СБОРНИК ЗАДАЧ ПО МЕТОДАМ ВЫЧИСЛЕНИЙ

Под редакцией
П. И. Монастырного

2-е издание

Допущено
Министерством образования
Республики Беларусь в качестве
учебного пособия для студентов
высших учебных заведений
по специальностям "Математика",
"Прикладная математика"

Минск
«Университетское»
2000

УДК 519.6(075.8)
ББК 22.19я73
С 23

Рецензенты:
кафедра вычислительной техники и прикладной математики Брестского политехнического института; В. В. Бобков, профессор, доктор физико-математических наук

Авторы:
А. И. Азаров, В. А. Басик, Ю. А. Кремень, И. Н. Мелешко, П. И. Монастырный, В. А. Радаева, Н. П. Феденко, В. С. Федосенко, А. С. Шибут, Т. С. Якименко

Сборник задач по методам вычислений: Учеб.
С 23 пособие / Под ред. П. И. Монастырного.—2-е изд.— Мн.: Университетское, 2000.— 311 с.
ISBN 985-09-0343-0.

Данный сборник является одним из наиболее полных учебных пособий по решению задач и упражнений по вычислительной математике.

Рекомендуется для студентов вузов, а также для широкого круга специалистов в области прикладной математики.

УДК 519.6 (075.8)
ББК 22.19я73

Учебное издание

Азаров Алексей Иванович
Басик Василий Алексеевич
Кремень Юрий Алексеевич и др.

СБОРНИК ЗАДАЧ ПО МЕТОДАМ ВЫЧИСЛЕНИЙ

Учебное пособие

Второе издание

Редактор *А. В. Новикова.* Художественный редактор *Н. Б. Ярота.* Технический редактор *В. П. Безбородова.* Корректоры *Л. Н. Макейчик, Т. В. Кульнис.*

Подписано в печать 26.10.2000 Формат 84×108 1/32 Бумага газетная Гарнитура литературная. Печать высокая Усл. печ. л 16,38 Уч.-изд. л 19,26. Тираж 280 экз. Заказ 5252.

Налоговая льгота — Общегосударственный классификатор Республики Беларусь ОКРБ 007-98, ч. 1, 22.11.20.600

Издательское республиканское унитарное предприятие «Университетское» Государственного комитета Республики Беларусь по печати Лицензия ЛВ № 9 от 31.12.97 220048, Минск, проспект Машерова, 11

Отпечатано с диапозитивов Республиканского унитарного предприятия «Полиграфический комбинат им. Я. Коласа» в типографии «Перамога». 222310, Молодечно, ул. Тавлая, 11.

ISBN 985-09-0343-0

ПРЕДИСЛОВИЕ КО ВТОРОМУ ИЗДАНИЮ

Пособие представляет собой сборник задач по основным разделам курса «Методы вычислений и вычислительный практикум» для студентов механико-математических факультетов и факультетов прикладной математики университетов. Оно может быть рекомендовано для студентов университетских специальностей «Математика» и «Прикладная математика», а также для широкого круга специалистов, применяющих компьютеры при решении задач науки и техники.

Сборник отражает в основном имеющийся в этой области опыт и опыт авторов в проведении вычислительного практикума и лабораторных занятий по методам вычислений. Эти занятия, по нашему мнению, могут преследовать следующие цели:

1) усвоение и закрепление основных алгоритмов, понятий и определений вычислительной математики;

2) практическое решение типичных задач вычислительной математики, требующих небольшого объема вычислений, которые могут быть проведены в вычислительных лабораториях с помощью ПЭВМ или других вычислительных средств;

3) решение достаточно сложных в вычислительном отношении задач, требующих для их численной реализации использования мощных современных компьютеров.

При таких требованиях к занятиям у студентов имеется возможность изучить теорию основных вычислительных алгоритмов и реально убедиться в их действительных возможностях и свойствах на примере численного решения типичных модельных и прикладных задач. Этим принципам в определенной мере подчинено и построение учебного пособия. В начале каждого параграфа приводится краткое изложение метода, даются основные результаты и оценки, рассматривается решение несколь-

ких примеров и задач, даются рекомендации по численной реализации методов на компьютерах. В конце каждого параграфа приводится большое количество примеров и задач для самостоятельного выполнения.

Сборник задач состоит из 12 глав и в достаточно полной мере иллюстрирует содержание университетского курса методов вычислений.

В заключение приводятся ответы к ряду задач и примеров и достаточно полная библиография, содержащая наименования основных учебников, учебных пособий и книг по вычислительной математике и компьютерным приложениям.

Авторы выражают глубокую благодарность заведующему кафедрой вычислительной математики МГУ им. М. В. Ломоносова академику РАН Н. С. Бахвалову, заведующему кафедрой вычислительной математики БГУ, доктору физико-математических наук, профессору В. В. Бобкову, члену-корреспонденту НАН Беларуси, доктору физико-математических наук, профессору Л. А. Яновичу, доктору физико-математических наук, профессору И. П. Мысовских, кандидату физико-математических наук, доценту В. Г. Афонину за доброжелательную поддержку, ценные критические замечания и советы.

ГЛАВА 1

ЭЛЕМЕНТЫ ТЕОРИИ ПОГРЕШНОСТЕЙ

Под погрешностью понимается некоторая величина, характеризующая точность результата. Существует три вида погрешностей: 1) неустранимая погрешность (возникающая из-за неточности исходной информации, например неточности измерений); 2) погрешность метода; 3) погрешность вычислений (возникающая из-за округлений).

Основная задача теории погрешностей — указание области неопределенности результата [3, 6, 24].

1.1. Вычислительная погрешность

Рассмотрим процесс округления чисел. Если число $c=2{,}967393$ надо округлить, например, до пяти десятичных знаков после запятой, то имеем $c^*=2{,}96739$, т. е. если старший отбрасываемый разряд меньше 5, то предшествующая ему цифра в числе не меняется. Таким образом:

$$c=249{,}647339, \quad c^*=249{,}647;$$
$$c=17{,}94649, \quad c^*=17{,}946;$$
$$c=250331, \quad c^*=2{,}50\cdot 10^5.$$

Если $c=2{,}697393$ и его надо округлить до четырех знаков после запятой, то $c^*=2{,}6974$, т. е. если старший отбрасываемый разряд больше 5, то предшествующая ему цифра в числе увеличивается на 1, поэтому

$$c=2{,}396785, \quad c^*=2{,}3968;$$
$$c=172{,}397, \quad c^*=172{,}4;$$
$$c=183296, \quad c^*=1{,}833\cdot 10^5.$$

Если старший отбрасываемый разряд равен 5, то по общепринятому соглашению предшествующая ему чет-

ная цифра в числе не меняется ($c=3{,}965$; $c^*=3{,}96$), а нечетная увеличивается на единицу ($c=3{,}915$; $c^*=3{,}92$), например:

$c=1{,}9396712$,	$c=245{,}351365$,
$c^*=1{,}939671$,	$c^*=245{,}35136$,
$c^*=1{,}93967$,	$c^*=245{,}3514$,
$c^*=1{,}9397$,	$c^*=245{,}351$,
$c^*=1{,}940$,	$c^*=245{,}35$,
$c^*=1{,}94$,	$c^*=245{,}4$,
$c^*=1{,}9$,	$c^*=245$,
$c^*=2$;	$c^*=2{,}4\cdot 10^2$,
	$c^*=2\cdot 10^2$.

При округлении целого числа отброшенные знаки не следует заменять нулями, надо применять умножение на соответствующие степени 10.

В основе процессов округления лежит идея минимальности разности числа c и его округленного значения c^*.

Поведение вычислительной погрешности зависит от правила округлений и алгоритма численного решения задачи.

Пример 1. $S=25{,}71\cdot 1{,}42-3{,}21\cdot 7{,}46+0{,}93\cdot 7{,}75-4{,}31\cdot 2{,}69$.

