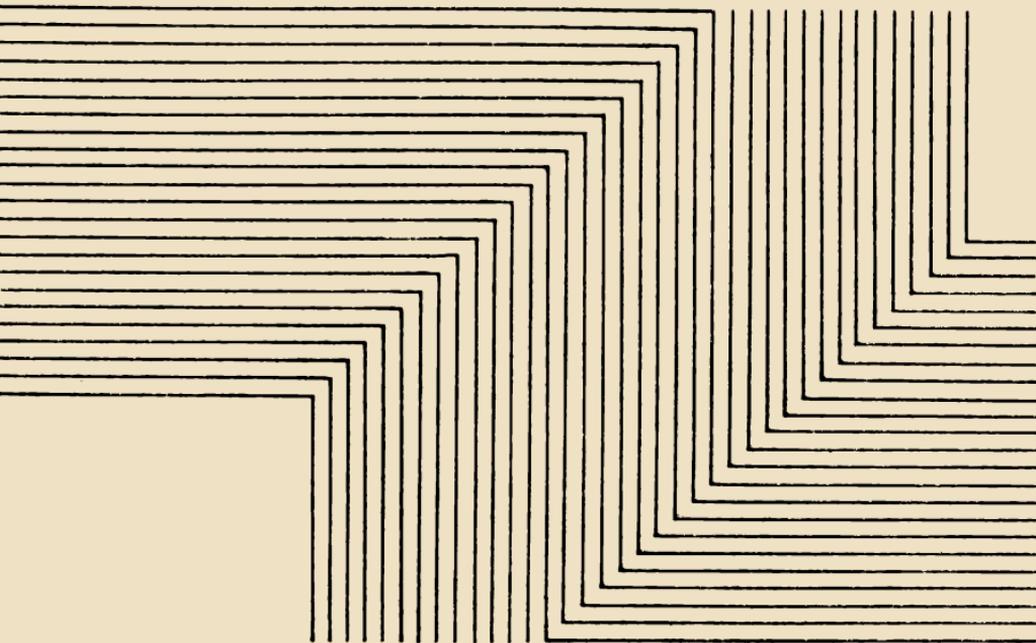


Ж. КУНЦМАН

# Численные методы



Ж. КУНЦМАН

---

# ЧИСЛЕННЫЕ МЕТОДЫ

Перевод с французского  
Е. И. СТЕЧКИНОЙ

под редакцией  
Д. П. КОСТОМАРОВА



МОСКВА «НАУКА»  
ГЛАВНАЯ РЕДАКЦИЯ  
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ  
1979

22.19  
К 91  
УДК 519.6

JEAN KUNTZMANN

# MÉTHODES NUMÉRIQUES

PIERMANN

Scan+DjVu: AlVaKo  
08/07/2022

**Численные методы.** Кунцман Ж. Перевод с франц. /Под ред. Д. П. Костомарова. — М.: Наука. Главная редакция физико-математической литературы, 1979.

Книга представляет собой элементарное введение в вычислительную математику. В ней содержатся понятие алгоритма, формы представления чисел, синтаксис алгебраических выражений. Значительное место уделено простейшим численным методам и методам табулирования.

Книга рассчитана на преподавателей средней школы, студентов педвузов, на учащихся школ и техникумов.

К  $\frac{20204-144}{053(02)-79}$  87-79. 1702070000

© Перевод на русский язык,  
Главная редакция  
физико-математической  
литературы  
издательства «Наука», 1979

## ОГЛАВЛЕНИЕ

Предисловие редактора . . . . .	5
Из предисловия автора . . . . .	9
<b>Г л а в а I. Алгоритмы . . . . .</b>	<b>9</b>
I. Общие сведения . . . . .	9
II. Запись чисел . . . . .	15
III. Действия над числами . . . . .	19
IV. Смена основания системы счисления . . . . .	26
Решения упражнений . . . . .	30
Решения задач . . . . .	32
<b>Г л а в а II. Синтаксис алгебраических выражений . . . . .</b>	<b>38</b>
I. Основные понятия . . . . .	38
II. Выражения без скобок . . . . .	40
III. Синтаксис выражений со скобками . . . . .	45
IV. Скобочные выражения . . . . .	49
V. Префиксная и постфиксная нотации . . . . .	55
Решения упражнений . . . . .	58
Решения задач . . . . .	59
<b>Г л а в а III. Приближения . . . . .</b>	<b>64</b>
I. Общие понятия . . . . .	64
II. Интервал приближения . . . . .	66
III. Применение нормы . . . . .	68
IV. Систематические десятичные приближения . . . . .	71
Решения упражнений . . . . .	78
Решения задач . . . . .	79
<b>Г л а в а IV. Понятие о численных методах . . . . .</b>	<b>81</b>
I. Система линейных алгебраических уравнений . . . . .	81
II. Многочлены. Интерполяция . . . . .	86
III. Квадратурные формулы . . . . .	94
IV. Дифференциальные уравнения с начальными условиями . . . . .	101
Решения упражнений . . . . .	105
Решения задач . . . . .	107
<b>Г л а в а V. Классификация и обработка погрешностей . . . . .</b>	<b>109</b>
I. Классификация погрешностей . . . . .	109
II. Распространение погрешностей . . . . .	112
III. Общие проблемы, относящиеся к погрешностям . . . . .	118

Решения упражнений . . . . .	121
Решения задач . . . . .	122
<b>Г л а в а VI. Проверка — контроль . . . . .</b>	<b>124</b>
I. Проверка . . . . .	124
II. Контроль . . . . .	134
Решения упражнений . . . . .	137
Решения задач . . . . .	138
<b>Г л а в а VII. Сведения о средствах вычислений и практиче- ские советы . . . . .</b>	<b>139</b>
I. Сведения о средствах вычислений . . . . .	139
II. Советы к выполнению вычислений . . . . .	143
Решения упражнений . . . . .	147
<b>Г л а в а VIII. Таблицы . . . . .</b>	<b>148</b>
Решения упражнений . . . . .	156
Решения задач . . . . .	156
Предметный указатель . . . . .	157

## ПРЕДИСЛОВИЕ РЕДАКТОРА

Одной из характерных особенностей нашего времени является широкое применение математических методов и электронно-вычислительных машин (ЭВМ) в самых различных областях человеческой деятельности. Бурный процесс математизации науки и техники начался в пятидесятых годах после появления и быстрого совершенствования ЭВМ. Он привел к формированию современной прикладной математики. В настоящее время математические методы и ЭВМ используются для решения больших научных, технических, экономических задач, для проектирования сложных объектов и управления их работой, для сбора и обработки информации в естественно-научных экспериментах, для поиска и реализации оптимальных режимов производственно-технологических процессов и т. д.

Однако вычислительные машины не работают сами. Они должны пройти соответствующее «обучение», т. е. получить программное обеспечение как общего, так и специально ориентированного характера. Общение с ними невозможно без знания языков программирования, алгоритмов, численных методов. По всем этим вопросам имеется много специальных книг, которые рассчитаны на профессионально подготовленных читателей, и очень мало популярной литературы. Предлагаемая книга Ж. Кунцмана в какой-то степени заполняет этот пробел. Рассчитанная на неспециалиста, она может быть использована для первого знакомства с широким кругом вопросов численного решения математических задач.

