

С.П. Шарый

Курс
ВЫЧИСЛИТЕЛЬНЫХ
МЕТОДОВ

Курс ВЫЧИСЛИТЕЛЬНЫХ МЕТОДОВ

С. П. ШАРЫЙ

Федеральный исследовательский центр
информационных и вычислительных технологий
Новосибирский государственный университет

Новосибирск – 2020

Книга является систематическим учебником по курсу вычислительных методов и написана на основе лекций, читаемых автором на механико-математическом факультете Новосибирского государственного университета. Особенностью книги является изложение методов интервального анализа и результатов конструктивной математики, связанных с традиционными разделами численного анализа.

Оглавление

Предисловие	9
Глава 1. Введение	10
1.1 Погрешности приближённых величин	12
1.2 Погрешности и вычисления	16
1.3 Компьютерная арифметика	21
1.4 Интервальная арифметика	27
1.5 Интервальные расширения функций	33
1.6 Обусловленность математических задач	37
1.7 Устойчивость алгоритмов	40
1.8 Элементы конструктивной математики	43
1.9 Сложность задач и трудоёмкость алгоритмов	45
1.10 Доказательные вычисления на ЭВМ	48
Литература к главе 1	51
Глава 2. Численные методы анализа	54
2.1 Введение	54
2.2 Интерполирование функций	57
2.2а Постановка задачи и её свойства	57
2.2б Интерполяционный полином Лагранжа	63
2.2в Разделённые разности и их свойства	66
2.2г Интерполяционный полином Ньютона	74
2.2д Погрешность алгебраической интерполяции	78
2.3 Полиномы Чебышёва	84
2.3а Определение и основные свойства	84
2.3б Применения полиномов Чебышёва	89
2.3в Обусловленность алгебраической интерполяции	91

2.4	Интерполяция с кратными узлами	94
2.5	Общие факты интерполяции	100
2.5а	Интерполяционный процесс	100
2.5б	Сводка результатов и обсуждение	102
2.6	Сплайны	110
2.6а	Элементы теории	110
2.6б	Интерполяционные кубические сплайны	114
2.6в	Погрешность интерполирования сплайнами	119
2.6г	Экстремальное свойство кубических сплайнов	121
2.7	Нелинейные методы интерполяции	123
2.8	Численное дифференцирование	129
2.8а	Интерполяционный подход	130
2.8б	Оценка погрешности дифференцирования	135
2.8в	Метод неопределённых коэффициентов	142
2.8г	Полная погрешность дифференцирования	144
2.9	Алгоритмическое дифференцирование	148
2.10	Приближение функций	151
2.10а	Обсуждение постановки задачи	151
2.10б	Существование и единственность приближения	154
2.10в	Приближение в евклидовом подпространстве	160
2.10г	Геометрия наилучшего приближения	168
2.10д	Приближение из линейной оболочки векторов	169
2.10е	Псевдорешения систем линейных уравнений	171
2.10ж	Среднеквадратичное приближение функций	177
2.11	Полиномы Лежандра	184
2.11а	Мотивация и определение	184
2.11б	Основные свойства полиномов Лежандра	190
2.12	Численное интегрирование	195
2.12а	Постановка и обсуждение задачи	195
2.12б	Простейшие квадратурные формулы	199
2.12в	Квадратурная формула Симпсона	205
2.12г	Интерполяционные квадратурные формулы	212
2.12д	Дальнейшие формулы Ньютона-Котеса	215
2.13	Квадратурные формулы Гаусса	220
2.13а	Задача оптимизации квадратурных формул	220
2.13б	Простейшие квадратуры Гаусса	222
2.13в	Выбор узлов для квадратурных формул Гаусса	226
2.13г	Практическое применение формул Гаусса	229
2.13д	Погрешность квадратур Гаусса	232

2.13е	Метод неопределённых коэффициентов	236
2.14	Составные квадратурные формулы	237
2.15	Сходимость квадратур	243
2.16	Вычисление интегралов методом Монте-Карло	248
2.17	Правило Рунге для оценки погрешности	253
	Литература к главе 2	254
Глава 3.	Численные методы линейной алгебры	260
3.1	Задачи вычислительной линейной алгебры	260
3.2	Теоретическое введение	263
3.2а	Необходимые сведения из линейной алгебры	263
3.2б	Основные понятия теории матриц	266
3.2в	Собственные числа и собственные векторы	274
3.2г	Разложения матриц, использующие их спектр	278
3.2д	Сингулярные числа и сингулярные векторы	280
3.2е	Сингулярное разложение матриц	287
3.2ж	Системы линейных алгебраических уравнений	290
3.3	Нормы векторов и матриц	291
3.3а	Векторные нормы	291
3.3б	Топология на векторных пространствах	295
3.3в	Матричные нормы	301
3.3г	Подчинённые матричные нормы	305
3.3д	Топология на множествах матриц	310
3.3е	Энергетическая норма	313
3.3ж	Спектральный радиус	317
3.3з	Матричный ряд Неймана	323
3.4	Приложения сингулярного разложения	325
3.4а	Исследование неособенности и ранга матриц	325
3.4б	Решение систем линейных уравнений	327
3.4в	Малоранговые приближения матрицы	329
3.4г	Метод главных компонент	332
3.5	Обусловленность систем линейных уравнений	333
3.5а	Число обусловленности матриц	333
3.5б	Примеры хорошо и плохо обусловленных матриц	339
3.5в	Матрицы с диагональным преобладанием	342
3.5г	Практическое применение числа обусловленности	347
3.6	Прямые методы решения линейных систем	351
3.6а	Основные понятия	351

3.6б	Решение треугольных и трапецевидных линейных систем	355
3.6в	Метод Гаусса для решения линейных систем . . .	357
3.6г	Матричная интерпретация метода Гаусса	360
3.6д	Метод Гаусса с выбором ведущего элемента	364
3.6е	Существование LU-разложения	368
3.6ж	Разложение Холецкого	373
3.6з	Метод Холецкого	375
3.7	Методы на основе ортогональных преобразований	380
3.7а	Обусловленность и матричные преобразования . .	380
3.7б	Ортогональность и матричные вычисления	384
3.7в	QR-разложение матриц	386
3.7г	Ортогональные матрицы отражения	389
3.7д	Метод Хаусхолдера	392
3.7е	Матрицы вращения и метод вращений	397
3.7ж	Процессы ортогонализации	401
3.8	Метод прогонки	406
3.9	Стационарные итерационные методы	413
3.9а	Краткая теория	413
3.9б	Сходимость стационарных одношаговых методов	417
3.9в	Подготовка системы к итерационному процессу .	423
3.9г	Оптимизация скалярного предобуславливателя . .	427
3.9д	Итерационный метод Якоби	432
3.9е	Итерационный метод Гаусса-Зейделя	437
3.9ж	Методы релаксации	443
3.10	Нестационарные итерационные методы	449
3.10а	Теоретическое введение	449
3.10б	Метод спуска для минимизации функций	455
3.10в	Наискорейший градиентный спуск	458
3.10г	Метод минимальных невязок	462
3.10д	Метод сопряжённых градиентов	466
3.11	Методы установления	479
3.12	Теория А.А. Самарского	481
3.13	Вычисление определителей и обратных матриц	485
3.14	Оценка погрешности приближённого решения	488
3.15	Линейная задача наименьших квадратов	492
3.15а	Постановка задачи и основные свойства	492
3.15б	Численные методы для линейной задачи наименьших квадратов	497

3.16	Проблема собственных значений	500
3.16а	Обсуждение постановки задачи	500
3.16б	Обусловленность проблемы собственных значений	505
3.16в	Коэффициенты перекося матрицы	513
3.16г	Круги Гершгорина	516
3.16д	Отношение Рэлея	519
3.16е	Предварительное упрощение матрицы	522
3.17	Численные методы для несимметричной проблемы собственных значений	525
3.17а	Степенной метод	525
3.17б	Обратные степенные итерации	533
3.17в	Сдвиги спектра	535
3.17г	Базовый QR-алгоритм	538
3.17д	Модификации QR-алгоритма	541
3.18	Численные методы для симметричной проблемы собственных значений	544
3.18а	Метод Якоби	545
3.18б	Численные методы сингулярного разложения	552
	Литература к главе 3	552
Глава 4.	Решение нелинейных уравнений и их систем	559
4.1	Обзор постановок задачи	559
4.2	Вычислительно-корректные задачи	561
4.2а	Предварительные сведения и определения	561
4.2б	Задача решения уравнений не является вычислительно-корректной	564
4.2в	ϵ -решения уравнений	566
4.2г	Недостаточность ϵ -решений	568
4.3	Векторные поля и их вращение	571
4.3а	Векторные поля	571
4.3б	Вращение векторных полей	573
4.3в	Индексы особых точек	576
4.3г	Устойчивость особых точек	577
4.3д	Вычислительно-корректная постановка	579
4.4	Классические методы решения уравнений	580
4.4а	Предварительная локализация решений	581
4.4б	Метод дихотомии	582
4.4в	Метод простой итерации	585
4.4г	Метод Ньютона и его модификации	588

4.4д	Методы Чебышёва	592
4.5	Классические методы решения систем уравнений	595
4.5а	Метод простой итерации	595
4.5б	Метод Ньютона и его модификации	596
4.6	Интервальные системы линейных уравнений	598
4.6а	Интервальный метод Гаусса-Зейделя	599
4.7	Интервальные методы решения уравнений	602
4.7а	Основы интервальной техники	604
4.7б	Одномерный интервальный метод Ньютона	607
4.7в	Многомерный интервальный метод Ньютона	612
4.7г	Метод Кравчика	614
4.8	Глобальное решение уравнений и систем	616
	Литература к главе 4	621
Обозначения		626
Краткий биографический словарь		630
Предметный указатель		638

Предисловие

Представляемая вниманию читателей 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 предшествующих учебниках по предмету. Условно материал книги можно 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.

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 краткий очерк нелинейного анализа (теории степени отображения), который необходим при тщательном анализе решения систем нелинейных уравнений. Наконец, читатель найдёт в книге изложение идей конструктивной математики и теории сложности вычислений, тесно связанных с предметом математики вычислительной.

Автор благодарен своей жене Ирине за любовь, моральную поддержку и неоценимую помощь в работе.

Глава 1

Введение

Курс методов вычислений является частью более широкой математической дисциплины — вычислительной математики, 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 получение конструктивных представлений (приближений) для различных математических объектов. С 70-х годов XX века, когда качественно нового уровня достигло 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 символичные 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 в общем дереве математических наук замечательно своей тесной связью с практикой. В Новом времени в связи с общим бурным развитием наук вычислительная математика выдели-

лась в самостоятельное научное направление в конце XIX – начале XX века.

Развитие вычислительной математики в различные исторические периоды имело свои особенности и акценты. Начиная с античности (вспомним Архимеда) и вплоть до Нового времени вычислительные методы гармонично входили в сферу научных интересов крупнейших математиков — И. Ньютона, Л. Эйлера, К.Ф. Гаусса, Н.И. Лобачевского, К.Г. Якоби, П.Л. Чебышёва и многих других, чьи имена остались в названиях популярных численных методов. Дальнейшее развитие и дифференциация математики, дробление её на ветви и 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 [13], вышедших в 1906 году и выдержавших семь изданий.

В XX веке, и особенно в его второй половине, на первый план выдвинулась разработка и применение конкретных практических алгоритмов для решения сложных задач математического моделирования (в основном, вычислительной физики, механики и управления). Необходимо было запускать и наводить ракеты, улучшать характеристики летательных аппаратов, судов, автомобилей и других сложных технических устройств, учиться управлять ими и т. п.

На развитие вычислительной математики очень большое влияние оказывали конкретные способы вычислений и вычислительные устройства, которые возникали по ходу развития технологий и применялись в процессах вычислений. В частности, огромное по своим последствиям влияние было испытано вычислительной математикой в середине XX века в связи с появлением электронных цифровых вычислительных машин, кратко называемых ныне «компьютерами».¹

Три типа задач, в основном, интересуют нас в связи с процессом вычислений:

- Как конструктивно найти (вычислить) тот или иной математический объект или его конструктивное приближение? К примеру, как найти производную, интеграл, решение дифференциального уравнения и т. п.?

¹Строго говоря, термин «компьютер» является более широким по значению, и электронные цифровые вычислительные машины являются одной из его разновидностей. Вообще, компьютеры могут быть не только электронными, но и механическими, оптическими, биологическими, квантовыми и т. п.

- Какова трудоёмкость нахождения тех или иных объектов? может ли она быть уменьшена и как именно?
- Если алгоритм для нахождения некоторого объекта уже известен, то как наилучшим образом организовать вычисления по этому алгоритму на том или ином конкретном вычислительном устройстве? Например, чтобы ускорить его выполнение. Чтобы при этом уменьшить погрешность вычисления и/или сделать его менее трудоёмким.

Вопросы из последнего пункта сделались особенно актуальными в связи с развитием различных архитектур электронных вычислительных машин, в частности, в связи с вхождением в нашу повседневную жизнь многопроцессорных и параллельных компьютеров.

Ясно, что все три отмеченных выше типа вопросов тесно связаны между собой. К примеру, если нам удастся 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 различные вычислительные методы для решения тех или иных конкретных задач относятся к своим специфичным математическим дисциплинам. Например, численные методы для отыскания экстремумов различных функций являются предметом вычислительной оптимизации, теории принятия решений и исследования операций.

1.1 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,

то за редким исключением они не могут быть произведены абсолютно точно. То же самое относится к результатам большинства вычислений.

Погрешностью приближённого значения \tilde{x} какой-либо величины называют разность между \tilde{x} и точным значением x^* этой величины, т. е. $\tilde{x} - x^*$. На практике точное значение x^* интересующей нас величины, 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.²

Во-вторых, вместо точной абсолютной погрешности приходится довольствоваться её приближёнными верхними оценками. Наилучшую возможную в данных условиях оценку сверху для абсолютной погрешности называют *предельной* (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 Δ — 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 неравенства 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.

На практике указание одной только абсолютной погрешности недостаточно для характеристики качества рассматриваемого приближения. К примеру, для $x^* = 10$ абсолютная погрешность, равная, скажем, единице, соответствует довольно грубому приближению, тогда как для

²В обывденной 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.

$x^* = 1000$ та же погрешность обеспечивается 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.

Идеальным было бы относить абсолютную погрешность к модулю точного значения x^* , т. е. определять относительную погрешность как

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

Но в условиях недоступности 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 практичности 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 $\tilde{\Delta}$. Как следствие, оперируют верхними оценками для $\tilde{\delta}$. *Предельной относительной погрешностью* некоторого приближённого значения называют число δ , в данных условиях наилучшим образом оценивающее сверху его относительную погрешность. Таким образом, если Δ — 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

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

Говоря про абсолютную или относительную погрешность, обычно опускают эпитет «предельная», поскольку именно предельные (граничные) погрешности являются реально доступными нам величинами,

с которыми и работают на практике.³ При этом именно относительная погрешность является наиболее адекватным эквивалентом популярного, но нестрогого понятия «точности» приближённой величины. Фактически, относительная погрешность показывает, сколько в приближённой величине истинного значения и насколько мы можем в ней сомневаться.

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», т. е. даёт существенную информацию о его погрешности. Например, в каждом из десятичных чисел 0.1234567, 1234567 и 1234.567 значащими являются по 7 цифр, начиная с 1.

Часто различают *верные* и *сомнительные* значащие цифры приближённого числа. Значащая цифра называется верной, если абсолютная погрешность числа не превосходит половины единицы разряда, который соответствует этой цифре. Очевидно, что большего мы требовать от значащей цифры не можем. Если это условие не выполнено, то значащая цифра 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 говорят о количестве его верных 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 в представлении чисел ненадёжные цифры. Обычно принимают за правило писать числа так, чтобы все их значащие цифры кроме, может быть, последней были верны, а последняя цифра была бы сомнительной не более чем на единицу. Например, согласно этому правилу число 1234000, у которого цифра 4 уже неточна и может быть равна 3, 4, 5 или 6, нужно записывать в виде $1.23 \cdot 10^6$.

³Они могут быть, к примеру, даны в спецификациях используемых технических устройств, могут быть а priori оценены из каких-либо содержательных соображений или же найдены из практики и т. п.

1.2 Погрешности и вычисления

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 вопрос решается с помощью так называемой интервальной арифметики, которая рассматривается далее в §1.4. Здесь мы приводим несколько иное решение вопроса, иногда более удобное для теории или практического использования, но имеющее приближённый характер для умножения и деления.

Предложение 1.2.1 *Абсолютная погрешность суммы или разности приближённых чисел равна сумме абсолютных погрешностей операндов.*

Доказательство. Если x_1^*, x_2^* — точные значения рассматриваемых чисел, \tilde{x}_1, \tilde{x}_2 — их приближённые значения, а Δ_1, Δ_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

дальше полная погрешность суммы уже не сможет стать меньше погрешности этого слагаемого, сколь бы точными не 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 1.2.2 *Если приближённые величины \tilde{x}_1, \tilde{x}_2 имеют абсолютные погрешности Δ_1 и Δ_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, кроме того, $\tilde{x}_2 \neq 0$ и известно, что точное значение этой величины x_2^* — ненулевое, а её относительная погрешность $\delta_2 = \Delta_2/|x_2^*|$ меньше единицы, то абсолютная погрешность частного \tilde{x}_1/\tilde{x}_2 не превосходит*

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

Интересно, что в формуле для абсолютной погрешности частного оказалось задействована относительная погрешность делителя. Как уже упоминалось, точные численные результаты для операций между приближёнными величинами даются формулами интервальной арифметики, рассматриваемой ниже в §1.4.

Доказательство. Имеем

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

что доказывает первое утверждение.

Доказательство второго утверждения:

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

Рассмотрим теперь эволюцию относительной погрешности в вычислениях.

Предложение 1.2.3 *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 относительных погрешностей.*

Доказательство. Пусть складываются две приближённые величины, значения которых равны \tilde{x}_1 и \tilde{x}_2 , а относительные погрешности суть δ_1 и δ_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

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

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 — модуль значения суммы, то полученное неравенство завершает доказательство предложения.

Адаптация этих рассуждений для нижней границы относительной погрешности очевидна. ■

Ситуация с относительной погрешностью принципиально меняется, когда в сумме слагаемые имеют разный знак, т. е. она является разностью. 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) делается очень маленьким и относительная погрешность результата вычитания может катастрофически возрасти.

Пример 1.2.1 Рассмотрим вычитание чисел 1001 и 1000, каждое из которых является приближённым и известным с абсолютной точностью 0.1. Таким образом, относительные точности обоих чисел примерно равны 0.01%.

Выполняя вычитание, получим результат 1, который имеет абсолютную погрешность $0.1 + 0.1 = 0.2$. Как следствие, относительная погрешность результата достигла 20%, т. е. увеличилась в 2000 (две тысячи) раз. ■

Отмеченное явление резкого увеличения относительной погрешности при вычитании называют *эффектом потери точности* или *эффектом потери значащих цифр*, поскольку его следствием является уменьшение количества верных значащих цифр в представлении результата.⁴ Как следствие, 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 вычислялись непосредственно, без вычисления самих этих величин.

Следующие два результата являются аналогом Предложения 1.2.1 для относительных погрешностей операндов.

Предложение 1.2.4 *Если погрешности приближённых чисел малы, то относительная погрешность их произведения приближённо равна (с точностью до членов более высокого порядка малости) сумме относительных погрешностей сомножителей.*

⁴Соответствующий английский термин — loss of significance.

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 $z^* := x_1^* x_2^* \dots x_n^*$ и $\tilde{z} := \tilde{x}_1 \tilde{x}_2 \dots \tilde{x}_n$.

Рассмотрим функцию

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

— произведение чисел x_1, x_2, \dots, x_n . Разлагая её в точке $(x_1^*, x_2^*, \dots, x_n^*)$ по формуле Тейлора с точностью до членов первого порядка, получим

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

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

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

что и требовалось. ■

Предложение 1.2.5 Если погрешности приближённых чисел малы, то относительная погрешность их частного приближённо (с точностью до членов более высокого порядка малости) равна сумме относительных погрешностей сомножителей.

Доказательство. Пусть x^*, y^* — точные значения рассматриваемых чисел, \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 первого 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

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

так что

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

Это и требовалось показать. ■

Мы оценили 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 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, механическое распространение полученных результатов на общий случай даёт довольно грубые оценки.

1.3 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, на которых в настоящее время выполняется подавляющая часть вычислений (и которые, как правило, обозначаются аббревиатурой ЭВМ), эти

детали реализации регламентируются специальными международными стандартами. Первый из них был принят в 1985 году Институтом инженеров по электротехнике и электронике⁵, 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 [27, 38], 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 всем этим стандартам, 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 β и p . *Числами с плавающей точкой* по основанию β 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 $0 \leq \alpha_i < \beta$, $i = 0, 1, \dots, p-1$. 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. Обычно число (1.7) обозначают условной записью

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

Отметим, что термин «мантисса» 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

⁵Чаще всего его называют английской аббревиатурой 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 (см. подробности в [27]).

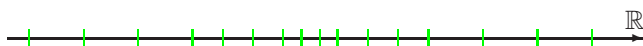


Рис. 1.1. Множество чисел, представимых в цифровой ЭВМ — дискретное конечное подмножество вещественной оси \mathbb{R} .

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 α_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 «+».⁶

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

⁶Таким образом, во внутреннем «машинном» представлении знак присутствует у любого числа, в том числе и у нуля.

мости 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 вполне удовлетворяет большинство научно-технических расчётов. В целом, числа с плавающей точкой обеспечивают практически фиксированную 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 $\pm\infty$ » или «к нулю».

Количество различных показателей степени t , равное 2046 для двойной точности, не исчерпывает всех возможных $2^{11} = 2048$ целых чисел, 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 ⁷, обладает свойствами математической бесконечности ∞ :

$$\begin{aligned} \text{Inf} \pm a &= \text{Inf}, & \text{Inf} + \text{Inf} &= \text{Inf}, \\ \text{Inf} \cdot a &= \text{Inf} \text{ для } a > 0, & \text{Inf} \cdot a &= -\text{Inf} \text{ для } a < 0. \end{aligned}$$

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

NaN означает невозможность придания результату операции какого-либо смысла 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,

$$\begin{aligned} \text{Inf} - \text{Inf} &= \text{NaN}, & \text{Inf} \cdot 0 &= \text{NaN}, \\ 0/0 &= \text{NaN}, & \text{Inf}/\text{Inf} &= \text{NaN}. \end{aligned}$$

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 ε » (машинное эpsilon), которое характеризует густоту (плотность) множества машинно-представимых чисел. Более точно, оно показывает расстояние между соседними машинно-представимыми числами. Это наименьшее такое положительное число $\varepsilon_{\text{маш}}$, что в компьютерной арифметике $1 + \varepsilon_{\text{маш}} \neq 1$ при округлении

⁷От латинского слова infinitum. Не следует путать его с точной нижней гранью множества, обозначаемой «inf».

«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 «машинное эпсилон» примерно равно $1.11 \cdot 10^{-16}$.

Удвоенное «машинное 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 β . Это вытекает из того отмеченного выше факта, что машинно-представимые числа расположены на вещественной оси равномерными участками.

Принципиальной особенностью компьютерной арифметики, вызванной дискретностью множества машинных чисел и наличием округлений, является невыполнение некоторых общеизвестных 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. Эта 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.

1.4 Интервальная арифметика

Исходной идеей создания интервальной арифметики является наблюдение о том, что всё в нашем мире неточно, и нам в реальности чаще всего приходится работать не с точными значениями величин, которые образуют основу классической «идеальной» математики, а с целыми диапазонами значений той или иной величины. Например, множество вещественных чисел, которые точно представляются в цифровых ЭВМ, конечно, и из-за присутствия округления каждое из этих чисел, в действительности, является представителем интервала значений обычной вещественной оси \mathbb{R} (см. Рис. 1.5–1.6).

Нельзя ли организовать операции и отношения между диапазонами-интервалами так, как это сделано для обычных точных значений? Чтобы можно было работать с ними, подобно обычным числам, 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 (см., к примеру, [25]).

Интервалом $[a, b]$ вещественной оси \mathbb{R} мы 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 мы

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.

Предположим, что нам даны переменные a и b , точные значения которых неизвестны, но мы знаем, что они могут находиться в интервалах $[\underline{a}, \bar{a}]$ и $[\underline{b}, \bar{b}]$. 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 $a + b \in [\underline{a} + \underline{b}, \bar{a} + \bar{b}]$.

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 $a - b \in [\underline{a} - \bar{b}, \bar{a} - \underline{b}]$.

Для умножения двух переменных $a \in [\underline{a}, \bar{a}]$ и $b \in [\underline{b}, \bar{b}]$ 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 $\phi(a, b) = a \cdot b$, будучи линейной по b при каждом фиксированном a , принимает минимальное и максимальное значения на концах интервала изменения переменной b . Это же верно и для экстремумов по $a \in [\underline{a}, \bar{a}]$ при любом фиксированном значении b . Наконец,

$$\begin{aligned}\min_{a \in [\underline{a}, \bar{a}], b \in [\underline{b}, \bar{b}]} \phi(a, b) &= \min_{a \in [\underline{a}, \bar{a}]} \min_{b \in [\underline{b}, \bar{b}]} \phi(a, b), \\ \max_{a \in [\underline{a}, \bar{a}], b \in [\underline{b}, \bar{b}]} \phi(a, b) &= \max_{a \in [\underline{a}, \bar{a}]} \max_{b \in [\underline{b}, \bar{b}]} \phi(a, b),\end{aligned}$$

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 $a \in [\underline{a}, \bar{a}]$ и $b \in [\underline{b}, \bar{b}]$ в самом деле

$$\min\{\underline{a}\underline{b}, \underline{a}\bar{b}, \bar{a}\underline{b}, \bar{a}\bar{b}\} \leq a \cdot b \leq \max\{\underline{a}\underline{b}, \underline{a}\bar{b}, \bar{a}\underline{b}, \bar{a}\bar{b}\}, \quad (1.8)$$

и нетрудно видеть, что эта оценка достижима с обеих сторон.

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 можно будет ввести свои собственные операции, отношения и т.п. Мы далее будем обозначать интервалы буквами жирного шрифта: $\mathbf{a}, \mathbf{b}, \mathbf{c}, \dots, \mathbf{x}, \mathbf{y}, \mathbf{z}$. Подчёркивание и надчёркивание — \underline{a} и \bar{a} — будут зарезервированы для обозначения нижнего и верхнего концов интервала, так что $\mathbf{a} = [\underline{a}, \bar{a}]$.

Рассмотрим множество всех вещественных интервалов $\mathbf{a} := [\underline{a}, \bar{a}] = \{a \in \mathbb{R} \mid \underline{a} \leq a \leq \bar{a}\}$, и бинарные операции — сложение, вычитание, умножение и деление — определим между ними «по представителям», т.е. в соответствии со следующим фундаментальным принципом:

$$\mathbf{a} \star \mathbf{b} := \{a \star b \mid a \in \mathbf{a}, b \in \mathbf{b}\} \quad (1.9)$$

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 отождествляются с интервалами нулевой ширины $[a, a]$, которые называются также *вырожденными интервалами*. Кроме того, через $(-\mathbf{a})$ условимся обозначать интервал $(-1) \cdot \mathbf{a}$.

Для интервальных арифметических операций развёрнутое определение, равносильное (1.9), как мы установили выше, задаётся следующими формулами:

$$\mathbf{a} + \mathbf{b} = [\underline{a} + \underline{b}, \bar{a} + \bar{b}], \quad (1.10)$$

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

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

$$\mathbf{a}/\mathbf{b} = \mathbf{a} \cdot [1/\bar{b}, 1/\underline{b}] \quad \text{для } \mathbf{b} \not\equiv 0. \quad (1.13)$$

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

$$\mu \cdot \mathbf{a} = \begin{cases} [\mu \underline{\mathbf{a}}, \mu \overline{\mathbf{a}}], & \text{если } \mu \geq 0, \\ [\mu \overline{\mathbf{a}}, \mu \underline{\mathbf{a}}], & \text{если } \mu \leq 0. \end{cases} \quad (1.14)$$

Алгебраическая система $\langle \mathbb{IR}, +, -, \cdot, / \rangle$, образованная множеством всех вещественных интервалов $\mathbf{a} := [\underline{\mathbf{a}}, \overline{\mathbf{a}}] = \{x \in \mathbb{R} \mid \underline{\mathbf{a}} \leq x \leq \overline{\mathbf{a}}\}$ с бинарными операциями сложения, вычитания, умножения и деления, которые определены формулами (1.10)–(1.13), называется *классической интервальной арифметикой*. Эпитет «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 операндов. Для этого выделим в \mathbb{IR} следующие подмножества:

$$\begin{aligned} \mathcal{P} &:= \{\mathbf{a} \in \mathbb{IR} \mid \underline{\mathbf{a}} \geq 0 \text{ и } \overline{\mathbf{a}} \geq 0\} && \text{— неотрицательные интервалы,} \\ \mathcal{Z} &:= \{\mathbf{a} \in \mathbb{IR} \mid \underline{\mathbf{a}} \leq 0 \leq \overline{\mathbf{a}}\} && \text{— нульсодержащие интервалы,} \\ -\mathcal{P} &:= \{\mathbf{a} \in \mathbb{IR} \mid -\mathbf{a} \in \mathcal{P}\} && \text{— неположительные интервалы.} \end{aligned}$$

В целом $\mathbb{IR} = \mathcal{P} \cup \mathcal{Z} \cup (-\mathcal{P})$. Тогда интервальное умножение (1.12) может быть описано с помощью Табл. 1.1, особенно удобной при реализации этой операции на ЭВМ.

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

$$(\mathbf{a} + \mathbf{b})\mathbf{c} \neq \mathbf{ac} + \mathbf{bc}.$$

Например,

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

Таблица 1.1. Интервальное умножение

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

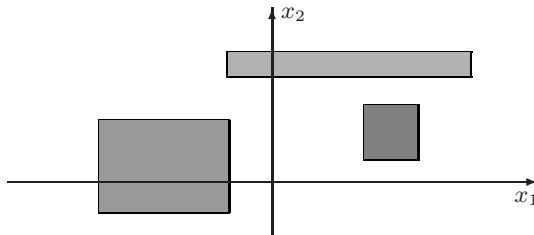
Тем не менее, имеет место более слабое свойство

$$a(b + c) \subseteq ab + ac \quad (1.15)$$

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

$$a(b + c) = ab + ac, \quad \text{если } a \text{ — вещественное число,} \quad (1.16)$$

$$a(b + c) = ab + ac, \quad \text{если } b, c \geq 0 \text{ или } b, c \leq 0. \quad (1.17)$$

Рис. 1.2. Интервальные векторы-брусы в \mathbb{R}^2 .

Интервальный вектор — это 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

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

а

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

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

Введём также 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 . Нетрудно 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 есть 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

$$\mathbf{A} \star \mathbf{B} = \square \{ \mathbf{A} \star \mathbf{B} \mid \mathbf{A} \in \mathbf{A}, \mathbf{B} \in \mathbf{B} \}, \quad \star \in \{ +, -, \cdot \}, \quad (1.18)$$

где \square — интервальная оболочка множества, наименьший по включению интервальный вектор-брус, который содержит его.

Интервальная арифметика, фактически, даёт точный и алгоритмизованный способ для оперирования с диапазонами значений и абсолютными погрешностями, который мы рассматривали в §1.2. Но формулы интервальной арифметики, будучи хорошо приспособленными для реализации на ЭВМ, всё-таки менее удобны для теоретического анализа, чем результаты Предложений 1.2.1–1.2.5.

1.5 Интервальные расширения функций

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

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

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 так называемого *интервального расширения функции*.

Определение 1.5.1 Пусть D — непустое подмножество пространства \mathbb{R}^n . Интервальная функция $\mathbf{f} : \mathbb{I}D \rightarrow \mathbb{I}\mathbb{R}^m$ называется интервальным продолжением точечной функции $f : D \rightarrow \mathbb{R}^m$, если $\mathbf{f}(x) = f(x)$ для всех $x \in D$.

Определение 1.5.2 Пусть D — непустое подмножество пространства \mathbb{R}^n . Интервальная функция $\mathbf{f} : \mathbb{I}D \rightarrow \mathbb{I}\mathbb{R}^m$ называется интервальным расширением точечной функции $f : D \rightarrow \mathbb{R}^m$, если

- 1) $\mathbf{f}(\mathbf{x})$ — интервальное продолжение $f(x)$,
- 2) $\mathbf{f}(\mathbf{x})$ монотонна по включению, т. е.
 $\mathbf{x}' \subseteq \mathbf{x}'' \Rightarrow \mathbf{f}(\mathbf{x}') \subseteq \mathbf{f}(\mathbf{x}'')$ на $\mathbb{I}D$.

Таким образом, если $\mathbf{f}(\mathbf{x})$ — интервальное расширение функции $f(x)$, то для области значений f на бресе $\mathbf{X} \subset D$ мы получаем следующую внешнюю (с помощью объемлющего множества) оценку:

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

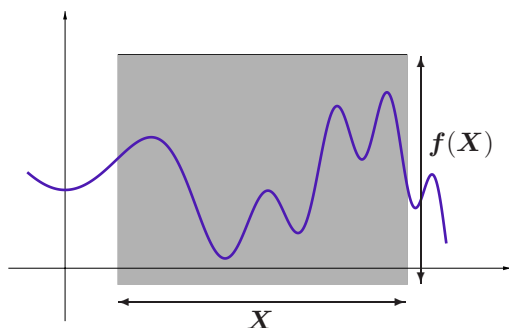


Рис. 1.3. Интервальное расширение функции даёт внешнюю оценку её области значений

Эффективное построение интервальных расширений функций — это важнейшая задача интервального анализа, поиски различных решений которой продолжаются и в настоящее время. Уместно привести в рамках нашего беглого обзора некоторые общезначимые результаты в этом направлении. Первый из них часто называют «основной теоремой интервальной арифметики»:

Теорема 1.5.1 Если для рациональной функции $f(x) = f(x_1, x_2, \dots, x_n)$ на брус $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ определён результат $\mathbf{f}_\pm(\mathbf{x})$ 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

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

т. е. $\mathbf{f}(\mathbf{x})$ содержит множество значений функции $f(x)$ на \mathbf{x} .

Нетрудно понять, что по отношению к рациональной функции $f(x)$ интервальная функция $\mathbf{f}_\pm(\mathbf{x})$, о которой идёт речь в Теореме 1.5.1, является интервальным расширением. Оно называется *естественным интервальным расширением* и вычисляется совершенно элементарно.

Пример 1.5.1 Для функции $f(x) = x/(x+1)$ на интервале $[1, 3]$ область значений в соответствии с результатом Теоремы 1.5.1 можно оценить извне как

$$\frac{[1, 3]}{[1, 3] + 1} = \frac{[1, 3]}{[2, 4]} = \left[\frac{1}{4}, \frac{3}{2}\right]. \quad (1.19)$$

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

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

Он более узок (т.е. более точен), чем (1.19), 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*:

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

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} .

В выписанном выше выражении $\mathbf{g}_i(\mathbf{x}, \tilde{x})$ 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, а интервальная функция \mathbf{f}_c называется дифференциальной центрированной формой интервального расширения.⁸

⁸По отношению к ней часто используют также термин «среднезначная форма», поскольку она может быть выведена из известной теоремы Лагранжа о среднем.

Пример 1.5.2 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

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

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

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

Как видим, этот результат значительно точнее естественного интервального расширения (1.19). ■

За дальнейшей информацией мы отсылаем заинтересованного читателя к книгам [1, 25, 29, 30], развёрнуто излагающим построение интервальных расширений функций. Важно отметить, что точность 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

$$\text{dist} (\mathbf{f}_\natural(\mathbf{x}), f(\mathbf{x})) \leq C \|\text{wid } \mathbf{x}\| \quad (1.20)$$

с некоторой константой C . Этот факт обычно выражают словами «естественное интервальное расширение имеет первый порядок точности» (см. Определение 2.8.1, стр. 137).

Для центрированной формы верно соотношение

$$\text{dist} (\mathbf{f}_c(\mathbf{x}, \tilde{x}), f(\mathbf{x})) \leq 2 (\text{wid } \mathbf{g}(\mathbf{x}, \tilde{x}))^\top | \mathbf{x} - \tilde{x} |, \quad (1.21)$$

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 случае, когда интервальные оценки для функций $\mathbf{g}_i(\mathbf{x}, \tilde{x})$ находятся с первым порядком точности, общий порядок точности центрированной формы согласно

(1.21) будет уже вторым. Вывод этих оценок заинтересованный читатель может найти, к примеру, в [25, 30].

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 очень элегантно записываются остаточные члены различных приближённых формул. В качестве двух содержательных примеров применения интервальных расширений функций мы рассмотрим решение уравнений и оценку константы Липшица для функций.

1.6 Обусловленность математических задач

Вынесенный в заголовок этого параграфа термин — *обусловленность* — означает меру чувствительности решения задачи к изменениям (возмущениям) её входных данных. Ясно, что любая информация подобного сорта 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 мы можем предпринимать необходимые меры для компенсации этого явления — повышать разрядность вычислений, наконец, модифицировать или вообще сменить выбранный вычислительный алгоритм и т. п.

Существует 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 §2.8г в качестве примера таких задач мы рассмотрим задачу численного дифференцирования. Во-вторых, в случае наличия этой непрерывности желательно иметь некоторую количественную меру чувствительности решения в зависимости от входных данных.

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} , мы будем называть *разрешающим отображением* (или *разрешающим оператором*). Отображение ϕ может быть выписано явным образом, если ответ к задаче задаётся каким-либо выражением. Часто

разрешающее отображение задаётся 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 выписать 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 ϕ было непрерывным относительно некоторого задания норм в \mathcal{D} и \mathcal{S} . 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 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 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 a часто привлекают отдельные частные производные $\frac{\partial \phi_i}{\partial a_j}$, т. е. элементы матрицы Якоби разрешающего отображения ϕ , которые при этом 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

$$\|\phi(a + \Delta a) - \phi(a)\| \leq C_1 \|\Delta a\| \quad (1.22)$$

и

$$\frac{\|\phi(a + \Delta a) - \phi(a)\|}{\|\phi(a)\|} \leq C_2 \frac{\|\Delta a\|}{\|a\|}. \quad (1.23)$$

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

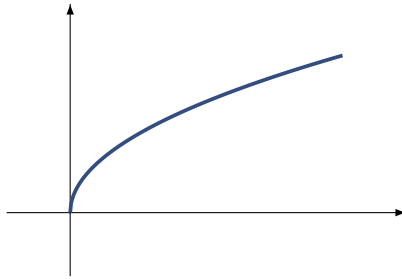


Рис. 1.4. Непрерывная функция $y = \sqrt{x}$ имеет бесконечную скорость роста при $x = 0$ и не является липшицевой

В связи с неравенствами (1.22)–(1.23) напомним, что вещественная функция $f : \mathbb{R}^n \supseteq D \rightarrow \mathbb{R}$ называется *непрерывной по Липшицу* (или просто *липшицевой*), если существует такая константа L , что

$$|f(x') - f(x'')| \leq L \cdot \text{dist}(x', x'') \quad (1.24)$$

для любых $x', x'' \in D$. Величину L называют при этом *константой Липшица* функции f на D . Понятие непрерывности по Липшицу формализует интуитивно понятное условие соразмерности изменения функции изменению аргумента. Именно, 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 обоих.

Нетрудно видеть, что искомые константы C_1 и C_2 в неравенствах (1.22) и (1.23), характеризующие чувствительность решения задачи по отношению к возмущениям входных данных — это не что иное, как константы Липшица для разрешающего отображения ϕ и произведение константы Липшица L_ψ отображения $\psi : \mathcal{D} \rightarrow \mathcal{S}$, действующего по правилу $a \mapsto \phi(a)/\|\phi(a)\|$ на норму $\|a\|$. В последнем случае

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

1.7 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 выполняемых с ним расчётов.

Если алгоритм устойчив, то он не позволяет неизбежным погрешностям входных данных задачи и погрешностям вычислений заметно исказить результат. С неустойчивым алгоритмом такого может не быть, и результат его работы подчас имеет мало общего с ответом к решаемой задаче. Конкретной мерой устойчивости алгоритма может быть, к примеру, чувствительность получаемых с его помощью результатов в зависимости от возмущений входных данных и погрешностей промежуточных расчётов. Устойчивый алгоритм характеризуется низкой чувствительностью к возмущениям и погрешностям во входных данных и промежуточных результатах, тогда как у неустойчивого алгоритма эта чувствительность высока.

Если обусловленность задачи характеризует объективный факт, касающийся самой постановки задачи, то устойчивость алгоритма — это свойство сконструированной и применяемой нами процедуры для решения этой задачи. Как следствие, мы должны стремиться сделать эту устойчивость наилучшей.

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 тех или иных явлений (другие модели часто нам просто недоступны). В этом случае говорят, что устойчивый алгоритм *регуляризует* исходную плохо обусловленную задачу.

Построение устойчивых алгоритмов для решения плохообусловленных и некорректных задач является предметом большой и важной научной дисциплины, основы которой были заложены А.Н. Тихоновым в середине XX века (см. [37]). За прошедшие десятилетия она получила чрезвычайно большое развитие и многочисленные практические применения, расширившие сферу применимости вычислительной математики.

Пример 1.7.1 (пример Бабушки-Витасека-Прагера [31, 36])

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

$$\begin{aligned} I_n &= \int_0^1 x^n d(e^{x-1}) = x^n e^{x-1} \Big|_0^1 - n \int_0^1 x^{n-1} e^{x-1} dx \\ &= 1 - n I_{n-1}. \end{aligned} \tag{1.25}$$

Кроме того,

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

Теперь уже нетрудно организовать вычисления, но ... поведение их результатов для растущих n делается странным. Таблица ниже приво-

дит 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 (1.25) 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. [31, 36]), 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 погреш-

ность I_n , вычисляемая по формуле (1.25), всё сильнее искажает результат из-за вычитания близких чисел (см. §1.2).

Итак, алгоритм вычисления I_n с помощью рекуррентной формулы (1.25) даёт пример неустойчивого алгоритма, в котором ошибки промежуточных вычислений не подавляются, а разрастаются неконтролируемым образом. ■

Надёжность результатов иногда можно проконтролировать с помощью интервальных вычислений, и для примера Бабушки-Витасека-Прагера это сделано в [36]. Большое увеличение ширины внешней интервальной оценки результата свидетельствует о неустойчивости вычислений.

1.8 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 развиваемым им «интуиционизмом».

В частности, теория алгоритмов и рекурсивных функций — это 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. Далее, *конструктивным объектом* называется объект, который может быть построен с помощью некоторой конечной последователь-

ности действий над каким-то конечным алфавитом. Таковы, например, рациональные числа. Строго говоря, конструктивные объекты и только они могут быть получены в качестве ответов при решении задачи на реальных цифровых ЭВМ с конечными быстродействием и объёмом памяти.

В частности, конечными машинами являются широко распространенные ныне электронные цифровые вычислительные машины: они способны представлять, по сути дела, только конечные множества чисел. Таким образом, обречены на неудачу любые попытки использовать их для выполнения арифметических абсолютно точных операций над числовыми полями \mathbb{R} и \mathbb{C} , которые являются бесконечными (и даже непрерывными) множествами, большинство элементов которых 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 (см. [34]), и он указывает на принципиальные границы возможностей алгоритмического подхода и ЭВМ в деле решения задач математического анализа.

Тем не менее, и в этом океане неконструктивности имеет смысл выделить объекты, которые могут быть «достаточно хорошо» приближены конструктивными объектами. На этом пути мы приходим к понятию *вычислимого вещественного числа* [23, 34]⁹: вещественное число α называется вычислимым, если существует алгоритм, дающий по всякому натуральному числу n рациональное приближение к α с погрешностью $\frac{1}{n}$. Множество всех вычислимых вещественных чисел образует *вычислимый континуум*. Соответственно, *вычислимая вещественная функция* определяется как отображение из вычислимого континуума в вычислимый континуум, задаваемая алгоритмом преобразования программы аргумента в программу значений.

Важно помнить, что и вычисляемое 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 трудно-

⁹Совершенно аналогичным является определение *конструктивного вещественного числа* у Б.А. Кушнера [33].

сти, связанные с задачей решения уравнений. Для вычислимых вещественных чисел и функций ряд традиционных постановок задач оказывается *алгоритмически неразрешимыми* в том смысле, что построение общих алгоритмов их решения принципиально невозможно.

Например, алгоритмически неразрешимыми являются задачи

- 1) распознавания равно нулю или нет произвольное вычислимое вещественного число [33, 34, 35], распознавания равенства двух вычислимых вещественных чисел [23, 26, 33, 34];
- 2) нахождения для каждой совместной системы линейных уравнений над полем конструктивных вещественных чисел какого-либо ее решения [33, 35];
- 3) нахождения нулей всякой непрерывной кусочно-линейной знакопеременной функции [35].

Приведённые выше результаты задают, как нам представляется, ту абсолютную и совершенно объективную мерку (в отличие от субъективных пристрастий), с которой мы должны подходить к оценке трудоёмкости тех или иных вычислительных методов. В Главе 4, к примеру, проводится 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! На фоне этого мрачного факта наличие даже экспоненциально трудного алгоритма с небольшим основанием «одноэтажной» экспоненты в оценке сложности (вроде 2^n) можно рассматривать как вполне приемлемый вариант разрешимости задачи. Именно это имеет место в ситуации с вычислением вращения векторного поля (степени отображения), которое требуется в новой формулировке задачи решения уравнений и систем уравнений.

Вычислительная математика тесно примыкает к конструктивной, хотя их цели и методы существенно разнятся.

1.9 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 тех или иных алгоритмов для их решения.

Например, 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 корней 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 для практики бесполезен.

Естественно измерять трудоёмкость алгоритма количеством «элементарных операций», требуемых для его исполнения. Следует лишь иметь в виду, что эти операции могут быть весьма различными. Скажем, сложение и умножение двух чисел «с плавающей точкой» требуют для своего выполнения разного количества тактов современных процессоров и, соответственно, разного времени. До определённой степени эти различия можно игнорировать и оперировать понятием усреднённой арифметической операции.

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.¹⁰

На качественном уровне полезно различать *полиномиальную трудоёмкость* и *экспоненциальную трудоёмкость*. Говорят, что алгоритм имеет полиномиальную трудоёмкость, если сложность его выполнения не превосходит значений некоторого полинома от длины входных данных. Напротив, про некоторый алгоритм говорят, что он имеет экспоненциальную трудоёмкость, если сложность его выполнения превосходит значения любого полинома от длины подаваемых ему на вход данных.

Получение оценок трудоёмкости задач является непростым делом. Если какой-то алгоритм решает поставленную задачу, то, очевидно,

¹⁰В связи с этим получили распространение также относительные единицы измерения трудоёмкости алгоритмов — через количество вычислений функции, правой части уравнения и т. п.

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 очень небольшого круга задач, которые имеют, скорее, теоретическое значение. В этих условиях широкое распространение получила альтернативная теория сложности, в основе которой лежат понятие сводимости задач друг к другу и вытекающее из него понятие эквивалентности задач по трудоёмкости.

Наибольшее распространение получило *полиномиальное сведение* одних задач к другим, под которым понимается такое преобразование одной задачи к другой, что данные и ответ исходной задачи переводятся в данные и ответ для другой, а трудоёмкость этого преобразования не превышает значений некоторого полинома от размера исходной задачи. Взаимная полиномиальная сводимость двух задач друг другу является отношением эквивалентности.

Этим рассуждениям можно придать более строгую форму, что приводит к так называемой теории NP-трудности, получившей существенное развитие в последние десятилетия. Её основным понятием является понятие *NP-трудной задачи* [10], и неформально подобные задачи можно охарактеризовать как “универсальные переборные задачи”.

Таким образом, теория NP-полноты не отвечает напрямую на вопрос о трудоёмкости решения тех или иных задач, но позволяет утверждать, что некоторые задачи «столь же трудны», как и другие известные задачи. Нередко знание уже этого одного 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 вариантов, 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. Но вот оценивание разброса решений систем линейных или нелинейных уравнений при варьировании параметров этих систем в самом общем случае, когда мы не ограничиваем себя величиной возмущений, является NP-трудной задачей. В частности, таковы интервальные системы уравнений.

1.10 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 [3], где доказательным вычислениям посвящён отдельный параграф, можно прочесть: «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.¹¹ К примеру, если мы находим число π , то доказательным ответом может быть установление гарантированных неравенств $\pi > 3.1415926$ или $\pi \geq 3.1415926$.

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. Например, в моделях двоичной компьютерной арифметики с плавающей точкой форматов «single» и «double» (см. §1.3) не представимы точно такие обыденные десятичные дроби как 0.1 и 0.2.

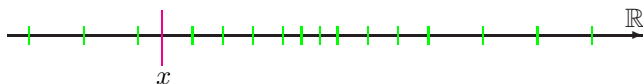


Рис. 1.5. Типичное вещественное число не представимо точно в цифровой ЭВМ с конечной разрядной сеткой

Ситуация, в действительности, ещё более серьёзна, так как неиз-

¹¹Термин «доказательные вычисления на ЭВМ» является хорошим русским эквивалентом таких распространённых английских оборотов как *verified computation*, *verification numerics* и др. Ранее для этой же цели применялся термин *validated numerics*, но он по ряду причин неудачен и сейчас не используется.

бежными погрешностями, как правило, сопровождаются ввод данных в ЭВМ и выполнение с ними любых арифметических операций. Хотя эти погрешности могут быть очень малы, но, накапливаясь, они способны существенно исказить ответ к решаемой задаче. Встаёт нетривиальная проблема их учёта в процессе счёта на ЭВМ...

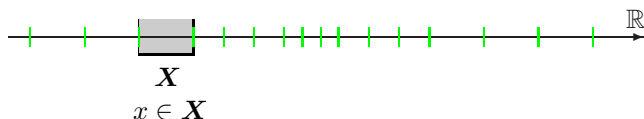


Рис. 1.6. Интервальное решение проблемы представления вещественных чисел в цифровой ЭВМ

Одним из средств доказательных вычислений на ЭВМ служит интервальная арифметика и, более общо, 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 . Далее с получающимися интервалами можно выполнять операции по правилам интервальной арифметики, рассмотренным в §1.4.

Концы интервалов, получающихся при расчётах по формулам (1.10)–(1.13), также могут оказаться вещественными числами, не представимыми в ЭВМ. В этом случае для обеспечения доказательности вычислений имеет смысл несколько расширить полученный интервал до ближайшего объёмлющего его интервала с машинно-представимыми концами. Подобная версия интервальной арифметики называется *машинной интервальной арифметикой* с направленным округлением (см. Рис. 1.7).

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

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 и логические операции, так и целые их последовательности, слагающиеся в крупные блоки алгоритма. При этом полная погрешность получается из погрешностей отдельных «элементарных шагов» по правилам исчисления из §1.2.

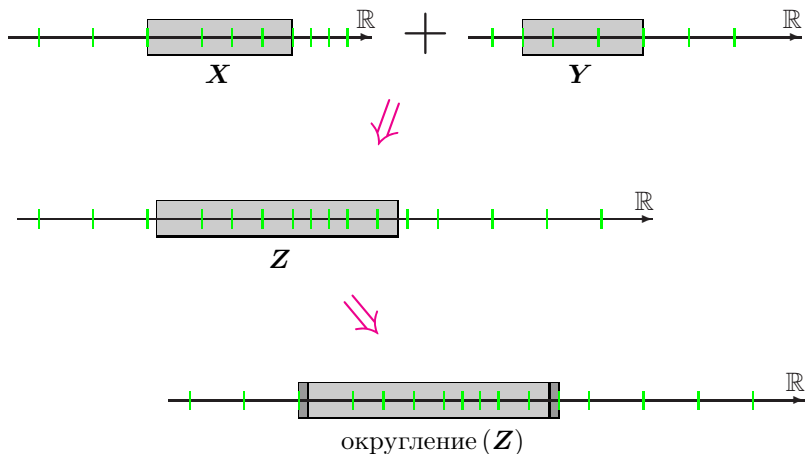


Рис. 1.7. Машинная интервальная арифметика с внешним направленным округлением

Очевидный недостаток такого способа организации оценки погрешностей состоит в том, что мы неявно привязываемся к конкретному алгоритму вычисления решения. При этом качество оценок, получаемых с помощью пошаговой парадигмы, существенно зависит от алгоритма, и «хороший» в обычном смысле алгоритм не обязательно хорош при анализе и оценивании его погрешностей.

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 [32].

Напротив, при апостериорном оценивании погрешности мы оцениваем погрешность окончательного результата уже *после* его получения. Иными словами, мы разделяем получение двусторонней оценки решения и установление её доказательности. Этот подход оказался существенно более практичным и плодотворным, чем пошаговый, и ниже в Главе 4 мы приведём примеры конкретных алгоритмов апостериор-

ного оценивания для доказательного решения некоторых популярных математических задач.

Литература к главе 1

Основная

- [1] АЛЕФЕЛЬД Г., ХЕРЦБЕРГЕР Ю. *Введение в интервальные вычисления*. – Москва: Мир, 1987.
- [2] БАРАХНИН В.Б., ШАПЕЕВ В.П. *Введение в численный анализ*. – Санкт-Петербург–Москва–Краснодар: Лань, 2005.
- [3] БАБЕНКО К.И. *Основы численного анализа*. – Москва: Наука, 1986.
- [4] БАУЭР Ф.Л., ГООЗ Г. *Информатика. В 2-х ч.* – Москва: Мир, 1990.
- [5] БАХВАЛОВ Н.С., ЖИДКОВ Н.П., КОВЕЛЬКОВ Г.М. *Численные методы*. – Москва: Бином, 2003, а также другие издания этой книги.
- [6] БАХВАЛОВ Н.С., КОРНЕВ А.А., ЧИЖОНКОВ Е.В. *Численные методы. Решения задач и упражнения*. – Москва: Дрофа, 2008.
- [7] БЕРЕЗИН И.С., ЖИДКОВ Н.П. *Методы вычислений. Т. 1–2*. – Москва: Наука, 1966.
- [8] ВЕРЖВИЦКИЙ В.М. *Численные методы. Части 1–2*. – Москва: «Оникс 21 век», 2005.
- [9] ВОЛКОВ Е.А. *Численные методы*. – Москва: Наука, 1987.
- [10] ГЭРИ М., ДЖОНСОН Д. *Вычислительные машины и труднорешаемые задачи*. – Москва: Мир, 1982.
- [11] ДЕМИДОВИЧ Б.П., МАРОН А.А. *Основы вычислительной математики*. – Москва: Наука, 1970.
- [12] КАЛИТКИН Н.Н. *Численные методы*. – Москва: Наука, 1978.
- [13] КРЫЛОВ А.Н. *Лекции о приближённых вычислениях*. – Москва: ГИТТЛ, 1954, а также более ранние издания.
- [14] КРЫЛОВ В.И., БОВКОВ В.В., МОНАСТЫРНЫЙ П.И. *Вычислительные методы. Т. 1–2*. – Москва: Наука, 1976.
- [15] КУНЦ К.С. *Численный анализ*. – Киев: Техника, 1964.
- [16] КУНЦМАН Ж. *Численные методы*. – Москва: Наука, 1979.
- [17] МАЦОКИН А.М., СОРОКИН С.Б. *Численные методы. Часть 1. Численный анализ*. – Новосибирск: НГУ, 2006.
- [18] МЕНЬШИКОВ Г.Г. *Локализуемые вычисления. Конспект лекций*. – Санкт-Петербург: СПбГУ, Факультет прикладной математики–процессов управления, 2003.
- [19] МИНЬКОВ С.Л., МИНЬКОВ Л.Л. *Основы численных методов*. – Томск: Издательство научно-технической литературы, 2005.

- [20] РАЙС Дж. *Матричные вычисления и математическое обеспечение*. – Москва: Мир, 1984.
- [21] САМАРСКИЙ А.А., ГУЛИН А.В. *Численные методы*. – Москва: Наука, 1989.
- [22] ТЫРТЫШНИКОВ Е.Е. *Методы численного анализа*. – Москва: Академия, 2007.
- [23] УСПЕНСКИЙ В.А., СЕМЁНОВ А.Л. *Теория алгоритмов: основные открытия и приложения*. – Москва: Наука, 1987.
- [24] ХАНСЕН Э., УОЛСТЕР ДЖ.У. *Глобальная оптимизация с помощью методов интервального анализа*. – Москва-Ижевск: Издательство «РХД», 2012.
- [25] ШАРЫЙ С.П. *Конечномерный интервальный анализ*. – ФИЦ ИВТ: Новосибирск, 2020 – Электронная книга, доступная на <http://www.nsc.ru/interval/Library/InteBooks/>)
- [26] АВЕРТН О. *Precise numerical methods using C++*. – San Diego: Academic Press, 1998.
- [27] GOLDBERG D. What every computer scientist should know about floating point arithmetic // *ACM Computing Surveys*. – 1991. – Vol. 23, No. 1. – P. 5–48.
- [28] KREINOVICH V., LAKEYEV A.V., ROHN J., KAHL P. *Computational complexity and feasibility of data processing and interval computations*. – Dordrecht: Kluwer, 1997.
- [29] MOORE R.E., KEARFOTT R.B., CLOUD M. *Introduction to interval analysis*. – Philadelphia: SIAM, 2009.
- [30] NEUMAIER A. *Interval methods for systems of equations*. – Cambridge: Cambridge University Press, 1990.

Дополнительная

- [31] БАБУШКА И., ВИТАСЕК Э., ПРАГЕР М. *Численные процессы решения дифференциальных уравнений*. – Москва: Мир, 1969.
- [32] 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.
- [33] КУШНЕР Б.А. *Лекции по конструктивному математическому анализу*. – Москва: Наука, 1973.
- [34] МАРТИН-ЛЁФ П. *Очерки по конструктивной математике*. – Москва: Наука, 1975.
- [35] *Математический Энциклопедический Словарь*. – Москва: Большая Российская Энциклопедия, 1995.
- [36] 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.
- [37] ТИХОНОВ А.Н., АРСЕНИН В.Я. *Методы решения некорректных задач*. – Москва: Наука, 1979.

- [38] *IEEE Std 754-1985. IEEE Standard for Binary Floating-Point Arithmetic.* – New York: IEEE, 1985.

Глава 2

Численные методы анализа

2.1 Введение

Под численными методами анализа обычно понимаются вычислительные методы решения ряда задач, возникающих в классическом математическом анализе. Традиционно сюда относят задачи интерполирования и приближения функций, задачи численного нахождения производных и интегралов, задачу суммирования рядов, а также вычислительные методы гармонического анализа. Кроме того, численные методы анализа охватывают задачу вычисления значений различных функций. Она относительно проста для функций, явно задаваемых несложными арифметическими выражениями, но становится нетривиальной в случае, когда функция задаётся неявно или с помощью операций, выходящих за пределы конечного набора элементарных арифметических действий.

В нашем курсе мы рассмотрим первые четыре из перечисленных выше задач, и сначала займёмся задачами интерполирования и приближения функций.

Задачи интерполирования¹ и приближения функций являются тесно связанными друг с другом задачами, которые укладываются в рамки следующей единой неформальной схемы. Пусть дана функция $f(x)$,

¹Наряду с термином «интерполирование» в равной мере используется его синоним «интерполяция».

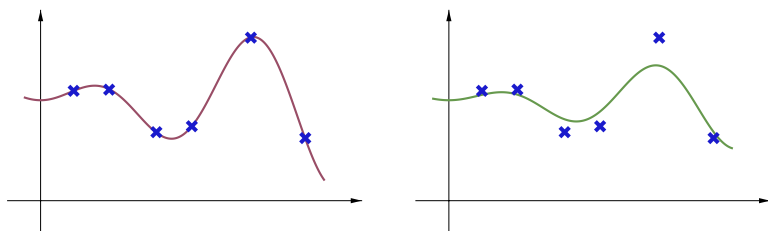


Рис. 2.1. Различие задач интерполяции и приближения функций.

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 полезно наделять рассматриваемые классы функций дополнительной структурой, например, считать, что они являются линейными векторными пространствами с нормой и т. п. Наконец, часто имеет место включение $\mathcal{G} \subset \mathcal{F}$.

Задача интерполирования 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 пересечения областей определения. От функции f при этом требуются лишь значения на этом множестве точек, и потому при постановке задачи интерполяции она сама часто 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.

Задача приближения функций (называемая также задачей *аппроксимации функций*) является частным случаем выписанной выше общей формулировки, в котором «близость» и «отклонение» функций f и g понимается как малое различие их значений друг от друга. При этом мы рассматриваем и сравниваем значения функций на некотором множестве X , которое является подмножеством их области определения. Оно может совпадать со всей этой областью определения, но может быть его небольшой частью, скажем, конечным набором точек. В последнем случае получается *дискретная* задача приближения, близкая

к задаче интерполирования. Если множество X совпадает со всей областью определения функций, то удобно рассматривать их «близость» или «отклонение» в терминах какого-то абстрактного расстояния (метрики), заданного на пересечении классов функций \mathcal{F} и \mathcal{G} (см. также §2.10а). 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.

Напомним, что на множестве Y , образованном элементами произвольной природы, *расстоянием* (называемым также *метрикой*) называется определённая на декартовом произведении $Y \times Y$ функция dist с неотрицательными вещественными значениями, удовлетворяющая для любых $f, g, h \in Y$ следующим условиям [12]:

- (1) $\text{dist}(f, g) = 0$ тогда и только тогда, когда $f = g$,
- (2) $\text{dist}(f, g) = \text{dist}(g, f)$ — симметричность,
- (3) $\text{dist}(f, h) \leq \text{dist}(f, g) + \text{dist}(g, h)$ — неравенство треугольника.

Разнообразные способы определения расстояния между функциями, возникающие в практике математического моделирования, приводят к различным математическим задачам приближения. Например, для функций $f, g : \mathbb{R} \supseteq [a, b] \rightarrow \mathbb{R}$ популярны равномерное (чебышёвское) расстояние, которое определяется как

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

или интегральное расстояние, определяемое как

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

В §2.10ж мы рассмотрим также задачу среднеквадратичного приближения функций, 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 вполне 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 задачах дискретного приближения функций «отклонение» или «близость» адекватно описывается понятием *псевдорасстояния* (псевдометрики), которое определяется почти так же, как обычное расстояние, но отличается от него ослаблением первой аксиомы: хотя всегда имеет место $\text{dist}(f, f) = 0$, но из $\text{dist}(f, g) = 0$ не обязательно следует, что $f = g$.² Тогда псевдорасстояние между двумя функциями, совпадающими на заданном наборе значений аргумента, будет равно нулю, даже если эти функции не равны друг другу, т.е. различаются при каких-то других аргументах.

2.2 Интерполирование функций

2.2а Постановка задачи и её свойства

Задача интерполирования — это задача восстановления (доопределения) функции, которая задана на дискретном множестве точек x_i , $i = 0, 1, \dots, n$. Для вещественных функций одного аргумента её постановка такова.

Задан интервал $[a, b] \subset \mathbb{R}$ и конечное множество несовпадающих точек $x_i \in [a, b]$, $i = 0, 1, \dots, n$, называемых *узлами интерполяции*. Совокупность всех узлов — множество $\{x_0, x_1, \dots, x_n\}$ — будем называть *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 при этом *интерполирующей функцией* или *интерполянт*ом.

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* .

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

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

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 анализов требует существенного времени, так что множество результатов этих анализов 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 предоставляют отдельные метеостанции.

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 (см. [39]). 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, найти значение интересующей 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?³

Здесь на помощь приходит интерполяция — восстановление значения функции в промежуточных точках по ряду известных значений в некоторых фиксированных опорных точках. Собственно, сам тер-

³Именно таковы, к примеру, популярные «Четырёхзначные математические таблицы» В.М. Брадиса [4] для средней школы.

мин «интерполирование» («интерполяция») был впервые употреблён в 1656 году Дж. Валлисом при составлении астрономических и математических таблиц. Он происходит от латинского слова «interpolo», означающего «переделывать», «подновлять», «ремонттировать».

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 затабулированы все основные функции, встречающиеся в математической и инженерной практике (см., к примеру, [39] и им 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 «на голом месте» 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, очень 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 затабулированных значений этих функций для фиксированных аргументов. Опираясь на них, строится значение в нужной нам точке.

Ещё один источник возникновения задачи интерполирования — 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{S} интерполирующих функций достаточно широк, то решение задачи интерполяции может быть неединственным (см. Рис. 2.2). Напротив, если \mathcal{S} узок, то у задачи интерполяции может вовсе не быть решений. На практике выбор класса \mathcal{S} обычно диктуется спецификой решаемой практической задачи.

В случае, когда, к примеру, заранее известно, что интерполируемая функция периодична, в качестве интерполантов естественно взять то-

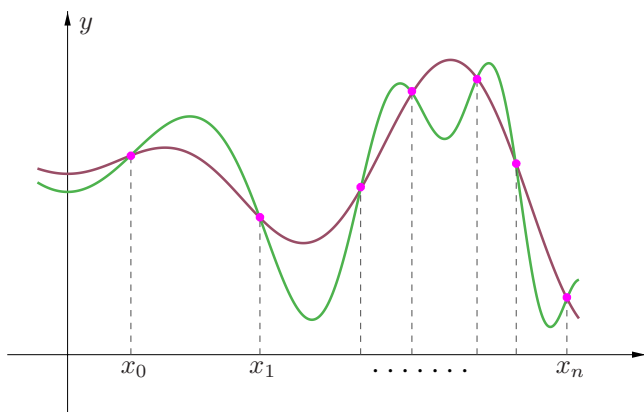


Рис. 2.2. Задача интерполяции может иметь неединственное решение.

же периодические функции, с тем же периодом. Ими могут быть, в частности, тригонометрические полиномы

$$a_0 + \sum_{k=1}^m (a_k \cos kx + b_k \sin kx) \quad (2.4)$$

для некоторого фиксированного 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

$$a_0 + a_1x + a_2x^2 + \dots + a_mx^m. \quad (2.5)$$

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 интерполирования. Вообще, проблема наиболее адекватного выбора класса интерполирующих функций \mathcal{G} 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

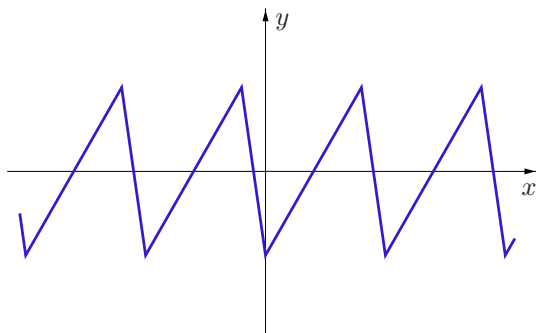


Рис. 2.3. Функция, которую лучше интерполировать с помощью периодических функций.

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

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

Как по интерполяционным данным (x_i, y_i) , $i = 0, 1, \dots, n$, найти интерполяционный полином вида (2.5), т.е. определить его коэффициенты a_0, a_1, \dots, a_m ?

Подставляя в выражение (2.5) значения аргумента x_0, x_1, \dots, x_n и учитывая, что получающиеся при этом значения полинома должны быть равны y_0, y_1, \dots, y_n соответственно, приходим к соотношениям

$$\begin{aligned}
 a_0 + a_1x_0 + a_2x_0^2 + \dots + a_mx_0^m &= y_0, \\
 a_0 + a_1x_1 + a_2x_1^2 + \dots + a_mx_1^m &= y_1, \\
 \vdots \quad \quad \quad \vdots \quad \quad \quad \ddots \quad \quad \quad \vdots \quad \quad \quad \vdots & \\
 a_0 + a_1x_n + a_2x_n^2 + \dots + a_mx_n^m &= y_n.
 \end{aligned} \tag{2.6}$$

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 (2.6) может не иметь решения, а если оно существует, то может быть неединственным. Имеется, тем не менее, важный частный случай задачи алгебраической интерполяции, для которого гарантируется однозначная разрешимость.

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

Доказательство. При $m = n$ в системе линейных алгебраических уравнений (2.6) число неизвестных совпадает с числом уравнений, а матрица этой системы — квадратная. Она имеет вид

$$V(x_0, x_1, \dots, x_n) = \begin{pmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{pmatrix} \quad (2.7)$$

и является так называемой матрицей Вандермонда (см., к примеру, [17, 35]). 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 (2.6) 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{S} 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, аналогично тому, как это получилось выше для алгебраических полиномов фиксированной степени.

Например, сказанное справедливо при тригонометрической интерполяции, т. е. с помощью функций, задаваемых выражениями (2.4), при интерполировании суммами экспонент вида

$$a_0 + a_1 e^{\beta_1 x} + \dots + a_m e^{\beta_m x}$$

с несовпадающими β_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 интерполанта тогда необходимо решать систему нелинейных уравнений.

2.26 Интерполяционный полином Лагранжа

Развитый в предшествующем разделе способ построения интерполанта 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. Далее, при значительном количестве узлов построение интерполанта посредством решения системы уравнений невыгодно в вычислительном отношении. Помимо того, что решение систем линейных уравнений само по себе не является тривиальной задачей, система (2.6) с матрицей Вандермонда оказывается весьма чувствительной к возмущениям данных или, как принято говорить, *плохо обусловленной* (см. §1.6; конкретные числовые оценки чувствительности решения системы (2.6) можно найти в §§3.5а–3.5б). Поэтому получаемый на этом пути интерполяционный полином может обладать большой погрешностью.

Систему линейных уравнений (2.6) 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 пройден, хотя и требует громоздких алгебраических преобразований.

На самом деле нам нечасто требуется знать для интерполяционного полинома именно каноническую форму (2.5). 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.⁴

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

$$P_n(x) = \sum_{i=0}^n y_i \phi_i(x), \quad (2.8)$$

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

$$\phi_i(x_j) = \delta_{ij} = \begin{cases} 0, & \text{при } i \neq j, \\ 1, & \text{при } i = j, \end{cases} \quad (2.9)$$

$i, j = 0, 1, \dots, n$, и посредством δ_{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 (2.8), действительно удовлетворяет условиям задачи.

Найдём теперь $\phi_i(x)$. Коль скоро этот полином зануляется в точках

⁴Сказанное можно выразить словами «оператор интерполирования линейен». В действительности, он даже является проектором, и эти наблюдения являются началом большого и плодотворного направления теории приближения функций.

$x_0, \dots, x_{i-1}, x_{i+1}, \dots, x_n$, то он имеет вид

$$\phi_i(x) = K_i (x - x_0) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n). \quad (2.10)$$

При этом K_i должен быть некоторым числовым множителем, так как в правой части равенства (2.10) произведение n линейных по x членов уже даёт полином степени n . Для определения этого множителя подставим в выражение (2.10) значение аргумента $x = x_i$, откуда в силу (2.9) 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

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

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 (2.9)). В целом, из (2.8) следует, что задачу алгебраической интерполяции решает полином

$$P_n(x) = \sum_{i=0}^n y_i \frac{\prod_{j \neq i} (x - x_j)}{\prod_{j \neq i} (x_i - x_j)}. \quad (2.11)$$

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

Далее нам потребуется его запись в несколько другом виде. Введём вспомогательную функцию

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

— полином $(n + 1)$ -й степени, зануляющийся во всех узлах интерполяции. Тогда

$$\phi_i(x) = \frac{\omega_n(x)}{\omega'_n(x_i)} = \frac{\omega_n(x)}{(x - x_i) \omega'_n(x_i)}, \quad (2.13)$$

и поэтому

$$P_n(x) = \sum_{i=0}^n y_i \frac{\omega_n(x)}{(x - x_i) \omega'_n(x_i)}. \quad (2.14)$$

Задача интерполяции полностью решается с помощью полиномов (2.11) и (2.14), которые находят широчайшее применение в вычислительной практике. Тем не менее, в ряде случаев и они оказываются 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.

2.2в Разделённые разности и их свойства

Определение 2.2.2 Пусть дана 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 $f^\angle(x_i, x_{i+1})$, называются отношения

$$f^\angle(x_i, x_{i+1}) := \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i}, \quad (2.15)$$

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

Разделённые разности второго порядка — это величины

$$f^\angle(x_i, x_{i+1}, x_{i+2}) := \frac{f^\angle(x_{i+1}, x_{i+2}) - f^\angle(x_i, x_{i+1})}{x_{i+2} - x_i}, \quad (2.16)$$

$i = 0, 1, \dots, n-2$, которые являются разделёнными разностями от разделённых разностей. Аналогичным образом вводятся разделённые раз-

ности высших порядков: *разделённая разность k -го порядка* от функции f есть, по определению,

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

$i = 0, 1, \dots, n - k$, 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. Ньютоном, 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 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

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

$i = 0, 1, \dots, n - k$. Это обозначение 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 [20, 28]).

Нетрудно увидеть геометрический смысл разделённой разности первого порядка. Будучи отношением приращения функции к приращению 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} . 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 $f'(x)$.

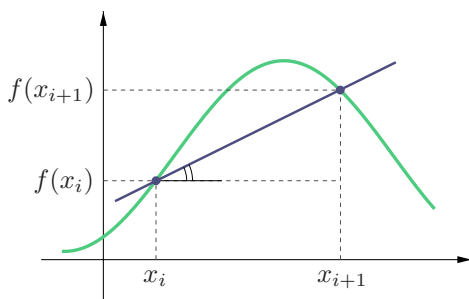


Рис. 2.4. Иллюстрация смысла разделённых разностей, как углового коэффициента секущей графика функции

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

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

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 (см. [15]). Разделённые разности-наклоны могут быть определены для функций многих переменных и даже для операторов, действующих из одного абстрактного пространства в дру-

гое. Интересно, что в начале XX века для обозначения этой конструкции использовался также термин «подъём функции» [79].

Для фиксированного набора узлов численные значения разделённых разностей любой функции нетрудно вычислить согласно определениям (2.15)–(2.16) или по формуле (2.19). 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 α, β справедливо

$$(\alpha f + \beta g)^\angle = \alpha f^\angle + \beta g^\angle \quad (2.18)$$

при одинаковых аргументах разделённых разностей. Это очевидно следует из определения для разделённой разности первого порядка, а для разделённых разностей высших порядков показывается 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*

$$f^\angle(x_i, x_{i+1}, \dots, x_{i+k}) = \sum_{j=i}^{i+k} \frac{f(x_j)}{\prod_{\substack{l=i \\ l \neq j}}^{i+k} (x_j - x_l)}. \quad (2.19)$$

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

$$(y_i, y_{i+1}, \dots, y_{i+k})^\angle = \sum_{j=i}^{i+k} \frac{y_j}{\prod_{\substack{l=i \\ l \neq j}}^{i+k} (x_j - x_l)}. \quad (2.20)$$

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

$$\begin{aligned}
 & f^{\angle}(x_i, x_{i+1}, \dots, x_{i+k+1}) \\
 &= \frac{f^{\angle}(x_{i+1}, x_{i+2}, \dots, x_{i+k+1}) - f^{\angle}(x_i, x_{i+1}, \dots, x_{i+k})}{x_{i+k+1} - x_i} \\
 &= \frac{1}{x_{i+k+1} - x_i} \cdot \left(\sum_{j=i+1}^{i+k+1} \frac{f(x_j)}{\prod_{\substack{l=i+1 \\ l \neq j}}^{i+k+1} (x_j - x_l)} - \sum_{j=i}^{i+k} \frac{f(x_j)}{\prod_{\substack{l=i \\ l \neq j}}^{i+k} (x_j - x_l)} \right) \\
 &\quad \text{согласно индукционному предположению} \\
 &= \frac{f(x_{i+k+1})}{(x_{i+k+1} - x_i) \prod_{l=i+1}^{i+k} (x_{i+k+1} - x_l)} \\
 &\quad + \frac{1}{x_{i+k+1} - x_i} \cdot \left(\sum_{j=i+1}^{i+k} \frac{f(x_j)}{\prod_{\substack{l=i+1 \\ l \neq j}}^{i+k+1} (x_j - x_l)} - \sum_{j=i+1}^{i+k} \frac{f(x_j)}{\prod_{\substack{l=i \\ l \neq j}}^{i+k} (x_j - x_l)} \right) \\
 &\quad - \frac{f(x_i)}{(x_{i+k+1} - x_i) \prod_{l=i+1}^{i+k} (x_i - x_l)}
 \end{aligned}$$

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

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 (2.19) или (2.20).

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

$$\frac{x^n - y^n}{x - y} = x^{n-1} + x^{n-2}y + \dots + xy^{n-2} + y^{n-1}. \quad (2.21)$$

Этот результат позволяет явно выписать разделённую разность для любой целой степени переменной. Для произвольного полинома далее можно воспользоваться свойством (2.18), т. е. линейностью разделённой разности.

Пример 2.2.1 Вычислим разделённые разности от полинома $g(x) = x^3 - 4x + 1$.

Будем искать по отдельности разделённые разности от мономов, образующих $g(x)$. В силу (2.21) имеем

$$\frac{x_2^3 - x_1^3}{x_2 - x_1} = x_2^2 + x_2x_1 + x_1^2.$$

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

$$g^\zeta(x_1, x_2) = x_2^2 + x_2x_1 + x_1^2 - 4.$$

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

$$\begin{aligned} g^\zeta(x_1, x_2, x_3) &= \frac{g^\zeta(x_2, x_3) - g^\zeta(x_1, x_2)}{x_3 - x_1} \\ &= \frac{(x_3^2 + x_3x_2 + x_2^2 - 4) - (x_2^2 + x_2x_1 + x_1^2 - 4)}{x_3 - x_1} \\ &= \frac{x_3^2 + (x_3 - x_1)x_2 - x_1^2}{x_3 - x_1} = x_1 + x_2 + x_3. \end{aligned}$$

Третья разделённая разность

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

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 (2.21). 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

$$f^{\prime\prime\prime}(x_0, x_1, \dots, x_n) = \frac{1}{n!} f^{(n)}(\xi) \quad (2.22)$$

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

Для разделённых разностей первого порядка этот факт непосредственно следует из теоремы Лагранжа о среднем (формулы конечных приращений), 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 §2.2д.

Существует более точное (хотя и более громоздкое) интегральное представление для разделённых разностей, о котором можно подробно узнать в [20, 64, 81].

2.2г Интерполяционный полином Ньютона

Выведем теперь другую форму интерполяционного полинома, которая в минимальной степени перестраивалась бы при смене набора узлов интерполяции.

Обозначим через $P_k(x)$ интерполяционный полином степени k , построенный по 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 очевидно следующее тождество

$$P_n(x) = P_0(x) + \sum_{k=1}^n (P_k(x) - P_{k-1}(x)). \quad (2.23)$$

Замечательность этого представления состоит в том, что при добавлении или удалении последних по номеру узлов интерполяции перестройке должны подвергнуться лишь те последние слагаемые суммы из правой части (2.23), которые вовлекает эти изменяемые узлы. Первые слагаемые в (2.23) зависят только от первых узлов интерполяции и останутся неизменными.⁵ Таким образом, стоящая перед нами задача окажется решённой, если будут найдены удобные и просто выписываемые 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 должны принимать 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 A_k вспомним, что по условию интерполяции $P_k(x_k) = y_k$. Следовательно

⁵Следует 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.

но,

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

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

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

в силу Предложения 2.2.1. Окончательно представление (2.23) принимает вид

$$P_n(x) = y_0 + (y_0, y_1)^{\angle} (x - x_0) + (y_0, y_1, y_2)^{\angle} (x - x_0)(x - x_1) + \dots + (y_0, y_1, \dots, y_n)^{\angle} (x - x_0)(x - x_1) \cdots (x - x_{n-1}). \quad (2.24)$$

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

$$P_n(x) = f(x_0) + f^{\angle}(x_0, x_1)(x - x_0) + f^{\angle}(x_0, x_1, x_2)(x - x_0)(x - x_1) + \dots + f^{\angle}(x_0, x_1, \dots, x_n)(x - x_0)(x - x_1) \cdots (x - x_{n-1}). \quad (2.25)$$

Выражения в правых частях равенств (2.24)–(2.25) 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 (2.23), на основе которого мы конструировали интерполяционный полином Ньютона, может быть уточнено и конкретизировано следующим образом:

$$P_n(x) = P_k(x) + f^{\angle}(x_0, x_1, \dots, x_{k+1})(x - x_0)(x - x_1) \cdots (x - x_k) + \dots + f^{\angle}(x_0, x_1, \dots, x_n)(x - x_0)(x - x_1) \cdots (x - x_{n-1}), \quad (2.26)$$

для любого k , такого что $0 \leq k \leq n - 1$. Образно выражаясь, формула (2.26) 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 $(x - x_0)(x - x_1) \dots (x - x_k)$, которые в случае равномерно расположенных и упорядоченных по возрастанию узлов x_0, x_1, \dots, x_k часто называют *обобщённой степенью* [10].

Практическое построение интерполяционного полинома Ньютона требует знания числовых значений всех разделённых разностей функции, и чаще всего наиболее удобно находить их по рекуррентным формулам (2.15)–(2.17).

Важнейший частный случай интерполирования относится к равномерному расположению узлов, когда величина $h_i = x_{i+1} - x_i$, 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

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

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

$$P_n(x) =$$

$$f(x_0) + \frac{1}{1!} \frac{\Delta f(x_0)}{h} (x - x_0) + \frac{1}{2!} \frac{\Delta^2 f(x_0)}{h^2} (x - x_0)(x - x_1) + \dots + \frac{1}{n!} \frac{\Delta^n f(x_0)}{h^n} (x - x_0)(x - x_1) \dots (x - x_{n-1}).$$

Таблица 2.1. Таблица конечных разностей функции

x	y	Δy	$\Delta^2 y$	\dots	$\Delta^n y$
x_0	y_0				
x_1	y_1	Δy_0	$\Delta^2 y_0$		
x_2	y_2	Δy_1	$\Delta^2 y_1$	\dots	$\Delta^n y_0$
x_3	y_3	Δy_2	$\Delta^2 y_2$	\dots	$\Delta^n y_1$
x_4	y_4	Δy_3	$\Delta^2 y_3$	\dots	$\Delta^n y_2$
\vdots	\vdots	\vdots	\vdots	\ddots	\vdots

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 (см. Табл. 2.1), где в дополнительных столбцах (третьем, четвёртом и т. д.), заполняемых последовательно один за другим слева направо, выписываются 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 [10, 21, 64, 79].

2.2д 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, на стр. 54. Именно, значения y_0, y_1, \dots, y_n принимают-

ся 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 интерполиант 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 интерполирования 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

$$R_n(f, z) = f^{\angle}(x_0, x_1, \dots, x_n, z) \cdot \omega_n(z), \quad (2.27)$$

где функция ω_n определяется посредством (2.12), т. е.

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

Доказательство. Выпишем для f интерполяционный полином Ньютона $(n+1)$ -й степени по узлам x_0, x_1, \dots, x_n, z . Согласно представлению (2.26)

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

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 (2.22)

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

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

Отметим прежде всего, что в (2.22) 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

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

заметим, что Предложение 2.2.2 и равенство (2.22) равносильны следующему утверждению: на $]x_0, x_n[$ существует точка ξ , 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 -ая производная по x от остаточного члена интерполяции $R_n(f, x) = f(x) - P_n(x)$, т. е.

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

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 $f^{\angle}(x_0, x_1, \dots, x_n)$ множитель является полиномом n -ой 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 -ой 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 интерполировании по несовпадающим узлам $x_0, x_1, \dots, x_n \in [a, b]$ с помощью полинома n -ой степени остаточный член $R_n(f, x)$ может быть представлен в виде

$$R_n(f, x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!} \cdot \omega_n(x), \quad (2.28)$$

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 интерполирования, то $R_n(f, x) = 0$, но в то же время и $\omega_n(x) = 0$. Поэтому в качестве ξ 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 интерполирования, то применяем Предложение 2.2.3, в котором разделённую разность выражаем через производную согласно результату Предложения 2.2.2. ■

Выражение (2.28) было получено О.Л. Коши в первой половине XIX века (см. [8]), 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, 81].

Если обозначить

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

— максимум абсолютного значения n -ой производной на рассматриваемом интервале, то нетрудно выписать огрублённые оценки, вытекающие из (2.28) и полезные при практическом вычислении погрешности интерполирования:

$$|R_n(f, x)| \leq \frac{M_{n+1}}{(n+1)!} \cdot |\omega_n(x)|, \quad (2.29)$$

или даже совсем простую

$$|R_n(f, x)| \leq \frac{M_{n+1}(b-a)^{n+1}}{(n+1)!}. \quad (2.30)$$

Если доступно явное выражение для $(n+1)$ -ой 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. §1.5).

Отметим, что полученные выше оценки — (2.28) и её следствия (2.29) и (2.30) — становятся неприменимыми, если функция f имеет гладкость, меньшую чем $n+1$. В то же время представление погрешности интерполирования через разделённые разности в виде (2.27) или в интегральной форме справедливо для любых функций.

В представлении (2.28) поведение полинома $\omega_n(x)$ при изменении x типично для полиномов с вещественными корнями 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 корней 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 корней полинома $\omega_n(x)$, то его абсолютная величина, а вместе с ней и погрешность алгебраической интерполяции, очень быстро растут. На Рис. 2.5 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 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 суще-

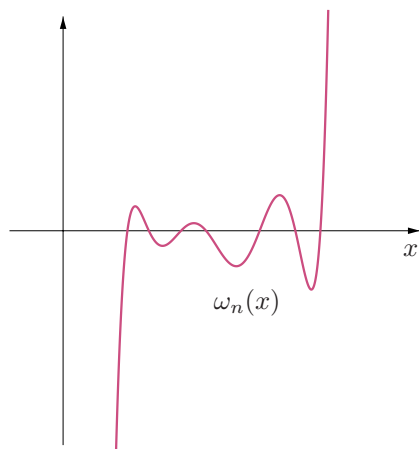


Рис. 2.5. Типичное поведение полинома $\omega_n(x)$: быстрый рост за пределами интервала узлов

ственными погрешностями, и потому не стоит использовать её слишком широко.

В рассмотренной выше постановке задачи интерполирования (§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 интерполирования, в которых мы можем управлять выбором узлов. При этом естественно возникает вопрос о том, как сделать этот выбор наилучшим образом, чтобы погрешность интерполирования 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 $f^{(j)}(x_0, x_1, \dots, x_n, z)$ и $f^{(n+1)}(\xi(x))/(n+1)!$ в выражениях для остаточного члена (2.27) или (2.28) 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. 2.23 из §2.10ж). 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.

2.3 Полиномы Чебышёва

2.3а Определение и основные свойства

Полиномы Чебышёва — это семейство алгебраических полиномов, обозначаемых по традиции⁶ как $T_n(x)$, и зависящих от неотрицательного целого параметра n . Они могут быть определены различными равносильными способами, и наиболее просто и наглядно их *тригонометрическое представление*:

$$T_n(x) = \cos(n \arccos x), \quad (2.31)$$

$x \in [-1, 1]$, $n = 0, 1, 2, \dots$. Как известно, всякий полином степени n однозначно определяется своими значениями в $(n + 1)$ точках, а формулой (2.31) мы фактически задаём значения функции в бесконечном множестве точек из $[-1, 1]$. Поэтому если посредством (2.31) 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 (2.31), — алгебраический полином степени n , и его старший коэффициент равен 2^{n-1} при $n \geq 1$.*

⁶С буквы «Т» начинаются немецкое (Tschebyshev) и французское (Tchebychev) написания фамилии П.Л. Чебышёва, открывшего эти полиномы в 1854 году.

Доказательство. Мы 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 проведения 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 (2.31)

$$T_{n+1}(x) = 2x T_n(x) - T_{n-1}(x) \quad (2.32)$$

для любых $n = 1, 2, \dots$.

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. ■

Полученные в доказательстве рекуррентная формула (2.32) позволяет, отправляясь от $T_0(x)$ и $T_1(x)$, последовательно выписать явные алгебраические выражения для полиномов Чебышёва:

$$\begin{aligned} T_0(x) &= 1, \\ T_1(x) &= x, \\ T_2(x) &= 2x^2 - 1, \\ T_3(x) &= 4x^3 - 3x, \\ T_4(x) &= 8x^4 - 8x^2 + 1, \\ T_5(x) &= 16x^5 - 20x^3 + 5x, \\ &\dots \qquad \dots \end{aligned} \quad (2.33)$$

По рекуррентной формуле (2.32) и следующим из неё явным выражениям (2.33) полиномы Чебышёва единообразно определяются для любых

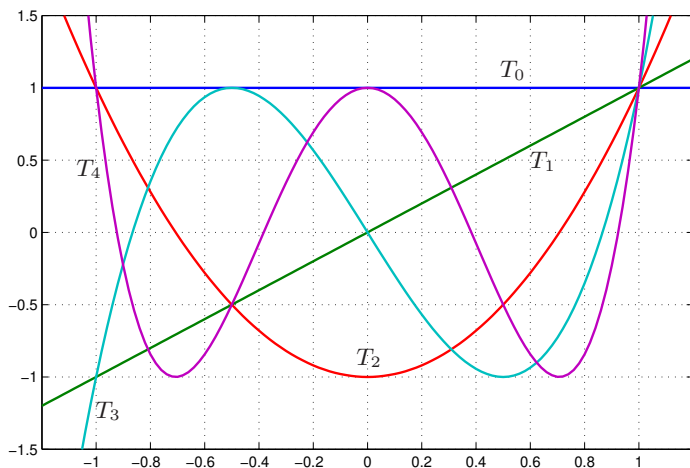


Рис. 2.6. Графики первых полиномов Чебышёва на интервале $[-1.2, 1.2]$.

значений аргумента x . Графики первых полиномов Чебышёва можно увидеть на Рис. 2.6.

Продолжая тему представления полиномов Чебышёва, заметим, что формула (2.31) справедлива, в действительности, при любых вещественных аргументах x , если для $\arccos x$ допустить комплексные значения и, соответственно, рассматривать косинус от комплексного аргумента. Можно показать, что

$$T_n(x) = \operatorname{ch}(n \operatorname{arch} x), \quad (2.34)$$

где $\operatorname{ch} z = \frac{1}{2}(e^z + e^{-z})$ — гиперболический косинус, а arch — обратная к нему функция. Определение (2.34) удобно применять для вещественных аргументов x , таких что $x \geq 1$.

Ещё одно полезное представление полиномов Чебышёва —

$$T_n(x) = \frac{1}{2} \left((x + \sqrt{x^2 - 1})^n + (x - \sqrt{x^2 - 1})^n \right), \quad (2.35)$$

$n = 0, 1, 2, \dots$ (см. [8, 29, 46, 64]). Для него нетрудно установить эквивалентность тригонометрическому представлению (2.31), сделав замену $x = \cos t$ для некоторого $t \in \mathbb{C}$. Тогда из (2.35) 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 (2.31).

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 (2.32).

Найдём корни полиномов Чебышёва на вещественном интервале $[-1, 1]$. Исходя из тригонометрического представления (2.31), вспомним, каковы нули косинуса. Должно быть

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

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

$$\hat{x}_k = \cos \frac{(2k+1)\pi}{2n}, \quad k = 0, 1, \dots, n-1, \quad (2.36)$$

всего n штук. А поскольку в силу основной теоремы алгебры у полинома $T_n(x)$ в поле комплексных чисел может быть не более n корней, то можем заключить, что других корней, отличных от (2.36), полином $T_n(x)$ не имеет.

Итак, все корни полинома Чебышёва $T_n(x)$ в самом деле находятся на интервале $[-1, 1]$ и выражаются в виде (2.36). Расположение корней полинома Чебышёва можно наглядно проиллюстрировать чертежом на Рис. 2.7, где эти корни соответствуют абсциссам точек пересечения единичной окружности с центром в начале координат с радиусами, откладываемыми через одинаковые доли развёрнутого угла в π радиан. Из этой иллюстрации хорошо видно, что корни полинома Чебышёва

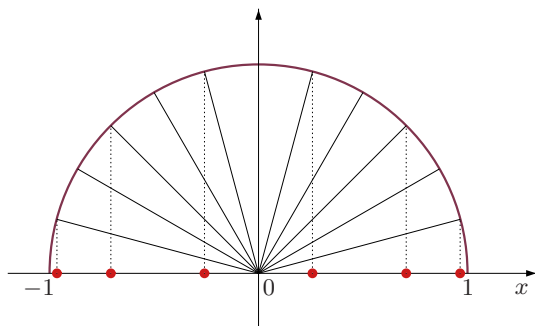


Рис. 2.7. Иллюстрация расположения корней полинома Чебышёва шестой степени.

расположены существенно неравномерно: они сгущаются к концам интервала $[-1, 1]$, а в его средней части более разрежены.

Из тригонометрического представления (2.31) 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 (2.31) должен достигать максимальных по модулю значений ± 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

$$\max_{x \in [-1, 1]} |Q_n(x)| \geq \max_{x \in [-1, 1]} |\tilde{T}_n(x)| = 2^{1-n}. \quad (2.37)$$

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, выполняется неравенство

$$\max_{x \in [-1, 1]} |Q_n(x)| < \max_{x \in [-1, 1]} |\tilde{T}_n(x)|, \quad (2.38)$$

которое противоположно по смыслу неравенству (2.37). 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 быть справедливо

$$\begin{aligned} \operatorname{sgn}(\tilde{T}_n(x_s) - Q_n(x_s)) &= \operatorname{sgn}((-1)^s 2^{1-n} - Q_n(x_s)) \\ &= \operatorname{sgn}((-1)^s 2^{1-n}) \quad \text{в силу (2.38)} \\ &= (-1)^s. \end{aligned}$$

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 обязан иметь корень. 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 корней 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 (2.38). ■

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 лишь косвенным образом использует то обстоятельство, что полиномы рассматриваются на интер-

вале $[-1, 1]$. 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

$$y = \frac{1}{2}(b + a) + \frac{1}{2}(b - a)x \quad (2.39)$$

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

$$x = \frac{2y - (b + a)}{(b - a)}, \quad (2.40)$$

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

$$y_k = \frac{1}{2}(b + a) + \frac{1}{2}(b - a) \cos \frac{(2k + 1)\pi}{2n}, \quad k = 0, 1, \dots, n - 1. \quad (2.41)$$

Свойство, аналогичное Предложению 2.3.2, будет верно на интервале $[a, b]$ для полинома, полученного из $T_n(x)$ с помощью линейной замены переменных (2.40) и масштабирования (нормировки).

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

$$2^{1-2n} (b - a)^n \cdot T_n \left(\frac{2y - (b + a)}{b - a} \right), \quad (2.42)$$

имеет старший коэффициент 1 и на интервале $[a, b]$ равномерно наименее уклоняется от нуля среди всех полиномов степени n со старшим коэффициентом 1.

Доказательство. Первое утверждение Предложения вытекает из того, что в результате замены переменной (2.40) из полинома n -ой 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 (2.42) 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 (2.39) 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 §2.2д 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

$$x_k = \frac{1}{2}(b+a) + \frac{1}{2}(b-a) \cos\left(\frac{(2k+1)\pi}{2(n+1)}\right), \quad k = 0, 1, \dots, n, \quad (2.43)$$

т.е. корнями полинома вида (2.42), который получается в результате замены переменных (2.40) из полинома Чебышёва $(n+1)$ -ой 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 (2.28) для остаточного члена интерполяции, совпадёт с полиномом $(n+1)$ -ой степени вида (2.42). 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. 83) будет наименьшей возможной. Узлы интерполяции (2.43) называют *чебышёвскими узлами* на интервале $[a, b]$, а в совокупности они образуют *чебышёвскую сетку* на $[a, b]$.

Помимо интерполирования полиномы Чебышёва и их обобщения имеют и другие важные применения в различных задачах вычислительной математики и анализа [54, 59, 72]. Полиномы Чебышёва образуют систему ортогональных функций относительно скалярного произведения из $\mathcal{L}^2[-1, 1]$ с весом $(1-x^2)^{-1/2}$ (см. §2.10ж). По этой причине очень важное значение имеют, к примеру, разложения функций в ряды по полиномам Чебышёва (см., к примеру, [59]).

2.3в Обусловленность задачи алгебраической интерполяции

Предположим, что при алгебраическом интерполировании функции её значения в узлах вычисляются с некоторой погрешностью. Как это

отразится на значениях интерполяционного многочлена?

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 вместо точных значений $y_i = f(x_i)$, $i = 0, 1, \dots, n$, имеются их приближённые значения \tilde{y}_i с абсолютной погрешностью ε , так что

$$|\tilde{y}_i - y_i| \leq \varepsilon.$$

Тогда вместо точного интерполяционного полинома $P_n(x)$ мы получим полином

$$\tilde{P}_n(x) = \sum_{i=0}^n \tilde{y}_i \phi_i(x),$$

где, как и прежде,

$$\phi_i(x) = \frac{\prod_{j \neq i} (x - x_j)}{\prod_{j \neq i} (x_i - x_j)}, \quad i = 1, 2, \dots, n,$$

— базисные интерполяционные полиномы Лагранжа. Абсолютная погрешность значения приближённого интерполяционного полинома \tilde{P}_n в точке x равна, следовательно,

$$\begin{aligned} |\tilde{P}_n(x) - P_n(x)| &= \left| \sum_{i=0}^n (\tilde{y}_i - y_i) \phi_i(x) \right| \\ &\leq \sum_{i=0}^n |\tilde{y}_i - y_i| |\phi_i(x)| \leq \varepsilon \sum_{i=0}^n |\phi_i(x)|. \end{aligned}$$

В целом для интервала $[a, b]$

$$\max_{x \in [a, b]} |\tilde{P}_n(x) - P_n(x)| \leq \varepsilon \max_{x \in [a, b]} \sum_{i=0}^n |\phi_i(x)|.$$

Таким образом, величину, стоящую множителем при ε в правой части 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 Для заданного интервала $[a, b] \subset \mathbb{R}$ и набора узлов (сетки) 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 -ой константой Лебега.⁷

Константа Лебега Λ_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

$$\Lambda_n > \frac{1}{8\sqrt{\pi}} \ln n.$$

Доказательство можно увидеть, к примеру, в [7].