1. Вычислить S точно.

2. Вычислить S и округлить его до двух знаков после запятой; результат обозначить S_1^*.

3. Вычислить каждое произведение с двумя знаками после запятой и просуммировать; результат обозначить S_2^*.

Установить различие между S, S_1^*, S_2^*. Сделать выводы.

Погрешности округлений возрастают в неустойчивых алгоритмах.

Пример 2. Вычислить $I_n=\frac{1}{e}\int\limits_0^1 x^n e^x dx$ ($n=0, 1, \ldots, 10$) по рекуррентной формуле $I_n=1-nI_{n-1}$, $I_0=1-1/e$ ($e\approx 2{,}718282$). Очевидно, что $0<I_{n+1}<I_n$, $I_n\to 0$ при $n\to\infty$. Выполнить вычисления и убедиться в неустойчивости алгоритма.

1.2. Абсолютная и относительная погрешности

Понятие об абсолютной и относительной погрешностях числа. Значащими цифрами числа называются все цифры в его записи, начиная с первой ненулевой слева, например:

1) $x=2{,}396029$ — все цифры (и нуль) значащие;

2) $x=0{,}00267$ — значащие только 2, 6, 7; первые три нуля незначащие, ибо они служат вспомогательной цели — определению положения цифр 2, 6, 7, поэтому может быть принята запись $x=2{,}67\cdot 10^{-3}$;

3) $x=2\,270\,000$ и $x=2{,}27\cdot 10^{6}$ — в первой записи все семь цифр (и последние четыре нуля) значащие, во второй значащие только 2, 2, 7.

Если известно, что c — число точное и $c=3200$, то для него нельзя использовать запись $c=3{,}2\cdot 10^{3}$, ибо тем самым два нуля переводятся в разряд незначащих цифр.

Пусть x — точное значение величины, а x^* — ее приближенное значение.

Абсолютной погрешностью числа x^* обычно называется величина Δx^*, удовлетворяющая условию $|x-x^*|\leqslant\Delta x^*$. *Относительной погрешностью* называется обычно некоторая величина δx^*, удовлетворяющая условию $\left|\frac{x-x^*}{x^*}\right|\leqslant\delta x^*$.

Точность результата лучше характеризует его относительная погрешность. Например, рассмотрим два числа: $\pi^*=3{,}14$ и $l^*=256\,795$. Известно, что $\pi=3{,}14159265...$ Значит, $\Delta\pi^*=0{,}0016$ (при записях Δx^* и δx^* более двух значащих цифр, как правило, не берут, так что в нашем случае $\Delta\pi^*=0{,}0016$). Тогда относительная погрешность $\delta\pi^*=0{,}0016/3{,}14=0{,}0005$, или 0,05 %. Известно, что $\Delta l^*=1$. Значит, $\delta l^*=1/256\,795=0{,}0000039$, или 0,00039 %. Хотя $\Delta\pi^*\ll\Delta l^*$, само число l^* определено точнее числа π^*.

Абсолютную и относительную погрешности числа принято округлять только в бóльшую сторону, так как при округлениях границы неопределенности числа, как правило, увеличиваются. По этой причине вычисления ведут с одним-двумя запасными знаками.

Верные значащие цифры. Значащая цифра называется *верной*, если абсолютная погрешность числа не превосходит 1/2 единицы разряда, соответствующего этой цифре.

Пример 1. Пусть $x^*=12{,}396$ и известно, что $\Delta x^*=0{,}03$. Сколько верных значащих цифр у числа x^*?

Имеем: $\Delta x^*>1/2\cdot 10^{-3}$; $\Delta x^*>1/2\cdot 10^{-2}$ и $\Delta x^*<1/2\cdot 10^{-1}$. Значит, у x^* верные знаки 1, 2, 3, а 9 и 6 сомнительные.

Пример 2. Пусть $x^*=0{,}037862$ и $\Delta x^*=0{,}007$.

Здесь $\Delta x^*<1/2\cdot 10^{-1}$. Значит, у числа x^* все цифры сомнительные.

Пример 3. Пусть $x^*=9{,}999785$ и $\Delta x^*=4\cdot 10^{-4}$.

Так как $\Delta x^*=0{,}4\cdot 10^{-3}<1/2\cdot 10^{-3}$, то у x^* три знака после запятой верные.

Если число имеет лишь верные цифры, его округленное значение имеет также лишь верные цифры.

Совпадение приближенного значения, имеющего все верные значащие цифры, с точным не обязательно.

При вычислениях желательно сохранить такое количество значащих цифр, чтобы их число не превышало числа верных цифр более чем на одну-две единицы. Эти одну-две последние цифры относят к разряду сомнительных цифр и называют *запасными* [6, 20].

ЗАДАЧИ

1. Округляя следующие числа до трех значащих цифр, определить абсолютную Δ и относительную δ погрешности полученных приближенных чисел:

1) 2,1514; 2) 0,16152; 3) 0,01204; 4) 1,225;
5) $-0{,}0015281$; 6) $-392{,}85$; 7) 0,1545; 8) 0,003922;
9) 625,55, 10) 94,525

2. Определить абсолютную погрешность следующих приближенных чисел по их относительным погрешностям:

1) $a=13\,267$, $\delta=0{,}1$ %, 2) $a=2{,}32$, $\delta=0{,}7$ %;
3) $a=35{,}72$, $\delta=1$ %; 4) $a=0{,}896$, $\delta=10$ %,
5) $a=232{,}44$, $\delta=1$ %.

3. При измерении некоторых углов получены числа $\alpha_1=21°37'3''$, $\alpha_2=45°$, $\alpha_3=75°20'44''$, $\alpha_4=1°10''$. Определить относительные погрешности чисел α_1, α_2, α_3, α_4, полагая абсолютную погрешность измерения равной $1''$.

4. Определить количество верных цифр в числе x, если известна его абсолютная погрешность:

1) $x=0{,}3941$, $\Delta x=0{,}25\cdot 10^{-2}$;
2) $x=0{,}1132$, $\Delta x=0{,}1\cdot 10^{-3}$;
3) $x=38{,}2543$, $\Delta x=0{,}27\cdot 10^{-2}$;
4) $x=293{,}481$, $\Delta x=0{,}1$;
5) $x=2{,}325$, $\Delta x=0{,}1\cdot 10^{-1}$;

6) $x=14{,}00231$, $\Delta x=0{,}1\cdot 10^{-3}$,
7) $x=0{,}0842$, $\Delta x=0{,}15\cdot 10^{-2}$;
8) $x=0{,}00381$, $\Delta x=0{,}1\cdot 10^{-4}$;
9) $x=-32{,}285$, $\Delta x=0{,}2\cdot 10^{-2}$;
10) $x=-0{,}2113$, $\Delta x=0{,}5\cdot 10^{-2}$.

5. Определить количество верных цифр в числе, если известна его относительная погрешность:

1) $a=1{,}8921$, $\delta a=0{,}1\cdot 10^{-2}$;
2) $a=0{,}2218$, $\delta a=0{,}2\cdot 10^{-1}$;
3) $a=22{,}351$, $\delta a=0{,}1$;
4) $a=0{,}02425$, $\delta a=0{,}5\cdot 10^{-2}$;
5) $a=0{,}000135$, $\delta a=0{,}15$;
6) $a=9{,}3598$, $\delta a=0{,}1\ \%$;
7) $a=0{,}11452$, $\delta a=10\ \%$;
8) $a=48361$, $\delta a=1\ \%$;
9) $a=592{,}8$, $\delta a=2\ \%$;
10) $a=14{,}9360$, $\delta a=1\ \%$.

1.3. Прямая задача теории погрешностей

Пусть в некоторой области G n-мерного числового пространства рассматривается непрерывно дифференцируемая функция $y=f(x_1, \ldots, x_n)$.

Предположим, что в точке $(x_1, \ldots, x_n)$ области G нужно вычислить значение $y=f(x_1, \ldots, x_n)$.