Несколько слов о ее содержании. В двух первых главах рассматриваются понятия алгоритма, алгоритмического языка, его синтаксиса и семантики, метаязыка,

как средства описания синтаксиса языков программирования. В трех следующих главах содержатся сведения о приближениях, численных методах и погрешностях, возникающих при проведении вычислений с конечнозначными числами. Три последних главы носят описательный характер. В них рассказывается о методах проверки и контроля, о средствах вычислений и о специальных вопросах, возникающих при составлении таблиц. Изложение ведется с привлечением большого числа примеров. Для закрепления материала автор предлагает много задач и упражнений, ответы на которые даны в конце каждой главы.

Книга не лишена недостатков. Изложение ряда вопросов дано слишком фрагментарно, формулировки отдельных теорем, задач, упражнений расплывчаты и неточны, доказательства порой не отличаются строгостью. В главе VII, посвященной средствам вычислений, ничего не сказано об электронно-вычислительных машинах. При работе над переводом некоторые неточности были устранены редакционным путем; кроме того, иногда приводятся подстрочные примечания, которые дают дополнительные пояснения читателям.

*Д. П. Костомаров*

## ИЗ ПРЕДИСЛОВИЯ АВТОРА

Математика — это та наука, которая имеет наибольшее применение как в других науках, так и в технике, и даже в повседневной жизни. Чтобы сделать применение математики более легким и эффективным, необходимо развивать ее аппарат. Например, десятичное исчисление настолько вошло в нашу жизнь, что никто не задумывается о том, что речь идет о математическом аппарате, предназначенном для облегчения практических действий с числами.

Математический инструмент имеет динамический характер. Так, говоря о наибольшем общем делителе чисел 72 и 32, мы имеем в виду число 8 в его отношении с этими двумя числами. Но можно также видеть процесс построения наибольшего общего делителя, на последнем этапе которого будет получено число 8. Динамический аспект приводит к двум важным следствиям:

— Человек не свободен от ошибок, и поэтому ошибки в рассуждении, в выкладках, в вычислении — это явления, к которым надо приспособиться; отсюда понятия проверки, контроля и т. д.

— Всякое человеческое действие имеет некоторую количественную оценку (во времени, в деньгах, в людях), откуда возникает понятие рентабельности и необходимость сравнивать методы.

Прикладные вопросы математики должны были бы составлять около четверти математического образования любого математика, ибо весьма важно иметь уравновешенный общий взгляд на науку, в которой работают. Еще более важно, чтобы преподаватели математики в школе были в достаточной мере знакомы с этими вопросами, поскольку именно с этой стороны дети начинают познавать математику. Кроме того, не следует забывать, что подавляющее

большинство учащихся будут использовать математику в практической деятельности.

Как преподнести преподавателям и учащимся сведения по прикладной математике? Существуют различные подходы. В настоящей книге предлагается следующее решение этого вопроса. Нет необходимости утяжелять программы, чтобы приучать учащихся к использованию математического аппарата. Оставаясь в рамках программы или в непосредственной ее окрестности, мы стараемся дать возможность посмотреть с новой точки зрения на числа, операции, формулы, на вычисления.

В первой главе представлены алгоритмы; запись чисел и операций над ними дают нам простейшие примеры цепочек. Работа с цепочками приводит к понятию синтаксиса. Этому посвящена вторая глава.

Как только хотят иметь дело с множеством действительных чисел, так сразу же приходится обращаться к понятию приближения и к понятию погрешности. Глава III показывает нам, как эти понятия вписываются в математику. В главе IV приводятся методы решения нескольких численных задач. Глава V показывает, как практически обрабатывать погрешности, а глава VI — как проводить проверки. Глава VII посвящена советам о практическом проведении вычислений. Глава VIII вводит нас в область построения таблиц.

Для записи алгоритмов используется язык, близкий к алголу. Это может вызвать недовольство тех, кто привык к этому языку. Однако несмотря на свои прекрасные качества, алгол — слишком строгий язык, и подчас немного замкнутый. Современная же тенденция (даже для общения с машинами) состоит в расширении и развитии языков, близких к естественному \*).

---

\*) Упрек автора в сторону алгола не обоснован. Современная тенденция расширения возможностей алгоритмических языков не означает, что правила грамматик этих языков станут менее строгими. (Прим. ред.)

## I. ОБЩИЕ СВЕДЕНИЯ

**1.1. Определение цепочки.** *Цепочка* есть конечное множество упорядоченных символов. Под этим мы понимаем то, что символы должны быть помещены один за другим на материальном носителе или в памяти машины. Например, множество чисел

$$\begin{array}{cccc} & & 4 & \\ & 1 & & 7 \\ 0 & & 3 & 8 \\ & & & 6 \\ 9 & 5 & & \end{array}$$

не является цепочкой. Для того чтобы оно стало цепочкой, надо расставить символы соответствующим образом.

**2.1. Разделитель.** Среди символов цепочки имеются вспомогательные знаки — разделители. (Их может быть достаточно много.) Приведем примеры разделителей:

- запятая в записи числа;
- пробелы, отделяющие слова в тексте;
- знак ■, используемый в полиграфии для указания места отсутствующей литеры;
- знаки препинания в тексте, такие как , ; . Напротив, такие знаки как ! и ? являются носителями некоторого оттенка и не могут рассматриваться как разделители.

**У п р а ж н е н и е 1.** Имеются ли разделители в цепочке  $a, b, c, d$ ?

**У п р а ж н е н и е 2.** Привести пример цепочки с несколькими типами разделителей.

**У п р а ж н е н и е 3.** Различны или нет цепочки

$$a + b + c \text{ и } a + c + b?$$

**2.2. Указатель.** *Указатель* есть вспомогательный знак, который перемещается вдоль цепочки либо слева направо (в этом случае он называется *прямым*), либо в обратном направлении (в этом случае он называется *обратным*). При этом указатель можно установить напротив определенной позиции цепочки, например, напротив разделителя или указанного знака. Указатель позволяет метить символ, с которым работают. Для одной и той же цепочки можно использовать различные указатели.

**2.3. Пример.** Пусть требуется найти в цепочке чисел наибольший или равный всем остальным элемент.

**1-й алгоритм.** Воспользуемся указателем  $P$ , пробегающим цепочку слева направо. Поместим в ячейку памяти  $M$  первый элемент цепочки, а затем установим указатель на начало цепочки. Если число  $a$ , на которое указывает указатель, больше числа, помещенного в  $M$ , то заменим число в  $M$  на  $a$ . Когда мы пробежим всю цепочку, в  $M$  окажется наибольший или равный всем остальным элемент в цепочке. В случае цепочки

1 4 2 18 5 13 11

$M$  будет содержать последовательно 1 4 18.

**З а м е ч а н и е.** В какой позиции цепочки находится наибольший элемент, заранее неизвестно.

**2-й алгоритм.** Воспользуемся указателями  $p$  и  $P$ . Сначала оба указателя устанавливаются на начало цепочки. Указатель  $p$  двигают вдоль цепочки. Когда число против указателя  $p$  больше числа против указателя  $P$ , указатель  $P$  ставится на место указателя  $p$ . Когда  $p$  пробежит цепочку,  $P$  будет указывать на наибольший элемент цепочки.

Пусть снова имеется цепочка 1 4 2 18 5 13 11.  $P$  будет последовательно указывать на 1 4 18.