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 [7, 61], а для левого неравенства — в [36].

Теорема 2.3.3 Для последовательности чебышёвских сеток на заданном интервале интерполяции справедливо неравенство

$$\Lambda_n \leq 8 + \frac{4}{\pi} \ln n.$$

⁷Иногда говорят «интерполяционной константой Лебега», так как существуют также «константы Лебега», относящиеся к сходимости рядов Фурье.

Результат Теоремы 2.3.3 принадлежит С.Н. Бернштейну, а его доказательство можно увидеть в [9] и для несколько упрощённой формулировки в [36]. 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. В отличие от оценки О. Коши (2.28), такой способ оценивания погрешности не опирается на гладкость интерполируемой функции и неравенства для её высших производных. Подробности интересующийся читатель может увидеть, к примеру, в [7, 9, 61].

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 *задачей алгебраической интерполяции с кратными узлами* мы будем называть следующую постановку. Даны несовпадающие точки x_i , $i = 0, 1, \dots, n$, — узлы интерполирования, в которых заданы значения $y_i^{(k)}$, $k = 0, 1, \dots, N_i - 1$, — их принимают интерполируемая функция 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,

$$H_m^{(k)}(x_i) = y_i^{(k)}, \quad i = 0, 1, \dots, n, \quad k = 0, 1, \dots, N_i - 1, \quad (2.44)$$

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.⁸ Эта постановка задачи алгебраической интерполяции с кратными узлами не является самой общей (порядки производных, для которых задаются значения в узлах, идут в ней «без пробелов»), но она достаточно практична и хорошо исследована.

Теорема 2.4.1 *Решение задачи эрмитовой интерполяции с кратными узлами при $m = N_0 + N_1 + \dots + N_n - 1$ 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 (2.44). Получим систему линейных

⁸Не следует путать «интерполяционный полином Эрмита» с известными ортогональными полиномами, которые тоже носят его имя и которые правильнее было бы называть «ортогональными полиномами Чебышёва-Эрмита».

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

$$Ga = y, \quad (2.45)$$

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 (2.44).

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 (2.45) 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 интерполанта становятся его корнями. Поэтому если правая часть в (2.45) — нулевая, то это означает, что полином $H_m(x)$ с учётом кратности имеет $N_0 + N_1 + \dots + N_n = m + 1$ корней, 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 $a = (a_0, a_1, \dots, a_m)^\top = (0, 0, \dots, 0)^\top$.

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 (2.45) 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. Тем не менее, аналогично случаю интерполяции с простыми узлами, желательно иметь её аналитическое решение в виде обозримого конечного выражения для интерполанта. Он может иметь форму Лагранжа либо форму Ньютона (см. подробности, к примеру, [3, 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.

Аналогично §2.2б, 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 α и β 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 (2.45). 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

$$\phi_{ik}^{(l)}(x_j) = \begin{cases} 0, & \text{при } i \neq j \text{ или } k \neq l, \\ 1, & \text{при } i = j \text{ и } k = l. \end{cases} \quad (2.46)$$

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 (2.45) 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 (2.46) в виде

$$\phi_{ik}^{(l)}(x_i) = \delta_{kl}, \quad k = 0, 1, \dots, N_i - 1, \quad (2.47)$$

$$\begin{aligned} \phi_{ik}^{(l)}(x_j) &= 0, & j &= 0, 1, \dots, i-1, i+1, \dots, n, \\ & & l &= 0, 1, \dots, N_i - 1. \end{aligned} \quad (2.48)$$

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. (2.47). И так далее.

Мы не будем завершать этого построения, так как дальнейшие выкладки весьма громоздки, а алгоритм нахождения полинома из приведённых рассуждения вполне ясен. Детали построения интерполяционного полинома Эрмита можно увидеть, к примеру, в книге [8].

Какова погрешность алгебраической 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

$$R_m(f, x) = f(x) - H_m(x) = \frac{f^{(m+1)}(\xi(x))}{(m+1)!} \cdot \prod_{i=0}^n (x - x_i)^{N_i}, \quad (2.49)$$

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 (2.49), т. е.

$$\Omega(x) := \prod_{i=0}^n (x - x_i)^{N_i}.$$

Это — аналог функции $\omega_n(x)$, введённой в §2.2д и широко используемой в различных рассуждениях.

Если $x = x_i$ для одного из узлов интерполирования, $i = 0, 1, \dots, n$, то $R_m(f, x) = 0$, но в то же время и $\Omega(x) = 0$. Поэтому в (2.49) в качестве $\xi(x)$ можно взять любую точку из интервала $]a, b[$.

Предположим теперь, что точка x из интервала интерполирования $[a, b]$ 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

$$\psi^{(m+1)}(z) = f^{(m+1)}(z) - K(m+1)!, \quad (2.50)$$

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, всего производная $\psi'(z)$ должна иметь 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, $(m + 1)$ -ая производная зануляется на $[a, b]$ хотя бы один раз.

Итак, на интервале $[a, b]$ обязательно найдётся по крайней мере одна точка ξ , зависящая, естественно, от x и такая что $\psi^{(m+1)}(\xi) = 0$. Поэтому в силу равенства (2.50) получаем

$$K = \frac{f^{(m+1)}(\xi)}{(m+1)!}.$$

Принимая во внимание определение константы K , немедленно получаем отсюда формулу (2.49). ■

Интересно, что при наличии одного узла кратности m интерполяционный полином Эрмита становится полиномом Тейлора, а формула (2.49) совпадает с известной формулой остаточного члена (в форме Лагранжа) для полинома Тейлора. Если же все узлы интерполяции простые, то (2.49) превращается в полученную ранее формулу погрешности простой интерполяции (2.28).

2.5 Общие факты интерполяции

2.5а Интерполяционный процесс

Как с теоретической, так и с практической точек зрения интересен вопрос о том, насколько малой может быть сделана погрешность

интерполирования при возрастании числа узлов. Вообще, сходятся ли интерполяционные полиномы к интерполируемой функции при неограниченном росте количества узлов? Конечно, по условиям интерполяции исходная функция равна своему интерполанту 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 интерполантов, необходимо формализовать некоторые необходимые понятия.

Определение 2.5.1 Пусть для интервала $[a, b]$ задана бесконечная треугольная матрица узлов

$$\begin{pmatrix} x_0^{(0)} & 0 & 0 & 0 & \cdots \\ x_0^{(1)} & x_1^{(1)} & 0 & 0 & \cdots \\ x_0^{(2)} & x_1^{(2)} & x_2^{(2)} & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}, \quad (2.51)$$

такая что в каждой её строке расположены различные точки интервала $[a, b]$, т. е. $x_i^{(n)} \in [a, b]$ для всех неотрицательных целых 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 -ой строки этой матрицы берутся в качестве узлов интерполяции, по которым строится последовательность интерполантов $g_n(x)$, $n = 0, 1, 2, \dots$.

Если в этом определении все интерполанты $g_n(x)$ являются алгебраическими полиномами, то употребляется термин *алгебраический интерполяционный процесс*.

Ясно, что Определение 2.5.1 предполагает наличие бесконечной треугольной матрицы, похожей на (2.51) и составленной из значений $y_i^{(n)}$, которые интерполанты $g_n(x)$ 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 $n \rightarrow \infty$. Интерполяционный процесс для функции f на интервале $[a, b]$ называется сходящимся равномерно, если $\max_{x \in [a, b]} |f(x) - g_n(x)| \rightarrow 0$ при $n \rightarrow \infty$ для порождаемой этим процессом последовательности интерполянтов $g_n(x)$.

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.

2.5б Сводка результатов и обсуждение

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.

Если $f : \mathbb{R} \supset [a, b] \rightarrow \mathbb{R}$ — непрерывная функция, то для всякого $\epsilon > 0$ существует полином $P_n(x)$ степени $n = n(\epsilon)$, равномерно приближающий функцию f с погрешностью, не большей ϵ , т. е. такой, что

$$\max_{x \in [a, b]} |f(x) - P_n(x)| \leq \epsilon.$$

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 интерполанта на данном множестве точек-узлов. В задаче приближения функций участки области определения, где сравниваемые функции совпадают друг с другом, не фиксированы, и совершенно аналогично участки, где функции отклоняются

друг от друга, могут «гулять» по интервалу. Таким образом, теорема Вейерштрасса не даёт ответа на конкретные вопросы о решении задачи интерполирования.

Как следует из результатов §2.2д 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, включающим и концы этого интервала. Не слишком трудными рассуждениями показывается (см. [8, 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 . Может показаться, что причиной расходимости интерполяционного процесса в примере С.Н. Бернштейна является отсутствие гладкости интерполируемой функции, но это верно лишь отчасти.

Предположим, что интерполируемая функция f имеет бесконечную гладкость, т.е. $f \in C^\infty[a, b]$, и при этом её производные растут «не слишком быстро». В последнее условие будем вкладывать следующий смысл:

$$\sup_{x \in [a, b]} |f^{(n)}(x)| < M^n, \quad n = 1, 2, \dots, \quad (2.52)$$

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

$$\frac{(M(b-a))^n}{n!},$$

то есть при $n \rightarrow \infty$ она очевидно 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 (2.52) 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* [48]. В теории функций показывается, что степенной ряд, о котором говорится в этом определении, в действительности является рядом Тейлора, и 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 [3, 64].

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. 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 при их принадлежности какой-то ограниченной области специальной формы, содержащей интервал интерполирования [8, 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 интерполирования. По-видимому, наиболее известный пример такого рода привёл немецкий математик К. Рунге в 1901 году [93].⁹

В примере Рунге функция

$$\Upsilon(x) = \frac{1}{1+x^2}$$

⁹Нередко его называют также «явлением Рунге» или «феноменом Рунге».

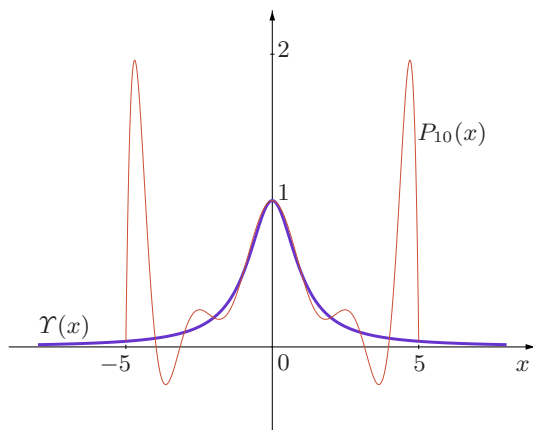


Рис. 2.8. Интерполяция полиномом 10-й степени в примере Рунге

на интервале $[-5, 5]$ интерполируется алгебраическими полиномами, которые построены на последовательности равномерных сеток с узлами $x_i = -5 + 10i/n$, $i = 0, 1, \dots, n$ ($n > 0$). Если $P_n(x)$ — интерполяционный полином n -ой степени, построенный по n -ой сетке, то оказывается, что

$$\lim_{n \rightarrow \infty} \max_{x \in [-1, 1]} |\mathcal{Y}(x) - P_n(x)| = \infty.$$

При этом вблизи концов интервала интерполирования $[-5, 5]$ у полиномов $P_n(x)$ с ростом n возникают сильные колебания (называемые *осцилляциями*), размах которых стремится к бесконечности (см. Рис. 2.8). Получается, что хотя в узлах интерполирования значения функции $\mathcal{Y}(x)$ совпадают со значениями интерполяционного полинома $P_n(x)$, между этим узлами $\mathcal{Y}(x)$ и $P_n(x)$ могут отличаться сколь угодно сильно, даже несмотря на плавный (бесконечно гладкий) характер изменения функции $\mathcal{Y}(x)$ и равномерное сгущение узлов интерполяции.¹⁰

Интересно, что на интервале $[-\kappa, \kappa]$, где $\kappa \approx 3.63$, рассматриваемый интерполяционный процесс равномерно сходится к $\mathcal{Y}(x)$ (см. [93]). Функция $\mathcal{Y}(x)$ имеет производные всех порядков для любого вещественного аргумента x , но у концов интервала интерполирования $[-5, 5]$ эти производные растут очень быстро и уже не удовлетворяют усло-

¹⁰Помимо оригинальной статьи К. Рунге [93] этот факт строго обосновывается, к примеру, в книге [79].

вию (2.52). Проявив некоторую изобретательность (детали описаны в пункте 116 тома 1 известной книги Г.М. Фихтенгольца [38]), можно показать, что при $x > 0$

$$\mathcal{Y}^{(n)}(x) = \left(\frac{1}{1+x^2} \right)^{(n)} = \frac{(-1)^n (n-2)!}{(1+x^2)^{(n-1)/2}} \cdot \sin \left((n-1) \arctg \frac{1}{x} \right). \quad (2.53)$$

Другой способ многократного дифференцирования функции из примера Рунге может быть основан на её комплексном представлении

$$\mathcal{Y}(x) = \frac{1}{1+x^2} = \frac{i}{2} \left(\frac{1}{x+i} - \frac{1}{x-i} \right).$$

Как следствие, легко получаем

$$\mathcal{Y}^{(n)}(x) = \left(\frac{1}{1+x^2} \right)^{(n)} = \frac{i}{2} (-1)^n n! \left(\frac{1}{(x+i)^{n+1}} - \frac{1}{(x-i)^{n+1}} \right), \quad (2.54)$$

откуда нетрудно вывести вещественную производную. Множители в виде факториалов в выражениях (2.53) и (2.54) определяет общее поведение производных при росте n . Таким образом, несмотря на простой вид, функция $\mathcal{Y}(x)$ 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, [48]).

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 δ , 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 δ . Можно показать (см. [29]), что $\omega_f(\delta)$ — неубывающая неотрицательная функция от δ , 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 непрерывна на конечном интервале

$[a, b]$, тогда и только тогда, когда её модуль непрерывности стремится к нулю при $\delta \rightarrow 0$. При этом по скорости убывания модуля непрерывности функции можно судить о весьма тонких свойствах 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 δ модуль непрерывности убывает быстрее, чем $|\ln \delta|^{-1}$. Оказывается, что если функция удовлетворяет условию Дини-Липшица, то на последовательности чебышёвских сеток из заданного интервала алгебраический интерполяционный процесс сходится к ней равномерно.¹¹ Обоснование этого результата читатель может найти в [9, 29, 61]. 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.

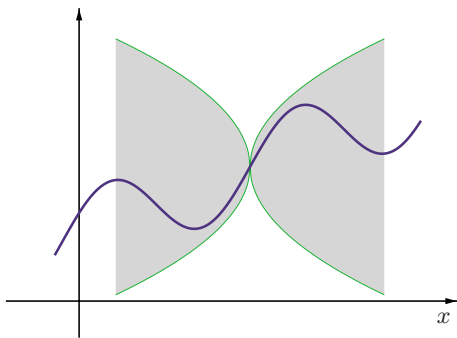


Рис. 2.9. Иллюстрация обобщённого условия Липшица

Удобным достаточным условием, при котором выполняется условие Дини-Липшица, является *обобщённое условие Липшица*: для любых x, y из области определения функции f имеет место

$$|f(x) - f(y)| \leq C|x - y|^\alpha, \quad (2.55)$$

¹¹Условие Дини-Липшица на функцию является также достаточным условием равномерной сходимости к ней ряда Фурье, см. [8].

с некоторыми константами C и α , $0 < \alpha \leq 1$ (см. Рис. 2.9). Как следствие, справедлива

Теорема 2.5.2 *Если функция удовлетворяет обобщённому условию Липшица (2.55), 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 [8, 9, 28]¹² *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 слишком 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, [38], том 2, пункт 444). Такой же 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 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

¹²Часто её называют «теоремой Фабера-Бернштейна».

дополнительных естественных условий, которые выделяли бы классы функций, непрерывных в более сильном смысле 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 условий Липшица, Дини-Липшица, обобщённого условия Липшица и ряда других им аналогичных.

С другой стороны, для общих непрерывных функций имеет место «оптимистический» результат, практическая ценность которого, правда, невелика:

Теорема Марцинкевича [8, 9, 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.*

Интересно, что ситуация со сходимостью интерполяционных процессов в среднеквадратичном смысле более благоприятна, чем для равномерной сходимости. Если рассматривается среднеквадратичное состояние между функциями (2.3) или более общее (2.112) с некоторым весом $\varrho(x)$, то, взяв бесконечную треугольную матрицу узлов из интервала интерполирования по корням так называемых ортогональных, для данного веса $\varrho(x)$, полиномов, мы получим сходимость интерполяционного процесса для любой непрерывной функции (см. подробности в [9, 29] и цитированной там литературе). Ортогональные полиномы будут обсуждаться далее в §2.10ж и §2.11, где мы детально рассмотрим полиномы Лежандра, ортогональные с единичным весом на интервале $[-1, 1]$ (см. §2.11). Сетки по их корням обеспечивают среднеквадратичную сходимость интерполяционного процесса для любой непрерывной функции.

Отметим, что полиномы Чебышёва — это тоже ортогональные полиномы, но с некоторым специальным весом. Чебышёвские сетки также обеспечивают сходимость интерполяционных процессов в среднеквадратичном смысле для любых непрерывных функций.

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 ин-

струменты интерполяции, т. е. использовать в качестве интерполянтов другие классы функций. Их рассмотрению будут посвящены следующие параграфы.

2.6 Сплайны

2.6а Элементы теории

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*.

Далее мы не будем рассматривать какие-либо другие сплайны помимо полиномиальных, и потому для краткости станем говорить просто о сплайнах, опуская прилагательное «полиномиальный».

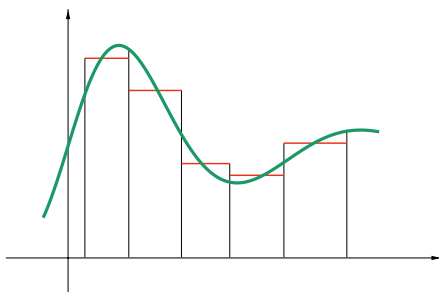


Рис. 2.10. Кусочно-постоянная интерполяция функции.

Почему именно кусочные полиномы? К идее их введения можно прийти, к примеру, отправляясь от следующих неформальных моти-

ваний. Содержательный (механический, физический, биологический и т. п.) смысл имеют, как правило, производные порядка не выше 2–4, и именно их мы можем видеть в различных математических моделях реальных явлений.¹³ Пятые 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 ощутимы 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 (см. Рис. 2.10). Далее мы восстанавливаем искомую функцию, последовательно применяя 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 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, изгибая, использовали чертёжники для проведения гладкой линии между данными фиксированными точками. Понятие «сплайн-функции» было введено И. Шёнбергом в 1946 году [94], 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 [89] явно использовал конструкции обычных сплайнов и так называемых *B*-сплайнов.¹⁴

С середины XX века по настоящее время сплайны нашли широкие применения в математике и её приложениях. В теории 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

¹³Характерный пример: в книге А.К. Маловичко, О.Л. Тарунина «Использование высших производных при обработке и интерпретации результатов геофизических наблюдений» (Москва, издательство «Недра», 1981 год) рассматриваются производные только второго и третьего порядков.

¹⁴Вклад Н.И. Лобачевского даже дал повод некоторым авторам называть *сплайнами Лобачевского* специальный вид сплайнов.

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 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.

Сплайны существенно лучше 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. Что касается сплайнов, то наличие у них кусочного 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,¹⁵ между которыми эта производная равна нулю. «В среднем» же $(q + 2)$ -я производная от сплайна оказывается не равной тождественно нулю!

Чем больше дефект сплайна, тем больше он отличается от полинома и тем более специфичны его свойства. Но слишком большой дефект приводит к существенному понижению общей гладкости сплайна. 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, будучи «непрерывно склеенным» в узлах x_i , $i = 1, 2, \dots, n - 1$. Иными словами, это — кусочно-линейная функция, имеющая, несмотря на

¹⁵Этому утверждению можно даже придать строгий математический смысл, привлекая понятие так называемой обобщённой функции.

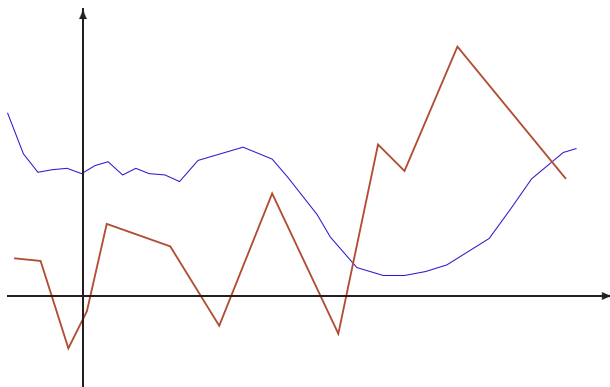


Рис. 2.11. Простейшие сплайны — кусочно-линейные функции.

свою простоту, богатые приложения в математике.¹⁶ Сплайны второй степени часто называют *параболическими сплайнами*.

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 .

Всего $d(n-1) + (n+1) = n(d+1) - (d-1)$ штук, и потому для 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, требуя, чтобы на каждом подинтервале $[x_{i-1}, x_i]$ сплайн являлся бы 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,

¹⁶Вспомним, к примеру, «ломанные Эйлера», которые применяются при доказательстве существования решения задачи Коши для обыкновенных дифференциальных уравнений [42].

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 интерполяционные сплайны нечётной степени 3, узлы которых совпадают с узлами интерполирования. 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, необходимом, 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, мы называем совокупность всех узлов *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 определения $4n$ неизвестных величин мы имеем всего $3(n-1) + (n+1) = 4n - 2$ условий. Два недостающих условия определяют 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), задающие на концах интервала $[a, b]$ первую или вторую производную 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.

Мы 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 : пусть

$$S(x) = \alpha_i + \beta_i(x - x_i) + \gamma_i \frac{(x - x_i)^2}{2} + \vartheta_i \frac{(x - x_i)^3}{6} \quad (2.56)$$

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. Ясно, что в такой форме представления сплайна величины β_0 и γ_0 совпадают по смыслу с теми, что даются в условиях (I)–(II) выше. Более того, из представления (2.56) вытекает, что

$$S''(x_i) = \gamma_i, \quad i = 1, 2, \dots, n-1.$$

Далее мы, во-первых, выведем из (2.56) 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 искомым сплайн при любых γ_0 и γ_n .

Заметим, что вторая производная $S''(x)$ является линейной функцией на $[x_i, x_{i+1}]$, и с учётом (2.56) должно быть

$$S''(x) = \gamma_i + \vartheta_i(x - x_i), \quad x \in [x_i, x_{i+1}]. \quad (2.57)$$

С другой стороны, вид этой линейной функции полностью задаётся двумя её крайними значениями γ_i и γ_{i+1} на концах соответствующего подинтервала $[x_i, x_{i+1}]$. Поэтому вместо (2.57) можно выписать более определённое представление, уже не задействующее ϑ_i . Именно, для $x \in [x_i, x_{i+1}]$, $i = 0, 1, \dots, n-1$, справедливо

$$S''(x) = \gamma_i \frac{x_{i+1} - x}{h_{i+1}} + \gamma_{i+1} \frac{x - x_i}{h_{i+1}}, \quad (2.58)$$

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 γ_0 и γ_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 (2.58), получим для $x \in [x_i, x_{i+1}]$

$$S(x) = \gamma_i \frac{(x_{i+1} - x)^3}{6h_{i+1}} + \gamma_{i+1} \frac{(x - x_i)^3}{6h_{i+1}} + C_1 x + C_2 \quad (2.59)$$

с какими-то константами C_1 и C_2 . Но нам будет удобно представить это выражение в несколько другом виде:

$$S(x) = \gamma_i \frac{(x_{i+1} - x)^3}{6h_{i+1}} + \gamma_{i+1} \frac{(x - x_i)^3}{6h_{i+1}} + K_1(x_{i+1} - x) + K_2(x - x_i), \quad (2.60)$$

где K_1 и K_2 — тоже константы.¹⁷ Насколько законен переход к такой форме? Из сравнения (2.59) и (2.60) следует, что C_1 и C_2 должны быть

¹⁷Строго говоря, константы C_1, C_2, K_1, K_2 нужно было бы снабдить ещё дополнительным индексом i , показывающими их зависимость от подинтервала $[x_i, x_{i+1}]$, к которому они относятся. Мы не делаем этого ради краткости изложения.

связаны с K_1 и K_2 посредством формул

$$\begin{aligned} C_1 &= -K_1 + K_2, \\ C_2 &= K_1 x_{i+1} - K_2 x_i. \end{aligned}$$

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 (2.59) и (2.60) совершенно равносильны друг другу.

Для определения K_1 и K_2 воспользуемся интерполяционными условиями. Подставляя в выражение (2.60) 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 (2.60) 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:

$$\begin{aligned} S(x) &= y_i \frac{x_{i+1} - x}{h_{i+1}} + y_{i+1} \frac{x - x_i}{h_{i+1}} \\ &+ \gamma_i \frac{(x_{i+1} - x)^3 - h_{i+1}^2(x_{i+1} - x)}{6h_{i+1}} + \gamma_{i+1} \frac{(x - x_i)^3 - h_{i+1}^2(x - x_i)}{6h_{i+1}}. \end{aligned} \quad (2.61)$$

Оно не содержит уже величин α_i , β_i и ϑ_i , которые фигурировали в исходном представлении (2.56) для $S(x)$, но неизвестными остались γ_1 , γ_2 , \dots , γ_{n-1} (напомним, что γ_0 и γ_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} :

$$S'(x_i - 0) = S'(x_i + 0), \quad i = 1, 2, \dots, n-1. \quad (2.62)$$

Продифференцировав по x формулу (2.61), будем иметь для подынтервала $[x_i, x_{i+1}]$ представление

$$S'(x) = \frac{y_{i+1} - y_i}{h_{i+1}} - \gamma_i \frac{3(x_{i+1} - x)^2 - h_{i+1}^2}{6h_{i+1}} + \gamma_{i+1} \frac{3(x - x_i)^2 - h_{i+1}^2}{6h_{i+1}}. \quad (2.63)$$

Следовательно, с учётом того, что $x_{i+1} - x_i = h_{i+1}$, получим

$$\begin{aligned} S'(x_i) &= \frac{y_{i+1} - y_i}{h_{i+1}} - \gamma_i \frac{3(x_{i+1} - x_i)^2 - h_{i+1}^2}{6h_{i+1}} - \gamma_{i+1} \frac{h_{i+1}^2}{6h_{i+1}} \\ &= \frac{y_{i+1} - y_i}{h_{i+1}} - \gamma_i \frac{h_{i+1}}{3} - \gamma_{i+1} \frac{h_{i+1}}{6}. \end{aligned} \quad (2.64)$$

С другой стороны, сдвигая все индексы в (2.63) на единицу назад, будем иметь для подынтервала $[x_{i-1}, x_i]$ 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

$$\begin{aligned} S'(x_i) &= \frac{y_i - y_{i-1}}{h_i} + \gamma_{i-1} \frac{h_i^2}{6h_i} + \gamma_i \frac{3(x_i - x_{i-1})^2 - h_i^2}{6h_i} \\ &= \frac{y_i - y_{i-1}}{h_i} + \gamma_{i-1} \frac{h_i}{6} + \gamma_i \frac{h_i}{3}. \end{aligned} \quad (2.65)$$

Приравнивание, согласно (2.62), производных (2.64) и (2.65), которые получены в узлах x_i с соседних подынтервалов $[x_{i-1}, x_i]$ и $[x_i, x_{i+1}]$, приводит к соотношениям

$$\begin{cases} \frac{h_i}{6} \gamma_{i-1} + \frac{h_i + h_{i+1}}{3} \gamma_i + \frac{h_{i+1}}{6} \gamma_{i+1} = \frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i}, \\ \gamma_0 \text{ и } \gamma_n \text{ заданы.} \end{cases} \quad i = 1, 2, \dots, n-1, \quad (2.66)$$

Это система линейных алгебраических уравнений относительно неизвестных переменных $\gamma_1, \gamma_2, \dots, \gamma_{n-1}$, имеющая матрицу

$$\frac{1}{6} \begin{pmatrix} 2(h_1 + h_2) & h_2 & & & 0 \\ h_2 & 2(h_2 + h_3) & h_3 & & \\ & h_3 & 2(h_3 + h_4) & \ddots & \\ & & \ddots & \ddots & \ddots \\ 0 & & & h_{n-1} & 2(h_{n-1} + h_n) \end{pmatrix}$$

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. §3.8). Кроме того, наша матрица зависит только от узлов x_i (но не от значений функции y_i), а в правых частях системы (2.66) 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. §2.8а).

Ещё одно полезное свойство матрицы системы (2.66) — это *диагональное преобладание* (см. §3.5в, стр. 343): стоящие на её главной диагонали элементы $\frac{1}{3}(h_i + h_{i+1})$ по модулю больше, чем суммы модулей внедиагональных элементов в соответствующих строках. В силу признака Адамара (он рассматривается далее в §3.5в, стр. 343) такие матрицы неособенны. Как следствие, система линейных уравнений (2.66) относительно неизвестных γ_i , $i = 1, 2, \dots, n-1$, однозначно разрешима при любой правой части, а искомый сплайн всегда существует и единствен.

Для нахождения решения системы (2.66) с трёхдиагональной матрицей может быть с успехом применён метод прогонки, описываемый ниже в §3.8. Найдя $\gamma_1, \gamma_2, \dots, \gamma_{n-1}$, подставим их в формулу (2.61), и это даст выражения для полиномов, определяющих искомый сплайн на каждом из отдельных подинтервалов $[x_i, x_{i+1}]$.

2.6в Погрешность интерполирования с помощью кубических сплайнов

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 $h = \max_{1 \leq i \leq n} h_i$.

При формулировке этого утверждения и далее в этой книге мы пользуемся символом $O(\cdot)$ — « O -большое», введённым Э. Ландау (вместе с « o -малым») и широко используемым в современной математике и её приложениях. Для двух переменных величин u и v пишут, что $u = O(v)$ в рассматриваемом процессе, 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 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.

Фактически, мы свели в формулировку Теоремы 2.6.1 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, 33]. 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 интерполирования, 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. Берн-

штейна и к функции $\mathcal{T}(x) = 1/(1+x^2)$ из примера Рунге, рассмотренным выше в §2.5. Обзор результатов о сходимости интерполяционных процессов со сплайнами и оценок погрешностей можно найти в журнальной работе [45]. Важно, что с повышением гладкости интерполируемой функции до определённого предела сходимость эта улучшается. В целом в задаче интерполирования полиномиальные сплайны оказываются во многих ситуациях лучше алгебраических полиномов как с точки зрения вычислительных удобств, так и с точки зрения качества приближения, обеспечивая минимально возможную погрешность для заданного размера сетки. Подробности заинтересованный читатель может увидеть, к примеру, в книге [55].

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).

Интерполирование сплайнами 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 [42]. Основной недостаток понятий насыщаемости / ненасыщаемости состоит в трудности практического определения гладкости данных, которые присутствуют в решаемой задаче.

2.6г Экстремальное свойство кубических сплайнов

Сплаины $S(x)$, удовлетворяющие на концах рассматриваемого интервала $[a, b]$ дополнительным условиям

$$S''(a) = S''(b) = 0, \quad (2.67)$$

называются *естественными* или *натуральными сплайнами*.

Одной из важных характеристик кривой является её *кривизна* — скорость изгибания 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

$$\frac{f''(x)}{(1 + (f'(x))^2)^{3/2}} \quad (2.68)$$

(см. подробности в [38, 73]). 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

$$\mathcal{E}(f) = \int_a^b (f''(x))^2 dx. \quad (2.69)$$

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 [88] или, к примеру, в книгах [2, 11, 36, 49]. 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 [40, 49].

Интеграл (2.69) приближённо пропорционален энергии деформации гибкой упругой линейки, форма которой описывается функцией $f(x)$ на интервале $[a, b]$. Краевые условия (2.67) соответствуют при этом линейке, свободно закреплённой на концах, где не приложены моменты каких-либо внешних сил.

В самом деле, потенциальная энергия изгибания малого участка упругого тела, как известно, пропорциональна квадрату его кривизны в данной точке. Поэтому в силу (2.68) энергия упругой деформации однородной линейки, принимающей форму кривой $y = f(x)$ на интервале $[a, b]$, выражается интегралом

$$\int_a^b \frac{\kappa}{(1 + (f'(x))^2)^3} (f''(x))^2 dx, \quad (2.70)$$

где κ — коэффициент, характеризующий свойства материала линейки. При условии приблизительного постоянства $f'(x)$ значения интеграла (2.70) пропорциональны значениям (2.69). 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.¹⁸ Оно служит началом большого и интересного направления теории, в котором сплайны вводятся и рассматриваются как решения некоторых задач на минимум. 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, задаваемые на различных участках своей области определения разными аналитическими формулами. Вариационный подход к сплайнам показывает, что это необходимо по существу дела.

2.7 Нелинейные методы интерполяции

Рассмотренные нами 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 интерполирования, для которых сформулированное выше

¹⁸Иногда также говорят о *вариационном свойстве* естественных сплайнов.

свойство не выполнено. Эти методы тоже широко применяются при практической интерполяции, так как 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

$$y = y(x) = \frac{a_0 + a_1x + a_2x^2 + \dots}{b_0 + b_1x + b_2x^2 + \dots}. \quad (2.71)$$

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 (2.71), 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 так называемыми полюсами. Этим термином в теории функций называют точки, где функция принимает бесконечно большие значения. Подобные полюса могут присутствовать у рассматриваемой 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 интерполирования 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. Хорошая полиномиальная интерполяция возможна 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 μ полинома числителя и степени ν полинома знаменателя в дроби (2.71). 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 (2.71) 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:

$$\begin{cases} \sum_{j=0}^{\mu} x_i^j a_j - \sum_{j=0}^{\nu} x_i^j y_i b_j = 0, \\ i = 0, 1, \dots, n. \end{cases} \quad (2.72)$$

Решение этой системы уравнений даёт искомым рациональный интерполянт.

Несмотря на технологическую простоту описанного решения задачи рациональной интерполяции, существование нетривиального решения, получаемого с его помощью, гарантировать нельзя. Это вызвано возможным занулением знаменателя дроби (2.71) и трудностью исследования системы линейных алгебраических уравнений (2.72), свойства которой, вообще говоря, могут быть плохими.

Пример 2.7.1 Построим рациональную интерполяцию данных

x	\parallel	-1	0	1
y	\parallel	1	0	1

Они принимаются функцией $y = |x|$.

Три узла соответствуют $n = 2$, что должно быть равно сумме степеней числителя и знаменателя интерполанта. Станем искать его в виде

$$y = \frac{a_0 + x}{b_0 + b_1 x},$$

т. е. зафиксировав значение $a_1 = 1$. Отсюда

$$a_0 + x = b_0 y + b_1 x y. \quad (2.73)$$

Подставляя в (2.73) 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 интерполантом является функция

$$y = \frac{x}{x}.$$

Оно непригодно в качестве решения задачи, так как при $x = 0$ не определено, и 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 интерполант, 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 (2.72), 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 искомым рациональный интерполянт имеет вид

$$g(x) = \frac{10 + 5x + x^2}{10 - 2x}. \quad (2.74)$$

Полученный интерполянт настолько хорош, что практически сливается 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 (2.74). ■

Опишем ещё один способ решения задачи рациональной интерполяции, эквивалентный изложенному выше, но, возможно, более удобный в некоторых ситуациях (см. [65]). Представление (2.71) равносильно тождеству

$$a_0 - b_0y + a_1x - b_1xy + a_2x^2 - b_2x^2y + \dots = 0. \quad (2.75)$$

Коль скоро при $x = x_i$ должно быть $y = y_i$, $i = 0, 1, \dots, n$, то получаем ещё $(n + 1)$ числовых равенств

$$a_0 - b_0y_i + a_1x_i - b_1x_iy_i + a_2x_i^2 - b_2x_i^2y_i + \dots = 0, \quad (2.76)$$

$i = 0, 1, \dots, n$. Соотношения (2.75)–(2.76) можно трактовать, как условие линейной зависимости, с коэффициентами $a_0, -b_0, a_1, -b_1, \dots$, для

вектор-столбцов

$$\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} y \\ y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \begin{pmatrix} x \\ x_0 \\ x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \begin{pmatrix} xy \\ x_0y_0 \\ x_1y_1 \\ \vdots \\ x_ny_n \end{pmatrix}, \quad \begin{pmatrix} x^2 \\ x_0^2 \\ x_1^2 \\ \vdots \\ x_n^2 \end{pmatrix}, \quad \begin{pmatrix} x^2y \\ x_0^2y_0 \\ x_1^2y_1 \\ \vdots \\ x_n^2y_n \end{pmatrix}, \quad \dots$$

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 числа. Затем разрешим полученное равенство нулю относительно переменной y , которая входит во все слагаемые в степени не выше первой, и мы 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 §3.13 мы рассмотрим соответствующие численные методы. 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 интересующийся читатель может найти, к примеру, в книге [95], §2.2.

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 (см. §§4.4г и 4.5б).

В настоящее время наиболее распространены три следующих способа вычисления производных:

- символьное (аналитическое) дифференцирование,
- численное дифференцирование,
- алгоритмическое (автоматическое) дифференцирование.

Символьным (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* оперируют 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-своему. 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 процедура определения значений этой функции не может быть выписана в виде выражения или детерминированной программы, 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 интерполирования, которой посвящены предшествующие параграфы, оказывается в высшей степени полезной при реализации такого подхода. Именно, 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 мы строим по двум соседним узлам сетки, т.е. по x_{i-1} и x_i , $i = 1, 2, \dots, n$:

$$\begin{aligned} P_{1,i}(x) &= \frac{x - x_i}{x_{i-1} - x_i} f_{i-1} + \frac{x - x_{i-1}}{x_i - x_{i-1}} f_i \\ &= \frac{f_i - f_{i-1}}{x_i - x_{i-1}} x + \frac{f_{i-1}x_i - f_i x_{i-1}}{x_i - x_{i-1}}, \end{aligned}$$

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

$$f'(x_i) \approx f_{\bar{x},i} := \frac{f_i - f_{i-1}}{x_i - x_{i-1}} = \frac{f_i - f_{i-1}}{h_i} \quad (2.77)$$

— разделённая *разность назад*,

$$f'(x_i) \approx f_{x,i} := \frac{f_{i+1} - f_i}{x_{i+1} - x_i} = \frac{f_{i+1} - f_i}{h_{i+1}} \quad (2.78)$$

— 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

$$\begin{aligned}
 P_{2,i}(x) &= \frac{(x - x_i)(x - x_{i+1})}{(x_{i-1} - x_i)(x_{i-1} - x_{i+1})} f_{i-1} + \frac{(x - x_{i-1})(x - x_{i+1})}{(x_i - x_{i-1})(x_i - x_{i+1})} f_i \\
 &\quad + \frac{(x - x_{i-1})(x - x_i)}{(x_{i+1} - x_{i-1})(x_{i+1} - x_i)} f_{i+1} \\
 &= \frac{x^2 - (x_i + x_{i+1})x + x_i x_{i+1}}{(x_{i-1} - x_i)(x_{i-1} - x_{i+1})} f_{i-1} \\
 &\quad + \frac{x^2 - (x_{i-1} + x_{i+1})x + x_{i-1} x_{i+1}}{(x_i - x_{i-1})(x_i - x_{i+1})} f_i \\
 &\quad + \frac{x^2 - (x_{i-1} + x_i)x + x_{i-1} x_i}{(x_{i+1} - x_{i-1})(x_{i+1} - x_i)} f_{i+1}.
 \end{aligned}$$

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

$$\begin{aligned}
 f'(x) \approx P'_{2,i}(x) &= \frac{2x - x_i - x_{i+1}}{h_i(h_i + h_{i+1})} f_{i-1} \\
 &\quad - \frac{2x - x_{i-1} - x_{i+1}}{h_i h_{i+1}} f_i + \frac{2x - x_{i-1} - x_i}{h_{i+1}(h_i + h_{i+1})} f_{i+1}.
 \end{aligned} \tag{2.79}$$

Формула (2.79) 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 наиболее 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 (2.79) получаем для первой производной формулу

$$f'(x_i) \approx f_{\bar{x},i} = \frac{f_{i+1} - f_{i-1}}{2h}, \tag{2.80}$$

называемую *формулой центральной разности*. Подставляя в (2.79) 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 (2.79) аргумент $x = x_{i+1}$ и сдвигая в получающемся результате индекс на (-1) , получим

$$f'(x_i) \approx \frac{f_{i-2} - 4f_{i-1} + 3f_i}{2h}.$$

Займёмся 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

$$f''(x_i) \approx \frac{f_{i-1} - 2f_i + f_{i+1}}{h^2}. \quad (2.81)$$

Эта формула широко используется в вычислительной математике, и по аналогии с (2.77)–(2.78) 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 —

$$f'(x_i) \approx \frac{1}{6h}(-11f_i + 18f_{i+1} - 9f_{i+2} + 2f_{i+3}), \quad (2.82)$$

$$f'(x_i) \approx \frac{1}{6h}(-2f_{i-1} - 3f_i + 6f_{i+1} - f_{i+2}), \quad (2.83)$$

$$f'(x_i) \approx \frac{1}{6h}(f_{i-2} - 6f_{i-1} + 3f_i + 2f_{i+1}), \quad (2.84)$$

$$f'(x_i) \approx \frac{1}{6h}(-2f_{i-3} + 9f_{i-2} - 18f_{i-1} + 11f_i), \quad (2.85)$$

для второй производной —

$$f''(x_i) \approx \frac{1}{h^2}(2f_i - 5f_{i+1} + 4f_{i+2} - f_{i+3}), \quad (2.86)$$

$$f''(x_i) \approx \frac{1}{h^2}(f_{i-1} - 2f_i + f_{i+1}), \quad (2.87)$$

$$f''(x_i) \approx \frac{1}{h^2}(-f_{i-3} + 4f_{i-2} - 5f_{i-1} + 2f_i). \quad (2.88)$$

В формуле (2.87) один из четырёх узлов, по которым строилась формула, никак не используется, а сама формула совпадает с формулой (2.81), полученной по трём точкам. Отметим красивую двойственность формул (2.82) и (2.85), (2.83) и (2.84), (2.86) и (2.88). 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.

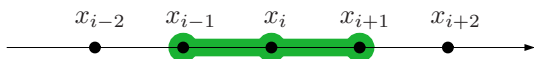


Рис. 2.12. Шаблон формулы второй разностной производной (2.81).

В связи с численным дифференцированием и во многих других вопросах вычислительной математики чрезвычайно полезно понятие шаблона (сеточной) формулы, под которым мы 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 (2.81) 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} (см. Рис. 2.12), в которых должны быть заданы f_{i-1} , f_i , f_{i+1} . Особенно разнообразны формы шаблонов в случае двух и более независимых переменных.

2.86 Оценка погрешности численного дифференцирования

Пусть для численного нахождения k -ой производной функции применяется формула численного дифференцирования Φ , имеющая шаблон Θ 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 Φ ? 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 нами ранее выражений (2.27) и (2.28) для погрешности интерполирования. Этот путь оказывается очень непростым, так как применение, к примеру, выражения (2.28) требует достаточной гладкости функции $\xi(x)$, о которой мы можем сказать немного. 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 (2.87), построенная по четырём точкам на основе полинома 3-й степени. Тем не менее, явные выражения для остаточного члена формул численного дифференцирования на этом пути можно получить методом, который напоминает вывод фор-

мулы для погрешности алгебраического интерполирования. Подробности изложены, к примеру, в книгах [20, 28].

Рассмотрим ниже детально более простой и достаточно универсальный способ оценивания погрешностей, основанный на разложениях по формуле Тейлора. Суть этого способа 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» (2.80):

$$f'(x_i) \approx f_{\bar{x},i} = \frac{f_{i+1} - f_{i-1}}{2h}.$$

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 (2.80) и разлагая относительно точки x_i по формуле Тейлора с остаточным членом в форме Лагранжа вплоть до членов второго порядка, получим

$$\begin{aligned} f_{\bar{x},i} &= \frac{1}{2h} \left(\left(f(x_i) + hf'(x_i) + \frac{h^2}{2} f''(x_i) + \frac{h^3}{6} f'''(\xi_+) \right) \right. \\ &\quad \left. - \left(f(x_i) - hf'(x_i) + \frac{h^2}{2} f''(x_i) - \frac{h^3}{6} f'''(\xi_-) \right) \right) \\ &= f'(x_i) + \frac{h^2}{12} f'''(\xi_+) + \frac{h^2}{12} f'''(\xi_-), \end{aligned}$$

где ξ_+ и ξ_- — некоторые точки из открытого интервала $]x_{i-1}, x_{i+1}[$. Поэтому

$$f_{\bar{x},i} - f'(x_i) = \frac{h^2}{12} (f'''(\xi_+) + f'''(\xi_-)) = \frac{\alpha h^2}{6},$$

где $\alpha = \frac{1}{2}(f'''(\xi_+) + f'''(\xi_-))$. В целом справедлива оценка

$$|f_{\bar{x},i} - f'(x_i)| \leq \frac{M_3}{6} h^2,$$

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 .

Определение 2.8.1 *Станем говорить, что приближённая формула (численного дифференцирования, интегрирования и т. п.) или же приближённый численный метод имеют p -ый 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.

Другое необходимое замечание состоит в том, что понятие порядка формулы или метода основывается на сравнении скорости убывания погрешности со скоростью убывания степенных функций $1, h, h^2, \dots, h^p, \dots$, то есть 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 в себя узлы $0, \pm h, \pm 2h$ и т. д. Для функции $y = g(x)$ рассмотрим интерполяцию значения $g(0)$ полусуммой

$$\frac{1}{2}(g(-h) + g(h)), \quad (2.89)$$

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 $g(x)$?¹⁹

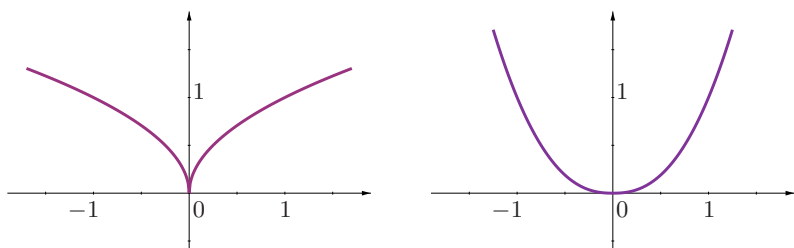


Рис. 2.13. Графики функции $y = |x|^\alpha$ при $0 < \alpha < 1$ и $\alpha > 1$.

Для функции $g(x) = |x|^\alpha$, $\alpha > 0$, погрешность интерполяции будет, очевидно, равна h^α , так что её порядок равен α . Он может быть нецелым числом (в частности, дробным) и даже сколь угодно малым.

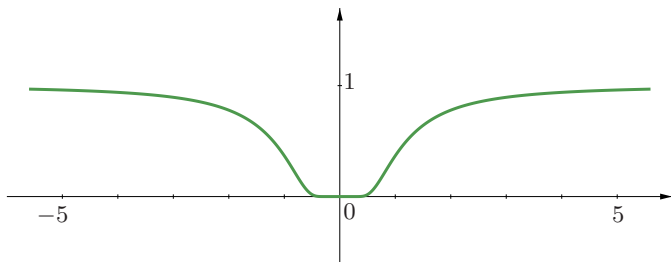


Рис. 2.14. График функции $y = \exp(-1/x^2)$.

Возьмём в качестве $g(x)$ функцию

$$g(x) = \begin{cases} \exp\left(-\frac{1}{x^2}\right), & \text{при } x \neq 0, \\ 0, & \text{при } x = 0, \end{cases}$$

известную в математическом анализе как пример бесконечно гладкой, но не аналитической (т. е. не разлагающейся в степенной ряд) функции. Погрешность интерполяции значения этой функции в нуле с помощью формулы (2.89) равна $\exp(-1/h^2)$, при $h \rightarrow 0$ она убывает быстрее любой степени h , так что порядок точности нашей интерполяции оказывается бесконечно большим. Но такой же бесконечно большой порядок точности интерполирования будет демонстрировать здесь функция

¹⁹Идея этого примера заимствована из пособия [51], задача 4.2.

$y = x^2 g(x)$, 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

$$|f_{x,i} - f'(x_i)| \leq \frac{M_2}{2} h, \quad |f_{\bar{x},i} - f'(x_i)| \leq \frac{M_2}{2} h, \quad (2.90)$$

где $M_2 = \max_{\xi} |f''(\xi)|$ по ξ из соответствующего интервала между узлами. Таким образом, разность вперёд (2.77) и разность назад (2.77) имеют всего лишь первый порядок точности. Отметим, что для дважды непрерывно дифференцируемых функций оценки (2.90) уже не могут быть улучшены и достигаются, к примеру, на функции $f(x) = x^2$.

Конспективно изложим другие результаты о точности формул численного 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 (2.81) 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

$$\begin{aligned} f_{x\bar{x},i} &= \frac{1}{h^2} \left(\left(f_i - hf'_i + \frac{h^2}{2} f''_i - \frac{h^3}{6} f'''_i + \frac{h^4}{24} f^{(4)}(\xi_-) \right) - 2f_i \right. \\ &\quad \left. + \left(f_i + hf'_i + \frac{h^2}{2} f''_i + \frac{h^3}{6} f'''_i + \frac{h^4}{24} f^{(4)}(\xi_+) \right) \right) \\ &= f''_i + \frac{h^2}{24} (f^{(4)}(\xi_-) + f^{(4)}(\xi_+)), \end{aligned}$$

где ξ_- , ξ_+ — 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 из C^4 .

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 (2.81) 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 разложений установить не сможем. Тот факт, что в этих условиях реальный порядок точности может быть в самом деле меньшим, чем для функций с высокой гладкостью, показывает следующий

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. 2.15.

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

$$g'(0) = \lim_{x \rightarrow 0} \frac{x|x|}{x} = 0.$$

Таким образом, 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.

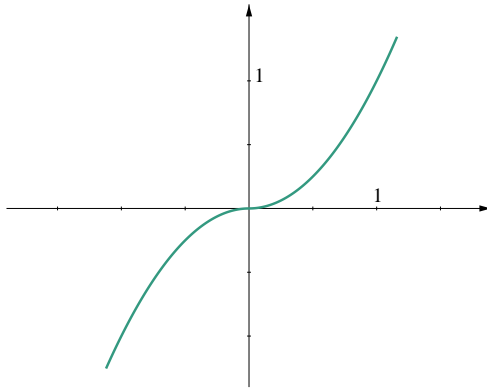


Рис. 2.15. График функции $y = x|x|$: увидеть разрыв её второй производной в нуле почти невозможно.

Воспользуемся для численного нахождения производной $g'(0)$ формулой центральной разности (2.80) на шаблоне с шагом h , симметричным 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. ■

2.8в 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.

Предположим, что задан шаблон из $p + 1$ штук точек x_0, x_1, \dots, x_p . Станем искать приближённое выражение для производной от функции в виде линейной формы от значений функции, т. е. как

$$f^{(k)}(x) \approx \sum_{i=0}^p c_i f(x_i). \quad (2.91)$$

Она мотивируется тем обстоятельством, что дифференцирование любого порядка является операций, линейной по значениям функции. Линейными формами от значений функции были, в частности, все полученные ранее формулы численного дифференцирования, начиная с (2.77) и кончая (2.88).

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 полиномы степени не выше заданной, либо тригонометрические полиномы (2.4) какого-то фиксированного порядка и т. п. Рассмотрим ниже подробно случай алгебраических полиномов.

Возьмём $f(x)$ равной последовательным степеням переменной x , т. е. $1, x, x^2, \dots, x^q$ для некоторого фиксированного q . Если формула (2.91) 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 для неизвестных коэффициентов c_i , и в целом мы приходим к системе линейных уравнений относительно c_i , $i = 0, 1, \dots, p$. Для разрешимости этой системы естественно взять число неизвестных равным числу уравнений, т. е. $q = p$.

Получающаяся система линейных уравнений имеет вид

$$\left\{ \begin{array}{lcl} c_0 & + & c_1 + \dots + c_p = 0, \\ c_0 x_0 & + & c_1 x_1 + \dots + c_p x_p = 0, \\ \vdots & & \ddots \quad \vdots \quad \vdots \\ c_0 x_0^{k-1} & + & c_1 x_1^{k-1} + \dots + c_p x_p^{k-1} = 0, \\ c_0 x_0^k & + & c_1 x_1^k + \dots + c_p x_p^k = k!, \\ c_0 x_0^{k+1} & + & c_1 x_1^{k+1} + \dots + c_p x_p^{k+1} = (k+1)!x, \\ \vdots & & \ddots \quad \vdots \quad \vdots \\ c_0 x_0^p & + & c_1 x_1^p + \dots + c_p x_p^p = p(p-1) \cdots (p-k+1) x^{p-k}. \end{array} \right. \quad (2.92)$$

В правых частях этой системы стоят k -е производные от $1, x, x^2, \dots, x^q$, а матрицей системы является матрица Вандермонда вида (2.7), которая неособенна для несовпадающих узлов x_0, x_1, \dots, x_p . При этом система линейных уравнений 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 (2.92) 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. §3.5б). Но на практике решение системы (2.92) — вручную или на компьютере — обычно не приводит к большим ошибкам, так как порядок системы (2.92), равный порядку производной, бывает, как правило, небольшим.²⁰

Пример 2.8.3 Построим формулу численного дифференцирования

Интересен вопрос о взаимоотношении метода неопределённых коэффициентов и рассмотренного ранее в §2.8а интерполяционного подхода к численному дифференцированию. К примеру, Ш.Е. Микеладзе

²⁰На стр. 111 мы уже обсуждали вопрос о том, каков наивысший порядок производных, всё ещё имеющих содержательный смысл.

в книге [64] утверждает, 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.

2.8г 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 δ 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:

$$\begin{aligned}
 E(h, \delta) &= \left| \frac{\tilde{f}_{i+1} - \tilde{f}_i}{h} - f'(x_i) \right| \\
 &\leq \left| \frac{\tilde{f}_{i+1} - \tilde{f}_i}{h} - \frac{f_{i+1} - f_i}{h} \right| + \left| \frac{f_{i+1} - f_i}{h} - f'(x_i) \right| \\
 &\leq \left| \frac{(\tilde{f}_{i+1} - f_{i+1}) + (f_i - \tilde{f}_i)}{h} \right| + \frac{M_2 h}{2} \\
 &\leq \frac{|f_{i+1} - \tilde{f}_{i+1}| + |f_i - \tilde{f}_i|}{h} + \frac{M_2 h}{2} = \frac{2\delta}{h} + \frac{M_2 h}{2}.
 \end{aligned} \tag{2.93}$$

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 (2.90) 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 δ 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

$$h^* = 2\sqrt{\delta/M_2}, \tag{2.94}$$

и брать меньший шаг численного дифференцирования смысла нет. Само 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. §1.3), 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 (2.94) имеем $h^* = 2\sqrt{\delta/M_2} = 2 \cdot 10^{-8}$, т. е. брать шаг сетки меньше 10^{-8} смысла не имеет. ■

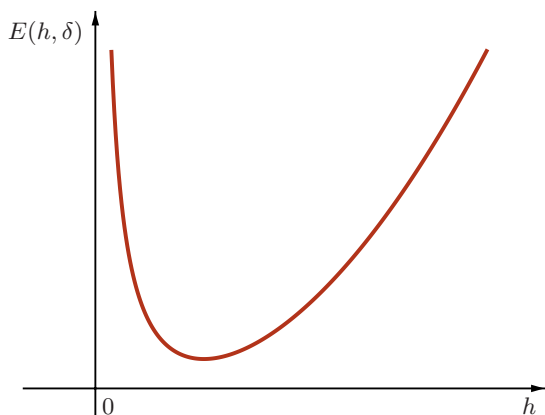


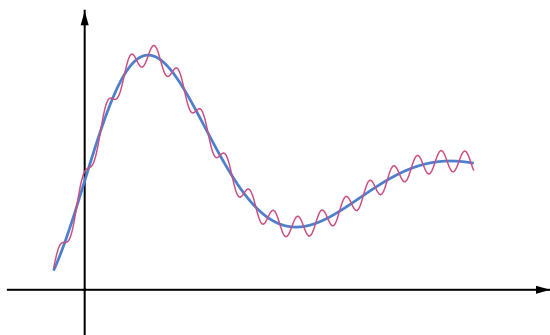
Рис. 2.16. Типичный график полной погрешности численного дифференцирования

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 вида²¹

$$f^{(k)}(x) = h^{-k} \sum_i c_i f(x_i) + R_k(f, x), \quad (2.95)$$

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 оста-

²¹Для примера можно взглянуть на те формулы, которые приведены в §2.8а.

Рис. 2.17. Возмущение функции добавкой $\frac{1}{n} \sin(nx)$.

точный 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 δ , то в правой части (2.95) возникает ещё член, абсолютная величина которого совершенно аналогично (2.93) 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. 2.16.

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. §1.6). 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 $f'(x)$ (см. Рис. 2.17). При возмущении исходной 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:

$$(u + v)' = u' + v', \quad (2.96)$$

$$(u - v)' = u' - v', \quad (2.97)$$

$$(uv)' = u'v + uv', \quad (2.98)$$

$$\left(\frac{u}{v}\right)' = \frac{u'v - uv'}{v^2}. \quad (2.99)$$

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 формул (2.96)–(2.99):

$$(u, u') + (v, v') = (u + v, u' + v'), \quad (2.100)$$

$$(u, u') - (v, v') = (u - v, u' - v'), \quad (2.101)$$

$$(u, u') \cdot (v, v') = (uv, u'v + uv'), \quad (2.102)$$

$$\left(\frac{u}{v}, \frac{u'v - uv'}{v^2}\right) = \left(\frac{u}{v}, \frac{u'v - uv'}{v^2}\right). \quad (2.103)$$

Первые члены пар преобразуются просто в соответствии с применяемой арифметической операцией, а операции над вторыми членами пар

— это в точности копии правил (2.96)–(2.99). Если для заданного выражения мы начнём вычисления по выписанным формулам (2.100)–(2.103), заменив исходную переменную x на пары $(x, 1)$, а константы c — на пары вида $(c, 0)$, то на выходе получим пару, состоящую из численных значений выражения и производной от него в точке x .

Это рассуждение очевидно 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

$$\exp((u, u')) = (\exp u, u' \exp u),$$

$$\sin((u, u')) = (\sin u, u' \cos u),$$

$$((u, u'))^2 = (u^2, 2uu'),$$

$$((u, u'))^3 = (u^3, 3u^2u') \text{ и т.д.}$$

Арифметику пар вида (u, u') с операциями (2.100)–(2.103) 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* [83], которое, в свою очередь, является простым частным случаем так называемых алгебр Клиффорда.

Строго говоря, мы рассмотрели один из возможных способов организации алгоритмического дифференцирования, который называют *прямым режимом*. Существует ещё и *обратный режим* алгоритмического дифференцирования.

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.

Идея алгоритмического дифференцирования может быть распространена на вычисление разделённых разностей (наклонов) функций

(см., к примеру, [92]), а также на вычисление интервальных расширений производных и наклонов [80].

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

$$u^{\angle} = \frac{u - \tilde{u}}{x - \tilde{x}}, \quad v^{\angle} = \frac{v - \tilde{v}}{x - \tilde{x}}.$$

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.

Арифметика наклонов, в отличие от дифференциальной, — это арифметика упорядоченных троек вида $(\tilde{u}, u, u^{\angle})$, а не пар. Они образованы двумя значениями функции, между аргументами которых берётся наклон, а также самим значением наклона.

Расчётные формулы арифметики наклонов, вытекающие из резуль-

татов предшествующих выкладок, имеют следующий вид

$$\begin{aligned}(\tilde{u}, u, u^\epsilon) + (\tilde{v}, v, v^\epsilon) &= (\tilde{u} + \tilde{v}, u + v, u^\epsilon + v^\epsilon), \\(\tilde{u}, u, u^\epsilon) - (\tilde{v}, v, v^\epsilon) &= (\tilde{u} - \tilde{v}, u - v, u^\epsilon - v^\epsilon), \\(\tilde{u}, u, u^\epsilon) \cdot (\tilde{v}, v, v^\epsilon) &= (\tilde{u}\tilde{v}, uv, u^\epsilon v + \tilde{u}v^\epsilon), \\(\tilde{u}, u, u^\epsilon) / (\tilde{v}, v, v^\epsilon) &= \left(\frac{\tilde{u}}{\tilde{v}}, \frac{u}{v}, \frac{u^\epsilon v - uv^\epsilon}{v\tilde{v}} \right).\end{aligned}$$

Расчётные формулы для умножения и деления могут иметь альтернативные варианты.

2.10 Приближение функций

2.10a Обсуждение постановки задачи

В этом разделе мы более подробно займёмся задачей приближения функций, постановка которой была предварительно рассмотрена в начале главы.

К задаче приближения функций естественно приходят в ситуациях, где методы интерполирования по различным причинам не удовлетворяют практику. Эти причины могут носить чисто технический характер. К примеру, гладкость интерполяционного сплайна может оказаться недостаточной, либо его построение — слишком сложным. Степень интерполяционного алгебраического полинома может быть неприемлемо высокой для данного набора узлов интерполяции (а высокая степень — это трудности при построении полинома и при работе с ним). 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. 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. $g(x_i) \in [\underline{f}_i, \overline{f}_i]$, $i = 0, 1, \dots, n$, $\underline{f}_i \leq \overline{f}_i$. 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» $[\underline{f}_i, \overline{f}_i]$; см. Рис. 2.18.

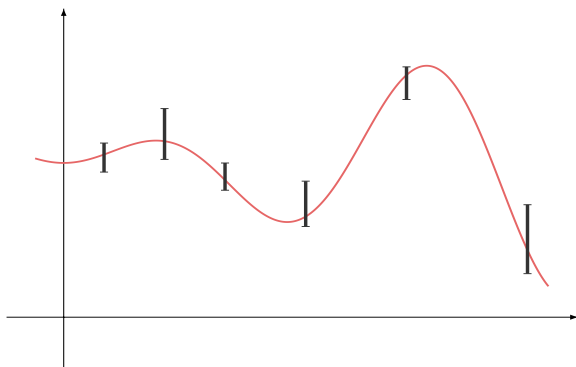


Рис. 2.18. Интерполяция функции, заданной с погрешностью (интервальная интерполяция)

Более общая постановка задачи предусматривает наличие некоторой метрики (расстояния), с помощью которой можно измерять отклонение вектора значений $(g(x_0), g(x_1), \dots, g(x_n))^T$ функции $g(x)$ в узлах сетки от вектора заданных значений $(f_0, f_1, \dots, f_n)^T$. 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 $\mathbf{f} = (f_0, f_1, \dots, f_n)^T$ и $\mathbf{g} = (g(x_0), g(x_1), \dots, g(x_n))^T$ не больше ϵ .

При этом $g(x)$ называют *приближающей* (аппроксимирующей) функцией. Важнейшей модификацией поставленной задачи служит *задача наилучшего приближения*, в которой величина ϵ не фиксируется и ищут приближающую (аппроксимирующую) функцию $g(x)$, которая доставляет минимум расстоянию между векторами \mathbf{f} и \mathbf{g} .

Согласно классификации §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 опре-

деления:

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, их мера взаимной близости, в приведённых выше формулировках могут быть самыми разнообразными, 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 dist (см. определение на стр. 56). Задача наилучшего приближения функций получает следующую развёрнутую постановку:

Для заданных функции $f(x)$ из класса функций \mathcal{F} и метрики dist найти функцию $g(x)$ из класса $\mathcal{G} \subset \mathcal{F}$, на которой достигается нижняя грань расстояний от $f(x)$ до функций из \mathcal{G} , то есть, для которой выполнено условие $\text{dist}(f, g) = \inf_{h \in \mathcal{G}} \text{dist}(f, h)$. (2.104)

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*, поскольку получаемая приближающая функция действительно «сглаживает» выбросы данных, вызванные случайными ошибками и т. п.

До сих пор ничего не было сказано о выборе классов функций \mathcal{F} и \mathcal{G} , и в наших формулировках они могут быть весьма произвольными. Помимо условия $\mathcal{F} \supset \mathcal{G}$ часто наделяют \mathcal{F} и \mathcal{G} структурой линейного векторного пространства с некоторой нормой $\|\cdot\|$. Именно в ней измеряют отклонение функций (непрерывного или дискретного аргумента) друг от друга, так что

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

Соответственно, в задаче наилучшего приближения функции f ищется такая функция $g \in \mathcal{G}$, на которой достигается $\inf_{h \in \mathcal{G}} \|f - h\|$.

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 §§2.10в–2.10ж, и теория равномерного приближения, когда отклонение функций оценивается в равномерной (чебышёвской) норме $\|f\| = \max_{x \in [a, b]} |f(x)|$ (см. [54, 74]). 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, к наиболее выгодному выбору класса приближающих функций и др.

2.10б Существование и единственность решения задачи 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. Таким образом, вместо классов функций в результатах этого раздела будет фигурировать линейное нормированное пространство и его подпространство, в котором мы выбираем приближение.

Теорема 2.10.1 (теорема Э. Бореля) *Пусть X — нормированное линейное пространство, U — его конечномерное линейное подпространство. Тогда для любого $f \in X$ существует его наилучшее приближение $u \in U$.*

Э. Борель доказал этот важный результат для случая равномерного (чебышёвского) приближения функций полиномами в книге [85], когда понятие нормированного линейного пространства ещё только зарождалось. Но и само утверждение, и метод его доказательства почти до-

словно обобщаются для общих нормированных пространств, что было сделано математиками 30-х годов XX века.

Доказательство. Пусть размерность подпространства U равна m . Зафиксируем в U некоторый базис $\{\phi_1, \phi_2, \dots, \phi_m\}$ и введём функцию $\mathcal{D} : \mathbb{R}^m \rightarrow \mathbb{R}_+$, задаваемую как

$$\mathcal{D}(a_1, a_2, \dots, a_m) = \left\| f - \sum_{j=1}^m a_j \phi_j \right\|,$$

где $\|\cdot\|$ — норма в X . Функция \mathcal{D} имеет своими значениями расстояние от элемента f до вектора из подпространства U , который определяется набором коэффициентов разложения a_1, a_2, \dots, a_m 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. Так как $\|x\| - \|y\| \leq \|x - y\|$ для любых векторов x и y , то имеем

$$\begin{aligned} & \left| \mathcal{D}(a_1, a_2, \dots, a_m) - \mathcal{D}(\tilde{a}_1, \tilde{a}_2, \dots, \tilde{a}_m) \right| \\ &= \left| \left\| f - \sum_{j=1}^m a_j \phi_j \right\| - \left\| f - \sum_{j=1}^m \tilde{a}_j \phi_j \right\| \right| \\ &\leq \left\| \left(f - \sum_{j=1}^m a_j \phi_j \right) - \left(f - \sum_{j=1}^m \tilde{a}_j \phi_j \right) \right\| \\ &\leq \left\| \sum_{j=1}^m (a_j - \tilde{a}_j) \phi_j \right\| \leq \sum_{j=1}^m |a_j - \tilde{a}_j| \|\phi_j\| \\ &\leq \max_{1 \leq j \leq m} |a_j - \tilde{a}_j| \cdot \sum_{j=1}^m \|\phi_j\|. \end{aligned}$$

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.

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 .

Возьмём какую-либо точку \tilde{u} из U . 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. 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

$$\left\| \sum_{j=1}^m a_j \phi_j \right\| \leq \|f - \tilde{u}\| + \|f\|. \quad (2.105)$$

Далее, зафиксируем в арифметическом пространстве \mathbb{R}^n векторов коэффициентов 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 на компактной единичной сфере пространства \mathbb{R}^m относительно нормы $\|\cdot\|'$, и, кроме того, она не равна нулю, если $\phi_1, \phi_2, \dots, \phi_m$ образуют базис в U . Как следствие, из (2.105) заключаем, что

$$\|a\| \leq \frac{1}{C} (\|f - \tilde{u}\| + \|f\|),$$

т. е. множество всех a из V в самом деле ограничено в \mathbb{R}^m .

Итак, V компактно в конечномерном пространстве \mathbb{R}^m , и потому 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 . ■

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 векторное пространство X называют строго нормированным, если в неравенстве треугольника в этом пространстве равенство достигается только на положительно пропорциональных элементах, т. е. если для произвольных $x, y \in X$ из равенства $\|x + y\| = \|x\| + \|y\|$ следует существование такого скаляра $\alpha \in \mathbb{R}_+$, что $y = \alpha x$.*

Важнейшим примером строго нормированных пространств являются линейные векторные пространства со скалярным произведением. Такие пространства часто называют *предгильбертовыми*, а если их размерность конечна, то *евклидовыми*. Если $\langle \cdot, \cdot \rangle$ — скалярное произведение, то норма задаётся как

$$\|x\| := \sqrt{\langle x, x \rangle},$$

и относительно неё линейное пространство является строго нормированным.

Действительно, для скалярного произведения в линейном пространстве справедливо *неравенство Коши-Буняковского* [12, 16, 35]

$$|\langle x, y \rangle| \leq \|x\| \|y\|.$$

Из него следует, что

$$\begin{aligned} \|x + y\|^2 &= \langle x + y, x + y \rangle \\ &= \|x\|^2 + 2\langle x, y \rangle + \|y\|^2 \\ &\leq \|x\|^2 + 2\|x\| \|y\| + \|y\|^2 = (\|x\| + \|y\|)^2. \end{aligned}$$

Так как все сравниваемые в выписанной цепочке выражения неотрицательны, то можем заключить, что

$$\|x + y\| \leq \|x\| + \|y\|$$

— выполняется и неравенство треугольника. Равенство в нём имеет место в том и лишь в том случае, когда оно выполнено в неравенстве Коши-Буняковского, т. е. в случае коллинеарности векторов x и y .

Пример 2.10.1 В арифметическом пространстве \mathbb{R}^n рассмотрим так называемую p -норму, обозначаемую $\|\cdot\|_p$ и задаваемую как

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

При $p > 1$ это в самом деле норма n -векторов, так как для неё выполнены все аксиомы нормы (см. §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, 35, 38, 41]). 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 $p > 1$.

В частности, строго нормировано пространство \mathbb{R}^n с 2-нормой, которая называется также *евклидовой нормой* и порождается стандартным скалярным произведением в \mathbb{R}^n (см. §3.3а). ■

Пример 2.10.2 В арифметическом пространстве \mathbb{R}^n рассмотрим нормы

$$\|x\|_1 = |x_1| + \dots + |x_n| \quad \text{и} \quad \|x\|_\infty = \max_{1 \leq i \leq n} |x_i|,$$

которые эквивалентны p -норме (см. §3.3б). Но они не делают \mathbb{R}^n строго нормированным пространством.

Возьмём, к примеру, $x = (0, \dots, 0, 1)^\top$ и $y = (1, \dots, 1, 1)^\top$. Векторы x и y , очевидно, неколлинеарны, но $\|x + y\|_\infty = \|x\|_\infty + \|y\|_\infty$. Аналогичный контрпример нетрудно построить и для нормы $\|\cdot\|_1$. ■

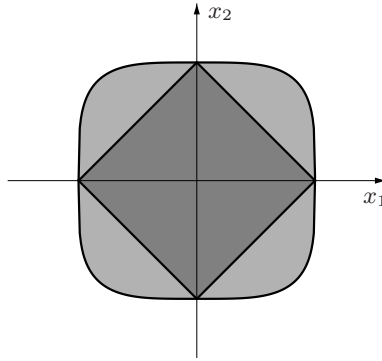


Рис. 2.19. Единичные шары 1-нормы (тёмного тона) и 4-нормы (светлого тона) в пространстве \mathbb{R}^2 .

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 μ — 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

$$\begin{aligned} \|f - \tfrac{1}{2}(u' + u'')\| &= \|\tfrac{1}{2}(f - u') + \tfrac{1}{2}(f - u'')\| \\ &\leq \tfrac{1}{2}\|f - u'\| + \tfrac{1}{2}\|f - u''\| = \mu. \end{aligned}$$

Строгого неравенства в выписанной цепочке отношений быть не может, так как оно означало бы существование элемента, приближающего f лучше, чем два наилучших 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

$$f - u' = \alpha(f - u'') \quad (2.106)$$

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 (2.106) 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 . ■

Дальнейшие результаты о существовании наилучших приближений в различных подмножествах линейных векторных пространств можно найти, например, в книге [62].

2.10в Приближение в евклидовом подпространстве

Рассмотрим подробно важный частный случай задачи о наилучшем приближении (2.104), в котором

- класс функций \mathcal{F} , для которых мы строим приближения, — линейное векторное пространство функций, на котором задано скалярное произведение $\langle \cdot, \cdot \rangle$, определяющее 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, порождаемой скалярным произведением.

Напомним, что вещественное конечномерное линейное векторное пространство, в котором определено скалярное произведение, называется *евклидовым пространством*. Бесконечномерное линейное векторное пространство со скалярным произведением называют *гильбертовым пространством*, если выполнено дополнительное условие *полно-*

ты: всякая фундаментальная последовательность относительно нормы, порождённой скалярным произведением, имеет в этом пространстве предел [12, 16]. Евклидовы пространства — это пространства с привычной нам геометрией, а гильбертовы пространства являются их ближайшим обобщением.

Из результатов предыдущего раздела §2.10б следует, что решение задачи наилучшего приближения в евклидовом подпространстве всегда существует и единственно. Кроме того, ниже мы покажем, что это наилучшее приближение может быть конструктивно найдено в результате решения некоторой специальной системы линейных алгебраических уравнений, построенной по приближаемому элементу и базису евклидова подпространства.

Приближение функций в норме, порождённой скалярным произведением, обычно называют *среднеквадратичным* или даже просто *квадратичным*.

В конечномерной ситуации скалярное произведение векторов $f = (f_0, f_1, \dots, f_n)^\top$ и $g = (g_0, g_1, \dots, g_n)^\top$ обычно определяется как

$$\langle f, g \rangle = f_0 g_0 + f_1 g_1 + \dots + f_n g_n = \sum_{i=0}^n f_i g_i,$$

и этим скалярным произведением задаётся норма

$$\|f\| := \left(\sum_{i=0}^n f_i^2 \right)^{1/2},$$

в которой фигурируют квадраты компонент вектора. Но во многих задачах скалярное произведение конечномерных векторов удобнее рассматривать в несколько модифицированном, хотя и совершенно эквивалентном виде —

$$\langle f, g \rangle = \frac{1}{n+1} \sum_{i=0}^n \varrho_i f_i g_i, \quad (2.107)$$

с нормирующим множителем $1/(n+1)$ при сумме и какими-то положительными весовыми множителями $\varrho_i > 0$ для отдельных компонент. Порождённая этим скалярным произведением норма $\|\cdot\|$ определяется как

$$\|f\| := \left(\frac{1}{n+1} \sum_{i=0}^n \varrho_i f_i^2 \right)^{1/2}, \quad (2.108)$$

а расстояние между функциями дискретного аргумента f и g есть

$$\text{dist}(f, g) = \|f - g\| = \left(\frac{1}{n+1} \sum_{i=0}^n \varrho_i (f_i - g_i)^2 \right)^{1/2}. \quad (2.109)$$

Под знаком степени « $1/2$ » в этих выражениях стоит не что иное как усреднение квадратов компонент векторов с весовыми множителями ϱ_i , $i = 0, 1, \dots, n$.

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 ϱ_i можно назначать так, чтобы отразить величину этой точности, сопоставляя больший вес более точным значениям f_i . Нормирующий множитель $\frac{1}{n+1}$ при суммах в (2.107), (2.108) и (2.109) 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

$$\langle f, g \rangle = \int_a^b \varrho(x) f(x) g(x) dx, \quad (2.110)$$

для некоторой весовой функции $\varrho(x) > 0$. Это выражение с точностью до множителя можно рассматривать как предел выражения (2.107) при $n \rightarrow \infty$, так как в (2.107) легко угадываются интегральные суммы Римана для интеграла (2.110) по интервалу $[a, b]$ единичной длины и его равномерном разбиении на подинтервалы. Тогда аналогом нормы (2.108) будет

$$\|f\| := \left(\int_a^b \varrho(x) (f(x))^2 dx \right)^{1/2}, \quad (2.111)$$

а расстояние между функциями вместо (2.109) определится как

$$\text{dist}(f, g) = \|f - g\| = \left(\int_a^b \varrho(x) (f(x) - g(x))^2 dx \right)^{1/2}. \quad (2.112)$$

Это *среднеквадратичная метрика*, которую мы упоминали в §2.1.

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.

В условиях постановки задачи, описанной в начале раздела, 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 приближение g в виде

$$g = \sum_{i=1}^m c_i \varphi_i, \quad (2.113)$$

где $c_i, i = 1, 2, \dots, m$, — неизвестные коэффициенты, подлежащие определению.

Если через Φ обозначить квадрат нормы отклонения f от g , то

$$\begin{aligned} \Phi &= \|f - g\|^2 = \langle f - g, f - g \rangle \\ &= \langle f, f \rangle - 2\langle f, g \rangle + \langle g, g \rangle \\ &= \langle f, f \rangle - 2 \sum_{i=1}^m c_i \langle f, \varphi_i \rangle + \sum_{i=1}^m \sum_{j=1}^m c_i c_j \langle \varphi_i, \varphi_j \rangle. \end{aligned} \quad (2.114)$$

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 Φ принимает только неотрицательные значения. При возрастании евклидовой нормы вектора коэффициентов $c = (c_1, c_2, \dots, c_m)$ в разложении (2.113) 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

$$v := \min_{\|c\|=1} \left\| \sum_{i=1}^m c_i \varphi_i \right\| \quad (2.115)$$

— минимум значений нормы вектора g из представления (2.113) по всем векторам коэффициентов $c = (c_1, c_2, \dots, c_m)$, имеющим единичную евклидову норму. Множество таких векторов компактно в \mathbb{R}^m , и потому значение v достигается непрерывной функцией, которой является норма вектора (2.113). Кроме того, $v > 0$, так как равенство нулю означало

бы наличие зануляющейся нетривиальной линейной комбинации векторов базиса $\{\varphi_i\}_{i=1}^m$.

Для любого другого вектора коэффициентов $c = (c_1, c_2, \dots, c_m)$ разложения (2.113) имеем

$$\|g\| = \left\| \sum_{i=1}^m c_i \varphi_i \right\| = \|c\| \cdot \left\| \sum_{i=1}^m \frac{c_i}{\|c\|} \varphi_i \right\| \geq \|c\| \cdot v,$$

и это значение может сделаться сколь угодно большим при возрастании $\|c\|$. 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 $\|f\| = \text{const}$.

Покажем, что $\Phi(c_1, c_2, \dots, c_m)$ в действительности достигает своего минимума на всём \mathbb{R}^m .

Прежде всего, найдём стационарные точки функции Φ , т.е. точки зануления её производной. Продифференцировав $\Phi(c_1, c_2, \dots, c_m)$ по c_i , $i = 1, 2, \dots, m$, и приравнявая полученные производные нулю, получим

$$\frac{\partial \Phi}{\partial c_i} = -2\langle f, \varphi_i \rangle + 2 \sum_{j=1}^m c_j \langle \varphi_i, \varphi_j \rangle = 0. \quad (2.116)$$

Здесь множитель 2 при сумме всех $c_j \langle \varphi_i, \varphi_j \rangle$ появляется оттого, что в двойной сумме из выражения (2.114) слагаемое с c_i возникает дважды: один раз с коэффициентом $\langle \varphi_i, \varphi_j \rangle$, а другой раз — с коэффициентом $\langle \varphi_j, \varphi_i \rangle$.

В целом, для определения неизвестных c_i , $i = 1, 2, \dots, m$, из равенств (2.116) получается система линейных алгебраических уравнений

$$\sum_{j=1}^m \langle \varphi_i, \varphi_j \rangle c_j = \langle f, \varphi_i \rangle, \quad i = 1, 2, \dots, m. \quad (2.117)$$

Матрица её коэффициентов

$$\Gamma(\varphi_1, \varphi_2, \dots, \varphi_m) = \begin{pmatrix} \langle \varphi_1, \varphi_1 \rangle & \langle \varphi_1, \varphi_2 \rangle & \dots & \langle \varphi_1, \varphi_m \rangle \\ \langle \varphi_2, \varphi_1 \rangle & \langle \varphi_2, \varphi_2 \rangle & \dots & \langle \varphi_2, \varphi_m \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \varphi_m, \varphi_1 \rangle & \langle \varphi_m, \varphi_2 \rangle & \dots & \langle \varphi_m, \varphi_m \rangle \end{pmatrix} \quad (2.118)$$

называется, как известно, *матрицей Грама* системы векторов $\varphi_1, \varphi_2, \dots, \varphi_m$. Из курса линейной алгебры и аналитической геометрии читателю должно быть известно, 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, [35]). При выполнении этого условия матрица Грама является ещё и положительно определённой. Таким образом, решение системы уравнений (2.117) существует и единственно, 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.

Дифференцируя функцию Φ второй раз, нетрудно найти её гессиан, т. е. матрицу вторых производных. Она постоянна и равна удвоенной матрице Грама (2.118), и поэтому является положительно определённой. Применяя известное из математического анализа достаточное условие экстремума функции многих переменных, которое основано на информации о вторых производных (см. [12, 38]), можем заключить, что в стационарной точке функции Φ , т. е. на решении системы (2.117), в самом деле достигается минимум. Этот минимум является глобальным, так как других стационарных точек гладкая функция Φ не имеет, а «на бесконечности», т. е. при неограниченном удалении аргумента (c_1, c_2, \dots, c_m) от нуля, значения Φ неограниченно возрастают.

Подведём итоги. Для нахождения наилучшего среднеквадратичного приближения нужно

- 1) по базису $\{\varphi_1, \varphi_2, \dots, \varphi_m\}$ 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 $b = (\langle f, \varphi_1 \rangle, \dots, \langle f, \varphi_2 \rangle, \dots, \langle f, \varphi_m \rangle)^\top$;
- 3) решить систему линейных уравнений $Gc = b$, определив коэффициенты разложения $c = (c_1, c_2, \dots, c_m)^\top$ наилучшего среднеквадратичного приближения;
- 4) по найденным коэффициентам разложения построить вектор наилучшего среднеквадратичного приближения $g = \sum_{i=1}^m c_i \varphi_i$.

Пример 2.10.3 Среди всех квадратичных функций $y(x) = \alpha x^2 + \beta$ с вещественными параметрами α, β найдём ту, которая имеет наименьшее среднеквадратичное отклонение от данных

$$\begin{array}{c|c|c|c} x & 1 & 2 & 3 \\ \hline y & 1 & 1 & 2 \end{array}.$$

Эта задача — частный случай полиномиального среднеквадратичного приближения функции по дискретному набору значений. Подобные задачи часто возникают в ситуациях, когда необходимо найти выражение для функциональной зависимости, наилучшим образом соответствующее данным измерениям или наблюдениям. 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 $x = 1, 2, 3$.

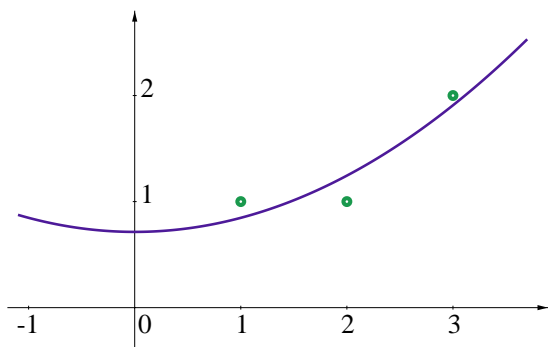


Рис. 2.20. График квадратного двучлена наилучшего среднеквадратичного приближения.

В качестве базиса в \mathcal{G} возьмём функции $\varphi_1 = 1$, $\varphi_2 = x^2$. На значениях аргументов 1, 2 и 3 эти функции принимают наборы значений $(1, 1, 1)$ и $(1, 4, 9)$. Скалярное произведение векторов в данном случае — это сумма произведений их компонент, т. е. значений при соответствующих аргументах, и потому матрица Грама этой системы векторов

$$\begin{pmatrix} \langle \varphi_1, \varphi_1 \rangle & \langle \varphi_1, \varphi_2 \rangle \\ \langle \varphi_2, \varphi_1 \rangle & \langle \varphi_2, \varphi_2 \rangle \end{pmatrix} = \begin{pmatrix} 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 1 & 1 \cdot 1 + 1 \cdot 4 + 1 \cdot 9 \\ 1 \cdot 1 + 4 \cdot 1 + 9 \cdot 1 & 1 \cdot 1 + 4 \cdot 4 + 9 \cdot 9 \end{pmatrix} = \begin{pmatrix} 3 & 14 \\ 14 & 98 \end{pmatrix}.$$

Вектор значений данных, который нам необходимо приближать, имеет вид $f = (1, 1, 2)^\top$, так что правой частью системы линейных уравнений

(2.117) является

$$\begin{pmatrix} \langle f, \varphi_1 \rangle \\ \langle f, \varphi_2 \rangle \end{pmatrix} = \begin{pmatrix} 1 \cdot 1 + 1 \cdot 1 + 2 \cdot 1 \\ 1 \cdot 1 + 1 \cdot 4 + 2 \cdot 9 \end{pmatrix} = \begin{pmatrix} 4 \\ 23 \end{pmatrix}.$$

Нетрудно найти решение системы уравнений (2.117) для определения коэффициентов разложения. Оно равно $(70/98, 13/98)^\top$, что соответствует $\alpha \approx 0.1326531$, $\beta \approx 0.7142857$. Итак, искомой функцией, наилучшим образом приближающей данные среди всех квадратных двучленов, будет

$$y(x) = 0.1326531 x^2 + 0.7142857,$$

а её график изображён на Рис. 2.20. Из чертежа видно, что построенная функция в самом деле даёт неплохое приближение к данным, которые изображены кружками. ■

Обратимся теперь к практическим аспектам реализации развитого выше метода и обсудим свойства системы уравнений (2.117). Наиболее простой вид матрица Грама имеет в случае, когда базисные функции φ_i ортогональны друг другу, т.е. когда $\langle \varphi_i, \varphi_j \rangle = 0$ при $i \neq j$. Тогда система линейных уравнений (2.117) становится диагональной и решается тривиально. Соответствующее наилучшее приближение имеет при этом вид суммы

$$g = \sum_{i=1}^m c_i \varphi_i, \quad \text{где } c_i = \frac{\langle f, \varphi_i \rangle}{\langle \varphi_i, \varphi_i \rangle}, \quad i = 1, 2, \dots, m. \quad (2.119)$$

Это представление, как известно, называется *рядом Фурье* для f по ортогональной системе векторов $\{\varphi_i\}_{i=1}^m$. Коэффициенты c_i из (2.119) называют при этом *коэффициентами Фурье* разложения функции f . В нашем случае ряд Фурье конечен, но в общем случае он может быть и бесконечным. См. подробности, например, в [12, 16].

Кроме того, в случае ортогонального и близкого к ортогональному базиса $\{\varphi_i\}_{i=1}^m$ решение системы (2.117) устойчиво к возмущениям в правой части и неизбежным погрешностям вычислений. Но в общем случае, если базис линейного подпространства \mathcal{G} сильно отличается от ортогонального, то свойства системы уравнений (2.117) могут быть плохими в том смысле, что её решение будет чувствительным к возмущениям данных и погрешностям вычислений.

2.10г Геометрия наилучшего приближения

В предшествующем разделе мы рассматривали задачу наилучшего приближения в евклидовом подпространстве с помощью аналитических инструментов. Но решение этой задачи имеет также красивую геометрическую интерпретацию.

Наилучшее приближение в евклидовом подпространстве — перпендикулярная (ортогональная) проекция на это подпространство

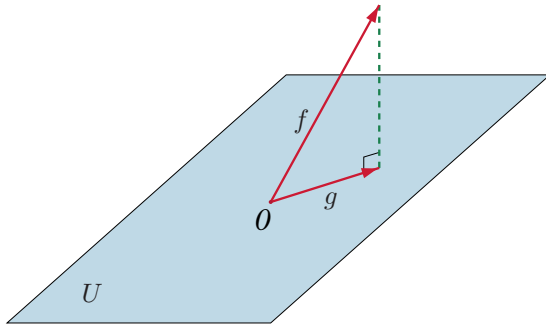


Рис. 2.21. Наилучшее приближение в евклидовом подпространстве — ортогональная проекция вектора на это подпространство.

Теорема 2.10.3 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 их разность $(f - g)$ ортогональна подпространству U .

Доказательство. Нам нужно показать, что $\|f - g\|^2 = \langle f - g, f - g \rangle$ достигает минимума, если и только если $(f - g) \perp U$.

Разложим разность $f - g$ на компоненты, ортогональную U и лежащую в 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 $u = 0$.

Имеем

$$\begin{aligned}\|f - g\|^2 &= \langle f - g, f - g \rangle = \langle u + v, u + v \rangle \\ &= \langle u, u \rangle + 2\langle u, v \rangle + \langle v, v \rangle = \|u\|^2 + \|v\|^2,\end{aligned}$$

так как $u \perp v$ по условию разложения разности $f - g$. Требование минимизации $\|f - g\|^2 = \|u\|^2 + \|v\|^2$ за счёт выбора $u \in U$ влечёт $\|u\|^2 = 0$.

Наоборот, если $u = 0$, то разность $f - g$ получает наименьшую евклидову норму. Таким образом, в любом случае $f - g = v$, т. е. разность $f - g$ должна быть ортогональна подпространству U , в котором мы находим приближение. ■

2.10д Среднеквадратичное приближение из линейной оболочки векторов

Рассмотрим теперь более общий, но вместе с тем 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 некоторых векторов, которые могут и не быть базисом. Напомним, что линейной оболочкой заданного набора векторов v_1, v_2, \dots, v_m называется множество

$$\text{lin} \{ v_1, v_2, \dots, v_m \} := \left\{ \sum_{i=1}^m \alpha_i v_i \mid \alpha_i \in \mathbb{R} \right\},$$

образованное всевозможными линейными комбинациями данных векторов. Эквивалентное определение: линейной оболочкой заданного набора векторов называется наименьшее по включению линейное подпространство, содержащее все эти векторы. То обстоятельство, что они не образуют базиса, означает «избыточность» этого набора, т. е. что некоторые из его векторов являются линейными комбинациями остальных векторов.

Итак, пусть теперь

$$\mathcal{G} = \text{lin} \{ \varphi_1, \varphi_2, \dots, \varphi_m \}$$

для каких-то $\varphi_1, \varphi_2, \dots, \varphi_m$. Такой набор векторов часто получается в результате наблюдений или измерений, но проверка его 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 среднеквадратичное приближение в виде линейной комбинации

$$g = \sum_{i=1}^m c_i \varphi_i, \quad (2.113)$$

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, и, кроме того, линейная комбинация (2.113) 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 с линейными членами и постоянным слагаемым. Функция Φ всегда неотрицательна, но при возрастании евклидовой нормы вектора коэффициентов (c_1, c_2, \dots, c_m) , когда вектор g устремляется «к бесконечности», функция Φ не обязательно принимает сколь угодно большие значения. Это может произойти при линейной зависимости $\varphi_1, \varphi_2, \dots, \varphi_m$, когда минимум (2.115) занулится.

Стационарные точки функции Φ являются решениями системы линейных алгебраических уравнений (2.117) с матрицей (2.118), которая служит матрицей Грама системы векторов $\varphi_1, \varphi_2, \dots, \varphi_m$. Но теперь нельзя определённо 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 не применимыми рассуждения из §2.10в, обосновывающие существование и единственность решения системы (2.117) для нахождения коэффициентов разложения (2.113).

Посмотрим, тем не менее, на нашу задачу с точки зрения общей теории из §2.10б. Линейная оболочка \mathcal{G} является конечномерным линейным подпространством в нормированном пространстве, и потому в

Назовём *невязкой* приближённого решения \tilde{x} уравнения (или системы уравнений) разность левой и правой частей при подстановке в них \tilde{x} . Нередко невязкой называют функцию разности левой и правой частей уравнения или системы уравнений.

Итак, вместо обычного решения системы уравнений, если оно не существует, можно рассмотреть такой вектор, на котором достигается «наименьшее значение» невязки. Это определение нуждается в уточнении, так как не вполне ясно, в каком именно смысле понимается «наименьшее значение» вектора. Обычно рассматривают какую-либо норму невязки, т. е. в качестве псевдорешения берётся вектор, на котором минимальна какая-то норма невязки $Ax - b$.

Определение 2.10.2 *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 и к отдельным уравнениям, и вместо нормы невязки тогда рассматривается просто модуль невязки.

Одной из наиболее популярных норм, относительно которых минимизируют невязку системы уравнений, является евклидова норма вектора (т. е. его обычная длина), часто называемая 2-нормой (см. §3.3а):

$$\|x\|_2 := \sqrt{\sum_{i=1}^n |x_i|^2} \quad \text{для } x = (x_1, x_2, \dots, x_n)^\top.$$

Евклидова норма порождается стандартным скалярным произведением $\langle \cdot, \cdot \rangle$ в \mathbb{R}^n

$$\langle x, y \rangle = \sum_{i=1}^n x_i y_i = x^\top y \quad \text{для } x, y \in \mathbb{R}^n,$$

так что $\|x\|_2 = \sqrt{\langle x, x \rangle}$. Поскольку для любой $m \times n$ -матрицы $A = (a_{ij})$ и произвольного n -вектора x

$$Ax = \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{pmatrix} x_1 + \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{pmatrix} x_2 + \cdots + \begin{pmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{pmatrix} x_n,$$

то задача нахождения псевдорешений системы линейных алгебраических уравнений $Ax = b$ в случае евклидовой нормы сводится к построению наилучшего среднеквадратичного приближения вектора правой части b из линейной оболочки вектор-столбцов матрицы A . Соответственно, здесь мы можем применить всю развитую выше в §2.10в и §2.10д теорию.

Псевдорешение системы линейных уравнений $Ax = b$ относительно евклидовой нормы — вектор коэффициентов разложения по столбцам матрицы системы A для наилучшего среднеквадратичного приближения правой части b . Эти коэффициенты разложения определяются из вспомогательной системы уравнений (2.117) с матрицей Грама вектор-столбцов A . Следовательно, в матрице системы уравнений (2.117) 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 -ый элемент матрицы $A^T A$.

В правой части системы уравнений (2.117) в качестве i -ой компоненты стоит скалярное произведение i -го столбца A и вектора b , т. е.

$$\sum_{k=1}^m a_{ki}b_k,$$

и это i -ая компонента вектора $A^T b$. Итак, система линейных алгебраических уравнений (2.117) для определения наилучшего приближения имеет в нашем случае вид

$$A^T Ax = A^T b.$$

Определение 2.10.3 Для заданной системы линейных алгебраических уравнений $Ax = b$ система вида $A^T Ax = A^T b$ называется нормальной системой уравнений.

Теорема 2.10.4 Для системы линейных алгебраических уравнений $Ax = b$ псевдорешение относительно евклидовой нормы является решением нормальной системы уравнений $A^T Ax = A^T b$.

Доказательство немедленно следует из результатов предшествующего раздела §2.10д.

Предложение 2.10.1 *Нормальная система уравнений $A^T Ax = A^T b$ всегда имеет решение.*

Фактически, мы обосновали это Предложение в §2.10д на основе теоремы Э. Бореля, но ниже будет дано ещё одно прямое доказательство, которое не опирается на результаты по среднеквадратичным приближениям.

Доказательство будет опираться на критерий разрешимости системы линейных алгебраических уравнений, известный как «теорема Фредгольма» (см. §3.2ж). 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 $A^T Ay = 0$, так как её матрица, очевидно, совпадает с симметричной матрицей нормальной системы.

Если $A^T A\tilde{y} = 0$ для некоторого \tilde{y} , то

$$0 = \tilde{y}^T (A^T A\tilde{y}) = (\tilde{y}^T A^T) (A\tilde{y}) = (A\tilde{y})^T (A\tilde{y}) = \|A\tilde{y}\|_2^2,$$

откуда следует, что $A\tilde{y} = 0$. Следовательно,

$$\langle \tilde{y}, A^T b \rangle = \tilde{y}^T (A^T b) = (\tilde{y}^T A^T) b = (A\tilde{y})^T b = 0,$$

т.е. любое решение \tilde{y} однородной транспонированной системы ортогонально вектору $A^T b$, стоящему в правой части нормальной системы уравнений. В силу альтернативы Фредгольма можем заключить, что система линейных уравнений $A^T Ax = A^T b$ в самом деле должна быть разрешимой. ■

Для заданного уравнения или системы уравнений псевдорешение может быть неединственным, т.е. возможно существование бесконечных множеств псевдорешений. Кроме того, псевдорешение неустойчиво к возмущениям элементов матрицы, если она имеет неполный ранг, т.е. меньший $\min\{m, n\}$. По этим причинам обычно выделяют из всех псевдорешений так называемые *нормальные псевдорешения*, которые имеют наименьшую евклидову норму. Из Теоремы 2.10.2 следует, что нормальное псевдорешение системы линейных алгебраических уравнений единственно, и, кроме того, можно показать, что оно более устойчиво к возмущениям элементов матрицы и правой части системы.

