей сомножителей.*

**Доказательство.** Пусть  $x_1^*, x_2^*, \dots, x_n^*$  — точные значения рассматриваемых чисел,  $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n$  — их приближённые значения. Обозначим также  $z^* := x_1^* x_2^* \cdots x_n^*$  и  $\tilde{z} := \tilde{x}_1 \tilde{x}_2 \cdots \tilde{x}_n$ .

Рассмотрим функцию

$$f(x_1, x_2, \dots, x_n) := x_1 x_2 \cdots x_n$$

— произведение  $x_1, x_2, \dots, x_n$ . Разлагая её в точке  $(x_1^*, x_2^*, \dots, x_n^*)$  по

формуле Тейлора с нулевым и первым членами, получим

$$\begin{aligned}\tilde{z} - z^* &= f(\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n) - f(x_1^*, x_2^*, \dots, x_n^*) \approx \\ &\approx \sum_{i=1}^n \frac{\partial f}{\partial x_i}(x_1^*, x_2^*, \dots, x_n^*) \cdot (\tilde{x}_i - x_i^*) = \\ &= \sum_{i=1}^n x_1^* \cdots x_{i-1}^* x_{i+1}^* \cdots x_n^* (\tilde{x}_i - x_i^*) = \\ &= \sum_{i=1}^n x_1^* x_2^* \cdots x_n^* \frac{\tilde{x}_i - x_i^*}{x_i^*}.\end{aligned}$$

Разделив на  $z^* = x_1^* x_2^* \cdots x_n^*$  обе части этого приближённого равенства и взяв от них абсолютное значение, с точностью до членов второго порядка малости получим

$$\left| \frac{\tilde{z} - z^*}{z^*} \right| \approx \sum_{i=1}^n \left| \frac{\tilde{x}_i - x_i^*}{x_i^*} \right|,$$

что и требовалось. ■

**Предложение 1.3.5** *Если погрешности приближённых чисел малы, то относительная погрешность их частного приближённо (с точностью до членов более высокого порядка малости) равна сумме относительных погрешностей делимого и делителя.*

**Доказательство.** Пусть  $x^*, y^*$  — точные значения рассматриваемых чисел, а  $\tilde{x}, \tilde{y}$  — их приближённые значения. Кроме того, обозначим  $z^* := x^*/y^*$  и  $\tilde{z} := \tilde{x}/\tilde{y}$ . Рассмотрим также функцию двух переменных

$$g(x, y) = x/y$$

— частное чисел  $x$  и  $y$  — и разложим её в точке  $(x^*, y^*)$  по формуле Тейлора с точностью до членов второго порядка:

$$\tilde{z} - z^* \approx \frac{\partial g}{\partial x}(\tilde{x} - x^*) + \frac{\partial g}{\partial y}(\tilde{y} - y^*) = \frac{(\tilde{x} - x^*)}{y^*} - \frac{x^*(\tilde{y} - y^*)}{(y^*)^2}.$$

Поэтому

$$\frac{\tilde{z} - z^*}{z^*} \approx \frac{\tilde{x} - x^*}{y^*} - \frac{x^*(\tilde{y} - y^*)}{(y^*)^2} = \frac{\tilde{x} - x^*}{x^*} - \frac{\tilde{y} - y^*}{y^*},$$

так что в целом

$$\left| \frac{\tilde{z} - z^*}{z^*} \right| \lesssim \left| \frac{\tilde{x} - x^*}{x^*} \right| + \left| \frac{\tilde{y} - y^*}{y^*} \right|.$$

Это и требовалось показать. ■

Мы оценили выше погрешности отдельно взятых арифметических операций. Они дают полезные сведения о поведении погрешностей и позволяют, в первом приближении, ориентироваться при анализе и проектировании вычислительных алгоритмов. Но решение почти любой практической задачи требует большого количества подобных операций, которые складываются в длинные цепочки вычислений. Можно ли применить полученные оценки погрешности к таким более сложным вычислениям?

Ответ на этот вопрос положителен, но лишь отчасти. В длинных цепочках вычислений погрешности отдельных операций могут взаимодействовать друг с другом весьма сложным образом, иногда усиливая, а иногда компенсируя друг друга. Как следствие, непосредственное распространение полученных результатов на вычисление арифметических выражений даёт довольно грубые оценки. Для более точного оценивания погрешностей вычислений и численного решения математических задач в настоящее время широко используют несколько различных подходов, в частности вероятностные модели погрешностей и методы интервального анализа.

## 1.4 Компьютерная арифметика

Для правильного учёта погрешностей реализации вычислительных методов на различных устройствах и для правильной организации этих методов нужно знать детали конкретного способа вычислений. В современных электронных цифровых вычислительных машинах (обозначаемых аббревиатурой ЭВМ), на которых сегодня выполняется подавляющее большинство вычислений, эти детали реализации регламентируются специальными международными стандартами. Первый из них

был принят в 1985 году Институтом инженеров по электротехнике и электронике<sup>6</sup>, профессиональной ассоциацией, объединяющей в своих рядах также специалистов по аппаратному обеспечению ЭВМ. Этот стандарт, коротко называемый IEEE 754, был дополнен и развит в 1995 году следующим стандартом IEEE 854 [19, 33], а затем в 2008 году появилась переработанная версия первого стандарта, которая получила наименование IEEE 754-2008. Работа по обновлению и дальнейшему развитию этих стандартов продолжается и сейчас, но вполне сформировалось некоторое устойчивое ядро, общее для всех этих стандартов, которое не изменится в ближайшем будущем. Его мы кратко рассмотрим в этом разделе.

Согласно стандартам IEEE 754/854 вещественные числа представляются в ЭВМ в виде «чисел с плавающей точкой». Они фактически являются специальной разновидностью чисел в экспоненциальной форме, которые записываются в виде произведения некоторого нормализованного множителя, называемого *мантиссой*, на степень основания системы счисления. Более точно, зафиксируем натуральные числа  $\beta$  и  $p$ . Числами с плавающей точкой по основанию  $\beta$  называются числа вида

$$\pm(\alpha_0 + \alpha_1\beta^{-1} + \alpha_2\beta^{-2} + \dots + \alpha_{p-1}\beta^{-(p-1)}) \cdot \beta^t, \quad (1.7)$$

где  $0 \leq \alpha_i < \beta$ ,  $i = 0, 1, \dots, p - 1$ . Обычно эти числа обозначают условной записью

$$\alpha_0.\alpha_1\alpha_2 \dots \alpha_{p-1} \cdot \beta^t.$$

На показатель степени  $t$  накладывается двустороннее ограничение

$$t_{\min} \leq t \leq t_{\max},$$

а величина  $p$ , отвечающая за количество значащих цифр мантиссы, — это *точность* или *разрядность* рассматриваемой числовой модели с плавающей точкой. Отметим, что термин «мантиssa» является общеупотребительным и имеет давнюю историю, но в текстах стандартов IEEE 754/854 множитель  $\alpha_0.\alpha_1\alpha_2 \dots \alpha_{p-1}$  называется  *significand* (который можно перевести как «значимое»).

Стандарты IEEE 754/854 предписывают для цифровых ЭВМ значения  $\beta = 2$  или  $\beta = 10$ , и в большинстве компьютеров используется  $\beta = 2$ , т. е. двоичная система счисления. С одной стороны, это вызвано

---

<sup>6</sup>Обычно его называют английским сокращением IEEE от Institute of Electrical and Electronics Engineers.

особенностями физической реализации современных ЭВМ, где 0 соответствует отсутствию сигнала (заряда и т. п.), а 1 — его наличию. С другой стороны, двоичная система оказывается реально выгодной при выполнении с ней приближённых вычислений [19].



Рис. 1.2. Множество чисел, представимых в цифровой ЭВМ  
— дискретное конечное подмножество вещественной оси  $\mathbb{R}$

Представление вещественного числа в виде с плавающей точкой, как правило, неединственно. Например,  $1.234 \cdot 10^5 = 0.1234 \cdot 10^6$  и т. д. Это безобидное, на первый взгляд, явление вызывает существенные неудобства реализации, и потому на вид чисел с плавающей точкой (1.7) часто накладывают ограничение  $\alpha_0 \neq 0$ . Удовлетворяющие этому условию числа называют *нормализованными* числами с плавающей точкой. Представление вещественного числа в нормализованном виде уже единственно, и именно такие числа главным образом используются в вычислениях по стандартам IEEE 754/854.

Если в выражении (1.7) зафиксировать показатель  $t$ , то, варьируя все коэффициенты  $\alpha_i$  в предписанных им пределах, получим дискретное множество чисел на вещественной оси, в котором расстояние между соседними числами постоянно и равно  $\beta^{-(p-1)} \cdot \beta^t$ . Для другого значения показателя  $t$  будет то же самое, с другим постоянным расстоянием между машинно представимыми числами. Таким образом, множество всех чисел с плавающей точкой является объединением равномерных участков, покрывающих более или менее обширную часть вещественной оси  $\mathbb{R}$ . Оно симметрично относительно нуля.

Для чисел с плавающей точкой стандарты IEEE 754/854 предусматривают «одинарную точность» и «двойную точность», а также «расширенные» варианты этих представлений. При этом для хранения чисел одинарной точности отводится 4 байта памяти ЭВМ, для двойной точности — 8 байтов. Из этих 32 или 64 битов один бит зарезервирован для указания знака числа: 0 соответствует «-», а 1 соответствует «+».<sup>7</sup>

Для одинарной точности (которая обозначается, как правило, ключевым словом «single») на показатель степени  $t$  отводится 8 битов па-

<sup>7</sup>Таким образом, во внутреннем «машинном» представлении знак присутствует у любого числа, в том числе и у нуля.

мяти и полагается  $t_{\min} = -126$ ,  $t_{\max} = 127$ . Для двойной точности, наиболее широко распространённой в современных расчётах (она обозначается ключевым словом «double»), на показатель степени  $t$  отводится 11 битов памяти и полагается  $t_{\min} = -1022$ ,  $t_{\max} = 1023$ .

На мантиссу чисел одинарной и двойной точности стандарты IEEE 754/854 отводят 23 бита и 52 бита соответственно. Но реально это соответствует разрядностям  $p = 24$  и  $p = 53$ , так как для представления чисел (1.7) кроме явно выделенных битов ещё неявно используется так называемый «скрытый бит». Дело в том, что для двоичной системы счисления условие нормализации  $\alpha_0 \neq 0$  необходимо влечёт  $\alpha_0 = 1$ . Поэтому соответствующий бит, постоянно равный единице, можно вообще не хранить в компьютерном представлении числа, используя как некоторую константу, присутствующую по умолчанию.

Как следствие, диапазон чисел одинарной точности, представимых в ЭВМ, простирается по абсолютной величине примерно от  $1.18 \cdot 10^{-38}$  до  $3.4 \cdot 10^{38}$ . Диапазон чисел двойной точности, представимых в ЭВМ, гораздо более широк, и по абсолютной величине охватывает числа примерно от  $2.22 \cdot 10^{-308}$  до  $1.79 \cdot 10^{308}$ . Как видим, числа одинарной точности могут быть недостаточны для современного математического моделирования, где даже значения некоторых физических констант приближаются к пределам представимости на ЭВМ. Модель чисел с плавающей точкой двойной точности обеспечивает 16 десятичных значащих цифр в представлении числа и вполне удовлетворяет большинство научно-технических расчётов. В целом числа с плавающей точкой обеспечивают практически постоянную относительную погрешность представления вещественных чисел и изменяющуюся абсолютную погрешность.

Переход от множества вещественных чисел из выписанных выше диапазонов к машинно представимым числам выполняется с помощью операции, называемой *округлением*. Оно может выполняться различными способами, и по умолчанию в компьютерных системах обычно установлен режим округления «к ближайшему». Это означает, что вещественное число, которое не представляется точно в ЭВМ, заменяется на ближайшее к нему машинно представимое число заданного формата. Но для решения специфических задач можно также установить специальными командами режимы округления «к  $-\infty$ », «к  $+\infty$ » или «к нулю».

Количество различных показателей степени  $t$ , равное 2046 для двойной точности, не исчерпывает всех возможных  $2^{11} = 2048$  целых чи-

сем, которые можно закодировать 11 битами. Аналогична ситуация и с одинарной точностью. Оставшиеся значения показателей  $t$ , которые не входят в целочисленный интервал  $[t_{\min}, t_{\max}]$ , стандарты IEEE 754/854 предназначают для кодирования некоторых специальных объектов, которые могут участвовать в вычислениях или быть их результатами. Прежде всего это нуль, который нельзя представить среди нормализованных чисел с плавающей точкой. Он поэтому кодируется специальным образом, как  $1.0 \cdot \beta^{t_{\min}-1}$ , т. е. в виде числа со всеми нулями в мантиссе (кроме скрытого бита) и показателем степени за пределами интервала  $[t_{\min}, t_{\max}]$ . Кроме того, стандарты IEEE 754/854 вводят «машинную бесконечность» и особый нечисловой объект под названием NaN (имя которого есть аббревиатура английской фразы «Not a Number»).

Машинная бесконечность, обычно обозначаемая Inf, обладает свойствами математической бесконечности  $\infty$ :

$$\text{Inf} \pm a = \text{Inf},$$

$$\text{Inf} + \text{Inf} = \text{Inf},$$

$$\text{Inf} \cdot a = \text{Inf} \text{ для } a > 0,$$

$$\text{Inf} \cdot a = -\text{Inf} \text{ для } a < 0.$$

Она необходима для сигнализации о том, что результат вычислений вышел за пределы машинно представимых чисел, и потому относиться к нему нужно по-особому.<sup>8</sup>

NaN означает, что результату операции невозможно придать какой-либо смысл вообще. Он полезен во многих ситуациях, в частности, может использоваться для сигнализации о нетипичных и исключительных событиях в процессе вычислений, неопределённых результатах и т. п. Например,

$$\text{Inf} - \text{Inf} = \text{NaN},$$

$$\text{Inf} \cdot 0 = \text{NaN},$$

$$0/0 = \text{NaN},$$

$$\text{Inf}/\text{Inf} = \text{NaN}.$$

Результатом любой операции с NaN также является NaN. Машинная бесконечность и NaN представляются в компьютере последовательностями битов, которые соответствуют показателям степени  $t_{\max} + 1$ .

Очень важной характеристикой множества машинных чисел является так называемое «машинное  $\varepsilon$ » (машинное эпсилон), которое характеризует расстояние между соседними машинно представимыми числами. Можно сказать, что это величина, обратная густоте (плотности)

---

<sup>8</sup>Обозначение машинной бесконечности взято от латинского слова infinitum. Не следует путать его с точной нижней гранью множества, обозначаемой «inf».

множества машинно-представимых чисел. Более точно, машинное  $\varepsilon$  — это наибольшее положительное число  $\varepsilon_{\text{маш}}$ , для которого в компьютерной арифметике  $1 + \varepsilon_{\text{маш}} = 1$  при округлении «к ближайшему». Из конструкции чисел с плавающей точкой следует тогда, что компьютер, грубо говоря, не будет различать чисел  $a$  и  $b$ , удовлетворяющих условию  $1 < a/b < 1 + \varepsilon_{\text{маш}}$ . Для двойной точности представления в стандарте IEEE 754/854 «машинальное эпсилон» примерно равно  $1.11 \cdot 10^{-16}$ .

Удвоенное «машинальное эпсилон» — это расстояние между соседними машинно-представимыми числами в районе единицы, справа от неё. В других местах вещественной оси это расстояние будет другим, но его можно легко найти из значения  $2\varepsilon_{\text{маш}}$  с помощью домножения на необходимый масштабирующий множитель, который является степенью  $\beta$ . Это вытекает из того отмеченного выше факта, что машинно-представимые числа расположены на вещественной оси равномерными участками.

Принципиальная особенность компьютерной арифметики, вызванная дискретностью множества машинных чисел и наличием округлений — невыполнение некоторых общезвестных свойств вещественной арифметики. Например, сложение чисел с плавающей точкой неассоциативно, т. е. в общем случае неверно, что

$$(a + b) + c = a + (b + c).$$

Читатель может проверить на любом компьютере, что в арифметике IEEE 754/854 двойной точности при округлении «к ближайшему»

$$(1 + 10^{-16}) + 10^{-16} \neq 1 + (10^{-16} + 10^{-16}).$$

Левая часть этого соотношения равна 1, тогда как правая — ближайшему к единице справа машинно-представимому числу. Аналогичная ситуация имеет место в любых приближённых вычислениях, которые сопровождаются округлениями, а не только при расчётах на современных цифровых ЭВМ.

Ещё один аналогичный по духу пример отсутствия ассоциативности в компьютерной арифметике

$$(10^{20} - 10^{20}) + 1 \neq 10^{20} + (-10^{20} + 1).$$

Из отсутствия ассоциативности следует, что результат суммирования длинных сумм вида  $x_1 + x_2 + \dots + x_n$  зависит от порядка, в котором

выполняется попарное суммирование слагаемых, или, как говорят, от расстановки скобок в сумме. Каким образом следует организовывать такое суммирование в компьютерной арифметике, чтобы получать наиболее точные результаты? Ответ на этот вопрос существенно зависит от значений слагаемых, но в случае суммирования уменьшающихся по абсолютной величине чисел суммировать их нужно «с конца». Именно так, к примеру, целесообразно находить суммы большинства рядов.

## 1.5 Интервальная арифметика

Исходной идеей создания интервальной арифметики является наблюдение о том, что почти всё в нашем мире неточно. В реальности нам чаще всего приходится работать не с точными значениями величин, которые образуют основу классической «точной» математики, а с целыми диапазонами значений тех или иных величин. Например, множество вещественных чисел, которые точно представляются в цифровых ЭВМ, является конечным, и каждое из этих чисел из-за присутствия округления в действительности служит представителем интервала значений обычной вещественной оси  $\mathbb{R}$  (рис. 1.7 и 1.8).

Нельзя ли организовать операции и отношения между диапазонами-интервалами так, как это сделано для обычных точных значений? Чтобы можно было работать с ними, как с обычным числами, опираясь на алгебраические преобразования, аналитические операции и т. п.? Результатом таких вычислений с диапазонами-интервалами станут оценки изменения интересующих нас величин, т. е. очень ценная и востребованная на практике информация. Ответы на поставленные вопросы в целом положительны, хотя и не столь просты, а свойства получающейся «интервальной арифметики» оказываются непохожими на привычные свойства операций с обычными числами. Дальнейшие исследования в этом направлении привели к появлению и развитию интервального анализа — одной из плодотворных ветвей современной вычислительной математики [1, 15, 16, 21, 35].

*Интервалом*  $[a, b]$  вещественной оси  $\mathbb{R}$  называем множество всех чисел, расположенных между заданными числами  $a$  и  $b$ , включая их самих, т. е.

$$[a, b] := \{ x \in \mathbb{R} \mid a \leq x \leq b \}.$$

При этом  $a$  и  $b$  называются *концами* интервала  $[a, b]$ , левым и правым соответственно, а множество всех интервалов обозначается символом

$\mathbb{IR}$ . В противоположность интервалам и интервальным величинам будем называть *точечными* те величины, значениями которых являются отдельные точки — вещественной оси, плоскости или, более общо, какого-либо пространства. Помимо замкнутых интервалов существуют также полуоткрытые и открытые интервалы, которым не принадлежат один или оба из их концов. Они обозначаются  $]a, b]$ ,  $[a, b[$  и  $]a, b[$  соответственно.

Далее будем обозначать интервалы и составленные из них объекты буквами жирного шрифта:  $a$ ,  $b$ ,  $c$ , …,  $x$ ,  $y$ ,  $z$ . Подчёркивание и надчёркивание —  $\underline{a}$  и  $\overline{a}$  — означают взятие нижнего и верхнего концов интервала, так что  $a = [\underline{a}, \overline{a}]$ .

Всякий интервал полностью задаётся двумя своими концами, но для его всестороннего описания очень важны и другие характеристики. Главными из них являются *середина* и *ширина* интервала:

$$\text{mid } a = \frac{1}{2}(\underline{a} + \overline{a}) \quad \text{— середина (центр),}$$

$$\text{wid } a = \overline{a} - \underline{a} \quad \text{— ширина.}$$

Середина интервала является его точкой, наименее удалённой от всех остальных точек интервала, так что при прочих равных условиях середину можно брать в качестве «наиболее представительной» точки из интервала. Ширина интервала характеризует его абсолютный размах, количественную меру неопределённости значений, представляющей этим интервалом.

Интервалы являются множествами, так что для них определены теоретико-множественные операции и отношения. В частности, очень большую роль играет в интервальном анализе отношение включения вещественных интервалов, одного в другой:

$$a \subseteq b \iff \underline{a} \geq \underline{b} \text{ и } \overline{a} \leq \overline{b}. \quad (1.8)$$

Во многих конструкциях необходима *внутренность* интервала  $a$ , которая обозначается  $\text{int } a$  и определяется как открытый интервал  $\text{int } a := [\underline{a}, \overline{a}[$  (см. также § 3.36).

Предположим, что нам даны переменные  $a$  и  $b$ , точные значения которых неизвестны, но мы знаем, что они могут находиться в интервалах  $a = [\underline{a}, \overline{a}]$  и  $b = [\underline{b}, \overline{b}]$ . Что можно сказать о значении суммы  $a + b$ ?

Складывая почленно неравенства

$$\begin{aligned}\underline{a} &\leq a \leq \bar{a}, \\ \underline{b} &\leq b \leq \bar{b},\end{aligned}$$

получим

$$\underline{a} + \underline{b} \leq a + b \leq \bar{a} + \bar{b},$$

так что  $a + b \in [\underline{a} + \underline{b}, \bar{a} + \bar{b}]$ . Концы получившегося интервала, очевидно, достигаются, если слагаемые  $a$  и  $b$  могут независимо друг от друга принимать значения в пределах своих интервалов  $\mathbf{a}$  и  $\mathbf{b}$ .

На аналогичный вопрос, связанный с областью значений разности  $a - b$ , можно ответить, складывая почленно неравенства

$$\begin{aligned}\underline{a} &\leq a \leq \bar{a}, \\ -\bar{b} &\leq -b \leq -\underline{b}.\end{aligned}$$

Имеем в результате  $a - b \in [\underline{a} - \bar{b}, \bar{a} - \underline{b}]$ , причём концы этого интервала достигаются на тех же условиях, что и для суммы.

Для умножения двух переменных  $a \in \mathbf{a}$  и  $b \in \mathbf{b}$  имеет место несколько более сложная оценка

$$a \cdot b \in [\min \{\underline{a}\underline{b}, \underline{a}\bar{b}, \bar{a}\underline{b}, \bar{a}\bar{b}\}, \max \{\underline{a}\underline{b}, \underline{a}\bar{b}, \bar{a}\underline{b}, \bar{a}\bar{b}\}].$$

Чтобы доказать её, заметим, что функция  $\phi : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ , задаваемая правилом  $\phi(a, b) = a \cdot b$ , является линейной по  $b$  при каждом фиксированном  $a$ . Поэтому она принимает минимальное и максимальное значения на концах интервала изменения переменной  $b$ . Это же верно и для экстремумов по  $a \in \mathbf{a}$  при любом фиксированном значении  $b$ . Наконец,

$$\min_{a \in \mathbf{a}, b \in \mathbf{b}} \phi(a, b) = \min_{a \in \mathbf{a}} \min_{b \in \mathbf{b}} \phi(a, b),$$

$$\max_{a \in \mathbf{a}, b \in \mathbf{b}} \phi(a, b) = \max_{a \in \mathbf{a}} \max_{b \in \mathbf{b}} \phi(a, b),$$

т. е. взятие минимума по совокупности аргументов может быть заменено повторным минимумом, а взятие максимума по совокупности аргументов — повторным максимумом, причём в обоих случаях порядок экстремумов несуществен. Следовательно, для  $a \in \mathbf{a}$  и  $b \in \mathbf{b}$  в самом деле

$$\min \{\underline{a}\underline{b}, \underline{a}\bar{b}, \bar{a}\underline{b}, \bar{a}\bar{b}\} \leq a \cdot b \leq \max \{\underline{a}\underline{b}, \underline{a}\bar{b}, \bar{a}\underline{b}, \bar{a}\bar{b}\}, \quad (1.9)$$

и нетрудно видеть, что эта оценка достижима с обеих сторон, если сомножители  $a$  и  $b$  могут независимо друг от друга принимать значения в пределах своих интервалов  $\mathbf{a}$  и  $\mathbf{b}$ .

Наконец, для частного двух ограниченных переменных несложно вывести оценки из неравенств для умножения и из того факта, что  $a/b = a \cdot (1/b)$ .

Проведённые выше рассуждения подсказывают идею — рассматривать интервалы вещественной оси как самостоятельные объекты, между которыми вводятся свои собственные операции, отношения и т. п. На множестве всех вещественных интервалов арифметические операции — сложение, вычитание, умножение и деление — определим «по представителям», т. е. в соответствии со следующим правилом:

$$\mathbf{a} \star \mathbf{b} := \{ a \star b \mid a \in \mathbf{a}, b \in \mathbf{b} \} \quad (1.10)$$

для всех интервалов  $\mathbf{a}$ ,  $\mathbf{b}$ , таких что выполнение точечной операции  $a \star b$ ,  $\star \in \{+, -, \cdot, /\}$ , имеет смысл для любых  $a \in \mathbf{a}$  и  $b \in \mathbf{b}$ . При этом вещественные числа  $a$  отождествляются с интервалами нулевой ширины  $[a, a]$ , которые называются также *вырожденными интервалами*. Кроме того, через  $(-\mathbf{a})$  условимся обозначать интервал  $(-1) \cdot \mathbf{a}$ .

Как показано выше, развернутое определение интервальных арифметических операций, равносильное (1.10), задаётся следующими формулами:

$$\mathbf{a} + \mathbf{b} = [\underline{\mathbf{a}} + \underline{\mathbf{b}}, \bar{\mathbf{a}} + \bar{\mathbf{b}}], \quad (1.11)$$

$$\mathbf{a} - \mathbf{b} = [\underline{\mathbf{a}} - \bar{\mathbf{b}}, \bar{\mathbf{a}} - \underline{\mathbf{b}}], \quad (1.12)$$

$$\mathbf{a} \cdot \mathbf{b} = [\min\{\underline{\mathbf{a}}\underline{\mathbf{b}}, \underline{\mathbf{a}}\bar{\mathbf{b}}, \bar{\mathbf{a}}\underline{\mathbf{b}}, \bar{\mathbf{a}}\bar{\mathbf{b}}\}, \max\{\underline{\mathbf{a}}\underline{\mathbf{b}}, \underline{\mathbf{a}}\bar{\mathbf{b}}, \bar{\mathbf{a}}\underline{\mathbf{b}}, \bar{\mathbf{a}}\bar{\mathbf{b}}\}], \quad (1.13)$$

$$\mathbf{a}/\mathbf{b} = \mathbf{a} \cdot [1/\bar{\mathbf{b}}, 1/\underline{\mathbf{b}}] \quad \text{для } \mathbf{b} \not\equiv 0. \quad (1.14)$$

В частности, при умножении интервала на число полезно помнить следующее простое правило:

$$\mu \cdot \mathbf{a} = \begin{cases} [\mu \underline{\mathbf{a}}, \mu \bar{\mathbf{a}}], & \text{если } \mu \geq 0, \\ [\mu \bar{\mathbf{a}}, \mu \underline{\mathbf{a}}], & \text{если } \mu \leq 0. \end{cases} \quad (1.15)$$

**Пример 1.5.1** Наиболее сложной интервальной арифметической операцией является умножение, особенно на интервал, содержащий нуль:

$$[-1, 2] \cdot [3, 4] = [-4, 8], \text{ но и } [-1, 2] \cdot [0, 4] = [-4, 8].$$

Хорошо видно, что по своим свойствам такое умножение существенно отличается от привычного умножения вещественных чисел, хотя для неотрицательных интервалов умножение распадается «по концам». ■

Алгебраическая система  $\langle \mathbb{IR}, +, -, \cdot, / \rangle$ , образованная множеством всех вещественных интервалов  $a := [\underline{a}, \bar{a}] = \{x \in \mathbb{R} \mid \underline{a} \leq x \leq \bar{a}\}$  с бинарными операциями сложения, вычитания, умножения и деления, которые определены формулами (1.11)–(1.14), называется *классической интервальной арифметикой*. Дополнительное определение «классическая» используется здесь потому, что существуют и другие интервальные арифметики, приспособленные для решения других задач.

Полезно выписать определение интервального умножения в виде так называемой таблицы Кэли, дающей представление результата операции в зависимости от различных комбинаций значений operandов. Для этого выделим в  $\mathbb{IR}$  следующие подмножества:

$$\mathcal{P} := \{a \in \mathbb{IR} \mid \underline{a} \geq 0 \text{ и } \bar{a} \geq 0\} \quad \text{— неотрицательные интервалы,}$$

$$\mathcal{Z} := \{a \in \mathbb{IR} \mid \underline{a} \leq 0 \leq \bar{a}\} \quad \text{— нульсодержащие интервалы,}$$

$$-\mathcal{P} := \{a \in \mathbb{IR} \mid -a \in \mathcal{P}\} \quad \text{— неположительные интервалы.}$$

В целом  $\mathbb{IR} = \mathcal{P} \cup \mathcal{Z} \cup (-\mathcal{P})$ . Тогда интервальное умножение (1.13) может быть описано с помощью табл. 1.1, особенно удобной при реализации этой операции на ЭВМ.

Таблица 1.1. Интервальное умножение

.	$b \in \mathcal{P}$	$b \in \mathcal{Z}$	$b \in -\mathcal{P}$
$a \in \mathcal{P}$	$[\underline{a}\underline{b}, \bar{a}\bar{b}]$	$[\bar{a}\underline{b}, \bar{a}\bar{b}]$	$[\bar{a}\underline{b}, \underline{a}\bar{b}]$
$a \in \mathcal{Z}$	$[\underline{a}\bar{b}, \bar{a}\bar{b}]$	$[\min\{\underline{a}\bar{b}, \bar{a}\underline{b}\}, \max\{\underline{a}\bar{b}, \bar{a}\bar{b}\}]$	$[\bar{a}\underline{b}, \underline{a}\bar{b}]$
$a \in -\mathcal{P}$	$[\underline{a}\bar{b}, \bar{a}\underline{b}]$	$[\underline{a}\bar{b}, \underline{a}\bar{b}]$	$[\bar{a}\bar{b}, \underline{a}\bar{b}]$

Именно по этой таблице реализовано интервальное умножение в подавляющем большинстве компьютерных систем, поддерживающих интервальную арифметику, так как в сравнении с исходными формулами такая реализация — значительно более быстрая.

Алгебраические свойства классической интервальной арифметики существенно беднее, чем у поля вещественных чисел  $\mathbb{R}$ . В частности, особенностью интервальной арифметики является отсутствие дистрибутивности умножения относительно сложения: в общем случае

$$a(b + c) \neq ab + ac.$$

Например,

$$[1, 2] \cdot (1 - 1) = 0 \neq [-1, 1] = [1, 2] \cdot 1 - [1, 2] \cdot 1.$$

Тем не менее имеет место более слабое свойство

$$a(b + c) \subseteq ab + ac, \quad (1.16)$$

называемое *субдистрибутивностью* умножения относительно сложения. В ряде частных случаев дистрибутивность всё-таки выполняется:

$$a(b + c) = ab + ac, \quad \text{если } a \text{ — вещественное число,} \quad (1.17)$$

$$a(b + c) = ab + ac, \quad \text{если } b, c \geq 0 \text{ или } b, c \leq 0. \quad (1.18)$$

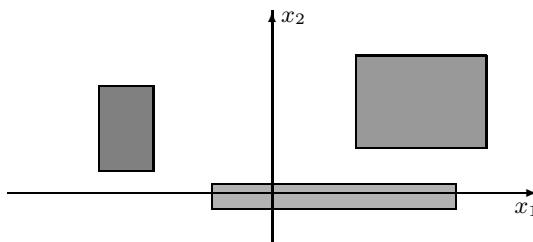


Рис. 1.3. Интервальные векторы-брюсы в  $\mathbb{R}^2$

Распространение интервальных конструкций на многомерный случай может быть выполнено несколькими возможными способами, и целесообразность выбора какого-то конкретного из них определяется

постановкой решаемой задачи, видом исходных данных, требованиями к ответу и т. п. Наиболее популярное определение заключается в том, что интервальный вектор — это прямое произведение интервалов по отдельным компонентам или, иными словами, упорядоченный кортеж из интервалов, расположенный вертикально (вектор-столбец) или горизонтально (вектор-строка). Таким образом, если  $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$  — некоторые интервалы, то

$$\mathbf{a} = \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{pmatrix} \text{ — интервальный вектор-столбец,}$$

а

$$\mathbf{a} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n) \text{ — интервальная вектор-строка.}$$

Интервальные векторы также называют *брусами*, имея в виду их геометрическую интерпретацию (рис. 1.3).

Множество интервальных векторов, компоненты которых принадлежат  $\mathbb{IR}$ , будем обозначать через  $\mathbb{IR}^n$ . При этом нулевые векторы, т. е. такие, все компоненты которых суть нули, традиционно обозначаются как « $\langle 0 \rangle$ ».

Интервальная матрица — матрица с интервальными элементами,

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_{11} & \mathbf{a}_{12} & \cdots & \mathbf{a}_{1n} \\ \mathbf{a}_{21} & \mathbf{a}_{22} & \cdots & \mathbf{a}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}_{m1} & \mathbf{a}_{m1} & \cdots & \mathbf{a}_{mn} \end{pmatrix},$$

где  $\mathbf{a}_{ij} \in \mathbb{IR}$ . Множество всех интервальных  $m \times n$ -матриц обычно обозначают  $\mathbb{IR}^{m \times n}$ . Интервальную матрицу можно рассматривать как множество всевозможных точечных матриц той же структуры, элементы которых принадлежат соответствующим интервальным элементам из интервальной матрицы.

Понятия середины и ширины распространяются на интервальные векторы (брuses) и матрицы покомпонентным и поэлементным образом, так что если, к примеру,  $\mathbf{a} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n)$ , то

$$\text{mid } \mathbf{a} = (\text{mid } \mathbf{a}_1, \text{mid } \mathbf{a}_2, \dots, \text{mid } \mathbf{a}_n),$$

$$\text{wid } \mathbf{a} = (\text{wid } \mathbf{a}_1, \text{wid } \mathbf{a}_2, \dots, \text{wid } \mathbf{a}_n).$$

Введём важное понятие интервальной оболочки множества. Если  $S$  — непустое ограниченное множество в  $\mathbb{R}^n$  или  $\mathbb{R}^{m \times n}$ , то его *интервальной оболочкой*  $\square S$  называется наименьший по включению интервальный вектор (или матрица), содержащий  $S$ . Легко понять, что это определение равносильно такому: интервальная оболочка множества  $S$  — это пересечение всех интервальных векторов, содержащих  $S$ , т. е.

$$\square S = \cap \{ \mathbf{a} \in \mathbb{IR}^n \mid \mathbf{a} \supseteq S \}.$$

Интервальная оболочка — это интервальный объект, наилучшим образом приближающий извне (т. е. объемлющий) рассматриваемое множество, и компоненты  $\square S$  являются проекциями множества  $S$  на координатные оси пространства.

Сумма (разность) двух интервальных матриц одинакового размера определяется как интервальная матрица того же размера, образованная поэлементными суммами (разностями) операндов. Если  $\mathbf{A} = (\mathbf{a}_{ij}) \in \mathbb{IR}^{m \times l}$  и  $\mathbf{B} = (\mathbf{b}_{ij}) \in \mathbb{IR}^{l \times n}$ , то произведение матриц  $\mathbf{A}$  и  $\mathbf{B}$  есть матрица  $\mathbf{C} = (\mathbf{c}_{ij}) \in \mathbb{IR}^{m \times n}$ , такая что

$$\mathbf{c}_{ij} := \sum_{k=1}^l \mathbf{a}_{ik} \mathbf{b}_{kj}.$$

Нетрудно показать, что для операций между матрицами выполняется соотношение

$$\mathbf{A} \star \mathbf{B} = \square \{ A \star B \mid A \in \mathbf{A}, B \in \mathbf{B} \}, \quad \star \in \{ +, -, \cdot \}, \quad (1.19)$$

и обоснование этого свойства можно увидеть в книгах [16, 22]. Точное равенство результата интервального матричного умножения и множества произведений «по представителям» невозможно, как показывают простейшие примеры, поскольку множество в правой части может не быть интервальной матрицей. По этой причине равенство (1.19) — это наибольшее, что можно получить в рассматриваемой ситуации.

Важнейший частный случай, когда достигается равенство результата интервального умножения и множества результатов умножений представителей, — это произведение интервальной матрицы на точечный вектор:

$$\mathbf{Ab} = \{ Ab \mid A \in \mathbf{A} \}. \quad (1.20)$$

Интервальная арифметика даёт алгоритмизированный способ оперирования с диапазонами значений и абсолютными погрешностями, и мы

уже рассматривали аналогичную технику в § 1.3. Но формулы интервальной арифметики, будучи хорошо приспособленными для реализации на ЭВМ, всё-таки менее удобны для теоретического анализа, чем результаты предложений 1.3.1–1.3.5.

## 1.6 Интервальные расширения функций

Пусть  $f : \mathbb{R} \rightarrow \mathbb{R}$  — некоторая функция. Если интервалы рассматриваются в виде самостоятельных объектов, то что следует понимать под значением функции от интервала? Естественно считать, что

$$f(\mathbf{x}) = \{f(x) \mid x \in \mathbf{x}\},$$

т. е. что значение функции для интервального аргумента — это множество всевозможных значений этой функции в точках из данного интервала. Оно само является интервалом, если функция непрерывна.

Задача об определении области значений функции на том или ином подмножестве области её определения, эквивалентная задаче оптимизации, в интервальном анализе принимает специфическую форму задачи о вычислении интервальных оценивающих функций и, в частности, интервального расширения функции.

**Определение 1.6.1** Пусть  $D$  — непустое подмножество пространства  $\mathbb{R}^n$ . Интервальная функция  $\mathbf{f} : \mathbb{I}D \rightarrow \mathbb{IR}^m$  называется интервальным продолжением точечной функции  $f : D \rightarrow \mathbb{R}^m$ , если  $\mathbf{f}(x) = f(x)$  для всех  $x \in D$ .

**Определение 1.6.2** Интервальная функция  $\mathbf{f} : \mathbb{I}D \rightarrow \mathbb{IR}^m$  называется внешней оценивающей функцией для точечной функции  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  на множестве  $D \subset \mathbb{R}^n$ , если для любого интервала  $\mathbf{X} \subseteq D$  имеет место

$$x \in \mathbf{X} \Rightarrow f(x) \in \mathbf{f}(\mathbf{X}),$$

или, что равносильно,

$$\{f(x) \mid x \in \mathbf{X}\} \subseteq \mathbf{f}(\mathbf{X}) \quad \text{для всех } \mathbf{X} \in \mathbb{I}D.$$

Если смысл оценивания ясен (внешнее оценивание), то можно говорить также — интервальная оценивающая функция или просто оценивающая функция.

**Определение 1.6.3** Интервальная функция  $f : \mathbb{IR}^n \rightarrow \mathbb{IR}^m$  называется интервальным расширением точечной функции  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  на множестве  $D \subset \mathbb{R}^n$ , если она является

- (i) интервальным продолжением для  $f$  на  $D$ ,
- (ii) внешней оценивающей функцией для  $f$  на  $D$  (рис. 1.4).

Одним из важнейших свойств интервальных функций является *моноотонность по включению*, когда для любых интервалов  $\mathbf{x}, \mathbf{y}$  из их области определения имеет место импликация

$$\mathbf{x} \subseteq \mathbf{y} \Rightarrow f(\mathbf{x}) \subseteq f(\mathbf{y}).$$

Монотонными по включению являются, например, все интервальные выражения, сконструированные из интервальных переменных с помощью правил интервальной арифметики. Нетрудно понять, что монотонное по включению интервальное продолжение точечной функции является её интервальным расширением: тогда для всякого интервала  $\mathbf{X}$  из принадлежности  $x \in \mathbf{X}$  следует

$$f(x) = f(x) \in f(\mathbf{X}).$$

Но в общем случае интервальные расширения не обязательно монотонны по включению. Эффективное построение интервальных расширений функций — это важнейшая задача интервального анализа, поиски различных решений которой продолжаются и в настоящее время. Приведём в рамках нашего беглого обзора некоторые общезначимые результаты в этом направлении.

Функцию, выражение для которой является конечной комбинацией символов переменных, констант и четырёх арифметических операций (сложения, вычитания, умножения и деления) будем называть *рациональной функцией*.

**Теорема 1.6.1** Если для рациональной функции  $f(x) = f(x_1, x_2, \dots, x_n)$  на брусе  $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathbb{IR}^n$  определён результат  $f_{\natural}(\mathbf{x})$  подстановки вместо её аргументов интервалов их изменения  $x_1, x_2, \dots, x_n$  и выполнения всех действий над ними по правилам интервальной арифметики, то

$$\{f(x) \mid x \in \mathbf{x}\} \subseteq f_{\natural}(\mathbf{x}),$$

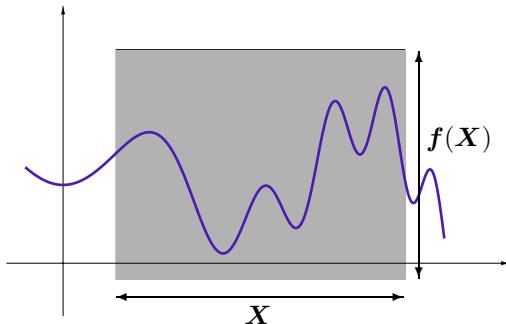


Рис. 1.4. Интервальное расширение функции даёт внешнюю оценку её области значений

*т. е.  $f_i(x)$  содержит множество значений функции  $f(x)$  на  $x$ . Если выражение для  $f(x)$  содержит не более одного входления каждой переменной в первой степени, то выписанное включение становится равенством.*

Первая часть утверждения теоремы вытекает из того, что  $f_i(x)$  — монотонное по включению интервальное продолжение вещественной функции  $f(x)$  (см. рассуждения на предыдущей странице). Развернутые доказательства теоремы можно увидеть, например, в [1, 15, 16, 21, 22], но корректное обоснование её второй части требует привлечения понятия зависимости/независимости интервальных величин, которое в необходимой полноте рассмотрено только в [16].

Итак, по отношению к рациональной функции  $f(x)$  интервальная функция  $f_i(x)$ , о которой идёт речь в теореме 1.6.1, является интервальным расширением. Оно называется *естественным интервальным расширением* и вычисляется совершенно элементарно.

Почему в основной теореме интервальной арифметики в общем случае утверждается включение области значений в результат естественного интервального расширения функции? Это вызывается так называемым эффектом зависимости, при котором интервальные величины, представляющие результаты оценивания промежуточных переменных в вычислениях, становятся зависимыми друг от друга (связанными), а потому в операциях классической интервальной арифметики концы результирующих интервалов уже могут не достигаться значениями функции. Читатель может увидеть подробный разбор этих вопросов в

книге [16].

**Пример 1.6.1** Для функции  $f(x) = x/(x+1)$  на интервале  $[1, 3]$  область значений в соответствии с результатом теоремы 1.6.1 можно оценить извне как

$$\frac{[1, 3]}{[1, 3] + 1} = \frac{[1, 3]}{[2, 4]} = \left[\frac{1}{4}, \frac{3}{2}\right]. \quad (1.21)$$

Но если предварительно переписать выражение для функции в виде

$$f(x) = \frac{1}{1 + 1/x},$$

разделив числитель и знаменатель дроби на  $x \neq 0$ , то интервальное оценивание даст уже результат

$$\frac{1}{1 + 1/[1, 3]} = \frac{1}{\left[\frac{4}{3}, 2\right]} = \left[\frac{1}{2}, \frac{3}{4}\right].$$

Он более узок (т. е. более точен), чем (1.21) и совпадает к тому же с областью значений. Как видим, качество интервального оценивания существенно зависит от вида выражения. ■

Использование естественного интервального расширения подчас даёт весьма грубые оценки областей значений функций, в связи с чем получили развитие более совершенные способы (формы) нахождения интервальных расширений. Одна из наиболее популярных — так называемая *центрированная форма*:

$$\mathbf{f}_c(\mathbf{x}, \tilde{x}) = f(\tilde{x}) + \sum_{i=1}^n \mathbf{g}_i(\mathbf{x}, \tilde{x})(\mathbf{x}_i - \tilde{x}_i), \quad (1.22)$$

где  $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n)$  — некоторая фиксированная точка, называемая «центром»,

$\mathbf{g}_i(\mathbf{x}, \tilde{x})$  — интервальные расширения некоторых функций  $g_i(x, \tilde{x})$ , которые строятся по  $f$  и зависят в общем случае как от  $\tilde{x}$ , так и от  $\mathbf{x}$ .

Центрированная форма является аналогом полинома Тейлора первой степени, полученного при разложении относительно центра  $\tilde{x}$ . В (1.22)

выражения для  $\mathbf{g}_i(\mathbf{x}, \tilde{x})$  могут быть внешними оценками коэффициентов наклона функции  $f$  на рассматриваемой области определения, взятыми относительно точки  $\tilde{x}$ , или же внешними интервальными оценками областей значений производных  $\partial f(x)/\partial x_i$  на  $\mathbf{x}$ . В последнем случае точка  $\tilde{x}$  никак не используется в  $\mathbf{g}_i(\mathbf{x}, \tilde{x})$ , а интервальная функция  $f_c$  называется *дифференциальной центрированной формой* интервального расширения.<sup>9</sup>

**Пример 1.6.2** Для оценивания функции  $f(x) = x/(x+1)$  на интервале  $\mathbf{x} = [1, 3]$  применим дифференциальную центрированную форму.

Так как

$$f'(x) = \frac{1}{(x+1)^2},$$

то интервальная оценка производной на заданном интервале области определения есть

$$\frac{1}{([1, 3] + 1)^2} = [\frac{1}{16}, \frac{1}{4}].$$

Поэтому если в качестве центра разложения взять середину интервала  $\text{mid } \mathbf{x} = 2$ , то

$$\begin{aligned} f(\text{mid } \mathbf{x}) + f'(\mathbf{x})(\mathbf{x} - \text{mid } \mathbf{x}) &= \frac{2}{3} + [\frac{1}{16}, \frac{1}{4}] \cdot [-1, 1] = \\ &= \frac{2}{3} + [-\frac{1}{4}, \frac{1}{4}] = [\frac{5}{12}, \frac{11}{12}]. \end{aligned}$$

Как видим, этот результат значительно точнее естественного интервального расширения (1.21). ■

За дальнейшей информацией мы отсылаем заинтересованного читателя к книгам [1, 16, 21, 22], развёрнуто излагающим построение интервальных расширений функций. Важно отметить, что погрешность интервального оценивания при использовании любой из форм интервального расширения критическим образом зависит от ширины интервала оценивания. Если обозначить через  $f(\mathbf{x})$  точную область значений целевой функции на  $\mathbf{x}$ , т. е.  $f(\mathbf{x}) = \{f(x) \mid x \in \mathbf{x}\}$ , то для естественного интервального расширения липшицевых функций имеет место неравенство

$$\text{dist} (\mathbf{f}_{\natural}(\mathbf{x}), f(\mathbf{x})) \leq C \|\text{wid } \mathbf{x}\| \quad (1.23)$$

---

<sup>9</sup>По отношению к ней часто используют также термин «среднезначная форма», поскольку она может быть выведена из известной теоремы Лагранжа о среднем.

с некоторой константой  $C$ . Этот факт обычно выражают словами «естественнное интервальное расширение имеет первый порядок точности» (см. определение 2.8.1, стр. 168).

Для центрированной формы верно соотношение

$$\text{dist} \left( f_c(\mathbf{x}, \tilde{x}), f(\mathbf{x}) \right) \leq 2 (\text{wid } \mathbf{g}(\mathbf{x}, \tilde{x}))^\top |\mathbf{x} - \tilde{x}|, \quad (1.24)$$

где  $\mathbf{g}(\mathbf{x}, \tilde{x}) = (\mathbf{g}_1(\mathbf{x}, \tilde{x}), \mathbf{g}_2(\mathbf{x}, \tilde{x}), \dots, \mathbf{g}_n(\mathbf{x}, \tilde{x}))$ . В случаее, когда интервальные оценки для функций  $\mathbf{g}_i(\mathbf{x}, \tilde{x})$  находятся с первым порядком точности, общий порядок точности центрированной формы согласно (1.24) будет уже вторым. Вывод этих оценок заинтересованный читатель может найти, к примеру, в [16, 22].

Интервальные оценки областей значений функций, которые находятся с помощью интервальных расширений, оказываются полезными в самых различных вопросах вычислительной математики. В частности, с помощью интервального языка элегантно записываются остаточные члены различных приближённых формул. В качестве двух содержательных примеров применения интервальных расширений функций мы рассмотрим решение уравнений и систем уравнений (см. главу 4), а также оценку константы Липшица для функций.

## 1.7 Обусловленность математических задач

Термин *обусловленность* означает меру чувствительности решения задачи к изменениям (возмущениям) её входных данных. Ясно, что любая информация подобного рода чрезвычайно важна при практических вычислениях, так как позволяет оценивать достоверность результатов, полученных в условиях приближённого характера этих вычислений. С другой стороны, зная о высокой чувствительности решения, можем предпринимать необходимые меры для компенсации этого явления — повышать разрядность вычислений, наконец, модифицировать или вообще сменить выбранный вычислительный алгоритм и т. п.

Есть несколько уровней рассмотрения поставленного вопроса.

Во-первых, следует знать, является ли вообще непрерывной зависимость решения задачи от входных данных. Задачи, решение которых не зависит непрерывно от их данных, называют *некорректными*. Далее в § 2.8д в качестве примера таких задач мы рассмотрим задачу численного дифференцирования. Некорректна задача определения собственных векторов матрицы (§ 3.17в) и некоторые другие.

Во-вторых, в случае наличия этой непрерывности желательно иметь какую-нибудь количественную меру чувствительности решения к изменению входных данных. Это может быть, к примеру, скорость изменения решения в зависимости от изменения входных данных. Тогда можно будет различать, насколько «хороша» или «плоха» решаемая задача, смотря по значению этой меры.

Переходя к формальным конструкциям, предположим, что в рассматриваемой задаче по значениям из множества  $\mathcal{D}$  входных данных мы должны вычислить решение задачи из множества ответов  $\mathcal{S}$ . Отображение  $\phi : \mathcal{D} \rightarrow \mathcal{S}$ , сопоставляющее всякому  $a$  из  $\mathcal{D}$  решение задачи из  $\mathcal{S}$ , будем называть *разрешающим отображением* (или *разрешающим оператором*). Отображение  $\phi$  может быть записано явным образом, если ответ к задаче задаётся каким-либо выражением. Часто разрешающее отображение задаётся алгоритмом или даже программой для компьютера. Наконец, иногда разрешающее отображение может быть задано неявно, как, например, при решении уравнения или системы уравнений

$$F(a, x) = 0$$

с входным параметром  $a$ .

Даже при неявном задании разрешающего отображения нередко можно теоретически записать его вид, как, например,  $x = A^{-1}b$  при решении системы линейных уравнений  $Ax = b$  с квадратной матрицей  $A$ . Но в любом случае удобно предполагать существование этого отображения и некоторые его свойства. Пусть также  $\mathcal{D}$  и  $\mathcal{S}$  являются линейными нормированными пространствами. Для простоты можно далее считать, что  $\mathcal{D}$  и  $\mathcal{S}$  конечномерны и являются арифметическими векторными пространствами (именно таковы многие задачи этой книги), т. е.  $\mathcal{D} = \mathbb{R}^p$  и  $\mathcal{S} = \mathbb{R}^q$  для некоторых натуральных  $p$  и  $q$ .

Очевидно, что самый первый вопрос, касающийся обусловленности задачи, требует, чтобы разрешающее отображение  $\phi$  было непрерывным относительно некоторого задания норм в  $\mathcal{D}$  и  $\mathcal{S}$ . Такие задачи будем называть *вычислительно-корректными* (см. § 4.2).

Если непрерывность разрешающего отображения имеет место, то для характеристики обусловленности задачи интересна скорость изменения его значений при возмущении исходных данных. Возможны два подхода к введению числовой меры обусловленности математических задач. Один из них условно может быть назван *дифференциальным*, а другой основан на оценивании *константы Липшица* разрешающего оператора.

Пусть разрешающее отображение  $\phi$  дифференцируемо по крайней мере в интересующей нас точке  $a$  из множества входных данных  $\mathcal{D}$ . Тогда можно считать, что

$$\phi(a + \Delta a) \approx \phi(a) + \phi'(a) \cdot \Delta a, \quad \phi'(a) \in \mathbb{R}^{q \times p},$$

и потому

$$\|\phi(a + \Delta a) - \phi(a)\| \lesssim \|\phi'(a)\| \|\Delta a\|,$$

где  $\|\cdot\|$  — нормы векторов или матриц соответственно, согласованные друг с другом (см. § 3.3д). По этой причине мерой чувствительности решения может служить  $\|\phi'(a)\|$ , т. е. норма матрицы Якоби  $\phi'(a)$ .

Для более детального описания зависимости различных компонент решения  $\phi(a)$  от исходных данных  $a$  часто привлекают отдельные частные производные  $\frac{\partial \phi_i}{\partial a_j}$ , т. е. элементы матрицы Якоби разрешающего отображения  $\phi$ . Их обычно называют *коэффициентами чувствительности*. Интересна также мера относительной чувствительности решения, которую можно извлечь из соотношения

$$\frac{\phi(a + \Delta a) - \phi(a)}{\|\phi(a)\|} \approx \left( \frac{\phi'(a)}{\|\phi(a)\|} \cdot \|a\| \right) \frac{\Delta a}{\|a\|}.$$

Второй подход к определению обусловленности требует нахождения как можно более точных констант  $C_1$  и  $C_2$  в неравенствах

$$\|\phi(a + \Delta a) - \phi(a)\| \leq C_1 \|\Delta a\|, \quad (1.25)$$

$$\frac{\|\phi(a + \Delta a) - \phi(a)\|}{\|\phi(a)\|} \leq C_2 \frac{\|\Delta a\|}{\|a\|}. \quad (1.26)$$

Величины этих констант, зависящие от задачи, а иногда и конкретных входных данных, берутся за меру обусловленности решения задачи.

В связи с неравенствами (1.25) и (1.26) напомним, что вещественная функция  $f : \mathbb{R}^n \supseteq D \rightarrow \mathbb{R}$  называется *непрерывной по Липшицу* (или удовлетворяет *условию Липшица*), если существует такая константа  $L$ , что

$$|f(x') - f(x'')| \leq L \cdot \text{dist}(x', x'') \quad (1.27)$$

для любых  $x', x'' \in D$ . Величину  $L$  называют при этом *константой Липшица* функции  $f$  на  $D$ . Понятие непрерывности по Липшичу формулирует интуитивно понятное условие соразмерности изменения функции

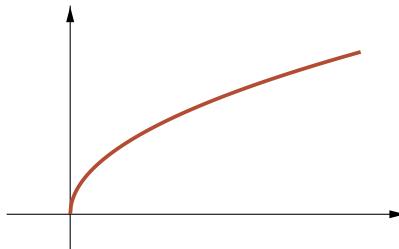


Рис. 1.5. Непрерывная функция  $y = \sqrt{x}$  имеет бесконечную скорость роста при  $x = 0$  и потому не является липшицевой

изменению аргумента. Более точно, приращение функции не должно превосходить приращение аргумента (по абсолютной величине или в некоторой заданной метрике) более чем в определённое фиксированное число раз. При этом сама функция может быть и негладкой, как, например, модуль числа в окрестности нуля. Отметим, что понятие непрерывности по Липшицу является более сильным свойством, чем просто непрерывность или даже равномерная непрерывность, так как влечёт за собой их обеих.

Нетрудно видеть, что искомые константы  $C_1$  и  $C_2$  в неравенствах (1.25) и (1.26), характеризующие чувствительность решения задачи по отношению к возмущениям входных данных — это не что иное, как константы Липшица для разрешающего отображения  $\phi$  и произведение константы Липшица  $L_\psi$  отображения  $\psi : \mathcal{D} \rightarrow \mathcal{S}$ , действующего по правилу  $a \mapsto \phi(a)/\|\phi(a)\|$  на норму  $\|a\|$ . В последнем случае

$$\frac{\|\phi(a + \Delta a) - \phi(a)\|}{\|\phi(a)\|} \lesssim L_\psi \|\Delta a\| \leq L_\psi \|a\| \cdot \frac{\|\Delta a\|}{\|a\|}.$$

## 1.8 Устойчивость алгоритмов

С понятием обусловленности математических задач тесно связано понятие устойчивости алгоритмов для их решения. Неформально *устойчивость* численного алгоритма — это его свойство не увеличивать существенно погрешностей исходных данных и погрешностей, вносимых в процесс его выполнения, так что окончательный результат расчётов сохраняет адекватность. Более формально устойчивость

алгоритма означает, что погрешность его результатов является «не слишком быстро растущей» функцией погрешностей исходных данных и промежуточных результатов расчётов по этому алгоритму. Строгие определения, соответствующие различным вычислительным задачам, можно увидеть в [3, 23, 30, 32].

Если алгоритм устойчив, то он не позволяет неизбежным погрешностям входных данных задачи и погрешностям вычислений сильно вырасти и заметно исказить результат. С неустойчивым алгоритмом такого может не быть, поскольку погрешности в нём не только накапливаются, но и существенно усиливаются, а результат его работы подчас имеет мало общего с ответом к решаемой задаче. Конкретной мерой устойчивости алгоритма может быть, к примеру, мера чувствительности получаемых с его помощью результатов в зависимости от возмущений входных данных и погрешностей промежуточных расчётов. Устойчивый алгоритм характеризуется низкой чувствительностью к таким возмущениям и погрешностям, тогда как у неустойчивого алгоритма эта чувствительность слишком высока.

Примером раздела вычислительной математики, в котором выработана очень развитая техника исследования устойчивости, основанная на специфических приёмах и различных количественных показателях устойчивости, является теория разностных схем и вообще численное решение дифференциальных уравнений [3, 4, 8, 23, 30]. В таких расчётах приходится проводить длинные (или даже очень длинные) цепочки вычислений по единообразным рекуррентным формулам, где неконтролируемое накопление погрешностей является одним из доминирующих эффектов. Соответственно, расчёты по неустойчивым разностным схемам практически малополезны и неинформативны.

Если обусловленность задачи характеризует объективный факт, зависящий самой постановки задачи, то устойчивость алгоритма — свойство сконструированной и применяемой нами процедуры для решения этой задачи. Как следствие, мы можем делать эту устойчивость лучше или хуже, модифицируя алгоритм, его вычислительную схему.

### Пример 1.8.1 (пример Бабушки–Витасека–Прагера [23, 29])

Пусть требуется вычислить для последовательных натуральных чисел  $n = 0, 1, 2, \dots$  интегралы

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

Ясно, что все искомые  $I_n$  положительны, меньше единицы и убывают с ростом номера  $n$ , так как при увеличении  $n$  подинтегральная функция уменьшается.

Для каждого фиксированного  $n$  первообразную для подинтегральной функции нетрудно найти в конечном виде с помощью нескольких интегрирований по частям, но сложность получающихся выражений быстро растёт с увеличением  $n$ . Это наводит на мысль воспользоваться для решения задачи рекуррентной формулой, которая несложно выводится из представления искомого интеграла:

$$\begin{aligned} I_n &= \int_0^1 x^n d(e^{x-1}) = x^n e^{x-1} \Big|_0^1 - n \int_0^1 x^{n-1} e^{x-1} dx = \\ &= 1 - n I_{n-1}. \end{aligned} \quad (1.28)$$

Кроме того,

$$I_0 = \int_0^1 e^{x-1} dx = 1 - e^{-1}.$$

Теперь уже нетрудно рекуррентно организовать вычисление интегралов, но ... поведение их результатов для растущих  $n$  делается странным. Таблица ниже приводит с шестью значащими цифрами результаты этих вычислений в стандартной арифметике с двойной точностью (их легко воспроизвести, к примеру, в любой системе компьютерной математики Scilab, MATLAB, Octave и т. п.):

$n$	Вычисленный $I_n$
1	0.367879
2	0.264241
3	0.207277
...	...
16	0.0554593
17	0.0571919
18	-0.0294537
19	1.55962
...	...

Как видим, 17-е вычисленное значение  $I_n$  больше предыдущего, а 18-е даже отрицательно, что явно нелепо. При дальнейшем увеличении  $n$  вычисленные по рекуррентной формуле (1.28) значения быстро

растут по абсолютной величине и совершенно не отражают истинное значение  $I_n$ . В вычислениях с другой точностью представления чисел в ЭВМ результаты могут отличаться от приведённых в таблице (см., к примеру, [23, 29]), но рано или поздно итоговый «развал» расчётов происходит всегда.

Причина отмеченного явления достаточно прозрачна. Погрешность вычисления  $I_{n-1}$ , какой бы малой она ни была, умножается в рекуррентной формуле  $I_n = 1 - nI_{n-1}$  на  $n$ , т. е. на увеличивающиеся целые числа. Таким образом, при вычислении  $I_n$  исходная погрешность в  $I_0$  получает множитель  $1 \cdot 2 \cdot 3 \cdots n = n!$ , погрешность следующих интегралов — немного меньшие, но тоже большие множители. При ограниченности  $I_n$  это приводит к полному искажению результатов с ростом  $n$ .

Более тонкий анализ примера Бабушки–Витасека–Прагера вскрывает ещё один источник погрешностей. Если рекуррентную формулу  $I_n = 1 - nI_{n-1}$  подставить в двойное неравенство  $0 < I_n < I_{n-1}$ , то получим  $0 < 1 - nI_{n-1} < I_{n-1}$ , откуда

$$\frac{1}{n+1} < I_{n-1} < \frac{1}{n}.$$

Следовательно,  $nI_{n-1} \rightarrow 1$  с ростом  $n$ , так что относительная погрешность значений  $I_n$ , вычисляемых по формуле (1.28), всё сильнее искажает результат из-за вычитания близких чисел: работает «эффект потери точности», рассмотренный в § 1.3.

Итак, алгоритм вычисления  $I_n$  с помощью рекуррентной формулы (1.28) даёт пример неустойчивого алгоритма, в котором ошибки промежуточных вычислений не подавляются, а разрастаются неконтролируемым образом. ■

Другие выразительные примеры неустойчивых вычислений, возникающих в самых обыденных ситуациях, можно увидеть в книге [32].

Неустойчивость алгоритма в примере Бабушки–Витасека–Прагера можно преодолеть, сменив направление рекуррентных вычислений на обратное. Станем последовательно вычислять не  $I_n$  по  $I_{n-1}$ , а  $I_{n-1}$  по  $I_n$  с помощью формулы, которая получается обращением (1.28):

$$I_{n-1} = \frac{1 - I_n}{n}. \quad (1.29)$$

Пусть требуется вычислить интегралы  $I_k$  для  $k = 1, 2, \dots, n$  при некотором фиксированном  $n$ . Возьмём достаточно большое натуральное

число  $N$ , значительно превосходящее  $n$ , и положим  $x_N = 0$ . Это можно сделать потому, что  $I_n$  стремятся к нулю. Естественно, при этом допускается некоторая погрешность, но она, как увидим, подавляется новым алгоритмом. Далее с помощью (1.29) найдём  $I_{N-1}$ , потом  $I_{N-2}$  и так далее, до  $I_n$  и  $I_1$ .

Например, начав с  $N = 10$  и  $I_{10} = 0$ , с помощью описанного выше модифицированного алгоритма получим  $I_0 = 0.632120535714286$  в стандартной арифметике двойной точности. Этот ответ в семи значащих цифрах совпадает с точным  $1 - e^{-1} = 0.632120558828558$ . Если же взять  $N = 17$  и  $I_{17} = 0$ , то  $I_0$  вычисляется уже со всеми 15 верными значащими цифрами.

Описанный выше приём обращения рекуррентной формулы для повышения устойчивости алгоритма носит общий характер и часто с успехом применяется в вычислительной математике. Хорошо известно, например, что при решении дифференциальных уравнений неявные разностные схемы, в которых на каждом временному шаге вычисления идут в направлении, противоположном направлению времени, более устойчивы, чем явные [3, 4, 8, 10, 30].

Для количественной характеристики устойчивости можно использовать те же идеи, что и при введении количественной меры обусловленности.

Ясно, что для хорошо обусловленных задач наилучшими являются устойчивые алгоритмы. Но другая естественная мысль, что для решения плохо обусловленных задач алгоритмы не могут быть лучше самих задач, которые они решают, является верной лишь отчасти. Иногда устойчивые алгоритмы стремятся построить и для плохо обусловленных задач, поскольку именно такие задачи получаются как наиболее подходящие модели интересующих нас явлений (а другие модели часто недоступны). В этом случае говорят, что устойчивый алгоритм *регуляризует* исходную плохообусловленную задачу.

Построение устойчивых алгоритмов для решения плохообусловленных и некорректных задач является предметом теории некорректных задач — важной математической дисциплины, основы которой были заложены А.Н. Тихоновым в середине XX века [31]. За прошедшие десятилетия она получила чрезвычайно большое развитие и многочисленные практические применения, расширявшие сферу применимости вычислительной математики.

Отметим, что устойчивость алгоритмов и достоверность их результатов можно иногда проконтролировать с помощью интервальных вы-

числений, точнее, с помощью машинной интервальной арифметики с внешним направленным округлением (см. § 1.11). В самом деле, запустим исследуемый численный алгоритм в такой интервальной арифметике, задав начальные данные в виде вырожденных интервалов и определив все операции на интервальные. Если полученный в результате интервал имеет аномально большую ширину, то это свидетельствует о возможной неустойчивости исходного алгоритма и его высокой чувствительности к погрешностям вычислений. Для примера Бабушки–Витасека–Прагера подобное исследование проведено, например, в [29].

## 1.9 Элементы конструктивной математики

«Конструктивная математика» — это неформальное название той части современной математики, тех математических дисциплин, — теории алгоритмов, теории сложности вычислений и ряда других, в которых главным объектом изучения являются процессы построения тех или иных математических объектов. Оформление конструктивной математики в отдельную ветвь общего математического дерева произошло на рубеже XIX и XX веков под влиянием обнаруженных к тому времени парадоксов теории множеств. Эти парадоксы заставили критически переосмыслить существовавшие в математике способы рассуждений и само понятие «существования» для математических объектов. Создание основ конструктивного направления в математике связано прежде всего с деятельностью Л.Э.Я. Брауэра и развивающим им «интуиционизмом».

В 30-е годы XX века возникла теория алгоритмов и рекурсивных функций — математическая дисциплина, исследующая конструктивные свойства различных математических объектов. Её основные понятия — это *алгоритм, конструктивный объект, вычислимость, разрешимость* и др.

Алгоритм — это конечная последовательность инструкций, записанных на некотором языке и определяющих процесс переработки исходных данных в искомые результаты (ответ решаемой задачи и т. п.). Алгоритм принципиально конечен и определяет собой конечный процесс. *Конструктивным объектом* называется объект, который может быть построен с помощью конечной последовательности действий над каким-то конечным множеством первичных объектов, элементарных «кирпичиков». Конструктивны, например, рациональные числа: они

получаются как дроби числителями и знаменателями в виде целых чисел.

Говорят, что математическая задача является *алгоритмически разрешимой*, если существует алгоритм, дающий её решение.

Значительным открытием теории алгоритмов стало установление того факта, что некоторые математические задачи *алгоритмически неразрешимы*, т. е. для них не существует никаких алгоритмов, с помощью которых можно было бы получать их решения. Таким задачами являются, например, проблема распознавания истинности формул элементарной арифметики, проблема тождества элементарных функций вещественного переменного, задача выяснения существования целочисленного решения произвольного алгебраического уравнения с целыми коэффициентами (известная также как «10-я проблема Гильберта») и многие другие (см. подробности, к примеру, в [14]).

Так или иначе, конструктивные объекты и только они могут быть получены в качестве ответов при решении задачи на реальных цифровых ЭВМ с конечными быстродействием и объёмом памяти. Широко распространенные ныне электронные цифровые вычислительные машины способны представлять и оперировать, по сути дела, только конечными множествами чисел. Такие машины в принципе не могут использоваться для выполнения абсолютно точных арифметических операций над числовыми полями  $\mathbb{R}$  и  $\mathbb{C}$ , которые являются бесконечными несчётными множествами (мощности континуума): большинство их элементов не представимы в цифровых ЭВМ.

Оказывается, что значительная часть объектов, с которыми работают современная математика и её приложения, не являются конструктивными. В частности, неконструктивным является традиционное понятие вещественного числа, подразумевающее бесконечную процедуру определения всех знаков его десятичного разложения (которое в общем случае непериодично). Факт неконструктивности вещественных чисел может быть обоснован строго математически [26], и он указывает на принципиальные границы возможностей алгоритмического подхода и ЭВМ в деле решения задач математического анализа.

Тем не менее в этом океане неконструктивности имеет смысл выделить объекты, которые могут быть «достаточно хорошо» приближены конструктивными объектами. На этом пути приходим к понятию *вычислимого вещественного числа* [14, 26]: вещественное число  $\alpha$  называется вычислимым, если существует алгоритм, дающий по всякому натуральному числу  $n$  рациональное приближение к  $\alpha$  с погрешностью не

более  $1/n$ . Множество всех вычислимых вещественных чисел образует *вычислимый континуум*. Соответственно, *вычислимая вещественная функция* определяется как отображение из вычислимого континуума в себя, которое задаётся алгоритмом преобразования программы аргумента в программу значений.<sup>10</sup>

Важно помнить, что и вычислимое вещественное число, и вычислимая функция — это уже не конструктивные объекты. Но, как выясняется, даже ценой ослабления наших требований к конструктивности нельзя вполне преодолеть принципиальные алгоритмические трудности, связанные с некоторыми «обычными» постановками задач. Для вычислимых вещественных чисел и функций ряд традиционных постановок задач оказывается алгоритмически неразрешимыми — построение общих алгоритмов их решения принципиально невозможно.

Например, алгоритмически неразрешимыми являются

- 1) задача распознавания равенства/неравенства нулю произвольного вычислимого вещественного числа [25, 26, 27];
- 2) задача распознавания равенства двух вычислимых вещественных чисел [14, 17, 25, 26];
- 3) задача нахождения какого-либо решения для совместной системы линейных алгебраических уравнений над полем конструктивных вещественных чисел [25, 27];
- 4) задача нахождения нулей всякой непрерывной кусочно-линейной знакопеременной функции [27].

Приведённые выше результаты задают, как нам представляется, ту абсолютную и совершенно объективную мерку, с которой следует подходить к оценке трудоёмкости (сложности выполнения) тех или иных вычислительных методов. В главе 4 этой книги (§ 4.2 и § 4.3а), к примеру, проводятся критическое переосмысление и переформулировка классической задачи решения уравнений и систем уравнений, и необходимость этого шага, как выясняется, связана ещё и с тем, что в традиционной постановке эти задачи оказываются алгоритмически неразрешимыми!

---

<sup>10</sup> В известной книге Б.А. Кушнера [25] введённые выше объекты называются немного иначе — *конструктивное вещественное число, конструктивный континуум, конструктивная функция*.

На фоне описанных выше «пессимистических» фактов наличие даже экспоненциально трудного алгоритма с небольшим основанием «одноэтажной» экспоненты в оценке сложности (вроде  $2^n$ ) можно рассматривать как вполне приемлемый вариант разрешимости задачи. Например, именно это имеет место в ситуации с вычислением вращения векторного поля (степени отображения), которое требуется в новой формулировке задачи решения уравнений и систем уравнений, предлагаемой в § 4.2 и 4.3д.

Резюмируя тему, можно сказать, что вычислительная математика тесно примыкает к конструктивной, они находятся во взаимной связи и проникают друг в друга, хотя их цели и методы существенно разнятся.

## 1.10 Сложность задач и трудоёмкость алгоритмов

Как правило, нас удовлетворит не всякий процесс решения поставленной задачи, а лишь только тот, который выполним за практически приемлемое время. Соответственно, помимо алгоритмической разрешимости задач огромную роль играет трудоёмкость (сложность выполнения) тех или иных алгоритмов для их решения. Некоторые алгоритмы, которые работают «слишком долго», имеют главным образом теоретическое значение и на практике бесполезны.

Один из наиболее популярных примеров такого рода — алгоритм Тарского, предназначенный для установления истинности или ложности любой замкнутой арифметической формулы первого порядка с переменными для вещественных чисел [28]. Теоретически его существование означает, что задачи элементарной алгебры (и элементарной геометрии) алгоритмически разрешимы, но практического значения алгоритм Тарского почти не имеет ввиду огромной трудоёмкости.

Другой пример. Множество вещественных чисел, точно представимых в цифровых ЭВМ в формате «с плавающей точкой» согласно стандарту IEEE 754/854, является конечным, и потому мы можем найти, скажем, приближённые значения нулей полинома (или убедиться в их отсутствии) за конечное время, просто перебрав все эти машинные числа и вычисляя в них значения полинома. Но, будучи принципиально выполнимым, такой алгоритм требует непомерных вычислительных затрат и практического значения не имеет.

Естественно измерять трудоёмкость алгоритма количеством «эле-

ментарных операций», требуемых для его исполнения. Следует лишь иметь в виду, что эти операции могут быть весьма различными. Скажем, на сложение и умножение двух чисел «с плавающей точкой» затрачивается разное количество тактов современных процессоров и, соответственно, разное время. Ещё более трудоёмкой операцией является деление чисел. Но до определённой степени эти различия можно игнорировать и оперировать понятием «усреднённой арифметической операции». Именно так определяется «флопс» (flops, flop/s), единица измерения производительности компьютеров, которая показывает, сколько операций с плавающей точкой в секунду выполняет вычислительная система.

Большую роль играет также объём данных, подаваемых на вход алгоритма. К примеру, входными данными могут быть небольшие целые числа, а могут и рациональные дроби с внушительными числителями и знаменателями. Ясно, что переработка больших объёмов данных должна потребовать больших трудозатрат от алгоритма, так что имеет смысл сложность исполнения алгоритма в каждом конкретном случае относить к сложности представления входных данных алгоритма.<sup>11</sup>

На качественном уровне полезно различать *полиномиальную трудоёмкость* и *экспоненциальную трудоёмкость*. Говорят, что алгоритм имеет полиномиальную трудоёмкость, если сложность его выполнения не превосходит значений некоторого алгебраического полинома от длины входных данных. Напротив, алгоритм имеет экспоненциальную трудоёмкость, если сложность его выполнения ограничена сверху экспонентой алгебраического полинома от длины подаваемых ему на вход данных. Строго говоря, существует также «субэкспоненциальная трудоёмкость», при которой сложность выполнения алгоритма растёт быстрее, чем для любого полинома, хотя всё-таки не является экспоненциальной. Но такие алгоритмы также считаются «медленными», и их объединяют с экспоненциально трудоёмкими.

Но свойство задачи иметь алгоритм решения с заданной трудоёмкостью или не иметь его является также объективной характеристикой самой задачи. Некоторые задачи принципиально не могут быть решены проще, чем с помощью алгоритмов, требующих какое-то определённое количество операций. В таких ситуациях говорят о трудоёмкости (сложности) самой задачи. Информация о трудоёмкости (сложности)

---

<sup>11</sup>В связи с этим получили распространение также относительные и косвенные единицы измерения трудоёмкости алгоритмов — через количество вычислений заданной функции, правой части уравнения и т. п.

решения) задач является чрезвычайно ценной, так как позволяет ориентироваться при конструировании конкретных алгоритмов.

Получение оценок трудоёмкости задач является непростым делом. Если какой-то алгоритм решает поставленную задачу, то, очевидно, его трудоёмкость может служить верхней оценкой сложности решения этой задачи. Но вот получение нижних оценок сложности решения задач является чрезвычайно трудным. В явном виде такие нижние оценки найдены лишь для небольшого круга задач, которые имеют, скорее, теоретическое значение. В этих условиях широкое распространение получила альтернативная теория сложности — теория NP-трудных задач и NP-полноты [6, 20]. В её основе лежит понятие *сводимости* задач друг к другу и вытекающие из него выводы об их сравнительной трудоёмкости.

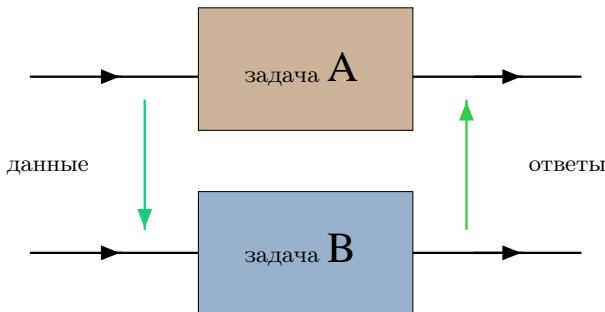


Рис. 1.6. Задача А сводится к задаче В

Задача А *сводится* к задаче В, если исходные данные задачи А можно преобразовать в данные задачи В, а ответы задачи В — в ответы к задаче А, так что с помощью этих преобразований решение задачи А заменяется на решение задачи В. При этом говорят также, что построена *сводимость* или *сведение* задачи А к задаче В.

Если задача А сводится к задаче В и это сведение является «достаточно простым», то тогда естественно считать, что задача В «не легче» задачи А. На этом соображении можно основывать сравнение задач по трудоёмкости.

Наибольшее распространение получило *полиномиальное сведение* — преобразование одной задачи к другой, трудоёмкость которого полиномиальна, т. е. не превышает значений некоторого полинома от разме-

ра исходной задачи. Взаимная полиномиальная сводимость двух задач друг к другу является отношением эквивалентности. Соответственно, все задачи разбиваются на классы эквивалентных по этому отношению задач, т. е. эквивалентных по трудоёмкости задач.

Этим рассуждениям можно придать строгую форму и развить дальше, что приводит к теории NP-трудных задач, получившей большое распространение в последние десятилетия [6]. Её главными объектами являются так называемые классы задач *P* и *NP*. Класс *P* — задачи, разрешимые за полиномиальное время на обычном вычислительном устройстве (машине Тьюринга, например). Класс *NP* — задачи, разрешимые за полиномиальное время на недетерминированном вычислительном устройстве. Можно также сказать, что класс *NP* образован задачами, для которых проверка их удовлетворения решением является полиномиально сложной.

В классе *NP* выявлены универсальные (так называемые *NP*-полные) задачи, к которым полиномиально сводится любая задача из *NP*. Эти «универсальные переборные задачи» (по удачной терминологии одного из создателей теории) как бы дают «эталон сложности». Немного более слабое свойство — *NP*-трудность.

Центральным для теории сводимости является вопрос « $P \neq NP?$ », на который пока строгого ответа нет. Если в самом деле  $P \neq NP$  (что, скорее всего, правильно), то алгоритмов с полиномиальным временем исполнения для *NP*-трудных задач не существует.

Но и в нынешнем не вполне окончательном виде эта теория помогает ориентироваться в сложности решения конкретных практических задач. Если какая-то *NP*-трудная или *NP*-полнная задача полиномиально сводится к интересующей нас задаче, то и нашу задачу также можно считать «труднорешаемой».

В целом теория *NP*-трудности не отвечает напрямую на вопрос о трудоёмкости решения тех или иных задач (по крайней мере, на данный момент, пока не получен определённый ответ в отношении гипотезы « $P = NP?$ »). Но эта теория позволяет утверждать, что некоторые задачи «столь же трудны», как и другие известные задачи, которые признаются «трудными», имеют откровенно переборный характер и т. п. Нередко знание уже одного этого факта бывает существенным для ориентировки создателям вычислительных технологий решения конкретных задач. Если известно, к примеру, что некоторая задача не проще, чем известные «переборные» задачи, которые, по-видимому, не могут быть решены лучше, чем полным перебором всех возможных ва-

риантов, то имеет смысл и для рассматриваемой задачи не стесняться конструирования алгоритмов «переборного» типа, имеющих экспоненциальную трудоёмкость.

Именно такова ситуация с некоторыми задачами, которые возникают в вычислительной математике. Известно, к примеру, что решение систем линейных алгебраических уравнений может быть получено алгоритмами с полиномиальной сложностью (см. главу 3). Но решение приближённых систем линейных или нелинейных уравнений, у которых коэффициенты, свободные члены и параметры не известны точно, в самом общем случае, когда мы не ограничиваем себя величиной неточностей и погрешностей, является NP-трудной задачей [20]. В частности, именно такой является задача оценивания множеств решений интервальных систем линейных алгебраических уравнений (§ 4.6). Созданные для её точного решения алгоритмы, использующие неявный полный перебор (метод ветвей и границ), как показывает теория, не могут быть принципиально улучшены.

## 1.11 Доказательные вычисления на ЭВМ

Термин «доказательные вычисления» был введён в 70-е годы XX века советским математиком К.И. Бабенко для обозначения вычислений, результат которых имеет такой же статус достоверности, как и результаты «чистой математики», полученные с помощью традиционных доказательств. В книге [2], где доказательным вычислениям посвящён отдельный параграф, можно прочитать: «Под *доказательными вычислениями* в анализе мы понимаем такие целенаправленные вычисления на ЭВМ, комбинируемые с аналитическими исследованиями, которые приводят к строгому установлению новых фактов (теорем)». В отношении задач, где ответом являются числа (набор чисел, вектор или матрица и т. п.), доказательность означает свойство гарантированности этих числовых ответов.<sup>12</sup> К примеру, если мы находим число  $\pi$ , то доказательным ответом может быть установление гарантированных неравенств  $\pi > 3.1415$  или

$$3.1415926535 \leq \pi \leq 3.1415926536.$$

---

<sup>12</sup>Численные расчёты такого рода часто называют «вычислениями с гарантированной точностью» или даже просто «гарантированными вычислениями» [24].

Термин «доказательные вычисления на ЭВМ» является хорошим русским эквивалентом таких распространённых английских терминов как validated numerics, verified computation, reliable computation и пр.

Основная трудность, с которой сталкиваются при проведении доказательных вычислений на современных цифровых ЭВМ, вытекает из невозможности адекватно отобразить непрерывную числовую ось  $\mathbb{R}$  в виде множества машинно представимых чисел. Таковых может быть лишь конечное число (либо потенциально счётное), тогда как вещественная ось  $\mathbb{R}$  является непрерывным континуумом. Как следствие, типичное вещественное число не представимо точно в цифровой ЭВМ с конечной разрядной сеткой (рис. 1.7). Например, в моделях двоичной компьютерной арифметики с плавающей точкой стандартных форматов «single» и «double» не представимы точно такие обыденные десятичные дроби, как 0.1, 0.2 и др.



Рис. 1.7. Типичное вещественное число не представимо точно в цифровой ЭВМ с конечной разрядной сеткой

Ситуация в действительности ещё более серьёзна, так как неизбежными погрешностями, как правило, сопровождаются ввод данных в ЭВМ и выполнение с ними любых арифметических операций. В целом мы почти всегда решаем на ЭВМ задачу, немного отличающуюся от исходной, и с помощью алгоритма, который тоже не вполне совпадает с идеальным. Хотя погрешности при вводе и вычислениях могут быть очень малы, но, накапливаясь, они способны существенно исказить ответ к решаемой задаче. Встаёт нетривиальная проблема их учёта в вычислениях на ЭВМ.



Рис. 1.8. Интервальное решение проблемы представления вещественных чисел в цифровой ЭВМ

Одним из средств доказательных вычислений на ЭВМ служит ин-

тервальная арифметика или, точнее, методы интервального анализа. В частности, вещественное число  $x$  в общем случае наиболее корректно представляется в цифровых ЭВМ интервалом, левый конец которого — наибольшее машинно-представимое число, не превосходящее  $x$ , а правый — наименьшее машинно-представимое число, не меньшее  $x$  (рис. 1.8). Далее с получающимися интервалами можно выполнять операции по правилам интервальной арифметики, рассмотренным в § 1.5.

Но концы интервалов, которые получаются при расчётах по формулам (1.11)–(1.14), также могут оказаться вещественными числами, не представимыми в ЭВМ. В этом случае для обеспечения доказательности вычислений имеет смысл несколько расширить полученный интервал до ближайшего объемлющего его интервала с машинно-представимыми концами. Подобная версия интервальной арифметики называется *машинной интервальной арифметикой* с направленным округлением (рис. 1.9).

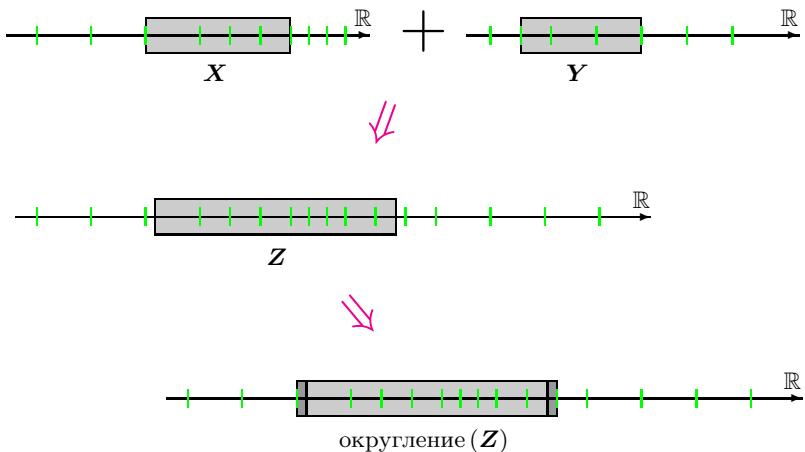


Рис. 1.9. Машинная интервальная арифметика  
с внешним направленным округлением

Машинная интервальная арифметика позволяет достичь гарантированности результатов расчётов, их двусторонних оценок, но ценой некоторого округления. Если общий объём таких вычислений не слишком велик, то часто приобретаемые погрешности вполне приемлемы и с ними можно мириться.

Решение современных задач математического моделирования требует нередко огромных объёмов вычислений, и применение в них всюду машинной интервальной арифметики, конечно же, нецелесообразно и просто бесполезно из-за грубоści получаемых окончательных результатов. Как правило, интервальную арифметику и интервальные методы применяют в ограниченной мере, для отдельных особо ответственных участков, либо непрямым образом, для окончательного контроля готового результата.

Существует несколько подходов к организации доказательных вычислений на ЭВМ, из которых наиболее известными являются *пошаговый способ оценки ошибок, апостериорное оценивание* и различные *интервальные методы*.

В пошаговом способе доказательных вычислений мы разбиваем алгоритм вычисления решения на «элементарные шаги», оцениваем погрешности на каждом шаге вычислений. «Элементарными шагами» здесь могут быть как отдельные арифметические операции, так и цепь их последовательности, слагающиеся в крупные блоки алгоритма. Полная погрешность получается при этом из погрешностей отдельных «элементарных шагов» по правилам исчисления из § 1.3 или их модификациям. Очевидный недостаток такого способа организации оценки погрешностей состоит в том, что мы неявно привязываемся к конкретному алгоритму вычисления решения. Таким образом, качество оценок, получаемых с помощью пошагового подхода, существенно зависит от алгоритма, и «хороший» в обычном смысле алгоритм не обязательно хорош при анализе его погрешностей и их оценивании.

При оценивании погрешностей простых «элементарных шагов» алгоритмов с помощью таких несложных средств, как классическая интервальная арифметика, получаемые оценки, как правило, отличаются невысоким качеством. Но изощрённые варианты пошагового способа оценки погрешностей могут показывать вполне удовлетворительные результаты даже для довольно сложных задач. Таковы, к примеру, вычислительные алгоритмы для решения систем линейных алгебраических уравнений, развиваемые в [24] и других аналогичных работах.

Напротив, при апостериорном оценивании погрешность окончательного результата вычисляется уже *после* его получения или одновременно с ним. При этом технологически можно разделить вычисление двухсторонней оценки решения и установление её доказательности. Более точно, полученный точечный результат окружает интервалом (интервальным вектором и т. п.) достаточно малого размера и далее с помо-

шью дополнительного исследования (не очень сложного) устанавливают присутствие в нём искомого точного решения. Альтернативный подход — начинать уточнение гарантированной двусторонней оценки решения с какого-то большого интервала (интервального вектора-бруса), который заведомо содержит это решение. Потом исходный брус сжимается, его лишние части отсекаются, и мы получаем более или менее точную двустороннюю оценку решения.

Апостериорный подход и интервальные методы оказались существенно более гибкими, практическими и плодотворными, чем пошаговый подход, так как они применимы к более широкому классу задач и позволяют получать существенно более точные результаты. Примеры этих методов (метод Кравчика, интервальный метод Ньютона и др.) и результаты их применения читатель может увидеть в главе 4. Обзоры современного состояния этих вычислительных технологий и дальнейшие ссылки можно найти, например, в [21, 35, 36].

## Литература к главе 1

### Основная

- [1] АЛЕФЕЛЬД Г., ХЕРЦБЕРГЕР Ю. *Введение в интервальные вычисления*. – М.: Мир, 1987.
- [2] БАБЕНКО К.И. *Основы численного анализа*. – М.: Наука, 1986.
- [3] БАХВАЛОВ Н.С., ЖИДКОВ Н.П., КОВЕЛЬКОВ Г.М. *Численные методы*. – М.: Бином, 2003, а также другие издания этой книги.
- [4] БЕРЕЗИН И.С., ЖИДКОВ Н.П. *Методы вычислений*. Т. 1–2. – М.: Наука, 1966.
- [5] ВЕРЖВИЦКИЙ В.М. *Численные методы. Части 1–2*. – М.: «Оникс 21 век», 2005.
- [6] ГЭРИ М., ДЖОНСОН Д. *Вычислительные машины и труднорешаемые задачи*. – М.: Мир, 1982.
- [7] ДЕМИДОВИЧ Б.П., МАРОН А.А. *Основы вычислительной математики*. – М.: Наука, 1970.
- [8] КАЛИТКИН Н.Н. *Численные методы*. – М.: Наука, 1978.
- [9] КРЫЛОВ А.Н. *Лекции о приближённых вычислениях*. – М.: ГИТГЛ, 1954, а также более ранние издания.
- [10] КРЫЛОВ В.И., БОБКОВ В.В., МОНАСТЫРНЫЙ П.И. *Вычислительные методы*. Т. 1–2. – М.: Наука, 1976.
- [11] МАТИЯСЕВИЧ Ю.В. Вещественные числа и ЭВМ // *Кибернетика и вычислительная техника*. – М.: Наука, 1986. – Вып. 2. – С. 104–133.
- [12] МЕНЬШИКОВ Г.Г. *Локализующие вычисления. Конспект лекций*. – СПб.: СПбГУ, Факультет прикладной математики–процессов управления, 2003.

- [13] Тыртышников Е.Е. Методы численного анализа. – М.: Академия, 2007.
- [14] Успенский В.А., Семёнов А.Л. Теория алгоритмов: основные открытия и приложения. – М.: Наука, 1987.
- [15] Хансен Э., Уолстер Дж.У. Глобальная оптимизация с помощью методов интервального анализа. – М.-Ижевск: Издательство «РХД», 2012.
- [16] Шарый С.П. Конечномерный интервальный анализ. – ФИЦ ИВТ: Новосибирск, 2024. – Электронная книга, доступная на <http://www.nsc.ru/interval/Library/InteBooks/>)
- [17] АБЕРТ О. *Introduction to precise numerical methods*. – Amsterdam-Boston-Heidelberg: Academic Press, Elsevier Science, 2007.
- [18] СЕВЕРИО М., КОСХЕЛЕВА О., КРЕИНОВИЧ В. How to describe relative approximation error? A new justification for Gustafson's logarithmic expression // Technical Report: UTEP-CS-22-28 at University of Texas at El Paso. Доступна на [https://scholarworks.utep.edu/cs\\_techrep/1667](https://scholarworks.utep.edu/cs_techrep/1667)
- [19] ГОРДБЕРГ Д. What every computer scientist should know about floating point arithmetic // *ACM Computing Surveys*. – 1991. – Vol. 23, No. 1. – P. 5–48.
- [20] КРЕИНОВИЧ В., ЛАКЕЙЕВ А.В., РОНН Ј., КАЛЬ П. *Computational complexity and feasibility of data processing and interval computations*. – Dordrecht: Kluwer, 1997.
- [21] МУРР Р.Е., КИАРФОРТ Р.Б., КЛУД М. *Introduction to interval analysis*. – Philadelphia: SIAM, 2009.
- [22] НЕУМАЙЕР А. *Interval methods for systems of equations*. – Cambridge: Cambridge University Press, 1990.

## Дополнительная

- [23] Бабушка И., Витасек Э., Прагер М. Численные процессы решения дифференциальных уравнений. – М.: Мир, 1969.
- [24] Годунов С.К., Антонов А.Г., Кирилюк О.Г., Костин В.И. Гарантированная точность решения систем линейных уравнений в евклидовых пространствах. – Новосибирск: Наука, 1988 и 1992.
- [25] Кушнер Б.А. Лекции по конструктивному математическому анализу. – М.: Наука, 1973.
- [26] Мартин-Лёф П. Очерки по конструктивной математике. – М.: Наука, 1975.
- [27] Математический Энциклопедический Словарь. – М.: Большая Российская Энциклопедия, 1995.
- [28] Матиясевич Ю.В. Алгоритм Тарского // Компьютерные инструменты в образовании. – 2008. – №6. – С. 4–14.
- [29] Меньшиков Г.Г. Демонстрационные возможности примера Бабушки-Витасека-Прагера в точечных и интервальных расчетах // Вестник Санкт-Петербургского университета. Серия 10. Прикладная математика. Информатика. Процессы управления. – 2005. – Вып. 2. – С. 179–183.

- [30] САМАРСКИЙ А.А., Гулин А.В. Устойчивость разностных схем. – М.: Наука, 1973.
- [31] Тихонов А.Н., Арсенин В.Я. Методы решения некорректных задач. – М.: Наука, 1979.
- [32] HIGHAM N.J. Accuracy and stability of numerical algorithms. – Philadelphia: SIAM, 2002.
- [33] IEEE Std 754-1985. IEEE Standard for Binary Floating-Point Arithmetic. – New York: IEEE, 1985.
- [34] Mathematics Subject Classification – MSC2020. – Доступна на <https://zbmath.org/classification/> или <https://mathscinet.ams.org/mathscinet/msc/msc2020.html>
- [35] MAYER G. Interval analysis and automatic result verification. – Berlin: De Gruyter, 2017.
- [36] RUMP S.M. Verification methods: rigorous results using floating-point arithmetic // Acta Numerica. – 2010. – Vol. 19. – P. 287–449.

## Глава 2

# Численные методы анализа

## 2.1 Введение

Численными методами анализа обычно называют вычислительные методы решения ряда задач, возникающих в классическом математическом анализе. Традиционно сюда относят задачи интерполяции и приближения функций, задачи численного нахождения производных и интегралов, задачу суммирования рядов, а также вычислительные методы гармонического анализа, т. е. методы, связанные с представлениями функций в виде рядов или интегралов Фурье. Кроме того, численные методы анализа охватывают задачу вычисления значений функций. Она относительно проста для функций, явно задаваемых несложными арифметическими выражениями, но становится нетривиальной и трудной в случае, когда функция задаётся неявно или с помощью формул, выводящих за пределы конечного набора элементарных арифметических действий.

В нашем курсе мы рассмотрим первые четыре из перечисленных выше задач. Их нередко относят также к математическим задачам *анализа данных*, поскольку на практике они часто встречаются при обработке результатов наблюдений и измерений. Сначала займёмся задачами интерполяции и приближения функций, которые возникают, к примеру, при восстановлении функциональных зависимостей по экспериментальным данным.

Задачи интерполяирования и задачи приближения функций являются тесно связанными друг с другом и укладываются в рамки следующей единой неформальной схемы.<sup>1</sup> Пусть дана функция  $f(x)$ , принадлежащая некоторому классу функций  $\mathcal{F}$ , и пусть также задан класс функций  $\mathcal{G}$ . Требуется найти функцию  $g(x)$  из  $\mathcal{G}$ , которая в определённом заранее смысле «достаточно близка» (или даже «наиболее близка») к данной функции  $f(x)$ .

В зависимости от смысла, который вкладывается в понятие «близости» функций, в зависимости от того, какие именно функции образуют классы  $\mathcal{F}$  и  $\mathcal{G}$ , здесь могут получаться различные конкретные постановки задач. При этом рассматриваемые классы функций нередко наделяются дополнительной структурой, облегчающей получение решения. Например, в некоторых задачах можно считать, что эти классы являются линейными векторными пространствами с нормой и т. п. Наконец, часто имеет место включение  $\mathcal{G} \subset \mathcal{F}$ , т. е. функции из  $\mathcal{F}$  приближаются с помощью функций из какого-то более узкого подкласса.

Задача интерполяирования получается из приведённой выше общей формулировки, когда «близость» функций  $f$  и  $g$  означает их совпадение на некотором дискретном множестве точек  $x_0, x_1, \dots, x_n$  из общей области определения. От функции  $f$  требуются лишь значения на этом множестве точек, и потому при постановке задачи интерполяции она сама иногда даже не фигурирует. Вместо  $f$  обычно задаются лишь её значения  $y_0, y_1, \dots, y_n$  в точках  $x_0, x_1, \dots, x_n$  соответственно.

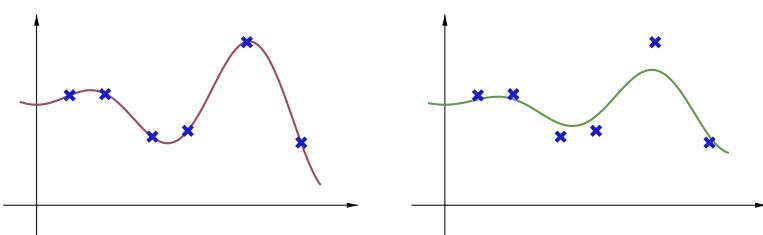


Рис. 2.1. Различие задач интерполяции и дискретного приближения функций

Задача приближения функций (называемая также задачей  *аппроксимации функций*) является частным случаем выписанной выше общей

---

<sup>1</sup>Наряду с термином «интерполяирование» в равной мере используется его синоним «интерполяция».

формулировки, в котором «близость» и «отклонение» функций  $f$  и  $g$  друг от друга понимаются как малое различие их значений на одних и тех же аргументах. При этом можно рассматривать и сравнивать значения этих функций как на всей общей области определения, так и на некотором более узком множестве (множестве сравнения), на котором, например, только и доступны значения функций в интересующей нас практической ситуации. Это множество сравнения может быть небольшой частью общей области определения сравниваемых функций, скажем, конечным набором точек, аналогично задаче интерполяции. В последнем случае получается *дискретная* задача приближения, близкая к задаче интерполяирования. Но в любом случае в задаче приближения, в отличие от задачи интерполяции, точное равенство функции  $g$  заданным значениям не требуется, и это наглядно иллюстрируется на рис. 2.1.

Обозначим множество сравнения посредством  $\mathcal{C}$ . Для постановки задачи приближения нужно определить «близость» функций или их «отклонение» друг от друга, опираясь на различие значений  $f(x)$  и  $g(x)$  для каждого  $x \in \mathcal{C}$ . Это может быть сделано разнообразными способами, которые определяются, как правило, практическим смыслом задачи, и они могут быть не вполне равносильны друг другу.

Пусть для функций  $f, g : \mathbb{R} \supseteq [a, b] \rightarrow \mathbb{R}$  множество сравнения  $\mathcal{C}$  совпадает со всей областью их определения, т. е. интервалом  $[a, b]$ . В качестве расстояния между функциями  $f$  и  $g$  очень популярно максимальное значение модуля разности  $f(x) - g(x)$ , т. е.

$$\max_{x \in [a, b]} |f(x) - g(x)|, \quad (2.1)$$

Оно называется *равномерным* или *чебышёвским* расстоянием.

Во многих физических задачах ясный смысл имеет интеграл от модуля функции, и ему соответствует *интегральное расстояние* между функциями, определяемое как

$$\int_a^b |f(x) - g(x)| dx. \quad (2.2)$$

В § 2.11 мы рассмотрим также задачу среднеквадратичного приближения, в которой расстояние между функциями  $f$  и  $g$  на интервале  $[a, b]$  полагается равным

$$\sqrt{\int_a^b (f(x) - g(x))^2 dx}. \quad (2.3)$$

Кроме перечисленных выше применяются и другие расстояния между функциями. Отметим, что расстояния (2.1)–(2.3) не эквивалентны друг другу в том смысле, что сходимость последовательности функций к какому-то пределу относительно одного из этих расстояний не обязательно влечёт сходимость относительно другого.

В дискретной задаче приближения функций, когда множество сравнения является набором точек  $x_0, x_1, \dots, x_n$ , аналогами расстояний (2.1)–(2.3) являются:

$$\max_{0 \leq i \leq n} |f(x_i) - g(x_i)| — \text{для чебышёвского расстояния,}$$

$$\sum_{i=0}^n |f(x_i) - g(x_i)| — \text{для интегрального расстояния,}$$

$$\sqrt{\sum_{i=0}^n (f(x_i) - g(x_i))^2} — \text{для среднеквадратичного расстояния.}$$

Иногда во второе и третье выражения добавляют усреднение — масштабирующий множитель  $1/n$  перед суммой (тогда квадратичное расстояние становится среднеквадратичным по существу).

Удобно формулировать задачу приближения функций, опираясь на понятие метрики (абстрактного расстояния), которая задана на рассматриваемом классе функций. Напомним, что *метрикой* на множестве  $\mathcal{V}$ , образованном элементами произвольной природы, называется определённая на декартовом произведении  $\mathcal{V} \times \mathcal{V}$  функция  $\text{dist}$  с неотрицательными вещественными значениями, удовлетворяющая для любых  $u, v, w \in \mathcal{V}$  следующим условиям [12, 32]:

- ▶  $\text{dist}(u, v) = 0$  тогда и только тогда, когда  $u = v$ ,
- ▶  $\text{dist}(u, v) = \text{dist}(v, u)$  — симметричность,
- ▶  $\text{dist}(u, w) \leq \text{dist}(u, v) + \text{dist}(v, w)$  — неравенство треугольника.

Само множество, на котором определена метрика (расстояние), называют *метрическим пространством*.

Разнообразные способы определения метрик на пространствах функций, которые возникают в практике математического моделирования, приводят к различным математическим задачам приближения. В задачах дискретного приближения функций нередко требуется характеризовать «расстояние» между функциями или их отклонение друг от друга.

га для всей области изменения непрерывного аргумента, а не только на множестве сравнения. Такое отклонение можно адекватно описать понятием *псевдорасстояния* (псевдометрики), которое определяется почти так же, как обычное расстояние, но отличается от него ослаблением первого условия-аксиомы: хотя всегда имеет место  $\text{dist}(f, f) = 0$ , но из  $\text{dist}(f, g) = 0$  не обязательно следует, что  $f = g$ . Тогда псевдорасстояние между двумя функциями, совпадающими на заданном наборе значений аргумента, будет равно нулю, даже если эти функции не равны в точности друг другу, т. е. различаются при каких-то других аргументах.<sup>2</sup>

Если пространство, на котором требуется рассматривать расстояние, несёт на себе линейную структуру, т. е. является линейным векторным пространством, то удобно задавать расстояние с помощью нормы.

Напомним, что *нормой* называется обобщение понятия абсолютной величины или модуля вещественного числа, которое может быть применено для векторов в общих линейных пространствах. Норму можно трактовать как «длину вектора», расстояние до начала координат и т. п. В самом общем случае норма обычно определяется аксиоматически. На линейном пространстве  $\mathcal{X}$  над полем вещественных или комплексных чисел нормой называется неотрицательная вещественновзначная функция, обозначаемая по традиции как  $\|\cdot\|$ , которая удовлетворяет для любых  $x, y, z \in \mathcal{X}$  следующим условиям [12, 15, 16, 32]:

- $\|x\| = 0$  тогда и только тогда, когда  $x = 0$ ,
- $\|\alpha x\| = |\alpha| \|x\|$  — абсолютная однородность,
- $\|x + y\| \leq \|x\| + \|y\|$  — «неравенство треугольника».

Само линейное пространство  $\mathcal{X}$ , снабжённое нормой, называют *нормированными линейным пространством*.

В этой книге нормы подробно рассматриваются в конечномерной ситуации в главе 3 (см. § 3.3) в связи с задачами вычислительной линейной алгебры. Но понятия нормы и нормированного линейного пространства являются популярными и широко используемыми конструкциями, которые встречается в самых разнообразных разделах математики и приложений. Этому способствовали главным образом работы С. Банаха в 20-х годах XX века. Именно в норме измеряют обычно отклонение функций (непрерывного или дискретного аргумента) друг от друга. В нормированном пространстве расстояние между элементами

---

<sup>2</sup>Для этого ослабленного расстояния можно встретить и другие термины. Так, в книге [15] используется термин «квазирасстояние».

$f$  и  $g$  естественно определяются как норма их отличия друг от друга:

$$\text{dist}(f, g) := \|f - g\|. \quad (2.4)$$

Нетрудно проверить, что введённое таким образом расстояние удовлетворяет всем аксиомам метрики, выписанным выше.

**Пример 2.1.1** Множество всех непрерывных вещественных функций, заданных на интервале  $[a, b] \subset \mathbb{R}$ , можно сделать линейным векторным пространством над полем вещественных чисел, если определить сложение функций и умножение на число поточечным образом, т. е. как

$$(f + g)(x) := f(x) + g(x),$$

$$(\alpha f)(x) := \alpha f(x).$$

В качестве норм на этом пространстве можно взять, например,

$$\begin{aligned} \|f\|_1 &:= \int_a^b |f(x)| \, dx, \\ \|f\|_2 &:= \left( \int_a^b |f(x)|^2 \, dx \right)^{1/2}, \\ \|f\|_\infty &:= \max_{x \in [a, b]} |f(x)|. \end{aligned}$$

Эти нормы с помощью определения (2.4) порождают соответствующие расстояния (2.1)–(2.3), которые рассматривались выше. ■

## 2.2 Интерполяирование функций

### 2.2a Постановка задачи и её свойства

Задача интерполяирования — это задача восстановления (доопределения) функции, которая задана на дискретном множестве точек из области своего определения, на какую-то более широкую область непрерывного изменения аргумента. Для вещественных функций одной переменной её постановка такова (рис. 2.2).

Заданы интервал  $[a, b] \subset \mathbb{R}$  и конечное множество несовпадающих точек  $x_i \in [a, b]$ ,  $i = 0, 1, \dots, n$ , называемых *узлами интерполяции*. Со вокупность всех узлов — множество  $\{x_0, x_1, \dots, x_n\}$  — будем называть

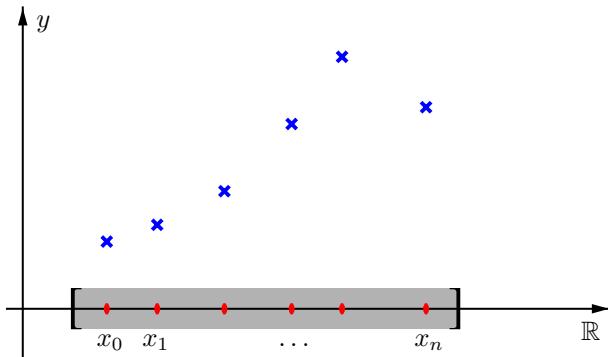


Рис. 2.2. Иллюстрация постановки задачи интерполяции.

*сеткой*. Даны также вещественные числа  $y_i, i = 0, 1, \dots, n$ . Требуется построить функцию  $g(x)$  от непрерывного аргумента  $x \in [a, b]$ , которая принадлежит заданному классу функций  $\mathcal{G}$  и в узлах  $x_i$  принимает значения  $y_i, i = 0, 1, \dots, n$ . Искомую функцию  $g(x)$  называют *интерполяющей функцией* или *интерполянтом*.

Часто значения  $y_0, y_1, \dots, y_n$  принимаются в заданных узлах  $x_0, x_1, \dots, x_n$  некоторой реальной функцией непрерывного аргумента  $f(x)$ . Как и ранее, требуется построить функцию  $g(x)$  от аргумента  $x \in [a, b]$ , которая принадлежит заданному классу функций  $\mathcal{G}$  и в узлах  $x_i$  принимает значения  $y = f(x_i), i = 0, 1, \dots, n$ . В этом случае будем говорить, что рассматривается *задача интерполяции функции  $f(x)$  по узлам  $x_0, x_1, \dots, x_n$* .

Узлы интерполяции  $x_0, x_1, \dots, x_n$ , фигурирующие в этой постановке, нередко называют также *простыми узлами*, так как информация о функции задаётся в них единожды, по одному разу. Соответственно, нашу задачу можно называть задачей интерполяции с простыми узлами. Но на практике возможны ситуации, когда в некоторых узлах информация о функции задаётся более одного раза — значениями самой функции и каких-то её производных и т. п. Такие узлы называются *кратными*, а про соответствующую задачу говорят, что в ней рассматривается интерполяция по кратным узлам (см. далее § 2.4).

Практическая значимость задачи интерполяции чрезвычайно велика. Она встречается всюду, где у функции непрерывного аргумента (который может быть временем, пространственной координатой и т. п.)

мы имеем возможность наблюдать лишь значения в дискретном множестве точек, но хотим восстановить по ним ход функции на всём множестве значений аргумента. Например, выполнение многих химических и биологических анализов требует существенного времени, так что множество результатов этих анализов на любом временнóм интервале по необходимости дискретно. Если нужно отслеживать по ним непрерывно изменяющийся параметр какого-либо процесса, то неизбежно потребуется интерполирование результатов анализов.

Очень часто дискретность множества точек, в которых наблюдаются на практике значения функции, вызвана ограниченностью ресурсов, которые мы можем выделить для сбора данных, или же вообще недоступностью этих данных. Именно это происходит при наблюдении за параметрами земной атмосферы (скоростью и направлением ветра, температурой, влажностью, и пр.) по данным их измерений, которые предоставляются отдельными метеостанциями.

В качестве ещё одного примера интерполирования упомянем вычисление различных функций, как элементарных —  $\sin$ ,  $\cos$ ,  $\exp$ ,  $\log$ , ..., так и более сложных, называемых «специальными функциями», которые часто встречаются в различных задачах естествознания [41]. С подобной задачей человеческая цивилизация столкнулась очень давно, столетия и даже тысячелетия назад, и типичным способом её решения в докомпьютерную эпоху было составление для нужд практики таблиц — *табулирование*. Этим термином называется вычисление значений интересующей нас функции при некоторых специальных фиксированных значениях аргумента, более или менее плотно покрывающих область определения, и сведение этих значений в структурированную таблицу.

Подобные таблицы составлялись квалифицированными вычислителями, иногда специально создаваемыми для этой цели организациями, а затем широко распространялись по научным и техническим центрам, по библиотекам и т. п., так что к ним всегда имели доступ люди, занимающиеся практическими вычислениями. Но как, имея подобную таблицу, вычислить значение интересующей нас функции для аргумента, который не представлен в таблице точно? Скажем, найти синус угла  $17^{\circ}23'$  по таблице, где аргумент идёт с шагом  $6'$ , т. е. шесть угловых минут (одна десятая градуса)?<sup>3</sup>

Здесь на помощь приходит интерполяция — нахождение значения функции в промежуточных точках по ряду известных значений в неко-

---

<sup>3</sup>Именно таковы, к примеру, популярные «Четырехзначные математические таблицы» В.М. Брадиса [3] для средней школы.

торых фиксированных опорных точках. Собственно, сам термин «интерполяция» («интерполяция») был впервые употреблён в 1656 году Дж. Валлисом при составлении астрономических и математических таблиц. Он происходит от латинского слова «*interpolo*», означающего «переделывать», «подновлять», «ремонтировать».

Для целей практических вычислений таблицы значений различных функций составлялись и издавались вплоть до середины XX века. Издаются они и сейчас, хотя и не столь интенсивно. Вершиной этой деятельности стал выпуск многих томов капитальных таблиц, в которых были тщательно затащированы все основные функции, встречающиеся в математической и инженерной практике (см., к примеру, [41] и аналогичные таблицы для других целей).

Интересно, что с появлением и развитием электронных цифровых вычислительных машин описанное применение интерполяции не кануло в лету. В начальный период развития ЭВМ преобладал алгоритмический подход к вычислению элементарных и специальных функций, когда основной упор делался на создании алгоритмов, способных вычислить функцию, исходя из какого-нибудь её аналитического представления, например, в виде быстросходящегося ряда и т. п. (см., к примеру, [23, 24]). Но затем, по мере увеличения объема памяти ЭВМ и повышения её быстродействия, постепенно распространился подход, сильно напоминающий старый добрый табличный способ, но уже на новом уровне. Хранение сотен килобайт или даже мегабайт цифровой информации и быстрый доступ к ним никаких проблем сейчас не представляет, и потому для современных компьютеров программы вычисления функций (элементарных и специальных), как правило, включают в себя библиотеки затащированных значений этих функций для фиксированных аргументов. Опираясь на них, строится значение в нужной нам точке.

Ещё один источник возникновения задачи интерполяирования — это желание иметь просто вычисляемое выражение для сложных функциональных зависимостей, заданных явно или неявно, которые в исходной форме требуют очень большого труда для своего вычисления.

Если класс  $\mathcal{G}$  интерполирующих функций достаточно широк, то решение задачи интерполяции может быть неединственным (см. рис. 2.3). Напротив, если  $\mathcal{G}$  узок, то у задачи интерполяции может вовсе не быть решений. На практике выбор класса  $\mathcal{G}$  обычно диктуется спецификой решаемой практической задачи.

В случае, когда, к примеру, заранее известно, что интерполируе-

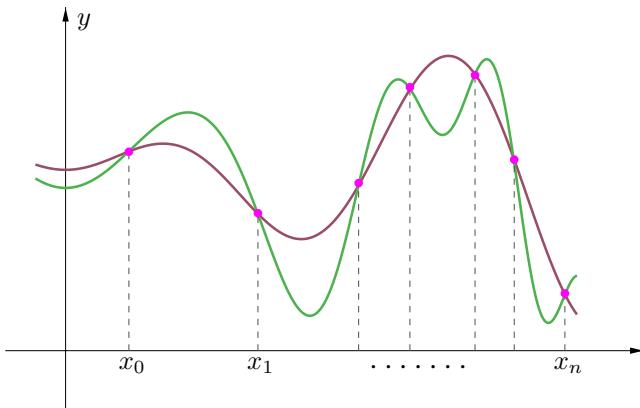


Рис. 2.3. Задача интерполяции может иметь неединственное решение

мая функция периодична, в качестве интерполянтов естественно взять тоже периодические функции с тем же периодом. Ими могут быть, в частности, тригонометрические полиномы

$$\frac{a_0}{2} + \sum_{k=1}^m (a_k \cos kx + b_k \sin kx) \quad (2.5)$$

для некоторой фиксированной степени  $m$  (там, где требуется гладкость), либо пилообразные функции или «ступеньки» (в импульсных системах) и т. п.

Ниже мы подробно рассмотрим ситуацию, когда в качестве интерполирующих функций берутся алгебраические полиномы —

$$a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m. \quad (2.6)$$

Они являются простым и хорошо изученным математическим объектом, а их вычисление реализуется несложно. При этом мы откладываем до § 2.5 рассмотрение вопроса о том, насколько подходящими такие полиномы являются для различных случаев интерполяирования. Вообще, проблема наиболее адекватного выбора класса интерполирующих функций  $\mathcal{F}$  не является тривиальной. Для её хорошего решения, как правило, необходимо, чтобы интерполирующие функции были «той же природы», что и интерполируемые функции из класса  $\mathcal{F}$  (который может даже не фигурировать в формальной постановке задачи). Если

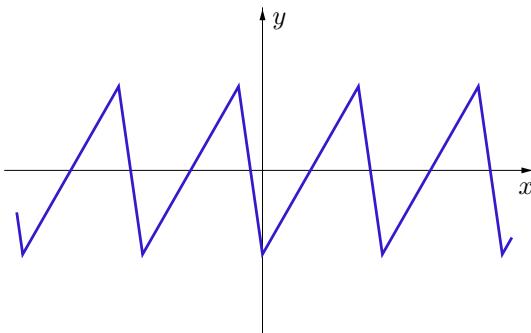


Рис. 2.4. Функция, которую лучше интерполировать с помощью периодических функций

это условие не выполнено, то задача интерполяции может решаться неудовлетворительно.

## 2.2б Алгебраическая интерполяция

**Определение 2.2.1** *Интерполярование функций с помощью алгебраических полиномов называют алгебраической интерполяцией. Алгебраический полином  $P_m(x) = a_0 + a_1x + a_2x^2 + \dots + a_mx^m$ , решający задачу алгебраической интерполяции, называется интерполяционным полиномом или алгебраическим интерполянтом.*

Впервые общая задача алгебраической интерполяции была рассмотрена Дж. Грегори в 1670 году.

Как по интерполяционным данным  $(x_i, y_i)$ ,  $i = 0, 1, \dots, n$ , найти интерполяционный полином вида (2.6), т. е. определить его коэффициенты  $a_0, a_1, \dots, a_m$ ?

Подставляя в выражение (2.6) значения аргумента  $x_0, x_1, \dots, x_n$  и учитывая, что получающиеся при этом значения полинома должны совпадать с  $y_0, y_1, \dots, y_n$  соответственно, приходим к равенствам

$$\begin{aligned} a_0 + a_1x_0 + a_2x_0^2 + \dots + a_mx_0^m &= y_0, \\ a_0 + a_1x_1 + a_2x_1^2 + \dots + a_mx_1^m &= y_1, \\ \vdots &\quad \vdots \quad \vdots \quad \ddots \quad \vdots \quad \vdots \\ a_0 + a_1x_n + a_2x_n^2 + \dots + a_mx_n^m &= y_n. \end{aligned} \tag{2.7}$$

Они образуют систему линейных алгебраических уравнений относительно неизвестных коэффициентов  $a_0, a_1, a_2, \dots, a_m$  искомого полинома. Решив её, можно построить и сам полином.

В самом общем случае, если мы не накладываем никаких ограничений на степень полинома  $m$  и количество узлов интерполяции  $n+1$ , система (2.7) может не иметь решения, а если оно существует, то может быть неединственным. Тем не менее имеется важный частный случай задачи алгебраической интерполяции, для которого гарантируется однозначная разрешимость системы (2.7).

**Теорема 2.2.1** *Если  $m = n$ , т. е. степень интерполяционного полинома на единицу меньше количества узлов, то решение задачи алгебраической интерполяции существует и единствено.*

**Доказательство.** При  $m = n$  в системе линейных алгебраических уравнений (2.7) число неизвестных совпадает с числом уравнений, а матрица этой системы — квадратная. Она имеет вид

$$V(x_0, x_1, \dots, x_n) = \begin{pmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{pmatrix} \quad (2.8)$$

и является так называемой матрицей Вандермонда (см., к примеру, [17, 22, 37]). Её определитель равен, как известно, произведению

$$\prod_{0 \leq i < j \leq n} (x_j - x_i),$$

и он не зануляется, если узлы интерполяции попарно отличны друг от друга. Следовательно, система линейных уравнений (2.7) однозначно разрешима тогда при любой правой части, т. е. при любых  $y_i$ ,  $i = 0, 1, \dots, n$ . ■

Теорема 2.2.1 и предшествующие ей рассуждения дают конструктивный способ построения интерполяционного полинома через решение системы линейных алгебраических уравнений, который вполне практичен, особенно при небольших  $n$ . Он носит общий характер и пригоден для других сходных случаев, когда применяются так называемые

*линейные методы интерполяции.* Этим термином мы будем называть способы интерполяции, в которых интерполирующие функции из класса  $\mathcal{G}$  линейно зависят от некоторых параметров. В частности, это имеет место, когда  $\mathcal{G}$  является линейным векторным пространством с заданным базисом. Если число параметров конечно, т. е. конечна размерность пространства  $\mathcal{G}$ , то условия удовлетворения интерполяционным данным приводят к необходимости решения системы линейных уравнений относительно этих параметров, аналогично тому, как получилось выше для алгебраических полиномов.

Например, сказанное справедливо при тригонометрической интерполяции, т. е. с помощью функций, задаваемых выражениями (2.5), при интерполяции суммами экспонент вида  $a_0 + a_1 e^{\beta_1 x} + \dots + a_m e^{\beta_m x}$  с несовпадающими  $\beta_k$ , а также в некоторых других практически важных ситуациях.

Если же интерполирующие функций из класса  $\mathcal{G}$  нельзя представить линейно зависящими от параметров, то соответствующую задачу интерполяции будем называть *нелинейной*. Для определения интерполянта тогда необходимо решать систему нелинейных уравнений.

**Пример 2.2.1** Рассмотрим алгебраическую интерполяцию степенной функции  $y = x^{0.4}$  на интервале  $[0, 5]$  по набору равномерно расположенных узлов  $\{0, 1, 2, 3, 4, 5\}$ . Интерполяционный полином имеет здесь пятую степень, и для отыскания его коэффициентов  $a_0, a_1, a_2, a_3, a_4$  и  $a_5$  нужно решить систему линейных алгебраических уравнений вида (2.7) с матрицей Вандермонда (2.8) по выбранной совокупности узлов:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 & 16 & 32 \\ 1 & 3 & 9 & 27 & 81 & 243 \\ 1 & 4 & 16 & 64 & 256 & 1024 \\ 1 & 5 & 25 & 125 & 625 & 3125 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{pmatrix} = \begin{pmatrix} 0^{0.4} \\ 1^{0.4} \\ 2^{0.4} \\ 3^{0.4} \\ 4^{0.4} \\ 5^{0.4} \end{pmatrix}.$$

Решение выписанной системы несложно получить в любой системе компьютерной математики или даже с помощью ручных вычислений (очевидно, что  $a_0 = 0$ , и система реально имеет размер  $5 \times 5$ ). Искомый интерполяционный полином имеет вид

$$1.7796 x - 1.1059 x^2 + 0.38831 x^3 - 0.066345 x^4 + 0.0043460 x^5 \quad (2.9)$$

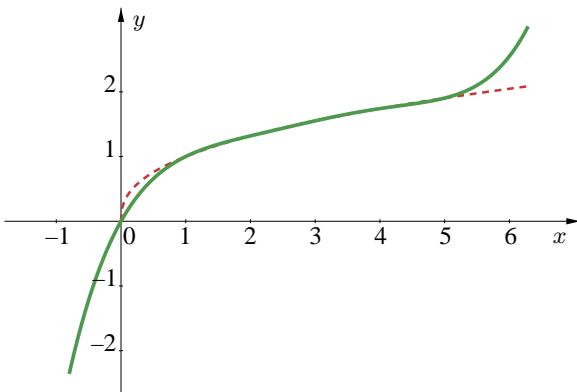


Рис. 2.5. График интерполяционного полинома (2.9).

(коэффициенты даны с точностью до пяти значащих цифр). Его график изображён на рис. 2.9 сплошной линией. График интерполируемой функции  $y = x^{0.4}$  построен там же штриховой линией.

Качество приближения вблизи нуля неудовлетворительно, и этот феномен, очевидно, вызван бесконечной производной исходной функции в нуле (и очень большими значениями около нуля). Алгебраические полиномы всюду имеют конечные производные и, естественно, не могут хорошо воспроизвести эту особенность интерполируемой функции.

Далее, на интервале  $[1, 5]$ , интерполяционный полином (2.9) хорошо приближает функцию, но вне интервала интерполяции, после узла  $x_5 = 5$ , значение полинома резко отклоняется от значений функции. ■

Качество приближения функции при алгебраической интерполяции будет более подробно рассмотрено в § 2.2e и § 2.5.

## 2.2в Интерполяционный полином Лагранжа

Развитый в предшествующем разделе способ построения интерполянта через решение системы уравнений в силу ряда причин может не удовлетворить практику. Например, иногда желательно иметь для интерполяционного полинома какое-либо явное аналитическое представление через исходные данные  $x_1, \dots, x_n, y_0, \dots, y_n$ , которого рассмотренный способ не даёт. Кроме того, при значительном количестве уз-

лов построение интерполянта посредством решения системы уравнений невыгодно в вычислительном отношении. Помимо того, что решение систем линейных уравнений само по себе не является тривиальной задачей, система (2.7) с матрицей Вандермонда оказывается весьма чувствительной к возмущениям данных или, как принято говорить, *плохо обусловленной* (см. § 1.7; конкретные числовые оценки чувствительности решения системы (2.7) можно найти в § 3.4а–3.4б). Поэтому получаемый на этом пути интерполяционный полином может обладать большой погрешностью.

Систему линейных уравнений (2.7) можно попытаться решить в общем виде с помощью правила Крамера, пользуясь удобным выражением для определителя матрицы Вандермонда в знаменателе и разложением определителей в числителе по столбцу свободных членов  $(y_0, y_1, \dots, y_n)^\top$ . Этот путь может быть успешно пройдён, хотя и требует громоздких алгебраических преобразований.

На самом деле для интерполяционного полинома нам нечасто требуется знать именно каноническую форму (2.6). Для большинства практических целей достаточно иметь какое-либо конструктивное представление интерполяционного полинома, позволяющее вычислять его значения в любой наперёд заданной точке.

Для отыскания такого представления заметим, что при фиксированных узлах  $x_0, x_1, \dots, x_n$  результат алгебраической интерполяции линейным образом зависит от значений  $y_0, y_1, \dots, y_n$ . Более точно, если полином  $P(x)$  решает задачу интерполяции по значениям  $y = (y_0, y_1, \dots, y_n)$ , а полином  $Q(x)$  решает задачу интерполяции с теми же узлами по значениям  $z = (z_0, z_1, \dots, z_n)$ , то для любых чисел  $\alpha, \beta \in \mathbb{R}$  полином  $\alpha P(x) + \beta Q(x)$  решает задачу интерполяции для значений  $\alpha y + \beta z = (\alpha y_0 + \beta z_0, \alpha y_1 + \beta z_1, \dots, \alpha y_n + \beta z_n)$  на той же совокупности узлов.<sup>4</sup>

Отмеченным свойством линейности можно воспользоваться для решения задачи интерполяции «по частям», которые удовлетворяют отдельным интерполяционным условиям в заданных узлах, а затем собрать эти части воедино. Именно, будем искать интерполяционный полином в виде

$$P_n(x) = \sum_{i=0}^n y_i \phi_i(x), \quad (2.10)$$

---

<sup>4</sup>Сказанное можно выразить словами «оператор интерполяирования линеен». В действительности он даже является проектором, и эти наблюдения служат началом большого и плодотворного направления теории приближения функций.

где  $\phi_i(x)$  — полином степени  $n$ , такой что

$$\phi_i(x_j) = \delta_{ij} = \begin{cases} 0 & \text{при } i \neq j, \\ 1 & \text{при } i = j, \end{cases} \quad (2.11)$$

$i, j = 0, 1, \dots, n$ , и посредством  $\delta_{ij}$  обозначен символ Кронекера. Тогда полином  $y_i \phi_i(x)$ ,  $i = 0, 1, \dots, n$ , имеет степень  $n$  и решает задачу интерполяции набора значений  $(0, \dots, 0, y_i, 0, \dots, 0)$  по узлам  $x_0, x_1, \dots, x_n$ . Как следствие, полином  $P_n(x)$ , задаваемый представлением (2.10), действительно удовлетворяет условиям задачи.

Найдём теперь  $\phi_i(x)$ . Коль скоро этот полином зануляется в точках  $x_0, \dots, x_{i-1}, x_{i+1}, \dots, x_n$ , то он имеет вид

$$\phi_i(x) = K_i (x - x_0) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n). \quad (2.12)$$

При этом  $K_i$  должен быть некоторым числовым множителем, так как в правой части равенства (2.12) произведение  $n$  линейных по  $x$  членов уже даёт полином степени  $n$ . Для определения этого множителя подставим в выражение (2.12) значение аргумента  $x = x_i$ , откуда в силу (2.11) получается

$$K_i (x_i - x_0) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n) = 1.$$

Следовательно,

$$K_i = \frac{1}{(x_i - x_0) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)},$$

и потому

$$\phi_i(x) = \frac{(x - x_0) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n)}{(x_i - x_0) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)}. \quad (2.13)$$

Полиномы  $\phi_i(x)$ ,  $i = 0, 1, \dots, n$ , называют *базисными полиномами Лагранжа*, а иногда также *полиномами влияния  $i$ -го узла* (последний термин объясняется условием (2.11)). В целом из (2.10) следует, что задачу алгебраической интерполяции решает полином

$$\begin{aligned} P_n(x) &= \sum_{i=0}^n y_i \frac{\prod_{j \neq i} (x - x_j)}{\prod_{j \neq i} (x_i - x_j)} = \\ &= \sum_{i=0}^n y_i \frac{(x - x_0) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n)}{(x_i - x_0) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)}. \end{aligned} \quad (2.14)$$

Его называют *интерполяционным полиномом в форме Лагранжа* или просто *интерполяционным полиномом Лагранжа*.

**Пример 2.2.2** Рассмотрим внимательнее базисные полиномы Лагранжа, эти элементарные кирпичики, из которых собирается общее решение задачи алгебраической интерполяции. Пусть, к примеру, задан набор простых узлов  $\{1, 2, 3, 4, 5, 6\}$ . Интерполяционный полином по нему имеет пятую степень, как и все базисные полиномы Лагранжа (2.13).

Первый и третий базисные полиномы,  $\phi_0(x)$  и  $\phi_2(x)$ , построенные относительно узлов  $x_0 = 1$  и  $x_2 = 3$ , выглядят следующим образом:

$$\phi_0(x) = -\frac{1}{120}(x-2)(x-3)(x-4)(x-5)(x-6), \quad (2.15)$$

$$\phi_2(x) = \frac{1}{12}(x-1)(x-2)(x-4)(x-5)(x-6), \quad (2.16)$$

и их графики изображены на рис. 2.6.

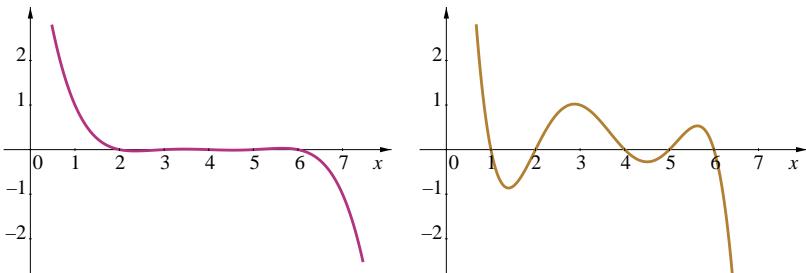


Рис. 2.6. Графики базисных полиномов Лагранжа (2.15) и (2.16)

Как и требуется условиями (2.11), эти базисные полиномы в узлах 1 и 3, соответственно, принимают единичные значения и зануляются в остальных узлах. Но интересно также их поведение вне узлов. Видно, что за пределами интервала расположения узлов значения обоих полиномов  $\phi_0$  и  $\phi_2$  неограниченно увеличиваются по абсолютной величине. Кроме того, между узлами базисный полином  $\phi_2$  также принимает заметные значения. Эти колебания значений полиномов (называемые также *осциляциями*) увеличиваются с ростом их степени. ■

Задача алгебраической интерполяции полностью решается с помощью полинома (2.14), который находит широчайшее применение в вычислительной практике. Но в некоторых случаях и он оказывается не

совсем удобными. Дело в том, что каждый из базисных полиномов Лагранжа  $\phi_i(x)$  зависит от всех узлов интерполяции сразу. По этой причине, работая с изменяющимся набором узлов, мы каждый раз должны будем перевычислять все  $\phi_i(x)$ . Иными словами, при смене набора узлов интерполяции полином Лагранжа претерпевает большое изменение и должен быть перевычислен заново.

Нельзя ли найти такую форму интерполяционного полинома, которая изменялась бы незначительно при небольших изменениях в наборе узлов интерполяции? Этот вопрос решается с помощью интерполяционного полинома в форме Ньютона, и для его построения нам будет необходима новая техника, основанная на понятии разделённой разности от функции.

## 2.2г Разделённые разности и их свойства

**Определение 2.2.2** Пусть даны функция  $f$  и несовпадающие точки  $x_0, x_1, \dots, x_n$  из её области определения, в которых функция принимает значения  $f(x_0), f(x_1), \dots, f(x_n)$ . Разделёнными разностями функции  $f$ , обозначаемыми  $f^{\wedge}(x_i, x_{i+1})$ , называются отношения

$$f^{\wedge}(x_i, x_{i+1}) := \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i}, \quad (2.17)$$

$i = 0, 1, \dots, n-1$ . Их называют также разделёнными разностями первого порядка.

Разделённые разности второго порядка — это величины

$$f^{\wedge}(x_i, x_{i+1}, x_{i+2}) := \frac{f^{\wedge}(x_{i+1}, x_{i+2}) - f^{\wedge}(x_i, x_{i+1})}{x_{i+2} - x_i}, \quad (2.18)$$

$i = 0, 1, \dots, n-2$ , которые являются разделёнными разностями от разделённых разностей. Аналогичным образом вводятся разделённые разности высших порядков: по определению разделённая разность  $k$ -го порядка от функции  $f$  есть

$$f^{\wedge}(x_i, x_{i+1}, \dots, x_{i+k}) := \frac{f^{\wedge}(x_{i+1}, \dots, x_{i+k}) - f^{\wedge}(x_i, \dots, x_{i+k-1})}{x_{i+k} - x_i}, \quad (2.19)$$

$i = 0, 1, \dots, n-k$ , т. е. она равна разделённой разности от разделённых разностей предыдущего  $(k-1)$ -го порядка. Порядок разделённой

разности нашими обозначениями специально не указывается; он определяется числом аргументов разделённой разности и на единицу его меньше. Для удобства и единообразия можно считать, что сами значения функции являются разделёнными разностями нулевого порядка, т. е.  $f^\angle(x_i) = f(x_i)$ ,  $i = 0, 1, \dots, n$ .

Разделённые разности введены в математику в начале XVIII века И. Ньютона, хотя сам термин для них установился уже в XIX веке. В математических текстах для разделённых разностей функции  $f$  по точкам  $x_i, x_{i+1}, \dots, x_{i+k}$  часто применяется идущее от классиков обозначение  $f[x_i, x_{i+1}, \dots, x_{i+k}]$ , а иногда используется даже маловыразительное  $f(x_i, x_{i+1}, \dots, x_{i+k})$ .

Разделённые разности можно определять не только для функций непрерывного аргумента, но и для функций дискретного аргумента, или, иначе говоря, для набора значений  $y_0, y_1, \dots, y_n$ , соответствующего узлам  $x_0, x_1, \dots, x_n$ . Назовём для него разделённой разностью первого порядка между узлами  $x_i$  и  $x_{i+1}$  величину

$$(y_i, y_{i+1})^\angle := \frac{y_{i+1} - y_i}{x_{i+1} - x_i}.$$

Разделённой разностью  $k$ -го порядка значений  $y_i, y_{i+1}, \dots, y_{i+k}$  по узлам  $x_i, x_{i+1}, \dots, x_{i+k}$  называется величина

$$(y_i, y_{i+1}, \dots, y_{i+k})^\angle := \frac{(y_{i+1}, \dots, y_{i+k})^\angle - (y_i, \dots, y_{i+k-1})^\angle}{x_{i+k} - x_i},$$

$i = 0, 1, \dots, n - k$ , т. е. разделённая разность от разделённых разностей предшествующего  $(k - 1)$ -го порядка. Обозначение  $(y_i, y_{i+1}, \dots, y_{i+k})^\angle$  не содержит явного указания на узлы  $x_i, x_{i+1}, \dots, x_{i+k}$ , относительно которых рассматривается набор  $(y_i, y_{i+1}, \dots, y_{i+k})$ , так что наличие определённых заданных узлов здесь подразумевается.

Отметим, что в определении разделённых разностей, вообще говоря, не накладывается никаких условий на взаимное расположение точек  $x_0, x_1, \dots, x_n$ . В частности, совсем не обязательно, чтобы  $x_i < x_{i+1}$ . Понятию разделённой разности от функции непрерывного аргумента можно придать смысл для случая совпадающих узлов  $x_i = x_{i+1}$ , если понимать его как результат предельного перехода при  $x_i \rightarrow x_{i+1}$ . Тогда разделённая разность, очевидно, превращается в производную от функции (см. подробности, к примеру, в книгах [20, 28]).

Нетрудно видеть геометрический смысл разделённой разности первого порядка. Будучи отношением приращения функции к прираще-

нию её аргумента, она даёт угловой коэффициент (тангенс угла наклона к оси абсцисс) секущей графика функции  $y = f(x)$ , которая взята между точками с аргументами  $x_i$  и  $x_{i+1}$  (рис. 2.7). В общем случае разделённая разность функции — это «средняя скорость» её изменения на рассматриваемом интервале, в отличие от «мгновенной скорости» изменения функции в точке, которая равна производной  $f'(x)$ .

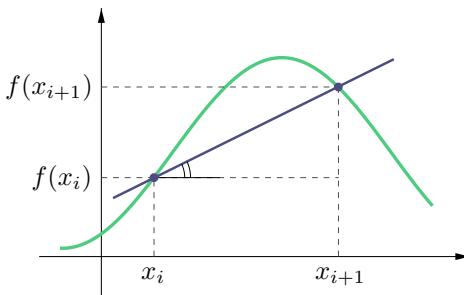


Рис. 2.7. Иллюстрация смысла разделённых разностей как углового коэффициента секущей графика функции

Если  $\check{x}$  — какая-то фиксированная точка, то для любой другой точки  $x$  имеет место равенство

$$f(x) = f(\check{x}) + f'(\check{x}, x)(x - \check{x}),$$

аналогичное формуле Тейлора, в которой удержаны лишь члены первого порядка. Но в отличие от формулы Тейлора выписанное равенство — абсолютно точное и не имеет никаких остаточных членов. Заметим также, что разделённую разность иногда называют *наклоном* функции между заданными точками [15]. Разделённые разности-наклоны могут быть определены для функций многих переменных и даже для операторов, действующих из одного абстрактного пространства в другое. Интересно, что в начале XX века для обозначения этой конструкции использовался также термин «подъём функции» [92].

Для фиксированного набора узлов численные значения разделённых разностей любой функции нетрудно вычислить согласно определениям (2.17), (2.18) или по формуле (2.21). Ещё один способ вычисления разделённых разностей — это алгоритмическое (автоматическое) дифференцирование, кратко рассматриваемое в § 2.9.

Операция взятия разделённой разности является линейной: для любых функций  $f, g$  и для любых скаляров  $\alpha, \beta$  справедливо

$$(\alpha f + \beta g)^\angle = \alpha f^\angle + \beta g^\angle \quad (2.20)$$

при одинаковых аргументах разделённых разностей. Это очевидно следует из определения для разделённой разности первого порядка, а для разделённых разностей высших порядков доказывается несложной индукцией по величине порядка. То же самое верно и для разделённых разностей от наборов значений по одним и тем же узлам:

$$(\alpha(y_i, \dots, y_{i+k}) + \beta(z_i, \dots, z_{i+k}))^\angle = \alpha(y_i, \dots, y_{i+k})^\angle + \beta(z_i, \dots, z_{i+k})^\angle.$$

Полезно иметь в виду, что любая разделённая разность от постоянной функции — тождественно нулевая.

**Предложение 2.2.1** *Имеет место представление*

$$f^\angle(x_i, x_{i+1}, \dots, x_{i+k}) = \sum_{j=i}^{i+k} \frac{f(x_j)}{\prod_{\substack{l=i \\ l \neq j}}^{i+k} (x_j - x_l)}. \quad (2.21)$$

Для разделённой разности от набора значений  $(y_0, y_1, \dots, y_n)$  по узлам  $x_0, x_1, \dots, x_n$  аналогичная формула выглядит следующим образом:

$$(y_i, y_{i+1}, \dots, y_{i+k})^\angle = \sum_{j=i}^{i+k} \frac{y_j}{\prod_{\substack{l=i \\ l \neq j}}^{i+k} (x_j - x_l)}. \quad (2.22)$$

**Доказательство.** Оно проводится индукцией по порядку  $k$  разделённой разности. Мы выпишем ниже подробные выкладки лишь для разделённых разностей от функций, так как для разделённых разностей от набора значений доказательство совершенно аналогично.

При  $k = 1$  доказываемая формула, как нетрудно проверить, совпадает с определением разделённой разности первого порядка.

Пусть предложение уже доказано для некоторого положительного целого  $k$ . Тогда по определению разделённой разности  $k + 1$ -го порядка

будем иметь

$$f'(x_i, x_{i+1}, \dots, x_{i+k+1}) =$$

$$= \frac{f(x_{i+1}, x_{i+2}, \dots, x_{i+k+1}) - f(x_i, x_{i+1}, \dots, x_{i+k})}{x_{i+k+1} - x_i} =$$

$$= \frac{1}{x_{i+k+1} - x_i} \cdot \left( \sum_{j=i+1}^{i+k+1} \frac{f(x_j)}{\prod_{\substack{l=i+1 \\ l \neq j}}^{i+k+1} (x_j - x_l)} - \sum_{j=i}^{i+k} \frac{f(x_j)}{\prod_{\substack{l=i \\ l \neq j}}^{i+k} (x_j - x_l)} \right) =$$

согласно индукционному предположению

$$= \frac{f(x_{i+k+1})}{(x_{i+k+1} - x_i) \prod_{l=i+1}^{i+k} (x_{i+k+1} - x_l)} =$$

$$+ \frac{1}{x_{i+k+1} - x_i} \cdot \left( \sum_{j=i+1}^{i+k} \frac{f(x_j)}{\prod_{\substack{l=i+1 \\ l \neq j}}^{i+k+1} (x_j - x_l)} - \sum_{j=i+1}^{i+k} \frac{f(x_j)}{\prod_{\substack{l=i \\ l \neq j}}^{i+k} (x_j - x_l)} \right) -$$

$$- \frac{f(x_i)}{(x_{i+k+1} - x_i) \prod_{l=i+1}^{i+k} (x_i - x_l)} =$$

$$= \frac{f(x_{i+k+1})}{(x_{i+k+1} - x_i) \prod_{l=i+1}^{i+k} (x_{i+k+1} - x_l)} +$$

$$+ \frac{1}{x_{i+k+1} - x_i} \cdot \sum_{j=i+1}^{i+k} f(x_j) \cdot \left( \frac{1}{\prod_{\substack{l=i+1 \\ l \neq j}}^{i+k+1} (x_j - x_l)} - \frac{1}{\prod_{\substack{l=i \\ l \neq j}}^{i+k} (x_j - x_l)} \right) -$$

$$- \frac{f(x_i)}{(x_{i+k+1} - x_i) \prod_{l=i+1}^{i+k} (x_i - x_l)}$$

после выведения из-под скобок последнего слагаемого первой суммы и первого слагаемого второй суммы.

В полученное выражение члены с  $f(x_{i+k+1})$  и  $f(x_i)$  — первый и последний — входят по одному разу, причём их коэффициенты уже имеют тот вид, который утверждается в предложении. Для остальных членов коэффициент при  $f(x_j)$  будет равен

$$\begin{aligned} \frac{1}{x_{i+k+1} - x_i} \cdot \left( \frac{1}{\prod_{\substack{l=i+1 \\ l \neq j}}^{i+k+1} (x_j - x_l)} - \frac{1}{\prod_{\substack{l=i \\ l \neq j}}^{i+k} (x_j - x_l)} \right) = \\ = \frac{(x_j - x_i) - (x_j - x_{i+k+1})}{(x_{i+k+1} - x_i) \prod_{\substack{l=i \\ l \neq j}}^{i+k+1} (x_j - x_l)} = \frac{1}{\prod_{\substack{l=i \\ l \neq j}}^{i+k+1} (x_j - x_l)}, \end{aligned}$$

что и требовалось показать. ■

**Следствие.** Разделённая разность как функция узлов  $x_0, x_1, \dots, x_k$  — симметричная функция своих аргументов. Иными словами, она не изменяется при любой их перестановке. Это непосредственно следует из симметричного вида выражения, стоящего в правой части (2.21) или (2.22).

Иногда требуется знать выражения для разделённых разностей, как функций узлов. Как правило, их сложность в общем случае быстро возрастает с ростом порядка разделённой разности. Тем не менее в случае алгебраических полиномов выражения для разделённых разностей относительно просто получаются из выражений для исходной функции. Вспомним известную формулу элементарной алгебры

$$(x - y)(x^{n-1} + x^{n-2}y + \dots + xy^{n-2} + y^{n-1}) = x^n - y^n,$$

из которой следует, что

$$\frac{x^n - y^n}{x - y} = x^{n-1} + x^{n-2}y + \dots + xy^{n-2} + y^{n-1}. \quad (2.23)$$

Этот результат позволяет явно выписать разделённую разность для любой целой степени переменной. Для произвольного полинома далее можно воспользоваться свойством (2.20), т. е. линейностью разделённой разности.

**Пример 2.2.3** Вычислим разделённые разности от полинома  $g(x) = x^3 - 4x + 1$ .

Будем искать по отдельности разделённые разности от мономов, образующих  $g(x)$ . В силу (2.23) имеем

$$\frac{x_2^3 - x_1^3}{x_2 - x_1} = x_2^2 + x_2 x_1 + x_1^2.$$

Для линейного монома  $(-4x)$  разделённая разность находится тривиально и равна  $(-4)$ , а для константы 1 она равна нулю. Следовательно, в целом

$$g^\wedge(x_1, x_2) = x_2^2 + x_2 x_1 + x_1^2 - 4.$$

Вычислим вторую разделённую разность от  $g(x)$ :

$$\begin{aligned} g^\wedge(x_1, x_2, x_3) &= \frac{g^\wedge(x_2, x_3) - g^\wedge(x_1, x_2)}{x_3 - x_1} = \\ &= \frac{(x_3^2 + x_3 x_2 + x_2^2 - 4) - (x_2^2 + x_2 x_1 + x_1^2 - 4)}{x_3 - x_1} = \\ &= \frac{x_3^2 + (x_3 - x_1)x_2 - x_1^2}{x_3 - x_1} = x_1 + x_2 + x_3. \end{aligned}$$

Третья разделённая разность

$$\begin{aligned} g^\wedge(x_1, x_2, x_3, x_4) &= \frac{g^\wedge(x_2, x_3, x_4) - g^\wedge(x_1, x_2, x_3)}{x_4 - x_1} = \\ &= \frac{(x_2 + x_3 + x_4) - (x_1 + x_2 + x_3)}{x_4 - x_1} = \\ &= \frac{x_4 - x_1}{x_4 - x_1} = 1, \end{aligned}$$

т. е. является постоянной. Четвёртая и последующие разделённые разности от  $g(x)$  будут, очевидно, тождественно нулевыми функциями. ■

Как видим, взятие разделённой разности от алгебраического полинома уменьшает его степень на единицу, так что разделённые разности

порядка более  $n$  от полинома степени  $n$  равны нулю. Это следует в общем случае из формулы (2.23). Сделанное наблюдение демонстрирует глубокую аналогию между разделёнными разностями и производными: каждое применение к полиному операции дифференцирования так же последовательно уменьшает его степень на единицу. В действительности эта связь видна даже из определения разделённой разности первого порядка, которую можно рассматривать как «неполную производную», поскольку у неё отсутствует предельный переход одного аргумента к другому.

**Предложение 2.2.2** (связь разделённых разностей с производными) *Пусть  $f \in C^n[a, b]$ , т. е. функция  $f$  непрерывно дифференцируема  $n$  раз на интервале  $[a, b]$ , где расположены узлы  $x_0, x_1, \dots, x_n$ , и пусть  $\underline{x} = \min\{x_0, x_1, \dots, x_n\}$ ,  $\bar{x} = \max\{x_0, x_1, \dots, x_n\}$ . Тогда*

$$f^{\angle}(x_0, x_1, \dots, x_n) = \frac{1}{n!} f^{(n)}(\xi) \quad (2.24)$$

для некоторой точки  $\xi \in ]\underline{x}, \bar{x}[$ .

Для разделённых разностей первого порядка этот факт непосредственно следует из теоремы Лагранжа о среднем (о конечном приращении), согласно которой

$$f(x_{i+1}) - f(x_i) = f'(\xi) \cdot (x_{i+1} - x_i)$$

для некоторой точки  $\xi \in ]x_i, x_{i+1}[$ . Для общего случая доказательство предложения 2.2.2 будет приведено несколько позже, в § 2.2e.

Существует более точное (хотя и более громоздкое) интегральное представление для разделённых разностей, о котором можно подробно узнать в книгах [20, 74, 94]. Оно также часто используется в теории и приложениях интерполяции.

## 2.2д Интерполяционный полином Ньютона

Выведем теперь другую форму интерполяционного полинома, которая минимальным образом перестраивалась бы при смене набора узлов интерполяции.

Предполагая заданным набор узлов интерполяции  $x_0, x_1, \dots, x_n$ , обозначим через  $P_k(x)$  интерполяционный полином степени  $k$ , построенный по первым узлам  $x_0, x_1, \dots, x_k$ . В частности,  $P_0(x) = y_0 = f(x_0)$

— интерполяционный полином нулевой степени, построенный по одному узлу  $x_0$ . Тогда рассматриваемое нами требование на форму интерполяционного полинома влечёт равенство

$$P_k(x) = P_{k-1}(x) + D_k(x), \quad k = 1, 2, \dots, n.$$

Иными словами, выражение для интерполяционного полинома по  $k$  узлам должно включать выражение для полинома по  $k - 1$  узлам и ещё добавочный член  $D_k(x)$ , с помощью которого учитывается узел  $x_k$ . Тогда  $D_k(x) = P_k(x) - P_{k-1}(x)$ ,  $k = 1, 2, \dots, n$ . Для интерполяционного полинома по всему набору узлов получаем следующее очевидное тождество:

$$\begin{aligned} P_n(x) &= P_0(x) + \sum_{k=1}^n (P_k(x) - P_{k-1}(x)) = \\ &= P_0(x) + \sum_{k=1}^n D_k(x). \end{aligned} \tag{2.25}$$

Замечательность этого представления состоит в том, что при добавлении или удалении последних по номеру узлов интерполяции перестройке должны подвергнуться лишь те последние слагаемые сумм из (2.25), которые вовлекают эти изменяющиеся узлы. Первые слагаемые в (2.25) зависят только от первых узлов интерполяции и останутся неизменными.<sup>5</sup> Таким образом, стоящая перед нами задача окажется решённой, если будут найдены удобные и просто записываемые выражения для разностей  $P_k(x) - P_{k-1}(x)$ .

Заметим, что разность  $(P_k(x) - P_{k-1}(x))$  есть полином степени  $k$ , который обращается в нуль в узлах  $x_0, x_1, \dots, x_{k-1}$ , общих для  $P_k(x)$  и  $P_{k-1}(x)$ , где эти полиномы принимают одинаковые значения  $y_0, y_1, \dots, y_{k-1}$ . Поэтому должно быть

$$P_k(x) - P_{k-1}(x) = A_k(x - x_0)(x - x_1) \cdots (x - x_{k-1}),$$

где  $A_k$  — некоторая константа, так как произведение следующих за ней линейных множителей образует полином степени  $k$ . Для определения

---

<sup>5</sup> Следует помнить, что нумерация узлов является в значительной мере условной: она может не отражать реальный порядок узлов на вещественной оси и вообще назначаться по нашему усмотрению для удобства работы с интерполянтом.

$A_k$  вспомним, что по условию интерполяции  $P_k(x_k) = y_k$ . Следовательно,

$$A_k = \frac{y_k - P_{k-1}(x_k)}{(x_k - x_0)(x_k - x_1) \cdots (x_k - x_{k-1})} = \frac{y_k - P_{k-1}(x_k)}{\prod_{l=0}^{k-1} (x_k - x_l)}.$$

Подставляя далее вместо  $P_{k-1}(x)$  выражение для интерполяционного полинома в форме Лагранжа, нетрудно вывести, что

$$\begin{aligned} A_k &= \frac{1}{\prod_{l=0}^{k-1} (x_k - x_l)} \cdot \left( y_k - \sum_{j=0}^{k-1} y_j \frac{\prod_{\substack{l=0 \\ l \neq j}}^{k-1} (x_k - x_l)}{\prod_{\substack{l=0 \\ l \neq j}}^{k-1} (x_j - x_l)} \right) = \\ &= \frac{y_k}{\prod_{l=0}^{k-1} (x_k - x_l)} - \sum_{j=0}^{k-1} \left( \frac{1}{\prod_{l=0}^{k-1} (x_k - x_l)} y_j \frac{\prod_{\substack{l=0 \\ l \neq j}}^{k-1} (x_k - x_l)}{\prod_{\substack{l=0 \\ l \neq j}}^{k-1} (x_j - x_l)} \right) = \\ &= \frac{y_k}{\prod_{l=0}^{k-1} (x_k - x_l)} - \sum_{j=0}^{k-1} \frac{y_j}{(x_k - x_j) \prod_{\substack{l=0 \\ l \neq j}}^{k-1} (x_j - x_l)} = \\ &\quad \text{после сокращения произведений} \\ &= \frac{y_k}{\prod_{\substack{l=0 \\ l \neq k}}^k (x_k - x_l)} + \sum_{j=0}^{k-1} \frac{y_j}{(x_j - x_k) \prod_{\substack{l=0 \\ l \neq j}}^{k-1} (x_j - x_l)} = \\ &= \sum_{j=0}^k \frac{y_j}{\prod_{\substack{l=0 \\ l \neq j}}^k (x_j - x_l)} = (y_0, y_1, \dots, y_k)^{\angle} \end{aligned}$$

в силу предложения 2.2.1. Окончательно представление (2.25) принимает вид

$$\begin{aligned} P_n(x) &= \\ &= y_0 + (y_0, y_1)^\angle (x - x_0) + (y_0, y_1, y_2)^\angle (x - x_0)(x - x_1) + \quad (2.26) \\ &\quad + \dots + (y_0, y_1, \dots, y_n)^\angle (x - x_0)(x - x_1) \cdots (x - x_{n-1}). \end{aligned}$$

Для задачи интерполяирования заданной функции  $f$  аналогичное выражение для интерполяционного полинома имеет вид

$$\begin{aligned} P_n(x) &= \\ &= f(x_0) + f^\angle(x_0, x_1)(x - x_0) + f^\angle(x_0, x_1, x_2)(x - x_0)(x - x_1) + \quad (2.27) \\ &\quad + \dots + f^\angle(x_0, x_1, \dots, x_n)(x - x_0)(x - x_1) \cdots (x - x_{n-1}). \end{aligned}$$

Выражения в правых частях равенств (2.26) и (2.27) называются интерполяционным полиномом *в форме Ньютона*, или просто *интерполяционным полиномом Ньютона*. Они являются равносильными формами записи интерполяционного полинома, широко применяемыми на практике, и особенно в ситуациях, где использование формы Лагранжа по тем или иным причинам оказывается неудобным.

Представление (2.25), на основе которого мы конструировали интерполяционный полином Ньютона, может быть уточнено и конкретизировано следующим образом:

$$\begin{aligned} P_n(x) &= \\ &P_k(x) + f^\angle(x_0, x_1, \dots, x_{k+1})(x - x_0)(x - x_1) \cdots (x - x_k) + \quad (2.28) \\ &\quad + \dots + f^\angle(x_0, x_1, \dots, x_n)(x - x_0)(x - x_1) \cdots (x - x_{n-1}), \end{aligned}$$

для любого  $k$ , такого что  $0 \leq k \leq n - 1$ . Образно выражаясь, формула (2.28) показывает, как интерполяционные полиномы Ньютона разных степеней вложены друг в друга наподобие «матрёшек».

Пусть  $f$  — вещественная  $n$  раз непрерывно дифференцируемая функция. С учётом результата предложения 2.2.2, т. е. равенства

$$f^\angle(x_0, x_1, \dots, x_n) = \frac{1}{n!} f^{(n)}(\xi),$$

хорошо видно, что интерполяционный полином Ньютона для гладкой функции непрерывного аргумента является прямым аналогом извест-

ного в математическом анализе полинома Тейлора (формулы Тейлора)

$$f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!} (x - x_0)^2 + \dots + \frac{f^{(n)}(x_0)}{n!} (x - x_0)^n.$$

При этом аналогами степеней переменной  $(x - x_0)^k$  являются произведения  $(x - x_0)(x - x_1) \cdots (x - x_{k-1})$ , которые в случае равномерно расположенных и упорядоченных по возрастанию узлов  $x_0, x_1, \dots, x_{k-1}$  часто называют *обобщённой степенью* [9].

Практическое построение интерполяционного полинома Ньютона требует знания числовых значений всех разделённых разностей функции, и чаще всего наиболее удобно находить их по рекуррентным формулам (2.17)–(2.19).

Важнейший частный случай интерполяирования относится к равномерному расположению узлов, когда разность  $h_i = x_i - x_{i-1}$ ,  $i = 1, 2, \dots, n$ , называемая *шагом сетки*  $\{x_0, x_1, \dots, x_n\}$ , постоянна и не зависит от  $i$ , т. е.  $h_i = h = \text{const}$ . Тогда вычисление разделённых разностей решительно упрощается, сводясь к оперированию с так называемыми *конечными разностями*. По определению конечной разностью (иногда добавляют — первого порядка) от функции  $f$  в точке  $x$  называется величина

$$\Delta y = \Delta f(x) = f(x + h) - f(x).$$

В частности, для независимого аргумента  $\Delta x = h$ . Конечные разности второго порядка  $\Delta^2 f(x)$  — это конечные разности от конечных разностей, и далее рекуррентно.

Индукцией по порядку разделённых и конечных разностей нетрудно показать, что они связаны друг с другом соотношением

$$f^\zeta(x_0, x_1, \dots, x_k) = \frac{\Delta^k f(x_0)}{k! h^k}, \quad k = 1, 2, \dots$$

Как следствие, интерполяционный полином Ньютона для равномерно расположенных узлов принимает вид

$$\begin{aligned} P_n(x) &= \\ &= f(x_0) + \frac{1}{1!} \frac{\Delta f(x_0)}{h} (x - x_0) + \frac{1}{2!} \frac{\Delta^2 f(x_0)}{h^2} (x - x_0)(x - x_1) + \\ &\quad + \dots + \frac{1}{n!} \frac{\Delta^n f(x_0)}{h^n} (x - x_0)(x - x_1) \cdots (x - x_{n-1}). \end{aligned}$$

Таблица 2.1. Таблица конечных разностей функции

$x$	$y$	$\Delta y$	$\Delta^2 y$	...	$\Delta^n y$
$x_0$	$y_0$				
$x_1$	$y_1$	$\Delta y_0$	$\Delta^2 y_0$		
$x_2$	$y_2$	$\Delta y_1$	$\Delta^2 y_1$	...	$\Delta^n y_0$
$x_3$	$y_3$	$\Delta y_2$	$\Delta^2 y_2$	...	$\Delta^n y_1$
$x_4$	$y_4$	$\Delta y_3$	$\Delta^2 y_3$	...	$\Delta^n y_2$
⋮	⋮	⋮	⋮	⋮	⋮

Он особенно сильно похож на полином Тейлора, и это сходство тем более разительно, если мы вспомним одно из классических обозначений для производных  $k$ -го порядка:  $f^{(k)} = \frac{d^k f}{dx^k}$ .

Вычисление конечных и разделённых разностей таблично заданной функции удобно оформлять также в виде таблицы (табл. 2.1). В ней в дополнительных столбцах (третьем, четвёртом и т. д.), заполняемых последовательно один за другим слева направо, записываются числовые значения конечных или разделённых разностей. Каждая из них получается из двух значений предшествующего столбца, которые расположены выше и ниже её.

В заключение темы стоит отметить, что помимо форм Лагранжа и Ньютона для интерполяционного полинома существуют и другие формы, особенно удобные в различных конкретных приложениях задачи интерполяции. Это интерполяционные формулы Гаусса, Стирлинга, Бесселя и др., подробности о которых можно узнать, к примеру, в [9, 21, 74, 92].

## 2.2e Погрешность алгебраической интерполяции с простыми узлами

Задача интерполяции, успешно решённая в предшествующих разделах, часто находится в более широком контексте, описанном во введении к этой теме (стр. 67). Значения  $y_0, y_1, \dots, y_n$  даны не сами по себе,

а принимаются в узлах  $x_0, x_1, \dots, x_n$  некоторой реальной функцией непрерывного аргумента  $f(x)$ , свойства которой (хотя бы отчасти) известны. Насколько сильно построенный нами интерполянт отличается от функции  $f$  на всей области определения, в частности, вне узлов интерполяции? Именно это отличие понимается под «погрешностью интерполяции».

**Определение 2.2.3** Пусть дана задача интерполяирования функции  $f$  по некоторому набору узлов. Остаточным членом или остатком интерполяции в этой задаче называется функция  $R(f, x) = f(x) - g(x)$ , являющаяся разностью рассматриваемой функции  $f(x)$  и интерполирующей её функции  $g(x)$ .

**Предложение 2.2.3** Если точка  $z$  не совпадает ни с одним из узлов  $x_0, x_1, \dots, x_n$ , то в задаче алгебраической интерполяции функции  $f$  по этим узлам значение остаточного члена в точке  $z$  равно

$$R_n(f, z) = f^\angle(x_0, x_1, \dots, x_n, z) \cdot \omega_n(z), \quad (2.29)$$

где функция  $\omega_n$  определяется как

$$\omega_n(x) = (x - x_0)(x - x_1) \cdots (x - x_n).$$

**Доказательство.** Выпишем для  $f$  интерполяционный полином Ньютона  $(n+1)$ -й степени по узлам  $x_0, x_1, \dots, x_n, z$ . Согласно представлению (2.28)

$$P_{n+1}(x) = P_n(x) + f^\angle(x_0, x_1, \dots, x_n, z) (x - x_0)(x - x_1) \cdots (x - x_n),$$

где  $P_n(x)$  — полином Ньютона для узлов  $x_0, x_1, \dots, x_n$ . Подставляя в это соотношение значение  $x = z$ , получим

$$P_{n+1}(z) = P_n(z) + f^\angle(x_0, x_1, \dots, x_n, z) (z - x_0)(z - x_1) \cdots (z - x_n).$$

Но  $P_{n+1}(z) = f(z)$  по построению полинома  $P_{n+1}$ . Поэтому

$$\begin{aligned} R_n(f, z) &= f(z) - P_n(z) = \\ &= f^\angle(x_0, x_1, \dots, x_n, z) (z - x_0)(z - x_1) \cdots (z - x_n), \end{aligned}$$

что и требовалось. ■

Полученный результат позволяет точно находить численное значение погрешности алгебраического интерполирования в конкретных точках, но он не слишком пригоден для исследования поведения погрешности «в целом», на всём интервале интерполирования. Чтобы получить более удобные оценки для остаточного члена, можно воспользоваться предложением 2.2.2 о связи разделённых разностей и производных, и ниже мы дадим его строгое доказательство.

**Доказательство** предложения 2.2.2, т. е. равенства (2.24)

$$f^{\wedge}(x_0, x_1, \dots, x_n) = \frac{1}{n!} f^{(n)}(\xi)$$

для некоторой точки  $\xi \in ]\underline{x}, \bar{x}[$ .

Отметим прежде всего, что в (2.24) без какого-либо ограничения общности можно считать узлы  $x_0, x_1, \dots, x_n$  упорядоченными по возрастанию индекса, т. е.  $x_0 < x_1 < \dots < x_n$ , поскольку разделённая разность есть симметричная функция узлов, по которым она берётся. Обозначив

$$\theta(x) := f^{(n)}(x) - n! f^{\wedge}(x_0, x_1, \dots, x_n),$$

заметим, что предложение 2.2.2 и равенство (2.24) равносильны следующему утверждению: на  $]x_0, x_n[$  существует точка  $\xi$ , которая является нулём функции  $\theta(x)$ .

По точкам  $x_0, x_1, \dots, x_n$  построим для функции  $f(x)$  интерполяционный полином  $P_n(x)$ . Оказывается, что введённая выше функция  $\theta(x)$  есть  $n$ -я производная по  $x$  от остаточного члена интерполяции  $R_n(f, x) = f(x) - P_n(x)$ , т. е.

$$\theta(x) = f^{(n)}(x) - n! f^{\wedge}(x_0, x_1, \dots, x_n) = R_n^{(n)}(f, x).$$

В этом можно убедиться непосредственным дифференцированием равенства

$$R_n(f, x) = f(x) - P_n(x),$$

где интерполяционный полином  $P_n(x)$  выписан в форме Ньютона.

В самом деле, в выражении для интерполяционного полинома Ньютона только у разделённой разности  $n$ -го порядка  $f^{\wedge}(x_0, x_1, \dots, x_n)$  множитель является полиномом  $n$ -й степени со старшим членом  $x^n$ . Множители у остальных разделённых разностей — это полиномы меньших степеней от  $x$ , которые исчезнут при  $n$ -кратном дифференцировании,

тогда как от полинома  $n$ -й степени со старшим членом  $x^n$  после такого дифференцирования останется число  $n!$ .

По условию предложения 2.2.2 функция  $R_n(f, x)$  является  $n$  раз непрерывно дифференцируемой на  $[a, b]$  и, кроме того, обращается в нуль в  $n + 1$  различных точках — узлах интерполяции  $x_0, x_1, \dots, x_n$ . В силу известной из математического анализа теоремы Ролля производная  $R'_n(f, x)$  обязана зануляться внутри каждого из  $n$  интервалов  $[x_0, x_1], [x_1, x_2], \dots, [x_{n-1}, x_n]$ , т. е. она имеет  $n$  нулей.

Далее, повторяя те же рассуждения в отношении второй производной  $R''_n(f, x)$ , приходим к выводу, что она должна иметь на  $]x_0, x_n[$  не менее  $n - 1$  нулей. Аналогично для третьей производной  $R'''_n(f, x)$  и т. д. вплоть до  $R^{(n)}_n(f, x)$ , которая должна иметь на  $]x_0, x_n[$  хотя бы один нуль. Это и требовалось доказать. ■

**Теорема 2.2.2** Пусть  $f \in C^{n+1}[a, b]$ , т. е. функция  $f(x)$  непрерывно дифференцируема  $n + 1$  раз на интервале  $[a, b]$ . При её интерполяции по несовпадающим узлам  $x_0, x_1, \dots, x_n \in [a, b]$  с помощью полинома  $n$ -й степени остаточный член  $R_n(f, x)$  может быть представлен в виде

$$R_n(f, x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!} \cdot \omega_n(x), \quad (2.30)$$

где  $\xi(x)$  — некоторая точка, принадлежащая открытому интервалу  $]a, b[$  и зависящая от  $x$ , а  $\omega_n = (x - x_0)(x - x_1) \dots (x - x_n)$ .

**Доказательство.** Если  $x = x_i$  для одного из узлов интерполяции, то  $R_n(f, x) = 0$ , но в то же время и  $\omega_n(x) = 0$ . Поэтому в качестве  $\xi$  в этом случае можно взять любую точку из открытого интервала  $]a, b[$ .

Если же аргумент  $x$  остаточного члена не совпадает ни с одним из узлов интерполяции, то применяем предложение 2.2.3, в котором разделённую разность  $(n + 1)$ -го порядка по точкам  $x_0, x_1, \dots, x_n$  и  $x$  из  $[a, b]$  выражаем через  $(n + 1)$ -ю производную функции согласно результату предложения 2.2.2. ■

Выражение (2.30) было получено О.Л. Коши в первой половине XIX века [7], и потому его обычно называют *остаточным членом алгебраической интерполяции в форме Коши*. Другое выражение для этого остаточного члена, не использующее неизвестную точку  $\xi(x)$  и основанное на интегральном представлении разделённых разностей, можно найти, к примеру, в книгах [20, 94].

Обозначим

$$M_n = \max_{\xi \in [a, b]} |f^{(n)}(\xi)|$$

— максимум абсолютного значения  $n$ -й производной на рассматриваемом интервале. С помощью этой константы нетрудно выписать огрублённые оценки, вытекающие из (2.30) и полезные при практическом вычислении погрешности интерполяции —

$$|R_n(f, x)| \leq \frac{M_{n+1}}{(n+1)!} \cdot |\omega_n(x)| \quad (2.31)$$

или даже совсем простую

$$|R_n(f, x)| \leq \frac{M_{n+1}(b-a)^{n+1}}{(n+1)!}. \quad (2.32)$$

Если доступно явное выражение для  $(n+1)$ -й производной функции  $f$ , то для оценивания  $M_{n+1}$  можно воспользоваться, к примеру, интервальными методами, взяв какое-либо интервальное расширение для  $f^{(n+1)}(x)$  на  $[a, b]$  (см. § 1.6).

Отметим, что полученные выше оценки — (2.30) и её следствия (2.31) и (2.32) — становятся неприменимыми, если функция  $f$  имеет гладкость, меньшую  $n+1$ . В то же время представление погрешности интерполяции через разделённые разности в виде (2.29) или в интегральной форме справедливо для любых функций.

В представлении (2.30) поведение полинома  $\omega_n(x)$  при изменении  $x$  типично для полиномов с вещественными нулями вообще. Пусть, как и ранее,  $\underline{x} = \min\{x_0, x_1, \dots, x_n\}$ ,  $\bar{x} = \max\{x_0, x_1, \dots, x_n\}$ . Если аргумент  $x$  находится на интервале  $[\underline{x}, \bar{x}]$  расположения нулей  $x_0, x_1, \dots, x_n$  или «не слишком далёко» от него, то  $\omega_n(x)$  принимает относительно умеренные значения, так как формирующие это произведение множители  $(x - x_i)$ ,  $i = 0, 1, \dots, n$ , «не слишком сильно» отличаются от нуля. Если же значения аргумента  $x$  находятся на существенном удалении от нулей полинома  $\omega_n(x)$ , то его абсолютная величина и вместе с ней погрешность алгебраической интерполяции очень быстро растут. На рис. 2.8 изображён пример графика такого полинома нечётной (седьмой) степени.

В связи со сказанным полезно на качественном уровне различать два случая интерполяции. Если значения интерполируемой функции ищутся в точках, далёких от интервала узлов интерполяции, используют термин *экстраполяция*. Ей противопоставляется *интерполяция* в

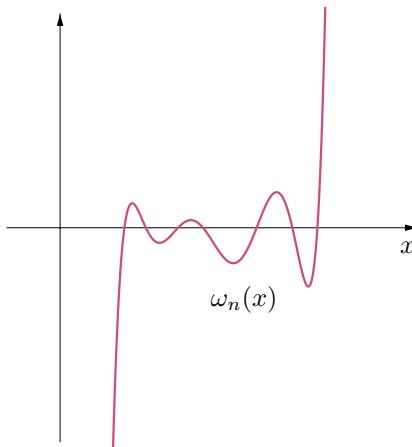


Рис. 2.8. Типичное поведение полинома  $\omega_n(x)$ : быстрый рост за пределами интервала узлов

узком смысле, когда значения функции восстанавливаются на интервале, где расположены узлы, или же вблизи от него. Из наших рассуждений следует, что экстраполяция, как правило, сопровождается существенными погрешностями и потому не стоит использовать её слишком широко.

В рассмотренной постановке задачи интерполяирования (§ 2.2а) расположение узлов считалось данным извне и фиксированным. Подобная ситуация характерна для тех практических задач, в которых, к примеру, измерения величины  $y_i$  могут осуществляться лишь в какие-то фиксированные моменты времени  $x_i$  либо в определённых выделенных точках пространства и т. п., то есть заданы каким-то внешним образом и не могут быть изменены по нашему желанию.

Но существуют задачи интерполяирования, в которых мы можем управлять выбором узлов. При этом естественно возникает вопрос о том, как сделать этот выбор наилучшим образом, чтобы погрешность интерполяирования была как можно меньшей. В наиболее общей формулировке эта задача является весьма трудной, и её решение существенно завязано на свойства интерполируемой функции  $f(x)$ . Но имеет смысл рассмотреть и упрощённую постановку, в которой на заданном интервале минимизируются значения полинома  $\omega_n(x)$ , тогда как множители

$f^\angle(x_0, x_1, \dots, x_n, z)$  и  $f^{(n+1)}(\xi(x))/(n+1)!$  в выражениях для остаточного члена (2.29) или (2.30) соответственно считаются огрублённо «приближёнными константами».

Фактически, ответ на поставленный вопрос сводится к подбору узлов  $x_0, x_1, \dots, x_n$  в пределах заданного интервала  $[a, b]$  так, чтобы полином  $\omega_n(x) = (x - x_0)(x - x_1) \dots (x - x_n)$  принимал «как можно меньшие значения» на  $[a, b]$ . Конкретный смысл, который вкладывается в это требование, может быть весьма различен, так как функция — полином  $\omega_n(x)$  в нашем случае — определяется своими значениями в бесконечном множестве аргументов, и малость одних значений функции может иметь место наряду с очень большими значениями при других аргументах (см., к примеру, рис. 2.30 из § 2.11). Ниже в § 2.3 мы рассмотрим ситуацию, когда «отклонение от нуля» понимается как равномерное (чебышёвское) расстояние (2.1) до нулевой функции, т. е. как максимум абсолютных значений функции на интервале. Это условие является одним из наиболее часто встречающихся в прикладных задачах.

## 2.2ж Тригонометрическая интерполяция

Задачей тригонометрической интерполяции, как уже отмечалось, называется задача построения для заданной функции  $f : \mathbb{R} \supset [a, b] \rightarrow \mathbb{R}$  интерполянта в виде тригонометрического полинома

$$Q_n(x) = \frac{a_0}{2} + \sum_{k=1}^n (a_k \cos(kx) + b_k \sin(kx)), \quad (2.33)$$

построенного по заданной сетке и интерполяционным данным. Тригонометрический полином является периодической функцией, с периодом  $2\pi$ , так что с его помощью интерполировать лучше всего тоже периодические функции с тем же периодом. Но по разным причинам бывает необходимо решать задачу тригонометрической интерполяции и для непериодических функций, заданных на каком-то вещественном интервале. В этом случае можно считать, например, что интерполируемая функция периодически продолжается за границы исходного интервала.

Присутствие одновременно синусов и косинусов в выражении (2.33) объясняется необходимостью сделать тригонометрические полиномы «более представительными» функциями, чем это могут обеспечить одни только синусы или одни косинусы. Всё синусы кратных аргументов

$\sin(kx)$  являются нечётными функциями, и потому их линейная комбинация также нечётна. Аналогично все косинусы кратных аргументов  $\cos(kx)$  являются чётными функциями, так что их линейная комбинация — чётная. По этой причине одни синусы или одни косинусы в выражении (2.33) порождают довольно специальную функцию, которая не сможет успешно интерполировать или приближать произвольные функции с более разнообразным характером изменения.

Для того чтобы сделать соизмеримыми период или, более общо, интервал характерного изменения интерполируемой функции и период тригонометрического полинома, обычно вводят в аргументы синусов и косинусов дополнительный масштабирующий множитель, так что вместо (2.33) рассматривается

$$Q_n(x) = \frac{a_0}{2} + \sum_{k=1}^n (a_k \cos(k\omega x) + b_k \sin(k\omega x)). \quad (2.34)$$

Параметр  $\omega$  называется в физике *циклической частотой* или *круговой частотой*. Подходящий выбор  $\omega$  — ответственное дело, которое должно выполняться ещё до решения собственно математической постановки задачи. Слагаемые интерполяционного полинома, представляющие синусы и косинусы кратных аргументов, часто называют *гармониками*.

Предположим, что на интервале  $[0, 2\pi]$  задана сетка с  $2n+1$  простыми узлами. Докажем существование и единственность решения задачи тригонометрической интерполяции по этим узлам, т. е. существование и единственность интерполяционного тригонометрического полинома вида (2.33), построенного по заданной сетке и интерполяционным данным.

Вспомним формулы Эйлера, дающие представление тригонометрических функций через экспоненту комплексного аргумента:

$$\cos(kx) = \frac{e^{ikx} + e^{-ikx}}{2}, \quad \sin(kx) = \frac{e^{ikx} - e^{-ikx}}{2i},$$

где  $e = 2.7182818\dots$  — число Эйлера (основание натуральных логарифмов),  $i = \sqrt{-1}$  — мнимая единица. С помощью этих формул тригонометрический полином можно преобразовать к виду

$$Q_n(x) = \sum_{k=-n}^n c_k e^{ikx}, \quad (2.35)$$

где

$$c_0 = \frac{a_0}{2}, \quad c_k = \frac{a_k - b_k i}{2}, \quad c_{-k} = \frac{a_k + b_k i}{2}, \quad k = 1, 2, \dots, n.$$

Обозначим узлы интерполяции через  $x_0, x_1, \dots, x_{2n}$ . Составив для получившегося полинома из экспонент (2.35) систему линейных алгебраических уравнений с неизвестными  $c_k$ ,  $k = -n, -n+1, \dots, n-1, n$ , удовлетворяющую интерполяционным данным по  $2n+1$  узлам, исследуем матрицу этой системы

$$\begin{pmatrix} e^{-ni x_0} & e^{-(n-1)i x_0} & \dots & e^{(n-1)i x_0} & e^{ni x_0} \\ e^{-ni x_1} & e^{-(n-1)i x_1} & \dots & e^{(n-1)i x_1} & e^{ni x_1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ e^{-ni x_{2n}} & e^{-(n-1)i x_{2n}} & \dots & e^{(n-1)i x_{2n}} & e^{ni x_{2n}} \end{pmatrix}.$$

Каков её определитель?

Вынося из строк выписанной матрицы последовательно множители  $e^{-ni x_0}, e^{-ni x_1}, \dots, e^{-ni x_{2n}}$  и вспоминая свойства определителя, можем заключить, что определитель матрицы равен

$$e^{-ni x_0} e^{-ni x_1} \dots e^{-ni x_{2n}} \cdot \det \begin{pmatrix} 1 & e^{ix_0} & \dots & e^{(2n-1)ix_0} & e^{2nix_0} \\ 1 & e^{ix_1} & \dots & e^{(2n-1)ix_1} & e^{2nix_1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & e^{ix_{2n}} & \dots & e^{(2n-1)ix_{2n}} & e^{2nix_{2n}} \end{pmatrix}.$$

В этом выражении в матрице под знаком определителя узнаётся матрица Вандермонда, которая неособенна в случае, когда её строки образованы степенями различных чисел.

Осталось лишь заметить, что переход от коэффициентов  $a_k, b_k$  в исходной форме тригонометрического полинома к коэффициентам  $c_k$  в (2.35) — это неособенное линейное преобразование, которое обратимо.

Проведённое доказательство конструктивно и фактически повторяет тот способ построения интерполянта, который был применён в § 2.26 для алгебраических полиномов. Сходство объясняется тем, что тригонометрическая интерполяция тоже относится к линейным методам интерполяции, так как тригонометрический полином (2.33) линейно зависит от своих коэффициентов  $a_k$  и  $b_k$ . На практике можно подставлять

значения узлов прямо в выражение (2.33) и приравнивать интерполяционным данным. Получится система линейных алгебраических уравнений, решив которую, найдём коэффициенты  $a_k$  и  $b_k$ .

Невыгодным качеством рассмотренного выше пути является необходимость решения системы уравнений для определения коэффициентов интерполяционного полинома. Оно преодолевается в другом способе построения тригонометрического интерполяционного полинома, который аналогичен описанному в § 2.2в подходу для полинома Лагранжа в алгебраической интерполяции.

Подробности построения тригонометрического аналога интерполяционного полинома Лагранжа в форме, предложенной О.Л. Коши, читатель может увидеть, к примеру, в книгах [2, 10, 58, 82]. Он выглядит следующим образом

$$P_n(x) = \sum_{i=0}^{2n} y_i \frac{\prod_{j \neq i} \sin\left(\frac{x - x_j}{2}\right)}{\prod_{j \neq i} \sin\left(\frac{x_i - x_j}{2}\right)}, \quad (2.36)$$

причём функции, задаваемые выражениями

$$\begin{aligned} & \frac{\prod_{j \neq i} \sin\left(\frac{x - x_j}{2}\right)}{\prod_{j \neq i} \sin\left(\frac{x_i - x_j}{2}\right)} = \\ & = \frac{\sin\left(\frac{x - x_0}{2}\right) \cdots \sin\left(\frac{x - x_{i-1}}{2}\right) \sin\left(\frac{x - x_{i+1}}{2}\right) \cdots \sin\left(\frac{x - x_n}{2}\right)}{\sin\left(\frac{x_i - x_0}{2}\right) \cdots \sin\left(\frac{x_i - x_{i-1}}{2}\right) \sin\left(\frac{x_i - x_{i+1}}{2}\right) \cdots \sin\left(\frac{x_i - x_n}{2}\right)}, \end{aligned}$$

$i = 0, 1, \dots, 2n$ , в самом деле являются тригонометрическими полиномами, т. е. имеют вид (2.33). Это следует из формул элементарной тригонометрии для взаимных произведений синусов и/или косинусов. В частности, эти функции — периодические. Они служат полными аналогами базисных интерполяционных полиномов Лагранжа (2.13), принимая единичное значение в узле  $x_i$  и зануляясь в остальных узлах.

**Пример 2.2.4** Рассмотрим задачу интерполяции степенной функции  $y = x^{0.4}$ , рассмотренную в примере 2.2.1.

В конструкциях этого параграфа используется нечётное количество узлов, и потому мы возьмём их семь штук — 0, 1, 2, 3, 4, 5, 6. Построив

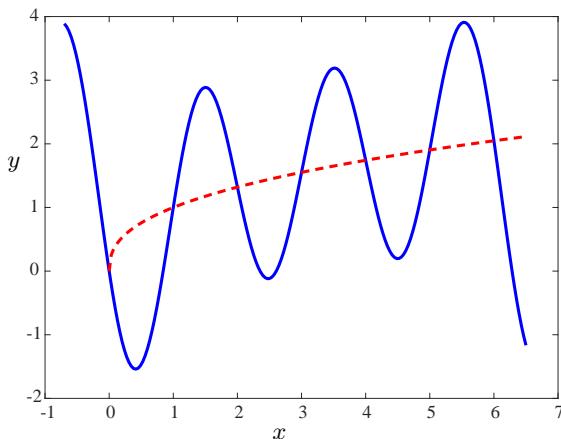


Рис. 2.9. Тригонометрическая интерполяция в случае, когда «характерный период» подобран неудачно

интерполяционный тригонометрический полином вида (2.36) и изобразив его на графике (рис. 2.9, где интерполянт представлен сплошной линией, а исходная функция — штриховой), мы обнаружим, что вне узлов он вообще никак не приближает интерполируемую функцию. Отличие исходной функции и её интерполянта шокирует и является разительным контрастом с поведением алгебраического интерполянта по тому же множеству узлов. Феномен можно интуитивно объяснить тем, что интервал интерполяции примерно равен периоду тригонометрического полинома и на этом интервале интерполируемая функция изменяется в целом довольно медленно. Как следствие, интерполяция её с помощью небольшого числа синусов и косинусов кратных аргументов, которые изменяются быстрее и сильнее, не приведёт к хорошему приближению.

Чтобы получить более точное приближение, необходимо увеличить степень тригонометрического полинома, увеличивая число узлов и число гармоник. На рис. 2.10 изображён результат интерполирования нашей функции на равномерной сетке из 21 узла. Погрешность стала меньше в сравнении с ситуацией на рис. 2.9. Нетрудно показать, что при увеличении количества равномерно расположенных узлов и соответствующих гармоник тригонометрический полином будет сходиться к ряду Фурье для интерполируемой функции по тригонометрической

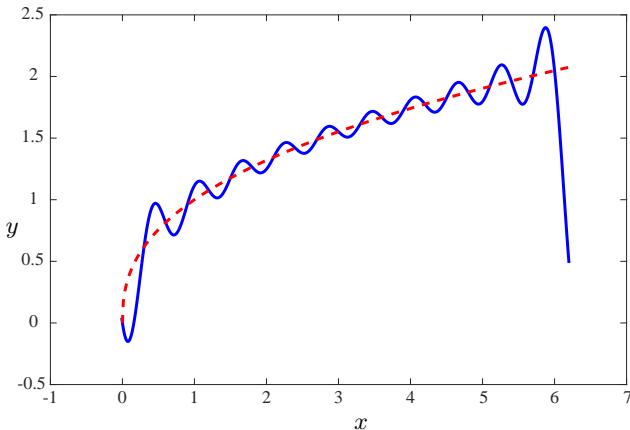


Рис. 2.10. Тригонометрическая интерполяция с увеличенным числом узлов

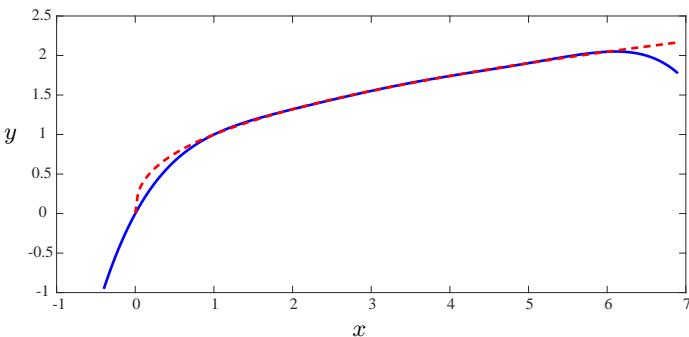


Рис. 2.11. Тригонометрическая интерполяция в случае, когда «характерный период» подобран удачно

системе (см. [40], том 3, или [74]). В свою очередь, в этой ситуации из общей теории рядов Фурье можно делать заключения о сходимости к интерполируемой функции в том или ином смысле.<sup>6</sup> Но этот путь часто нежелателен по техническим соображениям, так как увеличение числа узлов требует большей информации о функции и приводит к более сложному интерполянту.

<sup>6</sup>Эти вопросы весьма детально изложены, например, в книге [58], том 2.

Другой способ получения более качественных результатов тригонометрического интерполяирования может состоять в том, чтобы уменьшить циклическую частоту  $\omega$  в (2.34), если это допускается постановкой задачи. Тогда период синусов и косинусов увеличивается, а их изменение делается более медленным. Если, скажем, взять  $\omega = 1/5 = 0.2$  и снова построить интерполяционный полином (2.36), модифицированный соответствующим образом, то результаты такого тригонометрического интерполяирования для некоторых целей уже вполне удовлетворительны. Они представлены на рис. 2.11 (где, как и ранее, график функции  $y = x^{0.4}$  дан штриховой линией).

Аналогично случаю алгебраического интерполяирования хорошего приближения функции  $x^{0.4}$  вблизи нуля ожидать нельзя из-за её бесконечной производной. Но на оставшейся части интервала области определения качество приближения функции хорошее. Далее аргумента  $x = 7$  значения интерполянта «уютят» от значений исходной функции, и это уже неизбежно. ■

Помимо формулы Коши (2.36) существуют и другие формулы для интерполяционного тригонометрического полинома. Две таких формулы предложил К.Ф. Гаусс, и читатель может увидеть их в книге [39], стр. 266, или же в [58], том 2. Ещё одна альтернативная форма была дана Ш. Эрмитом [39, 88],

$$P_n(x) = \sum_{i=0}^n y_i \frac{\prod_{j \neq i} \sin(x - x_j)}{\prod_{j \neq i} \sin(x_i - x_j)}. \quad (2.37)$$

Фактически она соответствует форме Коши (2.36) с удвоенной циклической частотой. Отметим, что в этих альтернативных формулах тригонометрических интерполяционных полиномов, которые являются аналогами полинома Лагранжа, количество слагаемых не обязательно всегда нечётно. Как следствие, эти формулы можно успешно применять при любом числе узлов.

## 2.3 Полиномы Чебышёва

### 2.3а Определение и основные свойства

Полиномы Чебышёва — это семейство алгебраических полиномов, обозначаемых по традиции как  $T_n(x)$  и занумерованных неотрицатель-

ными целыми индексами  $n$ .<sup>7</sup> Они могут быть определены различными равносильными способами, и наиболее просто и наглядно их *тригонометрическое представление*:

$$T_n(x) = \cos(n \arccos x), \quad (2.38)$$

$x \in [-1, 1]$ ,  $n = 0, 1, 2, \dots$ . Как известно, всякий полином степени  $n$  однозначно определяется своими значениями в  $(n + 1)$  точках, а формулой (2.38) мы фактически задаём значения функции в бесконечном множестве точек из  $[-1, 1]$ . Поэтому если посредством (2.38) на  $[-1, 1]$  в самом деле задаются полиномы, то с помощью этой формулы они однозначно определяются на всей вещественной оси, а не только для значений аргумента  $x \in [-1, 1]$ .

**Предложение 2.3.1** *Функция  $T_n(x)$ , задаваемая формулой (2.38), — алгебраический полином степени  $n$ , и его старший коэффициент равен  $2^{n-1}$  при  $n \geq 1$ .*

**Доказательство.** Проведём его индукцией по номеру  $n$  полинома Чебышёва. При  $n = 0$  имеем  $T_0(x) = 1$ , при  $n = 1$  справедливо  $T_1(x) = x$ , так что база индукции установлена.

Для выполнения индукционного перехода заметим, что из известной тригонометрической формулы

$$\cos \alpha + \cos \beta = 2 \cos\left(\frac{\alpha + \beta}{2}\right) \cos\left(\frac{\alpha - \beta}{2}\right)$$

следует

$$\begin{aligned} \cos((n+1)\arccos x) + \cos((n-1)\arccos x) &= \\ &= 2 \cos(n\arccos x) \cos(\arccos x) = \\ &= 2x \cos(n\arccos x). \end{aligned}$$

Тогда в силу определения (2.38)

$$T_{n+1}(x) = 2x T_n(x) - T_{n-1}(x) \quad (2.39)$$

для любых  $n = 1, 2, \dots$

---

<sup>7</sup> С буквы «Т» начинаются немецкое (Tschebyschev) и французское (Tchebychev) написания фамилии П.Л. Чебышёва, открывшего эти полиномы в 1854 году.

Таким образом, если  $T_{n-1}(x)$  и  $T_n(x)$  являются полиномами степеней  $(n - 1)$  и  $n$  соответственно, то  $T_{n+1}(x)$  — тоже полином, степень которого на единицу выше степени  $T_n(x)$ , а старший коэффициент — в 2 раза больше. ■

Полученная в доказательстве рекуррентная формула (2.39) позволяет, отправляясь от  $T_0(x)$  и  $T_1(x)$ , последовательно выписывать явные алгебраические выражения для полиномов Чебышёва:

$$\begin{aligned} T_0(x) &= 1, \\ T_1(x) &= x, \\ T_2(x) &= 2x^2 - 1, \\ T_3(x) &= 4x^3 - 3x, \\ T_4(x) &= 8x^4 - 8x^2 + 1, \\ T_5(x) &= 16x^5 - 20x^3 + 5x, \\ &\dots \quad \dots \end{aligned} \tag{2.40}$$

По рекуррентной формуле (2.39) и следующим из неё явным выражениям (2.40) полиномы Чебышёва единообразно определяются для любых значений аргумента  $x$ . Графики первых полиномов Чебышёва можно увидеть на рис. 2.12.

Продолжая тему представления полиномов Чебышёва, заметим, что формула (2.38) справедлива в действительности при любых вещественных аргументах  $x$ , если для  $\arccos x$  допустить комплексные значения и, соответственно, рассматривать косинус от комплексного аргумента. В этих условиях можно показать, что

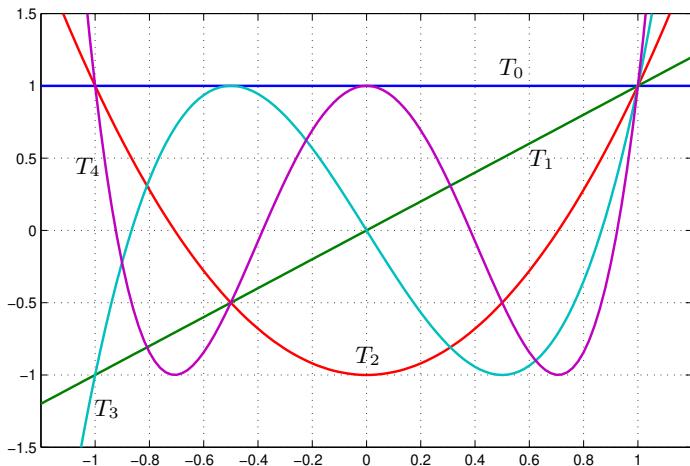
$$T_n(x) = \operatorname{ch}(n \operatorname{arch} x), \tag{2.41}$$

где  $\operatorname{ch} z = \frac{1}{2}(e^z + e^{-z})$  — гиперболический косинус, а  $\operatorname{arch}$  — обратная к нему функция. Определение (2.41) удобно применять для вещественных аргументов  $x$ , таких что  $x \geq 1$ , поскольку такова область определения  $\operatorname{arch}$  на вещественной оси.

Ещё одно полезное представление полиномов Чебышёва —

$$T_n(x) = \frac{1}{2} \left( (x + \sqrt{x^2 - 1})^n + (x - \sqrt{x^2 - 1})^n \right), \tag{2.42}$$

$n = 0, 1, 2, \dots$  (см. [7, 29, 52, 74]). Для него нетрудно установить эквивалентность тригонометрическому представлению (2.38), сделав замену

Рис. 2.12. Графики первых полиномов Чебышёва на интервале  $[-1.2, 1.2]$ 

$x = \cos t$  для некоторого  $t \in \mathbb{C}$ . Тогда из (2.42) следует, что

$$T_n(x) = \frac{1}{2}((\cos t + i \sin t)^n + (\cos t - i \sin t)^n).$$

В силу известной формулы Муавра  $(\cos t \pm i \sin t)^n = \cos nt \pm i \sin nt$ , и потому

$$T_n(x) = \frac{1}{2}(\cos nt + i \sin nt + \cos nt - i \sin nt) = \cos nt.$$

Наконец, поскольку  $t = \arccos x$ , немедленно получаем

$$T_n(x) = \cos(n \arccos x),$$

т. е. тригонометрическое представление (2.38).

Рассмотрим кратко основные свойства полиномов Чебышёва.

При чётном (нечётном)  $n$  полином Чебышёва  $T_n(x)$  есть чётная (нечётная) функция от  $x$ . Действительно, выражение для  $T_n(x)$  при чётном  $n$  содержит только чётные степени  $x$  (нуль считаем чётным числом), а при нечётном  $n$  — только нечётные степени  $x$ , что по индукции следует из рекуррентной формулы (2.39).

Найдём нули полиномов Чебышёва на интервале  $[-1, 1]$  вещественной оси. Исходя из тригонометрического представления (2.38) и фор-

мулы для нулей косинуса должно быть

$$n \arccos x = \frac{\pi}{2} + k\pi, \quad k \in \mathbb{Z}.$$

В этом равенстве  $k$  можно брать таким, чтобы область значений правой части не выходила за интервал  $[0, n\pi]$ , в котором принимает значения левая часть равенства. Следовательно, нулями полинома Чебышёва  $T_n(x)$  на  $[-1, 1]$  являются

$$\dot{x}_k = \cos \frac{(2k+1)\pi}{2n}, \quad k = 0, 1, \dots, n-1, \quad (2.43)$$

всего  $n$  штук. А поскольку в силу основной теоремы алгебры у полинома  $T_n(x)$  в поле комплексных чисел может быть не более  $n$  нулей, то можем заключить, что других нулей, отличных от (2.43), полином  $T_n(x)$  не имеет.

Итак, все нули полинома Чебышёва  $T_n(x)$  в самом деле находятся на интервале  $[-1, 1]$  и выражаются в виде (2.43). Расположение нулей полинома Чебышёва можно наглядно проиллюстрировать чертежом на рис. 2.13, где эти нули соответствуют абсциссам точек пересечения единичной окружности, имеющей центр в начале координат, с радиусами, откладываемыми через одинаковые доли развёрнутого угла в  $\pi$  радиан. Из этой иллюстрации хорошо видно, что нули полинома Чебышёва расположены существенно неравномерно: они сконцентрированы к концам интервала  $[-1, 1]$ , а в его средней части более разрежены.

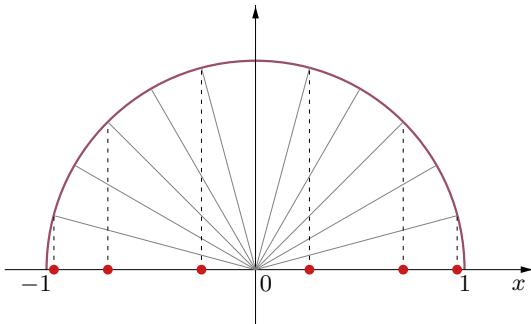


Рис. 2.13. Иллюстрация расположения нулей полинома Чебышёва шестой степени

Из тригонометрического представления (2.38) следует также, что

максимум модуля значений полинома Чебышёва на  $[-1, 1]$  равен 1, т. е.

$$\max_{x \in [-1, 1]} |T_n(x)| = 1.$$

Этот максимум достигается в точках  $x_s = \cos(s\pi/n)$ ,  $s = 0, 1, \dots, n$ , причём  $T_n(x_s) = (-1)^s$ ,  $s = 0, 1, \dots, n$ , так как внешний  $\cos$  в (2.38) должен достигать максимальных по модулю значений  $\pm 1$  в точках  $x_s$ , удовлетворяющих условию  $n \arccos x = s\pi$  при целочисленных  $s$ . Для  $x \in [-1, 1]$  область значений  $(n \arccos x)$  есть интервал  $[0, n\pi]$ , откуда вытекает, что  $s$  может быть равным  $0, 1, \dots, n$ .

Следующее свойство полиномов Чебышёва естественно основывается на предшествующих, и оно настолько важно, что мы оформим его как отдельное

**Предложение 2.3.2** *Среди полиномов степени  $n$ ,  $n \geq 1$ , со старшим коэффициентом, равным 1, полином  $\tilde{T}_n(x) := 2^{1-n} T_n(x)$  имеет на интервале  $[-1, 1]$  наименьшее равномерное отклонение от нуля. Иными словами, если  $Q_n(x)$  — полином степени  $n$  со старшим коэффициентом 1, то*

$$\max_{x \in [-1, 1]} |Q_n(x)| \geq \max_{x \in [-1, 1]} |\tilde{T}_n(x)| = 2^{1-n}. \quad (2.44)$$

Полиномы  $\tilde{T}_n(x) = 2^{1-n} T_n(x)$ , фигурирующие в предложении 2.3.2 и имеющие, согласно предложению 2.3.1, единичный старший коэффициент, называют *приведёнными полиномами Чебышёва*.

**Доказательство.** Предположим противное доказываемому, т. е. что для какого-то полинома  $Q_n(x)$ , имеющего старший коэффициент 1, справедливо неравенство

$$\max_{x \in [-1, 1]} |Q_n(x)| < \max_{x \in [-1, 1]} |\tilde{T}_n(x)|, \quad (2.45)$$

которое противоположно по смыслу неравенству (2.44). Тогда разность  $(\tilde{T}_n(x) - Q_n(x))$  есть полином степени не выше  $n - 1$ . В то же время в точках  $x_s = \cos(s\pi/n)$ ,  $s = 0, 1, \dots, n$ , доставляющих полиному Чебышёва максимумы модуля на  $[-1, 1]$ , должно выполняться

$$\begin{aligned} \operatorname{sgn} (\tilde{T}_n(x_s) - Q_n(x_s)) &= \operatorname{sgn} ((-1)^s 2^{1-n} - Q_n(x_s)) = \\ &= \operatorname{sgn} ((-1)^s 2^{1-n}) \text{ в силу (2.45)} = \\ &= (-1)^s. \end{aligned}$$

Как следствие, на каждом из открытых интервалов  $]x_s, x_{s+1}[$  полином  $(\tilde{T}_n(x) - Q_n(x))$  меняет знак, и потому в силу теоремы Больцано–Коши он обязан иметь нуль. Коль скоро это происходит для  $s = 0, 1, \dots, n-1$ , т. е. всего  $n$  раз, то полином  $(\tilde{T}_n(x) - Q_n(x))$  имеет  $n$  нулей на  $[-1, 1]$ . Степень этого полинома не превосходит  $n-1$ , так что полученные выводы можно примирить лишь при условии  $(\tilde{T}_n(x) - Q_n(x)) = 0$ , т. е. когда  $Q_n(x) = \tilde{T}_n(x)$ . Мы пришли к противоречию с допущением (2.45). ■

Доказанное свойство иногда называют *экстремальным свойством полиномов Чебышёва*, и оно имеет равносильные двойственные формулировки. Именно, среди всех многочленов заданной степени, значения которых на интервале  $[-1, 1]$  не превосходят по модулю 1, многочлен Чебышёва имеет наибольший старший коэффициент и наибольшее значение в любой точке за пределами  $[-1, 1]$ .

### 2.36 Применения полиномов Чебышёва

Доказательство предложения 2.3.2 использует тот факт, что полиномы рассматриваются на интервале  $[-1, 1]$ , лишь косвенным образом. Фактически мы опирались на свойство полиномов Чебышёва достигать своих знакопеременных экстремумов в  $n+1$  точках этого интервала. Если в качестве области определения полиномов необходимо взять интервал  $[a, b]$ , отличный от  $[-1, 1]$ , то линейной заменой переменной

$$y = \frac{1}{2}(b+a) + \frac{1}{2}(b-a)x \quad (2.46)$$

интервал  $[-1, 1]$  может быть преобразован в  $[a, b]$ . При этом обратное отображение  $[a, b] \rightarrow [-1, 1]$  задаётся формулой

$$x = \frac{2y - (b+a)}{(b-a)}, \quad (2.47)$$

а нулям полинома Чебышёва на  $[-1, 1]$  соответствуют тогда в интервале  $[a, b]$  точки

$$\dot{y}_k = \frac{1}{2}(b+a) + \frac{1}{2}(b-a) \cos \frac{(2k+1)\pi}{2n}, \quad k = 0, 1, \dots, n-1. \quad (2.48)$$

Свойство, аналогичное предложению 2.3.2, будет верно на интервале  $[a, b]$  для полинома, полученного из  $T_n(x)$  с помощью линейной замены переменных (2.47) и масштабирования (нормировки).

**Предложение 2.3.3** Если  $T_n(x)$  —  $n$ -й полином Чебышёва, то полином переменной  $y$ , задаваемый как

$$2^{1-2n} (b-a)^n \cdot T_n \left( \frac{2y-(b+a)}{b-a} \right), \quad (2.49)$$

имеет старший коэффициент 1 и на интервале  $[a, b]$  равномерно наименее уклоняется от нуля среди всех полиномов степени  $n$  со старшим коэффициентом 1. Чебышёвская норма этого полинома на интервале  $[a, b]$  равна  $2^{1-2n}(b-a)^n$ .

**Доказательство.** Первое утверждение предложения вытекает из того, что в результате замены переменной (2.47) из полинома  $n$ -й степени получается полином той же степени, но старший коэффициент приобретает дополнительный множитель  $2^n/(b-a)^n$ .

Далее, из свойств полиномов Чебышёва следует, что на  $[a, b]$  полином (2.49) достигает максимумов и минимумов, которые имеют чередующиеся знаки и одинаковые абсолютные значения  $2^{1-2n}(b-a)^n$  в точках

$$y_s = \frac{1}{2}(a+b) + \frac{1}{2}(b-a) \cos\left(\frac{s\pi}{n}\right), \quad s = 0, 1, \dots, n.$$

Они получаются с помощью линейного преобразования (2.46) из аргументов  $x_s = \cos(s\pi/n)$ ,  $s = 0, 1, \dots, n$ , доставляющих аналогичные максимумы модуля полиному Чебышёва на  $[-1, 1]$ . Дальнейшие рассуждения повторяют доказательство предложения 2.3.2, так как специфика интервала  $[-1, 1]$  там, фактически, никак не использовалась.

■

Обратимся к поставленной в конце § 2.2e задаче наиболее выгодного расположения узлов  $\{x_0, x_1, \dots, x_n\}$  алгебраического интерполянта степени  $n$  на заданном интервале  $[a, b]$ . Возьмём эти узлы как

$$x_k = \frac{1}{2}(b+a) + \frac{1}{2}(b-a) \cos\left(\frac{(2k+1)\pi}{2(n+1)}\right), \quad k = 0, 1, \dots, n, \quad (2.50)$$

т. е. нулям полинома вида (2.49), который получается в результате замены переменных (2.47) из полинома Чебышёва  $(n+1)$ -й степени  $T_{n+1}(x)$ . Тогда соответствующий полином

$$\omega_n(x) = (x - x_0)(x - x_1) \dots (x - x_n),$$

который фигурирует в формуле (2.30) для остаточного члена интерполяции, совпадёт с полиномом  $(n + 1)$ -й степени вида (2.49). При этом  $\omega_n(x)$  будет иметь наименьшее отклонение от нуля на  $[a, b]$  в равномерной (чебышёвской) метрике (2.1), и в смысле этой метрики погрешность интерполирования при прочих равных условиях (сформулированных на стр. 102) будет наименьшей возможной. Узлы интерполяции (2.50) называют *чебышёвскими узлами* на интервале  $[a, b]$ , а в совокупности они образуют *чебышёвскую сетку* на  $[a, b]$ .

Какой выигрыш достигается применением чебышёвских сеток? Ответ на этот вопрос зависит от того, с чем сравнивать. С точки зрения величины  $\|\omega_n\|_\infty$ , т. е. равномерного отклонения полинома  $\omega_n$  от нуля на заданном интервале  $[a, b]$ , наихудшими следует признать сетки, в которых все узлы «сбиваются» к одному из концов интервала,  $a$  или  $b$ . Тогда, если  $\tilde{x}$  становится равным другому концу интервала, то значение  $|\omega_n(\tilde{x})|$  близко к  $(b - a)^{n+1}$ . При этом разница между ними может быть сделана сколь угодно малой путём сближения всех узлов с концом интервала, и поэтому неравенство

$$\|\omega_n\|_\infty = \max_{x \in [a, b]} |\omega_n(x)| = |\omega_n(\tilde{x})| \leq (b - a)^{n+1} \quad (2.51)$$

является неулучшаемым (точным). В сравнении с этой оценкой та погрешность, которую обеспечивают чебышёвские сетки с  $n + 1$  узлом, т. е. оценка

$$\|\omega_n\|_\infty \leq 2^{-2n-1} (b - a)^{n+1},$$

меньше в  $2^{2n+1}$  раз. Но неравенство (2.51) соответствует очень экзотичным сеткам, которые совершенно нетипичны для практики. Что можно сказать о сравнении чебышёвских сеток с другими популярными сетками, например равномерной?

**Пример 2.3.1** Пусть на интервале  $[a, b]$  задана равномерная сетка, не включающая концы интервала:

$$x_k = a + \frac{b - a}{n + 1} k, \quad k = 1, 2, \dots, n.$$

Если  $\tilde{x} = a$  — левому концу рассматриваемого интервала, то

$$\tilde{x} - x_k = a - \left( a + \frac{b - a}{n + 1} k \right) = \frac{a - b}{n + 1} k, \quad k = 1, 2, \dots, n.$$

Тогда

$$\omega_n(\tilde{x}) = \omega_n(a) = \prod_{k=1}^n \frac{a-b}{n+1} k = \left( \frac{a-b}{n+1} \right)^n n!.$$

Это значение можно взять в качестве оценки снизу для  $\|\omega_n\|_\infty$  на рассматриваемой равномерной сетке (в действительности она точно равна этой норме).

Сравним полученную оценку со значением  $\|\omega_n\|_\infty$  по чебышёвской сетке из  $n$  узлов на  $[a, b]$ . Отношение этих величин равно

$$\frac{\left( \frac{a-b}{n+1} \right)^n n!}{2^{1-2n} (b-a)^n} = \frac{1}{2} \frac{n!}{\frac{(n+1)^n}{4^n}} = \frac{1}{2} \frac{n!}{\left( \frac{n+1}{n} \right)^n \frac{n^n}{4^n}} \approx \frac{1}{2e} \frac{n!}{\left( \frac{n}{4} \right)^n},$$

где  $e = 2.7182818\dots$  — число Эйлера, так как в силу известного из математического анализа факта

$$\left( \frac{n+1}{n} \right)^n \rightarrow e \text{ при } n \rightarrow \infty.$$

Вспомним теперь формулу Стирлинга для факториала [12, 40]:

$$n! \approx \sqrt{2\pi n} \left( \frac{n}{e} \right)^n \text{ при } n \rightarrow \infty.$$

Из неё следует, что отношение норм  $\|\omega_n\|_\infty$  для равномерной и чебышёвской сеток асимптотически равно

$$\frac{1}{2e} \frac{\sqrt{2\pi n} \left( \frac{n}{e} \right)^n}{\left( \frac{n}{4} \right)^n} = \frac{1}{e} \sqrt{\frac{\pi n}{2}} \left( \frac{4}{e} \right)^n$$

без зависимости от интервала  $[a, b]$ . В частности, с добавлением каждой новой точки погрешность для чебышёвской сетки в сравнении с погрешностью для равномерной уменьшается более чем в  $4/e \approx 1.47$  раза. На сетке из десяти точек выигрыш от использования чебышёвской сетки превосходит 70 раз.

Аналогичные оценки нетрудно получить также для равномерной сетки, включающей концы интервала. В качестве «пробной» точки  $\tilde{x}$ , которая используется для оценки нормы  $\|\omega_n\|_\infty$ , в этом случае можно взять середину подинтервала между концом интервала и соседним узлом. ■

Помимо интерполяции полиномы Чебышёва и их обобщения имеют и другие важные применения в различных задачах вычислительной математики и анализа [63, 69, 83]. Полиномы Чебышёва образуют систему ортогональных функций относительно интегрального скалярного произведения на интервале  $[-1, 1]$  с весом  $(1 - x^2)^{-1/2}$  (см. § 2.11). По этой причине очень важное значение имеют, к примеру, разложения функций в ряды Фурье по полиномам Чебышёва.

## 2.3в Обусловленность задачи алгебраической интерполяции

Предположим, что при алгебраическом интерполяции функции её значения в узлах вычисляются с некоторой погрешностью. Как она отразится на значениях интерполяционного полинома? Ответ на этот вопрос, вообще говоря, сильно зависит от формы интерполяционного полинома, от вида его выражения. Ниже мы рассмотрим интерполяционный полином в форме Лагранжа.

Пусть задана совокупность узлов интерполяции  $x_0, x_1, \dots, x_n$ , но вместо точных значений интерполируемой функции в узлах, т. е.  $y_i = f(x_i)$ , имеются их приближённые значения  $\tilde{y}_i$  с общей абсолютной погрешностью  $\varepsilon$ , так что

$$|\tilde{y}_i - y_i| \leq \varepsilon, \quad i = 0, 1, \dots, n.$$

Тогда вместо точного интерполяционного полинома  $P_n(x)$  получим полином

$$\tilde{P}_n(x) = \sum_{i=0}^n \tilde{y}_i \phi_i(x),$$

где, как и прежде,

$$\phi_i(x) = \frac{\prod_{j \neq i} (x - x_j)}{\prod_{j \neq i} (x_i - x_j)}, \quad i = 0, 1, \dots, n,$$

— базисные интерполяционные полиномы Лагранжа. Абсолютная погрешность значения приближённого интерполяционного полинома  $\tilde{P}_n$

в точке  $x$  равна, следовательно,

$$\begin{aligned} |\tilde{P}_n(x) - P_n(x)| &= \left| \sum_{i=0}^n (\tilde{y}_i - y_i) \phi_i(x) \right| \leq \\ &\leq \sum_{i=0}^n |\tilde{y}_i - y_i| |\phi_i(x)| \leq \varepsilon \sum_{i=0}^n |\phi_i(x)|. \end{aligned}$$

В целом для интервала  $[a, b]$  равномерное отклонение приближённого интерполяционного полинома  $\tilde{P}_n$  от точного может быть оценено как

$$\max_{x \in [a, b]} |\tilde{P}_n(x) - P_n(x)| \leq \varepsilon \max_{x \in [a, b]} \sum_{i=0}^n |\phi_i(x)|.$$

Величину, стоящую множителем при  $\varepsilon$  в правой части выписанного неравенства, можно рассматривать как коэффициент усиления ошибки в интерполяционных данных. Помимо оценивания погрешностей интерполяции она возникает при решении многих других математических вопросов и имеет собственное имя.

**Определение 2.3.1** Пусть заданы интервал  $[a, b] \subset \mathbb{R}$  и набор узлов (сетка)  $x_0, x_1, \dots, x_n$  на нём. Величина

$$\Lambda_n = \max_{x \in [a, b]} \sum_{i=0}^n |\phi_i(x)|,$$

где  $\phi_i(x)$ ,  $i = 0, 1, \dots, n$ , — базисные интерполяционные полиномы Лагранжа, называется  $n$ -й константой Лебега.<sup>8</sup>

Константа Лебега  $\Lambda_n$  существенно зависит от расположения узлов на интервале интерполирования. Следующий классический результат, полученный Г. Фабером и С.Н. Бернштейном, показывает, что константы Лебега неизбежно должны расти при увеличении числа узлов интерполяции, хотя скорость этого роста — довольно скромная.

**Теорема 2.3.1** Для любой бесконечной треугольной матрицы узлов из заданного интервала интерполяции справедливо неравенство

$$\Lambda_n > \frac{1}{8\sqrt{\pi}} \ln n.$$

---

<sup>8</sup>Иногда говорят «интерполяционной константой Лебега», так как существуют также «константы Лебега», относящиеся к сходимости рядов Фурье.

Доказательство можно увидеть, к примеру, в [6].

Но для конкретных сеток, применяемых на практике, оценка констант Лебега может расти с увеличением числа узлов чрезвычайно быстро. Этой неприятной особенностью обладают, в частности, популярные и удобные равномерные сетки.

**Теорема 2.3.2** *Для последовательности равномерных сеток на заданном интервале интерполяции справедливо неравенство*

$$\frac{1}{8n^{3/2}} 2^n \leq \Lambda_n \leq \frac{1}{2} 2^n, \quad n \geq 2.$$

Доказательство можно увидеть в книгах [6, 71], а для левого неравенства — в [38].

**Теорема 2.3.3** *Для последовательности чебышёвских сеток на заданном интервале интерполяции справедливо неравенство*

$$\Lambda_n \leq 8 + \frac{4}{\pi} \ln n.$$

Результат теоремы 2.3.3 принадлежит С.Н. Бернштейну, а его доказательство можно увидеть в [8] и для несколько упрощённой формулировки — в [38]. Фактически, теорема 2.3.3 означает, что в асимптотическом смысле чебышёвские сетки являются наилучшими, обеспечивая порядок роста констант Лебега, который диктуется теоремой 2.3.1 и принципиально не может быть улучшен.

Как видим, чебышёвские сетки не только уменьшают погрешность алгебраического интерполирования, но и обеспечивают лучшую обусловленность задачи, т. е. меньшую чувствительность решения по отношению к возмущениям в данных.

Константы Лебега, введённые в связи с анализом обусловленности задачи алгебраического интерполирования, оказываются также полезными при оценке погрешности решения этой задачи. Подробности интересующийся читатель может увидеть, к примеру, в книгах [6, 8, 71]. Этот способ имеет то преимущество перед оценкой О.Л. Коши (2.30), что оценивание погрешности не опирается на гладкость интерполируемой функции и неравенства для её высших производных.

## 2.4 Алгебраическая интерполяция с кратными узлами

*Кратным узлом* называют, по определению, узел, в котором информация о функции задаётся более одного раза. Помимо значения функции это может быть какая-либо дополнительная информация о ней, например значения производных и т. п. К задаче интерполяции с кратными узлами мы приходим, в частности, если степень интерполяционного полинома, который нужно однозначно построить по некоторым узлам, равна либо больше количества этих узлов.

Далее мы будем рассматривать следующую постановку задачи. Данные несовпадающие точки  $x_i$ ,  $i = 0, 1, \dots, n$ , — узлы интерполяирования, в которых заданы значения  $y_i^{(k)}$ ,  $k = 0, 1, \dots, N_i - 1$ , — их принимают интерполируемая функции  $f$  и её производные  $f^{(k)}(x)$ . При этом число  $N_i$  называют *кратностью* узла  $x_i$ . Требуется построить полином  $H_m(x)$  степени  $m$ , такой что

$$\begin{aligned} H_m(x_0) &= y_0^{(0)}, & H'_m(x_0) &= y_0^{(1)}, & \dots, & & H_m^{(N_0-1)}(x_0) &= y_0^{(N_0-1)}, \\ H_m(x_1) &= y_1^{(0)}, & H'_m(x_1) &= y_1^{(1)}, & \dots, & & H_m^{(N_1-1)}(x_1) &= y_1^{(N_1-1)}, \\ &\vdots &&\vdots &&\ddots &&\vdots \\ H_m(x_n) &= y_n^{(0)}, & H'_m(x_n) &= y_n^{(1)}, & \dots, & & H_m^{(N_n-1)}(x_n) &= y_n^{(N_n-1)} \end{aligned}$$

или кратко

$$H_m^{(k)}(x_i) = y_i^{(k)}, \quad i = 0, 1, \dots, n, \quad k = 0, 1, \dots, N_i - 1, \quad (2.52)$$

где полагается  $H_m^{(0)} = H_m$ . Иными словами, в узлах  $x_i$ ,  $i = 0, 1, \dots, n$ , как сам полином  $H_m(x)$ , так и все его производные  $H_m^{(k)}(x)$  вплоть до заданных порядков ( $N_i - 1$ ) должны принимать предписанные им значения  $y_i^{(k)}$ .

Задачу алгебраической интерполяции с кратными узлами в выписанной выше постановке часто называют также задачей *эрмитовой интерполяции*, а сам полином  $H_m(x)$ , решающий эту задачу, называют *интерполяционным полиномом Эрмита* по имени Ш. Эрмита, предложившего его во второй половине XIX века.<sup>9</sup> Эта постановка задачи

---

<sup>9</sup> Не следует путать «интерполяционный полином Эрмита» с известными ортогональными полиномами, которые тоже носят его имя и которые правильнее было бы называть «ортогональными полиномами Чебышёва—Эрмита».

алгебраической интерполяции с кратными узлами не является самой общей, так как порядки производных, для которых задаются значения в узлах, идут в ней последовательно друг за другом без пропусков. Тем не менее такая задача достаточно практична и хорошо исследована. Более общую постановку задачи алгебраической интерполяции с кратными узлами, где производные функции в узлах могут задаваться для произвольных фиксированных порядков, которые не идут один за другим, называют *интерполяцией Эрмита–Биркгофа*.

**Теорема 2.4.1** *Решение задачи эрмитовой интерполяции с кратными узлами при  $t = N_0 + N_1 + \dots + N_n - 1$  существует и единственно.*

**Доказательство.** В канонической форме полином  $H_m(x)$  имеет вид

$$H_m(x) = a_0 + a_1x + a_2x^2 + \dots + a_mx^m,$$

и для определения коэффициентов  $a_0, a_1, \dots, a_m$  станем подставлять в него и в его производные  $H'_m(x), H''_m(x), \dots$ , аргументы  $x_i, i = 0, 1, \dots, n$ , и использовать условия (2.52). Получим систему линейных алгебраических уравнений относительно  $a_0, a_1, \dots, a_m$ , в которой число уравнений равно  $N_0 + N_1 + \dots + N_n$ . При  $m = N_0 + N_1 + \dots + N_n - 1$  оно совпадает с числом неизвестных, равным  $m + 1$ .

Обозначим получившуюся систему линейных уравнений как

$$Ga = y, \tag{2.53}$$

где  $G$  — квадратная  $(m + 1) \times (m + 1)$ -матрица,

$a = (a_0, a_1, \dots, a_m)^\top \in \mathbb{R}^{m+1}$  — вектор неизвестных коэффициентов интерполяционного полинома,

$y = (y_0^{(0)}, y_0^{(1)}, \dots, y_0^{(N_0-1)}, y_1^{(0)}, y_1^{(1)}, \dots, y_n^{(N_n-1)})^\top \in \mathbb{R}^{m+1}$

— вектор, составленный из интерполяционных данных (2.52).