**У п р а ж н е н и е 4.** а) Что произойдет, если цепочка имеет несколько наибольших (равных между собой) элементов?

б) Что произойдет, если  $P$  ставится на место  $p$ , когда число против  $p$  было больше или равно числу, находившемуся против  $P$ ?

**2.4. Другой пример.** В предыдущем примере мы работали с одной цепочкой. Приведем пример, когда, исходя из одной цепочки, мы построим другую.

Пусть требуется умножить десятичное число на 3. Будем рассматривать число как цепочку цифр  $a$ ,  $i$ -й

элемент которой обозначается  $a_i$ . Запишем цепочку  $a$  справа налево. Результатом умножения будет цепочка  $b$ ,  $i$ -й член которой обозначим  $b_i$ . Пусть  $R$  — позиция вне цепочки, содержащая число единиц переноса \*) из одного разряда в другой, и пусть  $b_i, R_{i+1}$  определяются из  $a_i$  и  $R_i$  при помощи таблиц. Появление разделителя (запятой) в цепочке  $a$  ведет к появлению запятой в цепочке  $b$ . Если в конце работы число единиц переноса не равно нулю, то его помещают в цепочке слева от последнего занятого места.

Таблица  $b_i$

$R_i$	0	1	2
$a_i$			
0	0	1	2
1	3	4	5
2	6	7	8
3	9	0	1
4	2	3	4
5	5	6	7
6	8	9	0
7	1	2	3
8	4	5	6
9	7	8	9

Таблица  $R_{i+1}$

$R_i$	0	1	2
$a_i$			
0	0	0	0
1	0	0	0
2	0	0	0
3	0	1	1
4	1	1	1
5	1	1	1
6	1	1	2
7	2	2	2
8	2	2	2
9	2	2	2

**3.1. Стег.** В предыдущих примерах речь шла либо о заданных фиксированных цепочках, либо о цепочках, строившихся последовательно в естественном порядке следования своих элементов. Однако часто встречаются цепочки, над которыми при их построении надо выполнять следующие операции:

- добавление нового элемента после последнего элемента цепочки;
- замена последнего элемента;
- исключение последнего элемента.

---

\*) При умножении  $a_i$  на 3 может получиться двузначное число ( $a_i = 7; a_i \cdot 3 = 21$ ); тогда число единиц произведения (в нашем примере 1) мы записываем в позиции  $b_i$ , а число десятков (в нашем примере 2), которые и будут единицами переноса, переносим в позицию  $b_{i+1}$ . (Прим. ред.)

Такая цепочка называется *стеком* \*).

**3.2. Пример.** Пусть требуется построить французские слова, состоящие не более чем из пяти букв. Будем составлять их в алфавитном порядке. Пробуем

*a*, что уже является французским словом, затем *aa*, что тоже является французским словом, затем *aaa*, что не является французским словом. После *artuf*, что не является французским словом, пробуем *artug*, что не является французским словом. После *arzzz* пробуем

*as*, что уже есть французское слово, затем

*asa* — тоже французское слово, затем

*asb* — не являющееся французским словом. После проверки

*zzzzz* процесс заканчивается.

При работе со стеком может использоваться указатель. Встречаются также *бистеки*, т.е. цепочки, в которых можно исключать, добавлять, заменять не только последний элемент, но и первый.

**4.1. Теория алгоритмов.** Из приведенных выше примеров следует, что введенные понятия позволяют описывать многочисленные алгоритмы при помощи ограниченного словаря, содержащего:

- имена цепочек;
- имена разделителей;
- имена прямых или обратных указателей;
- имена операций, таких, как
- передвинуть (указатель);
- установить (указатель);
- для стеков и бистеков:
- поместить;
- заменить;
- исключить.

Необходимы, разумеется, также слова для обозначения локальных действий, производимых на уровне каждого элемента стека.

**4.2. Описание алгоритма.** Предыдущие алгоритмы мы описали на разговорном языке. Но такое описание становится трудным для восприятия, как только оно хотя не-

---

\*) Часто стек изображают в виде цепочки, записанной не справа налево, т. е. в строку, а сверху вниз. При этом последним элементом цепочки считается верхний элемент. Механизм работы стека аналогичен механизму работы магазина винтовки. (*Прим. ред.*)

много усложняется. Язык алгол \*), который мы здесь будем использовать, и притом достаточно свободно, поскольку речь не идет о непосредственном переходе к машине, позволяет упростить описание алгоритма. Обозначение

$x :=$  выражение

означает, что величине  $x$  приписывается в качестве значения результат вычисления выражения, стоящего справа. Например,  $x := x + 1$  означает, что предыдущее значение  $x$  увеличивается на 1.

**Пример.** Первый алгоритм из п. 2.3.

**Алгоритм.**

1)  $M := a_1; P := 1.$

2) Для  $P := 2, \dots, n$

взять  $M := \max(M, a_P).$

3) Результат  $:= M.$

**Второй алгоритм.**

1)  $p := P := 1.$

2) Для  $p := 2, \dots, n$

взять:

если  $a_p > a_P$ , то  $P := p.$

3) Результат  $:= a_P.$

**4.3. Блок-схемы.** В некоторых случаях для более ясного представления алгоритма можно использовать блок-схемы. Описание первого алгоритма из п. 2.3 см. на блок-схеме 1\*\*).

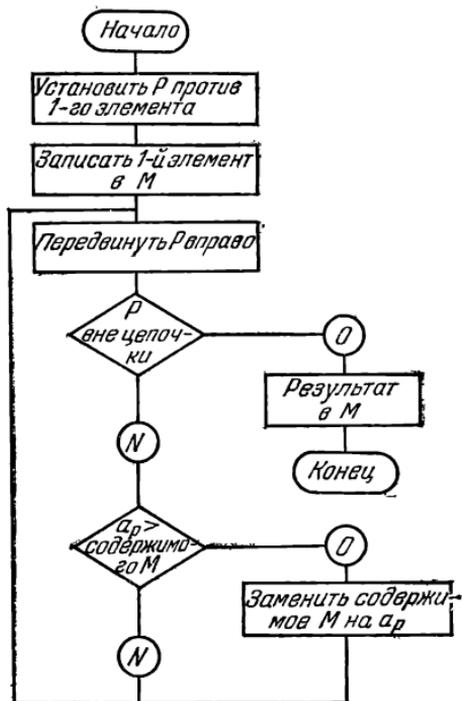
**Упражнение 5.**

Привести блок-схему второго алгоритма из п. 2.3.

**Упражнение 6. а)**

Описать алгоритм сложения двух многочленов, исходя из цепочек их коэффициентов.

**б)** Описать алгоритм умножения двух многочленов.



Блок - схема 1.

\*) Точнее, авторы используют некоторый алголоподобный язык, но не алгол. (Прим. ред.)

\*\*\*) На блок-схеме каждый ромбик соединен с двумя кружочками. В одном стоит буква  $N$ , в другом высказывание, написанное в ромбике, истинно, то мы движемся вдоль линии, проходящей через кружок с буквой  $O$ , иначе — через кружок с буквой  $N$ . (Прим. ред.)