Дальнейшее обсуждение темы и обзор численных методов можно найти в разделе 3.15.

Пример 2.10.4 Среди всех линейных функций вида $y(x) = \alpha x + \beta$ найдём ту, которая имеет наименьшее среднеквадратичное отклонение от данных

x	1	2	3
y	1	1	2

В данном случае можно считать набор параметров линейной зависимости, т. е. вектор $(\alpha, \beta)^\top$, псевдорешением относительно евклидовой нормы для системы линейных алгебраических уравнений

$$\begin{cases} \alpha + \beta = 1, \\ 2\alpha + \beta = 1, \\ 3\alpha + \beta = 2, \end{cases}$$

которая получается при подстановке в выражение для функции $y(x)$ значений аргументов $x = 1, 2, 3$ и приравниваем значениям функции.

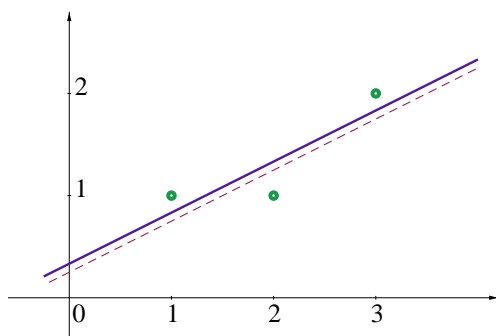


Рис. 2.22. Графики линейных функций наилучшего среднеквадратичного приближения и наилучшего чебышёвского приближения.

В векторно-матричной форме эта система имеет вид

$$\begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix},$$

и для неё нормальная система уравнений выглядит следующим образом

$$\begin{pmatrix} 14 & 6 \\ 6 & 3 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 9 \\ 4 \end{pmatrix}.$$

Нетрудно найти решение системы уравнений (2.117) для определения коэффициентов разложения. Оно равно $(\frac{1}{2}, \frac{1}{3})^T$, так что искомой линейной функцией, которая наилучшим образом приближает данные, будет

$$y(x) = \frac{1}{2}x + \frac{1}{3},$$

и её график изображён на Рис. 2.22 толстой сплошной прямой. Для сравнения на том же рисунке тонким пунктиром представлен график линейной функции наилучшего чебышёвского (равномерного) приближения тех же данных. Она задаётся выражением $\frac{1}{2}x + \frac{1}{4}$, у которого тот же угловой коэффициент, но другой свободный член. ■

Рассмотренный пример очевидным образом обобщается на многомерный случай и позволяет решать задачу построения линейной функции нескольких переменных, наименее уклоняющейся от заданного набора значений в среднеквадратичном смысле. Этот способ построения приближающей линейной функции называют *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.

Среднеквадратичные приближения и метод наименьших квадратов для решения переопределённых систем линейных алгебраических уравнений, которые возникают в связи с задачами обработки наблюдений, были почти одновременно предложены на рубеже XVIII–XIX веков А.-М. Лежандром и К.Ф. Гауссом. Современное название новому 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, [60]). 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 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 выполняется, то метод наименьших квадратов становится не самой лучшей возможностью решения задачи приближения или задачи восстановления функциональной зависимости.

2.10ж Среднеквадратичное приближение функций

В этом разделе мы применим развитый в §2.10в 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.

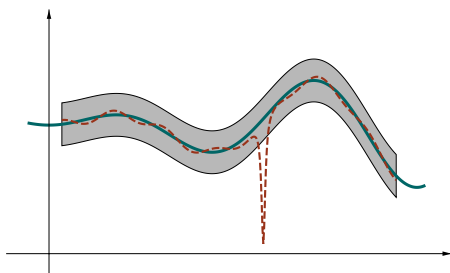


Рис. 2.23. Иллюстрация различия равномерного и интегрального (в частности, среднеквадратичного) отклонений функций

Пример 2.10.5 В качестве примера практического возникновения

задачи среднеквадратичного приближения рассмотрим тепловое действие тока $I(t)$ в проводнике сопротивлением R . Мгновенная тепловая мощность, как известно из теории электричества, равна при этом $I^2(t)R$, а полное количество теплоты, выделившееся между моментами времени a и b , равно

$$\int_a^b I^2(t)R dt.$$

Если мы хотим, скажем, минимизировать тепловыделение рассматриваемого участка электрической цепи, то нам нужно искать такой режим её работы, при котором достигался бы минимум выписанного интеграла, т. е. среднеквадратичного значения тока. В электротехнике его называют также *действующим* или *эффективным* значением силы тока, и именно его измеряют почти все амперметры переменного тока. ■

Как соотносится наша задача среднеквадратичного приближения функций с абстрактной постановкой из §2.10в и данным там же методом её решения? Класс приближаемых функций \mathcal{F} должен быть линейным векторным пространством со скалярным произведением (2.110). Иногда эти условия выполняются тривиально, но бывают ситуации, когда они не могут быть удовлетворены.

Пример 2.10.6 Известно, что эволюция некоторого физического процесса во времени описывается функцией

$$y = \frac{a+t}{b+t}, \quad \text{где } b > a > 0, \quad (2.120)$$

По измеренным данным для t и y нужно восстановить конкретный вид функциональной зависимости, т. е. найти a и b .

В принципе, можно рассмотреть среднеквадратичное приближение результатов измерений классом функций вида (2.120) для различных параметров a , b , и выбрать наилучшее приближение. Такое решение имеет теоретико-вероятностный смысл. Но метод, развитый в §2.10в, непригоден для его получения, так как функции (2.120) не зависят линейно от совокупности искомых параметров a и b . Следовательно, линейное пространство они не образуют. ■

С другой стороны, класс приближаемых функций \mathcal{F} со структурой линейного пространства часто бывает неявно определён самой поста-

новкой задачи. К примеру, мы работаем с непрерывным (гладкими и т. п.) функциями с поточечными операциями сложения и умножения на скаляр. Но и здесь выбор конкретного \mathcal{F} может быть предметом обсуждения. 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* (называемая также *функцией единичного скачка*, см. Рис. 2.24), которая задаётся как

$$\theta(x) := \begin{cases} 0, & \text{если } x < 0, \\ 1, & \text{если } x \geq 0. \end{cases}$$

Функция Хевисайда широко используется в математической теории управления и обработке сигналов.

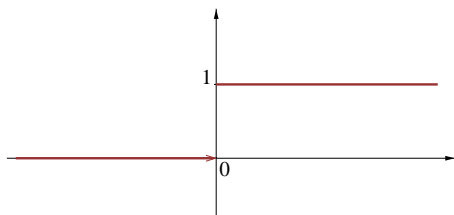


Рис. 2.24. График функции Хевисайда

В этих условиях в качестве класса приближаемых функций \mathcal{F} естественно взять множество всех вещественнозначных функций, определённых на интервале $[a, b] \subset \mathbb{R}$ и для которых определены скалярное произведение (2.110) и вытекающие из него конструкции. Обычно требуют, чтобы была конечная норма (2.111), т. е. для таких функций квадрат (т. е. степень 2) должен быть интегрируем на интервале $[a, b]$ с заданным положительным весом $\varrho(x)$. Тогда в силу очевидного неравенства

$$2|f(x)g(x)| \leq (f(x))^2 + (g(x))^2$$

мы получим также интегрируемость их произведения с той же весовой функцией.

Множество функций, для которых интеграл от квадрата с заданным весом конечен на $[a, b]$, называют пространством $\mathcal{L}^2[a, b]$. Операции сложения векторов-функций и умножения на скаляр задаются в нём обычным поточечным образом. Чтобы оказаться в условиях постановки задачи из раздела §2.10в и воспользоваться развитым там методом решения, нам необходимо 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.²²

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

$$\int_a^b \varrho(x) (f(x) + g(x))^2 dx =$$

$$\int_a^b \varrho(x) (f(x))^2 dx + 2 \int_a^b \varrho(x) f(x) g(x) dx + \int_a^b \varrho(x) (g(x))^2 dx,$$

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 курсах функционального анализа показывается, что если интегрирование понимается в смысле Лебега, то $\mathcal{L}^2[a, b]$ — гильбертово пространство, т. е. дополнительно обладает свойством полноты (см., к примеру, [16]). По этой причине оно очень популярно в самых различных математических дисциплинах, от теории уравнений в частных производных до математической статистики. Кроме пространства \mathcal{L}^2 существуют и часто применяются пространства \mathcal{L}^p , состоящие из функций, у которых интегрируема p -ая степень, $p \geq 1$.

В качестве пространства \mathcal{G} , в котором ищутся приближения для элементов из $\mathcal{L}^2[a, b]$, обычно рассматривается какое-то его конечномерное подпространство. Например, это могут быть алгебраические или тригонометрические полиномы заданной степени, пространства экспоненциальных сумм и т. п. В \mathcal{G} задаётся некоторый базис $\{\varphi_j(x)\}_{j=1}^m$ и

²²Буква « \mathcal{L} » в обозначении этого пространства связывается с именем А.Л. Лебега.

среднеквадратичное приближение ищется в виде

$$g(x) = \sum_{j=1}^m c_j \varphi_j(x). \quad (2.121)$$

В общем виде наша задача была решена в §2.10в, но интересно знать, как выглядит система линейных уравнений (2.117) для определения коэффициентов c_j наилучшего приближения в связи с конкретной постановкой этого раздела, т.е. когда мы приближаем вещественные функции на интервале. Это зависит как от подпространства $\mathcal{G} \subset \mathcal{L}^2[a, b]$, так и от базиса, выбранного в \mathcal{G} . Рассмотрим конкретные ситуации, которые могут здесь встретиться.

Пример 2.10.7 Пусть дана задача о среднеквадратичном приближении функций из $\mathcal{L}^2[0, 1]$ с единичным весом полиномами фиксированной степени m . Скалярное произведение определяется при этом как

$$\langle f, g \rangle = \int_0^1 f(x) g(x) dx,$$

а нормой берём

$$\|f\| = \left(\int_0^1 (f(x))^2 dx \right)^{1/2}.$$

Соответственно, расстояние между функциями определяется тогда как

$$\text{dist}(f, g) = \|f - g\| = \left(\int_0^1 (f(x) - g(x))^2 dx \right)^{1/2}.$$

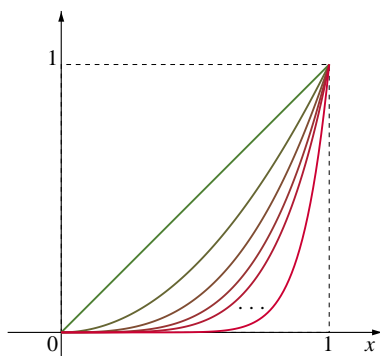
Если в качестве базиса в линейном подпространстве полиномов мы возьмём последовательные степени

$$1, \quad x, \quad x^2, \quad \dots, \quad x^m,$$

то на месте (i, j) в матрице Грама (2.118) 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).

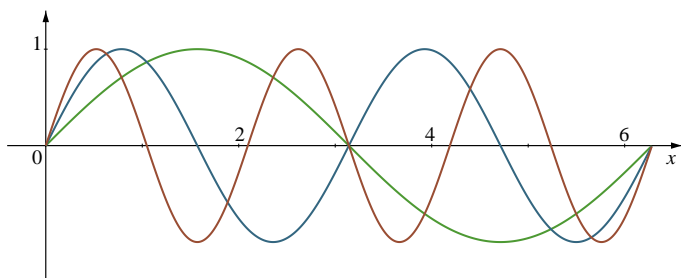
Рис. 2.25. Графики последовательных степеней переменной x .

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. §3.56). Иными словами, решение СЛАУ с этой матрицей 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 (см. Рис. 2.25). Совершенно то же происходит со строками (или столбцами) матрицы Гильберта, отличие которых с ростом номера делается всё меньшим. Поэтому хотя последовательные степени x^n линейно независимы, но базис из них «сплюснен», и это обстоятельство отражает их матрица Грама. ■

Рис. 2.26. Графики функций $\sin x$, $\sin 2x$ и $\sin 3x$.

Пример 2.10.8 Пусть k и l — натуральные числа. Поскольку

$$\int_0^{2\pi} \sin(kx) \cos(lx) dx = 0$$

для любых k, l , и

$$\int_0^{2\pi} \sin(kx) \sin(lx) dx = 0, \quad \int_0^{2\pi} \cos(kx) \cos(lx) dx = 0$$

для $k \neq l$, то базис из тригонометрических полиномов вида

$$1, \quad \cos(kx), \quad \sin(kx), \quad k = 1, 2, \dots, \quad (2.122)$$

является ортогональным на $[0, 2\pi]$ относительно скалярного произведения (2.110) с весом $\varrho(x) = 1$. Иными словами, в вычислительном отношении этот базис очень хорош для построения среднеквадратичных приближений. Ясно, что вместо интервала $[0, 2\pi]$ можно взять любой другой интервал, ширина которого равна периоду 2π , а подходящим масштабированием из него можно получить вообще любой интервал вещественной оси.

Из Рис. 2.26 видно, что поведение функций тригонометрической системы — совершенно другое, нежели у последовательных степеней на Рис. 2.25: функции тригонометрической системы «существенно отличаются» друг от друга, и это приводит к «хорошей» матрице Грама.

Отметим, что исторически первые ряды Фурье были построены Ж.Б. Фурье в начале XIX века именно как разложения по тригонометрической системе функций (2.122). ■

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, ортогональных по отношению к какому-то скалярному произведению (возможно, даже отличающемуся от того, относительно которого рассматривается задача приближения). 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 (2.121), которая берётся в виде линейной комбинации базисных, не подходит по тем или иным причинам (именно таков Пример 2.10.6). 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 системы линейных алгебраических уравнений (2.117), и для нахождения минимума нам нужно применять численные методы оптимизации. Обсуждение этого круга вопросов и дальнейшие ссылки можно найти в книге [50].

2.11 Полиномы Лежандра

2.11a Мотивация и определение

Примеры 2.10.7 и 2.10.8 из предшествующего раздела показывают, что выбор хорошего, т. е. ортогонального или почти ортогонального, базиса для среднеквадратичного приближения функций является очень важной задачей. Для её конструктивного решения можно воспользоваться, к примеру, известным из курса линейной алгебры процессом ортогонализации Грама-Шмидта или его модификациями (см. §3.7ж). Напомним, что по данной 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:

$$q_1 \leftarrow v_1, \quad (2.123)$$

$$q_k \leftarrow v_k - \sum_{i=1}^{k-1} \frac{\langle v_k, q_i \rangle}{\langle q_i, q_i \rangle} q_i, \quad k = 2, \dots, n. \quad (2.124)$$

Иногда получающийся ортогональный базис дополнительно нормируют.

В задаче среднеквадратичного приближения функций из §2.10ж ортогонализуемые элементы линейного пространства — это функции, а их скалярное произведение — интеграл (2.110). По этой причине процесс ортогонализации (2.123)–(2.124) довольно трудоёмок, а конкретный вид ортогональных в смысле $\mathcal{L}^2[a, b]$ функций, которые получаются в результате, зависит, во-первых, от интервала $[a, b]$, для которого рассматривается скалярное произведение (2.110), 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 $g((y-s)/r)$ переменной y будут ортогональны на $[a, b]$.

Рассмотрим среднеквадратичное приближение функций полиномами. В этом случае в качестве канонического интервала обычно берётся $[-1, 1]$, а произвольный интервал $[a, b]$ можно получить из него с помощью замены переменных (2.39)

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

(она уже встречалась нам в §2.3б). Ясно, что переменная y пробегает интервал $[a, b]$, если $x \in [-1, 1]$. Обратное преобразование даётся формулой (2.40)

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

которая позволяет строить ортогональные в смысле $\mathcal{L}^2[a, b]$ полиномы для любого интервала $[a, b] \subset \mathbb{R}$, зная их для интервала $[-1, 1]$.

Полиномами Лежандра называют семейство полиномов, зависящих от неотрицательного целого параметра n , таких что n -ый полином имеет степень n и в совокупности они образуют на интервале $[-1, 1]$ ортогональную систему относительно скалярного произведения (2.110) с единичным весом. Эти полиномы были введены в широкий оборот французским математиком А.-М. Лежандром в 1785 году.²³ Из общей теории скалярного произведения в линейных векторных пространствах следует, что такие полиномы существуют и единственны с точностью до постоянного множителя. Нормирование полиномов Лежандра обычно выполняют различными способами, наиболее подходящими для той или иной задачи.

Применяя к степеням $1, x, x^2, x^3, \dots$ последовательно формулы ортогонализации Грама-Шмидта (2.123)–(2.124) со скалярным произведением (2.110) на интервале $[-1, 1]$, получим

$$1, \quad x, \quad x^2 - \frac{1}{3}, \quad x^3 - \frac{3}{5}x, \quad \dots \quad (2.125)$$

(две первых степени оказываются изначально ортогональными).

Удобное альтернативное представление для полиномов Лежандра даёт *формула Родрига*:

$$L_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n, \quad n = 0, 1, 2, \dots \quad (2.126)$$

Очевидно, что функция $L_n(x)$, определяемая этой формулой, является алгебраическим полиномом n -ой 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 (2.126) 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, 64]).

Всюду далее посредством $L_n(x)$ мы будем обозначать полиномы Лежандра, определяемые формулой (2.126). Для её обоснования необходимо доказать

²³Иногда их называют *сферическими полиномами*, так как они естественно возникают при нахождении решений некоторых задач математической физики в сферических координатах. Именно так их использовал сам А.-М. Лежандр.

Предложение 2.11.1 Полиномы $L_n(x)$, $n = 0, 1, \dots$, задаваемые формулой Родрига (2.126), ортогональны друг другу в смысле скалярного произведения на $\mathcal{L}^2[-1, 1]$ 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 (2.126)

$$L_n(x) = \frac{1}{2^n n!} \psi^{(n)}(x), \quad n = 0, 1, 2, \dots$$

Поэтому, если $Q(x)$ является n -кратно непрерывно дифференцируемой функцией на $[-1, 1]$, то, последовательно применяя n раз формулу ин-

тегрирования по частям, получим

$$\begin{aligned}
 \int_{-1}^1 Q(x) L_n(x) dx &= \frac{1}{2^n n!} \int_{-1}^1 Q(x) \psi^{(n)}(x) dx \\
 &= \frac{1}{2^n n!} \int_{-1}^1 Q(x) d(\psi^{(n-1)}(x)) \\
 &= \frac{1}{2^n n!} \left(\left(Q(x) \psi^{(n-1)}(x) \right) \Big|_{-1}^1 - \int_{-1}^1 Q'(x) \psi^{(n-1)}(x) dx \right) \\
 &= -\frac{1}{2^n n!} \int_{-1}^1 Q'(x) \psi^{(n-1)}(x) dx \\
 &= \dots \dots \dots \\
 &= (-1)^n \frac{1}{2^n n!} \int_{-1}^1 Q^{(n)}(x) \psi(x) dx.
 \end{aligned} \tag{2.127}$$

Если $Q(x)$ — полином степени меньше n , то его n -ая производная $Q^{(n)}(x)$ равна тождественному нулю, а потому из полученной формулы тогда 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 (2.127) следует

$$\begin{aligned}\int_{-1}^1 L_n(x) L_n(x) dx &= (-1)^n \frac{(2n)!}{2^{2n}(n!)^2} \int_{-1}^1 \psi(x) dx \\ &= (-1)^n \frac{(2n)!}{2^{2n}(n!)^2} \int_{-1}^1 (x^2 - 1)^n dx.\end{aligned}\quad (2.128)$$

С другой стороны, последовательно интегрируя по частям n раз, получим

$$\begin{aligned}\int_{-1}^1 (x^2 - 1)^n dx &= \int_{-1}^1 (x - 1)^n (x + 1)^n dx \\ &= \frac{1}{n + 1} \int_{-1}^1 (x - 1)^n d((x + 1)^{n+1}) \\ &= \frac{1}{n + 1} \left((x - 1)^n (x + 1)^{n+1} \Big|_{-1}^1 - n \int_{-1}^1 (x - 1)^{n-1} (x + 1)^{n+1} dx \right) \\ &= \frac{(-1)^n n}{n + 1} \int_{-1}^1 (x - 1)^{n-1} (x + 1)^{n+1} dx \\ &= \dots \dots \\ &= \frac{(-1)^n n!}{(n + 1) \cdot \dots \cdot 2n} \int_{-1}^1 (x - 1)^0 (x + 1)^{2n} dx \\ &= \frac{(-1)^n (n!)^2}{(2n)!} \frac{(x + 1)^{2n+1}}{2n + 1} \Big|_{-1}^1 = \frac{(-1)^n (n!)^2}{(2n)!} \frac{2^{2n+1}}{2n + 1}.\end{aligned}$$

Комбинируя полученный результат с (2.128), будем иметь

$$\int_{-1}^1 L_n(x) L_n(x) dx = \frac{2}{2n + 1},$$

что завершает доказательство Предложения. ■

2.116 Основные свойства полиномов Лежандра

Выпишем первые полиномы Лежандра, как они даются формулой Родрига (2.126):

$$\begin{aligned}
 L_0(x) &= 1, \\
 L_1(x) &= x, \\
 L_2(x) &= \frac{1}{2}(3x^2 - 1), \\
 L_3(x) &= \frac{1}{2}(5x^3 - 3x), \\
 L_4(x) &= \frac{1}{8}(35x^4 - 30x^2 + 3), \\
 L_5(x) &= \frac{1}{8}(63x^5 - 70x^3 + 15x), \\
 &\dots
 \end{aligned} \tag{2.129}$$

Эти полиномы с точностью до множителя совпадают с результатом ортогонализации Грама-Шмидта (2.125). Графики полиномов (2.129) изображены на Рис. 2.27, и они похожи на графики полиномов Чебышёва.

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

$$(n+1)L_{n+1}(x) = (2n+1)xL_n(x) - nL_{n-1}(x).$$

Из него индукцией нетрудно показать, что полиномы Лежандра с чётными номерами являются чётными функциями, а с нечётными номерами — нечётными функциями. Детальное доказательство этих свойств можно найти, например, в книгах [29, 64]. Рекуррентная формула даёт практически удобный способ вычисления значений полиномов Лежандра, так как в их явном представлении (2.129) 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 на $[-1, 1]$ у полиномов Лежандра различны и

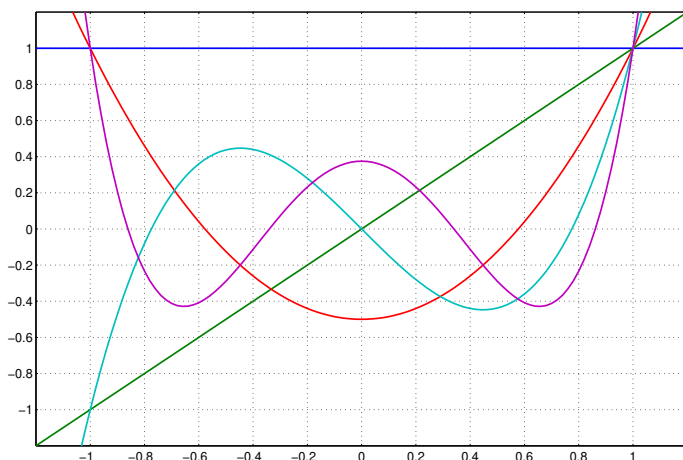


Рис. 2.27. Графики первых полиномов Лежандра на интервале $[-1.2, 1.2]$.

не могут быть сделаны одинаковыми ни при каком масштабировании. Тем не менее, сходство полиномов Лежандра и полиномов Чебышёва подтверждает следующее

Предложение 2.11.2 *Все нули полиномов Лежандра $L_n(x)$ вещественные, простые и находятся на интервале $[-1, 1]$.*

Доказательство. Предположим, что среди корней полинома $L_n(x)$, лежащих на $[-1, 1]$, имеется s штук различных корней $\theta_1, \theta_2, \dots, \theta_s$ нечётной кратности $\alpha_1, \alpha_2, \dots, \alpha_s$ соответственно. Поэтому

$$L_n(x) = (x - \theta_1)^{\alpha_1} (x - \theta_2)^{\alpha_2} \dots (x - \theta_s)^{\alpha_s} \gamma(x),$$

где в полиноме $\gamma(x)$ присутствуют корни $L_n(x)$, не лежащие на $[-1, 1]$, а также те корни $L_n(x)$ 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 определён $\mathcal{I} \neq 0$.

С другой стороны, выражение для \mathcal{I} есть скалярное произведение, в смысле $\mathcal{L}^2[-1, 1]$, 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 при этом 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 корни полинома $L_n(x)$ различны, просты и лежат на интервале $[-1, 1]$. ■

Отметим, что проведённое доказательство непосредственно переносится на общие скалярные произведения вида (2.110) 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 корни любых ортогональных в смысле \mathcal{L}^2 полиномов — 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 (2.126) достаточно поставить перед n -ой производной множитель, который компенсирует коэффициенты при старшем члене полинома $(x^2 - 1)^n$, возникающие в процессе n -кратного дифференцирования. Тогда

$$\begin{aligned}\tilde{L}_n(x) &= \frac{1}{2n(2n-1) \cdots (n+1)} \frac{d^n}{dx^n} (x^2 - 1)^n \\ &= \frac{n!}{(2n)!} \frac{d^n}{dx^n} (x^2 - 1)^n,\end{aligned}\tag{2.130}$$

$n = 1, 2, \dots$. 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 (2.130) и формулы Родрига (2.126) следует также, что

$$\tilde{L}_n(x) = \frac{2^n(n!)^2}{(2n)!} L_n(x). \quad (2.131)$$

Предложение 2.11.3 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

$$\int_{-1}^1 (Q_n(x))^2 dx \geq \int_{-1}^1 (\tilde{L}_n(x))^2 dx. \quad (2.132)$$

Доказательство. Если $Q_n(x) = x^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0$, то для отыскания наименьшего значения выражения

$$\begin{aligned} \mathcal{J}(a_0, a_1, \dots, a_{n-1}) &= \int_{-1}^1 (Q_n(x))^2 dx \\ &= \int_{-1}^1 (x^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0)^2 dx \end{aligned} \quad (2.133)$$

продифференцируем его по переменным a_0, a_1, \dots, a_{n-1} и приравняем полученные производные к нулю. Так как в данных условиях дифференцирование интеграла по параметру, от которого зависит подинтегральная функция, сводится к взятию интеграла от её производной (см. [12, 38]), имеем в результате

$$\begin{aligned} \frac{\partial \mathcal{J}}{\partial a_k} &= \int_{-1}^1 2(x^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0)x^k dx \\ &= 2 \int_{-1}^1 Q_n(x)x^k dx = 0, \quad k = 0, 1, \dots, n-1. \end{aligned} \quad (2.134)$$

То, что в точке, удовлетворяющей условиям (2.134), 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 скалярное произведение в $\mathcal{L}^2[-1, 1]$ 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 (2.134) означают, что полином $Q_n(x)$ ортогонален в смысле $\mathcal{L}^2[-1, 1]$ всем полиномам меньшей степени. Следовательно, при минимальном значении интеграла (2.133) 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 и затем необходимым масштабированием, аналогично тому, как это было сделано в задаче интерполирования для полиномов Чебышёва в §2.36. Обоснование этого способа следует из того, что при линейной замене переменной, как мы выяснили в §2.11а, из полиномов Лежандра получатся полиномы, ортогональные на $[a, b]$ с единичным весом.

Из Предложения 2.11.3 следует также, что для достижения наименьшей погрешности алгебраической интерполяции в среднеквадратичном смысле узлы интерполяции следует брать корнями 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, введённые в §2.3 полиномы Чебышёва являются ортогональными на интервале $[-1, 1]$ относительно скалярного произведения (2.110) с весом $(1 - x^2)^{-1/2}$.

Часто возникает необходимость воспользоваться ортогональными полиномами на бесконечных интервалах $[0, +\infty]$ или даже $[-\infty, \infty]$. Естественно, единичный вес $\varrho(x) = 1$ тут малоприменим, так как с ним интегралы по бесконечным интервалам окажутся, по большей части, расходящимися. Полиномы, ортогональные на интервалах $[0, +\infty]$ или $[-\infty, \infty]$ с быстроубывающими весами e^{-x} и e^{-x^2} называются полиномами Лагерра и полиномами Эрмита соответственно.²⁴ Они также находят многообразные применения в задачах приближения, и более

²⁴Иногда их называют полиномами Чебышёва-Лагерра и Чебышёва-Эрмита (см., к примеру, [47, 78]), поскольку они были известны ещё П.Л.Чебышёву.

подробные сведения на эту тему читатель может почерпнуть в [29, 78].

В §2.10ж и §2.11 мы рассматривали, главным образом, непрерывную задачу среднеквадратичного приближения, в которой рассматривались функции от непрерывного аргумента, а расстояние между функциями определялось посредством (2.112). Но дискретная задача наилучшего среднеквадратичного приближения, в которой рассматриваются функции на некоторой сетке, а расстояние между ними задаётся в виде (2.109), тоже важна и востребована на практике. Для её решения удобно воспользоваться ортогональными полиномами дискретной переменной, теория которых изложена, к примеру, в [71].

2.12 Численное интегрирование

2.12a Постановка и обсуждение задачи

Задача вычисления определённого интеграла

$$\int_a^b f(x) \, dx \quad (2.135)$$

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 т. п. механические, физические, химические и другие задачи. В математическом анализе обобщается *формула Ньютона-Лейбница*

$$\int_a^b f(x) \, dx = F(b) - F(a), \quad (2.136)$$

где $F(x)$ — первообразная для функции $f(x)$, т. е. такая, что $F'(x) = f(x)$. Она даёт удобный способ вычисления интегралов, который в значительной степени удовлетворяет потребности решения подобных задач. Тем не менее, возникают ситуации, когда для вычисления интеграла (2.135) требуются другие подходы.

Подобная задача возникает на практике в том случае, когда первообразная для интегрируемой функции не выражается через известные функции, элементарные и даже специальные функции. Даже если эта первообразная может быть найдена в конечном виде, её вычисление не всегда осуществляется просто (длинное и неустойчивое к ошиб-

кам 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 (2.135) на основе знания значений функции $f(x)$, без привлечения её первообразных и формулы Ньютона-Лейбница (2.136).

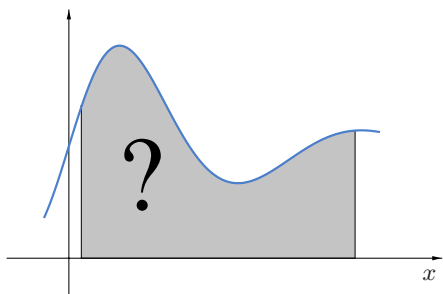


Рис. 2.28. Вычисление определённого интеграла необходимо при нахождении площадей фигур с криволинейными границами

Для нахождения интегралов наибольшее распространение в вычислительной практике получили формулы вида

$$\int_a^b f(x) dx \approx \sum_{k=0}^n c_k f(x_k), \quad (2.137)$$

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 *веса* 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*.²⁵ 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 (2.137) применяются также квадратурные формулы, использующие значения производных интегрируемой функции в узлах (см. [30]). Но мы не будем рассматривать их в этом курсе.

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, см. [12, 38]). 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. §2.2д).

Если для некоторой функции f или же для целого класса функций $\mathcal{F} \ni f$ 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

²⁵ «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».

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 2.12.1 *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 δ_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 δ_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

$$\sum_{k=0}^n |c_k|, \tag{2.138}$$

— 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 2.12.1, стр. 213). Следовательно, если мы хотим организовать вычисления по квадратурной формуле наиболее устойчивым образом, то все весовые коэффициенты c_k должны иметь один знак, т.е. быть положительными. Именно тогда при прочих равных условиях минимальна сумма модулей весов (2.138) 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, которое происходит в формуле, где присутствуют положительные и отрицательные веса.

2.126 Формулы Ньютона-Котеса. Простейшие квадратурные формулы

Простейший приём построения квадратурных формул — замена под-интегральной 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 построения 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 $c_k = \int_a^b \gamma_k(x) dx$, $k = 0, 1, \dots, n$.

В случае, когда подинтегральная функция $f(x)$ заменяется интерполантом 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 практичностью: 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*.

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 «левых прямоугольников», а если $x_0 = b$ — формула «правых прямоугольников» (см. Рис. 2.29).

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

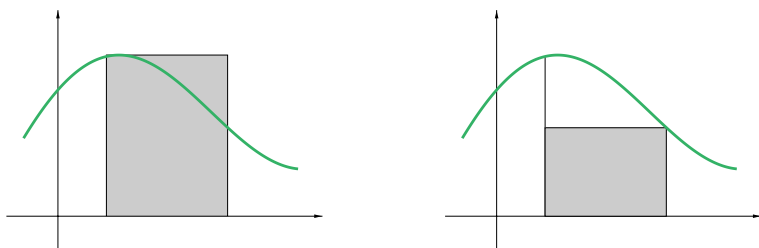


Рис. 2.29. Иллюстрация квадратурных формул левых и правых прямоугольников

квадратурной формуле

$$\int_a^b f(x) dx \approx (b-a) \cdot f\left(\frac{a+b}{2}\right),$$

называемой *формулой средних прямоугольников*: согласно ей интеграл берётся равным площади прямоугольника с основанием $(b-a)$ и высотой $f((a+b)/2)$ (см. Рис. 2.30). 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

$$f(x) = f\left(\frac{a+b}{2}\right) + f'\left(\frac{a+b}{2}\right) \cdot \left(x - \frac{a+b}{2}\right) + \frac{f''(\xi)}{2} \cdot \left(x - \frac{a+b}{2}\right)^2, \quad (2.139)$$

где ξ — зависящая от x точка интервала $[a, b]$, которую корректно обо-

значить 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

$$\int_a^b \left(x - \frac{a+b}{2}\right) dx = \int_{-\frac{b-a}{2}}^{\frac{b-a}{2}} t dt = 0,$$

— интеграл от первого члена разложения (2.139) 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

$$M_2 = \max_{x \in [a,b]} |g''(x)| = 2, \quad g\left(\frac{a+b}{2}\right) = 0,$$

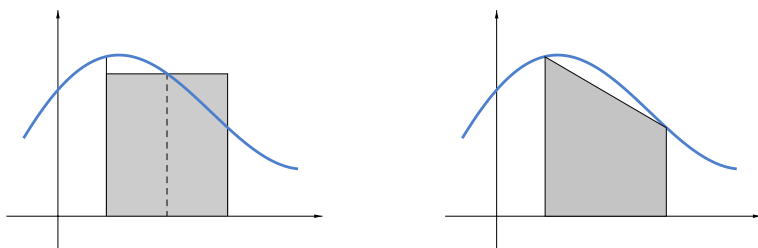


Рис. 2.30. Иллюстрация квадратурных формул средних прямоугольников и трапеций

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

$$|R(f)| \leq \frac{M_1(b-a)^2}{2}.$$

Рассмотрим теперь квадратурную формулу Ньютона-Котеса, соответствующую случаю $n = 1$, когда подынтегральная 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

$$\begin{aligned} \int_a^b P_1(x) dx &= \frac{f(a)}{a-b} \int_a^b (x-b) dx + \frac{f(b)}{b-a} \int_a^b (x-a) dx \\ &= \frac{f(a)}{a-b} \frac{(x-b)^2}{2} \Big|_a^b + \frac{f(b)}{b-a} \frac{(x-a)^2}{2} \Big|_a^b \\ &= \frac{b-a}{2} (f(a) + f(b)). \end{aligned}$$

Мы вывели *квадратурную формулу трапеций*

$$\boxed{\int_a^b f(x) dx \approx \frac{b-a}{2} \cdot (f(a) + f(b))}, \quad (2.140)$$

название которой тоже наваяно 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 $f(b)$ (см. Рис. 2.30).

Чтобы найти погрешность формулы трапеций, вспомним оценку (2.28) 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

$$\begin{aligned} |R(f)| &\leq \int_a^b \frac{|f''(\xi(x))|}{2} \cdot |(x-a)(x-b)| dx \\ &\leq \frac{M_2}{2} \cdot \left| \int_a^b (x-a)(x-b) dx \right|, \end{aligned}$$

где $M_2 = \max_{x \in [a,b]} |f''(x)|$. Далее

$$\begin{aligned} \int_a^b (x-a)(x-b) dx &= \int_a^b (x^2 - (a+b)x + ab) dx \\ &= \frac{x^3}{3} \Big|_a^b - (a+b) \frac{x^2}{2} \Big|_a^b + abx \Big|_a^b \\ &= \frac{1}{6} \left(2(b^3 - a^3) - 3(a+b)(b^2 - a^2) + 6ab(b-a) \right) \\ &= \frac{1}{6} (-b^3 + 3ab^2 - 3a^2b + a^3) = -\frac{(b-a)^3}{6}. \end{aligned} \tag{2.141}$$

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 $[a, b]$.

2.12в Квадратурная формула Симпсона

Построим квадратурную формулу Ньютона-Котеса для $n = 2$, т. е. для трёх равномерно расположенных узлов

$$x_0 = a, \quad x_1 = \frac{a+b}{2}, \quad x_2 = b$$

из интервала интегрирования $[a, b]$.

Для упрощения рассуждений выполним параллельный перенос криволинейной трапеции, площадь которой мы находим с помощью интегрирования, и сделаем точку a началом координат оси абсцисс (см.

Рис. 2.31). 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

$$\begin{cases} c_0 = f(a), \\ c_0 + c_1 \frac{w}{2} + c_2 \frac{w^2}{4} = f\left(\frac{a+b}{2}\right), \\ c_0 + c_1 w + c_2 w^2 = f(b). \end{cases} \quad (2.142)$$

Площадь, ограниченная графиком интерполяционного полинома $\check{P}_2(x)$, равна

$$\begin{aligned} \int_0^w (c_0 + c_1x + c_2x^2) \, dx &= c_0w + c_1 \frac{w^2}{2} + c_2 \frac{w^3}{3} \\ &= \frac{w}{6} (6c_0 + 3c_1w + 2c_2w^2). \end{aligned}$$

Фактически, для построения квадратурной формулы требуется подставить в полученное выражение c_0 , c_1 и c_2 , которые найдены в результате решения системы уравнений (2.142). Но можно выразить трёхчлен $6c_0 + 3c_1w + 2c_2w^2$ через значения подинтегральной функции f в узлах, не решая систему (2.142) явно.

Умножая второе уравнение системы (2.142) 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,

$$\begin{aligned} \int_0^w \check{P}_2(x) \, dx &= \int_0^w (c_0 + c_1x + c_2x^2) \, dx \\ &= \frac{w}{6} \left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right), \end{aligned}$$

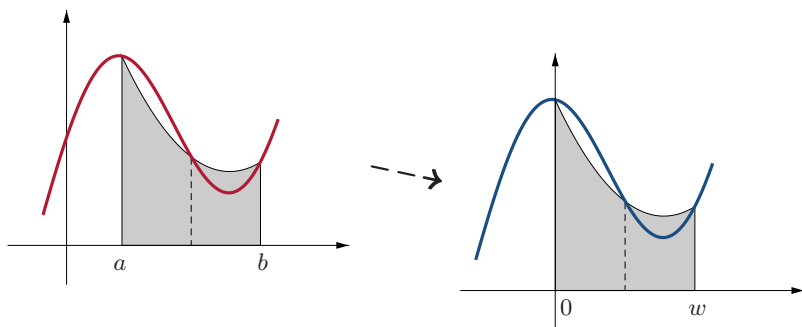


Рис. 2.31. Иллюстрация вывода квадратурной формулы Симпсона

что даёт приближённое равенство

$$\int_a^b f(x) dx \approx \frac{b-a}{6} \cdot \left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right). \quad (2.143)$$

Оно называется *квадратурной формулой Симпсона* или *формулой парабол* (см. Рис. 2.31), коль скоро основано на приближении подинтегральной функции подходящей параболой.²⁶

Читатель может самостоятельно убедиться, что та же самая формула получается (но только более длинно и утомительно) в результате интегрирования по $[a, b]$ интерполяционного полинома второй степени

²⁶Приведённый нами элегантный способ вывода формулы Симпсона заимствован из книги А.Н. Крылова [18], где он применяется даже к более общему случаю.

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 .

Предложение 2.12.1 (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 подынтегральной 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 §2.12б), желательно в соответствующем выражении использовать более высокую степень переменной. Иными словами, при оценке погрешности формулы Симпсона нужно взять для подынтегральной 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

$$H_3(a) = f(a), \quad H_3(b) = f(b), \quad (2.144)$$

$$H_3\left(\frac{a+b}{2}\right) = f\left(\frac{a+b}{2}\right), \quad H_3'\left(\frac{a+b}{2}\right) = f'\left(\frac{a+b}{2}\right), \quad (2.145)$$

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 (2.144)–(2.145) существует, и далее потребуется оценка его погрешности.

Существование и единственность решения подобных задач была установлена в §2.4, и там же обосновывается оценка его погрешности (2.49):

$$f(x) - H_m(x) = \frac{f^{(m+1)}(\xi(x))}{(m+1)!} \prod_{i=0}^n (x - x_i)^{N_i}, \quad (2.146)$$

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 (2.144)–(2.145) 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 (2.144)–(2.145) следуют равенства

$$\begin{aligned} \int_a^b H_3(x) dx &= \frac{b-a}{6} \cdot \left(H_3(a) + 4H_3\left(\frac{a+b}{2}\right) + H_3(b) \right) \\ &= \frac{b-a}{6} \cdot \left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right). \end{aligned} \quad (2.147)$$

Отсюда уже нетрудно вывести выражение для погрешности квадра-

турной формулы Симпсона:

$$\begin{aligned}
 R(f) &= \int_a^b f(x) \, dx - \frac{b-a}{6} \cdot \left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right) \\
 &= \int_a^b (f(x) - H_3(x)) \, dx \quad \text{в силу (2.147)} \\
 &= \int_a^b \frac{f^{(4)}(\xi(x))}{24} \cdot (x-a) \left(x - \frac{a+b}{2}\right)^2 (x-b) \, dx \quad \text{из (2.146)}.
 \end{aligned}$$

Из него следует оценка

$$\begin{aligned}
 |R(f)| &= \left| \int_a^b \frac{f^{(4)}(\xi(x))}{24} \cdot (x-a) \left(x - \frac{a+b}{2}\right)^2 (x-b) \, dx \right| \\
 &\leq \int_a^b \left| \frac{f^{(4)}(\xi(x))}{24} \right| \cdot \left| (x-a) \left(x - \frac{a+b}{2}\right)^2 (x-b) \right| \, dx \\
 &\leq \frac{M_4}{24} \cdot \left| \int_a^b (x-a) \left(x - \frac{a+b}{2}\right)^2 (x-b) \, dx \right|, \quad (2.148)
 \end{aligned}$$

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 (2.148), как обычно, обозначено $M_4 = \max_{x \in [a, b]} |f^{(4)}(x)|$.

Для вычисления интеграла из (2.148) 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

$$|R(f)| \leq \frac{M_4 (b-a)^5}{2880}.$$

Как видим, более тонкие рассуждения о свойствах формулы Симпсона позволили получить действительно более точную оценку её погрешности.

2.12г Общие интерполяционные квадратурные формулы

Напомним, что интерполяционными квадратурными формулами (или квадратурными формулами интерполяционного типа, см. §2.12б) мы назвали 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. 65).

Интерполяционная квадратурная формула должна получаться из приближённого равенства

$$\int_a^b f(x) dx \approx \int_a^b P_n(x) dx \quad (2.149)$$

в результате выполнения интегрирования в правой части. Как следствие, в представлении (2.137) весовые коэффициенты формулы имеют вид

$$\begin{aligned} c_k &= \int_a^b \phi_k(x) dx \\ &= \int_a^b \frac{(x - x_0) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_n)}{(x_k - x_0) \cdots (x_k - x_{k-1})(x_k - x_{k+1}) \cdots (x_k - x_n)} dx, \end{aligned} \quad (2.150)$$

$k = 0, 1, \dots, n$. 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 (2.150), 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 2.12.1 *Для того, чтобы квадратурная формула (2.137), 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 интерполантом.

Покажем достаточность: если квадратурная формула (2.137), построенная по $(n + 1)$ узлу, является точной для любого алгебраического полинома степени n , то её весовые коэффициенты вычисляются по

формулам (2.150), т. е. она является квадратурной формулой интерполяционного типа.

В самом деле, для базисных интерполяционных полиномов $\phi_i(x)$ выполнено свойство (2.9)

$$\phi_i(x_k) = \delta_{ik} = \begin{cases} 0, & \text{при } i \neq k, \\ 1, & \text{при } i = k, \end{cases}$$

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, имеет место равенства (2.150), что и требовалось доказать. ■

Следствие. Сумма весов интерполяционной квадратурной формулы равна ширине интервала интегрирования.

В самом деле, 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 утверждение.

Из (2.149) 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 §2.2д была получена оценка для $R_n(f, x)$ в форме Коши (2.28)

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

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

$$\begin{aligned} |R(f)| &\leq \frac{M_{n+1}}{(n+1)!} \int_a^b |\omega_n(x)| dx \\ &\leq \frac{M_{n+1} (b-a)^{n+2}}{(n+1)!}, \end{aligned} \quad (2.151)$$

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 (2.151), как видно из наших рассуждений, является простейшей, использующей лишь основные свойства алгебраического интерполанта. В некоторых случаях она может оказаться существенно завышенной, что мы могли видеть на примере формулы Симпсона.

Другое необходимое замечание состоит в том, что оценка (2.151), основанная на учёте погрешности 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 [30].

2.12д Дальнейшие формулы Ньютона-Котеса

В §2.12б и §2.12в 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 (2.150), 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

$$dx = h dt,$$

$$\begin{aligned} & (x - x_0^{(n)}) \cdots (x - x_{k-1}^{(n)}) (x - x_{k+1}^{(n)}) \cdots (x - x_n^{(n)}) \\ &= h^n t(t-1) \cdots (t-k+1)(t-k-1) \cdots (t-n), \\ & (x_k^{(n)} - x_0^{(n)}) \cdots (x_k^{(n)} - x_{k-1}^{(n)}) (x_k^{(n)} - x_{k+1}^{(n)}) \cdots (x_k^{(n)} - x_n^{(n)}) \\ &= (-1)^{n-k} h^n k!(n-k)!, \end{aligned}$$

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

$$B_k^{(n)} = \frac{(-1)^{n-k}}{n k!(n-k)!} \int_0^n t(t-1) \cdots (t-k+1)(t-k-1) \cdots (t-n) dt.$$

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 $n = 1$

$$B_0^{(1)} = - \int_0^1 (t-1) dt = - \left. \frac{(t-1)^2}{2} \right|_0^1 = \frac{1}{2},$$

$$B_1^{(1)} = \int_0^1 t dt = \left. \frac{t^2}{2} \right|_0^1 = \frac{1}{2}.$$

Мы вновь получили веса квадратурной формулы трапеций (2.140). 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 (2.143). 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. по этому поводу также [3, 25, 39, 82]).

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. 105). 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 (см. Рис. 2.32). Это снижает 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 году получил в работе [58] асимптотические формулы для коэффициентов Котеса²⁷, 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

²⁷Помимо оригинальной статьи Р.О. Кузьмина [58] эти асимптотические формулы излагаются, к примеру, в учебнике [20].

Таблица 2.2. Коэффициенты Котеса
для первых натуральных номеров

n	1	2	3	4	5	6	7	8	9	10
$k=0$	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{8}$	$\frac{7}{90}$	$\frac{19}{288}$	$\frac{41}{840}$	$\frac{751}{17280}$	$\frac{989}{28350}$	$\frac{2857}{89600}$	$\frac{16067}{598752}$
$k=1$	$\frac{1}{2}$	$\frac{4}{6}$	$\frac{3}{8}$	$\frac{16}{45}$	$\frac{25}{96}$	$\frac{9}{35}$	$\frac{3577}{17280}$	$\frac{5838}{28350}$	$\frac{15741}{89600}$	$\frac{106300}{598752}$
$k=2$		$\frac{1}{6}$	$\frac{3}{8}$	$\frac{2}{15}$	$\frac{25}{144}$	$\frac{9}{280}$	$\frac{1323}{17280}$	$-\frac{928}{28350}$	$\frac{1080}{89600}$	$-\frac{48525}{598752}$
$k=3$			$\frac{1}{8}$	$\frac{16}{45}$	$\frac{25}{144}$	$\frac{34}{105}$	$\frac{2989}{17280}$	$\frac{10496}{28350}$	$\frac{19344}{89600}$	$\frac{272400}{598752}$
$k=4$				$\frac{7}{90}$	$\frac{25}{96}$	$\frac{9}{280}$	$\frac{2989}{17280}$	$-\frac{4540}{28350}$	$\frac{5778}{89600}$	$-\frac{260550}{598752}$
$k=5$					$\frac{19}{288}$	$\frac{9}{35}$	$\frac{1323}{17280}$	$\frac{10496}{28350}$	$\frac{5778}{89600}$	$\frac{427368}{598752}$
$k=6$						$\frac{41}{840}$	$\frac{3577}{17280}$	$-\frac{928}{28350}$	$\frac{19344}{89600}$	$-\frac{260550}{598752}$
$k=7$							$\frac{751}{17280}$	$\frac{5838}{28350}$	$\frac{1080}{89600}$	$\frac{272400}{598752}$
$k=8$								$\frac{989}{28350}$	$\frac{15741}{89600}$	$-\frac{48525}{598752}$
$k=9$									$\frac{2857}{89600}$	$\frac{106300}{598752}$
$k=10$										$\frac{16067}{598752}$

сколь угодно велики (см. §2.12а). 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,

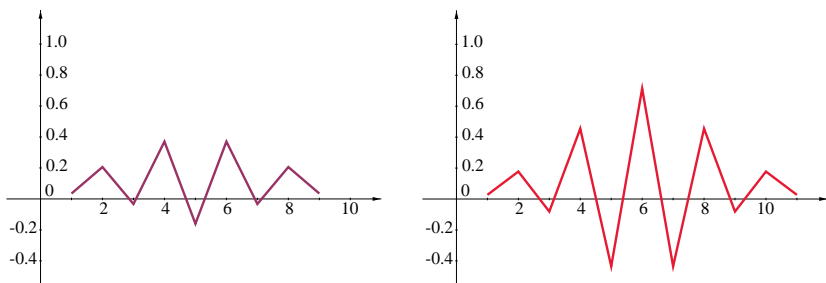


Рис. 2.32. Иллюстрация осцилляций коэффициентов Котеса для $n = 8$ (левый график) и $n = 10$ (правый график).

к примеру, в книгах [3, 20, 64]. 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 §2.13) и другим квадратурным формулам.

Из популярных квадратурных формул Ньютона-Котеса приведём ещё формулу «трёх восьмых», которая получается при замене подынтегральной функции интерполяционным полиномом 3-й степени:

$$\int_a^b f(x) dx \approx \frac{b-a}{8} \cdot \left(f(a) + 3f\left(\frac{2a+b}{3}\right) + 3f\left(\frac{a+2b}{3}\right) + f(b) \right).$$

Её погрешность оценивается как

$$|R(f)| \leq \frac{M_4 (b-a)^5}{6480},$$

где $M_4 = \max_{x \in [a,b]} |f^{(4)}(x)|$. Порядок точности этой формулы — такой же, как и у формулы Симпсона, хотя остаточный член существенно меньше. 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 [1, 3, 20].

2.13 Квадратурные формулы Гаусса

2.13а Задача оптимизации квадратурных формул

Параметрами квадратурной формулы (2.137)

$$\int_a^b f(x) dx \approx \sum_k c_k f(x_k)$$

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 (2.137) 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 корням полиномов Чебышёва существенно улучшает точность интерполирования (см. §2.3).

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 квадратурная формула, брать её алгебраическую степень точности (см. Определение 2.12.1).

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 (2.137), а как

$$\int_a^b f(x) dx \approx \sum_{k=1}^n c_k f(x_k), \quad (2.152)$$

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

$$\begin{cases} \sum_{k=1}^n c_k x_k^l = \frac{1}{l+1} (b^{l+1} - a^{l+1}), \\ l = 0, 1, 2, \dots, m, \end{cases}$$

или, в развёрнутом виде,

$$\begin{cases} c_1 + c_2 + \dots + c_n = b - a, \\ c_1 x_1 + c_2 x_2 + \dots + c_n x_n = \frac{1}{2}(b^2 - a^2), \\ \vdots \quad \quad \quad \ddots \quad \quad \quad \vdots \\ c_1 x_1^m + c_2 x_2^m + \dots + c_n x_n^m = \frac{1}{m+1}(b^{m+1} - a^{m+1}). \end{cases} \quad (2.153)$$

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 (2.153) переопределена, и в случае общего положения она оказывается неразрешимой.

Сделанное заключение можно обосновать строго.

Предложение 2.13.1 *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 (2.152). 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 (2.152) 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. Подынтегральная 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 (2.152) 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.

2.136 Простейшие квадратуры Гаусса

Перейдём к построению квадратурных формул Гаусса. При небольших n система уравнений (2.153) для узлов и весов может быть решена с помощью несложных аналитических преобразований.

Пусть $n = 1$, тогда $m = 2n - 1 = 1$, и система уравнений (2.153) принимает вид

$$\begin{cases} c_1 = b - a, \\ c_1 x_1 = \frac{1}{2}(b^2 - a^2). \end{cases}$$

Отсюда

$$\begin{aligned}c_1 &= b - a, \\x_1 &= \frac{1}{2c_1}(b^2 - a^2) = \frac{1}{2}(a + b).\end{aligned}$$

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. §2.126), 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 (2.153) 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

$$x_1 + x_2 = a + b. \quad (2.154)$$

Отсюда после возведения в квадрат имеем

$$x_1^2 + 2x_1x_2 + x_2^2 = a^2 + 2ab + b^2. \quad (2.155)$$

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 (2.155), будем иметь

$$x_1 x_2 = \frac{1}{6}(b^2 + 4ab + a^2). \quad (2.156)$$

Соотношения (2.154) и (2.156) на основе известной из элементарной алгебры теоремы Виета позволяют сделать вывод, что x_1 и x_2 являются корнями квадратного уравнения

$$x^2 - (a + b)x + \frac{1}{6}(b^2 + 4ab + a^2) = 0,$$

так что

$$x_{1,2} = \frac{1}{2}(a + b) \pm \frac{\sqrt{3}}{6}(b - a). \quad (2.157)$$

Удовлетворение полученными решениями четвёртого уравнения системы проверяется прямой подстановкой. Кроме того, поскольку

$$\frac{1}{2} > \frac{1}{2} \cdot \frac{1}{\sqrt{3}} = \frac{\sqrt{3}}{6},$$

то x_1 и x_2 действительно лежат на интервале $[a, b]$. В целом мы вывели *квадратурную формулу Гаусса с двумя узлами*

$$\int_a^b f(x) dx \approx \frac{b-a}{2} \cdot (f(x_1) + f(x_2)) \quad , \quad (2.158)$$

где x_1 и x_2 определяются посредством (2.157). Она очень похожа на формулу трапеций (2.140) и, фактически, является её модификацией. Согласно (2.158) приближённым значением интеграла тоже берётся площадь трапеции с высотой $(b - a)$, но основаниями, равными значениям интегрируемой функции в двух специально подобранных узлах.

Пример 2.13.1 Вычислим с помощью полученной выше формулы Гаусса с двумя узлами (2.158) 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 (2.157) и (2.158) имеем

$$\begin{aligned} \int_0^{\pi/2} \cos x \, dx &\approx \frac{\pi/2}{2} \cdot \left(\cos\left(\frac{\pi}{4} - \frac{\sqrt{3}}{6} \frac{\pi}{2}\right) + \cos\left(\frac{\pi}{4} + \frac{\sqrt{3}}{6} \frac{\pi}{2}\right) \right) \\ &= 0.998473. \end{aligned}$$

Формула Ньютона-Котеса с двумя узлами 0 и $\pi/2$ — формула трапеций — даёт для этого интеграла значение

$$\int_0^{\pi/2} \cos x \, dx \approx \frac{\pi}{2} \cdot \left(\cos 0 + \cos \frac{\pi}{2} \right) = 0.785398,$$

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 (2.143), использующая три узла — 0, $\pi/4$ и $\pi/2$, — приводит к результату

$$\begin{aligned} \int_0^{\pi/2} \cos x \, dx &\approx \frac{\pi/2}{6} \cdot \left(\cos 0 + 4 \cos \frac{\pi}{4} + \cos \frac{\pi}{2} \right) \\ &= \frac{\pi}{12} (1 + 2\sqrt{2}) = 1.00228, \end{aligned}$$

погрешность которого по порядку величины примерно равна погрешности ответа по формуле Гаусса (2.158), но всё-таки превосходит её в полтора раза. ■

С ростом n сложность системы уравнений (2.153) для узлов и весов формул Гаусса быстро нарастает, так что в общем случае не вполне ясно, будет ли она иметь вещественные решения при любом наперёд заданном n . Кроме того, эти решения системы (2.153), соответствующие узлам, должны быть различны и принадлежать интервалу интегрирования $[a, b]$.

Получение ответов на поставленные вопросы непосредственно из системы уравнений (2.153) в принципе возможно (см. учебник [10], Глава XVI, §9), но оно является громоздким и несколько искусственным. Мы рассмотрим другое, более элегантное решение задачи построения

формулы Гаусса, которое основано на расчленении общей задачи на отдельные подзадачи

- 1) построения узлов формулы и
- 2) вычисления её весовых коэффициентов.

Зная узлы формулы, можно подставить их в систему уравнений (2.153), 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 (2.153) (см., к примеру, [10]).

Другой способ решения подзадачи 2, когда узлы уже известны — это вычисление весовых коэффициентов по формулам (2.150), 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 2.12.1, 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 [3].

2.13в Выбор узлов для квадратурных формул Гаусса

Теорема 2.13.1 *Квадратурная формула (2.152)*

$$\int_a^b f(x) dx \approx \sum_{k=1}^n c_k f(x_k),$$

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 корнями такого полинома*

$$\omega(x) = (x - x_1)(x - x_2) \cdots (x - x_n),$$

что

$$\int_a^b \omega(x) q(x) dx = 0 \quad (2.159)$$

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 §2.10ж. На линейном пространстве $\mathcal{L}^2[a, b]$ интегрируемых с квадратом функций оно задаёт *скалярное произведение*, т.е. симметричную билинейную и положительно определённую форму. По этой причине утверждение Теоремы 2.13.1 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 , являлась точной на алгебраических полиномах степени $(2n - 1)$ необходимо и достаточно, чтобы эта формула была интерполяционной, а её узлы являлись корнями полинома, который в смысле $\mathcal{L}^2[a, b]$ 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 $n - 1$, то отсюда следует выполнение условия (б).

Справедливость условия (а) следует из Теоремы 2.12.1: если построенная по n узлам квадратурная формула (2.152) является точной для любого полинома степени не менее $n - 1$, то она — интерполяционная.

Достаточность. Пусть интерполяционная квадратурная формула построена по узлам $x_1, x_2, \dots, x_n \in [a, b]$, которые являются различными корнями полинома $\omega(x)$ степени n , удовлетворяющего условию ортогональности (2.159) 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

$$f(x) = \omega(x)q(x) + r(x), \quad (2.160)$$

в котором $q(x)$ и $r(x)$ — соответственно частное и остаток от деления $f(x)$ на $\omega(x)$. 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

$$\int_a^b f(x) dx = \int_a^b \omega(x)q(x) dx + \int_a^b r(x) dx = \int_a^b r(x) dx \quad (2.161)$$

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 2.12.1), в частности, на полиноме $r(x)$. Следовательно,

$$\begin{aligned} \int_a^b r(x) dx &= \sum_{k=1}^n c_k r(x_k) = \sum_{k=1}^n c_k (\omega(x_k)q(x_k) + r(x_k)) \\ &\quad \text{в силу равенств } \omega(x_k) = 0 \\ &= \sum_{k=1}^n c_k f(x_k), \quad \text{поскольку имеет место (2.160).} \end{aligned}$$

Сравнивая результаты этой выкладки с (2.161), 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 корнями некоторых специальных полиномов $\omega(x)$, удовлетворяющих условию (б) из Теоремы 2.13.1. В этих полиномах легко угадываются знакомые нам из §2.11 ортогональные полиномы, которые являются полиномами Лежандра для случая $[a, b] = [-1, 1]$ или соответствующим образом преобразованы из них для любого другого интервала интегрирования $[a, b]$.

2.13г 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 §2.11 мы указали рецепт построения из полиномов Лежандра полиномов, ортогональных с единичным весом, для любого интервала вещественной оси. Этой техникой и нужно воспользоваться в данном случае.

Итак, пусть

$$x = \frac{1}{2}(a + b) + \frac{1}{2}(b - a)y. \quad (2.162)$$

Переменная x будет пробегать интервал $[a, b]$, когда y изменяется в $[-1, 1]$. Если y_i , $i = 1, 2, \dots, n$, — корни полинома Лежандра, которые согласно Предложению 2.11.2 все различны и лежат на интервале $[-1, 1]$, то узлами квадратурной формулы Гаусса для интервала интегрирования $[a, b]$ являются

$$x_i = \frac{1}{2}(a + b) + \frac{1}{2}(b - a)y_i, \quad i = 1, 2, \dots, n. \quad (2.163)$$

Все они также различны и лежат на интервале интегрирования $[a, b]$.

Далее, веса c_k любой интерполяционной квадратурной формулы могут быть выражены в виде интегралов (2.150). В случае формул Гаусса (когда узлы нумеруются с единицы) они принимают вид

$$c_k = \int_a^b \phi_k(x) dx, \quad k = 1, 2, \dots, n,$$

где $\phi_k(x)$ — k -ый базисный интерполяционный полином Лагранжа (см. стр. 65), построенный по узлам (2.163):

$$\phi_i(x) = \frac{(x - x_1) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n)}{(x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)}.$$

Тогда, выполняя замену переменных (2.162), 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 -ый базисный полином Лагранжа, построенный по узлам y_i , $i = 1, 2, \dots, n$, которые являются корнями n -го полинома Лежандра. Мы обосновали

Предложение 2.13.2 *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. корни полиномов Лежандра) и их веса тщательно заатабулированы 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 [3, 64] и в специальных журнальных статьях. В частности, оказывается, что весовые коэффициенты формулы Гаусса с n узлами выражаются как

$$c_k = \frac{2}{(1 - x_k^2)(L'_n(x_k))^2}, \quad k = 1, 2, \dots, n,$$

где $L_n(x)$ — n -ый полином Лежандра в виде, даваемом формулой Родрига (2.126). Эта формула была впервые получена Э.Б. Кристоффелем в середине XIX века [86], 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 [2, 3, 10, 18, 19, 28] или в специализированных справочниках, например, в [39, 57]. В частности, в учебнике [3] 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 [39, 57] — с 20 значащими цифрами вплоть до $n = 96$ и $n = 48$. Таким образом, практическое применение квадратур Гаусса обычно не встречает затруднений.

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

При небольших значениях n можно дать точные аналитические вы-

ражения для узлов формул Гаусса, как корней полиномов Лежандра $L_n(x)$, имеющих явные представления (2.129). Так, 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 для $n = 3$ равны

$$x_1 = -\sqrt{\frac{3}{5}} = -0.77459\ 66692\ 41483\dots,$$

$$x_2 = 0,$$

$$x_3 = \sqrt{\frac{3}{5}} = 0.77459\ 66692\ 41483\dots$$

Для $n = 4$

$$L_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3),$$

и нахождение корней 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 решением небольших систем линейных уравнений, к которым редуцируется система (2.153) 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.

2.13д Погрешность квадратур Гаусса

Для исследования остаточного члена квадратурных формул Гаусса предположим, что подинтегральная функция $f(x)$ имеет достаточно высокую гладкость. 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 $2n - 1$, то для адекватного учёта этого факта степень полинома, интерполирующего подинтегральную функцию, тоже нужно взять равной $2n - 1$. Имея всего n узлов, мы находимся в ситуации, совершенно аналогичной той, что встрети́лась нам при анализе формулы Симпсона. Необходимая степень интерполяционного полинома для формулы Гаусса получится, если рассматривать на интервале интегрирования интерполяцию с кратными узлами. 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 (2.49) погрешность интерполирования 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 интерполирования $[a, b]$. 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

$$\begin{aligned} \int_a^b f(x) dx &= \int_a^b (H_{2n-1}(x) + R_{2n-1}(f, x)) dx \\ &= \int_a^b H_{2n-1}(x) dx + \int_a^b R_{2n-1}(f, x) dx \\ &= \sum_{i=1}^n c_i H_{2n-1}(x_i) + \int_a^b R_{2n-1}(f, x) dx \end{aligned}$$

так как формула точна на полиноме $H_{2n-1}(x)$

$$= \sum_{i=1}^n c_i f(x_i) + \frac{1}{(2n)!} \int_a^b f^{(2n)}(\xi(x)) (\omega(x))^2 dx.$$

Выражение для второго слагаемого последней суммы, т. е. для остаточного члена квадратуры, можно упростить, приняв во внимание зна-

копостоянство множителя $(\omega(x))^2$. В силу интегральной теоремы о среднем (см., к примеру, [12, 38]) 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 корни 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 (2.40)

$$x = \frac{2y - (b + a)}{(b - a)}.$$

Из-за того, что у полинома $\omega(x)$ старший коэффициент — единица, для вычисления нашего интеграла удобнее воспользоваться приведёнными полиномами Лежандра (2.130), скорректировав результат Предложения 2.11.1 с учётом соотношения (2.131). 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

$$R(f) = \frac{(n!)^4}{(2n+1)((2n)!)^3} (b-a)^{2n+1} f^{(2n)}(\theta), \quad (2.164)$$

где θ — 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 (2.157)–(2.158) с двумя узлами

$$|R_2(f)| \leq \frac{M_4(b-a)^5}{4320},$$

что даже лучше оценки погрешности для формулы Симпсона. На практике мы могли видеть это в Примере 2.13.1.

Отметим, что выведенная оценка (2.164) справедлива лишь при достаточной 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 (2.164) числитель числового коэффициента, т. е. $(n!)^4 (b-a)^{2n+1}$, с ростом n может быть сделан сколь угодно меньшим знаменателя $(2n+1)((2n)!)^3$. 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 сходимости, не зависящим от гладкости исходных данных (стр. 121). Таким образом, квадратурные формулы Гаусса дают пример *ненасыщаемого* численного метода, порядок точности которого может быть сделан любым в зависимости от того, насколько гладкими являются входные данные для этого метода.

2.13е Метод неопределённых коэффициентов

Снова обратимся к задаче построения квадратурной формулы

$$\int_a^b f(x) dx \approx \sum_{k=0}^n c_k f(x_k), \quad (2.137)$$

по заданным узлам x_0, x_1, \dots, x_n (они, как и в общем случае, нумеруются с нуля). Эта задача имеет большое практическое значение, так как часто расположение узлов квадратуры жёстко задаётся из технических, экономических и прочих соображений.

Если на квадратурную формулу наложить условие, что она — интерполяционная, то, как мы видели в разделе 2.12г, весовые коэффициенты c_k могут быть вычислены по формулам (2.150), как интегралы от базисных интерполяционных полиномов Лагранжа. Но это не единственный возможный способ определения весов.

Весовые коэффициенты c_k можно найти из условия зануления погрешности равенства (2.137) для какого-то «достаточно представительного» набора несложно интегрируемых функций $f_i(x)$, $i = 1, 2, \dots$. Каждое отдельное равенство является уравнением на неизвестные c_0, c_1, \dots, c_n , и потому, выписав достаточное число подобных уравнений, мы получим систему вида (2.153). Она линейна относительно c_0, c_1, \dots, c_n , и, решив её, мы сможем определить желаемые веса, т. е. построить квадратурную формулу (2.137). В этом — суть *метода неопределённых коэффициентов*. Он идейно похож, таким образом, на метод неопределённых коэффициентов для построения формул численного дифференцирования из §2.8в.

В качестве пробных функций $f_p(x)$, $p = 1, 2, \dots$ часто берут алгебраические полиномы. Для равномерно расположенных узлов при этом получаются знакомые нам квадратурные формулы Ньютона-Котеса.

Продemonстрируем работу метода неопределённых коэффициентов для тригонометрических полиномов

$$1, \quad \sin(px), \quad \cos(px), \quad p = 1, 2, \dots$$

В заключение темы следует сказать, что на практике нередко требуется включение во множество узлов квадратурной формулы каких-либо фиксированных точек интервала интегрирования. Это могут быть, к примеру, его концы (один или оба), либо какие-то выделенные внутренние точки. С другой стороны, мы готовы допустить некоторое снижение алгебраической степени точности (и без того весьма высокой для

формул Гаусса). Основная идея формул Гаусса может быть с успехом применена к построению таких квадратурных формул, которые называются *квадратурами Маркова* [3, 13, 64, 28] (иногда используют также термин *квадратуры Лобатто* [2, 39]).

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 являются наиболее адекватным мерилем качества квадратурной формулы. Например, можно развивать квадратуры наивысшей *тригонометрической степени точности*, которые окажутся практичнее при вычислении интегралов от осциллирующих и периодических функций [28, 69].

2.14 Составные квадратурные формулы

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 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 искомым интеграл приближается их суммой.²⁸

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

²⁸Интересно, что при этом подинтегральная функция, фактически, интерполируется на всём интервале интегрирования $[a, b]$ при помощи сплайна.

$[a, r_1], [r_1, r_2], \dots, [r_{N-1}, b]$. 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

$$\int_a^b f(x) dx \approx \sum_{i=0}^{N-1} \mathcal{J}_i. \quad (2.165)$$

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

$$|R(f)| \leq K(b-a)^p, \quad (2.166)$$

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, зависят косвенно 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, так что в наших обозначениях $K_{[a_1, b_1]} \leq K_{[a_2, b_2]}$, 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 (2.165) можно оценить сверху суммой погрешностей отдельных слагаемых (см. Предложение 1.2.1). Поэтому

$$\begin{aligned} |\tilde{R}(f)| &\leq \sum_{i=0}^{N-1} K_{[r_i, r_{i+1}]} \left(\frac{b-a}{N} \right)^p \leq \sum_{i=0}^{N-1} K_{[a, b]} \left(\frac{b-a}{N} \right)^p \\ &= N K_{[a, b]} \left(\frac{b-a}{N} \right)^p = \frac{K_{[a, b]} (b-a)^p}{N^{p-1}} \\ &= K_{[a, b]} (b-a) h^{p-1}. \end{aligned} \quad (2.167)$$

Так как $p-1 > 0$, то оценка погрешности (2.167) уменьшилась в N^{p-1} раз по сравнению с (2.166). И потенциально таким способом погрешность вычисления интеграла можно сделать сколь угодно малой.

В соответствии с Определением 2.8.1 (стр. 137) 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 анализ погрешности проводить тогда труднее.

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 $[r_0, r_1], [r_1, r_2], \dots, [r_{N-1}, r_N]$ длины $h = (b-a)/N$ каждая, в котором $a = r_0$ и $r_N = b$.

Составная формула средних прямоугольников

$$\int_a^b f(x) dx \approx h \sum_{i=1}^N f(r_{i-1/2}),$$

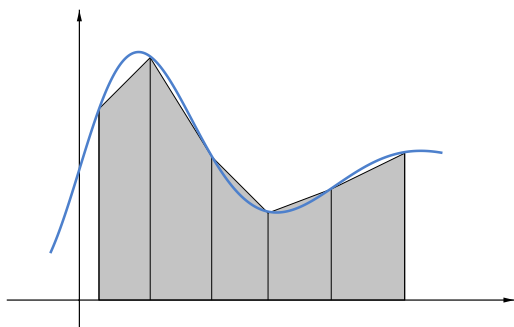


Рис. 2.33. Составная квадратурная формула трапеций.

где $r_{i-1/2} = r_i - h/2$. 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

$$\int_a^b f(x) dx \approx h \left(\frac{1}{2}f(a) + \sum_{i=1}^{N-1} f(r_i) + \frac{1}{2}f(b) \right).$$

Её полная погрешность

$$|\tilde{R}(f)| \leq M_2 \frac{(b-a)h^2}{12},$$

т.е. порядок точности тоже второй.

Составная формула Симпсона (формула парабол)

$$\int_a^b f(x) dx \approx \frac{h}{6} \sum_{i=1}^N (f(r_{i-1}) + 4f(r_{i-1/2}) + f(r_i)),$$

где $r_{i-1/2} = r_i - h/2$. Её полная погрешность

$$|\tilde{R}(f)| \leq M_4 \frac{(b-a)h^4}{2880},$$

т. е. формула имеет четвёртый порядок точности.

Аналогично конструируются и исследуются составные квадратурные формулы Гаусса, но мы не будем здесь развёртывать детали этого несложного построения. Отметим лишь, что вследствие оценок (2.164) и (2.167) порядок точности составной квадратурной формулы Гаусса по n узлам равен $2n$.

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.

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 порядка m и не точна для полиномов порядка $m + 1$.

Предложение 2.14.1 *Тригонометрическая степень точности квадратурной формулы, построенной по n узлам, не превосходит $n - 1$.*

Доказательство опускается, читатель может найти его в публикациях [69, 28]. Сравнивая этот результат с Предложением 2.13.1, можем видеть существенно более сильные ограничения на тригонометрическую степень точности квадратуры, чем на алгебраическую степень точности. Справедлив замечательный результат:

Теорема И.П. Мысовских [69] *Составная квадратурная формула прямоугольников*

$$\int_0^{2\pi} f(x) dx \approx \frac{2\pi}{n} \sum_{k=1}^n f\left(\tilde{x} + \frac{2\pi(k-1)}{n}\right),$$

где \tilde{x} — произвольная точка из подинтервала $[0, 2\pi/n]$,

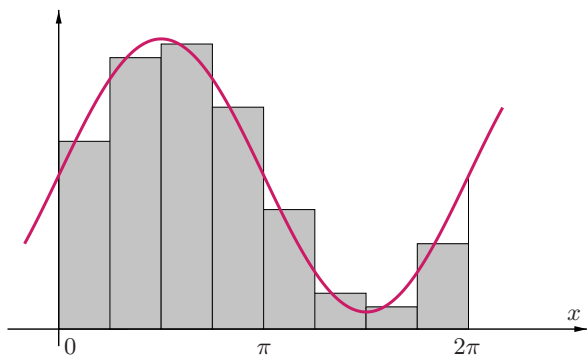


Рис. 2.34. Иллюстрация к теореме И.П. Мысовских: составная квадратурная формула прямоугольников для синусоиды

и только она является квадратурной формулой наивысшей тригонометрической степени точности $n - 1$.

Доказательство теоремы нетривиально, его можно увидеть в оригинальной статье [69] или в книге [28].²⁹ 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 (см. Рис. 2.34).

Идея разбиения области интегрирования на более мелкие части для повышения точности вычисления интеграла применима и 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 интегрирова-

²⁹Существуют результаты о методе Ньютона для решения нелинейных уравнений, которые иногда тоже называют «теоремами Мысовских».

ния 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. §2.16).

2.15 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

$$\begin{pmatrix} x_0^{(1)} & & & \cdots \\ x_0^{(2)} & x_1^{(2)} & & 0 & \cdots \\ x_0^{(3)} & x_1^{(3)} & x_2^{(3)} & & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}, \quad (2.168)$$

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

$$\begin{pmatrix} c_0^{(1)} & & & \cdots \\ c_0^{(2)} & c_1^{(2)} & & 0 & \cdots \\ c_0^{(3)} & c_1^{(3)} & c_2^{(3)} & & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}. \quad (2.169)$$

В случае задания бесконечных треугольных матриц (2.168)–(2.169), по которым организуется приближённое вычисление интегралов на после-

довательности сеток, будем говорить, что на интервале $[a, b]$ для функции f определён *квадратурный процесс*.

Определение 2.15.1 *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 (2.168)–(2.169), 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 предыдущем §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 [77], а впоследствии Д. Пойа [91] доказал также необходимость условий В.А. Стеклова.

Теорема 2.15.1 (теорема Стеклова-Пойа) *Квадратурный процесс, порождаемый матрицами узлов и весов (2.168)–(2.169), сходится для любой непрерывной на $[a, b]$ функции тогда и только тогда, когда*

- (1) *этот процесс сходится для полиномов,*

- (2) суммы абсолютных значений весов ограничены равномерно по n , т. е. существует такая константа C , что

$$\sum_{k=0}^n |c_k^{(n)}| \leq C \quad (2.170)$$

для всех $n = 0, 1, 2, \dots$

Покажем достаточность условий теоремы Стеклова-Пойа. С этой целью, задавшись каким-то $\epsilon > 0$, найдём полином $P_N(x)$, который равномерно с погрешностью ϵ 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 формулы:

$$\begin{aligned} R_n(f) &= \int_a^b f(x) dx - \sum_{k=0}^n c_k^{(n)} f(x_k^{(n)}) \\ &= \int_a^b (f(x) - P_N(x)) dx + \int_a^b P_N(x) dx - \sum_{k=0}^n c_k^{(n)} f(x_k^{(n)}) \\ &= \int_a^b (f(x) - P_N(x)) dx \\ &\quad + \left(\int_a^b P_N(x) dx - \sum_{k=0}^n c_k^{(n)} P_N(x_k^{(n)}) \right) \\ &\quad + \sum_{k=0}^n c_k^{(n)} \left(P_N(x_k^{(n)}) - f(x_k^{(n)}) \right). \end{aligned}$$

Отдельные слагаемые полученной суммы, расположенные выше в различных строчках, оцениваются при достаточно больших номерах n

следующим образом:

$$\left| \int_a^b (f(x) - P_N(x)) \, dx \right| \leq \epsilon (b - a), \quad \text{так как } P_N(x) \text{ приближает } f(x) \\ \text{равномерно с погрешностью } \epsilon \\ \text{на интервале } [a, b];$$

$$\left| \int_a^b P_N(x) \, dx - \sum_{k=0}^n c_k^{(n)} P_N(x_k^{(n)}) \right| \leq \epsilon, \quad \text{так как квадратуры} \\ \text{сходятся на полиномах};$$

$$\left| \sum_{k=0}^n c_k^{(n)} \left(P_N(x_k^{(n)}) - f(x_k^{(n)}) \right) \right| \leq \epsilon \sum_{k=0}^n |c_k^{(n)}| \leq \epsilon C \quad \text{в силу (2.170)}.$$

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-Пойа помимо оригинальной статьи [91] можно найти в книгах [3, 29].

В формулировке теоремы фигурирует величина (2.138)

$$\sum_{k=0}^n |c_k|, \quad ((2.138))$$

— сумма абсолютных значений весов, которая, как мы видели в §2.12а, является коэффициентом увеличения погрешности в данных и играет очень важную роль при оценке качества различных квадратурных формул. В §2.12д уже упоминался результат Р.О. Кузьмина [58] о том, что для формул Ньютона-Котеса величина (2.138) неограниченно увеличивается с ростом числа узлов n . Как следствие, на произвольных непрерывных функциях эти квадратурные формулы сходимостью не обладают.

Для квадратурных формул Гаусса ситуация другая. Справедливо

Предложение 2.15.1 *Весовые коэффициенты квадратурных формул Гаусса положительны.*

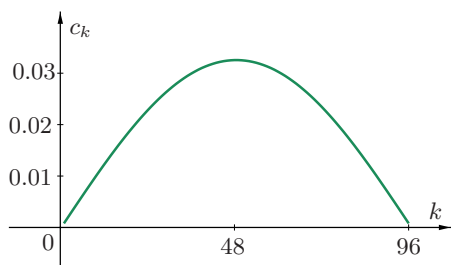


Рис. 2.35. Зависимость весовых коэффициентов от номера для квадратуры Гаусса 96-го порядка (числовые данные взяты из справочника [39])

Иллюстрацией этого утверждения может служить Рис. 2.35. 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).

Доказательство. Ранее мы уже выводили для весов интерполяционных квадратурных формул выражение (2.150). 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 $2n - 1$, то для полинома

$$P_i(x) = ((x - x_1) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n))^2$$

степени $2(n - 1)$ должно выполняться точное равенство

$$\int_a^b P_i(x) dx = \sum_{k=1}^n c_k P_i(x_k). \quad (2.171)$$

Но $P_i(x_k) = \delta_{ik}$ по построению полинома P_i , так что от суммы справа в (2.171) остаётся лишь одно слагаемое $c_i P_i(x_i)$:

$$\int_a^b P_i(x) dx = c_i P_i(x_i).$$

Следовательно,

$$c_i = \int_a^b P_i(x) dx / P_i(x_i).$$

Далее, $P_i(x) > 0$ всюду на интервале интегрирования $[a, b]$ за исключением конечного числа точек, и потому положителен интеграл в числителе выписанного выражения. Кроме того, $P_i(x_i) > 0$, 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. §2.12г). Следовательно, величина (2.138) при этом ограничена, и квадратурный процесс по формулам Гаусса всегда сходится.

Завершая тему, можно отметить, что ситуация со сходимостью квадратур оказывается в целом более благоприятной, чем для интерполяционных процессов.

2.16 Вычисление интегралов методом Монте-Карло

В *методе Монте-Карло*, называемом также *методом статистического моделирования*, искомое решение задачи представляется в виде какой-либо характеристики специально построенного случайного процесса. Затем этот процесс моделируется, с помощью ЭВМ или какими-то другими средствами, и по его реализациям мы вычисляем нужную характеристику, т.е. решение задачи. Наиболее часто решение задач представляется так называемым математическим ожиданием (средним значением) специально подобранной случайной величины.

В качестве примера рассмотрим задачу вычисления определённого интеграла

$$\int_a^b f(x) dx \tag{2.172}$$

от непрерывной функции $f(x)$. Согласно известной из интегрального исчисления теореме о среднем (см., к примеру, [12, 38])

$$\int_a^b f(x) dx = (b - a) f(c)$$

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 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», т. е. обнаруживают закономерности поведения в больших сериях повторяющихся испытаний. Одними из основных понятий теории вероятностей являются понятия *вероятности*, *случайной величины* и её *функции распределения*. Случайной величиной называется переменная величина, значения которой зависят от случая и для которой определена так называемая функция распределения вероятностей. Вероятность — это величина, выражающая относительную частоту интересующего нас события, которая обычно устанавливается в большой серии испытаний. Функция распределения показывает, следовательно, вероятность появления тех или иных значений этой случайной величины. Конкретное значение, которое случайная величина принимает в результате отдельного опыта, обычно называют реализацией случай-

ной 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 длину, попадает примерно одинаковое количество точек.

На этом пути мы и приходим к простейшему методу Монте-Карло для вычисления определённого интеграла (2.172):

$$\begin{aligned} & \text{фиксируем натуральное число } N; \\ & \text{организуем реализации } \xi_i, i = 1, 2, \dots, N, \text{ для} \\ & \text{случайной величины } \xi, \text{ имеющей равномерное} \\ & \text{распределение на интервале } [a, b]; \\ & (\text{искомый интеграл}) \leftarrow \frac{b-a}{N} \cdot \sum_{i=1}^N f(\xi_i); \end{aligned} \quad (2.173)$$

Получение равномерно распределённой случайной величины (как и других случайных распределений) является не вполне тривиальной 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 интервале.

Рассмотрим теперь задачу определения площади фигуры с криволинейными границами (Рис. 2.36). 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 объёмлющего 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.

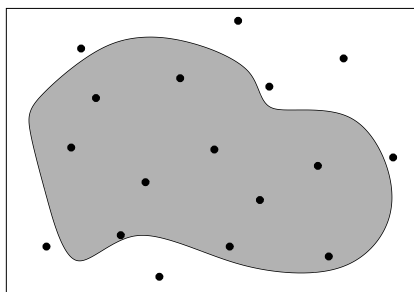


Рис. 2.36. Вычисление объёма области методом Монте-Карло.

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 (см. Рис. 2.37). 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 найденной относительной частоты на площадь большого прямоугольника.

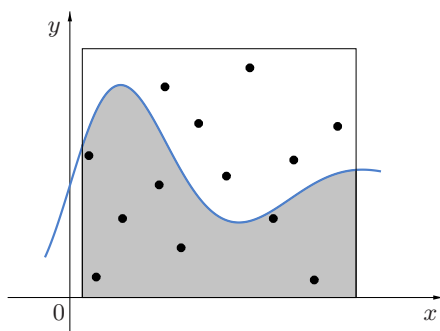


Рис. 2.37. Один из способов приближённого вычисления определённого интеграла методом Монте-Карло

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.

Результаты вычислений по методу Монте-Карло сами являются случайной величиной, и два результата различных решений одной и той же задачи интегрирования, полученные описанными выше способами, вообще говоря, могут отличаться друг от друга. Можно показать, 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 (называемый в теории вероятностей *дисперсией*) у него больше [32].

Сформулированные выше идеи и основанные на них алгоритмы в действительности применимы для интегрирования функций от произвольного количества переменных. Вероятностные оценки погрешности оказываются пропорциональными $N^{-1/2}$, 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 уже превосходят по своей эффективности классические кубатурные формулы и являются сегодня основным методом вычисления многомерных интегралов.

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 практики можно отсчитывать с конца XVIII века, когда Ж.-Л. Бюффон (в 1777 году) предложил способ определения числа π с помощью случайных бросаний иглы на бумагу, разграфлённую параллельными линиями.³⁰ 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.

В 1944 году в связи с работами по созданию атомной бомбы в США,

³⁰Наиболее известная «докомпьютерная» реализация метода Бюффона была осуществлена американским астрономом А. Холлом [87].

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.³¹ 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, [32, 66] и цитированную там литературу). Хороший популярный очерк методов Монте-Карло читатель может найти в книге [70].

2.17 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-либо задачу, где фигурирует сетка на интервале вещественной оси или в пространстве бóльшего 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 интерполирования 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. Аналогична ситуация и с другими задачами вычислительной математики и погрешностями их решения.

К. Рунге принадлежит идея использовать для определения значения константы C результаты нескольких расчётов на различных сетках. Далее, после того как величина C будет определена, мы можем

³¹Интересно, что примерно в те же самые годы в СССР решение аналогичных задач советского атомного проекта было успешно выполнено другими методами.

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{J}_h и $\mathcal{J}_{h/2}$, а точное решение есть \mathcal{J} . Тогда

$$\begin{aligned}\mathcal{J}_h - \mathcal{J} &\approx Ch^p, \\ \mathcal{J}_{h/2} - \mathcal{J} &\approx C \left(\frac{h}{2}\right)^p = C \frac{h^p}{2^p}.\end{aligned}$$

Вычитая второе равенство из первого, получим

$$\mathcal{J}_h - \mathcal{J}_{h/2} \approx Ch^p - C \frac{h^p}{2^p} = Ch^p \frac{2^p - 1}{2^p},$$

так что

$$C \approx \frac{2^p}{2^p - 1} \cdot \frac{\mathcal{J}_h - \mathcal{J}_{h/2}}{h^p}.$$

Зная константу C , можно уже находить оценку погрешности рассчитанных решений \mathcal{J}_h , $\mathcal{J}_{h/2}$ или любых других

Правило Рунге работает плохо, если главный член погрешности Ch^p не доминирует над последующими членами её разложения, которые соответствуют $(p+1)$ -ой и более высоким степеням шага сетки h . Это происходит, как правило, для сильно меняющихся решений.

Пример 2.17.1



Литература к главе 2

Основная

- [1] БАРАХНИН В.Б., ШАПЕЕВ В.П. *Введение в численный анализ*. – Санкт-Петербург–Москва–Краснодар: Лань, 2005.
- [2] БАХВАЛОВ Н.С., ЖИДКОВ Н.П., КОВЕЛЬКОВ Г.М. *Численные методы*. – Москва: Бином, 2003, а также другие издания этой книги.
- [3] БЕРЕЗИН И.С., ЖИДКОВ Н.П. *Методы вычислений*. Т. 1–2. – Москва: Наука, 1966.
- [4] БРАДИС В.М. *Четырехзначные математические таблицы*. – Москва: Дрофа, 2010, а также более ранние издания.

- [5] Вержвицкий В.М. *Численные методы. Части 1–2.* – Москва: «Оникс 21 век», 2005.
- [6] Волков Е.А. *Численные методы.* – Москва: Наука, 1987.
- [7] Гавриков М.Б., Таюрский А.А. *Функциональный анализ и вычислительная математика.* – Москва: URSS, Ленанд, 2016.
- [8] Гончаров В.Л. *Теория интерполирования и приближения функций.* – Москва: ГИТТЛ, 1954.
- [9] Даугавет И.К. *Введение в теорию приближения функций.* – Ленинград: Издательство Ленинградского университета, 1977.
- [10] Демидович Б.П., Марон А.А. *Основы вычислительной математики.* – Москва: Наука, 1970.
- [11] Завьялов Ю.С., Квасов Б.И., Мирошниченко В.Л. *Методы сплайн-функций.* – Москва: Наука, 1980.
- [12] Зорич В.А. *Математический анализ.* Т. 1. – Москва: Наука, 1981. Т. 2. – Москва: Наука, 1984, а также более поздние издания.
- [13] Калиткин Н.Н. *Численные методы.* – Москва: 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.* – Москва: Мир, 1969.
- [16] Колмогоров А.Н., Фомин С.В. *Элементы теории функций и функционального анализа.* – Москва: Наука, 1976, а также более поздние издания.
- [17] Кострикин А.Н. *Введение в алгебру. Часть 1. Основы алгебры.* – Москва: Физматлит, 2001.
- [18] Крылов А.Н. *Лекции о приближённых вычислениях.* – Москва: ГИТТЛ, 1954, а также более ранние издания.
- [19] Крылов В.И. *Приближённое вычисление интегралов.* – Москва: Наука, 1967.
- [20] Крылов В.И., Бобков В.В., Монастырный П.И. *Вычислительные методы. Т. 1–2.* – Москва: Наука, 1976.
- [21] Кунц К.С. *Численный анализ.* – Киев: Техника, 1964.
- [22] Курош А.Г. *Курс высшей алгебры.* – Москва: Наука, 1968.
- [23] Люстерник Л.А., Червоненкис О.А., Янпольский А.Р. *Математический анализ. Вычисление элементарных функций.* – Москва: ГИФМЛ, 1963.
- [24] Мак-Кракен Д., Дорн У. *Численные методы и программирование на ФОРТРАНе.* – Москва: 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*. – Санкт-Петербург: Издательство Санкт-Петербургского университета, 1998.
- [29] Натансон И.П. *Конструктивная теория функций*. – Москва–Ленинград: ГИТТЛ, 1949.
- [30] Никольский С.М. *Квадратурные формулы*. – Москва: Наука, 1988.
- [31] Самарский А.А., Гулин А.В. *Численные методы*. – Москва: Наука, 1989.
- [32] Соболев И.М. *Численные методы Монте-Карло*. – Москва: Наука, 1973.
- [33] Стечкин С.Б., Субботин Ю.Н. *Сплайны в вычислительной математике*. – Москва: Наука, 1976.
- [34] Тихонов А.Н., Арсенин В.Я. *Методы решения некорректных задач*. – Москва: Наука, 1979.
- [35] Тыртышников Е.Е. *Матричный анализ и линейная алгебра*. – Москва: Физматлит, 2007.
- [36] Тыртышников Е.Е. *Методы численного анализа*. – Москва: Академия, 2007.
- [37] Уиттекер Э., Робинсон Г. *Математическая обработка результатов наблюдений*. – Ленинград-Москва: ГТТИ, 1933.
- [38] Фихтенгольц Г.М. *Курс дифференциального и интегрального исчисления. Т. 1, 2*. – Москва: Наука, 1966; Москва: ФИЗМАТЛИТ, 2001; Санкт-Петербург: Лань, 2017.

Дополнительная

- [39] Абрамовиц М., Стиган И. *Таблицы специальных функций*. – Москва: Наука, 1979.
- [40] Алберг Дж., Нильсон Э., Уолш Дж. *Теория сплайнов и её приложения*. – Москва: Мир, 1972.
- [41] Ахиезер Н.И. *Лекции по теории аппроксимации*. – Москва: Наука, 1965.
- [42] Бабенко К.И. *Основы численного анализа*. – Москва: Наука, 1986.
- [43] Бахвалов Н.С., Корнев А.А., Чижонков Е.В. *Численные методы. Решение задач и упражнения*. – Москва: Дрофа, 2008.
- [44] Бердышев В.И., Петрак Л.В. *Аппроксимация функций, сжатие численной информации, приложения*. – Екатеринбург: УрО РАН, 1999.
- [45] 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.
- [46] Гельфонд А.О. *Исчисление конечных разностей*. – Москва: Наука, 1967, а также более поздние репринтные издания.
- [47] Геронимус Я.Л. *Теория ортогональных многочленов*. – Москва: Госуд. изд-во технико-теоретической литературы, 1950.
- [48] Гурвиц А., Курант Р. *Теория функций*. – Москва: Наука, Физматлит, 1968.

- [49] де Бор К. *Практическое руководство по сплайнам*. – Москва: Радио и связь, 1985.
- [50] Демиденко Е.З. *Оптимизация и регрессия*. – Москва: Наука, 1989.
- [51] Дробышевский В.И., Дымников В.П., Ривин Г.С. *Задачи по вычислительной математике*. – Москва: Наука, 1980.
- [52] КАХАНЕР Д., МОУЛЕР К., НЭШ С. *Численные методы и программное обеспечение*. – Москва: Мир, 1998.
- [53] КВАСОВ Б.И. *Методы изогеоиметрической аппроксимации сплайнами*. – Москва: Физматлит, 2006.
- [54] КОЛЛАТЦ Л., КРАВС В. *Теория приближений. Чебышёвские приближения и их приложения*. – Москва: Наука, 1978.
- [55] КОРНЕЙЧУК Н.П. *Сплайны в теории приближения*. – Москва: Наука, 1984.
- [56] КРОНРОД А.С. *Узлы и веса квадратурных формул. Шестнадцатизначные таблицы*. – Москва: Наука, 1964.
- [57] КРЫЛОВ В.И., ШУЛЬГИНА Л.Т. *Справочная книга по численному интегрированию*. – Москва: Наука, 1966.
- [58] КУЗЬМИН Р.О. К теории механических квадратур // *Известия Ленинградского политехнического института им. М.И. Калинина*. – 1931. – Т. 33. – С. 5–14.
- [59] ЛАНС Дж.Н. *Численные методы для быстродействующих вычислительных машин*. – Москва: Издательство иностранной литературы, 1962.
- [60] ЛИННИК Ю.В. *Метод наименьших квадратов и основы теории обработки наблюдений*. 2-е изд. – Москва: ГИФМЛ, 1962.
- [61] ЛОКУЦИЕВСКИЙ О.В., ГАВРИКОВ М.Б. *Начала численного анализа*. – Москва: ТОО «Янус», 1994.
- [62] ЛОРАН Ж.-П. *Аппроксимация и оптимизация*. – Москва: Мир, 1975.
- [63] МЕНЬШИКОВ Г.Г. *Локализуемые вычисления. Конспект лекций*. – Санкт-Петербург: СПбГУ, Факультет прикладной математики–процессов управления, 2003.
- [64] МИКЕЛАДЗЕ Ш.Е. *Численные методы математического анализа*. – Москва: ГИТТЛ, 1953.
- [65] МИЛН В.Э. *Численный анализ*. – Москва: Издательство иностранной литературы, 1951.
- [66] МИХАЙЛОВ Г.А., ВОЙТИШЕК А.В. *Численное статистическое моделирование. Методы Монте-Карло*. – Москва: Изд. центр «Академия», 2006.
- [67] МУДРОВ А.Е. *Численные методы для ПЭВМ на языках Бейсик, Фортран и Паскаль*. – Томск: МП «Раско», 1991.
- [68] МЫСОВСКИХ И.П. *Интерполяционные кубатурные формулы*. – Москва: Наука, 1981.
- [69] МЫСОВСКИХ И.П. *Квадратурные формулы наивысшей тригонометрической степени точности // Журнал вычислительной математики и математической физики*. – 1985. – Т. 25, №8. – С. 1246–1252.

- [70] НИВЕРГЕЛЬТ Ю., ФАРРАР ДЖ., РЕЙНГОЛД Э. *Машинный подход к решению математических задач.* – Москва: Мир, 1977.
- [71] НИКИФОРОВ А.Ф., СУСЛОВ С.К., УВАРОВ В.Б. *Классические ортогональные полиномы дискретной переменной.* – Москва: Наука, 1985.
- [72] ПАШКОВСКИЙ С. *Вычислительные применения многочленов и рядов Чебышёва.* – Москва: Наука, 1983.
- [73] ПОГОРЕЛОВ А.И. *Дифференциальная геометрия.* – Москва: Наука, 1974.
- [74] РЕМЕЗ Е.Я. *Основы численных методов чебышёвского приближения.* – Киев: Наукова думка, 1969.
- [75] СЕГЁ Г. *Ортогональные многочлены.* – Москва: Физматлит, 1962.
- [76] СОВОЛЕВ С.Л. *Введение в теорию кубатурных формул.* – Москва: Наука, 1974.
- [77] СТЕКЛОВ В.А. О приближённом вычислении определённых интегралов // *Известия Академии Наук.* – 1916. – Т. 10, №6. – С. 169–186.
- [78] СУЕТИН П.К. *Классические ортогональные многочлены.* – Москва: Наука, 1979.
- [79] СТЕФЕНСЕН И.Ф. *Теория интерполяции.* – Москва: Объединённое научно-техническое издательство НКТП СССР, 1935.
- [80] ХАНСЕН Э., УОЛСТЕР ДЖ.У. *Глобальная оптимизация с помощью методов интервального анализа.* – Москва-Ижевск: Издательство «РХД», 2012.
- [81] ХАУСХОЛДЕР А.С. *Основы численного анализа.* – Москва: Издательство иностранной литературы, 1956.
- [82] ХЕММИНГ Р.В. *Численные методы.* – Москва: Наука, 1972.
- [83] ЯГЛОМ И.М. *Комплексные числа и их применение в геометрии.* – Москва: Физматлит, 1963.
- [84] ABERTH O. *Precise numerical methods using C++.* – San Diego: Academic Press, 1998.
- [85] BOREL É. *Leçons sur les fonctions de variables réelles et les développements en séries de polynomes.* – Paris: Gauthier-Villars, 1905.
- [86] CHRISTOFFEL E.B. Über die Gaußsche Quadratur und eine Verallgemeinerung derselben // *Journal für die reine and angewandte Mathematik.* – 1858. – Issue 55. – S. 61–82.
- [87] HALL A. On an experimental determination of π // *Messenger of Mathematics.* – 1873. – Vol. 2. – P. 113–114.
- [88] HOLLADAY J.C. Smoothest curve approximation // *Mathematical Tables and Other Aids to Computation.* – 1957. – Vol. 11, No. 60. – P. 233–243.
- [89] LOBACHEVSKY N. Probabilité des résultats moyens tirés d'observations répétées // *Journal für die reine und angewandte Mathematik.* – 1842. – Bd. 24. – S. 164–170.
- [90] MOORE R.E., KEARFOTT R.B., CLOUD M. *Introduction to interval analysis.* – Philadelphia: SIAM, 2009.
- [91] POLYA G. Über Konvergenz von Quadraturverfahren // *Mathematische Zeitschrift.* – 1933. – Bd. 37. – S. 264–286.

