Университетский учебник

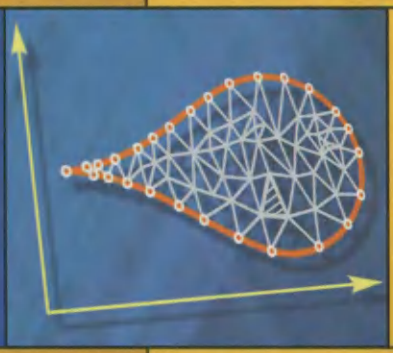
ЧИСЛЕННЫЕ МЕТОДЫ

В двух книгах

Книга 1

Н. Н. Калиткин
Е. А. Альшина

ЧИСЛЕННЫЙ АНАЛИЗ



Прикладная математика
и информатика

ACADEMIA

УНИВЕРСИТЕТСКИЙ УЧЕБНИК

Серия «Прикладная математика и информатика»

ЧИСЛЕННЫЕ МЕТОДЫ

В ДВУХ КНИГАХ

Книга 1

Н. Н. КАЛИТКИН, Е. А. АЛЬШИНА

ЧИСЛЕННЫЙ АНАЛИЗ

Допущено
Учебно-методическим объединением
по классическому университетскому образованию
в качестве учебника для студентов высших
учебных заведений, обучающихся по направлениям
«Прикладная математика и информатика»,
«Фундаментальная информатика и информационные технологии»

Москва
Издательский центр «Академия»
2013

УДК 51(075.8)
ББК 22.1я73
Ч-671

Рецензенты:
д-р физ.-мат. наук, проф. МГУ им. М. В. Ломоносова *А. В. Гулин*;
чл.-кор. РАН, зав. кафедрой вычислительной математики
Московского физико-технического института —
технического университета *А. С. Холодов*

Численные методы : в 2 кн. Кн. 1. Численный анализ :
Ч-671 учебник для студ. учреждений высш. проф. образования /
Н. Н. Калиткин, Е. А. Альшина. — М. : Издательский центр
«Академия», 2013. — 304 с. — (Университетский учебник. Сер.
Прикладная математика и информатика).

ISBN 978-5-7695-5089-8

В учебнике, состоящем из двух книг, изложены основные численные методы решения задач математического анализа, возникающих при исследовании прикладных проблем. Приведенные алгоритмы пригодны для расчетов как на ЭВМ, так и на калькуляторе. Особое внимание уделено нахождению точной оценки погрешности вычислений.

Для студентов учреждений высшего профессионального образования. Может быть полезен аспирантам, преподавателям, научным работникам и инженерам-исследователям, а также лицам, имеющим дело с численными расчетами.

УДК 51(075.8)
ББК 22.1я73

ISBN 978-5-7695-5089-8 (кн. 1)
ISBN 978-5-7695-5090-4

ПРЕДИСЛОВИЕ

Использование компьютеров для вычислений позволило от простейших расчетов и оценок различных конструкций или процессов перейти к новой стадии работы — детальному математическому моделированию (вычислительному эксперименту), которое существенно сокращает потребность в дорогостоящих, а нередко даже опасных натурных экспериментах.

В основе вычислительного эксперимента лежит решение уравнений математической модели численными методами.

Сложные вычислительные задачи, возникающие при исследовании различных физических и технических проблем, можно разделить на ряд элементарных: вычисление интеграла, решение дифференциального уравнения и т. п. Многие элементарные задачи являются несложными и хорошо изучены. Для этих задач уже разработаны методы численного решения и обычно имеются стандартные программы. Возникает вопрос: «Нужно ли практику-вычислителю изучать численные методы при наличии стандартных программ?»

Ответ на вопрос дает следующий пример. Известно, что $\sin x$ можно разложить в ряд Тейлора, который знакопеременен и сходится при любом значении x. Согласно классической математике, погрешность частичной суммы этого ряда не превышает первого отброшенного члена. Составленная программа для 64-разрядного компьютера обрывает вычисления, если очередной член ряда Тейлора меньше 10^{-8}. При задании угла $x = 2\,550°$ был получен ответ: $\sin 2\,550° = 29,5\ldots$!!! Из данного примера следует очевидный вывод — полагаться на мощность компьютера и незнакомую программу рискованно, следует знать численные методы, включая тонкости вычисления алгоритмов.

Для каждой задачи существует множество методов решения. Например, хорошо обусловленную систему линейных уравнений можно решать методами Гаусса, Жордана, оптимального исключения, окаймления, отражений, ортогонализации и др. Интерполяционный многочлен записывают в формах Лагранжа, Ньютона, Грегори — Ньютона, Бесселя, Стирлинга, Гаусса и Ла-

пласа — Эверетта. Подобные методы обычно являются вариациями одного-двух основных методов, и даже если в каких-то частных случаях они имеют преимущества, то незначительные. Кроме того, многие методы были созданы до появления компьютеров, и ряд из них в качестве существенного элемента включает интуицию вычислителя. Компьютерное вычисление потребовало переоценки существующих методов, поскольку эффективность многих методов сильно зависит от мелких деталей алгоритма, почти не поддающихся теоретическому анализу; поэтому окончательно отобрать лучшие методы можно лишь, используя большой опыт практических расчетов. Попытка такого отбора сделана авторами в данном учебнике на основе многолетнего опыта решения большого числа разнообразных задач математической физики. Для большинства рассмотренных в книге задач изложены только наиболее эффективные методы с широкой областью применимости. Несколько методов для одной и той же задачи даны в том случае, если они имеют существенно разные области применимости или если для этой задачи еще не разработаны достаточно удовлетворительные методы.

Изложению численных методов посвящено немало книг, однако большинство из них ориентировано на студентов и научных работников математического профиля. Данный же учебник предназначен для широкого круга читателей — как учебник для студентов и аспирантов физических и технических специальностей и как справочное пособие для научных сотрудников, инженеров и математиков-вычислителей. Авторы старались сочетать простоту изложения, разумную степень строгости, умеренный объем и широту охвата материала. Большое внимание в книге уделено рекомендациям по практическому применению алгоритмов; изложение пояснено рядом примеров. Для обоснования алгоритмов использован несложный математический аппарат, знакомый студентам физических и инженерных специальностей.

Книга является полным курсом численных методов для физических и инженерных специальностей; для математических специальностей вузов она может быть вводным курсом численных методов, после которого слушатели могут изучать углубленные спецкурсы.

Учебник написан на основе курса лекций, читавшихся вначале инженерам-конструкторам, а после переработки курса — студентам физического факультета МГУ им. М. В. Ломоносова и некоторых других вузов. Он дополнен рядом актуальных разделов (решение жестких систем, задачи в неограниченных обла-

стях и т. п.), состоит из двух книг: книга 1 содержит материал по численному анализу, книга 2 — численные методы решения дифференциальных (обыкновенных и в частных производных) и интегральных уравнений.

Учебник разделен на главы и подразделы. Формулы имеют двойную нумерацию: номер главы и порядковый номер формул в главе; то же относится к нумерации рисунков и таблиц. Конец доказательства теоремы отмечен знаком ■. Приведенные в списке литературы учебники и монографии рекомендуются для углубленного изучения отдельных разделов.

Общий подход к теории и практике вычислений, определивший стиль этой книги, сложился у авторов под влиянием многолетней совместной работы с А. А. Самарским и В. Я. Гольдиным. Ряд актуальных тем был включен по инициативе А. Г. Свешникова и В. Б. Гласко. Много ценных замечаний сделали В. Ф. Бутузов, А. В. Гулин, Б. Л. Рождественский, И. М. Соболь, И. В. Фрязинов, Е. В. Шикин. В оформлении рукописи большую помощь оказали Л. В. Кузьмина и Т. Г. Ермакова. Искренне благодарим всех названных лиц и особенно Александра Андреевича Самарского.

Глава 1

О ЧИСЛЕННОМ АНАЛИЗЕ

1.1. НЕМНОГО ИСТОРИИ

1.1.1. Развитие численных методов

Предположительно численные методы впервые были использованы в XIV — XV вв. (во времена существования Древнего Египта и Древней Греции) для подсчета доходов, площадей или объемов.

Начиная с XVII в. появляются принципиально новые задачи, связанные с возникновением небесной механики и дифференциального исчисления. Примером может служить грандиозная по тем временам исследовательская задача — построение геодезической сети на поверхности Земли (она требует аккуратного уравнивания сумм углов треугольников на сферической поверхности) — чрезвычайно важной для навигации. Эта задача решалась во Франции в годы правления Людовиков XV и XVI и закончена была уже после Великой французской революции. Необходимо отметить, что финансирование науки в те времена считалось приоритетным как для власти королей, так и революционеров. Даже мадам Помпадур жертвовала немалые суммы на нужды науки.

Аккуратный расчет движения небесных тел потребовал составления таблиц тригонометрических функций и логарифмов. В XIX в. появилась логарифмическая линейка.

В период Первой мировой войны повысились требования к скорости расчетов, так как стрельбу приходилось вести по движущимся целям — танкам и кораблям. В связи с этим были созданы первые механические, а затем и аналоговые вычислительные приборы.

Ко времени Второй мировой войны требования к быстроте вычислений многократно возросли, так как появились самолеты, скорость которых существенно превышала скорость танков.

Это вызвало появление первых электронных вычислительных приборов.

В 1945 г. в Лос-Аламосе (США) проводилась интенсивная разработка первой атомной бомбы тремя группами ученых одновременно. Первая группа состояла из одного человека — П. Ферми, который вел расчеты на листке бумаги, вторая (также из одного человека — барона Дж. фон Неймана) использовала логарифмическую линейку, третья подготовила программу для первой в мире ЭВМ «MANIAC», быстродействие которой было настолько слабым (сотни операций в секунду при мизерной оперативной памяти), что все три группы пришли к ответу практически одновременно. Какие именно цифры прогноза получила каждая из этих групп, американцы не сообщали (судя по пари, которые заключили между собой участники первого испытания, они различались в несколько раз).

Стремительное и продолжающееся по настоящее время развитие электронной вычислительной техники быстро изменило ситуацию и сделало применение ЭВМ повсеместным. Если еще лет 10 назад основным требованием к программным продуктам была экономичность, то сейчас быстродействие ЭВМ настолько велико, что на передний план выходит надежность. На данный момент даже в два раза более быстрый, но менее надежный метод не может считаться предпочтительным.

В МГУ им. М. В. Ломоносова разделение физико-математического факультета на физиков и математиков произошло лишь в 1933 г. Профессор А. Н. Тихонов, возглавлявший кафедру математики еще до Великой Отечественной войны, первым высказал мнение, что преподавание математики для физиков должно отличаться от классического преподавания математики на механико-математическом факультете, где упор делается на теоремы существования решения. Это мнение иллюстрирует высказывание Л. Ландау: «Зачем мне теорема существования электрона? Я и так знаю, что он существует!». По мнению А. Н. Тихонова, акцент преподавания математики для физиков должен быть сделан на доведение решения до конечного ответа, в частности на технику вычислений (что не исключает изложения теорем существования и единственности решения).

В конце 40-х годов XX в. активизировалась работа над созданием отечественной атомной бомбы. Когда стало ясно, что необходимый запас плутония вскоре будет накоплен, потребовалось рассчитать мощность взрыва. Поначалу задача решалась в рамках уравнений газовой динамики, описывающих в том числе

ударную волну. Однако встал вопрос о том, как математически оценить деление урана, движение возникших нейтронов, разлет вещества под действием выделившейся энергии. В группе Ландау данный процесс был представлен с помощью обыкновенных дифференциальных уравнений с усреднением по плотности для облегчения численного расчета. Но с какой точностью будет рассчитана мощность взрыва в рамках этой модели, оценить было сложно. В 1948 г. на совещании, посвященном этой проблеме, А. Н. Тихонов, впервые ознакомившись с проблематикой, высказал идею о том, как численно решить систему дифференциальных уравнений в частных производных с реальными (неусредненными) плотностями. Всего за несколько месяцев была разработана методика и проведены расчеты, которые почти совпали с данными эксперимента 1949 г. Поэтому данная область знаний получила широкое развитие, в частности, на кафедре математики физического факультета МГУ им. М. В. Ломоносова, работы велись под руководством А. Н. Тихонова и А. А. Самарского.

Таким образом, общая тенденция развития вычислительной техники и численных методов была такова: в США всегда были более мощные машины, чем в России, но российские ученые использовали более совершенные модели и численные методы.

1.1.2. Теории и модели

Уже в глубокой древности считалось, что наука занимается непреложными и неизменными истинами: естественные науки открывают законы природы, а математика придает этим законам совершенную форму. Древнегреческий ученый Архимед открыл закон рычага и выталкивающую силу. Сильнейшее подтверждение этой точки зрения предоставил И. Ньютон в труде «Математические начала натуральной философии» (1687), где сформулированы основные законы механики. Им же был создан математический аппарат — дифференциальное и интегральное исчисления (одновременно с ним и даже чуть раньше то же сделал Г. В. Лейбниц), что позволило дать строгие математические формулировки задач и многие из них точно решить.

Триумфом механики стали выведенные Ньютоном законы движения небесных тел. С этого момента началось бурное развитие механики. Формулировались новые частные задачи — одни точно решались, для других создавались численные методы решения. Появились новые разделы физики; для описания

задач сплошной среды разработан аппарат уравнений в частных производных. Блестящим завершением этого периода стали уравнения Дж. Максвелла для электромагнитного излучения (их долго не хотели признавать, но эксперименты Г. Герца по обнаружению электромагнитных волн и П. Н. Лебедева по измерению давления подтвердили точность теории). В начале XX в. Х. А. Лоренц, А. Пуанкаре и А. Эйнштейн создали специальную теорию относительности. Классическая ньютоновская механика оказалась частным случаем релятивистской механики, справедливым при скоростях много меньших скорости света. То, что в течение трех веков считалось точным законом природы, оказалось лишь неким приближением.

В 30-е годы XX в. Э. Шредингер и В. фон Гейзенберг заложили основы квантовой механики. Ньютоновская механика оказалась частным случаем квантовой, когда размеры тел становятся много больше атомных. Наконец, П. Дирак создал еще более общую теорию — релятивистскую квантовую механику, также немедленно подтвержденную экспериментами исключительно высокой точности. Кроме того, теория Дирака предсказала, что кроме электрона должна существовать частица с той же массой, но положительным зарядом — позитрон. И вскоре позитрон был обнаружен в экспериментах.

В результате ученые осознали, что абсолютно точных фундаментальных законов — истин в последней инстанции — нет. Любой закон является лишь приближенным описанием природы, частным случаем более общего закона (быть может, еще не открытого). Он достаточно точно выполняется лишь в определенных условиях, например при не слишком больших скоростях, или не слишком малых размерах тел и т. п. Нужно только четко формулировать эти условия применимости и следить за тем, чтобы не выходить за их границы.

Но ведь приближениями пользовались намного раньше. Например, уравнения газодинамики (уравнения Эйлера) достаточно сложные; точно они не решаются, а решать их численно довольно трудоемко. Но если нас интересуют не любые движения масс газа, а лишь небольшие колебания давления, то удается приближенно заменить эти уравнения гораздо более простым уравнением акустики. Его нетрудно точно решить, и оно превосходно описывает многие явления, например распространение звука. Но для описания сильного взрыва уравнение акустики уже не пригодно — здесь изменение давления велико, и нарушены условия применимости.

Такие приближения строились с XVIII в., проверялись, становились классическими — входили в золотой фонд науки, всесторонне изучались. Однако с 40-х годов XX в. стремительное развитие техники, основанной на сложных физических и химических принципах, потребовало аккуратного проектирования и тщательного расчета новых конструкций. Пришлось разрабатывать специальное приближение для каждой конкретной задачи или явления. Такие приближения стали называть ***моделями*** данного явления.

1.2. МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

1.2.1. Математическая модель

Различают две стадии построения модели. Сначала обсуждают разные стороны и процессы данного явления. При этом оценивают, какие процессы и факторы обязательно надо учесть, а какие пренебрежимо малы и могут быть отброшены. Отобранные факторы составляют ***предметную*** (механическую, физическую, химическую, биологическую, социологическую и т. п.) модель явления. Затем отобранные факторы описывают математическими уравнениями (алгебраическими, дифференциальными, интегральными и т. п.). Эту совокупность уравнений называют ***математической моделью***.

Поясним это на примерах баллистических задач.

Пример 1.1. Пусть тело брошено со скоростью v_0 под углом α к горизонту с высоты y_0 (рис. 1.1). Учтем только силу притяжения, действующую вертикально вниз, это и есть физическая модель. Тогда согласно ньютоновской механике, движение по горизонтали будет равномерным, а по вертикали — равноускоренным:

$$x = v_0 t \cos\alpha_0; \quad y = y_0 + v_0 t \sin\alpha_0 - gt^2/2, \tag{1.1}$$

где t — время; $g \approx 10$ м/с^2 — ускорение свободного падения.

Уравнения (1.1) являются математической моделью и одновременно дают решение в параметрической форме (роль параметра играет t). Можно исключить t из (1.1) и получить траекторию полета, которая оказывается параболой:

$$y = y_0 + x \operatorname{tg}\alpha_0 - gx^2/(2v_0^2 \cos^2\alpha_0). \tag{1.2}$$

Полагая $y = 0$, найдем дальность броска:

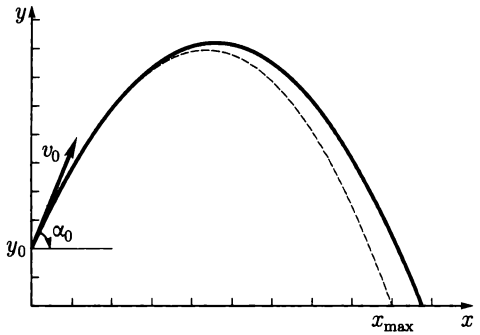


Рис. 1.1. Полет молота (сплошная линия) и волана (штриховая линия)

$$x_{\max} = v_0 \cos\alpha_0 \left(v_0 \sin\alpha_0 + \sqrt{v_0^2 \sin^2\alpha_0 + 2gy_0} \right) \Big/ g \xrightarrow[y_0 \to 0]{} \xrightarrow[y_0 \to 0]{} v_0^2 \sin 2\alpha_0 / g. \tag{1.3}$$

Дальность броска зависит от угла α_0. Можно даже явно найти оптимальный угол, обеспечивающий наибольшую дальность броска:

$$\sin\alpha_{\text{opt}} = 1/\sqrt{2(1 + gy_0/v_0^2)} \xrightarrow[y_0 \to 0]{} 1/\sqrt{2};$$
$$x_{\text{opt}} = v_0 \sqrt{v_0^2 + 2gy_0}/g \xrightarrow[y_0 \to 0]{} v_0^2/g. \tag{1.4}$$

При $y_0 = 0$ $\alpha_{\text{opt}} = 45°$ (что знают даже школьники), а при $y_0 > 0$ он меньше.

Эта модель очень проста. Но где она применима? Она хорошо описывает полет небольших сравнительно массивных тел с умеренными скоростями — бросок камня, толкание ядра, метание молота. Обработка спортивных киносъемок показывает, что лучшие метатели молота выпускают снаряд под углом 42 — 43° (бросок с высоты плеча $y_0 \approx 1,5$ м), а начальная скорость $v_0 \approx 20$ м/с обеспечивает дальность $x_{\text{opt}} \approx 41 - 42$ м.

Пример 1.2. Попытаемся применить модель к другим объектам. Круглая пуля времен войны 1812 г. летела со скоростью $v_0 > 100$ м/с, и, согласно формуле (1.4), дальность ее полета могла бы составлять 1 км и более. Однако дальнобойность ружей тогда не превышала 200 м. Очень нагляден полет волана в бадминтоне: с какой скоростью его ни посылай, он дальше ~ 36 м не полетит; вдобавок хорошо видно, что траектория его полета не

параболическая, а имеет более крутое снижение (см. штриховую линию на рис. 1.1).

Причина легко угадывается — надо учесть сопротивление воздуха. Сила сопротивления $\mathbf{F}$ направлена обратно скорости $\mathbf{v}$ и при средних (дозвуковых) скоростях примерно пропорциональна квадрату скорости, т. е. $\mathbf{F} \approx -k(v)\mathbf{v}$, где $k(v) \approx k_0 v$. Коэффициент k_0 зависит от размеров и формы тела и свойств воздуха (температуры и плотности). Примем эту физическую модель и запишем математическую модель — ньютоновские уравнения движения для координат x, y и компонент скоростей v_x, v_y:

$$\frac{dx}{dt} = v_x; \quad \frac{dy}{dt} = v_y; \quad \frac{dv_x}{dt} = -\frac{k(v)}{m} v_x; \quad \frac{dv_y}{dt} = -g - \frac{k(v)}{m} v_y, \tag{1.5}$$

где m — масса тела;

$$v = (v_x^2 + v_y^2)^{1/2}. \tag{1.6}$$

Уравнение нужно дополнить начальными условиями при $t = 0$:

$$x(0) = 0, \; y(0) = y_0; \; v_x(0) = v_0 \cos \alpha_0; \; v_y(0) = v_0 \sin \alpha_0. \tag{1.7}$$

Новая модель значительно сложнее, и решить задачу (1.5) — (1.7) явно уже не удается. Однако она существенно точнее. Нетрудно численно решить ее на компьютере и увидеть все те эффекты — несимметричность траектории и уменьшение дальности, о которых говорилось ранее. Для малых начальных скоростей $k(v) \approx 0$ модель (1.5) — (1.7) переходит в (1.1).

Пример 1.3. Заметим, что для модели (1.5) — (1.7) удается найти частный случай, где строится точное решение. Рассмотрим полет вертикально брошенного тела: $\alpha_0 = 90°$. Тогда $v_x(t) \equiv 0$ и $x(t) \equiv 0$, а оставшиеся уравнения принимают следующий вид:

$$\frac{dy}{dt} = v; \quad \frac{dv}{dt} = -g \mp \frac{k_0}{m} v^2; \quad y(0) = y_0; \quad v(0) = v_0 > 0, \tag{1.8}$$

здесь знак «$-$» соответствует полету вверх, а знак «$+$» — вниз.

Уравнение для скорости интегрируется точно:

- при движении вверх

$$v(t) = \frac{1}{\sigma} \frac{\sigma v_0 - \operatorname{tg}(\sigma g t)}{1 + \sigma v_0 \operatorname{tg}(\sigma g t)}; \quad \sigma = \sqrt{\frac{k_0}{mg}}; \quad 0 \leqslant t \leqslant t_{\text{up}}, \tag{1.9}$$

где время подъема

$$t_{\text{up}} = \frac{1}{\sigma g}\operatorname{arctg}(\sigma v_0) \xrightarrow[v_0\to\infty]{} \frac{\pi}{2\sigma g}; \tag{1.10}$$

- при движении вниз

$$v(t) = -\frac{1}{\sigma}\operatorname{th}[\sigma g(t - t_{\text{up}})]; \quad t \geqslant t_{\text{up}}; \quad v(+\infty) = -\frac{1}{\sigma}. \tag{1.11}$$

Время подъема оказалось конечным, сколь бы большой ни была начальная скорость, а скорость падения не превышает некоторой предельной, что кардинально отличается от модели (1.1).

Высота подъема находится интегрированием скорости (1.11)

$$y(t) = y_0 + \int_0^t v(\tau)d\tau. \tag{1.12}$$

Этот интеграл также точно берется, но формулы различны для стадий подъема $0 \leqslant t \leqslant t_{\text{up}}$ и падения $t_{\text{up}} \leqslant t$; первая из этих формул довольно громоздка. Значение $y_m = y(t_{\text{up}})$ определяет максимальную высоту подъема.

Пример 1.4. Рассмотрим еще один частный случай, когда модель (1.5)—(1.7) имеет точное решение. Это настильная стрельба (стрельба при малом начальном угле α_0). Если $\alpha_0 \leqslant 8°$, то $v_{y0} < 0,15v_{x0}$. В этом случае с точностью до 1 % можно считать, что $v \approx v_x$. Тогда уравнения модели (1.5)—(1.7) принимают следующий вид:

$$\frac{dx}{dt} = v_x; \quad \frac{dv}{dt} = -\frac{k_0 v_x^2}{m}; \quad \frac{dy}{dt} = v_y; \quad \frac{dv_y}{dt} = -g - \frac{k_0}{m}v_x v_y. \tag{1.13}$$

Уравнение для скорости v_x точно интегрируется и дает

$$v_x(t) = v_{x0}\Big/\left(1 + \frac{k_0}{m}v_{x0}t\right). \tag{1.14}$$

Интегрируя горизонтальную скорость (1.14), получим уравнение для горизонтальной координаты:

$$x(t) = \int_0^t v_x(\tau)d\tau = \frac{m}{k}\ln\left(1 + \frac{k_0}{m}v_{x0}t\right). \tag{1.15}$$

Уравнение для вертикальной скорости можно переписать в следующем виде:

$$\frac{dv_y}{dt} = -g - \frac{k_0 v_x(t)}{m} v_y,$$

где $v_x(t)$ определяется по формуле (1.14). Это линейное однородное уравнение, и оно также имеет точное решение:

$$v_y(t) = \left(v_{y0} - gt - \frac{g k_0 v_{x0}}{2m} t^2\right) \Big/ \left(1 + \frac{k_0 v_{x0}}{m} t\right). \tag{1.16}$$

Вертикальная координата также точно находится

$$y(t) = y_0 + \int_0^t v_y(\tau) d\tau = y_0 + \frac{m}{k_0 v_{x0}} \left(v_{y0} + \frac{mg}{2k_0 v_{x0}}\right) \times \\ \times \ln\left(1 + \frac{k_0 v_{x0}}{m} t\right) - \frac{mg}{2k_0 v_{x0}} t - \frac{g}{4} t^2. \tag{1.17}$$

Формулы (1.15) и (1.17) задают траекторию в параметрическом виде через параметр t.

Из условия $v_y(t) = 0$ можно найти время подъема:

$$t_{\text{up}} = \frac{2m}{g} \Big/ \left[1 + \left(1 + \frac{2k_0 v_{y0} v_{x0}}{mg}\right)^{1/2}\right] < \frac{v_{y0}}{g}. \tag{1.18}$$

Подставив (1.18) в (1.15) и (1.17), получим явное выражение координат верхней точки траектории. Можно также явно найти наклон любой точки траектории:

$$\operatorname{tg} \alpha(t) \equiv v_y(t)/v_x(t) = \frac{v_{y0}}{v_{x0}} - \frac{gt}{v_{x0}} - \frac{k_0 g}{2m} t^2. \tag{1.19}$$

Расчет по формулам (1.14) — (1.19) дает траекторию, соответствующую штриховой линии на рис. 1.1; ее нисходящая ветвь идет круче восходящей.

Пример 1.5. Современные винтовки и орудия изготовляют с высокой точностью. Обычная снайперская винтовка может поразить цель на расстоянии до 800 м, крупнокалиберная — на 1,5 — 2 км, а дальнобойное морское орудие при стрельбе на 30 км дает рассеивание ± 3 м. Однако для точного попадания в цель надо правильно установить прицел, т. е. верно определить угол α. Расчет этих углов выполняют по моделям типа (1.5), но дополненных еще рядом эффектов:

• в настоящее время начальные скорости v_0 достигают следующих значений: для пуль $\sim$ 1 км/с, для снарядов $\sim$ 2 км/с, поэтому зависимость $k(v)$ становится более сложной;

• коэффициент k_0 зависит от плотности и температуры воздуха, которые могут меняться вдоль трассы полета, что требует внесения поправок;

• нужно учитывать движение цели;

• необходимо вносить поправки на скорость и направление ветра;

• другие поправки (в морском бое вносят даже поправку на вращение Земли!).

Ранее такие поправки предварительно рассчитывали и распечатывали в виде книжечки, которую артиллеристы носили с собой. Но уже в Первую мировую войну у моряков появились механические устройства для автоматизации наводки. Сейчас соответствующие программы включены в компьютерные системы управления огнем. Но снайперы по-прежнему держат таблицу стрельб в голове.

Из рассмотренных примеров видна ***общая тенденция***. Если учитывать немного важнейших эффектов, то может получиться достаточно простая модель. Ее нетрудно будет численно рассчитать, а возможно и получить явное решение. Однако модель будет грубой и применимой лишь к сильно ограниченному кругу явлений.

Если учитывать слишком много эффектов, то модель окажется весьма точной, но очень сложной и неизвестно, сумеем ли мы провести по ней расчеты — найдутся ли подходящие алгоритмы и хватит ли мощности компьютера.

Поэтому оптимальная модель — разумный компромисс между требованиями к полноте и точности модели и вычислительными возможностями. К счастью, быстрое развитие компьютеров позволяет постоянно улучшать модели.

Однако заметим, что это лишь тенденция, а не закон. Известны примеры, когда более простая модель оказывалась и более точной. Так, все средневековые астрономы предсказывали положение планет, солнечные и лунные затмения по геоцентрической модели Птолемея, согласно которой планеты, Луна и Солнце движутся по малым кругам-эпициклам, центры которых вращаются вокруг Земли по большим кругам-циклам. И. Кеплер предложил простую гелиоцентрическую модель с эллиптическими орбитами, оказавшуюся более точной. И только Ньютон объяснил, почему она правильна.

Поэтому построение хорошей модели — не просто наука, но и искусство.

Возникает вопрос: «А зачем это нужно знать математику-прикладнику?» Ведь он только решает предложенную математическую задачу, а формулируют модель другие. Ответ на этот вопрос будет дан позже.

1.2.2. Модель — алгоритм — программа

После того как математическая модель построена, ее уравнения нужно решить. Для простейших моделей типа (1.1) удается получить решение в явном виде. Большинство моделей требует численного решения, и нужно выбрать или построить алгоритм расчета.

Для несложных моделей удается обойтись описанными в учебниках алгоритмами: одним или комбинацией нескольких. Так, для системы дифференциальных уравнений (1.5) — (1.7) хорошие результаты даст численное интегрирование по явным схемам Рунге — Кутты, желательно не ниже 4-го порядка точности. Можно воспользоваться стандартными программами, имеющимися во многих пакетах математических программ. Но при этом надо четко представлять, какова погрешность расчета, предусмотрен ли в программе контроль точности и насколько можно ему доверять. Во многих случаях данный программный контроль оказывается иллюзией (примеры этого будут приведены в книге 2 данного курса).

Для сложных моделей имеющиеся алгоритмы могут оказаться непригодными или малоэффективными. Тогда приходится разрабатывать оригинальные алгоритмы, обосновывать их точность и отлаживать программу. Для контроля правильной работы алгоритмов и программ полезно использовать частное точное решение типа (1.9) — (1.11), если их удается найти. Далее в книге будут показаны некоторые приемы такого контроля. Есть и способы проверки, не требующие знания точных решений.

Наконец, программа надежно отлажена и начались расчеты. Но успокаиваться еще рано. Надо вернуться к исходному явлению и взять несколько экспериментов, проведенных в заметно различающихся условиях (например, броски с разными скоростями v_0 под разными углами α). Если все расчеты хорошо совпадут с экспериментально измеренными длинами или киносъемками траекторий — мы справились с задачей. Если оказались заметные расхождения, то надо заново проверять все. При

этом заказчик уверен, что его модель правильна, а виноват математик — неверно написал алгоритм или программу («В любой сколь угодно малой программе есть по меньшей мере одна ошибка»). На самом деле и модель может чего-то существенного не учесть. Даже в не такой уж сложной модели полета снаряда (см. пример 1.5) список факторов можно продолжить: вращение снаряда, прецессия оси вращения, зависимость ускорения свободного падения g от высоты y (меняется расстояние от центра Земли), снижение плотности атмосферы с высотой. Поэтому математик должен разобраться и в модели. А возможно, контрольные эксперименты были проведены неаккуратно — их также нужно проверить.

1.3. ИСТОЧНИКИ ПОГРЕШНОСТИ

Термины «численные методы» и «приближенный анализ» — синонимы. Всякий раз точная задача заменяется приближенной. Например, по заданной величине w нужно вычислить u. Символически запишем операцию как

$$u = A(w). \tag{1.20}$$

Рассмотрим следующий пример: $u = A(w) = \int_a^b w(x)dx$. Интегралы даже от простых комбинаций элементарных функций далеко не всегда удается взять точно. Возможны следующие способы упрощения задачи:

1) $w(x) \approx \tilde{w}(x) = P(x)$.

По теореме Вейерштрасса всякую гладкую функцию можно приблизить полиномами, а интегралы от полиномов берутся точно;

2) можно приближенно заменить интеграл интегральной суммой:

$$\int_a^b w(x)dx = A \approx \tilde{A} = \sum w(x_i)h_i;\ \tilde{A} \to A \text{ при } h \to 0;$$

3) можно взять комбинацию первых двух способов $w(x) \approx \tilde{w}(x)$ и $A \approx \tilde{A}$.

Будем считать, что и решение приближенной задачи близко к точному $\tilde{u} \approx u$. Но как оценить точность $\tilde{u} - u$? Для этого понадобится понятие нормы.

1.3.1. Величины и нормы

Как исходные данные, так и решение могут быть величинами различных типов. Например, числа u, w; векторы разной размерности $\mathbf{u} = \{u_p, 1 \leqslant p \leqslant P\}$, $\mathbf{w} = \{w_q, 1 \leqslant q \leqslant Q\}$; матрицы; функции одной переменной $u(x)$, $w(y)$ или многих переменных; вектор-функции и т. п. При этом аргумент (аргументы) функции может быть непрерывным $a \leqslant x \leqslant b$ или дискретным $x \in \Omega$, где $\Omega = \{x_n, 1 \leqslant n \leqslant N\}$ есть некоторая сетка.

Проиллюстрируем изложенное некоторыми примерами:

- решается уравнение с одним неизвестным; u, w суть числа, вещественные или комплексные;
- решается система N линейных или нелинейных уравнений с таким же числом неизвестных; u, w суть векторы одинаковой размерности N;
- ищется определенный интеграл от $w(x)$; последняя есть функция непрерывного аргумента, u есть число;
- строится сплайн-аппроксимация функции, табулированной на сетке Ω; исходные данные — функция дискретного аргумента $w(x_n)$, которую также можно рассматривать как вектор $\{w_n\}$; решение $u(x)$ есть функция непрерывного аргумента;
- решается дифференциальное уравнение $du/dx{=}w(u, x)$; исходные данные — непрерывная функция двух аргументов $w(u, x)$, решение есть непрерывная функция одного аргумента $u(x)$. Однако для численного интегрирования вводится сетка $\{x_n\}$, так что численное решение оказывается функцией дискретного аргумента $u(x_n)$, т. е. вектором.

Количественной мерой точности является норма погрешности. Укажем некоторые наиболее употребительные нормы. Для числа u есть единственная норма

$$\|u\| = |u|.$$

Для ограниченных функций $u(x)$, $x \in [a, b]$, вводится чебышевская норма

$$\|u\|_C = \max_{x \in [a,b]} |u(x)|, \tag{1.21}$$

а для функций, интегрируемых с квадратом с весом $\rho(x)$ — гильбертова норма:

$$\|u\|_{L_2} = \left[\int_a^b u^2(x) \rho(x) dx \right]^{1/2}, \quad \rho(x) > 0. \tag{1.22}$$

Для функций дискретного аргумента $u(x_n)$ или векторов $\{u_n\}$ вводят дискретные аналоги норм (1.21) и (1.22):

$$\|u\|_C \max_{1\leqslant n\leqslant N} |u_n|; \quad \|u\|_{l_2} = \left(\sum_{n=1}^{N} \rho_n u_n^2\right)^{1/2}, \quad \rho_n > 0. \qquad (1.23)$$

Для матриц употребляют еще большее число норм.

Видно, что для одной и той же величины могут использоваться разные нормы.

Между нормами существуют определенные соотношения. Для функций непрерывного аргумента это односторонние неравенства; так, нетрудно проверить, что

$$\|u\|_C \left[\int_a^b \rho(x)dx\right]^{1/2} \geqslant \|u\|_{L_2}.$$

При этом из малости нормы, стоящей в левой части, следует малость нормы, стоящей в правой части, но не наоборот; то же относится к сходимости методов в данных нормах. Первую норму называют более ***сильной***, чем вторую. Наглядное отличие между нормами (1.21) и (1.22) таково: малость нормы C означает, что $u(x)$ мала во всех точках $[a, b]$; малость нормы L_2 означает, что $u(x)$ мала почти во всех точках, но на незначительной части $[a, b]$ может быть не мала.

Для функций дискретного (конечномерного) аргумента неравенства между нормами оказываются двусторонние. Например, для (1.23) выполняется

$$\|u\|_C \left(\sum_{n=1}^{N} \rho_n\right)^{1/2} \geqslant \|u\|_{l_2} \geqslant \|u\|_C \left(\min_{1\leqslant n\leqslant N} \rho_n\right)^{1/2}.$$

Поэтому из сходимости в одной норме следует сходимость в другой.

Такие нормы называют ***эквивалентными***. В конечномерных пространствах все нормы эквивалентны, но в бесконечномерном пространстве это не так.

Различают четыре источника и четыре составные части погрешности решения: погрешность модели, погрешность исходных данных, погрешность метода и погрешность округлений. Рассмотрим их подробнее.

1.3.2. Погрешность модели

Погрешности модели — это разница между физической и математической задачами. Как видно из примеров 1.1 — 1.5, учет или не учет тех или иных процессов приводит к разным уравнениям и различным результатам. Если математическая модель заведомо плоха (не учтены важные эффекты), то и численно решать задачу вовсе не стоит. Приведем некоторые известные примеры неудачных математических моделей.

1. Первый проект управляемого термоядерного синтеза предложен еще в 1956 г. Тем не менее результат до сих пор не получен. Идея проекта заключается в том, чтобы получить изолированную плазму, удерживаемую магнитным полем, и создать в ней высокую температуру. Но удовлетворяет ли плазма написанным для нее дифференциальным уравнениям? Вопреки теории плазменный шнур вырывается из магнитного поля, стенки сосуда испаряются и температура плазмы падает. Эту неустойчивость пытаются подавить, изменяя конструкцию магнитной ловушки. Однако каждый раз возникает новая неустойчивость. Известно и даже теоретически объяснено уже более 130 неустойчивостей плазмы в линейном приближении, но число их все растет. А все дело в том, что нет модели (уравнения), адекватно описывающей поведение разреженной плазмы в магнитном поле.

2. Озоновые дыры принято объяснять выбросами фреона в атмосферу. Модели построены на теории диффузии реагирующих газов в атмосфере. Однако система химических реакций плохо известна: нет полной цепочки реакций, константы реакций известны с погрешностями в десятки раз и более. Не удивительно, что вопреки развитой модели озоновая дыра в Антарктиде растет ничуть не меньше, чем над населенной частью Земли, хотя выбросов фреона там нет.

3. В парниковом эффекте также принято винить деятельность человека, сопряженную с повышенным выбросом диоксида углерода CO_2 и метана CH_4. Однако нет хорошей модели круговорота CO_2 в природе. Почвенные бактерии вырабатывают CO_2 в 5 — 10 раз больше, чем промышленность, и еще более мощным источником CO_2 в атмосфере являются вулканы.

Если модель плохая, то высокая точность численного расчета не улучшит ее, поэтому обычно придерживаются следующего правила: нецелесообразно добиваться большей точности расчетов, чем 10 % от погрешности модели.

1.3.3. Неустранимая погрешность

Часть погрешности решения, обусловленная ошибками исходных данных δw задачи (1.20), равна

$$\delta u = (dA/dw)\delta w; \quad \|\delta u\| \leqslant \left\| \frac{dA}{dw} \right\| \|\delta w\|. \tag{1.24}$$

Она зависит только от исходной задачи (1.20) и ошибок начальных данных, т. е. никакое искусство вычислителя не может ее уменьшить. Поэтому ее называют ***неустранимой***. Она тем больше, чем хуже обусловленность задачи (1.20) и чем больше погрешность начальных данных. Для оценки ее величины надо выяснить величины обоих факторов, т. е. норму погрешности входных данных $\|\delta w\|$ и норму производной $\|dA/dw\|$, являющуюся мерой обусловленности.

Задача (1.20) может быть как чисто математической, так и обработкой экспериментальных данных. Как ни удивительно, качественный характер исходных данных и их погрешности в обоих случаях схож. Математические данные могут быть функцией $w(x)$ или таблицей w_n; последнее обычно используют в случае, когда непосредственное вычисление $w(x)$ настолько трудоемко, что сделано заранее на некоторой сетке. Эксперимент может давать не только дискретную таблицу w_n; существует много экспериментов, регистрирующих непрерывную функцию $w(x)$ (например, осциллограммы).

Погрешность в каждой точке состоит из ***систематической*** и ***случайной*** ошибок. Для экспериментальных данных это общеизвестно. Например, при измерении напряжения вольтметром нуль шкалы может быть сбит, и тогда все полученные данные будут отличаться от истинных на одно и то же значение — это систематическая погрешность. Устранить эту погрешность можно после проверки прибора. При проведении нескольких измерений наш взгляд на стрелку вольтметра всякий раз падает под разными углами, что вносит случайную погрешность. С этим видом погрешности можно бороться, измеряя независимо одну и ту же величину несколько раз и усредняя результат.

Если имеется J измерений x_j, $1 \leqslant j \leqslant J$, величины $\bar{x}$, то более близким к математическому ожиданию будет среднее арифметическое:

$$Mx \approx \bar{x} = \frac{1}{J} \sum_{j=1}^{J} x_j. \tag{1.25}$$

Средняя величина погрешности единичного измерения равна корню квадратному из дисперсии (ее также называют стандартом или стандартным уклонением):

$$\delta x \equiv \sqrt{Dx} = \sqrt{\frac{1}{(J-1)} \sum_{j=1}^{J} (x_j - \bar{x})^2}. \qquad (1.26)$$

Среднее уклонение $\bar{x}$ от Mx будет в J раз меньше. Здесь сделано естественное предположение, что случайная величина распределена по закону Гаусса.

Приведенная погрешность. Пусть входные данные не скалярные величины, а некоторая зависимость $x(t)$, заданная парами значений (t_j, x_j). Считаем, что обе величины x, t измерены независимо со случайными погрешностями. Такие данные принято изображать на графике точками с «крестом» погрешностей (рис. 1.2). На самом деле истинное значение лежит внутри эллипса, главные оси которого равны ошибкам $\delta t, \delta x$ (если величины не коррелированы). В противном случае эллипс будет наклонен в ту или другую сторону в зависимости от знака корреляции.

Полная приведенная ошибка $\tilde{\delta} x$ величины $x(t)$ для некоррелированных погрешностей будет следующей:

$$\tilde{\delta} x \approx \sqrt{\left(\frac{dx}{dt}\delta t\right)^2 + (\delta x)^2}. \qquad (1.27)$$

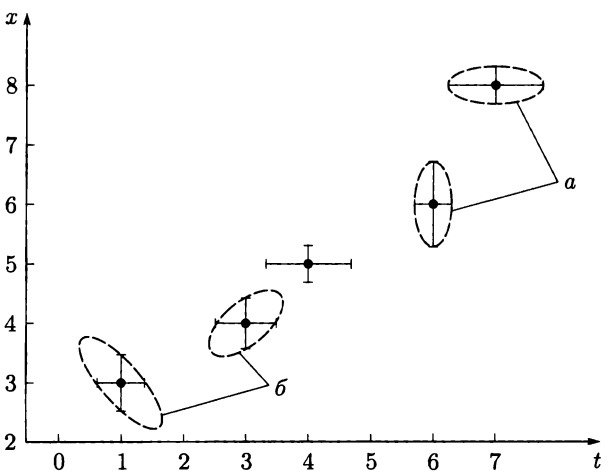


Рис. 1.2. Погрешность двух переменных, эллипсы ошибок:

а — для независимых измерений; *б* — для коррелированных

Это справедливо и для математических задач, если w само является результатом трудоемкого вспомогательного математического расчета: тогда роль систематической ошибки играет погрешность вспомогательного метода, а роль случайной ошибки — ошибка округления компьютерных вычислений (см. подразд. 1.2.3 и 1.2.4).

Относительная погрешность. Для коррелированных измерений вместо (1.27) необходимо записывать более сложное выражение, учитывающее величину и знак коэффициента корреляции.

Как для экспериментальных измерений, так и для численных расчетов величина dx/dt в (1.27) может оказаться заранее неизвестной. В этом случае через точки t_j, x_j проводят аппроксимирующую кривую методами, описанными в гл. 5. Дифференцируя полученную приближенную зависимость $x(t)$, ее подставляют в (1.27).

Величину погрешности удобно характеризовать отношением $\|\delta w\|/\|w\|$. Для простейших функций, математически вычисляемых на (80 — 64)-разрядных компьютерах, она может составлять $10^{-20} - 10^{-16}$. При сложных математических вычислениях относительная погрешность возрастает до $10^{-8} - 10^{-3}$. Для экспериментально измеряемых величин даже в астрономии и геодезии точность лучше 10^{-6} редко достигается; в технике она обычно составляет $10^{-4} - 10^{-2}$, а в передовых областях науки (физика плотной плазмы, химическая кинетика и т. п.) может становиться хуже 0,1.

Для оценки неустранимой погрешности (1.24) нужно знать $\|dA/dw\|$, что не часто удается определить теоретически.

На практике для определения $\|dA/dw\|$ можно применять следующий способ. Решим задачу (1.20) несколько раз, искусственно прибавляя к w различные вариации δw_j. В результате получим соответствующие им вариации δu_j. Составим отношения $c_j = \|\delta u_j\|/\|\delta w_j\|$; если они близки по порядку величины, то их среднее значение можно принять за $\|dA/dw\|$.

Описанный способ трудоемок и нестрог, ибо в нем оцениваются одновременно устойчивости задачи (1.20) и алгоритма ее решения. Поэтому его используют редко. Однако это пока единственный реальный способ оценки неустранимой погрешности, поэтому его следует рекомендовать для повседневного применения. К сожалению, на практике чаще всего ограничиваются интуитивной оценкой обусловленности, что ненадежно.

1.3.4. Погрешность метода

Многие алгоритмы строят так, чтобы у них были управляющие параметры. Например, у итерационного алгоритма это число итераций q, у разностного — шаг сетки h. Алгоритм строят так, чтобы при стремлении параметра к некоторому пределу ($q \to \infty$, $h \to 0$ и т. п.) численное решение стремилось бы к точному. Отличие численного решения от точного при конкретном значении параметра называют погрешностью метода.

Сам факт сходимости и скорость сходимости устанавливают теоретическими исследованиями для каждого метода отдельно. Есть и некоторые полуэмпирические способы исследования сходимости и ее скорости (например, по расчетам на сгущающихся сетках). Все это позволяет оценивать погрешность метода в конкретных расчетах и выбирать параметры метода для обеспечения заданной точности.

Параметры целесообразно выбирать так, чтобы погрешность метода была меньше неустранимой погрешности примерно в 10 раз. Заметно бóльшая погрешность снижает общую точность; заметно меньшая — увеличивает трудоемкость расчетов.

Существуют методы, дающие точный ответ за конечное число действий. Например, это явное решение уравнения в элементарных функциях, или решение систем линейных уравнений методом Гаусса. В них отсутствует погрешность метода. Такие методы называют ***прямыми***.

1.3.5. Погрешность округления

Все числа записываются и операции над ними производятся с конечным числом знаков, т. е. с ошибками. Число x с ошибкой записывают как

$$x \pm \Delta \quad \text{или} \quad x(1 \pm \delta),$$

где Δ — абсолютная ошибка; δ — относительная ошибка. Разумеется, это не точные значения ошибок; ошибки являются случайными величинами (их распределения можно считать гауссовыми), а Δ, δ — их стандартными уклонениями (стандартами). При сложении или вычитании чисел складываются квадраты стандартов их абсолютных ошибок:

$$(x_1 \pm \Delta_1) \pm (x_2 \pm \Delta_2) \to (x_1 \pm x_2) \pm \sqrt{\Delta_1^2 + \Delta_2^2}, \qquad (1.28)$$

а при умножении или делении то же касается относительных ошибок:

$$[x_1(1 \pm \delta_1)][x_2(1 \pm \delta_2)]^{\pm 1} \to (x_1 x_2^{\pm 1}) \left(1 \pm \sqrt{\delta_1^2 + \delta_2^2}\right). \quad (1.29)$$

Ошибка округления при компьютерной записи числа составляет половину последнего разряда мантиссы. Для компьютеров с 32-, 64- и 80-разрядными числами это составляет $\delta_0 \approx 10^{-8}$, 10^{-16} и 10^{-20} соответственно. Процессор Pentium производит вычисления с 80-разрядными числами, однако потребителю выдается результат в виде 64-разрядного числа. Получить высокие разряды возможно при использовании языка C++.

Показателен следующий тест математического обеспечения: выберем $a \approx 1$, $\varepsilon = 0,1; 0,01; \ldots, 10^{-16}$ и проведем, например, следующее вычисление: $\dfrac{\pi(a+\varepsilon) - \pi a}{\pi\varepsilon} \neq 1$. Мера отклонения конечного результата от 1 указывает точность, с которой проводятся арифметические операции. Нетрудно составить подобные тесты на вычислениях показательной, тригонометрической и других элементарных функций. Обычно при использовании четырех арифметических действий ошибка в таком тесте $\sim 10^{-16}$, при вычислении экспоненты даже лучше $\sim 10^{-18}$, а вот вычисление корня квадратного немного менее точная операция — ошибка $\sim 10^{-15}$.

При выполнении N последовательных умножений и делений относительная ошибка, согласно (1.29), увеличивается в $\sqrt{N}$ раз. Даже при огромном количестве действий, которые выполняют современные компьютеры, $\delta_0\sqrt{N}$ остается небольшой величиной. Казалось бы, ошибками округления в компьютерах можно пренебречь и учитывать их лишь при «ручных» расчетах с малым числом знаков.

Однако при сложениях и вычитаниях (1.28) величина $x_1 \pm x_2$ может стать очень малой и сопоставимой с ошибкой. Это наверняка случится при плохой обусловленности задачи (1.20), но может произойти и при хорошей обусловленности задачи, но неудачном построении алгоритма. Такие примеры будут приведены далее. Поэтому даже при вычислениях на многоразрядных компьютерах нельзя забывать об ошибках округления.

Основная рекомендация такова: суммарные ошибки округления должны быть менее погрешности метода, по крайней мере примерно в 10 раз. Для этого следует выбирать многоразрядный компьютер, проводить расчеты с двойной точностью и обращать внимание на тонкости математического обеспечения: не любое математическое обеспечение позволяет полностью использовать все разряды, имеющиеся в процессоре.

1.3.6. Корректность задачи

Понятие корректности математической задачи ввел Курант (назвав это корректностью по Адамару).

Определение 1.1. Задача $u = A(x)$ называется ***корректной***, если:

1) решение задачи существует при любом x, принадлежащем некоторому заданному множеству X;

2) это решение единственно для любого x;

3) решение непрерывно зависит от x (напомним нестрогое определение непрерывности: если вариация исходных данных $\|\delta x\| \to 0$, то вариация решения $\|\delta u\| \to 0$).

Пример 1.6. Возьмем однородное уравнение теплопроводности на конечном отрезке при нулевых граничных условиях:

$$\frac{\partial u}{\partial t} = \varkappa \frac{\partial^2 u}{\partial x^2}, \quad 0 \leqslant x \leqslant a; \quad u(0,t) = u(a,t) = 0. \tag{1.30}$$

Решение уравнения (1.30) легко строится методом разделения переменных:

$$\sum_{n=1}^{\infty} \alpha_n \exp(-\lambda_n t) \sin(\pi n x/a); \quad \lambda_n = \varkappa \pi^2 n^2/a^2, \tag{1.31}$$

где α_n — коэффициенты Фурье для профиля начальных данных $u(x,0)$. Видно, что при $t > 0$ все гармоники в (1.31) затухают, причем тем быстрее, чем больше номер гармоники. При этом любое возмущение начальных данных $\delta u(x,0)$ затухает со временем и задача (1.30) оказывается корректной.

Если же мы хотим восстановить распределение температуры до момента времени $t = 0$, т. е. проследить решение той же задачи при $t < 0$, то картина противоположная. Чем выше номер гармоники, тем больше временной множитель перед ней, причем при $n \to \infty$ эти множители неограниченно растут. Поэтому ничтожное возмущение старших коэффициентов Фурье начальных данных приводит к огромным возмущениям решения, т. е. задача (1.30) оказывается некорректной. По тем же причинам некорректными оказываются важные задачи геофизической разведки (сейсмическое и электромагнитное зондирование слоев земной коры).

Очевидно, прямое численное решение некорректных задач невозможно: даже если точно задать начальные данные, то неиз-

бежно присутствующие ошибки округления приведут к сильному искажению решения. Поэтому для численного решения некорректных задач разрабатывают специальные методы, называемые методами регуляризации. В них исходная задача видоизменяется введением дополнительных ограничений, превращающих задачу в корректную. Обоснованно выбрать такие ограничения непросто.

Даже формально корректная задача может оказаться очень трудной для численного решения. Например, пусть для точной задачи $\|\delta u\| = C\|\delta x\|$, где константа $C \gg 1$. Формально $\|\delta u\| \to 0$ при $\|\delta x\| \to 0$, т. е. задача корректна. Однако если $C = 10^{100}$, то даже ошибки компьютерного округления приводят к огромным ошибкам решения. Такие задачи называют ***плохо обусловленными***. На практике они мало отличаются от некорректных. Для их решения применяют способы, похожие на регуляризацию.

Глава 2

СИСТЕМЫ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

2.1. ЛИНЕЙНЫЕ СИСТЕМЫ

2.1.1. Задачи линейной алгебры

Необходимость решения систем алгебраических линейных уравнений большой размерности возникла в конце XVIII в. Тогда во Франции было предпринято прецизионное измерение длины дуги парижского меридиана. С севера на юг были построены триангуляционные вышки, образующие на земной сфере сеть треугольников. Углы этих треугольников были тщательно измерены. Для северного и южного триангуляционных пунктов астрономы измерили географические широты. Длины сторон некоторых треугольников измерялись мерными лентами. После этого определение длины дуги свелось к решению огромной системы линейных уравнений. Эту задачу решил Д. Ф. Араго, избранный за это в Парижскую академию наук.

Перечислим основные задачи линейной алгебры (пусть имеется квадратная матрица A порядка N).

Первая задача — решение системы линейных уравнений

$$A\mathbf{x} = \mathbf{b}, \tag{2.1}$$

где $\mathbf{x}, \mathbf{b}$ — векторы размерности N.

Систему (2.1) можно записать в покомпонетной форме:

$$\sum_{m=1}^{N} a_{nm}x_m = b_n, \quad 1 \leqslant n \leqslant N. \tag{2.2}$$

Напомним, что если $\det A \neq 0$, то система (2.2) имеет единственное решение. При $\det A = 0$ задача является некорректной (см. гл. 1). Она может быть также плохо обусловленной. У всех плохо обусловленных задач $\det A \approx 0$, однако и у многих хоро-

шо обусловленных задач $\det A \approx 0$. Вопрос о критерии плохой обусловленности будет обсужден далее.

Вторая задача — вычисление $\det A$.

Третья задача — нахождение обратной матрицы A^{-1}. Решение второй и третьей задач сводится к первой и будет изложено в этой главе.

Четвертая задача — нахождение спектра матрицы A — намного сложнее и ей посвящена отдельная глава (см. гл. 7).

Матрица A может быть плотно заполненной. Но в практике часто возникают специальные виды матриц, содержащие плотные массивы нулевых элементов (рис. 2.1).

Матрицу называют верхней треугольной, если все элементы ниже главной диагонали — нулевые: $a_{nm} = 0$, при $n > m$ (рис. 2.1, *а*). Аналогично определяют нижнюю треугольную матрицу. Верхней почти треугольной называют матрицу, у которой ниже главной диагонали отличны от нуля только несколько побочных диагоналей (их также называют кодиагоналями); если отлична от нуля только одна побочная диагональ, то матрицу называют матрицей Хессенберга (рис. 2.1, *б*). Аналогично опре-

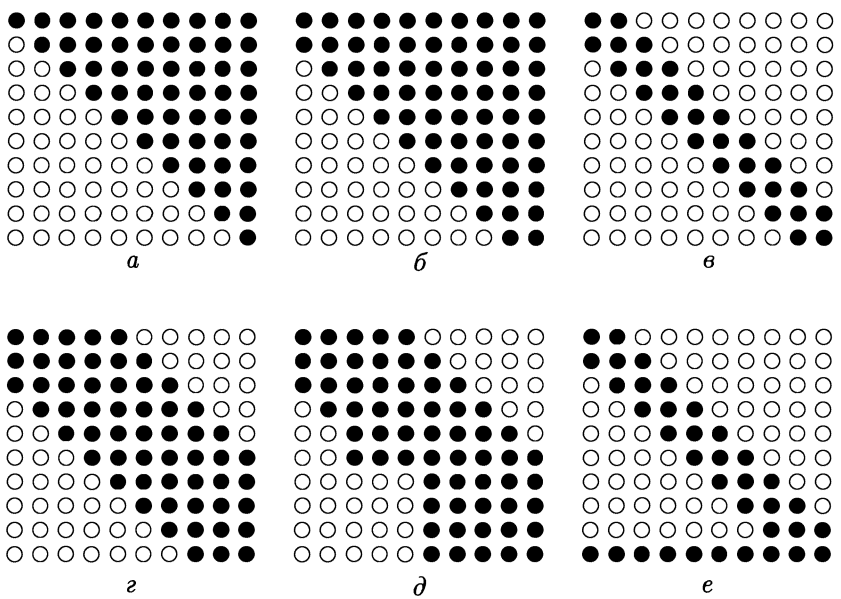


Рис. 2.1. Некоторые специальные виды матриц:

а — треугольная; *б* — почти треугольная; *в* — трехдиагональная; *г* — ленточная; *д* — ящичная; *е* — с массивом нулей (черные кружки — ненулевые элементы, светлые кружки — нулевые элементы)

деляют нижнюю почти треугольную матрицу. При решении краевых задач для дифференциальных уравнений второго порядка обычно появляются трехдиагональные матрицы: $a_{nm} \neq 0$ при $|n-m| \leqslant 1$ (рис. 2.1, *в*). Для дифференциальных уравнений четвертого порядка возникают пятидиагональные матрицы. Более общим является случай ленточной матрицы, вообще говоря, с несимметричной лентой (рис. 2.1, *г*).

Могут появляться ящичные матрицы, где ненулевые элементы имеются только в квадратных подматрицах, лежащих на главной диагонали (рис. 2.1, *д*). Ящичные матрицы можно рассматривать как частный случай ленточной матрицы, когда внутри ленты есть и заведомо нулевые элементы. Подобных специальных структур существует много. Примером является рис. 2.1, *е*, где помимо ленты имеется еще горизонтальная полоса ненулевых элементов. Прямые методы решения для плотно заполненных матриц, допускающие экономные модификации в случае матриц специального вида, будут описаны в этой главе.

При решении многомерных уравнений в частных производных и для некоторых других задач возникают сильно разреженные матрицы. Они имеют огромный порядок $N \sim 10^4 - 10^9$, но в каждой строке лишь несколько элементов отличны от нуля. Такие задачи решают итерационными методами, которые будут рассмотрены в других главах (см. 8.3).

2.1.2. Метод Гаусса

В 1810 г. К. Ф. Гаусс предложил метод решения систем линейных алгебраических уравнений, который экономичен и конкурентоспособен по сей день. Этот метод прямой, т. е. дает точное решение задачи за конечное заранее определенное число операций. У прямого метода отсутствует погрешность и существуют лишь ошибки округления.

Решение системы с треугольной матрицей находится исключительно просто. Главная идея метода Гаусса состоит в приведении матрицы системы к верхней треугольной форме. Это называется прямым ходом метода Гаусса. Решение полученной системы с верхней треугольной матрицей называют обратным ходом.

Умножение уравнений системы на то или иное число и вычитание одного уравнения из другого не меняет систему. Будем подбирать множители так, чтобы обратить в нуль в матрице элементы ниже главной диагонали. При этом построим метод так, чтобы он использовал информацию об уже имеющихся нулевых

элементах и не производил лишних операций с нулями. Это особенно просто делается для ленточной матрицы с несимметричной лентой (рис. 2.1, *г*):

$$a_{nm} \neq 0 \quad \text{для} \quad n - p \leqslant m \leqslant n + q, \tag{2.3}$$

где p, q — количество нижних и верхних кодиагоналей соответственно.

Очевидно, $p, q \leqslant N - 1$. Но (если $p = q = N - 1$, то матрица становится полностью заполненной) учет ленточной структуры дает существенную экономию в том случае, если $p \ll N$ и (или) $q \ll N$.

Прямой ход равносилен выполнению трех вложенных циклов. Самый внешний цикл — перебор всех столбцов матрицы для обращения в нуль элементов ниже главной диагонали.

Допустим, мы уже обратили в нуль элементы, лежащие ниже главной диагонали в первых $k - 1$ столбцах (рис. 2.2). Элементы матрицы системы к этому времени изменились, поэтому введем верхний индекс, указывающий номер цикла k. Этот индекс не относится к уже преобразованным столбцам ($m \leqslant k - 1$) и строкам ($n \leqslant k - 1$) матрицы и присваивается только оставшейся подматрице. С учетом ленточной структуры оставшиеся уравнения принимают вид:

$$\sum_{m=\max(k,n-p)}^{\min(n+q,N)} a_{nm}^{(k)} x_m = b_n^{(k)}, \quad k \leqslant n \leqslant N. \tag{2.4}$$

Для первого шага этого цикла нужно положить $a_{nm}^{(1)} = a_{nm}$, $b_n^{(1)} = b_n$.

Средний цикл — перебор элементов k-го столбца в строках $k + 1 \leqslant n \leqslant N$, которые мы хотим обратить в нуль. В случае ленточной структуры нижняя часть уже будет содержать заведомо нулевые элементы, так что для экономичной работы программы следует ограничиться значениями $k + 1 \leqslant n \leqslant \min(k + p, N)$.

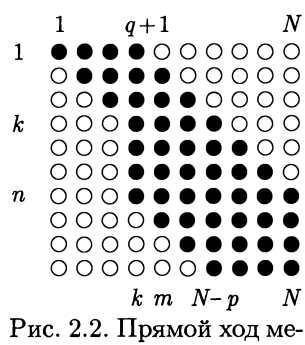


Рис. 2.2. Прямой ход метода Гаусса

Чтобы обратить в нуль элемент $a_{nk}^{(k)}$, умножим элементы k-й строки на величину

$$c_{nk} = a_{nk}^{(k)} / a_{kk}^{(k)}, \quad k + 1 \leqslant n \leqslant \min(k + p, N). \tag{2.5}$$

Прежде чем выполнять деление (2.5), следует провести проверку. Если $a_{nk}^{(k)} = 0$, то (2.5) выполнять не нужно, а n-я строка матрицы остается без изменений. Для многих типов матриц это дает существенную экономию.

Далее выполним внутренний цикл. Вычтем из n-й строки k-ю, умноженную на c_{nk}:

$$\begin{cases} b_n^{(k+1)} = b_n^{(k)} - c_{nk} b_k^{(k)}, \\ a_{nm}^{(k+1)} = a_{nm}^{(k)} - c_{nk} a_{km}^{(k)}, \quad k+1 \leqslant m \leqslant \min(n+q, N). \end{cases} \tag{2.6}$$

При программировании нужно придерживаться приведенных выше границ изменения индексов. Кроме того, величину $a_{nk}^{k+1} = 0$ не следует вычислять по (2.6), чтобы не накапливать ошибки округления; следует принудительно положить ее равной нулю.

Полезно запоминать коэффициенты c_{nk}, записывая их на место обращенных в нуль элементов a_{nk}. Эти коэффициенты сильно повышают экономичность, если приходится неоднократно решать систему с одной и той же матрицей, но разными правыми частями.

Прямой ход метода Гаусса требует $\sim 2/3N^3$ арифметических действий для плотно заполненной матрицы. Для почти треугольной матрицы число операций существенно меньше; например, для матрицы Хессенберга оно составляет $\sim 2N^2$. Для ленточной матрицы оно еще меньше $\sim (p+q)^2 N$.

Для плотно заполненной матрицы с $N \sim 1000$ (что нередко на практике) общее число действий прямого хода $\sim 10^9$. При этом ошибки округления, согласно статистике, достигают $\sim 10^5$. Таким образом, даже для хорошо обусловленных матриц теряется пять последних знаков! Поэтому вычисления нужно вести только с двойной точностью, используя максимальное число разрядов, допустимое в данном программном обеспечении.

Выбор главного элемента. Прямой ход метода Гаусса содержит только одно деление (2.5). Если $a_{kk}^{(k)} = 0$, то деление невыполнимо. Перебором элементов в k-м столбце подматрицы всегда можно найти элемент, отличный от нуля (иначе $\det A = 0$ и система неразрешима). Даже малые $a_{kk}^{(k)}$ опасны: из-за ошибок округления важно избегать деления на малые числа. Выберем максимальный по модулю элемент в k-м столбце подматрицы (он называется *главным*):

$$\max_n |a_{nk}^{(k)}|, \quad k \leqslant n \leqslant \min(k+p, N). \tag{2.7}$$

Переставим строки так, чтобы главный элемент оказался на главной диагонали. Такая операция застрахует нас от переполнения при делении на нуль и повысит устойчивость к ошибкам округления.

Если выбирать максимальный по модулю элемент не только в столбце, но и во всей подматрице, то устойчивость к ошибкам округления еще более возрастет. Но экономически это уже невыгодно, поскольку требуется переставлять не только строки, но и столбцы. Поэтому такой вариант не применяют.

Однако если матрица ленточная ($q < N-1$), то перестановка строк может нарушить структуру верхнего массива нулей и в ходе вычислений матрица перестанет быть ленточной. Поэтому для ленточных матриц выбор главного элемента не применяют.

Заметим, что в практике вычислений нередки симметричные ($a_{nm} = a_{mn}$) положительно или отрицательно определенные матрицы. Например, они возникают при решении задач аппроксимации методом наименьших квадратов. В линейной алгебре строго доказывается, что у таких матриц главный элемент всегда будет стоять на главной диагонали. Очевидно, для них также не нужно делать выбор главного элемента.

Обратный ход. После прямого хода результирующая система имеет треугольный вид

$$\sum_{m=n}^{\min(n+q,N)} a_{nm}^{(n)} x_m = b_n^{(n)}, \quad 1 \leqslant n \leqslant N. \tag{2.8}$$

Ее решение находится выполнением обратного цикла

$$x_n = \left(b_n^{(n)} - \sum_{m=n+1}^{\min(n+q,N)} a_{nm}^{(n)} x_m \right) \Big/ a_{nn}^{(n)}, \quad n = N, N-1, \ldots, 1. \tag{2.9}$$

В (2.9) при $n = N$ сумма отсутствует. Если проводился выбор главного элемента, то после каждого шага обратного хода нужно проводить обратную перестановку строк.

Для выполнения обратного хода требуется $\sim N^2$ действий для плотно заполненной матрицы и порядка qN для ленточной. Таким образом, обратный ход гораздо «дешевле» прямого и ошибки округления на нем меньше.

Метод Гаусса даже для плотно заполненной матрицы требует оперативной памяти для хранения $\approx N^2$ чисел. Для современных компьютеров это не вызывает затруднений.

В некоторых приложениях необходимо решать линейные системы с комплексными матрицами. Такие программы также существуют в современных математических обеспечениях.

2.1.3. Определитель и обратная матрица

Прямой ход метода Гаусса содержит вычитание из n-й строки k-й строки, домноженной на коэффициент c_{nk}. Из линейной алгебры известно, что такая операция не меняет величины определителя. Значит, $\det A$ равен определителю полученной треугольной матрицы, а ее определитель равен произведению диагональных элементов. Если на каждом шаге внешнего цикла прямого хода проводился выбор главного элемента и перестановка n-й и k-й строк, то определитель домножался на $(-1)^{n-k}$. Поэтому

$$\det A = \pm \prod_{n=1}^{N} a_{nn}^{(n)}.$$

Здесь знак выбирается в зависимости от четности суммы всех перестановок строк.

Таким образом, определитель матрицы вычисляется в качестве «бесплатного приложения» метода Гаусса. Однако это «бесплатное приложение» иногда дорого обходится. При большом порядке матрицы произведение даже небольших по модулю диагональных элементов оказывается огромным числом и приводит к переполнению. Известны случаи отказа стандартных подпрограмм решения систем линейных алгебраических уравнений именно из-за переполнения при вычислении определителя. Значит, несмотря на малые затраты вычисления определителя, не стоит включать эту операцию в программу без лишней необходимости.

По определению обратной матрицы $AA^{-1} = E$, где E — единичная матрица. Если обозначить элементы обратной матрицы через α_{ml}, то по правилу умножения матриц они удовлетворяют следующей системе уравнений:

$$\sum_{m=1}^{N} a_{nm}\alpha_{ml} = \delta_{nl} = \begin{cases} 1, & n = l, \\ 0, & n \neq l, \end{cases} \qquad 1 \leqslant n \leqslant N, \quad 1 \leqslant l \leqslant N. \quad (2.10)$$

Уравнение (2.10) следует трактовать следующим образом. Фиксируем l, т. е. рассмотрим l-й столбец матрицы A^{-1}: $\alpha_l = \{\alpha_{ml}\}$, $1 \leqslant l \leqslant N$. Тогда этот вектор удовлетворяет линейной системе

$$A\boldsymbol{\alpha}_l = \boldsymbol{\delta}_l. \tag{2.11}$$

Здесь $\delta_l = \{\delta_{ml}\}$, $1 \leqslant m \leqslant N$, — аналогичный вектор, l-я компонента которого равна 1, а остальные элементы — нули.

Для поиска обратной матрицы нужно N раз решать систему (2.11) с одной и той же матрицей A, но разными правыми частями $\boldsymbol{\delta}$. Прямой ход метода Гаусса нужно сделать всего один раз. Для преобразования правых частей пригодятся сохраненные c_{nk}. Трудоемкость поиска обратной матрицы при таком подходе $\sim 2N^3$ действий (для плотно заполненной матрицы), т. е. всего в три раза больше, чем трудоемкость прямого хода метода Гаусса. Требуемый объем памяти составляет $\approx 2N^2$ чисел (для хранения прямой и обратной матриц).

Аналогично действуют, если нужно решить несколько линейных систем с одной и той же матрицей, но разными правыми частями: прямой ход метода Гаусса делают только один раз. Это сильно сокращает объем вычислений. Заметим, что формально решение линейной системы (2.1) можно записать в виде $\mathbf{x} = A^{-1}\mathbf{b}$. Иногда неопытные вычислители пользуются этой формулой: находят по стандартной программе обратную матрицу A^{-1} и умножают ее на вектор $\mathbf{b}$. Это невыгодно, так как трудоемкость возрастает втрое по сравнению с прямым решением линейной системы.

2.1.4. Прочие методы

Есть очень много других прямых методов решения систем с плотно заполненной матрицей: метод оптимального исключения, метод окаймления, метод отражений, метод ортогонализации, метод Жордана и др. Все они имеют те же скорость и устойчивость к ошибкам округления, что и метод Гаусса. По объему памяти только метод Жордана при обращении матрицы дает небольшое преимущество: нужно хранить $\approx N^2$ чисел, так как обратная матрица ставится на место прямой. Но для современных компьютеров такой выигрыш несущественен.

Для эрмитовых матриц ($a_{nm} = a^*_{mn}$, где звездочка означает комплексно сопряженное число) имеется метод корня квадратного. Он явно использует симметрию матрицы и за счет этого требует вдвое меньше операций и памяти, чем метод Гаусса. Его устойчивость к ошибкам округления не лучше, чем у метода Гаусса.

В практике часто встречаются системы с трехдиагональной матрицей, которые принято записывать в специальной форме:

$$\begin{gathered} a_n x_{n-1} - b_n x_n + c_n x_{n+1} = d_n, \quad 1 \leqslant n \leqslant N; \\ a_1 = c_N = 0. \end{gathered} \tag{2.12}$$

Для их решения традиционно применяют метод прогонки. Это тоже исключение нижней кодиагонали. Формулы этого исключения немного отличаются от метода Гаусса. Прямой ход имеет вид

$$\begin{gathered} \xi_{n+1} = c_n/(b_n - a_n\xi_n), \\ \eta_{n+1} = (a_n\eta_n - d_n)/(b_n - a_n\xi_n), \quad 1 \leqslant n \leqslant N. \end{gathered} \tag{2.13}$$

Решение системы находится обратным ходом:

$$x_n = \xi_{n+1}x_{n+1} + \eta_{n+1}, \quad n = N, N-1, \ldots, 1, \quad x_{N+1} = 0. \tag{2.14}$$

Для возникающих в практике трехдиагональных матриц часто выполняется условие преобладания главной диагонали

$$|b_n| \geqslant |a_n| + |c_n|, \quad 1 \leqslant n \leqslant N, \tag{2.15}$$

причем хотя бы при одном n неравенство является строгим. Это условие обеспечивает существование и единственность решения (2.12), а также ненарастание ошибок округления в (2.13) и (2.14).

Однако прогонка имеет ту же трудоемкость и устойчивость к ошибкам округления, что и метод Гаусса для ленточных матриц, описанный в подразд. 2.1.2.

2.1.5. Плохо обусловленные системы

Примером плохо обусловленной матрицы может служить матрица Гильберта

$$H = \begin{pmatrix} 1 & 1/2 & 1/3 & 1/4 & 1/5 & \ldots \\ 1/2 & 1/3 & 1/4 & 1/5 & 1/6 & \ldots \\ 1/3 & 1/4 & 1/5 & 1/6 & 1/7 & \ldots \\ 1/4 & 1/5 & 1/6 & 1/7 & 1/8 & \ldots \\ 1/5 & 1/6 & 1/7 & 1/8 & 1/9 & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \end{pmatrix},$$

возникающая в задаче среднеквадратичной аппроксимации многочленами. Показателен следующий тест. Возьмем столбец из единиц $\mathbf{e} = (1, 1, \ldots)^T$. Прямым умножением на матрицу Гильберта вычислим $H\mathbf{e} = \mathbf{b}$; погрешность округления при этой процедуре невелика. Затем численно решим систему $H\mathbf{x} = \mathbf{b}$ методом Гаусса или любым другим. Сравним найденное решение $\mathbf{x}$ с $\mathbf{e}$.

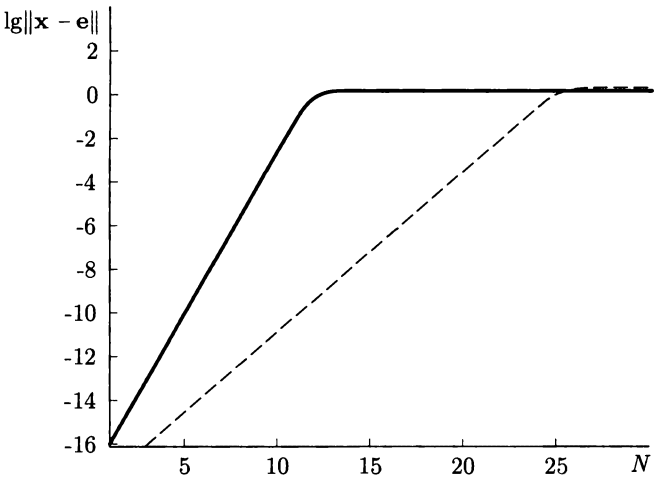


Рис. 2.3. Погрешность решения линейных систем порядка N с матрицей Гильберта (сплошная линия) и матрицей симметризованных степеней (штриховая линия)

На рис. 2.3 изображена зависимость погрешности $\|\mathbf{x} - \mathbf{e}\|$ от порядка N матрицы H (штриховая линия будет объяснена в гл. 5). Расчеты проводились с 64 разрядами. Для $N = 1$ ошибка не превышает 10^{-16} (что соответствует ошибке единичного округления). При увеличении N на 1 ошибка возрастает почти в 30 раз и при $N = 12$ в решении не остается ни одного верного знака! При дальнейшем увеличении N ошибка в хорошо написанной программе уже не нарастает.

Надежного количественного критерия плохой обусловленности до сих пор не выработано. Для компьютеров первого поколения, работающих с фиксированной точкой, за число обусловленности обычно принимали $\mathrm{Cond} = \|A\|\|A^{-1}\|$. И значения $\mathrm{Cond} > 10^4$ считали признаком плохой обусловленности.

Однако для вычислений с плавающей точкой этот критерий неприменим. Для диагональной матрицы наименьшая из норм — спектральная $\|D\| \max\limits_{1 \leqslant n \leqslant N} |d_{nn}|$. Но для обратной матрицы диагональные элементы будут равны d_{nn}^{-1}, так что $\|D^{-1}\| = \max\limits_{1 \leqslant n \leqslant N} |d_{nn}^{-1}| = 1/\min\limits_{1 \leqslant n \leqslant N} |d_{nn}|$. Таким образом, $\mathrm{Cond} = \max\limits_{1 \leqslant n \leqslant N} |d_{nn}| / \min\limits_{1 \leqslant n \leqslant N} |d_{nn}|$ может быть сделано очень большим при соответствующем подборе элементов. Однако для диагональной матрицы каждая компонента решения системы находится с помощью только одного деления. А эта операция при

плавающей точке выполняется с ошибкой $\sim 10^{-16}$ (для 64-разрядной техники).

Для плавающей точки сейчас наилучшим считается число обусловленности Дж. Ортеги

$$\mathrm{Cond} = \det A / \prod_{n=1}^{N} \left(\sum_{m=1}^{N} a_{nm}^2 \right)^{1/2}. \tag{2.16}$$

Его нестрого трактуют как отношение объема параллелепипеда, построенного на строках матрицы как на ребрах, к объему прямоугольного параллелепипеда с теми же ребрами. Для диагональной матрицы в (2.16) $\mathrm{Cond} = 1$. А большие значения Cond в (2.16) действительно соответствуют плохо обусловленным матрицам. Однако установить зависимость между величиной Cond (2.16) и количеством потерянных при решении системы знаков не удалось. Тестирование показало, что для одних матриц все 16 значащих цифр решения терялись при $\mathrm{Cond} = 10^{25}$, для других — при $\mathrm{Cond} = 10^{32}$.

Регуляризация. Повысить устойчивость решения плохо обусловленных систем можно методом регуляризации. Особенно просто регуляризация записывается, если матрица A эрмитова и положительно (или отрицательно) определенная. Второй случай сводится к первому заменой A на $-A$, поэтому будем считать, что $A = A^H > 0$. Тогда задача решения линейной системы

$$A\mathbf{x} = \mathbf{b} \tag{2.17}$$

эквивалентна нахождению минимума функционала:

$$(\mathbf{x}, A\mathbf{x}) - 2(\mathbf{x}, \mathbf{b}) = \min. \tag{2.18}$$

Пусть требуется найти решение, наиболее близкое к заданному $\mathbf{x}_0$, т. е. потребовать

$$(\mathbf{x} - \mathbf{x}_0, \mathbf{x} - \mathbf{x}_0) = \min. \tag{2.19}$$

Такое решение называют *нормальным*. Задачи (2.18) и (2.19) несовместны. Поэтому возьмем вместо них новую задачу:

$$(\mathbf{x}, A\mathbf{x}) - 2(\mathbf{x}, \mathbf{b}) + \alpha(\mathbf{x} - \mathbf{x}_0, \mathbf{x} - \mathbf{x}_0) = \min, \tag{2.20}$$

выбрав положительный достаточно малый параметр $0 < \alpha \ll 1$. Варьирование задачи приводит к уравнению

$$2(\delta\mathbf{x}, A\mathbf{x}) - 2(\delta\mathbf{x}, \mathbf{b}) + 2\alpha(\delta\mathbf{x}, \mathbf{x} - \mathbf{x}_0) = 0$$

или окончательно к линейной системе

$$(A + \alpha E)\mathbf{x} = \mathbf{b} + \alpha \mathbf{x}_0.$$

Результирующая система обусловлена гораздо лучше исходной благодаря добавлению члена αE. Такой прием — частный случай *регуляризации по Тихонову.*

Выбор $\mathbf{x}_0$ делают, опираясь на априорные сведения о решении системы или результаты предыдущих расчетов (например, данные за прошлый год в задачах планирования). Параметр α подбирают начиная с очень малых значений и постепенно увеличивают до тех пор, пока обусловленность матрицы $A + \alpha E$ не станет приемлемой. Чем больше α, тем дальше отстоит решение регуляризованной системы от истинного решения исходной задачи, но тем устойчивее оно вычисляется.

Регуляризацию можно провести и для неэрмитовых матриц. Домножим линейную исходную систему слева на матрицу A^H. Получим линейную систему

$$A^H A\mathbf{x} = A^H \mathbf{b}$$

с эрмитовой положительно определенной матрицей $A^H A$. Применив к ней описанный прием, имеем регуляризованную систему

$$(A^H A + \alpha E)\mathbf{x} = A^H \mathbf{b} + \alpha \mathbf{x}_0.$$

Однако матрица $A^H A$ гораздо хуже обусловлена, чем исходная матрица A. Поэтому для получения удовлетворительной регуляризации нужно брать существенно бóльшие значения α.

Пример 2.1. Рассмотрим следующую линейную систему:

$$Ax = b, \quad A = \begin{pmatrix} 1 + \varepsilon + \varepsilon^2/2 & -1 \\ -1 & 1 - \varepsilon + \varepsilon^2/2 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Для нее уже при умеренном $\varepsilon = 10^{-4}$ имеем $\det A < 10^{-16}$ (на уровне ошибок округления).

Решение нетрудно получить явно:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = -\frac{2}{\varepsilon^2} \begin{pmatrix} 2 - \varepsilon \\ 2 + \varepsilon \end{pmatrix}.$$

Видно, что при изменении малого параметра ε в 2 раза матрица системы практически не изменяется, а решение меняется в 8 раз!

Замечание. В учебниках и стандартных программах нередко предлагают следующий способ решения плохо обусловленных систем. Непосредственно решим исходную систему (2.17). Затем вычислим так называемую невязку, которая является вектором:

$$\mathbf{r} = A\mathbf{x} - \mathbf{b}. \tag{2.21}$$

Она будет отлична от нуля из-за ошибок округления. Подставим ее вместо $\mathbf{b}$ в уравнение (2.17) и снова решим систему. Полученную величину возьмем в качестве поправки к $\mathbf{x}$. Если точность окажется недостаточной, повторим этот процесс. Однако практика расчетов показала, что для удовлетворительно обусловленных систем точность решения и без того достаточна, а для очень плохо обусловленных систем улучшения точности фактически не происходит. При этом неясно, есть ли содержательный диапазон обусловленностей и где этот прием эффективен.

2.1.6. Переобусловленные системы

Иногда приходится решать системы, в которых число уравнений больше числа неизвестных (переопределенная система). Такие системы нередко возникают в задачах экономики, где при составлении оптимальных планов требуется удовлетворить огромному числу разных условий. Задача имеет вид

$$B\mathbf{x} = \mathbf{b}, \tag{2.22}$$

где B — прямоугольная матрица размера $N \times M$, у которой строк больше, чем столбцов: $N > M$. Вектор $\mathbf{x}$ имеет размерность M, а вектор $\mathbf{b}$ — размерность N. Система (2.22) в общем случае несовместна и не имеет решения.

Тогда по аналогии с (2.21) вводят невязку $\mathbf{r} = B\mathbf{x} - \mathbf{b}$. В этом случае невязка заведомо не близка к нулю. Потребуем минимума нормы этой невязки:

$$(\mathbf{r}, \mathbf{r}) = (B\mathbf{x} - \mathbf{b}, B\mathbf{x} - \mathbf{b}) = \min. \tag{2.23}$$

Решение задачи называют *квазирешением.*

В качестве дополнительного условия обычно берут близость к некоторому $\mathbf{x}_0$ и переходят к задаче

$$(B\mathbf{x} - \mathbf{b}, B\mathbf{x} - \mathbf{b}) + \alpha(\mathbf{x} - \mathbf{x}_0, \mathbf{x} - \mathbf{x}_0) = \min. \tag{2.24}$$

Такое квазирешение называют *нормальным.*

Варьируя $\mathbf{x}$ в (2.24), получим

$$(B\boldsymbol{\delta}\mathbf{x}, B\mathbf{x} - \mathbf{b}) + \alpha(\boldsymbol{\delta}\mathbf{x}, \mathbf{x} - \mathbf{x}_0) = 0. \tag{2.25}$$

Поскольку $(B\delta\mathbf{x}, B\mathbf{x} - \mathbf{b}) = (\delta\mathbf{x}, B^H B\mathbf{x} - B^H\mathbf{b})$, уравнение (2.25) приводится к виду

$$(B^H B + \alpha E)\mathbf{x} = B^H\mathbf{b} + \alpha\mathbf{x}_0. \tag{2.26}$$

Здесь слева стоит квадратная матрица $B^H B$ порядка M, а справа — вектор $B^H\mathbf{b}$ той же размерности с элементами

$$(B^H B)_{nm} = \sum_{l=1}^{N} b_{ln}^* b_{lm}, \quad (B^H\mathbf{b})_n = \sum_{l=1}^{N} b_{ln}^* b_l, \quad 1 \leqslant n, m \leqslant M.$$

Матрица системы (2.26) квадратная, эрмитова и положительно определенная. Поэтому ее решение существует и единственно.