**4.4. Схема Горнера.** Сейчас мы рассмотрим алгоритм, который будет часто использоваться. Пусть имеется цепочка чисел

$$a_0, a_1, \dots, a_n$$

и некоторое число  $x$ . Составим цепочку

$$b_0, b_1, \dots, b_n$$

$$b_0 := a_0$$

.....

$$b_{i+1} := b_i x + a_{i+1}$$

.....

Легко видеть, что

$$b_i = a_0 x^i + a_1 x^{i-1} + \dots + a_i.$$

Следовательно,  $b_n$  есть значение многочлена с заданными коэффициентами, для конкретного значения переменного  $x$ .

Этот процесс вычисления численного значения многочлена весьма эффективен, поскольку он насчитывает только  $n$  умножений и  $n$  сложений, тогда как прямое вычисление степеней  $x$  и членов многочлена насчитывает  $2n - 1$  умножений и  $n$  сложений.

**Задача 1.** Рассмотрим цепочку чисел и два указателя — прямой указатель  $d$  (пробегающий цепочку от начала к концу) и обратный указатель  $r$  (пробегающий цепочку от конца к началу). Если число, на которое указывает  $d$ , меньше или равно числу, на которое указывает  $r$ , то  $d$  сдвигают на следующий элемент цепочки; в противном случае перемещают  $r$ . Останавливаются, когда оба указателя указывают на один и тот же элемент.

- а) Что мы получили в результате такого алгоритма?
- б) Эффективен ли этот алгоритм?
- с) Сформулировать этот алгоритм на языке типа алгол.

**Задача 2.** Построить алгоритм, использующий цепочки, для нахождения НОД двух отличных от нуля чисел:

- а) последовательным делением;
- б) последовательным вычитанием;
- с) показать, что можно обойтись только двумя символами цепочки, а не использовать всю цепочку.

Определить алгоритм нахождения НОД  $n$  чисел:

- д) зная алгоритм нахождения НОД двух чисел;
- е) последовательным делением.

**Задача 3.** Описать алгоритм из п. 3.2:

а) посредством блок-схемы;

б) на языке типа алгол.

**Задача 4.** Описать алгоритм вычисления простых чисел, меньших  $A$ , при условии, что мы имеем цепочку делых чисел, меньших  $A$ . Предполагается, что мы умеем умножать и метить число из цепочки, равное результату.

**Задача 5.** Известно, что деление десятичных дробей  $a$  на  $b$  может привести, начиная с некоторого момента, к периодической последовательности десятичных дробей. Построить алгоритм с использованием цепочки, позволяющий найти этот период.

## II. ЗАПИСЬ ЧИСЕЛ

Мы применим приведенные выше понятия к некоторым простым вопросам, в частности, к изучению записи чисел.

**5.1. Замечания о цифровой записи чисел.** Нас интересуют числа, которые могут быть записаны в десятичной системе (или с основанием  $B$ ), и содержащие не более  $K$  символов. Некоторые из этих символов могут не быть цифрами; для их обозначения обычно используют знаки

+ - , .

Заметим, что таким способом можно записать только конечное число чисел. Иногда, в целях удобства, считают, что число помещено между двумя разделителями:

$F$  — помещается со стороны старших разрядов;

$f$  — помещается со стороны младших разрядов.

**5.2. Цифровая запись положительных целых чисел.** Положительное целое число записывают как цепочку цифр, следуя общему правилу: разделять цифры на триады, начиная с младшего разряда; разделители помещают между триадами.

Этот способ записи интересен только тем, что он позволяет сразу увидеть порядок величины числа. Впрочем, это так лишь в случае, если триад немного (скажем, не более 4).

**5.3. Название целых чисел.** Названия целых чисел подчиняются законам, подчас плохо объяснимым с логической точки зрения. В частности,

— некоторые числа имеют специальные названия (дюжина);

— названия триад высоких порядков плохо упорядочены.

Есть несколько способов чтения чисел; один из них состоит в следующем: цифры числа называются последовательно без указания их десятичного порядка:

3 562 087

читается как три, пять, шесть, два, нуль, восемь, семь вместо три миллиона пятьсот шестьдесят две тысячи семьдесят семь.

**6.1. Запись целых чисел.** Положительные и отрицательные целые числа можно обозначать по-разному. Так, в банках пишут положительные числа черными чернилами, а отрицательные — красными.

Обычный же способ обозначения состоит в использовании символов

+ и —

(при этом символ + можно опускать). Эти символы всегда помещаются на одном из концов цепочки, обычно со стороны старшего разряда.

**6.2. Случай числа нуль.** При записи числа нуль возникает трудность: это число можно записать как со знаком +, так и со знаком —, или вовсе без знака.

**6.3. Способы записи целых чисел без знака.** Относительные целые числа, состоящие не более чем из  $n$  цифр, можно записывать по основанию  $B$ , не указывая явно знака числа. Для этого достаточно взять  $n + 1$  позиций и записать отрицательное число  $a$  в виде

$$B^{n+1} + a.$$

Позиция наивысшего разряда содержит:

0 для положительного числа;

$B - 1$  для отрицательного числа.

П р и м е р.

$$\begin{array}{ll} n := 2 & \text{Основание } 10 \\ +36 & \text{записывается } 036 \\ -36 & \text{записывается } 964 \end{array}$$

Это соглашение используется, например, в арифмометрах. К тому же оно позволяет записывать нуль единственным образом.

**У п р а ж н е н и е 7.** а) Сколько различных относительных чисел, состоящих из  $K$  цифр, можно записать по основанию 10 при обычном способе обозначения?

б) То же самое при записи без знаков.

с) Сравнить и объяснить результаты п. п. а и б. Привести пример числа, которое может быть записано в одном соглашении и не может быть записано в другом.

**У п р а ж н е н и е 8.** Что экономичнее: записывать число по основанию 2 или 10, если стоимость десятичной цифры вчетверо больше стоимости двоичной цифры?

**7.1. Запись десятичных дробей.** Мы не будем говорить подробно об этой записи; отметим лишь то, что запятая в записи десятичных дробей рассматривается как разделитель.

**7.2. Запись очень больших или очень малых чисел.** В обычном виде запись очень больших и очень малых чисел неудобна. В самом деле, она приводит к продолжению цепочки значащих цифр вправо или влево с единственной целью — уточнить ранг единиц. Физики уже давно пользуются записью

$$a \cdot 10^b,$$

где  $b$  выбирается так, чтобы  $a$  содержало как можно меньше бесполезных нулей.

**П р и м е р.** Заряд электрона

$$e = 0,000\,000\,000\,000\,000\,000\,16 \text{ кул}$$

записывается  $1,6 \cdot 10^{-19}$  кул или  $16 \cdot 10^{-20}$  кул или  $0,16 \cdot 10^{-18}$  кул.

**8.1. Запись чисел в форме с плавающей запятой.** Предыдущее соглашение, т. е. запись чисел в виде  $a \cdot 10^b$ , оставляет одно неудобство — наличие запятой, что влечет:

— необходимость ставить ее;

— сложность работы с указателем, например, вправо или влево от запятой надо двигаться.

*Запись с плавающей запятой* исключает запятую, причем показатель  $b$  выбирается так, чтобы сделать число  $a$  меньшим 1. Таким образом, например, имеем  $0,16 \cdot 10^{-18}$  или  $0,016 \cdot 10^{-17}$ . Заметим, что символы 0 до запятой и, бесполезны, и можно писать просто  $16 \cdot 10^{-18}$  или  $016 \cdot 10^{-17}$ , условившись, что запятая помещается сразу слева от самой левой цифры. Число 16 (или 016) называется *мантиссой* (она может содержать знак). Число  $-18$  (или  $-17$ ) называется *порядком* (он также может содержать знак).