- [92] 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.
- [93] RUNGE C. Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten // *Zeitschrift für Mathematik und Physik*. – 1901. – Bd. 46. – S. 224–243.
- [94] 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.
- [95] 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 направление¹ 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 [47]. 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.

¹В англоязычной учебной и научной литературе для него часто используют термин «матричные вычисления», который 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 формулирует 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.

Постановку последней задачи мы будем обсуждать ниже в §3.2д. Вторая задача из этого списка — линейная задача наименьших квадратов — является одним из вариантов дискретной задачи о наилучшем среднеквадратичном приближении. Она возникает обычно в связи с решением переопределённых систем линейных алгебраических уравнений (СЛАУ), которые, к примеру, получаются при обработке экспериментальных данных.

Помимо перечисленных задач к сфере вычислительной линейной алгебры относится также решение разнообразных линейных матричных уравнений, т. е. уравнений, в которых неизвестными являются матрицы (см., к примеру, [81]). 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, которые, в свою очередь, могут быть вычислены по явной формуле;
- для вычисления собственных значений матрицы A нужно выписать её характеристическое (вековое) уравнение $\det(A - \lambda I) = 0$ и найти его корни λ ;

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, и потому из-за взрывного роста факториала² решение СЛАУ по правилу Крамера при $n \approx 20$ –30 делается невозможным даже на самых современных ЭВМ.

Производительность современных ЭВМ принято выражать в так называемых *флопах* (сокращение от английской фразы floating point operation), и 1 флоп — это одна усреднённая арифметическая операция в арифметике с плавающей точкой в секунду (см. §1.3). Для наиболее мощных на сегодняшний день ЭВМ скорость работы измеряется так называемым петафлопами, 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 эксафлоп = 10^{18} операций с плавающей точкой в секунду. Решение на такой вычислительной машине системы линейных алгебраических уравнений размера 30×30 по правилу Крамера, с раскрытием определителей по явной комбинаторной формуле, потребует времени

$$30 \text{ компонент решения} \cdot \frac{30 \cdot 30! \text{ операций}}{10^{18} \text{ флоп} \cdot 3600 \frac{\text{сек}}{\text{час}} \cdot 24 \frac{\text{час}}{\text{сутки}} \cdot 365 \frac{\text{сутки}}{\text{год}}},$$

т. е. примерно $7.57 \cdot 10^9$ лет. Для сравнения, возраст Земли в настоящее время оценивается в $4.5 \cdot 10^9$ лет.

Обращаясь к задаче вычисления собственных значений матрицы, напомним известную из алгебры теорему Абеля-Руффини³: для общих алгебраических уравнений степени выше четвёртой не существует конечной формулы, выражающей решения уравнения через коэффициенты с помощью арифметических операций и взятия корней произвольной степени. К этому добавляются трудности в раскрытии определителя, который входит в характеристическое уравнение матрицы. Таким образом, для матриц размера 5×5 и более мы по необходимо-

²Напомним 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.

³Иногда её называют просто «теоремой Абеля» (см., к примеру, [64]).

сти 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

3.2a 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

$$a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \quad \text{— это вектор-столбец,} \quad (3.1)$$

а

$$a = (a_1, a_2, \dots, a_n) \quad \text{— это вектор-строка.}$$

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».

Все три перечисленных выше смысла 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} , мы 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 можно складывать, вычитать, умножать на скаляр. Множество векторов с определёнными на нём операциями, которые подчиняются некоторым правилам (называемым также *аксиомами*; их полный список можно увидеть, например, в [6, 7, 24, 29, 43, 54, 68]), обычно рассматривают как специальную алгебраическую структуру — линейное векторное пространство.

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 $a = \alpha b$ для некоторого скаляра α . Иногда различают *сонаправленные* 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 с линейными векторными пространствами, 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.

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 $\text{lin} \{v_1, v_2, \dots, v_r\}$, так что

$$\text{lin} \{v_1, v_2, \dots, v_r\} := \left\{ \sum_{i=1}^r \alpha_i v_i \mid \alpha_i \in \mathbb{R} \text{ или } \mathbb{C} \right\}.$$

Говорят, что линейное пространство L есть *прямая сумма* своих подпространств L_1, L_2, \dots, L_m , и обозначают

$$L = L_1 \oplus L_2 \oplus \dots \oplus L_m,$$

если любой вектор $x \in L$ единственным образом представляется в виде суммы $x = x_1 + x_2 + \dots + x_m$, 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 .

На линейных пространствах \mathbb{R}^n и \mathbb{C}^n можно задать *скалярные произведения векторов*, которые мы будем обозначать угловыми скобками $\langle \cdot, \cdot \rangle$. Напомним, что в вещественном случае это положительно определённая симметричная и билинейная форма, а в комплексном — положительно определённая эрмитова форма. Обычно они задаются в следующем стандартном виде

$$\langle a, b \rangle = \sum_{i=1}^n a_i b_i = a^\top b = b^\top a \quad \text{для } a, b \in \mathbb{R}^n \quad (3.2)$$

или

$$\langle a, b \rangle = \sum_{i=1}^n a_i \bar{b}_i = b^* a = a^\top \bar{b} \quad \text{для } a, b \in \mathbb{C}^n, \quad (3.3)$$

где через \bar{b}_i обозначено комплексно-сопряжённое к b_i число. Наличие скалярного произведения позволяет говорить о величине угла между векторами и ввести очень важное понятие ортогональности векторов.

По определению векторы a и b называются *ортогональными*, если $\langle a, b \rangle = 0$. Для обозначения ортогональности мы будем также писать $a \perp b$. 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.

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{C}^n являются прямыми суммами любых своих линейных подпространств и их ортогональных дополнений (см. подробности, к примеру, в [7, 29]).

Для любого вектора a и подпространства L в \mathbb{R}^n или \mathbb{C}^n всегда существует единственное разложение $a = l + h$, $l \in L$, $h \in L^\perp$. 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 $\text{pr}_L a$.

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

$$A := \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}, \quad (3.4)$$

называя a_{ij} *элементами* матрицы $A = (a_{ij})$. Двойной индекс означает номер строки и номер столбца, в которых располагается рассматриваемый элемент. При этом мы можем отождествлять n -векторы с матрицами размера $n \times 1$ (вектор-столбцы) либо $1 \times n$ (вектор-строки).

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 -ым 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».

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. Во-вторых, с помощью матриц даётся удобное представление для линейных отображений конечномерных линейных векторных пространств. Кроме того, матрицы чрезвычайно полезны для компактной записи множества коэффициентов систем линейных алгебраических уравнений.

Подматрицей матрицы A называют матрицу, образованную элементами, находящимися на пересечении фиксированного множества строк и столбцов A с сохранением их исходного порядка. Определитель квадратной подматрицы порядка k матрицы A носит название *минора k -го порядка* матрицы A . Соответственно, *ведущей подматрицей* некоторой матрицы называется квадратная матрица, составленная из строк и столбцов с первыми номерами. *Ведущий минор* матрицы — это определитель ведущей подматрицы.

Транспонированной к $m \times n$ -матрице $A = (a_{ij})$ называется $n \times m$ -матрица A^T , в которой ij -ым элементом является a_{ji} . Иными словами, если A задаётся в виде (3.4), то

$$A^T := \begin{pmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & a_{22} & \dots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m} & a_{2m} & \dots & a_{nm} \end{pmatrix}.$$

Числовые матрицы можно складывать, вычитать и умножать друг на друга. Напомним, что сумма (разность) двух матриц одинакового размера есть матрица того же размера, образованная поэлементными суммами (разностями) операндов. 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.

Очевидное, но важное свойство, связывающее стандартное скалярное произведение с умножением на матрицу:

$$\langle Ax, y \rangle = \langle x, A^T y \rangle \quad \text{в вещественном случае,}$$

$$\langle Ax, y \rangle = \langle x, A^* y \rangle \quad \text{в комплексном случае.}$$

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.

Квадратная диагональная матрица I вида

$$\begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix},$$

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.⁴

⁴Буква I — от слова «identity», т. е. «тождественность».

Матрицы можно рассматривать как объекты, составленные из своих вектор-строк или же 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 столбцов. В курсах линейной алгебры показывается, что строчный и столбцовый ранги матрицы совпадают друг с другом и равны максимальному размеру ненулевого минора этой матрицы. Как следствие, мы 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*.

Квадратная матрица, все строки которой (или столбцы) линейно независимы, называется *неособенной* (регулярной, невырожденной). Её ранг равен, таким образом, её порядку. В противном случае квадратная матрица называется *особенной* (вырожденной).

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 обстоятельством стоит отметить, что неособенные матрицы часто называют *обратимыми*.

Квадратные матрицы A и B одинакового порядка называются *подобными*, если существует такая невырожденная матрица S того же порядка, что

$$B = S^{-1}AS.$$

Подобные матрицы получаются при задании одного и того же линейного преобразования матрицей в разных координатных системах.

В случае, когда нулевые и ненулевые элементы в матрице A структурированы определённым образом, по отношению к A будут употребляться дополнительные определяющие термины. Например,

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ & a_{22} & \dots & a_{2n} \\ & & \ddots & \vdots \\ 0 & & & a_{nn} \end{pmatrix} \quad \text{и} \quad \begin{pmatrix} a_{11} & & & 0 \\ a_{21} & a_{22} & & \\ \vdots & \vdots & \ddots & \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$$

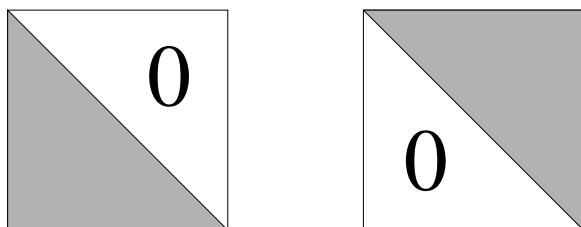


Рис. 3.1. Наглядные образы нижней треугольной и верхней треугольной матриц.

— это *верхняя треугольная* и *нижняя треугольная* матрицы соответственно (см. Рис. 3.1). Равнозначные термины — *правая треугольная* и *левая треугольная* матрицы. Выбор того или иного варианта названия обычно диктуется контекстом или сложившейся традицией.

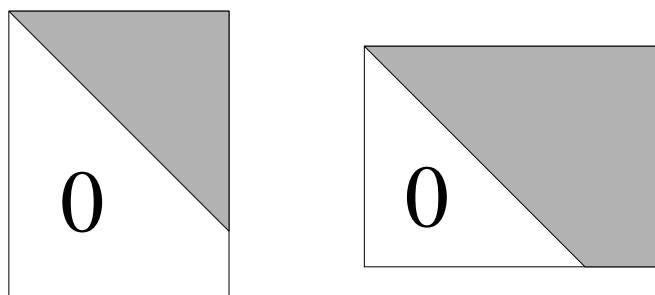


Рис. 3.2. Наглядные образы верхних (правых) трапецевидных матриц.

Обобщением понятия треугольных матриц на произвольный прямоугольный (неквадратный) случай являются *трапецевидные матрицы*. Именно, прямоугольная матрица с нулями выше (ниже) диагонали называется нижней (верхней) трапецевидной матрицей. Можно называть их также правой и левой трапецевидными матрицами.

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, строчные или столбцовые размеры которых вдоль одной строки или одного столбца одинаковы. Подматрицы A_{ij} называются тогда *блоками* рассматриваемой матрицы. Блочные матрицы вида

$$\begin{pmatrix} A_{11} & & & 0 \\ & A_{22} & & \\ 0 & & \ddots & \\ & & & A_{nn} \end{pmatrix} \quad \text{и} \quad \begin{pmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ & A_{22} & \dots & A_{2n} \\ 0 & & \ddots & \vdots \\ & & & A_{nn} \end{pmatrix},$$

где внедиагональные блоки или же блоки ниже главной диагонали являются нулевыми, назовём соответственно *блочно-диагональными* или *верхними блочно треугольными* (правыми блочно треугольными), см. Рис. 3.3. Аналогичным образом определяются нижние блочно треугольные (левые блочно треугольные) матрицы.

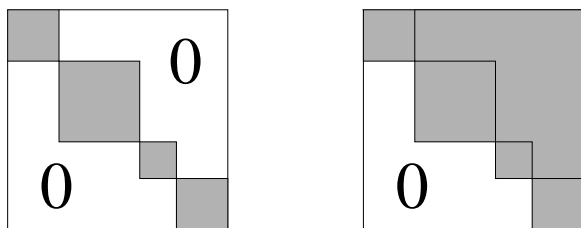


Рис. 3.3. Наглядные образы блочно-диагональной и верхней блочно-треугольной матриц.

Введение структурированных матриц и отдельное их изучение мотивируется тем, что многие операции с такими матрицами можно выполнить более специальным образом и существенно проще, чем в самом общем случае. В частности, для блочных матриц одинаковой структуры сложение и умножение выполняются «по блокам», т. е. совершенно аналогично операциям над обычными матрицами, но определённым

«поблочным» образом, когда блоки выступают как отдельные самостоятельные элементы.

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, так как в силу известной из математического анализа формулы Эйлера гармонические колебания с угловой частотой ω 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 -ым 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 элементов.

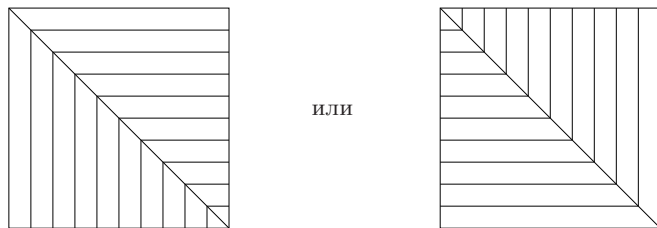


Рис. 3.4. Наглядные образы симметричной матрицы.

В линейной алгебре и её приложениях широко используются специальные типы матриц — эрмитовы, симметричные, косоэрмитовы, косо-симметричные, унитарные, ортогональные и т. п. Напомним, что *симметричными матрицами*⁵ называют матрицы, совпадающие со свои-

⁵Используют также термин *симметрическая матрица*.

ми транспонированными, т. е. удовлетворяющие $A^T = A$. Эрмитовыми матрицами называются такие комплексные матрицы A , что $A^* = A$. Матрица Q называется унитарной, если $Q^*Q = I$. Матрица Q называется ортогональной, если $Q^T Q = I$.

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* называют матрицы, которые не являются разреженными. Иными словами, в плотно заполненных матрицах большинство элементов не равны нулю.

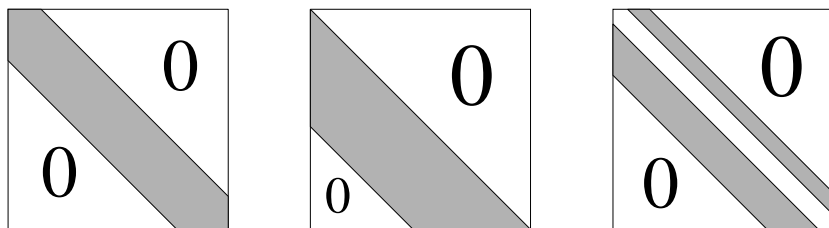


Рис. 3.5. Наглядные образы некоторых ленточных матриц.

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 $i - j > q$. В этом случае величина $p + q + 1$ называется *шириной ленты*. Простейшими и важными 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 встретятся нам в §3.8.

3.2в Собственные числа и собственные векторы матрицы

Как должно быть известно читателю, в теории и приложениях матриц огромную роль играют их *собственные значения* и *собственные векторы*. Если обозначить посредством λ собственное значение квадратной $n \times n$ -матрицы A , а x , $x \neq 0$, — её собственный вектор, то они удовлетворяют матричному уравнению

$$Ax = \lambda x. \quad (3.5)$$

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 λ , т.е. как растяжение или сжатие. Иными словами умножение вектора на всю матрицу в таких подпространствах эквивалентно умножению на один скаляр.

Собственные значения являются корнями так называемого *характеристического уравнения* матрицы, которое имеет вид

$$\det(A - \lambda I) = 0.$$

Для $n \times n$ -матрицы A это алгебраическое уравнение n -ой степени, в правой части которого стоит полином, называемый *характеристическим полиномом матрицы*. Очевидно, для полного и всестороннего исследования этого алгебраического полинома, а также разрешимости самого характеристического уравнения по существу требуется привлечение алгебраически полного поля комплексных чисел \mathbb{C} . Всякая $n \times n$ -матрица A зануляет свой характеристический полином, и этот результат является содержанием *теоремы Гамильтона-Кэли* (см. [9, 27, 37, 53]).

Совокупность собственных чисел матрицы называется её *спектром*, так что в общем случае спектр матрицы — подмножество комплексной плоскости. Напомним широко известный факт: собственные значения эрмитовых и симметричных матриц вещественны [9, 27, 53, 73].

Цель этого раздела — сообщить некоторые не общеизвестные свойства собственных значений и собственных векторов матриц, необходимые в дальнейшем изложении.

Предложение 3.2.1 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 нулём.

Доказательство. Пусть λ — какое-нибудь ненулевое собственное значение матрицы AB , так что

$$ABu = \lambda u \quad (3.6)$$

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 (3.6) необходимо должно быть $\lambda = 0$. Сказанное означает, что вектор Bu является собственным вектором матрицы BA , отвечающим такому же собственному значению λ .

И наоборот, если ненулевое μ есть собственное значение для BA , то, домножая слева равенство

$$BAv = \mu v$$

на матрицу A , получим

$$ABAv = AB(Av) = \mu(Av),$$

причём $Av \neq 0$. Как следствие, Av есть собственный вектор матрицы AB , отвечающий собственному значению μ . Иными словами, ненулевые собственные числа матриц AB и BA находятся во взаимнооднозначном 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.

Собственные векторы x , являющиеся решениями уравнения (3.5), называют также *правыми собственными векторами*, поскольку они умножаются на матрицу справа. Но нередко возникает необходимость рассмотрения *левых собственных векторов*, обладающих свойством,

аналогичным (3.5), 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. Ясно, что 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 3.2.2 *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,

$$\begin{aligned} \det(A - \lambda I) &= \det(A - \lambda I)^\top = \det(A^\top - \lambda I) \\ &= \overline{\det(A^\top - \lambda I)} = \overline{\det(A^* - \bar{\lambda} I)}. \end{aligned}$$

Отсюда мы можем заключить, что комплексное число z является корнем характеристического уравнения $\det(A - \lambda I) = 0$ для матрицы A тогда и только тогда, когда ему сопряжённое \bar{z} является корнем уравнения $\det(A^* - \lambda I) = 0$, характеристического для матрицы A^* . Это доказывает первое утверждение.

Пусть x и y — собственные векторы матриц A и A^* соответственно, а λ 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

$$\lambda \langle x, y \rangle - \bar{\varkappa} \langle x, y \rangle = 0,$$

то есть

$$(\lambda - \bar{\varkappa}) \langle x, y \rangle = 0.$$

Если x и y являются собственными векторами матриц A и A^* , отвечающим собственным значениям λ 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.

Обращаясь к определению правых и левых собственных векторов матрицы, можем утверждать, что если λ — правое собственное значение матрицы A , а μ — 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 3.2.3 Если λ — собственное число квадратной неособенной матрицы, то λ^{-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 $C^{-1}v = \lambda^{-1}v$. ■

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 α , на первой наддиагонали все единицы, а остальные элементы — нули, называют, как известно, *жордановой клеткой*, отвечающей значению α . Ясно, что α является собственным значением такой матрицы.

В линейной алгебре 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, 27, 41, 43, 53]). Иными словами для любой квадратной матрицы A существует такая неособенная матрица S , что

$$S^{-1}AS = J,$$

где

$$J = \left(\begin{array}{ccc|ccc|ccc} \lambda_1 & 1 & & & & & & & \\ & \lambda_1 & \ddots & & & & & & \\ & & \ddots & 1 & & & & & \\ & & & \lambda_1 & & & & & \\ \hline & & & & \lambda_2 & 1 & & & \\ & & & & & \ddots & \ddots & & \\ & & & & & & \lambda_2 & & \\ \hline & & & & & & & & \\ & & & & & & & & \\ & & & & & & & & \end{array} \right), \quad (3.7)$$

а $\lambda_1, \lambda_2, \dots$ — собственные значения матрицы A .

Соответственно, представление произвольной матрицы A в виде

$$A = SJS^{-1},$$

где J — матрица в жордановой нормальной форме, называют *жордановым разложением*.

Неприятной особенностью жордановой канонической формы и жорданова разложения является то, что они не зависят непрерывно от элементов матрицы, несмотря на то, что сами собственные значения матрицы непрерывно зависят от её элементов (теорема Островского, см. §3.16б). Размеры жордановых клеток-блоков и их расположение вдоль диагонали могут скачкообразно меняться при изменении элементов матрицы. Это делает жорданову форму малоприменимой 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. §3.16б), которые определяются как матрицы, подобные диагональным. Можно показать, что таких матриц — «большинство», т. е. типичная матрица имеет простую структуру. Жорданово разложение таких матриц превращается в более простое представление

$$A = SDS^{-1},$$

в котором D — диагональная матрица, у которой по диагонали стоят собственные значения A с учётом их кратности. Часто это представ-

ление 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 мы рассматриваем в §§3.17г–3.17д.

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 преобразования подобия любая 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 раздела 3.2в следует, что для определения собственных значений квадратной матрицы A и её левых и правых собственных

векторов необходимо решить относительно λ , x и y систему уравнений

$$\begin{cases} Ax = \lambda x, \\ y^* A = \lambda y^*. \end{cases} \quad (3.8)$$

Система уравнений (3.8) 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 (3.8) можно также по частям, отдельно для x и отдельно для y , что обычно и делают на практике. Если λ вещественно, т. е. $\lambda = \bar{\lambda}$, то системе (3.8), применив операцию эрмитова сопряжения ко второй части, можно придать следующий элегантный матричный вид

$$\begin{pmatrix} A & 0 \\ 0 & A^* \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix}. \quad (3.9)$$

Рассмотрим теперь аналогичную систему уравнений, порождаемую заданной матрицей A , которая получается изменением соотношений в (3.8) так, чтобы они «завязались» друг на друга:

$$\begin{cases} Ax = \sigma y, \\ y^* A = \sigma x^*, \end{cases} \quad (3.10)$$

— мы просто поменяли в правых частях векторы x и y местами друг с другом. Фигурально можно сказать, что в новой системе уравнений (3.10) векторы x и y становятся «право-левыми» и «лево-правыми собственными векторами» матрицы A . Как мы увидим вскоре, аналоги собственных чисел матрицы, которые мы переобозначили через σ , также получают новое содержание. Решения системы (3.8) 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 (3.10). Они тоже дают важную информацию о матрице, хотя и другого сорта, нежели (3.8).

Система уравнений (3.10) — это система алгебраических уравнений относительно σ , x , y , и потому естественно ожидать, что её решениями в самом общем случае, когда A — комплексная матрица, будут тоже комплексные числа σ и комплексные векторы x , y . Но оказывается, что σ всегда может быть взято вещественным.

В самом деле, если σ , удовлетворяющие системе (3.10), вещественны, то $\bar{\sigma} = \sigma$, и, беря эрмитово сопряжение второго матричного урав-

нения из (3.10), можем переписать всю эту систему в следующем равносильном виде:

$$\begin{cases} Ax = \sigma y, \\ A^*y = \sigma x. \end{cases} \quad (3.11)$$

Матричная форма этой системы —

$$\begin{pmatrix} 0 & A^* \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \sigma \begin{pmatrix} x \\ y \end{pmatrix}, \quad (3.12)$$

— находится в красивой двойственности с системой (3.9). Если A — вещественная матрица, то векторы x и y также могут быть взяты вещественными, а система уравнений (3.12) для определения сингулярных чисел и векторов принимает ещё более простой вид:

$$\begin{pmatrix} 0 & A^\top \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \sigma \begin{pmatrix} x \\ y \end{pmatrix}. \quad (3.13)$$

Сразу же можно заметить, что 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. Как следствие, матричные уравнения (3.12) и (3.13), которые определяют их собственные значения и собственные $(m+n)$ -векторы, имеют решения σ , x , y , причём σ — вещественные числа.

Определение 3.2.2 *Неотрицательные вещественные скаляры σ , которые являются решениями системы матричных уравнений (3.11), называются сингулярными числами матрицы A . Удовлетворяющие системе (3.11) векторы x называются правыми сингулярными векторами матрицы A , а векторы y — левыми сингулярными векторами матрицы A .*

Отметим, что и система (3.11), и данное выше определение имеют смысл уже для произвольных прямоугольных матриц, а не только для квадратных, как было в случае собственных значений и собственных векторов. Для $m \times n$ -матрицы A правые сингулярные векторы имеют размерность n , а левые — размерность m . Из уравнений (3.11)–(3.13)

видно также, что, в отличие от собственных значений, сингулярные числа характеризуют совместно как саму матрицу, так и её эрмитово-сопряжённую (транспонированную в вещественном случае).

Наша ближайшая цель — завершить исследование корректности Определения 3.2.2. Мы явно укажем решения σ , x , y для систем уравнений (3.11), (3.12) и (3.13) и наличие среди них вещественных неотрицательных σ .

Предложение 3.2.4 *Сингулярные числа матрицы A суть неотрицательные квадратные корни из собственных чисел матрицы A^*A или матрицы AA^* .*

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 §2.10e).

Формулировка Предложения 3.2.4 может потребовать пояснений, так как в случае 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. Соответственно, количество собственных чисел у них будет различным.

Но известно, что ранг произведения матриц не превосходит наименьшего из рангов перемножаемых матриц (см. [9, 24, 53]). 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 -ое — 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 -ое равны нулю. Таким образом, для $m \times n$ -матрицы A содержательный смысл имеет рассмотрение лишь $\min\{m, n\}$ штук собственных чисел матриц A^*A и AA^* , что устраняет отмеченную выше кажущуюся неоднозначность.

Другой 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 3.2.1 о совпадении ненулевых точек спектра произведений двух матриц, взятых в различном порядке. Впрочем, для частного случая матриц A^*A и AA^* этот факт будет обоснован в следующем ниже доказательстве.

Доказательство. Умножая обе части второго уравнения из (3.11) на

σ , получим $A^*(\sigma y) = \sigma^2 x$. Затем подставим сюда значение σy из первого уравнения (3.11):

$$A^*Ax = \sigma^2 x.$$

С другой стороны, умножая на σ обе части первого уравнения (3.11), получим $A(\sigma x) = \sigma^2 y$. Подстановка в это равенство значения σx из второго уравнения (3.10) даёт

$$AA^*y = \sigma^2 y.$$

Иными словами, числа σ^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 σ . Очевидно, это достаточно сделать лишь для одной из выписанных матриц, так как для другой рассуждения совершенно аналогичны.

Пусть λ — 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 осталось продемонстрировать, что арифметические квадратные корни из собственных значений матриц A^*A и AA^* вместе с их собственными векторами удовлетворяют системе уравнений (3.11)–(3.12).

Пусть u — собственный вектор матрицы A^*A , отвечающий собственному числу λ , так что $A^*Au = \lambda u$, причём $\lambda \geq 0$ в силу ранее доказанного. Обозначим $y := Au$ и $x := \sqrt{\lambda} u$. Тогда $\lambda u = \sqrt{\lambda} x$, и потому

$$Ax = A(\sqrt{\lambda} u) = \sqrt{\lambda} Au = \sqrt{\lambda} y,$$

$$A^*y = A^*Au = \lambda u = \sqrt{\lambda} x,$$

так что система (3.10)–(3.12) удовлетворяется при $\sigma = \sqrt{\lambda}$ с выбранными векторами x и y .

Аналогично, если v — собственный вектор матрицы AA^* , отвечающий её собственному числу μ , то $AA^*v = \mu v$, причём $\mu \geq 0$. Обозначим $x := A^*v$ и $y := \sqrt{\mu} v$. Тогда $\mu v = \sqrt{\mu} y$, и потому

$$Ax = AA^*v = \mu v = \sqrt{\mu} y,$$

$$A^*y = A^*(\sqrt{\mu} v) = \sqrt{\mu} A^*v = \sqrt{\mu} x,$$

так что система (3.11)–(3.12) действительно удовлетворяется при $\sigma = \sqrt{\mu}$ с выбранными векторами x и y . ■

Следствие. 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. ■

Итак, 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 доказательства Предложения 3.2.4 следует также, что правыми сингулярными векторами матрицы A являются правые собственные векторы матрицы A^*A , а левыми сингулярными векторами матрицы A — левые собственные векторы для A^*A или, что равносильно, эрмитово сопряжённые правых собственных векторов матрицы AA^* . Отметим также, что как левые, так и правые сингулярные векторы суть ортогональные системы векторов, коль скоро они являются собственными векторами эрмитовых матриц A^*A и AA^* .

Пример 3.2.1 Пусть A — это 1×1 -матрица, т. е. просто некоторое число a , вещественное или комплексное. Ясно, что единственное собственное число такой матрицы равно самому a . Сингулярное число у A также всего одно, и оно равно $\sqrt{\bar{a}a} = |a|$.

Пусть $A = (a_1, a_2, \dots, a_n)^T$ — 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. Далее мы ещё не раз увидим, как эта связь проявляется в самых неожиданных местах (см., в частности, Предложение 3.3.6).

Пример 3.2.2 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 3.2.3 Для 2×2 -матрицы

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \quad (3.14)$$

нетрудно выписать характеристическое уравнение

$$\det \begin{pmatrix} 1 - \lambda & 2 \\ 3 & 4 - \lambda \end{pmatrix} = \lambda^2 - 5\lambda - 2 = 0,$$

и найти его корни $\frac{1}{2}(5 \pm \sqrt{33})$ — собственные значения матрицы, приближённо равные -0.372 и 5.372 . Для определения сингулярных чисел образуем

$$A^\top A = \begin{pmatrix} 10 & 14 \\ 14 & 20 \end{pmatrix},$$

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

$$\begin{pmatrix} 1 & 2 \\ -3 & 4 \end{pmatrix}, \quad (3.15)$$

которая отличается от матрицы (3.14) 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 максимальные сингулярные числа рассмотренных матриц превосходят наибольшие из модулей их собственных чисел. Мы увидим ниже (см. §3.3ж), что это не случайно, и наибольшее сингулярное число всегда не меньше, чем максимум модулей собственных чисел матрицы.

Рассмотрим вопрос о том, как связаны сингулярные числа для взаимно обратных матриц.

Предложение 3.2.5 *Если σ — сингулярное число неособенной квадратной матрицы, то σ^{-1} — это сингулярное число обратной к ней матрицы.*

Доказательство. Вспомним, что собственные числа взаимно обратных матриц обратны друг другу (Предложение 3.2.3). 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. ■

3.2е 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 3.2.1 (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

$$A = U\Sigma V^* \quad (3.16)$$

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 (3.16) называется *сингулярным разложением матрицы* A . Если A — вещественная матрица, то U и V также являются вещественными ортогональными матрицами, и сингулярное разложение принимает вид

$$A = U \Sigma V^T.$$

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, 53, 73]. Рассмотрим подробно общий комплексный случай.

Как известно, любую эрмитову матрицу можно унитарными преобразованиями подобия привести к диагональному виду, так что $S = T^*DT$, где T — унитарная, а D — диагональная. 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 по диагонали в D стоят сингулярные числа матрицы A .

Исследуем произведение A^*A :

$$\begin{aligned} A^*A &= ((QT^*)DT)^*((QT^*)DT) \\ &= T^*D^*(QT^*)^*(QT^*)DT \\ &= T^*D^*DT = T^*D^2T. \end{aligned}$$

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.

Для случая общих прямоугольных матриц доказательство Теоремы 3.2.1 не очень сложно и может быть найдено, 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 растяжением 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 отличающимися друг от друга размерностями. Собственные числа матрицы полезны при изучении линейного преобразования векторного пространства в пространство той же размерности, в частности, самого в себя.

Другие примеры применения сингулярных чисел и сингулярных векторов матриц рассматриваются ниже в §3.4.

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 (см. [107]). Задача нахождения сингулярных чисел и сингулярных векторов матриц, последняя из списка на стр. 261, по видимости является частным случаем третьей задачи, относящейся к нахождению собственных чисел и собственных векторов. Но вычисление сингулярных чисел и векторов матриц сделалось в настоящее время очень важным как в теории, так и в приложениях 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 за-

дач. Это естественный процесс, в котором большую роль сыграло развитие вычислительной техники и информатики. Следует быть готовым к подобным изменениям и в будущем.

3.2ж Системы линейных алгебраических уравнений

В этом разделе конспективно излагаются сведения по теории систем линейных алгебраических уравнений, необходимые для понимания материала книги.

Систему уравнений

$$Ax = b$$

можно представить в равносильном виде

$$\begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{pmatrix} x_1 + \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{pmatrix} x_2 + \cdots + \begin{pmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{pmatrix} x_n = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}.$$

Из него 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.

Ниже мы будем опираться на известные результаты о существовании решений систем линейных алгебраических уравнений

Теорема 3.2.2 (теорема Кронекера-Капелли)

Система линейных алгебраических уравнений $Ax = b$ совместна 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 .

Доказательство можно увидеть в [7, 43, 68].

Теорема 3.2.3 (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 $A^\top y = 0$.

Эта формулировка теоремы Фредгольма является конечномерным вариантом более общего утверждения о разрешимости интегральных и некоторых операторных уравнений. Для систем линейных алгебраических уравнений подробное рассмотрение этого вопроса можно найти в [3, 6, 7, 68].

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 X называется вещественнозначная функция $\| \cdot \|$, удовлетворяющая следующим свойствам (называемым аксиомами нормы):*

$$(BN1) \quad \|a\| \geq 0 \quad \text{для любого } a \in X, \text{ причём } \|a\| = 0 \Leftrightarrow a = 0$$

— неотрицательность,

$$(BN2) \quad \|\alpha a\| = |\alpha| \cdot \|a\| \quad \text{для любых } a \in X \text{ и } \alpha \in \mathbb{R} \text{ или } \mathbb{C}$$

— абсолютная однородность,

$$(BN3) \quad \|a + b\| \leq \|a\| + \|b\| \quad \text{для любых } a, b \in X$$

— «неравенство треугольника».

Само пространство 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 практичны, 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

$$\begin{aligned}\|a\|_1 &:= \sum_{i=1}^n |a_i|, \\ \|a\|_2 &:= \left(\sum_{i=1}^n |a_i|^2 \right)^{1/2}, \\ \|a\|_\infty &:= \max_{1 \leq i \leq n} |a_i|.\end{aligned}$$

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 рассмотренных норм.

Замечательность евклидовой нормы $\|\cdot\|_2$ состоит в том, что она порождается стандартным скалярным произведением $\langle \cdot, \cdot \rangle$ в \mathbb{R}^n или \mathbb{C}^n . Более точно, если скалярное произведение задаётся как (3.2) или (3.3), 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 важное *неравенство Коши-Буняковского*

$$|\langle a, b \rangle| \leq \|a\|_2 \|b\|_2 \quad (3.17)$$

(см., к примеру, [7, 9, 14, 19, 23, 24, 43]).

Нормы $\|\cdot\|_1$ и $\|\cdot\|_2$ — это частные случаи более общей конструкции *p-нормы*

$$\|a\|_p = \left(\sum_{i=1}^n |a_i|^p \right)^{1/p} \quad \text{для } p \geq 1,$$

которую называют также *гёльдеровой нормой* (по имени О.Л. Гёльдера). 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, 53]. Чебышёвская норма тоже может быть получена из конструкции p -нормы с помощью предельного перехода по $p \rightarrow \infty$, что и объясняет индекс « ∞ » 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

$$\lim_{p \rightarrow \infty} \left(\sum_{i=1}^n |a_i|^p \right)^{1/p} = \max_{1 \leq i \leq n} |a_i|.$$

В нормированном пространстве \mathcal{X} *шаром* радиуса r с центром в точке a называется множество $\{x \in \mathcal{X} \mid \|x - a\| \leq r\}$. Геометрически наглядное представление о норме даётся её единичным шаром, т. е. множеством $\{x \mid \|x\| \leq 1\}$. На Рис. 3.6 нарисованы единичные шары для рассмотренных выше норм в \mathbb{R}^2 . Из аксиом нормы вытекает, что единичный шар любой нормы — это множество в линейном векторном

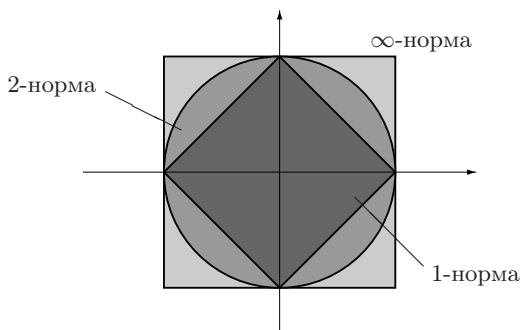


Рис. 3.6. Шары единичного радиуса в различных нормах.

пространстве, которое выпукло (следствие неравенства треугольника) и *уравновешено*, т. е. переходит в себя при умножении на любой скаляр α с $|\alpha| \leq 1$ (следствие абсолютной однородности).

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 (см. Рис. 3.7). Они являются важнейшим частным случаем многомерных интервалов (см. [98]), и в связи с этим обстоятельством взвешенная чебышёвская норма популярна в интервальном анализе.

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 §3.9б.

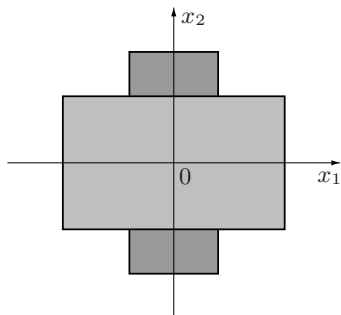


Рис. 3.7. Шары единичного радиуса во взвешенных чебышёвских нормах.

3.36 Топология на векторных пространствах

Говорят, что на множестве X задана *топологическая структура*, или просто *топология*⁶, если в X выделен класс подмножеств \mathcal{O} , содержащий вместе с каждым набором множеств их объединение и вместе с каждым конечным набором множеств — их пересечение. Множество, снабжённое топологической структурой, называется *топологическим пространством*, а множества выделенного класса \mathcal{O} — *открытыми множествами*. Подмножество топологического пространства называется *замкнутым*, если его теоретико-множественное дополнение до всего пространства открыто.

Окрестностью точки в топологическом пространстве называется всякое открытое множество, содержащее эту точку. Окрестностью подмножества топологического пространства называется всякое открытое множество, содержащее это подмножество. Задание окрестностей точек и множеств позволяет определять близость одного элемента множества к другому, предельные переходы, сходимости и т. п. понятия. 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)

⁶Топологией называется также математическая дисциплина, изучающая, главным образом, свойства объектов, инвариантные относительно непрерывных отображений (см., к примеру, [63, 70]).

или же с помощью различных норм. Преимущество этого пути состоит в том, что мы получаем в своё распоряжение количественную меру близости рассматриваемых объектов. При этом открытыми множествами считаются такие множества, каждая точка которых принадлежит множеству вместе с некоторым шаром с центром в этой точке.

Как известно, на нормированном пространстве X с нормой $\|\cdot\|$ расстояние (метрика) между элементами a и b может быть естественно задано как

$$\text{dist}(a, b) = \|a - b\|, \quad (3.18)$$

т.е. как «величина различия» элементов a и b . Непосредственной проверкой легко убедиться, что для введённой таким образом функции $\text{dist} : X \times X \rightarrow \mathbb{R}_+$ выполняются все аксиомы расстояния (мы приводили их ранее на стр. 56). 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 (3.18) линейное векторное пространство с нормой превращается в метрическое пространство. В целом, задание нормы на некотором линейном векторном пространстве X 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.

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.*

Помимо сходимости последовательностей и их пределов часто необходимо рассматривать сходимость непрерывно изменяющихся переменных величин. Для метрических пространств, как показывается в общей топологии, эти два понятия равносильны друг другу [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.

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, 55]):

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 $a \in X$*

$$C_1\|a\|' \leq \|a\|'' \leq C_2\|a\|'. \quad (3.19)$$

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 (3.19) 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 ∞ -норма эквивалентны друг другу.

Доказательство. Справедливость правого из первых неравенств следует из неравенства Коши-Буняковского (3.17), 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

$$\|a\|_2^2 = |a_1|^2 + |a_2|^2 + \dots + |a_n|^2,$$

$$\begin{aligned} \|a\|_1^2 &= |a_1|^2 + |a_2|^2 + \dots + |a_n|^2 \\ &\quad + 2|a_1 a_2| + 2|a_1 a_3| + \dots + 2|a_{n-1} a_n|, \end{aligned}$$

и все слагаемые $2|a_1 a_2|, 2|a_1 a_3|, \dots, 2|a_{n-1} a_n|$ 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:

$$\begin{aligned} \|a\|_2 &= \sqrt{|a_1|^2 + |a_2|^2 + \dots + |a_n|^2} \\ &\geq \sqrt{\max_i |a_i|^2} = \max_i |a_i| = \|a\|_\infty, \end{aligned}$$

$$\begin{aligned} \|a\|_2 &= \sqrt{|a_1|^2 + |a_2|^2 + \dots + |a_n|^2} \\ &\leq \sqrt{n \max_i |a_i|^2} = \sqrt{n} \max_i |a_i| = \sqrt{n} \|a\|_\infty, \end{aligned}$$

$$\begin{aligned} \|a\|_\infty &= \max_i |a_i| \\ &\leq |a_1| + |a_2| + \dots + |a_n| = \|a\|_1, \end{aligned}$$

$$\begin{aligned} \|a\|_1 &= |a_1| + |a_2| + \dots + |a_n| \\ &\leq n \max_i |a_i| \leq n \|a\|_\infty. \end{aligned}$$

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* (см., к примеру, [21, 43, 53]). Но содержание Предложения 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 останковки итераций и т. п.).

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 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 3.3.3 *Говорят, что в линейном векторном пространстве 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 $a_i^{(k)} \rightarrow a_i^*$ в \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 $\{e_i\}_{i=1}^n$. Тогда, разлагая $a^{(k)}$ и

a^* в этом базисе, получаем

$$\begin{aligned} \|a^{(k)} - a^*\| &= \left\| \sum_{i=1}^n a_i^{(k)} e_i - \sum_{i=1}^n a_i^* e_i \right\| = \left\| \sum_{i=1}^n (a_i^{(k)} - a_i^*) e_i \right\| \\ &\leq \sum_{i=1}^n \|(a_i^{(k)} - a_i^*) e_i\| = \sum_{i=1}^n |a_i^{(k)} - a_i^*| \|e_i\|. \end{aligned}$$

Как следствие, 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 норму

$$\|a\|_{\infty}^{\{e_i\}} := \max_{1 \leq i \leq n} |a_i|.$$

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 расчленивать задачу на отдельные компоненты, что иногда также упрощает рассуждения. В конце концов, в большинстве практических задач линейной алгебры векторы и матрицы — это структурированные 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 α .

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 -ых компонент векторов Ab и Ab^* :

$$\begin{aligned} |(Ab)_i - (Ab^*)_i| &= |(A(b - b^*))_i| = \left| \sum_{j=1}^n a_{ij}(b_j - b_j^*) \right| \\ &\leq \sqrt{\sum_{j=1}^n a_{ij}^2} \sqrt{\sum_{j=1}^n (b_j - b_j^*)^2} \end{aligned}$$

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 .

3.3в Матричные нормы

Помимо векторов основным объектом вычислительной линейной алгебре являются также матрицы. По этой причине нам будут нужны матричные нормы — для того, чтобы оценивать «величину» той или иной матрицы, и ещё для того, чтобы ввести расстояние между матрицами как

$$\text{dist}(A, B) := \|A - B\|, \quad (3.20)$$

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):

(МН1) $\|A\| \geq 0$ для любой матрицы A , причём $\|A\| = 0 \Leftrightarrow A = 0$
 — неотрицательность,

(МН2) $\|\alpha A\| = |\alpha| \cdot \|A\|$ для любых матриц A и $\alpha \in \mathbb{R}$ или $\alpha \in \mathbb{C}$
 — абсолютная однородность,

(МН3) $\|A + B\| \leq \|A\| + \|B\|$ для любых матриц A, B
 — «неравенство треугольника».

Но условия (МН1)–(МН3) 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», обозначающее множества с двумя взаимосогласованными бинарными операциями — сложением и умножением (см. [24, 43]). Связь нормы матриц с операцией их умножения отражает четвёртая аксиома матричной нормы:

(МН4) $\|AB\| \leq \|A\| \cdot \|B\|$ для любых матриц A, B
 — «субмультипликативность».⁷

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 МН4, понимаемая 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 .

⁷Приставка «суб-» означает «меньше», «ниже» и т. п. В этом смысле неравенства треугольника ВН3 и МН3 можно называть «субаддитивностью» норм.

Определение 3.3.4 Векторная норма $\|\cdot\|$ и матричная норма $\|\cdot\|'$ называются согласованными, если

$$\|Ax\| \leq \|A\|' \cdot \|x\| \quad (3.21)$$

для любой матрицы A и всех векторов x .

Рассмотрим примеры конкретных матричных норм.

Пример 3.3.1 Фробениусова норма матрицы $A = (a_{ij})$ определяется как

$$\|A\|_F = \left(\sum_{i,j} |a_{ij}|^2 \right)^{1/2}.$$

Ясно, что она удовлетворяет первым трём аксиомам матричной нормы просто потому, что задаётся совершенно аналогично евклидовой векторной норме $\|\cdot\|_2$. Для обоснования субмультипликативности рассмотрим

$$\|AB\|_F^2 = \sum_{i,j} \left| \sum_k a_{ik} b_{kj} \right|^2.$$

В силу неравенства Коши-Буняковского (3.17)

$$\left| \sum_k a_{ik} b_{kj} \right|^2 \leq \left(\sum_k a_{ik}^2 \right) \left(\sum_l b_{lj}^2 \right),$$

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. ■

Пример 3.3.2 Матричная норма

$$\|A\|_{\max} = n \max_{i,j} |a_{ij}|,$$

определённая 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 трех 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:

$$\begin{aligned} \|AB\|_{\max} &= n \max_{i,j} \left| \sum_{k=1}^n a_{ik} b_{kj} \right| \leq n \max_{i,j} \left(\sum_{k=1}^n |a_{ik}| |b_{kj}| \right) \\ &\leq n \left(\sum_{k=1}^n \max_{i,k} |a_{ik}| \max_{k,j} |b_{kj}| \right) \\ &\leq n^2 \max_{i,j} |a_{ij}| \max_{i,j} |b_{ij}| = \|A\|_{\max} \|B\|_{\max}. \end{aligned}$$

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. ■

В связи с последним примером следует отметить, что аксиома субмультипликативности 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 согласованными. Иными словами, справедливо

Предложение 3.3.4 *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 субмультипликативность матричной нормы, имеем

$$\begin{aligned} \|Av\| &= \|(Av, Av, \dots, Av)\|' = \|A \cdot (v, v, \dots, v)\|' \\ &\leq \|A\|' \cdot \|(v, v, \dots, v)\|' = \|A\|' \cdot \|v\|, \end{aligned}$$

так что требуемое согласование действительно будет достигнуто. \blacksquare

3.3г Подчинённые матричные нормы

В предшествующем пункте 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 (3.21), будут более точными и позволят получить более тонкие оценки результата. Например, конкретная оценка нормы погрешности может оказать 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 (3.21) вытекает неравенство для согласованной нормы матрицы $\|A\|$:

$$\|A\| \geq \|Ax\|/\|x\|, \quad (3.22)$$

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 (3.22) должно быть справедливым для любого ненулевого вектора x .

Предложение 3.3.5 *Для любой фиксированной векторной нормы $\|\cdot\|$ соотношением*

$$\|A\|' = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} \quad (3.23)$$

задаётся матричная норма.

Доказательство. Отметим прежде всего, что в случае конечномерных векторных пространств \mathbb{R}^n и \mathbb{C}^n вместо «sup» в выражении (3.23) 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, 53]. Непрерывная функция $\|Ay\|$ (см. §3.36) 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 (3.23) справедливость неравенства треугольника. Очевидно,

$$\|(A+B)y\| \leq \|Ay\| + \|By\|,$$

и потому

$$\begin{aligned} \max_{\|y\|=1} \|(A+B)y\| &\leq \max_{\|y\|=1} (\|Ay\| + \|By\|) \\ &\leq \max_{\|y\|=1} \|Ay\| + \max_{\|y\|=1} \|By\|, \end{aligned}$$

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 3.3.5 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 к норме $\|\cdot\|$ (или индуцированной нормой $\|\cdot\|$).

Иногда в отношении матричной нормы, задаваемой Определением 3.3.5, используют термин «операторная норма». Он мотивируется тем, что конструкция этой нормы хорошо отражает взгляд на матрицу как на оператор, задающий отображения 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 Предложения 3.3.6 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

$$\|I\| = \max_{\|y\|=1} \|Iy\| = \max_{\|y\|=1} \|y\| = 1. \quad (3.24)$$

Ответом на первые два вопроса является

Предложение 3.3.6 *Для векторной 1-нормы подчинённой матричной нормой $m \times n$ -матриц $A = (a_{ij})$ 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 (∞ -нормы) подчинённой матричной нормой $m \times n$ -матриц $A = (a_{ij})$ является

$$\|A\|_\infty = \max_{1 \leq i \leq m} \left(\sum_{j=1}^n |a_{ij}| \right)$$

— максимальная сумма модулей элементов по строкам.

Матричная норма, подчинённая евклидовой норме векторов $\|x\|_2$, есть $\|A\|_2 = \sigma_{\max}(A)$ — наибольшее сингулярное число матрицы A .

Доказательство. Для обоснования первой части предложения выпи-

шем следующую цепочку преобразований и оценок

$$\begin{aligned}
 \|Ax\|_1 &= \sum_{i=1}^m |(Ax)_i| = \sum_{i=1}^m \left| \sum_{j=1}^n a_{ij}x_j \right| \leq \sum_{i=1}^m \sum_{j=1}^n |a_{ij}x_j| \\
 &= \sum_{i=1}^m \sum_{j=1}^n |a_{ij}| |x_j| = \sum_{j=1}^n \sum_{i=1}^m |a_{ij}| |x_j| = \sum_{j=1}^n \left(|x_j| \sum_{i=1}^m |a_{ij}| \right) \\
 &\leq \left(\max_{1 \leq j \leq n} \sum_{i=1}^m |a_{ij}| \right) \cdot \sum_{j=1}^n |x_j| = \|A\|_1 \|x\|_1, \tag{3.25}
 \end{aligned}$$

из которой вытекает

$$\frac{\|Ax\|_1}{\|x\|_1} \leq \|A\|_1.$$

При этом все неравенства в цепочке (3.25) обращаются в равенства для вектора x в виде столбца единичной $n \times n$ -матрицы с тем номером j , на котором достигается $\max_j \sum_{i=1}^m |a_{ij}|$. Как следствие, на этом векторе достигается наибольшее значение отношения $\|Ax\|_1/\|x\|_1$ из определения подчинённой матричной нормы.

Аналогичным образом доказывается и вторая часть предложения, касающаяся $\|\cdot\|_\infty$.

Приступая к обоснованию последней части предложения 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, Λ — диагональная $n \times n$ -матрица. При этом у Λ на главной диагонали находятся числа σ_i^2 , $i = 1, 2, \dots, \min\{m, n\}$, которые являются квадратами сингулярных чисел σ_i матрицы A , и, возможно, ещё нули в случае $m < n$.

Далее имеем

$$\begin{aligned}
 \|A\|_2 &= \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max_{x \neq 0} \frac{\sqrt{x^* A^* A x}}{\sqrt{x^* x}} = \max_{x \neq 0} \frac{\sqrt{x^* U^* \Lambda U x}}{\sqrt{x^* U^* U x}} \\
 &= \max_{x \neq 0} \frac{\sqrt{(Ux)^* \Lambda (Ux)}}{\sqrt{(Ux)^* Ux}} = \max_{z \neq 0} \frac{\sqrt{z^* \Lambda z}}{\sqrt{z^* z}} = \max_{z \neq 0} \sqrt{\frac{\sum_i \sigma_i^2 |z_i|^2}{\sum_i |z_i|^2}} \\
 &\leq \max_{z \neq 0} \left(\sigma_{\max}(A) \sqrt{\frac{\sum_i |z_i|^2}{\sum_i |z_i|^2}} \right) = \sigma_{\max}(A),
 \end{aligned}$$

где в выкладках применена 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 Λ , 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. 3.3ж).

Отметим, что спектральная норма матриц не является абсолютной нормой (см. Пример 3.2.3), 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.

3.3д Топология на множествах матриц

Совершенно аналогично тому, как это было сделано для векторов, можно рассмотреть топологическую структуру на множестве матриц. Она может быть введена различными способами, но нам наиболее удобен *секвенциальный подход*, когда определение топологии осуществляется через сходимость последовательностей. 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 (3.19). 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 (3.19) играют огромную роль при выводе различных оценок, и их значения для важнейших норм даёт следующее

Предложение 3.3.7 *Для квадратных $n \times n$ -матриц*

$$\begin{aligned}\frac{1}{\sqrt{n}} \|A\|_2 &\leq \|A\|_1 \leq \sqrt{n} \|A\|_2, \\ \frac{1}{\sqrt{n}} \|A\|_\infty &\leq \|A\|_2 \leq \sqrt{n} \|A\|_\infty, \\ \frac{1}{n} \|A\|_1 &\leq \|A\|_\infty \leq n \|A\|_1.\end{aligned}$$

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

$$\|A\|_1 = \max_{\|y\|_1 \leq 1} \|Ay\|_1 \geq \max_{\|y\|_1 \leq 1} \|Ay\|_2. \quad (3.26)$$

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 (3.26), 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 *поэлементную сходимость* матриц, при которой одна матрица сходится к другой тогда и только тогда, когда все элементы первой матрицы сходятся к соответствующим элементам второй. Иными словами, для последовательности матриц $\{A^{(k)}\}_{k=0}^{\infty}$ положим

$$\begin{aligned} A^{(k)} = (a_{ij}^{(k)}) &\rightarrow A^* = (a_{ij}^*) \text{ поэлементно в } \mathbb{R}^{m \times n} \text{ или } \mathbb{C}^{m \times n} \\ &\Updownarrow \\ a_{ij}^{(k)} &\rightarrow a_{ij}^* \text{ в } \mathbb{R} \text{ или } \mathbb{C} \text{ для всех индексов } i, j. \end{aligned}$$

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 матриц в этой выкладке можно было бы взять, к примеру, ∞ -норму). Поэтому для любых индексов i и j верна оценка

$|a_{ij}| \leq C\|A\|$. Как следствие, сходимость последовательности матриц в любой норме влечёт поэлементную сходимость этой последовательности. Доказательство обратной импликации, т. е. того факта, что из поэлементной сходимости матриц следует сходимость по норме, совершенно аналогично первой части Предложения 3.3.3 (см. §3.36).

В целом для любой матричной нормы множество матриц с введённым на нём посредством (3.20) расстоянием является полным метрическим пространством, т. е. любая фундаментальная («сходящаяся в себе») последовательность имеет в нём предел. Это следует из предшествующего рассуждения и из факта полноты вещественной оси \mathbb{R} и комплексной плоскости \mathbb{C} .

В заключение этой темы отметим, что в вычислительной линейной алгебре нормы векторов и матриц широко используются с середины XX века. Пионерский вклад в развитие соответствующей математической техники внесли работа Дж. фон Неймана и Г. Голдстейна [117] и монография В. Н. Фаддеевой [94], которая предшествовала капитальной книге [47] и вошла в неё составной частью.

3.3e Энергетическая норма

Ещё одной важной и популярной конструкцией нормы является так называемая энергетическая норма векторов, которая порождается какой-либо симметричной положительно-определённой матрицей.⁸ 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 определить ортогональность, норму вектора x и т. п. В частности, нормой положим

$$\|x\|_A := \sqrt{\langle x, x \rangle_A} = \sqrt{\langle Ax, x \rangle}, \quad (3.27)$$

т. е. квадратный корень из произведения x на себя в этом скалярном произведении.

⁸Для комплексного случая обобщение очевидно, и мы не детализуем его лишь по причине экономии места.

Определение 3.3.6 Для симметричной 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 3.3.7 Для симметричной положительно определённой матрицы A векторная норма $\|\cdot\|_A$, задаваемая посредством (3.27), 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. §3.10а).

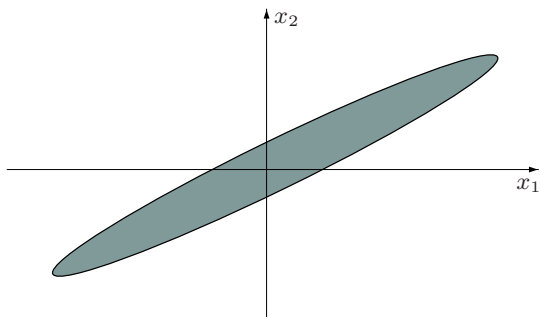


Рис. 3.8. Шар единичного радиуса в энергетической норме при значительном разбросе спектра порождающей матрицы

Так как симметричная матрица может быть приведена к диагональному виду ортогональными преобразованиями подобия, то

$$A = Q^T D Q,$$

где Q — ортогональная матрица, $D = \text{diag} \{ \lambda_1, \lambda_2, \dots, \lambda_n \}$ — диагональная матрица, на главной диагонали которой стоят положительные

собственные значения λ_i матрицы A . Поэтому

$$\begin{aligned}\|x\|_A &= \sqrt{\langle Ax, x \rangle} = \sqrt{\langle Q^\top D Q x, x \rangle} \\ &= \sqrt{\langle D Q x, Q x \rangle} = \sqrt{\langle D y, y \rangle} = \left(\sum_{i=1}^n \lambda_i y_i^2 \right)^{1/2},\end{aligned}\quad (3.28)$$

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 λ_i матрицы A , т.е. чем больше её число обусловленности $\text{cond}_2(A)$ (см. §3.5a).

Из сказанного вытекает характерная особенность энергетической нормы, которая в ряде случаев оборачивается её недостатком: возможность существенного искажения обычного геометрического масштаба объектов по разным направлениям (своеобразная 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. 3.8). 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 §3.10в, §3.10г и §3.10д.

Пример 3.3.3 Пусть

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}.$$
 (3.29)

Это положительно определённая матрица, которая может задавать энергетическое скалярное произведение и соответствующую A -норму.

Нетрудно проверить, что векторы

$$\begin{pmatrix} -2 \\ 1 \end{pmatrix} \quad \text{и} \quad \begin{pmatrix} 2 \\ 1 \end{pmatrix},$$

изображённые на Рис. 3.9, являются A -ортогональными для рассматриваемой матрицы A , хотя реальный угол между этими векторами — почти 127° . ■

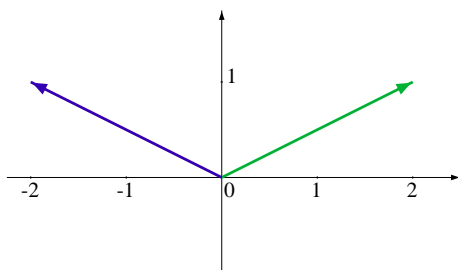


Рис. 3.9. A -ортогональные векторы относительно скалярного произведения, задаваемого матрицей (3.29).

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 (3.28) следует, что

$$\left(\min_i \lambda_i\right) \|x\|_2 \leq \|x\|_A \leq \left(\max_i \lambda_i\right) \|x\|_2,$$

где λ_i — собственные значения порождающей матрицы A . Другие двусторонние неравенства для энергетической нормы можно получить с помощью Предложения 3.3.7.

Выражения для матричных норм, которые подчинены энергетической норме векторов или просто согласованы с ней, выписываются непросто. Даже не всегда можно указать для них явный и несложно вычисляемый вид. Тем не менее, мы приведём полезный и красивый результат на эту тему, который будет далее использован при исследовании метода наискорейшего спуска в §3.10:

Предложение 3.3.8 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

$$\|Sx\|_A \leq \|S\|_2 \|x\|_A. \quad (3.30)$$

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 . Тогда

$$\begin{aligned} \|Sx\|_A^2 &= \langle ASx, Sx \rangle = \langle SAx, Sx \rangle \\ &= \langle Q\Sigma Q^\top Ax, Q\Sigma Q^\top x \rangle = \langle \Sigma Q^\top Ax, \Sigma Q^\top x \rangle \\ &\leq \left(\max_i s_i^2 \right) \langle Q^\top Ax, Q^\top x \rangle = \left(\max_i |s_i|^2 \right) \langle QQ^\top Ax, x \rangle \\ &= \left(\max_i (\sigma_i(S))^2 \right) \langle Ax, x \rangle = \|S\|_2^2 \|x\|_A^2 \end{aligned}$$

с учётом того, что $\max (\sigma_i(S))^2 = \|S\|_2^2$. ■

3.3ж Спектральный радиус

Определение 3.3.8 *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 (см. Рис. 3.10). 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 (см. [9, 37, 53]).

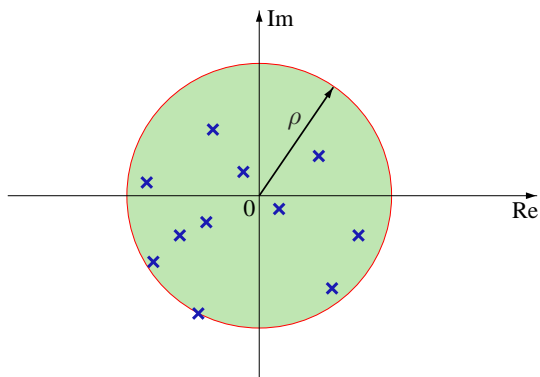


Рис. 3.10. Иллюстрация спектрального радиуса матрицы: крестиками обозначены точки спектра.

Теорема 3.3.1 *Спектральный радиус матрицы не превосходит любой её нормы.*

Доказательство. Рассмотрим сначала случай, когда матрица является комплексной.

Пусть λ — собственное значение матрицы A , а $v \neq 0$ — соответствующий собственный вектор, так что $Av = \lambda v$. Воспользуемся тем установленным в §3.3в фактом (Предложение 3.3.4), 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

$$\|A\| \cdot \|v\| \geq \|Av\| = \|\lambda v\| = |\lambda| \cdot \|v\|, \quad (3.31)$$

где $\|v\| > 0$, и потому сокращение на эту величину обеих частей неравенства (3.31) даёт $\|A\| \geq |\lambda|$. Коль скоро наше рассуждение справед-

ливо для любого собственного значения λ , то в самом деле $\max |\lambda| = \rho(A) \leq \|A\|$.

Рассмотрим теперь случай вещественной $n \times n$ -матрицы A . Если λ — её вещественное собственное значение, то проведённые выше рассуждения остаются полностью справедливыми. Если же λ — комплексное собственное значение матрицы A , то комплексным является и соответствующий собственный вектор v . Тогда цепочку соотношений (3.31) 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 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 (см. [10, 24, 37, 96]).

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

$$(x, y)^\top + (x', y')^\top = (x + x', y + y')^\top, \quad (3.32)$$

$$(\alpha + i\beta) \cdot (x, y)^\top = (\alpha x - \beta y, \alpha y + \beta x)^\top. \quad (3.33)$$

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

$$\mathcal{A} = \begin{pmatrix} A & 0 \\ 0 & A \end{pmatrix}. \quad (3.34)$$

Его блочно-диагональный вид объясняется тем, что согласно формуле (3.33) 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, так как такие нормы являются наименьшими из всех согласованных матричных норм (см. §3.3г). 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 $\|(x, y)^T\| = \|x\| + \|y\|$. Тогда ввиду (3.34) и с помощью рассуждений, аналогичных доказательству Предложения 3.3.6, нетрудно показать, что подчинённая матричная норма для \mathcal{A} во множестве $2n \times 2n$ -матриц есть $\|\mathcal{A}\| = \max\{\|A\|, \|A\|\} = \|A\|$. Кроме того, теперь для \mathcal{A} справедливы рассуждения о связи нормы и спектрального радиуса, проведённые в начале доказательства для случае 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 3.3.6 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 α 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

$$\begin{pmatrix} 0 & 1 & & 0 \\ & 0 & 1 & \\ & & \ddots & \ddots \\ 0 & & & 0 & 1 \\ & & & & 0 \end{pmatrix} \quad (3.35)$$

— жордановой клетки, отвечающей собственному значению 0, спектральный радиус равен нулю. Во-вторых, если A — матрица вида (3.35),

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 значений 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 3.3.9 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 λ — 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

$$(A^k)v = (\lambda^k)v. \quad (3.36)$$

Если последовательность степеней A^k , $k = 0, 1, 2, \dots$, ограничена, то при фиксированном векторе v ограничена также левая часть выписанного равенства (3.36). Как следствие, ограничена и правая часть в (3.36), 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 (3.36). Как следствие, к нулевому вектору должна сходиться и правая часть в (3.36), причём $v \neq 0$. Это возможно лишь в случае $|\lambda| < 1$. ■

Ниже в §3.9б мы увидим, что условие $\rho(A) < 1$ является, в действительности, достаточным для сходимости к нулю степеней матрицы A .

Рассуждения, с помощью которых доказано Предложение 3.3.9, можно продолжить и несложно вывести весьма тонкие свойства спектрального радиуса. Возьмём от обеих частей равенства (3.36) 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,

$$\rho(A) \leq \inf_{k \in \mathbb{N}} \|A^k\|^{1/k}. \quad (3.37)$$

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 (3.37) является *формула Гельфанда*

$$\rho(A) = \lim_{k \rightarrow \infty} \|A^k\|^{1/k},$$

которая верна для любой из матричных норм. Её доказательство можно найти, к примеру, в [53]. В целом, несмотря на то, что матрица является сложным составным объектом, нормы её степеней, как показывают Предложение 3.3.9, неравенство (3.37) и формула Гельфанда, ведут себя примерно так же, как геометрическая прогрессия со знаменателем, равным спектральному радиусу этой матрицы. Например, для

$n \times n$ -матрицы (3.35) или любой ей подобной n -ая степень зануляется, и это свойство обнаруживается спектральным радиусом.

3.3з 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 3.3.10 Пусть X — квадратная матрица и $\|X\| < 1$ в некоторой матричной норме. Тогда матрица $(I - X)$ неособенна, для обратной матрицы справедливо представление

$$(I - X)^{-1} = \sum_{k=0}^{\infty} X^k, \quad (3.38)$$

и имеет место оценка

$$\|(I - X)^{-1}\| \leq \frac{1}{1 - \|X\|}. \quad (3.39)$$

Аналог геометрической прогрессии для матриц, фигурирующий в правой части равенства (3.38), 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, беря от обеих частей этого равенства векторную норму, согласованную с матричной нормой, в которой $\|X\| < 1$ по условию Предложения, мы получим

$$\|X\| \|v\| \geq \|Xv\| = \|v\|.$$

В случае, когда $v \neq 0$, можем сократить обе части полученного 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 $S_N = \sum_{k=0}^N X^k$ — 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 , то 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. Эта важная и интересная тема, находящая многочисленные приложения; подробности можно увидеть, к примеру, в [9, 11, 24, 27].

3.4 Приложения сингулярного разложения

3.4a Исследование неособенности и ранга матриц

Рассмотренное в §3.2е сингулярное разложение матрицы может служить основой для вычислительных технологий решения многих важных 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 нулём. Но точное равенство определителя нулю искажается погрешностями, которые вносятся в процесс его вычисления. Кроме того, величина ненулевого определителя матрицы не является вполне адекватным признаком того, насколько близка матрица к особенной. Определитель очень сильно изменяется при умножении матрицы на число:

$$\det(\alpha A) = \alpha^n \cdot \det A \quad \text{для } n \times n\text{-матрицы } A.$$

Но 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 значении её 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. §3.166).

Наиболее надёжным в вычислительном отношении способом проверки особенности/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 (3.16) 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 мерилем 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. Хорошей количественной мерой особенности/неособенности, которая инвариантна относительно масштабирования матрицы, может также служить отношение её наибольшего и наименьшего сингулярных чисел, — так называемое число обусловленности матрицы относительно спектральной нормы (см. §3.5a).

Обсудим теперь задачу о вычислении ранга матрицы. Согласно 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. Как следствие, 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 минималь-

ный из рангов всех матриц.

Ранг диагональной матрицы равен числу её ненулевых диагональных элементов. Поэтому ранг произвольной матрицы равен количеству её ненулевых сингулярных чисел, что следует из сингулярного разложения $A = U\Sigma V^*$, и из того факта, что ортогональные преобразования сохраняют линейную зависимость или независимость. Следовательно, при нахождении приближённого ранга матрицы можно задаться каким-либо порогом малости ϵ , найти сингулярные числа матрицы и подсчитать, сколько из них больше или равны ϵ . 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.

3.46 Решение систем линейных уравнений

Если для матрицы A известно сингулярное разложение (3.16), то вещественная система линейных алгебраических уравнений $Ax = b$ может быть переписана эквивалентным образом как

$$U\Sigma V^T x = b.$$

Отсюда решение легко находится в виде

$$x = V\Sigma^{-1}U^T b.$$

Получается, что для вычисления решения мы должны умножить вектор правой части на ортогональную матрицу, затем разделить компоненты результата на сингулярные числа 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 (см. §3.6в) или других прямых методов решения СЛАУ. Но описанный путь безупречен с вычислительной точки зрения, так как позволяет без накопления ошибок

найти решение системы и, кроме того, проанализировать состояние её разрешимости, указав ранг матрицы системы (см. предшествующий пункт).

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. Ниже в §3.56 мы взглянем на этот факт с другой стороны. Отличие в поведении и результатах метода Гаусса и метода, основанного на сингулярном разложении, особенно зримо в случае, когда матрица системы «почти особенна».

Ещё большую пользу сингулярное разложение приносит при решении систем линейных алгебраических уравнений, в которых количество уравнений не совпадает с количеством неизвестных. Обычного решения такая система может не иметь, и это типично для переопределённых систем. Тогда находят *псевдорешения* системы линейных уравнений, которые минимизируют ту или иную норму невязки левой и правой частей, т. е. разности $Ax - b$. Сингулярное разложение матрицы системы является одним из главных инструментов нахождения псевдорешения таких систем для случая, когда минимизируется евклидова норма невязки, т. е. ищется псевдорешение в смысле «наименьших квадратов». Рассмотрим эту технологию более подробно.

Евклидова норма (2-норма) вектора не меняется при его умножении на ортогональную матрицу, и поэтому

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \|Ax - b\|_2 &= \min_{x \in \mathbb{R}^n} \|U \Sigma V^\top x - U U^\top b\|_2 \\ &= \min_{x \in \mathbb{R}^n} \|U(\Sigma V^\top x - U^\top b)\|_2 \\ &= \min_{x \in \mathbb{R}^n} \|\Sigma V^\top x - U^\top b\|_2 \\ &= \min_{y \in \mathbb{R}^n} \|\Sigma y - U^\top b\|_2, \end{aligned}$$

где выполнена неособенная замена переменной $V^\top x = y$. Если в системе уравнений $m \times n$ -матрица A такова, что $m \geq n$, то Σ — диагональная

матрица тех же размеров,

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix},$$

и $\sigma_1, \sigma_2, \dots, \sigma_n$ — 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

$$y_i^* = \frac{(U^\top b)_i}{\sigma_i}, \quad i = 1, 2, \dots, n. \quad (3.40)$$

Тогда это же верно для искомого $\min_{x \in \mathbb{R}^n} \|Ax - b\|_2$. Аргумент x^* искомого минимума, т.е. псевдорешение системы линейных уравнений $Ax = b$, находится обратной заменой $x^* = Vy^*$.

В формуле (3.40) 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 из постоянных величин, а y_r при этом можно взять произвольными.

Рассмотренная выше задача называется линейной задачей наименьших квадратов, и мы уже затрагивали её в §2.10е. Представленное здесь решение на основе сингулярного разложения матрицы является очень общим и весьма информативным, хотя его трудоёмкость больше, чем у других вычислительных методов. Их описанию посвящён §3.15.

3.4в Малоранговые приближения матрицы

Пусть A — $m \times n$ -матрица, u_k и v_k — это её k -ые нормированные левый и правый сингулярные векторы, а \mathcal{U}_k обозначает их внешнее произведение, т.е.

$$\mathcal{U}_k = u_k v_k^*.$$

Отметим, что Υ_k — $m \times n$ -матрица ранга 1. Тогда сингулярное разложение (3.16) матрицы A равносильно её представлению в виде суммы

$$A = \sum_{k=1}^n \sigma_k \Upsilon_k, \quad (3.41)$$

где σ_i , $i = 1, 2, \dots, \min\{m, n\}$, — сингулярные числа матрицы A . Если $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n$, и мы «обрубаем» выписанную сумму после p -го слагаемого ($p \leq n$), то получающаяся матрица

$$A_p = \sum_{k=1}^p \sigma_k \Upsilon_k, \quad (3.42)$$

называется *p -ранговым приближением* матрицы A .

Это в самом деле матрица ранга p , что следует из её сингулярного разложения, и погрешность, с которой она приближает исходную матрицу, равна

$$\sum_{k=p+1}^n \sigma_k \Upsilon_k.$$

Величина этой погрешности решающим образом зависит от величины сингулярных чисел $\sigma_{p+1}, \dots, \sigma_{\min\{m, n\}}$, соответствующих отброшенным слагаемым в (3.41). Более точно, погрешность p -рангового приближения характеризуется следующим замечательным свойством:

Теорема 3.4.1 Пусть σ_k , u_k и v_k — сингулярные числа, левые и правые сингулярные векторы $m \times n$ -матрицы A соответственно. Если $p < n$ и

$$A_p = \sum_{k=1}^p \sigma_k u_k v_k^*$$

— p -ранговое приближение матрицы A , то

$$\|A - A_p\|_2 = \min_{\substack{B \in \mathbb{C}^{m \times n} \\ \text{rank } B \leq p}} \|A - B\|_2 = \sigma_{p+1}.$$

Иными словами, относительно спектральной нормы p -ранговое приближение матрицы обеспечивает наименьшее отклонение от исходной матрицы среди всех матриц ранга не более p .

Доказательство. Предположим, что найдётся такая матрица B , имеющая ранг $\text{rank } B \leq p$, что $\|A - B\|_2 < \|A - A_p\|_2 = \sigma_{p+1}$. Тогда существует $(n - p)$ -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

$$\|Aw\|_2 = \|(A - B)w\|_2 \leq \|A - B\|_2 \|w\|_2 < \sigma_{p+1} \|w\|_2.$$

Таким образом, W является $(n - p)$ -мерным подпространством в \mathbb{C}^n , в котором $\|Aw\|_2 < \sigma_{p+1} \|w\|_2$.

Но в \mathbb{C}^n имеется $(p + 1)$ -мерное подпространство, образованное векторами v , для которых $\|Av\|_2 \geq \sigma_{p+1} \|v\|_2$. Это подпространство, являющееся линейной оболочкой первых $p + 1$ правых сингулярных векторов матрицы A . Поскольку сумма размерностей этого подпространства и подпространства W превосходит n , размерности всего пространства, то 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 3.4.1:

Теорема 3.4.2 (теорема Экарта-Янга [101]) Пусть σ_k , u_k и v_k — сингулярные числа и левые и правые сингулярные векторы $m \times n$ -матрицы A соответственно. Если $p < n$ и

$$A_p = \sum_{k=1}^p \sigma_k u_k v_k^*$$

— p -ранговое приближение матрицы A , то

$$\|A - A_p\|_F = \min_{\substack{B \in \mathbb{C}^{m \times n} \\ \text{rank } B \leq p}} \|A - B\|_F = \sigma_{p+1},$$

где $\|\cdot\|_F$ — фробениусова норма матриц. Иными словами, относительно фробениусовой нормы p -ранговое приближение матрицы обеспечивает наименьшее отклонение от исходной матрицы среди всех матриц ранга не более p .

Доказательство опускается.

Итак, если младшие сингулярные числа матрицы достаточно малы, то вместо неё можно взять p -ранговое приближение вида (3.42). Оно более «экономно», т. е. с меньшим числом параметров приближённо представляет исходную матрицу.

3.4г Метод главных компонент

В качестве важного практического примера, который иллюстрирует понятия ранга матрицы, сингулярных чисел и сингулярных векторов матрицы, а также результаты предыдущего пункта, рассмотрим *метод главных компонент*, широко применяемый в анализе данных и статистике. В этих дисциплинах решаются задачи обработки больших массивов числовых данных, характеризующих какой-либо объект или явление. Предположим для определённости, что рассматриваемый объект 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 вещественную $m \times n$ -матрицу, которую мы обозначим через A .

Нередко возникает необходимость сжатия данных, т. е. уменьшения числа n параметров объекта с тем, чтобы оставшиеся p признаков, $p < n$, 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 $\{e_1, e_2, \dots, e_p\}$, $p < n$, 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., к примеру, [87] или любой другой учебник по математической статистике). Для фробениусовой нормы матриц наша математическая задача ставится следующим образом. Нужно найти такой ортонормированный базис $\{e_1, e_2, \dots, e_p\}$ в \mathbb{R}^n , $p \leq n$, что квадратичное отклонение набора исходных векторов данных $A_i = (a_{i1}, a_{i2}, \dots, a_{in})^\top$ от их приближений $X^{(i)} = \sum_{j=1}^p x_{ij} e_j$ 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. Опирающаяся 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 $A^T A$, 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 перекрывались 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.

3.5 Обусловленность систем линейных уравнений

3.5a Число обусловленности матриц

В этом параграфе 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, развитые в §1.6, рассматриваются здесь в приложении к задаче решения систем линейных уравнений.

Рассмотрим систему линейных алгебраических уравнений

$$Ax = b \tag{3.43}$$

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 (3.43), получим

$$(\Delta A)x + (A + \Delta A)\Delta x = \Delta b, \quad (3.44)$$

или

$$(\Delta A)(x + \Delta x) + A\Delta x = \Delta b.$$

Вспоминая, что $x + \Delta x = \tilde{x}$, можно заключить

$$\Delta x = A^{-1}(-(\Delta A)\tilde{x} + \Delta b).$$

Для оценки величины изменения решения Δx 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

$$\begin{aligned} \frac{\|\Delta x\|}{\|\tilde{x}\|} &\leq \|A^{-1}\| \cdot \left(\|\Delta A\| + \frac{\|\Delta b\|}{\|\tilde{x}\|} \right) \\ &= \|A^{-1}\| \|A\| \cdot \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|A\| \cdot \|\tilde{x}\|} \right). \end{aligned} \quad (3.45)$$

Это весьма практичная *апостериорная оценка* относительной погрешности решения, которую удобно применять после того, как приближённое решение системы уже найдено.⁹ Коль скоро $\|A\| \cdot \|\tilde{x}\| \geq$

⁹От латинского словосочетания «a posteriori», означающего знание, полученное из опыта. Под «опытом» здесь, естественно, понимается процесс решения задачи.

$\|A\tilde{x}\| \approx \|b\|$, то знаменатель второго слагаемого в скобках из правой части неравенства «приблизительно не меньше», чем $\|b\|$. Поэтому полученной оценке (3.45) путём некоторого огрубления можно придать более элегантный вид

$$\frac{\|\Delta x\|}{\|\tilde{x}\|} \lesssim \|A^{-1}\| \|A\| \cdot \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|} \right), \quad (3.46)$$

в котором справа задействованы относительные погрешности в матрице A и правой части b .

Фигурирующая в оценках (3.45) и (3.46) величина $\|A^{-1}\| \|A\|$, на которую суммарно умножаются относительные ошибки в матрице и правой части, имеет своё собственное название, так как играет важнейшую роль в вычислительной линейной алгебре.

Определение 3.5.1 Для квадратной неособенной матрицы A величина $\|A^{-1}\| \|A\|$ называется её числом обусловленности (относительной выбранной нормы матрицы).

Понятие числа обусловленности введено А. Тьюрингом в 1948 году в работе [113]. Мы будем обозначать число обусловленности матрицы A посредством $\text{cond}(A)$, иногда с индексом, указывающим выбор нормы.¹⁰ 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.

Выведем теперь *априорную* оценку относительной погрешности ненулевого решения, которая не будет опираться на знание вычисленного решения и может быть применена *до* того, как мы начнём решать СЛАУ.¹¹

После вычитания точного уравнения из приближённого мы получили (3.44):

$$(\Delta A)x + (A + \Delta A)\Delta x = \Delta b.$$

¹⁰В математической литературе для числа обусловленности матрицы A иногда можно встретить обозначения $\mu(A)$ или $\kappa(A)$.

¹¹От латинского словосочетания «a priori», означающего в философии знание, полученное до опыта и независимо от него.

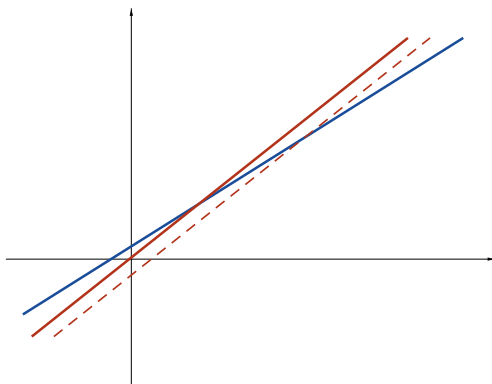


Рис. 3.11. Иллюстрация возмущения системы линейных уравнений с плохой обусловленностью матрицы: малые «шевеления» любой прямой приводят к большим изменениям в решении.

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

$$\|\Delta x\| \leq \|(I + A^{-1}\Delta A)^{-1}\| \cdot \|A^{-1}\| \cdot (\|\Delta A\| \|x\| + \|\Delta b\|),$$

откуда после деления обеих частей на $\|x\| > 0$:

$$\frac{\|\Delta x\|}{\|x\|} \leq \|(I + A^{-1}\Delta A)^{-1}\| \cdot \|A^{-1}\| \cdot \left(\|\Delta A\| + \frac{\|\Delta b\|}{\|x\|} \right).$$

Предположим, что возмущение ΔA 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 (3.38). Соответственно, мы можем воспользоваться вытекающей из этого оценкой (3.39). Тогда

$$\begin{aligned} \frac{\|\Delta x\|}{\|x\|} &\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \|\Delta A\|} \cdot \left(\|\Delta A\| + \frac{\|\Delta b\|}{\|x\|} \right) \\ &= \frac{\|A^{-1}\| \cdot \|A\|}{1 - \|A^{-1}\| \|\Delta A\|} \cdot \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|A\| \|x\|} \right) \\ &\leq \frac{\text{cond}(A)}{1 - \text{cond}(A) \cdot \frac{\|\Delta A\|}{\|A\|}} \cdot \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|} \right), \end{aligned} \quad (3.47)$$

поскольку $\|A\| \|x\| \geq \|Ax\| = \|b\|$.

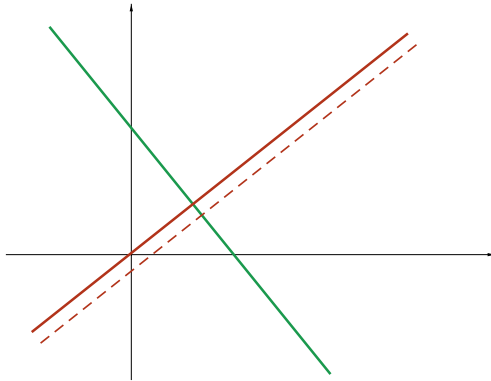


Рис. 3.12. Иллюстрация возмущения системы линейных уравнений с хорошей обусловленностью матрицы: «шевеления» прямых приводят к соизмеримым изменениям в решении.

Оценка (3.47) — 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 .

Понятие числа 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 преобладанием, рассматриваемые далее в §3.5в.

Существует также практически важный 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 3.2.5), 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

$$\text{cond}_2(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}. \quad (3.48)$$

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 §3.15 (см. [11, 13]). Вообще, соотношение (3.48) 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

$$\text{cond}_2(A) = \frac{\max_i |\lambda_i(A)|}{\min_i |\lambda_i(A)|} \quad (3.49)$$

— спектральное число обусловленности равно отношению наибольшего и наименьшего модулей собственных значений матрицы. Для симметричных положительно определённых матриц эта формула принимает совсем простой вид

$$\text{cond}_2(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}.$$

Отметим, что простые оценки минимальных сингулярных чисел существуют также для матриц с диагональным преобладанием [114, 115].

3.56 Примеры хорошообусловленных и плохообусловленных матриц

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.

Отметим, что для любой подчинённой матричной нормы

$$\text{cond}(A) = \|A^{-1}\| \|A\| \geq \|A^{-1}A\| = \|I\| = 1$$

в силу (3.24), и поэтому соответствующее число обусловленности матрицы 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, ∞ -нормы и спектральной нормы имеет единичная матрица.

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, $Q^{-1} = Q^T$ и тоже ортогональна, а потому $\|Q^{-1}\|_2 = 1$. Как следствие, $\text{cond}_2(Q) = 1$. Нетрудно также понять, что любая матрица, пропорциональная ортогональной, т. е. получающаяся из ортогональной умножением на ненулевое число, тоже имеет обусловленность 1 относительно спектральной нормы.

Самым популярным содержательным примером плохообусловленных матриц являются, пожалуй, матрицы Гильберта $H_n = (h_{ij})$, которые встретились нам в §2.10ж при обсуждении среднеквадратичного приближения алгебраическими полиномами на интервале $[0, 1]$. Это симметричные матрицы, 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:

$$\text{cond}_2(H_2) = 19.3,$$

$$\text{cond}_2(H_3) = 524,$$

...

$$\text{cond}_2(H_{10}) = 1.6 \cdot 10^{13},$$

...

Существует общая формула (см. [67, 118, 112]):

$$\text{cond}_2(H_n) = O\left(\frac{(1+\sqrt{2})^{4n}}{\sqrt{n}}\right) \approx O(34^n/\sqrt{n}),$$

где O — «о большое», известный из математического анализа символ Э.Ландау (см. стр. 120). Интересно, что матрицы, обратные к матрицам Гильберта могут быть вычислены явно с помощью аналитических выкладок [74, 102]. Они имеют целочисленные элементы, которые тоже очень быстро растут с размером матрицы.

Для матрицы Вандермонда (2.7) оценка снизу для числа обусловленности (см. [57])

$$\text{cond}_2(V(x_0, x_1, \dots, x_n)) \geq \sqrt{2} \frac{(1 + \sqrt{2})^{n-1}}{\sqrt{n+1}} \quad (3.50)$$

представляется существенно более скромной, хотя она всё-таки растёт экспоненциально с n .¹² Но оценка (3.50), не зависящая от значений x_0, x_1, \dots, x_n , является весьма грубой, и реальные матрицы Вандермонда, как правило, обусловлены гораздо хуже. По этой причине мы также относим матрицы Вандермонда к «плохообусловленным».

Последним примером рассмотрим верхнюю треугольную $n \times n$ -матрицу

$$U = \begin{pmatrix} 1 & -1 & -1 & \cdots & -1 \\ 0 & 1 & -1 & \cdots & -1 \\ 0 & 0 & 1 & \cdots & -1 \\ \vdots & \vdots & \ddots & \ddots & -1 \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}, \quad (3.51)$$

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 (3.70):

$$\begin{aligned} x_n &= y_n, \\ x_{n-1} &= y_{n-1} + y_n, \\ x_{n-2} &= y_{n-2} + y_{n-1} + 2y_n \\ x_{n-3} &= y_{n-3} + y_{n-2} + 2y_{n-1} + 4y_n \\ &\vdots \quad \quad \quad \ddots \quad \quad \quad \vdots \\ x_1 &= y_1 + y_2 + 2y_3 + \dots + 2^{n-2}y_n. \end{aligned}$$

¹²Аналогичные по смыслу, но более слабые экспоненциальные оценки снизу для числа обусловленности матрицы Вандермонда выводятся в книге [44].

Можем заключить, что обратная матрица U^{-1} равна

$$U^{-1} = \begin{pmatrix} 1 & 1 & 2 & 4 & \dots & 2^{n-3} & 2^{n-2} \\ 0 & 1 & 1 & 2 & \dots & 2^{n-4} & 2^{n-3} \\ 0 & 0 & 1 & 1 & \dots & 2^{n-5} & 2^{n-4} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}.$$

Как следствие, обусловленность $n \times n$ -матрицы (3.51) в подчинённой 1-норме, к примеру, равна

$$\|U\|_1 \|U^{-1}\|_1 = n (1 + 1 + 2 + 2^2 + \dots + 2^{n-2}) = n 2^{n-1}.$$

Точно такое же значение имеет обусловленность матрицы U относительно подчинённой чебышёвской нормы (∞ -нормы матриц). Для средних размеров матриц, возникающих в задачах математического моделирования (скажем, при $n \approx 100$) число $n 2^{n-1}$ уже очень велико, и соответствующие матрицы нужно считать плохообусловленными. Этот пример замечателен своей обыденностью и умеренными значениями элементов матрицы U , за которой, 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, 3.6з, 3.7д, 3.7е, 3.9е и др.).

3.5в Матрицы с диагональным преобладанием

В приложениях линейной алгебры и теории матриц часто возникают матрицы, в которых диагональные элементы в том или ином смысле «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 пункт.

Следует отметить, что сам смысл, вкладываемый в понятие «преобладания» может быть различен, и ниже мы рассмотрим простейший и наиболее популярный.

Определение 3.5.2 *Квадратную $n \times n$ -матрицу $A = (a_{ij})$ называют матрицей с диагональным преобладанием, если для любого $i = 1, 2, \dots, n$ имеет место*

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|. \quad (3.52)$$

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

$$|a_{ii}| \geq \sum_{j \neq i} |a_{ij}| \quad (3.53)$$

для любого $i = 1, 2, \dots, n$. Иногда в связи с условиями (3.52) и (3.53) 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 3.5.1 (признак неособенности Адамара)

Квадратная матрица с диагональным преобладанием неособенна.

Доказательство. Предположим, что, вопреки доказываемому, рассматриваемая матрица $A = (a_{ij})$ является особенной. Тогда её столбцы линейно зависимы, и для некоторого ненулевого n -вектора $y = (y_1, y_2, \dots, y_n)^\top$ выполняется равенство $Ay = 0$, т. е.

$$\sum_{j=1}^n a_{ij} y_j = 0, \quad i = 1, 2, \dots, n. \quad (3.54)$$

Выберем среди компонент вектора y ту, которая имеет наибольшее абсолютное значение. Пусть её номер — ν , 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 ν -го из равенств (3.54) является соотношение

$$-a_{\nu\nu} y_\nu = \sum_{j \neq \nu} a_{\nu j} y_j,$$

которое влечёт цепочку оценок

$$\begin{aligned} |a_{\nu\nu}| |y_\nu| &= \left| \sum_{j \neq \nu} a_{\nu j} y_j \right| \leq \sum_{j \neq \nu} |a_{\nu j}| |y_j| \\ &\leq \left(\max_{1 \leq j \leq n} |y_j| \right) \sum_{j \neq \nu} |a_{\nu j}| = |y_\nu| \sum_{j \neq \nu} |a_{\nu j}|. \end{aligned}$$

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 (3.52), т. е. наличию диагонального преобладания в матрице A . Итак, A действительно должна быть неособенной матрицей. ■

Доказанный выше результат часто именуют «теоремой Леви-Деспланка» (см., к примеру, [44, 53]), но мы придерживаемся здесь терминологии, принятой в [9, 91]. В книге М. Пароди [91] можно прочитать, в частности, некоторые сведения об истории вопроса.

Следствие. Матрица с диагональным преобладанием является строго регулярной (см. Определение 3.6.2, стр. 371). 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 обобщен на матрицы, удовлетворяющие нестрогому диагональному преобладанию (3.53).

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 3.5.2 (теорема Таусски) *Если для квадратной неразложимой матрицы A выполнены условия нестрогого диагонального преобладания (3.53), причём хотя бы одно из этих неравенств выполнено строго, то матрица A неособенна.*

Доказательство можно найти, к примеру, в [9].

Ещё одно полезное свойство матриц с диагональным преобладанием

Теорема 3.5.3 (теорема Алберга-Нильсона [62])¹³ Пусть $A = (a_{ij})$ — $n \times n$ -матрица с диагональным преобладанием и

$$\alpha := \min_{1 \leq i \leq n} \left\{ |a_{ii}| - \sum_{j \neq i} |a_{ij}| \right\}. \quad (3.55)$$

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

$$\alpha \|y\|_{\infty} \leq \|Ay\|_{\infty}. \quad (3.56)$$

¹³В англоязычной литературе этот результат нередко называют «теоремой Верз» по имени автора переоткрывшей его работы [114] (см. также [115]).

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 $|y_k| > 0$:

$$0 < \alpha |y_k| \leq |a_{kk}| |y_k| - \sum_{j \neq k} |a_{kj}| |y_k|.$$

Очевидно, что полученное неравенство только усилится, если заменить в сумме из правой части множителя $|y_k|$ 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

$$\begin{aligned} 0 < \alpha |y_k| &\leq |a_{kk} y_k| - \sum_{j \neq k} |a_{kj} y_j| \\ &\leq \left| \sum_{j=1}^n a_{kj} y_j \right| \leq \max_{1 \leq k \leq n} \left| \sum_{j=1}^n a_{kj} y_j \right| = \|Ay\|_\infty. \end{aligned}$$

Вспоминая, что $|y_k| = \|y\|_\infty$, можем заключить, что в самом деле выполняется неравенство (3.56). ■

Фактически, в теореме Алберга-Нильсона вводится количественная мера диагонального преобладания в виде величины α , задаваемой (3.55), и ∞ -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 [114, 115], где, в частности, даются оценки минимального сингулярного числа матрицы с диагональным преобладанием.

Пример 3.5.1 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

$$\frac{1}{6} \begin{pmatrix} 2(h_1 + h_2) & h_2 & & & 0 \\ h_2 & 2(h_2 + h_3) & h_3 & & \\ & h_3 & 2(h_3 + h_4) & \ddots & \\ & & \ddots & \ddots & \ddots \\ 0 & & & h_{n-1} & 2(h_{n-1} + h_n) \end{pmatrix}$$

с $h_i > 0$, и поэтому её подчинённая чебышёвская норма равна

$$\frac{1}{6} \max \left\{ 2h_1 + 3h_2, 3 \max_{2 \leq i \leq n-2} (h_i + h_{i+1}), 3h_{n-1} + 2h_n \right\}. \quad (3.57)$$

Мера диагонального преобладания этой матрицы в смысле теоремы Алберга-Нильсона

$$\frac{1}{6} \min \left\{ 2h_1 + h_2, \min_{2 \leq i \leq n-2} (h_i + h_{i+1}), h_{n-1} + 2h_n \right\}, \quad (3.58)$$

и отношение величин (3.57) и (3.58) даст 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 (3.57) имеем h , а вместо (3.58) — $\frac{1}{3}h$, и искомая обусловленность равна всего 3. Столь же невелико число обусловленности для сеток, которые не сильно отличаются от равномерных. ■

3.5г Практическое применение числа обусловленности матриц

Оценки (3.45) и (3.47) на возмущения решений систем линейных алгебраических уравнений являются неулучшаемыми на всём множестве матриц, векторов правых частей и их возмущений. Более точно, для системы с данной 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 (3.47) для спектральной нормы (см. рассуждения в §3.4б). 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 малоприспособно 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 (3.45) и (3.47) результаты типично завышены во много раз (иногда на порядки), и для решения упомянутой задачи более предпочтительны методы интервального анализа (см., к примеру, [98, 108]).

Пример 3.5.2 Рассмотрим 2×2 -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 си-

ствем 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 [98, 108], что задана *интервальная система линейных алгебраических уравнений*

$$\begin{pmatrix} [2, 4] & [-2, 0] \\ [-1, 1] & [2, 4] \end{pmatrix} x = \begin{pmatrix} [-1, 1] \\ [0, 2] \end{pmatrix}. \quad (3.59)$$

Её *множеством решений* называют множество, образованное всевозможными решениями систем линейных алгебраических уравнений того же вида, у которых коэффициенты матрицы и компонентны правой части принадлежат заданным интервалам. Множество решений рассматриваемой нами системы (3.59) изображено на Рис. 3.13.¹⁴ Мы более подробно рассматриваем интервальные линейные системы уравнений в §4.6.

