ГЛАВА I

ЧТО ТАКОЕ ЧИСЛЕННЫЕ МЕТОДЫ?

Глава I является вводной. В § 1 рассмотрены роль математики при решении физико-технических задач и место численных методов среди других математических методов и кратко изложена история численных методов. В § 2 разобраны основные понятия приближенного анализа: корректность постановки задач, определение близости точного и приближенного решений, структура погрешности.

§ 1. Математические модели и численные методы

1. Решение задачи. Физиков математика интересует не сама по себе, а как средство решения физических задач. Рассмотрим поэтому, как решается любая реальная задача — например, нахождение светового потока конструируемой лампы, производительности проектируемой химической установки или себестоимости продукции строящегося завода.

Одним из способов решения является эксперимент. Построим эту лампу, установку или завод и измерим интересующую нас характеристку. Если характеристика оказалась неудачной, то изменим проект и построим новый завод и т. д. Ясно, что мы получим достоверный ответ на вопрос, но слишком медленным и дорогим способом.

Другой способ — математический анализ конструкции или явления. Но такой анализ применяется не к реальным явлениям, а к некоторым математическим моделям этих явлений. Поэтому первая стадия работы — это *формулировка математической модели* (постановка задачи). Для физического процесса модель обычно состоит из уравнений, описывающих процесс; в эти уравнения в виде коэффициентов входят характеристики тел или веществ, участвующих в процессе. Например, скорость ракеты при вертикальном полете в вакууме определяется уравнением

$$\left(M-\int_0^t m(\tau)\,d\tau\right)\left(\frac{dv}{dt}+g\right)=cm(t), \qquad (1)$$

где M — начальная масса ракеты, $m(t)$ — заданный расход горю-

чего, g — ускорение поля тяготения, а c — скорость истечения газов, зависящая от калорийности топлива и среднего молекулярного веса продуктов сгорания.

Любое изучаемое явление бесконечно сложно. Оно связано с другими явлениями природы, возможно, не представляющими интереса для рассматриваемой задачи. Математическая модель должна охватывать важнейшие для данной задачи стороны явления. Наиболее сложная и ответственная работа при постановке задачи заключается в выборе связей и характеристик явления, существенных для данной задачи и подлежащих формализации и включению в математическую модель.

Если математическая модель выбрана недостаточно тщательно, то, какие бы методы мы ни применяли для расчета, все выводы будут недостаточно надежны, а в некоторых случаях могут оказаться совершенно неправильными. Так, уравнение (1) непригодно для запуска ракеты с поверхности Земли, ибо в нем не учтено сопротивление воздуха.

Вторая стадия работы — это *математическое исследование*. В зависимости от сложности модели применяются различные математические подходы. Для наиболее грубых и несложных моделей зачастую удается получить аналитические решения; это излюбленный путь многих физиков-теоретиков. Например, уравнение (1) легко интегрируется при $g = \text{const}$ и $m(t) = \text{const}$:

$$v = c \ln [M/(M - mt)] - gt.$$

Из-за грубости модели физическая точность этого подхода невелика; нередко такой подход позволяет оценить лишь порядки величин.

Для более точных и сложных моделей аналитические решения удается получить сравнительно редко. Обычно теоретики пользуются приближенными математическими методами (например, разложением по малому параметру), позволяющими получить удовлетворительные качественные и количественные результаты. Наконец, для наиболее сложных и точных моделей основными методами решения являются численные; как правило, они требуют проведения расчетов на ЭВМ. Эти методы зачастую позволяют добиться хорошего количественного описания явления, не говоря уже о качественном.

Во всех случаях математическая точность решения должна быть несколько (в 2 — 4 раза) выше, чем ожидаемая физическая точность модели. Более высокой математической точности добиваться бессмысленно, ибо общую точность ответа это все равно не повысит. Но более низкая математическая точность недопустима (для облегчения решения задачи нередко в ходе работы делают дополнительные математические упрощения; это снижает ценность результатов).

Наконец, третья стадия работы — это *осмысливание математического решения* и сопоставление его с экспериментальными данными. Если расчеты хорошо согласуются с контрольными экспериментами, то это свидетельствует о правильном выборе модели; такую модель можно использовать для расчета процессов данного типа. Если же расчет и эксперимент не согласуются, то модель необходимо пересмотреть и уточнить.

