В. И. Горбаченко

Вычислительная линейная алгебра с примерами на MATLAB

- *Теоретические основы численных методов*
- *Прямые методы решения систем линейных алгебраических уравнений*
- *Итерационные методы решения систем линейных алгебраических уравнений*
- *Собственные значения и векторы матриц*
- *Примеры на MATLAB*

УЧЕБНОЕ ПОСОБИЕ

bhv®

УДК 681.3.06+512.64(075.8)
ББК 32.973.26–018.2/22.143я73
Г67

Горбаченко В. И.

Г67 Вычислительная линейная алгебра с примерами на MATLAB. — СПб.: БХВ-Петербург, 2011. — 320 с.: ил. — (Учебное пособие)

ISBN 978-5-9775-0725-7

Излагаются теоретические основы численных методов, включая теорию погрешностей, особенности машинной арифметики, корректность и обусловленность вычислительных задач; современные прямые и итерационные методы решения больших систем линейных алгебраических уравнений. Основное внимание уделено современным итерационным методам на основе подпространств Крылова. Рассмотрено решение частичной и полной проблемы собственных значений, в том числе для больших разреженных матриц. Для основных вычислительных методов приведены реализации с использованием программ, разработанных автором, а также соответствующие функции системы MATLAB.

Для студентов и преподавателей высших учебных заведений

УДК 681.3.06+512.64(075.8)
ББК 32.973.26–018.2/22.143я73

Рецензенты:

О. Э. Яремко, канд. физ.-мат. наук, проф., завкафедрой математического анализа Пензенского государственного педагогического университета им. В. Г. Белинского;

Кафедра информационно-вычислительных систем Пензенского государственного университета, завкафедрой д-р техн. наук, проф. Ю. Н. Косников.

Группа подготовки издания:

Главный редактор	*Екатерина Кондукова*
Зам. главного редактора	*Евгений Рыбаков*
Зав. редакцией	*Григорий Добин*
Редактор	*Юрий Якубович*
Компьютерная верстка	*Натальи Караваевой*
Корректор	*Наталия Першакова*
Дизайн серии	*Инны Тачиной*
Оформление обложки	*Елены Беляевой*
Зав. производством	*Николай Тверских*

Лицензия ИД № 02429 от 24.07.00. Подписано в печать 26.07.11.
Формат $70\times100^{1}/_{16}$. Печать офсетная. Усл. печ. л. 25,8.
Тираж 1000 экз. Заказ №
"БХВ-Петербург", 190005, Санкт-Петербург, Измайловский пр., 29.

Санитарно-эпидемиологическое заключение на продукцию
№ 77.99.60.953.Д.005770.05.09 от 26.05.2009 г. выдано Федеральной службой
по надзору в сфере защиты прав потребителей и благополучия человека.

Отпечатано с готовых диапозитивов
в ГУП "Типография "Наука"
199034, Санкт-Петербург, 9 линия, 12.

ISBN 978-5-9775-0725-7

Оглавление

Глава 1

Теоретические основы численных методов

1.1. Математическое моделирование и вычислительный эксперимент

В современных науке и технике важную роль играет *математическое моделирование* [1—2][1], заменяющее эксперименты с реальными объектами экспериментами с их математическими моделями. Возник даже термин "*вычислительный эксперимент*" [1—3]. Вычислительный эксперимент имеет ряд преимуществ по сравнению с натурным экспериментом:

- ❒ экономичность, так как не тратятся ресурсы реальной системы;
- ❒ возможность моделирования гипотетических, то есть не реализованных в природе объектов (прежде всего на разных этапах проектирования);
- ❒ доступность тестирования режимов, опасных или трудновоспроизводимых в натуре (критический режим ядерного реактора, работа системы противоракетной обороны, природные и техногенные катастрофы);
- ❒ возможность изменения масштаба времени;
- ❒ простота многоаспектного анализа;
- ❒ большая прогностическая сила вследствие возможности выявления общих закономерностей.

Следует отметить, что моделирование сложных объектов, например атомных, космических и многих других, требует проведения колоссальных объемов вычислений. Например, для решения многих прикладных задач аэродинамики и ядерной физики требуется выполнение более 10^{13} арифметических операций [4].

Для практических задач довольно редко удается найти аналитическое решение уравнений, составляющих математическую модель явления. Поэтому приходится применять численные методы. Сущность применения численных методов рассмотрим на схеме вычислительного эксперимента [3, 5—6], показанной на рис. 1.1.

---

[1] Ссылки в тексте относятся к библиографическому списку в конце главы. — *Ред.*