Пример 2.2. Рассмотрим систему (2.22), где

$$B = \begin{pmatrix} 1 & \varepsilon - 1 \\ -1 & 1 \\ 1 + \varepsilon & -1 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} \varepsilon \\ 0 \\ -\varepsilon \end{pmatrix}.$$

Это система трех уравнений с двумя неизвестными. Она несовместна. Если взять первое и второе уравнения, то $\mathbf{x}_I = (1.1)^T$, а если второе и третье, то $\mathbf{x}_{II} = (-1, -1)^T$. Обе эти подсистемы второго порядка плохо обусловлены, так как их определители равны $\pm\varepsilon$. Квазирешения находим из системы (2.26), где

$$B^H B = \begin{pmatrix} 3 + 2\varepsilon + \varepsilon^2 & -3 \\ -3 & 3 - 2\varepsilon + \varepsilon^2 \end{pmatrix}, \quad B^H\mathbf{b} = \begin{pmatrix} -\varepsilon^2 \\ \varepsilon^2 \end{pmatrix}.$$

Обусловленность матрицы $B^H B$ еще хуже, так как $\det(B^H B) = 2\varepsilon^2 + \varepsilon^4$ еще меньше. Однако квазирешения легко находятся даже при $\alpha = 0$: и

$$\mathbf{x} = \left(\varepsilon\frac{1 - \varepsilon/2}{1 + \varepsilon^2/2}, \varepsilon\frac{1 + \varepsilon/2}{1 + \varepsilon^2/2}\right)^T \approx (\varepsilon, \varepsilon)^T.$$

Оно разумно лежит между x_I и x_{II}.

2.2. НЕЛИНЕЙНОЕ УРАВНЕНИЕ

2.2.1. Дихотомия

Одна из наиболее часто встречающихся в практике элементарных задач — проблема нахождения нулей функции, т. е. решение уравнения

$$f(x) = 0. \tag{2.27}$$

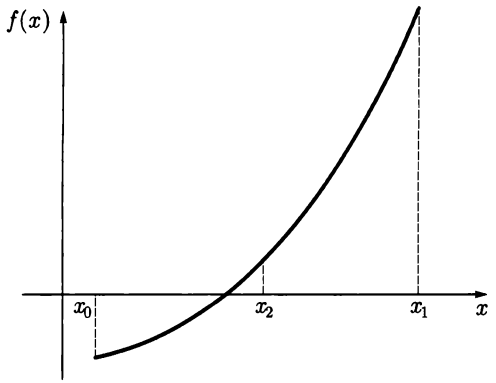


Рис. 2.4. Дихотомия

Оно может быть разрешено в явном виде лишь в редких частных случаях (например, если $f(x)$ является полиномом не выше 4-й степени). Естественное требование к функции $f(x)$ — ее непрерывность.

Качественное поведение функции $f(x)$ возможно установить методами классической математики. Разнообразные программные средства позволяют строить графики функции и тем самым определять примерное расположение и число корней уравнения (2.27). Если приближенное местонахождение корня известно, то его нахождение с требуемой точностью выполняют методами численного анализа.

Простейшим таким методом является дихотомия (ее также называют бисекцией или делением отрезка пополам). Для применения дихотомии достаточно лишь непрерывности функции $f(x)$. Метод основан на теореме классического анализа: если $f(x)$ непрерывна на отрезке $[x_0, x_1]$ и на концах отрезка принимает значения разных знаков, то на этом отрезке имеется по меньшей мере один корень уравнения (2.27) (рис. 2.4).

Пусть известен (например, из графика функции) отрезок $[x_0, x_1]$, на концах которого функция имеет разные знаки: $f(x_0) \times$ $\times f(x_1) < 0$. Формально можно написать $f(x_0)f(x_1) \leqslant 0$, но равенство означает, что один из концов отрезка является корнем (2.27).

Рассмотрим следующий алгоритм. Найдем середину отрезка

$$x_2 = (x_0 + x_1)/2. \tag{2.28}$$

Вычислим $f(x_2)$ и сравним его знак со знаками $f(x_0)$ и $f(x_1)$. Из двух половин отрезка выберем ту, на концах которой функция имеет разные знаки. Один шаг процесса завершен.

Далее повторяем описанную последовательность действий применительно к выбранной половине отрезка.

В этом алгоритме корень уравнения (2.27) всегда лежит внутри очередного выбранного отрезка. Тем самым длина отрезка на очередном шаге служит мерой погрешности.

С каждой новой итерацией точность повышается ровно вдвое. Значит, для уточнения трех верных знаков корня требуется около 10 итераций ($2^{10} = 1024 \approx 10^3$), а для всех 16 десятичных знаков на 64-разрядном компьютере достаточно примерно 53 итерации (при скорости современных компьютеров и низкой стоимости одного шага такое число итераций не представляет затруднений).

Критерий окончания счета в методе дихотомии таков. Итерации прерываются в случае, когда длина очередного отрезка меньше требуемого уровня точности.

В методе дихотомии для нахождения очередного приближения требуется знать две предыдущие точки. Такие алгоритмы называют *двухшаговыми*.

Метод дихотомии имеет следующие достоинства: 1) от функции требуется только непрерывность; 2) метод очень прост, исключительно надежен и всегда сходится к корню; 3) на каждой итерации значение функции вычисляется только один раз.

Недостатки метода следующие: 1) нужно заранее найти отрезок, на концах которого функция имеет разные знаки; 2) если у функции на отрезке $[x_0, x_1]$ имеется несколько корней, то заранее неизвестно, к какому из них сойдется дихотомия; 3) метод не позволяет найти корни четной кратности; 4) корни нечетной кратности находятся, но кратность их установить невозможно, а процесс тем более чувствителен к ошибкам округления, чем выше кратность корня; 5) для многочлена невозможно вычислить комплексные корни; 6) дихотомия не обобщается на случай функции многих переменных.

Из-за своей простоты и надежности метод дихотомии широко применяется в стандартных программах. Для устранения первого недостатка в программах предусматривают следующий прием. Пользователь может задать только одну точку x_0, в окрестности которой он ищет корень. Программа вычисляет значения $f(x)$ в нескольких точках вправо и влево от x_0, пока не найдет точку, в которой знак функции противоположен знаку $f(x_0)$. Затем производится деление найденного отрезка.

Однако описанный прием может не сработать, если корень уравнения лежит на узком экстремуме функции.

2.2.2. Метод Ньютона

Метод Ньютона называют также методом касательных. Для построения метода требуется существование первой непрерывной производной функции $f(x)$. Выбрав начальное приближение x_0, проведем касательную к графику функции, и точку пересечения касательной с осью абсцисс примем за следующее приближение к корню (рис. 2.5).

Математически этот процесс можно записать следующим образом:

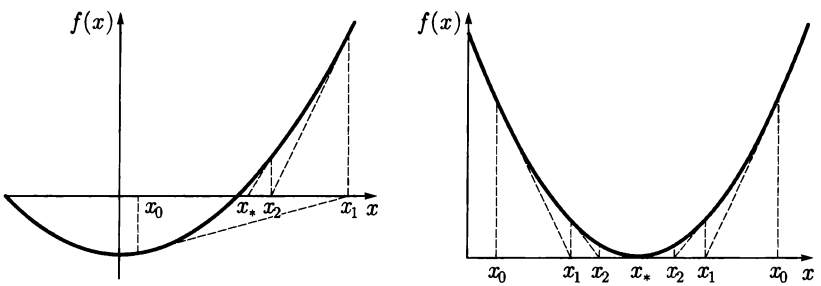
$$\begin{gathered} 0 \approx f(x_{s+1}) = \\ = f(x_s + (x_{s+1} - x_s)) \approx f(x_s) + f'(x_s)(x_{s+1} - x_s), \end{gathered} \tag{2.29}$$

откуда

$$x_{s+1} = x_s - \frac{f(x_s)}{f'(x_s)}. \tag{2.30}$$

Для вычисления очередного приближения x_{s+1} нужно знать только одно предыдущее значение x_s. Подобные алгоритмы называют *одношаговыми*.

По рис. 2.5 нетрудно сообразить, что метод Ньютона может сходиться к корню любой кратности: простому или кратному, как нечетной, так и четной кратности. Из рис. 2.5, *а* видно, что если корень простой, то с одной из сторон корня итерации сходятся монотонно; если выбрать нулевое приближение с другой стороны корня, то сначала происходит переброс на «правильную» сторону, а далее итерации сходятся монотонно. Если ко-



а *б*

Рис. 2.5. Метод Ньютона:

а — случай простого корня; *б* — случай кратного корня

рень кратный, то итерации с любой стороны сходятся монотонно; для корня четной кратности это видно из рис. 2.5, *б*, а для корня нечетной кратности картина аналогична.

Однако описанная картина справедлива в не слишком большой окрестности корня. Если x_0 выбрано далеко от корня x_*, а графиком функции является достаточно сложная кривая с несколькими экстремумами, то первая же итерация (2.30) может отбросить нас в сторону от корня и далее итерации могут не сходиться.

Сходимость. Аккуратное исследование сходимости можно провести лишь в небольшой окрестности корня x_*. Пусть это корень кратности p, а функция имеет p непрерывных ограниченных производных. Тогда вблизи корня x_* справедливы следующие представления:

$$f(x) \approx a(x - x_*)^p, \quad f'(x) \approx pa(x - x_*)^{p-1}. \tag{2.31}$$

Обозначим отклонение от корня $\Delta_s = x_s - x_*$. Вычтем x_* из обеих частей (2.30) и подставим в нее представление (2.31). Тогда для очередного отклонения получим

$$\Delta_{s+1} \approx \left(1 - \frac{1}{p}\right) \Delta_s. \tag{2.32}$$

Если корень простой $p = 1$, то выражение в скобках в (2.32) обращается в нуль и $|\Delta_{s+1}| \ll |\Delta_s|$. Сходимость метода Ньютона вблизи простого корня оказывается очень быстрой. Для этого вывода достаточно требования непрерывности и ограниченности первой производной.

Если при простом корне дополнительно потребовать непрерывности и ограниченности второй производной $f''(x)$, то можно уточнить оценку скорости сходимости. В этом случае к представлениям (2.31) можно добавить следующие степени: $f(x) \approx \approx a\Delta + b\Delta^2$, $f'(x) \approx a + 2b\Delta$ (здесь $p = 1$). Подстановка этих выражений в формулу Ньютона дает

$$\Delta_{s+1} \approx \frac{b}{a}\Delta_s^2 = O(\Delta_s^2).$$

Погрешность очередной итерации пропорциональна квадрату погрешности на предыдущем шаге. Такую сходимость называют *квадратичной*. На практике это означает, что можно останавливать итерации по условию

$$|x_{s+1} - x_s| < \varepsilon |x_{s+1}|, \quad \varepsilon \approx 10^{-10};$$

величина $|x_{s+1}|$ в правой части поставлена для того, чтобы ε соответствовала относительной точности получаемого числа. Это обеспечивает 15 — 20 верных знаков в значении очередного приближения x_{s+1}.

Пусть корень кратный $p \leqslant 2$, тогда из (2.32) видно, что итерации монотонно сходятся со скоростью геометрической прогрессии, знаменатель которой $q = 1 - 1/p$, $0,5 \leqslant q < 1$. Такую сходимость называют *линейной*. Линейная сходимость является гораздо более медленной, чем квадратичная. Даже при наименьшем из возможных здесь знаменателей $q = 0,5$ она будет такой же, как у дихотомии. Чем больше кратность корня p, тем ближе q к 1 и тем медленнее сходятся итерации.

Диагностика кратности. Знаменатель геометрической прогрессии можно приближенно определить из следующего соотношения:

$$q \approx q_s \equiv \frac{x_{s+1} - x_s}{x_s - x_{s-1}}. \tag{2.33}$$

Когда итерации приближаются к корню, $q_s \to q$. Если мы визуально наблюдаем стремление q_s к некоторому пределу q, то по величине этого предела нетрудно восстановить кратность корня:

$$p \approx 1/(1 - q_s).$$

Эта формула применима как для кратного корня, так и для простого. Дополнительный контроль очевиден: полученное p должно быть очень близким к целому числу.

Такой диагностики пока еще нет в стандартных программах.

Критерий остановки итераций. Для кратного корня погрешность очередного приближения можно оценить как сумму геометрической прогрессии, состоящей из еще не сделанных итераций. Это дает следующий критерий остановки итераций:

$$\left| \frac{q}{1 - q}(x_{s+1} - x_s) \right| < \varepsilon |x_{s+1}|, \tag{2.34}$$

где ε — требуемая относительная точность, а q вычисляется по (2.33). Однако здесь использование $\varepsilon \sim 10^{-10}$ дает именно такую точность, а не 15 — 20 верных знаков, как в случае простого корня.

Достоинства метода Ньютона следующие: 1) сходится для корней любой кратности; 2) квадратичная сходимость вблизи простого корня; 3) возможность диагностики кратности корня; 4) не нужно искать отрезок, на котором функция меняет знак; 5) обобщается на случай многих переменных.

Недостаток метода Ньютона в том, что при далеком начальном приближении сходимость не гарантирована.

Разностная производная. В формулу Ньютона входит производная. Даже если задано явное выражение $f(x)$, оно может быть сложным и содержать трудные для явного дифференцирования выражения. В этом случае можно приближенно заменить производную симметричной разностью:

$$f'(x_s) \approx [f(x_s + h) - f(x_s - h)]/(2h). \tag{2.35}$$

Надо брать достаточно малые значения h, чтобы погрешность такой замены не ухудшила сходимость. Однако слишком малое h недопустимо, так как при вычитании близких значений функции сокращаются первые значащие цифры и относительные ошибки округления становятся большими. Можно получить такую качественную оценку оптимального значения h:

$$h_{\text{opt}} \sim \delta^{1/3}|x_s - x_{s-1}|^{k(p)}, \quad k(1) = 1/3,\ k(2) = 2/3,\ k(p \geqslant 3) = 1,$$

где δ — относительная погрешность единичного округления (10^{-16} на 64-разрядном компьютере). Видно, что по мере сходимости итераций h_{opt} довольно быстро уменьшается, кратные корни требуют меньшего значения h, чем простой корень.

Не следует использовать вместо симметричной разности (2.35) одностороннюю $f'(x_s) \approx [f(x_s + h) - f(x_s)]/h$. Погрешность такой замены существенно больше, и потребуется использовать гораздо меньшие значения h, что увеличит ошибки округления и ухудшит окончательную точность.

Прием Гарвика. Пользоваться оптимальным шагом, меняющимся по мере расчета, довольно сложно. Поэтому на практике часто ограничиваются постоянным шагом h и страхуются от ошибок округления следующим приемом. Задают для критерия сходимости, помимо очень малого $\varepsilon \sim 10^{-10}$, более грубое $\varepsilon_1 \sim 10^{-5}$. Когда выполнится критерий сходимости (2.34) с ε_1, переходят на окончательное значение ε и одновременно проводят проверку дополнительного условия

$$|x_{s+1} - x_s| < |x_s - x_{s-1}|. \tag{2.36}$$

Итерации в малой окрестности корня должны сходиться монотонно; следовательно, если критерий (2.36) нарушился, это означает срыв процесса из-за ошибок округления. Поэтому теперь итерации ведутся до тех пор, пока либо выполнится критерий сходимости (2.34) с $\varepsilon \sim 10^{-10}$, либо нарушится условие (2.36).

Таблица 2.1

Извлечение корня квадратного из $a = 4$ по формуле (2.37)

s	x_s	s	x_s	s	x_s
0	10,0000	2	2,9846	4	2,0061
1	5,2000	3	2,1624	5	2,0000

Пример 2.3. Первичными операциями компьютера являются логические и четыре действия арифметики. Все остальное вычисляется с помощью стандартных программ, содержащих только указанные ранее операции. Опишем алгоритм, на котором основана стандартная программа вычисления корня квадратного из числа $a > 0$.

Задача сводится к нахождению нуля функции $f(x) = x^2 - a$. Подставив эту функцию в (2.30), получим

$$x_s = \frac{1}{2}\left(x_s + \frac{a}{x_s}\right). \tag{2.37}$$

Именно этот случай изображен на рис. 2.5, *а*. Видно, что корень простой, итерации сходятся при любом $x_0 > 0$, причем вблизи корня сходимость квадратичная. Это хорошо иллюстрируется табл. 2.1, где выбрано заведомо неудачное начальное приближение.

Заметим, что для данной задачи нетрудно выбрать неплохое x_0. Запишем число a в стандартном двоичном коде, отбросим половину старших знаков и возьмем получившееся число в качестве начального приближения.

2.2.3. Обобщенный метод Ньютона

Обобщенный метод Ньютона называют также непрерывным аналогом метода Ньютона. Метод был предложен в работах М. К. Гавурина, Е. П. Жидкова и др.

При неудачно выбранном начальном приближении движение по касательной начинается в сторону корня, но перебрасывает нас далеко за корень (см. рис. 2.5, *а*). Перепишем формулу Ньютона в новых обозначениях:

$$x_{s+1} = x_s + \xi_s, \quad \xi_s = -f(x_s)/f'(x_{s+1}). \tag{2.38}$$

Обобщим метод Ньютона, вводя дополнительный шаг τ:

$$x_{s+1} = x_s + \tau\xi_s, \quad 0 < \tau \leqslant 1. \tag{2.39}$$

Положительность τ означает движение по касательной в направлении корня. Выбор достаточно малого значения τ страхует от перескока за корень, но число итераций при этом окажется большим. Значение $\tau = 1$ соответствует классическому методу Ньютона. Оно выгодно вблизи корня, где сходимость быстрая. Значения $\tau > 1$ не следует выбирать: это ухудшает сходимость.

Целесообразно выбирать значения τ так, чтобы они были небольшими вдали от корня и стремились к 1 вблизи корня.

Оптимальный шаг. Были предложены разные методы выбора τ. Они основаны на правдоподобных, но нестрогих рассуждениях. Приведем один эффективный и очень простой способ.

Введем неотрицательную функцию $\varphi(\tau) = f^2(x_s + \tau\xi_s) \geqslant 0$. Ее минимальное значение равно 0 и достигается на корне исходного уравнения. Пусть τ возрастает от 0 до 1. При достаточно малых τ функция $\varphi(\tau)$ будет убывающей. Когда мы перескочили за корень, $\varphi(\tau)$ начинает возрастать. Если $\varphi(\tau) \gg \varphi(0)$, то это свидетельствует о далеком перескоке.

Будем выбирать τ по схеме предиктор-корректор. Нам заранее известно $\varphi(0) = f^2(x_s)$. В качестве предиктора возьмем ньютоновский шаг $\tau = 1$ и вычислим $\varphi(1) = f^2(x_s + \xi_s)$. Если $\varphi(1) \ll \varphi(0)$, то это свидетельствует о хорошем приближении к корню и нет необходимости корректировать ньютоновский шаг $\tau = 1$. Если же $\varphi(1) \gg \varphi(0)$, то произошел далекий перескок и нужно существенно уменьшить τ. Запишем несложную формулу корректировки шага:

$$\tau_s = \frac{\varphi(0)}{\varphi(0) + \varphi(1)}. \tag{2.40}$$

Очевидно, $0 < \tau_s < 1$, а сами значения τ_s качественно удовлетворяют сформулированным требованиям: если $\varphi(1) \ll \varphi(0)$, то $\tau_s \approx 1$; если же $\varphi(0) \ll \varphi(1)$, то $0 < \tau_s \ll 1$.

Последний случай приводит к очень малым шагам τ_s и медленной сходимости вдали от корня. Практика расчетов показала, что в (2.40) целесообразно ввести «кухонную» поправку:

$$\tau_s = \frac{\varphi(0) + \theta\varphi(1)}{\varphi(0) + \varphi(1)}, \quad 0 \leqslant \theta \leqslant 1, \tag{2.41}$$

где θ — настроечный параметр программы. Он ограничивает снизу величину шага $\theta < \tau_s \leqslant 1$. Очевидно, при $\theta = 1$ всегда $\tau_s = 1$, т. е. получается классический метод Ньютона.

Рекомендуется следующая стратегия расчета. Сначала полагают $\theta = 1$. Если за ~ 30 итераций сходимости нет (а это означа-

ет, что последовательные приближения хаотичны), то θ уменьшают в 10 раз и продолжают итерации с последнего найденного значения x_s. Если за следующие ~ 30 итераций сходимости по-прежнему нет, то снова уменьшают θ и т. д. Этот процесс повторяют, пока не дойдут до $\theta = 0,001$. Если при этом сходимость не достигнута, то считают, что при данном нулевом приближении сходимости нет. Следует выбрать другое нулевое приближение.

Обобщенный метод Ньютона в широкой окрестности корня любой кратности сходится линейно. В малой же окрестности корня его сходимость квадратична вблизи простого корня и линейна вблизи кратного.

Диагностика кратности. В обобщенном методе Ньютона также можно определить кратность корня. Для этого вычисляется такой же знаменатель сходимости

$$q_s = \frac{x_{s+1} - x_s}{x_s - x_{s-1}},$$

который вблизи корня стремится к пределу q. Однако можно показать, что при выборе шага согласно (2.41) этот предел связан с кратностью корня p более сложным соотношением:

$$q = 1 - \frac{1}{p}\frac{1 + \theta(1 - 1/p)^{2p}}{1 + (1 - 1/p)^{2p}}. \qquad (2.42)$$

Видно, что для простого корня $p = 1$ по-прежнему $q = 0$ и сходимость квадратичная. Для $p \geqslant 2$ сходимость линейна, а ее знаменатель немного ближе к 1, чем для классического метода Ньютона. Тем самым сходимость немного медленнее, но устойчивость метода лучше.

Простой корень $p = 1$ из (2.42) обнаруживается немедленно, но для кратных корней находить $p(q)$ из нелинейного уравнения (2.42) довольно сложно. Проще рассчитать таблицу значений $q(p)$ и отождествлять расчетный предел знаменателя q_s с ближайшим к нему числом этой таблицы.

2.2.4. Прочие методы

Существует очень много методов решения уравнения $f(x) = 0$. Кратко опишем некоторые наиболее распространенные методы, положенные в основу ряда стандартных программ.

Метод хорд. Его называют также методом секущих, хотя это название лучше приложимо к другому методу. Метод

хорд близок к дихотомии. Необходимо найти две точки x_0, x_1, в которых $f(x)$ принимает значения разных знаков. На отрезке $[x_0, x_1]$ расположен корень. Концы кривой на этом отрезке соединяют хордой и находят точку x_2, в которой хорда пересекается с осью абсцисс. Вычисляют $f(x_2)$ и сравнивают ее знак со знаками $f(x_0)$ и $f(x_1)$. Отбрасывают ту часть отрезка, на концах которой знаки одинаковы. Затем процесс повторяют.

Этот метод двухшаговый. Он сходится линейно, а его знаменатель может быть как меньше, так и больше 0,5. Поэтому метод не лучше, но и не хуже дихотомии.

Метод секущих. Это двухшаговый метод. Пусть известны две итерации x_s и x_{s-1}. Проведем через точки кривой $f(x_s)$ и $f(x_{s-1})$ секущую и найдем точку пересечения секущей с осью абсцисс:

$$x_{s+1} = x_s - \frac{(x_s - x_{s-1})f(x_s)}{f(x_s) - f(x_{s-1})}. \tag{2.43}$$

Формула (2.43) похожа на классический метод Ньютона (2.30), в котором $f'(x_s)$ заменена односторонним разностным отношением. Одна итерация этого метода более экономична, так как на ней нужно только однажды вычислить $f(x_s)$ и не вычислять производную.

Однако сходимость метода секущих хуже, чем метода Ньютона. Можно показать, что даже вблизи простого корня сходимость вместо квадратичной подчиняется закону

$$|x_{s+1} - x_s| < \text{const}\,|x_s - x_{s-1}|^\beta, \quad \beta = (1 + \sqrt{5})/2 \approx 1,618.$$

Это почти в 1,5 раза увеличивает число итераций. Сходимость даже вблизи корня может стать немонотонной. Вдобавок вычисления вблизи корня становятся чувствительными к ошибкам округления, так как происходит вычитание близких чисел. Все это обесценивает предполагаемую экономию.

Метод парабол. Это трехшаговый метод. Пусть известны три итерации x_s, x_{s-1}, x_{s-2}. Тогда через соответствующие три точки кривой $f(x)$ можно провести параболу. Эта парабола имеет две точки пересечения с осью абсцисс. За новое приближение x_{s+1} принимается та точка пересечения, которая лежит ближе к последнему приближению x_s.

Скорость сходимости этого метода меньше квадратичной. Даже вблизи простого корня она подчиняется зависимости

$$|x_{s+1} - x_s| < \text{const}\,|x_s - x_{s-1}|^\beta, \quad \beta \approx 1,84.$$

Метод парабол еще более чувствителен к ошибкам округления, чем метод секущих.

Интерес к этому методу был связан с тем, что парабола может и не пересекаться с осью абсцисс. В этом случае значение x_{s+1} оказывается комплексным. Это позволяет естественно находить комплексные корни многочленов, имеющих вещественные коэффициенты.

Однако для многочленов с вещественными коэффициентами комплексные корни можно найти и методом Ньютона. Для этого нужно выбрать комплексные значения x_0.

Простые итерации. Метод простой итерации также называют методом последовательных приближений. Пусть уравнение $f(x) = 0$ заменено эквивалентным уравнением специального вида $x = \varphi(x)$ с непрерывной $\varphi(x)$. Строится следующий итерационный процесс:

$$x_{s+1} = \varphi(x_s). \tag{2.44}$$

Если итерации сходятся, то $\lim\limits_{s\to\infty} x_s$ в силу непрерывности является корнем исходного уравнения x_*. Процесс (2.44) очень прост, а одна итерация не трудоемка. Однако для сходимости процесса хотя бы в малой окрестности корня необходимо $|\varphi'(x_*)| < 1$.

Нелегко бывает преобразовать исходное уравнение к нужной форме, чтобы это условие выполнялось. Кроме того, такое преобразование нужно индивидуально подбирать для каждого конкретного уравнения, что неприемлемо для стандартных программ.

2.2.5. Удаление корней

Уравнение $f(x) = 0$ может иметь несколько корней, как простых, так и кратных. Нередко требуется найти все корни уравнения или достаточно много его корней. Можно брать различные начальные приближения. Для выбора x_0 нередко используют случайные числа (см. подразд. 3.3.3). При этом велики шансы на то, что из разных нулевых приближений итерации сойдутся к разным корням. Однако возможно, что из разных нулевых приближений итерации будут сходиться к одному и тому же корню, минуя остальные. Поэтому такой способ не очень надежен.

Более надежным является исключение уже найденных корней. Пусть мы уже нашли x_*, являющийся p-кратным корнем уравнения $f(x) = 0$. Введем новую функцию:

$$g(x) = f(x)(x - x_*)^p. \tag{2.45}$$

Величина x_* уже не является нулем функции $g(x)$. Но все остальные нули функции $f(x)$ являются нулями $g(x)$. Поэтому нахождение оставшихся корней сводится к решению уравнения $g(x) = 0$.

Для описанного процесса очень важно при нахождении каждого корня x_* устанавливать его кратность. Казалось бы, можно поделить $f(x)$ на $x - x_*$. У полученной $g(x)$ снова будет корнем x_*. Итерации метода Ньютона для $g(x)$ снова сойдутся к x_*, мы его снова исключим и т. д. На самом деле из-за ошибок округления мы находим лишь приближенное значение корня $\tilde{x}_*$ и получаем $\tilde{g}(x) = f(x)/(x - \tilde{x}_*)$. Искаженная функция $\tilde{g}(x)$ будет иметь корень x_* и полюс в очень близкой к ней точке $\tilde{x}_*$. Для такой функции итерации метода Ньютона могут плохо сходиться. Поэтому обязательно нужно исключать найденный корень x_*, учитывая его кратность.

Определив какой-нибудь корень $g(x) = 0$ и установив его кратность, можно аналогично ввести новую функцию, исключив и этот корень. Процесс продолжается до тех пор, пока не будут найдены все нужные корни.

Заметим, что если мы хотим исключить корни, то нельзя пользоваться дихотомией. Корни четной кратности дихотомия вообще не обнаруживает. Корень нечетной кратности дихотомия обнаруживает, но не отличает его от простого. После деления на $(x - x_*)$ остается корень четной кратности; поэтому дальше дихотомия этот корень не замечает.

Улучшение точности. Даже метод Ньютона для простого корня не всегда позволяет найти все представимые в компьютере знаки; один-два последних десятичных знака нередко оказываются неверными. Для кратных корней неверными могут оказаться последние три — пять цифр. Поэтому отличие найденного корня $\tilde{x}_*$ от точного x_* реально существенно превышает ошибки округления. В результате расчет $g(x)$ по (2.45) с подстановкой $\tilde{x}_*$ ощутимо ухудшает точность.

Проводить таким способом многократное исключение корней рискованно. Накопление ошибки может стать серьезным. Опишем несложный прием, улучшающий точность.

Для приближенного значения корня $f(\tilde{x}_*) \neq 0$. Вместо формулы (2.45) определим $g(x)$ несколько иначе:

$$\tilde{g}(x) = \frac{f(x) - f(\tilde{x}_*)}{(x - \tilde{x}_*)^p}. \tag{2.46}$$

Теперь числитель обращается в нуль при $x = \tilde{x}_*$. Поэтому его можно делить на знаменатель. Строго говоря, точка $\tilde{x}_*$ будет

лишь простым нулем числителя, а не p-кратным. Однако на практике это оказывается малозаметным.

Далее надо находить корень $\tilde{g}(x)$, исключать его описанным способом и т. д. Данный прием существенно облегчает нахождение большого числа корней уравнения.

Корни многочлена. Особенно прост и надежен описанный процесс, если $f(x)$ является многочленом степени N. Тогда уравнение $f(x) = 0$ имеет ровно N корней с учетом кратности. Эти корни могут быть комплексными даже в случае многочлена с вещественными коэффициентами.

Как известно, для многочлена отношение (2.45) есть также многочлен степени $N - p$. Многочлен $g(x)$ целесообразно записать в явном виде, чтобы избежать ошибок округления при делении. Дадим способ явного нахождения этого многочлена для простого корня $p = 1$. Запишем оба многочлена в явном виде:

$$f(x) = \sum_{n=0}^{N} a_n x^n; \quad g(x) = \sum_{n=0}^{N-1} b_n x^n. \tag{2.47}$$

Очевидно $a_N \neq 0$, $b_{N-1} \neq 0$. Подставив эти суммы в соотношение $f(x) = (x - x_*)g(x)$ и приравняв коэффициенты при одинаковых степенях x, получим рекуррентную цепочку для вычисления коэффициентов $g(x)$:

$$b_{N-1} = a_N; \quad b_{n-1} = a_n + x_* b_n, \quad n = N-1, N-2, \ldots, 0. \tag{2.48}$$

При последнем значении $n = 0$ формально вычисляется b_{-1}. Оно должно быть нулем, но из-за ошибок округления этого не случается. Величина b_{-1} позволяет оценить влияние ошибок округления. При невысоких степенях многочлена это влияние незначительно и им можно пренебречь.

Заметим, что отбрасывание коэффициента b_{-1} эквивалентно общему приему вычитания $f(\tilde{x}_*)$ по формуле (2.46). Однако для многочленов можно предложить другой способ, при котором ошибки накапливаются слабее. Будем менять для компенсации ошибки не младший коэффициент исходного многочлена, а старший. Для этого положим

$$\tilde{a}_N = \left(\sum_{n=0}^{N-1} a_n \tilde{x}_*^n \right) \Big/ \tilde{x}_*^N. \tag{2.49}$$

Подставим измененный коэффициент $\tilde{a}_N$ в исходный многочлен (2.47) и поделим его на $x - \tilde{x}_*$ по формулам (2.48). Теперь получим $b_{-1} = 0$ с точностью до ошибок округления.

Если корень x_* является p-кратным, то для получения коэффициентов многочлена $g(x)$ нужно p раз последовательно выполнить описанное ранее деление. Можно сразу делить исходный многочлен на $(x - x_*)^p$, но соответствующие формулы более громоздки. Ошибки округления при этом увеличиваются тем сильнее, чем больше p. При очень высоких степенях многочлена влияние ошибок округления может оказаться существенным. Об этом свидетельствует следующий пример.

Пример 2.4. Рассмотрим многочлен высокой степени $f(x) = (x - x_*)^N$. Он имеет корень x_* кратности N. Слегка изменим этот многочлен, добавив к свободному члену малую величину δ, лежащую на уровне ошибок округления: $\tilde{f}(x) = (x - x_*)^N - \delta$. Измененный многочлен имеет N различных комплексных корней: $x_{*k} = x_* + \delta^{1/N} e^{2\pi i k/N}$, $0 \leqslant k \leqslant N - 1$. Если $\delta \sim 10^{-15}$ и $N \approx 30$, то $\delta^{1/N} \approx 0,3$. От незначительной ошибки округления корни изменились на существенную величину!

Очевидно, для уменьшения ошибок округления при исключении корней необходимо находить x_* с максимально возможной точностью.

Практика вычислений показала также, что ошибки округления существенно уменьшаются, если начинать процесс исключения с меньших по модулю корней. Для этого выбирают $x_0 = 0$; тогда итерации обычно сходятся к наименьшему по модулю корню. Исключив его, снова выбирают $x_0 = 0$ и т. д.

Если многочлен имеет вещественные коэффициенты, но нужно найти его комплексные корни, то выбирают x_0 комплексным. Комплексное число x_* может быть корнем многочлена с вещественными коэффициентами только в паре со своим комплексно-сопряженным. Так что практически сразу находится пара корней. Исключать также нужно сразу эту пару корней.

2.3. СИСТЕМЫ НЕЛИНЕЙНЫХ УРАВНЕНИЙ

2.3.1. Метод Ньютона

Одномерный метод дихотомии не обобщается на системы уравнений. Поэтому остается только один метод, пригодный для обобщения на системы нелинейных уравнений, — метод Ньютона. Рассмотрим систему N уравнений

$$f_n(x_1, x_2, \ldots, x_N) = 0, \quad 1 \leqslant n \leqslant N. \tag{2.50}$$

Ее можно записать также в векторной форме:

$$\mathbf{f}(\mathbf{x}) = 0, \quad \mathbf{f} = \{f_1, f_2, \ldots, f_N\}, \quad \mathbf{x} = \{x_1, x_2, \ldots, x_N\}. \quad (2.51)$$

Пусть уже найдено некоторое приближение $\mathbf{x}^{(s)} = \{x_1^{(s)}, x_2^{(s)}, \ldots, x_N^{(s)}\}$. Для этого приближения $\mathbf{f}(\mathbf{x}^{(s)})$ отличается от нуля сильнее, чем нас устраивает. Считаем, что для следующего приближения система (2.50) — (2.51) почти удовлетворяется: $\mathbf{f}(\mathbf{x}^{(s+1)}) \approx \approx 0$. Подставим сюда $\mathbf{x}^{(s+1)} = \mathbf{x}^{(s)} + (\mathbf{x}^{(s+1)} - \mathbf{x}^{(s)})$ и разложим в ряд Тейлора по малому приращению $\mathbf{x}^{(s+1)} - \mathbf{x}^{(s)}$. Ограничиваясь только первым членом разложения, получим

$$\mathbf{f}(\mathbf{x}^{(s)}) + \frac{\partial \mathbf{f}(\mathbf{x}^{(s)})}{\partial \mathbf{x}} \Delta^{(s)} = 0, \quad \Delta^{(s)} = \mathbf{x}^{(s+1)} - \mathbf{x}^{(s)}. \quad (2.52)$$

Здесь $\partial \mathbf{f}/\partial \mathbf{x} = \{\partial f_n / \partial x_m\}$ — матрица Якоби (матрица первых производных).

Выражение (2.52) можно переписать в виде

$$\frac{\partial \mathbf{f}(\mathbf{x}^{(s)})}{\partial \mathbf{x}} \Delta^{(s)} = -\mathbf{f}(\mathbf{x}^{(s)}), \quad \mathbf{x}^{(s+1)} = \mathbf{x}^{(s)} + \Delta^{(s)}. \quad (2.53)$$

Таким образом, для нахождения приращения $\Delta^{(s)}$ на каждой итерации нужно решить систему линейных уравнений. Ее можно записать в покоординатной форме

$$\sum_{m=1}^{N} \frac{\partial f_n(x_1^{(s)}, x_2^{(s)}, \ldots, x_N^{(s)})}{\partial x_m} \Delta_m^{(s)} = -f_n(x_1^{(s)}, x_2^{(s)}, \ldots, x_N^{(s)}), \quad (2.54)$$

$$1 \leqslant n \leqslant N;$$

$$x_n^{(s+1)} = x_n^{(s)} + \Delta_n^{(s)}.$$

Сходимость многомерного метода Ньютона исследована только для функций (2.50), имеющих вторые непрерывные производные, и простого корня $\mathbf{x}_*$ (корень называется простым, если в этой точке матрица Якоби неособенная, т. е. $\det[\partial \mathbf{f}(\mathbf{x}_*)/\partial \mathbf{x}] \neq 0$). В этом случае доказано, что в некоторой окрестности корня итерации сходятся, причем квадратично: $\|\mathbf{x}^{(s+1)} - \mathbf{x}_*\| \leqslant \text{const}\, \|\mathbf{x}^{(s)} - \mathbf{x}_*\|^2$. Таким образом, в окрестности простого корня многомерный метод Ньютона сходится так же быстро, как и одномерный. Поэтому критерием сходимости можно выбрать условие

$$\|\mathbf{x}^{(s+1)} - \mathbf{x}^{(s)}\| \leqslant \varepsilon \|\mathbf{x}^{(s)}\|, \quad \varepsilon \sim 10^{-10}. \quad (2.55)$$

Это обеспечивает получение 15 — 16 верных знаков.

В формуле (2.55) в качестве нормы вектора можно использовать любую из норм вектора: длину

$$\|x\|_{l_2} = (\sum_{n=1}^{N} x_n^2)^{1/2} \tag{2.56}$$

или максимальную по модулю компоненту

$$\|x\|_c = \max_{1 \leqslant n \leqslant N} |x_n|. \tag{2.57}$$

На практике целесообразно пользоваться длиной (2.56).

Замораживание. На каждой итерации метода Ньютона необходимо решать линейную систему с матрицей Якоби. Матрица производных содержит N^2 элементов. Ее вычисление как минимум в N раз более трудоемко, чем вычисление правых частей. Кроме того, само решение линейной системы требует $\sim N^3$ операций, что при больших N гораздо более трудоемко, чем вычисление матрицы производных.

Поэтому нередко предлагают следующий прием, рассчитанный на уменьшение трудоемкости итерации. Замораживают матрицу производных и делают с этой матрицей несколько итераций (по меньшей мере 3 — 5). При этом не приходится заново вычислять производные. Многократное решение линейной системы с одной и той же матрицей выполняется экономично, если запоминать коэффициенты c_{nm} и выполнять прямой ход метода Гаусса только один раз (см. подразд. 2.1.2). Этот вариант в литературе называют модифицированным методом Ньютона.

Однако мы не рекомендуем использовать этот способ. Расчет с замороженной матрицей обеспечивает не квадратичную, а лишь линейную сходимость. Поэтому число итераций существенно возрастает и общая трудоемкость расчета не уменьшается. Вдобавок критерий окончания (2.55) при линейной сходимости обеспечивает точность лишь $O(\varepsilon)$, а не $O(\varepsilon^2)$. Поэтому найти корень с высокой точностью становится намного труднее.

Разностная производная. Даже для одной переменной нахождение явного выражения производной нередко затруднительно. В многомерном случае это тем более нелегко сделать. В этой ситуации можно заменять производные симметричными разностями аналогично (2.35). При этом значение оптимального шага для каждой производной будет, вообще говоря, своим и уменьшающимся по мере сходимости итераций. Дать аккуратную оценку h_{opt} в многомерном случае достаточно трудно.

Поэтому на практике целесообразно ограничиться постоянным шагом h и страховаться от ошибок округления приемом Гарвика (см. подразд. 2.2.2). В формулах (2.36) вместо модулей берутся нормы векторов.

Начальное приближение. Даже одномерный метод Ньютона при неудачном начальном приближении плохо сходится. Для случая многих переменных хорошо выбрать начальное приближение еще труднее, поскольку компьютерный просмотр графиков практически невозможен. Поэтому часто пользуются так называемым случайным поиском.

Для этого берут прямоугольный N-мерный параллелепипед, в котором предполагается существование искомого решения. В нем выбирают некоторое число случайных равномерно распределенных точек (способ построения таких точек будет описан в подразд. 3.3.3). Каждую точку используют в качестве начального приближения. Если при некотором начальном приближении итерации плохо сходятся, процесс прекращают. Если они сходятся хорошо, то мы получаем одно из решений. При этом из разных начальных приближений возможна сходимость как к разным решениям, так и к одному и тому же.

Таким образом удается найти различные решения системы (2.53). Практика показала, что обычно достаточно выбрать $(5 \div 20)N$ случайных начальных приближений. Если же при этом не удается найти ни одного решения, то лучше не увеличивать число начальных приближений, а изменить область поиска.

2.3.2. Обобщенный метод Ньютона

Обобщенный метод Ньютона строится полностью аналогично одномерному случаю (см. подразд. 2.2.3). В формулу одной итерации вводится скалярный шаг τ:

$$\mathbf{x} = \mathbf{x}^{(s)} + \tau\Delta^{(s)}, \quad 0 < \tau \leqslant 1;$$

здесь $\Delta^{(s)}$ определяется из решения линейной системы (2.53). При $\tau = 1$ это дает нам обычную ньютоновскую итерацию. Вводится скалярная функция

$$\varphi(\tau) = \mathbf{f}^2(\mathbf{x}^{(s)} + \tau\Delta^{(s)}) = \sum_{n=1}^{N} f_n^2(\mathbf{x}^{(s)} + \tau\Delta^{(s)}).$$

Здесь $\varphi(0) = \mathbf{f}^2(\mathbf{x}^{(s)})$ соответствует исходной итерации, а $\varphi(1) = \mathbf{f}^2(\mathbf{x}^{(s)} + \Delta^{(s)})$ есть ньютоновский предиктор. Корректором будет значение

$$\tau_s = \frac{\varphi(0) + \theta\varphi(1)}{\varphi(0) + \varphi(1)}, \quad \theta \approx 0,1 \div 0,01.$$

Как и в одномерном случае, это обобщение расширяет область сходимости. Вдали от корня сходимость при этом лишь линейная. Вблизи простого корня $\tau_s \to 1$ и сходимость становится квадратичной. Поправочный коэффициент θ вводится для того, чтобы ограничить снизу значение шага: $\theta \leqslant \tau_s \leqslant 1$. По умолчанию рекомендуется выбирать $\theta \approx 0{,}1$. Если этого недостаточно для сходимости, то θ уменьшают, но лишь до $\theta \approx 0{,}01$.

Метод Ньютона и обобщенный метод Ньютона позволяют успешно решать нелинейные системы общего вида с $N \sim 100$. Известны применения этих методов и к некоторым задачам гораздо большей размерности.

Глава 3

ЧИСЛЕННОЕ ИНТЕГРИРОВАНИЕ

3.1. КВАДРАТУРНЫЕ ФОРМУЛЫ

3.1.1. Интегральная сумма

Пусть требуется вычислить однократный интеграл от непрерывной и достаточно гладкой функции $u(x)$:

$$U = \int_a^b u(x)dx. \tag{3.1}$$

Иногда в (3.1) удобнее выделить основную особенность $u(x)$ в виде весовой функции $\rho(x)$ и рассмотреть более общую задачу:

$$U = \int_a^b u(x)\rho(x)dx, \quad \rho(x) > 0.$$

Введем на отрезке $a \leqslant x \leqslant b$ сетку Ω с узлами x_n:

$$\Omega_N = \{a = x_0 < x_1 < x_2 < \ldots < x_N = b\}. \tag{3.2}$$

Шагом сетки назовем величину

$$h_n = x_n - x_{n-1}, \quad 1 \leqslant n \leqslant N. \tag{3.3}$$

Равномерной называют сетку, все шаги которой одинаковы:

$$h_n \equiv h = (b-a)/N. \tag{3.4}$$

Сетку с неодинаковыми шагами называют неравномерной. Отрезок

$$x_{n-1} \leqslant x \leqslant x_n, \quad 1 \leqslant n \leqslant N. \tag{3.5}$$

называют n-м ***интервалом*** сетки. Вводят также середину интервала

$$x_{n-1/2} = (x_n + x_{n-1})/2, \quad 1 \leqslant n \leqslant N. \tag{3.6}$$

Выберем в каждом интервале произвольную точку $\tilde{x}_n$ и составим интегральную сумму

$$U_N = \sum_{n=1}^{N} h_n u(\tilde{x}_n), \quad \tilde{x}_n \in [x_{n-1}, x_n], \quad 1 \leqslant n \leqslant N. \tag{3.7}$$

Устремим все шаги сетки к нулю. При этом для кусочно-непрерывной функции $u(x)$ независимо от выбора $\tilde{x}_n$ существует предел интегральных сумм U_N, который равен искомому интегралу U. Но скорость стремления к этому пределу зависит от выбора значений $\tilde{x}_n$. Далее рассмотрим, как выбрать $\tilde{x}_n$ для обеспечения хорошей сходимости.

3.1.2. Формула средних

Есть простое практическое правило, которое нетрудно обосновать: симметричное построение формулы при прочих равных условиях повышает ее точность. Для обеспечения симметрии выберем в качестве $\tilde{x}_n$ середину интервала $\tilde{x}_n = x_{n-1/2}$ и получим формулу средних (рис. 3.1):

$$U_N = \sum_{n=1}^{N} h_n u(x_{n-1/2}). \tag{3.8}$$

Из рис. 3.1 видно, что для линейной функции $u(x)$ формула средних дает точное значение интеграла. Оценим погрешность при вычислении интеграла по одному интервалу сетки, предполагая наличие непрерывной $u''(x)$.

Оценку можно получить строго путем разложения в ряд Тейлора с центром в середине интервала. Попробуем выписать главный член погрешности более простым способом из соображений размерности.

Обозначим размерность величины квадратными скобками. Размерность интеграла $[U] = [u] \times [h]$. Размерность погрешности

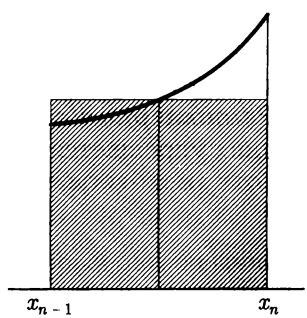


Рис. 3.1. Формула средних

$$R_N = U - U_N \tag{3.9}$$

должна совпадать с размерностью интеграла.

В оценку погрешности не может входить ни значение функции, ни значение ее первой производной, так как формула точна для линейной функции. Значит, в остаточный член формулы средних может входить только вторая производная и $R_N \sim u''$. Размерность производной $[u''] = [u]/[h^2]$. Чтобы получить размерность интеграла, надо домножить u'' на h^3. Тогда погрешность для одного интервала сетки есть $r_n = cu''_{n-1/2}h_n^3$, а полная погрешность $R_N = \sum_{n=1}^{N} r_n$ имеет ту же размерность. Здесь константа c безразмерна и не зависит от конкретного вида функции. Нужно ее найти. Чтобы вычисление константы было несложным, специально выберем интервал $x_{n-1} = 0$, $x_n = 1$ и функцию $u(x) = x^2$. Это простейшая функция, имеющая вторую производную, причем постоянную. Вместо отрезка $[a, b]$ рассмотрим только один интервал $[x_{n-1}, x_n]$. Легко вычислим точный интеграл (3.1), квадратуру средних (3.8) и погрешность (3.9):

$$U = \frac{1}{3}, \quad U_n = \frac{1}{4}, \quad r_n = U - U_n = \frac{1}{12}. \tag{3.10}$$

Отсюда $r_n = ch_n^3 u''_{n-1/2} = 2c$, $c = 1/24$ и главный член погрешности на произвольной неравномерной сетке приобретает следующий вид:

$$R_N = \frac{1}{24}\sum_{n=1}^{N} u''_{n-1/2}h_n^3. \tag{3.11}$$

Это асимптотически точное выражение: при измельчении шагов сетки $h_n \to 0$ погрешность стремится к значению (3.11), так как слагаемыми более высокого порядка малости при $h_n \to 0$ можно пренебречь.

Равномерная сетка. Пусть $h_n = h = \text{const}$, тогда формула средних упрощается:

$$U_N = h\sum_{n=1}^{N} u_{n-1/2}. \tag{3.12}$$

Преобразуем выражение для погрешности (3.11), вынося за знак суммы только h^2. Тогда остается интегральная сумма для второй производной, приближенно заменяемая интегралом:

$$R_N \approx \frac{1}{24} h^2 \sum_{n=1}^{N} h u''_{n-1/2} \approx \frac{h^2}{24} \int_a^b u''(x)dx = \frac{h^2}{24}[u'(b) - u'(a)]. \quad (3.13)$$

Приближенное равенство (3.13) выполнено с тем же порядком точности, что и в (3.11). Таким образом, погрешность интегральной формулы средних

$$R_N = O(h^2) = O(N^{-2}).$$

В таких случаях говорят, что метод имеет ***второй порядок точности***.

Формулы (3.11) и (3.13) дают главный член погрешности. Если $u(x)$ имеет много непрерывных производных, то можно рассмотреть следующие члены разложения в ряд Тейлора. Из симметричного построения формул видно, что члены с нечетными производными исчезнут: их вклад как в интеграл, так и в квадратурную формулу является нулевым в силу симметрии. Разложение погрешности будет содержать лишь четные производные с соответствующими степенями шагов: $u^{(2q)}_{n-1/2} h_n^{2q+1}$. Выражение (3.13) заменится суммой по четным степеням h^{2q}, $q = 1, 2, \ldots$

Если же ослабить требования на гладкость подынтегральной функции и потребовать только кусочной непрерывности $u''(x)$ то проведенные выше выкладки перестают быть справедливыми и удается получить не асимптотическую оценку, а мажорантную оценку погрешности:

$$|R_N| \leqslant \sum_{n=1}^{N} \frac{h_n^3}{24} \max_{x \in [x_{n-1}, x_n]} |u''(x)|.$$

На равномерной сетке это выражение упрощается:

$$|R_N| \leqslant \frac{h^2}{24} M_2, \quad M_2 = \max_{x \in [a,b]} |u''(x)|.$$

Мажорантная оценка менее удобна прежде всего тем, что она обычно сильно завышена.

Пример 3.1. Рассмотрим интеграл, который легко вычисляется точно, а функция на отрезке интегрирования имеет неограниченное число непрерывных производных:

$$u(x) = 1/\sqrt{x}, \quad a = 1, \quad b = 9, \quad U = 4. \quad (3.14)$$

Таблица 3.1

Вычисление тестового интеграла (3.14)

N	Формула средних			Формула трапеций		
	U_N	$\lg\lvert R_N\rvert$	R_{N-1}/R_N	U_N	$\lg\lvert R_N\rvert$	R_{N-1}/R_N
1	3,57771	−0,37		5,33333	0,12	
2	3,82126	−0,75	2,36	4,45552	−0,34	2,93
4	3,93782	−1,21	2,87	4,13839	−0,86	3,29
8	3,98166	−1,74	3,39	4,03810	−1,42	3,63
16	3,99511	−2,31	3,75	4,00988	−2,01	3,86
32	3,99875	−2,90	3,92	4,00250	−2,60	3,96
64	3,99969	−3,50	3,98	4,00063	−3,20	3,99
128	3,99992	−4,11	3,99	4,00016	−3,80	4,00
256	3,99998	−4,71	4,00	4,00004	−4,41	4,00

N	Формула Симпсона			Формула Эйлера		
	U_N	$\lg\lvert R_N\rvert$	R_{N-1}/R_N	U_N	$\lg\lvert R_N\rvert$	R_{N-1}/R_N
1	4,16292	−0,79		2,76543	0,09	
2	4,03268	−1,49	4,99	3,81355	−0,73	6,62
4	4,00467	−2,33	7,00	3,97790	−1,66	8,44
8	4,00047	−3,32	9,88	3,99798	−2,69	10,94
16	4,00004	−4,44	12,89	3,99985	−3,82	13,47
32	4,00000	−5,61	14,87	3,99999	−5,00	15,08
64	4,00000	−6,80	15,67	4,00000	−6,20	15,73
128	4,00000	−8,00	15,91	4,00000	−7,40	15,93
256	4,00000	−9,21	15,98	4,00000	−8,61	15,98

N	Формула Эйлера — Маклорена			Формула Гаусса		
	U_N	$\lg\lvert R_N\rvert$	R_{N-1}/R_N	N	U_N	$\lg\lvert R_N\rvert$
1	13,4272	0,97		1	3,57771	−0,37
2	4,47991	−0,32	19,64	2	3,91809	−1,09
4	4,01954	−1,71	24,56	3	3,98247	−1,76
8	4,00058	−3,23	33,52	4	3,99610	−2,41
16	4,00001	−4,89	45,77	5	3,99911	−3,05
32	4,00000	−6,65	56,29			
64	4,00000	−8,43	61,58	Гауссово-сеточный метод		
128	4,00000	−10,24	63,34	2×4	3,99969	−3,51
256	4,00000	−12,04	63,97			

Вычислим его, используя равномерные сетки с различным числом интервалов N. В табл. 3.1 приведены результаты расчетов на каждой сетке и значения погрешностей $U_N - U$, найденные по точному решению (3.14).

Видно, что погрешности быстро убывают при уменьшении шага. Поскольку шаг каждый раз уменьшается в 2 раза, то главный член погрешности формулы средних должен убывать в 4 раза. В табл. 3.1 приведены отношения погрешностей на соседних сетках. Видно, что отношения на грубых сетках заметно отличаются от 4 из-за примеси высших членов погрешности. Но это отношение монотонно стремится к 4 при сгущении сетки.

Такие тесты можно использовать для проверки правильности программ. Надо взять задачу с достаточно гладкой подынтегральной функцией и известным точным ответом, сосчитать погрешности при сгущении сеток в 2 раза и проанализировать отношения соседних погрешностей. Если эти отношения стремятся к теоретическому значению при сгущении сетки, то программа написана правильно.

3.1.3. Формула трапеций

Не всегда возможно или удобно использовать середины интервалов: например, функция может быть задана таблицей своих значений. В этом случае используют формулу трапеций на произвольной сетке (рис. 3.2); она также имеет симметричный вид:

$$U_N = \sum_{n=1}^{N} \frac{h_n}{2}(u(x_{n-1}) + u(x_n)). \tag{3.15}$$

Формула точна и для линейной функции. Следовательно, остаточный член $R_N = cu''h^3$. Для определения константы c, одинаковой для всех подынтегральных функций, снова воспользуемся описанным ранее приемом и получим

$$U = \int_0^1 x^2 dx = \frac{1}{3}, \quad U_n = \frac{1}{2}, \quad r_n = U - U_n = -\frac{1}{6},$$

$$r_n = ch_n^3 u''_{n-1/2} = 2c, \quad c = -\frac{1}{12}.$$

Погрешность формулы трапеций принимает вид

$$R_N \approx -\frac{1}{12}\sum_{n=1}^{N} u''_{n-1/2} h_n^3. \tag{3.16}$$

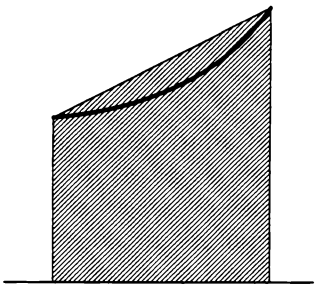


Рис. 3.2. Формула трапеций

Таким образом, погрешность формулы трапеций вдвое больше по модулю, чем в формуле средних, и имеет противоположный знак.

Для равномерной сетки формула трапеций и ее остаточный член упрощаются аналогично формуле средних:

$$U_N = h\left(\frac{u_0}{2} + \sum_{n=1}^{N-1} u_n + \frac{u_N}{2}\right), \quad (3.17)$$

$$R_N = -\frac{1}{12}h^2 \sum_{n=1}^{N} h u''_{n-1/2} \approx -\frac{h^2}{12}\int_a^b u''(x)dx = \\ = -\frac{h^2}{12}[u'(b) - u'(a)] = O(h^2). \quad (3.18)$$

Формула (3.18) асимптотически точна; она получена в предположении существования непрерывной $u''(x)$. Если $u''(x)$ только кусочно-непрерывная, то можно получить лишь мажорантную оценку погрешности:

$$|R_N| \leqslant \frac{1}{12}M_2 h^2, \quad M_2 = \max_{x\in[a,b]} |u''(x)|.$$

Если же $u(x)$ имеет много непрерывных производных, то погрешность представляется суммой членов $O(h^{2q})$, $q = 1, 2, \ldots$

Точность формул трапеций и средних не слишком высока, однако из-за своей простоты ими часто пользуются в численных расчетах.

Пример 3.2. Рассмотрим тест (3.14) и проведем вычисления по формуле трапеций. Результаты представлены в табл. 3.1. Качественное поведение погрешности такое же, как и для формулы средних. Сама погрешность имеет другой знак и при одинаковом шаге примерно вдвое больше, чем для формулы средних.

3.1.4. Формула Симпсона

Построим более точную квадратурную формулу, предполагая существование непрерывной $u''(x)$. Заметим, что на произвольной сетке остаточные члены формулы средних (3.11) и формулы трапеций (3.16) имеют одинаковый вид и различаются

лишь численным коэффициентом. Умножим формулу средних (3.8) на постоянный коэффициент 2/3, формулу трапеций — на постоянный коэффициент 1/3 и сложим их. Тогда главные члены погрешности точно компенсируются. При этом получается более точная формула Симпсона:

$$U_N = \sum_{n=1}^{N} \frac{h_n}{6}[u(x_{n-1}) + 4u(x_{n-1/2}) + u(x_n)]. \tag{3.19}$$

Она записана для произвольной неравномерной сетки. Ее вид также симметричен.

Оценим погрешность формулы (3.19) в предположении существования непрерывной $u^{(IV)}(x)$. Из симметрии формулы видно, что в погрешность должны входить только четные производные. Но член с $u''(x)$ уже исключен. Таким образом, с учетом требований размерности главный член погрешности имеет вид $r_n = cu^{(IV)}h^5$. Для нахождения константы c вновь возьмем один интервал $x_{n-1} = 0$, $x_n = 1$ и простейшую функцию с ненулевой четвертой производной: $u(x) = x^4$. Вычисления дают: $U = 1/5$, $U_N = 5/24$, $R_N = 24c$, откуда $c = -1/2\,880$ и главный член погрешности формулы Симпсона равен

$$R_N = -\frac{1}{2\,880} \sum_{n=1}^{N} u^{(IV)}_{n-1/2} h_n^5. \tag{3.20}$$

На равномерной сетке формула Симпсона и выражение для главного члена погрешности упрощаются:

$$U_N = \frac{h}{6}(u_0 + 4u_{1/2} + 2u_1 + 4u_{3/2} + 2u_2 + \ldots + \\ +2u_{N-1} + 4u_{N-1/2} + u_N); \tag{3.21}$$

$$R_N = -\frac{1}{2\,880} h^4 [u'''(b) - u'''(a)]. \tag{3.22}$$

Это также асимптотически точная оценка погрешности. Если $u^{(IV)}(x)$ лишь кусочно-непрерывная, эта оценка заменяется на мажорантную:

$$|R_N| \leqslant \frac{1}{2\,880} h^4 M_4, \quad M_4 = \max_{x\in[a,b]} |u^{(IV)}(x)|.$$

Наоборот, если $u(x)$ имеет много непрерывных производных, то погрешность формулы Симпсона разлагается в сумму по четным степеням шага сетки: $h^{(2q)}$, $q = 2, 3, \ldots$

Формула Симпсона применяется, когда точность формул трапеций и средних оказывается недостаточной.

Пример 3.3. Снова возьмем тест (3.14). Результаты вычисления по формуле Симпсона представлены в табл. 3.1. Видно, что точность существенно увеличилась по сравнению с формулами трапеций и средних. Фактические погрешности при сгущении сетки убывают гораздо быстрее. Отношение соседних погрешностей должно теперь стремиться к $2^4 = 16$. Однако близость к этому пределу наблюдается лишь при довольно мелких шагах.

3.1.5. Формулы Эйлера — Маклорена

Формулы Эйлера — Маклорена строят только на равномерных сетках. Их выводят как уточнение формулы трапеций или формулы средних.

Уточнение формулы трапеций. Пусть $u(x)$ имеет вторую непрерывную производную. Рассмотрим формулу трапеций и главный остаточный член ее погрешности на равномерной сетке при $h = \text{const}$ (3.17) и (3.18). Выражение для остаточного члена является асимптотически точным, поэтому его можно прибавить к формуле трапеций как поправку. Получается формула Эйлера:

$$U_N = h\left(\frac{u_0}{2} + \sum_{n=1}^{N-1} u_n + \frac{u_N}{2}\right) - \frac{h^2}{12}(u'_N - u'_0). \tag{3.23}$$

Очевидно, эта формула должна быть существенно точнее формулы трапеций. Но какова ее погрешность?

Проведем оценку погрешности, требуя существования непрерывной $u^{(IV)}(x)$. Используем описанный ранее стандартный прием. Мы явно учли главный член погрешности, пропорциональный $\sum u''h^3$. Из соображений симметрии нечетные производные в разложении погрешности выпадают. Неучтенный член пропорционален $\sum u^{(IV)}h^5$. Заменив эту сумму интегралом и взяв интеграл явно, получаем главный член погрешности формулы Эйлера (3.23): $R_N = Ch^4(u'''_N - u'''_0)$. Для определения константы C возьмем тестовую функцию $u(x) = x^4$ на одном интервале $x_0 = 0$, $x_1 = 1$ ($N = 1$). Тогда легко вычисляем $U = 1/5$, $U_N = 1/6$, $u'''_N - u'''_0 = 24$. Отсюда $C = 1/720$ и остаточный член формулы Эйлера равен

$$R_N = \frac{h^4}{720}(u'''_N - u'''_0). \tag{3.24}$$

Таблица 3.2

Коэффициенты в формуле Эйлера — Маклорена (3.25)

m	1	2	3	4	5	6
c_m	2	1	$\frac{4}{3}$	3	10	$\frac{691}{15}$

Таким образом, формула Эйлера имеет четвертый порядок точности, как и формула Симпсона (3.19). Коэффициент в остаточном члене формулы Эйлера в 4 раза больше, чем для формулы Симпсона, зато формула Эйлера требует вдвое меньшего числа расчетных точек (формула Симпсона включает полуцелые узлы).

Обобщение. Остаточный член формулы Эйлера (3.24) также можно учесть как поправку, повысив порядок точности формулы до шестого. Если $u(x)$ имеет достаточно много непрерывных производных, можно провести рекуррентное повышение точности, каждый раз вычисляя главный член погрешности очередной формулы и вводя его как новую поправку. Таким образом получают формулы Эйлера — Маклорена:

$$U_N = h\left(\frac{u_0}{2} + \sum_{n=1}^{N-1} u_n + \frac{u_N}{2}\right) + \\ + \sum_{m=1}^{M} \frac{(-1)^m c_m h^{2m}}{(2m+2)!}\left(u_N^{(2m-1)} - u_0^{(2m-1)}\right), \tag{3.25}$$

где значения первых коэффициентов c_m приведены в табл. 3.2.

Число слагаемых в сумме по m (3.25) определяется гладкостью подынтегральной функции. Если существует непрерывная $u^{(2M)}(x)$, то в сумме можно брать M слагаемых.

Уточнение формул средних. Процедура рекуррентного введения поправок в формулу средних приводит к следующему выражению:

$$U_N = h\sum_{n=1}^{N} u_{n-1/2} - \\ - \sum_{m=1}^{M} \left(1 - 2^{-(2m-1)}\right)\frac{(-1)^m c_m h^{2m}}{(2m+2)!}\left(u_N^{(2m-1)} - u_0^{(2m-1)}\right). \tag{3.26}$$

Здесь коэффициенты c_m те же, что и в (3.25); они приведены в табл. 3.2. Видно, что полный коэффициент в первой поправке

при h^2 для формулы средних вдвое меньше, чем в формуле трапеций. Но для следующих поправок эти коэффициенты быстро сближаются: $(1 - 2^{-(2m-1)}) \to 1$, оставаясь противоположных знаков.

Пример 3.4. Снова используем тест (3.14). Расчеты проводились для простейшей формулы Эйлера (3.23) и формулы Эйлера—Маклорена (3.25) с $M = 2$. Результаты приведены в табл. 3.1.

Видно, что формула Эйлера существенно точнее формулы трапеций и примерно вчетверо точнее формулы Симпсона при одинаковой трудоемкости (формула Симпсона содержит полуцелые узлы, поэтому она эквивалентна по трудоемкости формуле Эйлера с числом узлов $2N$).

Формулы Эйлера—Маклорена имеют еще более высокую точность, хотя по трудоемкости почти не отличаются от формулы трапеций. Однако выход на теоретический порядок точности в них происходит заметно медленнее (отношение двух соседних погрешностей должно стремиться к 64). Кроме того, при очень малом числе узлов погрешность может оказаться даже больше, чем у формулы трапеций.

Таким образом, введение в формулу трапеций небольших краевых поправок существенно повышает точность. Однако не следует применять эти формулы при абсурдно малом числе узлов (например, при $N = 1$): их преимущества проявляются при достаточно малых шагах.

Интегрирование периодических функций. Пусть $u(x)$ — достаточно гладкая периодическая функция, а интеграл берется по полному периоду $[a, b]$. Тогда в силу периодичности $u_N^{(2m-1)} = u_0^{(2m-1)}$ и поправочные члены в суммах обеих формул Эйлера—Маклорена (3.25), (3.26) обратятся в нуль. Это означает, что для интегралов по полному периоду формулы трапеций и средних на равномерной сетке имеют точность не $O(h^2)$, а выше.

Если $u(x)$ имеет $2M$ непрерывных производных, то сокращаются члены погрешности, включая $O(h^{2M})$; тогда остаточный член есть $o(h^{2M})$. Если же $u(x)$ имеет неограниченное число непрерывных производных, то в погрешности исчезают все степенные члены. В этом случае погрешность экспоненциально убывает с ростом числа узлов N.

Суммирование рядов. Случаев, когда можно точно вычислить бесконечную сумму, исключительно мало. Об этом свидетельствует хотя бы тот факт, что несколько сотен страниц спра-

вочника И. С. Градштейна и И. М. Рыжика посвящено интегралам и лишь несколько страниц — рядам.

Приведем пример приближенного вычисления суммы ряда

$$S_N = \sum_{n=N}^{\infty} \frac{1}{n^k}.$$

Точно сумма этого ряда вычисляется только для некоторых целочисленных k. Однако ряд можно рассматривать как квадратурную формулу средних для интегрирования $u(n) = n^{-k}$ с шагом $\Delta n = 1$ в пределах от $N - 1/2$ до ∞:

$$S_N \equiv U_N \approx U = \\ = \int_{N-1/2}^{\infty} \frac{dn}{n^k} = \frac{1}{(k-1)(N-1/2)^{k-1}}, \quad k > 1. \tag{3.27}$$

Интеграл точно берется при любом $k > 1$ (не обязательно целом), а полученный ответ отличается от интегральной суммы на $O(N^{-2})$. При достаточно больших N этот ответ является неплохим приближением.

Этот результат можно улучшить. Применим для вычисления (3.27) формулу средних с поправками Эйлера — Маклорена:

$$\int_{N-1/2}^{\infty} \frac{dn}{n^k} = \sum_{n=N}^{\infty} \frac{1}{n^k} + R_1 + R_2 + \dots \tag{3.28}$$

Учитывая при вычислении поправок Эйлера — Маклорена, что

$$u'_\infty = u'''_\infty = \dots = 0,$$

и вычисляя явно $u'_{N-1/2}, u'''_{N-1/2}, \dots$, получим

$$S_N \equiv \sum_{n=N}^{\infty} \frac{1}{n^k} = \frac{1}{(k-1)(N-1/2)^{k-1}} - \\ - \frac{1}{24} \frac{k}{(N-1/2)^{k+1}} + \frac{7}{5\,760} \frac{k(k+1)(k+2)}{(N-1/2)^{k+3}} - \dots \tag{3.29}$$

Выписанные члены обеспечивают точность $O(N^{-6})$. Уже при $N = 5$ это дает примерно 6 верных десятичных знаков.

Если исходное N невелико, точность формул (3.27) — (3.29) может оказаться недостаточной. Тогда первые несколько членов ряда суммируют явно, а оставшуюся часть заменяют интегралом.

Описанный прием используется достаточно часто. Если требуемый несобственный интеграл точно не вычисляется, то его приближенно вычисляют с помощью формул Гаусса — Кристоффеля (см. подразд. 3.1.6). В результате вместо ряда получается легко вычисляемая сумма небольшого количества слагаемых.

3.1.6. Формулы Гаусса — Кристоффеля

Формулы Гаусса — Кристоффеля называют также формулами наивысшей алгебраической точности. Рассмотрим интеграл с весовой функцией

$$U = \int_a^b u(x)\rho(x)dx, \quad \rho(x) > 0, \quad x \in (a, b).$$

От весовой функции требуется положительность, интегрируемость и непрерывность в открытом интервале (a, b). Вообще говоря, в граничных точках интервала весовая функция может обращаться в нуль или бесконечность.

Рассмотрим квадратурную формулу

$$U \approx U_N = \sum_{n=1}^{N} c_n u(x_n). \tag{3.30}$$

Значения c_n называют ***весами***, а x_n — ***узлами*** квадратурной формулы. Ранее в интегральных суммах в качестве c_n выбирались длины интервалов сетки, а x_n лежали внутри этих интервалов. Теперь будем варьировать c_n и x_n так, чтобы получить как можно более точную формулу. Число варьируемых параметров равно $2N$. Столько же коэффициентов содержит многочлен $Q_{2N-1}(x)$ степени $2N-1$. Подберем параметры так, чтобы квадратурная формула (3.30) была бы точна для любого полинома степени $2N-1$ (тем самым и для любого многочлена меньшей степени).

Для единичного веса $\rho(x) \equiv 1$ такие квадратурные формулы впервые получил Гаусс, а для произвольного веса — Кристоффель.

Из теории классических ортогональных полиномов известно, что для положительного, непрерывного и интегрируемого на (a,b) веса $\rho(x)$ существует система полиномов $\{P_k(x)\}$ ортогональных на $[a,b]$ с весом $\rho(x)$:

$$\int_a^b P_k(x)P_m(x)\rho(x)dx = \delta_{km}\|P_k(x)\|^2. \tag{3.31}$$

Все k нулей полинома $\{P_k(x)\}$ вещественны и лежат внутри $[a,b]$.

Теорема 3.1. Узлами формулы Гаусса — Кристоффеля (3.30) являются нули ортогонального многочлена $P_N(x)$ степени N системы (3.31).

Доказательство. Пусть x_n — нули квадратурной формулы Гаусса — Кристоффеля. Построим многочлен

$$\omega_N(x) = \prod_{n=1}^{N}(x - x_n). \tag{3.32}$$

Возьмем в качестве пробной функции $u(x) = P_m(x)\omega_N(x)$, $m \leqslant \leqslant N-1$. Эта функция есть многочлен степени не более $2N-1$. Для него формула Гаусса — Кристоффеля точна по определению, поэтому равенство

$$\int_a^b \omega_N(x)P_m(x)\rho(x)dx = \sum_{n=1}^{N} c_n\omega_N(x_n)P_m(x_n), \quad m \leqslant N-1 \tag{3.33}$$

будет точным. Но $\omega_N(x_n) = 0$, следовательно, правая часть (3.33) обращается в нуль. Это означает, что многочлен $\omega_N(x)$ ортогонален всем многочленам $P_m(x)$ системы (3.31) степени $m \leqslant N-1$. Значит, он с точностью до множителя совпадает с $P_N(x)$, а его нули являются нулями этого многочлена. ■

Вычисление весов. Будем выбирать в качестве подынтегральной функции следующие многочлены степени $N-1$:

$$u(x) = \psi_n(x) = \prod_{m=1, m\neq n}^{N} \frac{x - x_m}{x_n - x_m},$$

где точки x_n, x_m — узлы полинома $P_N(x)$.

Полиномы $\psi_n(x)$ имеют следующее свойство:

$$\psi_n(x_k) = \begin{cases} 0 \text{ при } k \neq n, \\ 0 \text{ при } k = n. \end{cases}$$

Для всех этих функций согласно определению квадратурная формула Гаусса — Кристоффеля точна. Это дает формулу для вычисления веса:

$$\int_a^b \psi_n(x)\rho(x)dx = \sum_{k=1}^{N} c_k \psi_n(x_k) = c_n. \tag{3.34}$$

Для некоторых весовых функций $\rho(x)$ (например, $\rho \equiv 1$) интегралы (3.34) удается вычислить точно через узлы x_n. В более сложных случаях они находятся прецизионным численным интегрированием. В литературе по ортогональным многочленам вычислены и приведены таблицы узлов и весов формулы (3.30) для наиболее важных весовых функций.

Из формулы (3.34) знак веса c_n не виден. Докажем, что все веса положительны. Для этого положим $u(x) = \psi_m^2(x)$. Это многочлен степени $2N-2$, для него формула Гаусса — Кристоффеля точна:

$$\int_a^b \psi_m^2(x)\rho(x)dx = \sum_{k=1}^{N} c_k \psi_m^2(x_k) = c_m.$$

Но под интегралом стоит положительная функция, так что $c_m > 0$, что и требовалось доказать.

Отметим еще одно свойство коэффициентов. Положим $u(x) \equiv 1$. Для нее формула (3.30) точна. Это дает

$$\int_a^b \rho(x)dx = \sum_{n=1}^{N} c_n.$$

Полученное равенство можно использовать для проверки точности вычисления c_n из (3.34).

Формула (3.30) точна для многочлена степени $2N-1$. Значит, ее погрешность не может содержать производных $u^{(m)}(x)$ с $m \leqslant 2N-1$. Тогда главный член погрешности будет содержать $u^{(2N)}(x)$. Из соображений размерности мажорантная оценка погрешности будет иметь вид

$$|R_N| \leqslant C_N M_{2N} (b-a)^{2N+1}; \quad M_{2N} = \max_{x \in [a,b]} |u^{(2N)}(x)|. \tag{3.35}$$

В оценку входит $b-a$ вместо привычного h, так как для формул Гаусса — Кристоффеля понятие шага отсутствует и есть только длина интервала. Из оценки (3.35) видно, что формулы Гаусса — Кристоффеля с N узлами следует применять только к функциям, имеющим не менее $2N$ непрерывных производных. Высокая точность формул Гаусса — Кристоффеля обеспечивается тем, что безразмерный множитель C_N очень быстро убывает с ростом N. Заметим, что аналогично мажорантной оценке (3.35) можно построить асимптотически точную оценку погрешности.

Важно помнить следующее. Для реализации высокой точности формул Гаусса — Кристоффеля необходимо задавать узлы и веса с максимально возможной точностью (с использованием всех доступных разрядов вычислительной техники). Поэтому особенно удобны для практики те частные случаи, когда известны явные выражения x_n, c_n через дроби или радикалы.

Частные случаи. Вычисление узлов и весов для произвольного $\rho(x)$ очень сложно. На практике в основном употребляют классические ортогональные многочлены, для которых эти величины давно найдены. Это пять случаев, перечисленных в табл. 3.3.

Узлы и веса формул Гаусса — Кристоффеля приведены в табл. 3.4. Первые три случая относятся к стандартному конечному отрезку интегрирования $[-1,1]$. Для интегрирования по произвольному отрезку $[a,b]$ можно провести линейное преобразование аргумента. Однако вместо этого можно сразу же написать x_n и c_n для произвольного отрезка $[a,b]$ через узлы ξ_n и веса γ_n стандартного отрезка $[-1,1]$:

$$c_n = \frac{b-a}{2}\gamma_n; \quad x_n = \frac{b+a}{2} + \frac{b-a}{2}\xi_n, \tag{3.36}$$

где γ_n, ξ_n — табличные веса и узлы на стандартном отрезке.

Отметим формулу, основанную на многочленах Чебышева I рода. В ней известны явные выражения для всех узлов и весов, что позволяет использовать ее для вычислений с большими значениями N. Она имеет вид

$$\int_{-1}^{1} \frac{u(x)dx}{\sqrt{1-x^2}} \approx \frac{\pi}{N}\sum_{n=1}^{N} u(x_n), \quad x_n = \cos\frac{\pi(n-0.5)}{N}, \tag{3.37}$$

и называется формулой Эрмита. Эту формулу используют при среднеквадратичной аппроксимации функций многочленами Чебышева.

Таблица 3.3

Сводная таблица ортогональных полиномов

Характеристики	Многочлен				
	Лежандра	Чебышева I рода	Чебышева II рода	Лагерра	Эрмита
Обозначение	$P_n(x)$	$T_n(x)$	$U_n(x)$	$L_n(x)$	$H_n(x)$
Отрезок	$[-1,1]$	$[-1,1]$	$[-1,1]$	$[0,+\infty)$	$(-\infty,+\infty)$
Вес	1	$\dfrac{1}{\sqrt{1-x^2}}$	$\sqrt{1-x^2}$	e^{-x}	e^{-x^2}

Формула, основанная на многочленах Чебышева II рода, полезна для вычислений двумерных интегралов по области с криволинейной границей.

Формулу, основанную на многочленах Лаггера, используют для вычисления несобственных интегралов на полупрямой $(0,\infty)$, если подынтегральное выражение убывает примерно как e^{-x}.

Для вычисления несобственных интегралов на прямой $(-\infty, +\infty)$ применяют формулу, основанную на полиномах Эрмита; подынтегральное выражение при этом должно убывать, как e^{-x^2}.

Пример 3.5. В табл. 3.1 приведены результаты расчета теста (3.14) по формуле Гаусса ($\rho(x) \equiv 1$). Они выполнены с учетом преобразования отрезка (3.36). Для $N = 1$ формула Гаусса совпадает с формулой средних. Но при дальнейшем увеличении N точность довольно быстро возрастает. При $N = 4$ она выше, чем у любой другой формулы из табл. 3.1, однако это превышение точности не столь значительно, как можно ожидать от формул наивысшей алгебраической точности. Объяснение в том, что в оценку погрешности (3.35) входит не шаг h (которого в этих формулах нет), а отнюдь не малая длина отрезка $b - a$. Использовать же большие N в этих формулах достаточно сложно, поэтому формулы Гаусса — Кристоффеля в чистом виде не часто применяются на практике.

Пример 3.6. Недостаточная точность расчета в примере 3.5 связана также с тем, что тестовая функция (3.14) имеет особенность в точке $x = 0$, расположенной довольно близко к отрезку интегрирования. Поэтому высокие производные $u(x)$, входящие в оценку погрешности, оказываются большими. Рассмотрим другой пример, где функция достаточно гладкая, т. е. ее высокие производные невелики. Вычислим интеграл

$$U = \int_{-1}^{1} \frac{u(x)dx}{\sqrt{1-x^2}}, \quad u(x) = e^x \tag{3.38}$$

по квадратурной формуле Эрмита (3.37). Результаты расчета для разных N приведены в табл. 3.5. Видно стремительное уве-

Таблица 3.4

Ортогональные многочлены

N	Многочлены Лежандра $P_n(x)$	
1	$x_1 = 0$	$c_1 = 2$
2	$x_{1,2} = \mp\sqrt{1/3}$	$c_{1,2} = 1$
3	$x_{1,3} = \mp\sqrt{3/5}, x_2 = 0$	$c_{1,3} = 5/9, c_2 = 8/9$
4	$x_{1,4} = \mp\sqrt{(15+2\sqrt{30})/35};$	$c_{1,4} = (18-\sqrt{30})/36;$
	$x_{2,3} = \mp\sqrt{(15-2\sqrt{30})/30}$	$c_{2,3} = (18+\sqrt{30})/36$
5	$x_{1,5} = \mp\sqrt{(35+2\sqrt{70})/63};$	$c_{1,5} = (322-13\sqrt{70})/900;$
	$x_{2,4} = \mp\sqrt{(35-2\sqrt{70})/63};$	$c_{2,4} = (322+13\sqrt{70})/900;$
	$x_3 = 0$	$c_3 = 128/225$

N	Многочлены Чебышева I рода $T_n(x)$	
n	$x_m^{(n)} = \cos\dfrac{\pi(2m-1)}{2n}$	$c_m^{(n)} = \pi/n, 1 \leqslant m \leqslant n$

N	Многочлены Чебышева II рода $U_n(x)$	
1	$x_m^{(n)} = \cos\dfrac{\pi m}{n+1}, 1 \leqslant m \leqslant n$	$c_1 = \pi/2$
2		$c_1 = c_2 = \pi/4$
3		$c_1 = c_3 = \pi/8, c_2 = \pi/4$
4		$c_1 = c_4 = \pi(5-\sqrt{5})/40;$ $c_2 = c_3 = \pi(5+\sqrt{5})/40$
5		$c_1 = c_5 = \pi/24;$ $c_2 = c_4 = \pi/8; c_3 = \pi/6$

N	Многочлены Лагерра $L_n(x)$	
1	$x_1 = 1$	$c_1 = 1$
2	$x_{1,2} = 2 \mp \sqrt{2}$	$c_{1,2} = (2 \pm \sqrt{2})/4$

Окончание табл. 3.4

N	Многочлены Эрмита $H_n(x)$	
1	$x_1 = 0$	$c_1 = \sqrt{\pi}$
2	$x_{1,2} = \mp\sqrt{1/2}$	$c_{1/2} = \sqrt{\pi}/2$
3	$x_{1,3} = \mp\sqrt{3/2}, x_2 = 0$	$c_{1,3} = \sqrt{\pi}/6; c_2 = 2\sqrt{\pi}/3$
4	$x_{1,4} = \mp\sqrt{(3+\sqrt{6})/2};$	$c_{1,4} = \sqrt{\pi}(3-\sqrt{6})/12;$
	$x_{2,3} = \mp\sqrt{(3-\sqrt{6})/2}$	$c_{2,3} = \sqrt{\pi}(3+\sqrt{6})/12$
5	$x_{1,5} = \mp\sqrt{(5+\sqrt{10})/2};$	$c_{1,5} = \sqrt{\pi}(7-2\sqrt{10})/60;$
	$x_{2,4} = \mp\sqrt{(5-\sqrt{10})/2};$	$c_{2,4} = \sqrt{\pi}(7+2\sqrt{10})/60;$
	$x_3 = 0$	$c_3 = 8\sqrt{\pi}/15$

Таблица 3.5

Вычисление интеграла (3.38) по формуле Эрмита (3.37)

N	U_N	N	U_N	N	U_N
1	3,14159265	3	3,97732194	5	3,97746325
2	3,96026605	4	3,97746262	6	3,97746326

личение точности при возрастании N. Уже $N = 5$ дает практически 9 верных знаков.

Таким образом, формулы Гаусса — Кристоффеля в чистом виде могут оказаться очень выгодными для функции высокой гладкости.

Гауссово-сеточный метод. Погрешность сеточных методов имеет вид $O(h^p) = O(N^{-p})$. В эту оценку входит величина производной $u^{(p)}(x)$, порядок которой p фиксирован, т. е. не меняется при $N \to \infty$. Это обеспечивает сходимость при $h \to 0$, т. е. при $N \to \infty$. В оценку погрешности формул Гаусса — Кристоффеля входят C_N и M_{2N}. При $N \to \infty$ величины $C_N \to 0$. Но поведение M_{2N} при этом неизвестно. Высокие производные $u(x)$ могут оказаться большими, поэтому трудно определить, какова будет скорость сходимости формул Гаусса — Кристоффеля при $N \to \infty$, и будет ли вообще сходимость.

В связи с этим нередко используют сочетание формул Гаусса — Кристоффеля с сеточными методами. На отрезке $[a, b]$ вводят вспомогательную сетку

$$\Omega_K = \{a = X_0 < X_1 < \ldots < X_K = b\}. \qquad (3.39)$$

Затем разбивают интеграл на сумму интегралов по отрезкам $[X_{k-1}, X_k]$ и на каждом отрезке используют формулы Гаусса — Кристоффеля с одним и тем же N. Примем для простоты, что сетка Ω_K равномерна: $h \equiv X_k - X_{k-1} = (b-a)/K$. Тогда в погрешности (3.35) по отрезкам надо подставить h вместо $b-a$. Суммирование этих погрешностей дает оценку

$$|R_N| \leqslant C_N(b-a)h^{2N}M_{2N}, \quad h = \frac{b-a}{K}. \tag{3.40}$$

Оценка погрешности (3.40) обеспечивает сходимость при $K \to \infty$ и фиксированном N. Здесь также можно построить не мажорантную, а асимптотически точную оценку погрешности.

Пример такого расчета для теста (3.14) приведен в табл. 3.1. Введена сетка всего из двух равных интервалов ($k = 2$) и выполнен расчет с $N = 4$. Полное число узлов равно 8. Видно, что точность при этом существенно повышается и становится лучше, чем у формулы Эйлера — Маклорена точности $O(h^6)$ с $N = 8$.

Этот вариант более пригоден для практического использования, чем формулы Гаусса — Кристоффеля в чистом виде.

3.1.7. Недостаточно гладкие функции

Все описанные квадратурные формулы были рассчитаны на подынтегральные функции, имеющие определенное число p непрерывных производных. Это обеспечивало погрешность $O(h^p)$. Если число непрерывных производных у функции было меньше, то это приводило к понижению порядка точности. При этом обычно существование лишь q-й непрерывной производной ($q < p$) понижает точность до $O(h^q)$. Например, если существует лишь кусочно-непрерывная $u'(x)$, то справедлива только следующая мажорантная оценка погрешности:

$$|R| \leqslant c(b-a)hM_1, \quad M_1 = \max|u'(x)|;$$

здесь $c = 1/4$ для формул трапеций и средних; $c = 5/18$ для формулы Симпсона и $c = 0{,}276$ для формулы Гаусса с $N = 4$. Поэтому формулы высокого порядка точности целесообразно использовать лишь для достаточно гладких функций.

Напротив, наличие у функции производной выше p-й не повышает порядка точности формул. Это свойство формул называют ***насыщением***. Аналогичное свойство имеется не только у квадратурных формул, но и у многих других сеточных методов.