2. Численные методы являются одним из мощных математических средств решения задачи. Простейшие численные методы мы используем всюду, например, извлекая квадратный корень на листке бумаги. Есть задачи, где без достаточно сложных численных методов не удалось бы получить ответа; классический пример — открытие Нептуна по аномалиям движения Урана.

В современной физике таких задач много. Более того, часто требуется выполнить огромное число действий за короткое время, иначе ответ будет не нужен. Например, суточный прогноз погоды должен быть вычислен за несколько часов; коррекцию траектории ракеты надо рассчитать за несколько минут (напомним, что для расчета орбиты Нептуна Леверье потребовалось полгода); режим работы прокатного стана должен исправляться за секунды. Это немыслимо без мощных ЭВМ, выполняющих тысячи или даже миллионы операций в секунду.

Современные численные методы и мощные ЭВМ дали возможность решать такие задачи, о которых полвека назад могли только мечтать. Но применять численные методы далеко не просто. Цифровые ЭВМ умеют выполнять только арифметические действия и логические операции. Поэтому помимо разработки математической модели, требуется еще разработка алгоритма, сводящего все вычисления к последовательности арифметических и логических действий. Выбирать модель и алгоритм надо с учетом скорости и объема памяти ЭВМ: чересчур сложная модель может оказаться машине не под силу, а слишком простая — не даст физической точности.

Сам алгоритм и программа для ЭВМ должны быть тщательно проверены. Даже проверка программы нелегка, о чем свидетельствует популярное утверждение: «В любой сколь угодно малой программе есть по меньшей мере одна ошибка». Проверка алгоритма еще более трудна, ибо для сложных алгоритмов не часто удается доказать сходимость классическими методами. Приходится использовать более или менее надежные «экспериментальные» проверки, проводя пробные расчеты на ЭВМ и анализируя их (смотри, например, главу IX, § 4, п. 3).

Строгое математическое обоснование алгоритма редко бывает исчерпывающим исследованием. Например, большинство доказательств сходимости итерационных процессов справедливо только при точном выполнении всех вычислений; практически же число

сохраняемых десятичных знаков редко происходит 5—6 при «ручных» вычислениях и 10—12 при вычислениях на ЭВМ. Плохо поддаются теоретическому исследованию «маленькие хитрости» — незначительные на первый взгляд детали алгоритма, сильно влияющие на его эффективность. Поэтому окончательную оценку метода можно дать только после опробования его в практических расчетах.

К чему приводит пренебрежение этими правилами — видно из принципа некомпетентности Питера: «ЭВМ многократно увеличивает некомпетентность вычислителя».

Для сложных задач разработка численных методов и составление программ для ЭВМ очень трудоемки и занимают от нескольких недель до нескольких лет. Стоимость комплекса отлаженных программ нередко сравнима со стоимостью экспериментальной физической установки. Зато проведение отдельного расчета по такому комплексу много быстрей и дешевле, чем проведение отдельного эксперимента. Такие комплексы позволяют подбирать оптимальные параметры исследуемых конструкций, что не под силу эксперименту.

Однако численные методы не всесильны. Они не отменяют все остальные математические методы. Начиная исследовать проблему, целесообразно использовать простейшие модели, аналитические методы и прикидки. И только разобравшись в основных чертах явления, надо переходить к полной модели и сложным численным методам; даже в этом случае численные методы выгодно применять в комбинации с точными и приближенными аналитическими методами.

Современный физик или инженер-конструктор для успешной работы должен одинаково хорошо владеть и «классическими» методами, и численными методами математики.

3. История прикладной математики. Раздел математики, имеющий дело с созданием и обоснованием численных алгоритмов для решения сложных задач различных областей науки, часто называют прикладной математикой; американцы применение численных методов к физическим задачам называют вычислительной физикой. Главная задача прикладной математики — фактическое нахождение решения с требуемой точностью; этим она отличается от классической математики, которая основное внимание уделяет исследованию условий существования и свойств решения.

В истории прикладной математики можно выделить три основных периода.

Первый начался 3—4 тысячи лет назад. Он был связан с ведением конторских книг, вычислением площадей и объемов, расчетами простейших механизмов; иными словами — с несложными задачами арифметики, алгебры и геометрии. Вычислительными средствами служили сначала собственные пальцы, а затем

— счеты. Исходные данные содержали мало цифр, и большинство выкладок выполнялось точно, без округлений.