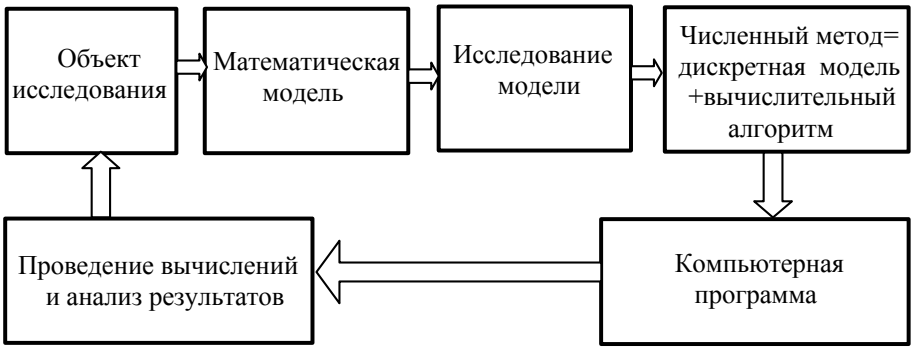

Рис. 1.1. Схема вычислительного эксперимента

Основу вычислительного эксперимента составляет триада: *модель — метод (алгоритм) — программа*. Сначала строится с некоторыми допущениями *математическая модель объекта.* Первоначально строится относительно простая, но достаточно полная с точки зрения экспериментальных данных модель. В ходе вычислительного эксперимента модель уточняется и дополняется. Поэтому можно говорить о *наборе математических моделей*, каждая из которых с различной точностью описывает объект или различные свойства объекта. Построение математических моделей выходит за рамки нашего курса. Более подробную информацию о построении математических моделей можно найти в книгах [2, 3].

После выбора или построения математической модели средствами прикладной математики проводится *предварительное качественное исследование модели*. Качественное исследование начинается с *приведения задачи к безразмерному виду* [6, 7]. Суть приведения задачи к безразмерному виду состоит в выделении характерных значений (величин) и масштабировании всех величин задачи. Приведение задачи к безразмерному виду сокращает общее число параметров математической модели, что упрощает анализ. Кроме того, в безразмерной задаче можно сравнивать параметры и упрощать задачу, отбрасывая малые параметры. Наконец, приведение к безразмерному виду уменьшает влияние ошибок округления при вычислениях на компьютере, так как все безразмерные величины изменяются примерно от −1 до +1. Приведение задачи к безразмерному виду представляет достаточно сложную задачу, так как основано на теории подобия [7] и требует знания особенностей процессов, протекающих в объекте исследования. В частности масштабы разных величин оказываются взаимосвязанными. Математическая модель часто оказывается очень сложной для качественного исследования. В этом случае для качественного исследования строят *модельные (упрощенные) задачи*. После приведения к безразмерному виду рассматриваются *существование* и *единственность* решения, основные *свойства решения*, влияние различных параметров задачи на решение, устойчивость решения относительно малых возмущений входных данных и другие вопросы. Качественное исследование модели зачастую представляет собой сложную самостоятельную задачу. Однако для многих прикладных задач, например для обыкновенных дифференциальных уравнений, качественный анализ уже проведен.

Затем необходимо решить систему уравнений, представляющую собой математическую модель объекта. Как уже говорилось, обычно приходится применять численные методы. Под *численным методом* понимается совокупность *дискретной модели*, реализуемой на компьютере, и *вычислительного алгоритма*, позволяющего решить дискретизированную задачу. Например, дискретной моделью вычисления определенного интеграла является вычисление суммы площадей прямоугольников, аппроксимирующих площадь криволинейной трапеции, являющейся геометрической интерпретацией задачи. Для реализации численного метода необходимо разработать *программу* на одном из языков программирования или применить готовый пакет прикладных программ. В настоящее время разработано несколько пакетов прикладных программ, или систем компьютерной математики, таких как MathCAD, MATLAB, Maple, Mathematica и других, которые позволяют решать большинство практически встречающихся задач. Однако грамотная постановка задачи, рациональный выбор метода решения, оценка погрешности и правильная интерпретация результатов требуют серьезных знаний численных методов. После отладки программы производятся вычисления на компьютере (обычно требуется провести много вариантов вычислений, для чего необходимо планировать вычислительный эксперимент) и анализ результатов. После получения результатов исследуется соответствие результатов вычислительного эксперимента процессу функционирования реального объекта (проверяется *адекватность модели*), и при необходимости уточняются компоненты схемы вычислительного эксперимента (рис. 1.1) до получения удовлетворительных результатов.

Можно выделить три основных типа вычислительного эксперимента [6]:

- *поисковый*, заключающийся в исследовании различных объектов и процессов посредством математических моделей;
- *оптимизационный*, направленный, например, на подбор оптимальных характеристик конструкций, технологических процессов;
- *диагностический*, когда по результатам натурных экспериментов определяются свойства объектов или явлений (при этом решаются *обратные задачи*).

Одной и той же математической модели можно поставить в соответствие множество дискретных моделей и вычислительных алгоритмов, то есть численных методов. Для выбора подходящих численных методов надо знать их основные свойства. Рассмотрение свойств численных методов начнем с анализа погрешностей вычислений.

1.2. Погрешности вычислений

1.2.1. Источники погрешностей вычислений

Исследование реального объекта методом вычислительного эксперимента носит приближенный характер. На каждом этапе вычислительного эксперимента (рис. 1.1) возникают погрешности. Можно выделить 5 источников погрешностей [5, 8—11]:

- погрешность математической модели;
- погрешность дискретизации;

- трансформированная погрешность (погрешность искажения);
- методическая погрешность;
- погрешность округления.

Неустранимая по отношению к численному методу *погрешность математической модели* обусловлена недостаточным знанием или чрезмерным упрощением моделируемых объектов и явлений. То есть причиной этого вида погрешности является неадекватность математической модели. Адекватность математической модели может быть оценена путем сравнения реального процесса и реализации на компьютере модели процесса. Это очень сложная задача, так как реализация модели на компьютере включает много составляющих погрешностей, исследование реального процесса также происходит с погрешностью. Оценка погрешности математической модели выходит за рамки данного курса.