Для случая непрерывной p-й производной мы строили асимптотически точную оценку для погрешности. Такая оценка не улучшаема; на ее основе можно построить алгоритмы расчетов с контролем точности (см. 3.2). Если же $u^{(p)}(x)$ лишь кусочно-непрерывна, то оценка погрешности становится мажорантной. Она обычно является сильно завышенной и менее удобна для практических приложений.

Однако в прикладных расчетах не часто возникают функции, у которых вообще отсутствуют высокие производные. Обычно функция и все ее производные существуют и непрерывны на всем отрезке, кроме отдельных точек. В этих точках либо сама функция, либо какие-то ее производные имеют разрывы или обращаются в бесконечность. В этом случае сходимость квадратурных формул можно существенно улучшить, выбирая специальные сетки.

Обозначим особые точки $u(x)$ (разрывы функции либо ее производных) через X_k. И причислим также к особым точкам концы отрезка. Введем специальную сетку Ω_K (3.39). Тогда внутри каждого отрезка $[X_{k-1}, X_k]$ функция $u(x)$ будет непрерывной вместе со всеми своими производными, а интеграл на этом отрезке можно будет вычислять по любой квадратурной формуле со своими узлами x_n, не опасаясь потери точности. Полный же интеграл равен сумме интегралов по всем отрезкам $[X_{k-1}, X_k]$.

Эту процедуру можно описать иначе. Совокупность сеток $\{x_n\}$ на всех отрезках $[X_{k-1}, X_k]$ составляет некоторую сетку на полном отрезке $[a, b]$ (даже если на каждом отрезке были построены равномерные сетки, суммарная сетка необязательно будет равномерной). При этом все особые точки X_k оказываются некоторыми узлами суммарной сетки $\{x_n\}$. Такие суммарные сетки называют ***специальными***. Это формулируют в виде правила:

если $u(x)$ или ее производная имеет особенность, поставь в эту точку узел сетки.

3.2. МЕТОД СГУЩЕНИЯ СЕТОК

3.2.1. Однократное сгущение

Сеточные методы особенно ценны для практических вычислений, потому что они позволяют не только получить ответ, но и оценить его точность.

Историческая справка. Первую такую оценку дал К. Рунге (1895) для численного интегрирования обыкновенных диффе-

ренциальных уравнений по своей новой схеме. Рекомендовалось провести два расчета: первый — на сетке с шагом h, второй — с шагом $h/2$. Оба расчета сравнивались, и совпадающие знаки считались верными.

Л. Ричардсон, занимавшийся уравнениями в частных производных, усовершенствовал этот способ (1910), но полное объяснение дал лишь в 1927 г. Он показал, как по двум расчетам с шагами h и $h/2$ по схеме p-го порядка точности можно получить хорошую оценку погрешности, а также повысить точность результата. Эти способы не требовали знания производных $u(x)$ и давали апостериорную оценку.

Эти методы применимы практически к любым задачам с сеточным представлением функции — к интерполяции, численному дифференцированию, численному интегрированию, решению задач Коши и краевых задач для обыкновенных дифференциальных уравнений, задач для уравнений в частных производных, для интегральных и интегродифференциальных уравнений. Более того, до сих пор это единственный способ получить для сеточных решений асимптотически точную оценку погрешности. Они служат надежной основой для построения программ с контролем точности. Далее будет показано, как это делается.

Оценка погрешности. Сначала будут рассмотрены равномерные сетки $h = (b-a)/N$. В 3.1 было построено много квадратурных формул и выведены априорные оценки главных членов их погрешностей. Все их можно записать в символическом виде

$$U = U_N + \bar{c}h^p + o(h^p) \equiv U_N + cN^{-p} + o(N^{-p}), \qquad (3.41)$$

где p — порядок точности формулы; c — постоянный коэффициент, являющийся некоторой комбинацией производных подынтегральной функции и не зависящий от шага сетки. Когда $N \to \infty$ и $h \to 0$, истинная погрешность очень близка к записанному здесь главному члену.

Проведем расчеты на двух сетках: первый — с числом интервалов N, второй — с числом интервалов rN (число r может быть нецелым). Для определенности полагаем $r > 1$, т. е. считаем вторую сетку более подробной. Формулу (3.41) можно записать в следующем виде:

$$\begin{aligned} U &= U_N + R_N + o(N^{-p}), \quad R_N = cN^{-p}; \\ U &= U_{rN} + R_{rN} + o(N^{-p}), \quad R_{rN} = cr^{-p}N^{-p}. \end{aligned} \qquad (3.42)$$

Здесь в обеих строках написаны одинаковые $o(N^{-p})$, поскольку это величины исчезающе малые по сравнению с главными чле-

нами погрешностей. Коэффициент c одинаков в обеих строках, поскольку он не зависит от шага сетки.

Величина точного значения U и коэффициент c в (3.42) нам неизвестны, однако из постоянства c следует, что $R_N = r^p R_{rN}$. Учитывая это и вычитая верхнюю строку (3.42) из нижней, получим первую формулу Ричардсона:

$$R_{rN} = \frac{U_{rN} - U_N}{r^p - 1} + o(N^{-p}). \tag{3.43}$$

Эта формула дает оценку погрешности на ***подробной*** сетке на основании результатов расчетов на двух сетках. Эта оценка не требует каких-либо сведений о подынтегральной функции и находится в ходе расчетов; такие оценки называют ***апостериорными.***

На достаточно подробных сетках член $o(N^{-p})$ пренебрежимо мал, поэтому оценка (3.43) ***асимптотически*** точная. Это означает, что на достаточно подробных сетках найденная оценка будет очень близка к истинной погрешности. Поэтому оценку по первой формуле Ричардсона можно выдавать в программах как гарантированную оценку погрешности для приближенного решения U_{rN}, полученного на подробной сетке.

Однако напомним, что все эти оценки справедливы лишь для достаточно гладкой подынтегральной функции. В задачах численного интегрирования для этого требовалось наличие непрерывной ограниченной $u^{(p)}(x)$. В других задачах, которые будут рассмотрены далее, требования к функции могут быть иными. Применение описанных оценок к недостаточно гладким функциям может привести к серьезным ошибкам.

Повышение точности. Поскольку оценка (3.43) асимптотически точна, ее можно учесть как поправку к приближенному решению U_{rN}. Подставив (3.43) во вторую строку (3.42), получим вторую формулу Ричардсона:

$$U = \tilde{U}_{rN} + o(N^{-p}); \quad \tilde{U}_{rN} = U_{rN} + \frac{U_{rN} - U_N}{r^p - 1}. \tag{3.44}$$

Эта формула позволяет по расчетам p-го порядка точности, выполненным на двух равномерных сетках с разным числом интервалов, получить результат $\tilde{U}_{rN}$ порядка точности выше p-го (какой именно порядок будет получен, рассмотрим в подразд. 3.2.2).

Из расчетов на двух сетках нельзя оценить погрешность уточненного значения $\tilde{U}_{rN}$. Однако в простейших программах

действует следующее нестрогое рассуждение. Предполагают, что погрешность уточненного решения $\tilde{U}_{rN}$ существенно меньше, чем погрешность неуточненного решения U_{rN} (в большинстве практических ситуаций это предположение выполняется). Выполняют расчеты на двух сетках, уточненное значение $\tilde{U}_{rN}$ (3.44) берут в качестве ответа, а величину R_{rN} (3.43) приводят как оценку погрешности уточненного решения. Разумеется, эта оценка уже не является асимптотически точной. Ее стоит рассматривать как подобие мажорантной оценки.

Многократное сгущение. Для практики нужны программы расчетов, обеспечивающие заданную точность ε. Простейшие программы целесообразно организовывать следующим образом. Выберем первоначально сетку с N_0 интервалами. Сгустим ее в r раз, получив сетку с числом узлов $N_1 = rN_0$, проведем расчет и оценим погрешность $R_{N_1} \equiv R_1$ (мы изменили обозначения). Если $|R_1| < \varepsilon$, то требуемая точность расчета достигнута и вычисления прекращаются. Если точность ε не достигнута, то снова сгустим сетку в r раз, получив сетку $N_2 = r^2N_0$. Теперь значения $U_{N_1} \equiv U_1$ и $U_{N_2} \equiv U_2$ на двух последних сетках надо рассматривать как расчеты на двух равномерных стеках с отношением шагов r и применять обе формулы Ричардсона. Такое сгущение сеток в r раз продолжают до тех пор, пока не будет достигнута требуемая точность: $|R_k| < \varepsilon$.

Эту процедуру важно дополнить следующим контролем. Метод Ричардсона применим, если можно пренебречь членом $o(N^{-p})$. Однако из табл. 3.1 видно, что это справедливо лишь на достаточно подробных сетках: отношения полных погрешностей на соседних сетках при малых N еще далеки от теоретического значения r^p. Поэтому в программе следует сравнивать фактические отношения соседних погрешностей R_{k-1}/R_k и оценивать их близость к теоретическому пределу. Особенно это удобно делать, вводя понятие ***эффективного порядка точности.*** Назовем p_k эффективным порядком точности, если выполняется $R_{k-1}/R_k = r^{p_k}$ или

$$p_k = \frac{\lg(R_{k-1}/R_k)}{\lg r}.$$

Если при последовательном сгущении сеток p_k монотонно стремится к теоретическому порядку точности p (3.41), то условия применимости формул Ричардсона выполнены (тогда описанная ранее процедура сгущения сетки работает правильно).

В программе необходима страховка от нештатных ситуаций. На очень грубых сетках поведение p_k может быть немонотон-

ным. Более того, отношение R_{k-1}/R_k может стать отрицательным. Это означает, что начальное число интервалов N_0 выбрано слишком малым и его необходимо увеличить. Вместо этого достаточно просто отбросить расчеты на первых сетках.

Немонотонность p_k или отрицательность R_{k-1}/R_k может снова возникнуть на очень подробных сетках. Это свидетельствует о выходе расчета на ошибки округления, когда последние разряды чисел беспорядочно меняются. Сами значения R_k при этом уже весьма малы. Если еще не выполнен критерий остановки расчета, то был задан нереалистично высокий уровень точности ε. При этом программа должна останавливать расчет и выдавать диагностику: «заданное ε недостижимо, наилучшая достижимая точность есть $|R_k|$».

Наконец, возможно стремление p_k к другому числу q, не совпадающему с теоретическим порядком точности: $q \neq p$. Если $q < p$, это означает недостаточную гладкость подынтегральной функции. По величине q можно оценить число существующих непрерывных производных $u(x)$. Если же $q > p$, то это означает, что запрограммированная формула на данной $u(x)$ по какой-либо причине имеет более высокий порядок точности (вспомним, например, что для периодических функций формула трапеций имеет порядок точности выше 2).

Удобным визуальным контролем работы программы служат графики. На первом графике строят зависимость $\lg|R_k|$ от $\lg N_k$. Средняя часть такого графика должна быть прямой линией с наклоном $\operatorname{tg}\alpha = -p$. Это свидетельствует о пренебрежимой малости $o(N^{-p})$ и применимости формул Ричардсона. Начальный участок должен монотонно выходить на эту прямую. На нижнем участке с некоторого момента погрешность ведет себя беспорядочно и в среднем не уменьшается (выход на ошибки округления).

Второй график — зависимость p_k от $\lg N_k$. Его средний участок должен быть константой p, а качественное поведение начального и конечного отрезков графика аналогично. Примеры таких графиков будут приведены далее.

Наборы сеток. Ранее отмечалось, что r может быть не целым. Целыми должны быть числа интервалов на любых сетках: $N_0, rN_0, r^2N_0, \ldots$ В каком же отношении целесообразно сгущать сетки? В практике наиболее часто употребляется значение $r = 2$. Тогда при любом N_0 значения N_k будут целыми, а ошибки округления при вычислении R_k (3.43) окажутся небольшими. Такие сетки особенно выгодны при решении дифференциальных урав-

нений для функции $u(x)$. При очередном сгущении каждый интервал сетки делится пополам, все узлы старой сетки становятся четными узлами новой сетки, а середины интервалов старой стеки — нечетными узлами новой сетки. Это позволяет сравнивать расчетные значения функции в совпадающих узлах соседних сеток.

Однако у сетки со значением $r = 2$ есть недостатки: 1) точки на описанных ранее графиках расположены довольно редко, что затрудняет визуальный контроль точности; 2) каждое сгущение на одномерных задачах вдвое увеличивает объем расчетов, на двумерных — вчетверо, на трехмерных — в восемь раз и т. д. Многомерные задачи в настоящее время являются вполне обычными в практике, и такое повышение трудоемкости зачастую нежелательно.

Существует несколько наборов целых чисел, у которых отношение соседних чисел почти одинаково. Их удобно использовать для построения сгущающихся сеток. Будем называть такие сетки ***магическими***:

$$N_0 = 5, \quad N_1 = 7, \quad N_2 = 10, \quad N_3 = 14, \quad N_4 = 20, \quad N_5 = 28, \ldots, \quad N_{k-1}/N_k = \sqrt{2} \pm 1\ \%; \tag{3.45}$$

$$N_0 = 12, \quad N_1 = 17, \quad N_2 = 24, \quad N_3 = 34, \quad N_4 = 48, \quad N_5 = 68, \ldots, N_{k-1}/N_k = \sqrt{2} \pm 0,2\ \%; \tag{3.46}$$

$$N_0 = 12, \quad N_1 = 15, \quad N_2 = 19, \quad N_3 = 24, \quad N_4 = 30, \quad N_5 = 38, \ldots, \quad N_{k-1}/N_k = \sqrt[3]{2} \pm 0,8\ \%; \tag{3.47}$$

$$N_0 = 10, \quad N_1 = 12, \quad N_2 = 14, \quad N_3 = 17, \quad N_4 = 20, \quad N_5 = 24, \ldots, \quad N_{k-1}/N_k = \sqrt[4]{2} \pm 2,1\ \%. \tag{3.48}$$

Такие наборы нельзя начинать с произвольного значения N_0; в качестве N_0 нужно выбирать одно из чисел каждого набора, но не обязательно первое. Среднее значение r в каждом наборе есть $2^{1/m}$ $(m = 2, 3, 4)$, что позволяет ставить точки на графиках погрешности в $2-4$ раза чаще, чем при $r = 2$.

Однако для аккуратного расчета R_k и p_k нельзя пользоваться средним значением r. В формулах Ричардсона (3.43), (3.44) для каждой пары соседних сеток надо брать именно их отношение $r_k = N_k/N_{k-1}$. Использование среднего отношения может вызвать небольшую искусственную немонотонность стремления R_k и p_k к пределам. Тогда расчет может необоснованно остановиться, не достигнув заданной точности.

В многомерных задачах использование наборов (3.45)—(3.48) дает многократное уменьшение трудоемкости расчетов.

Сделаем еще замечание. Формально можно брать соседние сетки с любыми числами интервалов, например, $N_1 = 100$ и $N_2 = 101$; это соответствует $r = 1{,}01$. Однако принимать $r \approx 1$ в методе Ричардсона не следует: при вычитании близких значений U_1 и U_2 сокращается много значащих цифр и сильно возрастают ошибки округления.

Метод Ричардсона можно использовать также при отладке программ. В качестве теста берут функцию с заведомо достаточной гладкостью, интеграл от которой известен. Проводят расчеты по программе на сгущающихся сетках. Если сеточные расчеты не сходятся к известному ответу при $N \to \infty$, то программа содержит грубую ошибку. Если сходимость к точному значению есть, но с порядком точности меньше теоретического (не тот наклон линий на графиках), то следует искать небольшую ошибку (например, сдвиг индекса суммирования или путаницу в коэффициентах).

Кусочная гладкость. Пусть функция и некоторое число ее производных кусочно-непрерывны. Если построить на отрезке $[a, b]$ равномерную сетку, то какие-то точки разрыва $u^{(q)}(x)$, $q \geqslant 0$, попадут внутрь интервалов. Сгущение сетки и применение метода Ричардсона при этом невозможно.

В этом случае нужно построить специальные сетки. Все точки разрыва надо взять в качестве узлов сетки, а между каждой парой соседних точек разрыва построить некоторую равномерную сетку. Если одновременно сгущать все эти сетки в одно и то же число раз, то условия применимости формул Ричардсона будут выполнены.

Такие сетки будем называть псевдоравномерными. Этот термин относится не к одной сетке, а ко всей совокупности сеток, получаемых сгущением начальной специальной сетки.

3.2.2. Рекуррентное уточнение

В подразд. 3.2.1 упоминалось, что при расчетах с заданной точностью приходится несколько раз сгущать сетку. Покажем, что такие расчеты позволяют кардинально улучшить точность, если сеточная функция достаточно гладкая, т. е. разложение погрешности по степеням шага содержит достаточно много членов.

Например, в формулах Эйлера—Маклорена поправочные члены можно рассматривать как разложение погрешности фор-

мул средних и трапеций по степеням шага h (3.27), (3.29). Обозначив через U_N расчет по квадратурной формуле трапеций или средних, получим

$$U = U_N + \sum_{m=0}^{M} c_m N^{-p-m\sigma} + o(N^{-p-M\sigma}). \tag{3.49}$$

Здесь коэффициенты c_m не зависят от шагов сетки. Для указанных формул $p = 2$ — порядок точности основной формулы средних или трапеций. Значения $\sigma = 2$, так как вследствие симметрии этих формул погрешность разлагается в ряд по четным степеням шага. Для несимметричных формул обычно $\sigma = 1$. Для справедливости разложения в задачах численного интегрирования требуется наличие непрерывной и ограниченной $(p + \sigma M)$-й производной $u(x)$.

Проведем серию расчетов, сгущая сетку в одно и то же число раз r. При справедливости разложения (3.49) в формулах Ричардсона вместо $o(N^{-p})$ появляется $O(N^{-p-\sigma})$. Тогда результат однократного уточнения (3.44) можно трактовать как результат применения метода более высокого порядка точности $p + \sigma$. К нему можно снова применить формулы Ричардсона с тем же r, но новым порядком точности $p + \sigma$.

Процедуру уточнения можно сделать рекуррентной. Обозначим решение на k-й сетке, даваемое исходной формулой, через $U_k^0 \equiv U_k$, а его погрешность — $R_k^0 \equiv R_k$. Однократное уточнение по Ричардсону обозначим $U_k^1 \equiv \tilde{U}_k$, второе уточнение — U_k^2 и т. д. Тогда формулы m-го уточнения примут следующий вид:

$$\begin{gathered} R_k^{m-1} = \frac{U_k^{m-1} - U_{k-1}^{m-1}}{r^{p+\sigma(m-1)} - 1}, \quad U_k^m = U_k^{m-1} + R_k^{m-1}, \\ p_k^{m-1} = \lg \frac{R_{k-1}^{m-1}}{R_k^{m-1}} / \lg r, \quad 1 \leqslant m \leqslant M. \end{gathered} \tag{3.50}$$

Основное время расчета занимает вычисление по исходной формуле: например, в квадратурных формулах нужно вычислять $u(x)$ в большом числе узлов, что является наиболее трудоемкой частью. Вычисления по формулам (3.50) требуют лишь небольшого числа простейших арифметических операций. Поэтому применение рекуррентного уточнения практически не увеличивает объема вычислений по сравнению с основной формулой, но существенно повышает порядок точности результата.

Таблица 3.6

Рекуррентное повышение точности

k	N_k	U_k^0	R_k^0	p_k^0	U_k^1	R_k^1	p_k^1	U_k^2	R_k^2	p_k^2
0	1	5,33333								
1	2	4,45552	$-2{,}9e-1$		4,16292					
2	4	4,13839	$-1{,}1e-1$	1,66	4,03268	$-8{,}7e-3$	2,21	4,02400		
3	8	4,03810	$-3{,}3e-2$	1,82	4,00467	$-1{,}9e-3$	2,74	4,00281	$-3{,}4e-4$	3,02
4	16	4,00988	$-9{,}4e-3$	1,94	4,00047	$-2{,}8e-4$	3,27	4,00019	$-4{,}0e-5$	3,82
5	32	4,00250	$-2{,}5e-4$	1,98	4,00004	$-3{,}0e-5$	3,67	4,00001	$-3{,}0e-6$	4,64
6	64	4,00063	$-6{,}2e-5$	2,00	4,00000	$-2{,}0e-6$	3,89	4,00000	$-1{,}0e-7$	5,33
	∞	4,00000	0	2,00	4,00000	0	4,0	4,00000	0	6,0

Пример 3.7. Возьмем в качестве исходной формулу трапеций (3.17) и рассмотрим тест (3.14), использованный ранее для разных квадратурных формул в табл. 3.1. Результаты рекуррентного уточнения удобно представлять в форме табл. 3.6. В ней первые два столбца — номер сетки k и число узлов N_k. Следующие три столбца соответствуют основному расчету по формуле трапеций, оценке ее погрешности по Ричардсону и эффективного порядка точности. Эта процедура описана в подразд. 3.2.1. Каждая следующая тройка столбцов — результат очередного уточнения. При каждом уточнении число остающихся сеток уменьшается на 1, поэтому полная таблица имеет вид треугольника: число строк в таблице равно числу использованных сеток, число групп столбцов ограничено гладкостью и не превышает M.

Расчет целесообразно проводить, добавляя к таблице по одной строке и проводя все доступные уточнения. Если хотя бы в одной из групп столбцов $\{U_k^{m-1}, R_k^{m-1}, p_k^{m-1}\}$ будет получено $|R_k^{m-1}| < \varepsilon$, то заданная точность достигнута и расчет прекращается. Значение U_k^m берется в качестве ответа, а $|R_k^{m-1}|$ считается мажорантной оценкой его погрешности.

Обязательно нужно использовать диагностику, описанную в подразд. 3.2.1. В каждой группе столбцов анализируется поведение p_k^{m-1}. До тех пор пока p_k^{m-1} монотонно стремится к теоретическому пределу, оценки погрешности R_k^{m-1} в этой группе асимптотически точны, а значения U_k^m достоверны. Нарушение монотонного стремления p_k^{m-1} к теоретическому пределу свидетельствует о выходе на ошибки округления. В этом случае нельзя продолжать вниз этот столбец и все лежащие правее него

группы столбцов. Однако в столбцах, лежащих левее, можно проводить дальнейшее сгущение сетки.

В данном примере видно, что p_k^0 быстро стремится к 2, поэтому первое уточнение является очень надежным. Значение p_k^1 стремится к 4 медленнее, но удовлетворительно. Поэтому второе уточнение также достаточно надежно. Стремление p_k^2 к 6 довольно медленное, поэтому к третьему уточнению нужно относиться осторожно (в табл. 3.6 оно не приведено). Однако уже второе уточнение при $N = 64$ обеспечивает высокую точность 10^{-7}. Тогда трудоемкость расчета практически не превышает трудоемкости формулы трапеций при том же $N = 64$. Такой метод получения высокой точности убедительно показывает эффективность рекуррентного уточнения.

Для визуального контроля работы программы удобны графики, описанные в подразд. 3.2.1. На первый график наносят зависимости $\lg|R_k^m|$ от $\lg N_k$ для всех m (рис. 3.3). Линия базового расчета ($m = 0$) является практически прямой с углом наклона $\operatorname{tg}\alpha_0 \approx -2$. Это соответствует второму порядку точности. График погрешности первого уточнения начинается на одну точку правее. Его начало искривлено, но он быстро выходит на прямую с наклоном $\operatorname{tg}\alpha_1 \approx -4$; это соответствует четвертому порядку точности. Для второго уточнения начальная и средняя части графика аналогичны предыдущему и при $N_k = 1\,024$ погрешность падает до $\approx 10^{-14}$. Однако при дальнейшем увеличе-

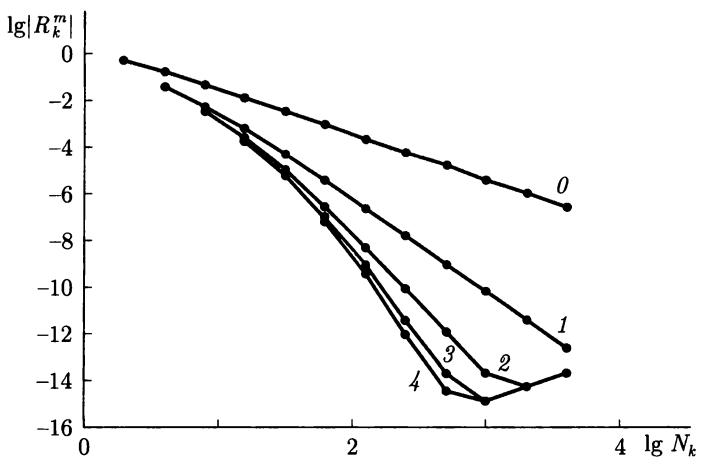


Рис. 3.3. Поведение погрешностей формулы трапеций при сгущении сетки и рекуррентных уточнениях (*0* — базовый расчет; *1* — *4* — номера уточнений)

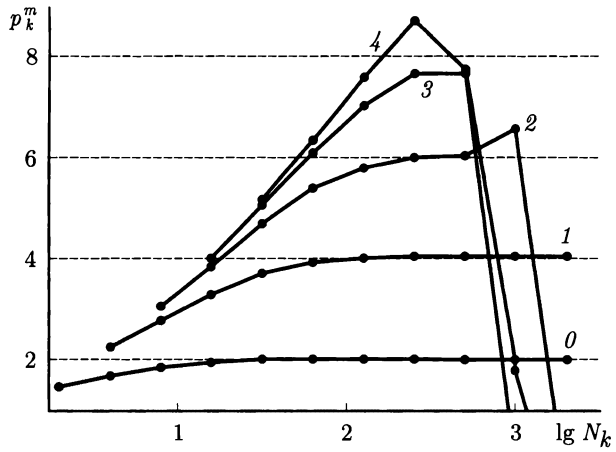


Рис. 3.4. Эффективные порядки точности формулы трапеций и рекуррентных уточнений при сгущении сетки (*0* — базовый уровень; *1* — *4* — номера уточнений)

нии N_k погрешность перестает убывать. Это означает выход на ошибки округления. Для следующих уточнений выход на ошибки округления происходит при том же уровне погрешности, но при соответственно меньших N_k. В данном примере третье уточнение еще повышает точность и позволяет достичь уровня 10^{-14} при $N_k = 512$. Но четвертое уточнение здесь уже практически бесполезно.

Для исходной формулы трапеций такой уровень точности 10^{-14} был бы достигнут лишь при $N_k > 10^7$, а для формулы Симпсона — при $N_k > 10^4$. Это наглядно иллюстрирует эффективность рекуррентных уточнений.

На рис. 3.4 показано поведение эффективного порядка точности в зависимости от шага сетки для разных уточнений. Видно, что базовая формула трапеций ($m = 0$) и первые два уточнения ($m = 1, 2$) успевают выйти на теоретические порядки точности $p = 2 + 2m$; при этом для второго уточнения график сразу после этого срывается в связи с выходом на ошибки округления. Для третьего уточнения ($m = 3$) ошибки округления сказываются несколько раньше выхода на предел, а график четвертого уточнения даже не успевает близко подойти к теоретическому пределу.

Выводы. Рекуррентное повышение точности является очень эффективным приемом по ряду причин: 1) позволяет получить ответ с гарантированной оценкой точности; 2) можно применять

простые формулы невысокого порядка точности в качестве исходных, получая в итоге ответ с высоким порядком точности; 3) требуемая точность достигается с использованием сравнительно небольшого числа узлов; 4) рекуррентное уточнение фактически является алгоритмом без насыщения, т. е. позволяет использовать все имеющиеся у функции непрерывные производные; 5) нередко позволяет установить гладкость исходной функции.

3.2.3. Квазиравномерные сетки

Одна-единственная сетка является либо равномерной, либо неравномерной; никаких альтернатив здесь нет. Но при сгущении сеток рассматривают не одну сетку, а некоторое семейство сеток Ω_N с разными числами интервалов N. Тогда можно построить какие-то семейства неравномерных сеток, обладающие нужными нам свойствами (например, в подразд. 3.2.1 упоминались псевдоравномерные сетки). Очень важным является семейство квазиравномерных сеток (впервые квазиравномерные сетки предложил и использовал в расчетах А. А. Самарский, 1952).

Пусть нас интересуют сетки по x с числом интервалов N на отрезке $a \leqslant x \leqslant b$; обозначим их через $\Omega_N[x] = \{a = x_0 < x_1 < < x_2 < \ldots < x_N = b\}$. Введем вспомогательную переменную ξ, $\alpha \leqslant \xi \leqslant \beta$. Рассмотрим некоторое преобразование $x(\xi)$, обладающее на отрезке $\alpha \leqslant \xi \leqslant \beta$ следующими тремя свойствами:

1) оно достаточно гладко, т. е. существует достаточно много непрерывных ограниченных производных:

$$|x^{(q)}(\xi)| \leqslant M_q, \quad q = 0, 1, \ldots, Q, \quad Q \gg 1; \quad \alpha \leqslant \xi \leqslant \beta; \tag{3.51}$$

2) строго монотонно:

$$x'(\xi) \geqslant \theta > 0; \quad \alpha \leqslant \xi \leqslant \beta; \tag{3.52}$$

3) преобразует отрезок $[\alpha, \beta]$ в отрезок $[a, b]$:

$$a = x(\alpha), \quad b = x(\beta). \tag{3.53}$$

Построим по переменной ξ на $[\alpha, \beta]$ равномерные сетки $\omega_N[\xi]$ со всевозможными числами интервалов $N = 1, 2, 3, \ldots$:

$$\xi_{nN} = \alpha + n\Delta, \quad \Delta = (\beta - \alpha)/N, \quad n = 0, 1, \ldots, N; \tag{3.54}$$

индекс N в ξ_{nN} обычно будем опускать, записав просто ξ_n. Каждой сетке $\omega_N[\xi]$ преобразование $x(\xi)$ ставит в соответствие некоторую сетку $\Omega_N[x]$:

$$x_{nN} \equiv x_n = x(\xi_n), \quad n = 0, 1, \ldots, N. \tag{3.55}$$

Таким образом, семейству равномерных сеток $\omega_N[\xi]$ сопоставлено некоторое семейство сеток $\Omega_N[x]$. Какими свойствами оно обладает?

Определение 3.1. Если преобразование $x(\xi)$ обладает свойствами (3.51)—(3.53), то семейство сеток $\Omega_N[x]$, порожденное семейством равномерных сеток $\omega_N[\xi]$, называется ***квазиравномерным***.

Для краткости слово «семейство» часто опускают и говорят просто о квазиравномерных сетках. Семейство содержит сетки со всевозможными $N = 1, 2, 3, \ldots$; для практического использования из него обычно выбирают какие-то последовательности N — например, сгущающиеся в одинаковое число раз сетки с $N_k = r^k N_0$.

На одном и том же отрезке $x \in [a, b]$ существует бесконечно много различных семейств квазиравномерных сеток, порожденных различными преобразованиями $x(\xi)$; те или иные преобразования подбирают в зависимости от задачи, которую предстоит решать на этих сетках.

Полуцелые точки. У порождающей равномерной сетки $\omega_N[\xi]$ середина и вообще любая дробная часть интервала $[\xi_{n-1}, \xi_n]$ определяется по обычным линейным формулам:

$$\begin{gathered} \xi_{n-1/2} = (\xi_{n-1} + \xi_n)/2 = \alpha + (\beta - \alpha)(n - 1/2)/N; \\ \xi_{n-\gamma} = \gamma\xi_{n-1} + (1 - \gamma)\xi_n, \quad 0 \le \gamma \le 1. \end{gathered} \tag{3.56}$$

Однако для квазиравномерной сетки середина и любая дробная часть интервала $[x_{n-1}, x_n]$ строится с помощью того же нелинейного порождающего преобразования:

$$x_{n-1/2} = x(\xi_{n-1/2}), \quad x_{n-\gamma} = x(\xi_{n-\gamma}), \tag{3.57}$$

поэтому для нее линейные соотношения неточны:

$$x_{n-1/2} \ne (x_{n-1} + x_n)/2, \quad x_{n-\gamma} \ne \gamma x_{n-1} + (1 - \gamma)x_n.$$

Разложением $x(\xi)$ в ряд Тейлора нетрудно проверить, что новое определение середины интервала (3.57) и традиционное весьма близки:

$$(x_n + x_{n-1})/2 - x_{n-1/2} = x''_{n-1/2}\Delta^2/8 = O(N^{-2}) = O(h_n^2), \tag{3.58}$$

где производная $x^{(q)}_{n-1/2} \equiv x^{(q)}(\xi_{n-1/2})$.

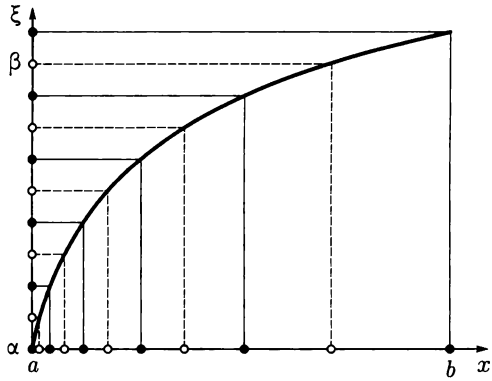


Рис. 3.5. Квазиравномерная сетка (точки — целые узлы сетки; кружки — полуцелые)

Все это удобно иллюстрировать с помощью графика функции, обратной к порождающему преобразованию $x(\xi)$. Отложим x по оси абсцисс, ξ — по оси ординат (рис. 3.5). На отрезке $[\alpha, \beta]$ оси ординат изображена равномерная сетка с целыми и полуцелыми узлами. По ней с помощью графика $x(\xi)$ на отрезке $[a, b]$ оси абсцисс построена квазиравномерная сетка.

Чтобы сгустить сетку $\Omega_N[x]$ в r раз, надо во столько же раз сгустить исходную сетку $\omega_N[\xi]$. Особенно наглядно сгущение вдвое: $r = 2$. Из формул (3.56), (3.57) и рис. 3.5 видно, что при этом все целые узлы начальной сетки становятся четными узлами удвоенной сетки, а полуцелые точки начальной сетки — нечетными узлами удвоенной сетки. Это подтверждает разумность данного способа введения полуцелых (дробных) узлов квазиравномерной сетки.

Шаги. Равномерная сетка $\omega_N[\xi]$ имеет постоянный шаг:

$$\Delta \equiv \xi_n - \xi_{n-1} = (\beta - \alpha)/N = \text{const} = O(N^{-1}). \qquad (3.59)$$

Длины шагов h_n квазиравномерной сетки $\Omega_N[x]$ неодинаковы. Используя разложение $x(\xi)$ в ряд Тейлора — Маклорена с центром в полуцелой точке $\xi_{n-1/2}$, получим для них следующее выражение:

$$h_n \equiv x_n - x_{n-1} = x'_{n-1/2}\Delta + \frac{1}{24}x'''_{n-1/2}\Delta^3 + \ldots = O(N^{-1}). \qquad (3.60)$$

Благодаря симметрии этой формулы разложение содержит степени Δ только одинаковой четности, а число членов в нем определяется гладкостью преобразования $x(\xi)$.

Найдем из (3.60) отношение двух шагов с разными номерами:

$$h_m/h_n = (x'_{m-1/2}/x'_{n-1/2})[1 + O(\Delta^2)]. \tag{3.61}$$

Здесь выражение в квадратной скобке практически всегда можно заменить единицей. Если интервалы соседние, т. е. $m = n + 1$, то выражение в круглой скобке также почти равно 1. Действительно, $x'_{n\pm1/2} = x'_n \pm 0,5x''_n\Delta + O(\Delta^2)$. Тогда (3.61) превращается в

$$h_{n+1}/h_n = 1 + [x''_n/x'_n]\Delta + O(\Delta^2) = 1 + O(N^{-1}). \tag{3.62}$$

Отношение соседних шагов стремится к 1 при $N \to \infty$; легко убедиться, что разность соседних шагов

$$h_{n+1} - h_n = O(N^{-2}) = O(h_n^2).$$

В то же время из (3.61) следует, что отношение далеких шагов может быть любым, большим или малым, и не стремится к 1 при увеличении числа узлов сетки. На рис. 3.5 хорошо видно, что первый и последний шаги h_n сильно различаются.

Заметим, что преобразование $x(\xi)$ вводит пользователь, ориентируясь на особенности конкретной задачи или класса задач. Удобно для несимметричных задач выбирать $\alpha = 0$, $\beta = 1$ и $0 \leqslant n \leqslant N$, а для симметричных — $\alpha = -1$, $\beta = 1$ и $-N \leqslant n \leqslant N$. В обоих случаях $\Delta = 1/N$.

Пример 3.8. Пусть рассматриваемая функция $u(x)$ быстро меняется вблизи левой границы отрезка $[a, b]$, но медленно — вдали от нее. Тогда надо сделать сетку $\Omega_N[x]$ густой вблизи левой границы, но вблизи правой границы сетка может быть достаточно редкой. Например, можно взять такое порождающее преобразование:

$$x(\xi) = a + (b - a)(e^{c\xi} - 1)/(e^c - 1), \quad c > 0, \quad 0 \leqslant \xi \leqslant 1. \tag{3.63}$$

Легко проверить, что преобразование (3.63) удовлетворяет всем требуемым условиям (3.51) — (3.53). Величина c есть управляющий параметр; чем он больше, тем сильнее сгущаются шаги h_n у левой границы. Для преобразования (3.63) справедливы соотношения

$$\begin{gathered} x'(\xi) = c(b - a)e^{c\xi}/(e^c - 1), \\ h_{n+1}/h_n = e^{c/N} = \text{const} \approx 1 + c/N, \\ h_1 \approx c(b - a)/[(e^c - 1)N], \quad h_N/h_1 \approx e^c. \end{gathered} \tag{3.64}$$

Отношения соседних шагов одинаковы, т. е. шаги образуют геометрическую прогрессию. Отношение наибольшего шага h_N к наименьшему h_1 уже при $c > 3$ становится очень большим, а первый шаг при этом очень мал: в $(e^c - 1)/c$ раз меньше шага равномерной сетки $h = (b - a)/N$.

Именно этот случай (при $c = 2$) изображен на рис. 3.5. Заметим, что если нужна сетка, густая вблизи правой границы и редкая вблизи левой, то можно также воспользоваться преобразованием (3.63), но взяв $c < 0$.

Пример 3.9. Нередко приходится решать задачи о процессах в слоистой среде с чередованием толстых и тонких слоев (например, о прохождении звука через оконный пакет из тонких стекол и довольно больших промежутков). Для правильной разностной аппроксимации уравнений сетки должны быть квазиравномерными и притом специальными (т. е. границы слоев должны быть узлами сетки). Для хорошей точности число интервалов в тонких слоях должно быть не малым. Но для экономичности нужно, чтобы в толстых слоях интервалов было не слишком много — примерно столько же, сколько и в тонких. Построим пример такой сетки.

Для простоты выберем трехслойную симметричную конфигурацию общей толщиной $2b$ с двумя граничными тонкими слоями размером $a \ll b$ и толстым средним слоем. Границы этих слоев — точки $x = -b,\ -b + a,\ b - a,\ b$. Рассмотрим следующее преобразование:

$$x(\xi) = c\xi/(1 + d\xi^2)^{1/2}, \quad x'(\xi)c/(1 + d\xi^2)^{3/2}, \\ c > 0, \quad d > 0, \quad -1 \leqslant \xi \leqslant 1. \tag{3.65}$$

Оно удовлетворяет требованиям (3.51) — (3.53).

Одно условие для подбора параметров c, d очевидно: $x(1) = b$. В качестве второго условия возьмем $x(\xi_*) = b - 1$, где ξ_* — некоторое выбранное нами число $(0 < \xi_* < 1)$. Это означает, что общее число интервалов сетки распределится так: доля ξ_* — в центральном толстом слое и по $(1 - \xi_*)/2$ в каждом тонком. Подстановка (3.65) в эти условия дает параметры

$$d = \frac{(b-a)^2 - b^2\xi_*^2}{a(2b-a)\xi_*^2} \approx \frac{b}{2a}(\xi_*^{-2} - 1) \gg 1, \\ c = \frac{b(b-a)}{\xi_*}\left[\frac{1-\xi_*^2}{a(2b-a)}\right]^{1/2} \approx [b^3(\xi_*^{-2} - 1)/(2a)]^{1/2} \gg b; \tag{3.66}$$

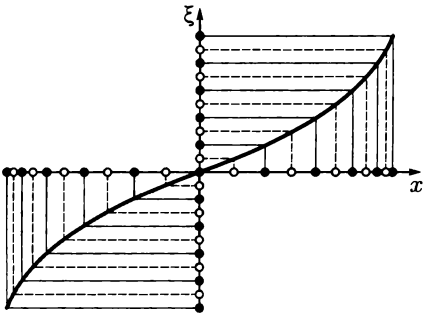


Рис. 3.6. Сетка (3.65) для трехслойной области (точки — узлы сетки; кружки — полуцелые точки)

здесь приближенные равенства относятся к случаю $b \gg a$, точные — к общему. Разумеется, величину ξ_* и полное число интервалов N надо выбирать так, чтобы $N(1-\xi_*)/2$ было целым числом. Только в этом случае узел сетки всегда будет попадать на границу раздела сред. На рис. 3.6 показан пример такой сетки для $a=1$, $b=6$, $\xi^*=0,5$, $N=5$ (полное число интервалов в этой симметричной задаче есть $2N$).

Пример 3.10. Возьмем преобразование $x=\xi^3$, $0 \leqslant \xi \leqslant 1$. Его график — гладкая монотонно возрастающая кривая. Однако это преобразование не порождает квазиравномерную сетку: $x'(0)=0$ и нарушено условие строгой монотонности (3.52). При этом $h_1=\Delta^3$, а $h_2=(2\Delta)^3$. Их отношение $h_2/h_1=8$ и не стремится к 1 при сгущении сеток. Это недопустимо: например, при написании разностных схем для уравнений в частных производных такое отношение приводит к уменьшению порядка аппроксимации и существенному снижению точности расчета.

Во всех рассмотренных примерах использовались преобразования $x(\xi)$, имеющие бесконечно много непрерывных производных. Это оптимальная ситуация.

Неограниченная область. Ранее неявно предполагалось, что a, b — ограниченный отрезок. Однако существует немало задач в неограниченной области. Простейшим примером служат несобственные интегралы от $u(x)$ на полупрямой и прямой. В краевых задачах для дифференциальных уравнений также возможна постановка краевых условий в бесконечно удаленной точке; например, в квантово-механических задачах о нахождении дискретного спектра это условие быстрого затухания волновой функции на бесконечности.

Очевидно, в неограниченной области невозможно ввести равномерную сетку с конечным числом интервалов. Это до сих пор существенно ограничивало применение сеточных методов к по-

добным задачам. Однако квазиравномерную сетку в неограниченной области нетрудно построить.

Например, рассмотрим следующее преобразование:

$$x(\xi) = c\xi/(1-\xi^2)^m, \quad c>0, \quad m>0, \quad -1 \leqslant \xi \leqslant 1, \qquad (3.67)$$

где c, m — управляющие параметры, причем m может быть нецелым.

Это преобразование переводит отрезок $\xi \in [-1,1]$ в прямую $x \in (-\infty, \infty)$. Такая сетка построена на рис. 3.7. Ее граничные узлы оказываются бесконечно удаленными точками: $x_{\pm N} = \pm\infty$. Соответственно крайние интервалы $(x_{-N}, x_{-N+1}]$ и $[x_{N-1}, x_N)$ неограниченны; разумеется, все остальные интервалы конечны. Однако середины крайних неограниченных интервалов $x_{-N+1/2}$ и $x_{N-1/2}$ оказываются конечными точками и то же относится к любой дробной точке. Это подтверждает разумность определения середины и дробной части интервала (3.57).

Опишем простейший априорный способ подбора параметров преобразования (3.67). Целесообразное число интервалов сетки N определяется мощностью компьютера. Если мы строим сетку, густую вблизи центра, то неявно предполагаем, что она наиболее важна для аккуратного решения задачи. Центральному шагу соответствует $\xi_1 = 1/N$, а четверти всех шагов — $\xi_{N/2} = 1/2$; подставляя их в (3.67), получаем

$$h_1 \approx c/N, \quad x_{N/2} = (c/2)(4/3)^m. \qquad (3.68)$$

Величина первого шага практически не зависит от m, поэтому c подбирают так, чтобы первый шаг оказался достаточно малым и

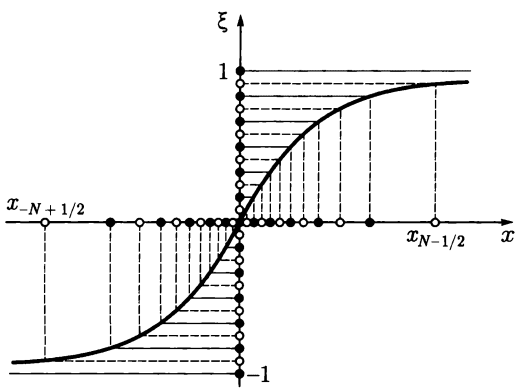


Рис. 3.7. Квазиравномерная сетка на прямой

обеспечивал нужную точность расчета. Найдя c, подбираем m из условия, чтобы половина интервалов сетки охватила бы наиболее важную область прямой; чем шире эта область, тем большее m приходится брать.

Если в (3.67) выбрать $0 \leqslant \xi \leqslant 1$, $0 \leqslant n \leqslant N$, то квазиравномерная сетка $\{x_n\}$ покрывает полупрямую.

Для квазиравномерной сетки в неограниченной области традиционное определение шага $h_n = x_n - x_{n-1}$ неприемлемо: при этом для бесконечных интервалов шаг оказывается бесконечным $h_N = h_{-N+1} = \infty$. Такие шаги нельзя использовать в формулах. Поэтому в неограниченной области целесообразно ввести иное определение шага. Для конечных интервалов справедливо соотношение (3.60). Постулируем это соотношение в качестве определения шага:

$$h_n = x'_{n-1/2}/N. \tag{3.69}$$

Такое определение дает конечную величину шага даже для бесконечных интервалов. Вводить новое определение шага (3.69) нужно одновременно для всех интервалов сетки.

Вычисление интеграла. Рассмотрим вычисление интеграла по формуле средних на квазиравномерной сетке. Для произвольной неравномерной сетки имеется формула (3.8) и выражение для ее погрешности (3.11)

$$R_N = \frac{1}{24} \sum_{n=1}^{N} u''_{n-1/2} h_n^3.$$

На квазиравномерной сетке в выражении погрешности можно произвести замену $h_n^3 \approx h_n x'^2_{n-1/2}/N^2$, тогда получим

$$R_N \approx \frac{1}{24N^2} \sum_{n=1}^{N} u''_{n-1/2} h_n x'^2_{n-1/2}. \tag{3.70}$$

Здесь полуцелая точка $x_{n-1/2}$ понимается теперь в смысле квазиравномерной сетки (3.57). Последнюю интегральную сумму можно заменить интегралом:

$$R_N \approx \frac{1}{24N^2} \int_a^b x'^2 u''(x)dx = \frac{\text{const}}{N^2}. \tag{3.71}$$

Здесь константа не зависит от числа узлов сетки. Все сделанные здесь приближенные замены вносили относительную погрешность $o(N^{-2})$.

Таким образом, формула средних на квазиравномерной сетке имеет погрешность $O(N^{-2})$. Форма ее остаточного члена (3.71) показывает, что к ней применим метод сгущения сеток с оценкой погрешности по методу Ричардсона и экстраполяционным повышением точности. Разумеется, теперь под коэффициентом сгущения r понимается отношение числа узлов соседних сеток.

Пусть существуют непрерывные производные $u^{(2M)}(x)$ и $x^{(2M-1)}(\xi)$. Тогда можно доказать, что погрешность формулы средних (3.11) разлагается в ряд по четным степеням N^{-1} до $O(N^{-2M})$ включительно. Пусть проведены расчеты на M квазиравномерных сетках с последовательным увеличением числа узлов в одинаковое число r раз. Тогда можно применять точно так же рекуррентное повышение точности, как это делалось для равномерных сеток.

Аналогичные выводы справедливы для формулы трапеций (3.15), (3.16) и для формулы Симпсона (3.19), (3.20) на квазиравномерных сетках. Это делает квазиравномерные сетки очень эффективным вспомогательным инструментом.

Однако есть формулы, к которым квазиравномерные сетки неприменимы: это формулы Эйлера — Маклорена, пригодные только для равномерной сетки, а также для формул Гаусса — Кристоффеля, в которых понятие шага сетки вообще отсутствует.

Интеграл на прямой. Квазиравномерные сетки дают удобный способ вычисления несобственных интегралов в неограниченной области. Рассмотрим следующий тест:

$$U = \int_0^\infty u(x)dx, \quad u(x) = \frac{2}{\pi(1+x^2)}, \quad U = 1. \tag{3.72}$$

Введем на полупрямой квазиравномерную сетку (3.67). Воспользуемся формулой средних (3.8) с определением шага (3.69), пригодным в неограниченной области. Получим следующую квадратурную формулу:

$$\begin{gathered} U_N = \frac{1}{N}\sum_{n=1}^{N} u(x(\xi_{n-1/2}))x'(\xi_{n-1/2}), \\ \xi_{n-1/2} = \frac{n-1/2}{N}, \quad x(\xi) = \frac{c\xi}{(1-\xi^2)^m}, \\ x'(\xi) = c\frac{1-(1-2m)\xi^2}{(1-\xi^2)^{m+1}}. \end{gathered} \tag{3.73}$$

Таблица 3.7

Вычисление несобственного интеграла (3.72)

N	$U_N^{(0)}$	$p_N^{(0)}$	$U_N^{(1)}$	$p_N^{(1)}$	$U_N^{(2)}$	$p_N^{(2)}$
2	1,01896					
4	1,00358		0,99846			
8	1,00084	2,488	0,99993		1,00003	
16	1,00021	2,110	1,00000	4,405	1,00000	
32	1,00005	2,023	1,00000	4,053	1,00000	4,885
64	1,00001	2,006	1,00000	4,002	1,00000	6,808
128	1,00000	2,001	1,00000	4,000	1,00000	6,080
256	1,00000	2,000	1,00000	4,000	1,00000	6,020
512	1,00000	2,000	1,00000	4,000	1,00000	6,055

Подынтегральная функция убывает довольно медленно, возьмем сравнительно небольшое значение параметра $m = 1$. Величина c слабо влияет на точность расчета, поэтому можно положить $c = 1$. Результаты расчетов на сгущающихся сетках приведены в табл. 3.7. Видна хорошая сходимость метода. Эффективные порядки точности основного расчета, первого и второго уточнений по Ричардсону быстро стремятся к своим теоретическим значениям (2, 4 и 6), а третье уточнение почти сразу выходит на ошибки округления. Таким образом, при втором уточнении на сетке $N = 512$ узлов гарантированно достигается очень высокая точность $\sim 10^{-15}$.

Этот способ вычисления несобственных интегралов универсален, прост и эффективен. Сравним его с другими способами вычисления интегралов в неограниченной области, описанными в учебниках.

Первый способ — обрезание верхнего предела и численный расчет в оставшейся конечной области по стандартным программам. В данном примере для точности $\sim 10^{-15}$ надо обрезать верхний предел на уровне $\sim 10^{15}$, что приводит к огромному объему расчетов на равномерных сетках, принятых в стандартных программах.

Второй способ — применение формул Гаусса — Кристоффеля в неограниченной области. Но узлы и веса таких формул известны только для функций с экспоненциальным убыванием. К рассмотренному примеру этот подход неприменим.

Третий способ — подбор замены переменных, преобразующей интеграл в собственный. Это нужно делать отдельно для каждого интеграла, причем поиск такой замены может оказаться достаточно сложным.

Таким образом, общепринятые способы оказываются гораздо менее пригодными для практических вычислений, чем метод квазиравномерных сеток.

3.2.4. Метод Эйткена

Метод Ричардсона хорошо работает для достаточно гладких функций, когда погрешность сеточного метода разлагается в ряд по заранее известным целым отрицательным степеням N. В практике существует немало задач для функций с особенностями. В них разложение погрешности может содержать нецелые отрицательные степени N. Формально к ним можно применять метод Ричардсона и даже рекуррентное повышение точности с нецелыми p и σ (3.49), если эти показатели заранее известны. Однако для априорного нахождения этих показателей для каждой конкретной задачи нужно провести непростое теоретическое исследование. Это далеко не всегда удается.

На практике удобен другой подход. Проведем расчет на серии сеток $\{N_k\}$, $k = 0, 1, 2, \ldots$ и вычислим сеточные значения $\{U_k\}$. Будем предполагать, что погрешность R_k разлагается в степенной ряд с неизвестными показателями степени (они могут быть нецелыми):

$$R_k \equiv U - U_k = \sum_{m=1} c_m N_k^{-p_m}, \quad 0 < p_1 < p_2 < p_3 < \ldots \qquad (3.74)$$

Если сгущать сетки в одно и то же число раз $r = N_k/N_{k-1} = \text{const}$, то главный член погрешности убывает от одной сетки к другой в $q_1 \equiv r^{p_1}$ раз. Тем самым эти погрешности убывают примерно в геометрической прогрессии со знаменателем $1/q_1$.

Оставим в погрешности (3.74) только главный член, а равенство заменим на приближенное. Запишем такие равенства для трех последовательных сеток:

$$\begin{gathered} U - U_k = R_k, \quad R_k \approx c_1 N_k^{-p_1}, \\ U - U_{k-1} = R_{k-1}, \quad R_{k-1} \approx q_1 R_k, \\ U - U_{k-2} = R_{k-2}, \quad R_{k-2} \approx q_1 R_{k-1} \approx q_1^2 R_k, \quad q_1 = r^{p_1}. \end{gathered} \qquad (3.75)$$

Попарно вычитая равенства (3.75), исключим неизвестное точное значение U:

$$U_k - U_{k-1} \approx (q_1 - 1)R_k, U_{k-1} - U_{k-2} \approx q_1(q_1 - 1)R_k. \quad (3.76)$$

Разделив соотношения (3.76), получим знаменатель геометрической прогрессии и эффективный порядок точности (показатель степени):

$$q_{1k} = \frac{U_{k-1} - U_{k-2}}{U_k - U_{k-1}}, \quad p_{1k} = \frac{\lg q_{1k}}{\lg r}. \quad (3.77)$$

Равенства (3.77) записаны как точные, но для эффективных q_{1k} и p_{1k}. При измельчении сетки эффективные показатели стремятся к точным: $q_{1k} \to q_1$ и $p_{1k} \to p_1$.

Подставив (3.77) в (3.75), получим асимптотическую оценку остаточного члена:

$$R_k \approx \frac{U_k - U_{k-1}}{q_{1k} - 1}. \quad (3.78)$$

Учитывая этот остаточный член как поправку в (3.75), получим уточненное значение

$$\tilde{U}_k = U_k + R_k \approx U_k + \frac{U_k - U_{k-1}}{q_{1k} - 1}. \quad (3.79)$$

Формулы (3.77) — (3.79) записаны в таком виде, чтобы ошибки округления сказывались как можно меньше; преобразовывать их к другим формам не рекомендуется.

Эти формулы получил Эйткен для ускорения суммирования медленно сходящейся геометрической прогрессии. Видно, что их можно применять для оценки погрешности сеточных методов и экстраполяционного повышения точности аналогично методу Ричардсона при неизвестных и нецелых показателях степеней $\{p_m\}$. Однако в методе Эйткена определяются два неизвестных параметра c_1, p_1 (3.74) (формула для c_1 явно не выписывалась). Для этого требуется не две сетки, как в методе Ричардсона, а три.

Несобственный интеграл. В качестве примера рассмотрим интегрирование неограниченной функции:

$$U = \int_0^4 u(x)dx, \quad u(x) = \frac{1}{\sqrt{x}}, \quad U = 4. \quad (3.80)$$

Поскольку $u(0) = \infty$, то нельзя применять формулы трапеций, Симпсона или Эйлера — Маклорена. Однако можно использовать формулу средних.

Таблица 3.8

Формула средних для несобственного интеграла (3.80)

N	U_N	$\lg\lvert R_N\rvert$	R_{N-1}/R_N
2	3,15470	-0,07	
4	3,39769	-0,22	1,4034
8	3,57292	-0,37	1,4103
16	3,69771	-0,52	1,4128
32	3,78618	-0,67	1,4137
64	3,84879	-0,82	1,4141
128	3,89307	-0,97	1,4141
256	3,92439	-1,12	1,4142
512	3,94653	-1,27	1,4141
1 024	3,96219	-1,42	1,4142
2 048	3,97327	-1,57	1,4141

Для простоты проведем серию расчетов на равномерных сетках с $N_0 = 2$ и последовательным удвоением числа узлов сетки $N_k = 2N_{k-1}$. Результаты приведены в табл. 3.8, где значения погрешности вычислены по точному значению погрешности $R_k = U - U_k$. Видно, что значения U_k стремятся к точному ответу $U = 4$, но очень медленно. Отношения соседних погрешностей близки к 1,4, а не к 4, следовательно, эффективный порядок точности далек от второго. Использовать метод Ричардсона при этом бессмысленно.

Воспользуемся методом Эйткена, взяв значения по формуле средних U_k из табл. 3.8. Расчет главного члена погрешности по формуле приведен на рис. 3.8 (помечен индексом «*0*»). Эта погрешность близка к истинной и убывает весьма медленно. Наклон графика $\operatorname{tg}\alpha_0 \approx 0,5$ указывает на нецелый порядок точности. Расчет эффективного порядка точности формулы средних по (3.77), приведенный на рис. 3.9 (также помечен индексом «*0*»), подтверждает это.

Это позволяет провести уточнение по формуле (3.79). Его погрешность также показана на рис. 3.8 (индекс «*1*»). Каждая точка этой линии вычислена по трем точкам базовой кривой. Погрешность уточненного решения много меньше, чем для исходной формулы средних. Ее график также практически прямая

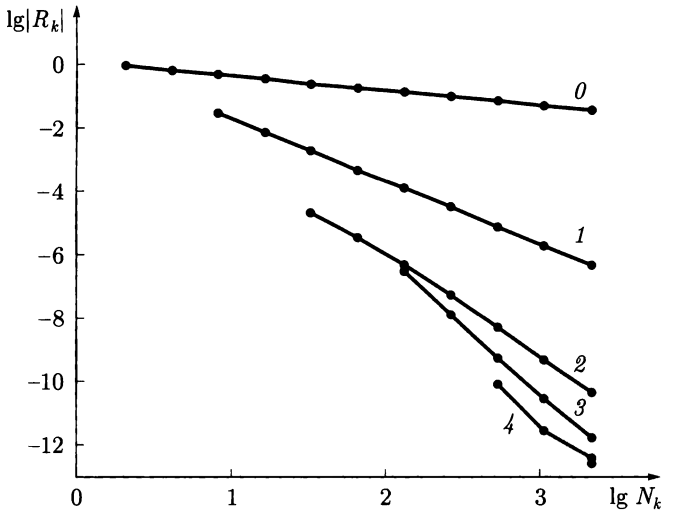


Рис. 3.8. Эффективные погрешности в методе Эйткена для примера (3.80): *0* — базовая формула средних; *1* — *4* — номера уточнений m

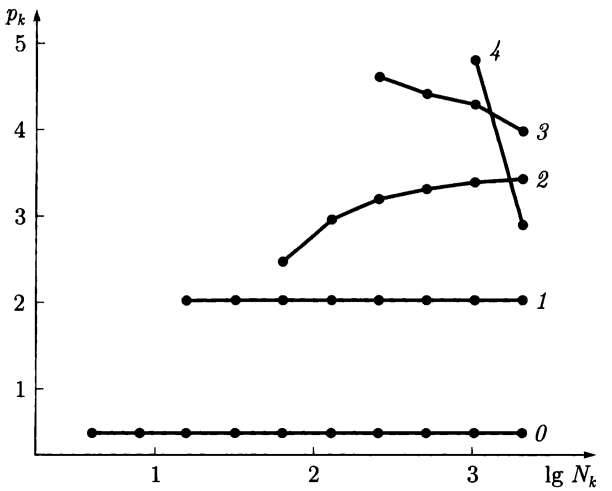


Рис. 3.9. Эффективные порядки точности в методе Эйткена для примера (3.80): *0* — базовая формула средних; *1* — *4* — номера уточнений m

линия, т. е. эта погрешность зависит от числа узлов N снова по степенному закону. Эта линия начинается на две точки правее базовой, так как при уточнении по Эйткену мы теряем две сетки.

Рекуррентное уточнение. При исключении главного члена погрешности (3.79) у нас остается следующий член, соответ-

ствующий показателю степени p_2. Поведение погрешности уточненного решения на рис. 3.8 соответствует эффективному порядку точности ≈ 2 (см. рис. 3.9). Поэтому можно провести второе уточнение по Эйткену, как это делалось в рекуррентном методе Ричардсона.

Погрешность второго уточнения также показана на рис. 3.8. Ее график снова сдвинут на две точки вправо. Теперь начало графика немного искривлено, но далее он выходит на наклон, соответствующий эффективному порядку точности $\approx 3,5$ (см. рис. 3.9). Второе уточнение здесь также сильно увеличивает точность.

На имеющихся сетках можно провести дальнейшие уточнения. Однако видно, что они уже не сильно повышают точность, причем четвертое уточнение выходит практически на ошибки округления при $N_k = 2\,048$. В итоге достигается погрешность $\approx 10^{-12}$, что в 10^{10} раз точнее базового расчета по формуле средних! Это показывает, что рекуррентный метод Эйткена на базе формулы средних является эффективным способом вычисления несобственных интегралов с высокой точностью.

В учебниках описан ряд способов вычисления несобственных интегралов.

Первый способ — обрезание пределов интегрирования вблизи особой точки и применение стандартных программ. При этом для высокой точности необходимо брать очень близкий предел обрезания и очень мелкие шаги интегрирования, что весьма трудоемко.

Второй способ — выделение в качестве веса основной особенности подынтегральной функции. Но узлы и веса известны только для единственного случая особенности — обращения в бесконечность $\sim 1/\sqrt{x}$, что делает этот способ малопригодным на практике.

Третий способ — нахождение замены переменных, преобразующей интеграл в собственный. Это делается индивидуально для каждой конкретной функции, и зачастую найти подходящее преобразование нелегко.

Метод Эйткена прост и достаточно универсален. Поэтому он является наиболее эффективным для вычисления несобственных интегралов широкого класса функций.

Выводы. 1. При каждом уточнении по Эйткену мы теряем две сетки, тогда как по методу Ричардсона лишь одну. Несмотря на это, метод Эйткена позволяет использовать достаточно высокие уточнения.

2. Метод Эйткена полезен не только для несобственных интегралов, но и при интегрировании ограниченных функций, имеющих неустранимые особенности производных. Например, заменим в тесте (3.80) подынтегральную функцию на $u(x) = \sqrt{x}$. Интеграл станет собственным, но главный член погрешности будет содержать $u'(0) = \infty$. В этом случае формулы средних или трапеций не будут иметь второго порядка точности и пользоваться методом Ричардсона нельзя. Метод Эйткена дает здесь отличные результаты.

3. Если для конкретного примера проведено исследование разложения погрешности в сумму (3.74) и найдены теоретические показатели степени $\{p_k\}$, то можно применить метод Ричардсона с показателями p_1 и даже провести одно уточнение с показателем p_2. Однако при первом уточнении по Ричардсону в сумме (3.74) исключается член $O(N^{-p_1})$, но дополнительно появляется $O(N^{-(2p_2-p_1)})$. При этом второе уточнение по Ричардсону можно делать с показателем p_2, но следующее уточнение уже выполняется с $\min\{p_3, 2p_2 - p_1\}$. Это осложняет использование метода Ричардсона даже при известных $\{p_k\}$. Метод же Эйткена применяется единообразно, так как в нем автоматически определяется степень главного на данный момент члена погрешности.

4. Точное разложение погрешности может содержать не только степенные члены, но и слагаемые $O(N^{-k} \ln N)$ или другие, близкие к степенным. В этом случае метод Эйткена не является строгим и его эффективность уменьшается. На практике это проявляется в том, что эффективный порядок точности медленно стремится к пределу, а очередное уточнение по Эйткену не сильно уменьшает погрешность. Но даже в этом случае эффективность метода Эйткена остается обычно достаточной для практического применения.

5. Метод Эйткена применим не только к вычислению интегралов, но и к решению широкого класса сеточных задач для функций.

3.3. КУБАТУРНЫЕ ФОРМУЛЫ

3.3.1. Метод средних

Интегралы невысокой кратности $(2-3)$ возникают при вычислениях на плоскости и в объеме. Для них дает хорошие результаты даже простейший метод ячеек (средних). Интегралы умеренной кратности $(4-6)$ появляются при вычислениях в ше-

стимерном пространстве координат—импульсов. Для них наиболее употребительным является произведение одномерных квадратурных формул. Интегралы высокой кратности ($\geqslant 7$) встречаются в задачах многократного рассеяния частиц, в экономических и других прикладных расчетах. Здесь сеточные методы требуют огромного объема вычислений и более экономичными оказываются статистические методы.

Обычно стараются преобразовать многомерную область интегрирования к прямоугольному параллелепипеду (единичному многомерному кубу). Задачи с криволинейной границей гораздо сложнее для расчетов. Здесь ограничимся двухмерными и трехмерными интегралами в областях простейшей формы.

Прямоугольная область. Рассмотрим двухмерный интеграл по прямоугольной области

$$U = \int_{\alpha}^{\beta} dy \int_{a}^{b} u(x,y)dx. \tag{3.81}$$

Введем сетку, образованную пересечением семейств линий (такую сетку называют регулярной):

$$\begin{gathered}\{x_n, 0 \leqslant n \leqslant N\}, x_0 = a, x_N = b, h_{xn} = x_n - x_{n-1}; \\ \{y_m, 0 \leqslant m \leqslant M\}, y_0 = \alpha, y_M = \beta, h_{ym} = y_m - y_{m-1}.\end{gathered} \tag{3.82}$$

Воспользуемся аддитивностью интеграла

$$\int_{\alpha}^{\beta} dy \int_{a}^{b} u(x,y)dx = \sum_{n=1}^{N}\sum_{m=1}^{M} \int_{x_{n-1}}^{x_n} dx \int_{y_{m-1}}^{y_m} u(x,y)dy.$$

Приближенно заменим интеграл по каждой ячейке сетки значением функции в центре ячейки (он же является ее центром тяжести), умноженным на площадь ячейки:

$$U \approx U_{NM} = \sum_{n=1}^{N}\sum_{m=1}^{M} h_{xn}h_{ym}u(x_{n-1/2}, y_{m-1/2}). \tag{3.83}$$

Оценка погрешности этого аналога интегральной формулы средних для многомерного случая может быть выполнена аналогично одномерному случаю. Очевидно, формула точна для линейной функции $u(x,y) = C_0 + C_1x + C_2y$. Значит, в оценку погрешности не могут входить производные ниже второго порядка.

В разложении по формуле Тейлора относительно центра ячейки фигурируют следующие слагаемые, содержащие вторые частные производные (мы предполагаем их существование и непрерывность):

$$(x - x_{n-1/2})^2 u_{xx},\ (x - x_{n-1/2})(y - y_{m-1/2})u_{xy},\ (y - y_{m-1/2})^2 u_{yy}.$$

Второе слагаемое при интегрировании по ячейке дает нуль в силу нечетности. Структура главного члена погрешности ясна из соображений размерности. Точное значение численного коэффициента $1/24$ нетрудно получить, если вычислить интеграл от квадратичной по обеим переменным функции и сравнить с точным значением, аналогично тому, как это делалось в подразд. 3.1.2. Отсюда следует

$$R_{NM} = \frac{1}{24} \sum_{n=1}^{N} \sum_{m=1}^{M} h_{xn} h_{ym} [h_{xn}^2 u_{xx}(x_{n-1/2}, y_{m-1/2}) + \\ + h_{ym}^2 u_{yy}(x_{n-1/2}, y_{m-1/2})].$$

На равномерной сетке $h_{xn} = (b - a)/N = \text{const}$, $h_{ym} = (\beta - \alpha)/M = \text{const}$ это выражение переходит в

$$R_{NM} = C_x h_x^2 + C_y h_y^2 = O(h_x^2 + h_y^2) = O(N^{-2} + M^{-2}), \qquad (3.84)$$

где

$$C_x = \frac{1}{24} \int_{\alpha}^{\beta} [u_x(b, y) - u_x(a, y)] dy,$$

$$C_y = \frac{1}{24} \int_{a}^{b} [u_y(x, \beta) - u_y(x, \alpha)] dx.$$

Тем самым, формула средних имеет *второй порядок точности по обеим переменным.*

Уточнение типа Эйлера—Маклорена в многомерном случае теоретически возможно, но практически очень громоздко и обычно не применяется.

Сгущение сетки. При выводе выражения для погрешности формулы средних предполагалось существование у подынтегральной функции непрерывных частных производных второго порядка. При наличии производных более высокого порядка можно сгущать сетки аналогично одномерному случаю и использовать прием Ричардсона для оценки погрешности и рекуррентного уточнения.

Начнем для простоты с равномерных сеток $h_{xn} = \text{const}$, $h_{ym} = \text{const}$. Сгущение многомерной сетки нужно проводить так, чтобы погрешность (3.84) убывала по обеим переменным в одно и то же число раз. Поскольку порядки точности в (3.84) одинаковы, то коэффициенты сгущения обеих сеток должен быть одинаковы: $r_x = r_y = r$.

Построение программы и диагностика погрешности и эффективных порядков точностей практически не отличаются от одномерного случая. В таблицах и графиках, аналогичных табл. 3.6 и рис. 3.3, 3.4, приводятся погрешности и эффективные порядки точности в зависимости от числа узлов по любой из переменных, поскольку числа узлов пропорциональны. Вся теория обобщается на случай вычисления интеграла по параллелепипеду произвольной размерности.

Удвоение числа узлов двумерной сетки по каждой переменной приводит к увеличению объема вычислений в четыре раза. Это не слишком экономично. Для уменьшения трудоемкости и постановки большего числа точек на графике можно использовать «магические» сетки (3.45) — (3.48), напомним, что они хороши для основного расчета и пригодны для первого уточнения по Ричардсону. Однако последующие уточнения нужно проводить очень осторожно (см. подразд. 3.2.1).

О других сетках. Однократный и рекуррентный методы Ричардсона можно использовать, если сетки по обеим переменным квазиравномерные. Порождающие преобразования по каждой переменной можно брать свои, исходя из конкретной задачи.

Пусть нужна сетка, подробная вблизи линии $y = \tilde{y}$, параллельной координатной оси x. Сетку по x оставляем равномерной. По переменной y строим квазиравномерную сетку, подробную вблизи $\tilde{y}$. Пусть нужно сгустить сетку вблизи точки $(\tilde{x}, \tilde{y})$. Тогда по каждой переменной строим квазиравномерную сетку, подробную соответственно вблизи точек $\tilde{x}, \tilde{y}$. Сгущение таких сеток производится одновременным увеличением числа узлов в одно и то же число раз.

Однако трудно построить квазиравномерную сетку, сгущающуюся вблизи прямой, не параллельной какой-нибудь оси координат. Для этого необходимо выбрать новую систему координат, одна из координатных линий которой совпадает с этой прямой. Еще труднее сгустить сетки вблизи кривой — для этого надо переходить к соответствующим криволинейным координатам.

Нередки задачи, в которых $u(x, y)$ имеет особенности (разрывы функции или ее производных) на линиях, параллельных од-

ной из координатных осей, например при $y = \tilde{y}$. Тогда хорошая сетка по y должна содержать узел $y_m = \tilde{y}$. В подразд. 3.2.1. описано, как сгущение таких сеток порождает псевдоравномерные сетки, на которых также возможно применение метода Ричардсона.

Однако если особенностью или областью наиболее значительного изменения функции является биссектриса координатного угла или некоторая кривая, то построение квази- или псевдоравномерной сетки является проблематичным.

Имеется один важный случай применения квазиравномерных сеток: вычисление интегралов в неограниченной области. Например, пусть нужно вычислить интеграл по полосе: $-\infty < x < +\infty$, $\alpha \leqslant y \leqslant \beta$. В этом случае сетку $\{y_m\}$ можно взять равномерной, но по переменной x — квазиравномерной (3.67). При этом надо использовать шаг $h_y = (\beta - \alpha)/M$, но по первой переменной нужно переопределить шаг и координату: $h_{xn} = x'(\xi_{n-1/2})/N$, $x_{n-1/2} = x(\xi_{n-1/2})$. Подстановка этих величин в кубатуру (3.83) дает погрешность $O(N^{-2} + M^{-2})$, т. е. второй порядок точности. Метод Ричардсона для сгущения такой сетки остается применимым.

Многоугольная область. Двумерная формула средних обобщается на случай многоугольной области. Всякий многоугольник можно разбить на конечное число треугольников. Сетка из таких треугольников в общем случае не будет регулярной. Треугольники в ней уже нельзя перенумеровать двумя индексами, каждый из которых относится к одной переменной. Перенумеруем такие треугольники единым индексом j $(1 \leqslant j \leqslant J)$. Для регулярной сетки было бы $J \sim NM$. Три вершины каждого треугольника пронумеруем в произвольном порядке индексом k $(1 \leqslant k \leqslant 3)$. Одна и та же точка треугольной сетки одновременно является вершиной нескольких треугольников, при этом в каждом треугольнике у нее может быть свой индекс k. Напомним, что центром тяжести j-го треугольника является точка пересечения его медиан. Она имеет координаты

$$\bar{x}_j = \frac{1}{3}\sum_{k=1}^{3} x_{jk}, \quad \bar{y}_j = \frac{1}{3}\sum_{k=1}^{3} y_{jk}.$$

Площадь этого треугольника равна

$$s_j = \sqrt{l(l-a)(l-b)(l-c)}, \quad l = (a+b+c)/2,$$

где a, b, c — длины сторон j-го треугольника (они легко выражаются через координаты его вершин).

В формуле средних вместо площади прямоугольной ячейки берется площадь треугольника s_j, а в качестве значения функции в средней точке — ее значение в центре масс треугольника $(\bar{x}, \bar{y})$

$$U_J = \sum_{j=1}^{J} s_j u(\bar{x}_j, \bar{y}_j). \qquad (3.85)$$

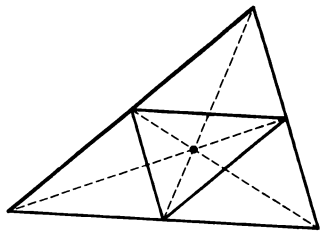


Рис. 3.10. Треугольная ячейка и ее дробление (штриховая линия — медианы; точка — центр тяжести)

Интеграл по каждому треугольнику по-прежнему точно берется для линейной функции $u(x, y) = C_0 + C_1 x + C_2 y$ в силу определения центра тяжести. При локальном разложении функции в ряды Тейлора погрешность внутри каждой ячейки будет содержать только вторые производные, причем лишь u_{xx}, u_{yy}. Следовательно, погрешность будет иметь второй порядок малости относительно сторон треугольной ячейки.

В этом случае также применим метод Ричардсона. Отрезками, соединяющими середины сторон, каждый треугольник разбивается на четыре равных треугольника (рис. 3.10). При этом сторона каждого треугольника уменьшается ровно в два раза, а главный член суммарной погрешности — в четыре раза (полное число ячеек J увеличивается тоже в четыре раза). Это означает, что для однократного метода Ричардсона надо принять коэффициент сгущения $r = 2$ и порядок точности $p = 2$.

Описанный процесс сгущения можно многократно повторять. Внутри каждого исходного треугольника сетка сгущается как равномерная. Поэтому рекуррентное уточнение по Ричардсону применимо. При каждом очередном сгущении $r = 2$, а эффективный порядок точности увеличивается на $\sigma = 2$. Однако треугольники исходной сетки, вообще говоря, не равны, поэтому в целом сгущенные сетки будут не равномерными, а псевдоравномерными. Вычислению интегралов это не препятствует (хотя для уравнений в частных производных такое сгущение может оказаться неприемлемым). Способов построения квазиравномерных треугольных сеток при произвольных исходных сетках пока не найдено.

Криволинейная граница. Эти задачи намного более трудны для вычислений. Накроем область с криволинейной границей регулярной сеткой (3.82) или нерегулярной треугольной сеткой. Ячейки сетки, целиком лежащие внутри границы, назовем

внутренними. Интегралы по таким ячейкам мы умеем вычислять со вторым порядком точности. Ячейки, целиком лежащие вне границы, назовем внешними; они не вносят вклада в интеграл. Ячейки, которые пересекает криволинейная граница, назовем граничными; они представляют основную трудность. Во-первых, нелегко определить, какие ячейки являются граничными, а какие — внутренними. Во-вторых, только часть граничной ячейки дает вклад в интеграл. Найти площадь этой части и ее центр тяжести достаточно сложно. Поэтому вычислить интеграл по граничной ячейке со вторым порядком точности практически невозможно и даже первого порядка точности трудно добиться.

Поэтому метод ячеек для криволинейной границы обычно дает лишь первый порядок точности относительно сторон ячейки. При этом оценка погрешности $O(h)$ будет не асимптотически точной, а лишь мажорантной. В этом случае при сгущении сетки нельзя пользоваться экстраполяционным методом Ричардсона для повышения порядка точности. Этот метод позволяет получить лишь ориентировочную оценку погрешности по порядку величины, причем надо использовать p =1, а не 2. Хорошие результаты можно получить лишь для границ простейшей формы (окружность, эллипс и т. п.), при которых специальным преобразованием переменных (например, полярных координат) область удается привести к прямоугольнику.

Многомерность. Метод ячеек естественно обобщается на произвольное число измерений L. Если область является прямоугольным параллелепипедом (линейным преобразованием сводится к единичному кубу), то для нее пишется аналог (3.83), где кратность суммы равна L, берется произведение шагов по всем переменным и центр соответствующей параллелепипедальной ячейки. Формула будет иметь второй порядок точности относительно числа узлов по каждой переменной.

Пусть трехмерная область является многогранником. Многогранную область можно разбить на трехгранные пирамиды. В этом случае применима формула (3.85), где под s_j надо понимать объем j-й пирамиды, а под $\bar{x}_j, \bar{y}_j, \bar{z}_j$ — центр тяжести пирамиды. Аналогичные формулы выписываются для многогранной области в L-мерном пространстве.

Оценим границы применимости метода. Если характерное число точек по каждой переменной $\sim N$, то полное число ячеек $J \sim N^L$. Второй порядок точности означает, что погрешность $R = O(N^{-2}) = O(J^{-2/L})$. Скорости современных компьютеров

позволяют брать для рядовых задач $J \sim 10^6 \div 10^9$. Будем считать математическую точность расчета по формуле средних приемлемой, если $R \sim 10^{-4} \div 10^{-6}$, это дает нам ограничения по размерности $L \leqslant 3$.

Таким образом, базовая формула метода средних обеспечивает разумную точность при приемлемом объеме вычислений для двумерных и трехмерных интегралов. В трехмерном случае для этого по каждой переменной берется $N \sim 100 \div 1\,000$ узлов. Использование сгущения сеток и метода Ричардсона позволяет увеличить точность и выполнять расчеты 4-мерных интегралов. Однако при большем числе измерений этот метод не обеспечивает хорошей точности.

3.3.2. Произведение квадратурных формул

Начнем рассмотрение с простейшего случая двумерного интеграла по прямоугольной области (3.81).

Введем обозначение

$$f(y) = \int_a^b u(x, y)dx,$$

тогда

$$U = \int_\alpha^\beta f(y)dy.$$

Вычислим последний интеграл по некоторой квадратурной (одномерной) формуле с весами c_m и узлами $\tilde{y}_m$:

$$U = \int_\alpha^\beta f(y)dy \approx \sum_{m=1}^{M} c_m f(\tilde{y}_m). \tag{3.86}$$

Можно брать любые квадратурные формулы, рассмотренные в подразд. 3.1.1. Нетрудно видеть, что для формулы средних надо выбрать сетку $\{y_m\}$, $(y_0 = \alpha, y_M = \beta)$ и положить узлы и веса равными

$$\tilde{y}_m = y_{m-1/2}, \quad c_m = h_{ym} = y_m - y_{m-1}. \tag{3.87}$$

Для формулы трапеций сумма берется от $m = 0$ до $m = M$:

$$\tilde{y}_m = y_m, \quad 0 \leqslant m \leqslant M;$$

$$c_0 = \frac{h_{y1}}{2}, \tag{3.88}$$

$$c_M = \frac{h_{yM}}{2}, \quad c_m = \frac{h_{ym} + h_{y,m+1}}{2}, \quad 1 \leqslant m \leqslant M-1.$$

Для некоторых формул Гаусса — Кристоффеля веса и узлы приведены в табл. 3.4. Можно использовать формулы Эйлера — Маклорена, но тогда следует взять равномерную сетку и к формуле трапеций (3.86) — (3.88) добавить члены с производными на границах.

Для приближенного вычисления $f(\tilde{y}_m)$ в узлах также применим одномерную квадратурную формулу (вообще говоря, другую):

$$f(\tilde{y}_m) \approx \sum_{n=1}^{N} d_n u(\tilde{x}_n, \tilde{y}_m),$$

здесь $\tilde{y}_m$ входит как параметр. В результате получаем *прямое произведение* одномерных квадратурных формул:

$$U \approx \sum_{m=1}^{M} \sum_{n=1}^{N} c_m d_n u(\tilde{x}_n, \tilde{y}_m).$$

Если по обеим координатам использована формула средних, то получаем ее двумерный аналог — метод ячеек (3.83).

Безусловным достоинством метода является большая свобода в выборе одномерных квадратурных формул по разным направлениям. При этом порядок точности по каждой из координат может быть свой.

Сгущение сеток. Для сгущения сеток в прямоугольной области надо выбирать коэффициенты сгущения по каждой из координат так, чтобы ошибки одномерных квадратурных формул убывали в одинаковое число раз. Если погрешность $R_{NM} = O(N^{-p} + M^{-q}) = c_x N^{-p} + c_y M^{-q}$, то для сгущенной сетки погрешность $R_{r_x N, r_y M} = c_x r_x^{-p} N^{-p} + c_y r_y^{-q} M^{-q}$. При $r_x^p = r_y^q$ оба основных члена погрешности пропорциональны, так что их можно исключить одновременно. Например, при $p = 2$ и $q = 4$ можно выбирать $r_x = 4$ и $r_y = 2$. При таком сгущении трудоемкость расчета на более подробной сетке в восемь раз больше, чем на исходной. Можно уменьшить трудоемкость, выбрав $r_x = 2$ и $r_y = \sqrt{2}$. При этом по переменной y нужно использовать соответствующие «магические» сетки.

Многомерность. Кубатурные формулы на основе произведения квадратур формально можно строить для любого числа измерений L. Оценим, какие формулы можно рекомендовать для практических расчетов.

Ранее было показано, что произведение одномерных формул средних пригодно для $L \approx 3$, а с учетом уточнения по Ричардсону — до $L \approx 4$. Одномерная формула трапеций имеет тот же порядок точности, что и формула средних. Поэтому оценка для нее аналогична.

Формула Симпсона имеет четвертый порядок точности. Однако она является результатом однократного уточнения Ричардсона, примененного к формуле трапеций. Следовательно, она пригодна до $L \approx 3 \div 4$. Проще применить не формулу Симпсона, а лишнее уточнение по Ричардсону для формулы трапеций.

Формулы Эйлера—Маклорена с использованием старших производных в граничных точках имеют высокий порядок точности и обычно эффективны на одномерных задачах. Однако произведение таких формул в многомерном случае записывается очень громоздко, а сами производные зачастую трудно вычислять. Поэтому такой метод практически не используется.

Одномерные формулы Гаусса с большим числом узлов оказались не слишком эффективными. Однако использование гауссово-сеточных квадратур кардинально улучшает точность. Из табл. 3.1 видно, что деление отрезка интегрирования всего на два интервала резко повысило точность гауссово-сеточного метода по сравнению с обычным методом Гаусса.

Оценим возможности применения произведения гауссово-сеточных квадратур для многомерных интегралов. Будем считать трудоемкость расчета допустимой вплоть до полного числа узлов $J = N^L \approx 10^9$. Для $L = 6$ это дает $N \approx 30$. Если ограничиться формулой Гаусса с пятью узлами (поскольку это последняя степень, для которой есть явная запись узлов и весов), то можно разбить каждую сторону прямоугольника на пять интервалов. Для функций с не слишком большими 10-ми производными это должно обеспечить высокую точность. Таким образом, произведение одномерных гауссово-сеточных квадратур должно оказаться эффективным способом вычисления интегралов в прямоугольном параллелепипеде до размерности $L \approx 6$.

Криволинейная граница. Рассмотрим двумерную область G с криволинейной гладкой границей Г (рис. 3.11). Максимальную и минимальную ординаты границы обозначим через β и α. Ограничимся для простоты случаем, когда граница образована

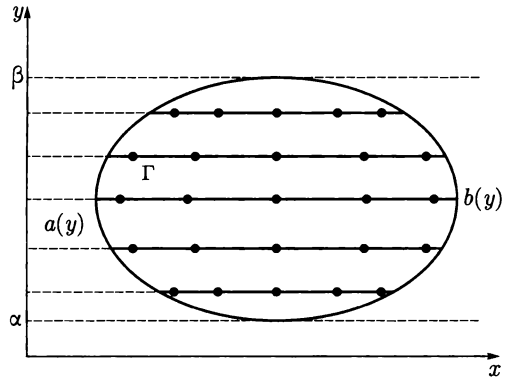


Рис. 3.11. Область с криволинейной границей

двумя однозначными функциями: левая — $x = a(y)$, правая — $x = b(y)$. В этом случае использование прямого произведения квадратурных формул вполне естественно.