Подсчитаем оценки возмущений, которые получаются для решения системы (3.59) на основе числа обусловленности. Можно рассматривать (3.59), как систему, получающуюся путём возмущения «средней системы»

$$\begin{pmatrix} 3 & -1 \\ 0 & 3 \end{pmatrix} x = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

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 возмущение правой части —

$$\Delta b = \begin{pmatrix} \Delta b_1 \\ \Delta b_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \|\Delta b\|_\infty \leq 1.$$

¹⁴Этот рисунок получен с помощью свободного пакета программ IntLinIncR2 [97].

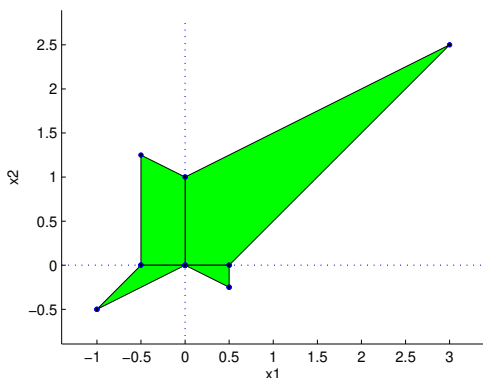


Рис. 3.13. Множество решений интервальной линейной системы (3.59).

Чебышёвская векторная норма (∞ -норма) используется здесь для оценки Δb потому, что она наиболее адекватно (без искажения формы) описывает возмущение правой части b . Соответствующая ∞ -норма для матрицы ΔA , подчинённая векторной ∞ -норме, также наиболее уместна в этой ситуации, поскольку она обеспечивает наиболее аккуратное согласование вычисляемых оценок (хотя и искажая немного форму множества возмущений).

Обусловленность средней матрицы относительно ∞ -нормы равна 1.778, ∞ -норма средней матрицы равна 4, а ∞ -норма средней правой части — это 1. Следовательно, по формуле (3.47) получаем

$$\frac{\|\Delta x\|}{\|x\|} \approx 24.$$

Поскольку решение средней системы есть $\tilde{x} = (\frac{1}{3}, \frac{1}{9})^T$, и оно имеет ∞ -норму $\frac{1}{3}$, то оценкой разброса решений рассматриваемой системы уравнений является $\tilde{x} \pm \Delta x$, где $\|\Delta x\|_\infty \leq 8$, т. е. двумерный брус¹⁵

$$\begin{pmatrix} [-7.667, 8.333] \\ [-7.889, 8.111] \end{pmatrix}.$$

По размерам он в более чем в 4 (четыре) раза превосходит оптимальные (точные) покоординатные оценки множества решений, которые удобно

¹⁵Читатель может проверить числовые данные этого примера в любой системе компьютерной математики: Scilab, МАТЛАВ, 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 (3.47), совершенно аналогичны своей грубостью оценивания возмущений решений. ■

Отметим в заключение этой темы, что задача оценивания разброса решений СЛАУ при вариациях входных данных является в общем случае NP-трудной [104, 105]. Иными словами, она требует для своего решения экспоненциально больших трудозатрат, если мы не накладываем никаких ограничений на величину возмущений в данных,

3.6 Прямые методы решения систем линейных алгебраических уравнений

3.6a Основные понятия

Решение систем линейных алгебраических уравнений вида

$$\left\{ \begin{array}{l} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = b_2, \\ \vdots \qquad \qquad \qquad \vdots \qquad \ddots \qquad \qquad \vdots \qquad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n = b_m, \end{array} \right. \tag{3.60}$$

с коэффициентами a_{ij} и свободными членами b_i , или, в краткой форме,

$$Ax = b \quad (3.61)$$

с $m \times n$ -матрицей $A = (a_{ij})$ и m -вектором правой части $b = (b_i)$, является важной математической задачей. Она часто встречается как сама по себе, так и в качестве составного элемента в технологической цепочке 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 (3.60)–(3.61) так называемыми прямыми методами.

Следует отметить, что системы линейных алгебраических уравнений не всегда предъявляются к решению в каноническом виде (3.60). 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 $\mathcal{D} = [\underline{x}_1, \overline{x}_1] \times [\underline{x}_2, \overline{x}_2]$, имеющая форму прямоугольника со сторонами, параллельными координатным осям. Рассмотрим в ней численное решение дифференциального уравнения Лапласа

$$\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = 0 \quad (3.62)$$

для функции двух переменных $u = u(x_1, x_2)$.

Уравнением Лапласа является одним из основных уравнений математической физики, с помощью которого описывается, к примеру, распределение температуры стационарного теплового поля, потенциал электростатического поля или же 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:

$$u(\underline{x}_1, x_2) = \underline{f}(x_2), \quad u(\overline{x}_1, x_2) = \overline{f}(x_2), \quad (3.63)$$

$$u(x_1, \underline{x}_2) = \underline{g}(x_1), \quad u(x_1, \overline{x}_2) = \overline{g}(x_1). \quad (3.64)$$

Рассматриваемую задачу определения функции $u(x_1, x_2)$, которая удовлетворяет уравнению (3.62) внутри области и условиям (3.63)–(3.64) на границе, называют *задачей Дирихле* для уравнения Лапласа.

Станем решать задачу (3.62)–(3.64) 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 $[\underline{x}_1, \overline{x}_1]$ на m частей, а интервал $[\underline{x}_2, \overline{x}_2]$ — на n частей. Вместо функции $u(x_1, x_2)$ непрерывного аргумента будем рассматривать её значения в узлах построенной сетки (см. Рис. 3.14), которые обозначим через x_{ij} , $i = 0, 1, \dots, m$, $j = 0, 1, \dots, n$.

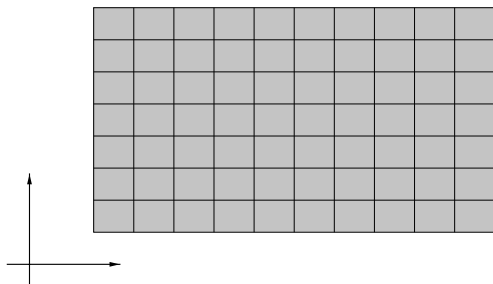


Рис. 3.14. Расчётная область и сетка для численного решения уравнения Лапласа (3.62).

Если обозначить через u_{ij} значение искомой функции u в точке x_{ij} , то после замены вторых производных формулами (2.81) получим систему соотношений вида

$$\frac{u_{i-1,j} - 2u_{ij} + u_{i+1,j}}{h_1^2} + \frac{u_{i,j-1} - 2u_{ij} + u_{i,j+1}}{h_2^2} = 0, \quad (3.65)$$

$i = 1, 2, \dots, m-1, j = 1, 2, \dots, n-1$, для внутренних узлов расчётной области. На границе области имеем условия

$$u_{i0} = \underline{f}_i, \quad u_{in} = \overline{f}_i, \quad (3.66)$$

$$u_{0j} = \underline{g}_j, \quad u_{mj} = \overline{g}_j, \quad (3.67)$$

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 $\underline{f}, \overline{f}, \underline{g}, \overline{g}$ в соответствующих узлах.

Соотношения (3.65) и (3.63)–(3.64) 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 (3.60), так как неизвестные имеют по два индекса. Конкретный вид (3.60), который получит решаемая 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, который соответствует (3.60), 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 основных 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 остаётся в матрице преобразованной системы, тем проще и устойчивее её решение, но, с другой стороны, 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, т. е. к невозможности найти значения всех неизвестных переменных.

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 (см., к примеру, [75]) или треугольному виду. Мы, в основном, рассмотрим методы, основанные на приведении к треугольному виду, так как именно они получили наибольшее распространение в практике вычислений.

Наконец, для простоты мы далее подробно разбираем системы линейных уравнений (3.60)–(3.61), в которых $m \times n$ -матрица коэффици-

ентов $A = (a_{ij})$ имеет полный ранг. В частности, A неособенна при $m = n$.

3.6б Решение треугольных и трапецевидных линейных систем

Напомним, что *треугольными матрицами* называют матрицы, у которых все элементы ниже главной диагонали либо все элементы выше главной диагонали — нулевые (так что и нулевые, и ненулевые элементы образуют треугольники):

$$U = \begin{pmatrix} \times & \times & \cdots & \times & \times \\ & \times & \ddots & \times & \times \\ & & \ddots & \vdots & \vdots \\ 0 & & & \times & \times \\ & & & & \times \end{pmatrix}, \quad L = \begin{pmatrix} \times & & & 0 & \\ \times & \times & & & \\ \times & \times & \ddots & & \\ \vdots & \vdots & \ddots & \times & \\ \times & \times & \cdots & \times & \times \end{pmatrix},$$

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. Соответственно, *треугольными* называются системы линейных алгебраических уравнений, матрицы которых имеют треугольный вид — верхний или нижний.

Рассмотрим для определённости линейную систему уравнений

$$Lx = b \tag{3.68}$$

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 (3.68) с нижней треугольной $n \times n$ -матрицей:

```
DO FOR  $i = 1$  TO  $n$ 
```

$$x_i \leftarrow \left(b_i - \sum_{j < i} l_{ij} x_j \right) / l_{ii} . \quad (3.69)$$

```
END DO
```

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 и 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
```

$$x_i \leftarrow \left(b_i - \sum_{j > i} u_{ij} x_j \right) / u_{ii} . \quad (3.70)$$

```
END DO
```

Трапецевидные матрицы — это обобщение треугольных матриц на прямоугольный случай, и нам далее особенно интересны верхние (пра-

вые) трапецевидные матрицы —

$$\begin{pmatrix} \times & \times & \cdots & \times & \times & \cdots & \times \\ & \times & \ddots & \times & \times & \cdots & \times \\ & & \ddots & \vdots & \vdots & \cdots & \times \\ 0 & & & \times & \times & \cdots & \times \\ & & & \times & \cdots & \times \end{pmatrix}, \quad \begin{pmatrix} \times & \times & \cdots & \times & \times \\ & \times & \ddots & \times & \times \\ & & \ddots & \vdots & \vdots \\ 0 & & & \times & \times \\ & & & \times \end{pmatrix},$$

и 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 «лежащие» матрицы), у которых $m < n$, необходимо предварительно перенести в правую часть члены с переменными x_{m+1}, \dots, x_n , сделав их свободными параметрами.

Для переопределённых систем, у которых $m > n$ (имеющих «стоячие» матрицы), точное решение существует лишь когда все правые части с $n + 1$ -ой по m -ую — нулевые. В этом случае мы также найдём его с помощью обратной подстановки (3.70). Если правые части с $n + 1$ -ой по m -ую — ненулевые, то система уравнений обычного решения не имеет, но с помощью обратной подстановки (3.70) легко находится её псевдорешение.

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 К.Ф. Гауссом в 1849 году, хотя письменные источники свидетельствуют о том, что он был известен как минимум за 250 лет до нашей эры.

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.

Выполнив $(n - 1)$ шагов подобного процесса — для 1-го, 2-го, \dots , $(n - 1)$ -го столбцов матрицы данной системы, мы получим, в конце концов, линейную систему с верхней треугольной матрицей, которая несложно решается с помощью обратной подстановки, рассмотренной ранее в §3.6б. Описанное преобразование системы линейных алгебраических уравнений к равносильному треугольному виду называется *прямым ходом* метода Гаусса, и его псевдокод выглядит следующим образом:

```
DO FOR  $j = 1$  TO  $n - 1$ 
  DO FOR  $i = j + 1$  TO  $n$ 
     $r_{ij} \leftarrow (-a_{ij}/a_{jj})$ 
    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
```

(3.71)

Он выражает процесс последовательного обнуления поддиагональных элементов j -го столбца матрицы системы, $j = 1, 2, \dots, n - 1$, и соответствующие преобразования вектора правой части. Матрица системы при этом приводится к верхнему треугольному виду. Отметим, что в псевдокоде (3.71) зануление поддиагональных элементов первых столбцов уже учтено нижней границей внутреннего цикла по k , которая равна j , а не 1.

После прямого хода метода Гаусса следует его *обратный ход*, на котором решается полученная верхняя треугольная система:


```
DO FOR  $i = n$  DOWNT0 1
```

$$x_i \leftarrow \left(b_i - \sum_{j>i} a_{ij}x_j \right) / a_{ii} . \quad (3.72)$$

```
END DO
```

Обратный ход является ни чем иным, как процессом обратной подстановки из §3.6б, в котором последовательно вычисляются, в обратном порядке, искомые значения неизвестных, начиная с n -ой.

Помимо изложенной выше вычислительной схемы существует много других версий метода Гаусса. 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

$$x_1 + \frac{a_{12}}{a_{11}} x_2 + \cdots + \frac{a_{1n}}{a_{11}} x_n = \frac{b_1}{a_{11}}. \quad (3.73)$$

Умножая затем уравнение (3.73) на a_{i1} , вычитают 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 (3.72).

Схема единственного деления совершенно эквивалентна алгоритму (3.71) 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 -ым уравнением могут быть представлены в матричном виде

как умножение обеих частей системы $Ax = b$ слева на матрицу

$$\begin{pmatrix} 1 & & & & \\ 0 & 1 & & & \\ \vdots & & \ddots & & \\ r_{i1} & & & 1 & \\ \vdots & & & & 1 \\ 0 & 0 & & & 1 \end{pmatrix},$$

которая отличается от единичной матрицы наличием одного дополнительного ненулевого элемента r_{i1} на месте $(i, 1)$.¹⁶ Исклучение поддиагональных элементов первого столбца матрицы СЛАУ в прямом ходе метода Гаусса (3.71) — это последовательное домножение обеих частей этой системы слева на матрицы

$$\begin{pmatrix} 1 & & & & \\ r_{21} & 1 & & & \\ 0 & & \ddots & & \\ \vdots & & & 1 & \\ 0 & 0 & & & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & & & & \\ 0 & 1 & & & \\ r_{31} & 0 & \ddots & & \\ \vdots & & & 1 & \\ 0 & 0 & & & 1 \end{pmatrix},$$

и так далее до

$$\begin{pmatrix} 1 & & & & \\ 0 & 1 & & & \\ \vdots & & \ddots & & \\ 0 & 0 & & 1 & \\ r_{n1} & 0 & & & 1 \end{pmatrix}.$$

Нетрудно убедиться, что умножение матриц выписанного выше ви-

¹⁶Матрицы такого вида называются *трансвекциями* [68].

да выполняется по простому правилу:

$$\begin{pmatrix} 1 & & & 0 \\ & 1 & & \\ r_{i1} & & \ddots & \\ & & & 1 \\ 0 & & \ddots & \\ & & & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & & & 0 \\ & 1 & & \\ & & \ddots & \\ r_{k1} & 0 & & \ddots \\ & & & 1 \end{pmatrix} = \begin{pmatrix} 1 & & & 0 \\ & 1 & & \\ r_{i1} & & \ddots & \\ & & & 1 \\ r_{k1} & 0 & \ddots & \\ & & & 1 \end{pmatrix}.$$

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

$$E_1 = \begin{pmatrix} 1 & & & 0 \\ r_{21} & 1 & & \\ r_{31} & 0 & 1 & \\ \vdots & & & \ddots \\ r_{n1} & 0 & & 1 \end{pmatrix}. \quad (3.74)$$

Аналогично, обнуление всех поддиагональных элементов j -го столбца матрицы СЛАУ и соответствующие преобразования правой части

можно интерпретировать как умножение системы слева на матрицу

$$E_j = \begin{pmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & 1 & & \\ 0 & r_{j+1,j} & 1 & & \\ & \vdots & & \ddots & \\ & r_{nj} & & & 1 \end{pmatrix}. \quad (3.75)$$

В целом метод Гаусса представляется как последовательность умножений обеих частей решаемой СЛАУ слева на матрицы E_j вида (3.75), $j = 1, 2, \dots, n-1$. При этом матрицей системы становится матрица

$$E_{n-1} \cdots E_2 E_1 A = U, \quad (3.76)$$

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 (3.76) 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*.¹⁷ Соответственно, преобразования матрицы A в прямом ходе метода Гаусса (3.71) можно трактовать как её разложение на нижний треугольный L и верхний треугольный U множители.

¹⁷От английских слов lower (нижний) и upper (верхний). Нередко для обозначения этого же понятия можно встретить кальки с иностранных терминов — «LU-факторизация» и «LU-декомпозиция».

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

$$\begin{cases} Ly = b, \\ Ux = y \end{cases} \quad (3.77)$$

с помощью прямой и обратной подстановок соответственно. Как уже отмечалось, LU-разложение, получаемое с помощью версии метода Гаусса (3.71)–(3.72), 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 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.¹⁸ В алгоритме (3.71) 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.

¹⁸Иногда в русской математической литературе его называют *главным* элементом.

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 -ом шаге прямого хода, тогда как первые $j-1$ строк и столбцов матрицы системы остаются уже неизменными.

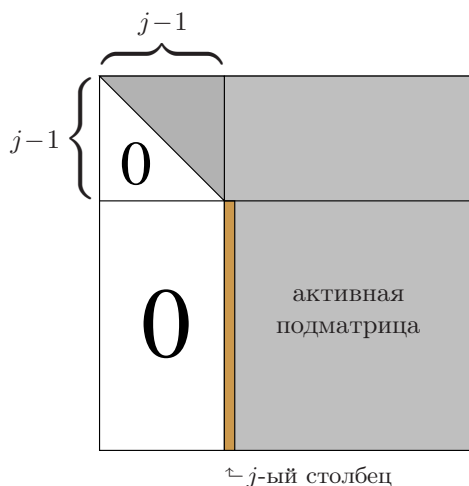


Рис. 3.15. Структура матрицы СЛАУ перед началом j -го шага прямого хода метода Гаусса.

Частичным выбором ведущего элемента на j -ом 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. Именно максимальным по модулю, а не просто ненулевым, ведущий элемент выбирается для того, чтобы обеспечить наибольшую численную устойчивость алгоритма в услови-

ях реальных вычислений с конечной точностью.

Предложение 3.6.1 *Метод Гаусса с частичным выбором ведущего элемента всегда выполним для систем линейных алгебраических уравнений с неособенными квадратными матрицами.*

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. 3.15, 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. ■

Каково матричное представление метода Гаусса с частичным выбором ведущего элемента?

Определение 3.6.1 *Элементарной матрицей перестановки называется матрица вида*

$$P = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 0 & \cdots & 1 \\ & & & 1 & \\ & \vdots & & & \ddots & \vdots \\ & & & & & 1 & \\ 1 & & \cdots & & & & 0 \\ & & & & & & & \ddots \\ & & & & & & & & 1 \end{pmatrix}, \quad \begin{array}{l} \leftarrow i\text{-ая строка} \\ \leftarrow j\text{-ая строка} \end{array} \quad (3.78)$$

которая получается из единичной матрицы перестановкой двух её строк (или столбцов). Матрицей перестановки называется матрица,

получающаяся 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 (3.78) (см. подробности, к примеру, в [7]).

Иногда для матриц (3.78) 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 -ой и j -ой 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

$$(E_{n-1}P_{n-1}) \cdots (E_1P_1)A = U,$$

где E_j — матрицы преобразований вида (3.75), введённые в предыдущем разделе, а P_1, P_2, \dots, P_{n-1} — элементарные матрицы перестановки (3.78), при помощи которых выполняется необходимая перестановка строк на 1-м, 2-м, \dots , $(n-1)$ -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 (3.71) ведущие элементы a_{jj} оказываются малыми, и потому коэффициенты $r_{ij} = -a_{ij}/a_{jj}$ получаются большими по абсолютной величине. Детальный анализ погрешностей вычислений в методе Гаусса (его можно увидеть, к примеру, в [50]) 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 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, изложенного в §3.6в.

Теорема 3.6.2 *Если $A = (a_{ij})$ — квадратная $n \times n$ -матрица, у которой все ведущие миноры порядков от 1 до $(n - 1)$ отличны от нуля,*

т. е.

$$a_{11} \neq 0, \quad \det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \neq 0, \quad \dots, \\ \det \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1,n-1} \\ a_{21} & a_{22} & \dots & a_{2,n-1} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n-1,1} & a_{n-1,2} & \dots & a_{n-1,n-1} \end{pmatrix} \neq 0.$$

то 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

$$A_{n-1} = L_{n-1}U_{n-1}, \quad (3.79)$$

$$z = L_{n-1}y, \quad (3.80)$$

$$v = xU_{n-1}, \quad (3.81)$$

$$a_{nn} = xy + l_{nn}u_{nn}. \quad (3.82)$$

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 x и y —

$$xU_{n-1} = v \quad \text{и} \quad L_{n-1}y = z,$$

которыми являются равенства (3.80)–(3.81), 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 (3.80)–(3.81) векторы x и y , мы сможем из соотношения (3.82) 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.

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)¹⁹, если она неособенна и все её ведущие миноры также отличны от нуля, т. е.

$$a_{11} \neq 0, \quad \det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \neq 0, \quad \dots, \quad \det A \neq 0.$$

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 доказали в Теореме 3.6.2.

Рис. 3.16. Блочное умножение в LU -разложении матрицы.

Для доказательства необходимости привлечём блочное представление треугольного разложения $A = LU$ (см. Рис. 3.16). 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 (3.79) равенства

$$A_k = L_k U_k, \quad k = 1, 2, \dots, n. \quad (3.83)$$

Они означают, что любая ведущая подматрица в A есть произведение ведущих подматриц соответствующих размеров из L и U .

¹⁹Соответствующие английские термины — strictly regular matrix, strictly nonsingular matrix.

Но L и U — неособенные треугольные матрицы, так что все их ведущие подматрицы L_k и U_k также неособенны. Поэтому из равенств (3.83) можно заключить неособенность всех ведущих подматриц A_k в A , т. е. строгую регулярность матрицы A . ■

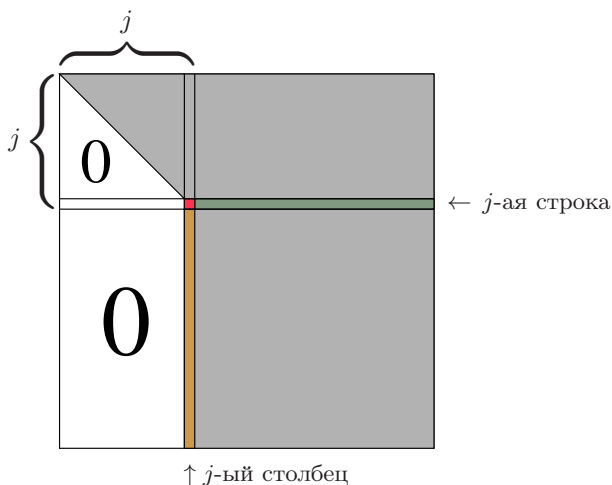


Рис. 3.17. Структура матрицы СЛАУ перед началом j -го шага прямого хода метода Гаусса: другой вид.

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 требуемое Теоремой 3.6.2 условие отличия от нуля ведущих миноров в матрице СЛАУ является достаточным для выполнимости рассмотренного в §3.6в варианта метода Гаусса.

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. В самом деле, к началу j -го шага прямого хода, на котором предстоит обнулить поддиагональные элементы j -го столбца матрицы СЛАУ, её ведущей $j \times j$ -подматрицей является треугольная

матрица, которая получена из исходной ведущей подматрицы преобразованиями предыдущих $j - 1$ шагов метода Гаусса (см. Рис. 3.17). 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 СЛАУ. В частности, при этом всегда $a_{jj} \neq 0$, так что деление на этот элемент в алгоритмах (3.71) и (3.72) 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 положительно определённых или отрицательно определённых матриц в силу известного критерия Сильвестера,
- для матриц с диагональным преобладанием в силу признака Адамара, см. §3.5в (если исходная матрица имеет диагональное преобладание, то его имеют и все ведущие подматрицы).

3.6ж Разложение Холецкого

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 $x^T A x > 0$ для любого $x \neq 0$. Ясно, что положительно-определённые матрицы неособенны.

Теорема 3.6.4 (теорема о разложении Холецкого) *Матрица A является симметричной положительно определённой тогда и только тогда, когда существует неособенная нижняя треугольная матрица C , такая что $A = CC^T$. При этом матрица C из выписанного представления единственна.*

Определение 3.6.3 *Представление $A = CC^T$ называется разложением Холецкого, а нижняя треугольная матрица C — множителем Холецкого 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.²⁰

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 мы можем заключить, что $A = LU$ для некоторых неособенных нижней треугольной матрицы $L = (l_{ij})$ и верхней треугольной матрицы U . Мы дополнительно потребуем, чтобы все диагональные элементы l_{ii} в L были единицами, так что это разложение будет даже однозначно определённым.

Так как

$$LU = A = A^\top = (LU)^\top = U^\top L^\top,$$

то

$$U = L^{-1}U^\top L^\top, \quad (3.84)$$

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 (3.84) вытекает

$$U = L^{-1}U^\top L^\top = DL^\top,$$

и потому

$$A = LU = LDL^\top. \quad (3.85)$$

²⁰Это рассуждение не использует треугольность C и на самом деле обосновывает общее утверждение: произведение неособенной квадратной матрицы на её транспонированную является симметричной положительно определённой матрицей.

Ясно, что 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 (3.85) 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 Холецкого, в действительности, единственно, так как по A при сделанных нами предположениях единственным образом определяется нижняя треугольная матрица L , а матричные преобразования, приведшие к формуле (3.85) и её следствиям, обратимы и также дают однозначно определённый результат. ■

3.63 Метод Холецкого

Основной результат предшествующего раздела мотивирует прямой метод решения систем линейных уравнений, который аналогичен методу (3.77) на основе LU-разложения. Именно, если найдено разложение Холецкого для матрицы A , то решение системы $Ax = b$, равносильной $CC^\top x = b$, сводится к решению двух треугольных систем линейных уравнений:


$$\begin{cases} Cy = b, \\ C^\top x = y. \end{cases} \quad (3.86)$$

Для решения первой системы применяем алгоритм прямой подстановки (3.69), а для решения второй системы — обратную подстановку (3.70).

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 $A = CC^T$, определяющее множитель Холесского, в развёрнутой форме с учётом симметричности A :

$$\begin{pmatrix} a_{11} & \begin{array}{c} \text{---} \end{array} & \begin{array}{c} \text{---} \end{array} \\ a_{21} & a_{22} & \begin{array}{c} \text{---} \end{array} \\ \vdots & \vdots & \ddots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} = \begin{pmatrix} c_{11} & \begin{array}{c} \text{---} \end{array} & \begin{array}{c} \text{---} \end{array} & \begin{array}{c} \text{---} \end{array} \\ c_{21} & c_{22} & \begin{array}{c} \text{---} \end{array} & \begin{array}{c} \text{---} \end{array} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \dots & c_{nn} \end{pmatrix} \cdot \begin{pmatrix} c_{11} & c_{21} & \dots & c_{n1} \\ \begin{array}{c} \text{---} \end{array} & c_{22} & \dots & c_{n2} \\ \begin{array}{c} \text{---} \end{array} & \begin{array}{c} \text{---} \end{array} & \ddots & \vdots \\ \begin{array}{c} \text{---} \end{array} & \begin{array}{c} \text{---} \end{array} & \begin{array}{c} \text{---} \end{array} & c_{nn} \end{pmatrix}, \quad (3.87)$$

где символом «» обозначены симметричные относительно главной диагонали элементы матрицы, которые несущественны в последующих рассуждениях. 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 (3.87) 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 (3.87) имеет вид, определяемый правилом умножения матриц и симметричностью A :

$$\sum_{k=1}^j c_{ik}c_{jk} = a_{ij} \quad \text{при } i \geq j. \quad (3.88)$$

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 (3.88) 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).

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} (см. Рис. 3.18), то все уравнения из (3.88) разбиваются на n следующих групп, которые удобно занумеровать столбцовым индексом $j = 1, 2, \dots, n$:

$$\begin{aligned} \text{для } j = 1 \quad & \begin{cases} c_{11}^2 = a_{11}, \\ c_{i1}c_{11} = a_{i1}, \quad i = 2, 3, \dots, n, \end{cases} \\ \text{для } j = 2 \quad & \begin{cases} c_{31}^2 + c_{22}^2 = a_{22}, \\ c_{i1}c_{21} + c_{i2}c_{22} = a_{i2}, \quad i = 3, 4, \dots, n, \end{cases} \\ \text{для } j = 3 \quad & \begin{cases} c_{31}^2 + c_{32}^2 + c_{33}^2 = a_{33}, \\ c_{i1}c_{31} + c_{i2}c_{32} + c_{i3}c_{33} = a_{i3}, \quad i = 4, 5, \dots, n, \end{cases} \\ \dots & \quad \quad \quad \dots \quad \quad \quad \dots \end{aligned}$$

В краткой записи получающаяся система может быть записана следующим образом:

$$\left\{ \begin{cases} c_{j1}^2 + c_{j2}^2 + \dots + c_{j,j-1}^2 + c_{jj}^2 = a_{jj}, \\ c_{i1}c_{j1} + c_{i2}c_{j2} + \dots + c_{ij}c_{jj} = a_{ij}, \quad i = j+1, \dots, n, \\ j = 1, 2, \dots, n, \end{cases} \right. \quad (3.89)$$

где считается, что $c_{ji} = 0$ при $j < i$.

Получается, что в уравнениях из (3.89) для j -го столбца множителя Холецкого 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 -ой 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.

$$C = \begin{pmatrix} \downarrow & & & & 0 \\ \downarrow & \downarrow & & & \\ \downarrow & \downarrow & \ddots & & \\ \vdots & \vdots & \ddots & \downarrow & \\ \curvearrowright & \curvearrowright & \dots & \curvearrowright & \times \end{pmatrix},$$

Рис. 3.18. Схема определения элементов треугольного множителя при разложении Холецкого.

В целом выписанная система уравнений (3.89) 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. 3.18. Более точно,

$$\text{при } j = 1 \quad \begin{cases} c_{11} = \sqrt{a_{11}}, \\ c_{i1} = a_{i1}/c_{11}, \quad i = 2, 3, \dots, n, \end{cases}$$

$$\text{при } j = 2 \quad \begin{cases} c_{22} = \sqrt{a_{22} - c_{21}^2}, \\ c_{i2} = (a_{i2} - c_{i1}c_{21})/c_{22}, \quad i = 3, 4, \dots, n, \end{cases}$$

$$\text{при } j = 3 \quad \begin{cases} c_{33} = \sqrt{a_{33} - c_{31}^2 - c_{32}^2}, \\ c_{i3} = (a_{i3} - c_{i1}c_{31} - c_{i2}c_{32})/c_{33}, \quad i = 4, 5, \dots, n, \end{cases}$$

и так далее для остальных j . Псевдокод этого процесса приведён в Табл. 3.1, где считается, что если нижний предел суммирования превосходит верхний, то сумма «пуста» и суммирование не выполняется.

Если A — симметричная положительно определённая матрица, то в силу теоремы о разложении Холецкого (Теорема 3.6.4) система уравнений (3.89) обязана иметь решение, и наш алгоритм успешно прорабатывает до конца, находя его. Если же матрица A не является положительно определённой, то алгоритм (3.90) аварийно прекращает работу при попытке извлечь корень из отрицательного числа либо разделить на нуль. Вообще, запуск алгоритма (3.90) — это самый экономичный способ проверки положительной определённости симметричной матрицы.

Таблица 3.1. Алгоритм разложения Холецкого
(прямой ход метода Холецкого)

$$\begin{array}{l}
 \text{DO FOR } j = 1 \text{ TO } n \\
 \quad c_{jj} \leftarrow \sqrt{a_{jj} - \sum_{k=1}^{j-1} c_{jk}^2} \\
 \quad \text{DO FOR } i = j + 1 \text{ TO } n \\
 \quad \quad c_{ij} \leftarrow \left(a_{ij} - \sum_{k=1}^{j-1} c_{ik} c_{jk} \right) / c_{jj} \\
 \quad \text{END DO} \\
 \text{END DO}
 \end{array} \tag{3.90}$$

Способ решения систем линейных алгебраических уравнений с симметричными положительно определёнными матрицами, который основан на нахождении их разложения Холецкого и использует алгоритм (3.90) и далее соотношения (3.86), называют *методом Холецкого*. Он был предложен в 1910 году А.-Л. Холецким 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 Холецкого можно рассматривать как специальную модификацию метода Гаусса, которая требует вдвое меньше времени и памяти ЭВМ, чем обычный метод Гаусса в общем случае. Замечательное свойство метода Холецкого состоит в том, что обусловленность множителей Холецкого, вообще говоря, является лучшей, чем у матрицы исходной СЛАУ: она равна корню квадратному из обусловленности матрицы системы (это следует из самого разложения Холецкого). То есть, в отличие от обычного метода Гаусса, треугольные системы

линейных уравнений из (3.86), к решению которых сводится задача, менее чувствительны к погрешностям, чем исходная линейная система. В следующем пункте мы увидим, что подобную ситуацию следует рассматривать как весьма нетипичную.

Если при реализации метода Холецкого использовать комплексную арифметику, то извлечение квадратного корня можно выполнять всегда, и потому такая модификация применима к симметричным неособенным матрицам, которые не являются положительно определёнными. При этом множители Холецкого становятся комплексными треугольными матрицами.

Другой способ распространения метода Холецкого на системы с произвольными симметричными матрицами состоит в том, чтобы ограничиться разложением (3.85), которое называется LDL^T -разложением матрицы. Если исходная матрица не является положительно определённой, то диагональные элементы в матричном множителе D могут быть отрицательными. Но LDL^T -разложение столь же удобно для решения систем линейных алгебраических уравнений, как и рассмотренные ранее треугольные разложения. Детали этих построений читатель может найти, к примеру, в [11, 15, 46, 77].

Отметим также, что существует возможность другой организации вычислений при решении системы уравнений (3.88), когда неизвестные элементы $c_{11}, c_{21}, c_{22}, \dots, c_{nn}$ последовательно находятся по строкам множителя Холецкого, а не по столбцам, как в (3.90). Этот алгоритм называется *схемой окаймления* [15], и он по своим свойствам примерно эквивалентен рассмотренному выше алгоритму (3.90).

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

$$\text{cond}(AB) = \|(AB)^{-1}\| \|AB\| \leq \text{cond } A \cdot \text{cond } B. \quad (3.91)$$

С другой стороны, если $C = AB$, то $A = CB^{-1}$, и в силу доказанного неравенства

$$\text{cond}(A) \leq \text{cond}(C) \cdot \text{cond}(B^{-1}) = \text{cond}(AB) \cdot \text{cond}(B),$$

когда 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

$$\text{cond}(AB) \geq \max \left\{ \frac{\text{cond}(A)}{\text{cond}(B)}, \frac{\text{cond}(B)}{\text{cond}(A)} \right\}. \quad (3.92)$$

Ясно, что её правая часть не меньше 1.

Неравенства (3.91)–(3.92) 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 (см. стр. 339)

$$\text{cond}_2(A) = \frac{\max_i |\lambda_i(A)|}{\min_i |\lambda_i(A)|}.$$

У матрицы A^2 собственные векторы, очевидно, совпадают с собственными векторами матрицы A , а собственные значения равны $\lambda_1^2, \lambda_2^2, \dots$. Как следствие, числом обусловленности матрицы A^2 становится

$$\text{cond}_2(A^2) = \frac{\max_i (\lambda_i(A))^2}{\min_i (\lambda_i(A))^2} = \frac{\max_i |\lambda_i(A)|^2}{\min_i |\lambda_i(A)|^2} = \left(\frac{\max_i |\lambda_i(A)|}{\min_i |\lambda_i(A)|} \right)^2,$$

и в верхней оценке (3.91) получаем равенство. Совершенно сходным образом можно показать, что для спектрального числа обусловленности оценка (3.91) достигается также на произведениях вида $A^T A$.

Нижняя оценка (3.92) достигается, к примеру, при $B = A^{-1}$ для чисел обусловленности, порождённых подчинёнными матричными нормами.

Практически наиболее важной является верхняя оценка (3.91), 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. Обнуление поддиагональных элементов первого столбца матрицы A — это умножение исходной СЛАУ слева на матрицу E_1 , имеющую вид (3.74), так что мы получаем систему

$$(E_1 A) x = E_1 b \quad (3.93)$$

с матрицей $E_1 A$, число обусловленности которой оценивается как

$$\text{cond}(E_1 A) \leq \text{cond}(E_1) \text{cond}(A).$$

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 (3.93), умножая её слева на матрицы E_2, E_3, \dots, E_{n-1} вида (3.75). В результате получаем верхнюю треугольную систему линейных уравнений

$$Ux = y,$$

в которой $U = E_{n-1} \dots E_2 E_1 A$, $y = E_{n-1} \dots E_2 E_1 b$, и число обусловленности матрицы U оценивается сверху как

$$\text{cond}(U) \leq \text{cond}(A) \cdot \text{cond}(E_1) \cdot \text{cond}(E_2) \cdot \dots \cdot \text{cond}(E_{n-1}). \quad (3.94)$$

Если E_j отлична от единичной матрицы, то $\text{cond}(E_j) > 1$, причём несмотря на специальный вид матриц E_j правая и левая части неравенства (3.94) могут отличаться не очень сильно (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 5×5 -системе линейных алгебраических уравнений первый столбец матрицы коэффициентов имеет вид $(1, 2, 3, 4, 5)^T$. Тогда обнуление поддиагональных элементов равносильно умножению слева на матрицу

$$E_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 & 0 \\ -3 & 0 & 1 & 0 & 0 \\ -4 & 0 & 0 & 1 & 0 \\ -5 & 0 & 0 & 0 & 1 \end{pmatrix},$$

и $\text{cond}_2(E_1) = 55.98$. ■

Пример 3.7.2 Для 2×2 -матрицы (3.14)

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

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 (3.15)

$$B = \begin{pmatrix} 1 & 2 \\ -3 & 4 \end{pmatrix},$$

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, где есть встроенная функция `cond` для расчёта числа обусловленности матрицы. ■

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. Соответствующую технологию, при которой двумя ортогональными преобразованиями матрица СЛАУ приводится к диагональному виду, мы обсуждали в §3.46. Но нахождение сингулярного разложения матрицы — задача более сложная и трудоёмкая, чем рассматриваемые нами прямые методы решения СЛАУ.

3.76 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 (евклидовы 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 ошибки и погрешности, вносимые в процесс вычисления 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

$$\|QA\tilde{x} - Qb\|_2 = \|Q(A\tilde{x} - b)\|_2 = \|A\tilde{x} - b\|_2.$$

Как следствие, 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. Имеено ортогональными преобразованиями можно приводить переопределённые системы линейных алгебраических уравнений к правой трапецевидной форме для вычисления псевдорешений, которые легко находятся обратной подстановкой.

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 *Отображение $\mathcal{A} : X \rightarrow X$ из метрического пространства X с расстоянием dist в себя, обладающее свойством*

$$\text{dist}(\mathcal{A}x, \mathcal{A}y) = \text{dist}(x, y),$$

называется изометрическим относительно расстояния dist или просто изометрией.

Если X линейное нормированное пространство, на котором расстояние задаётся нормой как (3.18), то линейное изометрическое отображение \mathcal{A} можно охарактеризовать проще, условием

$$\|\mathcal{A}x\| = \|x\| \quad \text{для любого } x \in X.$$

Линейные преобразования линейных пространств представляются матрицами, и потому в нашем исследовании особенно важна

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 перестановок.*

Доказательство опускается.

Но одни только матрицы перестановок не могут выполнять полноценные преобразования, приводящие к матрицам необходимой специальной структуры в задачах вычислительной линейной алгебры. Как следствие, прямые численные методы для нахождения псевдорешений систем линейных алгебраических уравнений относительно p -норм невозможны при $p \neq 2$. Эти численные методы обязаны быть итерационными (см., например, [92]), что, впрочем, не является каким-то существенным их недостатком. Они интенсивно разрабатываются в вычислительной оптимизации.

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 трапецивидную $m \times n$ -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 квадратной матрицы.*

Существует несколько способов доказательства этого результата, и почти все они имеют конструктивный характер, давая начало различным технологиям разложения матрицы на произведение ортогонального и треугольного сомножителей. Сначала мы приведём наиболее короткое доказательство, имеющее теоретический характер, а остальные будут изложены в соответствующих местах курса.

Доказательство. Если A — неособенная матрица, то, как было показано при доказательстве Теоремы 3.6.4, $A^\top A$ — симметричная положительно определённая матрица. Следовательно, существует её разложение Холесского

$$A^\top A = R^\top R,$$

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 куда-нибудь 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 $X \mapsto X^T X$) 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

$$\lim_{l \rightarrow \infty} (Q_{k_l}^T A_{k_l}) = \lim_{l \rightarrow \infty} Q_{k_l}^T \cdot \lim_{l \rightarrow \infty} A_{k_l} = Q^T A = R$$

— правой треугольной матрице, поскольку все $Q_{k_l}^T A_{k_l}$ 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 известно QR-разложение матрицы A , то решение исходной СЛАУ, равносильной

$$(QR)x = b$$

сводится к решению треугольной системы линейных алгебраических уравнений

$$Rx = Q^T b. \quad (3.95)$$

Ниже в §3.15 и §3.17г 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 Холецкого 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 отражения и матрицы вращения, описанию которых посвящены следующие разделы книги.

3.7г 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 $H(u)$.

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 (см. Рис. 3.19). 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.

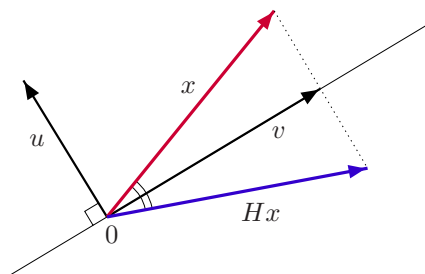


Рис. 3.19. Геометрическая интерпретация действия матрицы отражения.

Предложение 3.7.2 Пусть задан вектор $e \in \mathbb{R}^n$ единичной длины, $\|e\|_2 = 1$. Для любого ненулевого вектора $x \in \mathbb{R}^n$ существует матрица отражения, переводящая его в вектор, коллинеарный вектору e .

Доказательство. Если H — искомая матрица отражения, и u — порождающий её вектор Хаусхолдера, то утверждение предложения требует равенства

$$Hx = x - 2(uu^\top)x = \gamma e \quad (3.96)$$

с некоторым коэффициентом $\gamma \neq 0$. Тогда

$$2u(u^\top x) = x - \gamma e. \quad (3.97)$$

Отдельно рассмотрим два случая — когда векторы x и e неколлинеарны, и когда они коллинеарны друг другу.

В первом случае правая часть в (3.97) заведомо не равна нулю. Следовательно, числовой множитель $u^\top x$ в левой части этого равенства обязан быть ненулевым, и можно заключить, что

$$u = \frac{1}{2u^\top x} (x - \gamma e),$$

т. е. что вектор u , порождающий искомую матрицу отражения, должен быть коллинеарен вектору $(x - \gamma e)$.

Для определения коэффициента γ заметим, что ортогональная матрица H не изменяет длин векторов, так что $\|Hx\|_2 = \|x\|_2$. С другой стороны, взяв евклидову норму от обеих частей (3.96), получим $\|Hx\|_2 = |\gamma| \|e\|_2$. Сопоставляя оба равенства, получаем

$$\|x\|_2 = |\gamma| \|e\|_2, \quad \text{т. е. } \gamma = \pm \|x\|_2.$$

Следовательно, вектор Хаусхолдера u должен быть коллинеарен вектору

$$\tilde{u} = x \pm \|x\|_2 e, \quad (3.98)$$

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 теперь случай, когда x коллинеарен e . При этом предшествующая конструкция частично теряет смысл, так как вектор $\tilde{u} = x - \gamma e$ может занулиться при подходящем выборе множителя γ .

Но даже если $x - \gamma e = 0$ для какого-то одного из значений $\gamma = -\|x\|_2$ и $\gamma = \|x\|_2$, то для противоположного по знаку значения γ наверняка $x - \gamma e \neq 0$. Более формально можно сказать, что конкретный знак у множителя $\gamma = \pm \|x\|_2$ следует выбирать из условия максимизации нормы вектора $(x - \gamma e)$. Далее все рассуждения, следующие за формулой (3.97), остаются в силе и приводят к определению вектора Хаусхолдера.

Наконец, в случае коллинеарных векторов x и e мы можем просто указать явную формулу для вектора Хаусхолдера:

$$u = \frac{x}{\|x\|_2}.$$

При этом

$$u^\top x = \frac{x^\top x}{\|x\|_2} = \|x\|_2 \neq 0,$$

и для соответствующей матрицы отражения имеет место

$$Hx = x - 2(uu^\top)x = x - 2u(u^\top x) = x - 2\frac{x}{\|x\|_2}\|x\|_2 = -x.$$

Итак, вектор x снова переводится матрицей H в вектор, коллинеарный вектору e , т. е. условие Предложения удовлетворено и в этом случае.²¹



В доказательстве предложения присутствует неоднозначность в выборе знака в выражении $\tilde{u} = x \pm \|x\|_2 e$, если x и e неколлинеарны. В действительности, годится любой знак, и его конкретный выбор может определяться, как мы увидим, требованием устойчивости вычислительного алгоритма.

3.7д Метод 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 трапецевидной форме, а затем воспользоваться обратной подстановкой (3.72). Но теперь это приведение выполняется более глубокими, чем в методе Гаусса, преобразованиями матрицы, именно, путём последовательного умножения на специальным образом 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 u , возможно, единичных матриц, таких что матрица*

$$H_s H_{s-1} \cdots H_2 H_1 A = R$$

²¹Интересно, что этот тонкий случай доказательства имеет, скорее, теоретическое значение, так как на практике если вектор уже коллинеарен заданному, то с ним, как правило, можно вообще ничего не делать.

является правой треугольной или трапецевидной 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, меняется в зависимости от соотношения числа строк m и числа столбцов n в матрице A .

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.

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 $(1, 0, \dots, 0)^T$. 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

$$\tilde{A} = \left(\begin{array}{c|cccc} \times & \times & \times & \cdots & \times \\ \hline 0 & \times & \times & \cdots & \times \\ 0 & \times & \times & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \times & \times & \cdots & \times \end{array} \right).$$

где крестиками « \times » обозначены элементы, которые, возможно, не равны нулю. Прделаем теперь 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 теперь 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 $H_j(u)$ аннулировала 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.

Таблица 3.2. 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  $\check{y} \in \mathbb{R}^{n-j+1}$ ,
        порождающий отражение, которое переводит
        вектор  $A(j : n, j)$  в вектор, коллинеарный
        вектору  $(1, 0, \dots, 0)^T$ ;
         $\check{H} \leftarrow \check{I} - 2\check{y}\check{y}^T$ ;
    ELSE
         $\check{H} \leftarrow \check{I}$ ;
    END IF
     $A(j : n, j : n) \leftarrow \check{H} A(j : n, j : n)$ ;
END DO

```

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} , а \check{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 \check{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 \check{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

$$Q = (H_s H_{s-1} \cdots H_2 H_1)^{-1}.$$

Таким образом, мы получаем QR-разложение матрицы A , т.е. Предложения 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. 3.2.

Как следствие, исходная система уравнений $Ax = b$ становится равносильной системе уравнений

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

с несложно решаемыми составными частями. При практической реализации удобнее дополнить алгоритм Табл. 3.2 инструкциями, которые задают преобразования вектора правой части СЛАУ, и тогда результатом работы нового алгоритма будет правая треугольная система $Rx = y$. Её можно решать с помощью обратной подстановки (3.70).

Пример 3.7.3 Решим с помощью метода Хаусхолдера задачу построения линейной функции наилучшего среднеквадратичного приближения к данным, которая была рассмотрена в Примере 2.10.4. 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 трапецевидной форме с помощью отражений Хаусхолдера.

Для обнуления поддиагональных элементов первого столбца умножим слева обе части этой системы на матрицу

В целом получаем

$$Q = \begin{pmatrix} -0.26726 & 0.87287 & 0.40825 \\ -0.53452 & 0.21822 & -0.81650 \\ -0.80178 & -0.43644 & 0.40825 \end{pmatrix},$$

$$R = \begin{pmatrix} -3.74166 & -1.60357 \\ 0 & 0.65465 \\ 0 & 0 \end{pmatrix}.$$

$$Q^T b = \begin{pmatrix} -2.40535 \\ 0.21822 \\ 0.40825 \end{pmatrix}$$

Выполнив процесс обратной подстановки для треугольной 2×2 -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 $Rx = Q^T b$, получим

$$\alpha = 0.5, \quad \beta = 0.33333.$$

Это решение совпадает с тем, которое мы нашли в Примере 2.10.4 с помощью перехода к нормальной системе уравнений. ■

Согласно Предложению 3.7.2 вычисление вектора Хаусхолдера u в качестве первого шага требует нахождения из (3.98) 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

$$\tilde{u} = \begin{cases} A(j : n, j) + \|A(j : n, j)\|_2 e, & \text{если } a_{jj} \geq 0, \\ A(j : n, j) - \|A(j : n, j)\|_2 e, & \text{если } a_{jj} \leq 0, \end{cases}$$

где $e = (1, 0, \dots, 0)^\top$. Тогда вычисление первого элемента 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 в явном виде матрицу отражения \tilde{H} : умножение на неё можно выполнить по экономичной формуле

$$\begin{aligned} (I - 2uu^\top) A(j : n, j : n) \\ = A(j : n, j : n) - 2u (u^\top A(j : n, j : n)), \end{aligned}$$

в которой сама \tilde{H} не фигурирует.

При решении СЛАУ можно одновременно с разложением матрицы преобразовывать также правую часть, умножая её на матрицы \tilde{H} . Тогда к моменту окончания QR-разложения у нас уже будет известен вектор правой части $Q^\top b$ для системы с треугольной матрицей R .

Определённым недостатком метода Хаусхолдера и описываемого в следующем пункте 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 [95], которые могут оперировать рациональными дробями.

3.7e 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 θ , $0 \leq \theta < 2\pi$.

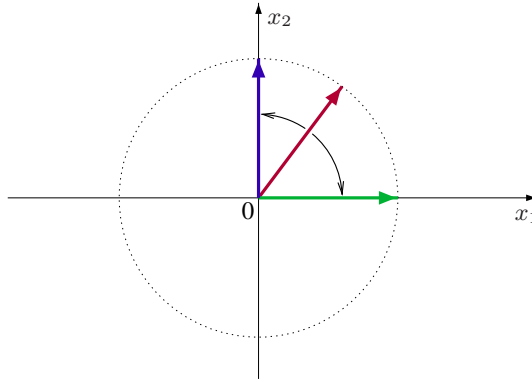


Рис. 3.20. Подходящим вращением можно занулить любую из компонент двумерного вектора.

Аналогично может быть занулена первая компонента вектора a , путём домножения на такую матрицу вращения (3.100), что

$$\cos \theta = \frac{a_2}{\|a\|_2}, \quad \sin \theta = \frac{a_1}{\|a\|_2}.$$

В общем случае умножение любой матрицы $A = (a_{ij})$ слева на матрицу вращения $G(k, l, \theta)$ приводит к тому, что в их произведении $\tilde{A} = (\tilde{a}_{ij}) := G(k, l, \theta) A$ строки k -ая и l -ая становятся линейными комбинациями строк с этими же номерами из A :

$$\begin{aligned} \tilde{a}_{kj} &\leftarrow a_{kj} \cos \theta - a_{lj} \sin \theta, \\ \tilde{a}_{lj} &\leftarrow a_{kj} \sin \theta + a_{lj} \cos \theta, \end{aligned} \quad (3.101)$$

$j = 1, 2, \dots, n$. 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 θ можно занулить любой элемент k -ой или l -ой строк матрицы $\tilde{A} = G(k, l, \theta) A$.

Как следствие, 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 последующих столбцов никак не испортит полученные ранее нулевые элементы предшествующих столбцов, так как линейное комбинирование нулей согласно формулам (3.101) даст снова нуль. На формальном языке можно сказать, что существует набор матриц вращения $G(1, 2), G(1, 3), \dots, G(1, n), G(2, 3), \dots, G(n-1, n)$, таких что

$$G(n-1, n) \cdots G(2, 3) G(1, n) \cdots G(1, 3) G(1, 2) A = R$$

— правая треугольная матрица. Отсюда

$$A = G(1, 2)^T G(1, 3)^T \cdots G(1, n)^T G(2, 3)^T \cdots G(n-1, n)^T R,$$

и мы получили QR-разложение матрицы A , поскольку произведение транспонированных матриц вращения также является ортогональной матрицей.

Использование преобразований вращения — ещё один конструктивный способ получения QR-разложения, технически даже более простой, чем метод отражений 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 лишь ненулевых элементов всего двух строк по формулам (3.101).

Для плотно заполненных матриц использование вращений в полтора раза более трудоёмко, чем получение QR-разложения с помощью матриц отражения, но зато вращения более предпочтительны для разреженных матриц в силу своей большей гибкости при занулении отдельных элементов.

Использование ортогональных матриц вращения можно положить в основу *метода вращений* для численного нахождения решения или псевдорешения систем линейных алгебраических уравнений с матрицами полного ранга. Если дана система линейных уравнений $Ax = b$, то с помощью матриц вращения выполним приведение матрицы A к правой (верхней) треугольной или трапецевидной форме, одновременно применяя преобразования к правой части b . Получим эквивалентную систему линейных уравнений с треугольной или трапецевидной матрицей, которая легко решается обратной подстановкой. Этот численный метод очень похож на метод Гаусса, но отличается от него лучшей устойчивостью и возможностью получать псевдорешения переопределённых систем уравнений относительно евклидовой нормы.

3.7ж Процессы ортогонализации

Ортогонализацией называют процесс построения по заданному базису линейного пространства некоторого ортогонального базиса, который имеет ту же самую линейную оболочку. Ввиду удобства ортогональных базисов для представления решений разнообразных задач и, как следствие, их важности во многих приложениях (см., к примеру, §2.10ж) огромное значение имеют и процессы ортогонализации.

Исторически первым процессом ортогонализации был алгоритм, который по традиции связывают с именами Й. Грама и Э. Шмидта.²³ По конечной линейно независимой системе векторов $\{v_1, v_2, \dots, v_n\}$ процесс Грама-Шмидта строит ортогональный базис $\{q_1, q_2, \dots, q_n\}$ линейной оболочки векторов $\{v_1, v_2, \dots, v_n\}$.

Возьмём в качестве первого вектора q_1 конструируемого ортогонального базиса вектор v_1 , первый из исходного базиса. Далее для построения q_2 можно использовать v_2 «как основу», но откорректировав его с учётом требования ортогональности к q_1 и принадлежности линейной оболочке векторов $q_1 = v_1$ и v_2 . Естественно положить $q_2 = v_2 - \alpha_{12}q_1$, где коэффициент α_{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 $q_3 = v_3 - \alpha_{13}q_1 - \alpha_{23}q_2$, и т. д.

В целом ортогонализация Грама-Шмидта выполняется в соответствии со следующими расчётными формулами:

$$q_1 \leftarrow v_1, \tag{3.102}$$

$$q_j \leftarrow v_j - \sum_{k=1}^{j-1} \frac{\langle q_k, v_j \rangle}{\langle q_k, q_k \rangle} q_k, \quad j = 2, 3, \dots, n. \tag{3.103}$$

В случае, когда очередной вычисленный вектор q_j — нулевой, ясно, что v_1, v_2, \dots, v_j линейно зависимы. Тогда построение для линейной оболочки $\text{lin} \{v_1, v_2, \dots, v_n\}$ ортогонального базиса, состоящего из того же количества векторов, невозможно.

²³Иногда этот процесс называют «ортогонализацией Сонина-Шмидта».

Получающиеся векторы ортогонального базиса, как правило, дополнительно нормируют сразу после их нахождения, добиваясь равенства $\langle q_k, q_k \rangle = 1$. Тогда в (3.103) упрощаются выражения для поправочных коэффициентов при q_k . Псевдокод соответствующего варианта ортогонализации Грама-Шмидта дан в Табл. 3.3.

Таблица 3.3. Ортогонализация Грама-Шмидта
системы векторов $\{v_1, v_2, \dots, v_n\}$

```

DO FOR  j = 1 TO n
    qj ← vj ;
    DO k = 1 TO j - 1
        αkj ← ⟨qk, vj⟩ ;
        qj ← qj - αkjqk ;
    END DO
    αjj ← ||qj||2 ;
    IF ( αjj = 0 ) THEN
        STOP, сигнализируя «vj линейно зависит
        от векторов v1, v2, ..., vj-1»
    END IF
    qj ← qj/αjj ;
END DO

```

Каково матричное представление процесса ортогонализации 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 мы организуем матрицу W . В результате ортогонализации мы должны получить ортогональную матрицу, в которой первый столбец — это нормированный первый вектор, второй столбец — это нормированная линейная комбинация первых двух вектор-столбцов, и т. д. Столбец с номером j результирующей ортогональной матрицы равен нормированной линейной комбинации первых j штук столбцов исходной матрицы. В целом процесс ортогонализации Грама-Шмидта равносильен умножению W справа на верхнюю треугольную матрицу, в результате чего должна получиться ортогональная матри-

ца.

Фактически, ортогонализацию Грама-Шмидта можно рассматривать как ещё один способ получения QR-разложения матрицы, наряду с методом отражений (§3.7д) или методом вращений (§3.7е).²⁴ Но устойчивость ортогонализации Грама-Шмидта существенно хуже. Если исходная система векторов близка к линейно зависимой, 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.

Причина этого явления довольно прозрачна. При построении QR-разложения с помощью матриц отражения или вращения ортогональность соответствующего матричного сомножителя специально контролируется в процессе работы алгоритма, тогда как в ортогонализации Грама-Шмидта мы идём обратным путём, от треугольной матрицы, и ортогональность получается как конечный продукт нетривиального вычислительного алгоритма. Неудивительно, что эта ортогональность и не имеет места для результата выполнения реального алгоритма, подверженного влиянию погрешностей.

Недостаточную устойчивость ортогонализации Грама-Шмидта можно до некоторой степени исправить, модифицировав её расчётные формулы так, чтобы вычисление поправочных коэффициентов α_{kj} выполнялось другим способом. Псевдокод модифицированной ортогонализации Грама-Шмидта приведён в Табл. 3.4.

В общем случае при ортогонализации Грама-Шмидта построение каждого следующего вектора требует привлечения всех ранее построенных векторов. Но если исходная система векторов имеет специальный вид, в определённом смысле согласованный с используемым скалярным произведением, то ситуация упрощается. Важнейший частный случай — ортогонализация так называемых *подпространств Крылова*.

Определение 3.7.4 Пусть даны квадратная $n \times n$ -матрица A и n -вектор r . Подпространствами Крылова $K_i(A, r)$, $i = 1, 2, \dots, n$, матрицы A относительно вектора r называются линейные оболочки векторов $r, Ar, \dots, A^{i-1}r$, т. е. $K_i(A, r) = \text{lin} \{r, Ar, \dots, A^{i-1}r\}$.

²⁴Верно и обратное: любое разложение матрицы на множители, в котором один из сомножителей ортогонален, соответствует некоторому процессу ортогонализации. Вопрос в том, насколько удобны и технологичны соответствующие алгоритмы.

Таблица 3.4. 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

```

Оказывается, что если A — симметричная положительно определённая матрица, то при ортогонализации подпространств Крылова построение каждого последующего вектора привлекает лишь два предшествующих вектора из строящегося базиса. Более точно, справедлива

Теорема 3.7.3 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 α_k и β_k вычисляются следующим

образом:

$$\alpha_k = \frac{\langle Ap_k, p_k \rangle}{\langle p_k, p_k \rangle}, \quad k = 0, 1, \dots, n-2,$$

$$\beta_k = \frac{\langle Ap_k, p_{k-1} \rangle}{\langle p_{k-1}, p_{k-1} \rangle} = \frac{\langle p_k, p_k \rangle}{\langle p_{k-1}, p_{k-1} \rangle}, \quad k = 1, 2, \dots, n-2.$$

Этот факт был открыт К. Ланцошем в 1952 году и имеет многочисленные применения в практике вычислений, так как более короткие вычислительные формулы меньше подвержены ошибкам вычислений и более устойчивы. В частности, результат Теоремы 3.7.3 является одной из теоретических основ метода сопряжённых градиентов для решения СЛАУ (см. §3.10д).

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 (3.102)–(3.103) следует, что для любого $k = 1, 2, \dots, n-1$

$$p_{k+1} = A^{k+1}r - \sum_{i=0}^k c_i^{(k)} A^i r, \quad c_i^{(k)} \in \mathbb{R}.$$

Как следствие, 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

$$p_{k+1} = Ap_k - \gamma_0^{(k)} p_0 - \dots - \gamma_k^{(k)} p_k$$

с какими-то коэффициентами $\gamma_0^{(k)}, \dots, \gamma_k^{(k)}$.

Домножая скалярно полученное соотношение 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 $\langle Ap_k, p_j \rangle = \langle p_k, Ap_j \rangle$, 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

$$\alpha_k = \gamma_k^{(k)} = \frac{\langle Ap_k, p_k \rangle}{\langle p_k, p_k \rangle},$$

$$\beta_k = \gamma_{k-1}^{(k)} = \frac{\langle Ap_k, p_{k-1} \rangle}{\langle p_{k-1}, p_{k-1} \rangle}.$$

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 3.7.3, называют *ортогонализацией Ланцоша*. Фактически, это удобный способ построения ортогонального базиса всего пространства, если подпространства Крылова линейно независимы.

3.8 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 большого числа систем линейных уравнений, встречающихся в практике математического моделирования, ненулевые элементы заполняют матрицу не полностью, образуя в ней те или иные 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.²⁵ Далее для краткости мы

²⁵В англоязычной литературе этот метод называют также «tridiagonal matrix algorithm» или «Thomas algorithm» (алгоритм Томаса).

нередко будем называть их просто «трёхдиагональными линейными системами». Это важный в приложениях случай СЛАУ, возникающий, к примеру, при решении многих краевых задач для дифференциальных уравнений. По определению трёхдиагональными называются матрицы, все ненулевые элементы которых сосредоточены на трёх диагоналях — главной и соседних с ней сверху и снизу. Иными словами, для трёхдиагональной матрицы $A = (a_{ij})$ неравенство $a_{ij} \neq 0$ может иметь место лишь при $i = j$ и $i = j \pm 1$.

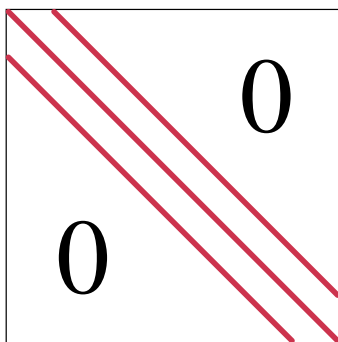


Рис. 3.21. Портрет трёхдиагональной матрицы.

Система n линейных алгебраических уравнений относительно неизвестных x_1, x_2, \dots, x_n , имеющая трёхдиагональную матрицу, в развёрнутом виде выглядит следующим образом:

$$\begin{cases} b_1 x_1 + c_1 x_2 = d_1, \\ a_i x_{i-1} + b_i x_i + c_i x_{i+1} = d_i, & 2 \leq i \leq n-1, \\ a_n x_{n-1} + b_n x_n = d_n. \end{cases} \quad (3.104)$$

Часто её записывают в следующем специальном каноническом виде, даже без обращения к матрично-векторной форме:

$$a_i x_{i-1} + b_i x_i + c_i x_{i+1} = d_i, \quad 1 \leq i \leq n, \quad (3.105)$$

где для единообразия 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 3.8.1 В §2.8 мы могли видеть, что на равномерной сетке с шагом h

$$u''(x_i) \approx \frac{u(x_{i-1}) - 2u(x_i) + u(x_{i+1}))}{h^2},$$

и 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 приводит к линейным системам уравнений с трёхдиагональными матрицами, у которых помимо главной диагонали заполнены только две соседние с ней. ■

Соотношения вида (3.105)

$$a_i x_{i-1} + b_i x_i + c_i x_{i+1} = d_i, \quad i = 1, 2, \dots,$$

называют также *трёхточечными разностными уравнениями* или *разностными уравнениями второго порядка*.

Пусть для СЛАУ с трёхдиагональной матрицей выполняется прямой ход метода Гаусса без перестановок 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

$$\begin{pmatrix} \times & \times & & 0 \\ & \times & \ddots & \\ & & \ddots & \times \\ 0 & & & \times & \times \\ & & & & \times \end{pmatrix}, \quad (3.106)$$

в которой ненулевые элементы (обозначенные крестиками) присутствуют лишь на главной диагонали и первой наддиагонали. Следовательно, формулы обратного хода метода Гаусса вместо (3.72) должны иметь следующий двучленный вид

$$x_i = \xi_{i+1} x_{i+1} + \eta_{i+1}, \quad i = n, n-1, \dots, 1, \quad (3.107)$$

где, как и в исходных уравнениях, в n -ом соотношении присутствует вспомогательная фиктивная неизвестная x_{n+1} . Оказывается, что величины ξ_i и η_i в соотношениях (3.107) можно несложным образом выразить через элементы исходной системы уравнений.

Уменьшим в (3.107) все индексы на единицу —

$$x_{i-1} = \xi_i x_i + \eta_i$$

— и подставим полученное соотношение в i -ое уравнение системы, что даёт

$$a_i(\xi_i x_i + \eta_i) + b_i x_i + c_i x_{i+1} = d_i.$$

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 эту формулу с двучленными расчётными формулами (3.107), можем заключить, что для $i = 1, 2, \dots, n$

$$\xi_{i+1} = -\frac{c_i}{a_i \xi_i + b_i},$$

$$\eta_{i+1} = \frac{d_i - a_i \eta_i}{a_i \xi_i + b_i}.$$

(3.108)

Это формулы *прямого хода* прогонки, целью которого является вычисление величин ξ_i и η_i , называемых *прогоночными коэффициентами*. Далее по формулам (3.107) 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 ξ_1 и η_1 в прямом ходе и x_{n+1} — в обратном. Формально они неизвестны, но фактически полностью определяются условием $a_1 =$

$c_n = 0$. Действительно, конкретные значения ξ_1 и η_1 не влияют на результаты решения, потому что в формулах (3.108) прямого хода прогонки они встречаются с множителем $a_1 = 0$. Кроме того, из формул прямого хода следует, что

$$\xi_{n+1} = -\frac{c_n}{a_n\xi_n + b_n} = -\frac{0}{a_n\xi_n + b_n} = 0,$$

а это коэффициент при x_{n+1} в обратном ходе прогонки. Поэтому и x_{n+1} может быть произвольным. Итак, для начала прогонки можно положить, к примеру,

$$\xi_1 = \eta_1 = x_{n+1} = 0. \quad (3.109)$$

Более удобна реализация прогонки, при которой прямой ход начинается с присваивания

$$\xi_2 = -c_1/b_1, \quad \eta_2 = d_1/b_1.$$

Оно вытекает как из первого уравнения системы (3.105), так и из формул (3.108) с $\xi_1 = \eta_1 = 0$. Далее находятся все прогоночные коэффициенты, а затем мы сразу полагаем

$$x_n = \eta_{n+1}.$$

После этого в обратном ходе прогонки находятся неизвестные x_{n-1}, \dots, x_2, x_1 .

Дадим теперь 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 (3.106) преобразованиями прямого хода метода Гаусса без перестановки строк или столбцов, так как эти преобразования являются ничем иным, как прямым ходом метода прогонки.

Здесь полностью применима теория, развитая в §3.6е, в частности, Предложение 3.6.2.

Предложение 3.8.1 *Если в системе линейных алгебраических уравнений с трёхдиагональной матрицей (3.104)–(3.105) 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 (3.109) является реализуемым.

Доказательство. Условие диагонального преобладания в матрице влечёт её строгую регулярность, как мы видели в §3.6е. 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, тем не менее, даётся ещё одно доказательство этого факта, которое позволяет помимо реализуемости установить ещё числовые оценки «запаса устойчивости» прогонки, т. е. того, насколько сильно знаменатели выражений (3.108) 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 $|\xi_i| < 1$.

Прежде всего, $\xi_1 = 0$ и потому база индукции выполнена: $|\xi_1| < 1$. Далее, предположим, что для некоторого индекса i уже установлена оценка $|\xi_i| < 1$. Если соответствующее $c_i = 0$, то из первой формулы (3.108) следует $\xi_{i+1} = 0$, и индукционный переход доказан. Поэтому пусть $c_i \neq 0$. Тогда справедлива следующая цепочка соотношений

$$\begin{aligned} |\xi_{i+1}| &= \left| -\frac{c_i}{a_i \xi_i + b_i} \right| = \frac{|c_i|}{|a_i \xi_i + b_i|} \\ &\leq \frac{|c_i|}{||b_i| - |a_i| \cdot |\xi_i||} \quad \text{из оценки снизу для модуля суммы} \\ &< \frac{|c_i|}{|a_i| + |c_i| - |a_i| \cdot |\xi_i|} \quad \text{в силу диагонального преобладания} \\ &= \frac{|c_i|}{|a_i|(1 - |\xi_i|) + |c_i|} \leq \frac{|c_i|}{|c_i|} = 1, \end{aligned}$$

где при переходе ко второй строке мы воспользовались известным неравенством для модуля суммы двух чисел:

$$|x + y| \geq ||x| - |y||. \quad (3.110)$$

Итак, неравенства $|\xi_i| < 1$ доказаны для всех прогоночных коэффициентов ξ_i , $i = 1, 2, \dots, n + 1$.

Как следствие, для знаменателей прогоночных коэффициентов ξ_i и η_i в формулах (3.108) имеем

$$\begin{aligned} |a_i \xi_i + b_i| &\geq ||b_i| - |a_i \xi_i|| \quad \text{по неравенству (3.110)} \\ &= |b_i| - |a_i| |\xi_i| \quad \text{в силу диагонального преобладания} \\ &> |a_i| + |c_i| - |a_i| \cdot |\xi_i| \quad \text{в силу диагонального преобладания} \\ &= |a_i|(1 - |\xi_i|) + |c_i| \\ &\geq |c_i| \geq 0 \quad \text{в силу оценки } |\xi_i| < 1. \end{aligned}$$

Иными словами, $|a_i \xi_i + b_i|$ строго отделены от нуля, что и требовалось доказать. Кроме того, чем больше является диагональное преобладание в матрице системы, тем «сильнее» выполняется строго неравенство в третьей строке выписанной цепочки, и тем больше отделён от нуля знаменатель прогоночных коэффициентов (3.108). ■

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 (3.53), но зато более жёсткие, чем в Предложении 3.8.1, условия на коэффициенты системы (см. [8, 40]). Весьма популярна, в частности, такая формулировка [93]:

Предложение 3.8.2 Пусть в трёхдиагональной матрице системы линейных алгебраических уравнений (3.104)–(3.105) 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 [31].

3.9 Стационарные итерационные методы для решения линейных систем

3.9а 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 мы рассмотрим этот вопрос в §3.14.

Общая схема итерационных методов выглядит следующим образом: выбираются одно или несколько *начальных приближений* $x^{(0)}$, $x^{(1)}$, ..., $x^{(\nu)}$, а затем по их известным значениям вычисляются последующие приближения

$$x^{(k+1)} \leftarrow T_k(x^{(0)}, x^{(1)}, \dots, x^{(k)}), \quad k = \nu, \nu + 1, \nu + 2, \dots, \quad (3.111)$$

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 процесс (3.111) называют *p-шаговым*, если его последующее приближение $x^{(k+1)}$ является функцией только от p предшествующих приближений, т. е. от $x^{(k)}, x^{(k-1)}, \dots, x^{(k-p+1)}$. В частности, *одношаговые* итерационные методы имеют вид

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

т. е. в них $x^{(k+1)}$ зависит лишь от значения одной предшествующей итерации $x^{(k)}$. Для начала работы одношаговых итерационных процессов нужно знать одно начальное приближение $x^{(0)}$.

Итерационный процесс 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 процессы

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

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 процессы с расчётными формулами того же вида. Более точно, *линейным p-шаговым итерационным процессом* будут называться итерации, в которых оператор перехода имеет вид

$$\begin{aligned} T_k(x^{(k)}, x^{(k-1)}, \dots, x^{(k-p+1)}) \\ = C^{(k,k)}x^{(k)} + C^{(k,k-1)}x^{(k-1)} + \dots + C^{(k,k-p+1)}x^{(k-p+1)} + d^{(k)} \end{aligned}$$

с какими-то коэффициентами $C^{(k,k)}, C^{(k,k-1)}, \dots, C^{(k,k-p+1)}$ 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 процесса.

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 более ясными и практичными. 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, связываемый с именем Герона Александрийского (см. Пример 4.4.3), был известен ещё древним вавилонянам. В Новом времени одним из первопроходцев 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. §4.4г).

Ясно, что 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, составляет важнейшую часть работы по конструированию и исследованию итерационных методов.

Другой источник возникновения итерационных методов — исследование по разрешимости различных уравнений и систем уравнений. Именно в таком качестве итерационные процессы, сходящиеся к решениям интегральных уравнений, появились в середине XIX века в работах Ж. Лиувилля и затем, уже во второй половине 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 (что мы увидим далее на примерах). 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, мы далее уже не возвращаемся к ней, а оперируем с её следствиями, которые никакой обратной связи от исходной системы не получают.²⁶

Как правило, итерационные процессы сравнительно несложно программируются, так как представляют собой повторяющиеся единообразные процедуры, применяемые к последовательным приближениям к решению. При решении СЛАУ с разреженными матрицами в итерационных процессах обычно можно легче, чем в прямых методах, учитывать структуру нулевых и ненулевых элементов матрицы и основывать на этом упрощённые формулы матрично-векторного умножения, которые существенно уменьшают общую трудоёмкость алгоритма.

Иногда системы линейных алгебраических уравнений задаются в операторном виде, рассмотренном нами в начале §3.6 (стр. 353) 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, преобразования матриц таких систем, которые являются основой прямых методов решения систем линейных уравнений, очень сложны или порой просто невозможны.

Наконец, быстро сходящиеся итерационные методы могут обеспечивать выигрыш по времени даже для СЛАУ общего вида, если требуют для практической сходимости 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, по которым решение задачи строится из последовательности, порождаемой методом. Интересный

²⁶Для исправления этого положения прямые методы решения СЛАУ в ответственных ситуациях часто дополняют процедурами итерационного уточнения. См., к примеру, пункт 67 главы 4 в [45].

пример дают методы Монте-Карло, в которых ответ получается как усреднение последовательности, порождаемой численным методом.

3.9б Сходимость стационарных одношаговых итерационных методов

Системы линейных уравнений вида

$$x = Cx + d,$$

в котором вектор неизвестных переменных выделен в одной из частей, мы будем называть *системами в рекуррентном виде*.

Теорема 3.9.1 Пусть система уравнений $x = Cx + d$ имеет единственное решение. Стационарный одношаговый итерационный процесс

$$x^{(k+1)} \leftarrow Cx^{(k)} + d, \quad k = 0, 1, 2, \dots, \quad (3.112)$$

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 $x = x$, имеющее решением любой вектор. Соответствующий итерационный процесс $x^{(k+1)} \leftarrow x^{(k)}$, $k = 0, 1, 2, \dots$, будет сходиться из любого начального приближения, хотя спектральный радиус матрицы перехода C равен единице.

Доказательство Теоремы 3.9.1 будет разбито на две части, результат каждой из которых представляет самостоятельный интерес.

Предложение 3.9.1 Если $\|C\| < 1$ в какой-нибудь матричной норме, то стационарный одношаговый итерационный процесс

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

сходится при любом начальном приближении $x^{(0)}$.

Доказательство. В формулировке предложения ничего не говорится о пределе, к которому сходится последовательность приближений

$\{x^{(k)}\}$, 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 3.3.10, стр. 323) 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 мы 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 $x^{(k)} = Cx^{(k-1)} + d$, $k = 1, 2, \dots$, получим

$$x^{(k)} - x^* = C(x^{(k-1)} - x^*).$$

Вспомним, что всякая матричная норма согласована с некоторой векторной нормой (Предложение 3.3.4), 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

$$\begin{aligned} \|x^{(k)} - x^*\| &\leq \|C\| \cdot \|x^{(k-1)} - x^*\| \\ &\leq \|C\|^2 \cdot \|x^{(k-2)} - x^*\| \\ &\leq \dots \dots \\ &\leq \|C\|^k \cdot \|x^{(0)} - x^*\|. \end{aligned} \quad (3.113)$$

Правая часть неравенства (3.113) сходится к нулю при $k \rightarrow \infty$ в силу условия $\|C\| < 1$, поэтому последовательность приближений $\{x^{(k)}\}_{k=0}^{\infty}$ действительно сходится к пределу x^* . ■

Побочным следствием доказательства Предложения 3.9.1 является прояснение роли нормы матрицы перехода $\|C\|$ как коэффициента подавления погрешности приближений к решению СЛАУ в согласованной векторной норме. Это следует из неравенств (3.113): чем меньше $\|C\|$, тем быстрее убывает эта погрешность на каждом отдельном шаге итерационного процесса.

Предложение 3.9.2 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

$$J = \left(\begin{array}{ccc|ccc|ccc} \lambda_1 & 1 & & & & & & & \\ & \lambda_1 & \ddots & & & & & & \\ & & \ddots & 1 & & & & & \\ & & & \lambda_1 & & & & & \\ \hline & & & & \lambda_2 & 1 & & & \\ & 0 & & & & \ddots & \ddots & & \\ & & & & & & \lambda_2 & & \\ \hline & & & & & & & & \\ & 0 & & & 0 & & & \ddots & \\ & & & & & & & & \ddots \end{array} \right),$$

а S — некоторая неособенная матрица, осуществляющая преобразование подобия. Положим

$$D_\epsilon := \text{diag} \{1, \epsilon, \epsilon^2, \dots, \epsilon^{n-1}\}$$

— диагональной $n \times n$ -матрице с числами $1, \epsilon, \epsilon^2, \dots, \epsilon^{n-1}$ по главной

диагонали. Тогда нетрудно проверить, что

$$(SD_\epsilon)^{-1}A(SD_\epsilon) = D_\epsilon^{-1}(S^{-1}AS)D_\epsilon$$

$$= D_\epsilon^{-1}JD_\epsilon = \left(\begin{array}{ccc|ccc|ccc} \lambda_1 & \epsilon & & & & & & & \\ & \lambda_1 & \ddots & & & & & & \\ & & \ddots & & & & & & \\ & & & \epsilon & & & & & \\ & & & \lambda_1 & & & & & \\ \hline & & & & \lambda_2 & \epsilon & & & \\ & & & & & \ddots & \ddots & & \\ & & & & & & \lambda_2 & & \\ \hline & & & & & & & \ddots & \\ & & & & & & & & \ddots \end{array} \right),$$

— матрица в «модифицированной» жордановой форме, которая отличается от обычной жордановой формы присутствием ϵ 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

$$\begin{aligned}
\|A\|_\epsilon &= \max_{x \neq 0} \frac{\|Ax\|_\epsilon}{\|x\|_\epsilon} = \max_{x \neq 0} \frac{\|(SD_\epsilon)^{-1}Ax\|_\infty}{\|(SD_\epsilon)^{-1}x\|_\infty} \\
&= \max_{y \neq 0} \frac{\|(SD_\epsilon)^{-1}A(SD_\epsilon)y\|_\infty}{\|y\|_\infty} \quad \text{после замены } y = (SD_\epsilon)^{-1}x \\
&= \max_{y \neq 0} \frac{\|(D_\epsilon^{-1}JD_\epsilon)y\|_\infty}{\|y\|_\infty} = \|D_\epsilon^{-1}JD_\epsilon\|_\infty \\
&= \text{максимум сумм модулей элементов в } D_\epsilon^{-1}JD_\epsilon \text{ по строкам} \\
&\leq \max_i |\lambda_i(A)| + \epsilon = \rho(A) + \epsilon,
\end{aligned}$$

где $\lambda_i(A)$ — i -ое собственное значение матрицы A . Неравенство при переходе к последней строке возникает по существу, так как матрица может иметь наибольшее по модулю собственное значение в жордановой клетке размера 1×1 , в которой нет элементов наддиагонали. ■

Доказательство Теоремы 3.9.1 о сходимости одношагового стационарного итерационного процесса.

Сначала покажем необходимость условия теоремы. Пусть порождаемая в итерационном процессе последовательность $\{x^{(k)}\}$ сходится. Её пределом при этом может быть только решение x^* системы $x = Cx + d$, т.е. должно быть $\lim_{k \rightarrow \infty} x^{(k)} = x^*$, в чём можно убедиться, переходя в соотношении

$$x^{(k+1)} = Cx^{(k)} + d$$

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

$$\begin{aligned}
 x^{(k)} - x^* &= C(x^{(k-1)} - x^*) \\
 &= C^2(x^{(k-2)} - x^*) \\
 &= \dots \dots \dots \\
 &= C^k(x^{(0)} - x^*).
 \end{aligned}$$

Так как левая часть этих 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 3.3.9 (стр. 321) заключаем, что спектральный радиус C должен быть строго меньше 1.

Достаточность. Если $\rho(C) < 1$, то, взяв положительное ϵ удовлетворяющим оценке $\epsilon < 1 - \rho(C)$, мы можем согласно Предложению 3.9.2 выбрать матричную норму $\|\cdot\|_\epsilon$ так, чтобы выполнялось неравенство $\|C\|_\epsilon < 1$. Далее в этих условиях применимо Предложение 3.9.1, которое утверждает сходимость итерационного процесса (3.112)

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

Это завершает доказательство Теоремы 3.9.1. ■

Доказанные результаты — теорема и два предложения — проясняют роль спектрального радиуса среди различных характеристик матрицы. Мы могли видеть в §3.3ж, что спектральный радиус не является матричной нормой, но, как выясняется, его 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 3.9.2. Степени матрицы A^k сходятся к нулевой матрице при $k \rightarrow \infty$ тогда и только тогда, когда $\rho(A) < 1$.

В самом деле, ранее мы установили (Предложение 3.3.9), что из сходимости степеней матрицы A^k при $k \rightarrow \infty$ к нулевой матрице вытекает $\rho(A) < 1$. Теперь результат Предложения 3.9.2 позволяет сказать,

что это условие на спектральный радиус является и достаточным: если $\rho(A) < 1$, то мы можем подобрать матричную норму так, чтобы $\|A\| < 1$, и тогда $\|A^k\| \leq \|A\|^k \rightarrow 0$ при $k \rightarrow \infty$.

С учётом Предложения 3.9.2 более точно переформулируются условия сходимости матричного ряда Неймана (Предложение 3.3.10): 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 3.9.1. 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 содержательных причин бывает известно некоторое 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. Беря $x^{(0)} = d$, мы экономим на этой итерации.

3.9в Подготовка линейной системы к итерационному процессу

В этом параграфе мы исследуем различные способы приведения системы линейных алгебраических уравнений

$$Ax = b \tag{3.114}$$

к равносильной системе в рекуррентном виде

$$x = Cx + d, \tag{3.115}$$

на основе которого можно организовывать одношаговый итерационный процесс для решения (3.114). Фактически, это вопрос о том, как связан предел стационарного одношагового итерационного процесса (3.112) с интересующим нас решением системы линейных алгебраических уравнений $Ax = b$. При этом практический интерес представляет, естественно, не всякое приведение системы (3.114) к виду (3.115), но лишь такое, которое удовлетворяет условию сходимости стационарного одношагового итерационного процесса. В предшествующем разделе мы показали, 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, мы рассмотрим ниже лишь несколько общих приёмов подготовки и организации итерационных процессов.

Простейший способ состоит в том, чтобы добавить к обеим частям исходной системы по вектору неизвестной переменной x , т. е.

$$x + Ax = x + b, \quad (3.116)$$

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 λ — собственное значение для A , то для матрицы $(I - A)$ собственным значением будет $1 - \lambda$, и тогда $1 - \lambda > 1$ при вещественных отрицательных λ . 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 мы тоже не получим.

Из предшествующих рассуждений можно ясно 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 3.9.1 *Предобуславливанием системы линейных алгебраических уравнений $Ax = b$ называется умножение слева обеих её частей на некоторую матрицу L . Сама эта матрица L называется предобуславливающей матрицей или, коротко, предобуславливателем.*

Цель предобуславливания — изменение (вообще говоря, улучшение) свойств матрицы A исходной системы $Ax = b$, вместо которой мы получаем систему

$$(LA)x = Lb.$$

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 (3.116), т.е. до прибавления вектора неизвестных x к обеим частям исходной СЛАУ. Поскольку тогда вместо системы $Ax = b$ будем иметь $(LA)x = Lb$, то далее получаем

$$x = (I - LA)x + Lb.$$

Теперь в этом рекуррентном виде с помощью подходящего выбора L можно добиваться требуемых свойств матрицы $(I - LA)$.

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 $L = A^{-1}$ 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.). 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 к главной диагонали вместе с поддиагональю и наддиагональю.

Другой способ приведения СЛАУ к рекуррентному виду основан на *расщеплении* матрицы системы.

Определение 3.9.2 *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

$$x^{(k+1)} \leftarrow G^{-1}Hx^{(k)} + G^{-1}b, \quad (3.117)$$

задавшись каким-то начальным приближением $x^{(0)}$.

Иногда по ряду причин невыгодно обращаться матрицу G явно, так что расчётные формулы итерационного метода основывают на равенстве

$$Gx = Hx + b.$$

Они могут выглядеть следующим образом

$$\begin{cases} y \leftarrow Hx^{(k)} + b, \\ x^{(k+1)} \leftarrow (\text{решение системы } Gx = y). \end{cases}$$

Итерационные методы с такой организацией 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 (3.117), 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 (3.117) 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 несложно 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 §3.9д и §3.9е мы подробно рассмотрим итерационные процессы, соответствующие первым двум пунктам из представленного списка. Детальный анализ некоторых типов расщеплений матриц читатель может увидеть в книгах [35, 116].

3.9г Скалярный предобуславливатель и его оптимизация

Напомним, что *скалярными матрицами* (из-за своего родства скалярам) называются матрицы, кратные единичным, т.е. имеющие вид τI , где $\tau \in \mathbb{R}$ или \mathbb{C} . Сейчас мы подробно исследуем описанную в предшествующем разделе возможность управления итерационным процессом на примере простейшего предобуславливания с помощью скалярной матрицы, когда $A = \tau I$, $\tau \in \mathbb{R}$ и $\tau \neq 0$.

Итак, рассматриваем итерационный процесс

$$x^{(k+1)} \leftarrow (I - \tau A) x^{(k)} + \tau b, \quad (3.118)$$

$\tau = \text{const.}$ Впервые он был предложен в 1910 году в работе [109] и обычно называется *методом Ричардсона* (см. далее §3.10а). Если λ_i ,

$i = 1, 2, \dots, n$, — собственные числа матрицы A (вообще говоря, они комплексны), то собственные числа матрицы $(I - \tau A)$ равны $(1 - \tau \lambda_i)$. Ясно, что в случае, когда среди λ_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 τ будет иметь как меньшие 1 значения для каких-то λ_i , так и большие чем 1 значения для некоторых других λ_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 τ будет невозможно.

Далее рассмотрим практически важный частный случай, когда A — симметричная положительно определённая матрица, так что все λ_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 $\mu > 0$.

Матрица $(I - \tau A)$ тогда тоже симметрична, и потому её спектральный радиус совпадает с 2-нормой. Чтобы обеспечить сходимость итерационного процесса и добиться её наибольшей скорости, нам нужно, согласно Теореме 3.9.1 и оценкам убывания погрешности (3.113), найти значение τ , 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 λ_i матрицы A , $i = 1, 2, \dots, n$. В условиях, когда о расположении λ_i ничего не известно кроме их принадлежности интервалу $[\mu, M]$, естественно заменить максимизацию по множеству всех λ_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

$$\min_{\tau} \|I - \tau A\|_2 = \min_{\tau} \max_{\lambda_i} |1 - \tau \lambda_i| \leq \Theta. \quad (3.119)$$



Рис. 3.22. Спектр положительно определённой матрицы системы и объемлющий его интервал.

Обозначив

$$g(\tau) := \max_{\mu \leq \lambda \leq M} |1 - \tau \lambda|,$$

обратимся для минимизации функции $g(\tau)$ к наглядной иллюстрации на Рис. 3.23. Пользуясь ею, мы исследуем поведение $g(\tau)$ при изменении аргумента τ .

При $\tau \leq 0$ выражение $(1 - \tau \lambda)$ не убывает по λ , и при положительных λ , очевидно, имеет значения не меньше 1 (на Рис. 3.23 этому случаю соответствует прямая, идущая от точки $(0, 1)$ вверх). Тогда итерационный процесс (3.118) сходиться не будет. Следовательно, в нашем анализе имеет смысл ограничиться только теми τ , для которых $(1 - \tau \lambda)$ убывает по λ . Это значения $\tau > 0$, и на Рис. 3.23 им соответствуют прямые, идущие от точки с координатами $(0, 1)$ вниз.

При $0 < \tau \leq M^{-1}$ выражение $(1 - \tau \lambda)$ на интервале $\lambda \in [\mu, M]$ неотрицательно и монотонно убывает по λ . Поэтому

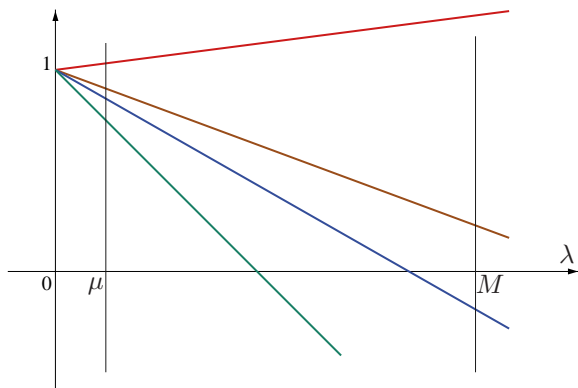
$$g(\tau) = \max_{\lambda} |1 - \tau \lambda| = 1 - \tau \mu,$$

где максимум по λ 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 τ , а на правом конце $-(1 - \tau M)$ растёт с ростом τ .

Рис. 3.23. Графики функций $1 - \tau\lambda$ для различных τ

При некотором $\tau = \tau_{\text{опт}}$ наступает момент, когда эти значения на концах интервала $[\mu, M]$ сравниваются друг с другом:

$$1 - \tau\mu = -(1 - \tau M).$$

Он и является моментом достижения оптимума, поскольку дальнейшее увеличение τ приводит к росту $-(1 - \tau M)$ на правом конце интервала, а уменьшение τ ведёт к росту $(1 - \tau\mu)$ на левом конце. В любом из этих случаев $g(\tau)$ возрастает. Отсюда

$$\tau_{\text{опт}} = \frac{2}{M + \mu}, \quad (3.120)$$

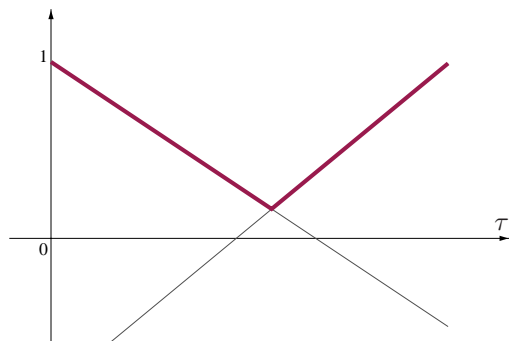
а значение оптимума $g(\tau)$ равно

$$\begin{aligned} \Theta &= \min_{\tau} \max_{\lambda \in [\mu, M]} |1 - \tau\lambda| = 1 - \tau_{\text{опт}}\mu \\ &= 1 - \frac{2}{M + \mu} \cdot \mu = \frac{M - \mu}{M + \mu}. \end{aligned}$$

Соответственно, в силу неравенства (3.119),

$$\|I - \tau_{\text{опт}}A\|_2 \leq \Theta = \frac{M - \mu}{M + \mu}, \quad (3.121)$$

и эту величину можно рассматривать, как коэффициент подавления евклидовой нормы погрешности (ввиду неравенств (3.113)). Она меньше

Рис. 3.24. График функции $g(\tau)$

единицы, т. е. даже с помощью простейшего скалярного предобуславливателя мы добились сходимости итерационного процесса.

Полезно оценить значение (3.121) через спектральное число обусловленности матрицы A . Так как $\mu \leq \lambda_{\min}(A)$ и $\lambda_{\max}(A) \leq M$, то для положительно определённой матрицы A справедливо

$$\text{cond}_2(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \leq \frac{M}{\mu}.$$

Поэтому, 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 медленнее, вообще говоря, сходимость нашего итерационного процесса. Иначе говоря, число обусловленности матрицы системы характеризует не только чувствительность её решения к возмущениям и погрешностям, но и скорость сходимости итерационных процессов. Мы увидим далее, что это характерно для поведения многих итерационных методов (см. §§3.10в, 3.10г, 3.10д).

Наибольшую трудность на практике представляет нахождение μ , 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

$$\tau_{\text{опт}} = \frac{2}{M + \mu} = \frac{2}{M}, \quad (3.122)$$

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 (3.122) является разумным значением τ , при котором обеспечивается сходимость итераций метода Ричардсона (3.118). Естественно, что о какой-либо оптимальности выбранного параметра говорить не приходится.

3.9д Итерационный метод Якоби

Пусть в системе линейных алгебраических уравнений $Ax = b$ диагональные элементы квадратной $n \times n$ -матрицы $A = (a_{ij})$ отличны от нуля, т.е. $a_{ii} \neq 0$, $i = 1, 2, \dots, n$. Это условие несколько не ограничит общность наших рассуждений, так как в неособенной матрице в каждой строке и каждом столбце должны присутствовать ненулевые элементы. Далее с помощью перестановки строк матрицы (соответствующей перестановке уравнений системы) всегда можно сделать её диагональные элементы ненулевыми.

Перепишем систему линейных алгебраических уравнений $Ax = b$ в развёрнутом виде:

$$\sum_{j=1}^n a_{ij} x_j = b_i, \quad i = 1, 2, \dots, n.$$

Та 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 $x = T(x)$, необходимым для организации одношаговых итераций

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

Более точно, можно взять

$$T(x) = (T_1(x), T_2(x), \dots, T_n(x))^T$$

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 3.5. Итерационный метод Якоби для решения СЛАУ

```

k ← 0;
выбираем начальное приближение  $x^{(0)}$ ;
DO WHILE ( метод не сошёлся )
    DO FOR i = 1 TO n
        
$$x_i^{(k+1)} \leftarrow \frac{1}{a_{ii}} \left( b_i - \sum_{j \neq i} a_{ij} x_j^{(k)} \right)$$