Погрешность дискретизации — это погрешность от замены непрерывной математической модели дискретной моделью. Примером дискретизации является замена дифференциального уравнения системой алгебраических уравнений. Решение системы алгебраических уравнений отличается от решения дифференциального уравнения, что и составляет погрешность дискретизации. Обычно дискретная модель зависит от некоторых параметров, изменением которых теоретически можно понизить погрешность дискретизации. В рассмотренном примере, выбирая более мелкий шаг дискретизации (и увеличивая порядок системы алгебраических уравнений), можно уменьшить погрешность дискретизации.

Трансформированная погрешность (погрешность искажения) [8] — это погрешность, возникающая за счет *погрешности исходных данных*. Исходные данные, как правило, являются результатом измерений некоторых величин, естественно, эти измерения производятся с некоторой погрешностью. Кроме того, из-за ограниченной разрядности компьютеров возникает *погрешность представления исходных данных* (погрешность округления исходных данных). Многие задачи являются весьма чувствительными к погрешностям исходных данных и к погрешностям, возникающим в ходе реализации алгоритма. В этом случае говорят, что задача является *плохо обусловленной*. Для плохо обусловленной задачи погрешности исходных данных, даже при отсутствии погрешностей вычислений, могут приводить к значительным искажениям результата.

Методическая погрешность возникает из-за применения приближенных алгоритмов. Например, решая систему алгебраических уравнений итерационным методом, теоретически можно получить точное решение лишь при бесконечно большом числе итераций. Однако число итераций приходится ограничивать, что приводит к погрешности метода.

Погрешность округления возникает из-за того, что все вычисления выполняются с ограниченным числом значащих цифр, то есть производится округление чисел. Погрешности округления накапливаются и при плохой обусловленности задачи могут привести к большим отклонениям результата.

Как выбирать допустимые погрешности? Известны различные рекомендации. В [5] предлагается обеспечивать одинаковый порядок всех погрешностей. То есть

не следует очень точно решать задачу с неточными исходными данными. Другими словами, погрешность решения *всей* задачи (всей последовательности действий вычислительного эксперимента) должна быть соизмерима с погрешностью исходных данных.

В [12] рекомендуется обеспечивать методическую погрешность в 2—10 раз меньшую, чем погрешность модели, а погрешность округления — на порядок меньше погрешности метода.

1.2.2. Приближенные числа. Абсолютная и относительная погрешность

Мерой точности вычислений являются погрешности. Пусть a — приближенное значение точного числа A. *Погрешностью*, или *ошибкой* Δa приближенного числа a называется разность

$$\Delta a = a - A .$$

В качестве меры погрешности используют абсолютную и относительную погрешность. *Абсолютная погрешность* приближенного числа a определяется формулой

$$\Delta\ a\ = |a - A| .$$

Так как точное значение числа A в большинстве случаев неизвестно, то используется оценка *предельной абсолютной погрешности* Δ_a, называемая также *границей абсолютной погрешности*:

$$\Delta\ a\ = |a - A| \le \Delta_a .$$

Относительная погрешность определяется формулой

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

Относительная погрешность часто выражается в процентах.

Предельной относительной погрешностью δ_a приближенного числа a называется число, заведомо не меньшее относительной погрешности этого числа

$$\delta\ a\ \le \delta_a .$$

Так как значение точного числа A неизвестно, то часто пользуются приближенными оценками предельных погрешностей

$$\delta_a \approx \frac{\Delta_a}{|a|}, \quad \Delta_a \approx |a|\delta_a .$$

Очень часто используется термин "точность решения". Хотя точность решения противоположна по смыслу погрешности решения, для измерения точности используются те же характеристики, что и для измерения погрешности. Когда говорят, что точность решения равна ε, то это означает, что принятая мера погрешности решения не превышает ε.

Абсолютная и относительная погрешности тесно связаны с понятием *верных значащих цифр*. Значащими цифрами числа называют все цифры в его записи, начиная

с первой ненулевой цифры слева. Например, число 0,000<u>12900</u> имеет пять значащих цифр (значащие цифры подчеркнуты). Значащая цифра называется *верной*, если абсолютная погрешность числа не превышает вес разряда, соответствующего этой цифре. Значащая цифра называется *верной в узком смысле слова*, если абсолютная погрешность числа не превышает половины веса разряда, соответствующего этой цифре. Пусть, например, число равно $a = 9348$, абсолютная погрешность числа равна $\Delta a = 15$. Записывая число в виде

$$9348 = 9 \cdot 10^3 + 3 \cdot 10^2 + 4 \cdot 10^1 + 8 \cdot 10^0,$$

имеем $0{,}5 \cdot 10^1 < \Delta a < 0{,}5 \cdot 10^2$, следовательно, число имеет две верных в узком смысле значащих цифр (9 и 3).

Верная значащая цифра может не совпадать с соответствующей цифрой в записи точного числа. Например, $A = 1.000$, $a = 0.999$, $\Delta a = 0.001$, тогда у приближенного числа a все значащие цифры верны, но не совпадают с цифрами точного числа A.

Количество верных значащих цифр числа связано со значением его относительной погрешности [12]. Если десятичное приближенное число a содержит n верных значащих цифр, то для относительной погрешности справедлива оценка

$$\delta a \le 10^{n-1} - 1^{-1} \approx 10^{-n+1}.$$

Чтобы число a содержало n верных значащих цифр, достаточно выполнения неравенства

$$\delta a \le 10^n + 1^{-1} \approx 10^{-n}.$$

Поэтому, если необходимо вычислить приближенное число a с *точностью* 10^{-n}, то необходимо сохранить верной значащую цифру, стоящую в n-м разряде после десятичной запятой.