Интеграл по области G можно записать в следующем виде:

$$U = \int_{\alpha}^{\beta} f(y)dy, \quad f(y) = \int_{a(y)}^{b(y)} u(x,y)dx. \tag{3.89}$$

Введем по оси ординат сетку $\{y_m\}$ $0 \leqslant m \leqslant M+1$; $y_0 = \alpha$, $y_{M+1} = \beta$. Проведем соответствующие хорды. Заменим интеграл по y некоторой квадратурной формулой.

Границами m-й хорды являются точки $a(y_m)$ и $b(y_m)$; длины всех хорд, вообще говоря, различны. Поэтому на каждой m-й хорде введем свой набор узлов $\{x_{nm}\}$, $1 \leqslant n \leqslant N$. Наиболее просто и выгодно взять число точек N на каждой хорде одинаковым и равным M, а в качестве самих точек $\{x_{nm}\}$ выбрать узлы квадратурной формулы Гаусса для веса ($\rho(x) \equiv 1$), при этом все узлы $\{x_{nm}\}$ лежат строго внутри области G. Тогда интегралы по хордам берутся по формуле Гаусса

$$f(y_m) = \sum_{n=1}^{N} c_{nm} u(x_{nm}, y_m),$$

здесь

$$x_{nm} = \frac{b(y_m) + a(y_m)}{2} + \frac{b(y_m) - a(y_m)}{2}\xi_n, \quad c_{nm} = \frac{b(y_m) - a(y_m)}{2}\gamma_n,$$

где ξ_n — нули многочленов Лежандра степени N; γ_n — соответствующие веса (для $N \leqslant 5$ они приведены в табл. 3.4). Напомним, что формула Гаусса точна для алгебраических многочленов от x степени не выше $2N-1$.

Очевидно, для интегрирования по y целесообразно применять квадратурную формулу, точную для алгебраических многочленов от y степени не выше $2N-1=2M-1$ (поскольку мы выбрали $N=M$). Однако интеграл по y имеет особенности даже для гладких функций $u(x,y)$ и гладкой границы Г. Причина в том, что $f(y_m)$ приблизительно пропорциональна длине хорды. При y, стремящемся к β или α, хорды вырождаются в точки, а их длины и интегралы по ним стремятся к нулю как корни квадратные:

$$f(y) \sim \sqrt{\beta - y} \quad \text{при} \quad y \to \beta, \quad f(y) \sim \sqrt{y - \alpha} \quad \text{при} \quad y \to \alpha.$$

В таком случае подынтегральную функцию в первом интеграле (3.89) можно представить в виде произведения: $f(y) = \varphi(y)\sqrt{(\beta - y)(y - \alpha)}$, где $\varphi(y)$ не имеет особенностей, а подкоренное выражение можно рассматривать как весовую функцию $\rho(y)$. Ей соответствует квадратурная формула Гаусса—Кристоффеля с многочленами Чебышева II рода:

$$\int_\alpha^\beta f(y)dy = \sum_{m=1}^{M} c_m \varphi(y_m), \quad \varphi(y) = \frac{f(y)}{\sqrt{(\beta - y)(y - \alpha)}},$$

здесь

$$y_m = \frac{\beta + \alpha}{2} + \frac{\beta - \alpha}{2} \cos \frac{\pi m}{M+1}, \quad 1 \leqslant m \leqslant M \quad (M = N),$$

$$c_m = \frac{\beta + \alpha}{2} + \frac{\beta - \alpha}{2} \tilde{\gamma}_m,$$

где $\tilde{\gamma}_m$ для $M \leqslant 5$ приведены в табл. 3.4.

Из изложенного следует, что произведение этих одномерных квадратур дает кубатурную формулу, точную для многочлена от x и y степени до $2N-1=2M-1$.

Замечания. 1. Описанный способ построен на основе формул Гаусса—Кристоффеля, поэтому применять для оценки погрешности метод Ричардсона здесь нельзя. Представление о точности можно получить, проводя последовательно вычисления с $N=M=1,2,\ldots$ Если визуально наблюдается стремление к пределу, то совпадающие знаки можно считать верными. Однако на практике дальше, чем $N=M=5$,

продвинуться трудно, потому что далее отсутствуют явные формулы для узлов и весов гауссовых квадратур.

2. Описанный метод не распространяется на случай большего числа измерений: уже в трехмерном случае помимо особенности с корнем квадратным возникают особенности с корнем четвертой степени. Для таких особенностей отсутствуют выражения узлов и весов формул Гаусса — Кристоффеля.

Метод Эйткена. Интеграл по области G с гладкой криволинейной границей Γ можно вычислять с использованием метода Эйткена. Для этого введем равномерную сетку по y с достаточно большим числом узлов $y_m = \alpha + m(\beta - \alpha)/M$, $0 \leqslant m \leqslant M$. Следует принимать $M = 2^k$: это обеспечивает возможность выбора из данной сетки последовательности вдвое сгущающихся вложенных сеток. На каждой хорде также введем равномерную сетку x_{nm} со своим числом узлов $x_{nm} = a(y_m) + n[b(y_m) - a(y_m)]/N_m$, $0 \leqslant n \leqslant N_m$, где в качестве N_m также целесообразно выбирать степень двойки.

Проведем серию вычислений интегралов по x вдоль каждой хорды по формуле трапеций с числами узлов $N_m, N_m/2, N_m/4, \ldots$ Читая эту серию в обратном направлении, получим серию расчетов на сгущающихся сетках с $r = 2$. Применим к ним рекуррентное уточнение по Ричардсону. Напомним, что для формулы трапеций порядок точности $p = 2$ и его повышение при каждом сгущении $\sigma = 2$. Выберем из промежуточных результатов наиболее точное значение $f(y_m)$ и оценку его погрешности ε_m.

Затем проведем серию вычислений интеграла от $f(y)$ также по формуле трапеций с числами узлов $M, M/2, M/4, \ldots$ Ранее отмечалось, что $f(y)$ имеет корневые особенности вблизи $y = \alpha$ и $y = \beta$. Поэтому применять метод Ричардсона к полученным результатам уже нельзя. Зато можно применить рекуррентный метод Эйткена и также найти наилучший результат и оценку его погрешности. Это и будет искомым ответом. Необходимо помнить, что при вычислении $f(y_m)$ роль ошибок округления будут играть величины ε_m, которые обычно много больше компьютерной точности 10^{-16}.

Описанный способ распространяется на многомерные интегралы по области с гладкой границей. Например, в трехмерном случае возникают секущие плоскости, выделяющие области с криволинейной гладкой границей. В каждой такой области проводятся хорды. Интегрирование по всем переменным выпол-

няется по методу трапеций. Рекуррентное уточнение по хордам проводится методом Ричардсона, но по всем остальным переменным — методом Эйткена. При этом нужно явно считать все сечения и их границы, поэтому реально этот метод можно применять в областях, границы которых описываются достаточно простыми явными формулами.

3.3.3. Статистические методы

Произведение одномерных квадратур с N узлами для вычисления L-кратного интеграла содержит N^L точек многомерной сетки. При увеличении размерности число точек и объем вычислений стремительно возрастают. Поэтому при $L > 6$ сеточные методы практически непригодны и требуются иные подходы. Они основаны на использовании случайных точек или других точек, близких к ним по свойствам.

Величина γ. Рассмотрим пример. Стрелок закрепляет винтовку в станке и выстреливает несколько раз подряд. Пули не попадут в одну и ту же точку из-за различных случайных факторов: масса пуль и зарядов немного отличается друг от друга, во время выстрела может подуть ветер и т. д. Вызванные этими факторами отклонения называют случайными, так как их причины стрелок не контролирует. Большинство пуль ляжет близко к точке прицеливания. Но будут и более значительные отклонения. Зависимость частоты возникновения отклонения от его величины называют распределением случайной величины. Точное математическое определение здесь не приводим (оно есть в курсах теории вероятностей).

Среди случайных величин особую роль играет случайная величина γ, равномерно распределенная на единичном отрезке. Ее формальное определение таково: *все значения величины γ принадлежат отрезку* $[0,1]$, *а плотность ее распределения* $\rho(\gamma) = 1$.

Последнее можно разъяснить так. Пусть концы интервала γ_1, γ_2 принадлежат отрезку $[0,1]$. Тогда вероятность того, что $\gamma_1 < \gamma < \gamma_2$ есть длина интервала $\gamma_2 - \gamma_1$ и не зависит от местоположения интервала (γ_1, γ_2) на единичном отрезке.

Возьмем достаточно большую выборку случайных чисел γ_j, $0 \leqslant j \leqslant J, J \gg 1$. Разобьем отрезок $[0,1]$ на равные половины, каждую половину — на $1/4$ и т. д. Вероятности попадания γ в каждую из половин отрезка одинаковы и равны 0,5. Поэтому примерно половина чисел последовательности $\{\gamma_j\}$ будет распо-

ложена в левой половине отрезка и примерно половина — в правой. В каждой четверти отрезка будет лежать примерно $1/4$ случайных точек и т. д. Это означает, что случайные точки будут равномерно разбросаны по единичному отрезку.

Теперь возьмем единичный квадрат $0 \leqslant x \leqslant 1$, $0 \leqslant y \leqslant 1$. Будем брать из последовательности $\{\gamma_j\}$ пары соседних чисел: $(\gamma_0, \gamma_1), (\gamma_1, \gamma_2), \ldots$ Каждую пару чисел можно рассматривать как координаты точки на плоскости. Аналогичными рассуждениями нетрудно убедиться, что эти точки будут равномерно распределены в единичном квадрате. Так же строятся случайные точки, равномерно распределенные в единичном L-мерном кубе. Для этого из последовательности $\{\gamma_j\}$ выбирают группы по L точек $(\gamma_{jL}, \gamma_{jL+1}, \ldots, \gamma_{jL+L-1}) = \gamma_j$, $j = 0, 1, \ldots$ Каждую группу рассматривают как координаты L-мерной точки.

Интеграл. Сначала рассмотрим одномерный интеграл по единичному отрезку:

$$U = \int_0^1 u(x)dx \tag{3.90}$$

и сравним его с такой суммой

$$U_J = \frac{1}{J+1} \sum_{j=0}^{J} u(\gamma_j), \tag{3.91}$$

где γ — равномерно распределенная случайная величина. Функцию $u(\gamma)$, зависящую от случайной величины γ, называют ***случайной функцией***, хотя вид самой функции u вполне определен.

Разобьем отрезок $[0, 1]$ на достаточно малые интервалы, но так, чтобы в каждый интервал попадало еще достаточно много точек последовательности $\{\gamma_j\}$. В каждом интервале значения $u(\gamma_j)$ будут почти совпадать между собой в силу малости интервала. Число таких значений примерно пропорционально длине интервала, а после деления на полное число точек $J+1$ оно примерно равно длине интервала. Поэтому вклад в сумму (3.91) от одного интервала будет близок к $u(x_{\text{ср}})dx$, где $x_{\text{ср}}$ — средняя точка интервала. В итоге сумма (3.91) будет мало отличаться от интегральной суммы для (3.90) и можно положить $U \approx U_J$. Такое равенство выполняется, как правило, тем лучше, чем больше число точек J. На языке статистики это означает, что интеграл (3.90) равен математическому ожиданию случайной функции $u(\gamma)$:

$$U = \mathsf{M}u(\gamma). \tag{3.92}$$

Сумма (3.91) является однократным разыгрыванием величины $\mathsf{M}u$ по выборке $\{\gamma_j\}$.

Описанный способ непосредственно переносится на вычисления интеграла в единичном кубе размерности L:

$$U = \int_0^1 \int_0^1 \dots \int_0^1 u(x_1, x_2, \dots, x_L) dx_1 dx_2 \dots dx_L = \mathsf{M}u(\gamma). \quad (3.93)$$

Математическое ожидание по-прежнему заменяют средним по выборке

$$\mathsf{M}u(\gamma) \approx U_J = \frac{1}{J+1} \sum_{j=0}^{J} u(\gamma_{jL}, \gamma_{jL+1}, \dots, \gamma_{jL+L-1}). \quad (3.94)$$

Никаких принципиальных трудностей при переходе ко многим измерениям не возникает. Однако на области более сложной формы, чем единичный куб (или прямоугольный параллелепипед), этот метод обобщить крайне сложно.

Погрешность. Ошибка в статистических методах также носит случайный характер. Среднее арифметическое по выборке U_J (3.91) само является новой случайной величиной. При единичном испытании (т. е. при выборе одного J и вычислении единственного соответствующего ему U_J) оно может даже заметно отличаться от математического ожидания $\mathsf{M}U$, хотя в большинстве случаев будет близко к нему (так опытный стрелок почти все пули кладет вблизи центра мишени, но и у него какой-то выстрел может уйти в «молоко»).

Поведение U_J определяется центральной предельной теоремой теории вероятности. Она утверждает, что:

1) математическое ожидание и дисперсия величины U_J связаны с математическим ожиданием $u(\gamma)$ соотношениями

$$\mathsf{M}U = \mathsf{M}u(\gamma), \quad \mathsf{D}U = \frac{1}{J+1}\mathsf{D}u(\gamma); \quad (3.95)$$

2) при $J \to \infty$ закон распределения величины U_J стремится к гауссовому (практически уже при $J \geqslant 30$ он достаточно близок к нему).

Для гауссова закона распределения вероятность отклонения единичного испытания от математического ожидания хорошо известна. Введем стандартное уклонение (*стандарт*)

$$\sigma = \sqrt{\mathsf{D}U}.$$

Вероятности того, что уклонение не превысит одного, двух или трех стандартов, равны соответственно

$$0{,}68 = 68\,\% \text{ для } 1\sigma, \quad 0{,}97 = 97\,\% \text{ для } 2\sigma \\ \text{и} \quad 0{,}995 = 99{,}5\,\% \text{ для } 3\sigma. \tag{3.96}$$

В статистических расчетах не удается достичь столь высокой точности, как в хороших сеточных. Обычно характерная погрешность составляет не менее 0,1 %, доходя до 10 % для наиболее трудных задач. Поэтому если работа не требует особой тщательности, то считают допустимым отклонение не более 1σ. При аккуратном выполнении физических и технических измерений в качестве оценки погрешности выбирают обычно 2σ. Если требуется высокая надежность, то в качестве допуска приводят величину 3σ. Вероятность заметно большего отклонения можно считать пренебрежимо малой.

Остается найти величину $\mathsf{D}U$. Она связана с $\mathsf{D}u(\gamma)$ соотношением (3.95). Воспользуемся известным соотношением из теории вероятностей:

$$\mathsf{D}u(\gamma) = \mathsf{M}u^2(\gamma) - [\mathsf{M}u(\gamma)]^2. \tag{3.97}$$

Второй член справа — это квадрат интеграла (3.93), приближенно вычисляемого по формуле (3.94). Но первый член справа точно равен интегралу от $u^2(x)$:

$$\mathsf{M}u^2(\gamma) = \int_0^1 \int_0^1 \dots \int_0^1 u^2(x_1, x_2, \dots, x_L) dx_1 dx_2 \dots dx_L.$$

Его тоже можно вычислить как среднее арифметическое. С учетом (3.95) это дает для искомой дисперсии величину

$$\sigma^2 = \mathsf{D}U \approx \sigma_J^2 = \\ = \frac{1}{J+1} \left\{ \frac{1}{J+1} \sum_{j=0}^{J} u^2(\gamma_j) - \left[\frac{1}{J+1} \sum_{j=0}^{J} u(\gamma_j) \right]^2 \right\}. \tag{3.98}$$

Расчет проводится следующим образом. Берут последовательность $\{\gamma_j\}$. Увеличивают значение J, и для каждого J вычисляют величины U_J и σ_J^2. (Для экономичности расчетов надо накапливать обе суммы в (3.98), а не вычислять их заново для каждого J.) Величина σ_J является приближенным значением стандарта σ и характеризует точность величины U_J.

Выражение в фигурной скобке в (3.98) близко к постоянной величине $\mathrm{D}u(\gamma)$. Поэтому выборочный стандарт равен

$$\sigma_J \approx \sqrt{\frac{\mathrm{D}u(\gamma)}{J}}$$

и при увеличении J стремится к нулю как $J^{-1/2}$. Вычисления прекращают, когда величина σ_J станет достаточно малой с точки зрения критерия (3.96).

Во все приведенные оценки не входила размерность пространства L, поэтому закон убывания погрешности $\sigma_J \sim J^{-1/2}$ справедлив для любой размерности пространства. Это позволяет применять статистические методы в отличие от сеточных методов к задачам огромной размерности. При малых размерностях $L \leqslant 6$ они менее экономичны, чем сеточные. Но при $L > 6$ требуют меньшего числа точек, чем сеточные, и с увеличением L их преимущество возрастает.

По скорости убывания σ_J метод статистических испытаний напоминает сеточный метод, имеющий порядок точности $p = 1/2$. Однако это стремление к нулю является не асимптотически точным, а лишь вероятностным, как отмечалось ранее. Поэтому применять метод Ричардсона к вычислениям с разными J совершенно недопустимо.

Псевдослучайные числа. Случайные величины являются математической абстракцией. Стопроцентно случайных величин в природе нет: те неконтролируемые факторы, которые отмечались в начале подраздела, можно было бы учесть при более внимательном рассмотрении как закономерные. Поэтому радиотехнические и другие физические датчики случайных чисел на самом деле не вполне надежны. При не очень больших J даваемые ими последовательности обычно удовлетворительны по качеству. Однако из-за малости J точность расчета может оказаться недостаточной. Но при больших J начнет сказываться неслучайность последовательности. Ошибки такого рода крайне трудно обнаружить по внешним проявлениям. Это делает их особенно опасными.

Математики с конца 1940-х годов занимались поисками способов чисто математического построения случайных чисел. Предлагалось много «кухонных» рецептов, например следующий: выберем какое-нибудь иррациональное число, а его последовательные десятичные знаки будем считать случайными целыми числами, равномерно распределенными на дискретном множестве $0 \div 9$. В качестве обоснования приводился такой ар-

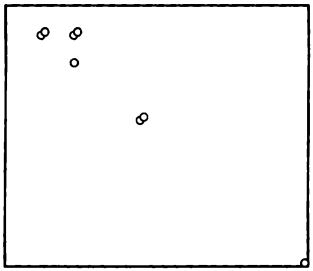


Рис. 3.12. Точки в квадрате, построенные по десятичным знакам числа e

гумент: если бы здесь существовала закономерность, то число свелось бы к периодической десятичной дроби, т. е. было бы рациональным.

Такими рецептами недопустимо пользоваться. Например, возьмем заведомо иррациональное число

$$e = 2{,}718281828459045\ldots$$

Рассмотрим последовательные пары знаков 27,18,28... как двузначные целые числа из дискретного множества $0 \div 99$. Видно, что слишком часто попадаются близкие и даже одинаковые числа. Кроме того, лишь одно число 90 лежит в правой половине отрезка, а остальные семь чисел — в левой. Еще хуже будет, если в каждой паре цифр первую и вторую цифры рассматривать как координаты (x, y) в квадрате со стороной 10. При этом все точки не заполняют квадрат, а группируются около верхней половины его диагонали (рис. 3.12). Подобные точки явно непригодны для статистических вычислений.

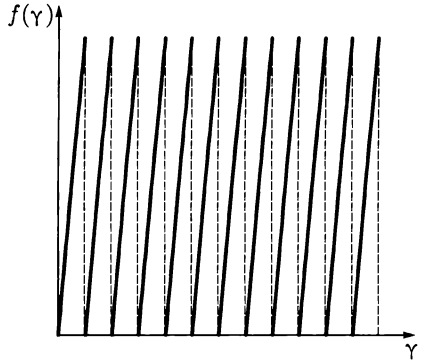
Более успешным оказалось конструирование таких функций $f(x)$, чтобы порождаемые ими последовательности $\gamma_{j+1} = f(\gamma_j)$ обладали бы свойствами случайных величин. Такие числа называют *псевдослучайными*. Качество таких последовательностей проверяют различными статистическими тестами. Например, для равномерно распределенных случайных чисел γ выборочное математическое ожидание должно быть близко к 0,5, парные корреляции любых чисел последовательности должны быть почти нулями, L-мерные случайные точки должны приблизительно равномерно заполнять единичный L-мерный куб и т. д.

Сконструировать хорошую функцию $f(x)$ трудно. В самом деле, построим последовательность $\gamma_{j+1} = f(\gamma_j)$ и будем рассматривать соседние пары чисел как координаты точки в единичном квадрате. Тогда все псевдослучайные точки будут лежать на кривой $y = f(x)$. Для того чтобы эти точки можно было считать равномерно распределенными в квадрате, график функции $f(x)$ должен равномерно и всюду плотно заполнять единичный квадрат!

Наиболее удачным для компьютерной реализации оказался алгоритм Д. Х. Лемера. Он был предложен для компьютеров большой разрядности. Его можно записать в следующем виде:

$$\gamma_{j+1} = [A\gamma_j],$$

Рис. 3.13. Алгоритм Лемера



где выражение в квадратных скобках обозначает целую часть, а константа A — целое число, близкое к наибольшему представимому на компьютере. Для малого $A = 10$ график такой функции показан на рис. 3.13. Это разрывная функция с числом участков, равным A. Очевидно, что при огромных A такая ломаная будет «почти» заполнять квадрат.

Реализацию алгоритма Лемера удобно проводить, используя операции с целыми числами m, M, G:

$$m_{j+1} = Gm_j (\mathrm{mod} M); \quad \gamma_{j+1} = \frac{m_j}{M}. \tag{3.99}$$

Здесь числа M, G близки к наибольшим представимым на компьютере целым числам, причем при их записи в двоичном коде расположение нулей и единиц должно выглядеть достаточно случайным. Значения G, M, m_0 специально подбирались для каждого типа компьютеров и тщательно тестировались. Очевидно, что такая последовательность не может быть абсолютно случайной. Члены последовательности (3.99) не превосходят M, рано или поздно последовательность зацикливается и ее период не превышает M. При неудачном выборе G, M, m_0 период может оказаться существенно меньше M.

Этот алгоритм хорошо работал на многоразрядных компьютерах в 50—70-е годы XX в., однако при неудачном выборе G, M, m_0 получались последовательности с более короткими периодами или другими неприятными свойствами. Так, широко распространенный датчик RANDO вместо равномерного заполнения для трехмерного куба группировал точки вблизи нескольких наклонных плоскостей. Отсюда видно, что можно пользоваться только тщательно оттестированными наборами G, M, m_0.

Первые персональные компьютеры 1980-х годов были 16-разрядными. Для них период последовательности оказывался

Таблица 3.9

Параметры модифицированного алгоритма Лемера

G	M	m_0
171	30 269	5
172	30 307	11
170	30 323	17

слишком коротким: $2^{16} \approx 65\,000$. Поэтому Б. А. Вичман и И. Д. Хилл (1982) предложили следующее видоизменение алгоритма Лемера: параллельно строились три последовательности

$$m_{j+1}^k = G^k m_j^k (\mathrm{mod}\, M^k), \quad k = 1, 2, 3, \tag{3.100}$$

и полагалось

$$\gamma_{j+1} = \left[\frac{m_j^1}{M^1} + \frac{m_j^2}{M^2} + \frac{m_j^3}{M^3} \right], \tag{3.101}$$

где квадратные скобки обозначают операцию взятия целой части. Рекомендуемый набор коэффициентов приведен в табл. 3.9. Он тщательно тестирован на длину последовательности 10^7 чисел. Период последовательности при этих константах не превышает $\approx 3 \cdot 10^{13}$. Если использовать эту последовательность для создания L-мерных случайных чисел, то из $\approx 10^7$ членов последовательности (3.101) можно построить $\approx 10^{7/L}$ надежных случайных точек.

В последующие годы появились многоразрядные персональные компьютеры, и можно было бы вернуться к одной последовательности (3.99). Но модифицированный алгоритм (3.101) оказался настолько удобным, что им по-прежнему продолжают пользоваться.

Квазислучайные точки. Построены специальные последовательности точек, совсем не случайных, но обладающих некоторыми свойствами случайных точек. Такие последовательности при вычислении интегралов имеют тот же закон убывания стандартного уклонения $\sigma = C(L) J^{-1/2}$, но с меньшим коэффициентом $C(L)$. Тем самым, точность получается выше. Наиболее хорошие результаты дают $\mathrm{ЛП}_\tau$-последовательности, предложенные И. М. Соболем (в зарубежной литературе их называют последовательностями Соболя). Построение этих точек поясним на двумерном случае.

Первая точка ставится в начало координат, вторая точка — в центр единичного квадрата так, что ее проекции на каждую ось равны $1/2$. Следующие две точки ставят так, чтобы одна имела проекции $1/4$ и $3/4$, а другая — наоборот. Следующие четыре точ-

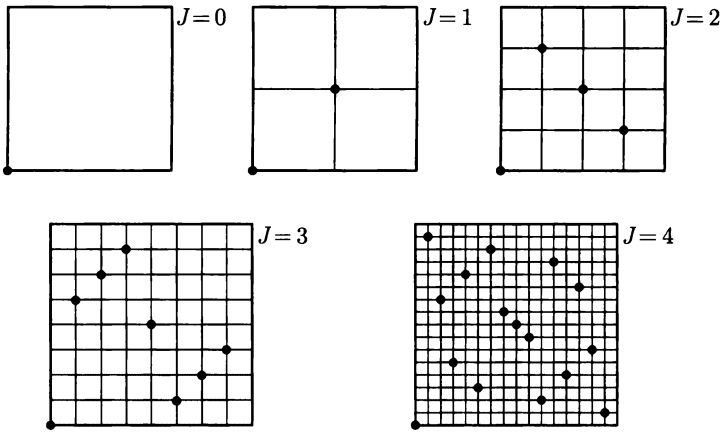


Рис. 3.14. Квазислучайные последовательности Соболя

ки должны иметь проекции, кратные $1/8$, причем одноименные координаты любых двух точек последовательности не должны совпадать. Потом добавляются еще восемь точек с проекциями, кратными $1/16$, и т. д. (рис. 3.14).

Такие точки дают существенное уменьшение дисперсии при числе измерений $L \leqslant 9$. С увеличением числа измерений коэффициент $C(L)$ возрастает и преимущество квазислучайных последовательностей становится не столь существенным. Однако до размерности $L = 13$ выигрыш заметен. Правила выбора этих точек для $L \leqslant 13$ приведены в монографии И. М. Соболя, 1973.

Особенно эффективны такие последовательности, если брать число точек $J = 2^k$, где $k \geqslant 10$ — целое число (при $k < 10$ число точек недостаточно для обеспечения хорошей точности). Это позволяет уменьшать объем расчета в несколько раз по сравнению с псевдослучайными точками.

Разбиения. Для размерности $L = 7 \div 9$ еще более эффективен следующий способ. Разобьем каждую сторону многомерного куба на k равных частей. Тогда весь куб разобьется на k^L одинаковых кубиков. В каждом таком кубике (рассматривая его как единичный) поставим одну псевдослучайную точку с помощью обобщенного алгоритма Лемера (3.101). Кроме того, возьмем еще вторую точку, симметричную первой относительно центра кубика. В единичном кубике это значит, что если координата первой точки есть γ, то соответствующая координата симметричной точки будет $1 - \gamma$.

Таким образом, в исходном кубе получается $J = 2k^L$ точек. По ним вычисляется среднее арифметическое (3.91). Если $u(x)$

достаточно гладкая, то стандартное уклонение в этом методе подчиняется закону $\sigma \sim J^{-1/2-2/L}$. Таким образом, дисперсия с увеличением числа точек убывает быстрее, чем для псевдослучайных или квазислучайных последовательностей. Однако при $L \geqslant 10$ этот способ по эффективности сравнивается с последовательностями Соболя.

Глава 4

ИНТЕРПОЛЯЦИЯ

4.1. ИНТЕРПОЛЯЦИОННЫЙ МНОГОЧЛЕН

4.1.1. Задачи интерполяции

Пусть функция $u(x)$ задана таблицей, т. е. на сетке $\{x_n, n = 0, 1, \ldots\}$ известны значения $u_n = u(x_n)$. Как восстановить ее значения в произвольной точке x? Разумеется, при этом мы должны требовать достаточно простого поведения $u(x)$: функция не должна иметь «всплесков» между соседними узлами. Математически это означает, что $u(x)$ должна иметь достаточное количество старших производных, не слишком больших по величине.

Простейший способ состоит в следующем. Выберем систему линейно независимых функций $\{\varphi_m(x), m = 0, 1, \ldots\}$. Линейную комбинацию таких функций называют обобщенным многочленом $\Phi(x)$. Попробуем приближенно заменить $u(x)$ обобщенным многочленом:

$$u(x) \approx \Phi_N(x) \equiv \sum_{m=0}^{N} c_m \varphi_m(x). \tag{4.1}$$

Коэффициенты c_m выберем из условия, чтобы обобщенный многочлен $\Phi_N(x)$, содержащий $N+1$ коэффициент, точно передавал табулированные значения функции в $(N+1)$-м узле:

$$\sum_{m=0}^{N} c_m \varphi_m(x_n) = u_n, \quad 0 \leqslant n \leqslant N. \tag{4.2}$$

Такой способ приближения функции называют интерполяцией.

Коэффициенты c_m находят из решения линейной системы (4.2). Для ее разрешимости необходимо, чтобы $\det[\varphi_m(x_n)] \neq 0$.

Заметим, что не всякая полная система даже линейно независимых функций пригодна для интерполяции. Например, четные степени $\varphi_m(x) = x^{2m}$, $m = 0, 1, \ldots$, линейно независимы, а их система полна в классе четных функций. Возьмем две функции $\varphi_0(x) = 1$, $\varphi_1(x) = x^2$. И выберем точки -1 и 1. При этом определитель системы (4.2) обращается в нуль, т. е. интерполяция невозможна.

Алгебраический многочлен. Выберем в качестве базисных функций систему всех степеней

$$\varphi_m(x) = x^m, \quad m = 0, 1, 2, \ldots$$

Будем считать все узлы x_n попарно различными. Докажем, что интерполяционный алгебраический многочлен $P_N(x)$ в этом случае является единственным. Предположим, что существуют два интерполяционных многочлена $P_N(x)$ и $\tilde{P}_N(x)$. Они совпадают в $N+1$ точках x_n. Тогда их разность $P_N(x) - \tilde{P}_N(x)$ также является многочленом степени не выше N и обращается в нуль в $(N+1)$-й точке, т. е. имеет $N+1$ корень. Последнее невозможно, поскольку у такого многочлена имеется не более N корней. Это доказывает единственность алгебраического интерполяционного многочлена.

Замечание. Помимо алгебраического многочлена, единственность обеспечивают и другие системы функций. В радиотехнических расчетах нередко используют интерполяцию тригонометрическими многочленами (так называемые формулы Бесселя); она будет описана в 5.2 как частный случай среднеквадратичной аппроксимации. В задачах квантово-механического расчета молекул используют систему экспонент $\varphi_m(x) = \exp(\alpha_m x)$, $m = 0, 1, \ldots$, но соответствующий алгоритм является специфическим.

4.1.2. Многочлен Ньютона

Алгебраический интерполяционный многочлен единственный, но существует много форм его записи. Они были предложены потому, что давали некоторые преимущества в каких-то частных случаях: например, если сетка равномерная или заранее известно число узлов, которое будет взято для построения многочлена, и т. п. Далее будет приведена форма Ньютона; она удобна тем, что пригодна на произвольной неравномерной сетке, позволяет увеличивать или уменьшать число узлов в ходе расчета, а также апостериорно оценивать достигнутую точность.

Разделенные разности. Определим разделенные разности первого, второго и так далее порядков функции $u(x)$ следующим образом:

$$u(x_0, x_1) = \frac{u(x_0) - u(x_1)}{x_0 - x_1};$$
$$u(x_0, x_1, x_2) = \frac{u(x_0, x_1) - u(x_1, x_2)}{x_0 - x_2}; \tag{4.3}$$
$$u(x_0, x_1, x_2, x_3) = \frac{u(x_0, x_1, x_2) - u(x_1, x_2, x_3)}{x_0 - x_3}$$

и т. д. Общий закон легко виден: справа в числителе стоят две разделенные разности предыдущего порядка со сдвинутыми на 1 индексами, а в знаменателе стоит разность крайних значений аргумента. Процесс построения разделенных разностей можно продолжать, пока у нас хватает точек. Заметим, что размерность разделенной разности k-го порядка (4.3) совпадает с размерностью $u^{(k)}(x)$.

Возьмем многочлен N-й степени $P_N(x)$ и $N + 1$ точку x_n, $0 \leqslant n \leqslant N$. Добавим к этим точкам искомую точку x, поместив ее в таблице впереди точки x_0 (хотя само значение x может быть любым). Построим разделенные разности этого многочлена по расширенной системе точек. Первая разделенная разность равна

$$P(x, x_0) = \frac{P(x) - P(x_0)}{x - x_0}. \tag{4.4}$$

Здесь числитель обращается в нуль при $x = x_0$, поэтому многочлен $P(x) - P(x_0)$ нацело делится на $x - x_0$. Следовательно, первая разделенная разность многочлена N-й степени является многочленом $(N - 1)$-й степени от x. Но правая часть (4.4) не меняется при перестановке x и x_0. Тем самым $P(x, x_0)$ является многочленом $(N-1)$-й степени относительно x_0. Значит, это многочлен по каждой из переменных, симметричный относительно перестановки переменных.

Вторая разделенная разность равна

$$P(x, x_0, x_1) = \frac{P(x, x_0) - P(x_0, x_1)}{x - x_1}. \tag{4.5}$$

Если $x = x_1$, то числитель также обращается в нуль, значит, он нацело делится на $x - x_1$. При этом получается многочлен степени $N - 2$ по каждому из трех аргументов. Он симметричен

относительно перестановки аргументов. Аналогично третья разделенная разность

$$P(x, x_0, x_1, x_2) = \frac{P(x, x_0, x_1) - P(x_0, x_1, x_2)}{x - x_2} \tag{4.6}$$

есть многочлен степени $N - 3$ от каждого из четырех аргументов.

Продолжим этот процесс. Когда подключается $(N-1)$-я точка, получается $(N-1)$-я разделенная разность, являющаяся многочленом нулевой степени, т. е. константной. Поэтому при подключении N-й точки правая часть обратится в нуль:

$$P(x, x_0, \ldots, x_N) = 0. \tag{4.7}$$

Таким образом, процесс построения разделенных разностей для многочлена исчерпывается: он не может использовать дальнейшие точки таблицы.

Форма Ньютона. Восстановим интерполяционный многочлен по его разделенным разностям. Для этого перепишем формулы (4.4) — (4.7) в следующем виде:

$$P(x) = P(x_0) + (x - x_0)P(x, x_0);$$
$$P(x, x_0) = P(x_0, x_1) + (x - x_1)P(x, x_0, x_1);$$
$$P(x, x_0, x_1) = P(x_0, x_1, x_2) + (x - x_2)P(x, x_0, x_1, x_2), \ldots;$$
$$P(x, x_0, x_1, \ldots, x_N) = 0.$$

Подставив здесь каждую последующую формулу в предыдущую, получим выражение многочлена через его разделенные разности:

$$P(x) = P(x_0) + \\ + \sum_{n=1}^{N} (x - x_0)(x - x_1)\ldots(x - x_{n-1})P(x_0, x_1, \ldots, x_n). \tag{4.8}$$

Здесь разделенные разности в правой части содержат только табулированные узлы; узел x в них не входит. Поскольку эти разделенные разности выражаются через табулированные выражения многочлена, это позволяет восстановить многочлен в $(N + 1)$-м узле.

Интерполяционный многочлен по определению совпадает с функцией $u(x)$ в $(N + 1)$-м выбранном узле. Тем самым его

разделенные разности совпадают с разделенными разностями (4.3) функции $u(x)$. Поэтому для построения интерполяционного многочлена (4.8) подставим разделенные разности (4.3). Это дает

$$\begin{aligned} u(x) &\approx u(x_0) + (x - x_0)u(x_0, x_1) + \\ &+ (x - x_0)(x - x_1)u(x_0, x_1, x_2) + \\ \ldots &+ (x - x_0)(x - x_1)\ldots(x - x_{N-1})u(x_0, x_1, \ldots, x_N) = \\ &= u(x_0) + \sum_{n=1}^{N} u(x_0, x_1, \ldots, x_n)\omega_n(x), \\ \omega_n(x) &= \prod_{k=0}^{n-1} (x - x_k). \end{aligned} \tag{4.9}$$

Видно, что формула (4.9) применима на произвольной неравномерной сетке, а число членов можно увеличивать или уменьшать, если исходные таблицы содержат достаточно много точек.

Для вычислений по формуле (4.9) берется исходная таблица из $N + 1$ узла x_n, $0 \leqslant n \leqslant N$. По каждой паре соседних узлов вычисляется разделенная разность первого порядка. Таких разностей будет N. По каждой паре соседних разделенных разностей первого порядка вычисляется разделенная разность второго порядка; их будет $N - 1$. Продолжив этот процесс, доходим до единственной разделенной разности N-го порядка. Таким образом, таблица разделенных разностей будет треугольной. Но для окончательных вычислений достаточно хранить лишь одну верхнюю строку разделенных разностей всех порядков, входящих в формулу (4.9). Остальная часть таблицы является промежуточным результатом и служит только для вычисления этой строки.

Пример 4.1. Составим набор формул для вычисления синуса в первой четверти. Расставим узлы равномерно с нулевого по N-й. При этом удобнее записать функцию как $u(x) = \sin(90°x/N)$, $x \in [0, N]$. Узлы при этом окажутся целыми числами $x_n = n$, $0 \leqslant n \leqslant N$. Вычислим для каждого N треугольные таблицы разделенных разностей (для $N = 3$ данные приведены в табл. 4.1). В каждой таблице возьмем верхнюю косую строку разделенных разностей возрастающего порядка и получим следующие интерполяционные формулы:

Таблица 4.1

Разделенные разности

$n = x_n$	u_n	Разделенные разности		
		I	II	III
0	0,0000			
		0,5000		
1	0,5000		−0,0670	
		0,3660		−0,0163
2	0,8660		0,1160	
		0,1340		
3	1,0000			

$$
\begin{gathered}
N = 1, \quad \sin(90°x) \approx 0 + 1{,}00(x-0), \\
x \in [0,1], \quad \delta = 0{,}24; \\
N = 2, \quad \sin(45°x) \approx 0 + 0{,}707(x-0) - \\
- 0{,}207(x-0)(x-1), \quad x \in [0,2], \quad \delta = 0{,}028; \\
N = 3, \quad \sin(30°x) \approx 0 + 0{,}5000(x-0) - \\
- 0{,}0670(x-0)(x-1) - 0{,}0163(x-0)(x-1)(x-2), \qquad (4.10) \\
x \in [0,3], \quad \delta = 0{,}0028; \\
N = 4, \quad \sin(22{,}5°x) \approx 0 + 0{,}38268(x-0) - \\
- 0{,}02913(x-0)(x-1) - 0{,}00823(x-0)(x-1)(x-2) + \\
+ 0{,}00068(x-0)(x-1)(x-2)(x-3), \\
x \in [0,4], \quad \delta = 0{,}00025.
\end{gathered}
$$

Для проверки точности сосчитаем по этим формулам значения синуса с шагом 1° и сверим с подробными тригонометрическими таблицами. Вычислим погрешности δ этих формул. Они также приведены в (4.10). Видно, что увеличение N на единицу улучшает точность интерполяции примерно в десять раз. Формула (4.10) с $N = 4$ может заменить таблицу значений синуса с тремя значащими цифрами. Аналогичная формула с $N = 5$ может уже заменить таблицу с четырьмя десятичными знаками.

4.1.3. Погрешность

Получим строгую оценку погрешности интерполяционного многочлена Ньютона. Для этого потребуем, чтобы существовала непрерывная ограниченная производная $u^{(N+1)}(x)$: $|u^{(N+1)}(x)| \leqslant \leqslant M_{N+1}$. Обозначим погрешность $\delta(x) = u(x) - P_N(x)$. По опре-

делению, погрешность обращается в нуль во всех узлах интерполяции: $\delta(x_n) = 0$, $0 \leqslant n \leqslant N$. Этим погрешность похожа на многочлен $\omega_{N+1}(x)$ (4.9)

$$\omega_{N+1}(x) = \prod_{k=0}^{N} (x - x_k). \tag{4.11}$$

Поэтому будем искать погрешность в следующем виде:

$$\delta(x) \equiv u(x) - P_N(x) = q(x)\omega_{N+1}(x). \tag{4.12}$$

Свойства функции $q(x)$ нам неизвестны. Введем вспомогательную функцию

$$r(x) = u(x) - P_N(x) - q(\xi)\omega_{N+1}(x), \tag{4.13}$$

здесь ξ рассматривается не как переменная, а как параметр, принадлежащий отрезку $\xi \in [\alpha, \beta]$, $\alpha = \min(x_0, x)$, $\beta = \max(x_N, x)$.

Исследуем поведение $r(x)$ также на отрезке $x \in [\alpha, \beta]$. Функция $u(x)$ дифференцируема $N + 1$ раз, а многочлены $P_N(x)$ и $\omega_{N+1}(x)$ дифференцируемы любое число раз. Поэтому $r(x)$ также дифференцируема $N + 1$ раз; дифференцируемости $q(\xi)$ не требуется, так как здесь ξ — параметр, по которому дифференцирование не производится.

На отрезке $[\alpha, \beta]$ функция $r(x)$ обращается в нуль $N + 2$ раз: в $(N + 1)$-м узле $x = x_n$, $0 \leqslant n \leqslant N$, в силу условий интерполяции, а также в точке $x = \xi$ в силу определения функции. Как известно, между каждой парой нулей $r(x)$ лежит нуль $r'(x)$, следовательно, на $[\alpha, \beta]$ имеется $N+1$ нуль $r'(x)$. Аналогично между каждой парой нулей $r'(x)$ лежит нуль $r''(x)$, значит, на $[\alpha, \beta]$ имеется N нулей $r''(x)$ и т. д. Продолжая этот процесс, получим, что на $[\alpha, \beta]$ имеется по меньшей мере один нуль $r^{(N+1)}(x)$. Обозначим эту точку через x_*.

Продифференцируем $N + 1$ раз соотношение (4.13), подставим в него x_* и учтем, что $P_N^{(N+1)}(x) = 0$, а $\omega_{N+1}^{(N+1)} = (N + 1)!$, поскольку это многочлены степеней N и $N + 1$ соответственно. Тогда получим

$$r^{(N+1)}(x_*) = 0 = u^{(N+1)}(x_*) - q(\xi)(N + 1)!$$

Отсюда получаем $q(\xi) = u^{(N+1)}(x_*)/(N + 1)!$ Тем самым $|q(x)| \leqslant \leqslant M_{N+1}/(N + 1)!$ С учетом этого неравенства соотношение (4.12) превращается в оценку

$$|\delta(x)| \equiv |u(x) - P_N(x)| \leqslant \frac{M_{N+1}}{(N + 1)!}|\omega_{N+1}(x)|. \tag{4.14}$$

Оценка (4.14) является строгой мажорантной априорной оценкой погрешности. Заметим, что величину M_{N+1} достаточно определять только по отрезку $[\alpha, \beta]$, так как в ходе доказательства за пределы этого отрезка не выходим.

Экстраполяция. В оценку (4.14) входит многочлен $\omega_{N+1}(x)$, который сравнительно невелик, если $x \in [x_0, x_N]$, т. е. лежит между узлами интерполяции. Этот случай называют *интерполяцией* в узком смысле этого слова.

Однако вне отрезка $[x_0, x_N]$ многочлен $\omega_{N+1}(x)$ быстро возрастает, так что погрешность может стать большой. Этот случай называют *экстраполяцией*. Видно, что экстраполяцией, особенно далеко за границы исходного отрезка, пользоваться опасно.

Порядок точности. Ограничимся случаем собственно интерполяции: $x \in [x_0, x_N]$. Величина погрешности зависит от расположения узлов. Наиболее часто приходится интерполировать функции, табулированные на равномерной сетке с шагом h. Тогда многочлен $\omega_{N+1}(x)$ содержит $N+1$ сомножитель, причем каждый не превышает Nh. Подстановка этой мажоранты в (4.14) ухудшает оценку, зато упрощает ее вид

$$|\delta(x)| < \frac{M_{N+1}}{(N+1)!}(Nh)^{N+1} = O(h^{N+1}). \tag{4.15}$$

Погрешность убывает, как $(N+1)$-я степень шага при $h \to 0$, т. е. имеет $(N+1)$-й порядок точности. Эта оценка показывает, что для получения хорошей точности интерполяции надо брать достаточно подробную сетку.

Оценку (4.15) можно существенно улучшить, если более детально учесть структуру экстремумов многочлена $\omega_{N+1}(x)$ и местоположение точки x. Если x расположена между средними узлами, то погрешность заметно меньше, чем при расположении x вблизи крайних узлов. Поэтому если мы выбираем узлы интерполяции из подробной таблицы, то следует выбирать примерно одинаковое число узлов справа и слева от искомой точки x.

Апостериорная оценка. Интерполяционный многочлен (4.9) содержит слагаемые с многочленами $\omega_{n-1}(x)$. Последний стоящий в сумме многочлен есть $\omega_N(x)$. Нетрудно заметить, что оценка погрешности (4.14) является мажорантной оценкой следующего слагаемого. Отсюда видно, что каждое слагаемое интерполяционного многочлена (4.9) является асимптотически точной оценкой погрешности суммы всех предыдущих слагаемых. Другими словами, погрешность интерполяционного много-

члена Ньютона примерно равна первому отброшенному слагаемому.

Разумеется, эти рассуждения справедливы для достаточно малого шага h. Зато такая оценка является апостериорной. Благодаря структуре формы Ньютона можно прибавлять к расчету один узел за другим и оценивать величины получаемых при этом членов. Если они быстро убывают, то ряд хорошо сходится. Можно остановиться, когда очередной член будет меньше требуемого уровня погрешности. Если же слагаемые убывают медленно, то шаг слишком велик и не позволяет получить надежной сходимости.

Так, в примере 4.1 при $N = 4$ видно, что разделенные разности быстро убывают с увеличением порядка, а многочлены $\omega_{n-1}(x)$ невелики. Поэтому слагаемые убывают достаточно быстро, обеспечивая удовлетворительную точность. При $N = 3$ убывание значительно хуже, а при $N = 2$ убывания фактически нет.

Сходимость. Насколько высокую точность можно получить с помощью интерполяционного многочлена? Пример 4.1 показывает, что точность неограниченно возрастает с увеличением степени многочлена. Однако это не всегда так. Рунге в 1901 г. предложил следующий пример:

$$u(x) = (1 + 25x^2)^{-1}, \quad x \in [-1, 1]; \tag{4.16}$$

на этом отрезке $u(x)$ имеет непрерывные ограниченные производные сколь угодно высокого порядка, так что условия справедливости оценки погрешности (4.14) выполнены. Рунге дробил отрезок $[-1, 1]$ на N интервалов и при каждом дроблении строил интерполяционный многочлен N-й степени. Наблюдалась странная картина. При увеличении N многочлены сходились к $u(x)$ в средней части отрезка, примерно при $|x| < 0{,}57$. Но на краях отрезка сходимость отсутствовала. Между узлами интерполяции многочлены сильно отличались от $u(x)$ (рис. 4.1.), а их экстремумы неограниченно возрастали при $N \to \infty$.

Этот пример показывает, что высокими степенями многочлена пользоваться опасно. Для хороших оценок при $N \to \infty$ надо требовать не просто ограниченности производных, а равномерной по N ограниченности. Последнее требование не всегда легко проверить.

Однако оценкой сходимости (4.15) можно пользоваться, если фиксировать число узлов и степень многочлена N и устремлять h к нулю. При этом погрешность неограниченно уменьшается

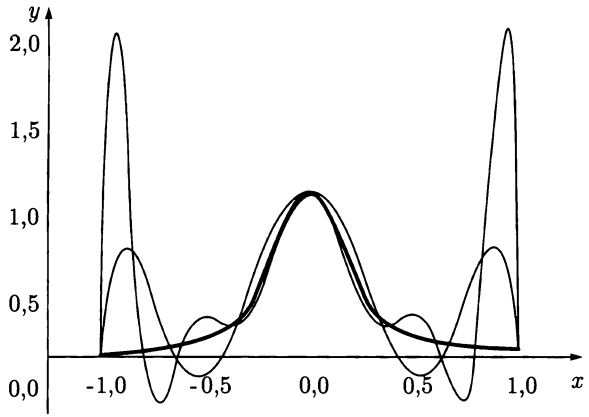


Рис. 4.1. Иллюстрация примера Рунге (жирная линия — функция (4.16); тонкие линии — интерполяционные многочлены для $N = 6$ и $N = 10$)

как $O(h^{N+1})$. При расчете по таблицам это означает, что мы переходим к таблицам со все более мелким шагом h, но при этом выбираем одно и то же число узлов $N + 1$. Напомним, что при этом рекомендуется брать эти узлы симметрично справа и слева относительно искомой точки x. Разумеется, этот процесс реально ограничен тем, насколько подробные таблицы нам доступны.

На практике поступают следующим образом. Ограничиваются не слишком высокой степенью многочлена $N \leqslant 4 \div 6$. Затем проводят расчет по имеющимся таблицам, выбирая в качестве шага h сначала расстояние между соседними узлами таблицы, потом вдвое бóльшую величину (т. е. берем точки таблицы через одну), затем в три-четыре раза бóльшую величину. Смотрят, при каком размере шага h слагаемые интерполяционного многочлена убывают достаточно быстро, а последнее слагаемое не превышает допустимой погрешности. По этим таблицам и с этим N производят все дальнейшие расчеты.

Ошибки округления. При вычислении разделенных разностей на сетке с малым шагом h происходит вычитание близких величин. При этом теряются значащие цифры, т. е. появляется заметная ошибка округления. С увеличением порядка разделенной разности ошибки округления стремительно нарастают. Чтобы сохранить приемлемое число знаков, вычисления надо проводить с высокой разрядностью. Но даже при 64-разрядных вычислениях разделенные разности 10 — 15-го порядков могут оказаться мало надежными. Это необходимо помнить при самосто-

ятельном использовании разделенных разностей (например, при разностном вычислении производных).

Однако при вычислении значений $u(x)$ по интерполяционному многочлену n-е разделенные разности умножаются на многочлены $\omega_n(x) - h^n$. Сам этот многочлен мал, поэтому вклад погрешности этого слагаемого в общую сумму будет соответствовать точности, с которой заданы табулированные значения u_n. Следовательно, ненадежность разделенных разностей высокого порядка не скажется на погрешности самой интерполяции.

4.1.4. Обратная интерполяция

В практике часто требуются обратные функции. Но даже если известно явное выражение функции $u(x)$, найти явное выражение для обратной функции $x(u)$ далеко не всегда возможно. Однако численно найти обратную функцию очень легко. Для этого выберем достаточно подробную сетку узлов $\{x_n\}$, покрывающую весь необходимый диапазон изменения аргумента. Вычисляя прямую функцию, составляем таблицу $u_n = u(x_n)$. Меняем столбцы этой таблицы местами и получаем таблицу (u_n, x_n). Если рассматривать u как аргумент, а x — как функцию, то это будет таблица значений обратной функции.

Тем самым обратная функция $x(u)$ табулирована. В узлах u_n она известна. Для промежуточных значений u ее можно найти, строя по обратной таблице интерполяционный многочлен Ньютона.

Этот прием называется обратной интерполяцией.

Решение уравнения. Обратную интерполяцию можно использовать для решения нелинейного уравнения $u(x) = 0$. Пусть уже известна неширокая окрестность корня x_*. Выберем в этой окрестности три — шесть узлов x_n. Составим таблицу (x_n, u_n), переставим столбцы и построим интерполяционный многочлен $P(u)$. Подставляя в него $u = 0$, получим приближенное значение искомого корня: $x_* \approx P(0)$.

Пример 4.2. Решим уравнение

$$u(x) \equiv (1 + x)e^{0{,}5x} - 2{,}5 = 0.$$

Не утруждая себя поиском малой окрестности корня, возьмем наугад три довольно удаленные от корня точки и составим таблицу обратной функции и ее разделенных разностей (табл. 4.2). Произведем вычисления, используя интерполяционный многочлен второй степени:

Таблица 4.2

Разделенные разности обратной функции

u_n	x_n	Разделенные разности	
		I	II
−1,500	0,000		
		0,540	
−0,574	0,500		−0,076
		0,365	
0,797	1,000		

$$\tilde{x}_* \equiv P(0) \approx x_0 + (0-u_0)x(u_0,u_1) +$$

$$+ (0-u_0)(0-u_1)x(u_0,u_1,u_2) = 0{,}744.$$

Точное решение есть $x_* = 0{,}732\ldots$, так что ошибка получилась небольшой. Для повышения точности в этом способе целесообразно взять новые узлы, близко расположенные к грубо найденному корню $\tilde{x}_*$. За центральный узел следует взять $\tilde{x}_*$ и поставить остальные узлы справа и слева от него. В данном случае можно взять следующие три точки: $\tilde{x}_*$, $\tilde{x}_* \pm 0{,}2$. Первоначальная сетка имела шаг $h = 0{,}5$, новая сетка имеет шаг $h = 0{,}2$. Погрешность $O(h^3)$ уменьшится примерно в 15 раз и составит менее 0,001.

4.1.5. Эрмитова интерполяция

Пусть в таблице заданы не только x_n и u_n, но и значения производных до некоторого порядка. Тогда можно потребовать, чтобы в узлах интерполяции совпадали не только значения функции и интерполяционного многочлена, но и их соответствующие производные. Такая интерполяция называется *эрмитовой*, и соответствующий интерполяционный полином будем обозначать $Q_N(x)$.

Ряд Тейлора. Поскольку значения функции и интерполяционного многочлена Ньютона в узлах совпадают, средние наклоны на участках между узлами также равны. Будем мысленно сближать соседние узлы $x_n \to x_{n-1}$, при этом средний наклон стремится к производной. Значит, после совпадения получим

многочлен, правильно передающий в точке x_0 не только значение функции, но и значение первой производной. Слияние трех узлов обеспечит правильную передачу не только значения функции и наклона, но и кривизны (второй производной) и т. д.

Проследим, как преобразуется интерполяционный многочлен Ньютона при слиянии двух соседних узлов:

$$u(x) \approx u(x_0) + \sum_{n=1}^{N} u(x_0, x_1, \ldots, x_n)\omega_n(x), \qquad \omega_n(x) = \prod_{k=0}^{n-1} (x - x_k). \tag{4.17}$$

Когда все узлы сливаются ($x_k \to x_0, k = 1, 2, \ldots$), то

$$\omega_n(x) \to (x - x_0)^n, \quad u(x_0, x_1, \ldots, x_n) \to u_0^{(n)}/n!$$

и (4.17) переходит в формулу Тейлора:

$$u(x) \approx \sum_{n=0}^{N} \frac{u_0^{(n)}}{n!}(x - x_0)^n.$$

Ряд Тейлора сходится внутри круга, радиус которого равен расстоянию от точки до ближайшей особой точки аналитического продолжения $u(x)$ в комплексную плоскость. Значит, большое значение $x - x_0$ брать неразумно. По этой причине многочленом Эрмита, построенным на основе ряда Тейлора, пользуются редко.

Локальный многочлен. Интерполяционный многочлен строится по ограниченному набору узлов x_n, $0 \leqslant n \leqslant N$. Поэтому формально он локальный. Однако он захватывает несколько интервалов. Будем называть локальным (в узком смысле слова) многочлен, который использует информацию о функции и ее производных только в паре соседних узлов.

Простейшим является случай, когда в узлах известны только значения функции. Очевидно, тогда это многочлен первой степени, совпадающий с соответствующим интерполяционным многочленом Ньютона. Однако сейчас его удобно записать в иной форме. Обозначим два используемых узла индексами $n = 0{,}1$ и введем локальную переменную

$$\xi = [x - 0{,}5(x_1 + x_0)]/h, \quad h = x_1 - x_0, \quad \xi \in [-0{,}5; 0{,}5]. \tag{4.18}$$

Тогда

$$Q_1(x) = 0{,}5(u_0 + u_1) + (u_1 - u_0)\xi. \tag{4.19}$$

Наиболее интересным является случай, когда в узлах заданы значения функции и ее первой производной. Четырем параметрам соответствует кубический интерполяционный многочлен. Будем искать его в следующей форме:

$$Q_3(x) = Q_1(x) + (\xi^2 - 1/4)(a + b\xi). \tag{4.20}$$

Второе слагаемое справа обращается в нуль на границах интервала при $\xi = \pm 0{,}5$, а $Q_1(x)$ в этих точках обеспечивает передачу значения функции в узлах x_0, x_1. Таким образом, $Q_3(x)$ также передает узловые значения функции. Остается передать узловые значения производной. Дифференцируем по x, учитывая, что $d\xi/dx = 1/h$. Затем подставляем граничные значения аргумента и получаем следующие соотношения:

$$u_1' = Q_3'(x_1) = (u_1 - u_0)/h + (2a + b)/(2h);$$

$$u_0' = Q_3'(x_0) = (u_1 - u_0)/h - (2a - b)/(2h).$$

Отсюда находим искомые коэффициенты:

$$a = (u_1' - u_0')h/2, \quad b = (u_1' + u_0')h + 2(u_1 - u_0). \tag{4.21}$$

Подставив (4.21) в (4.20), получим кубический интерполяционный многочлен Эрмита.

Этот прием обобщается на многочлены высших степеней. Например, пусть в узлах заданы дополнительно значения вторых производных. Тогда интерполяционный многочлен Эрмита 5-й степени ищем в следующем виде:

$$Q_5(x) = Q_3(x) + (\xi^2 - 1/4)^2(\alpha + \beta\xi).$$

Здесь $Q_3(x)$ определяется формулами (4.19) — (4.21). Аналогично предыдущему доказывается, что $Q_5(x)$ при любых α, β передает значения $u(x)$ и $u'(x)$ в обоих узлах x_0, x_1. Двукратно дифференцируя и приравнивая значения вторых производных в узлах, получаем систему линейных уравнений для α и β.

Погрешность. Формально многочлен Эрмита является многочленом Ньютона с кратными узлами. Поэтому для него справедлива априорная оценка погрешности интерполяционного многочлена Ньютона (4.14), где в многочлен $\omega_{N+1}(x)$ (4.11) нужно подставлять кратные узлы. Например, для кубического

многочлена узлы x_0, x_1 сольем в двукратный узел x_0, а узлы x_2, x_3 сделаем двукратным узлом x_1. Тогда получим мажорантную оценку:

$$|R_3| \leqslant \frac{M_4}{24}|\omega_4(x)|, \quad \omega_4 = (x - x_0)^2(x - x_1)^2.$$

Если мы не используем экстраполяцию, то $|\omega_4(x)| \leqslant h^4/16$, т. е.

$$|R_3| \leqslant \frac{M_4 h^4}{384}.$$

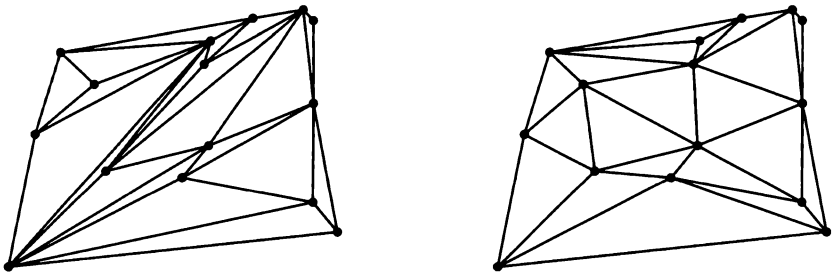
Видно, что кубический интерполяционный многочлен Эрмита имеет точность $O(h^4)$ по шагу сетки, как и кубический интерполяционный многочлен Ньютона. Но в интерполяционный многочлен Ньютона входит $\omega_4(x) = (x - x_0)(x - x_1)(x - x_2)(x - x_3)$. Легко проверить, что на равномерной сетке его центральный экстремум равен $9h^4/16$, а крайние — еще больше. Таким образом, интерполяционный многочлен Эрмита в девять раз точнее интерполяционного многочлена Ньютона при том же шаге сетки.

Однако это уточнение не бесплатное: для построения кубического интерполяционного многочлена Эрмита используется вдвое больше информации, чем для кубического интерполяционного многочлена Ньютона. Если информация все равно имеется, тогда выгоднее применять многочлен Эрмита. Если же использовать одинаковый объем информации, то для кубического многочлена Ньютона нужно брать вдвое меньший шаг. При этом он по точности будет почти вдвое (в 16/9 раз) превосходить интерполяционный многочлен Эрмита.

Аналогичные соображения верны и для многочленов высших степеней. Поэтому интерполяционный полином Эрмита применяется сравнительно редко.

4.1.6. Многомерная интерполяция

Существует немало функций от двух и более переменных, заданных таблично. Например, давление вещества является функцией двух переменных — его температуры и плотности; проводимость плазмы в электромагнитном поле зависит от трех переменных — температуры и плотности плазмы, а также поперечной компоненты магнитного поля. При этом сетки значений аргумента бывают существенно разных типов. Если исходные



а *б*

Рис. 4.2. Триангуляция:

а, б — варианты разбиения

таблицы получены расчетным путем, то значения аргументов при этом выбирают так, чтобы они образовывали регулярную сетку (например, в результате пересечения линий $x = \text{const}$ и $y = \text{const}$). Если же таблицы получены в экспериментах, то точки (x, y) нередко беспорядочно разбросаны на плоскости, т. е. образуют нерегулярную сетку.

Нерегулярная сетка. Пусть функция двух переменных $u(x, y)$ задана на нерегулярной сетке (x_n, y_n). В этом случае обычно ограничиваются только простейшими способами интерполяции.

Для этого сначала соединяют соседние точки так, чтобы вся плоскость разбилась на треугольники (рис. 4.2). Такое разбиение неоднозначно. Для интерполяции невыгодны вытянутые треугольники, у которых хотя бы одна из высот много меньше соответствующей стороны (рис. 4.2, *а*). Целесообразны разбиения, у которых все треугольники, насколько это возможно, близки к равносторонним (рис. 4.2, *б*).

Затем определяют треугольник, внутрь которого попадает искомая точка (x, y). Обозначим вершины этого треугольника через $\mathbf{r}_j = (x_j, y_j)$, $j = 1, 2, 3$. Нахождение $u(x, y)$ по этим значениям u_j является интерполяцией в узком смысле слова. Если точка (x, y) лежала бы вне данного треугольника, то имела бы место экстраполяция. По одномерной задаче известно, насколько это опасно.

Функцию на выбранном треугольнике можно заменить отрезком плоскости в трехмерном пространстве (u, x, y), проходящей через три узловые точки (u_j, x_j, y_j), $j = 1, 2, 3$, т. е. приближенно положить

$$u \approx a + bx + cy. \tag{4.22}$$

Для определения коэффициентов получим следующую систему трех уравнений:

$$\begin{cases} u_1 = a + bx_1 + cy_1, \\ u_2 = a + bx_2 + cy_2, \\ u_3 = a + bx_3 + cy_3. \end{cases} \tag{4.23}$$

Четыре уравнения (4.22) и (4.23) можно интерпретировать следующим образом. Четырехкомпонентный столбец $(u, u_1, u_2, u_3)^T$ является линейной комбинацией трех столбцов $(1,1,1,1)^T$, $(x, x_1, x_2, x_3)^T$, $(y, y_1, y_2, y_3)^T$ с некоторыми коэффициентами a, b, c. Но тогда определитель, составленный из этих четырех столбцов, обращается в нуль:

$$\begin{vmatrix} u & 1 & x & y \\ u_1 & 1 & x_1 & y_1 \\ u_2 & 1 & x_2 & y_2 \\ u_3 & 1 & x_3 & y_3 \end{vmatrix} = 0. \tag{4.24}$$

Раскрывая этот определитель по элементам первого столбца, получим

$$\begin{gathered} u\Delta(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3) - u_1\Delta(\mathbf{r}, \mathbf{r}_2, \mathbf{r}_3) + \\ + u_2\Delta(\mathbf{r}, \mathbf{r}_1, \mathbf{r}_3) - u_3\Delta(\mathbf{r}, \mathbf{r}_1, \mathbf{r}_2) = 0; \end{gathered} \tag{4.25}$$

здесь Δ — миноры элементов первого столбца (4.24):

$$\Delta(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3) = \begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix} \tag{4.26}$$

и т. д. Переставляя строки в минорах и меняя знаки, где необходимо, преобразуем (4.25) к следующему виду:

$$\begin{gathered} u = [u_1\Delta(\mathbf{r}, \mathbf{r}_2, \mathbf{r}_3) + u_2\Delta(\mathbf{r}_1, \mathbf{r}, \mathbf{r}_3) + \\ + u_3\Delta(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r})]/\Delta(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3). \end{gathered} \tag{4.27}$$

Эта формула определяет уравнение плоскости, проходящей через три заданные точки, т. е. является двумерной линейной интерполяцией по трем точкам. По аналогии с одномерным случаем ее погрешность есть $O(h^2)$, где h — размер наибольшей из сторон треугольника.

Многомерность. Описанный способ легко переносится на любое число измерений. Например, в трехмерном случае все пространство разбивается на тетраэдры. Надо найти тетраэдр,

в который попадает искомая точка (x, y, z), занумеровать его четыре вершины: (x_j, y_j, z_j), $1 \leqslant j \leqslant 4$. По ним можно однозначно построить линейную интерполяцию по аналогичным формулам, но с определителями порядка на 1 больше. Вместо определителя (4.26) появляется определитель 4-го порядка $\Delta(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \mathbf{r}_4)$, а в квадратную скобку в (4.27) войдет четвертое слагаемое.

Высокие порядки. Если точность $O(h^2)$ недостаточна, то можно построить интерполяцию более высокого порядка точности $O(h^3)$ и т. д. Параболическая интерполяция должна использовать квадратичную функцию двух переменных:

$$u \approx a + bx + cy + dx^2 + fxy + gy^2. \tag{4.28}$$

Она содержит шесть свободных коэффициентов, и для их определения нужно выбрать шесть точек (x_j, y_j), окружающих искомую точку. Фигура, составленная из этих точек, может иметь сложную форму. Поэтому хороший выбор этих точек — непростая задача.

Далее действуем аналогично линейному случаю. Вместо (4.22) возникает (4.28), а вместо трех уравнений (4.23) — линейная система из шести уравнений. Эти уравнения эквивалентны обращению в нуль определителя 7-го порядка, состоящего из столбцов $\{u\}, \{1\}, \{x\}, \{y\}, \{x^2\}, \{xy\}, \{y^2\}$. Он раскрывается по элементам первого столбца, но миноры при этом имеют более сложный вид.

Реализовать соответствующие формулы на компьютере несложно, но выбрать подходящие для интерполяции точки нелегко. Поэтому интерполяцией высокого порядка точности на нерегулярных сетках пользуются редко.

Регулярная сетка. Рассмотрим прямоугольную двумерную сетку, образованную линиями $x = x_n$ и $y = y_m$. Точки их пересечения (x_n, y_m) являются узлами регулярной сетки. Если функция табулирована на регулярной сетке, интерполировать ее намного проще, и нетрудно строить интерполяции высокого порядка точности как произведение одномерных интерполяций.

Например, построим простейшую интерполяцию точности $O(h^2)$. Для этого выберем такие пары соседних узлов, чтобы выполнялось $x_n < x < x_{n+1}$ и $y_m < y < y_{m+1}$. Тем самым найдены вершины четырехугольной ячейки, внутрь которой попадает искомая точка (x, y). Заметим, что процесс отыскания ячейки здесь несравненно проще, чем в случае нерегулярной сетки (рис. 4.3).

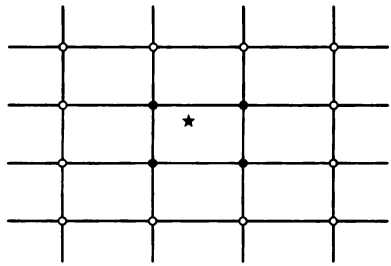


Рис. 4.3. Интерполяция на регулярной сетке (искомая точка — звездочка; по четырем точкам строится интерполяция $O(h^2)$; по точкам и кружкам — интерполяция $O(h^4)$)

Вдоль каждой строки $y = \text{const}$ проводим интерполяцию многочленом Ньютона по x. В данном случае это многочлен первой степени:

$$u(x, y_m) = u(x_n, y_m) + \frac{x - x_n}{x_{n+1} - x_n}[u(x_{n+1}, y_m) - u(x_n, y_m)] \quad (4.29)$$

и аналогично для y_{m+1}. Найденные значения интерполируем по вертикали по y:

$$u(x, y) = u(x, y_m) + \frac{y - y_m}{y_{m+1} - y_m}[u(x, y_{m+1}) - u(x, y_m)]. \quad (4.30)$$

Погрешность (4.29) есть $O(h_x^2)$, а (4.30) — $O(h_y^2)$. Поэтому полная погрешность данной аппроксимации составит $O(h_x^2 + h_y^2)$.

Порядок погрешности такой же, как при интерполяции по трем узлам на нерегулярной сетке. Хотя каждый из многочленов (4.29) и (4.30) имеет первую степень, подстановка (4.29) в (4.30) дает не только линейные члены, но и один квадратичный член $\sim xy$.

Аналогично можно строить интерполяцию более высокой точности. Например, для точности $O(h^4)$ надо взять по четыре точки по каждой координате (см. рис. 4.3). При этом по строчкам и вертикали строятся кубические интерполяционные многочлены Ньютона. Таким способом можно строить также интерполяцию, имеющую разную точность по разным координатам. Для этого по строчкам и столбцам берут различное количество точек.

4.2. СПЛАЙН-ИНТЕРПОЛЯЦИЯ

4.2.1. Историческая справка

Для построения графиков функции по точкам с помощью лекал стараются выбрать лекало, на которое попадает как мож-

но больше точек графика. Еще лучший результат достигается при использовании гибкого бруска (металлической линейки, поставленной на ребро), который прикладывают к точкам графика.

Уравнение гибкого бруска было написано еще Бернулли. Шонберг приближенно заменил решение этого уравнения полиномом третьей степени. Между каждой парой соседних узлов строился свой полином, а в узлах соседние полиномы «склеивались» так, чтобы обеспечить максимально возможную гладкость интерполяции.

Описанный прием быстро распространился и был обобщен на случай полинома большей степени и даже неполиномиальной интерполяции. Этот прием — интерполирование функции кусочно-полиномиальной функцией получил название сплайн-интерполяции (от англ. *spline* — гибкий брусок).

Пусть задана сетка $\{x_n, 0 \leqslant n \leqslant N\}$; точки x_n называют узлами сплайна. Полиномиальным сплайном $S_p(x)$ дефекта q называется функция, удовлетворяющая следующим требованиям:

1) $S_p(x)$ на каждом интервале $[x_{n-1}, x_n]$ является полиномом степени p;

2) эти полиномы «склеены» во внутренних узлах так, что сплайн остается непрерывным вместе со своими $p-q$ производными: $S_p^{(k)}(x_n - 0) = S_p^{(k)}(x_n + 0)$, $0 \leqslant k \leqslant p-q$, $1 \leqslant n \leqslant N-1$.

Чаще всего ограничиваются сплайнами дефекта $q = 1$, когда разрывна лишь старшая (p-я) производная, а все младшие производные непрерывны. В этом случае говорят просто о сплайне степени p, опуская упоминание о дефекте.

Если сплайн в заданных точках совпадает с табулированной функцией, то такой сплайн называется интерполяционным.

Приведем примеры интерполяционных сплайнов. Ломаная, проведенная через заданные точки, состоит из отрезков прямых. В узлах она непрерывна, но первая производная разрывна. Значит, это сплайн степени $p = 1$ дефекта $q = 1$. Кубический интерполяционный многочлен Эрмита (см. подразд. 4.1.5) склеен из кубических многочленов, а в узлах непрерывен вместе с первой производной. Это сплайн степени $p = 3$ дефекта $q = 2$.

4.2.2. Кубический сплайн

В этом случае на каждом интервале интерполяционный сплайн является многочленом третьей степени. Его удобно записать в следующем виде:

$$S_{3n}(x) = a_n + b_n(x - x_{n-1}) + c_n(x - x_{n-1})^2 + \\ + d_n(x - x_{n-1})^3, \quad x_{n-1} < x < x_n, \quad 1 \leqslant n \leqslant N, \tag{4.31}$$

здесь коэффициенты a_n, b_n, c_n, d_n свои для каждого интервала. Эти коэффициенты определяют из условий в узлах. Очевидно, каждый многочлен (4.31) в своем правом и левом узлах должен совпадать с табулированными значениями функции:

$$S_{3n}(x_{n-1}) \equiv a_n = u_{n-1}, \quad 1 \leqslant n \leqslant N, \tag{4.32}$$

$$S_{3n}(x_n) = a_n + b_n h_n + c_n h_n^2 + d_n h_n^3 = u_n, \\ h_n = x_n - x_{n-1}, \quad 1 \leqslant n \leqslant N. \tag{4.33}$$

Число этих уравнений вдвое меньше числа неизвестных коэффициентов, поэтому для определенности задачи нужны дополнительные условия. Для их получения вычислим первую и вторую производные многочлена (4.31):

$$S'_{3n}(x) = b_n + 2c_n(x - x_{n-1}) + 3d_n(x - x_{n-1})^2, \\ x_{n-1} < x < x_n, \quad 1 \leqslant n \leqslant N, \tag{4.34}$$

$$S''_{3n}(x) = 2c_n + 6d_n(x - x_{n-1}), \\ x_{n-1} < x < x_n, \quad 1 \leqslant n \leqslant N; \tag{4.35}$$

потребуем непрерывности этих производных (т. е. гладкости сплайна) во всех точках, включая узлы. Приравняв во внутреннем узле x_n правые и левые пределы производных, получим

$$b_{n+1} = b_n + 2c_n h_n + 3d_n h_n^2, \quad 1 \leqslant n \leqslant N - 1; \tag{4.36}$$

$$c_{n+1} = c_n + 3d_n h_n, \quad 1 \leqslant n \leqslant N. \tag{4.37}$$

Здесь введен дополнительный коэффициент c_{N+1}, имеющий смысл $0{,}5S''_3(x_N)$. Поэтому число уравнений (4.37) на одно больше, чем (4.36), а общее число неизвестных коэффициентов увеличилось на единицу.

Равенства (4.32), (4.33) и (4.36), (4.37) образуют систему $4N - 1$ уравнений для $4N + 1$ неизвестных коэффициентов. Недостает еще двух уравнений; но их выбор нетривиален, и его обсудим позже.

Преобразование уравнений. Сначала приведем систему уравнений к виду, содержащему только коэффициенты c_n. Уравнение (4.32) сразу дает нам коэффициенты a_n. Из уравнений (4.37) следует

$$d_n = (c_{n+1} - c_n)/(3h_n), \quad 1 \leqslant n \leqslant N. \tag{4.38}$$

Подставим (4.38) в (4.33), одновременно исключив оттуда $a_n = u_{n-1}$; тогда получим

$$b_n = (u_n - u_{n-1})/h_n - h_n(c_{n+1} + 2c_n)/3, \quad 1 \leqslant n \leqslant N. \tag{4.39}$$

Исключим теперь из (4.36) величины b_n и b_{n+1} с помощью (4.39), соответственно увеличивая во втором случае индекс на единицу, используя (4.38) для d_n. Остается система линейных уравнений для коэффициентов c_n, легко приводящаяся к следующему виду:

$$\begin{gathered} h_{n-1}c_{n-1} + 2(h_{n-1} + h_n)c_n + h_n c_{n+1} = \\ = 3[(u_n - u_{n-1})/h_n - (u_{n-1} - u_{n-2})/h_{n-1}], \quad 2 \leqslant n \leqslant N. \end{gathered} \tag{4.40}$$

Это система $N-1$ уравнения для $N+1$ неизвестных коэффициентов c_n. Уравнения имеют трехдиагональный вид. По-прежнему двух уравнений не хватает.

Кубический сплайн похож на кубический интерполяционный многочлен, поэтому следует ожидать от него точности $O(h^4)$. Если табулированы только значения функции, то дополнительные условия приходится подбирать так, чтобы не испортить этого порядка точности. Здесь возникают две различные ситуации: $u(x)$ может быть периодической либо непериодической.

Периодические граничные условия. Сначала рассмотрим периодическую и достаточно гладкую $u(x)$, для которой отрезок $[x_0, x_N]$ является периодом. Тогда на обоих концах отрезка функция и ее производные принимают одинаковые значения: $u^{(q)}(x_0) = u^{(q)}(x_N)$, $q = 0, 1, \ldots$ (при этом обязательно надо проверять исходную информацию: если $u_0 \neq u_N$, то данные ошибочны). Естественно также требовать от сплайна, чтобы он удовлетворял аналогичным условиям:

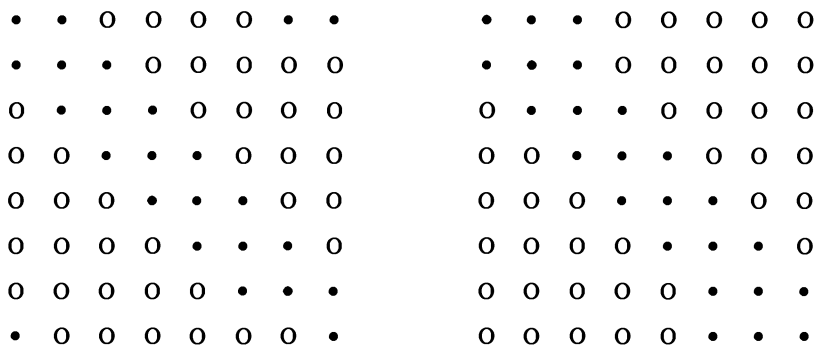
$$S_{31}^{(q)}(x_0 + 0) = S_{3N}^{(q)}(x_N - 0), \quad q = 0, 1, 2. \tag{4.41}$$

Требовать непрерывности третьей производной мы уже не можем. При $q = 0$ условие (4.41) автоматически следует из условия интерполяции. Значения $q = 1, 2$ дают два недостающих уравнения

$$b_1 = b_N + 2c_N h_N + 3d_N h_N^2 \quad \text{и} \quad c_1 = c_{N+1}.$$

Исключив отсюда b_1, b_N, d_N с помощью (4.38), (4.39), получим граничные условия в следующем виде:

$$\begin{gathered} (2c_1 + c_2)h_1 + (c_N + 2c_{N+1})h_N = \\ = 3[(u_1 - u_0)/h_1 - (u_N - u_{N-1})/h_N], \\ c_1 - c_{N+1} = 0. \end{gathered} \tag{4.42}$$



а *б*

Рис. 4.4. Структура матрицы для нахождения коэффициентов сплайна:

а — периодические граничные условия; *б* — естественные граничные условия

Заметим, что первое из этих уравнений является циклическим замыканием цепочки (4.40). Дополнив условиями (4.42) уравнения (4.40), получим систему $N+1$ линейных уравнений для определения $N+1$ коэффициентов c_n. Структура матрицы этой системы приведена на рис. 4.4, *а*.

Эта матрица имеет ненулевые три диагонали и три угловых элемента. Соответствующая система линейных уравнений легко решается методом Гаусса без выбора главного элемента с обходом нулей, или модификацией так называемой циклической прогонки (обыкновенная прогонка здесь не годится). Диагональные элементы матрицы преобладают, поэтому алгоритм очень устойчив.

Найдя коэффициенты c_n, оставшиеся коэффициенты сплайна вычисляют по формулам (4.32), (4.38) и (4.39).

Если периодическая $u(x)$ имеет непрерывную четвертую производную, то на равномерной сетке можно получить строгую оценку погрешности:

$$|u(x) - S_3(x)| \leqslant \frac{1}{384} M_4 h^4, \quad M_4 = \max |u^{(IV)}(x)|. \qquad (4.43)$$

Численный коэффициент в этой оценке очень мал. Точно такую же погрешность имеет локальный кубический интерполяционный многочлен Эрмита (4.20). Но для его построения нужно табулировать в узлах еще и производные u'_n, а для сплайна этого не требуется. Это наглядно демонстрирует преимущество сплайна.

Естественные граничные условия. Пусть $u(x)$ непериодическая функция. В этом случае постановка граничных условий неочевидна. Неудачные условия могут привести к существенному снижению точности. Например, Шонберг в 1946 г. предложил выводить из вариационного уравнения $\int [S_3''(x)]^2 dx = \min$ как само уравнение кубического сплайна, так и граничные условия к нему. Отсюда вытекали условия $S_3''(x_0) = 0$, $S_3''(x_N) = 0$. Эти условия рекомендуются во многих учебниках до сих пор. Точные значения $u''(x_0)$ и $u''(x_N)$, вообще говоря, не равны нулю. Значит, погрешность $|S_3''(x_0) - u''(x)|$ будет составлять $O(1)$ в граничной точке и, по непрерывности, вблизи нее. Соответственно погрешность $|S_3'(x_0) - u'(x)|$ составит $O(h)$, а погрешность $|S_3(x_0) - u(x)|$ будет $O(h^2)$ вместо $O(h^4)$!

Наиболее хорошим дополнительным условием является вариационное условие. Надо минимизировать разрывы третьих производных сплайна во внутренних узлах в смысле метода наименьших квадратов. Разрыв $S_3'''(x)$ в узле x_n пропорционален $d_{n+1} - d_n$. Наилучшие результаты дает вариационное условие

$$\begin{aligned} &\sum_{n=1}^{N-1} (d_{n+1} - d_n)^2/(h_n + h_{n+1}) = \\ &= \frac{1}{3} \sum_{n=1}^{N-1} [(c_{n+2} - c_{n+1})/h_{n+1} - \\ &-(c_{n+1} - c_n)/h_n]/(h_n + h_{n+1}) = \min. \end{aligned} \tag{4.44}$$

Условие (4.44) вместе с уравнениями (4.40) образует задачу на условный экстремум. Ее можно решить методом неопределенных множителей Лагранжа. Но это приводит к достаточно сложному алгоритму.

Есть один частный случай, когда вместо всей суммы (4.44) мы оставляем только крайние слагаемые с $n = 1$ и $n = N - 1$. Это эквивалентно требованию, чтобы $S_3'''(x)$ была непрерывна в приграничных узлах x_1, x_{N-1}. Это дает $d_1 = d_2$ и $d_{N-1} = d_N$. Выразив коэффициенты d_n через c_n по (4.38), получим соотношения

$$\begin{aligned} (c_2 - c_1)/h_1 - (c_3 - c_2)/h_2 = 0, \\ (c_{N+1} - c_N)/h_N - (c_N - c_{N-1})/h_{N-1} = 0. \end{aligned} \tag{4.45}$$

Соотношения (4.45) и (4.40) образуют систему линейных уравнений для коэффициентов c_n. Ее матрица отличается от трех-

диагональной только первой и последней строкой (рис. 4.4, *б*). Такая система легко решается методом Гаусса для ленточной матрицы без выбора главного элемента, так как диагональные элементы преобладают. Этот алгоритм очень устойчив. После нахождения коэффициентов c_n остальные коэффициенты выражаются через них так же, как коэффициенты периодического сплайна.

Погрешность естественного сплайна есть $O(h^4)$. На некотором расстоянии от граничных интервалов при этом справедлива оценка (4.43). Вблизи границы оценка аналогична, но с несколько бóльшим численным коэффициентом. Этот коэффициент быстро стремится к $1/384$ по мере удаления от границы.

Замечание. Выберем в качестве функции $u(x)$ кубический многочлен. Можно строго доказать, что кубический интерполяционный сплайн с естественными дополнительными условиями (4.44) или (4.40) воспроизводит этот многочлен точно. Использование других граничных условий не обеспечивает точного воспроизведения кубического многочлена.

4.2.3. Обобщения

Описанным способом можно строить полиномиальные сплайны произвольной степени p. Количество дополнительных (граничных) условий при этом равно $p-1$. Погрешность таких сплайнов при правильном выборе граничных условий будет $O(h^{p+1})$, как у интерполяционного многочлена p-й степени.

При $p = 1$ сплайн не требует граничных условий. Он просто является ломаной, проведенной через узлы, и на каждом интервале тождественно совпадает с интерполяционным многочленом первой степени. При $p > 1$ коэффициент в остаточном члене меньше, чем у аналогичного интерполяционного многочлена. Это означает количественный выигрыш в точности. Выигрыш будет тем сильнее, чем выше степень p. Однако с повышением степени сплайна сложность алгоритма для нахождения его коэффициентов быстро увеличивается, поэтому на практике чаще всего ограничиваются кубическим сплайном.

Кроме полиномиальных сплайнов существуют и другие. Например, для интерполяции быстро меняющихся функций вместо степеней используют комбинации экспонент или гиперболических функций.

4.3. НЕЛИНЕЙНАЯ ИНТЕРПОЛЯЦИЯ

4.3.1. Выравнивание

На успех интерполяции можно рассчитывать при приближении функций с не очень большими производными, когда можно использовать полиномы невысокой степени. Для быстро меняющихся функций, таких как e^x, $x \gg 1$, приходится выбирать неприемлемо малый шаг сетки h, иначе интерполяция дает плохую точность: велики значения производных, входящих в оценку погрешности.

Часто ситуацию можно поправить, выбрав удачную замену переменных так, чтобы новая функция изменялась слабо. Например, в задаче баллистики коэффициент сопротивления воздуха при малых значениях скорости пропорционален первой степени скорости: $k(v) \sim v$. При скоростях порядка скорости полета парашютиста $k(v) \sim v^2$, а при околозвуковых скоростях $k(v) \sim v^3$. В задачах о прохождении пучка света или частиц через поглощающую среду ослабление носит другой характер — экспоненциальный.