    END DO
    k ← k + 1;
END DO

```

Псевдокод соответствующего итерационного процесса представлен в Табл. 3.5 (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 следующего

приближения $x^{(k+1)}$ насчитываются независимо друг от друга по единообразным формулам, основанным на использовании лишь предыдущего приближения $x^{(k)}$.

В следующем параграфе будет рассмотрен итерационный процесс — метод Гаусса-Зейделя, устроенный несколько по-другому, в котором смещения-коррекции компонент очередного приближения к решению «не одновременны» в том смысле, что находятся последовательно одна за другой не только из предыдущего приближения, но ещё и друг из друга.

Пусть $A = \tilde{L} + D + \tilde{U}$, где

$$\tilde{L} = \begin{pmatrix} 0 & & & & \\ a_{21} & 0 & & & \\ a_{31} & a_{32} & \ddots & & \\ \vdots & \vdots & \ddots & 0 & \\ a_{n1} & a_{n2} & \cdots & a_{n,n-1} & 0 \end{pmatrix} \quad \begin{array}{l} \text{— строго нижняя} \\ \text{треугольная матрица,} \end{array}$$

$$D = \text{diag} \{a_{11}, a_{22}, \dots, a_{nn}\} \quad \text{— диагональ матрицы } A,$$

$$\tilde{U} = \begin{pmatrix} 0 & a_{12} & \cdots & a_{1,n-1} & a_{1n} \\ & 0 & \ddots & a_{2,n-1} & a_{2n} \\ & & \ddots & \vdots & \vdots \\ & & & 0 & a_{n-1,n} \\ \mathbf{0} & & & & 0 \end{pmatrix} \quad \begin{array}{l} \text{— строго верхняя} \\ \text{треугольная матрица.} \end{array}$$

Тогда итерационный метод Якоби может быть представлен как метод, основанный на таком расщеплении матрицы системы $A = G - H$ (см. §3.9в), что

$$G = D, \quad H = -(\tilde{L} + \tilde{U}).$$

Соответственно, в матричном виде метод Якоби записывается как

$$x^{(k+1)} \leftarrow -D^{-1}(\tilde{L} + \tilde{U})x^{(k)} + D^{-1}b, \quad k = 0, 1, 2, \dots$$

Теперь нетрудно дать условия его сходимости, основываясь на общем результате о сходимости стационарных одношаговых итераций

(Теорема 3.9.1). 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

$$\begin{pmatrix} 0 & a_{12}/a_{11} & \dots & a_{1n}/a_{11} \\ a_{21}/a_{22} & 0 & \dots & a_{2n}/a_{22} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}/a_{nn} & a_{n2}/a_{nn} & \dots & 0 \end{pmatrix}. \quad (3.123)$$

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 3.9.2 *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 (∞ -норму) матрицы $D^{-1}(\tilde{L} + \tilde{U})$, которая была выписана нами в (3.123). Таким образом,

$$\|D^{-1}(\tilde{L} + \tilde{U})\|_{\infty} < 1,$$

откуда, ввиду результата Предложения 3.9.1, 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.²⁷ Тем не менее, совсем забывать метод Якоби было бы преждевременным. Во-первых, он очень хорошо распараллеливается, и это благоприятствует его реализации на ЭВМ с современными архитектурами (см. [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 . В случае, когда $b(x)$ — нелинейная функция, никакие численные методы для решения СЛАУ здесь уже неприменимы, но для отыскания решения мы можем воспользоваться незначительной модификацией итераций Якоби

$$x_i^{(k+1)} \leftarrow \frac{1}{a_{ii}} \left(b_i(x^{(k)}) - \sum_{j \neq i} a_{ij} x_j^{(k)} \right), \quad i = 1, 2, \dots, n, \quad (3.124)$$

$k = 0, 1, 2, \dots$, с некоторым начальным приближением $x^{(0)}$.

Обозначим $n \times n$ -матрицу Якоби вектор-функции $b(x)$ через $b'(x)$. Если $b(x)$ изменяется «достаточно медленно», так что

$$\rho(D^{-1}(\tilde{L} + \tilde{U} + b'(x))) < 1$$

²⁷Примеры применения и детальные оценки скорости сходимости метода Якоби для решения модельных задач математической физики можно увидеть в [40].

для любых $x \in \mathbb{R}^n$, то итерационный процесс (3.124) сходится из произвольного начального приближения. Это нетрудно показать, применяя теорему о среднем для функции $b(x)$ и затем теорему Шрёдера о неподвижной точке (Теорема 4.4.5, стр. 588) к отображению, которое задаётся правой частью (3.124).

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 начальным приближением $x^{(0)}$, на очередном k -ом шаге последовательно находят решения \tilde{x}_i уравнений

$$F_i(x_1^{(k)}, \dots, x_{i-1}^{(k)}, x_i, x_{i+1}^{(k)}, \dots, x_n^{(k)}) = 0, \quad i = 1, 2, \dots, n$$

относительно x_i , а затем полагают $x_i^{(k+1)} \leftarrow \tilde{x}_i$, $i = 1, 2, \dots, n$.

3.9е Итерационный метод Гаусса-Зейделя

В итерационном методе Якоби при организации вычислений по инструкции

$$x_i^{(k+1)} \leftarrow \frac{1}{a_{ii}} \left(b_i - \sum_{j \neq i} a_{ij} x_j^{(k)} \right), \quad i = 1, 2, \dots, n, \quad (3.125)$$

компоненты очередного приближения $x^{(k+1)}$ находятся последовательно одна за другой, так что к моменту вычисления i -ой компоненты вектора $x^{(k+1)}$ уже найдены $x_1^{(k+1)}, x_2^{(k+1)}, \dots, x_{i-1}^{(k+1)}$. Но метод Якоби никак не использует эти новые значения, и при вычислении любой компоненты следующего приближения всегда опирается только на вектор $x^{(k)}$ предшествующего приближения. Если итерации сходятся к решению, то естественно ожидать, что все компоненты $x^{(k+1)}$ ближе к искомому решению, чем $x^{(k)}$, а посему немедленное вовлечение их в процесс вычислений будет способствовать ускорению сходимости.

Таблица 3.6. Итерационный метод Гаусса-Зейделя
для решения линейных систем уравнений

```

 $k \leftarrow 0;$ 
выбираем начальное приближение  $x^{(0)};$ 
DO WHILE ( метод не сошёлся )
  DO FOR  $i = 1$  TO  $n$ 
    
$$x_i^{(k+1)} \leftarrow \frac{1}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij} x_j^{(k)} \right)$$

  END DO
   $k \leftarrow k + 1;$ 
END DO

```

На этой идее основан *итерационный метод Гаусса-Зейделя*, идея которого была высказана сначала К.Ф. Гауссом, а в окончательной форме он появился в публикации Ф.Л. Зейделя в 1874 году.²⁸ Псевдокод этого метода представлен в Табл. 3.6 (где, как и ранее, k — счётчик итераций). В нём суммирование в формуле (3.125) для вычисления i -ой компоненты очередного приближения $x^{(k+1)}$ разбито на две части — по индексам, предшествующим i , и по индексам, следующим за i . Первая часть суммы использует новые вычисленные значения $x_1^{(k+1)}, \dots, x_{i-1}^{(k+1)}$, тогда как вторая — компоненты $x_{i+1}^{(k)}, \dots, x_n^{(k)}$ из старого приближения. Метод Гаусса-Зейделя иногда называют также итерационным методом «последовательных смещений», а его основная идея — немедленно вовлекать уже полученную информацию в вычислительный процесс — с успехом применима и для нелинейных итерационных схем.

Чтобы получить для метода Гаусса-Зейделя матричное представле-

²⁸По этой причине в отечественной литературе по вычислительной математике нередко используется термин «метод Зейделя».

ние, перепишем его расчётные формулы в виде

$$a_{ii}x_i^{(k+1)} + \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} = - \sum_{j=i+1}^n a_{ij}x_j^{(k)} + b_i, \quad i = 1, 2, \dots, n.$$

Используя введённые в §3.9д матрицы \tilde{L} , D и \tilde{U} , на которые разлагается A , можем записать эти формулы в виде

$$(D + \tilde{L})x^{(k+1)} = -\tilde{U}x^{(k)} + b,$$

т. е.

$$x^{(k+1)} = -(D + \tilde{L})^{-1}\tilde{U}x^{(k)} + (D + \tilde{L})^{-1}b, \quad k = 0, 1, 2, \dots \quad (3.126)$$

Таким образом, метод Гаусса-Зейделя можно рассматривать как итерационный метод, порождённый таким расщеплением матрицы СЛАУ в виде $A = G - H$ (см. §3.9в), что $G = D + \tilde{L}$, $H = -\tilde{U}$.

В силу Теоремы 3.9.1 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 3.9.3 *Если в системе линейных алгебраических 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, §3.5в). 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

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

можем заключить, что

$$z_i^{(k+1)} = \frac{1}{a_{ii}} \left(- \sum_{j=1}^{i-1} a_{ij} z_j^{(k+1)} - \sum_{j=i+1}^n a_{ij} z_j^{(k)} \right), \quad i = 1, 2, \dots, n.$$

Возьмём абсолютные значения от обеих частей этих равенств и воспользуемся неравенством треугольника для оценки сумм в правых частях. Мы будем иметь

$$\begin{aligned} |z_i^{(k+1)}| &\leq \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| \cdot |z_j^{(k+1)}| + \sum_{j=i+1}^n \left| \frac{a_{ij}}{a_{ii}} \right| \cdot |z_j^{(k)}| \\ &\leq \|z^{(k+1)}\|_{\infty} \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| + \|z^{(k)}\|_{\infty} \sum_{j=i+1}^n \left| \frac{a_{ij}}{a_{ii}} \right| \end{aligned} \quad (3.127)$$

для $i = 1, 2, \dots, n$.

С другой стороны, условие диагонального преобладания в матрице A решаемой системы уравнений, 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

$$\sum_{j \neq i} |a_{ij}| \leq \varkappa |a_{ii}|, \quad i = 1, 2, \dots, n. \quad (3.128)$$

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 (3.127), приходим к соотношениям

$$|z_i^{(k+1)}| \leq \|z^{(k+1)}\|_\infty \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| + \varkappa \|z^{(k)}\|_\infty \left(1 - \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| \right), \quad (3.129)$$

$i = 1, 2, \dots, n$.

Предположим, что $\max_{1 \leq i \leq n} |z_i^{(k+1)}|$ достигается при $i = l$, так что

$$\|z^{(k+1)}\|_\infty = |z_l^{(k+1)}|. \quad (3.130)$$

Рассмотрим теперь отдельно l -ое неравенство из (3.129). Привлекая равенство (3.130), можем утверждать, что

$$\|z^{(k+1)}\|_\infty \leq \|z^{(k+1)}\|_\infty \sum_{j=1}^{l-1} \left| \frac{a_{lj}}{a_{ll}} \right| + \varkappa \|z^{(k)}\|_\infty \left(1 - \sum_{j=1}^{l-1} \left| \frac{a_{lj}}{a_{ll}} \right| \right),$$

то есть

$$\|z^{(k+1)}\|_\infty \left(1 - \sum_{j=1}^{l-1} \left| \frac{a_{lj}}{a_{ll}} \right| \right) \leq \varkappa \|z^{(k)}\|_\infty \left(1 - \sum_{j=1}^{l-1} \left| \frac{a_{lj}}{a_{ll}} \right| \right). \quad (3.131)$$

Конечно, значение индекса l , на котором достигается равенство (3.130), может меняться в зависимости от номера итерации k . Но так как вплоть до оценки (3.129) мы отслеживали *все* компоненты погрешности $z_i^{(k+1)}$, то вне зависимости от k неравенство (3.131) должно быть справедливым для компоненты с номером l , определяемой условием (3.130).

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 (3.131). Окончательно получаем

$$\|z^{(k+1)}\|_\infty \leq \varkappa \|z^{(k)}\|_\infty,$$

что при $|\varkappa| < 1$ означает сходимость метода Гаусса-Зейделя. ■

Фактически, в доказательстве Предложения 3.9.3 мы получили даже оценку уменьшения чебышёвской нормы погрешности решения с помощью «меры диагонального преобладания» в матрице СЛАУ, в качестве которой выступает величина \varkappa , определённая посредством (3.128).

Теорема 3.9.4 *Если в системе линейных алгебраических уравнений $Ax = b$ матрица A является симметричной и положительно определённой, то метод Гаусса-Зейделя сходится к решению из любого начального приближения.*

Доказательство может быть найдено, к примеру, в [4, 11]. Теорема 3.9.4 является частным случаем теоремы Островского-Райха (теорема 3.9.5), которая, в свою очередь, может быть получена как следствие из более общей теории итерационных методов, развитой А.А. Самарским. Её начала мы излагаем в §3.12.

Метод Гаусса-Зейделя был сконструирован как модификация метода Якоби, и, казалось бы, должен работать лучше. Так оно и есть «в среднем», на случайно выбранных системах — метод Гаусса-Зейделя работает несколько быстрее, что можно показать математически строго при определённых допущениях на систему. Но в целом ситуация не столь однозначна. Для СЛАУ размера 3×3 и более существуют примеры, на которых метод Якоби расходится, а метод Гаусса-Зейделя сходится (см. Пример 3.9.1), так же как существуют и примеры другого свойства, когда метод Якоби сходится, а метод Гаусса-Зейделя расходится. В частности, для метода Якоби неверна Теорема 3.9.4, и он может расходиться для систем линейных уравнений с симметричными положительно-определёнными матрицами (Пример 3.9.1).

По поводу практического применения метода Гаусса-Зейделя можно сказать почти то же самое, что и о методе Якоби в §3.9д. 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. В последние годы очень большое значение приобрёл интервальный метод Гаусса-Зейделя, предназначенный для внешнего оценивания множеств решений интервальных систем линейных алгебраических уравнений (см. §4.6а).

3.9ж Методы релаксации

Одним из принципов, который кладётся в основу итерационных методов решения систем уравнений, является так называемый *принцип релаксации*.²⁹ 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 косвенных признаков, важнейшим среди которых является величина *невязки* решения. Невязка определяется как разность левой и правой частей уравнения после подстановки в него приближения к решению, и в нашем случае это $Ax^{(k)} - b$. При этом 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, поскольку 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, и без учёта влияния результатов вычисления этой компоненты 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 может казаться наиболее выгодной, но глобально, с точки зрения сходимости процесса в целом, тщательно подобранная неполная релаксация нередко приводит 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 $x^{(k)}$ и последующей $x^{(k+1)}$ итераций метода Гаусса-Зейделя. Более точно, зададимся веществен-

²⁹От латинского слова «relaxatio» — уменьшение напряжения, ослабление.

ным числом ω , которое будем называть *параметром релаксации*, и i -ую компоненту очередного $(k+1)$ -го приближения положим равной

$$\omega x_i^{(k+1)} + (1 - \omega)x_i^{(k)},$$

где $x_i^{(k)}$ — i -ая компонента приближения, полученного в результате k -го шага алгоритма, а $x_i^{(k+1)}$ — i -ая компонента приближения, которое было бы получено на основе $x^{(k)}$ и $x_1^{(k+1)}, \dots, x_{i-1}^{(k+1)}, x_{i+1}^{(k)}, \dots, x_n^{(k)}$ с помощью метода Гаусса-Зейделя. Псевдокод получающегося итерационного алгоритма, который обычно и называют методом релаксации для решения систем линейных алгебраических уравнений, представлен в Табл. 3.7.

Таблица 3.7. Псевдокод метода релаксации для решения систем линейных уравнений

```

 $k \leftarrow 0$ ;
выбираем начальное приближение  $x^{(0)}$ ;
DO WHILE ( метод не сошёлся )
  DO FOR  $i = 1$  TO  $n$ 
     $x_i^{(k+1)} \leftarrow (1 - \omega) x_i^{(k)}$ 
    
$$+ \frac{\omega}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij} x_j^{(k)} \right)$$

  END DO
   $k \leftarrow k + 1$ ;
END DO

```

Расчётные формулы этого метода перепишем в виде

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

$$i = 1, 2, \dots, n,$$

$k = 0, 1, 2, \dots$. Далее, используя введенные выше в §3.9е матрицы \tilde{L} , D и \tilde{U} , можно придать этим соотношениям более компактный вид

$$(D + \omega\tilde{L})x^{(k+1)} = ((1 - \omega)D - \omega\tilde{U})x^{(k)} + \omega b,$$

откуда

$$x^{(k+1)} = (D + \omega\tilde{L})^{-1}((1 - \omega)D - \omega\tilde{U})x^{(k)} + (D + \omega\tilde{L})^{-1}\omega b,$$

$k = 0, 1, 2, \dots$.

В зависимости от конкретного значения параметра релаксации принято различать три случая:

если $\omega < 1$, то говорят о «нижней релаксации»,

если $\omega = 1$, то имеем итерации Гаусса-Зейделя,

если $\omega > 1$, то говорят о «верхней релаксации».³⁰

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 ω и привлек внимание к случаю $\omega > 1$, который во многих ситуациях действительно обеспечивает существенное ускорение сходимости итераций в сравнении с методом Гаусса-Зейделя. Несколько упрощенное объяснение этого явления может состоять в том, что если направление от $x^{(k)}$ к $x^{(k+1)}$ оказывается удачным в том смысле, что приближает к искомому решению, то имеет смысл пройти по нему и дальше, за $x^{(k+1)}$. Это и соответствует случаю $\omega > 1$.

Пример 3.9.1 У системы линейных алгебраических уравнений

$$\begin{pmatrix} 6 & 2 & 2 \\ 2 & 3 & 4 \\ 2 & 4 & 8 \end{pmatrix} x = \begin{pmatrix} 6 \\ 3 \\ 6 \end{pmatrix} \quad (3.132)$$

матрица симметрична и положительно определена, а точное решение есть $(1, -1, 1)^T$. Применим к этой системе все рассмотренные нами стационарные итерационные процессы.

³⁰В англоязычной литературе по вычислительной линейной алгебре этот метод обычно обозначают аббревиатурой $\text{SOR}(\omega)$, которая происходит от термина «Successive OverRelaxation» — «последовательная перерелаксация».

Первым тестируем простейший итерационный метод Ричардсона из §3.9г. При оптимальном значении параметра предобуславливания (равном $\tau = 0.1638$) метод сходится из нулевого начального вектора за 148 шагов к приближению, которое обеспечивает невязку 10^{-6} в 1-норме.

Метод Якоби для системы (3.132) расходится из любого начального приближения, отличного от самого решения.

Метод Гаусса-Зейделя для системы (3.132) 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. ■

Важно отметить, что метод релаксации тоже укладывается в изложенную ранее в §3.9в схему итерационных процессов, порождаемых расщеплением матрицы системы уравнений. Именно, мы используем при этом 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 принимает поэтому вид

$$\rho(G_\omega^{-1} H_\omega) < 1.$$

Для некоторых классов специальных, но важных задач математической физики значение релаксационного параметра ω , при котором величина $\rho(G_\omega^{-1} H_\omega)$ достигает минимума или близка к нему, находится относительно просто. В более сложных задачах для оптимизации ω требуется весьма трудный анализ спектра матрицы перехода $G_\omega^{-1} H_\omega$ из представления (3.117). Обзоры состояния дел в этой области читатель может найти в [30, 35, 48, 52, 69, 89, 111, 116].

Предложение 3.9.3 (лемма Кэхэна) Пусть рассматривается метод релаксации с параметром ω , 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, как следствие, для сходимости метода релаксации из любого начального приближения необходимо выполнение неравенства $0 < \omega < 2$.

Доказательство. Прежде всего, преобразуем матрицу C_ω для придания ей более удобного для дальнейших выкладок вида:

$$\begin{aligned} C_\omega &= (D + \omega \tilde{L})^{-1} ((1 - \omega)D - \omega \tilde{U}) \\ &= (D(I + \omega D^{-1} \tilde{L}))^{-1} ((1 - \omega)D - \omega \tilde{U}) \\ &= (I + \omega D^{-1} \tilde{L})^{-1} D^{-1} ((1 - \omega)D - \omega \tilde{U}) \\ &= (I + \omega D^{-1} \tilde{L})^{-1} ((1 - \omega)I - \omega D^{-1} \tilde{U}). \end{aligned}$$

Желая исследовать расположение собственных чисел $\lambda_i(C_\omega)$ матрицы C_ω , 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 $\phi(0)$:

$$\begin{aligned} p_0 &= \det C_\omega = \det\left((I + \omega D^{-1} \tilde{L})^{-1} ((1 - \omega)I - \omega D^{-1} \tilde{U})\right) \\ &= \det((I + \omega D^{-1} \tilde{L})^{-1}) \cdot \det((1 - \omega)I - \omega D^{-1} \tilde{U}) \\ &= \det((1 - \omega)I - \omega D^{-1} \tilde{U}) = (1 - \omega)^n, \end{aligned}$$

когда 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, по теореме Виета свободный член характеристического полинома матрицы, делённый на старший коэффициент, равен произведению его корней, т. е. собственных чисел матрицы, умноженному на $(-1)^n$ (см., к примеру, [23]), 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. Это завершает доказательство Предложения. ■

Лемма Кэжэна даёт лишь необходимое условие сходимости метода релаксации, которое иногда не является достаточным. Это показывает следующий

Пример 3.9.2 Решением системы линейных алгебраических уравнений

$$\begin{pmatrix} 2 & 1 & -1 \\ 1 & -5 & 4 \\ 3 & 2 & 6 \end{pmatrix} x = \begin{pmatrix} 0 \\ 10 \\ 7 \end{pmatrix} \quad (3.133)$$

является $(1, -1, 1)^T$. 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 (3.133) расходится. ■

Один из популярных вариантов необходимых и достаточных условий сходимости метода релаксации —

Теорема 3.9.5 (теорема Островского-Райха) Пусть A — вещественная симметричная матрица с положительными диагональными элементами и $\omega \in]0, 2[$. Метод релаксации с параметром ω 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, 111, 116]. Обоснование теоремы Островского-Райха будет также дано ниже в §3.12 как следствие теоремы Самарского, дающей достаточные условия сходимости для итерационных методов весьма общего вида.

В заключение темы отметим, что на практике популярен так называемый симметричный метод релаксации (его английская аббревиатура — SSOR, от Symmetric Successive Overrelaxation).

3.10 Нестационарные итерационные методы для линейных систем

3.10а Теоретическое введение

В этом параграфе для решения систем линейных алгебраических уравнений мы рассмотрим нестационарные итерационные методы, которые распространены не меньше стационарных. В основу нестационарных итерационных методов могут быть положены различные идеи, так что соответствующие численные методы для решения СЛАУ отличаются огромным разнообразием. В нашем учебнике мы даём лишь беглый обзор основных идей и подходов.

В качестве первого примера рассмотрим простейший итерационный процесс Рундсона (3.118)

$$x^{(k+1)} \leftarrow (I - \tau A) x^{(k)} + \tau b, \quad k = 0, 1, 2, \dots,$$

исследованный нами в §3.9г. Если переписать его в виде

$$x^{(k+1)} \leftarrow x^{(k)} - \tau(Ax^{(k)} - b), \quad k = 0, 1, 2, \dots, \quad (3.134)$$

то расчёт каждой последующей итерации $x^{(k+1)}$ может трактоваться как вычитание из $x^{(k)}$ поправки, пропорциональной вектору текущей невязки $(Ax^{(k)} - b)$. Но при таком взгляде на итерационный процесс можно попытаться изменять параметр τ в зависимости от шага, т. е. взять $\tau = \tau_k$ переменным и рассмотреть итерации

$$x^{(k+1)} \leftarrow x^{(k)} - \tau_k(Ax^{(k)} - b), \quad k = 0, 1, 2, \dots \quad (3.135)$$

Этот простейший нестационарный итерационный метод тоже связывают с именем Л.Ф. Рундсона, который рассмотрел его в работе [109]. Он, к сожалению, не смог развить удовлетворительной теории выбора параметров τ_k , и для решения этого вопроса потребовалось ещё несколько десятилетий развития вычислительной математики. Отметим, что задача об оптимальном выборе параметров τ_k на группе из нескольких шагов приводит к так называемым чебышёвским циклическим итерационным методам (см. [40, 48, 88, 89]).

Можно пойти по намеченному выше пути дальше, рассмотрев нестационарное обобщение итерационного процесса

$$x^{(k+1)} \leftarrow (I - \Lambda A) x^{(k)} + \Lambda b, \quad k = 0, 1, 2, \dots,$$

который получен в результате матричного предобуславливания исходной системы линейных уравнений $Ax = b$. Переписав его вычислительную схему в виде

$$x^{(k+1)} \leftarrow x^{(k)} - \Lambda(Ax^{(k)} - b), \quad k = 0, 1, 2, \dots,$$

нетрудно увидеть возможность изменения предобуславливающей матрицы Λ в зависимости от номера шага. Таким образом, приходим к весьма общей схеме нестационарных линейных итерационных процессов

$$x^{(k+1)} \leftarrow x^{(k)} - \Lambda_k(Ax^{(k)} - b), \quad k = 0, 1, 2, \dots,$$

где $\{\Lambda_k\}_{k=0}^{\infty}$ — некоторая последовательность матриц. Выбор $\{\Lambda_k\}$, при котором этот процесс сходится, зависит, вообще говоря, от начального приближения $x^{(0)}$.

Другой популярный путь построения нестационарных итерационных методов для решения уравнений — использование *вариационных принципов*.

Интуитивно понятный термин «вариация» был введён в математику Ж.-Л. Лагранжем для обозначения малого изменения («шевеления») независимой переменной или рассматриваемой функции. Если в рассматриваемой точке эти вариации приводят к изменениям значений функции, знак которых одинаков, то имеем экстремум. Соответственно, метод исследования экстремумов, основанный на изучении зависимости функций и функционалов от вариаций их аргументов, получил название *метода вариаций*. Но со временем «вариационными» стали именовать методы решения различных уравнений, которые сводят исходную постановку к тем или иным задачам на нахождение экстремума. Согласно этой терминологии, *вариационными принципами* теперь называют переформулировки интересующих нас задач в виде каких-либо оптимизационных задач, т.е. задач на нахождение минимумов или максимумов. Тогда итерационные методы решения СЛАУ могут конструироваться как итерационные процессы для отыскания этих экстремумов тех или иных функционалов.

Вариационные принципы получаются весьма различными способами. 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» (см., к при-

меру, [80]). В последнее столетие имеется тенденция всё меньше связывать вариационные принципы с конкретным физическим содержанием, и они становятся абстрактным математическим инструментом решения разнообразных задач.

Строго говоря, в вычислительном отношении получающаяся в результате описанного выше сведения оптимизационная задача может быть не вполне эквивалентна исходной, так как задача нахождения устойчивого решения уравнения может превратиться в неустойчивую задачу о проверке точного равенства экстремума нулю (этот вопрос более подробно обсуждается далее в §4.26). 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 важным подспорьем практических вычислений. Именно такова ситуация с системами линейных алгебраических уравнений, разрешимость которых часто обеспечивается различными результатами из линейной алгебры.

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, таким образом, наименьшее возможное значение:

$$\boxed{\text{решение } Ax = b} \iff \boxed{\text{нахождение } \min \|Ax - b\|} .$$

Рассматривая конкретные векторные нормы, получаем различные вариационные переформулировки задачи решения системы линейных алгебраических уравнений.

Выбор той или иной нормы в этой конструкции существенно влияет на свойства получающейся оптимизационной задачи, и на практике чаще всего используют евклидову норму вектора невязки, которая порождена скалярным произведением и обладает рядом других хороших свойств. Наконец, желая получить 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 §2.10е и рассмотрим подробнее в §3.15.

Ещё одним фактом, который служит теоретической основой для ва-

риационных методов решения систем линейных алгебраических уравнений является

Теорема 3.10.1 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,

$$\boxed{\text{решение } Ax = b} \iff \boxed{\text{нахождение } \min \Psi(x)}.$$

Доказательство. Если A — симметричная положительно-определённая матрица, то решение x^* системы линейных уравнений $Ax = b$ существует и единственно. Другим следствием симметричности и положительной определённости A является то, что она порождает, как мы видели в §3.3а, так называемую 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 $x \in \mathbb{R}^n$ является решением системы уравнений тогда и только тогда, когда $x - x^* = 0$. Это, в свою очередь, равносильно занулению нормы погрешности, т. е. $\|x - x^*\|_A = 0$, что можно переформулировать также в следующем виде:

$$x \text{ есть решение системы } Ax = b \iff \frac{1}{2}\|x - x^*\|_A^2 = 0.$$

Преобразуем выражение из правой части этой эквивалентности, учитывая симметричность матрицы A , равенство $Ax^* = b$ и определение

энергетической нормы:

$$\begin{aligned}
 \frac{1}{2}\|x - x^*\|_A^2 &= \frac{1}{2}\langle A(x - x^*), x - x^* \rangle \\
 &= \frac{1}{2}\langle Ax, x \rangle - \frac{1}{2}\langle Ax, x^* \rangle - \frac{1}{2}\langle Ax^*, x \rangle + \frac{1}{2}\langle Ax^*, x^* \rangle \\
 &= \frac{1}{2}\langle Ax, x \rangle - \frac{1}{2}\langle Ax^*, x \rangle - \frac{1}{2}\langle Ax^*, x \rangle + \frac{1}{2}\|x^*\|_A^2 \\
 &= \frac{1}{2}\langle Ax, x \rangle - \langle b, x \rangle + \frac{1}{2}\|x^*\|_A^2 \\
 &= \Psi(x) + \frac{1}{2}\|x^*\|_A^2.
 \end{aligned}$$

Иными словами, 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. п. Для составных систем, образованных из нескольких частей, квадрат одной переменной в этих выражениях заменяется на квадратичную форму. Естественность присутствия множителя $\frac{1}{2}$ в выражении для $\Psi(x)$ получит также дополнительное обоснование в следующем разделе.

Равенство

$$\Psi(x) = \frac{1}{2}\|x - x^*\|_A^2 - \frac{1}{2}\|x^*\|_A^2 \quad (3.136)$$

— отдельное важное следствие доказательства Теоремы 3.10.1. Оно показывает, что функционал энергии лишь на константу отличается от энергетической A -нормы погрешности приближения к решению, и этот факт будет неоднократно использоваться далее.

Поскольку A — симметричная матрица, то ортогональным преобразованием подобия она может быть приведена к диагональной матрице:

$$A = Q^\top D Q,$$

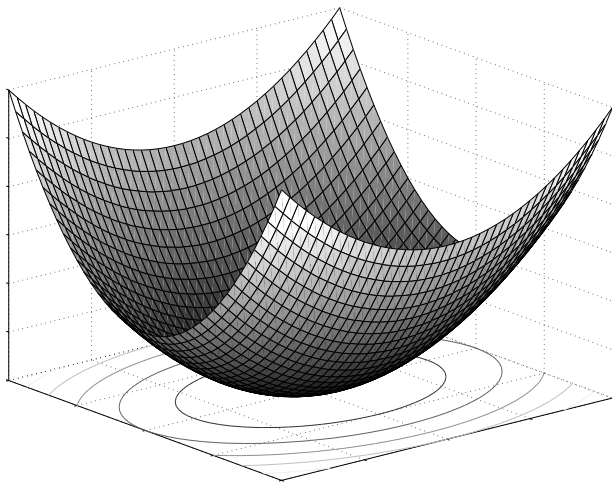


Рис. 3.25. Типичный график функционала энергии и его линии уровня.

где Q ортогональна, $D = \text{diag} \{ \lambda_1, \lambda_2, \dots, \lambda_n \}$, а λ_i — собственные значения матрицы A , причём все $\lambda_i > 0$ в силу положительной определённости A . Подставляя это представление в выражение для функционала энергии $\Psi(x)$, получим

$$\begin{aligned} \Psi(x) &= \frac{1}{2} \langle Q^T D Q x, x \rangle - \langle b, x \rangle = \frac{1}{2} \langle D(Qx), Qx \rangle - \langle Qb, Qx \rangle \\ &= \frac{1}{2} \langle Dy, y \rangle - \langle Qb, y \rangle = \frac{1}{2} \sum_{i=1}^n \lambda_i y_i^2 - \sum_{i=1}^n (Qb)_i y_i, \end{aligned} \quad (3.137)$$

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 с коэффициентами, равными собственным значениям матрицы A , минус линейные члены. Таким образом, график функционала энергии — это эллиптический параболоид, возможно, сдвинутый относительно начала координат и ещё повернутый, а 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 (3.137), т.е., согласно формуле (3.49), от спектрального числа обусловленности матрицы A . Чем больше эта обусловленность, тем сильнее сплюснены эллипсоиды уровня, так что для плохообусловленных СЛАУ решение находится на дне длинного и узкого «оврага».

3.106 Метод спуска для минимизации функций

В предшествующем пункте 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 $\{x^{(k)}\}$, 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*

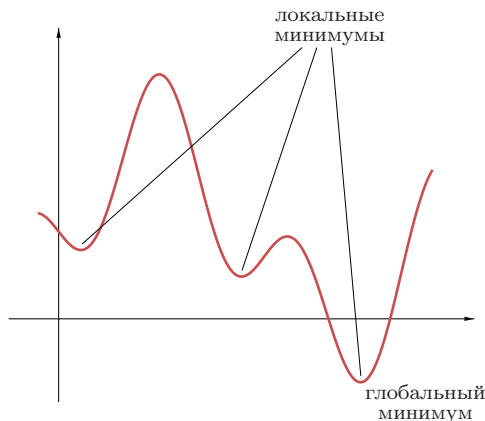


Рис. 3.26. Глобальные и локальные минимумы функции.

спуска, который заключается в следующем. Пусть уже найдено какое-то приближение $x^{(k)}$, $k = 0, 1, 2, \dots$, к точке минимума функции $f(x)$. Далее

- выбираем направление s , в котором целевая функция убывает в точке $x^{(k)}$,
- назначаем величину шага τ , на который сдвигаемся в выбранном направлении, полагая

$$x^{(k+1)} \leftarrow x^{(k)} + \tau s.$$

Конкретное значение τ находится из условия уменьшения целевой функции, т. е. так, чтобы $f(x^{(k+1)}) < f(x^{(k)})$.

Далее мы можем повторить этот шаг ещё раз и ещё ... столько, сколько нужно для достижения желаемого приближения к минимуму.

Выбор как направления спуска s , так и величины шага τ 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 в сторону решения.

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 $\Psi(x)$ достаточно хороши: он имеет один локальный минимум, который одновременно и глобален.

Пусть на очередном шаге метода спуска зафиксировано направление s . Определим шаг спуска по этому направлению, который будет обеспечивать наилучшую возможную минимизацию функционала энергии $\Psi(x)$. Далее мы будем называть подобный спуск *наискорейшим спуском*. Для этого подставим $x^{(k)} + \tau s$ в аргумент функционала энергии и продифференцируем получающееся выражение по τ . Имеем

$$\begin{aligned}\Psi(x^{(k)} + \tau s) &= \frac{1}{2} \langle A(x^{(k)} + \tau s), x^{(k)} + \tau s \rangle - \langle b, x^{(k)} + \tau s \rangle \\ &= \frac{1}{2} \langle Ax^{(k)}, x^{(k)} \rangle + \tau \langle Ax^{(k)}, s \rangle + \frac{1}{2} \tau_k^2 \langle As, s \rangle \\ &\quad - \langle b, x^{(k)} \rangle - \tau \langle b, s \rangle.\end{aligned}$$

При дифференцировании выписанного выражения по τ не зависящие от него члены исчезнут, и мы получим

$$\begin{aligned}\frac{d}{d\tau} \Psi(x^{(k)} + \tau s) &= \langle Ax^{(k)}, s \rangle + \tau \langle As, s \rangle - \langle b, s \rangle \\ &= \tau \langle As, s \rangle + \langle Ax^{(k)} - b, s \rangle \\ &= \tau \langle As, s \rangle + \langle r^{(k)}, s \rangle,\end{aligned}$$

где обозначено $r := Ax^{(k)} - b$ — невязка текущего приближения к решению. Таким образом, в точке экстремума по τ условие

$$\frac{d}{d\tau} \Psi(x^{(k)} + \tau s) = 0$$

необходимо влечёт

$$\tau = - \frac{\langle r, s \rangle}{\langle As, s \rangle}. \quad (3.138)$$

Легко видеть, что при найденном значении τ функционалом энергии действительно достигается минимум по выбранному направлению спуска. Это следует из того, что вторая производная по τ равна

$$\frac{d^2}{d\tau^2} \Psi(x^{(k)} + \tau s) = \langle As, s \rangle > 0,$$

в силу положительной определённости матрицы A .

3.10в Наискорейший градиентный спуск

Если целевая функция $f(x)$ дифференцируема, то, как известно, направление её наибольшего убывания в точке $x^{(k)}$ противоположно направлению вектора градиента

$$\nabla f(x) := f'(x^{(k)}) = (f'_1(x^{(k)}), f'_2(x^{(k)}), \dots, f'_n(x^{(k)}))^{\top}.$$

Это наблюдение даёт начало одному из распространённых вариантов метода спуска для минимизации функций — *методу градиентного спуска*, в котором направлением спуска берётся «антиградиент», т. е. вектор $-f'(x^{(k)})$, а очередной шаг выполняется по формуле

$$x^{(k+1)} \leftarrow x^{(k)} - \tau_k f'(x^{(k)}) \quad (3.139)$$

для некоторого $\tau_k \in \mathbb{R}$. Какой вид имеет градиентный спуск для минимизации функционала энергии $\Psi(x)$?

Вычислим градиент функционала энергии:

$$\frac{\partial \Psi(x)}{\partial x_l} = \frac{\partial}{\partial x_l} \left(\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j - \sum_{i=1}^n b_i x_i \right) = \sum_{j=1}^n a_{lj} x_j - b_l,$$

$l = 1, 2, \dots, n$. 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

$$\Psi'(x) = \left(\frac{\partial \Psi(x)}{\partial x_1}, \frac{\partial \Psi(x)}{\partial x_2}, \dots, \frac{\partial \Psi(x)}{\partial x_n} \right)^{\top} = Ax - b,$$

т. е. градиент функционала Ψ равен невязке решаемой системы линейных уравнений в рассматриваемой точке. Важнейшим выводом из этого факта является тот, что итерационный метод Ричардсона (3.134)

может быть представлен в виде

$$x^{(k+1)} \leftarrow x^{(k)} - \tau \Psi'(x^{(k)}), \quad k = 0, 1, 2, \dots,$$

т. е. он является не чем иным, как методом градиентного спуска (3.139) для минимизации функционала энергии Ψ , в котором шаг τ_k выбран постоянным и равным τ . В общем случае метод градиентного спуска (3.139) оказывается равносильным простейшему нестационарному итерационному методу (3.135).

Для метода градиентного спуска с постоянным шагом его трактовка как метода Ричардсона позволяет, опираясь на результат §3.9 об оптимизации скалярного предобуславливателя, выбрать шаг $\tau_k = \text{const}$, который наверняка обеспечивает сходимость процесса. Именно, если положительные числа μ и M — это нижняя и верхняя границы спектра положительно определённой матрицы A решаемой системы, то в соответствии с (3.120) для сходимости следует взять

$$\tau_k = \tau = \frac{2}{M + \mu}.$$

Другой способ выбора шага состоит в том, чтобы потребовать τ_k наибольшим возможным, обеспечивающим убывание функционала Ψ 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 τ_k в методе наискорейшего градиентного спуска воспользуемся результатом предшествующего раздела — формулой (3.138). Тогда величина шага должна быть

$$\tau_k = \frac{\langle r^{(k)}, r^{(k)} \rangle}{\langle Ar^{(k)}, r^{(k)} \rangle}.$$

В целом, псевдокод метода наискорейшего градиентного спуска для решения системы линейных алгебраических уравнений $Ax = b$ представлен в Табл. 3.8.

Теорема 3.10.2 *Если A — симметричная положительно определённая матрица, то последовательность $\{x^{(k)}\}$, порождаемая методом наискорейшего спуска, сходится к решению x^* системы уравнений $Ax = b$ из любого начального приближения $x^{(0)}$, и быстрота этой*

Таблица 3.8. Псевдокод метода наискорейшего спуска
для решения систем линейных уравнений

```

 $k \leftarrow 0$ ;
выбираем начальное приближение  $x^{(0)}$ ;
DO WHILE ( метод не сошёлся )
     $r^{(k)} \leftarrow Ax^{(k)} - b$ ;
     $\tau_k \leftarrow \frac{\|r^{(k)}\|_2^2}{\langle Ar^{(k)}, r^{(k)} \rangle}$ ;
     $x^{(k+1)} \leftarrow x^{(k)} - \tau_k r^{(k)}$ ;
     $k \leftarrow k + 1$ ;
END DO

```

сходимости оценивается неравенством

$$\|x^{(k)} - x^*\|_A \leq \left(\frac{M - \mu}{M + \mu} \right)^k \|x^{(0)} - x^*\|_A, \quad k = 0, 1, 2, \dots, \quad (3.140)$$

где μ , M — нижняя и верхняя границы спектра матрицы A .

Доказательство оценки (3.140) 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 -ый 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 τ

$$\Psi(x^{(k)}) \leq \Psi(\tilde{x}),$$

так как метод наискорейшего спуска обеспечивает наибольшее уменьшение функционала энергии на одном шаге итераций. Далее, из равенства (3.136)

$$\Psi(x) = \frac{1}{2}\|x - x^*\|_A^2 - \frac{1}{2}\|x^*\|_A^2$$

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.

$$\|x^{(k)} - x^*\|_A \leq \|\tilde{x} - x^*\|_A. \quad (3.141)$$

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 Рундсона (3.118) или (3.134) — имеем

$$\tilde{x} - x^* = (I - \tau A)(x^{(k)} - x^*), \quad k = 0, 1, 2, \dots$$

Матрица $(I - \tau A)$ является полиномом первой степени от матрицы A , и потому можем применить неравенство (3.30) из Предложения 3.3.8 (стр. 316):

$$\|\tilde{x} - x^*\|_A \leq \|I - \tau A\|_2 \|x^{(k)} - x^*\|_A.$$

При этом в силу (3.121) для метода наискорейшего спуска оценка погрешности заведомо не хуже этой оценки с произвольным значением τ . В частности, мы можем взять значение параметра $\tau = 2/(M + \mu)$, оптимальное для спуска с постоянным шагом. Тогда в соответствии с оценкой (3.121), выведенной при анализе скалярного предобуславливателя, получаем

$$\|x^{(k+1)} - x^*\|_A \leq \left(\frac{M - \mu}{M + \mu} \right) \|x^{(k)} - x^*\|_A, \quad k = 0, 1, 2, \dots,$$

откуда следует доказываемое неравенство (3.140). ■

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» — движение в сторону решения. Шаг в методе наискорейшего спуска идёт

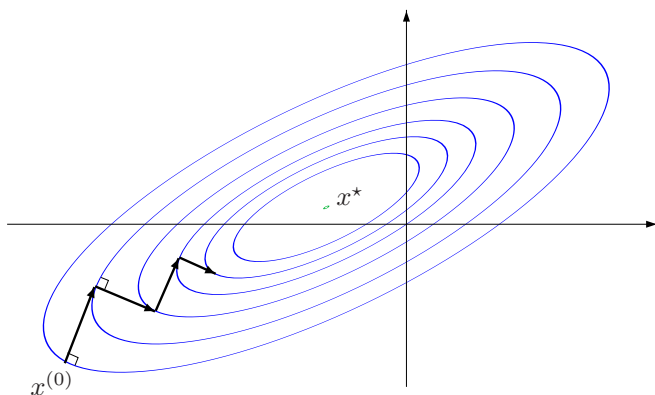


Рис. 3.27. Иллюстрация работы метода наискорейшего спуска.

на максимально возможную величину — до пересечения с касательным эллипсоидом. Поэтому траектория метода наискорейшего спуска является ломаной, звенья которой перпендикулярны друг другу (см. Рис. 3.27).

Хотя доказательство Теоремы 3.10.2 основано на мажоризации наискорейшего спуска итерационным методом Ричардсона и может показаться довольно грубым, оценка (3.140) 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.

3.10г Метод минимальных невязок

Пусть, как и ранее, дана система линейных алгебраических уравнений $Ax = b$ с симметричной положительно определённой матрицей A . Для нестационарного итерационного процесса (3.135)

$$x^{(k+1)} \leftarrow x^{(k)} - \tau_k (Ax^{(k)} - b), \quad k = 0, 1, 2, \dots,$$

ещё один популярный подход к выбору итерационных параметров τ_k был предложен С.Г. Крейном и М.А. Красносельским в работе [25] и назван ими *методом минимальных невязок*. Его псевдокод приведён в Табл. 3.9.

Каждый шаг этого метода минимизирует в направлении невязки k -го приближения, равной $r^{(k)} = Ax^{(k)} - b$, не функционал энергии $\Psi(x)$ из Теоремы 3.10.1, 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

$$\begin{aligned}\|Ax - b\|_2^2 &= \langle Ax - b, Ax - b \rangle = \langle Ax - Ax^*, Ax - Ax^* \rangle \\ &= \langle A(x - x^*), A(x - x^*) \rangle = \langle A^\top A(x - x^*), x - x^* \rangle \\ &= \|x - x^*\|_{A^\top A}^2.\end{aligned}\tag{3.142}$$

Таблица 3.9. Псевдокод метода минимальных невязок для решения систем линейных уравнений

```

 $k \leftarrow 0$ ;
выбираем начальное приближение  $x^{(0)}$ ;
DO WHILE ( метод не сошёлся )
     $r^{(k)} \leftarrow Ax^{(k)} - b$ ;
     $\tau_k \leftarrow \frac{\langle Ar^{(k)}, r^{(k)} \rangle}{\|Ar^{(k)}\|_2^2}$ ;
     $x^{(k+1)} \leftarrow x^{(k)} - \tau_k r^{(k)}$ ;
     $k \leftarrow k + 1$ ;
END DO

```

Если уже найдено $x^{(k)}$, и мы желаем выбрать параметр τ так, чтобы на очередном шаге итераций вектор $x^{(k)} - \tau r^{(k)}$ минимизировал 2-норму

невязки решения, то необходимо найти минимум по τ для выражения

$$\begin{aligned} \|A(x^{(k)} - \tau r^{(k)}) - b\|_2^2 &= \langle A(x^{(k)} - \tau r^{(k)}) - b, A(x^{(k)} - \tau r^{(k)}) - b \rangle \\ &= \tau^2 \langle Ar^{(k)}, Ar^{(k)} \rangle - 2\tau (\langle Ax^{(k)}, Ar^{(k)} \rangle - \langle b, Ar^{(k)} \rangle) \\ &\quad + \langle Ax^{(k)}, Ax^{(k)} \rangle + \langle b, b \rangle. \end{aligned}$$

Дифференцируя его по τ и приравнявая производную нулю, получим

$$2\tau \langle Ar^{(k)}, Ar^{(k)} \rangle - 2(\langle Ax^{(k)}, Ar^{(k)} \rangle - \langle b, Ar^{(k)} \rangle) = 0,$$

что с учётом равенства $Ax^{(k)} - b = r^{(k)}$ даёт

$$\tau \langle Ar^{(k)}, Ar^{(k)} \rangle - \langle r^{(k)}, Ar^{(k)} \rangle = 0.$$

Окончательно

$$\tau = \frac{\langle Ar^{(k)}, r^{(k)} \rangle}{\langle Ar^{(k)}, Ar^{(k)} \rangle} = \frac{\langle Ar^{(k)}, r^{(k)} \rangle}{\|Ar^{(k)}\|_2^2}.$$

Теорема 3.10.3 Если A — симметричная положительно определённая матрица, то последовательность $\{x^{(k)}\}$, порождаемая методом минимальных невязок, сходится к решению x^* системы уравнений $Ax = b$ из любого начального приближения $x^{(0)}$, и быстрота этой сходимости оценивается неравенством

$$\|x^{(k)} - x^*\|_{A^\top A} \leq \left(1 - \left(\frac{\mu}{M}\right)^2\right)^{k/2} \|x^{(0)} - x^*\|_{A^\top A}, \quad (3.143)$$

$k = 0, 1, 2, \dots$, где μ, M — нижняя и верхняя границы спектра матрицы A .

Доказательство теоремы можно найти, к примеру, в [25, 39], где для невязок $r^{(k)} = Ax^{(k)} - b$ доказывается оценка

$$\|r^{(k+1)}\|_2 \leq \sqrt{1 - \left(\frac{\mu}{M}\right)^2} \|r^{(k)}\|_2, \quad k = 0, 1, 2, \dots$$

С учётом выкладок (3.142) этот результат совершенно равносильен неравенству (3.143).

Полезным свойством метода минимальных невязок является монотонное убывание евклидовой нормы погрешности приближений на каждом шаге:

$$\|x^{(k+1)} - x^*\|_2 \leq \|x^{(k)} - x^*\|_2, \quad k = 0, 1, 2, \dots$$

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 тоже сходится. Но если матрица системы не является положительно определённой, метод может не сходиться к решению.

Пример 3.10.1 В системе линейных алгебраических уравнений

$$\begin{pmatrix} 2 & 2 \\ 1 & 0 \end{pmatrix} x = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

матрица не является ни симметричной, ни положительно определённой (её собственные значения приблизительно равны 2.732 и -0.7321). В применении к этой системе метод минимальных невязок с нулевым начальным приближением вскоре после начала работы устанавливается на векторе $(0.7530088, 0.2469912)^\top$, тогда как настоящее решение — это вектор $(1, 0)^\top$. 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, описываемую Теоремой 3.10.3.

Если сходимость методов наискорейшего спуска и минимальных невязок принципиально не лучше сходимости простейшего итерационного метода Рундсона, то имеют ли они какое-либо практическое значение? Ответ на этот вопрос положителен. Вспомним, что наша оптимизация метода Рундсона 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 «подстраиваются» под решаемую систему уравнений.

Метод минимальных невязок в представленной выше 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 — Generalized Minimal RESiduals — обобщённый метод минимальных невязок. Она была предложена Ю. Саадом [39] (см. также [46, 61]) и предназначена для решения разреженных систем линейных алгебраических уравнений большой размерности, возникающих при дискретизации уравнений в частных производных.

3.10д 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, ортогональных 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

$$x = x^{(0)} + \sum_{i=1}^n c_i s^{(i)},$$

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 , $i = 1, 2, \dots, n$.

Наиболее популярными представителями методов сопряжённых направлений являются *методы сопряжённых градиентов*, предложенные М.Р. Хестенсом и Э.Л. Штифелем в начале 50-х годов прошлого века. Как и почти любой содержательный численный метод, их можно представить, в зависимости от целей исследования и применения, несколькими различными способами. Нам будет удобно вывести ме-

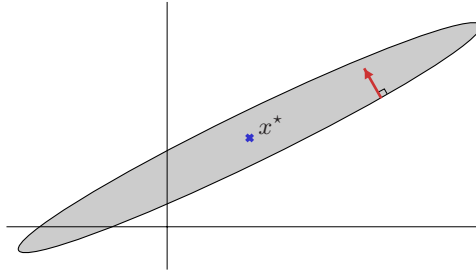


Рис. 3.28. Направление градиента функционала может быть плохим для спуска к его минимуму.

тоды сопряжённых градиентов как методы наискорейшего спуска для минимизации функционала энергии, порождённого системой линейных уравнений (см. §3.106), в которых направление спуска выбирается специальным и чрезвычайно удачным образом.

Мы видели в предшествующем параграфе, что направления антиградиентов функционала энергии, по которым осуществляется движение к решению в методе наискорейшего градиентного спуска, не очень удачны (см. Рис. 3.28), а шаги спуска могут сильно «вихлять» от шага к шагу. В целом траектория спуска к решению весьма нерациональна, и для нахождения решения затрачивается много лишней работы. Естественно попытаться сделать так, чтобы алгоритм шёл 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

- (i) $-\nabla\Psi(x^{(k-1)})$, т. е. антиградиента функционала энергии Ψ в точке $x^{(k-1)}$, который противоположен вектору невязки $r^{(k-1)} = Ax^{(k-1)} - b$,
- (ii) 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 мы виде-

ли в разделе 3.10в, не вполне удовлетворительно, и имеет смысл «смешать» его с дополнительной информацией о других возможных направлениях спуска. В качестве такового 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

$$\langle As^{(k)}, s^{(k-1)} \rangle = \langle s^{(k)}, As^{(k-1)} \rangle = 0, \quad k = 1, 2, \dots, n. \quad (3.144)$$

Последовательные направления спуска в методе наискорейшего градиентного спуска ортогональны друг другу. Условие (3.144) тоже означает ортогональность, но в скалярном произведении, порождённом матрицей решаемой системы уравнений. Его можно рассматривать как попытку скорректировать направления спуска так, чтобы они стали соответствовать той геометрии пространства, которая порождается решаемой системой. В целом, этим условием обеспечивается то, что направления спуска к минимуму «не слишком близки» друг к другу, т.е. некоторая защита от зигзагообразности траектории спуска к решению.

Имеем

$$\langle -r^{(k-1)} + v_k s^{(k-1)}, As^{(k-1)} \rangle = 0,$$

так что

$$-\langle r^{(k-1)}, As^{(k-1)} \rangle + v_k \langle s^{(k-1)}, As^{(k-1)} \rangle = 0.$$

Окончательно,

$$v_k = \frac{\langle r^{(k-1)}, As^{(k-1)} \rangle}{\langle s^{(k-1)}, As^{(k-1)} \rangle}, \quad (3.145)$$

Вычислительная схема метода сопряжённых градиентов

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

— такая же, как и у всех методов спуска, а величина шага τ_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 §3.10б и

можем воспользоваться полученной там формулой (3.138), которая в нашем случае даёт

$$\tau_k = -\frac{\langle r^{(k)}, s^{(k)} \rangle}{\langle As^{(k)}, s^{(k)} \rangle}. \quad (3.146)$$

Итак, метод сопряжённых градиентов можно представить следующей схемой.

Шаг 1. Выбираем начальное приближение $x^{(0)}$ и в качестве направления спуска на первом шаге берём

$$s^{(1)} = -\nabla\Psi(x^{(0)}) = -r^{(0)} = -(Ax^{(0)} - b)$$

— направление антиградиента функционала энергии в точке $x^{(0)}$. 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 (3.146), т. е. как

$$x^{(1)} \leftarrow x^{(0)} + \tau_1 s^{(1)},$$

где выбор

$$\tau_1 = -\frac{\langle r^{(0)}, s^{(1)} \rangle}{\langle As^{(1)}, s^{(1)} \rangle} = \frac{\langle r^{(0)}, r^{(0)} \rangle}{\langle Ar^{(0)}, r^{(0)} \rangle},$$

минимизирует функционал энергии на этом шаге.

Шаг k , $k = 2, 3, \dots$ Пусть уже известно приближение $x^{(k-1)}$ и направление спуска $s^{(k-1)}$, по которому оно было получено.

Выбираем следующим направлением спуска вектор

$$s^{(k)} = -r^{(k-1)} + v_k s^{(k-1)}, \quad (3.147)$$

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, т. е. в соответствии с формулой (3.145):

$$v_k = \frac{\langle As^{(k-1)}, r^{(k-1)} \rangle}{\langle As^{(k-1)}, s^{(k-1)} \rangle}. \quad (3.148)$$

Вычисляем 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, невязка на k -ом шаге метода сопряжённых градиентов (равная градиенту функционала энергии) изменяется следующим образом

$$\begin{aligned} r^{(k)} &= Ax^{(k)} - b \\ &= A(x^{(k-1)} + \tau_k s^{(k)}) - b \\ &= r^{(k-1)} + \tau_k As^{(k)}. \end{aligned} \tag{3.149}$$

Теперь исследуем построенный метод.

Предложение 3.10.1 *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 τ_1 .

Далее доказательство продолжается индукцией по k .

Рассмотрим k -ый шаг метода сопряжённых градиентов. Из формулы (3.149) 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 (3.146)

$$\tau_k = -\frac{\langle r^{(k-1)}, s^{(k)} \rangle}{\langle As^{(k)}, s^{(k)} \rangle}.$$

Поэтому на основе (3.147) и (3.149) 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 (3.144).

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 (3.147), 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

$$\langle r^{(k-1)}, r^{(k)} \rangle = 0.$$

Это завершает доказательство Предложения. ■

На самом деле, справедлив более общий результат

Теорема 3.10.4 В методе сопряжённых градиентов векторы направлений спуска $s^{(k)}$, $k = 1, 2, \dots$, являются A -ортгональными друг другу, т. е.

$$\langle s^{(i)}, s^{(j)} \rangle_A = \langle As^{(i)}, s^{(j)} \rangle = 0, \quad i \neq j.$$

В методе сопряжённых градиентов векторы невязок $r^{(k)} = Ax^{(k)} - b$, $k = 0, 1, 2, \dots$, ортгональны друг другу, т. е.

$$\langle r^{(i)}, r^{(j)} \rangle = 0, \quad i \neq j.$$

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.

Для удобства и без большого ограничения общности условимся считать, что $s^{(0)} = 0$, т. е. что направление спуска на «нулевом шаге», который виртуально привёл к нулевому приближению, — это нулевой вектор.

Доказательство 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 Предложения. Кроме того, A -ортогональны $s^{(0)}$ и $s^{(1)}$.

Предположив, что утверждение Теоремы справедливо для шага метода с номером k , покажем, что оно верно также для шага $k+1$. Иными словами, нам необходимо доказать

$$\begin{aligned}\langle s^{(k+1)}, s^{(j)} \rangle_A &= 0 \quad \text{для } j = 1, 2, \dots, k, \\ \langle r^{(k+1)}, r^{(j)} \rangle &= 0 \quad \text{для } j = 0, 1, 2, \dots, k.\end{aligned}$$

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 (3.147)

$$s^{(k+1)} = -r^{(k)} + v_{k+1}s^{(k)},$$

то

$$\langle s^{(k+1)}, s^{(j)} \rangle_A = \langle s^{(k+1)}, As^{(j)} \rangle = -\langle r^{(k)}, As^{(j)} \rangle + v_{k+1}\langle s^{(k)}, As^{(j)} \rangle,$$

где последнее слагаемое зануляется при $j < k$ в силу индукционного предположения. Итак,

$$\langle s^{(k+1)}, s^{(j)} \rangle_A = -\langle r^{(k)}, As^{(j)} \rangle. \quad (3.150)$$

С другой стороны, согласно (3.149) в методе сопряжённых градиентов невязка изменяется как

$$r^{(j)} = r^{(j-1)} + \tau_j A s^{(j)},$$

откуда

$$A s^{(j)} = \frac{1}{\tau_j} (r^{(j)} - r^{(j-1)}).$$

Подставив в правую часть равенства (3.150), будем иметь

$$\begin{aligned} \langle s^{(k+1)}, s^{(j)} \rangle_A &= -\langle r^{(k)}, A s^{(j)} \rangle \\ &= -\frac{1}{\tau_j} \left(\langle r^{(k)}, r^{(j)} \rangle - \langle r^{(k)}, r^{(j-1)} \rangle \right) = 0, \end{aligned}$$

так как оба слагаемых в больших скобках раны нулю в силу индукционного предположения. Это доказывает первое утверждение Теоремы — A -ортогональность направлений спуска.

Установим теперь ортогональность невязок $r^{(k)}$, $k = 0, 1, 2, \dots$, т. е.

$$\langle r^{(k+1)}, r^{(j)} \rangle = 0 \quad \text{для } j = 0, 1, 2, \dots, k.$$

Мы уже знаем, что

$$\langle r^{(k+1)}, r^{(k)} \rangle = 0$$

в силу доказанного ранее Предложения. Остаётся показать, что для $0 \leq j < k$ также имеем $\langle r^{(k+1)}, r^{(j)} \rangle = 0$.

В силу формулы (3.149)

$$r^{(k+1)} = r^{(k)} + \tau_{k+1} A s^{(k+1)},$$

и потому по индукционному предположению

$$\begin{aligned} \langle r^{(k+1)}, r^{(j)} \rangle &= \langle r^{(k)}, r^{(j)} \rangle + \tau_{k+1} \langle A s^{(k+1)}, r^{(j)} \rangle \\ &= \tau_{k+1} \langle A s^{(k+1)}, r^{(j)} \rangle. \end{aligned} \tag{3.151}$$

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 через эти направления можно выразить невязку $r^{(j)}$:

$$r^{(j)} = -s^{(j+1)} + v_{j+1}s^{(j)}.$$

Подставляя результат в выражение (3.151), 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.

Наконец, необходимо отдельно рассмотреть случай $j = 0$, для которого нельзя выписать использованное выше представление невязки $r^{(j)} = r^{(j-1)} + \tau_j As^{(j)}$. Тогда

$$r^{(0)} = -s^{(1)},$$

так что в самом деле

$$\langle r^{(k+1)}, r^{(0)} \rangle = -\tau_{k+1} \langle As^{(k+1)}, r^{(0)} \rangle = \tau_{k+1} \langle As^{(k+1)}, s^{(1)} \rangle = 0.$$

Это завершает доказательство Теоремы. ■

Предложение 3.10.2 *Метод сопряжённых градиентов для решения $n \times n$ -системы линейных уравнений с симметричной положительно-определённой матрицей находит точное решение не более чем за n шагов.*

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 и потому линейно независимы. Таким образом, ненулевых невязок может быть не более n штук. ■

На практике из-за неизбежных погрешностей вычислений метод сопряжённых градиентов может не прийти к точному решению системы

ровно за n шагов. Тогда целесообразно повторить цикл уточнения, превратив алгоритм при необходимости в итерационный. При этом

$$x = x^{(0)} + \sum_i c_i s^{(i)},$$

где $x^{(0)}$ — начальное приближение,
 $s^{(i)}$, $i = 1, 2, \dots, n$, — векторы направлений спуска,
 которые являются также векторами разложения
 решения,
 c_i — коэффициенты разложения решения, равные
 шагам спуска в соответствующих направлениях.

Верхний предел суммы может значительно превосходить n .

Теорема 3.10.5 Если A — симметричная положительно определённая матрица, 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

$$\|x^{(k)} - x^*\|_A \leq 2 \left(\frac{\sqrt{M} - \sqrt{\mu}}{\sqrt{M} + \sqrt{\mu}} \right)^k \|x^{(0)} - x^*\|_A, \quad (3.152)$$

$k = 0, 1, 2, \dots$, где μ , M — нижняя и верхняя границы спектра матрицы A .

Доказательство опускается. Его можно найти, к примеру, в [1, 44, 46, 39, 61]. Имеет смысл отметить, что оценка (3.152) 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.

В Табл. 3.10 представлен оптимизированный псевдокод метода сопряжённых градиентов. 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. 3.10 вычисляет длину 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

Таблица 3.10. Псевдокод метода сопряжённых градиентов
для решения систем линейных уравнений

```

 $k \leftarrow 0$ ;
выбираем начальное приближение  $x^{(0)}$ ;
 $r^{(0)} \leftarrow b - Ax^{(0)}$ ;       $s^{(0)} \leftarrow r^{(0)}$ ;
DO WHILE ( метод не сошёлся )
     $g \leftarrow As^{(k)}$ ;
     $\tau_k \leftarrow \frac{\langle r^{(k)}, r^{(k)} \rangle}{\langle s^{(k)}, g \rangle}$ ;
     $x^{(k+1)} \leftarrow x^{(k)} + \tau_k s^{(k)}$ ;
     $r^{(k+1)} \leftarrow r^{(k)} - \tau_k g$ ;
     $v_k \leftarrow \frac{\langle r^{(k+1)}, r^{(k+1)} \rangle}{\langle r^{(k)}, r^{(k)} \rangle}$ ;
     $s^{(k+1)} \leftarrow r^{(k+1)} + v_k s^{(k)}$ ;
     $k \leftarrow k + 1$ ;
END DO

```

математического итэлика, которым в тексте книги обозначается противоположная ему невязка $r^{(k)}$. Направление антиградиента функционала энергии в точке вновь найденного приближённого решения корректируется в четвёртой строке тела цикла. В следующих двух строках (перед увеличением счётчика k) вычисляется новое направление $s^{(k+1)}$ движения к решению. Кроме того, для удобства программирования нумерация переменных несколько изменена в сравнении с описанием метода, которое было дано ранее.

Для метода сопряжённых градиентов широко распространена также другая трактовка, представляющая его как метод построения A -ортогонального базиса пространства и одновременного нахождения коэффициентов разложения решения по нему.

Пусть требуется найти решение системы линейных алгебраических

уравнений

$$Ax = b$$

с симметричной и положительно определённой матрицей A . Такая матрица порождает, как мы видели, «энергетическое» скалярное произведение (см. §3.3е), или скалярное A -произведение векторов

$$\langle x, y \rangle_A := \langle Ax, y \rangle.$$

Соответственно, имеет смысл понятие A -ортогональности относительно этого нового скалярного произведения. Далее, пусть $s^{(1)}, s^{(2)}, \dots, s^{(n)}$ — базис \mathbb{R}^n , составленный из A -ортогональных векторов. Решение x^* системы уравнений можно искать в виде разложения по этому базису, т. е.

$$x^* = \sum_{i=1}^n x_i s^{(i)} \quad (3.153)$$

с какими-то неизвестными коэффициентами x_i , $i = 1, 2, \dots, n$. Умножая обе части этого равенства 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

$$\sum_{i=1}^n x_i \langle As^{(i)}, s^{(j)} \rangle = \langle b, s^{(j)} \rangle, \quad j = 1, 2, \dots, n. \quad (3.154)$$

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 (3.154) останется лишь

$$x_i \langle As^{(i)}, s^{(i)} \rangle = \langle b, s^{(i)} \rangle, \quad i = 1, 2, \dots, n.$$

Окончательно

$$x_i = \frac{\langle b, s^{(i)} \rangle}{\langle As^{(i)}, s^{(i)} \rangle}, \quad i = 1, 2, \dots, n,$$

откуда из (3.153) 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 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 Ланцоша (см. §3.7ж). Этот процесс ортогонализации конечен и завершается при некотором $k \leq n$, для которого $r^{(k)} = 0$, т.е. когда очередная невязка приближённого решения зануляется. Нетрудно понять, что в нашей 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.

Пример 3.10.2 Рассмотрим решение методом сопряжённых градиентов 8×8 -системы линейных алгебраических уравнений с гильбертовой матрицей. Её число обусловленности равно $\text{cond}_2(A) = 1.5 \cdot 10^{10}$. В качестве правой части возьмём вектор

$$(1, -1, 1, -1, 1, -1, 1, -1)^\top.$$

Метод сопряжённых градиентов справляется с этой плохообусловленной системой. При тех же требованиях на ответ, которые мы рассматривали в Примере 3.9.1, он выдаёт решение за 133 шага-итерации.

Метод Гаусса-Зейделя за миллион итераций так и не смог найти ничего более-менее похожего на решение ■

Пример 3.10.3 Рассмотрим теперь решение методом сопряжённых градиентов 10×10 -системы линейных алгебраических уравнений с гильбертовой матрицей. В качестве правой части возьмём вектор

$$(1, -1, 1, -1, 1, -1, 1, -1, 1, -1)^\top.$$

Строго говоря, так как при вводе чисел в компьютер и дальнейших операциях над ними допускаются неизбежные погрешности, то матрица системы не вполне совпадает с реальной гильбертовой матрицей, а является лишь «приближённо гильбертовой». Число обусловленности этой матрицы равно $\text{cond}_2(A) = 1.6 \cdot 10^{13}$.

Метод сопряжённых градиентов справляется с этой плохообусловленной системой. При тех же требованиях на ответ, которые мы рассматривали в Примере 3.9.1, он выдаёт решение за 133 шага-итерации.

Метод Гаусса-Зейделя за 10 миллионов итераций так и не смог найти ничего более-менее похожего на решение ■

3.11 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

$$\frac{\partial x(t)}{\partial t} + Ax(t) = b, \quad (3.155)$$

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 $\partial x / \partial t$ зануляется и соответствующие значения $x(t)$ являются решением исходной задачи

Наиболее часто задачу (3.155) 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 решения задачи (3.155) 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 (3.155) можно применять самые простые численные методы. Таким, к примеру, является явный

метод Эйлера (метод ломаных) с постоянным временным шагом τ , в котором производная заменяется на разделённую разность вперёд (см. [1, 4, 5, 26, 33, 40, 82]). Обозначая $x^{(k)} := x(t_k)$, $t_k = t_0 + \tau k$, $k = 0, 1, 2, \dots$, получим вместо (3.155)

$$\frac{x^{(k+1)} - x^{(k)}}{\tau} + Ax^{(k)} = b, \quad (3.156)$$

или

$$x^{(k+1)} = x^{(k)} - \tau(Ax^{(k)} - b), \quad k = 0, 1, 2, \dots,$$

то есть известный нам итерационный метод Рундсона (3.118) для решения системы уравнений $Ax = b$. При переменном шаге по времени, когда $\tau = \tau_k$, $k = 0, 1, 2, \dots$, получающийся метод Эйлера

$$\frac{x^{(k+1)} - x^{(k)}}{\tau_k} + Ax^{(k)} = b$$

эквивалентен

$$x^{(k+1)} = x^{(k)} - \tau_k(Ax^{(k)} - b), \quad k = 0, 1, 2, \dots,$$

т. е. простейшему нестационарному итерационному методу Рундсона (3.135).

Представление итерационного метода Рундсона в виде (3.156), как численного метода решения системы дифференциальных уравнений, даёт возможность понять суть ограничения на параметр τ . Это не что иное, как ограничение на величину шага по времени, вызванное требованием устойчивости метода. С другой стороны, если шаг по времени взят недостаточно большим, то до установления решения задачи (3.155) нам нужно сделать очень много таких мелких шагов, что даёт ещё одно объяснение невысокой вычислительной эффективности итераций Рундсона.

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*, идейно близкие друг другу (см. [99]).

Очевидно, что вместо (3.155) можно рассмотреть задачу более общего вида

$$B \frac{\partial x}{\partial t} + Ax(t) = b, \quad (3.157)$$

где B — некоторая неособенная матрица. Смысл её введения станет более понятен, если переписать (3.157) 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 преобуславливающей матрицы для системы $Ax = b$ (см. §3.9в).

Отметим в заключение темы, что для решения систем линейных алгебраических уравнений, возникающих при дискретизации уравнений в частных производных эллиптического типа, предельно эффективными являются *многосеточные методы*, предложенные Р.П. Федоренко в начале 60-х годов XX века.

3.12 Теория А.А. Самарского

Мы уже отмечали, что системы линейных алгебраических уравнений, которые необходимо решать на практике, часто бывают заданы неявно, в операторном виде. При этом мы не можем оперировать итерационными формулами вида (3.111) с явно заданным оператором T_k (наподобие (3.112)). Для подобных случаев А.А. Самарским была предложена специальная каноническая форма одношагового линейного итерационного процесса, предназначенного для решения систем линейных уравнений $Ax = b$:

$$B_k \frac{x^{(k+1)} - x^{(k)}}{\tau_k} + Ax^{(k)} = b, \quad k = 0, 1, 2, \dots, \quad (3.158)$$

где B_k , τ_k — некоторые последовательности матриц и скалярных параметров соответственно, причём $\tau_k > 0$. Мы будем называть её *канонической формой Самарского*. Если $x^{(k)}$ сходится к пределу, то при некоторых необременительных условиях на B_k и τ_k этот предел является решением системы линейных алгебраических уравнений $Ax = b$.

С учётом результатов предыдущего раздела нетрудно видеть, что форма Самарского навеяна представлением итерационных методов как процессов установления для решения систем уравнений.

Различные последовательности матриц B_k и итерационных параметров τ_k задают различные итерационные методы. Выбирая начальное значение $x^{(0)}$, находим затем из (3.158) последовательные приближения как решения уравнений

$$B_k x^{(k+1)} = (B_k - \tau_k A) x^{(k)} + \tau_k b, \quad k = 0, 1, 2, \dots$$

Ясно, что для однозначной разрешимости этой системы уравнений относительно $x^{(k+1)}$ необходимо, чтобы все матрицы B_k были неособенными. Итерационный метод в форме (3.158) естественно назвать *явным*, если $B_k = I$ — единичная матрица и выписанная выше система сводится к явной формуле для нахождения следующего итерационного приближения $x^{(k+1)}$. Иначе, если $B_k \neq I$, итерации (3.158) называются *неявными*. Неявные итерационные методы имеет смысл применять лишь в том случае, когда решение системы уравнений относительно $x^{(k+1)}$ существенно легче, чем решение исходной системы.

Выпишем представление в форме Самарского для рассмотренных ранее итерационных процессов. Итерационный метод Ричардсона из §3.9г принимает вид

$$\frac{x^{(k+1)} - x^{(k)}}{\tau} + Ax^{(k)} = b, \quad k = 0, 1, 2, \dots, \quad (3.159)$$

где $\tau = \tau_k = \text{const}$ — постоянный параметр, имеющий тот же смысл, что и в рассмотренных §3.9г. Переменный параметр τ_k в (3.159) приводит к нестационарному методу Ричардсона (3.135) (см. §3.10а). Если D и \tilde{L} — диагональная и строго нижняя треугольная части матрицы A соответственно (см. §3.9д), то методы Якоби и Гаусса-Зейделя можно записать в виде

$$D \frac{x^{(k+1)} - x^{(k)}}{1} + Ax^{(k)} = b,$$

и

$$(D + \tilde{L}) \frac{x^{(k+1)} - x^{(k)}}{1} + Ax^{(k)} = b.$$

Наконец, итерационный метод релаксации с релаксационным параметром ω (см. §3.9ж) в тех же обозначениях имеет форму Самарского

$$(D + \omega \tilde{L}) \frac{x^{(k+1)} - x^{(k)}}{\omega} + Ax^{(k)} = b, \quad k = 0, 1, 2, \dots$$

При исследовании сходимости итераций в форме Самарского удобно пользоваться матричными неравенствами, связанными со знакоопределённостью матриц. Условимся для вещественной $n \times n$ -матрицы G обозначать

$$G \triangleright 0, \quad \text{если } \langle Gx, x \rangle > 0 \quad \text{для всех ненулевых } n\text{-векторов } x,$$

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 (3.158) даёт

Теорема 3.12.1 (теорема Самарского) *Если A — симметричная положительно определённая матрица, $\tau > 0$ и $B \triangleright \frac{1}{2} \tau A$, то стационарный итерационный процесс*

$$B \frac{x^{(k+1)} - x^{(k)}}{\tau} + Ax^{(k)} = b, \quad k = 0, 1, 2, \dots,$$

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

$$B \frac{z^{(k+1)} - z^{(k)}}{\tau} + Az^{(k)} = 0, \quad k = 0, 1, 2, \dots \quad (3.160)$$

Исследуем поведение энергетической нормы погрешности. Покажем сначала, что в условиях теоремы числовая последовательность $\|z^{(n)}\|_A = \langle Az^{(n)}, z^{(n)} \rangle$ является невозрастающей.

Из соотношения (3.160) следует

$$z^{(k+1)} = (I - \tau B^{-1}A) z^{(k)}, \quad (3.161)$$

и

$$Az^{(k+1)} = (A - \tau AB^{-1}A)z^{(k)}.$$

Таким образом,

$$\begin{aligned} \langle Az^{(k+1)}, z^{(k+1)} \rangle &= \langle Az^{(k)}, z^{(k)} \rangle - \tau \langle AB^{-1}Az^{(k)}, z^{(k)} \rangle \\ &\quad - \tau \langle Az^{(k)}, B^{-1}Az^{(k)} \rangle + \tau^2 \langle AB^{-1}Az^{(k)}, AB^{-1}Az^{(k)} \rangle. \end{aligned}$$

Коль скоро матрица A симметрична,

$$\langle AB^{-1}Az^{(k)}, z^{(k)} \rangle = \langle Az^{(k)}, B^{-1}Az^{(k)} \rangle,$$

и потому

$$\begin{aligned} \langle Az^{(k+1)}, z^{(k+1)} \rangle &= \\ &= \langle Az^{(k)}, z^{(k)} \rangle - 2\tau \langle (B - \tfrac{1}{2}\tau A)B^{-1}Az^{(k)}, B^{-1}Az^{(k)} \rangle. \end{aligned} \quad (3.162)$$

Учитывая неравенство $B \succ \frac{1}{2}\tau A$, можем заключить, что вычитаемое в правой части полученного равенства всегда неотрицательно. По этой причине

$$\|z^{(k+1)}\|_A \leq \|z^{(k)}\|_A,$$

так что последовательность $\|z^{(k)}\|_A$ монотонно не возрастает и ограничена снизу нулём. В силу известной теоремы Вейерштрасса она имеет предел при $k \rightarrow \infty$.

Неравенство $B \succ \frac{1}{2}\tau A$, т. е. положительная определённости матрицы $(B - \frac{1}{2}\tau A)$, означает существование такого $\eta > 0$, что для любых $y \in \mathbb{R}^n$

$$\langle (B - \tfrac{1}{2}\tau A)y, y \rangle \geq \eta \langle y, y \rangle = \eta \|y\|_2^2.$$

Как итог, из (3.162) получаем

$$\|z^{(k+1)}\|_A^2 - \|z^{(k)}\|_A^2 + 2\eta\tau \|B^{-1}Az^{(k)}\|_2^2 \leq 0$$

для всех $k = 0, 1, 2, \dots$. Переходя в этом неравенстве к пределу по $k \rightarrow \infty$, мы видим, что при этом 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 3.9.5) о сходимости метода релаксации для СЛАУ с

симметричными положительно определёнными матрицами, а также, как её частный случай, Теорема 3.9.4 о сходимости метода Гаусса-Зейделя. В самом деле, пусть $A = \tilde{L} + D + \tilde{U}$ в обозначениях §3.9д, 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 $0 < \omega < 2$.

Дальнейшие результаты в этом направлении читатель может увидеть, к примеру, в [40, 93].

3.13 Вычисление определителей матриц и обратных матриц

Предположим, что для матрицы A выполняется LU-разложение. Как отмечалось, выполняемые в представленной нами версии метода Гаусса преобразования — линейное комбинирование строк — 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 .

Другая возможная трактовка этого результата состоит в том, что если $A = LU$ — треугольное разложение матрицы A , то, как известно из линейной алгебры,

$$\det A = \det L \cdot \det U.$$

Если разложение матрицы A выполнено так, что по диагонали в нижней треугольной матрице L стоят все единицы, то $\det L = 1$. Следовательно, как и ранее, $\det A = \det U$, а точнее — произведению всех диагональных элементов в верхней треугольной матрице U .

Совершенно аналогичные технологии можно организовать при использовании других матричных разложений. Пусть, например, нам удалось получить 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 . Наконец, искомым определитель $\det A$ вычисляется по R как произведение её диагональных элементов и ещё $\det Q$.

Рассмотрим теперь вычисление матрицы, обратной к данной матрице. Отметим, прежде всего, что в современных вычислительных технологиях это приходится делать не слишком часто. Один из примеров, когда подобное вычисление необходимо по существу, — нахождение дифференциала операции обращения матрицы $A \mapsto A^{-1}$, равного

$$d(A^{-1}) = -A^{-1}(dA)A^{-1}$$

(см., к примеру, [14]). 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 системы, см. §1.6) даются формулами

$$\frac{\partial x_\nu}{\partial a_{ij}} = -z_{\nu i}x_j, \quad \frac{\partial x_\nu}{\partial b_i} = z_{\nu i}, \quad \nu = 1, 2, \dots, n,$$

где $Z = (z_{ij}) = A^{-1}$ — обратная к матрице A .

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 заведомо лучше, чем вычисление $A^{-1}b$ через нахождение обратной A^{-1} , как по точности, так и по трудоёмкости.

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 способ нахождения обратной матрицы: нужно решить n штук систем линейных уравнений

$$Ax = e^{(j)}, \quad j = 1, 2, \dots, n, \quad (3.163)$$

где $e^{(j)}$ — j -ый столбец единичной матрицы I . Это можно сделать, к примеру, любым из рассмотренных нами 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 (3.77) (или (3.95)) для различных правых частей уравнений (3.163). Если матрица A — симметричная положительно определённая, то очень удобным может быть разложение Холецкого и последующее решение систем уравнений (3.163) с помощью представления (3.86).

В прямых методах решения СЛАУ прямой ход, т. е. приведение исходной системы к треугольному виду, является наиболее трудоёмкой частью всего алгоритма, которая требует обычно $O(n^3)$ 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: задавшись специальным начальным приближением $X^{(0)}$, выполняют итерации

$$X^{(k+1)} \leftarrow X^{(k)}(2I - AX^{(k)}), \quad k = 0, 1, 2, \dots \quad (3.164)$$

Метод Шульца — это не что иное как метод Ньютона для решения системы уравнений, применённый к $X^{-1} - A = 0$ (см. §4.5б).³¹ Его можно тоже рассматривать как матричную версию известной процедуры для вычисления обратной величины (см. [12], глава 3).

Предложение 3.13.1 *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

$$X^{(k+1)} = 2X^{(k)} - X^{(k)}AX^{(k)}.$$

³¹Иногда этот метод называют также *методом Хотеллинга*, так как одновременно с Г. Шульцем [110] его рассматривал американский экономист и статистик Г. Хотеллинг [103]. Кроме того, встречается (редко) название *метод Водевига*.

Умножим обе части этого равенства слева на $(-A)$ и добавим к ним по единичной матрице I , получим

$$I - AX^{(k+1)} = I - 2AX^{(k)} + AX^{(k)}AX^{(k)},$$

что равносильно

$$I - AX^{(k+1)} = (I - AX^{(k)})^2, \quad k = 0, 1, 2, \dots$$

Отсюда, в частности, следует, что

$$I - AX^{(k)} = (I - AX^{(0)})^{2^k}, \quad k = 0, 1, 2, \dots$$

Если $X^{(k)} \rightarrow A^{-1}$ при $k \rightarrow \infty$, то $(I - AX^{(0)})^{2^k} \rightarrow 0$ — последовательность степеней матрицы сходится к нулю. Тогда необходимо $\rho(I - AX^{(0)}) < 1$ в силу Предложения 3.3.9.

И наоборот, если $\rho(I - AX^{(0)}) < 1$, то $(I - AX^{(0)})^{2^k} \rightarrow 0$ при $k \rightarrow \infty$, и потому должна иметь место сходимость $X^{(k)} \rightarrow A^{-1}$. ■

Из доказательства предложения следует, что метод Шульца имеет квадратичную сходимость. 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.

3.14 Оценка погрешности приближённого решения

В этом параграфе мы рассмотрим практически важный вопрос об оценке погрешности приближённого решения систем линейных алгебраических уравнений. Будут предложены два простых способа ответить на этот вопрос, хотя в действительности существует довольно много различных подходов к оценке погрешности решения. Здесь уместно упомянуть об очень развитых интервальных методах доказательных вычислений и оценивания погрешностей.

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 $Ax = b$, тогда как x^* — её точное решение. Тогда, принимая во внимание, что $I = A^{-1}A$ и $Ax^* = b$,

$$\begin{aligned}\|\tilde{x} - x^*\| &= \|A^{-1}A\tilde{x} - A^{-1}Ax^*\| \\ &= \|A^{-1}(A\tilde{x} - Ax^*)\| \\ &\leq \|A^{-1}\| \|A\tilde{x} - b\|,\end{aligned}\tag{3.165}$$

где матричная и векторная нормы, естественно, должны быть согласованы. Величина $(A\tilde{x} - b)$ — это невязка приближённого решения \tilde{x} , которую мы обычно можем вычислять непосредственно по \tilde{x} . Как следствие, погрешность решения можно узнать, найдя каким-либо образом или оценив сверху норму обратной матрицы $\|A^{-1}\|$.

Иногда из практики можно получать какую-то информацию о значении $\|A^{-1}\|$. Например, если A — симметричная положительно определённая матрица и известна нижняя граница её спектра $\mu > 0$, то из Предложения 3.2.3 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 §3.9г. Описываемая ситуация типична в связи с численным решением некоторых популярных уравнений математической физики (уравнением Лапласа и его обобщениями, к примеру), для которых дискретные аналоги соответствующих дифференциальных операторов хорошо изучены и известны оценки их собственных значений (см. к примеру, [40, 17]).

Если матрица системы имеет диагональное преобладание, то для оценивания $\|A^{-1}\|$ можно воспользоваться теоремой Алберга-Нильсона (теорема 3.5.3, стр. 345).

В общем случае нахождение $\|A^{-1}\|$ или хотя бы разумной оценки для $\|A^{-1}\|$ в какой-то норме, которое было бы менее трудоёмким, чем решение исходной СЛАУ, является нетривиальным делом. Краткий обзор существующих численных процедур для этой цели, которые называются «оценщиками обусловленности», а также дальнейшие ссылки на литературу можно найти, к примеру, в книге [13], §2.4.3.

Для конкретных численных методов оценка погрешности приближённого решения иногда может быть выведена из свойств этих методов. Например, в стационарных одношаговых итерационных методах последовательность погрешностей приближений своими свойствами очень близка к геометрической прогрессии, и этим обстоятельством можно с успехом воспользоваться.

Пусть задан сходящийся стационарный одношаговый итерационный метод для решения системы линейных алгебраических уравнений. Мы рассмотрим его в каноническом виде

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

предполагая, что $\|C\| < 1$ для некоторой матричной нормы. Ясно, что ввиду результатов §3.96 о связи спектрального радиуса и матричных норм последнее допущение не ограничивает общности нашего рассмотрения. 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 $x^{(0)}, x^{(1)}, \dots, x^{(k)}, \dots$?

Как и прежде, имеем

$$\begin{aligned} x^{(k)} &= Cx^{(k-1)} + d, \\ x^* &= Cx^* + d. \end{aligned}$$

Вычитание второго равенства из первого даёт

$$x^{(k)} - x^* = C(x^{(k-1)} - x^*). \quad (3.166)$$

Перенесём $x^{(k)}$ в правую часть этого соотношения, а затем добавим к обеим частям по $x^{(k-1)}$:

$$x^{(k-1)} - x^* = x^{(k-1)} - x^{(k)} + C(x^{(k-1)} - x^*).$$

Возьмём теперь от обеих частей полученного равенства векторную норму, которая согласована с используемой матричной нормой для C . Применяя затем неравенство треугольника, приходим к оценке

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

Перенесение в левую часть второго слагаемого из правой части и последующее деление обеих частей неравенства на положительную величину $(1 - \|C\|)$ даёт

$$\|x^{(k-1)} - x^*\| \leq \frac{1}{1 - \|C\|} \|x^{(k)} - x^{(k-1)}\|. \quad (3.167)$$

С другой стороны, вспомним, что из (3.166) следует

$$\|x^{(k)} - x^*\| \leq \|C\| \cdot \|x^{(k-1)} - x^*\|.$$

Подставляя сюда вместо $\|x^{(k-1)} - x^*\|$ оценку сверху (3.167), получаем окончательно

$$\|x^{(k)} - x^*\| \leq \frac{\|C\|}{1 - \|C\|} \|x^{(k)} - x^{(k-1)}\|. \quad (3.168)$$

Выведенная оценка может быть использована на практике как для оценки погрешности какого-то приближения из итерационной последовательности, так и для определения момента окончания итераций, т. е. того, достигнута ли желаемая точность приближения к решению или нет.

Пример 3.14.1 Рассмотрим систему линейных алгебраических уравнений

$$\begin{pmatrix} 2 & 1 \\ 3 & 4 \end{pmatrix} x = \begin{pmatrix} 0 \\ 5 \end{pmatrix},$$

точное решение которой равно $(-1, 2)^\top$. Пусть для решения этой системы организован итерационный метод Гаусса-Зейделя с начальным приближением $x^{(0)} = (0, 0)^\top$. Через сколько итераций компоненты очередного приближения к решению станут отличаться от точного решения не более, чем на 10^{-3} ?

Исследуемый нами вопрос требует чебышёвской нормы $\|\cdot\|_\infty$ для измерения отклонения векторов друг от друга, и соответствующая подчинённая матричная норма задаётся выражением из Предложения 3.3.6. Матрица оператора перехода итерационного метода Гаусса-Зейделя согласно (3.126) есть

$$-\begin{pmatrix} 2 & 0 \\ 3 & 4 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -0.5 \\ 0 & 0.375 \end{pmatrix},$$

так что её ∞ -норма равна 0.5. Следовательно, в оценке (3.168) имеем

$$\frac{\|C\|}{1 - \|C\|} = \frac{0.5}{1 - 0.5} = 1,$$

и потому должно быть справедливым неравенство

$$\|x^{(k)} - x^*\|_\infty \leq \|x^{(k)} - x^{(k-1)}\|_\infty. \quad (3.169)$$

Оно показывает, что компоненты очередного приближения отличаются от компонент точного решения не более, чем компоненты приближений друг от друга.

Запустив итерации Гаусса-Зейделя, мы можем видеть, что

$$\begin{aligned}x^{(0)} &= (0, 0)^\top, \\x^{(1)} &= (0, 1.25)^\top, \\x^{(2)} &= (-0.625, 1.71875)^\top, \\&\dots \quad \dots \\x^{(8)} &= (-0.998957, 1.999218)^\top, \\x^{(9)} &= (-0.999609, 1.999707)^\top,\end{aligned}$$

т. е. 9-я итерация отличается от предыдущей 8-й меньше чем на 10^{-3} , и потому согласно оценке (3.169) на этой итерации мы получаем требуемую погрешность. То, что она действительно такова, можно убедиться из сравнения $x^{(9)}$ с известным нам точным решением $(-1, 2)^\top$. ■

Как хорошо видно из примера, практическая реализация рассматриваемой методики оценки погрешности итерационного решения может столкнуться с двумя трудностями. Во-первых, непростым является определение матрицы C (которая может и не задаваться в явном виде). Во-вторых, выбор нормы $\|\cdot\|$, в которой $\|C\| < 1$, также может быть неочевидным. Теоретически такая норма должна существовать, если итерационный процесс сходится из любого начального приближения, но её конкретный выбор в общем случае непрост.

3.15 Линейная задача наименьших квадратов

3.15a Постановка задачи и основные свойства

Для заданных $m \times n$ -матрицы A и m -вектора b *линейной задачей наименьших квадратов* называют задачу отыскания такого вектора x , который доставляет минимум квадратичной форме $\langle Ax - b, Ax - b \rangle$, или, что равносильно, квадрату евклидовой нормы невязки $\|Ax - b\|_2^2$.

Ясно, что для матриц A полного ранга в случае $m \leq n$, когда число строк матрицы не превосходит числа столбцов, искомым минимумом,

как правило, равен нулю. Для квадратной матрицы A 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.



Рис. 3.29. Структурная схема объекта идентификации.

Рассмотрим в качестве примера *задачу идентификации параметров* системы, 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 имеем величину $b \in \mathbb{R}$ (см. Рис. 3.29). Зависимость «вход-выход» является линейной, имеющей вид

$$b = a_1 x_1 + a_2 x_2 + \dots + a_n x_n \quad (3.170)$$

с некоторыми постоянными вещественными коэффициентами x_1, x_2, \dots, x_n . Задача идентификации — это задача определения (оценивания) значений x_k на основе данных о входе 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, которое связывает искомые x_1, x_2, \dots, x_n . Если серия измерений входа-выхода объекта является «достаточно представительной», то можно попытаться решить получающуюся систему уравнений относительно неизвестных x_k и найти их значения. Восстанавливаемую функциональную зависимость (3.170) обычно называют *регрессией* величины b по величинам a_1, a_2, \dots, a_n , а её график — *регрессионной линией* b по a_1, a_2, \dots, a_n .

Пусть e — вектор единичной длины, $\|e\|_2 = 1$. Найдём производную функции $\Phi(x) = \|Ax - b\|_2^2$ в точке x по направлению e . По определению

$$\begin{aligned}
 \frac{\partial \Phi(x)}{\partial e} &= \lim_{t \rightarrow 0} \frac{\Phi(x + te) - \Phi(x)}{t} \\
 &= \lim_{t \rightarrow 0} \frac{\langle A(x + te) - b, A(x + te) - b \rangle - \langle Ax - b, Ax - b \rangle}{t} \\
 &= \lim_{t \rightarrow 0} \frac{\langle A(te), Ax - b \rangle + \langle Ax - b, A(te) \rangle + \langle A(te), A(te) \rangle}{t} \\
 &= \lim_{t \rightarrow 0} \frac{2t \langle A^\top (Ax - b), e \rangle + t^2 \langle Ae, Ae \rangle}{t} \\
 &= \lim_{t \rightarrow 0} \frac{2t \langle A^\top (Ax - b), e \rangle}{t} + \lim_{t \rightarrow 0} \frac{t^2 \langle Ae, Ae \rangle}{t} \\
 &= 2 \langle A^\top Ax - A^\top b, e \rangle.
 \end{aligned}$$

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 $e \in \mathbb{R}^n$ единичной длины. По этой причине должно быть $A^\top Ax - A^\top b = 0$.

Система линейных алгебраических уравнений

$$A^\top Ax = A^\top b, \quad (3.172)$$

как известно, называется *нормальной системой уравнений* для линейной задачи наименьших квадратов с матрицей A и вектором b (см. §2.10е). Сам переход от исходной системы уравнений $Ax = b$ к нормальной системе (3.172) носит название *первой трансформации Гаусса*.³² Решение нормальной системы уравнений, как было показано в §2.10е, всегда существует и доставляет искомым минимум выражению $\Phi(x) = \|Ax - b\|_2^2$, что следует из проведённых выше выкладок и того факта, что гессианом функции Φ является положительно полуопределённая матрица $A^\top A$.

Исследуем единственность псевдорешения. Любое решение линейной задачи наименьших квадратов является также решением нормальной системы уравнений (3.172), так что наш вопрос сводится к следую-

³²Существует также «вторая трансформация Гаусса» систем линейных алгебраических уравнений.

щему: когда нормальная система уравнений имеет единственное решение? Как известно из курса 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 $A^T A x = 0$ имеет только нулевое решение?

Если ненулевой вектор \tilde{x} таков, что $A^T A \tilde{x} = 0$, то и

$$\tilde{x}^T (A^T A \tilde{x}) = 0,$$

и потому

$$\tilde{x}^T (A^T A \tilde{x}) = (\tilde{x}^T A^T)(A \tilde{x}) = (A \tilde{x})^T (A \tilde{x}) = \|A \tilde{x}\|_2^2 = 0.$$

Получаем, что

$$A \tilde{x} = 0.$$

Итак, однородная нормальная система уравнений $A^T A \tilde{x} = 0$ имеет только нулевое решение тогда и только тогда, когда однородная система $A x = 0$ имеет только нулевое решение. Как следствие, справедлива

Теорема 3.15.1 *Линейная задача наименьших квадратов с матрицей A имеет единственное решение тогда и только тогда, когда столбцы A являются линейно независимыми.*

Определение 3.15.2 *Линейная задача наименьших квадратов, у которой матрица имеет линейно независимые столбцы, называется 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 налагают на псевдорешение какие-нибудь условие. Например, это может быть требование того, чтобы псевдорешение имело наименьшую возможную евклидову норму.

Определение 3.15.3 *Псевдорешение системы линейных алгебраических уравнений с наименьшей евклидовой нормой (т. е. 2-нормой) называется нормальным псевдорешением.*

Теорема 3.15.2 *Нормальное псевдорешение системы линейных алгебраических уравнений единственно.*

Доказательство. Нормальное псевдорешение, которое мы обозначим через x_0 , — это перпендикуляр, построенный из начала координат пространства \mathbb{R}^n на плоскость $\hat{x} + \mathcal{K}$, образованную всеми решениями нормальной системы уравнений (3.172). Другими словами, нормальное псевдорешение x_0 — это ортогональная проекция начала координат \mathbb{R}^n , т. е. нуля, на плоскость $\hat{x} + \mathcal{K}$.

То, что вектор x_0 в самом деле искомый, следует из его принадлежности плоскости $\hat{x} + \mathcal{K}$ и того факта, что длина перпендикуляра (расстояние до ортогональной проекции точки) — действительно наименьшее среди расстояний до всех точек плоскости. Это одно из фундаментальных свойств перпендикуляра (ортогональной проекции). Единственность нормального псевдорешения также вытекает из единственности перпендикуляра. ■

3.15б Численные методы для линейной задачи наименьших квадратов

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 (3.172). 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 (3.172) равна квадрату обусловленности исходной, т. е. существенно хуже. В самом деле,

$$\text{cond}_2(A^\top A) = \|A^\top A\|_2 \|(A^\top A)^{-1}\|_2.$$

При этом

$$\|A^\top A\|_2 = \sigma_{\max}^2(A), \quad \text{и} \quad \|(A^\top A)^{-1}\|_2 = \sigma_{\max}^2(A^{-1}),$$

так что

$$\text{cond}_2(A^\top A) = \text{cond}_2^2(A).$$

Тем не менее, если размеры задачи не очень велики и обусловленность матрицы A исходной системы не слишком плоха, этим невыгодным обстоятельством можно пренебречь.

Другой идейно близкий способ — представить решение линейной задачи наименьших квадратов в виде решения расширенной системы линейных уравнений. В самом деле, нормальную систему уравнений (3.172) можно переписать в виде

$$\begin{cases} A^T y = A^T b, \\ Ax - y = 0, \end{cases}$$

который имеет блочно-матричную форму

$$\begin{pmatrix} 0 & A^T \\ A & -I \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} A^T b \\ 0 \end{pmatrix}. \quad (3.173)$$

С другой стороны, нормальную систему можно переписать несколько по-другому, в виде $A^T(b - Ax) = 0$, и затем как эквивалентную расширенную линейную систему

$$\begin{cases} b - Ax = z, \\ A^T z = 0. \end{cases}$$

с $(n + m)$ -вектором неизвестных $(x, z)^T$. Её блочно-матричная форма

$$\begin{pmatrix} I & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} z \\ x \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix}. \quad (3.174)$$

Расширенные СЛАУ (3.173)–(3.174) имеют размеры $(m+n) \times (m+n)$, но их достоинство по сравнению с нормальной системой уравнений (3.172) — отсутствие необходимости перемножения матриц A^T и A и меньший рост числа обусловленности.

Другие подходы к решению линейной задачи наименьших квадратов основаны на использовании QR-разложения матрицы A или её сингулярного разложения. Последний способ был рассмотрен в разделе о приложениях сингулярного разложения матрицы (см. §3.46), и далее мы подробно разберём решение задачи наименьших квадратов, основанное на QR-разложении матрицы.

Пусть известно QR-разложение $m \times n$ -матрицы A , т. е. представление

$$A = QR,$$

где Q — ортогональная $m \times m$ -матрица, R — трапецевидная (обобщённая треугольная) $m \times n$ -матрица. Евклидова норма (2-норма) вектора не меняется при его умножении на ортогональную матрицу, и поэтому

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \|Ax - b\|_2 &= \min_{x \in \mathbb{R}^n} \|QRx - QQ^\top b\|_2 \\ &= \min_{x \in \mathbb{R}^n} \|Q(Rx - Q^\top b)\|_2 \\ &= \min_{x \in \mathbb{R}^n} \|Rx - Q^\top b\|_2. \end{aligned}$$

Если $m \times n$ -матрица A такова, что $m \geq n$, то матрица R тех же размеров имеет вид

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_{nn} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Следовательно,

$$\|Rx - Q^\top b\|_2^2 = \sum_{i=1}^n \left(\sum_{j=1}^n r_{ij}x_j - (Q^\top b)_i \right)^2 + \sum_{i=n+1}^m ((Q^\top b)_i)^2. \quad (3.175)$$

В полученном выражении первая сумма — это квадрат 2-нормы невязки $n \times n$ -системы линейных алгебраических уравнений

$$\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_{nn} \end{pmatrix} x = \begin{pmatrix} (Q^\top b)_1 \\ (Q^\top b)_2 \\ \vdots \\ (Q^\top b)_n \end{pmatrix}. \quad (3.176)$$

Вторая сумма в (3.175) является постоянным слагаемым, так что минимум выражения (3.175) по x_i достигается при наименьшем значении

первой суммы, когда все её квадратичные слагаемые зануляются. Это достигается на решении x системы (3.176). Так как она является верхней (правой) треугольной, то искомое решение легко находится обратной подстановкой. Оно и будет решением линейной задачи наименьших квадратов.

В проведённых выше рассуждениях предполагается, что все $r_{ii} \neq 0$, т. е. матрица A имеет полный ранг. Если это не так и какие-то $r_{ii} = 0$, $i \in \{1, 2, \dots, n\}$, то нашу конструкцию нужно модифицировать. 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 позиции, соответствующие слагаемые переходят из первой суммы в (3.175) во вторую, и размер треугольной СЛАУ, которую необходимо решить, соответственно, уменьшается.

Технологические детали последних двух способов численного решения линейной задачи наименьших квадратов читатель может увидеть, например, в [13, 46].

3.16 Проблема собственных значений

3.16a 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:

$$Av = \lambda v. \quad (3.177)$$

Сам скаляр λ , 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 λ и n -векторы $v \neq 0$, удовлетворяющие условию (3.177).

Система уравнений (3.177) кажется недоопределённой, так как содержит $n + 1$ неизвестных, которые нужно найти из n уравнений. Но на самом деле можно замкнуть 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 (3.177) следует

$$(A - \lambda I)v = 0,$$

что при ненулевом векторе v означает наличие нетривиальной линейной зависимости между столбцами матрицы A . Итак, должно быть

$$\det(A - \lambda I) = 0. \quad (3.178)$$

Это уравнение относительно переменной λ называется, как известно, *характеристическим уравнением* для матрицы A .³³ Оно является алгебраическим уравнением степени n , что следует из формулы для разложения определителя

$$\det(A - \lambda I) = (-1)^n \lambda^n + p_{n-1} \lambda^{n-1} + \dots + p_1 \lambda + p_0,$$

где p_{n-1}, \dots, p_1, p_0 — какие-то выражения от элементов матрицы A . Переход от исходной задачи (3.177) к характеристическому уравнению (3.178) позволяет расчлнить 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

$$(A - \tilde{\lambda} I)x = 0 \quad (3.179)$$

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 (см. [23]) в поле комплексных чисел \mathbb{C} характеристическое уравнение имеет с учётом кратности n корней. Но, вообще говоря, вещественных решений характеристическое уравнение может не иметь.

³³В механике его называют также «вековым уравнением».

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 (3.179), всегда существует. Тем не менее, даже если рассматриваемая матрица A вещественна, могут не существовать вещественные λ и v , удовлетворяющие соотношению

$$Av = \lambda v.$$

В целом для математически наиболее полного исследования проблемы собственных значений необходим выход в поле комплексных чисел \mathbb{C} , которое алгебраически замкнуто.³⁴ Но полезность такого выхода для практического применения собственных чисел и собственных векторов матрицы в каждом конкретном случае должна рассматриваться отдельно.

Кратность собственного значения как корня характеристического уравнения матрицы называется *алгебраической кратностью* собственного значения. Часто её называют просто кратностью. Максимальное число линейно независимых собственных векторов, относящихся к собственному значению, называется *геометрической кратностью* собственного значения. Геометрическая кратность любого собственного значения не превосходит его алгебраической кратности. *Простыми собственными* значениями матрицы называют её собственные значения кратности 1.

Иногда при упоминании рассматриваемой задачи подчёркивают — «алгебраическая проблема собственных значений», чтобы уточнить, что 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

³⁴Напомним, что *алгебраически замкнутым* называется поле, в котором всякий полином ненулевой степени с коэффициентами из этого поля имеет хотя бы один корень [23].

примеру, наибольшего по модулю собственного значения, или нескольких наибольших по модулю собственных значений и соответствующих им собственных векторов.

Собственные значения матриц нужно знать во многих приложениях. Например, задача определения частот собственных колебаний механических систем (весьма актуальная при проектировании различных конструкций) сводится к нахождению 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.

Пример 3.16.1 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

$$x^{(k+1)} = Ax^{(k)} + b^{(k)}, \quad k = 0, 1, 2, \dots, \quad (3.180)$$

служат моделями разнообразных процессов окружающего нас мира, от биологии до экономики.

Общее решение такой системы есть сумма частного решения исходной системы (3.180) и общего решения однородной системы $x^{(k+1)} = Ax^{(k)}$ без свободного члена. Если искать нетривиальные решения однородной системы в виде $x^{(k)} = \lambda^k h$, где λ — ненулевой скаляр и h — n -вектор, то нетрудно убедиться, что λ должно быть собственным значением A , а h — собственным вектором матрицы A . ■

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 матрицы могут иметь и большую размерность, и в любом случае их знание доставляет важную информацию о рассматриваемом линейном операторе, позволяя упростить его представление. Пусть, например, \mathcal{S} — это l -мерное инвариантное подпространство матрицы A , так что $Ax \in \mathcal{S}$ для любого $x \in \mathcal{S}$, и базисом \mathcal{S} являются векторы v_1, v_2, \dots, v_l . Беря базис всего пространства \mathbb{R}^n так, чтобы его последними векторами были v_1, v_2, \dots, v_l (это, очевидно, можно сделать всегда), получим в нём блочно-треугольное представле-

ние рассматриваемого линейного оператора:

$$\begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}$$

с $l \times l$ -блоком A_{22} . 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} , даже для вещественных матриц, ещё одной особенностью проблемы собственных значений, осложняющей её решение, является нелинейный характер задачи, несмотря на традицию отнесения её к «вычислительной линейной алгебре». Это обстоятельство нетрудно осознать из рассмотрения основного соотношения (3.177)

$$Av = \lambda v,$$

которое является системой уравнений относительно λ и v , причём в его правой части суммарная степень неизвестных переменных равна *двум*: $2 = (1 \text{ при } \lambda) + (1 \text{ при } v)$.

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 вещественными или комплексными коэффициентами, является характеристическим полиномом для некоторой матрицы. Если, к примеру,

$$P(x) = p_n x^n + p_{n-1} x^{n-1} + \dots + p_1 x + p_0,$$

и этот полином имеет корни x_1, x_2, \dots, x_n , перечисленные с учётом кратности, то $P(x) = p_n(x - x_1)(x - x_2) \dots (x - x_n)$, а матрица $A_P = (-1)^n \text{diag} \{x_1, x_2, \dots, x_n\}$ и все ей подобные имеют характеристическим полиномом в точности $P(x)$. Вместо диагональной матрицы можно взять какую-нибудь жорданову каноническую форму вида (3.7), группируя в жордановы клетки размера 2×2 и более по несколько одинаковых собственных значений. Опять таки, преобразование подобия из жордановой формы легко получить плотно заполненную матрицу.

Напомним теперь известную в алгебре теорему Абеля-Руффини: для алгебраических полиномов степени 5 и выше не существует прямых (конечных) методов нахождения корней, в которых выражения

для этих корней даются в виде композиций четырёх арифметических действий и операции взятия корня [64]. Как следствие, мы не должны ожидать существования прямых методов решения проблемы собственных значений для произвольных матриц размера 5×5 и более, и потому подавляющее большинство методов решения проблемы собственных значений — существенно итерационные.

3.166 Обусловленность проблемы собственных значений

Под обусловленностью задачи мы понимаем степень чувствительности собственных значений и собственных векторов матрицы по отношению к возмущениям элементов матрицы. Оказывается, что их поведение существенно различно. Спектр матрицы, как множество точек комплексной плоскости \mathbb{C} , непрерывно зависит от элементов матрицы.

Теорема 3.16.1 (теорема Островского)

Пусть $A = (a_{ij})$ и $B = (b_{ij})$ — квадратные $n \times n$ -матрицы, и пусть также

$$M = \max\{|a_{ij}|, |b_{ij}|\}, \quad \delta = \frac{1}{nM} \sum_{i,j} |a_{ij} - b_{ij}|. \quad (3.181)$$

Тогда любому собственному значению $\lambda(B)$ матрицы B можно сопоставить такое собственное значение $\lambda(A)$ матрицы A , что выполнено неравенство

$$|\lambda(A) - \lambda(B)| \leq (n+2)M\delta^{1/n}.$$

Далее, можно так перенумеровать собственные числа матриц A и B , что для любого номера ν имеет место

$$|\lambda_\nu(A) - \lambda_\nu(B)| \leq 2(n+1)^2 M\delta^{1/n},$$

где λ_ν — ν -ое собственное значение.

Читатель может увидеть детальное изложение этой теории в книгах [16, 20, 27, 36, 44, 53].

Но собственные векторы (инвариантные подпространства) матрицы могут изменяться в зависимости от матрицы разрывным образом даже в совершенно обычных ситуациях.

Пример 3.16.2 [53] 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 α/β , можно придать первому собственному вектору любое направление, сколь бы малыми не являлись значения α и β .

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 β 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 δ из (3.181). Но в оценках теоремы эта δ присутствует в степени, меньшей единицы, вследствие чего правая часть оценок допускает неограниченную скорость изменения собственных значений в окрестности значения $\delta = 0$. Простые примеры показывают, что эта возможность в самом деле реализуется.

Пример 3.16.3 Рассмотрим матрицу

$$A = \begin{pmatrix} \alpha & 1 \\ \beta & \alpha \end{pmatrix}$$

— жорданову 2×2 -клетку с собственным значением α , возмущённую элементом β . Собственные значения этой матрицы суть $\alpha \pm \sqrt{\beta}$, так что мгновенная скорость их изменения в зависимости от β равна $\pm \frac{1}{2} \beta^{-1/2}$, и она бесконечна при $\beta = 0$. Это же явление имеет место и для произвольной жордановой клетки, размером 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 проблем при нахождении собственных значений и собственных векторов. Напротив, для матриц, каноническая форма которых 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 3.16.1 *Квадратные матрицы, подобные диагональным матрицам, называются матрицами простой структуры или диагонализуемыми матрицами.*

Матрицы простой структуры называют также *недефектными*, тогда как *дефектные матрицы* — это матрицы, у которых в канонической жордановой форме 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 3.16.2 *Квадратная матрица называется недефектной матрицей или матрицей простой структуры, 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 эрмитовой сопряжённой, т.е. такие матрицы A , что $AA^* = A^*A$. Можно показать [7, 9, 43, 53], 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.

Собственные числа матриц простой структуры зависят от возмущений гораздо более «плавным образом», чем в общем случае.

Теорема 3.16.2 (теорема Бауэра-Файка [100]) *Если A — квадратная матрица простой структуры, $\lambda_i(A)$ — её собственные числа, V — матрица из собственных векторов A , а $\tilde{\lambda}$ — собственное число возмущённой матрицы $A + \Delta A$, то*

$$\min_i |\tilde{\lambda} - \lambda_i(A)| \leq \text{cond}_2(V) \|\Delta A\|_2. \quad (3.182)$$

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 $\lambda_i(A)$, $i = 1, 2, \dots, n$. Если, согласно условию теоремы, A имеет простую структуру, 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. Следовательно, матрица $D - \tilde{\lambda}I$ неособенна.

С другой стороны, матрица $A + \Delta A - \tilde{\lambda}I$ является особенной по построению, так что особенна и матрица $V^{-1}(A + \Delta A - \tilde{\lambda}I)V$. Но

$$\begin{aligned} V^{-1}(A + \Delta A - \tilde{\lambda}I)V &= (D - \tilde{\lambda}I) + V^{-1}(\Delta A)V \\ &= (D - \tilde{\lambda}I)(I + (D - \tilde{\lambda}I)^{-1}V^{-1}(\Delta A)V), \end{aligned}$$

и потому матрица $(I + (D - \tilde{\lambda}I)^{-1}V^{-1}(\Delta A)V)$ также должна быть особенной. Как следствие, матрица

$$(D - \tilde{\lambda}I)^{-1}V^{-1}(\Delta A)V$$

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, это верно для спектральной нормы:

$$\|(D - \tilde{\lambda}I)^{-1}V^{-1}(\Delta A)V\|_2 \geq 1,$$

так что в силу субмультипликативности

$$\|(D - \tilde{\lambda}I)^{-1}\|_2 \|V^{-1}\|_2 \|(\Delta A)\|_2 \|V\|_2 \geq 1.$$

Но $(D - \tilde{\lambda}I)^{-1}$ — диагональная матрица, её спектральная норма равна наибольшему из чисел по диагонали, и поэтому получаем

$$\max_{1 \leq i \leq n} |(\lambda_i - \tilde{\lambda})^{-1}| \cdot \|V^{-1}\|_2 \|\Delta A\|_2 \|V\|_2 \geq 1.$$

Последнее неравенство равносильно

$$\min_{1 \leq i \leq n} |\lambda_i - \tilde{\lambda}| \leq \|V^{-1}\|_2 \|\Delta A\|_2 \|V\|_2,$$

или

$$\min_i |\tilde{\lambda} - \lambda_i(A)| \leq \text{cond}_2(V) \|\Delta A\|_2,$$

как и требовалось. ■

Теорема Бауэра-Файка показывает, что, каково бы ни было возмущение ΔA матрицы простой структуры A , для любого собственного значения $\tilde{\lambda}$ возмущённой матрицы $A + \Delta A$ найдётся собственное значение λ_i матрицы A , отличающееся от $\tilde{\lambda}$ не более чем на величину спектральной нормы возмущения $\|\Delta A\|_2$, умноженную 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

Важнейший частный случай применения теоремы Бауэра-Файка относится к симметричным 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 конечной скоростью и гораздо более умеренно, чем для матриц общего вида.

Об этом же свидетельствуют другие известные результаты теории матриц.

Теорема 3.16.3 (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 $|\tilde{\lambda}_i - \lambda_i| \leq \|B\|_2$.

Доказательство можно найти в [43, 53].

Теорема 3.16.4 (теорема Виландта-Хоффмана) 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]

Теоремы Вейля и Виландта-Хоффмана показывают, что собственные числа эрмитовых и симметричных матриц непрерывно зависят от элементов матрицы, и, самое главное, зависимость эта имеет довольно плавный характер.

Предложение 3.16.1 *Любая квадратная матрица сколь угодно малым возмущением её элементов может быть сделана матрицей простой структуры.*

Доказательство. Воспользуемся теоремой Шура о возможности приведения произвольной матрицы к верхней треугольной с помощью ортогонального преобразования подобия. Тогда

$$Q^T A Q = T,$$

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 ε 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^T . Тогда матрица

$$A + QEQ^T$$

ортогонально подобна матрице $T + E$, так что все её собственные значения различны, и она имеет простую структуру. Кроме того, норма возмущения

$$\|QEQ^T\| \leq \|Q\| \|E\| \|Q^T\| = \text{cond}(Q) \|E\|$$

может быть сделана сколь угодно малой подходящим выбором ε .

Для случая комплексной матрицы A доказательство адаптируется очевидным образом. ■

Следствие. Матрицы простой структуры образуют открытое всюду плотное подмножество во множестве всех квадратных матриц.

Как следствие, недиагонализуемые, дефектные матрицы, в канонической жордановой форме которых присутствуют клетки размера 2 и более, составляют весьма разреженное множество среди всех квадратных матриц. Более точно, это множество *первой бэровской категории* во множестве всех матриц. Подобные множества, называемые также *тощими*, являются в топологическом смысле наиболее бедными множествами (см. [19, 51]). Но на долю дефектных матриц приходится 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 симметричная матрица $A^T A$ (эрмитова матрица $A^* A$ в комплексном случае) всегда имеет простую структуру, т. е. диагонализуема.

3.16в 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 было отмечено в §3.16б, скорость изменения собственных значений в зависимости от возмущений элементов матрицы конечна. Более точно, из теоремы Бауэра-Файка (теорема 3.16.2) следует, что собственные значения непрерывны по Лишпицу в зависимости от элементов матрицы.

Пусть A — $n \times n$ -матрица с различными собственными значениями и ΔA — 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 λ_i собственные значения A , $x^{(i)}$ — соответствующие им собственные векторы, $i = 1, 2, \dots, n$. При этом последние образуют базис в \mathbb{R}^n , коль скоро по предположению A является матрицей простой структуры. Имеем

$$Ax^{(i)} = \lambda^{(i)} x^{(i)},$$

$$(A + \Delta A)(x^{(i)} + \Delta x^{(i)}) = (\lambda_i + \Delta \lambda_i)(x^{(i)} + \Delta x^{(i)}),$$

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

$$(\Delta A)x^{(i)} + A(\Delta x^{(i)}) = \lambda_i(\Delta x^{(i)}) + (\Delta\lambda_i)x^{(i)}, \quad (3.183)$$

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 (3.183) на y_j , получим

$$\begin{aligned} \langle (\Delta A)x^{(i)}, y^{(j)} \rangle + \langle A(\Delta x^{(i)}), y^{(j)} \rangle \\ = \lambda_i \langle \Delta x^{(i)}, y^{(j)} \rangle + (\Delta\lambda_i) \langle x^{(i)}, y^{(j)} \rangle. \end{aligned} \quad (3.184)$$

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 ν_i называются *коэффициентами перекоса* матрицы A , которые отвечают собственным значениям λ_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 φ_i угол между собственными векторами x_i и y_i исходной и эрмитовой сопряжённой матриц. Коэффициенты перекоса характеризуют, таким образом, обусловленность проблемы собственных значений (в смысле второго подхода из описанных в §1.6).

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, $\langle x^{(i)}, y^{(i)} \rangle = \langle x^{(i)}, x^{(i)} \rangle = \|x^{(i)}\|_2^2$, 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 (3.184), 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 3.2.2), 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 (3.184), будем иметь

$$\langle (\Delta A) x^{(i)}, y^{(j)} \rangle + \lambda_j \langle \Delta x^{(i)}, y^{(j)} \rangle = \lambda_i \langle \Delta x^{(i)}, y^{(j)} \rangle.$$

Поэтому

$$\langle \Delta x^{(i)}, y^{(j)} \rangle = \frac{\langle (\Delta A) x^{(i)}, y^{(j)} \rangle}{\lambda_i - \lambda_j}.$$

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 α_{ii} содержательного смысла не имеют, и можно даже положить $\alpha_{ii} = 0$ (напомним, что мы, в действительности, ищем 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 3.2.2. 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 разложения возмущений собственных векторов α_{ij} могут быть оценены 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

$$\|\Delta x^{(i)}\|_2 \leq \|\Delta A\|_2 \cdot \|x\|_2 \cdot \sum_{j \neq i} \frac{\nu_j}{|\lambda_i - \lambda_j|}. \quad (3.185)$$

Отметим значительную разницу в поведении возмущений собственных значений и собственных векторов матриц. Из оценки (3.185) 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 (3.185) присутствуют разности $\lambda_i - \lambda_j$, которые могут быть малыми при близких собственных значениях матрицы. Как следствие, собственные векторы при этом очень чувствительны к возмущениям в элементах матрицы. Это мы могли наблюдать в Примере 3.16.2. В частности, даже для симметричных (эрмитовых) матриц задача отыскания собственных векторов может оказаться плохообусловленной.

3.16г 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

$$Av = \lambda v \quad (3.186)$$

для некоторого собственного вектора $v \in \mathbb{C}^n$. 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 -ую компоненту векторного равенства (3.186):

$$\sum_{j=1}^n a_{lj} v_j = \lambda v_l,$$

что равносильно

$$\sum_{\substack{j=1 \\ j \neq l}}^n a_{lj} v_j = (\lambda - a_{ll}) v_l.$$

Следовательно,

$$\begin{aligned} |\lambda - a_{ll}| |v_l| &= \left| \sum_{j \neq l} a_{lj} v_j \right| \leq \sum_{j \neq l} |a_{lj} v_j| \\ &= \sum_{j \neq l} |a_{lj}| |v_j| \leq |v_l| \sum_{j \neq l} |a_{lj}|, \end{aligned}$$

потому что $|v_j| \leq |v_l|$. Наконец, поскольку $v \neq 0$, мы 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 выполненные нами рассуждения. Потому в целом, если λ — 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 λ 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. Гершгориним:

Теорема 3.16.5 (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 (см. подробности в [45, 53, 111]).

Нетрудно продемонстрировать, что теорема Гершгорина равносильна признаку Адамара неособенности матриц (Теорема 3.5.1). 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 λ — собственное значение матрицы $A = (a_{ij})$, то матрица $(A - \lambda I)$ особенна и потому не может иметь диагональное преобладание. Как следствие, 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 3.16.5 Для 2×2 -матрицы (3.14)

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix},$$

рассмотренной в Примере 3.2.3 (стр. 286), собственные значения суть $\frac{1}{2}(5 \pm \sqrt{33})$, они приблизительно равны -0.372 и 5.372 . На Рис. 3.30,

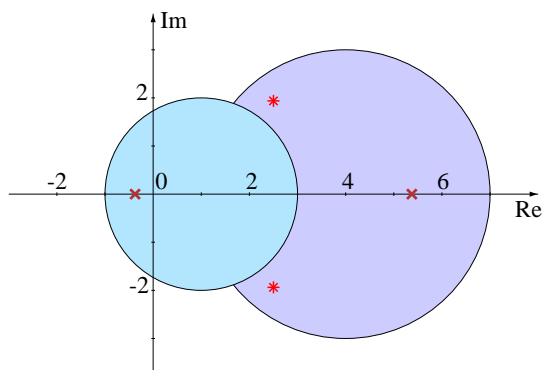


Рис. 3.30. Круги Гершгорина для матриц (3.14) и (3.15).

показывающем соответствующие матрице круги Гершгорина, эти собственные значения выделены крестиками.

Матрица (3.15)

$$\begin{pmatrix} 1 & 2 \\ -3 & 4 \end{pmatrix},$$

которая отличается от матрицы (3.14) 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. 3.30 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.

3.16д Отношение Рэля

Определение 3.16.3 Для квадратной $n \times n$ -матрицы A , вещественной или комплексной, отношением Рэля называется функционал $\mathcal{R}(x)$,

задаваемый как

$$\mathcal{R}(x) := \frac{\langle Ax, x \rangle}{\langle x, x \rangle},$$

который определён на множестве ненулевых векторов из \mathbb{R}^n или \mathbb{C}^n .

Область значений отношения Рэля, т. е. множество

$$\{\mathcal{R}(x) \mid x \neq 0\},$$

называется *областью значений* матрицы A . Можно показать, что оно является выпуклым подмножеством комплексной плоскости \mathbb{C} .

Перечислим основные свойства отношения Рэля.

Для любого скаляра α 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 λ , 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 доказанного свойства, можем заключить, что собственные числа матрицы принадлежат её полю значений.

Собственные векторы являются стационарными точками отношения Рэля, т. е. точками зануления его производной. Покажем это для вещественной симметричной матрицы, для которой отношение Рэля рассматривается для всех ненулевых вещественных векторов:

$$\frac{\partial \mathcal{R}(x)}{\partial x_i} = \frac{\partial}{\partial x_i} \left(\frac{\langle Ax, x \rangle}{\langle x, x \rangle} \right) = \frac{2(Ax)_i \langle x, x \rangle - \langle Ax, x \rangle \cdot 2x_i}{\langle x, x \rangle^2}.$$

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 Рэля для вычислительных методов состоит в том, что с его помощью можно легко получить приближение к собственному значению, если известен приближённый собственный вектор матрицы. Хотя отношение Рэля 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 $y = U^*x$. Поскольку

$$\frac{1}{\|y\|_2^2} \sum_{i=1}^n |y_i|^2 = \sum_{i=1}^n \frac{|y_i|^2}{\|y\|_2^2} = 1,$$

то для эрмитовой матрицы отношение Рэля на векторе y равно выпуклой комбинации, с коэффициентами $(|y_i|/\|y\|_2)^2$, её собственных значений. В целом же из проведённых выше выкладок следует, что область значения отношения Рэля для эрмитовой матрицы — это интервал $[\lambda_{\min}, \lambda_{\max}] \subset \mathbb{R}$, коль скоро все λ_i вещественны. Кроме того, для эрмитовых матриц отношение Рэля позволяет легко находить нетривиальные границы для наименьшего собственного значения сверху и наибольшего собственного значения снизу.

Пример 3.16.6 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:

$$\lambda(A) \in [-4, 12].$$

Посмотрим, что получится с помощью отношения Рэля. С этой целью случайно выберем какие-нибудь векторы $x \in \mathbb{R}^n$ и найдём для них значение отношения Рэля. 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

$$\mathcal{R}(x) := \frac{\langle Ax, x \rangle}{\langle x, x \rangle} = -0.66667.$$

Неплохие оценки для минимального и максимального собственных чисел! ■

В теории с помощью отношения Рэля нетрудно вывести полезные оценки для собственных и сингулярных чисел матриц. В частности, из свойств отношения Рэля следует теорема Вейля (теорема 3.16.3); см. подробности в [43, 53].

3.16е Предварительное упрощение матрицы

Естественная идея состоит в том, чтобы привести матрицу, для которой решается проблема собственных значений, к некоторой специальной форме, у которой собственные значения и/или собственные векторы могут быть найдены проще, чем для исходной. В частности, идеальным было бы приведение матрицы к диагональной или треугольной форме, по которым собственные числа находятся непосредственно.

Элементарными преобразованиями, с помощью которых должно выполняться это приведение, в данном случае должны быть, очевидно, такие, которые сохраняют неизменным спектр матрицы. Это преобразования подобия матрицы $A \mapsto S^{-1}AS$. Но они существенно сложнее действуют на матрицу, чем преобразования линейного комбинирования строк, которые использовались в прямых методах решения систем линейных алгебраических уравнений. 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 §3.16а: 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, в некоторых частных случаях идея предварительного упрощения матрицы для решения проблемы собственных значений может оказаться реализуемой и действительно способствует повышению эффективности численных алгоритмов. Её наиболее популярное воплощение — это так называемая почти треугольная (хессенбергова) форма для общих матриц, а также её частные случаи — трёхдиагональные симметричные и эрмитовы матрицы.

Определение 3.16.4 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:

$$H = \begin{pmatrix} \times & \times & \cdots & \times & \times \\ \times & \times & \cdots & \times & \times \\ & \times & \ddots & \vdots & \vdots \\ & & \ddots & \times & \times \\ 0 & & & \times & \times \end{pmatrix}.$$

Симметричная хессенбергова матрица — это, очевидно, трёхдиагональная матрица.

Предложение 3.16.2 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 $H = QAQ^T$ — хессенбергова матрица.

Доказательство. Рассмотрим для определённости преобразования с помощью матриц отражения.

Возьмём матрицу отражения $Q_1 = I - 2uu^T$ так, 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. занулялись бы элементы a_{31}, \dots, a_{n1} в первом столбце. 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 $Q_1 A$ умножается на $Q_1^\top = Q_1$ справа, то в ней не изменяются элементы первого столбца. Поэтому в матрице $Q_1 A Q_1^\top$, как и в $Q_1 A$, первый столбец имеет нули в позициях с 3-й по n -ую.

Далее выбираем матрицы отражения Q_2, Q_3, \dots, Q_{n-2} так, чтобы умножение слева на Q_i давало нули в позициях с $(i+2)$ -ой по n -ую 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 умножения справа на $Q_i^\top = Q_i$ тоже не портит возникающую почти треугольную структуру результирующей матрицы. Получающаяся в итоге матрица $Q A Q^\top$ с $Q = Q_{n-2} \dots Q_1$ действительно является верхней почти треугольной.

Для матриц отражения рассуждения доказательства совершенно аналогичны, хотя и более длинные. ■

Ниже мы увидим, что хессенбергова форма матрицы в самом деле помогает при реализации некоторых методов решения проблемы собственных значений, в частности, QR-алгоритма (см. §3.17г). Кроме того, хорошие численные методы решения проблемы собственных значений существуют для симметричных хессенберговских матриц, которые являются симметричными трёхдиагональными.

3.17 Численные методы несимметричной проблемы собственных значений

Существует большое количество разнообразных численных методов для решения общей несимметричной проблемы собственных значений. В нашем курсе рассматриваются лишь два основных и, пожалуй, наиболее популярных метода. Более подробную информацию о состоянии этой области вычислительной математики читатель может получить из более полных и специальных книг [72, 11, 13, 45, 46, 47, 61, 73, 75, 77] и др., а также из обзоров и специальных статей.

Выше мы видели, что для несимметричной проблемы собственных значений характерна плохая обусловленность, так что соответствующие задачи иногда являются вычислительно некорректными. Одним из направлений развития современных вычислительных методов линейной алгебры в настоящее время является пересомысление постановок несимметричной проблемы собственных значений. Вместо классической формулировки нахождения «точных 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 [106], можно исследовать расположение собственного значения относительно мнимой оси, или же внутри заданного круга и т. п. (см. [76]).

3.17a Степенной метод

Определение 3.17.1 Если у некоторой матрицы собственные значения λ_i , $i = 1, 2, \dots$, удовлетворяют неравенствам $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \dots$, то λ_1 называют доминирующим собственным значением, а соответствующий ему собственный вектор — доминирующим собственным вектором матрицы.

Степенной метод, описанию которого посвящён этот пункт, 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.

Лежащая в основе степенного метода идея чрезвычайно проста и состоит в том, что если у матрицы A имеется собственное значение λ_1 , превосходящее по модулю все остальные собственные значения, то при действии этой матрицей на произвольный вектор $x \in \mathbb{C}^n$ направление v_1 , отвечающее этому собственному значению λ_1 будет растягиваться сильнее остальных (при $\lambda_1 > 1$) или сжиматься меньше остальных (при $\lambda_1 \leq 1$). При повторном умножении A на результат Ax предшествующего умножения эта компонента ещё более удлинится в сравнении с остальными. Повторив рассмотренную процедуру умножения достаточное количество раз, мы получим вектор, в котором полностью преобладает направление v_1 , т. е. практически будет получен приближённый собственный вектор.

В качестве приближённого собственного значения матрицы A можно при этом взять «отношение» двух последовательных векторов, порождённых нашим процессом — $x^{(k+1)} = A^{k+1}x^{(0)}$ и $x^{(k)} = A^k x^{(0)}$, $k = 0, 1, 2, \dots$. Слово «отношение» взято здесь в кавычки потому, что употреблено не вполне строго: ясно, что векторы $x^{(k+1)}$ и $x^{(k)}$ 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 $x^{(k+1)}$ и $x^{(k)}$, т. е.

$$x_i^{(k+1)} / x_i^{(k)} \quad (3.187)$$

для некоторого $i \in \{1, 2, \dots, n\}$;

- 2) рассматривать отношение проекций последовательных приближений $x^{(k+1)}$ и $x^{(k)}$ на направление, задаваемое каким-нибудь вектором $l^{(k)}$, т. е.

$$\frac{\langle x^{(k+1)}, l^{(k)} \rangle}{\langle x^{(k)}, l^{(k)} \rangle}. \quad (3.188)$$

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 $x^{(k)}$, чтобы не занулился знаменатель в (3.188).

Последний способ кажется более предпочтительным в вычислительном отношении, поскольку позволяет избегать капризного поведения в одной отдельно взятой компоненте вектора $x^{(k)}$, когда она может сделаться очень малой по абсолютной величине или совсем занулиться,

3.17. Численные методы для несимметричной проблемы собственных значений

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

$$\frac{\langle x^{(k+1)}, x^{(k)} \rangle}{\langle x^{(k)}, x^{(k)} \rangle}, \quad (3.189)$$

где $x^{(k)} = A^k x^{(0)}$. Нетрудно увидеть, что это выражение совпадает с отношением Рэлея для приближения $x^{(k)}$ к собственному вектору (см. 3.16д).

Для организации вычислительного алгоритма степенного метода требуется разрешить ещё два тонких момента, связанных с реализацией на ЭВМ.

Во-первых, это возможное неограниченное увеличение (при $\lambda_1 > 1$) или неограниченное уменьшение (при $\lambda_1 < 1$) норм векторов $x^{(k)}$ и $x^{(k+1)}$, 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 (3.189) делается равным единице.

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 v_1 матрицы A . В противном случае произведения любых степеней матрицы A на $x^{(0)}$ будут также иметь нулевые проекции на v_1 , 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 v_1 , 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).

В псевдокоде, представленном в Табл. 3.11, $\tilde{\lambda}$ — это приближённое

Таблица 3.11. Степенной метод для нахождения доминирующего собственного значения матрицы