Матрица  $G$  зависит только от узлов  $x_0, x_1, \dots, x_n$  и никак не зависит от данных  $y_i^{(k)}, i = 0, 1, \dots, n, k = 0, 1, \dots, N_i - 1$ . Хотя эту матрицу даже можно выписать в явном виде, её прямое исследование весьма сложно, и для доказательства теоремы мы пойдём окольным путём.

Для определения свойств матрицы  $G$  рассмотрим однородную систему уравнений, отвечающую нулевой правой части  $y = 0$ , т. е.

$$Ga = 0.$$

В системе (2.53) вектор правой части  $y$  образован значениями интерполируемой функции и её производных  $y_i^{(k)}$  в узлах  $x_i$ ,  $i = 0, 1, \dots, n$ . Однородная система  $Ga = 0$  соответствует случаю  $y_i^{(k)} = 0$  для всех  $i = 0, 1, \dots, n$  и  $k = 0, 1, \dots, N_i - 1$ . Каким является вектор решений  $a$  этой системы?

Для нулевых интерполяционных данных узлы алгебраического интерполянта становятся его нулями. Поэтому если правая часть в (2.53) — нулевая, то это означает, что полином  $H_m(x)$  с учётом кратности имеет  $N_0 + N_1 + \dots + N_n = m + 1$  нулей, т. е. больше, чем его степень  $m$ . Это возможно лишь в случае, когда  $H_m(x)$  является тождественно нулевым, и тогда соответствующая однородная линейная система  $Ga = 0$  необходимо имеет лишь нулевое решение  $a = (a_0, a_1, \dots, a_n)^\top = (0, 0, \dots, 0)^\top$ .

Итак, линейная комбинация столбцов матрицы  $G$ , равная нулю, может быть только тривиальной, с нулевыми коэффициентами. Как следствие, матрица  $G$  должна быть неособенной, т. е.  $\det G \neq 0$ . Поэтому неоднородная система линейных уравнений (2.53) однозначно разрешима при любой правой части  $y$ , что и требовалось доказать. ■

Использованные при доказательстве теоремы 2.4.1 рассуждения, в которых построение интерполяционного полинома сводится к решению системы линейных алгебраических уравнений, носят конструктивный характер и позволяют практически решать задачу интерполяции с кратными узлами. Но аналогично случаю интерполяции с простыми узлами желательно иметь её аналитическое решение в виде обозримого конечного выражения для интерполянта. Он может иметь форму Лагранжа либо форму Ньютона (см. подробности, к примеру, в книгах [2, 28]). Наметим способ построения его лагранжевой формы, т. е. в виде линейной комбинации некоторых специальных базисных полиномов, каждый из которых отвечает за вклад отдельного узла.

Совершенно так же, как в § 2.2в, при фиксированном наборе узлов  $x_0, x_1, \dots, x_n$  результат решения рассматриваемой задачи интерполяции линейно зависит от значений  $y_0^{(0)}, y_0^{(1)}, \dots, y_0^{(N_0-1)}, y_1^{(0)}, y_1^{(1)}, \dots, y_n^{(N_n-1)}$ . Более точно, если полином  $P(x)$  решает задачу интерполяции по значениям  $y = (y_0^{(0)}, y_0^{(1)}, \dots, y_n^{(N_n-1)})$ , а полином  $Q(x)$  решает задачу интерполяции с теми же узлами по значениям  $z = (z_0^{(0)}, z_0^{(1)}, \dots, z_n^{(N_n-1)})$ , то для любых вещественных чисел  $\alpha$  и  $\beta$  полином  $\alpha P(x) + \beta Q(x)$  решает задачу интерполяции для значений  $\alpha y + \beta z =$

$(\alpha y_0^{(0)} + \beta z_0^{(0)}, \alpha y_0^{(1)} + \beta z_0^{(1)}, \dots, \alpha y_n^{(N_n-1)} + \beta z_n^{(N_n-1)})$  на той же совокупности узлов.

Отмеченное свойство можно также усмотреть из выписанного при доказательстве теоремы 2.4.1 представления вектора коэффициентов  $a = (a_0, a_1, \dots, a_n)^\top$  интерполяционного полинома как решения системы линейных уравнений (2.53). Из него следует, что  $a = G^{-1}y$ , т. е.  $a$  линейно зависит от вектора данных  $y$ , образованного значениями  $y_i^{(k)}$ ,  $k = 0, 1, \dots, N_i - 1$ ,  $i = 0, 1, \dots, n$ .

Итак, свойством линейности можно воспользоваться для решения задачи интерполяции с кратными узлами «по частям», которые удовлетворяют отдельным более простым интерполяционным условиям, а затем собрать эти части воедино. Иными словами, как и в случае интерполирования с простыми узлами, можно представить  $H_m(x)$  в виде линейной комбинации

$$H_m(x) = \sum_{i=0}^n \sum_{k=0}^{N_i-1} y_i^{(k)} \cdot \phi_{ik}(x),$$

где внешняя сумма берётся по узлам, внутренняя — по порядкам производной, а  $\phi_{ik}(x)$  — специальные «базисные» полиномы степени  $m$ , удовлетворяющие условиям

$$\phi_{ik}^{(l)}(x_j) = \begin{cases} 0 & \text{при } i \neq j \text{ или } k \neq l, \\ 1 & \text{при } i = j \text{ и } k = l. \end{cases} \quad (2.54)$$

У полинома  $\phi_{ik}(x)$  в узле  $x_i$  не равна нулю лишь одна из производных, порядок которой  $k$ , тогда как производные всех других порядков (среди которых может встретиться значение самого полинома) зануляются в  $x_i$ . Кроме того, полином  $\phi_{ik}(x)$  и все его производные равны нулю во всех остальных узлах, отличных от  $i$ -го. Фактически полином  $\phi_{ik}(x)$  отвечает системе уравнений (2.53) с вектор-столбцом правой части  $y$  вида  $(0, \dots, 0, 1, 0, \dots, 0)^\top$ , в котором все элементы нулевые, за исключением одного.

Каков конкретный вид этих базисных полиномов  $\phi_{ik}(x)$ ? Перепишем условия (2.54) в виде

$$\phi_{ik}^{(l)}(x_i) = \delta_{kl}, \quad k = 0, 1, \dots, N_i - 1, \quad (2.55)$$

$$\begin{aligned} \phi_{ik}^{(l)}(x_j) &= 0, & j &= 0, 1, \dots, i-1, i+1, \dots, n, \\ l &= 0, 1, \dots, N_i - 1. \end{aligned} \quad (2.56)$$

Из второго условия следует, что должно быть

$$\phi_{ik}(x) = (x - x_0)^{N_0} \dots (x - x_{i-1})^{N_{i-1}} (x - x_{i+1})^{N_{i+1}} \dots (x - x_n)^{N_n} Q_{ik}(x),$$

где  $Q_{ik}(x)$  — некоторый полином степени  $N_i - 1$ . Для его определения привлечём первое условие, т. е. (2.55). Оно означает, что узел  $x_i$  является нулём кратности  $k$  полинома  $\phi_{ik}$ . Поэтому

$$Q_{ik}(x) = (x - x_i)^k q_{ik}(x)$$

для какого-то полинома  $q_{ik}(x)$  степени  $N_i - k - 1$ . И так далее.

Мы не будем завершать это построение, так как дальнейшие выкладки весьма громоздки, а рассуждения нетривиальны. Детальное и полное построение интерполяционного полинома Эрмита можно увидеть, к примеру, в книгах [7, 25]. С помощью методов теории функций комплексного переменного формула для интерполяционного полинома Эрмита выводится в учебнике [20].

Какова погрешность эрмитовой интерполяции с кратными узлами? Она может быть представлена различными способами, и один из возможных вариантов ответа на этот вопрос даёт

**Теорема 2.4.2** Пусть  $f \in C^{m+1}[a, b]$ , т. е. функция  $f$  непрерывно дифференцируема  $m + 1$  раз на интервале  $[a, b]$ . Погрешность  $R_m(f, x)$  её эрмитовой интерполяции по несовпадающим узлам  $x_0, x_1, \dots, x_n \in [a, b]$  с кратностями  $N_0, N_1, \dots, N_n$  с помощью полинома  $H_m(x)$  степени  $m$  при условии  $m = N_0 + N_1 + \dots + N_n - 1$  может быть представлена в виде

$$R_m(f, x) = f(x) - H_m(x) = \frac{f^{(m+1)}(\xi(x))}{(m+1)!} \cdot \prod_{i=0}^n (x - x_i)^{N_i}, \quad (2.57)$$

где  $\xi(x)$  — некоторая точка из  $]a, b[$ , зависящая от  $x$ .

**Доказательство.** Для удобства обозначим через  $\Omega(x)$  произведение разностей со степенями, стоящее в правой части равенства (2.57), т. е.

$$\Omega(x) := \prod_{i=0}^n (x - x_i)^{N_i}.$$

$\Omega(x)$  — аналог полинома  $\omega_n(x)$ , введённого в § 2.2e и часто применяемого в конструкциях, относящихся к интерполяции.

Если  $x = x_i$  для одного из узлов интерполяции,  $i = 0, 1, \dots, n$ , то  $R_m(f, x) = 0$ , но в то же время и  $\Omega(x) = 0$ . Поэтому в (2.57) в качестве  $\xi(x)$  можно взять любую точку из интервала  $[a, b]$ .

Предположим теперь, что точка  $x$  из интервала интерполяции  $[a, b]$  не совпадает ни с одним из узлов  $x_i$ ,  $i = 0, 1, \dots, n$ . Введём вспомогательную функцию новой переменной  $z$

$$\psi(z) := f(z) - H_m(z) - K \Omega(z),$$

где числовую константу  $K$  для заданного  $x$  положим равной

$$K = \frac{f(x) - H_m(x)}{\Omega(x)}.$$

Следует отметить, что

$$\psi^{(m+1)}(z) = f^{(m+1)}(z) - K(m+1)!, \quad (2.58)$$

поскольку  $H_m(x)$  — полином степени  $m$  и  $H_m^{(m+1)}(z)$  — тождественный нуль, а  $\Omega(z)$  есть полином степени  $m+1$  со старшим коэффициентом 1.

Функция  $\psi(z)$  имеет нули в узлах  $x_0, x_1, \dots, x_n$  и, кроме того, по построению обращается в нуль при  $z = x$ , так что общее число нулей этой функции равно  $n+2$ . На основании теоремы Ролля можно заключить, что производная  $\psi'(z)$  должна обращаться в нуль по крайней мере в  $n+1$  точках, расположенных в интервалах между  $x, x_1, \dots, x_n$ . Но в узлах  $x_0, x_1, \dots, x_n$  функция  $\psi(z)$  имеет нули с кратностями  $N_0, N_1, \dots, N_n$  соответственно, что следует из условий интерполяции. Поэтому в  $x_0, x_1, \dots, x_n$  производная  $\psi'(z)$  имеет нули кратности  $N_0 - 1, N_1 - 1, \dots, N_n - 1$  (нулевая кратность означает отсутствие нуля в узле). Таким образом, производная  $\psi'(z)$  должна иметь всего, с учётом кратности, как минимум  $(n+1) + (N_0 - 1) + (N_1 - 1) + \dots + (N_n - 1) = N_0 + N_1 + \dots + N_n = m+1$  нулей на  $[a, b]$ .

Продолжая аналогичные рассуждения, получим, что вторая производная  $\psi''(z)$  будет иметь с учётом кратности по крайней мере  $m$  нулей на интервале  $[a, b]$  и т. д. При каждом последующем дифференцировании нули у производных функции  $\psi(z)$  могут возникать или исчезать, но, как следует из рассуждений предыдущего абзаца, их суммарная кратность уменьшается всякий раз на единицу. Наконец,  $(m+1)$ -я производная зануляется на  $[a, b]$  хотя бы один раз.

Итак, на интервале  $[a, b]$  обязательно найдётся по крайней мере одна точка  $\xi$ , зависящая, естественно, от  $x$  и такая, что  $\psi^{(m+1)}(\xi) = 0$ .

Поэтому в силу равенства (2.58) получаем

$$K = \frac{f^{(m+1)}(\xi)}{(m+1)!}.$$

Принимая во внимание определение константы  $K$ , немедленно приходим отсюда к формуле (2.57). ■

Интересно, что при наличии одного узла кратности  $m$  интерполяционный полином Эрмита становится полиномом Тейлора, а формула (2.57) совпадает с известной формулой остаточного члена (в форме Лагранжа) для полинома Тейлора. Если же все узлы интерполяции простые, то (2.57) превращается в полученную ранее формулу погрешности простой интерполяции (2.30).

## 2.5 Общие факты интерполяции

### 2.5а Интерполяционный процесс

Как с теоретической, так и с практической точек зрения интересен вопрос о том, можно ли при увеличении числа узлов уменьшить погрешность интерполирования и насколько именно. Вообще, сходятся ли интерполяционные полиномы к интерполируемой функции при неограниченном росте количества узлов? Конечно, по условиям интерполяции исходная функция равна своему интерполянту в узлах. Но не может ли оказаться, что между узлами, даже при их неограниченном сгущении, различие этих двух функций всё-таки будет неустранимым или даже увеличивающимся?

Чтобы строго сформулировать соответствующие вопросы и общие результаты о сходимости алгебраических интерполянтов, необходимо формализовать некоторые необходимые понятия.

**Определение 2.5.1** Пусть для интервала  $[a, b]$  задана бесконечная треугольная матрица узлов

$$\begin{pmatrix} x_0^{(0)} & 0 & 0 & 0 & \dots \\ x_0^{(1)} & x_1^{(1)} & 0 & 0 & \dots \\ x_0^{(2)} & x_1^{(2)} & x_2^{(2)} & 0 & \dots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}, \quad (2.59)$$

такая что в каждой её строке расположены различные точки интервала  $[a, b]$ , т. е.  $x_i^{(n)} \in [a, b]$  для всех неотрицательных целых чисел  $n$  и любых  $i = 0, 1, \dots, n$ , причём  $x_i^{(n)} \neq x_j^{(n)}$  для  $i \neq j$ . Говорят, что на интервале  $[a, b]$  задан интерполяционный процесс, если элементы  $n$ -й строки этой матрицы берутся в качестве узлов интерполяции, по которым строится последовательность интерполянтов  $g_n(x)$ ,  $n = 0, 1, 2, \dots$ .

Если в этом определении все интерполянты  $g_n(x)$  являются алгебраическими полиномами, то употребляется термин *алгебраический интерполяционный процесс*.

Ясно, что определение 2.5.1 предполагает наличие бесконечной треугольной матрицы, похожей на (2.59) и составленной из значений  $y_i^{(n)}$ , которые интерполянты  $g_n(x)$  принимают в узлах  $x_i^{(n)}$ ,  $i = 0, 1, \dots, n$ . Если при этом  $y_i^{(n)}$  являются значениями некоторой функции  $f$  в узлах  $x_i^{(n)}$ , т. е.  $y_i^{(n)} = f(x_i^{(n)})$ , то будем говорить, что интерполяционный процесс применяется к функции  $f$  (или для функции  $f$ ).

**Определение 2.5.2** *Интерполяционный процесс для функции  $f$  называется сходящимся в точке  $y \in [a, b]$ , если порождаемая им последовательность значений интерполянтов  $g_n(y)$  сходится к  $f(y)$  при  $n \rightarrow \infty$ .*

**Определение 2.5.3** *Интерполяционный процесс для функции  $f$  на интервале  $[a, b]$  называется сходящимся равномерно, если порождаемая им последовательность интерполянтов  $g_n(y)$  равномерно сходится к  $f(y)$  при  $n \rightarrow \infty$ , т. е.*

$$\lim_{n \rightarrow \infty} \max_{x \in [a, b]} |f(x) - g_n(x)| = 0.$$

Отметим, что помимо равномерной сходимости интерполяционного процесса, когда отклонение одной функции от другой измеряется в равномерной (чебышёвской) метрике (2.1), иногда необходимо рассматривать сходимость в других смыслах. Например, это может быть среднеквадратичная сходимость, задаваемая метрикой (2.3), или ещё какая-нибудь другая.

Определённую уверенность в положительном ответе на поставленные в начале параграфа вопросы о сходимости алгебраических интерполяционных процессов, даже в более сильном равномерном смысле, даёт известная из математического анализа

**Теорема Вейерштрасса о равномерном приближении** [2, 12, 74]  
 Если  $f : \mathbb{R} \supset [a, b] \rightarrow \mathbb{R}$  – непрерывная функция, то для всякого  $\epsilon > 0$  существует алгебраический полином  $\Pi_n(x)$  степени  $n = n(\epsilon)$ , равномерно приближающий функцию  $f$  с погрешностью, не большей  $\epsilon$ , т. е. такой, что

$$\max_{x \in [a, b]} |f(x) - \Pi_n(x)| \leq \epsilon.$$

Этот результат служит теоретической основой равномерного приближения непрерывных функций алгебраическими полиномами, обеспечивая существование полинома, который сколь угодно близок к заданной непрерывной функции в смысле чебышёвского расстояния (2.1).

Вместе с тем теорема Вейерштрасса относится к задаче приближения (аппроксимации) функций, а не к интерполированию, где требуется совпадение значений функции и её интерполянта на данном множестве точек-узлов. В задаче приближения функций участки области определения, где сравниваемые функции совпадают друг с другом, не фиксированы. Совершенно аналогично участки, где функции отклоняются друг от друга, тоже могут «гулять» по интервалу. Таким образом, теорема Вейерштрасса всё-таки не даёт ответов на конкретные вопросы о решении задачи интерполирования и сходимости интерполяционных процессов.

## 2.5б Сводка результатов и обсуждение

Как следует из результатов § 2.2e и § 2.3, огромное влияние на погрешность интерполяции оказывает расположение узлов. В частности, рассмотренные в § 2.3 чебышёвские сетки являются наилучшими возможными в условиях, когда неизвестна какая-либо дополнительная информация об интерполируемой функции.

Равномерные сетки, несмотря на их естественность и практическую удобность, ведут себя гораздо хуже. Для них один из первых примеров расходимости интерполяционных процессов привёл в 1910 году С.Н. Бернштейн, рассмотрев на интервале  $[-1, 1]$  алгебраическую интерполяцию функции  $f(x) = |x|$  по равноотстоящим узлам, включаяющим и концы этого интервала. Не слишком трудными рассуждениями показывается [7, 29], что с возрастанием числа узлов соответствующий интерполяционный полином не стремится к  $|x|$  ни в одной точке интервала  $[-1, 1]$ , отличной от  $-1, 0$  и  $1$  (рис. 2.14). Может показаться, что причиной расходимости интерполяционного процесса в приме-

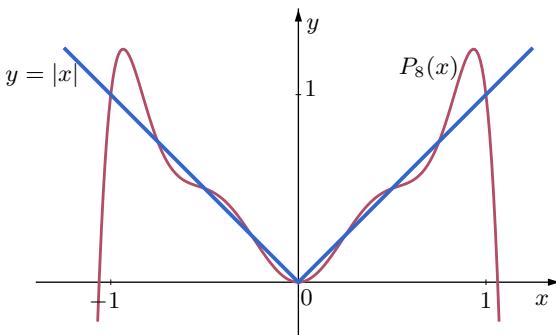


Рис. 2.14. Интерполяция полиномом 8-й степени в примере Бернштейна

ре С.Н. Бернштейна является отсутствие гладкости интерполируемой функции, но это верно лишь отчасти.

Предположим, что интерполируемая функция  $f$  имеет бесконечную гладкость,  $f \in C^\infty[a, b]$ , т. е. обладает непрерывными производными любого порядка, причём они растут «не слишком быстро». В последнее условие будем вкладывать следующий смысл:

$$\sup_{x \in [a, b]} |f^{(n)}(x)| < M^n, \quad n = 1, 2, \dots, \quad (2.60)$$

где константа  $M$  не зависит от  $n$ . Тогда из теоремы 2.2.2 следует, что погрешность алгебраического интерполирования по  $n$  узлам может быть оценена сверху как

$$\frac{(M(b-a))^n}{n!},$$

т. е. при  $n \rightarrow \infty$  она очевидным образом сходится к нулю вне зависимости от расположения узлов интерполяции. Иными словами, любой алгебраический интерполяционный процесс на интервале  $[a, b]$  будет равномерно сходиться к такой функции  $f$ .

Условие (2.60) влечёт сходимость ряда Тейлора для функции  $f$  в любой точке из  $[a, b]$ , и, отправляясь от этого наблюдения, можно дать простое достаточное условие сходимости интерполяционного процесса в терминах теории функций. Напомним, что если функция может быть представлена степенным рядом, который сходится при любых (вещественных или комплексных) значениях аргумента, то она называется

*целой функцией* (см., например, [85]). В теории функций показывается, что степенной ряд, о котором говорится в этом определении, в действительности является рядом Тейлора, а целые функции бесконечно дифференцируемы. Целые функции можно рассматривать как непосредственное обобщение многочленов, фактически как «многочлены бесконечной степени». Нетривиальные примеры целых функций — это экспонента, синус, косинус и т. п. Суммы, разности, произведения и суперпозиции целых функций также являются целыми.

**Теорема 2.5.1** *Если функция — целая, то интерполяционный процесс сходится к ней равномерно по любой последовательности сеток на заданном интервале.*

Заметим, что в условиях сформулированной теоремы расположение узлов даже несущественно. Доказательство этого результата можно найти, например, в [2, 74].

Значение теоремы 2.5.1 для практики не слишком велико, так как целые функции образуют достаточно узкий класс, который, как правило, недостаточен для многих задач математического моделирования. Например, логарифм, квадратный корень, дробно-рациональные функции не являются целыми функциями. Всё же отметим, что теорема 2.5.1 допускает обобщения на функции, разлагающиеся в степенные ряды, которые сходятся не при любых значениях аргумента, а лишь из какой-то ограниченной области специальной формы, содержащей интервал интерполяирования [7, 20].

В самом общем случае при алгебраическом интерполировании бесконечно гладких функций погрешность всё-таки может не сходить к нулю даже при «вполне разумном» расположении узлов, когда они всюду плотно покрывают интервал интерполяирования. По-видимому, наиболее известный пример такого рода привёл немецкий математик К. Рунге в 1901 году [107]. Нередко его называют также «явлением Рунге» или «феноменом Рунге».

В примере Рунге функция

$$\Upsilon(x) = \frac{1}{1+x^2}$$

на интервале  $[-5, 5]$  интерполируется алгебраическими полиномами, которые построены на последовательности равномерных сеток с узлами  $x_i = -5 + 10i/n$ ,  $i = 0, 1, \dots, n$ . Если  $P_n(x)$  — интерполяционный

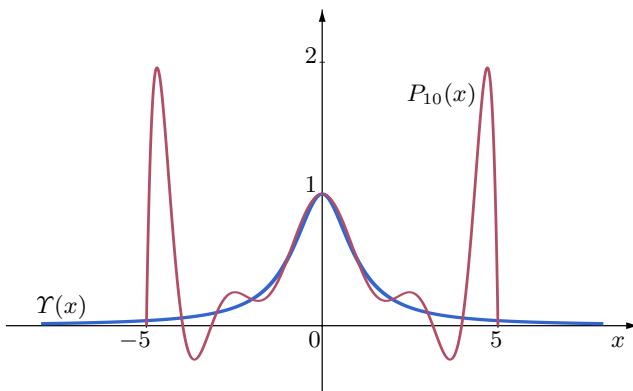


Рис. 2.15. Интерполяция полиномом 10-й степени в примере Рунге

полином  $n$ -й степени, построенный по  $n$ -й сетке, то оказывается, что

$$\lim_{n \rightarrow \infty} \max_{x \in [-5, 5]} |Y(x) - P_n(x)| = \infty.$$

При этом вблизи концов интервала интерполяирования  $[-5, 5]$  у полиномов  $P_n(x)$  с ростом  $n$  возникают сильные колебания (называемые *осцилляциями*), размах которых стремится к бесконечности (рис. 2.15). Получается, что хотя в узлах интерполяирования значения функции  $Y(x)$  совпадают со значениями полинома  $P_n(x)$ , между этими узлами  $Y(x)$  и  $P_n(x)$  могут различаться сколь угодно сильно, даже несмотря на плавный (бесконечно гладкий) характер изменения функции  $Y(x)$  и равномерное стущение узлов интерполяции.<sup>10</sup>

Интересно, что на интервале  $[-\kappa, \kappa]$ , где  $\kappa \approx 3.63$ , рассматриваемый интерполяционный процесс равномерно сходится к  $Y(x)$  [107]. Функция  $Y(x)$  имеет производные всех порядков для любого вещественного аргумента  $x$ , но у концов интервала интерполяирования  $[-5, 5]$  эти производные растут очень быстро и уже не удовлетворяют условию (2.60). Проявив некоторую изобретательность (детали описаны в п. 116 тома 1 известного учебника Г.М. Фихтенгольца [40]), можно показать, что при

---

<sup>10</sup>Помимо оригинальной статьи К. Рунге [107] этот факт строго обосновывается, к примеру, в книге [92].

$$x > 0$$

$$\Upsilon^{(n)}(x) = \left( \frac{1}{1+x^2} \right)^{(n)} = \frac{(-1)^n (n-2)!}{(1+x^2)^{(n-1)/2}} \cdot \sin \left( (n-1) \arctg \frac{1}{x} \right). \quad (2.61)$$

Другой способ многократного дифференцирования функции из примера Рунге может быть основан на её комплексном представлении

$$\Upsilon(x) = \frac{1}{1+x^2} = \frac{i}{2} \left( \frac{1}{x+i} - \frac{1}{x-i} \right).$$

Как следствие, легко получаем

$$\Upsilon^{(n)}(x) = \left( \frac{1}{1+x^2} \right)^{(n)} = \frac{i}{2} (-1)^n n! \left( \frac{1}{(x+i)^{n+1}} - \frac{1}{(x-i)^{n+1}} \right), \quad (2.62)$$

откуда нетрудно вывести вещественную производную. Множители в виде факториалов в выражениях (2.61) и (2.62) определяют общее поведение производных при росте  $n$ . Таким образом, несмотря на простой вид, функция  $\Upsilon(x)$  из примера Рунге своим поведением слишком непохожа на полиномы, производные от которых не растут столь быстро и, начиная с некоторого порядка, исчезают. Подробное рассмотрение этих интересных вопросов относится уже к предмету теории функций (см., к примеру, [85]).

Что касается чебышёвских сеток, то они обеспечивают сходимость алгебраических интерполяционных процессов для существенно более широких классов функций. Чтобы сформулировать соответствующие результаты, напомним, что *модулем непрерывности* функции  $f$  на интервале  $[a, b]$  (который может быть и бесконечным) называется функция  $\omega_f(\delta)$  неотрицательного аргумента  $\delta$ , определяемая как

$$\omega_f(\delta) := \sup \{ |f(x+h) - f(x)| \mid x, x+h \in [a, b], |h| \leq \delta \}.$$

Модуль непрерывности даёт точную верхнюю оценку отличия значений функции, для которых аргументы разнятся не более чем на  $\delta$ . Из самой конструкции модуля непрерывности следует, что он является глобальной характеристикой функции на всей её области определения.

Можно показать [29, 59], что  $\omega_f(\delta)$  — неубывающая неотрицательная функция от  $\delta$ , имеющая предел  $\omega_f(+0) = \lim_{\delta \rightarrow 0} \omega_f(\delta)$ . По этой причине обычно рассматривают модуль непрерывности при  $\delta \geq 0$ , полагая  $\omega_f(0) = \omega_f(+0)$ . Функция  $f$  равномерно непрерывна на интервале

$[a, b]$ , тогда и только тогда, когда её модуль непрерывности стремится к нулю при  $\delta \rightarrow 0$ . По скорости убывания модуля непрерывности можно судить о весьма тонких свойствах самой функции.

Говорят, что функция  $f$  удовлетворяет *условию Дини–Липшица* на заданном множестве, если

$$\lim_{\delta \rightarrow 0} \omega_f(\delta) \ln \delta = 0,$$

т. е. если при уменьшении  $\delta$  модуль непрерывности убывает быстрее, чем  $1/|\ln \delta|$ . Оказывается, что если функция удовлетворяет условию Дини–Липшица, то на последовательности чебышёвских сеток из заданного интервала алгебраический интерполяционный процесс сходится к ней равномерно.<sup>11</sup> Обоснование этого результата читатель может найти в [8, 29, 71]. Отметим, что из-за медленного роста модуля логарифма при стремлении его аргумента к нулю условие Дини–Липшица является очень слабым условием, которому заведомо удовлетворяют все непрерывные функции, встречающиеся в практике математического моделирования.

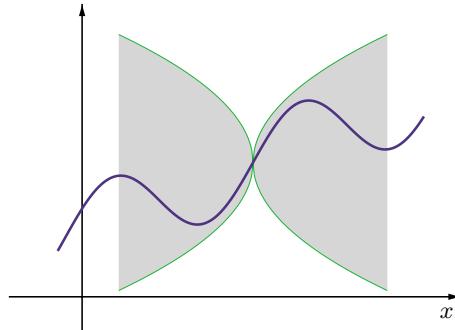


Рис. 2.16. Иллюстрация обобщённого условия Липшица

Удобным достаточным условием, при котором выполняется условие Дини–Липшица, является *условие Гёльдера* (иногда его называют также *обобщённым условием Липшица*): для любых  $x, y$  из области определения функции  $f$  имеет место

$$|f(x) - f(y)| \leq C |x - y|^\alpha \quad (2.63)$$

---

<sup>11</sup>Условие Дини–Липшица на функцию является также достаточным условием равномерной сходимости к ней ряда Фурье [7, 58].

с некоторыми константами  $C$  и  $\alpha$ ,  $0 < \alpha \leq 1$  (рис. 2.16). В самом деле, (2.63) равносильно тому, что

$$|f(x + h) - f(x)| \leq C|h|^\alpha,$$

и поэтому для значений модуля непрерывности  $\omega_f(\delta)$  можно дать оценку сверху в виде

$$\omega_f(\delta) \leq C|h|^\alpha.$$

Тогда  $|\omega_f(\delta) \ln \delta| \leq C|h|^\alpha |\ln \delta| \rightarrow 0$  и условие Дини–Липшица очевидно выполняется. Отметим, что условиями Дини–Липшица и Гёльдера допускаются бесконечные значения производных для функции (рис. 2.16).

**Теорема 2.5.2** *Если функция удовлетворяет условию Гёльдера (2.63), то на последовательности чебышёвских сеток алгебраический интерполяционный процесс сходится к этой функции равномерно.*

Обоснование этого утверждения можно увидеть, к примеру, в [29]. Тем не менее для общих непрерывных функций имеет место следующий отрицательный результат:

**Теорема Фабера** [7, 8, 28]<sup>12</sup> *Не существует бесконечной треугольной матрицы узлов из заданного интервала, такой что соответствующий ей алгебраический интерполяционный процесс сходился бы равномерно для любой непрерывной функции на этом интервале.*

В частности, даже на последовательности чебышёвских сеток из заданного интервала алгебраический интерполяционный процесс может *всюду* расходиться для некоторых непрерывных функций. Подробности можно найти в книге [29].

Но отрицательные результаты теоремы Фабера и примыкающих к ней примеров характеризуют, скорее, слишком большую общность математического понятия «непрерывной функции». Получается, что непрерывная функция может оказаться очень необычной и не похожей на то, что мы интуитивно вкладываем в смысл «непрерывности». Об этом же свидетельствуют парадоксальные примеры непрерывных нигде не дифференцируемых функций (примеры Вейерштрасса или ван дер Вардена; см., к примеру, [40], том 2, пункт 444). Такой же

---

<sup>12</sup>Часто её называют «теоремой Фабера–Бернштейна».

экзотичной является непрерывная функция, для которой расходится интерполяционный процесс по чебышёвским сеткам. Поэтому можно считать, что теорема Фабера утверждает лишь то, что класс непрерывных в классическом смысле функций (непрерывных «по Больцано–Коши») является слишком широким, чтобы для него существовал один (с точностью до преобразований интервала) интерполяционный процесс, обеспечивающий равномерную сходимость для любой функции.

Чересчур большая общность понятия непрерывной функции была осознана математиками почти сразу после своего появления, в первой половине XIX века. Она стимулировала работы по формулировке дополнительных естественных условий, которые выделяли бы классы функций, непрерывных в более сильных смыслах и позволяющих свободно выполнять те или иные операции анализа (например, взятие производной почти всюду в области определения и т. п.). Именно эти причины вызвали появление понятий равномерной непрерывности, абсолютной непрерывности, условий Липшица и Дини–Липшица, а также ряда других им аналогичных.

С другой стороны, для общих непрерывных функций имеет место «оптимистический» результат, практическая ценность которого, правда, невелика:

**Теорема Марцинкевича [7, 8, 28]** *Если функция непрерывна на заданном интервале, то существует такая бесконечная треугольная матрица узлов из этого интервала, что соответствующий ей алгебраический интерполяционный процесс для рассматриваемой функции сходится равномерно.*

Заметим, что построение матрицы интерполяционных узлов, т. е. последовательности сеток, о которой говорится в теореме Марцинкевича, является не менее трудным, чем практическая интерполяция заданной функции.

Интересно, что ситуация со сходимостью интерполяционных процессов в среднеквадратичном смысле более благоприятна, чем для равномерной сходимости. Если рассматривается среднеквадратичное расстояние между функциями (2.3) или более общее (2.117) с некоторым весом  $\varrho(x)$ , то, взяв бесконечную треугольную матрицу узлов из интервала интерполяирования по нулям ортогональных для данного веса  $\varrho(x)$  полиномов, мы получим сходимость интерполяционного процесса для любой непрерывной функции (см. подробности в [8, 29] и цитиро-

ванной там литературе). Ортогональные полиномы будут обсуждаться далее в § 2.11 и § 2.12, где мы детально рассмотрим полиномы Лежандра, ортогональные с единичным весом на интервале  $[-1, 1]$  (см. § 2.12). Сетки по их нулям обеспечивают среднеквадратичную сходимость интерполяционного процесса для любой непрерывной функции.

Как отмечалось в § 2.3б, полиномы Чебышёва — тоже ортогональные полиномы, но с некоторым специальным весом (см. также § 2.12в). Поэтому чебышёвские сетки тоже обеспечивают сходимость в среднеквадратичном смысле с этим весом для интерполяционных процессов с любыми непрерывными функциями.

Ещё один вывод из представленных выше примеров и результатов заключается в том, что алгебраические полиномы, несмотря на определённые удобства работы с ними, оказываются весьма капризным инструментом интерполирования достаточно общих непрерывных и даже гладких функций. Как следствие, нам нужно иметь более гибкие инструменты интерполяции, т. е. использовать в качестве интерполянтов другие классы функций. Их рассмотрению будут посвящены следующие параграфы.

## 2.6 Сплайны

### 2.6а Элементы теории

Сплайны являются функциями, которые обычно задаются способом, промежуточным между табличным способом, привычным инженерам и практикам, и заданием с помощью единой формулы, одним выражением, характерным для теоретической математики.

**Определение 2.6.1** Пусть задан некоторый интервал  $[a, b]$ , который разбит на подинтервалы  $[x_{i-1}, x_i]$ ,  $i = 1, 2, \dots, n$ , так что  $a = x_0$  и  $x_n = b$ . Полиномиальным сплайном на  $[a, b]$  называется функция, которая на каждом подинтервале  $[x_{i-1}, x_i]$  является алгебраическим полиномом и на всём интервале  $[a, b]$  непрерывна вместе со своими производными вплоть до некоторого порядка.

Максимальная на всём интервале  $[a, b]$  степень полиномов, задающих сплайн, называется *степенью сплайна*. Наивысший порядок производной сплайна, которая непрерывна на  $[a, b]$ , — это *гладкость сплайна*, а разность между степенью сплайна и его гладкостью называется

*дефектом сплайна.* Наконец, точки  $x_i$ ,  $i = 0, 1, \dots, n$ , — концы подинтервалов, на которые разбивается  $[a, b]$ , — называют *узлами сплайна*.

Помимо полиномиальных существуют также другие типы сплайнов — тригонометрические, экспоненциальные и т. п. Но далее мы рассмотрим только полиномиальные сплайны, и потому для краткости станем говорить просто о «сплайнах». Аналогично говорим просто о «полиномах», опуская прилагательное «алгебраический».

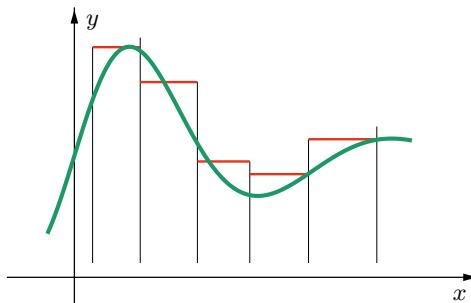


Рис. 2.17. Кусочно-постоянная интерполяция функции

Почему именно кусочные полиномы? К идею их введения можно прийти, к примеру, опираясь на следующие неформальные мотивации. Содержательный (механический, физический, биологический и т. п.) смысл имеют, как правило, производные порядка не выше 2–4, именно их мы можем видеть в математических формулировках различных законов природы или математических моделях реальных явлений.<sup>13</sup> Пятое производные — это уже экзотика, а производные шестого и более высоких порядков при описании реальности не встречаются. В частности, производные высоких «нефизических» порядков и их разрывы никак не ощущимы практически. Поэтому для сложно изменяющихся производных высоких порядков необходимые «нужные» значения в фиксированных узлах можно назначить, к примеру, с помощью простейшей кусочно-постоянной или кусочно-линейной интерполяции (рис. 2.17). Далее мы восстанавливаем искомую функцию, последовательно при-

<sup>13</sup>Характерный пример: в книге А.К. Маловичко и О.Л. Тарунина «Использование высших производных при обработке и интерпретации результатов геофизических наблюдений» (М., издательство «Недра», 1981 год) рассматриваются производные только второго и третьего порядков.

меняя необходимое число раз операцию интегрирования. При этом достигается желаемая гладкость функции на отдельных подинтервалах области определения, а если мы отслеживаем гладкость склейки этих кусков в единое целое, то получается и глобальная гладкость функции. Но при последовательном интегрировании константы возникают алгебраические полиномы, а в целом получающаяся функция — кусочно-полиномиальная.

Термин «сплайн» является удачным заимствованием из английского языка, где слово *spline* означает гибкую (обычно стальную) линейку, которую, изгиная, использовали чертёжники для проведения гладкой линии между фиксированными точками. Понятие сплайн-функции введено И. Шёнбергом в 1946 году [108], хотя различные применения тех объектов, которые впоследствии были названы «сплайнами», встречались в математике на протяжении предшествующей сотни лет. Пионером здесь следует назвать, по-видимому, Н.И. Лобачевского, который в статье [102] явно использовал конструкции обычных сплайнов и так называемых *B*-сплайнов.<sup>14</sup>

С середины XX века по настоящее время сплайны нашли широкое применение в математике и её приложениях. В теоретических дисциплинах и в вычислительных технологиях они могут использоваться для приближения и интерполирования функций, при численном решении дифференциальных и интегральных уравнений и т. п. Если сплайн применяется для решения задачи интерполяции, то он называется *интерполяционным*. Другими словами, интерполяционный сплайн — это сплайн, принимающий в заданных точках  $\tilde{x}_i$ ,  $i = 0, 1, \dots, r$ , — узлах интерполяции — требуемые значения  $y_i$ . Эти узлы интерполяции, вообще говоря, могут не совпадать с узлами сплайна  $x_i$ ,  $i = 0, 1, \dots, n$ , задающими интервалы полиномиальности.

Так как степень алгебраического полинома равна наивысшему порядку его ненулевой производной, то сплайны дефекта нуль — это функции, задаваемые на всём интервале  $[a, b]$  одной полиномиальной формулой. Таким образом, термин «дефект» весьма точно выражает то, сколько сплайну «не хватает» до полноценного полинома. С другой стороны, именно наличие дефекта обеспечивает сплайну большую гибкость в сравнении с полиномами и делает сплайны во многих ситуациях более удобным инструментом приближения и интерполирования функций.

---

<sup>14</sup> Вклад Н.И. Лобачевского даже дал повод некоторым авторам назвать сплайны *Лобачевского* специальный вид сплайнов.

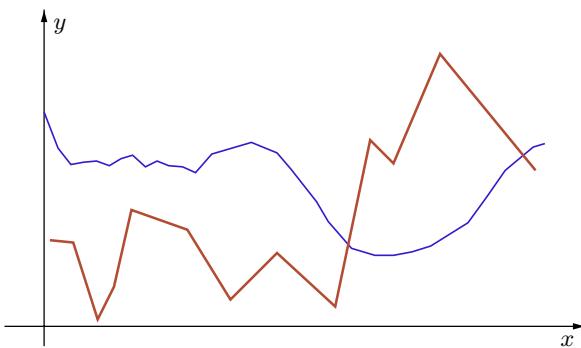


Рис. 2.18. Простейшие сплайны — кусочно-линейные функции

Сплайны существенно лучше алгебраических полиномов позволяют отслеживать специфику поведения многих функций. Дело в том, что у полиномов производные по мере увеличения их порядка имеют всё более медленный рост и в конце концов просто зануляются. Этим полиномы принципиально отличаются от всех других функций, производные которых при увеличении их порядка в нуль не обращаются или даже быстро растут. Что касается сплайнов, то наличие у них кусочно-го представления и точек склейки, где производные терпят разрывы, приводит к интересному эффекту. Пусть, к примеру,  $q$  — это гладкость сплайна дефекта 1. Тогда разрывы  $(q+1)$ -й производной в узлах сплайна можно трактовать как бесконечно большие значения следующей  $(q+2)$ -й производной в узлах,<sup>15</sup> между которыми эта производная равна нулю. «В среднем» же  $(q+2)$ -я производная от сплайна оказывается не равной тождественному нулю!

Чем больше дефект сплайна, тем больше он отличается от алгебраического полинома и тем более специфичны его свойства. Но слишком большой дефект при фиксированной степени сплайна приводит к существенному понижению его общей гладкости. В значительном числе приложений сплайнов вполне достаточными оказываются сплайны с минимально возможным дефектом 1, и только такие сплайны мы будем рассматривать далее в нашей книге.

Простейший «настоящий» сплайн имеет дефект 1 и степень 1, бу-

---

<sup>15</sup>Этому утверждению можно даже придать строгий математический смысл, привлекая понятие так называемой обобщённой функции.

дучи «непрерывно склеенным» в своих узлах  $x_i$ ,  $i = 1, 2, \dots, n - 1$ . Иными словами, это кусочно-линейная функция (рис. 2.18), имеющая, несмотря на свою простоту, богатые приложения в математике и других науках.<sup>16</sup> Сплайны второй степени часто называют *параболическими*.

Если степень сплайна равна  $d$ , то для его полного определения необходимо знать  $n(d + 1)$  значений коэффициентов полиномов, задающих сплайн на  $n$  подинтервалах  $[x_{i-1}, x_i]$ ,  $i = 1, 2, \dots, n$ . В то же время в случае дефекта 1 имеется

$d(n - 1)$  условий непрерывности самого сплайна и его производных вплоть до  $(d - 1)$ -го порядка в узлах  $x_1, x_2, \dots, x_{n-1}$ ,

$(n + 1)$  условие интерполяции в узлах  $x_0, x_1, \dots, x_n$ .

Каждое из этих условий является, фактически, уравнением относительно неизвестных коэффициентов полиномов, задающих сплайн на отдельных подинтервалах области определения. Соответственно, построение сплайна сводится к решению получающейся системы уравнений. Всего таких условий-уравнений набирается  $d(n - 1) + (n + 1) = n(d + 1) - (d - 1)$  штук, так что эта система уравнений недоопределенна. Для её однозначного решения и полного определения сплайна не хватает  $d - 1$  условий, которые обычно задают дополнительно на концах интервала  $[a, b]$ .

Сказанное имеет следующие важные следствия. Если решать задачу интерполяции с помощью сплайна чётной степени, когда на каждом подинтервале  $[x_{i-1}, x_i]$  сплайн должен рассматриваться полиномом чётной степени, то число  $(d - 1)$  подлежащих доопределению параметров оказывается нечётным. Поэтому на одном из концов интервала  $[a, b]$  приходится налагать больше условий, чем на другом. Это приводит, во-первых, к асимметрии задачи, и, во-вторых, может вызвать неустойчивость при определении параметров сплайна. Наконец, интерполяционный сплайн чётной степени при некоторых естественных краевых условиях (периодических, к примеру) может просто не существовать.

Отмеченные недостатки могут решаться, в частности, выбором узлов сплайна отличными от узлов интерполяции. Мы далее не будем останавливаться на преодолении этих затруднений и рассмотрим ин-

---

<sup>16</sup> Вспомним, к примеру, «ломаные Эйлера», которые применяются при доказательстве существования решения задачи Коши для обыкновенных дифференциальных уравнений [45].

терполяционные сплайны нечётной степени 3, узлы которых совпадают с узлами интерполяции. Последнее обстоятельство существенно упрощает процесс построения сплайна и работу с ним.

## 2.66 Интерполяционные кубические сплайны

В вычислительных технологиях решения различных задач одним из наиболее популярных инструментов являются полиномиальные сплайны третьей степени с дефектом 1, которые называются также *кубическими сплайнами*. Эту популярность можно объяснить относительной простотой этих сплайнов и тем обстоятельством, что они вполне достаточны для отслеживания непрерывности вторых производных функций. Это необходимо, например, во многих законах механики и физики.

Пусть задан набор узлов  $x_0, x_1, \dots, x_n \in [a, b]$ , такой что  $a = x_0 < x_1 < \dots < x_n = b$ . Как и прежде, совокупность всех узлов мы называем *сеткой*. Величину  $h_i = x_i - x_{i-1}$ ,  $i = 1, 2, \dots, n$ , назовём *шагом сетки*. Кубический интерполяционный сплайн на интервале  $[a, b]$  с определённой выше сеткой  $\{x_0, x_1, \dots, x_n\}$ , узлы которой являются также узлами интерполяции, — это функция  $S(x)$ , удовлетворяющая следующим условиям:

- 1)  $S(x)$  — полином третьей степени на каждом из подинтервалов  $[x_{i-1}, x_i]$ ,  $i = 1, 2, \dots, n$ ;
- 2)  $S(x) \in C^2[a, b]$ ;
- 3)  $S(x_i) = y_i$ ,  $i = 0, 1, 2, \dots, n$ .

Для построения такого сплайна  $S(x)$  нужно определить  $4n$  неизвестных величин — по 4 коэффициента полинома третьей степени на каждом из  $n$  штук подинтервалов  $[x_{i-1}, x_i]$ ,  $i = 1, 2, \dots, n$ .

Для решения поставленной задачи в нашем распоряжении имеются

$3(n - 1)$  условий непрерывности самой функции  $S(x)$ , её первой и второй производных во внутренних узлах  $x_1, x_2, \dots, x_{n-1}$ ;

$(n + 1)$  условие интерполяции  $S(x_i) = y_i$ ,  $i = 0, 1, 2, \dots, n$ .

Таким образом, для нахождения  $4n$  неизвестных величин мы имеем всего  $3(n - 1) + (n + 1) = 4n - 2$  условий. Два недостающих условия

определяются различными способами, среди которых часто используются, к примеру, такие:

$$(I) \quad S'(a) = \beta_0, \quad S'(b) = \beta_n,$$

$$(II) \quad S''(a) = \gamma_0, \quad S''(b) = \gamma_n,$$

$$(III) \quad S^{(k)}(a) = S^{(k)}(b), \quad k = 0, 1, 2,$$

где  $\beta_0, \beta_n, \gamma_0, \gamma_n$  — данные вещественные числа. Условия (I) и (II), которыми на концах интервала  $[a, b]$  назначаются первая или вторая производные искомого сплайна, определяют в этих точках его наклон или (с точностью до множителя) кривизну. Условие (III) — это условие гладкого периодического продолжения сплайна с интервала  $[a, b]$  на более широкое подмножество вещественной оси.

Рассмотрим подробно случай (II) задания краевых условий:

$$S''(a) = S''(x_0) = \gamma_0,$$

$$S''(b) = S''(x_n) = \gamma_n.$$

Будем искать кусочно-полиномиальное представление нашего кубического сплайна в специальном виде, привязанном к узлам сплайна  $x_i$ , когда переменными являются разности  $x - x_i$ . Более точно, пусть

$$S(x) = \alpha_i + \beta_i(x - x_i) + \gamma_i \frac{(x - x_i)^2}{2} + \vartheta_i \frac{(x - x_i)^3}{6} \quad (2.64)$$

для  $x \in [x_i, x_{i+1}]$ ,  $i = 0, 1, \dots, n - 1$ , где  $\alpha_i, \beta_i, \gamma_i, \vartheta_i$  — некоторые вещественные числа. Множители  $1/2$  и  $1/6$  введены для того, чтобы при дифференцировании они могли красиво сокращаться с показателями степени. Ясно, что в такой форме представления сплайна величины  $\beta_0$  и  $\gamma_0$  совпадают по смыслу с теми, что даются в условиях (I) и (II). Более того, из представления (2.64) вытекает, что

$$S''(x_i) = \gamma_i, \quad i = 1, 2, \dots, n - 1.$$

Далее мы, во-первых, выведем из (2.64) такое представление для сплайна на подинтервалах  $[x_i, x_{i+1}]$ , которое содержит в качестве неизвестных параметров только  $\gamma_i$ ,  $i = 1, 2, \dots, n - 1$ , и, во-вторых, составим для их нахождения систему линейных алгебраических уравнений. Её решение позволит однозначно построить искомый сплайн при любых  $\gamma_0$  и  $\gamma_n$ .

Заметим, что вторая производная  $S''(x)$  является линейной функцией на  $[x_i, x_{i+1}]$  и с учётом (2.64) должно быть

$$S''(x) = \gamma_i + \vartheta_i(x - x_i), \quad x \in [x_i, x_{i+1}]. \quad (2.65)$$

С другой стороны, вид этой линейной функции полностью задаётся двумя её крайними значениями  $\gamma_i$  и  $\gamma_{i+1}$  на концах соответствующего подинтервала  $[x_i, x_{i+1}]$ . Поэтому вместо (2.65) можно выписать более определённое представление, уже не задействующее  $\vartheta_i$ , но использующее  $\gamma_{i+1}$ . Итак, для  $x \in [x_i, x_{i+1}]$ ,  $i = 0, 1, \dots, n - 1$ , справедливо

$$S''(x) = \gamma_i \frac{x_{i+1} - x}{h_{i+1}} + \gamma_{i+1} \frac{x - x_i}{h_{i+1}}, \quad (2.66)$$

где  $h_{i+1} = x_{i+1} - x_i$  — шаг сетки. В этих формулах при  $i = 0$  и  $i = n - 1$  мы привлекаем известные нам из условия (II) значения  $\gamma_0$  и  $\gamma_n$  второй производной  $S''$  на левом и правом концах интервала  $[a, b]$ . Очевидно, что построенная таким образом функция  $S''(x)$  удовлетворяет условию «непрерывной склейки» в узлах  $x_1, x_2, \dots, x_{n-1}$ , т. е.

$$S''(x_i - 0) = S''(x_i + 0), \quad i = 1, 2, \dots, n - 1.$$

Чтобы восстановить  $S$  по  $S''$ , нужно теперь взять дважды первообразную (неопределённый интеграл) от  $S''(x)$ . Выполнив два раза интегрирование равенства (2.66), получим для  $x \in [x_i, x_{i+1}]$

$$S(x) = \gamma_i \frac{(x_{i+1} - x)^3}{6h_{i+1}} + \gamma_{i+1} \frac{(x - x_i)^3}{6h_{i+1}} + C_1 x + C_2 \quad (2.67)$$

с какими-то константами  $C_1$  и  $C_2$ . Но нам будет удобно представить это выражение в несколько другом виде:

$$S(x) = \gamma_i \frac{(x_{i+1} - x)^3}{6h_{i+1}} + \gamma_{i+1} \frac{(x - x_i)^3}{6h_{i+1}} + K_1(x_{i+1} - x) + K_2(x - x_i), \quad (2.68)$$

где  $K_1$  и  $K_2$  — тоже константы.<sup>17</sup> Насколько законен переход к такой форме? Из сравнения (2.67) и (2.68) следует, что  $C_1$  и  $C_2$  должны быть связаны с  $K_1$  и  $K_2$  посредством формул

$$\begin{aligned} C_1 &= -K_1 + K_2, \\ C_2 &= K_1 x_{i+1} - K_2 x_i. \end{aligned}$$

---

<sup>17</sup>Строго говоря, константы  $C_1, C_2, K_1, K_2$  нужно было бы снабдить ещё дополнительным индексом  $i$ , показывающим их зависимость от подинтервала  $[x_i, x_{i+1}]$ , к которому они относятся. Мы не делаем этого ради краткости изложения.

У выписанной системы линейных уравнений относительно  $K_1$  и  $K_2$  определитель равен  $x_i - x_{i+1} = -h_{i+1}$  и он не зануляется. Поэтому переход от  $C_1$  и  $C_2$  к  $K_1$  и  $K_2$  — это неособенная замена переменных. Следовательно, оба представления (2.67) и (2.68) совершенно равносильны друг другу.

Для определения  $K_1$  и  $K_2$  воспользуемся интерполяционными условиями. Подставляя в выражение (2.68) значения  $x = x_i$  и используя условия  $S(x_i) = y_i$ ,  $i = 0, 1, \dots, n - 1$ , будем иметь

$$\gamma_i \frac{(x_{i+1} - x_i)^3}{6h_{i+1}} + K_1(x_{i+1} - x_i) = y_i,$$

т. е.

$$\gamma_i \frac{h_{i+1}^2}{6} + K_1 h_{i+1} = y_i,$$

откуда

$$K_1 = \frac{y_i}{h_{i+1}} - \frac{\gamma_i h_{i+1}}{6}.$$

Совершенно аналогичным образом, подставляя в (2.68) значение  $x = x_{i+1}$  и используя условие  $S(x_{i+1}) = y_{i+1}$ , найдём

$$K_2 = \frac{y_{i+1}}{h_{i+1}} - \frac{\gamma_{i+1} h_{i+1}}{6}.$$

Выражение сплайна на подинтервале  $[x_i, x_{i+1}]$ ,  $i = 0, 1, \dots, n - 1$ , выглядит поэтому следующим образом:

$$\begin{aligned} S(x) &= y_i \frac{x_{i+1} - x}{h_{i+1}} + y_{i+1} \frac{x - x_i}{h_{i+1}} + \\ &+ \gamma_i \frac{(x_{i+1} - x)^3 - h_{i+1}^2(x_{i+1} - x)}{6h_{i+1}} + \gamma_{i+1} \frac{(x - x_i)^3 - h_{i+1}^2(x - x_i)}{6h_{i+1}}. \end{aligned} \quad (2.69)$$

Оно не содержит уже величин  $\alpha_i$ ,  $\beta_i$  и  $\vartheta_i$ , которые фигурировали в исходном представлении (2.64) для  $S(x)$ , но неизвестными остались  $\gamma_1$ ,  $\gamma_2, \dots, \gamma_{n-1}$  (напомним, что  $\gamma_0$  и  $\gamma_n$  даны по условию задачи).

Чтобы завершить определение вида сплайна, т. е. найти  $\gamma_1, \gamma_2, \dots, \gamma_{n-1}$ , можно воспользоваться условием непрерывности первой производной  $S'(x)$  в узлах  $x_1, x_2, \dots, x_{n-1}$ :

$$S'(x_i - 0) = S'(x_i + 0), \quad i = 1, 2, \dots, n - 1. \quad (2.70)$$

Продифференцировав по  $x$  формулу (2.69), будем иметь для подинтервала  $[x_i, x_{i+1}]$  представление

$$S'(x) = \frac{y_{i+1} - y_i}{h_{i+1}} - \gamma_i \frac{3(x_{i+1} - x)^2 - h_{i+1}^2}{6h_{i+1}} + \gamma_{i+1} \frac{3(x - x_i)^2 - h_{i+1}^2}{6h_{i+1}}. \quad (2.71)$$

Следовательно, с учётом того, что  $x_{i+1} - x_i = h_{i+1}$ , получим

$$\begin{aligned} S'(x_i) &= \frac{y_{i+1} - y_i}{h_{i+1}} - \gamma_i \frac{3(x_{i+1} - x_i)^2 - h_{i+1}^2}{6h_{i+1}} - \gamma_{i+1} \frac{h_{i+1}^2}{6h_{i+1}} = \\ &= \frac{y_{i+1} - y_i}{h_{i+1}} - \gamma_i \frac{h_{i+1}}{3} - \gamma_{i+1} \frac{h_{i+1}}{6}. \end{aligned} \quad (2.72)$$

С другой стороны, сдвигая все индексы в (2.71) на единицу назад, для подинтервала  $[x_{i-1}, x_i]$  будем иметь представление

$$S'(x) = \frac{y_i - y_{i-1}}{h_i} - \gamma_{i-1} \frac{3(x_i - x)^2 - h_i^2}{6h_i} + \gamma_i \frac{3(x - x_{i-1})^2 - h_i^2}{6h_i}.$$

Следовательно, с учётом того, что  $x_i - x_{i-1} = h_i$ , получим

$$\begin{aligned} S'(x_i) &= \frac{y_i - y_{i-1}}{h_i} + \gamma_{i-1} \frac{h_i^2}{6h_i} + \gamma_i \frac{3(x_i - x_{i-1})^2 - h_i^2}{6h_i} = \\ &= \frac{y_i - y_{i-1}}{h_i} + \gamma_{i-1} \frac{h_i}{6} + \gamma_i \frac{h_i}{3}. \end{aligned} \quad (2.73)$$

Приравнивание, согласно (2.70), производных (2.72) и (2.73), которые получены в узлах  $x_i$  с соседних подинтервалов  $[x_{i-1}, x_i]$  и  $[x_i, x_{i+1}]$ , приводит к соотношениям

$$\left\{ \begin{array}{l} \frac{h_i}{6} \gamma_{i-1} + \frac{h_i + h_{i+1}}{3} \gamma_i + \frac{h_{i+1}}{6} \gamma_{i+1} = \frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i}, \\ \gamma_0 \text{ и } \gamma_n \text{ заданы.} \end{array} \right. \quad i = 1, 2, \dots, n-1, \quad (2.74)$$

Это система линейных алгебраических уравнений с неизвестными

переменными  $\gamma_1, \gamma_2, \dots, \gamma_{n-1}$ , имеющая матрицу

$$\frac{1}{6} \begin{pmatrix} 2(h_1 + h_2) & h_2 & & & 0 \\ h_2 & 2(h_2 + h_3) & h_3 & & \\ & h_3 & 2(h_3 + h_4) & h_4 & \\ 0 & & \ddots & \ddots & \ddots \\ & & & h_{n-1} & 2(h_{n-1} + h_n) \end{pmatrix}$$

размера  $(n - 1) \times (n - 1)$ , в которой ненулевыми являются лишь главная диагональ и соседние с ней поддиагональ и наддиагональ. Такие матрицы называются *трёхдиагональными* (см. § 3.9). Кроме того, наша матрица зависит только от узлов  $x_i$  (но не от значений функции  $y_i$ ), а в правых частях системы (2.74) выражения

$$\frac{y_{i+1} - y_i}{h_{i+1}} \quad \text{и} \quad \frac{y_i - y_{i-1}}{h_i}$$

— это численные приближения к производным интерполируемой функции на подинтервалах  $[x_{i-1}, x_i]$  и  $[x_i, x_{i+1}]$  (см. § 2.8a).

Найдя из системы уравнений (2.74) значения  $\gamma_1, \gamma_2, \dots, \gamma_{n-1}$ , получим семейство формул (2.69) для определения искомого сплайна.

## 2.6в Интерполяционные кубические сплайны (продолжение)

Рассмотрим теперь построение кубического сплайна с краевыми условиями (I), т. е. когда заданы наклоны сплайна на концах интервала его определения:

$$S'(a) = \beta_0, \quad S'(b) = \beta_n.$$

Искать сплайн будем, как и ранее, в виде (2.64), который привязан к узлам сплайна. Этот вид далее был преобразован в (2.69), т. е.

$$\begin{aligned} S(x) &= y_i \frac{x_{i+1} - x}{h_{i+1}} + y_{i+1} \frac{x - x_i}{h_{i+1}} = \\ &+ \gamma_i \frac{(x_{i+1} - x)^3 - h_{i+1}^2(x_{i+1} - x)}{6h_{i+1}} + \gamma_{i+1} \frac{(x - x_i)^3 - h_{i+1}^2(x - x_i)}{6h_{i+1}}, \end{aligned}$$

который содержит параметры  $\gamma_0, \gamma_1, \dots, \gamma_n$ . Для краевых условий типа (II), рассмотренных в § 2.6б,  $\gamma_0$  и  $\gamma_n$  были известны, а неизвестными являлись  $\gamma_1, \gamma_2, \dots, \gamma_{n-1}$ . Теперь в силу краевых условий (I) значения  $\gamma_0$  и  $\gamma_n$  не даны по условию задачи и потому перешли в неизвестные. Чтобы организовать определённую систему уравнений для всех неизвестных  $\gamma_i, i = 0, 1, \dots, n$ , воспользуемся теми же соображениями, что и в предыдущем разделе, и ещё привлечём дополнительно краевые условия (I).

Ранее уже получена формула (2.71) для производной сплайна на подинтервале  $[x_i, x_{i+1}]$ ,  $i = 0, 1, \dots, n - 1$ :

$$S'(x) = \frac{y_{i+1} - y_i}{h_{i+1}} - \gamma_i \frac{3(x_{i+1} - x)^2 - h_{i+1}^2}{6h_{i+1}} + \gamma_{i+1} \frac{3(x - x_i)^2 - h_{i+1}^2}{6h_{i+1}}.$$

Возьмём её для  $i = 0$ , т. е. для подинтервала  $[x_0, x_1]$ , подставив  $x = x_0$ . С учётом того, что  $x_1 - x_0 = h_1$ , имеем

$$S'(x_0) = \frac{y_1 - y_0}{h_1} - \gamma_0 \frac{h_1}{3} - \gamma_1 \frac{h_1}{6}.$$

В силу краевых условий (I) эта производная должна быть равна  $S'(a) = \beta_0$ , так что получаем уравнение на  $\gamma_0$  и  $\gamma_1$ :

$$\frac{h_1}{3}\gamma_0 + \frac{h_1}{6}\gamma_1 = \frac{y_1 - y_0}{h_1} - \beta_0.$$

Возьмём теперь формулу (2.71) для  $i = n - 1$ , т. е. для подинтервала  $[x_{n-1}, x_n]$ , подставив  $x = x_n$ . С учётом того, что  $x_n - x_{n-1} = h_n$ , имеем

$$S'(x_n) = \frac{y_n - y_{n-1}}{h_n} + \gamma_{n-1} \frac{h_n}{6} + \gamma_n \frac{h_n}{3}.$$

В силу краевых условий (I) эта производная должна быть равна  $S'(b) = \beta_n$ , так что получаем ещё одно уравнение, теперь на  $\gamma_{n-1}$  и  $\gamma_n$ :

$$\frac{h_n}{6}\gamma_{n-1} + \frac{h_n}{3}\gamma_n = \beta_n - \frac{y_n - y_{n-1}}{h_n}.$$

Два новых уравнения вместе с уже полученными ранее  $n - 1$  уравнениями системы (2.74) образуют систему линейных алгебраических

уравнений

$$\left\{ \begin{array}{l} \frac{h_1}{3}\gamma_0 + \frac{h_1}{6}\gamma_1 = \frac{y_1 - y_0}{h_1} - \beta_0, \\ \frac{h_i}{6}\gamma_{i-1} + \frac{h_i + h_{i+1}}{3}\gamma_i + \frac{h_{i+1}}{6}\gamma_{i+1} = \frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i}, \\ \quad i = 1, 2, \dots, n-1, \\ \frac{h_n}{6}\gamma_{n-1} + \frac{h_n}{3}\gamma_n = \beta_n - \frac{y_n - y_{n-1}}{h_n}. \end{array} \right. \quad (2.75)$$

Матрица этой системы также трёхдиагональна и имеет вид

$$\frac{1}{6} \begin{pmatrix} 2h_1 & h_1 & & & & & 0 \\ h_1 & 2(h_1 + h_2) & h_2 & & & & \\ & h_2 & 2(h_2 + h_3) & h_3 & & & 0 \\ & & \ddots & \ddots & \ddots & & \\ & & & h_{n-1} & 2(h_{n-1} + h_n) & h_n & \\ 0 & & & & h_n & 2h_n & \end{pmatrix}$$

размера  $(n+1) \times (n+1)$ .

Полезное свойство матриц систем линейных алгебраических уравнений (2.74) и (2.75) — *диагональное преобладание* (см. § 3.4в, стр. 418): стоящие на их главных диагоналях элементы  $\frac{1}{3}h_1$ ,  $\frac{1}{3}(h_i + h_{i+1})$  и  $\frac{1}{3}h_n$  по модулю больше, чем суммы модулей внедиагональных элементов в соответствующих строках. В силу признака Адамара (он рассматривается далее в § 3.4в, стр. 419) такие матрицы неособенны. Как следствие, системы линейных уравнений (2.74) и (2.75) относительно неизвестных  $\gamma_i$  однозначно разрешимы при любых правых частях, а искомые кубические сплайны с краевыми условиями вида (I) или (II) всегда существуют и единственны.

Для нахождения решения систем (2.74) и (2.75) с трёхдиагональными матрицами может быть с успехом применён метод прогонки, описываемый ниже в § 3.9. Найдя неизвестные  $\gamma_i$ , подставим их в формулу (2.69), что даст выражения для полиномов, определяющих искомый сплайн на каждом из отдельных подинтервалов  $[x_i, x_{i+1}]$ .

## 2.6г Погрешность интерполяции с помощью кубических сплайнов

При описании погрешности интерполяции сплайнами и далее в этой книге будем пользоваться символом  $O(\cdot)$  — « $O$ -большое», который введён П. Бахманом и Э. Ландау и широко используется в математике и её приложениях для сравнения асимптотического поведения функций. Он естественно дополняет другой известный символ математического анализа — так называемое « $o$ -малое» [12, 40].

Для двух переменных величин  $u$  и  $v$ , участвующих в некотором асимптотическом процессе, пишут  $u = O(v)$ , если отношение  $u/v$  есть величина ограниченная. Иными словами,  $u = O(v)$  тогда и только тогда, когда существует константа  $C$ , такая что  $|u| \leq C|v|$  в этом процессе. В формулировке теоремы 2.6.1 ниже и в других ситуациях, где идёт речь о шаге сетки  $h$ , мы всюду имеем в виду  $h \rightarrow 0$ . Удобство использования символа  $O(\cdot)$  состоит в том, что, показывая качественный характер зависимости, он не требует явного выписывания констант, которые должны фигурировать в соответствующих отношениях.

**Теорема 2.6.1** Пусть  $f(x) \in C^p[a, b]$ ,  $p \in \{1, 2, 3, 4\}$ , а  $S(x)$  — интерполяционный кубический сплайн с краевыми условиями (I), (II) или (III), построенный по значениям  $f(x)$  на сетке  $a = x_0 < x_1 < \dots < x_n = b$  из интервала  $[a, b]$  с шагом  $h_i = x_i - x_{i-1}$ ,  $i = 1, 2, \dots, n$ , причём узлы интерполяции являются также узлами сплайна. Тогда для  $k \in \{0, 1, 2\}$ ,  $k \leq p$ , справедливо соотношение

$$\max_{x \in [a, b]} |f^{(k)}(x) - S^{(k)}(x)| = O(h^{p-k}),$$

где  $h = \max_{1 \leq i \leq n} h_i$ .

Фактически в формулировку теоремы 2.6.1 объединены несколько самостоятельных результатов, так что обоснование теоремы разбивается на ряд частных случаев, соответствующих различным значениям гладкости  $p$  и порядка производной  $k$ . Их доказательства можно увидеть, к примеру, в [11, 14, 35]. Там же читатель найдёт конкретные значения числовых констант, скрытых за символом  $O(\cdot)$  для различных частных случаев гладкости и порядка производной.

Повышение гладкости  $p$  интерполируемой функции  $f(x)$  выше, чем  $p = 4$ , уже не оказывает влияния на погрешность интерполяции, так как интерполяционный сплайн — кубический, т. е. имеет степень 3.

С другой стороны, свои особенности имеет также случай  $p = 0$ , когда интерполируемая функция всего лишь непрерывна, и мы не приводим здесь полную формулировку соответствующих результатов о погрешности (её можно найти, например, в книге [11]).

Отметим, что, в отличие от алгебраических интерполянтов, последовательность интерполяционных кубических сплайнов на равномерной сетке узлов всегда сходится к интерполируемой непрерывной функции. Это относится, в частности, к функции  $|x|$  из примера С.Н. Бернштейна и к функции  $\Upsilon(x) = 1/(1+x^2)$  из примера Рунге, рассмотренным выше в § 2.5. Обзор результатов о сходимости интерполяционных процессов со сплайнами и оценок погрешностей можно найти в журнальной работе [51]. Важно, что с повышением гладкости интерполируемой функции до определённого предела сходимость эта улучшается. В целом в задаче интерполяирования полиномиальные сплайны оказываются, как правило, лучшие алгебраических полиномов с точки зрения как вычислительных удобств, так и качества приближения, обеспечивая минимально возможную погрешность для заданного размера сетки. Подробности заинтересованный читатель может увидеть, к примеру, в книге [65].

В реальных задачах интерполяции для получения наилучших результатов приближения с помощью сплайнов следует аккуратно учитывать информацию о производных интерполируемой функции на концах интервала. Эти производные, первую или вторую, следует приближённо задать с необходимой точностью, пользуясь, к примеру формулами численного дифференцирования (см. § 2.8).

Интерполяирование с помощью сплайнов иллюстрирует интересное явление *насыщения* численных методов, когда, начиная с какого-то порядка, увеличение гладкости исходных данных задачи уже не приводит к увеличению точности результата. Соответствующие численные методы называют *насыщаемыми*.

Напротив, *ненасыщаемые численные методы* там, где их удаётся построить и применить, дают всё более точное решение при увеличении гладкости исходных данных [45]. Ненасыщаемые методы решения задач вычислительной математики очень привлекательны, но подчас весьма изощрённы и сложны в реализации. Целесообразность их использования в тех или иных конкретных ситуациях определяется трудностями практического определения гладкости данных, которые присутствуют в решаемой задаче. Примеры удачного применения ненасыщаемых алгоритмов относятся, как правило, к задачам, в которых

высокая гладкость входных данных известна априори, либо которые являются замкнутыми в себе и самодостаточными постановками, не привлекающими никаких дополнительных данных в виде правых частей или начальных (краевых) условий.

## 2.6д Экстремальное свойство кубических сплайнов

Сплайны  $S(x)$ , удовлетворяющие на концах рассматриваемого интервала  $[a, b]$  дополнительным условиям

$$S''(a) = S''(b) = 0, \quad (2.76)$$

называются *естественными* или *натуральными сплайнами*.

Одной из важных характеристик кривой является её *кривизна* — скорость изгибаия в зависимости от длины дуги кривой. Если плоская кривая является графиком функции  $y = f(x)$  в декартовой системе координат, то, как показывается в курсах дифференциальной геометрии, её кривизна в точке  $x$  равна

$$\frac{f''(x)}{\left(1 + (f'(x))^2\right)^{3/2}} \quad (2.77)$$

(см. подробности в [40, 84]). Таким образом, естественные сплайны — это сплайны с нулевой кривизной на концах интервала своей области определения.

Замечательное свойство естественных кубических сплайнов состоит в том, что они минимизируют функционал

$$\mathcal{E}(f) = \int_a^b (f''(x))^2 dx. \quad (2.78)$$

Более точно, справедлива

**Теорема 2.6.2** (теорема Холладея) *Если  $S(x)$  — естественный интерполяционный кубический сплайн, построенный на интервале  $[a, b]$  по узлам  $a = x_0 < x_1 < \dots < x_n = b$ , а  $\varphi(x)$  — любая другая дважды гладкая функция, принимающая в этих узлах те же значения, что и  $S(x)$ , то  $\mathcal{E}(\varphi) \geq \mathcal{E}(S)$ , причём неравенство строго для  $\varphi \neq S$ .*

Доказательство этого важного факта не очень сложно, его можно найти в оригинальной работе [101] или, к примеру, в книгах [1, 11, 38,

55]. Нетрудно показать, что утверждение теоремы Холладея выполняется также для более общих краевых условий на сплайн, которые не обязательно требуют зануления вторых производных на концах интервала. Соответствующие результаты можно увидеть в [42, 55].

Интеграл (2.78) приближённо пропорционален энергии деформации гибкой упругой линейки, форма которой описывается функцией  $f(x)$  на интервале  $[a, b]$ . Краевые условия (2.76) соответствуют при этом линейке, свободно закреплённой на концах, где не приложены моменты каких-либо внешних сил.

В самом деле, потенциальная энергия изгибаия малого участка упругого тела, как известно, пропорциональна квадрату его кривизны в данной точке. Поэтому в силу (2.77) энергия упругой деформации однородной линейки, принимающей форму кривой  $y = f(x)$  на интервале  $[a, b]$ , выражается интегралом

$$\int_a^b \frac{\kappa}{(1 + (f'(x))^2)^{3/2}} (f''(x))^2 dx, \quad (2.79)$$

где  $\kappa$  — коэффициент, характеризующий свойства материала линейки. При условии приблизительного постоянства  $f'(x)$  значения интеграла (2.79) пропорциональны значениям (2.78). Если упругая линейка закреплена в узлах интерполирования, позволяющих ей свободно изгибаться, то, будучи предоставленной самой себе, она принимает форму, которая, как известно из физики, должна минимизировать энергию своей деформации. Из теоремы Холладея следует, что эта форма очень близка к естественному кубическому сплайну.

С другой стороны, теорема Холладея указывает ещё одно преимущество сплайнов в практических задачах интерполяции и приближения. Это лучшая «физическая реализуемость» сплайнов, когда с их помощью строится решение задачи, «менее искривляемое» и более удобное в изготовлении, чем абстрактное математическое решение задачи, которое может, к примеру, слишком сильно изгибаться, быть менее технологичным в производстве и т. п.

Сформулированное в теореме Холладея свойство называют *экстремальным свойством* естественных сплайнов.<sup>18</sup> Оно служит началом большого и интересного направления — вариационной теории сплайнов, в котором сплайны вводятся и рассматриваются как решения некоторых задач на минимум [50, 72]. Например, интерполяционный ку-

---

<sup>18</sup>Иногда также говорят о *вариационном свойстве* естественных сплайнов.

бический сплайн вместо определения, данного в начале § 2.6б, можно равносильным образом определить как дважды непрерывно дифференцируемую функцию, которая в узлах удовлетворяет интерполяционным условиям и доставляет минимум интегралу (2.78). Аналогично определяются сплайны более высоких степеней, которые минимизируют интегралы вида

$$\int_a^b (f^{(q)}(x))^2 dx$$

для заданного натурального  $q$ . Этот подход к сплайнам позволяет рассматривать их с единой точки зрения, а также преодолевает то затруднение, что сплайны изначально появляются как функции, конструктивно не очень удобные или даже «не очень естественные», задаваемые на различных участках своей области определения разными аналитическими формулами. Вариационная теория сплайнов показывает, что это необходимо по существу дела.

В заключение темы отметим, что практика нередко требует от приближающей или интерполирующей функции удовлетворения некоторым глобальным геометрическим условиям — монотонности, выпуклости, наличию участков плато и т. п. Такие задачи приближения и интерполяции с дополнительными геометрическими условиями называют задачами *изогеометрического приближения* или *изогеометрической интерполяции* соответственно. Способы построения изогеометрических приближений с помощью сплайнов рассматриваются в [62].

## 2.7 Нелинейные методы интерполяции

Рассмотренные выше методы интерполяции (в частности, алгебраической) были *линейными* в том смысле, что результат решения задачи интерполяции при фиксированных узлах линейно зависел от данных. В этих условиях класс интерполирующих функций  $\mathcal{S}$  можно наделить структурой линейного векторного пространства над полем вещественных чисел  $\mathbb{R}$ : любая линейная комбинация функций тоже является функцией заданного вида, решающей задачу интерполяции для линейной комбинации данных. Но существуют и другие, нелинейные, методы интерполяирования, для которых сформулированное выше свойство не выполнено. Эти методы тоже широко применяются при практической интерполяции, так как они обладают многими важными достоинствами.

Итак, *нелинейными* называют методы интерполяции, в которых интерполирующая функция зависит от параметров, её определяющих, нелинейным образом. Решение задачи нелинейной интерполяции обычно сводится к решению системы нелинейных уравнений.

Рассмотрим подробнее важнейший частный случай нелинейных методов интерполяции — интерполяцию с помощью рациональных функций вида

$$y = y(x) = \frac{a_0 + a_1x + a_2x^2 + \dots}{b_0 + b_1x + b_2x^2 + \dots}. \quad (2.80)$$

Если в узлах  $x_0, x_1, \dots, x_n$  заданы значения функции  $y_0, y_1, \dots, y_n$ , то нужно найти рациональную дробь вида (2.80), такую что  $y_i = y(x_i)$ ,  $i = 0, 1, \dots, n$ .

Для целей интерполяции и приближения рациональные функции часто более предпочтительны, чем алгебраические полиномы, так как лучше способны передавать особенности поведения функций с особыми точками, которые являются так называемыми полюсами [85]. Особыми точками в теории функций называют точки, где функция принимает бесконечно большие значения, и для рациональных функций вида (2.80) они обычно являются нулями знаменателя. Подобные полюса могут присутствовать также у интерполируемой вещественной функции, но гораздо чаще встречается ситуация, когда функция конечна для любых конечных вещественных аргументов, но полюсы имеются у её аналитического продолжения в область комплексной плоскости, непосредственно примыкающую к интервалу интерполяирования на вещественной оси. Такие близкие полюса могут сильно ухудшить приближение и интерполирование с помощью алгебраических полиномов даже на вещественной оси, аналогично тому, как они разрушают сходимость бесконечных степенных рядов.

Из результатов теории функций следует, что для оценки успешности полиномиальной интерполяции нужно построить в комплексной плоскости круг, содержащий все узлы интерполяции, и определить его расположение относительно полюсов интерполируемой функции (см. подробности, например, в [7]). Хорошая полиномиальная интерполяция возможно лишь в случае, когда эти полюса находятся достаточно далеко от построенного круга. Напротив, приближение и интерполяция рациональной функцией будет успешной, если в её знаменателе имеются достаточно высокие степени аргумента для правильной передачи поведения функции в полюсах.

Поскольку дробь не меняется от умножения числителя и знаменателя на одно и то же ненулевое число, то для какого-нибудь одного из коэффициентов  $a_i$  или  $b_i$ ,  $i = 1, 2, \dots$ , может быть выбрано произвольное наперёд заданное значение. Кроме того, коэффициенты  $a_i$  и  $b_i$  должны быть такими, что удовлетворяются  $n + 1$  условий интерполяции в узлах. Как следствие, всего мы можем извлечь из постановки задачи  $n + 2$  условий на коэффициенты числителя и знаменателя. Этим задаётся общее число неизвестных, которое мы можем определить из задачи, т. е. сумма степени  $\mu$  полинома числителя и степени  $\nu$  полинома знаменателя в дроби (2.80). Должно выполняться соотношение  $(\mu + 1) + (\nu + 1) = n + 2$ , так что  $\mu + \nu = n$ .

Как найти коэффициенты  $a_0, a_1, \dots, a_\mu, b_0, b_1, \dots, b_\nu$ ? Умножая обе части равенства (2.80) на знаменатель дроби, получим

$$a_0 + a_1x + \dots + a_\mu x^\mu = (b_0 + b_1x + \dots + b_\nu x^\nu)y,$$

или

$$a_0 + a_1x + \dots + a_\mu x^\mu - (b_0y + b_1xy + \dots + b_\nu x^\nu y) = 0.$$

Подставляя в это равенство интерполяционные данные  $x_i$  и  $y_i$ ,  $i = 0, 1, \dots, n$ , получим систему из  $(n + 1)$ -го линейного алгебраического уравнения относительно неизвестных коэффициентов  $a_0, a_1, \dots, a_\mu, b_0, b_1, \dots, b_\nu$ , один из которых уже зафиксирован:

$$\left\{ \begin{array}{l} \sum_{j=0}^{\mu} x_i^j a_j - \sum_{j=0}^{\nu} x_i^j y_i b_j = 0, \\ i = 0, 1, \dots, n. \end{array} \right. \quad (2.81)$$

Решение этой системы уравнений определяет искомый рациональный интерполант.

Несмотря на технологическую простоту описанного выше решения задачи, мы не можем гарантировать, что с его помощью всегда будет получен нужный интерполант. Это вызвано возможным занулением знаменателя дроби (2.80) и трудностью исследования системы линейных алгебраических уравнений (2.81), свойства которой, вообще говоря, могут быть плохими.

**Пример 2.7.1** Построим рациональную интерполяцию данных

$x$	−1	0	1	
$y$	1	0	1	

Они принимаются функцией  $y = |x|$ .

Три узла соответствуют  $n = 2$ , что должно быть равно сумме степеней числителя и знаменателя интерполянта. Станем искать его в виде

$$y = \frac{a_0 + x}{b_0 + b_1 x},$$

т. е. зафиксировав значение  $a_1 = 1$ . Отсюда

$$a_0 + x = b_0 y + b_1 x y. \quad (2.82)$$

Подставляя в (2.82) интерполяционные данные, получим систему линейных алгебраических уравнений

$$\begin{cases} a_0 - 1 = b_0 - b_1, \\ a_0 = 0, \\ a_0 + 1 = b_0 + b_1. \end{cases}$$

Её решение —  $a_0 = 0$ ,  $b_0 = 0$ ,  $b_1 = 1$ , и потому искомым рациональным интерполянтом является функция

$$y = \frac{x}{x}.$$

Она непригодна в качестве решения задачи, так как при  $x = 0$  не определена. Даже если значением в нуле положить его предел при  $x \rightarrow 0$ , то получим 1, что значительно отличается от требуемого в нуле по условию. ■

**Пример 2.7.2** В качестве позитивного примера рассмотрим интерполяцию дробно-рациональной функцией таблицы значений

$x$	—1	0	1	2	
$y$	0.5	1	2	4	

Они принимаются функцией  $y = 2^x$ .

В данном случае  $n = 3$ , и мы можем взять дробно-рациональный интерполянт, к примеру, в виде

$$g(x) = \frac{a_0 + a_1 x + x^2}{b_0 + b_1 x},$$

зафиксировав значение  $a_2 = 1$ .

Составляем систему линейных уравнений (2.81), которая после равносильных преобразований принимает вид

$$\begin{cases} 1a_0 - 1a_1 - 0.5b_0 + 0.5b_1 = -1, \\ 1a_0 + 0a_1 - 1b_0 - 0b_1 = 0, \\ 1a_0 + 1a_1 - 2b_0 - 2b_1 = -1, \\ 1a_0 + 2a_1 - 4b_0 - 8b_1 = -4. \end{cases}$$

Её решением (которое можно быстро найти в какой-нибудь системе компьютерной математики) является  $(10, 5, 10, -2)^\top$ , так что искомый рациональный интерполянт имеет вид

$$g(x) = \frac{10 + 5x + x^2}{10 - 2x}. \quad (2.83)$$

Полученный интерполянт настолько хорош, что практически сливаются с графиком экспоненты  $2^x$  на интервале значений аргумента  $[-1.2, 2.2]$  (в чём можно убедиться с помощью любой программы построения графиков функций). В чебышёвской метрике его отклонение от функции  $2^x$  составляет всего 0.0025 на  $[-1, 2]$ .

Алгебраическим интерполянтом по данным примера является полином

$$1 + \frac{2}{3}x + \frac{1}{4}x^2 + \frac{1}{12}x^3,$$

для которого в чебышёвской метрике на интервале  $[-1, 2]$  отклонение от функции  $2^x$  равно 0.017. Это почти в 7 (семь) раз хуже, чем у рационального интерполянта (2.83). ■

Опишем ещё один способ решения задачи рациональной интерполяции, эквивалентный изложенному выше, но, возможно, более удобный в некоторых ситуациях [75]. Представление (2.80) равносильно тождеству

$$a_0 - b_0y + a_1x - b_1xy + a_2x^2 - b_2x^2y + \dots = 0. \quad (2.84)$$

Коль скоро при  $x = x_i$  должно быть  $y = y_i$ ,  $i = 0, 1, \dots, n$ , то получаем ещё  $(n+1)$  числовых равенств

$$a_0 - b_0y_i + a_1x_i - b_1x_iy_i + a_2x_i^2 - b_2x_i^2y_i + \dots = 0, \quad (2.85)$$

$i = 0, 1, \dots, n$ . Соотношения (2.84) и (2.85) можно трактовать как условие линейной зависимости, с коэффициентами  $a_0, -b_0, a_1, -b_1, \dots$ , для

вектор-столбцов

$$\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} y \\ y_0 \\ \vdots \\ y_1 \\ y_n \end{pmatrix}, \quad \begin{pmatrix} x \\ x_0 \\ \vdots \\ x_1 \\ x_n \end{pmatrix}, \quad \begin{pmatrix} xy \\ x_0y_0 \\ \vdots \\ x_1y_1 \\ x_ny_n \end{pmatrix}, \quad \begin{pmatrix} x^2 \\ x_0^2 \\ \vdots \\ x_1^2 \\ x_n^2 \end{pmatrix}, \quad \begin{pmatrix} x^2y \\ x_0^2y_0 \\ \vdots \\ x_1^2y_1 \\ x_n^2y_n \end{pmatrix}, \quad \dots$$

размера  $(n + 2)$ . Как следствие, определитель

$$\det \begin{pmatrix} 1 & y & x & xy & x^2 & x^2y & \dots \\ 1 & y_0 & x_0 & x_0y_0 & x_0^2 & x_0^2y_0 & \dots \\ 1 & y_1 & x_1 & x_1y_1 & x_1^2 & x_1^2y_1 & \dots \\ 1 & y_2 & x_2 & x_2y_2 & x_2^2 & x_2^2y_2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\ 1 & y_n & x_n & x_ny_n & x_n^2 & x_n^2y_n & \dots \end{pmatrix}$$

составленной из этих столбцов матрицы размера  $(n+2) \times (n+2)$  должен быть равен нулю. Разложим этот определитель по первой строке (см., к примеру, [22]), содержащей одночлены от переменных  $x$  и  $y$ . Коэффициенты этого разложения будут значениями определителей числовых матриц, т. е. просто числами. Полученное равенство нулю разрешим затем относительно переменной  $y$  — она входит во все слагаемые в степени не выше первой. В результате получим выражение для  $y$  в виде отношения двух многочленов от  $x$ .

Реализация описанного выше приёма требует нахождения значений определителя числовых  $(n + 1) \times (n + 1)$ -матриц, и далее в § 3.14 рассматриваются соответствующие численные методы решения этой задачи. Отметим, что в популярных системах компьютерной математики Scilab, MATLAB, Octave, Maple, Mathematica и др. для этого существует готовая встроенная функция `det`.

Дальнейшие сведения по рациональной интерполяции и её применением интересующийся читатель может найти, к примеру, в книгах [109], § 2.2, и [48].

## 2.8 Численное дифференцирование

Дифференцированием называется, как известно, процесс нахождения производной от заданной функции или же численного значения этой производной в заданной точке. Необходимость выполнения дифференцирования возникает весьма часто и вызвана огромным распространением этой операции в современной математике и её приложениях. Производная бывает нужна и сама по себе как мгновенная скорость тех или иных процессов и как вспомогательное средство для построения более сложных вычислительных технологий, например, в методе Ньютона для численного решения уравнений и систем уравнений (§ 4.4д и 4.5б).

В настоящее время наиболее распространены три следующих способа вычисления производных:

- символьное (аналитическое) дифференцирование,
- численное дифференцирование,
- алгоритмическое (автоматическое) дифференцирование.

*Символьным (или аналитическим) дифференцированием* называют процесс построения по функции, задаваемой каким-то выражением, производной функции, основываясь на известных из математического анализа правилах дифференцирования составных функций (суммы, разности, произведения, частного, композиции, обратной функции и т. п.) и известных производных для простейших функций. Основы символьного (аналитического) дифференцирования являются предметом математического анализа (точнее, дифференциального исчисления), а более продвинутые результаты по этой теме входят в курсы компьютерной алгебры.

При *алгоритмическом (автоматическом) дифференцировании* опирают не символьными представлениями выражений для функции и производных, как в символьном (аналитическом) дифференцировании, а их численными значениями при заданных значениях аргументов функции. Алгоритмическое (автоматическое) дифференцирование тоже требует знания выражения для функции (или хотя бы компьютерной программы для её вычисления), но использует это выражение по-иному. Мы кратко рассмотрим алгоритмическое дифференцирование в § 2.9.

*Численным дифференцированием* называется процесс нахождения

значения производной от функции, который использует значения этой функции в некотором наборе точек её области определения. Таким образом, если функция задана таблично (т. е. лишь на конечном множестве значений аргумента) либо определение значений этой функции не задаётся выражением или детерминированной программой, то альтернатив численному дифференцированию нет. Иногда в виде такого «чёрного ящика» мы вынуждены представлять вычисление значений функции, аналитическое выражение для которой существует, но является слишком сложным или неудобным для дифференцирования первыми двумя способами.

В основе методов численного дифференцирования лежат различные идеи. Самая первая состоит в том, чтобы доопределить (восстановить) таблично заданную функцию до функции непрерывного аргумента, к которой уже применима обычная операция дифференцирования. Теория интерполяирования, рассмотренная в предшествующих параграфах, оказывается в высшей степени полезной при реализации такого подхода. Таблично заданную функцию можно заменить её интерполяционным полиномом и его производные считать производными рассматриваемой функции. Для этого годится также интерполяция сплайнами или какими-либо другими функциями, а в целом описанный выше подход к численному дифференцированию называют *интерполяционным подходом*.

## 2.8а Интерполяционный подход

Итак, пусть задан набор узлов  $x_0, x_1, \dots, x_n \in [a, b]$ , т. е. сетка с шагом  $h_i = x_i - x_{i-1}$ ,  $i = 1, 2, \dots, n$ . Кроме того, заданы значения функции  $f_0, f_1, \dots, f_n$ , такие что  $f_i = f(x_i)$ ,  $i = 0, 1, \dots, n$ . Ниже мы рассмотрим простейший вариант интерполяционного подхода, в котором используется алгебраическая интерполяция.

Начнём со случая, когда применяется интерполяционный полином первой степени, который строится по двум соседним узлам сетки, т. е. по  $x_{i-1}$  и  $x_i$ ,  $i = 1, 2, \dots, n$ :

$$\begin{aligned} P_{1,i}(x) &= \frac{x - x_i}{x_{i-1} - x_i} f_{i-1} + \frac{x - x_{i-1}}{x_i - x_{i-1}} f_i = \\ &= \frac{f_i - f_{i-1}}{x_i - x_{i-1}} x + \frac{f_{i-1}x_i - f_ix_{i-1}}{x_i - x_{i-1}}, \end{aligned}$$

где у интерполяционного полинома добавлен дополнительный индекс « $i$ », указывающий на ту пару узлов, по которым он построен. Поэтому производная равна

$$P'_{1,i}(x) = \frac{f_i - f_{i-1}}{x_i - x_{i-1}} = \frac{f_i - f_{i-1}}{h_i}.$$

Это значение можно взять за приближение к производной от рассматриваемой функции на интервале  $[x_{i-1}, x_i]$ ,  $i = 1, 2, \dots, n$ .

Во внутренних узлах сетки —  $x_1, x_2, \dots, x_{n-1}$ , т. е. там, где встречаются два подинтервала, производную можно брать по любой из возможных формул

$$f'(x_i) \approx f_{\bar{x},i} := \frac{f_i - f_{i-1}}{x_i - x_{i-1}} = \frac{f_i - f_{i-1}}{h_i} \quad (2.86)$$

— разделённая разность назад,

$$f'(x_i) \approx f_{x,i} := \frac{f_{i+1} - f_i}{x_{i+1} - x_i} = \frac{f_{i+1} - f_i}{h_{i+1}} \quad (2.87)$$

— разделённая разность вперёд.

Обе они примерно равнозначны, и выбор конкретной из них может быть делом соглашения, удобства или целесообразности. Например, от направления этой разности может решающим образом зависеть устойчивость разностных схем для численного решения дифференциальных уравнений.

Построим теперь интерполяционные полиномы Лагранжа второй степени по трём соседним точкам сетки  $x_{i-1}, x_i, x_{i+1}$ ,  $i = 1, 2, \dots, n-1$ .

Имеем

$$\begin{aligned}
 P_{2,i}(x) &= \frac{(x - x_i)(x - x_{i+1})}{(x_{i-1} - x_i)(x_{i-1} - x_{i+1})} f_{i-1} + \frac{(x - x_{i-1})(x - x_{i+1})}{(x_i - x_{i-1})(x_i - x_{i+1})} f_i + \\
 &\quad + \frac{(x - x_{i-1})(x - x_i)}{(x_{i+1} - x_{i-1})(x_{i+1} - x_{i-1})} f_{i+1} = \\
 &= \frac{x^2 - (x_i + x_{i+1})x + x_i x_{i+1}}{(x_{i-1} - x_i)(x_{i-1} - x_{i+1})} f_{i-1} + \\
 &\quad + \frac{x^2 - (x_{i-1} + x_{i+1})x + x_{i-1} x_{i+1}}{(x_i - x_{i-1})(x_i - x_{i+1})} f_i + \\
 &\quad + \frac{x^2 - (x_{i-1} + x_i)x + x_{i-1} x_i}{(x_{i+1} - x_{i-1})(x_{i+1} - x_i)} f_{i+1}.
 \end{aligned}$$

Поэтому

$$\begin{aligned}
 P'_{2,i}(x) &= \frac{2x - (x_i + x_{i+1})}{(x_{i-1} - x_i)(x_{i-1} - x_{i+1})} f_{i-1} + \frac{2x - (x_{i-1} + x_{i+1})}{(x_i - x_{i-1})(x_i - x_{i+1})} f_i + \\
 &\quad + \frac{2x - (x_{i-1} + x_i)}{(x_{i+1} - x_{i-1})(x_{i+1} - x_i)} f_{i+1}.
 \end{aligned}$$

Воспользуемся теперь тем, что  $x_i - x_{i-1} = h_i$ ,  $x_{i+1} - x_i = h_{i+1}$ . Тогда  $x_{i+1} - x_{i-1} = h_i + h_{i+1}$ , а результат предшествующих выкладок может быть записан в виде

$$\begin{aligned}
 f'(x) \approx P'_{2,i}(x) &= \frac{2x - x_i - x_{i+1}}{h_i(h_i + h_{i+1})} f_{i-1} - \\
 &\quad - \frac{2x - x_{i-1} - x_{i+1}}{h_i h_{i+1}} f_i + \frac{2x - x_{i-1} - x_i}{h_{i+1}(h_i + h_{i+1})} f_{i+1}. \tag{2.88}
 \end{aligned}$$

Формула (2.88) может применяться при вычислении значения производной в произвольной точке  $x$  для случая общей неравномерной сетки. Предположим теперь для простоты, что сетка равномерна, т. е.  $h_i = h = \text{const}$ ,  $i = 1, 2, \dots, n$ . Кроме того, для таблично заданной функции на практике обычно интересны производные в тех же точках, где задана сама функция, т. е. в узлах  $x_0, x_1, \dots, x_n$ . В точке  $x = x_i$  из (2.88) для первой производной получаем формулу

$$f'(x_i) \approx f'_{\hat{x},i} = \frac{f_{i+1} - f_{i-1}}{2h}, \tag{2.89}$$

называемую *формулой центральной разности*. Подставляя в (2.88) аргумент  $x = x_{i-1}$  и сдвигая в получающемся результате индекс на +1, получим

$$f'(x_i) \approx \frac{-3f_i + 4f_{i+1} - f_{i+2}}{2h}.$$

Подставляя в (2.88) аргумент  $x = x_{i+1}$  и сдвигая в получающемся результате индекс на (-1), получим

$$f'(x_i) \approx \frac{f_{i-2} - 4f_{i-1} + 3f_i}{2h}.$$

Зайдёмся теперь выводом формул для второй производной. Используя интерполяционный полином второй степени, можно найти

$$f''(x_i) \approx P''_{2,i}(x) = \frac{2}{h_i(h_i + h_{i+1})} f_{i-1} - \frac{2}{h_i h_{i+1}} f_i + \frac{2}{h_{i+1}(h_i + h_{i+1})} f_{i+1}.$$

В частности, на равномерной сетке с  $h_i = h = \text{const}$ ,  $i = 1, 2, \dots, n$ , имеем

$$f''(x_i) \approx \frac{f_{i-1} - 2f_i + f_{i+1}}{h^2}. \quad (2.90)$$

Эта формула широко используется в вычислительной математике и по аналогии с (2.86) и (2.87) часто обозначается кратко как  $f_{x\bar{x}}$ . Естественно, что полученные выражения для второй производной не зависят от аргумента  $x$ .

Несмотря на то что проведённые выше рассуждения основывались на применении интерполяционного полинома Лагранжа, для взятия производных произвольных порядков на сетке общего вида удобнее использовать интерполяционный полином Ньютона, в котором члены являются полиномами возрастающих степеней.

Выпишем ещё без вывода формулы численного дифференцирования на равномерной сетке, полученные по четырём точкам, т. е. с применением интерполяционного полинома третьей степени: для первой

производной

$$f'(x_i) \approx \frac{1}{6h} (-11f_i + 18f_{i+1} - 9f_{i+2} + 2f_{i+3}), \quad (2.91)$$

$$f'(x_i) \approx \frac{1}{6h} (-2f_{i-1} - 3f_i + 6f_{i+1} - f_{i+2}), \quad (2.92)$$

$$f'(x_i) \approx \frac{1}{6h} (f_{i-2} - 6f_{i-1} + 3f_i + 2f_{i+1}), \quad (2.93)$$

$$f'(x_i) \approx \frac{1}{6h} (-2f_{i-3} + 9f_{i-2} - 18f_{i-1} + 11f_i), \quad (2.94)$$

для второй производной

$$f''(x_i) \approx \frac{1}{h^2} (2f_i - 5f_{i+1} + 4f_{i+2} - f_{i+3}), \quad (2.95)$$

$$f''(x_i) \approx \frac{1}{h^2} (f_{i-1} - 2f_i + f_{i+1}), \quad (2.96)$$

$$f''(x_i) \approx \frac{1}{h^2} (-f_{i-3} + 4f_{i-2} - 5f_{i-1} + 2f_i). \quad (2.97)$$

В формуле (2.96) один из четырёх узлов, по которым строилась формула, никак не используется, а сама формула совпадает с формулой (2.90), полученной по трём точкам. Отметим красивую двойственность формул (2.91) и (2.94), (2.92) и (2.93), (2.95) и (2.97). Неслучаен также тот факт, что сумма коэффициентов при значениях функции в узлах во всех формулах равна нулю: он является следствием того, что производная постоянной функции — нуль.



Рис. 2.19. Шаблон формулы второй разностной производной (2.90)

В связи с численным дифференцированием и во многих других вопросах вычислительной математики чрезвычайно полезно понятие шаблона (сеточной) формулы, под которым будем понимать совокупность охватываемых этой формулой узлов сетки. Более точно, *шаблон формулы* численного дифференцирования — это множество узлов

сетки, входящих в правую часть этой формулы, явным образом либо в качестве аргументов используемых значений функции. Например, шаблоном формулы (2.90) для вычисления второй производной на равномерной сетке

$$f''(x_i) \approx \frac{f_{i-1} - 2f_i + f_{i+1}}{h^2}$$

являются три точки —  $x_{i-1}$ ,  $x_i$ ,  $x_{i+1}$  (рис. 2.19), в которых должны быть заданы  $f_{i-1}$ ,  $f_i$ ,  $f_{i+1}$ . Особенno разнообразны формы шаблонов операторов дифференцирования в случае двух и более независимых переменных.

## 2.8б Оценка погрешности численного дифференцирования

Пусть для численного нахождения  $k$ -й производной функции применяется формула численного дифференцирования  $\Phi$ , имеющая шаблон  $\Theta$  и использующая значения функции в узлах этого шаблона. Если  $f(x)$  — дифференцируемая необходимое число раз функция, такая что  $f_i = f(x_i)$  для всех узлов  $x_i \in \Theta$ , то какова может быть погрешность вычисления  $f^{(k)}(x)$  по формуле  $\Phi$ ? Вопрос этот можно адресовать как к целому интервалу значений аргумента, так и локально, только к той точке  $x_i$ , которая служит аргументом левой части формулы численного дифференцирования.

Если рассматриваемая формула выведена в рамках интерполяционного подхода, то заманчивой идеей является получение ответа прямым дифференцированием полученных ранее выражений (2.29) и (2.30) для погрешности интерполирования. Этот путь оказывается очень непростым, так как применение, к примеру, выражения (2.30) требует достаточной гладкости функции  $\xi(x)$ , о которой мы можем сказать немногое. Даже если эта гладкость имеется у  $\xi(x)$ , полученные оценки будут содержать производные  $\xi'(x)$  и пр., о которых мы знаем ещё меньше. Наконец, шаблон некоторых формул численного дифференцирования содержит меньше точек, чем это необходимо для построения интерполяционных полиномов нужной степени. Такова, к примеру, формула «центральной разности» для первой производной или формула для второй производной (2.96), построенная по четырём точкам на основе полинома 3-й степени. Тем не менее явные выражения для остаточно-го члена формул численного дифференцирования на этом пути можно

получить методом, который напоминает вывод формулы для погрешности алгебраического интерполирования. Подробности изложены, к примеру, в книгах [20, 28].

Рассмотрим детально более простой и достаточно универсальный способ оценивания погрешностей — *метод локальных разложений*, который основан на применении формулы Тейлора для гладких функций. Этот способ заключается, во-первых, в выписывании по формуле Тейлора разложений для функций, входящих в правую часть формулы численного дифференцирования, и, во-вторых, в аккуратном учёте членов этих разложений с целью получить, по возможности, наиболее точное выражение для ошибки.

Поясним эту методику на примере оценки погрешности для формулы «центральной разности» (2.89):

$$f'(x_i) \approx f_{\dot{x},i} = \frac{f_{i+1} - f_{i-1}}{2h}.$$

Предположим, что  $f \in C^3[x_{i-1}, x_{i+1}]$ , т. е. функция  $f$  трижды непрерывно дифференцируема на интервале между узлами формулы. Подставляя её в (2.89) и разлагая относительно точки  $x_i$  по формуле Тейлора с остаточным членом в форме Лагранжа вплоть до членов второго порядка, получим

$$\begin{aligned} f_{\dot{x},i} &= \frac{1}{2h} \left( \left( f(x_i) + hf'(x_i) + \frac{h^2}{2} f''(x_i) + \frac{h^3}{6} f'''(\xi_+) \right) - \right. \\ &\quad \left. - \left( f(x_i) - hf'(x_i) + \frac{h^2}{2} f''(x_i) - \frac{h^3}{6} f'''(\xi_-) \right) \right) = \\ &= f'(x_i) + \frac{h^2}{12} f'''(\xi_+) + \frac{h^2}{12} f'''(\xi_-), \end{aligned}$$

где  $\xi_+$  и  $\xi_-$  — некоторые точки из открытого интервала  $]x_{i-1}, x_{i+1}[$ . Поэтому

$$f_{\dot{x},i} - f'(x_i) = \frac{h^2}{12} (f'''(\xi_+) + f'''(\xi_-)) = \frac{\alpha h^2}{6},$$

где  $\alpha = \frac{1}{2}(f'''(\xi_+) + f'''(\xi_-))$ . В целом справедлива оценка

$$|f_{\dot{x},i} - f'(x_i)| \leq \frac{M_3}{6} h^2,$$

в которой  $M_3 = \max_{\xi} |f'''(\xi)|$  для  $\xi \in ]x_{i-1}, x_{i+1}[$ . То есть на трижды непрерывно дифференцируемых функциях погрешность вычисления производной по формуле «центральной разности» равна  $O(h^2)$  для равномерной сетки шага  $h$ .

## 2.8в Порядок точности формул и методов

**Определение 2.8.1** Станем говорить, что приближённая формула (численного дифференцирования, интегрирования и т. п.) или приближённый численный метод имеют  $p$ -й порядок точности (порядок аппроксимации), если на равномерной сетке с шагом  $h$  их погрешность является величиной  $O(h^p)$ , т. е. не превосходит  $Ch^p$ , где  $C$  — константа, не зависящая от  $h$ .

Нередко понятие порядка точности распространяют и на неравномерные сетки, в которых шаг  $h_i$  меняется от узла к узлу. Тогда роль величины  $h$  играет какой-нибудь «характерный размер», описывающий данную сетку, например  $h = \max_i h_i$ . Порядок точности — важная количественная мера погрешности формулы или метода, и при прочих равных условиях более предпочтительной является та формула или тот метод, которые имеют более высокий порядок точности. Но следует чётко осознавать, что порядок точности имеет асимптотический характер и отражает поведение погрешности при стремлении шагов сетки к нулю. Если этого стремления нет и шаг сетки остаётся «достаточно большим», то вполне возможны ситуации, когда метод меньшего порядка точности даёт лучшие результаты, поскольку множитель при  $h^p$  в оценке погрешности у него меньше.

Итак, порядок точности приближённой формулы или вычислительного метода — это показатель степени  $h$  в главном члене погрешности. Таким образом, понятие порядка точности формулы или метода основывается на сравнении скорости убывания погрешности со скоростью убывания степенных функций  $1, h, h^2, \dots, h^p, \dots$ , т. е. существенно использует «степенную шкалу». Иногда (не слишком часто) эта шкала оказывается не вполне адекватной реальному поведению погрешности.

**Пример 2.8.1** Пусть на вещественной оси задана равномерная сетка шага  $h$ , включающая узлы  $0, \pm h, \pm 2h$  и т. д. Для функции  $y = g(x)$  рассмотрим интерполяцию значения  $g(0)$  полусуммой

$$\frac{1}{2}(g(-h) + g(h)), \quad (2.98)$$

т. е. простейшим интерполяционным полиномом первой степени по узлам  $-h$  и  $h$ . Каков будет порядок погрешности такой интерполяции в зависимости от  $h$  для различных функций  $g(x)$ ?<sup>19</sup>

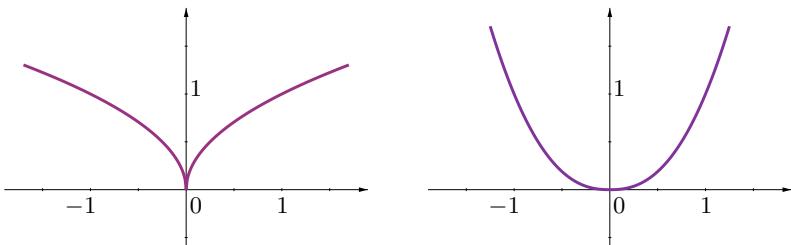


Рис. 2.20. Графики функции  $y = |x|^\alpha$  при  $0 < \alpha < 1$  и  $\alpha > 1$

Для функции  $g(x) = |x|^\alpha$ ,  $\alpha > 0$  (рис. 2.20), погрешность интерполяции будет, очевидно, равна  $h^\alpha$ , так что её порядок равен  $\alpha$ . Он может быть нецелым числом (в частности, дробным) и даже сколь угодно малым при приближении  $\alpha$  к нулю.

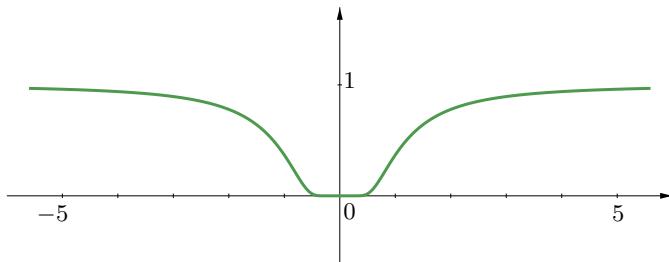


Рис. 2.21. График функции  $y = \exp(-1/x^2)$

Возьмём в качестве  $g(x)$  функцию

$$g(x) = \begin{cases} \exp(-1/x^2) & \text{при } x \neq 0, \\ 0 & \text{при } x = 0, \end{cases}$$

график которой изображён на рис. 2.21. Она известна в математическом анализе как пример бесконечно гладкой, но не аналитической

<sup>19</sup>Идея этого примера заимствована из пособия [57], задача 4.2.

(т. е. не разлагающейся в степенной ряд) функции. Погрешность интерполяции значения этой функции в нуле с помощью формулы (2.98) равна  $\exp(-1/h^2)$ , и при  $h \rightarrow 0$  она убывает быстрее любой степени  $h$ . Получается, что порядок точности нашей интерполяции оказывается бесконечно большим. Но такой же бесконечно большой порядок точности интерполирования будет демонстрировать в этих же условиях функция  $y = x^2 g(x)$ , хотя для неё погрешность  $h^2 \exp(-1/h^2)$  убывает существенно быстрее. ■

Из выкладок, проведённых для определения погрешности формулы «центральной разности», хорошо видна особенность метода разложений по формуле Тейлора: его *локальный* характер, вытекающий из свойств самой формулы Тейлора. Наши построения оказываются «привязанными» к определённому узлу (или узлам) сетки, относительно которого и следует строить все разложения, чтобы обеспечить взаимные уничтожения их ненужных членов. Как следствие, в этом специальном узле (узлах) мы можем быстро оценить погрешность. Но за пределами этого узла (узлов), в частности между узлами сетки, всё гораздо сложнее и не так красиво, поскольку взаимные уничтожения членов могут уже не происходить.

Какой порядок точности имеют другие формулы численного дифференцирования?

Методом разложений по формуле Тейлора для дважды гладкой функции  $f$  нетрудно получить оценки

$$|f_{x,i} - f'(x_i)| \leq \frac{M_2}{2} h, \quad |f_{\bar{x},i} - f'(x_i)| \leq \frac{M_2}{2} h, \quad (2.99)$$

где  $M_2 = \max_\xi |f''(\xi)|$  по  $\xi$  из соответствующего интервала между узлами. Таким образом, разность вперёд (2.86) и разность назад (2.87) имеют всего лишь первый порядок точности. Отметим, что для дважды непрерывно дифференцируемых функций оценки (2.99) уже не могут быть улучшены и достигаются, к примеру, на функции  $f(x) = x^2$ .

Конспективно изложим другие результаты о точности формул чис-

ленного дифференцирования:

$$f'(x_i) = \frac{1}{2h}(-3f_i + 4f_{i+1} - f_{i+2}) + O(h^2),$$

$$f'(x_i) = \frac{1}{2h}(f_{i-2} - 4f_{i-1} + 3f_i) + O(h^2),$$

$$f'(x_i) = \frac{1}{6h}(-2f_{i-1} - 3f_i + 6f_{i+1} - f_{i+2}) + O(h^3),$$

$$f'(x_i) = \frac{1}{6h}(f_{i-2} - 6f_{i-1} + 3f_i + 2f_{i+1}) + O(h^3).$$

Оценим теперь погрешность формулы (2.90) для второй производной

$$f''(x_i) \approx f_{x\bar{x},i} = \frac{f_{i-1} - 2f_i + f_{i+1}}{h^2}.$$

Обозначая для краткости  $f'_i = f'(x_i)$  и  $f''_i = f''(x_i)$ , получим

$$\begin{aligned} f_{x\bar{x},i} &= \frac{1}{h^2} \left( \left( f_i - hf'_i + \frac{h^2}{2} f''_i - \frac{h^3}{6} f'''_i + \frac{h^4}{24} f^{(4)}(\xi_-) \right) - 2f_i + \right. \\ &\quad \left. + \left( f_i + hf'_i + \frac{h^2}{2} f''_i + \frac{h^3}{6} f'''_i + \frac{h^4}{24} f^{(4)}(\xi_+) \right) \right) = \end{aligned}$$

$$= f''_i + \frac{h^2}{24}(f^{(4)}(\xi_-) + f^{(4)}(\xi_+)),$$

где  $\xi_-$ ,  $\xi_+$  — некоторые точки из открытого интервала  $]x_{i-1}, x_{i+1}[$ . Поэтому если  $f \in C^4[x_{i-1}, x_{i+1}]$ , то справедлива оценка

$$|f''(x_i) - f_{x\bar{x},i}| \leq \frac{M_4}{12} h^2,$$

где  $M_4 = \max_{\xi} |f^{(4)}(\xi)|$ . Таким образом, порядок точности этой формулы равен 2 на функциях с непрерывной четвёртой производной.

Приведём ещё без вывода результат о погрешности формулы для вычисления второй производной вблизи края сетки (таблицы):

$$f''(x_i) = \frac{1}{h^2}(2f_i - 5f_{i+1} + 4f_{i+2} - f_{i+3}) + O(h^2),$$

$$f''(x_i) = \frac{1}{h^2}(f_{i-3} - 4f_{i-2} + 5f_{i-1} - 2f_i) + O(h^2).$$

Порядок этих формул всего лишь второй, откуда видна роль симметричности шаблона в трёхточечной формуле (2.90) с тем же порядком точности.

Что произойдёт, если дифференцируемая функция не будет иметь достаточную гладкость? Тогда мы не сможем выписывать необходимое количество членов разложения по формуле Тейлора и потому полученный порядок точности формул с помощью метода локальных разложений установить будет нельзя. Тот факт, что в этих условиях реальный порядок точности может быть в самом деле меньшим, чем для функций с высокой гладкостью, показывает следующий пример.

**Пример 2.8.2** Рассмотрим функцию  $g(x) = x|x|$ , которую эквивалентным образом можно задать в виде

$$g(x) = \begin{cases} x^2, & \text{если } x \geq 0, \\ -x^2, & \text{если } x \leq 0. \end{cases}$$

Её график изображён на рис. 2.22.

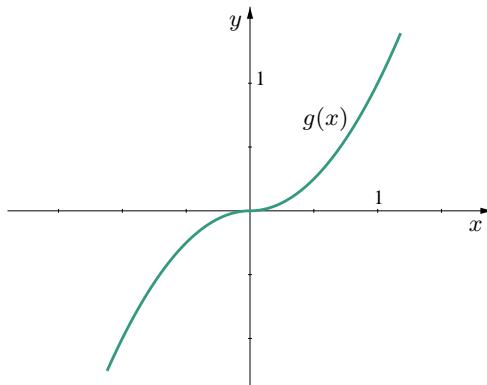


Рис. 2.22. График функции  $y = x|x|$ : увидеть разрыв её второй производной в нуле почти невозможно

Функция  $g(x)$  дифференцируема всюду на числовой оси. При  $x \neq 0$  она имеет производную, равную

$$g'(x) = (x|x|)' = x'|x| + x|x'| = |x| + x \operatorname{sgn} x = 2|x|,$$

а в нуле

$$g'(0) = \lim_{x \rightarrow 0} \frac{x|x|}{x} = 0.$$

В целом производная  $g'(x) = 2|x|$  всюду непрерывна. Но она недифференцируема в нуле, так что вторая производная  $g''(0)$  уже не существует. Как следствие,  $g(x) \in C^1$ , но  $g(x) \notin C^2$  на любом интервале, содержащем нуль. И заметить этот факт визуально, по графику функции на рис. 2.22, почти невозможно.

Воспользуемся для численного нахождения производной  $g'(0)$  формулой центральной разности (2.89) на шаблоне с шагом  $h$ , симметричном относительно нуля:

$$g'(0) \approx \frac{g(h) - g(-h)}{2h} = \frac{h|h| - (-h)|-h|}{2h} = \frac{h^2 + h^2}{2h} = h.$$

Таким образом, при  $h \rightarrow 0$  приближённое числовое значение производной стремится к  $g'(0) = 0$  с первым порядком по  $h$ , а не вторым, как мы установили это ранее для дважды гладких функций. ■

## 2.8г Метод неопределённых коэффициентов

*Метод неопределённых коэффициентов* — это другой подход к получению формул численного дифференцирования, особенно удобный в многомерном случае, когда построение интерполяционного полинома становится непростым.

Пусть задан шаблон из  $p + 1$  точек  $x_0, x_1, \dots, x_p$ . Станем искать приближённое выражение для  $k$ -й производной от функции в виде линейной формы от значений этой функции, т. е. как

$$f^{(k)}(x) \approx \sum_{i=0}^p c_i f(x_i). \quad (2.100)$$

Этот вид мотивируется тем обстоятельством, что дифференцирование любого порядка является операцией, линейной по значениям функции. Линейными формами от значений функции были, в частности, все полученные ранее формулы численного дифференцирования, начиная с (2.86) и кончая (2.97).

Коэффициенты  $c_i$  линейной формы постараемся подобрать так, чтобы эта формула являлась точной формулой для какого-то «достаточно представительного» набора функций. Например, в качестве таких

«пробных функций» можно взять все алгебраические полиномы степени не выше заданной либо тригонометрические полиномы (2.5) какой-то фиксированной степени и т. п. Рассмотрим ниже подробно случай алгебраических полиномов.

Возьмём функцию  $f(x)$  равной последовательным степеням переменной  $x$ , т. е.  $1, x, x^2, \dots, x^q$  для некоторого фиксированного  $q$ . Если формула (2.100) обращается в точное равенство на этих «пробных функциях», то с учётом её линейности можно утверждать, что она будет точной для любого алгебраического полинома степени не выше  $q$ .

Каждое условие, выписанное для какой-то определённой степени  $x^j$ ,  $j = 0, 1, \dots, q$ , является линейным соотношением на неизвестные коэффициенты  $c_i$ , и в целом мы приходим к системе линейных алгебраических уравнений относительно  $c_i$ ,  $i = 0, 1, \dots, p$ . Для её разрешимости естественно взять число неизвестных равным числу уравнений, т. е.  $q = p$ . Итак, получающаяся система линейных уравнений для определения коэффициентов формулы имеет вид

$$\left\{ \begin{array}{l} c_0 + c_1 + \dots + c_p = 0, \\ c_0x_0 + c_1x_1 + \dots + c_px_p = 0, \\ \vdots \quad \vdots \quad \ddots \quad \vdots \quad \vdots \\ c_0x_0^{k-1} + c_1x_1^{k-1} + \dots + c_px_p^{k-1} = 0, \\ c_0x_0^k + c_1x_1^k + \dots + c_px_p^k = k!, \\ c_0x_0^{k+1} + c_1x_1^{k+1} + \dots + c_px_p^{k+1} = (k+1)! x, \\ \vdots \quad \vdots \quad \ddots \quad \vdots \quad \vdots \quad \vdots \\ c_0x_0^p + c_1x_1^p + \dots + c_px_p^p = p(p-1)\cdots(p-k+1)x^{p-k}. \end{array} \right. \quad (2.101)$$

В правых частях этой системы стоят значения  $k$ -х производных от  $1, x, x^2, \dots, x^q$  в той точке  $x$ , где хотим найти приближение к  $f^{(k)}(x)$ . Матрицей системы является матрица Вандермонда вида (2.8), которая неосообщена для несовпадающих узлов  $x_0, x_1, \dots, x_p$ .

Система линейных уравнений (2.101) однозначно разрешима относительно  $c_0, c_1, \dots, c_p$  для любой правой части, но содержательным является лишь случай  $k \leq p$ . В противном случае, если  $k > p$ , правая часть системы (2.101) оказывается нулевой и, как следствие, система тоже имеет только бессодержательное нулевое решение. Этот факт имеет интуитивно ясное объяснение: нельзя построить формулу для

вычисления производной  $k$ -го порядка от функции, используя значения этой функции не более чем в  $k$  точках.

Матрицы Вандермонда в общем случае являются плохообусловленными (см. § 3.4б). Но на практике решение системы (2.101) — вручную или на компьютере — обычно не приводит к большим ошибкам, так как порядок системы (2.101), равный порядку производной, бывает, как правило, небольшим.<sup>20</sup>

**Пример 2.8.3** Исследуем функцию  $f : [0, 1] \rightarrow \mathbb{R}$ , для которой известны значения в дискретном наборе точек области определения, т. е. на сетке из  $[0, 1]$ . Предположим, что область вблизи правого конца нам особо интересна, так что сетка сгущается к ней, а соответствующие узлы равны 0.9, 0.95, 0.98, 1. Необходимо численно найти вторую производную функции на правом конце интервала, т. е. в точке 1, и для этого, желая достичь наибольшей возможной точности, можно привлечь значения функции в четырёх выписанных узлах.

Обозначим  $x_0 = 0.9$ ,  $x_1 = 0.95$ ,  $x_2 = 0.98$ ,  $x_3 = 1$ . Для определения коэффициентов формулы (2.100) составим систему вида (2.101):

$$\begin{cases} c_0 + c_1 + c_2 + c_3 = 0, \\ c_0x_0 + c_1x_1 + c_2x_2 + c_3x_3 = 0, \\ c_0x_0^2 + c_1x_1^2 + c_2x_2^2 + c_3x_3^2 = 2, \\ c_0x_0^3 + c_1x_1^3 + c_2x_2^3 + c_3x_3^3 = 6, \end{cases}$$

где в качестве значения  $x$  взята точка, в которой необходимо оценить производную, т. е. 1. Конкретно,

$$\begin{cases} c_0 + c_1 + c_2 + c_3 = 0, \\ 0.9c_0 + 0.95c_1 + 0.98c_2 + 1c_3 = 0, \\ 0.9^2c_0 + 0.95^2c_1 + 0.98^2c_2 + 1^2c_3 = 2, \\ 0.9^3c_0 + 0.95^3c_1 + 0.98^3c_2 + 1^3c_3 = 6. \end{cases}$$

Решение этой системы уравнений нетрудно найти как вручную, так и с помощью какой-нибудь системы компьютерной математики, и оно равно  $c_0 = -350$ ,  $c_1 = 3200$ ,  $c_2 = -6250$ ,  $c_3 = 3400$ . Поэтому искомая

---

<sup>20</sup>На стр. 139 мы уже обсуждали вопрос о том, каков наивысший порядок производных, всё ещё имеющих какой-либо смысл в практических задачах.

формула численного дифференцирования получается в виде

$$f''(1) \approx -350 f(0.9) + 3200 f(0.95) - 6250 f(0.98) + 3400 f(1).$$

В качестве иллюстрации её работы найдём численно значение второй производной от  $\sin x$  при  $x = 1$ . Согласно формуле оно оценивается как

$$-350 \sin 0.9 + 3200 \sin 0.95 - 6250 \sin 0.98 + 3400 \sin 1 = -0.84202,$$

тогда как точное значение для  $(\sin x)'' = -\sin x$  при  $x = 1$  равно  $-0.84147$ . Относительная погрешность полученного результата составляет 0.065 %. ■

Интересен вопрос о взаимоотношении метода неопределённых коэффициентов и рассмотренного ранее в § 2.8а интерполяционного подхода к численному дифференцированию. Ш.Е. Микеладзе в книге [74] утверждает, в частности, что любая формула численного дифференцирования, полученная методом неопределённых коэффициентов, может быть выведена с помощью интерполяционного подхода, отказывая методу неопределённых коэффициентов в оригинальности. Но нельзя отрицать, что метод неопределённых коэффициентов конструктивно проще и «технологичнее» в применении, и уже только это обстоятельство оправдывает его существование.

## 2.8д Полная вычислительная погрешность численного дифференцирования

Рассмотрим поведение полной погрешности численного дифференцирования при расчётах на реальных вычислительных устройствах. Под *полной погрешностью* мы понимаем суммарную ошибку численного нахождения производной, вызванную как приближённым характером самого метода, так и неточностями вычислений на цифровых ЭВМ из-за неизбежных ошибок округления и т. п.

Предположим, к примеру, что первая производная функции вычисляется по формуле «разность вперёд»

$$f'(x_i) \approx f_{x,i} = \frac{f_{i+1} - f_i}{h}.$$

Как мы уже знаем, её погрешность

$$|f_{x,i} - f'(x_i)| \leq \frac{M_2 h}{2},$$

где  $M_2 = \max_{\xi \in [a,b]} |f''(\xi)|$ . Если значения функции вычисляются с ошибками, то вместо точных  $f_i$  и  $f_{i+1}$  мы получаем их приближённые значения  $\tilde{f}_i$  и  $\tilde{f}_{i+1}$ , такие что

$$|f_i - \tilde{f}_i| \leq \delta \quad \text{и} \quad |f_{i+1} - \tilde{f}_{i+1}| \leq \delta,$$

где через  $\delta$  обозначена предельная абсолютная погрешность вычисления значений функции. Тогда в качестве приближённого значения производной мы должны взять

$$f'(x_i) \approx \frac{\tilde{f}_{i+1} - \tilde{f}_i}{h},$$

а предельную полную вычислительную погрешность  $E(h, \delta)$  нахождения первой производной функции можно оценить следующим образом:

$$\begin{aligned} E(h, \delta) &= \left| \frac{\tilde{f}_{i+1} - \tilde{f}_i}{h} - f'(x_i) \right| \leq \\ &\leq \left| \frac{\tilde{f}_{i+1} - \tilde{f}_i}{h} - \frac{f_{i+1} - f_i}{h} \right| + \left| \frac{f_{i+1} - f_i}{h} - f'(x_i) \right| \leq \\ &\leq \left| \frac{(\tilde{f}_{i+1} - f_{i+1}) + (f_i - \tilde{f}_i)}{h} \right| + \frac{M_2 h}{2} \leq \\ &\leq \frac{|f_{i+1} - \tilde{f}_{i+1}| + |f_i - \tilde{f}_i|}{h} + \frac{M_2 h}{2} = \frac{2\delta}{h} + \frac{M_2 h}{2}. \end{aligned} \tag{2.102}$$

Отметим, во-первых, что эта оценка достижима при подходящем сочетании знаков фигурирующих в неравенствах величин, коль скоро достижимо используемое в преобразованиях неравенство треугольника  $|a+b| \leq |a| + |b|$  и достижима оценка погрешности (2.99) для формулы «разность вперёд». Во-вторых, оценка не стремится к нулю при уменьшении шага  $h$ , так как первое слагаемое неограниченно увеличивается при  $h \rightarrow 0$ . В целом функция  $E(h, \delta)$  при фиксированном  $\delta$  имеет минимум, определяемый условием

$$\frac{\partial E(h, \delta)}{\partial h} = \frac{\partial}{\partial h} \left( \frac{2\delta}{h} + \frac{M_2 h}{2} \right) = -\frac{2\delta}{h^2} + \frac{M_2}{2} = 0.$$

То есть оптимальное значение шага численного дифференцирования, при котором достигается минимальная полная погрешность, равно

$$h^* = 2\sqrt{\delta/M_2}, \quad (2.103)$$

и брать меньший шаг численного дифференцирования смысла нет. Само наименьшее значение достигаемой при этом полной погрешности есть  $E(h^*, \delta) = 2\sqrt{\delta M_2}$ .

**Пример 2.8.4** Пусть в арифметике двойной точности с плавающей точкой, реализованной согласно стандарту IEEE 754/854 (см. § 1.4), численно находится производная функции, вычисление выражения для которой требует выполнения десяти арифметических операций с числами порядка единицы. Пусть также модуль второй производной ограничен сверху величиной  $M_2 = 10$ . Погрешность отдельной арифметической операции можно считать приближённо равной половине расстояния между соседними машинно представимыми числами, т. е. примерно  $10^{-16}$  в районе единицы. Наконец, пусть абсолютная погрешность вычисления функции складывается из сумм абсолютных погрешностей каждой операции, так что  $\delta \approx 10 \cdot 10^{-16} = 10^{-15}$  при аргументах порядка единицы.

Тогда в соответствии с формулой (2.103) имеем  $h^* = 2\sqrt{\delta/M_2} = 2 \cdot 10^{-8}$ , т. е. брать шаг сетки меньше  $10^{-8}$  смысла не имеет. ■

Совершенно аналогичная ситуация имеет место и при использовании других формул численного дифференцирования. Производная  $k$ -го порядка на равномерной сетке шага  $h$  определяется в общем случае формулой вида<sup>21</sup>

$$f^{(k)}(x) = h^{-k} \sum_i c_i f(x_i) + R_k(f, x), \quad (2.104)$$

где  $c_i = O(1)$  при  $h \rightarrow 0$ . Если эта формула имеет порядок точности  $p$ , то её остаточный член оценивается как  $R_k(f, x) \approx c(x) h^p$ . Этот остаточный член определяет «идеальную» погрешность численного дифференцирования в отсутствие ошибок вычисления функции, и он неограниченно убывает при  $h \rightarrow 0$ .

Но если погрешность вычисления значений функции  $f(x_i)$  в узлах равна  $\delta$ , то в правой части (2.104) возникает ещё член, абсолютная

---

<sup>21</sup>Для примера можно взглянуть на те формулы, которые приведены в § 2.8а.

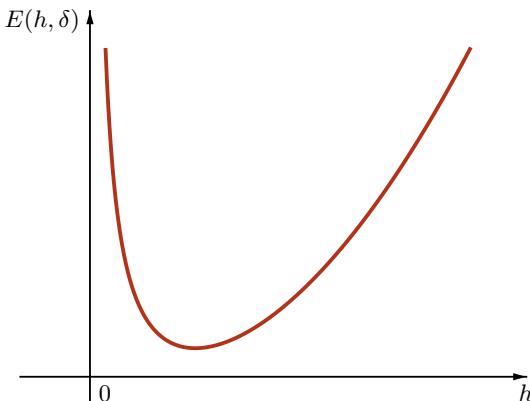


Рис. 2.23. Типичный график полной погрешности численного дифференцирования

величина которого совершенно аналогично (2.102) оценивается сверху как

$$\delta h^{-k} \sum_i |c_i|.$$

Она неограниченно возрастает при  $h \rightarrow 0$ . В целом график полной вычислительной погрешности численного дифференцирования выглядит в этом случае примерно так, как на рис. 2.23.

Практический вывод из сказанного состоит в том, что существует оптимальный шаг  $h$  численного дифференцирования, минимизирующий полную вычислительную погрешность, и брать слишком маленькое значение шага  $h$  в практических расчётах нецелесообразно.

Потенциально сколь угодно большое возрастание погрешности численного дифференцирования в действительности является отражением более глубокого факта *некорректности* задачи дифференцирования (см. § 1.7). Её решение не зависит непрерывно от входных данных, и это демонстрируют простые примеры. Если  $f(x)$  — исходная функция, производную которой нам требуется найти, то возмущённая функция  $f(x) + \frac{1}{n} \sin(nx)$  при  $n \rightarrow \infty$  будет равномерно сходиться к исходной, тогда как её производная

$$f'(x) + \cos(nx)$$

не сходится к производной  $f'(x)$  (рис. 2.24). При возмущении исход-

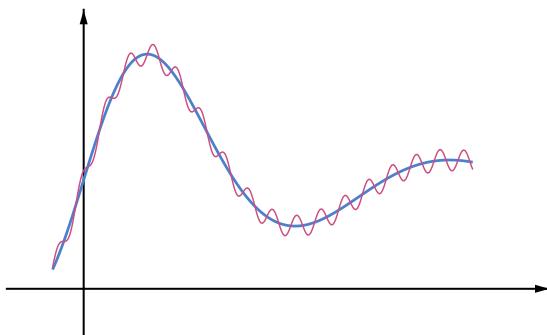


Рис. 2.24. Возмущение функции добавкой  $\frac{1}{n} \sin(nx)$

ной функции слагаемым  $\frac{1}{n} \sin(n^2 x)$  производная вообще может сколь угодно сильно отличаться от производной исходной функции.

## 2.9 Алгоритмическое дифференцирование

Пусть  $u = u(x)$  и  $v = v(x)$  — некоторые выражения от переменной  $x$ , из которых далее с помощью сложения, вычитания, умножения или деления конструируется более сложное выражение. Напомним правила дифференцирования выражений, образованных с помощью элементарных арифметических операций [12, 40]:

$$\begin{aligned}(u + v)' &= u' + v', \\ (u - v)' &= u' - v', \\ (uv)' &= u'v + uv', \\ \left(\frac{u}{v}\right)' &= \frac{u'v - uv'}{v^2}.\end{aligned}\tag{2.105}$$

Из них следует, что численное значение производной для сложного выражения мы можем найти, зная лишь значения образующих его подвыражений и их производных.

Сделанное наблюдение подсказывает идею ввести на множестве пар вида  $(u, u')$ , которые составлены из значений выражений и их производных, арифметические операции по правилам, следующим из фор-

мул (2.105):

$$\begin{aligned} (u, u') + (v, v') &= (u + v, u' + v'), \\ (u, u') - (v, v') &= (u - v, u' - v'), \\ (u, u') \cdot (v, v') &= (uv, u'v + uv'), \\ \frac{(u, u')}{(v, v')} &= \left( \frac{u}{v}, \frac{u'v - uv'}{v^2} \right). \end{aligned} \quad (2.106)$$

Первые члены пар преобразуются просто в соответствии с применяемой арифметической операцией, а операции над вторыми членами пар — это в точности копии правил (2.105). Проведём теперь для заданного выражения вычисления по выписанным формулам (2.105), заменив исходную переменную  $x$  на пару  $(x, 1)$ , а константы  $c$  — на пары вида  $(c, 0)$ , которые соответствуют значению переменной и её производной и значению константы и её производной. В результате получим пару, состоящую из численных значений выражения и производной от него для заданного значения переменной  $x$ .

Это рассуждение очевидным образом обобщается на случай, когда функция зависит от нескольких переменных.

Помимо арифметических операций интересующее нас выражение может содержать вхождения элементарных функций. Для них в соответствии с формулами дифференциального исчисления можем определить действия над парами следующим образом

$$\begin{aligned} \exp((u, u')) &= (\exp u, u' \exp u), \\ \sin((u, u')) &= (\sin u, u' \cos u), \\ ((u, u'))^2 &= (u^2, 2uu'), \\ ((u, u'))^3 &= (u^3, 3u^2u') \text{ и т. д.} \\ ((u, u'))^\alpha &= (u^\alpha, \alpha u^{\alpha-1}u'). \end{aligned}$$

Арифметику пар вида  $(u, u')$  с операциями (2.106) называют *дифференциальной арифметикой*, а основанный на её использовании способ вычисления значений производных носит название *алгоритмического дифференцирования*. Нередко используют также термин «автоматическое дифференцирование». С точки зрения абстрактной алгебры дифференциальная арифметика представляет собой множество *дуальных*

чисел или гиперкомплексных чисел параболического типа [96], которое, в свою очередь, является частным случаем так называемых алгебр Клиффорда.

**Пример 2.9.1** Рассмотрим процесс нахождения значения производной от функции, задаваемой выражением

$$y = \sqrt{2 - \sin x}$$

для значения аргумента  $x = 1$ . Наглядно процесс вычисления этого выражения (так называемое дерево Канторовича) изображается на рис. 2.25. Стрелка снизу обозначает корень дерева, т. е. ту его выделенную вершину, в которой получается значение выражения.

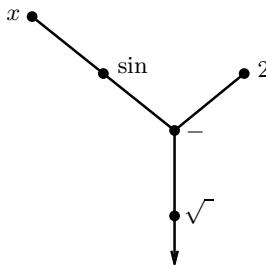


Рис. 2.25. Дерево Канторовича выражения  $\sqrt{2 - \sin x}$

Вычисления начнём с пары  $(1, 1)$  в левой ветви, где задаётся значение переменной, и с пары  $(2, 0)$  в правой ветви, где задаётся константа. Из первой пары после прохождения узла с синусом получится пара  $(\sin 1, \cos 1)$ , а далее в узле, где выполняется вычитание, получаем  $(2 - \sin 1, -\cos 1)$ . Наконец, в последнем узле, где берётся квадратный корень, результатом является пара

$$\left(\sqrt{2 - \sin 1}, -\frac{1}{2}(2 - \sin 1)^{-1/2} \cos 1\right).$$

Поэтому искомое значение производной равно  $-\frac{1}{2}(2 - \sin 1)^{-1/2} \cos 1$ , и для его вычисления не понадобилось находить явное аналитическое выражение для производной. ■

Отметим, что помимо дерева Канторовича существуют также другие способы представления процесса вычисления выражений. Иногда

они даже более удобны при компьютерной реализации алгоритмического дифференцирования, так как имеют «линейный характер», т. е. упорядочивают необходимые действия последовательно одно за другим «вдоль линии». Это *список инструкций* («code list» по-английски), в котором аналитические и логические формулы заменяются на список присваиваний некоторым величинам результатов выполнения операций с другими (предшествующими) величинами.<sup>22</sup>

Строго говоря, выше рассмотрен один из возможных способов организации алгоритмического дифференцирования, который называют *прямым режимом*. Существует также *обратный режим* алгоритмического дифференцирования, который основан на той же идее, но имеет определённые преимущества перед прямым режимом (см. подробности, к примеру, в [99]).

Описанную выше идею можно применить к вычислению вторых производных. Но теперь вместо дифференциальной арифметики пар чисел  $(u, u')$  нам необходимо будет оперировать с числовыми тройками вида  $(u, u', u'')$ , поскольку в формулах для вторых производных функции фигурируют значения самой функции и её первых и вторых производных. Эта конструкция распространяется также на производные более высоких порядков.

Идея алгоритмического дифференцирования может быть распространена на вычисление разделённых разностей (наклонов) функций (см., к примеру, [106]), на вычисление констант Липшица, а также на вычисление интервальных расширений производных и наклонов [93]. Эти интервальные расширения используются, к примеру, в интервальных методах оптимизации и при организации интервального метода Ньютона для решения уравнений и систем уравнений (см. § 4.7б и 4.7в).

Зафиксируем точки  $x$  и  $\tilde{x}$  в области определения функций  $u(x)$  и  $v(x)$ . Обозначим для краткости посредством  $u$ ,  $v$  и  $\tilde{u}$ ,  $\tilde{v}$  значения функций в этих точках. По определению разделённых разностей

$$u^\angle = \frac{u - \tilde{u}}{x - \tilde{x}}, \quad v^\angle = \frac{v - \tilde{v}}{x - \tilde{x}}.$$

---

<sup>22</sup>Он называется «расчленённой схемой» в оригинальной статье [60], где введена и эта конструкция, и «дерево Канторовича».

Тогда

$$\begin{aligned}
 (u+v)^\angle &= \frac{(u+v) - (\tilde{u}+\tilde{v})}{x-\tilde{x}} = \frac{(u-\tilde{u}) + (v-\tilde{v})}{x-\tilde{x}} = u^\angle + v^\angle, \\
 (u-v)^\angle &= \frac{(u-v) - (\tilde{u}-\tilde{v})}{x-\tilde{x}} = \frac{(u-\tilde{u}) - (v-\tilde{v})}{x-\tilde{x}} = u^\angle - v^\angle, \\
 (uv)^\angle &= \frac{uv - \tilde{u}\tilde{v}}{x-\tilde{x}} = \frac{uv - \tilde{u}v + \tilde{u}v - \tilde{u}\tilde{v}}{x-\tilde{x}} = \\
 &= \frac{(u-\tilde{u})v + \tilde{u}(v-\tilde{v})}{x-\tilde{x}} = u^\angle v + \tilde{u}v^\angle, \\
 \left(\frac{u}{v}\right)^\angle &= \frac{u/v - \tilde{u}/\tilde{v}}{x-\tilde{x}} = \frac{u\tilde{v} - \tilde{u}v}{(x-\tilde{x})v\tilde{v}} = \frac{u\tilde{v} - uv + uv - \tilde{u}v}{(x-\tilde{x})v\tilde{v}} = \\
 &= \frac{-u(v-\tilde{v}) + (u-\tilde{u})v}{(x-\tilde{x})v\tilde{v}} = \frac{u^\angle v - uv^\angle}{v\tilde{v}},
 \end{aligned}$$

Для умножения и деления выписанные выкладки можно провести другим способом, получив другие результаты, в целом аналогичные по своим свойствам, выписанным выше.

Арифметика разделённых разностей и наклонов, в отличие от дифференциальной, — это арифметика упорядоченных числовых троек вида  $(\tilde{u}, u, u^\angle)$ , а не арифметика пар. Эти тройки образованы двумя значениями функции, между аргументами которых берётся наклон, а также самим значением наклона.

Расчётные формулы арифметики наклонов, вытекающие из результатов предшествующих выкладок, имеют следующий вид:

$$\begin{aligned}
 (\tilde{u}, u, u^\angle) + (\tilde{v}, v, v^\angle) &= (\tilde{u} + \tilde{v}, u + v, u^\angle + v^\angle), \\
 (\tilde{u}, u, u^\angle) - (\tilde{v}, v, v^\angle) &= (\tilde{u} - \tilde{v}, u - v, u^\angle - v^\angle), \\
 (\tilde{u}, u, u^\angle) \cdot (\tilde{v}, v, v^\angle) &= (\tilde{u}\tilde{v}, uv, u^\angle v + \tilde{u}v^\angle), \\
 (\tilde{u}, u, u^\angle) / (\tilde{v}, v, v^\angle) &= \left( \frac{\tilde{u}}{\tilde{v}}, \frac{u}{v}, \frac{u^\angle v - uv^\angle}{v\tilde{v}} \right).
 \end{aligned}$$

Как отмечалось, расчётные формулы для умножения и деления могут иметь альтернативные варианты.

В настоящее время созданы готовые программные системы, позволяющие выполнять алгоритмическое дифференцирование почти так

же легко, как и вычисление стандартных математических функций. Таковы, например, библиотеки алгоритмического дифференцирования для языков программирования Fortran, C++, Питон, Julia, Scala и др.

## 2.10 Приближение функций

### 2.10а Обсуждение постановки задачи

В этом разделе мы более подробно займёмся задачей приближения функций, постановка которой была предварительно рассмотрена в начале главы. Цель решения задачи приближения (которую также называют «задачей аппроксимации») — та же, что и в случае интерполяции: дать для интересующего нас математического объекта подходящее приближение, в некотором определённом смысле, с помощью другого, более удобного (более простого и т. п.) объекта. Но делается это несколько иначе, не так, как в задаче интерполяции.

К задаче приближения функций естественно приходят в ситуациях, где методы интерполяции по различным причинам не удовлетворяют практику. Эти причины могут носить технический характер. К примеру, гладкость интерполяционного сплайна может оказаться недостаточной либо его построение — слишком сложным. Степень полинома, который мы должны построить по данным, может быть известной из каких-либо содержательных соображений (физических, химических, биологических и т. п.), но узлов для этого имеется слишком много. Кроме того, высокая степень полинома — это трудности при его построении и работе с ним. И так далее.

Но причины отказа от интерполяции могут иметь также принципиальный характер. Интерполяция теряет смысл, если значения функции в узлах известны неточно либо сами эти узлы нельзя указать явно и однозначно. Иногда даже наличие самих этих узлов, т. е. выделенных точек области определения, ничем не мотивируется. В этих условиях целесообразна коррекция постановки задачи.

Можно ослабить требование того, чтобы восстанавливаемая функция  $g$  была точно равна заданным значениям  $f_i$  в узлах  $x_0, x_1, \dots, x_n$ , допустив, к примеру, для  $g$  принадлежности её значений некоторым интервалам, т. е.

$$g(x_i) \in [\underline{f}_i, \bar{f}_i], \quad \underline{f}_i \leq \bar{f}_i, \quad i = 0, 1, \dots, n.$$

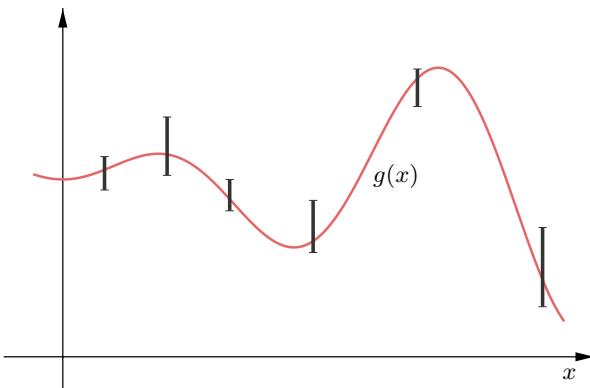


Рис. 2.26. Интерполяция функции, заданной с погрешностьюю (интервальная интерполяция). Точное задание узлов

Наглядно графически это означает построение функции  $g(x)$  из заданного класса  $\mathcal{G}$ , которая в каждом узле сетки  $x_i$ ,  $i = 0, 1, \dots, n$ , проходит через некоторый «коридор»  $[\underline{f}_i, \bar{f}_i]$ , как на рис. 2.26. Дальнейшее усложнение задачи происходит при неточным задании узлов, изображённом на рис. 2.27. Подобные постановки обычно возникают при обработке данных измерений и наблюдений [46].

Более общая постановка задачи предусматривает наличие некоторой метрики (расстояния), с помощью которой можно измерять отклонение вектора значений  $(g(x_0), g(x_1), \dots, g(x_n))^\top$  функции  $g(x)$  в узлах сетки от вектора заданных значений  $(f_0, f_1, \dots, f_n)^\top$ . Фактически в рассматриваемой ситуации задаётся какое-то расстояние на пространстве  $\mathbb{R}^{n+1}$  всех  $(n+1)$ -мерных вещественных векторов и соответствующая постановка задачи приближения (аппроксимации) формулируется следующим образом:

Для заданного набора узлов  $x_0, x_1, \dots, x_n$  на интервале  $[a, b]$ , соответствующих им значений  $f_0, f_1, \dots, f_n$  и  $\epsilon > 0$  найти такую функцию  $g(x)$  из класса  $\mathcal{G}$ , что расстояние между векторами  $\mathbf{f} = (f_0, f_1, \dots, f_n)^\top$  и  $\mathbf{g} = (g(x_0), g(x_1), \dots, g(x_n))^\top$  не больше  $\epsilon$ .

При этом  $g(x)$  называют *приближающей* (аппроксимирующей) функцией. Важнейшей модификацией поставленной задачи служит *задача наилучшего приближения*, в которой величина  $\epsilon$  не фиксируется

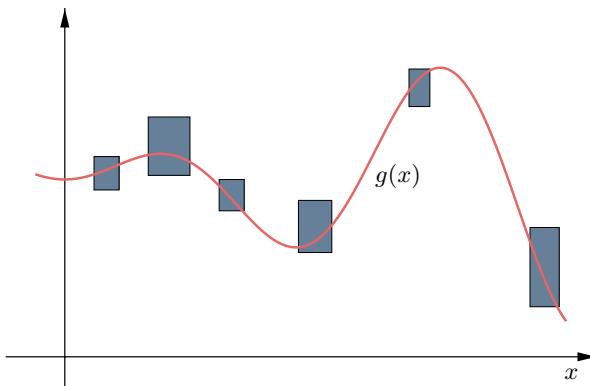


Рис. 2.27. Интерполяция функции, заданной с погрешностью (интервальная интерполяция). Неточное задание узлов

и ищут приближающую (аппроксимирующую) функцию  $g(x)$ , которая доставляет минимум расстоянию между векторами  $f$  и  $g$ .

Согласно классификации, описанной в § 2.1, выписанные выше формулировки являются дискретными вариантами общей задачи о приближении функции, в которой набор узлов  $x_0, x_1, \dots, x_n$  уже не фигурирует, а отклонение одной функции от другой измеряется на всей области их определения:

Пусть даны классы функций  $\mathcal{F}$  и  $\mathcal{G}$ . Для функции  $f(x)$  из класса  $\mathcal{F}$  найти функцию  $g(x)$  из класса  $\mathcal{G}$ , которая в заданном смысле достаточно близка к функции  $f(x)$ .

Соответствующая общая формулировка задачи о наилучшем приближении требует нахождения  $g(x)$ , «наиболее близкой» к  $f(x)$ .

Отклонение функций друг от друга, т. е. мера их взаимной близости, в приведённых выше формулировках может быть самым разнообразным, определяясь той или иной конкретной практической постановкой. Типична ситуация, когда класс функций  $\mathcal{G}$  является просто подмножеством класса  $\mathcal{F}$ , т. е.  $\mathcal{F} \supset \mathcal{G}$ . Тогда отклонение одной функции от другой может задаваться метрикой (расстоянием) на  $\mathcal{F}$ , которую мы обозначим через  $\text{dist}$  (см. определение на стр. 70). Задача наилучшего

приближения функций получает следующую развёрнутую постановку:

Для заданных функции  $f(x)$  из класса функций  $\mathcal{F}$  и метрики  $\text{dist}$  найти функцию  $g(x)$  из класса  $\mathcal{G} \subset \mathcal{F}$ , на которой достигается нижняя грань расстояний от функции  $f(x)$  до функций из  $\mathcal{G}$ , т. е., для которой выполняется условие  $\text{dist}(f, g) = \inf_{h \in \mathcal{G}} \text{dist}(f, h)$ . (2.107)

Решение  $g$  этой задачи, если оно существует, называется *наилучшим приближением* для  $f$  в классе  $\mathcal{G}$ . Отметим, что в каждом конкретном случае существование наилучшего приближения требует отдельного исследования.

Задачу приближения функций, значения которых могут быть не вполне точными, часто называют (особенно в практических приложениях) *задачей сглаживания*. Получаемая при этом приближающая функция действительно «сглаживает» погрешности и выбросы в данных, вызванные случайными ошибками и т. п.

Классы функций  $\mathcal{F}$  и  $\mathcal{G}$  в наших формулировках могут быть весьма разнообразными, и нередко мы находимся в условиях, когда  $\mathcal{F} \supset \mathcal{G}$ . В реальных практических постановках задач приближающая функция  $g$  обычно зависит от одного или нескольких параметров, т. е.  $g(x) = g(x, \beta_1, \beta_2, \dots)$ , и параметры  $\beta_1, \beta_2, \dots$ , необходимо выбрать из предписанных им множеств так, чтобы получить требуемое приближение или же наилучшее приближение. Тогда класс функций  $\mathcal{G}$  — это просто множество функций вида  $g(x, \beta_1, \beta_2, \dots)$  при всевозможных значениях параметров  $\beta_1, \beta_2, \dots$ .

Часто приближающие функции  $g(x) = g(x, \beta_1, \beta_2, \dots)$  зависят от параметров  $\beta_1, \beta_2, \dots$  линейным образом, как, например, алгебраические или тригонометрические полиномы. Факт линейности означает, что можно наделить  $\mathcal{G}$  (а вслед за ним и  $\mathcal{F}$ ) структурой линейного векторного пространства. Тогда в качестве количественной меры приближения обычно вводят на  $\mathcal{G}$  некоторую норму. Получающиеся линейные векторные пространства с нормой называют, как известно, *нормированными*. В нормированном пространстве расстояние естественно определяется с помощью стандартной конструкции (2.4),

$$\text{dist}(f, g) := \|f - g\|,$$

т. е. как норма отличия векторов друг от друга. Соответственно, в задаче наилучшего приближения функции  $f$  ищется такая функция  $g \in \mathcal{G}$ , на которой достигается  $\inf_{h \in \mathcal{G}} \|f - h\|$ .

Нужно отметить, что переход в линейные нормированные пространства вызывает и смену языка, на котором формулируются наши задачи приближения и их решения. Он становится абстрактным и геометрически ориентированным: функция теперь является «точкой» в подходящем пространстве, отклонение одной функции от другой — это расстояние между точками-функциями и т. д. Эта абстрактная, но и весьма плодотворная точка зрения на функции и способ их изучения была выработана в исследованиях М. Фреше и Д. Гильберта в начале XX века.

Рассмотренные выше постановки задач дают начало большим и важным разделам математики, в совокупности образующим теорию приближения функций (называемую также теорией аппроксимации). Её большими ветвями являются теория среднеквадратичного приближения (которой мы кратко коснёмся в § 2.10г и § 2.11) и теория равномерного приближения, когда отклонение функций оценивается в равномерной (чебышёвской) норме  $\|f\|_\infty$ ; см. книги [2, 8, 29, 44, 63, 86] и др. Выбор различных норм (т. е. различных мер отклонения функций друг от друга) и различных классов функций вызывает большое разнообразие задач теории приближения.

Кроме рассмотренных выше постановок в современной теории приближения функций интенсивно изучаются и другие задачи. Они относятся к приближению не отдельного элемента (функции), а целого множества функций, к наиболее выгодному выбору класса приближающих функций и др. Наконец, отметим существование нелинейных методов приближения функций (аналогично нелинейной интерполяции из § 2.7), в которых множества  $\mathcal{F}$  и  $\mathcal{G}$  не несут на себе линейной структуры. Эти нелинейные методы получили чрезвычайно широкое распространение в последние годы, так как они лежат в основе теории и приложений искусственных нейронных сетей.

## 2.10б Существование наилучшего приближения

Некоторые важные свойства решений задачи наилучшего приближения можно вывести уже из её абстрактной формулировки для линейных векторных пространств. В частности, это касается существования решения, а также его единственности при некоторых дополнительных условиях на норму. Как отмечалось, вместо функций и классов функ-

ций в результатах этого раздела будут фигурировать точки в линейном нормированном пространстве, а также его подпространство, в котором мы выбираем приближение.

**Теорема 2.10.1** (теорема Бореля) *Пусть  $X$  — нормированное линейное пространство,  $U$  — его конечномерное линейное подпространство. Тогда для любого  $f \in X$  существует его наилучшее приближение в  $U$ .*

Э. Борель доказал этот важный результат для случая равномерного (чебышёвского) приближения функций полиномами в книге [97], изданной в самом начале XX века, когда понятие нормированного линейного пространства ещё не сформировалось. Но и само утверждение, и метод его доказательства почти дословно переносятся на общие нормированные пространства, что и было сделано математиками 30-х годов XX века.

Напомним, что множество  $W$  в нормированном линейном пространстве называется *ограниченным*, если существует такое число  $K$ , что для любой точки  $x \in W$  имеет место неравенство  $\|x\| \leq K$ . Этим понятием формализуется интуитивно ясный смысл «ограниченности размеров» множества.

Далее важную роль играет также понятие компактного множества. *Компактными* называют множества, из любого покрытия которых открытыми множествами можно выбрать конечное подпокрытие [12, 32]. Для метрических пространств это определение эквивалентно следующему: множество является компактным тогда и только тогда, когда из любой последовательности его точек можно выделить подпоследовательность, сходящуюся к некоторой точке этого множества [12]. Компактные множества — это множества, которые фактически можно исчерпать «конечными средствами», и по этой причине они обладают многими замечательными свойствами.

**Доказательство.** Пусть размерность подпространства  $U$  равна  $m$ . Зададим в  $U$  некоторый базис  $\{\phi_1, \phi_2, \dots, \phi_m\}$  и введём функцию  $D : \mathbb{R}^m \rightarrow \mathbb{R}_+$ , задаваемую как

$$D(a_1, a_2, \dots, a_m) = \left\| f - \sum_{j=1}^m a_j \phi_j \right\|,$$

где  $\|\cdot\|$  — норма в  $X$ . Значениями функции  $\mathcal{D}$  являются расстояния от элемента  $f$  до векторов из подпространства  $U$ , которые определяются строками  $a = (a_1, a_2, \dots, a_m)$  коэффициентов их разложения по базису  $\{\phi_1, \phi_2, \dots, \phi_m\}$ . Теорема будет доказана, если мы обоснуем тот факт, что функция  $\mathcal{D}$  достигает своего наименьшего значения на  $\mathbb{R}^m$ .

Прежде всего покажем, что функция  $\mathcal{D}$  непрерывно зависит от своих аргументов. Пусть  $a = (a_1, a_2, \dots, a_m) \rightarrow \tilde{a} = (\tilde{a}_1, \tilde{a}_2, \dots, \tilde{a}_m)$ . Так как  $\|x\| - \|y\| \leq \|x - y\|$  для любых векторов  $x$  и  $y$ , то имеем

$$\begin{aligned} |\mathcal{D}(a_1, a_2, \dots, a_m) - \mathcal{D}(\tilde{a}_1, \tilde{a}_2, \dots, \tilde{a}_m)| &= \\ &= \left| \left\| f - \sum_{j=1}^m a_j \phi_j \right\| - \left\| f - \sum_{j=1}^m \tilde{a}_j \phi_j \right\| \right| \leq \\ &\leq \left\| \left( f - \sum_{j=1}^m a_j \phi_j \right) - \left( f - \sum_{j=1}^m \tilde{a}_j \phi_j \right) \right\| = \\ &= \left\| \sum_{j=1}^m (a_j - \tilde{a}_j) \phi_j \right\| \leq \sum_{j=1}^m |a_j - \tilde{a}_j| \|\phi_j\| \leq \\ &\leq \max_{1 \leq j \leq m} |a_j - \tilde{a}_j| \cdot \sum_{j=1}^m \|\phi_j\|. \end{aligned}$$

Следовательно, при  $a_j \rightarrow \tilde{a}_j$  разность между  $\mathcal{D}(a_1, a_2, \dots, a_m)$  и  $\mathcal{D}(\tilde{a}_1, \tilde{a}_2, \dots, \tilde{a}_m)$  тоже будет стремиться к нулю. Нетрудно понять, что приведённая выкладка обосновывает также непрерывную зависимость величины

$$\left\| \sum_{j=1}^m a_j \phi_j \right\|$$

от вектор-строки  $a = (a_1, a_2, \dots, a_m)$ , и этим фактом мы тоже воспользуемся в доказательстве теоремы.

Следующим шагом доказательства продемонстрируем, что непрерывная функция  $\mathcal{D}$  может достигать нижней грани своих значений лишь на некотором компактном подмножестве всего пространства  $\mathbb{R}^m$ . Тогда утверждение теоремы Бореля будет следовать из известного результата математического анализа, обобщающего теорему Вейерштраса.

са об экстремальных значениях: на компакте непрерывная веществен-нозначная функция достигает своих экстремумов, как минимума, так и максимума [12, 32].

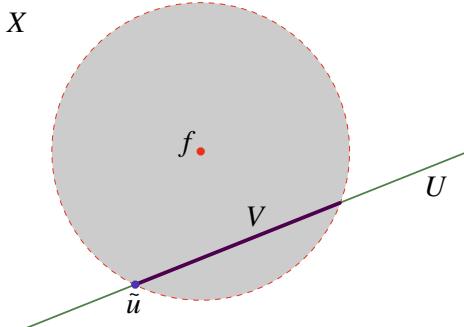


Рис. 2.28. Иллюстрация к теореме Бореля

Возьмём какую-либо точку  $\tilde{u}$  из  $U$  (рис. 2.28). Ясно, что нижняя грань расстояний от  $f$  до точек из  $U$  не больше, чем  $\|f - \tilde{u}\|$ , т. е.

$$\inf_{u \in U} \|f - u\| \leq \|f - \tilde{u}\|.$$

Поэтому

$$\inf_{a \in \mathbb{R}^m} \mathcal{D}(a) \leq \|f - \tilde{u}\|.$$

С учётом полученной оценки ясно, что эта нижняя грань достигается на множестве

$$V := \{a \in \mathbb{R}^m \mid \mathcal{D}(a) \leq \|f - \tilde{u}\|\},$$

которое замкнуто в  $\mathbb{R}^m$ , так как определяется нестрогим неравенством на непрерывную функцию  $\mathcal{D}$ . Кроме того, оно ещё и ограничено.

Действительно, из определения  $\mathcal{D}(a)$  следует

$$\mathcal{D}(a_1, a_2, \dots, a_m) = \left\| f - \sum_{j=1}^m a_j \phi_j \right\| \geq \left\| \sum_{j=1}^m a_j \phi_j \right\| - \|f\|.$$

Поэтому  $\mathcal{D}(a) \leq \|f - \tilde{u}\|$  влечёт

$$\left\| \sum_{j=1}^m a_j \phi_j \right\| \leq \|f - \tilde{u}\| + \|f\|. \quad (2.108)$$

Далее, зафиксируем в арифметическом пространстве  $\mathbb{R}^m$  векторов коэффициентов  $a = (a_1, a_2, \dots, a_m)$  какую-нибудь норму  $\|\cdot\|'$ . Тогда

$$\left\| \sum_{j=1}^m a_j \phi_j \right\| = \|a\|' \cdot \left\| \sum_{j=1}^m \frac{a_j}{\|a\|'} \phi_j \right\| \geq \|a\|' \cdot \min_{\|c\|'=1} \left\| \sum_{j=1}^m c_j \phi_j \right\|.$$

Величина

$$C := \min_{\|c\|'=1} \left\| \sum_{j=1}^m c_j \phi_j \right\|$$

достигается как экстремум непрерывной функции на компактной единичной сфере пространства  $\mathbb{R}^m$  относительно нормы  $\|\cdot\|'$ . Кроме того, она строго больше нуля, если  $\phi_1, \phi_2, \dots, \phi_m$  линейно независимы. Как следствие, из (2.108) заключаем, что

$$\|a\|' \leq \frac{1}{C} (\|f - \tilde{u}\| + \|f\|),$$

т. е. множество всех вектор-строк  $a = (a_1, a_2, \dots, a_m)$  из  $V$  в самом деле ограничено в  $\mathbb{R}^m$ .

Итак,  $V$  замкнуто и ограниченно, а потому компактно в конечномерном пространстве  $\mathbb{R}^m$  [12, 32]. По этой причине интересующее нас значение  $\inf_{a \in \mathbb{R}^m} \mathcal{D}(a) = \inf_{a \in V} \mathcal{D}(a)$  действительно достигается на каком-то определённом векторе  $a$ . Он даёт коэффициенты разложения наилучшего приближения для  $f$ . ■

## 2.10в Единственность наилучшего приближения

Наилучшее приближение, вообще говоря, может быть неединственным. Но при определённых условиях мы можем гарантировать его единственность, опираясь лишь на свойства пространства  $X$ .

**Определение 2.10.1** Нормированное линейное пространство  $X$  называют строго нормированным, если в неравенстве треугольника для нормы этого пространства равенство достигается только на положительно пропорциональных элементах, т. е. если для произвольных  $x, y \in X$  из равенства  $\|x + y\| = \|x\| + \|y\|$  следует, что  $y = \alpha x$  для некоторого скаляра  $\alpha \in \mathbb{R}_+$ .

**Теорема 2.10.2** Пусть  $X$  — строго нормированное линейное пространство, а  $U$  — его линейное подпространство. Для любого элемента из  $X$  существует не более одного наилучшего приближения в  $U$ .

**Доказательство.** Предположим, что для некоторого  $f \in X$  в подпространстве  $U$  существуют два наилучших приближения  $u'$  и  $u''$ , так что

$$\|f - u'\| = \|f - u''\| = \mu \geq 0,$$

где  $\mu$  — расстояние от  $f$  до  $u'$  и  $u''$ . Случай  $\mu = 0$  бессодержателен, так как он соответствует  $f = u' = u''$ . Следовательно, далее можем считать, что  $\mu > 0$ .

Взяв полусумму элементов  $u'$  и  $u''$ , т. е. точку  $\frac{1}{2}(u'+u'')$ , будем иметь

$$\begin{aligned} \left\| f - \frac{1}{2}(u' + u'') \right\| &= \left\| \frac{1}{2}(f - u') + \frac{1}{2}(f - u'') \right\| \leq \\ &\leq \frac{1}{2} \|f - u'\| + \frac{1}{2} \|f - u''\| = \mu. \end{aligned}$$

Строгое неравенства в выписанной цепочке отношений быть не может, так как оно означало бы существование элемента, приближающего  $f$  лучше, чем наилучшие приближения  $u'$  и  $u''$ . Поэтому необходимо должно выполняться

$$\|(f - u') + (f - u'')\| = \|f - u'\| + \|f - u''\|.$$

Но если пространство  $X$  — строго нормированное, то из полученного равенства следует

$$f - u' = \alpha(f - u'') \quad (2.109)$$

для некоторого вещественного  $\alpha > 0$ .

В случае  $\alpha = 1$  заключаем, что  $u' = u''$ , т. е. два наилучших приближения должны совпадать. В случае, когда  $\alpha \neq 1$ , из (2.109) вытекает

$$f = \frac{1}{1-\alpha} \cdot (u' - \alpha u''),$$

т. е.  $f$  представляется в виде линейной комбинации векторов подпространства  $U$ , а потому  $f \in U$ . Тогда должно быть  $\mu = 0$  и, как следствие, снова  $u' = u''$ .

Итак, для  $f$  в самом деле существует не более одного наилучшего приближения в  $U$ . ■

Какие нормы в линейных пространствах делают их строго нормированными пространствами? Рассмотрим примеры на эту тему.

**Пример 2.10.1** Возьмём линейное пространство функций, непрерывных на интервале вещественной оси, которое было введено в примере 2.10.1 (стр. 72). Не ограничивая общности, можно считать интервалом области определения  $[0, 1]$ , и пусть  $f(x) = x$  и  $g(x) = 1$ . Тогда

$$\|f\|_\infty = \max_{x \in [0, 1]} |f(x)| = 1 \quad \text{и} \quad \|g\|_\infty = \max_{x \in [0, 1]} |g(x)| = 1.$$

Далее,  $(f + g)(x) = x + 1$ , и потому  $\|f + g\|_\infty = 2$ . Как видим,

$$\|f + g\|_\infty = \|f\|_\infty + \|g\|_\infty,$$

но функции  $f$  и  $g$  не являются коллинеарными векторами рассматриваемого линейного пространства, т. е. нельзя сказать, что  $f(x) = \alpha g(x)$  для какого-то определённого  $\alpha$  и всех  $x \in [0, 1]$ .

Совершенно аналогичные примеры можно привести в арифметическом пространстве  $\mathbb{R}^n$  с нормами

$$\|x\|_1 = |x_1| + \dots + |x_n| \quad \text{и} \quad \|x\|_\infty = \max_{1 \leq i \leq n} |x_i|.$$

Возьмём, к примеру,  $x = (0, \dots, 0, 1)^\top$  и  $y = (1, \dots, 1, 1)^\top$ . Векторы  $x$  и  $y$ , очевидно, неколлинеарны, но

$$\|x + y\|_\infty = \|x\|_\infty + \|y\|_\infty.$$

Если же взять  $x = (1, 0, \dots, 0)^\top$  и  $y = (0, \dots, 0, 1)^\top$ , которые также неколлинеарны, то

$$\|x + y\|_1 = \|x\|_1 + \|y\|_1.$$

Иными словами, 1-норма и  $\infty$ -норма не делают  $\mathbb{R}^n$  строго нормированным пространством. ■

**Предложение 2.10.1** *Линейное векторное пространство со скалярным произведением  $\langle \cdot, \cdot \rangle$  является строго нормированным относительно стандартной нормы  $\|x\| = \sqrt{\langle x, x \rangle}$ .*

Вещественное линейное векторное пространство со скалярным произведением, имеющее конечную размерность, как известно, называют *евклидовым пространством*. Комплексное конечномерное линейное векторное пространство со скалярным произведением называется *унитарным пространством*. В общем случае, когда размерность может

быть бесконечной, пространства со скалярным произведением называются *гильбертовыми* или *предгильбертовыми* в зависимости от того, обладают ли они свойством полноты или нет (см. § 3.3б).

**Доказательство.** Для скалярного произведения, как известно, справедливо неравенство Коши–Буняковского [12, 16, 37]:

$$|\langle x, y \rangle| \leq \|x\| \|y\|,$$

причём равенство в нём достигается тогда и только тогда, когда векторы  $x$  и  $y$  коллинеарны или среди них есть нулевой. Правая часть неравенства всегда неотрицательна, так что из него следует более грубое соотношение

$$\langle x, y \rangle \leq \|x\| \|y\|,$$

причём равенство в нём достигается, как нетрудно понять, лишь при положительной пропорциональности векторов  $x$  и  $y$  или когда среди них есть нулевой.

Из неравенства Коши–Буняковского следует, что

$$\begin{aligned} \|x + y\|^2 &= \langle x + y, x + y \rangle = \\ &= \|x\|^2 + 2\langle x, y \rangle + \|y\|^2 \leq \\ &\leq \|x\|^2 + 2\|x\| \|y\| + \|y\|^2 = (\|x\| + \|y\|)^2. \end{aligned}$$

Так как все сравниваемые в выписанной цепочке выражения неотрицательны, можем заключить, что

$$\|x + y\| \leq \|x\| + \|y\|,$$

т. е. выполняется и неравенство треугольника. Равенство в нём при ненулевых  $x$  и  $y$  имеет место в том и лишь в том случае, когда оно выполнено в самом неравенстве Коши–Буняковского, т. е. при положительной пропорциональности векторов  $x$  и  $y$ . ■

В частности, строго нормировано пространство  $\mathbb{R}^n$  с 2-нормой,

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2},$$

порождаемой стандартным скалярным произведением в  $\mathbb{R}^n$  (см. § 3.3а). Эта норма называется также *евклидовой нормой* или даже просто *длинной вектора*.

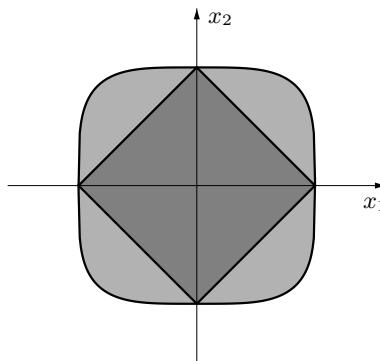


Рис. 2.29. Единичные шары 1-нормы (тёмного тона) и 4-нормы (светлого тона) в пространстве  $\mathbb{R}^2$

**Пример 2.10.2** В теории и на практике нередко используются так называемые *p*-нормы, обычно обозначаемые  $\|\cdot\|_p$ , которые обобщают введённые выше 1-норму, 2-норму и чебышёвскую норму ( $\infty$ -норму). В арифметическом пространстве  $\mathbb{R}^n$  (т. е., фактически, в пространстве функций дискретного аргумента) *p*-норму определяют следующим образом:

$$\|x\|_p = \left( \sum_{i=1}^n |x_i|^p \right)^{1/p}. \quad (2.110)$$

При  $p > 1$  это в самом деле норма  $n$ -векторов, так как для неё выполнены все аксиомы нормы (см. ниже § 3.3а). Неотрицательность и абсолютная однородность  $\|\cdot\|_p$  очевидны, а выполнение неравенства треугольника следует из *неравенства Минковского*

$$\left( \sum_{i=1}^n |x_i + y_i|^p \right)^{1/p} \leq \left( \sum_{i=1}^n |x_i|^p \right)^{1/p} + \left( \sum_{i=1}^n |y_i|^p \right)^{1/p}$$

(см. [12, 15, 16, 37, 40, 44]). Из свойств неравенства Минковского вытекает, кроме того, что равенство в нём при  $p > 1$  возможно лишь для коллинеарных  $x$  и  $y$ . По этой причине пространство  $\mathbb{R}^n$  с *p*-нормой является строго нормированным при  $p > 1$ . Рис. 2.29 изображает единичные шары *p*-нормы для некоторых *p* (см. также рис. 3.6).

В пространствах функций непрерывного аргумента *p*-нормы вво-

дятся обычно как

$$\|f\|_p = \left( \int |f(x)|^p dx \right)^{1/p},$$

где интеграл берётся по соответствующей области (интервалу вещественной оси, области плоскости или пространства и т. п.). Очевидно, что это обобщение определения (2.110). ■

Дальнейшие результаты о существовании наилучших приближений в различных подмножествах линейных векторных пространств можно найти, например, в книге [72].

## 2.10г Квадратичные и среднеквадратичные приближения

*Квадратичными и среднеквадратичными* называются приближения, в которых расстояние между функциями определяется через квадраты разностей их значений в заданных аргументах — либо как сумма, либо как интеграл от квадрата разности функций по заданной области. Такие приближения важны по целому ряду математических и практических причин. Это теоретико-вероятностные соображения, связанные с обработкой данных со случайными погрешностями, где среднеквадратичное отклонение означает дисперсию (меру рассеяния) случайной величины. Квадратичные выражения возникают также для различных видов энергии в физических системах и т. д. Отличие квадратичных приближений от среднеквадратичных выражается в том, что в последнем случае в определение расстояния (отклонения) дополнительно входит усреднение.

Если на классе приближаемых функций задана линейная структура, т. е. он является линейным пространством, то квадратичным (или среднеквадратичным) приближением называют также приближение в норме, порождённой скалярным произведением на этом пространстве. Поясним мотивы выбора такой терминологии.

В конечномерной ситуации скалярное произведение векторов  $f = (f_1, f_2, \dots, f_n)^\top$  и  $g = (g_1, g_2, \dots, g_n)^\top$  стандартно определяется как

$$\langle f, g \rangle = f_1 g_1 + f_2 g_2 + \cdots + f_n g_n = \sum_{i=1}^n f_i g_i.$$

Этим скалярным произведением задаётся 2-норма (евклидова норма)

$$\|f\| := \sqrt{\sum_{i=1}^n f_i^2}, \quad (2.111)$$

в которой фигурируют квадраты компонент вектора. Но во многих задачах скалярное произведение конечномерных векторов удобнее рассматривать в несколько модифицированном, хотя и совершенно эквивалентном виде —

$$\langle f, g \rangle = \frac{1}{n} \sum_{i=1}^n \varrho_i f_i g_i, \quad (2.112)$$

с нормирующим множителем  $1/n$  при сумме и какими-то положительными весовыми множителями  $\varrho_i > 0$  для отдельных компонент. Порождённая этим скалярным произведением норма  $\|\cdot\|$  определяется как

$$\|f\| := \sqrt{\frac{1}{n} \sum_{i=1}^n \varrho_i f_i^2}, \quad (2.113)$$

а расстояние между функциями дискретного аргумента, т. е. векторами  $f$  и  $g$ , есть

$$\text{dist}(f, g) = \|f - g\| = \sqrt{\frac{1}{n} \sum_{i=1}^n \varrho_i (f_i - g_i)^2}. \quad (2.114)$$

Под знаком корня в этих выражениях стоит не что иное, как усреднение квадратов компонент векторов с весовыми множителями  $\varrho_i$ ,  $i = 1, 2, \dots, n$ .

Весовые множители полезны для того, чтобы представить возможную неравноценность компонент вектора. Например, если известна информация о точности задания отдельных значений функции  $f_i$ , то веса  $\varrho_i$  можно назначать так, чтобы отразить величину этой точности, сопоставляя больший вес более точным значениям  $f_i$ . Нормирующий множитель  $\frac{1}{n}$  при суммах в (2.112), (2.113) и (2.114) удобно брать для того, чтобы с ростом размерности  $n$  (при росте количества наблюдений, измельчении сетки и т. п.) ограничить рост величины скалярного произведения и порождённой им нормы, обеспечив тем самым соизмеримость результатов при различных  $n$ .

Если  $f$  и  $g$  — функции непрерывного аргумента, то обычно полагают скалярное произведение равным

$$\langle f, g \rangle = \int_a^b \varrho(x) f(x) g(x) dx \quad (2.115)$$

для некоторой весовой функции  $\varrho(x) \geq 0$ . Будем называть его *интегральным скалярным произведением* функций  $f$  и  $g$  на интервале  $[a, b]$  с весом  $\varrho$ . Интервал интегрирования при необходимости может быть неограниченным, когда один или оба его конца бесконечны.

Выражение (2.115) с точностью до множителя можно рассматривать как предел выражения (2.112) при  $n \rightarrow \infty$ , так как в (2.112) легко угадываются интегральные суммы для интеграла Римана (2.115) по интервалу  $[a, b]$  единичной ширины и при его равномерном разбиении на подинтервалы. Тогда аналогом нормы (2.113) является

$$\|f\| := \sqrt{\int_a^b \varrho(x) (f(x))^2 dx}, \quad (2.116)$$

а расстояние между функциями вместо (2.114) определится как

$$\text{dist}(f, g) = \|f - g\| = \sqrt{\int_a^b \varrho(x) (f(x) - g(x))^2 dx}. \quad (2.117)$$

Это *среднеквадратичная метрика*, которую мы упоминали в § 2.1.

Весовая функция  $\varrho(x)$  является обобщением вектора весовых множителей, его предельным случаем, и естественно ожидать, что «почти все» её значения положительны. С другой стороны, условие строгой положительности значений нередко является чересчур жёстким и нереалистичным, так что обычно требуют, чтобы  $\varrho(x) \geq 0$  и нулевые значения весовая функция могла принимать лишь на «пренебрежимо малом» множестве аргументов. Математической формализацией последнего условия является тот факт, что множество точек, в которых допускается равенство  $\varrho(x) = 0$ , имеет меру нуль [12]: это множество можно покрыть не более чем счётным объединением интервалов сколь угодно малой общей ширины.

Итак, для задачи приближения функций или вообще элементов каких-то абстрактных пространств нахождение минимума нормы, порождённой скалярным произведением, является естественным обобщением

минимизации суммы квадратов отклонений компонент (монотонно возрастающая функция  $\sqrt{\cdot}$  никак не влияет на достижение этого минимума).

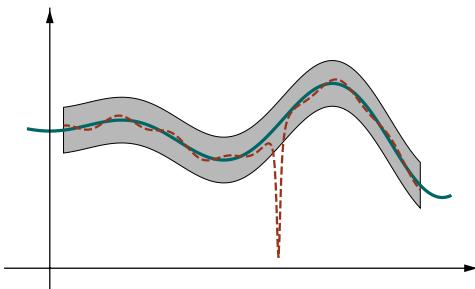


Рис. 2.30. Иллюстрация различия равномерного и интегрального (в частности, среднеквадратичного) отклонений функций

Свойства квадратичной (среднеквадратичной) метрики существенно отличаются от свойств равномерной (чебышёвской) метрики. Как следствие, в конкретных задачах может оказаться, что только какую-то одну из этих метрик и можно применять по смыслу самой постановки, тогда как другая метрика будет неадекватной. Квадратичная и среднеквадратичная метрики по самому своему построению являются «усредняющими», в которых отклонения функций друг от друга в каких-то отдельных аргументах складываются вместе и, как итог, могут компенсировать друг друга в получающейся сумме. Иными словами, малые отклонения на одних аргументах и большие отклонения на других аргументах уравновешивают друг друга, так что в целом полученное значение может оказаться приемлемым. Равномерная (чебышёвская) метрика работает по-другому, требуя, чтобы *все* отклонения соответствующих значений функций друг от друга были одинаково малы (рис. 2.30).

В целом среднеквадратичная и равномерная метрики на пространствах функций непрерывного аргумента не эквивалентны друг другу в том смысле, что сходимость в одной из них не равносильна сходимости в другой. Если последовательность функций  $g_k(x)$  сходится равномерно на  $[a, b] \subset \mathbb{R}$  к функции  $f(x)$ ,

$$\lim_{k \rightarrow \infty} \max_{x \in [a, b]} |g_k(x) - f(x)| = 0,$$

то из свойств интеграла следует

$$\lim_{k \rightarrow \infty} \sqrt{\int_a^b \varrho(x) (g_k(x) - f(x))^2 dx} = 0,$$

т. е. имеет место среднеквадратичная сходимость. Но обратное неверно: из среднеквадратичной сходимости функций не следует их равномерная сходимость.

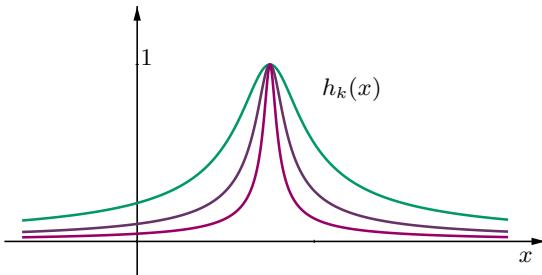


Рис. 2.31. Семейство функций с уменьшающейся среднеквадратичной нормой, у которых равномерная норма остаётся постоянной

**Пример 2.10.3** Пусть для простоты  $\varrho(x) = 1$ , т. е. весовая функция — тождественная единица. Для интервала  $[a, b]$  области определения обозначим его середину  $m := \frac{1}{2}(a + b)$ , его радиус  $r := \frac{1}{2}(b - a)$ , и определим последовательность функций

$$h_k(x) = \frac{1}{\sqrt{1 + (k(x - m))^2}}, \quad k = 1, 2, \dots$$

Это «всплески» единичной высоты вокруг середины интервала  $[a, b]$  (рис. 2.31).

Если  $f(x)$  — какая-то функция и  $g_k(x) = f(x) + h_k(x)$  — её возмущение с помощью описанного выше «всплеска», то

$$\begin{aligned} \int_a^b (g_k(x) - f(x))^2 dx &= \int_a^b (h_k(x))^2 dx = \int_a^b \frac{dx}{1 + (k(x - m))^2} = \\ &= \frac{1}{k} \operatorname{arctg}(kr) \rightarrow 0 \quad \text{при } k \rightarrow \infty. \end{aligned}$$

Иными словами, при  $k \rightarrow \infty$  среднеквадратичное отклонение функций  $g_k(x)$  и  $f(x)$  стремится к нулю. Но их равномерное отклонение равно единице при любых  $k$ , так как «высота» функций  $h_k(x)$  всегда 1.

Ценой минимального усложнения выкладок с тем же успехом вместо середины можно взять любую другую точку интервала. ■

В целом даже малое среднеквадратичное отклонение функций друг от друга может сопровождаться большой разницей их значений на каких-то небольших участках области определения, как это изображено на рис. 2.30, где красный штриховой график имеет узкий выброс по отношению к приближаемой функции. Те функции, которые равномерно приближают исходную функцию с зелёным графиком, попадают в коридор вокруг неё, залитый серым цветом.

Отметим, что аналогичный вывод об отсутствии эквивалентности справедлив также в отношении равномерной метрики и интегральной метрики (2.2) на пространствах функций. Пример 2.10.3 совершенно нетрудно адаптируется для этого случая, если взять квадраты функций  $h_k(x)$ .

## 2.11 Метод наименьших квадратов

### 2.11а Наилучшее приближение в евклидовом подпространстве

Рассмотрим подробно важный частный случай задачи о наилучшем приближении (2.107), где

- класс функций  $\mathcal{F}$ , для которых строятся приближения, — линейное пространство функций со скалярным произведением  $\langle \cdot, \cdot \rangle$ , определяющим в  $\mathcal{F}$  норму  $\|f\| = \sqrt{\langle f, f \rangle}$ ,
- класс функций  $\mathcal{G} \subseteq \mathcal{F}$ , из которого выбирается искомое наилучшее приближение для элементов из  $\mathcal{F}$ , является конечномерным линейным подпространством в  $\mathcal{F}$ .

Таким образом, наилучшее приближение ищется относительно нормы, порождающей скалярным произведением. Условимся далее называть такие приближения *квадратичными*, тогда как термин «среднеквадратичный» останется для обозначения приближений функций непрерывного аргумента в интегральной метрике (2.117).

Напомним, что конечномерное вещественное линейное векторное пространство, в котором определено скалярное произведение, называется *евклидовым*. Для постановки задачи приближения в евклидовом подпространстве, описанной выше, существование решения и его единственность следуют из общих результатов § 2.10б и § 2.10в — теоремы Бореля и теоремы 2.10.2. Но ниже мы дадим самостоятельное обоснование этих фактов вместе с конструктивным решением задачи.

Будем предполагать, что линейное подпространство  $\mathcal{G} \subseteq \mathcal{F}$  имеет размерность  $m$  и нам известен его базис  $\{\varphi_i\}_{i=1}^m$ . Для заданного  $f \in \mathcal{F}$  мы ищем наилучшее приближение  $g$  в виде

$$g = \sum_{i=1}^m c_i \varphi_i, \quad (2.118)$$

где  $c_i, i = 1, 2, \dots, m$ , — неизвестные коэффициенты, подлежащие определению.

Если через  $\Phi$  обозначить квадрат нормы отклонения  $f$  от  $g$ , то

$$\begin{aligned} \Phi &= \|f - g\|^2 = \langle f - g, f - g \rangle = \\ &= \langle f, f \rangle - 2\langle f, g \rangle + \langle g, g \rangle = \\ &= \langle f, f \rangle - 2 \sum_{i=1}^m c_i \langle f, \varphi_i \rangle + \sum_{i=1}^m \sum_{j=1}^m c_i c_j \langle \varphi_i, \varphi_j \rangle. \end{aligned} \quad (2.119)$$

Покажем, что  $\Phi(c_1, c_2, \dots, c_m)$  в действительности достигает своего минимума на всём  $\mathbb{R}^m$ .

Как видим,  $\Phi = \Phi(c_1, c_2, \dots, c_m)$  есть квадратичная форма от аргументов  $c_1, c_2, \dots, c_m$  плюс ещё некоторые линейные члены и постоянное слагаемое  $\langle f, f \rangle$ . Ясно, что  $\Phi$  принимает только неотрицательные значения. При возрастании евклидовой нормы (2-нормы) вектора коэффициентов  $c = (c_1, c_2, \dots, c_m)$  в разложении (2.118) сам вектор  $g$  неограниченно удаляется от начала координат, а функция  $\Phi(c_1, c_2, \dots, c_m)$  может принимать сколь угодно большие положительные значения.

Для обоснования последнего утверждения рассмотрим

$$v := \min_{\|c\|_2=1} \left\| \sum_{i=1}^m c_i \varphi_i \right\| \quad (2.120)$$

— минимум значений нормы вектора  $g$  из представления (2.118) по всем векторам коэффициентов, имеющим единичную евклидову норму. Множество таких векторов компактно в  $\mathbb{R}^m$  (ограничено и замкнуто), и потому значение  $v$  достигается непрерывной функцией, которой является норма вектора (2.118). Кроме того,  $v > 0$ , так как равенство  $v$  нулю означало бы существование нетривиальной зануляющейся линейной комбинации векторов базиса  $\{\varphi_i\}_{i=1}^m$ .

Тогда для любого вектора коэффициентов  $c = (c_1, c_2, \dots, c_m)$  разложения (2.118) имеем

$$\|g\| = \left\| \sum_{i=1}^m c_i \varphi_i \right\| = \|c\|_2 \cdot \left\| \sum_{i=1}^m \frac{c_i}{\|c\|} \varphi_i \right\| \geq \|c\|_2 \cdot v,$$

так что значение  $\|g\|$  при возрастании  $\|c\|_2$  может сделаться сколь угодно большим. Наконец, неограниченный рост функции  $\Phi(c) = \Phi(c_1, c_2, \dots, c_m)$  при росте  $\|g\|$  следует из неравенства

$$\Phi(c_1, c_2, \dots, c_m) = \|f - g\|^2 \geq (\|g\| - \|f\|)^2,$$

в котором  $\|f\| = \text{const}$ . По этой причине искомый  $\inf_{g \in \mathcal{G}} \|f - g\|$  не может достигаться при  $\|c\|_2 \rightarrow \infty$ .

Для отыскания минимума функции  $\Phi$  найдём её стационарные точки, т. е. точки зануления её производных. Продифференцировав выражение (2.119) по  $c_i$ ,  $i = 1, 2, \dots, m$ , и приравняв полученные производные нулю, получим

$$\frac{\partial \Phi}{\partial c_i} = -2\langle f, \varphi_i \rangle + 2 \sum_{j=1}^m c_j \langle \varphi_i, \varphi_j \rangle = 0. \quad (2.121)$$

Здесь множитель 2 при сумме всех  $c_j \langle \varphi_i, \varphi_j \rangle$  появляется оттого, что в двойной сумме из (2.119) слагаемое с  $c_i$  возникает дважды: один раз — с коэффициентом  $\langle \varphi_i, \varphi_j \rangle$ , а другой — с коэффициентом  $\langle \varphi_j, \varphi_i \rangle$ .

В целом для определения неизвестных  $c_i$ ,  $i = 1, 2, \dots, m$ , из равенств (2.121) получается система линейных алгебраических уравнений

$$\sum_{j=1}^m \langle \varphi_i, \varphi_j \rangle c_j = \langle f, \varphi_i \rangle, \quad i = 1, 2, \dots, m, \quad (2.122)$$

которую по традиции называют *системой нормальных уравнений* (или *нормальной системой уравнений*) квадратичного приближения. Объяснение этого термина даётся далее в § 2.11в. Матрица коэффициентов

этой системы имеет вид

$$\Gamma(\varphi_1, \varphi_2, \dots, \varphi_m) = \begin{pmatrix} \langle \varphi_1, \varphi_1 \rangle & \langle \varphi_1, \varphi_2 \rangle & \dots & \langle \varphi_1, \varphi_m \rangle \\ \langle \varphi_2, \varphi_1 \rangle & \langle \varphi_2, \varphi_2 \rangle & \dots & \langle \varphi_2, \varphi_m \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \varphi_m, \varphi_1 \rangle & \langle \varphi_m, \varphi_2 \rangle & \dots & \langle \varphi_m, \varphi_m \rangle \end{pmatrix} \quad (2.123)$$

и называется, как известно, *матрицей Грама* системы векторов  $\varphi_1, \varphi_2, \dots, \varphi_m$ .

В курсах линейной алгебры и аналитической геометрии показывается, что матрица Грама — это симметричная матрица, неособенная тогда и только тогда, когда векторы  $\varphi_1, \varphi_2, \dots, \varphi_m$  линейно независимы (см., к примеру, [37]). При выполнении этого условия матрица Грама является ещё и положительно определённой. Таким образом, решение системы нормальных уравнений (2.122) существует и единственno, если  $\varphi_1, \varphi_2, \dots, \varphi_m$  образуют базис в подпространстве  $\mathcal{G}$ . Тогда функция  $\Phi(c_1, c_2, \dots, c_m)$  имеет единственную стационарную точку, в которой зануляются все производные.

Найдём для функции  $\Phi$  гессиан, т. е. матрицу её вторых производных. Дифференцирование выражений (2.121) по  $c_1, c_2, \dots, c_m$  даёт матрицу вторых производных равной удвоенной матрице Грама (2.123), т. е. положительно определённой. Применяя известное из математического анализа достаточное условие экстремума функции многих переменных, которое основано на информации о вторых производных [12, 40], можем заключить, что в стационарной точке функции  $\Phi$ , т. е. на решении нормальной системы (2.122), в самом деле достигается минимум. Этот минимум является глобальным, так как других стационарных точек гладкая функция  $\Phi$  не имеет, а «на бесконечности», т. е. при неограниченном удалении аргумента  $(c_1, c_2, \dots, c_m)$  от нуля, значения  $\Phi$  неограниченно возрастают.

Подведём итоги. Для нахождения наилучшего квадратичного приближения нужно:

- 1) по базису  $\{\varphi_1, \varphi_2, \dots, \varphi_m\}$  приближающего подпространства  $\mathcal{G}$  организовать матрицу Грама  $G = \Gamma(\varphi_1, \varphi_2, \dots, \varphi_m)$ ;
- 2) по приближаемому вектору  $f$  и базису подпространства  $\mathcal{G}$  организовать вектор  $b = (\langle f, \varphi_1 \rangle, \langle f, \varphi_2 \rangle, \dots, \langle f, \varphi_m \rangle)^\top$ ;
- 3) решить систему нормальных уравнений  $Gc = b$ , определив коэффициенты разложения  $c = (c_1, c_2, \dots, c_m)^\top$  наилучшего

квадратичного приближения;

- 4) по найденным коэффициентам разложения построить вектор наилучшего квадратичного приближения  $g = \sum_{i=1}^m c_i \varphi_i$ .

Этот способ построения наилучших квадратичных приближений называют *методом наименьших квадратов*. Нередко этот термин используется вообще для любых методов построения наилучших квадратичных и среднеквадратичных приближений.

**Пример 2.11.1** Среди полиномов 2-й степени вида  $y(x) = \alpha x^2 + \beta$  с вещественными параметрами  $\alpha$  и  $\beta$  найдём тот, который имеет наименьшее квадратичное отклонение от данных

$$\begin{array}{c|ccc} x & | & 1 & 2 & 3 \\ \hline y & | & 1 & 1 & 2 \end{array} .$$

Эта задача — частный случай полиномиального квадратичного приближения функции по дискретному набору значений. Подобные задачи часто возникают в ситуациях, когда необходимо найти выражение для функциональной зависимости, наилучшим образом соответствующее данным измерений или наблюдений. Соответственно, их так и называют — *задачи восстановления зависимостей*.

В данном случае будем считать, что приближаемые функции  $\mathcal{F}$  и приближающие функции  $\mathcal{G}$  являются функциями дискретного аргумента  $x = 1, 2, 3$ . Иными словами, это просто трёхмерные векторы. Тогда мы решаем дискретный вариант задачи приближения функции из  $\mathcal{F}$  элементами двумерного линейного подпространства  $\mathcal{G}$  алгебраических полиномов второй степени вида  $y(x) = \alpha x^2 + \beta$  по значениям в точках  $x = 1, 2, 3$ .

В качестве базиса в  $\mathcal{G}$  возьмём функции  $\varphi_1 = x^2$ ,  $\varphi_2 = 1$ . На значениях аргументов 1, 2 и 3 эти функции принимают наборы значений  $(1, 4, 9)$  и  $(1, 1, 1)$ . Скалярное произведение векторов в данном случае — это сумма произведений их компонент, т. е. значений при соответствующих аргументах, и потому матрица Грама этой системы векторов

$$\begin{pmatrix} \langle \varphi_1, \varphi_1 \rangle & \langle \varphi_1, \varphi_2 \rangle \\ \langle \varphi_2, \varphi_1 \rangle & \langle \varphi_2, \varphi_2 \rangle \end{pmatrix} = \begin{pmatrix} 1 \cdot 1 + 4 \cdot 4 + 9 \cdot 9 & 1 \cdot 1 + 4 \cdot 1 + 9 \cdot 1 \\ 1 \cdot 1 + 1 \cdot 4 + 1 \cdot 9 & 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 1 \end{pmatrix} = \begin{pmatrix} 98 & 14 \\ 14 & 3 \end{pmatrix} .$$

Вектор значений данных, который нам необходимо приближать, имеет вид  $f = (1, 1, 2)^\top$ , так что правой частью системы нормальных уравнений

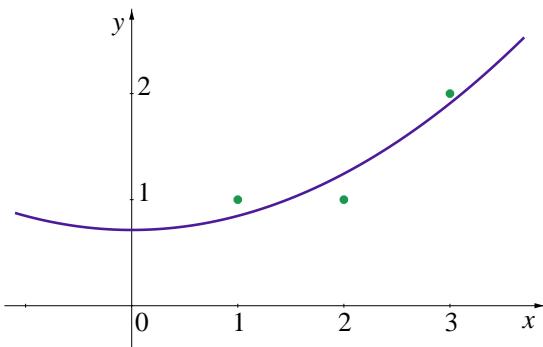


Рис. 2.32. График квадратного двучлена наилучшего квадратичного приближения

ний (2.122) является

$$\begin{pmatrix} \langle f, \varphi_1 \rangle \\ \langle f, \varphi_2 \rangle \end{pmatrix} = \begin{pmatrix} 1 \cdot 1 + 1 \cdot 4 + 2 \cdot 9 \\ 1 \cdot 1 + 1 \cdot 1 + 2 \cdot 1 \end{pmatrix} = \begin{pmatrix} 23 \\ 4 \end{pmatrix}.$$

Система нормальных уравнений (2.122) для определения коэффициентов разложения наилучшего приближения принимает в нашем случае вид

$$\begin{pmatrix} 98 & 14 \\ 14 & 3 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 23 \\ 4 \end{pmatrix}.$$

Её решение равно  $(13/98, 70/98)^\top$ , так что  $\alpha \approx 0.1326531$ ,  $\beta \approx 0.7142857$ . В целом искомой функцией, наилучшим образом приближающей данные среди всех квадратных двучленов, будет

$$y(x) = \frac{13}{98}x^2 + \frac{70}{98} \approx 0.1326531x^2 + 0.7142857,$$

и её график приведён на рис. 2.32. Из чертежа видно, что построенная функция в самом деле даёт неплохое приближение к данным, которые изображены кружками. ■

Рассмотренный пример легко обобщается на случай полинома произвольной степени с любым числом параметров-коэффициентов.

Обратимся теперь к практическим аспектам реализации развитого выше метода и обсудим свойства системы нормальных уравнений

(2.122). Наиболее простой вид матрица Грама имеет в случае, когда базисные функции  $\varphi_i$  ортогональны друг другу, т. е. когда  $\langle \varphi_i, \varphi_j \rangle = 0$  при  $i \neq j$ . Тогда система нормальных уравнений (2.122) становится диагональной и решается тривиально. Соответствующее наилучшее приближение имеет вид суммы

$$g = \sum_{i=1}^m c_i \varphi_i, \quad \text{где } c_i = \frac{\langle f, \varphi_i \rangle}{\langle \varphi_i, \varphi_i \rangle}, \quad i = 1, 2, \dots, m. \quad (2.124)$$

Это представление, как известно, называется *рядом Фурье* для  $f$  по ортогональной системе векторов  $\{\varphi_i\}_{i=1}^m$ . Коэффициенты  $c_i$  из (2.124) называют при этом *коэффициентами Фурье* разложения функции  $f$ . В нашем случае ряд Фурье конечен, но он может быть и бесконечным, если взять бесконечный базис  $\{\varphi_i\}$ , например, тригонометрическую систему из примера 2.11.6 (см. подробности в [12, 16, 32]).

Кроме того, в случае ортогонального и близкого к ортогональному базиса  $\{\varphi_i\}_{i=1}^m$  решение нормальной системы (2.122) устойчиво к возмущениям в правой части и неизбежным погрешностям вычислений. Но если базис линейного подпространства  $\mathcal{G}$  сильно отличается от ортогонального, то свойства системы уравнений (2.122) могут быть плохими в том смысле, что её решение будет чувствительным к возмущениям данных и погрешностям вычислений.

## 2.11б Квадратичное приближение из линейной оболочки векторов

Рассмотрим теперь более общий, но одновременно и более практический случай задачи наилучшего приближения в евклидовом пространстве. Будем считать, что множество  $\mathcal{G}$ , из которого мы должны выбрать наилучшее приближение, — это не линейное векторное подпространство с известным базисом, а линейная оболочка набора векторов, которые могут и не быть базисом. Напомним, что линейной оболочкой заданного набора векторов  $v_1, v_2, \dots, v_m$  называется множество

$$\text{span} \{ v_1, v_2, \dots, v_m \} := \left\{ \sum_{i=1}^m \alpha_i v_i \mid \alpha_i \in \mathbb{R} \right\},$$

образованное всевозможными линейными комбинациями данных векторов. Сами эти векторы будем называть *порождающими* для линей-

ной оболочки  $\text{span} \{ v_1, v_2, \dots, v_m \}$ . Эквивалентное определение: линейной оболочкой заданного набора векторов называется наименьшее по включению линейное подпространство, содержащее все эти векторы.<sup>23</sup> То обстоятельство, что порождающие векторы не образуют базиса, означает «избыточность» этого набора, т. е. что некоторые из его векторов являются линейными комбинациями остальных векторов.

Итак, пусть теперь

$$\mathcal{G} = \text{span} \{ \varphi_1, \varphi_2, \dots, \varphi_m \}$$

для каких-то  $\varphi_1, \varphi_2, \dots, \varphi_m$ . Такие наборы векторов часто получаются в результате наблюдений или измерений, но проверка их линейной зависимости или независимости является, как правило, самостоятельной нетривиальной задачей. Кроме того, наилучшее приближение нужно как-то находить даже при линейной зависимости этих векторов. Что изменится в наших конструкциях?

Как и раньше, можно искать наилучшее квадратичное приближение в виде линейной комбинации (2.118)

$$g = \sum_{i=1}^m c_i \varphi_i,$$

где  $c_i$  — некоторые неизвестные коэффициенты. Хотя векторы  $\varphi_1, \varphi_2, \dots, \varphi_m$  могут не образовывать базиса в  $\mathcal{G}$ , но ничего лучшего у нас нет, а линейная комбинация (2.118) всё-таки позволяет представить любой элемент из  $\mathcal{G}$ . Квадрат отклонения  $g$  от  $f$ , который является функцией от коэффициентов разложения  $c_1, c_2, \dots, c_m$  как и прежде равен

$$\begin{aligned} \Phi(c_1, c_2, \dots, c_m) &= \|f - g\|^2 = \\ &= \langle f, f \rangle - 2 \sum_{i=1}^m c_i \langle f, \varphi_i \rangle + \sum_{i=1}^m \sum_{j=1}^m c_i c_j \langle \varphi_i, \varphi_j \rangle. \end{aligned}$$

Он является квадратичной формой от аргументов  $c_1, c_2, \dots, c_m$ , дополненной линейными членами и постоянным слагаемым. Функция  $\Phi$  всегда неотрицательна, но при возрастании евклидовой нормы вектора коэффициентов  $(c_1, c_2, \dots, c_m)$ , когда вектор  $g$  устремляется «к бесконечности», функция  $\Phi$  не обязательно принимает сколь угодно большие

---

<sup>23</sup>Обозначение «span» является почти стандартным для этого объекта и происходит от английского термина linear span для линейной оболочки векторов.

значения. Доказательство этого факта, данное в § 2.11а, теперь не работает, так как при линейной зависимости векторов  $\varphi_1, \varphi_2, \dots, \varphi_m$  минимум (2.120) может занулиться.

Стационарными точками функции  $\Phi$ , как и ранее, служат решения системы нормальных уравнений (2.122) с матрицей (2.123), которая является матрицей Грама семейства векторов  $\varphi_1, \varphi_2, \dots, \varphi_m$ . Но теперь нельзя утверждать, что эта матрица неособенна и положительно определена. Она может быть особенной, если векторы  $\varphi_1, \varphi_2, \dots, \varphi_m$  линейно зависимы. Как следствие, становятся неприменимыми рассуждения из § 2.11а, обосновывающие существование и единственность решения нормальной системы (2.122) для нахождения коэффициентов разложения (2.118).

Посмотрим, тем не менее, на нашу задачу с точки зрения общей теории из § 2.10б и § 2.10в. Линейная оболочка  $\mathfrak{G}$  является конечномерным линейным подпространством в нормированном пространстве, и потому в силу теоремы Бореля (теорема 2.10.1) в  $\mathfrak{G}$  обязательно существует некоторый элемент наилучшего приближения  $g$ . Далее, линейное векторное пространство с нормой, порождённой скалярным произведением (в частности, евклидово), является строго нормированным. Поэтому из теоремы 2.10.2 следует, что решение  $g$  нашей задачи квадратичного приближение всегда единствено.

В точке минимума гладкой функции  $\Phi$  на  $\mathbb{R}^m$  её частные производные (2.121) обязательно зануляются, и поэтому сказанное влечёт разрешимость системы нормальных уравнений (2.122), из которой определяются коэффициенты  $c_1, c_2, \dots, c_m$  разложения наилучшего приближения  $g$ . Но теперь матрица этой системы может быть особенной, и потому единственность решения, т. е. единственность набора коэффициентов в представлении (2.118) для  $g$ , гарантировать уже нельзя. Впрочем, это не слишком большая потеря.

Покажем, что на всяком решении  $\tilde{c} = (\tilde{c}_1, \tilde{c}_2, \dots, \tilde{c}_m)$  системы нормальных уравнений (2.122) достигается минимум квадрата расстояния от приближающего элемента  $g$  до приближаемого  $f$ , т. е. функции  $\Phi(c) = \Phi(c_1, c_2, \dots, c_m) = \|f - g\|^2$ .

Разлагая гладкую функцию  $\Phi(c)$  в точке  $\tilde{c}$  по формуле Тейлора, получим сумму постоянной и членов первой и второй степеней относительно компонент вектора  $c = (c_1, c_2, \dots, c_m)^\top$ :

$$\Phi(c) = \Phi(\tilde{c}) + \Phi'(\tilde{c})(c - \tilde{c}) + (c - \tilde{c})^\top \Phi''(\tilde{c})(c - \tilde{c}), \quad (2.125)$$

где  $\Phi'(\tilde{c})$  — вектор-строка из частных производных функции  $\Phi$  (т. е.

градиент) в точке  $\tilde{c}$ , а  $\Phi''(\tilde{c})$  — матрица вторых производных (гессиан) функции  $\Phi$  в точке  $\tilde{c}$ . Членов третьего и более высоких порядков в этом представлении нет, так как уже вторые производные функции  $\Phi$  постоянны (и потому можно писать просто  $\Phi''$  вместо  $\Phi''(\tilde{c})$ ).

Но  $\Phi'(\tilde{c}) = 0$ , так как  $\tilde{c}$  — решение системы нормальных уравнений (2.122), которая выписана из условия равенства нулю производных  $\Phi$ . Кроме того, гессиан  $\Phi''(\tilde{c})$ , как мы выяснили в § 2.11а, равен удвоенной матрице Грама системы векторов  $\varphi_1, \varphi_2, \dots, \varphi_m$ . Матрица Грама всегда положительно полуопределенна, и поэтому для любых  $c$

$$(c - \tilde{c})^\top \Phi''(\tilde{c}) (c - \tilde{c}) \geq 0.$$

Как следствие, значение

$$\Phi(c) = \Phi(\tilde{c}) + (c - \tilde{c})^\top \Phi''(\tilde{c}) (c - \tilde{c}) \quad (2.126)$$

при любом  $c$  не меньше, чем  $\Phi(\tilde{c})$ , т. е. в точке  $\tilde{c}$  в самом деле имеется локальный минимум функции  $\Phi(c)$ . Другое обоснование этого факта будет дано в следующем § 2.11в.

## 2.11в Геометрия наилучшего квадратичного приближения

Задача наилучшего квадратичного приближения рассматривалась выше с помощью аналитических инструментов. Но решение этой задачи имеет также наглядную геометрическую интерпретацию, на которую указал А.Н. Колмогоров в 1946 году [64] (см. также [70]).

Пусть  $X$  — линейное пространство, в котором задано скалярное произведение  $\langle \cdot, \cdot \rangle$ . Как известно, наличие скалярного произведения позволяет ввести понятие ортогональности (перпендикулярности) векторов, обобщающее известное из элементарной геометрии и очень полезное свойство перпендикулярности отрезков или прямых. Векторы  $u$  и  $v$  из  $X$  называют *ортогональными*, если  $\langle u, v \rangle = 0$ . Этот факт обозначают также записью  $u \perp v$ . Нулевой вектор оказывается, таким образом, ортогонален любому другому вектору, но этот случай малосодержателен.

Набор ненулевых векторов  $\{u_i\}$  называется *ортогональным*, если  $\langle u_j, u_k \rangle = 0$  для  $j \neq k$ . Говорят также, что вектор  $u$  ортогонален множеству  $V \subseteq X$ , если  $u \perp v$  для любого  $v \in V$ .

Пусть  $U$  — линейное подпространство в  $X$ . Тогда для любого вектора  $x \in X$  всегда существует разложение  $x = u + v$ , в котором  $u \in U$ ,

а  $v \perp U$ . Вектор  $u$  в этом случае называется *ортогональной проекцией* вектора  $x$  на  $U$  (или перпендикулярной проекцией  $x$  на  $U$ ), а вектор  $v$  — *перпендикуляром*, опущенным из  $x$  на  $U$ .

Проекция и перпендикуляр вектора определяются однозначно, так как при существовании другого представления  $x = u' + v'$  из равенства  $u + v = u' + v'$  вытекало бы

$$u - u' = v' - v.$$

В этом равенстве слева стоит вектор из подпространства  $U$ , а справа — ему ортогональный, так что их равенство возможно лишь в случае, когда оба они — нулевые.

Как конструктивно находить ортогональную проекцию на линейное подпространство?

Пусть  $f$  — некоторый вектор из  $X$  и задано конечномерное подпространство  $U \subseteq X$ , которое является линейной оболочкой семейства порождающих векторов  $\{\varphi_i\}_{i=1}^m$ . Ортогональную проекцию  $f$  на  $U$  будем искать в виде линейной комбинации

$$g = \sum_{j=1}^m c_j \varphi_j, \quad (2.118)$$

где  $c_j$  — некоторые неизвестные коэффициенты. Условие ортогональности разности  $(f - g)$  и подпространства  $U$  требует, чтобы

$$\langle f - g, \varphi_i \rangle = 0, \quad i = 1, 2, \dots, m.$$

Подставив сюда вместо  $g$  его выражение (2.118) через порождающие векторы  $\varphi_i$ , будем иметь

$$\left\langle f - \sum_{j=1}^m c_j \varphi_j, \varphi_i \right\rangle = 0, \quad i = 1, 2, \dots, m.$$

Раскрывая угловые скобки согласно свойствам скалярного произведения и разнося члены в противоположные части равенств, получим

$$\sum_{j=1}^m \langle \varphi_i, \varphi_j \rangle c_j = \langle f, \varphi_i \rangle, \quad i = 1, 2, \dots, m. \quad (2.127)$$

Это система линейных алгебраических уравнений, которая в точности совпадает с системой нормальных уравнений (2.122) для определения коэффициентов  $c_j$ , выведенной ранее другим способом. Из этого факта вытекают разнообразные полезные следствия.

Во-первых, из разрешимости нормальной системы уравнений следует, что для любого вектора и любого конечномерного подпространства существуют и единственны его перпендикуляр и ортогональная проекция на это подпространство. Это утверждение имеет самостоятельную ценность и часто называется «теоремой о перпендикуляре» [37].

В самом деле, если система уравнений (2.127) разрешима относительно коэффициентов  $c_j$ , то по ним восстанавливается вектор  $g$ , т. е. ортогональная проекция  $f$ , а также перпендикуляр  $(f - g)$ . Они единственны, как установлено выше. Отметим также очевидный факт, вытекающий из этих рассуждений: линейная оболочка подпространства и любого вектора совпадает с линейной оболочкой подпространства и перпендикуляра этого вектора.

Во-вторых, справедлив следующий результат о геометрической характеристизации наилучшего приближения в евклидовом или унитарном подпространстве:

**Теорема 2.11.1** *Пусть  $X$  — линейное векторное пространство со скалярным произведением,  $U$  — его конечномерное линейное подпространство. Вектор  $g \in U$  является наилучшим приближением для  $f \in X$  относительно нормы, порождённой скалярным произведением в  $X$ , тогда и только тогда, когда он является ортогональной проекцией  $f$  на  $U$ .*

Обоснование этого результата фактически уже дано в предшествующих разделах, но мы дадим здесь другое доказательство, которое более глубоко иллюстрирует идею приближения и имеет большую общность.

**Доказательство.** Нам достаточно показать, что величина  $\|f - g\|^2 = \langle f - g, f - g \rangle$  достигает минимума, если и только если  $(f - g) \perp U$ .

Разложим разность  $(f - g)$  на компоненты, лежащую в  $U$  и ортогональную  $U$ :

$$f - g = u + v, \quad \text{где } u \in U, v \perp U,$$

и покажем, что для минимальности  $\|f - g\|^2$  необходимо и достаточно равенства  $u = 0$ . Это и означает, что  $g$  — ортогональная проекция  $f$ . На рисунке рис. 2.33, который иллюстрирует ситуацию теоремы, разность

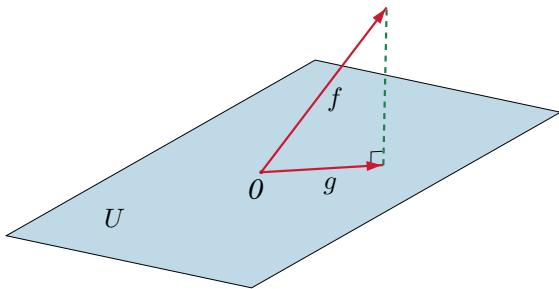


Рис. 2.33. Наилучшее приближение в евклидовом подпространстве — ортогональная проекция вектора на это подпространство

$(f - g)$  является вектором штриховой линии, идущим от плоскости  $U$  к концу вектора  $f$ .

Имеем

$$\begin{aligned} \|f - g\|^2 &= \langle f - g, f - g \rangle = \langle u + v, u + v \rangle = \\ &= \langle u, u \rangle + 2\langle u, v \rangle + \langle v, v \rangle = \|u\|^2 + \|v\|^2, \end{aligned} \quad (2.128)$$

так как  $u \perp v$  по условию разложения разности  $f - g$ . Ясно, что нахождение минимума  $\|f - g\|^2$  за счёт выбора  $g$ , лежащего в  $U$ , равносильно нахождению минимума  $\|f - g\|^2 = \|u\|^2 + \|v\|^2$  за счёт выбора  $u \in U$ . Оно требует  $\|u\|^2 = 0$ .

Наоборот, если  $u = 0$ , то из (2.128) следует, что разность  $(f - g)$  получает наименьшую норму. Таким образом, в любом случае  $f - g = v$ , т. е. разность  $(f - g)$  должна быть ортогональна подпространству  $U$ , в котором мы находим приближение. ■

Теорема 2.11.1 показывает, что каждое уравнение нормальной системы — это не что иное, как условие ортогональности (нормальности) разности  $(f - g)$  отдельным векторам, порождающим подпространство, в котором выбирается наилучшее приближение. Данный выше «геометрический» вывод нормальной системы уравнений одинаково пригоден как для линейно независимой системы векторов  $\{\varphi_i\}_{i=1}^m$ , так и для случая, когда в ней есть линейно зависимые векторы (мы специально рассматривали его в § 2.11б). Кроме того, теорема 2.11.1 и её доказательство верны также в унитарных пространствах, где дифференцирование

по комплексным коэффициентам является нетривиальным и где наш вывод метода наименьших квадратов из § 2.11а не работает.

## 2.11г Псевдорешения систем линейных алгебраических уравнений

Пусть дана система линейных алгебраических уравнений

$$\left\{ \begin{array}{l} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = b_2, \\ \vdots \qquad \vdots \qquad \ddots \qquad \vdots \qquad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n = b_m \end{array} \right.$$

с коэффициентами  $a_{ij}$  и свободными членами  $b_i$ , или в краткой форме

$$Ax = b$$

с  $m \times n$ -матрицей  $A = (a_{ij})$  и  $m$ -вектором правых частей  $b = (b_i)$ . Решение таких систем уравнений будет подробно рассматриваться в главе 3, а сейчас мы исследуем их с точки зрения теории приближения, развитой в предшествующих разделах. Она будет применена для нахождения так называемых псевдорешений.

Предположим, что в выписанной системе количество уравнений  $m$  не равно количеству неизвестных  $n$ , или же  $m = n$ , но ничего не известно о неособенности матрицы  $A$ . Тогда обычного решения системы может не иметь. Но во многих практических задачах, где возникает такая ситуация, точного равенства левой и правой частей системы на самом деле не требуется и заменой традиционному понятию решения может служить вектор, на котором правая и левая части системы имеют «наименьшее отличие» друг от друга.

**Определение 2.11.1** Невязкой приближённого решения  $\tilde{x}$  системы уравнений (или одного уравнения) называется разность левой и правой частей при подстановке в них  $\tilde{x}$ .

Нередко невязкой называют также функцию разности левой и правой частей уравнения или системы уравнений в зависимости от  $\tilde{x}$ .

Итак, вместо обычного решения системы уравнений, если оно не существует, можно рассмотреть такой вектор, на котором достигается

«наименьшая величина» невязки. Но нужно уточнить, в каком именно смысле понимается эта наименьшая «величина». Удобно использовать здесь какую-либо норму невязки, так что в качестве заменителя обычного решения системы берётся вектор, которому соответствует минимальная норма невязки.

**Определение 2.11.2** Псевдорешением системы уравнений относительно заданной нормы называется набор значений неизвестных переменных этой системы, на котором достигается наименьшее значение выбранной нормы её невязки.

Определение псевдорешения очевидным образом прилагается также к отдельным уравнениям, и вместо нормы невязки тогда рассматривается просто модуль невязки.

Поскольку для любой  $m \times n$ -матрицы  $A = (a_{ij})$  и произвольного  $n$ -вектора  $x = (x_1, x_2, \dots, x_n)^\top$  справедливо

$$Ax = \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{pmatrix} x_1 + \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{pmatrix} x_2 + \cdots + \begin{pmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{pmatrix} x_n,$$

то задача нахождения псевдорешений системы линейных алгебраических уравнений  $Ax = b$  сводится к построению наилучшего приближения, относительно выбранной нормы, для вектора правой части  $b$  из линейной оболочки вектор-столбцов матрицы  $A$ . Для случая евклидовой нормы (2-нормы), задаваемой как (2.111)

$$\|f\|_2 := \sqrt{\sum_{i=1}^n |f_i|^2},$$

можем применить теорию квадратичного приближения, развитую выше в § 2.11a и § 2.11b.

Псевдорешение системы линейных уравнений  $Ax = b$  относительно евклидовой нормы — вектор коэффициентов разложения по столбцам матрицы  $A$  для наилучшего квадратичного приближения правой части  $b$ . Эти коэффициенты разложения определяются из системы нормальных уравнений (2.122) с матрицей Грама для вектор-столбцов  $A$ .

Следовательно, в матрице системы уравнений (2.122) элемент на месте  $(i, j)$ , т. е. скалярное произведение  $i$ -го и  $j$ -го столбцов из  $A$ , равен

$$\sum_{k=1}^m a_{ki} a_{kj}.$$

Легко видеть, что это не что иное, как  $ij$ -й элемент матрицы  $A^\top A$ .

В правой части системы нормальных уравнений (2.122) в качестве  $i$ -й компоненты стоит скалярное произведение  $i$ -го столбца  $A$  и вектора  $b$ , т. е.

$$\sum_{k=1}^m a_{ki} b_k,$$

и это  $i$ -я компонента вектора  $A^\top b$ . Итак, система нормальных уравнений (2.122) для определения наилучшего приближения имеет в нашем случае вид

$$A^\top Ax = A^\top b.$$

**Определение 2.11.3** Для заданной системы линейных алгебраических уравнений  $Ax = b$  система  $A^\top Ax = A^\top b$ , полученная домножением обеих частей слева на матрицу  $A^\top$ , называется нормальной системой уравнений или же системой нормальных уравнений.

Отметим, что матрица  $A^\top A$  нормальной системы уравнений симметрична и положительно полуопределена.

**Теорема 2.11.2** Для системы линейных алгебраических уравнений  $Ax = b$  псевдорешения относительно евклидовой нормы и только они являются решениями нормальной системы уравнений  $A^\top Ax = A^\top b$ .

**Доказательство** немедленно следует из результатов раздела § 2.11б.

**Предложение 2.11.1** Нормальная система уравнений  $A^\top Ax = A^\top b$  всегда имеет решение.

Фактически мы обосновали это утверждение в § 2.11б на основе теоремы Бореля, исходя из того, что решение нормальной системы даёт наилучшее квадратичное приближение. Но ниже даётся одно прямое доказательство, которое не опирается на общие результаты теории приближений.

**Доказательство** будет использовать критерий разрешимости системы линейных алгебраических уравнений, известный как «теорема Фредгольма» (см. § 3.23). Он связывает разрешимость исходной системы со свойствами так называемой транспонированной однородной системы, у которой правая часть — нулевая, а матрица получена из матрицы исходной системы транспонированием. Для нормальной системы уравнений однородная транспонированная система имеет вид  $A^\top A\tilde{y} = 0$ , так как её матрица, очевидно, совпадает с симметричной матрицей нормальной системы.

Если  $A^\top A\tilde{y} = 0$  для некоторого  $\tilde{y}$ , то

$$0 = \tilde{y}^\top (A^\top A\tilde{y}) = (\tilde{y}^\top A^\top)(A\tilde{y}) = (A\tilde{y})^\top (A\tilde{y}) = \|A\tilde{y}\|_2^2,$$

откуда следует, что  $A\tilde{y} = 0$ . Следовательно,

$$\langle \tilde{y}, A^\top b \rangle = \tilde{y}^\top (A^\top b) = (\tilde{y}^\top A^\top)b = (A\tilde{y})^\top b = \langle A\tilde{y}, b \rangle = 0,$$

т. е. любое решение  $\tilde{y}$  однородной транспонированной системы ортогонально вектору  $A^\top b$ , стоящему в правой части нормальной системы уравнений. В силу альтернативы Фредгольма можем заключить, что система линейных уравнений  $A^\top Ax = A^\top b$  в самом деле должна быть разрешимой. ■

Предположим, что в исходной системе уравнений  $m \geq n$ , т. е. она квадратная или переопределённая (наиболее частый на практике случай), а матрица  $A$  имеет полный ранг. Тогда  $A^\top A$  — неособенная квадратная матрица размера  $n \times n$  и можно выписать решение нормальной системы в явном виде как

$$(A^\top A)^{-1} A^\top b.$$

В этом выражении матрица  $(A^\top A)^{-1} A^\top$  размера  $n \times m$  обладает свойствами обратной матрицы для  $A$ , и потому её называют *псевдообратной матрицей*.