Второй период начался с Ньютона. В этот период решались задачи астрономии, геодезии и расчета механических конструкций, сводящиеся либо к обыкновенным дифференциальным уравнениям, либо к алгебраическим системам с большим числом неизвестных. Вычисления выполнялись с округлением; нередко от результата требовалась высокая точность, так что приходилось сохранять до 8 значащих цифр.

Вычислительные средства стали разнообразнее: таблицы элементарных функций, затем — арифмометр и логарифмическая линейка; к концу этого периода появились неплохие клавишные машины с электромотором. Но скорость всех этих средств была невелика, и вычисления занимали дни, недели и даже месяцы.

Третий период начался примерно с 1940 г. Военные задачи — например, наводка зенитных орудий на быстро движущийся самолет — требовали недоступных человеку скоростей и привели к разработке электронных систем. Появились электронные вычислительные машины (ЭВМ).

Скорость даже простейших ЭВМ настолько превосходила скорость механических средств, что стало возможным проводить вычисления огромного объема. Это позволило численно решать новые классы задач; например, процессы в сплошных средах, описывающиеся уравнениями в частных производных.

Сначала для решения эти задач использовались численные методы, разработанные в «доэлектронный» период. Но применение ЭВМ быстро привело к переоценке методов. Многие старые методы оказались неподходящими для автоматизированных расчетов. Стали быстро разрабатываться новые методы, ориентированные прямо на ЭВМ (например, метод Монте-Карло).

Мощности ЭВМ быстро растут. Если в 50-е гг. в СССР вступила в строй первая «Стрела» со скоростью 2000 операций в секунду и памятью 1024 ячейки, то сейчас во многих вычислительных центрах страны работают БЭСМ-6 со скоростью в 300 раз больше и памятью в 30 раз больше. А наилучшие современные ЭВМ имеют скорость до 30 миллионов операций в секунду при практически неограниченной оперативной памяти с прямой адресацией. Становятся возможными расчеты все более сложных задач. Это служит стимулом для разработки новых численных методов.

§ 2. Приближенный анализ

1. Понятие близости. Если требуется определить некоторую величину y по известной величине x, то символически задачу можно записать в виде $y = A(x)$. Здесь и y, и x могут быть числами, совокупностью чисел, функцией одного или нескольких

переменных, набором функций и т.д. Если оператор A настолько сложен, что решение не удается явно выписать или точно вычислить, то задачу решают приближенно.

Например, пусть надо вычислить $y=\int_a^b x(t)\,dt$. Можно приближенно заменить $x(t)$ многочленом $\bar{x}(t)$ или другой функцией, интеграл от которой легко вычислить. А можно заменить интеграл суммой $\sum_i x(t_i)\,\Delta t_i$, вычислить которую тоже несложно. Таким образом, приближенный метод заключается в замене исходных данных на близкие данные $\bar{x}$ и (или) замене оператора на близкий оператор $\bar{A}$, так чтобы значение $\bar{y}=\bar{A}(\bar{x})$ легко вычислялось. При этом мы ожидаем, что значение $\bar{y}$ будет близко к искомому решению.

Но что такое «близко»? Очевидно, для двух чисел x_1 и x_2 надо требовать малости $|x_1-x_2|$; а близость двух функций можно определить разными способами. Эти вопросы рассматриваются в функциональном анализе, некоторые понятия которого будут сейчас изложены.

Множество элементов x любой природы называется *метрическим пространством,* если в нем введено расстояние $\rho(x_1, x_2)$ между любой парой элементов (*метрика*), удовлетворяющее следующим аксиомам:

$$\begin{aligned}
&\text{а) } \rho(x_1, x_2) \text{ — вещественное неотрицательное число,}\\
&\text{б) } \rho(x_1, x_2)=0, \text{ только если } x_1=x_2,\\
&\text{в) } \rho(x_1, x_2)=\rho(x_2, x_1),\\
&\text{г) } \rho(x_1, x_3)\leqslant\rho(x_1, x_2)+\rho(x_2, x_3).
\end{aligned}\qquad (2)$$

Последовательность элементов x_n метрического пространства называется *сходящейся* (по метрике) к элементу x, если $\rho(x_n, x)\to 0$ при $n\to\infty$. Последовательность x_n называется *фундаментальной*, если для любого $\varepsilon>0$ найдется такое $k(\varepsilon)$, что $\rho(x_n, x_m)<\varepsilon$ при всех n и $m>k$.