В таких ситуациях стараются найти такое преобразование переменных $\xi(x)$, $\eta(u)$, чтобы зависимость $\eta(\xi)$ была близка к линейной. Такие переменные называют ***выравнивающими***. Если выравнивающие переменные удалось найти, то исходную таблицу $\{x_n, u_n\}$ преобразуют в таблицу $\{\xi_n, \eta_n\}$, выполняют по ней интерполяцию и возвращаются к исходным переменным.

Разумеется, вид преобразования приходится подбирать с учетом специфики каждой задачи. Для выбора выравнивающих переменных стараются использовать априорные сведения о природе задачи. Например, в задаче баллистики зависимость $k(v)$ близка к степенной, поэтому в переменных $\xi = \ln v$, $\eta = \ln k$ зависимость $\eta(\xi)$ будет близка к линейной. Если таких соображений нет, то пробуют подобрать преобразование на глаз, строя при этом графики $\eta(\xi)$ и добиваясь их близости к прямой.

Пример 4.3. Рассмотрим функцию $u(x)$, заданную табл. 4.3. Видно, что функция очень сильно меняется от одного узла к другому и разделенные разности высоких порядков не малы. По этой таблице составляется интерполяционный кубический многочлен Ньютона:

$$P(x) = 1 + 4(x-0) + 11(x-0)(x-1) + \\ + 22(x-0)(x-1)(x-2). \tag{4.46}$$

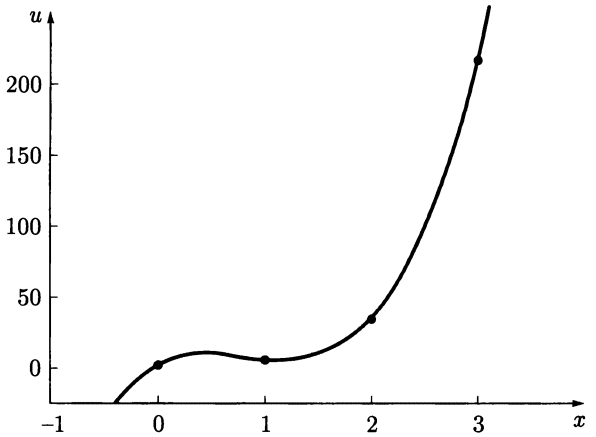


Рис. 4.5. Точки из табл. 4.3 и интерполяционный многочлен (4.46) в исходных переменных

Последовательные слагаемые в (4.46) не убывают, а скорее даже возрастают. Например, $P(1{,}5) = 1 + 6 + 8{,}25 - 8{,}25 = 7$. Это показывает отсутствие сходимости. Кроме того, этот интерполяционный многочлен оказывается немонотонным (рис. 4.5), несмотря на монотонность исходной функции. Поэтому ожидать правдоподобного результата не приходится.

Выберем в качестве выравнивающих переменных $\xi = x$, $\eta = \lg u$. Преобразованная табл. 4.3 выглядит существенно лучше. Разделенные разности быстро уменьшаются с увеличением порядка. В соответствующем интерполяционном многочлене

$$\eta(\xi) = 0 + 0{,}6990(\xi - 0) + 0{,}0467(\xi - 0)(\xi - 1) - \\ - 0{,}0080(\xi - 0)(\xi - 1)(\xi - 2) \tag{4.47}$$

Таблица 4.3

Интерполяция в исходных и выравнивающих переменных

x	u	Разности			$\xi = x$	$\eta = \lg u$	Разности		
0	1				0	0,0000			
		4					0,6990		
1	5		11		1	0,6990		0,0467	
		26		22			0,7924		−0,0080
2	31		77		2	1,4914		0,0203	
		180					0,8329		
3	211				3	2,3243			

слагаемые быстро убывают. Кроме того, новый интерполяционный многочлен монотонен в требуемом диапазоне изменений аргумента. Для того же аргумента $\xi = x = 1{,}5$ это дает $\eta(1{,}5) = 0 + 1{,}0485 + 0{,}0350 + 0{,}0033 = 1{,}0866$. В качестве погрешности можно принять модуль последнего слагаемого $\delta_\eta \approx 0{,}0033$. Возвращаясь к исходным переменным, получаем $u(1{,}5) = 12{,}21 \pm 0{,}8\,\%$.

Видно, что переход к выравнивающим переменным позволил получить результат с неплохой точностью.

Контроль. В данном примере контрольное значение $x = 1{,}5$ является серединой интервалов $(0, 3)$ и $(1, 2)$. Этим отрезкам соответствуют шаги $h = 3$ и $h = 1$. Линейная интерполяция по концам отрезка на середину интервала — это просто полусумма значений на концах. Результаты такой интерполяции в исходных и выравнивающих переменных представлены в табл. 4.4. Их можно рассматривать как результат расчета на сгущающихся сетках с коэффициентом сгущения $r = 3$. Порядок точности линейной интерполяции $p = 2$. Применив метод Ричардсона, получим экстраполяцию на $h = 0$; экстраполированное значение и ричардсоновская оценки погрешности приведены в последней строке табл. 4.4.

Видно, что в исходных переменных погрешность превышает само искомое значение, т. е. результат неприемлем. В выравнивающих переменных результат разумен, а оценка погрешности при переходе к исходным переменным составляет 2 %.

Это является дополнительным подтверждением целесообразности перехода к выравнивающим переменным.

Обобщение. Переход к выравнивающим переменным полезен в разнообразных ситуациях, если функция быстро меняется от одного узла сетки к другому. Его используют при интерполяции многочленами, сплайнами, другими видами функций, многомерной интерполяции, среднеквадратичной аппроксимации и т. д. Поэтому прежде чем использовать тот или иной способ интерполяции, стоит оценить, не выгоднее ли предварительно перейти к выравнивающим переменным.

Таблица 4.4

Контроль методом Ричардсона

h	$u(1{,}5)$	$\eta(1{,}5)$
3	106	1,1622
1	18	1,0952
0	7 ± 11	$1{,}0868 \pm 0{,}0084$

Например, при расчетах газовых турбин, аэродинамического обтекания, сопел реактивных двигателей требуется так назы-

ваемое уравнение состояния: зависимость давления p от температуры T и плотности ρ. Истинная зависимость очень сложна и обычно задается таблицей на регулярной сетке $p_{nm} = p(T_n, \rho_m)$. Зависимости функции от обоих аргументов здесь сравнительно близки к степенным. Поэтому в качестве выравнивающего преобразования выгодно брать двойное логарифмическое: $\zeta = \lg p$, $\xi = \lg T$, $\eta = \lg \rho$. В таких переменных зачастую двумерная интерполяция многочленами первой степени уже дает приемлемую точность.

4.3.2. Рациональная интерполяция

Многочлены, особенно высокой степени, быстро растут при увеличении аргумента. Сплайны — склейка из многочленов, поэтому к ним это в определенной степени тоже относится. Таким образом, многочленами и сплайнами трудно передать поведение функции на большом отрезке, даже если эта функция несложная.

Например, если отрезок $-a \leqslant x \leqslant a$ велик, то функция $u(x) = \operatorname{arctg} x$ вблизи его концов почти константа ($\pm\pi/2$). Попробуйте аппроксимировать такую функцию многочленом.

Многочленами и сплайнами невозможно аппроксимировать также функцию с полюсами, например $u(x) = \operatorname{tg} x$, $-\pi/2 < x < \pi/2$.

Возможности интерполяции существенно расширяются, если использовать аппроксимацию рациональной функцией — отношением многочленов:

$$u(x) \approx \frac{P_N(x)}{Q_M(x)}, \quad P_N = \sum_{n=0}^{N} a_n x^n, \quad Q_M = \sum_{m=0}^{M} b_m x^m. \qquad (4.48)$$

Поскольку числитель и знаменатель определены с точностью до общего множителя, то один из коэффициентов надо положить равным 1 (коэффициенты a_0 или b_0 могут обращаться в нуль, но оба старших коэффициента отличны от нуля, так как они определяют старшие степени многочленов). Далее будем полагать $b_M = 1$.

Рациональная функция может передать полюс $u(x)$ q-го порядка в точке $\bar{x}$: для этого достаточно, чтобы $Q_M(x)$ содержало множитель $(x - \bar{x})^q$. Можно передать горизонтальную асимптотику функции $u(x)$ при $x \to \infty$: для этого достаточно положить $N = M$.

Нахождение коэффициентов. Рациональная функция (4.48) содержит $N + M + 1$ свободный коэффициент. Для их нахождения выберем такое же число значений табулированной функции и занумеруем их подряд: $u_k = u(x_k)$, $0 \leqslant k \leqslant N + M$. Напишем условия интерполяции в выбранных узлах:

$$u_k = P_N(x_k)/Q_M(x_k), \quad 0 \leqslant k \leqslant N + M \quad (b_M = 1).$$

Эти уравнения легко преобразуются к следующей форме:

$$\sum_{n=0}^{N} a_n x_k^n - u_k \sum_{m=0}^{M-1} b_m x_k^m = u_k x_k^M, \quad 0 \leqslant k \leqslant N + M. \qquad (4.49)$$

Соотношения (4.49) являются системой $N + M + 1$ линейных уравнений для определения такого же числа коэффициентов рациональной интерполяции. Матрица системы плотно заполненная, и решать систему целесообразно методом Гаусса с выбором главного элемента.

Если функция табулирована на сетке с числом узлов больше, чем $N+M+1$, или если ее можно вычислить в любых требуемых нам точках, то встает вопрос о выборе узлов интерполяции $\{x_k\}$. Когда отрезок определения функции конечен и не слишком велик, то зачастую размещают узлы x_k равномерно по этому отрезку, включая в число узлов концы отрезка. Если на каких-то участках поведение функции более сложное, на этих участках берут более густую сетку. Если функция задана на неограниченном отрезке, то особенное внимание уделяют такому выбору степеней многочленов, чтобы правильно передать асимптотики функции. Здесь многое зависит от опыта вычислителя.

Матрица линейной системы (4.49) плотно заполнена. При увеличении числа коэффициентов ее обусловленность обычно быстро ухудшается. Поэтому большое число коэффициентов использовать не рекомендуется. По-видимому, при расчете с 64-разрядными числами целесообразно ограничиваться $N + M \leqslant 10$, хотя тщательного исследования этого вопроса не проводилось.

Пример 4.4. Рассмотрим построение рациональной интерполяции для $u(x) = \operatorname{arctg} x$ на полупрямой $0 \leqslant x < \infty$. Существует предел $u(+\infty) = \pi/2$. Это означает, что в формуле (4.48) нужно брать $N = M$. Кроме того, при $x \to 0$ функция $u(x) \approx x$ и разлагается в ряд по нечетным степеням x. Для того чтобы удовлетворить обоим требованиям, удобнее аппроксимировать

не $u(x)$, а ее квадрат: $u^2(x) \equiv \operatorname{arctg}^2 x \approx P_N(x^2)/Q_N(x^2)$. Для удовлетворения граничным условиям надо полагать

$$a_0 = 0, \quad a_1/b_0 = 1, \quad a_N = (\pi/2)^2, \quad b_N = 1.$$

Таким образом, свободными остаются $2N-2$ параметров. Столько же нужно брать узлов интерполяции на $(0, +\infty)$. При $N = 1$ дополнительных условий уже не требуется, что после извлечения корня квадратного и небольшого преобразования дает

$$\operatorname{arctg} x = x/\sqrt{1 + (2x/\pi)^2}.$$

Даже такое грубое приближение, где еще не взято ни одного узла интерполяции, дает погрешность не хуже 12 %. При введении узлов интерполяции точность быстро возрастает (напомним, что для каждого следующего уточнения узлы интерполяции нужно добавлять парами).

Пример 4.5. Аппроксимируем $u(x) = \operatorname{tg} x$ на отрезке $0 \leqslant x < \pi/2$. Функция имеет полюс первого порядка при $x = \pi/2$, причем $u(x) \approx (\pi/2 - x)^{-1}$; при $x \to 0$ снова $u(x) \approx x$ и разлагается в ряд по нечетным степеням x. Аналогично примеру 4.4 целесообразно аппроксимировать $u^2(x)$. При этом для передачи полюса знаменатель должен содержать множитель $(\pi/2 - x)^2$. Поэтому будем искать аппроксимацию в следующей форме:

$$u^2(x) \equiv \operatorname{tg}^2 x \approx \frac{x^2}{(1 - 4x^2/\pi^2)^2} \frac{P_N(x^2)}{Q_M(x^2)}; \quad N + M \geqslant 1.$$

Соотношение степеней этих многочленов может быть произвольным, но должно выполняться

$$\frac{P_N(0)}{Q_M(0)} = 1, \quad \frac{P_N(\pi^2/4)}{Q_M(\pi^2/4)} = \frac{64}{\pi^4}.$$

Например, даже не привлекая ни одного узла интерполяции, можно взять $P_1(x^2) = 1 - (4/\pi^2 - 256/\pi^6)x^2$ и $Q_0(x^2) \equiv 1$. Это дает точность не хуже 0,2 % на всем отрезке интерполирования! Обычно уже умеренные значения N и M позволяют получить уже очень высокую точность.

Рациональную интерполяцию часто используют при составлении стандартных программ для специальных функций. Если функция имеет какую-то известную нестепенную асимптотику (например, экспоненциальную), то ее предварительно выделяют в виде отдельного множителя и аппроксимируют оставшуюся часть.

Глава 5

СРЕДНЕКВАДРАТИЧНАЯ АППРОКСИМАЦИЯ

5.1. ОБЩИЙ СЛУЧАЙ

5.1.1. Выбор нормы

В гл.4 отмечалось, что для интерполяции табулированной функции обычно используют только небольшое число узлов сетки, ближайших к искомому значению аргумента x. Если рассматривать формулу интерполяции по этим узлам при других значениях x, то за пределами выбранных узлов эта формула быстро теряет точность. Таким образом, она оказывается пригодной лишь в малом диапазоне значений аргумента. Для другого диапазона значений аргумента нужно составлять аналогичную формулу, но уже по новой группе узлов, т. е. с другими численными коэффициентами.

Если нужна единая формула, имеющая хорошую точность на большом отрезке значений $a \leqslant x \leqslant b$, то интерполяцию с небольшим числом коэффициентов построить далеко не всегда удается. Даже сплайн-интерполяция, хотя она строится сразу на большом отрезке, обеспечивает высокую точность лишь за счет использования большого числа коэффициентов.

На практике также очень важна ситуация, когда функция в узлах задана с погрешностями: $u_n \pm \delta_n$. Если погрешности δ_n не очень малы (а для экспериментально измеренной функции это обычно так), то проводить аппроксимирующую функцию точно через точки u_n бессмысленно: на рис. 5.1 прямая аппроксимирует точки не хуже и более разумно, чем интерполяционный многочлен. Это еще один довод, требующий разработки других способов аппроксимации.

При наличии ошибок разумно проводить аппроксимирующую кривую лишь достаточно близко к точкам u_n, допуская отклонение от них на величины $\sim \delta_n$. Ошибки измерения или вычисления δ_n зачастую имеют статистическую природу. Поэтому ра-

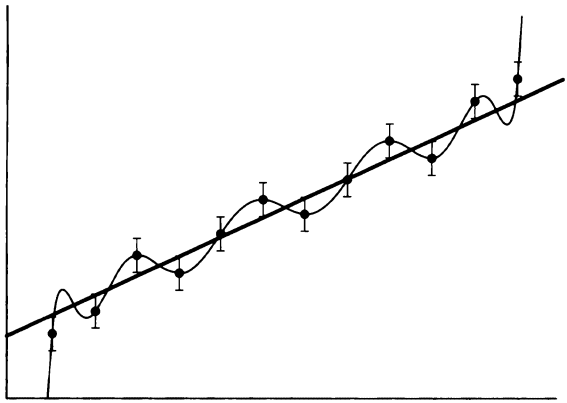


Рис. 5.1. Аппроксимация точек прямой (точки — экспериментальные точки; черточки — их погрешности; жирная линия (прямая) — аппроксимация в норме L_2; тонкая линия — интерполяционный многочлен высокой степени)

зумно понимать близость аппроксимирующей функции $\Phi(x)$ к искомой функции $u(x)$ в смысле какой-либо интегральной нормы. Например, условие близости в норме L_p есть требование малости

$$\|u-\Phi\|_{L_p} = R = \left[\frac{1}{b-a}\int\limits_a^b |u(x)-\Phi(x)|^p dx\right]^{1/p},$$

здесь p — натуральное число.

Такое приближение называют близостью *в среднем*. В гл. 1 отмечалось, что чем больше p, тем сильнее норма L_p; при $p \to \infty$ норма L_p переходит в норму C.

Покажем, что не любую норму L_p разумно выбирать. Например, выберем $p = 1$. Рассмотрим функцию $u(x) \equiv 1$, заданную на отрезке $[a = -1, b = 1]$. Попробуем найти приближение $u(x)$, наилучшее в норме L_1 функциями вида $\Phi(x) = cx$. Для этого надо решить задачу

$$\|u-\Phi\|_{L_1} \equiv \frac{1}{2}\int\limits_{-1}^{1} |1-cx|dx = \min.$$

Этот интеграл легко вычисляется; он равен

$$\|u-\Phi\|_{L_1} = \begin{cases} 1 & \text{при } |c| \leqslant 1; \\ \dfrac{c^2+1}{|2c|} > 1 & \text{при } |c| > 1. \end{cases}$$

Таким образом, любое значение $|c| \leqslant 1$ является минимизирующим. Наилучшее приближение существует, но оно не единственно. Практической пользы от такого «наилучшего» приближения очень мало.

Из-за подобных неприятностей норму L_1 в прикладной математике обычно не используют. Наиболее употребительной является норма L_2; ее называют также гильбертовой. Близость в смысле этой нормы называют *среднеквадратичной* близостью. Далее в этой главе мы будем пользоваться только нормой L_2.

Используем немного более общее определение нормы с весом $\rho(x)$. Для этого введем скалярное произведение двух комплекснозначных функций вещественного аргумента $u(x)$, $v(x)$, определенных на отрезке $a \leqslant x \leqslant b$:

$$(v,u) = \int_a^b v^*(x)u(x)\rho(x)dx, \quad \rho(x) > 0. \tag{5.1}$$

Здесь звездочка означает комплексно-сопряженную величину. Гильбертова норма выражается через скалярное произведение:

$$\|u\|_{L_2} = \sqrt{(u,u)}. \tag{5.2}$$

Заметим, что мы не вставляем делитель $b-a$ в (5.1), потому что его можно учесть множителем при выборе $\rho(x)$.

5.1.2. Аппроксимация обобщенным многочленом

Выберем на $[a, b]$ систему линейно независимых функций $\{\varphi_m(x), m = 0, 1, 2, \ldots\}$, образующих базис. Составим из начального отрезка этой системы обобщенный многочлен

$$\Phi_M(x) \equiv \sum_{m=0}^{M} c_m\varphi_m(x). \tag{5.3}$$

Построим наилучшее среднеквадратичное приближение искомой функции $u(x)$ таким многочленом. Для этого потребуем, чтобы выполнялось

$$\|\Phi_M - u\|_{L_2}^2 \equiv (\Phi_M - u, \Phi_M - u) = \min. \tag{5.4}$$

Многочлен $\Phi_M(x)$ линейно зависит от коэффициентов c_m, поэтому форма (5.4) является квадратичной функцией этих коэффициентов. Для нахождения ее минимума надо приравнять нулю ее частные производные по всем коэффициентам c_m.

Поскольку форма (5.4) симметрична относительно сомножителей скалярного произведения (с точностью до комплексного сопряжения, что несущественно), достаточно дифференцировать только первый из сомножителей. Учитывая, что $\partial \Phi_M(x)/\partial c_m = \varphi_m(x)$, получим уравнения

$$(\varphi_m, \Phi_M - u) \equiv (\varphi_m, \sum_{k=0}^{M} c_k \varphi_k - u) = 0, \quad 0 \leqslant m \leqslant M. \tag{5.5}$$

Эти уравнения преобразуются к следующему виду:

$$\sum_{k=0}^{M} (\varphi_m, \varphi_k) c_k = (\varphi_m, u), \quad 0 \leqslant m \leqslant M. \tag{5.6}$$

Получилась система из $M+1$ уравнения для определения такого же числа коэффициентов c_m.

Определитель системы (5.6) состоит из скалярных произведений (φ_m, φ_k), т. е. это определитель Грама для системы функций $\{\varphi_m(x), m = 0, 1, 2, \ldots\}$. Поскольку функции линейно независимы, определитель Грама отличен от нуля, так что линейная система (5.6) имеет решение, причем единственное. Тем самым наилучшее среднеквадратичное приближение обобщенным многочленом (5.3) всегда существует и единственно. Формулы для нахождения его коэффициентов несложны. Это показывает преимущество среднеквадратичного приближения перед другими подходами.

Известно, что матрица Грама является положительно определенной, и все ее главные миноры положительно определены. Это означает, что при решении системы (5.6) методом Гаусса главные элементы автоматически окажутся на главной диагонали, т. е. выбирать главный элемент не надо.

5.1.3. Неортогональные базисы

Пусть система функций $\{\varphi_m\}$ неортогональна в смысле скалярного произведения (5.1), порождающего норму (5.2): $(\varphi_m, \varphi_k) \neq 0$. Тогда матрица Грама будет плотно заполненной. Как правило, обусловленность ее будет плохой, причем очень быстро ухудшающейся с увеличением порядка M обобщенного многочлена. В этом случае при решении линейной системы методом Гаусса или любым другим будет очень сильно теряться

точность из-за ошибок округления. Вычисления при этом обязательно нужно проводить с полной разрядностью, доступной на компьютере. Но и этого может оказаться недостаточно. Расчеты для больших M при этом не удается провести. Поэтому неясно, насколько хорошую точность аппроксимации можно обеспечить.

Отметим еще одну неприятность. Коэффициенты обобщенного многочлена c_m зависят от выбранного порядка многочлена M; поэтому их следовало бы обозначать как c_{mM}, $0 \leqslant m \leqslant M$. Если мы увеличиваем порядок многочлена и находим следующее приближение, то при этом не просто добавляются новые слагаемые в сумму (5.3), но и меняются коэффициенты перед всеми предыдущими слагаемыми.

Ошибки округления при решении линейной системы сказываются в первую очередь на значениях коэффициентов c_{mM}. Если построить график зависимости коэффициентов c_{mM} от M $(M \geqslant m)$, то истинная зависимость должна быть плавной. Однако расчетная зависимость из-за ошибок округления становится неплавной уже при умеренных M. Несмотря на это, погрешность аппроксимации еще некоторое время продолжает уменьшаться с ростом M; на ней ошибки округления сказываются заметно позже.

Пример 5.1. Рассмотрим аппроксимацию $u(x)$ на отрезке $0 \leqslant x \leqslant 1$ системой степеней $\varphi_m(x) = x^m$, $m = 0, 1, 2, \ldots$ Скалярное произведение на этом отрезке возьмем с весом $\rho(x) \equiv 1$. Тогда

$$(\varphi_m, \varphi_k) = \int_0^1 x^{m+k} dx = \frac{1}{m+k+1} \neq 0, \quad m, k \geqslant 0. \tag{5.7}$$

Видно, что все базисные функции попарно неортогональны.

Матрица, составленная из элементов (5.7), называется матрицей Гильберта. Ее обусловленность стремительно ухудшается с ростом M. Прямые численные расчеты методом Гаусса показывают, что при увеличении M на единицу погрешность округления возрастает примерно в 30 раз, что хорошо видно из рис. 2.3. Это означает, что при 64-разрядных вычислениях уже для $M = 6$ коэффициенты c_{mM} вычисляются лишь с восемью верными знаками, а остальные восемь теряются. Для $M = 12$ в коэффициентах c_{mM} вообще не остается ни одного верного знака!

Тем самым система степеней на отрезке $0 \leqslant x \leqslant 1$ непригодна для получения аппроксимации с большим числом членов.

Реально следует ограничиваться степенями $M \approx 3 \div 4$, а это не обеспечивает высокой точности. Еще сложнее обстоит дело, если берется система степеней на отрезке $0 < a \leqslant x \leqslant b$. Обусловленность матрицы Грама будет тем хуже, чем ближе b/a к 1. Это обстоятельство настораживает, так как большинство практических обработок экспериментальных данных основано на аппроксимации системой степеней.

Пример 5.2. Для разложения по системе степеней на отрезке $a \leqslant x \leqslant b$ нетрудно предложить простой способ, существенно уменьшающий ошибки округления. Для этого начало координат переносят в точку $\bar{x} = (a+b)/2$ и масштабируют. Тогда задача сводится к аппроксимации по системе степеней на симметричном отрезке $-1 \leqslant x \leqslant 1$. Матричные элементы при этом равны

$$(\varphi_m, \varphi_k) = \int_{-1}^{+1} x^{m+k} dx = \begin{cases} 2/(m+k+1) & \text{при } m+k \text{ — четном;} \\ 0 & \text{при } m+k \text{ — нечетном.} \end{cases}$$

Базисные функции разной четности оказываются ортогональными, и половина элементов матрицы Грама обращается в нуль. Для этого случая увеличение M на единицу увеличивает ошибки округления при решении линейной системы лишь приблизительно в пять раз. (Именно эта матрица использована для рис. 2.3.) Все значащие цифры коэффициентов c_{mM} при вычислениях на 64-разрядном компьютере теряются при $M \approx 25$. В практических вычислениях можно без риска брать $M = 6 \div 8$ и с некоторой осторожностью $M = 10 \div 12$.

5.1.4. Ортогональные системы

Поиск наилучшей среднеквадратичной аппроксимации кардинально упрощается, если удается воспользоваться ортогональным базисом: $(\varphi_m, \varphi_k) = 0$ при $k \neq m$. В этом случае матрица Грама ортогональна, и решение линейной системы (5.6) сразу записывается в явном виде:

$$c_m = (\varphi_m, u)/(\varphi_m, \varphi_m). \tag{5.8}$$

Разложение по ортогональному базису называют обобщенным рядом Фурье, а коэффициенты c_m являются обобщенными коэффициентами Фурье.

Деление с плавающей точкой выполняется практически без ошибок округления. При ортогональном базисе можно брать любую степень обобщенного многочлена M, не теряя точности. Это

позволяет пользоваться большими M и добиваться очень малой погрешности. Подобные разложения очень выгодны при составлении прецизионных аппроксимаций для специальных или других трудно вычислимых функций.

Из формулы (5.8) видно, что коэффициенты c_m зависят только от своего индекса m и не зависят от M. Поэтому при увеличении M в обобщенный многочлен лишь добавляются новые слагаемые. Ранее найденные младшие коэффициенты при этом не меняются.

Погрешность. Если функция $u(x)$ удовлетворяет определенным требованиям (это могут быть требования непрерывности и ограниченности вместе с некоторым числом своих производных), то ее обобщенный ряд Фурье сходится к ней в норме L_2. При выполнении указанных требований запишем $u(x)$ в виде суммы бесконечного ряда и вычтем обобщенный многочлен. Учитывая независимость c_m от M, получим

$$\Phi_M(x) - u(x) = \sum_{m=M+1}^{\infty} c_m \varphi_m(x). \tag{5.9}$$

Возводя скалярно в квадрат правую и левую части этого равенства и учитывая ортогональность базиса, получим так называемое равенство Парсеваля

$$\|\Phi_M(x) - u(x)\|_{L_2}^2 = \sum_{m=M+1}^{\infty} |c_m|^2 (\varphi_m, \varphi_m). \tag{5.10}$$

Оно является оценкой погрешности аппроксимации. Видно, что чем быстрее убывают коэффициенты c_m, тем выше точность. Далее покажем, что скорость убывания коэффициентов c_m зависит от гладкости функции $u(x)$.

Существуют различные системы ортогональных функций. Для функций одной переменной наиболее известны классические ортогональные многочлены Лежандра, Чебышева первого и второго рода, Лаггера и Эрмита, а также тригонометрические многочлены. Имеются ортогональные функции от двух и трех переменных, часто применяемые в теоретической физике. Эти системы обычно используются, когда нужно провести точные (аналитические) вычисления.

Однако для численных расчетов большинство этих систем неудобно. Требуется находить многочлены высоких степеней $M \sim 20 \div 100$. Явные выражения для них зачастую записыва-

ются очень сложно. Существуют рекуррентные выражения многочленов высшей степени через низшие, но вычисления по соответствующим формулам приводят к очень большой потере точности из-за ошибок округления. Уже при $M \sim 10$ она заметна, а при $M \sim 20$ становится катастрофической. Кроме того, само вычисление многочлена высокой степени даже при точно известных его коэффициентах тоже может приводить к большой потере точности. Поэтому на практике удобны только те системы, в которых базисные функции высоких индексов можно вычислить на компьютере без сколько-нибудь заметной потери точности.

Таких систем очень мало. В первую очередь это тригонометрические многочлены. Стандартные программы вычисления $\sin mx$ и $\cos mx$ не дают заметных ошибок округления даже при очень больших m.

Среди алгебраических ортогональных многочленов удобны многочлены Чебышева первого рода, которые просто выражаются через тригонометрические функции. Многочлены Чебышева второго рода также выражаются через тригонометрические функции, но они менее удобны.

5.1.5. Метод наименьших квадратов

При обработке экспериментальных данных возникает следующая типовая задача. В точках $\{x_n, 0 \leqslant n \leqslant N\}$ измерены значения u_n искомой функции $u(x)$ с абсолютными погрешностями δ_n. Требуется аппроксимировать $u(x)$ обобщенным многочленом $\Phi_M(x)$ наилучшим образом. Подразумевается, что число экспериментальных точек достаточно велико, так что можно надеяться на хорошую аппроксимацию обобщенным многочленом невысокого порядка $M \ll N$.

Экспериментальную оценку погрешности δ_n получают, проводя несколько измерений в каждой точке x_n и выполняя усреднение (см. 1.2). Поэтому величина δ_n является не мажорантной оценкой погрешности, а статистической. Экспериментаторы принимают за δ_n одно (реже два) стандартных уклонения; далее будем считать, что принят один стандарт. Поэтому целесообразно рассматривать не абсолютное уклонение интерполяционного многочлена от измеренного значения, а так называемое нормированное:

$$\Delta_n = [\Phi_M(x_n) - u_n]/\delta_n. \tag{5.11}$$

Если $|\Delta_n| \leqslant 1$, то относительное уклонение не превышает стандартного, и аппроксимацию в данной точке можно считать хоро-

шей. Величина δ_n является не мажорантной оценкой погрешности u_n, а статистической (напомним, что мы приняли один стандарт). Поэтому по правилам статистики аппроксимацию можно считать достоверной, если $|\Delta_n|$ превышает 1 с вероятностью не более 38 %, 2 — с вероятностью 5 % и 3 — с вероятностью не более 0,5 %. Поэтому в качестве условия наилучшей аппроксимации целесообразно выбрать следующий критерий:

$$\Delta^2 \equiv \frac{1}{N-M}\sum_{n=0}^{N}\Delta_n^2 = \frac{1}{N-M}\sum_{n=0}^{N}\left[\frac{\Phi_M(x_n)-u_n}{\delta_n}\right]^2 = \min; \quad (5.12)$$

здесь Δ есть среднее нормированное уклонение в расчете на одну точку. В знаменателе из полного числа точек $N+1$ вычитается число свободных параметров $M+1$, так как значения этих параметров не являются абсолютно точными, а сами получаются при обработке того же экспериментального материала.

Задача является поиском наилучшего приближения, но не в гильбертовой норме L_2, а в его сеточном аналоге l_2. Эта норма определяется через сеточное скалярное произведение, которое обозначим угловыми скобками:

$$\langle v,u\rangle = \sum_{n=0}^{N}\frac{u_n v_n}{\delta_n^2}, \quad \|u\|_{l_2} = \sqrt{\langle u,u\rangle}. \quad (5.13)$$

Здесь не вводится знаменатель аналогично формуле (5.12), так как это постоянный общий делитель, не влияющий на последующие выкладки. Формула (5.13) написана для вещественных функций; для обобщения на комплексные функции в сумме надо заменить v_n на v_n^*.

Решение задачи минимизации с использованием скалярного произведения (5.13) проводится аналогично изложенному в подразд. 5.1.2 и дает линейную систему

$$\sum_{k=0}^{M}\langle\varphi_m,\varphi_k\rangle c_k = \langle\varphi_m,u\rangle, \quad 0\leqslant m\leqslant M.$$

Легко видеть, что матрица зависит только от величин x_n и δ_n, но не от u_n. Для произвольных наборов x_n и δ_n базисные функции оказываются неортогональными, а матрица Грама — плотно заполненной и чаще всего плохо обусловленной. Поэтому выбирать порядок обобщенного многочлена надо осторожно.

На практике рекомендуется следующая процедура. Выбирают некоторый базис $\{\varphi_m(x)\}$, стараясь учесть особенности поведения функции $u(x)$, если они известны. В противном случае пробуют найти выравнивающие переменные и работать в них. Затем выбирают для начала очень малое $M = 0$ или 1. Находят соответствующие коэффициенты c_m, получают обобщенный многочлен $\Phi_M(x)$, подставляют его в (5.12) и вычисляют среднее нормированное уклонение Δ.

Если оказалось $\Delta \leqslant 1$, то по правилам статистики полученная аппроксимация достоверно описывает экспериментальный материал (точнее, она достоверна с вероятностью 62 %), и вычисления можно прекратить.

Пусть $\Delta > 1$. Тогда аппроксимацию нельзя считать достоверной. Следует M увеличить на 1 и повторить описанную процедуру. При каждом увеличении M нормированное уклонение $\Delta(M)$ уменьшается, и в конце концов выполнится условие $\Delta \leqslant 1$. Полученный при этом обобщенный многочлен $\Phi_M(x)$ обеспечивает достоверную аппроксимацию экспериментального материала.

Далее увеличивать M не следует. С увеличением M многочлен $\Phi_M(x)$ будет все ближе к измеренным значениям функции, а при $M = N$ даже точно пройдет через все табулированные значения (см. рис. 5.1). Однако истинные значения $u(x_n)$ нам неизвестны, а приближать u_n с точностью больше, чем точность измерений, бессмысленно.

Если для найденного значения M выполняется $M \ll N$, причем ошибки округления еще не сказываются, то построенный $\Phi_M(x)$ следует считать хорошим приближением, а данное значение M — оптимальным. Если же процесс закончился при M, сравнимом с N ($M \sim N/2$), то результат следует считать неудачным. В этом случае система функций $\varphi_m(x)$ плохо приспособлена для аппроксимации исходной функции $u(x)$. Надо выбрать другой базис.

Кроме того, имеется еще одна трудность: даже при $M \ll N$ из-за неортогональности базиса ошибки округления могут существенно исказить коэффициенты c_m. На величине Δ ошибки округления сказываются заметно позднее, поэтому такая ситуация особенно трудна для диагностики. Для решения линейной системы рекомендуется применять те стандартные программы, в которых попутно указывается обусловленность матрицы.

Описанную процедуру называют методом наименьших квадратов. Именно ее наиболее часто используют при обработке экспериментальных данных.

5.2. ТРИГОНОМЕТРИЧЕСКИЙ РЯД ФУРЬЕ

5.2.1. Общие формулы

Если функция $u(x)$ периодическая и достаточно гладкая, то одним из лучших способов ее аппроксимации является разложение в тригонометрический ряд Фурье того же периода. Формально этот способ можно применять и к непериодической функции $u(x)$, заданной на отрезке $[a, b]$. Для этого нужно принять отрезок $[a, b]$ и периодически продолжить функцию за пределы этого отрезка. Правда, такое периодическое продолжение функции $u(x)$ будет в общем случае разрывно.

Выберем середину отрезка $[a, b]$ за начало координат и проведем масштабирование. Тогда можно считать, что функция задана на отрезке $[-\pi, \pi]$. Тригонометрический ряд Фурье на этом отрезке состоит из следующих функций:

$$\begin{aligned} &\varphi_{2m}(x) = \cos mx, \quad m = 0, 1, 2 \ldots \\ \text{и} \quad &\varphi_{2m-1}(x) = \sin mx, \quad m = 1, 2, 3 \ldots; \end{aligned} \tag{5.14}$$

очевидно $\varphi_0(x) = 1$. Легко проверить, что эти функции ортогональны на периоде с весом $\rho(x) \equiv 1$:

$$(\varphi_m, \varphi_k) = \int_{-\pi}^{\pi} \varphi_m(x)\varphi_k(x)dx = \begin{cases} 0 & \text{при } m \neq k; \\ \pi & \text{при } m = k \geqslant 1; \\ 2\pi & \text{при } m = k = 0. \end{cases} \tag{5.15}$$

Коэффициенты Фурье определяются общей формулой:

$$c_m = \frac{1}{(\varphi_m, \varphi_m)} \int_{-\pi}^{\pi} u(x)\varphi_m(x)dx. \tag{5.16}$$

Заметим, что базисные функции $\sin mx$ и $\cos mx$ отличаются друг от друга только фазой, но не длиной волны. Поэтому они, в сущности, равноправны. И подключать их в аппроксимацию нужно сразу парой, а не поодиночке. Тогда обобщенный многочлен целесообразно записать в следующем виде:

$$\Phi_M(x) = \sum_{m=0}^{M} (c_{2m-1} \sin mx + c_{2m} \cos mx), \tag{5.17}$$

несколько отличном от (5.3). Разумеется коэффициент $c_{-1} = 0$.

Выражение погрешности (5.10) принимает вид

$$\|\Phi_M - u\|_{L_2}^2 = \pi \sum_{m=M+1}^{\infty} (c_{2m-1}^2 + c_{2m}^2). \tag{5.18}$$

5.2.2. Сходимость

Пусть $u(x)$ вместе со своим периодическим продолжением имеет p производных, причем p-я производная кусочно-непрерывна и ограниченна, а младшие непрерывны. Получим для этого случая оценки скорости убывания коэффициентов. Ограничимся коэффициентами c_{2m-1}; оценки для коэффициентов c_{2m} полностью аналогичны. Многократно проинтегрируем по частям интеграл в выражении для коэффициентов (5.16). При этом учтем, что в силу периодичности и непрерывности $u^{(q)}(-\pi) = u^{(q)}(\pi)$, $0 \leqslant q \leqslant p-1$. Поэтому члены, остающиеся вне интегралов, точно сокращаются. Получим

$$\begin{aligned} \pi c_{2m-1} &= \int_{-\pi}^{\pi} u(x) \sin mx dx = \frac{1}{m} \int_{-\pi}^{\pi} u'(x) \cos mx dx = \\ &= -\frac{1}{m^2} \int_{-\pi}^{\pi} u''(x) \sin mx dx = \ldots = \\ &= \pm \frac{1}{m^p} \int_{-\pi}^{\pi} u^{(p)}(x) \begin{Bmatrix} \cos mx \\ \sin mx \end{Bmatrix} dx. \end{aligned} \tag{5.19}$$

Здесь выбор функции и знака в последнем выражении зависит от четности p.

Соотношение (5.19) позволяет сделать оценку: $|c_{2m-1}| \leqslant \leqslant 0{,}5M_p/m^p$, где $M_p = \max |u^{(p)}(x)|$. Более детальное рассмотрение с учетом кусочной непрерывности $u^{(p)}(x)$ позволяет сделать более сильную оценку. Разобьем последний интеграл в (5.19) на сумму интегралов по отрезкам непрерывности $u^{(p)}(x)$. Рассмотрим один такой отрезок $[\alpha, \beta]$. По теореме Вейерштрасса непрерывную на этом отрезке $u^{(p)}(x)$ можно со сколь угодно высокой точностью заменить многочленом достаточно высокой степени $P(x)$. Тогда интеграл по отрезку оценивается интегрированием по частям:

$$\int_\alpha^\beta u^{(p)}(x)\cos mx dx \approx \int_\alpha^\beta P(x)\cos mx dx =$$

$$= \frac{1}{m}P(x)\sin mx\Big|_\alpha^\beta - \frac{1}{m}\int_\alpha^\beta P'(x)\sin mx dx = \frac{\text{const}}{m}.$$

Суммируя такие оценки по отрезкам и подставляя в (5.19), получаем окончательную оценку:

$$c_{2m-1} = \text{const}/m^{p+1}. \tag{5.20}$$

Таким образом, коэффициенты Фурье убывают с ростом m тем быстрее, чем глаже функция $u(x)$. Если функция $u(x)$ лишь кусочно-непрерывна, т. е. $p = 0$, то $c_{2m-1} = O(1/m)$. Если $u(x)$ непрерывна, а ее первая производная кусочно-непрерывна, то $c_{2m-1} = O(1/m^2)$ и т. д.

Отсюда нетрудно сделать оценку скорости сходимости. Подставив коэффициенты (5.20) в равенство Парсеваля (5.18), получим

$$\|u - \Phi_M\|_{L_2}^2 \leqslant \text{const} \sum_{m=M+1}^{\infty} m^{-(2p+2)}. \tag{5.21}$$

Сумму нетрудно оценить, заменив ее по формуле средних интегралом (см. подразд. 3.1.5)

$$\int_{M+1/2}^{+\infty} \frac{dm}{m^{2p+2}} = \frac{1}{2p+1}(M+1/2)^{-(2p+1)}. \tag{5.22}$$

Пренебрегая разницей между M и $M + 1/2$ и извлекая корень квадратный, получим окончательную оценку

$$\|u - \Phi_M\|_{L_2} \leqslant \frac{\text{const}}{M^{p+1/2}}. \tag{5.23}$$

Порядок точности оказался полуцелым. Он тем выше, чем глаже функция $u(x)$. Для разрывной функции порядок точности в норме L_2 равен $1/2$, для непрерывной функции с конечным разрывом первой производной составляет $3/2$ и т. д.

Можно получить оценку и в норме C из общего равенства (5.9). Поскольку для тригонометрических функций $|\varphi_m(x)| \leqslant 1$, то из (5.9) и оценки (5.20) следует

$$\|u - \Phi_M\|_C = \max_x |u(x) - \Phi_M(x)| \leqslant \sum_{m=M+1}^{\infty} (|c_{2m-1}| + |c_{2m}|) \leqslant$$

$$\leqslant \text{const} \sum_{m=M+1}^{\infty} \frac{1}{m^{p+1}}.$$

Последняя сумма при $p = 0$ расходится. При $p \geqslant 1$ она аналогично оценивается интегралом. Это дает

$$\|u - \Phi_M\|_C \leqslant \frac{\text{const}}{M^p}, \quad p \geqslant 1. \tag{5.24}$$

Сходимость в норме C для $p = 0$ отсутствует. Для непрерывной функции с разрывной первой производной ($p = 1$) имеется сходимость в норме C с первым порядком точности и т. д.

Выводы. Если функция $u(x)$ периодическая и непрерывная, то лучшим способом ее аппроксимации является разложение в тригонометрический ряд Фурье. Непериодическую функцию $u(x)$ формально можно разложить в тригонометрический ряд Фурье, но это очень невыгодно. Периодическое продолжение такой функции, вообще говоря, разрывно. Сходимость тригонометрического ряда будет очень медленной, а вблизи точек разрыва — неравномерной. Например, рассмотрим $u(x) = x$, $-\pi \leqslant x \leqslant \pi$. Коэффициенты Фурье (5.16) в этом случае вычисляются точно:

$$c_{2m} = 0, \quad c_{2m-1} = 2(-1)^{m+1}/m. \tag{5.25}$$

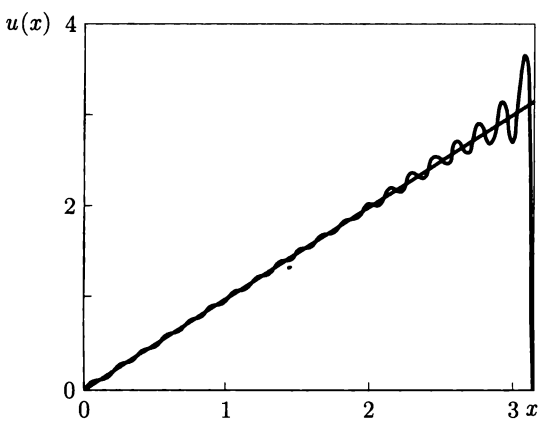


Рис. 5.2. Эффект Гиббса (показана половина отрезка и $\Phi_{40}(x)$)

Видно, что коэффициенты убывают медленно в соответствии с теоретической оценкой (5.20). На рис. 5.2 показаны функция $u(x)$ и ее тригонометрический многочлен $\Phi_{40}(x)$. В центральной части отрезка многочлен хорошо аппроксимирует функцию, но вблизи точек разрыва видны большие осцилляции. При увеличении M зона этих осцилляций уменьшается, но амплитуда максимальной осцилляции не убывает. Она стремится к $\sim 9\,\%$ от амплитуды разрыва функции периодического продолжения $u(x)$. Это явление называют эффектом Гиббса.

5.2.3. Вычисление коэффициентов

Для аппроксимации конкретной функции рядом Фурье нужно вычислить коэффициенты c_m. Они выражаются через интегралы (5.16) от произведения $u(x)$ и $\varphi_m(x)$. Как правило, вычислить эти интегралы в явном виде не удается. Приходится пользоваться какими-либо квадратурными формулами. Но вычисление интеграла по квадратурной формуле, т. е. замена интеграла дискретной суммой, означает введение нового скалярного произведения. Тогда и само наилучшее приближение нужно искать с данным определением скалярного произведения. Следовательно, необходимо элементы матрицы Грама (φ_m, φ_k) вычислять в смысле нового скалярного произведения $\langle \varphi_m, \varphi_k \rangle$. Несоблюдение этого требования приведет к тому, что найденное приближение окажется не наилучшим, а все оценки его сходимости станут нестрогими.

Равномерные сетки. Наиболее просты вычисления, если $u(x)$ уже табулирована (либо ее можно табулировать) на равномерной сетке $x_n = -\pi + nh$, $h = 2\pi/N$. Будем аппроксимировать любые интегралы формулой трапеций:

$$\int_{-\pi}^{\pi} v(x)dx = h\left(\frac{v_0}{2} + \sum_{n=1}^{N-1} v_n + \frac{v_N}{2}\right). \tag{5.26}$$

В подразд. 3.1.5 было показано, что если $v(x)$ — периодическая функция, имеющая p непрерывных производных, то погрешность формулы трапеций есть $O(h^{p+1})$.

Периодические базисные функции (5.14) имеют сколь угодно много непрерывных производных. Поэтому для них и их произведений формула трапеций (5.26) оказывается точной. Тем самым, при сеточном определении скалярного произведения мат-

ричные элементы $\langle\varphi_m, \varphi_k\rangle$ сохраняют старые значения (φ_m, φ_k) (5.15).

Остается вычислить интегралы от $v(x) = u(x)\varphi_m(x)$ по формуле трапеций (5.26). Это дает

$$c_{2m-1} = \frac{2}{N}\sum_{n=1}^{N-1} u_n \sin mx_n,$$

$$c_{2m} = \frac{2}{N}\left(\frac{u_0}{2}\cos mx_0 + \sum_{n=1}^{N-1} u_n \cos mx_n + \frac{u_N}{2}\cos mx_N\right), \quad (5.27)$$

$$c_0 = \frac{1}{N}\left(\frac{u_0}{2} + \sum_{n=1}^{N-1} u_n + \frac{u_N}{2}\right).$$

Заметим, что мы хотим определить $2M+1$ коэффициент c_m по $N+1$ табулированному значению u_n; если же $u(x)$ периодическая и непрерывная, то $u_0 = u_N$, так что фактическое число табулированных значений есть N. Поэтому следует требовать $2M+1 \leqslant N$. Для получения разумных результатов желательно более сильное требование $N/M \leqslant 4 \div 6$.

Квадратурная формула трапеций для непериодических функций относится к достаточно грубым. Однако для периодических p-гладких функций ее порядок точности есть $O(h^{p+1}) = O(N^{-p-1})$. Погрешность тригонометрического многочлена $\Phi_M(x)$ в норме C составляет $O(M^{-p})$. Поскольку $M < N/2$, дополнительная погрешность, вносимая квадратурными формулами при вычислении коэффициентов c_m, оказывается меньше погрешности самого многочлена $\Phi_M(x)$, т. е. не ухудшает общего порядка точности.

Формулы Бесселя. Если табулированные значения функции $u(x)$ известны с высокой точностью, то допустимо использовать не среднеквадратичную аппроксимацию, а интерполяцию. На равномерных сетках соответствующие формулы интерполяции получаются из (5.17) и (5.27), если N четно, и выбрано $M = N/2$. Такие формулы называют формулами Бесселя. Их часто используют в радиотехнических задачах для аппроксимации периодических сигналов, измеренных через равные промежутки времени.

Формулы Бесселя разумно применять только к функции, заданной на своем периоде. Если же $u_0 \neq u_N$, то применение этих формул дает неразумные результаты. График $\Phi_M(x)$ имеет значительные всплески вблизи границ интервала.

Неравномерные сетки. Нередко бывает, что функция уже табулирована на сетке $\{x_n, 0 \leqslant n \leqslant N\}$, которая является неравномерной. В этом случае для аппроксимации интегралов в скалярных произведениях приходится пользоваться квадратурной формулой трапеций на неравномерной сетке:

$$\int_{-\pi}^{\pi} v(x)dx \approx \frac{1}{2}\sum_{n=1}^{N}(u_n + u_{n-1})h_n. \tag{5.28}$$

Напомним, что погрешность этой формулы есть $O(M_2 h_{\max}^2)$. Более точной формулы на произвольной неравномерной сетке нет. В этом случае все скалярные произведения необходимо вычислять в смысле указанной формулы (5.28). В частности, для вычисления матричных элементов надо полагать $v(x) = \varphi_m(x)\varphi_k(x)$. Теперь все матричные элементы будут, вообще говоря, отличны от нуля. Тем самым базисные функции окажутся неортогональными в смысле сеточного скалярного произведения, матрица Грама будет плотно заполненной, а для вычисления коэффициентов c_m придется использовать общую линейную систему (5.6). Однако положение здесь не столь плохо, как для неортогональных систем функций. Будем считать, что сетка $\{x_n\}$ разумна, т. е. все шаги сетки достаточно малы: $\max h_n \ll 1$. Тогда элементы сеточной матрицы $\langle\varphi_m, \varphi_k\rangle$ отличаются от точных значений на величины $O(h_{\max}^2) \ll 1$. Значит, диагональные матричные элементы будут близки к π, а недиагональные — близки к нулю.

Такая матрица даже при больших порядках оказывается хорошо обусловленной. Система функций получается почти ортогональной. Тригонометрический многочлен $\Phi_M(x)$ вычисляется без затруднений, а влияние ошибок округления пренебрежимо мало. Однако теоретическую оценку точности дать в этом случае намного труднее.

Многомерность. Разложение в ряд Фурье легко обобщается на случай функции двух переменных $u(x, y)$, заданной на прямоугольнике. Линейным преобразованием переменных этот прямоугольник можно привести к квадрату $[-\pi \leqslant x \leqslant \pi, -\pi \leqslant y \leqslant \pi]$. В качестве базиса возьмем произведение одномерных рядов Фурье (5.14):

$$\psi_{mk}(x, y) = \varphi_m(x)\varphi_k(y).$$

Скалярное произведение естественно определяется через интеграл по квадрату. Нетрудно убедиться, что двумерный базис

ортогонален: $(\psi_{mk}, \psi_{m'k'}) = 0$, если $m \neq m'$ или $k \neq k'$. Поэтому матрица Грама диагональна, а коэффициенты двумерного разложения находятся без потери точности при любом числе членов ряда.

Описанный способ естественно обобщается на большее число переменных.

5.2.4. О равномерных приближениях

Из изложенного в 5.1 и 5.2 следует, что алгоритмы нахождения наилучших среднеквадратичных приближений очень просты. Точность наилучших среднеквадратичных приближений также оказывалась хорошей. По крайней мере, для ортогонального или почти ортогонального базиса. Хорошие оценки погрешности получались в норме L_2, а оценки в норме C были не намного хуже.

Однако норма C сильнее нормы L_2, поэтому в математической литературе, начиная с классических работ П. Л. Чебышева, много внимания уделяется построению аппроксимаций, наилучших в норме C. Такие аппроксимации называются наилучшими *равномерными*. Поэтому интересно сопоставить точности указанных двух типов наилучших приближений.

Мы обозначали тригонометрический многочлен наилучшего среднеквадратичного приближения через $\Phi_M(x)$. Обозначим аналогичный тригонометрический многочлен наилучшего равномерного приближения через $\Psi_M(x)$. По самому определению этих многочленов, каждый из них дает лучшую точность в своей норме:

$$\|u - \Phi_M\|_C \geqslant \|u - \Psi_M\|_C, \quad \text{но} \quad \|u - \Phi_M\|_{L_2} \leqslant \|u - \Psi_M\|_{L_2}.$$

Каким будет соотношение погрешностей, даваемых каждым многочленом в норме C? Для тригонометрических многочленов было показано, что

$$\|u - \Phi_M\|_C \leqslant (4{,}5 + \ln M)\|u - \Psi_M\|_C. \tag{5.29}$$

Для функций с достаточным числом непрерывных производных обычно уже порядки многочленов $M \approx 10 \div 20$ обеспечивают очень высокую точность. Даже для разрывных функций обычно хватает $M \leqslant 100$. При этом множитель в (5.29) не превышает $7 \div 9$. Поэтому наилучшие равномерные приближения даже в своей норме C дают лишь незначительный выигрыш в точности (а в норме L_2 они проигрывают).

Сходные оценки получались и для некоторых других систем базисных функций.

Алгоритмы же нахождения наилучших равномерных приближений достаточно сложны. Кроме того, дифференцирование многочленов наилучших равномерных приближений уже для первой производной дает обычно существенно худшую точность, чем для наилучших среднеквадратичных приближений. Для старших производных это еще заметнее. Поэтому на практике наиболее употребительны наилучшие среднеквадратичные приближения.

5.3. РЯДЫ ПО МНОГОЧЛЕНАМ ЧЕБЫШЕВА

5.3.1. Многочлены $T_m(x)$. Вычисление

Многочлены Чебышева I рода $T_m(x)$ легко и устойчиво вычисляются через тригонометрическую форму записи:

$$T_m(x) = \cos(m \arccos x), \quad x \in [-1, 1] \tag{5.30}$$

с использованием подпрограмм косинуса и арккосинуса. Эта форма годится только внутри отрезка $[-1, 1]$, за его пределами она приводит к вычислениям с комплексными величинами. Формула (5.30) очень устойчива даже при $m \sim 100$ и более. Убедимся, что она дает многочлен m-й степени. Введем следующие обозначения: $\theta = \arccos x$, $x = \cos\theta = (e^{i\theta} + e^{-i\theta})/2$. Последнее равенство можно записать в следующей форме: $e^{2i\theta} - 2xe^{i\theta} + 1 = 0$. Это квадратное уравнение относительно $e^{i\theta}$, его решением будет

$$e^{i\theta} = x \pm i\sqrt{1 - x^2}, \quad -1 \leqslant x \leqslant 1.$$

Здесь для $e^{i\theta}$ надо взять знак «+», а знак «−» соответствует $e^{-i\theta}$. С учетом изложенного,

$$\begin{aligned} T_m(x) = \cos m\theta = (e^{\mathrm{i}m\theta} + e^{-\mathrm{i}m\theta})/2 = \\ = [(x + i\sqrt{1 - x^2})^m + (x - i\sqrt{1 - x^2})^m]/2. \end{aligned} \tag{5.31}$$

Если возвести каждое выражение в круглых скобках в m-ю степень и сложить, то члены с нечетными степенями, содержащие мнимость и корни, имеют противоположные знаки и сокращаются. В результате останется многочлен m-й степени.

Но непосредственно пользоваться формулой (5.31) для вычисления $T_m(x)$ не следует. Она будет полезна, если $|x| > 1$. Тогда под корнем получится отрицательное число, и, внося i под знак корня, мы получим

$$T_m(x) = [(x + \sqrt{x^2-1})^m + (x - \sqrt{x^2-1})^m]/2, \quad |x| > 1. \quad (5.32)$$

Формула (5.32) — устойчивый способ вычисления многочлена Чебышева для $|x| > 1$ через подпрограммы корня и степени.

Не следует пользоваться записью $T_m(x)$ в форме обычного многочлена. Во-первых, получить такую запись для многочленов высоких степеней достаточно сложно. Во-вторых, вычисления по такой формуле при степенях $m > 10 \div 20$ приводят к неприемлемо большим ошибкам округления.

Ортогональность. Многочлены Чебышева ортогональны на отрезке $[-1, 1]$ с весом $\rho(x) = 1/\sqrt{1-x^2}$, который непрерывен внутри отрезка и обращается в бесконечность на его концах. Проверим это. Вычислим скалярное произведение с указанным весом:

$$(T_m, T_k) = \int_{-1}^{1} T_m(x) T_k(x) \frac{dx}{\sqrt{1-x^2}} = \\ = \int_0^{\pi} \cos m\theta \cos k\theta d\theta = \begin{cases} \pi & \text{при } k = m = 0; \\ \pi/2 & \text{при } k = m > 0; \\ 0 & \text{при } k \neq m. \end{cases} \quad (5.33)$$

Ортогональность многочленов $T_m(x)$ с указанным весом доказана.

Экстремальные свойства. Многочлен m-й степени имеет m нулей. Из формулы (5.30) видно, что все нули многочлена $T_m(x)$ вещественны, лежат внутри $(-1, 1)$ и определяются по формуле

$$x_{mn} = \cos(\pi(n - 1/2)/m), \quad 1 \leqslant n \leqslant m. \quad (5.34)$$

Эти нули расположены сравнительно редко вблизи середины интервала, и сгущаются вблизи его концов. Данный эффект выражен тем сильнее, чем выше степень многочлена m.

Экстремумы многочлена расположены между нулями (5.34) и попеременно принимают значения $1; -1$. Кроме того, на концах интервала многочлен принимает такие же значения: $T_m(1) = 1$, $T_m(-1) = (-1)^m$. Во всех остальных точках отрезка $[-1, 1]$ выполняется $|T_m(x)| < 1$. Поэтому многочлены $T_m(x)$ называют многочленами, наименее уклоняющимися от нуля. Если не выходить за пределы отрезка $[-1, 1]$, то многочлены Чебышева I рода оказываются очень похожими на Фурье-гармоники.

Вне отрезка $[-1,1]$ выполняется $|T_m(x)| > 1$, причем модуль многочлена быстро возрастает по мере удаления от границы отрезка. Дифференцируя многочлен в тригонометрической форме (5.30), нетрудно убедиться, что

$$\begin{aligned} T_m(\pm 1) &= (\pm 1)^m, \\ T'_m(\pm 1) &= (\pm 1)^{m+1} m^2, \\ T''_m(\pm 1) &= (\pm 1)^{m+2} m^2 (m^2-1)/3, \\ T'''_m(\pm 1) &= (\pm 1)^{m+3} m^2 (m^2-1)(m^2-4)/15. \end{aligned} \tag{5.35}$$

При больших m производные в граничных точках велики. Чем выше порядок производной, тем это сильнее выражено.

5.3.2. Разложение по $T_m(x)$

В подразд. 5.2.2 показано, что непериодическая функция $u(x)$ плохо разлагается в тригонометрический ряд Фурье. Сходимость медленная, а вблизи точек разрыва неравномерная. Многочлены $T_m(x)$ ортогональны, поэтому они гораздо лучше подходят для аппроксимации непериодической функции. Предварительно удобно отрезок задания функции линейным преобразованием аргумента перевести в отрезок $[-1,1]$. Тогда на этом отрезке разложение имеет вид

$$u(x) \approx \Phi_M(x) \equiv \sum_{m=0}^{M} c_m T_m(x), \tag{5.36}$$

где коэффициенты обобщенного ряда Фурье

$$c_m = \frac{(T_m, u)}{(T_m, T_m)} = \frac{1}{(T_m, T_m)} \int_{-1}^{1} u(x) T_m(x) \frac{dx}{\sqrt{1-x^2}}. \tag{5.37}$$

Рассмотрим, как вычислять интегралы (5.37). Явно их взять можно только в исключительных случаях. В остальных придется вводить сетку и использовать квадратурные формулы. При этом возникает два варианта.

Специальная сетка. Пусть функция $u(x)$ достаточно гладкая и можно сравнительно просто вычислить ее в любой точке. Тогда целесообразно использовать прецизионные квадратуры Эрмита — частный случай формул Гаусса—Кристоффеля

для веса $1/\sqrt{1-x^2}$ на отрезке $[-1,1]$ (см. подразд. 3.1.5). Возьмем сетку чебышевских узлов (5.34) для многочлена достаточно большой степени N. Если предполагается использовать сумму (5.36), содержащую $M+1$ многочлен Чебышева, то необходимо выбрать $N \geqslant M+1$; для получения хороших результатов рекомендуется более сильное требование $N > (2 \div 3)M$.

Напомним вид квадратурной формулы Эрмита:

$$\int_{-1}^{1} \frac{v(x)}{\sqrt{1-x^2}}dx \approx \frac{\pi}{N}\sum_{n=1}^{N} v(x_{Nn}), \qquad x_{Nn} = \cos\frac{\pi(n-1/2)}{N}, \quad 1 \leqslant n \leqslant N. \tag{5.38}$$

Погрешность этой формулы пропорциональна $\max|v^{(N+1)}(x)|$ с очень маленьким коэффициентом. Поэтому формула (5.38) дает хорошие результаты для $v(x)$ с не слишком большими высокими производными.

Мы взяли скалярное произведение (u, T_m) в смысле сеточного интеграла. Так как все скалярные произведения должны быть взяты в одинаковом смысле, то произведение (T_m, T_k) также надо брать в смысле квадратурной формулы (5.38). Но для пары многочленов Чебышева сеточный интеграл в смысле квадратурной формулы Эрмита совпадает с точным значением элемента матрицы Грама. Это легко проверяется тригонометрической заменой переменных: в этом случае сетка по x переходит в равномерную сетку по θ, а формула Эрмита — в формулу средних по переменной θ. Она так же точна на равномерной сетке, как и формула трапеций:

$$x_{Nn} = \cos\frac{\pi(n-1/2)}{N}, \quad \theta_{Nn} = \arccos x_{Nn} = \frac{\pi(n-1/2)}{N},$$

$$\langle T_m, T_k\rangle = \frac{\pi}{N}\sum_{n=1}^{N} T_m(x_{Nn})T_k(x_{Nn}) =$$

$$= \frac{\pi}{N}\sum_{n=1}^{N} \cos(m\theta_{Nn})\cos(k\theta_{Nn}) = (T_m, T_k).$$

Следовательно, можно пользоваться заранее вычисленными скалярными произведениями (5.33). Таким образом, специальная сетка для многочленов Чебышева I рода аналогична по свойствам равномерной сетке для тригонометрического ряда Фурье.

Тем самым непосредственно вычислять на этой сетке нужно только $\langle T_m, u\rangle$.

На специальной сетке многочлены Чебышева I рода остаются ортогональными, а линейная система для определения коэффициентов c_m по-прежнему ортогональна. Тем самым

$$c_m = \langle T_m, u\rangle / (T_m, T_m).$$

Ошибок округления при этом практически нет. Поэтому использование специальной сетки позволяет строить разложения (5.36) высокой точности с большим числом членов.

Произвольная сетка. Часто сетка $\{x_n, 1 \leqslant n \leqslant N\}$, на которой задана функция, определена заранее (практически всегда она включает граничные точки). Тогда положение заметно хуже: мы не можем пользоваться формулой Эрмита. Скалярное произведение $\langle T_m, u\rangle$ надо вычислять по некоторой другой квадратурной формуле. На краях отрезка подынтегральное выражение с учетом веса $\rho(x)$ обращается в бесконечность. Поэтому формулу трапеций применять нельзя, а других стандартных формул, включающих концы отрезка, для произвольной сетки у нас нет.

Разумнее всего перейти к переменной $\theta = \arccos x$. Тогда отрезок $-1 \leqslant x \leqslant 1$ переходит в $0 \leqslant \theta \leqslant \pi$, многочлены $T_m(x)$ — в $\cos m\theta$, а вместо веса $\rho(x) = 1/\sqrt{1-x^2}$ появляется вес $\rho(\theta) \equiv 1$. Задача сводится к аппроксимации тригонометрическим многочленом при скалярном произведении на сетке $\{\theta_n = \arccos x_n\}$. Эта задача уже рассмотрена в подразд. 5.2.3. Следует определить скалярное произведение по формуле трапеций:

$$\langle u, v\rangle = \sum_{n=1}^{N} \frac{\theta_n - \theta_{n-1}}{2}(u_n v_n - u_{n-1} v_{n-1}),$$

применяя ее к произведениям $\langle \cos m\theta, \cos k\theta\rangle$ и $\langle \cos m\theta, u\rangle$.

Функции $T_m(x)$ будут лишь почти ортогональны в смысле этой сетки. Но сработает то же правило, что и для обычных рядов Фурье: недиагональные элементы матрицы будут малы. Матрица Грама будет плотно заполнена, однако обусловленность линейной системы будет хорошей. Следовательно, можно брать несколько десятков членов ряда.

В этом случае точность формулы трапеций будет не очень хорошей. Наиболее часто встречается равномерная сетка $\{x_n\}$. При переходе к $\{\theta_n\}$ сетка станет резко неравномерной: вблизи

границ будут большие интервалы. Неравномерность выражена тем сильнее, чем больше N. Это обстоятельство затрудняет использование больших N и M, что препятствует получению очень высокой точности при произвольных сетках.

Сходимость. Переход к переменной θ позволяет исследовать сходимость разложения $u(x)$ по $T_m(x)$. Все доказательства проводятся по той же схеме, что и при разложении периодических функций в ряды Фурье с единственным уточнением. При многократном взятии интеграла по частям аналогично (5.19) граничные члены тоже сокращаются, но не всегда в силу периодичности.

В самом деле, теперь интегрируется по $d\theta$ выражение $v(\theta)\cos m\theta$, где $v(\theta) = u(\cos\theta)$. При интегрировании по частям вне интеграла возникают члены $v^{(2q)}(\theta)\sin m\theta$ или $v^{(2q-1)}(\theta)\cos m\theta$. Члены первого типа обращаются в нуль на пределах интегрирования $\theta = 0$ и $\theta = \pi$ благодаря обращению в нуль синуса. Члены второго типа содержат нечетные производные $v(\theta)$. Но $v(\theta) = u(\cos\theta)$ при $\theta \to 0$ разлагается в ряд, содержащий только четные степени θ; поэтому $v^{(2q-1)}(\theta)$ разлагается в ряд по нечетным степеням (θ), и обращается в нуль при $\theta \neq 0$. То же происходит при $\theta = \pi$. Поэтому члены второго типа также обращаются в нуль на границах интегрирования.

В результате доказывается, что для функции с $p-1$ непрерывной и p-й кусочно-непрерывной производной при точном вычислении интегралов в скалярных произведениях получается $c_m = O(m^{-p-1})$; погрешности аппроксимации в нормах C и L_2 оказываются $O(M^{-p})$ и $O(M^{-p-1/2})$ соответственно. Однако здесь не требуется периодического продолжения функции, поэтому несовпадение $u^{(q)}(-1)$ с $u^{(q)}(1)$ не рассматривается как разрыв и не ухудшает сходимости.

На специальной сетке (5.38) вклад погрешности квадратур мал и не снижает указанной скорости сходимости. Но погрешность на произвольной сетке может оказаться существенно больше.

Из приведенных здесь оценок видно, что сходимость разложения быстро улучшается при повышении гладкости функции $u(x)$, т. е. растет количество имеющихся у нее непрерывных производных. Поэтому разложение по многочленам Чебышева может оказаться особенно выгодным для функций высокой гладкости. В этом случае можно строить прецизионные аппроксимации (с погрешностью, близкой к ошибкам округления), причем при умеренных числах членов N.

5.4. МЕТОД ДВОЙНОГО ПЕРИОДА

5.4.1. Исключение разрывов

Разложение непериодической функции по многочленам Чебышева позволяет получить хорошие аппроксимации самой функции $u(x)$ на заданном отрезке. Однако многочлены Чебышева высоких порядков имеют очень большие производные вблизи границ исходного отрезка. Это приводит к двум неприятностям. Во-первых, ряд для производной $u(x)$, получающийся дифференцированием ряда для $u(x)$, плохо сходится вблизи границ отрезка. Во-вторых, ряд для $u(x)$ плохо экстраполируется за границы исходного отрезка.

Для периодических гладких функций эти трудности не возникают. Экстраполяция их ряда Фурье за пределы отрезка тривиальна в силу периодичности как функции, так и ряда. Почленное дифференцирование ряда Фурье не сильно ухудшает сходимость.

Поэтому заманчива идея — найти более удачный способ применения ряда Фурье к непериодическим функциям. Например, хорошо известен способ непрерывного периодического продолжения непрерывной непериодической функции $u(x)$, заданной на $a \leqslant x \leqslant b$. Для этого берут функцию с тем же граничным скачком:

$$[u(b) - u(a)](x - a)/(b - a). \tag{5.39}$$

Вычитая ее из исходной функции, получают

$$v(x) = u(x) - [u(b) - u(a)](x - a)/(b - a), \quad a \leqslant x \leqslant b. \tag{5.40}$$

Функция $v(x)$ непрерывна вместе со своим периодическим продолжением, поскольку $v(a) = v(b)$ (рис. 5.3).

Если разлагать $u(x)$ в ряд Фурье, то сходимость в норме C отсутствует, а в норме L_2 составляет $O(M^{-1/2})$. Однако ряд Фурье для $v(x)$ сходится на порядок лучше: погрешность в норме C составляет $O(M^{-1})$, а в норме L_2 есть $O(M^{-3/2})$. Добавляя к ряду Фурье для $v(x)$ функцию (5.39), получаем аппроксимацию $u(x)$ гораздо лучшей точности.

Периодическое продолжение $v(x)$ имеет разрыв первой производной, что не позволяет добиться более быстрой сходимости ряда Фурье. Пусть $u(x)$ имеет непрерывную производную, и нам известны граничные значения $u'(x)$. Тогда можно най-

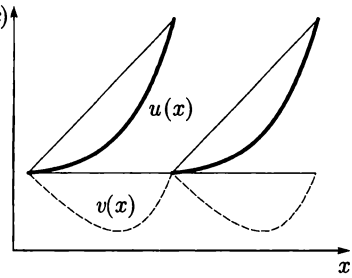


Рис. 5.3. Разложение функций в ряд Фурье (сплошная линия — $u(x)$, штриховая линия — $v(x)$ и их периодические продолжения)

ти скачок $v'(b) - v'(a)$ и удалить его. Для этого нужно домножить скачок уже не на линейную, а на квадратичную функцию $(x-b)(x-a)/(b-a)$. Получится остаток, периодическое продолжение которого непрерывно вместе с первой производной. Для этого остатка сходимость ряда Фурье еще лучше. Процесс можно продолжить, если известны более высокие производные на границах. Однако на практике обычно $u(x)$ задана только таблицей. В этом случае граничные производные неизвестны. Как обобщить прием рекуррентного исключения граничных разрывов на такой случай?

5.4.2. Двойной период

Пусть $u(x)$ — непериодическая функция, заданная на отрезке $a \leqslant x \leqslant b$, непрерывная на нем вместе со своими p производными. Линейным преобразованием аргумента переведем этот отрезок в $[-\pi/2, \pi/2]$. Мысленно продолжим $u(x)$ на вдвое бóльший отрезок $[-\pi, \pi]$ так, чтобы: а) это продолжение сохраняло p непрерывных производных; б) $u^{(q)}(\pi) = u^{(q)}(-\pi)$, $0 \leqslant q \leqslant p$. Тогда $u(x)$ будет иметь p-гладкое периодическое продолжение с периодом 2π. Такое продолжение можно сделать бесчисленным количеством способов.

Это продолжение $u(x)$ можно разложить в хорошо сходящийся ряд Фурье на отрезке $[-\pi, \pi]$. Однако возникает неопределенность с возможностью различных способов продолжения. Поэтому разобьем систему тригонометрических функций на две подсистемы:

$$\{\varphi_m(x), m = 0, 1, 2, \ldots\} = \\ = \{1, \sin 2x, \cos 2x, \sin 4x, \cos 4x, \sin 6x, \cos 6x, \ldots\} \tag{5.41}$$

$$\{\psi_k(x), k = 1, 2, \ldots, K\} = \\ = \{\sin x, \cos x, \sin 3x, \cos 3x, \sin 5x, \ldots\}. \tag{5.42}$$

Первая подсистема содержит только четные гармоники, вторая — только нечетные. Легко видеть, что подсистема (5.41) является тригонометрическим базисом на исходном отрезке $[-\pi/2, \pi/2]$. В принципе этой подсистемы достаточно для разложения искомой функции в ряд Фурье на этом отрезке без каких-либо искусственных продолжений. Подсистема (5.42) при этом не требуется. Однако сходимость ряда Фурье будет очень плохой, так как он соответствует разрывному периодическому продолжению.

Попробуем аппроксимировать исходную функцию, используя обе подсистемы:

$$u(x) \approx \Phi_M(x) + \Psi_K(x) = \sum_{m=0}^{2M} c_m \varphi_m(x) + \sum_{k=1}^{K} a_k \psi_k(x). \qquad (5.43)$$

Здесь первая сумма представляет собой ряд по функциям основного периода $[-\pi/2, \pi/2]$. Как отмечалось в 5.2, члены в этой сумме нужно добавлять парами: синус и косинус одинаковой частоты. Вторая сумма в (5.43) есть ряд по функциям только удвоенного периода $[-\pi, \pi]$. Их целесообразно добавлять не парами, а по одной.

Положим $K = 1$, т. е. возьмем только одну функцию удвоенного периода $\psi_1(x) = \sin x$. На концах основного периода $\psi_1(\pi/2) = 1$, $\psi_1(-\pi/2) = -1$. Положим $a_1 = [u(\pi/2) - u(-\pi/2)]/2$. Тогда слагаемое $a_1\psi_1(x)$ будет аналогичным (5.39): его граничный скачок таков же, как и для $u(x)$. Поэтому с его помощью можно исключить граничный разрыв аппроксимируемой функции $u(x)$ и улучшить гладкость остатка.

Аналогично с помощью $\psi_2(x)$ можно исключить граничный скачок $u'(x)$. С помощью следующих членов $\psi_k(x)$, но более сложным способом можно исключить граничные скачки высших производных.

Для непосредственного исключения нужна информация о граничных значениях $u^{(q)}(x)$. Ее, как правило, нет. Далее будет показано, что можно обойтись без нее. Вопрос о целесообразном выборе M и K также обсудим позже.

5.4.3. Наилучшее приближение

Будем искать коэффициенты обоих типов c_m и a_k из условия обеспечения наилучшей аппроксимации в норме L_2. Поскольку исходная функция задана только на $[-\pi/2, \pi/2]$, то скалярное произведение определим через интеграл по этому отрезку:

$$(u,v)=\int\limits_{-\pi/2}^{\pi/2}u(x)v(x)dx. \tag{5.44}$$

Условие наилучшего приближения принимает вид

$$\|\Phi_M+\Psi_K-u\|^2_{L_2}=(\Phi_M+\Psi_K-u,\Phi_M+\Psi_K-u)=\min. \tag{5.45}$$

Скалярное произведение (5.45) является квадратичной функцией от коэффициентов обоих типов. Для нахождения минимума нужно приравнять нулю производные по всем коэффициентам. Учитывая, что $\partial\Phi_M/\partial c_m=\varphi_m$ и $\partial\Psi_K/\partial a_k=\psi_k$, и выполняя выкладки, аналогичные приведенным в подразд. 5.1.2, получим две группы уравнений:

$$\begin{aligned}
&\sum_{l=0}^{2M}(\varphi_m,\varphi_l)c_l+\sum_{q=1}^{K}(\varphi_m,\psi_q)a_q=(\varphi_m,u),\quad 0\leqslant m\leqslant 2M;\\
&\sum_{l=0}^{2M}(\psi_k,\varphi_l)c_l+\sum_{q=1}^{K}(\psi_k,\psi_q)a_q=(\psi_k,u),\quad 1\leqslant k\leqslant K.
\end{aligned} \tag{5.46}$$

Эти группы образуют систему линейных уравнений порядка $2M+K+1$ для определения такого же числа неизвестных коэффициентов c_l,a_q. Все функции φ_m,ψ_k линейно независимы. Поэтому определитель матрицы Грама всей системы (5.46) отличен от нуля, а система имеет решение, причем единственное.

Рассмотрим структуру матрицы Грама. Она состоит из элементов трех типов и разбивается на четыре клетки (рис. 5.4). Эти элементы легко вычисляются. Подсистема $\{\varphi_m\}$ ортогональна, и в первой квадратной клетке заполнена только диагональ:

$$(\varphi_m,\varphi_l)=\begin{cases}\pi & \text{при } m=l=0;\\ \pi/2 & \text{при } m=l\neq 0;\\ 0 & \text{при } m\neq l.\end{cases} \tag{5.47}$$

Подсистема $\{\psi_k\}$ также ортогональна, а вторая квадратная клетка диагональна:

$$(\psi_k,\psi_q)=\begin{cases}\pi/2 & \text{при } k=q;\\ 0 & \text{при } k\neq q.\end{cases} \tag{5.48}$$

Однако подсистемы $\{\varphi_m\}$ и $\{\psi_k\}$ взаимно не полностью ортогональны. Синус каждой подсистемы ортогонален любому косину-

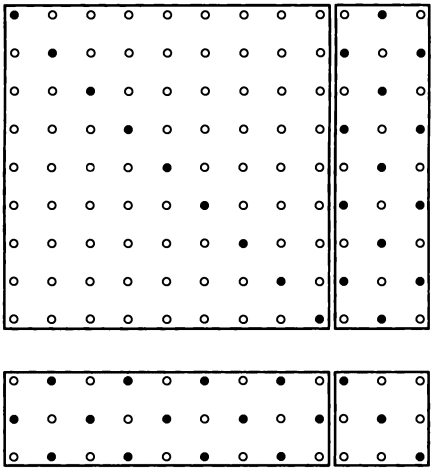


Рис. 5.4. Матрица Грама системы (5.46) $M = 4$, $K = 3$ (светлые кружки — нулевые элементы, темные — ненулевые)

су другой подсистемы в силу их различной четности. Остальные скалярные произведения не равны нулю:

$$
\begin{gathered}
(\varphi_{2m}, \psi_{2k-1}) = (\varphi_{2m-1}, \psi_{2k}) = 0, \\
(\varphi_{2m}, \psi_{2k}) = (\cos 2mx, \cos(2k-1)x) = \\
= \frac{(-1)^{m-k}(4k-2)}{4m^2 - (2k-1)^2}, \\
(\varphi_{2m-1}, \psi_{2k-1}) = (\sin 2mx, \sin(2k-1)x) = \\
= \frac{(-1)^{m-k} 4m}{4m^2 - (2k-1)^2}.
\end{gathered}
\tag{5.49}
$$

Таким образом, прямоугольные клетки взаимно симметричны и заполнены в шахматном порядке.

Напомним, что матрица Грама симметричная и положительно определенная вместе со своими главными минорами. Поэтому метод Гаусса для нее не требует выбора главного элемента. В ней очень много нулей, которые нужно обходить для экономичности вычислений. Для такой матрицы целесообразно составлять специальную программу метода исключений. Можно заметить, что четные и нечетные коэффициенты в системе (5.46) не «зацепляются» друг за друга. Поэтому можно выбрать отдельно четные и нечетные строки системы, и разбить ее на две независимые системы уравнений меньшей размерности. Такой способ решения существенно экономичнее и требует меньшего объема оперативной памяти.

Обусловленность. Напомним, что при полном отсутствии второй подсистемы ($K = 0$) уже только первая подсистема при

Таблица 5.1

Предельное M для метода двойного периода при 64-разрядных вычислениях ($\delta = 10^{-16}$)

ε	K							
	0	1	2	3	4	5	6	7
1	∞	171 000	399	41,0	14,4	7,7	5,1	3,8
10^{-4}	∞	7 940	71	14,7	6,7	4,2	3,0	2,4*
10^{-6}	∞	1 710	33	8,8	4,6	3,1	2,4*	2,0*

$M \to \infty$ обеспечивает сколь угодно точную аппроксимацию $u(x)$ в норме L_2. Значит, если взять $M \to \infty$ и $K > 0$, то задача нахождения наилучшего приближения (5.45) окажется переопределенной. Соответственно матрица Грама при больших M и даже небольших K будет плохо обусловленной.

Более подробный анализ показывает, что при увеличении M обусловленность медленно ухудшается, а с ростом K ухудшается катастрофически: ошибка единичного округления δ увеличивается в $\alpha_K(2M)^{2K-1}$ раз, где $\alpha_K \approx 1$. Расчет с выбранным K означает, что экстраполированная функция $u(x)$ имеет K непрерывных производных. Поэтому коэффициенты ее разложения по первой подсистеме убывают как β/m^{K+1}; здесь коэффициент β зависит от $u^{(K)}(x)$. Если мы хотим вычислять коэффициенты разложения с погрешностью ε, то для наименьшего, т. е. последнего, коэффициента разложения нужно приравнять желательную ошибку $\varepsilon\beta/M^{K+1}$ к погрешности округления $\delta(2M)^{2K-1}$. Отсюда получаем предельно допустимое M:

$$M \leqslant [(\varepsilon/\delta)(\beta/\alpha_K)]^{1/(3K)} \cdot 2^{-(2K-1)}/(3K). \qquad (5.50)$$

Отличием α_K от единицы здесь можно пренебречь. Величину β оценить трудно, но ее влияние на результат не очень велико из-за возведения в близкую нулю степень. Если нам нужно хорошо аппроксимировать лишь функцию, а значение коэффициентов ее разложения несущественно, то допустимо принять $\varepsilon \sim 1$. (Ошибка последнего коэффициента близка к нему самому.) Если нужно хорошо вычислять коэффициенты, целесообразно полагать $\varepsilon \sim 10^{-4} \div 10^{-6}$. Оценка предельного M для этих случаев при 64-разрядных вычислениях приведена в табл. 5.1.

Напомним, что число функций первой подсистемы $\{\varphi_m\}$ равно $2M + 1$ (в клетках табл. 5.1 они помечены звездочками), $2M + 1 \leqslant K$; ими пользоваться нецелесообразно.

Для компьютеров с большей разрядностью можно увеличивать значения M согласно оценке (5.50). Использовать вычисления с 32 разрядами не следует — ошибки округления при вычислении c_m, a_k неприемлемо велики и не позволяют достичь разумной точности.

Погрешность. В подразд. 5.4.2 отмечалось, что можно выбрать коэффициенты перед функциями $\psi_k(x)$ так, чтобы a_1 исключало граничный скачок самой функции, а a_2 — граничный скачок $u'(x)$ и т. д.

Пусть $u(x)$ имеет на отрезке $[-\pi/2, \pi/2]$ достаточно много непрерывных производных. Положим $K = 1$ и выберем a_1 так, чтобы исключить граничный скачок $u(x)$. Тогда периодическое продолжение остатка будет непрерывным, но с разрывом первой производной. Следовательно, при увеличении M первая подсистема обеспечит сходимость со скоростью $O(M^{-3/2})$ в норме L_2 и $O(M^{-1})$ в норме C.

Если для той же функции выбрать a_1 из условия минимизации (5.45), то погрешность в норме L_2 может только уменьшиться по сравнению с первым выбором, поскольку теперь мы строим наилучшее приближение в норме L_2. Значит, погрешность в норме L_2 будет не хуже, чем $O(M^{-3/2})$. Если учесть связь (5.26) между наилучшими приближениями в нормах C и L_2, то для нового способа выбора a_1 погрешность аппроксимации в норме C будет не хуже, чем $O(M^{-1} \ln M)$.

Аналогичные соображения справедливы для $K > 1$. На их основе можно доказать следующую теорему.

Теорема 5.1. Пусть $u(x)$ непрерывна со всеми своими производными до $u^{(p)}(x)$ включительно на отрезке $[-\pi/2, \pi/2]$, и выбрано $K \leqslant p + 1$. Тогда наилучшее приближение (5.45) имеет погрешность $O(M^{-K-1/2})$ в норме L_2 и $O(M^{-K} \ln M)$ в норме C на этом отрезке, включая граничные точки.

Видно, что увеличение K и M по-разному влияет на сходимость. Увеличение M эквивалентно увеличению числа членов классического ряда Фурье. Увеличение K от 0 до $p + 1$ эквивалентно повышению гладкости периодического продолжения $u(x)$, т. е. исключению граничных скачков функции и ее производных. Эта процедура повышает скорость сходимости классического ряда Фурье. Бессмысленно брать $K > p + 1$: гладкость периодического продолжения при этом не улучшается.

Выбор. Обсудим выбор M и K. Если проводить вычисления с бесконечной разрядностью, то следовало бы брать $K = p + 1$.

Однако при конечной разрядности вычислений фактор плохой обусловленности может стать решающим. Для хорошей точности необходимы еще и большие значения M, а они ограничены для каждого K (см. табл. 5.1).

Поэтому на практике пользуются небольшими K. Рекомендуется следующая процедура. Сначала полагают $K = 1$ и начинают увеличивать M. Для каждого M находят коэффициенты и непосредственно вычисляют полученную в норме L_2 погрешность по формулам (5.45) и (5.44).

Затем выбирают $K = 2$ и повторяют тот же процесс увеличения M, не переходя предельно допустимого значения из табл. 5.1. Проделав такие расчеты для разных K, выбирают наилучший с точки зрения точности и экономичности результат. Под экономичностью здесь понимается малость числа членов в окончательном разложении (5.43). Можно ограничиться упрощенной процедурой. Если нужна прецизионная аппроксимация плавно меняющейся функции, то целесообразно взять $K = 3$. Если функция имеет достаточно сложное поведение (много экстремумов, перегибов и т. п.), то лучше взять $K = 2$ или даже 1. В последнем случае точность аппроксимации вряд ли будет высокой. Разумеется, во всех случаях значения M ограничены табл. 5.1.