Тот факт, что число a является приближенным значением числа A с предельной абсолютной погрешностью Δ_a, записывают в виде

$$A = a \pm \Delta_a,$$

причем числа a и Δ_a записываются с одинаковым количеством цифр после запятой, например, $A = 2{,}347 \pm 0{,}002$ или $A = 2{,}347 \pm 2 \cdot 10^{-3}$.

Запись вида

$$A = a\ 1 \pm \delta_a$$

означает, что число a является приближенным значением числа A с предельной относительной погрешностью δ_a. Предельные абсолютные и относительные погрешности принято записывать с одной или двумя значащими цифрами. Бóльшая точность не имеет смысла, так как предельные погрешности представляют собой достаточно грубые оценки.

1.2.3. Особенности машинной арифметики

Одним из основных источников вычислительных погрешностей является приближенное представление чисел в компьютере, обусловленное конечностью разрядной сетки. При решении вычислительных задач обычно используют представление чисел в *форме с плавающей точкой (запятой)*. Число a в форме с плавающей точкой представляется в форме

$$a = \pm M r^{p},$$

где $\pm$ — знак числа; r — основание системы счисления (как правило, $r = 2$); p — порядок числа a; M — мантисса числа a, причем должно выполняться условие нормировки $r^{-1} \le M < 1$, означающее, что в компьютере хранятся только значащие цифры.

Диапазон изменения чисел в компьютере ограничен. Для всех представимых в компьютере нормализованных чисел x (за исключением нуля) справедливо неравенство

$$0 < X_0 \le |x| < X_\infty,$$

где X_0 — минимальное представимое в компьютере нормализованное число (*машинный ноль*) $X_0 = 2^{-p_{\max}+1}$; $p_{\max} = 2^{l+1} - 1$ — максимальное по абсолютной величине значение порядка; l — разрядность порядка; $X_\infty = M_{\max} 2^{p_{\max}} = (1 - 2^{-t})\, 2^{p_{\max}} \approx 2^{p_{\max}}$ — максимальное представимое в компьютере нормализованное число (*машинная бесконечность*); $M_{\max}$ — максимальное по абсолютной величине значение мантиссы; t — разрядность мантиссы.

В большинстве языков программирования получение машинной бесконечности приводит к аварийному останову выполнения программы по *переполнению*. Все числа, по модулю меньшие X_0, представляются как *ноль (исчезновение порядка)*.

Большинство современных компьютеров поддерживают *стандарт двоичной арифметики* IEEE 754-1985 (ANSI 754) Binary floating-point arithmetic (1985 г.)[1] [13]. Представление чисел с плавающей точкой в этом стандарте несколько отличается от рассмотренного "классического" представления. Под знак числа отводится один бит: 0 соответствует положительному числу, 1 — отрицательному. Мантисса нормализованных двоичных чисел удовлетворяет условию $1 \le M < 2$, то есть мантисса нормализованного числа всегда содержит единицу в целой части. Так как целая часть всех нормализованных чисел равна единице, то эта единица не хранится, а хранится только дробная часть мантиссы. Число получается прибавлением единицы к дробной части. Сэкономленный разряд используется для хранения еще

[1] IEEE — The Institute of Electrical and Electronics Engineers, Inc. (произносится "ай-трипл-и"), Институт инженеров по электротехнике и радиоэлектронике, ИИЭР (США) — крупнейшая в мире организация (**http://www.ieee.org/**), объединяющая более 300 тыс. технических специалистов из 147 стран, ведущая организация по стандартизации, отвечающая также за сетевые стандарты.

одного двоичного разряда. Математически представление мантиссы, принятое в стандарте, эквивалентно "классическому" представлению, но увеличивает разрядность мантиссы и диапазон представления чисел. Порядок может быть как положительным, так и отрицательным. Чтобы не вводить бит знака порядка, используют смещенный порядок, прибавляя к порядку смещение, равное $2^{l-1}-1$, где l — число разрядов, отведенное под смещенный порядок. Например, если под порядок отведен один байт, то смещение равно $2^7-1=127$. Если порядок равен +4, то смещенный порядок будет $4+127=131$. Если порядок равен −4, то смещенный порядок будет $-4+127=123$. Число с нулевым порядком и нулевой мантиссой считается нулем. При этом формально различаются $+0$ и -0.

Стандарт предусматривает два основных типа чисел с плавающей точкой: числа одинарной и двойной точности. В стандарте предусмотрены также расширенный одинарный (длиной 43 и более битов, редко используется) и расширенный двойной форматы (длиной 79 и более битов, обычно используется 80-битный формат Intel двойной точности).

Числа одинарной точности представляются 32 битами (4 байтами) в виде: 1 бит под знак, 8 бит под порядок, 23 бита под мантиссу. *Числа двойной точности* представляются 64 битами (8 байтами) в виде: 1 бит под знак, 11 бит под порядок, 52 бита под мантиссу.

IEEE-арифметика включает *субнормальные числа* — очень малые *ненормализованные* числа, расположенные между 0 и наименьшим по абсолютной величине нормализованным числом X_0. При таких условиях нормализация чисел не производится. Субнормальные числа имеют нулевой смещенный порядок. Наличие субнормальных чисел приводит к тому, что малые числа не превращаются в машинный ноль. Например, равенство $x-y=0$ возможно только при $x=y$.