Метрическое пространство называют *полным*, если любая фундаментальная последовательность его элементов сходится к элементу того же пространства. Примером неполного пространства является множество рациональных чисел $x=(n/m)$ с метрикой $\rho(x_1, x_2)=|x_1-x_2|$. Последовательность $x_k=(1+1/k)^k$ ему принадлежит, является фундаментальной, а сходится к иррациональному числу e, т.е. не к элементу данного пространства. Если переменные y, x принадлежат неполным пространствам, то обосновать сходимость численных методов очень трудно: даже если удается доказать, что при $x_n\to x$ последовательность y_n фундаментальная, то отсюда еще не следует, что она сходится к элементу данного пространства, т. е. к решению допустимого класса.

Элементами наших множеств будут числа, векторы, матрицы, функции и т. п. Сами множества обычно являются линейными нормированными пространствами, ибо в них определены операции сложения элементов и умножения их на число и введена *норма* каждого элемента $\|x\|$, причем выполнены следующие аксиомы:

$$x_1+x_2=x_2+x_1,\quad (x_1+x_2)+x_3=x_1+(x_2+x_3);$$

существует единственный элемент θ такой, что $x+\theta=x$ для любого x (будем использовать для θ обозначение 0); для всякого x существует единственный элемент $-x$ такой, что $x+(-x)=\theta$; (3)

$$a(x_1+x_2)=ax_1+ax_2;\quad (a+b)x=ax+bx;$$

$a(bx)=(ab)x$; $1\cdot x=x$; $0\cdot x=\theta$ единствен; $\|x\|\geqslant 0$ — вещественное число; $\|ax\|=|a|\cdot\|x\|$; $\|x\|=0$ только при $x=0$; $\|x_1+x_2\|\leqslant\|x_1\|+\|x_2\|$.

Линейное нормированное пространство есть частный случай метрического пространства, а норма определяется метрикой. Полное линейное нормированное пространство называется *банаховым*. Практически всегда величины, с которыми мы будем оперировать, являются элементами банаховых пространств; это важно при доказательстве сходимости численных методов.

Рассмотрим некоторые примеры банаховых пространств, с которыми нам часто придется встречаться. Выполнимость аксиом (3) и полноту читатели легко проверят сами.

а) Множество всех действительных чисел с нормой $\|x\|=|x|$.

б) Пространство C — множество функций $x(t)$, определенных и непрерывных при $0\leqslant t\leqslant 1$, с чебышевской нормой $\|x\|_c=\max|x(t)|$. Сходимость в этом пространстве называется *равномерной*. Условие $0\leqslant t\leqslant 1$ здесь и в следующем примере принято для удобства; оно не является существенным, и можно определять функции на любом конечном отрезке.

Класс непрерывных функций часто еще сужают, накладывая на функции дополнительные требования: липшиц-непрерывности, однократной или многократной дифференцируемости и т. д. Напомним некоторые определения.

Функция $x(t)$ называется *равномерно-непрерывной* на отрезке, если для сколь угодно малого $\omega>0$ найдется такое δ, что $|x(t_1)-x(t_2)|\leqslant\omega$ для любой пары точек отрезка, удовлетворяющих условию $|t_1-t_2|\leqslant\delta$. Таким образом, устанавливается функциональная связь между ω и δ. Величина $\omega(\delta)$ называется *модулем непрерывности* функции. Функция, непрерывная во всех

точках замкнутого отрезка $a \leqslant t \leqslant b$, является на этом отрезке ограниченной и равномерно-непрерывной (теорема Кантора); следовательно, пространство C — множество ограниченных и равномерно-непрерывных функций. Если $\omega(\delta) \leqslant K\delta$, где K — некоторая константа, то функцию называют *липшиц-непрерывной*. Нетрудно видеть, что если функция имеет ограниченную производную, то она липшиц-непрерывна, причем $K = \sup |x'(t)|$.