Пусть нам известны лишь приближенные значения $x_1^*, \ldots, x_n^*$ такие, что $(x_1^*, \ldots, x_n^*)\in G$, и их погрешности.

Вычислим приближенное значение $y^*=f(x_1^*, \ldots, x_n^*)$ и оценим его абсолютную погрешность.

Если воспользуемся формулой Лагранжа, то получим

$$\Delta y^*=|y-y^*|\lesssim\sum_{i=1}^{n}\Delta x_i^*\left|\frac{\partial}{\partial x_i}f(x_1^*, \ldots, x_n^*)\right|. \qquad (1.1)$$

Абсолютная погрешность дифференцируемой функции одного аргумента $y=f(x)$, вызываемая достаточно малой погрешностью аргумента Δx^*, оценивается величиной

$$\Delta y^*\lesssim|f'(x^*)|\Delta x^*. \qquad (1.2)$$

Погрешность результатов арифметических операций. *Погрешность суммы.* Пусть $f(x)\equiv x=x_1+\ldots+x_n$ и $x_i>$

> 0 $(i=1, \ldots, n)$. Здесь $f'x_i(x^*)=1$, поэтому из (1.1) имеем

$$\Delta y^* = \sum_{i=1}^{n} \Delta x_i^*.$$

Это означает, что абсолютная погрешность алгебраической суммы приближенных чисел равна сумме абсолютных погрешностей этих чисел. Очевидно, что абсолютная погрешность суммы не может быть меньше погрешности наименее точного из слагаемых. Поэтому сложение нескольких приближенных чисел, например $S=2{,}17+12{,}3971+1{,}198683+0{,}006732$, следует производить *по следующему правилу:*

1) выбираем число с наименьшим числом знаков после запятой: 2,17;

2) другие числа округляем до этого числа, сохранив два запасных знака: 12,3971; 1,1987; 0,0067;

3) суммируем: $S=15{,}7725$;

4) сумму S округляем на один знак: $\{S\}=15{,}772$.

Если $x=x_1+\ldots+x_n$ и $\Delta x_1^*=\ldots=\Delta x_n^*$, то $\Delta x^* = \sum_{i=1}^{n} \Delta x_i^* = n\Delta x_1^*$. При $n>10$ пользуются формулой Чеботарёва: $\Delta x^* = \sqrt{3n}\,\Delta x_1^*$.

Относительная погрешность суммы находится по правилу

$$\delta x^* = \sum_{i=1}^{n} \frac{x_i^*}{x^*}\,\delta x_i^*, \quad x^* = x_1^* + \ldots + x_n^*, \quad x_i^* > 0.$$

Пусть $\max\limits_i \delta x_i^* = M$, $\min\limits_i \delta x_i^* = m$; тогда

$$\delta x^* \leqslant \frac{x_1^* M + \ldots + x_n^* M}{x^*} = M,$$

$$\delta x^* \geqslant \frac{x_1^* m + \ldots + x_n^* m}{x^*} = m.$$

Таким образом, $m \leqslant \delta x^* \leqslant M$.

Погрешность вычитания. Пусть $x=x_1-x_2$ $(x_1>x_2>0)$ и известны Δx_1^*, Δx_2^*, x_1^* и x_2^*. Очевидно, что

$$\Delta x^* = \Delta x_1^* + \Delta x_2^*, \quad \delta x^* = \frac{x_1^*\delta x_1^* + x_2^*\delta x_2^*}{x^*}.$$

Если разность близка к нулю $(x^*=c\approx 0)$, то ее относительная погрешность значительно больше относительной

погрешности уменьшаемого и вычитаемого ($\delta x^* \gg \delta x_1^*$, δx_2^*). Имеет место потеря верных значащих цифр.

Пример 1. Найти разность чисел $x_1^* = 1{,}27569$, $x_2^* = 1{,}27531$. Известно, что у этих чисел четыре знака верные.

Разность $x^* = x_1^* - x_2^* = 0{,}00038$ не имеет ни одного верного знака, оба сомнительные; у суммы $x_1^* + x_2^*$ может быть четыре верных знака.

Пример 2. Найти разность $u = \sqrt{2{,}01} - \sqrt{2}$.

Если вычислять по формуле, взяв значения корней с десятью значащими цифрами: $\sqrt{2{,}01} = 1{,}417744688$, $\sqrt{2} = 1{,}414213562$, то в результате получим три значащие цифры: $u = 0{,}00353 = 3{,}53 \cdot 10^{-3}$. Преобразование

$$u = \frac{(\sqrt{2{,}01} - \sqrt{2})(\sqrt{2{,}01} + \sqrt{2})}{\sqrt{2{,}01} + \sqrt{2}} = \frac{0{,}01}{\sqrt{2{,}01} + \sqrt{2}}$$

позволяет получить тот же результат, вычисляя корни с тремя верными значащими цифрами.

Погрешность произведения. Пусть $f(x) \equiv x = x_1, \ldots, x_n$, известны Δx_i^* и x_i^* ($i = 1, \ldots, n$, $x_i > 0$); тогда абсолютная Δx^* и относительная δx^* погрешности произведения вычисляются по формулам [20, 24]

$$\Delta x^* = \sum_{i=1}^{n} \frac{x^*}{x_i^*} \Delta x_i^*, \quad \delta x^* = \sum_{i=1}^{n} \delta x_i^*.$$

При умножении приближенного числа на множитель $|k|$ относительная погрешность δx^* не меняется, а абсолютная увеличивается в $|k|$ раз.

Пример 3. Найти $u = x_1^* x_2^*$, где $x_1^* = 2{,}25$, $x_2^* = 1{,}0113$ (в этих числах все знаки верные).

Вычисления проведем *по следующему правилу:*

1) выбираем число с меньшим числом знаков после запятой: 2,25;

2) второе число округляем до этого числа, сохранив один знак запасной: 1,011;

3) вычисляем: $2{,}25 \cdot 1{,}011 = 2{,}27475$;

4) округляем: $u = 2{,}27$.

Если все сомножители имеют m верных значащих цифр и число сомножителей не более 10, то число верных значащих цифр произведения не более чем на две единицы меньше m.

Погрешность частного. Пусть $x=x_1/x_2$ и $x_1,\ x_2>0$. Очевидно, что

$$\Delta x^*=\frac{\Delta x_1^* x_2^*+\Delta x_2^* x_1^*}{(x_2^*)^2},\quad \delta x^*=\delta x_1^*+\delta x_2^*.$$

Практические рекомендации те же, что и при умножении приближенных чисел.

ЗАДАЧИ

6. Найти суммы приближенных чисел и указать их погрешности:

1) $0{,}145+321+78{,}2$ (все знаки верные);
2) $0{,}301+193{,}1+11{,}58$ (все знаки верные);
3) $398{,}5-72{,}28+0{,}34567$ (все знаки верные);
4) $x_1+x_2-x_3$, где $x_1=179{,}6$, $\Delta x_1=0{,}2$; $x_2=23{,}44$, $\Delta x_2=0{,}22$; $x_3=201{,}55$, $\Delta x_3=0{,}17$.

7. Указать правила оценки абсолютных и относительных погрешностей основных элементарных функций: $y=x^a$, $y=a^x$ $(a>0)$, $y=e^x$, $y=\ln x$, $y=\sin x$.

8. Найти произведение приближенных чисел и указать его погрешности (считать в исходных данных все знаки верными):

1) $3{,}49\cdot 8{,}6$; 2) $25{,}1\cdot 1{,}743$; 3) $0{,}02\cdot 16{,}5$;
4) $0{,}253\cdot 654\cdot 83{,}6$; 5) $1{,}78\cdot 9{,}1\cdot 1{,}183$; 6) $482{,}56\cdot 7256\cdot 0{,}0052$.