IEEE-арифметика поддерживает специальные символы $\pm\infty$ (в MATLAB обозначается inf) и NaN. Символ $\pm\infty$ генерируются при переполнении. Бесконечность содержит единицы во всех разрядах смещенного порядка и нули в разрядах мантиссы. Правила для символа $\pm\infty$: $x/\pm\infty=0$, $x/0=\pm\infty$, $\infty+\infty=+\infty$ и т. д. Любая операция, результат которой (конечный или бесконечный) не определен корректно, генерирует символ NaN (Not a Number — не число). Например, символ NaN генерируется при операциях: $\infty-\infty$, $\frac{\infty}{\infty}$, $\frac{0}{0}$, $NaN\odot x$, где $\odot$ — любая операция.

В случае если в результате арифметической операции получается неопределенность (NaN), машинная бесконечность, деление на ноль или машинный ноль, то процессор выставляет флаг исключительного состояния (exception flag). Большинство компиляторов по умолчанию не маскируют исключения с плавающей точкой. Это приводит к завершению программ при возникновении таких исключений. Программист должен маскировать и обрабатывать исключения с плавающей точкой самостоятельно, что может сильно усложнить программу. Система

MATLAB маскирует и обрабатывает исключительные ситуации, что существенно упрощает программирование.

В некоторых случаях оказывается недостаточно двойной точности стандарта IEEE 754-1985. Поэтому многие компиляторы поддерживают 80-битный формат Intel двойной точности (мантисса содержит 64 бита, порядок — 15 бит). Некоторые компиляторы программно поддерживают увеличенную разрядность вычислений (при этом существенно увеличивается время вычислений). В новой редакции стандарта IEEE 754-2008 предусмотрены числа учетверенной точности, мантисса которых содержит 112 бит, а порядок — 15 бит.

Для уменьшения погрешностей вычислений иногда используется *рациональная арифметика* — числа представляются в виде рациональных дробей и операции производятся как над обычными дробями. Рациональная арифметика реализована, например, в пакете Mathematica.

Существует также подход, известный как *интервальные вычисления* [14]. В интервальных вычислениях при выполнении каждой арифметической операции вычисляются границы ошибки. Однако алгоритмы интервальных вычислений оказываются сложными и медленными.

Найдем некоторые оценки точности представления чисел и выполнения арифметических операций. Для простоты используем простейшее представление чисел с плавающей точкой, то есть будем использовать условие нормировки мантиссы $2^{-1} \le M < 1$. Полученные результаты справедливы и для представления чисел в стандарте IEEE 754, если учесть, что в стандарте разрядность мантиссы фактически больше на единицу физически выделенной под мантиссу разрядности.

Из-за конечной разрядной сетки в компьютере можно представить не все числа. Число a, не представимое в компьютере, подвергается округлению, то есть заменяется близким числом $\tilde{a}$, представимым в компьютере точно. Известно несколько способов округления числа до n значащих цифр. Простейший способ — *усечение*. При усечении отбрасываются все цифры, расположенные справа от n-й значащей цифры. Несколько более сложный способ, *округление по дополнению*, получил наибольшее распространение. В простейшем варианте этого способа анализируется первая отбрасываемая цифра. Если эта цифра равна 1 (в двоичной системе счисления), то к младшей сохраняемой цифре прибавляется единица.

Найдем границу *относительной погрешности представления числа с плавающей точкой* [5]. Допустим, что применяется простейшее округление — усечение. Система счисления — двоичная. Пусть надо записать число, представляющее бесконечную двоичную дробь

$$a = \pm 2^p \cdot \underbrace{\left(\frac{a_1}{2} + \frac{a_2}{2^2} + \cdots + \frac{a_t}{2^t} + \frac{a_{t+1}}{2^{t+1}} + \cdots \right)}_{\text{мантисса}},$$

где $a_j = \begin{cases} 0 \\ 1 \end{cases}$, $j = 1, 2, \ldots$ — цифры мантиссы, а p — порядок.

Пусть под запись мантиссы отводится t двоичных разрядов. Отбрасывая лишние разряды, получим округленное число

$$\tilde{a} = \pm 2^p \left(\frac{a_1}{2} + \frac{a_2}{2^2} + \cdots + \frac{a_t}{2^t} \right).$$

Абсолютная погрешность округления в этом случае равна

$$\left| \tilde{a} - a \right| = 2^p \left(\frac{a_{t+1}}{2^{t+1}} + \frac{a_{t+2}}{2^{t+2}} + \cdots \right).$$

Наибольшая погрешность будет в случае $a_{t+1} = 1, \quad a_{t+2} = 1, \ldots$, тогда

$$\left| \tilde{a} - a \right| \le 2^p \frac{1}{2^{t+1}} \underbrace{\left(1 + \frac{1}{2} + \frac{1}{2^2} + \cdots \right)}_{=2} = 2^{p-t}.$$

Из условия нормировки для мантиссы всегда верно $a_1 = 1$, поэтому выполняется неравенство $M \ge 0{,}5$. Тогда $|a| \ge 2^p \cdot 2^{-1} = 2^{p-1}$, и относительную погрешность можно оценить следующим образом: $\delta\ a\ = \dfrac{|\tilde{a} - a|}{|a|} \le 2^{-t+1}$. Практически применяют более точные методы округления [11], и для погрешности представления чисел справедлива оценка

$$\delta\ a\ = \frac{|\tilde{a} - a|}{|a|} \le 2^{-t}, \tag{1.1}$$

то есть точность представления чисел определяется разрядностью мантиссы t.