Экстраполяция. Теорема дает оценку погрешности только на исходном отрезке $[-\pi/2, \pi/2]$. Пусть нам известно, что искомая функция достаточно гладкая и существует вне этого отрезка. Какова будет погрешность построенной аппроксимации при $|x| > \pi/2$, т. е. возможна ли разумная экстраполяция за пределы исходного отрезка?

Практика расчетов показывает, что разумная экстраполяция на небольшую долю длины исходного отрезка (10—20 %) зачастую возможна. Теоретически этот вопрос не исследован. По-видимому, допустимые пределы экстраполяции тем шире, чем дальше отстоят от исходного отрезка особые точки аналитического продолжения $u(x)$ в комплексную плоскость.

5.4.4. Вычисление скалярных произведений

Скалярные произведения (φ_m, u) и (ψ_k, u) в смысле интеграла (5.44) лишь в исключительных случаях удается вычислить точно. На практике приходится заменять интеграл квадратурной формулой. Напомним, что в этом случае необходимо по той же квадратурной формуле вычислять все элементы матрицы Грама (5.46).

Пусть функция $u(x)$ табулирована на произвольной неравномерной сетке $\{x_n, 0 \leqslant n \leqslant N\}$. При этом естественны следующие требования на сетку: а) точки x_0 и x_N являются границами отрезка; б) точки x_n расположены достаточно плотно на отрезке, т. е. максимальный из шагов невелик: $\max(x_n - x_{n-1}) \ll \pi$. В этом случае оптимальна формула трапеций:

$$\langle v, u\rangle = \sum_{n=1}^{N} (u_n v_n + u_{n-1} v_{n-1}) h_n/2, \quad h_n = x_n - x_{n-1}. \qquad (5.51)$$

Теперь любая пара базисных функций, вообще говоря, неортогональна в смысле сеточного скалярного произведения (5.51). Матрица Грама (5.46) станет плотно заполненной (но по-прежнему симметричной и положительно определенной). Решать ее придется полным методом Гаусса без выбора главного элемента.

Погрешность формулы трапеций (5.51) невелика: $O(\max h_n^2)$. Поэтому сеточные матричные элементы будут близки к их интегральным аналогам. В матрице на рис. 5.4 светлым кружкам будут соответствовать малые элементы $O(\max h_n^2)$, а значения элементов, соответствующих темным кружкам, почти не изменятся. Обусловленность матрицы Грама лишь немного ухудшится.

Для страховки от ошибок округления нужно останавливать расчет при несколько меньших значениях M, чем указано в табл. 5.1. Число узлов сетки должно быть больше числа свободных параметров: $N+1 > 2M+K+1$. Однако вычисления на грани этого неравенства рискованны из-за тех же ошибок округления. На практике рекомендуется использовать достаточно большое число узлов сетки, хотя бы $N \geqslant (2 \div 3)(2M + K)$.

Равномерные сетки. Более благоприятен для вычисления случай равномерной сетки $x_n = -\pi/2 + \pi n/N$. При этом элементы матрицы Грама вычисляются в явном виде аналогично тому, как это делалось для классического ряда Фурье. Соотношения ортогональности для базисных функций остаются теми же, что при интегральном скалярном произведении, так что структура матрицы Грама точно соответствует рис. 5.4. Для элементов (φ_m, φ_l) и (ψ_k, ψ_q) остаются справедливыми формулы (5.47) и (5.48). Формулы для элементов (φ_m, ψ_k) несколько усложняются по сравнению с (5.49), но их также можно записать в явном виде.

На практике рекомендуется проводить вычисления следующим образом. Сначала в матрице Грама нужно проставить нулевые элементы в соответствии с рис. 5.4, затем диагональные элементы $\langle\varphi_m, \varphi_l\rangle$ и $\langle\psi_k, \psi_q\rangle$, используя их интегральные значе-

ния (φ_m, φ_l), (ψ_k, ψ_q). Ненулевые элементы $\langle \varphi_m, \psi_k \rangle$ и правые части $\langle \varphi_m, u \rangle$, $\langle \psi_k, u \rangle$ вычисляют по формуле трапеций (5.51).

5.5. АППРОКСИМАЦИЯ СПЛАЙНАМИ

5.5.1. B-сплайны

В качестве аппроксимирующих функций можно брать сплайны. Однако для построения аппроксимаций, наилучших в норме L_2, локальная форма записи сплайна, приведенная в подразд. 4.2.2, неудобна. Алгоритм для нахождения коэффициентов наилучшего локального сплайна при этом оказывается исключительно громоздким. Универсальной записи такого алгоритма для произвольной степени сплайна p пока не построено. Гораздо проще строятся алгоритмы для так называемой базисной формы записи сплайна.

Базисным сплайном (или B-сплайном) называют сплайн с узлами $\{x_n, 0 \leqslant n \leqslant N\}$, отличный от нуля на минимальном количестве интервалов (или, как говорят, на минимальном носителе). Вне этого носителя он равен нулю, непрерывен вместе со своей $(p-1)$-й производной на носителе (включая переход в нулевой фон на концах носителя); а его p-я производная разрывна в узлах. Базисный сплайн p-й степени дефекта 1, носитель которого начинается в узле x_n, будем обозначать через $B_{pn}(x)$. Рассмотрим, как строятся такие сплайны разных степеней на сетке.

Начнем со степени сплайна $p = 0$. Такой сплайн на каждом интервале является многочленом нулевой степени, т. е. константой. Его дефект равен 1, следовательно, во всех узлах он разрывен. Можно построить сплайн нулевой степени $B_{0n}(x)$, отличный от 0 на одном интервале (т. е. его носителем является один интервал). Для этого полагаем, что левее точки x_n сплайн равен 0. На интервале $(x_n, x_{n+1}]$ это ненулевая константа. Правее точки x_{n+1} сплайн снова равен 0 (рис. 5.5). На всем отрезке $[x_0, x_N]$ существует N таких сплайнов, по одному на каждом интервале.

Сплайн первой степени $(B_{1n}(x))$ состоит из многочленов первой степени, т. е. линейных кусков. Он непрерывен, но в узлах имеет разрыв первой производной. Тем самым, это ломаная. Построим ее. Снова положим сплайн равным 0 левее точки x_n. Из точки x_n проведем первое звено ломаной до точки x_{n+1} (см. рис. 5.5). Продолжить нулем сплайн из точки x_{n+1} нельзя — получится разрывная линия вместо непрерывной. Поэтому из точки x_{n+1} построим второе звено ломаной, приходящее в нуль в

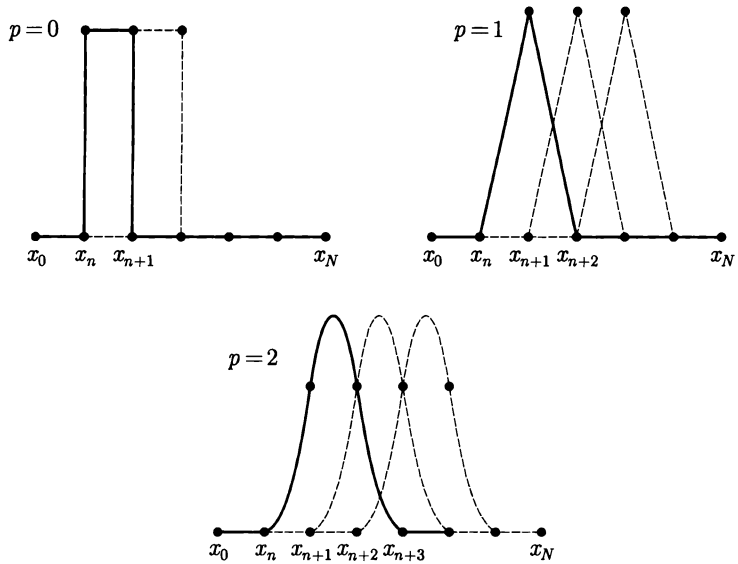


Рис. 5.5. B-сплайны для $p = 0, 1, 2$ (штриховые линии соответствуют сдвигам носителя)

точке x_{n+2}. Теперь можно положить сплайн равным нулю правее точки x_{n+2}: непрерывность соблюдена. Носитель такого сплайна состоит из двух соседних интервалов.

Однако теперь возникает некоторая неоднозначность. Сплайн $B_{1,N-1}(x)$, начинающийся в узле x_{N-1}, своей второй половинкой выходит правее точки x_N. Кроме того существует сплайн, кончающийся в точке x_1, его первая половина лежит левее точки x_0, а сам сплайн следует обозначать $B_{1,-1}(x)$. Таким образом, полное число сплайнов равно $N + 1$, т. е. больше числа интервалов сетки. Для построения «вылезающих» за границу сплайнов приходится дополнять сетку двумя фиктивными узлами x_{-1}, x_{N+1}.

Аналогично строится сплайн $B_{2n}(x)$ степени $p = 2$. Левее точки x_n берется горизонтальный нулевой участок. В точке x_n к нему гладко, но с разрывом второй производной «приклеивается» кусочек параболы, идущий вверх. В точке x_{n+1} «приклеивается» второй параболический участок. На этом нельзя остановиться, так как этот участок приходит к оси абсцисс с ненулевой первой производной. Поэтому к нему в точке x_{n+2} гладко «приклеивается» третий параболический участок. Он должен придти в точку x_{n+3} с нулевой производной. Тогда правее точки x_{n+3} он гладко продолжается нулем. Таким образом, параболический сплайн имеет носитель из трех соседних интервалов. Число та-

ких сплайнов равно $N+2$. Для их построения нужно вводить еще два фиктивных узла: x_{-2}, x_{N+2}.

Описанным способом можно строить сплайны более высоких степеней.

Явные формулы. Для удобства записи формул введем так называемую усеченную степень — обобщенную функцию Хевисайда (рис. 5.6):

$$\xi_+^p = \begin{cases} \xi^p & \text{при } \xi > 0, \\ 0 & \text{при } \xi \leqslant 0, \end{cases} \quad p = 0, 1, 2, \ldots \tag{5.52}$$

Эта функция непрерывна всюду вместе со своими производными вплоть до $(p-1)$-й. Ее p-я производная разрывна в точке $\xi = 0$. Поэтому ее называют также элементарным скачком p-й производной (или p-й степени).

B-сплайны легко записываются через обобщенную функцию Хевисайда. Для нулевой степени сплайн на своем носителе равен

$$B_{0n}(x) = \begin{cases} (x - x_n)_+^0 = 1 & \text{при } x_n < x \leqslant x_{n+1}, \\ 0 & \text{вне носителя,} \end{cases} \quad 0 \leqslant n \leqslant N-1. \tag{5.53}$$

Сплайн первой степени (линейный) на своем носителе равен

$$B_{1n}(x) = \frac{(x - x_n)_+^1}{x_{n+1} - x_n} - \frac{(x_{n+2} - x_n)(x - x_{n+1})_+^1}{(x_{n+1} - x_n)(x_{n+2} - x_{n+1})}, \quad x_n < x \leqslant x_{n+2} \tag{5.54}$$

и обращается в нуль вне носителя $(x_n, x_{n+2}]$. Параболический сплайн $(p=2)$ на своем носителе $(x_n, x_{n+3}]$ имеет вид

$$\begin{aligned} B_{2n}(x) = {} & \frac{(x - x_n)_+^2}{(x_{n+1} - x_n)(x_{n+2} - x_n)} - \\ & - \frac{(x_{n+3} - x_n)(x - x_{n+1})_+^2}{(x_{n+1} - x_n)(x_{n+2} - x_{n+1})(x_{n+3} - x_{n+1})} + \\ & + \frac{(x_{n+3} - x_n)(x - x_{n+2})_+^2 \cdot}{(x_{n+2} - x_n)(x_{n+2} - x_{n+1})(x_{n+3} - x_{n+2})}, \\ & x_n < x \leqslant x_{n+3}, \end{aligned} \tag{5.55}$$

и также обращается в нуль за пределами отрезка $(x_n, x_{n+3}]$. С увеличением степени сплайна эти частные формулы становятся все более громоздкими.

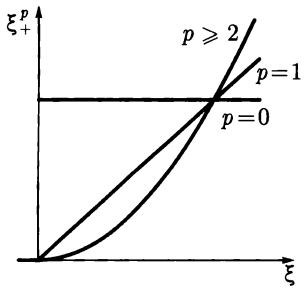


Рис. 5.6. Обобщенная функция Хевисайда (усеченная степень) для $p = 0, 1, 2, \ldots$

Можно записать подобные формулы в общем виде для произвольной степени сплайна p

$$B_{pn}(x) = \sum_{k=0}^{p} b_{pnk}(x - x_{n+k})_+^p \text{ при } x \in (x_n, x_{n+p+1}], \qquad (5.56)$$

которая обращается в нуль вне носителя $(x_n, x_{n+p+1}]$. Здесь

$$b_{pnk} = (x_{n+p+1} - x_n) \Big/ \prod_{l=0, l\neq k}^{p+1} (x_{n+l} - x_{n+k}), \qquad (5.57)$$
$$0 \leqslant k \leqslant p, \quad -p \leqslant n \leqslant N-1.$$

Видно, что сплайн p-й степени имеет носитель из $(p+1)$-го соседнего интервала. Число различных базисных сплайнов p-й степени на всем отрезке $[x_0, x_N]$ равно $N+p$, а для их построения надо ввести по p фиктивных узлов справа и слева: $\{x_n, -p \leqslant n \leqslant N + p\}$.

Анализируя знаки сомножителей в (5.57), нетрудно установить, что $\operatorname{sgn} b_{pnk} = (-1)^k$, т. е. нулевой член в сумме (5.56) положителен, а далее знаки чередуются. Очевидно, B-сплайн определяется с точностью до постоянного множителя. Этот множитель в формулах (5.53) — (5.57) выбран так, чтобы площадь под графиком сплайна равнялась

$$\int_{x_n}^{x_{n+p+1}} B_{pn}(x)dx = \frac{x_{n+p+1} - x_n}{p+1}. \qquad (5.58)$$

На равномерной сетке $h_n = x_{n+1} - x_n = \text{const} = h$ вид коэффициентов сильно упрощается:

$$b_{pnk} = \frac{(-1)^k(p+1)}{h^p k!(p+1-k)!}, \quad 0 \leqslant k \leqslant p, \quad -p \leqslant n \leqslant N-1, \quad (5.59)$$

а площадь под графиком сплайна просто равна h. Отсюда видно, что хотя длины носителей сплайнов разных степеней равны $(p+1)h$ и довольно сильно различаются, эффективные ширины графиков сплайнов на самом деле близки.

Расчет B-сплайнов по указанным формулам (5.56), (5.57) хорошо проводится для невысоких степеней $p \leqslant 5$. При дальнейшем увеличении степени начинают заметно возрастать ошибки округления из-за сложения знакопеременных членов в (5.56). При $p \geqslant 10$ эти ошибки сильно сказываются на вычислениях, а при $p \approx 20$ становятся катастрофически большими. Поэтому данные формулы целесообразно использовать лишь при небольших степенях сплайна.

Рекуррентный алгоритм. Существует рекуррентный алгоритм, выражающий базисный сплайн $(p+1)$-й степени через два базисных сплайна p-й степени. Приведем эту формулу без вывода:

$$B_{p+1,n}(x) = \frac{(x - x_n)B_{p,n}(x)}{x_{n+p+1} - x_n} + \frac{(x_{n+p+2} - x)B_{p,n+1}(x)}{x_{n+p+2} - x_{n+1}}. \qquad (5.60)$$

Формулу (5.60) можно непосредственно проверить на сплайнах вплоть до второй степени, используя (5.53) — (5.55).

Алгоритм (5.60) можно употреблять для вывода устойчивых явных формул вычисления сплайнов более высоких степеней. Однако такие формулы получаются слишком громоздкими, а их компактного общего вида не найдено.

Алгоритм (5.60) выгодно использовать совершенно иначе. Можно взять значения сплайнов нулевой степени (5.53) для заданного значения x, а затем рекуррентно повышая p на единицу, вычислять значения B-сплайна при данном x до тех пор, пока не дойдем до сплайна нужной степени. Такой алгоритм гораздо более устойчив, чем непосредственное вычисление сплайна по формулам типа (5.53) — (5.57). Его ошибки округления пренебрежимо малы, по крайней мере до $p \approx 20$, что превосходит потребности обычной практики.

5.5.2. Среднеквадратичная аппроксимация

Общее выражение для сплайна на всем отрезке $[x_0, x_N]$ составляется как линейная комбинация базисных сплайнов по всем носителям. При этом надо включать в сумму B-сплайны, выходящие за границы отрезка:

$$S_p(x) = \sum_{n=-p}^{N-1} c_n B_{pn}(x). \qquad (5.61)$$

Это выражение есть обобщенный многочлен, а его базисом является набор базисных сплайнов p-й степени. Хотя B-сплайны строятся на расширенной сетке $\{x_n, -p \leqslant n \leqslant N+p\}$, пользоваться формулой (5.61) следует только на отрезке $[x_0, x_N]$.

Аппроксимацию сплайнами, наилучшую в норме L_2, надо строить по обычным правилам, описанным в 5.1. Все скалярные произведения при этом берутся по исходному отрезку:

$$(B_{pn}, B_{pm}) = \int_{x_0}^{x_N} B_{pn}(x) B_{pm}(x) dx, \quad -p \leqslant n, m \leqslant N-1. \quad (5.62)$$

Аналогично записывается (B_{pn}, u). Для нахождения $N+p$ коэффициентов c_n получается система такого же числа линейных уравнений

$$\sum_{m=-p}^{N-1} (B_{pn}, B_{pm}) c_m = (B_{pn}, u), \quad -p \leqslant m \leqslant N-1.$$

Из рис. 5.5 видно, что сплайны нулевой степени ортогональны друг другу: носитель сплайна B_{0n} не перекрывается даже с носителем ближайшего соседа $B_{0,n+1}$, не говоря уже о более далеких. Поэтому скалярное произведение (5.62) любой пары различных сплайнов B_{0n} равно нулю. Матрица Грама при этом диагональна, и нахождение коэффициентов c_n производится без потери точности. Но этот случай неинтересен для практики из-за малой степени сплайна.

При $p = 1$ носитель каждого B-сплайна перекрывается с носителями одного правого и одного левого соседа. С более далекими соседями перекрытия носителя нет. Поэтому в каждой строке матрицы Грама отличными от нуля будут только три скалярных произведения: диагональное (B_{1n}, B_{1n}) и два соседних $(B_{1n}, B_{1,n\pm1})$. Матрица Грама оказывается трехдиагональной. Из слабого перекрытия графиков соседних сплайнов при $p = 1$ (см. рис. 5.5) видно, что скалярное произведение на главной диагонали будет сильно преобладающим (см. табл. 5.2), поэтому трехдиагональная линейная система для коэффициентов c_n устойчиво решается методом Гаусса или прогонкой.

Для $p = 2$ каждый B-сплайн перекрывается с двумя правыми и двумя левыми соседями. Ненулевые элементы будут уже на пяти диагоналях: (B_{2n}, B_{2n}), $(B_{2n}, B_{2,n\pm1})$ и $(B_{2n}, B_{2,n\pm2})$. Перекрытие с соседями более сильное; однако можно показать, что

Таблица 5.2

Обусловленность вычисления B-сплайна

p	1	2	3	4	5	6	7
$(B_n, B_n)/\sum_{n\neq k}(B_n, B_k)$	2,000	1,222	0,921	0,756	0,650	0,575	0,520
Допустимые N	∞	∞	200	125	80	65	50

преобладание главной диагонали еще сохраняется. Поэтому линейная система с пятидиагональной матрицей Грама устойчиво решается методом Гаусса.

При $p = 3$ матрица Грама семидиагональная. Перекрытие с соседями настолько значительно, что преобладания главной диагонали уже нет. При решении линейной системы накапливаются ошибки округления, но еще слабо. Их накопление может сказаться только при больших числах узлов N. Поэтому кубическим сплайном можно безопасно пользоваться.

При дальнейшем увеличении степени p количество диагоналей в матрице Грама возрастает (оно равно $2p+1$), а ее обусловленность ухудшается. Поэтому большими p надо пользоваться осторожно. В табл. 5.2 приведены отношения диагональных коэффициентов к сумме недиагональных для случая равномерной сетки, и допустимые числа N в зависимости от степени сплайна при 64-разрядных вычислениях.

Повышение указанных пределов N опасно из-за сильного накопления ошибок округления. Однако при меньших N устойчивость алгоритма остается хорошей.

Вычисление скалярных произведений. Элементы матрицы Грама можно вычислить точно, однако делать это нецелесообразно. Все равно для различных исходных функций $u(x)$ скалярные произведения (B_{pn}, u) точно удастся вычислить лишь в исключительных случаях. Обычно приходится пользоваться квадратурными формулами, особенно если функция $u(x)$ табулирована. Пусть функция задана своими табличными значениями $u_j = u(x_j)$, $0 \leqslant j \leqslant J$. Точки x_j никак не связаны с расположением узлов сплайна x_n (хотя точки с $j = 0$ и $j = J$ естественно должны быть концами отрезка). Как выбрать квадратурную формулу для скалярного произведения?

Если мы используем сплайн p-й степени, то неявно предполагаем, что существует непрерывная $u^{(p-1)}(x)$. Бессмысленно брать сплайн более гладким, чем исходная функция. В этом случае разумно пользоваться квадратурными формулами точности

вплоть до $O(h^{p-1})$. Если сетка $\{x_j\}$ равномерная, то такие формулы высокого порядка точности зачастую можно подобрать (например, формула Симпсона для $p=5$). Если же сетка неравномерна, то придется ограничиться формулой трапеций.

Для получения разумных результатов число табулированных значений функции должно быть много больше, чем число параметров сплайна $J \gg N+p$ (хотя бы в два-три раза).

Узлы сплайна. Если точки $\{x_j\}$ нам заданы, то узлы сплайна $\{x_n\}$ мы можем выбрать сами. Из предыдущего видно, что их число должно быть $N \ll J$. Однако этого недостаточно. Необходимо еще, чтобы внутрь каждого интервала $[x_n, x_{n+1}]$ попало достаточно много табулированных точек x_j (хотя бы две-три). Недопустимо, чтобы в какой-либо из интервалов $[x_n, x_{n+1}]$ не попало бы ни одной точки x_j.

Погрешность. Если скалярные произведения понимать в смысле точного интегрирования (5.62), а узлы сплайна $\{x_n\}$ образуют равномерную сетку с шагом h, то можно получить теоретические оценки погрешности наилучшего среднеквадратичного приближения. Пусть $u(x)$ имеет p непрерывных производных и аппроксимируется сплайном p-й степени. Тогда для самого сплайна и его q-х производных ($q \leqslant p-1$) справедливы следующие асимптотически точные при $h \to 0$ оценки погрешности:

а) в каждом интервале

$$\int\limits_{x_n}^{x_{n+1}} [S_p^{(q)}(x) - u^{(q)}(x)]^2 dx \approx (\beta_{p-q} h^{p-q+1})^2 \int\limits_{x_n}^{x_{n+1}} [u^{(p)}(x)]^2 dx; \tag{5.63}$$

б) на всем отрезке

$$\|S_p^{(q)} - u^{(q)}\|_{L_2} \approx \beta_{p-q} h^{p-q+1} \|u^{(p)}\|_{L_2}. \tag{5.64}$$

Коэффициенты этих формул (точнее, $1/\beta_{p-q}$) представлены в табл. 5.3. Известна асимптотика:

$$\beta_{p-q} \to \sqrt{2}/(2\pi)^{p-q+1} \quad \text{при} \quad p-q \to \infty.$$

Фактически она уже хорошо выполняется при $p-q \geqslant 1$. Таким образом, с увеличением степени сплайна не только возрастает порядок точности, но и быстро уменьшается остаточный коэффициент. Хотя сплайн p-й степени имеет такой же порядок точности, как и интерполяционный многочлен, но коэффициент в остаточном члене у него меньше; этот выигрыш тем сильнее, чем выше степень сплайна. Таким образом, среднеквадратичные сплайны являются эффективным средством аппроксимации.

Таблица 5.3

Коэффициенты в формулах (5.63), (5.64)

$p-q$	$1/\beta_{p-q}$	$p-q$	$1/\beta_{p-q}$
0	$2\sqrt{3}$	4	$144\sqrt{2\,310}$
1	$12\sqrt{5}$	5	$30\,240\sqrt{1\,430/691}$
2	$12\sqrt{210}$	6	$8\,640\sqrt{1\,001}$
3	$240\sqrt{21}$	7	$1\,209\,600\sqrt{7\,293/3\,617}$

Многомерность. В многомерном случае можно построить сплайн в виде произведения одномерных B-сплайнов. Например, двумерный B-сплайн первой степени есть произведение двух одномерных линейных сплайнов (см. рис. 5.4 для $p = 1$) по x и y. График такого сплайна имеет вид четырехгранной пирамиды с основанием 2×2 интервала.

Однако такие базисы существенно дальше от ортогональности, чем в одномерном случае. Например, носитель (основание описанной выше пирамиды) перекрывается не с двумя соседними носителями, как в одномерном случае, а с восемью. Процент заполнения матрицы Грама становится больше. Ее структура усложняется: для линейного сплайна в двумерном случае вместо трех диагоналей получается девять, причем расположенных не рядом, а образующих очень широкую ленту. Это не позволяет эффективно использовать ленточный вариант метода Гаусса.

Самое неприятное в другом: резко ухудшается обусловленность линейной системы. Уже при $p = 1$ отсутствует преобладание диагонального элемента, поэтому к прямому разложению по двумерному базису прибегать невыгодно. Лучше использовать последовательность одномерных разложений.

Для этого поступают следующим образом. Сначала разлагают функцию $u(x, y)$ по сплайнам одной переменной от x. Получают коэффициенты разложения $c_n(y)$, как функции второй переменной y. Потом проводят разложение этих коэффициентов $c_n(y)$ по одномерным сплайнам от переменной y.

Достоинства этой процедуры состоят в следующем: 1) она более устойчивая; 2) легко применяется к сплайнам произвольной степени; 3) естественно обобщается на случай большего числа переменных. Однако у нее имеется недостаток — произведение двух одномерных разложений не эквивалентно прямому двумерному разложению, поскольку базисные функции не ортогональны.

5.5.3. Конечные элементы

В литературе много внимания уделяют так называемому методу конечных элементов (его называют также проекционно-сеточным). В нем используют разложение одно-, двух- и трехмерных функций по базисным функциям, определенным на конечных носителях. Такие базисные функции называют конечными элементами.

Само разложение понимают в разных смыслах. Это может быть наилучшая аппроксимация в норме L_2, удовлетворение некоторому дифференциальному уравнению, минимизация некоторого специального функционала от данного разложения и т. п.

Видно, что B-сплайны являются частным случаем таких конечных элементов. Они имеют особенно простой вид и широко используются в данном методе. Особенно часто из-за своего очень простого вида употребляют линейные B-сплайны, несмотря на их невысокую гладкость. Конечные элементы высокой гладкости применяют заметно реже. По существу метод конечных элементов близок к разностным схемам. По историческим причинам он развивался сначала самостоятельно, но постепенно изложения этих методов все больше и больше сближаются.

5.6. АППРОКСИМАЦИЯ КРИВЫХ

5.6.1. Параметризация кривой

Отличие аппроксимации кривой от аппроксимации функции состоит в том, что кривая может оказаться замкнутой, с возвратом, иметь бесконечную производную и т. п. В этом случае невозможно аппроксимировать искомую кривую любой явной функцией, в том числе многочленом или отношением многочленов (рис. 5.7). Даже в случае однозначных быстроменяющихся функций простейшие явные аппроксимации могут оказаться непригодными. На рис. 4.5 (см. гл. 4) видно, что интерполяционный многочлен оказался немонотонным, хотя исходная зависимость была монотонной.

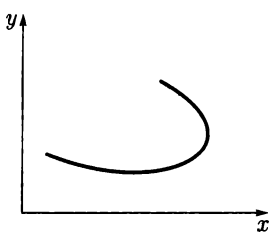


Рис. 5.7. Кривая, которую невозможно описать однозначной функцией

Рассмотрим для общности трехмерное пространство. Пусть кривая задана точками x_n, y_n, z_n, $0 \leqslant n \leqslant N$. Наиболее удоб-

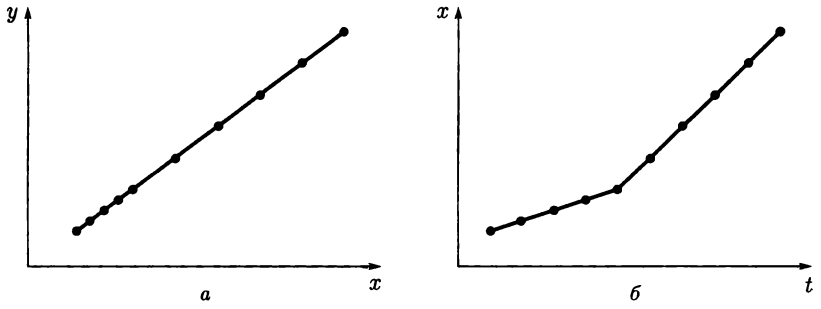


Рис. 5.8. Параметрическое представление кривой:

а — линия $y(x)$; *б* — ее неудачная параметризация $t_n = n$

но использовать параметрическое представление кривой. Введем параметр t, монотонно возрастающий вдоль кривой. Допишем к таблице его значения в узлах и разобьем ее на одномерные таблицы $(t_n, x_n), (t_n, y_n), (t_n, z_n)$. Далее строим аппроксимацию каждой таблицы $x(t), y(t), z(t)$. Изменяя t, можно вычислить любое число промежуточных точек. Поэтому идея ввести параметр и построить одномерные аппроксимации весьма привлекательна, тем более, что этот способ применим к любому числу измерений.

Разумеется, не всякий способ введения параметра пригоден. Например, нельзя брать в качестве параметра номер точки на кривой $t_n = n$. Возьмем отрезок простейшей зависимости $y(x) = x$. Пусть на первой части отрезка поставлены близко расположенные равноотстоящие точки, а на второй части — тоже равноотстоящие, но гораздо более редкие точки (рис. 5.8, *а*). Тогда графики $x(t)$ и $y(t)$ будут ломаными (рис. 5.8, *б*), хотя исходная кривая была гладкой.

Когда кривую рисуют по точкам с помощью лекала, лист бумаги можно вертеть, как нам удобно. Значит, мы интуитивно требуем, чтобы кривая не менялась при параллельном переносе или повороте осей координат. Математически это означает, что при описании кривой следует пользоваться чисто геометрическими понятиями: длиной, кривизной и т. п. Поэтому в качестве параметра целесообразно выбрать геометрическую величину.

Идеальный способ — взять за параметр длину дуги кривой. Но для этого нужно знать вид кривой. На практике берут одну из следующих характеристик кривой, которые нетрудно построить по точкам: длину хорды или длину дуги окружности (второй способ существенно точнее).

5.6.2. Хорда

Для простоты записи ограничимся случаем двух переменных и будем обозначать точки как координатами, так и радиусом-вектором $(x_n, y_n) = \mathbf{r}_n$. Простейшим приближением к длине дуги кривой $l_{n,n+1}$ между двумя соседними точками $\mathbf{r}_n$ и $\mathbf{r}_{n+1}$ будет длина хорды $\bar{l}_{n,n+1}$, соединяющей эти точки:

$$\bar{l}_{n,n+1} = |\mathbf{r}_{n+1} - \mathbf{r}_n| = \sqrt{(x_{n+1} - x_n)^2 + (y_{n+1} - y_n)^2}. \qquad (5.65)$$

Точное значение длины дуги между двумя точками равно

$$l_{n,n+1} = \int\limits_{t_n}^{t_{n+1}} \sqrt{x_t^2(t) + y_t^2(t)}dt = \int\limits_{x_n}^{x_{n+1}} \sqrt{1 + y_x^2(x)}dx. \qquad (5.66)$$

Сравним, насколько отличаются длины дуги и хорды.

Погрешность. Для нахождения погрешности введем локальную систему координат. За начало координат возьмем n-ю точку кривой, а за ось абсцисс — касательную к кривой в этой точке. Будем обозначать через $y(x)$, $y_x(x)$, $y_{xx}(x)$ и так далее текущие значения функции и ее производные вдоль кривой, а через y, y_x, y_{xx} и так далее — их значения в n-й точке. В силу выбора координат $y = 0$ и $y_x = 0$. Поэтому разложения функции и первой производной в ряд Тейлора принимают следующий вид:

$$\begin{aligned} y(x) &= \frac{1}{2}x^2 y_{xx} + \frac{1}{6}x^3 y_{xxx} + O(x^4), \\ y_x(x) &= xy_{xx} + \frac{1}{2}x^2 y_{xxx} + O(x^3). \end{aligned} \qquad (5.67)$$

Подставив разложение (5.67) в формулу (5.65), получим длину хорды для произвольного аргумента:

$$\bar{l}(x) = x + \frac{1}{8}x^3 y_{xx}^2 + \frac{1}{12}x^4 y_{xx}y_{xxx} + O(x^5). \qquad (5.68)$$

Такая же подстановка в (5.66) дает длину дуги:

$$l(x) = x + \frac{1}{6}x^3 y_{xx}^2 + \frac{1}{8}x^4 y_{xx}y_{xxx} + O(x^5). \qquad (5.69)$$

Сравнивая длины (5.68) и (5.69) между собой, найдем их отношение:

$$\bar{l}(x)/l(x) = 1 - \frac{1}{24}x^2 y_{xx}^2 + O(x^3). \qquad (5.70)$$

Это означает, что относительная погрешность длины при замене дуги хордой есть $O(x^2)$.

Видно, что приближение хордой (5.65) имеет второй порядок точности. Такая точность считается невысокой. Этот способ хорошо работает на участках малой кривизны, когда длина хорды стремится к длине дуги. Если кривизна большая (например, профиль крыла самолета в районе передней кромки), то этот способ становится не слишком аккуратным. На участках большой кривизны необходима очень подробная сетка. В противном случае аппроксимирующая кривая может иметь «выбросы» — маломасштабные формы, качественно не присущие исходной кривой.

Замечание. Второй порядок точности относился к вычислению самой длины. Однако параметризация кривой выполняется в первую очередь ради дальнейшей аппроксимации в параметрической форме. Заметим, что такая аппроксимация не изменится, если параметр умножить на константу. Это открывает возможность увеличения порядка точности последующей аппроксимации на 1.

Пусть сетки $\{t_n\}$ и, соответственно, $\{x_n\}$ квазиравномерны. Роль длины интервала в (5.70) играет величина x. Для квазиравномерных сеток отношения длин соседних интервалов равны $1 + O(x)$. Величины y_{xx} в соседних интервалах также равны с точностью до $O(x)$. Поэтому отношения $\bar{l}/l$ в соседних интервалах будут совпадать с точностью $O(x^3)$. Это означает, что на квазиравномерной сетке параметризацию длинами хорд можно использовать для построения аппроксимаций кривых, имеющих третий порядок точности.

5.6.3. Окружность

Более высокую точность по сравнению с хордой дает проведение окружности через две данные точки и третью, соседнюю справа или слева. Через эти три точки (считая, что они не лежат на одной прямой) можно провести единственную плоскость. Будем рассматривать задачу в этой плоскости, т. е. по-прежнему в двумерной постановке. Приведем вывод формул.

Возьмем три соседние точки с радиусами-векторами $\mathbf{r}_{n-1}$, $\mathbf{r}_n$, $\mathbf{r}_{n+1}$. Введем соединяющие их векторы: $\mathbf{r}_+ = \mathbf{r}_{n+1} - \mathbf{r}_n$, $\mathbf{r}_- = \mathbf{r}_n - \mathbf{r}_{n-1}$ и $\mathbf{r} = \mathbf{r}_{n+1} - \mathbf{r}_{n-1} = \mathbf{r}_- + \mathbf{r}_+$. Построим через эти точки окружность, центр которой обозначим через O, а радиус через R (рис. 5.9). Центральный угол, опирающийся на хорду $\mathbf{r}_+$, обозначим через $2\alpha_+$, а на хорду $\mathbf{r}_-$ — через $2\alpha_-$. Угол между векторами $\mathbf{r}$ и $\mathbf{r}_-$ опирается на хорду $\mathbf{r}_+$, а его вершина лежит на окружности. Поэтому он вдвое меньше соответствую-

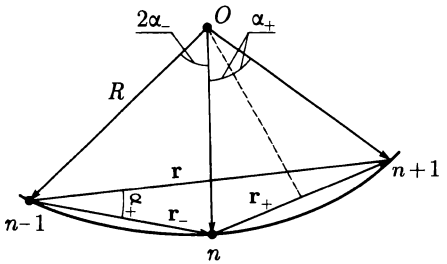


Рис. 5.9. Аппроксимация дугой окружности

щего центрального угла, опирающегося на ту же хорду, т. е. равен α_+.

Очевидно, длины хорд — это длины соответствующих векторов: $\bar{l}_- \equiv \bar{l}_{n-1,n} = |\mathbf{r}_-| = |\mathbf{r}_n - \mathbf{r}_{n-1}|$ и $\bar{l}_+ \equiv \bar{l}_{n,n+1} = |\mathbf{r}_+| = |\mathbf{r}_{n+1} - \mathbf{r}_n|$. Сами эти векторы и их длины легко вычисляются по координатам точек. Угол α_+ определяем из векторного произведения векторов $\mathbf{r}$ и $\mathbf{r}_-$: $[\mathbf{r}_-, \mathbf{r}] = |\mathbf{r}_-||\mathbf{r}| \sin\alpha_+$. Подставляя сюда $\mathbf{r} = \mathbf{r}_- + \mathbf{r}_+$ и учитывая, что векторное произведение $[\mathbf{r}_-, \mathbf{r}_-] = 0$, получим для искомого угла следующее выражение:

$$\sin\alpha_+ = [\mathbf{r}_-, \mathbf{r}_+]/(|\mathbf{r}_-||\mathbf{r}|). \tag{5.71}$$

Эта формула удобна тем, что при вычислениях по ней не происходит потери точности даже в случае подробной сетки, когда сами интервалы малы, а векторы хорд почти параллельны.

Из рис. 5.9 видно, что $|\mathbf{r}_+| = 2R\sin\alpha_+$. Отсюда радиус окружности равен

$$R = |\mathbf{r}_+|/2\sin\alpha_+. \tag{5.72}$$

Обозначим длины дуг окружностей через $\tilde{l}_- \equiv \tilde{l}_{n-1,n}$ и $\tilde{l}_+ \equiv \tilde{l}_{n,n+1}$. Очевидно,

$$\tilde{l}_+ = 2R\alpha_+ = |\mathbf{r}_+|\alpha_+/\sin\alpha_+. \tag{5.73}$$

По формулам (5.71) — (5.73) последовательно определяем α_+, R и $\tilde{l}_+$, тем самым находя длину дуги окружности, а попутно — кривизну $1/R$. Заменив в формулах (5.71) — (5.73) индекс «+» на «−», получим длину второго участка окружности $\tilde{l}_-$.

Погрешность. Сравним длину дуги окружности $\tilde{l}_+$ с истинной длиной дуги кривой l_+. Выберем ту же локальную систему координат, что и при исследовании погрешности длины хорды в подразд. 5.6.2. Приращения аргументов для $(n+1)$-й и $(n-1)$-й точек обозначим через x_+ и x_- соответственно (очевидно, знаки x_+ и x_- будут противоположными). Тогда векторные величины,

входящие в (5.71) — (5.73), выражают через координаты следующим образом:

$$[\mathbf{r}_-, \mathbf{r}_+] = xy(x_+) - x_+ y(x_-), \quad |\mathbf{r}_-| = \sqrt{x_-^2 + y^2(x_-)},$$
$$|\mathbf{r}_+| = \sqrt{x_+^2 + y^2(x_+)}, \tag{5.74}$$
$$|\mathbf{r}| = \sqrt{(x_+ - x_-)^2 + [y(x_+) - y(x_-)]^2}.$$

Эти формулы удобны для практических расчетов длины дуги. Разложим все функции $y(x_\pm)$ в ряды Тейлора согласно (5.67). Подставив эти разложения в (5.71) — (5.74), получим приближенные выражения для угла и дуги окружности

$$\sin\alpha_+ = \frac{1}{2}x_+ \left[y_{xx} + \frac{1}{3}(x_+ + x_-)y_{xxx} + O(x^2) \right],$$
$$\tilde{L}_+ = x_+ + \frac{1}{6}(x_+)^3 y_{xx}^2 + \frac{1}{36}x_+^3(4x_+ + x_-)y_{xx}y_{xxx} + O(x^5). \tag{5.75}$$

Для отношения длин дуг окружности (5.75) и точной кривой получаем следующее выражение:

$$\tilde{l}_+ / l_+ = 1 - \frac{1}{72}x_+^2(x_+ - 2x_-)y_{xx}y_{xxx} + O(x^4). \tag{5.76}$$

Это отношение отличается от 1 на величину $\sim x^3$ (в оценку этого члена входят оба приращения x_+ и x_-). Следовательно, дуга окружности аппроксимирует истинную дугу с третьим порядком точности.

Тем самым порядок точности (5.74) на единицу выше, чем для длины хорды. Это уже неплохая точность. Поскольку длина дуги окружности вычисляется просто, этот способ следует рекомендовать для практики как первоочередной.

Длину дуги $\tilde{l}_+ \equiv \tilde{l}_{n,n+1}$ можно вычислить этим способом дважды: один раз с использованием левого соседа — $(n-1)$-й точки, и второй раз — по правой соседней $(n+2)$-й точке. Обозначим эти два приближения к длине искомой дуги через $\tilde{l}'_+$ и $\tilde{l}''_+$ соответственно. Существует алгоритм, вычисляющий по этим приближениям и другой информации о точках более хороший результат относительной точности $O(l^4)$. Однако этот алгоритм гораздо более сложен.

Контроль. Построим несложный практический способ контроля достигаемой точности. Для этого используем длину хорды $\bar{l}_+ \equiv |\mathbf{r}_+|$ и оба приближения к дуге окружности $\tilde{l}'_+$ и $\tilde{l}''_+$.

Поскольку длина окружности есть гораздо лучшее приближение к истинной длине дуги, чем хорда, то асимптотически точной оценкой локальной погрешности хорды является величина $\bar{l}_+ - \tilde{l}_+$. Это не относительная погрешность, а абсолютная. В подразд. 5.6.2 показано, что относительная погрешность имеет второй порядок малости; следовательно, абсолютная погрешность имеет третий порядок малости:

$$\bar{l}_+ - \tilde{l}_+ = O(l_+^3) = O(\bar{l}_+^3).$$

Два определения дуги окружности $\tilde{l}'_+$ и $\tilde{l}''_+$ отличаются на величину более высокого порядка малости:

$$\tilde{l}'_+ - \tilde{l}''_+ = O(l_+^4) = O(\bar{l}_+^4).$$

Комбинируя эти оценки, получим следующий критерий:

$$|\tilde{l}''_+ - \tilde{l}'_+| \ll \tilde{l}_+ - \bar{l}_+ \ll (\bar{l}_+)^2.$$

При выполнении этого условия можно уверенно пользоваться любым из двух приближений к длине дуги окружности. В этом случае в качестве окончательного значения длины дуги и ее погрешности можно принять

$$\tilde{l}_+ = \frac{1}{2}(\tilde{l}'_+ + \tilde{l}''_+) \pm \frac{1}{2}|\tilde{l}'_+ - \tilde{l}''_+|.$$

Отметим следующее: 1) критерий можно записать только во внутренних интервалах, так как граничный интервал имеет лишь одного соседа; 2) левое неравенство критерия может не выполняться вблизи точки перегиба кривой, так как там радиус кривизны $R \to 0$ и хорда почти неотличима от окружности. Но в последнем случае и хорда, и окружность очень хорошо приближают истинную кривую, так что найденные длины и там обеспечивают хорошую точность.

Замечание. В замечании к подразд. 5.6.2 отмечалась улучшенная аппроксимация гладкой кривой, если сетка квазиравномерна. Это относится и к параметризации дугами окружности. На квазиравномерной сетке отношения (5.76) в соседних интервалах отличаются между собой на величину $O(x^4)$. В этой ситуации с помощью параметризации длиной окружности можно строить аппроксимации исходной кривой, имеющие четвертый порядок точности.

Многомерность. Вывод длин дуги окружности и хорды формально проводился для точек, лежащих в одной плоскости. Но для нахождения этих длин на самом деле строить эту плоскость нет необходимости. Все формулы написаны в векторном

виде, и в них прямо можно подставлять многомерные векторы. Поэтому описанный подход дает простые и удобные способы параметризации многомерной кривой. При этом найденные параметры являются геометрическими характеристиками кривой и не меняются при поворотах и трансляциях системы координат.

5.6.4. Аппроксимация

Вернемся к рассмотрению табулированной кривой (x_n,y_n,z_n), $0 \leqslant n \leqslant N$. Параметр t естественно ввести с помощью длин дуг окружностей: $t_0 = 0$, $t_{n+1} - t_n = \tilde{l}_{n,n+1}$ при $n = 0, 1, \ldots, N-1$. Этот параметр является монотонно возрастающим вдоль кривой. Исходная таблица разбивается на независимые таблицы $(t_n,x_n), (t_n,y_n), (t_n,z_n)$. Остается построить аппроксимацию каждой из этих таблиц. Для этого можно использовать различные способы интерполяции и среднеквадратичной аппроксимации, описанные в гл. 4 и этой главе.

Мы предполагаем, что исходная кривая достаточно гладкая. Напомним некоторые данные ранее рекомендации. Если число точек N мало, то лучше использовать интерполяцию, если велико — среднеквадратичную аппроксимацию. В случае замкнутой кривой $x(t), y(t), z(t)$ будут периодическими функциями и следует использовать разложение в тригонометрический ряд Фурье или сплайны с периодическими граничными условиями. Если функции непериодические, то лучше применять метод двойного периода, ряды по многочленам Чебышева или сплайны с естественными граничными условиями.

Опишем формальную процедуру аппроксимации $x(t)$ некоторым методом. Например, для определенности выберем аппроксимацию отрезком ряда Фурье

$$x(t) \approx \sum_{m=0}^{2M} c_m \varphi_m(t) \tag{5.77}$$

(см. 5.2). Коэффициенты Фурье c_m вычислим по квадратурной формуле трапеций на сетке $\{t_n\}$:

$$\begin{gathered} c_m = \frac{1}{\|\varphi_m\|^2} \sum_{n=1}^{N} \frac{t_n - t_{n-1}}{2} [x_n \varphi_m(t_n) + x_{n-1} \varphi_m(t_{n-1})], \\ \|\varphi_m\|^2 = \sum_{n=1}^{N} \frac{t_n - t_{n-1}}{2} [\varphi_m^2(t_n) + \varphi_m^2(t_{n-1})]. \end{gathered} \tag{5.78}$$

Из (5.78) видно, что c_m линейно зависят от узловых значений искомой функции $\{x_n\}$. Из (5.77) следует, что $x(t)$ линейно зависит от c_m и, следовательно, от величин $\{x_n\}$. Эту зависимость, вводя вектор $\mathbf{t} = \{t_0, t_1, \ldots t_N\}$, можно символически записать в следующем виде:

$$x(t) = \sum_{n=0}^{N} a_n(t; \mathbf{t}) x_n. \tag{5.79}$$

Формулы для коэффициентов $a_n(t; \mathbf{t})$ довольно громоздки и здесь не приводятся. Зависимость от t вошла в (5.77) из зависимости $\varphi_m(t)$, а зависимость от $\mathbf{t}$ — из квадратурной формулы (5.78). Важно отметить, что вид коэффициентов $a_n(t; \mathbf{t})$ не зависит от величин x_n. Он зависит только от выбранного способа аппроксимации (базиса $\varphi_m(t)$ и числа членов в этом базисе) и заданной сетки $\{t_n\}$.

Если для зависимостей $y(t)$ и $z(t)$ выбран тот же способ аппроксимации, причем с тем же числом членов, то окончательные формулы будут для них точно такими же и с теми же коэффициентами $a_n(t; \mathbf{t})$:

$$y(t) = \sum_{n=0}^{N} a_n(t; \mathbf{t}) y_n, \quad z(t) = \sum_{n=0}^{N} a_n(t; \mathbf{t}) z_n. \tag{5.80}$$

Если же для них выбрать либо другой вид аппроксимации, либо тот же вид аппроксимации, но с другим числом членов, то вместо $a_n(t; \mathbf{t})$ в формулах (5.80) будут стоять другие коэффициенты.

Формально мы имеем право для каждой зависимости $x(t)$, $y(t)$, $z(t)$ выбирать свой способ аппроксимации. Однако этого делать не целесообразно. Важно соблюдать следующее правило. Для всех трех зависимостей $x(t), y(t), z(t)$ следует: ***1) выбирать один и тот же вид аппроксимации; 2) использовать одинаковое число членов выбранного вида аппроксимации.***

Несоблюдение этого правила приведет к тому, что построенная кривая не будет удовлетворять требованию ротационной и трансляционной инвариантности.

Приведенные рассуждения справедливы для аппроксимации (5.77), линейной по коэффициентам. Правило не применимо к рациональной интерполяции (см. подразд. 4.3.2) и методу выравнивания (см. подразд. 4.3.1), поскольку в них зависимость от

коэффициентов нелинейная. Однако для всех остальных способов аппроксимации, рассмотренных в гл. 4 и данной главе, рассуждения остаются в силе.

5.6.5. Ротационная инвариантность

Будем считать, что аппроксимации всех координат построены единообразно согласно сформулированному в подразд. 5.5.4 правилу. Введем векторы-строки $\mathbf{x}, \mathbf{y}, \mathbf{z}$ табулированных значений координат и вектор-столбец $\mathbf{a}(t;\mathbf{t})$ коэффициентов аппроксимации размерности $N+1$:

$$\begin{matrix}\mathbf{x} = (x_0, x_1, \ldots, x_N),\\ \mathbf{y} = (y_0, y_1, \ldots, y_N),\\ \mathbf{z} = (z_0, z_1, \ldots, z_N),\end{matrix} \qquad \mathbf{a}(t;\mathbf{t}) = \begin{pmatrix} a_0(t;\mathbf{t}) \\ a_1(t;\mathbf{t}) \\ \ldots \\ a_N(t;\mathbf{t}) \end{pmatrix}.$$

Тогда аппроксимации можно записать в виде скалярных произведений

$$x(t) = [\mathbf{x}, \mathbf{a}(t;\mathbf{t})] = \sum_{n=0}^{N} x_n a(t;\mathbf{t}),$$

$$y(t) = [\mathbf{y}, \mathbf{a}(t;\mathbf{t})] = \sum_{n=0}^{N} y_n a_n(t,\mathbf{t}),$$

$$z(t) = [\mathbf{z}, \mathbf{a}(t;\mathbf{t})] = \sum_{n=0}^{N} z_n a_n(t,\mathbf{t}).$$

Из набора координат $x(t), y(t), z(t)$ составим трехмерный вектор-столбец:

$$\mathbf{r}(t) = \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix} = \begin{pmatrix} (\mathbf{x}, \mathbf{a}(t;\mathbf{t})) \\ (\mathbf{y}, \mathbf{a}(t;\mathbf{t})) \\ (\mathbf{z}, \mathbf{a}(t;\mathbf{t})) \end{pmatrix}. \tag{5.81}$$

Введем прямоугольную матрицу начальных данных R, имеющую размерность $3 \times (N+1)$:

$$R = \begin{pmatrix} x_0, x_1, \ldots, x_N \\ y_0, y_1, \ldots, y_N \\ z_0, z_1, \ldots, z_N \end{pmatrix}.$$

Вспоминая определение скалярных произведений, перепишем (5.81) в виде

$$\mathbf{r}(t) = R\mathbf{a}(t;\mathbf{t}); \tag{5.82}$$

умножение в правой части (5.82) производится по правилам перемножения прямоугольных матриц. Формула (5.82) дает нам искомую трехмерную кривую через матрицу табулированных координат и столбец коэффициентов аппроксимации.

Поворот листа бумаги, на котором поставлены исходные точки (или поворот трехмерного пространства), означает, что исходная ортогональная система координат преобразована в другую ортогональную систему координат с помощью унитарной (ортогональной) матрицы U порядка 3×3. Сами координаты и матрица табулированных координат при этом преобразуются следующим образом:

$$\mathbf{r}'(t) = U\mathbf{r}(t), \quad R' = UR.$$

Чтобы получить аппроксимацию в новых координатах, надо умножить матрицу преобразованных табулированных координат R' на тот же столбец коэффициентов аппроксимации:

$$\mathbf{r}'(t) = R'\mathbf{a}(t;\mathbf{t}) = UR\mathbf{a}(t;\mathbf{t}). \tag{5.83}$$

В силу ассоциативности умножения матриц в последнем произведении трех матриц сначала можно перемножить вторую и третью матрицы. Это дает

$$\mathbf{r}'(t) = U(R\mathbf{a}(t;\mathbf{t})) = U\mathbf{r}(t). \tag{5.84}$$

Последнее произведение в (5.84) означает применение матрицы поворота к той кривой, которая была построена в исходных координатах. Поэтому сравнение (5.83) и (5.84) означает следующее: если мы сначала построим аппроксимирующую кривую и потом повернем лист бумаги, то получим точно тот же результат, как если бы сначала повернуть координаты, а потом тем же способом построить в них аппроксимирующую кривую. Это свойство называется ротационной инвариантностью.

Мы доказали, что если для параметризации кривой выбрать ротационно-инвариантный параметр (такой, как длина дуги окружности или хорды) и применить для каждой координаты один и тот же линейный способ аппроксимации с одинаковым числом параметров, то полученная кривая будет ротационно-инвариантной. Нетрудно видеть, что такая кривая будет также

Таблица 5.4

Опорные точки для примера 5.3; x, y — исходная кривая; x^*, y^* — поворот координат на 45°

n	x_n	y_n	x_n^*	y_n^*
0	0	0	0	0
1	ξ	0,5	$(\xi - 1/2)/\sqrt{2}$	$(\xi + 1/2)/\sqrt{2}$
2	1	1	0	$\sqrt{2}$

трансляционно-инвариантной, т. е. не будет меняться при сдвигах начала координат.

Несмотря на то что выкладки были проведены для случая трех пространственных измерений, все формулы и выводы справедливы для пространства произвольной размерности.

Пример 5.3. Зададим кривую $y(x)$ на плоскости тремя опорными точками (табл. 5.4). Положение средней точки зависит от величины $\xi \in (-\infty, \infty)$. Для аппроксимации кривой выберем в качестве параметра длины хорд; это дает

$$t_0 = 0, \quad t_1 - t_0 = \sqrt{\xi^2 + 0{,}25}, \quad t_2 - t_1 = \sqrt{(1-\xi)^2 + 0{,}25}.$$

В подразд. 5.6.2 отмечалось, что такая параметризация позволяет строить аппроксимации третьего порядка точности. Поэтому для аппроксимации зависимостей $x(t)$, $y(t)$ возьмем интерполяционный многочлен Ньютона второй степени, обеспечивающий третий порядок точности, с учетом $t_0 = x_0 = y_0 = 0$ он принимает следующий вид:

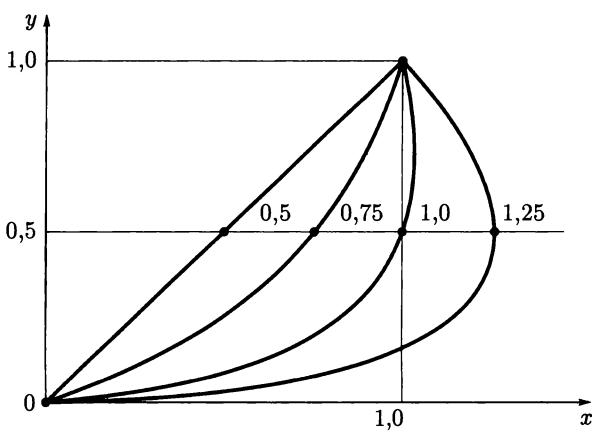


Рис. 5.10. Параметрическая аппроксимация кривой по трем точкам (см. пример 5.3)

$$x(t) = t\frac{x_1}{t_1} + \frac{t(t-t_1)}{t_2}\left(\frac{x_2 - x_1}{t_2 - t_1} - \frac{x_1}{t_1}\right)$$

и аналогично для $y(t)$. Соответствующие аппроксимационные кривые для серии значений $\xi = 0{,}5; 0{,}75; 1{,}0; 1{,}25$ приведены на рис. 5.10.

Видно, что все эти кривые имеют качественно разумный вид. Если же аппроксимировать параболическим многочленом непосредственную зависимость $y(x)$, соблюдая указанный в табл. 5.4 порядок точек, то для 3-й и 4-й кривых это сделать невозможно.

В данном примере выполнены условия ротационной инвариантности. Можно повернуть исходные координаты на некоторый угол; для поворота на 45°новые координаты точек также приведены в табл. 5.4. Если построить по этим точкам кривые тем же способом, то при обратном повороте на 45°они точно совпадут с кривыми, приведенными на рис. 5.10.

Глава 6

ЧИСЛЕННОЕ ДИФФЕРЕНЦИРОВАНИЕ

6.1. ПРОИЗВОДНАЯ МНОГОЧЛЕНА НЬЮТОНА

6.1.1. Общие формулы

Функцию приходится дифференцировать численно, если ее аналитическое выражение сложно или она задана таблицей своих значений. Когда способ вычисления функции задан, но очень сложен, также предварительно вычисляют некоторую таблицу ее значений. Таблицу аппроксимируют достаточно простой функцией (см. гл. 4 и 5) и дифференцируют уже эту функцию. При этом выделяют две группы способов: дифференцирование интерполяции и дифференцирование среднеквадратичной аппроксимации.

Пусть мы располагаем таблицей значений функции: (x_n, u_n) $0 \leqslant n \leqslant N$, число узлов сетки N невелико. В этом случае разумно прибегнуть к интерполяции. Проще всего использовать интерполяционный многочлен Ньютона:

$$\begin{gathered} u(x) \approx u(x_0) + \sum_{n=1}^{N} u(x_0, x_1, \ldots, x_n)\omega_n(x), \\ \omega_n(x) = \prod_{k=0}^{n-1} (x - x_k). \end{gathered} \tag{6.1}$$

Здесь $u(x_0, x_1, \ldots, x_n)$ — разделенные разности. Почленно продифференцируем q раз многочлен (6.1) и учтем, что производные многочленов $\omega_n^{(q)}(x)$ обращаются в нуль при $q > n$. Тогда получим

$$u^{(q)}(x) \approx \sum_{n=q}^{N} u(x_0, x_1, \ldots, x_n)\omega_n^{(q)}(x). \tag{6.2}$$

Очевидно, $\omega_q^{(q)} = q!$ Дифференцирование следующих членов суммы дает более громоздкие выражения. Для первой и второй производной почленное дифференцирование дает

$$\omega_n'(x) = \sum_{l=0}^{n-1} \prod_{0 \leqslant k \leqslant n-1, k \neq l} (x - x_k) \equiv \sum_{l=0}^{n-1} \frac{\omega_n(x)}{x - x_l}; \tag{6.3}$$

$$\omega_n''(x) = 2 \sum_{0 \leqslant l < m \leqslant n-1} \prod_{0 \leqslant k \leqslant n-1, k \neq l, m} (x - x_k). \tag{6.4}$$

Формулы для старших производных пишутся аналогично, но их вид более сложен, поэтому здесь их не приводим.

Пусть шаги сетки малы, не сильно различаются по величине и близки к $h \ll 1$. Тогда n-е слагаемые ($n \leqslant q$) в (6.2) убывают как $O(h^{n-q})$. Первое слагаемое есть $O(1)$, второе — $O(h)$ и т. д. Последнее слагаемое есть $O(h^{N-q})$. Чтобы получить выражение для q-й производной, необходимо $N \geqslant q$, т. е. надо воспользоваться не менее чем $q + 1$ точкой.

Если слагаемые в (6.2) убывают достаточно быстро, то по аналогии с самим многочленом Ньютона можно рассматривать первое из неучтенных слагаемых как асимптотически точную оценку погрешности. Для функций, имеющих достаточное число непрерывных ограниченных производных, это можно строго показать аналогично тому, как это было сделано в гл. 4. Однако узлы многочленов $\omega_n(x)$ расположены на малых расстояниях порядка h. Поэтому q-е производные этих многочленов будут быстро расти с увеличением q. Следовательно, убывание слагаемых в (6.2) будет тем медленнее, чем больше q. Это является следствием общего правила об ухудшении сходимости при почленном дифференцировании. Соответственно, если ограничиться в (6.2) только первым членом, то точность будет $O(h)$, если вторым — $O(h^2)$ и т. д. В частности, ограничиваясь в (6.2) одним слагаемым и учитывая, что $\omega_q^{(q)} = q!$, получим

$$u^{(q)}(x) = q! u(x_0, \ldots, x_q) + O(h). \tag{6.5}$$

Формула (6.5) дает приблизительную связь q-й разделенной разности с q-й производной.

Однако напомним, что интерполяционный многочлен Ньютона высокой степени может не только плохо сходиться, но и расходиться даже для функций с неограниченным числом непрерывных производных.

На практике безопасным считается использование $N \leqslant 6 \div 8$. Чтобы сумма (6.2) имела хотя бы два-три слагаемых для обеспечения удовлетворительного порядка точности, приходится ограничиваться вычислением только младших производных порядков $q = 1 \div 2$ и лишь изредка $q = 3 \div 4$.

Пример 6.1. В подразд. 4.1.2 строились интерполяционные многочлены для вычисления синуса в первой четверти. Поскольку для дифференцирования нужно больше узлов, воспользуемся формулой (4.10) с $N = 4$:

$$\sin(22{,}5^\circ x) \approx 0 + 0{,}38268(x-0) - 0{,}02913(x-0)(x-1) - \\ -0{,}00823(x-0)(x-1)(x-2) + \\ + 0{,}00068(x-0)(x-1)(x-2)(x-3), \quad x \in [0,4].$$

Дифференцируя последнее выражение, получаем

$$\frac{d\sin(22{,}5^\circ x)}{dx} \approx 0{,}38268 - 0{,}02913(2x-1) - \\ -0{,}00823(3x^2 - 6x + 2) + 0{,}00068(4x^3 - 18x^2 + 22x - 6),$$

$$\frac{d^2\sin(22{,}5^\circ x)}{dx^2} \approx -0{,}02913 \cdot 2 - 0{,}00823(6x-6) + \\ + 0{,}00068(12x^2 - 36x + 22).$$

В скобках стоят производные многочленов $\omega_n^{(q)}(x)$; видно их возрастание. Точность этих формул невысока. Очевидно, следующую производную искать по четырем точкам уже нецелесообразно.

6.1.2. Простейшие случаи

Запишем формулы для первой и второй производных на произвольной неравномерной сетке. Выпишем интерполяционный многочлен Ньютона, ограничиваясь небольшим числом членов

$$u(x) \approx u(x_0) + (x - x_0)u(x_0, x_1) + \\ + (x - x_0)(x - x_1)u(x_0, x_1, x_2) + O(h^3), \tag{6.6}$$

где низшие разделенные разности равны

$$u(x_0, x_1) = (u_0 - u_1)/(x_0 - x_1), \\ u(x_0, x_1, x_2) = [(u_0 - u_1)/(x_0 - x_1) - \\ - (u_1 - u_2)/(x_1 - x_2)]/(x_0 - x_2). \tag{6.7}$$

Дифференцируя (6.6), получим

$$u'(x) \approx u(x_0, x_1) + (2x - x_0 - x_1)u(x_0, x_1, x_2) + O(h^2), \quad (6.8)$$

$$u''(x) \approx 2u(x_0, x_1, x_2) + O(h). \quad (6.9)$$

Точность этих формул невысока, однако их достаточно часто употребляют при расчетах на неравномерных сетках.

Квазиравномерные сетки. При написании разностных схем приходится вычислять производную не в произвольной точке x, а в узле или полуцелой точке сетки. На произвольной неравномерной сетке формулы (6.8) и (6.9) сохраняют при этом указанную в них точность. Однако если сетка квазиравномерная (что предполагается далее), а производную находят в центре симметрии выбранной конфигурации узлов, то точность повышается. Обычно при этом нумеруют точки не от нулевой. Для построения второй производной выбирают узлы $n-1, n, n+1$ и записывают

$$u''(x_n) = \frac{2}{x_{n+1} - x_{n-1}} \left(\frac{u_{n+1} - u_n}{x_{n+1} - x_n} - \frac{u_n - u_{n-1}}{x_n - x_{n-1}} \right) + O(h^2). \quad (6.10)$$

Для первой производной ограничиваются первой разделенной разностью; используют два варианта записи:

$$u'(x_{n+1/2}) = \frac{u_{n+1} - u_n}{x_{n+1} - x_n} + O(h^2) \quad (6.11)$$

и

$$u'(x_n) = \frac{u_{n+1} - u_{n-1}}{x_{n+1} - x_{n-1}} + O(h^2). \quad (6.12)$$

В справедливости этих формул можно убедиться одним из двух способов: а) непосредственным разложением в ряд Тейлора; б) взять первый отброшенный член общих формул и убедиться в его соответствующей малости.

Например, для получения (6.11) было отброшено второе слагаемое в (6.8). В нем стоит множитель $(2x - x_0 - x_1) \to$ $\to (2x_{n+1/2} - x_{n+1} - x_n)$. В подразд. 3.2.3 показано, что на квазиравномерных сетках этот множитель составляет $O(h^2)$.

Употребительна также запись этих формул через шаги сетки $h_n = x_n - x_{n-1}$. Тогда первая производная в полуцелой точке запишется как

$$u'_{n-1/2} = (u_n - u_{n-1})/h_n + O(h^2),$$

причем второй порядок точности обеспечивается не только на квазиравномерной сетке, но и на произвольной неравномерной. Вторая производная в целом узле равна

$$u''_n = 2(u'_{n+1/2} - u'_{n-1/2})/(h_{n+1} + h_n) + O(h^2);$$

такая точность обеспечена лишь на квазиравномерной сетке, а на произвольной неравномерной сетке точность снижается до $O(h)$. Для первой производной в узле полезно также выражение

$$u'_n = (h_{n+1}u'_{n-1/2} + h_n u'_{n+1/2})/(h_{n+1} + h_n) + O(h^2),$$

причем такая точность сохраняется на произвольной неравномерной сетке. Иногда удобны экстраполяционные формулы

$$u'_{n+1} = u'_{n+1/2} + \frac{1}{2}h_{n+1}u''_n + O(h^2),$$

$$u'_{n-1} = u'_{n-1/2} - \frac{1}{2}h_n u''_n + O(h^2),$$

также сохраняющие второй порядок точности на произвольной неравномерной сетке.

Равномерные сетки. На равномерных сетках производные в узлах и полуцелых точках сетки записываются еще проще:

$$u''(x_n) = \frac{u_{n+1} - 2u_n + u_{n-1}}{h^2} + O(h^2), \quad h = \text{const}; \tag{6.13}$$

$$\begin{gathered} u'(x_{n+1/2}) = \frac{u_{n+1} - u_n}{h} + O(h^2); \\ u'(x_n) = \frac{u_{n+1} - u_{n-1}}{2h} + O(h^2). \end{gathered} \tag{6.14}$$

Эти формулы часто используют в вычислительной практике. Надо помнить, что применять их к неравномерным сеткам нельзя: уже на квазиравномерных сетках их точность ухудшается до $O(h)$, а на произвольной неравномерной сетке погрешность формулы (6.13) составляет $O(1)$!

6.1.3. Неограниченная область

Существует немало задач для дифференциальных уравнений в неограниченной области. В них обычно присутствуют первые и вторые производные. В подразд. 3.2.3 были построены квазиравномерные сетки с конечным числом узлов, покрывающие

неограниченную область. Можно использовать такие сетки для написания производных.

Напомним, что такие сетки строятся с помощью преобразования $x = x(\xi)$, имеющего полюс (полюса) на концах отрезка по ξ. Например, для полупрямой пригодно следующее преобразование:

$$0 \leqslant x < \infty, \quad x(\xi) = c\xi/(1-\xi^2)^m, \quad c > 0, \quad m > 0, \quad 0 \leqslant \xi \leqslant 1.$$

Равномерной сетке $\xi_n = n/N$ соответствует квазиравномерная сетка $x_n = x(\xi_n)$. Ее последний узел есть бесконечно удаленная точка $x_N = +\infty$.

Стандартные формулы для вычисления производных (6.10) — (6.12) в последнем интервале формально применять нельзя. Например, по ним получается $u'_{N-1/2} = (u_N - u_{N-1})/(x_N - x_{N-1}) = 0$ при любых значениях функции, поскольку $x_N = \infty$. То же происходит со второй производной: $u''_{N-1} = 0$.

Для преодоления этой трудности вводят иное определение шага сетки. Вместо $h_n = x_n - x_{n-1}$ и $h_n + h_{n+1} = x_{n+1} - x_{n-1}$ полагают

$$h_n = \frac{1}{N}\frac{dx(\xi_{n-1/2})}{d\xi} \equiv \frac{x'_{n-1/2}}{N}, \quad h_n + h_{n+1} = \frac{2}{N}x'_n. \tag{6.15}$$

Такое переопределение нужно сделать во всех интервалах квазиравномерной сетки (а не только вблизи бесконечно удаленной точки). Тогда переопределенные значения первой и второй разностных производных будут

$$u'_{n-1/2} \approx \frac{N}{x'_{n-1/2}}(u_n - u_{n-1}), \quad u'_n \approx \frac{N}{2x'_n}(u_{n+1} - u_{n-1}); \tag{6.16}$$

$$u''_n \approx \frac{N}{x'_n}(u'_{n+1/2} - u'_{n-1/2}). \tag{6.17}$$

Формулы (6.16), (6.17) имеют точность $O(N^{-2})$. Они позволяют найти производные во всех внутренних целых и полуцелых узлах.

Нередко требуется первая производная на границе u'_N. Получим для нее формулу точности $O(N^{-2})$. Для этого возьмем двучленную формулу (6.8), заменим в ней индексы 0, 1, 2 на $N-2$, $N-1$, N, положим $x = x_N$ и переопределим шаги согласно (6.15). Используя обозначения (6.16), получим

$$u'_N \approx u'_{N-1/2} + x'_{N-1/2}u''_{N-1}/(2N). \tag{6.18}$$

На левой границе можно написать аналогичную формулу для u'_0.

Отметим еще один способ вывода формул для производной в граничной точке (для определенности, в точке x_0). Подставим в формулы (6.3), (6.4) значение $x = x_0$. Тогда в произведениях обращаются в нуль все сомножители с $k, m = 0$. Это дает

$$\omega'_n(x_0) = \prod_{k=1}^{n-1}(x_0 - x_k); \quad \omega''_n(x_0) = 2\sum_{l=1}^{n-1}\prod_{1\leqslant k\leqslant n-1; k\neq l}(x_0 - x_k).$$

Для старших производных $\omega_n^{(q)}(x_0)$ выражения существенно сложнее. Однако подобным образом можно использовать для вычисления производной большее количество узлов.

6.1.4. Сгущение сеток

Точность простейших формул (6.10) — (6.14) сравнительно невысока, а построение формул более высокого порядка точности с помощью учета следующих членов интерполяционного многочлена Ньютона оказывается очень громоздким. Но на равномерных и квазиравномерных сетках можно получать более высокий порядок точности методом сгущения сеток.

Например, напишем формулу для первой производной, используя две разные выборки точек, симметричные относительно узла x_n:

$$\bar{u}'_n \approx \frac{u_{n+1} - u_{n-1}}{x_{n+1} - x_{n-1}}; \quad \tilde{u}'_n \approx \frac{u_{n+2} - u_{n-2}}{x_{n+2} - x_{n-2}}.$$

Обе они имеют порядок точности $p = 2$. На квазиравномерной сетке первая из них соответствует расчету с числом узлов N, а вторую можно рассматривать как расчет на сетке того же семейства, но с числом узлов $N/2$. Значит, они являются приближениями первой производной, вычисленными по одной и той же формуле, но на двух разных сетках с коэффициентом сгущения $r = 2$. Поэтому можно провести уточнение по формуле Ричардсона, полагая в ней $r = 2$ и $p = 2$. Это дает

$$u'_n \approx \frac{4}{3}\bar{u}'_n - \frac{1}{3}\tilde{u}'_n = \frac{4}{3}\frac{u_{n+1} - u_{n-1}}{x_{n+1} - x_{n-1}} - \frac{1}{3}\frac{u_{n+2} - u_{n-2}}{x_{n+2} - x_{n-2}}. \qquad (6.19)$$

Можно показать, что порядок точности повышается на 2. Поэтому формула (6.19) имеет точность $O(h^4)$. На равномерной сетке она переходит в следующее выражение:

$$u'_n = \frac{1}{h}\left[\frac{2}{3}(u_{n+1} - u_{n-1}) - \frac{1}{12}(u_{n+2} - u_{n-2})\right] + O(h^4).$$

Аналогичное уточнение можно провести в точке $x_{n+1/2}$. Здесь симметричными окружениями являются пары x_n, x_{n+1} и x_{n-1}, x_{n+2}. Второй отрезок втрое длиннее первого, поэтому $r = 3$. Уточненное значение производной равно

$$u'_{n+1/2} \approx \frac{9}{8}\frac{u_{n+1} - u_n}{x_{n+1} - x_n} - \frac{1}{8}\frac{u_{n+2} - u_{n-1}}{x_{n+2} - x_{n-1}}.$$

Нетрудно записать это выражение на равномерной сетке.

Для второй производной ограничимся случаем только равномерной сетки:

$$\bar{u}''_n \approx \frac{u_{n+1} - 2u_n + u_{n-1}}{h^2}; \quad \tilde{u}''_n \approx \frac{u_{n+2} - 2u_n + u_{n-2}}{4h^2}.$$

Эта пара формул имеет порядок точности $p = 2$, а коэффициент сгущения сетки составляет $r = 2$. Уточнение по Ричардсону дает

$$u''_n = \frac{4}{3}\bar{u}''_n - \frac{1}{3}\tilde{u}''_n + O(h^4) =$$

$$= \frac{4}{3}\frac{u_{n+1} - 2u_n + u_{n-1}}{h^2} - \frac{1}{12}\frac{u_{n+2} - 2u_n + u_{n-2}}{h^2} + O(h^4).$$

Напомним, что формулы, рассчитанные на $h = \text{const}$, недопустимо применять к неравномерным сеткам.

6.1.5. Старшие производные

В решениях таких прикладных задач, как изгиб упругого бруска, требуются производные вплоть до четвертой. Выражения для них на неравномерных сетках очень громоздкие. Однако на равномерных сетках соответствующие формулы имеют простой вид. Покажем, как их строить.

Третья производная может рассматриваться как производная от второй. Поэтому для нее можно написать следующую аппроксимацию в полуцелой точке:

$$u'''_{n+1/2} \approx \frac{u''_{n+1} - u''_n}{h}.$$

Подставив вместо вторых производных формулу (6.13), где для u''_{n+1} надо сдвинуть все индексы на 1, получим

$$u'''_{n+1/2} \approx \frac{u_{n+2} - 3u_{n+1} + 3u_n - u_{n-1}}{h^3}. \qquad (6.20)$$

Четвертую производную можно рассматривать как вторую от второй производной. Это дает в целом узле

$$u_n^{(4)} \approx \frac{u''_{n+1} - 2u''_n + u''_{n-1}}{h^2} \approx \approx \frac{u_{n+2} - 4u_{n+1} + 6u_n - 4u_{n-1} + u_{n-2}}{h^4}. \tag{6.21}$$

Легко заметить общую закономерность написания формул (6.13), (6.14) (6.20), (6.21): 1) производная нечетного порядка берется в полуцелом узле, четного – в целом; 2) окружающие узлы выбраны симметрично, а их число на единицу больше порядка производной q; 3) перед узловыми значениями u стоят биномиальные коэффициенты для степени q, знаки которых чередуются; 4) степень h в знаменателе равна порядку производной. Разложением в ряд Тейлора можно показать, что точность всех этих формул есть $O(h^2)$. Это очевидно из симметрии формул, поскольку члены нечетного порядка точно сокращаются.

6.2. ДИФФЕРЕНЦИРОВАНИЕ ИНЫХ АППРОКСИМАЦИЙ

6.2.1. Интерполяционный сплайн

Пусть функция задана в большом диапазоне значений аргумента, так что имеется подробная таблица (x_n, u_n), $0 \leqslant n \leqslant N$ с $N \gg 1$. В этом случае около каждого узла используется свой локальный многочлен Ньютона. Два интерполяционных многочлена, построенных по соседним группам узлов, стыкуются в общем узле непрерывно в силу условия интерполяции. Однако их производные в этом узле могут оказаться несовпадающими. Тогда аппроксимация производной на всем отрезке окажется разрывной.

Этих трудностей можно избежать, аппроксимируя $u(x)$ интерполяционным сплайном. Пусть для определенности взят кубический сплайн $S_3(x)$, построенный в подразд. 4.2.2. Его первая и вторая производные непрерывны всюду, включая узлы. Напомним выражения для сплайна и его производных:

$$S_{3n}(x) = a_n + b_n(x - x_{n-1}) + c_n(x - x_{n-1})^2 + d_n(x - x_{n-1})^3,$$
$$x_{n-1} < x < x_n, \quad 1 \leqslant n \leqslant N;$$
$$S'_{3n}(x) = b_n + 2c_n(x - x_{n-1}) + 3d_n(x - x_{n-1})^2,$$
$$x_{n-1} < x < x_n, \quad 1 \leqslant n \leqslant N;$$

$$S''_{3n}(x) = 2c_n + 6d_n(x - x_{n-1}), \quad x_{n-1} < x < x_n, \quad 1 \leqslant n \leqslant N.$$

Третья производная разрывна (кусочно-постоянна), и поэтому не представляет интереса. Погрешность самого сплайна при выборе периодических или естественных дополнительных условий составляет $O(h^4)$. Погрешность его первой и второй производных составляет $O(h^3)$ и $O(h^2)$ соответственно.

Достоинства этого способа состоят в следующем: 1) способ позволяет найти производные с указанным порядком точности в произвольной точке, а не только в узлах; 2) дает эту точность не только на равномерных сетках, но и на квазиравномерных, а также на достаточно произвольных неравномерных сетках, если они не содержат больших интервалов. Дифференцированием многочлена Ньютона таких результатов добиться не удается.

6.2.2. Метод выравнивания

Если функция быстро меняется от одного узла к другому, то для интерполяции следует подобрать выравнивающие переменные $\eta(u), \xi(x)$ (см. подразд. 4.3.1). Дифференцировать также следует в выравнивающих переменных. Найдя численно искомую точку ξ и $d\eta/d\xi$ в данной точке, перейдем к производной в исходных переменных с помощью точного соотношения

$$\frac{du}{dx} = \frac{d\eta}{d\xi}\frac{d\xi}{dx}\Big/\frac{d\eta}{du}. \tag{6.22}$$

Производные $d\xi/dx$ и $d\eta/du$ находят дифференцированием точных формул выравнивающих преобразований. Аналогично можно вычислять и более высокие производные. Однако уже для второй производной аналогом (6.22) будет весьма громоздкое выражение. Поэтому с помощью метода выравнивания на практике вычисляют только первую производную.

Пример 6.2. Вернемся к примеру 4.3. Для него в исходных переменных почленное дифференцирование кубического интерполяционного многочлена Ньютона дает следующее выражение:

$$P'(x) = 4 + 11(2x - 1) + 22(3x^2 - 6x + 2).$$

Коэффициенты в последовательных слагаемых не убывают, так что на разумный результат рассчитывать не приходится. Зато в выравнивающих переменных получаем

$$\eta'(\xi) = 0{,}6990 + 0{,}0467(2\xi - 1) - 0{,}0080(3\xi^2 - 6\xi + 2);$$
$$\eta'(1{,}5) = 0{,}7944.$$

Таблица 6.1

Дифференцирование в выравнивающих переменных

h	$u'(x)$	$\eta'(\xi)$
3	70	0,7748
1	26	0,7924
Экстремум	$20{,}5 \pm 5{,}5$	$0{,}7946 \pm 0{,}0022$

Здесь убывание коэффициентов слагаемых неплохое. Переход к исходным переменным предлагается читателям сделать самостоятельно.

Контроль. В примере 4.3 сетка содержала узлы $0, 1, 2, 3$. Точка $x = \xi = 1{,}5$ является серединой двух симметричных интервалов с отношением шагов $3 : 1$. В этом случае производную можно находить по простейшей формуле (6.14). Поэтому к производным в исходных и выравнивающих переменных можно применять уточнение по Ричардсону, полагая $p = 2$ и $r = 3$. Результаты представлены в табл. 6.1. Видно, что в исходных переменных погрешность огромная, а в выравнивающих – невелика и обеспечивает точность порядка 0,5 %.

Замечание. Для уточнений в точках, промежуточных между узлами, метод Ричардсона не всегда применим. В примере 6.2 точка $x = 1{,}5$ делила оба отрезка $[1, 2]$ и $[0, 3]$ точно пополам, то есть была для них полуцелой. В этих условиях метод Ричардсона использовать можно. Однако если взять другую точку, например $x = 1{,}2$, то она делит отрезок $[1, 2]$ в отношении $1 : 4$, а отрезок $[0, 3]$ в отношении $2 : 3$. Это разные отношения, и в такой ситуации применение метода Ричардсона было бы грубой ошибкой. Для его использования соседние сетки всегда должны быть подобны.

6.2.3. Среднеквадратичное приближение

Наилучшую аппроксимацию в норме L_2 дают кривые неточно проходящие через заданные точки; они слегка сглаживают мелкие осцилляции исходных кривых. Поэтому дифференцирование среднеквадратичных приближений обычно дает лучшие результаты, чем дифференцирование интерполяционных приближений. Общая формула дифференцирования обобщенного многочлена имеет вид

$$u^{(q)}(x) \approx \Phi_M^{(q)}(x) = \sum_{m=0}^{M} c_m \varphi_m^{(q)}(x), \qquad (6.23)$$

т. е. сумма дифференцируется почленно. Рассмотрим важнейшие случаи.

Ряд Фурье. Напомним, что для ряда Фурье $\varphi_{2m-1}(x) = \sin(mx)$, $\varphi_{2m}(x) = \cos(mx)$ и суммирование в (6.23) ведется не до M, а до $2M$. Поскольку

$$\frac{d\sin(mx)}{dx} = m\cos(mx); \quad \frac{d\cos(mx)}{dx} = -m\sin(mx), \qquad (6.24)$$

то каждое дифференцирование приводит к умножению коэффициентов ряда на m, замене синусов на косинусы и сменам знака. Тем самым коэффициенты Фурье для производных убывают медленнее, чем для исходной функции. Напомним, что для p раз непрерывно дифференцируемой $u(x)$ выполнялось $c_m = O(m^{-(p+1)})$. Значит, для производной $u^{(q)}(x)$, $q \leqslant p$, величины коэффициентов Фурье будут $O(m^{-(p-q+1)})$.

Сходимость ряда Фурье для производных соответственно ухудшается. Из полученной скорости убывания коэффициентов аналогично подразд. 5.2.2 следуют оценки

$$\|u^{(q)} - \Phi_M^{(q)}\|_{L_2} = O(M^{-(p-q+1/2)})$$

$$\text{и} \quad \|u^{(q)} - \Phi_M^{(q)}\|_C = O(M^{-(p-q)}), \quad q \leqslant p.$$

Отметим важное обстоятельство. Для периодической функции границ отрезка фактически нет, так как любую точку можно принять за начало периода. Поэтому ряд одинаково хорошо сходится в любой точке, включая граничные точки отрезка. Ухудшение сходимости в норме C имеет место лишь в точках разрыва старшей существующей производной. Это выгодно отличает тригонометрический ряд Фурье от других рядов.

Ряд по $T_m(x)$. Ряд для q-й производной находится по общей формуле (6.23). Выражения для первых производных записываются просто:

$$\varphi'_m(x) = \frac{dT_m(x)}{d\theta} \Big/ \frac{dx}{d\theta} = \frac{m\sin(m\theta)}{\sin\theta}; \quad \theta = \arccos x. \qquad (6.25)$$

Отношение синусов — это многочлен Чебышева II рода. Аналогично можно получить выражения для $\varphi''_m(x)$, однако оно будет более громоздким; поэтому двукратное дифференцирование используют заметно реже. По этой причине большие кратности дифференцирования практически не применяются.

Разложение самой функции в ряд по $T_m(x)$ может хорошо сходиться во всех точках. Однако в подразд. 5.3.1 показано,

что производные этих многочленов очень велики вблизи границ отрезка. Если для тригонометрического ряда Фурье однократное дифференцирование было эквивалентно умножению $\varphi_m(x)$ на m, то для многочленов Чебышева каждое дифференцирование эквивалентно умножению на m в средней части отрезка и на m^2 вблизи границ отрезка; это видно из (5.35). Для первой производной из (6.25) непосредственно видно общее умножение на m; а вблизи границы при $\theta \to 0$ и $\theta \to \pi$ раскрытие неопределенности отношения синусов по правилу Лопиталя дает дополнительный множитель m. Поэтому сходимость рядов для производных будет особенно быстро ухудшаться вблизи границ отрезка, причем тем сильнее, чем выше порядок производной.

Плохая сходимость рядов для производных вблизи границ отрезка обычно не позволяет рассчитывать на хорошую экстраполяцию за пределы отрезка.

Двойной период. Дифференцирование выполняется по общей формуле (6.23), где нахождение производных от тригонометрических функций $\varphi_m(x)$ и $\psi_k(x)$ аналогично (6.24). Остается выяснить скорость сходимости полученных рядов. Будем предполагать, что на исходном отрезке существует непрерывная $u^{(p)}(x)$, причем взято число функций двойного периода $K \leqslant p+1$, а число пар слагаемых функций основного периода $M \gg K$.