**8.2. Нормализованная запись числа с плавающей запятой.** Одно и то же число имеет бесконечно много форм записи с плавающей запятой. Выберем из них единствен-

ную, кроме случая числа нуль, уточнив, что самая левая цифра в записи числа должна быть отлична от 0. Величина заряда электрона в этой форме записи будет иметь вид  $16 \cdot 10^{-18}$ .

Эта запись называется *нормализованной записью* числа в форме с плавающей запятой.

**8.3. Случай числа нуль.** Следующее соглашение применимо лишь к числу нуль. Запись этого числа имеет нулевую мантиссу и произвольный порядок. Эту запись можно нормализовать, когда порядок имеет лишь конечное число допустимых значений, выбрав из них наименьшее.

**У п р а ж н е н и е 9.** Каковы наибольшее и наименьшее строго положительные числа, которые могут быть записаны по основанию 10, если имеются 6 позиций для мантиссы и 2 позиции для порядка (плюс одна позиция для знака)?

**У п р а ж н е н и е 10.** Построить алгоритм нормализации записи числа в форме с плавающей запятой с мантиссой из  $n$  цифр (принимая для 0 произвольный порядок).

**З а д а ч а 6.** а) Описать при помощи блок-схемы алгоритм, позволяющий указать триаду\*) тысяч целого числа, записанного в десятичной форме.

б) Как нужно изменить этот алгоритм, чтобы находить триаду миллионов?

**З а д а ч а 7.** а) Сколько потребуется слов, чтобы прочитать цифры чисел от 0 до 9999999?

б) Чтобы назвать эти числа?

**З а д а ч а 8.** а) Можно ли записывать относительные целые числа без знака посредством  $K$  цифр, придавая цифрам старших разрядов значения, отличные от 0 и  $B - 1$ ?

б) Исследовать случай четного  $B$  ( $B := 10$ ) и нечетного  $B$  ( $B := 5$ ).

**З а д а ч а 9.** Мы располагаем 8-ю позициями для десятичной записи положительных чисел. Сколько мы можем записать различных чисел:

а) целых;

б) в нормализованной форме с плавающей запятой, порядок которой имеет 2 цифры (без знака);

---

\*) Под *триадой* тысяч (миллионов и т. д.) понимается тройка последовательных цифр в записи числа, указывающая количество тысяч (миллионов и т. д.) в числе. (*Прим. ред.*)

с) в записи с плавающей запятой, порядок которой имеет 2 цифры (без знака);

д) в нормализованной записи с плавающей запятой со специальным символом 10, отделяющим мантиссу от порядка (без знака), принимая, что отсутствие порядка означает равенство его нулю?

### III. ДЕЙСТВИЯ НАД ЧИСЛАМИ

В качестве примеров использования понятия цепочки мы изучим перечисленные ниже задачи, не претендуя на исчерпывающее их исследование:

- распознавание равенства двух чисел;
- сравнение двух чисел;
- сложение двух чисел;
- вычитание двух чисел;
- умножение двух чисел;
- деление двух чисел.

**9.1. Равенство чисел.** Распознать равенство двух чисел может оказаться непросто, если эти числа, например, записаны в форме с плавающей запятой, и эта запись не нормализована.

Если же для каждого числа принята единственная запись, то проверка равенства состоит только в том, чтобы убедиться в совпадении символов, стоящих на соответствующих местах. В этом случае проверка может быть произведена в любом порядке.

**9.2. Сравнение чисел.** Мы рассмотрим только случай целых чисел со знаком (для нуля будет потребован один знак). Цель сравнения — выяснить, какое из соотношений

$$>, <, =$$

выполняется.

Отсюда сразу же выводится истинность или ложность дополнительных соотношений

$$\leq, \geq, \neq.$$

Отметим, что по отношению к сравнению чисел части числа классифицируются в порядке убывания важности следующим образом:

- знак;
- цифры в порядке убывания разрядов.

Таким образом, естественно предположить, что знак числа расположен слева от цифры самого старшего разряда.

**9.3. Нисходящее сравнение.** При этом способе сравниваются сначала знаки чисел, затем цифры каждой позиции в порядке убывания номера позиции. Для удобства сравнения оба числа надо записать так, чтобы каждое занимало  $n$  позиций, заменив недостающие цифры старших разрядов нулями.

Воспользуемся двумя указателями, отмечающими одну и ту же позицию в обоих числах. Первое появление различных символов в двух числах ведет к заключению

$>$  верно или  $<$  верно.

Попарное совпадение всех символов означает

$=$  верно.

**9.4. Восходящее сравнение положительных чисел.** При этом способе сравнения числа просматриваются, начиная с младших разрядов.

Для каждого числа берется свой указатель, причем оба указателя метят всегда позиции с одинаковыми номерами.

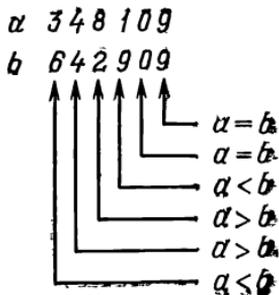
Пусть числа  $a$  и  $b$  записаны в виде

$$\begin{array}{l} a_n, a_{n-1} \dots a_1, \\ b_n, b_{n-1} \dots b_1. \end{array}$$

**А л г о р и т м.**

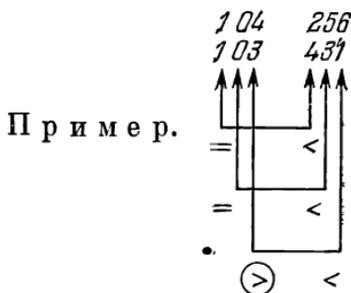
- 1)  $i := 0, a = b.$
- 2) Если  $i = n$ , то конец.
- 3)  $i := i + 1.$
- 4) Если  $a_i = b_i$ , то оставить предыдущее заключение и переходить к 2.
- 5) Если  $a_i > b_i$ , то  $a > b$ , иначе  $a < b$ .
- 6) Переходить к 2.
- 7) Результатом будет последнее полученное заключение.

**П р и м е р.**



**9.5. Ускоренное сравнение.** Можно сократить число шагов алгоритма, усложнив при этом сам алгоритм сравнения (восходящего или нисходящего). Вот как это делается.

Возьмем вторую пару указателей, перемещающихся в том же направлении, что и первая, но каждая пара указателей пробегает только половину цифр числа. Если в конце алгоритма указатели старших разрядов дадут  $a > b$  или  $b > a$ , то их заключение окончательно. Если же указатели старших разрядов дадут  $a = b$ , то окончательным будет заключение указателей, пробегающих младшие разряды.



**З а м е ч а н и е.** Увеличение сложности ради получения выигрыша во времени есть очень распространенный технический прием.

**У п р а ж н е н и е 11.** Пусть число записано в десятичной системе, но каждая из его цифр переведена в двоичную систему:

$$(0 \rightarrow 0000), \quad (1 \rightarrow 0001), \quad (9 \rightarrow 1001).$$

Показать, что можно осуществить восходящее или нисходящее сравнение непосредственно на двоично-десятичных цифрах.