в) Пространство L_p — множество функций $x(t)$, определенных при $0 \leqslant t \leqslant 1$ и интегрируемых по модулю с p-й степенью, если норма определена

$$\| x \|_{L_p} = \left[\int_0^1 |x(t)|^p \, dt \right]^{1/p}.$$

Сходимость в такой норме называют сходимостью *в среднем*. Пространство L_2 называют *гильбертовым*, а сходимость в нем — *среднеквадратичной*.

Разницу между равномерной близостью и близостью в среднем иллюстрирует рис. 1. Функция x_2 равномерно близка к функции x_1, а функция x_3 близка в среднем, т. е. мало отличается от x_1 на большей части отрезка, но может сильно отличаться от нее на небольших участках.

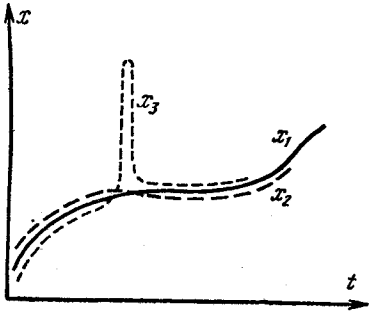

Рис. 1.

Выбирая метрические пространства, т. е. выбирая множества X, Y и определяя в них метрики, мы тем самым уславливаемся, в каких классах функций можно брать начальные данные и искать решение. Поэтому в конкретной задаче выбор пространств должен в первую очередь определяться физическим смыслом задачи, и лишь во вторую — чисто математическими соображениями (такими, например, как возможность доказать сходимость). Например, при расчете прочности самолета нужна равномерная близость приближенного решения к точному, а близости в среднем недостаточно: перенапряжение в маленьком участке может разрушить конструкцию. А в задаче о нагреве тела потоком тепла даже норма L_1 удовлетворительна, ибо температура тела определяется интегралом от потока по времени.

Нетрудно показать, что между разными нормами (если они существуют) выполняются определенные соотношения. Если функции $x(t)$ определены при $0 \leqslant t \leqslant 1$, тогда

$$\| x(t) \|_{L_1} \leqslant \| x(t) \|_{L_2} \leqslant \ldots \leqslant \| x(t) \|_C. \qquad (4)$$

В самом деле, например:

$$\| x(t) \|_{L_p}^p = \int_0^1 | x(t) |^p \, dt \leqslant \int_0^1 \max | x(t) |^p \, dt = \max | x(t) |^p = \| x(t) \|_C^p.$$

Следовательно, из равномерной сходимости вытекает сходимость в среднем, в частности — среднеквадратичная. Поэтому чебышевскую норму называют *более сильной*, чем гильбертова.

г) Координатные бесконечномерные пространства, элементами которых являются счетные множества чисел $x = \{x_1, x_2, \ldots\}$. По аналогии с пространствами функций, в них обычно вводят норму $\| x \|_c = \sup | x_i |$ или

$$\| x \|_{l_p} = \lim_{n \to \infty} \left(\frac{1}{n} \sum_{i=1}^{n} | x_i |^p \right)^{1/p},$$

а само пространство называют соответственно c или l_p.

д) Конечномерные пространства $c^{(n)}$, $l_p^{(n)}$, элементами которых являются группы из n чисел $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$; их можно считать координатами векторов в n-мерном пространстве, $l_2^{(n)}$ называют евклидовым. Нормы векторов вводят по аналогии со случаем (г), например,

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

Для конечномерных векторов между разными нормами существуют соотношения

$$\| \boldsymbol{x} \|_1 \leqslant \| \boldsymbol{x} \|_2 \leqslant \| \boldsymbol{x} \|_c \leqslant \sqrt{n} \, \| \boldsymbol{x} \|_2 \leqslant n \, \| \boldsymbol{x} \|_1, \tag{5}$$

которые легко проверить. Поэтому из сходимости в одной из этих норм следует сходимость во всех остальных нормах. Нормы, обладающие таким свойством, называют *эквивалентными*.

Отметим, что если последовательность векторов $\boldsymbol{x}_m$ не сходится, но $\boldsymbol{x}_m / \| \boldsymbol{x}_m \|$ сходится, то говорят о сходимости векторов *по направлению*.