9. Найти частное приближенных чисел и указать его погрешности (считать в исходных данных все знаки верными):

1) $5{,}684:5{,}032$; 2) $0{,}144:1{,}2$; 3) $216:4{,}2$;
4) $726{,}676:829$; 5) $754{,}9367:36{,}5$; 6) $7{,}3:4491$.

10. Стороны прямоугольника равны $4{,}02\pm0{,}01$, $4{,}96\pm0{,}01$ м. Вычислить площадь прямоугольника.

11. Катеты прямоугольного треугольника равны $12{,}10\pm0{,}01$, $25{,}21\pm0{,}01$ см. Вычислить тангенс угла, противолежащего первому катету.

12. При измерении радиуса R круга с точностью до 0,5 см получилось 12 см. Найти абсолютную и относительную погрешности при вычислении площади круга.

13. Каждое ребро куба, измеренное с точностью до 0,02 см, оказалось равным 8 см. Найти абсолютную и относительную погрешности при вычислении объема куба.

14. Высота h и радиус основания R цилиндра измерены с точностью до 0,5 %. Какова относительная погрешность при вычислении объема цилиндра?

1.4. Обратная задача теории погрешностей

Обратная задача теории погрешностей состоит в определении допустимой погрешности аргументов по допустимой погрешности функции.

Для функции одной переменной $y=f(x)$ абсолютную погрешность можно приближенно вычислить по формуле

$$\Delta x^*=\frac{1}{|f'(x^*)|}\Delta y^*,\ f'(x^*)\neq 0.$$

Для функций нескольких переменных $y=f(x_1, \ldots, x_n)$ задача решается при следующих ограничениях.

Если значение одного из аргументов значительно труднее измерить или вычислить с той же точностью, что и значения остальных аргументов, то погрешность именно этого аргумента надо согласовать с требуемой погрешностью функции.

Если значения всех аргументов можно одинаково легко определить с любой точностью, то применяют принцип равных влияний, т. е. считают, что все слагаемые $|\partial f/\partial x_i|\,\Delta x_i^*\ (i=1, \ldots, n)$ равны между собой. Тогда абсолютные погрешности всех аргументов определяются формулой

$$\Delta x_i^*=\frac{\Delta y^*}{n\,|\partial f/\partial x_i|},\ i=1, \ldots, n.$$

Пример 1. Сторона квадрата равна 1 м. С какой точностью ее надо измерить, чтобы погрешность площади была не больше 1 см^2?

Решение. Пусть $l^*=100\pm\Delta l^*$ см, тогда $S^*=(100\pm\Delta l^*)^2=100^2\pm 2\cdot 100\cdot\Delta l^*+(\Delta l^*)^2$ см^2; $S=100^2$ см^2; $\Delta S^*=|S-S^*|=|\pm 2\cdot 100\Delta l^*+(\Delta l^*)^2|\leqslant 1$ см^2. Отсюда видно, что измерение стороны l должно быть с точностью $\Delta l^*\leqslant\frac{1}{2}\cdot 10^{-2}$ см.

Пример 2. Корни уравнения $x^2-2x+\lg 2=0$ нужно получить с четырьмя верными значащими цифрами. С каким числом верных значащих цифр надо взять свободный член уравнения?

Решение. Имеем $x_1=1+\sqrt{1-\lg 2}$, $x_2=1-\sqrt{1-\lg 2}$. По смыслу задачи надо x_1^* определить так, чтобы было $\Delta x_1^*\leqslant\frac{1}{2}\cdot 10^{-3}$ и $\Delta x_2^*\leqslant\frac{1}{2}\cdot 10^{-4}$, ибо при вычитании в значении x_2^* возможна потеря знаков. Требование $\Delta x_2^*\leqslant\frac{1}{2}\cdot 10^{-4}$ показывает, что выражение под корнем следует определить с четырьмя верными знаками.

Имеем $\lg 2=0{,}301$. Далее получим $x_1=1+\sqrt{0{,}699}$.

Поэтому $x_1^* = 1{,}836$ и все знаки верные. При вычислении x_2^* возьмем $\lg 2 = 0{,}3010$, $x_2^* = 1 - \sqrt{1 - 0{,}3010} = 1 - \sqrt{0{,}6990} = 1 - 0{,}8361 = 0{,}1639$; здесь четыре знака верные.

Пример 3. В пятизначных логарифмических таблицах даны десятичные логарифмы чисел с точностью до $\frac{1}{2} \cdot 10^{-6}$. Как велика может быть погрешность при нахождении числа по его логарифму, если число заключено между 300 и 400?

Решение. Имеем $300 \leqslant M \leqslant 400$. В таблицах даны величины y, где $y = \log x$; y^* определены так, что $|y - y^*| \leqslant (1/2) \cdot 10^{-6}$; $\Delta y^* \leqslant 1/2 \cdot 10^{-6}$. Далее $x = 10^y$, $x = f(y)$. Отсюда $x^* = 10^{y^*}$. По формуле (1.1) имеем $\Delta x^* = 10^{y^*} \ln y^* \Delta y^*$, $300 \leqslant 10^{y^*} \leqslant 400$ $(2{,}477 \leqslant y^* \leqslant 2{,}602)$; следовательно, $0{,}907 \leqslant \ln y^* \leqslant 0{,}956$. Таким образом, $\Delta x^* \leqslant 400 \cdot 0{,}956 \cdot (1/2) \cdot 10^{-6} \leqslant 0{,}191 \cdot 10^{-3}$. Это означает, что число по логарифму может быть найдено как минимум с тремя верными значащими цифрами после запятой, а если оно будет лежать вблизи точки 300, то его можно будет определить и с четырьмя верными значащими цифрами после запятой.

ЗАДАЧИ

15. Углы x измерены с абсолютной погрешностью Δx. Определить абсолютную и относительную погрешности функций $y = \sin x$, $y = \cos x$, $y = \operatorname{tg} x$. Найти по таблицам значения функций, сохранив в результате лишь верные цифры:

1) $x = 10°20'$, $\Delta x = 1'$;
2) $x = 48°42'31''$, $\Delta x = 5''$;
3) $x = 45°$, $\Delta x = 1'$;
4) $x = 50°10'$, $\Delta x = 0{,}05°$;
5) $x = 0{,}45$, $\Delta x = 0{,}5 \cdot 10^{-2}$;
6) $x = 1{,}115$, $\Delta x = 0{,}1 \cdot 10^{-3}$.

16. Вычислить значения следующих функций при указанных значениях аргумента x. Определить абсолютную и относительную погрешности результатов:

1) $y = x^3 \sin x$ при $x = \sqrt{2}$, полагая $\sqrt{2} \approx 1{,}414$;
2) $y = x \ln x$ при $x = \pi$, полагая $\pi \approx 3{,}142$;
3) $y = e^x \cos x$ при $x = \sqrt{3}$, полагая $\sqrt{3} \approx 1{,}732$.

17. Вычислить значения следующих функций при указанных значениях переменных. Определить абсолютную и относительную погрешности результатов, считая все знаки исходных данных верными:

1) $u = \ln(x_1 + x_2^2)$, $x_1 = 0{,}97$, $x_2 = 1{,}132$,
2) $u = (x_1 + x_2^2)/x_3$, $x_1 = 3{,}28$, $x_2 = 0{,}932$, $x_3 = 1{,}132$;
3) $u = x_1 x_2 + x_1 x_3$, $x_1 = 2{,}104$, $x_2 = 1{,}935$, $x_3 = 0{,}845$.

18. Определить относительную погрешность при вычислении полной поверхности усеченного конуса, если радиусы его оснований R и r и образующая l, измеренные с точностью до 0,01 см, следующие. $R=23{,}64$ см, $r=17{,}31$ см, $l=10.21$ см

19. Длина периметра правильного вписанного 96-угольника, которым пользовался Архимед при вычислении π, выражается при $r=1$ формулой $p=96\sqrt{2-\sqrt{2+\sqrt{2+\sqrt{2+\sqrt{3}}}}}$. Если вычислять непосредственно по этой формуле, желая получить π с точностью до 0,001, то с какой точностью нужно производить вычисления подкоренных величин?