**10.1. Сложение.** Предварительные замечания. В то время как результат сравнения есть логическое значение, результат сложения есть цепочка.

Когда речь идет о сложении относительных целых чисел, записанных со знаком, то выполнение операции зависит от знаков чисел. А поскольку, как мы знаем, сложение и вычитание начинаются, как правило, с младших разрядов, то удобнее было бы писать

$$314+ \text{ вместо } +314.$$

**10.2. Сложение положительных целых чисел.** На практике сложение состоит в переносе единиц из младших разрядов в старшие.

Начнем прибавлять цифры одинакового порядка, составляя одну цепочку частных сумм единиц и вторую — десятков.

Пример.

$$\begin{array}{r}
 342748643 \\
 157270652 \\
 \hline
 499918295 \text{ единицы} \\
 000010100 \text{ десятки}
 \end{array}$$

Заметим, что в каждой позиции:

— цифра десятков есть 0 или 1;

— если цифра единиц есть 9, то цифра десятков есть 0.

Передвинем теперь цепочку десятков на одну позицию влево и произведем обычное сложение этих двух чисел (из которых второе имеет очень специальный вид). Для сложения можно использовать следующий алгоритм:

А л г о р и т м.

1)  $r_0 := 0$ .

2) Для  $i := 1, 2, 3, \dots, n + 1$  взять:

$r_i :=$  если  $u_{i-1} = 9$  и  $d_{i-1} = 1$  или  $r_{i-1} = 1$ , то 1, иначе 0;

$v_i := u_i + r_i + d_i \pmod{10}$ .

3) Результат есть цепочка из  $v_i$ .

Пример.

$$\begin{array}{r|cccccccc}
 u & & 4 & 9 & 9 & 9 & 1 & 8 & 2 & 9 & 5 \\
 d & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & \\
 \hline
 r & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
 \hline
 v & 0 & 5 & 0 & 0 & 0 & 1 & 9 & 2 & 9 & 5
 \end{array}$$

**У п р а ж н е н и е 12.** а) Показать, что можно осуществить сложение двух чисел, повторяя достаточное число раз перенос, как это было сделано в п. 10.2.

б) Показать, что для двух чисел из  $n$  цифр требуется самое большее  $n + 1$  переносов.

в) Показать, что могут быть случаи, когда необходимо осуществить все  $n + 1$  переносов.

**У п р а ж н е н и е 13.** а) Показать, что всегда  $r_i d_i = 0$ .

б) Сформулировать алгоритм из п. 10.2 при основании 2 и показать, что

$$\begin{aligned} r_i &= u_{i-1} (d_{i-1} + r_{i-1}), \\ v_i &= u_i + r_i + d_i \pmod{2}. \end{aligned}$$

11.1. Сложение чисел, записанных без знака. Сложение чисел, записанных без знака (см. соглашение п. 6.3), выполняется очень просто, поскольку  $a$  ( $b$ ) записывается в виде

$$\begin{aligned} a(b), \text{ если оно положительно,} \\ B^{n+1} + a(B^{n+1} + b), \text{ если оно отрицательно.} \end{aligned}$$

Следовательно,  $a + b$  записывается:

$$a + b \text{ или } a + b + B^{n+1}$$

или

$$a + b + 2B^{n+1}.$$

С этим способом записи сопряжена одна трудность. В самом деле, допустим, что абсолютное значение записанных чисел не превосходит  $B^n$ . Но условия

$$|a| < B^n, \quad |b| < B^n$$

не исключают

$$|a + b| \geq B^n.$$

Этот случай мы учтем следующим образом:

— если  $a$  и  $b$  положительны, то цифра в  $n + 1$ -й позиции результата есть 1;

— если  $a$  и  $b$  отрицательны, то цифра в  $n + 1$ -й позиции результата есть  $B - 2$ .

Но такая запись неприемлема. Результат не может быть представлен в соответствии с принятыми соглашениями. В этом случае говорят о переполнении емкости (переполнении разрядной сетки).

11.2. Случай основания 2. Основание 2 представляет собой особенность, при которой

$$1 = B - 1, \quad 0 = B - 2,$$

и переполнение емкости не происходит. Воспользуемся этим, записывая отрицательные числа в виде

$$B^{n+2} + a.$$

Тогда два самых старших разряда будут иметь вид:

$$\left. \begin{array}{l} 00 \text{ для положительного числа,} \\ 11 \text{ для отрицательного числа,} \\ 10 \\ 01 \end{array} \right\} \text{ в случае переполнения емкости.}$$

**12.1. Бинарное умножение.** Бинарное умножение получается последовательными сложениями и сдвигами со следующими особенностями:

Цифры множителя, равные 0 или 1, прибавляются не более 1 раза к множимому в каждой сдвигаемой позиции.

Пример.

$$\begin{array}{r}
 110101 \\
 \underline{10110} \\
 1101010 \\
 11010100 \\
 1101010000 \\
 \hline
 10010001110
 \end{array}$$

**12.2. Ускоренное умножение в десятичной записи.** В некоторых арифмометрах умножение на 9 требует 9 оборотов рукоятки (или мотора, если машина электрическая). В этих условиях экономичнее заменить 9 на

$$10-1,$$

в результате чего потребуются только два оборота рукоятки:

- один на уровне десятков,
- один в обратном направлении, на уровне единиц.

В записи числа при этом используются цифры  $-1, -2, -3, -4, 0, 1, 2, 3, 4, 5$  вместо обычных цифр от 0 до 9. Для определения такой записи используем указатель, пробегающий число

$$a_n a_{n-1} \dots a_1$$

начиная справа, и переменную  $r$  для запоминания единицы переноса в следующем алгоритме.

А л г о р и т м.

1)  $r_1 := 0$ .

2) Для  $i := 1, \dots, n + 1$  взять:

если  $a_i + r_i > 5$ , то  $b_i := -10 + a_i + r_i$ ;

$r_{i+1} := 1$ , иначе  $b_i := a_i + r_i, r_{i+1} := 0$ .

Примеры.

$a$	3	4	8	9	2	$a$	3	5	8	9	2
$r$	0	1	1	0	0	$r$	1	1	1	0	0
$b$	3	5	-1	-1	2	$b$	4	-4	-1	-1	2

**Задача 10.** а) Примем за  $T_B$  время сравнения двух цифр при основании  $B$ ; указать среднее время сравнения

двух чисел из  $n$  цифр ( $n$  велико), записанных при основании  $B$ , при условии, что сравнение выполняется слева направо и прекращается по получении результата сравнения.

б) Если  $T_{10} = 4T_2$ , то какое из двух оснований  $B := 10$  или  $B := 2$  предпочтительнее?

с) Каково при заданном основании  $B$  отношение средних времен сравнения справа налево и слева направо?

**Задача 11.** Требуется сравнить два числа из  $n$  цифр, разбивая цифры числа на  $p$  групп по  $q$  цифр, причем с каждой группой работает свой оператор. Обозначим через  $T$  время сравнения двух цифр. Предположим также, что столько же времени требуется для объявления результата сравнения одной группы. Через  $c$  обозначим стоимость работы оператора в течение времени  $T$ .

а) Каковы стоимость и продолжительность всей операции при работе справа налево? Сравнить предыдущий ответ со стоимостью и продолжительностью выполнения сравнения в случае работы одного оператора.