е) В пространстве квадратных матриц порядка n наиболее употребительны следующие нормы:

$$\begin{gathered} \| A \|_c = \max_i \left(\sum_{j=1}^{n} | a_{ij} | \right), \quad \| A \|_1 = \max_j \left(\sum_{i=1}^{n} | a_{ij} | \right), \\ \| A \|_M = n \cdot \max_{i,j} | a_{ij} |, \qquad \| A \|_E = \left(\sum_{i,j=1}^{n} | a_{ij} |^2 \right)^{1/2}, \\ \| A \|_2 = \sqrt{\max \mu_i}, \end{gathered} \tag{6}$$

где μ_i — собственные значения эрмитовой матрицы $A^H A$ (здесь A^H — матрица, эрмитово сопряженная по отношению к A). Первые две нормы не имеют специальных названий, третья называется максимальной, четвертая — сферической или евклидовой и пятая — спектральной. Между ними выполняются некоторые соотношения, аналогичные (5).

Интересна связь между нормами матриц и векторов, на которые матрицы действуют. Норма матрицы называется *согласованной* с нормой вектора, если $\|Ax\| \leqslant \|A\| \cdot \|x\|$. Наименьшая из норм матрицы, согласованных с данной нормой вектора: $\|A\| = \sup (\|Ax\| / \|x\|)$, называется нормой матрицы, *подчиненной* данной норме вектора.

Приведем пример подчиненной нормы. Из цепочки неравенств

$$\|Ax\|_c = \max_i \left| \sum_{j=1}^{n} a_{ij} x_j \right| \leqslant \max_i \left[\left(\max_j |x_j| \right) \sum_{k=1}^{n} |a_{ik}| \right] =$$

$$= \|x\|_c \cdot \max_i \left(\sum_{k=1}^{n} |a_{ik}| \right) = \|A\|_c \cdot \|x\|_c \qquad (7)$$

следует, что $\|A\|_c$ согласована с $\|x\|_c$. Кроме того, для любой матрицы A существует такой вектор x, что неравенство (7) обращается в равенство. Для его нахождения положим $x_j = \pm 1$; знаки выберем так, чтобы они совпадали со знаками элементов a_{ij} той строки матрицы i, в которой $\sum_{j=1}^{n} |a_{ij}|$ максимальна. Тогда именно сумма по этой строке будет максимальна в левой части (7), и неравенство превратится в равенство. Это означает, что $\|A\|_c$ есть наименьшая из норм, согласованных с $\|x\|_c$: если мы возьмем еще меньшую $\|A\|$, то при этом векторе x для нее знак неравенства (7) будет обратным, т. е. она не будет согласованной. Следовательно, $\|A\|_c$ подчинена $\|x\|_c$.

Без доказательства укажем, что $\|A\|_1$ подчинена $\|x\|_1$, и спектральная норма подчинена $\|x\|_2$. Сферическая норма согласована с $\|x\|_2$, а максимальная норма согласована со всеми рассмотренными выше векторными нормами.

2. Структура погрешности. Есть четыре источника погрешности результата: математическая модель, исходные данные, приближенный метод и округления при вычислениях. Погрешность математической модели связана с физическими допущениями и здесь рассматриваться не будет.

Исходные данные зачастую неточны; например, это могут быть экспериментально измеренные величины. В прецизионных физических измерениях точность доходит до 10^{-12}, но уже характерная астрономическая и геодезическая точность равна 10^{-6}, а во многих физических и технических задачах погрешность измерения бывает 1 — 10%. Погрешность исходных данных δx приводит к так называемой *неустранимой* (она не зависит от математика) погрешности решения $\delta y = A(x + \delta x) - A(x)$. В следующем пункте будут рассмотрены случаи, когда неустранимая погрешность может становиться недопустимо большой.

Погрешность метода связана с тем, что точные оператор и исходные данные заменяются приближенными. Например, заменяют интеграл суммой, производную — разностью, функцию — многочленом или строят бесконечный итерационный процесс и обрывают его после конечного числа итераций. Методы строятся обычно так, что в них входит некоторый параметр; при стремлении параметра к определенному пределу погрешность метода стремится к нулю, так что эту погрешность можно регулировать. Погрешность метода мы будем исследовать при рассмотрении конкретных методов.

Погрешность метода целесообразно выбирать так, чтобы она была в $2-5$ раз меньше неустранимой погрешности. Бóльшая погрешность метода снижает точность ответа, а заметно меньшая — невыгодна, ибо это обычно требует значительного увеличения объема вычислений.