20. Доказать, что предельная абсолютная погрешность минимальна при $y^*=(y_1+y_2)/2$, где

$$y_1=\inf_G f(a_1, \ldots, a_n),\ y_2=\sup_G f(a_1, \ldots, a_n).$$

1.5. О погрешностях вычислений на микрокалькуляторах

Результаты математических действий на восьмиразрядном микрокалькуляторе (МК) точны лишь в случае выполнения действий с целыми числами, если их мантиссы состоят не более чем из восьми цифр. Если промежуточные результаты попадут в область машинного нуля или бесконечности ($<10^{-99}$ или $>10^{99}$), то решение вообще может оказаться ошибочным. Обычно это происходит из-за неудачного выбора метода вычислений. Так, в примере [48]

$$\frac{1}{1-\sqrt{1-(0{,}0001)^2}}$$

восьмиразрядный МК после вычисления разности под знаком радикала удержит всего восемь цифр:

$$1-(0{,}0001)^2=1 \text{ вместо } 0{,}99999999,$$
$$\sqrt{1}=1,\ 1-\sqrt{1}=0,\ 1/0=\infty.$$

В результате вычислений индикатор покажет переполнение, хотя приведенное выражение имеет вполне определенное числовое значение. Воспользовавшись разложением в ряд Тейлора при $x\ll 1$, $\frac{1}{1-\sqrt{1-x^2}}\approx\frac{2}{x^2}$, при $x=0{,}0001$ получим

$$\frac{1}{1-\sqrt{1-(0{,}0001)^2}}\approx\frac{2}{(0{,}0001)^2}=2\cdot 10^8.$$

Погрешность полученного результата по сравнению с точным составляет 10^{-7} %, т. е. результат следует считать весьма точным.

Встроенные функции МК вычисляет при помощи рядов или итерационных схем, а максимальная погрешность дана в инструкции к МК. При выполнении математических действий МК удерживает и девятую цифру мантиссы числа, а при выводе на индикатор результат округляет.

Значение функций обычно вычисляют при помощи разложения в ряд Маклорена, причем суммирование прекращают, когда очередной член ряда оказывается меньше погрешности вычислений.

Обычно для вычисления одной и той же функции можно предложить несколько рядов, выбирая наиболее быстро сходящийся. Часто используют соотношения, связывающие эти функции, например,

$$\arcsin x = \operatorname{arctg} \frac{x}{\sqrt{1-x^2}}, \tag{1.3}$$

$$\arccos x = \frac{\pi}{2} - \arcsin x, \tag{1.4}$$

$$\operatorname{arctg} x = \frac{\pi}{2} - \operatorname{arcctg} x. \tag{1.5}$$

Показательную функцию самого общего вида можно вычислить по формуле

$$a^x = 1 + x \ln a + x^2 \frac{\ln^2 a}{2!} + \ldots + \frac{x^k \ln^k a}{k!} + \ldots . \tag{1.6}$$

Экспоненциальную функцию в случае $|x| < 1$ вычисляют при помощи ряда

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots + \frac{x^k}{k!} + \ldots . \tag{1.7}$$

Если же $|x| > 1$, то выделяют целую часть числа $n = [x]$, $x = n + \delta$, где $|\delta| < 1$, и вычисления проводят по правилу

$$e^x = e^{n+\delta} = \underbrace{e^{\delta} ee \ldots ee}_{n \text{ раз}},$$

где

$$e = 2{,}71828182459045\ldots, \quad 1/e = 0{,}36787944117144\ldots .$$

Величину e^{δ} вычисляют по разложению (1.7). Для вычисления логарифма при $|x| < 1$ используют ряд

$$\ln(1+x)=x-\frac{x^2}{2}+\frac{x^3}{3}-\frac{x^4}{4}+\frac{x^5}{5}\ldots. \qquad (1.8)$$

Ряд этот сходится очень медленно и используется лишь для малых значений x. Поэтому число x представляют в виде произведения целой части числа 2 на некоторое число z:

$$x=2^n z,\ 0{,}5\leqslant z\leqslant 1. \qquad (1.9)$$

Полагая

$$y=\frac{1-z}{1+z}, \qquad (1.10)$$

получают быстро сходящийся ряд

$$\ln x=n\ln 2-2\left(y+\frac{y^3}{3}+\ldots+\frac{y^{2k-1}}{2k-1}\right), \qquad (1.11)$$

где $\ln 2=C=0{,}693147180560\ldots$ — константа Эйлера [48].

Десятичные логарифмы вычисляют по формуле

$$\lg x=M\ln x,\ M=0{,}434294481903\ldots. \qquad (1.12)$$

Для вычисления прямых и обратных гиперболических функций используют разложения

$$\operatorname{sh} x=x+\frac{x^3}{3!}+\frac{x^5}{5!}+\frac{x^7}{7!}+\ldots, \qquad (1.13)$$

$$\operatorname{ch} x=1+\frac{x^2}{2!}+\frac{x^4}{4!}+\frac{x^6}{6!}+\frac{x^8}{8!}+\ldots, \qquad (1.14)$$

$$\operatorname{th} x=x-\frac{x^3}{3}+\frac{2x^5}{15}-\frac{17x^7}{315}+\frac{62x^9}{2835}-\ldots, \qquad (1.15)$$

$$\operatorname{arsh} x=x-\frac{1}{2\cdot 3}x^3+\frac{1\cdot 3}{2\cdot 4\cdot 5}x^5-\frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6\cdot 7}x^7+\ldots, \qquad (1.16)$$

$$\operatorname{arth} x=x+\frac{x^3}{3}+\frac{x^5}{5}+\frac{x^7}{7}+\frac{x^9}{9}+\ldots. \qquad (1.17)$$

При нахождении значения функции с помощью рядов удобно пользоваться двумя-тремя МК, выполняя на каждом одно арифметическое действие. Остается лишь считывать те или иные числа с одного и вводить в другой. Таким путем можно вычислить значение функций с 12—14 верными знаками, чего не могут дать обычные таблицы [48].

1.6. Погрешность округлений и запись чисел в ЭВМ

Вопросы представления чисел в ЭВМ и связанные с ними погрешности округления более подробно рассматриваются в [8, 38]. При ручном счете пользуются десятичной системой счисления, например $103{,}67 = 1 \cdot 10^2 + 0 \cdot 10^1 + 3 \cdot 10^0 + 6 \cdot 10^{-1} + 7 \cdot 10^{-2}$ (здесь 10 — основание системы счисления). ЭВМ работают, как правило, в двоичной системе, когда любое число записывают в виде последовательности нулей и единиц, например $0{,}0101 = 0 \cdot 2^0 + 0 \cdot 2^{-1} + 1 \cdot 2^{-2} + 0 \cdot 2^{-3} + 1 \cdot 2^{-4}$.

Как двоичная, так и десятичная система относятся к позиционным системам счисления. В позиционной системе с основанием r запись

$$a = \pm a_n a_{n-1} \ldots a_0, \; a_{-1} a_{-2} \ldots \tag{1.18}$$

означает, что

$$a = \pm (a_n r^n + a_{n-1} r^{n-1} + \ldots + a_0 r^0 + a_{-1} r^{-1} + a_{-2} r^{-2} + \ldots).$$

Запись вещественного числа в виде (1.18) называется его *представлением в форме числа с фиксированной запятой.* В ЭВМ чаще всего используется *представление чисел с плавающей запятой,* т. е. в виде

$$a = M r^p, \tag{1.19}$$

где r — основание системы счисления; p — целое число (положительное, отрицательное или нуль) и

$$r^{-1} \leqslant |M| < 1. \tag{1.20}$$

Число M представляется в форме числа с фиксированной запятой и называется *мантиссой* числа a. Число p называется *порядком* числа a.