Напомним, что подключение K функций двойного периода эквивалентно исключению разрыва в периодическом продолжении самой функции и ее $K-1$ производных. Тогда при фиксированном K и возрастании M ряд сходится не хуже, чем для периодической функции с $K-1$ непрерывной производной. Тем самым погрешности разложения будут

$$\|\Phi_M^{(q)} + \Psi_K^{(q)} - u^{(q)}\|_{L_2} = O(M^{-(K-q+1/2)});$$

$$\|\Phi_M^{(q)} + \Psi_K^{(q)} - u^{(q)}\|_C = O(M^{-(K-q)}), \quad q \leqslant K-1 \leqslant p.$$

Сходимость вблизи границ отрезка будет несколько хуже, чем у тригонометрического ряда Фурье для периодической функции, так как границы здесь являются истинными, а не формальными. Однако сходимость гораздо лучше, чем для разложения по многочленам Чебышева, а формулы для многократного дифференцирования просты. Поэтому метод двойного периода позволяет хорошо приближать старшие производные.

Среднеквадратичный сплайн. Для построения среднеквадратичных аппроксимаций использовались разложения

функции $u(x)$ по B-сплайнам p-й степени (см. 5.5). Дифференцируя разложение (5.61), получим

$$S_p^{(q)}(x) = \sum_{n=-p}^{N-1} c_n B_{pn}^{(q)}(x).$$

Заметим, что усеченная степень (5.52) дифференцируется практически как обычная:

$$\frac{d^q}{dx^q}\xi_+^p = \frac{p!}{(p-q)!}\xi_+^{p-q}, \quad q \leqslant p, \quad p = 0, 1, 2, \ldots$$

Тогда дифференцирование формулы (5.56) дает

$$B_{pn}^{(q)}(x) = \frac{p!}{(p-q)!} \begin{cases} \sum_{k=0}^{p} b_{pnk}(x - x_{n+k})_+^{p-q} & \text{при } x \in (x_n, x_{n+p+1}]; \\ 0 & \text{вне } (x_n, x_{n+p+1}], \end{cases}$$
$$q \leqslant p. \tag{6.26}$$

Расчет по формуле (6.26) устойчив при $p < 6$, для больших значений p возможно накопление ошибок округления. Коэффициенты b_{pnk} в (6.26) вычисляются по формуле (5.57).

Если необходимо провести расчеты для более высоких степеней сплайна, то надо дифференцировать рекуррентное соотношение (5.60). Для первой производной это дает

$$B'_{p+1,n}(x) = \frac{(x - x_n)B'_{pn}(x) + B_{pn}(x)}{x_{n+p+2} - x_n} + \\ + \frac{(x_{n+p+3} - x)B'_{p,n+1}(x) - B_{p,n+1}(x)}{x_{n+p+3} - x_{n+1}}. \tag{6.27}$$

Начинать расчет надо со сплайна первой степени, взяв в правой части $B_{0n}(x)$ согласно (5.53) и $B'_{0n}(x) \equiv 0$. Рекуррентные формулы для старших производных более громоздки. Их нетрудно записать, дифференцируя (6.25).

Напомним, что сплайн p-й степени целесообразно применять, если $u(x)$ имеет не менее $p + 1$ непрерывных производных. Погрешность q-й производной в этом случае есть $O(h^{p-q+1})$; асимптотически точная оценка этой погрешности приведена в (5.63), (5.64). Порядок точности этих формул таков же, как и для интерполяционных многочленов степени p, однако коэффициенты

в остаточных членах иные. Для старшей ненулевой производной $S_p^{(p)}(x)$ они таковы же, как для интерполяционного многочлена. Но для производных порядка $q < p$ эти коэффициенты меньше, чем у многочлена; чем меньше q, тем сильнее этот выигрыш в точности. Наибольший выигрыш достигается при $q = 0$, т. е. для самого сплайна $S_p(x)$. Чем выше степень сплайна p, тем больше выигрыш в точности. Поэтому, строя среднеквадратичные сплайны высоких степеней, можно хорошо аппроксимировать не только функцию, но и ее довольно высокие производные.

Замечание. В подразд. 5.2.4 отмечалось, что наилучшее среднеквадратичное приближение обычно выгоднее наилучших равномерных приближений для аппроксимации функций. Это преимущество еще сильнее при дифференцировании аппроксимаций. Ряды, дающие наилучшее равномерное приближение, плохо дифференцируются: почленное дифференцирование обычно приводит к медленно сходящимся или даже расходящимся рядам уже для первой производной. Для старших производных положение усугубляется.

6.3. НЕКОРРЕКТНОСТЬ ЧИСЛЕННОГО ДИФФЕРЕНЦИРОВАНИЯ

6.3.1. Дифференцирование интерполяционного многочлена

Операция дифференцирования является некорректной. Проиллюстрируем это на примере. Построим следующее бесконечно малое возмущение функции $u(x)$:

$$\delta u(x) = \frac{1}{\omega} \sin \omega^2 x, \quad \omega \to \infty.$$

Тогда возмущение первой производной равно $\delta u'(x) = \omega \cos \omega^2 x$. Видно, что

$$\|\delta u(x)\|_C = \frac{1}{\omega} \ll 1, \quad \|\delta u'(x)\|_C = \omega \gg 1.$$

Значит можно подобрать такое сколь угодно малое возмущение функции, что возмущение производной будет сколь угодно большим. Таким образом, операция дифференцирования является некорректной.

В численном дифференцировании с помощью интерполяционного многочлена некорректность проявляется следующим об-

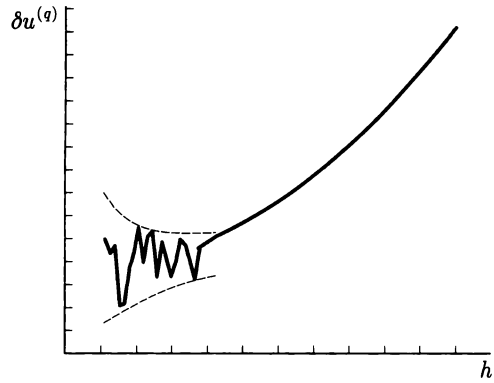


Рис. 6.1. Погрешность численного дифференцирования

разом. Пусть мы хотим найти $u'(x)$ по двум точкам. Если эти точки удалены друг от друга, то велики шаг h и погрешность аппроксимации (см. 6.1). Чтобы уменьшить погрешность, надо сближать точки. Но тогда значение функции в этих точках мало различаются. Вычисления проводятся с конечным числом знаков. Поэтому при вычитании близких значений сильно возрастают ошибки округления, и в разности может остаться малое число достоверных знаков.

С этой трудностью давно столкнулись вычислители при численном дифференцировании табулированных функций. Они выработали следующую рекомендацию: выбирать шаг h небольшим, чтобы ошибка аппроксимации была невелика; но не слишком малым, чтобы ошибки округления еще не сказывались бы. Оценим оптимальную величину шага h_0, для простоты предполагая сетку равномерной.

Пусть вычисляется производная $u^{(q)}(x)$ по формуле точности $O(h^p)$. Такая формула использует $p+q+1$ точку таблицы. Обозначим характерные величины ошибок округления через ε. Характерная зависимость погрешности $\delta u^{(q)}$ от шага показана на рис. 6.1. В правой части шаг достаточно велик, ошибки округления не сказываются, и погрешность обусловлена аппроксимацией: $\delta u^{(q)} \approx \alpha h^p$. В левой половине шаг настолько мал, что преобладают ошибки округления. Их можно оценить не асимптотически, а лишь по порядку величины.

Из формул (6.13), (6.14), (6.20), (6.21) видно, что малые разности узловых значений функции делятся на h^q, так что $\delta u^{(q)} \sim$ $\sim \varepsilon/h^q$. Это соответствует возрастанию ошибки при уменьшении h. Оптимален тот шаг, при котором погрешности аппроксимации

и погрешность округления приблизительно равны: $\varepsilon/h^q \sim \alpha h^p$. Это дает

$$h_0 \sim (\varepsilon/\alpha)^{1/(p+q)}, \quad \delta u^{(q)}(x) \sim \varepsilon^{p/(p+q)}\alpha^{q/(p+q)}. \tag{6.28}$$

При $h > h_0$ погрешность регулярно убывает с уменьшением шага. При $h < h_0$ погрешность возрастает с уменьшением шага, причем нерегулярно.

При 64-разрядных вычислениях $\varepsilon \approx 10^{-16}$. Если для $u'(x)$ взять формулу точности $O(h^2)$ с $\alpha \sim 1$, то получим $h_0 \sim 10^{-5}$, а $\delta u' \sim 10^{-10}$. Более высокой точности добиться невозможно. Если для той же производной взять формулу точности $O(h^4)$, то будет $h_0 \sim 10^{-3}$, а $\delta u' \sim 10^{-12}$; предельная точность возросла, но одновременно увеличился оптимальный шаг, а брать шаг меньше оптимального нельзя (об этом нередко забывают).

Контроль точности. Прежде чем проводить вычисления, необходимо оценить h_0 и работать с шагом заведомо бóльшим h_0. Еще более надежные результаты получаются при следующей тактике. Сначала проводят вычисления с заведомо грубым шагом h. Затем уменьшают h до тех пор, пока вычисленная $u^{(q)}(x)$ монотонно стремится к некоторому пределу. Скорость стремления целесообразно оценивать по методам Ричардсона или Эйткена. Она должна соответствовать p-му порядку точности. Когда этот закон нарушается, дальнейшее уменьшение шага следует прекращать.

Поскольку брать шаг меньше оптимального h_0 нельзя, а этот шаг сравнительно велик, то много раз сгустить сетку не удастся. Соответственно многократное повышение точности тоже нереально. Возьмем начальный шаг в 10 — 100 раз крупнее, чем грубо оцененный h_0. Например, для вычисления первой производной разумно взять $h_1 \sim 10^{-3}$. Уменьшать шаг сетки удобно в одно и то же число раз, но не обязательно в целое. Например, можно взять $r = \sqrt{2}$ или $r = 1{,}5$. Построим последовательность шагов

$$h_1 \sim 10^{-3}, \quad h_2 = h_1/r, \quad h_3 = h_2/r, \ldots \quad (h_k > h_0).$$

Пусть мы вычисляем первую производную. Возьмем для нее симметричную разность:

$$w_k \equiv u'(x; h_k) = \frac{u(x + h_k) - u(x - h_k)}{2h_k}, \quad k = 1, 2, \ldots \tag{6.29}$$

Ричардсоновское уточнение по двум соседним сеткам $\bar{w}_k$ и оценка погрешности δ_k равны

$$\bar{w}_k = w_k + \delta_k, \quad \delta_k = \frac{w_k - w_{k-1}}{r^p - 1}, \quad k = 2, 3, \ldots \tag{6.30}$$

Здесь $p = 2$ есть теоретический порядок точности исходной разностной формулы. Эффективный порядок точности вычисляется по трем сеткам:

$$p_k = \frac{\ln(\delta_{k-1}/\delta_k)}{\ln r}, \quad k = 3, 4, \ldots \tag{6.31}$$

Если значение p_k мало отличается от теоретического значения $p = 2$, то в качестве ответа можно брать значение $\bar{w}_k$ с оценкой погрешности δ_k (6.30). Можно уменьшать шаг h_k до тех пор, пока значение p_k близко к теоретическому. При этом оценки погрешности будут уменьшаться от одной сетки к другой приблизительно в r^p раз, т. е. точность результата будет повышаться.

Описанная процедура довольно громоздка и сравнительно трудоемка. Однако при скорости современных компьютеров это обычно несущественно. Зато она позволяет надежно вычислять разностные производные, давая высокую точность и гарантируя оценку достигнутой погрешности.

Выравнивание. Не следует забывать о переходе к выравнивающим переменным. Он существенно ослабляет некорректность задачи численного дифференцирования. В самом деле, если в исходных переменных функция сильно меняется от одного узла к другому, то в выравнивающих – слабо. Это эквивалентно тому, что коэффициент α в оценке ошибки аппроксимации уменьшается на порядки. Из (6.28) видно, что при этом оптимальный шаг h_0 заметно возрастает, а наилучшая достижимая точность существенно улучшается.

6.3.2. Дифференцирование рядов

Наилучшие среднеквадратичные приближения сглаживают исходную кривую, и этим уменьшают влияние некорректности дифференцирования. Однако в них есть другой источник некорректности. Неопытные вычислители для повышения точности стараются просуммировать возможно большее число членов ряда. Поскольку члены ряда могут иметь разные знаки, такое суммирование само приводит к накоплению ошибок округления (см. пример с компьютерным вычислением синуса в предисловии: $\sin 2\,550^\circ = 29{,}5$).

Коэффициенты обобщенного многочлена c_m вычисляются не точно, а с некоторыми погрешностями Δ_m. Часть этой погрешности связана с ошибками табулированной функции и ошибками при вычислении коэффициентов. Но наибольшая часть

этой ошибки обычно является погрешностью квадратурной формулы, с помощью которой вычисляют скалярные произведения. Она может на много порядков превышать уровень ошибок округления компьютера. В итоге сам обобщенный многочлен и его производные вычисляются с погрешностями

$$\delta\Phi_M^{(q)}(x) = \sum_{m=0}^{M} \Delta_m \varphi_m^{(q)}(x).$$

Когда мы прибавляем к многочлену дополнительное M-е слагаемое, то, с одной стороны, уменьшаем ошибку аппроксимации на $c_M\varphi_M^{(q)}(x)$, с другой — коэффициент известен с ошибкой, поэтому вносится дополнительная погрешность $\Delta_M\varphi_M^{(q)}(x)$. Величины коэффициентов c_m в среднем достаточно быстро убывают (иначе сходимость ряда не была бы обеспечена). Величины Δ_m обычно имеют примерно одинаковый уровень. Очевидно, имеет смысл суммировать ряд до тех пор, пока

$$|c_M| > |\Delta_M|, \tag{6.32}$$

т. е. суммировать до того момента, пока добавление новых слагаемых улучшает точность. Если критерий (6.32) нарушен, то прибавлять дальнейшие члены бессмысленно и опасно: погрешность может катастрофически возрасти.

Глава 7

СПЕКТР МАТРИЦЫ

7.1. ПРЕОБРАЗОВАНИЕ ПОДОБИЯ

7.1.1. Теория

Спектр. Наиболее трудной задачей линейной алгебры является так называемая алгебраическая проблема собственных значений. Пусть задана квадратная матрица A порядка N (ее элементы могут быть комплексными). Рассмотрим систему линейных уравнений

$$A\mathbf{x} = \lambda\mathbf{x}, \tag{7.1}$$

где $\mathbf{x}$ — N-мерный вектор; λ — число.

Если существует число λ, такое, что система (7.1) имеет нетривиальное решение ($\mathbf{x} \neq 0$), то λ называют собственным значением матрицы, а $\mathbf{x}$ — соответствующим собственным вектором.

Задача нахождения всех собственных значений и всех собственных векторов матрицы A называется полной проблемой собственных значений. Если нужны только некоторые λ и $\mathbf{x}$, то говорят о частичной проблеме.

Напомним некоторые элементы теории, известные из курса линейной алгебры. Перепишем задачу (7.1) в следующем виде: $(\mathbf{A} - \lambda E)\mathbf{x} = 0$. Поскольку $\mathbf{x} \neq 0$, то в круглых скобках стоит вырожденная матрица, и $\det(\mathbf{A} - \lambda E) = 0$. Но определитель N-го порядка такой структуры есть многочлен N-й степени от λ и задача сводится к нахождению его корней. Он имеет ровно N корней (с учетом кратности, если среди корней есть совпадающие). Таким образом, матрица N-го порядка имеет N собственных значений.

Собственные значения могут быть комплексными, даже если матрица A вещественная. Если матрица A эрмитова ($A = A^H$, т. е. $a_{nm} = a^*_{mn}$, где * означает комплексное сопряжение), то все ее собственные значения вещественны, даже если элементы матрицы комплексные. Напомним, что эрмитова матрица с веще-

ственными элементами — это просто вещественная симметричная матрица.

Косоэрмитовой называют матрицу, элементы которой удовлетворяют соотношению $a_{nm} = -a^*_{mn}$. Косоэрмитова матрица с вещественными элементами — это кососимметричная матрица. При умножении всех элементов на $\mathbf{i}$ такая матрица превращается в эрмитову. Поэтому все собственные значения косоэрмитовых (кососимметричных) матриц чисто мнимые.

Прикладные задачи. Имеются две большие группы прикладных задач, приводящих к нахождению спектра матрицы. Одна группа включает расчет конструкций, составленных из треугольников: ферм железнодорожных мостов, несущих конструкций корпусов кораблей, самолетов, ракет, автомобилей, купольных перекрытий и т. п. Собственные значения матриц в этом случае имеют смысл частот собственных колебаний этих конструкций. Резонанс хотя бы одной собственной частоты с внешним воздействием может привести к разрушению конструкции (флаттер стал причиной гибели многих самолетов). Поэтому в таких задачах необходимо вычислять весь спектр матрицы, а сама матрица является достаточно произвольной. Типичные порядки таких матриц $N \leqslant 100$.

Вторая группа задач — задачи о спектре дифференциальных уравнений с заданными краевыми условиями (колебания струны, упругого бруска и т. п.). Матрица возникает при аппроксимации дифференциального уравнения разностной схемой. Порядок матрицы равен числу узлов разностной сетки и может быть огромным: $N \sim 1\,000$ и более. Сами матрицы оказываются симметричными и слабо заполненными (нередко трехдиагональными). Их низшие собственные значения хорошо аппроксимируют собственные значения дифференциального уравнения; их вычислением обычно ограничиваются. Верхние собственные значения матрицы далеки от соответствующих собственных значений дифференциального уравнения, так что их вычислять бессмысленно. Для чисто математических целей бывает необходимо найти наибольшее и наименьшее собственное значение.

Собственные векторы. Очевидно, собственные векторы определены с точностью до численного множителя. Этот множитель обычно находят из условия нормировки $(\mathbf{x}, \mathbf{x}) = 1$. Даже при этом для вещественных собственных векторов множитель вычисляется с точностью до знака, а для комплексных векторов с точностью до множителя $e^{i\varphi}$.

Для любой матрицы каждому простому (некратному) собственному значению соответствует один и только один собственный вектор. Собственные векторы, соответствующие различным простым собственным значениям, линейно независимы (если матрица эрмитова, то эти векторы ортогональны). Поэтому если все собственные значения матрицы не кратные, ее собственные векторы образуют базис.

Собственному значению кратности q может соответствовать, в зависимости от типа матрицы, от одного до q линейно независимых собственных векторов. Если хотя бы у одного из кратных собственных значений матрицы число собственных векторов меньше кратности собственного значения, то собственные векторы этой матрицы не образуют базиса.

Если матрица эрмитова, то q-кратному собственному значению соответствует ровно q линейно независимых собственных векторов. Их всегда можно ортогонализировать. К прочим собственным векторам матрицы они будут ортогональны, поскольку соответствуют другому собственному значению. Таким образом, собственные векторы эрмитовой матрицы образуют ортогональный базис.

Жордановы подматрицы. В качестве примера рассмотрим такие матрицы четвертого порядка:

$$A = \begin{pmatrix} a & 0 & 0 & 0 \\ 0 & a & 0 & 0 \\ 0 & 0 & a & 0 \\ 0 & 0 & 0 & a \end{pmatrix}; \quad G_4 = \begin{pmatrix} a & 1 & 0 & 0 \\ 0 & a & 1 & 0 \\ 0 & 0 & a & 1 \\ 0 & 0 & 0 & a \end{pmatrix}; \quad B = \begin{pmatrix} a & 0 & 0 & 0 \\ 0 & a & 1 & 0 \\ 0 & 0 & a & 1 \\ 0 & 0 & 0 & a \end{pmatrix}. \tag{7.2}$$

У каждой из них характеристическое уравнение принимает вид $(a-\lambda)^4 = 0$. Следовательно, есть собственное значение $\lambda = a$ кратности $q = 4$. Однако у первой матрицы A есть четыре линейно независимых собственных вектора

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{e}_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}; \tag{7.3}$$

это легко проверить, поочередно подставляя векторы (7.3) в равенство (7.1).

У второй же матрицы имеется только один собственный вектор $\mathbf{e}_1$. В самом деле, пусть ее собственный вектор $\mathbf{x}$ имеет компоненты x_n; тогда уравнение (7.1) примет для нее вид

$$(G_4 - \lambda E)\mathbf{x} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} x_2 \\ x_3 \\ x_4 \\ 0 \end{pmatrix} = 0, \quad \lambda = a.$$

Отсюда $x_2 = x_3 = x_4 = 0$, а $x_1 = 1$ в силу условия нормировки. Вторую матрицу G_4 называют простой жордановой (или классической) матрицей.

Третья матрица имеет так называемую каноническую жорданову форму (по диагонали стоят либо числа, либо жордановы подматрицы, а остальные элементы равны нулю). Ее собственными векторами являются $\mathbf{e}_1$ и $\mathbf{e}_2$; в этом легко убедиться с помощью выкладки, аналогичной только что проделанной.

Обусловленность. Ошибки матричных элементов могут быть опасными. Рассмотрим такой пример. Возьмем жорданову матрицу G_N порядка N, аналогичную (7.2). Слегка изменим ее, положив угловой элемент $g_{N_1} = \varepsilon \ll 1$ вместо чистого нуля. Тогда характеристическое уравнение примет следующий вид:

$$(a - \lambda)^N - (-1)^N \varepsilon = 0,$$

что легко проверяется разложением по первому столбцу. Это характеристическое уравнение легко решается:

$$\lambda = a + \varepsilon^{1/N}. \tag{7.4}$$

Величину $\varepsilon^{1/N}$ надо понимать в смысле функции комплексного переменного: она имеет N различных комплексных значений. Тем самым, вместо N-кратного корня $\lambda = a$ возмущенная матрица имеет N различных собственных значений. Возьмем $\varepsilon = 10^{-16}$, что соответствует ошибкам округления при 64-разрядных вычислениях. Тогда при $N \sim 100$ возмущение собственного значения $|\varepsilon|^{1/N} = 10^{-0{,}16} \approx 0{,}7$. Оно совсем не мало. Это свидетельствует об очень плохой обусловленности задачи. При настолько плохой обусловленности задачу решать численно практически невозможно.

Однако рассмотренный случай был нетипичным. Если в жордановой матрице порядок невелик, или портится элемент, расположенный близко от диагонали, то обусловленность будет существенно лучше.

Собственные значения эрмитовой матрицы устойчивы по отношению к возмущению любых элементов.

Проблема устойчивости собственных векторов еще более сложна. В рассмотренном примере исходная жорданова матрица имела один собственный вектор, а возмущенная — N линейно независимых собственных векторов. Таким образом, имея дело с матрицей неизвестной структуры, надо соблюдать осторожность.

Подобие. Собственные значения очень легко найти, если матрица является диагональной ($a_{nm} = 0$ при $n \neq m$) или треугольной ($a_{nm} = 0$ при $n > m$ для верхней треугольной). Характеристическое уравнение такой матрицы имеет вид $\prod(a_{nn} - \lambda) = 0$, так что собственные значения равны диагональным элементам: $\lambda = a_{nn}$. Поиск собственных значений облегчается также, если матрица является трехдиагональной или почти треугольной. Поэтому возникает вопрос: «Нельзя ли упростить вид матрицы, не изменяя ее спектра?»

Такое упрощение можно делать преобразованием подобия. Возьмем некоторую неособенную матрицу F (т. е. $\det F \neq 0$). Такая матрица имеет обратную F^{-1}. Поэтому с помощью матрицы F можно проводить прямое и обратное преобразования базисов. Такому преобразованию базиса соответствует следующее преобразование матрицы A:

$$B = FAF^{-1}; \tag{7.5}$$

его называют преобразованием подобия.

Теорема 7.1. Преобразование подобия не меняет спектра матрицы.

Доказательство. Пусть μ и $\mathbf{y}$ — собственное значение и собственный вектор матрицы B. Последнее означает, что

$$B\mathbf{y} = \mu\mathbf{y}.$$

Подставив сюда (7.5), получаем $B\mathbf{y} = FAF^{-1}\mathbf{y} = \mu\mathbf{y}$. Умножив это равенство слева на F^{-1}, получим $AF^{-1}\mathbf{y} = \mu F^{-1}\mathbf{y}$. Последнее означает, что μ является собственным значением матрицы A, а $F^{-1}\mathbf{y}$ — соответствующим собственным вектором: $\mu = \lambda$, $F^{-1}\mathbf{y} = \mathbf{x}$. ■

Как известно, разложения по неортогональным базисам могут приводить к большим потерям точности. Во избежание этого стоит работать с ортогональными базисами. Поэтому удобны не

любые преобразования подобия. Следует выбирать такие матрицы F, которые преобразуют ортогональный базис также в ортогональный. Такие матрицы (вообще говоря, с комплексными элементами) называют унитарными и обозначают U. Вещественную унитарную матрицу называют ортогональной. Унитарные матрицы удовлетворяют соотношению $U^{-1} = U^H$.

Унитарное преобразование подобия не только хорошо обусловлено, оно еще сохраняет эрмитовость матриц; в самом деле, пусть $A = A^H$ — эрмитова матрица. Тогда

$$B^H = (UAU^{-1})^H = (U^{-1})^H A^H U^H = UAU^{-1} = B.$$

При этом использовалось правило взятия эрмитова сопряжения от произведения: переставить сомножители в обратном порядке и от каждого взять эрмитово сопряженное.

Зачастую для нужного упрощения матрицы A одного преобразования подобия оказывается недостаточно, и приходится проводить цепочку преобразований U_k. Это эквивалентно проведению одного преобразования подобия с матрицей $U = U_K \dots U_2 U_1$. Если каждое из преобразований было унитарным, то и результирующее преобразование будет унитарным.

7.1.2. Метод отражений

К сожалению, не существует такой конечной цепочки преобразований подобия, которая приводила бы произвольную матрицу A к диагональной или верхней треугольной форме. Однако имеются конечные цепочки, которыми можно преобразовать произвольную матрицу к верхней почти треугольной форме, а эрмитову — к трехдиагональной. Такие цепочки можно составить из двух классических видов унитарных преобразований: вращения системы координат и зеркального отражения. Самым быстрым из известных и одновременно наиболее устойчивым является метод отражений, предложенный Хаусхолдером (1958).

Отражение. Произведем в N-мерном векторном пространстве отражение относительно некоторой гиперплоскости, проходящей через начало координат. Преобразование полностью определяется заданием нормали $\mathbf{w}$ к гиперплоскости (в общем случае это комплексный вектор). Данная нормаль есть нормированный вектор-столбец. Условие нормировки имеет вид

$$(\mathbf{w}, \mathbf{w}) \equiv \mathbf{w}^H \mathbf{w} = \sum_{n=1}^{N} w_n^* w_n = 1, \qquad (7.6)$$

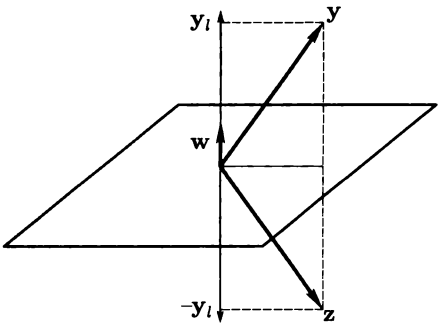


Рис. 7.1. Отражение вектора

где $\mathbf{w}^H$ — вектор-строка, эрмитово сопряженная к столбцу.

Возьмем произвольный вектор $\mathbf{y}$ и разложим его на две составляющие: параллельно нормали $\mathbf{y}_l = \mathbf{w}(\mathbf{w},\mathbf{y})$ и перпендикулярно ей. При отражении вектора его составляющая, перпендикулярная нормали, остается неизменной, а параллельная меняет знак (рис. 7.1), поэтому отраженный вектор $\mathbf{z}$ отличается от исходного на удвоенную величину параллельной компоненты:

$$\mathbf{z} = \mathbf{y} - 2\mathbf{w}(\mathbf{w},\mathbf{y}) \equiv \mathbf{y} - 2\mathbf{w}\mathbf{w}^H\mathbf{y} \equiv \mathbf{y} - 2(\mathbf{w}\mathbf{w}^H)\mathbf{y}.$$

Это преобразование вектора можно записать в канонической форме умножения на *матрицу отражения* R:

$$\mathbf{z} = R\mathbf{y}, \quad R = E - 2\mathbf{w}\mathbf{w}^H,$$

где умножение столбца $\mathbf{w}$ справа на строку той же длины $\mathbf{w}^H$ дает по правилам умножения прямоугольных матриц квадратную матрицу порядка N.

Исследуем свойства матрицы отражения. Эта матрица эрмитова, что непосредственно вытекает из следующей цепочки преобразований:

$$R^H = (E - 2\mathbf{w}\mathbf{w}^H)^H = E - 2(\mathbf{w}^H)^H\mathbf{w}^H = E - 2\mathbf{w}\mathbf{w}^H = R. \quad (7.7)$$

Возведем матрицу отражения в квадрат:

$$R^2 = (E - 2\mathbf{w}\mathbf{w}^H)(E - 2\mathbf{w}\mathbf{w}^H) = E - 4\mathbf{w}\mathbf{w}^H + 4\mathbf{w}\mathbf{w}^H\mathbf{w}\mathbf{w}^H.$$

Преобразуем последний член правой части, используя ассоциативность умножения матриц и условие нормировки (7.6):

$$\mathbf{w}\mathbf{w}^H\mathbf{w}\mathbf{w}^H = \mathbf{w}(\mathbf{w}^H\mathbf{w})\mathbf{w}^H = \mathbf{w}\mathbf{w}^H.$$

Тогда последний член сократится с предпоследним, и мы получим

$$RR = E, \quad \text{или} \quad R = R^{-1}, \quad (7.8)$$

т. е. матрица отражения равна своей обратной. Сравнив (7.7) и (7.8), убедимся, что $R^H = R^{-1}$, так что матрица отражений унитарна. Последнее свойство наиболее важно, поскольку унитарность обеспечивает сохранение эрмитовости и высокую устойчивость к ошибкам округления.

Цикл. Покажем, что для произвольной матрицы A можно подобрать такую конечную последовательность отражений, которая приводит матрицу к верхней почти треугольной форме. На q-м шаге при этом аннулируются элементы q-го столбца, лежащие ниже поддиагонали.

Будем считать, что уничтожен $q-1$ столбец. Разобьем матрицу A на клетки, как показано на рис. 7.2. Квадратная матрица A_1 есть верхняя почти треугольная, а в прямоугольной клетке A_3 только последний столбец отличен от нуля. Сделаем отражение с помощью вектора

$$\mathbf{w}^q = \{w_n^q\} = \begin{cases} w_n^q = 0 & \text{при } 1 \leqslant n \leqslant q; \\ w_n^q \neq 0 & \text{при } q+1 \leqslant n \leqslant N. \end{cases}$$

Первые q компонент вектора $\mathbf{w}^q$ нулевые. Далее верхний индекс q будем опускать.

Заметим, что если матрицу отражения R разбить на клетки аналогично матрице A, то она имеет следующий вид:

$$R = \begin{pmatrix} E_1 & 0 \\ 0 & W \end{pmatrix}, \quad W = E_4 - 2\mathbf{w}\mathbf{w}^H.$$

Из курса линейной алгебры известно, что если матрицы одинаковым образом разбиты на клетки, то они перемножаются по таким же правилам, как если бы эти клетки были обыкновенными элементами. Тогда искомое преобразование подобия принимает следующий вид:

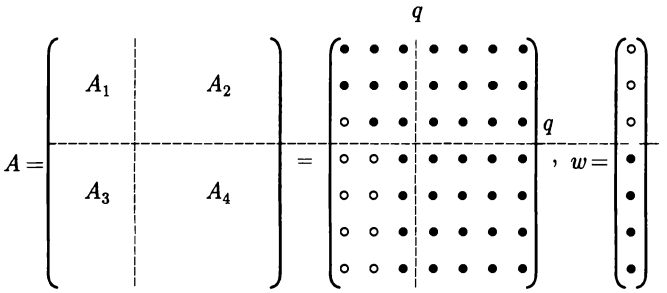


Рис. 7.2. Очередное отражение

$$B = RAR^{-1} = \begin{pmatrix} E_1 & 0 \\ 0 & W \end{pmatrix} \begin{pmatrix} A_1 & A_2 \\ A_3 & A_4 \end{pmatrix} \begin{pmatrix} E_1 & 0 \\ 0 & W \end{pmatrix} =$$

$$= \begin{pmatrix} A_1 & A_2 W \\ W A_3 & W A_4 W \end{pmatrix} \equiv \begin{pmatrix} B_1 & B_2 \\ B_3 & B_4 \end{pmatrix}.$$

Заметим, что клетку A_1 не нужно пересчитывать, так что клетка B_1 имеет нужную нам форму — верхнюю почти треугольную. Поскольку у клетки A_3 первые $q-1$ столбец нулевые, то и у клетки $B_3 = WA_3$ они также нулевые (тем самым их тоже не надо пересчитывать). Элементы q-го столбца клетки B_3 равны

$$b_{nq} = a_{nq} - 2w_n \sum_{m=q+1}^{N} w_m^* a_{mq} \equiv a_{nq} - cw_n, \quad q+1 \leqslant n \leqslant N, \tag{7.9}$$

где введено обозначение

$$c = 2 \sum_{m=q+1}^{N} w_m^* a_{mq}. \tag{7.10}$$

Надо так подобрать элементы вектора $\mathbf{w}$, чтобы все элементы столбца (7.9), кроме верхнего, обратились в нуль. Очевидно, для этого нужно положить

$$w_n = a_{nq}/c, \quad q+2 \leqslant n \leqslant N, \tag{7.11}$$

и найти, чему равна постоянная c. Заметим, что умножение вектора $\mathbf{w}$ на множитель $e^{i\varphi}$ не меняет матрицы отражения; тогда из (7.10) следует, что c также определена с точностью до этого множителя. Его всегда можно подобрать так, чтобы c была вещественной положительной, что и будем предполагать выполненным.

Введем вещественные положительные α и β с помощью соотношений

$$\alpha^2 = \sum_{n=q+2}^{N} |a_{nq}|^2; \quad \beta^2 = \sum_{n=q+1}^{N} |a_{nq}|^2 = \alpha^2 + |a_{q+1,q}|^2. \tag{7.12}$$

Затем выразим элемент w_{q+1} из условия нормировки (7.6) и определения (7.10) и преобразуем с помощью (7.11) и (7.12):

$$|w_{q+1}|^2 = 1 - \sum_{n=q+2}^{N} |w_n|^2 = 1 - \alpha^2/c^2; \tag{7.13}$$

$$w^*_{q+1}a_{q+1,q} = c/2 - \sum_{n=q+2}^{N} w^*_n a_{nq} = c/2 - \alpha^2/c. \qquad (7.14)$$

Возьмем квадрат модуля от соотношения (7.14) и исключим с его помощью $|w_{q+1}|^2$ из (7.13). Получим биквадратное уравнение для нахождения c:

$$c^4 - 4\beta^2 c^2 + 4\alpha^2\beta^2 = 0. \qquad (7.15)$$

Все его корни вещественны; нужные нам положительные корни равны

$$c = [2\beta(\beta \pm |a_{q+1,q}|)]^{1/2}. \qquad (7.16)$$

Выберем в (7.16) знак «+», чтобы $c \neq 0$ при $\alpha = 0$; таким образом мы избегаем возможного деления на нуль. Подставив выбранный корень в (7.11), находим последние компоненты $w_{q+2}, \ldots, w_N$ искомого вектора отражения $\mathbf{w}$. Подставив их в (7.14) и преобразуя с учетом вещественности β, получим первую компоненту $\mathbf{w}$:

$$w_{q+1} = \frac{1}{c}\left(a_{q+1,q} + \beta\frac{a_{q+1,q}}{|a_{q+1,q}|}\right) = \frac{c}{2\beta}\frac{a_{q+1,q}}{|a_{q+1,q}|}. \qquad (7.17)$$

В клетке B_3 остается единственный ненулевой элемент

$$b_{q+1,q} = a_{q+1,q} - cw_{q+1} = -\beta|a_{q+1,q}|/a^*_{q+1,q}. \qquad (7.18)$$

Заметим, что при этом получается

$$|b_{q+1,q}|^2 = \beta^2 = \sum_{n=q+1}^{N} |a_{nq}|^2,$$

т. е. сумма квадратов модулей элементов аннулируемого столбца сохраняется. Это означает, что в процессе расчета на нижней поддиагонали будут появляться большие элементы. На главной диагонали при этом существенных изменений величин элементов нет. В результате у итоговой верхней треугольной матрицы не будет преобладания главной диагонали даже в том случае, если эта матрица окажется эрмитовой, тем самым трехдиагональной. Обусловленность такой матрицы может оказаться не слишком хорошей.

С найденным вектором $\mathbf{w}$ проводится расчет клеток B_4 и B_2 (для эрмитовых матриц клетку B_2 вычислять не нужно: она эрмитово сопряжена клетке B_3).

Выводы. Выполняя отражения с $q = 1, 2, \ldots, N-2$, подобно преобразуем произвольную матрицу A к верхней почти треугольной форме (форме Хессенберга). Суммарное преобразование требует $(10/3)N^3$ арифметических операций (немногим больше, чем для обращения матрицы).

Если исходная матрица была эрмитовой, то ее эрмитовость сохраняется, и результирующая матрица оказывается трехдиагональной. В этом случае верхнюю половину матрицы нужно не вычислять, а заполнять по условию симметрии. Это уменьшает число арифметических операций до $(4/3)N^3$.

Если предполагается находить только собственные значения, то найденные на каждом шаге цикла векторы $\mathbf{w}$ и константы c хранить не надо. Если же нужно находить собственные векторы, то потребуется обратное преобразование. В этом случае $\mathbf{w}$ и c надо сохранять.

Практика показала, что метод Хаусхолдера очень устойчив по отношению к ошибкам округления. При 64-разрядных вычислениях он позволяет проводить расчеты с $N \sim 1\,000$ и более.

7.1.3. Другие методы

Метод Гивенса (1954). В этом методе используется цепочка унитарных преобразований — двумерных вращений координат. Каждое вращение происходит в плоскости, проходящей через какие-то две оси координат. При таком вращении преобразуются только элементы двух столбцов и двух одноименных строк матрицы A. Поэтому объем вычислений при одном повороте мал. Углы поворотов выбирают так, чтобы каждый поворот аннулировал определенный элемент матрицы A, лежащий ниже поддиагонали. Последовательность поворотов выбирают так, чтобы ранее аннулированный элемент оставался нулевым. После $(N-1)(N-2)/2$ поворотов аннулируются все элементы, лежащие ниже поддиагонали. Поэтому произвольная матрица A приводится к верхней почти треугольной форме.

Преобразования вращения сохраняют эрмитовость. Следовательно, если матрица A была эрмитовой, то цепочка преобразований Гивенса приводит матрицу к трехдиагональной форме.

По конечному результату метод Гивенса похож на метод Хаусхолдера. Он столь же устойчив, но в полтора раза медленнее ($5N^3$ арифметических действий для произвольной матрицы и $2N^3$ — для эрмитовой). В нем также поддиагональные элементы итоговой матрицы оказываются весьма большими.

Метод Якоби. Метод, предназначенный только для эрмитовых матриц A, был предложен в XVIII в. и возрожден с появлением компьютеров. В нем используются такие же двумерные вращения, как и в методе Гивенса, но угол поворота выбирается иначе. Аннулируется достаточно большой по модулю внедиагональный элемент. Симметричный ему внедиагональный элемент также обращается в нуль в силу сохранения эрмитовости. При этом ранее аннулированный элемент может снова стать ненулевым. Бесконечная цепочка таких поворотов в пределе приводит эрмитову матрицу к диагональному виду. При этом сразу находятся собственные значения: они равны диагональным элементам итоговой матрицы.

Такая цепочка преобразований не может быть конечной, т. е. метод Якоби — итерационный процесс. Он сходится всегда, но сходится достаточно медленно, так что объем вычислений для получения высокой точности в нем примерно в 20 раз больше, чем в методе Хаусхолдера. Поэтому на практике он применяется редко. Интересен он лишь тем, что исключительно устойчив и сразу дает собственные значения.

Элементарные преобразования. Возьмем произвольную (неэрмитову) матрицу A, уже приведенную методом отражений к верхней почти треугольной форме. Ее можно привести к трехдиагональной форме цепочкой так называемых элементарных преобразований подобия с матрицами вида

$$M = \begin{bmatrix} E_q & 0 \\ 0 & M_q \end{bmatrix}, \quad M_q = \begin{bmatrix} 1 & -v_{q+2} & -v_{q+3} & \dots & -v_N \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix},$$

причем $M^{-1}(v) = M(-v)$. Величины v_n выбирают так, чтобы аннулировать все элементы соответствующей строки матрицы A, лежащие правее верхней кодиагонали. Последовательность матриц M берут так, чтобы аннулировать сначала первую строку, затем вторую и т. д. Трудоемкость элементарных преобразований меньше, чем для преобразования отражения.

Однако матрицы M неунитарны, так что устойчивость этого метода хуже. В литературе отмечено, что обычно устойчивость метода удовлетворительна, но наблюдаются также случаи потери точности и даже срывы расчета. При мощности современных компьютеров надежность важнее небольшого выигрыша в трудоемкости.

7.2. ВЫЧИСЛЕНИЕ СПЕКТРА

7.2.1. Частичная проблема

Пусть требуется найти только некоторые собственные значения матрицы. При этом предполагается, что нам известны удовлетворительные приближения к нужным собственным значениям, и нужно лишь найти эти значения с высокой точностью. Наиболее просто и эффективно это делается методом обратных итераций с переменным сдвигом.

Обратные итерации со сдвигом. Будем предполагать, что собственные векторы $\mathbf{x}_m$, $1 \leqslant m \leqslant N$, матрицы A, соответствующие собственным значениям λ_m, образуют базис, а искомое собственное значение λ_M является простым. Пусть нам известно некоторое хорошее приближение к искомому собственному значению: $\tilde{\lambda} \approx \lambda_M$. Выберем в качестве нулевого приближения некоторый вектор $\mathbf{x}^{(0)}$ и построим следующий итерационный процесс:

$$(A - \tilde{\lambda}E)\mathbf{x}^{(s)} = \mathbf{x}^{(s-1)}, \quad s = 1, 2, \ldots \tag{7.19}$$

Для нахождения $\mathbf{x}^{(s)}$ требуется решить систему линейных уравнений с матрицей $(A - \tilde{\lambda}E)$. Поскольку $\tilde{\lambda}$ не точно равно собственному значению A, то $\det(A - \tilde{\lambda}E) \neq 0$, и система (7.19) имеет единственное решение.

Исследуем сходимость этих итераций. Для этого разложим векторы $\mathbf{x}^{(s)}$ по базису $\{\mathbf{x}_m\}$:

$$\mathbf{x}^{(s)} = \sum_{m=1}^{N} \alpha_m^{(s)} \mathbf{x}_m. \tag{7.20}$$

Подставим разложение (7.20) и аналогичное разложение для $\mathbf{x}^{(s-1)}$ в (7.19). Учитывая, что $A\mathbf{x}_m = \lambda_m \mathbf{x}_m$, и приравнивая коэффициенты в правой и левой частях при одинаковых базисных векторах $\mathbf{x}_m$, получим

$$\alpha_m^{(s)} = \alpha_m^{(s-1)}/(\lambda_m - \tilde{\lambda}) = \ldots = \alpha_m^{(0)}/(\lambda_m - \tilde{\lambda})^s, \quad 1 \leqslant m \leqslant N. \tag{7.21}$$

Поскольку $\tilde{\lambda} \approx \lambda_M$, то M-й знаменатель в (7.21) мал, а знаменатели с $m \neq M$ не малы. Поэтому $\alpha_M^{(s)}$ — коэффициент разложения $\mathbf{x}^{(s)}$ при собственном векторе $\mathbf{x}_m$ — будет быстро увеличиваться от одной итерации к другой, а остальные $\alpha_m^{(s)}, m \neq M$,

останутся небольшими. Значит, при достаточном числе итераций остальными компонентами разложения можно пренебречь, и вектор $\mathbf{x}^{(s)}$ станет почти параллельным собственному вектору $\mathbf{x}_M$ (это называется сходимостью по *направлению*).

Но если остальными компонентами можно пренебречь, то векторы $\mathbf{x}^{(s)}$ и $\mathbf{x}^{(s-1)}$ также будут почти параллельны, а отношения их компонент будут близки к $(\lambda_M - \tilde{\lambda})^{-1}$. Тем самым из отношения этих компонент можно определить величину λ_M.

Сходимость обратных итераций со сдвигом является линейной, т. е. довольно медленной. Фактическая скорость сходимости сильно зависит от того, насколько близко к λ_M приближение $\tilde{\lambda}$. Желателен более быстрый процесс.

Переменный сдвиг. Скорость сходимости можно кардинально улучшить, используя так называемый переменный сдвиг. Для этого в итерации (7.19) вместо постоянного сдвига $\tilde{\lambda}$ подставляют только что найденное приближение к λ_M. Формулы принимают следующий вид:

$$(A - \lambda^{(s-1)}E)\mathbf{x}^{(s)} = \mathbf{x}^{(s-1)}, \quad s = 1, 2, \ldots \tag{7.22}$$

За нулевое приближение к λ_M принимается $\lambda^{(0)} = \tilde{\lambda}$. По-прежнему для нахождения $\mathbf{x}^{(s)}$ надо решать линейную систему. Но теперь требуется построить формулы для определения $\lambda^{(s)}$. Очевидно, покомпонентное деление векторов на соседних итерациях — не слишком удобная процедура.

Разумно определять $\lambda^{(s)}$ из условия, чтобы на сколько возможно точно удовлетворялось бы соотношение $(A - \lambda^{(s)}E)\mathbf{x}^{(s)} \approx$ ≈ 0: если бы оно удовлетворялось точно, то $\lambda^{(s)}, \mathbf{x}^{(s)}$ были бы в точности собственным значением и собственным вектором матрицы A.

Это можно формулировать как требование

$$\begin{gathered}\|(A - \lambda^{(s)}E)\mathbf{x}^{(s)}\|_{l_2}^2 \equiv ((A - \lambda^{(s)}E)\mathbf{x}^{(s)}, \\ (A - \lambda^{(s)}E)\mathbf{x}^{(s)}) = \min.\end{gathered} \tag{7.23}$$

Поскольку $\mathbf{x}^{(s)}$ уже определено из (7.22), то в (7.23) неизвестно только $\lambda^{(s)}$. Для его определения надо приравнять нулю производную по $\lambda^{(s)}$ от (7.23). Это дает

$$\lambda^{(s)} = (\mathbf{x}^{(s)}, A\mathbf{x}^{(s)})/(\mathbf{x}^{(s)}, \mathbf{x}^{(s)}).$$

Подставляя сюда $A\mathbf{x}^{(s)} = \lambda^{(s-1)}\mathbf{x}^{(s)} + \mathbf{x}^{(s-1)}$ из (7.22), получим

$$\lambda^{(s)} = \lambda^{(s-1)} + (\mathbf{x}^{(s)}, \mathbf{x}^{(s-1)})/(\mathbf{x}^{(s)}, \mathbf{x}^{(s)}). \tag{7.24}$$

Формулы (7.22) и (7.24) полностью определяют итерационный процесс. Эти формулы справедливы для произвольных матриц.

По мере сходимости итераций длины векторов $\mathbf{x}^{(s)}$ очень быстро возрастают. Порядки чисел при этом могут выйти за пределы представления чисел на компьютере. Поэтому полезно ***после*** вычисления $\lambda^{(s)}$ нормировать вектор $\mathbf{x}^{(s)}$ на единицу.

Кроме того, чем ближе $\lambda^{(s)}$ к собственному значению λ_M, тем хуже обусловленность линейной системы (7.22). Поэтому следует прекращать итерации раньше, чем установятся все значащие цифры $\lambda^{(s)}$.

Сходимость. Возьмем разложение вектора $\mathbf{x}^{(s)}$ по базису $\{\mathbf{x}_m\}$ (7.20). Подставив это разложение в (7.22), получим аналогично (7.21) скорость возрастания коэффициентов разложения:

$$\alpha_m^{(s)} = \alpha_m^{(s-1)}/(\lambda_m - \lambda^{(s-1)}) = \ldots = \alpha_m^{(0)}/\prod_{\sigma=0}^{s-1}(\lambda_m - \lambda^{(\sigma)}). \qquad (7.25)$$

Для компонент с $m \neq M$ все множители в знаменателях будут близки к $(\lambda_m - \lambda_M)$ и не будут малыми. Для M-й компоненты множители в знаменателе будут быстро уменьшаться, так что эта компонента будет возрастать гораздо быстрее, чем при постоянном сдвиге.

Можно показать, что если собственные векторы образуют базис (а это справедливо не только для эрмитовых, но и для многих других классов матриц), то сходимость метода будет квадратичной. Если базис собственных векторов ортогональный (например, как у эрмитовых матриц), то сходимость будет кубической.

Построим такие оценки. Представим вектор $\mathbf{x}^{(s)}$ как сумму быстро возрастающей и медленной компонент:

$$\begin{gathered}\mathbf{x}^{(s)} = \mathbf{y}^{(s)} + \mathbf{z}^{(s)}, \quad \mathbf{y}^{(s)} = \alpha_M^{(0)}\mathbf{x}_M/\prod_{\sigma=0}^{s-1}(\lambda_M - \lambda^{(\sigma)}),\\ \mathbf{z}^{(s)} = \sum_{m\neq M}\alpha_m^{(0)}\mathbf{x}_m/\prod_{\sigma=0}^{s-1}(\lambda_m - \lambda^{(\sigma)}),\\ |\mathbf{y}^{(s)}| \gg |\mathbf{z}^{(s)}| \quad \text{при } s > 1;\end{gathered} \qquad (7.26)$$

разложение для $\mathbf{x}^{(s-1)}$ аналогично. Тогда скалярные произведения в (7.24) можно представить в следующем виде (с учетом комплексности векторов):

$$
\begin{gathered}
(\mathbf{x}^{(s)}, \mathbf{x}^{(s-1)}) = a_{s,s-1} + b_{s,s-1} + c_{s,s-1}, \\
a_{s,s-1} = (\mathbf{y}^{(s)}, \mathbf{y}^{(s-1)}) \sim \\
\sim \prod_{\sigma=0}^{s-1} (\lambda_M^* - \lambda^{(\sigma)*})^{-1} \times \prod_{\sigma=0}^{s-2} (\lambda_M - \lambda^{(\sigma)})^{-1}, \\
b_{s,s-1} = (\mathbf{y}^{(s)}, \mathbf{z}^{(s-1)}) + (\mathbf{z}^{(s)}, \mathbf{y}^{(s-1)}) \sim \\
\sim \prod_{\sigma=0}^{s-1} (\lambda_M^* - \lambda^{(\sigma)*})^{-1} + \prod_{\sigma=0}^{s-2} (\lambda_M - \lambda^{(\sigma)})^{-1}, \\
c_{s,s-1} = (\mathbf{z}^{(s)}, \mathbf{z}^{(s-1)}) \sim 1;
\end{gathered}
\tag{7.27}
$$

выражение для $(\mathbf{x}^{(s)}, \mathbf{x}^{(s)})$ записывается аналогично. Очевидно, $|a_{s,s-1}| \gg |b_{s,s-1}| \gg |c_{s,s-1}|$. Вычитая обе части равенства (7.24) из λ_M и подставляя туда (7.27), выделим главные члены:

$$
\begin{gathered}
\lambda_M - \lambda^{(s)} = \lambda_M - \lambda^{(s-1)} - \\
- \frac{a_{s,s-1}}{a_{ss}} \left(1 + \frac{b_{s,s-1} + c_{s,s-1}}{a_{s,s-1}}\right) \Big/ \left(1 + \frac{b_{ss} + c_{ss}}{a_{ss}}\right) = \\
= \left(\lambda_M - \lambda^{(s-1)}\right) \left[1 - \left(1 + \frac{b_{s,s-1} + c_{s,s-1}}{a_{s,s-1}}\right) \Big/ \left(1 + \frac{b_{ss} + c_{ss}}{a_{ss}}\right)\right],
\end{gathered}
\tag{7.28}
$$

поскольку $a_{s,s-1}/a_{ss} = \lambda_M - \lambda^{(s-1)}$.

Если собственные векторы ортогональны, то $(\mathbf{y}, \mathbf{x}) = 0$ и $b_{s,s-1} = b_{ss} = 0$. Тогда выражение в квадратных скобках в (7.28) примерно равно $c_{s,s-1}/a_{s,s-1}$, а отношение c_{ss}/a_{ss} уступает ему по величине на множитель $\lambda_M - \lambda^{(s)}$. Подставив в (7.28) искомое отношение из (7.27), получим

$$
\lambda_M - \lambda^{(s)} \approx \text{const} \prod_{\sigma=0}^{s-1} |\lambda_M - \lambda^{(\sigma)}|^2. \tag{7.29}
$$

Ищем решение этого уравнения в виде $|\lambda_M - \lambda^{(s)}| = \text{const}\, |\lambda_M - \lambda^{(s-1)}|^q$ или $|\lambda_M - \lambda^{(s-1)}| = \text{const}\, |\lambda_M - \lambda^{(s)}|^{1/q}$. Применяя последнее соотношение рекуррентно к правой части (7.29) и считая s не слишком малым, получим связь показателей степени:

$$
1 \approx 2(q^{-1} + q^{-2} + q^{-3} + \ldots) \approx 2/(q-1), \tag{7.30}
$$

откуда $q = 3$ и сходимость кубическая:

$$
|\lambda_M - \lambda^{(s)}| \approx \text{const}\, |\lambda_M - \lambda^{(s-1)}|^3.
$$

Если собственные векторы неортогональны, то $(\mathbf{y}, \mathbf{z}) \neq 0$, и отношение $b_{s,s-1}/a_{s,s-1}$ оказывается главным в (7.28), несколько уступает ему b_{ss}/a_{ss}, а оба отношения c/a пренебрежимо малы. Тогда для сходимости $|\lambda_M - \lambda^{(s)}|$ получается выражение вида (7.29), но в правой части вместо квадратов стоят первые степени. Соответственно, для q вместо (7.30) получаем $1 \approx q^{-1} + q^{-2} + q^{-3} + \ldots \approx 1/(q-1)$; это дает $q = 2$, т. е. квадратичную сходимость:

$$|\lambda_M - \lambda^{(s)}| \approx \text{const}\, |\lambda_M - \lambda^{(s-1)}|^2.$$

Когда искомое собственное значение λ_M кратное, тогда $\lambda^{(s)}$ по-прежнему сходятся к нему. Если у этого собственного значения несколько собственных векторов, то векторы $\mathbf{x}^{(s)}$ сходятся по направлению к одному из них (к какому именно — зависит от выбора $\mathbf{x}^{(0)}$). Чтобы получить другие собственные векторы, соответствующие кратному λ_M, надо выбрать другие $\mathbf{x}^{(0)}$.

Практика расчетов показывает, что обратные итерации с переменным сдвигом сходятся даже в том случае, когда собственные векторы матрицы не образуют базиса. Например, для жордановой матрицы имеется только один собственный вектор. При любом выборе $\mathbf{x}^{(0)}$ величины $\lambda^{(s)}$ сходятся к единственному собственному значению, а $\mathbf{x}^{(s)}$ — к единственному собственному вектору. Однако в этом случае длина $\mathbf{x}^{(s)}$ очень сильно возрастает от одной итерации к другой, и следует принимать меры, чтобы избежать переполнения.

В методе обратных итераций со сдвигом хороший выбор $\lambda^{(0)}$ является существенным. Итерации сходятся к тому из собственных значений λ_M, для которого $|\lambda_M - \lambda^{(0)}|$ минимален. Если $\lambda^{(0)}$ лежит вблизи границы раздела областей притяжения соседних собственных значений, то первые итерации сходятся медленно. Выбор же $\mathbf{x}^{(0)}$ почти не влияет на скорость сходимости. Только одного случая следует опасаться: если случайно коэффициент разложения $\alpha_M^{(0)}$ оказался нулем. Во избежание этого можно в качестве компонент вектора $\mathbf{x}^{(0)}$ брать случайные числа.

Метод обратных итераций можно непосредственно применять к произвольной матрице A, однако решение линейной системы (7.22) методом Гаусса требует $2N^3/3$ арифметических действий на каждой итерации. Гораздо выгоднее сначала преобразовать матрицу A методом Хаусхолдера к верхней почти треугольной форме (для эрмитовых матриц — к трехдиагональ-

ной). Тогда каждая итерация будет требовать $\approx 2N^2$ (для эрмитовых $\sim 10N$) арифметических действий.

7.2.2. Обобщенная проблема

Некоторым усложнением исходной задачи является так называемая обобщенная проблема собственных значений. Требуется найти нетривиальные решения задачи

$$B\mathbf{x} = \lambda C\mathbf{x}, \tag{7.31}$$

где B и C — квадратные матрицы порядка N, причем матрица C — неособенная (т. е. $\det C \neq 0$).

Умножая на C^{-1} слева, приводим задачу (7.31) к обычной проблеме собственных значений (7.1) с матрицей $A = C^{-1}B$. Для нее можно написать обычный процесс обратных итераций с переменным сдвигом (7.22); он принимает вид

$$(C^{-1}B - \lambda^{(s-1)}E)\mathbf{x}^{(s)} = \mathbf{x}^{(s-1)}. \tag{7.32}$$

Так можно поступать для произвольных матриц B и C.

Однако часто матрицы B и C уже имеют какую-то удобную структуру: являются почти треугольными или ленточными (в частности трехдиагональными). Поэтому удобно умножить (7.23) слева на C и записать итерации в следующей форме:

$$(B - \lambda^{(s-1)}C)\mathbf{x}^{(s)} = C\mathbf{x}^{(s-1)}. \tag{7.33}$$

Формула (7.33) эквивалентна (7.32), однако при отмеченных специальных структурах матриц вычисления по ней удобнее и дешевле.

Формулу для расчета $\lambda^{(s)}$ получаем аналогично тому, как это делалось для простой проблемы. Однако здесь возможны варианты. Если потребовать $\|(C^{-1}B - \lambda^{(s)}E)\mathbf{x}^{(s)}\| = \min$, то получим

$$\lambda^{(s)} = \lambda^{(s-1)} + (\mathbf{x}^{(s)}, \mathbf{x}^{(s-1)})/(\mathbf{x}^{(s)}, \mathbf{x}^{(s)}). \tag{7.34}$$

Эта формула совпадает с (7.24). Если же потребовать $\|(B - \lambda^{(s)}C)\mathbf{x}^{(s)}\| = \min$, то получается несколько другая формула:

$$\lambda^{(s)} = \lambda^{(s-1)} + (C\mathbf{x}^{(s)}, C\mathbf{x}^{(s-1)})/(C\mathbf{x}^{(s)}, C\mathbf{x}^{(s)}). \tag{7.35}$$

Любую из этих формул можно использовать в расчетах. Скорость сходимости обоих вариантов примерно одинакова и такова же, как для простой проблемы.

7.2.3. Полная проблема

Для нахождения всех собственных значений матрицы, т. е. для решения полной проблемы собственных значений, сейчас наиболее часто употребляют итерационный QR-алгоритм (предложен В. Н. Кублановской и Френсисом в 1961 г.). Алгоритм основан на подобном преобразовании матрицы к так называемой квазитреугольной форме (рис. 7.3). В этой форме каждому простому собственному значению соответствует матричный элемент на диагонали, а каждому k-кратному собственному значению квадратная подматрица («ящик») порядка k, стоящая на главной диагонали. Кратные собственные значения являются собственными значениями «ящиков». Верхняя половина матрицы заполнена, а нижняя, за исключением ящиков, не заполнена.

На практике порядки ящиков невелики, поэтому их собственные значения находятся легко. Например, можно применить метод обратных итераций с переменным сдвигом. Поскольку у «ящика» собственное значение одно, обратные итерации сходятся при любом нулевом приближении.

Итерации QR-алгоритма устойчивы и всегда сходятся. Скорость сходимости линейная со знаменателем $\mu = \max |\lambda_m/\lambda_{m+1}|$, где λ_m — собственные значения, расположенные в порядке воз-

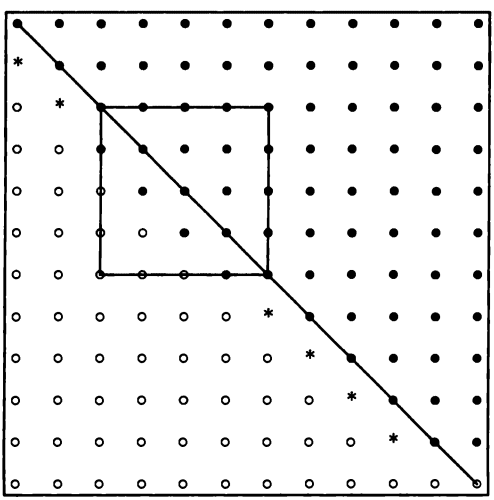


Рис. 7.3. QR-алгоритм (точки — ненулевые элементы квазитреугольной матрицы с одним ящиком пятого порядка; кружки — элементы, предварительно уничтожаемые методом отражений; звездочки — элементы, аннулируемые в ходе QR-итераций)

растания модулей (причем кратные значения считаются за одно). Поэтому при наличии близких по модулю собственных значений число итераций оказывается довольно большим. Однако этот метод является очень выгодным, если матрица уже приведена к верхней почти треугольной форме методом отражений. В ходе QR-итераций структура таких матриц не разрушается, и благодаря этому одна итерация требует всего $6N^2$ арифметических действий. Детали этого алгоритма хорошо отработаны, и существуют основанные на нем стандартные программы. Они надежно работают, и при 64-разрядных вычислениях допускают использование $N \sim 1\,000$.

Глава 8

ЗАДАЧИ МИНИМИЗАЦИИ

8.1. ОДНОМЕРНЫЙ МИНИМУМ

8.1.1. Золотое сечение

В методе наименьших квадратов и других аналогичных методах уже возникала задача минимизации. Имеется большое количество прикладных задач оптимизации: при моделировании производственных процессов, в экономике и др. Многие из этих задач можно сводить к решению систем уравнений. Однако у них есть свои особенности, поэтому для них разработан ряд специфических методов. Один из них — метод золотого сечения.

Рассмотрим скалярную функцию одного переменного $\Phi(x)$, заданную на конечном отрезке $[a, b]$. Наложим на нее очень слабые требования: будем считать ее кусочно-непрерывной и ограниченной. Будем искать минимум этой функции на заданном отрезке, то есть решать следующую задачу:

$$\Phi(x) = \min, \quad x \in [a, b]. \tag{8.1}$$

При заданных требованиях к $\Phi(x)$ она имеет по меньшей мере один минимум на $[a, b]$; может иметь несколько локальных минимумов. В этом случае рассматривают задачу отыскания одного или нескольких локальных минимумов, а также задачу нахождения глобального минимума.

Построим метод нахождения какого-то локального минимума, являющийся аналогом метода дихотомии. Это итерационный метод. Опишем одну итерацию. Обозначим концы отрезка через $x_0 = a$ и $x_1 = b$. Поставим внутри отрезка $[x_0, x_1]$ две точки $x_2 < x_3$ (рис. 8.1). Эти точки разбивают отрезок на три части. Вычислим значения $\Phi(x_2)$ и $\Phi(x_3)$, и сравним их между собой (значения функции на концах отрезка вычислять не надо!). Пусть для определенности $\Phi(x_2) < \Phi(x_3)$. Тогда из рис. 8.1 видно, что какой-то локальный минимум должен лежать на одной

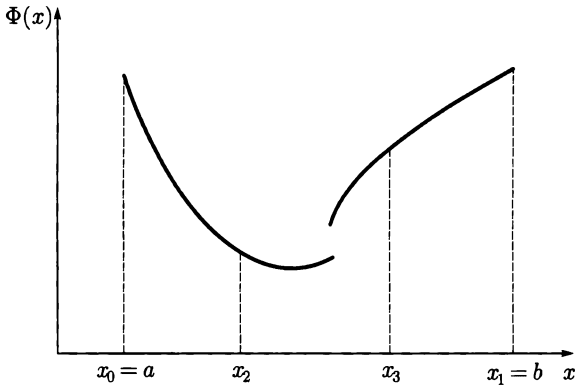


Рис. 8.1. Метод золотого сечения

из частей отрезка, примыкающих к x_2. Поэтому внешнюю часть $[x_3, x_1]$ можно отбросить, продолжив поиск на меньшем отрезке $[x_0, x_3]$.

На следующей итерации надо взять оставшийся отрезок $[x_0, x_3]$, поставить на нем две точки и повторить процесс. Но одна точка x_2 на нем уже стоит, и значение функции в ней вычислено. Поэтому достаточно поставить лишь одну дополнительную точку; тем самым на каждой серийной итерации будет вычисляться лишь одно значение функции.

Очевидно, на каждой итерации выгодно ставить точки симметрично. Для этого надо, чтобы каждое новое деление отрезка было бы подобно предыдущему. Примем длину отрезка $[x_0, x_1]$ за единицу: $x_1 - x_0 = 1$. Введем обозначение $\xi = x_2 - x_0$. В силу симметрии также будет $\xi = x_1 - x_3$, а $x_3 - x_2 = 1 - 2\xi$. Поэтому деление исходного отрезка можно записать так:

$$1 = (\xi) + (1 - 2\xi) + (\xi).$$

На следующей итерации отрезок имеет длину $x_3 - x_0 = 1 - \xi$. Теперь точка x_2 стала правой внутренней точкой, а отрезок $[x_2, x_3]$ — внешним правым. Отсюда легко найти долевое разбиение нового отрезка:

$$1 - \xi = (1 - 2\xi) + (3\xi - 1) + (1 - 2\xi).$$

Составим из подобных отрезков пропорцию. Например, для полных отрезков и их крайних отрезков это будет

$$1/(1 - \xi) = \xi/(1 - 2\xi),$$

что дает квадратное уравнение $\xi^2 - 3\xi + 1 = 0$. Нам нужен тот его корень, который принадлежит отрезку $[0; 0{,}5]$:

$$\xi = (3 - \sqrt{5})/2 \approx 0{,}3820. \tag{8.2}$$

В такой пропорции надо делить отрезок, и проводить итерации до тех пор, пока длина окончательного отрезка не станет меньше заданной точности ε.

За одну итерацию длина отрезка умножается на $q = 1 - \xi \approx 0{,}6180$. Это означает, что метод сходится линейно со знаменателем q. Поскольку $\lg q = -0{,}2090$, то за пять итераций длина отрезка сокращается примерно в 10 раз. Для получения восьми верных десятичных знаков требуется примерно 40 итераций. Заметим, что для достижения высокой точности необходимо задавать q с максимально доступным числом верных знаков.

Сходимость довольно медленная, но при скорости современных персональных компьютеров это не так уж важно. Достоинства метода заключаются в следующем: 1) минимальные требования к функции; 2) низкая стоимость и простота реализации итераций; 3) гарантированная сходимость к одному из локальных минимумов (в том числе, если точкой минимума является один из концов отрезка); 4) гораздо меньшая, чем у более быстрых методов, чувствительность к погрешностям вычисления функции. Поэтому метод золотого сечения часто применяют в практических вычислениях (особенно в тех случаях, когда погрешность вычисления функции существенна).

Погрешность. Пусть значения функции заданы с погрешностями $\pm\delta$. Оценим, на какую точность ε нахождения местоположения точки минимума можно при этом рассчитывать. Если существует $f''(x)$, то отклонению на ε от точки минимума $x_{\min}$ соответствует изменение функции на $0{,}5f''(x_{\min})\varepsilon^2$. Приравняв это изменение величине погрешности функции δ, получим оценку предельно достижимой точности:

$$\varepsilon \approx \sqrt{2\delta / f''(x_{\min})}. \tag{8.3}$$

Оценка справедлива для невырожденного минимума $f''(x_{\min}) \neq 0$. Пусть погрешность значения функции предельно мала — на уровне погрешности округления 64-разрядного компьютера $\delta \sim 10^{-16}$. Получается относительно невысокая точность нахождения минимума $\varepsilon \sim 10^{-8}$. Для вырожденного минимума с $f''(x_{\min}) = 0$, $f^{(IV)}(x_{\min}) > 0$ оценка будет еще более неблагоприятной.

Бесполезно добиваться точности выше (8.3). Итерационный процесс собьется, так как невозможно будет правильно определять соотношения значений функции в сравниваемых точках. Таким образом, метод золотого сечения позволяет вычислить точку минимума лишь с не очень высокой точностью.

8.1.2. Метод Ньютона

Если $\Phi(x)$ удовлетворяет более жестким требованиям, то для нее можно построить быстро сходящиеся итерационные процессы. Пусть функция $\Phi(x)$ непрерывна вместе с первой и второй производными. Тогда точки минимума функции удовлетворяют уравнению $\Phi'(x) = 0$. Но этому же уравнению удовлетворяют точки максимума и точки перегиба с горизонтальной касательной. Поэтому задачу нахождения минимума можно сформулировать следующим образом:

$$\Phi'(x) = 0 \quad \text{при} \quad \Phi''(x) > 0. \tag{8.4}$$

Уравнение (8.4) можно решать методом Ньютона (см. подразд. 2.2.2, где всюду $f(x)$ нужно заменить на $\Phi'(x)$). Приведем соответствующие формулы:

$$x^{(s+1)} = x^{(s)} - \Phi'(x^{(s)})/\Phi''(x^{(s)}). \tag{8.5}$$

Вблизи корня надо дополнительно проверять знак $\Phi''(x)$. Если $\Phi''(x) > 0$, то итерации сходятся к минимуму. В противном случае итерации можно прекращать, не добиваясь сходимости. На первых итерациях вдали от корня такая проверка не нужна.

Вблизи корня полезно проверять его кратность (см. подразд. 2.2.2). Если корень простой, то минимум невырожденный. Вырожденному минимуму соответствуют корни нечетной кратности $(3, 5, \ldots)$. Однако заметим, что наличие кратных корней возможно лишь при существовании у $\Phi(x)$ непрерывных производных еще более высокого порядка.

В методе Ньютона нет сравнения значений функции $\Phi(x)$. Вместо этого используются точно вычисленные производные, поэтому ошибки округления самой функции $\Phi(x)$ не опасны. Метод Ньютона, в отличие от метода золотого сечения, позволяет вычислить точку минимума с высокой точностью (на уровне ошибок округления компьютера).

Обобщенный метод. Напомним, что сходимость метода Ньютона очень чувствительна к выбору нулевого приближения.

Ослабляет эту зависимость переход к обобщенному методу Ньютона (см. подразд. 2.2.3). Его формулы для данной задачи принимают следующий вид:

- находим ньютоновское приращение аргумента:

$$\Delta^{(s)} = \Phi'(x^{(s)})/\Phi''(x^{(s)});$$

- вычисляем невязку исходной итерации

$$\varphi(0) = [\Phi'(x^{(s)})]^2$$

и ньютоновский предиктор невязки

$$\varphi(1) = [\Phi'(x^{(s)}) - \Delta^{(s)}]^2;$$

- находим корректирующий шаг

$$\tau^{(s)} = \frac{\varphi(0) + \theta\varphi(1)}{\varphi(0) + \varphi(1)}, \quad 0 < \theta \leqslant 1,$$

где θ — настроечный параметр метода (стратегия его выбора изложена в подразд. 2.2.3);

- проводим окончательный расчет новой итерации по формуле

$$x^{(s+1)} = x^{(s)} - \tau^{(s)}\Delta^{(s)}. \tag{8.6}$$

Если положить $\theta = 1$, то эти формулы переходят в классический метод Ньютона, поэтому можно сразу составлять программу для обобщенного метода Ньютона.

Разностные производные. Лучше всего вычислять $\Phi'(x)$ и $\Phi''(x)$ аналитически. Это страхует от ошибок округления компьютера. Однако использовать аналитические выражения не всегда удается. В этом случае заменяют производные какими-либо разностными аппроксимациями. Наиболее просто и точно ввести вспомогательный шаг h и воспользоваться симметричными аппроксимациями точности $O(h^2)$:

$$\begin{aligned} \Phi'(x^{(s)}) &\approx [\Phi(x^{(s)} + h) - \Phi(x^{(s)} - h)]/(2h); \\ \Phi''(x^{(s)}) &\approx [\Phi(x^{(s)} + h) - 2\Phi(x^{(s)}) + \Phi(x^{(s)} - h)]/h^2. \end{aligned} \tag{8.7}$$

Переход к разностным производным существенно увеличивает влияние ошибок округления по сравнению с использованием аналитических выражений. Рассмотрим, как целесообразно выбрать шаг h для получения наилучших результатов. В подразд. 6.3.1 давались оценки для оптимального шага h_0 при численном дифференцировании:

$$h_0 \sim \delta^{1/(p+q)},$$

где δ — погрешность вычисления функции; p — порядок точности аппроксимирующей производную формулы; q — порядок производной.

При этом достигается точность вычисления производной $\sim \delta^{p/(p+q)}$. Для формул (8.7) $q = 1$ и $q = 2$ соответственно, а $p = 2$. Считая основным источником ошибок компьютерное округление $\delta \sim 10^{-16}$, получаем $h_0 \sim 10^{-5}$ для вычисления первой производной и $h_0 \sim 10^{-4}$ для второй.

Казалось бы, вторая производная заставляет выбирать более крупный шаг. Однако наиболее важна точность вычисления $\Phi'(x)$, так как именно условие $\Phi'(x) = 0$ дает нам точку минимума. Ошибка же в вычислении $\Phi''(x)$ лишь несколько снизит скорость сходимости ньютоновских итераций. Поэтому целесообразно брать $h \sim 10^{-5}$ для вычисления обеих производных. Погрешность вычисления $\Phi'(x)$ при этом составляет $\sim 10^{-10} \div 10^{-11}$, что приводит к такой же ошибке вычисления точки минимума $\varepsilon \sim 10^{-10} \div 10^{-11}$. Такая погрешность существенно больше, чем при использовании аналитических выражений, но много лучше, чем в методе золотого сечения.

Еще более высокой точности можно добиться, применяя процедуру вычисления производных на сгущающихся сетках, описанную в конце подразд. 6.3.1.

Критерий сходимости. Если производные вычисляются аналитически, то итерации вблизи невырожденного минимума можно останавливать по критерию

$$|x^{(s)} - x^{(s-1)}| < \varepsilon. \tag{8.8}$$

Сходимость классического и обобщенного методов Ньютона квадратичная, поэтому при выполнении критерия отличие $x^{(s)}$ от точки экстремума будет много меньше ε. Таким же критерием можно пользоваться и в случае разностного вычисления производных, однако при этом надо дополнительно следить, чтобы разностные производные вычислялись с достаточной точностью.

8.1.3. Случай многих экстремумов

Функция $\Phi(x)$ на отрезке $[a, b]$ может иметь много локальных минимумов. Полное исследование задачи в этом случае провести практически невозможно. Обычно используют следующую процедуру. На отрезке $[a, b]$ выбирают в качестве начальных приближений $10 - 100$ случайных точек (с помощью псевдослучайных

чисел, см. подразд. 3.3.3). Из каждого начального приближения ищут ближайший экстремум обобщенным методом Ньютона. Затем сравнивают найденные экстремумы между собой. Наименьшее значение принимают за глобальный минимум. Дополнив эту процедуру построением графика $\Phi(x)$, можно получить достаточно достоверную картину экстремумов функции.

8.2. МНОГОМЕРНЫЙ МИНИМУМ

8.2.1. Рельеф функции

Пусть $\Phi(x)$ есть скалярная достаточно гладкая функция от N-мерного вектора $\mathbf{x} = \{x_n, 1 \leqslant n \leqslant N\}$, т. е. функция от многих скалярных переменных. Аналогично одномерному случаю формулируется задача поиска минимума в многомерной области G:

$$\Phi(\mathbf{x}) = \min, \quad \mathbf{x} \in G. \tag{8.9}$$

Геометрически эту задачу можно интерпретировать следующим образом. Возьмем $(N+1)$-мерное пространство с координатами $x_1, x_2, \ldots, x_N, \Phi$. Тогда $\Phi(x)$ есть гиперповерхность в этом пространстве. Решение задачи (8.9) означает низшую точку этой гиперповерхности в пределах области G.

Качественное поведение гиперповерхности характеризуется ее рельефом. Нетрудно пояснить рельеф на примере функции двух переменных. В этом случае гиперповерхность есть просто поверхность в трехмерном пространстве (x, y, Φ) (рис. 8.2). Проведем равноотстоящие плоскости, параллельные плоскости (x, y). Они пересекут поверхность $\Phi(x, y)$ по некоторым линиям; их называют линиями уровня. Проекции этих линий на плоскость (x, y) дают карту рельефа поверхности. Эта картина аналогична изображению рельефа местности на топографических картах. На картах направления убывания функции обозначают штрихом около линии уровня.

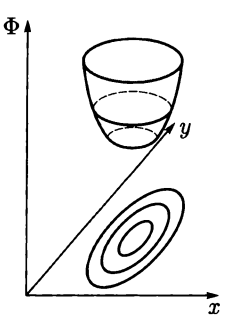


Рис. 8.2. Поверхность и ее изолинии уровня

Рассмотрим некоторые типичные формы рельефа. Простейшей формой является котловина (рис. 8.3, *а*). Например, возьмем квадратичную форму

$$\Phi(x, y) = (x - x_0)^2/a^2 + (y - y_0)^2/b^2;$$

линии уровня этой функции — эллипсы. В этой котловине есть единственный мини-

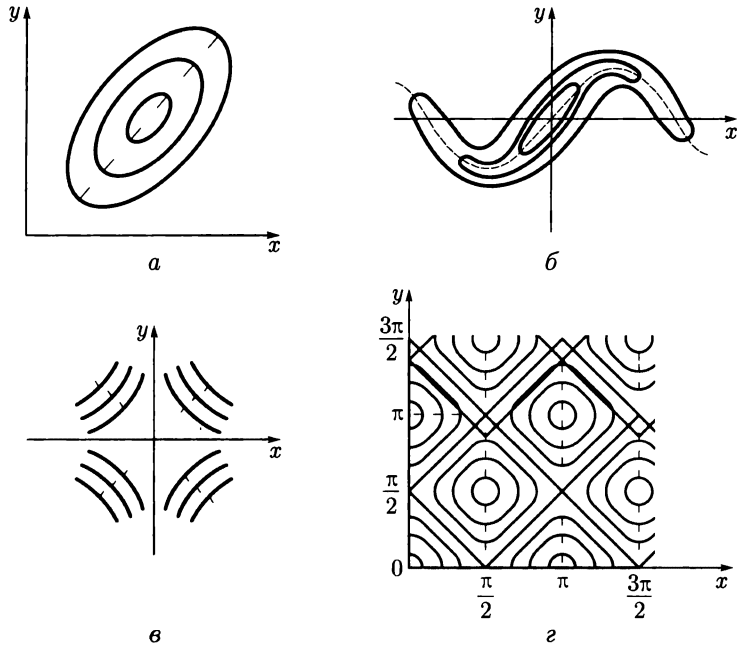


Рис. 8.3. Изолинии уровня для разных видов рельефа:

а — котловинный; *б* — овражный; *в* — седловинный; *г* — многоэкстремальный

мум. На таком рельефе пролитая вода соберется в точку минимума. Если $a \sim b$, то эллипсы слабо вытянуты. Этот случай наиболее легок для численных методов. Если же a и b очень сильно отличаются, то котловина сильно вытянута. Алгоритмы при этом гораздо медленнее работают, особенно при большом числе измерений.

Еще более сложен овраг, похожий на вытянутую котловину с изгибающимся дном (рис. 8.3, *б*). Например, рассмотрим функцию

$$\Phi(x, y) = 100(y - \sin x)^2 + 0{,}01x^2.$$

Пролитая на такой рельеф вода быстро стекает на дно оврага, а потом медленно течет к минимуму по извилистому дну. Примерно так же ведут себя траектории многих итерационных методов.

Если изменить знак у написанных выше функций Φ на противоположный, то направления наклонов изменятся, и минимумы станут максимумами. Это не вносит принципиально нового.

Если взять $\Phi(x, y) = (x - x_0)^2/a^2 - (y - y_0)^2/b^2$, то получится седловинный рельеф (рис. 8.3, *в*). По одним направлениям от седловой точки функция возрастает, по другим убывает.

Наконец, многоэкстремальный рельеф (рис. 8.3, *г*) содержит ряд минимумов, максимумов и седловых точек.

8.2.2. Обобщенный метод Ньютона

Метод золотого сечения не обобщается на многомерный случай. Наиболее работоспособным методом поиска локального минимума многомерных функций достаточно общего вида является обобщенный метод Ньютона (см. подразд. 2.3.2). Задача минимизации при этом заменяется системой нелинейных уравнений $\partial\Phi/\partial\mathbf{x} = 0$. Тогда формулы метода примут следующий вид:

- на s-й итерации решается система линейных уравнений для определения приращений $\Delta^{(s)}$ ньютоновского метода:

$$\frac{\partial^2\Phi(\mathbf{x}^{(s)})}{\partial\mathbf{x}^2}\Delta^{(s)} = \frac{\partial\Phi(\mathbf{x}^{(s)})}{\partial\mathbf{x}}. \tag{8.10}$$

Здесь первая производная скаляра по вектору есть вектор $\partial\mathbf{\Phi}/\partial\mathbf{x} = (\partial\mathbf{\Phi}/\partial x_n, 1 \leqslant n \leqslant N)^T$, а вторая производная скаляра по вектору есть матрица $\partial^2\Phi/\partial\mathbf{x}^2 = (\partial^2\Phi/\partial x_n\partial x_m, 1 \leqslant n \leqslant N, 1 \leqslant m \leqslant N)$;

- вычисляется невязка исходной итерации

$$\varphi(0) = \left[\frac{\partial\Phi(\mathbf{x}^{(s)})}{\partial\mathbf{x}}\right]^2 \equiv \sum_{n=1}^{N}\left[\frac{\partial\Phi(\mathbf{x}^{(s)})}{\partial x_n}\right]^2,$$

ньютоновский предиктор $\mathbf{x}^{(s)} - \mathbf{\Delta}^{(s)}$ и невязка ньютоновского предиктора

$$\varphi(1) = \left[\frac{\partial\Phi(\mathbf{x}^{(s)} - \mathbf{\Delta}^{(s)})}{\partial\mathbf{x}}\right]^2 \equiv \sum_{n=1}^{N}\left[\frac{\partial\Phi(\mathbf{x}^{(s)} - \mathbf{\Delta}^{(s)})}{\partial x_n}\right]^2;$$

- по соотношению невязок определяется корректирующий шаг:

$$\tau^{(s)} = \frac{\varphi(0) + \theta\varphi(1)}{\varphi(0) + \varphi(1)}, \quad 0 < \theta \leqslant 1,$$

где θ — настроечный параметр метода;

- проводится окончательный расчет новой итерации по формуле

$$\mathbf{x}^{(s+1)} = \mathbf{x}^{(s)} - \tau^{(s)}\mathbf{\Delta}^{(s)}. \tag{8.11}$$

Если положить $\theta = 1$, то эти формулы переходят в классический метод Ньютона.

Классический метод Ньютона хорошо работает на котловинах, дно которых почти неизогнуто, т. е. линии уровня близки к эллипсам: в этом случае уравнения системы $\partial\Phi/\partial\mathbf{x} = 0$ близки к линейным, даже если котловина сильно вытянутая.

Если дно котловины сильно изогнуто, т. е. это овраг, то траектория классического метода Ньютона трудно предсказуема. Очередное приращение $\Delta^{(s)}$ может оказаться очень большим, проскочить ближайший изгиб оврага и выбросить траекторию за пределы оврага.

Обобщенный метод Ньютона на котловинах ведет себя так же хорошо, как и классический метод. На оврагах его траектория быстро приводит на дно оврага. Далее приращения $\Delta^{(s)}$ довольно сильно уменьшаются, и процесс требует очень большого числа шагов. Траектория уверенно идет вблизи дна оврага в направлении точки минимума. Поэтому в многомерном случае обобщенный метод Ньютона еще более выгоден, чем в одномерном.

Отличие минимумов от максимумов естественно производится в ходе расчета. Если видно, что итерации $\mathbf{x}^{(s)}$ сходятся к некоторому пределу, а значения $\Phi(\mathbf{x}^{(s)})$ при этом монотонно уменьшаются, то процесс сходится к минимуму функции.

Разностные производные. В методе Ньютона и его обобщении есть две процедуры, трудоемкие при больших N: вычисление матрицы вторых производных (гессиана) и решение системы линейных уравнений.

В некоторых учебниках для уменьшения трудоемкости предлагается замораживать матрицу $\partial^2\Phi/\partial\mathbf{x}^2$ хотя бы на нескольких соседних итерациях; этого категорически не рекомендуется делать, ибо сходимость метода может резко ухудшиться. Очень желательно также пользоваться точными аналитическими выражениями для вычисления первой и второй производных: без этого невозможно обеспечить высокую точность нахождения точки минимума. Однако аналитические выражения оказываются очень громоздкими, а само их написание и программирование занимает много времени и сопряжено с ошибками. В этой ситуации прибегают к разностному вычислению производных, аналогично одномерному случаю (см. подразд. 8.1.2).

Однако в многомерном случае требования к точности и надежности вычисления производных существенно выше, чем в одномерном случае: придется решать систему линейных уравнений, а погрешность матричных элементов при больших размерностях может заметно сказаться на точности результата. Поэто-

му целесообразно использовать процедуру со сгущениями сетки, описанную в конце подразд. 6.3.1, несмотря на ее значительно бóльшую трудоемкость.