Если в описанной выше ситуации ( $m \geq n$ ) матрица системы имеет неполный ранг, т. е. меньший  $\min\{m, n\}$ , или же исходная система недопределена ( $m < n$ ), то псевдорешение может быть неединственным. Кроме того, псевдорешение оказывается неустойчивым к возмущениям элементов матрицы системы, если она имеет неполный ранг. По

этим причинам из всех псевдорешений обычно выделяют так называемые *нормальные псевдорешения*, которые имеют наименьшую евклидову норму. Из теоремы 2.10.2 следует, что нормальное псевдорешение системы линейных алгебраических уравнений единственno, так как евклидова норма делает линейное пространство строго нормированным. Тем не менее нормальные псевдорешения также могут скачкообразно меняться при возмущениях элементов матрицы системы, которые приводят к изменению её ранга.

Дальнейшее обсуждение темы и обзор численных методов можно найти в § 3.16.

## 2.11д Приложения к анализу данных

Применим полученные выше результаты к одной из популярных задач обработки данных — к задаче восстановления зависимостей по данным измерений или наблюдений. Характерной особенностью любых экспериментальных данных является то, что они почти всегда не вполне точны. Но теория квадратичного приближения, рассмотренная в предшествующих разделах, позволяет решать задачу построения линейной функции нескольких переменных вида

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n, \quad (2.129)$$

которая наилучшим образом приближает неточные данные в квадратичном смысле.

Пусть заданы  $m$  наборов значений независимых переменных и соответствующих им значений зависимой переменной

$$\begin{array}{cccccc} x_1^{(1)}, & x_2^{(1)}, & \dots, & x_n^{(1)}, & y^{(1)}, \\ x_1^{(2)}, & x_2^{(2)}, & \dots, & x_n^{(2)}, & y^{(2)}, \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_1^{(m)}, & x_2^{(m)}, & \dots, & x_n^{(m)}, & y^{(m)}. \end{array}$$

На практике они получаются обычно в результате отдельных измерений (наблюдений) исследуемой функциональной зависимости, а верхний индекс в скобках означает номер измерения или наблюдения. Необходимо найти такие значения свободного члена  $\beta_0$  и коэффициентов  $\beta_1, \dots, \beta_n$  линейной функции (2.129), чтобы принимаемые ею при  $x_1 = x_1^{(k)}, x_2 = x_2^{(k)}, \dots, x_n = x_n^{(k)}$  значения в совокупности наименее отличались в квадратичном смысле от значений  $y^{(k)}, k = 1, 2, \dots, m$ .

В статистике разности измеренных значений функции  $y^{(k)}$  и значений конструируемой функции на заданных аргументах  $x_1^{(k)}, x_2^{(k)}, \dots, x_n^{(k)}$ , т. е.

$$y^{(k)} - (\beta_0 + \beta_1 x_1^{(k)} + \dots + \beta_n x_n^{(k)}), \quad k = 1, 2, \dots, m,$$

называются *остатками*. Они характеризуют отклонение значений конструируемой функциональной зависимости от реальных данных, и в задаче восстановления зависимости мы должны добиться их наиболее сильного уменьшения с помощью подходящего выбора параметров функции. В нашем случае квадратичного приближения искомые  $\beta_0, \beta_1, \dots, \beta_n$  должны доставлять минимум выражению

$$\sqrt{\sum_{k=1}^m (y^{(k)} - (\beta_0 + \beta_1 x_1^{(k)} + \dots + \beta_n x_n^{(k)}))^2}. \quad (2.130)$$

Введём  $m \times (n+1)$ -матрицу  $X$  и  $m$ -вектор  $y$  следующим образом:

$$X = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \dots & x_n^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \dots & x_n^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(m)} & x_2^{(m)} & \dots & x_n^{(m)} \end{pmatrix}, \quad y = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{pmatrix}.$$

Нетрудно понять, что решением поставленной задачи, т. е. искомым вектором параметров  $\beta = (\beta_0, \beta_1, \dots, \beta_n)^\top$  линейной функции (2.129), является псевдорешение системы линейных алгебраических уравнений

$$X\beta = y, \quad (2.131)$$

минимизирующее евклидову норму её невязки (2.130). Оно, как следует из результатов предыдущего раздела, является решением нормальной системы уравнений

$$(X^\top X)\beta = X^\top y.$$

Если  $m > n$ , т. е. количество измерений  $m$  не меньше числа оцениваемых параметров  $n+1$ , и матрица  $X$  имеет полный ранг, то для оценки  $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_n)^\top$  вектора параметров линейной зависимости (2.129) можно выписать явное выражение в матричной форме

$$\hat{\beta} = (X^\top X)^{-1} X^\top y.$$

Но в конкретных задачах лучше использовать не эту формулу (которая имеет главным образом теоретическое значение), а численные методы для решения нормальной системы уравнений или даже специализированные методы, которые предназначены для нахождения псевдорешений исходной системы (2.131). Некоторые из них обсуждаются в главе 3.

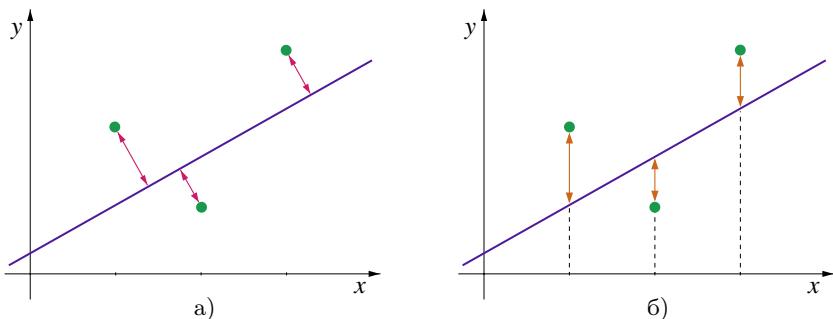


Рис. 2.34. Различные способы определения отклонения данных от графика функциональной зависимости

Рассмотренный способ конструирования линейной функции наилучшего квадратичного приближения соответствует чертежу на рис. 2.34б. Как и в случае общей задачи приближения из § 2.11а, он называется *методом наименьших квадратов* (очень распространена также аббревиатура «МНК»). В математической статистике и при анализе данных метод наименьших квадратов является одним из главных инструментов так называемого регрессионного анализа — дисциплины, которая изучает влияние одной или нескольких независимых переменных на зависимую переменную. Кроме того, метод наименьших квадратов широко применяется также в машинном обучении.

Возникает интересный методический вопрос: почему мы не строим прямую, приближающую данные, таким способом, который изображён на рис. 2.34а, когда минимизируется обычное расстояние от этой прямой до точек данных? почему отклонение от точек до прямой берётся только «по вертикали»?

Дело в том, что по разным осям системы координат, в которой построен график функции, могут откладываться совершенно разные по смыслу величины, смешивать которые смысла не имеет. Они являются

значениями независимой переменной и значениями интересующей нас функции, у которых даже физические размерности, возможно, различаются (граммы и секунды и т. п.). Кроме того, это могут быть, к примеру, измерения и наблюдения, выполняемые где-то очень далеко друг от друга, как в пространстве, так и во времени. Но при определении расстояния от точек данных до приближающей их линии (прямой), как на рис. 2.34а, которое выполняется «по косой», мы берём, фактически, длину линейной комбинации аргументов функции и зависимой переменной. Часто этой линейной комбинации разнородных данных ничего практически разумного не соответствует.

Способ приближения, показанный на рис. 2.34а, иногда тоже имеет смысл и реально применяется на практике,<sup>24</sup> но в задачах, где данные по различным осям выражают какие-то однокачественные признаки. Например, это могут быть пространственные координаты тела (материальной точки) и т. п.

**Пример 2.11.2** Среди всех линейных функций вида

$$y(x) = \beta_1 x + \beta_0$$

найдём ту, которая в заданных аргументах имеет наименьшее квадратичное отклонение от указанных значений:

$x$	1	2	3	
$y$	1	1	2	

В данном случае можно считать набор параметров линейной зависимости, т. е. вектор  $(\beta_0, \beta_1)^\top$ , псевдорешением относительно евклидовой нормы для системы линейных алгебраических уравнений

$$\begin{cases} \beta_0 + \beta_1 = 1, \\ \beta_0 + 2\beta_1 = 1, \\ \beta_0 + 3\beta_1 = 2, \end{cases}$$

которая получается при подстановке в выражение для функции  $y(x)$  значений аргументов  $x = 1, 2, 3$  и приравниваем значениям функции.

В векторно-матричной форме эта система имеет вид

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix},$$

---

<sup>24</sup>Часто его называют *ортогональной регрессией*.

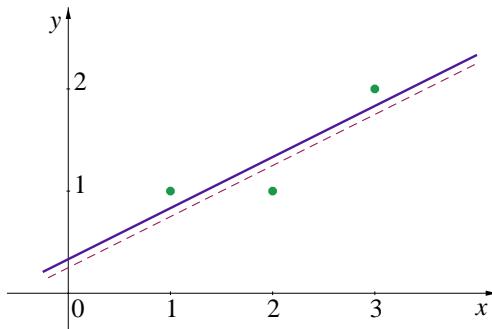


Рис. 2.35. Графики линейных функций наилучшего квадратичного приближения и наилучшего чебышёвского приближения

и для неё нормальная система уравнений выглядит следующим образом:

$$\begin{pmatrix} 3 & 6 \\ 6 & 14 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 4 \\ 9 \end{pmatrix}.$$

Нетрудно найти решение системы нормальных уравнений (2.122) для определения коэффициентов разложения. Оно равно  $(\frac{1}{3}, \frac{1}{2})^\top$ , так что искомой линейной функцией, которая наилучшим образом приближает данные, будет

$$y(x) = \frac{1}{2}x + \frac{1}{3}.$$

Её график изображён на рис. 2.35 сплошной прямой. Для сравнения на том же рисунке тонким пунктиром представлен график линейной функции наилучшего чебышёвского (равномерного) приближения тех же данных. Она задаётся выражением  $\frac{1}{2}x + \frac{1}{4}$  с одинаковым угловым коэффициентом, но другим свободным членом. ■

Квадратичные приближения и метод наименьших квадратов для решения переопределённых систем уравнений, которые возникают в связи с задачами обработки наблюдений, были почти одновременно предложены на рубеже XVIII–XIX веков А.-М. Лежандром, Р. Эдрейном и К.Ф. Гауссом. Современное название этому подходу тоже дал А.-М. Лежандр.

На практике метод наименьших квадратов находит широчайшее применение в силу двух главных причин. Во-первых, его применение

бывает вызвано содержательным смыслом задачи, в которой в качестве меры отклонения возникает именно сумма квадратов разностей компонент или интеграл от квадрата разности функций. Именно так обстоит дело в теоретико-вероятностном обосновании метода наименьших квадратов (см., к примеру, [70]). Впервые оно было дано К.Ф. Гауссом и далее доведено до современного состояния в трудах П.С. Лапласа, П.Л. Чебышёва, А.А. Маркова, А.Н. Колмогорова и многих других математиков.

Во-вторых, в линейных задачах метод наименьших квадратов сводит построение наилучшего приближения к решению системы линейных уравнений, т. е. к хорошо разработанной вычислительной задаче. Если для измерения расстояния между функциями применяются какие-то другие метрики, отличные от квадратичной или среднеквадратичной, то их минимизация требует решения задачи вычислительной оптимизации, что может оказаться более трудным или неудобным для решения.

В целом, если какое-либо одно или оба из выписанных условий не выполняются, то метод наименьших квадратов становится не самой лучшей возможностью решения задачи приближения или задачи восстановления функциональной зависимости.

## 2.11e Среднеквадратичное приближение функций

В этом разделе мы применим развитый в § 2.11a общий подход к конкретной задаче наилучшего среднеквадратичного приближения функций, заданных на интервале вещественной оси. Помимо математической элегантности среднеквадратичное приближение в прикладных задачах, как правило, имеет ясный содержательный смысл.

**Пример 2.11.3** Рассмотрим тепловое действие электрического тока в проводнике. Если величина силы тока в зависимости от времени  $t$  описывается функцией  $I(t)$ , а сопротивление проводника равно  $R$ , то мгновенная тепловая мощность, выделяемая в проводнике, как известно из теории электричества, составляет  $I^2(t)R$ . Полное количество теплоты, выделившееся между моментами времени  $a$  и  $b$ , равно интегралу

$$\int_a^b I^2(t)R dt \quad \text{или} \quad R \int_a^b I^2(t) dt \quad \text{при } R = \text{const.}$$

Если хотим минимизировать тепловыделение рассматриваемого участка электрической цепи (а это популярная конструкторская задача), то нам нужно искать такой режим её работы, при котором достигался бы минимум выписанного интеграла, т. е. среднеквадратичного значения тока.

Отметим, что в электротехнике аналогичная величина

$$I_{\text{эфф}} = \sqrt{\frac{1}{T} \int_0^T I^2(t) dt},$$

т. е. среднеквадратичное значение периодического переменного тока за его период  $T$ , называется также *действующим* или *эффективным* значением силы тока. Именно его измеряют почти все амперметры переменного тока. ■

Как соотносится наша задача среднеквадратичного приближения функций с абстрактной постановкой из § 2.11а и данным там же методом её решения? Прежде всего, множество приближаемых функций должно быть линейным пространством с интегральным скалярным произведением (2.115). Далее, тот факт, что функция, которую ищем в качестве приближения, берётся из линейного векторного подпространства, означает, что она выражается линейно, относительно некоторых параметров, через какие-то базисные функции. Иногда эти условия выполняются тривиально, но бывают ситуации, когда они не могут быть удовлетворены.

**Пример 2.11.4** В естественных науках эволюция некоторых процессов во времени описывается дробно-линейными функциями вида

$$y = \frac{a+t}{b+t}, \quad \text{где } t \text{ — время, } a \text{ и } b \text{ — некоторые константы.} \quad (2.132)$$

В частности, таково известное уравнение Михаэлиса–Ментен, описывающее зависимость скорости реакции, катализируемой ферментом, от концентрации субстрата:

$$y = \frac{cx}{x+d}$$

с положительными параметрами  $c$  и  $d$ .

Часто требуется приблизить функциями выписанного вида реальные экспериментальные кривые, и для этого, в принципе, можно рассмотреть их среднеквадратичное приближение классом функций вида

(2.132) для различных параметров  $a, b$  и выбрать наилучшее приближение. Такое решение имеет теоретико-вероятностный смысл. Но метод, развитый в § 2.11а, напрямую непригоден для его получения, так как  $b$  входит в знаменатель дроби в (2.132), вследствие чего функции (2.132) не зависят линейно от совокупности искомых параметров  $a$  и  $b$ . Иными словами, линейное пространство функций вида (2.132) не образуют. ■

С другой стороны, класс приближаемых функций  $\mathcal{F}$  со структурой линейного пространства бывает неявно определён самой постановкой задачи. Это происходит, к примеру, когда мы работаем с непрерывным (гладкими и т. п.) функциями, для которых физический смысл имеют поточечные операции их сложения и умножения на скаляр (см. пример 2.1.1). Но и в этих случаях могут возникать вопросы при математической постановке задачи.

В некоторых задачах математического моделирования желательно работать по возможности с наиболее широким классом функций, который позволяет адекватно описывать различные явления. В частности, необходимо иметь в этом классе функции с особенностями и разрывные функции, с помощью которых могут моделироваться различные переключательные процессы. Например, у инженеров популярна *функция Хевисайда* (рис. 2.36), которая задаётся как

$$\theta(x) := \begin{cases} 0, & \text{если } x < 0, \\ 1, & \text{если } x \geq 0. \end{cases}$$

Она называется также *функцией единичного скачка* и является интегралом (в обобщённом смысле) от известной дельта-функции Дирака. Функция Хевисайда широко используется в математической теории управления и обработке сигналов.

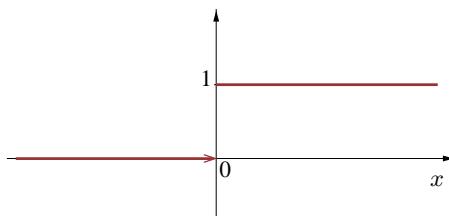


Рис. 2.36. График функции Хевисайда

Естественное решение вопроса о выборе класса приближаемых функций  $\mathcal{F}$  может состоять в том, что он берётся как множество всех функций, интегрируемых по Риману, т. е. в традиционном смысле. Известно, что сумма, произведение на число и произведение таких функций тоже интегрируемы по Риману [12, 40]. Следовательно, квадрат интегрируемой по Риману функции тоже интегрируем, и их среднеквадратичные нормы (2.116) существуют для случая единичного веса. В целом множество интегрируемых по Риману функций образует линейное векторное пространство. Кроме того, читателю должно быть известно, что такие функции могут иметь счётное число точек разрыва, т. е. они достаточно представительны и могут описывать разрывные процессы.

Определённым недостатком множества интегрируемых по Риману функций является тот факт, что оно не является полным метрическим пространством: не всякая фундаментальная (сходящаяся в себе) последовательность интегрируемых функций сходится к функции, которая тоже интегрируема по Риману. Это не вполне удобно в некоторых важных математических построениях.

Наиболее широкий запас функций для среднеквадратичного приближения получается в случае, когда интегрирование понимается в смысле Лебега [16, 32]), хотя его конструкция несколько более сложна, чем у интеграла Римана. Известно, что любая интегрируемая по Риману функция интегрируема по Лебегу, но не наоборот. При таком подходе в качестве класса приближаемых функций  $\mathcal{F}$  естественно взять множество всех вещественнозначных функций на интервале  $[a, b] \subset \mathbb{R}$ , для которых определены интегральное скалярное произведение (2.115) и вытекающие из него конструкции. Обычно требуют, чтобы норма (2.116) была конечной, т. е. для таких функций квадрат (степень 2) должен быть интегрируем на интервале  $[a, b]$  с заданным положительным весом  $\varrho(x)$ . Тогда в силу очевидного неравенства

$$2 |f(x) g(x)| \leq (f(x))^2 + (g(x))^2$$

и свойств интеграла Лебега [32] мы получим также интегрируемость произведения функций с тем же весом.

Множество функций, квадрат которых интегрируем с заданным весом на  $[a, b]$  в смысле Лебега, называют пространством  $\mathcal{L}^2[a, b]$  (см., например, [30]). Операции сложения векторов-функций и умножения на скаляр задаются в нём обычным поточечным образом. Чтобы оказаться в условиях постановки задачи из § 2.11а и воспользоваться развитым там методом решения, нужно ещё показать, что пространство  $\mathcal{L}^2[a, b]$

в самом деле является линейным векторным пространством, т. е. что умножение на скаляр и сложение функций из  $\mathcal{L}^2[a, b]$  не выводят за его пределы.<sup>25</sup> Ясно, что если  $f \in \mathcal{L}^2[a, b]$ , то для любого скаляра  $c$  функция  $cf(x)$  тоже интегрируема с квадратом на  $[a, b]$ . Далее, для  $f, g \in \mathcal{L}^2[a, b]$  можем представить

$$\begin{aligned} & \int_a^b \varrho(x) (f(x) + g(x))^2 dx = \\ & = \int_a^b \varrho(x) (f(x))^2 dx + 2 \int_a^b \varrho(x) f(x) g(x) dx + \int_a^b \varrho(x) (g(x))^2 dx, \end{aligned}$$

где каждый из интегралов в правой части равенства существует. Как следствие, сумма  $f(x) + g(x)$  также имеет интегрируемый с весом  $\varrho(x)$  квадрат, что завершает проверку линейности пространства  $\mathcal{L}^2[a, b]$ .

В теории функций вещественной переменной показывается (см., к примеру, [16, 30, 32]), что  $\mathcal{L}^2[a, b]$  как метрическое пространство обладает свойством полноты. По этой причине оно очень популярно в самых различных математических дисциплинах, от теории уравнений в частных производных до математической статистики. Отметим, что существуют и часто применяются пространства  $\mathcal{L}^p$ , состоящие из функций, у которых интегрируема  $p$ -я степень,  $p \geq 1$ .

Итак, в качестве пространства  $\mathfrak{F}$  будем рассматривать какое-либо линейное векторное пространство функций со скалярным произведением вида (2.115) и нормой (2.116). В качестве пространства  $\mathfrak{G}$ , в котором ищутся приближения для элементов из  $\mathfrak{F}$ , обычно рассматривается какое-то его конечномерное подпространство. Например, это могут быть алгебраические или тригонометрические полиномы заданной степени, суммы экспонент с разными показателями и т. п. В  $\mathfrak{G}$  задаётся некоторый базис  $\{\varphi_j(x)\}_{j=1}^m$  и среднеквадратичное приближение ищется в виде

$$g(x) = \sum_{j=1}^m c_j \varphi_j(x) \quad (2.133)$$

по технологии из § 2.11а. Как выглядит и как решается система нормальных уравнений (2.122) для определения коэффициентов  $c_j$  наилучшего приближения в связи с задачей этого раздела, т. е. когда мы приближаем вещественные функции на интервале? Это зависит как от

---

<sup>25</sup>Буква « $\mathcal{L}$ » в обозначении этого пространства связывается с именем А.Л. Лебега.

подпространства  $\mathcal{G} \subset \mathcal{F}$ , так и от базиса, выбранного в  $\mathcal{G}$ . Конкретные ситуации, которые могут здесь встретиться, рассмотрим в следующем разделе.

### 2.11ж Базисы для среднеквадратичных приближений

**Пример 2.11.5** Пусть дана задача о среднеквадратичном приближении непрерывных функций на  $[0, 1]$  с единичным весом полиномами фиксированной степени  $m$ . Тогда скалярное произведение определяется как

$$\langle f, g \rangle = \int_0^1 f(x) g(x) dx,$$

а нормой берём

$$\|f\| = \sqrt{\langle f, f \rangle} = \left( \int_0^1 (f(x))^2 dx \right)^{1/2}.$$

Соответственно, расстояние между функциями определяется как

$$\text{dist}(f, g) = \|f - g\| = \left( \int_0^1 (f(x) - g(x))^2 dx \right)^{1/2}.$$

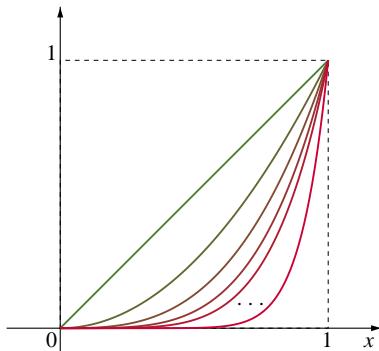


Рис. 2.37. Графики последовательных степеней переменной  $x$

Если в качестве базиса в линейном подпространстве полиномов возьмём последовательные степени

$$1, \quad x, \quad x^2, \quad \dots, \quad x^m,$$

то на месте  $(i, j)$  в матрице Грама (2.123) размера  $(m+1) \times (m+1)$  будет стоять элемент

$$\int_0^1 x^{i-1} x^{j-1} dx = \left. \frac{x^{i+j-1}}{i+j-1} \right|_0^1 = \frac{1}{i+j-1}, \quad i, j = 1, 2, \dots, m+1$$

(сдвиг показателей степени на  $(-1)$  вызван тем, что строки и столбцы матрицы нумеруются, начиная с единицы, а не с нуля, как последовательность степеней  $x$ ).

Матрица  $H = (h_{ij})$  с элементами  $h_{ij} = 1/(i+j-1)$ , имеющая вид

$$\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{m+1} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{m+2} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \cdots & \frac{1}{m+3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{m+1} & \frac{1}{m+2} & \frac{1}{m+3} & \cdots & \frac{1}{2m+1} \end{pmatrix},$$

называется *матрицей Гильберта*, и она является исключительно плохо обусловленной матрицей (см. § 3.46). Решение систем линейных уравнений с такими матрицами, необходимое для построения наилучших приближений в нашей постановке, является непростой задачей, которая очень чувствительна к влиянию погрешностей в данных и вычислениях.

Плохая обусловленность матрицы Гильберта неформально объясняется тем, что последовательные степени переменной  $x^n$  с ростом  $n$  отличаются друг от друга всё меньше и меньше (рис. 2.37). Совершенно то же происходит со строками (или столбцами) матрицы Гильберта, отличие которых с ростом номера делается всё меньшим. Поэтому хотя последовательные степени  $x^n$  в теории линейно независимы, но базис из них с ростом  $n$  всё более и более «сплющивается» и приближается к линейно зависимой системе векторов. ■

**Пример 2.11.6** Пусть  $k$  и  $l$  — неотрицательные целые числа. Тогда

в силу известных формул элементарной тригонометрии

$$\int_0^{2\pi} \sin(kx) \cos(lx) dx = \frac{1}{2} \int_0^{2\pi} (\sin((k+l)x) + \sin((k-l)x)) dx = 0$$

для любых  $k, l$ . Кроме того,

$$\int_0^{2\pi} \sin(kx) \sin(lx) dx = \frac{1}{2} \int_0^{2\pi} (\cos((k-l)x) - \cos((k+l)x)) dx = 0,$$

$$\int_0^{2\pi} \cos(kx) \cos(lx) dx = \frac{1}{2} \int_0^{2\pi} (\cos((k+l)x) + \cos((k-l)x)) dx = 0$$

и для  $k \neq l$ . Как следствие, функции вида

$$1, \quad \cos(kx), \quad \sin(kx), \quad k = 1, 2, \dots, \quad (2.134)$$

являются ортогональными на  $[0, 2\pi]$  относительно интегрального скалярного произведения (2.115) с единичным весом. Из ортогоональности, в свою очередь, следует их линейная независимость.

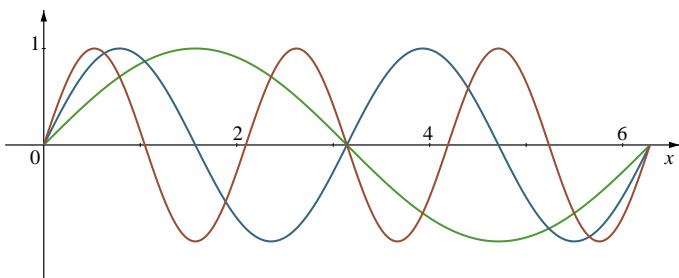


Рис. 2.38. Графики функций  $\sin x$ ,  $\sin 2x$  и  $\sin 3x$

Семейство функций (2.134) называется *тригонометрической системой функций*, и в вычислительном отношении базис из них очень хорош для построения среднеквадратичных приближений. Ясно, что вместо интервала  $[0, 2\pi]$  можно взять любой другой интервал, ширина которого равна периоду  $2\pi$ , а подходящим масштабированием из него можно получить вообще любой интервал вещественной оси.

Из рис. 2.38 видно, что поведение функций тригонометрической системы — совершенно другое, нежели у последовательных степеней на

рис. 2.37: функции тригонометрической системы «существенно отличаются» друг от друга, и это приводит к «хорошей» матрице Грама, имеющей «малую» внеdiagональную часть.

Отметим, что исторически первые ряды Фурье были построены в начале XIX века в работах Ж.Б. Фурье именно как разложения по тригонометрической системе функций (2.134). Для периодической функции  $f(x)$ , имеющей период  $2\pi$ , *тригонометрическим рядом Фурье* называется разложение

$$f(x) = \frac{a_0}{2} + \sum_k (a_k \cos(kx) + b_k \sin(kx)),$$

где

$$a_k = \frac{1}{\pi} \int_0^{2\pi} f(t) \cos(kt) dt, \quad b_k = \frac{1}{\pi} \int_0^{2\pi} f(t) \sin(kt) dt, \quad k = 0, 1, \dots,$$

— коэффициенты разложения, которые называются *коэффициентами Фурье* [12, 32]. Они равны интегральным скалярным произведениям приближаемой функции на базисные функции (2.134). ■

Рассмотренные в предшествующих примерах системы функций — последовательные степени переменной и тригонометрические функции кратных аргументов — обладают важным свойством *полноты*: любая интегрируемая с квадратом функция может быть сколь угодно точно приближена их линейными комбинациями.<sup>26</sup> Для некоторых практически важных ортогональных систем функций приходится добиваться выполнения этого свойства дополнительными средствами.

**Пример 2.11.7** Определим семейство функций  $r_k(x)$ , занумерованных неотрицательными целыми индексами  $k$ , следующим образом. Положим

$$r_0(x) = \begin{cases} 1, & \text{если } x \in [0, \frac{1}{2}[ , \\ -1, & \text{если } x \in [\frac{1}{2}, 1[ . \end{cases}$$

Продолжим функцию  $r_0(x)$  с периодом 1 на всю положительную полусось  $[0, +\infty]$ . Каждую следующую функцию  $r_k(x)$  определим как сжатие

---

<sup>26</sup> Термин «полнота» применяется также в других смыслах, в частности в отношении метрических пространств. См. § 3.36.

по горизонтальной оси предыдущей по номеру функции в два раза, так что

$$r_k(x) := r_{k-1}(2x), \quad k = 1, 2, \dots$$

Введённые выше функции  $r_k(x)$ ,  $k = 1, 2, \dots$ , называются *функциями Радемахера* по имени немецкого математика, предложившего их в 1922 году [105]. Графики первых из этих функций для интервала  $[0, 1]$  изображены на рис. 2.39. В целом функции Радемахера образуют ортогональную систему кусочно-постоянных функций на интервале  $[0, 1]$ , которые могут принимать лишь два различных значения.

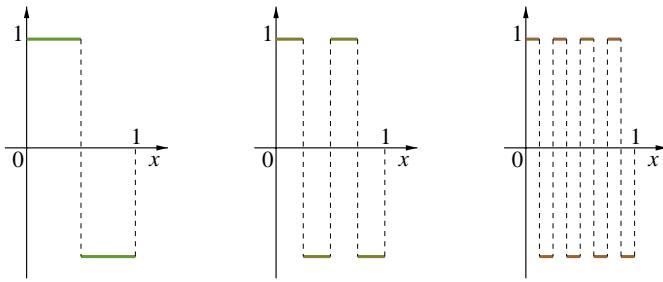


Рис. 2.39. Графики первых функций Радемахера

Существует также аналитическое задание функций Радемахера

$$r_n(x) = \operatorname{sgn}(\sin(2^{n+1}\pi x)).$$

Недостатком системы функций Радемахера является тот факт, что она не обладает свойством полноты: линейной комбинацией функций системы (даже бесконечной) нельзя представить любую достаточно произвольную функцию. Этот факт легко понять, заметив, что графики всех функций Радемахера симметричны относительно середины интервала  $[0, 1]$ , т. е. точки плоскости с координатами  $(0.5, 0)$ . Это же свойство сохраняется при масштабировании функций Радемахера. Как следствие, любая линейная комбинация функций Радемахера тоже будет обладать этим свойством, и произвольную функцию в таком виде представить нельзя.

Исправить недостаток функций Радемахера можно разными способами, и один из самых популярных приводит к так называемым функциям Уолша. По определению *функциями Уолша* называются функции, которые являются произведениями конечного числа функций Радемахера (рис. 2.40).

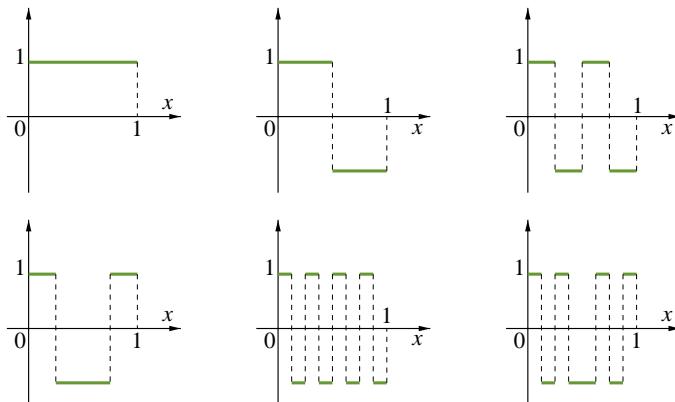


Рис. 2.40. Графики первых функций Уолша

Ясно, что функции Уолша также являются кусочно-постоянными и принимают только значения  $-1$  и  $1$ . Но теперь эти их значения расположены гораздо менее регулярно и в самом деле позволяют приближать достаточно произвольные функции. Нетрудно также понять, что функции Уолша образуют счётное семейство, и оно может быть занумеровано различными способами (см. подробности в [54]). Одной из наиболее популярных является нумерация, в которой номер функции соответствует так называемому коду Грея номера функции Уолша.

Особенностью функций Радемахера, функций Уолша и тригонометрических функций (из предыдущего примера) является сравнительная простота их получения в электротехнике, в электронных и радиотехнических устройствах. Систему функций Уолша часто используют для анализа и синтеза импульсных электрических сигналов конечной длительности [54]. Она также применяется при компьютерной обработке информации. ■

Более детальный теоретический анализ и практический опыт показывают, что в методе наименьших квадратов в качестве базиса  $\varphi_1, \varphi_2, \dots, \varphi_m$  линейного подпространства  $\mathcal{G} \subset \mathcal{F}$  имеет смысл брать системы векторов, хотя бы «приближённо ортогональные» или даже «не слишком далёкие» от ортогональных. Это служит гарантией «разумной малости» внедиагональных элементов матрицы Грама и, как следствие, её не слишком плохой обусловленности.

Нередко при поиске среднеквадратичных приближений форма приближающей функции (2.133), которая берётся в виде линейной комбинации базисных, не подходит по тем или иным причинам (именно таков пример 2.10.4). Тогда приходится прибегать к *нелинейному методу наименьших квадратов*, в котором приближающая функция  $g(x)$  выражается нелинейным образом через параметры  $c_j$ ,  $j = 1, 2, \dots, m$ . Соответственно, минимизация среднеквадратичного отклонения  $f$  от  $g$  уже не сводится к решению нормальной системы линейных алгебраических уравнений (2.122), и для нахождения минимума нам нужно применять численные методы оптимизации. Обсуждение этого круга вопросов и дальнейшие ссылки можно найти в книге [56].

## 2.12 Полиномы Лежандра

### 2.12а Мотивация и определение

Выбор хорошего, т. е. ортогонального или почти ортогонального, базиса для среднеквадратичного приближения функций, как показывают примеры 2.11.5 и 2.11.6 из предшествующего раздела, является очень важной задачей. Чтобы решить её и построить необходимый базис из функций заданного вида, можно воспользоваться известным из курса линейной алгебры процессом ортогонализации Грама–Шмидта или его модификациями (см. § 3.8).

Напомним, что по конечной линейно независимой системе векторов  $v_1, v_2, \dots, v_n$  этот процесс строит ортогональный базис  $q_1, q_2, \dots, q_n$  для линейной оболочки векторов  $v_1, v_2, \dots, v_n$ . Его расчётные формулы таковы:

$$q_1 \leftarrow v_1, \tag{2.135}$$

$$q_k \leftarrow v_k - \sum_{i=1}^{k-1} \frac{\langle v_k, q_i \rangle}{\langle q_i, q_i \rangle} q_i, \quad k = 2, \dots, n. \tag{2.136}$$

Фактически на каждом шаге этого процесса строится перпендикуляр к линейной оболочке векторов, обработанных на предыдущих шагах. Иногда получающийся ортогональный базис дополнительно нормируют, т. е. масштабируют его векторы к единичной норме.

В задаче среднеквадратичного приближения функций из § 2.11а ортогонализуемые элементы линейного пространства — это функции, а

их скалярное произведение  $\langle \cdot, \cdot \rangle$  — это интеграл (2.115). По этой причине процесс ортогонализации (2.135)–(2.136) довольно трудоёмок, а конкретный вид ортогональных функций, которые получатся в результате, зависит, во-первых, от интервала  $[a, b]$ , для которого рассматривается интегральное скалярное произведение (2.115), и, во-вторых, от весовой функции  $\varrho(x)$ .

Для частного случая единичного веса, когда  $\varrho(x) = 1$ , мы можем существенно облегчить свою задачу, если найдём семейство ортогональных функций для какого-нибудь одного интервала  $[\alpha, \beta]$ , который выбран в качестве «канонического». Для любого другого интервала  $[a, b]$  затем воспользуемся формулой линейной замены переменной  $y = rx + s$  со специально подобранными константами  $r, s \in \mathbb{R}$ ,  $r \neq 0$ . Тогда  $x = (y - s)/r$  и для  $a = r\alpha + s$ ,  $b = r\beta + s$  имеем равенство

$$\int_{\alpha}^{\beta} f(x) g(x) dx = \frac{1}{r} \int_a^b f\left(\frac{y-s}{r}\right) g\left(\frac{y-s}{r}\right) dy,$$

вытекающее из формулы замены переменных в определённом интеграле. Поэтому равный нулю интеграл по каноническому интервалу  $[\alpha, \beta]$  останется нулевым и при линейной замене переменных. Как следствие, получающиеся при такой замене функции  $f((y - s)/r)$  и  $g((y - s)/r)$  от переменной  $y$  будут ортогональны на  $[a, b]$ .

Рассмотрим среднеквадратичное приближение функций полиномами. В этом случае в качестве канонического интервала обычно берётся  $[-1, 1]$ , а любой другой интервал  $[a, b]$  можно получить из него с помощью замены переменных (2.46),

$$y = \frac{1}{2}(b - a)x + \frac{1}{2}(a + b),$$

которая уже встречалась нам в § 2.3б. Ясно, что переменная  $y$  пробегает интервал  $[a, b]$ , если  $x \in [-1, 1]$ . Обратное преобразование даётся формулой (2.47),

$$x = \frac{1}{b-a}(2y - (a+b)),$$

которая позволяет строить полиномы, ортогональные в смысле интегрального скалярного произведения, для любого интервала  $[a, b] \subset \mathbb{R}$ , зная их для интервала  $[-1, 1]$ .

*Полиномами Лежандра* называют семейство алгебраических полиномов, занумерованных неотрицательными целыми числами и таких, что  $n$ -й полином имеет степень  $n$ , а полиномы с различными номерами

ортогональны относительно интегрального скалярного произведения (2.115) с единичным весом на интервале  $[-1, 1]$ . Эти полиномы были введены в широкий оборот в 1785 году французским математиком А.-М. Лежандром. Иногда их называют *сферическими полиномами*, так как они естественно возникают при нахождении решений некоторых задач математической физики в сферических координатах. Именно в таком качестве их использовал сам А.-М. Лежандр.

Из общей теории скалярного произведения в линейных векторных пространствах следует, что такие полиномы существуют и единственны с точностью до постоянного множителя. В частности, «теорема о перпендикуляре», рассмотренная в § 2.11в, утверждает, что для любого вектора и конечномерного подпространства существуют и единственны перпендикуляр и ортогональная проекция этого вектора на подпространство. Тогда  $k$ -й полином Лежандра — это перпендикуляр на линейное подпространство всех алгебраических полиномов степени не более  $k - 1$ .

Если формулы ортогонализации Грама–Шмидта (2.135)–(2.136) с интегральным скалярным произведением (2.115) на интервале  $[-1, 1]$  применить к степеням  $1, x, x^2, x^3, \dots$ , то получим явные выражения для первых полиномов Лежандра:

$$1, \quad x, \quad x^2 - \frac{1}{3}, \quad x^3 - \frac{3}{5}x, \quad \dots \quad (2.137)$$

(две первых степени оказываются изначально ортогональными). Необходимое нормирование полиномов Лежандра обычно выполняют различными способами, наиболее подходящими для той или иной задачи.

## 2.126 Формула Родрига

Удобное альтернативное представление для полиномов Лежандра даёт *формула Родрига*

$$L_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n, \quad n = 0, 1, 2, \dots \quad (2.138)$$

Очевидно, что функция  $L_n(x)$ , определяемая этой формулой, является алгебраическим полиномом  $n$ -й степени со старшим коэффициентом, не равным нулю, так как при  $n$ -кратном дифференцировании полинома  $(x^2 - 1)^n = x^{2n} - nx^{2(n-1)} + \dots + (-1)^n$  степень понижается в точности на  $n$ . Коэффициент  $1/(2^n n!)$  перед производной в (2.138) взят с той

целью, чтобы удовлетворить условию  $L_n(1) = 1$ . Кроме того, нетрудно показать, что

$$L_n(-1) = (-1)^n, \quad n = 1, 2, \dots$$

(см. подробности в [29, 74]).

Всюду далее посредством  $L_n(x)$  мы будем обозначать полиномы Лежандра, определяемые формулой (2.138). Но для её обоснования необходимо доказать

**Предложение 2.12.1** *Полиномы  $L_n(x)$ ,  $n = 0, 1, \dots$ , задаваемые формулой Родрига (2.138), ортогональны друг другу в смысле интегрального скалярного произведения на  $[-1, 1]$  с единичным весом. Более точно*

$$\int_{-1}^1 L_m(x) L_n(x) dx = \begin{cases} 0, & \text{если } m \neq n, \\ \frac{2}{2n+1}, & \text{если } m = n. \end{cases}$$

**Доказательство.** Обозначая

$$\psi(x) = (x^2 - 1)^n,$$

можно заметить, что для производных порядка  $k = 0, 1, \dots, n-1$  от функции  $\psi(x)$  справедливо равенство

$$\psi^{(k)}(x) = \frac{d^k}{dx^k} (x^2 - 1)^n = 0 \quad \text{при } x = \pm 1.$$

Это следует из зануления множителей  $(x^2 - 1)$ , присутствующих во всех слагаемых выражений для  $\psi^{(k)}(x)$ ,  $k = 0, 1, \dots, n-1$ . Кроме того, в силу формулы Родрига (2.138)

$$L_n(x) = \frac{1}{2^n n!} \psi^{(n)}(x), \quad n = 0, 1, 2, \dots$$

Поэтому, если  $Q(x)$  является  $n$ -кратно непрерывно дифференцируемой функцией на  $[-1, 1]$ , то, последовательно применяя  $n$  раз формулу ин-

тегрирования по частям, получим

$$\begin{aligned}
 \int_{-1}^1 Q(x) L_n(x) dx &= \frac{1}{2^n n!} \int_{-1}^1 Q(x) \psi^{(n)}(x) dx = \\
 &= \frac{1}{2^n n!} \int_{-1}^1 Q(x) d(\psi^{(n-1)}(x)) = \\
 &= \frac{1}{2^n n!} \left( \left( Q(x) \psi^{(n-1)}(x) \right) \Big|_{-1}^1 - \int_{-1}^1 Q'(x) \psi^{(n-1)}(x) dx \right) = \\
 &= -\frac{1}{2^n n!} \int_{-1}^1 Q'(x) \psi^{(n-1)}(x) dx = \\
 &= \dots = \\
 &= (-1)^n \frac{1}{2^n n!} \int_{-1}^1 Q^{(n)}(x) \psi(x) dx. \tag{2.139}
 \end{aligned}$$

Если  $Q(x)$  — полином степени меньше  $n$ , то его  $n$ -я производная  $Q^{(n)}(x)$  равна тождественному нулю, а потому из полученной формулы следует

$$\int_{-1}^1 Q(x) L_n(x) dx = 0.$$

В частности, это верно и в случае, когда вместо  $Q(x)$  берётся полином  $L_m(x)$  степени  $m$ , меньшей  $n$ , что доказывает ортогональность этих полиномов с разными номерами.

Найдём теперь скалярное произведение полинома Лежандра на самого себя. Для этой цели в предшествующих рассуждениях положим  $Q(x) = L_n(x)$  и заметим, что тогда

$$Q^{(n)}(x) = \left( \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n \right)^{(n)} = \frac{1}{2^n n!} \frac{d^{2n}}{dx^{2n}} (x^2 - 1)^n = \frac{(2n)!}{2^n n!}.$$

По этой причине из (2.139) следует

$$\begin{aligned} \int_{-1}^1 L_n(x) L_n(x) dx &= (-1)^n \frac{(2n)!}{2^{2n}(n!)^2} \int_{-1}^1 \psi(x) dx = \\ &= (-1)^n \frac{(2n)!}{2^{2n}(n!)^2} \int_{-1}^1 (x^2 - 1)^n dx. \end{aligned} \quad (2.140)$$

С другой стороны, последовательно интегрируя по частям  $n$  раз, получим

$$\begin{aligned} \int_{-1}^1 (x^2 - 1)^n dx &= \int_{-1}^1 (x - 1)^n (x + 1)^n dx = \\ &= \frac{1}{n+1} \int_{-1}^1 (x - 1)^n d((x + 1)^{n+1}) = \\ &= \frac{1}{n+1} \left( (x - 1)^n (x + 1)^{n+1} \Big|_{-1}^1 - n \int_{-1}^1 (x - 1)^{n-1} (x + 1)^{n+1} dx \right) = \\ &= \frac{(-1)n}{n+1} \int_{-1}^1 (x - 1)^{n-1} (x + 1)^{n+1} dx = \\ &= \dots = \\ &= \frac{(-1)^n n!}{(n+1) \cdot \dots \cdot 2n} \int_{-1}^1 (x - 1)^0 (x + 1)^{2n} dx = \\ &= \frac{(-1)^n (n!)^2}{(2n)!} \frac{(x + 1)^{2n+1}}{2n+1} \Big|_{-1}^1 = \frac{(-1)^n (n!)^2}{(2n)!} \frac{2^{2n+1}}{2n+1}. \end{aligned}$$

Комбинируя полученный результат с (2.140), будем иметь

$$\int_{-1}^1 L_n(x) L_n(x) dx = \frac{2}{2n+1},$$

что завершает доказательство предложения. ■

Выпишем первые полиномы Лежандра, как они даются формулой Родрига (2.138):

$$L_0(x) = 1,$$

$$L_1(x) = x,$$

$$L_2(x) = \frac{1}{2}(3x^2 - 1),$$

$$L_3(x) = \frac{1}{2}(5x^3 - 3x), \quad (2.141)$$

$$L_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3),$$

$$L_5(x) = \frac{1}{8}(63x^5 - 70x^3 + 15x),$$

... .

Эти полиномы с точностью до множителя совпадают с результатами ортогонализации Грама–Шмидта (2.137). Графики полиномов (2.141) изображены на рис. 2.41, и они похожи на графики полиномов Чебышёва.

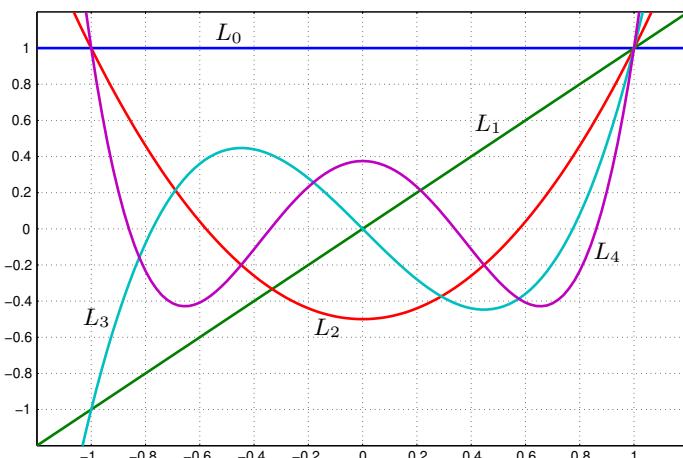


Рис. 2.41. Графики первых полиномов Лежандра на интервале  $[-1, 1]$

## 2.12в Основные свойства полиномов Лежандра

Аналогично полиномам Чебышёва нули полиномов Лежандра тоже сгущаются к концам интервала  $[-1, 1]$ . Кроме того, нули полинома Лежандра  $L_n(x)$  перемежаются с нулями полинома  $L_{n+1}(x)$ . Наконец, справедливо рекуррентное представление

$$(n+1)L_{n+1}(x) = (2n+1)xL_n(x) - nL_{n-1}(x).$$

Его можно записать также в виде

$$L_{n+1}(x) = \frac{2n+1}{n+1} xL_n(x) - \frac{n}{n+1} L_{n-1}(x), \quad (2.142)$$

который совершенно аналогичен по форме рекуррентному представлению полиномов Чебышёва (2.39). Детальные доказательства этих свойств можно найти, например, в книгах [29, 74].

Из (2.142) индукцией нетрудно показать, что полиномы Лежандра с чётными номерами являются чётными функциями, а с нечётными номерами — нечётными функциями. Кроме того, рекуррентные формулы дают практически удобный способ вычисления значений полиномов Лежандра. В их явном представлении (2.141) коэффициенты растут экспоненциально быстро в зависимости от номера полинома, и, как следствие, прямые вычисления с ними могут дать большую погрешность.

В одном существенном моменте полиномы Лежандра всё же отличаются от полиномов Чебышёва: абсолютные значения локальных минимумов и максимумов у полиномов Лежандра различны и не могут быть сделаны одинаковыми ни при каком масштабировании. Тем не менее глубокое сходство полиномов Лежандра и полиномов Чебышёва существует. Для его формулировки напомним, что *простыми нулями* (или *нулями кратности 1*) непрерывно дифференцируемой функции  $f(x)$  называются точки, в которых  $f(x) = 0$ , но  $f'(x) \neq 0$ , т. е. в которых функция зануляется, а её производная — нет. В простых нулях график функции пересекает ось абсцисс под ненулевым углом.

**Предложение 2.12.2** Все нули полиномов Лежандра  $L_n(x)$  вещественные, простые и находятся на интервале  $[-1, 1]$ .

**Доказательство.** Предположим, что среди нулей полинома  $L_n(x)$ , лежащих на  $[-1, 1]$ , имеется  $s$  штук различных нулей  $\theta_1, \theta_2, \dots, \theta_s$  нечёт-

ной кратности  $\alpha_1, \alpha_2, \dots, \alpha_s$  соответственно. Поэтому

$$L_n(x) = (x - \theta_1)^{\alpha_1} (x - \theta_2)^{\alpha_2} \cdots (x - \theta_s)^{\alpha_s} \gamma(x),$$

где в полиноме  $\gamma(x)$  присутствуют нули  $L_n(x)$ , не лежащие на  $[-1, 1]$ , а также те нули  $L_n(x)$  из  $[-1, 1]$ , которые имеют чётную кратность. Таким образом,  $\gamma(x)$  уже не меняет знака на интервале  $[-1, 1]$ . Ясно, что  $s \leq n$ , и наша задача — установить равенство  $s = n$ .

Рассмотрим интеграл

$$\begin{aligned} \mathcal{I} &= \int_{-1}^1 L_n(x) (x - \theta_1)(x - \theta_2) \cdots (x - \theta_s) dx = \\ &= \int_{-1}^1 (x - \theta_1)^{\alpha_1+1} (x - \theta_2)^{\alpha_2+1} \cdots (x - \theta_s)^{\alpha_s+1} \gamma(x) dx. \end{aligned}$$

Теперь  $\alpha_1+1, \alpha_2+1, \dots, \alpha_s+1$  — чётные числа, так что подинтегральное выражение не меняет знак на  $[-1, 1]$ . Это выражение равно нулю лишь в конечном множестве точек, и потому определено  $\mathcal{I} \neq 0$ .

С другой стороны, выражение для  $\mathcal{I}$  есть интегральное скалярное произведение на  $[-1, 1]$  полинома  $L_n(x)$  на полином  $(x - \theta_1)(x - \theta_2) \cdots (x - \theta_s)$  степени не более  $n-1$ , если выполнено условие  $s < n$ . Следовательно, в силу свойств полиномов Лежандра должно быть  $\mathcal{I} = 0$ .

Полученное противоречие может быть снято только в случае  $s = n$ , т. е. когда равенство  $\mathcal{I} = 0$  невозможно. При этом все нули полинома  $L_n(x)$  различны, просты и лежат на интервале  $[-1, 1]$ . ■

Отметим, что проведённое доказательство непосредственно переносится на интегральные скалярные произведения вида (2.115) с произвольными весовыми функциями  $\varrho(x)$ . Кроме того, нигде не использовался в явном виде тот факт, что интервал интегрирования есть  $[-1, 1]$ . Фактически это доказательство годится даже для бесконечных пределов интегрирования. Оно показывает, что нули любых полиномов, ортогональных относительно интегрального скалярного произведения, — вещественные и простые.

Введём так называемые *приведённые полиномы Лежандра*  $\tilde{L}_n(x)$ , старший коэффициент у которых равен единице. Чтобы получить явное представление для  $\tilde{L}_n(x)$ , в формуле Родрига (2.138) достаточно поставить перед  $n$ -й производной множитель, который компенсирует

коэффициенты при старшем члене полинома  $(x^2 - 1)^n$ , возникающие в процессе  $n$ -кратного дифференцирования. Тогда

$$\begin{aligned}\tilde{L}_n(x) &= \frac{1}{2n(2n-1)\cdots(n+1)} \frac{d^n}{dx^n} (x^2 - 1)^n = \\ &= \frac{n!}{(2n)!} \frac{d^n}{dx^n} (x^2 - 1)^n, \quad n = 1, 2, \dots\end{aligned}\tag{2.143}$$

Как и исходная формула Родрига, выражение после второго равенства имеет смысл при  $n = 0$ , если под производной нулевого порядка от функции понимать её саму. Из (2.143) и формулы Родрига (2.138) следует также, что

$$\tilde{L}_n(x) = \frac{2^n (n!)^2}{(2n)!} L_n(x).\tag{2.144}$$

**Предложение 2.12.3** *Среди всех полиномов степени  $n$ ,  $n \geq 1$ , со старшим коэффициентом, равным 1, полином  $\tilde{L}_n(x)$  имеет на интервале  $[-1, 1]$  наименьшее среднеквадратичное отклонение от нуля. Иными словами, если  $Q_n(x)$  — полином степени  $n$  со старшим коэффициентом 1, то*

$$\int_{-1}^1 (Q_n(x))^2 dx \geq \int_{-1}^1 (\tilde{L}_n(x))^2 dx.\tag{2.145}$$

**Доказательство.** Если  $Q_n(x) = x^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0$ , то для отыскания наименьшего значения выражения

$$\begin{aligned}\mathcal{J}(a_0, a_1, \dots, a_{n-1}) &= \int_{-1}^1 (Q_n(x))^2 dx = \\ &= \int_{-1}^1 (x^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0)^2 dx\end{aligned}\tag{2.146}$$

продифференцируем его по переменным  $a_0, a_1, \dots, a_{n-1}$  и приравняем полученные производные к нулю. В данном случае дифференцирование интеграла по параметру, от которого зависит подинтегральная функция, сводится к взятию интеграла от её производной [12, 40], и

поэтому имеем в результате

$$\begin{aligned} \frac{\partial \mathcal{J}}{\partial a_k} &= \int_{-1}^1 2(x^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0) x^k dx = \\ &= 2 \int_{-1}^1 Q_n(x) x^k dx = 0, \quad k = 0, 1, \dots, n-1. \end{aligned} \quad (2.147)$$

То, что в точке, удовлетворяющей условиям (2.147), в самом деле достигается минимум, следует из рассмотрения матрицы вторых производных (гессиана) функции  $\mathcal{J}(a_0, a_1, \dots, a_{n-1})$ , образованной элементами

$$\frac{\partial^2 \mathcal{J}}{\partial a_k \partial a_l} = 2 \int_{-1}^1 x^k x^l dx.$$

Интеграл в правой части выписанного равенства — это не что иное, как удвоенное интегральное скалярное произведение на  $[-1, 1]$  с единичным весом функций  $x^k$  и  $x^l$ . Получающаяся матрица Грама положительно определена в силу линейной независимости степеней  $x^k$ ,  $k = 0, 1, \dots, n-1$ .

Но условия (2.147) означают, что полином  $Q_n(x)$  ортогонален всем полиномам меньшей степени. Следовательно, при минимальном значении интеграла (2.146) полином  $Q_n(x)$  обязан совпадать с приведённым полиномом Лежандра  $\tilde{L}_n(x)$ . ■

Если необходимо построить полином, который имеет наименьшее среднеквадратичное отклонение от нуля на произвольном интервале  $[a, b]$ , а не на  $[-1, 1]$ , то можно воспользоваться линейной заменой переменной (2.46)–(2.47) и затем необходимым масштабированием, аналогично тому, как это было сделано в задаче интерполяции для полиномов Чебышёва в § 2.3б. Обоснование этого способа следует из того, что при линейной замене переменной, как мы выяснили в § 2.12а, из полиномов Лежандра получаются полиномы, ортогональные на  $[a, b]$  с единичным весом.

Из предложения 2.12.3 вытекают также интересные следствия в отношении задачи алгебраической интерполяции, рассмотренной в § 2.2–2.5. Из формулы (2.30) для остаточного члена алгебраической интерполяции

$$R_n(f, x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!} \cdot \omega_n(x)$$

следует, что в условиях, когда  $n+1$ -я производная  $f^{(n+1)}$  изменяется не слишком сильно, основной вклад в погрешность вносится полиномом  $\omega_n(x) = (x - x_0)(x - x_1) \dots (x - x_n)$ , который имеет единичный старший коэффициент. Для достижения наименьшей погрешности алгебраической интерполяции в среднеквадратичном смысле узлы интерполяции следует брать нулями соответствующего полинома Лежандра или же полинома, который получается из него линейным преобразованием на необходимый нам интервал.

Помимо полиномов Лежандра существуют и другие семейства ортогональных полиномов, широко используемые в теории и практических вычислениях. В частности, полиномы Чебышёва  $T_n(x)$  являются ортогональными относительно интегрального скалярного произведения (2.115) на интервале  $[-1, 1]$  с весовой функцией  $\varrho(x) = 1/\sqrt{1-x^2}$ , т. е.

$$\int_{-1}^1 \frac{T_m(x) T_n(x)}{\sqrt{1-x^2}} dx = \begin{cases} 0, & \text{если } m \neq n, \\ \pi/2, & \text{если } m = n \neq 0, \\ \pi, & \text{если } m = n = 0. \end{cases}$$

Действительно, подстановка  $x = \cos \theta$  даёт  $dx = -\sin \theta d\theta$  и  $\sqrt{1-x^2} = \sin \theta$ . Из тригонометрического представления полиномов Чебышёва поэтому получаем

$$\int_{-1}^1 \frac{T_m(x) T_n(x)}{\sqrt{1-x^2}} dx = \int_0^\pi \cos(m\theta) \cos(n\theta) d\theta,$$

откуда следует требуемое.

Полиномы Лежандра и полиномы Чебышёва являются частными случаями более общей конструкции — *полиномов Якоби*, обладающих многими красивыми и полезными свойствами. Для них, в частности, справедливо обобщение формулы Родрига и т. п. [29, 53, 87, 91].

Часто возникает необходимость воспользоваться ортогональными полиномами на бесконечных интервалах  $[0, +\infty]$  или даже  $[-\infty, \infty]$ . Естественно, единичный вес  $\varrho(x) = 1$  тут малопригоден, так как с ним интегралы от алгебраических полиномов по бесконечным интервалам окажутся расходящимися. Полиномы, ортогональные на интервалах  $[0, +\infty]$  и  $[-\infty, \infty]$  с быстроубывающими весами  $e^{-x}$  и  $e^{-x^2}$ , называются *полиномами Лагерра* (иногда также *полиномами Сонина–Лагерра* или *полиномами Сонина*) и *полиномами Эрмита* соответственно.<sup>27</sup> Они

---

<sup>27</sup>Иногда их называют также полиномами Чебышёва–Лагерра и Чебышёва–Эрмита (см., к примеру, [53, 91]), поскольку они были известны ещё П.Л.Чебышёву.

тоже находят многообразные применения в задачах приближения, и более подробные сведения на эту тему читатель может почерпнуть в книгах [29, 91].

В § 2.11а и § 2.12 мы изучали главным образом непрерывную задачу среднеквадратичного приближения, в которой рассматривались функции от непрерывного аргумента, а расстояние между функциями определялось с помощью интегральной метрики (2.117). Но дискретная задача наилучшего среднеквадратичного приближения, в которой рассматриваются функции на некоторой сетке, а расстояние между ними задаётся с помощью взвешенной евклидовой метрики (2.114), тоже важна и востребована на практике. Для её решения удобно воспользоваться ортогональными полиномами дискретной переменной, теория которых изложена, к примеру, в книге [81].

## 2.13 Численное интегрирование

### 2.13а Постановка и обсуждение задачи

Задача вычисления определённого интеграла

$$\int_a^b f(x) dx \quad (2.148)$$

является одной из важнейших математических задач, к которой сводится большое количество различных вопросов теории и практики. Это нахождение площадей криволинейных фигур, центров тяжести и моментов инерции тел, работы переменной силы и другие механические, физические, химические задачи. В математическом анализе обосновывается *формула Ньютона–Лейбница*

$$\int_a^b f(x) dx = F(b) - F(a), \quad (2.149)$$

где  $F(x)$  — первообразная для функции  $f(x)$ , т. е. такая, что  $F'(x) = f(x)$ . Эта формула даёт удобный способ вычисления интегралов, который в значительной степени удовлетворяет потребности решения подобных задач. Тем не менее, возникают ситуации, когда для вычисления интеграла (2.148) требуются другие подходы.

Они необходимы в случае, когда первообразная для интегрируемой функции не выражается через известные функции, элементарные или специальные. Примеры таких функций —  $e^{-x^2}$ ,  $e^x/x$ ,  $(\sin x)/x$  и т. д.

Далее, даже если эта первообразная может быть найдена в конечном виде, её вычисление не всегда осуществляется просто (длинное и неустойчивое к ошибкам округления выражение и т. п.). Наконец, подинтегральная функция  $f(x)$  нередко задаётся не аналитической формулой, а таблично, т. е. своими значениями в дискретном наборе точек, либо алгоритмически, т. е. с помощью какой-либо программы.

Все эти причины вызывают необходимость развития численных методов для нахождения определённых интегралов. Соответственно, задачей *численного интегрирования* называют задачу нахождения определённого интеграла (2.148) на основе знания значений функции  $f(x)$  на некоторых аргументах, без привлечения её первообразных и использования формулы Ньютона–Лейбница (2.149).

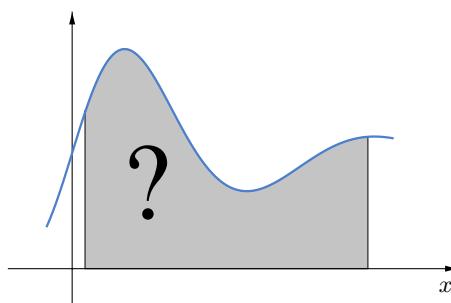


Рис. 2.42. Вычисление определённого интеграла необходимо при нахождении площадей фигур с криволинейными границами

Для нахождения интегралов наибольшее распространение в вычислительной практике получили формулы вида

$$\int_a^b f(x) dx \approx \sum_{k=0}^n c_k f(x_k), \quad (2.150)$$

где  $c_k$  — некоторые постоянные коэффициенты,  $x_k$  — точки из интервала интегрирования  $[a, b]$ ,  $k = 0, 1, \dots, n$ . Такие формулы называются *квадратурными формулами*, их коэффициенты  $c_k$  — это *весовые коэффициенты* или просто *весы* квадратурной формулы, а точки  $x_k$  — её

*узлы.* В многомерном случае аналогичные приближённые равенства

$$\int_D f(x) dx \approx \sum_{k=0}^n c_k f(x_k),$$

где  $x_k \in D \subset \mathbb{R}^m$ ,  $D$  — область в  $\mathbb{R}^m$ ,  $m \geq 2$ ,

называют *кубатурными формулами*.<sup>28</sup> Нередко узлы и веса квадратурной или кубатурной формулы нумеруют с единицы, а не с нуля. Естественное условие принадлежности узлов  $x_k$  области интегрирования вызвано тем, что за её пределами подинтегральная функция может быть просто не определена.

Помимо формул вида (2.150) применяются также квадратурные формулы, использующие значения производных интегрируемой функции в узлах [31, 74]). Мы не будем систематически рассматривать их в этом курсе (за исключением примера в § 2.16).

Тот факт, что квадратурные и кубатурные формулы являются линейными выражениями от значений интегрируемой функции в узлах, объясняется линейным характером зависимости самого интеграла от подинтегральной функции. С другой стороны, квадратурные формулы можно рассматривать как обобщения интегральных сумм Римана (через которые интеграл Римана и определяется [12, 40]). Так, простейшие составные квадратурные формулы прямоугольников просто совпадают с этими интегральными суммами.

Как и ранее, совокупность узлов  $x_0, x_1, \dots, x_n$  квадратурной (кубатурной) формулы называют *сеткой*. Разность

$$R(f) = \int_a^b f(x) dx - \sum_{k=0}^n c_k f(x_k)$$

называется *погрешностью квадратурной формулы* или её *остаточным членом*. Это число, зависящее от подинтегральной функции  $f$ , в отличие от остаточного члена интерполяции, который является ещё функцией точки (см. § 2.2e).

Если для некоторой функции  $f$  или же для целого класса функций

---

<sup>28</sup> «Квадратура» в оригинальном смысле, восходящем ещё к античности, означала построение квадрата, равновеликого заданной фигуре. Но в эпоху Возрождения этот термин стал означать вычисление площадей фигур. Аналогично с «кубатурой».

$\mathcal{F} \ni f$  имеет место точное равенство

$$\int_a^b f(x) dx = \sum_{k=0}^n c_k f(x_k),$$

то будем говорить, что квадратурная формула *точна* (является точной) на  $f$  или для класса функций  $\mathcal{F}$ . То, насколько широким является класс функций, на котором точна рассматриваемая формула, может служить косвенным признаком её точности вообще. Очень часто в качестве класса «пробных функций»  $\mathcal{F}$ , для которых исследуется совпадение результата квадратурной формулы и искомого интеграла, берут алгебраические полиномы. В этой связи полезно

**Определение 2.13.1** Алгебраической степенью точности квадратурной формулы называют наибольшую степень алгебраических полиномов, для которых эта квадратурная формула является точной.

Соответственно, с учётом специфики задачи из двух квадратурных формул более предпочтительной можно считать ту, которая имеет большую алгебраическую степень точности. Неформальным обоснованием этого критерия служит тот факт, что с помощью полиномов более высокой степени можно получать более точные приближения функций, как локально (с помощью формулы Тейлора), так и глобально (к примеру, с помощью разложения по полиномам Чебышёва или Лежандра).

Рассмотрим теперь влияние погрешностей реальных вычислений на ответ, получаемый с помощью квадратурных формул. Предположим, что значения  $f(x_k)$  интегрируемой функции в узлах  $x_k$  вычисляются неточно, с погрешностями  $\delta_k$ . Тогда по квадратурной формуле получим

$$\sum_{k=0}^n c_k (f(x_k) + \delta_k) = \sum_{k=0}^n c_k f(x_k) + \sum_{k=0}^n c_k \delta_k.$$

Если для всех  $k = 0, 1, \dots, n$  знаки погрешностей  $\delta_k$  совпадают со знаками весов  $c_k$ , то общая абсолютная погрешность результата, полученного по квадратурной формуле, становится равной  $\sum_k |c_k| |\delta_k|$ , причём

$$\sum_{k=0}^n |c_k| |\delta_k| \leq \max_{0 \leq k \leq n} |\delta_k| \sum_{k=0}^n |c_k|$$

и оценка справа, очевидно, достижима. Получается, что величину

$$\sum_{k=0}^n |c_k| \quad (2.151)$$

— сумму модулей весов квадратурной формулы — можно рассматривать как коэффициент усиления погрешности при вычислениях с этой формулой.

Далее мы узнаем, что для большинства популярных квадратурных формул сумма весовых коэффициентов равна ширине интервала интегрирования (см. следствие к теореме 2.13.1, стр. 266). По этой причине, если мы хотим организовать вычисления по квадратурной формуле наиболее устойчивым образом, все весовые коэффициенты  $c_k$  должны иметь один знак, т. е. быть положительными. Именно тогда при прочих равных условиях минимальна сумма модулей весов (2.151) и возможное усиление погрешностей вычислений.

Сказанному можно придать и другой смысл: в случае интегрирования функций, принимающих значения одного знака, использование квадратурных формул только с положительными весами позволяет избежать потери точности при вычитании (см. § 1.3), которая происходит в формуле, где присутствуют положительные и отрицательные веса.

## 2.136 Простейшие квадратурные формулы. Формулы Ньютона–Котеса

Простейший способ построения квадратурных формул — замена подинтегральной функции  $f(x)$  на интервале интегрирования  $[a, b]$  на «более простую», легче интегрируемую функцию, которая интерполирует или приближает  $f(x)$  по заданным узлам  $x_0, x_1, \dots, x_n$ . Если для нахождения функции  $g(x)$ , близкой к  $f(x)$ , используются линейные методы интерполяции или приближения, то получаем общее представление

$$f(x) \approx g(x) = \sum_{k=0}^n f(x_k) \gamma_k(x),$$

где  $\gamma_k(x)$  — некоторые функции. Интегрирование этого приближённого равенства даёт приближённое выражение для определённого интеграла

$$\int_a^b f(x) dx \approx \sum_{k=0}^n f(x_k) \int_a^b \gamma_k(x) dx.$$

Оно и является квадратурной формулой с узлами  $x_0, x_1, \dots, x_n$  и весами  $c_k = \int \gamma_k(x) dx, k = 0, 1, \dots, n$ .

В случае, когда подинтегральная функция  $f(x)$  заменяется интерполянтом и все рассматриваемые узлы — простые, говорят о квадратурных формулах интерполяционного типа, или, что равносильно, об *интерполяционных квадратурных формулах*. Наиболее часто подинтегральную функцию интерполируют алгебраическими полиномами, и в нашем курсе мы будем рассматривать главным образом именно такие интерполяционные квадратурные формулы.

Популярность и развитость интерполяционных квадратурных формул объясняется их практичесностью: в большинстве реальных задач, требующих вычисления интегралов, сами подинтегральные функции задаются лишь набором своих значений в ряде точек-узлов. Таким образом, интерполяционные квадратурные формулы неявно выполняют ещё и работу по восстановлению интегрируемой функции, что чрезвычайно удобно на практике.

*Формулами Ньютона–Котеса* называют интерполяционные квадратурные формулы, которые получены с помощью алгебраической интерполяции подинтегральной функции на равномерной сетке с простыми узлами. В зависимости от того, включаются ли концы интервала интегрирования  $[a, b]$  в множество узлов квадратурной формулы или нет, различают формулы Ньютона–Котеса *замкнутого типа* и *открытого типа*. В вычислительной практике используются как первые, так и вторые.

Далее мы построим и исследуем формулы Ньютона–Котеса для  $n = 0, 1, 2$ , причём будем строить наиболее популярные формулы замкнутого типа. Исключением станет случай  $n = 0$ , когда имеется всего один узел и замкнутая квадратурная формула просто невозможна.

Если  $n = 0$ , то подинтегральная функция  $f(x)$  интерполируется полиномом нулевой степени, т. е. какой-то константой, равной значению  $f(x)$  в единственном узле  $x_0 \in [a, b]$ . Соответствующая квадратурная формула — «формула прямоугольников» — имеет вид

$$\int_a^b f(x) dx \approx (b - a) \cdot f(x_0).$$

Если взять  $x_0 = a$ , то при этом получается квадратурная формула

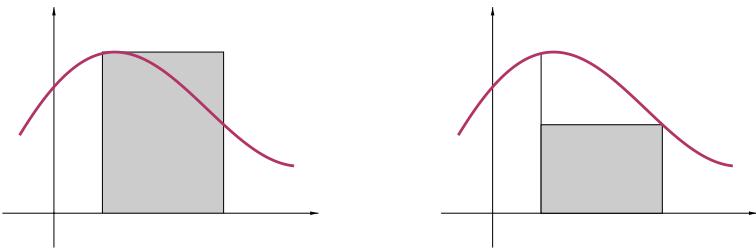


Рис. 2.43. Иллюстрация квадратурных формул левых и правых прямоугольников

«левых прямоугольников», а если  $x_0 = b$  — формула «правых прямоугольников» (рис. 2.43).

Ещё один естественный вариант выбора единственного узла —

$$x_0 = \frac{1}{2}(a + b),$$

т. е. как середины интервала интегрирования  $[a, b]$ . Тогда приходим к квадратурной формуле

$$\boxed{\int_a^b f(x) dx \approx (b - a) \cdot f\left(\frac{a + b}{2}\right)},$$

называемой *формулой средних прямоугольников*: согласно ей интеграл берётся равным площади прямоугольника с основанием  $(b - a)$  и высотой  $f((a + b)/2)$  (рис. 2.44). Эту формулу нередко называют просто «формулой прямоугольников», так как она является часто используемым и наиболее точным вариантом рассмотренных простейших квадратурных формул.

Оценим погрешность формулы средних прямоугольников методом локальных разложений, который ранее был использован при исследовании численного дифференцирования. Разлагая  $f(x)$  в окрестности точки  $x_0 = \frac{1}{2}(a + b)$  по формуле Тейлора с точностью до членов первого порядка, получим

$$f(x) = f\left(\frac{a + b}{2}\right) + f'\left(\frac{a + b}{2}\right) \cdot \left(x - \frac{a + b}{2}\right) + \frac{f''(\xi)}{2} \cdot \left(x - \frac{a + b}{2}\right)^2, \quad (2.152)$$

где  $\xi$  — зависящая от  $x$  точка интервала  $[a, b]$ , которую корректно обозначить через  $\xi(x)$ . Остаточный член квадратуры равен

$$\begin{aligned} R(f) &= \int_a^b f(x) dx - (b-a) \cdot f\left(\frac{a+b}{2}\right) = \\ &= \int_a^b \left( f(x) - f\left(\frac{a+b}{2}\right) \right) dx = \\ &= \int_a^b \left( f'\left(\frac{a+b}{2}\right) \cdot \left(x - \frac{a+b}{2}\right) + \frac{f''(\xi(x))}{2} \cdot \left(x - \frac{a+b}{2}\right)^2 \right) dx = \\ &= \int_a^b \frac{f''(\xi(x))}{2} \cdot \left(x - \frac{a+b}{2}\right)^2 dx, \end{aligned}$$

поскольку

$$\int_a^b \left(x - \frac{a+b}{2}\right) dx = \int_{-\frac{b-a}{2}}^{\frac{b-a}{2}} t dt = 0,$$

т. е. интеграл от первого члена разложения (2.152) зануляется. Следовательно, с учётом принятого нами ранее обозначения

$$M_p := \max_{x \in [a, b]} |f^{(p)}(x)|$$

можно выписать оценку

$$\begin{aligned} |R(f)| &\leq \int_a^b \left| \frac{f''(\xi)}{2} \right| \cdot \left(x - \frac{a+b}{2}\right)^2 dx \leq \frac{M_2}{2} \int_a^b \left(x - \frac{a+b}{2}\right)^2 dx = \\ &= \frac{M_2}{2} \cdot \frac{1}{3} \left(x - \frac{a+b}{2}\right)^3 \Big|_a^b = \frac{M_2(b-a)^3}{24}. \end{aligned}$$

Отсюда, в частности, следует, что для полиномов степени не выше 1 формула (средних) прямоугольников даёт точное значение интеграла, коль скоро вторая производная подинтегральной функции тогда зануляется и  $M_2 = 0$ .

Полученная оценка точности неулучшаема, так как достигается на функции  $g(x) = \left(x - \frac{1}{2}(a+b)\right)^2$ . При этом

$$M_2 = \max_{x \in [a, b]} |g''(x)| = 2, \quad g\left(\frac{a+b}{2}\right) = 0,$$

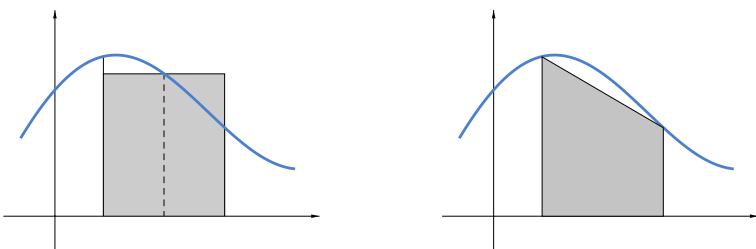


Рис. 2.44. Иллюстрация квадратурных формул средних прямоугольников и трапеций

и потому

$$\int_a^b g(x) dx - (b-a) \cdot g\left(\frac{a+b}{2}\right) = \frac{(b-a)^3}{12} = \frac{M_2(b-a)^3}{24},$$

т. е. имеем точное равенство на погрешность.

Нетрудно показать, что для других формул прямоугольников, когда единственный узел  $x_0$  не совпадает с серединой интервала интегрирования  $[a, b]$ , оценка погрешности имеет вид

$$|R(f)| \leq \frac{M_1(b-a)^2}{2}.$$

Это заметно хуже, чем у формулы средних прямоугольников для нешироких интервалов интегрирования.

Рассмотрим теперь квадратурную формулу Ньютона–Котеса, соответствующую случаю  $n = 1$ , когда подинтегральная функция приближается интерполяционным полиномом первой степени. Для формулы замкнутого типа построим его по узлам  $x_0 = a$  и  $x_1 = b$ , совпадающим с концами интервала интегрирования:

$$P_1(x) = \frac{x-b}{a-b} f(a) + \frac{x-a}{b-a} f(b).$$

Интегрируя это равенство, получим

$$\begin{aligned} \int_a^b P_1(x) dx &= \frac{f(a)}{a-b} \int_a^b (x-b) dx + \frac{f(b)}{b-a} \int_a^b (x-a) dx = \\ &= \frac{f(a)}{a-b} \left. \frac{(x-b)^2}{2} \right|_a^b + \frac{f(b)}{b-a} \left. \frac{(x-a)^2}{2} \right|_a^b = \\ &= \frac{b-a}{2} (f(a) + f(b)). \end{aligned}$$

Мы вывели *квадратурную формулу трапеций*

$$\int_a^b f(x) dx \approx \frac{b-a}{2} \cdot (f(a) + f(b))$$

(2.153)

название которой тоже навеяно геометрическим образом. Фактически согласно этой формуле точное значение интеграла заменяется на значение площади трапеции (стоящей боком на оси абсцисс) с высотой  $(b-a)$  и основаниями, равными  $f(a)$  и  $f(b)$  (рис. 2.44).

Чтобы найти погрешность формулы трапеций, вспомним оценку (2.30) для погрешности интерполяционного полинома. Из неё следует, что

$$f(x) - P_1(x) = \frac{f''(\xi(x))}{2} \cdot (x-a)(x-b)$$

для некоторой точки  $\xi(x) \in [a, b]$ . Таким образом, для формулы трапеций остаточный член есть

$$R(f) = \int_a^b (f(x) - P_1(x)) dx = \int_a^b \frac{f''(\xi(x))}{2} \cdot (x-a)(x-b) dx,$$

но вычисление полученного интеграла на практике нереально из-за неизвестного вида  $\xi(x)$ . Как обычно, имеет смысл вывести какие-то более удобные оценки погрешности, хотя они, возможно, будут не столь точны.

Поскольку выражение  $(x-a)(x-b)$  всюду на интервале  $[a, b]$ , кроме

его концов, сохраняет один и тот же знак, то

$$\begin{aligned} |R(f)| &\leq \int_a^b \frac{|f''(\xi(x))|}{2} \cdot |(x-a)(x-b)| dx \leq \\ &\leq \frac{M_2}{2} \cdot \left| \int_a^b (x-a)(x-b) dx \right|, \end{aligned}$$

где  $M_2 = \max_{x \in [a, b]} |f''(x)|$ . Далее

$$\begin{aligned} \int_a^b (x-a)(x-b) dx &= \int_a^b (x^2 - (a+b)x + ab) dx = \\ &= \frac{x^3}{3} \Big|_a^b - (a+b) \frac{x^2}{2} \Big|_a^b + abx \Big|_a^b = \\ &= \frac{1}{6} \left( 2(b^3 - a^3) - 3(a+b)(b^2 - a^2) + 6ab(b-a) \right) = \\ &= \frac{1}{6} (-b^3 + 3ab^2 - 3a^2b + a^3) = -\frac{(b-a)^3}{6}. \end{aligned} \quad (2.154)$$

Поэтому окончательно

$$|R(f)| \leq \frac{M_2(b-a)^3}{12}.$$

Эта оценка погрешности квадратурной формулы трапеций неулучшаема, поскольку достигается при интегрировании функции  $g(x) = (x-a)^2$  по интервалу  $[a, b]$ .

## 2.13в Квадратурная формула Симпсона

Построим квадратурную формулу Ньютона–Котеса для  $n = 2$ , т. е. для трёх равномерно расположенных узлов

$$x_0 = a, \quad x_1 = \frac{1}{2}(a+b), \quad x_2 = b$$

из интервала интегрирования  $[a, b]$ .

Для упрощения рассуждений выполним параллельный перенос криволинейной трапеции, площадь которой находим с помощью интегрирования, и сделаем точку  $a$  началом координат оси абсцисс (рис. 2.45).

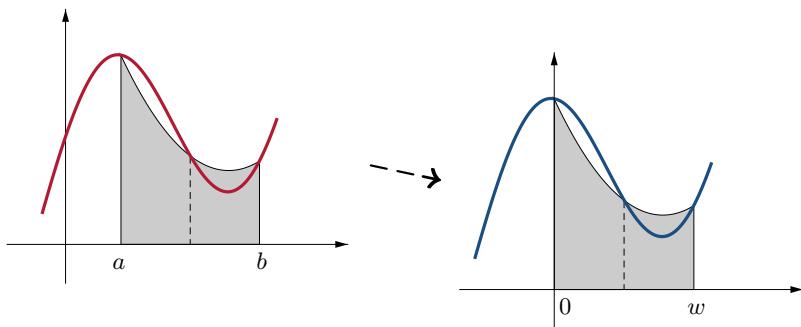


Рис. 2.45. Иллюстрация вывода квадратурной формулы Симпсона

Тогда интервалом интегрирования станет  $[0, w]$ , где  $w = b - a$  — ширина исходного интервала интегрирования.

Пусть

$$\check{P}_2(x) = c_0 + c_1x + c_2x^2$$

— полином второй степени, интерполирующий сдвинутую подинтегральную функцию по узлам  $0$ ,  $w/2$  и  $w$ . Если график  $\check{P}_2(x)$  проходит через точки плоскости  $Oxy$  с координатами

$$(0, f(a)), \quad \left( \frac{w}{2}, f\left(\frac{a+b}{2}\right) \right), \quad (w, f(b)),$$

то

$$\begin{cases} c_0 = f(a), \\ c_0 + c_1 \frac{w}{2} + c_2 \frac{w^2}{4} = f\left(\frac{a+b}{2}\right), \\ c_0 + c_1 w + c_2 w^2 = f(b). \end{cases} \quad (2.155)$$

Площадь, ограниченная графиком интерполяционного полинома  $\check{P}_2(x)$ , равна

$$\begin{aligned} \int_0^w (c_0 + c_1 x + c_2 x^2) dx &= c_0 w + c_1 \frac{w^2}{2} + c_2 \frac{w^3}{3} = \\ &= \frac{w}{6} (6c_0 + 3c_1 w + 2c_2 w^2). \end{aligned}$$

Для построения квадратурной формулы в полученное выражение требуется подставить  $c_0$ ,  $c_1$  и  $c_2$ , которые найдены в результате решения системы уравнений (2.155). Но можно выразить трёхчлен  $6c_0 + 3c_1w + 2c_2w^2$  через значения подинтегральной функции  $f$  в узлах, не решая систему (2.155) явно.

Умножая второе уравнение системы (2.155) на 4 и складывая с первым и третьим уравнением, получим

$$6c_0 + 3c_1w + 2c_2w^2 = f(a) + 4f\left(\frac{a+b}{2}\right) + f(b).$$

Таким образом,

$$\begin{aligned} \int_0^w \check{P}_2(x) dx &= \int_0^w (c_0 + c_1x + c_2x^2) dx = \\ &= \frac{w}{6} \left( f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right), \end{aligned}$$

что даёт приближённое равенство

$$\int_a^b f(x) dx \approx \frac{b-a}{6} \cdot \left( f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right)$$

(2.156)

Оно называется *квадратурной формулой Симпсона* или *формулой парабол* (рис. 2.45), коль скоро основано на приближении подинтегральной функции подходящей параболой.

Интересна история квадратурной формулы Симпсона. Она была известна И. Кеплеру и Б. Кавальери ещё в начале XVII века. Затем Дж. Грегори опубликовал её в 1668 году, а Т. Симпсон вывел в своей диссертации 1743 года (см. разъяснения в книге [39]). В немецкой математической литературе формула Симпсона (формула парабол) называется «Keplersche Fassregel», что буквально переводится как «бочковое правило Кеплера» или «правило Кеплера для бочек». Дело в том, что сам И. Кеплер успешно применял его для решения тогдашней важной практической задачи определения объёмов бочек.

Приведённый выше элегантный способ вывода формулы Симпсона заимствован из известной книги А.Н. Крылова [18], где он применяется даже к более общему случаю. Читатель может самостоятельно убедиться, что та же самая формула получается (но только более длинно

и громоздко) в результате интегрирования по  $[a, b]$  интерполяционного полинома второй степени в форме Лагранжа

$$\begin{aligned} P_2(x) &= \frac{\left(x - \frac{a+b}{2}\right)(x-b)}{\left(a - \frac{a+b}{2}\right)(a-b)} f(a) + \frac{(x-a)(x-b)}{\left(\frac{a+b}{2} - a\right)\left(\frac{a+b}{2} - b\right)} f\left(\frac{a+b}{2}\right) + \\ &\quad + \frac{(x-a)\left(x - \frac{a+b}{2}\right)}{(b-a)\left(b - \frac{a+b}{2}\right)} f(b) = \\ &= \frac{2}{(b-a)^2} \left( \left(x - \frac{a+b}{2}\right)(x-b) f(a) - 2(x-a)(x-b) f\left(\frac{a+b}{2}\right) + \right. \\ &\quad \left. + (x-a)\left(x - \frac{a+b}{2}\right) f(b) \right), \end{aligned}$$

который строится для подинтегральной функции по узлам  $a$ ,  $(a+b)/2$  и  $b$ .

## 2.13г Погрешность формулы Симпсона

**Предложение 2.13.1** (лемма Кеплера) Алгебраическая степень точности квадратурной формулы Симпсона равна 3, т. е. эта формула является точной для любого полинома степени не выше третьей.

**Доказательство.** То, что квадратурная формула Симпсона точна для полиномов степени не выше 2, непосредственно следует из построения этой формулы как интерполяционной, в которой подинтегральная функция интерполируется полиномом второй степени. Поэтому достаточно показать, что формула Симпсона точна для монома  $x^3$ , но не является точной для более высоких степеней переменной.

При интегрировании  $x^3$  получаем

$$\int_a^b x^3 dx = \frac{b^4 - a^4}{4}.$$

С другой стороны, согласно формуле Симпсона

$$\begin{aligned} \frac{b-a}{6} \left( a^3 + 4\left(\frac{a+b}{2}\right)^3 + b^3 \right) &= \frac{b-a}{6} \left( a^3 + \frac{a^3 + 3a^2b + 3ab^2 + b^3}{2} + b^3 \right) = \\ &= \frac{b-a}{6} \cdot \frac{3a^3 + 3a^2b + 3ab^2 + 3b^3}{2} = \\ &= \frac{b-a}{4} (a^3 + a^2b + ab^2 + b^3) = \frac{b^4 - a^4}{4}, \end{aligned}$$

что совпадает с результатом точного интегрирования.

Для монома  $x^4$  длинными, но несложными выкладками нетрудно проверить, что результат, даваемый формулой Симпсона для интеграла по интервалу  $[a, b]$ , т. е.

$$\frac{b-a}{6} \left( a^4 + 4\left(\frac{a+b}{2}\right)^4 + b^4 \right),$$

отличается от точного значения интеграла

$$\int_a^b x^4 dx = \frac{b^5 - a^5}{5}$$

на величину  $(b-a)^5/120$ . Она не зануляется при  $a \neq b$ , так что на полиномах четвёртой степени формула Симпсона уже не точна. ■

Итак, несмотря на то, что формула Симпсона основана на интерполяции подинтегральной функции полиномом степени 2, фактическая точность формулы оказывается более высокой, чем та, что обеспечивается полиномом второй степени. В этой ситуации для аккуратной оценки погрешности формулы Симпсона с помощью известной погрешности алгебраической интерполяции (аналогично тому, как это делалось для формулы трапеций в § 2.13б), желательно аккуратно использовать отмеченный факт. Иными словами, при оценке погрешности формулы Симпсона нужно взять для подинтегральной функции интерполяционный полином третьей степени, а не второй, на основе которого она была построена. При наличии всего трёх узлов мы оказываемся тогда в условиях задачи интерполяции с кратными узлами.

Предполагая существование производной  $f'$  в среднем узле  $x_1 = (a+b)/2$ , можно считать, к примеру, что именно он является кратным

узлом. При этом формально нам необходим такой интерполяционный полином 3-й степени  $H_3(x)$ , что

$$H_3(a) = f(a), \quad H_3(b) = f(b), \quad (2.157)$$

$$H_3\left(\frac{a+b}{2}\right) = f\left(\frac{a+b}{2}\right), \quad H'_3\left(\frac{a+b}{2}\right) = f'\left(\frac{a+b}{2}\right), \quad (2.158)$$

хотя конкретное значение производной в средней точке  $(a+b)/2$  далее никак не будет использоваться. Здесь нам важно лишь то, что при любом значении этой производной решение задачи интерполяции (2.157), (2.158) существует и известна оценка его погрешности.

Существование и единственность решения подобных задач была установлена в § 2.4, и там же обосновывается оценка его погрешности (2.57):

$$f(x) - H_m(x) = \frac{f^{(m+1)}(\xi(x))}{(m+1)!} \prod_{i=0}^n (x - x_i)^{N_i}, \quad (2.159)$$

где  $N_i$  — кратности узлов,  $m = N_0 + N_1 + \dots + N_n - 1$  — степень интерполяционного полинома, а  $\xi(x)$  — некоторая точка из  $[a, b]$ , зависящая от  $x$ . Для решения задачи (2.157)–(2.158) справедливо

$$f(x) - H_3(x) = \frac{f^{(4)}(\xi(x))}{24} \cdot (x - a) \left(x - \frac{a+b}{2}\right)^2 (x - b).$$

Далее, из того, что формула Симпсона точна для полиномов третьей степени, а также из условий (2.157)–(2.158) следуют равенства

$$\begin{aligned} \int_a^b H_3(x) dx &= \frac{b-a}{6} \cdot \left( H_3(a) + 4H_3\left(\frac{a+b}{2}\right) + H_3(b) \right) = \\ &= \frac{b-a}{6} \cdot \left( f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right). \end{aligned} \quad (2.160)$$

Отсюда уже нетрудно вывести выражение для погрешности квадра-

турной формулы Симпсона:

$$\begin{aligned}
 R(f) &= \int_a^b f(x) dx - \frac{b-a}{6} \cdot \left( f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right) = \\
 &= \int_a^b (f(x) - H_3(x)) dx \quad \text{в силу (2.160)} = \\
 &= \int_a^b \frac{f^{(4)}(\xi(x))}{24} \cdot (x-a)\left(x-\frac{a+b}{2}\right)^2(x-b) dx \quad \text{из (2.159).}
 \end{aligned}$$

Из него следует оценка

$$\begin{aligned}
 |R(f)| &= \left| \int_a^b \frac{f^{(4)}(\xi(x))}{24} \cdot (x-a)\left(x-\frac{a+b}{2}\right)^2(x-b) dx \right| = \\
 &\leq \int_a^b \left| \frac{f^{(4)}(\xi(x))}{24} \right| \cdot \left| (x-a)\left(x-\frac{a+b}{2}\right)^2(x-b) \right| dx = \\
 &\leq \frac{M_4}{24} \cdot \left| \int_a^b (x-a)\left(x-\frac{a+b}{2}\right)^2(x-b) dx \right|, \tag{2.161}
 \end{aligned}$$

поскольку в интегрируемой функции подвыражение

$$(x-a)\left(x-\frac{a+b}{2}\right)^2(x-b)$$

не меняет знак на интервале интегрирования  $[a, b]$ . В (2.161), как обычно, обозначено  $M_4 = \max_{x \in [a, b]} |f^{(4)}(x)|$ .

Для вычисления интеграла из (2.161) сделаем замену переменных

$$t = x - \frac{a+b}{2},$$

тогда

$$\begin{aligned} \int_a^b (x-a) \left( x - \frac{a+b}{2} \right)^2 (x-b) dx &= \\ &= \int_{-\frac{b-a}{2}}^{\frac{b-a}{2}} \left( t + \frac{b-a}{2} \right) t^2 \left( t - \frac{b-a}{2} \right) dt = \\ &= \int_{-\frac{b-a}{2}}^{\frac{b-a}{2}} t^2 \left( t^2 - \frac{(b-a)^2}{4} \right) dt = -\frac{(b-a)^5}{120}. \end{aligned}$$

Окончательно

$$|R(f)| \leq \frac{M_4 (b-a)^5}{2880}. \quad (2.162)$$

Как видим, более тонкие рассуждения о свойствах формулы Симпсона позволили получить действительно более точную оценку её погрешности.

## 2.13д Общие интерполяционные квадратурные формулы

Напомним, что интерполяционными квадратурными формулами (или квадратурными формулами интерполяционного типа, см. § 2.13б) мы называли формулы, получающиеся в результате замены подинтегральной функции  $f(x)$  интерполяционным полиномом  $P_n(x)$ , который построен по некоторой совокупности простых узлов  $x_0, x_1, \dots, x_n$  из интервала интегрирования. Выпишем для общего случая этот полином в форме Лагранжа:

$$P_n(x) = \sum_{k=0}^n f(x_k) \phi_k(x),$$

где

$$\phi_k(x) = \frac{(x-x_0) \cdots (x-x_{k-1})(x-x_{k+1}) \cdots (x-x_n)}{(x_k-x_0) \cdots (x_k-x_{k-1})(x_k-x_{k+1}) \cdots (x_k-x_n)},$$

— базисные полиномы Лагранжа (стр. 82).

Интерполяционная квадратурная формула должна получаться из приближённого равенства

$$\int_a^b f(x) dx \approx \int_a^b P_n(x) dx \quad (2.163)$$

в результате выполнения интегрирования в правой части. Как следствие, в представлении (2.150) весовые коэффициенты формулы имеют вид

$$\begin{aligned} c_k &= \int_a^b \phi_k(x) dx = \\ &= \int_a^b \frac{(x - x_0) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_n)}{(x_k - x_0) \cdots (x_k - x_{k-1})(x_k - x_{k+1}) \cdots (x_k - x_n)} dx, \end{aligned} \quad (2.164)$$

$k = 0, 1, \dots, n$ . Эти значения весов  $c_k$ , однозначно определяемые по узлам  $x_0, x_1, \dots, x_n$ , являются отличительным характеристическим признаком именно интерполяционной квадратурной формулы. Если для заданного набора узлов у какой-либо квадратурной формулы весовые коэффициенты равны (2.164), то можно считать, что они таким образом и вычислены, а сама квадратурная формула построена на основе алгебраической интерполяции подинтегральной функции по данным узлам, взятым с единичной кратностью.

**Теорема 2.13.1** Для того, чтобы квадратурная формула (2.150), построенная по  $(n + 1)$  несовпадающим узлам, была интерполяционной, необходимо и достаточно, чтобы её алгебраическая степень точности была не меньшей  $n$ .

В качестве замечания к формулировке нужно отметить, что в условиях теоремы квадратурная формула на самом деле может иметь алгебраическую степень точности выше  $n$ , как, например, формула средних прямоугольников или формула Симпсона.

**Доказательство.** Необходимость условий теоремы очевидна: интерполяционная квадратурная формула на  $n + 1$  узлах, конечно же, точна на полиномах степени  $n$ , поскольку тогда подинтегральная функция совпадает со своим алгебраическим интерполянтом.

Покажем достаточность: если квадратурная формула (2.150), построенная по  $(n + 1)$  узлам, является точной для любого алгебраического полинома степени  $n$ , то её весовые коэффициенты вычисляются

по формулам (2.164), т. е. она является квадратурной формулой интерполяционного типа.

В самом деле, для базисных интерполяционных полиномов  $\phi_i(x)$  выполнено свойство (2.11)

$$\phi_i(x_k) = \delta_{ik} = \begin{cases} 0 & \text{при } i \neq k, \\ 1 & \text{при } i = k, \end{cases}$$

и они имеют степень  $n$ . Следовательно, применяя рассматриваемую квадратурную формулу для вычисления интеграла от  $\phi_i(x)$ , получим

$$\int_a^b \phi_i(x) dx = \sum_{k=0}^n c_k \phi_i(x_k) = \sum_{k=0}^n c_k \delta_{ik} = c_i,$$

и это верно для всех  $i = 0, 1, \dots, n$ . Иными словами, имеют место равенства (2.164), что и требовалось доказать. ■

**Следствие.** Сумма весов интерполяционной квадратурной формулы равна ширине интервала интегрирования.

Дело в том, что любая интерполяционная квадратурная формула должна быть точной на полиномах нулевой степени — константах, так как она построена не менее чем по одному узлу. Поэтому, применив интерполяционную квадратурную формулу к вычислению интеграла от полинома нулевой степени  $P_0(x) = 1$ , получим равенство

$$b - a = \int_a^b 1 dx = \sum_{k=0}^n c_k.$$

Оно и доказывает сформулированное свойство.

Из (2.163) ясно, что погрешность интерполяционных квадратурных формул равна

$$R(f) = \int_a^b R_n(f, x) dx,$$

где  $R_n(f, x)$  — остаточный член алгебраической интерполяции. В § 2.2e была получена оценка для  $R_n(f, x)$  в форме Коши (2.30)

$$R_n(f, x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!} \cdot \omega_n(x),$$

где  $\xi(x) \in [a, b]$ . Поэтому

$$R(f) = \frac{1}{(n+1)!} \int_a^b f^{(n+1)}(\xi(x)) \omega_n(x) dx.$$

Справедливы огрублённые оценки

$$\begin{aligned} |R(f)| &\leq \frac{M_{n+1}}{(n+1)!} \int_a^b |\omega_n(x)| dx \leq \\ &\leq \frac{M_{n+1} (b-a)^{n+2}}{(n+1)!}, \end{aligned} \quad (2.165)$$

где  $M_{n+1} = \max_{x \in [a, b]} |f^{(n+1)}(x)|$ . Они ещё раз показывает, что квадратурная формула интерполяционного типа, построенная по  $(n+1)$  узлам, является точной для любого полинома степени не более  $n$ , поскольку тогда  $M_{n+1} = 0$ .

Оценка (2.165), как видно из наших рассуждений, является простейшей, использующей лишь основные свойства алгебраического интерполянта. В некоторых случаях она может оказаться существенно завышенной, что мы могли видеть на примере формулы Симпсона.

Другое необходимое замечание состоит в том, что оценка (2.165), основанная на формуле Коши для погрешности алгебраической интерполяции, может оказаться практически не очень удобной, так как требует информацию о производных довольно высоких порядков при числе узлов, большем чем 2. Это можно было почувствовать уже на примере формулы Симпсона. Оценки погрешности квадратурных формул, основанные на производных первых порядков от подинтегральной функции, можно найти в книге [31]. Кроме того, популярны оценки погрешности квадратур через разделённые разности, см. [88].

## 2.13e Дальнейшие формулы Ньютона–Котеса

В § 2.13б и § 2.13в простейшие квадратурные формулы Ньютона–Котеса — формулы прямоугольников и трапеций, формула Симпсона — были выведены и исследованы средствами, индивидуальными для каждой отдельной формулы. В этом разделе мы взглянем на формулы Ньютона–Котеса с более общих позиций.

Зафиксировав натуральное число  $n$ ,  $n \geq 1$ , возьмём на интервале интегрирования  $[a, b]$  равноотстоящие друг от друга узлы

$$x_k^{(n)} = a + kh, \quad k = 0, 1, \dots, n, \quad h = \frac{b - a}{n}.$$

Для определения весов формул Ньютона–Котеса необходимо вычислить величины (2.164), т. е. интегралы от базисных интерполяционных полиномов Лагранжа. Обозначим их для рассматриваемого случая как

$$A_k^{(n)} = \int_a^b \frac{(x - x_0^{(n)}) \cdots (x - x_{k-1}^{(n)})(x - x_{k+1}^{(n)}) \cdots (x - x_n^{(n)})}{(x_k^{(n)} - x_0^{(n)}) \cdots (x_k^{(n)} - x_{k-1}^{(n)})(x_k^{(n)} - x_{k+1}^{(n)}) \cdots (x_k^{(n)} - x_n^{(n)})} dx,$$

$k = 0, 1, \dots, n$ . Сделаем в этом интеграле замену переменных  $x = a + th$ , где  $t$  пробегает интервал  $[0, n]$ . Тогда

$$dx = h dt,$$

$$(x - x_0^{(n)}) \cdots (x - x_{k-1}^{(n)})(x - x_{k+1}^{(n)}) \cdots (x - x_n^{(n)}) = \\ = h^n t(t-1) \cdots (t-k+1)(t-k-1) \cdots (t-n),$$

$$(x_k^{(n)} - x_0^{(n)}) \cdots (x_k^{(n)} - x_{k-1}^{(n)})(x_k^{(n)} - x_{k+1}^{(n)}) \cdots (x_k^{(n)} - x_n^{(n)}) = \\ = (-1)^{n-k} h^n k!(n-k)!,$$

где считается, что  $0! = 1$ . Окончательно

$$A_k^{(n)} = h \frac{(-1)^{n-k}}{k!(n-k)!} \int_0^n t(t-1) \cdots (t-k+1)(t-k-1) \cdots (t-n) dt,$$

$k = 0, 1, \dots, n$ . Чтобы придать результату не зависящий от интервала интегрирования вид, положим

$$A_k^{(n)} = (b - a) B_k^{(n)},$$

где

$$B_k^{(n)} = \frac{(-1)^{n-k}}{k!(n-k)!n} \int_0^n t(t-1) \cdots (t-k+1)(t-k-1) \cdots (t-n) dt.$$

Теперь уже величины  $B_k^{(n)}$  не зависят от  $h$  и  $[a, b]$ . Они носят название *коэффициентов Котеса* и фактически являются весами квадратурных формул Ньютона–Котеса для интервала интегрирования  $[0, 1]$ .

К примеру, для  $n = 1$

$$B_0^{(1)} = - \int_0^1 (t-1) dt = - \frac{(t-1)^2}{2} \Big|_0^1 = \frac{1}{2},$$

$$B_1^{(1)} = \int_0^1 t dt = \frac{t^2}{2} \Big|_0^1 = \frac{1}{2}.$$

Мы вновь получили веса квадратурной формулы трапеций (2.153). Для случая  $n = 2$

$$B_0^{(2)} = \frac{1}{4} \int_0^2 (t-1)(t-2) dt = \frac{1}{4} \left( \frac{t^3}{3} - 3 \frac{t^2}{2} + 2t \right) \Big|_0^2 = \frac{1}{6},$$

$$B_1^{(2)} = -\frac{1}{2} \int_0^2 t(t-2) dt = -\frac{1}{2} \left( \frac{t^3}{3} - t^2 \right) \Big|_0^2 = \frac{4}{6},$$

$$B_2^{(2)} = \frac{1}{4} \int_0^2 t(t-1) dt = \frac{1}{4} \left( \frac{t^3}{3} - \frac{t^2}{2} \right) \Big|_0^2 = \frac{1}{6}.$$

Полученные коэффициенты соответствуют формуле Симпсона (2.156). И так далее.

За три с лишним столетия, прошедших с момента изобретения квадратурных формул Ньютона–Котеса, коэффициенты Котеса были тщательно вычислены для значений  $n$  из начального отрезка натурального ряда. В табл. 2.2, заимствованной из книги [18], приведены коэффициенты Котеса для  $n \leq 10$  (см. также [2, 25, 41, 95]).

Можно видеть, что с ростом  $n$  значения коэффициентов Котеса  $B_k^{(n)}$  в зависимости от номера  $k$  начинают всё сильнее и сильнее «осциллировать» (напоминая в чём-то пример Рунге, стр. 132). Результатом этого явления оказывается то необычное и противоестественное обстоятельство, что среди весов формул Ньютона–Котеса при числе узлов  $n = 8$  и  $n \geq 10$  встречаются отрицательные (рис. 2.46). Это снижает практическую ценность соответствующих формул, так как при интегрировании знакопостоянных функций может приводить к вычитанию близких чисел и потере точности.

Уже в XX веке выяснилось, что отмеченный недостаток типичен для формул Ньютона–Котеса высоких порядков. Р.О. Кузьмин в 1930

Таблица 2.2. Коэффициенты Котеса  
для первых натуральных номеров

$n$	1	2	3	4	5	6	7	8	9	10
$k=0$	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{8}$	$\frac{7}{90}$	$\frac{19}{288}$	$\frac{41}{840}$	$\frac{751}{17280}$	$\frac{989}{28350}$	$\frac{2857}{89600}$	$\frac{16067}{598752}$
$k=1$	$\frac{1}{2}$	$\frac{4}{6}$	$\frac{3}{8}$	$\frac{16}{45}$	$\frac{25}{96}$	$\frac{9}{35}$	$\frac{3577}{17280}$	$\frac{5838}{28350}$	$\frac{15741}{89600}$	$\frac{106300}{598752}$
$k=2$		$\frac{1}{6}$	$\frac{3}{8}$	$\frac{2}{15}$	$\frac{25}{144}$	$\frac{9}{280}$	$\frac{1323}{17280}$	$-\frac{928}{28350}$	$\frac{1080}{89600}$	$-\frac{48525}{598752}$
$k=3$			$\frac{1}{8}$	$\frac{16}{45}$	$\frac{25}{144}$	$\frac{34}{105}$	$\frac{2989}{17280}$	$\frac{10496}{28350}$	$\frac{19344}{89600}$	$\frac{272400}{598752}$
$k=4$				$\frac{7}{90}$	$\frac{25}{96}$	$\frac{9}{280}$	$\frac{2989}{17280}$	$-\frac{4540}{28350}$	$\frac{5778}{89600}$	$-\frac{260550}{598752}$
$k=5$					$\frac{19}{288}$	$\frac{9}{35}$	$\frac{1323}{17280}$	$\frac{10496}{28350}$	$\frac{5778}{89600}$	$\frac{427368}{598752}$
$k=6$						$\frac{41}{840}$	$\frac{3577}{17280}$	$-\frac{928}{28350}$	$\frac{19344}{89600}$	$-\frac{260550}{598752}$
$k=7$							$\frac{751}{17280}$	$\frac{5838}{28350}$	$\frac{1080}{89600}$	$\frac{272400}{598752}$
$k=8$								$\frac{989}{28350}$	$\frac{15741}{89600}$	$-\frac{48525}{598752}$
$k=9$									$\frac{2857}{89600}$	$\frac{106300}{598752}$
$k=10$										$\frac{16067}{598752}$

году получил в работе [68] асимптотические формулы для коэффициентов Котеса<sup>29</sup>, из которых следует, что сумма их модулей, т. е.

$$\sum_{k=0}^n |B_k^{(n)}|,$$

неограниченно возрастает с ростом  $n$ . Отсюда вытекает, во-первых, что погрешности вычислений с формулами Ньютона–Котеса могут быть

<sup>29</sup>Помимо оригинальной статьи Р.О. Кузьмина [68] эти асимптотические формулы излагаются, к примеру, в учебнике [20].

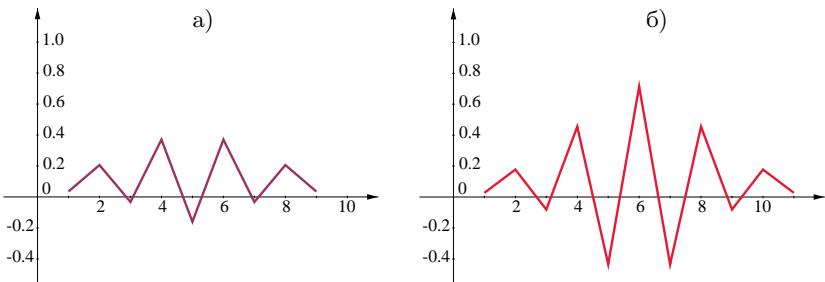


Рис. 2.46. Осцилляции коэффициентов Котеса для  $n = 8$  (а) и  $n = 10$  (б).

сколь угодно велики (см. § 2.13а). Во-вторых, так как дополнительно

$$\sum_{k=0}^n B_k^{(n)} = \frac{1}{b-a} \sum_{k=0}^n A_k^{(n)} = \frac{1}{b-a} \int_a^b 1 \, dx = 1,$$

то при достаточно больших  $n$  среди коэффициентов  $B_k^{(n)}$  обязательно должны быть как положительные, так и отрицательные. Доказательство упрощённого варианта этого результата можно найти в [29].

Общую теорию квадратурных формул Ньютона–Котеса вместе с тщательным исследованием их погрешностей читатель может увидеть, к примеру, в книгах [2, 20, 74]. Следует сказать, что формулы Ньютона–Котеса высоких порядков не очень употребительны. Помимо отмеченной выше численной неустойчивости они проигрывают по точности результатов на одинаковом количестве узлов формулам Гаусса (изучаемым далее в § 2.15) и другим квадратурным формулам.

Из квадратурных формул Ньютона–Котеса приведём ещё формулу «трёх восьмых», которая получается при замене подинтегральной функции интерполяционным полиномом 3-й степени:

$$\boxed{\int_a^b f(x) \, dx \approx \frac{b-a}{8} \cdot \left( f(a) + 3f\left(\frac{2a+b}{3}\right) + 3f\left(\frac{a+2b}{3}\right) + f(b) \right)}.$$

Её погрешность оценивается как

$$|R(f)| \leq \frac{M_4 (b-a)^5}{6480},$$

где  $M_4 = \max_{x \in [a, b]} |f^{(4)}(x)|$ , т. е. порядок точности этой формулы — такой же, как и у формулы Симпсона. Хотя её остаточный член меньше, чем у формулы Симпсона, это достигается ценой дополнительного вычисления подинтегральной функции. А если пересчитать погрешность на одно вычисление функции, то эта формула даже уступает формуле Симпсона (2.156).<sup>30</sup>

Существенно более выгодной является *квадратурная формула Булья*

$$\int_a^b f(x) dx \approx \frac{b-a}{90} \cdot$$

$$\cdot \left( 7f(a) + 32f\left(\frac{3a+b}{4}\right) + 12f\left(\frac{a+b}{2}\right) + 32f\left(\frac{a+3b}{4}\right) + 7f(b) \right)$$

Её погрешность оценивается как

$$|R(f)| \leq \frac{M_6 (b-a)^7}{1935\,360},$$

где  $M_6 = \max_{x \in [a, b]} |f^{(6)}(x)|$ .

Вообще, можно показать, что формулы Ньютона–Котеса с нечётным числом узлов, один из которых приходится на середину интервала интегрирования, имеют (как формула Симпсона) повышенный порядок точности. Подробности читатель может увидеть в [2, 20].

## 2.14 Составные квадратурные формулы

### 2.14а Общая идея и её обоснование

Рассмотренные выше квадратурные формулы дают приемлемую погрешность в случае, когда ширина интервала интегрирования  $[a, b]$  невелика и подинтегральная функция имеет на нём не слишком большие производные. Но если ширина  $(b - a)$  относительно велика или интегрируемая функция имеет большие производные тех порядков, которые входят в оценки остаточного члена, то погрешность вычисления интеграла делается значительной и неприемлемой для практики. Тогда для

<sup>30</sup>Эту формулу иногда называют «второй формулой Симпсона». Учитывая её невысокие качества, Дж. Скарборо в известной книге [88] даёт даже такой категоричный совет: «Никогда не используйте вторую формулу Симпсона».

получения требуемой точности вычисления интеграла применяют *составные квадратурные формулы*, основанные на разбиении интервала интегрирования на подинтервалы меньшей длины. По каждому из полученных подинтервалов вычисляется значение «элементарной квадратуры», а затем искомый интеграл приближается их суммой.

Итак, пусть необходимо найти

$$\int_a^b f(x) dx.$$

Зафиксировав некоторую квадратурную формулу, разобьём интервал интегрирования точками  $a = r_0, r_1, r_2, \dots, r_{N-1}, r_N = b$  на  $N$  частей  $[a, r_1], [r_1, r_2], \dots, [r_{N-1}, b]$ . Тогда в силу аддитивности интеграла

$$\int_a^b f(x) dx = \sum_{i=0}^{N-1} \int_{r_i}^{r_{i+1}} f(x) dx.$$

Пользуясь этим свойством, можно вычислить с помощью выбранной формулы интегралы по отдельным подинтервалам

$$\mathcal{J}_i \approx \int_{r_i}^{r_{i+1}} f(x) dx, \quad i = 0, 1, \dots, N-1,$$

а затем положить

$$\int_a^b f(x) dx \approx \sum_{i=0}^{N-1} \mathcal{J}_i. \quad (2.166)$$

Покажем, что погрешность такого способа вычисления интеграла существенно меньше, чем в результате применения квадратурной формулы ко всему интервалу  $[a, b]$ .

Можно считать, что у используемой квадратурной формулы остаточный член  $R(f)$  для интервала интегрирования  $[a, b]$  имеет оценку

$$|R(f)| \leq K (b-a)^p, \quad (2.167)$$

где  $K$  — константа, зависящая от типа квадратурной формулы и поведения интегрируемой функции на  $[a, b]$ ,

$p$  — положительное число.

Отметим, что  $p \geq 2$  для любых известных нам квадратурных формул. К примеру, для формулы средних прямоугольников  $K = M_2/24$ ,

$M_2 = \max_{\xi \in [a, b]} |f''(\xi)|$  и  $p = 3$ , а для формулы Симпсона  $K = M_4/2880$ ,  $M_4 = \max_{\xi \in [a, b]} |f^{(4)}(\xi)|$  и  $p = 5$ . Константа  $K$ , строго говоря, зависит от интервала интегрирования, и потому в ответственных рассуждениях имеет смысл обозначать эту зависимость, например, указывая интервал в индексе —  $K_{[a, b]}$ ,  $K_{[r_i, r_{i+1}]}$  и т. п. Из определения рассматриваемых констант следует, что они монотонно зависят от интервала интегрирования, т. е.  $K_{[a_1, b_1]} \leq K_{[a_2, b_2]}$ , если  $[a_1, b_1] \subseteq [a_2, b_2]$ . В частности,  $K_{[r_i, r_{i+1}]} \leq K_{[a, b]}$ .

Предположим для простоты, что точки разбиения интервала  $[a, b]$  расположены на нём равномерно, так что все подинтервалы  $[r_i, r_{i+1}]$  имеют одинаковую ширину  $h = (b - a)/N$ . При интегрировании по каждому из подинтервалов  $[r_i, r_{i+1}]$

$$|R(f)| \leq K_{[r_i, r_{i+1}]} \left( \frac{b - a}{N} \right)^p = K_{[r_i, r_{i+1}]} h^p,$$

а полную погрешность интегрирования  $\tilde{R}(f)$  при использовании представления (2.166) можно оценить сверху суммой погрешностей отдельных слагаемых (см. предложение 1.3.1). Поэтому

$$\begin{aligned} |\tilde{R}(f)| &\leq \sum_{i=0}^{N-1} K_{[r_i, r_{i+1}]} \left( \frac{b - a}{N} \right)^p \leq \sum_{i=0}^{N-1} K_{[a, b]} \left( \frac{b - a}{N} \right)^p = \\ &= N K_{[a, b]} \left( \frac{b - a}{N} \right)^p = \frac{K_{[a, b]} (b - a)^p}{N^{p-1}} = \\ &= K_{[a, b]} (b - a) h^{p-1}. \end{aligned} \tag{2.168}$$

Так как  $p - 1 > 0$ , оценка погрешности (2.168) уменьшилась в  $N^{p-1}$  раз по сравнению с (2.167). В принципе, таким способом погрешность вычисления интеграла можно сделать сколь угодно малой.

В соответствии с определением 2.8.1 (стр. 168) число  $(p - 1)$  является *порядком точности* составной квадратурной формулы, построенной на основе элементарной квадратуры порядка точности  $p$  с помощью равномерного разбиения интервала интегрирования. Ясно, что основная идея составных квадратурных формул работает и в случае неравномерного разбиения интервала интегрирования на более мелкие части, но тогда анализ погрешности проводить труднее.

Отметим, что некоторые составные квадратурные формулы обладают свойствами, которые качественно превосходят свойства элементарных квадратур, на которых они основаны. Так, несмотря на свою простоту, квадратурные формулы прямоугольников обладают замечательным свойством наивысшей тригонометрической степени точности (см. § 2.15e).

### 2.146 Некоторые конкретные составные квадратурные формулы

Для равномерного разбиения интервала интегрирования составные квадратурные формулы выглядят особенно просто. Выпишем их явный вид для рассмотренных выше простейших квадратур Ньютона–Котеса и разбиения интервала интегрирования  $[a, b]$  на  $N$  равных частей  $[x_0, x_1], [x_1, x_2], \dots, [x_{N-1}, x_N]$  ширины  $h = (b - a)/N$  каждая, в котором  $a = x_0$  и  $x_N = b$ .

Составная формула средних прямоугольников —

$$\int_a^b f(x) dx \approx h \sum_{i=1}^N f(x_{i-1/2}),$$

где  $x_{i-1/2} = x_i - h/2$ . Её полная погрешность

$$|\tilde{R}(f)| \leq M_2 \frac{(b-a)h^2}{24},$$

т.е. она имеет второй порядок точности. Эта формула, как нетрудно видеть, совпадает с интегральной суммой Римана для интеграла от  $f(x)$  по интервалу  $[a, b]$ .

Составная формула трапеций —

$$\int_a^b f(x) dx \approx h \left( \frac{1}{2}f(a) + \sum_{i=1}^{N-1} f(x_i) + \frac{1}{2}f(b) \right).$$

Её полная погрешность

$$|\tilde{R}(f)| \leq M_2 \frac{(b-a)h^2}{12},$$

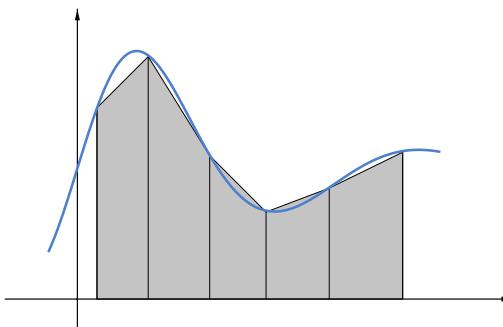


Рис. 2.47. Составная квадратурная формула трапеций

т. е. порядок точности тоже второй.

В квадратурной формуле Симпсона (формуле парабол) интервал интегрирования фактически делится средним узлом на два подинтервала равной ширины. Поэтому, чтобы выписать составную формулу Симпсона, удобно предполагать, что на всём интервале интегрирования  $[a, b]$  задана равномерная сетка с нечётным числом узлов:

$$x_0, x_1, \dots, x_{2N}, \quad \text{где } x_0 = a \text{ и } x_{2N} = b.$$

Тогда элементарные формулы Симпсона строятся на подинтервалах вида  $[x_{2i-2}, x_{2i}]$ ,  $i = 1, 2, \dots, N$ , а соответствующая составная формула формально может быть выписана в следующем виде:

$$\int_a^b f(x) dx \approx \frac{h}{6} \sum_{i=1}^N (f(x_{2i-2}) + 4f(x_{2i-1}) + f(x_{2i})).$$

Но правую часть лучше преобразовать для получения более экономного выражения, в котором подинтегральная функция не вычисляется дважды в стыковочных узлах. Имеем

$$\begin{aligned} & \sum_{i=1}^N (f(x_{2i-2}) + 4f(x_{2i-1}) + f(x_{2i})) = \\ & = (f(a) + 4f(x_1) + f(x_2) + 4f(x_3) + f(x_4) + \dots) = \\ & = (f(a) + 4(f(x_1) + f(x_3) + \dots) + 2(f(x_2) + f(x_4) + \dots) + f(b)). \end{aligned}$$

Итак, составная формула Симпсона (составная формула парабол) имеет следующий вид:

$$\int_a^b f(x) dx \approx \frac{h}{6} \left( f(a) + 4(f(x_1) + f(x_3) + \dots + f(x_{2N-1})) + 2(f(x_2) + f(x_4) + \dots + f(x_{2N-2})) + f(b) \right)$$

Полная погрешность составной формулы Симпсона

$$|\tilde{R}(f)| \leq M_4 \frac{(b-a)h^4}{2880},$$

т. е. формула имеет четвёртый порядок точности.

Отметим, что разбиение интервала интегрирования на равные по ширине части просто реализуется, но не является самым выгодным с точки зрения эффективности вычислений, т. е. достижения заданной точности при наименьших трудозатратах. Для функций, характер поведения которых сильно меняется на интервале интегрирования, погрешности вычисления интегралов на разных подинтервалах равномерного разбиения могут существенно отличаться, тогда как решающий вклад в общую погрешность вносят наиболее неточные результаты. В таких ситуациях гораздо выгоднее использовать неравномерное разбиение исходного интервала интегрирования, при котором погрешности интегрирования по подинтервалам сделаны примерно одинаковыми (см. подробности, к примеру, в книге [1], глава 3).

Идея разбиения области интегрирования на более мелкие части для повышения точности вычисления интеграла применима также к кубатурным формулам, т. е. при вычислении многомерных интегралов. Но при увеличении размерности мы сталкиваемся с новыми эффектами.

В составных квадратурных формулах увеличение точности вычисления интеграла достигается ценой дополнительных трудозатрат. В рассмотренном нами одномерном случае эти трудозатраты растут всего лишь линейно, хотя и здесь необходимость вычисления сложной подинтегральной функции может иногда быть весьма обременительной. Но при возрастании размерности интеграла, когда необходимо прибегнуть к составным кубатурным формулам, рост трудозатрат делается уже значительным, имея тот же порядок, что и размерность пространства.

Так же растёт и погрешность суммирования результатов интегрирования по отдельным подобластям общей области интегрирования. Поэтому увеличение точности составной формулы при возрастании размерности становится всё менее ощутимым.

Как следствие, для вычисления интегралов в пространствах размерности 7–8 и выше обычно используются принципиально другие методы (см. § 2.18).

## 2.15 Квадратурные формулы наивысшей степени точности

### 2.15а Задача оптимизации квадратурных формул

Параметрами квадратурной формулы (2.150)

$$\int_a^b f(x) dx \approx \sum_k c_k f(x_k)$$

являются узлы  $x_k$  и весовые коэффициенты  $c_k$ ,  $k = 0, 1, \dots, n$ . Но при построении квадратурных формул Ньютона–Котеса мы заранее задавали положение узлов, равномерное на интервале интегрирования, и потом по ним находили веса. Таким образом, возможности общей формулы (2.150) были использованы не в полной мере, поскольку для достижения наилучших результатов можно было бы управлять ещё и положением узлов. Лишь в формуле средних прямоугольников положение единственного узла было выбрано из соображений симметрии, и это привело к повышению её точности. Напомним для примера, что специальное неравномерное расположение узлов интерполяции по нулям полиномов Чебышёва или Лежандра существенно улучшает точность интерполирования (см. § 2.3б и § 2.12в).

В связи со сказанным возникает важный методический вопрос: как измерять это «улучшение» квадратурной формулы? Что брать критерием того, насколько точной она является? В идеальном случае желательно было бы минимизировать погрешность квадратурной формулы для тех или иных классов функций, но в такой общей постановке задача делается довольно сложной (хотя и не неразрешимой). Один из возможных естественных ответов на поставленный вопрос состоит в том, чтобы в качестве меры того, насколько хороша и точна квадра-

турная формула, брать её алгебраическую степень точности (см. определение 2.13.1).

Как следствие, сформулированную в начале этого параграфа задачу оптимизации узлов можно поставить, к примеру, следующим образом: для заданного фиксированного числа узлов из интервала интегрирования нужно построить квадратурную формулу, т. е. выбрать её узлы и веса так, чтобы эта формула имела наивысшую алгебраическую степень точности, или, иными словами, была точной на полиномах наиболее высокой степени. Нетривиальное решение этой задачи действительно существует, как будет показано в ближайших разделах книги. Формулы наивысшей алгебраической степени точности называются *квадратурными формулами Гаусса*, поскольку впервые они были рассмотрены в начале XIX века К.Ф. Гауссом.

Далее для удобства мы будем записывать квадратурные формулы Гаусса не в виде (2.150), а как

$$\int_a^b f(x) dx \approx \sum_{k=1}^n c_k f(x_k), \quad (2.169)$$

нумеруя узлы с  $k = 1$ , а не с нуля. Требование точного равенства для любого полинома степени  $m$  в этой формуле в силу её линейности эквивалентно тому, что формула является точной для одночленов  $f(x) = x^l$ ,  $l = 0, 1, 2, \dots, m$ , т. е.

$$\int_a^b x^l dx = \sum_{k=1}^n c_k x_k^l, \quad l = 0, 1, 2, \dots, m.$$

Интегралы от степеней переменной вычисляются тривиально, так что в целом получаем

$$\left\{ \begin{array}{l} \sum_{k=1}^n c_k x_k^l = \frac{1}{l+1} (b^{l+1} - a^{l+1}), \\ l = 0, 1, 2, \dots, m, \end{array} \right.$$

или, в развернутом виде,

$$\left\{ \begin{array}{l} c_1 + c_2 + \dots + c_n = b - a, \\ c_1 x_1 + c_2 x_2 + \dots + c_n x_n = \frac{1}{2} (b^2 - a^2), \\ \vdots \qquad \vdots \qquad \ddots \qquad \vdots \qquad \vdots \\ c_1 x_1^m + c_2 x_2^m + \dots + c_n x_n^m = \frac{1}{m+1} (b^{m+1} - a^{m+1}). \end{array} \right. \quad (2.170)$$

Это система из  $(m + 1)$  нелинейных уравнений с  $2n$  неизвестными величинами  $c_1, c_2, \dots, c_n, x_1, x_2, \dots, x_n$ . Число уравнений совпадает с числом неизвестных при  $m + 1 = 2n$ , т. е. при  $m = 2n - 1$ , и, вообще говоря, это максимальное возможное значение  $m$  для фиксированного  $n$ . При больших значениях  $m$  система уравнений (2.170) переопределена, и в случае общего положения она оказывается неразрешимой.

Сделанное заключение можно обосновать строго.

**Предложение 2.15.1** *Алгебраическая степень точности квадратурной формулы, построенной по  $n$  узлам, не может превосходить  $2n - 1$ .*

**Доказательство.** Пусть  $x_1, x_2, \dots, x_n$  — узлы квадратурной формулы (2.169). Рассмотрим интегрирование по интервалу  $[a, b]$  функции

$$g(x) = ((x - x_1)(x - x_2) \cdots (x - x_n))^2,$$

которая является полиномом степени  $2n$ . Если квадратурная формула (2.169) точна для  $g(x)$ , то

$$\sum_{k=1}^n c_k g(x_k) = \sum_{k=1}^n c_k \cdot 0 = 0.$$

С другой стороны, значение интеграла от  $g(x)$  очевидно не равно нулю. Подинтегральная функция  $g(x)$  всюду на  $[a, b]$  положительна, за исключением лишь конечного множества точек — узлов  $x_1, x_2, \dots, x_n$ , и поэтому  $\int_a^b g(x) dx > 0$ .

Полученное противоречие показывает, что квадратурная формула (2.169) не является точной для полиномов степени  $2n$ . ■

Итак, наивысшая алгебраическая степень точности квадратурной формулы, построенной по  $n$  узлам, в общем случае может быть равна не более  $2n - 1$ , и это немало. Для двух узлов получаем 3, при трёх узлах — 5 и т. д. Для сравнения напомним, что алгебраические степени точности формул трапеций и Симпсона, построенных по двум и трём узлам соответственно, равны всего 1 и 3. При возрастании числа узлов этот выигрыш в алгебраической степени точности формул Гаусса, достигаемый за счёт разумного расположения узлов, нарастает.

## 2.15б Простейшие квадратуры Гаусса

Перейдём к построению квадратурных формул Гаусса. При небольших  $n$  система уравнений (2.170) для узлов и весов может быть решена с помощью несложных аналитических преобразований.

Пусть  $n = 1$ , тогда  $m = 2n - 1 = 1$ , и система уравнений (2.170) принимает вид

$$\begin{cases} c_1 = b - a, \\ c_1 x_1 = \frac{1}{2}(b^2 - a^2). \end{cases}$$

Отсюда

$$c_1 = b - a, \quad x_1 = \frac{1}{2c_1}(b^2 - a^2) = \frac{1}{2}(a + b).$$

Как легко видеть, получающаяся квадратурная формула — это формула (средних) прямоугольников

$$\int_a^b f(x) dx \approx (b - a) \cdot f\left(\frac{a + b}{2}\right).$$

Нам в самом деле известно (см. § 2.13б), что она резко выделяется своей точностью среди родственных квадратурных формул.

Пусть  $n = 2$ , тогда алгебраическая степень точности соответствующей квадратурной формулы равна  $m = 2n - 1 = 3$ . Система уравнений (2.170) для узлов и весов принимает вид

$$\begin{cases} c_1 + c_2 = b - a, \\ c_1 x_1 + c_2 x_2 = \frac{1}{2}(b^2 - a^2), \\ c_1 x_1^2 + c_2 x_2^2 = \frac{1}{3}(b^3 - a^3), \\ c_1 x_1^3 + c_2 x_2^3 = \frac{1}{4}(b^4 - a^4). \end{cases}$$

Она обладает определённой симметрией: одновременная перемена местами  $x_1$  с  $x_2$  и  $c_1$  с  $c_2$  оставляет систему неизменной. По этой причине, учитывая вид первого уравнения, будем искать решение, в котором  $c_1 = c_2$ . Это даёт

$$c_1 = c_2 = \frac{1}{2}(b - a),$$

и из второго уравнения тогда получаем

$$x_1 + x_2 = a + b. \tag{2.171}$$

Отсюда после возвведения в квадрат имеем

$$x_1^2 + 2x_1x_2 + x_2^2 = a^2 + 2ab + b^2. \quad (2.172)$$

В то же время с учётом найденных значений  $c_1$  и  $c_2$  из третьего уравнения системы следует

$$x_1^2 + x_2^2 = \frac{2}{3}(b^2 + ab + a^2),$$

и, вычитая это равенство из (2.172), получим

$$x_1x_2 = \frac{1}{6}(b^2 + 4ab + a^2). \quad (2.173)$$

Соотношения (2.171) и (2.173) на основе известной из элементарной алгебры теоремы Виета позволяют сделать вывод, что  $x_1$  и  $x_2$  являются решениями квадратного уравнения

$$x^2 - (a + b)x + \frac{1}{6}(b^2 + 4ab + a^2) = 0,$$

так что

$$x_{1,2} = \frac{1}{2}(a + b) \pm \frac{\sqrt{3}}{6}(b - a). \quad (2.174)$$

Удовлетворение найденными решениями четвёртого уравнения системы проверяется прямой подстановкой. Кроме того, поскольку

$$\frac{1}{2} > \frac{1}{2} \cdot \frac{1}{\sqrt{3}} = \frac{\sqrt{3}}{6},$$

$x_1$  и  $x_2$  действительно лежат на интервале  $[a, b]$ . В целом мы вывели *квадратурную формулу Гаусса с двумя узлами*

$$\int_a^b f(x) dx \approx \frac{b-a}{2} \cdot (f(x_1) + f(x_2)), \quad (2.175)$$

где  $x_1$  и  $x_2$  определяются посредством (2.174). Она очень похожа на формулу трапеций (2.153) и фактически является её модификацией. Согласно (2.175) приближённым значением интеграла тоже берётся площадь трапеции с высотой  $(b - a)$ , но основаниями, равными значениям интегрируемой функции в двух специально подобранных узлах.

**Пример 2.15.1** Вычислим с помощью полученной выше формулы Гаусса по двумя узлам (2.175) интеграл

$$\int_0^{\pi/2} \cos x \, dx,$$

точное значение которого согласно формуле Ньютона–Лейбница равно  $\sin(\pi/2) - \sin 0 = 1$ . В соответствии с (2.174) и (2.175) имеем

$$\begin{aligned} \int_0^{\pi/2} \cos x \, dx &\approx \frac{\pi/2}{2} \cdot \left( \cos\left(\frac{\pi}{4} - \frac{\sqrt{3}}{6}\frac{\pi}{2}\right) + \cos\left(\frac{\pi}{4} + \frac{\sqrt{3}}{6}\frac{\pi}{2}\right) \right) = \\ &= 0.99847. \end{aligned}$$

Формула Ньютона–Котеса с двумя узлами 0 и  $\pi/2$  — формула трапеций — даёт для этого интеграла значение

$$\int_0^{\pi/2} \cos x \, dx \approx \frac{\pi/2}{2} \cdot \left( \cos 0 + \cos \frac{\pi}{2} \right) = 0.78540,$$

точность которого весьма низка.

Чтобы получить с формулами Ньютона–Котеса точность вычисления рассматриваемого интеграла, сравнимую с той, что даёт формула Гаусса, приходится брать больше узлов. Так, формула Симпсона (2.156), использующая три узла — 0,  $\pi/4$  и  $\pi/2$ , приводит к результату

$$\begin{aligned} \int_0^{\pi/2} \cos x \, dx &\approx \frac{\pi/2}{6} \cdot \left( \cos 0 + 4 \cos \frac{\pi}{4} + \cos \frac{\pi}{2} \right) = \\ &= \frac{\pi}{12} (1 + 2\sqrt{2}) = 1.0023, \end{aligned}$$

погрешность которого по порядку величины примерно равна погрешности ответа по формуле Гаусса (2.175), но всё-таки превосходит её в полтора раза. ■

С ростом  $n$  сложность системы уравнений (2.170) для узлов и весов формул Гаусса быстро увеличивается, так что в общем случае не вполне ясно, будет ли она иметь вещественные решения при любом на-перёд заданном  $n$ . Кроме того, эти решения системы (2.170), соответствующие узлам, должны быть различны и принадлежать интервалу интегрирования  $[a, b]$ .

Получение ответов на поставленные вопросы непосредственно из системы уравнений (2.170) в принципе возможно (см. учебник [9], глава XVI, § 9), но оно является громоздким и несколько искусственным. Мы рассмотрим другое, более элегантное решение задачи построения формул Гаусса, которое основано на расчленении общей задачи на отдельные подзадачи:

- 1) нахождения узлов формулы;
- 2) вычисления её весовых коэффициентов.

Зная узлы формулы, можно подставить их в систему уравнений (2.170), которая в результате решительно упростится, превратившись в систему линейных алгебраических уравнений относительно  $c_1, c_2, \dots, c_n$ . Она будет переопределённой, но нам достаточно рассматривать подсистему из первых  $n$  уравнений, матрица которой является транспонированной матрицей Вандермонда относительно узлов  $x_1, x_2, \dots, x_n$ . Решение этой подсистемы даст искомые веса квадратурной формулы Гаусса. Можно показать, что они будут удовлетворять оставшимся  $n$  уравнениям системы (2.170) (см., к примеру, [9]).

Другой способ решения подзадачи 2, когда узлы уже известны, — это вычисление весовых коэффициентов по формулам (2.164) путём интегрирования базисных интерполяционных полиномов Лагранжа, построенных по известным узлам. В этом случае мы пользуемся тем фактом, что конструируемая квадратурная формула Гаусса оказывается квадратурной формулой интерполяционного типа. Это прямо следует из теоремы 2.13.1, коль скоро формула Гаусса, построенная по  $n$  узлам, является точной для полиномов степени  $n-1$ . Детали этого построения и конкретные выкладки читатель может найти, к примеру, в [2].

## 2.15в Выбор узлов для квадратурных формул Гаусса

**Теорема 2.15.1** *Квадратурная формула (2.169)*

$$\int_a^b f(x) dx \approx \sum_{k=1}^n c_k f(x_k),$$

*построенная по  $n$  узлам, является точной на алгебраических полиномах степени  $(2n-1)$  тогда и только тогда, когда*

- (а) она является интерполяционной квадратурной формулой;

(б) её узлы  $x_1, x_2, \dots, x_n$  являются нулями такого полинома

$$\omega(x) = (x - x_1)(x - x_2) \cdots (x - x_n),$$

что

$$\int_a^b \omega(x) q(x) dx = 0 \quad (2.176)$$

для любого полинома  $q(x)$  степени не выше  $(n - 1)$ .

Выражение

$$\int_a^b \omega(x) q(x) dx$$

— интеграл от произведения двух функций, уже встречалось нам в § 2.11. Это интегральное скалярное произведение на интервале  $[a, b]$  с единичным весом. По этой причине утверждение теоремы 2.15.1 часто формулируют так: для того чтобы квадратурная формула

$$\int_a^b f(x) dx \approx \sum_{k=1}^n c_k f(x_k),$$

построенная по  $n$  узлам  $x_1, x_2, \dots, x_n$ , была точной на алгебраических полиномах степени  $(2n - 1)$ , необходимо и достаточно, чтобы эта формула была интерполяционной, а её узлы являлись нулями полинома, который относительно интегрального скалярного произведения на  $[a, b]$  с единичным весом ортогонален любому полиному степени не выше  $(n - 1)$ .

**Доказательство.** Необходимость. Пусть рассматриваемая квадратурная формула точна на полиномах степени  $(2n - 1)$ . Таковым является, в частности, полином  $\omega(x) q(x)$ , имеющий степень не выше  $n + (n - 1)$ , если степень  $q(x)$  не превосходит  $(n - 1)$ . Тогда имеет место равенство

$$\int_a^b \omega(x) q(x) dx = \sum_{k=1}^n c_k \omega(x_k) q(x_k) = 0,$$

поскольку все  $\omega(x_k) = 0$ . Так как этот результат верен для любого полинома  $q(x)$  степени не выше  $n - 1$ , отсюда следует выполнение условия (б).

Справедливость условия (а) следует из теоремы 2.13.1: если построенная по  $n$  узлам квадратурная формула (2.169) является точной для любого полинома степени не менее  $n - 1$ , то она — интерполяционная.

Достаточность. Предположим, что интерполяционная квадратурная формула построена по узлам  $x_1, x_2, \dots, x_n \in [a, b]$ , которые являются различными нулями полинома  $\omega(x)$  степени  $n$ , удовлетворяющего условию ортогональности (2.176) с любым полиномом  $q(x)$  степени не выше  $(n - 1)$ . Покажем, что эта квадратурная формула будет точна на алгебраических полиномах степени  $2n - 1$ .

Если  $f(x)$  — произвольный полином степени  $2n - 1$ , то, поделив его на полином  $\omega(x)$ , получим представление

$$f(x) = \omega(x) q(x) + r(x), \quad (2.177)$$

в котором  $q(x)$  и  $r(x)$  — соответственно частное и остаток от деления  $f(x)$  на  $\omega(x)$  [17, 22]. При этом полином  $q(x)$  имеет степень  $(2n - 1) - n = n - 1$ , а степень полинома-остатка  $r(x)$  по определению меньше степени  $\omega(x)$ , т. е. не превосходит  $n - 1$ . Отсюда

$$\int_a^b f(x) dx = \int_a^b \omega(x) q(x) dx + \int_a^b r(x) dx = \int_a^b r(x) dx \quad (2.178)$$

в силу сделанного нами предположения об ортогональности  $\omega(x)$  всем полиномам степени не выше  $n - 1$ .

Но по условиям теоремы рассматриваемая квадратурная формула является интерполяционной и построена по  $n$  узлам. Поэтому она является точной на полиномах степени  $n - 1$  (см. теорему 2.13.1), в частности, на полиноме  $r(x)$ . Следовательно,

$$\begin{aligned} \int_a^b r(x) dx &= \sum_{k=1}^n c_k r(x_k) = \sum_{k=1}^n c_k (\omega(x_k) q(x_k) + r(x_k)) = \\ &\quad \text{в силу равенств } \omega(x_k) = 0 \\ &= \sum_{k=1}^n c_k f(x_k), \text{ поскольку имеет место (2.177).} \end{aligned}$$

Сравнивая результаты этой выкладки с (2.178), получим

$$\int_a^b f(x) dx = \sum_{k=1}^n c_k f(x_k),$$

т. е. исследуемая квадратурная формула действительно является точной на полиномах степени  $2n - 1$ . ■

Подведём промежуточные итоги. Процедура построения квадратурных формул Гаусса разделена нами на две отдельные задачи — нахождения узлов и вычисления весов. В свою очередь, узлы квадратурной формулы, как выясняется, можно взять нулями некоторых специальных полиномов  $\omega(x)$ , удовлетворяющих условию (б) из теоремы 2.15.1. В этих полиномах легко угадываются знакомые нам из § 2.12 ортогональные полиномы, которые являются полиномами Лежандра для случая  $[a, b] = [-1, 1]$  или соответствующим образом преобразованы из них для любого другого интервала интегрирования  $[a, b]$ .

## 2.15г Практическое применение формул Гаусса

Отдельное нахождение узлов и весов формул Гаусса для каждого конкретного интервала интегрирования является весьма трудозатратным. Если бы нам нужно было проделывать эту процедуру всякий раз при смене интервала интегрирования, то практическое применение формул Гаусса значительно потеряло бы свою привлекательность. Естественная идея состоит в том, чтобы найти узлы и веса формул Гаусса для какого-то одного «канонического» интервала, а затем получать их для любого другого интервала с помощью несложных преобразований.

В качестве канонического интервала интегрирования обычно берут  $[-1, 1]$ , т. е. именно тот интервал, для которого строятся ортогональные полиномы Лежандра. Этот интервал удобен также симметричностью относительно нуля, которая позволяет более просто использовать свойство симметрии узлов и весовых коэффициентов квадратурной формулы. В § 2.12 мы указали рецепт построения из полиномов Лежандра алгебраических полиномов, которые ортогональны с единичным весом, для любого интервала вещественной оси. Этой техникой и нужно воспользоваться в данном случае.

Итак, пусть

$$x = \frac{1}{2}(a + b) + \frac{1}{2}(b - a)y. \quad (2.179)$$

Переменная  $x$  будет пробегать интервал  $[a, b]$ , когда  $y$  изменяется в  $[-1, 1]$ . Если  $y_i$ ,  $i = 1, 2, \dots, n$ , — нули полинома Лежандра, которые согласно предложению 2.12.2 все различны и лежат на интервале  $[-1, 1]$ ,

то узлами квадратурной формулы Гаусса для интервала интегрирования  $[a, b]$  являются

$$x_i = \frac{1}{2}(a + b) + \frac{1}{2}(b - a)y_i, \quad i = 1, 2, \dots, n. \quad (2.180)$$

Все они также различны и лежат на интервале интегрирования  $[a, b]$ .

Веса  $c_k$  любой интерполяционной квадратурной формулы, как было показано в § 2.13д, могут быть выражены в виде интегралов (2.164). В случае формул Гаусса (когда узлы нумеруются с единицы) они принимают вид

$$c_k = \int_a^b \phi_k(x) dx, \quad k = 1, 2, \dots, n, \quad (2.181)$$

где  $\phi_k(x)$  —  $k$ -й базисный интерполяционный полином Лагранжа (см. стр. 82), построенный по узлам (2.180):

$$\phi_i(x) = \frac{(x - x_1) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n)}{(x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)}.$$

Тогда, выполняя в интеграле (2.181) замену переменных (2.179), получим

$$dx = d\left(\frac{1}{2}(a + b) + \frac{1}{2}(b - a)y\right) = \frac{1}{2}(b - a) dy,$$

и потому

$$c_k = \int_a^b \phi_k(x) dx = \frac{1}{2}(b - a) \int_{-1}^1 \phi_k(y) dy, \quad k = 1, 2, \dots, n,$$

где  $\phi_k(y)$  —  $k$ -й базисный полином Лагранжа, построенный по узлам  $y_i$ ,  $i = 1, 2, \dots, n$ , которые являются нулями  $n$ -го полинома Лежандра. Мы обосновали

**Предложение 2.15.2** *Веса квадратурной формулы Гаусса для произвольного интервала интегрирования  $[a, b]$  равны произведениям весов для канонического интервала  $[-1, 1]$  на радиус интервала интегрирования, т. е. на  $\frac{1}{2}(b - a)$ .*

Для интервала  $[-1, 1]$  узлы квадратурных формул Гаусса (т. е. нули полиномов Лежандра) и их веса тщательно затабулированы для первых натуральных чисел  $n$  вплоть до нескольких десятков. Обсуждение вычислительных формул и других деталей численных процедур для их

нахождения читатель может найти, к примеру, в книгах [2, 74] и в специальных журнальных статьях. В частности, оказывается, что весовые коэффициенты формулы Гаусса с  $n$  узлами выражаются как

$$c_k = \frac{2}{(1 - x_k^2)(L'_n(x_k))^2}, \quad k = 1, 2, \dots, n, \quad (2.182)$$

где  $x_k$  — узлы формул Гаусса, а  $L_n(x)$  —  $n$ -й полином Лежандра в виде, даваемом формулой Родрига (2.138). Эта формула была впервые получена Э.Б. Кристоффелем в середине XIX века [98], и потому весовые коэффициенты квадратурных формул Гаусса называют ещё *числами Кристоффеля*.

Конкретные числовые значения узлов и весов квадратур Гаусса приводятся в подробных руководствах по вычислительным методам [1, 2, 9, 18, 19, 28] или в специализированных справочниках, например в [41, 67]. В частности, в учебнике [2] значения весов и узлов формул Гаусса приведены для небольших  $n$  с 16 значащими цифрами, в книге [19] — с 15 значащими цифрами вплоть до  $n = 16$ , а в справочниках [41, 67] — с 20 значащими цифрами вплоть до  $n = 96$  и  $n = 48$ . Таким образом, практическое применение квадратур Гаусса обычно не встречает затруднений.

При небольших значениях  $n$  можно дать точные числовые выражения для узлов формул Гаусса как нулей полиномов Лежандра  $L_n(x)$ , имеющих явные представления (2.141). В частности, для  $n = 1$  или  $n = 2$  нули полиномов Лежандра  $L_1(x)$  и  $L_2(x)$  соответствуют узлам формулы средних прямоугольников и квадратурной формулы Гаусса с двумя узлами (2.175), которую мы построили ранее другим способом.

Далее, для  $n = 3$

$$L_3(x) = \frac{1}{2}(5x^3 - 3x) = \frac{1}{2}x(5x^2 - 3).$$

Поэтому для канонического интервала интегрирования  $[-1, 1]$  узлы квадратурной формулы Гаусса с тремя узлами равны

$$\begin{aligned} x_1 &= -\sqrt{\frac{3}{5}} = -0.77459\ 66692\ 41483\dots, \\ x_2 &= 0, \\ x_3 &= \sqrt{\frac{3}{5}} = 0.77459\ 66692\ 41483\dots. \end{aligned}$$

Для  $n = 4$

$$L_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3),$$

Таблица 2.3. Узлы и веса квадратурных формул Гаусса

Узлы	Веса
$n = 2$	
$\pm 0.57735 \ 02691 \ 89626$	1.00000 00000 00000
$n = 3$	
0.00000 00000 00000	0.88888 88888 88889
$\pm 0.77459 \ 66692 \ 41483$	0.55555 55555 55556
$n = 4$	
$\pm 0.33998 \ 10435 \ 84856$	0.65214 51548 62546
$\pm 0.86113 \ 63115 \ 94053$	0.34785 48451 37454
$n = 5$	
0.00000 00000 00000	0.56888 88888 88889
$\pm 0.53846 \ 93101 \ 05683$	0.47862 86704 99366
$\pm 0.90617 \ 98459 \ 38664$	0.23692 68850 56189

и нахождение нулей этого биквадратного полинома труда не представляет. Аналогично и для  $n = 5$ , когда

$$L_5(x) = \frac{1}{8}(63x^5 - 70x^3 + 15x) = \frac{1}{8}x(63x^4 - 70x^2 + 15).$$

Соответствующие весовые коэффициенты можно легко найти с помощью формулы Кристоффеля (2.182) или решением небольших систем линейных уравнений, к которым редуцируется система (2.170) после подстановки в неё известных значений узлов.

Численные значения узлов и весов квадратурных формул Гаусса для  $n = 2, 3, 4, 5$  сведены в табл. 2.3. Видно, что узлы располагаются симметрично относительно середины интервала интегрирования, а равноотстоящие от неё весовые коэффициенты одинаковы. Симметрия расположения узлов очевидно следует из того, что любой полином Лежандра является, в зависимости от номера, либо чётной, либо нечётной

функцией.

## 2.15д Погрешность квадратур Гаусса

Для исследования остаточного члена квадратурных формул Гаусса предположим, что подинтегральная функция  $f(x)$  имеет достаточно высокую гладкость, т. е. достаточно высокий порядок непрерывных производных.

Желая воспользоваться результатом о погрешности алгебраической интерполяции, построим для  $f(x)$  интерполяционный полином, принимающий в узлах  $x_1, x_2, \dots, x_n$  значения  $f(x_1), f(x_2), \dots, f(x_n)$ . Поскольку квадратурная формула Гаусса точна на полиномах степени  $2n - 1$ , для адекватного учёта этого факта и получения наиболее точной оценки погрешности степень полинома, интерполирующего подинтегральную функцию, тоже нужно взять равной  $2n - 1$ . Имея всего  $n$  узлов, мы находимся в ситуации, совершенно аналогичной той, что встречалась при анализе формулы Симпсона. Необходимая степень интерполяционного полинома для формулы Гаусса получится, если рассматривать для подинтегральной функции интерполяцию с кратными узлами (см. § 2.4). В данном случае суммарная кратность узлов интерполяции должна быть равна  $2n$ , и её можно получить, например, назначив кратность всех  $n$  узлов равной двум. Иными словами, будем предполагать заданными в  $x_1, x_2, \dots, x_n$  значения функции  $f(x_1), f(x_2), \dots, f(x_n)$  и некоторые «виртуальные» значения производных  $f'(x_1), f'(x_2), \dots, f'(x_n)$ .

Тогда согласно (2.57) погрешность интерполирования подинтегральной функции  $f(x)$  полиномом Эрмита  $H_{2n-1}(x)$  равна

$$\begin{aligned} R_{2n-1}(f, x) &= f(x) - H_{2n-1}(x) = \\ &= \frac{f^{(2n)}(\xi(x))}{(2n)!} \cdot \prod_{i=1}^n (x - x_i)^2 = \frac{f^{(2n)}(\xi(x))}{(2n)!} \cdot (\omega(x))^2, \end{aligned}$$

где  $\omega(x) = (x - x_1)(x - x_2) \cdots (x - x_n)$  и  $\xi(x)$  — некоторая точка, зависящая от  $x$ , из интервала интерполяции  $[a, b]$ . По условиям интерполяции  $H_{2n-1}(x_i) = f(x_i)$ ,  $i = 1, 2, \dots, n$ , и, следовательно, если  $c_i$  —

веса квадратурной формулы Гаусса, то

$$\begin{aligned}
 \int_a^b f(x) dx &= \int_a^b (H_{2n-1}(x) + R_{2n-1}(f, x)) dx = \\
 &= \int_a^b H_{2n-1}(x) dx + \int_a^b R_{2n-1}(f, x) dx = \\
 &= \sum_{i=1}^n c_i H_{2n-1}(x_i) + \int_a^b R_{2n-1}(f, x) dx = \\
 &\quad \text{из-за того, что формула точна на полиноме } H_{2n-1}(x) \\
 &= \sum_{i=1}^n c_i f(x_i) + \frac{1}{(2n)!} \int_a^b f^{(2n)}(\xi(x)) (\omega(x))^2 dx.
 \end{aligned}$$

Выражение для второго слагаемого последней суммы, т. е. для остаточного члена квадратуры, можно упростить, приняв во внимание значение множителя  $(\omega(x))^2$ . В силу интегральной теоремы о среднем [12, 40]) имеем

$$\int_a^b f^{(2n)}(\xi(x)) (\omega(x))^2 dx = f^{(2n)}(\theta) \int_a^b (\omega(x))^2 dx$$

для некоторой точки  $\theta \in ]a, b[$ . Таким образом, погрешность квадратурной формулы Гаусса, построенной по  $n$  узлам  $x_1, x_2, \dots, x_n \in [a, b]$ , равна

$$R(f) = \frac{f^{(2n)}(\theta)}{(2n)!} \int_a^b (\omega(x))^2 dx,$$

где  $\theta \in ]a, b[$ . Это выражение можно упростить и дальше.

Узлы  $x_1, x_2, \dots, x_n$  — это нули полинома, полученного из полинома Лежандра линейной заменой переменных, а интеграл от квадрата — это его скалярное произведение на себя. По этой причине интеграл в полученной формуле для погрешности можно найти точно, приведя его к интервалу  $[-1, 1]$  заменой переменных (2.47), т. е.

$$x = \frac{2y - (b + a)}{(b - a)}.$$

Из-за того, что у полинома  $\omega(x)$  старший коэффициент — единица, для вычисления нашего интеграла удобнее воспользоваться приведёнными

полиномами Лежандра (2.143), скорректировав результат предложения 2.12.1 с учётом соотношения (2.144). После несложных выкладок это даёт

$$\int_a^b (\omega(x))^2 dx = \frac{(n!)^4}{(2n+1)((2n)!)^2} (b-a)^{2n+1}.$$

В конце концов для остаточного члена получаем представление

$$R(f) = \frac{(n!)^4}{(2n+1)((2n)!)^3} (b-a)^{2n+1} f^{(2n)}(\theta), \quad (2.183)$$

где  $\theta$  — некоторая внутренняя точка из интервала интегрирования  $[a, b]$ . Более практична грубая оценка

$$|R(f)| \leq \frac{(n!)^4}{(2n+1)((2n)!)^3} M_{2n} (b-a)^{2n+1},$$

в которой, как обычно, обозначено  $M_{2n} = \max_{x \in [a, b]} |f^{(2n)}(x)|$ .

В частности, для квадратурной формулы Гаусса (2.175) с двумя узлами (2.174) имеем

$$|R_2(f)| \leq \frac{M_4(b-a)^5}{4320},$$

что даже лучше оценки погрешности для формулы Симпсона. На практике мы могли видеть это в примере 2.15.1.

Отметим, что выведенная оценка (2.183) справедлива лишь при достаточноной гладкости подинтегральной функции  $f(x)$ . Вообще, квадратурные формулы Гаусса с большим числом узлов целесообразно применять лишь для функций, обладающих значительной гладкостью.

Другое важное наблюдение состоит в том, что в выражении (2.183) знаменатель числового коэффициента, т. е.  $(2n+1)((2n)!)^3$ , с ростом  $n$  может быть сделан сколь угодно большим числителем  $(n!)^4 (b-a)^{2n+1}$ . В самом деле, знаменатель можно грубо оценить снизу как

$$\begin{aligned} (2n+1)((2n)!)^3 &= (2n+1)(n! \cdot (n+1) \cdots 2n)^3 > \\ &> (2n+1)(n! \cdot n!)^3 = (2n+1)(n!)^6. \end{aligned}$$

По этой причине

$$\frac{(n!)^4}{(2n+1)((2n)!)^3} (b-a)^{2n+1} < \frac{1}{(2n+1)(n!)^2} (b-a)^{2n+1}.$$

При увеличении  $n$  выражение в правой части может быть сделано сколь угодно малым, так как факториал  $n!$  становится, в конце концов, неограниченно большим значений показательной функции с любым фиксированным основанием.

Как следствие, если производные подинтегральной функции не растут «слишком быстро» с ростом их порядка, то при увеличении числа узлов и гладкости интегрируемой функции порядок точности квадратурных формул Гаусса может быть сделан сколь угодно высоким. В этом квадратуры Гаусса принципиально отличаются, к примеру, от интерполяции с помощью сплайнов, которая сталкивается с ограничением на порядок точности, не зависящим от гладкости исходных данных (стр. 151). Таким образом, квадратурные формулы Гаусса дают пример *ненасыщаемого* численного метода, у которого порядок точности (поправка сходимости) может быть сделан любым в зависимости от того, насколько гладкими являются входные данные для этого метода.

Составные квадратурные формулы Гаусса конструируются и исследуются совершенно аналогично составным формулам других типов, рассмотренным выше в § 2.14, и даже проще. Не будем здесь разворачивать детали этого несложного построения и лишь отметим, что узлы формул Гаусса не приходятся на концы интервала интегрирования, и потому составная формула Гаусса является простой суммой элементарных квадратур. Вследствие оценок (2.183) и (2.168) порядок точности составной квадратурной формулы Гаусса по  $n$  узлам равен  $2n$ .

Практическим недостатком квадратурных формул Гаусса является нетривиальное расположение узлов, выражения для которых к тому же содержат иррациональности. Из-за этого при применении формул Гаусса для таблично заданных функций, как правило, необходимо интерполирование известных значений функции, которое требует дополнительных трудозатрат и уменьшает точность задания интегрируемой функции в узлах формулы.

В заключение темы отметим, что на практике нередко требуется включать во множество узлов квадратурной формулы какие-либо фиксированные точки интервала интегрирования. Ими могут быть, к примеру, его концы (один или оба) либо какие-то выделенные внутренние точки. С другой стороны, мы готовы допустить некоторое ухудшение алгебраической степени точности и следующей из неё оценки погрешности (и без того весьма высоких для формул Гаусса). Основная идея формул Гаусса может быть с успехом применена к построению таких квадратурных формул, которые называются *квадратурами Маркова*.

[2, 13, 28, 74] (иногда используют также термин *квадратуры Лобатто* [1, 41]).

## 2.15e Теорема И.П. Мысовских

Построение квадратурных формул Гаусса основывалось на оптимизации алгебраической степени точности квадратур. Эта идея может быть модифицирована и приспособлена к другим ситуациям, когда точность результата для алгебраических полиномов уже не является наиболее адекватным мерилом качества квадратурной формулы. Например, можно развивать квадратуры наивысшей *тригонометрической степени точности*, которые будут точными на тригонометрических полиномах, т. е функциях вида

$$\frac{a_0}{2} + \sum_{k=1}^m (a_k \cos kx + b_k \sin kx).$$

Ясно, что они окажутся практичеснее при вычислении интегралов от осциллирующих и периодических функций [28, 79].

При рассмотрении тригонометрических полиномов и связанных с ними вопросов обычно считают областью рассмотрения интервал периода основных тригонометрических функций, т. е.  $[0, 2\pi]$ . Ясно, что это допущение непринципиально и является делом технического удобства. Предположим, что для численного интегрирования мы применяем квадратурную формулу вида

$$\int_0^{2\pi} f(x) dx \approx \sum_{k=1}^n c_k f(x_k).$$

Говорят, что эта формула имеет *тригонометрическую степень точности*, равную  $m$ , если она точна для любого тригонометрического полинома степени  $m$  и не точна для полиномов степени  $m+1$ .

**Предложение 2.15.3** *Тригонометрическая степень точности квадратурной формулы, построенной по  $n$  узлам, не превосходит  $n-1$ .*

Доказательство опускается, читатель может найти его в публикациях [28, 79]. Интересно сравнить этот результат с леммой Кеплера (предложение 2.15.1) и другими известными фактами, касающимися алгебраической степени точности квадратур (в частности, для формул

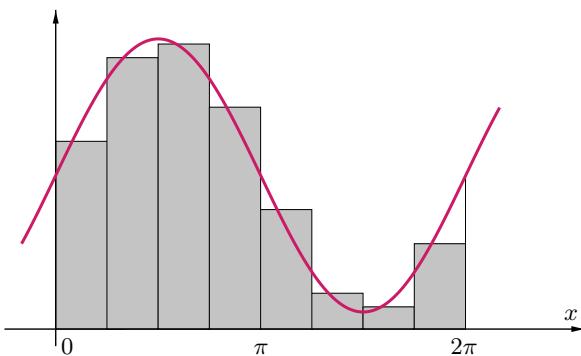


Рис. 2.48. Иллюстрация к теореме И.П. Мысовских: составная квадратурная формула прямоугольников для синусоиды

Гаусса). Это сравнение показывает, что ограничения на тригонометрическую степень точности существенно более сильны, чем на алгебраическую степень точности. Но справедлив следующий замечательный результат:

**Теорема И.П. Мысовских** [79] *Составная квадратурная формула прямоугольников*

$$\int_0^{2\pi} f(x) dx \approx \frac{2\pi}{n} \sum_{k=1}^n f\left(\check{x} + \frac{2\pi(k-1)}{n}\right),$$

где  $\check{x}$  — произвольная точка из подинтервала  $[0, 2\pi/n]$ ,

и только она является квадратурной формулой наивысшей тригонометрической степени точности  $n - 1$ .

Доказательство теоремы нетривиально, его можно увидеть в оригинальной статье [79] или в книге [28].<sup>31</sup> При построении этой составной формулы прямоугольников важно то, что в отдельных подинтервалах узлы выбираются одинаково: все они получаются из первого узла сдвигом на величину, кратную ширине подинтервала (рис. 2.48).

<sup>31</sup>Существуют результаты о методе Ньютона для решения нелинейных уравнений, которые иногда тоже называют «теоремами Мысовских».

Внимательный взгляд на рис. 2.48 помогает также понять причину точного взятия интеграла от тригонометрического полинома. Неформально говоря, при выбранном расположении узлов достигаемая на одних подинтервалах отрицательная погрешность компенсируется положительной на других, так что в целом формула в самом деле получается очень точной. Иными словами, составной характер квадратурной формулы позволяет здесь достичь качественно лучших характеристик в сравнении с отдельными элементарными квадратурами.<sup>32</sup>

Примерно такова же причина того, что составная квадратурная формула трапеций, полученная при равномерном разбиении интервала интегрирования  $[0, 2\pi]$  на  $n$  подинтервалов ширины  $2\pi/n$ , является точной для тригонометрических полиномов степени  $m < n$ .

## 2.16 Метод неопределённых коэффициентов

Опишем ещё один, отличный от интерполяционного, способ построения квадратурных формул

$$\int_a^b f(x) dx \approx \sum_{k=0}^n c_k f(x_k), \quad (2.150)$$

где некоторые (или все) узлы  $x_0, x_1, \dots, x_n$  и/или весовые коэффициенты  $c_0, c_1, \dots, c_n$  заданы по условию задачи.

Если квадратурная формула является интерполяционной и заданы её узлы, то, как мы видели в § 2.13д, весовые коэффициенты  $c_k$  могут быть вычислены по формулам (2.164) как интегралы от базисных интерполяционных полиномов Лагранжа. Но это не единственный возможный способ определения весов.

Как сами узлы  $x_k$ , так и весовые коэффициенты  $c_k$  можно найти из условия зануления погрешности равенства (2.150) для какого-то «достаточно представительного» набора несложного интегрируемых пробных функций  $g_l(x)$ ,  $l = 1, 2, \dots$ , линейной комбинацией которых можно «достаточно хорошо» приблизить подинтегральную функцию.

---

<sup>32</sup> «Целое больше, чем сумма его частей» (Аристотель, «Метафизика»).

Каждое отдельное равенство вида

$$\sum_{k=0}^n c_k g_l(x_k) = \int_a^b g_l(x) dx, \quad l = 1, 2, \dots,$$

является уравнением на неизвестные  $x_k$  и  $c_k$ , и потому, выписывая эти соотношения, мы получаем систему уравнений. В общем случае она является системой нелинейных уравнений, которая линейна относительно весов  $c_0, c_1, \dots, c_n$ . Решив её, определим желаемые узлы и/или веса, т. е. построим квадратурную формулу (2.150). В этом суть *метода неопределённых коэффициентов*. Он идейно аналогичен, таким образом, методу неопределённых коэффициентов для построения формул численного дифференцирования из § 2.8г.

В качестве пробных функций  $g_l(x)$ ,  $l = 1, 2, \dots$ , часто берут алгебраические полиномы, тригонометрические полиномы, семейства экспоненциальных функций с разными коэффициентами в показателях и т. д. Фактически именно так мы и поступали, выписывая для определения квадратур Гаусса систему уравнений (2.170), и в той ситуации не были заданы ни узлы формулы, ни её веса. Напомним, что эта система для нахождения весовых коэффициентов и узлов имеет вид

$$\left\{ \begin{array}{l} c_0 + c_1 + \dots + c_n = b - a, \\ c_0 x_0 + c_1 x_1 + \dots + c_n x_n = \frac{1}{2}(b^2 - a^2), \\ \vdots \qquad \vdots \qquad \ddots \qquad \vdots \qquad \vdots \\ c_0 x_0^m + c_1 x_1^m + \dots + c_n x_n^m = \frac{1}{m+1}(b^{m+1} - a^{m+1}). \end{array} \right. \quad (2.184)$$

Если некоторые из  $c_0, c_1, \dots, c_n, x_0, x_1, \dots, x_n$  уже известны, то реально нужно решать подсистему уравнений из (2.184), образованную первыми уравнениями, число которых, для определённости, должно совпадать с числом неизвестных параметров квадратуры.

Нетрудно показать, что если узлы  $x_0, x_1, \dots, x_n$  все заданы равномерно расположеными на интервале интегрирования, то в результате решения (2.184) относительно  $c_0, c_1, \dots, c_n$  получаются знакомые нам квадратурные формулы Ньютона–Котеса.

**Пример 2.16.1** Построим методом неопределённых коэффициентов квадратурную формулу

$$\int_{-1}^1 f(x) dx \approx c_0 f(x_0) + c_1 f\left(\frac{1}{2}\right)$$

с двумя узлами  $x_0$  и  $x_1$ , из которых второй фиксирован:  $x_1 = 1/2$ . Критерием построения выберем достижение наивысшей алгебраической степени точности, так что в качестве пробных функций будем брать последовательные степени переменной, начиная с нулевой.

Соответствующая система уравнений (2.184) выглядит следующим образом:

$$\begin{cases} c_0 + c_1 = 2, \\ c_0 x_0 + \frac{1}{2} c_1 = 0, \\ c_0 x_0^2 + \frac{1}{4} c_1 = \frac{1}{12}, \\ c_0 x_0^3 + \frac{1}{8} c_1 = 0, \\ \dots \quad \dots \quad \dots . \end{cases}$$

Из первого уравнения следует, что  $c_1 = -c_0 + 2$ . Подставив это выражение во второе и третье уравнения, получим

$$\begin{cases} c_0 x_0 - \frac{1}{2} c_0 + 1 = 0, \\ c_0 x_0^2 - \frac{1}{4} c_0 + \frac{5}{12} = 0. \end{cases} \quad (2.185)$$

Умножив первое из этих равенств на  $x_0$  и вычитая из него второе, будем иметь

$$-\frac{1}{2} c_0 x_0 + \frac{1}{4} c_0 + x_0 - \frac{5}{12} = 0.$$

Наконец, удвоим обе части полученного равенства и сложим с первым равенством из (2.185), тогда

$$2x_0 + \frac{1}{6} = 0 \quad \Rightarrow \quad x_0 = -\frac{1}{12}.$$

Подставляя найденное значение  $x_0$  во второе уравнение системы, получим вместе с первым уравнением  $2 \times 2$ -систему линейных уравнений относительно  $c_0$  и  $c_1$ . Она несложно решается, что даёт  $c_0 = 12/7$  и  $c_1 = 2/7$ .

Таким образом, искомая квадратурная формула имеет вид

$$\int_{-1}^1 f(x) dx \approx \frac{12}{7} f\left(\frac{1}{12}\right) + \frac{2}{7} f\left(\frac{1}{2}\right).$$

Третье и последующие уравнения системы (2.16) уже не выполняются, так что в целом формула имеет вторую алгебраическую степень точности. Для двух узлов это совсем неплохо: напомним, что аналогичная формула трапеций имеет алгебраическую степень точности 1. ■

**Пример 2.16.2** Предположим, что для вещественной функции  $f(t)$  известны её значение в нуле, при  $t = 0$ , а также значения её производной — мгновенной скорости изменения — на интервале  $[0, T]$ . Нужно численно найти интеграл от  $f(t)$  на  $[0, T]$ , т. е. построить квадратурную формулу, использующую значения  $f(0)$  и производной функции в каких-то точках интервала  $[0, T]$ .

Подобная задача возникает, к примеру, при необходимости определить перемещение за время  $T$  для тела, движущегося вдоль некоторой прямой, если известна скорость тела в начальный момент времени (в «момент старта»), а далее ускорение движения этого тела измеряется встроенным акселерометром. В самом деле, если  $f(t)$  — скорость тела в момент  $t$ , то его перемещение за время  $[0, T]$  равно интегралу

$$\int_0^T f(t) dt.$$

Производная  $f'(t)$  — это ускорение тела в момент  $t$ , известное из показаний акселерометра. Фактически описанная выше задача является упрощённой версией математических задач, решаемых при инерциальной навигации летательных аппаратов, ракет, судов и т. п.

Рассмотрим простейшую ситуацию, когда кроме узла  $t_0 = 0$ , в котором известно значение функции, мы можем использовать только ещё один узел  $t_1$ , так что конструируемая квадратурная формула должна иметь вид

$$\int_0^T f(t) dt \approx c_0 f(0) + c_1 f'(t_1).$$

Для нахождения  $c_0$ ,  $c_1$  и  $t_1$  применим метод неопределённых коэффициентов.

Если квадратурная формула точна на константах, т. е. на полиномах нулевой степени, то, подставляя в неё  $f(t) = 1$ , получим

$$T = c_0 \cdot 1 + c_1 \cdot 0,$$

откуда  $c_0 = T$ .

Если квадратурная формула точна на полиномах первой степени, то, подставляя в неё  $f(t) = t$ , получим

$$T^2/2 = c_0 \cdot 0 + c_1 \cdot 1,$$

откуда  $c_1 = T^2/2$ .

Если квадратурная формула точна на полиномах второй степени, то, подставляя в неё  $f(t) = t^2$ , получим

$$T^3/3 = c_0 \cdot 0 + c_1 \cdot (2t_1),$$

откуда  $2c_1 t_1 = T^3/3$ . Следовательно,  $t_1 = T/3$ .

Итак, искомая квадратурная формула имеет вид

$$\int_0^T f(t) dt \approx T f(0) + \frac{T^2}{2} f'\left(\frac{T}{3}\right).$$

Для  $f(t) = t^3$ , как нетрудно проверить, она уже не является точной, так что алгебраическая степень точности построенной формулы — два. Отметим, что при другом расположении узла  $t_1$  степень точности будет меньше.

Если применить построенную квадратурную формулу к вычислению интеграла из примера 2.15.1, то получим

$$\int_0^{\pi/2} \cos x dx \approx (\pi/2) \cdot \cos 0 + \frac{(\pi/2)^2}{2} \left(-\sin\left(\frac{\pi/2}{3}\right)\right) = 0.95395,$$

что гораздо точнее результата формулы трапеций, хотя заметно уступает формулам Симпсона и Гаусса с двумя узлами. ■

## 2.17 Сходимость квадратур

С теоретической точки зрения интересен вопрос о сходимости квадратур при неограниченном возрастании числа узлов. Иными словами, верно ли, что

$$\sum_{k=0}^n c_k f(x_k) \rightarrow \int_a^b f(x) dx$$

при  $n \rightarrow \infty$  (здесь узлы и веса квадратурных формул нумеруются с нуля)?

Похожий вопрос вставал при исследовании интерполяционного процесса, и мы обсуждали его в § 2.5. Но в случае квадратурных формул

помимо бесконечной треугольной матрицы узлов

$$\begin{pmatrix} x_0^{(0)} & & & & \cdots \\ x_0^{(1)} & x_1^{(1)} & & & \cdots \\ x_0^{(2)} & x_1^{(2)} & x_2^{(2)} & & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}, \quad (2.186)$$

таких что  $x_k^{(n)}$  лежат на интервале интегрирования  $[a, b]$  и  $x_i^{(n)} \neq x_j^{(n)}$  при  $i \neq j$ , необходимо задавать ещё и треугольную матрицу весовых коэффициентов квадратурных формул

$$\begin{pmatrix} c_0^{(0)} & & & & \cdots \\ c_0^{(1)} & c_1^{(1)} & & & \cdots \\ c_0^{(2)} & c_1^{(2)} & c_2^{(2)} & & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}. \quad (2.187)$$

В случае задания бесконечных треугольных матриц (2.186)–(2.187), по которым организуется приближённое вычисление интегралов на последовательности сеток, будем говорить, что на интервале  $[a, b]$  для функции  $f$  определён *квадратурный процесс*.

**Определение 2.17.1** *Квадратурный процесс, задаваемый зависящим от целочисленного параметра  $n$  семейством квадратурных формул*

$$\int_a^b f(x) dx \approx \sum_{k=0}^n c_k^{(n)} f(x_k^{(n)}), \quad n = 0, 1, 2, \dots,$$

которые определяются матрицами узлов и весов (2.186), (2.187), будем называть *сходящимся для функции  $f(x)$  на интервале  $[a, b]$ , если*

$$\lim_{n \rightarrow \infty} \sum_{k=0}^n c_k^{(n)} f(x_k^{(n)}) = \int_a^b f(x) dx,$$

*т. е. если при неограниченном возрастании числа узлов  $n$  предел результатов квадратурных формул равен точному интегралу от функции  $f$  по  $[a, b]$ .*

Необходимо оговориться, что в практическом плане вопрос о сходимости квадратур решается положительно с помощью составных формул, рассмотренных в § 2.14. При интегрировании достаточно общих функций путём построения составной квадратурной формулы всегда можно добиться сходимости приближённого значения интеграла к точному (для составной формулы прямоугольников это следует из самого определения интегрируемости по Риману). Обсуждаемый ниже круг вопросов относится больше к теоретическим качествам тех или иных «чистых» квадратурных формул, их предельному поведению при неограниченном возрастании числа узлов.

Весьма общие достаточные условия для сходимости квадратур были сформулированы и обоснованы В.А. Стекловым [90], а впоследствии Д. Пойа [104] доказал также необходимость условий В.А. Стеклова.

**Теорема 2.17.1** (теорема Стеклова–Пойа) *Квадратурный процесс, порождаемый матрицами узлов и весов (2.186), (2.187), сходится для любой непрерывной на  $[a, b]$  функции тогда и только тогда, когда*

- (1) *этот процесс сходится для полиномов,*
- (2) *суммы абсолютных значений весовых коэффициентов квадратурных формул ограничены равномерно по  $n$ , т. е. существует такая константа  $C$ , что*

$$\sum_{k=0}^n |c_k^{(n)}| \leq C \quad (2.188)$$

*для всех  $n = 0, 1, 2, \dots$*

**Доказательство.** Покажем достаточность условий теоремы. С этой целью, задавшись каким-то  $\epsilon > 0$ , найдём полином  $P_N(x)$ , который равномерно с погрешностью  $\epsilon$  приближает непрерывную подинтегральную функцию  $f(x)$  на рассматриваемом интервале  $[a, b]$ . Существование такого полинома обеспечивается теоремой Вейерштрасса (см. § 2.5). Далее преобразуем выражение для остаточного члена квадратурной фор-

мульты:

$$\begin{aligned}
 R_n(f) &= \int_a^b f(x) dx - \sum_{k=0}^n c_k^{(n)} f(x_k^{(n)}) = \\
 &= \int_a^b (f(x) - P_N(x)) dx + \int_a^b P_N(x) dx - \sum_{k=0}^n c_k^{(n)} f(x_k^{(n)}) = \\
 &= \int_a^b (f(x) - P_N(x)) dx + \\
 &\quad + \left( \int_a^b P_N(x) dx - \sum_{k=0}^n c_k^{(n)} P_N(x_k^{(n)}) \right) + \\
 &\quad + \sum_{k=0}^n c_k^{(n)} (P_N(x_k^{(n)}) - f(x_k^{(n)})).
 \end{aligned}$$

Отдельные слагаемые полученной суммы, расположенные выше в различных строках, оцениваются при достаточно больших номерах  $n$  следующим образом:

$$\left| \int_a^b (f(x) - P_N(x)) dx \right| \leq \epsilon(b-a), \quad \text{так как } P_N(x) \text{ приближает } f(x) \text{ равномерно с погрешностью } \epsilon \text{ на интервале } [a, b];$$

$$\left| \int_a^b P_N(x) dx - \sum_{k=0}^n c_k^{(n)} P_N(x_k^{(n)}) \right| \leq \epsilon, \quad \text{так как квадратуры сходятся на полиномах};$$

$$\left| \sum_{k=0}^n c_k^{(n)} (P_N(x_k^{(n)}) - f(x_k^{(n)})) \right| \leq \epsilon \sum_{k=0}^n |c_k^{(n)}| \leq \epsilon C \quad \text{в силу (2.188).}$$

Поэтому в целом, если  $n$  достаточно велико, имеем

$$|R_n(x)| \leq \epsilon(b-a+1+C).$$

Это и означает сходимость рассматриваемого квадратурного процесса.

Доказательство необходимости условия теоремы Стеклова–Пойя помимо оригинальной статьи [104] можно найти в книгах [2, 29]. ■

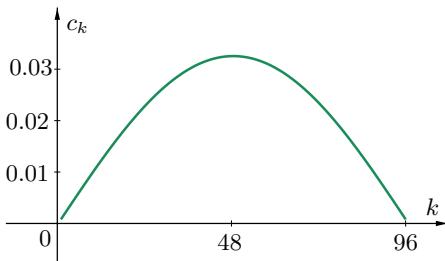


Рис. 2.49. Зависимость весовых коэффициентов от номера для квадратуры Гаусса 96-го порядка (числовые данные взяты из справочника [41])

В формулировке теоремы фигурирует величина (2.151)

$$\sum_{k=0}^n |c_k| \quad (2.151)$$

— сумма абсолютных значений весов, которая, как мы видели в § 2.13а, является коэффициентом увеличения погрешности в данных и играет очень важную роль при оценке качества различных квадратурных формул. В § 2.13е уже упоминался результат Р.О. Кузьмина [68] о том, что для формул Ньютона–Котеса величина (2.151) неограниченно увеличивается с ростом числа узлов  $n$ . Как следствие, на произвольных непрерывных функциях эти квадратурные формулы сходимостью не обладают.

Для квадратурных формул Гаусса ситуация иная, и сумма (2.151) ограничена для них равномерно по  $n$ . Для обоснования этого факта покажем сначала, что справедливо

**Предложение 2.17.1** *Весовые коэффициенты квадратурных формул Гаусса положительны.*

Иллюстрацией этого утверждения может служить рис. 2.49. Он показывает также плавное изменение весового коэффициента формулы Гаусса в зависимости от его номера, что резко контрастирует с поведением весов формул Ньютона–Котеса (см. табл. 2.2 и рис. 2.46).

**Доказательство.** Ранее мы уже выводили для весов интерполяционных квадратурных формул выражение (2.164). Зафиксировав индекс

$i \in \{1, 2, \dots, n\}$ , дадим другое явное представление для весового коэффициента  $c_i$  квадратурной формулы Гаусса, из которого и будет следовать доказываемое предложение.

Пусть  $x_1, x_2, \dots, x_n$  — узлы квадратурной формулы Гаусса на интервале интегрирования  $[a, b]$ . Так как формулы Гаусса имеют алгебраическую степень точности  $2n - 1$ , для полинома

$$\Pi_i(x) = ((x - x_1) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n))^2$$

степени  $2(n - 1)$  должно выполняться точное равенство

$$\int_a^b \Pi_i(x) dx = \sum_{k=1}^n c_k \Pi_i(x_k). \quad (2.189)$$

Но  $\Pi_i(x_k) = 0$  при  $i \neq k$  по построению полинома  $\Pi_i$ , так что от суммы справа в (2.189) остаётся лишь одно слагаемое  $c_i \Pi_i(x_i)$ :

$$\int_a^b \Pi_i(x) dx = c_i \Pi_i(x_i).$$

Следовательно,

$$c_i = \int_a^b \Pi_i(x) dx / \Pi_i(x_i).$$

Далее,  $\Pi_i(x) > 0$  всюду на интервале интегрирования  $[a, b]$ , за исключением конечного числа точек, и потому положителен интеграл в числителе выписанного выражения. Кроме того,  $\Pi_i(x_i) > 0$ , откуда можно заключить, что  $c_i > 0$ . ■

Напомним, что сумма весов формул Гаусса равна длине интервала интегрирования (как и для всех интерполяционных квадратурных формул, см. § 2.13д). Следовательно, величина (2.151) при этом ограничена, и квадратурный процесс по формулам Гаусса всегда сходится. В целом можно отметить, что ситуация со сходимостью квадратур оказывается более благоприятной, чем для интерполяционных процессов.

## 2.18 Вычисление интегралов методом Монте-Карло

В *методе Монте-Карло*, называемом также *методом статистических испытаний*, искомое решение задачи представляется в виде какой-

либо вероятностной характеристики специально построенного случайногопроцесса.<sup>33</sup> Затем этот процесс моделируется, с помощью ЭВМ или какими-то другими средствами, и по его реализациям вычисляется статистическая оценка нужной характеристики, т. е. оценка решения задачи. Наиболее часто решение задач представляется так называемым математическим ожиданием (средним значением) специально подобранный случайной величины.

В качестве примера рассмотрим задачу вычисления определённого интеграла

$$\int_a^b f(x) dx \quad (2.190)$$

от непрерывной функции  $f(x)$ . Согласно известной из интегрального исчисления теореме о среднем [12, 40]

$$\int_a^b f(x) dx = (b - a) f(c)$$

для некоторой точки  $c \in [a, b]$ . Смысл «средней точки»  $c$  можно понять глубже с помощью следующего рассуждения. Пусть интервал интегрирования  $[a, b]$  разбит на  $N$  равных подинтервалов. По определению интеграла Римана, если  $x_i$  — точки из этих подинтервалов, то

$$\int_a^b f(x) dx \approx \sum_{i=1}^N \frac{b-a}{N} f(x_i) = (b-a) \cdot \frac{1}{N} \sum_{i=1}^N f(x_i)$$

для достаточно больших  $N$ . Сумма в правой части — это произведение ширины интервала интегрирования  $(b-a)$  на среднее арифметическое значений подинтегральной функции  $f$  в точках  $x_i$ ,  $i = 1, 2, \dots, N$ . Таким образом, интеграл от  $f(x)$  по  $[a, b]$  есть не что иное, как «среднее значение» функции  $f(x)$  на интервале  $[a, b]$ , умноженное на ширину этого интервала.

Но при таком взгляде на искомый интеграл нетрудно заметить, что «среднее значение» функции  $f(x)$  можно получить каким-либо существенно более эффективным способом, чем простое увеличение количества равномерно расположенных точек  $x_i$ . Например, можно попытаться раскидывать эти точки случайно по  $[a, b]$ , но «приблизительно

---

<sup>33</sup>Название «метод Монте-Карло» происходит от географического названия средиземноморского городка, известного своими игорными заведениями, где, как считается, случайные явления играют доминирующую роль.

равномерно». Резон в таком образе действий следующий: случайный, но «равномерно случайный», выбор точек  $x_i$  позволит в пределе иметь то же «среднее значение» функции, но, возможно, полученное быстрее, так как при случайном бросании есть надежда, что будут легче учтены почти все «представительные» значения функции на  $[a, b]$ .

Для формализации высказанных идей целесообразно привлечь аппарат теории вероятностей. Эта математическая дисциплина исследует случайные явления, которые подчиняются свойству «статистической устойчивости» и обнаруживают закономерности поведения в больших сериях повторяющихся испытаний. Одними из основных понятий теории вероятностей являются понятия *вероятности, случайной величины и её функции распределения*.

Вероятность — это числовая характеристика степени возможности появления рассматриваемого события в тех или иных условиях, могущих повторяться большое (потенциально неограниченное) число раз. Вероятность, как правило, отражает относительную частоту (частость) интересующего нас события, которая обычно устанавливается в большой серии испытаний. Случайной величиной называется переменная величина, значения которой зависят от случая и для которой определена так называемая функция распределения вероятностей. В свою очередь, функция распределения показывает вероятность появления тех или иных значений этой случайной величины. Конкретное значение, которое случайная величина принимает в результате отдельного опыта (испытания), обычно называют *реализацией* случайной величины.

Случайные и «приблизительно равномерные» точки моделируются так называемым равномерным вероятностным распределением, в котором при большом количестве испытаний (реализаций) в любые подинтервалы исходного интервала  $[a, b]$ , имеющие равную ширину, попадает примерно одинаковое количество точек. На этом пути приходим к простейшему методу Монте-Карло для вычисления определённого интеграла (2.190), приведённому в табл. 2.4.

На языке теории вероятностей интеграл (2.190) является математическим ожиданием случайной величины  $f(\xi)$ , если  $\xi$  — равномерно распределённая случайная величина. Из этого факта и из закона больших чисел выводится так называемая сходимость по вероятности получаемой оценки к точному значению интеграла при  $N \rightarrow \infty$  [34].

Получение равномерно распределённой случайной величины (как и других случайных распределений) является не вполне тривиальной

Таблица 2.4. Простейший метод Монте-Карло  
для вычисления определённого интеграла

фиксируем натуральное число  $N$ ;

организуем реализации  $\xi_i$ ,  $i = 1, 2, \dots, N$ , для случайной величины  $\xi$ , имеющей на интервале  $[a, b]$  равномерное вероятностное распределение;

вычисляем значения подинтегральной функции  $f(\xi_i)$ ;

$$\left( \text{искомый интеграл от } f \text{ по } [a, b] \right) \leftarrow \frac{b - a}{N} \cdot \sum_{i=1}^N f(\xi_i).$$

задачей. Но она удовлетворительно решена на существующем уровне развития вычислительной техники и информатики. Так, практически во всех современных языках программирования имеются средства для моделирования простейших случайных величин, в частности равномерного распределения на заданном интервале. Обычно соответствующая функция имеет имя `rand` (от английского слова `random` — случайный) или производные от него.

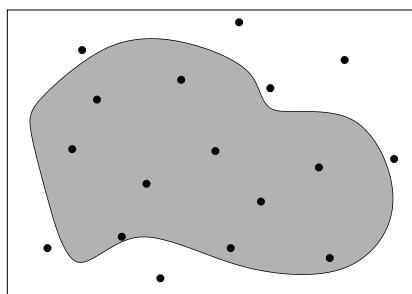


Рис. 2.50. Вычисление площади области методом Монте-Карло

Рассмотрим теперь задачу определения площади фигуры с криволинейными границами (рис. 2.50). Погрузим её в прямоугольник со

сторонами, параллельными координатным осям, имеющий известные размеры, и станем случайным образом раскидывать точки внутри этого прямоугольника. Ясно, что при равномерном распределении случайных бросаний вероятность попадания точки в рассматриваемую фигуру равна отношению площадей этой фигуры и объемлющего её прямоугольника. С другой стороны, это отношение будет приблизительно равно относительной доле количества точек, которые попали в фигуру. Оно может быть вычислено в достаточно длинной серии случайных бросаний точек в прямоугольник.

На основе сформулированной выше идеи можно реализовать ещё один способ вычисления интеграла от неотрицательной функции одной переменной. Помещаем криволинейную трапецию, ограниченную графиком интегрируемой функции, в прямоугольник на плоскости  $Oxy$  (рис. 2.51). Затем организуем равномерное случайное бросание точек в этом прямоугольнике и подсчитываем относительную частоту точек, попадающих ниже графика интегрируемой функции, т. е. в интересующую нас криволинейную трапецию. Искомый интеграл равен произведению эмпирически найденной относительной частоты на площадь большого прямоугольника.

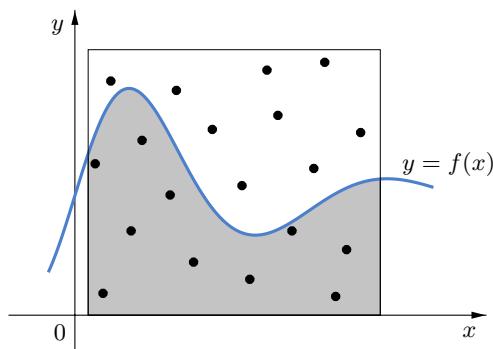


Рис. 2.51. Один из способов приближённого вычисления определённого интеграла методом Монте-Карло

Произвольную подинтегральную функцию всегда можно сделать неотрицательной, прибавив к ней достаточно большую константу. Затем из найденного интеграла следует лишь вычесть поправку, которая учитывает вклад этой константы.

Результат вычислений по методу Монте-Карло сам является слу-

чайной величиной, и два результата различных решений одной и той же задачи интегрирования, полученные описанными выше способами, вообще говоря, будут отличаться друг от друга. Можно показать [34], что второй (геометрический) способ вычисления интеграла методом Монте-Карло уступает по качеству результатов первому способу, основанному на нахождении «среднего значения» функции, так как среднеквадратичный разброс получаемых оценок у него больше.<sup>34</sup>

Сформулированные выше идеи и основанные на них алгоритмы в действительности применимы для интегрирования функций от произвольного количества переменных. Вероятностные оценки погрешности оказываются пропорциональными  $1/\sqrt{N}$ , где  $N$  — количество испытаний, т. е. вычислений подинтегральной функции в первом способе и бросаний точки в прямоугольнике в двух последних алгоритмах. Замечательное свойство этих оценок состоит в том, что они не зависят от размерности  $n$  пространства, в котором берётся интеграл, тогда как для традиционных детерминистских методов интегрирования оценки погрешности ухудшаются с ростом  $n$ . Начиная с 7–8 переменных методы Монте-Карло начинают уже превосходить по своей эффективности классические кубатурные формулы и являются сегодня основным методом вычисления многомерных интегралов.

Ещё одно хорошее свойство методов Монте-Карло для нахождения определённых интегралов — их нечувствительность к гладкости интегрируемой функции. В то же время порядок точности традиционных квадратурных формул очень существенно зависит от гладкости подинтегральной функции.

В заключение параграфа — краткий исторический очерк. Идея моделирования случайных явлений и его применения для решения различных задач очень стара. В современной истории науки использование статистического моделирования для решения конкретных задач можно отсчитывать с конца XVIII века, когда Ж.-Л. Бюффон (в 1777 году) предложил способ определения числа  $\pi$  с помощью случайных бросаний иглы на бумагу, разграфлённую параллельными линиями.<sup>35</sup> Тем не менее идея использования случайности при решении прикладных задач не получила большого развития вплоть до Второй мировой войны, т. е. до середины XX века.

<sup>34</sup> В теории вероятностей квадрат среднеквадратичного отклонения случайной величины от среднего значения называется *дисперсией*.

<sup>35</sup> Наиболее известная «докомпьютерная» реализация метода Бюффона была осуществлена американским астрономом А. Холлом [100].

В 1944 году в связи с работами по созданию атомной бомбы в США, поставившими ряд очень больших и сложных задач, С. Улам и Дж. фон Нейман предложили широко использовать для их решения статистическое моделирование и аппарат теории вероятностей.<sup>36</sup> Этому способствовало появление к тому времени электронных вычислительных машин, позволивших быстро выполнять многократные статистические испытания (Дж. фон Нейман тоже принимал активное участие в создании первых цифровых ЭВМ).

С конца 40-х годов XX века начинается широкое развитие метода Монте-Карло и методов статистического моделирования во всём мире. В настоящее время их успешно применяют для решения самых разнообразных задач практики (см., к примеру, [34, 76] и цитированную там литературу). Хороший популярный очерк методов Монте-Карло читатель может найти, например, в [13, 80], а сравнительный анализ метода Монте-Карло и традиционных детерминистских численных методов даётся в учебнике [1].

## 2.19 Правило Рунге для оценки погрешности

Предположим, что нам необходимо численно найти интеграл или производную функции, либо решение дифференциального или интегрального уравнения, т. е. решить какую-то задачу, где фигурирует сетка на интервале вещественной оси или в пространстве большего числа измерений. Пусть для решения этой задачи применяется численный метод порядка  $p$ , так что главный член его погрешности равен  $Ch^p$ , где  $h$  — шаг рассматриваемой сетки, а  $C$  — величина, напрямую от  $h$  не зависящая. Как правило, значение  $C$  не известно точно, и его нахождение непосредственно из исходных данных задачи является делом трудным и малоперспективным. Мы могли видеть, к примеру, что для задач интерполяции и численного интегрирования выражение для этой константы вовлекает оценки для производных высоких порядков от рассматриваемой функции либо её разделённые разности. Во многих случаях их практическое вычисление не представляется возможным, так что оценки эти носят главным образом теоретический характер.

---

<sup>36</sup>Интересно, что примерно в те же самые годы в СССР решение аналогичных задач Советского атомного проекта было успешно выполнено другими методами.

Аналогична ситуация и с другими задачами вычислительной математики и погрешностями их решения.

К. Рунге принадлежит идея использовать для определения константы  $C$  в реальных вычислениях результаты нескольких расчётов на различных сетках. Далее, после того как величина  $C$  будет определена, мы можем использовать её значение для практического оценивания погрешности приближённых решений нашей задачи, которые получаются с помощью выбранного численного метода.

Предположим для простоты анализа, что численные решения рассматриваемой задачи рассчитаны на сетках с шагом  $h$  и  $h/2$  и равны соответственно  $\mathcal{I}_h$  и  $\mathcal{I}_{h/2}$ , а точное решение есть  $\mathcal{I}$ . Тогда

$$\mathcal{I}_h - \mathcal{I} \approx Ch^p,$$

$$\mathcal{I}_{h/2} - \mathcal{I} \approx C \left(\frac{h}{2}\right)^p = C \frac{h^p}{2^p}.$$

Вычитая второе равенство из первого, получим

$$\mathcal{I}_h - \mathcal{I}_{h/2} \approx Ch^p - C \frac{h^p}{2^p} = Ch^p \frac{2^p - 1}{2^p},$$

так что

$$C \approx \frac{2^p}{2^p - 1} \cdot \frac{\mathcal{I}_h - \mathcal{I}_{h/2}}{h^p}. \quad (2.191)$$

Зная константу  $C$  и порядок точности используемого метода, можно уже находить оценку погрешности рассчитанных решений  $\mathcal{I}_h$ ,  $\mathcal{I}_{h/2}$  или любых других

Правило Рунге работает плохо, если главный член погрешности  $Ch^p$  не доминирует над последующими членами её разложения, которые соответствуют  $(p+1)$ -й и более высоким степеням шага сетки  $h$ . Это происходит, как правило, для сильно меняющихся решений.

Порядок точности численного метода, как правило, бывает известен из теории, но на практике встречаются и нестандартные ситуации. Рассмотрим

**Пример 2.19.1** Найдём численно интеграл

$$\int_0^1 \sqrt{x} \, dx$$

с помощью составной квадратурной формулы Симпсона.

Точное значение интеграла легко вычислить с помощью формулы Ньютона–Лейбница, и оно равно  $2/3$ .

Написав на каком-нибудь из языков программирования несложную программу, реализующую составную формулу Симпсона, получим в результате расчётов в арифметике двойной точности примерно следующее (в результатах, даваемых квадратурной формулой, показаны пять значащих цифр, а в погрешности — три):

Количество подинтервалов	Приближённое значение интеграла	Погрешность
1	0.63807	$-2.86 \cdot 10^{-2}$
2	0.65653	$-1.01 \cdot 10^{-2}$
4	0.66308	$-3.59 \cdot 10^{-3}$
8	0.66540	$-1.27 \cdot 10^{-3}$
16	0.66622	$-4.48 \cdot 10^{-4}$
32	0.66651	$-1.59 \cdot 10^{-4}$
64	0.66661	$-5.61 \cdot 10^{-5}$

Зная точное значение интеграла, можно просто по определению рассчитать порядок точности составной квадратурной формулы Симпсона, который она демонстрирует в этой задаче. Порядок точности получается равным примерно 1.5, что разительно отличается от теоретически выведенного порядка точности 4 (см. § 2.14б). В чём причина такого несоответствия?

Вспомним, что теоретическая оценка погрешности квадратурной формулы Симпсона, которой мы занимались в § 2.13г, привлекает значение максимума 4-й производной подинтегральной функции. Но в данной ситуации оно бесконечно, и потому теоретическая оценка (2.162) здесь неприменима.

При практическом применении правила Рунге для оценки погрешности интеграла в качестве значения  $p$  нужно брать, конечно же, реальный порядок точности, то есть 1.5, а не теоретический порядок 4. Тогда мы получим реалистичную оценку погрешности

Например, возьмём  $h = 1/8$  и  $h/2 = 1/16$ . Тогда формула (2.191) даёт  $C = -0.028702$ . Следовательно, правило Рунге приближённо оценивает погрешность как

$$Ch^p = -0.0012685,$$

и это отлично согласуется с реальным значением погрешности результата составной формулы Симпсона с 8 подинтервалами. Если бы мы взяли  $p = 4$ , оценка погрешности была бы далёкой от действительности. ■

## Литература к главе 2

### Основная

- [1] Бахвалов Н.С., Жидков Н.П., Кобельков Г.М. *Численные методы*. – М.: Бином, 2003, а также другие издания этой книги.
- [2] Березин И.С., Жидков Н.П. *Методы вычислений*. Т. 1–2. – М.: Наука, 1966.
- [3] Брадис В.М. *Четырехзначные математические таблицы*. – М.: Дрофа, 2010, а также более ранние издания.
- [4] Вержбицкий В.М. *Численные методы. Части 1–2*. – М.: «Оникс 21 век», 2005.
- [5] Волков Е.А. *Численные методы*. – М.: Наука, 1987.
- [6] Гавриков М.Б., Таюрский А.А. *Функциональный анализ и вычислительная математика*. – М.: URSS, Ленанд, 2016.
- [7] Гончаров В.Л. *Теория интерполирования и приближения функций*. – М.: ГИТТЛ, 1954.
- [8] Даугавет И.К. *Введение в теорию приближения функций*. – Л.: Издательство Ленинградского университета, 1977.
- [9] Демидович Б.П., Марон А.А. *Основы вычислительной математики*. – М.: Наука, 1970.
- [10] Демидович Б.П., Марон А.А., Шувалова Э.З. *Численные методы анализа*. – М.: Наука, 1967.
- [11] Завьялов Ю.С., Квасов Б.И., Мирошниченко В.Л. *Методы сплайн-функций*. – М.: Наука, 1980.
- [12] Зорич В.А. *Математический анализ*. Т. 1. – М.: Наука, 1981. Т. 2. – М.: Наука, 1984, а также более поздние издания.
- [13] Калиткин Н.Н. *Численные методы*. – М.: Наука, 1978.
- [14] Ковков В.В., Шокин Ю.И. *Сплайн-функции в численном анализе*. – Новосибирск: Издательство НГУ, 1983.
- [15] Коллатц Л. *Функциональный анализ и вычислительная математика*. – М.: Мир, 1969.
- [16] Колмогоров А.Н., Фомин С.В. *Элементы теории функций и функционального анализа*. – М.: Наука, 1976, а также более поздние издания.
- [17] Кострикин А.Н. *Введение в алгебру. Часть 1. Основы алгебры*. – М.: Физматлит, 2001.

- [18] Крылов А.Н. *Лекции о приближённых вычислениях*. – М.: ГИТТЛ, 1954, а также более ранние издания.
- [19] Крылов В.И. *Приближённое вычисление интегралов*. – М.: Наука, 1967.
- [20] Крылов В.И., Бобков В.В., Монастырный П.И. *Вычислительные методы. Т. 1–2*. – М.: Наука, 1976.
- [21] Кунц К.С. *Численный анализ*. – Киев: Техника, 1964.
- [22] Курош А.Г. *Курс высшей алгебры*. – М.: Наука, 1975.
- [23] Люстерник Л.А., Червоненкис О.А., Янпольский А.Р. *Математический анализ. Вычисление элементарных функций*. – М.: ГИФМЛ, 1963.
- [24] МАК-КРАКЕН Д., ДОРН У. *Численные методы и программирование на ФОРТРАНе*. – М.: Мир, 1977.
- [25] Марков А.А. *Исчисление конечных разностей*. – Одесса: Mathesis, 1910.
- [26] Мацокин А.М., Сорокин С.Б. *Численные методы. Часть 1. Численный анализ*. – Новосибирск: НГУ, 2006.
- [27] Миньков С.Л., Миньков Л.Л. *Основы численных методов*. – Томск: Издательство научно-технической литературы, 2005.
- [28] Мысовских И.П. *Лекции по методам вычислений*. – СПб.: Издательство Санкт-Петербургского университета, 1998.
- [29] Натансон И.П. *Конструктивная теория функций*. – М.–Л.: ГИТТЛ, 1949.
- [30] Натансон И.П. *Теория функций вещественной переменной*. – М.: Наука, 1974.
- [31] Никольский С.М. *Квадратурные формулы*. – М.: Наука, 1988.
- [32] Рудин У. *Основы математического анализа*. – М.: Мир, 1976.
- [33] Самарский А.А., Гулин А.В. *Численные методы*. – М.: Наука, 1989.
- [34] Соболь И.М. *Численные методы Монте-Карло*. – М.: Наука, 1973.
- [35] Стечкин С.Б., Субботин Ю.Н. *Сплайны в вычислительной математике*. – М.: Наука, 1976.
- [36] Тихонов А.Н., Арсенин В.Я. *Методы решения некорректных задач*. – М.: Наука, 1979.
- [37] Тыртышников Е.Е. *Матричный анализ и линейная алгебра*. – М.: Физматлит, 2007.
- [38] Тыртышников Е.Е. *Методы численного анализа*. – М.: Академия, 2007.
- [39] Уиттекер Э., Робинсон Г. *Математическая обработка результатов наблюдений*. – Л.–М.: ГТТИ, 1933.
- [40] Фихтенгольц Г.М. *Курс дифференциального и интегрального исчисления. Т. 1–3*. – М.: Наука, 1966; М.: ФИЗМАТЛИТ, 2001; СПб.: Лань, 2017.

## Дополнительная

- [41] Абрамович М., Стиган И. *Таблицы специальных функций*. – М.: Наука, 1979.

- [42] Алберг Дж., Нильсон Э., Уолш Дж. Теория сплайнов и её приложения. – М.: Мир, 1972.
- [43] Александрова Н.В. История математических терминов, понятий, обозначений. Словарь-справочник. – М.: URSS, 2018.
- [44] Ахиезер Н.И. Лекции по теории аппроксимации. – М.: Наука, 1965.
- [45] Бабенко К.И. Основы численного анализа. – М.: Наука, 1986.
- [46] Баженов А.Н., Жилин С.И., Кумков С.И., Шарый С.П. Обработка и анализ интервальных данных. – Ижевск-М.: Издательство «ИКИ», 2024.
- [47] Бахвалов Н.С., Корнев А.А., Чижонков Е.В. Численные методы. Решения задач и упражнения. – М.: Дрофа, 2008.
- [48] Бейкер Дж., мл., Грейвс-Моррис П. Аппроксимации Паде. – М.: Мир, 1986.
- [49] Бердышев В.И., Петрак Л.В. Аппроксимация функций, сжатие численной информации, приложения. – Екатеринбург: УрО РАН, 1999.
- [50] Василенко В.А. Сплайн-функции: теория, алгоритмы, программы. – Новосибирск: Наука, 1983.
- [51] Волков Ю.С., Субботин Ю.Н. 50 лет задаче Шёнберга о сходимости сплайн-интерполяции // Труды Института математики и механики УрО РАН. – 2014. – Т. 20, №1. – С. 52–67.
- [52] Гельфонд А.О. Исчисление конечных разностей. – М.: Наука, 1967, а также более поздние репринтные издания.
- [53] Геронимус Я.Л. Теория ортогональных многочленов. – М.: Госуд. изд-во технико-теоретической литературы, 1950.
- [54] Голубов Б.И., Ефимов А.В., Скворцов В.А. Ряды и преобразования Уолша. Теория и приложения. – М.: Наука, 1987.
- [55] де Бор К. Практическое руководство по сплайнам. – М.: Радио и связь, 1985.
- [56] Демиденко Е.З. Оптимизация и регрессия. – М.: Наука, 1989.
- [57] Дровышевич В.И., Дымников В.П., Ривин Г.С. Задачи по вычислительной математике. – М.: Наука, 1980.
- [58] Зигмунд А. Тригонометрические ряды. Т. 1–2. – М.: Мир, 1965.
- [59] Ильин В.А., Садовничий В.А., Сендов Бл.Х. Математический анализ. Начальный курс. 2-е изд. – М.: Издательство МГУ, 1985.
- [60] Канторович Л.В. О проведении численных и аналитических вычислений на машинах с программным управлением // Известия Академии Наук Армянской ССР. – 1957. – Т. X, №2. – С. 3–16.
- [61] Кахранер Д., Моулер К., Нэш С. Численные методы и программное обеспечение. – М.: Мир, 1998.
- [62] Квасов Б.И. Методы изогеометрической аппроксимации сплайнами. – М.: Физматлит, 2006.
- [63] Коллатц Л., Крабс В. Теория приближений. Чебышёвские приближения и их приложения. – М.: Наука, 1978.

- [64] Колмогоров А.Н. К обоснованию метода наименьших квадратов // Успехи математических наук. – 1946. – Т. 1, вып 1(11). – С. 57–70.
- [65] Корнейчук Н.П. Сплайны в теории приближения. – М.: Наука, 1984.
- [66] Кронрод А.С. Узлы и весы квадратурных формул. Шестнадцатизначные таблицы. – М.: Наука, 1964.
- [67] Крылов В.И., Шульгина Л.Т. Справочная книга по численному интегрированию. – М.: Наука, 1966.
- [68] Кузьмин Р.О. К теории механических квадратур // Известия Ленинградского политехнического института им. М.И. Калинина. – 1931. – Т. 33. – С. 5–14.
- [69] Ланс Дж.Н. Численные методы для быстродействующих вычислительных машин. – М.: Издательство иностранной литературы, 1962.
- [70] Линник Ю.В. Метод наименьших квадратов и основы теории обработки наблюдений. 2-е изд. – М.: ГИФМЛ, 1962.
- [71] Локуциевский О.В., Гавриков М.Б. Начала численного анализа. – М.: ТОО «Янус», 1994.
- [72] Лоран Ж.-П. Аппроксимация и оптимизация. – М.: Мир, 1975.
- [73] Меньшиков Г.Г. Локализующие вычисления. Конспект лекций. – СПб.: СПбГУ, Факультет прикладной математики–процессов управления, 2003.
- [74] Микеладзе Ш.Е. Численные методы математического анализа. – М.: ГИТТЛ, 1953.
- [75] Милн В.Э. Численный анализ. – М.: Издательство иностранной литературы, 1951.
- [76] Михайлов Г.А., Войтишек А.В. Численное статистическое моделирование. Методы Монте-Карло. – М.: Изд. центр «Академия», 2006.
- [77] Мудров А.Е. Численные методы для ПЭВМ на языках Бейсик, Фортран и Паскаль. – Томск: МП «Раско», 1991.
- [78] Мысовских И.П. Интерполяционные кубатурные формулы. – М.: Наука, 1981.
- [79] Мысовских И.П. Квадратурные формулы наивысшей тригонометрической степени точности // Журнал вычислительной математики и математической физики. – 1985. – Т. 25, №8. – С. 1246–1252.
- [80] Нивергельт Ю., Фаррар Дж., Рейнгольд Э. Машинный подход к решению математических задач. – М.: Мир, 1977.
- [81] Никифоров А.Ф., Суслов С.К., Уваров В.Б. Классические ортогональные полиномы дискретной переменной. – М.: Наука, 1985.
- [82] Никольский С.М. Приближение функций многих переменных и теоремы вложения. – М.: Наука, 1977.
- [83] Пащковский С. Вычислительные применения многочленов и рядов Чебышёва. – М.: Наука, 1983.
- [84] Погорелов А.И. Дифференциальная геометрия. – М.: Наука, 1974.

- [85] ПРИВАЛОВ И.И. Введение в теорию функций комплексного переменного. – М.: Наука, 1977, а также другие издания.
- [86] РЕМЕЗ Е.Я. Основы численных методов чебышёвского приближения. – Киев: Наукова думка, 1969.
- [87] СЕГЁ Г. Ортогональные многочлены. – М.: Физматлит, 1962.
- [88] СКАРБОРО Дж. Численные методы математического анализа. – М.–Л.: ГТТИ, 1934.
- [89] СОВОЛЕВ С.Л. Введение в теорию кубатурных формул. – М.: Наука, 1974.
- [90] СТЕКЛОВ В.А. О приближённом вычислении определённых интегралов // Известия Академии Наук. – 1916. – Т. 10, №6. – С. 169–186.
- [91] СУЕТИН П.К. Классические ортогональные многочлены. – М.: Наука, 1979.
- [92] СТЕФЕНСЕН И.Ф. Теория интерполяции. – М.: Объединённое научно-техническое издательство НКТП СССР, 1935.
- [93] ХАНСЕН Э., УОЛСТЕР Дж.У. Глобальная оптимизация с помощью методов интервального анализа. – М.-Ижевск: Издательство «РХД», 2012.
- [94] ХАУСХОЛДЕР А.С. Основы численного анализа. – М.: Издательство иностранной литературы, 1956.
- [95] ХЕММИНГ Р.В. Численные методы. – М.: Наука, 1972.
- [96] ЯГЛОМ И.М. Комплексные числа и их применение в геометрии. – М.: Физматлит, 1963.
- [97] BOREL É. Leçons sur les fonctions de variables réelles et les développements en séries de polynômes. – Paris: Gauthier-Villars, 1905.
- [98] CHRISTOFFEL E.B. Über die Gaußische Quadratur und eine Verallgemeinerung derselben // Journal für die reine und angewandte Mathematik. – 1858. – Issue 55. – S. 61–82.
- [99] GRIEWANK A., WALTHER A. Evaluating derivatives: principles and techniques of algorithmic differentiation. Second Edition. – Philadelphia: SIAM, 2008.
- [100] HALL A. On an experimental determination of  $\pi$  // Messenger of Mathematics. – 1873. – Vol. 2. – P. 113–114.
- [101] HOLLADAY J.C. Smoothest curve approximation // Mathematical Tables and Other Aids to Computation. – 1957. – Vol. 11, No. 60. – P. 233–243.
- [102] LOBACHEVSKY N. Probabilité des résultats moyens tirés d'observations répétées // Journal für die reine und angewandte Mathematik. – 1842. – Bd. 24. – S. 164–170.
- [103] MOORE R.E., KEARFOTT R.B., CLOUD M. Introduction to interval analysis. – Philadelphia: SIAM, 2009.
- [104] POLYA G. Über Konvergenz von Quadraturverfahren // Mathematische Zeitschrift. – 1933. – Bd. 37. – S. 264–286.
- [105] RADEMACHER H. Einige Sätze über Reihen von allgemeinen Orthogonalfunktionen // Mathematische Annalen. – 1922. – Bd. 87, Nr. 1-2. – S. 112–138.

- [106] RALL L.B., REPS T.W. Algorithmic differencing // Perspectives on Enclosure Methods / U. Kulisch, R. Lohner, A. Facius (eds.) – Vienna: Springer-Verlag, 2001. – P. 133–147.
- [107] RUNGE C. Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten // *Zeitschrift für Mathematik und Physik*. – 1901. – Bd. 46. – S. 224–243.
- [108] SCHOENBERG I.J Contributions to the problem of approximation of equidistant data by analytic functions. Part A: On the problem of smoothing or graduation. A first class of analytic approximation formulae. Part B: On the problem of osculatory interpolation. A second class of analytic approximation formulae // *Quart. Appl. Math.* – 1946. – Vol. 4. – P. 45–99, 112–141.
- [109] STOER J., BULIRSCH R. *Introduction to numerical analysis*. – Berlin-Heidelberg-New York: Springer-Verlag, 1993.

## Глава 3

# Численные методы линейной алгебры

## 3.1 Задачи вычислительной линейной алгебры

Численные методы линейной алгебры — это один из классических разделов вычислительной математики, который в середине XX века вычленился даже в отдельное научное направление<sup>1</sup> в связи с бурным развитием математических вычислений на ЭВМ. Традиционный исторически сложившийся список задач вычислительной линейной алгебры по состоянию на 50–60-е годы прошлого века можно найти в капитальной книге Д.К. Фаддеева и В.Н. Фаддеевой [48]. Он включал:

- решение систем линейных алгебраических уравнений,
- вычисление определителей матриц,
- нахождение обратной матрицы,
- нахождение собственных значений и собственных векторов матриц,

а также многочисленные разновидности этих основных задач.

---

<sup>1</sup> В англоязычной учебной и научной литературе для него часто используют термин «матричные вычисления». Но он гораздо уже по объёму, не охватывая, к примеру, такую часть вычислительной линейной алгебры, как тензорные вычисления.

Но «всё течёт, всё меняется». По мере развития науки и технологий в фокусе развития вычислительной линейной алгебры оказались новые задачи. Вот как формулировал список важнейших задач в 2001 году американский математик Дж. Деммель в книге [13]:

- решение систем линейных алгебраических уравнений;
- линейная задача наименьших квадратов —  
найти вектор  $x$ , минимизирующий  $\langle Ax - b, Ax - b \rangle$   
для заданных  $m \times n$ -матрицы  $A$  и  $m$ -вектора  $b$ ;
- нахождение собственных значений и собственных  
векторов матриц;
- нахождение сингулярных чисел и сингулярных  
векторов матриц.

Последняя задача будет подробно обсуждаться ниже, в частности в § 3.2д. Вторая задача из этого списка — линейная задача наименьших квадратов — является одним из вариантов дискретной задачи о наилучшем среднеквадратичном приближении (см. § 2.10). Она возникает обычно в связи с решением переопределённых систем линейных алгебраических уравнений (СЛАУ), которые, к примеру, получаются при обработке экспериментальных данных. Мы уже занимались этой задачей в § 2.11г.

Помимо перечисленных задач к сфере вычислительной линейной алгебры относится также решение разнообразных линейных матричных уравнений, т. е. уравнений, в которых неизвестными являются матрицы [85]. Таковы матричные уравнения Сильвестра, Ляпунова и др., которые возникают, к примеру, при исследовании устойчивости решений дифференциальных уравнений, в теории автоматического управления и т. п.

С точки зрения классических разделов математики решение выписанных задач даётся вполне конструктивными способами и как будто не встречает больших затруднений:

- ▶ решение квадратной СЛАУ получается покомпонентно по формуле Крамера как частное двух определителей [9, 54], которые, в свою очередь, могут быть вычислены по явным формулам;
- ▶ для вычисления собственных значений матрицы  $A$  нужно выписать её характеристическое (вековое) уравнение  $\det(A - \lambda I) = 0$  и найти его решения  $\lambda$ .

И так далее. Но практическая реализация этих теоретических рецептов наталкивается на почти непреодолимые трудности.

К примеру, явная формула для определителя  $n \times n$ -матрицы выражает его как сумму  $n!$  слагаемых, каждое из которых есть произведение  $n$  элементов из разных строк и столбцов матрицы. Раскрытие определителя по этой формуле требует  $n! (n - 1)$  умножений и  $(n! - 1)$  сложений, т. е. всего примерно  $n! n$  арифметических операций. Поэтому при таком вычислении определителя из-за взрывного роста факториала<sup>2</sup> решение СЛАУ по правилу Крамера для  $n \approx 20-30$  делается невозможным даже на самых современных ЭВМ.

Производительность современных ЭВМ принято выражать в так называемых *флопсах* (сокращение от английской фразы *floating point operation*), и 1 флопс — это одна усреднённая арифметическая операция в арифметике с плавающей точкой в секунду (см. § 1.4). Для наиболее мощных на сегодняшний день ЭВМ скорость работы измеряется так называемым петафлопсами, т. е.  $10^{15}$  операций с плавающей точкой в секунду. Для круглого счёта (и с прицелом на перспективу) можно даже взять производительность нашего гипотетического компьютера равной 1 экзафлопс =  $10^{18}$  операций с плавающей точкой в секунду. Решение на такой вычислительной машине системы линейных алгебраических уравнений размера  $30 \times 30$  по правилу Крамера, с раскрытием определителей по явной комбинаторной формуле, потребует времени

$$30 \text{ компонент решения} \cdot \frac{30 \cdot 30! \text{ операций}}{10^{18} \text{ флопс} \cdot 3600 \frac{\text{сек}}{\text{час}} \cdot 24 \frac{\text{час}}{\text{сутки}} \cdot 365 \frac{\text{сутки}}{\text{год}}},$$

т. е. примерно  $7.57 \cdot 10^9$  лет. Для сравнения: возраст Земли в настоящее время оценивается в  $4.5 \cdot 10^9$  лет.

Конечно, непрактичность явной комбинаторной формулы для определителя осознана давно, и реальные расчёты по ней почти никто не выполняет. Вместо этой формулы можно использовать, к примеру, различные разложения по строкам или столбцам матрицы (см., к примеру, [30]). Но современные численные методы линейной алгебры для вычисления определителей или решения систем линейных уравнений работают ещё быстрее.

Обращаясь к задаче вычисления собственных значений матрицы, вспомним известную из алгебры теорему Абеля–Руффини<sup>3</sup>: для общих

---

<sup>2</sup>Напомним в этой связи известную в математическом анализе асимптотическую формулу Стирлинга —  $n! \approx \sqrt{2\pi n} (n/e)^n$ , где  $e = 2.7182818\dots$  — число Эйлера.

<sup>3</sup>Иногда её называют просто «теоремой Абеля» [64].

алгебраических уравнений степени пять и выше не существует конечной формулы, выражающей решения уравнения через коэффициенты с помощью арифметических операций и извлечения корней произвольной степени. К этому добавляются трудности в раскрытии определителя, который входит в характеристическое уравнение матрицы. Таким образом, для матриц размера  $5 \times 5$  и более мы по необходимости должны развивать для нахождения собственных значений какие-то приближённые численные методы.

Наконец, помимо неприемлемой трудоёмкости ещё одной причиной непригодности для реальных вычислений некоторых широко известных алгоритмов из «чистой математики» является сильное влияние на их результаты неизбежных погрешностей счёта и ввода данных. Например, очень неустойчиво к погрешностям решение СЛАУ по правилу Крамера.

## 3.2 Теоретическое введение

### 3.2а Необходимые сведения из линейной алгебры

Термин «вектор» имеет несколько значений. Прежде всего, это направленный отрезок на прямой, плоскости или в пространстве. Далее, термин «вектор» может обозначать упорядоченный кортеж из чисел либо объектов какой-то другой природы, расположенный вертикально (вектор-столбец) или горизонтально (вектор-строка). Таким образом, если  $a_1, a_2, \dots, a_n$  — некоторые числа, то

$$a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \text{ — это вектор-столбец,} \quad (3.1)$$

а

$$a = (a_1, a_2, \dots, a_n) \text{ — это вектор-строка.} \quad (3.2)$$

Этот смысл термина «вектор» широко используется в информатике и программировании. Наконец, векторами называются элементы абстрактных «векторных пространств», т. е. некоторых аксиоматически определяемых алгебраических систем (структур). В современной математике и её приложениях огромное применение находят, к примеру,

линейные векторные пространства, об элементах которых мы привычно говорим как о некоторых «векторах».

Напомним, что *линейное векторное пространство* или просто *линейное пространство* — это множество объектов произвольной природы с определёнными на нём операциями сложения и умножения на скаляры из некоторого поля, которые подчиняются специальным правилам (называемым также *аксиомами*). Чтобы подчеркнуть особый статус поля скаляров часто говорят про линейное векторное «пространство над полем». Полный список аксиом линейного пространства можно увидеть, например, в [6, 7, 24, 30, 43, 55, 68].