В виде (1.19) можно единственным образом представить любое вещественное число, кроме нуля. Единственность обеспечивается *условием нормировки* (1.20).

Например, в ЭВМ БЭСМ-6 для записи числа, представленного в форме числа с плавающей запятой, отводится 48 двоичных разрядов, которые распределяются следующим образом:

Знак порядка	Порядок		Знак мантиссы	Мантисса	
48	47	42	41	40	1

Отсюда легко найти диапазон чисел, представимых в ЭВМ БЭСМ-6 (от 2^{-63} до 2^{63}, т. е. от 10^{-19} до 10^{19}). Ту же 48-разрядную сетку можно использовать для представления чисел с фиксированной запятой:

$$\underbrace{11\ldots1}_{23\text{ разр}},\ \underbrace{11\ldots1}_{24\text{ разр}}<2^{23}\approx10^{7}.$$

Следовательно, в данном случае диапазон допустимых чисел в 10^{12} раз меньше, чем при использовании представления с плавающей запятой.

1. Округление чисел в ЭВМ. Минимальное положительное число M_0, которое может быть представлено в ЭВМ с плавающей запятой, называется *машинным нулем* (для ЭВМ БЭСМ-6 $M_0\approx10^{-19}$). Число $M_\infty=M_0^{-1}$ называется *машинной бесконечностью.* Все вещественные числа, которые могут быть представлены в данной ЭВМ, расположены по абсолютной величине в диапазоне от M_0 до M_∞. Если в процессе счета какой-либо задачи появится вещественное число, меньшее по модулю, чем M_0, то ему присваивается нулевое значение. При появлении в процессе счета вещественного числа, большего по модулю, чем M_∞, происходит переполнение разрядной сетки, после чего ЭВМ прекращает счет задачи. Отметим, что нуль и целые числа представляются в ЭВМ особым образом — так, что они могут выходить за пределы диапазона $M_0\div M_\infty$. Число a, не представимое в ЭВМ точно, подвергается округлению, т. е. оно заменяется близким ему числом $\tilde{a}$, представимым в ЭВМ точно. Точность представления в ЭВМ чисел с плавающей запятой характеризуется относительной погрешностью $|a-\tilde{a}|/|a|$.

Величина относительной погрешности зависит от способа округления. Простейшим, но не самым точным способом округления является отбрасывание всех разрядов мантиссы числа a, которые выходят за пределы разрядной сетки.

Найдем границу относительной погрешности при таком способе округления. Пусть для записи мантиссы в ЭВМ отводится t двоичных разрядов. Предположим, что надо записать число, представленное в виде бесконечной двоичной дроби

$$a=\pm2^{p}\left(\frac{a_1}{2}+\frac{a_2}{2^2}+\ldots+\frac{a_t}{2^t}+\frac{a_{t+1}}{2^{t+1}}+\ldots\right), \qquad (1.21)$$

где каждое из a_j равно 0 или 1. Отбрасывая все лишние разряды, получаем округленное число

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

Таким образом, для погрешности округления

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

справедлива оценка

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

Далее, заметим, что из условия нормировки $|M| \geqslant 0{,}5$ следует, что в разложении (1.21) всегда $a_1 = 1$. Поэтому $|a| \geqslant 2^p \cdot 2^{-1} = 2^{p-1}$, и для относительной погрешности округления получим оценку

$$|a - \tilde{a}| / |a| \leqslant 2^{-t+1}.$$

При более точных способах округления можно изменить погрешность по крайней мере в 2 раза и добиться, чтобы выполнялась оценка

$$|a - \tilde{a}| / |a| \leqslant 2^{-t}.$$

Итак, относительная точность представления в ЭВМ чисел с плавающей запятой определяется числом разрядов t, отводимых для записи мантиссы. Точное число a и отвечающее ему округленное число $\tilde{a}$ связаны равенством

$$\tilde{a} = a(1 + \varepsilon), \qquad (1.22)$$

где $|\varepsilon| \leqslant 2^{-t}$. Число ε называется иногда *машинным эпсилоном*. Оно характеризует относительную точность представления чисел в ЭВМ. Для ЭВМ БЭСМ-6 имеем $t = 40$, $2^{-t} \approx 10^{-12}$, т. е. относительная точность представления чисел составляет 12 десятичных знаков.

Соотношение (1.22) справедливо лишь в случае $|a| \geqslant M_0$, где M_0 — машинный нуль. Если же число a мало, а именно $|a| < M_0$, то полагаем $\tilde{a} = 0$, что соответствует $\varepsilon = -1$ в формуле (1.22).

2. Накопление погрешностей округления. В процессе проведения вычислений погрешности округления могут

накапливаться, так как выполнение каждой из четырех арифметических операций вносит некоторую погрешность [38].

Будем обозначать округленное в системе с плавающей запятой число, соответствующее точному числу x, через $\mathrm{fl}(x)$. Считается, что выполнение каждой арифметической операции вносит относительную погрешность не более 2^{-t}. Это можно записать в виде

$$\mathrm{fl}(a * b) = a * b(1+\varepsilon), \tag{1.23}$$

где звездочка означает любую из операций $+$, $-$, $\times$, $:$ и $|\varepsilon| \leqslant 2^{-t}$.

Если результат выполнения арифметической операции является машинным нулем, то в формуле (1.23) надо положить $\varepsilon = -1$.

Может показаться, что предположение (1.23) не обосновано, так как, согласно (1.22), каждое из чисел a и b записывается с относительной погрешностью 2^{-t}; следовательно, погрешность результата может достигнуть 2^{-t+1}. Однако ЭВМ обладает возможностью проводить промежуточные вычисления с удвоенной значностью, т. е. с мантиссой, содержащей $2t$ разрядов, причем округлению до t разрядов подвергается лишь окончательный результат. Это обстоятельство позволяет добиться выполнения соотношения (1.23).

Для оценки влияния погрешностей округления на результат того или иного вычислительного алгоритма очень часто используется предположение о том, что результат вычислений, искаженный погрешностями округления, совпадает с результатом точного выполнения этого же алгоритма, но с иными входными данными.

Рассмотрим, например, процесс вычисления суммы трех положительных чисел:

$$z = y_1 + y_2 + y_3. \tag{1.24}$$

Пусть сначала вычисляется сумма $y_1 + y_2$. Тогда, согласно (1.23), получим

$$z_1 = \mathrm{fl}(y_1 + y_2) = (y_1 + y_2)(1+\varepsilon_1), \quad |\varepsilon_1| \leqslant 2^{-t}.$$

Затем в результате сложения z_1 и y_3 получим число $\tilde{z} = \mathrm{fl}(z_1 + y_3) = (z_1 + y_3)(1+\varepsilon_2)$, где $|\varepsilon_2| \leqslant 2^{-t}$. Таким образом, вместо точного значения суммы z получаем приближенное значение

$$\tilde{z} = (y_1 + y_2)(1+\varepsilon_1)(1+\varepsilon_2) + y_3(1+\varepsilon_2).$$

Отсюда видно, что результат выполнения алгоритма (1.24), искаженный погрешностями округления, совпадает с результатом точного выполнения того же алгоритма (1.24), примененного к другим исходным данным:

$$\tilde{y}_i=(1+\varepsilon_1)(1+\varepsilon_2)\,y_i,\ i=1,\ 2,\ \tilde{y}_3=(1+\varepsilon_2)\,y_3.$$

Этот же пример показывает, что результирующая погрешность зависит от порядка выполнения операций, так что вычисление суммы (1.24) в обратном порядке $(y_3+y_2)+y_1$ может привести к другому результату.

Практический интерес представляют оценки результирующей погрешности в зависимости от числа выполненных арифметических действий.

3. Оценки погрешностей округления. Приведем примеры оценок погрешностей округления, возникающих в результате выполнения вычислительных алгоритмов. Нас будет интересовать зависимость результирующей погрешности от числа арифметических действий n и от величины $\varepsilon=2^{-t}$, определяемой разрядностью ЭВМ.