Формулы для первых и вторых производных по каждому аргументу приведены в подразд. 8.1.2, но в матрице вторых производных требуются еще и смешанные вторые производные. Приведем базовую формулу для простейшего случая функции двух переменных $\Phi(x, y)$. Выберем разные шаги h_x, h_y по разным переменным и запишем

$$\frac{\partial^2\Phi}{\partial x\partial y} = \frac{\partial}{\partial x}\left(\frac{\partial\Phi}{\partial y}\right) \approx \frac{\partial}{\partial x}\left[\frac{\Phi(x, y+h_y) - \Phi(x, y-h_y)}{2h_y}\right] \approx$$
$$\approx \frac{\Phi(x+h_x, y+h_y) - \Phi(x+h_x, y-h_y) - \Phi(x-h_x, y+h_y) + \Phi(x-h_x, y-h_y)}{4h_xh_y}. \tag{8.12}$$

Если для повышения точности применяется процедура сгущения сеток, то оба шага в (8.12) надо уменьшать одновременно в одинаковое число раз.

Критерий сходимости. Если производные вычисляются аналитически, то итерации вблизи невырожденного минимума можно останавливать по критерию

$$|\mathbf{x}^{(s)} - \mathbf{x}^{(s-1)}| < \varepsilon. \tag{8.13}$$

Сходимость классического и обобщенного методов Ньютона квадратичная, поэтому при выполнении критерия отличие $\mathbf{x}^{(s)}$ от точки экстремума будет много меньше ε. Таким же критерием можно пользоваться и в случае разностного вычисления производных. Однако при этом надо дополнительно следить, чтобы разностные производные вычислялись с достаточной точностью.

8.2.3. Многоэкстремальность

Если функция имеет много экстремумов, то вопрос выбора нулевого приближения для их отыскания намного менее ясен, чем в одномерном случае. Пусть исходная N-мерная область имеет объем V и эффективный диаметр $D \sim V^{1/N}$. Если выбрать в ней M случайных точек в качестве нулевого приближения, то в среднем на каждую точку придется объем V/M с эффективным диаметром $d \sim (V/M)^{1/N} \sim D/M^{1/N}$. Когда размерность задачи велика $N \sim 100$, то даже при огромном числе $M \sim 10^{10}$ мы получим $d \approx 10^{-0,1}D \approx 0{,}8D$, т. е. расстояние между случайными точками будет почти равно диаметру области! Вероятность того, что хоть одна случайная точка окажется близко к какому-то минимуму, ничтожно мала.

Практики выработали следующие рекомендации. Выбирают в исходной области $\approx (5 \div 20)N$ случайных точек. Из каждой случайной точки ищут экстремум обобщенным методом Ньютона, ограничиваясь при этом умеренной точностью ε схождения итераций. Каждая траектория спуска приводит на дно котловины или оврага, принадлежащего ближайшему минимуму. При не слишком малом ε конечная точка траектории может оказаться далеко от точки минимума, но достигнутое значение $\Phi(\mathbf{x})$ будет близко к соответствующему минимальному (отличие $\sim \varepsilon^2$). Это уже позволяет сравнивать конечные точки траекторий, проведенных из разных нулевых приближений.

Когда требуется найти глобальный минимум, выбирают конец той траектории, на котором значение $\Phi(\mathbf{x})$ оказалось наименьшим. Из него продолжают спуск, добиваясь сходимости уже с малым ε. Конечную точку спуска принимают за положение глобального минимума.

Если требуется несколько минимумов, удовлетворяющих определенным критериям, то после сравнения выбирают несколько подходящих концов траекторий и продолжают спуски из них.

8.3. РЕШЕНИЕ СЕТОЧНЫХ УРАВНЕНИЙ

8.3.1. Градиентные спуски

При решении уравнений в частных производных эллиптического типа сеточными методами возникает система линейных уравнений огромной размерности

$$A\mathbf{x} = \mathbf{b}. \tag{8.14}$$

Если сетка по каждой переменной содержит N узлов, то в двумерном случае матрица A имеет порядок N^2 (т. е. размер матрицы равен $N^2 \times N^2$), а в трехмерном — порядок N^3. Однако эта матрица имеет специфический вид: она очень слабо заполнена. Например, в двумерном случае из N^2 элементов каждой строки матрицы лишь $5-9$ ненулевые; в трехмерном из N^3 элементов строки ненулевые $7-27$. При этом положение ненулевых элементов фиксировано и описывается несложным правилом: обычно это диагональ, две кодиагонали и т. п. Такие матрицы называют разреженными. Ненулевые элементы в них не образуют плотной ленты, и применять метод Гаусса или другие

прямые методы для их решения крайне невыгодно. Для их решения строят специальные методы, основанные на минимизации квадратичной функции.

В подобных задачах матрица A непроизвольная. Она всегда вещественна и положительна $A > 0$ (это значит, что для любого $\mathbf{x} \neq 0$ выполняется $(\mathbf{x}, A\mathbf{x}) > 0$). Во многих задачах она также симметрична. Пока ограничимся именно этим случаем. Тогда (8.14) эквивалентна нахождению минимума положительно определенной квадратичной формы

$$\Phi(\mathbf{x}) \equiv \frac{1}{2}(\mathbf{x}, A\mathbf{x}) - (\mathbf{x}, \mathbf{b}) = \min. \tag{8.15}$$

Доказательство этого приведено в подразд. 2.1.5. Задачу (8.15) решать методом Ньютона бессмысленно: он требует прямого решения исходной системы (8.14).

Для таких задач предлагались различные методы градиентного спуска. Они основаны на идее потока воды, стекающего по рельефу в поисках низшей точки. Вода стекает в направлении наиболее крутого склона. Но функция быстрее всего убывает в направлении, противоположном градиенту $\operatorname{grad}\Phi \equiv \nabla\Phi \equiv \equiv \partial\Phi/\partial\mathbf{x}$. Тогда, если известно некоторое приближение $\mathbf{x}^{(s)}$, то следующее приближение ищется на прямой

$$\mathbf{x}(\tau) = \mathbf{x}^{(s)} - \tau \operatorname{grad}\Phi(\mathbf{x}^{(s)}), \tag{8.16}$$

где τ — параметр, определяющий шаг вдоль прямой.

Для квадратичной формы (8.15) градиент попросту равен невязке линейной системы (8.14): $\operatorname{grad}\Phi(\mathbf{x}) = A\mathbf{x} - \mathbf{b} \equiv \mathbf{r}$. Поэтому для нее формула (8.16) переходит в

$$\mathbf{x}(\tau) = \mathbf{x}^{(s)} - \tau\mathbf{r}^{(s)}, \quad \mathbf{r}^{(s)} = A\mathbf{x}^{(s)} - \mathbf{b}. \tag{8.17}$$

Остается определить величину шага τ на каждой итерации.

Выгодность подобного подхода связана с малой трудоемкостью каждой итерации. На ней требуется умножать матрицу A на вектор $\mathbf{x}^{(s)}$. Это делают экономично, включая в расчет только ненулевые элементы матрицы A, местоположение которых заранее точно известно. То же относится к операции сложения векторов.

8.3.2. Наискорейший спуск

В этом методе спуск по прямой (8.16) продолжается до тех пор, пока значения функционала (8.15) вдоль нее убывают.

С учетом симметрии матрицы A проведем следующие преобразования:

$$\begin{aligned}\Phi(\mathbf{x}(\tau)) &= \frac{1}{2}(\mathbf{x}^{(s)} - \tau\mathbf{r}^{(s)},\\ A(\mathbf{x}^{(s)} - \tau\mathbf{r}^{(s)})) &- (\mathbf{x}^{(s)} - \tau\mathbf{r}^{(s)}, \mathbf{b}) =\\ = \frac{\tau^2}{2}(\mathbf{r}^{(s)}, A\mathbf{r}^{(s)}) &- \tau(\mathbf{r}^{(s)}, \mathbf{r}^{(s)}) + \Phi(\mathbf{x}^{(s)}).\end{aligned} \tag{8.18}$$

Для определения низшей точки вдоль прямой приравниваем нулю производную от (8.18) по τ и получаем шаг метода наискорейшего спуска:

$$\tau_{\text{нс}}^{(s)} = (\mathbf{r}^{(s)}, \mathbf{r}^{(s)})/(\mathbf{r}^{(s)}, A\mathbf{r}^{(s)}). \tag{8.19}$$

Здесь числитель положителен в силу свойств скалярного произведения, а знаменатель — в силу положительности матрицы A, так что $\tau_{\text{нс}}^{(s)} > 0$. Движение действительно происходит против градиента.

Сходимость. Поскольку матрица $A > 0$ симметрична, все ее собственные значения вещественны и положительны. Для эллиптических уравнений этот спектр имеет специфическое распределение собственных значений, причем $\lambda_{\max} \gg \lambda_{\min}$ (отношение $\lambda_{\max}/\lambda_{\min} \sim N^2$, где N — число узлов разностной сетки по каждой координате).

В этом случае доказывается, что метод наискорейшего спуска сходится линейно с очень близким к единице знаменателем:

$$\begin{aligned}|\mathbf{x}^{(s+1)} - \mathbf{x}^{(s)}| &\approx q|\mathbf{x}^{(s)} - \mathbf{x}^{(s-1)}|,\\ q = (\lambda_{\max} - \lambda_{\min})/(\lambda_{\max} &+ \lambda_{\min}) \approx\\ \approx 1 - 2\lambda_{\min}/\lambda_{\max} &= 1 - O(N^{-2}).\end{aligned} \tag{8.20}$$

Отсюда видно, что сходимость итераций очень медленная. Для нахождения точки минимума с точностью ε нужно огромное число итераций S:

$$S \approx [\lambda_{\max}/(2\lambda_{\min})] \ln(1/\varepsilon) \sim N^2. \tag{8.21}$$

Метод наискорейшего спуска оказывается одним из самых медленных методов.

Из-за близости q к единице критерий прекращения итераций следует выбирать осторожно. Нельзя использовать критерий $|\mathbf{x}^{(s+1)} - \mathbf{x}^{(s)}| < \varepsilon$. Следует определять эффективное значение q в ходе расчета по трем соседним итерациям и использовать улучшенный критерий линейной сходимости:

$$|\mathbf{x}^{(s+1)} - \mathbf{x}^{(s)}| < \varepsilon(1 - q^{(s)}),$$
$$q^{(s)} = |\mathbf{x}^{(s+1)} - \mathbf{x}^{(s)}|/|\mathbf{x}^{(s)} - \mathbf{x}^{(s-1)}|. \tag{8.22}$$

Из (8.22) видно, что получить последние 4 — 5 значащих цифр точного решения этим методом нельзя.

8.3.3. Минимальные невязки

Возьмем ту же антиградиентную прямую (8.16), что и в методе наискорейшего спуска. Однако шаг выберем из другого условия — чтобы новая невязка была минимальной в смысле нормы вектора: $(\mathbf{r}^{(s+1)}, \mathbf{r}^{(s+1)}) = \min$. Это приводит к иной задаче на минимум квадратичной функции от τ:

$$(\mathbf{r}^{(s+1)}, \mathbf{r}^{(s+1)}) = (A\mathbf{x}^{(s+1)} - \mathbf{b}, A\mathbf{x}^{(s+1)} - \mathbf{b}) =$$
$$= (A(\mathbf{x}^{(s)} - \tau\mathbf{r}^{(s)}) - \mathbf{b}, A(\mathbf{x}^{(s)} - \tau\mathbf{r}^{(s)}) - \mathbf{b}) =$$
$$= \tau^2(A\mathbf{r}^{(s)}, A\mathbf{r}^{(s)}) - 2\tau(\mathbf{r}^{(s)}, A\mathbf{r}^{(s)}) + (\mathbf{r}^{(s)}, \mathbf{r}^{(s)}) = \min.$$

Приравняв нулю производную по τ, получим шаг, обеспечивающий минимальную невязку вдоль антиградиентной прямой:

$$\tau_{\text{мн}}^{(s)} = (\mathbf{r}^{(s)}, A\mathbf{r}^{(s)})/(A\mathbf{r}^{(s)}, A\mathbf{r}^{(s)}). \tag{8.23}$$

Метод с таким выбором шага называют методом минимальных невязок. Знаменатель положителен в силу свойств скалярного произведения, а числитель также положителен, поскольку $A > 0$. Следовательно, $\tau_{\text{мн}}^{(s)} > 0$.

Собственные векторы симметричной матрицы $A > 0$ образуют ортогональный базис. Разложением $\mathbf{x}^{(s)}$ нетрудно показать, что

$$1/\lambda_{\min} \geqslant \tau_{\text{нс}}^{(s)} \geqslant \tau_{\text{мн}}^{(s)} \geqslant 1/\lambda_{\max} > 0.$$

Таким образом, шаг метода минимальных невязок всегда меньше, чем у наискорейшего спуска. Несмотря на это можно доказать, что метод минимальных невязок сходится точно с такой же скоростью, как и метод наискорейшего спуска. Знаменатель его линейной сходимости и требуемое число итераций описываются формулами (8.20) и (8.22).

Метод минимальных невязок имеет одно преимущество перед методом наискорейшего спуска. Формула его шага применима даже в том случае, когда матрица A не является знакоопределенной. Вероятность того, что в ходе итераций $A\mathbf{r}^{(s)} = 0$,

нулевая. Если же $A\mathbf{r}^{(s)} \neq 0$, то знаменатель в (8.23) всегда положителен. Но при этом возможен отрицательный числитель и $\tau_{\text{мн}}^{(s)} < 0$, что допустимо, ибо по направлению антиградиента обязан уменьшаться только $\Phi(\mathbf{x})$, а не невязка.

8.3.4. Усеченный спуск

Оба описанных метода сходятся настолько медленно, что полный объем вычислений для достижения точности ε неприемлемо велик, несмотря на дешевизну одной итерации. Однако есть простое видоизменение этих методов, резко повышающее скорость сходимости в подавляющем большинстве практически важных случаев.

Будем умножать расчетный шаг $\tau^{(s)}$, найденный по формулам (8.19) или (8.23), на постоянный поправочный коэффициент c, и получим формулы метода усеченного градиентного спуска:

$$\tau^{(s)} = c\tau_{\text{нс}}^{(s)} \quad \text{или} \quad \tau^{(s)} = c\tau_{\text{мн}}^{(s)}. \tag{8.24}$$

На большом числе тестовых матриц с различными спектрами была исследована зависимость числа итераций S, необходимых для получения точности ε, от величины поправочного коэффициента c. Типичная картина приведена на рис. 8.4.

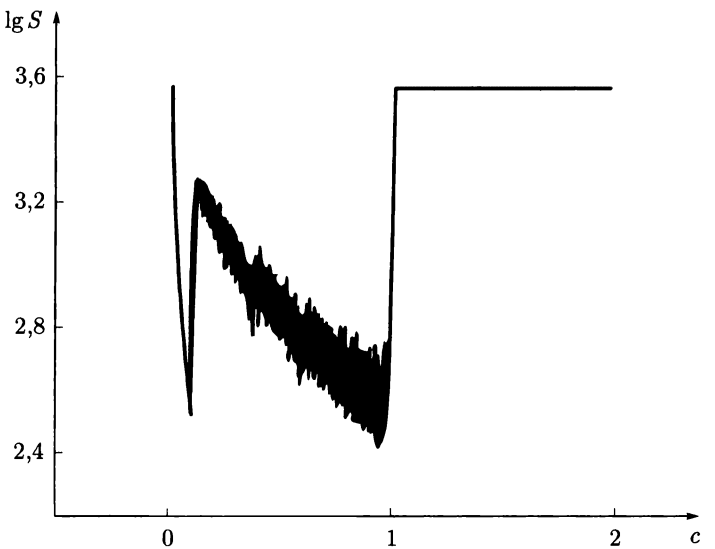


Рис. 8.4. Усеченный градиентный спуск. Зависимость числа итераций S от поправочного коэффициента c

При $c = 1$ метод переходит в классические методы наискорейшего спуска или минимальных невязок. При увеличении c в интервале $1 \leqslant c \leqslant 2$ число итераций остается практически неизменным (хотя для метода наискорейшего спуска $c > 1$ означает уже движение на подъем по противоположному склону котловины). При $c = 2$ достигается линия уровня $\Phi(\mathbf{x})$, соответствующая начальной точке спуска. При дальнейшем увеличении $c > 2$ число итераций резко возрастает. Наоборот, при уменьшении $c < 1$ число итераций сначала резко падает. В интервале $0{,}80 \leqslant c \leqslant 0{,}95$ оно примерно стабилизируется, причем на очень низком уровне. При дальнейшем уменьшении c оно снова возрастает.

Таким образом, выбирая диапазон $0{,}85 \leqslant c \leqslant 0{,}90$, мы надежно остаемся в районе наименьшего числа итераций. Само это число хорошо описывается формулой

$$S \approx \sqrt{\lambda_{\max}/\lambda_{\min}}\,\ln \varepsilon^{-1} = O(N). \tag{8.25}$$

В расчетах практических задач для эллиптических уравнений это означает уменьшение числа итераций в $\sim 100 \div 1\,000$ раз по сравнению с классическими методами спуска. При этом метод не сложнее классических и автоматически настраивается на границы спектра матрицы A (т. е. не требует знания $\lambda_{\min}, \lambda_{\max}$). Предпочтительнее работать, умножая на c шаг метода минимальных невязок (8.25), так как этот метод применим к знакопеременным матрицам A. Проверка на тестах показала, что метод сохраняет скорость сходимости (8.23) даже на несимметричных матрицах A, если они возникают из задач аппроксимации уравнений в частных производных эллиптического (но не гиперболического!) типа.

Введение коэффициента c является эмпирическим приемом, не имеющим строгого обоснования. Более того, можно построить примеры симметричных положительных матриц с таким спектром (например, когда все собственные значения матрицы одинаковы), что метод усеченного градиентного спуска будет сходиться существенно хуже оценки (8.25). Однако это нетипичные случаи. В большинстве же практических задач метод оказался весьма эффективным.

8.3.5. Сопряженные градиенты

Существуют методы, которые при произвольном спектре симметричной положительной матрицы A обеспечивают быструю

сходимость со скоростью (8.25). Это методы сопряженных градиентов, сопряженных невязок и некоторые другие. Поясним идею их построения.

В методе наискорейшего спуска движение по направлению $\mathbf{r}^{(s)}$ продолжается, пока функционал не достигнет минимума на этой прямой. Это означает, что линия спуска коснулась эллипсоида — гиперповерхности уровня $\Phi(\mathbf{x})$ (8.18). Направление следующего спуска $\mathbf{r}^{(s+1)}$ перпендикулярно гиперповерхности уровня, проходящей через точку $\mathbf{x}^{(s+1)}$. Тем самым оно перпендикулярно направлению предыдущего спуска: $(\mathbf{r}^{(s)}, \mathbf{r}^{(s+1)}) = 0$. Аналогично будет $\mathbf{r}^{(s+1)}$ перпендикулярно $\mathbf{r}^{(s+2)}$. Однако $\mathbf{r}^{(s)}$ и $\mathbf{r}^{(s+2)}$ не обязаны быть перпендикулярными.

В методе сопряженных спусков очередные направления выбираются так, чтобы они были перпендикулярны сразу всем предыдущим. При этом каждый раз ищется минимум не вдоль последнего направления, а во всем подпространстве, образованном всеми уже найденными направлениями. При этом формулы метода построены так, что помнить все предыдущие направления не требуется. Достаточно сохранять лишь информацию о текущем шаге $\mathbf{x}^{(s)}$ и предыдущем $\mathbf{x}^{(s-1)}$. Вывод этих формул весьма громоздок, однако сами окончательные формулы достаточно просты.

Есть несколько вариантов этого метода: сопряженные градиенты, сопряженные невязки и другие, причем с рекуррентным или нерекуррентным вычислением невязки. Приведем формулы первого из них.

Помимо векторов решения $\mathbf{x}$, правой части $\mathbf{b}$ и невязки $\mathbf{r}$ вводят два вспомогательных вектора $\mathbf{p}, \mathbf{q}$ и четыре скаляра $\alpha, \beta, R \equiv \equiv (\mathbf{r}, \mathbf{r}), Q$. В качестве нулевого приближения выбирают

$$\mathbf{x}^{(0)} \text{ любое}, \quad \mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)}, \quad \mathbf{p}^{(0)} = 0, \quad R^{(0)} = (\mathbf{r}^{(0)}, \mathbf{r}^{(0)}).$$

Дальнейшие итерации $s = 1, 2, \ldots$ выполняют по следующим формулам:

$$\beta^{(1)} = 1 \text{ и } \beta^{(s)} = R^{(s-1)}/R^{(s-2)} \quad \text{при} \quad s \geqslant 2;$$

$$\mathbf{p}^{(s)} = \mathbf{r}^{(s-1)} + \beta^{(s)}\mathbf{p}^{(s-1)}, \quad \mathbf{q}^{(s)} = A\mathbf{p}^{(s)}, \quad Q^{(s)} = (\mathbf{p}^{(s)}, \mathbf{q}^{(s)});$$

$$\alpha^{(s)} = R^{(s-1)}/Q^{(s)}, \quad \mathbf{x}^{(s)} = \mathbf{x}^{(s-1)} + \alpha^{(s)}\mathbf{p}^{(s)};$$

$$\mathbf{r}^{(s)} = \mathbf{r}^{(s-1)} - \alpha^{(s)}\mathbf{q}^{(s)}, \quad R^{(s)} = (\mathbf{r}^{(s)}, \mathbf{r}^{(s)}).$$

Итерации проводят до тех пор, пока не выполнится условие сходимости $R^{(s)} < \varepsilon^2 R^{(0)}$, где ε — требуемая относительная погрешность.

8.3.6. Нелинейность

Квадратичная функция большой размерности возникала в задачах решения линейных систем большой размерности. Для минимизации произвольных нелинейных функций очень большой размерности эффективных методов нет. Однако существует важный класс задач, где функция большой размерности в некотором смысле близка к квадратичной. Они возникают из решения эллиптических уравнений, коэффициенты которых зависят от самого решения, но не слишком сильно. Такие задачи можно записать в форме

$$\Phi(\mathbf{x}) \equiv (\mathbf{x}, A(\mathbf{x})\mathbf{x}) - 2(\mathbf{b}(\mathbf{x}), \mathbf{x}) = \min, \tag{8.26}$$

где зависимости $A(\mathbf{x}), \mathbf{b}(\mathbf{x})$ можно считать достаточно слабыми.

Далее будем рассматривать задачу $\Phi(\mathbf{x}) = \min$, не требуя формального вида (8.26), но неявно его предполагая. Построим итерационный алгоритм по аналогии с методами спуска для квадратичной функции. В этом случае невязкой будет вектор $\mathbf{r}^{(s)} = \operatorname{grad} \Phi(\mathbf{x}^{(s)}) \equiv \partial\Phi(\mathbf{x}^{(s)})/\partial\mathbf{x} \equiv \Phi_x$. Аналогом матрицы A будет матрица вторых производных $\partial^2\Phi(\mathbf{x}^{(s)})/\partial\mathbf{x}\partial\mathbf{x} \equiv \Phi_{xx}$. Тогда формулы градиентного спуска приобретают следующий вид:

$$\mathbf{x}^{(s+1)} = \mathbf{x}^{(s)} - \tau^{(s)}\Phi_x(\mathbf{x}^{(s)}). \tag{8.27}$$

Здесь шаг $\tau^{(s)}$ для аналога метода наискорейшего спуска равен

$$\tau^{(s)} = (\Phi_x(\mathbf{x}^{(s)}), \Phi_x(\mathbf{x}^{(s)}))/(\Phi_x(\mathbf{x}^{(s)}), \Phi_{xx}(\mathbf{x}^{(s)})\Phi_x(\mathbf{x}^{(s)})); \tag{8.28}$$

для аналога метода минимальных невязок составляет

$$\tau^{(s)} = \frac{(\Phi_x(\mathbf{x}^{(s)}), \Phi_{xx}(\mathbf{x}^{(s)})\Phi_x(\mathbf{x}^{(s)}))}{(\Phi_{xx}(\mathbf{x}^{(s)})\Phi_x(\mathbf{x}^{(s)}), \Phi_{xx}(\mathbf{x}^{(s)})\Phi_x(\mathbf{x}^{(s)}))}. \tag{8.29}$$

Поскольку сами по себе эти методы сходятся медленно, целесообразно проводить усеченные спуски, ускоряющие сходимость. Для этого шаги (8.28) и (8.29) умножают на численный коэффициент, приблизительно равный 0,85.

8.4. ЗАДАЧИ С ОГРАНИЧЕНИЯМИ

8.4.1. Наложение связей

В экономике и технике существует много задач на поиск минимума $\Phi(\mathbf{x})$ с дополнительными ограничениями. Такие ограничения могут быть равенствами или неравенствами. Например,

рассмотрим, как выгодно разместить заказ на изготовление однотипной продукции между несколькими заводами одной фирмы. Надо минимизировать себестоимость изготовления, т. е. это задача на минимум. Казалось бы, надо отдать весь заказ заводу, на котором себестоимость выпуска минимальная. Но у каждого завода имеется ограниченное число станков и рабочих. Отсюда возникают ограничения типа неравенств. По каждому заводу должны выполняться балансы получения и расхода электроэнергии и сырья. Это приводит к ограничениям типа равенств.

Сначала рассмотрим ограничения типа равенств. Тогда N-мерная задача с M связями (ограничениями типа равенств) принимает следующий вид:

$$\Phi(\mathbf{x}) = \min, \quad \mathbf{x} = \{x_n, 1 \leqslant n \leqslant N\}; \tag{8.30}$$

$$\varphi_m(\mathbf{x}) = 0, \quad 1 \leqslant m \leqslant M. \tag{8.31}$$

Каждое скалярное ограничение (8.31) означает выделение из некоторого пространства подпространства размерности на единицу меньше. Все ограничения в совокупности уменьшают размерность пространства на M. Таким образом, задача (8.30), (8.31) есть поиск минимума N-мерной скалярной функции в подпространстве размерности $N - M$. При $M = N$ это подпространство сводится к точке (или дискретной совокупности точек). В этом случае задача фактически перестает быть задачей на минимум. Поэтому содержательны лишь задачи с $M < N$.

Если все ограничения (8.31) линейны, то можно записать явное выражение подпространства ограничений. Но даже в этом простейшем случае при больших M, N такое явное выражение будет громоздким и неудобным для практического использования. В нелинейном же случае явное выражение подпространства найти не удается. Поэтому используют метод неопределенных множителей Лагранжа. Напомним его.

Выберем M свободных параметров λ_m, $1 \leqslant m \leqslant M$. Рассмотрим задачу на минимум следующей функции:

$$F(\mathbf{x}, \lambda) \equiv \Phi(\mathbf{x}) + \sum_{m=1}^{M} \lambda_m \varphi_m(\mathbf{x}) = \min. \tag{8.32}$$

Это функция $N + M$ переменных $x_1, x_2, \ldots, x_N, \lambda_1, \lambda_2, \ldots, \lambda_M$. Ее минимум ищется во всем $(N + M)$-мерном пространстве. Задача (8.32) есть задача на экстремум без ограничений. Для ее решения нужно приравнять нулю частные производные $F(\mathbf{x}, \lambda)$ по

всем аргументам. Дифференцирование (8.32) по λ_m дает уравнения, являющиеся исходными ограничениями (8.31). Дифференцирование же (8.32) по x_n дает следующие уравнения:

$$\frac{\partial F}{\partial x_n} = \frac{\partial \Phi(\mathbf{x})}{\partial x_n} + \sum_{m=1}^{M} \lambda_m \frac{\partial \varphi_m(\mathbf{x})}{\partial x_n} = 0, \quad n = 1, 2, \ldots, N. \tag{8.33}$$

Уравнения (8.31) и (8.33) образуют систему $M + N$ уравнений для определения такого же числа неизвестных $\{x_n, \lambda_m\}$. Таким образом, задача на экстремум со связями свелась к обычной задаче минимизации большей размерности.

Для произвольного вида функций и не слишком больших значений M, N полученную задачу в форме системы уравнений (8.31), (8.33) можно решать классическим или обобщенным методами Ньютона. Если N и M очень велики (что типично для задач экономики), то задачу удается решить лишь для специфических функций (квадратичных, линейных или очень близких к ним). В этом случае полезны методы градиентного спуска, описанные ранее.

8.4.2. Ограниченная область

Рассмотрим задачу поиска минимума в N-мерном пространстве с L ограничениями типа неравенств:

$$\Phi(\mathbf{x}) = \min, \quad \mathbf{x} = \{x_n, 1 \leqslant n \leqslant N\}, \tag{8.34}$$

$$\psi_l(\mathbf{x}) \leqslant 0, \quad 1 \leqslant l \leqslant L. \tag{8.35}$$

Если бы ограничение было равенством, оно выделяло бы гиперповерхность размерности $N - 1$. Неравенство же означает, что берется N-мерное полупространство, отсекаемое этой гиперповерхностью. Совокупность всех ограничений вырезает из исходного пространства N-мерную область G. Таким образом, (8.34), (8.35) есть задача поиска минимума не во всем N-мерном пространстве, а в N-мерной области G.

Поскольку ограничения типа неравенств не понижают размерности пространства, их число L может быть произвольным. Однако может оказаться, что какие-то из ограничений фактически не работают (например, возьмем ограничение $x_1 \leqslant 2$ и $x_1 \leqslant 3$). Еще сложнее ситуация, когда система ограничений несовместна (например, $x_1 \leqslant 2$ и $x_1 \geqslant 3$), так что область G не

существует. Проверить возникновение таких ситуаций в многомерном случае далеко не всегда удается.

Задачи с ограничениями типа неравенств труднее, чем задачи со связями. Во-первых, в них искомый минимум нередко достигается на границе области G, а не внутри нее. Во-вторых, градиентные спуски легко могут вывести нас за границу области.

Для решения таких задач можно применять метод штрафных функций. Рассмотрим следующую задачу:

$$\Psi(\mathbf{x}) \equiv \Phi(\mathbf{x}) + \sum_{l=1}^{L} c_l \{\max[0, \psi_l(\mathbf{x})]\}^2 = \min; \tag{8.36}$$

здесь $c_l > 0$ — некоторые произвольно выбранные коэффициенты. Если l-е ограничение (8.35) выполняется, то соответствующая фигурная скобка в (8.36) обращается в нуль. Если же это ограничение не выполняется, то соответствующий член в (8.36) положителен, т. е. увеличивает $\Psi(\mathbf{x})$. Его можно рассматривать как штраф за нарушение ограничений (8.35).

Очевидно, внутри области G никаких штрафов нет. Поэтому внутри этой области $\Psi(\mathbf{x}) = \Phi(\mathbf{x})$. Значит, если минимум (8.36) достигается внутри области G, то он является искомым условным экстремумом: все ограничения выполнены (сумма в (8.36) обращается в нуль) и $\Phi(\mathbf{x})$ минимальна.

Рассмотрим ситуацию, когда минимум $\mathbf{x}_\Psi$ функции $\Psi(\mathbf{x})$ лежит вне области G. Тогда минимум самой функции $\Phi(\mathbf{x})$ для задачи без ограничений также лежит вне области G, а для задачи с ограничениями этот минимум $\mathbf{x}_\Phi$ должен лежать на границе. В этом случае можно показать, что при достаточно больших коэффициентах $c_l > 0$ оба минимума $\mathbf{x}_\Psi$ и $\mathbf{x}_\Phi$ будут близки.

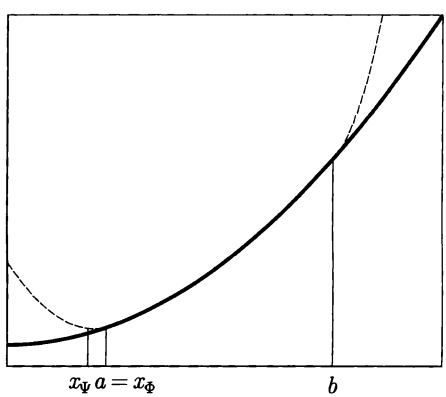


Рис. 8.5. Метод штрафных функций ($\Phi(x)$ — сплошная линия, $\Psi(x)$ — штриховая линия)

Поясним это простым примером. Рассмотрим монотонную функцию одной переменной $\Phi(x)$, показанную на рис. 8.5. Введем два ограничения: 1) $x \geqslant a$; 2) $x \leqslant b$, где $a < b$. Эти ограничения выделяют в качестве области G отрезок $[a, b]$. Значит, требуется найти минимум на этом отрезке. Видно, что он достигается на левой границе отрезка, т. е. $x_\Phi = a$. Штрафная сумма в (8.36) равна нулю на отрезке $[a, b]$. Левее точки a она сводится к слагаемому $c_1(x-a)^2$, а правее точки b — к слагаемому $c_2(x-b)^2$. Если коэффициенты c_1, c_2 достаточно велики, то на некотором удалении от границ отрезка $\Psi(x)$ будет быстро возрастающей. Ее минимум x_Ψ лежит близко к точке a, т. е. к искомому экстремуму x_Φ.

Погрешность. Пусть решение $\mathbf{x}_\Phi$ задачи с ограничениями (8.34), (8.35) достигается на границе области G. Обозначим вектор отклонения от этой точки $\Delta = \mathbf{x} - \mathbf{x}_\Phi$. Разложим $\Psi(\mathbf{x})$ вблизи точки $\mathbf{x}_\Phi$, удерживая в разложении $\Phi(\mathbf{x})$ и $\psi_l(\mathbf{x})$ только линейные члены. Учитывая, что $\psi_l(\mathbf{x})$ на границе области G обращаются в нуль, получим

$$\Psi(\mathbf{x}) \approx \Phi(\mathbf{x}_\Phi) + \left(\frac{\partial \Phi}{\partial \mathbf{x}}, \Delta\right) + \sum_l{}^* c_l \left(\frac{\partial \psi_l}{\partial \mathbf{x}}, \Delta\right)^2. \tag{8.37}$$

Здесь все производные берутся в точке $\mathbf{x}_\Phi$, а знак «$*$» при сумме означает, что суммируются по l только те слагаемые, которые «работают» вблизи данной точки границы (см. рис. 8.5, где вблизи точки минимума «работало» только первое штрафное слагаемое).

Минимум $\Psi(\mathbf{x})$ найдем, приравнивая нулю ее производную по вектору Δ. Получим следующую систему линейных уравнений:

$$\sum_l{}^* c_l \left[\frac{\partial \psi_l}{\partial \mathbf{x}} \left(\frac{\partial \psi_l}{\partial \mathbf{x}}\right)^T\right] \Delta = -\frac{1}{2}\frac{\partial \Phi}{\partial \mathbf{x}}. \tag{8.38}$$

Здесь $\partial\psi_l/\partial\mathbf{x}$ — вектор-столбец; умножение его на транспонированную вектор-строку дает матрицу. Эта матрица симметрична и положительна. Поскольку все $c_l > 0$, то сумма по l также будет симметричной положительной матрицей. Тогда можно оценить величину решения этой линейной системы (8.38): $|\Delta| < \text{const}/\min\{c_l\}$. При больших c_l это решение дает нам хорошее приближение к значению $\mathbf{x}_\Psi$. Отсюда получаем оценку

$$|\mathbf{x}_\Psi - \mathbf{x}_\Phi| < \text{const}/\min\{c_l\}. \tag{8.39}$$

Таким образом, если выбрать достаточно большие коэффициенты c_l, то искомое решение будет найдено с хорошей точностью.

Стратегия. Для функций и ограничений достаточно общего вида искать минимум $\Psi(\mathbf{x})$ можно методами градиентного спуска. Однако это нелегко. Штрафные функции «приклеиваются» к $\Phi(\mathbf{x})$ так, что $\Psi(\mathbf{x})$ непрерывна вместе с первыми производными, но вторые производные на границе области G разрывны. Методы же градиентного спуска дают действительно хорошие результаты при непрерывных вторых производных, а в данной ситуации могут вести себя хуже.

Этот эффект тем сильнее, чем больше c_l, так как при этом увеличиваются скачки вторых производных на границе. Но для получения хорошей близости согласно оценке (8.39) надо брать как можно бóльшие c_l. Поэтому целесообразна следующая стратегия расчета.

Сначала берут не слишком большие $c_l^{(0)}$. Выбирают некоторое нулевое приближение $\mathbf{x}^{(0)}$, проводят градиентный спуск и находят минимум $\mathbf{x}_\Psi^{(1)}$. Он будет довольно далек от $\mathbf{x}_\Phi$, поскольку $c_l^{(0)}$ не слишком велики. Косвенно об этом можно судить, подставляя $\mathbf{x}_\Psi^{(1)}$ в ограничения (8.35) и количественно оценивая их невыполнение:

$$d^{(1)} = \sum_{l=1}^{L} \max[0, \psi_l(\mathbf{x}_\Psi^{(1)})]. \tag{8.40}$$

Затем увеличим все коэффициенты c_l в штрафных функциях (в одинаковое число раз):

$$c_l^{(1)} = \mu^{(1)} c_l^{(0)}, \quad 1 \leqslant l \leqslant L.$$

Это затруднит градиентный спуск, но теперь в качестве нулевого приближения возьмем вектор $\mathbf{x}_\Psi^{(1)}$, сравнительно близкий к исходному минимуму. Это облегчает проведение расчета. Новый найденный минимум обозначим $\mathbf{x}_\Psi^{(2)}$. Он будет ближе к $\mathbf{x}_\Phi$. Количественно об этом можно судить, подставляя $\mathbf{x}_\Psi^{(2)}$ в оценку (8.40) и вычисляя $d^{(2)}$.

Будем повторять этот процесс, выбирая на s-й итерации коэффициенты $c_l^{(s)} = \mu^{(s)} c_l^{(0)}$, $1 \leqslant l \leqslant L$. Если $\mu^{(s)} > 0$ образуют неограниченно возрастающую последовательность, процесс будет сходиться к исходному минимуму $\mathbf{x}_\Phi$. По скорости убывания $d^{(s)} \geqslant 0$ можно судить о скорости сходимости. Если $d^{(s)}$ убывают

примерно как $1/\mu^{(s)}$, то можно оценить величину μ, необходимую для достижения хорошей точности.

Однако следует помнить, что приближения $\mathbf{x}_\Psi^{(s)}$, найденные методом штрафных функций, будут лежать вне заданной области G.

Выбор коэффициентов. В общем случае выбор начального распределения коэффициентов c_l (т. е. отношений коэффициентов с различными l) не вполне ясен. Однако есть более простой случай, когда градиент каждого ограничения $\partial\psi_l/\partial\mathbf{x}$ слабо меняется на том участке границы области G, где это ограничение «работает». Тогда целесообразно выбирать

$$c_l \approx c/(\partial\psi_l/\partial\mathbf{x})^2, \quad 1 \leqslant l \leqslant L.$$

Здесь $c > 0$ — некоторый общий коэффициент, не зависящий от l.

8.4.3. Общий случай

Во многих прикладных проблемах одновременно присутствуют ограничения обоих типов. Тогда задача принимает следующий вид:

$$\Phi(\mathbf{x}) = \min, \quad \mathbf{x} = \{x_n, 1 \leqslant n \leqslant N\}; \tag{8.41}$$

$$\varphi_m(\mathbf{x}) = 0, \quad 1 \leqslant m \leqslant M, \quad M < N; \tag{8.42}$$

$$\psi_l(\mathbf{x}) \leqslant 0, \quad 1 \leqslant l \leqslant L. \tag{8.43}$$

Связи (8.42) выделяют из N-мерного пространства $(N-M)$-мерное подпространство. Неравенства (8.43) вырезают в N-мерном пространстве область G. При этом в $N-M$-мерном подпространстве также вырезается некоторая область G^*. Таким образом, требуется найти минимум функции $\Phi(\mathbf{x})$ в области G^*.

Эта задача труднее двух предыдущих, так как любые методы спуска работают во всем N-мерном пространстве. Поэтому даже если начальное приближение принадлежит $(N-M)$-мерному подпространству, траектория спуска сейчас же выйдет за пределы этого подпространства. Существуют два подхода к решению задачи (8.41) — (8.43).

Штрафная функция. Можно ввести единую штрафную функцию, штрафуя за нарушение не только неравенств (8.43), но и равенств (8.42). Тогда исходная задача заменяется задачей на безусловный экстремум:

$$\Psi(\mathbf{x}) \equiv \Phi(\mathbf{x}) + \sum_{m=1}^{M} a_m \varphi_m^2(\mathbf{x}) + \\ + \sum_{l=1}^{L} c_l \{\max[0, \psi_l(\mathbf{x})]\}^2 = \min. \tag{8.44}$$

Здесь $a_m > 0, c_l > 0$ — штрафные коэффициенты. Эта задача решается аналогично подразд. 8.4.2, причем можно получить оценку

$$|\mathbf{x}_\Psi - \mathbf{x}_\Phi| < \text{const}/\min(a_m, c_l).$$

Итерационный процесс начинается с выбора начальных коэффициентов. Если градиенты ограничений слабо меняются, то целесообразно выбирать

$$a_m \approx c/(\partial\varphi_m/\partial\mathbf{x})^2, \quad 1 \leqslant m \leqslant M,$$

$$c_l \approx c/(\partial\psi_l\partial\mathbf{x})^2, \quad 1 \leqslant l \leqslant L.$$

Далее в ходе итераций все коэффициенты умножаются на один и тот же множитель $\mu^{(s)}$, увеличивающийся от одной итерации к другой.

На каждой итерации вычисляется величина, описывающая нарушение ограничений:

$$d^{(s)} = \sum_{m=1}^{M} |\varphi_m(\mathbf{x}_{Psi^{(s)}})| + \sum_{l=1}^{L} \max[0, \psi_l(\mathbf{x}_\Psi^{(s)})].$$

Если эта величина убывает примерно как $d^{(s)} \sim 1/\mu^{(s)}$, то процесс сходится удовлетворительно.

Смешанная стратегия. Штрафные функции неудобны тем, что их введение приводит к разрывам второй производной минимизируемой функции. Однако задачу (8.41) — (8.43) можно решить иным способом — комбинацией метода неопределенных множителей Лагранжа с методом штрафных функций. Для этого связи учитывают методом неопределенных множителей Лагранжа, переходя к функции $F(\mathbf{x}, \boldsymbol{\lambda})$ аналогично (8.32). Для этой функции увеличенного числа переменных $N + M$ рассматривают задачу с ограничениями типа неравенств (8.43). Теперь в функцию $\Psi(\mathbf{x},\boldsymbol{\lambda})$ войдут штрафы только за нарушение неравенств. Лишь они будут вносить разрывы во вторые производные. Это облегчает решение задачи минимизации, несмотря на увеличивающееся число переменных.

8.5. МИНИМИЗАЦИЯ ФУНКЦИОНАЛА

8.5.1. Прикладные проблемы

Функционалом $\Phi[u]$ называют скалярную величину, зависящую от функции $u(x)$. Многие прикладные задачи изначально формулируются именно в виде функционала $\Phi[u]$, который надо минимизировать. Такие задачи широко распространены в механике, гидродинамике, электродинамике, квантовой механике, экономике. Рассмотрим некоторые примеры.

Пример 8.1. В 4.2 упоминалось, что сплайны были построены как приближения к упругому бруску. Точное уравнение для гибкого бруска было написано Д. Бернулли (1742). Вывод основан на том, что потенциальная энергия изгиба пропорциональна квадрату кривизны бруска ρ. Тогда полная потенциальная энергия бруска равна $E = \int \rho^2 dl$, где l — длина дуги. Форму изогнутого бруска опишем функцией $u(x)$. Напомним, что через нее выражаются кривизна $\rho = u_{xx}/(1+u_x^2)^{3/2}$ и элемент длины дуги $dl = (1+u_x^2)^{1/2}dx$. При равновесии потенциальная энергия минимальна. Отсюда получаем задачу

$$\Phi[u] \equiv \int_a^b \frac{u_{xx}^2}{(1+u_x^2)^{5/2}} dx = \min, \tag{8.45}$$

где a и b — абсциссы начала и конца бруска. Мы получили задачу на минимум функционала.

Будем считать, что брусок изогнут достаточно слабо. Тогда $u_x \approx \text{const}$ и знаменатель в (8.45) почти постоянен, так что его можно отбросить. Задача упругого бруска сильно упрощается:

$$\int_a^b u_{xx}^2 dx = \min. \tag{8.46}$$

Именно это уравнение было взято Шонбергом (1946) в качестве исходного. Напомним, что вариационное уравнение Л. Эйлера для функционала (8.46) имеет вид $u^{(IV)}(x) = 0$. Его решением является произвольный кубический многочлен. Поэтому Шонберг в задаче интерполяции считал функцию кубическим многочленом между каждой парой соседних узлов, и

«склеивал» соседние многочлены в узлах. Отсюда видно, что кубический сплайн является хорошим приближением к упругому бруску лишь для слабо изогнутых, т. е. близких к прямой графиков $u(x)$. Это сильно ограничивает возможности сплайнов.

Заметим, что если из условия минимума функционала (8.46) выводить граничные условия, то получается $u''(a) = u''(b) = 0$. Однако в 4.2 отмечалось, что в задачах интерполяции эти граничные условия являются весьма неудачными и сильно понижающими точность вблизи границ.

Интересно было бы решить задачу упругого бруска (8.45), однако для нее вариационное уравнение Эйлера оказывается довольно сложным дифференциальным уравнением четвертого порядка, неразрешимым в элементарных функциях.

Пример 8.2. Пусть однородная стальная струна натянута горизонтально и прогибается под действием собственной массы и внешней нагрузки. Обозначим отклонение струны от горизонтальной линии через $u(x)$. Тогда ее суммарная энергия выражается следующим интегралом:

$$\int_0^a \left[\frac{\varkappa}{2} u_x^2(x) + f(x)u(x)\right] dx, \tag{8.47}$$

где $\varkappa$ — сила горизонтального натяжения струны; $f(x)$ — сумма массы и внешней вертикальной нагрузки единицы длины струны.

Первое слагаемое в интеграле является потенциальной энергией отклонения натянутой струны от горизонтали, а второе — потенциальная энергия, связанная с массой и внешней силой. Это также задача на минимум функционала, и ее вариационное уравнение Эйлера имеет следующий вид:

$$\varkappa u_{xx}(x) = f(x). \tag{8.48}$$

Граничные условия для этого уравнения также следует выводить не из вариационной задачи, а из дополнительных физических требований. Например, если оба конца закреплены на горизонтальной линии, то надо полагать $u(0) = u(a) = 0$.

Пример 8.3. В предыдущих примерах функционал квадратично зависел от функции. Это наиболее простой для решения случай. Но встречается немало функционалов с более сложной зависимостью. Например, в физике высоких давлений используют статистическую модель атома Томаса — Ферми. В ней

энергия атома зависит от электронной плотности $\rho(\mathbf{x})$, где $\mathbf{x}$ — радиус-вектор:

$$E[\rho(\mathbf{x})] = \int \rho^{5/3}(\mathbf{x})dv - z\int \frac{\rho(\mathbf{x})}{x}dv + \frac{1}{2}\iint \frac{\rho(\mathbf{x})\rho(\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}dvdv'. \quad (8.49)$$

Здесь интегрирование производится по объему атомной ячейки. Второй интеграл есть потенциальная энергия взаимодействия электронов с ядром заряда z. Третий интеграл соответствует потенциальной энергии взаимодействия электронов между собой. Первый интеграл приближенно описывает кинетическую энергию электронов. Минимизация функционала полной энергии дает условие стационарного состояния атома. Из-за степени $5/3$ в первом интеграле задача (8.49) намного труднее квадратичной.

Задачи. Наиболее общей задачей является следующая: найти минимальное значение функционала $\Phi_{\min}$ и минимизирующую функцию $u_{\min}(x)$, на которой этот минимум достигается, т. е.

$$\Phi[u] = \min; \quad \Phi_{\min} =? \quad u_{\min}(x) =? \quad (8.50)$$

Вблизи минимума зависимость функционала от функции является приблизительно квадратичной. Поэтому небольшие изменения $u(x)$ вблизи $u_{\min}(x)$ должны слабо влиять на величину функционала. Это означает, что найти $u_{\min}(x)$ с хорошей точностью будет заметно труднее, чем найти $\Phi_{\min}$.

В некоторых случаях достаточно найти только $\Phi_{\min}$ без нахождения соответствующей функции $u_{\min}(x)$. Например, может потребоваться лишь полная энергия атома (8.49) без знания электронной плотности. Из изложенного видно, что эта задача существенно проще. Хотя во всех методах расчета электронная плотность все равно попутно определяется, но не нужно находить ее с высокой точностью.

Существуют также задачи, где надо вычислять другие функционалы, зависящие от $u_{\min}(x)$. В сжатом атоме таким функционалом может быть давление, выражающееся через электронную плотность. В этих случаях $u_{\min}(x)$ надо определять с высокой точностью.

8.5.2. Сеточный метод

Сеточный метод является наиболее универсальным. Он применим практически к любым функционалам, как квадратичным, так и более сложным. При правильном его использовании

можно получать результаты с гарантированной оценкой точности.

Идея метода несложна. Ограничимся для простоты записи одномерными задачами. Пусть $u(x)$ определена на отрезке $[a, b]$. Введем на этом отрезке сетку $\{x_n, 0 \leqslant n \leqslant N, x_0 = a, x_N = b\}$. Функционал может содержать операции дифференцирования и интегрирования. Заменим все производные некоторыми разностными выражениями от $u_n \equiv u(x_n)$, а все интегралы — квадратурными суммами. Тогда функционал аппроксимируется некоторой функцией от сеточной функции $\{u_n\}$:

$$\Phi[u(x)] \approx F(\mathbf{u}) \equiv F(u_0, u_1, \ldots, u_N).$$

Решим задачу $F(\mathbf{u}) = \min$. Это задача на минимум функции от многих переменных, рассмотренная в 8.2. Найденный вектор $\mathbf{u}_{min}$ будет приближенным выражением для $u_{\min}(x)$ в узлах сетки. Соответствующее минимальное значение $F_{\min}(\mathbf{u}) = F(\mathbf{u}_{\min})$ будет приближением к $\Phi_{\min}$.

Рассмотрим реализацию этого метода для двух типичных случаев — квадратичного и более сложного функционалов.

Квадратичный функционал. Рассмотрим задачу на стационарный прогиб гибкой струны (8.47), закрепленной на концах: $u(0) = 0, u(a) = 0$. Ограничимся для простоты равномерной сеткой $x_n = nh, h = a/N$. Разностную производную $u(x)$ определим в середине интервала:

$$(du/dx)_{n-1/2} \approx (u_n - u_{n-1})/h, \quad 1 \leqslant n \leqslant N.$$

Первое слагаемое в интеграле (8.47) по интервалу $[x_{n-1}, x_n]$ возьмем по формуле средних. Это дает вклад отрезка:

$$(du/dx)^2_{n-1/2} h \approx (u_n - u_{n-1})^2/h, \quad 1 \leqslant n \leqslant N.$$

Эти интегралы надо просуммировать по всем N интервалам.

Интегрирование второго слагаемого удобно проводить по формуле средних, выбирая отрезки $[x_{n-1/2}, x_{n+1/2}]$; вклад от граничных отрезков будет нулевым, так как $u_0 = u_N = 0$. Таким образом, второе слагаемое надо суммировать лишь по внутренним узлам $1 \leqslant n \leqslant N - 1$. В результате имеем

$$F(\mathbf{u}) = \frac{\varkappa}{2h} \sum_{n=1}^{N} (u_n - u_{n-1})^2 + h \sum_{n=1}^{N-1} f_n u_n. \tag{8.51}$$

Мы получили алгебраическую функцию, являющуюся аппроксимацией исходного функционала. Остается минимизировать ее.

Исходный функционал был квадратичным, так что функция (8.51) получилась также квадратичной. Минимизация функции заключается в приравнивании нулю частных производных по всем неизвестным u_n.

Поскольку функция квадратична, это дает систему линейных уравнений для определения неизвестных величин u_n:

$$-(u_{n-1} - 2u_n + u_{n+1})/h + hf_n = 0, \quad 1 \leqslant n \leqslant N-1. \qquad (8.52)$$

В первом уравнении присутствует u_0, а в последнем — u_N, которые равны нулю в силу граничных условий. Таким образом, (8.52) есть система $(N-1)$ уравнений для такого же числа неизвестных. Она имеет трехдиагональную матрицу и легко решается прогонкой.

Если функционал имеет более сложный вид, но остается квадратичным, то данный метод дает квадратичную функцию $F(\mathbf{u})$. Поэтому для определения сеточной функции $\mathbf{u}$ попрежнему получается система линейных уравнений. Ее матрица может содержать большее число диагоналей или быть плотно заполненной. Но во всех этих случаях сеточное решение находится без труда.

Сложные функционалы. Случай неквадратичного функционала приводит к более сложному алгоритму. Рассмотрим атом Томаса — Ферми (8.49) со следующими упрощениями. Будем считать атом сферически симметричным, а атомную ячейку — шаром радиуса R. Предположим сферически симметричное распределение электронов $\rho(\mathbf{x}) = \rho(x)$. Проинтегрируем по углам в сферических координатах, учитывая, что

$$|\mathbf{x} - \mathbf{x}'| = \sqrt{x^2 + x'^2 - 2(\mathbf{x}, \mathbf{x}')}, \quad (\mathbf{x}, \mathbf{x}') = xx' \cos\theta,$$

$$\int_0^\pi \frac{\sin\theta d\theta}{|\mathbf{x} - \mathbf{x}'|} = \frac{1}{xx'}(x + x' - |x - x'|).$$

Получим одномерное выражение

$$E[\rho(x)] = 4\pi \int_0^R [\rho^{5/3}(x)x^2 - z\rho(x)x]dx + \\ +4\pi^2 \int_0^R \int_0^R \rho(x)\rho(x')(x + x' - |x - x'|)xx'dxdx'. \qquad (8.53)$$

Это выражение не содержит производных. Введем равномерную сетку $x_n = nR/N$. Выберем в качестве неизвестных значения в полуцелых узлах $\rho_{n-1/2} \equiv \rho(x_{n-1/2})$. Вычисляя интегралы по формуле средних, получим

$$E(\rho) = \frac{4\pi R}{N} \sum_{n=1}^{N} (\rho_{n-1/2}^{5/3} x_{n-1/2}^2 - z\rho_{n-1/2} x_{n-1/2}) + \\ + \frac{4\pi^2 R^2}{N^2} \sum_{n=1}^{N} \sum_{m=1}^{N} \rho_{n-1/2}\rho_{m-1/2} x_{n-1/2} x_{m-1/2} \times \\ \times (x_{n-1/2} + x_{m-1/2} - |x_{n-1/2} - x_{m-1/2}|) = \min, \\ x_{n-1/2} = \frac{R(n-1/2)}{N}. \tag{8.54}$$

Теперь подлежащая минимизации алгебраическая функция получилась не квадратичной. Дифференцируя по $\rho_{n-1/2}$ и сокращая множители, получим систему нелинейных алгебраических уравнений:

$$\frac{5}{3}\rho_{n-1/2}^{2/3} x_{n-1/2} - z + \frac{2\pi R}{N} \sum_{m=1}^{N} \rho_{m-1/2} x_{m-1/2} \times \\ \times (x_{n-1/2} + x_{m-1/2} - |x_{n-1/2} - x_{m-1/2}|) = 0. \tag{8.55}$$

Ввиду нелинейности системы (8.55), ее решение необходимо находить классическим или обобщенным методом Ньютона. Это может оказаться самостоятельной сложной задачей из-за большого числа неизвестных $\rho_{n-1/2}$. Для обеспечения сходимости метода при этом необходимо достаточно хорошо выбирать нулевое приближение, что непросто.

Контроль точности. Будем предполагать, что точное решение исходной задачи минимизации (8.1) является достаточно гладким. Теоретические оценки точности численных методов достаточно сложны, особенно для нелинейных задач. Однако оценку погрешности нужно проводить методом Ричардсона, многократно сгущая сетку.

Пусть для сеточной аппроксимации функционала использованы формулы p-го порядка точности. Рассмотрим случай, когда необходимо аккуратно найти само решение $u(x)$. Тогда выберем такой коэффициент сгущения сетки r, чтобы значения узлов на соседних сетках совпадали. Так, в задаче прогиба струны (8.52) используются узловые значения u_n. Если выбрать $r = 2$,

то четные узлы сгущенной сетки совпадут с узлами редкой сетки. В задаче атома (8.54) фигурируют значения $\varrho_{n-1/2}$ в полуцелых точках. Если выбрать $r = 3$, то середина каждого третьего интервала подробной сетки будет совпадать с серединой интервала редкой сетки.

Взяв значения на редкой сетке $u(x; N)$ и на более подробной сетке $u(x; rN)$ в одной и той же пространственной точке x, получим асимптотически точную оценку погрешности:

$$\Delta(x; rN) = [u(x; rN) - u(x; N)]/(r^p - 1)$$

и решение повышенной точности

$$\tilde{u}(x; rN) = u(x; rN) + \Delta(x; rN).$$

Эти значения можно вычислить не во всех узлах подробной сетки, а только в тех, которые совпадают с узлами редкой сетки. Если требуется точность ε, то сгущение надо продолжать до тех пор, пока не будет получена $\|\Delta\| \leqslant \varepsilon$. На практике используют нормы $\|\cdot\|_c$ или $\|\cdot\|_{l_2}$, в зависимости от требований задачи.

При многократном сгущении сетки можно применять рекуррентное уточнение по Ричардсону (см. подразд. 3.2). Это позволяет добиться высокой точности при умеренном числе узлов сетки. Обоснование контроля точности в сеточных методах будет дано в Ч. II этого курса. Напомним, что при рекуррентном применении метода Ричардсона каждое очередное сгущение сетки повышает порядок точности на 2 для симметрично построенных схем, и на 1 для несимметричных схем.

Если достаточно найти лишь минимальное значение $\Phi_{\min}$, а само минимизирующее решение $u_{\min}(x)$ не требуется, то можно использовать любой коэффициент сгущения сетки r, в том числе нецелый.

Таким образом, сеточный метод единообразно применяется к функционалам произвольного вида, а использование сгущения сеток при этом позволяет получить асимптотически точную оценку погрешности. Метод является простым и универсальным, и наиболее употребителен в настоящее время.

Нулевые приближения. Если функционал квадратичен, то минимизирующая система уравнений линейна, и сеточное решение легко вычисляется методом Гаусса. Для более сложных функционалов сеточное решение удовлетворяет системе нелинейных уравнений. Ее решают итерационными методами, а для них надо задать достаточно близкое нулевое приближение. Метод сгущения сетки помогает решить и эту проблему.

Возьмем начальную сетку с небольшим числом узлов N. Тогда количество неизвестных u_n будет малым. Даже если нулевое приближение задано не слишком удачно, и число итераций окажется большим, общая трудоемкость будет невелика.

Решение, найденное на грубой сетке, может быть еще не очень близким к точному. Однако оно уже будет хорошим нулевым приближением для итерационной процедуры на более подробной сетке. Надо только интерполировать найденные значения с грубой сетки на все узлы сгущенной сетки. Как правило, при этом классический ньютоновский процесс сходится за две — четыре итерации (а прибегать к обобщенному методу Ньютона нет необходимости). При каждом очередном сгущении нулевое приближение нужно брать с предыдущей сетки. Тогда суммарный объем вычислений при многократном сгущении будет эквивалентен $4-8$ итерациям на самой подробной сетке. Это еще одно преимущество сеточного метода.

8.5.3. Метод Ритца

Выберем некоторый набор базисных функций $\varphi_m(x)$. Функции должны быть линейно независимыми, а их система — полной в некотором классе функций, которому принадлежит искомое решение. Аппроксимируем решение обобщенным многочленом:

$$u(x) \approx u_M(x) \equiv \sum_{m=0}^{M} c_m \varphi_m(x). \tag{8.56}$$

Подставим выражение (8.56) в исходный функционал и выполним все дифференцирования и интегрирования по аргументу x. Тогда функционал окажется некоторой функцией F большого числа неизвестных параметров c_m:

$$\Phi[u_M(x)] = \Phi\left[\sum_{m=0}^{M} c_m \varphi_m(x)\right] = F(c_0, c_1, c_2, \ldots, c_M).$$

Определим параметры c_m из условия минимизации:

$$F(c_0, c_1, c_2, \ldots, c_M) = \min. \tag{8.57}$$

Таким образом, задача свелась к поиску минимума функции многих переменных, способы решения которой описаны в подразд. 8.2.

Квадратичные функционалы. Если функционал $\Phi[u(x)]$ квадратичен, то функция F будет квадратично зависеть от коэффициентов c_m. В этом случае задача на минимум после приравнивания нулю производных по c_m сводится к системе линейных уравнений для c_m. Даже если матрица системы будет плотно заполненной, ее нетрудно решить методом Гаусса. Поэтому для квадратичных функционалов можно выбирать много параметров ($M \sim 100 \div 1\,000$ и более) и добиваться достаточно высокой точности. Именно такой метод был предложен Ритцем.

В качестве примера рассмотрим задачу прогиба струны (8.47) с закрепленными концами: $u(0) = 0, u(a) = 0$. Выберем полную систему функций, удовлетворяющую данным граничным условиям:

$$\varphi_m(x) = \sin(\pi m x/a), m = 1, 2, \ldots \tag{8.58}$$

Эти функции ортогональны. Подставим их в функционал (8.47) и выполним все дифференцирования и интегрирования. С учетом ортогональности получим

$$F(c_1, c_2, \ldots, c_M) = \sum_{m=1}^{M} \left[\frac{\varkappa}{a} \left(\frac{\pi m}{2} \right)^2 c_m^2 + c_m f_m \right] = \min,$$

$$f_m = \int_0^a f(x) \sin(\pi m x/a) dx.$$

Дифференцирование по c_m дает здесь систему линейных уравнений с диагональной матрицей. Поэтому сразу можно написать ее решение:

$$c_m = -2a f_m/(\varkappa \pi^2 m^2). \tag{8.59}$$

Такое упрощение возникает благодаря ортогональности базисной системы. Если вместо базиса (8.58) взять базис из неортогональных многочленов $\varphi_m(x) = x^m(a - x)$, также удовлетворяющих нулевым граничным условиям, то вместо (8.59) получится система с плотно заполненной матрицей (вдобавок она будет плохо обусловлена). Отсюда видно, что целесообразно использовать ортогональные базисы.

Сложные функционалы. Если функционал неквадратичный, то подстановка в него обобщенного многочлена (8.56) приведет к достаточно сложной неквадратичной функции F. Минимизирующая система уравнений для c_m будет нелинейной и так-

же достаточно сложной. В этом случае возникает много трудностей. Во-первых, у нелинейной системы решение может не существовать или быть неединственным. Во-вторых, для вычисления решения (итерационным методом Ньютона) требуется достаточно хорошее нулевое приближение. Поэтому обычно c_m удается вычислить лишь для небольших значений M. Но небольшое M может не обеспечить необходимой точности аппроксимации.

Поэтому, хотя метод Ритца формально обобщается на неквадратичные функционалы, фактически он к ним плохо применим.

Точность. Сходимость метода Ритца обосновывается из следующих соображений. Пусть система $\{\varphi_m(x)\}$ полна в том классе функций, которому принадлежит решение $u_{\min}(x)$. Тогда можно аппроксимировать $u_{\min}(x)$ некоторым обобщенным многочленом $\tilde{u}_M(x)$ со сколь угодно высокой точностью, если выбрать достаточно большое M. Естественно также считать, что функционал непрерывно и гладко зависит от функции. Тогда разность $\Phi[\tilde{u}_M(x)] - \Phi[u_{\min}(x)]$ будет неотрицательной и сколь угодно малой. При фиксированном M мы находим c_m и обобщенный многочлен $u_M(x)$ из (8.57). Поэтому на всех обобщенных многочленах с данным M функционал принимает наименьшее значение на $u_M(x)$. Следовательно, $\Phi[u_M(x)] \leqslant \Phi[\tilde{u}_M(x)]$, и справедливо двустороннее неравенство:

$$\Phi_{\min} = \Phi(u_{\min}) \leqslant \Phi[u_M(x)] \leqslant \Phi[\tilde{u}_M(x)].$$

Отсюда видно, что решение задачи $u_M(x)$, являющееся решением (8.57), аппроксимирует $u_{\min}(x)$ с достаточно высокой точностью при достаточно большом M.

Пространство обобщенных многочленов $u_M(x)$ можно рассматривать как M-мерное пространство коэффициентов c_m. Увеличивая M, получаем последовательность таких пространств, вложенных друг в друга. Поэтому последовательность значений $\Phi[u_M(x)]$ при увеличении M будет невозрастающей и стремящейся к $\Phi_{\min}$. Это означает сходимость метода.

Теоретически оценить скорость сходимости можно лишь в частных случаях. Однако можно провести серию расчетов с M, визуально наблюдать сходимость к некоторому предельному значению и попытаться численно оценить фактическую скорость сходимости. Практически такую методику удается применять лишь к квадратичным функционалам, так как лишь для них удается проводить вычисления с достаточно большим M.

8.5.4. Конечные элементы

Метод конечных элементов является частным случаем метода Ритца. В качестве базиса берутся функции $\varphi_m(x)$ на конечных носителях. Для их построения выбирают сетку $\{x_n, 0 \leqslant n \leqslant N\}$. В качестве носителя каждой базисной функции берут минимально возможное количество соседних интервалов сетки. По существу метод является гибридом метода Ритца и сеточного метода. Поэтому сейчас метод конечных элементов часто называют ***проекционно-сеточным методом***. Как и метод Ритца, он сравнительно просто реализуется лишь для квадратичных функционалов.

Рассмотрим простейший пример — задачу прогиба струны (8.47). От базисных функций здесь требуется непрерывность и гладкость за исключением отдельных точек (так как надо вычислять u_x). Простейшими подходящими конечными элементами являются B-сплайны первой степени, описанные в 5.5. На интервале сетки $[x_{n-1}, x_n]$ отличны от нуля только два линейных B-сплайна: $B_{n-1}(x) = (x - x_{n-1})/h_n$, $B_{n-2}(x) = (x_n - x)/h_n$, $x \in [x_{n-1}, x_n]$. Поэтому на данном интервале искомое решение имеет следующий вид:

$$\begin{aligned} u(x) &= c_{n-2}B_{n-2}(x) + c_{n-1}B_{n-1}(x) = \\ &= [c_{n-2}(x_n - x) + c_{n-1}(x - x_{n-1})]/h_n, \\ u_x &= (c_{n-1} - c_{n-2})/h_n, \quad x \in [x_{n-1}, x_n]. \end{aligned} \tag{8.60}$$

Подставляя (8.60) в интеграл (8.47) и суммируя по всем интервалам сетки, получим

$$\begin{gathered} F(c_{-1}, c_0, \ldots, c_{N-1}) \equiv \\ \equiv \sum_{n=1}^{N} \left[\frac{\varkappa}{2h_n}(c_{n-1} - c_{n-2})^2 + \frac{1}{h_n}(c_{n-2}\tilde{f}_n + c_{n-1}\bar{f}_n) \right] = \min, \\ \tilde{f}_n = \int\limits_{x_{n-1}}^{x_n} f(x)(x_n - x)dx, \quad \bar{f}_n = \int\limits_{x_{n-1}}^{x_n} f(x)(x - x_{n-1})dx. \end{gathered} \tag{8.61}$$

Напомним, что индексы линейных B-сплайнов меняются от -1 до $N - 1$.

Теперь надо минимизировать квадратичную функцию (8.61) от коэффициентов c_m. Приравняем нулю производную (8.61) по коэффициенту c_m и учтем, что каждый коэффициент входит в

два члена суммы. Поэтому получим систему линейных уравнений с трехдиагональной матрицей:

$$-\frac{\varkappa}{h_{n+1}}(c_n - c_{n-1}) + \frac{\varkappa}{h_n}(c_{n-1} - c_{n-2}) = \frac{\bar{f}_n}{h_n} + \frac{\tilde{f}_{n+1}}{h_{n+1}}.$$

Для этой системы выполнено условие преобладания диагонального элемента, так что она устойчиво решается методом Гаусса или прогонкой. Эта система очень похожа на ту, что получилась для сеточного метода в 8.5.2.

Сравним метод конечных элементов с обычным сеточным методом. В приведенном примере аппроксимирующие выражения имеют точность $O(h^2)$, как и в простейшем сеточном методе (см. подразд. 8.5.2). Для получения более высокой точности надо использовать конечные элементы более высокой гладкости — например, параболические или кубические B-сплайны. Но тогда все формулы существенно усложняются. Интеграл от u_x^2 удается вычислить аналитически, но интеграл от $u(x)f(x)$ аналитически вычислить, вообще говоря, уже невозможно. Для них нужно конструировать специальные прецизионные квадратурные формулы. В классическом же сеточном методе формулы высокого порядка точности строятся существенно легче. Для неквадратичных функционалов применение метода конечных элементов, как и вообще метода Ритца, крайне затруднительно. Для классических сеточных методов усложнение оказывается не столь значительным.

Поэтому следует рекомендовать для использования обычные сеточные методы. В настоящее время методы конечных элементов используются в прикладных расчетах достаточно часто. Но это объясняется тем, что для наиболее употребительных прикладных задач уже созданы широко доступные и хорошо отлаженные программные пакеты.

8.5.5. Пробные функции

Пусть из аналитического исследования исходной задачи приблизительно известен качественный вид искомого решения. Выберем некоторую функцию $v(x; c_1, \ldots, c_M)$, где зависимость от x качественно близка к ожидаемой, а коэффициенты c_m — свободные параметры. Подставляя выбранную функцию в функционал и проведя все дифференцирования и интегрирования по x, получим задачу на минимум функции от коэффициентов:

$$F(c_1, \ldots, c_M) \equiv \Phi[v(x; c_1, \ldots, c_M)] = \min. \qquad (8.62)$$

Остается ее решить и найти искомые коэффициенты. Этот метод похож на метод Ритца с одним отличием: зависимость пробной функции $v(x; c_1, \ldots, c_M)$ от коэффициентов выбирается обычно нелинейной. Тем самым функция $F(c_1, \ldots, c_M)$ в (8.62) не будет квадратичной даже для квадратичного функционала. Поэтому задача минимизации (8.62) оказывается достаточно сложной и на ее успешное решение можно рассчитывать лишь при небольшом числе коэффициентов M.

Метод пробных функций обычно применяют к достаточно сложным функционалам. В нем можно рассчитывать на нахождение решения с невысокой точностью.

Например, рассмотрим задачу атома (8.53). Из вида сеточного уравнения можно заметить, что $\rho(x) \approx (3z/5x)^{3/2}$ при $x \to 0$. Поэтому можно выбрать пробную функцию следующего вида:

$$\rho(x) \approx v(x; c_1, c_2, c_3) \equiv [(3z/5x)^3(1 + c_1 x + c_2 x^2 + c_3 x^3)]^{1/2}.$$

Эта функция имеет правильное качественное поведение при $x \to$ $\to 0$ и остается конечной при $x \to \infty$. В ней всего три свободных параметра. Поэтому решение $\rho(x)$ она позволяет найти лишь очень грубо. Однако значение минимизируемого функционала (энергию атома) удается определить с точностью $\sim 1\,\%$.

СПИСОК ЛИТЕРАТУРЫ

Основной

Бахвалов Н. С. Численные методы / Н. С. Бахвалов, Н. П. Жидков, Г. М. Кобельков. — М. : БИНОМ, Лаборатория знаний, 2004.

Калиткин Н. Н. Вычисления на квазиравномерных сетках / Н. Н. Калиткин [и др.] — М. : Физматлит, 2005.

Калиткин Н. Н. Численные методы. — М. : Наука, 1978.

Марчук Г. И. Повышение точности решения разностных схем / Г. И. Марчук, В. В. Шайдуров. — М. : Наука, 1979.

Плис А. И. Лабораторный практикум по высшей математике / А. И. Плис, Н. А. Сливина. — М. : Высшая школа, 1994.

Самарский А. А. Численные методы / А. А. Самарский, А. В. Гулин. — М. : Наука, 1989.

Дополнительный

Амосов А. А. Вычислительные методы для инженеров / А. А. Амосов, Ю. А. Дубинский, Н. В. Копченова. — М. : Изд-во МЭИ, 2003.

Артемьев С. С. Некоторые проблемы вычислительной математики / С. С. Артемьев, Г. В. Демидов. — Новосибирск : Наука, 1975.

Бабенко К. И. Основы численного анализа. — М. : Наука, 1986.

Бабенко К. И. Основы численного анализа / под ред. А. Д. Брюно. — Ижевск : РХД, 2002.

Бахвалов Н. С. Численные методы в задачах и упражнениях / Н. С. Бахвалов, А. В. Лапин, Е. В. Чижонков. — М. : Высшая школа, 2002.

Бахвалов Н. С. Численные методы, алгебра, обыкновенные дифференциальные уравнения. — М. : Наука, 1973.

Березин И. С. Методы вычислений: в 2 т. / И. С. Березин, Н. П. Жидков. — М. : Физматгиз, 1959.

Воеводин В. В. Математические модели и методы в параллельных процессах. — М. : Наука, 1986.

Воеводин В. В. Линейная алгебра. — М. : Наука, 1979.

Воеводин В. В. Вычислительные процессы с теплицевыми матрицами / В. В. Воеводин, Е. Е. Тыртышников. — М. : Наука, 1987.

Волков Е. А. Численные методы. — М. : Наука, 1987.

Годунов С. К. Введение в теорию разностных схем / С. К. Годунов, В. С. Рябенький. — М. : Физматлит, 1962.

Голуб Дж. Матричные вычисления / Дж. Голуб, Ч. Ван Лоун. — М. : Мир, 1999.

Гончаров В. Л. Теория интерполирования и приближения функций. — М. : Гостехиздат, 1954.

Дэннис Дж. Численные методы безусловной оптимизации и решения нелинейных уравнений / Дж. Дэннис, Р. Шнабель — М. : Мир, 1988.

Дорн У. Численные методы и программирование на Фортране / У. Дорн, Д. Мак-Кракен. — М. : Мир, 1970.

Дробышевич В. И. Задачи по вычислительной математике / В. И. Дробышевич, В. П. Дымников, Г. С. Гивин. — М. : Наука, 1980.

Доргатт Дж. Математические методы математических вычислений / Дж. Доргатт, М. Малькольм, К. Моулер. — М. : Мир, 1980.

Калиткин Н. Н. Введение в численный анализ / Н. Н. Калиткин, Н. А. Гольцов. — М. : Изд-во МГУЛ, 2003.

Коллатц Л. Численные методы решения дифференциальных уравнений. — М. : ИЛ, 1953.

Коллатц Л. Функциональный анализ и вычислительная математика. — М. : Мир, 1969.

Костомаров Д. П. Вводные лекции по численным методам / Д. П. Костомаров, А. П. Фаворский. — М. : Логос, 2004.

Крылов А. Н. Лекции о приближенных вычислениях. — М. : Наука, 1954.

Крылов В. И. Начала теории вычислительных методов: в 5 т. / В. И. Крылов, В. В. Бобков, П. И. Монастырный. — Минск : Наука и техника, 1982 — 1986. Т. 1. Дифференциальные уравнения, 1982; Т. 2. Интерполирование и интегрирование, 1983; Т. 3. Интегральные уравнения, некорректные задачи и улучшение сходимости, 1984;Т. 4. Линейная алгебра и нелинейные уравнения, 1985; Т. 5. Уравнения в частных производных, 1986.

Ланцош К. А. Практические методы прикладного анализа. — М. : Физматгиз, 1961.

Лебедев В. И. Функциональный анализ и вычислительная математика. — М. : Физматлит, 2000.

Ляшко И. И. Методы вычислений / И. И. Ляшко, В. Л. Макаров, А. А. Скоробогатько. — Киев : Высшая школа, 1972.

Макаров В. Л. Сплайн–аппроксимация функций / В. Л. Макаров, В. В. Хлобыстов. — М. : Высшая школа, 1983.

Марчук Г. И. Методы вычислительной математики. — М. :Наука, 1989.

Милн В. Э. Численное решение дифференциальных уравнений. — М. : ИЛ, 1955.

Моисеев Н. Н. Численные методы теории оптимального управления. — М. : Изд-во МГУ, 1968.

Ортега Дж. Введение в параллельные и векторные методы решения линейных систем. — М. : Мир, 1991.

Ортега Дж. Итерационные методы решения нелинейных систем уравнений со многими неизвестными / Дж. Ортега, В. Рейнболдт: пер. с англ. — М. : Мир, 1975.

Пирумов У. Г. Численные методы. — М. : Изд-во МАИ, 1998.

Рябенький В. С. Введение в вычислительную математику. — М. : Наука, 1994.

Самарский А. А. Введение в численные методы. — М. : Наука, 1987.

Самарский А. А. Математическое моделирование: Идеи, методы, примеры / А. А. Самарский, А. П. Михайлов. — М. : Наука, 1997.

Соболь И. М. Численные методы Монте-Карло. — М. : Наука, 1973.

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

Турчак Л. И. Основы численных методов / Л. И. Турчак, П. В. Плотников. — М. : Физматлит, 2002.

Уилкинсон Дж. Х. Алгебраическая проблема собственных значений. — М. : Наука, 1970.

Фаддеев Д. К. Вычислительные методы линейной алгебры / Д. К. Фаддеев, В. Н. Фаддеева. — М. : Физматиз, 1963.

Федоренко Р. П. Введение в вычислительную физику. — М. : Изд-во МФТИ, 1994.

Форсайт Дж. Машинные методы математических вычислений / Дж. Форсайт, М. Малькольм, К. Моулер. — М. : Мир, 1980.

Хемминг Р. В. Численные методы. — М. : Наука, 1972.

ОГЛАВЛЕНИЕ

Предисловие 3

Глава 1. **О численном анализе** 6

1.1. Немного истории 6
1.1.1. Развитие численных методов 6
1.1.2. Теории и модели 8
1.2. Математическое моделирование 10
1.2.1. Математическая модель 10
1.2.2. Модель — алгоритм — программа 16
1.3. Источники погрешности 17
1.3.1. Величины и нормы 18
1.3.2. Погрешность модели 20
1.3.3. Неустранимая погрешность 21
1.3.4. Погрешность метода 24
1.3.5. Погрешность округления 24
1.3.6. Корректность задачи 26

Глава 2. **Системы алгебраических уравнений** 28

2.1. Линейные системы 28
2.1.1. Задачи линейной алгебры 28
2.1.2. Метод Гаусса 30
2.1.3. Определитель и обратная матрица 34
2.1.4. Прочие методы 35
2.1.5. Плохо обусловленные системы 36
2.1.6. Переобусловленные системы 40
2.2. Нелинейное уравнение 41
2.2.1. Дихотомия 41
2.2.2. Метод Ньютона 44
2.2.3. Обобщенный метод Ньютона 48
2.2.4. Прочие методы 50
2.2.5. Удаление корней 52
2.3. Системы нелинейных уравнений 55
2.3.1. Метод Ньютона 55
2.3.2. Обобщенный метод Ньютона 58

Глава 3. **Численное интегрирование** 60

3.1. Квадратурные формулы 60
3.1.1. Интегральная сумма 60

3.1.2. Формула средних . . . 61
3.1.3. Формула трапеций . . . 65
3.1.4. Формула Симпсона . . . 66
3.1.5. Формулы Эйлера — Маклорена . . . 68
3.1.6. Формулы Гаусса — Кристоффеля . . . 72
3.1.7. Недостаточно гладкие функции . . . 79
3.2. Метод сгущения сеток . . . 80
3.2.1. Однократное сгущение . . . 80
3.2.2. Рекуррентное уточнение . . . 86
3.2.3. Квазиравномерные сетки . . . 91
3.2.4. Метод Эйткена . . . 101
3.3. Кубатурные формулы . . . 106
3.3.1. Метод средних . . . 106
3.3.2. Произведение квадратурных формул . . . 113
3.3.3. Статистические методы . . . 119

Глава 4. **Интерполяция** . . . 129

4.1. Интерполяционный многочлен . . . 129
4.1.1. Задачи интерполяции . . . 129
4.1.2. Многочлен Ньютона . . . 130
4.1.3. Погрешность . . . 134
4.1.4. Обратная интерполяция . . . 139
4.1.5. Эрмитова интерполяция . . . 140
4.1.6. Многомерная интерполяция . . . 143
4.2. Сплайн-интерполяция . . . 147
4.2.1. Историческая справка . . . 147
4.2.2. Кубический сплайн . . . 148
4.2.3. Обобщения . . . 153
4.3. Нелинейная интерполяция . . . 154
4.3.1. Выравнивание . . . 154
4.3.2. Рациональная интерполяция . . . 157

Глава 5. **Среднеквадратичная аппроксимация** . . . 160

5.1. Общий случай . . . 160
5.1.1. Выбор нормы . . . 160
5.1.2. Аппроксимация обобщенным многочленом . . . 162
5.1.3. Неортогональные базисы . . . 163
5.1.4. Ортогональные системы . . . 165
5.1.5. Метод наименьших квадратов . . . 167
5.2. Тригонометрический ряд Фурье . . . 170
5.2.1. Общие формулы . . . 170
5.2.2. Сходимость . . . 171
5.2.3. Вычисление коэффициентов . . . 174
5.2.4. О равномерных приближениях . . . 177
5.3. Ряды по многочленам Чебышева . . . 178
5.3.1. Многочлены $T_m(x)$. Вычисление . . . 178

5.3.2. Разложение по $T_m(x)$ 180
5.4. Метод двойного периода 184
5.4.1. Исключение разрывов 184
5.4.2. Двойной период 185
5.4.3. Наилучшее приближение 186
5.4.4. Вычисление скалярных произведений 191
5.5. Аппроксимация сплайнами 193
5.5.1. B-сплайны 193
5.5.2. Среднеквадратичная аппроксимация 197
5.5.3. Конечные элементы 202
5.6. Аппроксимация кривых 202
5.6.1. Параметризация кривой 202
5.6.2. Хорда 204
5.6.3. Окружность 205
5.6.4. Аппроксимация 209
5.6.5. Ротационная инвариантность 211

Глава 6. **Численное дифференцирование** 215

6.1. Производная многочлена Ньютона 215
6.1.1. Общие формулы 215
6.1.2. Простейшие случаи 217
6.1.3. Неограниченная область 219
6.1.4. Сгущение сеток 221
6.1.5. Старшие производные 222
6.2. Дифференцирование иных аппроксимаций 223
6.2.1. Интерполяционный сплайн 223
6.2.2. Метод выравнивания 224
6.2.3. Среднеквадратичное приближение 225
6.3. Некорректность численного дифференцирования 229
6.3.1. Дифференцирование интерполяционного многочлена 229
6.3.2. Дифференцирование рядов 232

Глава 7. **Спектр матрицы** 234

7.1. Преобразование подобия 234
7.1.1. Теория 234
7.1.2. Метод отражений 239
7.1.3. Другие методы 244
7.2. Вычисление спектра 246
7.2.1. Частичная проблема 246
7.2.2. Обобщенная проблема 251
7.2.3. Полная проблема 252

Глава 8. **Задачи минимизации** 254
8.1. Одномерный минимум 254
8.1.1. Золотое сечение 254
8.1.2. Метод Ньютона 257
8.1.3. Случай многих экстремумов 259
8.2. Многомерный минимум 260
8.2.1. Рельеф функции 260
8.2.2. Обобщенный метод Ньютона 262
8.2.3. Многоэкстремальность 264
8.3. Решение сеточных уравнений 265
8.3.1. Градиентные спуски 265
8.3.2. Наискорейший спуск 266
8.3.3. Минимальные невязки 268
8.3.4. Усеченный спуск 269
8.3.5. Сопряженные градиенты 270
8.3.6. Нелинейность 272
8.4. Задачи с ограничениями 272
8.4.1. Наложение связей 272
8.4.2. Ограниченная область 274
8.4.3. Общий случай 278
8.5. Минимизация функционала 280
8.5.1. Прикладные проблемы 280
8.5.2. Сеточный метод 282
8.5.3. Метод Ритца 287
8.5.4. Конечные элементы 289
8.5.5. Пробные функции 290
Список литературы 293

Учебное издание

Калиткин Николай Николаевич,
Альшина Елена Александровна

ЧИСЛЕННЫЕ МЕТОДЫ

В двух книгах

КНИГА 1

ЧИСЛЕННЫЙ АНАЛИЗ

Учебник

Редактор *Л. В. Честная*
Технический редактор *Н. И. Горбачева*
Компьютерная верстка: *Т. А. Клименко*
Корректоры *Г. Н. Петрова*, *В. А. Жилкина*

Изд. № 101114030. Подписано в печать 22.05.2013. Формат 60 × 90/16.
Бумага офс. № 1. Печать офсетная. Гарнитура «Таймс». Усл. печ. л. 19,0.
Тираж 1 200 экз. Заказ № 34417.

ООО «Издательский центр «Академия». www.academia-moscow.ru
129085, Москва, пр-т Мира, 101В, стр. 1.
Тел./факс: (495) 648-0507, 616-00-29.
Санитарно-эпидемиологическое заключение № РОСС RU. АЕ51. Н 16476 от 05.04.2013.

Отпечатано в соответствии с качеством предоставленных издательством
электронных носителей в ОАО «Саратовский полиграфкомбинат».
410004, г. Саратов, ул. Чернышевского, 59. www.sarpk.ru

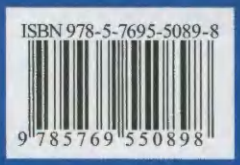
ЧИСЛЕННЫЕ МЕТОДЫ

В двух книгах

Книга 1

ЧИСЛЕННЫЙ АНАЛИЗ

ISBN 978-5-7695-5089-8



Издательский центр «Академия»
www.academia-moscow.ru