Все три перечисленных выше смысла термина «вектор» тесно связаны между собой и взаимно проникают друг в друга. Мы в равной степени будем пользоваться всеми ими, предполагая, что контекст изложения не даст повода к недоразумениям. По умолчанию, если не оговорено противное, условимся считать, что «векторами» во втором смысле являются вектор-столбцы (3.1), а сами числа  $a_1, a_2, \dots, a_n$  станем называть «компонентами» вектора  $a$ . Множество векторов вида (3.1), компоненты которых принадлежат вещественной оси  $\mathbb{R}$  или комплексной плоскости  $\mathbb{C}$ , будем обозначать через  $\mathbb{R}^n$  или  $\mathbb{C}^n$ . При этом нулевые векторы, т. е. векторы, все компоненты которых суть нули, традиционно обозначаем через « $0$ ».

Векторы-столбцы вида (3.1) с элементами из  $\mathbb{R}$  или  $\mathbb{C}$  можно складывать, вычитать, умножать на скаляры из этих же полей:

$$\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} a_1 + b_1 \\ a_2 + b_2 \\ \vdots \\ a_n + b_n \end{pmatrix}, \quad \alpha \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} \alpha a_1 \\ \alpha a_2 \\ \vdots \\ \alpha a_n \end{pmatrix}.$$

Совершенно аналогично определяются эти операции для вектор-строк (3.2). Нетрудно показать, что множества вектор-столбцов или вектор-строк с выписанными выше операциями являются линейными векторными пространствами над  $\mathbb{R}$  или  $\mathbb{C}$ . Они называются *арифметическими векторными пространствами* и играют основную роль в этой главе. Особую значимость этим пространствам придаёт то обстоятельство, что, как показывается в алгебре, конечномерные линейные векторные пространства одинаковой размерности над одним и тем же полем изоморфны друг другу, т. е. устроены структурно одинаково. Иными словами, арифметические векторные пространства являются типичными

представителями линейных векторных пространств одной с ними размерности.

Если в линейном пространстве  $L$  некоторое подмножество  $L' \subseteq L$  само образует линейное векторное пространство относительно операций, определённых на  $L$ , то  $L'$  называют *линейным подпространством* в  $L$ .

Ненулевые векторы  $a$  и  $b$  называются *коллинеарными*, если  $a = ab$  для некоторого скаляра  $\alpha$ . Иногда различают *соправленные* коллинеарные векторы, отвечающие случаю  $\alpha > 0$ , и *противоположно направленные*, для которых  $\alpha < 0$ . Нулевой вектор по определению коллинеарен любому вектору.

Вообще, в линейной алгебре, при работе в линейных векторных пространствах, большую роль играют линейные выражения вида

$$\alpha_1 v_1 + \alpha_2 v_2 + \dots + \alpha_r v_r,$$

где  $\alpha_1, \alpha_2, \dots, \alpha_r$  — некоторые скаляры, а  $v_1, v_2, \dots, v_r$  — векторы из рассматриваемого пространства. Такие выражения называются *линейными комбинациями* векторов  $v_1, v_2, \dots, v_r$ . Говорят также, что линейная комбинация *нетривиальная*, если хотя бы один из коэффициентов  $\alpha_1, \alpha_2, \dots, \alpha_r$  не равен нулю. Если все коэффициенты  $\alpha_i$  неотрицательны, а их сумма равна единице, то такую линейную комбинацию называют *выпуклой*. Название объясняется тем, что такая линейная комбинация даёт точку из выпуклой оболочки множества векторов  $v_1, v_2, \dots, v_r$ , т. е. наименьшего выпуклого множества, которое содержит всех их.

Векторы  $v_1, v_2, \dots, v_r$  называются *линейно зависимыми*, если равна нулю некоторая их нетривиальная линейная комбинация. Иначе, если любая нетривиальная линейная комбинация векторов не равна нулю, то эти векторы называются *линейно независимыми*.

*Линейной оболочкой* векторов  $v_1, v_2, \dots, v_r$  называют множество всевозможных линейных комбинаций этих векторов, т. е. наименьшее линейное подпространство, содержащее эти векторы  $v_1, v_2, \dots, v_r$ . Мы будем обозначать линейную оболочку посредством  $\text{span} \{v_1, v_2, \dots, v_r\}$ , так что

$$\text{span} \{v_1, v_2, \dots, v_r\} := \left\{ \sum_{i=1}^r \alpha_i v_i \mid \alpha_i \in \mathbb{R} \text{ или } \mathbb{C} \right\}.$$

Пусть  $L$  — линейное пространство, а  $L_1, L_2, \dots, L_m$  — его линейные подпространства. Говорят, что линейное пространство  $L$  есть *прямая*

сумма своих подпространств  $L_1, L_2, \dots, L_m$ , и обозначают

$$L = L_1 \oplus L_2 \oplus \cdots \oplus L_m,$$

если любой вектор  $x \in L$  единственным образом представляется в виде суммы  $x = x_1 + x_2 + \cdots + x_m$ , где  $x_i \in L_i$  для  $i = 1, 2, \dots, m$ .

Пусть линейное векторное пространство  $L$  представимо в виде прямой суммы двух своих подпространств  $L_1$  и  $L_2$ , т. е.  $L = L_1 \oplus L_2$ , так что любой вектор  $x \in L$  однозначно записывается в виде  $x = x_1 + x_2$ , где  $x_1 \in L_1$  и  $x_2 \in L_2$ . Тогда  $x_1$  называют *проекцией* вектора  $x$  на подпространство  $L_1$  вдоль подпространства  $L_2$ . Аналогично  $x_2$  есть проекция  $x$  на  $L_2$  вдоль  $L_1$ .

Помимо формальных определений полезно уяснить содержательный смысл понятия проекции вектора и операции проектирования. Фактически проекция — это некоторый вектор, который представляет заданный вектор пространства в более узком подпространстве, имеющем меньшую размерность. Естественно, что это представление сильно зависит от способа, которым мы хотим представлять вектор, и в нашем случае он задаётся выбором разложения пространства в прямую сумму  $L = L_1 \oplus L_2$ .

Отображение, которое ставит в соответствие каждому вектору пространства его проекцию на данное подпространство, называют *проектором*. Нетрудно показать, что это линейный оператор, который дополнительно удовлетворяет условию  $P^2 = P$ , называемому *идемпотентностью*. Верно и обратное: любой идемпотентный линейный оператор является проектором, т. е. оператором проектирования на подпространство, которое является его образом.

На линейных пространствах  $\mathbb{R}^n$  и  $\mathbb{C}^n$  можно задать *скалярное произведение векторов* — операцию, которая ставит в соответствие двум векторам скаляр, т. е. число из поля, над которым рассматривается это пространство. Далее будем обозначать скалярное произведение угловыми скобками  $\langle \cdot, \cdot \rangle$ , где операнды разделены запятой. В вещественном случае скалярное произведение формально определяется как симметричная, билинейная и положительно определённая форма, а в комплексном — это эрмитова положительно определённая форма [7, 24, 30, 43, 47]. Неформально скалярное произведение можно описать как специальную функцию от двух векторов, которая показывает их «взаимное расположение» с учётом длин этих векторов. Как должно быть известно читателю, для векторов единичной длины  $a$  и  $b$ , как направленных отрезков на плоскости или в пространстве, их скалярное

произведение характеризует угол  $\hat{ab}$  между ними, так как

$$\langle a, b \rangle = |a| \cdot |b| \cdot \cos \hat{ab}, \quad (3.3)$$

где  $|\cdot|$  означает длину вектора.

Скалярные произведения на  $\mathbb{R}^n$  и  $\mathbb{C}^n$  могут задаваться различным образом, но в качестве их стандартного вида обычно рассматривают следующие выражения:

$$\langle a, b \rangle = \sum_{i=1}^n a_i b_i \quad \text{для } a, b \in \mathbb{R}^n \quad (3.4)$$

или

$$\langle a, b \rangle = \sum_{i=1}^n a_i \bar{b}_i \quad \text{для } a, b \in \mathbb{C}^n, \quad (3.5)$$

где через  $\bar{b}_i$  обозначено комплексно-сопряжённое к  $b_i$  число. Нетрудно проверить, что в прямоугольной декартовой системе координат в  $\mathbb{R}^2$  и  $\mathbb{R}^3$  выражение (3.5) соответствует исходному определению скалярного произведения (3.3). Наличие в линейном пространстве скалярного произведения позволяет говорить о длине векторов, о величине угла между векторами, а также ввести очень важное понятие ортогональности векторов.

Векторы  $a$  и  $b$  называются *ортогональными*, если  $\langle a, b \rangle = 0$ . Для обозначения ортогональности мы будем также писать  $a \perp b$ . Геометрически в  $\mathbb{R}^2$  и  $\mathbb{R}^3$  ортогональность означает перпендикулярность этих векторов, т. е. прямой угол  $\pi/2$  между ними.

Система векторов называется *ортогональной*, если она либо состоит из одного вектора, либо все её векторы попарно ортогональны. В целом введение скалярного произведения превращает пространство  $\mathbb{R}^n$  в так называемое *евклидово пространство*, а  $\mathbb{C}^n$  — в *унитарное пространство*. Для евклидовых и унитарных пространств справедливы многие красивые и важные свойства, существенно обогащающие математические рассуждения.

*Длиной вектора*  $a$  в евклидовом или унитарном пространстве называется, согласно (3.3), величина  $|a| = \sqrt{\langle a, a \rangle}$ , т. е. квадратный корень из скалярного произведения вектора на себя (см. также § 3.3а). Векторы единичной длины, с помощью которых обычно задают направления, часто называют *нормированными*. Ясно, что любой вектор  $a$  можно сделать нормированным, умножив его на скаляр  $1/|a|$ , и это несложное преобразование называется *нормировкой*. Очень часто нормировку

вектора понимают в более общем смысле — как масштабирование к вектору, какая-то заданная норма которого, не обязательно евклидова, равна единице.

Ортогональную систему векторов, которые дополнительно нормированы, называют *ортонормальной* или *ортонормированной*.

Пусть  $M$  — непустое подмножество векторов пространства  $\mathbb{R}^n$  или  $\mathbb{C}^n$ . Совокупности всех векторов этих пространств, ортогональных к  $M$ , называются *ортогональным дополнением* множества  $M$  и обозначаются  $M^\perp$ . Нетрудно показать, что ортогональное дополнение любого непустого множества  $M$  является линейным подпространством в  $\mathbb{R}^n$  или  $\mathbb{C}^n$ . Кроме того, унитарное пространство  $\mathbb{R}^n$  и евклидово пространство  $\mathbb{R}^n$  являются прямыми суммами любых своих линейных подпространств и их ортогональных дополнений [7, 24, 30, 43].

Для любого вектора  $a$  и подпространства  $L$  в  $\mathbb{R}^n$  или  $\mathbb{C}^n$  всегда существует единственное разложение  $a = l + h$ , где  $l \in L$ ,  $h \in L^\perp$ . При этом вектор  $l$  называется *ортогональной проекцией* вектора  $a$  на подпространство  $L$ , а вектор  $h$  — *перпендикуляром*, опущенным из  $a$  на  $L$ . Мы будем обозначать ортогональную проекцию как  $\text{pr}_L a$ . Такие проекции обладают многими замечательными свойствами и наиболее популярны, но нередко используются и «наклонные» неортогональные проекции векторов.

Операция взятия проекции является линейным отображением некоторого специального вида, которое играет важнейшую роль в геометрии и при вычислениях с векторами.

### 3.26 Основные понятия теории матриц

*Матрицей* в математике называют прямоугольную таблицу, составленную из чисел или каких-либо других объектов. Если она имеет  $m$  строк и  $n$  столбцов, то обычно её записывают в виде

$$A := \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}, \quad (3.6)$$

и называют  $a_{ij}$  *элементами* матрицы  $A = (a_{ij})$ . Двойной индекс означает номер строки и номер столбца, в которых располагается рассматриваемый элемент. Говорят также, что матрица (3.6) имеет *размер*

$m \times n$ . При этом мы можем отождествлять  $n$ -векторы с матрицами размера  $n \times 1$  (вектор-столбцы) либо  $1 \times n$  (вектор-строки). Матрица называется *квадратной*, если количество её строк равно количеству столбцов, т.е.  $m = n$ . Иначе, если  $m \neq n$ , матрица называется *прямоугольной*.

Понятие матрицы оформилось в математике в середине XIX века, в основном после работ Дж. Сильвестра и А. Кэли, и сейчас матрицы широко используются для самых разнообразных целей. В частности, если дана система, составленная из конечного числа объектов (подсистем), то взаимодействие в ней  $i$ -го объекта с  $j$ -м можно описывать матрицей, элементы которой суть  $a_{ij}$ . В простейшем случае эти элементы принимают значения 1 или 0, соответствующие ситуациям «связь есть» и «никак не связано», и такую матрицу называют *матрицей смежности*.

Для нашего курса особенно важны применения матриц, связанные с различными конструкциями линейной алгебры.

Во-первых, матрица может представлять какой-либо набор векторов арифметических пространств  $\mathbb{R}^n$  или  $\mathbb{C}^n$ , когда упорядоченные кортежи чисел располагаются рядом друг с другом как единое целое. Иными словами, всякая числовая матрица есть упорядоченный набор своих вектор-строк или вектор-столбцов.

Во-вторых, с помощью матриц даётся удобное представление для набора коэффициентов линейных отображений конечномерных линейных векторных пространств с отмеченными в них базисами. Матрицы чрезвычайно полезны также для компактной записи наборов коэффициентов при неизвестных и правых частей уравнений в системах линейных алгебраических уравнений.

В-третьих, с помощью матриц удобно представлять наборы коэффициентов билинейных, полуторалинейных и квадратичных форм, и некоторые важные свойства этих форм равносильны свойствам их матриц коэффициентов.

Две матрицы одинаковых размеров считаются *равными*, если все их соответствующие элементы равны между собой. Матрица, все элементы которой равны нулю, называется *нулевой*.

*Главной диагональю* (или просто *диагональю*) матрицы  $A = (a_{ij})$  называется множество её элементов с совпадающими индексами —  $a_{11}, a_{22}, \dots, a_{kk}, \dots$  Наглядно геометрически диагонали квадратной матрицы действительно соответствует диагональ, идущая из левого верхнего угла в правый нижний угол. Побочной диагональю квадратной матри-

цы называется диагональ, идущая из правого верхнего угла в левый нижний угол. Матрица, в которой ненулевыми являются только элементы главной диагонали, называется *диагональной*.

*Подматрицей* матрицы  $A$  называют матрицу, которая образована элементами, находящимися на пересечении фиксированных множеств строк и столбцов  $A$  с сохранением их исходного порядка. *Ведущей подматрицей* (или угловой) некоторой матрицы называется квадратная матрица, составленная из строк и столбцов с первыми номерами. Подматрица, расположенная в столбцах и строках с одинаковыми номерами, называется *главной*.

*Транспонированной* к  $m \times n$ -матрице  $A = (a_{ij})$  называется  $n \times m$ -матрица  $A^\top$ , в которой  $ij$ -м элементом является  $a_{ji}$ . Таким образом, если  $A$  задаётся в виде (3.6), то

$$A^\top := \begin{pmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ a_{12} & a_{22} & \dots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{pmatrix}.$$

Числовые матрицы можно складывать, вычитать и умножать друг на друга. Эти операции соответствуют сумме, разности и композиции линейных отображений, задаваемых этими матрицами. Напомним, что сумма (разность) двух матриц одинакового размера есть матрица того же размера, образованная поэлементными суммами (разностями) операндов. Как следствие, сложение матриц коммутативно (перестановочно) и ассоциативно:

$$A + B = B + A, \quad (A + B) + C = A + (B + C).$$

Кроме того,

$$(A + B)^\top = A^\top + B^\top.$$

В качестве нейтрального элемента при сложении и вычитании выступает *нулевая матрица*, все элементы которой — нули.

Если  $A = (a_{ij})$  —  $m \times l$ -матрица и  $B = (b_{ij})$  —  $l \times n$ -матрица, то произведение матриц  $A$  и  $B$  есть такая  $m \times n$ -матрица  $C = (c_{ij})$ , что

$$c_{ij} := \sum_{k=1}^l a_{ik} b_{kj}.$$

Частными случаями этого определения являются определения умножения матрицы на вектор-столбец и умножения вектор-строки на матрицу. Кроме того, из определения матричного умножения следует

$$(AB)^\top = B^\top A^\top.$$

Умножение матриц в общем случае неперестановочно (некоммутативно), т. е.  $AB \neq BA$ . Но имеет место ассоциативность матричного умножения: для любых матриц  $A, B, C$  согласованных размеров

$$(AB)C = A(BC).$$

Следствием этого свойства является то обстоятельство, что в длинных произведениях матриц мы можем не заботиться о расстановке скобок, назначающих приоритет тех или иных умножений: при любом их порядке получается один и тот же результат. Кроме того, имеет место дистрибутивность матричного умножения по сложению:

$$(A + B)C = AC + BC, \quad A(B + C) = AB + AC.$$

Квадратная диагональная матрица  $I$  вида

$$\begin{pmatrix} 1 & & 0 \\ & 1 & \\ 0 & & \ddots & \\ & & & 1 \end{pmatrix},$$

у которой по диагонали стоят единицы, называется *единичной матрицей*. В матричном умножении она выполняет роль нейтрального элемента:

$$AI = A, \quad IA = A$$

для любой матрицы  $A$ , с которой имеют смысл выписанные произведения матриц.<sup>4</sup>

Матрицы можно рассматривать как объекты, составленные из своих вектор-строк или вектор-столбцов. *Строчным рангом* числовой матрицы (или рангом по строкам) называется количество её линейно независимых строк. *Столбцовым рангом* матрицы (или рангом по столбцам) называется максимальное количество её линейно независимых

---

<sup>4</sup>Буква  $I$  — от слова «identity», т. е. «тождественность».

столбцов. В курсах линейной алгебры показывается, что строчный и столбцовый ранги матрицы совпадают друг с другом и равны максимальному размеру ненулевого минора этой матрицы (см. определение ниже). По этой причине можем говорить просто о ранге матрицы. Мы будем обозначать его  $\text{rank } A$ .

Различают матрицы полного и неполного ранга. Более точно,  $m \times n$ -матрица, ранг которой равен  $\min\{m, n\}$ , т. е. максимально возможному для этой матрицы числу, называется *матрицей полного ранга*. Иначе матрица имеет *неполный ранг*.

Столбцовый и строчный ранги в действительности являются практически очень важными характеристиками матрицы, которые показывают, сколько строк или столбцов матрицы могут порождать линейную оболочку из *всех* её вектор-строк или вектор-столбцов. Нахождение минимального числа порождающих элементов важно, например, в анализе данных, когда необходимо оценить количество параметров, которыми может определяться вся совокупность имеющихся данных.

Квадратная матрица, все строки которой (или столбцы) линейно независимы, называется *неособенной* (регулярной, неособой). Её ранг равен, таким образом, её порядку. В противном случае квадратная матрица называется *особенной* (особой). Часто используемые для обозначения этого свойства термины «вырожденная матрица» и «невырожденная матрица» неудачны, поскольку матрица не перестаёт быть матрицей даже при линейной зависимости её строк (столбцов) и какого-то «вырождения» матрицы как таковой не происходит.

В качестве числовой меры особенности или неособенности квадратной матрицы часто применяется *определитель матрицы*  $\det A$  — полилинейная кососимметрическая функция строк (или столбцов) матрицы  $A$ , равная нулю тогда и только тогда, когда матрица  $A$  особенна. Определитель задаётся как алгебраическая сумма всевозможных произведений элементов матрицы, по одному из каждой строки и каждого столбца. Если столбцовые индексы элементов произведения образуют чётную перестановку при условии, что сами элементы расположены в порядке возрастания номеров строк, то это произведение берётся в сумме с положительным знаком, а если нечётную перестановку — с отрицательным. В формальных математических терминах

$$\det A = \sum_{\tau \in S_n} \operatorname{sgn} \tau \cdot a_{1,\tau(1)} a_{2,\tau(2)} \cdots a_{n,\tau(n)}, \quad (3.7)$$

где  $S_n$  — множество перестановок  $n$  элементов, а  $\operatorname{sgn} \tau$  означает знак

перестановки  $\tau$ . В частности, определитель квадратной диагональной матрицы равен произведению её диагональных элементов.

Определитель матрицы меняет знак, если поменять местами в матрице две её строки или два столбца. Определитель матрицы не меняется при добавлении к её строке (или столбцу) другой строки (столбца), умноженной на некоторое число. Определитель матрицы не меняется при её транспонировании. Наконец, определитель произведения матриц равен произведению их определителей, т. е.

$$\det AB = \det A \cdot \det B$$

для любых квадратных матриц  $A$  и  $B$  одинаковых размеров.

Определитель квадратной  $k \times k$ -подматрицы из матрицы  $A$  носит название *минора  $k$ -го порядка* матрицы  $A$ . Соответственно, *ведущий минор* матрицы — это определитель ведущей подматрицы.

Помимо явного выражения (3.7) для определителя матрицы активно применяются его представления через определители подматриц меньшего порядка. Наиболее важными являются формулы разложения определителя по строке и по столбцу (их называют *разложения Лапласа*):

$$\det A = \sum_{j=1}^n (-1)^{i+j} a_{ij} \det A_{ij} \quad \text{для фиксированного } i, \quad (3.8)$$

$$\det A = \sum_{i=1}^n (-1)^{i+j} a_{ij} \det A_{ij} \quad \text{для фиксированного } j, \quad (3.9)$$

где  $A_{ij}$  — подматрица размера  $(n-1) \times (n-1)$  в  $A = (a_{ij})$ , полученная из  $A$  вычёркиванием  $i$ -го столбца и  $j$ -й строки [9, 54, 47].

Помимо индикации особенности/неособенности матриц их определители полезны также во многих других конструкциях матричного анализа. В частности, определитель является ориентированным объёмом параллелепипеда, задаваемого вектор-строками матрицы [6, 47].

Если квадратная матрица  $A$  неособенна, то для неё существует *обратная матрица*, обозначаемая  $A^{-1}$  и имеющая те же размеры, такая что

$$AA^{-1} = I, \quad A^{-1}A = I.$$

В связи с этим неособенные матрицы часто называют *обратимыми*. Обратная матрица в самом деле играет роль обратного элемента в опе-

рации матричного умножения. Элементы обратных матриц можно выразить через определители самой матрицы и её подматриц с помощью формул, которые читатель может увидеть, к примеру, в [9, 54]. Из определения матричного умножения следует, что

$$(AB)^{-1} = B^{-1}A^{-1}, \quad (A^\top)^{-1} = (A^{-1})^\top.$$

Квадратные матрицы  $A$  и  $B$  одинакового порядка называются *подобными*, если существует такая невырожденная матрица  $S$  того же порядка, что

$$B = S^{-1}AS.$$

Такие матрицы получаются при задании одного и того же линейного преобразования в разных координатных системах. В этом случае  $S$  — матрица, определяющая преобразование координат. Преобразование подобия обладает ценным свойством сохранения спектра матрицы (см. § 3.2в).

Полезной характеристикой квадратных матриц является понятие *следа*. Для  $n \times n$ -матрицы  $A = (a_{ij})$  её следом называется величина

$$\operatorname{tr} A = a_{11} + a_{22} + \dots + a_{nn}$$

— сумма всех диагональных элементов матрицы. След является линейной функцией матрицы. Кроме того, след произведения матриц не зависит от порядка сомножителей, и потому у подобных матриц следы равны.

В случае, когда нулевые и ненулевые элементы в матрице  $A$  структурированы определённым образом, по отношению к  $A$  будут употребляться дополнительные определяющие термины. Например,

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ & a_{22} & \dots & a_{2n} \\ 0 & \ddots & \vdots & \\ & & & a_{nn} \end{pmatrix} \text{ и } \begin{pmatrix} a_{11} & & & 0 \\ a_{21} & a_{22} & & \\ \vdots & \vdots & \ddots & \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$$

— это *верхняя треугольная* и *нижняя треугольная* матрицы соответственно (рис. 3.1). Равносильные термины — *правая треугольная* и *левая треугольная* матрицы. Выбор того или иного варианта названия обычно диктуется контекстом или сложившейся традицией.

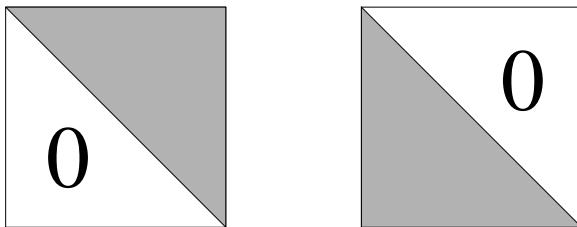


Рис. 3.1. Наглядные образы верхней треугольной и нижней треугольной матриц

Иногда возникает необходимость работать с треугольными матрицами, имеющими также нулевую диагональ. В них треугольник ненулевых элементов лежит строго выше или строго ниже главной диагонали, так что будем называть их *строго верхней* или *строго нижней* треугольными матрицами.

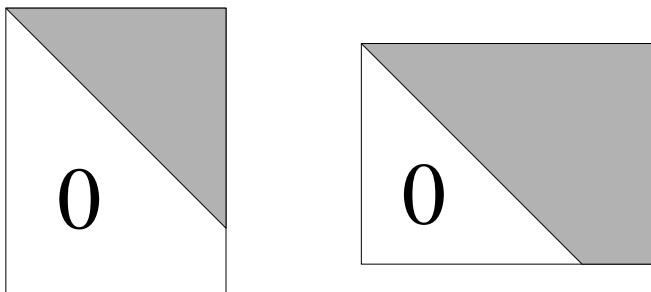


Рис. 3.2. Наглядные образы верхних (правых) трапециевидных матриц

Обобщением понятия треугольной матрицы на произвольный прямоугольный (неквадратный) случай являются *трапециевидные матрицы*. Именно, прямоугольная матрица с нулями ниже (выше) диагонали называется верхней (нижней) трапециевидной матрицей (рис. 3.2). Можно называть их также правой и левой трапециевидными матрицами. Вид области ненулевых элементов для прямоугольных матриц существенно отличается в зависимости от того, является  $m \times n$ -матрица «лежащей»

$(m < n)$  или «стоячей»  $(m > n)$ .

Блочными называются матрицы вида

$$\begin{pmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \dots & A_{mn} \end{pmatrix},$$

у которых элементы  $A_{ij}$ , в свою очередь, тоже являются матрицами, такими что их строчные и столбцовые размеры вдоль одной строки и одного столбца одинаковы. Подматрицы  $A_{ij}$  называются тогда *блоками* рассматриваемой матрицы. Блочные матрицы вида

$$\begin{pmatrix} A_{11} & & 0 \\ & A_{22} & \\ 0 & \ddots & A_{nn} \end{pmatrix} \quad \text{и} \quad \begin{pmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2n} \\ \vdots & \ddots & \vdots \\ 0 & & & A_{nn} \end{pmatrix},$$

где внедиагональные блоки или блоки ниже главной диагонали являются нулевыми, назовём соответственно *блочно-диагональными* или *верхними блочно-треугольными* (или правыми блочно треугольными) (рис. 3.3). Аналогичным образом определяются нижние блочно-треугольные (левые блочно-треугольные) матрицы.

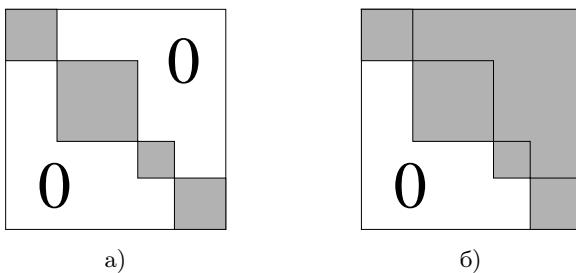


Рис. 3.3. Наглядные образы блочно-диагональной (а) и верхней блочно-треугольной (б) матриц

Введение структурированных матриц и отдельное их изучение мотивируется тем, что многие операции с такими матрицами можно выполнить более специальным образом или даже существенно проще,

чем в самом общем случае. В частности, сумма и произведение нижних (верхних) треугольных матриц есть нижняя (верхняя) треугольная матрица. Обратная к нижней (верхней) треугольной матрице также является нижней (верхней) треугольной матрицей. Для блочных матриц одинаковой структуры, имеющих одни и те же размеры и расположение своих блоков, сложение и умножение выполняются «по блокам», т. е. совершенно аналогично операциям над обычными матрицами, но «поблочным» образом, когда эти блоки выступают как отдельные самостоятельные элементы. Соответственно, обращение блочных матриц можно выполнить также «по блокам». Наконец, определитель блочной матрицы с квадратными блоками можно тоже вычислять «по блокам», оперируя их определителями, как если бы эти блоки были отдельными целостными элементами.

Линейная алгебра и её численные методы в некоторых ситуациях по существу требуют выхода в поле комплексных чисел  $\mathbb{C}$ , алгебраически пополняющее вещественную ось  $\mathbb{R}$ . Это необходимо, к примеру, в связи с понятиями собственных чисел и собственных векторов матриц, но может также диктоваться исходной содержательной постановкой задачи. В частности, привлечение комплексных чисел бывает необходимым при исследовании колебательных режимов в различных системах. В силу известной из математического анализа формулы Эйлера гармонические колебания с угловой частотой  $\omega$ , т. е. функции  $\cos \omega t$  и  $\sin \omega t$ , где  $t$  — время, обычно представляются в виде комплексной экспоненты  $\exp(i\omega t)$ .

Эрмитово-сопряжённой к  $m \times n$ -матрице  $A = (a_{ij})$  называют  $n \times m$ -матрицу  $A^*$ , в которой  $ij$ -м элементом является комплексно-сопряжённый  $\bar{a}_{ji}$ . Иными словами,

$$A^* := \begin{pmatrix} \bar{a}_{11} & \bar{a}_{21} & \dots & \bar{a}_{n1} \\ \bar{a}_{12} & \bar{a}_{22} & \dots & \bar{a}_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{a}_{1m} & \bar{a}_{2m} & \dots & \bar{a}_{nm} \end{pmatrix},$$

и эрмитово сопряжение матрицы есть композиция транспонирования и комплексного сопряжения её элементов.

В теории матриц и её приложениях широко используются специальные типы матриц, обладающие какими-либо специальными свойствами. Это эрмитовы, симметричные, косоэрмитовы, кососимметричные, унитарные, ортогональные и т. п. Напомним, что *симметричными*

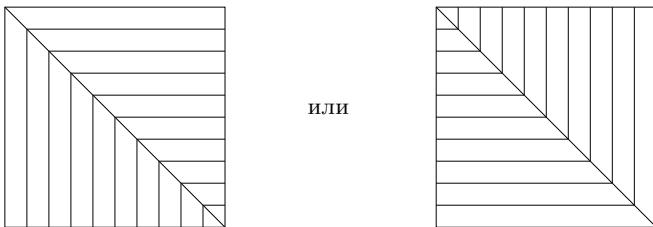


Рис. 3.4. Наглядные образы симметричной матрицы

*матрицами*<sup>5</sup> называют матрицы, совпадающие со своими транспонированными, т. е. удовлетворяющие  $A^\top = A$  (рис. 3.4). *Эрмитовыми матрицами* называются такие комплексные матрицы  $A$ , что  $A^* = A$ .

Матрица  $Q$  называется *унитарной*, если  $Q^*Q = I$  или, что равносильно,  $Q^{-1} = Q^*$ . Вещественные унитарные матрицы называют *ортогональными* матрицами. Таким образом, матрица  $Q$  — ортогональная, если  $Q^\top Q = I$  или  $Q^{-1} = Q^\top$ . Эквивалентное определение унитарных и ортогональных матриц состоит в том, что это матрицы, вектор-строки и вектор-столбцы которых взаимно ортогональны друг другу относительно стандартного скалярного произведения и имеют единичную евклидову норму.

Важнейшее математическое применение матриц — представление с их помощью наборов коэффициентов билинейных, полуторалинейных и квадратичных форм. Напомним, что *билинейной формой* называется выражение вида

$$\sum_{i,j} a_{ij} x_i y_j,$$

которое линейно по каждому из векторных аргументов  $(x_i)$  и  $(y_i)$  из  $\mathbb{R}^n$  или  $\mathbb{C}^n$ . В комплексном случае нередко удобнее работать с выражениями вида

$$\sum_{i,j} a_{ij} x_i \bar{y}_j,$$

аналогичным билинейным формам, но с комплексным сопряжением по второму аргументу. Они называются *полуторалинейными формами*

---

<sup>5</sup>Используют также термин *симметрическая матрица*.

или *эрмитово билинейными формами*. Таким образом, скалярное произведение (3.4) является простейшей вещественной билинейной формой, а скалярное произведение (3.5) — простейшей комплексной полуторалинейной формой.

*Квадратичной формой* называют выражение вида

$$\sum_{i,j} a_{ij} x_i x_j,$$

которое является суммой членов второй степени относительно компонент вектора  $x = (x_i)$ . Опять таки, в комплексном случае более удобными иногда оказываются *эрмитовы квадратичные формы*, которые определяются выражениями вида

$$\sum_{i,j} a_{ij} x_i \bar{x}_j$$

с сопряжением второго сомножителя в слагаемых.

Организуем коэффициенты  $a_{ij}$  билинейной, полуторалинейной или квадратичной формы в квадратную матрицу  $A = (a_{ij})$ . Тогда вещественные билинейная и квадратичная формы могут быть кратко записаны в виде

$$y^\top A x = \langle Ax, y \rangle \quad \text{и} \quad x^\top A x = \langle Ax, x \rangle$$

соответственно, где  $\langle \cdot, \cdot \rangle$  — стандартное скалярное произведение (3.4). В комплексном случае полуторалинейная и эрмитова квадратичная формы могут быть записаны в виде

$$y^* A x = \langle Ax, y \rangle \quad \text{и} \quad x^* A x = \langle Ax, x \rangle$$

соответственно, где  $\langle \cdot, \cdot \rangle$  — стандартное скалярное произведение (3.5). В преобразованиях таких выражений полезно очевидное свойство:

$$\langle Ax, y \rangle = \langle x, A^\top y \rangle \quad \text{в вещественном случае,}$$

$$\langle Ax, y \rangle = \langle x, A^* y \rangle \quad \text{в комплексном случае.}$$

Вещественные квадратные матрицы  $A$  и  $B$  одинакового размера называются *конгруэнтными*, если существует такая неособая матрица  $S$  того же размера, что

$$B = S^\top A S.$$

В комплексном случае это условие заменяется на  $B = S^*AS$ . Говорят также, что  $B$  получена из  $A$  с помощью преобразования конгруэнции с матрицей  $S$ . Им задаётся, в частности преобразование матрицы коэффициентов квадратичной формы при линейной замене переменных, определяемой матрицей  $S$ . Преобразование конгруэнции обладает ценным свойством сохранения симметричности матрицы.

Матрица  $A$  называется *положительно определённой*, если порождаемая ею квадратичная форма  $\langle Ax, x \rangle$  положительно определена, т. е.  $\langle Ax, x \rangle > 0$  для любых ненулевых  $x$ . Если же справедливо только нестрогое неравенство  $\langle Ax, x \rangle \geq 0$ , то матрица  $A$  называется *положительно полуопределённой*.

Отметим, что в данных выше определениях матрица  $A$  — квадратная, но не обязательно симметричная (эрмитова). Симметричная матрица положительно определена тогда и только тогда, когда все её ведущие миноры положительны (*критерий Сильвестра*). Симметричная матрица положительно определена тогда и только тогда, когда все её собственные значения положительны.

Важнейшей симметричной и положительно полуопределённой матрицей является *матрица Грама*, которая строится для заданного набора векторов  $\{v_1, v_2, \dots, v_m\}$  и образована их взаимными скалярными произведениями:

$$\Gamma(v_1, v_2, \dots, v_m) = \begin{pmatrix} \langle v_1, v_1 \rangle & \langle v_1, v_2 \rangle & \dots & \langle v_1, v_m \rangle \\ \langle v_2, v_1 \rangle & \langle v_2, v_2 \rangle & \dots & \langle v_2, v_m \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle v_m, v_1 \rangle & \langle v_m, v_2 \rangle & \dots & \langle v_m, v_m \rangle \end{pmatrix}$$

Матрица Грама положительно определена (т. е. неособенна), если векторы  $v_1, v_2, \dots, v_m$  линейно независимы.

*Разреженными* называются матрицы, большинство элементов которых равны нулю. Такие матрицы довольно часто встречаются в математическом моделировании, поскольку описывают системы или модели, в которых каждый элемент связан с относительно немногими другими элементами системы. Это происходит, например, если связи между элементами системы носят локальный характер. В противоположность этому *плотно заполненными* (или просто *плотными*) называют матрицы, которые не являются разреженными. Иными словами, в плотно заполненных матрицах большинство элементов не равны нулю.

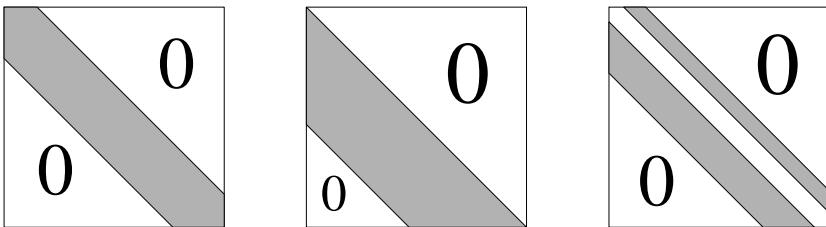


Рис. 3.5. Наглядные образы некоторых ленточных матриц

В разреженных матрицах нулевые и ненулевые элементы часто образуют какие-то регулярные структуры, и в этих случаях для названия соответствующих матриц употребляют более специальные термины. В частности, ленточными матрицами называют матрицы, у которых ненулевые элементы образуют выраженную «ленту» вокруг главной диагонали. В формальных терминах, матрица  $A = (a_{ij})$  называется *ленточной*, если существуют такие натуральные числа  $p$  и  $q$ , что  $a_{ij} = 0$  при  $j - i > p$  и  $i - j > q$  (рис. 3.5). В этом случае величина  $p + q + 1$  называется *шириной ленты*. Простейшими и важнейшими из ленточных матриц являются *трёхдиагональные матрицы*, для которых  $p = q = 1$ , и *двухдиагональные матрицы*, для которых  $p = 0$  и  $q = 1$  или  $p = 1$  и  $q = 0$ . Такие матрицы встречаются нам в § 3.9, 3.17ж и 3.19.

### 3.2в Собственные числа и собственные векторы матрицы

Как должно быть известно читателю, в теории и приложениях матриц огромную роль играют их *собственные значения* (собственные числа) и *собственные векторы*. Если обозначить посредством  $\lambda$  собственное значение квадратной  $n \times n$ -матрицы  $A$ , а  $x$ ,  $x \neq 0$ , — её собственный вектор, то они удовлетворяют матрично-векторному уравнению

$$Ax = \lambda x. \quad (3.10)$$

Содержательный смысл этого равенства состоит в том, что на одномерном линейном подпространстве в  $\mathbb{R}^n$  или  $\mathbb{C}^n$ , которое порождено

собственным вектором  $x$ , задаваемое матрицей  $A$  линейное преобразование действует как умножение на скаляр  $\lambda$ , т. е. как растяжение или сжатие. Иными словами, умножение вектора на всю матрицу в таких подпространствах эквивалентно умножению на один скаляр.

Из (3.10) следует  $(A - \lambda I)x = 0$ , что при  $x \neq 0$  означает особенность матрицы  $A - \lambda I$ . Соответственно, собственные значения матрицы  $A$  являются решениями так называемого *характеристического уравнения* матрицы, которое имеет вид

$$\det(A - \lambda I) = 0. \quad (3.11)$$

Для  $n \times n$ -матрицы  $A$  это алгебраическое уравнение  $n$ -й степени, в левой части которого стоит полином от переменной  $\lambda$ , называемый *характеристическим полиномом матрицы*. Очевидно, для всестороннего исследования этого алгебраического полинома и разрешимости самого характеристического уравнения по существу требуется привлечение алгебраически полного поля комплексных чисел  $\mathbb{C}$ , даже если сама матрица  $A$  — вещественная.

Всякая  $n \times n$ -матрица  $A$  зануляет свой характеристический полином, и этот результат является содержанием *теоремы Гамильтона–Кэли* [9, 28, 37, 44, 54]. Более точно, если  $A$  — произвольная квадратная матрица, вещественная или комплексная, и  $p(\lambda) = \det(A - \lambda I)$  — её характеристический полином, то верно матричное равенство  $p(A) = 0$ , где в правой части стоит нулевая матрица.

Так как всякое алгебраическое уравнение имеет в поле комплексных чисел  $\mathbb{C}$  хотя бы одно решение (основная теорема алгебры), то любая матрица имеет хотя бы одно собственное значение, в общем случае комплексное. Как следствие, любая матрица имеет хотя бы один собственный вектор из  $\mathbb{C}^n$ . Совокупность собственных чисел матрицы называется её *спектром*, так что в общем случае спектр матрицы — множество точек комплексной плоскости. Нетрудно показать, что произведение всех собственных значений матрицы равно её определителю. Напомним также широко известный факт: для эрмитовых и симметричных матриц собственные значения вещественны [9, 28, 54, 73].

Цель этого раздела — сообщить некоторые необщезвестные свойства собственных значений и собственных векторов матриц, необходимые в дальнейшем изложении.

Собственные векторы  $x$ , являющиеся решениями уравнения (3.10), называют также *правыми собственными векторами*, поскольку они умножаются на матрицу справа. Но нередко возникает необходимость

рассмотрения левых собственных векторов, обладающих свойством, аналогичным (3.10), но при умножении на матрицу слева. Очевидно, это должны быть собственные вектор-строки, но, имея в качестве основного пространство вектор-столбцов  $\mathbb{C}^n$ , нам будет удобно записать условие на левые собственные векторы в виде

$$y^* A = \mu y^*$$

для  $y \in \mathbb{C}^n$  и некоторого  $\mu \in \mathbb{C}$ . Применяя к этому равенству эрмитово сопряжение, получим

$$A^* y = \bar{\mu} y,$$

т.е. левые собственные векторы матрицы  $A$  являются правыми собственными векторами эрмитово сопряжённой матрицы  $A^*$ . Эта простая взаимосвязь объясняет редкость самостоятельного использования понятий левого и правого собственных векторов. Очевидно, при этом  $\det(A^* - \bar{\mu}I) = 0$ .

Исследуем подробнее так называемую *сопряжённую задачу* на собственные значения. Этим термином называют задачу нахождения собственных чисел и собственных векторов для эрмитово сопряжённой матрицы  $A^*$ :

$$A^* y = \varkappa y,$$

где  $\varkappa \in \mathbb{C}$  — собственное значение матрицы  $A^*$  и  $y \in \mathbb{C}^n$  — соответствующий собственный вектор. Как связаны между собой собственные значения и собственные векторы исходной  $A$  и сопряжённой  $A^*$  матриц? Для ответа на этот вопрос нам понадобится

**Определение 3.2.1** *Два набора из одинакового количества векторов  $\{r_1, r_2, \dots, r_m\}$  и  $\{s_1, s_2, \dots, s_m\}$  в евклидовом или унитарном пространстве называются биортогональными, если  $\langle r_i, s_j \rangle = 0$  при  $i \neq j$ .*

Приставка «би» в термине «биортогональность» означает, что введённое свойство относится к *двум* наборам векторов.

Выполнение свойства биортогональности существенно зависит от порядка нумерации векторов в пределах каждого из наборов, так что в определении биортогональности неявно предполагается, что необходимые нумерации существуют и рассматриваемые наборы упорядочены в соответствии с ними. Нетрудно также понять, что если какой-либо набор векторов биортогонален сам себе, то он ортогонален в обычном смысле.

**Предложение 3.2.1** *Собственные значения эрмитово-сопряжённых матриц попарно комплексно сопряжены друг другу. Собственные векторы эрмитово сопряжённых матриц биортогональны.*

**Доказательство.** Определитель матрицы, как известно, не меняется при её транспонировании, т. е.  $\det A^\top = \det A$ . С другой стороны, комплексное сопряжение элементов матрицы влечёт комплексное сопряжение её определителя,  $\det \bar{A} = \overline{\det A}$ . Следовательно,

$$\begin{aligned}\det(A - \lambda I) &= \det(A - \lambda I)^\top = \det(A^\top - \lambda I) = \\ &= \overline{\det(\bar{A}^\top - \bar{\lambda} I)} = \overline{\det(\bar{A}^* - \bar{\lambda} I)}.\end{aligned}$$

Отсюда мы можем заключить, что комплексное число  $z$  является решением характеристического уравнения  $\det(A - \lambda I) = 0$  для матрицы  $A$  тогда и только тогда, когда ему сопряжённое  $\bar{z}$  является решением уравнения  $\det(\bar{A}^* - \bar{\lambda} I) = 0$ , характеристического для матрицы  $A^*$ . Это доказывает первое утверждение.

Пусть  $x$  и  $y$  — собственные векторы матриц  $A$  и  $A^*$  соответственно, а  $\lambda$  и  $\varkappa$  — отвечающие этим векторам собственные числа матриц  $A$  и  $A^*$ . Для доказательства второго утверждения выпишем следующую цепочку преобразований:

$$\lambda \langle x, y \rangle = \langle \lambda x, y \rangle = \langle Ax, y \rangle = \langle x, A^* y \rangle = \langle x, \varkappa y \rangle = \bar{\varkappa} \langle x, y \rangle.$$

Поэтому

$$\lambda \langle x, y \rangle - \bar{\varkappa} \langle x, y \rangle = 0,$$

т. е.

$$(\lambda - \bar{\varkappa}) \langle x, y \rangle = 0.$$

Если  $x$  и  $y$  являются собственными векторами матриц  $A$  и  $A^*$ , отвечающими собственным значениям  $\lambda$  и  $\varkappa$ , которые не сопряжены комплексно друг другу, то в левой части полученного равенства первый сомножитель  $(\lambda - \bar{\varkappa}) \neq 0$ . По этой причине необходимо  $\langle x, y \rangle = 0$ , что и требовалось доказать. ■

**Следствие.** Собственные значения симметричных и эрмитовых матриц вещественны. Собственные векторы симметричных и эрмитовых матриц, отвечающие различным собственным значениям, ортогональны друг другу.

Результат предложения 3.2.1 может показаться парадоксальным, так как в применении к вещественным эрмитово-сопряжённым матрицам, которые являются просто транспонированными, он означает, что их собственные значения попарно комплексно сопряжены друг другу. Но собственные значения транспонированных матриц просто совпадают друг с другом. Противоречие разрешается тем, что у вещественных матриц характеристический полином тоже вещественный, так что его нули — это комплексно-сопряжённые пары. Следовательно, у вещественных матриц собственные значения либо вещественны, либо имеют комплексно-сопряжённых «двойников», которые переходят друг в друга при операции сопряжения.

Обращаясь к определению правых и левых собственных векторов матрицы, можем утверждать, что если  $\lambda$  — правое собственное значение матрицы  $A$ , а  $\mu$  — левое собственное значение, то  $\bar{\lambda} = \bar{\mu}$ . Иными словами, правые и левые собственные значения матрицы совпадают друг с другом. Поэтому их можно не различать и говорить просто о собственных значениях матрицы. Что касается правых и левых собственных векторов матрицы, то они биортогональны друг другу.

**Предложение 3.2.2** *Если  $\lambda$  — собственное число квадратной неособенной матрицы, то  $\lambda^{-1}$  — это собственное число обратной матрицы, отвечающее тому же собственному вектору.*

**Доказательство.** Если  $C$  — неособенная  $n \times n$ -матрица и  $Cv = \lambda v$ , то  $v = \lambda C^{-1}v$ . Далее, так как  $\lambda \neq 0$  в силу неособенности  $C$ , получаем отсюда  $C^{-1}v = \lambda^{-1}v$ . ■

**Предложение 3.2.3** *Пусть  $A$  —  $m \times n$ -матрица,  $B$  —  $n \times m$ -матрица, так что одновременно определены произведения  $AB$  и  $BA$ . Спектры матриц  $AB$  и  $BA$  могут различаться только нулевыми собственными значениями.*

**Доказательство.** Пусть  $\lambda$  — какое-нибудь ненулевое собственное значение матрицы  $AB$ , так что

$$ABu = \lambda u \tag{3.12}$$

с некоторым вектором  $u \neq 0$ . Умножая это равенство слева на матрицу  $B$ , получим

$$B(ABu) = B(\lambda u),$$

или

$$BA(Bu) = \lambda(Bu),$$

причём  $Bu \neq 0$ , так как иначе в исходном соотношении (3.12) должно быть  $\lambda = 0$ . Сказанное означает, что вектор  $Bu$  является собственным вектором матрицы  $BA$ , отвечающим такому же собственному значению  $\lambda$ .

И наоборот, если ненулевое  $\mu$  есть собственное значение для  $BA$ , то, домножая слева равенство

$$BAv = \mu v$$

на матрицу  $A$ , получим

$$ABA v = AB(Av) = \mu(Av),$$

причём  $Av \neq 0$ . По этой причине  $Av$  есть собственный вектор матрицы  $AB$ , отвечающий собственному значению  $\mu$ . Иными словами, ненулевые собственные числа матриц  $AB$  и  $BA$  находятся во взаимно однозначном соответствии друг с другом. ■

Другой вывод этого результата можно найти, к примеру, в [2, 45]. Особая роль нулевого собственного значения в этом результате объясняется тем, что если  $A$  и  $B$  — прямоугольные матрицы, то из двух матриц  $AB$  и  $BA$  по крайней мере одна имеет неполный ранг — та, чьи размеры больше. Она, соответственно, особенна и имеет нулевое собственное значение. Но меньшая по размерам матрица особенной при этом может и не быть.

### 3.2г Разложения матриц, использующие их спектр

Квадратную матрицу вида

$$\begin{pmatrix} \alpha & 1 & & 0 \\ & \alpha & 1 & \\ & & \ddots & \ddots & \\ 0 & & & \alpha & 1 \\ & & & & \alpha \end{pmatrix},$$

у которой по диагонали стоит  $\alpha$ , на первой наддиагонали — все единицы, а остальные элементы — нули, называют, как известно, *жордановой*

*клеткой*, отвечающей значению  $\alpha$ . Ясно, что  $\alpha$  является собственным значением такой матрицы.

В теории матриц показывается, что с помощью подходящего преобразования подобия любая квадратная матрица может быть приведена к *жордановой канонической форме* — блочно-диагональной матрице, на главной диагонали которой стоят жордановы клетки, отвечающие собственным значениям рассматриваемой матрицы (см., к примеру, [7, 9, 24, 28, 41, 43, 54]). Иначе говоря, для любой квадратной матрицы  $A$  существует такая неособенная матрица  $S$ , что

$$S^{-1}AS = J,$$

где

$$J = \left( \begin{array}{cc|cc|c} \lambda_1 & 1 & & & 0 \\ \lambda_1 & \ddots & 0 & & 0 \\ \ddots & 1 & \lambda_1 & & \\ \hline 0 & & \lambda_2 & 1 & 0 \\ & & & \ddots & \ddots \\ & & & & \lambda_2 \\ \hline 0 & & 0 & & \ddots \\ & & & & \ddots \end{array} \right), \quad (3.13)$$

а  $\lambda_1, \lambda_2, \dots$  — собственные значения матрицы  $A$ . Квадратные матрицы, имеющие выписанный выше вид (3.13), называют матрицами в *жордановой форме*.

Соответственно, представление произвольной матрицы  $A$  в виде

$$A = SJS^{-1},$$

где  $J$  — матрица в жордановой форме, называют *жордановым разложением*.

Неприятной особенностью жордановой канонической формы и жорданова разложения является то, что они не зависят непрерывно от элементов матрицы. Размеры жордановых клеток-блоков и их количество могут скачкообразно меняться при изменении элементов матрицы. В

то же время сами собственные значения матрицы непрерывно зависят от её элементов (теорема Островского, см. § 3.17в). Отмеченное обстоятельство делает жорданову форму малопригодной при решении многих практических задач, где входные данные носят приближённый и неточный характер.

В связи со сказанным большое значение имеют так называемые *матрицы простой структуры*, называемые также *диагонализуемыми* или *недефектными* матрицами (см. § 3.17в), которые определяются как матрицы, подобные диагональным. Можно показать, что таких матриц — «большинство», т. е. типичная матрица имеет простую структуру (см. предложение 3.17.1). Жорданово разложение таких матриц превращается в более простое представление

$$A = SDS^{-1},$$

где

$$D = \begin{pmatrix} \lambda_1 & & & & & \\ & \lambda_1 & & & & \\ & & \ddots & & & \\ & & & \lambda_1 & & \\ \hline & 0 & & & \lambda_2 & \\ & & & & & \ddots & \\ \hline & 0 & & & 0 & & \lambda_2 \\ & & & & & & \\ & 0 & & & 0 & & \ddots & \\ & & & & & & & \ddots \end{pmatrix},$$

— диагональная матрица, у которой по диагонали стоят собственные значения  $A$  с учётом их кратности. Часто это представление называют *спектральным разложением* матрицы (или соответствующего ей линейного оператора).

Другое популярное разложение матриц, использующее информацию о спектре матрицы — это разложение Шура.

Пусть  $A$  — комплексная  $n \times n$ -матрица и зафиксирован некоторый порядок её собственных значений  $\lambda_1, \lambda_2, \dots, \lambda_n$ . Существует такая унитарная  $n \times n$ -матрица  $U$ , что матрица  $T = U^*AU$  является верхней треугольной матрицей с диагональными элементами  $\lambda_1, \lambda_2, \dots, \lambda_n$ . Иными словами, любая комплексная квадратная матрица  $A$  унитарно

подобна треугольной матрице, в которой диагональные элементы являются собственными значениями для  $A$ , записанными в произвольном заранее заданном порядке. Если же  $A$  — это вещественная матрица и все её собственные значения вещественны, то  $U$  можно выбрать вещественной ортогональной матрицей. Представление

$$A = UTU^*$$

с верхней треугольной матрицей  $T$  и унитарными (ортогональными) матрицами  $U$  и  $U^*$  называют *разложением Шура* матрицы  $A$ . В отличие от жорданова разложения и жордановой нормальной формы, оно устойчиво к возмущениям элементов матрицы  $A$ . Конструктивным способом получения разложения Шура является QR-алгоритм, который рассматривается в § 3.18г–3.18д.

Для симметричных (эрмитовых в комплексном случае) матриц в выписанном представлении матрица  $T$  также должна быть симметричной (эрмитовой). Как следствие, в этом случае справедлив более сильный результат: с помощью ортогонального (унитарного) преобразования подобия любая симметричная (эрмитова) матрица может быть приведена к диагональному виду, с собственными значениями по диагонали. Тогда для соответствующего линейного оператора спектральное разложение даёт его представление в виде линейной комбинации операторов проектирования на взаимно ортогональные оси.

### 3.2д Сингулярные числа и сингулярные векторы матрицы

Из результатов § 3.2в следует, что для определения собственных значений квадратной матрицы  $A$  и её левых и правых собственных векторов необходимо решить относительно скаляра  $\lambda$  и векторов  $x$  и  $y$  систему уравнений

$$\begin{cases} Ax = \lambda x, \\ y^* A = \lambda y^*. \end{cases} \quad (3.14)$$

Система уравнений (3.14) является «распавшейся»: в ней первая половина уравнений (соответствующая  $Ax = \lambda x$ ) никак не зависит от второй половины уравнений (соответствующей  $y^* A = \lambda y^*$ ). Поэтому решать систему (3.14) можно также по частям, отдельно для  $x$  и отдельно для  $y$ , что обычно и делают на практике. Если  $\lambda$  вещественно, т. е.  $\lambda = \bar{\lambda}$ , то системе (3.14), применив операцию эрмитова сопряжения

ко второй части, можно придать следующий элегантный матричный вид

$$\begin{pmatrix} A & 0 \\ 0 & A^* \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix}. \quad (3.15)$$

Рассмотрим теперь аналогичную систему уравнений, порождаемую заданной матрицей  $A$ , которая получается изменением соотношений в (3.14) так, чтобы они «заязались» друг на друга:

$$\begin{cases} Ax = \sigma y, \\ y^* A = \sigma x^*. \end{cases} \quad (3.16)$$

В сравнении с системой (3.14) векторы  $x$  и  $y$  здесь просто поменялись местами. Фигурально можно сказать, что в новой системе уравнений (3.16) векторы  $x$  и  $y$  становятся «право-левыми» и «лево-правыми собственными векторами» матрицы  $A$ . Как увидим вскоре, аналоги собственных чисел матрицы, которые мы переобозначили через  $\sigma$ , также получают новое содержание. Решения системы (3.14) давали ценную информацию о матрице и задаваемом ею линейном преобразовании пространства, и то же самое, как будет показано ниже, справедливо в отношении решений новой системы уравнений (3.16). Они тоже дают важную информацию о матрице, хотя и другого сорта, нежели (3.14).

Система уравнений (3.16) — это система алгебраических уравнений относительно  $\sigma$ ,  $x$ ,  $y$ , и потому естественно ожидать, что её решениями в самом общем случае, когда  $A$  — комплексная матрица, будут тоже комплексные числа  $\sigma$  и комплексные векторы  $x$ ,  $y$ . Но оказывается, что  $\sigma$  всегда может быть взято вещественным.

В самом деле, если  $\sigma$ , удовлетворяющие системе (3.16), вещественны, то  $\bar{\sigma} = \sigma$ , и, взяв эрмитово сопряжение второго матричного уравнения из (3.16), можем переписать всю эту систему в следующем равносильном виде:

$$\begin{cases} Ax = \sigma y, \\ A^* y = \sigma x. \end{cases} \quad (3.17)$$

Матричная форма этой системы

$$\begin{pmatrix} 0 & A^* \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \sigma \begin{pmatrix} x \\ y \end{pmatrix} \quad (3.18)$$

находится в красивой двойственности с системой (3.15). Если  $A$  — вещественная матрица, то векторы  $x$  и  $y$  также могут быть взяты вещественными, а система уравнений (3.18) для определения сингулярных чисел и векторов принимает ещё более простой вид:

$$\begin{pmatrix} 0 & A^\top \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \sigma \begin{pmatrix} x \\ y \end{pmatrix}. \quad (3.19)$$

Немедленно можно заметить, что система (3.17) имеет смысл для произвольных прямоугольных матриц, а не только для квадратных, как было в случае собственных значений и собственных векторов. Так или иначе, матрицы

$$\begin{pmatrix} 0 & A^* \\ A & 0 \end{pmatrix} \quad \text{и} \quad \begin{pmatrix} 0 & A^\top \\ A & 0 \end{pmatrix}$$

размера  $(m+n) \times (m+n)$  являются эрмитовой и симметричной соответственно. Поэтому матричные уравнения (3.18) и (3.19), которые определяют их собственные значения и собственные  $(m+n)$ -векторы, имеют решения  $\sigma$ ,  $x$ ,  $y$ , в которых  $\sigma$  должно быть вещественным (см. следствие из предложения 3.2.1). Таким образом, сделанное выше допущение о вещественности  $\sigma$  в самом деле реализуется. В следующем разделе, в предложении 3.2.4, мы покажем, что других  $\sigma$  быть не может.

Отметим характерные особенности системы уравнений (3.17). Как и аналогичная система уравнений (3.14) для собственных чисел и собственных векторов, она всегда имеет тривиальное решение  $\sigma = 0$ ,  $x = 0$ ,  $y = 0$ , которое большого интереса не представляет и обычно не рассматривается. Но если какой-то из членов этой тройки — ненулевой, то такое решение имеет содержательный смысл и должно учитываться.

Далее, если тройка  $\sigma$ ,  $x$ ,  $y$  является решением системы (3.17), то решением также являются тройки  $(-\sigma)$ ,  $(-x)$ ,  $y$  и  $(-\sigma)$ ,  $x$ ,  $(-y)$ . В любом случае у системы уравнений (3.17) одновременно с решением  $\sigma$  присутствует также  $(-\sigma)$ .

Ещё одно наблюдение состоит в том, что векторы  $(x, y)^\top \in \mathbb{R}^{m+n}$ , которые являются решениями системы (3.18), отвечающими одному и тому же  $\sigma$ , образуют линейное подпространство в  $\mathbb{R}^{m+n}$  и оно имеет размерность по меньшей мере 1. Ясно, что из этого линейного подпространства имеет смысл брать линейно независимые решения системы уравнений (3.18).

Но если матрица  $A$  — прямоугольная с размерами  $m \times n$ , то система уравнений (3.18) имеет  $|m - n|$  заведомо нулевых решений  $\sigma$ , которые соответствуют ненулевым линейно независимым векторам  $(x, y)^\top$ . В самом деле, пусть для определённости  $m > n$ , так что матрица  $A$  — «стоячая». Тогда  $A^*$  — «лежачая» матрица, у которой ранг не превосходит  $n$ , и поэтому существуют ненулевые  $m$ -векторы  $y$ , удовлетворяющие равенству  $A^*y = 0$ . Соответствующее им  $\sigma$  в системе уравнений (3.17) должно быть нулевым. Количество таких линейно независимых векторов  $y$  равно  $m - n$ . Для случая матрицы  $A$ , у которой  $m < n$ , рассуждения аналогичны.

В силу отмеченных причин для  $m \times n$ -матрицы содержательный смысл имеет рассматривать  $m + n - |m - n| = 2 \min\{m, n\}$  значений  $\sigma$ , удовлетворяющих уравнению (3.17), или даже в половину меньше — всего  $\min\{m, n\}$ , принимая во внимание обязательное присутствие среди таких  $\sigma$  пар чисел с противоположными знаками.

**Определение 3.2.2** Неотрицательные вещественные скаляры  $\sigma$ , которые являются решениями системы матричных уравнений (3.17), называются сингулярными числами матрицы  $A$ . Удовлетворяющие системе (3.17) векторы  $x$  называются правыми сингулярными векторами матрицы  $A$ , а векторы  $y$  — левыми сингулярными векторами матрицы  $A$ .

Как отмечалось выше, и система (3.17) и данное выше определение имеют смысл для произвольных прямоугольных матриц. Для  $m \times n$ -матрицы  $A$  правые сингулярные векторы имеют размерность  $n$ , а левые — размерность  $m$ . Из уравнений (3.17)–(3.19) видно также, что, в отличие от собственных значений и собственных векторов, сингулярные числа и сингулярные векторы характеризуют совместно как саму матрицу, так и её эрмитово-сопряжённую (транспонированную в вещественном случае).

**Пример 3.2.1** Пусть  $A$  — это  $1 \times 1$ -матрица, т.е. просто некоторое число  $a$ , вещественное или комплексное. Ясно, что единственное собственное число такой матрицы равно самому  $a$ . Для нахождения сингулярных чисел образуем матричное уравнение вида (3.18)

$$\begin{pmatrix} 0 & \bar{a} \\ a & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \sigma \begin{pmatrix} x \\ y \end{pmatrix},$$

где  $x, y$  — числа. Таким образом,  $\sigma$  оказывается собственным значением матрицы

$$\begin{pmatrix} 0 & \bar{a} \\ a & 0 \end{pmatrix},$$

и для его нахождения можем воспользоваться решением характеристического уравнения (3.11), которое принимает вид  $\sigma^2 - \bar{a}a = 0$ . Следовательно, сингулярное число у матрицы  $A = (a)$  также всего одно, и оно равно  $\sigma = \sqrt{\bar{a}a} = |a|$ . ■

### 3.2e Свойства сингулярных чисел и векторов

**Предложение 3.2.4** *Сингулярные числа матрицы  $A$  суть неотрицательные квадратные корни из совпадающих собственных чисел матриц  $A^*A$  и  $AA^*$ .*

Нормированные в евклидовой норме собственные векторы матрицы  $A^*A$  являются правыми сингулярными векторами матрицы  $A$ , а нормированные в евклидовой норме собственные векторы матрицы  $AA^*$  являются левыми сингулярными векторами для  $A$ .

Отметим, что матрица  $A^*A$  — это матрица Грама (т. е. матрица взаимных скалярных произведений) для вектор-столбцов матрицы  $A$ , а  $AA^*$  — матрица Грама для вектор-строк матрицы  $A$  (эти факты использовались в § 2.11г).

Неочевидный момент формулировки предложения 3.2.4 — взаимоотношение собственных чисел матриц  $A^*A$  и  $AA^*$ . Здесь можно вспомнить доказанный выше общий результат линейной алгебры — предложение 3.2.3 о совпадении точек спектра произведений двух матриц, взятых в различном порядке, кроме, возможно, нулевых значений. Впрочем, для рассматриваемого частного случая совпадение собственных чисел матриц  $A^*A$  и  $AA^*$  будет выведено в следующем ниже доказательстве.

Формулировка предложения 3.2.4 требует пояснений ещё и потому, что для общей прямоугольной  $m \times n$ -матрицы  $A$  размеры квадратных матриц  $A^*A$  и  $AA^*$  различны: первая из них — это  $n \times n$ -матрица, а вторая —  $m \times m$ -матрица. Поэтому количество собственных чисел у них тоже различно.

Но известно, что ранг произведения матриц не превосходит наименьшего из рангов перемножаемых матриц [9, 24, 54]. Отсюда следует, что если  $m < n$ , то  $n \times n$ -матрица  $A^*A$  имеет неполный ранг, не

превосходящий  $m$ , а потому её собственные числа с  $(m+1)$ -го по  $n$ -е — заведомо нулевые. Аналогично, если  $m > n$ , то неполный ранг, который не превосходит  $n$ , имеет  $m \times m$ -матрица  $AA^*$ , и её собственные числа с  $(n+1)$ -го по  $m$ -е равны нулю. Не имеет большого смысла рассматривать эти заведомо нулевые собственные числа матриц  $A^*A$  и  $AA^*$ . По этой причине в связи с сингулярными числами  $m \times n$ -матрицы  $A$  учитывают лишь  $\min\{m, n\}$  штук общих собственных чисел матриц  $A^*A$  и  $AA^*$ , что устраняет отмеченную выше кажущуюся неоднозначность.

**Доказательство.** Умножая обе части второго уравнения из (3.17) на  $\sigma$ , получим  $A^*(\sigma y) = \sigma^2 x$ . Затем подставим сюда значение  $\sigma y$  из первого уравнения (3.17):

$$A^*Ax = \sigma^2 x.$$

С другой стороны, умножая на  $\sigma$  обе части первого уравнения (3.17), получим  $A(\sigma x) = \sigma^2 y$ . Подстановка в это равенство значения  $\sigma x$  из второго уравнения (3.16) даёт

$$AA^*y = \sigma^2 y.$$

Иными словами, числа  $\sigma^2$  являются собственными значениями как для  $A^*A$ , так и для  $AA^*$ .

Покажем теперь, что собственные значения у матриц  $A^*A$  и  $AA^*$  неотрицательны, чтобы иметь возможность извлекать из них квадратные корни для окончательного определения  $\sigma$ . Очевидно, это достаточно сделать лишь для одной из выписанных матриц, так как для другой рассуждения совершенно аналогичны.

Пусть  $\lambda$  — собственное значение матрицы  $A^*A$ , а  $u$  — соответствующий ему собственный вектор,  $u \neq 0$ . Произведение  $(Au)^*(Au)$  является суммой квадратов модулей компонент вектора  $Au$ , и потому неотрицательно. Кроме того,  $(Au)^*(Au) = u^*(A^*Au) = u^*\lambda u = \lambda(u^*u)$ , откуда в силу  $u^*u > 0$  следует  $\lambda \geq 0$ .

Для завершения доказательства продемонстрируем, что арифметические квадратные корни из собственных значений матриц  $A^*A$  и  $AA^*$  вместе с соответствующими нормированными собственными векторами удовлетворяют системе уравнений (3.17)–(3.18).

Пусть  $u$  — собственный вектор матрицы  $A^*A$ , отвечающий её собственному числу  $\lambda = \sigma^2$  и нормированный в евклидовой норме, так что

$A^*Au = \sigma^2 u$ . Если взять  $v := Au/\sigma$ , то

$$\begin{aligned} AA^*v &= AA^*Au/\sigma = A(A^*Au)/\sigma = A(\sigma^2 u)/\sigma = \sigma v, \\ \langle v, v \rangle &= \langle Au, Au \rangle / \sigma^2 = \langle A^*Au, u \rangle / \sigma^2 = \langle \sigma^2 u, u \rangle / \sigma^2 = 1, \end{aligned}$$

и потому вектор  $v$  оказывается нормированным собственным вектором матрицы  $AA^*$ . Кроме того, по самому построению  $u$  и  $v$  имеем

$$\begin{aligned} Au &= \sigma v, \\ A^*v &= A^*Au/\sigma = \sigma^2 u/\sigma = \sigma u, \end{aligned}$$

и система (3.16)–(3.18) для сингулярных чисел и сингулярных векторов удовлетворяется.

Пусть  $v$  — нормированный собственный вектор матрицы  $AA^*$ , отвечающий её собственному числу  $\mu = \sigma^2$ , так что  $AA^*v = \sigma^2 v$ . Возьмём  $u := A^*v/\sigma$ , тогда

$$\begin{aligned} A^*Au &= A^*AA^*v/\sigma = A^*(AA^*v)/\sigma = A^*(\sigma^2 v)/\sigma = \sigma u, \\ \langle u, u \rangle &= \langle A^*v, A^*v \rangle / \sigma^2 = \langle AA^*v, v \rangle / \sigma^2 = \langle \sigma^2 v, v \rangle / \sigma^2 = 1, \end{aligned}$$

и потому вектор  $u$  оказывается нормированным собственным вектором матрицы  $A^*A$ . Кроме того, по самому построению  $u$  и  $v$  имеем

$$\begin{aligned} Au &= AA^*v/\sigma = \sigma^2 v/\sigma = \sigma v, \\ A^*v &= \sigma u, \end{aligned}$$

и система (3.17)–(3.18) действительно удовлетворяется с выбранными векторами  $u$  и  $v$ . ■

Подведём промежуточные итоги. Задаваемые определением 3.2.2 сингулярные числа вещественной или комплексной  $m \times n$ -матрицы — это набор из  $\min\{m, n\}$  неотрицательных вещественных чисел, которые обычно нумеруют в порядке убывания:

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_{\min\{m, n\}} \geq 0.$$

Таким образом,  $\sigma_1 = \sigma_1(A)$  — это наибольшее сингулярное число матрицы  $A$ . Мы будем также обозначать наибольшее и наименьшее сингулярные числа матрицы посредством  $\sigma_{\max}(A)$  и  $\sigma_{\min}(A)$ .

Из предложения 3.2.4 следует, что сингулярные числа эрмитовых матриц (симметричных в вещественном случае) равны абсолютным значениям их собственных чисел, так как  $A^*A = AA^* = A^2$  для таких матриц.

Ещё одно следствие из предложения 3.2.4 состоит в том, что семейства левых и правых сингулярных векторов матрицы суть ортогональные системы векторов, коль скоро они являются собственными векторами эрмитовых матриц  $A^*A$  и  $AA^*$ . Кроме того, предложение 3.2.4 показывает, что правые и левые сингулярные векторы можно одновременно брать нормированными, что совершенно неочевидно из их определения.

**Пример 3.2.2** Пусть  $A = (a_1, a_2, \dots, a_n)^\top$  — это  $n \times 1$ -матрица, т. е. просто вектор-столбец. Тогда матрица  $A^*A$  является скаляром  $\bar{a}_1a_1 + \bar{a}_2a_2 + \dots + \bar{a}_na_n = |a_1|^2 + |a_2|^2 + \dots + |a_n|^2$ , и поэтому единственное сингулярное число матрицы  $A$  равно евклидовой норме вектора  $(a_1, a_2, \dots, a_n)^\top$ . То же самое верно для  $1 \times n$ -матрицы, то есть вектор-строки  $(a_1, a_2, \dots, a_n)$ . ■

Разобранный пример демонстрирует связь сингулярных чисел с евклидовой нормой векторов и матриц (для матриц она называется *фробениусовой нормой*). Далее мы ещё не раз увидим, как эта связь проявляется в самых неожиданных местах (см., в частности, предложение 3.3.8).

**Пример 3.2.3** Для единичной матрицы  $I$  все сингулярные числа очевидно равны единицам.

Но все единичные сингулярные числа имеет не только единичная матрица. Если  $U$  — унитарная комплексная матрица (ортогональная в вещественном случае), то  $U^*U = I$ , и потому все сингулярные числа для  $U$  также равны единицам. ■

**Пример 3.2.4** Для  $2 \times 2$ -матрицы

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \quad (3.20)$$

характеристическим уравнением является

$$\det \begin{pmatrix} 1 - \lambda & 2 \\ 3 & 4 - \lambda \end{pmatrix} = \lambda^2 - 5\lambda - 2 = 0.$$

Его решения  $\frac{1}{2}(5 \pm \sqrt{33})$  — собственные значения матрицы, приближённо равные  $-0.372$  и  $5.372$ . Для определения сингулярных чисел обратим

$$A^\top A = \begin{pmatrix} 10 & 14 \\ 14 & 20 \end{pmatrix}$$

и вычислим её собственные значения. Они равны  $15 \pm \sqrt{221}$ , и потому получается, что сингулярные числа матрицы  $A$  суть  $\sqrt{15 \pm \sqrt{221}}$ , т. е. примерно  $0.366$  и  $5.465$  (с точностью до трёх знаков после запятой).

С другой стороны, для матрицы

$$\begin{pmatrix} 1 & 2 \\ -3 & 4 \end{pmatrix}, \quad (3.21)$$

которая отличается от матрицы (3.20) лишь противоположным знаком элемента на месте  $(2, 1)$ , собственные значения — это комплексно-сопряжённая пара  $\frac{1}{2}(5 \pm i\sqrt{15}) \approx 2.5 \pm 1.936i$ , а сингулярные числа суть  $\sqrt{15 \pm \sqrt{125}}$ , т. е. приблизительно  $1.954$  и  $5.117$ . ■

Можно заметить, что модули собственных чисел рассмотренных матриц не превосходят их максимальных сингулярных чисел и в то же время не меньше, чем минимальные сингулярные числа. Это не случайно, и справедлива

**Теорема 3.2.1** (теорема Э. Брауна [109]) *Для любой квадратной матрицы  $A$  модули её собственных значений  $\lambda(A)$  заключены между минимальным и максимальным сингулярными числами  $A$ :*

$$\sigma_{\min}(A) \leq |\lambda(A)| \leq \sigma_{\max}(A). \quad (3.22)$$

**Доказательство.** Пусть  $v \in \mathbb{C}^n$  — собственный вектор, отвечающий собственному значению  $\lambda$  матрицы  $A$ . Можно сделать его нормированным в евклидовой норме, так что  $v^*v = 1$ . Пусть также эрмитова матрица  $A^*A$  с помощью унитарного преобразования подобия приведена к диагональному виду, т. е. разложена в произведение

$$A^*A = U^*S U$$

с унитарной матрицей  $U$  и диагональной матрицей  $S$  (см. § 3.2Г). При этом, как мы уже знаем из предложения 3.2.4, в  $S$  по диагонали стоят квадраты сингулярных чисел  $\sigma_i$  матрицы  $A$ , т. е.

$$S = \text{diag} \{ \sigma_1^2, \sigma_2^2, \dots, \sigma_n^2 \}.$$

Рассмотрим произведение  $(Av)^*(Av)$ , результат которого является числом. С одной стороны,

$$(Av)^*(Av) = (\lambda v)^*(\lambda v) = (\lambda^* \lambda)(v^* v) = |\lambda|^2. \quad (3.23)$$

С другой стороны,

$$\begin{aligned} (Av)^*(Av) &= v^* A^* Av = v^* U^* S U v = (Uv)^* S (Uv) = \\ &= w^* S w = \sum_{i=1}^n \sigma_i^2 w_i^* w_i = \sum_{i=1}^n \sigma_i^2 |w_i|^2, \end{aligned}$$

где сделана замена переменных  $w = Uv$ . При этом

$$w^* w = \sum_{i=1}^n |w_i|^2 = 1,$$

так как матрица  $U$  унитарна и  $v^* v = 1$ . Поэтому ясно, что

$$\sigma_{\min}^2(A) \leq (Av)^*(Av) = \sum_{i=1}^n \sigma_i^2 |w_i|^2 \leq \sigma_{\max}^2(A).$$

Сопоставляя это неравенство с (3.23), получим доказываемое. ■

**Предложение 3.2.5** *Сингулярные числа матрицы не меняются при умножении её на унитарную матрицу (ортогональную в вещественном случае).*

**Доказательство.** Пусть  $A$  — исходная матрица, а  $U$  унитарна. Тогда

$$(AU)^*(AU) = (U^* A^*)(AU) = U^* (A^* A) U = U^{-1} (A^* A) U,$$

и потому матрица  $(AU)^*(AU)$  подобна матрице  $A^* A$ . Как следствие, она имеет те же собственные значения. Кроме того,

$$(AU)(AU)^* = AUU^* A^* = AA^*.$$

Привлекая предложение 3.2.4, можем заключить, что сингулярные числа матриц  $AU$  и  $A$  тоже должны совпадать.

Для умножения слева, т. е. для матрицы  $UA$ , обоснование аналогично. ■

**Предложение 3.2.6** *Если  $\sigma$  — сингулярное число неособенной квадратной матрицы, то  $\sigma^{-1}$  — это сингулярное число обратной к ней матрицы.*

**Доказательство.** Вспомним, что собственные числа взаимно обратных матриц обратны друг другу (предложение 3.2.2). Применяя это соображение к матрице  $A^*A$ , можем заключить, что если  $\lambda_1, \lambda_2, \dots, \lambda_n$  — её собственные значения, то у обратной матрицы  $(A^*A)^{-1} = A^{-1}(A^*)^{-1}$  собственными значениями являются  $\lambda_1^{-1}, \lambda_2^{-1}, \dots, \lambda_n^{-1}$ . Но  $A^{-1}(A^*)^{-1} = A^{-1}(A^{-1})^*$ , а потому в силу предложения 3.2.4 выписанные числа  $\lambda_1^{-1}, \lambda_2^{-1}, \dots, \lambda_n^{-1}$  образуют набор квадратов сингулярных чисел матрицы  $A^{-1}$ . Это и требовалось показать. ■

### 3.2ж Сингулярное разложение матриц

Важнейший результат, касающийся сингулярных чисел и сингулярных векторов матриц, который служит одной из основ их широкого применения в разнообразных вопросах математики и её приложений — это

**Теорема 3.2.2** (теорема о сингулярном разложении матрицы)

*Для любой комплексной  $m \times n$ -матрицы  $A$  существуют унитарные  $m \times m$ -матрица  $U$  и  $n \times n$ -матрица  $V$ , такие что*

$$A = U\Sigma V^* \quad (3.24)$$

*с диагональной  $m \times n$ -матрицей*

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \end{pmatrix},$$

где  $\sigma_1, \sigma_2, \dots, \sigma_{\min\{m,n\}}$  — сингулярные числа матрицы  $A$ , а столбцы матриц  $U$  и  $V$  являются соответственно левыми и правыми сингулярными векторами матрицы  $A$ .

Представление (3.24) называется *сингулярным разложением матрицы*  $A$ . Если  $A$  — вещественная матрица, то  $U$  и  $V$  также являются вещественными ортогональными матрицами и сингулярное разложение принимает вид

$$A = U \Sigma V^\top.$$

Для квадратных матриц доказательство сингулярного разложения может быть легко выведено из известного полярного разложения матрицы, т. е. её представления в виде

$$A = QS,$$

где  $Q$  — ортогональная матрица, а  $S$  — симметричная положительно полуопределённая (в комплексном случае  $Q$  унитарна, а  $S$  эрмитова); см., к примеру, [9, 24, 54, 73]. Рассмотрим подробно этот вывод для случая комплексных квадратных матриц.

Как известно, любую эрмитову матрицу можно унитарными преобразованиями подобия привести к диагональному виду, так что  $S = T^*DT$ , где  $T$  — унитарная, а  $D$  — диагональная. При этом на диагонали в  $D$  стоят вещественные собственные числа матрицы  $S$ . Поэтому  $A = (QT^*)DT$ . Это уже почти требуемое представление для  $A$ , поскольку произведение унитарных матриц  $Q$  и  $T^*$  тоже унитарно. Нужно лишь убедиться в том, что диагональные элементы в  $D$  — это сингулярные числа матрицы  $A$ .

Исследуем произведение  $A^*A$ :

$$\begin{aligned} A^*A &= ((QT^*)DT)^*((QT^*)DT) = \\ &= T^*D^*(QT^*)^*(QT^*)DT = \\ &= T^*D^*DT = T^*D^2T = T^{-1}D^2T. \end{aligned}$$

Как видим, матрица  $A^*A$  подобна диагональной матрице  $D^2$ , их собственные числа поэтому совпадают. Следовательно, собственные числа  $A^*A$  суть квадраты диагональных элементов  $D$ . Это и требовалось доказать.

**Доказательство** теоремы о сингулярном разложении для произвольных матриц основано на результатах предшествующего § 3.2д — предложении 3.2.4 и следствии из него.

Пусть  $A$  —  $m \times n$ -матрица, причём для определённости предположим, что  $m \geq n$ . Для неё существуют сингулярные числа  $\sigma_1, \sigma_2, \dots, \sigma_n$ , а также соответствующие им левые и правые сингулярные векторы, которые будем обозначать

$$u^{(1)}, u^{(2)}, \dots, u^{(n)} \quad \text{и} \quad v^{(1)}, v^{(2)}, \dots, v^{(n)}.$$

Размерность левых сингулярных векторов равна  $m$ , а правых —  $n$ . Кроме того, согласно предложению 3.2.4 эти семейства векторов можно взять ортогональными и нормированными в евклидовой норме.

В силу основной системы уравнений (3.17), определяющей сингулярные числа и векторы,

$$Av^{(1)} = \sigma_1 u^{(1)}, \quad Av^{(2)} = \sigma_2 u^{(2)}, \quad \dots, \quad Av^{(n)} = \sigma_n u^{(n)}. \quad (3.25)$$

Эти  $n$  штук матрично-векторных равенств можно записать в виде одного матричного равенства

$$AV = \tilde{U}\tilde{\Sigma}, \quad (3.26)$$

если ввести диагональную  $n \times n$ -матрицу  $\tilde{\Sigma} = \text{diag}\{\sigma_1, \sigma_2, \dots, \sigma_n\}$  и матрицы  $\tilde{U}$  и  $V$ , составленные из векторов  $u^{(i)}$  и  $v^{(i)}$  как из столбцов. При этом  $\tilde{U}$  — матрица размера  $m \times n$ , а  $V$  — матрица размера  $n \times n$ .

Матрица  $\tilde{U}$  имеет  $n$  нормированных ортогональных столбцов размерности  $m$  и  $m \geq n$ . Если  $m > n$ , то мы можем дополнить существующие  $n$  столбцов до ортонормального базиса пространства  $\mathbb{R}^m$ , получив в целом унитарную  $m \times m$ -матрицу  $U$  (ортогональную в вещественном случае). Тогда все равенства (3.25), как и (3.26), равносильны одному матричному равенству

$$AV = U\Sigma,$$

где  $m \times n$ -матрица  $\Sigma$  определена следующим образом:

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix},$$

В целом получаем равенство  $A = U\Sigma V^*$ , в котором все сомножители правой части удовлетворяют условиям теоремы.

Случай  $m \leq n$ , т. е. когда матрица — квадратная или «лежачая», рассматривается совершенно аналогично. ■

Другие доказательства теоремы о сингулярном разложении для случая общих прямоугольных матриц можно найти, к примеру, в книгах [11, 41, 43]. Фактически этот результат показывает, как с помощью сингулярных чисел матрицы элегантно представляется действие соответствующего линейного оператора из одного векторного пространства в другое. Именно, для любого линейного отображения можно выбрать ортонормированный базис в пространстве области определения и ортонормированный базис в пространстве области значений так, чтобы в этих базисах рассматриваемое отображение представлялось растяжениями вдоль координатных осей. Сингулярные числа матрицы оказываются, как правило, адекватным инструментом её исследования, когда соответствующее линейное отображение действует из одного векторного пространства в другое, возможно, с отличающейся размерностью. Собственные числа матрицы полезны при изучении линейного преобразования векторного пространства в пространство той же размерности, в частности, самого в себя. Дальнейшие примеры применения сингулярных чисел и сингулярных векторов матриц рассматриваются ниже в § 3.5.

Для получения сингулярного разложения матриц разработано немало эффективных вычислительных алгоритмов, которые воплощены в надёжные программы для ЭВМ. Большинство из них входит в состав популярных систем компьютерной математики, пакетов программ вычислительной линейной алгебры и др. Как правило, эти программы для вычисления сингулярного разложения называются единообразным стандартным именем **svd**, означающим «singular value decomposition».

**Пример 3.2.5** Для  $3 \times 2$ -матрицы

$$A = \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 2 \end{pmatrix}$$

применение в системе компьютерной математики Octave процедуры **svd** даёт следующее разложение (то же самое — в системах MATLAB, Scilab, Maple и некоторых других):

```
>> [U,S,V] = svd(A)
```

```
U =
-0.3105  0.7551 -0.5774
-0.4987 -0.6465 -0.5774
-0.8092  0.1087  0.5774
```

```
S =
Diagonal Matrix
4.4552      0
  0   0.3888
  0      0
```

```
V =
-0.8385 -0.5449
-0.5449  0.8385
```

(результат выведен с точностью пяти знаков, что делается в большинстве систем по умолчанию).

Как видим, сингулярные числа матрицы равны 4.4552 и 0.3888, а сингулярные векторы — это столбцы матрицы  $U$  (левые сингулярные векторы) и строки матрицы  $V$  (правые сингулярные векторы). Строки вместо столбцов в последнем случае взяты потому, что интерфейс процедуры `svd` не вполне согласуется с теоремой о сингулярном разложении, где матрица  $V$  берётся с эрмитовым сопряжением (или транспонированием). ■

Сингулярное разложение матриц впервые возникло во второй половине XIX века в трудах Э. Бельтрами и К. Жордана, но термин «*valeurs singulières*» — «сингулярные значения» — впервые использовал французский математик Э. Пикар около 1910 года в работе по интегральным уравнениям [121]. Задача нахождения сингулярных чисел и сингулярных векторов матриц, последняя из списка на стр. 323, как может показаться, является частным случаем третьей задачи, относящейся к нахождению собственных чисел и собственных векторов. Но это не так ни в теории, ни на практике, где отыскание сингулярных чисел и сингулярных векторов матриц сделались в настоящее время чрезвычайно важными приложениями вычислительной линейной алгебры. С другой стороны, соответствующие численные методы весьма специализированы, так что эта задача в общем списке задач уже выделяется отдельным пунктом.

Комментируя современный список задач вычислительной линейной алгебры из § 3.1, можно также отметить, что на первые места в нём выдвинулась линейная задача наименьших квадратов. А некоторые старые и популярные ранее задачи как бы отошли на второй план, что стало отражением значительных изменений в математических моделях и вычислительных технологиях решения современных практических задач. Это естественный процесс, в котором большую роль сыграло развитие вычислительной техники и информатики. Следует быть готовым к подобным изменениям и в будущем.

### 3.2з Системы линейных алгебраических уравнений

В этом разделе конспективно излагаются сведения по теории систем линейных алгебраических уравнений, необходимые для понимания материала книги.

Систему линейных алгебраических уравнений

$$\left\{ \begin{array}{l} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = b_2, \\ \vdots \qquad \vdots \qquad \ddots \qquad \vdots \qquad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n = b_m \end{array} \right.$$

относительно неизвестных переменных  $x_1, x_2, \dots, x_n$ , имеющую коэффициенты  $a_{ij}$  и правые части  $b_i$ , можно записать в краткой матрично-векторной форме как

$$Ax = b,$$

где  $A = (a_{ij})$  —  $m \times n$ -матрица коэффициентов и  $b = (b_1, b_2, \dots, b_m)^\top$  —  $m$ -вектор правых частей,  $x = (x_1, x_2, \dots, x_n)^\top$  —  $n$ -вектор неизвестных.

Система линейных алгебраических уравнений называется *однородной*, если все её правые части нулевые, т. е.  $b = 0$ . Иначе, если хотя бы одна правая часть не равна нулю, система называется *неоднородной*.

*Решением* системы линейных алгебраических уравнений называется набор значений неизвестных переменных  $x_1, x_2, \dots, x_n$ , удовлетворяющих каждому из уравнений системы. Если система имеет хотя бы одно решение, она называется *разрешимой* или *совместной*. Иначе, если решений у системы нет, она называется *неразрешимой* или *несовместной*.

Если система линейных алгебраических уравнений является разрешимой (совместной), то всякое её отдельное решение называется *частным решением*. Множество всех частных решений называется *общим решением*.

Для однородной системы линейных алгебраических уравнений общее решение образует линейное подпространство в арифметическом пространстве  $\mathbb{R}^n$  или  $\mathbb{C}^n$ . Любой базис этого подпространства решений называется фундаментальной системой решений. Общее решение неоднородной системы получается как сумма любого его частного решения и общего решения однородной линейной системы с той же матрицей и нулевой правой частью.

Другой равносильный вид системы линейных алгебраических уравнений —

$$\begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{pmatrix} x_1 + \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{pmatrix} x_2 + \cdots + \begin{pmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{pmatrix} x_n = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix},$$

где матрица коэффициентов расчленена на отдельные вектор-столбцы. Из этого представления видно, что система разрешима тогда и только тогда, когда вектор её правой части принадлежит линейной оболочке вектор-столбцов матрицы системы. Коэффициентами соответствующей линейной комбинации являются компоненты искомого вектора решения  $x$ , если оно существует.

Отмеченный факт обосновывает следующий фундаментальный результат о разрешимости систем линейных уравнений:

**Теорема 3.2.3** (теорема Кронекера–Капелли) *Система линейных алгебраических уравнений  $Ax = b$  имеет решение тогда и только тогда, когда ранг матрицы  $A$  коэффициентов системы равен рангу расширенной матрицы  $(A | b)$ , полученной приписыванием к матрице  $A$  вектор-столбца правой части  $b$ .*

Детальные доказательства можно увидеть в [7, 24, 43, 68].

Если рассматривается система линейных алгебраических уравнений  $Ax = b$  с квадратной матрицей  $A$ , у которой число строк равно числу столбцов, то для её разрешимости достаточно того, чтобы в матрице коэффициентов  $A$  все столбцы были линейно независимы, т. е. матрица была неособенной,  $\det A \neq 0$ . Необходимым это условие в общем случае

не является, но если решение единственное, то матрица  $A$  обязана быть неособенной.

В целом, если квадратная матрица  $A$  неособенна, т. е.  $\det A \neq 0$ , то система линейных алгебраических уравнений  $Ax = b$  имеет единственное решение для любой правой части  $b$ . Оно может быть выражено как  $A^{-1}b$  через обратную к  $A$  матрицу. Кроме того, явные формулы для решения даются *правилом Крамера*:  $i$ -е неизвестное равно

$$x_i = \frac{\det A_i}{\det A}, \quad i = 1, 2, \dots, n,$$

где  $A_i$  — матрица, полученная из  $A$  заменой  $i$ -го столбца на столбец правых частей  $b$ .

Для общих, не обязательно квадратных, систем линейных алгебраических уравнений, полезна

**Теорема 3.2.4** (теорема Фредгольма) *Система линейных алгебраических уравнений  $Ax = b$  совместна тогда и только тогда, когда вектор правой части  $b$  ортогонален каждому решению транспонированной однородной системы уравнений  $A^\top y = 0$ .*

Доказательство можно найти в [3, 6, 7, 68]. Отметим, что эта формулировка теоремы Фредгольма является конечномерным вариантом более общего утверждения о разрешимости интегральных и некоторых операторных уравнений. Иногда результат формулируют в виде «альтернативы Фредгольма»: либо уравнение  $Ax = b$  имеет решение при любой правой части  $b$ , либо однородное транспонированное уравнение  $A^\top y = 0$  имеет нетривиальное решение.

### 3.3 Нормы векторов и матриц

#### 3.3а Векторные нормы

Норму можно рассматривать как обобщение понятия абсолютной величины числа на многомерный и абстрактный случаи. Вообще, и норма, и абсолютная величина являются понятиями, которые формализуют интуитивно ясное свойство «размера» объекта, его «величины», т. е. того, насколько он мал или велик безотносительно к его расположению в пространстве или к другим второстепенным качествам. Такова, например, длина вектора как направленного отрезка в привычном нам евклидовом пространстве.

Формальное определение нормы даётся следующим образом:

**Определение 3.3.1** Нормой в вещественном или комплексном линейном векторном пространстве  $\mathcal{X}$  называется вещественнозначная функция  $\|\cdot\|$ , удовлетворяющая следующим свойствам (называемым аксиомами нормы):

$$(BH1) \quad \|a\| \geq 0 \quad \text{для любого } a \in \mathcal{X}, \text{ причём } \|a\| = 0 \Leftrightarrow a = 0$$

— неотрицательность;

$$(BH2) \quad \|\alpha a\| = |\alpha| \cdot \|a\| \quad \text{для любых } a \in \mathcal{X} \text{ и } \alpha \in \mathbb{R} \text{ или } \mathbb{C}$$

— абсолютная однородность;

$$(BH3) \quad \|a + b\| \leq \|a\| + \|b\| \quad \text{для любых } a, b \in \mathcal{X}$$

— «неравенство треугольника».

Само пространство  $\mathcal{X}$  с нормой называется при этом нормированным линейным пространством.

Далее в качестве конкретных линейных векторных пространств у нас, как правило, всюду рассматриваются арифметические пространства  $\mathbb{R}^n$  или  $\mathbb{C}^n$ .

Не все нормы, удовлетворяющие выписанным аксиомам одинаково практически, и часто от нормы требуют выполнения ещё тех или иных дополнительных условий. К примеру, удобно иметь дело с *абсолютной нормой*, значение которой зависит лишь от абсолютных значений компонент векторов. В общем случае норма вектора этому условию может и не удовлетворять.

Приведём примеры наиболее часто используемых норм векторов в  $\mathbb{R}^n$  и  $\mathbb{C}^n$ . Если  $a = (a_1, a_2, \dots, a_n)^\top$ , то обозначим

$$\|a\|_1 := \sum_{i=1}^n |a_i|,$$

$$\|a\|_2 := \left( \sum_{i=1}^n |a_i|^2 \right)^{1/2},$$

$$\|a\|_\infty := \max_{1 \leq i \leq n} |a_i|.$$

Вторая из этих норм часто называется *евклидовой*, а третья — *чебышёвской* или *максимум-нормой*. Евклидова норма вектора, как направленного отрезка, — это его обычная длина, в связи с чем евклидову норму часто называют также *длиной вектора*. Нередко можно встретить и другие названия выписанных выше норм.

Замечательность евклидовой нормы  $\|\cdot\|_2$  состоит в том, что она порождается стандартным скалярным произведением в  $\mathbb{R}^n$  или  $\mathbb{C}^n$ . Более точно, если скалярное произведение  $\langle \cdot, \cdot \rangle$  задаётся как (3.5) или (3.4), то

$$\|a\|_2 = \sqrt{\langle a, a \rangle}.$$

Иными словами, 2-норма является составной частью более богатой и содержательной структуры на пространствах  $\mathbb{R}^n$  и  $\mathbb{C}^n$ , чем мы будем неоднократно пользоваться. Напомним, в частности, важное *неравенство Коши–Буняковского*

$$|\langle a, b \rangle| \leq \|a\|_2 \|b\|_2 \quad (3.27)$$

(см. [7, 9, 14, 19, 23, 24, 43]).

Нормы  $\|\cdot\|_1$  и  $\|\cdot\|_2$  — это частные случаи более общей конструкции так называемой *p-нормы*

$$\|a\|_p = \left( \sum_{i=1}^n |a_i|^p \right)^{1/p} \quad \text{для } p \geq 1. \quad (3.28)$$

Неравенство треугольника для неё имеет вид

$$\left( \sum_{i=1}^n |a_i + b_i|^p \right)^{1/p} \leq \left( \sum_{i=1}^n |a_i|^p \right)^{1/p} + \left( \sum_{i=1}^n |b_i|^p \right)^{1/p},$$

оно называется *неравенством Минковского* и имеет самостоятельное значение в различных разделах математики [7, 14, 19, 54]. Чебышёвская норма тоже может быть получена из конструкции *p-нормы* с помощью предельного перехода по  $p \rightarrow \infty$ , что и объясняет индекс « $\infty$ » в её обозначении.

В самом деле,

$$\left( \sum_{i=1}^n |a_i|^p \right)^{1/p} \leq \left( n \left( \max_{1 \leq i \leq n} |a_i| \right)^p \right)^{1/p} = n^{1/p} \max_{1 \leq i \leq n} |a_i|.$$

С другой стороны,

$$\left( \sum_{i=1}^n |a_i|^p \right)^{1/p} \geq \left( \left( \max_{1 \leq i \leq n} |a_i| \right)^p \right)^{1/p} = \max_{1 \leq i \leq n} |a_i|,$$

так что в целом

$$\max_{1 \leq i \leq n} |a_i| \leq \left( \sum_{i=1}^n |a_i|^p \right)^{1/p} \leq n^{1/p} \max_{1 \leq i \leq n} |a_i|.$$

При переходе в этом двойном неравенстве к пределу по  $p \rightarrow \infty$  оценки снизу и сверху сливаются друг с другом, и потому действительно

$$\lim_{p \rightarrow \infty} \left( \sum_{i=1}^n |a_i|^p \right)^{1/p} = \max_{1 \leq i \leq n} |a_i|.$$

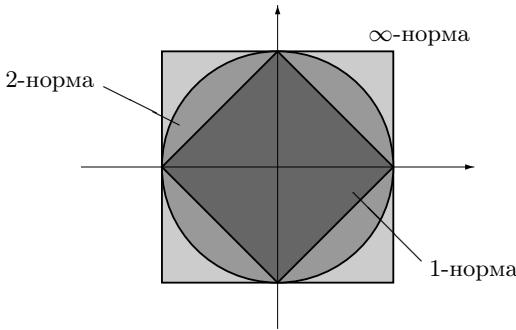


Рис. 3.6. Шары единичного радиуса в различных нормах

В нормированном пространстве  $\mathcal{X}$  шаром радиуса  $r$  с центром в точке  $a$  называется множество

$$\mathbb{S}_r(a) := \{ x \in \mathcal{X} \mid \|x - a\| \leq r \}. \quad (3.29)$$

Геометрически наглядное представление о норме даётся её единичным шаром, т. е. множеством  $\{ x \in \mathcal{X} \mid \|x\| \leq 1 \}$ . На рис. 3.6 изображены единичные шары для рассмотренных выше норм в  $\mathbb{R}^2$ . Из аксиом нормы вытекает, что единичный шар любой нормы — это множество в

линейном векторном пространстве, которое выпукло (следствие неравенства треугольника) и *уравновешено*, т. е. переходит в себя при умножении на любой скаляр  $\alpha$  с  $|\alpha| \leq 1$  (следствие абсолютной однородности). Единичный шар нормы своей формой показывает изменение значений нормы в зависимости от направления вектора, давая общее наглядное представление о её характере, и это может помочь при выборе конкретной нормы, наиболее подходящей для той или иной задачи.

Нередко используются взвешенные (масштабированные) варианты норм векторов, в выражениях для которых каждая компонента берётся с каким-то положительным весовым коэффициентом, отражающим его индивидуальный вклад в рассматриваемую модель. В частности, взвешенная чебышёвская норма определяется для положительного весового вектора  $(\gamma_1, \gamma_2, \dots, \gamma_n)$ ,  $\gamma_i > 0$  как

$$\|a\|_{\infty, \gamma} := \max_{1 \leq i \leq n} |\gamma_i a_i|.$$

Её единичные шары — различные прямоугольные брусы с гранями, параллельными координатным осям, т. е. прямые произведения интервалов вещественной оси (рис. 3.7). Они являются важнейшим частным случаем многомерных интервалов [105], в связи с чем взвешенная чебышёвская норма популярна в интервальном анализе.

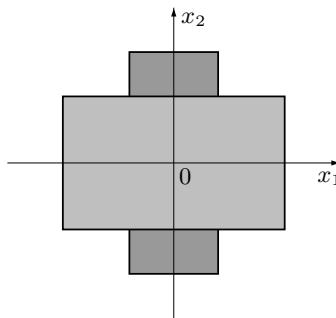


Рис. 3.7. Шары единичного радиуса для взвешенных чебышёвских норм

Обобщением конструкции взвешенных норм может служить норма, связанная с некоторой фиксированной неособенной матрицей. Именно, если  $\|\cdot\|$  — какая-либо векторная норма в  $\mathbb{R}^n$  или  $\mathbb{C}^n$ , а  $S$  — неособенная  $n \times n$ -матрица, то можно определить норму векторов как  $\|x\|_S = \|Sx\|$ .

Нетрудно проверить, что все аксиомы векторной нормы удовлетворяются для  $\|\cdot\|_S$ . Мы воспользуемся такой нормой ниже в § 3.106.

### 3.3б Топология на векторных пространствах

*Топологической структурой* или просто *топологией*<sup>6</sup> на заданном множестве называют специальную структуру, позволяющую говорить об «окрестностях» точек, «сходимости», «неограниченном приближении» точки к чему-либо, организовывать предельные переходы и тому подобные конструкции. В абстрактной форме для наделения множества топологической структурой достаточно каким-либо образом выделить в нём класс «открытых подмножеств», т. е. таких, что каждая их точка «со всех сторон окружена» точками этого же множества. Обычно это делают, опираясь на формальные свойства открытых множеств — сохранение свойства открытости при объединениях и конечных пересечениях.

*Окрестностью* точки в топологическом пространстве называется всякое открытое множество, содержащее эту точку. Окрестностью подмножества топологического пространства называется всякое содержащее его открытое множество. Задание окрестностей точек и множеств позволяет определять приближение одного элемента множества к другому, предельные переходы, сходимость и другие подобные понятия [14, 19]. Топологическую структуру (топологию) можно задавать различными способами, например простым описанием того, какие именно множества считаются открытыми.

В практике математического моделирования более распространено задание топологической структуры не сформулированным выше абстрактным способом, а при помощи функции расстояния (метрики) или с помощью различных норм. Преимущество этого пути состоит в том, что мы получаем в своё распоряжение количественную меру близости рассматриваемых объектов, а некоторые важные топологические конструкции решительно упрощаются.

В свою очередь, на нормированном линейном пространстве  $X$  с нормой  $\|\cdot\|$  расстояние (метрика) между элементами  $a$  и  $b$  может быть естественно задано как

$$\text{dist}(a, b) = \|a - b\|, \quad (3.30)$$

---

<sup>6</sup>Топологией называется также математическая дисциплина, изучающая главным образом свойства объектов, инвариантные относительно непрерывных отображений (см., к примеру, [63, 70]).

т. е. как «величина различия» элементов  $a$  и  $b$ . Непосредственной проверкой легко убедиться, что для введённой таким образом функции  $\text{dist} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$  выполняются все аксиомы расстояния (мы приводили их ранее на стр. 70).

Таким образом, нормы будут нужны нам как сами по себе, для оценивания «величины» тех или иных объектов, так и для измерения «отклонения» одного вектора от другого, иными словами, расстояния между векторами. С помощью определения (3.30) линейное векторное пространство с нормой превращается в метрическое пространство. В целом задание нормы на некотором линейном векторном пространстве автоматически определяет на нём и топологию, т. е. запас открытых и замкнутых множеств, структуру близости, с помощью которой можно будет, в частности, выполнять предельные переходы.

В метрическом пространстве  $\mathcal{X}$  шар радиуса  $r$  с центром в точке  $a$  определяется как множество  $\mathbb{W}_r(a) = \{x \in \mathcal{X} \mid \text{dist}(x, a) \leq r\}$ , образованное всеми точками, находящимися на расстоянии не более  $r$  от  $a$ . Аналогично, как (3.29), определяется шар в нормированном пространстве. Соответственно, *открытыми множествами* метрического или нормированного пространства считаются такие множества, каждая точка которых принадлежит множеству вместе с некоторым шаром с центром в этой точке. Антиподами открытых множеств являются *замкнутые множества*, которые неформально можно описать как «множества с чётко очерченной границей». Они определяются как теоретико-множественные дополнения открытых множеств.

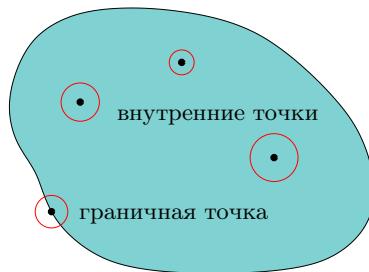


Рис. 3.8. Внутренние и граничные точки множества с их окрестностями в виде шаров евклидовой нормы

Точка множества называется *внутренней*, если она со всех сторон окружена точками этого же множества. Формально математически

точка является внутренней, если любой шар достаточно малого радиуса с центром в этой точке целиком лежит во множестве (рис. 3.8). Совокупность внутренних точек множества  $X$  называют *внутренностью*  $X$ , и мы будем обозначать её  $\text{int } X$  (от латинского и английского терминов «*interior*»).

Точка называется *граничной* для множества, если она окружена как точками самого множества, так и точками не из данного множества. Формально математически точка является граничной, если любой шар с центром в этой точке содержит как точки из множества, так и точки, не принадлежащие множеству (рис. 3.8). Совокупность граничных точек множества  $X$  называют его *границей* и обычно обозначают  $\partial X$ . Границные точки могут не принадлежать самому множеству, но если множество содержит свою границу, то оно является замкнутым. Соответственно, *замыканием* множества  $X$ , обозначаемым  $\text{cl } X$  (от английского термина «*closure*»), называется объединение множества со всеми его граничными точками.

**Пример 3.3.1** Внутренностью интервала  $\mathbf{a} = [\underline{a}, \bar{a}]$  вещественной оси является открытый интервал  $(\underline{a}, \bar{a}) \subset \mathbb{R}$ , т. е.

$$\text{int } \mathbf{a} = \text{int } [\underline{a}, \bar{a}] = (\underline{a}, \bar{a}) = \{ a \in \mathbb{R} \mid \underline{a} < a < \bar{a} \}.$$

Границей интервала является двухточечное множество  $\{\underline{a}, \bar{a}\}$ , состоящее из его концов.

Внутренностью интервального вектора-брюса  $\mathbf{a} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n)^\top$  является прямое декартово произведение открытых интервалов компонент:

$$\text{int } \mathbf{a} = \text{int } \mathbf{a}_1 \times \text{int } \mathbf{a}_2 \times \dots \times \text{int } \mathbf{a}_n.$$

Мы записываем его в теоретико-множественной форме, а не в виде вектора, так как формально не определяли подобные объекты.

Границей интервального вектора-брюса является объединение всех возможных множеств вида

$$\mathbf{a}_1 \times \dots \times \mathbf{a}_{i-1} \times \underline{a}_i \times \mathbf{a}_{i+1} \times \dots \times \mathbf{a}_n, \quad \mathbf{a}_1 \times \dots \times \mathbf{a}_{i-1} \times \bar{a}_i \times \mathbf{a}_{i+1} \times \dots \times \mathbf{a}_n,$$

т. е. декартовых произведений  $n - 1$  интервалов-компонент на концы интервалов одной оставшейся компоненты. ■

Классификация точек множества на внутренние и граничные важна с практической точки зрения. Внутренние точки остаются во множестве при любых достаточно малых возмущениях, тогда как граничные

точки ведут себя существенно по-другому: они могут покинуть множество при сколь угодно малых возмущениях. Получается, что принадлежность внутренней точки множеству «устойчива» при возмущениях, а граничной — нет.

Вообще, язык общей топологии, элементы которого изложены выше в этом разделе, оказывается чрезвычайно полезным при описании различных свойств математических объектов и их совокупностей. Например, чтобы подчеркнуть, что некоторое множество является очень малым по составу и «непредставительным», часто может быть использовано понятие *нигде не плотного подмножества*. Так называют множества у которых внутренность замыкания пуста. Получается, что даже после замыкания такие множества не содержат ни одного шара (хотя бы даже очень малого радиуса). Например, именно таким является множество всех особенных матриц в пространстве всех квадратных матриц.

Напротив, точками *всюду плотного подмножества*, замыкание которого совпадает со всем рассматриваемым множеством, можно сколь угодно точно приблизить его. Например, нетрудно показать, что во множестве всех квадратных матриц неособые матрицы образуют всюду плотное подмножество.

**Определение 3.3.2** Говорят, что в нормированном пространстве  $X$  с нормой  $\|\cdot\|$  последовательность  $\{a^{(k)}\}_{k=1}^{\infty}$  сходится к пределу  $a^*$  по норме (или относительно рассматриваемой нормы), если числовая последовательность  $\|a^{(k)} - a^*\|$  сходится к нулю.

Важнейшее место среди метрических пространств занимают полные метрические пространства, которые неформально можно определить как пространства «без дырок» и без выколотых точек, т. е. такие, что всякая «сходящаяся в себе» последовательность точек этого пространства имеет в нём предел. Напомним, что «сходящаяся в себе» последовательность в метрическом пространстве (называемая также фундаментальной последовательностью или последовательностью Коши) — это последовательность, в которой изменения членов становятся сколь угодно малыми с ростом их номера и которая явно должна сходиться к какому-то пределу, если судить по её поведению и внутренним свойствам.

Математически условие «сходимости в себе» обычно формализуют следующим образом: последовательность  $\{x_k\}$  в пространстве  $V$  с расстоянием  $\text{dist}$  называется *фундаментальной* (сходящейся в себе или

последовательностью Коши), если для любого  $\varepsilon > 0$  существует такое натуральное число  $N$ , что  $\text{dist}(x_m, x_n) < \varepsilon$  для любых  $m, n > N$ .

**Определение 3.3.3** Метрическое пространство, в котором последовательность элементов сходится к некоторому пределу из этого пространства тогда и только тогда, когда она является фундаментальной (последовательностью Коши и т. п.), называется полным.

Как показывается в курсах математического анализа, любая сходящаяся в себе последовательность вещественных чисел имеет пределом вещественное же число. Иными словами, вещественная ось  $\mathbb{R}$  с естественным расстоянием между числами  $\text{dist}(x, y) = |x - y|$  является полным метрическим пространством [14]. С другой стороны, множество рациональных чисел этим свойством уже не обладает, так как некоторые последовательности рациональных чисел, удовлетворяющие условиям критерия Коши, рационального предела не имеют. Таковы, например, последовательности рациональных приближений к решению уравнения  $x^2 - 2 = 0$ , равному  $\sqrt{2}$ .

Арифметические векторные пространства  $\mathbb{R}^n$  и  $\mathbb{C}^n$  являются полными метрическими пространствами, будучи прямыми произведениями нескольких экземпляров вещественной оси  $\mathbb{R}$  или комплексной плоскости  $\mathbb{C}$ .

Помимо сходимости последовательностей и их пределов часто необходимо рассматривать сходимость непрерывно изменяющихся переменных величин. Для метрических пространств, как показывается в общей топологии, эти две конструкции вполне равносильны и могут быть выведены одна из другой [70]. Формулировки на языке сходимости непрерывно изменяющихся величин иногда бывают всё же более удобны, но мы не будем приводить здесь формального описания соответствующих понятий, так как они существенно сложнее определения сходимости для последовательностей.

### 3.3в Эквивалентные векторные нормы

Нормы в линейном векторном пространстве называются *топологически эквивалентными* (или просто *эквивалентными*), если эквивалентны порождаемые ими топологии, т. е. любое открытое (замкнутое) относительно одной нормы множество является открытым (замкнутым) также в другой норме, и наоборот. При условии эквивалентности

норм, в частности, наличие предела в одной из них влечёт существование того же предела в другой, и обратно. Из математического анализа известен простой критерий эквивалентности двух норм (см., к примеру, [7, 43, 56]):

**Предложение 3.3.1** *Нормы  $\|\cdot\|'$  и  $\|\cdot\|''$  на линейном векторном пространстве  $X$  эквивалентны тогда и только тогда, когда существуют такие положительные константы  $C_1$  и  $C_2$ , что для любых  $a \in X$*

$$C_1\|a\|' \leq \|a\|'' \leq C_2\|a\|'. \quad (3.31)$$

Формулировка этого предложения имеет кажущуюся асимметрию, так как для значений одной из эквивалентных норм предъявляется двусторонняя «вилка» из значений другой нормы с подходящими множителями-константами. Но нетрудно видеть, что из (3.31) вытекает

$$\frac{1}{C_2}\|a\|'' \leq \|a\|' \leq \frac{1}{C_1}\|a\|'',$$

так что существование «вилки» для одной нормы автоматически подразумевает существование аналогичной «вилки» и для другой.  $C_1$  и  $C_2$  обычно называют *константами эквивалентности* норм  $\|\cdot\|'$  и  $\|\cdot\|''$ .

Содержательный смысл предложения 3.3.1 совершенно прозрачен. Если  $C_1\|a\|' \leq \|a\|''$ , то в любой шар ненулевого радиуса в норме  $\|\cdot\|''$  можно вложить некоторый шар в норме  $\|\cdot\|'$ . Если же  $\|a\|'' \leq C_2\|a\|'$ , то верно и обратное: в любой шар относительно нормы  $\|\cdot\|''$  можно поместить какой-то шар относительно нормы  $\|\cdot\|'$ . Как следствие, множество, открытое относительно одной нормы, тоже будет открытым относительно другой, и наоборот. По этой причине одинаковыми окажутся запасы окрестностей любой точки, так что топологические структуры, порождаемые этими двумя нормами, будут эквивалентны друг другу. Наконец, из предложения 3.3.1 немедленно следует, что при эквивалентности двух норм сходимость векторов относительно любой из них в самом деле влечёт сходимость относительно другой нормы.

**Предложение 3.3.2** *В векторных пространствах  $\mathbb{R}^n$  или  $\mathbb{C}^n$*

$$\|a\|_2 \leq \|a\|_1 \leq \sqrt{n}\|a\|_2,$$

$$\|a\|_\infty \leq \|a\|_2 \leq \sqrt{n}\|a\|_\infty,$$

$$\frac{1}{n}\|a\|_1 \leq \|a\|_\infty \leq \|a\|_1,$$

*т. е. векторные 1-норма, 2-норма и  $\infty$ -норма эквивалентны друг другу.*

**Доказательство.** Справедливость правого из первых неравенств следует из неравенства Коши–Буняковского (3.27), применённого к случаю  $b = (\operatorname{sgn} a_1, \operatorname{sgn} a_2, \dots, \operatorname{sgn} a_n)^\top$ . Для обоснования левого из первых неравенств заметим, что в силу определений 2-нормы и 1-нормы

$$\begin{aligned}\|a\|_2^2 &= |a_1|^2 + |a_2|^2 + \dots + |a_n|^2, \\ \|a\|_1^2 &= |a_1|^2 + |a_2|^2 + \dots + |a_n|^2 + \\ &\quad + 2|a_1a_2| + 2|a_1a_3| + \dots + 2|a_{n-1}a_n|\end{aligned}$$

и все слагаемые  $2|a_1a_2|, 2|a_1a_3|, \dots, 2|a_{n-1}a_n|$  неотрицательны. В частности, равенство  $\|a\|_2^2 = \|a\|_1^2$  и ему равносильное  $\|a\|_2 = \|a\|_1$  возможны лишь в случае, когда у вектора  $a$  все компоненты равны нулю, за исключением одной.

Обоснование остальных неравенств даётся следующими несложными выкладками:

$$\begin{aligned}\|a\|_2 &= \sqrt{|a_1|^2 + |a_2|^2 + \dots + |a_n|^2} \geq \\ &\geq \sqrt{\max_i |a_i|^2} = \max_i |a_i| = \|a\|_\infty, \\ \|a\|_2 &= \sqrt{|a_1|^2 + |a_2|^2 + \dots + |a_n|^2} \leq \\ &\leq \sqrt{n \max_i |a_i|^2} = \sqrt{n} \max_i |a_i| = \sqrt{n} \|a\|_\infty, \\ \|a\|_\infty &= \max_i |a_i| \leq \\ &\leq |a_1| + |a_2| + \dots + |a_n| = \|a\|_1, \\ \|a\|_1 &= |a_1| + |a_2| + \dots + |a_n| \leq \\ &\leq n \max_i |a_i| \leq n \|a\|_\infty.\end{aligned}$$

Нетрудно видеть, что все эти неравенства достижимые (точные). ■

Доказанный выше вывод об эквивалентности конкретных норм является частным случаем общего результата математического анализа: *в конечномерном линейном векторном пространстве все нормы топологически эквивалентны друг другу* [21, 43, 54]. Но содержание пред-

ложеия 3.3.2 состоит ещё и в указании конкретных констант эквивалентности норм, от которых существенно зависят различные числовые оценки и вытекающие из них действия по численному решению задач (условия остановки итераций и т. п.).

### 3.3г Покомпонентная сходимость

Любой вектор однозначно представляется своим разложением по какому-то фиксированному базису линейного пространства или, иными словами, своими компонентами-числами в этом базисе. В связи с этим помимо определённой выше в § 3.3б сходимости по норме имеет смысл рассматривать *покомпонентную сходимость*, при которой один вектор считается сходящимся к другому тогда и только тогда, когда все компоненты первого вектора сходятся к соответствующим компонентам второго. Формализацией этих соображений является

**Определение 3.3.4** Говорят, что в линейном векторном пространстве  $\mathcal{X}$  последовательность  $\{a^{(k)}\}_{k=1}^{\infty}$  сходится к пределу  $a^*$  покомпонентно (покомпонентным образом) относительно некоторого базиса, если при разложении  $a^{(k)}$  по этому базису для каждого индекса  $i$  имеет место сходимость соответствующей компоненты  $a_i^{(k)} \rightarrow a_i^* \in \mathbb{R}$  или  $\mathbb{C}$  при  $k \rightarrow \infty$ .

Интересен вопрос о том, как соотносятся между собой сходимость по норме и сходимость всех компонент вектора.

**Предложение 3.3.3** В конечномерных линейных векторных пространствах сходимость по норме и покомпонентная сходимость векторов равносильны друг другу.

**Доказательство.** Пусть  $\{a^{(k)}\}$  — последовательность векторов из  $n$ -мерного линейного пространства и она сходится к пределу  $a^*$  в покомпонентном смысле относительно базиса  $\{e_i\}_{i=1}^n$ . Разлагая  $a^{(k)}$  и  $a^*$  в этом базисе, получаем

$$\begin{aligned} \|a^{(k)} - a^*\| &= \left\| \sum_{i=1}^n a_i^{(k)} e_i - \sum_{i=1}^n a_i^* e_i \right\| = \left\| \sum_{i=1}^n (a_i^{(k)} - a_i^*) e_i \right\| \leq \\ &\leq \sum_{i=1}^n \|(a_i^{(k)} - a_i^*) e_i\| = \sum_{i=1}^n |a_i^{(k)} - a_i^*| \|e_i\|. \end{aligned}$$

Тогда, если  $a_i^{(k)}$  сходятся к  $a_i^*$  для любого индекса  $i = 1, 2, \dots, n$ , то и  $\|a^{(k)} - a^*\| \rightarrow 0$ .

Обратно, предположим, что имеет место сходимость  $a^{(k)}$  к  $a^*$  относительно какой-то нормы  $\|\cdot\|$ . Если  $a_i$  — коэффициенты разложения вектора  $a$  по рассматриваемому базису  $\{e_i\}_{i=1}^n$ , то определим связанную с этим базисом величину

$$\|a\|_{\infty}^{\{e_i\}} := \max_{1 \leq i \leq n} |a_i|.$$

Нетрудно убедиться, что она тоже является нормой. Кроме того, из факта эквивалентности норм  $\|\cdot\|$  и  $\|\cdot\|_{\infty}^{\{e_i\}}$  в конечномерном линейном векторном пространстве следует существование такой положительной константы  $C$ , что

$$\max_i |a_i^{(k)} - a_i^*| = \|a^{(k)} - a^*\|_{\infty}^{\{e_i\}} \leq C \|a^{(k)} - a^*\|.$$

Поэтому при  $\|a^{(k)} - a^*\| \rightarrow 0$  обязательно должна быть сходимость компонент  $a_i^{(k)}$  к  $a_i^*$  для всех индексов  $i$ . ■

Хотя сходимость по норме и покомпонентная сходимость равносильны друг другу, нередко бывает удобнее воспользоваться какой-нибудь одной из них. Норма является одним числом, указывающим на близость к пределу, и работать с ней поэтому проще. Но рассмотрение сходимости в покомпонентном смысле позволяет расчленить задачу на отдельные числовые компоненты, что иногда также упрощает анализ сходимости. В конце концов в большинстве практических задач линейной алгебры векторы и матрицы — это структурированные конечные массивы чисел, которые мы воспринимаем по их отдельным элементам и компонентам.

Введение на линейном пространстве нормы и, как следствие, задание топологической структуры позволяют говорить о непрерывности тех или иных отображений этого пространства в себя или в другие пространства и множества. Что можно сказать о непрерывности привычных и часто встречающихся отображений?

Покажем непрерывность сложения и умножения на скаляр относительно нормы. Пусть  $a \rightarrow a^*$  и  $b \rightarrow b^*$ , так что  $\|a - a^*\| \rightarrow 0$  и

$\|b - b^*\| \rightarrow 0$ . Тогда

$$\|(a + b) - (a^* + b^*)\| = \|(a - a^*) + (b - b^*)\| \leq \|a - a^*\| + \|b - b^*\| \rightarrow 0,$$

$$\|\alpha a - \alpha a^*\| = \|\alpha(a - a^*)\| = |\alpha| \|a - a^*\| \rightarrow 0$$

для любого скаляра  $\alpha$ .

Умножение на матрицу тоже непрерывно в конечномерном линейном векторном пространстве. Если  $A — m \times n$ -матрица и  $b$  — такой  $n$ -вектор, что  $b \rightarrow b^*$ , то, зафиксировав индекс  $i \in \{1, 2, \dots, m\}$ , оценим разность  $i$ -х компонент векторов  $Ab$  и  $Ab^*$ :

$$\begin{aligned} |(Ab)_i - (Ab^*)_i| &= |(A(b - b^*))_i| = \left| \sum_{j=1}^n a_{ij}(b_j - b_j^*) \right| \leq \\ &\leq \sqrt{\sum_{j=1}^n a_{ij}^2} \sqrt{\sum_{j=1}^n (b_j - b_j^*)^2} \end{aligned}$$

в силу неравенства Коши–Буняковского. Поэтому  $(Ab)_i \rightarrow (Ab^*)_i$  при  $b \rightarrow b^*$  для любого номера  $i$ . Аналогичной выкладкой нетрудно показать непрерывность стандартного скалярного произведения в  $\mathbb{R}^n$  и  $\mathbb{C}^n$ .

### 3.3д Матричные нормы

Помимо векторов основным объектом вычислительной линейной алгебры являются также матрицы. По этой причине нам будут нужны матричные нормы — для того, чтобы оценивать «величину» той или иной матрицы и, кроме того, чтобы ввести расстояние между матрицами как

$$\text{dist}(A, B) := \|A - B\|, \quad (3.32)$$

где  $A, B$  — вещественные или комплексные матрицы.

Множество матриц само является линейным векторным пространством, а матрица — это составной многомерный объект, в значительной степени аналогичный вектору. Поэтому вполне естественно прежде всего потребовать от матричной нормы тех же свойств, что и для векторной нормы. Формально матричной нормой на множестве вещественных или комплексных  $m \times n$ -матриц называют вещественнозначную функцию  $\|\cdot\|$ , удовлетворяющую следующим условиям (аксиомам нормы):

- (MH1)  $\|A\| \geq 0$  для любой матрицы  $A$ , причём  $\|A\| = 0 \Leftrightarrow A = 0$   
— неотрицательность,
- (MH2)  $\|\alpha A\| = |\alpha| \cdot \|A\|$  для любых матриц  $A$  и  $\alpha \in \mathbb{R}$  или  $\alpha \in \mathbb{C}$   
— абсолютная однородность,
- (MH3)  $\|A + B\| \leq \|A\| + \|B\|$  для любых матриц  $A, B$   
— «неравенство треугольника».

Но условия (MH1)–(MH3) выражают взгляд на матрицу как на «вектор размерности  $m \times n$ ». Они явно недостаточны, если мы хотим учесть специфику матриц как объектов, между которыми определена также операция умножения. Вообще, множество всех квадратных матриц фиксированного размера наделено более богатой структурой, нежели линейное векторное пространство. Обычно в связи с ним используют уже термины «кольцо» или «алгебра», обозначающие множества с двумя взаимосогласованными бинарными операциями — сложением и умножением [24, 43]. Связь нормы матриц с операцией их умножения отражает четвёртая аксиома матричной нормы

- (MH4)  $\|AB\| \leq \|A\| \cdot \|B\|$  для любых матриц  $A, B$   
— «субмультипликативность».<sup>7</sup>

Особую ценность и в теории, и на практике представляют ситуации, когда нормы векторов и матриц, которые рассматриваются совместно друг с другом, существуют не сами по себе, но в некотором смысле согласованы друг с другом. Речь идёт прежде всего об операциях, в которые они вступают вместе друг с другом, т. е. об умножении матрицы на вектор. Инструментом такого согласования может как раз таки выступать аксиома субмультипликативности MH4, которая понимается в расширенном смысле, т. е. для любых матриц  $A$  и  $B$  таких размеров, что произведение  $AB$  имеет смысл. В частности, она должна быть верна для  $n \times 1$ -матриц  $B$ , являющихся векторами из  $\mathbb{R}^n$  или  $\mathbb{C}^n$ .

**Определение 3.3.5** Векторная норма  $\|\cdot\|$  и матричная норма  $\|\cdot\|'$  называются согласованными, если

$$\|Ax\| \leq \|A\|' \cdot \|x\| \tag{3.33}$$

для любой матрицы  $A$  и всех векторов  $x$ .

---

<sup>7</sup>Приставка «суб-» означает «меньше», «ниже» и т. п. В этом смысле неравенства треугольника BH3 и MH3 можно называть «субаддитивностью» норм.

Рассмотрим примеры конкретных матричных норм.

**Пример 3.3.2** Фробениусова норма матрицы  $A = (a_{ij})$  определяется как

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

т. е. как корень из суммы квадратов всех элементов матрицы. Нередко её называют также *нормой Шура*.

Ясно, что она удовлетворяет первым трём аксиомам матричной нормы просто потому, что задаётся совершенно аналогично евклидовой векторной норме  $\|\cdot\|_2$ , т. е. как норма вектора размерности  $m n$ . Для обоснования субмультиплликативности рассмотрим

$$\|AB\|_F^2 = \sum_{i,j} \left| \sum_k a_{ik} b_{kj} \right|^2.$$

В силу неравенства Коши–Буняковского (3.27)

$$\left| \sum_k a_{ik} b_{kj} \right|^2 \leq \left( \sum_k |a_{ik}|^2 \right) \left( \sum_l |b_{lj}|^2 \right),$$

поэтому

$$\begin{aligned} \|AB\|_F^2 &\leq \sum_{i,j} \left( \sum_k |a_{ik}|^2 \right) \left( \sum_l |b_{lj}|^2 \right) = \\ &= \sum_{i,j,k,l} |a_{ik}|^2 |b_{lj}|^2 = \left( \sum_{i,k} |a_{ik}|^2 \right) \left( \sum_{l,j} |b_{lj}|^2 \right) = \\ &= \|A\|_F^2 \|B\|_F^2, \end{aligned}$$

что и требовалось.

Если считать, что  $B$  — это матрица размера  $n \times 1$ , т. е. вектор размерности  $n$ , то выполненные оценки показывают, что фробениусова норма матрицы согласована с евклидовой векторной нормой  $\|\cdot\|_2$ , с которой она совпадает для векторов. ■

Фробениусова норма обладает рядом замечательных свойств, которые мы будем интенсивно использовать в дальнейшем.

**Предложение 3.3.4** Фробениусова норма матрицы не изменяется при умножении на ортогональные матрицы слева или справа.

**Доказательство.** Напомним, что следом квадратной матрицы называется сумма всех её диагональных элементов (стр. 336), и привлечение этого понятия позволяет удобно представить определение фробениусовой нормы матрицы. Если в общей сумме квадратов элементов матрицы выполнить сначала суммирование по строкам, а затем по столбцам, то получим

$$\|A\|_F = \left( \sum_{i,j} a_{ij}^2 \right)^{1/2} = \left( \sum_{j=1}^m \left( \sum_{i=1}^n a_{ij} a_{ij} \right) \right)^{1/2} = (\operatorname{tr}(A^\top A))^{1/2},$$

где  $\operatorname{tr}$  — символ следа матрицы. Если же в общей сумме квадратов элементов матрицы выполнить суммирование сначала по столбцам, а потом по строкам, то получим

$$\|A\|_F = \left( \sum_{i,j} a_{ij}^2 \right)^{1/2} = \left( \sum_{i=1}^n \left( \sum_{j=1}^m a_{ij} a_{ij} \right) \right)^{1/2} = (\operatorname{tr}(AA^\top))^{1/2}.$$

Видим, что следы матриц  $AA^\top$  и  $A^\top A$  совпадают, а фробениусова норма матрицы  $A$  равна корню квадратному из их общего значения.

Далее, для любой ортогональной матрицы  $Q$  справедливо

$$\begin{aligned} \|QA\|_F &= \left( \operatorname{tr}((QA)^\top(QA)) \right)^{1/2} = \\ &= \left( \operatorname{tr}(A^\top Q^\top QA) \right)^{1/2} = (\operatorname{tr}(A^\top A))^{1/2} = \|A\|_F, \end{aligned}$$

$$\begin{aligned} \|AQ\|_F &= \left( \operatorname{tr}((AQ)(AQ)^\top) \right)^{1/2} = \\ &= \left( \operatorname{tr}(AQQ^\top A^\top) \right)^{1/2} = (\operatorname{tr}(AA^\top))^{1/2} = \|A\|_F, \end{aligned}$$

что завершает доказательство предложения. ■

**Следствие.** Фробениусова норма матрицы не меняется при ортогональных преобразованиях подобия.

Другое интересное следствие предложения 3.3.4 и его доказательства состоит в том, что для любой  $m \times n$ -матрицы

$$\|A\|_F = \sqrt{\sigma_1^2(A) + \sigma_2^2(A) + \dots + \sigma_{\min\{m,n\}}^2(A)},$$

т. е. фробениусова норма матрицы равна евклидовой норме вектора из её сингулярных чисел. Это вытекает из сингулярного разложения матрицы.

### Пример 3.3.3 Матричная норма

$$\|A\|_{\max} = n \max_{i,j} |a_{ij}|,$$

определенная на множестве квадратных  $n \times n$ -матриц, является аналогом чебышёвской нормы векторов  $\|\cdot\|_\infty$ , отличаясь от неё лишь постоянным множителем для матриц фиксированного размера. По этой причине выполнение первых трёх аксиом матричной нормы для  $\|A\|_{\max}$  очевидно. То, что в выражении для  $\|A\|_{\max}$  множитель перед  $\max |a_{ij}|$  равен именно  $n$ , объясняется необходимостью удовлетворить аксиоме субмультипликативности:

$$\begin{aligned} \|AB\|_{\max} &= n \max_{i,j} \left| \sum_{k=1}^n a_{ik} b_{kj} \right| \leq n \max_{i,j} \left( \sum_{k=1}^n |a_{ik}| |b_{kj}| \right) \leq \\ &\leq n \left( \sum_{k=1}^n \max_{i,j} \{ |a_{ik}| |b_{kj}| \} \right) \leq \\ &\leq n \left( \sum_{k=1}^n \max_i |a_{ik}| \cdot \max_j |b_{kj}| \right) \leq \\ &\leq n^2 \max_{i,j} |a_{ij}| \max_{i,j} |b_{ij}| = \|A\|_{\max} \|B\|_{\max}. \end{aligned}$$

Ясно, что без этого множителя выписанная выше цепочка неравенств была бы неверной.

Небольшая модификация проведённых выкладок показывает также, что норма  $\|A\|_{\max}$  согласована с чебышёвской нормой векторов. Кроме того, несложно устанавливается, что  $\|A\|_{\max}$  согласована с евклидовой векторной нормой. ■

Последний пример показывает, что аксиома субмультиплекативности МН4 накладывает на матричные нормы более серьёзные ограничения, чем может показаться на первый взгляд. В частности, матричные нормы, в отличие от векторных, нельзя произвольно масштабировать, умножая на какое-то число.

Оказывается, среди матричных норм квадратных матриц нет таких, которые не были бы ни с чем не согласованными. Иными словами, справедливо

**Предложение 3.3.5** Для любой нормы квадратных матриц можно подобрать подходящую норму векторов, с которой матричная норма будет согласована.

**Доказательство.** Для данной нормы  $\|\cdot\|'$  на множестве  $n \times n$ -матриц определим норму  $\|v\|$  для  $n$ -вектора  $v$  как  $\|(v, v, \dots, v)\|'$ , т. е. как норму матрицы  $(v, v, \dots, v)$ , составленной из  $n$  штук векторов  $v$  как из столбцов. Выполнение всех аксиом векторной нормы для  $\|v\|$  очевидным образом следует из аналогичных свойств рассматриваемой нормы матрицы.

Опираясь на субмультиплекативность матричной нормы, имеем

$$\begin{aligned} \|Av\| &= \|(Av, Av, \dots, Av)\|' = \|A \cdot (v, v, \dots, v)\|' \leq \\ &\leq \|A\|' \cdot \|(v, v, \dots, v)\|' = \|A\|' \cdot \|v\|, \end{aligned}$$

так что требуемое согласование действительно будет достигнуто. ■

### 3.3e Подчинённые матричные нормы

В предшествующем разделе мы могли видеть, что с заданной векторной нормой согласованы различные матричные нормы. И наоборот, для матричной нормы возможна согласованность со многими векторными нормами. В этих условиях при проведении различных преобразований и выводе оценок наиболее выгодно оперировать согласованными матричными нормами, которые принимают как можно меньшие значения. Тогда неравенства, получающиеся в результате применения в выкладках соотношения (3.33), будут более точными и позволят получить более тонкие оценки результата. Например, конкретная оценка нормы погрешности может оказывать сильное влияние на количество итераций, которые мы должны будем сделать в итерационном численном методе для достижения заданной точности приближённого решения.

Пусть дана векторная норма  $\|\cdot\|$  и зафиксирована матрица  $A$ . Из требования согласованности (3.33) вытекает неравенство для согласованной нормы матрицы:

$$\|A\| \geq \|Ax\|/\|x\|, \quad (3.34)$$

где  $x$  — произвольный вектор. Как следствие, значения всех матричных норм от  $A$ , согласованных с данной векторной нормой  $\|\cdot\|$ , ограничены снизу выражением

$$\sup_{x \neq 0} \frac{\|Ax\|}{\|x\|},$$

поскольку (3.34) должно быть справедливым для любого ненулевого вектора  $x$ .

**Предложение 3.3.6** Для любой фиксированной векторной нормы  $\|\cdot\|$  соотношением

$$\|A\|' = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} \quad (3.35)$$

задаётся матричная норма.

**Доказательство.** Отметим прежде всего, что в случае конечномерных векторных пространств  $\mathbb{R}^n$  и  $\mathbb{C}^n$  вместо «sup» в выражении (3.35) можно брать «max». В самом деле,

$$\sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{x \neq 0} \left\| A \frac{x}{\|x\|} \right\| = \sup_{\|y\|=1} \|Ay\|,$$

а задаваемая условием  $\|y\|=1$  единичная сфера любой нормы замкнута и ограничена, т. е. компактна в  $\mathbb{R}^n$  или  $\mathbb{C}^n$  [43, 54]. Непрерывная функция  $\|Ay\|$  (см. § 3.3б) достигает на этом компактном множестве своего максимума. Таким образом, в действительности

$$\|A\|' = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|y\|=1} \|Ay\|.$$

Проверим теперь для нашей конструкции выполнение аксиом нормы. Неотрицательность значений  $\|\cdot\|'$  очевидна. Далее, если  $A \neq 0$ , то найдётся ненулевой вектор  $u$ , такой что  $Au \neq 0$ . Ясно, что его можно считать нормированным, т. е.  $\|u\|=1$ . Тогда  $\|Au\| > 0$ , и потому  $\max_{\|y\|=1} \|Ay\| > 0$ , что доказывает для  $\|\cdot\|'$  первую аксиому нормы.

Абсолютная однородность для  $\|\cdot\|'$  доказывается тривиально. Покажем для (3.35) справедливость неравенства треугольника. Очевидно,

$$\|(A + B)y\| \leq \|Ay\| + \|By\|,$$

и потому

$$\begin{aligned} \max_{\|y\|=1} \|(A + B)y\| &\leq \max_{\|y\|=1} (\|Ay\| + \|By\|) \leq \\ &\leq \max_{\|y\|=1} \|Ay\| + \max_{\|y\|=1} \|By\|, \end{aligned}$$

что и требовалось.

Приступая к обоснованию субмультипликативности, отметим, что по самому построению  $\|Ax\| \leq \|A\|' \|x\|$  для любого вектора  $x$ . По этой причине

$$\begin{aligned} \|AB\|' &= \max_{\|y\|=1} \|(AB)y\| = \|ABv\| \quad \text{для некоторого } v \text{ с } \|v\| = 1 \\ &\leq \|A\|' \cdot \|Bv\| \leq \|A\|' \cdot \max_{\|z\|=1} \|Bz\| = \|A\|' \|B\|'. \end{aligned}$$

Это завершает доказательство предложения. ■

Доказанный результат мотивирует

**Определение 3.3.6** Для заданной векторной нормы  $\|\cdot\|$  матричная норма  $\|\cdot\|'$ , определяемая как

$$\|A\|' = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|y\|=1} \|Ay\|,$$

называется матричной нормой, подчинённой норме  $\|\cdot\|$  (или индуцированной нормой  $\|\cdot\|$ ).

Иногда в отношении матричной нормы, которая задаётся определением 3.3.6, используют термин «операторная норма». Он мотивируется тем, что конструкция этой нормы хорошо отражает взгляд на матрицу как на оператор, определяющий отображение линейных векторных пространств. Операторная норма показывает максимальную величину растяжения по норме, которую получает в сравнении с исходным вектором его образ при действии данного оператора.

На основе рассмотренной конструкции можно также определять подчинённые нормы для прямоугольных матриц: для этого требуется взять две векторные нормы — в пространствах, соответствующих строчной и столбцовой размерностям матрицы. Удобно вообще не различать их (допуская определённую вольность речи), если эти две нормы представляют собой варианты одной и той же нормы для разных размерностей. Именно так следует понимать формулировку предложений 3.3.7 и 3.3.8 ниже.

Итак, подчинённые матричные нормы — это минимальные по значениям из согласованных матричных норм. Но, несмотря на хорошие свойства подчинённых матричных норм, их определение не отличается большой конструктивностью, так как привлекает операцию взятия максимума. Естественно задаться вопросом о том, существуют ли вообще достаточно простые и обозримые выражения для матричных норм, подчинённых тем или иным векторным нормам. Какими являются подчинённые матричные нормы для популярных векторных норм  $\|\cdot\|_1$ ,  $\|\cdot\|_2$  и  $\|\cdot\|_\infty$ ? С другой стороны, являются ли рассмотренные выше матричные нормы  $\|A\|_F$  (Фробениусова) и  $\|A\|_{\max}$  подчинёнными для каких-либо векторных норм?

Ответ на последний вопрос отрицателен. В самом деле, для единичной  $n \times n$ -матрицы  $I$  имеем

$$\|I\|_F = \sqrt{n}, \quad \|I\|_{\max} = n,$$

тогда как из определения подчинённой нормы следует, что должно быть

$$\|I\| = \max_{\|y\|=1} \|Iy\| = \max_{\|y\|=1} \|y\| = 1. \quad (3.36)$$

Ответом на первые два вопроса является

**Предложение 3.3.7** Для векторной 1-нормы подчинённой матричной нормой  $m \times n$ -матрицы  $A = (a_{ij})$  является

$$\|A\|_1 = \max_{1 \leq j \leq n} \left( \sum_{i=1}^m |a_{ij}| \right),$$

т. е. максимальная сумма модулей элементов по столбцам.

Для чебышёвской векторной нормы ( $\infty$ -нормы) подчинённой матричной нормой  $m \times n$ -матрицы  $A = (a_{ij})$  является

$$\|A\|_\infty = \max_{1 \leq i \leq m} \left( \sum_{j=1}^n |a_{ij}| \right)$$

— максимальная сумма модулей элементов по строкам.

**Доказательство.** Для обоснования первой части предложения выпишем цепочку преобразований и оценок

$$\begin{aligned}
 \|Ax\|_1 &= \sum_{i=1}^m |(Ax)_i| = \sum_{i=1}^m \left| \sum_{j=1}^n a_{ij} x_j \right| \leq \sum_{i=1}^m \sum_{j=1}^n |a_{ij} x_j| = \\
 &= \sum_{i=1}^m \sum_{j=1}^n |a_{ij}| |x_j| = \sum_{j=1}^n \sum_{i=1}^m |a_{ij}| |x_j| = \sum_{j=1}^n \left( |x_j| \sum_{i=1}^m |a_{ij}| \right) \leq \\
 &\leq \left( \max_{1 \leq j \leq n} \sum_{i=1}^m |a_{ij}| \right) \cdot \sum_{j=1}^n |x_j| = \|A\|_1 \|x\|_1,
 \end{aligned} \tag{3.37}$$

из которой вытекает

$$\frac{\|Ax\|_1}{\|x\|_1} \leq \|A\|_1.$$

При этом все неравенства в цепочке (3.37) обращаются в равенства для вектора  $x$  в виде столбца единичной  $n \times n$ -матрицы с тем номером  $j$ , на котором достигается  $\max_j \sum_{i=1}^m |a_{ij}|$ . Следовательно, на этом векторе достигается наибольшее значение отношения  $\|Ax\|_1/\|x\|_1$  из определения подчинённой матричной нормы.

Аналогичным образом доказывается и вторая часть предложения, касающаяся  $\|\cdot\|_\infty$ . ■

**Предложение 3.3.8** Для евклидовой векторной нормы ( $2$ -нормы) подчинённой нормой матрицы  $A$  является

$$\|A\|_2 = \sigma_{\max}(A)$$

— наибольшее сингулярное число матрицы.

**Доказательство.** Если матрица  $A$  имеет размеры  $m \times n$ , рассмотрим  $n \times n$ -матрицу  $A^*A$ . Она является эрмитовой, её собственные числа вещественны и неотрицательны, будучи квадратами сингулярных чисел матрицы  $A$  и, возможно, ещё нулями (см. предложение 3.2.4). Унитарным преобразованием подобия (ортогональным в вещественном случае) матрица  $A^*A$  может быть приведена к диагональному виду:

$A^*A = U^*\Lambda U$ , где  $U$  — унитарная  $n \times n$ -матрица,  $\Lambda$  — диагональная  $n \times n$ -матрица. При этом у  $\Lambda$  на главной диагонали находятся числа  $\sigma_i^2$ , которые являются квадратами сингулярных чисел  $\sigma_i$  матрицы  $A$  и, возможно, ещё нулями в случае  $m < n$ .

Далее имеем

$$\begin{aligned} \|A\|_2 &= \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max_{x \neq 0} \frac{\sqrt{x^* A^* A x}}{\sqrt{x^* x}} = \max_{x \neq 0} \frac{\sqrt{x^* U^* \Lambda U x}}{\sqrt{x^* U^* U x}} = \\ &= \max_{x \neq 0} \frac{\sqrt{(Ux)^* \Lambda (Ux)}}{\sqrt{(Ux)^* Ux}} = \max_{z \neq 0} \frac{\sqrt{z^* \Lambda z}}{\sqrt{z^* z}} = \max_{z \neq 0} \sqrt{\frac{\sum_i \sigma_i^2 |z_i|^2}{\sum_i |z_i|^2}} \leq \\ &\leq \max_{z \neq 0} \left( \sigma_{\max}(A) \sqrt{\frac{\sum_i |z_i|^2}{\sum_i |z_i|^2}} \right) = \sigma_{\max}(A), \end{aligned}$$

где в выкладках использована замена переменных  $z = Ux$ . Кроме того, полученная для  $\|A\|_2$  оценка достижима: достаточно взять в качестве вектора  $z$  столбец единичной  $n \times n$ -матрицы с номером, равным месту элемента  $\sigma_{\max}^2(A)$  на диагонали в  $\Lambda$ , а в самом начале выкладок положить  $x = U^*z$ . ■

Норму матриц  $\|\cdot\|_2$ , подчинённую евклидовой векторной норме, часто называют также *спектральной нормой* матриц. Для симметричных матриц она равна наибольшему из модулей собственных чисел и совпадает с так называемым спектральным радиусом матрицы (см. § 3.3и).

Отметим, что спектральная норма матриц не является абсолютной нормой (пример 3.2.4), т. е. она зависит не только от абсолютных значений элементов матрицы. В то же время  $\|\cdot\|_1$  и  $\|\cdot\|_\infty$  — это абсолютные матричные нормы, что следует из вида их выражений.

### 3.3ж Топология на множествах матриц

Совершенно аналогично тому, как это было сделано для векторов, необходимо рассмотреть топологическую структуру на множестве матриц, и она может быть введена различными способами. Например, после введения расстояния (метрики) с помощью (3.32) множество матриц одного размера превращается в метрическое пространство, в ко-

тором можно рассматривать открытые и замкнутые множества, компактность, понятия сходимости и непрерывности различных матричных функций и операций. В частности, для введения сходимости удобен так называемый секвенциальный подход — с помощью последовательностей, уже использованный в § 3.36.

**Определение 3.3.7** Будем говорить, что последовательность матриц  $\{A^{(k)}\}_{k=0}^{\infty}$  сходится к пределу  $A^*$  относительно фиксированной нормы матриц  $\|\cdot\|$  (сходится по норме), если числовая последовательность  $\|A^{(k)} - A^*\|$  сходится к нулю. При этом пишут

$$\lim_{n \rightarrow \infty} A^{(k)} = A^* \quad \text{или} \quad A^{(k)} \rightarrow A^*.$$

Матричные нормы назовём *топологически эквивалентными* (или просто *эквивалентными*), если предельный переход в одной норме влечёт существование того же предела в другой, и обратно. Эквивалентность двух матричных норм равносильна выполнению для них двустороннего неравенства, аналогичного (3.31). Наконец, в силу известного факта из математического анализа в конечномерном линейном пространстве матриц одинаковых размеров все нормы эквивалентны. Тем не менее конкретные константы эквивалентности из неравенства (3.31) играют огромную роль при выводе различных оценок, и их значения для важнейших норм даёт следующее

**Предложение 3.3.9** Для квадратных  $n \times n$ -матриц

$$\frac{1}{\sqrt{n}} \|A\|_2 \leq \|A\|_1 \leq \sqrt{n} \|A\|_2,$$

$$\frac{1}{\sqrt{n}} \|A\|_\infty \leq \|A\|_2 \leq \sqrt{n} \|A\|_\infty,$$

$$\frac{1}{n} \|A\|_1 \leq \|A\|_\infty \leq n \|A\|_1.$$

**Доказательство.** Нам потребуется несколько модифицировать определение подчинённой матричной нормы: вместо  $\|A\|' = \max_{\|y\|=1} \|Ay\|$ , как нетрудно понять, можно написать

$$\|A\|' = \max_{\substack{\|y\| \leq 1 \\ y \neq 0}} \|Ay\|,$$

формально расширив множество, по которому берётся max.

Докажем первое двустороннее неравенство. Правая оценка первого двустороннего неравенства из предложения 3.3.2 имеет следствием, во-первых, что

$$\|A\|_1 = \max_{\|y\|_1 \leq 1} \|Ay\|_1 \leq \max_{\|y\|_1 \leq 1} (\sqrt{n} \|Ay\|_2),$$

и, во-вторых, что множество векторов  $y$ , удовлетворяющих  $\|y\|_1 \leq 1$ , включается во множество векторов, определяемых условием  $\|y\|_2 \leq 1$ . По этой причине

$$\max_{\|y\|_1 \leq 1} \|Ay\|_2 \leq \max_{\|y\|_2 \leq 1} \|Ay\|_2 = \|A\|_2,$$

так что в целом действительно  $\|A\|_1 \leq \sqrt{n} \|A\|_2$ .

С другой стороны, в силу левой оценки первого неравенства из предложения 3.3.2

$$\|A\|_1 = \max_{\|y\|_1 \leq 1} \|Ay\|_1 \geq \max_{\|y\|_1 \leq 1} \|Ay\|_2. \quad (3.38)$$

Но правая оценка того же первого неравенства означает, что множество векторов  $y$ , удовлетворяющих  $\|y\|_1 \leq 1$ , не более чем в  $\sqrt{n}$  меньше множества векторов, удовлетворяющих  $\|y\|_2 \leq 1$ :

$$\sqrt{n} \cdot \{y \mid \|y\|_1 \leq 1\} \subseteq \{y \mid \|y\|_2 \leq 1\}.$$

Следствием абсолютной однородности нормы является тогда неравенство

$$\max_{\|y\|_1 \leq 1} \|Ay\|_2 \geq \frac{1}{\sqrt{n}} \max_{\|y\|_2 \leq 1} \|Ay\|_2 = \frac{1}{\sqrt{n}} \|A\|_2.$$

Сопоставляя выписанное неравенство с (3.38), получаем левую оценку первого неравенства доказываемого предложения.

Доказательства второго и третьего двусторонних неравенств аналогичны, они следуют из двусторонних неравенств для соответствующих векторных норм, которые приведены в предложении 3.3.2. ■

Как и для векторов, помимо сходимости по норме введём также

**Определение 3.3.8** Будем говорить, что последовательность матриц  $\{A^{(k)}\}_{k=1}^{\infty}$  одинаковых размеров,  $A^{(k)} = (a_{ij}^{(k)})$ , сходится поэлементно к матрице  $A^* = (a_{ij}^*)$ , если для каждой пары индексов  $i, j$  имеет

место сходимость соответствующего элемента  $a_{ij}^{(k)} \rightarrow a_{ij}^* \in \mathbb{R}$  или  $\mathbb{C}$  при  $k \rightarrow \infty$ .

Иными словами, для последовательности матриц  $\{A^{(k)}\}_{k=0}^\infty$  полагаем

$$\begin{aligned} A^{(k)} = (a_{ij}^{(k)}) \rightarrow A^* = (a_{ij}^*) \text{ поэлементно в } \mathbb{R}^{m \times n} \text{ или } \mathbb{C}^{m \times n} \\ \Updownarrow \\ a_{ij}^{(k)} \rightarrow a_{ij}^* \text{ в } \mathbb{R} \text{ или } \mathbb{C} \text{ для всех индексов } i, j. \end{aligned}$$

Из эквивалентности матричных норм следует существование для любой нормы  $\|\cdot\|$  такой константы  $C$ , что

$$\max_{i,j} |a_{ij}| \leq \max_{1 \leq j \leq n} \left( \sum_{i=1}^m |a_{ij}| \right) = \|A\|_1 \leq C\|A\|$$

(вместо 1-нормы в этой выкладке можно взять, к примеру,  $\infty$ -норму). Поэтому для любых индексов  $i$  и  $j$  верна оценка  $|a_{ij}| \leq C\|A\|$ , т. е. сходимость последовательности матриц в любой норме влечёт поэлементную сходимость этой последовательности. Доказательство обратной импликации, т. е. того факта, что поэлементная сходимость матриц приводит к сходимости по норме, совершенно аналогично первой части предложения 3.3.3 (см. § 3.3б).

В целом для любой матричной нормы множество матриц с введённым на нём посредством (3.32) расстоянием является полным метрическим пространством, т. е. любая фундаментальная («сходящаяся в себе») последовательность имеет в нём предел. Это вытекает из предшествующего рассуждения и из полноты вещественной оси  $\mathbb{R}$  и комплексной плоскости  $\mathbb{C}$ .

**Предложение 3.3.10** *Операция умножения матриц является непрерывным отображением.*

Доказательство аналогично тому, что было проделано в § 3.3б.

В заключение темы отметим, что в вычислительной линейной алгебре нормы векторов и матриц широко используются с середины XX века. Пionерский вклад в развитие соответствующей математической техники внесли работы Дж. фон Неймана и Г. Голдстайна [134] и монография В.Н. Фаддеевой [101], которая предшествовала капитальной книге [48] и вошла в неё составной частью.

### 3.33 Энергетическая норма

Ещё одной важной и популярной конструкцией нормы является так называемая энергетическая норма векторов, которая порождается каким-либо симметричной положительно определённой матрицей.<sup>8</sup> Если  $A$  — такая матрица, то выражение  $\langle Ax, y \rangle$ , как нетрудно проверить, есть симметричная билинейная положительно определённая форма, т. е. скалярное произведение векторов  $x$  и  $y$ . Обычно обозначают его  $\langle x, y \rangle_A$ , т. е.

$$\langle x, y \rangle_A := \langle Ax, y \rangle.$$

Следовательно, относительно этого нового скалярного произведения можно рассматривать ортогональность, норму векторов и т. п. В частности, определим норму стандартным образом

$$\|x\|_A := \sqrt{\langle x, x \rangle_A} = \sqrt{\langle Ax, x \rangle}, \quad (3.39)$$

т. е. как квадратный корень из произведения  $x$  на себя в этом скалярном произведении.

**Определение 3.3.9** Для симметричной и положительно определённой матрицы  $A$  векторы  $x$  и  $y$ , удовлетворяющие условию

$$\langle x, y \rangle_A = \langle Ax, y \rangle = 0,$$

будем называть ортогональными относительно скалярного произведения, задаваемого матрицей  $A$ , или же просто  $A$ -ортогональными.

**Определение 3.3.10** Для симметричной положительно определённой матрицы  $A$  векторная норма  $\|\cdot\|_A$ , задаваемая посредством (3.39), называется энергетической нормой относительно матрицы  $A$  или же просто  $A$ -нормой.

Энергетическую норму  $\|\cdot\|_A$  часто называют также  $A$ -нормой векторов, если в задаче имеется в виду какая-то конкретная симметричная положительно определённая матрица  $A$ . Термин «энергетическая» происходит из-за аналогии выражения для этой нормы с выражениями для различных видов энергии в физических системах (см. § 3.11а).

---

<sup>8</sup>Для комплексного случая обобщение очевидно, и мы не детализуем его лишь по причине экономии места.

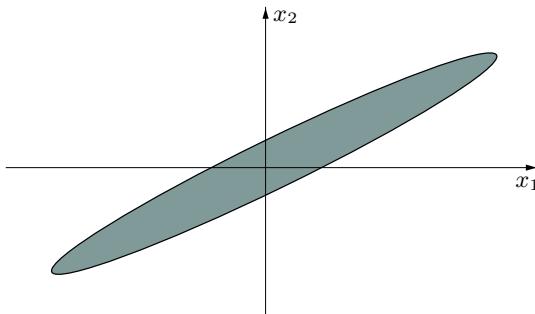


Рис. 3.9. Шар единичного радиуса в энергетической норме при значительном разбросе спектра порождающей матрицы

Так как симметричная матрица может быть приведена к диагональному виду ортогональными преобразованиями подобия, то

$$A = Q^\top D Q,$$

где  $Q$  — ортогональная матрица,  $D = \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\}$  — диагональная матрица, на главной диагонали которой стоят положительные собственные значения  $\lambda_i$  матрицы  $A$ . Поэтому

$$\begin{aligned} \|x\|_A &= \sqrt{\langle Ax, x \rangle} = \sqrt{\langle Q^\top D Q x, x \rangle} = \\ &= \sqrt{\langle D Q x, Q x \rangle} = \sqrt{\langle D y, y \rangle} = \left( \sum_{i=1}^n \lambda_i y_i^2 \right)^{1/2}, \end{aligned} \quad (3.40)$$

где  $y = Qx$ . Таким образом, в системе координат, которая получается из исходной ортогональным преобразованием  $x = Q^\top y$ , поверхности уровня энергетической нормы, задаваемые уравнениями  $\|x\|_A = \text{const}$ , являются эллипсоидами в  $\mathbb{R}^n$ . Они тем более вытянуты, чем больше различаются между собой собственные значения  $\lambda_i$  матрицы  $A$ , т. е. чем больше её число обусловленности  $\text{cond}_2(A)$  (см. § 3.4а).

Из сказанного вытекает характерная особенность энергетической нормы, которая в ряде случаев оборачивается её недостатком: возможность существенного искажения обычного геометрического масштаба объектов по разным направлениям (своебразная анизотропия). Она вызывается разбросом собственных значений порождающей матрицы

$A$  и приводит к тому, что векторы из  $\mathbb{R}^n$ , имеющие одинаковую энергетическую норму, существенно различны по обычной евклидовой длине, и наоборот (рис. 3.9). С другой стороны, использование энергетической нормы, которая порождена матрицей, фигурирующей в постановке задачи (системе линейных алгебраических уравнений, задаче на собственные значения и т. п.), часто является удобным и оправданным, а альтернативы ему очень ограничены. Мы встретимся с интенсивным использованием энергетических норм в § 3.11в, 3.11г и 3.11е.

**Пример 3.3.4** Пусть

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}. \quad (3.41)$$

Это положительно определённая матрица, с помощью которой можно задавать энергетическое скалярное произведение и соответствующую  $A$ -норму.

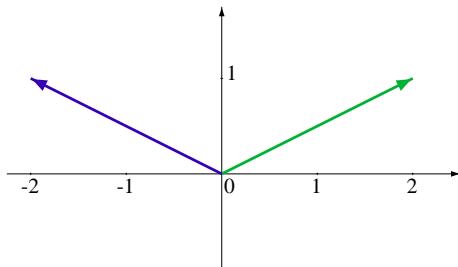


Рис. 3.10.  $A$ -ортогональные векторы относительно скалярного произведения, задаваемого матрицей (3.41)

Нетрудно проверить, что векторы

$$\begin{pmatrix} -2 \\ 1 \end{pmatrix} \quad \text{и} \quad \begin{pmatrix} 2 \\ 1 \end{pmatrix},$$

изображённые на рис. 3.10, являются  $A$ -ортогональными для рассматриваемой матрицы  $A$ , хотя реальный угол между этими векторами — почти  $127^\circ$ . ■

Из общего факта эквивалентности любых норм в конечномерном линейном пространстве следует, что энергетическая норма эквивалентна рассмотренным выше векторным нормам  $\|\cdot\|_1$ ,  $\|\cdot\|_2$ ,  $\|\cdot\|_\infty$  и  $\|\cdot\|_p$ .

Но интересно знать конкретные константы эквивалентности. Из выражения (3.40) следует, что

$$\left( \min_i \sqrt{\lambda_i} \right) \|x\|_2 \leq \|x\|_A \leq \left( \max_i \sqrt{\lambda_i} \right) \|x\|_2,$$

где  $\lambda_i$  — собственные значения порождающей матрицы  $A$ . Другие двусторонние неравенства для энергетической нормы можно получить с помощью предложения 3.3.9.

Выражения для матричных норм, которые подчинены энергетической норме векторов или просто согласованы с нею, выписываются непросто. Даже не всегда можно указать для них явный и несложный вычисляемый вид. Тем не менее мы приведём полезный и красивый результат на эту тему, который будет далее использован при исследовании метода наискорейшего спуска в § 3.11в:

**Предложение 3.3.11** *Пусть  $A$  — симметричная положительно определённая матрица, порождающая энергетическую норму  $\|\cdot\|_A$  в  $\mathbb{R}^n$ . Если  $S$  — матрица, которая является значением некоторого полинома от матрицы  $A$ , то для любого вектора  $x \in \mathbb{R}^n$  справедливо*

$$\|Sx\|_A \leq \|S\|_2 \|x\|_A. \quad (3.42)$$

Фактически в предложении 3.3.11 утверждается, что спектральная матричная норма (см. определение 3.3.5) согласована с энергетической нормой векторов. Этим можно пользоваться при проведении различных выкладок, оценок и т. п. Но это верно лишь для некоторого частного класса матриц, родственных той матрице, которая порождает энергетическую норму.

**Доказательство.** Прежде всего обоснуем вспомогательный факт, который будет использован в доказательстве предложения.

Умножение матриц в общем случае некоммутативно, но если в произведении двух матриц один из сомножителей является значением какого-то алгебраического полинома от второго сомножителя, то эти матрицы перестановочны. В самом деле, пусть  $S = \alpha_0 I + \alpha_1 A + \dots + \alpha_p A^p$ , тогда

$$\begin{aligned} AS &= A(\alpha_0 I + \alpha_1 A + \dots + \alpha_p A^p) = \\ &= \alpha_0 A + \alpha_1 A^2 + \dots + \alpha_p A^{p+1} = \\ &= (\alpha_0 I + \alpha_1 A + \dots + \alpha_p A^p)A = SA. \end{aligned}$$

Переходя к доказательству предложения, заметим, что матрица  $S$  симметрична одновременно с  $A$ . Выполним её разложение в виде  $S = Q\Sigma Q^\top$ , где  $Q$  — ортогональная матрица, а  $\Sigma = \text{diag}\{s_1, s_2, \dots, s_n\}$  — диагональная матрица, имеющая по диагонали собственные числа  $S$ . Их модули являются сингулярными числами  $\sigma_i(S)$  матрицы  $S$ . Поэтому

$$\begin{aligned} \|Sx\|_A^2 &= \langle ASx, Sx \rangle = \langle SAx, Sx \rangle = \\ &= \langle Q\Sigma Q^\top Ax, Q\Sigma Q^\top x \rangle = \langle \Sigma Q^\top Ax, \Sigma Q^\top x \rangle \leq \\ &\leq \left( \max_i s_i^2 \right) \langle Q^\top Ax, Q^\top x \rangle = \left( \max_i |s_i|^2 \right) \langle QQ^\top Ax, x \rangle = \\ &= \left( \max_i (\sigma_i(S))^2 \right) \langle Ax, x \rangle = \|S\|_2^2 \|x\|_A^2 \end{aligned}$$

с учётом того, что  $\max_i (\sigma_i(S))^2 = \|S\|_2^2$ . ■

### 3.3и Спектральный радиус

**Определение 3.3.11** Спектральным радиусом квадратной матрицы называется наибольший из модулей её собственных чисел.

Эквивалентное определение: спектральным радиусом матрицы называется наименьший из радиусов кругов комплексной плоскости  $\mathbb{C}$  с центрами в нуле, которые содержат весь спектр матрицы. Эта трактовка хорошо объясняет и сам термин. Обычно спектральный радиус матрицы  $A$  обозначают  $\rho(A)$ .

Спектральный радиус матрицы — неотрицательное число, которое в общем случае может не совпадать ни с одним из собственных значений (рис. 3.11). Но если матрица неотрицательна, т. е. все её элементы — неотрицательные вещественные числа, то наибольшее по модулю собственное значение такой матрицы тоже неотрицательно и, таким образом, равно спектральному радиусу матрицы. Кроме того, неотрицательным может быть выбран соответствующий собственный вектор. Эти утверждения составляют содержание теоремы Перрона–Фробениуса, одного из главных результатов теории неотрицательных матриц [9, 37, 54].

**Теорема 3.3.1** Спектральный радиус матрицы не превосходит любой её нормы.

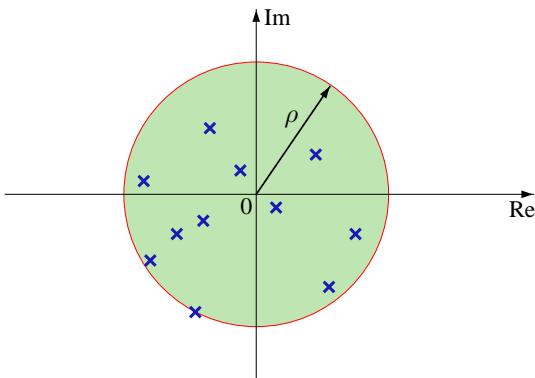


Рис. 3.11. Иллюстрация спектрального радиуса матрицы:  
крестиками обозначены точки спектра

**Доказательство.** Рассмотрим сначала случай, когда матрица является комплексной.

Пусть  $\lambda$  — собственное значение матрицы  $A$ , а  $v \neq 0$  — соответствующий собственный вектор, так что  $Av = \lambda v$ . Воспользуемся тем установленным в § 3.3д фактом (предложение 3.3.5), что любая матричная норма согласована с некоторой векторной нормой, и возьмём от обеих частей равенства  $Av = \lambda v$  норму, согласованную с рассматриваемой нормой матрицы, т. е. с  $\|A\|$ . Получим

$$\|A\| \cdot \|v\| \geq \|Av\| = \|\lambda v\| = |\lambda| \cdot \|v\|, \quad (3.43)$$

где  $\|v\| > 0$ , и потому сокращение на эту величину обеих частей неравенства (3.43) даёт  $\|A\| \geq |\lambda|$ . Коль скоро наше рассуждение справедливо для любого собственного значения  $\lambda$ , то в самом деле  $\max |\lambda| = \rho(A) \leq \|A\|$ .

Рассмотрим теперь случай вещественной  $n \times n$ -матрицы  $A$ . Если  $\lambda$  — её вещественное собственное значение, то проведённые выше рассуждения остаются полностью справедливыми. Если же  $\lambda$  — комплексное собственное значение матрицы  $A$ , то комплексным является и соответствующий собственный вектор  $v$ . Тогда цепочку соотношений (3.43) выписать нельзя, поскольку согласованная векторная норма определена лишь для вещественных векторов из  $\mathbb{R}^n$ .

Выполним *комплексификацию* рассматриваемого линейного пространства, т. е. вложим его в более широкое линейное векторное про-

пространство над полем комплексных чисел. В формальных терминах мы переходим от  $\mathbb{R}^n$  к пространству  $\mathbb{R}^n \oplus i\mathbb{R}^n$ , где  $i$  — мнимая единица (т. е. скаляр, обладающий свойством  $i^2 = -1$ ),  $i\mathbb{R}^n$  — это множество всех произведений  $iy$  для  $y \in \mathbb{R}^n$ , а « $\oplus$ » означает прямую сумму линейных пространств [10, 24, 37, 103].

Элементами линейного пространства  $\mathbb{R}^n \oplus i\mathbb{R}^n$  служат упорядоченные пары  $(x, y)^\top$ , где  $x, y \in \mathbb{R}^n$ . Сложение и умножение на скаляр  $(\alpha + i\beta) \in \mathbb{C}$  определяются для них следующим образом

$$(x, y)^\top + (x', y')^\top = (x + x', y + y')^\top, \quad (3.44)$$

$$(\alpha + i\beta) \cdot (x, y)^\top = (\alpha x - \beta y, \alpha y + \beta x)^\top. \quad (3.45)$$

Введённые пары векторов  $(x, y)^\top$  обычно записывают в виде  $x + iy$ , причём  $x$  и  $y$  называются соответственно вещественной и мнимой частями вектора из  $\mathbb{R}^n \oplus i\mathbb{R}^n$ . Линейный оператор, действующий на  $\mathbb{R}^n \oplus i\mathbb{R}^n$  и продолжающий линейное преобразование  $\mathbb{R}^n$  с матрицей  $A$ , сам может быть представлен в матричном виде как

$$\mathcal{A} = \begin{pmatrix} A & 0 \\ 0 & A \end{pmatrix}. \quad (3.46)$$

Его блочно-диагональный вид объясняется тем, что согласно формуле (3.45) для любого  $\alpha \in \mathbb{R}$

$$\alpha \cdot (x, y)^\top = (\alpha x, \alpha y)^\top,$$

и потому вещественная матрица  $A$  независимо действует на вещественную и мнимую части векторов из построенного комплексного пространства  $\mathbb{R}^n \oplus i\mathbb{R}^n$ . Для доказательства важно, что матрица  $\mathcal{A}$  имеет тот же спектр, что  $A$ .

Без какого-либо ограничения общности можно считать, что рассматриваемая нами норма матрицы, т. е.  $\|A\|$ , является подчинённой (операторной) нормой, поскольку такие нормы являются наименьшими из всех согласованных матричных норм (см. § 3.3e). Если предложение будет обосновано для подчинённых матричных норм, то оно тем более будет верным для всех прочих норм матриц.

Пусть  $\|\cdot\|$  — векторная норма в  $\mathbb{R}^n$ , которой подчинена наша матричная норма. Зададим в  $\mathbb{R}^n \oplus i\mathbb{R}^n$  норму векторов как  $\|(x, y)^\top\| = \|x\| + \|y\|$ . Тогда ввиду (3.46) и с помощью рассуждений, аналогичных доказательству предложения 3.3.7, нетрудно показать, что подчинённая матричная норма для  $\mathcal{A}$  во множестве  $2n \times 2n$ -матриц есть

$\|\mathcal{A}\| = \max\{\|A\|, \|A^\top\|\} = \|A\|$ . Кроме того, теперь для  $\mathcal{A}$  справедливы выводы о связи нормы и спектрального радиуса, полученные в начале доказательства для случая комплексной матрицы, т. е.

$$\rho(A) = \rho(\mathcal{A}) \leq \|\mathcal{A}\| = \|A\|.$$

Это и требовалось доказать. ■

Для симметричных и эрмитовых матриц спектральный радиус есть норма, которая совпадает со спектральной матричной нормой  $\|\cdot\|_2$ . Это следует из предложения 3.3.8 и того факта, что для симметричных и эрмитовых матриц сингулярные числа равны абсолютным значениям собственных чисел. Но для матриц общего вида спектральный радиус матричной нормой не является. Хотя для любого скаляра  $\alpha$  справедливо

$$\rho(\alpha A) = |\alpha| \rho(A),$$

т. е. спектральный радиус обладает абсолютной однородностью, аксиома неотрицательности матричной нормы (МН1) и неравенство треугольника (МН3) для него не выполняются.

Во-первых, для ненулевой матрицы

$$\begin{pmatrix} 0 & 1 & & 0 \\ & 0 & 1 & \\ & & \ddots & \ddots \\ 0 & & & 0 & 1 \\ & & & & 0 \end{pmatrix} \quad (3.47)$$

— жордановой клетки, отвечающей собственному значению 0, спектральный радиус равен нулю. Во-вторых, если  $A$  — матрица вида (3.47), то  $\rho(A^\top) = \rho(A) = 0$ , но  $\rho(A + A^\top) > 0$ . Последнее вытекает из того, что симметричная матрица  $A + A^\top$  — ненулевая, поэтому  $\|A + A^\top\|_2 > 0$ , и, как следствие, наибольший из модулей её собственных чисел строго больше нуля. Получается, что неверно «неравенство треугольника»

$$\rho(A + A^\top) \leq \rho(A) + \rho(A^\top).$$

Тем не менее спектральный радиус является важной характеристикой матрицы, которая описывает асимптотическое поведение её степеней.

Как известно, для вещественного или комплексного числа  $q$  поведение степеней  $q^n$  при неограниченном возрастании  $n$  полностью определяется абсолютным значением  $|q|$ :

- если  $|q| < 1$ , то  $q^n \rightarrow 0$  при  $n \rightarrow \infty$ ,
- если  $|q| > 1$ , то  $q^n \rightarrow \infty$  при  $n \rightarrow \infty$ ,
- если  $|q| = 1$ , то  $q^n$  ограничено при  $n \rightarrow \infty$ .

Для квадратной матрицы наиболее адекватной характеристикой, описывающей асимптотику её степеней, оказывается не норма — непосредственное обобщение абсолютного значения, а спектральный радиус.

**Предложение 3.3.12** *Пусть  $A$  — квадратная матрица, вещественная или комплексная. Если последовательность  $\{A^k\}_{k=0}^\infty$  из степеней матрицы ограничена, то  $\rho(A) \leq 1$ , т. е. спектральный радиус матрицы  $A$  не превосходит 1. Если  $\lim_{k \rightarrow \infty} A^k = 0$  — степени матрицы  $A$  сходятся к нулевой матрице, то  $\rho(A) < 1$ , т. е. спектральный радиус матрицы  $A$  меньше 1.*

**Доказательство.** Пусть  $\lambda$  — собственное число матрицы  $A$  (возможно, комплексное), а  $v \neq 0$  — соответствующий ему собственный вектор (который тоже может быть комплексным). Тогда  $Av = \lambda v$ , и потому

$$\begin{aligned} A^2v &= A(Av) = A(\lambda v) = \lambda(Av) = \lambda^2v, \\ A^3v &= A(A^2v) = A(\lambda^2v) = \lambda^2(Av) = \lambda^3v, \\ &\dots \quad \dots \quad , \end{aligned}$$

так что в целом

$$(A^k)v = (\lambda^k)v. \quad (3.48)$$

Если последовательность степеней  $A^k$ ,  $k = 0, 1, 2, \dots$ , ограничена, то при фиксированном векторе  $v$  ограничена также левая часть выписанного равенства (3.48). Поэтому ограничена и правая часть в (3.48), причём  $v \neq 0$ . Это возможно лишь в случае  $|\lambda| \leq 1$ .

Если последовательность степеней  $A^k$ ,  $k = 0, 1, 2, \dots$ , сходится к нулевой матрице, то при фиксированном векторе  $v$  нулевой предел имеет вся левая часть равенства (3.48). Поэтому к нулевому вектору должна сходиться и правая часть в (3.48), причём  $v \neq 0$ . Это возможно лишь в случае  $|\lambda| < 1$ . ■

Ниже в § 3.10б мы увидим, что условие  $\rho(A) < 1$  является также и достаточным для сходимости к нулю степеней матрицы  $A$ .

Рассуждения, с помощью которых проведено доказательство предложения 3.3.12, можно продолжить, получив дальнейшие тонкие свойства спектрального радиуса. Возьмём от обеих частей равенства (3.48) какую-нибудь векторную норму:

$$\|A^k v\| = \|\lambda^k v\|.$$

Поэтому  $\|A^k\| \|v\| \geq |\lambda^k| \|v\|$  для согласованной матричной нормы  $\|A\|$ , так что после сокращения на  $\|v\| \neq 0$  получаем

$$\|A^k\| \geq |\lambda|^k \quad \text{для всех } k = 0, 1, 2, \dots$$

По этой причине для любого собственного значения матрицы имеет место оценка

$$|\lambda| \leq \inf_{k \in \mathbb{N}} \|A^k\|^{1/k},$$

или, иными словами,

$$\rho(A) \leq \inf_{k \in \mathbb{N}} \|A^k\|^{1/k}. \quad (3.49)$$

Так как всякая матричная норма всегда согласована с какой-то векторной, то выведенное неравенство справедливо для любой матричной нормы. Оно является обобщением теоремы 3.3.1, переходя в него при  $k = 1$ .

Уточнением неравенства (3.49) является *формула Гельфанд*

$$\rho(A) = \lim_{k \rightarrow \infty} \|A^k\|^{1/k},$$

которая верна для любой из матричных норм. Её доказательство можно найти в книге [54]. В целом, предложение 3.3.12, неравенство (3.49) и формула Гельфанд показывают, что хотя матрица и является сложным составным объектом, нормы её степеней ведут себя примерно так же, как геометрическая прогрессия со знаменателем, равным спектральному радиусу этой матрицы. Например, для  $n \times n$ -матрицы (3.47) или любой ей подобной  $n$ -я степень зануляется, и это свойство обнаруживается спектральным радиусом.

### 3.3к Матричный ряд Неймана

Как известно из математического анализа, операцию суммирования можно обобщить на случай бесконечного числа слагаемых, и такие

бесконечные суммы называются *рядами*. Суммой ряда называют предел (если он существует) для сумм конечного числа слагаемых ряда, когда это число неограниченно возрастает. Совершенно аналогичная конструкция применима также к суммированию векторов и матриц, а не только чисел. Именно, суммой матричного ряда

$$\sum_{k=0}^{\infty} A^{(k)},$$

где  $A^{(k)}$ ,  $k = 0, 1, 2, \dots$ , — матрицы одного размера, мы будем называть предел частичных сумм  $\sum_{k=0}^N A^{(k)}$  при  $N \rightarrow \infty$ . В этом определении  $A^{(k)}$  могут быть и векторами.

**Предложение 3.3.13** Пусть  $X$  — квадратная матрица и  $\|X\| < 1$  в некоторой матричной норме. Тогда матрица  $(I - X)$  неособенна, для обратной матрицы справедливо представление

$$(I - X)^{-1} = \sum_{k=0}^{\infty} X^k \quad (3.50)$$

и имеет место оценка

$$\|(I - X)^{-1}\| \leq \frac{1}{1 - \|X\|}. \quad (3.51)$$

Аналог геометрической прогрессии для матриц, стоящий в правой части равенства (3.50), называется *матричным рядом Неймана*.

**Доказательство.** Покажем, что матрица  $(I - X)$  неособенна. Если это не так, то  $(I - X)v = 0$  для некоторого ненулевого вектора  $v$ . Тогда  $Xv = v$ , и, взяв от обеих частей этого равенства векторную норму, согласованную с матричной нормой, в которой  $\|X\| < 1$  по условию, получим

$$\|X\| \|v\| \geq \|Xv\| = \|v\|.$$

В случае, когда  $v \neq 0$ , можем сократить обе части неравенства на положительную величину  $\|v\|$ , что даёт  $\|X\| \geq 1$ . Следовательно, при условии  $\|X\| < 1$  и ненулевых  $v$  равенство  $(I - X)v = 0$  невозможно.

Обозначим посредством  $S_N = \sum_{k=0}^N X^k$  частичную сумму матричного ряда Неймана. Коль скоро

$$\begin{aligned}\|S_{N+p} - S_N\| &= \left\| \sum_{k=N+1}^{N+p} X^k \right\| \leq \sum_{k=N+1}^{N+p} \|X^k\| \leq \sum_{k=N+1}^{N+p} \|X\|^k = \\ &= \|X\|^{N+1} \cdot \frac{1 - \|X\|^p}{1 - \|X\|} \rightarrow 0\end{aligned}$$

при  $N \rightarrow \infty$  и любых целых положительных  $p$ , последовательность  $S_N$  является фундаментальной (последовательностью Коши) в полном метрическом пространстве квадратных матриц с расстоянием, которое порождено рассматриваемой нормой  $\|\cdot\|$ . Следовательно, частичные суммы  $S_N$  ряда Неймана имеют предел  $S = \lim_{N \rightarrow \infty} S_N$ , причём

$$(I - X)S_N = (I - X)(I + X + X^2 + \dots + X^N) = I - X^{N+1} \rightarrow I$$

при  $N \rightarrow \infty$ , поскольку тогда  $\|X^{N+1}\| \leq \|X\|^{N+1} \rightarrow 0$ . Так как этот предел  $S$  удовлетворяет соотношению  $(I - X)S = I$ , можем заключить, что  $S = (I - X)^{-1}$ .

Наконец,

$$\|(I - X)^{-1}\| = \left\| \sum_{k=0}^{\infty} X^k \right\| \leq \sum_{k=0}^{\infty} \|X^k\| \leq \sum_{k=0}^{\infty} \|X\|^k = \frac{1}{1 - \|X\|},$$

где для бесконечных сумм неравенство треугольника может быть обосновано предельным переходом по аналогичным неравенствам для конечных сумм. Это завершает доказательство предложения. ■

Матричный ряд Неймана является простейшим из матричных степенных рядов, т. е. сумм вида

$$\sum_{k=0}^{\infty} c_k X^k,$$

где  $X$  — квадратная матрица и  $c_k$ ,  $k = 0, 1, 2, \dots$ , — счётный набор коэффициентов. С помощью матричных степенных рядов можно определять значения аналитических функций от матрицы (например, экспоненту, логарифм, синус и косинус и т. п.), просто подставляя эту матрицу вместо аргумента в степенные разложения для соответствующих функций. Это важная и интересная тема, находящая многочисленные приложения; подробности можно увидеть, к примеру, в [9, 11, 24, 28].

## 3.4 Обусловленность систем линейных уравнений

### 3.4a Число обусловленности матриц

В этом разделе мы вводим количественную меру чувствительности решения системы линейных алгебраических уравнений по отношению к возмущениям (или изменениям) матрицы и вектора правой части. Фактически общие идеи и понятия, намеченные в § 1.7, развиваются здесь в приложении к задаче решения систем линейных уравнений.

Рассмотрим систему линейных алгебраических уравнений

$$Ax = b \quad (3.52)$$

с неособенной квадратной матрицей  $A$  и вектором правой части  $b \neq 0$ , а также систему

$$(A + \Delta A) \tilde{x} = b + \Delta b,$$

где  $\Delta A \in \mathbb{R}^{n \times n}$  и  $\Delta b \in \mathbb{R}^n$  — возмущения матрицы и вектора правой части. Насколько сильно ненулевое решение  $\tilde{x}$  возмущённой системы может отличаться от решения  $x$  исходной системы уравнений?

Пусть это отличие есть  $\Delta x = \tilde{x} - x$ , так что  $\tilde{x} = x + \Delta x$ , и потому

$$(A + \Delta A)(x + \Delta x) = b + \Delta b.$$

Вычитая из этого равенства исходную невозмущённую систему уравнений (3.52), получим

$$(\Delta A)x + (A + \Delta A)\Delta x = \Delta b, \quad (3.53)$$

т. е.

$$(\Delta A)(x + \Delta x) + A\Delta x = \Delta b.$$

Вспоминая, что  $x + \Delta x = \tilde{x}$ , можем заключить

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

Для оценки величины изменения решения  $\Delta x$  воспользуемся какой-нибудь подходящей по условиям задачи векторной нормой. Применяя её к обеим частям полученного соотношения, будем иметь

$$\|\Delta x\| \leq \|A^{-1}\| \cdot (\|\Delta A\| \|\tilde{x}\| + \|\Delta b\|)$$

при согласовании используемых векторных и матричных норм. Предполагая, что возмущённое решение  $\tilde{x}$  не равно нулю, можем поделить обе части на  $\|\tilde{x}\| > 0$ , придя к неравенству

$$\begin{aligned} \frac{\|\Delta x\|}{\|\tilde{x}\|} &\leq \|A^{-1}\| \cdot \left( \|\Delta A\| + \frac{\|\Delta b\|}{\|\tilde{x}\|} \right) = \\ &= \|A^{-1}\| \|A\| \cdot \left( \frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|A\| \cdot \|\tilde{x}\|} \right). \end{aligned} \quad (3.54)$$

Это практическая *апостериорная оценка* относительной погрешности решения, которую удобно применять после того, как приближённое решение системы уже найдено.<sup>9</sup> Ввиду того, что  $\|A\| \cdot \|\tilde{x}\| \geq \|A\tilde{x}\| \approx \|b\|$ , знаменатель второго слагаемого в скобках из правой части неравенства «приблизительно не меньше», чем  $\|b\|$ . Поэтому полученной оценке (3.54) путём некоторого огрубления можно придать более элегантный вид

$$\frac{\|\Delta x\|}{\|\tilde{x}\|} \lesssim \|A^{-1}\| \|A\| \cdot \left( \frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|} \right), \quad (3.55)$$

в котором справа задействованы относительные погрешности в матрице  $A$  и правой части  $b$ .

Фигурирующая в оценках (3.54) и (3.55) величина  $\|A^{-1}\| \|A\|$ , на которую суммарно умножаются относительные ошибки в матрице и правой части, имеет своё собственное название, так как она играет важнейшую роль в вычислительной линейной алгебре.

**Определение 3.4.1** Для квадратной неособенной матрицы  $A$  величина  $\|A^{-1}\| \|A\|$  называется её *числом обусловленности* (относительно выбранной нормы матриц).

Понятие числа обусловленности введено А. Тьюрингом в 1948 году в работе [130]. Мы будем обозначать число обусловленности матрицы  $A$  посредством  $\text{cond}(A)$ , иногда с индексом, указывающим выбор нормы.<sup>10</sup> Если же матрица  $A$  — особенная, то удобно положить  $\text{cond}(A) = +\infty$ . Это соглашение оправдывается тем, что обычно  $\|A^{-1}\|$  неограниченно возрастает при приближении матрицы  $A$  к множеству особенных матриц.

---

<sup>9</sup>От латинского словосочетания «*a posteriori*», означающего знание, полученное из опыта. Под «опытом» здесь, естественно, понимается процесс решения задачи.

<sup>10</sup>В математической литературе для числа обусловленности матрицы  $A$  иногда можно встретить обозначения  $\mu(A)$  или  $\kappa(A)$ .

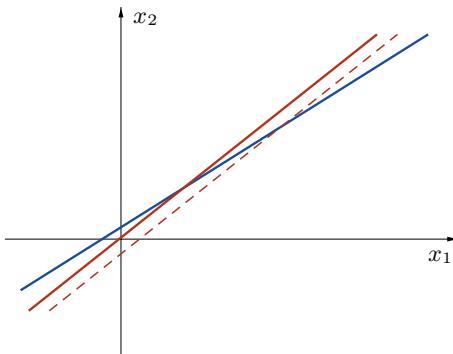


Рис. 3.12. Иллюстрация возмущения системы линейных уравнений с плохой обусловленностью матрицы: малые «шевеления» любой прямой приводят к большим изменениям в решении

Выведем теперь *aприорную* оценку относительной погрешности не-нулевого решения, которая не будет опираться на знание вычислительного решения и может быть применена *до* того, как мы начнём решать СЛАУ.<sup>11</sup>

После вычитания точного уравнения из приближённого мы получили (3.53):

$$(\Delta A)x + (A + \Delta A)\Delta x = \Delta b.$$

Отсюда

$$\begin{aligned} \Delta x &= (A + \Delta A)^{-1}(-(\Delta A)x + \Delta b) = \\ &= (A(I + A^{-1}\Delta A))^{-1}(-(\Delta A)x + \Delta b) = \\ &= (I + A^{-1}\Delta A)^{-1}A^{-1}(-(\Delta A)x + \Delta b). \end{aligned}$$

Беря интересующую нас векторную норму от обеих частей этого равенства и пользуясь далее условием согласования с матричной нормой, субмультипликативностью и неравенством треугольника, получим

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

---

<sup>11</sup>От латинского словосочетания «*a priori*», означающего в философии знание, полученное до опыта и независимо от него.

После деления обеих частей неравенства на  $\|x\| > 0$  будем иметь

$$\frac{\|\Delta x\|}{\|x\|} \leq \| (I + A^{-1}\Delta A)^{-1} \| \cdot \|A^{-1}\| \cdot \left( \|\Delta A\| + \frac{\|\Delta b\|}{\|x\|} \right).$$

Предположим, что возмущение  $\Delta A$  матрицы  $A$  не слишком велико, так что выполнено условие

$$\|\Delta A\| \leq \frac{1}{\|A^{-1}\|}.$$

Тогда

$$\|A^{-1}\Delta A\| \leq \|A^{-1}\| \|\Delta A\| < 1$$

и обратная матрица  $(I + A^{-1}\Delta A)^{-1}$  разлагается в матричный ряд Неймана (3.50). Соответственно, мы можем воспользоваться вытекающей из этого оценкой (3.51). Следовательно,

$$\begin{aligned} \frac{\|\Delta x\|}{\|x\|} &\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \|\Delta A\|} \cdot \left( \|\Delta A\| + \frac{\|\Delta b\|}{\|x\|} \right) = \\ &= \frac{\|A^{-1}\| \cdot \|A\|}{1 - \|A^{-1}\| \|\Delta A\|} \cdot \left( \frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|A\| \|x\|} \right) = \\ &\leq \frac{\text{cond}(A)}{1 - \text{cond}(A) \cdot \frac{\|\Delta A\|}{\|A\|}} \cdot \left( \frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|} \right), \end{aligned} \quad (3.56)$$

поскольку  $\|A\| \|x\| \geq \|Ax\| = \|b\|$ .

Оценка (3.56) — важная априорная оценка относительной погрешности численного решения системы линейных алгебраических уравнений через оценки относительных погрешностей её матрицы и правой части. Если величина  $\|\Delta A\|$  достаточно мала, то множитель усиления относительной ошибки в данных

$$\frac{\text{cond}(A)}{1 - \text{cond}(A) \cdot \frac{\|\Delta A\|}{\|A\|}}$$

близок к числу обусловленности матрицы  $A$ .

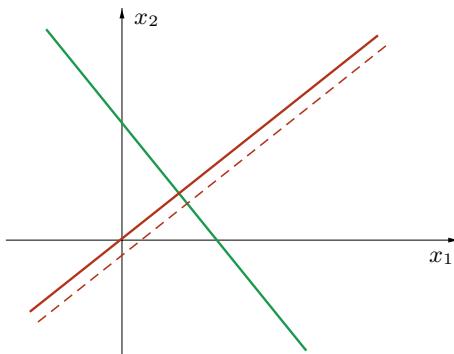


Рис. 3.13. Иллюстрация возмущения системы линейных уравнений с хорошей обусловленностью матрицы: «шевеления» прямых приводят к соизмеримым изменениям в решении

Число обусловленности матрицы и полученные с его помощью оценки имеют большое теоретическое значение, но их практическая полезность напрямую зависит от наличия эффективных способов вычисления или хотя бы приближённого оценивания числа обусловленности матриц. Фактически определение числа обусловленности требует знания некоторых характеристик обратной матрицы, и в самом общем случае решение задачи оценивания  $\text{cond}(A)$  весьма непросто. Определённым исключением являются различные специальные типы матриц, в частности матрицы с диагональным преобразованием, рассматриваемые далее в § 3.4в.

Существует также важный для практики частный случай, когда число обусловленности матрицы имеет элегантное явное выражение, на основе которого можно достаточно эффективно организовать его вычисление. Это случай спектральной матричной нормы  $\|\cdot\|_2$ , подчинённой евклидовой норме векторов.

Напомним (предложение 3.2.6), что для любой неособенной квадратной матрицы  $A$  справедливо равенство  $\sigma_{\max}(A^{-1}) = \sigma_{\min}^{-1}(A)$ , и поэтому относительно спектральной нормы число обусловленности матрицы есть

$$\text{cond}_2(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}. \quad (3.57)$$

Выражение в правой части этого равенства имеет смысл не только

для квадратных матриц, но и для общих прямоугольных, так как для них сингулярные числа тоже определены. В этом случае отношение наибольшего и наименьшего сингулярных чисел матрицы СЛАУ даёт количественную меру обусловленности линейной задачи наименьших квадратов, рассматриваемой далее в § 3.16 [11, 13]. Вообще, соотношение (3.57) помогает понять большую роль сингулярных чисел в современной вычислительной линейной алгебре и важность алгоритмов для их нахождения. В совокупности с ясным геометрическим смыслом евклидовой векторной нормы (2-нормы) эти обстоятельства вызывают преимущественное использование этих норм для многих задач теории и практики.

Если квадратная  $n \times n$ -матрица  $A$  симметрична (эрмитова), то её сингулярные числа  $\sigma_i(A)$  совпадают с модулями собственных значений  $\lambda_i(A)$ ,  $i = 1, 2, \dots, n$ , и тогда

$$\operatorname{cond}_2(A) = \frac{\max_i |\lambda_i(A)|}{\min_i |\lambda_i(A)|} \quad (3.58)$$

— спектральное число обусловленности равно отношению наибольшего и наименьшего модулей собственных значений матрицы. Для симметричных положительно определённых матриц эта формула принимает совсем простой вид

$$\operatorname{cond}_2(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}.$$

Далее в § 3.4в мы увидим, что простые оценки числа обусловленности существуют также для матриц с диагональным преобладанием.

### 3.4б Хорошо обусловленные и плохо обусловленные матрицы

Условимся называть матрицу *хорошо обусловленной*, если её число обусловленности невелико. Напротив, если число обусловленности матрицы велико, станем говорить, что матрица *плохо обусловлена*. Естественно, что эти определения имеют неформальный характер, так как зависят от нестрогих понятий «невелико» и «велико». Тем не менее они весьма полезны в практическом отношении, в частности потому, что позволяют сделать наш язык более выразительным. Эти же термины — хорошо обусловленная и плохо обусловленная — будем применять в отношении систем линейных алгебраических уравнений с соответствующими матрицами (рис. 3.12 и 3.13).

Отметим, что для любой подчинённой матричной нормы

$$\operatorname{cond}(A) = \|A^{-1}\| \|A\| \geq \|A^{-1}A\| = \|I\| = 1$$

в силу (3.36), и поэтому число обусловленности матрицы в таких нормах всегда не меньше единицы. Для произвольных матричных норм полученное неравенство тем более верно в силу того факта, что подчинённые нормы принимают наименьшие значения среди всех согласованных матричных норм.

Наименьшее возможное число обусловленности относительно 1-нормы,  $\infty$ -нормы и спектральной нормы имеют единичная матрица и кратные ей.

**Пример 3.4.1** Нетривиальным примером матриц, которые обладают наилучшей возможной обусловленностью относительно спектральной нормы, являются ортогональные матрицы (унитарные в комплексном случае).

Действительно, если  $Q$  ортогональна, то  $\|Qx\|_2 = \|x\|_2$  для любого вектора  $x$ . Следовательно,  $\|Q\|_2 = 1$ . Кроме того,  $Q^{-1} = Q^\top$  и тоже ортогональна, а потому  $\|Q^{-1}\|_2 = 1$ . Как следствие,  $\operatorname{cond}_2(Q) = 1$ .

Ясно также, что любая матрица, пропорциональная ортогональной, т. е. получающаяся из ортогональной умножением на ненулевое число, тоже имеет обусловленность 1 относительно спектральной нормы. ■

Нетрудно сообразить, что системы линейных алгебраических уравнений с ортогональными матрицами соответствуют ситуации, изображённой на рис. 3.13. Но этому рисунку соответствуют также другие системы, у которых матрицы не являются полноценно ортогональными, хотя и имеют ортогональные вектор-строки. Например, такова матрица

$$\begin{pmatrix} 1000 & 1000 \\ -1 & 1 \end{pmatrix}.$$

Её спектральное число обусловленности равно 1000, а не единице, как у матриц, пропорциональных ортогональным. Таким образом, помимо направлений вектор-строк, составленных из коэффициентов уравнений, важна соизмеримость их длин. Это же наблюдение справедливо и для общих матриц, строки которых не обязательно ортогональны: слишком большое различие длин вектор-строк матрицы может служить причиной её плохой обусловленности.

Из сказанного вытекает полезность процедуры *балансировки* строк матрицы системы (называемая также *масштабированием*). Отдельные уравнения системы предварительно домножают на специально подобранные числа так, чтобы нормы (длины) вектор-строк матрицы коэффициентов сделались примерно одинаковыми или хотя бы различались «не слишком сильно» [51]. Тогда обусловленность матрицы системы будет наименьшей при прочих равных условиях, а численное решение этой системы уравнений окажется менее подверженным погрешностям.

**Пример 3.4.2** Самым популярным содержательным примером плохо обусловленных матриц являются, пожалуй, матрицы Гильберта, которые встретились нам в § 2.11 при обсуждении среднеквадратичного приближения алгебраическими полиномами на интервале  $[0, 1]$ . Это симметричные матрицы  $H_n = (h_{ij})$ , образованные элементами

$$h_{ij} = \frac{1}{i + j - 1}, \quad i, j = 1, 2, \dots, n,$$

так что, к примеру,

$$H_3 = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{pmatrix}.$$

Число обусловленности матриц Гильберта исключительно быстро растёт в зависимости от их размера  $n$ . Воспользовавшись какими-либо стандартными процедурами для вычисления числа обусловленности матриц (встроенными, к примеру, в системы компьютерной математики Scilab, MATLAB, Octave, Maple и им подобные), нетрудно найти следующие числовые данные:

$$\operatorname{cond}_2(H_2) = 19.3,$$

$$\operatorname{cond}_2(H_3) = 524,$$

...

$$\operatorname{cond}_2(H_{10}) = 1.6 \cdot 10^{13},$$

...

Существует общая формула [67, 129, 135]:

$$\operatorname{cond}_2(H_n) = O\left(\frac{(1 + \sqrt{2})^{4n}}{\sqrt{n}}\right) \approx O(34^n / \sqrt{n}),$$

где  $O$  — «о большое», известный из математического анализа символ Э. Ландау (см. стр. 150). Интересно, что матрицы, обратные к матрицам Гильберта, могут быть вычислены явно с помощью аналитических выкладок [76, 114]. Они имеют целочисленные элементы, которые тоже очень быстро растут с размером матрицы. ■

**Пример 3.4.3** Для матрицы Вандермонда (2.8) оценка снизу для числа обусловленности [57]

$$\operatorname{cond}_2 V(x_0, x_1, \dots, x_n) \geq \sqrt{2} \frac{(1 + \sqrt{2})^{n-1}}{\sqrt{n+1}} \quad (3.59)$$

представляется существенно более скромной, хотя она всё-таки растёт экспоненциально с  $n$ . Аналогичные по смыслу, но более слабые экспоненциальные оценки снизу для числа обусловленности матрицы Вандермонда выводятся в книге [44]. Но оценка (3.59) и ей подобные, не зависящие от значений  $x_0, x_1, \dots, x_n$ , являются весьма грубыми, и реальные матрицы Вандермонда, как правило, обусловлены гораздо хуже.

Например, для матриц размера  $6 \times 6$  оценка (3.59) даёт всего лишь 19.6, но для реальной матрицы Вандермонда

$$\begin{pmatrix} 1 & 1 & 1^2 & 1^3 & 1^4 & 1^5 \\ 1 & 1.2 & 1.2^2 & 1.2^3 & 1.2^4 & 1.2^5 \\ 1 & 1.4 & 1.4^2 & 1.4^3 & 1.4^4 & 1.4^5 \\ 1 & 1.6 & 1.6^2 & 1.6^3 & 1.6^4 & 1.6^5 \\ 1 & 1.8 & 1.8^2 & 1.8^3 & 1.8^4 & 1.8^5 \\ 1 & 2 & 2^2 & 2^3 & 2^4 & 2^5 \end{pmatrix},$$

которая получается при решении задачи алгебраической интерполяции на равномерной сетке  $\{1, 1.2, 1.4, 1.6, 1.8, 2\}$  (покрывающей интервал  $[1, 2]$ ), число обусловленности относительно спектральной нормы равно  $8.9 \cdot 10^5$ .

Для  $6 \times 6$ -матрицы Вандермонда, которая возникла в примере 2.2.1 в связи с интерполяцией по набору узлов  $\{0, 1, 2, 3, 4, 5\}$ , число обусловленности равно  $5.8 \cdot 10^4$ . И так далее.

В целом матрицы Вандермонда можно неформально классифицировать как «умеренно плохообусловленные». ■

**Пример 3.4.4** Рассмотрим верхнюю треугольную  $n \times n$ -матрицу

$$U = \begin{pmatrix} 1 & -1 & -1 & \cdots & -1 \\ 0 & 1 & -1 & \cdots & -1 \\ 0 & 0 & 1 & \cdots & -1 \\ \vdots & \vdots & \ddots & \ddots & -1 \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}, \quad (3.60)$$

у которой по главной диагонали стоят единицы, а все остальные элементы выше главной диагонали равны  $-1$ . Если  $y = (y_1, y_2, \dots, y_n)^\top$ , то решение системы уравнений  $Ux = y$  нетрудно выписать явно, используя формулы алгоритма обратной подстановки (3.82):

$$\begin{aligned} x_n &= y_n, \\ x_{n-1} &= y_{n-1} + y_n, \\ x_{n-2} &= y_{n-2} + y_{n-1} + 2y_n \\ x_{n-3} &= y_{n-3} + y_{n-2} + 2y_{n-1} + 4y_n \\ &\vdots \quad \vdots \quad \ddots \quad \vdots \\ x_1 &= y_1 + y_2 + 2y_3 + \dots + 2^{n-2}y_n. \end{aligned}$$

В силу произвольности  $y$  можем заключить, что обратная матрица  $U^{-1}$ , тоже верхняя треугольная, равна

$$U^{-1} = \begin{pmatrix} 1 & 1 & 2 & 4 & \cdots & 2^{n-3} & 2^{n-2} \\ 0 & 1 & 1 & 2 & \cdots & 2^{n-4} & 2^{n-3} \\ 0 & 0 & 1 & 1 & \cdots & 2^{n-5} & 2^{n-4} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & 1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Число обусловленности  $n \times n$ -матрицы (3.60), к примеру, в подчинённой 1-норме равно

$$\|U\|_1 \|U^{-1}\|_1 = n(1 + 1 + 2 + 2^2 + \dots + 2^{n-2}) = n2^{n-1}.$$

Точно таким же является число обусловленности матрицы  $U$  относительно подчинённой чебышёвской нормы ( $\infty$ -нормы матриц). Для средних размеров матриц, возникающих в задачах математического моделирования (скажем, при  $n \approx 100$ ), значение  $n^{2n-1}$  уже весьма велико, и соответствующие матрицы плохо обусловлены. ■

Последний пример замечателен своей обыденностью и умеренными значениями элементов матрицы  $U$ , за которыми тем не менее скрывается плохая обусловленность. Кроме того, сама матрица — треугольная, и системы линейных уравнений с такими матрицами часто возникают в виде промежуточных результатов многих алгоритмов линейной алгебры (см. § 3.6в, 3.6и, 3.7е, 3.7г, 3.10е и др.). Дальнейшие примеры конкретных матриц с их числами обусловленности читатель может найти в справочнике [114].

### 3.4в Матрицы с диагональным преобладанием

В приложениях линейной алгебры и теории матриц часто возникают матрицы, в которых диагональные элементы в том или ином смысле доминируют (преобладают) над остальной, недиагональной частью матрицы. Это обстоятельство может быть, к примеру, следствием особенностей рассматриваемой математической модели, в которой связи составляющих её частей с самими собой (они и выражаются диагональными элементами) сильнее, чем с остальными. Такие матрицы обладают рядом замечательных свойств, и изложению некоторых из них посвящён этот раздел.

Следует отметить, что сам смысл, вкладываемый в понятие «диагонального преобладания», может быть различен, и ниже мы рассмотрим простейший и наиболее популярный.

**Определение 3.4.2** Квадратную  $n \times n$ -матрицу  $A = (a_{ij})$  называют матрицей с диагональным преобладанием, если для любого  $i = 1, 2, \dots, n$  имеет место

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|. \quad (3.61)$$

Матрицы, удовлетворяющие этому определению, некоторые авторы называют матрицами со «строгим диагональным преобладанием». Со

своей стороны, мы будем говорить, что  $n \times n$ -матрица  $A = (a_{ij})$  имеет *нестрогое диагональное преобладание* в случае выполнения неравенств

$$|a_{ii}| \geq \sum_{j \neq i} |a_{ij}| \quad (3.62)$$

для любого  $i = 1, 2, \dots, n$ . Иногда в связи с условиями (3.61) и (3.62) необходимо уточнять, что речь идёт о диагональном преобладании «по строкам», поскольку имеет также смысл диагональное преобладание «по столбцам», которое определяется совершенно аналогичным образом с суммированием внедиагональных элементов по столбцам.

**Теорема 3.4.1** (признак неособенности Адамара)

*Квадратная матрица с диагональным преобладанием неособенна.*

**Доказательство.** Предположим вопреки утверждению теоремы, что рассматриваемая матрица  $A = (a_{ij})$  — особая. Тогда её столбцы линейно зависимы и для некоторого ненулевого вектора  $y = (y_1, y_2, \dots, y_n)^\top$  выполняется равенство  $Ay = 0$ , т. е.

$$\sum_{j=1}^n a_{ij} y_j = 0, \quad i = 1, 2, \dots, n. \quad (3.63)$$

Выберем среди компонент вектора  $y$  ту, которая имеет наибольшее абсолютное значение. Пусть её номер  $\nu$ , так что  $|y_\nu| = \max_{1 \leq j \leq n} |y_j|$ , причём  $|y_\nu| > 0$  в силу сделанного выше предположения о том, что  $y \neq 0$ . Следствием  $\nu$ -го из равенств (3.63) является соотношение

$$-a_{\nu\nu} y_\nu = \sum_{j \neq \nu} a_{\nu j} y_j,$$

которое влечёт цепочку оценок

$$\begin{aligned} |a_{\nu\nu}| |y_\nu| &= \left| \sum_{j \neq \nu} a_{\nu j} y_j \right| \leq \sum_{j \neq \nu} |a_{\nu j}| |y_j| \leq \\ &\leq \left( \max_{1 \leq j \leq n} |y_j| \right) \sum_{j \neq \nu} |a_{\nu j}| = |y_\nu| \sum_{j \neq \nu} |a_{\nu j}|. \end{aligned}$$

Сокращая теперь обе части полученного неравенства на  $|y_\nu| > 0$ , будем иметь

$$|a_{\nu\nu}| \leq \sum_{j \neq \nu} |a_{\nu j}|,$$

что противоречит неравенствам (3.61), т. е. диагональному преобладанию в матрице  $A$ . Итак,  $A$  действительно должна быть неособой матрицей. ■

Доказанный результат часто именуют «теоремой Леви–Деспланка» [44, 54]), но мы придерживаемся здесь терминологии, принятой в [9, 97]. В книге М. Пароди [97] можно прочитать, в частности, некоторые сведения об истории вопроса.

**Следствие.** Матрица с диагональным преобладанием является строго регулярной (см. определение 3.6.2, стр. 461). В самом деле, если исходная матрица имеет диагональное преобладание, то его имеют также все ведущие подматрицы.

Внимательное изучение доказательства признака Адамара показывает, что в нём нигде не использовался факт принадлежности элементов матрицы и векторов какому-то конкретному числовому полю —  $\mathbb{R}$  или  $\mathbb{C}$ . Таким образом, признак Адамара справедлив и для комплексных матриц. Кроме того, он может быть отчасти обобщён на матрицы, удовлетворяющие нестрогому диагональному преобладанию (3.62).

Вещественная или комплексная  $n \times n$ -матрица  $A = (a_{ij})$  называется *разложимой*, если существует разбиение множества  $\{1, 2, \dots, n\}$  первых  $n$  натуральных чисел на два непересекающихся подмножества  $I$  и  $J$ , таких что  $a_{ij} = 0$  при  $i \in I$  и  $j \in J$ . Эквивалентное определение: матрица  $A \in \mathbb{R}^{n \times n}$  разложима, если путём перестановок строк и столбцов она может быть приведена к блочно-треугольному виду

$$\begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}$$

с квадратными блоками  $A_{11}$  и  $A_{22}$ . Матрицы, не являющиеся разложимыми, называются *неразложимыми*. Важнейший пример неразложимых матриц — это матрицы, все элементы которых не равны нулю, в частности, положительны.

Обобщением признака Адамара является

**Теорема 3.4.2** (теорема Таусски) *Если для квадратной неразложимой матрицы  $A$  выполнены условия нестрогого диагонального преобразования (3.62), причём хотя бы одно из этих неравенств выполнено строго, то матрица  $A$  неособенна.*

Доказательство можно найти, к примеру, в [9].

Ещё одним полезным свойством матриц с диагональным преобразованием является возможность оценки нормы обратной матрицы.

**Теорема 3.4.3** (теорема Алберга–Нильсона [62])<sup>12</sup> *Пусть  $A = (a_{ij})$  —  $n \times n$ -матрица с диагональным преобразованием и*

$$\alpha := \min_{1 \leq i \leq n} \left\{ |a_{ii}| - \sum_{j \neq i} |a_{ij}| \right\}. \quad (3.64)$$

Тогда  $\|A^{-1}\|_\infty \leq \alpha^{-1}$ .

**Доказательство.** Прежде всего заметим, что

$$\|A^{-1}\|_\infty = \max_{x \neq 0} \frac{\|A^{-1}x\|_\infty}{\|x\|_\infty} = \max_{y \neq 0} \frac{\|y\|_\infty}{\|Ay\|_\infty} = \left( \min_{y \neq 0} \frac{\|Ay\|_\infty}{\|y\|_\infty} \right)^{-1},$$

где использована замена  $y = A^{-1}x$ . Поэтому для доказательства теоремы достаточно установить, что  $\|Ay\|_\infty/\|y\|_\infty \geq \alpha$  для любого ненулевого вектора  $y$ . Это неравенство, в свою очередь, равносильно

$$\alpha \|y\|_\infty \leq \|Ay\|_\infty. \quad (3.65)$$

Пусть компонента вектора  $y$  с наибольшим абсолютным значением имеет номер  $k$ , так что  $|y_k| = \max_{1 \leq i \leq n} |y_i| = \|y\|_\infty > 0$  в силу  $y \neq 0$ . По условию теоремы

$$0 < \alpha \leq |a_{kk}| - \sum_{j \neq k} |a_{kj}|,$$

и мы можем умножить это неравенство почленно на  $|y_k| > 0$ :

$$0 < \alpha |y_k| \leq |a_{kk}| |y_k| - \sum_{j \neq k} |a_{kj}| |y_k|.$$

---

<sup>12</sup>В англоязычной литературе этот результат нередко называют «теоремой Верэ» по имени автора переоткрывшей его работы [131] (см. также [132]).

Очевидно, что полученное неравенство только усилится, если заменить в сумме из правой части множители  $|y_k|$  на меньшие или равные им  $|y_j|$ ,  $|y_j| \leq |y_k|$ :

$$0 < \alpha |y_k| \leq |a_{kk}| |y_k| - \sum_{j \neq k} |a_{kj}| |y_j|.$$

Далее

$$\begin{aligned} 0 < \alpha |y_k| &\leq |a_{kk}y_k| - \sum_{j \neq k} |a_{kj}y_j| \leq \\ &\leq \left| \sum_{j=1}^n a_{kj}y_j \right| \leq \max_{1 \leq k \leq n} \left| \sum_{j=1}^n a_{kj}y_j \right| = \|Ay\|_\infty. \end{aligned}$$

Вспоминая, что  $|y_k| = \|y\|_\infty$ , можем заключить, что в самом деле выполняется неравенство (3.65). ■

Фактически в теореме Алберга–Нильсона вводится количественная мера диагонального преобладания — величина  $\alpha$ , задаваемая (3.64). Далее  $\infty$ -норма обратной матрицы просто оценивается через эту меру. Как следствие, для матриц с диагональным преобладанием справедлива оценка

$$\text{cond}_\infty(A) \leq \alpha^{-1} \|A\|_\infty.$$

Полезные обобщения теоремы Алберга–Нильсона можно найти в работах [131, 132], где, в частности, даются оценки минимального сингулярного числа матрицы с диагональным преобладанием.

**Пример 3.4.5** Необходимость решения системы линейных уравнений с матрицей, имеющей диагональное преобладание, возникает при построении интерполяционного кубического сплайна (см. § 2.6б). Оценим число обусловленности этой матрицы относительно подчинённой чебышёвской нормы с помощью теоремы Алберга–Нильсона и следующих из неё результатов.

Напомним, что матрица системы имеет вид

$$\frac{1}{6} \begin{pmatrix} 2(h_1 + h_2) & h_2 & & & 0 \\ h_2 & 2(h_2 + h_3) & h_3 & & \\ & h_3 & 2(h_3 + h_4) & & \ddots \\ & & & \ddots & \ddots \\ 0 & & & & h_{n-1} + 2h_n \end{pmatrix}$$

с  $h_i > 0$ , и поэтому её подчинённая чебышёвская норма равна

$$\frac{1}{6} \max \left\{ 2h_1 + 3h_2, 3 \max_{2 \leq i \leq n-2} (h_i + h_{i+1}), 3h_{n-1} + 2h_n \right\}. \quad (3.66)$$

Мера диагонального преобладания этой матрицы в смысле теоремы Алберга–Нильсона

$$\frac{1}{6} \min \left\{ 2h_1 + h_2, \min_{2 \leq i \leq n-2} (h_i + h_{i+1}), h_{n-1} + 2h_n \right\}, \quad (3.67)$$

и отношение величин (3.66) и (3.67) даёт оценку числа обусловленности матрицы.

В простейшем и наиболее важном случае равномерной сетки, когда  $h_i = h = \text{const}$ , выписанные выражения решительно упрощаются, так что вместо (3.66) имеем  $h$ , а вместо (3.67) —  $\frac{1}{3}h$ , и искомая обусловленность равна всего 3. Столь же невелико число обусловленности рассматриваемой матрицы для сеток, которые не сильно отличаются от равномерных. ■

### 3.4г Практическое применение числа обусловленности матриц

Оценки (3.54) и (3.56) на возмущения решений систем линейных алгебраических уравнений являются неулучшаемыми на всём множестве матриц, векторов правых частей и их возмущений. Более точно, для системы с заданной матрицей эти оценки достигаются на каких-то векторах правой части и возмущениях матрицы и правой части. Но «плохая обусловленность» матрицы не всегда означает высокую чувствительность решения *конкретной* системы по отношению к тем или иным *конкретным* возмущениям. Если, к примеру, правая часть имеет нулевые компоненты в направлении сингулярных векторов, отвечающих наименьшим сингулярным числам матрицы системы, то решение

СЛАУ зависит от возмущений этой правой части гораздо слабее, чем показывает оценка (3.56) для спектральной нормы (см. рассуждения в § 3.5б). И определение того, какова конкретно правая часть по отношению к матрице СЛАУ — плохая или не очень, не менее трудно, чем само решение данной системы линейных уравнений.

Из сказанного должна вытекать известная осторожность и осмотрительность по отношению к выводам, которые делаются о практической разрешимости и достоверности решений какой-либо системы линейных уравнений лишь на основании того, велико или мало число обусловленности их матрицы. Тривиальный пример: решение СЛАУ с диагональными матрицами почти никаких проблем не вызывает, но число обусловленности диагональной матрицы может быть при этом сколь угодно большим!

Наконец, оценка погрешности решений через число обусловленности выводилась при условии малости ошибок в элементах СЛАУ. По этой причине число обусловленности малопригодно для оценки разброса решения СЛАУ при значительных и больших изменениях элементов матрицы и правой части (начиная с нескольких процентов от исходного значения). Получаемые при этом с помощью оценок (3.54) и (3.56) результаты типично завышены во много раз (иногда на порядки), и для решения такой задачи более предпочтительны методы интервального анализа [105, 122].

**Пример 3.4.6** Рассмотрим  $2 \times 2$ -систему линейных уравнений

$$\begin{pmatrix} 3 & -1 \\ 0 & 3 \end{pmatrix} x = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

в которой элементы матрицы и правой части заданы неточно, с абсолютной погрешностью 1, так что в действительности можно было бы записать эту систему в неформальном виде как

$$\begin{pmatrix} 3 \pm 1 & -1 \pm 1 \\ 0 \pm 1 & 3 \pm 1 \end{pmatrix} x = \begin{pmatrix} 0 \pm 1 \\ 1 \pm 1 \end{pmatrix}.$$

Фактически мы имеем совокупность эквивалентных по точности систем линейных уравнений

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} x = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix},$$

у которых элементы матрицы и правой части могут принимать значения из интервалов

$$\begin{aligned} a_{11} &\in [2, 4], & a_{12} &\in [-2, 0], & b_1 &\in [-1, 1], \\ a_{12} &\in [-1, 1], & a_{22} &\in [2, 4], & b_2 &\in [0, 2]. \end{aligned}$$

При этом обычно говорят, что задана *интервальная система линейных алгебраических уравнений* [105, 122], в данном случае

$$\begin{pmatrix} [2, 4] & [-2, 0] \\ [-1, 1] & [2, 4] \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} [-1, 1] \\ [0, 2] \end{pmatrix}. \quad (3.68)$$

Её *множеством решений* называют множество, образованное всевозможными решениями систем линейных алгебраических уравнений того же вида, у которых элементы матрицы и компоненты правой части принадлежат заданным интервалам. Множество решений рассматриваемой нами системы (3.68) изображено на рис. 3.14.<sup>13</sup> Мы более подробно рассматриваем интервальные линейные системы уравнений в § 4.6.

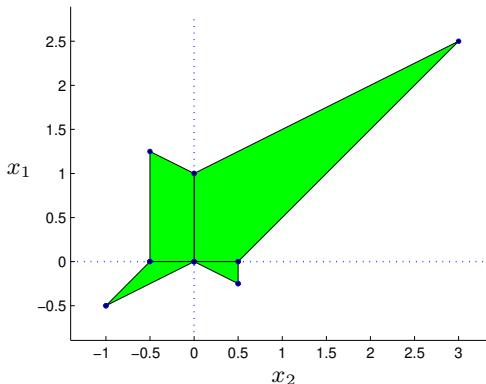


Рис. 3.14. Множество решений интервальной линейной системы (3.68)

Подсчитаем оценки возмущений, которые получаются для решения системы (3.68) на основе числа обусловленности. Можно рассматривать (3.68) как систему, получающуюся путём возмущения «средней

---

<sup>13</sup>Он построен с помощью свободного пакета программ IntLinIncR2 [104].

системы»

$$\begin{pmatrix} 3 & -1 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

в которой возмущением матрицы является

$$\Delta A = \begin{pmatrix} \Delta a_{11} & \Delta a_{12} \\ \Delta a_{21} & \Delta a_{22} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad \|\Delta A\|_\infty \leq 2,$$

а возмущением правой части —

$$\Delta b = \begin{pmatrix} \Delta b_1 \\ \Delta b_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \|\Delta b\|_\infty \leq 1.$$

Чебышёвская векторная норма ( $\infty$ -норма) используется здесь для оценки  $\Delta b$  потому, что она наиболее адекватно (без искажения формы) описывает возмущение правой части  $b$ . Соответствующая  $\infty$ -норма для матрицы  $\Delta A$ , подчинённая векторной  $\infty$ -норме, также наиболее уместна в этой ситуации, поскольку обеспечивает наиболее аккуратное согласование вычисляемых оценок (хотя и искажая немного форму множества возмущений).

Обусловленность средней матрицы относительно  $\infty$ -нормы равна 1.778,  $\infty$ -норма средней матрицы равна 4, а  $\infty$ -норма средней правой части — это 1. Следовательно, по формуле (3.56) получаем

$$\frac{\|\Delta x\|}{\|x\|} \lesssim 24.$$

Поскольку решение средней системы есть  $\tilde{x} = \left(\frac{1}{3}, \frac{1}{9}\right)^\top$  и оно имеет  $\infty$ -норму  $\frac{1}{3}$ , оценкой разброса решений рассматриваемой системы уравнений является  $\tilde{x} \pm \Delta x$ , где  $\|\Delta x\|_\infty \leq 8$ , т. е. двумерный брус<sup>14</sup>

$$\begin{pmatrix} [-7.667, 8.333] \\ [-7.889, 8.111] \end{pmatrix}.$$

По размерам он в более чем в 4 (четыре) раза превосходит оптимальные (точные) покоординатные оценки множества решений, которые удобно

---

<sup>14</sup>Читатель может проверить числовые данные этого примера в любой системе компьютерной математики: Scilab, MATLAB, Octave, Maple и т. п.

описать интервальным вектором

$$\begin{pmatrix} [-1, 3] \\ [-0.5, 2.5] \end{pmatrix}.$$

При использовании других норм результаты, даваемые формулой (3.56), совершенно аналогичны своей грубостью оценивания возмущений решений. Решение задачи получается проще и точнее с помощью интервальных методов (пример 4.6.2). ■

Отметим, что задача оценивания разброса решений СЛАУ при вариациях входных данных является NP-трудной, если не накладывать никаких ограничений на величину возмущений в данных [118, 119]. Это означает, что трудозатраты на её решение в худшем случае растут экспоненциально в зависимости от размера системы.

В заключение темы нужно сообщить читателю, что в математической литературе вместо числа обусловленности матриц не раз предлагались различные конструкции, которые лучше подходят для каких-то узких конкретных целей (см., к примеру, работу [87]). Их поиск продолжится и в будущем, хотя число обусловленностиочно обосновалось в теории и практике вычислений.

## 3.5 Приложения сингулярного разложения

### 3.5а Исследование неособенности и ранга матриц

Сингулярное разложение матриц, рассмотренное в § 3.2ж, может служить основой для вычислительных технологий решения многих математических задач. Рассмотрим первую задачу об определении того, особенна или неособенна матрица.