Пример 1. Вычислить произведение

$$z_n=\prod_{j=1}^{n} y_j$$

вещественных чисел по формуле

$$z_j=y_jz_{j-1},\quad j=1,\ \ldots,\ n,\quad z_0=1. \tag{1.25}$$

Предположим, что в результате округления точного значения z_{j-1} получено приближенное значение $\tilde{z}_{j-1}$. Тогда, согласно (1.23), вместо $y_j\tilde{z}_{j-1}$ получим величину

$$\mathrm{fl}\,(y_j\tilde{z}_{j-1})=y_j\tilde{z}_{j-1}(1+\varepsilon_j),\ |\varepsilon_j|\leqslant\varepsilon=2^{-t}.$$

Таким образом, вместо z_j получаем

$$\tilde{z}_j=(1+\varepsilon_j)\,y_j\tilde{z}_{j-1},$$

т. е. приближенное значение $\tilde{z}_j$ удовлетворяет рекуррентному соотношению

$$\tilde{z}_j=\tilde{y}_j\tilde{z}_{j-1},\ j=1,\ \ldots,\ n,\ \tilde{z}_0=1, \tag{1.26}$$

где $\tilde{y}_j=y_j(1+\varepsilon_j)$. Результирующая погрешность равна

$$z_n-\tilde{z}_n=\prod_{j=1}^{n} y_j-\prod_{j=1}^{n}(1+\varepsilon_j)\,y_j,$$

поэтому относительная погрешность есть

$$\frac{z_n-\tilde{z}_n}{z_n}=1-\prod_{j=1}^{n}(1+\varepsilon_j).$$

Для оценки относительной погрешности заметим, что

$$|1+\varepsilon_j|\leqslant 1+\varepsilon,\ j=1,\ \ldots,\ n,\ \varepsilon=2^{-t},$$

поэтому с точностью до величин второго порядка малости относительно ε можно считать, что

$$\left|\frac{z_n-\tilde{z}_n}{z_n}\right|\leqslant n\varepsilon=n\cdot 2^{-t}. \qquad (1.27)$$

При выводе оценки (1.27) предполагалось, что $\varepsilon=2^{-t}$, т. е. при перемножении не возникает чисел, меньших машинного нуля или больших машинной бесконечности. Однако может оказаться, что на каком-то этапе вычислений в качестве промежуточного результата будет получен либо машинный нуль M_0, либо машинная бесконечность M_∞. Поскольку оба указанных случая приводят к неверному окончательному результату, необходимо видоизменить вычислительный алгоритм. Здесь существенным оказывается порядок действий.

Пусть, например, $M_0=2^{-p}$ и $M_\infty=2^p$ при некотором $p>0$. Предположим, что надо перемножить пять чисел: $y_1=2^{p/2}$, $y_2=2^{p/4}$, $y_3=2^{3p/4}$, $y_4=2^{-p/2}$, $y_5=2^{-3p/4}$. Каждое из этих чисел и их произведение $2^{p/4}$ принадлежат допустимому диапазону чисел $(M_0,\ M_\infty)$. Однако произведение $y_1y_2y_3$ равно $2^{3p/2}>M_\infty$, поэтому при указанном порядке действий дальнейшее выполнение алгоритма становится невозможным. Если проводить вычисление в порядке $y_5y_4y_3y_2y_1$, то получаем $y_5y_4=2^{-5p/4}<M_0$, следовательно, $\mathrm{fl}\,(y_5y_4)=0$ и все произведение окажется равным нулю, т. е. получим неверный результат. В данном примере к верному результату приводит вычисление произведения в порядке $y_5y_3y_1y_4y_2$.

В случае произвольного числа n сомножителей можно предложить следующий алгоритм вычисления произведения (см. [38]). Предположим, что $|y_1|\leqslant|y_2|\leqslant\ldots\leqslant|y_n|$, причем $|y_1|\leqslant 1$, $|y_n|\geqslant 1$. Будем сначала проводить умножение в порядке $y_1y_ny_{n-1}\ldots$ до тех пор, пока впервые не получим число, большее единицы. Затем полученное частичное произведение будем последовательно умножать на y_2, y_3 и т. д. до тех пор, пока новое

частичное произведение не станет меньше единицы. Процесс повторяем до тех пор, пока все оставшиеся сомножители не станут либо только большими единицы по модулю, либо только меньшими. Далее умножение проводится в произвольном порядке.

Пример 2. Вычислить сумму

$$z_n = y_1 + y_2 + \ldots + y_n. \qquad (1.28)$$

Предположим, что все y_j положительны и больше машинного нуля; тогда в процессе вычислений не может появиться нулевой результат.

Получим уравнение, которому удовлетворяет приближенное решение $\tilde{z}_j$. Предположим, что вместо точного значения z_{j-1} в результате накопления погрешностей округления получено приближенное значение $\tilde{z}_{j-1}$. Тогда, согласно (1.23), вместо z_j получим число

$$\tilde{z}_j = \mathrm{fl}\,(\tilde{z}_{j-1} + y_j) = (1+\varepsilon_j)(\tilde{z}_{j-1} + y_j),$$

где $|\varepsilon_j| \leqslant 2^{-t}$.

Таким образом, приближенное значение $\tilde{z}_j$ удовлетворяет разностному уравнению

$$\tilde{z}_j = q_j \tilde{z}_{j-1} + \tilde{y}_j,\ j = 1, \ldots, n,\ \tilde{z}_0 = 0, \qquad (1.29)$$
$$q_j = 1 + \varepsilon_j,\ \tilde{y}_j = (1+\varepsilon_j)\, y_j.$$

Имеем

$$\tilde{z}_n - z_n = \sum_{k=1}^{n} E_{nk} y_k, \qquad (1.31)$$

$$E_{nk} = \begin{cases} q_n - 1, & k = n, \\ q_n q_{n-1} \ldots q_{k+1} q_k - 1, & k = 1, \ldots, n-1. \end{cases} \qquad (1.30)$$

Коэффициент E_{nk} в формуле (1.30) указывает, какую долю погрешности вносит k-е слагаемое суммы (1.28) в общую погрешность. Покажем, что чем меньше номер k, тем большая погрешность вносится за счет y_k. Для этого оценим приближенно величины E_{nk}. Так как $q_j = 1 + \varepsilon_j$ и $|\varepsilon_j| < \varepsilon = 2^{-t}$, то $|q_n| \leqslant 1 + \varepsilon$, $|q_n q_{n-1} \ldots q_{k+1} q_k| < (1+\varepsilon)^{n-k+1}$.

Отбрасывая величины второго порядка малости относительно ε, можно считать, что $|q_n q_{n-1} \ldots q_k| \leqslant 1 + (n-k+1)\varepsilon$, и тогда

$$|E_{nk}| \leqslant (n-k+1)\,\varepsilon, \quad k=1, \ldots, n. \qquad (1.32)$$

Из формулы (1 30) легко получить оценку относительной погрешности $|\tilde{z}_n - z_n|/|z_n|$. Заметим, что для положительных $y_1, \ldots, y_n$ последовательность z_j неотрицательная и монотонно возрастающая, т. е. $0 \leqslant z_{k-1} \leqslant z_k$ $(k=1, \ldots, n)$. Поэтому для $y_k = z_k - z_{k-1}$ справедливо неравенство $0 \leqslant y_k \leqslant |z_k| + |z_{k+1}| \leqslant 2|z_n|$ $(k=1, \ldots, n)$. Отсюда и из (1.30) получим оценку

$$|\tilde{z}_n - z_n| \leqslant 2|z_n| \sum_{k=1}^{n} |E_{nk}|.$$

Учитывая приближенное неравенство (1.32), приходим к следующей оценке относительной погрешности:

$$\left| \frac{\tilde{z}_n - z_n}{z_n} \right| \leqslant \varepsilon n\,(n+1), \quad \varepsilon = 2^{-t}.$$

Следовательно, относительная погрешность, возникающая при суммировании n положительных чисел, оценивается примерно как $n^2 \cdot 2^{-t}$, где t — число разрядов, отводимое для записи мантиссы. Например, при $2^{-t} = 10^{-12}$, $n = 10^3$ получаем, что результирующая относительная погрешность не превзойдет 10^{-6}.