Тогда приближенно представленное в компьютере число можно записать в виде $\tilde{a} = a\ 1 \pm \varepsilon$, где $\varepsilon = 2^{-t}$ — так называемый машинный эпсилон или машинная точность (часто обозначается macheps) — относительная погрешность представления чисел (1.1). В системе MATLAB переменная `eps` обозначает машинную точность, `eps=2^-52`, что примерно равно `2.2204e-016`.

При вычислениях с плавающей точкой операция округления может потребоваться после выполнения любой арифметической операции. Так, умножение или деление двух чисел сводится к умножению или делению мантисс и к сложению или вычитанию порядков. Так как в общем случае количество разрядов мантисс произведений и частных больше допустимой разрядности мантиссы, то требуется округление мантиссы результата. При сложении или вычитании чисел с плавающей точкой операнды должны быть предварительно приведены к одному порядку. Это осуществляется сдвигом вправо мантиссы числа, имеющего меньший порядок, и увеличением в соответствующее число раз порядка этого числа. Сдвиг мантиссы вправо может привести к потере младших разрядов мантиссы, то есть появляется погрешность округления.

Обозначим округленное в системе с плавающей точкой число, соответствующее точному числу x, через $\mathrm{fl}\ x$ (от англ. floating — плавающий). Известно [5], что

выполнение каждой арифметической операции вносит относительную погрешность, не большую, чем погрешность представления чисел с плавающей точкой (1.1). Тогда можно записать

$$\text{fl}\ a \square b\ = a \square b\ 1 \pm \delta\ ,$$

где $\square$ — любая из арифметических операций, а $\delta \le 2^{-t}$.

Таким образом, относительная погрешность арифметической операции над числами, представленными в форме с плавающей точкой, не превышает машинный эпсилон.

1.2.4. Трансформированные погрешности арифметических операций

Рассмотрим *трансформированные погрешности арифметических операций*. Будем считать, что арифметические операции проводятся над приближенными числами, ошибку арифметических операций учитывать не будем (эту ошибку легко учесть, прибавив ошибку округления соответствующей операции к вычисленной ошибке).

Рассмотрим сложение и вычитание приближенных чисел. Покажем, что абсолютная погрешность алгебраической суммы приближенных чисел не превосходит суммы абсолютных погрешностей слагаемых.

Действительно, пусть сумма точных чисел равна

$$S = A_1 + A_2 + \cdots + A_n\ .$$

Абсолютная погрешность суммы равна

$$\begin{aligned} \Delta\ S\ &= \Delta\ a_1 + a_2 + \cdots + a_n\ = \left|\ a_1 + a_2 + \cdots + a_n\ -\ A_1 + A_2 + \cdots + A_n\ \right| = \\ &= \left|\ a_1 - A_1\ +\ a_2 - A_2\ + \cdots +\ a_n - A_n\ \right| \le \\ &\le \left|a_1 - A_1\right| + \left|a_2 - A_2\right| + \cdots + \left|a_n - A_n\right| = \Delta\ a_1\ + \Delta\ a_2\ + \cdots + \Delta\ a_n\ . \end{aligned} \tag{1.2}$$

Из (1.2) получаем оценку с использованием предельных абсолютных погрешностей:

$$\Delta\ S\ \le \Delta_{a_1} + \Delta_{a_2} + \ldots + \Delta_{a_n}\ .$$

Для относительной погрешности суммы нескольких чисел справедлива оценка

$$\begin{aligned} \delta\ S\ &= \frac{\Delta\ S}{|S|} \le \frac{|a_1|}{|S|}\left(\frac{\Delta\ a_1}{|a_1|}\right) + \frac{|a_2|}{|S|}\left(\frac{\Delta\ a_2}{|a_2|}\right) + \cdots \\ &= \frac{|a_1|\delta\ a_1\ + |a_2|\delta\ a_2\ + \ldots}{|S|}\ , \end{aligned} \tag{1.3}$$

где $\delta\ a_i\ ,\quad i = 1, 2, ..., n$ — относительные погрешности представления чисел.

Из (1.3) следует, что относительная погрешность суммы нескольких чисел *одного знака* не превышает наибольшей из относительных погрешностей слагаемых:

$$\delta\ S\ \le \max_k \delta\ a_k\ ,\quad k = 1, 2, \ldots, n\ .$$

Пусть A и B — *ненулевые числа одного знака*. Тогда для соответствующих приближенных чисел справедливы неравенства

$$\delta\ a+b\ \le\delta_{\max},\quad \delta\ a-b\ \le\gamma\delta_{\max}, \tag{1.4}$$

где $\delta_{\max}=\max\ \delta\ a\ ,\delta\ b\ \ ,\quad \gamma=\left|A+B\right|/\left|A-B\right|$.

Для доказательства, учитывая (1.2), рассмотрим выражение

$$\begin{aligned}&\left|A\pm B\right|\delta\ a\pm b\ =\Delta\ a\pm b\ \le\Delta\ a\ +\Delta\ b\ =\\&=\left|A\right|\delta\ a\ +\left|B\right|\delta\ b\ \le\ \left|A\right|+\left|B\right|\ \delta_{\max}=\left|A+B\right|\delta_{\max}.\end{aligned} \tag{1.5}$$

Из выражения (1.5) следуют оценки (1.4). Из второго неравенства (1.4) следует, что *при сложении чисел разного знака или вычитании чисел одного знака относительная погрешность может быть очень большой* (если числа близки между собой). Это происходит из-за того, что величина γ в этой ситуации может быть очень большой. Поэтому вычислительные алгоритмы необходимо строить таким образом, чтобы избегать вычитания близких чисел.