Вычисления как на бумаге, так и на ЭВМ выполняют с определенным числом значащих цифр. Это вносит в ответ *погрешность округления*, которая накапливается в ходе вычислений.

Рассмотрим накопление погрешности при простейших вычислениях. Пусть исходные данные x_i известны с относительной погрешностью $\Delta_i > 0$, т. е. заключены между $x_i(1-\Delta_i)$ и $x_i(1+\Delta_i)$; их абсолютные погрешности равны $\Delta_i|x_i|$. Тогда при сложении или вычитании двух чисел результат равен $x_1 \pm x_2$ с абсолютной погрешностью не более $\Delta_1|x_1|+\Delta_2|x_2|$, т. е. при этих операциях абсолютные погрешности складываются. При умножении (делении) результат равен x_1x_2 (x_1/x_2) с относительной погрешностью не более $\Delta_1+\Delta_2$, т. е. складываются относительные погрешности. На современных ЭВМ числа записываются с 10—12 десятичными знаками, поэтому в расчете на них погрешность единичного округления $\Delta = 10^{-10} \div 10^{-12}$ обычно пренебрежимо мала по сравнению с погрешностью метода и неустранимой погрешностью.

При решении больших задач выполняются миллиарды действий. Казалось бы, начальные ошибки возрастут в 10^9 раз и погрешность ответа будет огромной. Однако при отдельных действиях фактические погрешности чисел могут иметь разные знаки и компенсировать друг друга. Согласно статистике при N одинаковых действиях среднее значение суммарной ошибки превышает единичную примерно в $\sqrt{N}$ раз, а вероятность заметного уклонения суммарной ошибки от среднего значения очень мала. Значит, если нет систематических причин, то случайное накопление ошибок не слишком существенно.

Систематические причины возникают, например, если алгоритм таков, что в нем есть вычитание близких по величине чисел: хотя абсолютная ошибка при этом невелика, относительная ошибка $\Delta = (\Delta_1|x_1|+\Delta_2|x_2|)/(x_1-x_2)$ может стать большой. Например, при нахождении корней квадратного уравнения по

обычной формуле

$$x^2+px-q=0, \quad x_{1,2}=-0{,}5p\pm\sqrt{0{,}25p^2+q}$$

в случае, когда $0<q\ll p$, относительная ошибка округления для положительного корня x_1 велика. Это надо заранее предусмотреть и преобразовать формулу так, чтобы избавиться от подобных вычитаний:

$$x_1=q/(0{,}5p+\sqrt{0{,}25p^2+q}).$$

Этот пример очень прост. Существуют гораздо более сложные алгоритмы, где ошибки округления очень опасны: например, нахождение корней многочлена очень высокой степени (глава V, § 2, п.8) или итерационное решение разностных схем для эллиптических уравнений при помощи чебышевского набора параметров (глава XII, § 1). В этих случаях только после серьезного исследования удалось так видоизменить алгоритм, чтобы довести ошибки округления до безопасного уровня.

Отметим, что в большинстве подобных задач неприятностей можно избежать, проводя расчет с двойной или тройной точностью. Такая возможность реализована в хороших математических обеспечениях ЭВМ; это в несколько раз увеличивает время расчета, зато позволяет пользоваться уже известными алгоритмами, а не разрабатывать новые.

При любых расчетах справедливо правило: надо удерживать столько значащих цифр, чтобы погрешность округления была существенно меньше всех остальных погрешностей.

3. Корректность. Задача $y=A(x)$ называется *корректно поставленной*, если для любых входных данных x из некоторого класса решение y существует, единственно и устойчиво по входным данным. Рассмотрим это определение подробнее.

Чтобы численно решать задачу $y=A(x)$, надо быть уверенным в том, что искомое решение существует. Естественно также требовать единственности решения точной задачи: численный алгоритм — однозначная последовательность действий, и она может привести к одному решению. Но этого мало.

Нас интересует решение y, соответствующее входным данным x. Но реально мы имеем входные данные с погрешностью $x+\delta x$ и находим $y+\delta y=A(x+\delta x)$. Следовательно, неустранимая погрешность решения равна $\delta y=A(x+\delta x)-A(x)$. Если решение непрерывно зависит от входных данных, т. е. всегда $\|\delta y\|\to 0$ при $\|\delta x\|\to 0$, то задача называется *устойчивой* по входным данным; в противном случае задача неустойчива по входным данным.