Для проверки степени влияния погрешностей округления на конечный результат рекомендуется произвести одни и те же вычисления по компьютерным программам с разным количеством значащих цифр, например сначала с одинарной, а затем с удвоенный точностью (с 7 и с 16 значащими цифрами соответственно), и сравнить полученные результаты.

Этот чрезвычайно простой и естественный прием позволяет во многих случаях практически оценить устойчивость алгоритма к погрешностям округления и меру обусловленности данной математический задачи.

ГЛАВА 2

ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ СИСТЕМ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

Методы решения систем линейных алгебраических уравнений (ЛАУ) можно разделить на две группы: 1) точные методы; 2) методы последовательных приближений [6, 24, 37, 38].

ЛИТЕРАТУРА

1. *Алберг Д., Нильсон Э., Уолш Д.* Теория сплайнов и ее приложения. М.: Мир, 1972. 318 с.

2. *Бахвалов Н. С.* Численные методы. М.: Наука, 1975. 632 с.

3. *Бахвалов Н С., Лапин А. В., Чижонков Е. В.* Численные методы в задачах и примерах. М.: Высш. шк., 2000. 190 с.

4. *Бахвалов Н. С., Жидков Н. П., Кобельков Г. М.* Численные методы. М.. Наука, 1987. 600 с.

5. *Березин И. С., Жидков Н. П.* Методы вычислений. Т. 1. М. Наука, 1966. 464 с.

6. *Березин И. С., Жидков Н. П.* Методы вычислений Т 2. М.: Физматгиз, 1962. 640 с.

7. *Бухтияров А. М.* и др. Практикум по программированию на Фортране (ОСЕС ЭВМ). М.: Наука, 1983. 304 с.

8 *Воеводин В. В., Кузнецов Ю. А.* Матрицы и вычисления. М.: Наука, 1984. 318 с.

9. *Годунов С. К., Рябенький В. С* Разностные схемы. М.: Наука, 1977. 440 с.

10. *Гринчишин Я. Т., Ефимов В. И., Ломакович А. Н.* Алгоритмы и программы на Бейсике. М.: Просвещение, 1988. 159 с

11. *Даугавет И. К.* Введение в теорию приближения функций. Л.: Изд-во ЛГУ, 1977. 184 с.

12. *Демидович Б. П., Марон И. А.* Основы вычислительной математики. М.: Наука, 1970. 664 с.

13. *Дробышевич В. И., Дымников В. П., Ривин Г. С.* Задачи по вычислительной математике. М.: Наука, 1980. 144 с.

14. *Дьяконов В. П.* Справочник по алгоритмам и программам на языке Бейсик для персональных ЭВМ. М.. Наука, 1987. 239 с.

15. *Завьялов Ю. С., Квасов Б. И., Мирошниченко В. Л.* Методы сплайн-функций. М : Наука, 1980. 352 с.

16. *Калиткин Н. Н.* Численные методы. М.: Наука, 1978. 512 с.

17. *Карпов В. Я., Корягин Д. А.* Пакеты прикладных программ. М.: Знание, 1983. 64 с.

18. *Касьянов В Н., Сабельфельд В. К.* Сборник заданий по практикуму на ЭВМ. М.: Наука, 1986. 272 с.

19. *Коллатц Л., Альбрехт Ю.* Задачи по прикладной математике. М.: Мир, 1978. 168 с.

20. *Копченова Н. В., Марон И. А.* Вычислительная математика в примерах и задачах. М : Наука, 1972. 368 с.

21. *Крылов В. И.* Приближенное вычисление интегралов. М : Наука, 1967. 500 с.

22. *Крылов В. И., Бобков В. В., Монастырный П. И.* Вычислительные методы высшей математики. Т. 1. Мн.: Вышэйш. шк., 1972. 584 с.

23. *Крылов В. И., Бобков В. В., Монастырный П. И.* Вычислительные методы высшей математики. Т. 2. Мн.: Вышэйш. шк., 1975. 672 с.

24. *Крылов В. И., Бобков В. В., Монастырный П. И.* Вычислительные методы. Т. 1. М.: Наука, 1976. 304 с

25. *Крылов В. И., Бобков В. В., Монастырный П. И.* Вычислительные методы. Т. 2 М Наука, 1977. 400 с.

26. *Крылов В. И., Бобков В. В., Монастырный П. И.* Начала теории вычислительных методов. Дифференциальные уравнения Мн.: Наука и техника, 1982. 286 с.

27. *Крылов В. И., Бобков В. В., Монастырный П. И.* Начала теории вычислительных методов. Уравнения в частных производных. Мн: Наука и техника, 1986. 311 с.

28. *Кудряшов И. А.* Программирование, отладка и решение задач на ЭВМ единой серии. Язык Фортран. Л.: Энергоатомиздат, 1988. С. 203.

29. *Кушниренко А. Г.* и др. Практикум по программированию/Под ред Н. С. Бахвалова, А В. Михалёва. М.· Изд-во МГУ, 1986. 434 с.

30. *Ляшко И. И., Макаров В. Л., Скоробогатько А. А.* Методы вычислений. Киев. Выщ. шк., 1977 406 с.

31. *Марчук Г. И.* Методы вычислительной математики. М.: Наука, 1989. 608 с.

32. *Марчук Г. И., Лебедев В. И.* Численные методы в теории переноса нейтронов. М.: Атомиздат, 1981. 453 с.

33. *Мысовских И. П.* Лекции по методам вычислений. М.: Наука, 1993. 496 с.

34. *На Ц.* Вычислительные методы решения прикладных граничных задач. М.: Мир, 1982. 294 с.

35. *Ортега Дж., Пул У.* Введение в численные методы решения дифференциальных уравнений М: Наука, 1986. 288 с.

36. *Петровский И. Г.* Лекции по теории интегральных уравнений. М.: Наука, 1965 128 с.

37. *Самарский А. А.* Введение в численные методы. М.: Наука, 1987. 288 с.

38. *Самарский А. А.* Теория разностных схем М.: Наука, 1983. 616 с.

39. *Самарский А. А., Гулин А. В.* Численные методы М.: Наука, 1989. 432 с.

40. *Самарский А. А., Николаев Е. С.* Методы решения сеточных уравнений. М.· Наука, 1987. 600 с.

41. *Светозарова Г И., Мельников А. А., Козловский А. В.* Практикум по программированию на языке Бейсик М.: Наука, 1988. 368 с.

42. *Скарборо Дж* Численные методы математического анализа М.: ИЛ, 1934 252 с.

43. *Трифонов Н. П, Пасхин Е. Н.* Практикум работы на ЭВМ. М.: Наука, 1982. 288 с.

44. *Уилкинсон Дж. Х., Райш К.* Справочник алгоритмов на языке Алгол. Линейная алгебра. М.: Машиностроение, 1976. 390 с.

45. *Фаддеев Д. К., Фаддеева В. П.* Вычислительные методы линейной алгебры. М.· Физматгиз, 1963. 386 с.

46. *Холл Дж., Уотт Дж.* Современные численные методы решения обыкновенных дифференциальных уравнений. М . Мир, 1979. 312 с

47 *Шаманский В. Е.* Методы численного решения краевых задач. Ч. 2. Киев: Наук. думка, 1982. 166 с.

48. *Шелест А. Е.* Микрокалькуляторы в физике: М.: Наука, 1988. 241 с.

ОГЛАВЛЕНИЕ

ISBN 985-09-0343-0