Для предельных относительных погрешностей справедливо равенство

$$\delta_S=\frac{\left|a_1\right|\delta_{a_1}+\left|a_2\right|\delta_{a_2}+\ldots}{\left|S\right|}\ .$$

Отметим также, что *погрешности вычислений зависят от порядка вычислений*. Рассмотрим пример сложения трех приближенных чисел

$$\begin{aligned}S&=x_1+x_2+x_3,\\ \tilde{S}_1&=\ x_1+x_2\ \ 1\pm\delta\ ,\\ \tilde{S}_2&=\ \tilde{S}_1+x_3\ \ 1\pm\delta\ =\ x_1+x_2\ \ 1\pm\delta\ \ 1\pm\delta\ +x_3\ 1\pm\delta\ ,\end{aligned} \tag{1.6}$$

где δ — предельная относительная погрешность округления.

При другой последовательности действий результат будет другим:

$$\begin{aligned}\tilde{S}_3&=\ x_3+x_2\ \ 1\pm\delta\ ,\\ \tilde{S}_4&=\ x_3+x_2\ \ 1\pm\delta\ \ 1\pm\delta\ +x_1\ 1\pm\delta\ .\end{aligned}$$

Из (1.6) видно, что результат выполнения некоторого алгоритма, искаженный погрешностями округления, совпадает с результатом выполнения того же алгоритма, но с неточными исходными данными. Поэтому можно применять *обратный анализ*, то есть свести влияние погрешностей округления к возмущению исходных данных. Другими словами, вместо (1.6) можно записать

$$\tilde{S}=\tilde{x}_1+\tilde{x}_2+\tilde{x}_3\ ,$$

где $\tilde{x}_1=x_1\ 1\pm\delta\ \ 1\pm\delta\ ,\quad \tilde{x}_2=x_2\ 1\pm\delta\ \ 1\pm\delta\ ,\quad \tilde{x}_3=x_3\ 1+\delta\ .$

В *прямом анализе ошибок* пытаются оценить влияние погрешностей исходных данных, метода и округлений на результат. В *обратном анализе ошибок* приближенное решение рассматривается как точное решение задачи, но при возмущенных

исходных данных. В обратном анализе рассматривается влияние ошибок на результат, а не оценивается погрешность результата. Однако оценить эквивалентное возмущение исходных данных весьма сложно, поэтому на практике обратный анализ ошибок пока не получил широкого распространения.

Для относительной трансформированной погрешности произведения справедлива оценка

$$\delta\ ab\ \le \delta\ a\ + \delta\ b\ + \delta\ a\ \delta\ b\ . \tag{1.7}$$

Для доказательства рассмотрим выражение

$$\begin{aligned}
\left|AB\right|\delta\ ab\ &= \Delta\ ab\ = \left|ab - AB\right| = \\
&= \left|\ A - a\ \ B + \ B - b\ \ A - \ A - a\ \ \ B - b\ \right| \le \\
&\le \left|B\right|\Delta\ a\ + \left|A\right|\Delta\ b\ + \Delta\ a\ \Delta\ b\ = \\
&= \left|B\right|\Delta\ a\ \frac{\left|A\right|}{\left|A\right|} + \left|A\right|\Delta\ b\ \frac{\left|B\right|}{\left|B\right|} + \Delta\ a\ \Delta\ b\ \frac{\left|A\right|}{\left|A\right|}\frac{\left|B\right|}{\left|B\right|} = \\
&= \left|AB\right|\ \delta\ a\ + \delta\ b\ + \delta\ a\ \delta\ b\ \ ,
\end{aligned}$$

из которого следует (1.7).

Для относительной погрешности частного справедлива оценка

$$\delta\left(\frac{a}{b}\right) \le \frac{\delta\ a\ + \delta\ b}{1 - \delta\ b}.$$

Действительно,

$$\begin{aligned}
\delta\left(\frac{a}{b}\right) &= \frac{\left|\frac{a}{b} - \frac{A}{B}\right|}{\left|\frac{A}{B}\right|} = \frac{\left|aB - Ab\right|}{\left|Ab\right|} = \frac{\left|B\ \ a - A\ \ - A\ \ b - B\ \right|}{\left|Ab\right|} \le \\
&\le \frac{\left|B\right|\Delta\ a\ + \left|A\right|\Delta\ b}{\left|AB\right|\ 1 - \delta\ b} = \frac{\delta\ a\ + \delta\ b}{1 - \delta\ b}.
\end{aligned} \tag{1.8}$$

В выражении (1.8) учтено неравенство

$$\left|b\right| = \left|B + \ b - B\ \right| \ge \left|B\right| - \Delta\ b\ = \left|B\right|\ 1 - \delta\ b\ \ .$$

Так как для предельных погрешностей справедливы оценки $\delta_a \ll 1$ и $\delta_b \ll 1$, то для оценки предельных трансформированных погрешностей умножения и деления на практике используются оценки

$$\delta_{ab} \approx \delta_a + \delta_b, \qquad \delta_{a/b} \approx \delta_a + \delta_b.$$