```

 $k \leftarrow 0;$ 
выбираем вектор  $x^{(0)} \neq 0;$ 
нормируем  $x^{(0)} \leftarrow x^{(0)} / \|x^{(0)}\|_2;$ 
DO WHILE ( метод не сошёлся )
     $y^{(k+1)} \leftarrow Ax^{(k)};$ 
     $\tilde{\lambda} \leftarrow \langle y^{(k+1)}, x^{(k)} \rangle;$ 
     $x^{(k+1)} \leftarrow y^{(k+1)} / \|y^{(k+1)}\|_2;$ 
     $k \leftarrow k + 1;$ 
END DO

```

доминирующее собственное значение матрицы A , а $x^{(k)}$ — текущее приближение к нормированному доминирующему собственному вектору.

Теорема 3.17.1 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 $\text{lin} \{v_2, \dots, v_n\}$ 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 λ_1 — доминирующее собственное значение A . Матрица V

3.17. Численные методы для несимметричной проблемы собственных значений

составлена из собственных векторов v_i матрицы A как из столбцов:

$$V = (v_1 \ v_2 \ \dots \ v_n) = \left(\begin{array}{c|c|c|c} (v_1)_1 & (v_2)_1 & \dots & (v_n)_1 \\ (v_1)_2 & (v_2)_2 & \dots & (v_n)_2 \\ \vdots & \vdots & \ddots & \vdots \\ (v_1)_n & (v_2)_n & \dots & (v_n)_n \end{array} \right),$$

где через $(v_i)_j$ обозначена j -ая компонента i -го собственного вектора матрицы A . При этом можно считать, что $\|v_i\|_2 = 1$. Следовательно,

$$\begin{aligned} A^k x^{(0)} &= (VDV^{-1})^k x^{(0)} = \underbrace{(VDV^{-1})(VDV^{-1}) \dots (VDV^{-1})}_{k \text{ раз}} x^{(0)} \\ &= VD(V^{-1}V)D(V^{-1}V) \dots (V^{-1}V)DV^{-1}x^{(0)} \\ &= VD^k V^{-1}x^{(0)} = VD^k z \\ &= V \begin{pmatrix} \lambda_1^k z_1 \\ \lambda_2^k z_2 \\ \vdots \\ \lambda_n^k z_n \end{pmatrix} = (\lambda_1^k z_1) V \begin{pmatrix} 1 \\ (\lambda_2/\lambda_1)^k (z_2/z_1) \\ \vdots \\ (\lambda_n/\lambda_1)^k (z_n/z_1) \end{pmatrix}, \end{aligned}$$

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 λ_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

$$\begin{pmatrix} 1 \\ (\lambda_2/\lambda_1)^k (z_2/z_1) \\ \vdots \\ (\lambda_n/\lambda_1)^k (z_n/z_1) \end{pmatrix} \quad (3.190)$$

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 λ_1 . Вектор $x^{(k)}$, который отличается от $A^k x^{(0)}$ лишь нормировкой, сходится к собственному вектору v_1 , а величина $\tilde{\lambda} = \langle y^{(k+1)}, x^{(k)} \rangle$ 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 (3.190). 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 отделённость, тем быстрее сходимость степенного метода.

Пример 3.17.1 Для матрицы (3.14)

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

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

3.17. Численные методы для несимметричной проблемы собственных значений

Быстрая сходимость объясняется малостью величины $|\lambda_2/\lambda_1|$, которая, как мы могли видеть в Примере 3.2.3, для рассматриваемой матрицы равна всего лишь 0.069.

Для матрицы (3.15)

$$\begin{pmatrix} 1 & 2 \\ -3 & 4 \end{pmatrix},$$

при тех же исходных условиях степенной метод порождает последовательность значений $\tilde{\lambda}$, 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 3.2.3). ■

Отметим, что для симметричных (эрмитовых) положительно определённых матриц в степенном методе в качестве приближения к доминирующему собственному значению можно брать отношение

$$\frac{\|x^{(k+1)}\|_2}{\|x^{(k)}\|_2}, \quad x^{(k+1)} = Ax^{(k)}$$

(см. [89]).

Наконец, необходимое 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 φ , 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. Либо кроме нормировки собственных векторов следует предусмотреть ещё приведение их к такой форме, в которой координатные представления будут определяться более «жёстко», например, требованием, чтобы первая компонента вектора была бы чисто вещественной.

Пример 3.17.2 Рассмотрим работу степенного метода в применении

к матрице

$$\begin{pmatrix} 1 & 2i \\ 3 & 4i \end{pmatrix},$$

имеющей собственные значения

$$\lambda_1 = -0.4308405 - 0.1485958i,$$

$$\lambda_2 = 1.4308405 + 4.1485958i.$$

Доминирующим собственным значением здесь является λ_2 .

Начав итерирование с вектора $x^{(0)} = (1, 1)^T$, уже через 7 итераций мы получим 6 правильных десятичных знаков в вещественной и мнимой частях собственного значения λ_2 . Но вот в порождаемых алгоритмом нормированных векторах $x^{(k)}$ —

$$x^{(9)} = \begin{pmatrix} -0.01132 - 0.43223i \\ -0.11659 - 0.89413i \end{pmatrix},$$

$$x^{(10)} = \begin{pmatrix} 0.40491 - 0.15163i \\ 0.80725 - 0.40175i \end{pmatrix},$$

$$x^{(11)} = \begin{pmatrix} 0.27536 + 0.33335i \\ 0.64300 + 0.63215i \end{pmatrix},$$

$$x^{(12)} = \begin{pmatrix} 0.22535 + 0.36900i \\ -0.38795 + 0.81397i \end{pmatrix},$$

и так далее — нелегко «невооружённым глазом» узнать один и тот же собственный вектор, который «крутится» 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, 47]. Наиболее неблагоприятен

3.17. Численные методы для несимметричной проблемы собственных значений

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 3.17.3 Рассмотрим работу степенного метода в применении к матрице

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

т. е. к жордановой 2×2 -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

То есть, 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 принципиально не меняется и делается ещё более медленной. ■

3.176 Обратные степенные итерации

Обратными степенными итерациями для матрицы A называют описанный в прошлом параграфе степенной метод, применённый к обратной матрице A^{-1} , в котором вычисляется отношение результатов предыдущей итерации к последующей, т. е. величина, обратная к (3.187) или (3.188). Явное нахождение обратной матрицы A^{-1} при этом не требуется, так как в степенном методе используется лишь результат

$x^{(k+1)}$ её умножения на вектор $x^{(k)}$ очередного приближения, а это, как известно (см., в частности, §3.13), эквивалентно решению системы линейных уравнений $Ax^{(k+1)} = x^{(k)}$.

Так как собственные значения матриц A и A^{-1} взаимно обратны, то обратные степенные итерации будут сходиться к наименьшему по абсолютной величине собственному значению A и соответствующему собственному вектору.

Чтобы в обратном к (3.188) отношении

$$\frac{\langle x^{(k)}, l^{(k)} \rangle}{\langle x^{(k+1)}, l^{(k)} \rangle},$$

которое необходимо вычислять в обратных степенных итерациях, знаменатель не занулялся, удобно брать $l^{(k)} = x^{(k+1)}$. Тогда очередным приближением к наименьшему по модулю собственному значению матрицы A является

$$\frac{\langle x^{(k)}, x^{(k+1)} \rangle}{\langle x^{(k+1)}, x^{(k+1)} \rangle},$$

где $Ax^{(k+1)} = x^{(k)}$. Псевдокод получающегося метода представлен в Табл. 3.12.

Таблица 3.12. Обратные степенные итерации для нахождения наименьшего по модулю собственного значения матрицы A

```

 $k \leftarrow 0;$ 
выбираем вектор  $x^{(0)} \neq 0;$ 
DO WHILE ( метод не сошёлся )
    найти  $y^{(k+1)}$  из системы  $Ay^{(k+1)} = x^{(k)};$ 
     $\tilde{\lambda} \leftarrow \langle x^{(k)}, y^{(k+1)} \rangle / \langle y^{(k+1)}, y^{(k+1)} \rangle;$ 
     $x^{(k+1)} \leftarrow y^{(k+1)} / \|y^{(k+1)}\|_2;$ 
     $k \leftarrow k + 1;$ 
END DO

```

3.17. Численные методы для несимметричной проблемы собственных значений

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 (3.77) или (3.95). Их решение сводится к выполнению прямой и обратной подстановок для треугольных СЛАУ.

Пример 3.17.4 Рассмотрим работу обратных степенных итераций для знакомой нам матрицы

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix},$$

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 $x^{(0)} = (1, 1)^T$, за 7 итераций получим 7 верных значащих цифр наименьшего по модулю собственного числа 0.3722813 . Скорость сходимости здесь получается такой же, как в Примере 3.14.1 для доминирующего собственного значения этой матрицы, что неудивительно ввиду одинакового значения знаменателя геометрической прогрессии λ_2/λ_1 . ■

Обратные степенные итерации особенно эффективны в случае, когда имеется хорошее приближение к собственному значению и требуется найти соответствующий собственный вектор.

3.17в Сдвиги спектра

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 ϑ . Если $\lambda_i(A)$ — собственные значения матрицы A , то собственными значениями матрицы $A + \vartheta I$ являются числа $\lambda_i(A) + \vartheta$, тогда как собственные векторы остаются неизменными. Цель сдвига — преобразование 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 станет наибольшим по модулю, и теперь уже единственным, т. е. доминирующим. Соответственно, степенной метод сделается применимым к новой матрице.

Пример 3.17.5 Для матрицы (3.15)

$$\begin{pmatrix} 1 & 2 \\ -3 & 4 \end{pmatrix},$$

как было отмечено в Примере 3.17.1, простейший степенной метод расходится из-за существования двух наибольших по абсолютной величине собственных значений.

Но если сдвинуть эту матрицу на $2i$, то её спектр (см. Рис. 3.30) поднимется «вверх», абсолютные величины собственных значений перестанут совпадать, и степенной метод окажется применимым.

Степенные итерации для «сдвинутой» матрицы

$$\begin{pmatrix} 1 + 2i & 2 \\ -3 & 4 + 2i \end{pmatrix} \quad (3.191)$$

довольно быстро сходятся к наибольшему по модулю собственному значению $\frac{5}{2} + (2 + \frac{1}{2}\sqrt{15})i \approx 2.5 + 3.9364917i$. Детальная картина сходимости при вычислениях с двойной точностью и начальным вектором $x^{(0)} = (1, 1)^T$ 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	$2.5000 + 3.9365i$

В данном случае для матрицы (3.191) имеем $|\lambda_2/\lambda_1| \approx 0.536$.

Ещё большее ускорение сходимости степенного метода можно получить при сдвиге исходной матрицы на $(-2 + 2i)$, когда отношение модулей собственных значений становится равным всего 0.127 . ■

3.17. Численные методы для несимметричной проблемы собственных значений

Поскольку спектр симметричной (эрмитовой) матрицы лежит на вещественной оси, то к таким матрицам имеет смысл применять вещественные сдвиги. В частности, при этом для симметричных вещественных матриц алгоритмы будут реализовываться в более простой вещественной арифметике.

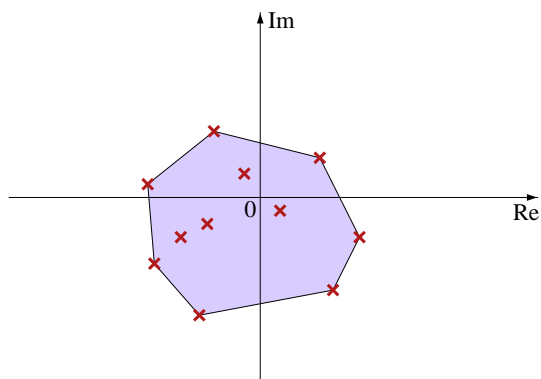


Рис. 3.31. С помощью подходящего сдвига матрицы любую крайнюю точку выпуклой оболочки спектра можно сделать наибольшей по модулю.

С помощью сдвигов матрицы можно любое её собственное значение, которое является крайней точкой выпуклой оболочки спектра, сделать наибольшим по модулю, обеспечив, таким образом, сходимость к нему итераций степенного метода (см. Рис. 3.31). Но как добиться сходимости к другим собственным значениям, которые лежат «внутри» спектра, а не «с краю»? Здесь на помощь приходят обратные степенные итерации.

Обратные степенные итерации сходятся к ближайшей к нулю точке спектра матрицы, и такой точкой с помощью подходящего сдвига может быть сделано любое собственное число. В этом — важное 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 ϑ можно добиться того, чтобы

$$\left| \frac{\lambda_2 + \vartheta}{\lambda_1 + \vartheta} \right|$$

было меньшим, чем $|\lambda_2/\lambda_1|$, ускорив тем самым степенные итерации. Совершенно аналогичный эффект оказывает удачный выбор сдвига на отношение $|\lambda_n/\lambda_{n-1}|$, которое определяет скорость сходимости обратных степенных итераций.

3.17г 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 первой,³⁵ а Дж. Фрэнсис более полно разработал практическую версию QR-алгоритма.

QR-алгоритм является наиболее успешным представителем большого семейства родственных методов решения полной проблемы собственных значений, основанных на разложении исходной матрицы на простые сомножители. QR-алгоритму предшествовал LR-алгоритм Х. Рунтисхаузера. На практике применяются также ортогональный степенной метод, предложенный В.В. Воеводиным [71], и различные другие близкие вычислительные процессы.

Вспомним теорему о QR-разложении (Теорема 3.7.2, стр. 387): всякая квадратная матрица представима в виде произведения ортогональной и правой (верхней) треугольной матриц. Ранее в нашем курсе мы уже обсуждали конструктивные способы выполнения этого разложения — с помощью матриц отражения Хаусхолдера и с помощью матриц вращений. Следовательно, далее можно считать, что QR-разложение выполнимо и основывать на этом факте свои построения.

Вычислительная схема базового QR-алгоритма для решения проблемы собственных значений представлена в Табл. 3.13: мы разлагаем матрицу $A^{(k)}$, полученную на k -м шаге алгоритма, $k = 0, 1, 2, \dots$, на ортогональный $Q^{(k)}$ и правый треугольный $R^{(k)}$ сомножители и далее, поменяв их местами, умножаем друг на друга, образуя следующее приближение $A^{(k+1)}$.

Прежде всего отметим, что поскольку

$$A^{(k+1)} = R^{(k)}Q^{(k)} = (Q^{(k)})^\top (Q^{(k)}R^{(k)})Q^{(k)} = (Q^{(k)})^\top A^{(k)}Q^{(k)},$$

³⁵Упомянув вклад В.Н. Кублановской в изобретение QR-алгоритма, нередко ссылаются на её статью 1961 года в «Журнале вычислительной математики и математической физики» [85]. Но самое первое сообщение о QR-алгоритме было опубликовано ею раньше — в Дополнении к изданию 1960 года книги [47].

3.17. Численные методы для несимметричной проблемы собственных значений

Таблица 3.13. QR-алгоритм для нахождения собственных значений матрицы A

```
 $k \leftarrow 0;$   
 $A^{(0)} \leftarrow A;$   
DO WHILE ( метод не сошёлся )  
    вычислить QR-разложение  $A^{(k)} = Q^{(k)} R^{(k)};$   
     $A^{(k+1)} \leftarrow R^{(k)} Q^{(k)};$   
     $k \leftarrow k + 1;$   
END DO
```

то все матрицы $A^{(k)}$, $k = 0, 1, 2, \dots$, ортогонально подобны друг другу и исходной матрице A . Как следствие, 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.

Это означает, что предельная матрица, к которой сходится QR-алгоритм, является верхней треугольной либо верхней блочно-треугольной, причём 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 2×2 -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 3.17.6 Проиллюстрируем работу QR-алгоритма на примере матрицы

$$\begin{pmatrix} 1 & -2 & 3 \\ 4 & 5 & -6 \\ -7 & 8 & 9 \end{pmatrix}, \quad (3.192)$$

имеющей собственные значения

$$\begin{aligned} &2.7584 \\ &6.1207 \pm 8.04789i \end{aligned}$$

Читатель может провести на компьютере этот увлекательный эксперимент самостоятельно, воспользовавшись системами Scilab, MATLAB или им подобными: все они имеют встроенную процедуру для QR-разложения матриц.³⁶

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 2×2 -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 1×1 -блоке, который расположен на месте $(3, 3)$, и оно равно 2.7584. ■

³⁶В Scilab'e, MATLAB'e и Octave она так и называется — `qr`.

Пример 3.17.7 Для ортогональной матрицы

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad (3.193)$$

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 (3.193) имеет собственные значения, равные ± 1 , так что в данном случае QR-алгоритм не работает. ■

3.17д Модификации QR-алгоритма

Представленная в Табл. 3.13 версия QR-алгоритма на практике обычно снабжается рядом модификаций, которые существенно повышают её эффективность. Главными из этих модификаций являются

- 1) сдвиги матрицы, рассмотренные нами в §3.17в, и
- 2) предварительное приведение матрицы к специальной верхней почти треугольной форме.

Можно показать (см. теорию в книгах [13, 44]), что, аналогично степенному методу, сдвиги тоже помогают ускорению QR-алгоритма. Но в QR-алгоритме их традиционно организуют способом, представленным в Табл. 3.14.

Особенность организации 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 $A^{(k)}$ и $A^{(k+1)}$ ортогонально подобны, совершенно так же, как и в исходной версии QR-алгоритма:

$$\begin{aligned} A^{(k+1)} &= R^{(k)} Q^{(k)} + \vartheta_k I = (Q^{(k)})^\top Q^{(k)} R^{(k)} Q^{(k)} + \vartheta_k (Q^{(k)})^\top Q^{(k)} \\ &= (Q^{(k)})^\top (Q^{(k)} R^{(k)} + \vartheta_k I) Q^{(k)} = (Q^{(k)})^\top A^{(k)} Q^{(k)}. \end{aligned}$$

Представленная организация сдвигов позволяет сделать их в одно и то же время локальными и динамическими по характеру, т. е. изменяющимся от шага к шагу. По этой причине их можно корректировать на основе результатов промежуточных вычислений.

Таблица 3.14. QR-алгоритм со сдвигами для нахождения собственных значений матрицы A

```

 $k \leftarrow 0$ ;
 $A^{(0)} \leftarrow A$ ;
DO WHILE ( метод не сошёлся )
    выбрать сдвиг  $\vartheta_k$ , приближённо равный
    собственному значению  $A$ ;
    вычислить QR-разложение сдвинутой
    матрицы  $A^{(k)} - \vartheta_k I = Q^{(k)} R^{(k)}$ ;
     $A^{(k+1)} \leftarrow R^{(k)} Q^{(k)} + \vartheta_k I$ ;
     $k \leftarrow k + 1$ ;
END DO

```

Пример 3.17.8 Проиллюстрируем работу QR-алгоритма со сдвигами на знакомой нам матрице (3.192)

$$\begin{pmatrix} 1 & -2 & 3 \\ 4 & 5 & -6 \\ -7 & 8 & 9 \end{pmatrix}$$

из предыдущего примера.

Запустим сначала для этой матрицы QR-алгоритм с нулевым сдвигом. Через 5 итераций получим матрицу

$$\begin{pmatrix} 4.6508 & -11.925 & 4.2996 \\ 5.6124 & 7.578999 & 2.4637651 \\ 0.015767 & -0.0055973 & 2.7702 \end{pmatrix},$$

откуда можно ясно увидеть постепенное выделение блочно-треугольной формы с ведущим 2×2 -блоком: элементы на местах $(2, 2)$ и $(3, 3)$ делаются маленькими в сравнении с другими элементами матрицы. Как следствие, элемент на месте $(3, 3)$ должен быть близок к собственному значению, и на его величину можно сделать сдвиг.

3.17. Численные методы для несимметричной проблемы собственных значений

Положив $\theta_k = 2.77$, получим резкое ускорение сходимости, так что после 8-й итерации

$$\begin{pmatrix} 3.1391 & -10.546 & 4.8526 \\ 6.9846 & 9.1025 & 1.0639 \\ 4.13 \cdot 10^{-11} & -2.77 \cdot 10^{-12} & 2.7584 \end{pmatrix}.$$

Сходимость ускорилась бы ещё значительно, если бы мы динамически подстраивали параметр θ на этих шагах. ■

Предложение 3.17.1 Матрица, имеющая хессенбергову форму, сохраняет эту форму при выполнении с ней QR-алгоритма.

Доказательство. Предположим, что к началу k -го шага алгоритма получена хессенбергова матрица $(A^{(k)} - \vartheta I)$. При её QR-разложении

$$(A^{(k)} - \vartheta I) = Q^{(k)} R^{(k)}$$

в качестве ортогонального сомножителя $Q^{(k)}$ получается также хессенбергова матрица.

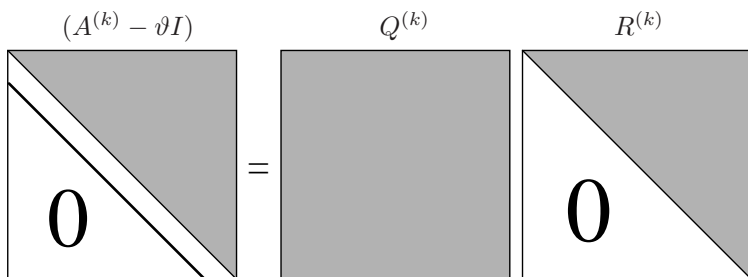


Рис. 3.32. Разложение матрицы $(A^{(k)} - \vartheta I)$ в QR-алгоритме.

В самом деле, поскольку матрица $R^{(k)}$ — правая треугольная, то j -ый столбец в $(A^{(k)} - \vartheta I)$ есть линейная комбинация первых j столбцов матрицы $Q^{(k)}$ с коэффициентами, равными элементам из j -го столбца $R^{(k)}$ (см. Рис. 3.32). Отсюда следует, что первый столбец матрицы $Q^{(k)}$ должен выглядеть совершенно так же, как первый столбец в $(A^{(k)} - \vartheta I)$, т. е. иметь нулевыми элементы на местах $(3,1)$, $(4,1)$ и т. д.

Переходя ко второму столбцу матрицы $(A^{(k)} - \vartheta I)$, 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 совершенно так же, как второй столбец в $(A^{(k)} - \vartheta I)$, т. е. иметь нулевыми элементы на местах (4,2), (5,2) и т. д. Продолжая эти рассуждения для последующих столбцов матриц $Q^{(k)}$ и $(A^{(k)} - \vartheta I)$, 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 ϑI не изменяет верхней почти треугольной формы матрицы. Таким образом, к началу $(k+1)$ -го шага QR-алгоритма снова получается матрица в хессенберговой форме. ■

Смысл предварительного приведения к хессенберговой форме заключается в следующем. Хотя это приведение матрицы требует $O(n^3)$ операций, дальнейшее выполнение одной итерации QR-алгоритма с хессенберговой формой будет теперь стоить всего $O(n^2)$ операций, так что общая трудоёмкость QR-алгоритма составит $O(n^3)$. Для исходной версии QR-алгоритма, которая оперирует с плотно заполненной матрицей, трудоёмкость равна $O(n^4)$, поскольку на каждой итерации алгоритма выполнение QR-разложения требует $O(n^3)$ операций.

3.18 Численные методы для симметричной проблемы собственных значений и сингулярного разложения

Симметричная проблема собственных значений, как мы могли видеть, своими свойствами выделяется из общей проблемы собственных значений матриц. Для симметричных (эрмитовых в комплексном случае) матриц все собственные значения вещественны, а собственные векторы ортогональны. Симметричная проблема собственных значений обладает хорошей обусловленностью. Эти обстоятельства позволяют выделить её в качестве важнейшей отдельной части общей проблемы собственных значений. Кроме того, с симметричной проблемой

собственных значений для матриц

$$A^*A, \quad AA^*, \quad \begin{pmatrix} 0 & A^* \\ A & 0 \end{pmatrix}$$

тесно связана с задачей нахождения сингулярного разложения матрицы A (см. §3.2д). Как следствие, логично рассматривать вместе численные методы для симметричной проблемы собственных значений и сингулярного разложения.

3.18a Метод Якоби для решения симметричной проблемы собственных значений

В этом параграфе мы рассмотрим численный метод для решения симметричной проблемы собственных значений, т. е. для вычисления собственных чисел и собственных векторов симметричных матриц. Он был впервые применён К.Г. Якоби в 1846 году к конкретной 7×7 -матрице, а затем был забыт на целое столетие и вновь переоткрыт лишь после Второй мировой войны, когда началось бурное развитие вычислительной математики.

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 §3.7е. Почему именно ортогональные матрицы и почему вращений? Ответ на эти вопросы станет ясен позднее при анализе работы алгоритма.

Итак, положим $A^{(0)} := A$. Если матрица $A^{(k)}$, $k = 0, 1, 2, \dots$, уже вычислена, то подберём матрицу вращений $G(p, q, \theta)$ вида (3.99) таким образом, чтобы сделать нулями пару внедиагональных элементов в позициях (p, q) и (q, p) в матрице $A^{(k+1)} := G(p, q, \theta)^\top A^{(k)} G(p, q, \theta)$. Желая

достичь этой цели, мы должны добиться выполнения равенства

$$\begin{pmatrix} a_{pp}^{(k+1)} & a_{pq}^{(k+1)} \\ a_{qp}^{(k+1)} & a_{qq}^{(k+1)} \end{pmatrix} = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}^{\top} \begin{pmatrix} a_{pp}^{(k)} & a_{pq}^{(k)} \\ a_{qp}^{(k)} & a_{qq}^{(k)} \end{pmatrix} \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \\ = \begin{pmatrix} \times & 0 \\ 0 & \times \end{pmatrix},$$

где, как обычно, 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 . Этот эффект будет проанализирован ниже в Предложении 3.18.2.

Опуская индексы, обозначающие номер итерации и приняв сокращённые обозначения $c = \cos \theta$, $s = \sin \theta$, получим

$$\begin{pmatrix} \times & 0 \\ 0 & \times \end{pmatrix} = \begin{pmatrix} a_{pp}c^2 + a_{qq}s^2 + 2sca_{pq} & sc(a_{qq} - a_{pp}) + a_{pq}(c^2 - s^2) \\ sc(a_{qq} - a_{pp}) + a_{pq}(c^2 - s^2) & a_{pp}s^2 + a_{qq}c^2 - 2sca_{pq} \end{pmatrix}.$$

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 τ , т. е. $\tau = 1/\operatorname{tg}(2\theta)$.

Пусть

$$t := \frac{\sin \theta}{\cos \theta} = \operatorname{tg} \theta.$$

3.18. Численные методы для симметричной проблемы собственных значений

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 корнем квадратного уравнения

$$t^2 + 2\tau t - 1 = 0. \quad (3.194)$$

Его дискриминант $(4\tau^2 + 4)$ положителен и, следовательно, уравнение (3.194) всегда имеет вещественные корни

$$t_{1,2} = -\tau \pm \sqrt{\tau^2 + 1}.$$

При этом из двух корней мы берём наименьший по абсолютной величине, равный

$$\begin{aligned} t &= -\tau + \operatorname{sgn} \tau \cdot \sqrt{\tau^2 + 1}, & \text{если } \tau \neq 0, \\ t &= \pm 1, & \text{если } \tau = 0. \end{aligned}$$

Первую формулу для улучшения численной устойчивости лучше записать в виде, освобождённом от вычитания близких чисел, так что соответствующий корень есть

$$\begin{aligned} t &= \frac{(-\tau + \operatorname{sgn} \tau \cdot \sqrt{\tau^2 + 1})(\tau + \operatorname{sgn} \tau \cdot \sqrt{\tau^2 + 1})}{(\tau + \operatorname{sgn} \tau \cdot \sqrt{\tau^2 + 1})} \\ &= \frac{1}{(\tau + \operatorname{sgn} \tau \cdot \sqrt{\tau^2 + 1})} \quad \text{при } \tau \neq 0. \end{aligned}$$

Наконец, на основе известных тригонометрических формул, выражающих косинус и синус через тангенс, находим c и s :

$$c = \frac{1}{\sqrt{t^2 + 1}}, \quad s = t \cdot c.$$

Займёмся теперь обоснованием сходимости метода Якоби для решения симметричной проблемы собственных значений.

Предложение 3.18.1 Фробениусова норма матрицы A , т. е.

$$\|A\|_F = \left(\sum_{i,j=1}^n a_{ij}^2 \right)^{1/2},$$

не изменяется при умножениях на ортогональные матрицы слева или справа.

Доказательство. Напомним, что *следом матрицы* $A = (a_{ij})$, обозначаемым $\text{tr } A$, называется сумма всех её диагональных элементов:

$$\text{tr } A = \sum_{i=1}^n a_{ii}.$$

Нетрудно проверить, что привлечение понятия следа позволяет переписать определение фробениусовой нормы матрицы таким образом

$$\|A\|_F = \left(\sum_{j=1}^n \left(\sum_{i=1}^n a_{ij} a_{ij} \right) \right)^{1/2} = (\text{tr } (A^\top A))^{1/2}.$$

Следовательно, для любой ортогональной матрицы Q справедливо

$$\begin{aligned} \|QA\|_F &= \left(\text{tr } ((QA)^\top (QA)) \right)^{1/2} \\ &= \left(\text{tr } (A^\top Q^\top QA) \right)^{1/2} = (\text{tr } (A^\top A))^{1/2} = \|A\|_F. \end{aligned}$$

Для доказательства аналогичного соотношения с умножением на ортогональную матрицу справа заметим, что фробениусова норма не меняется при транспонировании матрицы. Следовательно,

$$\|AQ\|_F = \|(Q^\top A^\top)^\top\|_F = \|Q^\top A^\top\|_F = \|A^\top\|_F = \|A\|_F,$$

что завершает доказательство Предложения. ■

Следствие. Фробениусова норма матрицы не меняется при ортогональных преобразованиях подобия.

Для более точного описания меры близости матриц $A^{(k)}$, которые порождаются конструируемым нами методом, к диагональной матрице введём величину

$$ND(A) = \left(\sum_{j \neq i} a_{ij}^2 \right)^{1/2}$$

— фробениусову норму внедиагональной части матрицы. Ясно, что матрица A диагональна тогда и только тогда, когда $ND(A) = 0$.

Предложение 3.18.2 Пусть преобразование подобия матрицы A с помощью матрицы вращений G таково, что в матрице $B = G^\top A G$ аннулируются элементы в позициях (p, q) и (q, p) . Тогда

$$ND^2(B) = ND^2(A) - 2a_{pq}^2. \quad (3.195)$$

Итак, в сравнении с матрицей A в матрице B изменились элементы строк и столбцов с номерами p и q , но фробениусова норма недиагональной части изменилась при этом так, как будто кроме аннулирования элементов a_{pq} и a_{qp} ничего не произошло.

Доказательство. Для 2×2 -подматрицы

$$\begin{pmatrix} a_{pp} & a_{pq} \\ a_{qp} & a_{qq} \end{pmatrix}$$

из матрицы A и соответствующей ей 2×2 -подматрицы

$$\begin{pmatrix} b_{pp} & 0 \\ 0 & b_{qq} \end{pmatrix}$$

в матрице B справедливо соотношение

$$a_{pp}^2 + a_{qq}^2 + 2a_{pq}^2 = b_{pp}^2 + b_{qq}^2,$$

так как ортогональным преобразованием подобия фробениусова норма матрицы не изменяется. Но, кроме того, $\|A\|_F^2 = \|B\|_F^2$, 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} . ■

Таблица 3.15. Метод Якоби для вычисления собственных значений симметричной матрицы

<p style="text-align: center;">Вход</p> <p>Симметричная матрица A.</p>
<p style="text-align: center;">Выход</p> <p>Матрица, на диагонали которой стоят приближения к собственным значениям A.</p>
<p style="text-align: center;">Алгоритм</p> <p>DO WHILE (метод не сошёлся)</p> <p style="padding-left: 40px;">выбрать ненулевой внедиагональный элемент a_{pq} в A ;</p> <p style="padding-left: 40px;">обнулить a_{pq} и a_{qp} преобразованием подобия с матрицей вращения $G(p, q, \theta)$;</p> <p>END DO</p>

Теперь можно ответить на вопрос о том, почему в методе Якоби для преобразований подобия применяются именно ортогональные матрицы. Как следует из результатов Предложений 3.18.1 и 3.18.2, умножение на ортогональные матрицы обладает замечательным свойством сохранения фробениусовой нормы матрицы и, как следствие, «перекачивания» её величины с внедиагональных элементов на диагональ в результате специально подобранных цепочек таких умножений. При других преобразованиях подобия добиться этого было бы едва ли возможно.

Итак, всё готово для организации метода вращений Якоби — итерационного процесса приведения симметричной матрицы к диагональному виду, при котором внедиагональные элементы последовательно подавляются. Как уже отмечалось, занулённые на каком-то шаге алгоритма элементы могут впоследствии вновь сделаться ненулевыми. Но результат Предложения 3.18.2 показывает, что норма внедиагональной части матрицы при этом всё равно монотонно уменьшается. В Табл. 3.15 схематично представлен простейший вариант метод Якоби.

Рассмотрим вопрос о критерии остановки метода Якоби, тесно связанный 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 $\epsilon > 0$. Но более элегантное решение этого вопроса может быть основано на теореме Гершгорина (см. §3.16г). Если A — симметричная $n \times n$ -матрица, полученная на каком-то шаге метода вращений Якоби, то ясно, что её собственные значения локализованы в интервалах $[a_{ii} - \Delta, a_{ii} + \Delta]$, $i = 1, 2, \dots, n$, где

$$\Delta = \max_{1 \leq i \leq n} \sum_{j \neq i} |a_{ij}|.$$

При общем уменьшении величины внедиагональных элементов значение Δ также становится маленьким и обеспечивает точность оценивания собственных значений, лучшую чем с помощью ND .

Различные способы выбора ненулевых 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 этого элемента имеет трудоёмкость $n(n-1)/2$, что может оказаться относительно дорогостоящим, особенно для матриц больших размеров. Преобразование подобия с матрицей вращений обходится всего в $O(n)$ операций! Чаще 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.

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., к примеру, рассуждения в [90]). Дальнейшее развитие не подтвердило эти пессимистичные прогнозы. Выяснилось, что метод Якоби почти не имеет конкурентов по точности нахождения малых собственных значений, тогда как методы, основанные на трёхдиагонализации исходной матрицы,

могут терять точность (соответствующие примеры приведены в [13]). Кроме того, метод Якоби оказался хорошо распараллеливаемым, т. е. подходящим для расчётов на современных многопроцессорных ЭВМ.

3.186 Численные методы сингулярного разложения

Сингулярные числа зависят от элементов матрицы существенно более плавным образом, нежели собственные числа. Это непосредственно следует из теорем Вейля (теорема 3.16.3) и Виландта-Хоффмана (теорема 3.16.4). В частности, из теоремы Вейля вытекает

Следствие. Пусть A и B — произвольные матрицы одинакового размера, причём $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n$ — сингулярные числа матрицы A , а $\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq \dots \geq \tilde{\sigma}_n$ — сингулярные числа матрицы $\tilde{A} = A + B$. Тогда $|\tilde{\sigma}_i - \sigma_i| \leq \|B\|_2$.

Итак, 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 3.16.3), тогда как скорость изменения сингулярных чисел соразмерна норме матрицы.

Простейшие методы нахождения сингулярных чисел матриц основаны на том, что они являются собственными числами матриц $A^T A$ и AA^T ($A^* A$ и AA^* в комплексном случае).

Алгоритм Голуба-Кахана [11]

Метод Якоби для сингулярного разложения матрицы [71, 11]

Литература к главе 3

Основная

- [1] Бахвалов Н.С., Жидков Н.П., Кобельков Г.М. *Численные методы*. — Москва: «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*. — Москва: Дрофа, 2008.
- [3] Беклемишев Д.В. *Дополнительные главы линейной алгебры*. — Москва: Наука, 1983.
- [4] Березин И.С., Жидков Н.П. *Методы вычислений. Т. 1–2*. — Москва: Наука, 1966.

- [5] Вержвицкий В.М. *Численные методы. Части 1–2.* – Москва: «Оникс 21 век», 2005.
- [6] Воеводин В.В. *Линейная алгебра.* – Москва: Наука, 1980.
- [7] Воеводин В.В., Воеводин Вл.В. *Энциклопедия линейной алгебры. Электронная система ЛИНЕАЛ.* – Санкт-Петербург: БХВ-Петербург, 2006.
- [8] Волков Е.А. *Численные методы.* – Москва: Наука, 1987.
- [9] ГАНТМАХЕР Ф.Р. *Теория матриц.* – Москва: Наука, 1988.
- [10] ГЛАЗМАН И.М., ЛЮВИЧ Ю.И. *Конечномерный линейный анализ.* – Москва: Наука, 1969.
- [11] ГОЛУБ Дж., ван ЛОУН Ч. *Матричные вычисления.* – Москва: Мир, 1999.
- [12] ДЕМИДОВИЧ Б.П., МАРОН А.А. *Основы вычислительной математики.* – Москва: Наука, 1970.
- [13] ДЕММЕЛЬ Дж. *Вычислительная линейная алгебра.* – Москва: Мир, 2001.
- [14] ЗОРИЧ В.А. *Математический анализ.* Т. 1. – Москва: Наука, 1981. Т. 2. – Москва: Наука, 1984, а также более поздние издания.
- [15] ИКРАМОВ Х.Д. *Численные методы для симметричных линейных систем.* – Москва: Наука, 1988.
- [16] ИКРАМОВ Х.Д. *Несимметричная проблема собственных значений.* – Москва: 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.* – Москва: Наука, 1985.
- [19] КАНТОРОВИЧ Л.В., АКИЛОВ Г.П. *Функциональный анализ.* – Москва: Наука, 1984.
- [20] КАТО Т. *Теория возмущений линейных операторов.* – Москва: Мир, 1972.
- [21] КОЛЛАТЦ Л. *Функциональный анализ и вычислительная математика.* – Москва: 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.* – Москва: Физматлит, 2001.
- [24] КОСТРИКИН А.Н. *Введение в алгебру. Часть 2. Линейная алгебра.* – Москва: 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., БОБКОВ В.В., МОНАСТЫРНЫЙ П.И. *Вычислительные методы. Т. 1–2.* – Москва: Наука, 1976.
- [27] ЛАНКАСТЕР П. *Теория матриц.* – Москва: Наука, 1978.

- [28] Лоусон Ч., Хенсон Р. Численное решение задач методом наименьших квадратов. – Москва: Наука, 1986.
- [29] Мальцев А.И. Основы линейной алгебры. – Москва: Наука, 1975.
- [30] Марчук Г.И., Кузнецов Ю.А. Итерационные методы и квадратичные функционалы. – Новосибирск: Наука, 1972.
- [31] Матрицы и квадратичные формы. Основные понятия. Терминология / Академия Наук СССР. Комитет научно-технической терминологии. – Москва: 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. – Санкт-Петербург: Издательство Санкт-Петербургского университета, 1998.
- [35] Ортега Дж. Введение в параллельные и векторные методы решения линейных систем. – Москва: Мир, 1991.
- [36] Островский А.М. Решение уравнений и систем уравнений. – Москва: Издательство иностранной литературы, 1963.
- [37] Прасолов В.В. Задачи и теоремы линейной алгебры. – Москва: Наука-Физматлит, 1996.
- [38] Райс Дж. Матричные вычисления и математическое обеспечение. – Москва: Мир, 1984.
- [39] Саад Ю. Итерационные методы для разреженных линейных систем. Учебное пособие в 2-х томах. – Москва: Издательство Московского университета, 2013–2014.
- [40] Самарский А.А., Гулин А.В. Численные методы. – Москва: Наука, 1989.
- [41] Стренг Г. Линейная алгебра и её применения. – Москва: Мир, 1980.
- [42] Тихонов А.Н., Арсенин В.Я. Методы решения некорректных задач. – Москва: Наука, 1979.
- [43] Тыртышников Е.Е. Матричный анализ и линейная алгебра. – Москва: Физматлит, 2007.
- [44] Тыртышников Е.Е. Методы численного анализа. – Москва: Академия, 2007.
- [45] Уилкинсон Дж. Алгебраическая проблема собственных значений. – Москва: Наука, 1970.
- [46] Уоткинс Д. Основы матричных вычислений. – Москва: «БИНОМ. Лаборатория знаний», 2009.
- [47] Фаддеев Д.К., Фаддеева В.Н. Вычислительные методы линейной алгебры. – Москва–Ленинград: Физматлит, 1960 (первое издание) и 1963 (второе издание).

- [48] ФЕДОРЕНКО Р.П. Итерационные методы решения разностных эллиптических уравнений // *Успехи Математических Наук*. – 1973. – Т. 28, вып. 2 (170). – С. 121–182.
- [49] ФОРСАЙТ Дж.Э. Что представляют собой релаксационные методы? // *Современная математика для инженеров под ред. Э.Ф.Беккенбаха*. – Москва: Издательство иностранной литературы, 1958. – С. 418–440.
- [50] ФОРСАЙТ Дж., МОЛЕР К. Численное решение систем линейных алгебраических уравнений. – Москва: Мир, 1969.
- [51] ХАУСДОРФ Ф. Теория множеств. – Москва: УРСС Эдиториал, 2007.
- [52] ХЕЙГЕМАН Л., ЯНГ Д. Прикладные итерационные методы. – Москва: Мир, 1986.
- [53] ХОРН Р., ДЖОНСОН Ч. Матричный анализ. – Москва: Мир, 1989.
- [54] ШИЛОВ Г.Е. Математический анализ. Конечномерные линейные пространства. – Москва: Наука, 1969.
- [55] ШИЛОВ Г.Е. Математический анализ. Функции одного переменного. Часть 3. – Москва: Наука, 1970.
- [56] АВЕРТН О. *Precise numerical methods using C++*. – San Diego: Academic Press, 1998.
- [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 — The Free Platform 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] АЛБЕРГ Дж., НИЛЬСОН Э., УОЛШ Дж. Теория сплайнов и её приложения. – Москва: Мир, 1972.
- [63] АЛЕКСАНДРОВ П.С. Введение в теорию множеств и общую топологию. – Санкт-Петербург: Лань, 2010.
- [64] АЛЕКСЕЕВ В.Б. Теорема Абеля в задачах и решениях. – Москва: Московский Центр непрерывного математического образования, 2001.
- [65] АЛЕКСЕЕВ Е.Р., ЧЕСНОКОВА О.В., РУДЧЕНКО Е.А. *Scilab. Решение инженерных и математических задач*. – Москва: Alt Linux – «БИНОМ. Лаборатория знаний», 2008.

- [66] АЛЕФЕЛЬД Г., ХЕРЦБЕРГЕР Ю. *Введение в интервальные вычисления*. – Москва: Мир, 1987.
- [67] БАБЕНКО К.И. *Основы численного анализа*. – Москва: Наука, 1986; Ижевск-Москва: 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*. – Москва: «БИНОМ. Лаборатория знаний», 2012.
- [70] ВИРО О.Я., ИВАНОВ О.А., НЕЦВЕТАЕВ Н.Ю., ХАРЛАМОВ В.М. *Элементарная топология*. – Москва: Московский центр непрерывного математического образования, 2010 и 2012.
- [71] ВОЕВОДИН В.В. *Численные методы алгебры. Теория и алгоритмы*. – Москва: Наука, 1966.
- [72] ВОЕВОДИН В.В. *Вычислительные основы линейной алгебры*. – Москва: Наука, 1977.
- [73] ВОЕВОДИН В.В., КУЗНЕЦОВ Ю.А. *Матрицы и вычисления*. – Москва: Наука, 1984.
- [74] ГАВРИКОВ М.Б., ТАЮРСКИЙ А.А. *Функциональный анализ и вычислительная математика*. – Москва: URSS, 2016.
- [75] 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.
- [76] ГОДУНОВ С.К. *Современные аспекты линейной алгебры*. – Новосибирск: Научная книга, 1997.
- [77] ГОРБАЧЕНКО В.И. *Вычислительная линейная алгебра с примерами на MATLAB*. – Санкт-Петербург: «БХВ-Петербург», 2011.
- [78] ДЖОРДЖ А., ЛЮ ДЖ. *Численное решение больших разреженных систем уравнений*. – Москва: Мир, 1984.
- [79] ДРОБИШЕВИЧ В.И., ДЫМНИКОВ В.П., РИВИН Г.С. *Задачи по вычислительной математике*. – Москва: Наука, 1980.
- [80] ЗЕЛЬДОВИЧ Я.Б., МЫШКИС А.Д. *Элементы прикладной математики*. – Москва: Наука, 1972.
- [81] ИКРАМОВ Х.Д. *Численное решение матричных уравнений*. – Москва: Наука, 1984.
- [82] КАЛИТКИН Н.Н. *Численные методы*. – Москва: Наука, 1978.
- [83] КРЫЛОВ А.Н. *Лекции о приближённых вычислениях*. – Москва: ГИТТЛ, 1954, а также более ранние издания.
- [84] КРЫЛОВ В.И., БОБКОВ В.В., МОНАСТЫРНЫЙ П.И. *Вычислительные методы высшей математики. Т. 1*. – Минск: «Вышэйшая школа», 1972.
- [85] КУБЛАНОВСКАЯ В.Н. О некоторых алгоритмах для решения полной проблемы собственных значений // *Журнал вычисл. матем. и мат. физики*. – 1961. – Т. 1, № 4. – С. 555–570.

- [86] Кузнецов Ю.А. Метод сопряжённых градиентов, его обобщения и применения // Вычислительные процессы и системы. – Москва: Наука, 1983 – Вып. 1. – С. 267–301.
- [87] Лагутин М.Б. *Наглядная математическая статистика*. 2-е изд. – Москва: «БИНОМ. Лаборатория знаний», 2011.
- [88] Лебедев В.И. *Функциональный анализ и вычислительная математика*. – Москва: Физматлит, 1989.
- [89] Марчук Г.И. *Методы вычислительной математики*. – Москва: Наука, 1989.
- [90] Парлетт Б. *Симметричная проблема собственных значений. Численные методы*. – Москва: Мир, 1983.
- [91] Пароди М. *Локализация характеристических чисел матриц и её применения*. – Москва: Издательство иностранной литературы, 1960.
- [92] Ремез Е.Я. *Основы численных методов чебышевского приближения*. – Киев: Наукова думка, 1969.
- [93] Самарский А.А., Николаев Е.С. *Методы решения сеточных уравнений*. – Москва: Наука, 1978.
- [94] Фаддеева В.Н. *Вычислительные методы линейной алгебры*. – Москва–Ленинград: Гостехиздат, 1950.
- [95] Флэнаган Д., Мацумото Ю. *Язык программирования Ruby*. – Санкт-Петербург: Питер, 2011.
- [96] Халмош П. *Конечномерные векторные пространства*. – Москва: ГИФМЛ, 1963.
- [97] Шарая И.А. IntLinIncR2 — пакет программ для визуализации множеств решений интервальных линейных систем с двумя неизвестными. Версия для MATLAB. 2014. http://www.nsc.ru/interval/Programing/MCodes/IntLinIncR2_UTF8.zip
- [98] Шарый С.П. *Конечномерный интервальный анализ*. – Электронная книга, 2012 (см. <http://www.nsc.ru/interval/Library/InteBooks>)
- [99] Яненко Н.Н. *Метод дробных шагов решения многомерных задач математической физики*. – Новосибирск: Наука, 1967.
- [100] BAUER F.L., FIKE C.T. Norms and exclusion theorems // *Numerische Mathematik*. – 1960. – Vol. 2. – P. 137–141.
- [101] ECKART C., YOUNG G. The approximation of one matrix by another of lower rank // *Psychometrika*. – 1936. – Vol. 1. – P. 211–218.
- [102] GREGORY R.T., KARNEY D.L. *A collection of matrices for testing computational algorithms*. – Hantington, New York: Robert E. Krieger Publishing Company, 1978.
- [103] 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.
- [104] KREINOVICH V., LAKEYEV A.V, NOSKOV S.I. Approximate linear algebra is intractable // *Linear Algebra and its Applications*. – 1996. – Vol. 232. – P. 45–54.

- [105] KREINOVICH V., LAKEYEV A.V., ROHN J., KAHL P. *Computational complexity and feasibility of data processing and interval computations*. – Dordrecht: Kluwer, 1997.
- [106] MAYER G. *Interval analysis and automatic result verification*. – Berlin: De Gruyter, 2017.
- [107] MOLER C. Professor SVD // *The MathWorks News & Notes*. – October 2006. – P. 26–29.
- [108] MOORE R.E., KEARFOTT R.B., CLOUD M. *Introduction to interval analysis*. – Philadelphia: SIAM, 2009.
- [109] 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.
- [110] SCHULZ G. Iterative Berechnung der reziproken Matrix // *Z. Angew. Math. Mech.* – 1933. – Bd. 13 (1). – S. 57–59.
- [111] STOER J., BULIRSCH R. *Introduction to numerical analysis*. – Berlin-Heidelberg-New York: Springer-Verlag, 1993.
- [112] 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.
- [113] TURING A.M. Rounding-off errors in matrix processes // *Quarterly Journal of Mechanics and Applied Mathematics*. – 1948. – Vol. 1. – P. 287–308.
- [114] 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.
- [115] VARGA R.S. On diagonal dominance arguments for bounding $\|A^{-1}\|_{\infty}$ // *Linear Algebra and its Applications*. – 1976. – Vol. 14. – P. 211–217.
- [116] VARGA R.S. *Matrix iterative analysis*. – Berlin, Heidelberg, New York: Springer Verlag, 2000, 2010.
- [117] 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.
- [118] 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 системы уравнений

$$\left\{ \begin{array}{lcl} F_1(x_1, x_2, \dots, x_n) & = & 0, \\ F_2(x_1, x_2, \dots, x_n) & = & 0, \\ \vdots & \ddots & \vdots \\ F_n(x_1, x_2, \dots, x_n) & = & 0, \end{array} \right. \quad (4.1)$$

над полем вещественных чисел \mathbb{R} , или, кратко,

$$F(x) = 0, \quad (4.2)$$

где $x = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ — вектор неизвестных переменных,

$F_i(x)$, $i = 1, 2, \dots, n$, — вещественнозначные функции,

$F(x) = (F_1(x), F_2(x), \dots, F_n(x))^T$ — вектор-столбец функций F_i .

Для переменных x_1, x_2, \dots, x_n нужно найти набор значений $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n$, называемый *решением системы*, который одновременно обращает в равенства все уравнения системы (4.1). В некоторых случаях желательно найти все такие возможные наборы, т. е. все решения системы,

а иногда достаточно какого-то одного. В случае, когда система уравнений (4.1)–(4.2) не имеет решений, нередко требуется предоставить обоснование этого заключения 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 приближениями функций, вычислением интегралов, задача решения уравнений и систем уравнений является одной из классических задач вычислительной математики.

Всюду далее мы предполагаем, что функции $F_i(x)$ по меньшей мере непрерывны, а количество уравнений в системе (4.1)–(4.2) совпадает с количеством неизвестных переменных. Помимо записи систем уравнений в каноническом виде (4.1)–(4.2) часто встречаются и другие формы их представления, например,

$$G(x) = H(x) \quad (4.3)$$

с какими-то функциями G, H . Чрезвычайно важным частным случаем этой формы является *рекуррентный вид* системы уравнений (или одного уравнения),

$$x = G(x). \quad (4.4)$$

в котором неизвестная переменная выражена через саму себя. В этом случае решение системы уравнений (или уравнения) есть *неподвижная точка* отображения G , т. е. такой элемент области определения G , который переводится этим отображением сам в себя. Кроме того, рекуррентный вид уравнения или системы хорош тем, что позволяет довольно просто организовать итерационный процесс для нахождения решения, что 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 (4.1)–(4.2) с неподвижными точками отображений, т. е. с решениями уравнений в рекуррентном виде (4.4). Ясно, что

$$F(x) = 0 \quad \Longleftrightarrow \quad x = x - \Lambda F(x),$$

где Λ — 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 весьма общую природу, то математические свойства уравнений (4.2) и (4.4) могут существенно различаться, так что при этом формы записи (4.2) и (4.4), строго говоря, не вполне равносильны друг другу. По этой причине для их обозначения часто употребляют отдельные термины — *уравнение первого рода* и *уравнение второго рода* соответственно.

Обращаясь к решению нелинейных уравнений и их систем, мы обнаруживаем себя в гораздо более сложных условиях, нежели при решении систем линейных алгебраических уравнений (3.60)–(3.61). Стройная и весьма полная теория разрешимости систем линейных уравнений, базирующаяся на классических результатах линейной алгебры, обеспечивала в необходимых нам случаях уверенность в существовании решения систем 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», и потому отличаются огромным разнообразием. Из общих нелинейных уравнений и систем уравнений принято выделять *алгебраические* уравнения и системы уравнений, в которых функции $F_i(x)$ являются алгебраическими полиномами относительно неизвестных переменных x_1, x_2, \dots, x_n .

4.2 Вычислительно-корректные задачи

4.2a Предварительные сведения и определения

Напомним общеизвестный факт: 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} . Для цифровых вычислительных машин это заключение следует 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,
- 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, которые поэтому подразумеваются целыми или, на худой конец, рациональными с произвольной длиной числителя и знаменателя (см. [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, т. е. устойчивы по отношению к возмущениям в этих начальных 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.

Под *массовой задачей* [13] будем понимать некоторый общий вопрос, формулировка которого содержит несколько свободных переменных — *параметров* — могущих принимать значения в пределах предписанных им множеств. В целом массовая задача Π определяется

- 1) указанием её входных данных, т. е. общим списком всех *параметров* с областями их определения,
- 2) формулировкой тех свойств, которым должен удовлетворять *ответ*, т. е. решение задачи.

Индивидуальная задача I получается из массовой задачи Π путём присваивания всем параметрам задачи Π каких-то конкретных значений. Наконец, *разрешающим отображением* задачи Π мы называем отображение, сопоставляющее каждому набору входных данных-параметров ответ соответствующей индивидуальной задачи (см. §1.6). 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.

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 §3.16в для собственных значений некоторых матриц, которые являются «практически разрывными» функциями элементов матрицы. 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. §3.5а), оценивающие сверху границы изменения решений x через изменения элементов матрицы A , правой части b и число обусловленности матрицы A . ■

Пример 4.2.2 Вычисление ранга матрицы — вычислительно некорректная задача. Дело в том, что в основе понятия ранга лежит линейная зависимость строк или столбцов матрицы, т. е. свойство, которое нарушается при сколь угодно малых возмущениях 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* [8]. Напомним её определение.

Если задано метрическое пространство с метрикой ϱ , 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 целыми множествами.

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

$$x^2 + px + q = 0 \tag{4.5}$$

для

$$p^2 = 4q \quad (4.6)$$

имеет лишь одно решение $x = -p/2$. Но при любых сколь угодно малых возмущениях коэффициента p и свободного члена q , нарушающих равенство (4.6), уравнение (4.5) теряет это единственное решение или же приобретает ещё одно (см. Рис. 4.1).

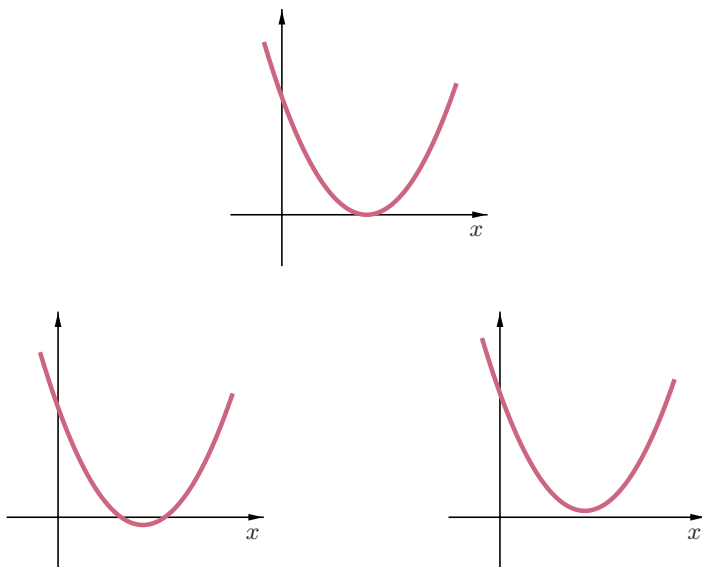


Рис. 4.1. Неустойчивая зависимость решений уравнения (4.5)–(4.6) от сколь угодно малых шевелений его коэффициентов.

Аналогичным образом ведёт себя решение двумерной системы уравнений, эквивалентной (4.5),

$$\begin{cases} x + y = r, \\ xy = s \end{cases}$$

при $s = r^2/4$. При этом раздвоение решения не является большим грехом, коль скоро мы можем рассматривать хаусдорфово расстояние между целостными множествами решений. Но вот исчезновение единственного решения, при котором расстояние между множествами реше-

ний скачком меняется до $+\infty$ — это чрезвычайное событие, однозначно указывающее на разрывность разрешающего отображения.

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 выписывая соотношение

$$F(x) = 0 \quad (4.2)$$

и ему подобные, имеют весьма различную природу. Это и будет отправной точкой нашей ревизии постановки задачи решения уравнений и систем уравнений.

4.2в ε -решения уравнений

В ряде практических задач пользователю требуется не точное равенство некоторого выражения нулю, а лишь его «исчезающая малость» в сравнении с каким-то а priori установленным порогом. С аналогичной точки зрения часто имеет смысл рассматривать соотношения вида (4.3) или (4.4), выражающие 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 самом 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 точностью, превосходящей 0.001 градуса, вообще пробле-

матично в силу принципиальных соображений. Наконец, ограниченная точность, с которой известны абсолютно все физические константы¹, также воздвигает 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 (4.2) 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 ε , т. е. $\|F(x)\| < \varepsilon$.

Решением этой задачи является, как правило, множество точек, которые мы будем называть ε -решениями или *почти решениями*, если порог этой пренебрежимой малости не оговорён явно или несуществен. Фактически, понятие *псевдорешения* уравнений и систем уравнений, которое рассматривалось в Главах 2 и 3, является дальнейшим развитием и обобщений ε -решений и почти решений.

Нетрудно понять, 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 «о нахождении почти решений» является вычислительно-корректной.

Как уже отмечалось выше, в некоторых задачах система уравнений более естественно записывается не как (4.2), а в виде (4.3)

$$G(x) = H(x),$$

¹В лучшем случае относительная погрешность известных на сегодняшний день значений физических констант равна 10^{-10} , см. [66].

и требуется обеспечить с относительной погрешностью ε равенство её левой и правой частей:

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 ε , т. е.

$$\frac{\|G(x) - H(x)\|}{\max\{\|G(x)\|, \|H(x)\|\}} < \varepsilon .$$

Решения этой задачи мы тоже будем называть ε -решениями системы уравнений вида (4.3).

Математические понятия, определения которых привлекают малый допуск ε , не являются чем-то экзотическим. Таковы, к примеру, ε -энтропия множеств в метрических пространствах, ε -субдифференциал функции, ε -оптимальные решения задач оптимизации и т. п. Одним из частных случаев ε -решений являются точки ε -спектра матрицы, предложенные для обобщения традиционного понятия собственного значения матрицы [11, 49, 67]. Говорят, что точка z на комплексной плоскости принадлежит ε -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 ε » равенство $Av = zv$.

4.2г Недостаточность ε -решений

Но есть и принципиально другой тип задач, которые образно могут быть названы задачами «об определении перехода через нуль» и не сводятся к нахождению ε -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.

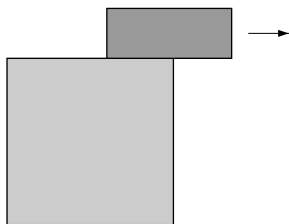


Рис. 4.2. Когда кирпич упадёт с подставки?

Рассмотрим следующую ситуацию, для анализа которой достаточно знание элементарной физики. Пусть кирпич лежит на опоре (см. Рис. 4.2), 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 ε -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 систему линейных дифференциальных уравнений с постоянными коэффициентами

$$\frac{dx}{dt} = Ax, \quad (4.7)$$

матрица которой $A = A(\theta)$ зависит от параметра θ (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 (4.7) устойчивы по Ляпунову (и даже

асимптотически устойчивы). При каких значениях параметра θ рассматриваемая система делается неустойчивой?

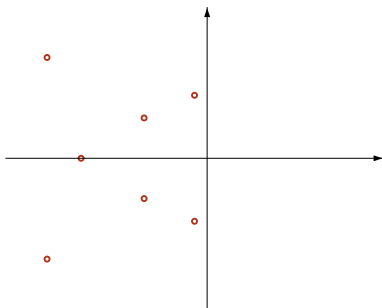


Рис. 4.3. Срыв устойчивости в динамической системе (4.7) происходит, когда собственные значения A «переходят» через мнимую ось.

Традиционно отвечают на этот вопрос следующим образом. Срыв устойчивости в системе (4.7) произойдет при $\operatorname{Re} \lambda(A(\theta)) = 0$ 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 ϵ -решения из §4.2в или ϵ -спектра из работ [11, 49, 67].

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», см. [40]). И здесь принципиален именно переход через некоторый порог, а не

близость к нему, на которую делается упор в понятиях ϵ -решения и ϵ -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 Векторные поля и их вращение

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* Φ . 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, [8, 52]), 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 является *непрерывным*, если непрерывно отображение $\Phi(x)$. Например, на Рис. 4.4 изображены векторные поля

$$\Phi(x) = \Phi(x_1, x_2) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \text{и} \quad \Psi(x) = \Psi(x_1, x_2) = \begin{pmatrix} x_1 \\ -x_2 \end{pmatrix}, \quad (4.8)$$

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.

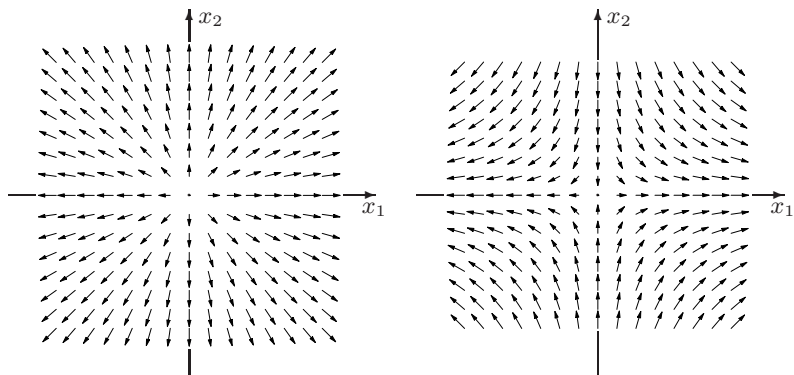


Рис. 4.4. Векторные поля $\Phi(x)$ и $\Psi(x)$, задаваемые формулами (4.8).

Связь векторных полей и их особых точек с основным предметом этой главы очевидна: особая точка поля $\Phi: M \rightarrow \mathbb{R}^n$ — это решение системы n уравнений

$$\begin{cases} \Phi_1(x_1, x_2, \dots, x_n) = 0, \\ \Phi_2(x_1, x_2, \dots, x_n) = 0, \\ \vdots \quad \ddots \quad \vdots \\ \Phi_n(x_1, x_2, \dots, x_n) = 0, \end{cases}$$

лежащее в M . Будем говорить, что векторное поле Φ вырождено, если у него есть особые точки. Иначе Φ называется невырожденным. К примеру, векторные поля Рис. 4.4 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 Φ . 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 $\Delta(\lambda, x)$ 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.

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.

4.36 Вращение векторных полей

Пусть D — ограниченная область в \mathbb{R}^n с границей ∂D . Через $\text{cl } D$ мы обозначим её топологическое замыкание. Оказывается, каждому невырожденному на ∂D векторному полю Φ можно сопоставить целочисленную характеристику — *вращение векторного поля Φ на ∂D* , — обозначаемую $\gamma(\Phi, D)$ и удовлетворяющую следующим условиям:

(А) Гомотопные на ∂D векторные поля имеют одинаковое вращение.

- (В) Пусть $D_i, i = 1, 2, \dots$, — непересекающиеся области, лежащие в D (их может быть бесконечно много). Если непрерывное векторное поле Φ 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

$$\gamma(\Phi, D) = \gamma(\Phi, D_1) + \gamma(\Phi, D_2) + \dots$$

- (С) Если $\Phi(x) = x - a$ для некоторой точки $a \in D$, то вращение Φ на ∂D равно $(+1)$, т. е.

$$\gamma(\Phi, D) = 1.$$

Нетрудно понять, что определённая так величина вращения поля устойчива к малым шевелениям как области (это следует из (В)), так и векторного поля (это вытекает из (А)).

Условиями (А)–(В)–(С) вращение векторного поля задаётся однозначно, но недостаток такого определения — в отсутствии конструктивности. Можно показать (см. подробности в [62]), что сформулированное определение равносильно следующему конструктивному. Зафиксируем некоторую параметризацию поверхности ∂D , т. е. задание её в виде

$$x_1 = x_1(u_1, u_2, \dots, u_{n-1}),$$

$$x_2 = x_2(u_1, u_2, \dots, u_{n-1}),$$

$$\vdots \quad \quad \quad \ddots \quad \quad \quad \vdots$$

$$x_n = x_n(u_1, u_2, \dots, u_{n-1}),$$

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 ∂D области D равно значению поверхностного интеграла

$$\frac{1}{S_n} \int_{\partial D} \frac{1}{\|\Phi(x)\|^n} \cdot \det \begin{pmatrix} \Phi_1(x) & \frac{\partial \Phi_1(x)}{\partial u_1} & \dots & \frac{\partial \Phi_1(x)}{\partial u_{n-1}} \\ \Phi_2(x) & \frac{\partial \Phi_2(x)}{\partial u_1} & \dots & \frac{\partial \Phi_2(x)}{\partial u_{n-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_n(x) & \frac{\partial \Phi_n(x)}{\partial u_1} & \dots & \frac{\partial \Phi_n(x)}{\partial u_{n-1}} \end{pmatrix} du_1 du_2 \dots du_n, \quad (4.9)$$

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, совершаемое при движении точки аргумента в положительном направлении по рассматриваемой границе области [52, 57, 58, 30, 60]. В многомерном случае такой наглядности уже нет, но величина вращения векторного поля Φ всё равно может быть истолкована как «число раз, которое отображение $\Phi : \partial D \rightarrow \Phi(\partial D)$ накрывает образ $\Phi(\partial D)$ ».

Рассмотрим примеры. На любой окружности с центром в нуле поле, изображенное на левой половине Рис. 4.4, имеет вращение $+1$, а поле на правой половине Рис. 4.4 — вращение -1 . Векторные поля Рис. 4.5, 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 $x = (x_1, x_2)$, ψ — 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* [57, 58, 59, 30, 31, 62, 60]. Именно, вращение поля Φ на границе области D есть топологическая степень такого отображения ϕ границы ∂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 характеристики «прохождения через нуль» многомерной функции, и теоретической основой этого шага служат следующие результаты:

Предложение 4.3.1 [57, 58, 30, 60] *Если векторное поле Φ невырождено на замыкании ограниченной области D , то вращение $\gamma(\Phi, D) = 0$.*

Теорема 4.3.1 (теорема Кронекера) [57, 58, 30] *Пусть векторное поле Φ невырождено на границе ограниченной области D и непрерывно на её замыкании. Если $\gamma(\Phi, D) \neq 0$, то поле Φ имеет в D по крайней мере одну особую точку.*

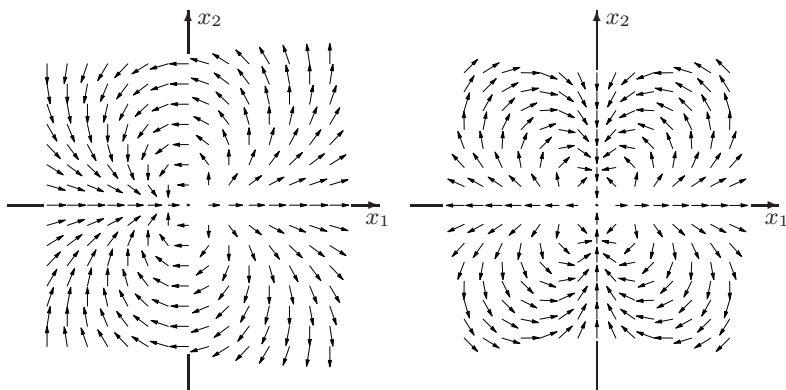


Рис. 4.5. Векторные поля, имеющие вращения $+2$ (левый чертёж) и $+3$ (правый чертёж) на любой окружности с центром в нуле.

Теорема Кронекера обладает большой общностью, но требует для своей проверки и большой вычислительной работы. Часто она 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 ± 1 .

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 Φ или *индексом нуля* \tilde{x} поля Φ , 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 [57, 58, 30] *Если A — невырожденное линейное преобразование пространства \mathbb{R}^n , то его единственная особая точка — нуль — имеет индекс $\text{ind}(0, A) = \text{sgn det } A$.*

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 [57, 58, 30] *Если \tilde{x} — регулярная особая точка дифференцируемого векторного поля Φ , то $\text{ind}(\tilde{x}, \Phi) = \text{sgn det } \Phi'(\tilde{x})$.*

Таким образом, регулярные (не критические) особые точки векторных полей имеют индекс ± 1 , а в прочих случаях значение индекса может быть весьма произвольным.

Например, индексы расположенного в начале координат нуля векторных полей, которые изображены на Рис. 4.4, равны $+1$ и (-1) , причём поля эти всюду дифференцируемы. Индексы нуля полей Рис. 4.5 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 Φ называется устойчивой, если для любого $\tau > 0$ можно найти такое $\eta > 0$, что всякое поле, отличающееся от Φ меньше чем на η , имеет особую точку, удалённую от z менее, чем на τ . Иначе особая точка z называется неустойчивой.*

Легко понять, что в связи с задачей решения систем уравнений нас интересуют именно устойчивые особые точки, поскольку задача поиска только таких точек является вычислительно-корректной.

Вторым основным результатом, ради которого мы затевали обзор теории вращения векторных полей, является следующее

Предложение 4.3.4 [58] *Изолированная особая точка непрерывного векторного поля устойчива тогда и только тогда, когда её индекс отличен от нуля.*

Например, неустойчивое решение квадратного уравнения (4.5)–(4.6) имеет индекс 0, а у векторных полей, изображённых на рисунках 4.4 и 4.5, 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 вещах: 1) в линейных задачах индекс решения, как правило, равен ± 1 , а в нелинейных может быть как нулевым, так и отличным от ± 1 , и, как следствие, 2) 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, путем шевеления одной устойчивой особой точки можно получить сразу *несколько* особых точек, и это легко видеть на примере полей Рис. 4.5. Любая сколь угодно малая постоянная добавка к полю, изображённому на левом чертеже Рис. 4.5, 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. 4.5, приводит к распадению нулевой особой точки на три особые точки индекса 1. Таким образом, свойство единственности решения 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 Φ имеет в D конечное число особых точек x_1, x_2, \dots, x_s и невырождено на границе ∂D . Тогда*

$$\gamma(\Phi, D) = \text{ind}(x_1, \Phi) + \text{ind}(x_2, \Phi) + \dots + \text{ind}(x_s, \Phi).$$

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.

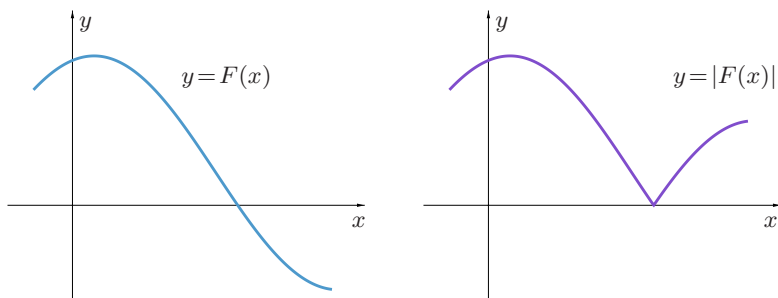


Рис. 4.6. Устойчивый нуль функции превращается в неустойчивый после взятия нормы функции.

Наконец, сделаем ещё одно важное замечание. Нередко на практике для решения систем нелинейных уравнений исходную задачу переформулируют как оптимизационную, пользуясь, например, тем, что справедливы следующие математические эквивалентности:

$$F(x) = 0 \quad \Leftrightarrow \quad \min_x \|F(x)\| = 0$$

и

$$F(x) = 0 \quad \Leftrightarrow \quad \min_x \|F(x)\|^2 = 0.$$

Далее имеющимися стандартными пакетами программ ищется решение задачи минимизации нормы $\|F(x)\|$ (или $\|F(x)\|^2$, чтобы обеспечить гладкость целевой функции) и результат сравнивается с нулём. 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 общем случае), так что нуль функции исчезнет.

4.3д Вычислительно-корректная постановка задачи

Теперь все готово для вычислительно-корректной переформулировки задачи решения уравнений и систем уравнений. Она должна выгля-

деть следующим образом:

Для заданного $\varepsilon > 0$ и системы уравнений

$$F(x) = 0$$

найти на данном множестве $D \subseteq \mathbb{R}^n$

- 1) гарантированные двусторонние границы
всех решений ненулевого индекса,
- 2) множество ε -решений.

(4.10)

Мы не требуем единственности решения в выдаваемых брусах, так как свойство решения быть единственным не является, как мы могли видеть, устойчивым к малым возмущениям задачи.

4.4 Классические методы решения уравнений

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, в форме дуги окружности. Балки, поддерживающие такую кровлю, должны иметь форму круговых сегментов, и для того, чтобы придать им правильную форму, нужно знать, какой именно радиус закругления крыши при этом получится (см. Рис. 4.7).

Обозначим искомый радиус крыши через R . Если 2α — угловая величина дуги (в радианах), соответствующей крыше, то

$$\frac{l}{2\alpha} = R.$$

С другой стороны, из рассмотрения прямоугольного треугольника с катетом $h/2$ и гипотенузой R получаем

$$R \sin \alpha = h/2.$$

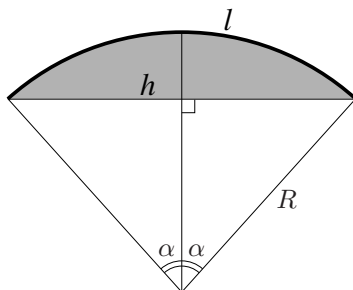


Рис. 4.7. Проектирование круглой крыши.

Исключая из этих двух соотношений R , получим уравнение относительно одной неизвестной α :

$$l \sin \alpha = \alpha h. \quad (4.11)$$

Решив его, легко найдём и радиус закругления крыши R .

Но решение уравнения (4.11) не может быть выражено в виде какой-либо явной конечной формулы от l и h , и потому далее мы обсудим возможности его численного решения. ■

Уравнение (4.11) является простейшим нелинейным *трансцендентным* уравнением. Так называют уравнения и системы уравнений, не являющиеся алгебраическими, т.е. такие, в которых в обеих частях уравнений не стоят алгебраические выражения относительно неизвестных переменных.

4.4а Предварительная локализация решений

Обычно первым этапом численного решения уравнений и систем уравнений является предварительная локализация, т.е. уточнение местонахождения искомых решений. Это вызвано тем, что большинство численных методов для поиска решений имеют локальный характер, т.е. сходятся к этим решениям лишь из достаточно близких начальных приближений.

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 совершенно пренебрегать и *графическими* методами

локализации решений, основанными на построении и исследовании графиков функций, которые фигурируют в уравнении. Графические методы не обладают большой строгостью и полнотой, но очень наглядны и тоже способны внести свою лепту в практическое решение уравнений.

Особенно много аналитических результатов существует о локализации решений алгебраических уравнений (корней полиномов), что, конечно, имеет причину в очень специальном виде этих уравнений, допускающем исследование с помощью аналитических выкладок и т. п. инструментов.

Теорема 4.4.1 Пусть для алгебраического уравнения вида

$$a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0 = 0$$

обозначено

$$\alpha = \max\{a_0, \dots, a_{n-1}\}, \quad \beta = \max\{a_1, \dots, a_n\}.$$

Тогда все решения этого уравнения принадлежат кольцу в комплексной плоскости, определяемому условием

$$\frac{1}{1 + \beta/|a_0|} \leq |x| \leq 1 + \frac{\alpha}{|a_n|}.$$

Полезно правило знаков Декарта, утверждающее, что число положительных корней полинома с вещественными коэффициентами равно числу перемен знаков в ряду его коэффициентов или на чётное число меньше этого числа. При этом корни считаются с учётом кратности, а нулевые коэффициенты при подсчёте числа перемен знаков не учитываются. Если, к примеру, заранее известно, что все корни данного полинома вещественны, то правило знаков Декарта даёт точное число корней. Рассматривая полином с переменной $(-x)$ можно с помощью этого же результата найти число отрицательных корней исходного полинома.

4.46 Метод дихотомии

Этот метод часто называют также *методом бисекции* или *методом половинного деления*. Он заключается в последовательном делении пополам интервала локализации корня уравнения, на концах которого функция принимает значения разных знаков. Теоретической основой метода дихотомии является следующий факт, хорошо известный в математическом анализе:

Теорема 4.4.2 (теорема Больцано-Коши) *Если функция $f : \mathbb{R} \rightarrow \mathbb{R}$ непрерывна на интервале $X \subset \mathbb{R}$ и на его концах принимает значения 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» (см., к примеру, [42]), 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 переменная знака обязана сохраняться. Её мы и оставляем в результате очередной итерации метода дихотомии.

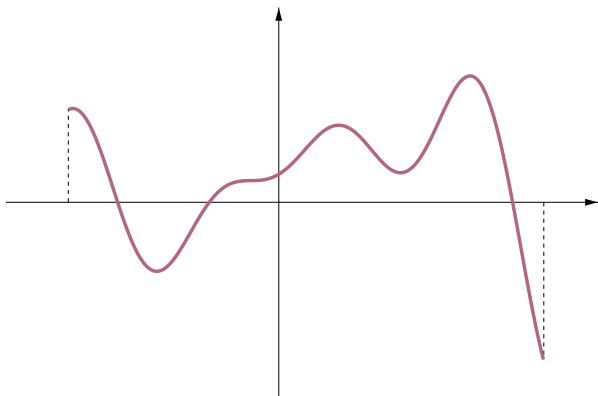


Рис. 4.8. Иллюстрация метода дихотомии (половинного деления).

На вход алгоритму подаются функция f , принимающая на концах интервала $[a, b]$ значения разных знаков, и точность ϵ , с которой необходимо локализовать решение уравнения $f(x) = 0$. На выходе получаем интервал $[\underline{x}, \bar{x}]$ шириной не более ϵ , содержащий решение уравнения.

Недостаток этого простейшего варианта метода дихотомии — возможность потери решений для функций, аналогичных изображенной на Рис. 4.8. На левой половине исходного интервала функция знака не меняет, 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 фигурирующей в нём функции. В общем случае потери нулей можно избежать, если не отбрасывать подынтервалы, на которых доказательно не установлено отсутствие решений. Последовательная реализация

Таблица 4.1. Метод дихотомии решения уравнений