Исследование особенности или неособенности матрицы обычно проводят с помощью вычисления её определителя и сравнения его с нулем. При этом точное равенство определителя нулю искажается погрешностями, которые вносятся в процесс его вычисления. Но более важен тот факт, что величина ненулевого определителя матрицы не является вполне адекватным признаком того, насколько близка матрица к особенной. Определитель очень сильно изменяется при умножении матрицы на число:

$$\det(\alpha A) = \alpha^n \cdot \det A \quad \text{для } n \times n\text{-матрицы } A.$$

В то же время ясно, что мера линейной независимости столбцов матрицы  $A$  или её строк при таких преобразованиях должна быть либо неизменной, либо изменяющейся не столь сильно.

Более подходящая характеристика особенности или неособенности матрицы может быть основана на учёте собственных значений. Отличие минимального по модулю собственного числа от нуля — это более адекватная мера близости матрицы к особенным. К сожалению, собственные значения несимметричных матриц могут быть очень неустойчивыми, а их вычисление — очень ненадёжным и трудоёмким (см. § 3.17в).

Наиболее надёжным в вычислительном отношении способом проверки особенности или неособенности матрицы является исследование её сингулярных чисел. Квадратная диагональная матрица неособенна тогда и только тогда, когда все её диагональные элементы не равны нулю. Из сингулярного разложения матрицы (3.24) следует, что произвольная квадратная матрица неособенна тогда и только тогда, когда её сингулярные числа — ненулевые. Таким образом, величина наименьшего сингулярного числа матрицы и его отличие от нуля могут служить мерилом того, насколько эта матрица особенна или нет. Хотя нахождение сингулярных чисел матрицы несколько более трудоёмко, чем вычисление её определителя, описанная технология гораздо более предпочтительна в силу существенно лучшей устойчивости к погрешностям вычислений и большей адекватности ответа.

Наилучшей количественной мерой особенности/неособенности матрицы, которая инвариантна относительно её масштабирования, может служить отношение её наибольшего и наименьшего сингулярных чисел, — число обусловленности матрицы относительно спектральной нормы (см. § 3.4а).

Как связано число обусловленности матрицы с её определителем? Это разные характеристики матрицы, каждая из которых служит для своей цели.

**Пример 3.5.1** [112] Пусть

$$A = \begin{pmatrix} \alpha + \beta & \alpha \\ \alpha & \alpha - \beta \end{pmatrix}.$$

Если  $\beta \neq 0$ , то  $\det A = -\beta^2 \neq 0$ . Собственные числа симметричной матрицы  $A$  равны

$$\lambda_{1,2} = \alpha \pm \sqrt{\alpha^2 + \beta^2},$$

а их модули — это сингулярные числа матрицы. Потому спектральное число обусловленности матрицы  $A$  равно (см. (3.58))

$$\frac{|\lambda_2|}{|\lambda_1|} = \frac{\alpha + \sqrt{\alpha^2 + \beta^2}}{|\alpha - \sqrt{\alpha^2 + \beta^2}|} = 1 + \frac{2}{\beta^2} (\alpha^2 + |\alpha| \sqrt{\alpha^2 + \beta^2}).$$

Если  $\beta = \text{const}$ , то и определитель матрицы  $A$  остаётся постоянным, но при этом с ростом  $|\alpha|$  число обусловленности матрицы  $A$  растёт с квадратичной асимптотикой. Ясно, что сама матрица тогда приближается к особенной, так как её строки (или столбцы) становятся всё менее различными.

С другой стороны, если рассмотреть матрицу

$$B = \begin{pmatrix} \alpha + \beta & \alpha \\ \alpha & \alpha + \beta \end{pmatrix},$$

то при  $\beta \neq 0$  и  $2\alpha + \beta \neq 0$  определитель

$$\det B = \beta (2\alpha + \beta) \neq 0.$$

Собственные числа матрицы  $B$  равны

$$\lambda_{1,2} = \alpha + \beta \pm \alpha.$$

Следовательно, если  $\alpha$  и  $\beta$  имеют одинаковый знак, то число обусловленности  $B$  есть

$$1 + 2 |\alpha/\beta|.$$

При тех же условиях, которые были наложены на матрицу  $A$ , т. е.  $\beta = \text{const}$ , определитель  $\det B$  и число обусловленности  $B$  одновременно растут по модулю при увеличении  $|\alpha|$ . Матрица  $B$  при этом также приближается к особенной из-за уменьшающегося различия её строк (столбцов).

Как видим, число обусловленности гораздо лучше определителя соответствует смыслу количественной меры особенности или неособенности матрицы. ■

Обсудим теперь задачу о вычислении ранга матрицы. По определению, ранг — это количество линейно независимых вектор-строк или вектор-столбцов матрицы, с помощью которых можно линейным комбинированием породить всю матрицу. Фактически ранг — число независимых параметров, задающих матрицу. При таком взгляде на ранг

хорошо видна важность этого понятия в задачах обработки данных, когда нам необходимо выявить какие-то закономерности в числовых массивах, полученных в результате наблюдений или опытов. С помощью ранга можно увидеть, к примеру, что все рассматриваемые данные являются линейными комбинациями немногих порождающих.

Ранг матрицы не зависит непрерывно от её элементов. Выражаясь языком, который развивается в главе 4 (§ 4.2), можно сказать, что задача вычисления ранга матрицы не является вычислительно-корректной. Поэтому совершенно точное определение ранга в условиях «зашумлённых» данных, которые искажены случайными помехами и погрешностями измерений, не имеет смысла. Нам нужно, как правило, знать «приближённый ранг», и при прочих равных условиях для его нахождения более предпочтителен тот метод, который менее чувствителен к погрешностям и возмущениям в данных. Под «приближённым рангом» естественно понимать ранг матрицы, приближённо равной исходной в смысле некоторой нормы. Здесь, правда, следует иметь в виду, что матрицы, «приближённо равные» данной в пределах указанной точности, могут иметь разный ранг. Если требуется знать ранг, который гарантированно имеют все матрицы из рассматриваемого множества, то в качестве приближённого ранга имеет смысл взять минимальный из рангов всех этих матриц.

Ранг диагональной матрицы, квадратной или прямоугольной, равен числу её ненулевых диагональных элементов. Поэтому ранг произвольной матрицы равен количеству её ненулевых сингулярных чисел. Это следует из сингулярного разложения  $A = U\Sigma V^*$ , где  $\Sigma$  диагональна, и того факта, что ортогональные преобразования с матрицами  $U$  и  $V^*$  сохраняют линейную зависимость или независимость. Следовательно, при нахождении приближённого ранга матрицы можно задаться каким-либо порогом малости  $\epsilon$ , найти сингулярные числа матрицы и подсчитать, сколько из них больше или равны  $\epsilon$ . Это и будет приближённый ранг матрицы.

Другой способ нахождения ранга матрицы может состоять в приведении её к так называемому строчно-ступенчатому виду с помощью преобразований, которые использовались в прямом ходе метода Гаусса. Но в условиях неточных данных и неточных арифметических операций на ЭВМ строчно-ступенчатая форма является не очень надёжным инструментом из-за своей неустойчивости. Использование сингулярного разложения — более трудоёмкий, но зато существенно более надёжный подход к определению ранга матрицы.

### 3.56 Решение систем линейных уравнений

Пусть дана система линейных алгебраических уравнений  $Ax = b$ , в которой количество переменных совпадает с числом уравнений, т. е. матрица  $A$  — квадратная. Если для  $A$  известно сингулярное разложение (3.24), то система уравнений  $Ax = b$  может быть переписана эквивалентным образом как

$$U\Sigma V^\top x = b.$$

Отсюда

$$x = V\Sigma^{-1}U^\top b.$$

Получается, что для вычисления решения мы должны умножить вектор правой части на ортогональную матрицу, затем разделить компоненты результата на сингулярные числа (если они ненулевые) и, наконец, ещё раз умножить получившийся вектор на другую ортогональную матрицу. С учётом того, что сингулярное разложение матрицы системы нужно ещё найти, вычислительной работы здесь существенно больше, чем при реализации, к примеру, метода исключения Гаусса или других прямых методов решения СЛАУ. Но описанный путь безупречен с вычислительной точки зрения, так как позволяет найти решение системы с минимальным накоплением погрешностей и, кроме того, проанализировать состояние её разрешимости, указав ранг матрицы системы (см. предшествующий раздел).

Напомним, что с геометрической точки зрения преобразования, осуществляемые ортогональными матрицами, являются обобщениями поворотов и отражений: они сохраняют длины и углы. Поэтому в вычислительном отношении умножения на ортогональные матрицы обладают очень хорошими свойствами, так как не увеличивают ошибок округлений и других погрешностей. Ранее в § 3.4б мы взглянули на этот факт с другой стороны, установив, что ортогональные матрицы имеют наименьшее возможное число обусловленности. Отличие в поведении и результатах метода, основанного на сингулярном разложении, и других численных методов особенно зримо в случае, когда матрица системы «почти особенна».

Ещё большую пользу сингулярное разложение приносит при решении переопределённых систем линейных алгебраических уравнений, в которых количество уравнений больше числа неизвестных, так что матрица системы — прямоугольная «стоячая». Обычного решения такая система может не иметь, и тогда находят их *псевдорешения*, которые

минимизируют ту или иную норму невязки левой и правой частей, т. е. разности  $Ax - b$ . В § 2.11г мы уже рассматривали решение этой задачи для случая псевдорешений в евклидовой норме (2-норме), т. е. в смысле «наименьших квадратов». Сингулярное разложение матрицы системы также применяется для нахождения 2-псевдорешений, и преимущество этого пути в том, что он позволяет полностью исследовать задачу и выдавать искомое псевдорешение для любых систем. Опишем эту технологию более подробно.

Евклидова норма (2-норма) вектора не меняется при его умножении на ортогональную матрицу, и поэтому

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \|Ax - b\|_2 &= \min_{x \in \mathbb{R}^n} \|U\Sigma V^\top x - UU^\top b\|_2 = \\ &= \min_{x \in \mathbb{R}^n} \|U(\Sigma V^\top x - U^\top b)\|_2 = \\ &= \min_{x \in \mathbb{R}^n} \|\Sigma V^\top x - U^\top b\|_2 = \min_{y \in \mathbb{R}^n} \|\Sigma y - U^\top b\|_2, \end{aligned}$$

где выполнена неособынная замена переменной  $V^\top x = y$ . Если в системе уравнений  $m \times n$ -матрица  $A$  такова, что  $m \geq n$ , то  $\Sigma$  — диагональная матрица тех же размеров

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix},$$

а  $\sigma_1, \sigma_2, \dots, \sigma_n$  — сингулярные числа  $A$ . Следовательно,

$$\|\Sigma y - U^\top b\|_2^2 = \sum_{i=1}^n (\sigma_i y_i - (U^\top b)_i)^2 + \sum_{i=n+1}^m ((U^\top b)_i)^2,$$

и минимум этого выражения по  $y_i$  достигается при наименьшем значении первой суммы, когда все её слагаемые зануляются. Это происходит при

$$y_i^* = \frac{(U^\top b)_i}{\sigma_i}, \quad i = 1, 2, \dots, n. \quad (3.69)$$

Возвращаясь к исходному  $\min_{x \in \mathbb{R}^n} \|Ax - b\|_2$ , можем найти аргумент  $x^*$ , на котором он достигается, с помощью обратной замены  $x^* = Vy^*$ . Это и есть псевдорешение системы линейных уравнений  $Ax = b$ .

В формуле (3.69) предполагается, что все  $\sigma_i \neq 0$ , т. е. матрица  $A$  имеет полный ранг. Если это не так и какие-то  $\sigma_r = 0$ ,  $r \in \{1, 2, \dots, n\}$ , то соответствующие слагаемые из первой суммы перейдут во вторую сумму постоянных величин, а  $y_r$  при этом можно взять произвольными.

Рассмотренная выше задача называется линейной задачей наименьших квадратов, и она рассматривалась в § 2.11г. Представленное здесь решение на основе сингулярного разложения матрицы является очень общим и весьма информативным, хотя его трудоёмкость больше, чем у других вычислительных методов. Некоторые из них описываются далее в § 3.16.

### 3.5в Малоранговые приближения матрицы

Пусть  $A — m \times n$ -матрица,  $u_k$  и  $v_k$  — это её  $k$ -е нормированные левый и правый сингулярные векторы, а  $\Upsilon_k$  обозначает их внешнее произведение, т. е.

$$\Upsilon_k = u_k v_k^*.$$

Отметим, что  $\Upsilon_k — m \times n$ -матрица ранга 1. Тогда сингулярное разложение (3.24) матрицы  $A$  равносильно её представлению в виде суммы

$$A = \sum_k \sigma_k \Upsilon_k, \quad (3.70)$$

в которой содержится  $\min\{m, n\}$  слагаемых, а множители  $\sigma_i$  — сингулярные числа матрицы  $A$ . Предположим для определённости, что  $\min\{m, n\} = n$ , т. е.  $m \geq n$  и матрица  $A$  — «стоячая».

Если  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n$  и мы «обрубаем» сумму (3.70) после  $s$ -го слагаемого ( $s \leq n$ ), то получающаяся матрица

$$A_s = \sum_{k=1}^s \sigma_k \Upsilon_k \quad (3.71)$$

называется *s-rangовым приближением* матрицы  $A$  или просто *малоранговым приближением*  $A$ .

Это в самом деле матрица ранга  $s$ , что следует из её сингулярного разложения, а погрешность, с которой она приближает исходную

матрицу, равна

$$\sum_{k=s+1}^n \sigma_k \gamma_k.$$

Величина этой погрешности решающим образом зависит от величины сингулярных чисел  $\sigma_{s+1}, \dots, \sigma_n$ , соответствующих отброшенным слагаемым в (3.70). Более точно, погрешность  $s$ -рангового приближения характеризуется следующим замечательным свойством:

**Теорема 3.5.1** (теорема о малоранговом приближении матрицы)

Пусть  $\sigma_k, u_k$  и  $v_k$  — сингулярные числа, левые и правые сингулярные векторы  $m \times n$ -матрицы  $A$  соответственно. Если  $s < n$  и

$$A_s = \sum_{k=1}^s \sigma_k u_k v_k^*$$

—  $s$ -ранговое приближение матрицы  $A$ , то

$$\|A - A_s\|_2 = \min_{\substack{B \in \mathbb{C}^{m \times n} \\ \text{rank } B \leq s}} \|A - B\|_2 = \sigma_{s+1}.$$

Иными словами, относительно спектральной нормы  $s$ -ранговое приближение матрицы обеспечивает наименьшее отклонение от исходной матрицы среди всех матриц ранга не более  $s$ .

**Доказательство.** Предположим, что найдётся такая матрица  $B$ , имеющая ранг  $\text{rank } B \leq s$ , что  $\|A - B\|_2 < \|A - A_s\|_2 = \sigma_{s+1}$ . Тогда существует  $(n - s)$ -мерное подпространство  $W \subset \mathbb{C}^n$ , для которого справедливо  $w \in W \Rightarrow Bw = 0$ . При этом для любого  $w \in W$  мы имеем  $Aw = (A - B)w$ , так что

$$\|Aw\|_2 = \|(A - B)w\|_2 \leq \|A - B\|_2 \|w\|_2 < \sigma_{s+1} \|w\|_2.$$

Таким образом,  $W$  является  $(n - s)$ -мерным подпространством в  $\mathbb{C}^n$ , в котором  $\|Aw\|_2 < \sigma_{s+1} \|w\|_2$ .

Но в  $\mathbb{C}^n$  имеется  $(s + 1)$ -мерное подпространство, образованное векторами  $v$ , для которых  $\|Av\|_2 \geq \sigma_{s+1} \|v\|_2$ . Это подпространство является линейной оболочкой первых  $s + 1$  правых сингулярных векторов матрицы  $A$ . Поскольку сумма размерностей этого подпространства и

подпространства  $W$  превосходит  $n$ , т. е. размерность всего пространства, должен существовать ненулевой вектор, лежащий в них обоих. Это приводит к противоречию. ■

Совершенно аналогичный результат справедлив для фробениусовой нормы матриц, и исторически он был обнаружен даже раньше, чем теорема 3.5.1.

**Теорема 3.5.2** (теорема Экарта–Янга [111])

Пусть  $\sigma_k$ ,  $u_k$  и  $v_k$  — сингулярные числа и левые и правые сингулярные векторы  $m \times n$ -матрицы  $A$  соответственно. Если  $s < n$  и

$$A_s = \sum_{k=1}^s \sigma_k u_k v_k^*$$

—  $s$ -ранговое приближение матрицы  $A$ , то

$$\|A - A_s\|_F = \min_{\substack{B \in \mathbb{C}^{m \times n} \\ \text{rank } B \leq s}} \|A - B\|_F = \sigma_{s+1},$$

где  $\|\cdot\|_F$  — фробениусова норма матриц. Иными словами, относительно фробениусовой нормы  $s$ -ранговое приближение матрицы обеспечивает наименьшее отклонение от исходной матрицы среди всех матриц ранга не более  $s$ .

Доказательство опускается.

Итак, если младшие сингулярные числа матрицы достаточно малы, то вместо неё можно взять  $s$ -ранговое приближение вида (3.71). Оно более «экономно» при небольших  $s$ , т. е. с меньшим числом параметров приближённо представляет исходную матрицу.

### 3.5г Метод главных компонент

В качестве важного практического примера, который иллюстрирует понятия ранга матрицы, сингулярных чисел и сингулярных векторов матрицы, а также результаты предыдущего раздела, рассмотрим *метод главных компонент*, широко применяемый в анализе данных и статистике [92]. В этих дисциплинах решают, в частности, задачи обработки больших массивов числовых данных. Предположим для определённости, что рассматриваемый объект или явление характеризуется

некоторым набором параметров (свойств, признаков и т. п.), которые образуют вектор-строку из  $n$  чисел, и мы имеем  $m$  штук таких векторов, относящихся, к примеру, к отдельным сеансам измерений. Полученные данные образуют числовую  $m \times n$ -матрицу, которую обозначим посредством  $A$ .

Нередко возникает необходимость сжатия данных, т. е. уменьшения числа  $n$  параметров объекта с тем, чтобы оставшиеся  $s$  признаков,  $s < n$ , всё-таки «наиболее полно» описывали всю совокупность накопленной об объекте информации, содержащейся в матрице  $A$ . В более формализованном виде этот вопрос звучит следующим образом: можно ли найти в  $\mathbb{R}^n$  ортонормированный базис  $\{e_1, e_2, \dots, e_s\}$ ,  $s < n$ , в котором рассматриваемые нами данные, содержащиеся в матрице  $A$ , будут представлены в более экономичной, хотя и приближённой, форме?

В качестве меры «близости» матриц мы можем брать различные расстояния, получая различные постановки задач. Одним из практически наиболее важных является расстояние, порождённое фробениусовой нормой матриц. Оно имеет ясный вероятностно-статистический смысл, так как с точностью до множителя совпадает с так называемой выборочной дисперсией набора данных (см. [92] или любой другой учебник по математической статистике). Для фробениусовой нормы матриц наша математическая задача ставится следующим образом. Нужно найти такой ортонормированный базис  $\{e_1, e_2, \dots, e_s\}$  в  $\mathbb{R}^n$ ,  $s \leq n$ , что квадратичное отклонение набора исходных векторов данных  $A_i := (a_{i1}, a_{i2}, \dots, a_{in})^\top$  от их приближений  $X^{(i)} = \sum_{j=1}^s x_{ij} e_j$  в этом базисе было бы наименьшим возможным для всех  $i = 1, 2, \dots, m$ .

Приведённая выше теорема Экарта–Янга даёт математическую основу для решения поставленной задачи. Опираясь на неё процедура малоранговых приближений матрицы данных, которая предварительно «центрирована» путём вычитания из каждого столбца его среднего значения, называется *методом главных компонент*. При этом компонентами называются правые сингулярные векторы  $v_k$ , а масштабированные левые сингулярные векторы  $\sigma_k u_k$  носят название *долей*. Метод главных компонент обычно описывают в терминах собственных чисел и собственных векторов так называемой ковариационной матрицы  $A^\top A$ , но подход, основанный на сингулярном разложении, лучше с вычислительной точки зрения.

Другая ситуация, в которой часто прибегают к методу главных компонент и которая не связана с необходимостью сжатия данных, вызывается желанием выделить из этих данных наиболее значимые *факторы*,

т. е. комбинации переменных, наиболее существенные для рассматриваемого объекта или явления. Здесь и пригождается понятие ранга матрицы или же приближённого ранга для случая неточных данных.

Следует отметить, что соответствующие результаты неоднократно переоткрывались статистиками и, по-видимому, впервые метод главных компонент применял К. Пирсон в начале XX века. В настоящее время метод главных компонент получил широчайшее распространение как один из основных методов анализа многомерных данных и статистики.

## 3.6 Прямые методы решения систем линейных алгебраических уравнений

### 3.6a Основные понятия

Рассмотрим решение систем линейных алгебраических уравнений вида

$$\left\{ \begin{array}{l} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = b_2, \\ \vdots \qquad \vdots \qquad \ddots \qquad \vdots \qquad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n = b_m \end{array} \right. \quad (3.72)$$

с коэффициентами  $a_{ij}$  и свободными членами  $b_i$ , или в краткой форме

$$Ax = b \quad (3.73)$$

с  $m \times n$ -матрицей  $A = (a_{ij})$  и  $m$ -вектором правой части  $b = (b_i)$ . Оно является важной математической задачей, которая повсеместно встречается как сама по себе, так и в качестве составной части технологических цепочек решения более сложных задач. Например, решение нелинейных уравнений или систем уравнений часто сводится к последовательности решений линейных уравнений (см. метод Ньютона в главе 4). В этом и следующем разделах мы рассмотрим задачи нахождения решений и псевдорешений систем линейных алгебраических уравнений (3.72)–(3.73) так называемыми прямыми методами.

Следует отметить, что системы линейных алгебраических уравнений не всегда предъявляются к решению в каноническом виде (3.72).

Процесс решения таких систем в «неканоническом» виде имеет дополнительную специфику, которая иногда жёстко диктует выбор подходящих численных методов.

**Пример 3.6.1** Пусть в  $\mathbb{R}^2$  задана область  $\mathcal{D} = [\underline{x}_1, \bar{x}_1] \times [\underline{x}_2, \bar{x}_2]$ , имеющая форму прямоугольника со сторонами, параллельными координатным осям. Рассмотрим в ней численное решение дифференциального уравнения Лапласа

$$\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = 0 \quad (3.74)$$

для функции двух переменных  $u = u(x_1, x_2)$ .

Уравнение Лапласа является одним из основных и часто встречающихся уравнений математической физики. С помощью него описываются, к примеру, распределение температуры стационарного теплового поля, потенциал электростатического поля при заданном распределении зарядов, течение несжимаемой жидкости и т. п. Для определения конкретного решения этого уравнения задают ещё какие-либо краевые условия на границе расчётной области. Мы будем считать заданными значения искомой функции  $u(x_1, x_2)$  на границе прямоугольника  $[\underline{x}_1, \bar{x}_1] \times [\underline{x}_2, \bar{x}_2]$ :

$$u(\underline{x}_1, x_2) = \underline{f}(x_2), \quad u(\bar{x}_1, x_2) = \bar{f}(x_2), \quad (3.75)$$

$$u(x_1, \underline{x}_2) = \underline{g}(x_1), \quad u(x_1, \bar{x}_2) = \bar{g}(x_1). \quad (3.76)$$

Рассматриваемую задачу определения функции  $u(x_1, x_2)$ , которая удовлетворяет уравнению (3.74) внутри области и условиям (3.75), (3.76) на границе, называют *задачей Дирихле* для уравнения Лапласа.

Станем решать задачу (3.74)–(3.76) с помощью *конечно-разностного метода*, в котором искомая функция заменяется своим дискретным аналогом, а производные в решаемом уравнении заменяются на разностные отношения. Введём на области  $\mathcal{D}$  равномерную прямоугольную сетку, разбив узлами интервал  $[\underline{x}_1, \bar{x}_1]$  на  $m$  частей, а интервал  $[\underline{x}_2, \bar{x}_2]$  — на  $n$  частей. Вместо функции  $u(x_1, x_2)$  непрерывных аргументов  $x_1$  и  $x_2$  будем рассматривать её значения в узлах построенной сетки (рис. 3.15), которые обозначим через  $x_{ij}$ ,  $i = 0, 1, \dots, m$ ,  $j = 0, 1, \dots, n$ .

Если обозначить через  $u_{ij}$  значение искомой функции  $u$  в точке  $x_{ij}$ , то после замены вторых производных формулами (2.90) получим

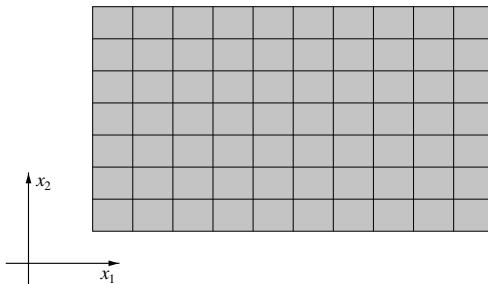


Рис. 3.15. Расчёчная область и сетка для численного решения уравнения Лапласа (3.74)

систему соотношений вида

$$\frac{u_{i-1,j} - 2u_{ij} + u_{i+1,j}}{h_1^2} + \frac{u_{i,j-1} - 2u_{ij} + u_{i,j+1}}{h_2^2} = 0, \quad (3.77)$$

$i = 1, 2, \dots, m - 1, j = 1, 2, \dots, n - 1$ , для внутренних узлов расчётной области. На границе области имеем условия

$$u_{i0} = \underline{f}_i, \quad u_{in} = \overline{f}_i, \quad (3.78)$$

$$u_{0j} = \underline{g}_j, \quad u_{mj} = \overline{g}_j, \quad (3.79)$$

где  $i = 1, 2, \dots, m - 1, j = 1, 2, \dots, n - 1$ , а  $\underline{f}_i, \overline{f}_i, \underline{g}_j, \overline{g}_j$  — значения функций  $f, \overline{f}, g, \overline{g}$  в соответствующих узлах.

Соотношения (3.77) и (3.75)–(3.76) образуют, очевидно, систему линейных алгебраических уравнений относительно неизвестных  $u_{ij}$ ,  $i = 1, 2, \dots, m - 1, j = 1, 2, \dots, n - 1$ , но она не имеет канонический вид (3.72), так как неизвестные имеют по два индекса. Конкретный вид (3.72), который получит эта система уравнений, зависит от способа выбора базиса в пространстве векторов неизвестных, в частности от способа перенумерации этих неизвестных, при котором мы образуем из них вектор с компонентами, имеющими один индекс.

Ясно, что рассмотренный пример может быть сделан ещё более выразительным в трёхмерном случае, когда нам необходимо численно решать трёхмерное уравнение Лапласа. ■

Системы линейных алгебраических уравнений, аналогичные рассмотренной в примере 3.6.1, где матрица и вектор неизвестных не заданы в явном виде, соответствующем (3.73), будем называть системами в *операторной форме*. Не все из изложенных ниже методов решения СЛАУ могут быть непосредственно применены к системам подобного вида.

По характеру вычислительного алгоритма методы решения уравнений и систем уравнений традиционно разделяют на *прямые* и *итерационные*. В прямых методах искомое решение получается в результате выполнения конечной последовательности действий, так что эти методы нередко называют ещё *конечными* или даже *точными*. Напротив, в итерационных методах решение достигается как предел некоторой последовательности приближений, которая конструируется по решаемой системе уравнений.

Одна из главных идей, лежащих в основе прямых методов для решения систем линейных алгебраических уравнений, состоит в том, чтобы эквивалентными преобразованиями привести решаемую систему к наиболее простому виду, из которого решение находится уже непосредственно. В качестве таких простейших могут выступать системы с диагональными, двухдиагональными, треугольными и т. п. матрицами. Чем меньше ненулевых элементов остается в матрице преобразованной системы, тем проще и устойчивее процесс её решения, но в то же время тем сложнее и неустойчивее приведение к такому виду. С другой стороны, диагональный вид матрицы системы позволяет более полно исследовать её и найти значения каждой отдельной неизвестной переменной независимо от других. При треугольной или трапецевидной форме матрицы системы неизвестные находятся друг за другом в цепочке, и обрыв её на какой-то переменной приводит к аварийному завершению всего процесса. Как следствие, тогда мы не сможем найти значения неизвестных, которые получаются в оставшейся части этой цепочки.

На практике обычно стремятся к компромиссу между очерченными выше противоположными ценностями, и в зависимости от целей, преследуемых при решении СЛАУ, приводят её к диагональному (метод Гаусса–Йордана), двухдиагональному [79] или треугольному виду. Мы главным образом рассмотрим методы, основанные на приведении к треугольному виду, так как именно они получили наибольшее распространение в практике вычислений.

Другая плодотворная идея, лежащая в основе прямых методов ре-

шения СЛАУ — разложение решения по специальному базису, который связан с матрицей системы, и нахождение коэффициентов этого разложения по простым выражениям. Например, это может быть разложение в конечный ряд Фурье вида (2.124) по ортогональному базису из собственных векторов симметричной матрицы системы. Существуют также другие подходы и идеи, которые затрагивать уже не будем.

Далее для простоты мы подробно разбираем системы линейных алгебраических уравнений (3.72)–(3.73), в которых  $m \times n$ -матрица коэффициентов  $A = (a_{ij})$  имеет полный ранг. В частности,  $A$  неособенна при  $m = n$ .

### 3.66 Решение треугольных и трапециевидных линейных систем

Напомним, что *треугольными матрицами* называют квадратные матрицы, у которых все элементы ниже главной диагонали либо все элементы выше главной диагонали — нулевые (так что и нулевые, и ненулевые элементы образуют треугольники):

$$U = \begin{pmatrix} \times & \times & \cdots & \times & \times \\ & \times & \ddots & \times & \times \\ & & \ddots & \vdots & \vdots \\ 0 & & & \times & \times \\ & & & & \times \end{pmatrix}, \quad L = \begin{pmatrix} \times & & & & \\ \times & \times & & & 0 \\ \times & \times & \ddots & & \\ \vdots & \vdots & \ddots & \times & \\ \times & \times & \cdots & \times & \times \end{pmatrix},$$

где крестиками « $\times$ » обозначены ненулевые элементы. В первом случае говорят о *верхней* (или *правой*) треугольной матрице, а во втором — о *нижней* (или *левой*) треугольной матрице. Чаще всего эти матрицы обозначают прописными буквами  $L$  и  $U$  — от английских слов «lower» (нижний) и «upper» (верхний). Соответственно, *треугольными* называются системы линейных алгебраических уравнений, матрицы которых имеют треугольный вид — верхний или нижний.

Рассмотрим для определённости линейную систему уравнений

$$Lx = b \tag{3.80}$$

с неособенной нижней треугольной матрицей  $L = (l_{ij})$ , так что  $l_{ij} = 0$  при  $j > i$  и  $l_{ii} \neq 0$  для всех  $i = 1, 2, \dots, n$ . Её первое уравнение содержит

только одну неизвестную переменную  $x_1$ , второе уравнение содержит две неизвестных переменных  $x_1$  и  $x_2$  и т. д., так что в  $i$ -е уравнение входят лишь переменные  $x_1, x_2, \dots, x_i$ . Найдём из первого уравнения значение  $x_1$  и подставим его во второе уравнение системы, в котором в результате останется всего одна неизвестная переменная  $x_2$ . Вычислим  $x_2$  и затем подставим известные значения  $x_1$  и  $x_2$  в третье уравнение, из которого определится  $x_3$ . И так далее.

Описанной выше последовательности действий соответствует следующий простой алгоритм решения линейной системы (3.80) с нижней треугольной  $n \times n$ -матрицей:

```

DO FOR i = 1 TO n
     $x_i \leftarrow \left( b_i - \sum_{j < i} l_{ij} x_j \right) / l_{ii}$ 
END DO

```

(3.81)

Он позволяет последовательно друг за другом вычислить искомые значения неизвестных переменных, начиная с первой. Этот процесс называется *прямой подстановкой*, поскольку он выполняется по возрастанию индексов компонент вектора  $x$ , а его главным содержанием является подстановка на очередном шаге уже найденных значений неизвестных в следующее уравнение.

Для решения систем линейных уравнений  $Ux = b$  с неособенной верхней треугольной матрицей  $U = (u_{ij})$  существует аналогичный процесс, который называется *обратной подстановкой* — он идёт в обратном направлении, т. е. от  $x_n$  к  $x_1$ . Его псевдокод имеет следующий вид:

```
DO FOR i = n DOWNT0 1
```

$$x_i \leftarrow \left( b_i - \sum_{j>i} u_{ij} x_j \right) / u_{ii} . \quad (3.82)$$

```
END DO
```

*Трапециевидные матрицы* — это обобщение треугольных матриц на прямоугольный случай, и нам далее особенно интересны верхние (правые) трапециевидные матрицы

$$\begin{pmatrix} \times & \times & \cdots & \times & \times & \cdots & \times \\ & \times & \ddots & \times & \times & \cdots & \times \\ & & \ddots & \vdots & \vdots & \cdots & \times \\ 0 & & & \times & \times & \cdots & \times \\ & & & & \times & \cdots & \times \end{pmatrix}, \quad \begin{pmatrix} \times & \times & \cdots & \times & \times \\ & \times & \ddots & \times & \times \\ & & \ddots & \vdots & \vdots \\ 0 & & & \times & \times \\ & & & & \times \end{pmatrix},$$

а также системы линейных алгебраических уравнений с ними. Разрешимость и неразрешимость систем линейных алгебраических уравнений с такими матрицами, а также их решения или псевдорешения могут быть легко найдены с помощью процесса обратной подстановки.

Для недоопределённых систем линейных уравнений (имеющих «лежачие» матрицы), у которых  $m < n$ , необходимо предварительно перенести в правую часть члены с переменными  $x_{m+1}, \dots, x_n$ , сделав их свободными параметрами. В результате получаем СЛАУ с треугольной  $m \times m$ -матрицей и вектором правой части с параметрами  $x_{m+1}, \dots, x_n$ .

Для переопределённых трапециевидных систем линейных уравнений, у которых  $m > n$ , т. е. имеющих «стоячие» матрицы (рис. 3.16), точное решение существует, если все правые части с  $n+1$ -й по  $m$ -ю — нулевые. В этом случае оно также находится с помощью обратной подстановки (3.82). Если же правые части с  $n+1$ -й по  $m$ -ю — ненулевые, то такая система уравнений обычного решения не имеет. Но применив процесс

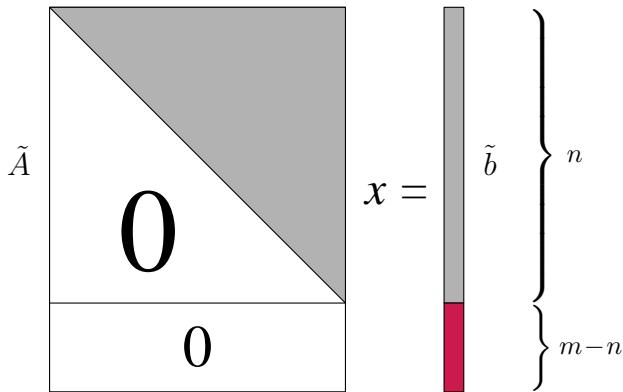


Рис. 3.16. Переопределённая трапециевидная система линейных уравнений и её расчленение перед обратной подстановкой

обратной подстановки (3.82), мы можем легко найти её псевдорешение относительно любой из  $p$ -норм (3.28), в том числе и для  $p = \infty$ , т. е. для чебышёвской нормы.

Действительно, выражение для  $p$ -нормы невязки приближённого решения можно представить в следующем виде:

$$\begin{aligned} \|Ax - b\|_p &= \left( \sum_{i=1}^m |(Ax)_i - b_i|^p \right)^{1/p} = \\ &= \left( \sum_{i=1}^n |(Ax)_i - b_i|^p + \sum_{i=n+1}^m |(Ax)_i - b_i|^p \right)^{1/p}. \end{aligned}$$

Далее, так как в матрице  $A$  строки с номерами  $n+1, \dots, m$  — целиком нулевые, то и  $(Ax)_i = 0$  для  $i = n+1, \dots, m$ . Следовательно,

$$\|Ax - b\|_p = \left( \sum_{i=1}^n |(Ax)_i - b_i|^p + \sum_{i=n+1}^m |b_i|^p \right)^{1/p}.$$

Желая найти  $\min \|Ax - b\|_p$  по всем  $x \in \mathbb{R}^n$ , примем во внимание, что вторая сумма в выражении под степенью  $1/p$  постоянна и не зависит от переменной  $x$ . Поэтому  $\min \|Ax - b\|_p$  достигается одновременно с

минимумом для первой суммы, т. е.

$$\sum_{i=1}^n |(Ax)_i - b_i|^p, \quad (3.83)$$

которая берётся по первым  $n$  уравнениям решаемой системы.

Нетрудно понять, что (3.83) — это  $p$ -я степень  $p$ -нормы невязки квадратной подсистемы линейных алгебраических уравнений  $\tilde{A}x = \tilde{b}$  с матрицей  $\tilde{A}$ , образованной первыми  $n$  строками из  $A$ , и правой частью  $\tilde{b}$ , образованной первыми  $n$  компонентами вектора  $b$  (рис. 3.16). Матрица  $\tilde{A}$  неособенна в силу условия полного ранга, наложенного на всю  $A$ . Минимум невязки приближённого решения для системы  $\tilde{A}x = \tilde{b}$  достигается поэтому на её решении, где он равен нулю. В свою очередь, это решение может быть найдено с помощью процесса обратной подстановки (3.82), поскольку  $\tilde{A}$  — верхняя треугольная.

Насколько полезен полученный результат в общем случае для систем линейных алгебраических уравнений с матрицами общего вида? Это зависит от того, существуют ли удобные эквивалентные преобразования, сохраняющие  $p$ -норму невязки, с помощью которых можно приводить общие линейные системы к трапециевидным, которые выглядят так, как на рис. 3.16. Ниже в § 3.7 мы увидим, что этот вопрос решается положительно для  $p = 2$ , т. е. в случае поиска псевдорешений относительно 2-нормы векторов (евклидовой нормы). Если же  $p \neq 2$ , то удовлетворительного решения поставленного вопроса, к сожалению, не существует. Тем не менее методы нахождения 2-псевдорешений систем линейных алгебраических уравнений, основанные на изложенной выше идее, имеют большое практическое значение и очень популярны при решении реальных задач математического моделирования.

### 3.6в Метод Гаусса для решения линейных систем уравнений

Описываемый в этом разделе *метод Гаусса* для решения систем линейных алгебраических уравнений впервые в Новом времени был описан И. Ньютоном в конце XVII века, хотя письменные источники свидетельствуют о том, что он был известен китайским математикам как минимум за 150 лет до нашей эры. В XIX веке К.Ф. Гаусс разработал удобную вычислительную схему для решения этим методом систем нормальных уравнений (см. § 2.10г). По-видимому, это и послужило причиной называть метод его именем.

Хорошо известно, что умножение какого-либо уравнения системы на ненулевое число, а также замена уравнения на его сумму с другим уравнением системы приводят к равносильной системе уравнений, т. е. имеющей те же самые решения. Воспользуемся этими свойствами для преобразования решаемой системы линейных алгебраических уравнений к более простому виду.

Пусть дана система линейных алгебраических уравнений

$$\left\{ \begin{array}{l} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = b_2, \\ \vdots \qquad \vdots \qquad \ddots \qquad \vdots \qquad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n = b_m, \end{array} \right.$$

в которой коэффициент  $a_{11}$  — ненулевой, т. е.  $a_{11} \neq 0$ . Умножим первое уравнение системы на  $(-a_{21}/a_{11})$  и сложим со вторым уравнением. В результате коэффициент  $a_{21}$  во втором уравнении занулятся, а получившаяся система будет совершенно равносильна исходной.

Проделаем описанное преобразование с остальными — 3-м, 4-м и т. д. до  $m$ -го уравнениями системы, т. е. будем умножать первое уравнение на  $(-a_{i1}/a_{11})$  и складывать с  $i$ -м уравнением системы,  $i = 3, 4, \dots$ . В результате получим равносильную исходной систему линейных алгебраических уравнений, в которой неизвестная переменная  $x_1$  присутствует лишь в первом уравнении. Матрица получившейся СЛАУ станет выглядеть следующим образом:

$$\left( \begin{array}{c|ccccc} a_{11} & \times & \times & \cdots & \times \\ \hline 0 & \times & \times & \cdots & \times \\ 0 & \times & \times & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \times & \times & \cdots & \times \end{array} \right),$$

где посредством « $\times$ » обозначены элементы, возможно, не равные нулю.

Рассмотрим в преобразованной системе уравнения со 2-го по  $m$ -е. Они образуют подсистему линейных уравнений размера  $(m-1) \times (n-1)$ , в которой неизвестная переменная  $x_1$  уже не присутствует и которую можно решать отдельно, никак не обращаясь к первому уравнению исходной системы. Если элемент на месте  $(2, 2)$  не сделался равным

нулю, к этой системе можно заново применить описанную выше процедуру исключения неизвестного. Её результатом будет обнуление поддиагональных элементов  $j$ -го столбца матрицы СЛАУ. И так далее. В зависимости от того, как соотносятся  $m$  и  $n$ , дальнейшее выполнение алгоритма может пойти по-разному и иметь два возможных исхода.

Если  $m \leq n$ , то, выполнив  $(m - 1)$  шагов процесса — для 1-го, 2-го, ...,  $(m - 1)$ -го столбцов матрицы системы, из квадратной или недопределённой СЛАУ мы получим, в конце концов, линейную систему с верхней треугольной матрицей. Она несложно решается с помощью обратной подстановки, рассмотренной в § 3.6б.

Если  $m > n$ , то, выполнив  $n$  шагов процесса — для 1-го, 2-го, ...,  $n$ -го столбцов матрицы системы, из переопределённой СЛАУ получим систему с верхней трапециевидной матрицей. Она также решается с помощью обратной подстановки, рассмотренной в § 3.6б, но может оказаться несовместной.

Таблица 3.1. Прямой ход метода Гаусса

```

DO FOR  $j = 1$  TO  $m - 1$ 
    DO FOR  $i = j + 1$  TO  $m$ 
         $r_{ij} \leftarrow (-a_{ij}/a_{jj})$ 
        DO FOR  $k = j$  TO  $n$ 
             $a_{ik} \leftarrow a_{ik} + r_{ij}a_{jk}$ 
        END DO
         $b_i \leftarrow b_i + r_{ij}b_j$ 
    END DO
END DO

```

Описанное выше преобразование системы линейных алгебраических уравнений к равносильной треугольной или трапециевидной форме называется *прямым ходом* метода Гаусса. Для случая  $m \leq n$  его псевдокод представлен в табл. 3.1. Он выражает процесс последовательного обнуления поддиагональных элементов  $j$ -го столбца матрицы системы,  $j = 1, 2, \dots, m - 1$ , и соответствующие преобразования вектора правой

части. Матрица системы при этом приводится к верхнему треугольному виду. Отметим, что в псевдокоде табл. 3.1 зануление поддиагональных элементов первых столбцов уже учтено нижней границей внутреннего цикла по  $k$ , которая равна  $j$ , а не 1. Аналогичный вид имеет прямой ход при решении переопределённых систем, когда  $m > n$ , с тем единственным отличием, что внешний цикл выполняется до  $j = m$ , а не до  $j = m - 1$ .

После прямого хода метода Гаусса следует его *обратный ход*, на котором решается полученная верхняя треугольная система. Обратный ход является не чем иным, как процессом обратной подстановки из § 3.6б, в котором в обратном порядке последовательно вычисляются искомые значения неизвестных, начиная с  $n$ -й.

Таблица 3.2. Обратный ход метода Гаусса

```

DO FOR i = n DOWNTO 1
     $x_i \leftarrow \left( b_i - \sum_{j>i} a_{ij}x_j \right) / a_{ii}$ 
END DO

```

Трудоёмкость прямого хода метода Гаусса, как нетрудно подсчитать, равна  $O(m^2n)$  арифметических операций (флопсов), а обратного хода —  $O(n^2)$  операций. Для квадратных  $n \times n$ -систем линейных уравнений общая трудоёмкость метода Гаусса оценивается как  $O(n^3)$  флопсов, с главным членом  $\frac{2}{3}n^3$  (подробности вычислений можно увидеть, к примеру, в [40, 46]).

Помимо изложенной выше вычислительной схемы существует много других версий метода Гаусса [48]. Весьма популярной является, к примеру, *схема единственного деления*. При выполнении её прямого хода сначала делят первое уравнение системы на  $a_{11} \neq 0$ , что даёт

$$x_1 + \frac{a_{12}}{a_{11}}x_2 + \cdots + \frac{a_{1n}}{a_{11}}x_n = \frac{b_1}{a_{11}}. \quad (3.84)$$

Умножая затем уравнение (3.84) на  $a_{i1}$ , вычтают результат из  $i$ -го

уравнения системы для  $i = 2, 3, \dots, n$ , добиваясь обнуления поддиагональных элементов первого столбца. Затем процедура повторяется в отношении 2-го уравнения и 2-го столбца получившейся СЛАУ, и так далее. Обратный ход для решения окончательной верхней треугольной системы совпадает с псевдокодом табл. 3.2.

Схема единственного деления совершенно эквивалентна алгоритму табл. 3.1 и отличается от него лишь тем, что для каждого столбца деление в ней выполняется действительно только один раз, тогда как все остальные операции — это умножение и сложение. С другой стороны, уравнения преобразуемой системы в схеме единственного деления дополнительно масштабируются диагональными коэффициентами при неизвестных, и в некоторых случаях это бывает нежелательно.

### 3.6г Матричная интерпретация метода Гаусса

Умножение первого уравнения системы на  $r_{i1} = -a_{i1}/a_{11}$  и сложение его с  $i$ -м уравнением могут быть представлены как умножение обеих частей системы уравнений  $Ax = b$  слева на матрицу

$$\begin{pmatrix} 1 & & & & 0 \\ 0 & 1 & & & \\ \vdots & & \ddots & & \\ r_{i1} & & & 1 & \\ \vdots & & & & 1 \\ 0 & 0 & & & 1 \end{pmatrix},$$

которая отличается от единичной матрицы наличием одного дополнительного ненулевого элемента  $r_{i1}$  на месте  $(i, 1)$ . Матрицы такого вида называются *трансвекциями* [47, 68]. Исключение поддиагональных элементов первого столбца матрицы СЛАУ в прямом ходе метода Гаусса (табл. 3.1) — это последовательное домножение обеих частей этой

системы слева на матрицы

$$\begin{pmatrix} 1 & & & 0 \\ r_{21} & 1 & & \\ 0 & & \ddots & \\ \vdots & & & 1 \\ 0 & 0 & & & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & & & 0 \\ 0 & 1 & & \\ r_{31} & 0 & \ddots & \\ \vdots & 0 & & 1 \\ 0 & & & 1 \end{pmatrix},$$

и так далее до

$$\begin{pmatrix} 1 & & & 0 \\ 0 & 1 & & \\ \vdots & & \ddots & \\ 0 & 0 & & 1 \\ r_{n1} & & & & 1 \end{pmatrix}.$$

Нетрудно убедиться, что умножение трансвекций выписанного выше специального вида выполняется по простому правилу

$$\begin{pmatrix} 1 & & & 0 \\ & 1 & & \\ r_{i1} & & \ddots & \\ & & 1 & \\ 0 & & \ddots & & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & & & 0 \\ & 1 & & \\ & & \ddots & \\ r_{k1} & & & 1 \\ & & & \ddots & & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & & & 0 \\ & 1 & & \\ r_{i1} & & \ddots & \\ & & 1 & \\ r_{k1} & & \ddots & & 1 \\ 0 & & & & \ddots & & 1 \end{pmatrix}.$$

Оно также остаётся верным в случае, когда у матриц-сомножителей на несовпадающих местах в первом столбце присутствует более одного ненулевого элемента. Следовательно, обнуление всех поддиагональных элементов первого столбца и соответствующие преобразования правой части в методе Гаусса — это не что иное, как умножение обеих частей СЛАУ слева на матрицу

$$E_1 = \begin{pmatrix} 1 & & & & 0 \\ r_{21} & 1 & & & \\ r_{31} & 0 & 1 & & \\ \vdots & & & \ddots & \\ r_{n1} & 0 & & & 1 \end{pmatrix}. \quad (3.85)$$

Аналогично обнуление всех поддиагональных элементов  $j$ -го столбца матрицы СЛАУ и соответствующие преобразования правой части можно интерпретировать как умножение системы слева на матрицу

$$E_j = \begin{pmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & 1 & & \\ & 0 & r_{j+1,j} & 1 & \\ & & \vdots & & \ddots \\ & & & r_{nj} & 1 \end{pmatrix}. \quad (3.86)$$

В целом метод Гаусса представляется как последовательность умножений обеих частей решаемой СЛАУ слева на матрицы  $E_j$  вида (3.86),  $j = 1, 2, \dots, n - 1$ . При этом матрицей системы становится матрица

$$E_{n-1} \cdots E_2 E_1 A = U, \quad (3.87)$$

которая является верхней треугольной.

Коль скоро все  $E_j$  — нижние треугольные матрицы, их произведение тоже является нижним треугольным. Кроме того, все  $E_j$  неособенны (нижние треугольные с единицами по главной диагонали). Поэтому неособенно и их произведение  $E_{n-1} \cdots E_2 E_1$ . Если определить

$$L = (E_{n-1} \cdots E_2 E_1)^{-1},$$

то, как нетрудно понять,  $L$  — тоже нижняя треугольная матрица с единицами по главной диагонали. Для этой матрицы в силу (3.87) справедливо равенство

$$A = LU.$$

Получается, что исходная матрица СЛАУ оказалась представленной в виде произведения нижней треугольной  $L$  и верхней треугольной  $U$  матриц. Это представление называют *треугольным разложением* матрицы или *LU-разложением*.<sup>15</sup> Соответственно, преобразования матрицы  $A$  в прямом ходе метода Гаусса из табл. 3.1 можно трактовать как её разложение на нижний треугольный  $L$  и верхний треугольный  $U$  множители.

Отметим, что если LU-разложение матрицы  $A$  уже дано, то система  $Ax = b$  может быть переписана в равносильной форме

$$L(Ux) = b.$$

Тогда её решение сводится к решению двух треугольных систем линейных алгебраических уравнений

$$\begin{cases} Ly = b, \\ Ux = y \end{cases} \quad (3.88)$$

с помощью прямой и обратной подстановок соответственно. LU-разложение, получаемое с помощью версии метода Гаусса из табл. 3.1 и 3.2), как уже отмечалось, обладает тем свойством, что в нижней треугольной матрице  $L$  по диагонали стоят все единицы. При реализации такого метода Гаусса на компьютере для экономии машинной памяти можно хранить треугольные сомножители  $L$  и  $U$  на месте  $A$ , так как диагональ в  $L$  имеет фиксированный вид.

### 3.6д Метод Гаусса с выбором ведущего элемента

И в прямом и в обратном ходе метода Гаусса встречаются операции деления, которые не выполнимы в случае нулевого делителя. Тогда не может быть выполнен и метод Гаусса в целом. Этот раздел посвящен

---

<sup>15</sup>Нередко для обозначения этого же понятия можно встретить кальки с иностранных терминов — «LU-факторизация» и «LU-декомпозиция».

тому, как модифицировать метод Гаусса, чтобы он был применим для решения любых СЛАУ с неособенными матрицами.

*Ведущим элементом* в методе Гаусса называют элемент матрицы решаемой системы, на который в прямом ходе выполняется деление при исключении поддиагональных элементов очередного столбца.<sup>16</sup> В алгоритме табл. 3.1 из предыдущего раздела ведущим всюду берётся фиксированный диагональный элемент  $a_{jj}$ , вне зависимости от его значения. Необходимо модифицировать метод Гаусса так, чтобы ведущий элемент, по возможности, всегда был отличен от нуля. С другой стороны, при решении конкретных СЛАУ, даже в случае  $a_{jj} \neq 0$ , по соображениям устойчивости алгоритма более предпочтительным может оказаться выбор другого элемента в качестве ведущего.

Отметим, что любое изменение порядка уравнений в системе приводит к равносильной системе уравнений, хотя при этом в матрице СЛАУ переставляются строки и она заметно меняется. Этим наблюдением можно воспользоваться для организации успешного выполнения метода Гаусса.

Назовём *активной подматрицей*  $j$ -го шага прямого хода метода Гаусса квадратную подматрицу, которая образована строками и столбцами с номерами  $j, j+1, \dots, n$  в матрице СЛАУ, полученной в результате  $(j-1)$  шагов прямого хода. Именно эта подматрица подвергается преобразованиям на  $j$ -м шаге прямого хода, тогда как первые  $j-1$  строк и столбцов матрицы системы остаются уже неизменными.

*Частичным выбором* ведущего элемента на  $j$ -м шаге прямого хода метода Гаусса называют его выбор как максимального по модулю элемента из всех элементов  $j$ -го столбца, лежащих не выше диагонали. Соответственно, частичный выбор ведущего элемента сопровождается необходимой перестановкой строк матрицы и компонент правой части (т. е. уравнений СЛАУ), при которых этот максимальный по модулю элемент становится диагональным. В методе Гаусса на него выполняется деление, и потому именно максимальным по модулю, а не просто ненулевым, ведущий элемент имеет смысл выбирать для того, чтобы обеспечить, по возможности, наименьшее накопление погрешностей в реальных вычислениях с конечной точностью (см. § 1.2).

**Предложение 3.6.1** *Метод Гаусса с частичным выбором ведущего элемента всегда выполним для систем линейных алгебраических уравнений с неособенными квадратными матрицами.*

---

<sup>16</sup>Иногда его называют также *главным элементом*.

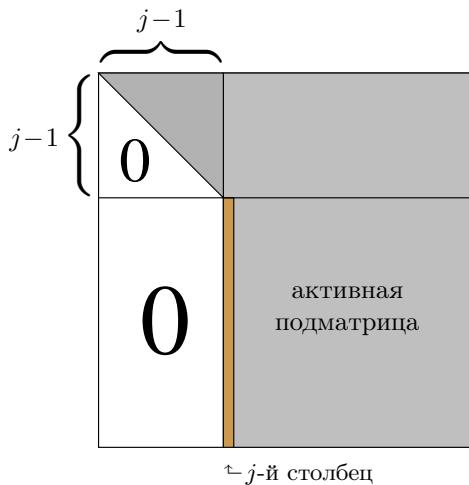


Рис. 3.17. Структура матрицы СЛАУ перед началом  $j$ -го шага прямого хода метода Гаусса

**Доказательство.** Преобразования прямого хода метода Гаусса сохраняют свойство определителя матрицы системы быть неравным нулю. Перед началом  $j$ -го шага прямого хода эта матрица имеет блочно-треугольный вид, изображённый на рис. 3.17, и поэтому её определитель равен произведению определителей диагональных блоков, т. е. определителей ведущей подматрицы порядка  $(j - 1)$  и активной подматрицы порядка  $n - j + 1$ . Как следствие, активная подматрица имеет ненулевой определитель, так что в первом её столбце обязан найтись хотя бы один ненулевой элемент. Максимальный по модулю из этих ненулевых элементов — также ненулевой, и его мы делаем ведущим. Итак, прямой ход метода Гаусса выполним.

Обратный ход тоже не встречает деления на нуль, поскольку полученная в прямом ходе верхняя треугольная матрица неособенна, т. е. все её диагональные элементы должны быть ненулевыми. ■

Чтобы выписать матричное представление метода Гаусса с частичным выбором ведущего элемента, напомним

**Определение 3.6.1** Элементарной матрицей перестановки называет-

ся матрица вида

$$P = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & 0 & \cdots & & 1 \\ & & 1 & & \\ \vdots & & \ddots & & \vdots \\ & & & 1 & \\ 1 & \cdots & & 0 & \\ & & & & \ddots \\ & & & & 1 \end{pmatrix}, \quad \begin{array}{l} \leftarrow i\text{-я строка} \\ \leftarrow j\text{-я строка} \end{array} \quad (3.89)$$

которая получается из единичной матрицы взаимной перестановкой местами двух её строк (или столбцов). Матрицей перестановки называется матрица, которая получается из единичной матрицы перестановкой произвольного числа её строк (или столбцов).

Фактически матрица перестановки — это квадратная матрица, образованная элементами 0 и 1, в каждой строке и столбце которой находится ровно один единичный элемент. Матрица перестановки может быть представлена как произведение нескольких элементарных матриц перестановки вида (3.89) [7].

Иногда для матриц (3.89) используют также термин *матрица транспозиции*. Если элементарная матрица перестановки отличается от единичной строками (столбцами) с номерами  $i$  и  $j$ , то умножение её слева на любую матрицу приводит к перестановке в этой матрице  $i$ -й и  $j$ -й строк, а при умножении справа — к перестановке  $i$ -го и  $j$ -го столбцов. Тогда для прямого хода метода Гаусса с частичным выбором ведущего элемента справедливо следующее матричное представление:

$$(E_{n-1} P_{n-1}) \cdots (E_1 P_1) A = U,$$

где  $E_j$  — матрицы преобразований вида (3.86), введённые в предыдущем разделе, а  $P_1, P_2, \dots, P_{n-1}$  — элементарные матрицы перестановок (3.89), при помощи которых выполняется необходимая «перетасовка» строк на 1-м, 2-м, …,  $(n - 1)$ -м шагах прямого хода метода Гаусса.

Несмотря на то что метод Гаусса с частичным выбором ведущего элемента теоретически работоспособен для любых СЛАУ с неособенными матрицами, на практике для некоторых «плохих» систем он всё-

таки может работать недостаточно устойчиво. Это происходит в случаях, когда на прямом ходе (табл. 3.1) ведущие элементы  $a_{jj}$  оказываются малыми в сравнении с другими  $a_{ij}$  в столбце. Тогда при вычислении коэффициентов  $r_{ij} = -a_{ij}/a_{jj}$  деление на  $a_{jj}$  может сопровождаться большими погрешностями (см. § 1.3), а сами  $r_{ij}$  получаются большими по абсолютной величине.

Детальный анализ погрешностей вычислений в методе Гаусса (его можно увидеть, к примеру, в [51]) показывает, что для общих матриц желательно поддерживать коэффициенты  $r_{ij}$  по модулю не большими единицы. По этим причинам для обеспечения лучшей вычислительной устойчивости метода Гаусса иногда имеет смысл выбирать ведущий элемент более тщательно, чем это делается при описанном выше частичном выборе.

Вспомним, что ещё одним простым способом равносильного преобразования системы уравнений является перенумерация переменных. Ей соответствует перестановка столбцов матрицы, тогда как вектор правых частей при этом неизменен. Полным выбором ведущего элемента называют способ его выбора как максимального по модулю элемента из всей активной подматрицы (а не только из её первого столбца, что характерно при частичном выборе). Полный выбор ведущего элемента сопровождается соответствующей перестановкой строк и столбцов матрицы и компонент правой части. Прямой ход метода Гаусса с полным выбором ведущего элемента имеет следующее матричное представление:

$$(E_{n-1}\check{P}_{n-1}) \cdots (E_1\check{P}_1)A\hat{P}_1 \cdots \hat{P}_{n-1} = U,$$

где  $\check{P}_i$  — элементарные матрицы перестановки, при помощи которых выполняется перестановка строк,  $\hat{P}_j$  — элементарные матрицы перестановки, с помощью которых выполняется перестановка столбцов на соответствующих шагах прямого хода метода Гаусса.

**Теорема 3.6.1** Для неособенной матрицы  $A$  существуют матрицы перестановки  $\check{P}$  и  $\hat{P}$ , такие что

$$\check{P}A\hat{P} = LU,$$

где  $L$ ,  $U$  — нижняя и верхняя треугольные матрицы, причём диагональными элементами в  $L$  являются единицы. В этом представлении можно ограничиться лишь одной из матриц —  $\check{P}$  или  $\hat{P}$ .

Этот результат показывает, что можно один раз переставить строки и столбцы в исходной матрице и потом уже выполнять LU-разложение прямым ходом метода Гаусса без какого-либо специального выбора ведущего элемента. Доказательство теоремы можно найти в [11, 13, 38].

### 3.6e Алгоритмы Дулитла и Кроута

Существуют ли какие-то другие способы получения LU-разложения матрицы кроме прямого хода метода Гаусса? Для этого разработаны, в частности, алгоритм Дулитла и алгоритм Кроута, теоретической основой которых является следующий результат:

**Теорема 3.6.2** Пусть  $A = (a_{ij})$  — квадратная  $n \times n$ -матрица, у которой все ведущие миноры порядков от 1 до  $(n-1)$  отличны от нуля, т. е.

$$a_{11} \neq 0, \quad \det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \neq 0, \quad \dots,$$

$$\det \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1,n-1} \\ a_{21} & a_{22} & \dots & a_{2,n-1} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n-1,1} & a_{n-1,2} & \dots & a_{n-1,n-1} \end{pmatrix} \neq 0.$$

Тогда для  $A$  существует LU-разложение, т. е. представление её в виде

$$A = LU$$

— произведения нижней треугольной  $n \times n$ -матрицы  $L$  и верхней треугольной  $n \times n$ -матрицы  $U$ . Это LU-разложение для  $A$  единственно при условии, что диагональными элементами в  $L$  являются единицы.

**Доказательство** проводится индукцией по порядку  $n$  матрицы  $A$ .

Если  $n = 1$ , то утверждение теоремы очевидно. Тогда искомые матрицы  $L = (l_{ij})$  и  $U = (u_{ij})$  являются просто числами и достаточно взять  $l_{11} = 1$  и  $u_{11} = a_{11}$ .

Пусть теорема верна для матриц размера  $(n-1) \times (n-1)$ . Если  $A$

—  $n \times n$ -матрица, то представим её в блочном виде:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} = \begin{pmatrix} A_{n-1} & z \\ v & a_{nn} \end{pmatrix},$$

где  $A_{n-1}$  — ведущая  $(n-1) \times (n-1)$ -подматрица из  $A$ ,

$z$  — вектор-столбец размера  $n-1$ ,

$v$  — вектор-строка размера  $n-1$ ,

такие что

$$z = \begin{pmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{n-1,n} \end{pmatrix}, \quad v = (a_{n1} \ a_{n2} \ \dots \ a_{n,n-1}).$$

Требование разложения  $A$  на треугольные множители диктует равенство

$$A = \begin{pmatrix} A_{n-1} & z \\ v & a_{nn} \end{pmatrix} = \begin{pmatrix} L_{n-1} & 0 \\ x & l_{nn} \end{pmatrix} \cdot \begin{pmatrix} U_{n-1} & y \\ 0 & u_{nn} \end{pmatrix},$$

где  $L_{n-1}, U_{n-1}$  — нижняя и верхняя треугольные  $(n-1) \times (n-1)$ -матрицы,

$x$  — вектор-строка размера  $n-1$ ,

$y$  — вектор-столбец размера  $n-1$ .

Следовательно, используя правила перемножения матриц по блокам, необходимо имеем

$$A_{n-1} = L_{n-1}U_{n-1}, \tag{3.90}$$

$$z = L_{n-1}y, \tag{3.91}$$

$$v = xU_{n-1}, \tag{3.92}$$

$$a_{nn} = xy + l_{nn}u_{nn}. \tag{3.93}$$

Первое из полученных соотношений выполнено в силу индукционного предположения, причём оно должно однозначно определять  $L_{n-1}$  и  $U_{n-1}$ , если потребовать по диагонали в  $L_{n-1}$  единичные элементы. Далее, по условию теоремы  $\det A_{n-1} \neq 0$ , а потому матрицы  $L_{n-1}$  и

$U_{n-1}$  тоже должны быть неособенны. По этой причине системы линейных уравнений относительно  $y$  и  $x$  —

$$L_{n-1}y = z \quad \text{и} \quad xU_{n-1} = v,$$

которыми являются равенства (3.91)–(3.92), однозначно разрешимы. Стоит отметить, что именно в этом месте доказательства индукционный переход неявно опирается на условие теоремы, которое требует, чтобы в матрице  $A$  все ведущие миноры порядков, меньших чем  $n$ , были ненулевыми.

Найдя из (3.91)–(3.92) векторы  $y$  и  $x$ , мы сможем из соотношения (3.93) восстановить  $l_{nn}$  и  $u_{nn}$ . Если дополнительно положить  $l_{nn} = 1$ , то значение  $u_{nn}$  находится однозначно и равно  $(a_{nn} - xy)$ . ■

В теореме 3.6.2 не требуется неособенность всей матрицы  $A$ . Из доказательства нетрудно видеть, что при наложенных на  $A$  условиях её LU-разложение будет существовать даже при  $\det A = 0$ , но тогда в матрице  $U$  последний элемент  $u_{nn}$  будет равен нулю.

Доказательство теоремы 3.6.2 является совершенно конструктивным и его легко воплотить в компьютерный алгоритм, который находит LU-разложение матрицы способом, альтернативным прямому ходу метода Гаусса. Он называется *алгоритмом Дулитла*, а его псевдокод приведён в табл. 3.3.

В алгоритме Дулитла внешний цикл выполняется по параметру  $i$ , от 1 до  $n$ , и в нём на основе треугольных матриц размера  $(i-1) \times (i-1)$ , построенных на предыдущем шаге, достраиваются треугольные  $i \times i$ -матрицы  $U = (u_{ij})$  и  $L = (l_{ij})$ . Для этого в двух внутренних циклах по  $j$  с помощью прямой подстановки решаются треугольные линейные системы вида (3.91) и (3.92) с матрицами, полученными с предыдущего шага. В результате для матриц  $U$  и  $L$  очередного шага находятся неполный  $i$ -й столбец и неполная  $i$ -я строка соответственно (они обозначаются как  $y$  и  $x$  в доказательстве теоремы 3.6.2). При этом полагаем  $l_{ii} = 1$ , а элемент  $u_{ii}$  находим из равенства, аналогичного (3.93) (эта инструкция включена в первый внутренний цикл).

Заметим, что в доказательстве теоремы 3.6.2 элементы  $l_{nn}$  и  $u_{nn}$  могут быть определены из равенства (3.93), вообще говоря, бесчисленным количеством способов. Назначение  $l_{nn} = 1$  является лишь одним из возможных вариантов, часто практически наиболее удобным. Если зафиксировать не  $l_{nn}$ , а элемент  $u_{nn}$  каким-нибудь ненулевым значением, то из равенства (3.93) однозначно определится  $l_{nn}$ .

Таблица 3.3. Алгоритм Дулитла для вычисления LU-разложения матрицы

```

DO FOR  $i = 1$  TO  $n$ 
    DO FOR  $j = i$  TO  $n$ 
         $u_{ij} \leftarrow a_{ij} - \sum_{k=1}^{i-1} l_{ik} u_{kj}$ 
    END DO
     $l_{ii} \leftarrow 1$ 
    DO FOR  $j = i + 1$  TO  $n$ 
         $l_{ji} \leftarrow \frac{1}{u_{ii}} \left( a_{ji} - \sum_{k=1}^{i-1} l_{jk} u_{ki} \right)$ 
    END DO
END DO

```

(3.94)

Если полагать  $u_{ii} = 1$  при алгоритмической реализации доказательства теоремы 3.6.2, то получим так называемый *алгоритм Кроута* для LU-разложения. В нём единичную диагональ получает не нижний треугольный сомножитель  $L$ , а верхняя треугольная матрица  $U$ , что иногда может оказаться более предпочтительным на практике. Мы не приводим отдельного псевдокода алгоритма Кроута, так как он очень похож на алгоритм Дулитла: порядок внутренних циклов противоположен тому, что использован в табл. 3.3, и вместо инструкции  $l_{ii} \leftarrow 1$  присутствует  $u_{ii} \leftarrow 1$ .

Наконец, если потребовать  $l_{ii} = u_{ii}$ ,  $i = 1, 2, \dots, n$ , а также равенства соответствующих элементов строк и столбцов треугольных сомножителей  $L$  и  $U$ , то тогда  $U = L^\top$  и получаем разложение Холесского, которому посвящены отдельные § 3.63 и 3.6и.

Нетрудно подсчитать, что для квадратных  $n \times n$ -матриц трудоёмкость выполнения алгоритмов Дулитла и Кроута оценивается как  $O(n^3)$  арифметических операций.

### 3.6ж Существование LU-разложения

В методе Гаусса с выбором ведущего элемента перестановка строк и столбцов может привести к существенному изменению исходной матрицы системы, что иногда нежелательно. Естественно задаться вопросом о достаточных условиях реализуемости метода Гаусса без перестановки строк и столбцов. Этот вопрос тесно связан с условиями получения LU-разложения матрицы посредством прямого хода «немодифицированного» метода Гаусса из § 3.6в либо с помощью алгоритмов Дулитла или Кроута из предыдущего раздела. Частично мы ответили на него в предшествующем разделе теоремой 3.6.2, но имеет смысл исследовать тему глубже.

В связи с матрицами, имеющими ненулевые ведущие миноры, полезно следующее

**Определение 3.6.2** Квадратная матрица  $A = (a_{ij})$  называется строго регулярной (или строго неособенной)<sup>17</sup>, если все её ведущие миноры, включая определитель самой матрицы, отличны от нуля, т. е. если

$$a_{11} \neq 0, \quad \det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \neq 0, \quad \dots, \quad \det A \neq 0.$$

**Теорема 3.6.3** Пусть  $A$  — квадратная неособенная матрица. Для существования её LU-разложения необходимо и достаточно, чтобы она была строго регулярной.

**Доказательство.** Достаточность мы уже доказали в теореме 3.6.2.

Для доказательства необходимости привлечём блочное представление треугольного разложения  $A = LU$  (рис. 3.18). Задавая различные размеры ведущих  $k \times k$ -подматриц  $A_k$ ,  $L_k$  и  $U_k$  в матрицах  $A$ ,  $L$  и  $U$  и применяя правила умножения блочных матриц, получим аналогичные (3.90) равенства

$$A_k = L_k U_k, \quad k = 1, 2, \dots, n. \quad (3.95)$$

Они означают, что любая ведущая подматрица в  $A$  есть произведение ведущих подматриц соответствующих размеров из  $L$  и  $U$ .

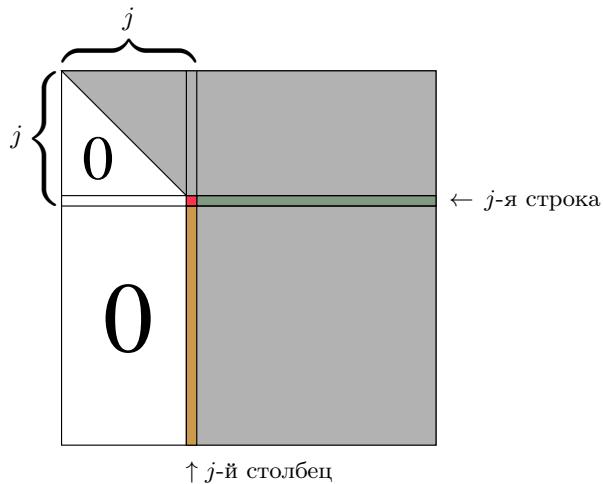
---

<sup>17</sup>Соответствующие английские термины — strictly regular matrix, strictly nonsingular matrix.

$$\begin{array}{|c|c|} \hline A_k & \\ \hline \end{array} = \begin{array}{|c|c|} \hline L_k & 0 \\ \hline \end{array} \cdot \begin{array}{|c|c|} \hline U_k & \\ \hline 0 & \\ \hline \end{array}$$

Рис. 3.18. Блочное умножение в LU-разложении матрицы

Но  $L$  и  $U$  — неособенные треугольные матрицы, так что все их ведущие подматрицы  $L_k$  и  $U_k$  также неособенны. Поэтому из равенств (3.95) можно заключить неособенность всех ведущих подматриц  $A_k$  в  $A$ , т. е. строгую регулярность матрицы  $A$ . ■

Рис. 3.19. Структура матрицы СЛАУ перед началом  $j$ -го шага прямого хода метода Гаусса: другой вид

В формулировке теоремы 3.6.2 ничего не говорится о том, реализуем ли метод Гаусса для соответствующей системы линейных алгебраических уравнений. Но нетрудно понять, что в действительности требуе-

мое теоремой 3.6.2 условие отличия от нуля ведущих миноров в матрице СЛАУ является достаточным для выполнимости варианта метода Гаусса, рассмотренного в § 3.6в.

**Предложение 3.6.2** *Если в системе линейных алгебраических уравнений  $Ax = b$  матрица  $A$  — квадратная и строго регулярная, то метод Гаусса реализуем в применении к этой системе без перестановки строк и столбцов.*

**Доказательство.** К началу  $j$ -го шага прямого хода, на котором предстоит обнулить поддиагональные элементы  $j$ -го столбца матрицы системы, её ведущая  $j \times j$ -подматрица является треугольной, и она получена из исходной ведущей подматрицы преобразованиями предыдущих  $j - 1$  шагов метода Гаусса (рис. 3.19). Эти преобразования — линейное комбинирование строк — не изменяют свойство определителя матрицы быть неравным нулю. Поэтому отличие от нуля какого-либо ведущего минора влечёт отличие от нуля всех диагональных элементов ведущей треугольной подматрицы того же размера в преобразованной матрице системы. В частности, всегда  $a_{jj} \neq 0$ , так что деление на этот элемент в алгоритмах из табл. 3.1 и 3.2 выполнимо. ■

В общем случае проверка условий теоремы 3.6.2 или строгой регулярности матрицы является весьма непростой, поскольку вычисление ведущих миноров матрицы требует немалых трудозатрат и по существу ничуть не проще самого метода Гаусса. Тем не менее условия теоремы 3.6.2 заведомо выполнены, к примеру, в двух важных частных случаях:

- для симметричных матриц, которые положительно определены или же отрицательно определены, в силу известного критерия Сильвестра,
- для матриц с диагональным преобладанием в силу признака Адамара, см. § 3.4в (если исходная матрица имеет диагональное преобладание, то его имеют и все ведущие подматрицы).

### 3.6з Разложение Холесского

Напомним, что квадратная  $n \times n$ -матрица называется *положительно определённой*, если  $\langle Ax, x \rangle > 0$  для любых ненулевых  $n$ -векторов  $x$ ,

или, иными словами  $x^\top Ax > 0$  для любого  $x \neq 0$ . Ясно, что положительно-определенные матрицы неособенны.

**Теорема 3.6.4** (теорема о разложении Холесского) *Матрица  $A$  является симметричной положительно определённой тогда и только тогда, когда существует неособенная нижняя треугольная матрица  $C$ , такая что  $A = CC^\top$ . При этом матрица  $C$  из выписанного представления единственна.*

**Определение 3.6.3** *Представление  $A = CC^\top$  называется разложением Холесского, а нижняя треугольная матрица  $C$  — множителем Холесского для  $A$ .*

**Доказательство.** Пусть  $A = CC^\top$  и  $C$  неособенна. Тогда неособенна матрица  $C^\top$ , и для любого ненулевого вектора  $x \in \mathbb{R}^n$  имеем

$$\begin{aligned}\langle Ax, x \rangle &= (Ax)^\top x = (CC^\top x)^\top x = \\ &= x^\top CC^\top x = (C^\top x)^\top (C^\top x) = \|C^\top x\|_2^2 > 0,\end{aligned}$$

поскольку  $C^\top x \neq 0$ . Кроме того,  $A$  симметрична по построению. Таким образом, она является симметричной положительно определённой матрицей.<sup>18</sup>

Обратно, пусть матрица  $A$  симметрична и положительно определена. В силу критерия Сильвестра все её ведущие миноры положительны, а потому на основании теоремы 3.6.2 о существовании LU-разложения можем заключить, что  $A = LU$  для некоторых неособенных нижней треугольной матрицы  $L = (l_{ij})$  и верхней треугольной матрицы  $U$ . Если дополнительно потребовать, чтобы все диагональные элементы  $l_{ii}$  в  $L$  были единицами, то из теоремы 3.6.2 будет следовать, что это разложение однозначно определено.

Так как

$$LU = A = A^\top = (LU)^\top = U^\top L^\top,$$

то

$$U = L^{-1}U^\top L^\top, \tag{3.96}$$

---

<sup>18</sup>Это рассуждение не использует треугольность  $C$  и на самом деле обосновывает общее утверждение: произведение неособенной квадратной матрицы на её транспонированную является симметричной положительно определённой матрицей.

и далее

$$U(L^\top)^{-1} = L^{-1}U^\top.$$

Слева в этом равенстве стоит произведение верхних треугольных матриц, а справа — произведение нижних треугольных. Равенство, следовательно, возможно лишь в случае, когда левая и правая его части — это диагональная матрица, которую мы обозначим через  $D$ , так что

$$D := \text{diag}\{d_1, d_2, \dots, d_n\} = U(L^\top)^{-1} = L^{-1}U^\top.$$

Тогда из (3.96) вытекает

$$U = L^{-1}U^\top L^\top = DL^\top,$$

и потому

$$A = LU = LDL^\top. \quad (3.97)$$

В силу неособенности  $L$  и  $U$  матрица  $D$  также неособенна, так что по диагонали у неё стоят ненулевые элементы  $d_i$ ,  $i = 1, 2, \dots, n$ . Более того, мы покажем, что все  $d_i$  положительны.

Из (3.97) следует, что  $D = L^{-1}A(L^\top)^{-1} = L^{-1}A(L^{-1})^\top$ . Следовательно, для любого ненулевого вектора  $x$

$$\begin{aligned} \langle Dx, x \rangle &= x^\top Dx = x^\top L^{-1}A(L^{-1})^\top x = \\ &= ((L^{-1})^\top x)^\top A((L^{-1})^\top x) = \langle A(L^{-1})^\top x, (L^{-1})^\top x \rangle > 0, \end{aligned}$$

так как  $(L^{-1})^\top x \neq 0$  в силу неособенности матрицы  $(L^{-1})^\top$ . Иными словами, диагональная матрица  $D$  положительно определена одновременно с  $A$ . Но тогда её диагональные элементы обязаны быть положительными. В противном случае, если предположить, что  $d_i \leq 0$  для некоторого  $i$ , то, беря вектор  $x$  равным  $i$ -му столбцу единичной матрицы, получим

$$\langle Dx, x \rangle = (Dx)^\top x = x^\top Dx = d_i \leq 0.$$

Это противоречит положительной определённости матрицы  $D$ .

Как следствие, из диагональных элементов матрицы  $D$  можно извлекать квадратные корни. Если обозначить получающуюся при этом диагональную матрицу через  $\sqrt{D} := \text{diag}\{\sqrt{d_1}, \sqrt{d_2}, \dots, \sqrt{d_n}\}$ , то окончательно можем взять  $C = L\sqrt{D}$ . Это представление для множителя

Холесского в действительности единственны, так как по  $A$  при сделанных нами предположениях единственным образом определяется нижняя треугольная матрица  $L$ , а матричные преобразования, приведшие к формуле (3.97) и её следствиям, обратимы и также дают однозначно определённый результат. ■

### 3.6и Метод Холесского

Основной результат предшествующего раздела мотивирует прямой метод решения систем линейных уравнений, который аналогичен методу (3.88) на основе LU-разложения. Именно, если найдено разложение Холесского для матрицы  $A$ , то решение системы  $Ax = b$ , равносильной  $CC^\top x = b$ , сводится к решению двух треугольных систем линейных уравнений:

$$\begin{cases} Cy = b, \\ C^\top x = y. \end{cases} \quad (3.98)$$

Для решения первой системы применяем алгоритм прямой подстановки (3.81), а для решения второй системы — обратную подстановку (3.82).

Но как практически найти разложение Холесского? Теорема 3.6.4 носит конструктивный характер и в принципе может служить основой для соответствующего алгоритма. Недостатком этого подхода является существенная опора на LU-разложение матрицы, и потому желательно иметь более прямой способ нахождения разложения Холесского.

Выпишем равенство  $A = CC^\top$ , определяющее множитель Холесского, в развёрнутой форме с учётом симметричности  $A$ :

$$\begin{aligned} & \begin{pmatrix} a_{11} & & & \\ a_{21} & a_{22} & & \\ \vdots & \vdots & \ddots & \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} = \\ & = \begin{pmatrix} c_{11} & & & \\ c_{21} & c_{22} & & \\ \vdots & \vdots & \ddots & \\ c_{n1} & c_{n2} & \dots & c_{nn} \end{pmatrix} \cdot \begin{pmatrix} c_{11} & c_{21} & \cdots & c_{n1} \\ c_{21} & c_{22} & \cdots & c_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & & & c_{nn} \end{pmatrix}, \end{aligned} \quad (3.99)$$

где символом « $\square$ » обозначены симметричные относительно главной диагонали элементы матрицы, которые несущественны в последующих рассмотрениях. Можно рассматривать это равенство как систему уравнений относительно неизвестных переменных  $c_{11}, c_{21}, c_{22}, \dots, c_{nn}$  — элементов нижнего треугольника множителя Холесского. Всего их  $1 + 2 + \dots + n = \frac{1}{2}n(n + 1)$  штук. Для их определения имеем столько же соотношений, вытекающих в матричном равенстве (3.99) из выражений для элементов  $a_{ij}$ ,  $i \geq j$ , которые образуют диагональ и поддиагональный треугольник симметричной матрицы  $A = (a_{ij})$ .

В поэлементной форме система уравнений (3.99) имеет вид, определяемый правилом умножения матриц и симметричностью  $A$ :

$$\sum_{k=1}^j c_{ik}c_{jk} = a_{ij} \quad \text{при } j \leq i. \quad (3.100)$$

Выписанные соотношения образуют, фактически, двумерный массив, в котором уравнения имеют двойные индексы —  $i$  и  $j$ , но их можно линейно упорядочить таким образом, что система уравнений (3.100) получит специальный вид, очень напоминающий треугольные СЛАУ. Далее эта система может быть решена с помощью процесса, сходного с прямой подстановкой для треугольных СЛАУ (см. § 3.66).

$$C = \begin{pmatrix} \downarrow & & & & & 0 \\ \downarrow & \downarrow & & & & \\ \downarrow & \downarrow & \ddots & & & \\ \vdots & \vdots & \ddots & & \downarrow & \\ \curvearrowleft & \curvearrowleft & \cdots & \curvearrowleft & \curvearrowleft & \times \end{pmatrix}$$

Рис. 3.20. Схема определения элементов треугольного множителя при разложении Холесского

В самом деле, если выписывать выражения для элементов  $a_{ij}$  по столбцам матрицы  $A$ , начиная в каждом столбце с диагонального элемента  $a_{jj}$  и идя сверху вниз до  $a_{jn}$  (рис. 3.20), то все уравнения из (3.100) разбиваются на  $n$  следующих групп, которые удобно занумеро-

вать столбцовым индексом  $j = 1, 2 \dots, n$ :

$$\text{для } j = 1 \quad \begin{cases} c_{11}^2 = a_{11}, \\ c_{i1}c_{11} = a_{i1}, \quad i = 2, 3, \dots, n, \end{cases}$$

$$\text{для } j = 2 \quad \begin{cases} c_{21}^2 + c_{22}^2 = a_{22}, \\ c_{i1}c_{21} + c_{i2}c_{22} = a_{i2}, \quad i = 3, 4, \dots, n, \end{cases}$$

$$\text{для } j = 3 \quad \begin{cases} c_{31}^2 + c_{32}^2 + c_{33}^2 = a_{33}, \\ c_{i1}c_{31} + c_{i2}c_{32} + c_{i3}c_{33} = a_{i3}, \quad i = 4, 5, \dots, n, \end{cases}$$

...      ...      .

В краткой записи получающаяся система может быть записана следующим образом:

$$\left\{ \begin{array}{l} \left\{ \begin{array}{l} c_{j1}^2 + c_{j2}^2 + \dots + c_{j,j-1}^2 + c_{jj}^2 = a_{jj}, \\ c_{i1}c_{j1} + c_{i2}c_{j2} + \dots + c_{ij}c_{jj} = a_{ij}, \quad i = j+1, \dots, n, \end{array} \right. \\ j = 1, 2, \dots, n, \end{array} \right. \quad (3.101)$$

где считается, что  $c_{ji} = 0$  при  $j < i$ .

Получается, что в уравнениях из (3.101) для  $j$ -го столбца множители Холлесского присутствуют все элементы  $j$ -го и предшествующих столбцов. Если последовательно рассматривать группы уравнений в порядке возрастания номера  $j$ , то реально неизвестными к моменту обработки  $j$ -го столбца (т. е. решения  $j$ -й группы уравнений) являются только  $(n - j + 1)$  элементов  $c_{ij}$  именно этого  $j$ -го столбца, которые к тому же выражаются несложным образом через известные элементы и друг через друга.

В целом выписанная система уравнений (3.101) действительно имеет очень специальный вид, пользуясь которым можно находить элементы  $c_{ij}$  матрицы  $C$  последовательно друг за другом по столбцам в

порядке, который наглядно изображён на рис. 3.20. Более точно,

$$\text{при } j = 1 \quad \begin{cases} c_{11} = \sqrt{a_{11}}, \\ c_{i1} = a_{i1}/c_{11}, \quad i = 2, 3, \dots, n, \end{cases}$$

$$\text{при } j = 2 \quad \begin{cases} c_{22} = \sqrt{a_{22} - c_{11}^2}, \\ c_{i2} = (a_{i2} - c_{i1}c_{11})/c_{22}, \quad i = 3, 4, \dots, n, \end{cases}$$

$$\text{при } j = 3 \quad \begin{cases} c_{33} = \sqrt{a_{33} - c_{11}^2 - c_{22}^2}, \\ c_{i3} = (a_{i3} - c_{i1}c_{11} - c_{i2}c_{22})/c_{33}, \quad i = 4, 5, \dots, n, \end{cases}$$

и так далее для остальных  $j$ . Псевдокод этого процесса приведён в табл. 3.4, где считается, что если нижний предел суммирования пре-восходит верхний, то сумма «пуста» и суммирование не выполняется.

Таблица 3.4. Алгоритм разложения Холесского  
(прямой ход метода Холесского)

<pre> DO FOR <math>j = 1</math> TO <math>n</math>     <math>c_{jj} \leftarrow \sqrt{a_{jj} - \sum_{k=1}^{j-1} c_{jk}^2}</math>     DO FOR <math>i = j + 1</math> TO <math>n</math>         <math>c_{ij} \leftarrow \left( a_{ij} - \sum_{k=1}^{j-1} c_{ik}c_{jk} \right) / c_{jj}</math>     END DO END DO </pre>	(3.102)
---	---------

Если  $A$  — симметричная положительно определённая матрица, то в силу теоремы о разложении Холесского (теорема 3.6.4) система уравнений (3.101) обязана иметь решение, и наш алгоритм успешно прорабатывает до конца, находя его. Если же матрица  $A$  не является положи-

тельно определённой, то алгоритм (3.102) аварийно прекращает работу при попытке извлечь корень из отрицательного числа либо разделить на нуль. Вообще, запуск алгоритма (3.102) — это самый экономичный способ проверки положительной определённости симметричной матрицы.

Способ решения систем линейных алгебраических уравнений с симметричными положительно определёнными матрицами, который основан на нахождении их разложения Холесского и использует алгоритм (3.102) и далее соотношения (3.98), называют *методом Холесского*. Он был предложен в 1910 году А.-Л. Холесским в неопубликованной рукописи, которая тем не менее сделалась широко известной во французской геодезической службе, где решались такие системы уравнений. Позднее метод неоднократно переоткрывался, и потому иногда в связи с ним используются также термины «метод квадратного корня», «метод квадратных корней» или даже другие имена, данные его позднейшими авторами.

Метод Холесского можно рассматривать как специальную модификацию метода исключения Гаусса. Нетрудно показать, что он имеет ту же асимптотическую трудоёмкость  $O(n^3)$ , но с константой, которая в два раза меньше константы для метода Гаусса [40].

Замечательное свойство метода Холесского состоит в том, что обусловленность множителей Холесского, вообще говоря, является лучшей, чем у матрицы исходной СЛАУ: она равна корню квадратному из обусловленности матрицы системы. Это следует из предложения 3.2.4, применённого к множителям Холесского, и самого разложения Холесского. Иными словами, в отличие от обычного метода Гаусса, треугольные системы линейных уравнений из (3.98), к решению которых сводится задача, менее чувствительны к погрешностям, чем исходная линейная система. В следующем разделе мы увидим, что подобную ситуацию следует рассматривать как весьма нетипичную.

Если при реализации метода Холесского использовать комплексную арифметику, то извлечение квадратного корня можно выполнять всегда, и потому такая модификация применима к СЛАУ с симметричными неособенными матрицами, которые не являются положительно определёнными. При этом множители Холесского становятся комплексными треугольными матрицами.

Другой способ распространения метода Холесского на системы с произвольными симметричными матрицами состоит в том, чтобы ограничиться разложением (3.97), которое называется  *$LDL^\top$ -разложением*

матрицы. Если исходная матрица не является положительно определённой, то диагональные элементы в матричном множителе  $D$  могут быть отрицательными. Но  $LDL^\top$ -разложение столь же удобно для решения систем линейных алгебраических уравнений, как и рассмотренные ранее треугольные разложения. Детали этих построений читатель может найти в [11, 15, 46, 80].

Отметим также, что существует возможность другой организации вычислений при решении системы уравнений (3.100), когда неизвестные элементы  $c_{11}, c_{21}, c_{22}, \dots, c_{nn}$  последовательно находятся по строкам множителя Холесского, а не по столбцам, как в (3.102). Этот алгоритм называется *схемой окаймления* [15], и он по своим свойствам примерно эквивалентен рассмотренному выше алгоритму (3.102). Реализации метода Холесского и его модификаций присутствуют во всех библиотеках программ вычислительной линейной алгебры и большинстве систем компьютерной математики.

## 3.7 Прямые методы на основе ортогональных преобразований

### 3.7а Число обусловленности и матричные преобразования

Пусть  $A$  и  $B$  — неособенные квадратные матрицы, и матрица  $A$  умножается на матрицу  $B$ . Как связано число обусловленности произведения  $AB$  с числами обусловленности сомножителей  $A$  и  $B$ ?

Справедливы соотношения

$$\begin{aligned}\|AB\| &\leq \|A\| \|B\|, \\ \|(AB)^{-1}\| &= \|B^{-1}A^{-1}\| \leq \|A^{-1}\| \|B^{-1}\|,\end{aligned}$$

и поэтому

$$\operatorname{cond}(AB) = \|(AB)^{-1}\| \|AB\| \leq \operatorname{cond} A \cdot \operatorname{cond} B. \quad (3.103)$$

С другой стороны, если  $C = AB$ , то  $A = CB^{-1}$ , и в силу доказанного неравенства

$$\operatorname{cond}(A) \leq \operatorname{cond}(C) \cdot \operatorname{cond}(B^{-1}) = \operatorname{cond}(AB) \cdot \operatorname{cond}(B),$$

коль скоро  $\text{cond}(B^{-1}) = \text{cond}(B)$ . Поэтому

$$\text{cond}(AB) \geq \text{cond}(A)/\text{cond}(B).$$

Аналогичным образом из  $B = CA^{-1}$  следует

$$\text{cond}(AB) \geq \text{cond}(B)/\text{cond}(A).$$

Объединяя полученные неравенства, в целом получаем оценку

$$\text{cond}(AB) \geq \max \left\{ \frac{\text{cond}(A)}{\text{cond}(B)}, \frac{\text{cond}(B)}{\text{cond}(A)} \right\}. \quad (3.104)$$

Ясно, что её правая часть не меньше 1.

Неравенства (3.103)–(3.104) кажутся грубыми, но они достижимы. В самом деле, пусть  $A$  — неособенная симметричная матрица с собственными значениями  $\lambda_1, \lambda_2, \dots$ , так что её спектральное число обусловленности равно (стр. 413)

$$\text{cond}_2(A) = \frac{\max_i |\lambda_i(A)|}{\min_i |\lambda_i(A)|}.$$

У матрицы  $A^2$  собственные векторы, очевидно, совпадают с собственными векторами матрицы  $A$ , а собственные значения равны  $\lambda_1^2, \lambda_2^2$  и т. д. Поэтому числом обусловленности матрицы  $A^2$  является

$$\text{cond}_2(A^2) = \frac{\max_i (\lambda_i(A))^2}{\min_i (\lambda_i(A))^2} = \frac{\max_i |\lambda_i(A)|^2}{\min_i |\lambda_i(A)|^2} = \left( \frac{\max_i |\lambda_i(A)|}{\min_i |\lambda_i(A)|} \right)^2,$$

а в верхней оценке (3.103) получаем равенство. Совершенно сходным образом можно показать, что для спектрального числа обусловленности оценка (3.103) достигается также на произведениях вида  $A^T A$ .

Нижняя оценка (3.104) достигается, к примеру, при  $B = A^{-1}$  для чисел обусловленности, соответствующих подчинёнными матричными нормам.

Практически наиболее важной является верхняя оценка (3.103), и она показывает, что при преобразованиях и разложениях матриц число обусловленности может существенно расти. Рассмотрим, к примеру, решение системы линейных алгебраических уравнений  $Ax = b$  методом Гаусса в его матричной интерпретации (см. § 3.6г). Обнуление поддиагональных элементов первого столбца матрицы  $A$  — это умножение

исходной СЛАУ слева на матрицу  $E_1$ , имеющую вид (3.85), так что мы получаем систему

$$(E_1 A) x = E_1 b \quad (3.105)$$

с матрицей  $E_1 A$ , число обусловленности которой оценивается как

$$\operatorname{cond}(E_1 A) \leq \operatorname{cond}(E_1) \operatorname{cond}(A).$$

Перестановка строк или столбцов матрицы, выполняемая для поиска ведущего элемента, может незначительно изменить эту оценку в сторону увеличения, так как матрицы перестановки ортогональны и имеют небольшие числа обусловленности. Далее мы обнуляем поддиагональные элементы второго, третьего и т. д. столбцов матрицы системы (3.105), умножая её слева на матрицы  $E_2, E_3, \dots, E_{n-1}$  вида (3.86). В результате получаем верхнюю треугольную систему линейных уравнений

$$Ux = y,$$

в которой  $U = E_{n-1} \dots E_2 E_1 A$ ,  $y = E_{n-1} \dots E_2 E_1 b$ , а число обусловленности матрицы  $U$  оценивается сверху как

$$\operatorname{cond}(U) \leq \operatorname{cond}(A) \cdot \operatorname{cond}(E_1) \cdot \operatorname{cond}(E_2) \cdot \dots \cdot \operatorname{cond}(E_{n-1}). \quad (3.106)$$

Если  $E_j$  отлична от единичной матрицы, то  $\operatorname{cond}(E_j) > 1$ , причём правая и левая части неравенства (3.106) могут отличаться не очень сильно, несмотря на специальный вид матриц  $E_j$  (см. примеры ниже). Как следствие, обусловленность матриц, в которые матрица  $A$  исходной СЛАУ преобразуется на промежуточных шагах прямого хода метода Гаусса, а также обусловленность итоговой верхней треугольной матрицы  $U$  могут быть существенно хуже, чем у матрицы  $A$ .

**Пример 3.7.1** Предположим, что в  $5 \times 5$ -системе линейных алгебраических уравнений первый столбец матрицы коэффициентов имеет вид  $(1, 2, 3, 4, 5)^\top$ . Тогда обнуление поддиагональных элементов равносильно умножению слева на матрицу

$$E_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 & 0 \\ -3 & 0 & 1 & 0 & 0 \\ -4 & 0 & 0 & 1 & 0 \\ -5 & 0 & 0 & 0 & 1 \end{pmatrix}$$

(см. подробности в § 3.6г). Нетрудно проверить, что  $\text{cond}_2(E_1) = 55.98$ .

■

**Пример 3.7.2** Для  $2 \times 2$ -матрицы (3.20)

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

число обусловленности равно  $\text{cond}_2(A) = 14.93$ . Выполнение для неё преобразований прямого хода метода Гаусса приводит к матрице

$$\tilde{A} = \begin{pmatrix} 1 & 2 \\ 0 & -2 \end{pmatrix},$$

число обусловленности которой  $\text{cond}_2(\tilde{A}) = 4.27$ , т. е. уменьшается.

С другой стороны, для матрицы (3.21)

$$B = \begin{pmatrix} 1 & 2 \\ -3 & 4 \end{pmatrix},$$

число обусловленности  $\text{cond}_2(B) = 2.62$ . Преобразования метода Гаусса превращают её в матрицу

$$\tilde{B} = \begin{pmatrix} 1 & 2 \\ 0 & 10 \end{pmatrix},$$

для которой число обусловленности уже равно  $\text{cond}_2(\tilde{B}) = 10.4$ , т. е. существенно возрастает.

Аналогичные изменения претерпевают числа обусловленности матриц  $A$  и  $B$  относительно других норм. Числовые данные этого и предыдущего примеров читатель может воспроизвести с помощью систем компьютерной математики, таких как Scilab, MATLAB, Octave и им аналогичных. Все они имеют встроенную функцию `cond` для расчёта чисел обусловленности матрицы относительно различных норм.

■

Фактически, ухудшение обусловленности и, как следствие, всё большая чувствительность решения к погрешностям в данных — это плата за приведение матрицы (и всей СЛАУ) к удобному для решения виду и простоту алгоритма приведения. Можно ли уменьшить эту плату? И если да, то как?

Хорошей идеей является привлечение для матричных преобразований ортогональных матриц, которые имеют наименьшую возможную обусловленность в спектральной норме (и небольшие числа обусловленности в других нормах). Умножение на такие матрицы, по крайней мере, не будет ухудшать обусловленность получающихся систем линейных уравнений и устойчивость их решений к погрешностям вычислений. Единичную обусловленность относительно спектральной нормы имеют также матрицы, пропорциональные ортогональным, но именно ортогональные матрицы наиболее предпочтительны для преобразований векторно-матричных уравнений потому, что они не увеличивают евклидову норму невязок и погрешностей приближённых решений.

По-видимому, с точки зрения устойчивости наилучшим инструментом численного решения систем линейных алгебраических уравнений на цифровых ЭВМ с конечной точностью представления данных является сингулярное разложение матрицы системы. Соответствующая технология, при которой двумя ортогональными преобразованиями матрица СЛАУ приводится к диагональному виду, основана на сингулярном разложении матрицы — задача более сложная и трудоёмкая, чем рассматриваемые прямые методы решения СЛАУ.

### 3.7б Ортогональные преобразования и матричные вычисления

Более пристальное рассмотрение ортогональных матриц, которое мотивируется выводами предшествующего раздела, приводит к мысли о том, что они обладают важными или даже уникальными свойствами, которые позволяют успешно применять их для решения ряда задач вычислительной линейной алгебры.

С геометрической точки зрения преобразования пространства  $\mathbb{R}^n$ , выполняемые с помощью ортогональных матриц, являются обобщениями поворотов и отражений. Они сохраняют длины отрезков и векторов (евклидовые нормы), углы между прямыми и т. п., что следует из равенства

$$\|Qx\|_2 = \|x\|_2 \quad \text{для любой ортогональной матрицы } Q.$$

Как следствие, для матричных вычислений ортогональные матрицы являются очень «дружественными», поскольку они не увеличивают по-

грешности, вносимые в реальные вычислительные процессы округлениями, неточностью данных и прочими источниками.

Более того, для систем линейных алгебраических уравнений ортогональные преобразования сохраняют евклидову норму невязки приближённого решения. Если от системы уравнений  $Ax = b$  мы приходим в процессе преобразований к системе  $QAx = Qb$  и матрица  $Q$  ортогональна, то для любого вектора  $\tilde{x} \in \mathbb{R}^n$  справедливо

$$\|QA\tilde{x} - Qb\|_2 = \|Q(A\tilde{x} - b)\|_2 = \|A\tilde{x} - b\|_2.$$

Поэтому псевдорешения относительно евклидовой нормы для систем линейных уравнений  $Ax = b$  и  $QAx = Qb$  одинаковы.

Отмеченное свойство открывает широкие возможности для применения ортогональных преобразований при решении линейной задачи наименьших квадратов и нахождении псевдорешений систем линейных алгебраических уравнений относительно евклидовой нормы. Именно, ортогональными преобразованиями можно приводить переопределённые системы линейных алгебраических уравнений к правой (верхней) трапециевидной форме, для которой псевдорешения легко находятся алгоритмом обратной подстановки (см. § 3.6б).

Есть ли возможность распространить эти результаты и технологии на другие матричные нормы и псевдорешения относительно других норм? К сожалению, нет. Ортогональные матрицы являются в некотором роде уникальными.

Для изложения дальнейших результатов напомним

**Определение 3.7.1** Пусть  $X$  — метрическое пространство с расстоянием  $\text{dist}$ . Отображение  $\mathcal{A} : X \rightarrow X$  называется изометрическим относительно расстояния  $\text{dist}$  или просто изометрией, если

$$\text{dist}(\mathcal{A}x, \mathcal{A}y) = \text{dist}(x, y)$$

для любых  $x, y \in X$ , т. е. если отображение  $\mathcal{A}$  сохраняет неизменным расстояние между любыми двумя точками.

Если  $X$  линейное нормированное пространство, на котором расстояние задаётся нормой с помощью стандартной конструкции (3.30), т. е. как

$$\text{dist}(a, b) = \|a - b\|,$$

то линейное изометрическое отображение  $\mathcal{A}$  на  $X$  можно охарактеризовать проще, условием

$$\|\mathcal{A}x\| = \|x\| \quad \text{для любого } x \in X.$$

В частности, изометрическими относительно евклидовой нормы являются ортогональные линейные преобразования.

При эквивалентных преобразованиях уравнений, неравенств и их систем изометрические преобразования особенно цепны тем, что, аналогично ортогональным, не увеличивают погрешности. В более общей задаче поиска псевдорешений уравнений и систем уравнений изометрические преобразования и только они являются эквивалентными преобразованиями, сохраняющими псевдорешения, так как оставляют неизменной норму невязки.

В вычислительных методах линейной алгебры особенно важны линейные преобразования арифметических пространств  $\mathbb{R}^n$  и  $\mathbb{C}^n$ , задаваемые умножением на различные матрицы. По этой причине особую значимость приобретает

**Теорема 3.7.1** [43] Для любого  $p \neq 2$  множество матриц, задающих линейное изометрическое относительно  $p$ -нормы преобразование пространства  $\mathbb{R}^n$ , совпадает с множеством матриц перестановки.

**Доказательство** опускается.

Но одни только матрицы перестановки не могут выполнять полноценные преобразования, приводящие к матрицам необходимой специальной структуры в задачах вычислительной линейной алгебры. Как следствие, при  $p \neq 2$  не существуют прямые численные методы для нахождения псевдорешений систем линейных алгебраических уравнений относительно  $p$ -норм. Соответствующие численные методы обязаны быть итерационными, что, впрочем, не является каким-то существенным их недостатком. Они интенсивно разрабатываются в вычислительной оптимизации [98].

### 3.7в QR-разложение матриц

**Определение 3.7.2** Для матрицы  $A$  представление  $A = QR$  в виде произведения ортогональной матрицы  $Q$  и правой треугольной матрицы  $R$  называется QR-разложением.

По поводу этого определения следует пояснить, что правая треугольная матрица — это то же самое, что верхняя треугольная матрица, которую мы условились обозначать  $U$ . Другая терминология обусловлена здесь историческими причинами, и частичное её оправдание состоит в том, что QR-разложение матрицы действительно «совсем другое», нежели LU-разложение. Впрочем, в математической литературе можно встретить тексты, где LU-разложение матрицы называется «LR-разложением» (от английских слов left-right), т. е. разложением на «левую и правую треугольные матрицы».

QR-разложение матриц определяют также для общих прямоугольных матриц, не обязательно квадратных. Если  $A$  — это  $m \times n$ -матрица, то представление  $A = QR$  может трактоваться как произведение ортогональной  $m \times m$ -матрицы  $Q$  на трапециевидную  $m \times n$ -матрицу  $R$  или же как произведение  $m \times n$ -матрицы  $Q$  с ортогональными строками (столбцами) на правую треугольную  $n \times n$ -матрицу  $R$ . На практике встречаются оба вида разложений.

**Теорема 3.7.2** QR-разложение существует для любой матрицы.

Существует несколько способов доказательства этого результата, и почти все они имеют конструктивный характер, давая начало различным технологиям разложения матрицы. Сначала мы приведём теоретическое доказательство для важного частного случая квадратных матриц, а остальные будут изложены в соответствующих местах курса.

**Доказательство.** Если  $A$  — неособенная матрица, то в доказательстве теоремы 3.6.4 было показано, что  $A^\top A$  — симметричная положительно определённая матрица. В силу того же результата существует её разложение Холлесского

$$A^\top A = R^\top R,$$

где  $R$  — правая (верхняя) треугольная матрица. При этом  $R$ , очевидно, неособенна. Тогда матрица  $Q := AR^{-1}$  ортогональна, поскольку

$$\begin{aligned} Q^\top Q &= (AR^{-1})^\top AR^{-1} = (R^{-1})^\top A^\top A R^{-1} = \\ &= (R^{-1})^\top (R^\top R) R^{-1} = ((R^{-1})^\top R^\top)(RR^{-1}) = I. \end{aligned}$$

Следовательно, в целом  $A = QR$ , где определённые выше сомножители  $Q$  и  $R$  удовлетворяют условиям теоремы.

Рассмотрим теперь случай особенной матрицы  $A$ . Известно, что любую особенную матрицу можно приблизить последовательностью неособенных. Например, это можно сделать с помощью матриц  $A_k = A + \frac{1}{k}I$ , начиная с достаточно больших натуральных номеров  $k$ . При этом собственные значения матриц  $A_k$  суть  $\lambda(A_k) = \lambda(A) + \frac{1}{k}$ , и если величина  $\frac{1}{k}$  меньше расстояния от нуля до ближайшего ненулевого собственного значения матрицы  $A$ , то  $A_k$  неособенна.

В силу уже доказанного для всех матриц из последовательности  $\{A_k\}$  существуют QR-разложения:

$$A_k = Q_k R_k,$$

где все  $Q_k$  ортогональны, а  $R_k$  — правые треугольные матрицы. В качестве ортогонального разложения для  $A$  можно было бы взять пределы матриц  $Q_k$  и  $R_k$ , если таковые существуют. Но сходятся ли куданибудь последовательности этих матриц при  $k \rightarrow \infty$ , когда  $A_k \rightarrow A$ ? Ответ на это вопрос может быть отрицательным, а потому приходится действовать более тонко, выделяя из  $\{A_k\}$  подходящую подпоследовательность.

Множество ортогональных матриц компактно, поскольку является замкнутым (прообраз единичной матрицы  $I$  при непрерывном отображении  $X \mapsto X^\top X$ ) и ограничено ( $\|X\|_2 \leq 1$ ). Поэтому из последовательности ортогональных матриц  $\{Q_k\}$  можно выбрать сходящуюся подпоследовательность  $\{Q_{k_l}\}_{l=1}^\infty$ . Ей соответствуют подпоследовательности  $\{A_{k_l}\}$  и  $\{R_{k_l}\}$ , причём первая из них также сходится, как подпоследовательность сходящейся последовательности  $\{A_k\}$ .

Обозначим  $Q := \lim_{l \rightarrow \infty} Q_{k_l}$ , и это тоже ортогональная матрица. Тогда

$$\lim_{l \rightarrow \infty} (Q_{k_l}^\top A_{k_l}) = \lim_{l \rightarrow \infty} Q_{k_l}^\top \cdot \lim_{l \rightarrow \infty} A_{k_l} = Q^\top A = R$$

— правой треугольной матрице, поскольку все  $Q_{k_l}^\top A_{k_l}$  были правыми треугольными матрицами  $R_{k_l}$ . Таким образом, в целом снова  $A = QR$  с ортогональной  $Q$  и правой треугольной  $R$ , как и требовалось. ■

Если для системы линейных алгебраических уравнений  $Ax = b$  известно QR-разложение матрицы  $A$ , то она равносильна

$$(QR)x = b,$$

и её решение сводится к решению треугольной системы

$$Rx = Q^\top b. \quad (3.107)$$

Ниже в § 3.16 и § 3.18г мы встретимся и с другими важными применениями QR-разложения матриц — при численном решении линейной задачи наименьших квадратов и проблемы собственных значений.

Как видим, для неособенных матриц доказательство теоремы 3.7.2 конструктивно и опирается на разложение Холлесского матрицы  $A^T A$ . В принципе, нахождение намеченным способом QR-разложения — это путь возможный, но чреватый многими опасностями. Главная из них состоит в том, что приближённый характер вычислений на цифровых ЭВМ будет приводить к тому, что ортогональная матрица в получающемся QR-разложении не вполне ортогональна. На практике основным инструментом получения QR-разложения является техника, использующая так называемые матрицы вращения и матрицы отражения, описаннию которых посвящены следующие разделы книги.

### 3.7г Матрицы вращения и метод вращений

Пусть даны натуральные числа  $k, l$ , не превосходящие  $n$ , т. е. размерности пространства  $\mathbb{R}^n$ , и пусть задано значение угла  $\theta$ ,  $0 \leq \theta < 2\pi$ . *Матрицей вращения* называется  $n \times n$ -матрица  $G(k, l, \theta)$  вида

$$\begin{array}{c} k\text{-я строка} \\ l\text{-я строка} \end{array} \left( \begin{array}{cccccc} 1 & & & & & & \\ & \ddots & & & & & \\ & & \cos \theta & \cdots & -\sin \theta & & \\ & & & 1 & & & \\ & & & & \ddots & \vdots & \\ & & & & & 1 & \\ & & & & & & \ddots \\ & & & & & & & 1 \end{array} \right), \quad (3.108)$$

где все не выписанные явно элементы вне главной диагонали равны нулю. Таким образом,  $G(k, l, \theta)$  — это матрица, которая отличается от единичной матрицы лишь элементами, находящимися в позициях  $(k, k)$ ,  $(k, l)$ ,  $(l, k)$  и  $(l, l)$ . Нетрудно проверить, что она ортогональна и её определитель равен единице, т. е.  $\det G(k, l, \theta) = 1$ .

Матрица

$$\begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \quad (3.109)$$

задаёт, как известно, вращение двумерной плоскости  $0x_1x_2$  на угол  $\theta$  вокруг начала координат.<sup>19</sup> Матрица  $G(k, l, \theta)$  тоже задаёт вращение пространства  $\mathbb{R}^n$  на угол  $\theta$  вокруг оси, проходящей через начало координат и ортогональной гиперплоскости  $0x_kx_l$ . Матрицы вращения  $G(k, l, \theta)$  называют также *матрицами Гивенса*. Мы будем иногда обозначать их посредством  $G(k, l)$ , когда конкретная величина угла  $\theta$  несущественна.

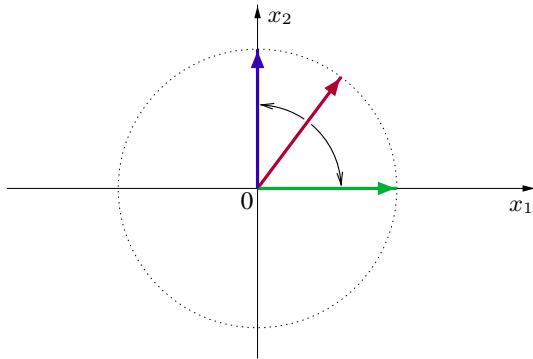


Рис. 3.21. Подходящим вращением можно занулить любую из компонент двумерного вектора

Если вектор  $a = (a_1, a_2)^\top$  — ненулевой, то, взяв

$$\cos \theta = \frac{a_1}{\|a\|_2}, \quad \sin \theta = \frac{-a_2}{\|a\|_2}, \quad \text{где } \|a\|_2 = \sqrt{a_1^2 + a_2^2},$$

мы можем с помощью матрицы двумерного вращения (3.109) занулить вторую компоненту этого вектора:

$$\begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \|a\|_2 \\ 0 \end{pmatrix}.$$

Аналогично первая компонента вектора  $a$  может быть занулена умножением на такую матрицу вращения (3.109), что

$$\cos \theta = \frac{a_2}{\|a\|_2}, \quad \sin \theta = \frac{a_1}{\|a\|_2}.$$

---

<sup>19</sup>Напомним, что положительным направлением вращения плоскости считается вращение «против часовой стрелки».

Пусть матрица  $A = (a_{ij})$  умножается слева на матрицу вращения  $G(k, l, \theta)$  и результатом является матрица  $\tilde{A}$ ,

$$\tilde{A} = (\tilde{a}_{ij}) := G(k, l, \theta) A.$$

Строки  $k$ -я и  $l$ -я в ней становятся линейными комбинациями строк с этими же номерами из  $A$ :

$$\begin{aligned}\tilde{a}_{kj} &\leftarrow a_{kj} \cos \theta - a_{lj} \sin \theta, \\ \tilde{a}_{lj} &\leftarrow a_{kj} \sin \theta + a_{lj} \cos \theta,\end{aligned}\quad j = 1, 2, \dots, n. \quad (3.110)$$

Остальные элементы матрицы  $\tilde{A}$  совпадают с элементами матрицы  $A$ . Из рассуждений предшествующего абзаца вытекает, что с помощью умножения на матрицу вращения со специально подобранным углом  $\theta$  можно занулить любой элемент  $k$ -й или  $l$ -й строк матрицы  $\tilde{A} = G(k, l, \theta) A$ .

Следовательно, любая квадратная матрица  $A$  может быть приведена к правому треугольному виду с помощью последовательности умножений слева на матрицы вращения. Более точно, мы можем один за другим занулить поддиагональные элементы первого столбца, потом второго, третьего и т. д., аналогично тому, как это делалось в прямом ходе метода Гаусса. При этом зануление поддиагональных элементов второго и следующих столбцов никак не испортит полученные ранее нулевые элементы предшествующих столбцов, так как линейное комбинирование нулей согласно формулам (3.110) даст снова нуль. Говоря формально, существует такой набор матриц вращения  $G(1, 2), G(1, 3), \dots, G(1, n), G(2, 3), \dots, G(n - 1, n)$ , что

$$G(n - 1, n) \cdots G(2, 3) G(1, n) \cdots G(1, 3) G(1, 2) A = R$$

является правой треугольной матрицей. Отсюда

$$A = G(1, 2)^\top G(1, 3)^\top \cdots G(1, n)^\top G(2, 3)^\top \cdots G(n - 1, n)^\top R,$$

т. е. получено QR-разложение матрицы  $A$ , поскольку произведение транспонированных матриц вращения также является ортогональной матрицей.

В одном важном моменте преобразования вращения всё таки отличаются от элементарных преобразований матрицы в прямом ходе метода Гаусса (см. § 3.6в и 3.6г). В результате элементарных преобразований метода Гаусса изменяется лишь *одна* строка матрицы, к которой

мы прибавляем другую, умноженную на какое-то число. Но в преобразованиях вращения, как следует из формул (3.110), изменяются *две* строки матрицы. Каждая из них становится линейной комбинацией двух исходных строк. Это обстоятельство приводит к важным последствиям для различных матричных алгоритмов, особенно при решении задачи на собственные значения.

Использование преобразований вращения — конструктивный способ получения QR-разложения, и технически он более прост, чем излагаемый в следующем разделе способ, основанный на отражениях Хаусхолдера. При его реализации организовывать полноценные матрицы вращения  $G(k, l, \theta)$  и матричные умножения с ними, конечно, нецелесообразно, так как большинством элементов в  $G(k, l, \theta)$  являются нули. Результат умножения слева на матрицу вращения разумно находить путём перевычисления элементов всего двух строк по формулам (3.110).

*Метод вращений* — это прямой численный метод для нахождения решений или евклидовых псевдорешений (в смысле наименьших квадратов) систем линейных алгебраических уравнений, основанный на использовании QR-разложения матрицы системы с помощью ортогональных преобразований вращения. Для системы линейных уравнений  $Ax = b$  с помощью матриц вращения выполним, как описано выше, приведение матрицы  $A$  к правой (верхней) треугольной или трапецевидной форме, одновременно применяя те же преобразования к правой части  $b$ . Это прямой ход метода, на котором получается равносильная система линейных уравнений с треугольной или трапецевидной матрицей. Далее на обратном ходе она решается с помощью обратной подстановки (см. § 3.6б). В целом численный метод очень похож на метод Гаусса, но отличается от него лучшей устойчивостью и возможностью находить псевдорешения переопределённых систем уравнений относительно 2-нормы, т. е. решать линейную задачу наименьших квадратов.

Если матрица системы имеет полный ранг, то в правой треугольной или трапецевидной матрице, полученной на прямом ходе, все диагональные элементы должны быть ненулевыми. Поэтому обратный ход метода вращений выполним и решение будет получено всегда.

Но даже если матрица системы имеет неполный ранг, то модификацией алгоритма можно всё-таки добиться решения системы. В этом случае можно выполнить QR-разложение с выбором ведущего элемента, аналогично тому, как это делается в методе Гаусса. Нулевые диагональные элементы в матрице  $R$  помещаются тогда на последние пози-

ции, соответствующие слагаемые переходят в правую часть, и размер треугольной или трапециевидной СЛАУ, которую необходимо решить, соответственно, уменьшается.

Трудоёмкость метода вращений для квадратных  $n \times n$ -систем составляет  $O(n^3)$  операций, но константа, которая скрывается за « $O$ -большим», в три раза больше той, что присутствует в оценке для метода Гаусса и в полтора раза больше, чем у метода отражений Хаусхолдера, описываемого в следующих разделах.

### 3.7д Ортогональные матрицы отражения

**Определение 3.7.3** Для вектора  $u \in \mathbb{R}^n$  с единичной евклидовой нормой,  $\|u\|_2 = 1$ , матрица  $H = H(u) = I - 2uu^\top$  называется матрицей отражения или матрицей Хаусхолдера. Вектор  $u$  называется порождающим или вектором Хаусхолдера для матрицы отражения  $H(u)$ .

Отметим, что в этом определении произведение  $uu^\top$  — одноранговая  $n \times n$ -матрица, так что если  $u = (u_1, u_2, \dots, u_n)^\top$ ,  $\|u\|_2 = 1$ , то порождаемая этим вектором матрица отражений выглядит следующим образом:

$$\begin{pmatrix} 1 - 2u_1^2 & -2u_1u_2 & -2u_1u_3 & \cdots & -2u_1u_n \\ -2u_2u_1 & 1 - 2u_2^2 & -2u_2u_3 & \cdots & -2u_2u_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -2u_nu_1 & -2u_nu_2 & -2u_nu_3 & \cdots & 1 - 2u_n^2 \end{pmatrix}.$$

**Предложение 3.7.1** Матрицы отражения являются симметричными ортогональными матрицами. Кроме того, для матрицы  $H(u)$

порождающий вектор  $u$  является собственным вектором, отвечающим собственному значению  $(-1)$ , т. е.  $H(u) \cdot u = -u$ ;

любой вектор  $v$ , ортогональный порождающему вектору  $u$ , является собственным вектором, отвечающим собственному значению  $1$ , т. е.  $H(u) \cdot v = v$ .

Определитель матрицы отражения равен  $-1$ , т. е.  $\det H(u) = -1$ .

**Доказательство** проводится непосредственной проверкой.

Симметричность матрицы  $H(u)$ :

$$\begin{aligned} H^\top &= (I - 2uu^\top)^\top = I^\top - (2uu^\top)^\top = \\ &= I - 2(u^\top)^\top u^\top = I - 2uu^\top = H. \end{aligned}$$

Ортогональность:

$$\begin{aligned} H^\top H &= (I - 2uu^\top)(I - 2uu^\top) = \\ &= I - 2uu^\top - 2uu^\top + 4uu^\top uu^\top = \\ &= I - 4uu^\top + 4u(u^\top u)u^\top = I, \quad \text{так как } u^\top u = \|u\|_2^2 = 1. \end{aligned}$$

Собственные векторы и собственные значения:

$$\begin{aligned} H(u) \cdot u &= (I - 2uu^\top)u = u - 2u(u^\top u) = u - 2u = -u; \\ H(u) \cdot v &= (I - 2uu^\top)v = v - 2u(u^\top v) = v, \quad \text{если } u^\top v = 0. \end{aligned}$$

Последнее свойство матриц отражения следует из того, что определитель любой матрицы равен произведению её собственных значений. ■

Из свойств собственных векторов и собственных значений матриц отражения следует геометрическая интерпретация, которая мотивирует их название. Эти матрицы действительно осуществляют преобразование отражения относительно гиперплоскости, ортогональной порождающему вектору  $u$ . В самом деле, представим произвольный вектор  $x$  в виде  $\alpha u + v$ , где  $\alpha \in \mathbb{R}$ ,  $u$  — порождающий матрицу отражения вектор, а  $v$  — ему ортогональный (рис. 3.22). Тогда

$$H(u) \cdot x = H(u) \cdot (\alpha u + v) = -\alpha u + v,$$

т. е. в векторе, преобразованном матрицей  $H(u)$ , та компонента, которая ортогональна рассматриваемой гиперплоскости, сменила направление на противоположное. Это и соответствует отражению относительно неё.

**Предложение 3.7.2** Для любых двух ненулевых векторов в  $\mathbb{R}^n$  существует матрица отражения, которая переводит первый вектор в вектор, коллинеарный второму.

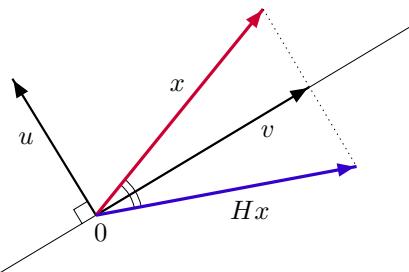


Рис. 3.22. Геометрическая интерпретация действия матрицы отражения

**Доказательство.** Обозначим данные в условии предложения векторы как  $x, y \in \mathbb{R}^n$ .

Если  $H$  — искомая матрица отражения и  $u$  — порождающий её вектор Хаусхольдера, то утверждение предложения требует равенства

$$Hx = x - 2(uu^\top)x = \gamma y \quad (3.111)$$

с некоторым коэффициентом  $\gamma \neq 0$ . Для его определения заметим, что ортогональная матрица  $H$  не изменяет евклидову норму векторов, так что  $\|Hx\|_2 = \|x\|_2$ . С другой стороны, взяв евклидову норму от обеих частей (3.111), получим  $\|Hx\|_2 = |\gamma| \|y\|_2$ . Сопоставляя оба равенства, будем иметь

$$\|x\|_2 = |\gamma| \|y\|_2, \quad \text{т. е. } \gamma = \pm \frac{\|x\|_2}{\|y\|_2}. \quad (3.112)$$

Из (3.111) следует

$$2u(u^\top x) = x - \gamma y. \quad (3.113)$$

Далее рассмотрим отдельно два случая — когда векторы  $x$  и  $y$  неколлинеарны и когда они коллинеарны друг другу.

В первом случае правая часть в (3.113) заведомо не равна нулю. Поэтому числовой множитель  $u^\top x$  в левой части этого равенства обязан быть ненулевым и можно заключить, что

$$u = \frac{1}{2u^\top x} (x - \gamma y),$$

т. е. вектор  $u$ , порождающий искомую матрицу отражения, коллинеарен вектору  $(x - \gamma y)$ . Более точно, с учётом (3.112) вектор Хаусхольде-

ра  $u$  должен быть коллинеарен вектору

$$\tilde{u} = x \pm \frac{\|x\|_2}{\|y\|_2} y, \quad (3.114)$$

где вместо « $\pm$ » выбран какой-то один определённый знак. Для окончательного нахождения  $u$  остаётся лишь применить нормировку:

$$u = \frac{\tilde{u}}{\|\tilde{u}\|_2},$$

и тогда  $H = I - 2uu^\top$  — искомая матрица отражения.

Обсудим случай, когда  $x$  коллинеарен  $y$ . При этом предшествующая конструкция частично теряет смысл, так как вектор  $\tilde{u} = x - \gamma y$  может занулиться при подходящем выборе множителя  $\gamma$ .

Но даже если  $x - \gamma y = 0$  для какого-то одного из значений  $\gamma$ , определяемых в (3.112), то для противоположного по знаку значения  $\gamma$  на-верняка  $x - \gamma y \neq 0$ . Более формально можно сказать, что конкретный знак у множителя  $\gamma = \pm \|x\|_2/\|y\|_2$  следует выбирать из условия максимизации нормы вектора  $(x - \gamma y)$ . Далее все рассуждения, следующие за формулой (3.113), остаются в силе и приводят к определению вектора Хаусхольдера.

Наконец, в случае коллинеарных векторов  $x$  и  $y$  мы можем просто указать явную формулу для вектора Хаусхольдера:

$$u = \frac{x}{\|x\|_2}.$$

При этом

$$u^\top x = \frac{x^\top x}{\|x\|_2} = \|x\|_2 \neq 0,$$

и для соответствующей матрицы отражения справедливо

$$Hx = x - 2(uu^\top)x = x - 2u(u^\top x) = x - 2\frac{x}{\|x\|_2}\|x\|_2 = -x.$$

Итак, вектор  $x$  снова переводится матрицей  $H$  в вектор, коллинеарный вектору  $y$ .<sup>20</sup> ■

---

<sup>20</sup>Интересно, что этот тонкий случай доказательства имеет скорее теоретическое значение, так как если вектор уже коллинеарен заданному, то на практике с ним, как правило, можно вообще ничего не делать.

В доказательстве предложения присутствует неоднозначность в выборе знака в выражении  $\tilde{u} = x \pm \|x\|_2 y$ , если  $x$  и  $y$  неколлинеарны. В действительности годится любой знак, и его конкретный выбор может определяться, как мы увидим, требованием устойчивости вычислительного алгоритма.

### 3.7e Метод отражений Хаусхолдера

*Метод Хаусхолдера* — это прямой численный метод для нахождения решений и псевдорешений систем линейных алгебраических уравнений с матрицами полного ранга, использующий матрицы отражения Хаусхолдера. Иногда его называют также *методом отражений*. В его основе лежит та же самая идея, что и в методе Гаусса: привести эквивалентными преобразованиями исходную систему к правой (верхней) треугольной или трапециевидной форме, а затем воспользоваться обратной подстановкой (табл. 3.2). Но теперь это приведение выполняется более глубокими, чем в методе Гаусса, преобразованиями матрицы — путём последовательного умножения на специально подобранные матрицы отражения.

**Предложение 3.7.3** Для любой матрицы  $A$  существует конечная последовательность  $H_1, H_2, \dots, H_s$ ,  $s \in \{n-1, n\}$ , состоящая из матриц отражения и, возможно, единичных матриц, таких что матрица

$$H_s H_{s-1} \cdots H_2 H_1 A = R$$

является правой треугольной или трапециевидной матрицей.

Раздельное упоминание матриц отражения и единичных матриц вызвано здесь тем, что единичная матрица не является матрицей отражения. Длина  $s$  конечной последовательности матриц отражения, очевидно, равна  $n - 1$  или  $n$  в зависимости от того, является матрица  $A$  квадратной или лежачей (тогда  $s = n - 1$ ) или стоячей (тогда  $s = n$ ).

Доказательство предложения конструктивно и для формального описания алгоритма, который фактически строится в нём, очень удобно применять систему обозначений матрично-векторных объектов, укоренившуюся в языках программирования Fortran, MATLAB, Scilab и им подобных. Согласно ей посредством  $A(p : q, r : s)$  обозначается *сечение* массива  $A$ , которое определяется как массив с тем же количеством измерений и элементами, которые стоят на пересечении строк с номерами

с  $p$  по  $q$  и столбцов с номерами с  $r$  по  $s$ . То есть, запись  $A(p : q, r : s)$  указывает в индексах матрицы  $A$  не отдельные значения, а целые диапазоны изменения индексов элементов, из которых образуется новая матрица, как подматрица исходной  $A$ .

**Доказательство.** Пусть  $A = (a_{ij})$ . Если хотя бы один из элементов  $a_{21}, a_{31}, \dots, a_{n1}$  не равен нулю, то, используя результат предложения 3.7.2, возьмём в качестве  $H_1$  матрицу отражения, которая переводит 1-й столбец  $A$  в вектор, коллинеарный  $(1, 0, \dots, 0)^\top$ . Иначе полагаем  $H_1 = I$ . Затем переходим ко второму шагу.

В результате выполнения первого шага матрица СЛАУ приводится, как и в методе Гаусса, к виду

$$\tilde{A} = \left( \begin{array}{c|ccccc} \times & \times & \times & \cdots & \times \\ \hline 0 & \times & \times & \cdots & \times \\ 0 & \times & \times & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \times & \times & \cdots & \times \end{array} \right),$$

где крестиками « $\times$ » обозначены элементы, которые, возможно, не равны нулю. Проделаем затем то же самое с матрицей  $\tilde{A}(2:n, 2:n)$ , обнулив у неё подходящим отражением поддиагональные элементы первого столбца, который является вторым во всей большой матрице  $\tilde{A}$ . И так далее до  $(n - 1)$ -го столбца.

Для формального описания алгоритма определим матрицу  $H_j = H_j(u)$ ,  $j = 2, 3, \dots, n - 1$ , как  $n \times n$ -матрицу отражения, порождаемую вектором Хаусхолдера  $u \in \mathbb{R}^n$ , который имеет нулевыми первые  $j - 1$  компонент и подобран так, чтобы  $H_j(u)$  обнуляла в матрице  $\tilde{A} = H_{j-1} \cdots H_2 H_1 A$  поддиагональные элементы  $j$ -го столбца, если среди них существуют ненулевые. Иначе, если в преобразуемой матрице  $\tilde{A} = (\tilde{a}_{ij})$  все элементы  $\tilde{a}_{j+1,j}, \tilde{a}_{j+2,j}, \dots, \tilde{a}_{nj}$  нулевые, то полагаем  $H_j = I$  — единичной  $n \times n$ -матрице.

Можно положить в блочной форме

$$H_j := \left( \begin{array}{c|c} \check{I} & 0 \\ \hline 0 & \check{H}_j \end{array} \right),$$

где в верхнем левом углу стоит единичная  $(j - 1) \times (j - 1)$ -матрица  $\check{I}$ ,

Таблица 3.5. QR-разложение матрицы  
с помощью отражений Хаусхолдера

```

DO FOR  $j = 1$  TO  $n - 1$ 
     $\check{I} \leftarrow$  единичная матрица размера  $(n - j + 1)$ ;
    IF ( вектор  $A((j+1) : n, j)$  ненулевой ) THEN
        вычислить вектор Хаусхолдера  $\check{u} \in \mathbb{R}^{n-j+1}$ ,
        порождающий отражение, которое переводит
        вектор  $A(j : n, j)$  в вектор, коллинеарный
        вектору  $(1, 0, \dots, 0)^\top$ ;
         $\check{H} \leftarrow \check{I} - 2\check{u}\check{u}^\top$ ;
    ELSE
         $\check{H} \leftarrow \check{I}$ ;
    END IF
     $A(j : n, j : n) \leftarrow \check{H} A(j : n, j : n)$ ;
END DO

```

а  $\tilde{H}_j$  — матрица размера  $(n - j + 1) \times (n - j + 1)$ , которая переводит вектор  $\tilde{A}(j : n, j)$  в  $(n - j + 1)$ -вектор, коллинеарный  $(1, 0, \dots, 0)^\top$ , т. е. обнуляет поддиагональные элементы  $j$ -го столбца в  $\tilde{A}$ . Если хотя бы один из элементов  $\tilde{a}_{j+1,j}, \tilde{a}_{j+2,j}, \dots, \tilde{a}_{nj}$  не равен нулю, то  $\tilde{H}_j$  — матрица отражения, способ построения которой описывается в предложении 3.7.2. Иначе, если  $(\tilde{a}_{j+1,j}, \tilde{a}_{j+2,j}, \dots, \tilde{a}_{nj})^\top = 0$ , то  $\tilde{H}_j$  — единичная  $(n - j + 1) \times (n - j + 1)$ -матрица. ■

Отметим, что из представления

$$H_s H_{s-1} \cdots H_2 H_1 A = R$$

вытекает равенство  $A = QR$  с ортогональной матрицей

$$Q = (H_s H_{s-1} \cdots H_2 H_1)^\top.$$

Таким образом, мы получаем QR-разложение матрицы  $A$ , т. е. пред-

ложения 3.7.2 и 3.7.3 дают в совокупности ещё одно конструктивное доказательство теоремы 3.7.2. Соответствующий псевдокод алгоритма для вычисления QR-разложения матрицы приведён в табл. 3.5 для случая квадратной или лежачей матрицы, когда  $s = n - 1$ .

Как следствие, исходная система уравнений  $Ax = b$  становится равносильной системе уравнений

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

с несложными решаемыми составными частями. При практической реализации этой идеи удобнее дополнить алгоритм табл. 3.5 инструкциями, которые задают преобразования вектора  $b$  правой части СЛАУ, что в совокупности даёт «прямой ход» метода отражений Хаусхолдера. Более точно, между последней и предпоследней строками псевдокода в табл. 3.5 нужно поместить присваивание результата умножения подвектора в  $b$  на  $\check{H}$ :

$$b(j : n) \leftarrow \check{H} b(j : n)$$

Тогда к моменту окончания QR-разложения у нас уже будет известен вектор правой части  $y = Q^\top b$  для системы уравнений  $Rx = y$ . Её решение или псевдорешение могут быть найдены с помощью обратной подстановки (3.82), так как матрица  $R$  — правая треугольная или трапециевидная.

Согласно предложению 3.7.2 вычисление вектора Хаусхолдера  $u$  в качестве первого шага требует нахождения из (3.114) вектора  $\tilde{u}$ , в котором имеется неоднозначность выбора знака второго слагаемого. При вычислениях на цифровых ЭВМ в стандартной арифметике с плавающей точкой имеет смысл брать

$$\tilde{u} = \begin{cases} A(j : n, j) + \|A(j : n, j)\|_2 e, & \text{если } a_{jj} \geq 0, \\ A(j : n, j) - \|A(j : n, j)\|_2 e, & \text{если } a_{jj} \leq 0, \end{cases} \quad (3.115)$$

где  $e = (1, 0, \dots, 0)^\top$  — вектор размерности  $n - j + 1$ . Тогда вычисление первого элемента  $a_{jj}$  в столбце  $A(j : n, j)$ , т. е. того единственного элемента из всего столбца, который останется ненулевым, не будет сопровождаться вычитанием чисел одного знака и, как следствие, возможной потерей точности.

Ещё одно соображение по практической реализации описанного алгоритма QR-разложения состоит в том, что в действительности даже не

нужно формировать матрицу отражения  $\check{H}$  в явном виде. Умножение на неё можно выполнить по экономичным формулам

$$\begin{aligned} (I - 2uu^\top) A(j : n, j : n) &= \\ &= A(j : n, j : n) - 2u(u^\top A(j : n, j : n)), \quad (3.116) \\ (I - 2uu^\top) b(j : n) &= b(j : n) - 2u(u^\top b(j : n)), \end{aligned}$$

в которых сама матрица  $\check{H}$  не фигурирует.

Нетрудно показать [45, 71, 73], что для квадратных  $n \times n$ -систем трудоёмкость метода отражений Хаусхолдера составляет  $O(n^3)$  операций. Но здесь константа, которая скрывается за «*О-большим*», в два раза больше той, что присутствует в оценке для метода Гаусса и в два раза меньше, чем для метода вращений.

**Пример 3.7.3** Решим с помощью метода Хаусхолдера задачу построения линейной функции наилучшего среднеквадратичного приближения к данным, которая была рассмотрена в примере 2.10.2. Она сводится к нахождению псевдорешения системы линейных алгебраических уравнений

$$\begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}.$$

Приведём эту систему к верхней трапециевидной форме с помощью отражений Хаусхолдера. В числовых результатах, приводимых ниже, удерживается по пять значащих цифр после десятичной точки, и читатель может повторить наши вычисления в любой системе компьютерной математики.

Для обнуления поддиагональных элементов первого столбца умножим слева обе части этой системы на  $3 \times 3$ -матрицу отражений, которая переводит первый столбец  $(1, 2, 3)^\top$  в вектор, коллинеарный  $(1, 0, 0)^\top$ . Используя предложение 3.7.2 и рецепт выбора знака (3.115), получим вектор Хаусхолдера  $(0.79601, 0.33575, 0.50363)^\top$ . Далее, формулы редуцированного умножения (3.116) дают систему

$$\begin{pmatrix} -3.74166 & -1.60357 \\ 0.00000 & -0.09817 \\ 0.00000 & -0.64725 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} -2.40535 \\ -0.43635 \\ -0.15453 \end{pmatrix}.$$

Теперь обнуляем поддиагональные элементы второго столбца. При этом мы, фактически, работаем не со всей выписанной выше системой, а с её активной  $2 \times 1$ -подсистемой

$$\begin{pmatrix} -0.09817 \\ -0.64725 \end{pmatrix} (\beta) = \begin{pmatrix} -0.43635 \\ -0.15453 \end{pmatrix}.$$

Мы должны умножить слева обе части этой системы уравнений на  $2 \times 2$ -матрицу отражений, которая переводит столбец  $(-0.09817, -0.64725)^\top$  в вектор, коллинеарный  $(1, 0)^\top$ . Используя предложение 3.7.2 и рецепт выбора знака (3.115), получим порождающий вектор Хаусхолдера  $(-0.75827, -0.65194)^\top$ . Далее, формулы редуцированного умножения (3.116) дают систему

$$\begin{pmatrix} 0.65465 \\ 0.00000 \end{pmatrix} (\beta) = \begin{pmatrix} 0.21822 \\ 0.40825 \end{pmatrix}.$$

В целом преобразования отражения Хаусхолдера приводят исходную линейную систему к правой трапециевидной форме

$$\begin{pmatrix} -3.74166 & -1.60357 \\ 0 & 0.65465 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} -2.40535 \\ 0.21822 \\ 0.40825 \end{pmatrix}. \quad (3.117)$$

Выполнив процесс обратной подстановки для треугольной  $2 \times 2$ -системы линейных уравнений,

$$\begin{pmatrix} -3.74166 & -1.60357 \\ 0 & 0.65465 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} -2.40535 \\ 0.21828 \end{pmatrix},$$

которая получается выделением квадратной подсистемы из системы (3.117) (см. подробности в § 3.66), найдём

$$\alpha = 0.5, \quad \beta = 0.33333.$$

Это решение совпадает с тем, которое мы нашли в примере 2.10.2 с помощью перехода к нормальной системе уравнений. Но количество операций, затрачиваемое в этом способе решения, меньше, а сам алгоритм более устойчив. ■

Для плотно заполненных матриц получение QR-разложения с помощью отражений примерно в полтора раза менее трудоёмко, чем с

помощью матриц вращения. Но зато вращения более предпочтительны для разреженных матриц в силу своей большей гибкости при занулении отдельных элементов.

Определённым недостатком метода Хаусхолдера и описанного ранее метода вращений в сравнении с методом Гаусса является привлечение неарифметической операции извлечения квадратного корня, которая приводит к иррациональностям. Это не позволяет точно (без округлений) реализовать соответствующие алгоритмы в поле рациональных чисел, к примеру, в программных системах так называемых «безошибочных вычислений» или языках программирования типа Ruby [102], которые могут оперировать рациональными дробями.

### 3.8 Процессы ортогонализации

*Ортогонализацией* называют процесс построения по заданному семейству векторов линейного пространства некоторого ортогонального семейства векторов, которое имеет ту же самую линейную оболочку. Ввиду удобства ортогональных систем и ортогональных базисов для представления решений разнообразных задач и, как следствие, их важности во многих приложениях (см., к примеру, § 2.11) огромное значение имеют и процессы ортогонализации.

Но помимо своего явного назначения процессы ортогонализации применяются также для других целей, главная из которых — выяснение линейной зависимости или независимости заданной системы векторов. Ортогональные векторы, как известно, всегда линейно независимы. Поэтому если в процессе ортогонализации получено семейство ортогональных векторов, которое меньше исходного набора векторов, то тогда некоторые его векторы являются линейными комбинациями оставшихся, т. е. исходный набор векторов линейно зависим. Такое исследование вектор-строк или вектор-столбцов числовой матрицы необходимо при определении её ранга. Конечно, для определения линейной зависимости/независимости векторов можно применять и другие методы, но ортогонализация и связанные с ней процедуры являются наиболее устойчивыми для этой цели.

Исторически первым процессом ортогонализации был алгоритм, который по традиции связывают с именами Й. Грама и Э. Шмидта.<sup>21</sup> По конечному семейству векторов  $\{v_1, v_2, \dots, v_n\}$  процесс Грама–Шмидта

---

<sup>21</sup> Иногда этот процесс называют «ортогонализацией Сонина–Шмидта».

строит ортогональный базис  $\{q_1, q_2, \dots\}$  линейной оболочки для  $\{v_1, v_2, \dots, v_n\}$ , выясняя по ходу своего исполнения, является это семейство линейно зависимым или независимым.

Возьмём в качестве первого вектора  $q_1$  конструируемого ортогонального базиса вектор  $v_1$ , первый из исходного семейства. Далее для построения  $q_2$  можно использовать  $v_2$  «как основу», но откорректировав его с учётом требования ортогональности к  $q_1$  и принадлежности линейной оболочки векторов  $q_1 = v_1$  и  $v_2$ . Естественно положить  $q_2 = v_2 - \alpha_{12}q_1$ , где коэффициент  $\alpha_{12}$  подлежит определению из условия ортогональности

$$\langle q_1, v_2 - \alpha_{12}q_1 \rangle = 0.$$

Отсюда

$$\alpha_{12} = \frac{\langle q_1, v_2 \rangle}{\langle q_1, q_1 \rangle}.$$

Далее аналогичным образом находится вектор  $q_3 = v_3 - \alpha_{13}q_1 - \alpha_{23}q_2$  и т. д.

В целом ортогонализация Грама–Шмидта выполняется в соответствии со следующими расчётными формулами:

$$q_1 \leftarrow v_1, \tag{3.118}$$

$$\begin{cases} \alpha_{kj} \leftarrow \frac{\langle q_k, v_j \rangle}{\langle q_k, q_k \rangle}, & k = 1, \dots, j-1, \\ q_j \leftarrow v_j - \sum_{k=1}^{j-1} \alpha_{kj} q_k, & j = 2, 3, \dots, n. \end{cases} \tag{3.119}$$

В случае, когда очередной вычисленный вектор  $q_j$  оказывается нулевым, ясно, что  $v_1, v_2, \dots, v_j$  линейно зависимы. Тогда процесс завершается, полученное ортогональное семейство векторов  $\{q_1, q_2, \dots, q_{j-1}\}$  выдаётся в качестве базиса линейной оболочки  $\text{span}\{v_1, v_2, \dots, v_n\}$ , а относительно исходных векторов  $\{v_1, v_2, \dots, v_n\}$  выводится сообщение об их линейной зависимости. Если же процесс Грама–Шмидта дорабатывает до своего естественного завершения, то исходные векторы линейно независимы и, как и прежде, выдаётся ортогональный базис для  $\text{span}\{v_1, v_2, \dots, v_n\}$ .

Получающиеся векторы ортогонального базиса, как правило, дополнительно нормируют сразу после их нахождения, добиваясь равенства  $\langle q_k, q_k \rangle = 1$ . Тогда в (3.119) упрощаются выражения для поправоч-

ных коэффициентов  $\alpha_{kj}$ . Псевдокод соответствующего варианта ортогонализации Грама–Шмидта приведён в табл. 3.6.

Таблица 3.6. Ортогонализация Грама–Шмидта  
семейства векторов  $\{v_1, v_2, \dots, v_n\}$

```

DO FOR  $j = 1$  TO  $n$ 
     $q_j \leftarrow v_j$  ;
    DO  $k = 1$  TO  $j - 1$ 
         $\alpha_{kj} \leftarrow \langle q_k, v_j \rangle$  ;
         $q_j \leftarrow q_j - \alpha_{kj} q_k$  ;
    END DO
     $\alpha_{jj} \leftarrow \|q_j\|_2$  ;
    IF ( $\alpha_{jj} = 0$ ) THEN
        STOP, сигнализируя « $v_j$  линейно зависит
        от векторов  $v_1, v_2, \dots, v_{j-1}$ »
    END IF
     $q_j \leftarrow q_j / \alpha_{jj}$  ;
END DO

```

Интересно дать матричное представление процесса Грама–Шмидта. Пусть векторы  $v_1, v_2, \dots, v_n$  заданы своими координатными представлениями в некотором базисе, и из вектор-столбцов этих координатных представлений организована матрица  $W$ :

$$W = (v_1, v_2, \dots, v_n).$$

В результате ортогонализации мы должны получить ортогональную матрицу, в которой первый столбец — это нормированный первый вектор  $v_1$ , второй столбец — это нормированная линейная комбинация первых двух вектор-столбцов  $v_1$  и  $v_2$ , и т. д. Столбец с номером  $j$  результирующей ортогональной матрицы равен нормированной линейной комбинации первых  $j$  штук столбцов исходной матрицы  $W$ . Иначе говоря, процесс ортогонализации Грама–Шмидта равносителен умножению  $W$  справа на верхнюю треугольную матрицу  $Y = (\alpha_{kj})$ , в результа-

те чего должна получиться ортогональная матрица  $Q = WY$ . Тогда  $W = Y^{-1}Q$ .

Фактически ортогонализацию Грама–Шмидта можно рассматривать как ещё один способ получения QR-разложения матрицы, наряду с методом отражений (§ 3.7e) или методом вращений (§ 3.7г).<sup>22</sup> Но устойчивость ортогонализации Грама–Шмидта существенно хуже. Если исходное семейство векторов близко к линейно зависимому, то, выполняя алгоритм Грама–Шмидта с помощью приближённых вычислений (например, в арифметике с плавающей точкой современных ЭВМ), мы можем получить базис, который существенно отличается от ортогонального: попарные скалярные произведения его векторов будут заметно отличны от нуля.

Таблица 3.7. Модифицированный алгоритм  
ортогонализации Грама–Шмидта

```

DO FOR  $j = 1$  TO  $n$ 
     $q_j \leftarrow v_j$  ;
    DO  $k = 1$  TO  $j - 1$ 
         $\alpha_{kj} \leftarrow \langle q_k, q_j \rangle$  ;
         $q_j \leftarrow q_j - \alpha_{kj} q_k$  ;
    END DO
     $\alpha_{jj} \leftarrow \|q_j\|_2$  ;
    IF ( $\alpha_{jj} = 0$ ) THEN
        STOP, сигнализируя « $v_j$  линейно зависит
        от векторов  $v_1, v_2, \dots, v_{j-1}$ »
    END IF
     $q_j \leftarrow q_j / \alpha_{jj}$  ;
END DO

```

Причина этого явления довольно прозрачна. При построении QR-разложения с помощью матриц отражения или вращения ортогональ-

<sup>22</sup> Верно и обратное: любое разложение матрицы на множители, один из которых ортогонален, соответствует некоторому процессу ортогонализации. Вопрос в том, насколько удобны и технологичны соответствующие алгоритмы.

ность соответствующего матричного сомножителя специально организуется с самого начала и контролируется в процессе работы алгоритма, тогда как в ортогонализации Грама–Шмидта мы идём обратным путём, от треугольной матрицы, и ортогональность получается как конечный продукт более или менее длительного вычислительного процесса. Неудивительно, что эта ортогональность искажается для результата выполнения реального алгоритма, подверженного влиянию погрешностей вычислений.

Недостаточную устойчивость ортогонализации Грама–Шмидта до некоторой степени можно исправить, модифицировав расчётные формулы для вычисления поправочных коэффициентов  $\alpha_{kj}$ . Псевдокод алгоритма модифицированной ортогонализации Грама–Шмидта приведён в табл. 3.7. В нём формулы для  $\alpha_{kj}$  в силу свойств конструируемого семейства векторов  $\{q_1, q_2, \dots\}$  математически эквивалентны тем, что используются в исходной версии, но более устойчивы.

В общем случае при ортогонализации Грама–Шмидта построение каждого следующего вектора требует привлечения всех ранее построенных векторов. Но если исходное семейство векторов имеет специальный вид, в определённом смысле согласованный с используемым скалярным произведением, то ситуация упрощается. Важнейший частный случай — ортогонализация так называемых подпространств Крылова.

**Определение 3.8.1** Пусть даны квадратная  $n \times n$ -матрица  $A$  и  $n$ -вектор  $r$ . Подпространствами Крылова  $\mathcal{K}_i(A, r)$ ,  $i = 1, 2, \dots, n$ , матрицы  $A$  относительно вектора  $r$  называются линейные оболочки векторов  $r, Ar, \dots, A^{i-1}r$ , т. е.  $\mathcal{K}_i(A, r) := \text{span}\{r, Ar, \dots, A^{i-1}r\}$ .

Ограничение сверху  $n$  на количество подпространств Крылова вызвано тем, что по теореме Гамильтона–Кэли всякая матрица зануляет свой характеристический полином, который имеет степень  $n$ . Как следствие, векторы  $r, Ar, \dots, A^n r$  будут заведомо линейно зависимы, а  $n + 1$ -е подпространство Крылова  $\mathcal{K}_{n+1}(A, r)$ , являющееся их линейной оболочкой, не отличается от  $\mathcal{K}_n(A, r)$  и особого содержательного смысла не несёт.

Оказывается, если  $A$  — симметричная положительно определённая матрица, то при ортогонализации подпространств Крылова построение каждого следующего вектора привлекает лишь два предшествующих вектора из строящегося базиса. Более точно, справедлива

**Теорема 3.8.1** Пусть  $A$  — симметричная положительно определённая матрица и векторы  $r, Ar, A^2r, \dots, A^{n-1}r$  линейно независимы. Если векторы  $p_0, p_1, \dots, p_{n-1}$  получены из них с помощью процесса ортогонализации Грама–Шмидта, то они выражаются рекуррентными соотношениями

$$\begin{aligned} p_0 &= r, \\ p_1 &= Ap_0 - \alpha_0 p_0, \\ p_{k+1} &= Ap_k - \alpha_k p_k - \beta_k p_{k-1}, \quad k = 1, 2, \dots, n-2, \end{aligned}$$

где коэффициенты ортогонализации  $\alpha_k$  и  $\beta_k$  вычисляются следующим образом:

$$\begin{aligned} \alpha_k &= \frac{\langle Ap_k, p_k \rangle}{\langle p_k, p_k \rangle}, \quad k = 0, 1, \dots, n-2, \\ \beta_k &= \frac{\langle Ap_k, p_{k-1} \rangle}{\langle p_{k-1}, p_{k-1} \rangle} = \frac{\langle p_k, p_k \rangle}{\langle p_{k-1}, p_{k-1} \rangle}, \quad k = 1, 2, \dots, n-2. \end{aligned}$$

Этот факт был открыт К. Ланцшем в 1952 году и имеет многочисленные применения, так как более короткие вычислительные формулы меньше подвержены влиянию погрешностей вычислений и более устойчивы. Соответственно, получаемые с их помощью ортогональные семейства векторов обладают более высоким «качеством ортогональности». С другой стороны, подпространства Крылова естественно возникают во многих алгоритмах линейной алгебры, в частности в различных проекционных методах решения СЛАУ (см. § 3.11а, стр. 547).

Наконец, симметричной и положительно определённой матрицей  $A$  задаётся скалярное произведение в  $\mathbb{R}^n$  (см. § 3.3з). Поэтому результат теоремы 3.8.1 можно рассматривать также как способ построения в пространстве базиса, который  $A$ -ортогонален относительно скалярного произведения, порождаемого матрицей  $A$ . И тогда теорема 3.8.1 становится теоретической основой метода сопряжённых градиентов для решения СЛАУ (см. § 3.11д).

**Доказательство.** Если векторы  $p_0, p_1, \dots, p_{n-1}$  получены из  $r, Ar, A^2r, \dots, A^{n-1}r$  в результате ортогонализации Грама–Шмидта, то из

формул (3.118), (3.119) следует, что для любого  $k = 1, 2, \dots, n - 1$

$$p_{k+1} = A^{k+1}r - \sum_{i=0}^k c_i^{(k)} A^i r, \quad \text{где } c_i^{(k)} \in \mathbb{R}.$$

Отсюда вытекает, что вектор  $p_{k+1} - Ap_k$  принадлежит подпространству, являющемуся линейной оболочкой векторов  $r, Ar, \dots, A^k r$  или, что то же самое, линейной оболочкой векторов  $p_0, p_1, \dots, p_k$ . По этой причине  $p_{k+1}$  выражается через предшествующие векторы как

$$p_{k+1} = Ap_k - \gamma_0^{(k)} p_0 - \dots - \gamma_k^{(k)} p_k \quad (3.120)$$

с какими-то коэффициентами  $\gamma_0^{(k)}, \dots, \gamma_k^{(k)}$ .

Домножая скалярно выписанное равенство на векторы  $p_0, p_1, \dots, p_k$  и привлекая условие ортогональности вектора  $p_{k+1}$  всем  $p_0, p_1, \dots, p_k$ , получим

$$\gamma_j^{(k)} = \frac{\langle Ap_k, p_j \rangle}{\langle p_j, p_j \rangle}, \quad j = 0, 1, \dots, k.$$

Но при  $j = 0, 1, \dots, k - 2$  справедливо  $\langle Ap_k, p_j \rangle = 0$ , так как  $\langle Ap_k, p_j \rangle = \langle p_k, Ap_j \rangle$  в силу симметричности  $A$ , а вектор  $Ap_j$  есть линейная комбинация векторов  $p_0, p_1, \dots, p_{j+1}$ , каждый из которых ортогонален к  $p_k$  при  $j + 1 < k$ , т. е.  $j \leq k - 2$ .

Итак, из коэффициентов  $\gamma_j^{(k)}$  ненулевыми остаются лишь два коэффициента

$$\alpha_k = \gamma_k^{(k)} = \frac{\langle Ap_k, p_k \rangle}{\langle p_k, p_k \rangle}, \quad \beta_k = \gamma_{k-1}^{(k)} = \frac{\langle Ap_k, p_{k-1} \rangle}{\langle p_{k-1}, p_{k-1} \rangle},$$

а формула (3.120) принимает вид

$$p_{k+1} = Ap_k - \alpha_k p_k - \beta_k p_{k-1}.$$

Далее,

$$\langle Ap_k, p_{k-1} \rangle = \langle p_k, Ap_{k-1} \rangle = \langle p_k, p_k + \alpha_{k-1} p_{k-1} + \beta_{k-1} p_{k-2} \rangle = \langle p_k, p_k \rangle,$$

и поэтому

$$\beta_k = \frac{\langle p_k, p_k \rangle}{\langle p_{k-1}, p_{k-1} \rangle}.$$

Это завершает доказательство теоремы. ■

Рассмотренный алгоритм ортогонализации подпространств Крылова, использующий расчётные формулы из теоремы 3.8.1, называют *ортогонализацией Ланцоша*. Фактически, это удобный способ построения ортогонального базиса всего пространства, если подпространства Крылова линейно независимы.

## 3.9 Метод прогонки

Решая системы линейных алгебраических уравнений, мы до сих пор не делали никаких дополнительных предположений о структуре нулевых и ненулевых элементов в матрице системы. Но для многих систем линейных уравнений, встречающихся в практике математического моделирования, ненулевые элементы заполняют матрицу не полностью, образуя в ней разнообразные правильные структуры — ленты, блоки, их комбинации и т. п. Естественно попытаться использовать это обстоятельство при конструировании более эффективных численных методов для решения СЛАУ с такими матрицами. *Метод прогонки*, предложенный в 1952 году И.М. Гельфандом и О.В. Локуциевским, предназначен для решения линейных систем уравнений с трёхдиагональными матрицами.<sup>23</sup>

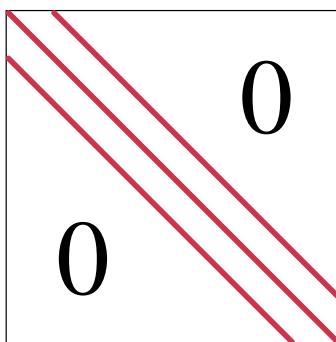


Рис. 3.23. Портрет трёхдиагональной матрицы

По определению *трёхдиагональными* называются матрицы, у ко-

---

<sup>23</sup> В англоязычной литературе этот метод называют также «tridiagonal matrix algorithm» или «Thomas algorithm» (алгоритм Томаса).

торых все ненулевые элементы сосредоточены на трёх диагоналях — главной и соседних с ней сверху и снизу. Иными словами, для трёхдиагональной матрицы  $H = (h_{ij})$  неравенство  $h_{ij} \neq 0$  может иметь место лишь при  $i = j$  и  $i = j \pm 1$ . Далее для краткости будем называть системы линейных алгебраических уравнений с такими матрицами просто «трёхдиагональными линейными системами». Это важный в приложениях случай СЛАУ, возникающий, к примеру, при решении многих краевых задач для дифференциальных уравнений.

Систему  $n$  линейных алгебраических уравнений относительно неизвестных  $x_1, x_2, \dots, x_n$ , имеющую трёхдиагональную матрицу, удобно представлять в специальном виде, даже не обращаясь к матрично-векторной форме:

$$\begin{cases} b_1x_1 + c_1x_2 = d_1, \\ a_ix_{i-1} + b_ix_i + c_ix_{i+1} = d_i, & 2 \leq i \leq n-1, \\ a_nx_{n-1} + b_nx_n = d_n. \end{cases} \quad (3.121)$$

Он равносителен

$$a_ix_{i-1} + b_ix_i + c_ix_{i+1} = d_i, \quad 1 \leq i \leq n, \quad (3.122)$$

где для единства обозначения полагают  $a_1 = c_n = 0$  в качестве коэффициентов при фиктивных переменных  $x_0$  и  $x_{n+1}$ . Подобный вид и обозначения оправдываются тем, что соответствующие СЛАУ получаются действительно «локально», как дискретизация дифференциальных уравнений, связывающих значения искомых величин тоже локально, в окрестности какой-либо рассматриваемой точки.

**Пример 3.9.1** Пусть  $u(x)$  — дважды непрерывно дифференцируемая функция. В § 2.8 мы видели, что на равномерной сетке с шагом  $h$

$$u''(x_i) \approx \frac{u(x_{i-1}) - 2u(x_i) + u(x_{i+1})}{h^2}.$$

Правая часть этой формулы помимо самого узла  $x_i$ , в котором берётся производная, вовлекает ещё только соседние узлы  $x_{i-1}$  и  $x_{i+1}$ . Поэтому решение конечно-разностными методами краевых задач для различных дифференциальных уравнений второго порядка часто приводит к линейным системам уравнений с трёхдиагональными матрицами, у которых помимо главной диагонали заполнены только две соседние с ней. ■

Соотношения вида (3.122)

$$a_i x_{i-1} + b_i x_i + c_i x_{i+1} = d_i, \quad i = 1, 2, \dots,$$

называют также *трёхточечными разностными уравнениями* или *разностными уравнениями второго порядка*.

Пусть для СЛАУ с трёхдиагональной матрицей выполняется прямой ход метода Гаусса без перестановки строк и столбцов матрицы, т. е. без специального выбора ведущего элемента. Если он успешно прорабатывает до конца, то приводит к системе с двухдиагональной матрицей вида

$$\begin{pmatrix} \times & \times & & & 0 \\ & \times & \ddots & & \\ & & \ddots & \times & \\ 0 & & & \times & \times \\ & & & & \times \end{pmatrix}, \quad (3.123)$$

в которой ненулевые элементы (обозначенные крестиками) присутствуют лишь на главной диагонали и первой наддиагонали. Следовательно, формулы обратного хода метода Гаусса вместо тех, что даны в табл. 3.2, должны иметь более простой вид, в котором очередная независимая переменная выражается не более чем через одну предшествующую переменную. Этим формулам можно придать следующий единообразный двучленный вид

$$x_i = \xi_i x_{i+1} + \eta_i, \quad i = n, n-1, \dots, 1. \quad (3.124)$$

Здесь в  $n$ -ом соотношении, как и в исходных уравнениях, присутствует вспомогательная фиктивная неизвестная  $x_{n+1} = 0$ . Оказывается, что величины  $\xi_i$  и  $\eta_i$  в соотношениях (3.124) можно несложным образом выразить через элементы исходной системы уравнений.

Уменьшим в (3.124) все индексы на единицу:

$$x_{i-1} = \xi_i x_i + \eta_i.$$

Подставив полученное соотношение в  $i$ -е уравнение системы, будем иметь

$$a_i(\xi_i x_i + \eta_i) + b_i x_i + c_i x_{i+1} = d_i.$$

Отсюда

$$x_i = -\frac{c_i}{a_i \xi_i + b_i} x_{i+1} + \frac{d_i - a_i \eta_i}{a_i \xi_i + b_i}.$$

Сравнивая равенство с аналогичными двучленными расчётными формулами (3.124), можем заключить, что для  $i = 1, 2, \dots, n$

$$\boxed{\begin{aligned}\xi_{i+1} &= -\frac{c_i}{a_i\xi_i + b_i}, \\ \eta_{i+1} &= \frac{d_i - a_i\eta_i}{a_i\xi_i + b_i}.\end{aligned}} \quad (3.125)$$

Это формулы *прямого хода* прогонки, целью которого является вычисление величин  $\xi_i$  и  $\eta_i$ , называемых *прогоночными коэффициентами*. Далее по формулам (3.124) выполняется *обратный ход*:

$$x_i = \xi_{i+1}x_{i+1} + \eta_{i+1}, \quad i = n, n-1, \dots, 1.$$

На нём находятся искомые значения неизвестных. Совместно два эти этапа — прямой ход и обратный ход — определяют метод прогонки для решения системы линейных алгебраических уравнений с трёхдиагональной матрицей.

Для начала расчётов по выведенным формулам требуется знать величины  $\xi_1$  и  $\eta_1$  в прямом ходе. Формально они неизвестны, но на самом деле полностью определяются условием  $a_1 = c_n = 0$ . Действительно, конкретные значения  $\xi_1$  и  $\eta_1$  не влияют на результаты решения, потому что в формулах (3.125) прямого хода прогонки они встречаются с множителем  $a_1 = 0$ . Кроме того,  $x_{n+1} = 0$ . Для начала прогонки можно положить, к примеру,

$$\xi_1 = \eta_1 = x_{n+1} = 0. \quad (3.126)$$

На практике более удобна другая эквивалентная реализация прогонки, при которой прямой ход начинается с присваиваний

$$\xi_2 = -c_1/b_1, \quad \eta_2 = d_1/b_1. \quad (3.127)$$

Они вытекают из первого уравнения системы (3.122) и формул (3.125) с  $\xi_1 = \eta_1 = 0$ . Далее находятся все прогоночные коэффициенты, а затем сразу полагаем

$$x_n = \eta_{n+1}. \quad (3.128)$$

После этого в обратном ходе прогонки находятся неизвестные  $x_{n-1}, \dots, x_2, x_1$ .

Дадим достаточные условия выполнимости метода прогонки, т. е. того, что знаменатели в расчётных формулах прямого хода не обращаются в нуль. Эти условия фактически будут также обосновывать возможность приведения трёхдиагональной матрицы исходной СЛАУ к двухдиагональному виду (3.123) преобразованиями прямого хода метода Гаусса без перестановки строк или столбцов. Эти преобразования являются не чем иным, как прямым ходом метода прогонки.

Здесь полностью применима теория, развитая в § 3.6ж, в частности, предложение 3.6.2.

**Предложение 3.9.1** *Если в матрице трёхдиагональной системы линейных алгебраических уравнений (3.121)–(3.122) имеет место диагональное преобладание, т. е.*

$$|b_i| > |a_i| + |c_i|, \quad i = 1, 2, \dots, n,$$

*то метод прогонки с выбором начальных значений согласно (3.126) или (3.127)–(3.128) является реализуемым.*

**Доказательство.** Диагональное преобладание в матрице влечёт её строгую регулярность, как мы видели в § 3.6ж. Поэтому в силу теоремы 3.6.2 существует LU-разложение такой матрицы, а предложение 3.6.2 утверждает, что оно может быть получено с помощью прямого хода метода Гаусса без перестановки строк и столбцов. Это и означает реализуемость метода прогонки. ■

Ниже даётся ещё одно доказательство этого факта, которое позволяет помимо реализуемости установить ещё числовые оценки «запаса устойчивости» прогонки, т. е. насколько сильно знаменатели выражений (3.125) для прогоночных коэффициентов отличны от нуля в зависимости от элементов матрицы СЛАУ.

**Доказательство.** Покажем по индукции, что в рассматриваемой реализации прогонки для всех индексов  $i$  справедливо неравенство  $|\xi_i| < 1$ .

Если прогонка начинается с помощью (3.126), то  $\xi_1 = 0$  и потому  $|\xi_1| < 1$ . Если прогонка начинается формулами (3.127), то в силу диагонального преобладания тоже  $|\xi_2| < 1$ , так что база индукции выполняется.

Далее предположим, что для некоторого индекса  $i$  уже установлена оценка  $|\xi_i| < 1$ . Если соответствующее  $c_i = 0$ , то из первой формулы (3.125) следует  $\xi_{i+1} = 0$ , и индукционный переход доказан. Поэтому пусть  $c_i \neq 0$ . Тогда справедлива следующая цепочка соотношений

$$\begin{aligned} |\xi_{i+1}| &= \left| -\frac{c_i}{a_i \xi_i + b_i} \right| = \frac{|c_i|}{|a_i \xi_i + b_i|} \leq \\ &\leq \frac{|c_i|}{||b_i| - |a_i| \cdot |\xi_i||} \quad \text{из оценки снизу для модуля суммы} < \\ &< \frac{|c_i|}{|a_i| + |c_i| - |a_i| \cdot |\xi_i|} \quad \text{в силу диагонального преобладания} = \\ &= \frac{|c_i|}{|a_i|(1 - |\xi_i|) + |c_i|} \leq \frac{|c_i|}{|c_i|} = 1, \end{aligned}$$

где при переходе ко второй строке используется известное неравенство для модуля суммы двух чисел:

$$|x + y| \geq ||x| - |y||. \quad (3.129)$$

Итак, неравенства  $|\xi_i| < 1$  доказаны для всех прогоночных коэффициентов  $\xi_i$ ,  $i = 1, 2, \dots, n+1$ .

Как следствие, для знаменателей прогоночных коэффициентов  $\xi_i$  и  $\eta_i$  в формулах (3.125) имеем

$$\begin{aligned} |a_i \xi_i + b_i| &\geq ||b_i| - |a_i \xi_i|| \quad \text{по неравенству (3.129)} = \\ &= |b_i| - |a_i| |\xi_i| \quad \text{в силу диагонального преобладания} > \\ &> |a_i| + |c_i| - |a_i| \cdot |\xi_i| \quad \text{в силу диагонального преобладания} = \\ &= |a_i|(1 - |\xi_i|) + |c_i| \geq \\ &\geq |c_i| \geq 0 \quad \text{в силу оценки } |\xi_i| < 1. \end{aligned}$$

Иными словами,  $|a_i \xi_i + b_i|$  строго отделены от нуля, что и требовалось доказать. Кроме того, чем большим является диагональное преобладание в матрице системы, тем «сильнее» выполняется строгое неравенство в третьей строке выписанной цепочки, и тем больше отделён от нуля знаменатель прогоночных коэффициентов (3.125). ■

Отметим, что существуют и другие условия реализуемости метода прогонки, основанные на диагональном преобладании в матрице СЛАУ. Например, некоторые из них требуют от матрицы более мягкое нестрогое диагональное преобладание (3.62), но зато более жёсткие, чем в предложении 3.9.1, условия на коэффициенты системы [8, 40]. Весьма популярна, в частности, такая формулировка [99]:

**Предложение 3.9.2** *Пусть в трёхдиагональной матрице системы линейных алгебраических уравнений (3.121)–(3.122) все элементы поддиагонали, за исключением, может быть, последнего, и все элементы наддиагонали, за исключением, возможно, первого, не равны нулю, т. е.  $a_i \neq 0$ ,  $c_i \neq 0$ ,  $i = 2, 3, \dots, n - 1$ , и, кроме того,  $b_1 \neq 0$ ,  $b_n \neq 0$ . Если матрица системы имеет нестрогое диагональное преобладание,*

$$|b_i| \geq |a_i| + |c_i|, \quad i = 1, 2, \dots, n,$$

*но хотя бы для одного индекса  $i$  это неравенство является строгим, то метод прогонки реализуем.*

Нетрудно убедиться, что реализация прогонки требует линейного в зависимости от размера системы количества арифметических операций (примерно  $8n$ ), т. е. весьма экономична.

На сегодняшний день разработано немало модификаций метода прогонки, которые хорошо приспособлены для решения различных специальных систем уравнений, как трёхдиагональных, так и более общих, имеющих ленточные или даже блочно-ленточные матрицы [18]. В частности, существует метод матричной прогонки [33].

## 3.10 Стационарные итерационные методы для решения линейных систем

### 3.10а Краткая теория

*Итерационные методы* решения уравнений и систем уравнений — это методы, порождающие последовательность приближений  $\{x^{(k)}\}_{k=0}^{\infty}$  к искомому решению  $x^*$ , которое получается как предел

$$x^* = \lim_{k \rightarrow \infty} x^{(k)}.$$

Допуская некоторую вольность речи, обычно говорят, что «итерационный метод сходится», если к пределу сходится конструируемая им последовательность приближений  $\{x^{(k)}\}$ .

Естественно, что на практике переход к пределу по  $k \rightarrow \infty$  невозможен в силу конечности объёма вычислений, который мы можем произвести. Поэтому при реализации итерационных методов вместо  $x^*$  обычно довольствуются нахождением какого-то достаточно хорошего приближения  $x^{(k)}$  к  $x^*$ . Здесь важно правильно выбрать условие остановки итераций, при котором мы прекращаем порождать очередные приближения и выдаём  $x^{(k)}$  в качестве решения. Подробнее этот вопрос рассматривается в § 3.15.

Общая схема итерационных методов выглядит следующим образом:

- ▶ выбираются одно или несколько *начальных приближений*  $x^{(0)}, x^{(1)}, \dots, x^{(\nu)}$ , для нахождения которых в рамках общего алгоритма, возможно, организуются отдельные вычисления,
- ▶ затем по известным начальным или уже найденным приближениям вычисляются следующие приближения

$$x^{(k)} \leftarrow T_k(x^{(0)}, x^{(1)}, \dots, x^{(k)}), \quad k = \nu + 1, \nu + 2, \dots, \quad (3.130)$$

где  $T_k$  — отображение, называемое *оператором перехода* или *оператором шага* (иногда уточняют, что  $k$ -го).

Конечно, в реальных итерационных процессах каждое следующее приближение, как правило, зависит не от всех предшествующих приближений, а лишь от какого-то их фиксированного конечного числа. Более точно, итерационный метод (3.130) называют *p-шаговым*, если его очередное приближение  $x^{(k)}$  является функцией только от  $p$  предпоследних приближений, т. е. от  $x^{(k-1)}, x^{(k-2)}, \dots, x^{(k-p)}$ . В частности, *одношаговые* итерационные методы имеют вид

$$x^{(k)} \leftarrow T_k(x^{(k-1)}), \quad k = 1, 2, \dots,$$

т. е. в них  $x^{(k)}$  зависит лишь от значения одной предшествующей итерации  $x^{(k-1)}$ . Для начала работы одношаговых итерационных методов нужно знать одно начальное приближение  $x^{(0)}$ .

Итерационный метод называется *стационарным*, если оператор перехода  $T_k$  не зависит от номера шага  $k$ , т. е.  $T_k = T$ , и *нестационарным*

в противном случае. Стационарные одношаговые итерационные методы

$$x^{(k)} \leftarrow T(x^{(k-1)}), \quad k = 1, 2, \dots,$$

с неизменным оператором  $T$  определяют наиболее простые итерационные методы для решения разнообразных задач, и часто в отношении них используют обобщённое (хотя и не вполне точное) название *методы простой итерации*.

В этой главе мы занимаемся линейными задачами, для решения которых в первую очередь будут строиться итерационные методы с расчётыми формулами того же вида. Более точно, *линейным p-шаговым итерационным методом* будут называться итерации, в которых оператор перехода имеет вид

$$\begin{aligned} T_k(x^{(k-1)}, x^{(k-2)}, \dots, x^{(k-p)}) &= \\ &= C^{(k,k-1)}x^{(k-1)} + C^{(k,k-2)}x^{(k-2)} + \dots + C^{(k,k-p)}x^{(k-p)} + d^{(k)} \end{aligned}$$

с какими-то коэффициентами  $C^{(k,k-1)}$ ,  $C^{(k,k-2)}$ , …,  $C^{(k,k-p)}$  и свободным членом  $d^{(k)}$ . В случае векторной неизвестной переменной  $x$  все  $C^{(k,l)}$  являются матрицами подходящих размеров, а  $d^{(k)}$  — вектор той же размерности, что и  $x$ . Матрицы  $C^{(k,l)}$  часто называют *матрицами перехода* рассматриваемого итерационного метода.

Итерационные методы были представлены выше в абстрактной манере, как некоторые конструктивные процессы, которые порождают последовательности, сходящиеся к искомому решению. В действительности мотивации возникновения и развития итерационных методов являлись ясными и практическими. Итерационные методы решения уравнений и систем уравнений возникли как уточняющие процедуры, которые позволяли за небольшое (удовлетворяющее практику) количество шагов получить приемлемое по точности приближённое решение задачи. Многие из классических итерационных методов явно несут отпечаток этих взглядов и ценностей.

История итерационных методов — не менее древняя, чем у прямых. Например, итерационный метод вычисления квадратного корня, связанный с именем Герона Александрийского (см. пример 4.4.8), был известен ещё древним вавилонянам. В Новом времени одним из первоходцев итерационных методов стал И. Ньютон, который предложил вычислительный алгоритм решения уравнений, носящий ныне его имя и являющийся одним из наиболее эффективных инструментов вычислительной математики (см. § 4.4д).

Для коррекции приближённого решения необходимо знать, насколько и как именно оно нарушает точное равенство обеих частей уравнения. На этом пути возникает важное понятие *невязки* приближённого решения  $\tilde{x}$ , которая определяется как разность левой и правой частей уравнения (системы уравнений) после подстановки в него  $\tilde{x}$ . Исследование этой величины, отдельных её компонент (в случае системы уравнений) и решение вопроса о том, как можно на основе этой информации корректировать приближение к решению, составляют важнейшую часть работы по конструированию и использованию итерационных методов.

Другой источник возникновения итерационных методов — исследования по разрешимости различных уравнений и систем уравнений общей природы. Именно в таком качестве итерационные процессы, сходящиеся к решениям интегральных уравнений, появились в середине XIX века в работах Ж. Лиувилля и затем, уже ближе к концу XIX века, в работах К.Г. Неймана. Конструктивный характер этих процессов позволял в некоторых случаях получать решение в явном виде, и по мере развития вычислительной математики идеи Ж. Лиувилля и К.Г. Неймана послужили основой для создания эффективных итерационных численных методов для решения различных задач.

Мы подробно рассматриваем различные итерационные методы для решения нелинейных уравнений и систем уравнений в главе 4, а здесь основное внимание будет уделено итерационному решению систем линейных алгебраических уравнений и проблемы собственных значений.

Причины, по которым для решения систем линейных уравнений итерационные методы могут оказаться более предпочтительными, чем прямые, заключаются в следующем. Большинство итерационных методов являются *самоисправляющимися*, т. е. такими, в которых погрешность, допущенная в вычислениях, при сходимости исправляется в ходе итерирования и не отражается на окончательном результате. Это следует из конструкции оператора перехода, в котором обычно по самому его построению присутствует информация о решаемой системе уравнений (см. конкретные примеры в этом и следующем разделах книги). При выполнении алгоритма эта информация на каждом шаге вносится в итерационный процесс и оказывает влияние на его ход. Напротив, прямые методы решения СЛАУ этим свойством не обладают: оттолкнувшись от исходной системы, алгоритм уже не возвращается к ней, а оперирует с её системами-следствиями, которые никакой обратной

связи от исходной системы не получают.<sup>24</sup>

Как правило, итерационные процессы сравнительно несложно программируются, поскольку представляют собой повторяющиеся единобразные процедуры, применяемые к последовательным приближениям к решению. Для СЛАУ с разреженными матрицами в итерационных процессах обычно легче, чем в прямых методах, учитывать структуру нулевых и ненулевых элементов матрицы и основываться на этом упрощённые формулы матрично-векторного умножения, которые существенно уменьшают общую трудоёмкость алгоритма.

Иногда системы линейных алгебраических уравнений задаются в операторном виде, рассмотренном в § 3.6а, т. е. так, что их матрица и правая часть не выписываются явно. Вместо этого задаётся действие такой матрицы (линейного оператора) на любой вектор, и это позволяет строить и использовать итерационные методы. С другой стороны, для таких систем преобразования матриц, которые являются основой прямых методов решения, довольно сложны или порой просто невозможны.

Наконец, быстро сходящиеся итерационные методы могут обеспечивать выигрыши по времени даже для СЛАУ общего вида, если требуют для нахождения ответа небольшое число итераций.

То обстоятельство, что искомое решение получается как (топологический) предел последовательности, порождаемой методом, является характерной чертой именно итерационных методов решения уравнений. Существуют и другие конструкции, с помощью которых решение конструируется по последовательности результатов отдельных шагов алгоритма. Интересный пример дают методы Монте-Карло, в которых ответ получается как усреднение последовательности, порождаемой численным методом.

### 3.10б Сходимость стационарных одношаговых итерационных методов

Системы линейных уравнений вида

$$x = Cx + d,$$

---

<sup>24</sup>Для исправления этого положения прямые методы решения СЛАУ в ответственных ситуациях часто дополняют процедурами итерационного уточнения. См., к примеру, пункт 67 главы 4 в [45].

в которых вектор неизвестных переменных выделен в одной из частей, мы будем называть *системами в рекуррентном виде* (или *рекуррентной форме*).

**Предложение 3.10.1** *Если  $\|C\| < 1$  в какой-нибудь матричной норме, то стационарный односторонний итерационный процесс*

$$x^{(k)} \leftarrow Cx^{(k-1)} + d, \quad k = 1, 2, \dots, \quad (3.131)$$

*сходится при любом начальном приближении  $x^{(0)}$ .*

**Доказательство.** В формулировке предложения ничего не говорит о пределе, к которому сходится последовательность приближений  $\{x^{(k)}\}_{k=0}^{\infty}$ , порождаемых итерационным процессом. Но мы можем указать его в явном виде и строить доказательство с учётом этого знания.

Если  $\|C\| < 1$  для какой-нибудь матричной нормы, то в силу результата о матричном ряде Неймана (предложение 3.3.13, стр. 406) матрица  $(I - C)$  неособенна и имеет обратную. Следовательно, система уравнений  $(I - C)x = d$ , как и равносильная ей  $x = Cx + d$ , имеют единственное решение, которое обозначим  $x^*$ . Покажем, что в условиях предложения это и есть предел последовательных приближений  $x^{(k)}$ .

В самом деле, если

$$x^* = Cx^* + d,$$

то, вычитая это равенство из соотношений  $x^{(k)} = Cx^{(k-1)} + d$ , будем иметь

$$x^{(k)} - x^* = C(x^{(k-1)} - x^*), \quad k = 1, 2, \dots$$

Вспомним, что всякая матричная норма согласована с некоторой векторной нормой (предложение 3.3.5), и именно эту норму мы применим к обеим частям последнего равенства. Получим

$$\|x^{(k)} - x^*\| = \|C(x^{(k-1)} - x^*)\| \leq \|C\| \|x^{(k-1)} - x^*\|.$$

Повторное применение этой оценки погрешности для  $x^{(k-1)}, x^{(k-2)}, \dots$  и т. д. вплоть до  $x^{(1)}$  приводит к цепочке неравенств

$$\begin{aligned} \|x^{(k)} - x^*\| &\leq \|C\| \cdot \|x^{(k-1)} - x^*\| \leq \\ &\leq \|C\|^2 \cdot \|x^{(k-2)} - x^*\| \leq \\ &\leq \dots \dots \leq \\ &\leq \|C\|^k \cdot \|x^{(0)} - x^*\|. \end{aligned} \quad (3.132)$$

Правая часть неравенства (3.132) сходится к нулю при  $k \rightarrow \infty$  в силу условия  $\|C\| < 1$ , поэтому последовательность приближений  $\{x^{(k)}\}$  действительно сходится к пределу  $x^*$ . ■

Побочным следствием доказательства предложения 3.10.1 является прояснение роли нормы матрицы перехода  $\|C\|$  как коэффициента давления погрешности приближений к решению СЛАУ. Это следует из неравенств (3.132): чем меньше  $\|C\|$ , тем быстрее убывает погрешность на каждом отдельном шаге итерационного процесса.

**Предложение 3.10.2** Для любой квадратной матрицы  $A$  и любого  $\epsilon > 0$  существует такая подчинённая матричная норма  $\|\cdot\|_\epsilon$ , что

$$\rho(A) \leq \|A\|_\epsilon \leq \rho(A) + \epsilon.$$

**Доказательство.** Левое из выписанных неравенств было обосновано ранее в теореме 3.3.1, и потому содержанием сформулированного результата является правое неравенство. Оно даёт фактически оценку снизу для спектрального радиуса с помощью некоторой специальной матричной нормы.

С помощью преобразования подобия приведём матрицу  $A$  к жордановой канонической форме:

$$S^{-1}AS = J,$$

где

$$J = \left( \begin{array}{cc|c|c} \lambda_1 & 1 & 0 & 0 \\ \lambda_1 & \ddots & \ddots & 0 \\ \ddots & 1 & \lambda_1 & 0 \\ \hline 0 & & \lambda_2 & 1 \\ & & & \ddots & \ddots & 0 \\ & & & & \lambda_2 & \\ \hline 0 & & 0 & & & \ddots \\ & & & & & \ddots \end{array} \right),$$

а  $S$  — некоторая неособенная матрица, осуществляющая преобразование подобия. Положим

$$D_\epsilon := \text{diag} \{1, \epsilon, \epsilon^2, \dots, \epsilon^{n-1}\},$$

т. е.  $D_\epsilon$  — диагональная  $n \times n$ -матрица с числами  $1, \epsilon, \epsilon^2, \dots, \epsilon^{n-1}$  по главной диагонали. Тогда нетрудно проверить, что

$$(SD_\epsilon)^{-1} A (SD_\epsilon) = D_\epsilon^{-1} (S^{-1} AS) D_\epsilon$$

$$= D_\epsilon^{-1} J D_\epsilon = \left( \begin{array}{ccc|cc|c} \lambda_1 & \epsilon & & 0 & & 0 \\ \lambda_1 & \ddots & & 0 & & 0 \\ \ddots & \ddots & \epsilon & & & \\ & & \lambda_1 & & & \\ \hline 0 & & & \lambda_2 & \epsilon & 0 \\ & & & & \ddots & \ddots \\ & & & & & \lambda_2 \\ \hline 0 & & & 0 & & \ddots \\ & & & & & \ddots \end{array} \right)$$

— матрица в «модифицированной» жордановой форме, которая отличается от обычной жордановой формы присутствием  $\epsilon$  вместо 1 на наддиагонали каждой жордановой клетки.

Действительно, умножение на диагональную матрицу слева — это умножение строк матрицы на соответствующие диагональные элементы, а умножение на диагональную матрицу справа равносильно умножению столбцов на элементы диагонали. Два таких умножения — на  $D_\epsilon^{-1} = \text{diag} \{1, \epsilon^{-1}, \epsilon^{-2}, \dots, \epsilon^{1-n}\}$  слева и на  $D_\epsilon = \text{diag} \{1, \epsilon, \epsilon^2, \dots, \epsilon^{n-1}\}$  справа — компенсируют друг друга на главной диагонали матрицы  $J$ . Но на наддиагонали, где ненулевые элементы имеют индексы  $(i, i+1)$ , от этих умножений остаётся множитель  $\epsilon^{-i} \epsilon^{i+1} = \epsilon$ ,  $i = 0, 1, \dots, n-1$ .

Определим теперь векторную норму

$$\|x\|_\epsilon := \|(SD_\epsilon)^{-1} x\|_\infty.$$

Тогда для подчинённой ей матричной нормы справедлива следующая

цепочка оценок:

$$\begin{aligned}
 \|A\|_\epsilon &= \max_{x \neq 0} \frac{\|Ax\|_\epsilon}{\|x\|_\epsilon} = \max_{x \neq 0} \frac{\|(SD_\epsilon)^{-1}Ax\|_\infty}{\|(SD_\epsilon)^{-1}x\|_\infty} = \\
 &= \max_{y \neq 0} \frac{\|(SD_\epsilon)^{-1}A(SD_\epsilon)y\|_\infty}{\|y\|_\infty} \text{ после замены } y = (SD_\epsilon)^{-1}x = \\
 &= \max_{y \neq 0} \frac{\|(D_\epsilon^{-1}JD_\epsilon)y\|_\infty}{\|y\|_\infty} = \|D_\epsilon^{-1}JD_\epsilon\|_\infty = \\
 &= \text{максимум сумм модулей элементов в } D_\epsilon^{-1}JD_\epsilon \text{ по строкам} \leq \\
 &\leq \max_i |\lambda_i(A)| + \epsilon = \rho(A) + \epsilon,
 \end{aligned}$$

где  $\lambda_i(A)$  —  $i$ -е собственное значение матрицы  $A$ . Неравенство при переходе к последней строке выкладок возникает по существу, так как матрица может иметь наибольшее по модулю собственное значение в жордановой клетке размера  $1 \times 1$ , в которой нет элементов наддиагонали. ■

Хотя доказательство предложения 3.10.2 опирается на жорданову форму матрицы, оно нечувствительно к возможным скачкообразным перестроениям этой формы при изменениях элементов матрицы. Приведённое доказательство зависит лишь от собственных значений матрицы, но не от структуры её жордановых клеток, так как замена  $\epsilon$  на нуль или наоборот на наддиагонали ничего не меняет в доказательстве.

**Теорема 3.10.1** Пусть система уравнений  $x = Cx + d$  имеет единственное решение. Стационарный одношаговый итерационный процесс

$$x^{(k)} \leftarrow Cx^{(k-1)} + d, \quad k = 1, 2, \dots, \quad (3.133)$$

сходится при любом начальном приближении  $x^{(0)}$  тогда и только тогда, когда  $\rho(C) < 1$ , т. е. когда спектральный радиус матрицы  $C$  меньше единицы.

Оговорка о единственности решения существенна. Если взять, к примеру,  $C = I$  и  $d = 0$ , то рассматриваемая система обратится в тождество  $x = x$ , которому удовлетворяет любой вектор. Соответствующий

итерационный процесс  $x^{(k)} \leftarrow x^{(k-1)}$ ,  $k = 1, 2, \dots$ , будет сходиться из любого начального приближения, хотя спектральный радиус матрицы перехода  $C$  равен единице.

**Доказательство.** Сначала покажем необходимость условия теоремы. Предположим, что порождаемая в итерационном процессе последовательность  $\{x^{(k)}\}$  сходится. Её пределом может быть только решение  $x^*$  системы  $x = Cx + d$ , т. е. необходимо  $\lim_{k \rightarrow \infty} x^{(k)} = x^*$ , в чём можно убедиться, переходя в соотношении

$$x^{(k)} = Cx^{(k-1)} + d$$

к пределу по  $k \rightarrow \infty$ . Далее, вычитая почленно равенство для точного решения  $x^* = Cx^* + d$  из расчётной формулы итерационного процесса  $x^{(k)} = Cx^{(k-1)} + d$ , получим

$$x^{(k)} - x^* = C(x^{(k-1)} - x^*), \quad k = 1, 2, \dots,$$

откуда

$$\begin{aligned} x^{(k)} - x^* &= C(x^{(k-1)} - x^*) = C^2(x^{(k-2)} - x^*) = \\ &= \dots \dots = C^k(x^{(0)} - x^*). \end{aligned}$$

Так как левая часть выписанных равенств при  $k \rightarrow \infty$  сходится к нулю, то должна сходиться к нулю и правая, причём для любого вектора  $x^{(0)}$ . В силу единственности и, следовательно, фиксированности решения  $x^*$ , вектор  $(x^{(0)} - x^*)$  тоже может быть произвольным. Но тогда сходимость погрешности к нулю возможна лишь при  $C^k \rightarrow 0$ . На основании предложения 3.3.12 (стр. 404) заключаем, что спектральный радиус  $C$  должен быть строго меньше 1.

Достаточность. Если  $\rho(C) < 1$ , то, взяв положительное  $\epsilon$  удовлетворяющим оценке  $\epsilon < 1 - \rho(C)$ , мы можем согласно предложению 3.10.2 выбрать матричную норму  $\|\cdot\|_\epsilon$  так, чтобы выполнялось неравенство  $\|C\|_\epsilon < 1$ . Далее в этих условиях применимо предложение 3.10.1, которое утверждает сходимость итерационного процесса (3.133)

$$x^{(k)} \leftarrow Cx^{(k-1)} + d, \quad k = 1, 2, \dots$$

Это завершает доказательство теоремы 3.10.1. ■

Доказанные результаты — два предложения и объединяющая их теорема — проясняют также роль спектрального радиуса среди различных характеристик матрицы. В § 3.3и показано, что спектральный радиус не является матричной нормой. Но, как выясняется, спектральный радиус с любой степенью точности можно приблизить некоторой подчинённой матричной нормой. Кроме того, понятие спектрального радиуса оказывается чрезвычайно полезным при исследовании итерационных процессов и вообще степеней матрицы.

**Следствие** из предложения 3.10.2. Степени матрицы  $A^k$  сходятся к нулевой матрице при  $k \rightarrow \infty$  тогда и только тогда, когда  $\rho(A) < 1$ .

В самом деле, в предложении 3.3.12 ранее было установлено, что из сходимости степеней матрицы  $A^k$  при  $k \rightarrow \infty$  к нулевой матрице вытекает  $\rho(A) < 1$ . Теперь результат предложения 3.10.2 позволяет сказать, что это условие на спектральный радиус является и достаточным: если  $\rho(A) < 1$ , то можем подобрать матричную норму так, чтобы  $\|A\| < 1$ , и тогда  $\|A^k\| \leq \|A\|^k \rightarrow 0$  при  $k \rightarrow \infty$ .

С учётом предложения 3.10.2 более точно переформулируются условия сходимости матричного ряда Неймана (предложение 3.3.13): он сходится для матрицы  $A$  тогда и только тогда, когда  $\rho(A) < 1$ , а условие  $\|A\| < 1$  является всего лишь достаточным.

Заметим, что для несимметричных матриц нормы, близкие к спектральному радиусу, могут оказаться очень экзотичными и даже неестественными. Это видно из доказательства теоремы 3.10.1. Как правило, исследовать сходимость итерационных процессов лучше всё-таки в обычных нормах, часто имеющих практический смысл.

Интересен вопрос о выборе начального приближения для итерационных методов решения СЛАУ. Иногда его решают из каких-то содержательных соображений, когда в силу физических и прочих причин уже известно какое-то хорошее приближение к решению, а итерационный метод предназначен для его уточнения. При отсутствии таких условий начальное приближение нужно выбирать на основе других идей.

Например, если в рекуррентном виде  $x = Cx + d$ , исходя из которого строятся сходящиеся итерации, матрица  $C$  имеет «малую» норму (относительно неё мы вправе предполагать, что  $\|C\| < 1$ ), то тогда членом  $Cx$  можно пренебречь. Как следствие, точное решение не сильно отличается от вектора свободных членов  $d$ , и поэтому можно взять

$x^{(0)} = d$ . Этот вектор привлекателен также тем, что получается как первая итерация при нулевом начальном приближении. Взяв  $x^{(0)} = d$ , мы сэкономим на этой итерации.

### 3.10в Подготовка линейной системы к итерационному процессу

В этом параграфе исследуются различные способы приведения системы линейных алгебраических уравнений

$$Ax = b \quad (3.134)$$

к равносильной системе в рекуррентном виде

$$x = Cx + d, \quad (3.135)$$

на основе которого можно организовывать одношаговый итерационный процесс для решения (3.134). Фактически это вопрос о том, как связан предел стационарных одношаговых итераций (3.133) с интересующим нас решением системы уравнений  $Ax = b$ . При этом практический интерес представляет, естественно, не всякое приведение системы (3.134) к виду (3.135), но лишь такое, которое удовлетворяет условию сходимости стационарного одношагового итерационного процесса. В предшествующем разделе показано, что им является неравенство  $\rho(C) < 1$ .

Существует большое количество различных способов приведения исходной СЛАУ к виду, допускающему применение итераций, большое разнообразие способов организации этих итерационных процессов и т. п. Не претендуя на всеохватную теорию, рассмотрим ниже лишь несколько общих приёмов подготовки и организации итераций.

Простейший способ состоит в том, чтобы добавить к обеим частям исходной системы по вектору неизвестной переменной  $x$ , т. е.

$$x + Ax = x + b, \quad (3.136)$$

а затем член  $Ax$  перенести в правую часть:

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

Иногда этот приём работает, но весьма часто он непригоден, так как спектральный радиус матрицы  $C = I - A$  оказывается не меньшим единицы.

В самом деле, если  $\lambda$  — собственное значение для  $A$ , то для матрицы  $(I - A)$  собственным значением будет  $1 - \lambda$ , и тогда  $1 - \lambda > 1$  при вещественных отрицательных  $\lambda$ . С другой стороны, если у матрицы  $A$  есть собственные значения, вещественные или комплексные, большие по модулю, чем 2, т. е. если  $|\lambda| > 2$ , то

$$|1 - \lambda| = |\lambda - 1| \geq ||\lambda| - 1| > 1,$$

и сходимости стационарных итераций тоже не получим.

Из предшествующих рассуждений можно видеть, что необходим активный способ управления свойствами матрицы  $C$  в получающейся системе рекуррентного вида  $x = Cx + d$ . Одним из важнейших инструментов такого управления служит *предобуславливание* исходной системы.

**Определение 3.10.1** *Предобуславливанием системы линейных алгебраических уравнений  $Ax = b$  называется умножение слева обеих её частей на некоторую матрицу  $\Lambda$ . Сама эта матрица  $\Lambda$  называется предобуславливающей матрицей или, коротко, предобуславливателем.*

Цель предобуславливания — изменение (вообще говоря, улучшение) свойств матрицы  $A$  исходной системы  $Ax = b$ , вместо которой мы получаем систему

$$(\Lambda A)x = \Lambda b.$$

Продуманный выбор предобуславливателя может изменить выгодным нам образом расположение спектра матрицы  $A$ , так необходимое для организации сходящихся итерационных процессов.

Естественно выполнить предобуславливание до перехода к системе (3.136), т. е. до прибавления вектора неизвестных  $x$  к обеим частям исходной СЛАУ. Поскольку тогда вместо системы  $Ax = b$  будем иметь  $(\Lambda A)x = \Lambda b$ , то далее получаем

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

Теперь в этом рекуррентном виде с помощью подходящего выбора  $\Lambda$  можно добиваться требуемых свойств матрицы  $(I - \Lambda A)$ .

Каким образом следует выбирать предобуславливатели? Совершенно общего рецепта на этот счёт не существует, и теория разбивается здесь на набор рекомендаций для ряда более или менее конкретных важных случаев.

Например, если в качестве предобуславливающей матрицы взять  $\Lambda = A^{-1}$  или хотя бы приближённо равную обратной к  $A$ , то вместо системы  $Ax = b$  получим  $(A^{-1}A)x = A^{-1}b$ , т. е. систему уравнений

$$Ix = A^{-1}b$$

или близкую к ней. Её матрица обладает всеми возможными достоинствами (хорошим диагональным преобладанием, малой обусловленностью и т. п.). Кроме того, элементы разности  $I - \Lambda A$  малы, а потому малой является норма этой матрицы, что обеспечивает сходимость соответствующего итерационного процесса (3.133). Ясно, что нахождение подобного предобуславливателя ненамного легче, чем решение исходной системы, но сама идея примера весьма плодотворна. На практике в качестве предобуславливателей часто берут несложно вычисляемые обратные матрицы для какой-то «существенной» части матрицы  $A$ , к примеру, для главной диагонали матрицы или же для главной диагонали вместе с поддиагональю и наддиагональю. Другой подход, связанный с предобуславливанием при помощи усечённых рядов, описывается в книге [35].

Ещё один способ приведения СЛАУ к рекуррентному виду основан на *расщеплении* матрицы системы.

**Определение 3.10.2** *Расщеплением квадратной матрицы  $A$  называется её представление в виде  $A = G + (-H) = G - H$ , где  $G$  – неособенная матрица.*

Если известно некоторое расщепление матрицы  $A$ ,  $A = G - H$ , то вместо исходной системы  $Ax = b$  мы можем рассмотреть

$$(G - H)x = b,$$

которая равносильна

$$Gx = Hx + b,$$

так что

$$x = G^{-1}Hx + G^{-1}b.$$

На основе полученного рекуррентного вида можно организовать итерации

$$x^{(k)} \leftarrow G^{-1}Hx^{(k-1)} + G^{-1}b, \quad k = 1, 2, \dots, \quad (3.137)$$

задавшись каким-то начальным приближением  $x^{(0)}$ .

Иногда по ряду причин невыгодно обращать матрицу  $G$  явно, так что расчётные формулы итерационного метода основывают на равенстве

$$Gx = Hx + b.$$

Тогда они выглядят следующим образом

$$\begin{cases} y \leftarrow Hx^{(k-1)} + b, \\ x^{(k)} \leftarrow (\text{решение системы } Gx = y), \end{cases} \quad k = 1, 2, \dots$$

Итерационные методы с такой организацией, в которых очередное приближение находится из вспомогательной СЛАУ, называют *неявными*. В целом можно сказать, что всякое расщепление матрицы СЛАУ помогает конструированию итерационных процессов.

Но практическое значение имеют не все расщепления, а лишь те, в которых матрица  $G$  обращается «относительно просто», чтобы организация итерационного процесса не сделалась более сложной задачей, чем решение исходной СЛАУ. Другое требование к матрицам, образующим расщепление, состоит в том, чтобы норма обратной для  $G$ , т. е.  $\|G^{-1}\|$ , была «достаточно малой». При этом мы скорее добьёмся сходимости итерационного процесса (3.137), так как  $\|G^{-1}H\| \leq \|G^{-1}\| \|H\|$ . Наоборот, если норма матрицы  $G^{-1}$  не мала, её элементы велики, то может оказаться  $\rho(G^{-1}H) > 1$ , и сходимости у итерационного процесса (3.137) не будет.

Очень популярный способ расщепления матрицы  $A$  состоит в том, чтобы сделать элементы в  $G = (g_{ij})$  и  $H = (h_{ij})$  взаимнодополнительными, т. е. такими, что  $g_{ij}h_{ij} = 0$  для любых индексов  $i$  и  $j$ . Тогда ненулевые элементы матриц  $G$  и  $(-H)$  совпадают с ненулевыми элементами  $A$ .

В качестве примеров несложного обращаемых матриц можно указать

- 1) диагональные матрицы,
- 2) треугольные матрицы,
- 3) трёхдиагональные матрицы,
- 4) ... .

Обратная матрица несложно находится также для некоторых других классов матриц (например, для ортогональных), но если эта обратная практически не меняет норму матриц, на которые она умножается, то соответствующие расщепления почти не используются в организации итерационных процессов.

Ниже в § 3.10д и § 3.10е мы подробно рассмотрим итерационные процессы, которые соответствуют первым двум пунктам из представленного списка. Детальный анализ некоторых типов расщеплений матриц читатель может увидеть в книгах [35, 133].

### 3.10г Метод Ричардсона и его оптимизация

Напомним, что *скалярными матрицами* (из-за своего родства с скалярам) называются матрицы, кратные единичным, т. е. имеющие вид  $\tau I$ , где  $\tau \in \mathbb{R}$  или  $\mathbb{C}$ . В этом разделе подробно исследуется описанная в § 3.10в возможность управления итерационным процессом на примере простейшего предобуславливания с помощью скалярной матрицы, когда  $\Lambda = \tau I$ ,  $\tau \in \mathbb{R}$  и  $\tau \neq 0$ .

Рассмотрим итерационный процесс

$$x^{(k)} \leftarrow (I - \tau A) x^{(k-1)} + \tau b, \quad k = 1, 2, \dots, \quad (3.138)$$

впервые предложенный в 1910 году в работе [124], который обычно называют *методом Ричардсона* (см. также § 3.11а). Если  $\lambda_i$ ,  $i = 1, 2, \dots, n$ , — собственные числа матрицы  $A$  (которые, вообще говоря, комплексные), то собственные числа матрицы  $(I - \tau A)$  равны  $(1 - \tau \lambda_i)$ . Ясно, что в случае, когда среди  $\lambda_i$  имеются числа с разным знаком вещественной части  $\operatorname{Re} \lambda_i$ , выражение

$$\operatorname{Re}(1 - \tau \lambda_i) = 1 - \tau \operatorname{Re} \lambda_i$$

при любом фиксированном вещественном  $\tau$  будет иметь как меньшие 1 значения для каких-то  $\lambda_i$ , так и большие чем 1 значения для некоторых других  $\lambda_i$ . Следовательно, добиться локализации всех значений  $(1 - \tau \lambda_i)$  в единичном круге комплексной плоскости с центром в нуле, т. е. соблюдения условия  $\rho(I - \tau A) < 1$ , никаким выбором  $\tau$  будет невозможно.

Далее рассмотрим практически важный частный случай, когда  $A$  — симметричная положительно определённая матрица, так что все  $\lambda_i$ ,  $i = 1, 2, \dots, n$ , вещественны и положительны. Обычно они не бывают известными, но нередко более или менее точно известен интервал их расположения на вещественной полуоси  $\mathbb{R}_+$ . Будем предполагать, что  $\lambda_i \in [\mu, M]$ ,  $i = 1, 2, \dots, n$ , и  $[\mu, M] \subset \mathbb{R}_+$ , т. е.  $\mu > 0$ .

Матрица  $(I - \tau A)$  тогда тоже симметрична, и потому её спектральный радиус совпадает с 2-нормой. Чтобы обеспечить сходимость итерационного процесса и сделать её наиболее быстрой в евклидовой норме,

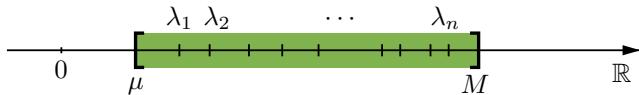


Рис. 3.24. Спектр положительно определённой матрицы системы и объемлющий его интервал

нам нужно, согласно теореме 3.10.1 и оценкам убывания погрешности (3.132), найти значение  $\tau$ , которое доставляет минимум величине

$$\|I - \tau A\|_2 = \max_{\lambda_i} |1 - \tau \lambda_i|.$$

Здесь максимум в правой части берётся по дискретному множеству собственных значений  $\lambda_i$  матрицы  $A$ ,  $i = 1, 2, \dots, n$ . В условиях, когда о расположении  $\lambda_i$  ничего не известно кроме их принадлежности интервалу  $[\mu, M]$ , естественно заменить максимизацию по множеству всех  $\lambda_i$ ,  $i = 1, 2, \dots, n$ , на максимизацию по объемлющему его интервалу  $[\mu, M]$ . Тогда

$$\|I - \tau A\|_2 = \max_{\lambda_i} |1 - \tau \lambda_i| \leq \max_{\lambda \in [\mu, M]} |1 - \tau \lambda|,$$

и мы будем искать оптимальное значение  $\tau = \tau_{\text{опт}}$ , на котором достигается наименьшее значение правой части этого неравенства, а также сам этот минимум, т. е.

$$\Theta = \min_{\tau} \left( \max_{\lambda \in [\mu, M]} |1 - \tau \lambda| \right).$$

Ясно, что

$$\min_{\tau} \|I - \tau A\|_2 = \min_{\tau} \max_{\lambda_i} |1 - \tau \lambda_i| \leq \Theta. \quad (3.139)$$

Обозначим

$$g(\tau) := \max_{\mu \leq \lambda \leq M} |1 - \tau \lambda|$$

и обратимся для минимизации функции  $g(\tau)$  к наглядной иллюстрации на рис. 3.25. Пользуясь ею, исследуем поведение функции  $g(\tau)$  при изменении аргумента  $\tau$ .

При  $\tau \leq 0$  выражение  $(1 - \tau \lambda)$  не убывает по  $\lambda$  и при положительных  $\lambda$ , очевидно, имеет значения не меньше 1 (на рис. 3.25 этому случаю

соответствует прямая, идущая от точки  $(0, 1)$  вправо-вверх). Тогда итерационный процесс (3.138) сходиться не будет. Следовательно, в нашем анализе имеет смысл ограничиться только теми  $\tau$ , для которых  $(1 - \tau\lambda)$  убывает по  $\lambda$ . Это значения  $\tau > 0$ , и на рис. 3.25 им соответствуют прямые, идущие от точки с координатами  $(0, 1)$  вправо-вниз.

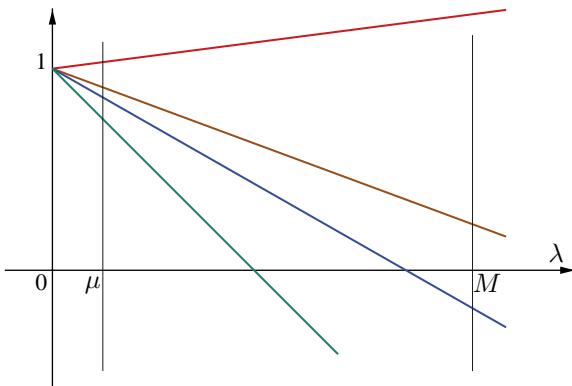


Рис. 3.25. Графики функций  $1 - \tau\lambda$  для различных  $\tau$

При  $0 < \tau \leq M^{-1}$  выражение  $(1 - \tau\lambda)$  на интервале  $\lambda \in [\mu, M]$  неотрицательно и монотонно убывает по  $\lambda$ . Поэтому

$$g(\tau) = \max_{\lambda} |1 - \tau\lambda| = 1 - \tau\mu,$$

где максимум по  $\lambda$  достигается на левом конце интервала  $[\mu, M]$ .

При  $\tau > M^{-1}$  величина  $1 - \tau M$  отрицательна, так что график функции  $1 - \tau\lambda$  на интервале  $\lambda \in [\mu, M]$  пересекает ось абсцисс. Тогда

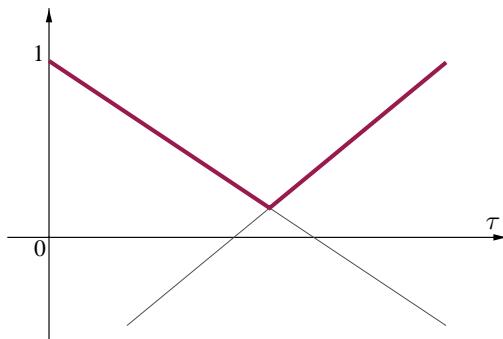
$$g(\tau) = \max \{1 - \tau\mu, -(1 - \tau M)\},$$

причём на левом конце  $(1 - \tau\mu)$  убывает с ростом  $\tau$ , а на правом конце  $-(1 - \tau M)$  растёт с ростом  $\tau$ .

При дальнейшем увеличении  $\tau$  наступает момент, когда значения функции  $|1 - \tau\lambda|$  на концах интервала  $[\mu, M]$  сравниваются друг с другом:

$$1 - \tau\mu = -(1 - \tau M). \quad (3.140)$$

Соответствующее  $\tau = \tau_{\text{опт}}$  доставляет искомый оптимум, поскольку дальнейшее увеличение  $\tau$  приводит к росту  $-(1 - \tau M)$  на правом конце

Рис. 3.26. График функции  $g(\tau)$ 

интервала, а уменьшение  $\tau$  ведёт к росту  $(1 - \tau\mu)$  на левом конце. В любом из этих случаев  $g(\tau)$  возрастает.

Из сказанного и из равенства (3.140) следует

$$\tau_{\text{опт}} = \frac{2}{M + \mu}, \quad (3.141)$$

а значение оптимума  $g(\tau)$  равно

$$\Theta = \min_{\tau} \max_{\lambda \in [\mu, M]} |1 - \tau\lambda| = 1 - \tau_{\text{опт}}\mu = 1 - \frac{2}{M + \mu} \cdot \mu = \frac{M - \mu}{M + \mu}.$$

Соответственно, в силу неравенства (3.139)

$$\|I - \tau_{\text{опт}} A\|_2 \leq \Theta = \frac{M - \mu}{M + \mu}, \quad (3.142)$$

и эту величину можно рассматривать, как коэффициент подавления евклидовой нормы погрешности (ввиду неравенств (3.132)). Она меньше единицы, т. е. даже с помощью простейшего скалярного предобуславливателя мы добились сходимости итерационного процесса.

Итогом наших рассуждений является

**Теорема 3.10.2** *Если в системе линейных алгебраических уравнений  $Ax = b$  матрица  $A$  симметрична, положительно определена и все её собственные значения лежат на интервале  $[\mu, M] \subset \mathbb{R}_+$ , то последовательность  $\{x^{(k)}\}$ , порождаемая стационарным методом Ричардсона с параметром  $\tau = 2/(M + \mu)$ , сходится к решению системы  $x^*$ .*

из любого начального приближения  $x^{(0)}$ . Быстрота этой сходимости оценивается неравенством

$$\|x^{(k)} - x^*\|_2 \leq \left(\frac{M-\mu}{M+\mu}\right)^k \|x^{(0)} - x^*\|_2, \quad k = 0, 1, 2, \dots \quad (3.143)$$

Полезно оценить значение (3.142) через спектральное число обусловленности матрицы  $A$ . Так как  $\mu \leq \lambda_{\min}(A)$  и  $\lambda_{\max}(A) \leq M$ , для положительно определённой матрицы  $A$  справедливо

$$\text{cond}_2(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \leq \frac{M}{\mu}.$$

Принимая во внимание тот факт, что функция

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

возрастает при положительных  $x$ , можем заключить, что

$$\frac{M-\mu}{M+\mu} = \frac{M/\mu - 1}{M/\mu + 1} \geq \frac{\text{cond}_2(A) - 1}{\text{cond}_2(A) + 1}.$$

Получается, что чем больше  $\text{cond}_2(A)$ , т. е. чем хуже обусловленность матрицы  $A$  исходной системы, тем ближе отношение  $\frac{M-\mu}{M+\mu}$  к единице и тем медленнее, вообще говоря, сходится наш итерационный процесс. Итак, число обусловленности матрицы системы характеризует не только чувствительность её решения к возмущениям и погрешностям в данных, но и скорость сходимости итерационных процессов к этому решению. Мы увидим далее, что это характерно для поведения многих итерационных методов (см. § 3.11в, 3.11г, 3.11е).

Наибольшую трудность на практике представляет нахождение  $\mu$ , т. е. нижней границы спектра матрицы СЛАУ. Иногда мы даже можем ничего не знать о её конкретной величине кроме того, что  $\mu \geq 0$ . В этих условиях развитая нами теория применима лишь частично, но добиться сходимости итераций мы всё-таки можем.

При  $\mu = 0$  непосредственно применять формулу

$$\tau_{\text{опт}} = \frac{2}{M+\mu} = \frac{2}{M} \quad (3.144)$$

уже нельзя, так как если  $M$  — точная верхняя граница спектра симметричной положительно определённой матрицы  $A$ , то соответствующее значение нормы оператора перехода  $\|I - \tau_{\text{опт}} A\|_2$  может стать равным единице. Но если эта верхняя граница спектра  $M$  не точна и оценивает его с некоторым запасом (чего можно добиться «ручной» корректировкой  $M$  вверх), то (3.144) является разумным значением  $\tau$ , при котором обеспечивается сходимость итераций метода Ричардсона (3.138). Естественно, что о какой-либо оптимальности выбранного параметра говорить не приходится.

### 3.10д Итерационный метод Якоби

Пусть в системе линейных алгебраических уравнений  $Ax = b$  матрица  $A = (a_{ij})$  — квадратная, размера  $n \times n$ , и её диагональные элементы отличны от нуля, т. е.  $a_{ii} \neq 0$ ,  $i = 1, 2, \dots, n$ . Это условие нисколько не ограничит общность наших рассуждений, так как в неособенной матрице в каждой строке и каждом столбце должны присутствовать ненулевые элементы. С помощью перестановки строк такой матрицы (соответствующей перестановке уравнений системы) всегда можно сделать её диагональные элементы ненулевыми.

Перепишем систему линейных алгебраических уравнений  $Ax = b$  в развёрнутом виде:

$$\sum_{j=1}^n a_{ij} x_j = b_i, \quad i = 1, 2, \dots, n.$$

Так как  $a_{ii} \neq 0$ , то из  $i$ -го уравнения мы можем выразить  $i$ -ю компоненту вектора неизвестных:

$$x_i = \frac{1}{a_{ii}} \left( b_i - \sum_{j \neq i} a_{ij} x_j \right), \quad i = 1, 2, \dots, n.$$

Нетрудно понять, что эти соотношения дают представление исходной СЛАУ в рекуррентном виде  $x = T(x)$ , необходимом для организации одношаговых итераций вида  $x^{(k)} \leftarrow T(x^{(k-1)})$ ,  $k = 1, 2, \dots$ . Более точно, можно взять

$$T(x) = (T_1(x), T_2(x), \dots, T_n(x))^\top$$

и

$$T_i(x) = \frac{1}{a_{ii}} \left( b_i - \sum_{j \neq i} a_{ij} x_j \right), \quad i = 1, 2, \dots, n.$$

Таблица 3.8. Итерационный метод Якоби для решения СЛАУ

```

 $k \leftarrow 1;$ 
выбираем начальное приближение  $x^{(0)}$ ;
DO WHILE ( метод не сопёлся )
    DO FOR  $i = 1$  TO  $n$ 
         $x_i^{(k)} \leftarrow \frac{1}{a_{ii}} \left( b_i - \sum_{j \neq i} a_{ij} x_j^{(k-1)} \right)$ 
    END DO
     $k \leftarrow k + 1;$ 
END DO

```

Псевдокод соответствующего итерационного процесса представлен в табл. 3.8, где вспомогательная переменная  $k$  — это счётчик числа итераций. Он был предложен ещё в середине XIX века К.Г. Якоби и часто (особенно в старых книгах по численным методам) называется «методом одновременных смещений». Под «смещениями» здесь имеются в виду коррекции компонент очередного приближения к решению, выполняемые на каждом шаге итерационного метода. Смещения-коррекции «одновременны» потому, что все компоненты очередного приближения  $x^{(k)}$  насчитываются независимо друг от друга по единообразным формулам, основанным на использовании лишь предыдущего приближения  $x^{(k-1)}$ .

В следующем разделе рассматривается итерационный метод Гаусса–Зейделя, устроенный несколько иначе. В нём смещения-коррекции компонент очередного приближения к решению «не одновременны» в том смысле, что находятся последовательно одна за другой не только из предыдущего приближения, но и друг из друга.

Пусть  $A = \tilde{L} + D + \tilde{U}$ , где

$$\tilde{L} := \begin{pmatrix} 0 & & & & \\ a_{21} & 0 & & & \\ a_{31} & a_{32} & \ddots & & \\ \vdots & \vdots & \ddots & 0 & \\ a_{n1} & a_{n2} & \cdots & a_{n,n-1} & 0 \end{pmatrix} \quad \text{— строго нижняя треугольная матрица,}$$

$$D := \text{diag}\{a_{11}, a_{22}, \dots, a_{nn}\} \quad \text{— диагональ матрицы } A,$$

$$\tilde{U} := \begin{pmatrix} 0 & a_{12} & \cdots & a_{1,n-1} & a_{1n} \\ 0 & \ddots & a_{2,n-1} & a_{2n} \\ \ddots & \ddots & \vdots & \vdots \\ 0 & & 0 & a_{n-1,n} \\ & & & 0 \end{pmatrix} \quad \text{— строго верхняя треугольная матрица.}$$

Тогда итерационный метод Якоби может быть представлен как метод, основанный на расщеплении матрицы системы  $A = G - H$  (см. § 3.10в), в котором

$$G = D, \quad H = -(\tilde{L} + \tilde{U}).$$

Соответственно, в матричном виде метод Якоби записывается как

$$x^{(k)} \leftarrow -D^{-1}(\tilde{L} + \tilde{U})x^{(k-1)} + D^{-1}b, \quad k = 1, 2, \dots$$

Теперь нетрудно дать условия его сходимости, основываясь на общем результате о сходимости стационарных одношаговых итераций (теорема 3.10.1). Именно, метод Якоби сходится из любого начального приближения тогда и только тогда, когда

$$\rho(D^{-1}(\tilde{L} + \tilde{U})) < 1.$$

Матрица  $D^{-1}(\tilde{L} + \tilde{U})$  просто выписывается по исходной системе и имеет вид

$$\begin{pmatrix} 0 & a_{12}/a_{11} & \dots & a_{1n}/a_{11} \\ a_{21}/a_{22} & 0 & \dots & a_{2n}/a_{22} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}/a_{nn} & a_{n2}/a_{nn} & \dots & 0 \end{pmatrix}. \quad (3.145)$$

Но нахождение её спектрального радиуса является задачей, сравнимой по сложности с выполнением самого итерационного процесса, и потому применять его для исследования сходимости метода Якоби непрактично. Для быстрой и грубой оценки спектрального радиуса можно воспользоваться какой-нибудь матричной нормой и результатом теоремы 3.3.1.

Полезен также следующий достаточный признак сходимости:

**Теорема 3.10.3** *Если в системе линейных алгебраических уравнений  $Ax = b$  квадратная матрица  $A$  имеет диагональное преобладание, то метод Якоби для решения этой системы сходится при любом начальном приближении.*

**Доказательство.** Диагональное преобладание в матрице  $A = (a_{ij})$  означает, что

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|, \quad i = 1, 2, \dots, n.$$

Следовательно,

$$\sum_{j \neq i} \left| \frac{a_{ij}}{a_{ii}} \right| < 1, \quad i = 1, 2, \dots, n,$$

что равносильно

$$\max_{1 \leq i \leq n} \left( \sum_{j \neq i} \left| \frac{a_{ij}}{a_{ii}} \right| \right) < 1.$$

В выражении, стоящем в левой части неравенства, легко угадать подчинённую чебышёвскую норму ( $\infty$ -норму) матрицы  $D^{-1}(\tilde{L} + \tilde{U})$ , которая выписана в (3.145). Таким образом,

$$\|D^{-1}(\tilde{L} + \tilde{U})\|_\infty < 1,$$

откуда ввиду результата предложения 3.10.1 следует доказываемое. ■

Итерационный метод Якоби был изобретён в середине XIX века и сейчас при практическом решении систем линейных алгебраических уравнений используется нечасто, так как существенно проигрывает по эффективности более современным численным методам.<sup>25</sup> Но совсем

---

<sup>25</sup> Примеры применения и детальные оценки скорости сходимости метода Якоби для решения модельных задач математической физики можно увидеть в [40].

забывать метод Якоби было бы преждевременным. Во-первых, он очень хорошо распараллеливается, и это благоприятствует его реализации на ЭВМ с современными архитектурами [35]. Во-вторых, лежащая в его основе идея выделения из оператора системы уравнений «диагональной части» достаточно плодотворна и может быть с успехом применена в различных ситуациях.

Рассмотрим, к примеру, систему уравнений

$$Ax = b(x),$$

в которой  $A$  —  $n \times n$ -матрица,  $b(x)$  — некоторая вектор-функция от неизвестной переменной  $x \in \mathbb{R}^n$ . В случае, когда  $b(x)$  — нелинейная функция, численные методы для решения СЛАУ здесь едва ли применимы, но для отыскания решения мы можем воспользоваться незначительной модификацией итераций Якоби:

$$x_i^{(k)} \leftarrow \frac{1}{a_{ii}} \left( b_i(x^{(k-1)}) - \sum_{j \neq i} a_{ij} x_j^{(k-1)} \right), \quad i = 1, 2, \dots, n, \quad (3.146)$$

$k = 1, 2, \dots$ , с некоторым начальным приближением  $x^{(0)}$ .