Рассмотрим классический пример неустойчивости — задачу Коши для эллиптического уравнения в полуплоскости $y\geqslant 0$:

$$u_{xx}+u_{yy}=0, \quad u(x,\ 0)=0, \quad u_y(x,\ 0)=\varphi(x). \tag{8}$$

Входными данными является $\varphi(x)$. Если $\bar{\varphi}(x) = 0$, то задача имеет только тривиальное решение $\bar{u}(x, y) = 0$. Если же $\varphi_n(x) = \frac{1}{n}\cos nx$, то решением будет

$$u_n(x, y) = \frac{1}{n^2}\cos nx \cdot \operatorname{sh} ny.$$

Очевидно, $\varphi_n(x)$ равномерно сходятся к $\bar{\varphi}(x)$ при $n \to \infty$; но при этом если $y \neq 0$, то $u_n(x, y)$ неограничено и никак не может сходиться к $\bar{u}(x, y)$. Этот пример связан с физической задачей о тяжелой жидкости, налитой поверх легкой; при этом действительно возникает так называемая релей-тейлоровская неустойчивость.

Отсутствие устойчивости обычно означает, что даже сравнительно небольшой погрешности δx соответствует весьма большое δy, т. е. получаемое в расчете решение будет далеко от искомого. Непосредственно к такой задаче численные методы применять бессмысленно, ибо погрешности, неизбежно появляющиеся при численном расчете, будут катастрофически нарастать в ходе вычислений.

Правда, сейчас развиты методы решения многих некорректных задач. Но они основаны на решении не исходной задачи, а близкой к ней вспомогательной корректно поставленной задачи, содержащей параметр α; при $\alpha \to 0$ решение вспомогательной задачи должно стремиться к решению исходной задачи. Примеры таких методов (называемых регуляризацией) даны в следующих двух главах, а их строгое обоснование приведено в главе XIV, § 2.

На практике даже не всякую устойчивую задачу легко решить. Пусть $\|\delta y\| \leqslant C\|\delta x\|$, причем константа C очень велика. Задача формально устойчива, но фактическая неустранимая ошибка может быть большой. Этот случай называют *слабой* устойчивостью (или плохой обусловленностью). Примером является такая задача:

$$y''(x) = y(x), \tag{9а}$$

$$y(0) = 1, \quad y'(0) = -1. \tag{9б}$$

Общее решение дифференциального уравнения (9а) есть:

$$y(x) = 0{,}5\,[y(0) + y'(0)]\,e^x + 0{,}5\,[y(0) - y'(0)]\,e^{-x}.$$

Начальным условиям (9б) соответствует точное решение $y(x) = e^{-x}$; но небольшая погрешность начальных данных может привести к тому, что в решении добавится член вида εe^x, который при больших аргументах много больше искомого решения.

Очевидно, для хорошей практической устойчивости расчета константа C должна быть не слишком велика. Так, если начальные данные известны точно, т. е. могут быть заданы с точностью до ошибок округления $\Delta \sim 10^{-12}$, то необходимо, чтобы $C \ll 10^{12}$. Если же начальные данные найдены из эксперимента с точностью

$\delta x \sim 0{,}001$, а требуемая точность решения $\delta y \sim 0{,}1$, то допустимо $C \leqslant 100$.

Даже если задача устойчива, то численный алгоритм может быть неустойчивым. Например, если производные заменяются разностями, то приходится вычитать близкие числа и сильно теряется точность. Эти неточные промежуточные результаты используются в дальнейших вычислениях, и ошибки могут сильно нарастать.

По аналогии можно говорить о корректности алгоритма $\bar{y} = \bar{A}(\bar{x})$, подразумевая существование и единственность приближенного решения для любых входных данных $\bar{x}$ некоторого класса, и устойчивость относительно всех ошибок в исходных данных и промежуточных выкладках. Однако в общем случае этим определением трудно пользоваться; только в теории разностных схем (глава IX) оно применяется успешно.

ЗАДАЧИ

1. Доказать выполнимость всех соотношений (4). Рассмотреть, как меняется форма записи этих соотношений при задании функции на произвольном конечном отрезке $a \leqslant t \leqslant b$.

2. Доказать утверждения о согласованности и подчиненности норм матриц, приведенные в конце п. 1 § 2.