б) Разумно ли брать  $p$  большим?

с) Ответить на те же вопросы, но для случая сравнения слева направо (использовать результат задачи 9). Представляет ли этот метод интерес?

**Задача 12.** Предположим, что производится сложение двух чисел из  $n$  цифр, разделенных на  $p$  групп по  $q$  цифр, и с каждой группой работает свой оператор. Для каждой группы, кроме самой правой, оператор будет выполнять операцию дважды, чтобы учесть возможный перенос 1 из предыдущих групп (допустим, что время второго сложения равно половине времени обычного сложения).

Пусть  $T$  — время сложения двух чисел из одной цифры каждое,  $T_0$  — время чтения одной цифры (и время замены операции). Найти:

а) продолжительность операции для одной группы;

б) время проделанной работы.

**Задача 13.** Возьмем снова условия задачи 7 для случая  $B := 10$ . Какие значения может принимать цифра в  $n + 1$ -й позиции, если мы хотим предотвратить переполнение емкости и иметь возможность:

а) различать очень большие положительные числа от очень малых отрицательных;

б) выдавать предупреждение, когда случается переполнение емкости.

**Задача 14.** В Древней Греции использовалась запись чисел, при которой (отличные от нуля) цифры единиц были представлены 9 первыми буквами алфавита, десятков — следующими 9 буквами и сотен — последними 9 буквами.

а) Насколько просто сравнить два числа в этой записи?

б) Тот же вопрос для сложения.

с) Тот же вопрос для умножения.

**Задача 15.** а) Найти алгоритм, позволяющий делить на 2 целое число, начиная справа.

б) Та же задача для деления на 3.

**Задача 16.** а) Разделить число на 2 с использованием нескольких операторов, каждый из которых работает с одной триадой.

б) Та же задача для деления на 3.

с) Каково отношение времен и стоимостей (в каждом случае) для той же работы, проведенной одним оператором?

#### IV. СМЕНА ОСНОВАНИЯ СИСТЕМЫ СЧИСЛЕНИЯ

**13.1. Кодирование, декодирование.** Смена основания системы счисления производится довольно часто, поскольку люди работают в системе счисления с основанием 10, а машины — почти всегда в системе с основанием 2.

Смена основания может происходить в двух направлениях: от старого основания, т. е. того, с которым постоянно работают, к новому (тогда говорят о кодировании), либо от нового к старому (в этом случае говорят о декодировании)\*).

**14.1. Кодирование целых чисел.** Пусть требуется записать в двоичной системе число, заданное в десятичной записи. Предполагается, что работа будет проводиться в десятичной системе. Двоичная цифра единиц есть оста-

---

\*) Для человека «старым» основанием будет 10, а «новым» 2. Поэтому переход из десятичной системы в двоичную, с точки зрения человека, есть кодирование, а в обратном направлении — декодирование. В то же время, с точки зрения машины, переход из десятичной системы в двоичную есть декодирование, а из двоичной в десятичную — кодирование. Это объясняется тем, что машина работает в двоичной системе счисления, а стало быть «старым» основанием для нее является 2. (Прим. ред.)

ток от деления числа на 2. Для получения следующей цифры берем остаток от деления частного на 2 и так далее. Получаем цепочку двоичных цифр с младшим разрядом в начале.

**Пример.**

$$\begin{array}{r|l}
 3851 & 1 \\
 1925 & 1 \\
 962 & 0 \\
 481 & 1 \\
 240 & 0 \\
 120 & 0 \\
 60 & 0 \\
 30 & 0 \\
 15 & 1 \\
 7 & 1 \\
 3 & 1 \\
 1 & 1
 \end{array}$$

Число имеет вид 111100001011.

**Упражнение 14.** а) Объяснить приведенный ниже процесс умножения (египетское умножение):

$$37 \times 41$$

$$\begin{array}{r|l}
 37 & 41 \\
 \cancel{74} & 20 \\
 \cancel{148} & 10 \\
 296 & 5 \\
 \cancel{592} & 2 \\
 \hline
 1184 & 1 \\
 \hline
 1517 &
 \end{array}$$

б) Объяснить, как умножить число  $a$ , записанное в десятичной системе, на целое число  $b$ , записанное в двоичной системе (результат представить в десятичной системе).

**Упражнение 15.** а) Преобразовать в недели, дни, часы, минуты и секунды 3 8 4 7 1 8 3 секунд.

б) Какое отношение к системам счисления имеет п. а)?

**14.2. Декодирование целого числа.** Исследуем обратную задачу.

Пусть имеется число, записанное в двоичной системе:

$$111100001011.$$

Найдем его десятичную запись, работая в десятичной системе.

Берем цифру старшего разряда, умножаем ее на 2, прибавляем к следующей цифре, умножаем на 2 и так далее. Получаем в точности промежуточные результаты из п. 14.1.

Можно заметить, что это сводится к вычислению по схеме Горнера значения при  $x := 2$  многочлена

$$1 + 1x + 0x^2 + 1x^3 + 0x^4 + 0x^5 + 0x^6 + 0x^7 + 1x^8 + \\ + 1x^9 + 1x^{10} + 1x^{11}.$$

**У п р а ж н е н и е 16.** а) Объяснить, как переводится в секунды продолжительность, выраженная в неделях, днях, часах, минутах, секундах.

б) Можно ли рассматривать это как вычисление значения многочлена?

**15.1. Кодирование числа без целой части.** Пусть требуется записать в двоичной системе число

$$0,1483.$$

Будем работать в десятичной системе. Это число в двоичной системе запишется в виде  $0,xyz\dots$

Для получения  $x$  достаточно заметить, что это есть целая часть удвоенного числа. Выделим эту целую часть. Действуя снова так же, получим  $y$ , затем  $z$  и т. д. Итак, получаем цепочку со старшим разрядом во главе.

**П р и м е р.**

0	1483	Искомое число
0	2966	0,0010010111
0	5932	
1	1864	
0	3728	
0	7456	
1	4912	
0	9824	
1	9648	
1	9296	
1	8592	

**З а м е ч а н и е.** Число без целой части, имеющее конечную десятичную запись, вообще говоря, может не иметь конечной двоичной записи.

**У п р а ж н е н и е 17.** В каком случае переход от основания  $B$  к основанию  $B'$  сохраняет конечность записи?

**15.2. Декодирование числа без целой части.** Если двоичная запись числа содержит  $n$  цифр после запятой, то это так и для десятичной. Но 10 двоичных цифр дают почти такую же точность, как 3 десятичных, поскольку

$$2^{10} \approx 10^3.$$

Таким образом, во всех вопросах, куда входят приближения, мы, вообще говоря, не будем сохранять все найденные десятичные цифры.

**15.3. Практика декодирования.** Пусть требуется двоичное число

$$0,0010010111,$$

найденное в п. 15.1, записать в десятичной системе, работая в десятичной системе.