```

 $\underline{x} \leftarrow a; \quad \overline{x} \leftarrow b;$ 
DO WHILE ( $\overline{x} - \underline{x} > \epsilon$ )
     $y \leftarrow \frac{1}{2}(\underline{x} + \overline{x});$ 
    IF ( $f(\underline{x}) < 0$  и  $f(y) > 0$ ) или ( $f(\underline{x}) > 0$  и  $f(y) < 0$ )
         $\overline{x} \leftarrow y$ 
    ELSE
         $\underline{x} \leftarrow y$ 
    END IF
END DO

```

этой идеи приводит к «методу ветвлений и отсечений», который подробно рассматривается далее в §4.8.

Многомерное обобщение теоремы Больцано-Коши было опубликовано более чем столетием позже в заметке [46]:

Теорема 4.4.3 (теорема Миранды) Пусть $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$, $f(x) = (f_1(x), f_2(x), \dots, f_n(x))^T$ — функция, непрерывная на брус $\mathbf{X} \subset \mathbb{R}^n$ со сторонами, параллельными координатным 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 брус \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 брусках. Удобное средство для решения этой задачи предоставляют методы интервального анализа. Задача об определении области значений функции на брусках из области её определения эквивалентна задаче оптимизации, но в интервальном анализе она принимает специфическую форму задачи о вычислении так называемого *интервального расширения функции* (см. §1.5).

4.4в Метод простой итерации

Методом простой итерации обычно называют стационарный одношаговый итерационный процесс, который организуется после того, как исходное уравнение каким-либо способом приведено к равносильному рекуррентному виду $x = \Phi(x)$. Далее, после выбора некоторого начального приближения $x^{(0)}$, запускается итерационный процесс

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

При благоприятных обстоятельствах последовательность $\{x^{(k)}\}$ сходится, и её пределом является решение исходного уравнения. Но в общем случае и характер сходимости, и вообще её наличие существенно зависят как от отображения Φ , так и от начального приближения к решению.

Пример 4.4.2 Уравнение (4.11) 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

$$\alpha^{(k+1)} \leftarrow \frac{l}{h} \sin \alpha^{(k)}, \quad k = 0, 1, 2, \dots, \quad (4.12)$$

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).

Итерационный процесс (4.12) сходится к решению α^* не из любого начального приближения. Если $\alpha^{(0)} = \pi l$, $l \in \mathbb{Z}$, то выполнение итераций (4.12) с идеальной точностью даёт $\alpha^{(k)} = 0$, $k = 1, 2, \dots$. Если же $\alpha^{(0)}$ таково, что синус от него отрицателен, то итерации (4.12) сходятся к решению $(-\alpha^*)$ уравнения (4.11). И нулевое, и отрицательное решения очевидно не имеют содержательного смысла.

С другой стороны, переписывание исходного уравнения (4.12) 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

$$\alpha^{(k+1)} \leftarrow \frac{1}{l} \arcsin(\alpha^{(k)} h), \quad k = 0, 1, 2, \dots,$$

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 α^* 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*, если существует такая окрестность Ω точки x^* , что итерационный процесс $x^{(k+1)} \leftarrow \Phi(x^{(k)})$ сходится к x^* из любого начального приближения $x^{(0)} \in \Omega$.

Неподвижная точка x^* функции $\Phi(x)$ называется *отталкивающей*, если существует такая окрестность Ω точки x^* , что итерационный процесс $x^{(k+1)} \leftarrow \Phi(x^{(k)})$ не сходится к x^* при любом начальном приближении $x^{(0)} \in \Omega$.

Ясно, что простые итерации $x^{(k+1)} \leftarrow \Phi(x^{(k)})$ 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, и самыми популярными из них являются так называемые условия сжимаемости (сжатия) образа.

Напомним, что отображение $g : X \rightarrow X$ метрического пространства X с расстоянием $\text{dist} : X \rightarrow \mathbb{R}_+$ называется *сжимающим* (или просто *сжатием*), если существует такая положительная постоянная $\alpha < 1$, что для любой пары элементов $x, y \in X$ имеет место неравенство

$$\text{dist}(g(x), g(y)) \leq \alpha \cdot \text{dist}(x, y).$$

Теорема 4.4.4 (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 методом последовательных приближений*

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

при любом начальном приближении $x^{(0)} \in X$.

Доказательство этого результата можно найти, к примеру, в [16, 18, 20, 24, 31]. Особенно ценен в теореме Банаха её конструктивный характер, позволяющий организовать численные методы для нахождения неподвижной точки.

Иногда бывает полезно работать с векторнозначным расстоянием — *мультиметрикой*, — которая вводится на \mathbb{R}^n как

$$\text{Dist}(x, y) := \begin{pmatrix} \text{dist}(x_1, y_1) \\ \vdots \\ \text{dist}(x_n, y_n) \end{pmatrix} \in \mathbb{R}_+^n. \quad (4.13)$$

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 4.4.1 *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*

$$\text{Dist}(g(x), g(y)) \leq P \cdot \text{Dist}(x, y). \quad (4.14)$$

Следует отметить, что математики, к сожалению, не придерживаются здесь единой терминологии. Ряд авторов (см. [48]) за матрицей P из (4.14) закрепляют отдельное понятие «оператора Липшица (матрицы Липшица) отображения g », и в условиях Определения 4.4.1 говорят, что «оператор Липшица для g сжимающий».

Теорема 4.4.5 (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 Dist . Тогда для любого $x^{(0)}$ последовательность итераций*

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

сходится к единственной неподвижной точке x^ отображения g в X и имеет место оценка*

$$\text{Dist}(x^{(k)}, x^*) \leq (I - P)^{-1} P \cdot \text{Dist}(x^{(k)}, x^{(k-1)}).$$

Доказательство можно найти, например, в книгах [1, 19, 31, 48]

4.4г Метод Ньютона и его модификации

Предположим, что для уравнения $f(x) = 0$ с вещественнозначной функцией f известно некоторое приближение \tilde{x} к решению x^* . Если f — плавно меняющаяся (гладкая функция), то естественно приблизить её в окрестности точки \tilde{x} линейной функцией, т.е.

$$f(x) \approx f(\tilde{x}) + f'(\tilde{x})(x - \tilde{x}),$$

и далее для вычисления следующего приближения к x^* решать линейное уравнение

$$f(\tilde{x}) + f'(\tilde{x})(x - \tilde{x}) = 0.$$

Отсюда очередное приближение к решению

$$x = \tilde{x} - \frac{f(\tilde{x})}{f'(\tilde{x})}.$$

Итерационный процесс

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

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).

В англоязычной литературе метод Ньютона иногда называют также «методом Ньютона-Рафсона» (см., к примеру, [23, 61]), так как именно Дж. Рафсон придал форму, близкую к современной, 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.

Пример 4.4.3 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

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})} = x^{(k)} - \frac{(x^{(k)})^2 - a}{2x^{(k)}} = \frac{x^{(k)}}{2} + \frac{a}{2x^{(k)}}.$$

Итерационный процесс для нахождения \sqrt{a} , определяемый как

$$x^{(k+1)} \leftarrow \frac{1}{2} \left(x^{(k)} + \frac{a}{x^{(k)}} \right), \quad k = 0, 1, 2, \dots,$$

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 неприемлемым или труднодостижимым. Одна из очевидных модификаций метода Ньютона состоит в том, чтобы «заморозить» производную в некоторой точке и вести итерации по формуле

$$x^{(k+1)} \leftarrow x^{(k)} - \frac{f(x^{(k)})}{f'(\tilde{x})}, \quad k = 0, 1, 2, \dots,$$

где \tilde{x} — фиксированная точка, в которой берётся производная. Получаем стационарный итерационный процесс, который существенно проще в реализации, но он имеет качественно более медленную сходимость.

Несмотря на большую популярность метода Ньютона и его хорошие свойства, существуют примеры его плохого и даже патологического поведения.

Пример 4.4.4 (пример Донована-Миллера-Морэланда [64])

Рассмотрим численное решение, с помощью метода Ньютона, уравнения

$$h(x) = \sqrt[3]{x} e^{-x^2} = 0.$$

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 $|x^{(k+1)} - x^{(k)}| < \epsilon$ для заданного малого порога ϵ .

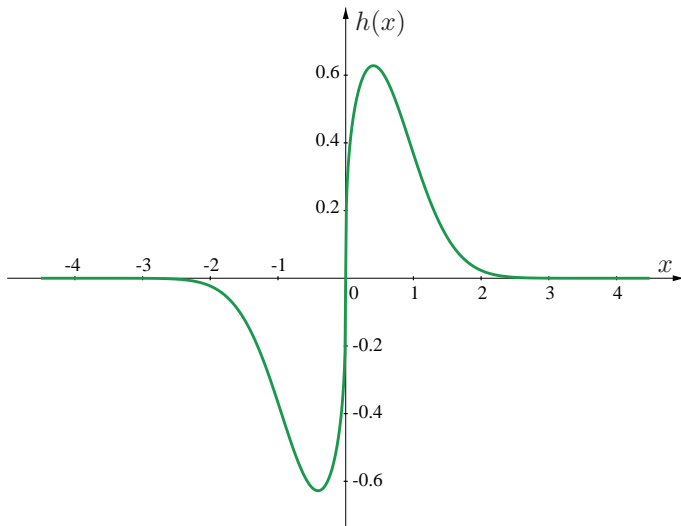


Рис. 4.9. График функции из примера Донована-Миллера-Морэланда.

В самом деле, итерации метода Ньютона в данном случае имеют

вид

$$x^{(k+1)} = x^{(k)} - \frac{h(x^{(k)})}{h'(x^{(k)})} = x^{(k)} - \frac{x^{(k)}}{\frac{1}{3} - 2(x^{(k)})^2}, \quad (4.15)$$

и они определены всюду за исключением стационарных точек функции $h(x)$ (точек зануления её производной), равных $\pm 1/\sqrt{6}$. Если $x^{(k)} \in] -1/\sqrt{6}, 1/\sqrt{6}[$ и $x^{(k)} \neq 0$, то

$$\left| \frac{x^{(k+1)}}{x^{(k)}} \right| = \left| 1 - \frac{1}{\frac{1}{3} - 2(x^{(k)})^2} \right| > 2.$$

Как следствие, последовательность, порождаемая методом Ньютона (4.15) 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 $x^{(k)} \notin [-1/\sqrt{6}, 1/\sqrt{6}]$ достаточно проводить лишь в случае $x^{(k)} > 0$, т. е. когда

$$x^{(k)} > \frac{1}{\sqrt{6}}.$$

Для отрицательных $x^{(k)}$ рассуждения будут аналогичными из-за нечётности функции $h(x)$.

При $x^{(k)} > 1/\sqrt{6}$ имеем $\frac{1}{3} - 2(x^{(k)})^2 < 0$, а потому в итерациях метода Ньютона (4.15) последовательные приближения монотонно возрастают, т. е. $x^{(k+1)} > x^{(k)}$. Если для заданного $\epsilon > 0$ условие останова $|x^{(k+1)} - x^{(k)}| < \epsilon$ никогда не выполнено, то последовательные итерации (4.15) отличаются друг от друга не менее, чем на ϵ , и потому $x^{(k+m)} \geq x^{(k)} + m\epsilon$. 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 (4.15) следует, что

$$|x^{(k+1)} - x^{(k)}| = \left| \frac{x^{(k)}}{\frac{1}{3} - 2(x^{(k)})^2} \right|.$$

Выражение в правой части этого равенства должно стремиться к нулю при $x^{(k)} \rightarrow \infty$, что противоречит нашему допущению $|x^{(k+1)} - x^{(k)}| \geq \epsilon$.

Следовательно, для какого-то номера в итерациях (4.15) будет выполнено условие останова $|x^{(k+1)} - x^{(k)}| < \epsilon$. 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. ■

Для надёжного определения погрешности приближённого решения \tilde{x} и контроля точности вычислений можно применять следующее условие:

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, то

$$|\tilde{x} - x^*| \leq \frac{|f(\tilde{x})|}{\min_{\xi \in [a, b]} |f'(\xi)|}. \quad (4.16)$$

Доказательство следует из теоремы Лагранжа о среднем (формулы конечных приращений)

$$f(\tilde{x}) - f(x^*) = f'(\xi) \cdot (\tilde{x} - x^*),$$

в которой ξ — некоторая точка, заключённая между \tilde{x} и x^* . Ясно, что тогда

$$|f(\tilde{x}) - f(x^*)| \geq \min_{\xi} |f'(\xi)| \cdot |\tilde{x} - x^*|,$$

и при $\min_{\xi \in [a, b]} |f'(\xi)| \neq 0$ получаем оценку (4.16). Отметим её очевидную аналогию с оценкой (3.165) для погрешности решения систем линейных уравнений: в обоих случаях невязка приближённого решения умножается на норму обратного отображения.

На практике нахождение точного минимума $\min_{\xi \in [a, b]} |f'(\xi)|$ может быть затруднительным, и тогда вместо него в (4.16) можно взять какую-нибудь положительную оценку снизу для области значений производной на $[a, b]$.

4.4.4 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} , но практически обычно ограничиваются небольшими порядками.

Предположим, что вещественная функция f является гладкой и монотонной на интервале $[a, b]$, так что она взаимно однозначно отображает этот интервал в некоторый интервал $[\alpha, \beta]$. Как следствие, существует обратная к f функция $g = f^{-1} : [\alpha, \beta] \rightarrow [a, b]$, 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 ξ — 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

$$x^{(k+1)} \leftarrow x^{(k)} + \sum_{l=1}^p (-1)^l g^{(l)}(f(\tilde{x})) \frac{(f(\tilde{x}))^l}{l!}, \quad k = 0, 1, 2, \dots,$$

которые и 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

$$\begin{aligned} g'(f(x)) f'(x) &= 1, \\ g''(f(x)) (f'(x))^2 + g'(f(x)) f''(x) &= 0, \\ g'''(f(x)) (f'(x))^3 + g''(f(x)) \cdot 2f'(x)f''(x) \\ &\quad + g'(f(x)) f'(x) f''(x) + g'(f(x)) f'''(x) = 0, \\ &\quad \dots \qquad \dots \qquad \dots, \end{aligned}$$

или

$$\begin{aligned} g'(f(x)) f'(x) &= 1, \\ g''(f(x)) (f'(x))^2 + g'(f(x)) f''(x) &= 0, \\ g'''(f(x)) (f'(x))^3 + 3g''(f(x)) f'(x) f''(x) + g'(f(x)) f'''(x) &= 0, \\ \dots \qquad \dots \qquad \dots \end{aligned}$$

Относительно неизвестных значений производных $g'(f(x))$, $g''(f(x))$, $g'''(f(x))$ и т. д. эта система соотношений имеет треугольный вид, позволяющий найти их последовательно одну за другой:

$$\begin{aligned} g'(f(x)) &= \frac{1}{f'(x)}, \\ g''(f(x)) &= -\frac{g(f(x)) f''(x)}{(f'(x))^2} = -\frac{f''(x)}{(f'(x))^3}, \\ g'''(f(x)) &= -\frac{3g''(f(x)) f'(x) f''(x) + g'(f(x)) f'''(x)}{(f'(x))^3} \\ &= -3 \frac{(f''(x))^2}{(f'(x))^5} - \frac{f'''(x)}{(f'(x))^4} \end{aligned}$$

и так далее.

Для $p = 1$ расчётные формулы метода Чебышёва имеют вид

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

что совпадает с методом Ньютона.

Для $p = 2$ расчётные формулы метода Чебышёва таковы

$$x^{(k+1)} \leftarrow x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})} - \frac{f''(x^{(k)}) (f(x^{(k)}))^2}{2(f'(x^{(k)}))^3}, \quad k = 0, 1, 2, \dots$$

Наиболее часто методом Чебышёва называют именно этот итерационный процесс, так как методы более высокого порядка из этого семейства на практике используются редко.

4.5 Классические методы решения систем уравнений

4.5a Метод простой итерации

Схема применения метода простой итерации для систем уравнений в принципе не отличается от случая одного уравнения. Исходная система уравнений $F(x) = 0$ должна быть каким-либо способом приведена к равносильному рекуррентному виду, например,

$$x = x - \Lambda F(x),$$

где Λ — неособенная матрица, и далее, после выбора некоторого начального приближения $x^{(0)}$, запускается итерационный процесс

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

где $\Phi(x) = x - \Lambda F(x)$. При благоприятных обстоятельствах последовательность $\{x^{(k)}\}$ сходится, и её пределом является искомое решение системы уравнений. Для обеспечения сходимости итераций к решению стараются удовлетворить теореме Банаха о неподвижной точке или же её аналогу — теореме Шрёдера.

4.56 Метод Ньютона и его модификации

Пусть для системы уравнений

$$\begin{cases} F_1(x_1, x_2, \dots, x_n) = 0, \\ F_2(x_1, x_2, \dots, x_n) = 0, \\ \vdots \quad \ddots \quad \vdots \\ F_n(x_1, x_2, \dots, x_n) = 0, \end{cases}$$

или, кратко, $F(x) = 0$, известно некоторое приближение \tilde{x} к решению x^* . Если F — плавно меняющаяся (гладкая функция), то естественно приблизить её в окрестности точки \tilde{x} линейной функцией, т. е.

$$F(x) \approx F(\tilde{x}) + F'(\tilde{x})(x - \tilde{x}),$$

где $F'(\tilde{x})$ — матрица Якоби отображения F в точке \tilde{x} . Далее для вычисления следующего и более точного приближения к решению системы уравнений естественно взять решение системы линейных алгебраических уравнений

$$F(\tilde{x}) + F'(\tilde{x})(x - \tilde{x}) = 0,$$

которая получается из разложения F в окрестности \tilde{x} . Следовательно, очередным приближением к решению можно взять

$$x = \tilde{x} - (F'(\tilde{x}))^{-1}F(\tilde{x}).$$

Итерационный процесс

$$x^{(k+1)} \leftarrow x^{(k)} - (F'(x^{(k)}))^{-1}F(x^{(k)}), \quad k = 0, 1, 2, \dots,$$

называют *методом Ньютона*.

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 \tilde{x} , в которой вычисляется эта матрица производных, то получим упрощённый стационарный итерационный процесс

$$x^{(k+1)} \leftarrow x^{(k)} - (F'(\tilde{x}))^{-1}F(x^{(k)}), \quad k = 0, 1, 2, \dots,$$

который часто называют *модифицированным методом Ньютона*. В нём решение систем линейных уравнений с одинаковыми матрицами $F'(\tilde{x})$ можно проводить по упрощённым алгоритмам, к примеру, найдя один раз LU-разложение матрицы $F'(\tilde{x})$ и далее используя его.

У метода Ньютона существует много различных вариантов и модификаций, и заинтересованный читатель может найти информацию о них, к примеру, в [15, 31].

Один из наиболее часто используемых результатов о сходимости метода Ньютона — это

Теорема 4.5.1 (теорема Л.В. Канторовича о методе Ньютона)

Пусть 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 $\text{cl } D$. Пусть, кроме того, существует такой непрерывный линейный оператор $\Gamma_0 = (F'(x_0))^{-1}$, что $\|\Gamma_0(F(x_0))\| \leq \eta$ и $\|\Gamma_0 F''(x)\| < K$ для всех $x \in \text{cl } D$ и некоторых констант η и K . Если

$$h = K\eta \leq \frac{1}{2}$$

и

$$r \geq r_0 = \frac{1 - \sqrt{1 - 2h}}{h} \eta,$$

то уравнение $F(x) = 0$ имеет решение x^ , к которому сходится метод Ньютона, как исходный, так и модифицированный. При этом*

$$\|x^* - x_0\| \leq r_0.$$

Для исходного метода Ньютона сходимость описывается оценкой

$$\|x^* - x_k\| \leq \frac{\eta}{2^k h} (2h)^{2^k}, \quad k = 0, 1, 2, \dots,$$

а для модифицированного метода верна оценка

$$\|x^* - x_k\| \leq \frac{\eta}{h} (1 - \sqrt{1 - 2h})^{k+1}, \quad k = 0, 1, 2, \dots,$$

при условии $h < \frac{1}{2}$.

Доказательство и дальнейшие результаты на эту тему можно найти в книге [18].

4.6 Интервальные системы линейных уравнений

Предметом рассмотрения настоящего пункта являются интервальные системы линейных алгебраических уравнений (ИСЛАУ) вида

$$\mathbf{A}x = \mathbf{b}, \quad (4.17)$$

где $\mathbf{A} = (a_{ij})$ — это интервальная $m \times n$ -матрица и $\mathbf{b} = (b_i)$ — интервальный m -вектор. Для интервальных уравнений решения и множества решений могут быть определены разнообразными способами (см. [41]), но ниже мы ограничимся так называемым *объединённым множеством решений* для (4.17), которое образовано всевозможными решениями x точечных систем $Ax = b$, когда матрица A и вектор b независимо пробегают \mathbf{A} и \mathbf{b} соответственно. Объединённое множество решений определяется строго как

$$\Xi(\mathbf{A}, \mathbf{b}) = \{x \in \mathbb{R}^n \mid (\exists A \in \mathbf{A})(\exists b \in \mathbf{b})(Ax = b)\}, \quad (4.18)$$

и ниже мы будем называть его просто *множеством решений* интервальной линейной системы (4.17), так как другие множества решений нами не исследуются. Точное описание множества решений может расти экспоненциально с размерностью вектора неизвестных n , а потому является практически невозможным уже при n , превосходящем несколько десятков. С другой стороны, в большинстве реальных постановок задач точное описание на самом деле и не нужно. На практике бывает вполне достаточно нахождения *оценки* для множества решений, т. е. приближенного описания, удовлетворяющего содержательно-му смыслу рассматриваемой задачи.

Приведём полезный технический результат, который часто используется в связи с исследованием и оцениванием множества решений интервальных систем линейных алгебраических уравнений.

Теорема 4.6.1 (характеризация Бекка) *Если $\mathbf{A} \in \mathbb{IR}^{m \times n}$, $\mathbf{b} \in \mathbb{IR}^m$, то*

$$\begin{aligned} \Xi(\mathbf{A}, \mathbf{b}) &= \{x \in \mathbb{R}^n \mid \mathbf{A}x \cap \mathbf{b} \neq \emptyset\} \\ &= \{x \in \mathbb{R}^n \mid 0 \in \mathbf{A}x - \mathbf{b}\}. \end{aligned}$$

Доказательство. Если $\tilde{x} \in \Xi(\mathbf{A}, \mathbf{b})$, то $\tilde{A}\tilde{x} = \tilde{b}$ для некоторых $\tilde{A} \in \mathbf{A}$, $\tilde{b} \in \mathbf{b}$. Следовательно, по крайней мере $\tilde{b} \in \mathbf{A}\tilde{x} \cap \mathbf{b}$, так что действительно $\mathbf{A}\tilde{x} \cap \mathbf{b} \neq \emptyset$.

Наоборот, если $\mathbf{A}\tilde{x} \cap \mathbf{b} \neq \emptyset$, то это пересечение $\mathbf{A}\tilde{x} \cap \mathbf{b}$ содержит вектор $\tilde{b} \in \mathbb{R}^m$, 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 $0 \in \mathbf{A}\tilde{x} - \mathbf{b}$. ■

Теорема 4.6.2 (характеризация Оеттли-Прагера) *Для объединённого множества решений ИСЛАУ имеет место*

$$x \in \Xi(\mathbf{A}, \mathbf{b}) \quad \Leftrightarrow \quad |(\text{mid } \mathbf{A})x - \text{mid } \mathbf{b}| \leq \text{rad } \mathbf{A} \cdot |x| + \text{rad } \mathbf{b}, \quad (4.19)$$

где неравенство между векторами понимается покомпонентным образом.

Доказательство. Для любых интервальных векторов-брусков \mathbf{p} и \mathbf{q} включение $\mathbf{p} \subseteq \mathbf{q}$ равносильно покомпонентному неравенству

$$|\text{mid } \mathbf{q} - \text{mid } \mathbf{p}| \leq \text{rad } \mathbf{q} - \text{rad } \mathbf{p}.$$

Следовательно, условие характеристики Бекка, т. е. $0 \in \mathbf{A}\tilde{x} - \mathbf{b}$, может быть переписано в следующем виде:

$$|\text{mid } (\mathbf{A}\tilde{x} - \mathbf{b})| \leq \text{rad } (\mathbf{A}\tilde{x} - \mathbf{b}).$$

С учётом правил преобразования середины и радиуса получаем

$$\begin{aligned} \text{mid } (\mathbf{A}\tilde{x} - \mathbf{b}) &= (\text{mid } \mathbf{A})\tilde{x} - \text{mid } \mathbf{b}, \\ \text{rad } (\mathbf{A}\tilde{x} - \mathbf{b}) &= (\text{rad } \mathbf{A}) \cdot |\tilde{x}| + \text{rad } \mathbf{b}, \end{aligned}$$

откуда вытекает требуемое. ■

4.6a Интервальный итерационный метод Гаусса-Зейделя

Интервальный метод Гаусса-Зейделя, которому посвящён этот параграф, является итерационной процедурой для уточнения уже известной внешней оценки множества решений ИСЛАУ вида (4.17), либо для

нахождения внешней оценки части множества решений, ограниченной некоторым брусом. Обычно его применяют после предварительного предобуславливания системы.

Предположим, что дана интервальная система линейных уравнений $\mathbf{A}\mathbf{x} = \mathbf{b}$ и известен некоторый брус \mathbf{x} , содержащий множество решений $\Xi(\mathbf{A}, \mathbf{b})$ или же некоторую его часть, которая интересует нас по условиям задачи. Пусть также в интервальной матрице $\mathbf{A} = (a_{ij})$ элементы главной диагонали не содержат нуля, т. е. $0 \notin a_{ii}$ для $i = 1, 2, \dots, n$. 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, наконец, поделим обе части равенств на \tilde{a}_{ii} . Получим

$$\tilde{x}_i = \left(\tilde{b}_i - \sum_{j \neq i} \tilde{a}_{ij} \tilde{x}_j \right) / \tilde{a}_{ii}, \quad i = 1, 2, \dots, n. \quad (4.20)$$

Полагая

$$\mathbf{x}'_i := \left(\mathbf{b}_i - \sum_{j \neq i} a_{ij} \mathbf{x}_j \right) / a_{ii}, \quad i = 1, 2, \dots, n,$$

мы должны признать, что

$$\tilde{x}_i \in \mathbf{x}'_i, \quad i = 1, 2, \dots, n,$$

так как выражения для \mathbf{x}'_i являются естественными интервальными расширениями выражений (4.20) по $\tilde{a}_{ij} \in a_{ij}$, $\tilde{b}_i \in \mathbf{b}_i$ и $\tilde{x}_j \in \mathbf{x}_j$. Итак, $\tilde{x} \in \mathbf{x}'$, и это верно для любой точки $\tilde{x} \in \Xi(\mathbf{A}, \mathbf{b})$, так что в целом $\Xi(\mathbf{A}, \mathbf{b}) \subseteq \mathbf{x}'$, т. е. брус \mathbf{x}' является новой внешней оценкой множества решений рассматриваемой ИСЛАУ.

Чтобы взять лучшее от обеих внешних оценок — новой \mathbf{x}' и старой \mathbf{x} , естественно выполнить их пересечение, коль скоро $\Xi(\mathbf{A}, \mathbf{b}) \subseteq \mathbf{x}' \cap \mathbf{x}$.

Таблица 4.2. Интервальный метод Гаусса-Зейделя
для внешнего оценивания множеств решений ИСЛАУ

<p style="text-align: center;">Вход</p> <p>Интервальная линейная система уравнений $\mathbf{Ax} = \mathbf{b}$. 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 $\epsilon > 0$.</p>
<p style="text-align: center;">Выход</p> <p>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}».</p>
<p style="text-align: center;">Алгоритм</p> <p>$q \leftarrow +\infty$; 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</p>

Далее процесс построения новой внешней оценки для множества решений можно повторить, отправляясь от \mathbf{x}' и потом снова взяв пересечение полученной внешней 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 других компонент следующего приближения.

Может ли в процессе описанного уточнения-пересечения встретиться ситуация $\mathbf{x}'_i \cap \mathbf{x}_i = \emptyset$ 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.

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} ? Исследование интервального метода Гаусса-Зейделя для внешнего оценивания объединённых множеств решений ИСЛАУ было выполнено многими исследователями, и краткое резюме их результатов

4.7 Интервальные методы решения уравнений и систем уравнений

Интервальные методы позволяют придать конструктивный характер некоторым известным результатам математического анализа, которые раньше рассматривались как «чистые» теоремы существования. Мы уже могли видеть это для теоремы Миранды в §4.46. Другим ярким примером является

Теорема 4.7.1 (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 [16, 18, 30, 31, 59, 62].

С учётом сказанного выше во введении к главе (стр. 560) о равносильности рекуррентного вида систем уравнений (4.4) и их канонической формы (4.1)–(4.2) для вычислительной математики чрезвычайно полезными оказываются результаты анализа, утверждающие существование неподвижных точек отображений. Фактически, теорема Брауэра является именно таким результатом. Она аналогична по смыслу теоремам Банаха и Шрёдера о неподвижной точке (см. §4.4в), но имеет свою специфику. В ней не требуется, чтобы отображение было глобально сжимающим на всём пространстве. Достаточно и того, что некоторое условие (аналогичное, по сути, условию «сжатия») выполняется для образа какого-то интересующего нас выпуклого компакта, в котором нужно обосновать присутствие решения уравнения.

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 в процессе вычислений на компьютере.

Ещё большую конструктивную силу имеет модификация теоремы Брауэра:

Теорема 4.7.2 (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 ∂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 [18, 30, 59].

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.

4.7a 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, 31, 35, 44]). 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)

(4.21)

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. §1.10).

Задача (4.21) оказывается трудной, и в классическом численном анализе почти полностью отсутствуют развитые методы для её решения. Из часто используемых подходов, имеющих ограниченный успех, следует упомянуть *аналитическое исследование, мультистарт, методы продолжения* [31].

Итак, пусть к решению предъявлена система уравнений (4.2)

$$F(x) = 0$$

на брус $\mathbf{X} \subset \mathbb{R}^n$. Существование решения этой системы 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 весьма полезной при решении рассматриваемой задачи. В частности, если нуль содержится во внутренней интервальной оценке множества значений $\text{ran}(F, \mathbf{X})$ отображения F , то на брус \mathbf{X} гарантированно находится решение системы (4.2). С другой стороны, если в нашем распоряжении имеется интервальное расширение \mathbf{F} функции F на \mathbf{X} , то $\mathbf{F}(\mathbf{X}) \supseteq \text{ran}(F, \mathbf{X})$. Поэтому если $0 \notin \mathbf{F}(\mathbf{X})$, то 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 (4.2) всегда можно переписать в равносильной рекуррентной форме

$$x = T(x) \tag{4.22}$$

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 Λ , либо как-нибудь ещё. Переход к рекуррентной форме даёт некоторые дополнительные возможности. Пусть $\mathbf{T} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ — интервальное расширение отображения T . Ясно, что

решения системы (4.22) могут лежать лишь в пересечении $X \cap T(X)$. Поэтому если

$$X \cap T(X) = \emptyset,$$

то в X нет решений системы уравнений (4.22). Коль скоро искомое решение содержится и в $T(X)$, то для дальнейшего уточнения бруса, в котором может присутствовать решение, мы можем организовать итерации с пересечением

$$X^{(0)} \leftarrow X, \quad (4.23)$$

$$X^{(k+1)} \leftarrow T(X^{(k)}) \cap X^{(k)}, \quad k = 0, 1, 2, \dots \quad (4.24)$$

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 X выполнено

$$T(X) \subseteq X,$$

то по теореме Брауэра о неподвижной точке (стр. 602) в X гарантированно находится решение системы (4.22). Для уточнения этого бруса мы снова можем воспользоваться итерациями (4.23)–(4.24). Таким образом, наихудшим, с точки зрения уточнения информации о решении системы, является случай

$$T(X) \supsetneq X. \quad (4.25)$$

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 ($X \cap T(X)$) либо предел последовательности (4.23)–(4.24). 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 брус решения системы уравнений нет.

В действительности, каждый из изложенных выше приёмов уточнения решения допускает далеко идущие модификации и улучшения. Например, это относится к итерациям вида (4.23)–(4.24), которые могут быть последовательно применены не к целым брусам $X^{(k)}$, а к отдельным их компонентам в комбинации с различными способами приведения исходной системы к рекуррентному виду (4.22). На этом пути

мы приходим к чрезвычайно эффективным алгоритмам, которые получили наименование *методов распространения ограничений* (см., к примеру, [34]).

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 (4.2) и T в (4.22) 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. §1.5).

4.76 Одномерный интервальный метод Ньютона

В этом параграфе мы рассмотрим простейший случай одного уравнения с одним неизвестным.

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

$$f(\tilde{x}) - f(x^*) = (\tilde{x} - x^*) \cdot f'(\xi), \quad (4.26)$$

где ξ — 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 $\mathbf{f}'(\mathbf{X})$ является какой-либо интервальной оценкой производной от функции $f(x)$ на \mathbf{X} , то $f'(\xi) \in \mathbf{f}'(\mathbf{X})$ и, интервализуя выписанное равенство, получим включение

$$x^* \in \tilde{x} - \frac{f(\tilde{x})}{\mathbf{f}'(\mathbf{X})} \quad (4.27)$$

в случае $0 \notin \mathbf{f}'(\mathbf{X})$. Иными словами, для корня x^* уравнения $f(x) = 0$ получается новый интервал локализации в виде правой части включения (4.27). Интервальное выражение, фигурирующее в правой части (4.27), 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

$$\mathcal{N} : \mathbb{I}\mathbb{R} \times \mathbb{R} \rightarrow \mathbb{I}\mathbb{R},$$

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 .

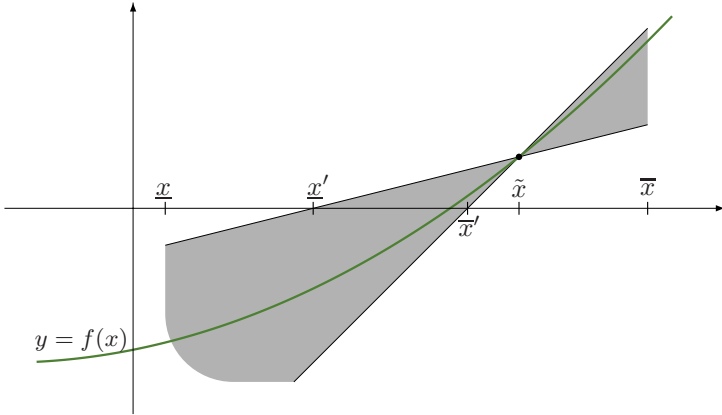


Рис. 4.10. Иллюстрация работы одномерного интервального метода Ньютона.

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. 4.10, где обозначено $\mathbf{X}' = \mathcal{N}(\mathbf{X}, \tilde{x})$. Далее, положив $\mathbf{X}^{(0)} = \mathbf{X}$, естественно организовать итерационное уточнение

$$\mathbf{X}^{(k+1)} \leftarrow \mathbf{X}^{(k)} \cap \mathcal{N}(\mathbf{X}^{(k)}, \tilde{x}^{(k)}), \quad k = 0, 1, 2, \dots, \quad (4.28)$$

задавшись каким-то правилом выбора точек $\tilde{x}^{(k)} \in \mathbf{X}^{(k)}$. Итерации (4.28) 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 (4.28) 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.² 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

$$\mathbf{X}^{(k+1)} = \mathbf{X}^{(k)} \cap \mathcal{N}(\mathbf{X}^{(k)}, \tilde{x}) = \mathbf{X}^{(k)}.$$

Ясно, что тогда и все последующие итерации метода будут равны $\mathbf{X}^{(k)}$, 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 $\mathbf{f}'(\mathbf{X})$ не содержит нуля, интервальный метод Ньютона обладает рядом замечательных качеств. Если $0 \notin \mathbf{f}'(\mathbf{X})$ 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.

В разделе §4.4г мы рассматривали пример Донована-Миллера-Мореланда, 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 успешно решает этот пример, что впервые было отмечено в работе [63]. Таким образом, интервальный метод Ньютона оказывается даже более сильным, чем его прародитель.

²В 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 $\mathbf{f}'(\mathbf{X}) \not\equiv 0$. Если для некоторой точки \tilde{x} справедливо включение $N(\mathbf{X}, \tilde{x}) \subseteq \mathbf{X}$, 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 найдётся такая точка $\xi \in \square\{\tilde{x}, y\} \subset \mathbf{X}$, что

$$f(\tilde{x}) - f(y) = f'(\xi)(\tilde{x} - y). \quad (4.29)$$

Чтобы подчеркнуть зависимость этой точки от \tilde{x} и y , мы обозначим её как $\xi(\tilde{x}, y)$. Коль скоро $f'(\xi) \in \mathbf{f}'(\mathbf{X})$, то ясно, что $f'(\xi(\tilde{x}, y)) \neq 0$ при любых \tilde{x} и y . По этой причине мы можем определить функцию

$$g(y) = y - \frac{f(y)}{f'(\xi(\tilde{x}, y))}. \quad (4.30)$$

По условиям Предложения функция $g(y)$ непрерывна. Кроме того, из равенства (4.29) 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

$$g(y) = \tilde{x} - \frac{f(\tilde{x})}{f'(\xi(\tilde{x}, y))} \in \tilde{x} - \frac{f(\tilde{x})}{\mathbf{f}'(\mathbf{X})} = N(\mathbf{X}, \tilde{x}) \subseteq \mathbf{X}.$$

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 (4.30) тогда вытекает, что $f(y^*) = 0$, т. е. y^* является решением уравнения $f(x) = 0$. ■

Рассмотрим теперь случай $0 \in \mathbf{f}'(\mathbf{X})$. Он встречается, когда на интервале \mathbf{X} имеется кратный корень x^* , в котором $f'(x^*) = 0$, либо когда

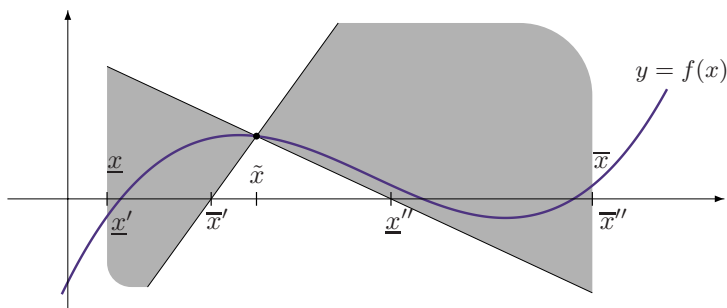


Рис. 4.11. Иллюстрация работы одномерного интервального метода Ньютона. Случай нульсодержащего интервала производной.

интервал \mathbf{X} настолько широк, что содержит более одного корня. В этом случае мы тоже можем придать смысл интервальному оператору Ньютона, воспользовавшись для выполнения деления $f(\tilde{x})/f'(\mathbf{X})$ специальной интервальной арифметикой — так называемой интервальной арифметикой Кахана, допускающей деление на нульсодержащие интервалы. В действительности, эта модификация даже усилит интервальный метод Ньютона, так как мы получим возможность отделять различные решения друг от друга: в результате выполнения шага интервального метода Ньютона при $0 \in \text{int } f'(\mathbf{X})$ часто получаются два непересекающихся интервала. Эта ситуация иллюстрируется на Рис. 4.11.

В арифметике Кахана дополнительно определено деление интервалов \mathbf{a} и \mathbf{b} с $0 \in \mathbf{b}$, которое и приводит к бесконечным интервалам. Для удобства мы выпишем соответствующие результаты в развёрнутой

форме:

$$a/b = \frac{[\underline{a}, \bar{a}]}{[\underline{b}, \bar{b}]} = \begin{cases} \underline{a} \cdot [1/\bar{b}, 1/\underline{b}], & \text{если } 0 \notin \underline{b}, \\] - \infty, +\infty[, & \text{если } 0 \in \underline{a} \text{ и } 0 \in \underline{b}, \\ [\bar{a}/\underline{b}, +\infty[, & \text{если } \bar{a} < 0 \text{ и } \underline{b} < \bar{b} = 0, \\] - \infty, \bar{a}/\bar{b}] \cup [\bar{a}/\underline{b}, +\infty[, & \text{если } \bar{a} < 0 \text{ и } \underline{b} < 0 < \bar{b}, \\] - \infty, \bar{a}/\bar{b}], & \text{если } \bar{a} < 0 \text{ и } 0 = \underline{b} < \bar{b}, \\] - \infty, \underline{a}/\underline{b}], & \text{если } 0 < \underline{a} \text{ и } \underline{b} < \bar{b} = 0, \\] - \infty, \underline{a}/\underline{b}] \cup [\underline{a}/\bar{b}, +\infty[, & \text{если } 0 < \underline{a} \text{ и } \underline{b} < 0 < \bar{b}, \\ [\underline{a}/\bar{b}, +\infty[, & \text{если } 0 < \underline{a} \text{ и } 0 = \underline{b} < \bar{b}, \\ \emptyset, & \text{если } 0 \notin \underline{a} \text{ и } 0 = \underline{b}. \end{cases} \quad (4.31)$$

В заключение — необходимый комментарий о реализации интервального метода Ньютона на ЭВМ. Значение $f(\tilde{x})$, несмотря на точность 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.

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 ещё и способ внешнего оценивания множества решений интервальной линейной системы, к которой приводится оценивание бруса решения. В оставшейся части этого параграфа

мы рассмотрим простейшую форму многомерного интервального метода Ньютона, а его более специальным версиям, которые связываются с именами Кравчика и Хансена-Сенгупты, будут посвящены отдельные параграфы.

Определение 4.7.2 [48] Для отображения $F : \mathbb{R}^n \supseteq D_0 \rightarrow \mathbb{R}^m$ матрица $\mathbf{A} \in \mathbb{IR}^{m \times n}$ называется интервальной матрицей наклонов на $D \subseteq D_0$, если для любых $x, y \in D$ равенство

$$F(y) - F(x) = \mathbf{A}(y - x)$$

имеет место с некоторой вещественной $m \times n$ -матрицей $A \in \mathbf{A}$.

Предположим, что на брусе \mathbf{x} к решению предъявлена система нелинейных уравнений

$$F(x) = 0. \quad (4.32)$$

Если \mathbf{S} — интервальная матрица наклонов отображения F на \mathbf{x} , то для любых точек $x, \tilde{x} \in \mathbf{x}$ справедливо представление

$$F(x) \in F(\tilde{x}) + \mathbf{S}(x - \tilde{x}).$$

В частности, если x — решение системы уравнений (4.32), т. е. $F(x) = 0$, то

$$0 \in F(\tilde{x}) + \mathbf{S}(x - \tilde{x}). \quad (4.33)$$

Вспомним характеристику Бекка для объединённого множества решений ИСЛАУ (Теорема 4.6.1): получается, что точка x удовлетворяет включению (4.33) тогда и только тогда, когда она принадлежит объединённому множеству решений интервальной линейной системы

$$\mathbf{S}(x - \tilde{x}) = -F(\tilde{x}). \quad (4.34)$$

Далее, если $Encl$ — процедура внешнего оценивания множества решений ИСЛАУ, то справедливо включение

$$x - \tilde{x} \in Encl(\mathbf{S}, -F(\tilde{x})),$$

так что

$$x \in \tilde{x} + Encl(\mathbf{S}, -F(\tilde{x})).$$

Определение 4.7.3 Пусть для внешнего оценивания множеств решений ИСЛАУ зафиксирована процедура Encl , а для отображения $F : \mathbb{R}^n \supseteq D \rightarrow \mathbb{R}^n$ известна интервальная матрица наклонов $\mathbf{S} \in \mathbb{IR}^{n \times n}$. Отображение

$$\mathcal{N} : \mathbb{ID} \times \mathbb{R}^n \rightarrow \mathbb{IR}^n,$$

задаваемое правилом

$$\mathcal{N}(\mathbf{x}, \tilde{x}) = \tilde{x} + \text{Encl}(\mathbf{S}, -F(\tilde{x})),$$

называется интервальным оператором Ньютона на \mathbb{ID} относительно точки \tilde{x} .

Как лучше выбирать центр разложения \tilde{x} ? 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

$$\mathbf{S}(\mathbf{x} - \tilde{x}) = -F(\tilde{x}),$$

которое мы должны пересекать с исходным брусом. Может быть, 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.

Наиболее неблагоприятной ситуацией при работе интервального метода Ньютона является, конечно, появление включения

$$\mathcal{N}(\mathbf{x}, \tilde{x}) \supseteq \mathbf{x}.$$

Тогда все последующие шаги зацикливаются на брусе \mathbf{x} и не дают никакой дополнительной информации об искомым решениях системы. Как поступать в этом случае? Ответ на этот вопрос рассматривается в следующем §4.8.

4.7г Метод Кравчика

Пусть на брусе $\mathbf{x} \in \mathbb{IR}^n$ задана система n нелинейных уравнений с n неизвестными

$$F(\mathbf{x}) = 0,$$

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 $\mathbf{S} \in \mathbb{R}^{n \times n}$ — интервальная матрица наклонов отображения F на брус \mathbf{x} . Если x — это точка решения системы, то

$$0 \in F(\tilde{x}) + \mathbf{S}(x - \tilde{x}). \quad (4.33)$$

Но далее, в отличие от интервального метода Ньютона, мы не будем переходить к рассмотрению интервальной линейной системы (4.34), а домножим обе части этого включения слева на точечную $n \times n$ -матрицу, которую нам будет удобно обозначить как $(-A)$:

$$0 \in -AF(\tilde{x}) - A\mathbf{S}(x - \tilde{x}).$$

Добавление к обеим частям получившегося соотношения по $(x - \tilde{x})$ приводит к

$$x - \tilde{x} \in -AF(\tilde{x}) - A\mathbf{S}(x - \tilde{x}) + (x - \tilde{x}),$$

что равносильно

$$x \in \tilde{x} - AF(\tilde{x}) + (I - A\mathbf{S})(x - \tilde{x}),$$

так как для неинтервального общего множителя $(x - \tilde{x})$ можно воспользоваться дистрибутивным соотношением (1.16). 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, придя к соотношению

$$x \in \tilde{x} - AF(\tilde{x}) + (I - A\mathbf{S})(\mathbf{x} - \tilde{x}),$$

Определение 4.7.4 Пусть определены некоторые правила, сопоставляющие всякому брусу $\mathbf{x} \in \mathbb{IR}^n$ точку $\tilde{x} \in \mathbf{x}$ и вещественную $n \times n$ -матрицу A и пусть $\mathbf{S} \in \mathbb{IR}^{n \times n}$ — интервальная матрица наклонов отображения $F : \mathbb{R}^n \supseteq D \rightarrow \mathbb{R}^n$ на D . Отображение

$$\mathcal{K} : \mathbb{ID} \times \mathbb{R} \rightarrow \mathbb{IR}^n,$$

задаваемое выражением

$$\mathcal{K}(\mathbf{x}, \tilde{x}) := \tilde{x} - AF(\tilde{x}) + (I - A\mathbf{S})(\mathbf{x} - \tilde{x}),$$

называется оператором Кравчика на \mathbb{ID} относительно точки \tilde{x} .

Теорема 4.7.3 Пусть $F : \mathbb{R}^n \supseteq D \rightarrow \mathbb{R}^n$ — непрерывное по Липшицу отображение, S — его интервальная матрица наклонов и $\tilde{x} \in \mathbf{x} \subseteq \mathbb{ID}$. Тогда

- (i) каждое решение системы $F(x) = 0$ 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 $F(x) = 0$;
- (iv) если $\tilde{x} \in \text{int } \mathbf{x}$ и $\emptyset \neq \mathcal{K}(\mathbf{x}, \tilde{x}) \subseteq \text{int } \mathbf{x}$, то интервальная матрица S сильно неособенна и в $\mathcal{K}(\mathbf{x}, \tilde{x})$ содержится в точности одно решение системы $F(x) = 0$.

Оператор Кравчика — это не что иное, как центрированная форма интервального расширения отображения $\Phi(x) = x - LF(x)$, возникающего в правой части системы уравнений после её приведения к рекуррентному виду

$$x = \Phi(x).$$

Отсюда легко вывести свойства (i)–(iii) оператора Кравчика из сформулированной выше теоремы.

4.8 Глобальное решение уравнений и систем уравнений

Если ширина бруса \mathbf{X} велика, то на нём описанные в предшествующем параграфе методики уточнения решения могут оказаться малоуспешными в том смысле, что мы получим включение (4.25), из которого нельзя вывести никакого определённого заключения ни о существовании решения на брусе \mathbf{X} , ни о его отсутствии. Кроме того, сам этот брус, как область потенциально содержащая решение, несколько не будет уточнён (уменьшен).

Тогда практикуют принудительное дробление \mathbf{X} на более мелкие подбрусы. Наиболее популярна при этом бисекция — разбиение бруса \mathbf{X} на две (равные или неравные) части вдоль какой-нибудь грани,

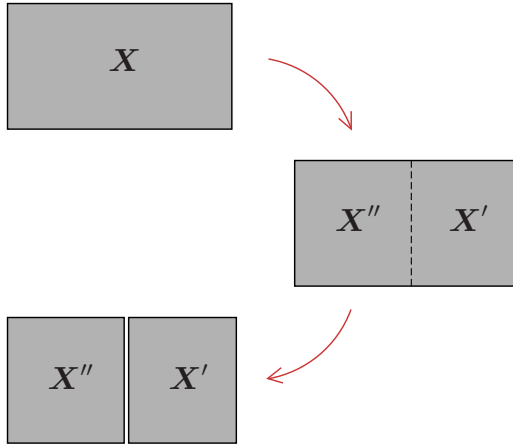


Рис. 4.12. Принудительное дробление бруса.

например, на половинки

$$\begin{aligned} X' &= (X_1, \dots, [\underline{X}_\iota, \text{mid } X_\iota], \dots, X_n), \\ X'' &= (X_1, \dots, [\text{mid } X_\iota, \overline{X}_\iota], \dots, X_n) \end{aligned}$$

для некоторого номера $\iota \in \{1, 2, \dots, n\}$. При этом подбрусы X' и X'' называются *потомками* бруса X . Далее эти потомки можно разбить ещё раз, и ещё \dots — столько, сколько необходимо для достижения желаемой малости их размеров, при которой мы сможем успешно выполнять на этих брусах рассмотренные выше тесты существования решений.

Если мы не хотим упустить при этом ни одного решения системы, то должны хранить все возникающие в процессе такого дробления подбрусы, относительно которых тестом существования не доказано строго, что они не содержат решений. Организуем поэтому *рабочий список* \mathcal{L} из всех потомков начального бруса X , подозрительных на содержание решений. Хотя мы называем эту структуру данных «списком», в смысле программной реализации это может быть не обязательно список, а любое хранилище брусов, организованное, к примеру, как *стек* (магазин) или *куча* и т. п. (см. [4]) В целом же алгоритм глобального доказательного решения системы уравнений организуем в виде повторяющейся последовательности следующих действий:

- извлечение некоторого бруса из списка \mathcal{L} ,
- дробление этого бруса на потомки,
- проверка существования решений в каждом из подбрус-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) брусков δ , 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 подбрусы (фактически, разбиениям исходной задачи на подзадачи), а отсечения — это отбрасывание бесперспективных подбрусков исходной области поиска.³

Неизбежные ограничения на вычислительные ресурсы ЭВМ могут не позволить решить этим алгоритмом конкретную задачу (4.21) «до конца», поскольку возможны ситуации, когда

- 1) размеры обрабатываемого бруса уже меньше δ , но нам ещё не удаётся ни доказать существование в нём решений, ни показать их отсутствие;
- 2) размеры обрабатываемого бруса всё ещё больше δ , но вычислительные ресурсы уже не позволяют производить его обработку дальше: исчерпались выделенное время, память и т. п.

³Стандартный английский термин для обозначения подобного типа алгоритмов — «branch-and-prune». С ними тесно связаны *методы ветвей и границы*, широко применяемые в вычислительной оптимизации.

Таблица 4.3. Интервальный метод ветвлений и отсечений
для глобального доказательного решения уравнений

<p style="text-align: center;">Вход</p> <p>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.</p>
<p style="text-align: center;">Выход</p> <p>Список НавернякаРешения из брусов размера менее δ, которые гарантированно содержат решения системы уравнений в \mathbf{X}. Список ВозможноРешения из брусов размера менее δ, которые могут содержать решения системы уравнений в \mathbf{X}. Список Недообработанные из брусов размера более δ, которые могут содержать решения системы уравнений в \mathbf{X}.</p>
<p style="text-align: center;">Алгоритм</p> <p>инициализируем рабочий список \mathcal{L} исходным брусом \mathbf{X} ; DO WHILE (($\mathcal{L} \neq \emptyset$) и (не исчерпаны ресурсы ЭВМ)) извлекаем из рабочего списка \mathcal{L} брус \mathbf{Y} ; применяем к \mathbf{Y} тест существования решения, его результат обозначаем также через \mathbf{Y} ; IF (в \mathbf{Y} доказано отсутствие решений) THEN удаляем брус \mathbf{Y} из рассмотрения ELSE IF ((размер бруса \mathbf{Y}) $< \delta$) THEN заносим \mathbf{Y} в соответствующий из списков НавернякаРешения или ВозможноРешения ELSE рассекаем \mathbf{Y} на потомки \mathbf{Y}' и \mathbf{Y}'' и заносим их в рабочий список \mathcal{L} END IF END IF END DO все брусы из \mathcal{L} перемещаем в список Недообработанные;</p>

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 брусков:

список **НавернякаРешения**, состоящий из брусков шириной меньше δ , которые гарантированно содержат решения,

список **ВозможноРешения**, состоящий из брусков шириной меньше δ , подозрительных на содержание решения, и

список **Недообработанные**, состоящий брусков, которые алгоритму не удалось обработать «до конца» и которые имеют ширину не меньше δ .

При этом все решения рассматриваемой системы уравнений, не принадлежащие брусам из списка **НавернякаРешения**, содержатся в брусах из списков **ВозможноРешения** и **Недообработанные**.

Практика эксплуатации интервальных методов для доказательного глобального решения уравнений и систем уравнений выявила ряд проблем и трудностей. Во многих случаях (особенно при наличии так называемых кратных корней) задачу не удаётся решить до конца и предъявить все гарантированные решения уравнения. Список брусков-ответов с неопределённым статусом (**ВозможноРешения** в псевдокоде Табл. 4.3) часто никак не собираются исчезать ни при увеличении точности вычислений, ни при выделении дополнительного времени счета и т.п. Нередко он разрастается до огромных размеров, хотя большинство образующих его брусков возможных решений являются «фантомами» немногих реальных решений. Но эти феномены могут быть успешно 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 нами при доказательном решении систем уравнений инструменты приспособлены для выявления особых точек различных типов? Нетрудно понять, что интервальный метод Ньютона, методы Кравчика и Хансена-Сенгупты,

тесты существования Мура и Куи, основывающиеся на теоремах Лерэ-Шаудера и Брауэра и наиболее часто используемые при практических доказательных вычислениях решений уравнений, охватывают только случаи индекса ± 1 особой точки F . Если же решение системы является критической точкой соответствующего отображения с индексом, не равным ± 1 , то доказать его существование с помощью вышеупомянутых результатов принципиально не получится. Это объясняет, почему многие существующие практические интервальные алгоритмы для доказательства глобального решения систем уравнений не могут достичь «полного успеха» в общем случае.

Помимо вышеназванной причины необходимо отметить, что список **ВозможноРешения** может соответствовать неустойчивым решениям системы уравнений, имеющим нулевой индекс. 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 кратный корень квадратного уравнения (4.5)–(4.6), и хорошо известно, что он плохо находится численно как традиционными, так и интервальными подходами.

Алгоритмы ветвлений и отсечений, дополненные различными усовершенствованиями и приёмами, ускоряющими сходимость, получили большое развитие в интервальном анализе в последние десятилетия (см., например, книги [39, 43, 47, 48]), 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*. – Москва: Мир, 1987.
- [2] АКРИТАС А. *Основы компьютерной алгебры с приложениями*. – Москва: Мир, 1994.
- [3] БАРАХНИН В.Б., ШАПЕЕВ В.П. *Введение в численный анализ*. – Санкт-Петербург–Москва–Краснодар: Лань, 2005.
- [4] БАУЭР Ф.Л., ГООЗ Г. *Информатика. В 2-х ч.* – Москва: Мир, 1990.
- [5] БАХВАЛОВ Н.С., ЖИДКОВ Н.П., КОБЕЛЬКОВ Г.М. *Численные методы*. – Москва: Бином, 2003, а также другие издания этой книги.

- [6] БАХВАЛОВ Н.С., КОРНЕВ А.А., ЧИЖОНКОВ Е.В. *Численные методы. Решение задач и упражнения.* – Москва: Дрофа, 2008.
- [7] БЕРЕЗИН И.С., ЖИДКОВ Н.П. *Методы вычислений. Т. 1–2.* – Москва: Наука, 1966.
- [8] БЕРЖЕ М. *Геометрия. Т. 1, 2.* – Москва: Наука, 1984.
- [9] ВЕРЖБИЦКИЙ В.М. *Численные методы. Части 1–2.* – Москва: «Оникс 21 век», 2005.
- [10] ВОЛКОВ Е.А. *Численные методы.* – Москва: Наука, 1987.
- [11] ГОДУНОВ С.К. *Современные аспекты линейной алгебры.* – Новосибирск: Научная книга, 1997.
- [12] ГОДУНОВ С.К., АНТОНОВ А.Г., КИРИЛЛЮК О.П., КОСТИН В.И. *Гарантированная точность решения систем линейных уравнений в евклидовых пространствах.* – Новосибирск: Наука, 1992.
- [13] ГЭРИ М., ДЖОНСОН Д. *Вычислительные машины и труднорешаемые задачи.* – Москва: Мир, 1982.
- [14] ДЕМИДОВИЧ Б.П., МАРОН А.А. *Основы вычислительной математики.* – Москва: Наука, 1970.
- [15] ДЭННИС ДЖ., мл., ШНАБЕЛЬ Р. *Численные методы безусловной оптимизации и решения нелинейных уравнений.* – Москва: Мир, 1988.
- [16] ЗОРИЧ В.А. *Математический анализ. Т. 1.* – Москва: Наука, 1981. *Т. 2.* – Москва: Наука, 1984, а также другие издания.
- [17] КАЛИТКИН Н.Н. *Численные методы.* – Москва: Наука, 1978.
- [18] КАНТОРОВИЧ Л.В., АКИЛОВ Г.П. *Функциональный анализ.* – Москва: Наука, 1984.
- [19] КОЛЛАТЦ Л. *Функциональный анализ и вычислительная математика.* – Москва: Мир, 1969.
- [20] КОЛМОГОРОВ А.Н., ФОМИН С.В. *Элементы теории функций и функционального анализа.* – Москва: Физматлит, 2004, а также другие издания книги.
- [21] КРЫЛОВ А.Н. *Лекции о приближённых вычислениях.* – Москва: ГИТТЛ, 1954, а также более ранние издания.
- [22] КРЫЛОВ В.И., БОВКОВ В.В., МОНАСТЫРНЫЙ П.И. *Вычислительные методы. Т. 1–2.* – Москва: Наука, 1976.
- [23] КУНЦ К.С. *Численный анализ.* – Киев: Техника, 1964.
- [24] ЛЕБЕДЕВ В.И. *Функциональный анализ и вычислительная математика.* – Москва: Физматлит, 2000.
- [25] МАЦОКИН А.М. *Численный анализ. Вычислительные методы линейной алгебры. Конспекты лекций для преподавания в III семестре ММФ НГУ.* – Новосибирск: НГУ, 2009–2010.
- [26] МАЦОКИН А.М., СОРОКИН С.Б. *Численные методы. Часть 1. Численный анализ.* – Новосибирск: НГУ, 2006.

- [27] Меньшиков Г.Г. *Локализуемые вычисления. Конспект лекций.* – Санкт-Петербург: СПбГУ, Факультет прикладной математики–процессов управления, 2003.
- [28] Миньков С.Л., Миньков Л.Л. *Основы численных методов.* – Томск: Издательство научно-технической литературы, 2005.
- [29] Мысовских И.П. *Лекции по методам вычислений.* – Санкт-Петербург: Издательство Санкт-Петербургского университета, 1998.
- [30] Опойцев В.И. *Нелинейная системостатика.* – Москва: Наука, 1986.
- [31] Ортега Дж., Рейнболдт В. *Итерационные методы решения нелинейных систем уравнений со многими неизвестными.* – Москва: Мир, 1975.
- [32] Островский А.М. *Решение уравнений и систем уравнений.* – Москва: Издательство иностранной литературы, 1963.
- [33] Самарский А.А., Гулин А.В. *Численные методы.* – Москва: Наука, 1989.
- [34] 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.
- [35] Траув Дж. *Итерационные методы решения уравнений.* – Москва: Мир, 1985.
- [36] Тыртышников Е.Е. *Методы численного анализа.* – Москва: Академия, 2007.
- [37] Успенский В.А., Семёнов А.Л. *Теория алгоритмов: основные открытия и приложения.* – Москва: Наука, 1987.
- [38] Фихтенгольц Г.М. *Курс дифференциального и интегрального исчисления. Т. I.* – Москва: Наука, 1966.
- [39] Хансен Э., Уолстер Дж.У. *Глобальная оптимизация с помощью методов интервального анализа.* – Москва-Ижевск: Издательство «РХД», 2012.
- [40] Холодниок М., Клич А., Кубичек М., Марек М. *Методы анализа нелинейных динамических моделей.* – Москва: Мир, 1991.
- [41] Шарый С.П. *Конечномерный интервальный анализ.* – Электронная книга, 2019 (см. <http://www.nsc.ru/interval/Library/InteBooks>)
- [42] Шилов Г.Е. *Математический анализ. Функции одного переменного. Ч. 1–2.* – Москва: Наука, 1969.
- [43] KEARFOTT R.B. *Rigorous global search: Continuous problems.* – Dordrecht: Kluwer, 1996.
- [44] KELLEY C.T. *Iterative methods for linear and nonlinear equations.* – Philadelphia: SIAM, 1995.
- [45] KREINOVICH V., LAKEYEV A.V., ROHN J., KAHL P. *Computational complexity and feasibility of data processing and interval computations.* – Dordrecht: Kluwer, 1997.
- [46] MIRANDA C. Un' osservazione su un teorema di Brouwer // *Bollet. Unione Mat. Ital. Serie II.* – 1940. – Т. 3. – С. 5–7.

- [47] MOORE R.E., KEARFOTT R.B., CLOUD M. *Introduction to interval analysis*. – Philadelphia: SIAM, 2009.
- [48] NEUMAIER A. *Interval methods for systems of equations*. – Cambridge: Cambridge University Press, 1990.
- [49] TREFETHEN L.N. Pseudospectra of linear operators // *SIAM Review*. 1997. – Vol. 39, No. 3. – P. 383–406.
- [50] TREFETHEN L.N., BAU D. III *Numerical linear algebra*. – Philadelphia: SIAM, 1997.

Дополнительная

- [51] АБАФФИ Й., СПЕДИКАТО Э. *Математические методы для линейных и нелинейных уравнений. Проекционные ABS-алгоритмы*. – Москва: Мир, 1996.
- [52] АРНОЛЬД В.И. *Обыкновенные дифференциальные уравнения*. – Москва: Наука, 1984.
- [53] БАБЕНКО К.И. *Основы численного анализа*. – Москва: Наука, 1986.
- [54] ГАНШИН Г.С. *Методы оптимизации и решение уравнений*. – Москва: Наука, 1987.
- [55] ЕРМАКОВ С.М. *Метод Монте-Карло и смежные вопросы*. – Москва: Наука, 1975.
- [56] ЗАГУСКИН В.Л. *Справочник по численным методам решения алгебраических и трансцендентных уравнений*. – Москва: Физматгиз, 1960.
- [57] КРАСНОСЕЛЬСКИЙ М.А., ЗАБРЕЙКО П.П. *Геометрические методы нелинейного анализа*. – Москва: Наука, 1975.
- [58] КРАСНОСЕЛЬСКИЙ М.А., ПЕРОВ А.И., ПОВОЛОЦКИЙ А.И., ЗАБРЕЙКО П.П. *Векторные поля на плоскости*. – Москва: Физматлит, 1963.
- [59] НИРЕНБЕРГ Л. *Лекции по нелинейному функциональному анализу*. – Москва: Мир, 1977.
- [60] ОПОЙЦЕВ В.И. *Школа Опойцева: Математический анализ*. – Москва: URSS, 2016.
- [61] СКАРБОРО ДЖ. *Численные методы математического анализа*. – Москва–Ленинград: ГТТИ, 1934.
- [62] АВЕРТН О. *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] MAYER G. *Interval analysis and automatic result verification*. – Berlin: De Gruyter, 2017.

- [66] The NIST reference on constants, units, and uncertainty. – <http://physics.nist.gov/cuu/Constants>
- [67] Pseudospectra gateway. – <http://web.comlab.ox.ac.uk/projects/pseudospectra/>
- [68] Scilab — The Free Platform for Numerical Computation. <http://www.scilab.org>

Обозначения

\Rightarrow	логическая импликация
\Longleftrightarrow	логическая равносильность
$\&$	логическая конъюнкция, связка «и»
\rightarrow	отображение множеств; предельный переход
\mapsto	правило сопоставления элементов при отображении
\leftarrow	оператор присваивания в алгоритмах
\circ	знак композиции отображений
\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
∂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{R}	множество интервалов вещественной оси \mathbb{R}
\mathbb{R}^n	множество вещественных n -мерных векторов
\mathbb{C}^n	множество комплексных n -векторов
\mathbb{IR}^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
$\mathbb{IR}^{m \times n}$	множество интервальных $m \times n$ -матриц
$:=$	равенство по определению
\approx	приблизительно равно
\lesssim	приблизительно меньше или равно
\gtrsim	приблизительно больше или равно
δ_{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 числа $a \in \mathbb{R}$
$[a, b]$	интервал с нижним концом a и верхним 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 $X \subseteq \mathbb{R}^n$
dist	метрика (расстояние)
Dist	мультиметрика (векторнозначное расстояние)
$\operatorname{dom} f$	область определения функции f
$\operatorname{ran}(f, X)$	область значений функции f на X
f^\perp	разделённая разность от функции f
$f(x) _a^b$	разность значений функции f между $x = a$ и $x = b$
$\operatorname{d}f$	дифференциал функции f

$\frac{\partial f}{\partial x_i}$	частная производная функции f по переменной x_i
I	единичная матрица соответствующих размеров
$\ \cdot\ $	векторная или матричная норма
$\langle \cdot, \cdot \rangle$	скалярное произведение векторов
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
$\mathcal{K}_i(A, r)$	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
$\text{lin}\{v_1, \dots, v_n\}$	линейная оболочка векторов v_1, \dots, v_n
$C^p[a, b]$	класс функций, непрерывно дифференцируемых вплоть до p -го порядка на интервале $[a, b]$
$\mathcal{L}^2[a, b]$	класс функций, интегрируемых с квадратом на интервале $[a, b]$
\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{IR} (см. §1.4).

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.

Бабенко, Константин Иванович (1919–1987)

— советский математик и механик.

Банах, Стефан (Stefan Bahach, 1892–1945)

— польский математик.

Бауэр, Фридрих Людвиг (Friedrich Ludwig Bauer, род. 1924)

— немецкий математик.

Бельтрами, Эудженио (Eugenio Beltrami, 1835–1900)

— итальянский математик.

Бернштейн, Сергей Натанович (1880–1968)

— российский и советский математик.

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, Лейтзен Эгберт Ян (Luitzen Egbertus Jan Brouwer, 1881–1966)
— голландский математик.

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 Theophill 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.

Виет, Франсуа (François Viète, 1540–1603)
— французский математик.

Виландт, Хельмут (Helmut Wielandt, 1910–2001)
— немецкий математик.

Гамильтон, Уильям Роуэн (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. *Ἡρώων ο Αλεξανδρεὺς*, около 1 в. н.э.)
— греческий 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.

Дини, Улисс (Ulisse Dini, 1845–1918)

— итальянский математик.

Евклид, или Эвклид (др.-греч. *Ευκλείδης*, 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.⁴

Канторович, Леонид Витальевич (1912–1986)

— советский математик и экономист, известный пионерским вкладом в линейное программирование.

Кеплер, Иоганн (Johannes Kepler, 1571–1630)

— немецкий математик, астроном и механик.

Кнут, Дональд Эрвин (Donald Ervin Knuth, род. 1938)

— американский математик и специалист по информатике и программированию.

Колмогоров, Андрей Николаевич (1903–1987)

— советский математик, внёсший большой вклад во многие разделы современной математики, от 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, род. 1920)

— немецкий математик.

Крамер, Габриэль (Gabriel Cramer, 1704–1752)

— швейцарский математик.

⁴Не 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.
- Кронекер, Леопольд (Leopold Kronecker, 1823–1891)
— немецкий 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.
- Лагранж, Жозеф Луи (Joseph Louis Lagrange, 1736–1813)
— французский математик и механик.
- Ландау, Эдмунд (Edmund Landau, 1877–1938)
— немецкий математик.
- Ланцош, 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.
- 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.

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.

Локуцкий, Олег Вячеславович (1922–1990)

— советский математик.

Ляпунов, Александр Михайлович (1857–1918)

— русский математик и механик, основоположник математической теории устойчивости.

Марков, Андрей Андреевич (1856–1922)

— русский математик.

Марцинкевич, Юзеф (Józef Marcinkiewicz, 1910–1941)

— польский математик.

Микеладзе, Шалва Ефимович (1895–1976)

— советский математик.

Минковский, Герман (Hermann Minkowski, 1864–1909)

— немецкий математик.

Миранда, Карло (Carlo Miranda, 1912–1982)

— итальянский математик.

Муавр, Абрахам де (Abraham de Moivre, 1667–1754)

— английский математик французского происхождения.

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.⁵

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.

⁵Его именем назван спектральный признак устойчивости разностных схем.

- Перрон, Оскар (Oskar Perron, 1880–1975)
— немецкий математик.
- 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.
- 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.
- Рунге, Карл Давид (Karl David Runge, 1856–1927)
— немецкий физик и математик.
- Рутисхаузер, Хайнц (Heinz Rutishauser, 1918–1970)
— швейцарский математик.
- Руффини, Паоло (Paolo Ruffini, 1765–1822)
— итальянский математик.
- Рэлей, Джон Уильям (John William Reyleigh, 1842–1919)
— английский физик.
- 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.
- Таусски, Ольга (Olga Tausski, 1906–1995)
— американский математик.
- Тейлор, Брук (Brook Taylor, 1685–1731)
— английский математик.
- Тихонов, Андрей Николаевич (1906–1993)
— советский математик.
- Томас, Левелин Хиллет (Llewellyn Hilleth Thomas, 1903–1992)
— английский и американский физик и математик.
- Тьюринг, Алан Мэтисон (Alan Mathison Turing, 1912–1954)
— английский математик, логик, криптограф.
- Улам, Станислав (Stanislaw Marcin Ulam, 1909–1984)
— американский математик польского происхождения.
- Фабер, Георг (Georg Faber, 1877–1966)
— немецкий математик.
- Фаддеев, Дмитрий Константинович (1907–1989)
— советский математик.
- Фаддеева, Вера Николаевна (1906–1983)
— советский математик.
- Файк, (С.Т. 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.
- Фишер, Эрнст Сигизмунд (Ernst Sigismund Fischer, 1875–1954)
— немецкий математик.⁶
- Фробениус, Фердинанд Георг (Ferdinand Georg Frobenius, 1849–1917)
— немецкий математик.

⁶Примерно к этому же времени относится жизнь и деятельность известного английского статистика и биолога Рональда Э. Фишера (1890–1962).

- Фрэнсис, Джон (John G.F. Francis, род. 1934)
— английский математик и программист.
- Фурье, Жан Батист (Jean Baptiste Fourier, 1768–1830)
— французский математик и физик.
- 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.
- Холлесский, Андре-Луи (André-Louis Cholesky, 1875–1918)
— французский геодезист и математик.⁷
- Хопф, Хайнц (Heinz Hopf, 1896–1971)
— немецкий и швейцарский математик.
- Хоффман, Алан Джером (Alan Jerome Hoffman, род. 1924)
— американский математик.⁸
- 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.

⁷В русской научной литературе его фамилия нередко транслитерируется как «Холецкий» или даже «Халецкий».

⁸Иногда его фамилию транслитерируют как «Гоффман».

Шур, Исай (Issai Schur, 1875–1941)

— немецкий и израильский математик.

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.

Предметный указатель

- A -норма, 312
- A -ортогональность, 312
- ∞ -норма, 290
- $\mathcal{L}^2[a, b]$, 180
- \mathcal{L}^p , 180
- p -норма, 290
- O -большое, 120
- P -сжатие, 585
- LDL[⊤]-разложение, 378
- ε -решения, 565
- p -ранговое приближение матрицы, 328
- 1-норма, 290
- 2-норма, 172, 290

- LU-разложение, 361

- QR-алгоритм, 537, 540
- QR-разложение, 384

- абсолютная погрешность, 13
- автоматическое дифференцирование, 129, 149
- алгебраическая степень точности, 198
- алгебраический интерполянт, 61
- алгоритм Томаса, 404
- алгоритмическое дифференцирование, 129, 149

- арифметика дифференциальная, 149

- биортогональность, 274

- ведущая подматрица, 265
- ведущий минор, 265
- ведущий элемент, 362
- векторная норма, 289
- верная значащая цифра, 15
- вырожденный интервал, 29

- гёльдерова норма, 291
- главный элемент, 362

- дефект сплайна, 110
- дефектная матрица, 505
- диагонализуемая матрица, 505
- диагональное преобладание, 341
- дифференциальная арифметика, 149
- дифференцирование автоматическое, 129, 149
- дифференцирование алгоритмическое, 129, 149
- дифференцирование символьное, 129
- дифференцирование численное, 130
- длина вектора, 290

доминирующее собственное значение, 523
доминирующий собственный вектор, 523
евклидова норма, 172, 290
естественный сплайн, 121
жорданова форма матрицы, 276
жорданово разложение, 277
задача восстановления зависимостей, 166, 491
задача вычислительно корректная, 561
задача интерполяции, 57
задача интерполяции функции, 57
задача наименьших квадратов линейная, 449, 491
задача некорректная, 37, 148
задача приближения функции, 153
задача сглаживания, 153
задача численного интегрирования, 196
значащая цифра, 15
значимое, 22
индуцированная норма, 305
интегральная метрика, 56
интервал, 27
интервал открытый, 28
интервал полуоткрытый, 28
интервальная арифметика, 29
интервальное расширение, 33
интервальный метод Ньютона, 607
интерполирование, 57
интерполянт, 57
интерполянт алгебраический, 61
интерполяционная квадратурная формула, 199, 211

интерполяция, 57
интерполяция эрмитова, 95
итерационные методы, 352
каноническая форма СЛАУ, 350
каноническая форма Самарского, 479
квадратичное приближение, 161
квадратурная формула, 196
квадратурная формула интерполяционная, 199, 211
классическая интервальная арифметика, 30
ковариационная матрица, 331
коллинеарные векторы, 262
комплексификация, 317
конечные методы, 352
константа Лебега, 93
коэффициент чувствительности, 38
коэффициента Котеса, 215
коэффициенты Фурье, 167
коэффициенты перекося, 513
кратность собственного значения, 500
кратность узла, 95
круги Гершгорина, 516
кубатурная формула, 197
лемма Кеплера, 208
лемма Кэхэна, 444
линейная задача наименьших квадратов, 449, 491
линейная оболочка, 169, 263
линейное подпространство, 262
линейный метод интерполяции, 63
максимум-норма, 290
мантисса, 22
матрица Вандермонда, 62, 339
матрица Гильберта, 182, 338

- матрица Грама, 164
матрица Уилкинсона, 508
матрица вращения, 396
матрица дефектная, 505
матрица диагонализуемая, 505
матрица наклонов интервальная, 611
матрица недефектная, 505
матрица неособенная, 267
матрица неразложимая, 343
матрица нормальная, 506
матрица особенная, 267
матрица отражения, 387
матрица перестановки, 365
матрица почти треугольная, 521
матрица преобуславливающая, 423
матрица простой структуры, 505
матрица разложимая, 342
матрица регулярная, 267
матрица скалярная, 425
матрица строго верхняя треугольная, 432
матрица строго нижняя треугольная, 432
матрица строго регулярная, 369
матрица трёхдиагональная, 405
матрица транспозиции, 365
матрица трапецевидная, 268
матричная норма, 300
матричный ряд Неймана, 321
машинная интервальная арифметика, 49
машинное эписилон, 26
мера диагонального преобладания, 440
метод Гаусса, 357
метод Гаусса-Зейделя, 436
метод Гаусса-Зейделя интервальный, 598
метод Гаусса-Йордана, 352
метод Герона, 587
метод Кравчика, 614
метод Ньютона, 587, 594
метод Ньютона интервальный, 607
метод Ричардсона, 425, 447
метод Хаусхолдера, 390
метод Холесского, 377
метод Шульца, 485
метод Эйлера, 478
метод Якоби, 431
метод ветвлений и отсечений, 616
метод вращений, 398
метод градиентного спуска, 456
метод квадратного корня, 377
метод минимальных невязок, 461
метод наименьших квадратов, 176
метод наискорейшего спуска, 458
метод отражений, 390
метод прогонки, 407
метод простой итерации, 412
метод релаксации, 442
метод сопряжённых градиентов, 474
метод спуска, 454
метод установления, 477
метрика, 56
метрика среднеквадратичная, 162
множество решений, 596
множество тощее, 511
множитель Холесского, 371
модуль непрерывности, 106
мультиметрика, 585
насыщение численного метода, 121
натуральный сплайн, 121
невязка, 413, 441
недефектная матрица, 505
нелинейная интерполяция, 63
ненасыщаемый метод, 121, 235
непрерывность по Липшицу, 39, 108

неравенство Коши-Буняковского, 290
 неравенство Минковского, 158, 291
 нестационарный итерационный процесс, 412
 неявный итерационный метод, 424
 норма, 289
 норма индуцированная, 305
 норма операторная, 305
 норма подчинённая, 305
 норма согласованная, 301
 норма энергетическая, 312
 нормальная матрица, 506
 нормальная система уравнений, 173, 493
 нормальное псевдорешение, 174, 494
 область значений матрицы, 518
 обобщённая степень, 77
 обратная матрица, 267
 обратные степенные итерации, 532
 оператор Кравчика, 613
 оператор Ньютона интервальный, 606
 операторная норма, 305
 операторная форма СЛАУ, 352
 ортогонализация Грама-Шмидта, 184, 399
 ортогонализация Ланцоша, 404
 ортогональная проекция, 264
 ортогональное дополнение, 264
 ортогональность, 263
 основная теорема интервальной арифметики, 34
 остаточный член квадратурной формулы, 197
 осцилляции, 105, 216
 относительная погрешность, 14
 отношение Рэлея, 517

ошибка, 13
 перпендикуляр, 264
 погрешность абсолютная, 13
 погрешность относительная, 14
 подобная матрица, 267
 подпространства Крылова, 401
 подчинённая норма, 305
 полином интерполяционный, 61
 полином интерполяционный Лагранжа, 65
 полином интерполяционный Ньютона, 76
 полиномиальная трудоёмкость, 46
 полиномы Лежандра, 186
 полиномы Чебышёва, 84
 порядок аппроксимации, 137
 порядок точности формулы, 137, 238
 потеря значащих цифр, 19
 потеря точности, 19
 почти решения, 565
 правило Рунге, 252
 предобуславливание, 423
 предобуславливатель, 423
 приближение
 среднеквадратичное, 161
 приведённые полиномы Чебышёва, 89
 признак Адамара, 341
 пример Бабушки-Витасека-Прагера, 41
 пример Бернштейна, 103
 пример Донована-Миллера-Мореланда, 588
 пример Рунге, 104
 принцип вариационный, 448
 принцип релаксации, 441
 проекция, 263
 проекция ортогональная, 264

- простое собственное значение, 500
пространство строго
 нормированное, 157
прямая сумма, 263
прямые методы, 352
псевдометрика, 57
псевдорасстояние, 57
псевдорешение, 172, 494
псевдорешение нормальное, 174
- равномерная метрика, 56
разделённая разность, 66
разложение Жордана, 277
разложение Холесского, 371
разложение Шура, 278
разложение сингулярное, 285
разложение спектральное, 278
разложение треугольное, 361
разностные уравнения
 трёхточечные, 406
разность вперёд, 131
разность назад, 131
расстояние, 56
расщепление матрицы, 424
регрессионная линия, 491
регрессия, 491
рекуррентный вид системы, 415, 558
рекуррентный вид уравнения, 558
ряд Фурье, 167
- сдвиг спектра, 533
сетка, 57, 197
сжатие, 585
сжимающее отображение, 585
символьное дифференцирование, 129
сингулярные векторы, 280
сингулярные числа, 280
система трёхдиагональная, 405
скалярные произведения, 263
след матрицы, 546
собственное значение, 498
собственное число, 498
собственный вектор, 498
согласованная норма, 301
сомнительная значащая цифра, 15
спектр матрицы, 272
спектральная норма, 308
спектральное разложение, 278
спектральный радиус, 315
сплайн, 110
среднеквадратичная метрика, 56, 162
среднеквадратичное
 приближение, 161
стационарный итерационный процесс, 412
стационарный метод Ричардсона, 425
степенной метод, 526
степень сплайна, 110
степень точности алгебраическая, 198
степень точности
 тригонометрическая, 240
строго нормированное пространство, 157
строго регулярная матрица, 369
субдистрибутивность, 31
схема единственного деления, 358
сходимость по норме, 294, 308
сходимость поэлементная, 310
- табулирование, 58
теорема Абеля-Руффини, 503
теорема Алберга-Нильсона, 343
теорема Банаха о неподвижной точке, 585
теорема Бауэра-Файка, 506
теорема Больцано-Коши, 581
теорема Брауэра о неподвижной точке, 600
теорема Вейерштрасса, 102

- теорема Вейля, 509
 теорема Виландта-Хофмана, 509
 теорема Гершгорина, 516
 теорема Кронекера-Капелли, 288
 теорема Леви-Деспланка, 342
 теорема Марцинкевича, 109
 теорема Миранды, 582
 теорема Мысовских, 240
 теорема Островского, 503
 теорема Островского-Райха, 446
 теорема Самарского, 481
 теорема Стеклова-Пойа, 243
 теорема Таусски, 343
 теорема Фабера, 108
 теорема Фредгольма, 288
 теорема Холладея, 122
 теорема Шрёдера о неподвижной точке, 586
 теорема Э. Бореля, 154
 теорема Экарта-Янга, 329
 теорема о разложении Холесского, 371
 теорема о сингулярном разложении, 285
 тест существования решения, 604
 точечная величина, 28
 тощее множество, 511
 трёхдиагональная матрица, 119
 трансекция, 359
 трансформация Гаусса, 493
 трапецевидная матрица, 268
 треугольное разложение, 361
 тригонометрическая степень точности, 240
 тригонометрические полиномы, 60
 узлы сплайна, 110
 условие Дини-Липшица, 107
 устойчивость алгоритма, 40
 формула Муавра, 87
 формула Ньютона-Лейбница, 195
 формула Родрига, 186
 формула Симпсона, 207
 формула квадратурная, 196
 формула кубатурная, 197
 формула парабол, 207
 формула прямоугольников, 200, 240
 формула средних прямоугольников, 200
 формула трапеций, 204
 формулы Гаусса, 220
 формулы Лобатто, 235
 формулы Маркова, 235
 формулы Ньютона-Котеса, 200
 формулы численного дифференцирования, 131, 134
 функционал энергии, 451
 функция Хевисайда, 179
 функция единичного скачка, 179
 функция целая, 104
 характеристика Бекка, 596
 характеристика Оеттли-Прагера, 597
 характеристический полином матрицы, 272
 характеристическое уравнение матрицы, 272, 499
 хессенбергова форма, 521
 целая функция, 104
 целевая функция, 453
 чебышёвская метрика, 56
 чебышёвская норма, 290
 чебышёвская сетка, 91
 чебышёвские узлы, 91
 числа Кристоффеля, 230
 численное дифференцирование, 130
 число обусловленности, 333
 число с плавающей точкой, 22

шаблон, 134
шаг сетки, 77

эквивалентные нормы, 295, 309
экспоненциальная трудоёмкость,
46

экстраполяция, 82
экстремум глобальный, 453
экстремум локальный, 453
элементарная матрица
перестановки, 364

энергетическая норма, 312
энергии функционал, 451
эрмитова интерполяция, 95