Обозначим  $n \times n$ -матрицу матрицу Якоби (матрицу частных производных) вектор-функции  $b(x)$  через  $b'(x)$ . Если  $b(x)$  изменяется «достаточно медленно», так что

$$\rho(D^{-1}(\tilde{L} + \tilde{U} + b'(x))) < 1$$

для любых  $x \in \mathbb{R}^n$ , то итерационный процесс (3.146) сходится из произвольного начального приближения. Это нетрудно показать, применяя теорему о конечном приращении для функции  $b(x)$  и затем теорему Шрёдера о неподвижной точке (теорема 4.3.5, стр. 729) к отображению, которое задаётся правой частью (3.146).

Вообще, *нелинейный итерационный процесс Якоби* в применении к системе уравнений

$$\begin{cases} F_1(x_1, x_2, \dots, x_n) = 0, \\ F_2(x_1, x_2, \dots, x_n) = 0, \\ \vdots \quad \ddots \quad \vdots \\ F_n(x_1, x_2, \dots, x_n) = 0 \end{cases}$$

может заключаться в следующем. Задавшись каким-то вектором начального приближения  $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, \dots, x_n^{(0)})^\top$ , на очередном  $k$ -ом шаге,  $k = 1, 2, \dots$ , последовательно находят решения  $\tilde{x}_i$  уравнений

$$F_i(x_1^{(k-1)}, \dots, x_{i-1}^{(k-1)}, x_i, x_{i+1}^{(k-1)}, \dots, x_n^{(k-1)}) = 0, \quad i = 1, 2, \dots, n,$$

относительно  $x_i$ , а затем полагают  $x^{(k)} \leftarrow (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n)^\top$ . Эти итерации являются «нелинейным аналогом» метода Якоби из табл. 3.8.

### 3.10e Итерационный метод Гаусса–Зейделя

В итерационном методе Якоби при организации вычислений по инструкции

$$x_i^{(k)} \leftarrow \frac{1}{a_{ii}} \left( b_i - \sum_{j \neq i} a_{ij} x_j^{(k-1)} \right), \quad i = 1, 2, \dots, n, \quad (3.147)$$

компоненты очередного приближения  $x^{(k)}$  находятся последовательно одна за другой, так что к моменту вычисления  $i$ -й компоненты вектора  $x^{(k)}$  уже найдены  $x_1^{(k)}, x_2^{(k)}, \dots, x_{i-1}^{(k)}$ . Но метод Якоби никак не использует эти новые значения, и при вычислении любой компоненты следующего приближения всегда опирается только на вектор  $x^{(k-1)}$ . Если итерации сходятся к решению, то естественно ожидать, что все компоненты  $x^{(k)}$  ближе к искомому решению, чем у  $x^{(k-1)}$ , а потому немедленное вовлечение их в процесс вычислений будет способствовать ускорению сходимости.

Сформулированные выше соображения являются основой *итерационного метода Гаусса–Зейделя*, идея которого была высказана сначала К.Ф. Гауссом, а затем окончательно реализована Ф.Л. Зейделем в публикации 1874 года.<sup>26</sup> Псевдокод метода Гаусса–Зейделя представлен в табл. 3.9, где  $k$  — счётчик итераций. В нём предполагается, как и в методе Якоби, что  $a_{ii} \neq 0$  в результате предварительной подготовки системы уравнений.

В методе Гаусса–Зейделя суммирование в формуле (3.147) для вычисления  $i$ -й компоненты очередного приближения  $x^{(k)}$  разбито на две части — по индексам, предшествующим  $i$ , и по индексам, следующим за  $i$ . Первая часть суммы использует новые вычисленные значения  $x_1^{(k)}$ ,

<sup>26</sup>По этой причине в отечественной литературе по вычислительной математике нередко используется термин «метод Зейделя».

Таблица 3.9. Итерационный метод Гаусса–Зейделя  
для решения линейных систем уравнений

```

 $k \leftarrow 1;$ 
выбираем начальное приближение  $x^{(0)}$ ;
DO WHILE ( метод не сошёлся )
    DO FOR  $i = 1$  TO  $n$ 
         $x_i^{(k)} \leftarrow \frac{1}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^n a_{ij}x_j^{(k-1)} \right)$ 
    END DO
     $k \leftarrow k + 1;$ 
END DO

```

$\dots, x_{i-1}^{(k)}$ , тогда как вторая — компоненты  $x_{i+1}^{(k-1)}, \dots, x_n^{(k-1)}$  из старого приближения. Метод Гаусса–Зейделя иногда называют также итерационным методом «последовательных смещений», а его основная идея — немедленно вовлекать уже полученную информацию в вычисления — с успехом применима и для нелинейных итерационных процессов.

Чтобы получить для метода Гаусса–Зейделя матричное представление, перепишем его расчётные формулы в виде

$$a_{ii}x_i^{(k)} + \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} = - \sum_{j=i+1}^n a_{ij}x_j^{(k-1)} + b_i, \quad i = 1, 2, \dots, n.$$

Используя введённые в § 3.10д матрицы  $\tilde{L}$ ,  $D$  и  $\tilde{U}$ , на которые разлагается  $A$ , можем записать эти равенства следующим образом:

$$(D + \tilde{L})x^{(k)} = -\tilde{U}x^{(k-1)} + b,$$

т. е.

$$x^{(k)} = -(D + \tilde{L})^{-1}\tilde{U}x^{(k-1)} + (D + \tilde{L})^{-1}b, \quad k = 1, 2, \dots \quad (3.148)$$

Итак, метод Гаусса–Зейделя можно рассматривать как итерационный метод, порождённый таким расщеплением матрицы СЛАУ в виде  $A = G - H$  (см. § 3.10в), что  $G = D + \tilde{L}$ ,  $H = -\tilde{U}$ .

В силу теоремы 3.10.1 необходимым и достаточным условием сходимости метода Гаусса–Зейделя из любого начального приближения является неравенство

$$\rho((D + \tilde{L})^{-1}\tilde{U}) < 1.$$

Но, как и в случае аналогичного условия для метода Якоби, оно имеет главным образом теоретическое значение.

**Теорема 3.10.4** *Если в системе линейных уравнений  $Ax = b$  матрица  $A$  имеет диагональное преобладание, то метод Гаусса–Зейделя для решения этой системы сходится при любом начальном приближении.*

**Доказательство.** Отметим, прежде всего, что в условиях диагонального преобладания в  $A$  решение  $x^*$  линейной системы  $Ax = b$  всегда существует и единственно (вспомним признак неособенности Адамара, § 3.4в). Пусть, как и ранее,  $x^{(k)}$  — приближение к решению, полученное на  $k$ -ом шаге итерационного процесса. Исследуем поведение погрешности решения  $z^{(k)} := x^{(k)} - x^*$  в зависимости от номера итерации  $k$ .

Чтобы получить формулу для  $z^{(k)}$ , перепишем соотношения, которым удовлетворяет точное решение  $x^*$ : вместо

$$\sum_{j=1}^n a_{ij} x_j^* = b_i, \quad i = 1, 2, \dots, n,$$

придадим им следующий эквивалентный вид

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

Вычитая затем почленно эти равенства из расчётных формул метода Гаусса–Зейделя, т. е. из

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

можем заключить, что

$$z_i^{(k)} = \frac{1}{a_{ii}} \left( - \sum_{j=1}^{i-1} a_{ij} z_j^{(k)} - \sum_{j=i+1}^n a_{ij} z_j^{(k-1)} \right), \quad i = 1, 2, \dots, n.$$

Возьмём абсолютные значения от обеих частей этих равенств и воспользуемся неравенством треугольника для оценки сумм в правых частях. Будем иметь

$$\begin{aligned} |z_i^{(k)}| &\leq \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| \cdot |z_j^{(k)}| + \sum_{j=i+1}^n \left| \frac{a_{ij}}{a_{ii}} \right| \cdot |z_j^{(k-1)}| \leq \\ &\leq \|z^{(k)}\|_\infty \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| + \|z^{(k-1)}\|_\infty \sum_{j=i+1}^n \left| \frac{a_{ij}}{a_{ii}} \right| \end{aligned} \quad (3.149)$$

для  $i = 1, 2, \dots, n$ .

С другой стороны, диагональное преобладания в матрице  $A$ , т. е.

$$\sum_{j \neq i} |a_{ij}| < |a_{ii}|, \quad i = 1, 2, \dots, n,$$

означает существование константы  $\varkappa$ ,  $0 \leq \varkappa < 1$ , такой что

$$\sum_{j \neq i} |a_{ij}| \leq \varkappa |a_{ii}|, \quad i = 1, 2, \dots, n. \quad (3.150)$$

По этой причине

$$\sum_{j \neq i} \left| \frac{a_{ij}}{a_{ii}} \right| \leq \varkappa, \quad i = 1, 2, \dots, n,$$

откуда следует

$$\sum_{j=i+1}^n \left| \frac{a_{ij}}{a_{ii}} \right| \leq \varkappa - \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| \leq \varkappa - \varkappa \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| = \varkappa \left( 1 - \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| \right).$$

Подставляя полученную оценку в неравенства (3.149), приходим к соотношениям

$$|z_i^{(k)}| \leq \|z^{(k)}\|_\infty \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| + \varkappa \|z^{(k-1)}\|_\infty \left( 1 - \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| \right), \quad (3.151)$$

$i = 1, 2, \dots, n$ .

Предположим, что  $\max_{1 \leq i \leq n} |z_i^{(k)}|$  достигается при  $i = l$ , так что

$$\|z^{(k)}\|_\infty = |z_l^{(k)}|. \quad (3.152)$$

Рассмотрим теперь отдельно  $l$ -е неравенство из (3.151). Привлекая равенство (3.152), можем утверждать, что

$$\|z^{(k)}\|_{\infty} \leq \|z^{(k)}\|_{\infty} \sum_{j=1}^{l-1} \left| \frac{a_{lj}}{a_{ll}} \right| + \varkappa \|z^{(k-1)}\|_{\infty} \left( 1 - \sum_{j=1}^{l-1} \left| \frac{a_{lj}}{a_{ll}} \right| \right),$$

то есть

$$\|z^{(k)}\|_{\infty} \left( 1 - \sum_{j=1}^{l-1} \left| \frac{a_{lj}}{a_{ll}} \right| \right) \leq \varkappa \|z^{(k-1)}\|_{\infty} \left( 1 - \sum_{j=1}^{l-1} \left| \frac{a_{lj}}{a_{ll}} \right| \right). \quad (3.153)$$

Конечно, значение индекса  $l$ , на котором достигается равенство (3.152), может меняться в зависимости от номера итерации  $k$ . Но так как вплоть до оценки (3.151) мы отслеживали все компоненты погрешности  $z_i^{(k)}$ , то вне зависимости от  $k$  неравенство (3.153) должно быть справедливым для компоненты с номером  $l$ , определяемой условием (3.152).

Далее, в силу диагонального преобладания в матрице  $A$

$$1 - \sum_{j=1}^{l-1} \left| \frac{a_{lj}}{a_{ll}} \right| > 0,$$

и на эту положительную величину можно сократить обе части неравенства (3.153). Окончательно получаем

$$\|z^{(k)}\|_{\infty} \leq \varkappa \|z^{(k-1)}\|_{\infty},$$

что при  $|\varkappa| < 1$  означает сходимость метода Гаусса–Зейделя. ■

Фактически в доказательстве предложения 3.10.4 получена оценка уменьшения чебышёвской нормы погрешности решения с помощью «меры диагонального преобладания» в матрице СЛАУ, в качестве которой выступает величина  $\varkappa$ , определённая посредством (3.150).

**Теорема 3.10.5** *Если в системе линейных алгебраических уравнений  $Ax = b$  матрица  $A$  является симметричной и положительно определённой, то метод Гаусса–Зейделя сходится к решению этой системы из любого начального приближения.*

**Доказательство** может быть найдено, к примеру, в [4, 11]. Теорема 3.10.5 является частным случаем теоремы Островского–Райха (теорема 3.10.6), которая, в свою очередь, может быть получена как следствие из более общей теории итерационных методов, развитой А.А. Смарским. Её начала излагаются в § 3.13. ■

Метод Гаусса–Зейделя был сконструирован как модификация метода Якоби и, казалось бы, должен работать лучше. Так оно и есть «в среднем», и на случайно выбранных системах метод Гаусса–Зейделя работает несколько быстрее, что можно показать математически строго при естественных допущениях на систему. В частности, известная теорема Штейна–Розенберга [133] утверждает, что методы Якоби и Гаусса–Зейделя одновременно сходятся или одновременно расходятся для систем с матрицами специального вида, имеющим на главной диагонали неотрицательные элементы, а вне диагонали — отрицательные. Но при этом в случае сходимости метод Гаусса–Зейделя является более быстрым.

В целом же ситуация не столь однозначна. Для СЛАУ размера  $3 \times 3$  и более существуют примеры, на которых метод Якоби расходится, а метод Гаусса–Зейделя сходится. В частности, для метода Якоби неверна теорема 3.10.5, и он может расходиться для систем линейных уравнений с симметричными положительно определёнными матрицами (пример 3.10.1). Но существуют и примеры другого свойства, когда метод Якоби сходится, а метод Гаусса–Зейделя расходится.

По поводу практического применения метода Гаусса–Зейделя можно сказать почти то же самое, что и о методе Якоби в § 3.10д. Для решения систем линейных алгебраических уравнений он используется в настоящее время нечасто, но его идея не утратила своего значения и успешно применяется при построении различных итерационных процессов для решения линейных и нелинейных систем уравнений [83, 95]. Очень большое значение приобрёл интервальный метод Гаусса–Зейделя, предназначенный для внешнего оценивания множеств решений интервальных систем линейных уравнений (см. § 4.7). Наконец, в последние десятилетия XX века основная идея метода Гаусса–Зейделя вкупе с интервальными вычислениями оригинально претворилась в так называемых методах распространения ограничений, в общем и чрезвычайно мощном подходе к решению систем уравнений, неравенств и различных других соотношений.<sup>27</sup>

<sup>27</sup>Их англоязычное название — constraint propagation methods [115].

Метод Гаусса–Зейделя и рассматриваемый ниже метод релаксации можно применять для решения систем линейных алгебраических уравнений в операторной форме. В частности, в методе Гаусса–Зейделя для системы (3.77), которая возникает при решении двумерного уравнения Лапласа, очередное приближение удобно пересчитывается на основе значений функции в четырёх соседних узлах. Этот метод известен с 20-х годов XX века под именем *метода Либмана*.

### 3.10ж Методы релаксации

Одним из принципов, который кладётся в основу итерационных методов решения систем уравнений, является так называемый *принцип релаксации*.<sup>28</sup> Он понимается как специальная организация итераций, при которой на каждом шаге процесса уменьшается какая-либо величина, характеризующая погрешность очередного приближения  $x^{(k)}$  к решению системы.

Поскольку само решение  $x^*$  нам неизвестно, то оценить напрямую погрешность  $(x^{(k)} - x^*)$  не представляется возможным. По этой причине о степени близости  $x^{(k)}$  к  $x^*$  судят на основании каких-то косвенных признаков. Важнейшим из них является величина *невязки*, которая определяется как разность левой и правой частей уравнения после подстановки в него приближения к решению. В нашем случае она равна  $Ax^{(k)} - b$ . Конкретное применение принципа релаксации может заключаться в том, что на каждом шаге итерационного процесса стремятся уменьшить абсолютные значения компонент вектора невязки либо её норму, либо какую-то зависящую от них величину. В этом смысле методы Якоби и Гаусса–Зейделя можно рассматривать как итерационные процессы, в которых также осуществляется релаксация: на каждом их шаге компоненты очередного приближения вычисляются из условия зануления соответствующих компонент невязки на основе уже полученной информации о решении. Правда, это делается «локально», для отдельно взятой компоненты, а также без учёта её влияния на другие компоненты невязки.

Различают релаксацию *полную* и *неполную*, в зависимости от того, добиваемся ли мы на каждом отдельном шаге итерационного процесса (или его подшаге) наибольшего возможного улучшения рассматриваемой функции от погрешности или нет. Локально полная релаксация

---

<sup>28</sup>От латинского слова «relaxatio» — уменьшение напряжения, ослабление.

может казаться наиболее выгодной, но глобально, с точки зрения сходимости процесса в целом, тщательно подобранные неполная релаксации нередко приводят к более эффективным методам.

Популярной реализацией высказанных выше общих идей является итерационный метод решения систем линейных алгебраических уравнений, в котором для улучшения сходимости берётся «взвешенное среднее» значений компонент предшествующей  $x^{(k-1)}$  и текущей  $x^{(k)}$  итераций метода Гаусса–Зейделя. Более точно, зададимся вещественным числом  $\omega$ , которое назовём *параметром релаксации*, и  $i$ -ю компоненту очередного  $k$ -го приближения положим равной

$$\omega x_i^{(k)} + (1 - \omega)x_i^{(k-1)},$$

где  $x_i^{(k-1)}$  —  $i$ -я компонента приближения  $x^{(k-1)}$ , полученного в результате  $(k-1)$ -го шага алгоритма, а  $x_i^{(k)}$  —  $i$ -я компонента приближения, которое было бы получено на основе значений компонент  $x_1^{(k)}, \dots, x_{i-1}^{(k)}, x_{i+1}^{(k-1)}, \dots, x_n^{(k-1)}$  с помощью расчётной формулы метода Гаусса–Зейделя. Псевдокод получающегося итерационного алгоритма, который обычно называют *методом релаксации* для решения систем линейных алгебраических уравнений, представлен в табл. 3.10.

Расчётные формулы этого метода можно переписать в виде

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

$$i = 1, 2, \dots, n, \quad k = 1, 2, \dots$$

Далее, используя введённые выше в § 3.10e матрицы  $\tilde{L}$ ,  $D$  и  $\tilde{U}$ , можно придать этим соотношениям более компактный вид

$$(D + \omega \tilde{L}) x^{(k)} = ((1 - \omega)D - \omega \tilde{U}) x^{(k-1)} + \omega b,$$

откуда

$$x^{(k)} = (D + \omega \tilde{L})^{-1} ((1 - \omega)D - \omega \tilde{U}) x^{(k-1)} + (D + \omega \tilde{L})^{-1} \omega b, \quad k = 1, 2, \dots$$

В зависимости от конкретного значения параметра релаксации принято различать три случая:

если  $\omega < 1$ , то говорят о «нижней релаксации»,

Таблица 3.10. Метод релаксации для решения систем линейных алгебраических уравнений

```

 $k \leftarrow 1;$ 
выбираем начальное приближение  $x^{(0)}$ ;
DO WHILE ( метод не сопшлся )
    DO FOR  $i = 1$  TO  $n$ 
         $x_i^{(k)} \leftarrow (1 - \omega) x_i^{(k-1)}$ 
         $+ \frac{\omega}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^n a_{ij} x_j^{(k-1)} \right)$ 
    END DO
     $k \leftarrow k + 1;$ 
END DO

```

если  $\omega = 1$ , то имеем итерации Гаусса–Зейделя,  
если  $\omega > 1$ , то говорят о «верхней релаксации».<sup>29</sup>

Различные варианты методов релаксации развивались с переменным успехом с начала XX века. В 1950 году Д.М. Янг дал их строгий анализ, предложил эффективную процедуру выбора параметра релаксации  $\omega$  и привлек внимание к случаю  $\omega > 1$ , который во многих ситуациях действительно обеспечивает существенное ускорение сходимости итераций в сравнении с методом Гаусса–Зейделя. Несколько упрощённое объяснение этого явления может состоять в том, что если направление от  $x^{(k-1)}$  к  $x^{(k)}$  оказывается удачным в том смысле, что приближает к искомому решению, то имеет смысл пройти по нему и дальше, за  $x^{(k)}$ . Это и соответствует случаю  $\omega > 1$ .

Отметим, что метод релаксации тоже укладывается в изложенную

---

<sup>29</sup>В англоязычной литературе по вычислительной линейной алгебре этот метод обычно обозначают аббревиатурой SOR( $\omega$ ), которая происходит от термина «Successive OverRelaxation» — «последовательная перерелаксация». Популярен также симметричный метод релаксации, английская аббревиатура которого — SSOR, от Symmetric Successive OverRelaxation.

ранее в § 3.10в схему итерационных процессов, порождаемых расщеплением матрицы системы уравнений. При этом используется представление  $A = G_\omega - H_\omega$  с матрицами

$$G_\omega = D + \omega \tilde{L}, \quad H_\omega = (1 - \omega)D - \omega \tilde{U}.$$

Необходимое и достаточное условие сходимости метода релаксации из любого начального приближения принимает поэту вид

$$\rho(G_\omega^{-1} H_\omega) < 1.$$

Для некоторых классов специальных, но важных задач математической физики значение релаксационного параметра  $\omega$ , при котором величина  $\rho(G_\omega^{-1} H_\omega)$  достигает минимума или близка к нему, находится относительно просто. В более сложных задачах для оптимизации  $\omega$  требуется весьма трудный анализ спектра матрицы перехода  $G_\omega^{-1} H_\omega$  из представления (3.137). Обзоры состояния дел в этой области читатель может найти в [35, 49, 53, 69, 94, 128, 133].

**Предложение 3.10.3** (лемма Кэхэна) *Пусть рассматривается метод релаксации с параметром  $\omega$ , так что матрицей оператора перехода является*

$$C_\omega = (D + \omega \tilde{L})^{-1} ((1 - \omega)D - \omega \tilde{U}).$$

*Тогда  $\rho(C_\omega) \geq |\omega - 1|$ , и потому для сходимости метода релаксации из любого начального приближения необходимо выполнение неравенства  $0 < \omega < 2$ .*

**Доказательство.** Преобразуем матрицу  $C_\omega$ , чтобы придать ей вид, более удобный для дальнейших выкладок:

$$\begin{aligned} C_\omega &= (D + \omega \tilde{L})^{-1} ((1 - \omega)D - \omega \tilde{U}) = \\ &= (D(I + \omega D^{-1} \tilde{L}))^{-1} ((1 - \omega)D - \omega \tilde{U}) = \\ &= (I + \omega D^{-1} \tilde{L})^{-1} D^{-1} ((1 - \omega)D - \omega \tilde{U}) = \\ &= (I + \omega D^{-1} \tilde{L})^{-1} ((1 - \omega)I - \omega D^{-1} \tilde{U}). \end{aligned}$$

Желая исследовать расположение собственных чисел  $\lambda_i(C_\omega)$  матрицы перехода  $C_\omega$ , рассмотрим её характеристический полином

$$\begin{aligned}\phi(\lambda) &= \det(C_\omega - \lambda I) = \\ &= \det\left((I + \omega D^{-1}\tilde{L})^{-1}((1 - \omega)I - \omega D^{-1}\tilde{U}) - \lambda I\right) = \\ &= p_n\lambda^n + p_{n-1}\lambda^{n-1} + \dots + p_1\lambda + p_0,\end{aligned}$$

в котором  $p_n = (-1)^n$  по построению. Свободный член  $p_0$  характеристического полинома может быть найден как  $\phi(0)$ :

$$\begin{aligned}p_0 &= \det C_\omega = \det\left((I + \omega D^{-1}\tilde{L})^{-1}((1 - \omega)I - \omega D^{-1}\tilde{U})\right) = \\ &= (\det(I + \omega D^{-1}\tilde{L}))^{-1} \cdot \det((1 - \omega)I - \omega D^{-1}\tilde{U}) = \\ &= \det((1 - \omega)I - \omega D^{-1}\tilde{U}) = (1 - \omega)^n,\end{aligned}$$

коль скоро матрица  $(I + \omega D^{-1}\tilde{L})$  — нижняя треугольная и диагональными элементами имеет единицы, а  $((1 - \omega)I - \omega D^{-1}\tilde{U})$  — верхняя треугольная с элементами  $(1 - \omega)$  по главной диагонали.

С другой стороны, согласно теореме Виета (формулам Виета) свободный член характеристического полинома матрицы, делённый на старший коэффициент, равен произведению нулей полинома, умноженному на  $(-1)^n$  [7, 23, 43]. Но нули характеристического полинома являются собственными значениями матрицы, и поэтому

$$\prod_{i=1}^n \lambda_i(C_\omega) = (1 - \omega)^n.$$

Отсюда необходимо следует

$$\max_{1 \leq i \leq n} |\lambda_i(C_\omega)| \geq |\omega - 1|,$$

так как в противном случае произведение всех собственных чисел было бы меньшим единицы.

Если  $\omega \notin ]0, 2[$ , то спектральный радиус матрицы перехода заведомо не меньше единицы, а потому в силу теоремы 3.10.1 итерационный метод не может сходиться из любого начального приближения. ■

Лемма Кэхэна даёт лишь необходимое условие сходимости метода релаксации, которое иногда не является достаточным. Это показывает следующий

**Пример 3.10.1** Решением системы линейных алгебраических уравнений

$$\begin{pmatrix} 2 & 1 & -1 \\ 1 & -5 & 4 \\ 3 & 2 & 6 \end{pmatrix} x = \begin{pmatrix} 0 \\ 10 \\ 7 \end{pmatrix} \quad (3.154)$$

является  $(1, -1, 1)^\top$ . Метод Гаусса–Зейделя за 126 итераций сходится из нулевого вектора к приближённому решению этой системы с невязкой  $10^{-6}$  в 1-норме.

Для метода релаксации оптимальное значение параметра находится в районе  $\omega = 0.8$ , и тогда для достижения той же точности приближённого решения требуется всего 15 итераций. Но с любым параметром  $\omega \gtrapprox 1.035598$  метод релаксации для системы (3.154) расходится. ■

Один из популярных вариантов необходимых и достаточных условий сходимости метода релаксации —

**Теорема 3.10.6** (теорема Островского–Райха) *Пусть  $A$  — вещественная симметричная матрица с положительными диагональными элементами и  $\omega \in ]0, 2[$ . Метод релаксации с параметром  $\omega$  сходится к решению системы линейных алгебраических уравнений  $Ax = b$  с произвольной правой частью и из любого начального приближения тогда и только тогда, когда матрица  $A$  положительно определена.*

**Доказательство** опускается. Читатель может найти его, к примеру, в книгах [13, 128, 133]. Обоснование теоремы Островского–Райха будет также представлено ниже в § 3.13 как следствие теоремы Самарского, дающей достаточные условия сходимости для итерационных методов весьма общего вида.

**Пример 3.10.2** У системы линейных алгебраических уравнений

$$\begin{pmatrix} 6 & 2 & 2 \\ 2 & 3 & 4 \\ 2 & 4 & 8 \end{pmatrix} x = \begin{pmatrix} 6 \\ 3 \\ 6 \end{pmatrix} \quad (3.155)$$

матрица симметрична и положительно определена, а точное решение есть  $(1, -1, 1)^\top$ . Применим к этой системе все рассмотренные выше стационарные итерационные процессы.

Первым тестируем простейший итерационный метод Ричардсона из § 3.10г. Если считать точно известными границы спектра матрицы системы, то оптимальное значение параметра предобусловливания оказывается равным  $\tau = 0.1638$ . Для него метод Ричардсона сходится из нулевого начального вектора за 148 шагов к приближению, которое обеспечивает невязку  $10^{-6}$  в 1-норме.

Метод Якоби для системы (3.155) расходится из любого начального приближения, отличного от самого решения.

Метод Гаусса–Зейделя для системы (3.155) сходится из любого начального приближения. Для достижения 1-нормы невязки приближённого решения, меньшей  $10^{-6}$ , ему потребовалось сделать из нулевого начального приближения 42 шага.

Методу релаксации с параметром  $\omega = 1.288$  для достижения того же результата потребовалось 16 шагов, т. е. ещё в два с половиной раза меньше. Оптимальность этого значения  $\omega$  нетрудно обнаружить с помощью бисекции интервала  $[0, 2]$  и тестовых прогонов с различными  $\omega$ , хотя на практике подобный выбор параметра релаксации часто нецелесообразен. По этой причине имеет смысл рассмотреть другие близкие значения  $\omega$ .

Для достижения 1-нормы невязки приближённого решения, меньшей  $10^{-6}$ , методу релаксации при  $\omega = 1.1$  потребовалось 33 итерации, при  $\omega = 1.2$  потребовалось 25 итераций, при  $\omega = 1.3$  потребовалось 18 итераций, а при  $\omega = 1.4$  — 21 итерация. Таким образом, в окрестности минимума зависимость количества итераций от параметра  $\omega$  изменяется слабо. ■

**Пример 3.10.3** Рассмотрим решение  $8 \times 8$ -системы линейных алгебраических уравнений с гильбертовой матрицей (см. § 3.4б). В качестве правой части возьмём вектор  $(1, -1, 1, -1, 1, -1, 1, -1, 1, -1)^\top$ .

Так как при вводе чисел в компьютер и дальнейших операциях над ними допускаются неизбежные погрешности, то матрица системы, строго говоря, не вполне совпадает с реальной гильбертовой матрицей, а является лишь «приближённо гильбертовой». Спектральное число обусловленности этой матрицы равно  $1.5 \cdot 10^{10}$ .

Несмотря на теоретический результат о сходимости (теорема 3.10.6), ни метод Гаусса–Зейделя, ни метод релаксации не справляются с этой плохообусловленной системой. Порождаемые ими последовательности приближений за миллионы итераций никуда не сходятся, а невязка приближённого решения системы не становится при этом меньше

какого-то существенного порога (порядка нескольких единиц). ■

## 3.11 Нестационарные итерационные методы для линейных систем

### 3.11а Теоретическое введение

В этом разделе для решения систем линейных алгебраических уравнений рассматриваются нестационарные итерационные методы, которые распространены не меньше стационарных. В основу нестационарных итерационных методов могут быть положены различные идеи, так что соответствующие вычислительные алгоритмы для решения СЛАУ отличаются огромным разнообразием. В нашем учебнике даётся лишь беглый обзор основных идей и подходов.

#### Обобщение метода Ричардсона

В качестве первого примера рассмотрим простейший итерационный процесс Ричардсона (3.138)

$$x^{(k)} \leftarrow (I - \tau A) x^{(k-1)} + \tau b, \quad k = 1, 2, \dots,$$

исследованный в § 3.10г. Если переписать его в виде

$$x^{(k)} \leftarrow x^{(k-1)} - \tau (Ax^{(k-1)} - b), \quad k = 1, 2, \dots, \quad (3.156)$$

то расчёт каждой следующей итерации  $x^{(k)}$  может трактоваться как вычитание из  $x^{(k-1)}$  поправки, пропорциональной вектору его невязки  $(Ax^{(k-1)} - b)$ . Но при таком взгляде на итерационный процесс можно изменять параметр  $\tau$  в зависимости от шага, т. е. взять  $\tau = \tau_k$  переменным и рассмотреть итерации

$$x^{(k)} \leftarrow x^{(k-1)} - \tau_k (Ax^{(k-1)} - b), \quad k = 1, 2, \dots \quad (3.157)$$

Этот простейший нестационарный итерационный метод тоже связывают с именем Л.Ф. Ричардсона, который рассмотрел его в работе [124]. Он, к сожалению, не смог развить удовлетворительной теории выбора параметров  $\tau_k$ , и для решения этого вопроса потребовалось ещё несколько десятилетий развития вычислительной математики. Отметим, что задача об оптимальном выборе параметров  $\tau_k$  на группе из

нескольких шагов приводит к так называемым чебышёвским циклическим итерационным методам. Подробную информацию о них читатель найдёт в книгах [40, 49, 93, 94].

Можно пойти по намеченному пути дальше, рассмотрев нестационарное обобщение итерационного процесса

$$x^{(k)} \leftarrow (I - \Lambda A) x^{(k-1)} + \Lambda b, \quad k = 1, 2, \dots,$$

который получен в результате матричного предобуславливания исходной системы линейных уравнений  $Ax = b$ . Если переписать его вычислительную схему в виде

$$x^{(k)} \leftarrow x^{(k-1)} - \Lambda(Ax^{(k-1)} - b), \quad k = 1, 2, \dots,$$

нетрудно увидеть возможность изменения предобуславливающей матрицы  $\Lambda$  в зависимости от номера шага. Таким образом, приходим к весьма общей схеме нестационарных линейных итерационных процессов

$$x^{(k)} \leftarrow x^{(k-1)} - \Lambda_k(Ax^{(k-1)} - b), \quad k = 1, 2, \dots,$$

где  $\{\Lambda_k\}_{k=1}^{\infty}$  — некоторая последовательность матриц. Выбор  $\{\Lambda_k\}$ , при котором этот процесс сходится, зависит, вообще говоря, от начального приближения  $x^{(0)}$ .

## Проекционные методы

Другой путь к построению нестационарных итерационных методов приводит к так называемым *проекционным методам*. В них искомое решение получается итогом последовательности приближений, которые строятся как решения «проекций» исходного уравнения на линейные подпространства меньшей размерности. В отношении проекционных методов нередко используется термин «метод моментов» [74, 77], но он не столь выразителен и менее удачен, в частности, потому, что перегружен другими смыслами (существует метод моментов в математической статистике и т. д.).

В § 2.11в уже обсуждалось использование проекции элемента линейного векторного пространства на его подпространство. Проекция является, в некотором роде, «наилучшим представлением» рассматриваемого элемента в данном подпространстве. По этой причине можно изучать свойства интересующего нас элемента не во всём «большом»

пространстве, а в меньшем подпространстве, что обычно проще и технически удобнее. Аналогичным образом можно строить «проекцию» уравнения (системы уравнений) в линейном пространстве и решать её в подпространстве с помощью более простых инструментов.

Пусть задано линейное операторное уравнение

$$\mathcal{A}x = b, \quad (3.158)$$

т. е. уравнение, в котором  $\mathcal{A}$  — линейный оператор, действующий из линейного пространства  $U$  в линейное пространство  $V$ . Необходимо найти его решение, т. е. такой элемент  $\tilde{x} \in U$  (или его достаточно хорошее приближение), что он превращает (3.158) в истинное равенство. Пространства  $U$  и  $V$  могут иметь, вообще говоря, произвольную природу, в частности быть бесконечномерными пространствами функций и т. п. Линейный оператор  $\mathcal{A}$  может быть, к примеру, оператором дифференцирования, оператором интегрирования и т. д. Соответственно, в описываемую общую схему укладываются широкие классы линейных дифференциальных уравнений, линейных интегральных уравнений, интегро-дифференциальные и другие важнейшие уравнения. Для построения приближённого решения уравнения (3.158) «спроектируем» его в подпространство (обычно меньшей размерности) и в нём построим решение «проектированного уравнения». Если эти подпространства имеют конечную размерность, то для получения «проектированных решений» могут быть применены известные численные методы для систем линейных уравнений и т. п. Опишем этот подход более подробно.

Пусть в  $U$  задано подпространство  $\mathcal{U}$ , в  $V$  — подпространство  $\mathcal{V}$ , и определён оператор проектирования (проектор)  $P : V \rightarrow \mathcal{V}$ . Для операторного уравнения (3.158) *проекционным уравнением* назовём операторное уравнение вида

$$P(\mathcal{A}x) = Pb$$

на подпространстве  $\mathcal{U} \subseteq U$ . Если  $\mathcal{U}$  «достаточно представительно» в  $U$ , то можно надеяться, что решение проекционного уравнения будет «достаточно близко» к решению исходного. Обычно в качестве  $\mathcal{U}$  и  $\mathcal{V}$  берут конечномерные подпространства в  $U$  и  $V$  соответственно, так что тогда проекционное уравнение тоже является операторным уравнением в конечномерных пространствах, т. е. фактически приводится к системе линейных алгебраических уравнений.

Проекционное уравнение можно переписать равносильным образом как

$$P(\mathcal{A}x - b) = 0. \quad (3.159)$$

При проверке этого условия на первый план выходят свойства проектора  $P : V \rightarrow \mathcal{V}$  и выполняемой им проекции. Наиболее простой и элегантной проверка равенства (3.159) становится в случае, когда  $V$  — пространство со скалярным произведением (в частности, гильбертово пространство), а  $P$  — оператор ортогонального проектирования на конечномерное подпространство  $\mathcal{V}$  из  $V$ . Тогда, взяв в  $\mathcal{V}$  какой-нибудь базис  $\{v_1, v_2, \dots, v_m\}$ , можем обеспечить равенство (3.159), если потребуем, чтобы

$$\langle \mathcal{A}x - b, v_i \rangle = 0, \quad i = 1, 2, \dots, m. \quad (3.160)$$

Эти условия называются *условиями Петрова–Галёркина* по именам механиков, впервые применивших их в первой половине XX века.

Если  $\mathcal{U}$  — конечномерное подпространство в  $U$ , которое имеет базис  $\{u_1, u_2, \dots, u_n\}$ , то любой элемент  $x$  из  $\mathcal{U}$  можно представить в виде

$$x = \sum_{j=1}^n x_j u_j$$

с какими-то коэффициентами  $x_j$ ,  $j = 1, 2, \dots, n$ . Подставляя это выражение в левую часть (3.160), получим

$$\begin{aligned} \left\langle \mathcal{A} \left( \sum_{j=1}^n x_j u_j \right) - b, v_i \right\rangle &= \left\langle \sum_{j=1}^n x_j (\mathcal{A}u_j) - b, v_i \right\rangle = \\ &= \sum_{j=1}^n \langle \mathcal{A}u_j, v_i \rangle x_j - \langle b, v_i \rangle \end{aligned}$$

для  $i = 1, 2, \dots, m$ . Условия Петрова–Галёркина (3.160) будут равносильны тогда системе линейных алгебраических уравнений

$$\sum_{j=1}^n \langle \mathcal{A}u_j, v_i \rangle x_j = \langle b, v_i \rangle, \quad i = 1, 2, \dots, m. \quad (3.161)$$

Важнейший частный случай рассмотренной конструкции проекционного уравнения относится к операторному уравнению  $\mathcal{A}x = b$ , в котором линейный оператор  $\mathcal{A}$  действует из пространства  $U$  в него же,

т. е.  $\mathcal{A} : U \rightarrow U$ . При этом на  $U$  предполагается заданным скалярное произведение. Пусть, как и ранее,  $\mathcal{U}$  — конечномерное линейное подпространство в  $U$ , имеющее базис  $\{u_1, u_2, \dots, u_n\}$ . Тогда условия (3.160) удовлетворения проекционному уравнению получают вид

$$\langle \mathcal{A}x - b, u_i \rangle = 0, \quad i = 1, 2, \dots, n.$$

Эти более специальные соотношения называют *условиями Галёркина* или *условиями Бубнова–Галёркина*.<sup>30</sup> Любой элемент из  $\mathcal{U}$  можно представить в виде разложения по базису

$$x = \sum_{j=1}^n x_j u_j.$$

Соответственно, для определения коэффициентов  $x_i$  разложения решения  $x$  проекционного уравнения нужно решить квадратную систему линейных алгебраических уравнений

$$\sum_{j=1}^n \langle \mathcal{A}u_j, u_i \rangle x_j = \langle b, u_i \rangle, \quad i = 1, 2, \dots, n,$$

которая аналогична (3.161).

Наибольшую выгоду от применения описанного выше перехода к проекционному уравнению можно получить при решении функциональных уравнений в бесконечномерных пространствах. Тогда для более точного приближения к решению обычно строят последовательность конечномерных подпространств  $\mathcal{V}_k \subset V$ ,  $k = 1, 2, \dots$ , и последовательность соответствующих проекционных уравнений, решения которых сходились бы к точному решению исходного операторного уравнения. Вообще, метод Галёркина и его модификации и адаптации являются популярнейшими численными методами для решения краевых задач для дифференциальных уравнений и решения интегральных уравнений. Фактически специальной разновидностью метода Галёркина является метод конечных элементов, успешно применяемый инженерами более столетия.

Но немалую пользу от проекционной схемы можно извлечь и при решении конечномерных линейных операторных уравнений, т. е. систем

---

<sup>30</sup>Инженер И.Г. Бубнов первым практически применял этот подход при расчётах конструкций кораблей, а чуть позже механик и математик Б.Г. Галёркин развил теорию и дальнейшие приложения метода.

линейных алгебраических уравнений. Обычно точное решение исходной системы при таком подходе получается как последний член последовательности приближённых решений проекционных уравнений, построенных по последовательности вложенных подпространств увеличивающейся размерности. Решение каждого предыдущего уравнения служит хорошим приближением к решению следующего, так что по существу получается нестационарный итерационный процесс.

В качестве последовательности расширяющихся подпространств в настоящее время наиболее часто используются так называемые подпространства Крылова, порождённые матрицей системы и её редуцированной правой частью (см. § 3.8). В последние десятилетия XX века было показано, что такой выбор обладает определёнными свойствами оптимальности [44, 110, 123], которые делают применение подпространств Крылова особенно выгодным. С другой стороны, многие популярные алгоритмы вычислительной линейной алгебры (в частности, метод сопряжённых градиентов, см. § 3.11д) можно представить как проекционные методы на подпространства Крылова. Детальное изложение современного состояния численных методов решения СЛАУ, основанных на использовании подпространств Крылова, даётся в книге [39].

## Вариационные итерационные методы

*Вариационными методами* называют методы решения математических задач, в которых исходная постановка сводится к задаче нахождение экстремума, минимума или максимума. Вариационные итерационные методы для решения уравнений и систем уравнений — это итерационные методы, сконструированные как методы нахождения некоторых экстремумов, по которым определяется решение исходной задачи. Они обычно основаны на использовании так называемых вариационных принципов.

В свою очередь, *вариационными принципами* называют переформулировки интересующих нас задач в виде каких-либо оптимизационных задач, т. е. задач на нахождение минимумов или максимумов. Они получаются весьма различными способами, и некоторые из них вытекают из содержательного (физического, механического и пр.) смысла решаемой задачи. Например, в классической механике хорошо известны «принцип наименьшего действия Лагранжа», «принцип наименьшего действия Гамильтона» (или Гамильтона–Остроградского), в оптике существует «принцип Ферма» [84]. В последнее столетие имеется тен-

денция всё меньше связывать вариационные принципы с конкретным физическим содержанием, и они становятся абстрактным математическим инструментом решения разнообразных задач.

Интуитивно понятный термин «вариация» был введён в математику Ж.-Л. Лагранжем для обозначения малого изменения («шевеления») независимой переменной. Если в рассматриваемой точке эти вариации приводят к изменениям значений функции, знак которых одинаков, то имеем экстремум. И наоборот, если некоторые вариации аргумента приводят к возрастанию функции, а другие — к убыванию, то экстремума в этой точке нет. Соответственно, метод исследования минимумов и максимумов, основанный на изучении зависимости функций от вариаций их аргументов, получил название *метода вариаций*.

Строго говоря, в вычислительном отношении оптимизационная задача, получающаяся в результате применения вариационного принципа, может быть не вполне эквивалентна исходной. В частности, задача нахождения устойчивого решения уравнения может превратиться в неустойчивую задачу о проверке точного равенства экстремума нулю (этот вопрос более подробно обсуждается далее в § 4.2б). Но если существование решения уравнения известно априори, до того, как мы приступаем к его нахождению (например, на основе каких-либо теорем существования), то вариационные методы становятся нашим важным подспорьем. Именно такова ситуация с системами линейных алгебраических уравнений, разрешимость которых часто обеспечивается различными результатами из линейной алгебры.

Наконец, вариационные принципы естественно возникают в случаях, когда по смыслу практической задачи необходимо найти не обычное решение уравнения (или системы уравнений), а какое-либо псевдорешение. Само его определение которого требует минимизации различия левой и правой частей уравнения.

Как именно можно переформулировать задачу решения СЛАУ в виде оптимизационной задачи? Самый простой и естественный способ может основываться на том факте, что точное решение  $x^*$  зануляет норму невязки  $\|Ax - b\|$ , доставляя ей наименьшее возможное значение:

решение  $Ax = b$

$\iff$

нахождение  $\min \|Ax - b\|$

Рассматривая конкретные векторные нормы, получаем различные ва-

риационные переформулировки задачи решения системы линейных алгебраических уравнений.

Выбор той или иной нормы в этой конструкции существенно влияет на свойства получающейся оптимизационной задачи. На практике чаще всего используют евклидову норму вектора невязки, которая порождена скалярным произведением и обладает рядом других хороших свойств. Наконец, желая иметь глобальную гладкость получаемого функционала по неизвестной переменной  $x$  и избавиться от операции взятия корня, обычно берут квадрат евклидовой нормы, т. е. скалярное произведение  $\langle Ax - b, Ax - b \rangle$ . Получающаяся задача минимизации величины  $\|Ax - b\|_2^2$  называется *линейной задачей наименьших квадратов*. Мы уже касались её в § 2.11г и рассмотрим подробнее в § 3.16.

Ещё одним фактом, который служит теоретической основой для вариационных методов решения систем линейных алгебраических уравнений является

**Теорема 3.11.1** Вектор  $x^* \in \mathbb{R}^n$  является решением системы линейных алгебраических уравнений  $Ax = b$  с симметричной положительно определённой матрицей  $A$  тогда и только тогда, когда он доставляет минимум функционалу  $\Psi(x) = \frac{1}{2}\langle Ax, x \rangle - \langle b, x \rangle$ .

Иными словами,

решение  $Ax = b$

$\iff$

нахождение  $\min \Psi(x)$

**Доказательство.** Если  $A$  — симметричная положительно определённая матрица, то решение  $x^*$  системы линейных уравнений  $Ax = b$  существует и единственno. Другим следствием симметричности и положительной определённости  $A$  является то, что она порождает (см. § 3.3а) энергетическую норму  $\|\cdot\|_A$  векторов из  $\mathbb{R}^n$ :

$$\|x\|_A = \sqrt{\langle Ax, x \rangle}.$$

В этой норме мы будем рассматривать погрешность приближений к решению системы.

Из единственности  $x^*$  следует, что некоторый вектор  $\tilde{x} \in \mathbb{R}^n$  является решением системы уравнений тогда и только тогда, когда  $\tilde{x} - x^* = 0$ .

В свою очередь, это равносильно занулению нормы погрешности, т. е.  $\|\tilde{x} - x^*\|_A = 0$ , что можно переформулировать также в следующем виде:

$$\tilde{x} \text{ есть решение системы } Ax = b \iff \frac{1}{2} \|\tilde{x} - x^*\|_A^2 = 0.$$

Преобразуем равенство из правой части этой эквивалентности, учитывая симметричность матрицы  $A$ , равенство  $Ax^* = b$  и определение энергетической нормы:

$$\begin{aligned} \frac{1}{2} \|\tilde{x} - x^*\|_A^2 &= \frac{1}{2} \langle A(\tilde{x} - x^*), \tilde{x} - x^* \rangle = \\ &= \frac{1}{2} \langle A\tilde{x}, \tilde{x} \rangle - \frac{1}{2} \langle A\tilde{x}, x^* \rangle - \frac{1}{2} \langle Ax^*, \tilde{x} \rangle + \frac{1}{2} \langle Ax^*, x^* \rangle = \\ &= \frac{1}{2} \langle A\tilde{x}, \tilde{x} \rangle - \frac{1}{2} \langle Ax^*, \tilde{x} \rangle - \frac{1}{2} \langle Ax^*, \tilde{x} \rangle + \frac{1}{2} \|x^*\|_A^2 = \\ &= \frac{1}{2} \langle A\tilde{x}, \tilde{x} \rangle - \langle b, \tilde{x} \rangle + \frac{1}{2} \|x^*\|_A^2 = \\ &= \Psi(\tilde{x}) + \frac{1}{2} \|x^*\|_A^2. \end{aligned}$$

Результат выкладки показывает, что функционал  $\Psi(x)$  отличается от половины квадрата энергетической нормы погрешности приближённого решения лишь постоянным слагаемым  $\frac{1}{2} \|x^*\|_A^2$  (которое, вообще говоря, неизвестно из-за незнания нами  $x^*$ ). Как следствие,  $\Psi(x)$  действительно достигает своего единственного минимума при том же значении аргумента, что и  $\|x - x^*\|_A^2$ , т. е. на точном решении  $x^*$  рассматриваемой линейной системы. ■

Функционал  $\Psi(x) = \frac{1}{2} \langle Ax, x \rangle - \langle b, x \rangle$ , который является квадратичной формой от вектора переменных  $x$ , обычно называют *функционалом энергии* из-за его сходства с выражениями для различных видов энергии в физических системах. К примеру, кинетическая энергия тела массы  $m$ , движущегося со скоростью  $v$ , равна  $\frac{1}{2}mv^2$ . Энергия упругой деформации пружины с жёсткостью  $k$ , растянутой или сжатой на величину  $x$ , равна  $\frac{1}{2}kx^2$  и т. д. Для более сложных физических систем, образованных из нескольких составных частей, квадрат одной переменной в этих выражениях заменяется на квадратичную форму от нескольких переменных. Естественность присутствия множителя  $\frac{1}{2}$  в выражении для  $\Psi(x)$  получит также дополнительное обоснование далее в разделе § 3.11в.

Равенство

$$\Psi(x) = \frac{1}{2} \|x - x^*\|_A^2 - \frac{1}{2} \|x^*\|_A^2 \quad (3.162)$$

— отдельное важное следствие доказательства теоремы 3.11.1. Оно показывает, как отмечалось, что функционал энергии лишь на константу отличается от энергетической  $A$ -нормы погрешности приближения к решению. Этот факт будет неоднократно использоваться далее.

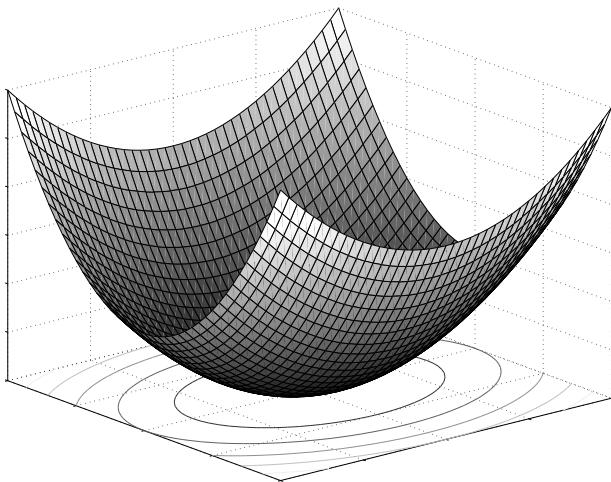


Рис. 3.27. Типичный график функционала энергии и его линии уровня

Поскольку  $A$  — симметричная матрица, ортогональным преобразованием подобия она может быть приведена к диагональной матрице:

$$A = Q^T D Q,$$

где  $Q$  ортогональна,  $D = \text{diag} \{ \lambda_1, \lambda_2, \dots, \lambda_n \}$ , а  $\lambda_i$  — собственные значения матрицы  $A$ , причём все  $\lambda_i > 0$  в силу положительной определённости  $A$ . Подставляя это выражение в функционал энергии  $\Psi(x)$ , будем иметь

$$\begin{aligned} \Psi(x) &= \frac{1}{2} \langle Q^T D Q x, x \rangle - \langle b, x \rangle = \frac{1}{2} \langle D(Qx), Qx \rangle - \langle Qb, Qx \rangle = \\ &= \frac{1}{2} \langle Dy, y \rangle - \langle Qb, y \rangle = \frac{1}{2} \sum_{i=1}^n \lambda_i y_i^2 - \sum_{i=1}^n (Qb)_i y_i, \end{aligned} \quad (3.163)$$

где обозначено  $y = Qx$ .

Итак, в изменённой системе координат, которая получается с помощью ортогонального линейного преобразования переменных, выражение для функционала энергии  $\Psi(x)$  есть половина суммы квадратов переменных с коэффициентами, равными собственным значениям матрицы  $A$ , минус линейные члены. Поэтому график функционала энергии — эллиптический параболоид, возможно, сдвинутый относительно начала координат и ещё повёрнутый. Его поверхности уровня (линии уровня в двумерном случае) — эллипсоиды (эллипсы), в центре которых находится искомое решение системы уравнений. При этом форма эллипсоидов уровня находится в зависимости от разброса коэффициентов при квадратах переменных в выражении (3.163), т. е. от спектрального числа обусловленности матрицы  $A$  согласно формуле (3.58). Чем больше эта обусловленность, тем сильнее сплющены эллипсоиды уровня, так что для плохо обусловленных СЛАУ решение находится на дне длинного и узкого «оврага».

### 3.116 Метод спуска для минимизации функций

В предшествующем разделе были предложены две вариационные переформулировки задачи решения системы линейных алгебраических уравнений. Как находить минимум соответствующих функционалов? Прежде чем строить конкретные численные алгоритмы, рассмотрим общую схему.

Пусть  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  — некоторая функция, ограниченная снизу на всём пространстве  $\mathbb{R}^n$  и принимающая своё наименьшее значение в  $x^*$ , так что

$$f(x) \geq f(x^*) = \min_{x \in \mathbb{R}^n} f(x) \quad \text{для любых } x \in \mathbb{R}^n.$$

Нам нужно найти точку  $x^*$ . При этом саму функцию  $f$ , для которой ищется экстремум, в теории оптимизации называют *целевой функцией*.

Различают экстремумы *локальные* и *глобальные*. Локальными называют экстремумы, в которых значения целевой функции лучше, чем в некоторой окрестности рассматриваемой точки. Глобальные экстремумы доставляют функции значения, лучшие среди значений функции на всей её области определения. В связи с задачей минимизации функционала энергии нас интересуют, конечно, его глобальные минимумы.

Типичным подходом к решению задач оптимизации является итерационное построение последовательности значений аргумента  $\{x^{(k)}\}$ ,

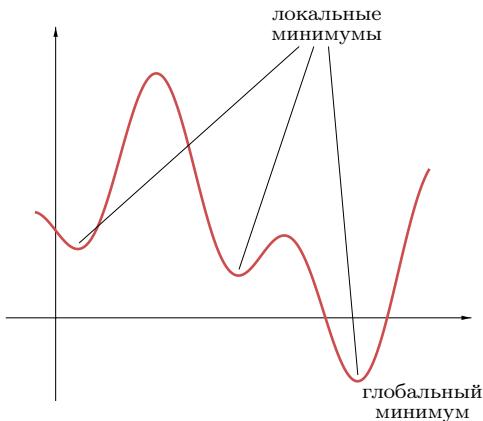


Рис. 3.28. Глобальные и локальные минимумы функции

которая «минимизирует» функцию  $f$  в том смысле, что

$$\lim_{k \rightarrow \infty} f(x^{(k)}) = \min_{x \in \mathbb{R}^n} f(x).$$

Если построенная последовательность  $\{x^{(k)}\}$  сходится к некоторому пределу, то он и является решением задачи  $x^*$  в случае непрерывной функции  $f$ .

Одним из популярных методов построения минимизирующей последовательности для широких классов целевых функций является *метод спуска*, который заключается в следующем. Пусть в результате выполнения  $k - 1$  шагов метода,  $k = 1, 2, \dots$ , уже найдено какое-то приближение  $x^{(k-1)}$  к точке минимума функции  $f(x)$ . Далее на  $k$ -ом шаге

- выбираем направление  $s^{(k)}$ , в котором целевая функция убывает в точке  $x^{(k-1)}$ ,
- назначаем величину шага  $\tau_k$ , на который сдвигаемся в выбранном направлении, полагая

$$x^{(k)} \leftarrow x^{(k-1)} + \tau_k s^{(k)}.$$

Конкретное значение  $\tau_k$  находится из условия уменьшения целевой функции, т. е. так, чтобы  $f(x^{(k)}) < f(x^{(k-1)})$ .

Далее мы можем повторить этот шаг ещё раз и ещё ... столько, сколько нужно для достижения желаемого приближения к минимуму.

Выбор как направления спуска  $s^{(k)}$ , так и величины шага  $\tau_k$  является очень ответственным делом, так как от них зависит и наличие сходимости, и её скорость. Как правило, спуск по подходящему направлению обеспечивает убывание целевой функции лишь при достаточно малых шагах, и потому при неудачно большой величине шага мы можем попасть в точку, где значение функционала не меньше, чем в текущей точке. С другой стороны, слишком малый шаг приведёт к очень медленному движению к решению задачи.

Если целевая функция имеет более одного локального экстремума, то метод спуска может сходиться к какому-нибудь одному из них, который не обязательно является глобальным. Тогда предпринимают многократный запуск метода спуска из различных начальных точек («мультистарт»), применяют другие модификации, которые помогают преодолеть локальный характер метода спуска. Но в случае минимизации функционала энергии  $\Psi(x)$ , порождаемого системой линейных алгебраических уравнений с симметричной положительно определённой матрицей, подобный феномен, к счастью, случиться не может. Свойства функционала  $\Psi(x)$  достаточно хороши: он имеет один локальный минимум, который одновременно и глобален.

Пусть на очередном  $k$ -ом шаге метода спуска зафиксировано направление  $s$ . Определим шаг спуска по этому направлению так, чтобы он обеспечивал наилучшую возможную минимизацию функционала энергии  $\Psi(x)$ . Далее мы будем называть подобный вариант рассматриваемого метода *наискорейшим спуском*, уточняя при необходимости направление этого спуска.

Для определения  $\tau_k$  подставим  $x^{(k-1)} + \tau_k s$  в аргумент функционала энергии и продифференцируем получившееся выражение по  $\tau_k$ . После подстановки будем иметь

$$\begin{aligned} \Psi(x^{(k-1)} + \tau_k s) &= \frac{1}{2} \langle A(x^{(k-1)} + \tau_k s), x^{(k-1)} + \tau_k s \rangle - \langle b, x^{(k-1)} + \tau_k s \rangle = \\ &= \frac{1}{2} \langle Ax^{(k-1)}, x^{(k-1)} \rangle + \tau_k \langle Ax^{(k-1)}, s \rangle + \frac{1}{2} \tau_k^2 \langle As, s \rangle - \\ &\quad - \langle b, x^{(k-1)} \rangle - \tau_k \langle b, s \rangle. \end{aligned}$$

При дифференцировании выписанного выражения по  $\tau_k$  не зависящие

от него члены исчезнут, и мы получим

$$\begin{aligned} \frac{d}{d\tau_k} \Psi(x^{(k-1)} + \tau_k s) &= \langle Ax^{(k-1)}, s \rangle + \tau_k \langle As, s \rangle - \langle b, s \rangle = \\ &= \tau_k \langle As, s \rangle + \langle Ax^{(k-1)} - b, s \rangle = \\ &= \tau_k \langle As, s \rangle + \langle r^{(k-1)}, s \rangle, \end{aligned}$$

где обозначено  $r^{(k-1)} := Ax^{(k-1)} - b$  — невязка используемого приближения к решению. Таким образом, в точке экстремума по  $\tau_k$  условие

$$\frac{d}{d\tau_k} \Psi(x^{(k-1)} + \tau_k s) = 0$$

необходимо влечёт

$$\tau_k = -\frac{\langle r^{(k-1)}, s \rangle}{\langle As, s \rangle}. \quad (3.164)$$

Легко видеть, что при найденном значении  $\tau_k$  функционалом энергии действительно достигается минимум по выбранному направлению спуска. Это следует из того, что вторая производная по  $\tau_k$  равна

$$\frac{d^2}{d\tau_k^2} \Psi(x^{(k-1)} + \tau_k s) = \langle As, s \rangle > 0$$

в силу положительной определённости матрицы  $A$ .

### 3.11в Наискорейший градиентный спуск

Если целевая функция  $f(x)$  дифференцируема, то, как известно, направление её наибольшего убывания в точке  $x^{(k-1)}$  противоположно направлению вектора градиента

$$\nabla f(x) := f'(x^{(k-1)}) = (f'_1(x^{(k-1)}), f'_2(x^{(k-1)}), \dots, f'_n(x^{(k-1)}))^\top.$$

Этот факт служит основой одного из популярных вариантов метода спуска для минимизации функций — *метода градиентного спуска*, в котором направлением спуска берётся «антиградиент», т. е. вектор  $-f'(x^{(k)})$ , а очередной шаг выполняется по формуле

$$x^{(k)} \leftarrow x^{(k-1)} - \tau_k f'(x^{(k-1)}) \quad (3.165)$$

для некоторого  $\tau_k \in \mathbb{R}$ . Какой вид имеет градиентный спуск для минимизации функционала энергии  $\Psi(x)$ ?

Чтобы найти градиент функционала энергии, вычислим его частные производные:

$$\frac{\partial \Psi(x)}{\partial x_l} = \frac{\partial}{\partial x_l} \left( \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j - \sum_{i=1}^n b_i x_i \right) = \sum_{j=1}^n a_{lj} x_j - b_l,$$

$l = 1, 2, \dots, n$ . Множитель  $1/2$  исчезает в результате потому, что в двойной сумме помимо квадратичных слагаемых  $a_{ii}x_i^2$  остальные слагаемые присутствуют парами, как  $a_{ij}x_i x_j$  и  $a_{ji}x_j x_i$ , причём  $a_{ij} = a_{ji}$ . В целом

$$\Psi'(x) = \left( \frac{\partial \Psi(x)}{\partial x_1}, \frac{\partial \Psi(x)}{\partial x_2}, \dots, \frac{\partial \Psi(x)}{\partial x_n} \right)^T = Ax - b, \quad (3.166)$$

т. е. градиент функционала  $\Psi$  равен невязке решаемой системы линейных уравнений в рассматриваемой точке.

Важнейшим выводом из полученного результата является тот факт, что итерационный метод Ричардсона (3.156) для систем с симметричными и положительно определёнными матрицами может быть представлен в виде

$$x^{(k)} \leftarrow x^{(k-1)} - \tau \Psi'(x^{(k-1)}), \quad k = 1, 2, \dots,$$

т. е. он является не чем иным, как методом градиентного спуска (3.165) для минимизации функционала энергии  $\Psi$ , в котором шаг  $\tau_k$  выбран постоянным и равным  $\tau$ . В общем случае метод градиентного спуска (3.165) для таких СЛАУ оказывается равносильным простейшему нестационарному итерационному методу (3.157).

Для метода градиентного спуска с постоянным шагом его трактовка как метода Ричардсона позволяет, опираясь на результат об оптимизации скалярного предобуславливателя из § 3.10г, выбрать шаг  $\tau_k = \text{const}$ , который наверняка обеспечивает сходимость процесса. Именно, если положительные числа  $\mu$  и  $M$  — это нижняя и верхняя границы спектра положительно определённой матрицы  $A$  решаемой системы, то в соответствии с (3.141) для сходимости следует взять

$$\tau_k = \tau = \frac{2}{M + \mu}.$$

Другой способ выбора шага состоит в том, чтобы потребовать  $\tau_k$  наибольшим возможным, обеспечивающим убывание функционала  $\Psi$  вдоль выбранного направления спуска по антиградиенту. При этом получается *метод наискорейшего градиентного спуска*, теория которого была разработана в конце 40-х годов XX века Л.В. Канторовичем.

Для определения конкретной величины шага  $\tau_k$  в методе наискорейшего градиентного спуска воспользуемся результатом предшествующего раздела — формулой (3.164). Тогда величина шага в (3.165) должна быть

$$\tau_k = \frac{\langle r^{(k-1)}, r^{(k-1)} \rangle}{\langle Ar^{(k-1)}, r^{(k-1)} \rangle},$$

где  $r^{(k-1)} = Ax^{(k-1)} - b$  — невязка  $(k-1)$ -го приближения к решению, и для удобства сменён знак в сравнении с (3.164). В целом псевдокод метода наискорейшего градиентного спуска для решения системы линейных алгебраических уравнений  $Ax = b$  представлен в табл. 3.11.

Таблица 3.11. Метод наискорейшего градиентного спуска  
для решения систем линейных уравнений

```

 $k \leftarrow 1;$ 
выбираем начальное приближение  $x^{(0)}$ ;
DO WHILE ( метод не сопротивляется )
     $r^{(k-1)} \leftarrow Ax^{(k-1)} - b;$ 
     $\tau_k \leftarrow \frac{\|r^{(k-1)}\|_2^2}{\langle Ar^{(k-1)}, r^{(k-1)} \rangle};$ 
     $x^{(k)} \leftarrow x^{(k-1)} - \tau_k r^{(k-1)};$ 
     $k \leftarrow k + 1;$ 
END DO

```

**Теорема 3.11.2** *Если в системе линейных алгебраических уравнений  $Ax = b$  матрица  $A$  симметрична и положительно определена, то последовательность  $\{x^{(k)}\}$ , порождаемая методом наискорейшего градиентного спуска, сходится к решению системы  $x^*$  из любого начального приближения  $x^{(0)}$ .*

чального приближения  $x^{(0)}$ . Быстрота этой сходимости оценивается неравенством

$$\|x^{(k)} - x^*\|_A \leq \left(\frac{M-\mu}{M+\mu}\right)^k \|x^{(0)} - x^*\|_A, \quad (3.167)$$

$k = 0, 1, 2, \dots$ , если все собственные значения матрицы  $A$  лежат в интервале  $[\mu, M] \subset \mathbb{R}_+$ .

**Доказательство** оценки (3.167) и теоремы в целом будет получено путём сравнения метода наискорейшего спуска с методом градиентного спуска с постоянным оптимальным шагом, т. е. с методом Ричардсона.

Пусть в результате выполнения  $(k-1)$  шагов метода наискорейшего спуска получено приближение  $x^{(k-1)}$ , и мы делаем  $k$ -й шаг, который даёт  $x^{(k)}$ . Обозначим также через  $\tilde{x}$  результат выполнения с  $x^{(k-1)}$  одного шага итерационного метода Ричардсона, так что

$$\tilde{x} = x^{(k-1)} - \tau(Ax^{(k-1)} - b).$$

Из развитой в начале раздела теории вытекает, что при любом выборе параметра  $\tau$

$$\Psi(x^{(k)}) \leq \Psi(\tilde{x}),$$

так как метод наискорейшего спуска обеспечивает наибольшее уменьшение функционала энергии на одном шаге итераций. Далее, из равенства (3.162)

$$\Psi(x) = \frac{1}{2}\|x - x^*\|_A^2 - \frac{1}{2}\|x^*\|_A^2$$

с постоянным вычитаемым  $\frac{1}{2}\|x^*\|_A^2$  следует, что

$$\frac{1}{2}\|x^{(k)} - x^*\|_A^2 \leq \frac{1}{2}\|\tilde{x} - x^*\|_A^2,$$

т. е.

$$\|x^{(k)} - x^*\|_A \leq \|\tilde{x} - x^*\|_A. \quad (3.168)$$

Иными словами, метод, обеспечивающий лучшее убывание значения функционала энергии одновременно обеспечивает лучшее приближение к решению в энергетической норме.

В методе градиентного спуска с постоянным шагом, совпадающем с итерационным методом Ричардсона (3.138) или (3.156), имеем

$$\tilde{x} - x^* = (I - \tau A)(x^{(k-1)} - x^*), \quad k = 1, 2, \dots$$

Матрица  $(I - \tau A)$  является полиномом первой степени от матрицы  $A$ , и потому можем применить неравенство (3.42) из предложения 3.3.11 (стр. 399):

$$\|\tilde{x} - x^*\|_A \leq \|I - \tau A\|_2 \|x^{(k-1)} - x^*\|_A.$$

При этом в силу (3.142) для метода наискорейшего спуска оценка погрешности заведомо не хуже этой оценки с произвольным значением  $\tau$ . В частности, мы можем взять значение параметра  $\tau = 2/(M + \mu)$ , оптимальное для спуска с постоянным шагом. Тогда в соответствии с оценкой (3.142), выведенной при анализе скалярного предобуславливателя, получаем

$$\|x^{(k)} - x^*\|_A \leq \left(\frac{M - \mu}{M + \mu}\right) \|x^{(k-1)} - x^*\|_A, \quad k = 1, 2, \dots,$$

откуда следует доказываемое неравенство (3.167). ■

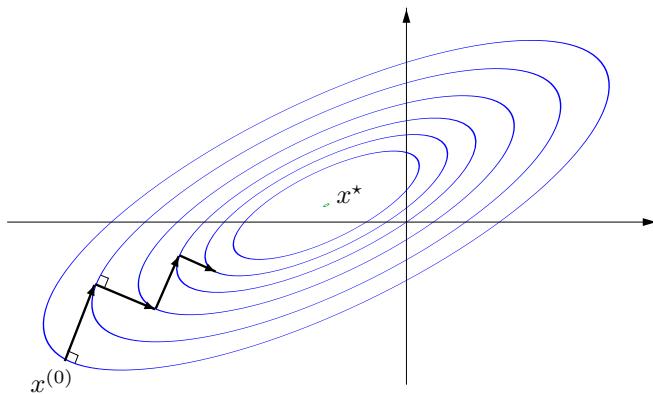


Рис. 3.29. Иллюстрация работы метода наискорейшего градиентного спуска

Интересно и поучительно рассмотреть геометрическую иллюстрацию работы метода наискорейшего спуска.

Градиент функционала энергии нормален к его поверхностям уровня, и именно по этим направлениям осуществляется «спуск», т. е. движение в сторону решения. Шаг в методе наискорейшего спуска идёт на максимально возможную величину, до пересечения с касательным эллипсоидом. Поэтому траектория метода наискорейшего спуска является ломаной, звенья которой перпендикулярны друг другу (рис. 3.29).

Доказательство теоремы 3.11.2 основано на «мажоризации» (оценивании сверху) наискорейшего спуска итерационным методом Ричардсона и может показаться довольно грубым. Но итоговая оценка (3.167) в действительности весьма точно передаёт особенности поведения метода, а именно, замедление сходимости при  $M \gg \mu$ . При этом матрица системы плохообусловлена, а зигзагообразное движение к решению в методе наискорейшего спуска весьма далеко от оптимального. Этот факт подтверждается вычислительной практикой и может быть понят на основе геометрической интерпретации. Искомое решение находится тогда на дне глубокого и вытянутого оврага, а метод «рыскает» от одного склона оврага к другому вместо того, чтобы идти напрямую к глубочайшей точке — решению.

### 3.11г Метод минимальных невязок

Пусть дана система линейных алгебраических уравнений  $Ax = b$  с положительно определённой матрицей  $A$ , которая не обязательно симметрична. Для нестационарного итерационного процесса (3.157)

$$x^{(k)} \leftarrow x^{(k-1)} - \tau_k (Ax^{(k-1)} - b), \quad k = 1, 2, \dots,$$

ещё один популярный подход к выбору итерационных параметров  $\tau_k$  был предложен С.Г. Крейном и М.А. Красносельским в работе [25] и назван ими *методом минимальных невязок*. Его псевдокод приведён в табл. 3.12.

Каждый шаг этого метода минимизирует в направлении невязки  $(k-1)$ -го приближения, равной  $r^{(k-1)} = Ax^{(k-1)} - b$ , не функционал энергии  $\Psi(x)$ , как это делалось в методе наискорейшего градиентного спуска, а евклидову норму невязки  $\|Ax - b\|_2$  или, что равносильно,  $\|Ax - b\|_2^2$ . Оказывается, что это эквивалентно наибольшему возможному уменьшению погрешности приближённого решения в энергетической норме, которая порождена матрицей  $A^\top A$ . В самом деле, если  $x^*$  — точное решение системы уравнений, то  $Ax^* = b$ , и потому

$$\begin{aligned} \|Ax - b\|_2^2 &= \langle Ax - b, Ax - b \rangle = \langle Ax - Ax^*, Ax - Ax^* \rangle = \\ &= \langle A(x - x^*), A(x - x^*) \rangle = \langle A^\top A(x - x^*), x - x^* \rangle = \\ &= \|x - x^*\|_{A^\top A}^2. \end{aligned} \tag{3.169}$$

Предположим, что  $\tilde{x}$  — приближение к решению, которому соответствует невязка  $\tilde{r} = A\tilde{x} - b$ . Если желаем выбрать параметр шага  $\tau$

Таблица 3.12. Метод минимальных невязок для решения систем линейных алгебраических уравнений

```

 $k \leftarrow 1;$ 
выбираем начальное приближение  $x^{(0)}$ ;
DO WHILE ( метод не сопшёлся )
     $r^{(k-1)} \leftarrow Ax^{(k-1)} - b;$ 
     $\tau_k \leftarrow \frac{\langle Ar^{(k-1)}, r^{(k-1)} \rangle}{\|Ar^{(k-1)}\|_2^2};$ 
     $x^{(k)} \leftarrow x^{(k-1)} - \tau_k r^{(k-1)};$ 
     $k \leftarrow k + 1;$ 
END DO

```

так, чтобы очередное приближение  $(\tilde{x} - \tau \tilde{r})$ , выбранное по направлению  $(-\tilde{r})$ , минимизировало 2-норму невязки, то необходимо найти минимум по  $\tau$  для выражения

$$\begin{aligned} \|A(\tilde{x} - \tau \tilde{r}) - b\|_2^2 &= \langle A(\tilde{x} - \tau \tilde{r}) - b, A(\tilde{x} - \tau \tilde{r}) - b \rangle = \\ &= \tau^2 \langle A\tilde{r}, A\tilde{r} \rangle - 2\tau (\langle A\tilde{x}, A\tilde{r} \rangle - \langle b, A\tilde{r} \rangle) + \\ &\quad + \langle A\tilde{x}, A\tilde{x} \rangle + \langle b, b \rangle. \end{aligned}$$

Дифференцируя его по  $\tau$  и приравнивая производную нулю, получим

$$2\tau \langle A\tilde{r}, A\tilde{r} \rangle - 2(\langle A\tilde{x}, A\tilde{r} \rangle - \langle b, A\tilde{r} \rangle) = 0,$$

что с учётом равенства  $A\tilde{x} - b = \tilde{r}$  даёт

$$\tau \langle A\tilde{r}, A\tilde{r} \rangle - \langle \tilde{r}, A\tilde{r} \rangle = 0.$$

Отсюда вытекает

$$\tau = \frac{\langle A\tilde{r}, \tilde{r} \rangle}{\langle A\tilde{r}, A\tilde{r} \rangle} = \frac{\langle A\tilde{r}, \tilde{r} \rangle}{\|A\tilde{r}\|_2^2}.$$

Этот выбор  $\tau$  реализован в методе минимальных невязок из табл. 3.12.

**Теорема 3.11.3** Пусть в системе линейных алгебраических уравнений  $Ax = b$  матрица  $A$  положительно определена,  $\mu$  — минимальное собственное значение матрицы  $(A + A^\top)/2$  и  $M := \|A\|_2$ . Последовательность  $\{x^{(k)}\}$ , порождаемая методом минимальных невязок, сходится к решению системы уравнений  $Ax = b$  из любого начального приближения и быстрота убывания невязок  $r^{(k)} := Ax^{(k)} - b$  оценивается неравенством

$$\|r^{(k)}\|_2 \leq \sqrt{1 - \left(\frac{\mu}{M}\right)^2} \|r^{(k-1)}\|_2, \quad k = 1, 2, \dots$$

**Доказательство** теоремы можно найти, к примеру, в книге [39].

Отметим, что с учётом выкладок (3.169) неравенство из формулировки теоремы совершенно равносильно следующей оценке погрешности очередного приближения в методе минимальных невязок:

$$\|x^{(k)} - x^*\|_{A^\top A} \leq \left(1 - \left(\frac{\mu}{M}\right)^2\right)^{k/2} \|x^{(0)} - x^*\|_{A^\top A}.$$

Для линейных систем с симметричной положительно определённой матрицей полезным свойством метода минимальных невязок является монотонное убывание евклидовой нормы погрешности приближений на каждом шаге:

$$\|x^{(k)} - x^*\|_2 \leq \|x^{(k-1)} - x^*\|_2, \quad k = 1, 2, \dots$$

Обоснование этого неравенства нетривиально, и его можно найти в оригинальной работе [25].

Для систем линейных уравнений с несимметричными матрицами, которые положительно определены, метод минимальных невязок успешно сходится. Но если матрица системы не является положительно определённой, сходимости к решению может не быть.

**Пример 3.11.1** В системе линейных алгебраических уравнений

$$\begin{pmatrix} 2 & 2 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

матрица не является симметричной, но она положительно определена. Метод минимальных невязок сходится из любого начального приближения к точному решению  $(-1, 1)^\top$ .

В системе линейных алгебраических уравнений

$$\begin{pmatrix} 2 & 2 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

матрица не является ни симметричной, ни положительно определённой (её собственные значения приблизительно равны 2.732 и  $-0.7321$ ). В применении к этой системе у метода минимальных невязок с нулевым начальным приближением все итерации оказываются нулевыми, тогда как настоящее решение — это вектор  $(1, -1)^\top$ . Из других начальных приближений метод будет сходиться к другим векторам, которые также не совпадают с этим точным решением. ■

Практически важной особенностью метода минимальных невязок является быстрая сходимость к решению на первых шагах, которая затем замедляется и выходит на асимптотическую скорость, описанную теоремой 3.11.3.

Если сходимость методов наискорейшего спуска и минимальных невязок принципиально не лучше сходимости простейшего итерационного метода Ричардсона, то имеют ли они какое-либо практическое значение? Ответ на этот вопрос положителен. Напомним, что оптимизация метода Ричардсона в § 3.10г основывалась на знании границ спектра симметричной положительно определённой матрицы СЛАУ. Для работы методов наискорейшего спуска и минимальных невязок этой информации не требуется. Они фактически сами «подстраивают» под решаемую систему уравнений.

Метод минимальных невязок в представленной выше простейшей версии не отличается большой эффективностью. Но он послужил основой для создания многих популярных современных методов решения СЛАУ. В частности, большое распространение на практике получила модификация метода минимальных невязок, известная под англоязычной аббревиатурой GMRES (от Generalized Minimal REsidulas — обобщённый метод минимальных невязок), предложенная Ю. Саадом [39] (см. также [46, 61]).

### 3.11д Метод сопряжённых градиентов

*Методами сопряжённых направлений* для решения систем линейных алгебраических уравнений вида  $Ax = b$  называют методы, в которых решение ищется в виде линейной комбинации векторов, орто-

гональных в каком-то специальном скалярном произведении. Обычно оно порождено матрицей системы  $A$  или же какой-либо матрицей, связанной с матрицей системы. Таким образом, решение представляется в виде

$$x = x^{(0)} + \sum_{i=1}^n c_i s^{(i)}, \quad (3.170)$$

где  $x^{(0)}$  — начальное приближение,  $s^{(i)}$ ,  $i = 1, 2, \dots, n$ , — векторы «сопряжённых направлений»,  $c_i$  — коэффициенты разложения решения по ним.

Термин «сопряжённые направления» имеет происхождение в аналитической геометрии, где направления, задаваемые векторами  $u$  и  $v$ , называются сопряжёнными относительно поверхности второго порядка, которая определяется уравнением  $\langle Rx, x \rangle = \text{const}$  с симметричной матрицей  $R$ , если  $\langle Ru, v \rangle = 0$ . В методах сопряжённых направлений последовательно строится базис из  $A$ -ортогональных векторов  $s^{(i)}$  и одновременно находятся коэффициенты  $c_i$ ,  $i = 1, 2, \dots, n$ .

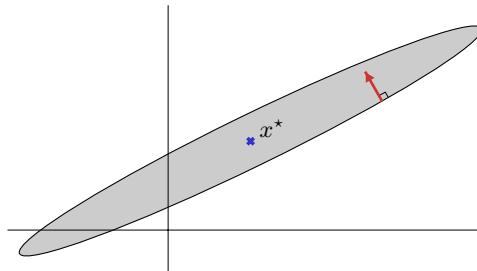


Рис. 3.30. Направление антиградиента функционала может быть плохим для спуска к его минимуму

Наиболее популярным представителем методов сопряжённых направлений является *метод сопряжённых градиентов*, предложенный М.Р. Хестенсом и Э.Л. Штифелем в начале 50-х годов прошлого века. Как почти любой достаточно сложный численный метод, его можно представить несколькими различными способами в зависимости от целей исследования и применения. Нам будет удобно вывести метод сопряжённых градиентов как метод наискорейшего спуска для минимизации функционала энергии (см. § 3.11б), в котором направление спуска выбирается специальным и чрезвычайно удачным образом.

В предшествующем разделе мы могли видеть, что направления антиградиентов функционала энергии, по которым осуществляется движение к решению в методе наискорейшего градиентного спуска, не очень удачны (рис. 3.30), а спуск по ним может сильно «вихлять» от шага к шагу. В целом эта траектория спуска к решению весьма нерациональна и для нахождения решения затрачивает много лишней работы. Естественно попытаться организовать алгоритм так, чтобы он вёл к решению более прямым путём, и один из возможных способов сделать это состоит в коррекции направления спуска.

Пусть к началу  $k$ -го шага поиска решения с помощью спуска по направлению  $s^{(k-1)}$  уже найдено приближение к решению  $x^{(k-1)}$ . Следующее направление спуска  $s^{(k)}$ , на котором будет получено  $x^{(k)}$ , возьмём как линейную комбинацию векторов

- $-\nabla\Psi(x^{(k-1)})$ , т. е. антиградиента функционала энергии в  $x^{(k-1)}$ , противоположного невязке  $r^{(k-1)} = Ax^{(k-1)} - b$ ,
- предыдущего направления спуска  $s^{(k-1)}$ .

На языке формул

$$s^{(k)} := -r^{(k-1)} + v_k s^{(k-1)}$$

с некоторым коэффициентом  $v_k \in \mathbb{R}$ . Этот выбор вполне естественен, так как одно лишь «чистое» направление антиградиента, как показано в § 3.11в, не вполне удовлетворительно, и имеет смысл «смешать» его с дополнительной информацией о других возможных направлениях спуска. В качестве такого берётся направление спуска предыдущего шага.

Отличительной особенностью именно метода сопряжённых градиентов является такой выбор коэффициента  $v_k$ , при котором направления спуска  $s^{(k)}$  и  $s^{(k-1)}$  являются  $A$ -ортогональными:

$$\langle s^{(k)}, s^{(k-1)} \rangle_A = 0 ,$$

то есть

$$\langle As^{(k)}, s^{(k-1)} \rangle = \langle s^{(k)}, As^{(k-1)} \rangle = 0, \quad k = 1, 2, \dots, n. \quad (3.171)$$

Напомним, что последовательные направления спуска в методе наискорейшего градиентного спуска ортогональны друг другу в обычном

смысле. Условие (3.171) тоже требует ортогональность, но в скалярном произведении, порождённом матрицей решаемой системы уравнений. Его можно рассматривать как попытку скорректировать направления спуска таким образом, чтобы они стали лучше соответствовать конкретной решаемой системе.

С другой стороны, можно трактовать (3.171) как выбор в качестве последовательных направлений спуска наиболее представительного семейства векторов, «покрывающих» всё пространство (фактически его базис) и согласованных с матрицей решаемой системы. Эта идея вполне естественна для несложных целевых функций (каким и является функционал энергии).

Из (3.171) следует

$$\langle -r^{(k-1)} + v_k s^{(k-1)}, As^{(k-1)} \rangle = 0,$$

так что

$$-\langle r^{(k-1)}, As^{(k-1)} \rangle + v_k \langle s^{(k-1)}, As^{(k-1)} \rangle = 0.$$

Окончательно,

$$v_k = \frac{\langle r^{(k-1)}, As^{(k-1)} \rangle}{\langle s^{(k-1)}, As^{(k-1)} \rangle} = \frac{\langle As^{(k-1)}, r^{(k-1)} \rangle}{\langle As^{(k-1)}, s^{(k-1)} \rangle}. \quad (3.172)$$

Вычислительная схема метода сопряжённых градиентов

$$x^{(k)} \leftarrow x^{(k-1)} + \tau_k s^{(k)}, \quad k = 1, 2, \dots,$$

— такая же, как и у всех методов спуска, а величина шага  $\tau_k$  берётся так, чтобы обеспечить наибольшее убывание функционала энергии вдоль направления спуска. Задачу выбора шага в методе наискорейшего спуска вдоль заданного направления мы уже решали в § 3.11б и можем воспользоваться полученной там формулой (3.164). В нашем случае она даёт

$$\tau_k = -\frac{\langle r^{(k-1)}, s^{(k)} \rangle}{\langle As^{(k)}, s^{(k)} \rangle}. \quad (3.173)$$

Итак, метод сопряжённых градиентов можно представить следующей схемой:

Шаг 1. Выбираем начальное приближение  $x^{(0)}$  и в качестве направления спуска на первом шаге берём

$$s^{(1)} = -\nabla \Psi(x^{(0)}) = -r^{(0)} = -(Ax^{(0)} - b),$$

т. е. направление антиградиента функционала энергии в точке  $x^{(0)}$ , равное минус невязке в силу равенства (3.166). Первый шаг не имеет предшествующего, так что процедура выбора направления спуска на этом заканчивается.

Следующее приближение  $x^{(1)}$  строится как шаг от  $x^{(0)}$  по направлению  $s^{(1)}$  на величину, определяемую формулой (3.173) при  $k = 1$ , т. е. как

$$x^{(1)} \leftarrow x^{(0)} + \tau_1 s^{(1)},$$

где

$$\tau_1 = -\frac{\langle r^{(0)}, s^{(1)} \rangle}{\langle As^{(1)}, s^{(1)} \rangle} = \frac{\langle r^{(0)}, r^{(0)} \rangle}{\langle Ar^{(0)}, r^{(0)} \rangle}.$$

Из общей теории следует, что такой выбор шага минимизирует функционал энергии вдоль выбранного направления.

Шаг  $k$ ,  $k = 2, 3, \dots$ . Пусть уже известно приближение  $x^{(k-1)}$  и направление спуска  $s^{(k-1)}$ , по которому оно было получено на предыдущем шаге.

Выбираем следующим направлением спуска вектор

$$s^{(k)} = -r^{(k-1)} + v_k s^{(k-1)}, \quad (3.174)$$

где

$$r^{(k-1)} = Ax^{(k-1)} - b \quad \text{— невязка в точке } x^{(k-1)},$$

а коэффициент  $v_k$  выбирается из условия  $A$ -ортогональности направлений спуска  $s^{(k)}$  и  $s^{(k-1)}$ , т. е. в соответствии с формулой (3.172):

$$v_k = \frac{\langle As^{(k-1)}, r^{(k-1)} \rangle}{\langle As^{(k-1)}, s^{(k-1)} \rangle}.$$

Вычисляем затем следующее приближение к решению

$$x^{(k)} \leftarrow x^{(k-1)} + \tau_k s^{(k)},$$

где

$$\tau_k = -\frac{\langle r^{(k-1)}, s^{(k)} \rangle}{\langle As^{(k)}, s^{(k)} \rangle}.$$

Как следствие, на  $k$ -ом шаге метода сопряжённых градиентов невязка (равная градиенту функционала энергии) изменяется следующим

образом

$$\begin{aligned} r^{(k)} &= Ax^{(k)} - b = \\ &= A(x^{(k-1)} + \tau_k s^{(k)}) - b = \\ &= r^{(k-1)} + \tau_k As^{(k)}. \end{aligned} \quad (3.175)$$

Последняя формула даёт альтернативное рекуррентное выражение для невязки, которое полезно и в теории, и на практике.

### 3.11e Сходимость метода сопряжённых градиентов

**Предложение 3.11.1** В методе сопряжённых градиентов векторы невязок ортогональны направлениям спуска на текущем и предыдущем шагах:

$$\begin{aligned} \langle r^{(1)}, s^{(1)} \rangle &= 0, \\ \langle r^{(k)}, s^{(k)} \rangle &= \langle r^{(k)}, s^{(k-1)} \rangle = 0, \quad k = 2, 3, \dots \end{aligned}$$

Векторы невязок, получаемые на следующих друг за другом шагах метода сопряжённых градиентов, ортогональны друг другу:

$$\langle r^{(k-1)}, r^{(k)} \rangle = 0, \quad k = 1, 2, \dots,$$

**Доказательство.** Поскольку

$$x^{(1)} = x^{(0)} - \tau_1 r^{(0)},$$

то

$$\begin{aligned} -\langle r^{(1)}, s^{(1)} \rangle &= \langle r^{(1)}, r^{(0)} \rangle = \langle Ax^{(1)} - b, r^{(0)} \rangle = \\ &= \langle Ax^{(0)} - \tau_1 Ar^{(0)} - b, r^{(0)} \rangle = \langle r^{(0)} - \tau_1 Ar^{(0)}, r^{(0)} \rangle = \\ &= \langle r^{(0)}, r^{(0)} \rangle - \tau_1 \langle Ar^{(0)}, r^{(0)} \rangle = 0 \end{aligned}$$

в силу определения  $\tau_1$ .

Далее доказательство продолжается индукцией по  $k$ .

Рассмотрим  $k$ -й шаг метода сопряжённых градиентов. Из формулы (3.175) следует, что

$$\langle r^{(k)}, s^{(k)} \rangle = \langle r^{(k-1)}, s^{(k)} \rangle + \tau_k \langle As^{(k)}, s^{(k)} \rangle = 0,$$

так как в силу (3.173)

$$\tau_k = -\frac{\langle r^{(k-1)}, s^{(k)} \rangle}{\langle As^{(k)}, s^{(k)} \rangle}.$$

Поэтому на основе (3.175) можем заключить, что

$$\begin{aligned}\langle r^{(k)}, s^{(k-1)} \rangle &= \langle r^{(k-1)}, s^{(k-1)} \rangle + \tau_k \langle As^{(k)}, s^{(k-1)} \rangle = \\ &= \langle r^{(k-1)}, s^{(k-1)} \rangle + \tau_k \langle s^{(k)}, s^{(k-1)} \rangle_A.\end{aligned}$$

Первое слагаемое здесь равно нулю по индукционному предположению, а второе — по построению метода сопряжённых градиентов, т. е. в силу равенства (3.171).

Обоснование второй части предложения выведем из доказанного равенства

$$\langle s^{(k)}, r^{(k)} \rangle = 0, \quad k = 1, 2, \dots$$

Если подставить сюда определение  $s_k$  из формулы (3.174), получим

$$-\langle r^{(k-1)}, r^{(k)} \rangle + v_k \langle s^{(k-1)}, r^{(k)} \rangle = 0.$$

Выше мы показали, что  $\langle r^{(k)}, s^{(k-1)} \rangle = \langle s^{(k-1)}, r^{(k)} \rangle = 0$ . Таким образом, в самом деле

$$\langle r^{(k-1)}, r^{(k)} \rangle = 0,$$

что завершает доказательство предложения. ■

На самом деле справедлив более общий результат

**Теорема 3.11.4** В методе сопряжённых градиентов векторы направлений спуска  $s^{(k)}$ ,  $k = 1, 2, \dots$ , являются  $A$ -ортогональными друг другу, т. е.

$$\langle s^{(i)}, s^{(j)} \rangle_A = \langle As^{(i)}, s^{(j)} \rangle = 0 \quad \text{при } i \neq j.$$

В методе сопряжённых градиентов векторы невязок  $r^{(k)} = Ax^{(k)} - b$ ,  $k = 0, 1, 2, \dots$ , ортогональны друг другу, т. е.

$$\langle r^{(i)}, r^{(j)} \rangle = 0 \quad \text{при } i \neq j.$$

Иными словами,

- $A$ -ортогональны не только следующие друг за другом направления спуска, как требуется по построению метода, но все эти направления вообще взаимно  $A$ -ортогональны друг другу;
- ортогональны не только следующие друг за другом вектора невязок, но все эти невязки вообще взаимно ортогональны друг другу.

Напомним, что в методе наискорейшего градиентного спуска направления спуска на соседних шагах тоже ортогональны друг другу, но в целом ортогональности нет. То же самое верно для невязок метода наискорейшего градиентного спуска (которые, как известно, противоположны направлениям спуска): ортогональны последовательные невязки на соседних шагах, но общей ортогональности нет. Метод сопряжённых градиентов в этом отношении качественно отличается от метода наискорейшего градиентного спуска. Как уже отмечалось, в методе сопряжённых градиентов из последовательных направлений спуска фактически строится  $A$ -ортогональный базис пространства решений. Шаги вдоль этих направлений спуска дают коэффициенты  $c_i$  разложения решения по конструируемому базису в представлении (3.170) (см. § 3.11ж).

Для удобства условимся считать, что  $s^{(0)} = 0$ , т. е. что существует виртуальный «нулевой шаг», который привёл к нулевому приближению, и направление спуска на нём — это нулевой вектор.

**Доказательство** теоремы проводится индукцией по номеру  $k$  шага алгоритма. Прежде всего построим базу индукции.

Для  $k = 1$  невязки  $r^{(0)}$  и  $r^{(1)}$  ортогональны в силу доказанного выше предложения 3.11.1. Кроме того,  $A$ -ортогональны  $s^{(0)}$  и  $s^{(1)}$ , просто потому, что нулевой вектор ортогонален любому вектору.

Предположив, что утверждение теоремы справедливо для  $k$ -го шага метода, покажем, что оно верно также для шага  $k + 1$ . Иначе говоря, необходимо доказать равенства

$$\langle s^{(k+1)}, s^{(j)} \rangle_A = 0 \quad \text{для } j = 1, 2, \dots, k, \quad (3.176)$$

$$\langle r^{(k+1)}, r^{(j)} \rangle = 0 \quad \text{для } j = 0, 1, 2, \dots, k. \quad (3.177)$$

Заметим, что

$$\langle s^{(k+1)}, s^{(k)} \rangle_A = \langle s^{(k+1)}, As^{(k)} \rangle = 0$$

по построению метода сопряжённых градиентов. Далее, так как согласно (3.174)

$$s^{(k+1)} = -r^{(k)} + v_{k+1}s^{(k)},$$

должно выполняться

$$\begin{aligned} \langle s^{(k+1)}, s^{(j)} \rangle_A &= \langle s^{(k+1)}, As^{(j)} \rangle = \\ &= -\langle r^{(k)}, As^{(j)} \rangle + v_{k+1} \langle s^{(k)}, As^{(j)} \rangle = \\ &= -\langle r^{(k)}, As^{(j)} \rangle + v_{k+1} \langle s^{(k)}, s^{(j)} \rangle_A, \end{aligned}$$

где последнее слагаемое зануляется при  $j < k$  в силу индукционного предположения. Итак,

$$\langle s^{(k+1)}, s^{(j)} \rangle_A = -\langle r^{(k)}, As^{(j)} \rangle. \quad (3.178)$$

С другой стороны, из соотношения (3.175) для  $j$ -й невязки в методе сопряжённых градиентов

$$r^{(j)} = r^{(j-1)} + \tau_j As^{(j)}$$

следует, что

$$As^{(j)} = \frac{1}{\tau_j} (r^{(j)} - r^{(j-1)}).$$

Подставив полученное выражение в правую часть равенства (3.178), будем иметь

$$\begin{aligned} \langle s^{(k+1)}, s^{(j)} \rangle_A &= -\langle r^{(k)}, As^{(j)} \rangle = \\ &= -\frac{1}{\tau_j} (\langle r^{(k)}, r^{(j)} \rangle - \langle r^{(k)}, r^{(j-1)} \rangle) = 0, \end{aligned}$$

поскольку оба слагаемых в больших скобках равны нулю в силу индукционного предположения. Это доказывает (3.176), т. е. первое утверждение теоремы об  $A$ -ортогональности всех направлений спуска.

Установим теперь равенства (3.177). Мы знаем, что  $\langle r^{(k+1)}, r^{(k)} \rangle = 0$  в силу доказанного предложения 3.11.1. Остаётся показать, что для  $0 \leq j < k$  также  $\langle r^{(k+1)}, r^{(j)} \rangle = 0$ .

В силу формулы (3.175)

$$r^{(k+1)} = r^{(k)} + \tau_{k+1} As^{(k+1)},$$

и потому

$$\begin{aligned} \langle r^{(k+1)}, r^{(j)} \rangle &= \langle r^{(k)}, r^{(j)} \rangle + \tau_{k+1} \langle As^{(k+1)}, r^{(j)} \rangle = \\ &= \tau_{k+1} \langle As^{(k+1)}, r^{(j)} \rangle \end{aligned} \quad (3.179)$$

согласно индукционному предположению. В методе сопряжённых градиентов направления спуска на  $j$ -ом и  $(j+1)$ -ом шагах связаны соотношением

$$s^{(j+1)} = -r^{(j)} + v_{j+1}s^{(j)},$$

так что с помощью этих направлений можно выразить невязку  $r^{(j)}$ :

$$r^{(j)} = -s^{(j+1)} + v_{j+1}s^{(j)}.$$

Подставляя результат в выражение (3.179), получим

$$\begin{aligned} \langle r^{(k+1)}, r^{(j)} \rangle &= \tau_{k+1} \langle As^{(k+1)}, r^{(j)} \rangle = \\ &= \tau_{k+1} \left( -\langle As^{(k+1)}, s^{(j+1)} \rangle + v_{j+1} \langle As^{(k+1)}, s^{(j)} \rangle \right). \end{aligned}$$

Но мы уже доказали, что все направления спуска  $A$ -ортогональны друг другу. Следовательно, для  $j < k$

$$\langle As^{(k+1)}, s^{(j+1)} \rangle = 0 \quad \text{и} \quad \langle As^{(k+1)}, s^{(j)} \rangle = 0.$$

В целом имеем,

$$\langle r^{(k+1)}, r^{(j)} \rangle = 0,$$

как и требовалось.

Если придерживаться нашего соглашения  $s^{(0)} = 0$ , то проведённые рассуждения верны также в случае  $j = 0$ , когда  $r^{(0)} = -s^{(1)}$ .

Теперь теорема полностью доказана. ■

**Предложение 3.11.2** *Метод сопряжённых градиентов в применении к  $n \times n$ -системе линейных алгебраических уравнений с симметричной положительно определённой матрицей находит точное решение не более чем за  $n$  шагов.*

Необходимая оговорка к формулировке теоремы заключается в том, что в ней подразумевается выполнение метода сопряжённых градиентов в идеальной арифметике вещественных чисел, а не в арифметике реальных цифровых ЭВМ.

**Доказательство** следует из того, что размерность пространства  $\mathbb{R}^n$  равна  $n$ , а невязки  $r^{(k)} = Ax^{(k)} - b$ ,  $k = 0, 1, 2, \dots$ , — ортогональные в силу доказанной теоремы. Поэтому они линейно независимы. Таким образом, ненулевых невязок может быть не более  $n$  штук, и с какого-то шага метода невязка необходимо занулится. ■

На практике из-за неизбежных погрешностей вычислений метод сопряжённых градиентов может не прийти к точному решению системы за  $n$  или менее шагов. Тогда целесообразно повторить цикл уточнения, возможно, даже не один раз, превратив алгоритм при необходимости в итерационный. При этом справедливо представление, обобщающее (3.170):

$$x = x^{(0)} + \sum_i c_i s^{(i)},$$

где  $x^{(0)}$  — начальное приближение,  
 $s^{(i)}$ ,  $i = 1, 2, \dots$ , — векторы направлений спуска, которые являются также векторами разложения решения,  
 $c_i$  — коэффициенты разложения решения по векторам  $s^{(i)}$ , равные шагам спуска в соответствующих направлениях.

Верхний предел выписанной суммы может значительно превосходить  $n$ .

**Теорема 3.11.5** *Если в системе линейных алгебраических уравнений  $Ax = b$  матрица  $A$  симметрична и положительно определена, то последовательность  $\{x^{(k)}\}$ , порождаемая методом сопряжённых градиентов, сходится к решению  $x^*$  системы уравнений  $Ax = b$  из любого начального приближения  $x^{(0)}$ . Быстрота этой сходимости оценивается неравенством*

$$\|x^{(k)} - x^*\|_A \leq 2 \left( \frac{\sqrt{M} - \sqrt{\mu}}{\sqrt{M} + \sqrt{\mu}} \right)^k \|x^{(0)} - x^*\|_A, \quad (3.180)$$

$k = 1, 2, \dots$ , если все собственные значения матрицы  $A$  лежат в интервале  $[\mu, M] \subset \mathbb{R}_+$ .

**Доказательство** опускается. Его можно найти, к примеру, в [1, 39, 44, 46, 61]. Нужно отметить, что оценка (3.180) не отражает все существенные особенности сходимости метода сопряжённых градиентов. Большое значение имеет конкретное расположение собственных чисел

матрицы системы, а не только наименьшее и наибольшее из них. Но оценка (3.180) всё-таки существенно лучше аналогичных оценок (3.143) и (3.167), так как в ней фигурируют  $\sqrt{M}$  и  $\sqrt{\mu}$  вместо их первых степеней, а потому коэффициент подавления погрешности на отдельном шаге в методе сопряжённых градиентов лучше.

Разделим числитель и знаменатель дроби  $\frac{\sqrt{M}-\sqrt{\mu}}{\sqrt{M}+\sqrt{\mu}}$  на  $\sqrt{\mu}$  и применим рассуждения, аналогичные тем, с помощью которых анализировалась оценка (3.143). Они приводят к заключению, что количество итераций метода сопряжённых градиентов, необходимых для уменьшения погрешности в заданное число раз, пропорционально квадратному корню из числа обусловленности, а не самому этому числу.

Таблица 3.13. Метод сопряжённых градиентов для решения систем линейных алгебраических уравнений

```

 $k \leftarrow 1;$ 
выбираем начальное приближение  $x^{(0)}$ ;
 $r^{(0)} \leftarrow b - Ax^{(0)}$ ;  $s^{(1)} \leftarrow r^{(0)}$ ;
DO WHILE ( метод не сопёлся )
 $g \leftarrow As^{(k)}$ ;
 $\tau_k \leftarrow \frac{\langle r^{(k-1)}, r^{(k-1)} \rangle}{\langle s^{(k)}, g \rangle}$ ;
 $x^{(k)} \leftarrow x^{(k-1)} + \tau_k s^{(k)}$ ;
 $r^{(k)} \leftarrow r^{(k-1)} - \tau_k g$ ;
 $v_{k+1} \leftarrow \frac{\langle r^{(k)}, r^{(k)} \rangle}{\langle r^{(k-1)}, r^{(k-1)} \rangle}$ ;
 $s^{(k+1)} \leftarrow r^{(k)} + v_{k+1}s^{(k)}$ ;
 $k \leftarrow k + 1$ ;
END DO

```

В табл. 3.13 представлен оптимизированный псевдокод метода сопряжённых градиентов, отточенный за десятилетия его эксплуатации.

В нём присутствует вспомогательная величина  $g := As^{(k)}$ , введённая по той причине, что произведение  $As^{(k)}$  используется в алгоритме более одного раза. Вторая строка основного цикла псевдокода табл. 3.13 вычисляет величину очередного шага метода, а третья даёт следующее приближение к решению. Переменная  $r^{(k)}$  в тексте псевдокода означает антиградиент функционала энергии в точке  $x^{(k)}$ , и она специально набрана прямым шрифтом, отличным от стандартного математического наклонного шрифта («итэлик»), которым в тексте книги обозначается противоположная ему невязка  $r^{(k)}$ . Эта «антиневязка» или направление антиградиента функционала энергии в точке вновь найденного приближённого решения корректируется в четвёртой строке тела цикла согласно формуле (3.175). Вычисление невязки по определяющей её формуле не производится, так как оно гораздо менее устойчиво. В следующих двух строках (перед увеличением счётчика  $k$ ) вычисляется новое направление  $s^{(k+1)}$  движения к решению, которое будет использовано на следующем шаге.

Для ускорения метода сопряжённых градиентов и для расширения сферы его применимости часто используют специальное симметризующее предобусловливание системы. Его подробное описание читатель может увидеть, например, в книгах [11, 13].

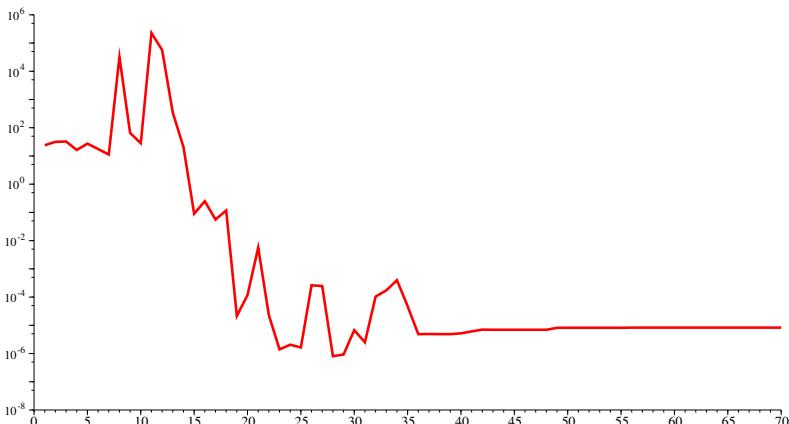


Рис. 3.31. График изменения невязки приближённого решения плохообусловленной системы в методе сопряжённых градиентов

**Пример 3.11.2** Рассмотрим решение методом сопряжённых градиентов задачи из примера 3.10.3, т. е.  $8 \times 8$ -системы линейных алгебраических уравнений с матрицей Гильберта и вектором правой части  $(1, -1, 1, -1, 1, -1, 1, -1)^\top$ . Спектральное число обусловленности матрицы системы  $\approx 1.5 \cdot 10^{10}$ , оно весьма велико, так что матрицу можно считать «почти особенной».

Метод сопряжённых градиентов справляется с этой плохообусловленной системой. График убывания невязки приближённого решения в алгоритме из табл. 3.13, который был реализован в системе компьютерной математики Scilab версии 6, показан в логарифмической шкале на рис. 3.31. Результаты работы того же самого алгоритма в других программных системах могут несколько отличаться от представленного, но главные качественные черты поведения метода сопряжённых градиентов будут примерно такими же.

Если потребовать, чтобы невязка приближённого решения стала меньшей  $10^{-5}$  в 1-норме, то метод сопряжённых градиентов выдаёт такое решение за 23 шага-итерации:

$$\begin{pmatrix} 868359.9672 \\ -46383622.2462 \\ 604807537.1852 \\ -3271357196.9926 \\ 8806671389.9489 \\ -12463050134.3752 \\ 8871750557.4972 \\ -2503935626.9622 \end{pmatrix}. \quad (3.181)$$

Его получение заняло три полных цикла метода. После этого момента можно лишь несущественно уменьшить норму невязки, а при дальнейшем итерировании норма невязки стабилизируется на уровне примерно  $5 \cdot 10^{-6}$  и уже не улучшается. В арифметике с плавающей точкой двойной точности, которая реализована в системе Scilab и других аналогичных, поддерживается примерно 16 знаков десятичного представления вещественных чисел (см. § 1.4). По этой причине ничего лучшего для рассматриваемой системы уравнений, решение которой имеет компоненты порядка  $10^9$ – $10^{10}$ , мы уже не сможем достичь.

В целом результат очень неплох. Напомним (пример 3.10.3), что метод Гаусса–Зейделя и метод релаксации, несмотря на оптимистичные теоретические результаты об их сходимости, на практике вообще не сходятся к чему-либо, хоть отдалённо напоминающему решение

рассматриваемой системы уравнений. Далее в примере 4.7.3 найденное выше решение успешно верифицируется с помощью доказательных интервальных методов. ■

### 3.11ж Другой подход к методу сопряжённых градиентов

Для метода сопряжённых градиентов широко распространена также другая трактовка, которая представляет его как проекционный метод (см. § 3.11а). В нём выполняется построение  $A$ -ортогонального базиса пространства и одновременное нахождение коэффициентов разложения решения по нему, т. е. строится последовательность расширяющихся подпространств, в каждом из которых ищется приближение к решению. Их последовательность за конечное число шагов даёт точное решение.

Пусть требуется найти решение системы линейных алгебраических уравнений

$$Ax = b$$

с симметричной и положительно определённой матрицей  $A$ . Пусть также  $s^{(1)}, s^{(2)}, \dots, s^{(n)}$  — базис  $\mathbb{R}^n$ , составленный из  $A$ -ортогональных векторов. Решение  $x^*$  системы уравнений можно искать в виде разложения по этому базису, т. е.

$$x^* = \sum_{i=1}^n x_i s^{(i)} \quad (3.182)$$

с какими-то неизвестными коэффициентами  $x_i, i = 1, 2, \dots, n$ .

Умножая обе части равенства (3.182) слева на матрицу  $A$  и учитывая, что  $Ax^* = b$ , будем иметь

$$\sum_{i=1}^n x_i (As^{(i)}) = b.$$

Если далее умножить скалярно это равенство на  $s^{(j)}, j = 1, 2, \dots, n$ , то получим  $n$  штук соотношений

$$\sum_{i=1}^n x_i \langle As^{(i)}, s^{(j)} \rangle = \langle b, s^{(j)} \rangle, \quad j = 1, 2, \dots, n. \quad (3.183)$$

Но в силу  $A$ -ортогональности системы векторов  $s^{(1)}, s^{(2)}, \dots, s^{(n)}$

$$\langle As^{(i)}, s^{(j)} \rangle = \langle s^{(i)}, s^{(j)} \rangle_A = \delta_{ij} = \begin{cases} 0, & \text{если } i \neq j, \\ 1, & \text{если } i = j, \end{cases}$$

так что от равенств (3.183) останется лишь

$$x_i \langle As^{(i)}, s^{(i)} \rangle = \langle b, s^{(i)} \rangle, \quad i = 1, 2, \dots, n.$$

Окончательно

$$x_i = \frac{\langle b, s^{(i)} \rangle}{\langle As^{(i)}, s^{(i)} \rangle}, \quad i = 1, 2, \dots, n, \quad (3.184)$$

откуда на основе (3.182) нетрудно восстановить искомое решение СЛАУ. Но для практического применения этого элегантного результата нужно уметь эффективно строить  $A$ -ортогональный базис  $s^{(1)}, s^{(2)}, \dots, s^{(n)}$  пространства  $\mathbb{R}^n$ .

Его построение можно выполнить, к примеру, как процесс  $A$ -ортогонализации невязок  $r^{(0)}, r^{(1)}, \dots, r^{(n-1)}$  последовательных приближений к решению  $x^{(0)}, x^{(1)}, \dots, x^{(n-1)}$ , и для реализации этой идеи идеально подходит ортогонализация Ланцшона (см. § 3.8). Вычислительный процесс, в котором выполняются ортогонализация невязок и одновременное нахождение коэффициентов разложения (3.184) по получающемуся базису, является конечным. Он завершается при некотором  $k \leq n$ , для которого  $r^{(k)} = 0$ , т. е. когда очередная невязка приближённого решения зануляется. Нетрудно понять, что в этой трактовке метода сопряжённых градиентов векторы  $s^{(1)}, s^{(2)}, \dots, s^{(n)}$ , образующие  $A$ -ортогональный базис пространства  $\mathbb{R}^n$ , — это не что иное, как последовательные направления спуска к минимуму функционала энергии.

Изложенный выше взгляд на метод сопряжённых градиентов ясно показывает роль циклов из  $n$  шагов как последовательных циклов построения ортогональных базисов пространства решений.

## 3.12 Методы установления

*Методы установления* — общее название для большой группы методов, в основе которых лежит идея искать решение рассматриваемой

стационарной задачи как предела по времени  $t \rightarrow \infty$  для решения связанной с ней вспомогательной нестационарной задачи. Этот подход к решению различных задач математической физики был развит в 30-е годы XX века А.Н. Тихоновым.

Пусть требуется решить систему линейных алгебраических уравнений

$$Ax = b,$$

где  $x \in \mathbb{R}^n$  — вектор-столбец неизвестных. Наряду с этой системой рассмотрим также систему дифференциальных уравнений

$$\frac{\partial x(t)}{\partial t} + Ax(t) = b, \quad (3.185)$$

в которой  $n$ -вектор неизвестных переменных  $x$  зависит от времени  $t$ . Ясно, что если в какой-то части своей области определения функция  $x(t)$  не изменяется в зависимости от переменной  $t$ , то производная  $dx/dt$  зануляется и соответствующие значения  $x(t)$  являются решением исходной задачи

Наиболее часто задачу (3.185) рассматривают на бесконечном интервале  $[t_0, \infty)$  и ищут её устанавливающееся решение, т. е. такое, что существует конечный  $\lim_{t \rightarrow \infty} x(t) = x^*$ . Тогда из свойств задачи (3.185) следует, что

$$\lim_{t \rightarrow \infty} \frac{\partial x}{\partial t} = 0,$$

и потому  $x^*$  является искомым решением для  $Ax = b$ .

При поиске значений  $x(t)$ , установившихся в пределе  $t \rightarrow \infty$ , нам не слишком интересны  $x(t)$  при конечных  $t$ . По этой причине для решения системы дифференциальных уравнений (3.185) можно применять самые простые численные методы. Таким, к примеру, является явный метод Эйлера (метод ломаных) с постоянным временным шагом  $\tau$ , в котором производная заменяется на разделённую разность вперёд [1, 4, 5, 26, 33, 40, 86]. Обозначая

$$t_k := t_0 + \tau k, \quad x^{(k)} := x(t_k), \quad k = 0, 1, 2, \dots,$$

получим вместо (3.185)

$$\frac{x^{(k)} - x^{(k-1)}}{\tau} + Ax^{(k-1)} = b, \quad (3.186)$$

или

$$x^{(k)} = x^{(k-1)} - \tau(Ax^{(k-1)} - b), \quad k = 1, 2, \dots$$

Это известный нам итерационный метод Ричардсона (3.138) для решения системы уравнений  $Ax = b$ . При переменном шаге по времени, когда  $\tau = \tau_k$ ,  $k = 1, 2, \dots$ , получающийся метод Эйлера

$$\frac{x^{(k)} - x^{(k-1)}}{\tau_k} + Ax^{(k-1)} = b$$

эквивалентен

$$x^{(k)} = x^{(k-1)} - \tau_k(Ax^{(k-1)} - b), \quad k = 1, 2, \dots,$$

т. е. простейшему нестационарному итерационному методу Ричардсона (3.157).

Представление итерационного метода Ричардсона в виде (3.186), как численного метода решения системы дифференциальных уравнений, даёт возможность понять суть ограничения на параметр  $\tau$ . Это не что иное, как ограничение на величину шага по времени, вызванное требованием устойчивости метода Эйлера. С другой стороны, если шаг по времени взят недостаточно большим, то до установления решения задачи (3.185) нам нужно сделать очень много таких мелких шагов, что даёт ещё одно объяснение невысокой вычислительной эффективности итераций Ричардсона.

Более быструю сходимость к решению можно достичь, взяв шаг по времени большим, но для этого нужно преодолеть ограничение на устойчивость метода. Реализация этой идеи действительно приводит к более эффективным численным методам решения некоторых специальных систем линейных уравнений  $Ax = b$ , встречающихся при дискретизации дифференциальных уравнений с частными производными. Таковы *методы переменных направлений*, *методы расщепления* и *методы дробных шагов*, идеально близкие друг другу [106].

Очевидно, что вместо (3.185) можно рассмотреть задачу более общего вида

$$B \frac{\partial x}{\partial t} + Ax(t) = b, \quad (3.187)$$

где  $B$  — некоторая неособенная матрица. Смысл её введения станет более понятен, если переписать (3.187) в равносильном виде

$$\frac{\partial x}{\partial t} + B^{-1}Ax(t) = B^{-1}b.$$

Тогда в пределе, при занулении  $\partial x / \partial t$ , имеем

$$B^{-1}Ax = B^{-1}b,$$

откуда видно, что матрица  $B$  выполняет роль, аналогичную роли пре-  
дубуславливающей матрицы для системы  $Ax = b$  (см. § 3.10в).

Отметим в заключение темы, что для решения систем линейных ал-  
гебраических уравнений, возникающих при дискретизации уравнений  
в частных производных эллиптического типа, предельно эффективны-  
ми являются *многосеточные методы*, предложенные Р.П. Федоренко и  
Н.С. Бахваловым в 60-е годы XX века.<sup>31</sup>

### 3.13 Теория А.А. Самарского

Системы линейных алгебраических уравнений, которые необходимо  
решать на практике, часто бывают заданы неявно, в операторном виде  
(см. § 3.6а). При этом невозможно работать с итерационными форму-  
лами вида (3.130) с явно заданным оператором  $T_k$  (наподобие (3.133)).  
Для таких случаев А.А. Самарским была предложена специальная ка-  
ноническая форма одношагового линейного итерационного процесса,  
предназначенного для решения систем линейных уравнений  $Ax = b$ :

$$B_k \frac{x^{(k)} - x^{(k-1)}}{\tau_k} + Ax^{(k-1)} = b, \quad k = 1, 2, \dots, \quad (3.188)$$

где  $B_k$ ,  $\tau_k$  — некоторые последовательности матриц и скалярных па-  
раметров соответственно, причём  $\tau_k > 0$ . Будем называть представление  
(3.188) *канонической формой Самарского*. Если  $x^{(k)}$  сходится к пределу,  
то при некоторых необременительных условиях на  $B_k$  и  $\tau_k$  этот предел,  
как легко убедиться, является решением системы линейных алгебраи-  
ческих уравнений  $Ax = b$ .

С учётом результатов предыдущего раздела нетрудно видеть, что  
форма Самарского фактически отражает взгляд на итерационные ме-  
тоды как на процессы установления для решения систем уравнений.

Различные последовательности матриц  $B_k$  и итерационных па-  
раметров  $\tau_k$  задают различные итерационные методы. Выбирая началь-  
ное значение  $x^{(0)}$ , находим затем из (3.188) последовательные прибли-  
жения как решения систем уравнений

$$B_k x^{(k)} = (B_k - \tau_k A) x^{(k-1)} + \tau_k b, \quad k = 1, 2, \dots \quad (3.189)$$

---

<sup>31</sup>В первых публикациях они по традиции назывались «релаксационными» [49].  
Осознание того, что это принципиально новый подход, которому требуется и новое  
название, пришло позже.

Ясно, что для однозначной разрешимости этой системы уравнений относительно  $x^{(k)}$  необходимо, чтобы все матрицы  $B_k$  были неособенными. Итерационный метод в форме (3.188) естественно назвать *явным*, если  $B_k = I$  — единичная матрица и выписанная выше система сводится к явной формуле для нахождения очередного итерационного приближения  $x^{(k)}$ . Иначе, если  $B_k \neq I$ , итерации (3.188) называются *неявными*. Неявные итерационные методы имеет смысл применять лишь в том случае, когда решение системы уравнений (3.189) относительно  $x^{(k)}$  существенно легче, чем решение исходной системы.

Выпишем представление в форме Самарского для рассмотренных ранее итерационных процессов. Итерационный метод Ричардсона из § 3.10г принимает вид

$$\frac{x^{(k)} - x^{(k-1)}}{\tau} + Ax^{(k-1)} = b, \quad k = 1, 2, \dots, \quad (3.190)$$

где  $\tau = \tau_k = \text{const}$  — постоянный параметр, имеющий тот же смысл, что и в § 3.10г. Переменный параметр  $\tau_k$  в (3.190) приводит к нестационарному методу Ричардсона (3.157) (§ 3.11а). Если  $D$  и  $\tilde{L}$  — соответственно диагональная и строго нижняя треугольная части матрицы  $A$  (см. § 3.10д), то методы Якоби и Гаусса–Зейделя можно записать в виде

$$D \frac{x^{(k)} - x^{(k-1)}}{1} + Ax^{(k-1)} = b \quad \text{и} \quad (D + \tilde{L}) \frac{x^{(k)} - x^{(k-1)}}{1} + Ax^{(k-1)} = b.$$

Итерационный метод релаксации с параметром  $\omega$  (см. § 3.10ж) в тех же обозначениях имеет форму Самарского

$$(D + \omega \tilde{L}) \frac{x^{(k)} - x^{(k-1)}}{\omega} + Ax^{(k-1)} = b, \quad k = 1, 2, \dots$$

При исследовании сходимости итераций в форме Самарского удобно пользоваться матричными неравенствами, связанными со знакоопределённостью матриц. Условимся для вещественной  $n \times n$ -матрицы  $G$  обозначать

$$G > 0, \quad \text{если } \langle Gx, x \rangle > 0 \quad \text{для всех ненулевых } n\text{-векторов } x,$$

т. е. если матрица  $G$  положительно определена. Из этого неравенства следует также существование такой константы  $\mu > 0$ , что  $\langle Gx, x \rangle >$

$\mu \langle x, x \rangle$ . Неравенство  $G \triangleright H$  будем понимать как  $\langle Gx, x \rangle > \langle Hx, x \rangle$  для всех  $x$ , что равносильно  $G - H \triangleright 0$ .

Достаточное условие сходимости итерационного процесса в форме Самарского (3.188) даёт

**Теорема 3.13.1** (теорема Самарского) *Если  $A$  — симметричная положительно определённая матрица,  $\tau > 0$  и  $B \triangleright \frac{1}{2}\tau A$ , то стационарный итерационный процесс*

$$B \frac{x^{(k)} - x^{(k-1)}}{\tau} + Ax^{(k-1)} = b, \quad k = 1, 2, \dots,$$

*сходится к решению системы уравнений  $Ax = b$  из любого начального приближения.*

**Доказательство.** Пусть  $x^*$  — решение системы уравнений  $Ax = b$ , так что

$$B \frac{x^* - x^*}{\tau} + Ax^* = b.$$

Если обозначить через  $z^{(k)} := x^{(k)} - x^*$  погрешность  $k$ -го приближения, то, как нетрудно проверить, она удовлетворяет однородному соотношению

$$B \frac{z^{(k)} - z^{(k-1)}}{\tau} + Az^{(k-1)} = 0, \quad k = 1, 2, \dots \quad (3.191)$$

Исследуем поведение погрешности в энергетической норме, порождающей матрицей  $A$ . Сначала покажем, что в условиях теоремы числовая последовательность  $\|z^{(k)}\|_A = \langle Az^{(k)}, z^{(k)} \rangle^{1/2}$ ,  $k = 1, 2, \dots$ , является невозрастающей.

Из соотношения (3.191) следует

$$z^{(k)} = (I - \tau B^{-1} A) z^{(k-1)}, \quad (3.192)$$

и

$$Az^{(k)} = (A - \tau AB^{-1} A) z^{(k-1)}.$$

Таким образом,

$$\begin{aligned} \langle Az^{(k)}, z^{(k)} \rangle &= \langle Az^{(k-1)}, z^{(k-1)} \rangle - \tau \langle AB^{-1} Az^{(k-1)}, z^{(k-1)} \rangle - \\ &- \tau \langle Az^{(k-1)}, B^{-1} Az^{(k-1)} \rangle + \tau^2 \langle AB^{-1} Az^{(k-1)}, AB^{-1} Az^{(k-1)} \rangle. \end{aligned}$$

Коль скоро матрица  $A$  симметрична,

$$\langle AB^{-1}Az^{(k-1)}, z^{(k-1)} \rangle = \langle Az^{(k-1)}, B^{-1}Az^{(k-1)} \rangle,$$

и потому

$$\begin{aligned} \langle Az^{(k)}, z^{(k)} \rangle &= \\ &= \langle Az^{(k-1)}, z^{(k-1)} \rangle - 2\tau \left\langle \left(B - \frac{1}{2}\tau A\right) B^{-1} Az^{(k-1)}, B^{-1} Az^{(k-1)} \right\rangle. \end{aligned} \quad (3.193)$$

Учитывая неравенство  $B \triangleright \frac{1}{2}\tau A$ , можем заключить, что вычитаемое в правой части полученного равенства всегда неотрицательно. По этой причине из (3.193) следует

$$\|z^{(k)}\|_A^2 \leq \|z^{(k-1)}\|_A^2,$$

так что последовательность  $\|z^{(k-1)}\|_A$  монотонно не возрастает и ограничена снизу нулём. В силу известной теоремы Вейерштрасса она имеет предел при  $k \rightarrow \infty$ .

Неравенство  $B \triangleright \frac{1}{2}\tau A$ , т. е. положительная определённость матрицы  $(B - \frac{1}{2}\tau A)$ , означает существование такого  $\eta > 0$ , что для любых  $y \in \mathbb{R}^n$

$$\left\langle \left(B - \frac{1}{2}\tau A\right) y, y \right\rangle \geq \eta \langle y, y \rangle = \eta \|y\|_2^2.$$

Как итог, из (3.193) получаем

$$\|z^{(k)}\|_A^2 - \|z^{(k-1)}\|_A^2 + 2\eta\tau \|B^{-1}Az^{(k-1)}\|_2^2 \leq 0$$

для всех  $k = 1, 2, \dots$ . Переходя в этом неравенстве к пределу по  $k \rightarrow \infty$ , видим, что должно быть  $\|B^{-1}Az^{(k)}\|_2 \rightarrow 0$ . Для неособенной матрицы  $B^{-1}A$  это возможно лишь при  $z^{(k)} \rightarrow 0$ . Итак, вне зависимости от выбора начального приближения итерационный процесс в самом деле сходится. ■

Отметим, что из теоремы Самарского следует теорема Островского–Райха (теорема 3.10.6) о сходимости метода релаксации для СЛАУ с симметричными положительно определёнными матрицами, а также, как её частный случай, теорема 3.10.5 о сходимости метода Гаусса–Зейделя. В самом деле, пусть  $A = \tilde{L} + D + \tilde{U}$  в обозначениях § 3.10д, т. е.  $\tilde{L}$  и  $\tilde{U}$  – строго нижняя и строго верхняя треугольные части матрицы  $A$ , а  $D$  – её диагональная часть. Если  $A$  симметрична, то  $\tilde{L} = \tilde{U}^\top$ , и поэтому

$$\langle Ax, x \rangle = \langle \tilde{L}x, x \rangle + \langle Dx, x \rangle + \langle \tilde{U}x, x \rangle = \langle Dx, x \rangle + 2\langle \tilde{L}x, x \rangle.$$

Тогда

$$\begin{aligned} \langle Bx, x \rangle - \frac{1}{2}\omega \langle Ax, x \rangle &= \langle (D + \omega \tilde{L})x, x \rangle - \frac{1}{2}\omega (\langle Dx, x \rangle + 2\langle \tilde{L}x, x \rangle) = \\ &= (1 - \frac{1}{2}\omega) \langle Dx, x \rangle > 0 \end{aligned}$$

при  $0 < \omega < 2$ , т. е. условии, требуемом леммой Кэхэна (см. § 3.10ж).

Дальнейшие результаты в этом направлении читатель может увидеть в книгах [40, 99].

### 3.14 Вычисление определителей матриц и обратных матриц

В некоторых задачах необходимо вычислять определитель квадратной матрицы. Сумма (3.7), которая выражает его определение, для  $n \times n$ -матрицы состоит из  $n!$  слагаемых, являющихся произведениями из  $n$  элементов матрицы. Соответственно, вычисление определителя по формуле (3.7) возможно лишь при небольших  $n$  из-за взрывного роста количества слагаемых и затрат на выполнение алгоритма. Как можно найти определитель матрицы с приемлемыми трудозатратами?

Предположим, что с матрицей  $A$  выполняется прямой ход метода исключения Гаусса для решения системы линейных алгебраических уравнений. Преобразования, выполняемые в той версии метода Гаусса, которая представлена в § 3.6, — это линейное комбинирование строк, и они не изменяют величины определителя матрицы. Следовательно,  $\det A$  равен определителю получающейся в итоге верхней треугольной матрицы  $U$ , т. е.  $\det A$  есть произведение диагональных элементов  $U$ .

Более общо, пусть  $A = LU$  — треугольное разложение матрицы  $A$ , которое можно получить с помощью прямого хода метода Гаусса, с помощью методов Дулитла и Кроута (§ 3.6e) или какими-то другими способами. Тогда, как известно из линейной алгебры,

$$\det A = \det L \cdot \det U,$$

а определители треугольных матриц  $L$  и  $U$  легко находятся перемножением их диагональных элементов. Если разложение матрицы  $A$  выполнено так, что по диагонали в нижней треугольной матрице  $L$  стоят все единицы (это достигается в методе Гаусса и в методе Дулитла), то  $\det L = 1$  и  $\det A = \det U$ . Если разложение матрицы  $A$  выполнено так,

что по диагонали в верхней треугольной матрице  $U$  стоят все единицы (это происходит в методе Кроута), то  $\det U = 1$  и  $\det A = \det L$ .

Совершенно аналогичные технологии можно организовать при использовании других матричных разложений. Пусть, например, найдено QR-разложение  $A = QR$ , т. е. представление исходной матрицы в виде произведения ортогональной  $Q$  и правой треугольной  $R$ . Тогда  $\det Q = \pm 1$  и, как правило, мы знаем свойства матрицы  $Q$ , т. е. в виде произведения какого количества каких элементарных ортогональных матриц — отражения или вращения — она получена. По этой причине нам точно известен её определитель, равный  $+1$  или  $-1$ . Искомый определитель  $\det A$  вычисляется по  $R$  как произведение её диагональных элементов и ещё  $\det Q$ .

Рассмотрим теперь вычисление матрицы, обратной к данной матрице. Отметим, что в современных вычислительных технологиях это приходится делать не слишком часто. Один из примеров, когда подобное вычисление необходимо по существу, — это нахождение дифференциала операции обращения матрицы  $A \mapsto A^{-1}$ , равного [14]

$$d(A^{-1}) = -A^{-1}(dA)A^{-1}.$$

Тогда производные решения системы уравнений  $Ax = b$  по элементам матрицы и правой части (т. е. коэффициенты чувствительности решения по отношению к коэффициентам и правым частям СЛАУ; см. § 1.7) даются формулами

$$\frac{\partial x_\nu}{\partial a_{ij}} = -z_{\nu i} \tilde{x}_j, \quad \frac{\partial x_\nu}{\partial b_i} = z_{\nu i}, \quad \nu = 1, 2, \dots, n,$$

где  $Z = (z_{ij}) = A^{-1}$  — обратная к матрице  $A$ ,  $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n)^\top$  — решение рассматриваемой системы уравнений.

Гораздо чаще встречается необходимость вычисления произведения обратной матрицы  $A^{-1}$  на какой-то вектор  $b$ , и это произведение всегда следует находить как решение системы уравнений  $Ax = b$  какими-либо из методов для решения СЛАУ. Такой способ как по точности, так и по трудоёмкости заведомо лучше, чем вычисление  $A^{-1}b$  с помощью явного нахождения обратной  $A^{-1}$ .

Матрица  $A^{-1}$ , обратная к данной матрице  $A$ , является решением матричного уравнения

$$AX = I.$$

Но это уравнение распадается на  $n$  уравнений относительно векторных неизвестных, соответствующих отдельным столбцам неизвестной матрицы  $X$ , и потому мы можем решать получающиеся уравнения порознь.

Из сказанного следует общий способ нахождения обратной матрицы: нужно решить  $n$  штук систем линейных уравнений

$$Ax = e^{(j)}, \quad j = 1, 2, \dots, n, \quad (3.194)$$

где  $e^{(j)}$  —  $j$ -й столбец единичной матрицы  $I$ . Это можно сделать, к примеру, любым из рассмотренных выше методов, причём прямые методы здесь особенно удобны в своей матричной трактовке. В самом деле, сначала мы можем выполнить один раз LU-разложение или QR-разложение исходной матрицы  $A$ , а затем хранить его и использовать посредством схемы (3.88) или (3.107) для различных правых частей уравнений (3.194). Если матрица  $A$  — симметричная положительно определённая, то весьма выгодно её разложение Холесского и последующее решение систем уравнений (3.194) с помощью представления (3.98).

В прямых методах решения СЛАУ прямой ход, т. е. приведение исходной системы к треугольному виду, является наиболее трудоёмкой частью всего алгоритма, которая требует в типичных случаях  $O(n^3)$  арифметических операций. Обратный ход (обратная подстановка) — существенно более лёгкая часть алгоритма, требующая всего  $O(n^2)$  операций. По этой причине изложенный выше рецепт однократного LU-разложения матрицы (или других разложений) позволяет сохранить общую трудоёмкость  $O(n^3)$  для алгоритма вычисления обратной матрицы.

Другой подход к обращению матриц — конструирование чисто матричных процедур, не опирающихся на методы решения систем линейных уравнений с векторными неизвестными. Известен итерационный метод Шульца для обращения матриц,<sup>32</sup> в котором после задания матрицы начального приближения  $X^{(0)}$  выполняют итерации

$$X^{(k)} \leftarrow X^{(k-1)}(2I - AX^{(k-1)}), \quad k = 1, 2, \dots \quad (3.195)$$

Последовательность матриц  $\{X^{(k)}\}$  сходится к обратной матрице  $A^{-1}$ , если начальное приближение  $X^{(0)}$  достаточно близко к  $A^{-1}$ . Метод

---

<sup>32</sup>Иногда этот метод называют также *методом Хотеллинга*, так как одновременно с Г. Шульцем [127] его рассматривал американский экономист и статистик Г. Хотеллинг [117]. Кроме того, встречается (редко) название *метод Бодевига*.

Шульца — это не что иное как метод Ньютона для решения системы уравнений, применённый к  $X^{-1} - A = 0$  (см. § 4.56).

**Предложение 3.14.1** *Метод Шульца сходится тогда и только тогда, когда его начальное приближение  $X^{(0)}$  удовлетворяет условию  $\rho(I - AX^{(0)}) < 1$ .*

**Доказательство.** Расчётную формулу метода Шульца можно переписать в виде

$$X^{(k)} = 2X^{(k-1)} - X^{(k-1)}AX^{(k-1)}, \quad k = 1, 2, \dots$$

Умножив обе части этого равенства слева на  $(-A)$  и добавив к ним по единичной матрице  $I$ , получим

$$I - AX^{(k)} = I - 2AX^{(k-1)} + AX^{(k-1)}AX^{(k-1)}.$$

Это равносильно

$$I - AX^{(k)} = (I - AX^{(k-1)})^2, \quad k = 1, 2, \dots,$$

откуда следует, что

$$I - AX^{(k)} = (I - AX^{(0)})^{2^k}, \quad k = 0, 1, 2, \dots \quad (3.196)$$

Если  $X^{(k)} \rightarrow A^{-1}$  при  $k \rightarrow \infty$ , то в левой части выписанного равенства  $I - AX^{(k)} \rightarrow 0$ . Соответственно, справа

$$(I - AX^{(0)})^{2^k} \rightarrow 0,$$

т. е. последовательность степеней матрицы сходится к нулю. Тогда необходимо  $\rho(I - AX^{(0)}) < 1$  в силу предложения 3.3.12.

Наоборот, если  $\rho(I - AX^{(0)}) < 1$ , то в правой части равенства (3.196) должно быть

$$(I - AX^{(0)})^{2^k} \rightarrow 0 \quad \text{при } k \rightarrow \infty$$

в силу следствия из предложения 3.10.2 (стр. 517). По этой причине должна иметь место сходимость  $X^{(k)} \rightarrow A^{-1}$ . ■

Из приведённого выше доказательства нетрудно вывести, что метод Шульца имеет квадратичную сходимость:

$$\|X^{(k)} - A^{-1}\| \leq C \|X^{(k-1)} - A^{-1}\|^2,$$

где  $C$  — некоторая константа. Эта оценка следует также из теоремы Канторовича о методе Ньютона (§ 4.5б).

В качестве достаточного условия сходимости из начального приближения  $X^{(0)}$  можно взять неравенство  $\|I - AX^{(0)}\| < 1$  для какой-нибудь удобной матричной нормы. Но в целом можно сказать, что метод Шульца лучше рассматривать как быструю уточняющую процедуру, так как он требует для своей сходимости выполнения довольно сильных условий на близость начального приближения к искомой обратной матрице.

### 3.15 Оценка погрешности приближённого решения

В этом параграфе рассматривается практически важный вопрос об оценке погрешности приближённого решения систем линейных алгебраических уравнений. Будут предложены три практических способа ответить на этот вопрос, хотя в действительности существует довольно много различных подходов к оценке погрешности решения.

#### Оценка с помощью нормы обратного оператора

Первый из излагаемых нами способов носит общий характер и может применяться в любых ситуациях, в частности, не обязательно в связи с какими-то конкретными численными методами.

Пусть  $\tilde{x}$  — приближённое решение системы уравнений  $Ax = b$ , а  $x^*$  — её точное решение. Принимая во внимание, что  $I = A^{-1}A$  и  $Ax^* = b$ , можем оценить погрешность  $\tilde{x}$  следующим образом:

$$\begin{aligned}\|\tilde{x} - x^*\| &= \|A^{-1}A\tilde{x} - A^{-1}Ax^*\| = \\ &= \|A^{-1}(A\tilde{x} - Ax^*)\| \leq \\ &\leq \|A^{-1}\| \|A\tilde{x} - b\|,\end{aligned}\tag{3.197}$$

где матричная и векторная нормы предполагаются согласованными. Величина  $(A\tilde{x} - b)$  — это невязка приближённого решения  $\tilde{x}$ , которая вычисляется непосредственно по  $\tilde{x}$ . Как следствие, погрешность решения можно узнать, найдя каким-либо образом или оценив сверху норму обратной матрицы  $\|A^{-1}\|$ .

Иногда на практике бывает известна информация о матрице  $A$ , которую можно использовать при оценке  $\|A^{-1}\|$ . Например, если рассматривается спектральная норма матрицы (2-норма), то

$$\|A^{-1}\|_2 = \sigma_{\max}(A^{-1}) = 1/\sigma_{\min}(A)$$

в силу предложения 3.2.6. Поэтому если известны сингулярные числа матрицы, то оценка нормы обратной к ней тривиальна.

Сингулярные числа матрицы равны абсолютным значениям собственных чисел для симметричных матриц. Далее, если  $A$  — симметричная положительно определённая матрица и известна нижняя граница её спектра  $\mu > 0$ , то из предложения 3.2.2 следует, что

$$\|A^{-1}\|_2 = \lambda_{\max}(A^{-1}) = (\lambda_{\min}(A))^{-1} \leq \mu^{-1}.$$

Напомним, что аналогичную информацию о спектре матрицы СЛАУ мы использовали при оптимизации скалярного предобуславливателя в § 3.10г. Описываемая ситуация типична при численном решении некоторых краевых задач для популярных уравнений математической физики, к примеру, для уравнения Лапласа и его обобщений. Для них дискретные аналоги соответствующих дифференциальных операторов хорошо изучены и известны оценки их собственных значений [17, 40].

**Пример 3.15.1** Предположим, что необходимо численно решить краевую задачу для дифференциального уравнения второго порядка на интервале  $[0, l]$ :

$$\begin{aligned} u''(x) &= f(x) \quad \text{на } [0, l], \\ u(0) &= a, \quad u(l) = b. \end{aligned} \tag{3.198}$$

Организуем на рассматриваемом интервале равномерную сетку  $\{x_i\}$  с шагом  $h$ , т. е.  $x_i = ih$ ,  $i = 0, 1, \dots, n$ , где  $h = l/n$ . Представим далее функцию непрерывного аргумента  $u(x)$  её дискретным образом — вектором значений  $(u_0, u_1, \dots, u_n)^\top$  в точках сетки. При этом  $u_0$  и  $u_n$  известны из условия задачи, так что вектором неизвестных является  $(u_1, u_2, \dots, u_{n-1})^\top$ .

Если для функции  $u(x)$  взять конечно-разностный аналог второй производной в виде (2.90), т. е.

$$u''(x_i) \approx \frac{u_{i-1} - 2u_i + u_{i+1}}{h^2},$$

то дискретным представлением краевой задачи (3.198) станет система линейных алгебраических уравнений

$$\frac{1}{h^2} \begin{pmatrix} -2 & 1 & & 0 \\ 1 & -2 & 1 & \\ & \ddots & \ddots & \ddots \\ 0 & & & 1 & -2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n-1} \end{pmatrix} = \begin{pmatrix} f(x_1) - a/h^2 \\ f(x_2) \\ \vdots \\ f(x_{n-1}) - b/h^2 \end{pmatrix},$$

матрица которой — симметричная трёхдиагональная. Численное решение нашей задачи сводится к нахождению решения этой системы.

Спектр матрицы системы линейных уравнений хорошо исследован (см., к примеру, учебник [40], § 1.2 главы 3 в части III), и собственные числа равны

$$\lambda_k = \frac{4}{h^2} \sin^2 \left( \frac{k\pi h}{2l} \right), \quad k = 1, 2, \dots, n-1.$$

Там же показано, что для собственных значений выполняются неравенства

$$\frac{9}{l^2} \leq \lambda_k < \frac{4}{h^2}, \quad k = 1, 2, \dots, n-1.$$

Нижняя граница собственных значений позволяет оценить сверху спектральную норму обратной матрицы системы линейных алгебраических уравнений как  $l^2/9$ , независимо от шага  $h$ . ■

Если матрица системы имеет диагональное преобладание, то для оценивания  $\|A^{-1}\|$  можно воспользоваться теоремой Алберга–Нильсона (теорема 3.4.3, стр. 421).

В общем случае нахождение  $\|A^{-1}\|$  или хотя бы разумной оценки для  $\|A^{-1}\|$  в какой-то норме, которые были бы менее трудоёмкими, чем решение исходной СЛАУ, является нетривиальным. Численные процедуры для этой цели обычно называют «оценщиками обусловленности», так как число обусловленности матрицы по определению тесно связано с нормой её обратной:

$$\operatorname{cond}(A) = \|A^{-1}\| \|A\|.$$

откуда следует

$$\|A^{-1}\| = \operatorname{cond}(A)/\|A\|.$$

Вычисление нужной нормы самой матрицы (или её оценки) обычно несложно и трудностей не представляет. Для этого можно использовать, например, какие-то удобные и легко вычислимые нормы, а затем неравенства эквивалентности из предложения 3.3.9 или им аналогичные. Поэтому из известной обусловленности матрицы или её оценки сверху нетрудно вывести оценку для нормы обратной матрицы. Краткий обзор существующих «оценщиков обусловленности», а также дальнейшие ссылки на литературу можно найти, к примеру, в книгах [13, 116].

### Оценка на основе исследования оператора перехода

Для некоторых численных методов оценка погрешности приближённого решения может быть получена как побочный продукт работы этих методов. Например, в стационарных однопшаговых итерационных методах последовательность погрешностей приближений своими свойствами очень близка к геометрической прогрессии, и этим обстоятельством можно с успехом воспользоваться.

Рассмотрим сходящийся стационарный однопшаговый итерационный метод в каноническом виде (3.131):

$$x^{(k)} \leftarrow Cx^{(k-1)} + d, \quad k = 1, 2, \dots$$

Как оценить отклонение по норме очередного приближения  $x^{(k)}$  от предела  $x^* := \lim_{k \rightarrow \infty} x^{(k)}$ , не зная самого этого предела и наблюдая лишь за итерационной последовательностью  $x^{(0)}, x^{(1)}, \dots, x^{(k)}, \dots$ ?

Зафиксируем какую-нибудь векторную норму и рассмотрим величину

$$\frac{\|x^{(k)} - x^*\|}{\|x^{(k-1)} - x^*\|}$$

— отношение норм погрешностей текущего и предыдущего приближений. Конечно, оно непостоянно, зависит от номера итерации и начального приближения, и поэтому удобнее использовать его верхний предел

$$\varkappa := \overline{\lim}_{k \rightarrow \infty} \frac{\|x^{(k)} - x^*\|}{\|x^{(k-1)} - x^*\|},$$

который всегда существует и который назовём *коэффициентом подавления погрешности* в рассматриваемом итерационном процессе отно-

сительно выбранной нормы. Из этого определения следует приближённое неравенство

$$\|x^{(k)} - x^*\| \lesssim \varkappa \|x^{(k-1)} - x^*\|, \quad (3.199)$$

что оправдывает название  $\varkappa$  и показывает, что для сходимости итерационного процесса достаточно  $\varkappa < 1$ . Пусть это неравенство выполняется далее.

Рассмотрим очевидное равенство

$$x^{(k-1)} - x^* = (x^{(k-1)} - x^{(k)}) + (x^{(k)} - x^*),$$

из которого в силу неравенства треугольника вытекает

$$\|x^{(k-1)} - x^*\| \leq \|x^{(k-1)} - x^{(k)}\| + \|x^{(k)} - x^*\|,$$

так что из определения  $\varkappa$  имеем

$$\|x^{(k-1)} - x^*\| \lesssim \|x^{(k-1)} - x^{(k)}\| + \varkappa \|x^{(k-1)} - x^*\|.$$

Перенесение в левую часть второго слагаемого из правой части и последующее деление обеих частей неравенства на положительную величину  $(1 - \varkappa)$  даёт

$$\|x^{(k-1)} - x^*\| \lesssim \frac{1}{1 - \varkappa} \|x^{(k)} - x^{(k-1)}\|.$$

Подставляя вместо  $\|x^{(k-1)} - x^*\|$  оценку (3.199), получаем окончательно

$$\|x^{(k)} - x^*\| \lesssim \frac{\varkappa}{1 - \varkappa} \|x^{(k)} - x^{(k-1)}\|, \quad (3.200)$$

где  $\varkappa$  — коэффициент подавления погрешности в рассматриваемом итерационном методе ( $\varkappa < 1$ ).

Выведенное неравенство может быть использовано на практике как для оценивания погрешности какого-то приближения из итерационной последовательности, так и для определения момента окончания итераций, т. е. того, достигнута ли желаемая погрешность приближения к решению. Но применение оценки (3.200) решающим образом зависит

от того, насколько эффективно и точно определяется коэффициент  $\kappa$  или же оценка для него сверху, меньшая единицы.

В стационарном одностадийном итерационном процессе

$$x^{(k)} = Cx^{(k-1)} + d \quad \text{и} \quad x^* = Cx^* + d.$$

Вычитание второго равенства из первого даёт

$$x^{(k)} - x^* = C(x^{(k-1)} - x^*), \quad (3.201)$$

а после применения к обеим частям векторной нормы получаем

$$\|x^{(k)} - x^*\| \leq \|C\| \|x^{(k-1)} - x^*\|, \quad (3.202)$$

где  $\|C\|$  — норма матрицы, согласованная с используемой векторной нормой. В случае, когда эта матричная норма является не просто согласованной, а ещё и подчинённой, неравенство становится достижимым на множестве всех значений разности  $x^{(k-1)} - x^*$  (см. определение 3.3.6). Если матрица  $C$  доступна в явном виде и к тому же  $\|C\| < 1$ , то можно применять неравенство (3.200), взяв в качестве верхней оценки для коэффициента подавления погрешности  $\kappa$  значение  $\|C\|$ .

Тем не менее такой выбор  $\kappa$  является довольно грубым и не самым удобным. Во-первых, непростым является определение матрицы  $C$  (которая может и не задаваться в явном виде). Во-вторых, конструкция нормы  $\|\cdot\|$ , в которой  $\|C\| < 1$ , также может быть неочевидной. Из результатов раздела § 3.106 следует, что теоретически такая норма всегда существует, если итерационный процесс сходится из любого начального приближения, но её конкретная конструкция в общем случае непроста. В-третьих, неравенство (3.202) является не самым точным, как будет видно из дальнейшего.

Более тонкую оценку коэффициента подавления погрешности можно дать на основе анализа последовательности приближений, порождаемой алгоритмом. Из (3.201) видно, что для всех  $k = 1, 2, \dots$  очередная разность  $(x^{(k)} - x^*)$  получается из предыдущей  $(x^{(k-1)} - x^*)$  умножением на одну и ту же матрицу  $C$ . Каждое такое умножение увеличивает в векторе долю направления, отвечающего доминирующему собственному значению, так что через достаточно большое число итераций сами эти векторы становятся почти совпадающими с соответствующими собственными векторами  $C$ . Этот эффект хорошо известен и тщательно изучен, он лежит в основе степенного метода нахождения доминирующего собственного значения матрицы (см. подробности в § 3.18a). Для

наших целей важно то, что при некоторых дополнительных условиях на матрицу перехода векторы  $(x^{(k)} - x^*)$  и  $(x^{(k-1)} - x^*)$  с ростом  $k$  делаются почти коллинеарными, и коэффициент их пропорциональности близок к наибольшему по модулю собственному значению. Тогда отношение их норм — это спектральный радиус матрицы. Итак, при достаточно больших  $k$  коэффициент подавления погрешности можно положить  $\varkappa = \rho(C)$  — спектральному радиусу матрицы перехода  $C$ .

Как найти или оценить  $\rho(C)$ ? Снова рассмотрим последовательность векторов  $\{x^{(k)}\}$ , порождаемую итерационным процессом (3.131),

$$x^{(k)} \leftarrow Cx^{(k-1)} + d, \quad k = 1, 2, \dots,$$

так что

$$x^{(k)} = Cx^{(k-1)} + d \quad \text{и} \quad x^{(k+1)} = Cx^{(k)} + d.$$

Вычитая первое равенство из второго, получим

$$x^{(k+1)} - x^{(k)} = C(x^{(k)} - x^{(k-1)}), \quad (3.203)$$

т. е. векторы последовательных разностей  $(x^{(k+1)} - x^{(k)})$  и  $(x^{(k)} - x^{(k-1)})$  для всех  $k = 1, 2, \dots$  получаются умножением на одну и ту же постоянную матрицу  $C$ . В частности,

$$x^{(k+1)} - x^{(k)} = C^k (x^{(1)} - x^{(0)}).$$

Про поведение последовательности разностей  $(x^{(k)} - x^{(k-1)})$  можно сказать то же самое, что и выше: она сходится к нулю при любом  $x^{(0)}$  тогда и только тогда, когда  $\rho(C) < 1$ . Сами разности  $(x^{(k)} - x^{(k-1)})$  с ростом  $k$  становятся почти коллинеарными друг другу, если у матрицы существует доминирующее собственное значение, в направлении которого происходит преимущественное растяжение этих векторов на каждом шаге. Тогда отношение норм

$$\frac{\|x^{(k+1)} - x^{(k)}\|}{\|x^{(k)} - x^{(k-1)}\|}, \quad k = 1, 2, \dots, \quad (3.204)$$

должно приближаться к модулю этого доминирующего собственного значения матрицы перехода  $C$ , т. е. к её спектральному радиусу. Конкретный выбор векторной нормы в выписанном отношении роли уже не играет.

Определив отношение (3.204) из трёх последовательных итераций при достаточно большом  $k$ , мы сможем воспользоваться найденным значением коэффициента  $\kappa$  в неравенстве (3.200) для оценки погрешности.

**Пример 3.15.2** Рассмотрим систему линейных алгебраических уравнений (3.155) из примера 3.10.1,

$$\begin{pmatrix} 6 & 2 & 2 \\ 2 & 3 & 4 \\ 2 & 4 & 8 \end{pmatrix} x = \begin{pmatrix} 6 \\ 3 \\ 6 \end{pmatrix},$$

решением которой является  $(1, -1, 1)^\top$ . Предположим, что для решения этой системы организован итерационный метод Гаусса–Зейделя с нулевым начальным приближением. Когда компоненты очередного вычисленного приближения к решению станут отличаться от точного решения не более чем на  $10^{-3}$ ?

Ответ на поставленный вопрос требует чебышёвской нормы  $\|\cdot\|_\infty$  для измерения отклонения векторов друг от друга, и выражение для соответствующей подчинённой матричной нормы даётся в предложении 3.3.7.

Запустив итерации Гаусса–Зейделя, мы увидим, что

$$\begin{aligned} x^{(0)} &= (0, 0, 0)^\top, \\ x^{(1)} &= (1, 0.333333, 0.333333)^\top, \\ x^{(2)} &= (0.777778, 0.037037, 0.537037)^\top, \\ &\dots &&\dots &&\dots \\ x^{(20)} &= (0.999495, -0.998122, 0.999187)^\top, \\ x^{(21)} &= (0.999645, -0.998679, 0.999429)^\top, \\ x^{(22)} &= (0.999750, -0.999072, 0.999598)^\top \end{aligned}$$

(результаты даны с шестью значащими цифрами). Отношение (3.204), оценивающее спектральный радиус матрицы перехода, уже с  $k = 8$  устанавливается на значении примерно 0.7031. Подставляя его в неравенство (3.200) вместо  $\kappa$ , получим

$$\frac{\kappa}{1 - \kappa} = \frac{0.7031}{1 - 0.7031} \approx 2.368,$$

и потому в итерационном методе Гаусса–Зейделя для нашей системы уравнений при достаточно больших  $k$  должна выполняться оценка

$$\|x^{(k)} - x^*\|_\infty \lesssim 2.368 \|x^{(k)} - x^{(k-1)}\|_\infty.$$

Из неё следует, что необходимая разность между двумя последовательными приближениями должна быть не больше  $10^{-3}/2.368$ , и она достигается только на разности 21-й и 22-й итераций. Итак, 22-я итерация даёт необходимое приближение к точному решению системы  $(1, -1, 1)^\top$ , что можно видеть непосредственно.

Чтобы использовать в оценке коэффициент подавления погрешности в виде нормы матрицы перехода, нужно сначала построить эту матрицу, что также требует трудозатрат. Согласно (3.148) для итерационного метода Гаусса–Зейделя она равна

$$-\begin{pmatrix} 6 & 0 & 0 \\ 2 & 3 & 0 \\ 2 & 4 & 8 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 2 & 2 \\ 0 & 0 & 4 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -0.33333 & -0.33333 \\ 0 & 0.22222 & -1.11111 \\ 0 & -0.02778 & 0.63889 \end{pmatrix}.$$

Её спектральный радиус — 0.703075, и это совершенно согласуется с результатами численных расчётов выше. Но чебышёвская норма матрицы перехода равна 1.33333, она больше единицы, а потому использовать её значение в нашей технологии нельзя. Аналогична ситуация и с другими популярными матричными нормами. ■

В принципе, возможны ситуации, когда последовательность отношений (3.204) ведёт себя хаотически и никуда не сходится. Это свидетельствует о том, что в матрице перехода нет одного доминирующего собственного значения, а есть нескольких собственных значений с одним и тем же модулем, превосходящим модули остальных собственных чисел. Тогда требуется более тонкий анализ результатов итераций, который всё-таки позволяет получать информацию о спектральном радиусе матрицы перехода. Можно прочитать о нём в книге [48] в главе о степенном методе.

### Апостериорное оценивание

Ещё один способ оценки погрешности приближённых решений систем линейных алгебраических уравнений опирается на методы интервального анализа, но его суть проста и естественна. Часто его называют *апостериорным оцениванием*.

Этот способ заключается в том, что найденное приближённое решение  $\tilde{x}$  окружают интервальным бруском  $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n)$ ,  $\mathbf{X} \ni \tilde{x}$ , для которого затем доказывается численно, на основе некоторых результаты анализа (теорема Брауэра, теорема Миранды и др.; см. главу 4), гарантированное присутствие решения рассматриваемой системы уравнений.

Ширина компонент бруса  $\mathbf{X}$  будет при этом оценкой погрешности приближения  $\tilde{x}$ . Для систем линейных алгебраических уравнений области значений линейных функций, которые необходимы для применения теорем Брауэра, Миранды и пр. легко получаются с помощью естественного интервального расширения (см. § 1.6). Построение бруса  $\mathbf{X}$ , который должен гарантированно содержать точное решение СЛАУ, выполняется обычно с помощью процедуры, которая «раздувает», иногда за несколько шагов, найденное приближение  $\tilde{x}$  до телесного бруса, содержащего  $\tilde{x}$ . Она называется  $\epsilon$ -раздутьем, так как зависит от параметра  $\epsilon$ , подстраиваемого под задачу.

В заключение отметим, что интервальные методы доказательных вычислений и оценивания погрешностей являются очень развитыми и имеют в своём арсенале много практических и эффективных подходов. В частности, метод апостериорного оценивания погрешностей успешно обобщается на нелинейные уравнения и системы уравнений. В книге эти методы не рассматриваются детально, но читатель может получить некоторое представление об их организации, работе и характере результатов по § 4.7. Современные обзоры по теме можно найти, к примеру, в [120, 125].

## 3.16 Линейная задача наименьших квадратов

Для заданных  $m \times n$ -матрицы  $A$  и  $m$ -вектора  $b$  линейной задачей наименьших квадратов называют задачу отыскания  $n$ -вектора  $x$ , который доставляет минимум евклидовой норме невязки  $\|Ax - b\|_2$  системы линейных алгебраических уравнений  $Ax = b$  или, что равносильно, квадратичной форме  $\langle Ax - b, Ax - b \rangle$ .

Ясно, что для матриц  $A$  полного ранга в случае  $m \leq n$ , когда матрица является квадратной или лежачей, искомый минимум, как правило, равен нулю. Для квадратной матрицы  $A$ , когда  $m = n$ , линейная задача наименьших квадратов фактически равносильна решению системы

линейных алгебраических уравнений  $Ax = b$  и несёт особую специфику лишь когда  $A$  имеет неполный ранг, т. е. особенна. Теоретически и практически наиболее важный случай линейной задачи наименьших квадратов соответствует  $m \geq n$ . Он находит многочисленные и разнообразные применения при обработке данных, и мы занимались им в главе 2.



Рис. 3.32. Структурная схема объекта идентификации.

Рассмотрим в качестве примера *задачу идентификации параметров* объекта, часто называемую также *задачей восстановления зависимостей*. Предположим, что имеется объект, на вход которому подаются воздействия, описываемые вектором  $a = (a_1, a_2, \dots, a_n) \in \mathbb{R}^n$ , а на выходе получается величина  $b \in \mathbb{R}$  (рис. 3.32). Внутреннее устройство исследуемого объекта неизвестно в деталях, но для работы с ним часто вполне достаточно знания функциональной зависимости  $b$  от переменных  $a_1, a_2, \dots, a_n$ . Простейший и часто встречающийся случай — однородная линейная зависимость «вход-выход», имеющая вид

$$b = x_1 a_1 + x_2 a_2 + \dots + x_n a_n \quad (3.205)$$

с некоторыми постоянными коэффициентами  $x_1, x_2, \dots, x_n$ . Задача идентификации — это задача определения (оценивания) значений  $x_j$  на основе данных о входах и выходе объекта, т. е. по ряду соответствующих друг другу значений  $(a_1, a_2, \dots, a_n)$  и  $b$ .

Каждое наблюдение (измерение) входов и выходов объекта порождает соотношение вида (3.205), которое связывает искомые  $x_1, x_2, \dots, x_n$ . Если серия измерений «входы-выход» объекта является «достаточно представительной», то можно попытаться решить получившуюся систему уравнений относительно неизвестных  $x_j$  и найти их значения. Восстанавливаемую функциональную зависимость (3.205) в статистике часто называют *регрессией* величины  $b$  по величинам  $a_1, a_2, \dots, a_n$ ,

а её график — *регрессионной линией* или *поверхностью регрессии*  $b$  по  $a_1, a_2, \dots, a_n$  (в зависимости от размерности  $n$ ).

Для удобства выкладок условимся обозначать результат  $i$ -го измерения входов нашего объекта через  $(a_{i1}, a_{i2}, \dots, a_{in})$ , а выхода — через  $b_i$ ,  $i = 1, 2, \dots, m$ . Тогда решение задачи идентификации — это вектор  $x = (x_1, x_2, \dots, x_n)^\top$ , который удовлетворяет всем соотношениям вида (3.205), получившимся в результате отдельных наблюдений. Иными словами, результат идентификации  $x = (x_1, x_2, \dots, x_n)^\top$  является решением системы уравнений

$$\left\{ \begin{array}{l} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = b_2, \\ \vdots \qquad \vdots \qquad \ddots \qquad \vdots \qquad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n = b_m, \end{array} \right. \quad (3.206)$$

где  $m$  — общее количество измерений (наблюдений). Найдя его, получим решение рассматриваемой задачи идентификации.

Но на практике описанная выше ясная и стройная схема почти не работает, так как система уравнений (3.206), как правило, решений не имеет. Это вызвано тем, что она обычно переопределена: измерений стремятся получить как можно больше, так как каждое из них может содержать какую-то информацию об интересующих нас параметрах. В то же время данные измерений (наблюдений) содержат случайные ошибки, которые делают невозможными точные равенства (3.205).

Кроме того, назначая избранный вид зависимости между входными и выходной переменными (в нашем случае — линейный), мы тоже можем совершать идеализацию, желая, к примеру, выявить только «главную часть» реальной функциональной зависимости, которая связывает  $b$  и  $a_1, a_2, \dots, a_n$ . В этих условиях достижение точных равенств в уравнениях выписанной системы в принципе невозможно.

Следовательно, для системы уравнений (3.206) мы должны искать решение в каком-то обобщённом смысле, который подразумевает выполнение более мягких условий, чем точное равенство в (3.205). Чаще всего в качестве такого решения ищут набор значений  $x_1, x_2, \dots, x_n$ , минимизирующий норму невязки системы (3.206), т. е. разности между её правой и левой частями. В случае, когда рассматривается евклидова норма невязки, приходим к линейной задаче наименьших квадратов.

**Определение 3.16.1** Вектор, на котором достигается минимум евклидовой нормы невязки системы линейных алгебраических уравнений

ний, называют псевдорешением относительно евклидовой нормы или псевдорешением в смысле наименьших квадратов.

Выведем условия, которым удовлетворяет псевдорешение системы линейных алгебраических уравнений относительно евклидовой нормы. Предположим, что  $e$  —  $n$ -вектор единичной длины,  $\|e\|_2 = 1$ . Введём функцию

$$\Phi(x) = \|Ax - b\|_2^2 = \langle Ax - b, Ax - b \rangle$$

— квадрат евклидовой нормы невязки, и найдём её производную в точке  $x$  по направлению вектора  $e$ . С квадратом евклидовой нормы невязки работать удобнее, так как его выражение не содержит корня, но его минимум достигается при тех же условиях, что и для самой нормы. Согласно определению [14]

$$\begin{aligned} \frac{\partial \Phi(x)}{\partial e} &= \lim_{t \rightarrow 0} \frac{\Phi(x + te) - \Phi(x)}{t} = \\ &= \lim_{t \rightarrow 0} \frac{\langle A(x + te) - b, A(x + te) - b \rangle - \langle Ax - b, Ax - b \rangle}{t} = \\ &= \lim_{t \rightarrow 0} \frac{\langle A(te), Ax - b \rangle + \langle Ax - b, A(te) \rangle + \langle A(te), A(te) \rangle}{t} = \\ &= \lim_{t \rightarrow 0} \frac{2t \langle A^\top(Ax - b), e \rangle + t^2 \langle Ae, Ae \rangle}{t} = \\ &= \lim_{t \rightarrow 0} \frac{2t \langle A^\top(Ax - b), e \rangle}{t} + \lim_{t \rightarrow 0} \frac{t^2 \langle Ae, Ae \rangle}{t} = \\ &= 2 \langle A^\top Ax - A^\top b, e \rangle. \end{aligned}$$

В точке минимума производная функции по любому направлению равна нулю, так что  $\langle A^\top Ax - A^\top b, e \rangle = 0$  для любого вектора  $e$ . По этой причине должно быть  $A^\top Ax - A^\top b = 0$ .

Система линейных алгебраических уравнений

$$A^\top Ax = A^\top b, \tag{3.207}$$

как известно, называется *нормальной системой уравнений* для линейной задачи наименьших квадратов с матрицей  $A$  и вектором  $b$  (см. § 2.11г). Таким образом, выше ещё раз показано, в дополнение к результатам § 2.11г, что минимум евклидовой нормы невязки достигается на

решении нормальной системы уравнений. Это решение, как установлено в § 2.11г, всегда существует и в самом деле доставляет минимум выражению  $\Phi(x) = \|Ax - b\|_2^2$ , поскольку матрица вторых производных (гессиан) функции  $\Phi(x)$  является положительно полуопределённой матрицей  $A^\top A$ .

Переход от исходной системы уравнений  $Ax = b$  к нормальной системе (3.207) носит название *первой трансформации Гаусса*.<sup>33</sup>

Исследуем единственность псевдорешения. Любое решение линейной задачи наименьших квадратов является также решением нормальной системы уравнений (3.207), так что наш вопрос сводится к следующему: когда нормальная система уравнений имеет единственное решение? Как известно из линейной алгебры, общее решение системы линейных алгебраических уравнений есть сумма частного решения этой системы и общего решения однородной системы. Следовательно, вопрос упрощается до такого: когда однородная нормальная система уравнений  $A^\top A\tilde{x} = 0$  имеет только нулевое решение?

Если ненулевой вектор  $\tilde{x}$  таков, что  $A^\top A\tilde{x} = 0$ , то и

$$\tilde{x}^\top (A^\top A\tilde{x}) = 0,$$

и потому

$$\tilde{x}^\top (A^\top A\tilde{x}) = (\tilde{x}^\top A^\top)(A\tilde{x}) = (A\tilde{x})^\top (A\tilde{x}) = \|A\tilde{x}\|_2^2 = 0.$$

Получаем

$$A\tilde{x} = 0.$$

Итак, однородная нормальная система уравнений  $A^\top A\tilde{x} = 0$  имеет только нулевое решение тогда и только тогда, когда однородная система  $Ax = 0$  имеет только нулевое решение. Как следствие, справедлива

**Теорема 3.16.1** *Линейная задача наименьших квадратов с матрицей  $A$  имеет единственное решение в том и лишь том случае, когда столбцы  $A$  являются линейно независимыми.*

**Определение 3.16.2** *Линейная задача наименьших квадратов, у которой матрица имеет линейно независимые столбцы, называется*

---

<sup>33</sup>Существует также «вторая трансформация Гаусса» систем линейных алгебраических уравнений, при которой вместо системы  $Ax = b$  мы решаем систему  $AA^\top y = b$ , а затем находим решение исходной системы как  $x = A^\top y$ .

*линейной задачей полного ранга. Линейная задача наименьших квадратов с матрицей, столбцы которой линейно зависимы, называется линейной задачей неполного ранга.*

Отметим особенность терминологии: ранг  $m \times n$ -матрицы совпадает с рангом линейной задачи наименьших квадратов с этой матрицей в случае  $m \geq n$ , тогда как при  $m < n$  эти понятия различны.

В случае неединственности псевдорешения для выделения единственного решения линейной задачи наименьших квадратов дополнительно налагаются на псевдорешение какие-нибудь условия. Например, это может быть требование, чтобы псевдорешение имело наименьшую возможную евклидову норму.

**Определение 3.16.3** *Псевдорешение системы линейных алгебраических уравнений с наименьшей евклидовой нормой (т. е. 2-нормой) называется нормальным псевдорешением.*

**Предложение 3.16.1** *Нормальное псевдорешение системы линейных алгебраических уравнений единственно.*

Доказательство немедленно следует из теоремы Бореля (теорема 2.10.1, стр. 190) или из теоремы о перпендикуляре (§ 2.11в).

На практике применяется несколько подходов к решению линейной задачи наименьших квадратов. Самым первым способом, восходящим ещё к К.Ф. Гауссу, является непосредственное решение нормальной системы уравнений (3.207). Матрица нормальной системы уравнений симметрична и положительно определена, если задача имеет полный ранг. Это позволяет применять к ней такие эффективные алгоритмы как метод Холесского, метод сопряжённых градиентов и другие. Недостаток этого способа состоит в том, что обусловленность нормальной системы (3.207) равна квадрату обусловленности исходной, т. е. существенно ухудшается. В самом деле,

$$\operatorname{cond}_2(A^\top A) = \|A^\top A\|_2 \|(A^\top A)^{-1}\|_2.$$

При этом  $\|A^\top A\|_2 = \sigma_{\max}^2(A)$  и  $\|(A^\top A)^{-1}\|_2 = \sigma_{\max}^2(A^{-1})$ , так что

$$\operatorname{cond}_2(A^\top A) = \operatorname{cond}_2^2(A).$$

Но если размеры задачи не очень велики и обусловленность матрицы  $A$  исходной системы не слишком плоха, этим невыгодным обстоятельством можно пренебречь.

Ещё один идейно близкий способ — представить решение линейной задачи наименьших квадратов в виде решения расширенной системы линейных уравнений. В самом деле, обозначив  $y = Ax$ , нормальную систему уравнений (3.207) можно переписать в виде

$$\begin{cases} A^\top y = A^\top b, \\ Ax - y = 0. \end{cases}$$

Если ввести  $(n+m)$ -вектор неизвестных  $(x, y)^\top$ , то выписанной системе можно придать блочно-матричную форму

$$\begin{pmatrix} 0 & A^\top \\ A & -I \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} A^\top b \\ 0 \end{pmatrix}. \quad (3.208)$$

С другой стороны, нормальную систему можно переписать несколько по-другому, в виде  $A^\top(b - Ax) = 0$ , и затем как эквивалентную расширенную линейную систему

$$\begin{cases} b - Ax = z, \\ A^\top z = 0. \end{cases}$$

Её блочно-матричная форма с  $(m+n)$ -вектором неизвестных  $(z, x)^\top$  выглядит следующим образом:

$$\begin{pmatrix} I & A \\ A^\top & 0 \end{pmatrix} \begin{pmatrix} z \\ x \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix}. \quad (3.209)$$

Расширенные системы (3.208)–(3.209) имеют симметричные матрицы размера  $(m+n) \times (m+n)$ , но их достоинство по сравнению с нормальной системой уравнений (3.207) — отсутствие необходимости перемножения матриц  $A^\top$  и  $A$  и меньший рост числа обусловленности.

Другими подходами к решению линейной задачи наименьших квадратов являются прямые методы на основе ортогональных преобразований, описанные ранее в этой главе. Более точно, это метод вращений из § 3.7г и метод отражений Хаусхольдера из § 3.7е, в обратном ходе которых используется обратная подстановка для трапециевидных линейных систем из § 3.6б. Технологические детали этих способов численного решения линейной задачи наименьших квадратов читатель может увидеть, например, в [13, 29, 46].

Наконец, ещё один подход к решению линейной задачи наименьших квадратов основан на сингулярном разложении матрицы системы и рассмотрен в § 3.5б. Он более сложен технически, но обеспечивает наиболее полный анализ задачи. Его подробности описаны, например, в книге [29].

## 3.17 Матричная проблема собственных значений

### 3.17а Обсуждение постановки задачи

Ненулевой вектор  $v$  называется *собственным вектором* квадратной матрицы  $A$ , если в результате умножения на эту матрицу он переходит в коллинеарный себе, т. е. отличающийся от исходного только некоторым скалярным множителем:

$$Av = \lambda v. \quad (3.210)$$

Сам скаляр  $\lambda$ , который является коэффициентом пропорциональности исходного вектора и его образа при действии матрицы, называют *собственным значением* или *собственным числом* матрицы. Соответственно, *проблемой собственных значений* называется задача определения собственных значений и собственных векторов матриц: для заданной  $n \times n$ -матрицы  $A$  найти числа  $\lambda$  и  $n$ -векторы  $v \neq 0$ , удовлетворяющие условию (3.210).

Собственные значения и собственные векторы матриц нужно знать во многих приложениях. Например, задача определения частот собственных колебаний механических систем (весьма актуальная при проектировании различных конструкций) требует нахождения собственных значений так называемых матриц жёсткости этих систем. Особую важность собственным значениям придаёт то обстоятельство, что соответствующие им частоты собственных колебаний являются непосредственно наблюдаемыми из опыта физическими величинами. Это тон звучания тронутой гитарной струны и т. п. Хороший обзор мотиваций и приложений проблемы собственных значений читатель может увидеть, например, в последней главе книги [126].

**Пример 3.17.1** Пусть  $A$  —  $n \times n$ -матрица,  $x^{(k)}$  и  $b^{(k)}$ ,  $k = 0, 1, 2, \dots$ , — семейства  $n$ -векторов. Линейные динамические системы с дискретным

временем вида

$$x^{(k)} = Ax^{(k-1)} + b^{(k)}, \quad k = 1, 2, \dots, \quad (3.211)$$

служат моделями разнообразных процессов окружающего нас мира, от биологии до экономики. Фактически, (3.211) означает, что в  $k$ -й момент времени интересующая нас величина (объём производства некоторого продукта, численность биологического вида и т. п.) является линейной функцией от своего значения в  $k$ -й момент времени, причём коэффициенты и свободный член этой линейной зависимости постоянны.

Общее решение такой системы есть сумма частного решения исходной системы (3.211) и общего решения однородной системы  $x^{(k)} = Ax^{(k-1)}$  без свободного члена. Если искать нетривиальные решения однородной системы в виде  $x^{(k)} = \lambda^k h$ , где  $\lambda$  — ненулевой скаляр и  $h$  —  $n$ -вектор, то нетрудно убедиться, что  $\lambda$  должно быть собственным значением  $A$ , а  $h$  — собственным вектором матрицы  $A$ . ■

Система уравнений (3.210) кажется недоопределённой, так как содержит  $n+1$  неизвестных, которые нужно найти из  $n$  уравнений. Но на самом деле можно дополнить её, к примеру, каким-нибудь условием нормировки собственных векторов ( $\|v\| = 1$  в какой-то норме) или требованием, чтобы какая-либо компонента  $v$  принимала бы заданное значение. Последнее условие иногда даже более предпочтительно ввиду своей линейности.

Из (3.210) следует

$$(A - \lambda I)v = 0,$$

что при ненулевом векторе  $v$  означает наличие нетривиальной линейной зависимости между столбцами матрицы  $A - \lambda I$ . Итак, должно быть

$$\det(A - \lambda I) = 0. \quad (3.212)$$

Это уравнение относительно переменной  $\lambda$  называется, как известно, *характеристическим уравнением* для матрицы  $A$ .<sup>34</sup> Оно является алгебраическим уравнением степени  $n$ , что следует из формулы для разложения определителя

$$\det(A - \lambda I) = (-1)^n \lambda^n + p_{n-1} \lambda^{n-1} + \dots + p_1 \lambda + p_0,$$

---

<sup>34</sup> В механике его называют также «вековым уравнением».

где  $p_{n-1}, \dots, p_1, p_0$  — какие-то выражения от элементов матрицы  $A$ . Переход от исходной задачи (3.210) к характеристическому уравнению (3.212) позволяет расчленить задачу и избавиться от неизвестного собственного вектора  $v$ .

Если собственное значение  $\tilde{\lambda}$  матрицы  $A$  уже найдено, то определение соответствующих собственных векторов сводится к решению системы линейных алгебраических уравнений

$$(A - \tilde{\lambda}I)x = 0 \quad (3.213)$$

с особенной матрицей. Но на практике часто предпочитают пользоваться для нахождения собственных векторов специализированными вычислительными процедурами. Многие из них позволяют вычислять собственные векторы одновременно с собственными значениями матриц.

В силу основной теоремы алгебры [23, 43, 47] в поле комплексных чисел  $\mathbb{C}$  характеристическое уравнение имеет с учётом кратности  $n$  решений. Но, вообще говоря, вещественных решений характеристическое уравнение может не иметь.

Таким образом, всякая  $n \times n$ -матрица в поле комплексных чисел имеет с учётом кратности ровно  $n$  собственных чисел. Собственных векторов может быть меньше, но хотя бы один, являющийся решением системы (3.213), всегда существует. Тем не менее, даже если рассматриваемая матрица  $A$  вещественна, могут не существовать вещественные  $\lambda$  и  $v$ , удовлетворяющие соотношению

$$Av = \lambda v.$$

В целом для математически полного исследования проблемы собственных значений необходим выход в поле комплексных чисел  $\mathbb{C}$ , в котором всякий алгебраический полином степени  $\geq 1$  с коэффициентами из этого поля имеет хотя бы один нуль. Напомним, что такие поля называются *алгебраически замкнутыми* [23, 47]. Только в  $\mathbb{C}$  можно полноценно «увидеть» все собственные значения матрицы и её собственные векторы. Но полезность такого выхода для практического применения собственных чисел и собственных векторов матрицы в каждом конкретном случае должна рассматриваться отдельно.

Нередко при упоминании рассматриваемой задачи подчёркивают — «матричная проблема собственных значений»<sup>35</sup>, чтобы уточнить, что

---

<sup>35</sup>Раньше использовался также термин «алгебраическая проблема собственных значений», см. классическую книгу [45].

речь идёт о матрицах конечных размеров, конечномерной ситуации и т. п. в отличие, скажем, от аналогичной задачи нахождения собственных значений операторов в бесконечномерных пространствах функций. Слово «проблема» тоже уместно в этом контексте, поскольку задача сложна и имеет много различных вариантов и частных случаев.

Различают *полную проблему* собственных значений и *частичную проблему* собственных значений. В полной проблеме требуется нахождение всех собственных чисел и собственных векторов. Частичная проблема собственных значений — это задача нахождения некоторых собственных чисел матрицы и/или некоторых собственных векторов. К примеру, наибольшего по модулю собственного значения, или нескольких наибольших по модулю собственных значений и соответствующих им собственных векторов.

Ясно, что собственные векторы матрицы определяются неоднозначно, с точностью до скалярного множителя. В связи с этим часто говорят о нахождении одномерных *инвариантных подпространств* матрицы. Инвариантные подпространства могут иметь и большую размерность, и в любом случае их знание доставляет важную информацию о задаваемом матрицей линейном операторе, позволяя упростить его представление. Пусть, например,  $\mathcal{S}$  — это  $l$ -мерное линейное подпространство в  $\mathbb{R}^n$ , которое является инвариантным для матрицы  $A$ , так что  $Ax \in \mathcal{S}$  для любого  $x \in \mathcal{S}$ . Пусть базис в  $\mathcal{S}$  образуют векторы  $v_1, v_2, \dots, v_l$ . Взяв базис всего пространства  $\mathbb{R}^n$  так, чтобы его первыми векторами были  $v_1, v_2, \dots, v_l$  (это, очевидно, можно сделать всегда), получим в нём блочно-треугольное представление рассматриваемого линейного оператора:

$$\begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix}$$

с  $l \times l$ -блоком  $A_{11}$ . Если  $\mathcal{S}'$  — дополнительное к  $\mathcal{S}$  подпространство, т. е.  $\mathbb{R}^n = \mathcal{S} \oplus \mathcal{S}'$ , и оно также инвариантно для матрицы  $A$ , то нетрудно понять, что блочно-треугольное представление матрицы превратится в блочно-диагональное:

$$\begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix}$$

В последние десятилетия задача определения для матрицы тех или иных инвариантных подпространств, не обязательно одномерных, также включается в «проблему собственных значений».

Помимо необходимости выхода в общем случае в комплексную плоскость  $\mathbb{C}$  ещё одной особенностью проблемы собственных значений, которая осложняет её решение, является нелинейный характер задачи, несмотря на традицию отнесения её к «вычислительной линейной алгебре». Это обстоятельство нетрудно осознать из рассмотрения основного соотношения (3.210)

$$Av = \lambda v,$$

которое является системой уравнений относительно  $\lambda$  и  $v$ , причём в его правой части суммарная степень неизвестных переменных равна двум:  $2 = (1 \text{ при } \lambda) + (1 \text{ при } v)$ . Но нелинейность уравнений — это возможная неустойчивость их решений (см. § 4.2б), которая приводит также к сложностям при их численным нахождении.

В заключение нашего обсуждения коснёмся алгоритмического аспекта проблемы собственных значений. Нахождение собственных значений матрицы сводится к решению алгебраического характеристического уравнения. С другой стороны, любой алгебраический полином с единичным старшим коэффициентом, имеющий вещественные или комплексные коэффициенты, является характеристическим полиномом для некоторой матрицы, в общем случае комплексной. Если, к примеру,

$$P(x) = x^n + p_{n-1}x^{n-1} + \dots + p_1x + p_0,$$

и этот полином имеет нули  $x_1, x_2, \dots, x_n$ , перечисленные с учётом кратности, то  $P(x) = (x - x_1)(x - x_2)\dots(x - x_n)$ . У матрицы  $A_P = (-1)^n \operatorname{diag}\{x_1, x_2, \dots, x_n\}$  и у всех подобных к ней матриц характеристическим полиномом является  $P(x)$ . Вместо диагональной матрицы можно взять какую-нибудь жорданову каноническую форму вида (3.13), группируя в жордановы клетки размера  $2 \times 2$  и более по несколько одинаковых собственных значений. Опять таки, преобразованием подобия из жордановой формы легко получить плотно заполненную матрицу.

Ещё одна популярная конструкция, с помощью которой можно построить по алгебраическому полиному матрицу, собственные значения которой совпадают с нулями полинома, — это так называемая сопровождающая матрица. Детали читатель может увидеть, например, в [9, 28, 43, 54].

Напомним теперь известную в алгебре теорему Абеля–Руффини: для алгебраических уравнений степени 5 и выше не существует конечных формул, выражающих их решения через коэффициенты уравнения.

ний с помощью четырёх арифметических действий и операции взятия корня произвольной степени [64, 75]. Как следствие, мы не должны ожидать существования прямых методов решения проблемы собственных значений для произвольных матриц размера  $5 \times 5$  и более, и потому подавляющее большинство методов решения проблемы собственных значений — существенно итерационные.

### 3.176 Матрицы простой структуры

Кратность собственного значения как решения характеристического уравнения матрицы называется *алгебраической кратностью* собственного значения. Часто её называют просто кратностью. Но одному собственному значению могут соответствовать несколько собственных векторов. Максимальное число линейно независимых собственных векторов, относящихся к собственному значению, называется *геометрической кратностью* собственного значения. Известно, что геометрическая кратность любого собственного значения не превосходит его алгебраической кратности [7, 54]. Если алгебраическая кратность собственного значения не равна его геометрической кратности (строго меньше), то это собственное значение называется *дефектным*.

*Простым собственным значением* матрицы называют её собственное значение кратности 1, т. е. простой нуль характеристического полинома этой матрицы. Ему соответствует единственный собственный вектор матрицы. Для матриц, у которых все собственные значения просты, проблема собственных значений решается наиболее просто. Но столь же несложными являются несколько более общие матрицы, у которых собственные значения могут быть кратными, но их алгебраическая кратность равна геометрической. Жорданова форма таких матриц — диагональная.

Сложность решения проблемы собственных значений для матрицы существенно зависит от структуры её жордановой канонической формы. Оказывается, что вычисление собственных векторов матрицы неустойчиво, если в её жордановой канонической форме соответствующие собственные значения находятся в жордановых клетках размера более 1. Напротив, для матриц, у которых жорданова каноническая форма таких клеток не имеет, то есть матриц, приводимых с помощью преобразования подобия к диагональному виду, проблема собственных значений ставится и решается существенно проще. Для их терминологического выделения вводится

**Определение 3.17.1** Квадратные матрицы, подобные диагональным матрицам, называются матрицами простой структуры или диагонализуемыми матрицами.

Матрицы простой структуры называют также *недефектными*, тогда как *дефектные матрицы* — это матрицы, которые не являются матрицами простой структуры (диагонализуемыми). Иначе говоря, дефектными называются матрицы, у которых некоторые собственные значения дефектны. В канонической жордановой форме таких матриц присутствуют нетривиальные клетки, имеющие размер 2 или более.

Если  $A$  — матрица простой структуры, то для некоторой неособенной матрицы  $V$  и диагональной матрицы  $D$

$$V^{-1}AV = D.$$

Тогда  $AV = VD$  и столбцы матрицы  $V$  являются собственными векторами для  $A$ . Они линейно независимы в силу неособенности  $V$  и всего их  $n$  штук. Из этого замечания вытекает ещё одно равносильное определение матриц простой структуры.

**Определение 3.17.2** Квадратную матрицу будем называть матрицей простой структуры или недефектной матрицей, если она обладает полным линейно независимым набором собственных векторов.

Так как всякому собственному значению матрицы соответствует хотя бы один собственный вектор, причём собственные векторы, отвечающие различным собственным значениям, линейно независимы, то матрица имеет простую структуру, если все её собственные значения различны. Обратное неверно, и матрица простой структуры может иметь совпадающие собственные числа (такова, например, единичная матрица).

Другой важный пример матриц простой структуры — это *нормальные матрицы*, которые перестановочны со своей эрмитово сопряжённой, т. е. такие матрицы  $A$ , что  $AA^* = A^*A$ . Можно показать [7, 9, 43, 54], что нормальные матрицы приводятся к диагональному виду унитарными (ортогональными в вещественном случае) преобразованиями подобия. Нормальными матрицами являются, в частности, симметричные и эрмитовы матрицы, кососимметричные и косоэрмитовы, ортогональные и унитарные, и все они — матрицы простой структуры.

Покажем, что матрицы простой структуры являются «типичными» и потому составляют в определённом смысле «большинство» во множестве всех квадратных матриц.

**Предложение 3.17.1** Любая квадратная матрица сколь угодно малым возмущением её элементов может быть сделана матрицей простой структуры.

**Доказательство.** Пусть  $A$  — рассматриваемая матрица. Воспользуемся теоремой Шура о возможности приведения произвольной матрицы к верхней треугольной с помощью ортогонального преобразования подобия. Тогда

$$Q^\top A Q = T,$$

где  $Q$  — некоторая ортогональная матрица,  $T$  — верхняя треугольная матрица, у которой по диагонали стоят собственные значения  $\lambda_1, \lambda_2, \dots, \lambda_n$  матрицы  $A$ .

Если для достаточно малого  $\varepsilon$  к  $n \times n$ -матрице  $T$  прибавить возмущающую диагональную матрицу тех же размеров

$$E = \begin{pmatrix} \varepsilon & & & 0 & \\ & \varepsilon/2 & & & \\ & & \varepsilon/3 & & \\ & & & \ddots & \\ 0 & & & & \varepsilon/n \end{pmatrix},$$

то треугольная матрица  $T + E$  будет иметь различные собственные числа.

Рассмотрим теперь возмущение исходной матрицы  $A$ , которое получается из  $E$  путём преобразования подобия, обратного по отношению к тому, что переводит  $A$  к треугольной форме, т.е.  $QEQ^\top$ . Тогда матрица

$$A + QEQ^\top$$

ортогонально подобна матрице  $T + E$ , причём все её собственные значения различны и она имеет простую структуру. Кроме того, любая норма возмущающей матрицы оценивается как

$$\|QEQ^\top\| \leq \|Q\| \|E\| \|Q^\top\| = \text{cond}(Q) \|E\|,$$

и она может быть сделана сколь угодно малой подходящим выбором  $\varepsilon$ . Это очевидно для нормы  $\|\cdot\|_{\max}$  (см. § 3.3д), а для других матричных норм следует из факта их эквивалентности.

Для случая комплексной матрицы  $A$  доказательство адаптируется очевидным образом. ■

**Следствие.** Матрицы простой структуры образуют открытое всюду плотное подмножество во множестве всех квадратных матриц.

Напомним, что множество  $E$  называется *всюду плотным* в  $X$ , если каждая точка множества  $X$  является предельной точкой множества  $E$  или же принадлежит множеству  $E$  (или то и другое вместе). Всюду плотное подмножество — это очень «представительная» часть множества, которая, образно выражаясь, «проникает всюду» и точками которой можно сколь угодно точно приблизить любую точку исходного множества.

Как следствие, недиагонализуемые, т. е. дефектные матрицы, в канонической жордановой форме которых присутствуют клетки размера 2 и более, составляют дополнение до всюду плотного множества. Это весьма разреженное множество среди всех квадратных матриц. Более точно, такие множества называют *множествами первой бэрковской категории*. Другой равносильный термин — *тощие множества*, так как в топологическом смысле они являются наиболее бедными множествами [19, 52].

Ещё одно соображение, показывающее типичность матриц простой структуры, может быть основано на понятии меры множества. Напомним, что мерой множества называют неотрицательную величину, обобщающую понятия длины, площади, объёма. Существует несколько примерно равносильных конструкций меры, применяемых в математике [88], и можно оценить меру множества дефектных матриц среди всего множества матриц. Дефектные матрицы образуют подмножество во множестве матриц, имеющих как минимум одно кратное собственное значение. В то же время множество жордановых форм  $n \times n$ -матриц с кратными собственными значениями определяется числом параметров  $< n$  во множестве всех жордановых форм (включающих также диагональные матрицы), и потому оно имеет меру нуль. То же самое верно в отношении дефектных матриц, что свидетельствует об их «исключительности» и «нетипичности».

Тем не менее на долю дефектных матриц приходятся главные трудности, с которыми сталкиваются при решении проблемы собственных значений. В этом отношении задача нахождения сингулярных чисел и сингулярных векторов является принципиально другой, так как сим-

метрическая матрица  $A^\top A$  (эрмитова матрица  $A^*A$  в комплексном случае) всегда имеет простую структуру, т. е. всегда диагонализуема.

### 3.17в Обусловленность проблемы собственных значений

Под обусловленностью проблемы собственных значений понимается степень чувствительности собственных значений и собственных векторов матрицы по отношению к возмущениям элементов матрицы. Их поведение оказывается существенно различным, так что обусловленность соответствующих подзадач общей проблемы собственных значений нужно рассматривать отдельно друг от друга. Спектр матрицы, как множество точек комплексной плоскости  $\mathbb{C}$ , непрерывно зависит от элементов матрицы, тогда как собственные векторы могут меняться скачкообразно.

**Теорема 3.17.1** (теорема Островского)

Пусть  $A = (a_{ij})$  и  $B = (b_{ij})$  — квадратные  $n \times n$ -матрицы, и пусть также

$$M = \max\{|a_{ij}|, |b_{ij}|\}, \quad \delta = \frac{1}{nM} \sum_{i,j} |a_{ij} - b_{ij}|. \quad (3.214)$$

Тогда любому собственному значению  $\lambda(B)$  матрицы  $B$  можно сопоставить такое собственное значение  $\lambda(A)$  матрицы  $A$ , что выполнено неравенство

$$|\lambda(A) - \lambda(B)| \leq (n+2)M\delta^{1/n}.$$

Можно так перенумеровать собственные числа матриц  $A$  и  $B$ , что для любого номера  $\nu = 1, 2, \dots, n$  имеет место

$$|\lambda_\nu(A) - \lambda_\nu(B)| \leq 2(n+1)^2 M\delta^{1/n},$$

где  $\lambda_\nu$  —  $\nu$ -е собственное значение.

Читатель может увидеть детали доказательства в книге [36], а соответствующая теория излагается также в [16, 20, 28, 36, 44, 54].

Но собственные векторы матрицы при изменении её элементов, вообще говоря, не изменяются непрерывно, а могут претерпевать скачки даже в совершенно обычных ситуациях.

**Пример 3.17.2** [54] Рассмотрим матрицу

$$A = \begin{pmatrix} 1 + \alpha & \beta \\ 0 & 1 \end{pmatrix}.$$

Её собственные значения суть числа  $1$  и  $1 + \alpha$ , и при  $\alpha\beta \neq 0$  соответствующими нормированными собственными векторами являются

$$\frac{1}{\sqrt{\alpha^2 + \beta^2}} \begin{pmatrix} -\beta \\ \alpha \end{pmatrix} \quad \text{и} \quad \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Выбирая подходящим образом отношение  $\alpha/\beta$ , можно придать первому собственному вектору любое направление, сколь бы малыми не являлись значения  $\alpha$  и  $\beta$ .

Если положить  $\alpha = 0$ , то

$$A = \begin{pmatrix} 1 & \beta \\ 0 & 1 \end{pmatrix}.$$

При  $\beta \neq 0$  у матрицы  $A$  будет всего один собственный вектор, хотя при надлежащем  $\beta$  её можно сделать сколь угодно близкой к единичной матрице, имеющей два линейно независимых собственных вектора. ■

Теорема Островского даёт явную оценку для изменения собственных чисел в зависимости от изменения элементов матрицы, мерой которого выступает величина  $\delta$  из (3.214). Но в оценках теоремы эта  $\delta$  присутствует в степени  $1/n$ , т. е. меньшей единицы, вследствие чего правая часть оценок допускает неограниченную скорость изменения собственных значений в окрестности нулевого  $\delta$ . Простые примеры показывают, что эта возможность в самом деле реализуется.

**Пример 3.17.3** Рассмотрим матрицу

$$A = \begin{pmatrix} \alpha & 1 \\ \beta & \alpha \end{pmatrix}$$

— жорданову  $2 \times 2$ -клетку с собственным значением  $\alpha$ , возмущённую элементом  $\beta$ . Собственные значения этой матрицы суть  $\alpha \pm \sqrt{\beta}$ , так что мгновенная скорость их изменения в зависимости от  $\beta$  равна  $\pm \frac{1}{2}\beta^{-1/2}$  и она бесконечна при  $\beta = 0$ . Это же явление имеет место и для произвольной жордановой клетки размера более двух. ■

Итак, несмотря на непрерывную зависимость собственных значений от элементов матрицы, скорость их изменения может быть сколь угодно большой (даже для матриц фиксированного размера). Это происходит в случае, если в канонической жордановой форме матрицы эти собственные значения находятся в жордановых клетках размера 2 и более, т. е. соответствуют так называемым нелинейным элементарным делителям матрицы. В целом, наличие нетривиальных жордановых клеток в канонической жордановой форме матрицы является источником основных проблем при нахождении собственных значений и собственных векторов. Собственные числа матриц простой структуры зависят от возмущений гораздо более «плавным образом», чем в общем случае.

**Теорема 3.17.2** (теорема Бауэра–Файка [108]) *Пусть  $A$  — квадратная  $n \times n$ -матрица простой структуры,  $\lambda_i$ ,  $i = 1, 2, \dots, n$ , — её собственные числа,  $V$  — матрица, составленная из собственных векторов  $A$  как из столбцов,  $B$  — произвольная  $n \times n$ -матрица. Тогда для всякого собственного значения  $\tilde{\lambda}$  возмущённой матрицы  $A + B$ , справедливо*

$$|\tilde{\lambda} - \lambda_i| \leq \text{cond}_2(V) \|B\|_2. \quad (3.215)$$

**Доказательство.** Если  $\tilde{\lambda}$  совпадает с каким-то из собственных значений исходной матрицы  $A$ , то левая часть доказываемого неравенства зануляется и оно очевидно справедливо. Будем поэтому предполагать, что  $\tilde{\lambda}$  не совпадает ни с одним из  $\lambda_i$ ,  $i = 1, 2, \dots, n$ . Если  $A$  имеет простую структуру согласно условию теоремы, то

$$V^{-1}AV = D,$$

где  $D = \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\}$  — диагональная матрица с собственными числами матрицы  $A$  по диагонали. По этой причине матрица  $D - \tilde{\lambda}I$  неособенна.

С другой стороны, матрица  $A + B - \tilde{\lambda}I$  является особенной по построению, так что особенна и матрица  $V^{-1}(A + B - \tilde{\lambda}I)V$ . Но

$$\begin{aligned} V^{-1}(A + B - \tilde{\lambda}I)V &= (D - \tilde{\lambda}I) + V^{-1}BV = \\ &= (D - \tilde{\lambda}I)(I + (D - \tilde{\lambda}I)^{-1}V^{-1}BV), \end{aligned}$$

и потому матрица  $(I + (D - \tilde{\lambda}I)^{-1}V^{-1}BV)$  также должна быть особенной. Следовательно, матрица

$$(D - \tilde{\lambda}I)^{-1}V^{-1}BV$$

имеет собственное значение  $-1$ , и потому из соотношения между спектральным радиусом и нормой матрицы (теорема 3.3.1) можем заключить, что любая норма этой матрицы должна быть не меньше 1.

В частности, сделанный вывод справедлив для спектральной нормы:

$$\|(D - \tilde{\lambda}I)^{-1}V^{-1}BV\|_2 \geq 1.$$

Поэтому в силу субмультипликативности

$$\|(D - \tilde{\lambda}I)^{-1}\|_2 \|V^{-1}\|_2 \|B\|_2 \|V\|_2 \geq 1.$$

Но  $(D - \tilde{\lambda}I)^{-1}$  — диагональная матрица, её спектральная норма равна наибольшему из модулей чисел по диагонали, и поэтому получаем

$$\max_{1 \leq i \leq n} |(\lambda_i - \tilde{\lambda})^{-1}| \cdot \|V^{-1}\|_2 \|B\|_2 \|V\|_2 \geq 1.$$

Последнее неравенство равносильно

$$\min_{1 \leq i \leq n} |\lambda_i - \tilde{\lambda}| \leq \|V^{-1}\|_2 \|B\|_2 \|V\|_2,$$

или

$$\min_{1 \leq i \leq n} |\tilde{\lambda} - \lambda_i| \leq \text{cond}_2(V) \|B\|_2,$$

как и требовалось. ■

Теорема Бауэра–Файка показывает, грубо говоря, что возмущение собственных значений матрицы простой структуры не превосходит величины спектральной нормы возмущения, умноженной на число обусловленности матрицы собственных векторов. Таким образом, скорость изменения собственных значений матриц простой структуры всегда конечна, а число обусловленности матрицы из собственных векторов может служить мерой обусловленности проблемы собственных значений.

То, что сделанный вывод следует применять с осторожностью и оговорками, демонстрирует следующий

**Пример 3.17.4** вещественная  $20 \times 20$ -матрица

$$\begin{pmatrix} 20 & 20 & & 0 \\ & 19 & 20 & \\ & 18 & 20 & \\ & \ddots & \ddots & \\ \varepsilon & & 2 & 20 \\ & & & 1 \end{pmatrix}$$

называется *матрицей Уилкинсона* [45]. В ней ненулевыми являются главная диагональ и наддиагональ, а также элемент  $\varepsilon$ , стоящий первым в последней строке. При  $\varepsilon = 0$  эта матрица имеет, очевидно, различные собственные значения  $1, 2, \dots, 18, 19, 20$ . Но в общем случае характеристическое уравнение матрицы Уилкинсона —

$$(20 - \lambda)(19 - \lambda) \dots (1 - \lambda) - 20^{19}\varepsilon = 0,$$

и его свободный член, который равен  $20! - 20^{19}\varepsilon$ , зануляется при  $\varepsilon = 20^{-19} \cdot 20! \approx 4.64 \cdot 10^{-7}$ . Матрица будет иметь при этом нулевое собственное значение, т. е. сделается особенной.

Величина возмущения, изменившего наименьшее собственное значение с 1 до 0, сопоставима по порядку с расстоянием между машинно-представимыми числами одинарной точности в районе единицы согласно стандартам IEEE 754/854. Ясно, что увеличением размера матрицы можно сделать эту величину критического возмущения сколь угодно малой.<sup>36</sup>

Итак, несмотря на то, что все собственные числа матрицы различны и, следовательно, являются гладкими функциями от элементов матрицы, скорость их изменения настолько велика, что практически мы как будто имеем дело с разрывными функциями. ■

Практическую ценность теоремы Бауэра–Файка в целом и неравенства (3.215) в частности снижает то обстоятельство, что собственные векторы матрицы определены с точностью до скалярного множителя, и потому  $\text{cond}_2(V)$  есть величина, заданная не вполне однозначно.

---

<sup>36</sup>Сам Дж. Уилкинсон решил, очевидно, остановиться на размере  $20 \times 20$ , так как такого размера вполне хватало для его иллюстративных целей в 50-60-е годы XX века, когда писалась книга [45] и длина мантиссы чисел в ЭВМ была небольшой.

Наилучшим выбором для  $\text{cond}_2(V)$  в неравенстве был бы, очевидно, минимум чисел обусловленности матриц из собственных векторов, но его нахождение является в общем случае сложной задачей. Тем не менее, прикидочные оценки и качественные выводы на основе теоремы Бауэра–Файка делать можно.

Важнейший частный случай применения теоремы Бауэра–Файка относится к симметричным (или эрмитовым) матрицам. Они имеют простую структуру и, кроме того, собственные векторы симметричных матриц ортогональны друг другу. Как следствие, матрица собственных векторов  $V$  может быть взята ортогональной, с числом обусловленности 1. Получаем следующий результат: если  $\lambda_i(A)$  — собственные числа симметричной матрицы  $A$ , а  $\tilde{\lambda}$  — собственное число возмущённой матрицы  $A + \Delta A$ , то

$$\min_i |\tilde{\lambda} - \lambda_i(A)| \leq \|\Delta A\|_2.$$

Иными словами, при возмущении симметричных матриц их собственные числа изменяются на величину, не превосходящую спектральной нормы возмущения, т. е. с гораздо меньшей конечной скоростью, нежели для матриц общего вида. Это же самое верно для эрмитовых матриц.

Сделанные выводы подкрепляются другими известными результатами теории матриц.

**Теорема 3.17.3** (теорема Вейля) *Пусть  $A$  и  $B$  — эрмитовы  $n \times n$ -матрицы, причём  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$  — собственные значения матрицы  $A$  и  $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \dots \geq \tilde{\lambda}_n$  — собственные значения матрицы  $\tilde{A} = A + B$ . Тогда  $|\tilde{\lambda}_i - \lambda_i| \leq \|B\|_2$  для каждого  $i = 1, 2, \dots, n$ .*

**Доказательство** можно найти в [43, 54].

**Теорема 3.17.4** (теорема Виландта–Хофмана) *Пусть  $A$  и  $B$  — эрмитовы  $n \times n$ -матрицы, причём  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$  — собственные значения матрицы  $A$  и  $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \dots \geq \tilde{\lambda}_n$  — собственные значения матрицы  $\tilde{A} = A + B$ . Тогда*

$$\left( \sum_{i=1}^n (\tilde{\lambda}_i - \lambda_i)^2 \right)^{1/2} \leq \|B\|_F,$$

где  $\|\cdot\|_F$  — фробениусова норма матрицы.

**Доказательство** можно найти в [44, 45]

Специфика теорем Вейля и Виландта–Хоффмана состоит в том, что в них, в отличие от теоремы Бауэра–Файка, рассматриваются возмущения эрмитовых (симметричных) матриц, которые тоже эрмитовы (симметричны), т. е. не выводят за пределы множества всех эрмитовых (симметричных) матриц. В этой ситуации получаются и более точные оценки возмущений собственных значений. Они также показывают, что собственные числа эрмитовых и симметричных матриц непрерывно зависят от элементов матрицы, и, самое главное, зависимость эта имеет довольно плавный характер.

### 3.17г Коэффициенты перекоса матрицы

Целью этого раздела является детальное исследование устойчивости решения проблемы собственных значений в упрощённой ситуации, когда у матрицы все собственные значения различны. В этом случае матрица имеет простую структуру (диагонализуема), и потому, как отмечалось в § 3.17в, скорость изменения собственных значений в зависимости от возмущений элементов матрицы конечна. Более точно, из теоремы Бауэра–Файка (теорема 3.17.2) следует, что собственные значения непрерывны по Лишицу в зависимости от элементов матрицы.

Пусть  $A = n \times n$ -матрица с различными собственными значениями и  $\Delta A$  — её возмущение, так что  $A + \Delta A$  — это близкая к  $A$  возмущённая матрица. Как изменятся собственные значения и собственные векторы матрицы  $A + \Delta A$  в сравнении с собственными значениями и собственными векторами  $A$ ?

Обозначим через  $\lambda_i$  собственные значения  $A$ ,  $x^{(i)}$  — соответствующие им собственные векторы,  $i = 1, 2, \dots, n$ . Эти векторы образуют базис в  $\mathbb{R}^n$ , коль скоро по предположению  $A$  является матрицей простой структуры. Имеем

$$\begin{aligned} Ax^{(i)} &= \lambda^{(i)}x^{(i)}, \\ (A + \Delta A)(x^{(i)} + \Delta x^{(i)}) &= (\lambda_i + \Delta\lambda_i)(x^{(i)} + \Delta x^{(i)}), \end{aligned}$$

где  $\Delta\lambda_i$  и  $\Delta x^{(i)}$  — изменения  $i$ -го собственного значения и  $i$ -го собственного вектора матрицы. Вычитая из второго равенства первое получим

$$(\Delta A)x^{(i)} + A\Delta x^{(i)} + \Delta A\Delta x^{(i)} = \lambda_i\Delta x^{(i)} + \Delta\lambda^{(i)}x^{(i)} + \Delta\lambda^{(i)}\Delta x^{(i)}.$$

Если пренебречь членами второго порядка малости, т. е.  $\Delta\lambda^{(i)}\Delta x^{(i)}$  и  $\Delta A \Delta x^{(i)}$ , то приходим к приближённому соотношению

$$\langle (\Delta A) x^{(i)} + A(\Delta x^{(i)}) \rangle = \lambda_i(\Delta x^{(i)}) + (\Delta\lambda_i) x^{(i)}, \quad (3.216)$$

которое отражает поведение возмущений «в главном».

Пусть  $y^{(1)}, y^{(2)}, \dots, y^{(n)}$  — собственные векторы эрмитово сопряжённой матрицы  $A^*$ , соответствующие её собственным значениям  $\bar{\lambda}_1, \bar{\lambda}_2, \dots, \bar{\lambda}_n$ . Умножая скалярно равенство (3.216) на  $y^{(j)}$ , получим

$$\begin{aligned} \langle (\Delta A) x^{(i)}, y^{(j)} \rangle + \langle A(\Delta x^{(i)}), y^{(j)} \rangle &= \\ &= \lambda_i \langle \Delta x^{(i)}, y^{(j)} \rangle + (\Delta\lambda_i) \langle x^{(i)}, y^{(j)} \rangle. \end{aligned} \quad (3.217)$$

В частности, при  $j = i$  имеем

$$\langle (\Delta A) x^{(i)}, y^{(i)} \rangle + \langle A(\Delta x^{(i)}), y^{(i)} \rangle = \lambda_i \langle \Delta x^{(i)}, y^{(i)} \rangle + (\Delta\lambda_i) \langle x^{(i)}, y^{(i)} \rangle,$$

где соседние со знаком равенства члены можно взаимно уничтожить. Они оказываются одинаковыми, коль скоро

$$\langle A(\Delta x^{(i)}), y^{(i)} \rangle = \langle \Delta x^{(i)}, A^* y^{(i)} \rangle = \langle \Delta x^{(i)}, \bar{\lambda}_i y^{(i)} \rangle = \lambda_i \langle \Delta x^{(i)}, y^{(i)} \rangle.$$

Следовательно,

$$\langle (\Delta A) x^{(i)}, y^{(i)} \rangle = (\Delta\lambda_i) \langle x^{(i)}, y^{(i)} \rangle,$$

и потому

$$\Delta\lambda_i = \frac{\langle (\Delta A) x^{(i)}, y^{(i)} \rangle}{\langle x^{(i)}, y^{(i)} \rangle}.$$

Теперь можно дать оценку возмущений собственных значений. Из полученной формулы для приращения  $\Delta\lambda_i$  и из неравенства Коши–Буняковского следует

$$|\Delta\lambda_i| \leq \frac{\|\Delta A\|_2 \|x^{(i)}\|_2 \|y^{(i)}\|_2}{\langle x^{(i)}, y^{(i)} \rangle} = \nu_i \|\Delta A\|_2,$$

где обозначено

$$\nu_i := \frac{\|x^{(i)}\|_2 \|y^{(i)}\|_2}{\langle x^{(i)}, y^{(i)} \rangle}, \quad i = 1, 2, \dots, n.$$

Величины  $\nu_i$  называются *коэффициентами перекоса* матрицы  $A$ , отвечающими собственным значениям  $\lambda_i$ ,  $i = 1, 2, \dots, n$ .

Ясно, что  $\nu_i \geq 1$  и можно интерпретировать коэффициенты перекоса как

$$\nu_i = \frac{1}{\cos \varphi_i},$$

где  $\varphi_i$  угол между собственными векторами  $x_i$  и  $y_i$  исходной и эрмитово сопряжённой матриц. Коэффициенты перекоса характеризуют, таким образом, обусловленность проблемы собственных значений (в смысле второго подхода из описанных в § 1.7).

Для симметричной (или, более общо, эрмитовой) матрицы коэффициенты перекоса равны 1. В самом деле, сопряжённая к ней задача на собственные значения совпадает с ней самой, и потому в наших обозначениях  $x^{(i)} = y^{(i)}$ ,  $i = 1, 2, \dots, n$ . Следовательно,  $\langle x^{(i)}, y^{(i)} \rangle = \langle x^{(i)}, x^{(i)} \rangle = \|x^{(i)}\|_2 \|y^{(i)}\|_2$ , откуда и следует  $\nu_i = 1$ . Это наименьшее возможное значение коэффициентов перекоса, так что численное нахождение собственных значений симметричных (эрмитовых в комплексном случае) матриц является наиболее устойчивым.

Продолжим преобразования с равенством (3.217), но теперь уже для случая  $j \neq i$ . Тогда  $\langle x^{(i)}, y^{(j)} \rangle = 0$  в силу биортогональности систем векторов  $\{x^{(i)}\}$  и  $\{y^{(j)}\}$  (см. предложение 3.2.1), и потому

$$\langle A(\Delta x^{(i)}), y^{(j)} \rangle = \langle \Delta x^{(i)}, A^* y^{(j)} \rangle = \langle \Delta x^{(i)}, \bar{\lambda}_j y^{(j)} \rangle = \lambda_j \langle \Delta x^{(i)}, y^{(j)} \rangle.$$

Подставляя этот результат в (3.217), будем иметь

$$\langle (\Delta A) x^{(i)}, y^{(j)} \rangle + \lambda_j \langle \Delta x^{(i)}, y^{(j)} \rangle = \lambda_i \langle \Delta x^{(i)}, y^{(j)} \rangle.$$

Следовательно,

$$\langle \Delta x^{(i)}, y^{(j)} \rangle = \frac{\langle (\Delta A) x^{(i)}, y^{(j)} \rangle}{\lambda_i - \lambda_j}.$$

Чтобы оценить возмущения  $\Delta x^{(i)}$  собственных векторов  $x^{(i)}$  матрицы  $A$  (напомним, они образуют базис в  $\mathbb{R}^n$ ), разложим по ним  $\Delta x^{(i)}$ :

$$\Delta x^{(i)} = \sum_{j=1}^n \alpha_{ij} x^{(j)}.$$

Так как собственные векторы матрицы задаются с точностью до множителя, то в этом разложении коэффициенты  $\alpha_{ii}$  содержательного

смысла не имеют, и можно даже положить  $\alpha_{ii} = 0$  (напомним, что в действительности ищется возмущение одномерного инвариантного подпространства матрицы). Для остальных коэффициентов имеем

$$\langle \Delta x^{(i)}, y^{(j)} \rangle = \alpha_{ij} \langle x^{(j)}, y^{(j)} \rangle$$

опять таки в силу предложения 3.2.1. Следовательно, для  $i \neq j$

$$\alpha_{ij} = \frac{\langle (\Delta A) x^{(i)}, y^{(j)} \rangle}{(\lambda_i - \lambda_j) \langle x^{(j)}, y^{(j)} \rangle}.$$

Коэффициенты  $\alpha_{ij}$  разложения возмущений собственных векторов оцениваются сверху как

$$|\alpha_{ij}| \leq \frac{\|(\Delta A) x^{(i)}\|_2 \|y^{(j)}\|_2}{|\lambda_i - \lambda_j| \cdot |\langle x^{(j)}, y^{(j)} \rangle|} \leq \frac{\|\Delta A\|_2}{|\lambda_i - \lambda_j|} \nu_j,$$

и потому имеет место неравенство

$$\|\Delta x^{(i)}\|_2 \leq \|\Delta A\|_2 \cdot \|x\|_2 \cdot \sum_{j \neq i} \frac{\nu_j}{|\lambda_i - \lambda_j|}. \quad (3.218)$$

Отметим значительную разницу в поведении возмущений собственных значений и собственных векторов матриц. Из оценки (3.218) следует, что на чувствительность отдельного собственного вектора влияют коэффициенты перекоса *всех* собственных значений матрицы, а не только того, которое отвечает этому вектору. Кроме того, в знаменателях слагаемых из правой части (3.218) присутствуют разности  $\lambda_i - \lambda_j$ , которые могут быть малыми при близких собственных значениях матрицы. Собственные векторы при этом очень чувствительны к возмущениям в элементах матрицы, что мы могли наблюдать в примере 3.17.2. В частности, даже для симметричных (эрмитовых) матриц задача отыскания собственных векторов может оказаться плохообусловленной.

### 3.17д Круги Гершгорина

Пусть  $A = (a_{ij})$  — квадратная матрица из  $\mathbb{R}^{n \times n}$  или  $\mathbb{C}^{n \times n}$ . Если  $\lambda \in \mathbb{C}$  — её собственное значение, то

$$Av = \lambda v \quad (3.219)$$

для некоторого собственного вектора  $v \in \mathbb{C}^n$ ,  $v \neq 0$ . Предположим, что в  $v$  наибольшее абсолютное значение имеет компонента с номером  $l$ , так что  $|v_l| = \max_{1 \leq j \leq n} |v_j|$ .

Рассмотрим  $l$ -ю компоненту векторного равенства (3.219):

$$\sum_{j=1}^n a_{lj} v_j = \lambda v_l.$$

Выписанное равенство равносильно

$$\sum_{\substack{j=1 \\ j \neq l}}^n a_{lj} v_j = (\lambda - a_{ll}) v_l,$$

откуда

$$\begin{aligned} |\lambda - a_{ll}| |v_l| &= \left| \sum_{j \neq l} a_{lj} v_j \right| \leq \sum_{j \neq l} |a_{lj} v_j| = \\ &= \sum_{j \neq l} |a_{lj}| |v_j| \leq |v_l| \sum_{j \neq l} |a_{lj}|, \end{aligned}$$

поскольку  $|v_j| \leq |v_l|$ . Наконец, так как  $v \neq 0$ , можем сократить обе части полученного неравенства на положительную величину  $|v_l|$ . Это даёт

$$|\lambda - a_{ll}| \leq \sum_{j \neq l} |a_{lj}|.$$

Не зная собственного вектора  $v$ , мы не располагаем и номером  $l$  его наибольшей по модулю компоненты. Но можно действовать наверняка, рассмотрев дизъюнкцию (объединение) соотношений выписанного выше вида для всех  $l = 1, 2, \dots, n$ , так как хотя бы для одного из них непременно справедливы наши рассуждения. Поэтому в целом, если  $\lambda$  — какое-либо собственное значение рассматриваемой матрицы  $A$ , то должно выполняться хотя бы одно из неравенств

$$|\lambda - a_{ll}| \leq \sum_{j \neq l} |a_{lj}|, \quad l = 1, 2, \dots, n.$$

Каждое из этих соотношений на  $\lambda$  определяет на комплексной плоскости  $\mathbb{C}$  круг с центром в точке  $a_{ll}$  и радиусом, равным  $\sum_{j \neq l} |a_{lj}|$ . Как следствие, мы приходим к результату, который был установлен в 1931 году С.А. Гершгориным:

**Теорема 3.17.5** (теорема Гершгорина) Для любой вещественной или комплексной  $n \times n$ -матрицы  $A = (a_{ij})$  все собственные значения  $\lambda(A)$  расположены в объединении кругов комплексной плоскости с центрами  $a_{ii}$  и радиусами  $\sum_{j \neq i} |a_{ij}|$ ,  $i = 1, 2, \dots, n$ , т. е.

$$\lambda(A) \in \bigcup_{i=1}^n \left\{ z \in \mathbb{C} \mid |z - a_{ii}| \leq \sum_{j \neq i} |a_{ij}| \right\}.$$

Фигурирующие в условиях теоремы круги комплексной плоскости

$$\left\{ z \in \mathbb{C} \mid |z - a_{ii}| \leq \sum_{j \neq i} |a_{ij}| \right\}, \quad i = 1, 2, \dots, n,$$

называются *кругами Гершгорина* матрицы  $A = (a_{ij})$ . Можно дополнительно показать, что если объединение кругов Гершгорина распадается на несколько связных, но непересекающихся частей, то каждая такая часть содержит столько собственных значений матрицы, сколько кругов её составляют [45, 54, 128].

Теорема Гершгорина равносильна признаку Адамара неособенности матриц (теорема 3.4.1), т. е. каждый из этих результатов может быть выведен из другого. В самом деле, если матрица имеет диагональное преобладание, то её круги Гершгорина не захватывают начала координат комплексной плоскости, а потому в условиях теоремы Гершгорина матрица должна быть неособенной. Обратно, пусть верен признак Адамара. Если  $\lambda$  — собственное значение матрицы  $A = (a_{ij})$ , то матрица  $(A - \lambda I)$  особенна и потому не может иметь диагональное преобладание. По этой причине хотя бы для одного  $i = 1, 2, \dots, n$  должно быть выполнено

$$|\lambda - a_{ii}| \leq \sum_{j \neq i} |a_{ij}|, \quad i = 1, 2, \dots, n.$$

Этими условиями и определяются круги Гершгорина.

**Пример 3.17.5** Для  $2 \times 2$ -матрицы (3.20)

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix},$$

рассмотренной в примере 3.2.4 (стр. 358), собственные значения суть  $\frac{1}{2}(5 \pm \sqrt{33})$ , они приблизительно равны  $-0.372$  и  $5.372$ . На рис. 3.33,

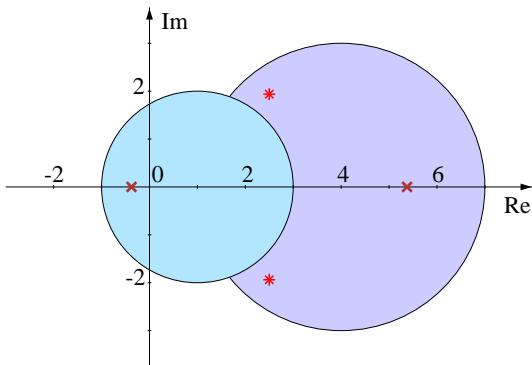


Рис. 3.33. Круги Гершгорина и спектры матриц (3.20) и (3.21)

показывающим соответствующие матрице круги Гершгорина, эти собственные значения выделены крестиками.

Матрица (3.21)

$$\begin{pmatrix} 1 & 2 \\ -3 & 4 \end{pmatrix},$$

которая отличается от матрицы (3.20) лишь противоположным знаком элемента на месте (2, 1), имеет те же самые круги Гершгорина. Но собственные значения у неё комплексные, равные  $\frac{1}{2}(5 \pm i\sqrt{15})$ , т. е. приблизительно  $2.5 \pm 1.936i$ . Они выделены на рис. 3.33 звёздочками, целиком находясь в одном из кругов Гершгорина. ■

Бросается в глаза «избыточность» кругов Гершгорина, которые в качестве области локализации собственных значений очерчивают довольно большую область комплексной плоскости. Это характерно для матриц с существенной внедиагональной частью. Но если недиагональные элементы матрицы малы сравнительно с диагональными, то информация, даваемая кругами Гершгорина, становится весьма точной.

### 3.17e Отношение Рэлея

Проблема собственных значений для  $n \times n$ -матрицы  $A$  требует решения системы уравнений (3.210),

$$Av = \lambda v,$$

относительно  $\lambda$  и  $v$ . Если собственный вектор  $v$  уже известен из каких-либо соображений или получен ранее, то собственное число  $\lambda$  можно найти, сравнив коллинеарные векторы  $Av$  и  $v$  и вычислив их отношение. Но такая идеальная ситуация реализуется редко, хотя бы даже потому, что собственный вектор матрицы почти всегда находится не абсолютно точно, а с некоторой погрешностью. Что делать, если известно какое-то приближение  $x$  к собственному вектору  $v$  и необходимо на основе этой информации вычислить приближение к соответствующему собственному значению?

Естественно попытаться взять отношение какой-нибудь нормы от  $Ax$  и  $x$ , т. е. приблизить собственное значение дробью  $\|Ax\|/\|x\|$ . Но существует также другое элегантное решение вопроса.

После подстановки  $x$  вместо  $v$  в соотношение (3.210) и изменения порядка членов получим

$$x \lambda = Ax, \quad (3.220)$$

и это фактически система  $n$  уравнений относительно одного неизвестного  $\lambda$ . Точного решения она наверняка не имеет, но наилучшим приближением к собственному значению матрицы можно взять псевдорешение этой системы, т. е. то значение  $\lambda$ , при котором левая и правая части имеют наименьшее отличие в какой-то норме. Псевдорешение удобно искать в 2-норме, порождённой скалярным произведением, привлекая теорию из главы 2.

В § 2.11 эта теория была развита для вещественных линейных пространств и вещественных линейных систем, но она легко распространяется и на комплексный случай, в котором мы находимся при решении проблемы собственных значений. В частности, геометрическое обоснование метода наименьших квадратов в § 2.11в не зависит от того, является линейное векторное пространство со скалярным произведением вещественным или комплексным. Нахождение псевдорешения всё равно сводится к организации системы нормальных уравнений (2.127) и т. д.

В нашей ситуации с системой линейных уравнений (3.220) соответствующая нормальная система превращается в одно уравнение относительно неизвестного  $\lambda$ ,

$$x^* x \lambda = x^* A x,$$

где эрмитово сопряжение применяется к  $x$  потому, что он, вообще го-

воля, комплексный. Решение этого уравнения равно

$$\tilde{\lambda} = \frac{x^*Ax}{x^*x},$$

и оно обеспечивает искомое приближение к собственному значению  $\lambda$ . Сказанное выше мотивирует

**Определение 3.17.3** Для квадратной  $n \times n$ -матрицы  $A$ , вещественной или комплексной, отношением Рэлея называется величина  $\mathcal{R}(x)$ , задаваемая как

$$\mathcal{R}(x) := \frac{\langle Ax, x \rangle}{\langle x, x \rangle} = \frac{x^*Ax}{x^*x},$$

которая определена на множестве ненулевых векторов  $x$  из арифметического пространства  $\mathbb{R}^n$  или  $\mathbb{C}^n$ .