Рассмотренные аналитические оценки громоздки и малоприменимы для оценки погрешностей многократно повторенных операций. Кроме того, эти оценки рассчитаны на наихудшие случаи и не учитывают частичную компенсацию погрешностей при выполнении большого числа операций. Ведь часть операций будет иска-

жать результат в большую сторону, а часть — в меньшую. При большом числе n арифметических операций можно пользоваться *приближенной статистической оценкой погрешности арифметических операций*, учитывающей частичную компенсацию погрешностей разных знаков [15]: $\delta_\Sigma \approx \delta_{fl}\sqrt{n}$, где δ_Σ — суммарная относительная погрешность; $\delta_{fl} \le \varepsilon$ — относительная погрешность выполнения операций с плавающей точкой; ε — погрешность представления чисел с плавающей точкой. Статистическую оценку погрешности проиллюстрируем на примере суммы большого числа слагаемых $S = \sum_{i=1}^{n} a_i$. Вследствие погрешностей слагаемые a_i можно рассматривать как случайные величины. Если слагаемые a_i независимы и их среднеквадратические отклонения равны σ_a, то из теории вероятностей известно, что $\sigma_S = \sigma_a\sqrt{n}$. Если принять, что погрешности распределены по нормальному закону с нулевым математическим ожиданием и среднеквадратическим отклонением σ_a, то часто считают, что погрешность не превышает трех σ_a: $\Delta_a \le 3\sigma_a$ ("правило трех сигма"). Умножая предыдущее равенство на 3, получаем для предельных погрешностей $\Delta_S = \Delta_a\sqrt{n}$.

1.2.5. Трансформированные погрешности вычисления функций

Рассмотрим *трансформированную погрешность вычисления значений функций*. Пусть $y = f\ x_1, x_2, ..., x_m$ — дифференцируемая функция многих переменных и известны приближенные значения аргументов $\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_m$. Тогда трансформированная погрешность значения функции равна

$$\Delta y = f\ \tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_m\ - f\ x_1, x_2, ..., x_m\ .$$

(1.9)

Выражение (1.9) представляет собой полное приращение функции. Главной линейной частью приращения (1.9) является полный дифференциал, тогда

$$\Delta y \approx dy = \sum_{i=1}^{m} \frac{\partial y}{\partial x_i} dx_i\,. \qquad (1.10)$$

Из соотношения (1.10) получаем оценку *абсолютной трансформированной погрешности функции многих переменных*

$$\Delta\ y\ \approx |dy| = \left|\sum_{i=1}^{m} \frac{\partial y}{\partial x_i} dx_i\right| \le \sum_{i=1}^{m} \left|\frac{\partial y}{\partial x_i}\right| |\tilde{x}_i - x_i| = \sum_{i=1}^{m} \left|\frac{\partial y}{\partial x_i}\right| \Delta\ x_i\ \le \sum_{i=1}^{m} \left|\frac{\partial y}{\partial x_i}\right| \Delta_{x_i}\,, \qquad (1.11)$$

где Δ_{x_i} — предельная абсолютная погрешность переменной x_i.

Относительная трансформированная погрешность функции многих переменных равна

$$\delta\ y\ = \frac{\Delta\ y}{|y|} \le \sum_{i=1}^{m} \left| \frac{\partial y}{\partial x_i} \right| \frac{\Delta\ x_i}{|y|} = \sum_{i=1}^{m} \left| \frac{\partial y}{y \partial x_i} \right| \Delta\ x_i\ = \\ = \sum_{i=1}^{m} \left| \frac{\partial\ \ln y}{\partial x_i} \right| \Delta\ x_i\ \le \sum_{i=1}^{m} \left| \frac{\partial\ \ln y}{\partial x_i} \right| \Delta_{x_i}. \tag{1.12}$$

Трансформированные погрешности функции одной переменной получаются из (1.11) и (1.12) при $m = 1$.

С практической точки зрения важно определить *допустимую погрешность аргументов по допустимой погрешности функции* (обратная задача). Эта задача имеет однозначное решение только для дифференцируемых функций одной переменной $y = f\ x\ $ в случае, когда $f'\ x\ \neq 0$:

$$\Delta\ x\ = \frac{1}{|f'\ x\ |} \Delta\ y\ .$$

Для *функций многих переменных задача не имеет однозначного решения*, и необходимо ввести дополнительные ограничения. Например, если функция $y = f(x_1, x_2, \ldots, x_m)$ наиболее критична к погрешности $\Delta\ x_i\ $, то, пренебрегая погрешностью других аргументов, получаем

$$\Delta\ x_i\ = \Delta\ y\ \Big/ \left| \frac{\partial f}{\partial x_i} \right|.$$

Если вклад погрешностей всех аргументов примерно одинаков, то применяют *принцип равных влияний*: $\Delta\ x_i\ = \Delta\ y\ \Big/ m \left| \frac{\partial f}{\partial x_i} \right|, \quad i = \overline{1,m}.$

1.3. Свойства вычислительных задач и алгоритмов

1.3.1. Корректность вычислительной задачи

Пусть решается операторное уравнение общего вида

$$Ay = f\ , \tag{1.13}$$

где $A : Y \to F$ — некоторый оператор, действующий из метрического пространства Y в метрическое пространство F. Тогда задача нахождения элемента $y \in Y$ по заданному элементу $f \in F$ из уравнения (1.13) называется *корректно поставленной по Адамару*, или просто *корректной*, если [17]:

- решение $y \in Y$ существует при каждом $f \in F$;
- это решение единственно;