Таким образом, отношение Рэлея — это наилучшее, в некотором определённом смысле, приближение к собственному значению матрицы для заданного фиксированного приближения к соответствующему собственному вектору.

Поскольку

$$\frac{x^*Ax}{x^*x} = \frac{x^*Ax}{\|x\|_2^2} = \left( \frac{x^*}{\|x\|_2} \right) A \left( \frac{x}{\|x\|_2} \right),$$

то отношение Рэлея на самом деле является не совсем отношением, а просто даёт значения квадратичной формы, порождаемой матрицей  $A$ , на единичной сфере 2-нормы. Это наблюдение мотивирует следующее понятие: множество

$$\{ \mathcal{R}(x) \mid x \neq 0 \},$$

т. е. область значений отношения Рэлея, называется *областью значений* квадратной матрицы  $A$ . Нередко его называют также *числовым образом* матрицы.

**Теорема 3.17.6** (теорема Тёплица–Хаусдорфа) *Область значений квадратной матрицы — выпуклое подмножество вещественной оси  $\mathbb{R}$  или комплексной плоскости  $\mathbb{C}$ .*

**Доказательство** можно увидеть в [44, 78].

Перечислим основные свойства отношения Рэлея. Для любого ненулевого скаляра  $\alpha$  справедливо

$$\mathcal{R}(\alpha x) = \mathcal{R}(x),$$

что устанавливается непосредственной проверкой.

Если  $v$  — собственный вектор матрицы  $A$ , то  $\mathcal{R}(v)$  равен собственному значению матрицы, отвечающему  $v$ . В самом деле, если обозначить это собственное значение посредством  $\lambda$ , то  $Av = \lambda v$ . По этой причине

$$\mathcal{R}(v) = \frac{\langle Av, v \rangle}{\langle v, v \rangle} = \frac{\langle \lambda v, v \rangle}{\langle v, v \rangle} = \frac{\lambda \langle v, v \rangle}{\langle v, v \rangle} = \lambda.$$

Как следствие, собственные числа матрицы принадлежат её области значений.

Собственные векторы являются стационарными точками отношения Рэлея, т. е. точками зануления его частных производных по отдельным компонентам вектора. Покажем это для вещественной симметричной матрицы, для которой отношение Рэлея рассматривается на ненулевых вещественных векторах:

$$\frac{\partial \mathcal{R}(x)}{\partial x_i} = \frac{\partial}{\partial x_i} \left( \frac{\langle Ax, x \rangle}{\langle x, x \rangle} \right) = \frac{2(Ax)_i \langle x, x \rangle - \langle Ax, x \rangle \cdot 2x_i}{\langle x, x \rangle^2}.$$

Если  $x = v = (v_1, v_2, \dots, v_n)^\top$  — собственный вектор матрицы  $A$ , то числитель последней дроби равен  $2\lambda v_i \langle v, v \rangle - \langle \lambda v, v \rangle \cdot 2v_i = 0$ .

Практическое значение отношения Рэлея для вычислительных методов состоит в том, что с его помощью можно легко получить приближение к собственному значению, если известен приближённый собственный вектор матрицы. Уточнение собственных значений играет большую роль при организации сдвигов матрицы в обратных степенных итерациях, QR-алгоритме и других. Подробности излагаются далее в § 3.18б и 3.19в, 3.18в, 3.18д.

Хотя отношение Рэлея имеет смысл и применяется для произвольных матриц, особую красоту и богатство содержания оно приобретает для эрмитовых (симметричных в вещественном случае) матриц. Если  $A$  — эрмитова  $n \times n$ -матрица, то, как известно,

$$A = UDU^*,$$

где  $D = \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\}$  — диагональная матрица с вещественными собственными значениями матрицы  $A$  по диагонали,  $U$  — некоторая

унитарная  $n \times n$ -матрица (ортогональная в вещественном случае). Тогда

$$\mathcal{R}(x) = \frac{\langle Ax, x \rangle}{\langle x, x \rangle} = \frac{\langle UDU^*x, x \rangle}{\langle x, x \rangle} = \frac{\langle DU^*x, U^*x \rangle}{\langle U^*x, U^*x \rangle} = \frac{\sum_{i=1}^n \lambda_i |y_i|^2}{\|y\|_2^2},$$

где  $(y_1, y_2, \dots, y_n)^\top = U^*x$ . Поскольку

$$\frac{1}{\|y\|_2^2} \sum_{i=1}^n |y_i|^2 = \sum_{i=1}^n \frac{|y_i|^2}{\|y\|_2^2} = 1,$$

отношение Рэлея оказывается равным выпуклой комбинации собственных значений  $\lambda_i$  с коэффициентами  $(|y_i|/\|y\|_2)^2$ . Но все  $\lambda_i$  вещественны, и потому область значений отношения Рэлея для эрмитовой матрицы — это интервал  $[\lambda_{\min}, \lambda_{\max}] \subset \mathbb{R}$ , от минимального собственного значения матрицы до максимального.

Сделанное наблюдение позволяет с помощью отношения Рэлея легко находить для симметричных (эрмитовых) матриц нетривиальные оценки их минимальных собственных значений сверху и максимальных собственных значений снизу. Они противоположны по смыслу тем оценкам, которые получаются с помощью теоремы Гершгорина, т. е. фактически дополняют их.

**Пример 3.17.6** Рассмотрим симметричную матрицу

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 4 \\ 3 & 4 & 5 \end{pmatrix}.$$

Её собственные значения равны  $-0.78765$ ,  $-0.54420$  и  $9.3319$ . Грубые внешние границы спектра можно найти с помощью кругов Гершгорина:

$$\lambda(A) \in [\min \{1 - 5, 2 - 6, 5 - 7\}, \max \{1 + 5, 2 + 6, 5 + 7\}] = [-4, 12].$$

Посмотрим, что получится с помощью отношения Рэлея. С этой целью случайно выберем какие-нибудь векторы  $x \in \mathbb{R}^n$  и найдём для них значение отношения Рэлея. Возьмём, к примеру,

$$x = (1, 2, 3)^\top,$$

получим

$$\mathcal{R}(x) = \frac{\langle Ax, x \rangle}{\langle x, x \rangle} = 9.1429.$$

А если взять

$$x = (1, 1, -1)^\top,$$

то получим

$$\mathcal{R}(x) = \frac{\langle Ax, x \rangle}{\langle x, x \rangle} = -0.6667.$$

Неплохие прикидочные оценки для минимального и максимального собственных чисел нашей матрицы! ■

В теории с помощью отношения Рэлея нетрудно вывести полезные оценки для собственных и сингулярных чисел матриц. В частности, из свойств отношения Рэлея следует теорема Вейля (теорема 3.17.3); см. подробности в [43, 54].

### 3.17ж Предварительное упрощение матрицы

Естественная идея состоит в том, чтобы при решении проблемы собственных значений привести матрицу к некоторой специальной форме, для которой собственные значения и/или собственные векторы могут быть найдены проще, чем для исходной. То же самое верно для задачи нахождения сингулярных чисел и сингулярных векторов матрицы.

В задаче на собственные значения идеальным было бы приведение матрицы к диагональной или треугольной форме, по которым собственные числа находятся непосредственно. Элементарными преобразованиями, с помощью которых выполняется это приведение, должны быть, очевидно, те, что сохраняют неизменным спектр матрицы. Это преобразования подобия матрицы, имеющие вид  $A \mapsto S^{-1}AS$ . Но они существенно сложнее действуют на матрицу, чем линейное комбинирование строк, которое использовалось в прямых методах решения систем линейных алгебраических уравнений. Преобразования подобия линейно комбинируют как строки, так и столбцы. По этой причине нельзя уже столь просто управлять обнулением тех или иных элементов матрицы, как в прямом ходе метода Гаусса, в методе Хаусхолдера или методе вращений. Невозможность полной реализации идеи упрощения матрицы следует также из теоремы Абеля–Руффини, которую мы обсуждали в § 3.17а. Если бы это упрощение было осуществимым,

то оно привело бы к конечному алгоритму решения алгебраических уравнений произвольной степени, что в общем случае невозможно.

Тем не менее идея предварительного упрощения матрицы для решения проблемы собственных значений и для нахождении сингулярного разложения является частично реализуемой и во многих случаях действительно способствует повышению эффективности численных алгоритмов. Её наиболее популярное воплощение для задачи на собственные значения — это так называемая почти треугольная (хессенбергова) форма для общих матриц, а также её частные случаи — трёхдиагональные симметричные и эрмитовы матрицы. При нахождении сингулярных чисел и сингулярных векторов упрощённой формой матриц являются двухдиагональные матрицы. Почти треугольная, трёхдиагональная и двухдиагональная матрицы содержат существенно меньше ненулевых элементов, чем плотно заполненная матрица, и, что самое важное, их специальная форма сохраняется при работе некоторых численных методов (например, в QR-алгоритме).

**Определение 3.17.4** Матрица  $H = (h_{ij})$  называется верхней почти треугольной или хессенберговой матрицей (в форме Хессенберга), если  $h_{ij} = 0$  при  $i > j + 1$ .

Наглядный «портрет» хессенберговой матрицы выглядит следующим образом:

$$H = \begin{pmatrix} \times & \times & \cdots & \times & \times \\ \times & \times & \cdots & \times & \times \\ & \times & \ddots & \vdots & \vdots \\ 0 & & \ddots & \times & \times \\ & & & \times & \times \end{pmatrix}.$$

Симметричная хессенбергова матрица — это, очевидно, трёхдиагональная матрица. Нередко используется также нижняя почти треугольная матрица (нижняя хессенбергова), и её определение совершенно аналогично.

**Предложение 3.17.2** Любая квадратная матрица с помощью ортогональных преобразований подобия может быть приведена к хессенберговой форме. Более точно, для любой квадратной матрицы  $A$  существует такая ортогональная матрица  $Q$ , которая является произведением конечного числа матриц отражения или матриц вращения, что  $H = QAQ^\top$  — хессенбергова матрица.

**Доказательство.** Рассмотрим сначала приведение с помощью матриц отражения Хаусхолдера (см. § 3.7д).

Возьмём матрицу отражения  $Q_1 = I - 2uu^\top$  такой, чтобы первая компонента её вектора Хаусхолдера  $u$  была нулевой и при этом

$$Q_1 \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \\ \vdots \\ a_{n1} \end{pmatrix} = \begin{pmatrix} a_{11} \\ a'_{21} \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

т. е. при умножении  $A$  на  $Q_1$  слева занулялись бы элементы  $a_{31}, \dots, a_{n1}$  в первом столбце  $A$ . Нетрудно видеть, что  $Q_1$  выглядит следующим образом

$$Q_1 = \left( \begin{array}{c|ccccc} 1 & 0 & \cdots & 0 & 0 \\ \hline 0 & \times & \cdots & \times & \times \\ 0 & \times & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \times & \times \\ 0 & \times & \cdots & \times & \times \end{array} \right).$$

Когда  $A$  умножается на такую матрицу  $Q_1$  слева, то в ней не изменяются элементы первой строки. Когда матрица  $Q_1A$  умножается на  $Q_1^\top = Q_1$  справа, то в ней не изменяются элементы первого столбца. Поэтому в матрице  $Q_1AQ_1^\top$ , как и в  $Q_1A$ , первый столбец имеет нули в позициях с 3-й по  $n$ -ю.

Далее выбираем матрицы отражения  $Q_2, Q_3, \dots, Q_{n-2}$  такими, чтобы умножение слева на  $Q_i$  давало нули в позициях с  $(i+2)$ -й по  $n$ -ю в  $i$ -ом столбце. Эти матрицы имеют вид

$$Q_i = \left( \begin{array}{c|c} I & 0 \\ \hline 0 & \tilde{Q}_i \end{array} \right),$$

где в верхнем левом углу стоит единичная матрица размера  $i \times i$ , а  $\tilde{Q}_i$  — матрица отражения размера  $(n-i) \times (n-i)$ . При этом последующее умножение справа на  $Q_i^\top = Q_i$  тоже не портит возникающую почти треугольную структуру результирующей матрицы. Получающаяся в итоге матрица  $QAQ^\top$  с  $Q = Q_{n-2} \dots Q_1$  действительно является верхней почти треугольной.

Для матриц вращения доказательство совершенно аналогично. Обнулим элементы  $a_{31}, a_{41}, \dots, a_{n1}$  с помощью умножений слева на специально подобранные матрицы вращений  $G(2, 3), G(2, 4), \dots, G(2, n)$  (см. детали в § 3.7г). Затем станем умножать справа на обратные (транспонированные) к этим матрицам. При умножении справа на  $G(2, 3)^\top$  будет вычислена линейная комбинация 2-го и 3-го столбцов, а первый останется неизменным, т. е. нулевым. То же самое произойдёт при умножении на  $G(2, 4)^\top$  справа и т. д. вплоть до  $G(2, n)^\top$ .

Далее обнулим во втором столбце элементы  $a_{42}, a_{52}, \dots, a_{n2}$  с помощью подходящих матриц вращения  $G(3, 4), G(3, 5), \dots, G(3, n)$  и затем умножим на транспонированные к ним справа. И снова нули в позициях  $(4, 2), (5, 2), \dots, (n, 2)$  останутся, поскольку умножения справа сводятся к линейному комбинированию третьего столбца с последующими. И так далее. В результате через  $n - 2$  шага описанного процесса получится нижняя почти треугольная матрица. ■

Ниже мы увидим, что хессенбергова форма матрицы существенно помогает реализации некоторых методов решения проблемы собственных значений, в частности, QR-алгоритма (см. § 3.18г). Кроме того, развитые численные методы решения проблемы собственных значений существуют для симметричных хессенберговых матриц, которые являются симметричными трёхдиагональными.

Законный вопрос по поводу доказанного результата заключается в том, зачем преобразования подобия выполняются с помощью ортогональных матриц? В принципе, для этой цели можно было бы применять и другие матрицы. Но применение именно ортогональных матриц оправдывается как минимум двумя причинами.

Во-первых, умножение на ортогональную матрицу — это изометрическое преобразование арифметического пространства  $\mathbb{R}^n$ , которое сохраняет расстояния между точками и углы между прямыми. Поэтому семейство собственных векторов матрицы после ортогонального преобразования подобия не ухудшает (сохраняет) меру своей линейной зависимости или независимости. Более точно, спектральное число обусловленности матрицы из собственных векторов не увеличивается. По теореме Бауэра–Файка (теорема 3.17.2) именно эта величина может быть взята как количественная характеристика обусловленности проблемы собственных значений для матриц простой структуры (см. § 3.17в).

Во-вторых, ортогональные преобразования подобия являются также преобразованиями конгруэнции и потому сохраняют симметрич-

ность матрицы, если она имеет место. Другие подобия могут нарушать свойство симметричности, что очень невыгодно.

В задаче нахождения сингулярного разложения матрицы элементарными преобразованиями, которые не изменяют сингулярные числа, являются односторонние — слева или справа — умножения на ортогональные матрицы (см. предложение 3.2.5).

**Предложение 3.17.3** *Любая матрица с помощью умножений слева и справа на ортогональные матрицы может быть приведена к двухдиагональной форме. Более точно, для любой матрицы  $A$  существуют такие ортогональные матрицы  $U$  и  $V$ , которые являются произведениями конечного числа матриц отражения или матриц вращения, что  $T = U A V$  — двухдиагональная матрица.*

**Доказательство.** Оно вполне аналогично доказательству предложения 3.17.2, и мы проведём его для матриц вращения (см. § 3.7г).

Умножением на матрицы вращения слева зануляем поддиагональные элементы первого столбца. Затем умножением на матрицы вращения справа обнуляем элементы  $a_{13}, \dots, a_{1n}$  в первой строке.

Почему зануляем элементы, начиная с  $a_{13}$ , а не с  $a_{12}$ ? Тогда умножение на матрицу вращения  $G(1, 2)$  справа привело бы к изменению целиком 1-го и 2-го столбцов, которые сделались бы линейными комбинациями своих предшествующих значений. По этой причине занулить элемент  $a_{12}$  с помощью умножения на матрицу вращения справа мы не можем без побочного эффекта — исчезновения ранее полученного нуля в позиции  $(2, 1)$ . Это следует из формул (3.110), и мы разбирали эту особенность вращений в § 3.7г. Именно поэтому обнуление элементов строки начинается не с элемента, непосредственно идущего за диагональным, а со следующего за ним, у которого индексная пара имеет вид  $(i, i + 2)$ .

На следующем шаге снова умножаем полученную матрицу на последовательность матриц вращения слева, чтобы обнулить поддиагональные элементы второго столбца. После этого умножениями на матрицы вращения справа обнуляем элементы второй строки, начиная с  $a_{24}$ . И так далее.

Процесс продолжается вплоть до предпоследнего столбца с номером  $n - 1$ , для которого обнулять элементы соответствующей строки уже не нужно. В результате получаем двухдиагональную матрицу с главной диагональю и одной наддиагональю, а также две последова-

тельности ортогональных матриц вращения, на которые умножали исходную матрицу слева и справа. Их произведения дают матрицы  $U$  и  $V$  соответственно.

Нетрудно видеть, что если обнулять сначала задиагональные элементы строки матрицы, а потом уже элементы столбца, то с тем же успехом можно получить двухдиагональную матрицу с ненулевой поддиагональю. ■

Произведение двухдиагональной матрицы на её транспонированную даёт симметричную трёхдиагональную матрицу, так что и в этом случае приходим к ситуации, аналогичной проблеме собственных значений. Это замечание актуально при нахождении сингулярных чисел и сингулярных векторов матриц (см. § 3.19д).

## 3.18 Численные методы несимметричной проблемы собственных значений

Существует очень большое количество разнообразных численных методов для решения общей несимметричной проблемы собственных значений. В нашем курсе рассматриваются лишь несколько основных и, пожалуй, наиболее популярных методов. Более подробную информацию о состоянии этой области вычислительной математики читатель может получить из более полных и специальных книг [11, 13, 45, 46, 48, 61, 72, 73, 79, 80] и др., а также из обзоров и журнальных статей.

Выше мы видели, что несимметричная проблема собственных значений может иметь плохую обусловленность, так что соответствующие задачи иногда являются вычислительно некорректными. По этой причине одним из направлений развития современных вычислительных методов линейной алгебры в настоящее время является переосмысление постановок несимметричной проблемы собственных значений. Вместо классической формулировки «нахождения значений» для собственных чисел и собственных векторов предлагается уточнять области их локализации на комплексной плоскости. Например, можно искать принадлежность собственных значений тем или иным интервалам комплексной плоскости [120], можно исследовать расположение собственного значения относительно какой-либо прямой, или же внутри заданного круга и т. п. [78].

### 3.18а Степенной метод

**Определение 3.18.1** Если у некоторой матрицы собственные значения  $\lambda_i$ ,  $i = 1, 2, \dots$ , удовлетворяют неравенствам  $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \dots$ , то  $\lambda_1$  называют доминирующим собственным значением, а соответствующий ему собственный вектор — доминирующим собственным вектором матрицы.

Степенной метод, описанию которого посвящён этот раздел книги, предназначен для решения частичной проблемы собственных значений — нахождения доминирующих собственного значения и собственного вектора матрицы. Его нередко используют для вычисления спектрального радиуса матрицы, который является не чем иным, как модулем её доминирующего собственного значения, когда оно может быть выделено в матрице.

Лежащая в основе степенного метода идея чрезвычайно проста и состоит следующем. Если у матрицы  $A$  имеется одно собственное значение  $\hat{\lambda}$ , превосходящее по модулю все остальные собственные значения, то при умножении этой матрицы на произвольный вектор  $x^{(0)}$  направление  $\hat{v}$ , отвечающее этому собственному значению  $\hat{\lambda}$  будет растягиваться сильнее остальных (при  $\hat{\lambda} > 1$ ) или сжиматься меньше остальных (при  $\hat{\lambda} \leq 1$ ). При повторном умножении  $A$  на результат  $Ax^{(0)}$  предшествующего умножения эта компонента ещё более удлиняется в сравнении с остальными. Повторив рассмотренную процедуру умножения достаточное количество раз, мы получим вектор, в котором полностью преобладает направление  $\hat{v}$ , т. е. практически будем иметь приближённый собственный вектор.

В качестве приближённого собственного значения матрицы  $A$  можно при этом взять «отношение» двух последовательных векторов, порождённых нашим процессом —  $x^{(k)} = A^k x^{(0)}$  и  $x^{(k-1)} = A^{k-1} x^{(0)}$ ,  $k = 1, 2, \dots$ . Слово «отношение» взято здесь в кавычки потому, что употреблено не вполне строго: ясно, что векторы  $x^{(k)}$  и  $x^{(k-1)}$  могут оказаться неколлинеарными, и тогда их «отношение» смысла иметь не будет. Возможны следующие пути решения этого вопроса:

- 1) рассматривать отношение каких-нибудь фиксированных компонент векторов  $x^{(k)}$  и  $x^{(k-1)}$ , т. е.

$$x_i^{(k)} / x_i^{(k-1)} \quad (3.221)$$

для некоторого  $i \in \{1, 2, \dots, n\}$ ;

- 2) рассматривать отношение проекций последовательных приближений  $x^{(k)}$  и  $x^{(k-1)}$  на направление, задаваемое каким-нибудь вектором  $l^{(k)}$ , т. е.

$$\frac{\langle x^{(k)}, l^{(k)} \rangle}{\langle x^{(k-1)}, l^{(k)} \rangle}. \quad (3.222)$$

Во втором случае мы обозначили направление проектирования через  $l^{(k)}$ , чтобы подчеркнуть его возможную зависимость от номера шага  $k$ . Ясно также, что это направление  $l^{(k)}$  не должно быть ортогональным вектору  $x^{(k-1)}$ , чтобы не занулился знаменатель в (3.222).

Последний способ кажется более предпочтительным в вычислительном отношении. Он позволяет избегать капризного поведения в одной отдельно взятой компоненте вектора  $x^{(k)}$ , когда она может сделаться очень малой по абсолютной величине или совсем зануиться, хотя в целом вектор  $x^{(k)}$  будет иметь значительную длину. Наконец, в качестве вектора, задающего направление проектирования во втором варианте, естественно взять сам  $x^{(k)}$ , вычисляя на каждом шаге отношение

$$\frac{\langle x^{(k)}, x^{(k-1)} \rangle}{\langle x^{(k-1)}, x^{(k-1)} \rangle}, \quad (3.223)$$

где  $x^{(k)} = A^k x^{(0)}$ . Нетрудно увидеть, что это выражение совпадает с отношением Рэлея для приближения  $x^{(k-1)}$  к собственному вектору (см. § 3.17e).

Для организации вычислительного алгоритма степенного метода требуется разрешить ещё два тонких момента, связанных с реализацией на ЭВМ.

Во-первых, это возможное неограниченное увеличение (при  $\hat{\lambda} > 1$ ) или неограниченное уменьшение (при  $\hat{\lambda} < 1$ ) норм векторов  $x^{(k-1)}$  и  $x^{(k)}$ , порождаемых в нашем процессе. Разрядная сетка современных цифровых ЭВМ, как известно, конечна и позволяет представлять числа из ограниченного диапазона. Чтобы избежать проблем, вызванных выходом за этот диапазон («переполнением» или «исчезновением порядка»), имеет смысл нормировать  $x^{(k)}$ . При этом наиболее удобна нормировка в евклидовой норме  $\|\cdot\|_2$ , так как тогда знаменатель отношения (3.223) становится равным единице.

Нормировка полезна также для того, чтобы «выровнять» участвующие в арифметических операциях числа, сделать их соизмеримыми друг другу по величине, так как именно при таких условиях достига-

ется наибольшая точность результатов сложения и умножения в арифметике с плавающей точкой на ЭВМ.

Во-вторых, при выводе степенного метода мы неявно предполагали, что начальный вектор  $x^{(0)}$  выбран так, что он имеет ненулевую проекцию на направление доминирующего собственного вектора  $\hat{v}$  матрицы  $A$ . В противном случае произведения любых степеней матрицы  $A$  на  $x^{(0)}$  будут также иметь нулевые проекции на  $\hat{v}$ , и никакой дифференциации длины компонент  $A^k x^{(0)}$ , на которой и основывается степенной метод, не произойдёт. Это затруднение может быть преодолено с помощью какой-нибудь априорной информации о доминирующем собственном векторе матрицы. Кроме того, при практической реализации степенного метода на цифровых ЭВМ неизбежные ошибки округления, как правило, приводят к появлению ненулевых компонент в направлении  $\hat{v}$ , которые затем в процессе итерирования растянутся на нужную величину. Но, строго говоря, это может не происходить в некоторых исключительных случаях, и потому при ответственных вычислениях рекомендуется многократный запуск степенного метода с различными начальными векторами (так называемый мультистарт).

Таблица 3.14. Степенной метод для нахождения доминирующего собственного значения матрицы и соответствующего ему собственного вектора

```

 $k \leftarrow 1;$ 
выбираем вектор  $x^{(0)} \neq 0$  ;
нормируем  $x^{(0)} \leftarrow x^{(0)} / \|x^{(0)}\|_2$  ;
DO WHILE ( метод не сошёлся )
     $y^{(k)} \leftarrow Ax^{(k-1)}$  ;
     $\hat{\lambda} \leftarrow \langle y^{(k)}, x^{(k-1)} \rangle$  ;
     $x^{(k)} \leftarrow y^{(k)} / \|y^{(k)}\|_2$  ;
     $k \leftarrow k + 1$  ;
END DO

```

В псевдокоде, представленном в табл. 3.14,  $\hat{\lambda}$  — это приближённое доминирующее собственное значение матрицы  $A$ , а  $x^{(k)}$  — текущее приближение к нормированному доминирующему собственному вектору.

**Теорема 3.18.1** Пусть  $n \times n$ -матрица  $A$  является матрицей простой структуры (т. е. диагонализуема) и у неё имеется простое доминирующее собственное значение. Если начальный вектор  $x^{(0)}$  не лежит в линейной оболочке  $\text{span}\{v_2, \dots, v_n\}$  собственных векторов  $A$ , которые не являются доминирующими, то степенной метод сходится.

**Доказательство.** При сделанных нами предположениях о матрице  $A$  она может быть представлена в виде

$$A = VDV^{-1},$$

где  $D = \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\}$  — диагональная матрица с собственными значениями  $\lambda_1, \lambda_2, \dots, \lambda_n$  по диагонали, а  $V$  — матрица, осуществляющая преобразование подобия, причём без ограничения общности можно считать, что  $\lambda_1$  — доминирующее собственное значение  $A$ . Матрица  $V$  составлена из собственных векторов  $v_i$  матрицы  $A$  как из столбцов:

$$V = (v_1 \ v_2 \ \cdots \ v_n) = \begin{pmatrix} (v_1)_1 & (v_2)_1 & \cdots & (v_n)_1 \\ (v_1)_2 & (v_2)_2 & \cdots & (v_n)_2 \\ \vdots & \vdots & \ddots & \vdots \\ (v_1)_n & (v_2)_n & \cdots & (v_n)_n \end{pmatrix},$$

где через  $(v_i)_j$  обозначена  $j$ -я компонента  $i$ -го собственного вектора

матрицы  $A$ . При этом можно считать, что  $\|v_i\|_2 = 1$ . Следовательно,

$$\begin{aligned} A^k x^{(0)} &= (VDV^{-1})^k x^{(0)} = \underbrace{(VDV^{-1})(VDV^{-1}) \cdots (VDV^{-1})}_{k \text{ раз}} x^{(0)} = \\ &= VD(V^{-1}V)D(V^{-1}V) \cdots (V^{-1}V)DV^{-1}x^{(0)} = \\ &= VD^k V^{-1}x^{(0)} = VD^k z = \\ &= V \begin{pmatrix} \lambda_1^k z_1 \\ \lambda_2^k z_2 \\ \vdots \\ \lambda_n^k z_n \end{pmatrix} = (\lambda_1^k z_1) V \begin{pmatrix} 1 \\ (\lambda_2/\lambda_1)^k(z_2/z_1) \\ \vdots \\ (\lambda_n/\lambda_1)^k(z_n/z_1) \end{pmatrix}, \end{aligned}$$

где обозначено  $z = V^{-1}x^{(0)}$ . Необходимое условие последнего преобразования этой цепочки —  $z_1 \neq 0$  — выполнено потому, что в условиях теоремы вектор  $x^{(0)} = Vz$  должен иметь ненулевую первую компоненту при разложении по базису из собственных векторов  $A$ , т. е. столбцов матрицы  $V$ .

Коль скоро  $\lambda_1$  — доминирующее собственное значение матрицы  $A$ , т. е.

$$|\lambda_1| > |\lambda_2| \geq \dots \geq |\lambda_n|,$$

то  $|\lambda_1| > 0$  и все частные  $\lambda_2/\lambda_1, \lambda_3/\lambda_1, \dots, \lambda_n/\lambda_1$  существуют и по модулю меньше единицы. Поэтому при  $k \rightarrow \infty$  вектор

$$\begin{pmatrix} 1 \\ (\lambda_2/\lambda_1)^k(z_2/z_1) \\ \vdots \\ (\lambda_n/\lambda_1)^k(z_n/z_1) \end{pmatrix} \tag{3.224}$$

сходится к вектору  $(1, 0, 0, \dots, 0)^\top$ . Соответственно, произведение

$$V \begin{pmatrix} 1 \\ (\lambda_2/\lambda_1)^k(z_2/z_1) \\ \vdots \\ (\lambda_n/\lambda_1)^k(z_n/z_1) \end{pmatrix}$$

сходится к первому столбцу матрицы  $V$ , т. е. к собственному вектору, отвечающему  $\lambda_1$ . Вектор  $x^{(k)}$ , который отличается от  $A^k x^{(0)}$  лишь нормировкой, сходится к собственному вектору  $v_1$ , а величина  $\hat{\lambda} = \langle y^{(k)}, x^{(k-1)} \rangle$  сходится к  $\langle Av_1, v_1 \rangle = \langle \lambda_1 v_1, v_1 \rangle = \lambda_1$ . ■

Из проведённых выше выкладок следует, что быстрота сходимости степенного метода определяется отношениями  $|\lambda_i/\lambda_1|$ ,  $i = 2, 3, \dots, n$ , — знаменателями геометрических прогрессий, стоящих в качестве элементов вектора (3.224). Фактически решающее значение имеет наибольшее из этих отношений, т. е.  $|\lambda_2/\lambda_1|$ , зависящее от того, насколько модуль доминирующего собственного значения отделён от модуля остальной части спектра. Чем больше эта отдалённость, тем быстрее сходимость степенного метода.

**Пример 3.18.1** Для матрицы (3.20)

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

при вычислениях с двойной точностью степенной метод с начальным вектором  $x^{(0)} = (1, 1)^\top$  за 7 итераций даёт семь верных знаков доминирующего собственного значения  $\frac{1}{2}(5 + \sqrt{33}) \approx 5.3722813$ . Детальная картина сходимости показана в следующей табличке:

Номер итерации	Приближение к собственному значению
1	5.0
2	5.3448276
3	5.3739445
4	5.3721649
5	5.3722894
6	5.3722808
7	5.3722814

Быстрая сходимость объясняется малостью величины  $|\lambda_2/\lambda_1|$ , которая, как мы могли видеть в примере 3.2.4, для рассматриваемой матрицы равна всего лишь 0.069.

Для матрицы (3.21)

$$\begin{pmatrix} 1 & 2 \\ -3 & 4 \end{pmatrix},$$

при тех же исходных условиях степенной метод порождает последовательность значений  $\hat{\lambda}$ , которая случайно колеблется от примерно 0.9 до 4 с лишним и очевидным образом не имеет предела. Причина — наличие у матрицы двух одинаковых по абсолютной величине комплексно-сопряжённых собственных значений  $2.5 \pm 1.936i$  (см. пример 3.2.4). ■

Если в матрице нет доминирующего собственного значения, то развитая выше теория степенного метода не работает, а сам он может никуда не сходить (как в примере выше). Но и в этих случаях из последовательности векторов, порождаемой степенным методом, с помощью специальных приёмов иногда всё-таки возможно извлечь информацию о наибольших по модулю собственных значениях матрицы. Детали читатель может увидеть в книгах [4, 48], где описаны, в частности, приёмы уточнения пары комплексно-сопряжённых собственных значений с наибольшим модулем и пары вещественных собственных значений с наибольшим модулем, но противоположными знаками.

Отметим, что для симметричных (эрмитовых) и положительно полуопределённых матриц в степенном методе в качестве приближения к доминирующему собственному значению можно брать отношение

$$\frac{\|x^{(k)}\|_2}{\|x^{(k-1)}\|_2}, \quad \text{где } x^{(k)} = Ax^{(k-1)}$$

(см. [94]) или вообще отношение любых норм векторов  $x^{(k)}$  и  $x^{(k-1)}$ . Это обосновывается тем, что последовательные векторы  $x^{(k-1)}$  и  $x^{(k)}$  при достаточно больших  $k$  становятся почти сонаправленными.

Наконец, необходимо замечание о сходимости степенного метода в комплексном случае. Так как комплексные числа описываются парами вещественных чисел, то комплексные одномерные инвариантные пространства матрицы имеют вещественную размерность 2. Даже будучи нормированными, векторы из такого подпространства могут отличаться на скалярный множитель  $e^{i\varphi}$  для какого-то аргумента  $\varphi$ , так что если не принять специальных мер, то в степенном методе видимой стабилизации координатных представлений комплексных собственных векторов может не наблюдаться. Тем не менее о факте сходимости или расходимости можно при этом судить по стабилизации приближения к собственному значению. Другой способ преодолеть затруднение состоит в том, чтобы кроме нормировки собственных векторов предусмотреть ещё приведение их к такой форме, в которой координатные

представления будут определяться более «жёстко». Например, требованием, чтобы первая компонента вектора была чисто вещественной.

**Пример 3.18.2** Рассмотрим работу степенного метода в применении к комплексной матрице

$$\begin{pmatrix} 1 & 2i \\ 3 & 4i \end{pmatrix},$$

имеющей собственные значения

$$\lambda_1 = -0.4308405 - 0.1485958i, \quad \lambda_2 = 1.4308405 + 4.1485958i.$$

Доминирующим собственным значением здесь очевидно является  $\lambda_2$ .

Начав итерирование с вектора  $x^{(0)} = (1, 1)^\top$ , уже через 7 итераций мы получим 6 правильных десятичных знаков в вещественной и мнимой частях собственного значения  $\lambda_2$ . Порождаемые алгоритмом нормированные векторы  $x^{(k)}$  выглядят следующим образом:

$$x^{(9)} = \begin{pmatrix} -0.01132 - 0.43223i \\ -0.11659 - 0.89413i \end{pmatrix},$$

$$x^{(10)} = \begin{pmatrix} 0.40491 - 0.15163i \\ 0.80725 - 0.40175i \end{pmatrix},$$

$$x^{(11)} = \begin{pmatrix} 0.27536 + 0.33335i \\ 0.64300 + 0.63215i \end{pmatrix},$$

$$x^{(12)} = \begin{pmatrix} 0.22535 + 0.36900i \\ -0.38795 + 0.81397i \end{pmatrix} \text{ и так далее.}$$

В них нелегко «невооружённым глазом» узнать один и тот же собственный вектор, который «крутился» в одномерном комплексном инвариантном подпространстве. Но если поделить все получающиеся векторы на их первую компоненту, то получим один и тот же результат

$$\begin{pmatrix} 1. \\ 2.07430 - 0.21542i \end{pmatrix},$$

и теперь уже налицо факт сходимости собственных векторов. ■

Как ведёт себя степенной метод в случае, когда матрица  $A$  является дефектной, т. е. не имеет простой структуры? Полный анализ ситуации

можно найти, например, в книгах [45, 48]. Наиболее неблагоприятен при этом случай, когда доминирующее собственное значение находится в жордановой клетке размера два и более. Теоретически степенной метод всё таки сходится к этому собственному значению, но уже медленнее любой геометрической прогрессии.

**Пример 3.18.3** Рассмотрим работу степенного метода в применении к матрице

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

т. е. к жордановой  $2 \times 2$ -клетке с собственным значением 1.

Запустив степенной метод из начального вектора  $x^{(0)} = (1, 1)^\top$ , будем иметь следующее

Номер итерации	Приближение к собственному значению
1	1.5
3	1.3
10	1.0990099
30	1.0332963
100	1.009999
300	1.0033333
1000	1.001

Как видим, для получения  $n$  верных десятичных знаков собственного значения приходится делать примерно  $10^{n-1}$  итераций, что, конечно же, непомерно много. Дальнейшее итерирование демонстрирует ту же картину. При увеличении размера жордановой клетки сходимость степенного метода становится ещё более медленной, хотя принципиально не меняется. ■

Дальнейшим обобщением степенного метода является метод итерирования подпространства [13, 46, 48, 96, 126].<sup>37</sup> В нём уточняется не одно доминирующее собственное значение, а целое семейство таких собственных значений вместе с подпространством, которое является линейной оболочкой соответствующих собственных векторов.

<sup>37</sup>По-английски — subspace iteration. Распространены также альтернативные термины «ортогональное итерирование» и «одновременное итерирование» [13, 45, 46, 107]. Аналогичный по духу приём кратко описан в книге [48].

### 3.18б Обратные степенные итерации

*Обратными степенными итерациями* для матрицы  $A$  называют описанный в предыдущем разделе степенной метод, применённый к обратной матрице  $A^{-1}$ , в котором вычисляется отношение результатов предыдущей итерации к следующей, т. е. величина, обратная к (3.221) или (3.222). Явное нахождение обратной матрицы  $A^{-1}$  при этом не требуется, так как в степенном методе используется лишь результат  $x^{(k)}$  её умножения на вектор  $x^{(k-1)}$  очередного приближения, а это эквивалентно решению системы линейных уравнений  $Ax^{(k)} = x^{(k-1)}$  (см. § 3.14).

Собственные значения матриц  $A$  и  $A^{-1}$  взаимно обратны, а собственные векторы одинаковы (предложение 3.2.2). Поэтому обратные степенные итерации будут сходиться к наименьшему по абсолютной величине собственному значению  $A$  и соответствующему собственному вектору.

Чтобы в обратном к (3.222) отношении

$$\frac{\langle x^{(k-1)}, l^{(k)} \rangle}{\langle x^{(k)}, l^{(k)} \rangle},$$

которое необходимо вычислять в обратных степенных итерациях, знаменатель не занулялся, удобно брать  $l^{(k)} = x^{(k)}$ . Тогда очередным приближением к наименьшему по модулю собственному значению матрицы  $A$  является

$$\check{\lambda} = \frac{\langle x^{(k-1)}, x^{(k)} \rangle}{\langle x^{(k)}, x^{(k)} \rangle},$$

где  $Ax^{(k)} = x^{(k-1)}$ . Псевдокод получающегося численного метода представлен в табл. 3.15.

На каждом шаге обратных степенных итераций нужно решать систему линейных алгебраических уравнений с одной и той же матрицей (5-я строка псевдокода). Практическая реализация этого решения может быть сделана достаточно эффективной, если предварительно выполнить LU- или QR-разложение матрицы, а затем на каждом шаге метода использовать равносильные представления системы в виде (3.88) или (3.107). Их решение сводится к выполнению прямой и обратной подстановок для треугольных СЛАУ в случае LU-разложения или умножению на ортогональную матрицу и обратную подстановку в случае QR-разложения.

Таблица 3.15. Обратные степенные итерации для нахождения наименьшего по модулю собственного значения матрицы  $A$  и соответствующего ему собственного вектора

```

 $k \leftarrow 1;$ 
выбираем вектор  $x^{(0)} \neq 0$ ;
DO WHILE ( метод не сопрёлся )
    найти  $y^{(k)}$  из системы  $Ay^{(k)} = x^{(k-1)}$  ;
     $\lambda \leftarrow \langle x^{(k-1)}, y^{(k)} \rangle / \langle y^{(k)}, y^{(k)} \rangle$  ;
     $x^{(k)} \leftarrow y^{(k)} / \|y^{(k)}\|_2$  ;
     $k \leftarrow k + 1$  ;
END DO

```

**Пример 3.18.4** Рассмотрим работу обратных степенных итераций для знакомой нам матрицы (см. примеры 3.17.5 и 3.18.1)

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix},$$

собственные значения которой суть  $\frac{1}{2}(5 \pm \sqrt{33})$ , приблизительно равные  $-0.372$  и  $5.372$ .

Запустив обратные степенные итерации из начального вектора  $x^{(0)} = (1, 1)^\top$ , за 7 итераций получим 7 верных значащих цифр наименьшего по модулю собственного числа  $0.3722813$ . Скорость сходимости здесь получается такой же, как в примере 3.18.1 для доминирующего собственного значения этой матрицы, что неудивительно ввиду одинакового значения знаменателя геометрической прогрессии  $\lambda_2/\lambda_1$ . ■

Обратные степенные итерации особенно эффективны в комбинации со сдвигами матрицы, которые описываются в следующем разделе, когда имеется хорошее приближение к собственному значению и требуется найти соответствующий собственный вектор. Дальнейшим развитием обратных степенных итераций являются итерации с отнoshением

Рэлея, которым посвящён § 3.17e.

### 3.18в Сдвиги спектра, исчерпывание и понижение порядка

В названии этого раздела перечислены популярные приёмы преобразования матриц, которые помогают свести решение проблемы собственных значений к решению этой же задачи для другой матрицы, которая более удобна, менее сложна или же обладает какими-то другими необходимыми свойствами.

*Сдвигом* матрицы называют прибавление к ней скалярной матрицы, т. е. матрицы, пропорциональной единичной матрице. При этом вместо матрицы  $A$  мы получаем матрицу  $A + \vartheta I$  для некоторого вещественного или комплексного числа  $\vartheta$ . Если  $\lambda$  — собственное значение матрицы  $A$ , то

$$Av = \lambda v$$

для ненулевого собственного вектора  $v$ . Поэтому

$$(A + \vartheta I)v = Av + \vartheta v = \lambda v + \vartheta v = (\lambda + \vartheta)v,$$

так что число  $\lambda + \vartheta$  становится собственным значением матрицы  $A + \vartheta I$ , а соответствующие ему собственные векторы остаются неизменными. Сдвиги часто применяют для преобразования спектра матрицы с тем, чтобы улучшить работу некоторых алгоритмов решения проблемы собственных значений.

Если, к примеру, у матрицы  $A$  наибольшими по абсолютной величине были два собственных значения  $-2$  и  $2$ , то прямое применение к ней степенного метода не приведёт к успеху. Но у матрицы  $A + I$  эти собственные значения перейдут в  $-1$  и  $3$ , второе собственное число сделается наибольшим по модулю и теперь уже единственным доминирующим. Соответственно, степенной метод станет применимым к новой матрице, и после его работы нужно вычесть из результата величину сдвига.

**Пример 3.18.5** Для матрицы (3.21)

$$\begin{pmatrix} 1 & 2 \\ -3 & 4 \end{pmatrix},$$

как отмечено в примере 3.18.1, простейший степенной метод расходится из-за существования двух наибольших по абсолютной величине собственных значений.

Но если сдвинуть эту матрицу на  $2i$ , то её спектр (рис. 3.33) поднимется «вверх» на комплексной плоскости, абсолютные величины собственных значений перестанут совпадать, и степенной метод окажется применимым. Степенные итерации для «сдвинутой» матрицы

$$\begin{pmatrix} 1 + 2i & 2 \\ -3 & 4 + 2i \end{pmatrix} \quad (3.225)$$

довольно быстро сходятся к наибольшему по модулю собственному значению  $\frac{5}{2} + (2 + \frac{1}{2}\sqrt{15})i \approx 2.5 + 3.93649i$ . Детальная картина сходимости при вычислениях с двойной точностью и начальным вектором  $x^{(0)} = (1, 1)^\top$  показана в следующей табличке:

Номер итерации	Приближение к собственному значению
1	$2.0 + 2.0i$
3	$2.0413 + 4.3140i$
5	$2.7022 + 3.9373i$
10	$2.5005 + 3.9455i$
20	$2.5000 + 3.9365i$

Для матрицы (3.225) модуль отношения второго собственного значения к доминирующему, который определяет скорость сходимости, равен  $|\lambda_2/\lambda_1| \approx 0.536$ . Ещё большее ускорение сходимости степенного метода можно получить при сдвиге исходной матрицы на  $(-2 + 2i)$ , когда отношение модулей этих собственных значений становится равным всего 0.127. ■

Поскольку спектр симметричной (эрмитовой) матрицы лежит на вещественной оси, то к таким матрицам имеет смысл применять вещественные сдвиги. В частности, для симметричных вещественных матриц алгоритмы будут реализовываться при этом в более простой вещественной арифметике.

С помощью сдвигов матрицы можно любое её собственное значение, которое является вершиной многоугольника, описывающего спектр,<sup>38</sup>

<sup>38</sup>Более точно, крайней точкой выпуклой оболочки спектра.

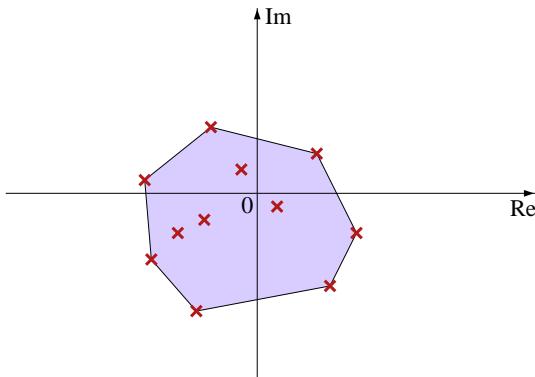


Рис. 3.34. С помощью подходящего сдвига матрицы любую крайнюю точку выпуклой оболочки спектра можно сделать наибольшей по модулю

сделать наибольшим по модулю, обеспечив, таким образом, сходимость к нему итераций степенного метода (рис. 3.34). Но как добиться сходимости к другим собственным значениям, которые лежат «внутри» многоугольника спектра, а не «с краю»? Здесь могут помочь обратные степенные итерации.

Обратные степенные итерации сходятся к ближайшей к нулю точке спектра матрицы, и такой точкой с помощью подходящего сдвига можно сделать любое собственное значение. В этом — принципиальное преимущество сдвигов для обратных степенных итераций.

Другое важное следствие сдвигов — изменение отношения  $|\lambda_2/\lambda_1|$ , величина которого влияет на скорость сходимости степенного метода. Обычно с помощью подходящего выбора величины сдвига  $\vartheta$  можно добиться того, чтобы

$$\left| \frac{\lambda_2 + \vartheta}{\lambda_1 + \vartheta} \right|$$

стало меньшим, чем  $|\lambda_2/\lambda_1|$ , ускорив тем самым степенные итерации. Это мы могли видеть в примере 3.18.5. Совершенно аналогичный эффект оказывает удачный выбор сдвига на отношение  $|\lambda_n/\lambda_{n-1}|$ , которое определяет скорость сходимости обратных степенных итераций.

Опишем теперь два популярных метода *исчерпывания матрицы*. Так называют процедуры для построения вспомогательных матриц,

которые имеют спектром все собственные значения исходной матрицы за исключением какого-то выделенного или уже найденного.<sup>39</sup>

Предположим, что для матрицы  $A$  известно собственное значение  $\lambda$  и соответствующий ему правый собственный вектор  $v$ . Они могут быть как вещественными, так и комплексными. Пусть также  $u$  — какой-нибудь вектор, удовлетворяющий условию  $u^*v = 1$ . Задавшись каким-нибудь числом  $\kappa$ , образуем матрицу

$$\tilde{A} = A - \kappa(vu^*),$$

где, напомним,

$$vu^* = \begin{pmatrix} v_1\bar{u}_1 & v_1\bar{u}_2 & \cdots & v_1\bar{u}_n \\ v_2\bar{u}_1 & v_2\bar{u}_2 & \cdots & v_2\bar{u}_n \\ \vdots & \vdots & \ddots & \vdots \\ v_n\bar{u}_1 & v_n\bar{u}_2 & \cdots & v_n\bar{u}_n \end{pmatrix}.$$

Оказывается, что спектр матрицы  $\tilde{A}$  совпадает со спектром матрицы  $A$  за исключением собственного значения  $\lambda$ , вместо которого появляется собственное значение  $\lambda - \kappa$ .

В самом деле, если  $v$  — собственный вектор, отвечающий собственному значению  $\lambda$  в  $A$ , то

$$\tilde{A}v = (A - \kappa(vu^*))v = Av - \kappa v(u^*v) = \lambda v - \kappa v = (\lambda - \kappa)v$$

в силу условия, наложенного на  $u$ . С другой стороны, если  $\mu$  — какое-то собственное число матрицы  $A$ , отличное от  $\lambda$ , а  $w$  — отвечающий ему левый собственный вектор, то

$$w^*\tilde{A} = w^*A - w^*(\kappa vu^*) = w^*A - \kappa w^*vu^* = \mu w^*$$

в силу ортогональности векторов  $v$  и  $w$  (см. § 3.2в, стр. 347). Так как левые и правые собственные значения матрицы совпадают друг с другом, то  $\mu$  должно быть в спектре матрицы  $\tilde{A}$ .

Описанное исчерпывание называют *исчерпыванием Виландта* (по-английски — *Wielandt deflation*) по имени предложившего её математика. Вычислительная практика показывает, что оно может быть не

---

<sup>39</sup>Общее англоязычное название процедур исчерпывания — matrix deflation, несмотря на то, что технически они могут сильно различаться.

очень устойчивым для некоторых «плохих» матриц. Из проведённого выше обоснования исчерпывания Виландта следует также, что оно оставляет неизменными все левые собственные векторы матрицы  $A$  и правый собственный вектор, отвечающий «исчерпываемому» собственному значению.

Ясно, что существует бесконечно много способов выбора вектора  $u$ , такого что  $u^*v = 1$ . Один из наиболее популярных состоит в том, чтобы взять  $u$  в виде левого собственного вектора матрицы  $A$ , который отвечает тому же собственному значению  $\lambda$ , и нормированного необходимым способом. Преимуществом такого выбора  $u$  является то обстоятельство, что левые и правые собственные векторы матриц  $A$  и  $\tilde{A}$  оказываются равными друг другу. В частности, этот приём особенно удобен в случае, когда матрица  $A$  симметрична (эрмитова) и её левые и правые собственные векторы совпадают.

Наконец, рассмотрим процедуру исчерпывания, которую называют *понижением порядка*. Пусть  $v = (v_1, v_2, \dots, v_n)^\top$  — собственный вектор, соответствующий собственному значению  $\lambda$ . Рассмотрим матрицу

$$E = \begin{pmatrix} v_1 & 0 & \cdots & 0 \\ v_2 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ v_n & 0 & \cdots & 1 \end{pmatrix},$$

отличающуюся от единичной только первым столбцом, в котором выписаны компоненты собственного вектора. Нетрудно вычислить обратную к ней матрицу:

$$E^{-1} = \begin{pmatrix} 1/v_1 & 0 & \cdots & 0 \\ -v_2/v_1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -v_n/v_1 & 0 & \cdots & 1 \end{pmatrix}.$$

В её первом столбце, начиная со второго элемента, выписаны отношения компонент собственного вектора к его первой компоненте, взятые

с противоположным знаком.<sup>40</sup> Тогда произведение  $E^{-1}AE$  имеет вид

$$E^{-1}AE = \begin{pmatrix} \lambda & \frac{a_{12}}{v_1} & \cdots & \frac{a_{1n}}{v_1} \\ 0 & a_{22} - \frac{v_2}{v_1}a_{12} & \cdots & a_{2n} - \frac{v_2}{v_1}a_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2} - \frac{v_n}{v_1}a_{12} & \cdots & a_{nn} - \frac{v_n}{v_1}a_{1n} \end{pmatrix}. \quad (3.226)$$

Оно является блочно-треугольным с блоками  $1 \times 1$  и  $(n-1) \times (n-1)$  по главной диагонали.

С другой стороны, матрица  $E^{-1}AE$  подобна исходной матрице  $A$ , так что их собственные значения совпадают. Следовательно,  $(n-1) \times (n-1)$ -матрица

$$B = \begin{pmatrix} a_{22} - \frac{v_2}{v_1}a_{12} & \cdots & a_{2n} - \frac{v_2}{v_1}a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n2} - \frac{v_n}{v_1}a_{12} & \cdots & a_{nn} - \frac{v_n}{v_1}a_{1n} \end{pmatrix},$$

т.е. второй диагональный блок в (3.226), имеет собственными значениями все те собственные значения матрицы  $A$ , которые отличаются от  $\lambda$ . Для их нахождения можно привлекать любые доступные численные методы, например, вычислить доминирующее собственное значение степенным методом.

Пусть  $u$  — собственный  $(n-1)$ -вектор матрицы  $B$ , отвечающий её собственному значению  $\mu$ , которое является также собственным значением для исходной матрицы  $A$ . Предположим, что он уже известен, и для удобства положим  $u = (u_2, \dots, u_n)^\top$ . Будем теперь искать соответствующий собственный вектор матрицы  $E^{-1}AE$  в виде

$$\left( \frac{u_1}{u} \right) = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix},$$

---

<sup>40</sup>Матрицы  $E$  и  $E^{-1}$  совпадают по структуре с матрицами вида (3.85), которые возникали в § 3.6 при анализе метода Гаусса, и поэтому обозначаются аналогично.

где необходимо определить число  $u_1$ .

Согласно определению собственного значения и собственного вектора должно выполняться векторное равенство

$$\left( \begin{array}{c|ccc} \lambda & a_{12} & \cdots & a_{1n} \\ \hline 0 & & & \\ \vdots & & B & \\ 0 & & & \end{array} \right) \begin{pmatrix} u_1 \\ u \end{pmatrix} = \mu \begin{pmatrix} u_1 \\ u \end{pmatrix}.$$

Приравнивая в нём первые компоненты, найдём

$$\lambda u_1 + a_{12}u_2 + \dots + a_{1n}u_n = \mu u_1,$$

откуда

$$u_1 = \frac{a_{12}u_2 + \dots + a_{1n}u_n}{\mu - \lambda}.$$

Собственный вектор исходной матрицы  $A$  получается тогда равным

$$E \begin{pmatrix} u_1 \\ u \end{pmatrix} = \begin{pmatrix} u_1 v_1 \\ u_1 v_2 + u_2 \\ \vdots \\ u_1 v_n + u_n \end{pmatrix}.$$

О других конструкциях исчерпывания матрицы читатель может узнать, например, из книг [45, 48, 96, 107, 126]. Теоретически с помощью исчерпываний можно найти все собственные числа матрицы, но на практике некоторые из этих процедур приводят к существенному понижению точности, особенно для малых по абсолютной величине собственных значений. Как следствие, процедуры исчерпывания применяют ограниченным образом.

### 3.18г Базовый QR-алгоритм

QR-алгоритм, изложению которого посвящён этот параграф, является одним из наиболее эффективных численных методов для решения полной проблемы собственных значений. Он был изобретён независимо

В.Н. Кублановской (1960 год) и Дж. Фрэнсисом (1961 год). Публикация В.Н. Кублановской появилась раньше,<sup>41</sup> а Дж. Фрэнсис более полно развил некоторые практические модификации QR-алгоритма.

QR-алгоритм — представитель большого семейства родственных методов решения полной проблемы собственных значений, которые основаны на разложении матриц, получающихся на последовательных шагах алгоритма, на простые сомножители. В частности, QR-алгоритму предшествовал LR-алгоритм Х. Рутисхаузера [48]. На практике применяются также предложенный В.В. Воеводиным ортогональный степенной метод [71] и другие близкие вычислительные процессы.

Вспомним теорему о QR-разложении (теорема 3.7.2, стр. 478): всякая квадратная матрица представима в виде произведения ортогональной и правой (верхней) треугольной матриц. Ранее мы также обсуждали конструктивные способы выполнения этого разложения — с помощью матриц вращений (§ 3.7д) и с помощью матриц отражения Хаусхолдера (§ 3.7е). Следовательно, далее можно считать, что QR-разложение всегда выполнимо и основывать на этом факте свои построения.

Вычислительная схема базового QR-алгоритма для решения проблемы собственных значений представлена в табл. 3.16: мы разлагаем матрицу  $A^{(k-1)}$ , поступившую на  $k$ -й шаг алгоритма,  $k = 1, 2, \dots$ , на ортогональный  $Q^{(k)}$  и правый треугольный  $R^{(k)}$  сомножители и далее, поменяв их местами, умножаем друг на друга, образуя следующее приближение  $A^{(k)}$ .

Прежде всего отметим, что поскольку

$$A^{(k)} = R^{(k)}Q^{(k)} = (Q^{(k)})^\top (Q^{(k)}R^{(k)})Q^{(k)} = (Q^{(k)})^\top A^{(k-1)}Q^{(k)}, \quad (3.227)$$

то все матрицы  $A^{(k)}$ ,  $k = 1, 2, \dots$ , ортогонально подобны друг другу и исходной матрице  $A$ . Поэтому собственные значения всех матриц  $A^{(k)}$  совпадают с собственными значениями  $A$ . Результат о сходимости QR-алгоритма неформальным образом может быть резюмирован в следующем виде: если  $A$  — неособенная вещественная матрица, то последовательность порождаемых QR-алгоритмом матриц  $A^{(k)}$  сходится «по форме» к верхней блочно-треугольной матрице. Определим это понятие более точно.

---

<sup>41</sup> Упоминая о вкладе В.Н. Кублановской в изобретение QR-алгоритма, обычно ссылаются на её статью 1961 года в «Журнале вычислительной математики и математической физики» [91]. Но самое первое сообщение о QR-алгоритме было опубликовано ею раньше — в Дополнении к изданию 1960 года книги [48].

Таблица 3.16. Базовый QR-алгоритм для нахождения собственных значений матрицы  $A$

```

 $k \leftarrow 1;$ 
 $A^{(0)} \leftarrow A;$ 
DO WHILE ( метод не сопшёлся )
    вычислить QR-разложение  $A^{(k-1)} = Q^{(k)}R^{(k)}$ ;
     $A^{(k)} \leftarrow R^{(k)}Q^{(k)};$ 
     $k \leftarrow k + 1;$ 
END DO

```

Пусть даны верхняя (правая) блочно-треугольная матрица

$$B = \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1,n-1} & B_{1n} \\ 0 & B_{22} & \cdots & B_{2,n-1} & B_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & B_{mn} \end{pmatrix}$$

и последовательность матриц  $A^{(k)}$ ,  $k = 1, 2, \dots$ , одинаковых с  $B$  размеров. Разобьём матрицы  $A^{(k)}$  на блоки согласно блочной структуре матрицы  $B$ :

$$A^{(k)} = \begin{pmatrix} A_{11}^{(k)} & A_{12}^{(k)} & \cdots & A_{1,n-1}^{(k)} & A_{1n}^{(k)} \\ A_{21}^{(k)} & A_{22}^{(k)} & \cdots & A_{2,n-1}^{(k)} & A_{2n}^{(k)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ A_{m1}^{(k)} & A_{m2}^{(k)} & \cdots & A_{m,n-1}^{(k)} & A_{mn}^{(k)} \end{pmatrix}.$$

Будем говорить, что последовательность матриц  $\{A^{(k)}\}$  *сходится по форме* к блочно-треугольной матрице  $B$ , если все элементы матриц  $A^{(k)}$  равномерно по  $k$  ограничены по модулю, а блоки ниже диагонали сходятся к нулю (нулевым матрицам). Аналогично можно определить

сходимость по форме к нижней (левой) блочно-треугольной матрице или даже к блочно-диагональной матрице.

В применении к общим матрицам QR-алгоритм сходится по форме либо к верхней треугольной матрице, либо к верхней блочно-треугольной. При этом размеры диагональных блоков зависят, во-первых, от типа собственных значений матрицы (кратности и принадлежности вещественной оси  $\mathbb{R}$ ), и, во-вторых, от того, в вещественной или комплексной арифметике выполняется QR-алгоритм.

Если алгоритм выполняется в вещественной (комплексной) арифметике и все собственные значения матрицы вещественны (комплексны) и различны по модулю, то предельная матрица — верхняя треугольная. Если алгоритм выполняется в вещественной (комплексной) арифметике и некоторое собственное значение матрицы вещественно (комплексно) и имеет кратность  $p$ , то в предельной матрице ему соответствует диагональный блок размера  $p \times p$ . Если алгоритм выполняется для вещественной матрицы в вещественной арифметике, то простым комплексно-сопряжённым собственным значениям (они имеют равные модули) отвечают диагональные  $2 \times 2$ -блоки в предельной матрице. Наконец, если некоторое комплексное собственное значение вещественной матрицы имеет кратность  $p$ , так что ему соответствует ещё такое же комплексно-сопряжённое собственное значение кратности  $p$ , то при выполнении QR-алгоритма в вещественной арифметике предельная матрица получит диагональный блок размера  $2p \times 2p$ .

**Пример 3.18.6** Проиллюстрируем работу QR-алгоритма на примере матрицы

$$\begin{pmatrix} 1 & -2 & 3 \\ 4 & 5 & -6 \\ -7 & 8 & 9 \end{pmatrix}, \quad (3.228)$$

имеющей собственные значения

$$2.7584 \quad \text{и} \quad 6.1207 \pm 8.04789i.$$

Читатель может провести на компьютере этот увлекательный эксперимент самостоятельно, воспользовавшись системами Scilab, Octave, MATLAB или им подобными: все они имеют встроенную процедуру для QR-разложения матриц.<sup>42</sup>

---

<sup>42</sup> В Scilab'e, MATLAB'e и Octave она так и называется — qr.

Через 20 итераций QR-алгоритм выдаёт матрицу

$$\begin{pmatrix} 6.0821 & -5.2925 & -3.3410 \\ 12.238 & 6.1594 & 3.6766 \\ -8.04 \cdot 10^{-11} & 2.24 \cdot 10^{-11} & 2.7584 \end{pmatrix},$$

в которой угадывается блочно-диагональная матрица с ведущим с  $2 \times 2$ -блоком

$$\begin{pmatrix} 6.0821 & -5.2925 \\ 12.238 & 6.1594 \end{pmatrix}.$$

Этот блок «скрывает» два комплексно-сопряжённых собственных значения —  $6.1208 \pm 8.0479i$ , которые легко получаются из решения квадратного характеристического уравнения. Третье собственное значение матрицы находится в  $1 \times 1$ -блоке, который расположен на месте  $(3, 3)$ , и оно равно 2.7584.

При дальнейшем итерировании элементы матрицы на местах  $(3, 1)$  и  $(3, 2)$  стремятся к нулю, но ведущий  $2 \times 2$ -блок никак не распадается, а его элементы не стабилизируются, принимая кажущиеся хаотичными значения. Таким образом, в данном случае QR-алгоритм сходится по форме к блочно-треугольной матрице с блоками размера  $2 \times 2$  и  $1 \times 1$  по диагонали. ■

**Пример 3.18.7** Для ортогональной матрицы

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad (3.229)$$

QR-разложением является произведение её самой на единичную матрицу. Поэтому в результате одного шага QR-алгоритма мы снова получим исходную матрицу, которая, следовательно, и будет пределом итераций. В то же время, матрица (3.229) имеет собственные значения, равные  $\pm 1$ , так что в данном случае QR-алгоритм формально не работает. Ситуация исправляется сдвигами матрицы, описываемыми в следующем разделе.

Совершенно аналогична ситуация с  $3 \times 3$ -матрицей

$$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

и такими же матрицами больших размеров. ■

### 3.18д Модификации QR-алгоритма

Простейшая версия QR-алгоритма, представленная в табл. 3.16, на практике обычно снабжается рядом модификаций, которые существенно повышают её эффективность и расширяют сферу применимости. Главными из этих модификаций являются

- 1) сдвиги матрицы, рассмотренные ранее в § 3.18в, и
- 2) предварительное приведение матрицы к хессенберговой, т. е. верхней почти треугольной форме, описанное в § 3.17ж.

Аналогично степенному методу, QR-алгоритм также сильно ускоряется с помощью сдвигов, что можно обосновать и теоретически [13, 44]. При практической реализации QR-алгоритма сдвиги часто организуют способом, представленным в табл. 3.17.

Таблица 3.17. QR-алгоритм со сдвигами для нахождения собственных значений матрицы  $A$

```

 $k \leftarrow 1;$ 
 $A^{(0)} \leftarrow A;$ 
DO WHILE ( метод не сопшёлся )
    выбрать сдвиг  $\vartheta_k$ , приближённо равный
    собственному значению  $A$  ;
    вычислить QR-разложение сдвинутой
    матрицы  $A^{(k-1)} - \vartheta_k I = Q^{(k)} R^{(k)}$ ;
     $A^{(k)} \leftarrow R^{(k)} Q^{(k)} + \vartheta_k I$ ;
     $k \leftarrow k + 1$ ;
END DO

```

Особенность выполнения сдвигов в этом псевдокоде — присутствие обратных сдвигов (в строке 8 алгоритма) сразу же вслед за прямыми (в 6-й и 7-й строках). Из-за этого в получающемся алгоритме последовательно вычисляемые матрицы  $A^{(k-1)}$  и  $A^{(k)}$  ортогонально подобны,

совершенно так же, как и в исходной версии QR-алгоритма:

$$\begin{aligned} A^{(k)} &= R^{(k)}Q^{(k)} + \vartheta_k I = (Q^{(k)})^\top Q^{(k)} R^{(k)} Q^{(k)} + \vartheta_k (Q^{(k)})^\top Q^{(k)} = \\ &= (Q^{(k)})^\top (Q^{(k)} R^{(k)} + \vartheta_k I) Q^{(k)} = (Q^{(k)})^\top A^{(k-1)} Q^{(k)}. \end{aligned}$$

Представленная организация сдвигов позволяет сделать их в одно и то же время локальными и динамическими по характеру, т. е. изменяющимися от шага к шагу. По этой причине их можно корректировать на основе результатов промежуточных вычислений. Отметим, что традиционные сдвиги «без восстановления» также широко используются при реализации QR-алгоритма.

**Пример 3.18.8** Проиллюстрируем работу QR-алгоритма со сдвигами на знакомой нам матрице (3.228)

$$\begin{pmatrix} 1 & -2 & 3 \\ 4 & 5 & -6 \\ -7 & 8 & 9 \end{pmatrix}$$

из примера в предыдущем разделе.

Запустим сначала для этой матрицы QR-алгоритм без сдвигов. Через 5 итераций получается матрица

$$\begin{pmatrix} 4.6508 & -11.925 & 4.2996 \\ 5.6124 & 7.578999 & 2.4637651 \\ 0.015767 & -0.0055973 & 2.7702 \end{pmatrix},$$

из которой можно ясно увидеть постепенное выделение блочно-треугольной формы с ведущим  $2 \times 2$ -блоком: элементы на местах (3, 1) и (3, 2) делаются маленькими в сравнении с другими элементами матрицы. Поэтому элемент на месте (3, 3) должен быть близок к собственному значению, и на его величину можно сделать сдвиг.

Положив  $\vartheta_k = 2.77$  в алгоритме табл. 3.17, получим резкое ускорение сходимости, так что после 8-й итерации достигается матрица

$$\begin{pmatrix} 3.1391 & -10.546 & 4.8526 \\ 6.9846 & 9.1025 & 1.0639 \\ 4.13 \cdot 10^{-11} & -2.77 \cdot 10^{-12} & 2.7584 \end{pmatrix}.$$

Напомним, что в примере 3.18.6 для получения сравнимого результата потребовалось 20 итераций базового QR-алгоритма. Сходимость ускорялась бы ещё значительно, если динамически подстраивать параметр сдвига  $\vartheta_k$  на отдельных шагах QR-алгоритма. ■

Теоретической основой второй модификации QR-алгоритма служит следующий результат.

**Предложение 3.18.1** *Матрица, имеющая хессенбергову форму, сохраняет эту форму при выполнении с ней QR-алгоритма.*

**Доказательство.** Предположим, что к началу  $k$ -го шага алгоритма (табл. 3.17) получена хессенбергова матрица  $A^{(k-1)}$ . Затем её сдвигают и организуют QR-разложение

$$A^{(k-1)} - \vartheta_k I = Q^{(k)} R^{(k)},$$

причём в качестве ортогонального сомножителя  $Q^{(k)}$  получается также хессенбергова матрица.

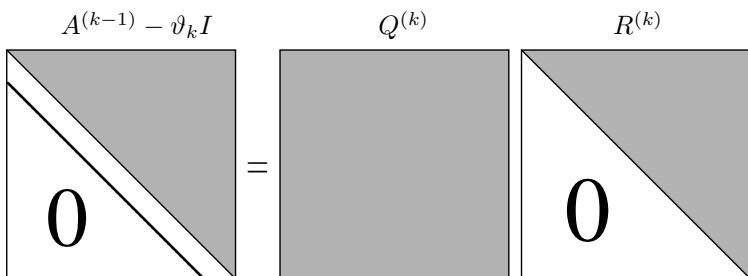


Рис. 3.35. Разложение матрицы  $(A^{(k-1)} - \vartheta_k I)$  в QR-алгоритме

В самом деле, поскольку матрица  $R^{(k)}$  — правая треугольная, то  $j$ -й столбец в  $(A^{(k-1)} - \vartheta_k I)$  есть линейная комбинация первых  $j$  столбцов матрицы  $Q^{(k)}$  с коэффициентами, равными элементам из  $j$ -го столбца  $R^{(k)}$  (рис. 3.35). Отсюда следует, что первый столбец матрицы  $Q^{(k)}$  должен выглядеть совершенно так же, как первый столбец в матрице  $(A^{(k-1)} - \vartheta_k I)$ , т. е. иметь нулевыми элементы на местах (3,1), (4,1) и т. д.

Переходя ко второму столбцу матрицы  $(A^{(k-1)} - \vartheta I)$ , мы видим, что он имеет нулевыми элементы на местах (4,2), (5,2) и т. д., будучи в то же время линейной комбинацией двух первых столбцов матрицы  $Q^{(k)}$ . Так как мы уже знаем, что первый столбец  $Q^{(k)}$  имеет нули в позициях 3-й, 4-й и т. д., то никакого вклада в линейную комбинацию со вторым столбцом  $Q^{(k)}$  в компонентах 3-й, 4-й и т. д. он не вносит. Итак, второй столбец матрицы  $Q^{(k)}$  должен выглядеть аналогично второму столбцу в  $(A^{(k-1)} - \vartheta I)$ , т. е. иметь нулевыми элементы на местах (4,2), (5,2) и т. д. Продолжая эти рассуждения для следующих столбцов матриц  $Q^{(k)}$  и  $(A^{(k-1)} - \vartheta I)$ , можем заключить, что  $Q^{(k)}$  тоже является хессенберговой.

В свою очередь, матрица  $R^{(k)}Q^{(k)}$  — произведение после перестановки сомножителей — опять получается хессенберговой. Добавление диагонального слагаемого  $\vartheta_k I$  не изменяет верхней почти треугольной формы матрицы. Таким образом, к началу следующего  $k + 1$ -го шага QR-алгоритма снова получается матрица в хессенберговой форме. ■

Предварительное приведение матрицы к хессенберговой форме существенно ускоряет QR-алгоритм. Хотя это приведение требует  $O(n^3)$  операций, дальнейшее выполнение одного шага QR-алгоритма с хессенберговой формой будет теперь стоить всего  $O(n^2)$  операций. Для сходимости QR-алгоритму обычно требуется  $O(n)$  шагов, и потому его общая трудоёмкость составит  $O(n^3)$ . Но у исходной версии QR-алгоритма, которая оперирует с плотно заполненной матрицей, трудоёмкость равна примерно  $O(n^4)$ , поскольку на каждой итерации алгоритма выполнение QR-разложения требует  $O(n^3)$  операций.

### 3.19 Численные методы для симметричной проблемы собственных значений и сингулярного разложения

Симметричная проблема собственных значений, как было показано в § 3.17б, 3.17в и 3.17г, своими свойствами заметно выделяется из общей проблемы собственных значений матриц. Для симметричных (в комплексном случае — эрмитовых) матриц все собственные значения вещественны, а собственные векторы ортогональны. Симметричная проблема собственных значений обладает хорошей обусловленностью. Кроме

того, умелая эксплуатация симметричной (эрмитовой) структуры матрицы также повышает эффективность решения задачи. Эти обстоятельства позволяют вычленить её в качестве отдельной части общей проблемы собственных значений.

С симметричной проблемой собственных значений тесно связана задача нахождения сингулярных чисел и сингулярных векторов матрицы (см. § 3.2д). Поэтому логично рассматривать вместе численные методы для симметричной проблемы собственных значений и сингулярного разложения.

Ниже кратко и весьма фрагментарно описаны некоторые характерные подходы к вычислительному решению симметричной проблемы собственных значений и нахождению сингулярного разложения. Более полную информацию по теме читатель может найти в книгах [11, 13, 96] и специализированных журнальных публикациях.

### 3.19а Симметричный QR-алгоритм

*Симметричным QR-алгоритмом* называют QR-алгоритм, описанный в § 3.18г и 3.18д, в применении к симметричным и эрмитовым матрицам. Этот вариант общего алгоритма имеет свои особенности — простоту и малую трудоёмкость.

Во-первых, в процессе работы такого QR-алгоритма порождается последовательность симметричных (эрмитовых) матриц. Это следует из того, что все они ортогонально подобны исходной, т. е. конгруэнтны ей (см. выкладку на стр. 658). Если исходная матрица является симметричной (эрмитовой), то на каждом шаге QR-алгоритма тоже получаются симметричные (эрмитовы) матрицы.

Во-вторых, у симметричных матриц собственные значения вещественны, и потому вещественный QR-алгоритм сходится к правой треугольной матрице без каких-либо диагональных  $2 \times 2$ -блоков, которые соответствовали бы комплексным собственным значениям. С учётом первого замечания это означает, что симметричный QR-алгоритм сходится к диагональной матрице.

В-третьих, симметричный QR-алгоритм допускает элегантное решение вопроса с критерием остановки. Он может быть основан на теореме Гершгорина (теорема 3.17.5, стр. 628). Из неё следует, что если  $A^{(k)} = (a_{ij}^{(k)})$  —  $n \times n$ -матрица, полученная на  $k$ -ом шаге QR-алгоритма, то её собственные значения локализованы в интервалах

$[a_{ii}^{(k)} - \Delta, a_{ii}^{(k)} + \Delta]$ ,  $i = 1, 2 \dots, n$ , где

$$\Delta = \max_{1 \leq i \leq n} \sum_{j \neq i} |a_{ij}^{(k)}|.$$

При общем уменьшении величины внедиагональных элементов значение  $\Delta$  также становится маленьким и обеспечивает хорошую гарантированную точность оценивания собственных значений.

В-четвёртых, предварительное приведение симметричной матрицы к хессенберговой форме, описанное в § 3.17ж, даёт симметричную трёхдиагональную матрицу. Как следствие, один шаг симметричного QR-алгоритма в применении к такой матрице выполняется всего за  $b_n$  арифметических операций. И хотя в целом трудоёмкость симметричного QR-алгоритма оценивается как  $O(n^2)$ , т. е. так же, как у исходного QR-алгоритма, за «О-большим» стоит теперь существенно меньшая константа. Трёхдиагональность матрицы упрощает также критерий остановки, описанный в предыдущем пункте.

В качестве сдвигов в симметричном QR-алгоритме можно использовать последний элемент текущей матрицы, т. е.  $a_{nn}^{(k)}$ . В [13, 45] показано, что он почти всегда обеспечивает *кубическую сходимость* к наименьшему по модулю собственному значению матрицы: если  $\varrho_k$  — приближение к собственному значению  $\lambda$ , получаемое на  $k$ -м шаге алгоритма, то

$$|\varrho_k - \lambda| \leq C |\varrho_{k-1} - \lambda|^3, \quad k = 1, 2, \dots,$$

с некоторой константой  $C$ . Это очень быстрая сходимость, при которой количество верных значащих цифр приближения к собственному значению утраивается после каждого шага алгоритма.

Но существуют также примеры расходимости симметричного QR-алгоритма с таким выбором сдвига [96]. Более изощрённая процедура сдвига, которая обеспечивает глобальную сходимость, состоит в том, что величиной сдвига берётся то собственное значение подматрицы

$$\begin{pmatrix} a_{n-1,n-1}^{(k)} & a_{n-1,n}^{(k)} \\ a_{n,n-1}^{(k)} & a_{n,n}^{(k)} \end{pmatrix},$$

которое ближе к  $a_{nn}^{(k)}$ . Такой сдвиг называется *сдвигом Уилкинсона* и он также обеспечивает кубическую сходимость почти во всех случаях. Кроме него иногда применяются и другие рецепты, описание которых читатель может найти в книге [96].

**Пример 3.19.1** (пример Бодевига) Рассмотрим решение симметричной проблемы собственных значений с помощью QR-алгоритма для вещественной матрицы

$$\begin{pmatrix} 2 & 1 & 3 & 4 \\ 1 & -3 & 1 & 5 \\ 3 & 1 & 6 & -2 \\ 4 & 5 & -2 & -1 \end{pmatrix}.$$

В справочнике [114] её собственные значения даны в следующем виде:

$$\begin{aligned}\lambda_1 &= -8.02857835, \\ \lambda_2 &= 7.93290471, \\ \lambda_3 &= 5.66886437, \\ \lambda_4 &= -1.57319073.\end{aligned}$$

После приведения к хессенберговой форме получаем трёхдиагональную матрицу

$$\begin{pmatrix} 2.0 & -5.09901951 & 0 & 0 \\ -5.09901951 & 1.26923077 & -4.40330027 & 0 \\ 0 & -4.40330027 & -4.30821757 & -3.29080690 \\ 0 & 0 & -3.29080690 & 5.03898680 \end{pmatrix}.$$

Базовый QR-алгоритм для этой матрицы (как и для исходной) сходит медленно из-за наличия близких по модулю собственных чисел  $\lambda_1$  и  $\lambda_2$ . В частности, для достижения элементом на месте (3,2) значения, меньшего  $10^{-15}$ , требуется более сотни итераций. Ситуацию не улучшает принципиально переход к итерированию с подматрицами меньших размеров, которые можно выделить из исходной матрицы после оформления блочной структуры.

Организуем теперь сдвиги на последний элемент текущей матрицы, сопровождаемый переходом алгоритма к подматрицам, которые выделяются по ходу его работы.

После 4 шагов работы такого алгоритма в применении к исходной матрице получаем собственное значение  $\lambda_3$  с 14 установленными значащими цифрами, из которых первые девять совпадают с данными справочника [114].

Выделяем из матрицы ведущую  $3 \times 3$ -матрицу и снова запускаем симметричный QR-алгоритм со сдвигами на последний элемент. За следующие 3 итерации выделяется собственное значение  $\lambda_2$  и оно равно 7.932904717870015. Видно, что последний 9-й знак собственного значения из справочника [114] не вполне точен с учётом следующих за ним цифр.

Ещё 3 итерации требуются на выделение  $\lambda_4 = -1.573190738303504$ , причём последний знак значения из справочника тоже оказывается неверным. Одновременно из полученной  $2 \times 2$ -матрицы находится и  $\lambda_1$ , причём все значащие цифры из справочника [114] подтверждаются.

Как видим, сходимость симметричного QR-алгоритма со сдвигами на последний элемент и выделением подматриц в самом деле чрезвычайно быстрая, что согласуется с теоретическими результатами. ■

Собственные векторы матрицы можно найти как произведение всех ортогональных матриц-сомножителей  $Q^{(k)}$ , получающихся в процессе работы симметричного QR-алгоритма. В самом деле, из равенства (3.227) следует

$$\begin{aligned} A^{(k)} &= (Q^{(k)})^\top \cdots (Q^{(1)})^\top A Q^{(1)} \cdots Q^{(k)} = \\ &= (Q^{(1)} \cdots Q^{(k)})^\top A Q^{(1)} \cdots Q^{(k)}. \end{aligned}$$

Переходя в этом соотношении к пределу по  $k \rightarrow \infty$ , найдём

$$D = \tilde{Q}^\top A \tilde{Q},$$

где  $\tilde{Q} = \lim_k (Q^{(1)} \cdots Q^{(k)})$  и  $D$  — диагональная матрица с собственными значениями  $A$  по диагонали. Умножая обе части полученного равенства слева на  $\tilde{Q}$  и пользуясь ортогональностью этой матрицы, будем иметь

$$A \tilde{Q} = \tilde{Q} D.$$

Это означает, что столбцы из  $\tilde{Q}$  являются правыми собственными векторами матрицы  $A$ .

### 3.196 Метод Якоби для симметричной проблемы собственных значений

В этом параграфе рассмотрим численный метод для нахождения собственных чисел и собственных векторов симметричных плотно заполненных матриц. Он был впервые применён К.Г. Якоби в 1846 году к

конкретной  $7 \times 7$ -матрице,<sup>43</sup> а затем оказался забыт на целое столетие. Его переоткрытие состоялось лишь после Второй мировой войны после начала бурного развития вычислительной математики.

## Теоретические основы

Идея метода Якоби состоит в том, чтобы подходящими преобразованиями подобия от шага к шагу уменьшать норму внедиагональной части матрицы. Получающиеся при этом матрицы имеют тот же спектр, что и исходная матрица, но будут стремиться к диагональной матрице с собственными значениями на главной диагонали. Инструментом реализации этого плана выступают элементарные ортогональные матрицы вращений, рассмотренные в § 3.7г. Почему именно ортогональные матрицы и почему вращений? Ответ на первый вопрос заключается в том, что только ортогональные преобразования подобия являются одновременно преобразованиями конгруэнции, т. е. оставляют матрицу симметричной. Иначе ценное свойство симметричности может быть потеряно при других преобразованиях подобия. Ответ на второй вопрос станет ясен позднее, после детального исследования подобий с помощью вращений.

Положим  $A^{(0)} := A$ . Если уже известна матрица  $A^{(k-1)} = (a_{ij}^{(k-1)})$ ,  $k = 1, 2, \dots$ , то матрица  $A^{(k)} = (a_{ij}^{(k)})$  вычисляется как результат преобразования подобия для  $A^{(k-1)}$ ,

$$A^{(k)} := G^\top A^{(k-1)} G. \quad (3.230)$$

Оно выполняется с помощью такой матрицы вращений  $G$  вида (3.108), что для заранее выбранных неравных целых чисел  $p$  и  $q$  внедиагональные элементы в позициях  $(p, q)$  и  $(q, p)$  матрицы  $A^{(k)}$  становятся нулевыми.

Из формул умножения на матрицу вращений слева (3.110) и им аналогичных для умножения справа следует, что матрица вращений  $G$  должна иметь синусы и косинусы в позициях  $(p, p)$ ,  $(q, q)$ ,  $(p, q)$  и  $(q, p)$ , т. е. в обозначениях (3.108) быть матрицей  $G(p, q, \theta)$  для некоторого угла  $\theta$ . При этом удобно представить результаты преобразования подобия (3.230) для интересующих нас четырёх элементов в виде операций с  $2 \times 2$ -матрицами.

---

<sup>43</sup>Любопытно, что тогда ещё не существовало самого термина «матрица».

Желая занулить элементы в позициях  $(p, q)$  и  $(q, p)$ , мы должны добиться выполнения равенства

$$\begin{pmatrix} a_{pp}^{(k)} & a_{pq}^{(k)} \\ a_{qp}^{(k)} & a_{qq}^{(k)} \end{pmatrix} = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}^\top \begin{pmatrix} a_{pp}^{(k-1)} & a_{pq}^{(k-1)} \\ a_{qp}^{(k-1)} & a_{qq}^{(k-1)} \end{pmatrix} \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} = \\ = \begin{pmatrix} \times & 0 \\ 0 & \times \end{pmatrix},$$

где посредством « $\times$ » обозначены какие-то элементы, конкретное значение которых несущественно. Строго говоря, в результате рассматриваемого преобразования подобия в матрице  $A^{(k)}$  изменятся и другие элементы, находящиеся в строках и столбцах с номерами  $p$  и  $q$ . Но этот эффект не препятствует общей сходимости метода, и ниже мы проанализируем его в предложении 3.19.1.

Опуская индексы, обозначающие номер итерации, примем сокращённые обозначения  $c = \cos \theta$ ,  $s = \sin \theta$ . Получим тогда

$$\begin{pmatrix} \times & 0 \\ 0 & \times \end{pmatrix} = \\ = \begin{pmatrix} a_{pp}c^2 + a_{qq}s^2 + 2sca_{pq} & sc(a_{qq} - a_{pp}) + a_{pq}(c^2 - s^2) \\ sc(a_{qq} - a_{pp}) + a_{pq}(c^2 - s^2) & a_{pp}s^2 + a_{qq}c^2 - 2sca_{pq} \end{pmatrix}.$$

Приравнивание внедиагональных элементов нулю даёт

$$\frac{a_{pp} - a_{qq}}{a_{pq}} = \frac{c^2 - s^2}{sc}.$$

Поделив обе части этой пропорции пополам, воспользуемся тригонометрическими формулами двойных углов:

$$\frac{a_{pp} - a_{qq}}{2a_{pq}} = \frac{c^2 - s^2}{2sc} = \frac{\cos(2\theta)}{\sin(2\theta)} = \frac{1}{\operatorname{tg}(2\theta)}.$$

Результат проведённой выкладки для удобства дальнейшего использования обозначим через  $\tau$ , т. е.  $\tau := 1/\operatorname{tg}(2\theta)$ .

Пусть

$$t := \frac{\sin \theta}{\cos \theta} = \operatorname{tg} \theta.$$

Вспоминая тригонометрическую формулу для тангенса двойного угла,

$$\operatorname{tg}(2\theta) = \frac{2 \operatorname{tg} \theta}{1 - \operatorname{tg}^2 \theta},$$

мы можем прийти к выводу, что  $t$  является решением квадратного уравнения

$$t^2 + 2\tau t - 1 = 0. \quad (3.231)$$

Его дискриминант  $(4\tau^2 + 4)$  положителен и, следовательно, уравнение (3.231) всегда имеет вещественные решения

$$t_{1,2} = -\tau \pm \sqrt{\tau^2 + 1}.$$

При этом из двух решений мы берём наименьшее по абсолютной величине, равное

$$\begin{aligned} t &= -\tau + \operatorname{sgn} \tau \cdot \sqrt{\tau^2 + 1}, && \text{если } \tau \neq 0, \\ t &= \pm 1, && \text{если } \tau = 0. \end{aligned}$$

Первую формулу для улучшения численной устойчивости лучше переписать в виде, освобождённом от вычитания близких чисел (вспомним эффект потери точности из § 1.3). Для этого умножим выражение в правой части на дробь с одинаковыми числителем и знаменателем  $(\tau + \operatorname{sgn} \tau \cdot \sqrt{\tau^2 + 1})$ , получив

$$\begin{aligned} t &= \frac{(-\tau + \operatorname{sgn} \tau \cdot \sqrt{\tau^2 + 1})(\tau + \operatorname{sgn} \tau \cdot \sqrt{\tau^2 + 1})}{(\tau + \operatorname{sgn} \tau \cdot \sqrt{\tau^2 + 1})} = \\ &= \frac{1}{(\tau + \operatorname{sgn} \tau \cdot \sqrt{\tau^2 + 1})} \quad \text{при } \tau \neq 0. \end{aligned}$$

Выбор меньшего по модулю решения уравнения (3.231) вызван тем, что он не приводит к перемене местами диагональных элементов в матрице [96]. На заключительном этапе работы метода Якоби подобное явление крайне нежелательно, поскольку мешает отслеживать сходимость к решению.

Наконец, по известному  $t$  находим  $c$  и  $s$  с помощью тригонометрических формул, выраждающих косинус и синус через тангенс:

$$c = \frac{1}{\sqrt{t^2 + 1}}, \quad s = t \cdot c.$$

Этим завершается построение матрицы вращений  $G(p, q, \theta)$ , необходимой для выполнения одного шага метода Якоби для решения симметричной проблемы собственных значений.

Займёмся теперь обоснованием сходимости метода Якоби. Для количественного описания близости матриц, которые порождаются методом Якоби, к диагональной матрице, введём величину

$$ND(A) = \left( \sum_{j \neq i} a_{ij}^2 \right)^{1/2}$$

— фробениусову норму внедиагональной части матрицы  $A = (a_{ij})$ . Ясно, что матрица  $A$  диагональна тогда и только тогда, когда  $ND(A) = 0$ .

**Предложение 3.19.1** Пусть преобразование подобия матрицы  $A = (a_{ij})$  с помощью матрицы вращений  $G$  таково, что в полученной матрице  $B = G^\top AG$  зануляются элементы в позициях  $(p, q)$  и  $(q, p)$ . Тогда

$$ND^2(B) = ND^2(A) - 2a_{pq}^2. \quad (3.232)$$

Итак, в сравнении с матрицей  $A$  в матрице  $B$  изменяются элементы строк и столбцов с номерами  $p$  и  $q$ , но фробениусова норма недиагональной части изменяется при этом так, как будто кроме зануления элементов  $a_{pq}$  и  $a_{qp}$  ничего не происходит.

**Доказательство.**  $2 \times 2$ -подматрица

$$\begin{pmatrix} a_{pp} & a_{pq} \\ a_{qp} & a_{qq} \end{pmatrix}$$

из матрицы  $A$  и соответствующая ей  $2 \times 2$ -подматрица

$$\begin{pmatrix} b_{pp} & 0 \\ 0 & b_{qq} \end{pmatrix}$$

в матрице  $B$  являются ортогонально подобными. По этой причине

$$a_{pp}^2 + a_{qq}^2 + 2a_{pq}^2 = b_{pp}^2 + b_{qq}^2,$$

так как от умножения на ортогональные матрицы фробениусова норма не изменяется (предложение 3.3.4 и следствие из него, стр. 385). Но это же верно для самих матриц  $A$  и  $B$ , т. е.  $\|A\|_F^2 = \|B\|_F^2$ , и потому

$$\begin{aligned} ND^2(B) &= \|B\|_F^2 - \sum_{i=1}^n b_{ii}^2 = \\ &= \|A\|_F^2 - \left( \sum_{i=1}^n a_{ii}^2 - (a_{pp}^2 + a_{qq}^2) + (b_{pp}^2 + b_{qq}^2) \right) = \\ &= ND^2(A) - 2a_{pq}^2, \end{aligned}$$

поскольку на диагонали у матрицы  $A$  изменились только два элемента —  $a_{pp}$  и  $a_{qq}$ . ■

Теперь можно дополнить ответ на вопрос о том, почему в методе Якоби применяются именно ортогональные матрицы вращения. Как следует из предложения 3.19.1, специальными преобразованиями подобия с матрицами вращения суммарная величина внедиагональных элементов матрицы «перекачивается» на диагональ, уменьшая внедиагональную часть. Можно ли сделать что-то похожее с помощью других преобразований подобия — не вполне ясно.

### Алгоритм и его сходимость

В табл. 3.18 схематично представлен простейший вариант метода вращений Якоби — итерационного процесса приведения симметричной матрицы к диагональному виду, при котором внедиагональные элементы последовательно подавляются преобразованиями подобия с матрицами вращения. Занулённые на каком-то шаге алгоритма элементы, как отмечалось, могут впоследствии вновь сделаться ненулевыми. Но результат предложения 3.19.1 показывает, что норма внедиагональной части матрицы при этом всё равно уменьшается.

Рассмотрим вопрос о критерии остановки метода Якоби, т. е. о том, когда матрица становится «достаточно близкой к диагональной». Он тесно связан с оценкой точности получающихся приближений к собственным значениям. Очевидной идеей является использование нормы  $ND$  внедиагональной части матрицы, когда итерации останавливаются при достижении неравенства  $ND(A) < \epsilon$  для заданного допуска  $\epsilon > 0$ . Тогда оценки близости диагональных элементов матрицы

Таблица 3.18. Метод Якоби для вычисления собственных значений симметричной матрицы

<b>Вход</b>
Симметричная матрица $A = (a_{ij})$ .
<b>Выход</b>
Матрица, на диагонали которой стоят приближения к собственным значениям $A$ .
<b>Алгоритм</b>
<pre> DO WHILE ( <math>A</math> недостаточно близка к диагональной )     выбрать ненулевой внедиагональный     элемент <math>a_{pq}</math> в <math>A</math> ;     обнулить элементы <math>a_{pq}</math> и <math>a_{qp}</math> преобразованием     подобия с матрицей вращения <math>G(p, q, \theta)</math>,     построение которой описано в начале раздела END DO </pre>

к искомым собственным значениям даются с помощью теорем Вейля или Виландта–Хоффмана (теоремы 3.17.3 и 3.17.4). Но более удобный и более точный критерий остановки может быть основан на теореме Гершгорина аналогично тому, как сделано для симметричного QR-алгоритма в § 3.19а.

Различные способы выбора внедиагональных элементов, подлежащих обнулению, приводят к различным практическим версиям метода Якоби. Выбор наибольшего по модулю внедиагонального элемента — наилучшее для отдельно взятого шага алгоритма решение. Но поиск такого элемента требует сравнения  $n(n - 1)/2$  элементов матрицы, лежащих выше (или ниже) главной диагонали. Это может оказаться относительно дорогостоящим, особенно для матриц больших размеров. Преобразование подобия с матрицей вращений обходится всего в  $O(n)$  арифметических операций! Часто применяют циклический обход столбцов (или строк) матрицы, и наибольший по модулю элемент берут

в пределах рассматриваемого столбца (строки).

Наконец, ещё одна популярная версия — это так называемый «барьерный метод Якоби», в котором назначают величину «барьера» на значение модуля внедиагональных элементов матрицы, и алгоритм обнуляет все элементы, модуль которых превосходит этот барьер. Затем барьер понижается, процесс обнуления повторяется заново, и так до тех пор, пока не будет достигнута требуемая точность.

Покажем сходимость метода Якоби при выборе очередного зануляемого элемента как наибольшего по модулю внедиагонального элемента матрицы. Пусть в матрице  $A^{(k)}$ , полученной в результате  $k$  шагов метода Якоби, наибольшим по модулю внедиагональным элементом является  $a_{i_k j_k}^{(k)}$ . В результате  $k + 1$ -го шага вычисляется матрица  $A^{(k+1)}$ , для которой согласно предложению 3.19.1

$$ND^2(A^{(k+1)}) = ND^2(A^{(k)}) - 2(a_{i_k j_k}^{(k)})^2.$$

Очевидно также, что

$$ND^2(A^{(k)}) \leq n(n-1)(a_{i_k j_k}^{(k)})^2,$$

откуда следует

$$(a_{i_k j_k}^{(k)})^2 \geq \frac{1}{n(n-1)} ND^2(A^{(k)}).$$

Можем заключить

$$ND^2(A^{(k+1)}) \leq ND^2(A^{(k)}) \left(1 - \frac{2}{n(n-1)}\right),$$

и потому в целом

$$ND^2(A^{(k)}) \leq ND^2(A) \left(1 - \frac{2}{n(n-1)}\right)^k, \quad k = 0, 1, 2, \dots,$$

так что метод Якоби сходится не медленнее, чем со скоростью геометрической прогрессии. Более детальный анализ показывает, что после того как внедиагональные элементы матрицы достаточно уменьшаются, метод Якоби приобретает квадратичную сходимость [13, 48, 96].

Собственные векторы матрицы в методе Якоби получаются как столбцы произведения всех ортогональных матриц  $G$ , с помощью которых выполняются преобразования (3.230) на каждом отдельном шаге

метода. В самом деле, если обозначить верхним индексом у матрицы вращения  $G$  номер шага, на котором она используется, то из (3.230) следует, что

$$\begin{aligned} A^{(1)} &= (G^{(1)})^\top AG^{(1)}, \\ A^{(2)} &= (G^{(2)})^\top (G^{(1)})^\top AG^{(1)}G^{(2)}, \\ &\dots \end{aligned}$$

Пусть  $\tilde{G} = G^{(1)}G^{(2)}\dots$  — произведение всех  $G^{(k)}$ , тогда по итогу работы метода Якоби

$$(\tilde{G})^\top A\tilde{G} = D,$$

где  $D$  диагональна. В силу ортогональности  $\tilde{G}$  это равносильно

$$A\tilde{G} = \tilde{G}D.$$

Слева в полученном равенстве матрица  $A$  умножается на столбцы из  $\tilde{G}$ , а справа эти же самые столбцы умножаются на элементы диагонали  $D$ , равные собственным значениям матрицы. Это и означает, что столбцы  $\tilde{G}$  суть собственные векторы  $A$ .

К 70-м годам прошлого века, когда было разработано немало эффективных численных методов для решения симметричной проблемы собственных значений, стало казаться, что метод Якоби устарел и будет вытеснен из широкой вычислительной практики (см. рассуждения в [96]). Дальнейшее развитие не подтвердило эти пессимистичные прогнозы. Выяснилось, что метод Якоби почти не имеет конкурентов по точности нахождения малых собственных значений, тогда как в методах, основанных на предварительной трёхдиагонализации исходной матрицы (§ 3.17ж, 3.19а и далее § 3.19г), точность малых собственных значений может заметно ухудшаться. Соответствующие примеры читатель может увидеть в [13]. Кроме того, метод Якоби оказался хорошо распараллеливаемым, т. е. подходящим для расчётов на современных многопроцессорных ЭВМ [11].

### 3.19в Итерации с отношением Рэлея

Для решения симметричной проблемы собственных значений можно использовать все те методы, которые были разработаны для общего несимметричного случая. При этом учёт специальной структуры матрицы часто позволяет достигать качественно большего эффекта при их

применении или реализовывать такие модификации, которые в общем случае невозможны.

Выше в § 3.18б рассматривались обратные степенные итерации. Модификации, изложенные затем в § 3.18в, превращают обратные степенные итерации в простой и очень эффективный подход к решению проблем собственных значений. Их дальнейшее естественное развитие может состоять в том, чтобы сделать сдвиги динамическими, т. е. изменяющимися от шага к шагу и использующими информацию об уточнении собственных значений и собственных векторов. Хорошей идеей является, в частности, определение величины сдвигов из отношения Рэлея

$$\mathcal{R}(x) = \frac{\langle Ax, x \rangle}{\langle x, x \rangle} = \frac{x^\top Ax}{x^\top x}$$

(см. подробности в § 3.17е). Если  $x$  — нормированный в евклидовой норме вектор, то это выражение принимает более простой вид

$$\mathcal{R}(x) = \langle Ax, x \rangle = x^\top Ax,$$

который и используется в алгоритме ниже.

Псевдокод обратных степенных итераций с отношением Рэлея, которые называют также просто *итерациями с отношением Рэлея*, приведён в табл. 3.19. В нём числа  $\varrho_k$  — это сдвиги исходной матрицы, уточняемые по мере работы алгоритма, которые также дают приближения к собственным значениям. Векторы  $x^{(k)}$  являются приближениями к нормированным собственным векторам. Условие остановки алгоритма по величине нормы  $\|y^{(k)}\|$  (строка 6 псевдокода) указано для того, чтобы завершить исполнение при возможной расходимости, когда векторы  $y^{(k)}$  сильно удаляются от начала координат. Критерием остановки алгоритма в цикле **DO WHILE** можно взять, например, удовлетворение неравенству

$$\|Ax^{(k)} - \varrho_k x^{(k)}\| \leq \epsilon$$

для какого-то заданного порога  $\epsilon$ .

При создании серьёзной программной реализации итераций с отношением Рэлея следует дополнить алгоритм из табл. 3.19 различными модификациями, которые учитывают как особенности машинных вычислений с плавающей точкой, так и исключительные случаи. В частности, нужно предусмотреть возможность обработки ситуаций, когда на каком-то шаге алгоритма число  $\varrho_k$  сделается очень близким к собственному значению. Хотя критерий остановки ещё не будет выполнен,

Таблица 3.19. Итерации с отношением Рэлея для решения симметричной проблемы собственных значений

```

 $k \leftarrow 0; \quad \varrho_0 \leftarrow 0;$ 
выбрать вектор  $x^{(0)}$ , такой что  $\|x^{(0)}\|_2 = 1$ ;
DO WHILE ( метод не сорвался )
     $k \leftarrow k + 1;$ 
    найти  $y^{(k)}$  из системы  $(A - \varrho_{k-1}I) y^{(k)} = x^{(k-1)}$ ;
    IF (  $\|y^{(k)}\|$  велика ) THEN
        STOP
    END IF
    выполнить нормировку  $x^{(k)} \leftarrow y^{(k)} / \|y^{(k)}\|_2$ ;
     $\varrho_k \leftarrow (x^{(k)})^\top A x^{(k)}$ ;
END DO

```

обусловленность системы линейных уравнений с матрицей  $A - \varrho_{k-1}I$  в строке 5 псевдокода может сделаться непомерно большой для устойчивого и надёжного решения.

Для обозначения алгоритма из табл. 3.19 часто используют термин «RQ-итерации» от английской фразы «Rayleigh quotient», означающей «отношение Рэлея» [13], хотя он весьма двусмыслен и может приводить к путанице с QR-алгоритмом.

**Пример 3.19.2** Рассмотрим симметричную матрицу

$$\begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix},$$

которая имеет собственные значения  $2 \pm \sqrt{5}$ .

Из начального вектора  $(1/\sqrt{2}, 1/\sqrt{2})^\top$  итерации с отношением Рэлея сходятся к большему по модулю собственному значению, и уже через 3 (три) итерации выдают верными его 10 значащих цифр!

Из начального вектора  $(0, 1)^\top$  итерации с отношением Рэлея с тем

же успехом сходятся к собственному значению  $2 - \sqrt{5}$ , меньшему по модулю. ■

Поведение итераций с отношением Рэлея хорошо исследовано теоретически. В частности, в книге [96] показывается, что евклидова норма невязки  $\|Ax^{(k)} - \varrho_k x^{(k)}\|_2$  монотонно убывает вне зависимости от выбора начального вектора  $x^{(0)}$ . Как следствие этого факта в [96] обосновано, что итерации с отношениями Рэлея всегда сходятся к собственному значению матрицы, а собственные векторы сходятся почти всегда, т. е. за исключением пренебрежимого множества начальных векторов. Далее, если начальный вектор  $x^{(0)}$  выбран достаточно близким к собственному вектору матрицы, а соответствующее собственное значение  $\lambda$  — простое, то итерации с отношениями Рэлея имеют кубическую скорость сходимости [13, 96]. Мы видели этот эффект в численном примере выше.

В принципе, отношение Рэлея можно также использовать для организации сдвигов в прямых степенных итерациях, но этот способ не имеет той общности и гибкости, как в обратных степенных итерациях.

### 3.19г Трёхдиагональные матрицы

В § 3.17ж рассматривалось предварительное упрощение матриц общего вида перед численным решением проблемы собственных значений. Напомним, что для симметричной проблемы собственных значений удобной упрощённой формой матрицы можно считать симметричную трёхдиагональную, к которой произвольная симметричная матрица приводится ортогональными преобразованиями подобия. Для задачи вычисления сингулярных чисел и сингулярных векторов упрощённой формой, к которой обычно приводят матрицы общего вида, являются двухдиагональные матрицы, и из них далее также получаются трёхдиагональные (см. § 3.19д).

Простая структура трёхдиагональных матриц имеет следствием существование несложных алгоритмов для вычисления их собственных чисел и собственных векторов. Один из наиболее эффективных подходов — применение QR-алгоритма. Специальная структура трёхдиагональных матриц сохраняется в процессе работы этого алгоритма, а её аккуратный учёт позволяет существенно экономить вычисления.

Опишем здесь специализированные методы для трёхдиагональных матриц, основанные на других идеях.

Если  $A$  — трёхдиагональная матрица, то тогда и  $A - \lambda I$  тоже является трёхдиагональной. Обозначим посредством  $D_m(\lambda)$  ведущий минор  $m$ -го порядка матрицы  $A - \lambda I$ , а также соответствующую ведущую подматрицу. Выделим в ней две последние строки и два последних столбца:

$$D_m(\lambda) = \left( \begin{array}{c|cc|c} D_{m-2}(\lambda) & 0 & 0 \\ & \vdots & \vdots \\ & 0 & 0 \\ & a_{m-2,m-1} & 0 \\ \hline 0 & \dots & 0 & a_{m-1,m-2} & a_{m-1,m-1} - \lambda & a_{m-1,m} \\ 0 & \dots & 0 & 0 & a_{m,m-1} & a_{m,m} - \lambda \end{array} \right).$$

Если разложить  $D_m(\lambda)$  по последней  $m$ -й строке (см. (3.8)), то, учитывая, что в ней находятся всего два ненулевых элемента, получаем

$$D_m(\lambda) = (a_{mm} - \lambda)D_{m-1}(\lambda) - a_{m,m-1}B_{m,m-1}(\lambda), \quad (3.233)$$

где  $B_{m,m-1}(\lambda)$  — минор, дополняющий элемент  $a_{m,m-1}$ . Этот минор содержит в последнем столбце только один ненулевой элемент  $a_{m-1,m}$ , так что удобно дальше разложить  $B_{m,m-1}(\lambda)$  по этому последнему столбцу:

$$B_{m,m-1}(\lambda) = a_{m-1,m}D_{m-2}(\lambda).$$

Подставляя полученное равенство в (3.233), получим рекуррентное соотношение, которое выражает ведущий минор трёхдиагональной матрицы через миноры меньшего порядка:

$$D_m(\lambda) = (a_{mm} - \lambda)D_{m-1}(\lambda) - a_{m,m-1}a_{m-1,m}D_{m-2}(\lambda), \quad (3.234)$$

$$m = 2, 3, \dots, n.$$

Для начала расчётов по этой рекуррентной формуле нужно задать два первых минора, и удобно положить

$$D_0(\lambda) = 1, \quad D_1(\lambda) = a_{11} - \lambda,$$

введя фиктивный минор нулевого порядка  $D_0(\lambda)$ . Так как сама матрица  $A - \lambda I$  является своей ведущей  $n \times n$ -подматрицей, то в результате выполнения рекуррентных инструкций (3.234) мы на  $n$ -ом шаге получим выражение для  $\det(A - \lambda I)$ , т. е. характеристический полином исходной матрицы  $A$ . Вычислив его нули, найдём собственные значения матрицы.

Практическая реализация описанного выше рецепта разумна лишь для матриц  $A$  малого порядка, так как у полиномов высоких степеней нули очень чувствительны к неточностям в задании коэффициентов и погрешностям вычислений. Если порядок матрицы  $A$  значителен, то формулы (3.234) более выгодно использовать не для получения явного аналитического выражения характеристического полинома от переменной  $\lambda$ , а для вычисления его значений в конкретных точках.

Нахождение численного значения определителя  $\det(A - \lambda I)$  для фиксированного значения  $\lambda$  по формулам (3.234) требует  $5n$  арифметических операций, причём среди них нет делений. Таким образом, получен быстрый и устойчивый способ вычисления значений характеристического полинома. Далее на него можно опираться при нахождении нулей этого полинома, т. е. собственных значений матрицы, с помощью каких-либо численных методов для решения уравнений, описанных далее в главе 4. Например, неплохо могут подойти для этой цели метод половинного деления (бисекции), метод секущих и метод парабол (метод Мюллера), изложению которых посвящены § 4.4б и § 4.4г соответственно.

Недостаток этого подхода в самом общем случае может состоять в том, что многие численные методы находят только вещественные нули характеристического полинома, т. е. только вещественные собственные значения матрицы. Тогда особенно выгодно применение метода парабол (метода Мюллера), который способен находить даже комплексные нули вещественных полиномов. Но для симметричных вещественных матриц описанная проблема не возникает, так как их собственные значения всегда вещественны. Отметим, что выше нигде явно не использовалась симметричность исходной матрицы, и потому вся описанная методика применима к общим трёхдиагональным матрицам.

Аналогичной по духу является техника уточнения вещественных собственных значений трёхдиагональной матрицы, основанная на применении известной из алгебры теоремы Штурма о количестве нулей полиномов [23, 27, 47]. Ценой дополнительных трудозатрат она позволяет получать гораздо больше информации о собственных значениях, и подробности читатель может увидеть в книгах [45, 79]. Кроме того, в книге [79] подробно описана машинная реализация метода последовательностей Штурма, аккуратно учитывающая погрешности вычислений в арифметике с плавающей точкой.

### 3.19д Численные методы сингулярного разложения

Сингулярные числа зависят от элементов матрицы существенно более плавным образом, нежели собственные числа. Это непосредственно следует из теоремы Бауэра–Файка (теорема 3.17.2), теорем Вейля (теорема 3.17.3) и Виландта–Хоффмана (теорема 3.17.4), применённых к симметричным матрицам  $A^*A$  и  $AA^*$ . В частности, как из теоремы Бауэра–Файка, так и из теоремы Вейля, вытекает

**Предложение 3.19.2** Пусть  $A$  и  $B$  – произвольные матрицы одинакового размера, вещественные или комплексные, причём  $\sigma_1 \geq \sigma_2 \geq \dots$  – сингулярные числа матрицы  $A$ , а  $\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq \dots$  – сингулярные числа матрицы  $\tilde{A} = A + B$ . Тогда  $|\tilde{\sigma}_i - \sigma_i| \leq \|B\|_2$  для всех  $i$ .

Таким образом, одно из важнейших отличий сингулярных чисел матрицы от её собственных чисел состоит в том, что собственные числа могут изменяться в зависимости от элементов матрицы сколь угодно быстро (см. пример 3.17.3), тогда как изменение сингулярных чисел соправмерно величине возмущений матрицы.

Ограничимся далее случаем вещественных матриц, так как на комплексный случай всё сказанное ниже распространяется очевидным образом.

Вычисление сингулярного разложения матрицы  $A$ , как показано в § 3.2ж, может быть сведено к симметричной проблеме собственных значений для какой-либо из матриц

$$A^\top A, \quad AA^\top, \quad \begin{pmatrix} 0 & A^\top \\ A & 0 \end{pmatrix}. \quad (3.235)$$

На этом факте могут быть основаны простейшие методы нахождения сингулярных чисел и сингулярных векторов матриц. Недостаток первых двух матриц из (3.235) – возможная потеря точности при умножении, а также возвведение в квадрат числа обусловленности исходной матрицы. Недостаток последней матрицы из (3.235) (существенно более мягкий) – двойной размер.

Если матрица предварительно приведена к двухдиагональному виду

ду (см. § 3.17ж), то задача существенно упрощается. Более точно, если

$$B = \begin{pmatrix} \alpha_1 & \beta_1 & & & 0 \\ & \alpha_2 & \beta_2 & & \\ & & \ddots & \ddots & \\ 0 & & & \alpha_{n-1} & \beta_{n-1} \\ & & & & \alpha_n \end{pmatrix}$$

— двухдиагональная матрица, то, как нетрудно проверить,  $B^\top B$  и  $BB^\top$  являются симметричными трёхдиагональными матрицами:

$$B^\top B = \begin{pmatrix} \alpha_1^2 & \alpha_1\beta_1 & & & 0 \\ \alpha_1\beta_1 & \alpha_2^2 + \beta_1^2 & \alpha_2\beta_2 & & \\ & \ddots & \ddots & \ddots & \\ 0 & & \alpha_{n-2}\beta_{n-2} & \alpha_{n-1}^2 + \beta_{n-2}^2 & \alpha_{n-1}\beta_{n-1} \\ & & & \alpha_{n-1}\beta_{n-1} & \alpha_n^2 + \beta_{n-1}^2 \end{pmatrix}$$

и

$$BB^\top = \begin{pmatrix} \alpha_1^2 + \beta_1^2 & \alpha_2\beta_1 & & & 0 \\ \alpha_2\beta_1 & \alpha_2^2 + \beta_2^2 & \alpha_3\beta_2 & & \\ & \ddots & \ddots & \ddots & \\ 0 & & \alpha_{n-1}\beta_{n-2} & \alpha_{n-1}^2 + \beta_{n-1}^2 & \alpha_n\beta_{n-1} \\ & & & \alpha_n\beta_{n-1} & \alpha_n^2 \end{pmatrix}.$$

Матрица

$$\begin{pmatrix} 0 & B^\top \\ B & 0 \end{pmatrix} \quad (3.236)$$

трёхдиагональной, очевидно, не является, но путём перестановки её строк и столбцов она может быть сделана трёхдиагональной. Так как матрицы перестановки ортогональны, то сингулярные числа при таком преобразовании останутся неизменными.

Пусть  $e_j$  обозначает  $j$ -й столбец единичной матрицы, т. е.

$$e_j := (0, \dots, \underset{\leftarrow j\text{-е место}}{1}, \dots, 0)^\top.$$

Обозначим также

$$P := (e_1, e_{n+1}, e_2, e_{n+2}, \dots, e_n, e_{2n})$$

— матрицу, составленную из столбцов единичной матрицы, переупорядоченных так, как указано индексами столбцов. Очевидно, что это матрица перестановки (см. определение 3.6.1, стр. 454), и она замечательна тем, что порождаемое ею преобразование подобия переводит матрицу (3.236) в трёхдиагональную матрицу:

$$P^\top \begin{pmatrix} 0 & B^\top \\ B & 0 \end{pmatrix} P = \begin{pmatrix} 0 & \alpha_1 & & & & 0 \\ \alpha_1 & 0 & \beta_1 & & & \\ & \beta_1 & 0 & \alpha_2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \alpha_{n-1} & 0 & \beta_{n-1} \\ 0 & & & & \beta_{n-1} & 0 & \alpha_n \\ & & & & & \alpha_n & 0 \end{pmatrix}.$$

На главной диагонали этой матрицы стоят нули, а наддиагональными и поддиагональными элементами являются числа  $\alpha_1, \beta_1, \alpha_2, \beta_2, \dots, \alpha_{n-1}, \beta_{n-1}, \alpha_n$ .

К настоящему времени для решения симметричной проблемы собственных значений с трёхдиагональными матрицами разработаны эффективные численные методы. Это, в частности, симметричный QR-алгоритм (§ 3.19а), специализированный алгоритм из § 3.19г и ряд других. Они с успехом применяются также для нахождения сингулярного разложения матрицы.

Одним из основных инструментов для получения сингулярного разложения произвольных матриц является алгоритм Голуба–Кэхэна [11], предложенный в 1965 году. Он является, по существу, модификацией симметричного QR-алгоритма, которая применяется неявно к матрице  $A^\top A$ .

Наконец, стоит упомянуть метод Якоби для сингулярного разложения матрицы, детальное описание которого можно найти в [11, 71].

## Литература к главе 3

### Основная

- [1] Бахвалов Н.С., Жидков Н.П., Кобельков Г.М. Численные методы. – М.: «БИНОМ. Лаборатория знаний», 2003, а также другие издания этой книги.
- [2] Бахвалов Н.С., Корнев А.А., Чижонков Е.В. Численные методы. Решения задач и упражнения. – М.: Дрофа, 2008.
- [3] Беклемишев Д.В. Дополнительные главы линейной алгебры. – М.: Наука, 1983.
- [4] Березин И.С., Жидков Н.П. Методы вычислений. Т. 1–2. – М.: Наука, 1966.
- [5] Вержбицкий В.М. Численные методы. Части 1–2. – М.: «Оникс 21 век», 2005.
- [6] Воеводин В.В. Линейная алгебра. – М.: Наука, 1980.
- [7] Воеводин В.В., Воеводин Вл.В. Энциклопедия линейной алгебры. Электронная система ЛИНЕАЛ. – СПб.: БХВ-Петербург, 2006.
- [8] Волков Е.А. Численные методы. – М.: Наука, 1987.
- [9] Гантмахер Ф.Р. Теория матриц. – М.: Наука, 1988.
- [10] Глазман И.М., Любич Ю.И. Конечномерный линейный анализ. – М.: Наука, 1969.
- [11] Голуб Дж., ван Лоун Ч. Матричные вычисления. – М.: Мир, 1999.
- [12] Демидович Б.П., Марон А.А. Основы вычислительной математики. – М.: Наука, 1970.
- [13] Деммель Дж. Вычислительная линейная алгебра. – М.: Мир, 2001.
- [14] Зорич В.А. Математический анализ. Т. 1. – М.: Наука, 1981. Т. 2. – М.: Наука, 1984, а также более поздние издания.
- [15] Икрамов Х.Д. Численные методы для симметричных линейных систем. – М.: Наука, 1988.
- [16] Икрамов Х.Д. Несимметричная проблема собственных значений. – М.: Наука, 1991.
- [17] Ильин В.П. Методы и технологии конечных элементов. – Новосибирск: Издательство ИВМиМГ СО РАН, 2007.
- [18] Ильин В.П., Кузнецов Ю.И. Трёхдиагональные матрицы и их приложения. – М.: Наука, 1985.
- [19] Канторович Л.В., Акилов Г.П. Функциональный анализ. – М.: Наука, 1984.
- [20] Като Т. Теория возмущений линейных операторов. – М.: Мир, 1972.
- [21] Коллатц Л. Функциональный анализ и вычислительная математика. – М.: Мир, 1969.
- [22] Коновалов А.Н. Введение в вычислительные методы линейной алгебры. – Новосибирск: Наука, 1993.
- [23] Кострикин А.Н. Введение в алгебру. Часть 1. Основы алгебры. – М.: Физматлит, 2001.
- [24] Кострикин А.Н. Введение в алгебру. Часть 2. Линейная алгебра. – М.: Физматлит, 2001.

- [25] Красносельский М.А., Крейн С.Г. Итеративный процесс с минимальными невязками // Математический Сборник. – 1952. – Т. 31 (73), №2. – С. 315–334.
- [26] Крылов В.И., Бовков В.В., Монастырный П.И. Вычислительные методы. Т. 1–2. – М.: Наука, 1976.
- [27] Курош А.Г. Курс высшей алгебры. – М.: Наука, 1975.
- [28] Ланкастер П. Теория матриц. – М.: Наука, 1978.
- [29] Лоусон Ч., Хенсон Р. Численное решение задач методом наименьших квадратов. – М.: Наука, 1986.
- [30] Мальцев А.И. Основы линейной алгебры. – М.: Наука, 1975.
- [31] Матрицы и квадратичные формы. Основные понятия. Терминология / Академия Наук СССР. Комитет научно-технической терминологии. – М.: Наука, 1990. – (Сборники научно-нормативной терминологии; Вып. 112).
- [32] Мацокин А.М. Численный анализ. Вычислительные методы линейной алгебры. Конспекты лекций для преподавания в III семестре ММФ НГУ. – Новосибирск: НГУ, 2009–2010.
- [33] Миньков С.Л., Миньков Л.Л. Основы численных методов. – Томск: Издательство научно-технической литературы, 2005.
- [34] Мысовских И.П. Лекции по методам вычислений. – СПб.: Издательство Санкт-Петербургского университета, 1998.
- [35] Орtega Дж. Введение в параллельные и векторные методы решения линейных систем. – М.: Мир, 1991.
- [36] Островский А.М. Решение уравнений и систем уравнений. – М.: Издательство иностранной литературы, 1963.
- [37] Прасолов В.В. Задачи и теоремы линейной алгебры. – М.: Наука-Физматлит, 1996.
- [38] Райс Дж. Матричные вычисления и математическое обеспечение. – М.: Мир, 1984.
- [39] Саад Ю. Итерационные методы для разреженных линейных систем. Учебное пособие в 2-х томах. – М.: Издательство Московского университета, 2013–2014.
- [40] Самарский А.А., Гулин А.В. Численные методы. – М.: Наука, 1989.
- [41] Стренг Г. Линейная алгебра и её применения. – М.: Мир, 1980.
- [42] Тихонов А.Н., Арсенин В.Я. Методы решения некорректных задач. – М.: Наука, 1979.
- [43] Тыртышников Е.Е. Матричный анализ и линейная алгебра. – М.: Физматлит, 2007.
- [44] Тыртышников Е.Е. Методы численного анализа. – М.: Академия, 2007.
- [45] Уилкинсон Дж. Алгебраическая проблема собственных значений. – М.: Наука, 1970.
- [46] Уоткинс Д. Основы матричных вычислений. – М.: «БИНОМ. Лаборатория знаний», 2009.

- [47] ФАДДЕЕВ Д.К. *Лекции по алгебре*. – М.: Наука, 1984.
- [48] ФАДДЕЕВ Д.К., ФАДДЕЕВА В.Н. *Вычислительные методы линейной алгебры*. – М.-Л.: Физматлит, 1960 (первое издание) и 1963 (второе издание).
- [49] ФЕДОРЕНКО Р.П. Итерационные методы решения разностных эллиптических уравнений // Успехи математических наук. – 1973. – Т. 28, вып. 2 (170). – С. 121–182.
- [50] ФОРСАЙТ Дж.Э. Что представляют собой релаксационные методы? // Современная математика для инженеров под ред. Э.Ф.Беккенбаха. – М.: Издательство иностранной литературы, 1958. – С. 418–440.
- [51] ФОРСАЙТ Дж., МОЛЕР К. Численное решение систем линейных алгебраических уравнений. – М.: Мир, 1969.
- [52] ХАУСДОРФ Ф. Теория множеств. – М.: УРСС Эдиториал, 2007.
- [53] ХЕЙГЕМАН Л., ЯНГ Д. Прикладные итерационные методы. – М.: Мир, 1986.
- [54] ХОРН Р., Джонсон Ч. Матричный анализ. – М.: Мир, 1989.
- [55] ШИЛОВ Г.Е. Математический анализ. Конечномерные линейные пространства. – М.: Наука, 1969.
- [56] ШИЛОВ Г.Е. Математический анализ. Функции одного переменного. Часть 3. – М.: Наука, 1970.
- [57] BECKERMANN B. The condition number of real Vandermonde, Krylov and positive definite Hankel matrices // Numerische Mathematik. – 2000. – Vol. 85, No. 4. – P. 553–577.
- [58] KELLEY C.T. Iterative methods for linear and nonlinear equations. – Philadelphia: SIAM, 1995.
- [59] Scilab — Open source software for numerical computation. <http://www.scilab.org>
- [60] TEMPLE G. The general theory of relaxation methods applied to linear systems // Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences. – 1939. – Vol. 169, No. 939. – P. 476–500.
- [61] TREFETHEN L.N., BAU D. III Numerical linear algebra. – Philadelphia: SIAM, 1997.

### **Дополнительная**

- [62] АЛЬБЕРГ Дж., Нильсон Э., Уолш Дж. Теория сплайнов и её приложения. – М.: Мир, 1972.
- [63] АЛЕКСАНДРОВ П.С. Введение в теорию множеств и общую топологию. – СПб.: Лань, 2010.
- [64] АЛЕКСЕЕВ В.Б. Теорема Абеля в задачах и решениях. – М.: Московский Центр непрерывного математического образования, 2001.
- [65] АЛЕКСЕЕВ Е.Р., Дога К.В., Чеснокова О.В. Scilab. Решение инженерных и математических задач. – М.: «ДМК Пресс», 2024.
- [66] АЛЕФЕЛЬД Г., ХЕРЦБЕРГЕР Ю. Введение в интервальные вычисления. – М.: Мир, 1987.

- [67] Бабенко К.И. Основы численного анализа. – М.: Наука, 1986; Ижевск-М.: Издательство «РХД», 2002.
- [68] Бардаков В.Г. Лекции по алгебре Ю.И. Мерзлякова. – Новосибирск: Издательство НГУ, 2012.
- [69] Быченков Ю.В., Чижонков Е.В. Итерационные методы решения седловых задач. – М.: «БИНОМ. Лаборатория знаний», 2012.
- [70] Виро О.Я., Иванов О.А., Неизвестен Н.Ю., Харламов В.М. Элементарная топология. – М.: Московский центр непрерывного математического образования, 2010 и 2012.
- [71] Воеводин В.В. Численные методы алгебры. Теория и алгорифмы. – М.: Наука, 1966.
- [72] Воеводин В.В. Вычислительные основы линейной алгебры. – М.: Наука, 1977.
- [73] Воеводин В.В., Кузнецов Ю.А. Матрицы и вычисления. – М.: Наука, 1984.
- [74] Воробьев Ю.В. Метод моментов в прикладной математике. – М.: Физматлит, 1958.
- [75] Высшая алгебра. Справочная математическая библиотека. – М.: Физматгиз, 1962.
- [76] Гавриков М.Б., Таюрский А.А. Функциональный анализ и вычислительная математика. – М.: URSS, 2016.
- [77] Гавурин М.К. Лекции по методам вычислений. – М.: Наука, 1971.
- [78] Годунов С.К. Современные аспекты линейной алгебры. – Новосибирск: Научная книга, 1997.
- [79] Годунов С.К., Антонов А.Г., Кирилюк О.Г., Костин В.И. Гарантированная точность решения систем линейных уравнений в евклидовых пространствах. – Новосибирск: Наука, 1988 и 1992.
- [80] Горбаченко В.И. Вычислительная линейная алгебра с примерами на MATLAB. – СПб.: «БХВ-Петербург», 2011.
- [81] Джордж А., Лю Дж. Численное решение больших разреженных систем уравнений. – М.: Мир, 1984.
- [82] Дробышевич В.И., Дымников В.П., Ривин Г.С. Задачи по вычислительной математике. – М.: Наука, 1980.
- [83] Загускин В.Л. Справочник по численным методам решения уравнений. – М.: Физматгиз, 1960.
- [84] Зельдович Я.Б., Мышикис А.Д. Элементы прикладной математики. – М.: Наука, 1972.
- [85] Икрамов Х.Д. Численное решение матричных уравнений. – М.: Наука, 1984.
- [86] Калиткин Н.Н. Численные методы. – М.: Наука, 1978.
- [87] Калиткин Н.Н., Юхно Л.Ф., Кузьмина Л.В. Количественный критерий обусловленности систем линейных алгебраических уравнений // Математическое моделирование. – 2011. – Т. 23, №2. – С. 3–26.

- [88] Колмогоров А.Н., Фомин С.В. Элементы теории функций и функционального анализа. – М.: Наука, 1976, а также более поздние издания.
- [89] Крылов А.Н. Лекции о приближённых вычислениях. – М.: ГИТТЛ, 1954, а также более ранние издания.
- [90] Крылов В.И., Бобков В.В., Монастырный П.И. Вычислительные методы высшей математики. Т. 1. – Минск: «Вышешшая школа», 1972.
- [91] Кублановская В.Н. О некоторых алгорифмах для решения полной проблемы собственных значений // Журнал вычисл. матем. и мат. физики. – 1961. – Т. 1, № 4. – С. 555–570.
- [92] Лагутин М.Б. Наглядная математическая статистика. 2-е изд. – М.: «БИ-НОМ. Лаборатория знаний», 2011.
- [93] Лебедев В.И. Функциональный анализ и вычислительная математика. – М.: Физматлит, 1989.
- [94] Марчук Г.И. Методы вычислительной математики. – М.: Наука, 1989.
- [95] Орtega Дж., Рейнболдт В. Итерационные методы решения нелинейных систем уравнений со многими неизвестными. – М.: Мир, 1975.
- [96] Парлетт Б. Симметричная проблема собственных значений. Численные методы. – М.: Мир, 1983.
- [97] Паради М. Локализация характеристических чисел матриц и её применения. – М.: Издательство иностранной литературы, 1960.
- [98] Ремез Е.Я. Основы численных методов чебышевского приближения. – Киев: Наукова думка, 1969.
- [99] Самарский А.А., Николаев Е.С. Методы решения сеточных уравнений. – М.: Наука, 1978.
- [100] Уилкинсон Дж., Райнш К. Справочник алгоритмов на языке Алгол. Линейная алгебра. – М.: Машиностроение, 1976.
- [101] Фаддеева В.Н. Вычислительные методы линейной алгебры. – М.–Л.: Гостехиздат, 1950.
- [102] Флэнаган Д., Мацумото Ю. Язык программирования Ruby. – СПб.: Питер, 2011.
- [103] Халмош П. Конечномерные векторные пространства. – М.: ГИФМЛ, 1963.
- [104] Шарай И.А. IntLinIncR2 — пакет программ для визуализации множеств решений интервальных линейных систем с двумя неизвестными. Версия для MATLAB. 2014. Свободно доступно на [http://www.nsc.ru/interval/Programing/MCodes/IntLinIncR2\\_UTF8.zip](http://www.nsc.ru/interval/Programing/MCodes/IntLinIncR2_UTF8.zip)
- [105] Шарый С.П. Конечномерный интервальный анализ. – Электронная книга, 2024. Свободно доступна на <http://www.nsc.ru/interval/Library/InteBooks>
- [106] Яненко Н.Н. Метод дробных шагов решения многомерных задач математической физики. – Новосибирск: Наука, 1967.
- [107] ACKLEH A.S., ALLEN E.J., KEARFOTT R.B., SESHAIYER P. Classical and modern numerical analysis. – Boca Raton–London–New York: CRC press, 2010.

- [108] BAUER F.L., FIKE C.T. Norms and exclusion theorems // *Numerische Mathematik*. – 1960. – Vol. 2. – P. 137–141.
- [109] BROWNE E.T. Limits to the characteristic roots of a matrix // *American Mathematical Monthly*. – 1939. – Vol. 46, No. 5. – P. 252–265.
- [110] CHOW A.W. On the optimality of Krylov information // *Journal of Complexity*. – 1987. – Vol. 3. – P. 26–40.
- [111] ECKART C., YOUNG G. The approximation of one matrix by another of lower rank // *Psychometrika*. – 1936. – Vol. 1. – P. 211–218.
- [112] GALPERIN E.A. The condition problem in solution of linear multistage systems // *Interval Mathematics 1975*, edited by K. Nickel. – Berlin - Heidelberg - New York: Springer-Verlag, 1975. – P. 191–195. – (Lecture Notes in Computer Science; vol. 29)
- [113] GNU Octave — Scientific Programming Language. <https://octave.org/>
- [114] GREGORY R.T., KARNEY D.L. *A collection of matrices for testing computational algorithms*. – Hantington, New York: Robert E. Krieger Publishing Company, 1978.
- [115] *Handbook of Constraint Programming*. F. Rossi, P. van Beek, T. Walsh, eds. – Amsterdam: Elsevier, 2006.
- [116] HIGHAM N.J. *Accuracy ans stability of numerical algorithms*. – Philadelphia: SIAM, 2002.
- [117] HOTELLING H. Analysis of a complex of statistical variables into principal components // *J. Educ. Psych.* – 1933 – Vol. 24. – Part I: pp. 417–441, Part II: pp. 498–520.
- [118] KREINOVICH V., LAKEYEV A.V., NOSKOV S.I. Approximate linear algebra is intractable // *Linear Algebra and its Applications*. – 1996. – Vol. 232. – P. 45–54.
- [119] KREINOVICH V., LAKEYEV A.V., ROHN J., KAHL P. *Computational complexity and feasibility of data processing and interval computations*. – Dordrecht: Kluwer, 1997.
- [120] MAYER G. *Interval analysis and automatic result verification*. – Berlin: De Gruyter, 2017.
- [121] MOLER C. Professor SVD // *The MathWorks News & Notes*. – October 2006. – P. 26–29.
- [122] MOORE R.E., KEARFOTT R.B., CLOUD M. *Introduction to interval analysis*. – Philadelphia: SIAM, 2009.
- [123] NEMIROVSKY A.S. On optimality of Krylov's information when solving linear operator equation // *Journal of Complexity*. – 1991. – Vol. 7. – P. 121–130.
- [124] RICHARDSON L.F. The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam // *Philosophical Transactions of the Royal Society A*. – 1910. – Vol. 210. – P. 307–357.
- [125] RUMP S.M. Verification methods: rigorous results using floating-point arithmetic // *Acta Numerica*. – 2010. – Vol. 19. – P. 287–449.

- [126] SAAD Y. *Numerical methods for large eigenvalue problems. Second edition.* – Philadelphia: SIAM, 2011.
- [127] SCHULZ G. Iterative Berechnung der reziproken Matrix // *Z. Angew. Math. Mech.* – 1933. – Bd. 13 (1). – S. 57–59.
- [128] STOER J., BULIRSCH R. *Introduction to numerical analysis.* – Berlin-Heidelberg-New York: Springer-Verlag, 1993.
- [129] TODD J. The condition number of the finite segment of the Hilbert matrix // *National Bureau of Standards, Applied Mathematics Series.* – 1954. – Vol. 39. – P. 109–116.
- [130] TURING A.M. Rounding-off errors in matrix processes // *Quarterly Journal of Mechanics and Applied Mathematics.* – 1948. – Vol. 1. – P. 287–308.
- [131] VARAH J.M. A lower bound for the smallest singular value of a matrix // *Linear Algebra and its Applications.* – 1975. – Vol. 11. – P. 3–5.
- [132] VARGA R.S. On diagonal dominance arguments for bounding  $\|A^{-1}\|_\infty$  // *Linear Algebra and its Applications.* – 1976. – Vol. 14. – P. 211–217.
- [133] VARGA R.S. *Matrix iterative analysis.* – Berlin, Heidelberg, New York: Springer Verlag, 2000, 2010.
- [134] VON NEUMANN J., GOLDSTINE H.H. Numerical inverting of matrices of high order // *Bulletin of the American Mathematical Society.* – 1947. – Vol. 53, No. 11. – P. 1021–1099.
- [135] WILF H.S. *Finite sections of some classical inequalities.* – Heidelberg: Springer, 1970.

## Глава 4

# Решение нелинейных уравнений и их систем

## 4.1 Обзор постановок задачи

В этой главе рассматривается задача решения уравнений

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

и систем уравнений

$$\begin{cases} F_1(x_1, x_2, \dots, x_n) = 0, \\ F_2(x_1, x_2, \dots, x_n) = 0, \\ \vdots \quad \ddots \quad \vdots \\ F_n(x_1, x_2, \dots, x_n) = 0, \end{cases} \quad (4.2)$$

где  $x, x_1, \dots, x_n$  — вещественные неизвестные переменные,  $f(x)$  и  $F_i(x_1, x_2, \dots, x_n)$  — вещественные функции. Систему уравнений (4.2) можно также записать кратко в виде

$$F(x) = 0, \quad (4.3)$$

где  $x = (x_1, x_2, \dots, x_n)^\top \in \mathbb{R}^n$  — вектор неизвестных переменных,

$F(x) = (F_1(x), F_2(x), \dots, F_n(x))^\top$  — вектор-столбец функций  $F_i$ .

*Решением уравнения* (4.1) называется значение переменной  $x$  (или все такие значения), которое обращает (4.1) в истинное равенство.

*Решением системы уравнений* (4.2)–(4.3) называется набор значений переменных  $x_1, x_2, \dots, x_n$ , которые обращают в истинные равенства одновременно все уравнения системы (4.2)–(4.3). В некоторых случаях желательно найти все такие возможные наборы, т. е. все решения системы, а иногда достаточно какого-то одного. Если система уравнений (4.2), (4.3) не имеет решений, нередко требуется предоставить обоснование этого факта или его подробный вывод, и им может быть программа для ЭВМ и протокол её работы и т. п.

Наряду с задачами, рассмотренными в главе 2, то есть интерполяцией и приближением функций, вычислением интегралов и др., задача решения уравнений и систем уравнений является одной из классических задач вычислительной математики.

Не все встречающиеся на практике системы уравнений можно представить в виде (4.2) с явно выписываемыми функциями  $F_i(x)$ ,  $i = 1, 2, \dots, n$ . Инженерам приходится решать большое количество таких систем уравнений, в которых эти функции задаются, к примеру, таблицами значений на некоторых сетках или графически, и даже семействами графиков, которые изображают ход функций нескольких переменных. Подобные задачи практически очень важны, но мы не будем их рассматривать.

Всюду далее предполагается, что функции  $F_i(x)$  по крайней мере непрерывны, а количество уравнений в системе (4.2)–(4.3) совпадает с количеством неизвестных переменных. Решение недоопределённых или переопределённых систем уравнений в этой главе систематически не затрагивается.

Помимо записи систем уравнений в каноническом виде (4.2)–(4.3) часто встречаются и другие формы их представления, например,

$$G(x) = H(x) \tag{4.4}$$

с какими-то функциями  $G$ ,  $H$ . Чрезвычайно важным частным случаем этой формы является *рекуррентный вид* системы уравнений (или одного уравнения),

$$x = G(x), \tag{4.5}$$

в котором неизвестная переменная выражена через саму себя. В этом случае решение системы уравнений (или уравнения) есть *неподвижная точка* отображения  $G$ , т. е. такая точка  $x^*$  из области определения  $G$ ,

которая переводится этим отображением сама в себя,  $x^* = G(x^*)$ . Кроме того, рекуррентный вид уравнения или системы хорош тем, что позволяет довольно просто организовать итерационный процесс для нахождения решения. Это мы могли видеть в главе 3 на примере систем линейных алгебраических уравнений.

Как правило, системы уравнений различного вида могут быть приведены друг к другу равносильными преобразованиями. В частности, несложно установить связь решений уравнений и систем уравнений вида (4.2)–(4.3) с неподвижными точками отображений, т. е. с решениями уравнений в рекуррентном виде (4.5). Ясно, что

$$F(x) = 0 \quad \iff \quad x = x - \Lambda F(x),$$

где  $\Lambda$  — ненулевой скаляр в одномерном случае или же неособенная  $n \times n$ -матрица в случае вектор-функции  $F$ . Поэтому решение уравнения

$$F(x) = 0$$

является неподвижной точкой отображения

$$G(x) := x - \Lambda F(x).$$

Если неизвестная  $x$  не является конечномерным вектором, а отображения  $F$  и  $G$  имеют общую природу, то математические свойства уравнений (4.3) и (4.5) могут существенно различаться. При этом формы записи (4.3) и (4.5), строго говоря, не вполне равносильны друг другу. По этой причине для их обозначения нередко употребляют отдельные термины — *уравнение первого рода* и *уравнение второго рода* соответственно, которые пришли из теории интегральных уравнений.

Решение уравнения  $f(x) = 0$  называется *кратным*, если помимо самой функции  $f$  на этом решении зануляется производная  $f'$ . Говорят также, что решение имеет *кратность*  $p$ , если в нём зануляются как сама функция  $f(x)$ , так и все её производные последовательных порядков вплоть до  $p - 1$ . Иначе, если производная  $f'$  не зануляется на решении, то оно называется *простым*. Таким образом, вблизи кратного решения функция «стелется», прежде чем принять нулевое значение, и её график пересекает ось абсцисс с нулевым углом (рис. 4.1). Это вызывает трудности в численном нахождении таких решений. Простые решения соответствуют пересечению оси абсцисс и графика функции под ненулевым углом, так что они устойчивы при малых «шевелениях» (возмущениях) уравнения и вычислять их проще (рис. 4.1).

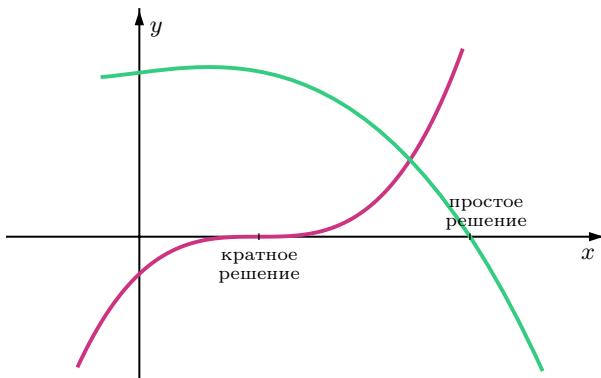


Рис. 4.1. Иллюстрация кратного и простого решений уравнения

Решение системы уравнений  $F(x) = 0$  называется *кратным*, если в этой точке зануляется функция  $F$ , а её матрица Якоби  $F'$ , составленная из частных производных отдельных компонент  $F$  по переменным, является особенной матрицей. Иначе, если матрица Якоби  $F'$  — неособая, то решение системы уравнений называется *простым*. Совершенно аналогично случаю скалярных уравнений численное нахождение кратных решений систем уравнений представляет большие трудности. Таким решениям соответствует в многомерном пространстве пересечение под нулевым углом поверхностей, отвечающих отдельным уравнениям системы. Это может иметь следствием неустойчивый характер решений, которые то ли существуют, то ли нет (см. § 4.2б).

Обращаясь к нахождению решений нелинейных уравнений и их систем, мы обнаруживаем себя в гораздо более сложных условиях, нежели при решении систем линейных алгебраических уравнений (3.72)–(3.73). Стойкая и весьма полная теория разрешимости систем линейных уравнений, базирующаяся на классических результатах линейной алгебры, обеспечивала в необходимых нам случаях уверенность в существовании решения системы линейных уравнений и его единственности. Для нелинейных уравнений столь общей и простой теории не существует. Напротив, нелинейные уравнения и их системы имеют в качестве общего признака лишь отрицание линейности, т. е. то, что все они «не линейны». Как следствие, нелинейные уравнения отличаются поэтом огромным разнообразием.

Из общих нелинейных уравнений и систем уравнений принято выделять *алгебраические*, в которых функции  $F_i(x)$  являются алгебраическими полиномами относительно неизвестных переменных  $x_1, x_2, \dots, x_n$ . Для систем алгебраических уравнений существуют *символьные методы решения*, основанные на преобразованиях алгебраических полиномов. Часто их называют *методами компьютерной алгебры*, и информацию о них читатель может найти в специализированной литературе [2, 53].

В заключение конспективно изложим результаты по вычислительной сложности решения алгебраических уравнений и систем уравнений, которые приведены и доказаны в [42] и других публикациях.

Для систем линейных алгебраических уравнений существуют полиномиально сложные алгоритмы вычисления решений. Это, например, метод исключения Гаусса и другие, рассмотренные в главе 3 нашей книги.

Для алгебраических уравнений 2-й степени (квадратичных) хорошо известны несложные формулы нахождения решений. Но для систем уравнений 2-й степени ситуация другая. Для всякого  $\epsilon > 0$  каждый алгоритм, который находит решение для любой системы квадратичных уравнений с погрешностью не более  $\epsilon$  требует хотя бы для одной такой системы по крайней мере экспоненциальных трудозатрат. Аналогично — для систем полиномиальных уравнений степени более двух, т. е. кубических и выше [42].

Уравнение или систему уравнений назовём *разрешимыми*, если они имеют решения. Существует полиномиально сложный алгоритм для проверки того, разрешима ли система линейных алгебраических уравнений (это тривиально). Но задача проверки разрешимости произвольной системы квадратных уравнений и систем уравнений более высоких степеней является NP-трудной [42] (см. также § 1.10).

Приведённые выше результаты описывают системы с различными степенями переменной, и в них не накладывается никаких ограничений на величину коэффициентов полиномов, входящих в систему уравнений, количество переменных и количество уравнений. Часто возникающие дополнительные условия хотя бы на один из этих параметров могут изменить ситуацию, а могут и вовсе не влиять на неё.

Если наложить какие-то априорные ограничения на коэффициенты полиномов, то результаты по вычислительной сложности не изменятся даже если потребовать, чтобы все коэффициенты полиномов принимали только значения 0 или 1. Такие полиномы называют 0-1-

*полиномами*, а соответствующие уравнения — *0-1-полиномиальными уравнениями*.

Оказывается, что для каждого  $\epsilon > 0$  всякий алгоритм, который находит  $\epsilon$ -приближённое решение любой системы 0-1-квадратных уравнений требует хотя бы для одной такой системы по крайней мере экспоненциальных трудозатрат. Задача проверки разрешимости системы 0-1-квадратных уравнений является NP-трудной.

Наложим ограничение на количество переменных  $n$ . Тогда для всякого  $n$  существует полиномиально сложный алгоритм, который находит  $\epsilon$ -приближённое решение произвольной системы полиномиальных уравнений с  $n$  неизвестными. Для всякого  $n$  существует полиномиально сложный алгоритм, который проверяет разрешимость произвольной системы полиномиальных уравнений с  $n$  неизвестными.

Если вместо числа неизвестных зафиксировать количество уравнений, то картина изменится. В частности, для одного линейного уравнения уравнения существует алгоритм с линейной вычислительной сложностью, который решает это уравнение (это тривиально). Существуют алгоритмы с полиномиальной трудоёмкостью, которые проверяют разрешимость любого квадратного или кубического уравнения. Задача проверки разрешимости произвольных полиномиальных уравнений степени 4 и выше является NP-трудной.<sup>1</sup>

Удобно свести сформулированные результаты в одну табличку [42], где в первом столбце указан тип полинома, а следующих двух столбцах — трудоёмкость решения одного уравнения и системы уравнений этого типа соответственно:

тип полинома	одно уравнение	система уравнений
линейный	линейная	полиномиальная
квадратичный	полиномиальная	NP-трудна
кубический	полиномиальная	NP-трудна
4-й степени и выше	NP-трудна	NP-трудна

Отметим, что описываемые далее в § 4.4, 4.5 и 4.7 численные методы нахождения решений уравнений и систем уравнений имеют по-

<sup>1</sup>Здесь уместно снова вспомнить теорему Абеля–Руффини и сопоставить с ней сформулированный результат о трудоёмкости выяснения разрешимости.

линомиальную трудоёмкость, что как будто противоречит информации из таблицы. Кажущийся парадокс объясняется тем, что численные методы носят локальный характер: они сходятся к решению лишь из достаточно близкого начального приближения. Сходимость из произвольного начального приближения ими не гарантируется, как и доказательство отсутствия решений. Лишь интервальные методы для глобального решения уравнений и систем уравнений из § 4.8 способны находить все решения или выдавать доказательное заключение об их отсутствии на данной области. Но эти численные методы имеют экспоненциальную в зависимости от размера задачи трудоёмкость.

## 4.2 Вычислительно-корректные задачи

### 4.2а Предварительные сведения и определения

Напомним общеизвестный факт: вещественные числа не являются конструктивным объектом, так как в общем случае сложность представления некоторых вещественных чисел (иррациональных) бесконечна.<sup>2</sup> Следствием этого обстоятельства является то, что на вычислительных машинах (как электронных, так и механических, как цифровых, так и аналоговых) мы можем выполнять, как правило, лишь приближённые вычисления над полем вещественных чисел  $\mathbb{R}$  или, точнее, лишь вычисления с приближениями вещественных чисел. Для цифровых вычислительных машин это заключение вытекает из того, что они являются дискретными и конечными устройствами, так что и ввод вещественных чисел в такую вычислительную машину, и выполнение с ними различных арифметических операций сопровождаются неизбежными ошибками, вызванными конечным характером представления чисел, конечностью исполнительных устройств и т. п. Для аналоговых вычислительных машин данные также не могут быть введены абсолютно точно, и процесс вычислений тоже не абсолютно точен. Потенциально все отмеченные погрешности могут быть сделаны сколь угодно малыми, но в принципе избавиться от них не представляется возможным. Получается, что реально

- мы решаем на вычислительной машине не исходную математическую задачу, а более или менее близкую к ней,

---

<sup>2</sup>Вещественные числа не являются конструктивными объектами также потому, что их множество несчётно.

- сам процесс решения на ЭВМ отличается от своего идеального математического прообраза, т. е. от результатов вычислений в  $\mathbb{R}$  или  $\mathbb{C}$  по тем формулам, которые его задают.

Возникновение и бурное развитие компьютерной алгебры с её «безошибочными» вычислениями едва ли опровергает высказанный выше тезис, так как исходные постановки задач для систем символьных преобразований требуют *точную* представимость входных данных. Эти данные поэтому подразумеваются целыми или, на худой конец, рациональными с произвольной длиной числителя и знаменателя [2], а все преобразования над ними не выводят за пределы поля рациональных чисел.

Как следствие, в условиях приближённого представления входных числовых данных и приближённого характера вычислений над полем вещественных чисел  $\mathbb{R}$  мы в принципе можем решать лишь те постановки задач, ответы которых «не слишком резко» меняются при изменении входных данных. Для этого, по крайней мере, должна иметь место непрерывная зависимость решения от входных данных.

Для формализации высказанных выше соображений нам необходимо точнее определить ряд понятий.

*Массовой задачей* [11] будем называть некоторый общий вопрос, формулировка которого содержит несколько свободных переменных — *входных данных* (или исходных данных), которые могут принимать значения в пределах предписанных им множеств. В целом массовая задача  $\Pi$  определяется

- 1) указанием её входных данных вместе с областями их определения,
- 2) формулировкой тех условий, которым должен удовлетворять *ответ*, т. е. решение этой задачи.

Неформально, массовая математическая задача — это семейство однотипных задач. Индивидуальная задача  $I$  получается из массовой задачи  $\Pi$  путём присваивания всем переменным входных данных задачи  $\Pi$  каких-то конкретных значений. Решением массовой задачи является общий метод (алгоритм), дающий для каждой из составляющих её единичных задач решение этой задачи. *Разрешающим отображением* задачи  $\Pi$  называем отображение, сопоставляющее каждому набору исходных данных ответ соответствующей индивидуальной задачи (см. § 1.7).

**Определение 4.2.1** Станем говорить, что массовая математическая задача является вычислительно корректной, если её разрешающее отображение  $\mathcal{P} \rightarrow \mathcal{A}$  из множества входных данных  $\mathcal{P}$  во множество  $\mathcal{A}$  ответов задачи непрерывно относительно некоторых топологий на  $\mathcal{P}$  и  $\mathcal{A}$ , определяемых содержательным смыслом задачи. Иначе массовую математическую задачу называем вычислительно некорректной.

Это определение является ослабленной версией классического определения корректности математических задач, данного в начале XX века Ж.Адамаром: решение задачи должно существовать, быть единственным и непрерывно зависеть от исходных данных [31]. Часто его так и называют — *корректность по Адамару*. Но в ситуациях, когда решение к задаче по существу может быть неединственным (решение нелинейного уравнения и т. п.) нет смысла жёстко требовать выполнение второго условия.

Те задачи, ответы на которые неустойчивы по отношению к возмущениям входных данных, могут решаться на ЭВМ с конечной разрядной сеткой лишь опосредованно, после проведения мероприятий, необходимых для защиты от этой неустойчивости или её нейтрализации.

Конечно, скорость изменения решения в зависимости от изменений входных данных может быть столь большой, что эта зависимость, даже будучи непрерывной и сколь угодно гладкой, становится похожей на разрывную. Это мы могли видеть в § 3.17г для собственных значений некоторых матриц, которые являются чрезвычайно быстро изменяющими функциями элементов матрицы, так что уже «практически разрывны». Но определением вычислительно корректной задачи выделяются те задачи, для которых хотя бы в принципе возможно добиться сколь угодно точного приближения к идеальному математическому ответу, например, увеличением количества значащих цифр при вычислениях и т. п.

**Пример 4.2.1** Задача решения систем линейных уравнений  $Ax = b$  с неособенной квадратной матрицей  $A$  является вычислительно-корректной. Если топология на пространстве  $\mathbb{R}^n$  её решений задаётся обычным евклидовым расстоянием и подобным же традиционным образом задаётся расстояние между векторами правой части и матрицами, то существуют хорошо известные неравенства (см. § 3.4а), оценивающие сверху

границы изменения решений  $x$  через изменения элементов матрицы  $A$ , правой части  $b$  и число обусловленности матрицы  $A$ . ■

**Пример 4.2.2** Нахождение ранга матрицы — вычислительно некорректная задача. Дело в том, что в основе понятия ранга лежит линейная зависимость строк или столбцов матрицы, т. е. свойство их линейной комбинации быть равной нулю или неравной нулю. Оно нарушается при сколь угодно малых возмущениях элементов матрицы. ■

Разрывная зависимость решения от входных данных задачи может возникать вследствие присутствия в алгоритме вычисления функции условных операторов вида IF ... THEN ... ELSE, приводящих к ветвлению. Такова хорошо известная функция знака числа

$$\operatorname{sgn} x = \begin{cases} -1, & \text{если } x < 0, \\ 0, & \text{если } x = 0, \\ 1, & \text{если } x > 0. \end{cases}$$

Аналогична функция модуля числа  $|x|$ , с которой в обычных и внешне простых выражениях могут быть замаскированы разрывы и ветвления. Например, таково частное  $\sin x / |x|$ , которое ведёт себя в окрестности нуля примерно как  $\operatorname{sgn} x$ .

Для систем нелинейных уравнений, могущих иметь неединственное решение, топологическую структуру на множестве ответов  $\mathcal{A}$  нужно задавать уже каким-либо расстоянием между множествами, например, с помощью так называемой *хаусдорфовой метрики* [7]. Напомним её определение.

Если задано метрическое пространство с метрикой  $\varrho$ , то *расстоянием* точки  $a$  до множества  $X$  называется величина  $\varrho(a, X)$ , определяемая как  $\inf_{x \in X} \varrho(a, x)$ . *Хаусдорфовым расстоянием* между компактными множествами  $X$  и  $Y$  называют величину

$$\varrho(X, Y) = \max \left\{ \max_{x \in X} \varrho(x, Y), \max_{y \in Y} \varrho(y, X) \right\}.$$

При этом  $\varrho(X, Y) = +\infty$ , если  $X = \emptyset$  или  $Y = \emptyset$ . Введённая таким образом величина действительно обладает всеми свойствами расстояния и может быть использована для задания топологии на пространствах решений тех задач, ответы к которым неединственны, т. е. являются уже множествами, а не отдельными точками.

## 4.26 Задача решения уравнений не является вычислительно-корректной

Уже простейшие примеры показывают, что задача решения уравнений и систем уравнений не является вычислительно-корректной. Например, квадратное уравнение

$$x^2 + px + q = 0, \quad (4.6)$$

как хорошо известно, имеет решения, выражаемые формулой

$$x_{1,2} = -\frac{p}{2} \pm \sqrt{\left(\frac{p}{2}\right)^2 - q}.$$

В частности, при условии

$$p^2 = 4q \quad (4.7)$$

существует лишь одно вещественное решение  $x = -p/2$  (верхний чертёж на рис. 4.2). Но при любых сколь угодно малых возмущениях коэффициента  $p$  и свободного члена  $q$ , нарушающих равенство (4.7), уравнение (4.6) либо теряет это единственное решение, либо приобретает ещё одно (нижние чертежи на рис. 4.2).

Аналогичным образом ведёт себя решение системы двух уравнений, равносильной (4.6),

$$\begin{cases} x + y = r, \\ xy = s \end{cases}$$

при  $s = r^2/4$ . При этом раздвоение решения не является большим грехом, коль скоро мы можем рассматривать хаусдорфово расстояние между нетривиальными множествами решений. Но вот исчезновение единственного решения, при котором расстояние между множествами решений скачком меняется до  $+\infty$ , — это чрезвычайное событие, однозначно указывающее на то, что разрешающее отображение не является непрерывным.

Как видим, математическую постановку задачи нахождения решений уравнений нужно «исправить», заменив какой-нибудь вычислительно-корректной постановкой задачи. Приступая к поиску ответа на этот математический вопрос, отметим, прежде всего, что с точки зрения практических приложений задачи, которые мы обычно формулируем в виде решения уравнений или систем уравнений, традиционно записывая соотношение (4.3),

$$F(x) = 0,$$

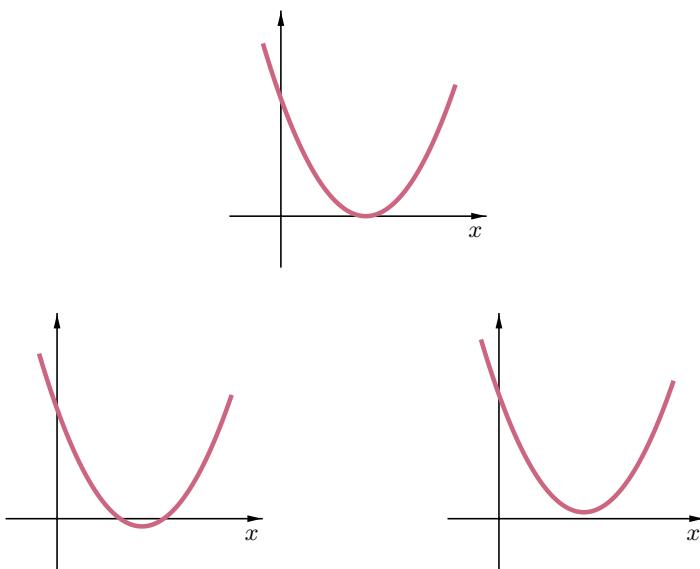


Рис. 4.2. Неустойчивая зависимость решений квадратного уравнения (4.6)–(4.7) от его коэффициентов

и ему подобные, имеют весьма различную природу. Это и будет отправной точкой нашей ревизии постановки задачи.

## 4.2в $\varepsilon$ -решения уравнений

В ряде практических задач заказчикам требуется не точное равенство некоторого выражения нулю, а лишь его «исчезающая малость» в сравнении с каким-то a priori установленным порогом. С аналогичной точки зрения часто имеет смысл рассматривать соотношения вида (4.4) или (4.5), которые фиксируют равенство двух каких-то выражений.

Таковы, например, в большинстве физических, химических и других естественнонаучных расчётов уравнения материального баланса, вытекающие из закона сохранения массы и закона сохранения заряда. Точное равенство левой и правой частей уравнения здесь неявным образом и не требуется, так как погрешность этого равенства всегда ограничена снизу естественными пределами делимости материи. В са-

мом деле, масса молекулы, масса и размеры атома, заряд элементарной частицы и т. п. величины, с точностью до которых имеет смысл рассматривать конкретные уравнения баланса — все они имеют вполне конечные (хотя и весьма малые) значения.

Например, не имеет смысла требовать, чтобы закон сохранения заряда выполнялся с погрешностью, меньшей чем величина элементарного электрического заряда (т. е. заряда электрона, равного  $1.6 \cdot 10^{-19}$  Кл). Также бессмысленно требовать, чтобы погрешность изготовления или подгонки деталей оптических систем была существенно меньшей длины световой волны (от  $4 \cdot 10^{-7}$  м до  $7.6 \cdot 10^{-7}$  м в зависимости от цвета). А что касается температуры, то при обычных земных условиях определение её с погрешностью, не превосходящей 0.001 градуса, вообще проблематично в силу принципиальных соображений. Наконец, ограниченная точность, с которой известны абсолютно все физические константы<sup>3</sup>, также возводит границы для требований равенства в физических соотношениях.

Совершенно аналогична ситуация с экономическими балансами, как в стоимостном выражении, так и в натуральном: требовать, чтобы они выполнялись с погрешностью, меньшей, чем одна копейка (наименьшая денежная величина) или чем единица неделимого товара (телевизор, автомобиль и т. п.) просто бессмысленно.

Во всех вышеприведённых примерах под решением уравнения понимается значение переменной, которое доставляет левой и правой частям уравнения пренебрежимо отличающиеся значения. В применении к уравнениям вида (4.3) соответствующая формулировка выглядит следующим образом:

Для заданных отображения  $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  и  $\varepsilon > 0$  найти значения неизвестной переменной  $x$ , такие что  $F(x) \approx 0$  с абсолютной погрешностью  $\varepsilon$ , т. е.  $\|F(x)\| \leq \varepsilon$ .

Решением этой задачи является, как правило, множество точек, которые мы будем называть  $\varepsilon$ -решениями или почти решениями, если порог этой пренебрежимой малости не оговорён явно или несуществен.

<sup>3</sup>В лучшем случае относительная погрешность известных на сегодняшний день значений физических констант равна  $10^{-10}$ , см. [71].

Фактически, понятие *псевдорешения* уравнений и систем уравнений, которое рассматривалось в главах 2 и 3, является дальнейшим развитием и обобщений  $\varepsilon$ -решений и почти решений.

Из известных результатов математического анализа следует, что условием  $\|F(x)\| < \varepsilon$  задаётся открытое множество, если отображение  $F$  непрерывно. Любая точка из этого множества устойчива к малым возмущениям исходных данных, а задача нахождения «почти решений» является вычислительно-корректной.

Как уже отмечалось выше, в некоторых задачах система уравнений более естественно записывается не как (4.3), а в виде (4.4)

$$G(x) = H(x).$$

Если требуется обеспечить с относительной погрешностью  $\varepsilon$  равенство её левой и правой частей, то задача формулируется следующим образом:

Для заданных отображений  $G, H : \mathbb{R}^n \rightarrow \mathbb{R}^n$  и  $\varepsilon > 0$   
найти значения неизвестной переменной  $x$ , такие что  
 $G(x) \approx H(x)$  с относительной погрешностью  $\varepsilon$ , т. е.

$$\frac{\|G(x) - H(x)\|}{\max\{\|G(x)\|, \|H(x)\|\}} \leq \varepsilon.$$

Решения этой задачи мы тоже будем называть  $\varepsilon$ -*решениями* системы уравнений вида (4.4).

Математические понятия, определения которых привлекают малый допуск  $\varepsilon$ , не являются чем-то экзотическим. Таковы, к примеру,  $\varepsilon$ -энтропия множеств в метрических пространствах,  $\varepsilon$ -субдифференциал функций,  $\varepsilon$ -оптимальные решения задач оптимизации и т. п. Одним из частных случаев  $\varepsilon$ -решений являются точки  $\varepsilon$ -спектра матрицы, предложенные для обобщения традиционного понятия собственного значения матрицы [9, 46, 73]. Говорят, что точка  $z$  на комплексной плоскости принадлежит  $\varepsilon$ -спектру матрицы  $A$ , если существует комплексный вектор  $v$  единичной длины, такой что  $\|(A - zI)v\| \leq \varepsilon$ , где  $\|\cdot\|$  — какая-то векторная норма. Иными словами, при условии  $\|v\| = 1$  здесь рассматривается приближённое «с точностью до  $\varepsilon$ » равенство  $Av = zv$ .

## 4.2г Недостаточность $\varepsilon$ -решений

Но есть принципиально другой тип задач, возникающих для уравнений и систем уравнений, которые образно могут быть названы задачами «об определении перехода через нуль», и они не сводятся к нахождению  $\varepsilon$ -решений. Таковы задачи, в которых требуется гарантированно отследить переход функции к значениям противоположного знака (или, более общо, переход через некоторое критическое значение). При этом, в частности, в любой окрестности решения должны присутствовать как положительные значения функции, так и её отрицательные значения, тогда как в задачах нахождения «почти решений» это условие может и не выполняться.

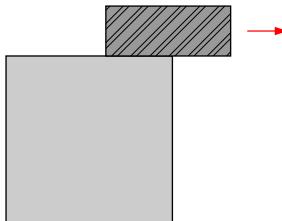


Рис. 4.3. Когда кирпич упадёт с подставки?

Рассмотрим следующую ситуацию, для анализа которой достаточно знание элементарной физики. Пусть кирпич лежит на жёсткой опоре (рис. 4.3), и мы потихоньку сдвигаем его к краю. Когда он упадёт? Для ответа на этот вопрос приравнивают момент силы тяжести, действующей на свисающую часть кирпича, и момент силы тяжести, действующей на ту часть, которая лежит на опоре.

Но в случае точного их равенства кирпич *ещё не упадёт!* Эта ситуация называется в физике «неустойчивым равновесием», и в отсутствие каких-либо воздействий на кирпич он не будет падать, а зависнет на грани опоры. Для падения кирпича именно нужен его переход чуть дальше этого положения неустойчивого равновесия (либо какое-то дополнительное внешнее воздействие). Но  $\varepsilon$ -решения для анализа этой ситуации совершенно не годятся по существу дела.

Другой пример. Фазовый переход в физической системе (плавление, кристаллизация и т. п.) — типичная задача такого сорта, так как в процессе фазового перехода температура системы не меняется. Если

мы хотим узнать, прошёл ли фазовый переход полностью, то нужно зафиксировать момент достижения множества состояний, лежащего по другую сторону от границы раздела различных состояний!

Ещё один пример. Рассмотрим динамическую систему, описываемую линейными дифференциальными уравнениями с постоянными коэффициентами:

$$\frac{dx}{dt} = Ax, \quad (4.8)$$

где  $x = (x_1, x_2, \dots, x_n)^\top$ , а  $n \times n$ -матрица  $A = A(\theta)$  зависит от параметра  $\theta$  (возможно, векторного). Пусть при некотором начальном значении  $\theta = \theta_0$  собственные значения  $\lambda(A)$  матрицы  $A$  имеют отрицательные вещественные части, так что все решения системы (4.8) устойчивы по Ляпунову и даже асимптотически устойчивы [59]. При каких значениях параметра  $\theta$  рассматриваемая система сделается неустойчивой?

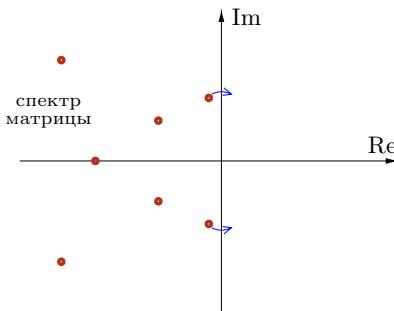


Рис. 4.4. Срыв устойчивости в динамической системе (4.8) происходит, когда собственные числа матрицы  $A$  «переходят» через мнимую ось

Традиционно отвечают на этот вопрос следующим образом. Срыв устойчивости в системе (4.8) произойдет при  $\text{Re } \lambda(A(\theta)) = 0$  для какого-то собственного значения, так что для определения этого момента нужно найти решение выписанного уравнения. Но такой ответ неправилен, так как для потери устойчивости необходимо не точное равенство нулю действительных частей некоторых собственных чисел матрицы, а переход их через нуль в область положительного знака. Без этого перехода через мнимую ось и «ещё чуть-чуть дальше» система останется устойчивой, сколь бы близко мы не придвинули собственные значения к мнимой оси или даже достигли бы её. Здесь важен именно переход

«через и за» критическое значение, в отсутствие которого качественное изменение в поведении системы не совершится, и этот феномен совершенно не ухватывается понятиями  $\epsilon$ -решения из § 4.2в или  $\epsilon$ -спектра из работ [9, 46, 73].

Рассмотренная ситуация, в действительности, весьма типична для динамических систем, где условием совершения многих типов структурных перестроек и изменений установившихся режимов работы систем — так называемых *бифуркаций* — является переход некоторого параметра через определённое *бифуркационное значение*. К примеру, при переходе через мнимую ось пары комплексных собственных чисел матрицы линеаризованной системы происходит бифуркация Андронова–Хопфа, называемая также «бифуркацией рождения цикла» [37]. И здесь принципиален именно переход через некоторый порог, а не близость к нему, на которую делается упор в понятиях  $\epsilon$ -решения и  $\epsilon$ -спектра.

Нетрудно понять, что такое «переход через нуль» для непрерывной функции одного переменного  $f : \mathbb{R} \rightarrow \mathbb{R}$ . Но в многомерной ситуации мы сталкиваемся с методическими трудностями, возникающими из необходимости иметь для нестрогого понятия «прохождение функции через нуль» чисто математическое определение. Из требования вычислительной корректности следует, что в любой окрестности такого решения каждая из компонент  $F_i(x)$  вектор-функции  $F(x)$  должна принимать как положительные, так и отрицательные значения. Но как именно? Какими должны (или могут) быть значения компонент  $F_j(x)$ ,  $j \neq i$ , если  $F_i(x) > 0$  или  $F_i(x) < 0$ ?

В разрешении этого затруднения нам на помощь приходят нелинейный анализ и алгебраическая топология. В следующем параграфе мы приведём краткий набросок возможного решения этого вопроса.

## 4.3 Существование решений уравнений и систем уравнений

Этот раздел посвящён краткому обзору математических результатов, касающихся существования решений уравнений и систем уравнений. Эти решения можно рассматривать с различных позиций, и, по-видимому, важнейшими являются представления решений как особых точек векторных полей и неподвижных точек отображений.

### 4.3а Векторные поля

Если  $M$  — некоторое множество в  $\mathbb{R}^n$  и задано отображение

$$\Phi : M \rightarrow \mathbb{R}^n,$$

то часто удобно представлять значение  $\Phi(x)$  как вектор, торчащий из точки  $x \in M$ . При этом говорят, что на множестве  $M$  задано *векторное поле*  $\Phi$ . Любопытно, что это понятие было введено около 1830 года М. Фарадеем в связи с необходимостью построения теории электрических и магнитных явлений. Затем соответствующий язык проник в математическую физику, теорию дифференциальных уравнений и теорию динамических систем (см., к примеру, [7, 49]), и в настоящее время широко используется в современном естествознании. Мы воспользуемся соответствующими понятиями и результатами для наших целей анализа решений систем уравнений, численных методов и коррекции постановки задачи.

Векторное поле называется *непрерывным*, если непрерывно отображение  $\Phi(x)$ . Например, на рис. 4.5 изображены векторные поля

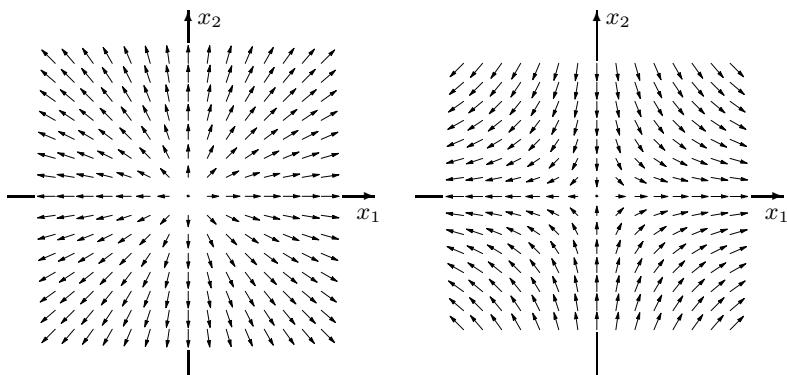
$$\Phi(x) = \Phi(x_1, x_2) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \text{и} \quad \Psi(x) = \Psi(x_1, x_2) = \begin{pmatrix} x_1 \\ -x_2 \end{pmatrix}, \quad (4.9)$$

которые непрерывны и даже дифференцируемы.

**Определение 4.3.1** Пусть задано векторное поле  $\Phi : \mathbb{R}^n \supseteq M \rightarrow \mathbb{R}^n$ . Точки  $x \in M$ , в которых поле обращается в нуль, т. е.  $\Phi(x) = 0$ , называются *нулями поля* или же *его особыми точками*.

Объяснение термина «особая» в отношении точки зануления векторного поля легко понять из того факта, что в такой точке никакого определённого направления векторное поле не задаёт. Если векторным полем определяется, скажем, направление движения (эволюции объекта) по фазовой плоскости, то в особой точке мы встречаемся с неопределенностью, и она может разрешаться очень специфичными способами.

Связь векторных полей и их особых точек с основным предметом этой главы очевидна: особая точка поля  $\Phi : M \rightarrow \mathbb{R}^n$  — это решение

Рис. 4.5. Векторные поля  $\Phi(x)$  и  $\Psi(x)$ , задаваемые формулами (4.9)

системы  $n$  уравнений

$$\left\{ \begin{array}{l} \Phi_1(x_1, x_2, \dots, x_n) = 0, \\ \Phi_2(x_1, x_2, \dots, x_n) = 0, \\ \vdots \quad \ddots \quad \vdots \\ \Phi_n(x_1, x_2, \dots, x_n) = 0, \end{array} \right.$$

лежащее в  $M$ . Будем говорить, что векторное поле  $\Phi$  *вырождено*, если у него есть особые точки. Иначе  $\Phi$  называется *невырожденным*. К примеру, векторные поля рис. 4.5 вырождены на всём  $\mathbb{R}^2$  и имеют единственными особыми точками начало координат.

**Определение 4.3.2** Пусть  $\Phi(x)$  и  $\Psi(x)$  — векторные поля на множестве  $M \subseteq \mathbb{R}^n$ . Непрерывная функция

$$\Delta(\lambda, x) : \mathbb{R} \times M \rightarrow \mathbb{R}^n$$

от параметра  $\lambda \in [0, 1]$  и вектора  $x \in \mathbb{R}^n$ , такая что  $\Phi(x) = \Delta(0, x)$  и  $\Psi(x) = \Delta(1, x)$ , называется деформацией векторного поля  $\Phi(x)$  в векторное поле  $\Psi(x)$ .

Достаточно прозрачна связь деформаций с возмущениями векторного поля, т. е. отображения  $\Phi$ . Но в качестве инструмента исследования решений систем уравнений и особых точек векторных полей нам

нужны деформации, которые не искажают свойство поля быть невырожденным.

**Определение 4.3.3** Деформацию  $\Delta(\lambda, x)$  векторного поля назовём невырожденной, если  $\Delta(\lambda, x) \neq 0$  для всех  $\lambda \in [0, 1]$  и  $x \in M$ .

Ясно, что невырожденные деформации могут преобразовывать друг в друга (соединять) только невырожденные векторные поля. Примерами невырожденных деформаций векторных полей, заданных на всём  $\mathbb{R}^n$ , являются растяжение, поворот относительно некоторой точки, параллельный перенос.

**Определение 4.3.4** Если векторные поля можно соединить невырожденной деформацией, то они называются гомотопными.

В частности, любая достаточно малая деформация невырожденного векторного поля приводит к гомотопному полю, что следует из соображений непрерывности.

Нетрудно понять, что отношение гомотопии векторных полей рефлексивно, симметрично и транзитивно, будучи поэтому *отношением эквивалентности*. Как следствие, непрерывные векторные поля, невырожденные на фиксированном множестве  $M \subseteq \mathbb{R}^n$ , распадается на классы гомотопных между собой полей.

### 4.3б Вращение векторных полей

Пусть  $D$  — ограниченная область в  $\mathbb{R}^n$  с границей  $\partial D$ . Через  $\text{cl } D$  мы обозначим топологическое замыкание  $D$ . Оказывается, каждому невырожденному на  $\partial D$  векторному полю  $\Phi$  можно сопоставить целочисленную характеристику — *вращение векторного поля*  $\Phi$  на  $\partial D$ , — обозначаемую  $\gamma(\Phi, D)$  и удовлетворяющую следующим условиям:

- (A) Гомотопные на  $\partial D$  векторные поля имеют одинаковое вращение.
- (B) Пусть  $D_i$ ,  $i = 1, 2, \dots$ , — непересекающиеся области, лежащие в  $D$  (их может быть бесконечно много). Если непрерывное векторное поле  $\Phi$  невырождено на теоретико-множественной разности

$$\text{cl } D \setminus \left( \bigcup_i D_i \right),$$

то вращения  $\gamma(\Phi, D_i)$  отличны от нуля лишь для конечного набора  $D_i$  и

$$\gamma(\Phi, D) = \gamma(\Phi, D_1) + \gamma(\Phi, D_2) + \dots$$

- (C) Если  $\Phi(x) = x - a$  для некоторой точки  $a \in D$ , то вращение  $\Phi$  на  $\partial D$  равно  $(+1)$ , т. е.

$$\gamma(\Phi, D) = 1.$$

Нетрудно понять, что определённая так величина вращения поля устойчива к малым шевелениям (возмущениям) как области (это следует из (B)), так и векторного поля (это вытекает из (A)). Условие (C) задаёт «калибровку» вращения, определяя, что вращение тождественного отображения, сдвинутого на вектор  $a$  из области  $D$ , равно 1. Фактически, с точностью до сдвига такое векторное поле совпадает с полем, изображённым на левом чертеже рис. 4.5.

Условиями (A)–(B)–(C) вращение векторного поля задаётся однозначно, но недостаток такого определения — отсутствие конструктивности. Можно показать (см. подробности в [62]), что сформулированное определение равносильно следующему конструктивному. Зафиксируем некоторую параметризацию поверхности  $\partial D$ , т. е. задание её в виде

$$x_1 = x_1(u_1, u_2, \dots, u_{n-1}),$$

$$x_2 = x_2(u_1, u_2, \dots, u_{n-1}),$$

$$\vdots \quad \vdots \quad \ddots \quad \vdots$$

$$x_n = x_n(u_1, u_2, \dots, u_{n-1}),$$

где  $u_1, u_2, \dots, u_{n-1}$  — параметры,  $x_i(u_1, u_2, \dots, u_{n-1})$ ,  $i = 1, 2, \dots, n$ , — функции, определяющие одноименные координаты точки  $x = (x_1, x_2, \dots, x_n) \in \partial D$ . Тогда вращение поля  $\Phi(x)$  на границе  $\partial D$  области  $D$  равно значению поверхностного интеграла

$$\frac{1}{S_n} \int_{\partial D} \frac{1}{\|\Phi(x)\|^n} \cdot \det \begin{pmatrix} \Phi_1(x) & \frac{\partial \Phi_1(x)}{\partial u_1} & \dots & \frac{\partial \Phi_1(x)}{\partial u_{n-1}} \\ \Phi_2(x) & \frac{\partial \Phi_2(x)}{\partial u_1} & \dots & \frac{\partial \Phi_2(x)}{\partial u_{n-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_n(x) & \frac{\partial \Phi_n(x)}{\partial u_1} & \dots & \frac{\partial \Phi_n(x)}{\partial u_{n-1}} \end{pmatrix} du_1 du_2 \dots du_n, \quad (4.10)$$

где  $S_n$  — площадь поверхности единичной сферы в  $\mathbb{R}^n$ . Этот интеграл обычно называют *интегралом Кронекера*.

В двумерном случае вращение векторного поля имеет простую геометрическую интерпретацию: это количество полных оборотов вектора поля, совершающееся при движении точки аргумента в положительном направлении по рассматриваемой границе области [26, 49, 55, 56, 58]. В многомерном случае такой наглядности уже нет, но величина вращения векторного поля  $\Phi$  всё равно может быть истолкована как «число раз, которое отображение  $\Phi : \partial D \rightarrow \Phi(\partial D)$  накрывает образ  $\Phi(\partial D)$ ».

**Пример 4.3.1** На любой окружности с центром в нуле поле, изображённое на левой половине рис. 4.5, имеет вращение +1, а поле на правой половине рис. 4.5 — вращение -1.

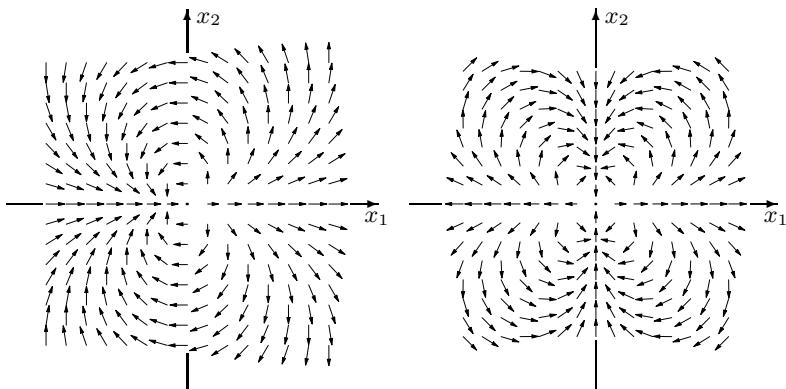


Рис. 4.6. Векторные поля, имеющие вращения +2 (левый чертёж) и +3 (правый чертёж) на любой окружности с центром в нуле

Векторные поля рис. 4.6, которые задаются формулами

$$\begin{cases} x_1 = r \cos(N\psi), \\ x_2 = r \sin(N\psi), \end{cases}$$

где  $r = \sqrt{x_1^2 + x_2^2}$  — длина радиус-вектора точки  $x = (x_1, x_2)$ ,  $\psi$  — его

угол с положительным лучом оси абсцисс, при  $N = 2$  и  $N = 3$  имеют вращения  $+2$  и  $+3$  на окружностях с центром в нуле. ■

С вращением векторного поля тесно связана другая известная глобальная характеристика отображений — *топологическая степень* [26, 27, 55, 56, 57, 58, 62]. Именно, вращение поля  $\Phi$  на границе области  $D$  есть топологическая степень такого отображения  $\phi$  границы  $\partial D$  в единичную сферу пространства  $\mathbb{R}^n$ , что

$$\phi(x) = \|\Phi(x)\|^{-1}\Phi(x).$$

Зачем нам понадобилось понятие вращения векторного поля? Мы собираемся использовать его для характеризации «прохождения через нуль» многомерной функции, и теоретической основой этого шага служат следующие результаты:

**Предложение 4.3.1** [26, 55, 56, 58] *Если векторное поле  $\Phi$  невырождено на замыкании ограниченной области  $D$ , то вращение  $\gamma(\Phi, D) = 0$ .*

**Теорема 4.3.1** (теорема Кронекера) [26, 55, 56] *Пусть векторное поле  $\Phi$  невырождено на границе ограниченной области  $D$  и непрерывно на её замыкании. Если  $\gamma(\Phi, D) \neq 0$ , то поле  $\Phi$  имеет в  $D$  по крайней мере одну особую точку.*

Теорема Кронекера обладает большой общностью, но проверка её условий, основанная на прямом вычислении вращения поля по формуле (4.10) требует большой вычислительной работы. Тем не менее она может быть проведена, и соответствующие численные методы для нахождения интеграла (4.10) используются в ответственных случаях [68, 72, 75].

Чаще всего теорема Кронекера применяется не напрямую, а служит основой для конкретных и несложно проверяемых достаточных условий существования нулей поля или решений систем уравнений. Например, доказательство теоремы Миранды (см. § 4.4б) сводится, фактически, к демонстрации того, что на границе области вращение векторного поля, соответствующего исследуемому отображению, равно  $\pm 1$ .

Другим ярким примером является

**Теорема 4.3.2** (теорема Брауэра о неподвижной точке)

*Пусть  $D$  — выпуклое компактное множество в  $\mathbb{R}^n$ . Если непрерывное отображение  $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$  переводит  $D$  в себя,  $g(D) \subseteq D$ , то оно имеет на  $D$  неподвижную точку  $x^*$ , т. е. такую что  $x^* = g(x^*)$ .*

Её доказательство и обсуждение можно найти, к примеру, в [14, 16, 26, 27, 57, 62]. Интервальные методы позволяют придать конструктивный характер этому результату, который раньше рассматривался скорее как «чистая» теоремы существования. Если вместо произвольных выпуклых компактов ограничиться интервальными векторами-брусами в  $\mathbb{R}^n$ , а для оценивания области значений применять его внешнюю оценку в виде интервального расширения, то условия теоремы Брауэра могут быть конструктивно проверены с помощью вычислений на компьютере.

Ещё большую конструктивную силу имеет следующая модификация теоремы Брауэра:

**Теорема 4.3.3** (усиленная теорема Брауэра о неподвижной точке)  
*Пусть  $D$  — выпуклое компактное множество в  $\mathbb{R}^n$ . Если непрерывное отображение  $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$  переводит границу  $\partial D$  множества  $D$  в само это множество,  $g(\partial D) \subseteq D$ , то  $g$  имеет на  $D$  неподвижную точку  $x^*$ , т. е. такую что  $x^* = g(x^*)$ .*

**Доказательство** можно найти в книгах [16, 26, 57].

Применение усиленной теоремы Брауэра требует оценивания области значений отображения на границе рассматриваемого множества, и в случае интервальных векторов-брусов в  $\mathbb{R}^n$  эта граница распадается на  $2n$  интервальных векторов размерности  $n - 1$ , имеющих меньшие размеры. Соответственно, погрешность нахождения оценки области значений интервальными методами при этом будет существенно меньшей, чем для всего множества.

### 4.3в Индексы особых точек

Станем говорить, что особая точка является *изолированной*, если в некоторой её окрестности нет других особых точек рассматриваемого векторного поля. Таким образом, вращение поля одинаково на сферах достаточно малых радиусов с центром в изолированной особой точке  $\tilde{x}$ . Это общее вращение называют *индексом* особой точки  $\tilde{x}$  поля  $\Phi$  или *индексом нуля*  $\tilde{x}$  поля  $\Phi$ , и обозначают  $\text{ind}(\tilde{x}, \Phi)$ .

Итак, оказывается, что особые точки векторных полей (и решения систем уравнений) могут быть существенно разными, отличаясь друг от друга своим индексом, и различных типов особых точек существует

столько же, сколько и целых чисел, т. е. счётное множество. Какими являются наиболее часто встречающиеся особые точки и, соответственно, решения систем уравнений? Ответ на этот вопрос даётся следующими двумя результатами:

**Предложение 4.3.2** [26, 55, 56] *Если  $A$  — невырожденное линейное преобразование пространства  $\mathbb{R}^n$ , то его единственная особая точка — нуль — имеет индекс  $\text{ind}(0, A) = \text{sgn } \det A$ , равный знаку определителя  $A$ .*

**Определение 4.3.5** Точка области определения отображения дифференцируемого отображения  $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  называется критической, если в ней якобиан  $F'$  является особенной матрицей. Иначе говорят, что эта точка — регулярная.

**Предложение 4.3.3** [26, 55, 56] *Если  $\tilde{x}$  — регулярная особая точка дифференцируемого векторного поля  $\Phi$ , то  $\text{ind}(\tilde{x}, \Phi) = \text{sgn } \det \Phi'(\tilde{x})$ .*

Таким образом, регулярные (не критические) особые точки векторных полей имеют индекс  $\pm 1$ , а в прочих случаях значение индекса может быть весьма произвольным.

Например, индексы расположенного в начале координат нуля векторных полей, которые изображены на рис. 4.5, равны  $+1$  и  $(-1)$ , при чём поля эти всюду дифференцируемы. Индексы нуля полей рис. 4.6 равны  $+2$  и  $+3$ , и в начале координат эти поля не дифференцируемы. Векторное поле на прямой, задаваемое рассмотренным в § 4.2б квадратичным отображением  $x \mapsto x^2 + px + q$  при  $p^2 = 4q$  имеет особую точку  $x = -p/2$  нулевого индекса.

### 4.3г Устойчивость особых точек

**Определение 4.3.6** Особая точка  $z$  поля  $\Phi$  называется устойчивой, если для любого  $\tau > 0$  можно найти такое  $\eta > 0$ , что всякое поле, отличающееся от  $\Phi$  меньше чем на  $\eta$ , имеет особую точку, удалённую от  $z$  менее, чем на  $\tau$ . Иначе особая точка  $z$  называется неустойчивой.

Ясно, что в связи с задачей решения уравнений и систем уравнений нас интересуют именно устойчивые особые точки, поскольку задача поиска только таких точек является вычислительно-корректной.

Вторым основным результатом, ради которого здесь представлен обзор теории вращения векторных полей, является следующее

**Предложение 4.3.4** [56] *Изолированная особая точка непрерывного векторного поля устойчива тогда и только тогда, когда её индекс отличен от нуля.*

Например, неустойчивое решение квадратного уравнения (4.6)–(4.7) имеет индекс 0, а у векторных полей, изображённых на рис. 4.5 и рис. 4.6, начало координат является устойчивой особой точкой.

Интересно отметить, что отличие линейных уравнений от нелинейных, как следует из всего сказанного, проявляется не только в форме и структуре, но и в более глубоких фактах:

в линейных задачах индекс решения, как правило, равен  $\pm 1$ , а в нелинейных может быть как нулевым, так и отличным от  $\pm 1$ , и, как следствие,

в типичных линейных задачах изолированное решение устойчиво, а в нелинейных может быть неустойчивым.

Отметим отдельно, что результат об устойчивости особой точки ненулевого индекса ничего не говорит о количестве особых точек, близких к возмущаемой особой точке. В действительности, путём шевеления одной устойчивой особой точки можно получить сразу *несколько* особых точек, и это легко видеть на примере одномерного векторного поля, порождаемого функцией  $y = x^3$ . Небольшая добавка, скажем, линейных членов приводит к тому, что вместо единственного нуля появляются три (рис. 4.7).

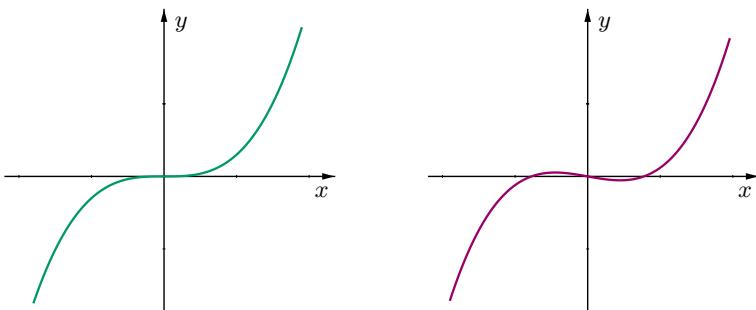


Рис. 4.7. Единственный нуль кубической параболы при возмущении функции распадается на три нуля

Аналогичное явление происходит с векторными полями рис. 4.6. Любая сколь угодно малая постоянная добавка к полю, изображённому на левом чертеже рис. 4.6, приводит к распадению нулевой особой точки индекса 2 на две особые точки индекса 1. Аналогично, любая сколь угодно малая постоянная добавка к полю, изображённому на правом чертеже рис. 4.6, приводит к распадению нулевой особой точки на три особые точки индекса 1. Таким образом, свойство единственности решения в общем случае неустойчиво и требовать его наличия нужно со специальными оговорками.

Если в области  $D$  находится конечное число особых точек, то сумму их индексов называют *алгебраическим числом особых точек*.

**Предложение 4.3.5** *Пусть непрерывное векторное поле  $\Phi$  имеет в  $D$  конечное число особых точек  $x_1, x_2, \dots, x_s$  и невырождено на границе  $\partial D$ . Тогда*

$$\gamma(\Phi, D) = \text{ind} (x_1, \Phi) + \text{ind} (x_2, \Phi) + \cdots + \text{ind} (x_s, \Phi).$$

Алгебраическое число особых точек устойчиво к малым возмущениям области и векторного поля, так как охватывает совокупную сумму индексов вне зависимости от рождения и уничтожения отдельных точек. Иллюстрацией может служить, опять таки, рис. 4.7.

#### 4.3д Вычислительно-корректная постановка задачи

Теперь все готово для вычислительно-корректной переформулировки задачи решения уравнений и систем уравнений. Она должна выглядеть следующим образом:

Для заданного  $\varepsilon > 0$  и системы уравнений

$$F(x) = 0$$

найти на данном множестве  $D \subseteq \mathbb{R}^n$

(4.11)

1) гарантированные двусторонние границы

всех решений ненулевого индекса,

2) множество  $\varepsilon$ -решений.

Мы не требуем единственности решения в пределах выдаваемых двусторонних границ, так как свойство решения быть единственным в общем случае не является, как мы могли видеть, устойчивым к малым возмущениям задачи. С другой стороны, в некоторых задачах единственность решения может быть установлена, и ниже мы обсудим такие ситуации.

Но решение задачи (4.11) — программа-максимум, которая может быть реализована очень нечасто и с большими трудозатратами, необходимыми для вычисления индекса особой точки через интеграл Кронекера. На практике с помощью классических методов решения уравнений обычно вычисляют решения «как таковые», без разделения их на типы. То, что найденное решение является точкой перехода через нуль или же  $\varepsilon$ -решением, часто можно заключить из свойств самого уравнения или моделируемого им процесса. Более полное решение сформулированной выше задачи (4.11) может быть получено с помощью интервальных методов.

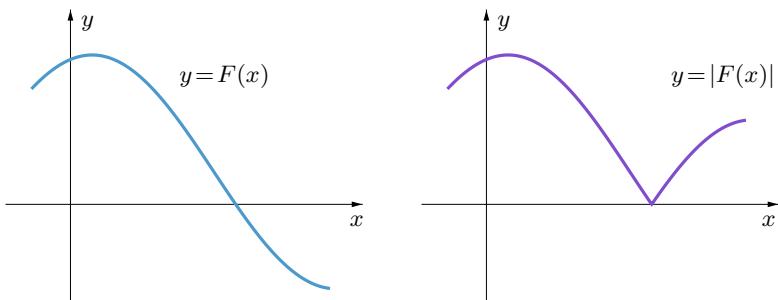


Рис. 4.8. Устойчивый нуль функции превращается в неустойчивый после взятия нормы функции

В заключение темы сделаем ещё важное замечание по поводу одной популярной переформулировки задачи решения уравнений и систем уравнений. Нередко на практике её представляют как оптимизационную, пользуясь тем, что справедливы следующие математические

эквивалентности: для одного уравнения —

$$\begin{aligned} f(x) = 0 &\Leftrightarrow \min_x |f(x)| = 0, \\ \text{или} \quad f(x) = 0 &\Leftrightarrow \min_x (f(x))^2 = 0, \end{aligned}$$

и для системы уравнений —

$$F(x) = 0 \Leftrightarrow \min_x \|F(x)\| = 0$$

где  $\|\cdot\|$  — какая-то векторная норма. Далее имеющимися стандартными методами (или пакетами программ) ищется решение задачи минимизации модуля  $|f(x)|$  или нормы  $\|F(x)\|$  (либо каких-то монотонных функций от них, чтобы обеспечить гладкость целевой функции), и затем результат сравнивается с нулем. Например,

$$F(x) = 0 \Leftrightarrow \min_x \left( (F_1(x))^2 + (F_2(x))^2 + \dots + (F_n(x))^2 \right) = 0.$$

Напомним, что близкие по духу вариационные переформулировки задачи решения системы линейных алгебраических уравнений мы успешно использовали в разделе 3.11.

С учётом наших знаний о задаче решения систем уравнений хорошо видна вычислительная неэквивалентность такого приведения: устойчивая особая точка *всегда* превращается при подобной трансформации в неустойчивое решение редуцированной задачи! Именно, любая сколь угодно малая добавка к  $|F(x)|$  может приподнять график функции  $y = |f(x)|$  над осью абсцисс (плоскостью нулевого уровня в случае системы уравнений), так что нуль функции исчезнет. Как следствие, вариационную переформулировку имеет смысл применять лишь в тех случаях, когда мы заранее знаем о существовании решения уравнения или системы уравнений, либо когда существование этого решения следует из самой задачи. Это типично, к примеру, для линейных уравнений и их систем.

#### 4.3e Теоремы о сжимающих отображениях

Выше в этом разделе мы рассматривали решения уравнений и систем уравнений как особые точки соответствующих векторных полей. Но эти решения можно исследовать и с других позиций, из которых

особую ценность имеет представление их в виде неподвижных точек некоторых отображений. Этот вопрос уже обсуждался вкратце в § 4.1.

Пусть  $X$  — множество произвольной природы. Напомним, что *неподвижной точкой* отображения  $G : X \rightarrow X$  называется такой элемент  $x^* \in X$ , что  $G(x^*) = x^*$ . Неподвижных точек у отображения может быть много (рис. 4.9), а может и не быть вовсе. В любом случае неподвижные точки — решения некоторых уравнений, имеющих специальную форму (рекуррентный вид), но очень часто к нему могут быть приведены уравнения довольно общего вида. Поэтому методы исследования и нахождения неподвижных точек отображений являются, по существу, методами решения уравнений.

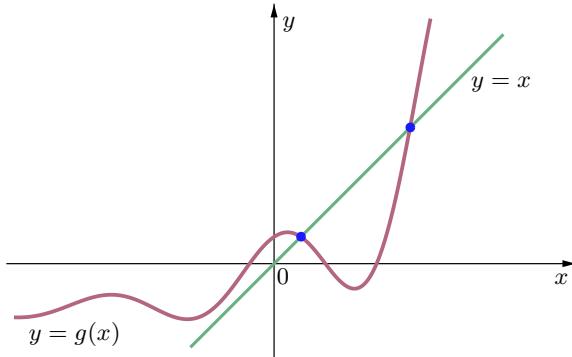


Рис. 4.9. Вещественная функция и её неподвижные точки, как пересечения графика с прямой  $y = x$

Существуют простые и практические условия существования неподвижных точек отображений, которые можно положить в основу численных методов решения уравнений.

**Определение 4.3.7** Пусть  $X$  — метрическое пространство с расстоянием  $\text{dist} : X \times X \rightarrow \mathbb{R}_+$ . Отображение  $g : X \rightarrow X$  называется *сжимающим* (или просто *сжатием*), если существует такая положительная постоянная  $\alpha < 1$ , что для любых элементов  $x, y \in X$  имеет место неравенство

$$\text{dist}(g(x), g(y)) \leq \alpha \cdot \text{dist}(x, y). \quad (4.12)$$

Иными словами, отображение является сжимающим, если расстояние между образами любых двух точек не более чем в  $\alpha$  раз пре-

восходит расстояние между прообразами, причём  $\alpha < 1$ . Сжимающее отображение, очевидно, непрерывно, так как сходимость  $x$  к  $\tilde{x}$  означает  $\text{dist}(x, \tilde{x}) \rightarrow 0$ , откуда следует, что  $\text{dist}(g(x), g(\tilde{x}))$  также стремится к нулю, т. е.  $g(x)$  сходится к  $g(\tilde{x})$ .

Свойство отображения быть сжимающим (которое будем кратко называть также «сжимаемость») родственно условию Липшица (стр. 47). В частности, относительно стандартного расстояния на вещественной оси,

$$\text{dist}(x, y) = |x - y|,$$

сжимаемость вещественных функций равносильна непрерывности по Липшицу с константой, которая меньше единицы. Выделение класса сжимающих отображений вызвано рядом их замечательных свойств, важнейшее из которых формализует

**Теорема 4.3.4** (теорема Банаха о неподвижной точке). *Сжимающее отображение  $g : X \rightarrow X$  полного метрического пространства  $X$  в себя имеет единственную неподвижную точку. Она может быть найдена как предел последовательных приближений*

$$x^{(k)} \leftarrow g(x^{(k-1)}), \quad k = 1, 2, \dots, \quad (4.13)$$

при любом начальном приближении  $x^{(0)} \in X$ . Скорость сходимости  $x^{(k)}$  к пределу — неподвижной точке  $x^*$  — описывается оценкой

$$\text{dist}(x^{(k)}, x^*) \leq \frac{\alpha^k}{1 - \alpha} \text{dist}(x^{(1)}, x^{(0)}).$$

Хотя теорема Банаха носит весьма общий характер, всё же напомним, что и вещественная ось  $\mathbb{R}$ , и арифметические пространства  $\mathbb{R}^n$  и  $\mathbb{C}^n$  являются полными метрическими пространствами.

Приведённая выше формулировка теоремы о сжимающем отображении была впервые дана С. Банахом в работе 1922 года, хотя лежащая в её основе идея применялась математиками и раньше. В частности, в конце XIX века Э. Пикар использовал итерации со сжимающим отображением для доказательства существования решения задачи Коши для дифференциальных уравнений (этот результат известен как «теорема Пикара» или «теорема Пикара-Линдлёфа»). В теореме Банаха особенно ценен конструктивный характер, позволяющий на её основе организовывать численные методы для нахождения неподвижных точек, которые являются решениями конкретных уравнений.

**Доказательство.** Если в условиях теоремы отображение  $g$  имеет две неподвижные точки  $x'$  и  $x''$ , то

$$\text{dist}(x', x'') = \text{dist}(g(x'), g(x'')) \leq \alpha \text{dist}(x', x'') < \text{dist}(x', x''),$$

что явно абсурдно при  $\text{dist}(x', x'') \neq 0$ . Единственная возможность устраниТЬ противоречие состоит в том, чтобы принять равенство

$$\text{dist}(x', x'') = 0,$$

т. е. совпадение этих неподвижных точек.

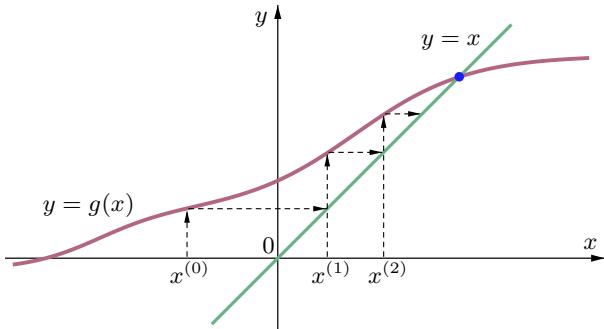


Рис. 4.10. Иллюстрация итераций со сжимающим отображением на вещественной оси

Взяв произвольный  $x^{(0)}$ , организуем последовательность по правилу (4.13) (рис. 4.10). Покажем, что она является фундаментальной последовательностью (последовательностью Коши) в пространстве  $X$ .

Справедливы неравенства

$$\text{dist}(x^{(2)}, x^{(1)}) \leq \alpha \text{dist}(x^{(1)}, x^{(0)}),$$

$$\text{dist}(x^{(3)}, x^{(2)}) \leq \alpha \text{dist}(x^{(2)}, x^{(1)}) \leq \alpha^2 \text{dist}(x^{(1)}, x^{(0)}),$$

$$\vdots \quad \ddots \quad \vdots$$

$$\text{dist}(x^{(k+1)}, x^{(k)}) \leq \alpha^k \text{dist}(x^{(1)}, x^{(0)}),$$

из которых вытекает

$$\begin{aligned}
 \text{dist}(x^{(k+p)}, x^{(k)}) &\leq \text{dist}(x^{(k+p)}, x^{(k+p-1)}) + \\
 &\quad + \text{dist}(x^{(k+p-1)}, x^{(k+p-2)}) + \\
 &\quad + \cdots \quad \cdots \quad + \\
 &\quad + \text{dist}(x^{(k+1)}, x^{(k)}) \leq \\
 &\leq (\alpha^{p-1} + \alpha^{p-2} + \cdots + \alpha + 1) \alpha^k \text{dist}(x^{(1)}, x^{(0)}) \leq \\
 &\leq \frac{\alpha^k}{1-\alpha} \text{dist}(x^{(1)}, x^{(0)}). \tag{4.14}
 \end{aligned}$$

Следовательно,  $\text{dist}(x^{(k+p)}, x^{(k)}) \rightarrow 0$  при  $k \rightarrow \infty$  для любых  $p$ , и последовательность  $\{x^{(k)}\}$  в самом деле является фундаментальной. Она сходится к некоторому пределу  $x^* = \lim_k x^{(k)}$ , поскольку  $X$  — полное метрическое пространство.

В силу непрерывности отображения  $g$  переход к пределу по  $k \rightarrow \infty$  в расчётом соотношении (4.13) даёт

$$x^* = g(x^*),$$

т. е.  $x^*$  является неподвижной точкой отображения  $g$ . Оценка скорости сходимости к ней последовательности приближений  $\{x^{(k)}\}$  получается переходом к пределу по  $p \rightarrow \infty$  в неравенстве (4.14). ■

Для существования и единственности неподвижной точки важны все условия из формулировки доказанной теоремы, как показывают простые примеры.

Рассмотрим тождественное отображение пространства  $X$  на себя. Для него  $\alpha = 1$  в неравенстве (4.12), так что сжимающим отображение не является. Но все точки из  $X$  являются для него неподвижными.

Рассмотрим в  $\mathbb{R}^n$  отображение сдвига на постоянный вектор  $a$ ,

$$x \mapsto x + a.$$

Очевидно, что неподвижных точек оно вообще не имеет. В то же время, это отображение не является и сжимающим, так как  $\alpha = 1$  в неравенстве (4.12), если расстояние в  $\mathbb{R}^n$  задаётся стандартным образом с помощью какой-либо нормы, как (3.30).

Какие отображения являются сжимающими?

Если вещественная функция  $g : \mathbb{R} \supset [a, b] \rightarrow \mathbb{R}$  непрерывно дифференцируема, то в силу теоремы Лагранжа о конечном приращении

$$g(x) - g(y) = g'(\xi)(x - y)$$

для некоторой точки  $\xi$  между  $x$  и  $y$ . Поэтому

$$|g(x) - g(y)| \leq \max_{\xi \in [a, b]} |g'(\xi)| |x - y|.$$

Таким образом, достаточным условием сжимаемости отображения  $g$  на интервале  $[a, b] \subseteq \mathbb{R}$  относительно стандартного расстояния вещественной оси является неравенство

$$\max_{\xi \in [a, b]} |g'(\xi)| < 1 \quad .$$

(4.15)

Если интервал  $[a, b]$  бесконечен, то в этом условии « $\max$ » нужно заменить на « $\sup$ ». Наглядная иллюстрация сжимающего отображения вещественной оси и теоремы Банаха о неподвижной точке дана на рис. 4.10.

Рассмотрим многомерный случай, когда  $g : \mathbb{R}^n \supset D \rightarrow \mathbb{R}^n$  — непрерывно дифференцируемое отображение из области  $D \subseteq \mathbb{R}^n$  в  $\mathbb{R}^n$ . Пусть расстояние на  $\mathbb{R}^n$  задаётся с помощью какой-то нормы  $\|\cdot\|$  как (3.30):

$$\text{dist}(x, y) = \|x - y\|.$$

Обозначим посредством  $g'(x)$  матрицу Якоби отображения  $g$ , т. е.

$$g'(x) := \left( \frac{\partial g_i(x)}{\partial x_j} \right)_{i,j=1,\dots,n} .$$

В дифференциальном исчислении функций многих переменных известна теорема о конечном приращении (или «формула конечных приращений»), которая является обобщением теоремы Лагранжа о среднем значении и даёт оценку приращения функции через приращение аргумента и оценку матрицы Якоби [14, 16, 18, 27]. Из неё в нашей ситуации следует, что для любых точек  $x, y \in D$  справедливо неравенство

$$\|g(x) - g(y)\| \leq \sup_{\xi} \|g'(\xi)\| \|x - y\|,$$

где к матрице Якоби  $g'(\xi)$  применяется согласованная матричная норма, а  $\sup$  берётся по всем точкам отрезка прямой, соединяющего  $x$  и  $y$ . Следовательно, если выполнено условие

$$\boxed{\sup_{\xi \in D} \|g'(\xi)\| < 1}, \quad (4.16)$$

т. е. если норма матрицы Якоби отображения  $g$  на  $D$  меньше единицы, то это отображение — сжимающее.

**Пример 4.3.2** Исследуем разрешимость уравнения

$$x - \frac{1}{2} \sin(x + \mu) + \nu = 0$$

и количество его решений в зависимости от значений вещественных параметров  $\mu$  и  $\nu$ .

Если переписать уравнение в рекуррентной форме

$$x = \frac{1}{2} \sin(x + \mu) - \nu,$$

то его решение станет неподвижной точкой отображения  $g$ , при котором  $x \mapsto \frac{1}{2} \sin(x + \mu) - \nu$ . Оно является сжимающим относительно стандартного расстояния на  $\mathbb{R}$ , так как производная

$$g'(x) = \frac{1}{2} \cos(x + \mu)$$

ограничена по модулю сверху числом  $\frac{1}{2}$ , меньшим единицы, т. е. выполнено неравенство (4.15). Следовательно, в силу теоремы Банаха о неподвижной точке при любых  $\mu$  и  $\nu$  исследуемое уравнение имеет единственное решение. ■

Рассмотрим применение теоремы Банаха о неподвижной точке к исследованию систем линейных алгебраических уравнений в рекуррентной форме

$$x = Cx + d.$$

Решение этого уравнения было всесторонне исследовано в § 3.10, но здесь нам будет важно то, что оно является неподвижной точкой отображения  $x \mapsto Cx + d$ .

Предположим, что в какой-то матричной норме справедливо неравенство  $\|C\| < 1$ . Выберем в  $\mathbb{R}^n$  векторную норму, согласованную с этой матричной нормой (что можно сделать согласно предложению 3.3.5), обозначая её тем же символом  $\|\cdot\|$ . С помощью этой нормы стандартным способом (3.30) можно задать расстояние в  $\mathbb{R}^n$ , и относительно него отображение  $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$ , определяемое как  $x \mapsto Cx + d$ , является сжимающим:

$$\|g(x) - g(y)\| = \|Cx - Cy\| \leq \|C\| \cdot \|x - y\|, \quad \text{причём } \|C\| < 1.$$

По теореме Банаха о неподвижной точке существует единственный  $x^* \in \mathbb{R}^n$ , такой что  $x^* = g(x^*)$ , т. е.  $x^* = Cx^* + d$ . К этому решению системы из любого начального приближения будет сходиться итерационный процесс  $x^{(k)} \leftarrow Cx^{(k-1)} + d$ ,  $k = 1, 2, \dots$ . Фактически, мы доказали другим способом предложение 3.10.1 (стр. 512) о сходимости стационарных одношаговых итерационных методов.

Иногда бывает полезно работать с векторнозначным расстоянием — *мультиметрикой*, которую мы будем обозначать той же аббревиатурой, но с большой буквы — Dist. Самый простой и естественный способ её введения на  $\mathbb{R}^n$ , по-видимому, состоит в следующем:

$$\text{Dist}(x, y) := \begin{pmatrix} \text{dist}(x_1, y_1) \\ \vdots \\ \text{dist}(x_n, y_n) \end{pmatrix} = \begin{pmatrix} |x_1 - y_1| \\ \vdots \\ |x_n - y_n| \end{pmatrix} \in \mathbb{R}_+^n, \quad (4.17)$$

т. е. векторное расстояние формируется как вектор из расстояний между отдельными компонентами рассматриваемых векторов.

Для мультиметрических пространств аналогом теоремы Банаха о неподвижной точке для сжимающих отображений является приводимая ниже теорема Шрёдера о неподвижной точке. Перед тем, как дать её точную формулировку, введём

**Определение 4.3.8** *Отображение  $g : X \rightarrow X$  мультиметрического пространства  $X$  с мультиметрикой  $\text{Dist} : X \rightarrow \mathbb{R}_+^n$  называется  $P$ -сжимающим (или просто  $P$ -сжатием), если существует неотрицательная  $n \times n$ -матрица  $P$  со спектральным радиусом  $\rho(P) < 1$ , такая что для всех  $x, y \in X$  имеет место*

$$\text{Dist}(g(x), g(y)) \leq P \cdot \text{Dist}(x, y). \quad (4.18)$$

Ряд авторов (в частности [45]) за матрицей  $P$  из (4.18) закрепляют отдельное понятие «оператора Липшица (матрицы Липшица) отображения  $g$ » из-за сходства неравенства (4.18) с классическим условием Липшица. Тогда в условиях определения 4.3.8 говорят, что «оператор Липшица для  $g$  — сжимающий».

**Теорема 4.3.5** (теорема Шрёдера о неподвижной точке) *Пусть отображение  $g : \mathbb{R}^n \supseteq X \rightarrow \mathbb{R}^n$  является  $P$ -сжимающим на замкнутом подмножестве  $X$  пространства  $\mathbb{R}^n$  с мультиметрикой  $\text{Dist}$ . Тогда  $g$  имеет единственную неподвижную точку, и для любого  $x^{(0)}$  последовательность итераций*

$$x^{(k)} = g(x^{(k-1)}), \quad k = 1, 2, \dots,$$

*сходится к неподвижной точке  $x^*$  отображения  $g$  в  $X$ , причём имеет место оценка*

$$\text{Dist}(x^{(k)}, x^*) \leq (I - P)^{-1}P \cdot \text{Dist}(x^{(k)}, x^{(k-1)}).$$

В действительности, выше дан частный конечномерный случай более общей формулировки теоремы Шрёдера, которую можно увидеть в [17]. Доказательства читатель найдёт, например, в книгах [1, 17, 27, 45].

В применении к системам линейных уравнений в рекуррентной форме  $x = Cx + d$  теорема Шрёдера относительно мультиметрики (4.17) приводит к достаточному условию  $\rho(C) < 1$  для существования и единственности решения. Другими словами, она сразу же обеспечивает доказательство достаточности в теореме 3.10.1, главном результате о сходимости стационарных линейных одношаговых итерационных методов.

## 4.4 Классические методы решения уравнений

Уравнения и системы уравнений часто классифицируют по типу функциональных зависимостей, которые фигурируют в них. *Алгебраическими* называются уравнения и системы уравнений, в обеих частях которых стоят алгебраические полиномы от неизвестных переменных (в частности, нулевой полином). *Дробно-рациональными* называются уравнения и системы уравнений, образованные дробно-рациональными функциями, т. е. частными алгебраических полиномов. Уравнения и

системы уравнений, в которых встречаются функции, не являющиеся алгебраическими полиномами или их частными, например, экспонента, логарифм, синус, косинус, арксинус и т. п., называют *трансцендентными*. Рассмотрим практический пример возникновения простейшего такого уравнения.

**Пример 4.4.1** Рабочие имеют кусок кровельного материала шириной  $l = 3.3$  метра и хотят покрыть им пролёт шириной  $h = 3$  метра, сделав крышу круглой, как часть поверхности кругового цилиндра. Балки, поддерживающие такую кровлю, должны иметь форму круговых сегментов (выделены серым на рис. 4.11), и для того, чтобы придать им правильную форму, нужно знать, какой именно радиус закругления крыши при этом получится. Эта задача возникает, например, при проектировании теплиц из прозрачного листового поликарбоната, который нельзя сильно сгибать. Крыша из такого материала должна быть гладкой, т. е. иметь непрерывные производные, которые к тому же меняются не слишком быстро. Кроме того, для стока талых и дождевых вод крышу теплицы всё равно необходимо делать скатной, с несложной формой.

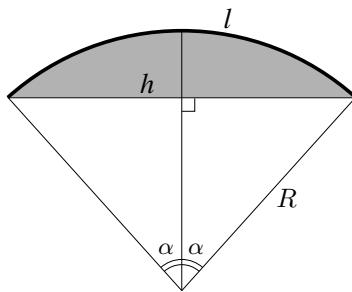


Рис. 4.11. Иллюстрация математической задачи проектирования круглой крыши

Обозначим искомый радиус крыши через  $R$ . Если  $2\alpha$  — угловая величина дуги (в радианах), соответствующей крыше, то

$$\frac{l}{2\alpha} = R.$$

С другой стороны, из рассмотрения прямоугольного треугольника с ка-

тетом  $h/2$  и гипотенузой  $R$  получаем

$$R \sin \alpha = h/2.$$

Исключая из этих двух соотношений  $R$ , получим уравнение относительно одной неизвестной  $\alpha$ :

$$l \sin \alpha = \alpha h. \quad (4.19)$$

Решив его, легко найдём и радиус закругления крыши  $R$ .

Но решение уравнения (4.19) не может быть выражено в виде какой-либо явной конечной формулы от  $l$  и  $h$ , использующей арифметические операции и элементарные функции. Сколь угодно точные приближения к искомому решению можно найти лишь численными методами, и мы обсудим их далее в оставшейся части главы. ■

Далее в этой главе рассматриваются основные и наиболее популярные численные методы для решения уравнений и систем уравнений, но, конечно, изложение не является всеохватным и полным. Мы не касаемся, например, специализированных методов, предназначенных для алгебраических полиномиальных уравнений. Это большое семейство численных методов, созданных за предшествующие столетия, которые имеют довольно специализированный характер и описаны, к примеру, в [52, 54]. Не рассматриваются так называемые ABS-методы [48], квазиньютоновские методы [13] и немало других эффективных инструментов вычислительной математики. Кроме того, полностью «за бортом» изложения остались современные методы компьютерной алгебры, предназначенные для решения алгебраических уравнений и их систем [2, 53]): эти методы основаны на символьных преобразованиях с полиномами и по сути не являются численными.

#### 4.4а Предварительная локализация решений

Обычно первым этапом численного решения уравнений и систем уравнений является предварительная локализация искомых решений, т. е. уточнение их местонахождения. В результате могут получаться как более или менее точные приближения, так и ограничение областей, где эти решения могут находиться. Необходимость этапа локализации решений вызвана тем, что большинство численных методов для их поиска имеют локальный характер, т. е. сходятся к этим решениям лишь

из достаточно близких начальных приближений. Качественное преимущество здесь имеют интервальные методы для решения уравнений и систем уравнений, рассматриваемые ниже в § 4.6–4.7. Они находят все решения, и хотя также требуют для своей работы указания некоторого интервала (или бруса), который содержит искомые решения, но размеры его обычно не слишком критичны.

Для локализации решений могут применяться как численные, так и аналитические методы, а также их смесь — гибридные методы, которые (следуя Д. Кнуту) можно назвать *получисленными* или *полуаналитическими*. Нельзя пренебрегать и *графическими* методами локализации решений, основанными на построении и исследовании графиков функций, которые фигурируют в уравнении. Решение уравнения получается как точка пересечения двух графиков (левой и правой частей), либо как точка пересечения графика с осью абсцисс (если в правой части нуль). Графические методы не обладают большой строгостью и точностью, но они просты, наглядны и тоже часто используются в практическом решении уравнений.

**Пример 4.4.2** Рассмотрим уравнение (4.19), которое перепишем в несколько изменённом виде:

$$\sin \alpha = \frac{h}{l} \alpha. \quad (4.20)$$

Характер изменения функций в левой и правой частях выписанного уравнения хорошо известен. Значения синуса в левой части всегда ограничены интервалом  $[-1, 1]$ , тогда как линейная функция в правой части монотонно возрастает и может принимать сколь угодно большие значения с ростом аргумента  $\alpha$  (рис. 4.12). Следовательно все решения уравнения (4.20) находятся в ограниченном множестве вещественной оси. Можно найти его размеры, основываясь на том соображении, что при  $|((h/l) \alpha)| > 1$  прямая не пересекает график синуса. Поэтому все решения уравнений (4.19) и (4.20) должны лежать во множестве

$$|\alpha| \leq l/h.$$

Напомним, что практическая постановка задачи, которая привела к рассматриваемому уравнению, требует нахождения его положительных решений на интервале  $[0, \pi]$ . Нетрудно показать, что при  $h < l$  такое решение обязательно существует и единственno. Из графиков,

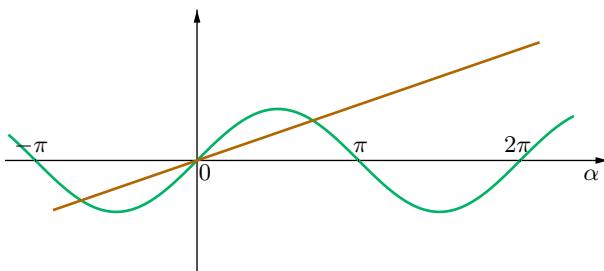


Рис. 4.12. Графики функций левой и правой частей уравнения (4.20)

аналогичных рис. 4.12, можно даже найти начальное приближение к решению, которое будет использовано в описываемых далее численных методах. ■

Особенно много аналитических результатов существует о локализации решений алгебраических уравнений (нулей полиномов), что, конечно, имеет причину в очень специальном виде этих уравнений, допускающем исследование с помощью выкладок, символьных преобразований и т. п. инструментов.

**Теорема 4.4.1** (теорема Декарта [22]) Для алгебраического полинома с вещественными коэффициентами число перемен знаков в последовательности его коэффициентов (при подсчёте которого нулевые значения не учитываются) равно количеству положительных нулей этого полинома с учётом кратности или же на чётное число большее этого количества.

Если, к примеру, заранее известно, что все нули данного полинома вещественны, то правило знаков Декарта даёт точное число нулей. Рассматривая полином с переменной  $(-x)$  можно с помощью этого же результата найти число отрицательных нулей исходного полинома.

**Теорема 4.4.2** Для алгебраического уравнения вида

$$a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0 = 0,$$

с вещественными или комплексными коэффициентами, все решения по модулю меньшие величины

$$1 + \frac{A}{|a_n|}$$

где  $A := \max\{|a_0|, \dots, |a_{n-1}|\}$ .

Доказательство можно увидеть, например, в книге [22], и оно основано на сравнении роста модуля старшего члена полинома и модуля остальных его членов. Ясно, что этот результат применим также к оценке вещественных решений алгебраических уравнений (если они существуют). В [22] показано, как с помощью замены переменной и несложных преобразований теорему 4.4.2 можно адаптировать для определения нижней границы вещественных положительных решений, а также нижней и верхней границ отрицательных решений.

Другие полезные результаты о локализации корней алгебраических полиномов, такие как теорема Бюдана-Фурье, способ Маклорена и другие, можно найти в книгах [22, 52, 54]. Очень мощным инструментом надёжного вычисления нулей алгебраических полиномов является метод полиномов Штурма [22].

Эффективные инструменты для локализации решений уравнений предоставляет интервальный анализ. Предположим, что необходимо исследовать наличие решений уравнения  $f(x) = 0$  на интервале  $\mathbf{X} \subset \mathbb{R}$ . Построим какое-нибудь интервальное расширение  $\mathbf{f}$  для функции  $f$  (см. § 1.6) и вычислим  $\mathbf{f}(\mathbf{X})$ . Полученный интервал даёт внешнюю оценку области значений функции  $f$  на  $\mathbf{X}$ , и потому если  $0 \notin \mathbf{f}(\mathbf{X})$ , то на  $\mathbf{X}$  нет решений уравнения  $f(x) = 0$ . Соответственно, интервал  $\mathbf{X}$  можно исключить из области поиска решения.

**Пример 4.4.3** Пусть необходимо найти вещественные решения уравнения

$$x^3 - 2x^2 + x - 2 = 0.$$

Известно, что алгебраическое уравнение нечётной степени всегда имеет хотя бы одно вещественное решение, и в данном случае его границы можно найти из теоремы 4.4.2: это интервал  $[-3, 3]$ .

В качестве интервального расширения  $\mathbf{f}$  функции из левой части уравнения возьмём естественное интервальное расширение полинома, переписанного по схеме Горнера, с вынесением общих множителей:

$$\mathbf{f}(\mathbf{X}) = ((\mathbf{X} - 2)\mathbf{X} + 1)\mathbf{X} - 2.$$

В силу субдистрибутивности (см. § 1.5) такой вид выражения для полинома позволяет вычислять, вообще говоря, более узкие интервальные оценки.

Для интервала  $[-3, 3]$  построенное выражение равно  $[-50, 46]$ , т. е. содержит нуль, но вот для его левой половины  $[-3, 0]$  справедливо

$$f([-3, 0]) = [-50, -2] \not\ni 0,$$

так что решений уравнения на интервале  $[-3, 0]$  быть не может.

Суженный интервал локализации решения  $[0, 3]$  можно исследовать и далее. Последовательно отсекая от него подинтервалы слева и справа и вычисляя на них интервальные оценки области значений, получим

$$f([0, 1.5]) = [-5, -0.5] \not\ni 0,$$

$$f([1.5, 1.9]) = [-1.925, -0.385] \not\ni 0,$$

$$f([2.1, 3]) = [0.541, 10] \not\ni 0.$$

Таким образом, все интервалы, выписанные выше аргументами интервального расширения  $f$ , не содержат решений уравнения, и потому интервал локализации решения сузился до  $[1.9, 2.1]$ . Ясно, что этот процесс уточнения двусторонних оценок решения можно алгоритмизовать и продолжить. Хорошим приближением к искомому решению будет середина интервала локализации.

Точное значение вещественного решения рассматриваемого уравнения, как нетрудно проверить, равно 2. ■

Другие интервальные приёмы локализации и уточнения решений уравнений читатель может увидеть в § 4.7.

#### 4.4б Метод половинного деления (бисекции)

Этот численный метод заключается в последовательном делении пополам (или иногда на большее число частей) интервала локализации решения уравнения и последующей проверке (и возможной отбраковке) полученных подинтервалов. Слово «бисекция», часто используемое для его названия, — латинизм, означающий деление на две части, но иногда можно встретить и другие термины — «метод дихотомии» [8] или даже «метод вилки» [5].

Теоретической основой метода половинного деления является следующий факт, хорошо известный в математическом анализе:

**Теорема 4.4.3** (теорема Больцано–Коши) *Если функция  $f : \mathbb{R} \rightarrow \mathbb{R}$  непрерывна на интервале  $X \subset \mathbb{R}$  и на его концах принимает значения*

Таблица 4.1. Метод половинного деления для решения уравнений

```

 $\underline{x} \leftarrow a; \quad \bar{x} \leftarrow b;$ 
DO WHILE ( $\bar{x} - \underline{x} > \epsilon$ )
     $\mu \leftarrow \frac{1}{2}(\underline{x} + \bar{x});$ 
    IF ( $f(\underline{x}) < 0$  и  $f(\mu) > 0$ ) или ( $f(\underline{x}) > 0$  и  $f(\mu) < 0$ )
         $\bar{x} \leftarrow \mu$ 
    ELSE
         $\underline{x} \leftarrow \mu$ 
    END IF
END DO

```

разных знаков, то внутри интервала  $X$  существует нуль функции  $f$ , т. е. точка  $\tilde{x} \in X$ , в которой  $f(\tilde{x}) = 0$ .

Часто её называют просто «теоремой Больцано» [39], так как именно Б. Больцано первым обнаружил это замечательное свойство непрерывных функций.

Очевидно, что из двух половин интервала, на котором функция меняет знак, хотя бы на одной эта переменна знака обязана сохраняться. Её мы и оставляем в результате очередной итерации метода половинного деления, а затем снова подвергаем дроблению и исследованию, и т. д.

На вход алгоритму подаются функция  $f$ , принимающая на концах интервала  $[a, b]$  значения разных знаков, и точность  $\epsilon$ , с которой необходимо локализовать решение уравнения  $f(x) = 0$ . На выходе получаем интервал  $[\underline{x}, \bar{x}]$  шириной не более  $\epsilon$ , содержащий решение уравнения. Псевдокод алгоритма представлен в табл. 4.1.

Недостаток этого простейшего варианта метода половинного деления — возможность потери нулей для функций, аналогичных изображённой на рис. 4.13. На левой половине исходного интервала  $x^{(0)}$  функция не меняет знак, но там находятся два нуля функции. Чтобы убедиться в единственности решения или в его отсутствии, можно привлекать дополнительную информацию об уравнении, к примеру, о

производной функции из левой части. Если она сохраняет знак на рассматриваемом интервале (функция монотонна), то наличие решения равносильно перемене знака значений, и решение единствено.

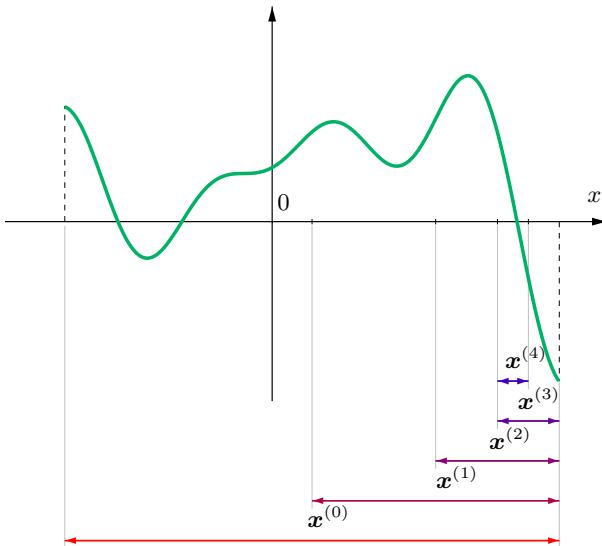


Рис. 4.13. Иллюстрация работы метода половинного деления (бисекции)

В общем случае потери нулей можно также избежать, если не отбрасывать подинтервалы, на которых доказательно не установлено отсутствие решений, а сохранять их и дополнительно обрабатывать дальше, чтобы более определённо установить их статус. Последовательная реализация этой идеи с привлечением инструментов интервального анализа приводит к «методу ветвлений и отсечений», который подробно рассматривается далее в § 4.8.

Многомерное обобщение теоремы Больцано–Коши было найдено и опубликовано более чем столетием позже в заметке [43]:

#### Теорема 4.4.4 (теорема Миранды)

Пусть  $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ ,  $f(x) = (f_1(x), f_2(x), \dots, f_n(x))^\top$  — функция, непрерывная на брусе  $\mathbf{X} \subset \mathbb{R}^n$  с гранями, параллельными координат-

ным осям, и для любого  $i = 1, 2, \dots, n$  имеет место либо

$$f_i(\mathbf{X}_1, \dots, \mathbf{X}_{i-1}, \underline{\mathbf{X}}_i, \mathbf{X}_{i+1}, \dots, \mathbf{X}_n) \leq 0$$

$$\text{и } f_i(\mathbf{X}_1, \dots, \mathbf{X}_{i-1}, \overline{\mathbf{X}}_i, \mathbf{X}_{i+1}, \dots, \mathbf{X}_n) \geq 0,$$

либо

$$f_i(\mathbf{X}_1, \dots, \mathbf{X}_{i-1}, \underline{\mathbf{X}}_i, \mathbf{X}_{i+1}, \dots, \mathbf{X}_n) \geq 0$$

$$\text{и } f_i(\mathbf{X}_1, \dots, \mathbf{X}_{i-1}, \overline{\mathbf{X}}_i, \mathbf{X}_{i+1}, \dots, \mathbf{X}_n) \leq 0,$$

т. е. области значений каждой компоненты функции  $f(x)$  на соответствующих противоположных гранях бруса  $\mathbf{X}$  имеют разные знаки. Тогда на брусе  $\mathbf{X}$  существует нуль функции  $f$ , т. е. точка  $x^* \in \mathbf{X}$ , в которой  $f(x^*) = 0$ .

Характерной особенностью теоремы Миранды является специальная форма множества, на котором утверждается существование нуля функции: оно должно быть бруском с гранями, параллельными координатным осям, т. е. интервальным вектором. Для полноценного применения теоремы Миранды нужно уметь находить или как-то оценивать области значений функций на таких бру сах. Удобное средство для решения этой задачи представляют методы интервального анализа. Задача об определении области значений функции эквивалентна задаче оптимизации, но в интервальном анализе, когда областями определения являются интервальные векторы-брюсы, внешнее оценивание области значений принимает специфическую форму задачи о вычислении так называемого *интервального расширения функции* (см. § 1.6). Таким образом, для проверки неравенств из условия теоремы Миранды необходимо взять интервальные расширения функции  $f(x)$  на гранях бруса  $\mathbf{X}$ , которых  $2n$  штук.

Опираясь на теорему Миранды и интервальные вычисления, можно организовать многомерный вариант метода половинного деления. Но при этом следует иметь в виду, что из-за неизбежных погрешностей интервальных оценок на некоторых бру сах определить наличие нулей функции или же их отсутствие будет невозможно, так как неравенства из теоремы Миранды не будут выполнены, как и неравенства, означающие отсутствие решений. Чтобы не упустить эти решения, необходимо сохранять все такие брусы с неопределенным статусом, полученные в ходе дробления, пока их дальнейшая обработка на следующих шагах

алгоритма не прояснит наверняка наличие или отсутствие в них решений. В целом, как и в одномерном случае, удобно оформить вычислительную схему в виде «метода ветвлений и отсечений», рассматриваемого в § 4.8.

#### 4.4в Метод простой итерации

*Методом простой итерации* обычно называют стационарный одноступенчатый итерационный процесс, который организуется после того, как исходное уравнение  $f(x) = 0$  каким-либо способом приведено к равносильному рекуррентному виду  $x = \Phi(x)$ . Далее, выбрав некоторое начальное приближение  $x^{(0)}$ , запускаем итерирование

$$x^{(k)} \leftarrow \Phi(x^{(k-1)}), \quad k = 1, 2, \dots$$

При благоприятных обстоятельствах конструируемая последовательность  $\{x^{(k)}\}$  сходится, и её пределом является неподвижная точка отображения  $\Phi$ , т. е. решение исходного уравнения. Часто эту схему называют также *методом последовательных приближений*. В общем случае и характер сходимости, и вообще её наличие существенно зависят как от отображения  $\Phi$ , так и от начального приближения к решению.

**Пример 4.4.4** Уравнение (4.19) из примера 4.4 нетрудно привести к рекуррентному виду

$$\alpha = \frac{l}{h} \sin \alpha,$$

где  $l = 3.3$  и  $h = 3$ . Далее, взяв в качестве начального приближения, например,  $\alpha^{(0)} = 1$ , через 50 итераций вида

$$\alpha^{(k)} \leftarrow \frac{l}{h} \sin \alpha^{(k-1)}, \quad k = 1, 2, \dots, \quad (4.21)$$

получаем пять верных знаков точного решения  $\alpha^* = 0.748986642697\dots$  (читатель легко может самостоятельно проверить все числовые данные этого примера с помощью любой системы компьютерной математики). Это ответ в радианах, что в более привычных угловых градусах соответствует примерно  $42.914^\circ$ . Общий угловой размер кругового сектора крыши, следовательно, равен  $85.828^\circ$ .

Итерационный процесс (4.21) сходится к решению  $\alpha^*$  не из любого начального приближения. Если  $\alpha^{(0)} = \pi l$ ,  $l \in \mathbb{Z}$ , то выполнение итераций (4.21) с идеальной точностью даёт  $\alpha^{(k)} = 0$ ,  $k = 1, 2, \dots$ . Если

же  $\alpha^{(0)}$  таково, что синус от него отрицателен, то итерации (4.21) сходятся к решению  $(-\alpha^*)$  уравнения (4.19). И нулевое, и отрицательное решения очевидно не имеют содержательного смысла.

С другой стороны, переписывание исходного уравнения (4.21) в другом рекуррентном виде —

$$\alpha = \frac{1}{l} \arcsin(\alpha h)$$

— приводит к тому, что характер сходимости метода простой итерации совершенно меняется. Из любого начального приближения, меньшего по модулю чем примерно 0.226965, итерации

$$\alpha^{(k)} \leftarrow \frac{1}{l} \arcsin(\alpha^{(k-1)} h), \quad k = 1, 2, \dots,$$

сходятся лишь к нулевому решению. Большие по модулю начальные приближения быстро выводят за границы области определения вещественного арксинуса, переводя итерации в комплексную плоскость, где они снова сходятся к нулевому решению. Таким образом, искомого решения  $\alpha^*$  мы при этом никак не получаем. ■

Рассмотренный пример хорошо иллюстрирует различный характер неподвижных точек отображений и мотивирует следующие определения.

Неподвижная точка  $x^*$  функции  $\Phi(x)$  называется *притягивающей* (или точкой притяжения), если существует такая окрестность  $\Omega$  этой точки  $x^*$ , что итерационный процесс  $x^{(k)} \leftarrow \Phi(x^{(k-1)})$ ,  $k = 1, 2, \dots$ , сходится к  $x^*$  из любого начального приближения  $x^{(0)} \in \Omega$ .

Неподвижная точка  $x^*$  функции  $\Phi(x)$  называется *отталкивающей*, если существует такая окрестность  $\Omega$  точки  $x^*$ , что итерационный процесс  $x^{(k)} \leftarrow \Phi(x^{(k-1)})$ ,  $k = 1, 2, \dots$ , не сходится к  $x^*$  при любом начальном приближении  $x^{(0)} \in \Omega$ .

Ясно, что простые итерации  $x^{(k)} \leftarrow \Phi(x^{(k-1)})$  непригодны для нахождения отталкивающих неподвижных точек. Здесь возникает интересный вопрос о том, какими преобразованиями уравнений и систем уравнений отталкивающие точки можно сделать притягивающими.

Наиболее часто существование притягивающих неподвижных точек можно гарантировать у отображений, которые удовлетворяют тем или иным дополнительным условиям. Самыми популярными из них являются так называемые условия сжимаемости (сжатия) образа, которые

обсуждались в § 4.3e. Нетрудно понять, что достаточным условием того, что неподвижная точка — притягивающая, является сжимаемость отображения  $\Phi$  в окрестности этой точки.

Пусть  $x^*$  — простое решение уравнения  $f(x) = 0$ , так что  $f'(x^*) \neq 0$ . Тогда производная по  $x$  от выражения

$$x - f(x)/f'(x^*)$$

зануляется в  $x^*$ , а потому существует окрестность  $\Omega$  точки  $x^*$ , в которой эта производная строго меньше единицы. Следовательно, в силу условия (4.15) отображение  $\Phi$ , задаваемое как

$$\Phi(x) = x - \frac{1}{f'(x^*)} f(x),$$

является сжимающим на  $\Omega$ . Естественно, на практике мы не знаем точного решения, но вместо  $x^*$  с таким же успехом можно брать точку, достаточно близкую к решению, которая найдена, например, с помощью грубой прикидки или графическим способом.

#### Пример 4.4.5 Рассмотрим уравнение

$$x^2 + \sin(3x) - 1 = 0 \quad \text{на интервале } [-2, 2].$$

Оно имеет 4 решения, одно отрицательное и три положительных, равные примерно  $-1$ ,  $0.4$ ,  $1.2$  и  $1.3$  (рис. 4.14 изображает график функции из левой части уравнения).

Преобразуя выражение из левой части, можем переписать уравнение в различных рекуррентных формах:

$$x = \sqrt{1 - \sin(3x)}, \quad x = \frac{1 - \sin(3x)}{x}, \quad x = \frac{1}{3} \arcsin(1 - x^2), \quad \dots$$

Экспериментально нетрудно обнаружить, что первая и вторая формулы не обеспечивают сходимость метода простой итерации, и только для третьей формулы итерации сходятся к одному решению  $x_2 = 0.35462$ , причём промежуточные приближения выходят при этом в комплексную плоскость. Можно ли найти остальные решения уравнения?

Чтобы организовать сходящиеся итерационные формулы, воспользуемся другим рецептом, который изложен перед нашим примером.

Предположим, что нам нужно уточнить самое левое (т. е. наименьшее) решение уравнения, приблизительно равное  $-1$ , о существовании

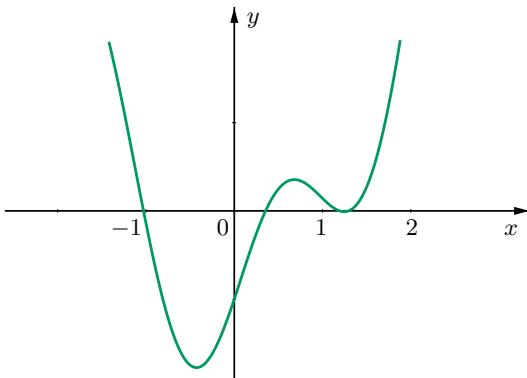


Рис. 4.14. График функции  $y = x^2 + \sin(3x) - 1$

которого и примерном расположении можно получить информацию из предварительно построенного графика (рис. 4.14). Организуем рекуррентную форму решаемого уравнения в виде

$$x = x - \Lambda f(x),$$

где  $f(x) = x^2 + \sin(3x) - 1$  — левая часть решаемого уравнения,  $\Lambda = 1/f'(\tilde{x}) \approx -0.97$ ,  $\tilde{x} = 1$  — приближение к решению. При выборе любого начального приближения из интервала  $[-5, 0]$  соответствующий итерационный процесс

$$x^{(k)} \leftarrow x^{(k-1)} + 0.97 f(x^{(k-1)}), \quad k = 1, 2, \dots,$$

сходится к решению  $x_1 = -1.028153788714888$ . Аналогично — с другими решениями уравнения. ■

Несмотря на успешное нахождение всех решений уравнения в рассмотренном примере, это потребовало кропотливой предварительной работы, фактически, «ручной настройки» метода. У читателя может естественно возникнуть вопрос о её целесообразности в каких-то практических ситуациях. Рассматриваемые в следующих разделах методы решения уравнений имеют заметно большие удобство и эффективность, но метод простых итераций также обладает рядом хороших свойств и полезных особенностей, из-за которых списывать со счетов

его не стоит. В силу теоремы Банаха о неподвижной точке он, например, обеспечивает доказательство единственности решения для уравнений в рекуррентной форме, имеющих сжимающие отображения.

#### 4.4г Интерполяционные методы

В главе 2 для решения задач численного дифференцирования и интегрирования с успехом применялась алгебраическая интерполяция, которая помогала заменить исходную задачу на близкую, но более простую и решаемую в явном виде. Похожую идею можно использовать также при решении уравнений, интерполируя входящие в них функции и решая затем алгебраические уравнения.

Но реализация этой идеи сталкивается с принципиальным ограничением, связанным с возможностью эффективного решения полиномиальных уравнений. Теорема Абеля–Руффини утверждает, что невозможно с помощью явных формул решать алгебраические уравнения степени выше пятой. Но для четвёртой, и даже для третьей степеней эти формулы для решений громоздки и не очень удобны для быстрого применения в качестве составной части более сложных численных методов. Тем не менее высказанную идею вполне можно реализовать для линейного и квадратичного интерполянтов. При этом получаются метод секущих и метод парабол (метод Мюллера), которые очень популярны в практических вычислениях.

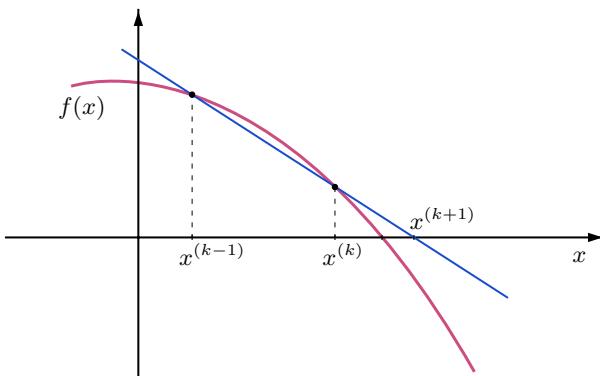


Рис. 4.15. Графическая иллюстрация итерации метода секущих

Пусть необходимо решить уравнение  $f(x) = 0$ . В *методе секущих* конструируется последовательность приближений, в которой каждые два последовательных члена служат для построения линейного интерполянта функции  $f$ , а следующее приближение получается как решение соответствующего линейного уравнения. Если заданы  $x^{(k-1)}$  и  $x^{(k)}$ , то алгебраический интерполант первой степени для функции  $f(x)$  по этим узлам имеет вид

$$P_1(x) = f(x^{(k)}) + f'(x^{(k)}, x^{(k-1)})(x - x^{(k)}),$$

где  $f'(x^{(k)}, x^{(k-1)})$  — разделённая разность функции  $f$  между точками  $x^{(k)}$  и  $x^{(k-1)}$  (см. § 2.2д). Решение  $\tilde{x}$  уравнения  $P_1(x) = 0$  выражается формулой

$$\tilde{x} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)}, x^{(k-1)})}.$$

Как итог, расчётные формулы метода секущих выглядят следующим образом:

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

Из них следует, что метод секущих — двухшаговый, так что для начала его работы необходимо задать два первых приближения  $x^{(0)}$  и  $x^{(1)}$ . Кроме того, программа метода секущих должна уметь преодолевать исключительный случай, когда  $f(x^{(k)}) = f(x^{(k-1)})$  и знаменатель дроби из выражения в расчётной формуле зануляется. Тогда можно, например, немного изменить  $x^{(k)}$  и/или  $x^{(k-1)}$ .

Иногда описанный выше численный метод называют также *методом хорд*. Но нередко термины «метод секущих» и «метод хорд» используют и в других смыслах, обозначая ими способы решения уравнений, основанные на идее линейной интерполяции, но в которых учитываются знаки функции в очередных приближениях [3, 8, 12].

**Пример 4.4.6** Решим с помощью метода секущих уравнение (4.19), которое перепишем в виде

$$l \sin \alpha - \alpha h = 0,$$

с  $l = 3.3$  и  $h = 3$ . Оно было получено во введении к § 4.4 и решено методом простой итерации в § 4.4в.

Если одно из двух начальных приближений — нулевое или близко к нулю, то метод сходится к нулевому решению уравнения. Но если оба приближения положительны и заметно отличны от нуля, получаем интересующее нас решение 0.748986642697341.

Например, при запуске метода секущих из начальных приближений  $\alpha^{(0)} = 1$  и  $\alpha^{(1)} = 2$  он быстро сходится к выписанному выше точному решению, причём пять верных знаков достигаются уже после шести шагов. Все 15 значащих цифр получаются после 9 шагов. Существенное увеличение эффективности в сравнении с примером 4.4.4! ■

В *методе парабол* уравнение  $f(x) = 0$  заменяется на каждом шаге квадратным уравнением, с квадратным трёхчленом, который интерполирует функцию  $f(x)$  по трём точкам. Метод парабол часто называют также *методом Мюллера* по имени американского математика, который предложил его в 1956 году [70].

Пусть уже известны три последние приближения  $x^{(k-2)}$ ,  $x^{(k-1)}$  и  $x^{(k)}$  к решению уравнения. Построим по ним для функции  $f(x)$  интерполяционный квадратный трёхчлен в форме Ньютона (см. § 2.2д):

$$\begin{aligned} P_2(x) = & f(x^{(k)}) + f^{\wedge}(x^{(k)}, x^{(k-1)})(x - x^{(k)}) + \\ & + f^{\wedge}(x^{(k)}, x^{(k-1)}, x^{(k-2)})(x - x^{(k)})(x - x^{(k-1)}). \end{aligned}$$

Ясно, что

$$\begin{aligned} P_2(x) = & f(x^{(k)}) + f^{\wedge}(x^{(k)}, x^{(k-1)})(x - x^{(k)}) + \\ & + f^{\wedge}(x^{(k)}, x^{(k-1)}, x^{(k-2)})(x - x^{(k)})( (x - x^{(k)}) + (x^{(k)} - x^{(k-1)})) \\ = & f(x^{(k)}) + \\ & + (f^{\wedge}(x^{(k)}, x^{(k-1)}) + (x^{(k)} - x^{(k-1)}) f^{\wedge}(x^{(k)}, x^{(k-1)}, x^{(k-2)})) \cdot \\ & \cdot (x - x^{(k)}) + \\ & + f^{\wedge}(x^{(k)}, x^{(k-1)}, x^{(k-2)})(x - x^{(k)})^2. \end{aligned}$$

Приравнивая последнее выражение нулю, получим квадратное уравнение, которое удобно записать относительно новой переменной  $y := x - x_k$ :

$$ay^2 + by + c = 0,$$

где

$$\begin{aligned} a &= f''(x^{(k)}, x^{(k-1)}, x^{(k-2)}), \\ b &= f'(x^{(k)}, x^{(k-1)}) + (x^{(k)} - x^{(k-1)}) a, \\ c &= f(x^{(k)}). \end{aligned}$$

Формулы решений этого вспомогательного уравнения хорошо известны из школьного курса алгебры:

$$y_{1,2}^* = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Они могут комплексными, если дискриминант  $b^2 - 4ac$  отрицателен, и это вполне нормальная ситуация для метода Мюллера. После появления какого-то комплексного приближения дальнейшие итерации, скорее всего, останутся комплексными, что не должно сильно смущать: если исходное уравнение имеет вещественное решение, то комплексные последовательные приближения метода Мюллера всё равно будут сходиться к нему. Для полноценной реализации метода Мюллера, таким образом, требуется комплексный тип данных и комплексная арифметика.

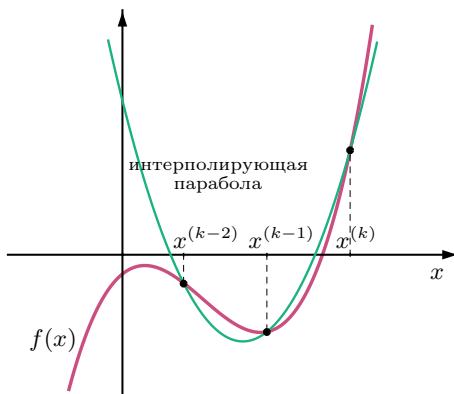


Рис. 4.16. Наглядная иллюстрация идеи метода Мюллера

Обычно берут меньшее по модулю решение  $y^*$ , так как ему соответствует новое приближение к решению, которое ближе к текущему

$x^{(k)}$ . В целом в методе Мюллера следующее приближение к решению исходного уравнения определяется как

$$x^{(k+1)} = x^{(k)} + y^*.$$

Поскольку для расчёта очередного шага необходимо задать три точки  $x^{(k-2)}$ ,  $x^{(k-1)}$  и  $x^{(k)}$ , то итерационный метод Мюллера является трёхшаговым. Начальные точки, необходимые для обычно запуска итераций, выбирают случайно, либо равномерно покрывая интервал локализации решения.

При реализации метода Мюллера могут встречаться исключительные ситуации. Если значения функции из левой части уравнения в каких-то трёх последовательных приближениях совпадают, то метод Мюллера не может быть продолжен, поскольку через три таких точки вместо параболы можно провести лишь прямую, параллельную оси абсцисс. Программа, реализующая метод Мюллера, должна уметь преодолевать описанную проблему, например, с помощью небольшого «шагования» приближений.

Замечательное качество метода Мюллера состоит в том, что он способен находить комплексные решения полиномиальных уравнений с вещественными коэффициентами. Это особенно полезно, например, при вычислении комплексных собственных значений вещественных матриц (см. § 3.19г).

**Пример 4.4.7** Рассмотрим уравнение

$$x^3 - 2x^2 + x - 2 = 0.$$

Поскольку

$$x^3 - 2x^2 + x - 2 = (x - 2)(x^2 + 1) = (x - 2)(x - i)(x + i),$$

то ясно, что в поле комплексных чисел рассматриваемое уравнение имеет три решения — 2 и  $\pm i$ .

Если начальными точками взять тройку 1, 2, 3, то почти при любой точности остановки получаем вещественное решение 2, к которому метод Мюллера сходится за 4 итерации.

Если начальными точками взять тройку 3, 4, 5, то метод Мюллера сходится к тому же решению 2, но в ответе появляется очень маленькая мнимая часть. Это объясняется тем, что промежуточные приближения

были существенно комплексными, т. е. метод шёл к вещественному решению уравнения через комплексную плоскость.

Наконец, если начальными точками взять тройку  $-1, 0, 1$ , то метод Мюллера сходится к комплексному решению  $0 + i$ . Поскольку исходное полиномиальное уравнение имеет вещественные коэффициенты, то его комплексные решения должны присутствовать сопряжёнными парами. Отсюда можем сделать вывод, что уравнение, помимо найденного, должно также иметь решение  $0 - i$ . ■

Последнее соображение примера очень важно при практическом применении метода Мюллера к полиномиальным уравнениям с вещественными коэффициентами, так как позволяет существенно экономить усилия при их решении.

Метод секущих и метод Мюллера находят решения уравнений лишь при условии, что начальные приближения «достаточно близки» к этим решениям. Таким образом, эти методы носят локальный характер и сами по себе не предназначены для нахождения *всех* решений.

#### 4.4д Метод Ньютона и его модификации

Предположим, что для уравнения  $f(x) = 0$  с вещественной функцией  $f$  известно некоторое приближение  $\tilde{x}$  к решению  $x^*$ . Если  $f$  — дифференцируемая, то можно приблизить её в окрестности точки  $\tilde{x}$  линейной функцией, как

$$f(x) \approx f(\tilde{x}) + f'(\tilde{x})(x - \tilde{x}).$$

Для вычисления следующего приближения к  $x^*$  решим тогда линейное уравнение

$$f(\tilde{x}) + f'(\tilde{x})(x - \tilde{x}) = 0,$$

приближающее исходное уравнение. Иными словами, следующее приближение к решению уравнения естественно взять в виде

$$\tilde{x} - \frac{f(\tilde{x})}{f'(\tilde{x})}.$$

Итерационный метод решения уравнения, основанный на его локальной линеаризации в точке очередного приближения и выполняемый по правилу

$$x^{(k)} \leftarrow x^{(k-1)} - \frac{f(x^{(k-1)})}{f'(x^{(k-1)})}, \quad k = 1, 2, \dots,$$

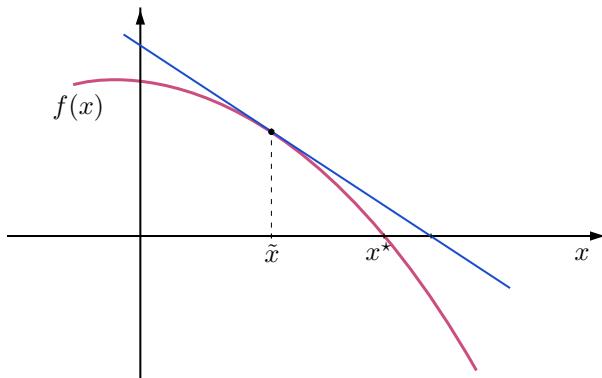


Рис. 4.17. Иллюстрация итерации метода Ньютона

называют *методом Ньютона*. Он является одним из популярнейших и наиболее эффективных численных методов решения уравнений и имеет многочисленные обобщения, в том числе на многомерный случай, т. е. в применении к решению систем уравнений (см. § 4.7в).

Оценим быстроту сходимости метода Ньютона к простому решению уравнения. Разложим по формуле Тейлора функцию  $f$  в  $x^{(k-1)}$ :

$$0 = f(x^*) = f(x^{(k-1)}) + f'(x^{(k-1)})(x^* - x^{(k-1)}) + \frac{1}{2} f''(\xi) (x^* - x^{(k-1)})^2,$$

где  $\xi$  лежит между  $x^*$  и  $x^{(k-1)}$ . Следовательно,

$$\frac{f(x^{(k-1)})}{f'(x^{(k-1)})} = x^{(k-1)} - x^* - \frac{1}{2} \frac{f''(\xi)}{f'(x^{(k-1)})} (x^* - x^{(k-1)})^2,$$

и поэтому в силу расчётной формулы метода Ньютона

$$x^{(k)} - x^* = x^{(k-1)} - x^* - \frac{f(x^{(k-1)})}{f'(x^{(k-1)})} = \frac{1}{2} \frac{f''(\xi)}{f'(x^{(k-1)})} (x^* - x^{(k-1)})^2.$$

Обозначим  $M_2 := \max_{x \in [\underline{x}, \bar{x}]} |f''(x)|$  и  $m_1 := \min_{x \in [\underline{x}, \bar{x}]} |f'(x)|$ , где  $[\underline{x}, \bar{x}]$  — интервал локализации решения и приближений к нему, на котором производная  $f'$  не меняет знак. Тогда

$$|x^{(k)} - x^*| \leq \frac{M_2}{2m_1} |x^{(k-1)} - x^*|^2, \quad k = 1, 2, \dots$$

В целом,

$$|x^{(k)} - x^*| \leq \frac{M_2}{2m_1} |x^{(0)} - x^*|^{2^k},$$

и это очень быстрая сходимость, которая называется *квадратичной сходимостью*. Она превосходит скорость сходимости геометрической прогрессии (её называют *линейной*), типичную для стационарных итерационных методов.

В англоязычной литературе метод Ньютона иногда называют также «методом Ньютона–Рафсона» [13, 21, 60], так как именно Дж. Рафсон придал современную форму тому способу, которым И. Ньютон решал полиномиальные уравнения. Окончательное оформление метод Ньютона получил в середине XVIII века у Т. Симпсона, который применял этот метод для произвольных, не обязательно алгебраических, уравнений и затем к системам двух уравнений с двумя неизвестными.

**Пример 4.4.8** Рассмотрим уравнение  $x^2 - a = 0$ , решением которого является квадратный корень из числа  $a$ . Если  $f(x) = x^2 - a$ , то  $f'(x) = 2x$ , так что в методе Ньютона для нахождения решения рассматриваемого уравнения имеем

$$\begin{aligned} x^{(k)} &= x^{(k-1)} - \frac{f(x^{(k-1)})}{f'(x^{(k-1)})} = \\ &= x^{(k-1)} - \frac{(x^{(k-1)})^2 - a}{2x^{(k-1)}} = \frac{x^{(k-1)}}{2} + \frac{a}{2x^{(k-1)}}. \end{aligned}$$

Итерационный процесс для нахождения  $\sqrt{a}$ , определяемый формулой

$$x^{(k)} \leftarrow \frac{1}{2} \left( x^{(k-1)} + \frac{a}{x^{(k-1)}} \right), \quad k = 1, 2, \dots,$$

известен ещё с античности и часто называется *методом Герона*. Для любого положительного начального приближения  $x^{(0)}$  он порождает убывающую, начиная с  $x^{(1)}$ , последовательность, которая быстро сходится к арифметическому значению  $\sqrt{a}$ . ■

Метод Ньютона требует вычисления на каждом шаге производной от функции  $f$ , что может оказаться трудным или вообще невозможным. Одна из очевидных модификаций метода Ньютона состоит в том,

чтобы «заморозить» производную в некоторой точке и вести итерации по формуле

$$x^{(k)} \leftarrow x^{(k-1)} - \frac{f(x^{(k-1)})}{f'(\tilde{x})}, \quad k = 1, 2, \dots,$$

где  $\tilde{x}$  — фиксированная точка, в которой берётся производная. Получаем стационарный итерационный процесс, который существенно проще в реализации, но он имеет качественно более медленную линейную сходимость. Фактически, это метод простой итерации, организованный с помощью конструкции из § 4.4в с  $\Lambda = 1/f'(\tilde{x})$ .

Несмотря на большую популярность метода Ньютона и его хорошие свойства, он является локальным, аналогично методу секущих и методу Мюллера из предыдущего раздела. В некоторых ситуациях поведение метода Ньютона является плохим или даже патологическим. Например, работа метода Ньютона ухудшается на кратных решениях уравнений. По мере приближения к такому решению производная становится всё меньше и меньше, и деление на неё приводит к неустойчивому счёту, а сам метод Ньютона замедляется.

Существуют относительно простые уравнения, в применении к которым метод Ньютона расходится. Это имеет место, в частности, при нахождении единственного нулевого решения уравнения  $x^{1/3} = 0$  из начальных приближений, не равных самому нулю, что можно показать несложными аналитическими выкладками. Дальнейшим развитием этого примера является выразительный

#### Пример 4.4.9 (пример Донована–Миллера–Морелэнда [64])

Рассмотрим численное решение, с помощью метода Ньютона, уравнения

$$h(x) = x^{1/3} e^{-x^2} = 0.$$

Оно имеет единственное решение, равное нулю, но метод Ньютона не сходится к нему ни из какого начального приближения, отличного от нуля. Другое замечательное свойство этого примера состоит в том, что он демонстрирует недостаток нередко используемого условия остановки итераций  $|x^{(k)} - x^{(k-1)}| < \epsilon$  для заданного малого порога  $\epsilon$ .

В самом деле,

$$h'(x) = \left(\frac{1}{3}x^{-2/3} - 2x^{4/3}\right)e^{-x^2},$$

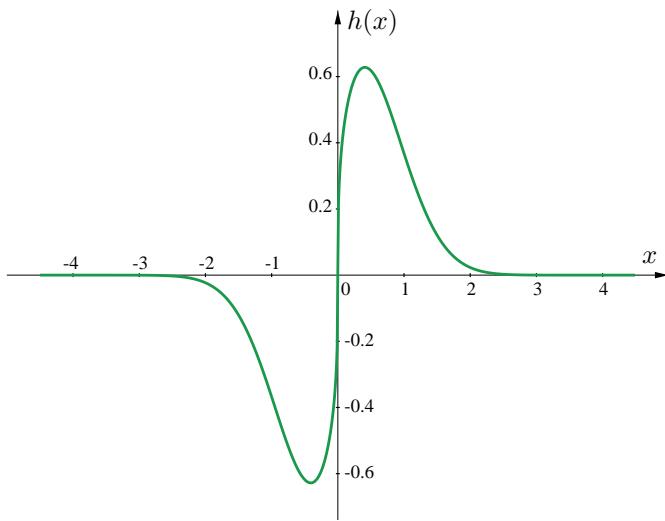


Рис. 4.18. График функции из примера Донована–Миллера–Морелэнда

и итерации метода Ньютона в данном случае имеют вид

$$x^{(k)} \leftarrow x^{(k-1)} - \frac{h(x^{(k-1)})}{h'(x^{(k-1)})} = x^{(k-1)} - \frac{x^{(k-1)}}{\frac{1}{3} - 2(x^{(k-1)})^2}, \quad (4.22)$$

$$k = 1, 2, \dots$$

Они определены всюду за исключением стационарных точек функции  $h(x)$  (точек зануления её производной), равных  $\pm 1/\sqrt{6}$ . Если  $x^{(k-1)} \in ]-1/\sqrt{6}, 1/\sqrt{6}[$  и  $x^{(k-1)} \neq 0$ , то

$$\left| \frac{x^{(k)}}{x^{(k-1)}} \right| = \left| 1 - \frac{1}{\frac{1}{3} - 2(x^{(k-1)})^2} \right| > 2.$$

Как следствие, каждое следующее приближение метода Ньютона (4.22) находится на расстоянии от нуля, более чем вдвое превышающем его для предыдущего приближения. Последовательность, порождаемая методом Ньютона с любым ненулевым начальным приближением, вместо сходимости «разбалтывается» и, в конце концов, выходит за пределы интервала  $[-1/\sqrt{6}, 1/\sqrt{6}]$ .

Дальнейший анализ ситуации для  $x^{(k-1)} \notin [-1/\sqrt{6}, 1/\sqrt{6}]$  достаточно провести лишь для случая  $x^{(k-1)} > 0$ , т. е. когда

$$x^{(k-1)} > \frac{1}{\sqrt{6}}.$$

Для отрицательных  $x^{(k-1)}$  рассуждения будут аналогичными из-за того, что функция  $h(x)$  — нечётная.

При  $x^{(k-1)} > 1/\sqrt{6}$  имеем  $\frac{1}{3} - 2(x^{(k-1)})^2 < 0$ , а потому в итерациях метода Ньютона (4.22) последовательные приближения монотонно возрастают, т. е.  $x^{(k)} > x^{(k-1)}$ . Если предположить, что для заданного  $\epsilon > 0$  условие остановки  $|x^{(k)} - x^{(k-1)}| < \epsilon$  никогда не выполняется, то последовательные итерации (4.22) отличаются друг от друга не менее, чем на  $\epsilon$ , и потому  $x^{(k+m)} \geq x^{(k-1)} + (m+1)\epsilon$ . Ясно, что при достаточно больших  $m$  правая часть этого неравенства может быть сделана сколь угодно большой, что означает  $x^{(k)} \rightarrow \infty$  при неограниченном росте  $k$ .

С другой стороны, из расчётной формулы (4.22) следует, что

$$|x^{(k)} - x^{(k-1)}| = \left| \frac{x^{(k-1)}}{\frac{1}{3} - 2(x^{(k-1)})^2} \right|.$$

Выражение в правой части этого равенства должно стремиться к нулю при  $x^{(k)} \rightarrow \infty$ , что противоречит нашему допущению  $|x^{(k)} - x^{(k-1)}| \geq \epsilon$ .

Следовательно, для какого-то номера в итерациях (4.22) будет выполнено условие остановки  $|x^{(k)} - x^{(k-1)}| < \epsilon$ . Но настоящего решения уравнения мы не получим, так как функция  $h(x)$  не зануляется ни при каких  $x > 0$ , хотя и может принимать очень малые значения. ■

Интересно, что интервальный метод Ньютона (см. § 4.7б), который реализован с помощью интервальной арифметики Кэхэна, аккуратно выполняющей деление на интервалы с нулём, справляется с примером Донована–Миллера–Морелэнда. Этот и другие аналогичные примеры подробно рассматривается в работе [63].

#### 4.4e Методы Чебышёва

Методы Чебышёва для решения уравнения  $f(x) = 0$  основаны на разложении по формуле Тейлора функции  $f^{-1}$ , обратной к  $f$ . Они могут иметь произвольно высокий порядок точности, определяемый количеством членов разложения для  $f^{-1}$ , хотя практически обычно ограничиваются небольшими порядками. Переход к обратной функции

устраняет необходимость решения вспомогательного уравнения, как в методах секущих и Мюллера, и, как следствие, позволяет обойти трудности, вызванные теоремой Абеля–Руффини (см. § 4.4г).

Предположим, что на интервале  $[a, b]$  вещественная функция  $f$  является достаточно гладкой и монотонной, так что она взаимно однозначно отображает  $[a, b]$  в некоторый интервал  $[\alpha, \beta]$ . Как следствие, известные результаты математического анализа обеспечивают существование обратной к  $f$  функции  $g := f^{-1} : [\alpha, \beta] \rightarrow [a, b]$ , которая имеет ту же гладкость, что и функция  $f$ .

Итак, пусть известно некоторое приближение  $\tilde{x}$  к решению  $x^*$  уравнения  $f(x) = 0$ . Обозначив  $y = f(\tilde{x})$ , разложим обратную функцию  $g$  в точке  $y$  по формуле Тейлора с остаточным членом в форме Лагранжа:

$$\begin{aligned} g(0) &= g(y) + g'(y)(0 - y) + g''(y) \frac{(0 - y)^2}{2} + \dots + g^{(p)}(y) \frac{(0 - y)^p}{p!} \\ &\quad + g^{(p+1)}(\xi) \frac{(0 - y)^{p+1}}{(p+1)!} = \\ &= g(y) + \sum_{l=1}^p (-1)^l g^{(l)}(y) \frac{y^l}{l!} + (-1)^{p+1} g^{(p+1)}(\xi) \frac{y^{p+1}}{(p+1)!}, \end{aligned}$$

где  $\xi$  — какая-то точка между 0 и  $y$ . Возвращаясь к переменной  $x$ , будем иметь

$$x^* = \tilde{x} + \sum_{l=1}^p (-1)^l g^{(l)}(f(\tilde{x})) \frac{(f(\tilde{x}))^l}{l!} + (-1)^{p+1} g^{(p+1)}(\xi) \frac{(f(\tilde{x}))^{p+1}}{(p+1)!}.$$

В качестве следующего приближения к решению мы можем взять, отбросив остаточный член, значение

$$\tilde{x} + \sum_{l=1}^p (-1)^l g^{(l)}(f(\tilde{x})) \frac{(f(\tilde{x}))^l}{l!}.$$

Подытоживая сказанное, определим итерации

$$x^{(k)} \leftarrow x^{(k-1)} + \sum_{l=1}^p (-1)^l g^{(l)}(f(x^{(k-1)})) \frac{(f(x^{(k-1)}))^l}{l!}, \quad k = 1, 2, \dots,$$

которые называются *методом Чебышёва*  $p$ -го порядка.

Как на практике найти производные обратной функции  $g$ ?

Мы можем выразить их из известных значений производных функции  $f$ . В самом деле, последовательно дифференцируя тождество  $x = g(f(x))$ , получим

$$\begin{aligned} g'(f(x)) f'(x) &= 1, \\ g''(f(x)) (f'(x))^2 + g'(f(x)) f''(x) &= 0, \\ g'''(f(x)) (f'(x))^3 + g''(f(x)) \cdot 2f'(x)f''(x) + \\ &+ g''(f(x)) f'(x)f''(x) + g'(f(x)) f'''(x) = 0, \\ &\dots && \dots, \end{aligned}$$

или

$$\begin{aligned} g'(f(x)) f'(x) &= 1, \\ g''(f(x)) (f'(x))^2 + g'(f(x)) f''(x) &= 0, \\ g'''(f(x)) (f'(x))^3 + 3g''(f(x)) f'(x)f''(x) + g'(f(x)) f'''(x) &= 0, \\ &\dots && \dots \end{aligned}$$

Относительно неизвестных значений производных  $g'(f(x))$ ,  $g''(f(x))$ ,  $g'''(f(x))$  и т. д. эта система соотношений имеет специфическую форму, позволяющую найти их последовательно одну за другой, аналогично прямой подстановке для решения СЛАУ с нижними треугольными матрицами:

$$\begin{aligned} g'(f(x)) &= \frac{1}{f'(x)}, \\ g''(f(x)) &= -\frac{g(f(x)) f''(x)}{(f'(x))^2} = -\frac{f''(x)}{(f'(x))^3}, \\ g'''(f(x)) &= -\frac{3g''(f(x)) f'(x)f''(x) + g'(f(x)) f'''(x)}{(f'(x))^3} = \\ &= -3 \frac{(f''(x))^2}{(f'(x))^5} - \frac{f'''(x)}{(f'(x))^4} \end{aligned}$$

и так далее.

Для  $p = 1$  расчётные формулы метода Чебышёва имеют вид

$$x^{(k)} \leftarrow x^{(k-1)} - \frac{f(x^{(k-1)})}{f'(x^{(k-1)})}, \quad k = 1, 2, \dots,$$

что совпадает с методом Ньютона.

Для  $p = 2$  расчётные формулы метода Чебышёва таковы

$$x^{(k)} \leftarrow x^{(k-1)} - \frac{f(x^{(k-1)})}{f'(x^{(k-1)})} - \frac{f''(x^{(k-1)}) (f(x^{(k-1)}))^2}{2(f'(x^{(k-1)}))^3}, \quad k = 1, 2, \dots \quad (4.23)$$

Наиболее часто методом Чебышёва называют именно этот итерационный процесс, так как методы более высокого порядка из этого семейства на практике используются редко.

Отметим, что, аналогично методу Ньютона, на примере Донована–Миллера–Морелэнда из предыдущего раздела метод Чебышёва (4.23) расходится из любого ненулевого начального приближения. В этом легко убедиться, например, численными экспериментами.

#### 4.4ж Оценка погрешности приближённого решения

Важной частью процессов приближённого решения уравнений и их систем является оценивание погрешности, необходимое, например, для корректной остановки итерационных методов. Для систем линейных уравнений мы уже рассматривали этот вопрос в разделе § 3.15.

Отметим, что популярное условие остановки итерационных методов как достижение малости отличия двух последовательных приближений,  $|x^{(k)} - x^{(k-1)}|$ , часто может привести к преждевременному завершению итераций, при котором полученное приближение оказывается далёким от точного решения. Рассмотрим для примера последовательность частичных сумм гармонического ряда

$$S_n = \sum_{k=1}^n \frac{1}{k}, \quad n = 1, 2, \dots$$

Разность членов этой последовательности  $S_n - S_{n-1} = \frac{1}{n}$ , и она может быть сделана меньшей любого положительного числа при достаточно

большом  $n$ . Но гармонический ряд, как известно, расходится:  $S_n \rightarrow \infty$  при неограниченном росте  $n$ .

Конечно, последовательности приближений, порождаемые сходящимися итерационными процессами, являются другими по своей природе. Но даже для сходящейся последовательности  $\{x^{(k)}\}$  судить о близости к пределу по величине разности  $|x^{(k)} - x^{(k-1)}|$  можно лишь после поправки, вытекающей из учёта свойств этой последовательности.

**Пример 4.4.10** Пусть  $C$  — положительная константа, большая единицы, так что  $0 \leq 1 - 1/C \leq 1$ . Геометрическая прогрессия с первым членом 1 и знаменателем  $(1 - 1/C)$ , т. е.

$$x^{(k)} = (1 - 1/C)^k, \quad k = 0, 1, 2, \dots,$$

очевидно, сходится к нулю. Но разность её двух соседних членов

$$x^{(k)} - x^{(k-1)} = (1 - 1/C)^k - (1 - 1/C)^{k-1} = -(1 - 1/C)^{k-1}/C$$

в  $C$  раз меньше по абсолютной величине расстояния  $(k-1)$ -го члена до предела последовательности.

Например, случай  $C = 1000$ , когда различие достигает 1000 раз, получается для геометрической прогрессии со знаменателем 0.999. Это различие, в принципе, можно сделать вообще сколь угодно большим.

С другой стороны, оценив по нескольким членам геометрической прогрессии её знаменатель с помощью техники, описанной в § 3.15, можно рассчитать величину поправки, на которую нужно умножить различие между двумя членами, чтобы получить расстояние до предела. ■

Оценку погрешности приближения к решению уравнения  $f(x) = 0$  можно основывать на различных идеях и, по-видимому, самой простой является следующая. Если на концах некоторого интервала  $\mathbf{X} \subset \mathbb{R}$  функция  $f$  принимает значения разных знаков, то, как известно, по теореме Больцано–Коши внутри этого интервала обязательно должно быть решение уравнения. Тогда

в качестве приближённого решения лучше всего взять середину интервала,  $\text{mid } \mathbf{X}$ ,

радиус интервала  $\text{rad } \mathbf{X}$  является естественной мерой погрешности этого приближённого решения.

Обратим эти соображения. Пусть дано приближение  $\tilde{x}$  к решению уравнения  $f(x) = 0$ . Построим вокруг него интервал  $[\tilde{x} - \Delta, \tilde{x} + \Delta]$ , где  $\Delta$  — заданный малый порог, и затем проверим знаки функции  $f$  в концах построенного интервала, вычислив  $f(\tilde{x} - \Delta)$  и  $f(\tilde{x} + \Delta)$ . Если их знаки — разные, то приближение  $\tilde{x}$  имеет погрешность не более  $\Delta$ .

Для того, что чтобы получить описанным способом гарантированные (доказательные) результаты, нужно вычислять значения функции в точках  $\tilde{x} - \Delta$  и  $\tilde{x} + \Delta$  тоже с гарантией. Для этого, строго говоря, необходима машинная интервальная арифметика с внешним направленным округлением.

Ещё один популярный способ оценивания погрешности приближённого решения уравнения основан на использовании производной от входящей в него функции.

**Предложение 4.4.1** *Пусть для уравнения  $f(x) = 0$  точное решение равно  $x^*$ , дано приближение к нему  $\tilde{x}$ , и они лежат на интервале  $[a, b] \subset \mathbb{R}$ . Если  $f$  — непрерывно дифференцируемая функция, у которой производная не зануляется на  $[a, b]$ , то*

$$|\tilde{x} - x^*| \leq \frac{|f(\tilde{x})|}{\min_{\xi \in [a, b]} |f'(\xi)|}. \quad (4.24)$$

**Доказательство** следует из теоремы Лагранжа о среднем (формулы конечных приращений):

$$f(\tilde{x}) - f(x^*) = f'(\xi) \cdot (\tilde{x} - x^*),$$

где  $\xi$  — некоторая точка, заключённая между  $\tilde{x}$  и  $x^*$ . Ясно, что тогда

$$|f(\tilde{x}) - f(x^*)| \geq \min_{\xi \in [a, b]} |f'(\xi)| \cdot |\tilde{x} - x^*|,$$

и при  $\min_{\xi} |f'(\xi)| \neq 0$  получаем оценку (4.24). Отметим её очевидную аналогию с оценкой (3.197) для погрешности решения систем линейных уравнений: в обоих случаях невязка приближённого решения умножается на норму обратного отображения. ■

Формула (4.24) показывает, что чем меньше производная функции в окрестности решения, т. е. чем сильнее «стелется» функция около своего нуля, тем дальше может быть расстояние точки до этого нуля

при заданном абсолютном значении функции. Можно также сказать, что при малых значениях производной функции решение уравнения является «почти кратным» (рис. 4.1).

На практике нахождение точного минимума  $\min_{\xi} |f'(\xi)|$  по интервалу с решением может быть затруднительным, и тогда вместо него в (4.24) можно взять какую-нибудь положительную оценку снизу для области значений производной на  $[a, b]$ . Когда заранее известно, что вычисляемое решение не является кратным и потому производная  $f'$  не зануляется в нём, можно прибегнуть к приближённому способу оценки погрешности:

$$|\tilde{x} - x^*| \approx \frac{|f(\tilde{x})|}{|f'(\tilde{x})|},$$

если  $\tilde{x}$  достаточно близко к  $x^*$ .

## 4.5 Классические методы решения систем уравнений

### 4.5а Метод простой итерации

Схема применения метода простой итерации (метода последовательных приближений) для систем уравнений в принципе не отличается от случая одного уравнения, который был рассмотрен в § 4.4в. Исходная система уравнений  $F(x) = 0$  вида (4.2)–(4.3) должна быть каким-либо способом приведена к равносильному рекуррентному виду

$$x = \Phi(x)$$

и далее, после выбора некоторого начального приближения  $x^{(0)}$ , запускается стационарный итерационный процесс

$$x^{(k)} \leftarrow \Phi(x^{(k-1)}), \quad k = 1, 2, \dots \quad (4.25)$$

При благоприятных обстоятельствах последовательность  $\{x^{(k)}\}$  сходится, и её пределом является искомое решение системы уравнений.

Приведение исходной системы уравнений к рекуррентному виду может быть выполнено самыми различными способами. Во-первых, это могут быть аналитические преобразования уравнений системы, при которых в левых частях выделяются «в чистом виде» отдельные неизвестные переменные. Во-вторых, это может быть переход от системы

$F(x) = 0$  к системе вида

$$x = x - \Lambda F(x),$$

где  $\Lambda$  — неособенная матрица (предобуславливающая матрица). Иными словами, полагаем  $\Phi(x) = x - \Lambda F(x)$ .

Матрицу  $\Lambda$  во втором способе приведения нужно выбирать так, чтобы удовлетворялись (хотя бы локально) условия сходимости итераций (4.25), описываемые теоремой Банаха о неподвижной точке или же её аналогом — теоремой Шрёдера (см. § 4.4в). Матрица Якоби отображения  $\Phi$  есть  $I - \Lambda F'(x)$ , и она примерно равна матрице Липшица для  $\Phi$ , если  $F$  является гладким. Следовательно, желательно сделать матрицу  $I - \Lambda F'(x)$ , по-возможности, меньшей, чтобы уменьшить её спектральный радиус (оцениваемый сверху любой матричной нормой). В окрестности решения системы  $x^*$  локально наилучшим выбором является поэтому  $\Lambda = (F'(x^*))^{-1}$ , при котором матрица  $I - \Lambda F'(x)$  близка к нулевой недалеко от  $x^*$ .

**Пример 4.5.1** Рассмотрим решение системы уравнений

$$\begin{cases} x_1^2 + x_2^2 - 4 = 0, \\ x_1^3 - x_1 - x_2 - 1 = 0. \end{cases} \quad (4.26)$$

Первое уравнение системы задаёт в плоскости  $0x_1x_2$  окружность радиуса 2 с центром в начале координат, второе — график кубического полинома. Эти кривые имеют два пересечения, соответствующие двум решениям системы уравнений, примерно равным  $(-1.22, -1.59)$  и  $(1.56, 1.25)$ .

Чтобы переписать систему (4.26) в рекуррентном виде, выразим  $x_1$  из второго уравнения, а  $x_2$  — из первого. Получим

$$\begin{cases} x_1 = (x_1 + x_2 + 1)^{1/3}, \\ x_2 = \sqrt{4 - x_1^2}. \end{cases}$$

Таким образом, можно организовать итерационный процесс

$$\begin{aligned} x_1^{(k)} &= \left( x_1^{(k-1)} + x_2^{(k-1)} + 1 \right)^{1/3}, \\ x_2^{(k)} &= \sqrt{4 - \left( x_1^{(k-1)} \right)^2}, \quad k = 1, 2, \dots \end{aligned} \quad (4.27)$$

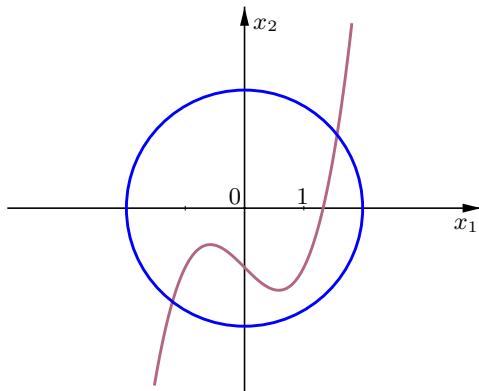


Рис. 4.19. Графики функций из левых частей уравнений системы (4.26)

Из начального приближения  $x_1^{(0)} = x_2^{(0)} = 1$  он сходится к положительному решению системы, и за 40 шагов мы получаем установление итераций на

$$x_1^* = 1.562003397594915, \quad x_2^* = 1.249057799263885.$$

К этому же решению рассматриваемый итерационный метод сходится из любого другого начального приближения, но получающиеся при этом последовательные приближения к решению могут выходить в комплексную плоскость. Но второе решение системы,  $(-0.81, -1.53)$ , с помощью итераций (4.27) мы никогда не получим.

Чтобы найти это отрицательное решение, воспользуемся альтернативным способом приведения системы к рекуррентному виду. Обозначим  $x := (x_1, x_2)^\top$  и

$$F(x) := \begin{pmatrix} x_1^2 + x_2^2 - 4 \\ x_1^3 - x_1 - x_2 - 1 \end{pmatrix}$$

— вектор-функцию из левых частей системы (4.26). Пусть также

$$A = \frac{1}{6} \begin{pmatrix} -1 & 2 \\ -2 & -2 \end{pmatrix}$$

— обратная к матрице Якоби отображения  $F$  в точке  $(-1, -1)^\top$ . Тогда метод простой итерации (4.25) с  $\Phi(x) = x - AF(x)$  из начального

приближения  $(-1, -1)^\top$  сходится к отрицательному решению системы уравнений (4.26).

Если же в качестве  $\Lambda$  взять обратную к матрице Якоби отображения  $F$  в точке  $(2, 2)^\top$ , то организованный так метод простой итерации из начального приближения  $(-1, -1)^\top$  будет сходиться к положительному решению системы. ■

Из разобранного примера хорошо видны особенности метода простой итерации и его недостатки. Схема его применения является очень гибкой, но это приводит к тому, что использование метода простой итерации из технологии может подчас превращаться в искусство. Метод простой итерации носит, вообще говоря, локальный характер, и нередко не обеспечивает глобальную сходимость к решению. Метод простой итерации сам по себе не предназначен для нахождения всех решений системы уравнений. Для этого требуется привлечение дополнительных инструментов, например, тех, что описаны в § 4.8.

## 4.56 Метод Ньютона и его модификации

Рассмотрим систему уравнений

$$\left\{ \begin{array}{l} F_1(x_1, x_2, \dots, x_n) = 0, \\ F_2(x_1, x_2, \dots, x_n) = 0, \\ \vdots \quad \ddots \quad \vdots \\ F_n(x_1, x_2, \dots, x_n) = 0, \end{array} \right. \quad (4.28)$$

или, кратко,

$$F(x) = 0,$$

где  $F(x) = (F_1(x), \dots, F_n(x))^\top$ ,  $x = (x_1, x_2, \dots, x_n)^\top$ . Предположим, что известно некоторое приближение  $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n)$  к решению  $x^* = (x_1^*, x_2^*, \dots, x_n^*)$  этой системы. Если функция  $F$  дифференцируема, то можно воспользоваться формулой Тейлора и приблизить её в окрестности точки  $\tilde{x}$  линейной функцией:

$$F(x) \approx F(\tilde{x}) + F'(\tilde{x})(x - \tilde{x}),$$

где  $F'(\tilde{x})$  — матрица Якоби для  $F$  в точке  $\tilde{x}$ . Далее для вычисления следующего и более точного приближения к решению естественно решить

систему линейных алгебраических уравнений

$$F(\tilde{x}) + F'(\tilde{x})(x - \tilde{x}) = 0,$$

которая близка к исходной системе в окрестности  $\tilde{x}$ . Следовательно, очередным приближением к решению можно взять

$$\tilde{\tilde{x}} = \tilde{x} - (F'(\tilde{x}))^{-1}F(\tilde{x}).$$

Это новое приближение можно снова улучшить по той же формуле и т. д.

Итерационный процесс

$$x^{(k)} \leftarrow x^{(k-1)} - (F'(x^{(k-1)}))^{-1}F(x^{(k-1)}), \quad k = 1, 2, \dots, \quad (4.29)$$

называют *методом Ньютона*. Он является многомерным обобщением метода Ньютона для решения уравнений, который был рассмотрен в § 4.4д. При реализации итераций (4.29) не нужно вычислять на каждом шаге обратную к матрице Якоби  $F'(x^{(k-1)})$ , так как требуется не она сама, а её произведение на вектор  $F(x^{(k-1)})$ . Оно равно решению системы линейных алгебраических уравнений с матрицей  $F'(x^{(k-1)})$  и вектором правой части  $F(x^{(k-1)})$ , и нахождение этого решения является более простым и устойчивым.

Как видим, метод Ньютона требует вычисления на каждом шаге матрицы производных функции  $F$  и решения системы линейных алгебраических уравнений с этой матрицей, которая изменяется от шага к шагу. Нередко подобные трудозатраты могут стать излишне обременительными. Если зафиксировать точку  $\check{x}$ , в которой вычисляется матрица производных  $F'$ , то получим упрощённый стационарный итерационный процесс

$$x^{(k)} \leftarrow x^{(k-1)} - (F'(\check{x}))^{-1}F(x^{(k-1)}), \quad k = 1, 2, \dots,$$

который часто называют *модифицированным методом Ньютона*. В нём решение систем линейных уравнений с одинаковыми матрицами  $F'(\check{x})$  можно проводить по упрощённым алгоритмам, к примеру, найдя один раз LU-разложение матрицы  $F'(\check{x})$  и далее используя его на каждом шаге (см. § 3.6г). Фактически, модифицированный метод Ньютона является специальной формой метода простой итерации из предыдущего раздела.

У метода Ньютона существует много различных вариантов и модификаций, и заинтересованный читатель может найти информацию о них, к примеру, в [13, 27].

Один из наиболее полных и часто используемых результатов о сходимости метода Ньютона — это теорема Л.В. Канторовича о методе Ньютона. Адаптируясь к тематике этой главы, мы сформулируем её для конечномерного случая.

#### Теорема 4.5.1 (теорема Канторовича о методе Ньютона)

Пусть для системы уравнений (4.28) и вектора начального приближения  $x^{(0)}$  к её решению выполнены условия:

отображение  $F : \mathbb{R}^n \supset D \rightarrow \mathbb{R}^n$  определено в открытой области  $D \subset \mathbb{R}^n$  и имеет непрерывную вторую производную  $F''$  в замыкании  $\text{cl } D$ ;

существует непрерывный линейный оператор  $\Gamma = (F'(x^{(0)}))^{-1}$ , такой что  $\|\Gamma(F(x^{(0)}))\| \leq \eta$  и  $\|\Gamma F''(x)\| < K$  для всех  $x \in \text{cl } D$  и некоторых констант  $\eta$  и  $K$ .

Если

$$h = K\eta \leq \frac{1}{2} \quad u \quad r \geq r_0 = \frac{1 - \sqrt{1 - 2h}}{h} \eta,$$

то система уравнений  $F(x) = 0$  имеет решение  $x^*$ , к которому сходится метод Ньютона, как основной, так и модифицированный. При этом

$$\|x^{(0)} - x^*\| \leq r_0.$$

Для основного метода Ньютона сходимость описывается оценкой

$$\|x^{(k)} - x^*\| \leq \frac{\eta}{2^k h} (2h)^{2^k}, \quad k = 0, 1, 2, \dots,$$

а для модифицированного метода верна оценка

$$\|x^{(k)} - x^*\| \leq \frac{\eta}{h} (1 - \sqrt{1 - 2h})^{k+1}, \quad k = 0, 1, 2, \dots,$$

при условии  $h < \frac{1}{2}$ .

Доказательство и дальнейшие результаты на эту тему можно найти в книге [16]. Конечномерный частный случай теоремы Канторовича рассматривается и доказывается также в [12].

Практическая проверка условий теоремы Канторовича непроста, так что сама она имеет, скорее, теоретический характер, обеспечивая уверенность в тех или иных свойства метода. Если условия теоремы Канторовича могут быть проверены, то она обосновывает существование, единственность и локализацию самого решения системы уравнений без его фактического нахождения. Отметим также, что неравенства, которые оценивают быстроту сходимости в теореме Канторовича, означают, что модифицированный метод Ньютона сходится со скоростью геометрической прогрессии, а исходный метод Ньютона — гораздо быстрее. Такая скорость сходимость называется *квадратичной*, и она совершенно такая же, как у сходимости метода Ньютона для уравнений (см. § 4.4д).

**Пример 4.5.2** В применении к системе уравнений (4.26) из предыдущего раздела метод Ньютона позволяет найти оба решения, но для этого нужно подобрать «хорошие» начальные приближения. В целом же сходимость метода Ньютона из начальных приближений, которые не близки к решениям, является нетривиальной.

Например, если взять начальным вектором  $(-3, 3)^T$ , то метод Ньютона сходится к решению  $(-1.2174, -1.5868)^T$ , а если начальным приближением взять  $(-10, -10)^T$ , то к решению  $(1.5620, 1.2491)^T$ . Но из начального приближения  $(-30, -30)^T$  получаем сходимость снова к решению  $(-1.2174, -1.5868)^T$  и т. д.

В общем случае часто бывает неясно, все ли решения системы уравнений были найдены с помощью варьирования начального приближения или какие-то решения остались неохваченными. Это сложный вопрос, который требует привлечения новых идей и методов. В заключительных разделах книги мы рассмотрим интервальные методы для решения этой задачи, наиболее развитые и эффективные в настоящее время. ■

## 4.6 Интервальные системы линейных уравнений

### 4.6а Интервальные уравнения и их решения

Предположим, что в системе линейных алгебраических уравнений рассмотренного нами стандартного вида (3.72)–(3.73) коэффициенты

при неизвестных переменных и правые части известны неточно и могут меняться в пределах некоторых интервалов, т. е.

$$a_{ij} \in \mathbf{a}_{ij}, \quad b_i \in \mathbf{b}_i$$

для всех возможных индексов  $i, j$ . Будем говорить тогда, что задана *интервальная система линейных алгебраических уравнений* вида

$$\left\{ \begin{array}{l} \mathbf{a}_{11}x_1 + \mathbf{a}_{12}x_2 + \dots + \mathbf{a}_{1n}x_n = \mathbf{b}_1, \\ \mathbf{a}_{21}x_1 + \mathbf{a}_{22}x_2 + \dots + \mathbf{a}_{2n}x_n = \mathbf{b}_2, \\ \vdots \qquad \vdots \qquad \ddots \qquad \vdots \qquad \vdots \\ \mathbf{a}_{n1}x_1 + \mathbf{a}_{n2}x_2 + \dots + \mathbf{a}_{nn}x_n = \mathbf{b}_m, \end{array} \right. \quad (4.30)$$

с интервальными коэффициентами  $\mathbf{a}_{ij}$  и свободными членами  $\mathbf{b}_i$ . Её можно записать в краткой матрично-векторной форме

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \quad (4.31)$$

где  $\mathbf{A} = (\mathbf{a}_{ij})$  — это интервальная  $m \times n$ -матрица и  $\mathbf{b} = (\mathbf{b}_i)$  — интервальный  $m$ -вектор, т. е. матрица и вектор составленные из интервалов.

Интервальные уравнения и системы уравнений — это, по существу, параметрические уравнения и системы, но имеющие параметры весьма специального вида. Это обстоятельство позволяет во многих ситуациях выполнять анализ таких уравнений или их решение более полно и глубоко, опираясь на арифметические и прочие операции между интервалами. Интервальные уравнения полезны сами по себе, так как естественно возникают при математическом моделировании многих явлений, в описании которых присутствуют неточности и неопределённости. Но интервальные уравнения и системы уравнений могут также появиться в виде вспомогательного объекта в задачах, где интервальность изначально не присутствует. Например, желая найти доказательные границы решений «обычных точечных» уравнений, мы приходим к необходимости организовывать интервальные уравнения, в которых числовые константы и коэффициенты заменены на гарантированно содержащие их машинно представимые интервалы, как это показывалось в § 1.11 (см. также § 4.7 и 4.8).

Для интервальных уравнений решения и множества решений могут быть определены разнообразными способами [38], но ниже мы ограничимся так называемым *объединённым множеством решений* для

(4.30) и (4.31), которое образовано всевозможными решениями точечных систем  $Ax = b$  того же размера, когда матрица  $A$  и вектор  $b$  независимо пробегают  $\mathbf{A}$  и  $\mathbf{b}$  соответственно. Объединённое множество решений определяется строго как

$$\Xi(\mathbf{A}, \mathbf{b}) := \{ x \in \mathbb{R}^n \mid Ax = b \text{ для некоторых } A \in \mathbf{A} \text{ и } b \in \mathbf{b} \} \quad (4.32)$$

или, совсем формально,

$$\Xi(\mathbf{A}, \mathbf{b}) := \{ x \in \mathbb{R}^n \mid (\exists A \in \mathbf{A})(\exists b \in \mathbf{b})(Ax = b) \}. \quad (4.33)$$

Ниже мы называем его просто *множеством решений* интервальной линейной системы (4.30), так как другие множества решений в нашем учебнике не встречаются.

Совершенно аналогичным образом можно определить интервальные уравнения и системы уравнений общего вида, не обязательно линейные, а также их множества решений.

Предположим, что задана система уравнений

$$\left\{ \begin{array}{l} F_1(a_1, a_2, \dots, a_l, x_1, x_2, \dots, x_n) = 0, \\ F_2(a_1, a_2, \dots, a_l, x_1, x_2, \dots, x_n) = 0, \\ \vdots \quad \ddots \quad \vdots \quad \vdots \\ F_m(a_1, a_2, \dots, a_l, x_1, x_2, \dots, x_n) = 0, \end{array} \right.$$

с неизвестными переменными  $x_1, x_2, \dots, x_n$  и параметрами  $a_1, a_2, \dots, a_l$ , где  $F_i(a_1, a_2, \dots, a_l, x_1, x_2, \dots, x_n)$ ,  $i = 1, 2, \dots, m$ , — некоторые функции. Эту систему уравнений можно также записать кратко в виде

$$F(a, x) = 0,$$

где  $x = (x_1, x_2, \dots, x_n)^\top \in \mathbb{R}^n$  — вектор неизвестных переменных,

$a = (a_1, a_2, \dots, a_l)^\top \in \mathbb{R}^l$  — вектор параметров,

$$F(a, x) = (F_1(a, x), \dots, F_m(a, x))^\top \text{ — вектор-столбец из } F_i.$$

Если параметры  $a_1, a_2, \dots, a_l$  могут изменяться в пределах некоторых интервалов  $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_l$  соответственно, то будем говорить, что задана интервальная система уравнений

$$\left\{ \begin{array}{l} F_1(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_l, x_1, x_2, \dots, x_n) = 0, \\ F_2(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_l, x_1, x_2, \dots, x_n) = 0, \\ \vdots \quad \ddots \quad \vdots \quad \vdots \\ F_m(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_l, x_1, x_2, \dots, x_n) = 0, \end{array} \right. \quad (4.34)$$

с неизвестными переменными  $x_1, x_2, \dots, x_n$  и интервальными параметрами  $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_l$ . Иначе говоря, интервальная система уравнений — это совокупность обычных (точечных) систем уравнений той же структуры и размера, параметры которых могут принимать значения из заданных для них интервалов.

*Объединённым множеством решений* интервальной системы уравнений (4.34) называется множество

$$\Xi(F, \mathbf{a}) := \{x \in \mathbb{R}^n \mid F(a, x) = 0 \text{ для некоторых } a \in \mathbf{a}\},$$

состоящее из всех решений точечных систем  $F(a, x) = 0$ , для которых  $a \in \mathbf{a} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_l)^\top$ . На языке формальной логики это определение равносильно записывается как

$$\Xi(F, \mathbf{a}) := \{x \in \mathbb{R}^n \mid (\exists a \in \mathbf{a})(F(a, x) = 0)\}. \quad (4.35)$$

Ниже мы называем  $\Xi(F, \mathbf{a})$  просто *множеством решений* интервальной системы уравнений (4.34), так как другие множества решений не рассматриваем.

Определения множеств решений (4.32), (4.33) и (4.35) выписаны с помощью формального логического языка, который даёт ясную интерпретацию, но не очень удобен для исследования и для дальнейших приложений. Необходимо иметь представление множеств решений интервальных систем уравнений через стандартные операции алгебры и анализа, точечные или интервальные. Одним из наиболее популярных результатов на эту тему, относящимся к интервальным линейным системам, является

**Теорема 4.6.1** (характеризация Бекка) *Пусть  $\mathbf{A}$  — интервальная  $m \times n$ -матрица,  $\mathbf{b}$  — интервальный  $m$ -вектор. Точка  $\tilde{x}$  из  $\mathbb{R}^n$  принадлежит множеству решений  $\Xi(\mathbf{A}, \mathbf{b})$  интервальной линейной системы уравнений  $A\tilde{x} = \mathbf{b}$  тогда и только тогда, когда  $\mathbf{A}\tilde{x} \cap \mathbf{b} \neq \emptyset$  или, что равносильно,  $0 \in \mathbf{A}\tilde{x} - \mathbf{b}$ .*

**Доказательство.** Если  $\tilde{x} \in \Xi(\mathbf{A}, \mathbf{b})$ , то  $\tilde{A}\tilde{x} = \tilde{b}$  для некоторых  $\tilde{A} \in \mathbf{A}$ ,  $\tilde{b} \in \mathbf{b}$ . Тогда  $\tilde{A}\tilde{x} \in \mathbf{A}\tilde{x}$  и, следовательно, множества  $\mathbf{A}\tilde{x}$  и  $\mathbf{b}$  содержат общий элемент — вектор  $\tilde{b}$ . Поэтому действительно  $\mathbf{A}\tilde{x} \cap \mathbf{b} \neq \emptyset$ .

Наоборот, пусть  $\mathbf{A}\tilde{x} \cap \mathbf{b} \neq \emptyset$ , так что пересечение  $\mathbf{A}\tilde{x} \cap \mathbf{b}$  содержит некоторый вектор  $\tilde{b} \in \mathbb{R}^m$ . Поскольку  $\mathbf{A}\tilde{x} = \{A\tilde{x} \mid A \in \mathbf{A}\}$  в силу свойства (1.20) интервального матрично-векторного умножения,

то найдётся  $\tilde{A} \in \mathbf{A}$ , для которого должно иметь место равенство  $\tilde{b} = \tilde{A}\tilde{x}$  с некоторой  $\tilde{A} \in \mathbf{A}$ . Итак,  $\tilde{x} \in \Xi(\mathbf{A}, \mathbf{b})$ .

Второе равенство следует из того, что  $\mathbf{A}\tilde{x} \cap \mathbf{b} \neq \emptyset$  тогда и только тогда, когда  $0 \in \mathbf{A}\tilde{x} - \mathbf{b}$ . ■

Теоретико-множественные отношения между интервальными векторами (включение и непустота пересечения), которые фигурируют в характеризации Бекка, могут быть выражены через неравенства (1.8) между концами интервальных компонент векторов. Произведение интервальной матрицы  $\mathbf{A}$  на вектор  $\tilde{x}$  — это интервальный вектор, концы компонент которого выражаются произведениями каких-то концевых точечных матриц из  $\mathbf{A}$  на  $\tilde{x}$ . По этой причине характеризация Бекка в каждом отдельном ортанте пространства  $\mathbb{R}^n$  эквивалентна системе линейных алгебраических неравенств, т. е. задаёт выпуклое многогранное множество. В целом множество решений (4.32)–(4.33) является объединением этих выпуклых многогранников (подробности можно увидеть в книгах [38, 51]).

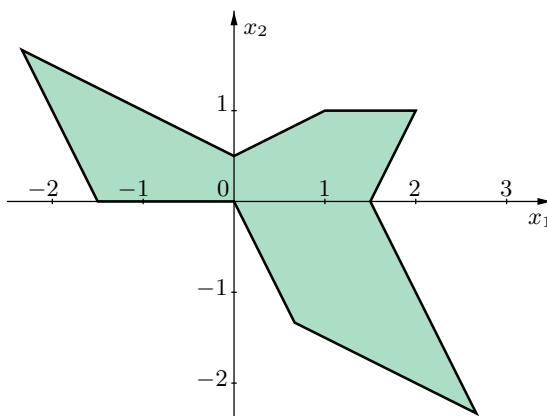


Рис. 4.20. Множество решений интервальной линейной системы (4.36)

**Пример 4.6.1** На рис. 4.20 изображено множество решений интер-

вальной системы линейных уравнений<sup>4</sup>

$$\begin{pmatrix} [2, 4] & [-1, 1] \\ [-1, 1] & [2, 4] \\ [0, 2] & [1, 2] \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} [-3, 3] \\ [-2, 1] \\ [0, 1] \end{pmatrix}. \quad (4.36)$$

Количество уравнений в ней больше числа неизвестных, но множество решений всё таки непусто, причём оно остаётся непустым при небольшом изменении элементов матрицы и правой части. В этом — принципиальное отличие интервальных систем уравнений от обычных точечных. Неинтервальные переопределённые системы уравнений не имеют решений в случае общего положения, а если решение такой системы всё-таки существует, то оно обычно неустойчиво к возмущениям уравнений. ■

Точное описание множества решений интервальной линейной системы уравнений, при котором скрупулёзно выписываются все его грани (ограничивающие его гиперплоскости), может расти экспоненциально с числом неизвестных  $n$  (как общее число ортантов  $2^n$ ). Поэтому оно является практически невозможным уже при  $n$ , достигающем нескольких десятков. С другой стороны, в большинстве реальных постановок задач такое точное описание в действительности и не нужно. На практике бывает вполне достаточно нахождения некоторой *оценки* для множества решений, т. е. приближённого описания, удовлетворяющего содержательному смыслу рассматриваемой задачи. Эти приближённые оценки множеств решений могут иметь разный смысл (рис. 4.21), в частности, быть внутренними оценками (в виде подмножеств) или внешними оценками (в виде объемлющих множеств) или какими-то ещё.

Существование различных множеств решений и большое разнообразие способов их оценивания, которое в самом деле требуется в задачах теории принятия решений, интервального анализа данных и т. п., имеют следствием большое количество постановок задач для интервальных уравнений и систем уравнений. По этой причине не вполне корректно говорить о «решении интервальных систем уравнений» вообще, без привязки к какой-то конкретной постановке задачи.

---

<sup>4</sup>Технология отрисовки таких множеств решений описана, к примеру, в книге [51], § 4.14.2. Для практической работы можно использовать свободные пакеты *IntLinIncR2* и *IntLinIncR3* [61] или процедуры библиотеки *IntvalPy*.

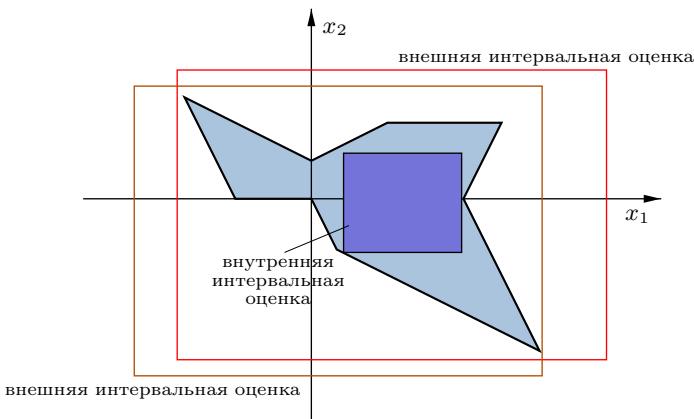


Рис. 4.21. Различные способы оценивания множества решений

Для полного описания постановки задачи необходимо указать тип множества решений, способ его оценивания (чем и как) и другие подробности, которые в совокупности определяют «решение» интервальной системы уравнений и процесс его получения.

Трудоёмкость этих задач тоже весьма различна. Для объединённого множества решений задача распознавания пустоты или непустоты и задача внешнего интервального оценивания с заданной абсолютной или относительной погрешностью являются труднорешаемыми (NP-трудными); см. подробности в [42]. Иными словами, решение этих задач может потребовать трудозатрат, которые растут экспоненциально с размером задачи, а сами алгоритмы для их решения в полной постановке неизбежно должны носить переборный характер [38]. Для других множеств решений аналогичные задачи распознавания и оценивания могут быть полиномиально разрешимыми (например, это верно для так называемого допускового множества решений).

#### 4.66 Численные методы для интервальных линейных систем уравнений

К настоящему времени разработано большое количество численных методов для решения различных постановок задач для интервальных систем линейных алгебраических уравнений вида (4.30) или (4.31). Они

включают в себя разнообразные методы для внутреннего и внешнего оценивания тех или иных множеств решений, методы для распознавания этих множеств решений, исчерпывающего оценивания и т. п. Наиболее развитыми из них являются численные методы для внешнего интервального оценивания множества решений интервальных линейных систем, так как они находят наибольшее применение в традиционных задачах вычислительной математики и смежных дисциплинах.

Из труднорешаемости задач распознавания и оценивания множества решений следует, что имеет смысл классифицировать методы решения этих задач по трудоёмкости — полиномиальной либо экспоненциальной. Полиномиальные алгоритмы («быстрые») дают ответ к задаче за приемлемое время, но без каких-либо гарантий точности и т. п. Те алгоритмы, которые нацелены на получение гарантированной точности ответов («точные»), неизбежно должны иметь экспоненциальную трудоёмкость. Традиционное разделение методов на прямые и итерационные, которое характерно для классической вычислительной математики, в этой ситуации делается не слишком актуальным.

В этом разделе рассмотрим бегло некоторые «быстрые» алгоритмы для внешнего оценивания множеств решений интервальных систем линейных алгебраических уравнений. «Точные» алгоритмы устроены существенно по-другому и читатель может увидеть их, например, в книге [38].

### Интервальный метод Гаусса

Интервальный метод Гаусса является естественным интервальным расширением обычного метода Гаусса для систем линейных алгебраических уравнений. Он предназначен для нахождения внешней интервальной оценки объединённого множества решений, и его детальное описание вместе с исследованием свойств и особенностей можно найти в [38, 45, 67]. Использованию интервального метода Гаусса часто предшествует предварительное предобуславливание интервальной линейной системы (см. подраздел далее).

Традиционный неинтервальный метод Гаусса обычно применяют для квадратных систем линейных уравнений, в которых число уравнений равно числу неизвестных. Но в интервальном случае, как впервые заметили Э. Хансен и У. Уолстер в [69], метод хорошо применим и к переопределённому случаю, когда число уравнений больше числа неизвестных и матрица системы стоячая.

**Пример 4.6.2** Рассмотрим решение с помощью интервального метода Гаусса системы (4.36). Ответ

$$\begin{pmatrix} [-2.6667, 2.6667] \\ [-2.3333, 1.6667] \end{pmatrix}$$

является очень хорошей внешней оценкой множества решений, представленного на рис. 4.20. Оптимальность оценки по последней компоненте является следствием того, что система имеет всего два неизвестных, и в общем случае это не имеет места. ■

Так как интервальный метод Гаусса является естественным интервальным расширением обычного метода Гаусса, то погрешность получаемых им результатов может быть оценена с помощью неравенства (1.23). Она имеет, таким образом, первый порядок по ширине интервалов в матрице и правой части системы уравнений. В целом результаты интервального метода Гаусса не отличаются высоким качеством оценивания множеств решений, но зато сфера применимости этого метода заметно шире, чем у других более тонких методов. Интервальный метод Гаусса нередко позволяет получать хоть какие-то оценки в ситуациях, когда другие подходы неприменимы.

### Интервальный метод Гаусса–Зейделя

Интервальный метод Гаусса–Зейделя является итерационной процедурой для уточнения внешней оценки множества решений интервальной линейной системы вида (4.30) либо части множества решений, ограниченной некоторым бруском. Он был предложен в 70-е годы XX века как обобщение обычного неинтервального метода Гаусса–Зейделя (см. § 3.10e). Как правило, применяют интервальный метод Гаусса–Зейделя после предварительного предобуславливания системы уравнений.

Предположим, что дана интервальная система линейных алгебраических уравнений  $\mathbf{A}\mathbf{x} = \mathbf{b}$  и известен некоторый брус  $\mathbf{X}$ ,

$$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n)^\top,$$

содержащий множество решений  $\Xi(\mathbf{A}, \mathbf{b})$  или же некоторую его часть, которая интересует нас по условиям задачи. Пусть также в интервальной матрице  $\mathbf{A} = (\mathbf{a}_{ij})$  элементы главной диагонали не содержат нуля,

т. е.  $0 \notin a_{ii}$  для  $i = 1, 2, \dots, n$ . Для неособенной интервальной матрицы этому условию всегда можно удовлетворить перестановкой строк (уравнений системы).

Если  $\tilde{x} \in \Xi(\mathbf{A}, \mathbf{b}) \cap \mathbf{X}$ , то

$$\tilde{A}\tilde{x} = \tilde{b}$$

для некоторых  $\tilde{A} = (\tilde{a}_{ij}) \in \mathbf{A}$  и  $\tilde{b} = (\tilde{b}_i) \in \mathbf{b}$ , или, в развёрнутом виде,

$$\sum_{j=1}^n \tilde{a}_{ij} \tilde{x}_j = \tilde{b}_i, \quad i = 1, 2, \dots, n.$$

Оставим в левых частях этих равенств слагаемые, соответствующие диагональным элементам матрицы, а остальные перенесём в правые части и затем поделим обе части равенств на  $\tilde{a}_{ii}$ . Получим

$$\tilde{x}_i = \left( \tilde{b}_i - \sum_{j \neq i} \tilde{a}_{ij} \tilde{x}_j \right) / \tilde{a}_{ii}, \quad i = 1, 2, \dots, n. \quad (4.37)$$

Полагая

$$\tilde{\mathbf{X}}_i := \mathbf{X}_i \cap \left( \mathbf{b}_i - \sum_{j \neq i} \mathbf{a}_{ij} \mathbf{X}_j \right) / \mathbf{a}_{ii}, \quad i = 1, 2, \dots, n, \quad (4.38)$$

мы должны признать, что

$$\tilde{x}_i \in \tilde{\mathbf{X}}_i, \quad i = 1, 2, \dots, n,$$

так как выражения для  $\tilde{\mathbf{X}}_i$  являются пересечениями  $\mathbf{X}_i$  с естественными интервальными расширениями выражений (4.37) по  $\tilde{a}_{ij} \in \mathbf{a}_{ij}$ ,  $\tilde{b}_i \in \mathbf{b}_i$  и  $\tilde{x}_j \in \mathbf{X}_j$ . Итак,  $\tilde{x} \in \tilde{\mathbf{X}}$ , и это верно для любой точки  $\tilde{x} \in \Xi(\mathbf{A}, \mathbf{b}) \cap \mathbf{X}$ , так что в целом

$$\Xi(\mathbf{A}, \mathbf{b}) \cap \mathbf{X} \subseteq \tilde{\mathbf{X}},$$

т. е. брус  $\tilde{\mathbf{X}}$  является новой внешней оценкой интересующей нас части множества решений рассматриваемой интервальной системы.

Чтобы взять лучшее от обеих внешних оценок — старой  $\mathbf{X}$  и новой  $\tilde{\mathbf{X}}$ , естественно выполнить их пересечение. Это в действительности уже делается при вычислении по формуле (4.38). Далее процесс построения новой внешней оценки для множества решений можно повторить, отправляясь от  $\tilde{\mathbf{X}}$  и потом снова взяв пересечение полученной внешней

Таблица 4.2. Интервальный метод Гаусса–Зейделя  
для внешнего оценивания множеств решений  
интервальных линейных систем уравнений

**Вход**

Интервальная линейная система уравнений  $\mathbf{A}x = \mathbf{b}$ .

Брус  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^\top \in \mathbb{IR}^n$ , ограничивающий желаемую часть объединённого множества решений  $\Xi(\mathbf{A}, \mathbf{b})$ .

Некоторая константа  $\epsilon > 0$ .

**Выход**

Уточнённая внешняя оценка  $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}_1, \dots, \tilde{\mathbf{X}}_n)^\top \supseteq \Xi(\mathbf{A}, \mathbf{b}) \cap \mathbf{X}$  для части множества решений, содержащейся в  $\mathbf{X}$ , либо информация «множество  $\Xi(\mathbf{A}, \mathbf{b})$  не пересекает брус  $\mathbf{X}$ ».

**Алгоритм**

$q \leftarrow +\infty;$

DO WHILE (  $q \geq \epsilon$  )

DO FOR  $i = 1$  TO  $n$

$$\tilde{\mathbf{X}}_i \leftarrow \mathbf{X}_i \cap \left( \mathbf{b}_i - \sum_{j=1}^{i-1} \mathbf{a}_{ij} \tilde{\mathbf{X}}_j - \sum_{j=i+1}^n \mathbf{a}_{ij} \mathbf{X}_j \right) / \mathbf{a}_{ii};$$

IF (  $\tilde{\mathbf{X}}_i = \emptyset$  ) THEN

STOP, сигнализируя «множество решений  $\Xi(\mathbf{A}, \mathbf{b})$  не пересекает брус  $\mathbf{X}$ »

END IF

END DO

$q \leftarrow$  расстояние между векторами  $\mathbf{X}$  и  $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}_1, \dots, \tilde{\mathbf{X}}_n)^\top$ ;

$\mathbf{X} \leftarrow \tilde{\mathbf{X}}$ ;

END DO

оценки с предшествующей и т. д. Наконец, поскольку в этом итерационном процессе компоненты нового внешнего приближения множества решений насчитываются последовательно друг за другом, начиная с самой первой, то мы можем организовывать пересечение старой и новой оценок тоже по мере вычисления компонент и сразу же привлекать уточнённые новые компоненты для расчёта оставшихся компонент следующего приближения.

Может ли в процессе описанного уточнения-пересечения встретиться ситуация  $\mathbf{X}_i \cap \tilde{\mathbf{X}}_i = \emptyset$  для некоторого  $i$ ? Из наших рассуждений следует, что это возможно лишь при нарушении исходного допущения о том, что начальный брус содержит точки множества решений, т. е. когда  $\Xi(\mathbf{A}, \mathbf{b}) \cap \mathbf{X} = \emptyset$ . Таким образом, развитую выше методику можно применять для произвольного начального бруса  $\mathbf{X}$ , но получение в качестве промежуточного результата пустого множества будет свидетельствовать о том, что  $\mathbf{X}$  вообще не пересекает оцениваемого множества решений.

Псевдокод итоговой вычислительной схемы интервального метода Гаусса–Зейделя приведён в табл. 4.2.

По самому построению интервального метода Гаусса–Зейделя результатом его работы является брус, не более широкий, чем начальное приближение  $\mathbf{X}$ . Когда он действительно уже, чем  $\mathbf{X}$ ? Одно из достаточных условий того, что брус начального приближения будет улучшен интервальным методом Гаусса–Зейделя, — диагональное преобладание в матрице интервальной линейной системы. Оно означает, что все точечные матрицы из данной интервальной матрицы имеют диагональное преобладание [38, 45].

**Пример 4.6.3** Рассмотрим внешнее оценивание множества решений интервальной системы линейных уравнений (3.68), которая встречалась в § 3.4г:

$$\begin{pmatrix} [2, 4] & [-2, 0] \\ [-1, 1] & [2, 4] \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} [-1, 1] \\ [0, 2] \end{pmatrix}.$$

Напомним, что брус оптимальной внешней оценки её множества решений равен  $([-1, 3], [-0.5, 2.5])^\top$  (рис. 3.14).

Интервальный метод Гаусса–Зейделя для этой системы из начального приближения  $([-20, 20], [-20, 20])^\top$  довольно быстро сходится к брусу  $([-2, 3], [-1.5, 2.5])^\top$  (за 11 итераций получаются уже по 2 вер-

ных знака после десятичной точки). Этот результат гораздо точнее ответа, полученного в примере 3.4.6 с помощью неравенства с числом обусловленности.

Усложним задачу, заменив интервал на месте  $(2,1)$  в матрице на более широкий  $[-1, 2]$ . Множество решений новой интервальной линейной системы

$$\begin{pmatrix} [2, 4] & [-2, 0] \\ [-1, 2] & [2, 4] \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} [-1, 1] \\ [0, 2] \end{pmatrix} \quad (4.39)$$

также расширится, хотя брус оптимальной внешней оценки останется прежним. Но интервальный метод Гаусса-Зейделя станет сходиться существенно хуже и будет выдавать менее точные оценки. Например, нетрудно проверить (и даже показать строго с помощью выкладок), что любой интервальный брус вида  $([-C, C], [-C, C])^\top$ ,  $C > 0$ , никак не улучшается интервальным методом Гаусса-Зейделя. Некоторые другие начальные брусы могут улучшаться, но в целом качество оценок становится неудовлетворительным. ■

## Предобуславливание

Предобуславливанием интервальной системы линейных алгебраических уравнений называют умножение слева матрицы системы и её правой части на специально подобранный точечной матрицы. Таким образом, вместо системы  $\mathbf{A}\mathbf{x} = \mathbf{b}$  получаем систему  $(\Lambda\mathbf{A})\mathbf{x} = \Lambda\mathbf{b}$ , где  $\Lambda$  — предобуславливающая матрица.

Цель предобуславливания — улучшение свойств матрицы системы (в частности, получение диагонального преобразования и т. п.), в результате которого к этой системе могут быть применены те или иные интервальные численные методы. Поскольку

$$\Lambda\mathbf{A} \supseteq \{\Lambda A \mid A \in \mathbf{A}\} \quad \text{и} \quad \Lambda\mathbf{b} \supseteq \{\Lambda b \mid b \in \mathbf{b}\},$$

то

$$\Xi(\mathbf{A}, \mathbf{b}) \subseteq \Xi(\Lambda\mathbf{A}, \Lambda\mathbf{b}),$$

т. е. множество решений интервальной линейной системы в результате предобуславливания расширяется. По этой причине внешняя оценка для множества решений предобуславленной системы может служить также внешней оценкой для множества решений исходной.

Пусть  $\Lambda = (\text{mid } \mathbf{A})^{-1}$ , тогда

$$\begin{aligned}\Lambda \mathbf{A} &= (\text{mid } \mathbf{A})^{-1} (\text{mid } \mathbf{A} + [-1, 1] \cdot \text{rad } \mathbf{A}) = \\ &= I + (\text{mid } \mathbf{A})^{-1} \cdot [-1, 1] \cdot \text{rad } \mathbf{A}.\end{aligned}$$

Если матрица  $\text{rad } \mathbf{A}$  является «малой» в том смысле, что мала  $\|\text{rad } \mathbf{A}\|$ , то мало и произведение, в котором она встречается (второе слагаемое полученной суммы), так что матрица  $\Lambda \mathbf{A}$  не сильно отличается от единичной. Тогда она обладает диагональным преобладанием и прочими достоинствами, а интервальный метод Гаусса-Зейделя и другие интервальные методы становятся применимыми к предобусловленной системе уравнений.

Рассмотренное домножение обеих частей системы на  $\Lambda = (\text{mid } \mathbf{A})^{-1}$  называют *предобуславливанием «обратной средней»*, и оно является наиболее популярным для интервальных линейных систем с «неширокими» матрицами коэффициентов.

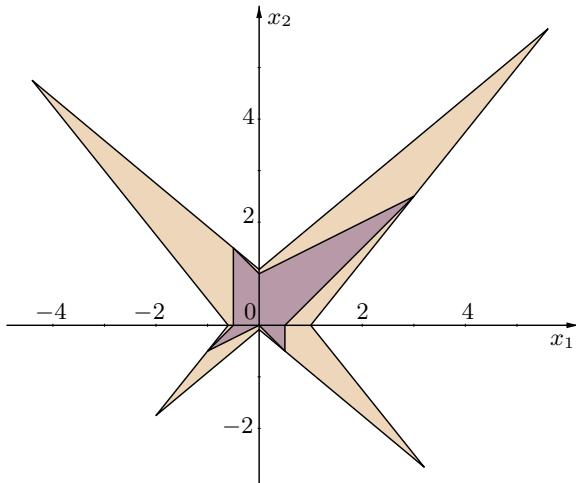


Рис. 4.22. Множества решений интервальных линейных систем (4.39) и (4.40)

**Пример 4.6.4** Снова рассмотрим интервальную линейную систему (4.39), с широкими интервалами в матрице, на которой интервальный

метод Гаусса-Зейделя работает плохо. Выполним её предобусловливание с помощью «обратной средней» матрицы, которая равна

$$\begin{pmatrix} 0.3158 & 0.1053 \\ -0.0526 & 0.3158 \end{pmatrix}$$

(здесь и ниже для числовых результатов оставляем по четыре знака после десятичной точки). Получаем интервальную линейную систему

$$\begin{pmatrix} [0.5263, 1.47368] & [-0.4211, 0.4211] \\ [-0.526, 0.5263] & [0.6316, 1.3684] \end{pmatrix} x = \begin{pmatrix} [-0.3158, 0.5263] \\ [-0.0526, 0.6842] \end{pmatrix}, \quad (4.40)$$

множество решений которой существенно расширяется. На рис. 4.22 оно изображено светлым тоном, тогда как более тёмное и меньшее множество внутри него — это множество решений исходной системы (4.39). Но для системы (4.40) интервальный метод Гаусса-Зейделя работает лучше. Из бруса начального приближения  $([-20, 20], [-20, 20])^\top$ , в частности, он сходится к интервальному вектору

$$\begin{pmatrix} [-5.2, 5.6] \\ [-4.75, 5.75] \end{pmatrix},$$

который ненамного шире оптимальной внешней оценки множества решений системы (4.40) — бруса  $([-4.4, 5.6], [-2.75, 5.75])^\top$ .

Полученный результат не является слишком уж качественной оценкой множества решений системы (4.39), но всё-таки заметно лучше того, что найден в примере 3.4.6 для более узкой интервальной системы с помощью традиционной техники, основанной на числе обусловленности.

## 4.7 Интервальные методы решения уравнений и систем уравнений

### 4.7а Основы интервальной техники

Задача решения уравнений и систем уравнений является одной из классических задач вычислительной математики, для решения которой развито немало эффективных подходов — метод простой итерации, метод Ньютона, их модификации и т.п. Преимущества и недостатки этих классических методов мы обсудили выше в § 4.4–4.5 (см. также [5, 27, 32, 41]). Для дальнейшего нам важны два факта:

- Для уравнений, в которых фигурируют функции, не обладающие «хорошими» глобальными свойствами, все традиционные методы имеют *локальный характер*, т. е. обеспечивают отыскание решения, находящегося в некоторой (иногда достаточно малой) окрестности начального приближения. Задача нахождения *всех* решений уравнения или системы уравнений, как правило, рассматривается лишь в специальных руководствах и методы её решения оказываются очень сложными.
- Гарантированные оценки погрешности найденного приближения к решению в традиционных методах дать весьма непросто.

Указание приближённого значения величины и его максимальной погрешности равносильно тому, что мы знаем левую и правую границы возможных значений этой величины, и поэтому можно переформулировать нашу задачу в следующем усиленном виде —

Найти все решения системы уравнений

$$F(x) = 0$$

на данном множестве  $D \subseteq \mathbb{R}^n$  и указать для каждого гарантированные двусторонние границы (по-возможности, наиболее точные)

. (4.41)

Эту постановку будем называть *задачей доказательного глобального решения* системы уравнений. Эпитет «доказательный» означает здесь, что получаемый нами ответ к задаче — границы решений и т. п. — имеет статус математически строго доказанного утверждения о расположении решений при условии, что ЭВМ работает корректно (см. § 1.11).

Задача (4.41) оказывается трудной (см. обзор в конце § 4.1), и в классическом численном анализе существует очень немного развитых методов для её решения. Из часто используемых подходов, имеющих ограниченный успех, следует упомянуть *аналитическое исследование, мультистарт, методы продолжения* [27].

Итак, пусть необходимо найти решения системы уравнений (4.3),

$$F(x) = 0$$

на брусе  $\mathbf{X} \subset \mathbb{R}^n$ . Обозначим посредством  $\text{ran}(F, \mathbf{X})$  область значений отображения  $F$  на  $\mathbf{X}$ , т. е.

$$\text{ran}(F, \mathbf{X}) := \{F(x) \mid x \in \mathbf{X}\}.$$

Существование решения интересующей нас системы уравнений на  $\mathbf{X}$  можно переписать в виде равносильного условия

$$\text{ran}(F, \mathbf{X}) \ni 0,$$

и потому техника интервального оценивания множеств значений функций оказывается чрезвычайно полезной при решении рассматриваемой задачи.

Предположим, что найдена внутренняя (в виде подмножества) интервальная оценка множества значений  $\text{ran}(F, \mathbf{X})$  отображения  $F$ , и она содержит нуль. Тогда тем более  $0 \in \text{ran}(F, \mathbf{X})$ , и, следовательно, на брусе  $\mathbf{X}$  гарантированно находится решение системы (4.3).

С другой стороны, если в нашем распоряжении имеется интервальное расширение  $\mathbf{F}$  функции  $F$  на  $\mathbf{X}$ , то оно даёт внешнюю (в виде объемлющего множества) оценку области значений, т. е.  $\mathbf{F}(\mathbf{X}) \supseteq \text{ran}(F, \mathbf{X})$ . Поэтому если  $0 \notin \mathbf{F}(\mathbf{X})$ , то тогда тем более  $0 \notin \text{ran}(F, \mathbf{X})$  и потому на  $\mathbf{X}$  нет решений рассматриваемой системы уравнений. Отметим, что эти соображения справедливы вообще для любых уравнений и систем уравнений, недоопределённых, переопределённых и пр., т. е. у которых количество неизвестных не обязательно совпадает с числом уравнений.

Далее, если которых количество неизвестных равно числу уравнений, то исходную систему (4.3) всегда можно переписать в равносильной рекуррентной форме

$$x = T(x) \tag{4.42}$$

с некоторым отображением  $T : \mathbb{R}^n \rightarrow \mathbb{R}^n$ . Оно может быть взято, к примеру, в виде

$$T(x) = x - F(x)$$

либо

$$T(x) = x - \Lambda F(x),$$

с неособенной  $n \times n$ -матрицей  $\Lambda$ , либо каким-нибудь другим способом. Переход к рекуррентной форме даёт некоторые дополнительные возможности. Пусть  $\mathbf{T} : \mathbb{IR}^n \rightarrow \mathbb{IR}^n$  — интервальное расширение отображения  $T$ . Ясно, что решения системы (4.42) могут лежать лишь в

пересечении  $\mathbf{X} \cap \mathbf{T}(\mathbf{X})$ . Поэтому если

$$\mathbf{X} \cap \mathbf{T}(\mathbf{X}) = \emptyset,$$

то в  $\mathbf{X}$  нет решений системы уравнений (4.42). Коль скоро искомое решение содержится и в  $\mathbf{T}(\mathbf{X})$ , то для дальнейшего уточнения бруса, в котором может присутствовать решение, мы можем организовать итерации с пересечением

$$\mathbf{X}^{(0)} \leftarrow \mathbf{X}, \quad (4.43)$$

$$\mathbf{X}^{(k)} \leftarrow \mathbf{T}(\mathbf{X}^{(k-1)}) \cap \mathbf{X}^{(k-1)}, \quad k = 1, 2, \dots \quad (4.44)$$

Следует особо отметить, что в получающихся при этом брусах наличие решения, вообще говоря, не гарантируется. Они являются лишь «подозрительными» на существование решения.

Но вот если для бруса  $\mathbf{X}$  выполнено

$$\mathbf{T}(\mathbf{X}) \subseteq \mathbf{X},$$

то по теореме Брауэра о неподвижной точке (стр. 715) в  $\mathbf{X}$  гарантированно находится решение системы (4.42). Для уточнения этого бруса мы снова можем воспользоваться итерациями (4.43)–(4.44). Таким образом, наихудшим, с точки зрения уточнения информации о решении системы, является случай

$$\mathbf{T}(\mathbf{X}) \supsetneq \mathbf{X}. \quad (4.45)$$

Приведённую выше последовательность действий по обнаружению решения системы уравнений и уточнению его границ мы будем называть далее кратко *тестом существования* (решения). Условимся считать, что его результатом является брус пересечения ( $\mathbf{X} \cap \mathbf{T}(\mathbf{X})$ ) либо предел последовательности (4.43)–(4.44). Если этот брус непуст, то он либо наверняка содержит решение системы уравнений, либо является подозрительным на наличие в нём решения. Если же результат теста существования пуст, то в исходном брусе решений системы уравнений нет.

В действительности, каждый из описанных приёмов уточнения решения допускает далеко идущие модификации и улучшения. Например, это относится к итерациям вида (4.43)–(4.44), которые могут быть последовательно применены не к целым брусам  $\mathbf{X}^{(k)}$ , а к отдельным

их компонентам в комбинации с различными способами приведения исходной системы к рекуррентному виду (4.42). На этом пути мы приходим к чрезвычайно эффективным алгоритмам, которые получили наименование *методов распространения ограничений* [30].

Как простейший тест существования, так и его более продвинутые варианты без особых проблем реализуются на ЭВМ и работают тем лучше, чем более качественно вычисляются интервальные расширения функций  $F$  в (4.3) и  $T$  в (4.42) и чем меньше ширина бруса  $\mathbf{X}$ . Последнее связано с тем, что погрешность оценивания области значений функции посредством любого интервального расширения убывает с уменьшением размеров бруса, на котором производится это оценивание. (см. § 1.6).

## 4.76 Одномерный интервальный метод Ньютона

В этом параграфе мы рассмотрим простейший случай одного уравнения с одним неизвестным и интервальную версию метода Ньютона для его решения. Впервые она была предложена Т. Сунагой в работе [76], а затем получила популярность и большое развитие во многих направлениях.

Предположим, что  $f : \mathbb{R} \supseteq \mathbf{X} \rightarrow \mathbb{R}$  — функция, имеющая нуль  $x^*$  на рассматриваемом интервале  $\mathbf{X}$  и дифференцируемая на нём. Тогда для любой точки  $\tilde{x} \in \mathbf{X}$  из этого же интервала в силу теоремы Лагранжа о среднем значении (о конечном приращении)

$$f(\tilde{x}) - f(x^*) = (\tilde{x} - x^*) \cdot f'(\xi), \quad (4.46)$$

где  $\xi$  — некоторая точка между  $\tilde{x}$  и  $x^*$ . Но так как  $f(x^*) = 0$ , то при  $f'(\xi) \neq 0$  отсюда следует

$$x^* = \tilde{x} - \frac{f(\tilde{x})}{f'(\xi)}.$$

Если  $f'(\mathbf{X})$  является какой-либо интервальной оценкой производной от функции  $f(x)$  на  $\mathbf{X}$ , то  $f'(\xi) \in f'(\mathbf{X})$  и, интервализуя выписанное равенство, получим включение

$$x^* \in \tilde{x} - \frac{f(\tilde{x})}{f'(\mathbf{X})} \quad (4.47)$$

при условии  $0 \notin f'(\mathbf{X})$ . Иными словами, для решения  $x^*$  уравнения  $f(x) = 0$  получается новый интервал локализации в виде правой части

включения (4.47). Интервальное выражение, фигурирующее в правой части (4.47), играет важную роль в интервальном анализе и потому выделяется в самостоятельное понятие.

**Определение 4.7.1** Пусть заданы функция  $f : \mathbb{R} \rightarrow \mathbb{R}$  и интервальная оценивающая функция  $\mathbf{f}'$  для её производной. Отображение

$$\mathcal{N} : \mathbb{IR} \times \mathbb{R} \rightarrow \mathbb{IR},$$

действующее по правилу

$$\mathcal{N}(\mathbf{X}, \tilde{x}) := \tilde{x} - \frac{f(\tilde{x})}{\mathbf{f}'(\mathbf{X})},$$

называется (одномерным) интервальным оператором Ньютона для  $f$ .

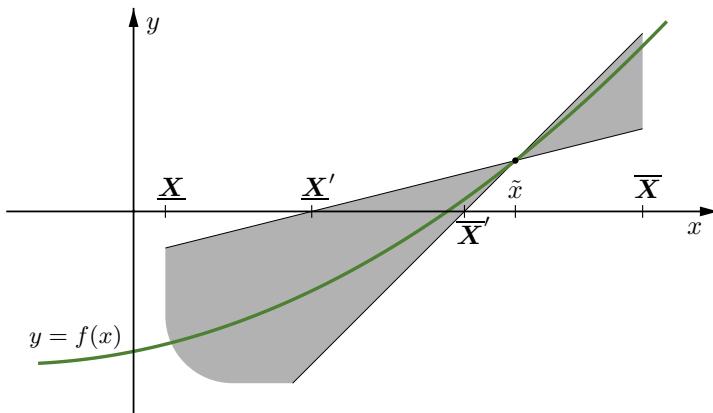


Рис. 4.23. Иллюстрация работы интервального метода Ньютона

Итак, пусть  $0 \notin \mathbf{f}'(\mathbf{X})$ , так что  $\mathcal{N}(\mathbf{X}, \tilde{x})$  является вполне определённым конечным интервалом. Поскольку любой нуль функции  $f(x)$  на  $\mathbf{X}$  лежит также в  $\mathcal{N}(\mathbf{X}, \tilde{x})$ , то разумно взять в качестве следующего более точного интервала локализации решения пересечение

$$\mathbf{X} \cap \mathcal{N}(\mathbf{X}, \tilde{x}),$$

которое окажется, по крайней мере, не хуже  $\mathbf{X}$ . Эта ситуация иллюстрируется на рис. 4.23, где обозначено  $\mathbf{X}' = \mathcal{N}(\mathbf{X}, \tilde{x})$ . Далее, положив  $\mathbf{X}^{(0)} = \mathbf{X}$ , естественно организовать итерационное уточнение

$$\mathbf{X}^{(k)} \leftarrow \mathbf{X}^{(k-1)} \cap \mathcal{N}(\mathbf{X}^{(k-1)}, \tilde{x}^{(k-1)}), \quad k = 1, 2, \dots, \quad (4.48)$$

задавшись каким-то правилом выбора точек  $\tilde{x}^{(k-1)} \in \mathbf{X}^{(k-1)}$ . Итерации (4.48) называются *интервальным методом Ньютона*. В благоприятном случае он порождает последовательность интервалов  $\mathbf{X}^{(k)}$  уменьшающейся ширины, которые содержат искомое решение уравнения. Критерием остановки итераций при этом может быть достижение требуемой точности локализации решения, т. е. ширины  $\mathbf{X}^{(k)}$ .

Ещё одним вариантом развития итераций (4.48) является возникновение на каком-то шаге пустого пересечения  $\mathbf{X}^{(k)} \cap \mathcal{N}(\mathbf{X}^{(k)}, \tilde{x})$ . При этом необходимо прекращать выполнение алгоритма, коль скоро арифметические операции с пустым множеством не определены.<sup>5</sup> С другой стороны, тогда по построению интервального оператора Ньютона мы должны заключить, что на  $\mathbf{X}^{(k)}$ , а значит и на исходном интервале  $\mathbf{X}$ , решений уравнения  $f(x)$  нет.

Наконец, наименее благоприятным с точки зрения уточнения информации о решении является «застаивание» итераций интервального метода Ньютона, когда на каком-то шаге получаем  $\mathbf{X}^{(k)} \subseteq \mathcal{N}(\mathbf{X}^{(k)}, \tilde{x})$ , так что

$$\mathbf{X}^{(k)} = \mathbf{X}^{(k-1)} \cap \mathcal{N}(\mathbf{X}^{(k-1)}, \tilde{x}) = \mathbf{X}^{(k-1)}.$$

Ясно, что тогда все последующие итерации метода будут равны  $\mathbf{X}^{(k-1)}$ , и решение никак не уточнится. Ниже в § 4.8 мы обсудим, как преодолевать это затруднение.

В случае, когда производная функции  $f$ , фигурирующей в левой части уравнения, на интервале  $\mathbf{X}$  не зануляется и её оценка  $f'(\mathbf{X})$  не содержит нуля, интервальный метод Ньютона обладает рядом замечательных качеств. Если  $0 \notin f'(\mathbf{X})$  для некоторого  $\mathbf{X}$ , то на следующем шаге метода будет исключена по крайней мере половина  $\mathbf{X}$ . При этом асимптотический порядок сходимости метода к нулю функции  $f$  на интервале  $\mathbf{X}$  является квадратичным, т. е. таким же, как у обычного неинтервального метода Ньютона.

---

<sup>5</sup> В некоторых компьютерных реализациях результат любой операции с пустым множеством полагается равным также пустому множеству, что почти равносильно.

**Предложение 4.7.1** Пусть функция  $f$  непрерывно дифференцируема и на интервале  $\mathbf{X}$  имеет место  $f'(\mathbf{X}) \not\equiv 0$ . Если для некоторой точки  $\tilde{x}$  справедливо включение  $\mathcal{N}(\mathbf{X}, \tilde{x}) \subseteq \mathbf{X}$ , то интервал  $\mathbf{X}$  содержит решение уравнения  $f(x) = 0$ .

**Доказательство.** Помимо  $\tilde{x}$  рассмотрим ещё точку  $y \in \mathbf{X}$ . Согласно теореме Лагранжа о среднем значении найдётся такая точка  $\xi \in \square\{\tilde{x}, y\} \subset \mathbf{X}$ , что

$$f(\tilde{x}) - f(y) = f'(\xi)(\tilde{x} - y). \quad (4.49)$$

Чтобы подчеркнуть зависимость этой точки от  $\tilde{x}$  и  $y$ , мы обозначим её как  $\xi(\tilde{x}, y)$ . Коль скоро  $f'(\xi) \in f'(\mathbf{X})$ , то ясно, что  $f'(\xi(\tilde{x}, y)) \neq 0$  при любых  $\tilde{x}$  и  $y$ . По этой причине мы можем определить функцию

$$g(y) = y - \frac{f(y)}{f'(\xi(\tilde{x}, y))}. \quad (4.50)$$

По условиям предложения функция  $g(y)$  непрерывна. Кроме того, из равенства (4.49) следует

$$\tilde{x} - \frac{f(\tilde{x})}{f'(\xi(\tilde{x}, y))} = y - \frac{f(y)}{f'(\xi(\tilde{x}, y))},$$

так что верно альтернативное представление функции  $g$ :

$$g(y) = \tilde{x} - \frac{f(\tilde{x})}{f'(\xi(\tilde{x}, y))}.$$

Как следствие, после интервализации этого выражения по  $y \in \mathbf{X}$ , получаем

$$g(y) = \tilde{x} - \frac{f(\tilde{x})}{f'(\xi(\tilde{x}, y))} \in \tilde{x} - \frac{f(\tilde{x})}{f'(\mathbf{X})} = \mathcal{N}(\mathbf{X}, \tilde{x}) \subseteq \mathbf{X}.$$

Так как это включение справедливо для любого  $y \in \mathbf{X}$ , то получается, что непрерывное отображение  $g$  переводит интервал  $\mathbf{X}$  в себя. Следовательно, в силу теоремы Брауэра о неподвижной точке, существует такое  $y^* \in \mathbf{X}$ , что  $g(y^*) = y^*$ . Из (4.50) тогда вытекает, что  $f(y^*) = 0$ , т. е.  $y^*$  является решением уравнения  $f(x) = 0$ . ■

**Пример 4.7.1** Применим интервальный метод Ньютона к решению уравнения (4.19) для проектирования крыши, переписав его в виде

$$l \sin \alpha - \alpha h = 0,$$

с  $l = 3.3$  и  $h = 3$ . Оно было решено ранее методом простой итерации в § 4.4в и методом секущих в § 4.4г. Для реализации интервальных вычислений в машинной арифметике двойной точности воспользуемся языком Julia и сопровождающим его пакетом `JuliaInterval` [66]. При вычислении оператора Ньютона в итерациях (4.48) в качестве точки  $\tilde{x}$  возьмём середину интервала  $\underline{X}$ , т. е.  $\tilde{x} = \text{mid } \underline{X}$ .

В качестве начального приближения нужно взять интервал, который гарантированно содержит искомое решение, но не включает неинтересное нам нулевое решение. Начав итерирование с интервала  $[0.5, 2]$ , за 7 (семь) итераций получаем интервал локализации решения в виде  $[0.7489866426973398, 0.7489866426973414]$ . Дальнейшее итерирование к уточнению уже не приводит. Полученный в примере 4.4.4 точечный ответ, как легко видеть, принадлежит найденному интервалу.

Ширина интервала ответа оказывается равной  $1.55 \cdot 10^{-15}$ , тогда как расстояние между соседними машинно-представимыми числами в районе решения — примерно  $1.1 \cdot 10^{-16}$ . Превышение в 14 раз объясняется тем, что при реализации интервального метода Ньютона для сохранения доказательности все вычисления между вещественными величинами необходимо выполнять в машинной интервальной арифметике с внешним округлением. В частности, нужно объявлять интервальным типом результаты вычисления середины  $\text{mid } \underline{X}$ . Эти предосторожности увеличивают погрешность. ■

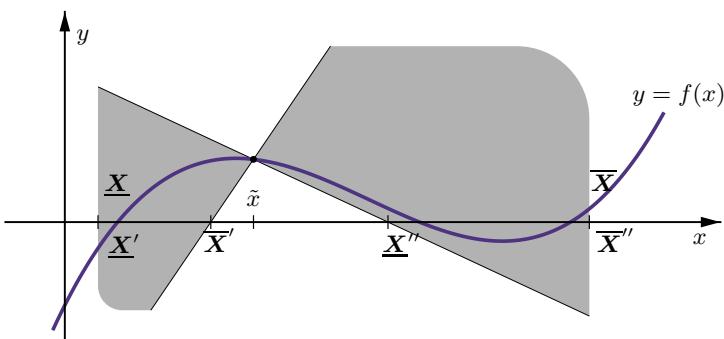


Рис. 4.24. Иллюстрация работы интервального метода Ньютона.  
Случай нульсодержащего интервала производной

Рассмотрим теперь случай  $0 \in f'(\mathbf{X})$ . Он встречается, когда на интервале  $\mathbf{X}$  имеется кратное решение  $x^*$ , в котором  $f'(x^*) = 0$ , либо когда интервал  $\mathbf{X}$  настолько широк, что содержит более одного решения. В этом случае мы тоже можем придать смысл интерваль-ному оператору Ньютона, воспользовавшись для выполнения деления  $f(\tilde{x})/f'(\mathbf{X})$  специальной интервальной арифметикой — так называемой интервальной арифметикой Кэхэна, допускающей деление на нульсо-держащие интервалы [38]. В действительности, эта модификация даже усиливает интервальный метод Ньютона, так как мы получим возмож-ность отделять различные решения друг от друга. Дело в том, что в результате выполнения шага интервального метода Ньютона при попадании нуля во внутренность интервала-делителя, когда  $0 \in \text{int } f'(\mathbf{X})$ , часто получаются два непересекающиеся интервала. Эта ситуация иллюстрируется на рис. 4.24.

Интервальная арифметика Кэхэна является расширением класси-ческой интервальной арифметики  $\mathbb{IR}$  и помимо традиционных интер-валов из  $\mathbb{IR}$  имеет своими элементами бесконечные и полубесконечные интервалы вида  $[-\infty, p]$ ,  $[q, +\infty]$ ,  $[-\infty, p] \cup [q, +\infty]$ . Мы пишем по тра-диции бесконечные концы как бы принадлежащими интервалу, хотя на самом деле это условность и  $\pm\infty$  — просто символы. Арифметические операции между традиционными интервалами в арифметике Кэхэна и в  $\mathbb{IR}$  совершенно совпадают друг с другом. Но в арифметике Кэхэна дополнительно определено деление интервалов  $\mathbf{a}$  и  $\mathbf{b}$  с  $0 \in \mathbf{b}$ , кото-рое и приводит к бесконечным интервалам. Для удобства мы выпишем соответствующие результаты в развернутой форме:

$$\mathbf{a}/\mathbf{b} = \frac{[\underline{a}, \bar{a}]}{[\underline{b}, \bar{b}]} =$$

$$= \begin{cases} a \cdot [1/\bar{b}, 1/\underline{b}], & \text{если } 0 \notin b, \\ ]-\infty, +\infty[, & \text{если } 0 \in a \text{ и } 0 \in b, \\ [\bar{a}/\underline{b}, +\infty[, & \text{если } \bar{a} < 0 \text{ и } \underline{b} < \bar{b} = 0, \\ ]-\infty, \bar{a}/\bar{b}] \cup [\bar{a}/\underline{b}, +\infty[, & \text{если } \bar{a} < 0 \text{ и } \underline{b} < 0 < \bar{b}, \\ ]-\infty, \bar{a}/\bar{b}], & \text{если } \bar{a} < 0 \text{ и } 0 = \underline{b} < \bar{b}, \\ ]-\infty, \underline{a}/\underline{b}], & \text{если } 0 < \underline{a} \text{ и } \underline{b} < \bar{b} = 0, \\ ]-\infty, \underline{a}/\bar{b}] \cup [\underline{a}/\bar{b}, +\infty[, & \text{если } 0 < \underline{a} \text{ и } \underline{b} < 0 < \bar{b}, \\ [\underline{a}/\bar{b}, +\infty[, & \text{если } 0 < \underline{a} \text{ и } 0 = \underline{b} < \bar{b}, \\ \emptyset, & \text{если } 0 \notin a \text{ и } 0 = b. \end{cases} \quad (4.51)$$

Эта операция деления на нульсодержащий интервал реализована в некоторых языках программирования и библиотеках интервальных вычислений, например, в библиотеке `IntervalArithmetic` языка Julia [66].

В заключение — необходимый комментарий о реализации интервального метода Ньютона на ЭВМ. При вычислении интервального оператора Ньютона (правой части включения (4.47)) значение  $f(\tilde{x})$ , несмотря на точечность аргумента  $\tilde{x}$ , для достижения доказательности вычислений следует находить с помощью машинной интервальной арифметики с внешним направленным округлением. Иначе возможны потеря решений и другие нежелательные феномены.

В § 4.4 мы рассматривали пример Донована–Миллера–Морелэнда, где традиционный метод Ньютона не мог найти решения уравнения с гладкой функцией ни при каком начальном приближении, отличном от самого решения. Но интервальный метод Ньютона с интервальной арифметикой Кэхэна успешно решает этот пример, что впервые было отмечено в работе [63]. Таким образом, интервальный метод Ньютона оказывается даже более сильным, чем его прародитель.

## 4.7в Многомерный интервальный метод Ньютона

Переходя к решению систем нелинейных уравнений, следует отметить, что многомерные версии интервального метода Ньютона гораздо более многочисленны, чем одномерные, и отличаются очень большим

разнообразием. В многомерном случае мы можем варьировать не только выбор точки  $\tilde{x}$ , вокруг которой осуществляется разложение, форму интервального расширения производных или наклонов функции, как это было в одномерном случае, но и способ внешнего оценивания множества решений интервальной линейной системы, к которой приводится оценивание бруса решения. В этом разделе мы рассмотрим простейшую форму многомерного интервального метода Ньютона, а также его более продвинутую версию, которую связывают с именами Хансена и Сенгупты.

Основой традиционного метода Ньютона является линеаризация функции, фигурирующей в уравнении, и последующая замена этого уравнения на линейное. Идея интервального метода Ньютона аналогична, как мы видели для одномерного случая в § 4.7б. Для построения многомерного интервального метода Ньютона необходимо ввести важную техническую конструкцию.

**Определение 4.7.2** [38, 45] Для отображения  $F : \mathbb{R}^n \supseteq D \rightarrow \mathbb{R}^m$  матрица  $S \in \mathbb{IR}^{m \times n}$  называется интервальной матрицей наклонов на брусе  $x \subseteq D$ , если для любых  $x, y \in x$  равенство

$$F(y) - F(x) = S(y - x)$$

имеет место с некоторой вещественной  $m \times n$ -матрицей  $S \in S$ .

Так как интервальная матрица наклонов  $S$  зависит, вообще говоря, от бруса  $x$ , в некоторых ситуациях необходимо обозначать её как  $S(x)$ .

С помощью интервальной матрицы наклонов можно обеспечить линеаризованное включение для приращения функции, фактически, его внешнее оценивание. На практике в качестве интервальной матрицы наклонов можно взять, например, внешнюю интервальную оценку для матрицы Якоби рассматриваемого отображения по заданному брусу. Это следует из теоремы Лагранжа о конечном приращении для вещественнонезначимых функций от нескольких переменных [14], применённой к каждой отдельной компоненте отображения  $F(x)$ .

Предположим, что на брусе  $X$  нужно найти решения системы нелинейных уравнений

$$F(x) = 0. \quad (4.52)$$

Если  $S$  — интервальная матрица наклонов отображения  $F$  на  $X$ , то для любых точек  $x, \tilde{x} \in X$  справедливо представление

$$F(x) \in F(\tilde{x}) + S(x - \tilde{x}).$$

В частности, если  $x = x^*$  — решение системы уравнений (4.52), т. е.  $F(x^*) = 0$ , то

$$0 \in F(\tilde{x}) + \mathbf{S}(x^* - \tilde{x}). \quad (4.53)$$

Вспомним характеризацию Бекка для объединённого множества решений интервальной линейной системы (теорема 4.6.1): получается, что точка  $x^*$  удовлетворяет включению (4.53) тогда и только тогда, когда она принадлежит объединённому множеству решений интервальной системы уравнений

$$\mathbf{S}(x - \tilde{x}) = -F(\tilde{x}) \quad (4.54)$$

с неизвестной переменной  $x$ .

Пусть  $\text{Encl}$  — процедура внешнего оценивания множества решений интервальной системы линейных уравнений, т. е. какой-то алгоритм, дающий внешнюю интервальную оценку для этого множества решений. Фактически, это отображение  $\text{Encl} : \mathbb{IR}^{n \times n} \times \mathbb{IR}^n \rightarrow \mathbb{IR}^n$  из множества всех интервальных  $n \times n$ -матриц и  $n$ -векторов во множество интервальных  $n$ -векторов, такое что  $\text{Encl}(\mathbf{A}, \mathbf{b}) \supseteq \Xi(\mathbf{A}, \mathbf{b})$ . Реально это может быть, например, какой-то интервальный численный метод из тех, что были кратко рассмотрены в § 4.6б, либо любой другой аналогичный. Тогда справедливо включение

$$x^* - \tilde{x} \in \text{Encl}(\mathbf{S}, -F(\tilde{x})),$$

так что

$$x^* \in \tilde{x} + \text{Encl}(\mathbf{S}, -F(\tilde{x})).$$

**Определение 4.7.3** Пусть для внешнего оценивания множеств решений интервальных линейных систем уравнений зафиксирована процедура  $\text{Encl}$ . Пусть также дано некоторое правило, которое брусу  $\mathbf{X}$  сопоставляет точку  $\tilde{x} \in \mathbf{X}$ , и для отображения  $F : \mathbb{R}^n \supseteq D \rightarrow \mathbb{R}^n$  на произвольном брусе  $\mathbf{X} \subseteq D$  известна интервальная  $n \times n$ -матрица наклонов  $\mathbf{S}(\mathbf{X})$ . Интервальнозначное отображение

$$\mathcal{N} : \mathbb{ID} \times \mathbb{R}^n \rightarrow \mathbb{IR}^n,$$

задаваемое правилом

$$\mathcal{N}(\mathbf{X}, \tilde{x}) = \tilde{x} + \text{Encl}(\mathbf{S}(\mathbf{X}), -F(\tilde{x})),$$

называется интервальным оператором Ньютона на  $\mathbb{ID}$  относительно точки  $\tilde{x}$ .

Итак, основное свойство интервального оператора Ньютона состоит в том, что для заданного бруса  $\mathbf{X}$ , в котором локализовано решение, он строит другой брус  $\mathcal{N}(\mathbf{X}, \tilde{x})$ , в котором тоже должно находиться решение системы уравнений. Следовательно, положение искомого решения можно уточнить пересечением  $\mathbf{X} \cap \mathcal{N}(\mathbf{X}, \tilde{x})$ . Если же оно пусто, то на брусе  $\mathbf{X}$  решений нет.

Дальнейшее уточнение бруса, содержащего решение системы уравнений (4.52), можно организовать с помощью итераций

$$\mathbf{X}^{(k+1)} \leftarrow \mathbf{X}^{(k)} \cap \mathcal{N}(\mathbf{X}^{(k)}, \tilde{x}^{(k)}), \quad k = 0, 1, 2, \dots,$$

для какой-то последовательности точек  $\tilde{x}^{(k)} \in \mathbf{X}^{(k)}$ . Этот итерационный процесс называется *интервальным методом Ньютона*. Чтобы обеспечить доказательность вычислений с оператором Ньютона, значение функции  $F(\tilde{x})$  следует находить с помощью машинной интервальной арифметики с внешним направленным округлением.

Как лучше выбирать центр разложения  $\tilde{x}$ ? Простейший и наиболее популярный выбор — центр бруса  $\mathbf{X}$ , т. е.  $\tilde{x} = \text{mid } \mathbf{X}$ . Если есть желание и возможность выбирать точку  $\tilde{x}$  более тщательно, то имеет смысл делать это так, чтобы величина  $\|F(\tilde{x})\|$  была, по-возможности, меньшей. Чем меньше будет норма вектор-функции  $F(\tilde{x})$ , тем меньшим будет норма векторов, образующих множество решений интервальной линейной системы

$$\mathbf{S}(x - \tilde{x}) = -F(\tilde{x}),$$

которое мы должны пересекать с исходным бруском. Скорее всего, мы получим при этом более узкую внешнюю оценку множества решений исходной нелинейной системы и более точно определим статус исследуемого бруса. Численные эксперименты как будто подтверждают этот вывод. Процедуру для уточнения центра разложения можно организовать как метод типа Ньютона, коль скоро нам известна интервальная матрица наклонов.

Внимательный взгляд на итерации интервального метода Ньютона приводит к выводу, что для его организации не нужно оценивание всего множества решений интервальной линейной системы (4.54) (которое может быть даже неограниченным, если  $\mathbf{S}$  содержит особые матрицы). Достаточно находить внешнюю оценку той части множества решений, которая принадлежит брусу  $\mathbf{X}$ . Учёт этой априорной информации позволит выполнять более точное оценивание и успешно обрабатывать ситуации с особенной матрицей  $\mathbf{S}$ . Для решения модифицированной за-

дачии хорошо подходит интервальный метод Гаусса–Зейделя, рассмотренный в § 4.6б.

Численный метод, получающийся встраиванием интервального метода Гаусса–Зейделя в описанную выше вычислительную схему вместо процедуры *Encl*, называют *методом Хансена–Сенгупты*. Он в самом деле несколько более эффективен чем исходный интервальный метод Ньютона (см. [36, 38, 40, 44, 45]), так как позволяет быстрее отсекать от исходного бруса части, которые не содержат решений.

Наиболее неблагоприятной ситуацией при работе интервального метода Ньютона и его модификаций является, конечно, появление на каком-то шаге включения

$$\mathcal{N}(\mathbf{X}^{(k)}, \tilde{x}^{(k)}) \supseteq \mathbf{X}^{(k)}$$

Тогда все последующие шаги зацикливаются на брусе  $\mathbf{X}^{(k)}$  и не дают никакого уточнения решений системы. Как поступать в этом случае? Ответ на этот вопрос рассматривается далее в § 4.8.

#### 4.7г Метод Кравчика

Пусть на брусе  $\mathbf{X}$  из  $\mathbb{R}^n$  задана система  $n$  нелинейных уравнений с  $n$  неизвестными вида (4.2)–(4.3), т. е.

$$F(x) = 0,$$

для которой требуется уточнить двусторонние границы решений. Возьмём какую-нибудь точку  $\tilde{x} \in \mathbf{X}$  и организуем относительно неё разложение функции  $F$ :

$$F(x) \in F(\tilde{x}) + \mathbf{S}(x - \tilde{x}),$$

где  $\mathbf{S} \in \mathbb{IR}^{n \times n}$  — интервальная матрица наклонов отображения  $F$  на брусе  $\mathbf{X}$  (см. § 4.7в). Если  $x^*$  — решение системы, то  $F(x^*) = 0$  и

$$0 \in F(\tilde{x}) + \mathbf{S}(x^* - \tilde{x}). \quad (4.53)$$

Но далее, в отличие от интервального метода Ньютона, мы не будем переходить к рассмотрению интервальной линейной системы (4.54), а домножим обе части этого включения слева на специально подобранную точечную  $n \times n$ -матрицу, которую нам будет удобно обозначить как  $(-\Lambda)$ :

$$0 \in -\Lambda F(\tilde{x}) - \Lambda \mathbf{S}(x^* - \tilde{x}).$$

Добавляя к обеим частям получившегося соотношения по  $(x^* - \tilde{x})$ , приходим к включению

$$x^* - \tilde{x} \in -\Lambda F(\tilde{x}) + (x^* - \tilde{x}) - \Lambda S(x^* - \tilde{x}),$$

что равносильно

$$x^* \in \tilde{x} - \Lambda F(\tilde{x}) + (I - \Lambda S)(x^* - \tilde{x}),$$

так как для неинтервального общего множителя  $(x^* - \tilde{x})$  можно воспользоваться частным случаем дистрибутивности (1.17). Наконец, если решение  $x^*$  системы уравнений предполагается принадлежащим брусу  $\mathbf{X}$ , мы можем взять интервальное расширение по  $x^* \in \mathbf{X}$  правой части полученного включения, и это даёт

$$x^* \in \tilde{x} - \Lambda F(\tilde{x}) + (I - \Lambda S)(\mathbf{X} - \tilde{x}).$$

**Определение 4.7.4** Пусть зафиксированы правила, которые всяко-му брусу  $\mathbf{X} \in \mathbb{IR}^n$  сопоставляют точку  $\tilde{x} \in \mathbf{X}$  и вещественную  $n \times n$ -матрицу  $\Lambda$ , и пусть  $S(\mathbf{X}) \in \mathbb{IR}^{n \times n}$  — интервальная матрица наклонов отображения  $F : \mathbb{R}^n \supseteq D \rightarrow \mathbb{R}^n$  на брусе  $\mathbf{X}$  из  $D$ . Отображение

$$\mathcal{K} : \mathbb{ID} \times \mathbb{R} \rightarrow \mathbb{IR}^n,$$

задаваемое выражением

$$\mathcal{K}(\mathbf{X}, \tilde{x}) := \tilde{x} - \Lambda F(\tilde{x}) + (I - \Lambda S(\mathbf{X}))(\mathbf{X} - \tilde{x}),$$

называется оператором Кравчика на  $\mathbb{ID}$  относительно точки  $\tilde{x}$ .

Вывод оператора Кравчика и сама его конструкция могут показаться несколько изощрёнными. Но его основная идея становится совершенно прозрачной, если заметить, что оператор Кравчика — это не что иное, как дифференциальная центрированная форма интервального расширения (см. § 1.6), взятая относительно точки  $\tilde{x}$ , для отображения  $\Phi(x) = x - \Lambda F(x)$ , которое возникает в правой части системы уравнений после её приведения к рекуррентному виду

$$x = \Phi(x).$$

Соответственно,  $\Lambda$  — предобуславливающая матрица, которую желательно брать такой, чтобы якобиан  $\Phi'(x) = I - \Lambda F'(x)$  был «как можно меньшим» на рассматриваемом брусе. Неплохим выбором является  $\Lambda \approx (\text{mid } \mathbf{S})^{-1}$ , т. е. обратная к середине матрицы наклонов.

В силу сказанного выше оператор Кравчика можно использовать во всех построениях и приёмах уточнения решений из § 4.7а.

Свойства оператора Кравчика сведём для удобства в одну формулировку:

**Теорема 4.7.1** Пусть  $F : \mathbb{R}^n \supseteq D \rightarrow \mathbb{R}^n$  — непрерывное по Липшичу отображение,  $\tilde{x} \in \mathbf{X} \subseteq \mathbb{I}D$  и определением 4.7.4 на  $\mathbb{I}D$  задаётся соответствующий оператор Кравчика. Тогда

- (i) каждое решение системы уравнений  $F(x) = 0$  на брусе  $\mathbf{X}$  лежит также в  $\mathcal{K}(\mathbf{X}, \tilde{x})$ ;
- (ii) если  $\mathbf{X} \cap \mathcal{K}(\mathbf{X}, \tilde{x}) = \emptyset$ , то в  $\mathbf{X}$  нет решений системы  $F(x) = 0$ ;
- (iii) если  $\mathcal{K}(\mathbf{X}, \tilde{x}) \subseteq \mathbf{X}$ , то в  $\mathbf{X}$  находится хотя бы одно решение системы уравнений  $F(x) = 0$ ;
- (iv) если  $\tilde{x} \in \text{int } \mathbf{X}$ ,  $\mathcal{K}(\mathbf{X}, \tilde{x}) \neq \emptyset$  и  $\mathcal{K}(\mathbf{X}, \tilde{x}) \subseteq \text{int } \mathbf{X}$ , то в  $\mathcal{K}(\mathbf{X}, \tilde{x})$  содержится в точности одно решение системы  $F(x) = 0$ .

Свойства (i)–(iii) оператора Кравчика нетрудно вывести из сделанного выше наблюдения о его конструкции. В частности, (iii) вытекает из теоремы Брауэра о неподвижной точке. Свойство (iv) доказывается более сложно, и подробности читатель может увидеть, например, в книге [38].

Дальнейшее уточнение решения можно организовать совершенно так же, как и в интервальном методе Ньютона, положив  $\mathbf{X}^{(0)} = \mathbf{X}$  и запуская итерации

$$\mathbf{X}^{(k+1)} \leftarrow \mathbf{X}^{(k)} \cap \mathcal{K}(\mathbf{X}^{(k)}, \tilde{x}^{(k)}), \quad k = 0, 1, 2, \dots, \quad (4.55)$$

для какой-то последовательности точек  $\tilde{x}^{(k)} \in \mathbf{X}^{(k)}$ . Этот интервальный итерационный процесс называется *методом Кравчика*. Аналогично интервальному методу Ньютона, чтобы обеспечить доказательность вычислений с оператором Кравчика, значение функции  $F(\tilde{x})$  следует находить с помощью машинной интервальной арифметики с внешним направленным округлением.

**Пример 4.7.2** Рассмотрим решение системы уравнений (4.26)

$$\begin{cases} x_1^2 + x_2^2 - 4 = 0, \\ x_1^3 - x_1 - x_2 - 1 = 0, \end{cases}$$

которая уже встречалась в § 4.5а. Для организации метода Кравчика интервальную матрицу наклонов возьмём в виде интервального расширения матрицы Якоби отображения, задаваемого левыми частями решаемой системы уравнений, т. е.

$$\begin{pmatrix} 2\mathbf{X}_1 & 2\mathbf{X}_2 \\ 3\mathbf{X}_1^2 - 1 & -1 \end{pmatrix}.$$

В качестве предобуславливающей матрицы  $\Lambda$  возьмём обратную к середине для этой матрицы, а в качестве точек  $\hat{x}^{(k)}$  — середины брусов  $\mathbf{X}^{(k)}$ ,  $k = 0, 1, 2, \dots$ .

Метод Кравчика, запущенный на брусе  $([-2, -1], [-2, 0])^\top$  за 9 итераций даёт

$$\left( \begin{bmatrix} -1.217385223395195, -1.217385223395193 \\ -1.586812281859148, -1.586812281859146 \end{bmatrix} \right),$$

а запущенный на брусе  $([0, 2], [1, 2])^\top$  за 11 итераций даёт

$$\left( \begin{bmatrix} 1.562003397594915, 1.562003397594916 \\ 1.249057799263884, 1.249057799263886 \end{bmatrix} \right).$$

Это гарантированные двусторонние оценки двух решений системы, с 15 совпадающими значащими цифрами в левой и правой границах интервалов каждой компоненты.

В то же время, для бруса  $([0, 1], [0, 2])^\top$  метод Кравчика на самой первой итерации выдаёт  $(\emptyset, [0, 1.5])^\top$ , пустое пересечение, что означает отсутствие решений. Аналогично метод Кравчика доказывает отсутствие решений системы в брусе  $([-1, 0], [-2, 0])^\top$ .

Интервальные вычисления этого примера проводились в свободной системе компьютерной математики Octave [65] с пакетом `Interval`. ■

Метод Кравчика можно применять не только к нелинейным системам уравнений, но также и к линейным, желая, к примеру, получить доказательные интервальные границы их решений.

**Пример 4.7.3** Рассмотрим  $8 \times 8$ -систему линейных алгебраических уравнений с гильбертовой матрицей, которая решалась различными численными методами в примере 3.10.3 (стр. 545) и в примере 3.11.2 (стр. 579). В последнем случае система была более или менее успешно решена методом сопряжённых градиентов и получен ответ (3.181). Но как он соотносится с идеальным математическим решением задачи?

Дело в том, что даже ввод в цифровую ЭВМ матрицы Гильbertа, элементы которой являются рациональными дробями, сопровождается неизбежными погрешностями округления, так что в действительности мы решаем систему линейных уравнений с матрицей, приближённо равной гильбертовой. Чтобы аккуратно учесть этот эффект, воспользуемся какой-нибудь системой программирования с интервальными типами данных и при формировании матрицы Гильbertа вычислим её элементы с помощью машинной интервальной арифметики с внешним направлением округлением (см. § 1.11). Получится интервальная матрица, гарантированно содержащая идеальную гильбертову матрицу. То же самое проделаем с вектором правой части, так что в целом в ЭВМ будем иметь интервальную систему линейных алгебраических уравнений, которая гарантированно содержит решаемую систему. Далее к ней можно применять различные интервальные методы для внешнего оценивания множества решений, и мы рассмотрим решение этой задачи с помощью метода Кравчика.

Воспользуемся снова системой компьютерной математики Octave [65] и её пакетом `interval`, реализующим машинную интервальную арифметику и удобные операции с интервальными векторами и матрицами. Для организации оператора Кравчика возьмём точку  $\tilde{x}$  приближённым решением точечной системы уравнений, а  $\Lambda$  — приближением к обратной точечной матрице, в качестве которой можно взять обратную к «приближённо гильбертовой» матрице, введённой без интерваллизации. И  $\tilde{x}$ , и  $\Lambda$  мы вычислим стандартными средствами из Octave, с помощью процедур `linsolve` и `inverse` соответственно, что вполне оправдано поставленными целями получения гарантированных двусторонних границ решения. Традиционные «точечные» вычислительные методы используются в качестве составных блоков в интервальном алгоритме для получения результата с качественно новыми свойствами доказательности.

Матрица наклонов отображения  $F(x) = Ax - b$  будет равна, очевидно, самой матрице системы  $A$ , и для доказательности расчётов мы должны взять эту матрицу интервальной, какой она ввелаась в ЭВМ с

помощью встроенных процедур ввода и интервальной арифметики.

Начинём с большого бруса, заведомо содержащего решение системы, —  $[-10^{20}, 10^{20}]$  по каждой компоненте. Метод Кравчика через 5 (пять) итераций стабилизируется на следующих гарантированных двухсторонних границах решения нашей системы:

$$\left( \begin{array}{l} [868359.8295714197, 868360.2540801974] \\ [-46383637.64426394, -46383614.94996058] \\ [604807442.4624705, 604807738.5775404] \\ [-3271358288.597981, -3271356686.088945] \\ [8806670017.258436, 8806674333.124724] \\ [-12463054304.38583, -12463048194.54187] \\ [8871749178.132976, 8871753528.636025] \\ [-2503936466.152937, -2503935237.984345] \end{array} \right)$$

В этом интервальном векторе левые и правые концы компонент имеют по 5–6 совпадающих значащих цифр, что весьма неплохо, учитывая обусловленность матрицы, равную  $1.5 \cdot 10^{10}$ . Результат, полученный в примере 3.11.2 с помощью метода сопряжённых градиентов, как нетрудно видеть, находится в пределах этого интервального вектора, т. е. также вполне достоверен. ■

## 4.8 Глобальное решение уравнений и систем уравнений

Если ширина бруса  $X$  велика, то на нём описанные в предшествующем параграфе методики уточнения решения могут оказаться малоуспешными в том смысле, что мы получим включение (4.45), из которого нельзя вывести никакого определённого заключения ни о существовании решения на брусе  $X$ , ни о его отсутствии. Кроме того, сам этот брус, как область потенциально содержащая решение, нисколько не будет уточнён (уменьшен).

Тогда практикуют принудительное дробление  $X$  на более мелкие подбрюсы. Выбор их количества и размеров зависит, строго говоря, от удобства реализации алгоритма и от архитектуры вычислительной системы (числа параллельных процессоров и т. п.) Далее для простоты и определённости рассмотрим *бисекцию* — разбиение бруса  $X$  на две

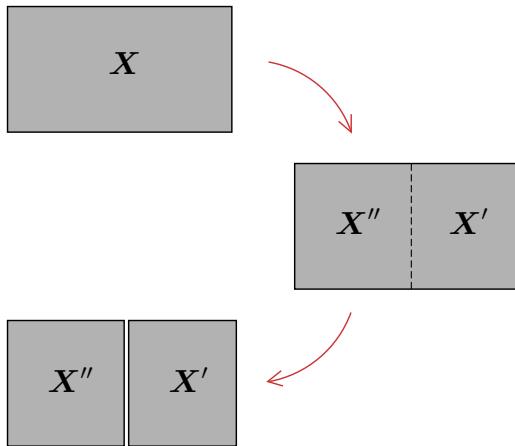


Рис. 4.25. Принудительное дробление бруса

(равные или неравные) части вдоль какой-нибудь грани, например, на половинки

$$\mathbf{X}' = (\mathbf{X}_1, \dots, [\underline{\mathbf{X}}_\iota, \text{mid } \mathbf{X}_\iota], \dots, \mathbf{X}_n),$$

$$\mathbf{X}'' = (\mathbf{X}_1, \dots, [\text{mid } \mathbf{X}_\iota, \overline{\mathbf{X}}_\iota], \dots, \mathbf{X}_n)$$

для некоторого номера  $\iota \in \{1, 2, \dots, n\}$ . При этом подбрюсы  $\mathbf{X}'$  и  $\mathbf{X}''$  называются *потомками* бруса  $\mathbf{X}$ . Далее эти потомки можно разбить ещё раз, и ещё ... — столько, сколько необходимо для достижения желаемой малости их размеров, при которой мы сможем успешно выполнять на этих брусах рассмотренные в предыдущем разделе интервальные тесты существования решений.

Если мы не хотим упустить при этом ни одного решения системы, то должны хранить все возникающие в процессе такого дробления подбрюсы, относительно которых тестом существования не доказано строго, что они не содержат решений. Организуем поэтому *рабочий список*  $\mathcal{L}$  из всех потомков начального бруса  $\mathbf{X}$ , подозрительных на содержание решений. Хотя мы называем эту структуру данных «списком», в смысле программной реализации это может быть любое хранилище брусов, организованное, к примеру, как *стек* (магазин) или *куча* и т. п. [4]. В целом же алгоритм глобального доказательного решения систе-

мы уравнений организуем в виде повторяющейся последовательности следующих действий:

- извлечение некоторого бруса из списка  $\mathcal{L}$ ,
- дробление этого бруса на потомки меньших размеров,
- проверка существования решений в каждом из подбрусов-потомков, по результатам которой мы
  - либо выдаём этот подбрус в качестве ответа к решаемой задаче,
  - либо заносим его в рабочий список  $\mathcal{L}$  для последующей обработки,
  - либо исключаем из дальнейшего рассмотрения, как не содержащий решений рассматриваемой системы.

Кроме того, чтобы обеспечить ограниченность времени работы алгоритма, на практике имеет смысл задаться некоторым порогом мелкости (малости размеров) брусов  $\delta$ , при достижении которого дальше дробить брус уже не имеет смысла. В табл. 4.3 приведён псевдокод получающегося алгоритма, который называется *методом ветвлений и отсечений*: ветвления соответствуют разбиениям исходного бруса на подбрюсы (фактически, разбиениям исходной задачи на подзадачи), а отсечения — это отбрасывание бесперспективных подбрюсов исходной области поиска.<sup>6</sup>

Неизбежные ограничения на вычислительные ресурсы ЭВМ могут не позволить решить этим алгоритмом конкретную задачу (4.41) «до конца», поскольку возможны ситуации, когда

- 1) размеры обрабатываемого бруса уже меньше  $\delta$ , но нам ещё не удается ни доказать существование в нём решений, ни показать их отсутствие;
- 2) размеры обрабатываемого бруса всё ещё больше  $\delta$ , но вычислительные ресурсы уже не позволяют производить его обработку дальше: исчерпались выделенное время, память и т. п.

---

<sup>6</sup>Стандартный английский термин для обозначения подобного типа алгоритмов — «branch-and-prune». С ними тесно связаны *методы ветвей и границ* (branch-and-bound methods), широко применяемые в вычислительной оптимизации.

Таблица 4.3. Интервальный метод ветвлений и отсечений  
для глобального доказательного решения уравнений

<b>Вход</b>
Система уравнений $F(x) = 0$ . Брус $\mathbf{X} \in \mathbb{IR}^n$ .
Интервальное расширение $\mathbf{F} : \mathbb{IX} \rightarrow \mathbb{IR}^n$ функции $F$ .
Заданная точность $\delta > 0$ локализации решений системы.
<b>Выход</b>
Список НавернякаРешения из брусов размера менее $\delta$ , которые гарантированно содержат решения системы уравнений в $\mathbf{X}$ .
Список ВозможнРешения из брусов размера менее $\delta$ , которые могут содержать решения системы уравнений в $\mathbf{X}$ .
Список Недообработанные из брусов размера более $\delta$ , которые могут содержать решения системы уравнений в $\mathbf{X}$ .
<b>Алгоритм</b>
<pre> инициализируем рабочий список <math>\mathcal{L}</math> исходным бруском <math>\mathbf{X}</math> ; DO WHILE ( ( <math>\mathcal{L} \neq \emptyset</math> ) и ( не исчерпаны ресурсы ЭВМ ) )     извлекаем из рабочего списка <math>\mathcal{L}</math> брус <math>\mathbf{Y}</math> ;     применяем к <math>\mathbf{Y}</math> тест существования решения,         его результат обозначаем также через <math>\mathbf{Y}</math> ;     IF ( в <math>\mathbf{Y}</math> доказано отсутствие решений ) THEN         удаляем брус <math>\mathbf{Y}</math> из рассмотрения     ELSE         IF ( (размер бруса <math>\mathbf{Y}</math>) &lt; <math>\delta</math> ) THEN             заносим <math>\mathbf{Y}</math> в соответствующий из списков                 НавернякаРешения или ВозможнРешения         ELSE             рассекаем <math>\mathbf{Y}</math> на потомки <math>\mathbf{Y}'</math> и <math>\mathbf{Y}''</math>             и заносим их в рабочий список <math>\mathcal{L}</math>         END IF     END IF END DO все брусы из <math>\mathcal{L}</math> перемещаем в список Недообработанные;</pre>

В реальных вычислениях остановка алгоритма табл. 4.3 может происходить поэтому не только при достижении пустого рабочего списка  $\mathcal{L}$  (когда исчерпана вся область поиска решений), но и, к примеру, при достижении определённого числа шагов или времени счёта и т. п. Тогда все брусы, оставшиеся в рабочем списке  $\mathcal{L}$ , оказываются не до конца обработанными, и мы условимся так и называть их — «недообработанные». Итак, в общем случае результатом работы нашего алгоритма должны быть три списка брусов:

список **НавернякРешения**, состоящий из брусов шириной меньше  $\delta$ , которые гарантированно содержат решения,

список **ВозможнРешения**, состоящий из брусов шириной меньше  $\delta$ , подозрительных на содержание решения, и

список **Недообработанные**, состоящий брусов, которые алгоритму не удалось обработать «до конца» и которые имеют ширину не меньше  $\delta$ .

При этом все решения рассматриваемой системы уравнений, не принадлежащие брусьям из списка **НавернякРешения**, содержатся в брусьях из списков **ВозможнРешения** и **Недообработанные**.

**Пример 4.8.1** Найдём решения системы уравнений (4.26) —

$$\begin{cases} x_1^2 + x_2^2 - 4 = 0, \\ x_1^3 - x_1 - x_2 - 1 = 0, \end{cases}$$

которая уже рассматривалась в § 4.5а и § 4.5б.

Алгоритм ветвлений и отсечений из табл. 4.3, в котором тест существования взят как однократное применение оператора Кравчика и пересечение результата с исходным бруском, будучи запущенным из начального бруса  $\mathbf{X} = ([-5, 6], [-6, 5])^\top$ , за 143 шага выдаёт два бруса из списка **НавернякРешения**:

$$\left( \begin{array}{l} [1.562003397594914, 1.562003397594917] \\ [1.249057799263884, 1.249057799263887] \end{array} \right)$$

и

$$\left( \begin{array}{l} [-1.217385223395196, -1.217385223395192] \\ [-1.586812281859149, -1.586812281859145] \end{array} \right).$$

Списки ВозможноРешения и Недообработанные оказываются пустыми. Полученные брусы содержат приближения к решению, найденные ранее с помощью обычных точечных методов в § 4.5а и § 4.5б. Но теперь мы можем доказательно утверждать, что других решений система уравнений не имеет, а найденные интервалы являются гарантированными границами настоящих решений.

Насколько эффективно использовать в качестве теста существования решения всего лишь однократное применение оператора Кравчика? Если он позволяет уменьшить размеры очередного бруса из рабочего списка, то, может быть, стоит организовать дальнейшее итерирование вида (4.55), отложив дробление бруса на потомки? Это технологические вопросы, которые невозможно определённо разрешить раз и навсегда, для всех возможных задач. Ответ на них зависит от конкретного вида системы уравнений, от её размерности и т. д.

Действительно, имеет смысл по максимуму использовать итерации (4.55) для уменьшения бруса, но если они сужают брус слишком медленно, то его дробление на более мелкие части приобретает смысл. Нередко оно сразу же качественно ускоряет процесс сужения подбрусов с помощью итераций с оператором Кравчика. Иногда же ситуация развивается противоположным образом.

Возьмём, к примеру, решение системы (4.26) с тестом существования в виде полноценных итераций (4.55), где условием остановки является уменьшение размера бруса (нормы вектора ширин компонент) не более чем на  $\theta = 20\%$ . Тогда наш алгоритм ветвлений и отсечений завершается за 73 шага, на которых оператор Кравчика был вычислен 112 раз. Налицо значительное улучшение эффективности.

При дальнейшем уменьшении порога  $\theta$  общее число шагов алгоритма ветвлений и отсечений уменьшается, но количество вычислений оператора Кравчика увеличивается, т. е. в среднем на каждом шаге начинает тратиться больше ресурсов на итерации (4.55). При  $\theta = 1\%$  алгоритм делает 41 шаг, но оператор Кравчика вычисляет 170 раз. При  $\theta = 0.1\% - 41$  шаг, и 239 вычислений оператора Кравчика, и т. д. Как видим, общая эффективность постепенно падает, если итерациям с оператором Кравчика позволять работать слишком долго. Эта картина несколько меняется с ростом размерности системы, но отмеченная тенденция сохраняется. ■

Практика эксплуатации интервальных методов для доказательного глобального решения уравнений и систем уравнений выявила также

ряд проблем и трудностей. Во многих случаях (особенно при наличии так называемых кратных решений) задачу не удается решить до конца и предъявить все гарантированные решения уравнения. Список брусов-ответов с неопределенным статусом (*ВозможноРешения* в псевдо-коде табл. 4.3) нередко не собирается исчезать ни при увеличении точности вычислений, ни при выделении дополнительного времени счета и т. п. Иногда он разрастается до огромных размеров, хотя большинство образующих его брусов возможных решений являются «фантомами» немногих реальных решений. Но эти феномены могут быть отчасти объяснены на основе теории, изложенной в § 4.3.

Решения уравнений и систем уравнений — это особые точки соответствующих векторных полей, которые, как мы могли видеть, отличаются большим разнообразием. Насколько используемые при доказательном решении систем уравнений инструменты приспособлены для выявления особых точек различных типов?

Интервальный метод Ньютона, метод Кравчика и другие тесты существования решений, которые основаны на теореме Брауэра, теореме Миранды и теореме Банаха о сжимающем отображении и которые наиболее часто используются при практических доказательных вычислениях решений уравнений, охватывают только случаи индекса  $\pm 1$  для особых точек. Если же решение системы является критической особой точкой соответствующего отображения, индекс которого не равен  $\pm 1$ , то доказать его существование с помощью вышеупомянутых результатов принципиально не получится. Это одна из причин того, почему некоторые практические интервальные алгоритмы для доказательного глобального решения уравнений и систем уравнений не могут достичь «полного успеха» в общем случае. Фактически, в этих ситуациях нужно вычислять интеграл Кронекера (4.10), что требует больших трудозатрат и часто его не реализуют в полной мере.

Помимо вышеназванной причины необходимо отметить, что список *ВозможноРешения* может соответствовать неустойчивым решениям системы уравнений, имеющим нулевой индекс (см. § 4.3в). Эти решения разрушаются при сколь угодно малых возмущениях уравнений и потому не могут быть идентифицированы никаким приближенным вычислительным алгоритмом с конечной точностью представления данных. К примеру, таковым является кратное решение квадратного уравнения (4.6)–(4.7), и хорошо известно, что он плохо находится численно как традиционными, так и интервальными подходами.

Алгоритмы ветвлений и отсечений, дополненные различными ус-

вершенствованиями и приёмами, ускоряющими сходимость, получили большое развитие в интервальном анализе в последние десятилетия (см. [36, 38, 40, 44, 45]), а реализованные на их основе программные комплексы существенно продвинули практику численного решения уравнений и систем уравнений.

## Литература к главе 4

### Основная

- [1] АЛЕФЕЛЬД Г., ХЕРЦБЕРГЕР Ю. *Введение в интервальные вычисления*. – М.: Мир, 1987.
- [2] АКРИТАС А. *Основы компьютерной алгебры с приложениями*. – М.: Мир, 1994.
- [3] БАРАХНИН В.Б., ШАПЕЕВ В.П. *Введение в численный анализ*. – СПб.–М. – Краснодар: Лань, 2005.
- [4] БАУЭР Ф.Л., Гооз Г. *Информатика. В 2-х ч.* – М.: Мир, 1990.
- [5] БАХВАЛОВ Н.С., ЖИДКОВ Н.П., КОВЕЛЬКОВ Г.М. *Численные методы*. – М.: Бином, 2003, а также другие издания этой книги.
- [6] БЕРЕЗИН И.С., ЖИДКОВ Н.П. *Методы вычислений. Т. 1–2*. – М.: Наука, 1966.
- [7] БЕРЖЕ М. *Геометрия. Т. 1, 2*. – М.: Наука, 1984.
- [8] ВЕРЖВИЦКИЙ В.М. *Численные методы. Части 1–2*. – М.: «Оникс 21 век», 2005.
- [9] ГОДУНОВ С.К. *Современные аспекты линейной алгебры*. – Новосибирск: Научная книга, 1997.
- [10] ГОДУНОВ С.К., АНТОНОВ А.Г., КИРИЛЛЮК О.П., КОСТИН В.И. *Гарантированная точность решения систем линейных уравнений в евклидовых пространствах*. – Новосибирск: Наука, 1992.
- [11] ГЭРИ М., ДЖОНСОН Д. *Вычислительные машины и труднорешаемые задачи*. – М.: Мир, 1982.
- [12] ДЕМИДОВИЧ Б.П., МАРОН А.А. *Основы вычислительной математики*. – М.: Наука, 1970.
- [13] ДЭННИС Дж., мл., ШНАВЕЛЬ Р. *Численные методы безусловной оптимизации и решения нелинейных уравнений*. – М.: Мир, 1988.
- [14] ЗОРИЧ В.А. *Математический анализ. Т. 1*. – М.: Наука, 1981. Т. 2. – М.: Наука, 1984, а также другие издания.
- [15] КАЛИТКИН Н.Н. *Численные методы*. – М.: Наука, 1978.
- [16] КАНТОРОВИЧ Л.В., АКИЛОВ Г.П. *Функциональный анализ*. – М.: Наука, 1984.
- [17] КОЛЛАТЦ Л. *Функциональный анализ и вычислительная математика*. – М.: Мир, 1969.
- [18] КОЛМГОРОВ А.Н., ФОМИН С.В. *Элементы теории функций и функционального анализа*. – М.: Физматлит, 2004, а также другие издания книги.

- [19] Крылов А.Н. *Лекции о приближённых вычислениях*. – М.: ГИТТЛ, 1954, а также более ранние издания.
- [20] Крылов В.И., Бовков В.В., Монастырный П.И. *Вычислительные методы. Т. 1–2*. – М.: Наука, 1976.
- [21] Кунц К.С. *Численный анализ*. – Киев: Техника, 1964.
- [22] Курош А.Г. *Курс высшей алгебры*. – М.: Наука, 1975.
- [23] Лебедев В.И. *Функциональный анализ и вычислительная математика*. – М.: Физматлит, 2000.
- [24] Меньшиков Г.Г. *Локализующие вычисления. Конспект лекций*. – СПб.: СПбГУ, Факультет прикладной математики–процессов управления, 2003.
- [25] Мысовских И.П. *Лекции по методам вычислений*. – СПб.: Издательство Санкт-Петербургского университета, 1998.
- [26] Опойцев В.И. *Нелинейная системостатистика*. – М.: Наука, 1986.
- [27] Орtega Дж., Рейнболдт В. *Итерационные методы решения нелинейных систем уравнений со многими неизвестными*. – М.: Мир, 1975.
- [28] Островский А.М. *Решение уравнений и систем уравнений*. – М.: Издательство иностранной литературы, 1963.
- [29] Самарский А.А., Гулин А.В. *Численные методы*. – М.: Наука, 1989.
- [30] Семёнов А.Л., Важев И.В., Кашеварова Т.П. и др. Интервальные методы распространения ограничений и их приложения // *Системная информатика*. – Новосибирск: Издательство СО РАН, 2004. – Вып. 9. – С. 245–358.
- [31] Тихонов А.Н., Арсенин В.Я. *Методы решения некорректных задач*. – М.: Наука, 1979, 1986; М.: URSS, 2022.
- [32] Трауб Дж. *Итерационные методы решения уравнений*. – М.: Мир, 1985.
- [33] Тыртышников Е.Е. *Методы численного анализа*. – М.: Академия, 2007.
- [34] Успенский В.А., Семёнов А.Л. *Теория алгоритмов: основные открытия и приложения*. – М.: Наука, 1987.
- [35] Фихтенгольц Г.М. *Курс дифференциального и интегрального исчисления. Т. 1*. – М.: Наука, 1966.
- [36] Хансен Э., Уолстер Дж.У. *Глобальная оптимизация с помощью методов интервального анализа*. – М.-Ижевск: Издательство «РХД», 2012.
- [37] Холодниок М., Клич А., Кубичек М., Марек М. *Методы анализа нелинейных динамических моделей*. – М.: Мир, 1991.
- [38] Шарый С.П. *Конечномерный интервальный анализ*. – Новосибирск: XYZ, 2025. – Электронная книга, доступная на <http://www.nsc.ru/interval/Library/InteBooks/>
- [39] Шилов Г.Е. *Математический анализ. Функции одного переменного. Ч. 1–2*. – М.: Наука, 1969.
- [40] Kearfott R.B. *Rigorous global search: Continuous problems*. – Dordrecht: Kluwer, 1996.

- [41] KELLEY C.T. *Iterative methods for linear and nonlinear equations.* – Philadelphia: SIAM, 1995.
- [42] KREINOVICH V., LAKEYEV A.V., ROHN J., KAHL P. *Computational complexity and feasibility of data processing and interval computations.* – Dordrecht: Kluwer, 1997.
- [43] MIRANDA C. Un' osservazione su un teorema di Brouwer // *Bullet. Unione Mat. Ital. Serie II.* – 1940. – T. 3. – C. 5–7.
- [44] MOORE R.E., KEARFOTT R.B., CLOUD M. *Introduction to interval analysis.* – Philadelphia: SIAM, 2009.
- [45] NEUMAIER A. *Interval methods for systems of equations.* – Cambridge: Cambridge University Press, 1990.
- [46] TREFETHEN L.N. Pseudospectra of linear operators // *SIAM Review.* 1997. – Vol. 39, No. 3. – P. 383–406.
- [47] TREFETHEN L.N., BAU D. *III Numerical linear algebra.* – Philadelphia: SIAM, 1997.

### Дополнительная

- [48] АБАФФИ Й., СПЕДИКАТО Э. Математические методы для линейных и нелинейных уравнений. Проекционные ABS-алгоритмы. – М.: Мир, 1996.
- [49] АРНОЛЬД В.И. Обыкновенные дифференциальные уравнения. – М.: Наука, 1984.
- [50] БАВЕНКО К.И. Основы численного анализа. – М.: Наука, 1986.
- [51] БАЖЕНОВ А.Н., ЖИЛИН С.И., КУМКОВ С.И., ШАРЫЙ С.П. Обработка и анализ интервальных данных. – М.–Ижевск: Издательство «ИКИ», 2024.
- [52] Высшая алгебра. Справочная математическая библиотека. – М.: Физматгиз, 1962.
- [53] ДЭВЕНПОРТ Дж., СИРЭ И., ТУРНЬЕ Э. Компьютерная алгебра. Системы и алгоритмы алгебраических вычислений. – М.: Мир, 1991.
- [54] ЗАГУСКИН В.Л. Справочник по численным методам решения алгебраических и трансцендентных уравнений. – М.: Физматгиз, 1960.
- [55] КРАСНОСЕЛЬСКИЙ М.А., ЗАБРЕЙКО П.П. Геометрические методы нелинейного анализа. – М.: Наука, 1975.
- [56] КРАСНОСЕЛЬСКИЙ М.А., ПЕРОВ А.И., ПОВОЛОЦКИЙ А.И., ЗАБРЕЙКО П.П. Векторные поля на плоскости. – М.: Физматлит, 1963.
- [57] НИРЕНБЕРГ Л. Лекции по нелинейному функциональному анализу. – М.: Мир, 1977.
- [58] ОПОЙЦЕВ В.И. Школа Опойцева: Математический анализ. – М.: URSS, 2016.
- [59] ПЕТРОВСКИЙ И.Г. Лекции по теории обыкновенных дифференциальных уравнений. – М.: Наука, 1970.
- [60] СКАРБОРО Дж. Численные методы математического анализа. – М.–Л.: ГТТИ, 1934.

- [61] ШАРАЯ И.А. IntLinIncR2 и IntLinIncR3 — пакеты программ для визуализации множеств решений интервальных линейных систем отношений. Версия для MATLAB. 2014. Свободно доступно на [http://www.nsc.ru/interval/Programing/MCodes/IntLinIncR2\\_UTF8.zip](http://www.nsc.ru/interval/Programing/MCodes/IntLinIncR2_UTF8.zip) и [http://www.nsc.ru/interval/Programing/MCodes/IntLinIncR3\\_UTF8.zip](http://www.nsc.ru/interval/Programing/MCodes/IntLinIncR3_UTF8.zip)
- [62] ABERTH O. *Precise numerical methods using C++*. – San Diego: Academic Press, 1998.
- [63] AKYILDIZ Y., AL-SUWAIYEL M.I. No pathologies for interval Newton's method // *Interval Computations*. – 1993. – No. 1. – P. 60–72.
- [64] DONOVAN G.C., MILLER A.R., MORELAND T.J. Pathological functions for Newton's method // *The American Mathematical Monthly*. – 1993. – Vol. 100, No. 1. – P. 53–58.
- [65] GNU Octave — Scientific Programming Language. <https://octave.org/>
- [66] SANDERS D.P., BENET L. JuliaInterval — interval arithmetic package for Julia. – 2014. <https://github.com/JuliaIntervals/IntervalArithmetic.jl>
- [67] MAYER G. *Interval analysis and automatic result verification*. – Berlin: De Gruyter, 2017.
- [68] KEARFOTT B. An efficient degree-computation method for a generalized method of bisection // *Numerische Mathematik*. – 1979. – Vol. 32. – P. 109–127.
- [69] HANSEN E., WALSTER G.W. Solving overdetermined systems of interval linear equations // *Reliable Computing*. – 2006. – Vol. 12. – P. 239–243.
- [70] MULLER D.E. A method for solving algebraic equations using an automatic computer // *Mathematics of Computation*. – 1956. – Vol. 10. – P. 208–215.
- [71] The NIST reference on constants, units, and uncertainty. – <http://physics.nist.gov/cuu/Constants>
- [72] O'NEIL T., THOMAS J. W. The calculation of the topological degree by quadrature // *SIAM Journal on Numerical Analysis*. – 1975. – Vol. 12, No. 5. – P. 673–680.
- [73] Pseudospectra gateway. – Электронный ресурс, см. [urlhttp://web.comlab.ox.ac.uk/projects/pseudospectra/](http://web.comlab.ox.ac.uk/projects/pseudospectra/)
- [74] Scilab — Open source software for numerical computation. <http://www.scilab.org>
- [75] STENGER F. Computing the topological degree of a mapping in  $\mathbb{R}^n$  // *Numerische Mathematik*. – 1975. – Vol. 25. – P. 23–38.
- [76] SUNAGA N. Theory of an interval algebra and its application to numerical analysis // *RAAG Memoirs*. – 1958. – Vol. 2, Misc. II. – P. 547–564.

# Обозначения

$\Rightarrow$	логическая импликация
$\iff$	логическая равносильность
$\&$	логическая конъюнкция, связка «и»
$\rightarrow$	отображение множеств; предельный переход
$\mapsto$	правило сопоставления элементов при отображении
$\leftarrow$	оператор присваивания в алгоритмах
$\emptyset$	пустое множество
$x \in X$	элемент $x$ принадлежит множеству $X$
$x \notin X$	элемент $x$ не принадлежит множеству $X$
$X \cup Y$	объединение множеств $X$ и $Y$
$X \cap Y$	пересечение множеств $X$ и $Y$
$X \setminus Y$	разность множеств $X$ и $Y$
$X \subseteq Y$	множество $X$ есть подмножество множества $Y$
$X \times Y$	прямое декартово произведение множеств $X$ и $Y$
$\text{int } X$	топологическая внутренность множества $X$
$\text{cl } X$	топологическое замыкание множества $X$
$\partial X$	граница множества $X$
$\mathbb{N}$	множество натуральных чисел
$\mathbb{R}$	множество вещественных (действительных) чисел
$\mathbb{R}_+$	множество неотрицательных вещественных чисел
$\mathbb{C}$	множество комплексных чисел

$\mathbb{I}\mathbb{R}$	множество интервалов вещественной оси $\mathbb{R}$
$\mathbb{R}^n$	множество вещественных $n$ -мерных векторов
$\mathbb{C}^n$	множество комплексных $n$ -векторов
$\mathbb{I}\mathbb{R}^n$	множество $n$ -мерных интервальных векторов
$\mathbb{R}^{m \times n}$	множество вещественных $m \times n$ -матриц
$\mathbb{C}^{m \times n}$	множество комплексных $m \times n$ -матриц
$\mathbb{I}\mathbb{R}^{m \times n}$	множество интервальных $m \times n$ -матриц
$\coloneqq$	равенство по определению
$\approx$	приблизительно равно
$\lessapprox$	приблизительно меньше или равно
$\gtrapprox$	приблизительно больше или равно
$\delta_{ij}$	символ Кронекера: 1 при $i = j$ и 0 иначе
$i$	мнимая единица
$\bar{z}$	комплексно сопряжённое к числу $z \in \mathbb{C}$
$\operatorname{sgn} a$	знак вещественного числа $a$
$[a, b]$	интервал с левым концом $a$ и правым $b$
$]a, b[$	открытый интервал с концами $a$ и $b$
$\underline{a}, \inf a$	левый конец интервала $a$
$\overline{a}, \sup a$	правый конец интервала $a$
$\operatorname{mid} a$	середина интервала $a$
$\operatorname{wid} a$	ширина интервала $a$
$\square X$	интервальная оболочка множества $X \subseteq \mathbb{R}^n$
$\operatorname{dist}$	метрика (расстояние)
$\operatorname{Dist}$	мультиметрика (векторнозначное расстояние)
$\operatorname{dom} f$	область определения функции $f$
$\operatorname{ran}(f, X)$	область значений функции $f$ на $X$
$f^\angle$	разделённая разность от функции $f$
$f(x) _a^b$	разность значений функции $f$ между $x = a$ и $x = b$
$\mathrm{d}f$	дифференциал функции $f$
$\frac{\partial f}{\partial x_i}$	частная производная функции $f$ по переменной $x_i$

$C^p[a, b]$	класс функций, непрерывно дифференцируемых вплоть до $p$ -го порядка на интервале $[a, b]$
$\mathcal{L}^2[a, b]$	класс функций, интегрируемых в смысле Лебега с квадратом на интервале $[a, b]$
$O(\cdot)$	$O$ -большое, символ Ландау
$I$	единичная матрица соответствующих размеров
$\ \cdot\ $	векторная или матричная норма
$\langle \cdot, \cdot \rangle$	скалярное произведение векторов
$u \perp v$	векторы $u$ и $v$ ортогональны (перпендикулярны)
$A^\top$	матрица, транспонированная к матрице $A$
$A^*$	матрица, эрмитово сопряжённая к матрице $A$
$A^{-1}$	матрица, обратная к матрице $A$
$\rho(A)$	спектральный радиус матрицы $A$
$\lambda(A), \lambda_i(A)$	собственные числа матрицы $A$
$\sigma(A), \sigma_i(A)$	сингулярные числа матрицы $A$
$\text{cond } A$	число обусловленности матрицы $A$
$\text{rank } A$	ранг матрицы $A$
$\det A$	определитель матрицы $A$
$\text{tr } A$	след матрицы $A$
$\mathcal{K}_i(A, r)$	подпространство Крылова матрицы $A$
$\text{diag}\{z_1, \dots, z_n\}$	диагональная $n \times n$ -матрица с элементами $z_1, \dots, z_n$ по главной диагонали
$\text{span}\{v_1, \dots, v_n\}$	линейная оболочка векторов $v_1, \dots, v_n$
$\min, \max$	операции взятия минимума и максимума
$\sum$	символ суммы нескольких слагаемых
$\prod$	символ произведения нескольких сомножителей

Интервалы и другие интервальные величины (векторы, матрицы и др.) всюду в тексте обозначаются жирным математическим шрифтом, например,  $\mathbf{A}, \mathbf{B}, \mathbf{C}, \dots, \mathbf{x}, \mathbf{y}, \mathbf{z}$ , тогда как неинтервальные (точечные) величины никак специально не выделяются. Арифметические операции с интервальными величинами — это операции классической интервальной арифметики  $\mathbb{I}\mathbb{R}$  (см. §1.5).

Если не оговорено противное, под векторами (точечными или интервальными) всюду понимаются вектор-столбцы.

Конец доказательства теоремы или предложения и конец примера выделяются в тексте стандартным знаком «■».

Значительная часть описываемых в книге алгоритмов снабжается псевдокодами на неформальном алгоритмическом языке, основные конструкции и ключевые слова которого должны быть понятны читателю из начального курса программирования. В частности, операторные скобки

DO FOR ... END DO означают оператор цикла со счётчиком, который задаётся после FOR,

DO WHILE ... END DO означают оператор цикла с предусловием, стоящим после WHILE,

IF ... THEN ... END IF или IF ... THEN ... ELSE ... END IF означают условные операторы с условием, стоящим после IF.

В циклах «DO FOR» ключевое слово «TO» означает увеличение счётчика итераций от начального значения до конечного (положительный шаг), а ключевое слово «DOWNT0» — уменьшение счётчика итераций (отрицательный шаг). По умолчанию значения счётчика изменяется на единицу.

# Краткий биографический словарь

Абель, Нильс Хенрик (Niels Henrik Abel, 1802–1829)

— норвежский математик.

Адамар, Жак Саломон (Jacques Salomon Hadamard, 1865–1963)

— французский математик.

Андронов, Александр Александрович (1901–1952)

— советский физик и механик.

Аристотель (др.-греч. *Αριστοτελης*, 384–322 годы до н.э.)

— древнегреческий философ и эрудит.

Архимед (др.-греч. *Αρχιμηδης*, 287–212 годы до н. э.)

— древнегреческий математик, естествоиспытатель и инженер.

Бабенко, Константин Иванович (1919–1987)

— советский математик и механик.

Бабушка, Иво (Ivo M. Babuška, 1926–2023)

— чешский и американский математик.

Банах, Стефан (Stefan Banach, 1892–1945)

— польский математик.

Бауэр, Фридрих Людвиг (Friedrich Ludwig Bauer, 1924–2015)

— немецкий математик.

Бахвалов, Николай Сергеевич (1934–2005)

— советский и российский математик.

Бахман, Пауль (Paul Bachmann, 1837–1920)

— немецкий математик.

Бельтрами, Эудженио (Eugenio Beltrami, 1835–1900)  
 — итальянский математик.

Бернштейн, Сергей Натаевич (1880–1968)  
 — российский и советский математик.

Биркгоф, Джордж Дэвид (George David Birkhoff, 1884–1944)  
 — американский математик.<sup>7</sup>

Больцано, Бернард (Bernard Bolzano, 1781–1848)  
 — чешский теолог, философ и математик.

Борель, Эмиль (Émile Borel, 1871–1956)  
 — французский математик и политический деятель.

Брадис, Владимир Модестович (1890–1975)  
 — русский и советский математик и педагог.

Брауэр, Лейтzen Эйберт Ян (Luitzen Egbertus Jan Brouwer, 1881–1966)  
 — голландский математик.

Браун, Эдвард (Edward Tankard Browne)  
 — американский математик.

Бубнов, Иван Григорьевич (1872–1919)  
 — русский корабельный инженер, математик и механик.

Буль, Джордж (George Boole, 1815–1864)  
 — английский математик и логик.

Бюффон, Жорж-Луи Леклерк де (Georges-Louis Leclerc de Buffon,  
 1707–1788) — французский естествоиспытатель.

Валлис, Джон (John Wallis, 1616–1703)  
 — английский математик.

ван дер Варден, Бартель Леендерт (Bartel Leendert van der Waerden,  
 1903–1996) — голландский математик.

Вандермонд, Александр Теофиль (Alexandre Theophil Vandermonde,  
 1735–1796) — французский музыкант и математик.

Вейерштрасс, Карл Теодор (Karl Theodor Weierstrass, 1815–1897)  
 — немецкий математик.

Вейль, Герман (Hermann Weyl, 1885–1955)  
 — немецкий и американский математик.

---

<sup>7</sup>Известен также американский математик Гаррет Биркгоф (1911–1996), его сын.

Виет, Франсуа (François Viète, 1540–1603)

— французский математик.

Виландт, Хельмут (Helmut Wielandt, 1910–2001)

— немецкий математик.

Галёркин, Борис Григорьевич (1871–1945)

— русский и советский механик и математик.

Гамильтон, Уильям Роэн (William Rowan Hamilton, 1805–1865)

— ирландский математик, механик и физик.

Гаусс, Карл Фридрих (Carl Friedrich Gauss, 1777–1855)

— немецкий математик, внёсший также фундаментальный вклад в численные методы, астрономию и геодезию.

Гельфанд, Израиль Моисеевич (1913–2009)

— советский математик; с 1989 года жил и работал в США.

Герон Александрийский (др.-греч. *Ηρων ο Αλεξανδρευς*, около 1 в. н.э.)

— древнегреческий математик и механик.

Гершгорин, Семён Аронович (1901–1933)

— советский математик, живший и работавший в Ленинграде.

Гёльдер, Людвиг Отто (Ludwig Otto Hölder, 1859–1937)

— немецкий математик.

Гивенс, Джеймс Уоллес (James Wallace Givens, 1910–1993)

— американский математик.

Гильберт, Давид (David Hilbert, 1862–1943)

— немецкий математик.

Грам, Йорген Педерсен (Jorgen Pedersen Gram, 1850–1916)

— датский математик.

Грегори, Джеймс (James Gregory, 1638–1675)

— шотландский математик и астроном.

Дйни, Улисс (Ulisse Dini, 1845–1918)

— итальянский математик.

Дирак, Пол (Paul Dirac, 1902–1984)

— британский физик, один из создателей квантовой теории.

Дулиттл, Майрик (Myrick Doolittle, 1830–1911)

— американский математик.

Евклид, или Эвклид (др.-греч. Εὐκλείδης, около 300 г. до н. э.)  
 — древнегреческий математик.

Жордан, Мари Энмон Камилл (Marie Ennemond Camille Jordan, 1838–1922) — французский математик.

Зейдель, Филипп Людвиг (Philipp Ludwig Seidel, 1821–1896)  
 — немецкий астроном и математик.

Йордан, Вильгельм (Wilhelm Jordan, 1842–1899)  
 — немецкий геодезист.<sup>8</sup>

Кавальери, Бонавентура (Bonaventura Cavalieri, 1598–1647)  
 — итальянский математик.

Канторович, Леонид Витальевич (1912–1986)  
 — советский математик и экономист, известный пионерским вкладом  
 в линейное программирование и функциональный анализ.

Кеплер, Иоганн (Johannes Kepler, 1571–1630)  
 — немецкий математик, астроном и механик.

Клиффорд, Уильям Кингдон (William Kingdon Clifford, 1845–1879)  
 — английский математик и философ.

Кнут, Дональд Эрвин (Donald Ervin Knuth, род. 1938)  
 — американский математик и специалист по информатике и  
 программированию.

Колмогоров, Андрей Николаевич (1903–1987)  
 — советский математик, внёсший фундаментальный вклад во многие  
 разделы математики, от логики и топологии до теории вероятностей.

Котес, Роджер (Roger Cotes, 1682–1716)  
 — английский математик.

Коши, Огюстен Луи (Augustin Louis Cauchy, 1789–1857)  
 — французский математик и механик.

Кравчик, Рудольф (Rudolf Krawczyk, 1921–2019)  
 — немецкий математик.

Крамер, Габриэль (Gabriel Cramer, 1704–1752)  
 — швейцарский математик.

---

<sup>8</sup>Не следует путать его с Паскуалем Йорданом (Pascual Jordan, 1902–1980),  
 немецким физиком и математиком.

Красносельский, Марк Александрович (1920–1997)  
— советский и российский математик.

Крейн, Селим Григорьевич (1917–1999)  
— советский и российский математик.

Кристоффель, Эльвин (Elwin Christoffel, 1829–1900)  
— немецкий математик.

Кронекер, Леопольд (Leopold Kronecker, 1823–1891)  
— немецкий математик.

Кроут, Прескотт (Prescott Durand Crout, 1907–1984)  
— американский математик.

Крылов, Алексей Николаевич (1863–1945)  
— русский и советский математик, механик и кораблестроитель.

Кублановская, Вера Николаевна (1920–2012)  
— советский и российский математик.

Кузьмин, Родион Осиевич (1891–1949)  
— русский и советский математик.

Курант, Рихард (Richard Courant, 1888–1972)  
— немецкий и американский математик.

Кэли, Артур (Arthur Cayley, 1821–1895)  
— английский математик.

Кэхэн, Уильям Мортон (William Morton Kahan, род. 1933)  
— канадский математик и специалист по компьютерам.<sup>9</sup>

Лагерр, Эдмон Никола (Edmond Nicolas Laguerre, 1834–1886)  
— французский математик.

Лагранж, Жозеф Луи (Joseph Louis Lagrange, 1736–1813)  
— французский математик и механик.

Ландау, Эдмунд (Edmund Landau, 1877–1938)  
— немецкий математик.

Ланцш, Корнелий (Cornelius Lanczos, 1893–1974)  
— американский физик и математик венгерского происхождения.

Лаплас, Пьер-Симон (Pierre-Simon Laplace, 1749–1827)  
— французский математик, механик, физик и астроном.

---

<sup>9</sup>Его фамилию нередко транслитерируют на русский язык как «Кахан».

Лебег, Анри Леон (Henri Léon Lebesgue, 1875–1941)  
 — французский математик.

Лежандр, Адриен-Мари (Adrien-Marie Legendre, 1752–1833)  
 — французский математик и механик.

Лейбниц, Готфрид Вильгельм (Gottfried Wilhelm Leibnitz, 1646–1716)  
 — немецкий философ, математик и физик, один из создателей дифференциального и интегрального исчисления.

Линдёлф, Эрнст Леонард (Ernst Leonard Lindelöf, 1870–1946)  
 финский математик.

Липшиц, Рудольф (Rudolf Lipschitz, 1832–1903)  
 — немецкий математик.

Лиувилль, Жозеф (Joseph Liouville, 1809–1882)  
 — французский математик.

Лобачевский, Николай Иванович (1792–1856)  
 — русский математик, создатель неевклидовой геометрии, внёсший также заметный вклад в численные методы алгебры.

Локуциевский, Олег Вячеславович (1922–1990)  
 — советский математик.

Ляпунов, Александр Михайлович (1857–1918)  
 — русский математик и механик, основоположник математической теории устойчивости, работавший также в теории вероятностей.

Марков, Андрей Андреевич (1856–1922)  
 — русский математик, внёсший фундаментальный вклад в теорию вероятностей и теорию случайных процессов.<sup>10</sup>

Марцинкевич, Юзеф (Józef Marcinkiewicz, 1910–1941)  
 — польский математик.

Ментен, Мод Леонора (Maud Leonora Menten, 1879–1960)  
 — канадский биохимик и гистохимик.

Микеладзе, Шалва Ефимович (1895–1976)  
 — советский математик.

Минковский, Герман (Hermann Minkowski, 1864–1909)  
 — немецкий математик.

---

<sup>10</sup>Известен также его сын, математик Андрей Андреевич Марков младший (1903–1979), основоположник советской школы конструктивной математики.

Миранда, Карло (Carlo Miranda, 1912–1982)

— итальянский математик.

Михаэлис, Леонор (Leonor Michaelis, 1875–1949)

— немецкий биохимик, физикохимик и физик.

Муавр, Абрахам де (Abraham de Moivre, 1667–1754)

— английский математик французского происхождения.

Мысовских, Иван Петрович (1921–2007)

— советский и российский математик.

Мюллер, Дэвид Юджин (David Eugene Muller, 1924–2008)

— американский математик.

Нейман, Карл Готфрид (Karl Gottfried Neumann, 1832–1925)

— немецкий математик.

фон Нейман, Джон (John von Neumann, 1903–1957)

— американский математик венгерского происхождения, известный также работами по развитию первых цифровых ЭВМ.<sup>11</sup>

Ньютон, Исаак (Isaac Newton, 1643–1727)

— английский физик и математик, заложивший основы дифференциального и интегрального исчисления и механики.

Островский, Александр Маркович (Alexander M. Ostrowski, 1893–1986)

— немецкий и швейцарский математик русского происхождения.

Паскаль, Блез (Blaise Pascal, 1623–1662)

— французский математик, физик и философ.

Перрон, Оскар (Oskar Perron, 1880–1975)

— немецкий математик.

Петров, Георгий Иванович (1912–1987)

— советский механик.

Пикар, Шарль Эмиль (Picard, Charles Émile, 1856–1941)

— французский математик.

Пирсон, Карл (Чарльз) (Karl (Charles) Pearson, 1857–1936)

— английский математик, биолог и философ.

Пойа (Полиа), Дьёрдь (иногда Джордж) (György Polya, 1887–1985)

— венгерский и американский математик.

<sup>11</sup>Его именем назван спектральный признак устойчивости разностных схем.

Радемахер, Ганс (Hans Rademacher, 1892–1969)  
— немецкий и американский математик.

Рафсон, Джозеф (Joseph Raphson, ≈1648–1715)  
— английский математик.

Риман, Бернхард (Georg-Friedrich-Bernhard Riemann, 1826–1866)  
— немецкий математик, механик и физик.

Ричардсон, Льюис Фрай (Lewis Fry Richardson, 1881–1953)  
— английский математик, физик и метеоролог.

Родриг, Бенжамен Оленд (Benjamin Olinde Rodrigues, 1795–1851)  
— французский математик и банкир.

Ролль, Мишель (Michel Rolle, 1652–1719)  
— французский математик.

Рунге, Карл Давид (Karl David Runge, 1856–1927)  
— немецкий физик и математик.

Руффини, Паоло (Paolo Ruffini, 1765–1822)  
— итальянский математик.

Рэлей, Джон Уильям (John William Reyleigh, 1842–1919)  
— английский физик.

Саад, Юсеф (Jousef Saad, род. 1950)  
— американский математик алжирского происхождения.

Самарский, Александр Андреевич (1919–2008)  
— советский и российский математик.

Сильвестр, Джеймс Джозеф (James Joseph Sylvester, 1814–1897)  
— английский математик.

Симпсон, Томас (Thomas Simpson, 1710–1761)  
— английский математик.

Сонин Николай Яковлевич (1849–1915)  
— русский математик.

Стеклов, Владимир Андреевич (1863–1926)  
— русский и советский математик и механик.

Стирлинг, Джеймс (James Stirling, 1692–1770)  
— шотландский математик.

Сунага, Теруо (Teruo Sunaga, 1929–1995)

— японский математик.

Тарский, Альфред (Alfred Tarski, 1901–1983)

— польский и американский математик и логик.

Таусски, Ольга (Olga Taussky, 1906–1995)

— американский математик.

Тейлор, Брук (Brook Taylor, 1685–1731)

— английский математик.

Тёплиц, Отто (Otto Toeplitz, 1881–1940)

немецкий математик.

Тихонов, Андрей Николаевич (1906–1993)

— советский математик.

Томас, Левелин (Llewellyn Thomas, 1903–1992)

— английский и американский физик и математик.

Тьюринг, Аллан (Alan Turing, 1912–1954)

— английский математик, логик, криптограф.

Улам, Станислав (Stanislaw Ulam, 1909–1984)

— американский математик польского происхождения.

Уолш, Джозеф (Joseph Walsh, 1895–1973)

— американский математик.

Фабер, Георг (Georg Faber, 1877–1966)

— немецкий математик.

Фаддеев, Дмитрий Константинович (1907–1989)

— советский математик.

Фаддеева, Вера Николаевна (1906–1983)

— советский математик.

Файк, (C.T. Fike, –)

— американский математик.

Фарадей, Майкл (Michael Faraday, 1791–1867)

— английский физик и химик.

Федоренко, Радий Петрович (1930–2009)

— советский и российский математик.

Ферма, Пьер (Pierre Fermat, 1601–1665)  
 — французский математик.

Фихтенгольц, Григорий Михайлович (1888–1959)  
 русский и советский математик.

Фишер, Эрнст Сигизмунд (Ernst Sigismund Fischer, 1875–1954)  
 — немецкий математик.<sup>12</sup>

Фредгольм, Эрик Ивар (Erik Ivar Fredholm, 1866–1927)  
 — шведский математик.

Фреше, Морис Рене (Maurice René Fréchet, 1878–1973)  
 — французский математик.

Фробениус, Фердинанд Георг (Ferdinand Georg Frobenius, 1849–1917)  
 — немецкий математик.

Фрэнсис, Джон (John G.F. Francis, род. 1934)  
 — английский математик и программист.

Фурье, Жан Батист (Jean Baptiste Fourier, 1768–1830)  
 — французский математик и физик.

Хансен, Элдон (Elton Robert Hansen, род. 1927)  
 — американский математик.

Хаусдорф, Феликс (Felix Hausdorff, 1868–1942)  
 — немецкий математик.

Хаусхолдер, Элстон (Alston Scott Householder, 1904–1993)  
 — американский математик.

Хевисайд, Оливер (Oliver Heaviside, 1850–1925)  
 — английский инженер, математик и физик.

Хессенберг, Карл Адольф (Karl Adolf Hessenberg, 1904–1959)  
 — немецкий математик и инженер.

Хестенс, Магнус (Magnus R. Hestenes, 1906–1991)  
 — американский математик.

Холесский, Андре-Луи (André-Louis Cholesky, 1875–1918)  
 — французский геодезист и математик.<sup>13</sup>

<sup>12</sup>Примерно к этому же времени относится жизнь и деятельность известного английского статистика и биолога Рональда Э. Фишера (1890–1962).

<sup>13</sup>В русской научной литературе его фамилия нередко транслитерируется как «Холецкий» или даже «Халецкий».

Хопф, Хайнц (Heinz Hopf, 1896–1971)

— немецкий и швейцарский математик.

Хоффман, Алан Джером (Alan Jerome Hoffman, 1924–2021)

— американский математик.<sup>14</sup>

Чебышёв, Пафнутий Львович (1821–1894)

— русский математик и механик, внёсший основополагающий вклад, в частности, в теорию приближений и теорию вероятностей.

Шёнберг, Исаак Якоб (Isaac Jacob Schönberg, 1903–1990)

— румынский и американский математик.

Шмидт, Эрхард (Erhard Schmidt, 1876–1959)

— немецкий математик.

Шрёдер, Иоганн (Johann Schröder, 1925–2007)

— немецкий математик.

Штифель, Эдуард (Eduard L. Stiefel, 1909–1978)

— швейцарский математик.

Штурм, Шарль Франсуа (Jacques Charles François Sturm, 1803–1855)

— французский математик.

Шульц, Гюнтер (Günther Schulz, 1903–1962)

— немецкий математик.

Шур, Исаи (Issai Schur, 1875–1941)

— немецкий и израильский математик.

Эдрейн, Роберт (Robert Adrain, 1775–1843)

— американский математик ирландского происхождения.

Эйлер, Леонард (Leonhard Euler, 1707–1783)

— российский математик швейцарского происхождения, внёсший фундаментальный вклад практически во все разделы математики.

Эрмит, Шарль (Charles Hermite, 1822–1901)

— французский математик.

Якоби, Карл Густав (Carl Gustav Jacobi, 1804–1851)

— немецкий математик.

Яненко, Николай Николаевич (1921–1984)

— советский математик и механик.

---

<sup>14</sup>Иногда его фамилию транслитерируют как «Гоффман».

# Предметный указатель

- $A$ -норма, 396  
 $A$ -ортогональность, 396  
 $\epsilon$ -раздупие, 601  
 $\infty$ -норма, 369  
 $L^2[a, b]$ , 229  
 $p$ -норма, 197, 370  
 $svd$ , 364  
 $O$ -большое, 150  
 $P$ -сжатие, 728  
 $LDL^\top$ -разложение, 471  
 $\varepsilon$ -решения, 706  
 $s$ -ранговое приближение матрицы, 433  
1-норма, 369  
2-норма, 199, 369  
LU-разложение, 452  
NP-полная задача, 59  
NP-трудная задача, 59, 698  
QR-алгоритм, 659, 662, 679  
QR-разложение, 477  
RQ-алгоритм, 679  
абсолютная погрешность, 15  
автоматическое дифференцирование, 160, 181  
активная подматрица, 454  
алгебра, 383  
алгебраическая кратность, 613  
алгебраическая степень точности, 251  
алгебраический интерполянт, 77  
алгебраический полином, 76  
алгоритм, 12  
алгоритм «быстрый», 772  
алгоритм «точный», 772  
алгоритм Дулигла, 459  
алгоритм Кроута, 460  
алгоритм Томаса, 501  
алгоритм неустойчивый, 49  
алгоритм устойчивый, 49  
алгоритмическое дифференцирование, 160, 181  
апостериорное оценивание, 600  
аппроксимация функций, 68  
арифметика дифференциальная, 181  
арифметическое векторное пространство, 195, 326  
балансировка матрицы, 415  
биортогональность, 345  
бисекция, 736, 799  
брюс, 38  
вариационное свойство, 153  
ведущая подматрица, 332  
ведущий минор, 335  
ведущий элемент, 453  
вектор, 326

- векторная норма, 369  
векторное поле, 710  
верная значащая цифра, 18  
верхняя трапециевидная матрица,  
    443  
верхняя треугольная матрица,  
    336, 351, 445  
веса квадратурной формулы, 250  
весовая функция, 200  
весовые множители, 199  
внешняя оценивающая функция,  
    40  
внутренность, 33, 375  
вращение векторного поля, 712  
выпуклая комбинация, 327  
вырожденный интервал, 35  
гармоника, 103  
геометрическая кратность, 613  
гильбертово пространство, 196  
главный элемент, 453  
граница, 375  
дерево Канторовича, 182  
дефект сплайна, 138  
дефектная матрица, 614  
дефектное собственное значение,  
    613  
диагонализуемая матрица, 350,  
    614  
диагональ матрицы, 331  
диагональное преобладание, 418  
дифференциальная арифметика,  
    181  
дифференциальная  
    центрированная форма,  
        44  
дифференцирование  
    автоматическое, 160, 181  
дифференцирование  
    алгоритмическое, 160,  
        181  
дифференцирование символьное,  
    160
- дифференцирование численное,  
    161  
длина вектора, 196, 329, 370  
доминирующее собственное  
    значение, 640  
доминирующий собственный  
    вектор, 640  
евклидова норма, 196, 199, 369  
евклидово пространство, 196, 329  
естественное интервальное  
    расширение, 42  
естественный сплайн, 152  
жорданова форма матрицы, 349  
жорданово разложение, 349  
задача Дирихле, 438  
задача восстановления  
    зависимостей, 207, 602  
задача вычислительно  
    корректная, 701  
задача вычислительно  
    некорректная, 701  
задача идентификации, 602  
задача интерполяции, 73  
задача интерполяции функции,  
    73  
задача наименьших квадратов  
    линейная, 552, 602  
задача некорректная, 45, 179  
задача приближения функции,  
    187  
задача сглаживания, 188  
задачи анализа данных, 67  
замыкание, 375  
значащая цифра, 17  
значимое, 27  
идемпотентность, 328  
индекс особой точки, 716  
индуцированная норма, 389  
интеграл Кронекера, 714  
интегральная метрика, 69  
интегральное скалярное  
    произведение, 200

- интегрирование численное, 249  
 интервал, 32  
 интервал открытый, 33  
 интервал полуоткрытый, 33  
 интервальная арифметика, 35  
 интервальная арифметика  
     Кэхэна, 788  
 интервальная линейная система,  
     766  
 интервальная матрица, 38  
 интервальная оболочка, 39  
 интервальная система уравнений,  
     767  
 интервальное продолжение, 40  
 интервальное расширение, 41  
 интервальный вектор, 38  
 интервальный метод Ньютона,  
     785, 792  
 интерполирование, 72  
 интерполянт, 73  
 интерполянт алгебраический, 77  
 интерполяционная квадратурная  
     формула, 253, 265  
 интерполяционный процесс, 128,  
     131, 135, 151  
 интерполяционный  
     тригонометрический  
         полином, 105, 108  
 интерполяция, 72  
 интерполяция Эрмита–Биркгофа,  
     122  
 интерполяция эрмитова, 121  
 истинное значение, 15  
 исчерпывание, 653  
 исчерпывание Виландта, 654  
 итерации с отношением Рэлея,  
     678  
 итерационные методы, 440  
 каноническая форма СЛАУ, 438  
 каноническая форма Самарского,  
     584  
 квадратичная метрика, 70  
 квадратичная сходимость, 591,  
     750, 765  
 квадратичное приближение, 198  
 квадратурная формула, 249  
 квадратурная формула  
     интерполяционная, 253,  
     265  
 квадратурный процесс, 303, 304  
 квазирасстояние, 71  
 классическая интервальная  
     арифметика, 36  
 ковариационная матрица, 436  
 коллинеарные векторы, 327  
 кольцо, 383  
 компактное множество, 190  
 комплексификация, 402  
 конгруэнтность матриц, 341  
 конечные методы, 440  
 константа Лебега, 119  
 константа Липшица, 47  
 корректность задачи, 701  
 коэффициент подавления  
     погрешности, 525, 596  
 коэффициент чувствительности,  
     47, 589  
 коэффициенты Котеса, 269  
 коэффициенты Фурье, 209, 233  
 коэффициенты перекоса, 625  
 кратное решение, 695, 696  
 кратность решения, 695  
 кратность собственного значения,  
     613  
 кратность узла, 121  
 кратный узел, 73, 121  
 критерий Сильвестра, 342  
 круги Гершгорина, 628  
 кубатурная формула, 250  
 кубическая сходимость, 667  
 лемма Кеплера, 261  
 лемма Кэхэна, 541  
 линейная зависимость, 327  
 линейная задача наименьших

- квадратов, 552, 602  
 линейная комбинация, 327  
 линейная оболочка, 210, 327  
 линейная сходимость, 750  
 линейное подпространство, 190,  
     193, 203, 327  
 линейное пространство, 154, 189,  
     190, 326  
 линейный метод интерполяции,  
     79  
 локализация, 731  
 максимум-норма, 369  
 малоранговое приближение  
     матрицы, 433  
 мантисса, 27  
 матрица, 330  
 матрица Вандермонда, 78, 416  
 матрица Гивенса, 481  
 матрица Гильберта, 231, 415  
 матрица Грама, 206, 342  
 матрица Уилкинсона, 621  
 матрица Хаусхольдера, 484  
 матрица вращения, 480  
 матрица двухдиагональная, 343,  
     503, 638  
 матрица дефектная, 614  
 матрица диагонализуемая, 614  
 матрица диагональная, 332  
 матрица единичная, 333  
 матрица квадратная, 331  
 матрица лежачая, 338, 443  
 матрица ленточная, 343  
 матрица наклонов интервальная,  
     790  
 матрица недефектная, 614  
 матрица неособенная, 334  
 матрица неполного ранга, 219,  
     334  
 матрица неразложимая, 420  
 матрица нормальная, 614  
 матрица нулевая, 331  
 матрица ортогональная, 340  
 матрица особенная, 334  
 матрица отражения, 484  
 матрица перестановки, 455  
 матрица перехода, 509  
 матрица плотно заполненная, 342  
 матрица полного ранга, 334  
 матрица положительно  
     определенная, 342  
 матрица почти треугольная, 635  
 матрица предобуславливающая,  
     519  
 матрица простой структуры, 350,  
     614  
 матрица прямоугольная, 331  
 матрица псевдообратная, 219  
 матрица разложимая, 420  
 матрица разреженная, 342  
 матрица регулярная, 334  
 матрица скалярная, 522  
 матрица стоячая, 338, 443  
 матрица строго верхняя  
     треугольная, 529  
 матрица строго нижняя  
     треугольная, 529  
 матрица строго регулярная, 461  
 матрица транспозиции, 455  
 матрица трапециевидная, 337, 443  
 матрица треугольная, 441  
 матрица трёхдиагональная, 343,  
     502  
 матрица унитарная, 340  
 матричная норма, 383  
 матричный ряд Неймана, 406  
 машинная интервальная  
     арифметика, 62  
 машинное эпсилон, 31  
 мера диагонального  
     преобладания, 536  
 метод Гаусса, 447  
 метод Гаусса–Зейделя, 533  
 метод Гаусса–Зейделя  
     интервальный, 773

- метод Гаусса–Йордана, 440  
 метод Герона, 750  
 метод Кравчика, 795  
 метод Либмана, 538  
 метод Монте–Карло, 308  
 метод Мюллера, 745  
 метод Ньютона, 749, 763  
 метод Ньютона интервальный,  
     785  
 метод Ньютона–Рафсона, 750  
 метод Ричардсона, 522, 546  
 метод Хансена–Сенгупты, 793  
 метод Хаусхольдера, 488  
 метод Холесского, 470  
 метод Чебышёва, 753, 756  
 метод Шульца, 590  
 метод Эйлера, 582  
 метод Якоби, 528  
 метод бисекции, 736  
 метод ветвлений и отсечений, 800  
 метод вращений, 483  
 метод градиентного спуска, 558  
 метод исчерпывания, 653  
 метод квадратного корня, 470  
 метод локальных разложений,  
     167, 254  
 метод минимальных невязок, 564  
 метод моментов, 546  
 метод наименьших квадратов,  
     207, 222  
 метод наискорейшего спуска, 557,  
     560  
 метод неопределённых  
     коэффициентов, 173,  
     299  
 метод одношаговый, 508  
 метод отражений, 488  
 метод парабол, 745  
 метод половинного деления, 736  
 метод последовательных  
     приближений, 739, 759  
 метод прогонки, 504  
 метод простой итерации, 509, 739,  
     759  
 метод релаксации, 540  
 метод самоисправляющийся, 510  
 метод сопряжённых градиентов,  
     577  
 метод спуска, 556  
 метод статистических испытаний,  
     308  
 метод стационарный, 509  
 метод установления, 582  
 метод хорд, 744  
 метрика, 70  
 метрика среднеквадратичная, 200  
 метрическое пространство, 70  
 минор, 335  
 множество всюду плотное, 616  
 множество замкнутое, 374  
 множество компактное, 190  
 множество ограниченное, 190  
 множество открытое, 374  
 множество решений, 767, 768  
 множество сравнения, 69  
 множество тощее, 616  
 множитель Холесского, 464  
 модуль непрерывности, 133  
 монотонность по включению, 41  
 мультиметрика, 728  
 насыщение численного метода,  
     151  
 натуральный сплайн, 152  
 невязка, 216, 510, 538  
 недефектная матрица, 350, 614  
 нелинейная интерполяция, 79  
 ненадёжная значащая цифра, 18  
 ненасыщаемый метод, 151, 295  
 неподвижная точка, 695, 722, 723,  
     729  
 неподвижная точка  
     отталкивающая, 740  
 неподвижная точка  
     притягивающая, 740

- непрерывность по Липшицу, 47  
неравенство Коши–Буняковского, 196, 370  
неравенство Минковского, 197, 370  
неравенство треугольника, 70, 71, 369  
нестационарный итерационный метод, 508  
неявный итерационный метод, 521  
нижняя треугольная матрица, 336  
норма, 71, 369  
норма Фробениуса, 384, 385  
норма Шура, 384  
норма евклидова, 196, 199  
норма индуцированная, 389  
норма операторная, 389  
норма подчинённая, 389  
норма согласованная, 383  
норма спектральная, 392  
норма энергетическая, 396  
нормальная матрица, 614  
нормальная система уравнений, 205, 218, 604  
нормальное псевдорешение, 220, 606  
нормированное пространство, 71, 189, 190, 193, 369  
нормировка, 329  
нулевая матрица, 332  
нуль кратности 1, 243  
область значений матрицы, 631  
область значений функции, 781  
обобщённая степень, 95  
обратная матрица, 335  
обратная подстановка, 442  
обратные степенные итерации, 650  
обратный ход, 448  
ограниченное множество, 190  
одношаговый метод, 508  
оператор Кравчика, 794  
оператор Ньютона интервальный, 784  
оператор перехода, 508  
операторная норма, 389  
операторная форма СЛАУ, 440  
операторное уравнение, 547  
определитель, 334, 427  
ортогонализация Грама–Шмидта, 236, 495  
ортогонализация Ланцша, 501  
ортогональная проекция, 213, 330  
ортогональная система, 329  
ортогональное дополнение, 330  
ортогональность, 212, 329  
ортонормальная система, 330  
ортонормированная система, 330  
основная теорема интервальной арифметики, 41  
особые точки, 710  
остатки, 221  
остаточный член интерполяции, 97  
остаточный член квадратурной формулы, 250  
осцилляции, 132, 270  
относительная погрешность, 16  
отношение Рэлея, 631, 678  
ошибка, 15  
перпендикуляр, 213, 330  
поверхность регрессии, 603  
погрешность абсолютная, 15  
погрешность относительная, 16  
подматрица, 332  
подматрица активная, 454  
подобие матриц, 336  
подобные матрицы, 336  
подпространства Крылова, 498, 550  
подчинённая норма, 389  
полином алгебраический, 76

- полином интерполяционный, 77  
 полином интерполяционный  
     Лагранжа, 83  
 полином интерполяционный  
     Ньютона, 94  
 полином тригонометрический, 76  
 полиномиальная трудоёмкость, 57  
 полиномы Лагерра, 247  
 полиномы Лежандра, 238  
 полиномы Сонина, 248  
 полиномы Чебышёва, 108  
 полиномы Эрмита, 247  
 полиномы Якоби, 247  
 полное метрическое  
     пространство, 377  
 полнота системы функций, 233  
 полный выбор ведущего  
     элемента, 456  
 положительно определённая  
     матрица, 342  
 полярное разложение, 362  
 понижение порядка, 655  
 порядок аппроксимации, 168  
 порядок точности метода, 168  
 порядок точности формулы, 168,  
     275  
 последовательность Коши, 724  
 потеря значащих цифр, 23  
 потеря точности, 23  
 почти решения, 706  
 правая трапецевидная матрица,  
     443  
 правило Крамера, 368  
 правило Рунге, 314  
 предобусловливание, 519, 778  
 предобусловливатель, 519  
 приближение изогеометрическое,  
     154  
 приближение квадратичное, 198  
 приближение функций, 68  
 приведённые полиномы  
     Чебышёва, 113  
 признак Адамара, 419  
 пример Бабушки–Витасека–  
     Прагера, 49  
 пример Бернштейна, 130  
 пример Донована–Миллера–  
     Морелэнда, 751  
 пример Рунге, 131  
 принцип вариационный, 550  
 принцип релаксации, 538  
 проблема собственных значений,  
     608  
 проектор, 328  
 проекционное уравнение, 547  
 проекционный метод, 546  
 проекция, 328  
 проекция ортогональная, 330  
 простое решение, 695  
 простое собственное значение, 613  
 простой нуль, 243  
 простой узел, 73  
 пространство гильбертово, 196  
 пространство евклидово, 196, 329  
 пространство метрическое, 70  
 пространство нормированное,  
     189, 369  
 пространство строго  
     нормированное, 193  
 пространство унитарное, 195, 329  
 прямая подстановка, 442  
 прямая сумма, 328  
 прямой ход, 447  
 прямые методы, 440  
 псевдометрика, 71  
 псевдообратная матрица, 219  
 псевдорасстояние, 71  
 псевдорешение, 217, 606  
 псевдорешение нормальное, 220  
 равномерная метрика, 69, 70  
 разделённая разность, 84  
 разложение Жордана, 349  
 разложение Холесского, 464  
 разложение Шура, 351

- разложение полярное, 362  
 разложение сингулярное, 361  
 разложение спектральное, 350  
 разложение треугольное, 452  
 размер матрицы, 331  
 разностные уравнения  
     трёхточечные, 503  
 разность вперёд, 162  
 разность назад, 162  
 разность центральная, 163  
 разрешимость, 697  
 ранг матрицы, 333, 429  
 ранг столбцовий, 334  
 ранг строчный, 334  
 расстояние, 70  
 расщепление матрицы, 520  
 рациональная функция, 41  
 регрессионная линия, 603  
 регрессия, 603  
 регуляризация, 52  
 рекуррентный вид системы, 512,  
     694  
 рекуррентный вид уравнения, 694  
 решение общее, 367  
 решение системы уравнений, 366,  
     694  
 решение уравнения, 694  
 решение частное, 367  
 ряд Фурье, 209  
 сведение задачи, 58  
 сводимость, 58  
 сдвиг Уилкинсона, 667  
 сдвиг спектра, 651  
 середина интервала, 33, 38  
 сетка, 73, 250  
 сечение массива, 489  
 скатие, 722  
 сжимающее отображение, 722  
 символьное дифференцирование,  
     160  
 симметричный QR-алгоритм, 666  
 сингулярное разложение, 361  
 сингулярное число, 354  
 сингулярный вектор, 354  
 система линейных  
     алгебраических  
     уравнений, 366, 437  
 система неоднородная, 366  
 система неразрешимая, 366  
 система несовместная, 366  
 система нормальных уравнений,  
     205, 218  
 система однородная, 366  
 система разрешимая, 366  
 система совместная, 366  
 система уравнений, 693  
 скалярное произведение, 198, 203,  
     286, 328  
 след матрицы, 336, 385  
 собственное значение, 344, 608  
 собственное число, 344, 608  
 собственный вектор, 344, 608  
 согласованная норма, 383  
 сомнительная значащая цифра,  
     18  
 составная формула Симпсона, 278  
 составная формула  
     прямоугольников, 276,  
     297  
 составная формула трапеций, 276  
 спектр матрицы, 344  
 спектральная норма, 392  
 спектральное разложение, 350  
 спектральный радиус, 400, 640  
 сплайн, 137  
 среднеквадратичная метрика, 69,  
     200  
 среднеквадратичное  
     приближение, 198  
 стационарный итерационный  
     метод, 508  
 стационарный метод Ричардсона,  
     522  
 степенной метод, 642

- степень отображения, 715  
 степень сплайна, 137  
 степень точности алгебраическая, 251  
 степень точности  
     тригонометрическая, 296  
 строго верхняя треугольная  
     матрица, 337, 529  
 строго нижняя треугольная  
     матрица, 337, 529  
 строго нормированное  
     пространство, 193  
 строго регулярная матрица, 461  
 субдистрибутивность, 37  
 схема единственного деления, 448  
 сходимость квадратичная, 750,  
     765  
 сходимость кубическая, 667  
 сходимость линейная, 750  
 сходимость по норме, 376, 393  
 сходимость по форме, 660  
 сходимость покомпонентная, 380  
 сходимость поэлементная, 395  
 табулирование, 74  
 теорема Абеля–Руффини, 325,  
     612, 743  
 теорема Алберга–Нильсона, 421  
 теорема Банаха о неподвижной  
     точке, 723  
 теорема Бауэра–Файка, 619  
 теорема Больцано–Коши, 735  
 теорема Бореля, 190  
 теорема Брауэра о неподвижной  
     точке, 715  
 теорема Вейерштрасса, 129  
 теорема Вейля, 622  
 теорема Виландта–Хофмана,  
     622  
 теорема Гамильтона–Кэли, 344  
 теорема Гершгорина, 628  
 теорема Канторовича о методе  
     Ньютона, 764  
 теорема Кронекера, 715  
 теорема Кронекера–Капелли, 367  
 теорема Леви–Деспланка, 420  
 теорема Марцинкевича, 136  
 теорема Миранды, 737  
 теорема Мысовских, 297  
 теорема Островского, 617  
 теорема Островского–Райха, 543  
 теорема Самарского, 586  
 теорема Стеклова–Пойа, 304  
 теорема Таусски, 421  
 теорема Тёплица–Хаусдорфа, 631  
 теорема Фабера, 135  
 теорема Фредгольма, 368  
 теорема Холладея, 152  
 теорема Шрёдера о неподвижной  
     точке, 729  
 теорема Штейна–Розенберга, 537  
 теорема Экарта–Янга, 435  
 теорема о QR-разложении, 478  
 теорема о малоранговом  
     приближении, 434  
 теорема о перпендикуляре, 214,  
     238  
 теорема о разложении  
     Холлесского, 464  
 теорема о сингулярном  
     разложении, 361  
 теорема о сходимости  
     интерполяционного  
     процесса, 135  
 теорема об LU-разложении, 457  
 тест существования решения, 782  
 топологическая структура, 373  
 топология, 373, 702  
 точечная величина, 33  
 точка внутренняя, 374  
 точка притяжения, 740  
 тощее множество, 616  
 трансвекция, 449  
 трансформация Гаусса, 605

- трапецевидная матрица, 337, 443  
треугольная матрица, 441  
треугольное разложение, 452  
тригонометрическая  
    интерполяция, 102  
тригонометрическая система  
    функций, 232  
тригонометрическая степень  
    точности, 296  
тригонометрический полином, 76  
тригонометрический ряд Фурье,  
    233  
трёхдиагональная матрица, 147  
угловая подматрица, 332  
узел кратный, 73, 121  
узел простой, 73  
узлы интерполяции, 73  
узлы квадратурной формулы,  
    250, 253, 279  
узлы сплайна, 138  
унитарное пространство, 195, 329  
уравнение Лапласа, 438, 538, 593  
уравнение алгебраическое, 697,  
    730  
уравнение второго рода, 695  
уравнение первого рода, 695  
уравнение трансцендентное, 730  
условие Гёльдера, 134  
условие Дини–Липшица, 134  
условие Липшица, 47  
условие остановки, 756  
условия Бубнова–Галёркина, 549  
условия Галёркина, 549  
условия Петрова–Галёркина, 548  
устойчивость алгоритма, 48  
флопс, 57, 324  
формула Муавра, 111  
формула Ньютона–Лейбница, 248  
формула Родрига, 238  
формула Симпсона, 260  
формула Стирлинга, 117, 324  
формула Эрмита, 108  
формула квадратурная, 249  
формула кубатурная, 250  
формула парабол, 260  
формула прямоугольников, 253  
формула средних  
    прямоугольников, 254  
формула трапеций, 257  
формулы Гаусса, 280  
формулы Лобатто, 296  
формулы Маркова, 296  
формулы Ньютона–Котеса, 253  
формулы численного  
    дифференцирования,  
    162, 165  
фробениусова норма, 384, 385  
фундаментальная  
    последовательность, 724  
функции Радемахера, 234  
функции Уолша, 234  
функционал энергии, 553  
функция Хевисайда, 227  
функция единичного скачка, 227  
функция рациональная, 41  
функция целая, 131  
характеризация Бекка, 768  
характеристический полином  
    матрицы, 344  
характеристическое уравнение  
    матрицы, 344, 609  
хессенбергова форма, 635  
целая функция, 131  
целевая функция, 555  
центральная разность, 164  
центрированная форма, 43  
частичный выбор ведущего  
    элемента, 453  
чебышёвская метрика, 69, 70  
чебышёвская норма, 369  
чебышёвская сетка, 116  
чебышёвские узлы, 116  
числа Кристоффеля, 290  
численное дифференцирование,

- 160  
численное интегрирование, 249  
численные методы анализа, 67  
число обусловленности, 409  
число с плавающей точкой, 27  
числовой образ матрицы, 631  
шаблон формулы, 166  
шаг сетки, 95, 142  
шар, 371, 374  
ширина интервала, 33, 38  
эквивалентные нормы, 377, 393  
экспоненциальная трудоёмкость,  
    57  
экстраполяция, 100  
экстремальное свойство  
    сплайнов, 154  
экстремум глобальный, 555  
экстремум локальный, 555  
элементарная матрица  
    перестановки, 455  
энергетическая норма, 396  
энергии функционал, 553  
эрмитова интерполяция, 122