Заметим, что искомое число есть значение для  $x := 1/2$  многочлена

$$0x + 0x^2 + 1x^3 + 0x^4 + 0x^5 + 1x^6 + 0x^7 + 1x^8 + 1x^9 + 1x^{10}.$$

Применим схему Горнера, начиная с вычисления младших разрядов. Берем цифру младшего разряда, умножаем на 0,5, прибавляем к следующей, умножаем на 0,5 и т. д. Получаем

$$\begin{array}{r} 1 \\ 1,5 \\ 1,75 \\ 0,875 \\ 1,4375 \\ 0,71875 \\ 0,359375 \\ 1,1796875 \\ 0,58984375 \\ 0,294921875 \\ 0,1474609375 \end{array}$$

**З а м е ч а н и е.** Результат кажется отличным от числа, из которого мы исходили в п. 15.1. В действительности же  $8,5/10^4 < 1/1024$ , и значит, мы находимся в допустимых пределах.

**З а д а ч а 17.** а) Имеется простое соотношение между записью числа в двоичной системе и в восьмеричной.

Применить это соотношение к числу 3417 (записанному по основанию 8).

б) Назовем двоично-пятеричной запись, в которой для последовательных порядков используются поочередно основания 5 и 2. Как получить в двоично-пятеричной записи число, записанное в десятичной системе?

с) Проверить на числе 3724 (записанном по основанию 10).

## РЕШЕНИЯ УПРАЖНЕНИЙ ГЛАВЫ I

1) Да, запятые в этой цепочке могут рассматриваться как разделители.

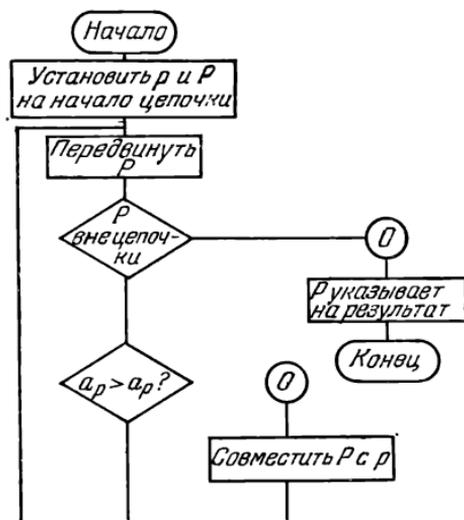
2) Запись десятичного числа триадами (между триадами либо ставится запятая, либо оставляется пробел).

3) Различны.

4) а) Будет найден тот из этих элементов, который стоит первым в цепочке.

б) Будет найден элемент, больший или равный всем другим, который стоит последним в цепочке.

5) См. блок-схему 2.



Блок-схема 2.

б) а) Построим цепочку  $c$  на основе цепочек  $a$  и  $b$ :

$$c_i := a_i + b_i.$$

Порядок вычисления  $c_i$  безразличен.

б) Воспользуемся двумя указателями:  $p_1, q_1$  для первого многочлена и  $p_2, q_2$  для второго;  $q_1$  перемещается влево от  $p_1$ ,  $q_2$  перемещается вправо от  $p_2$ .

Вычисляем

$$\sum a_{q_1} b_{q_2}$$

до тех пор, пока это возможно.

Сначала  $p_1$  и  $p_2$  находятся в самых левых позициях. После вычисления каждой  $\Sigma$  перемещаем  $p_1$  вправо или, если это невозможно, перемещаем вправо  $p_2$ . Вычисление заканчивается, когда  $p_1$  и  $p_2$  находятся в самых правых позициях.

7) а)  $2 \cdot 10^{K-1} - 1$ .

б)  $2 \cdot 10^{K-1}$ .

с) Разница равна 1. Это происходит потому, что при обычном соглашении число нуль имеет две записи. Число  $-10^{K-1}$  не может быть записано в первом соглашении. Оно записывается: 9000 во втором (для  $K := 4$ ).

8) Чтобы записать число по основанию  $B$ , необходимо около  $\log_B n + 1$  цифр. Но

$$\frac{4 \cdot \log_{10} n}{\log_2 n} = 4 \cdot \log_{10} 2 \approx 1,2.$$

Двоичная запись экономичнее.

9)  $0,9999999 \cdot 10^{99} \approx 10^{99}$ ,

$0,000001 \cdot 10^{-99} \approx 10^{-105}$ .

10)  $P$  — указатель, пробегающий мантиссу начиная слева, а  $e$  — показатель.

А л г о р и т м.

1)  $P := 1$ .

2) Если  $P := n$ , то конец.

3) Если  $a_P \neq 0$ , то конец.

4) Если  $e$  — наименьшее возможное, то нет решения.

5)  $P := P + 1$ ;  $e := e - 1$ ; перейти к 2.

11) Отображение: «десятичное число  $\rightarrow$  двоичное число» возрастает.

12) а), б) Можно рассматривать разбиение единиц и десятков как сложение. На каждом этапе по крайней мере одна позиция, считая справа, становится окончательной.

с)  $(10^n - 1) + 1$  требует  $n + 1$  шагов.

13) а)  $r_i = 1$  требует  $u_{i-1} = 9$ , что влечет  $d_i = 0$ .

б) Проверить все возможные случаи.

14) а) Колонка справа представляет собой последовательность частных, дающая двоичную запись числа 41. Колонка слева содержит  $37 \cdot 2^n$ . Незачеркнутые числа соответствуют позициям двоичной записи числа 41, содержащим 1.

б) Составим  $a \cdot 2^n$ . Сложим среди них те, которые соответствуют цифрам 1 целого числа.

15) а) 6 недель 2 дня 12 часов 39 минут 43 секунды.

б) Запись целого числа, последовательные порядки, имеющие различные основания: 60, 60, 24, 7 (и кроме того, опять, разумеется, основание 10).

16) а)  $s, s \cdot 7 + j, (s \cdot 7 + j) \cdot 24 + h, \dots$

б) Это есть значение многочлена от  $\alpha, \beta, \gamma, \delta$ :

$$s\alpha\beta\gamma\delta + j\beta\gamma\delta + h\gamma\delta + m\delta + s$$

при  $\alpha := 7, \beta := 24, \gamma := 60, \delta := 60$ .

17) Необходимо (и достаточно), чтобы любой простой сомножитель числа  $B$  был простым сомножителем числа  $B^t$ .

## РЕШЕНИЯ ЗАДАЧ ГЛАВЫ I

Задача 1. а) То же элемент, что и в упражнении 4, б).

б) Разумеется, поскольку число перемещений указателя является наименьшим из всех возможных.

в) Алгоритм.

1)  $d := 1, r := n$ .

2) Если  $d = r$ , то перейти к 5.

3) Если  $a_d \leq a_r$ , то  $d := d + 1$ , иначе  $r := r - 1$ .

4) Перейти к 2.

5) Результат :=  $a_d$ .

Задача 2. а) Построим цепочку, начинающуюся с заданных чисел  $a_0$  и  $a_1$ .

Алгоритм.

1)  $r := 1$ .

2)  $a_{r+1} :=$  остаток от деления  $a_{r-1}$  на  $a_r$ .

3) Если  $a_{r+1} = 0$ , то перейти к 5.

4)  $r := r + 1$ ; перейти к 2.

5) Результат :=  $a_r$ .

б) Алгоритм.

1)  $r := 1$ .

2) Если  $a_r = 0$ , то перейти к 5.

3)  $a_{r+1} :=$  если  $a_{r-1} < a_r$ , то  $a_{r-1}$ , иначе  $a_{r-1} - a_r$ .

4)  $r := r + 1$ ; перейти к 2.

5) Результат :=  $a_{r-1}$ .

в) Мы всегда используем только  $a_{r-1}$  и  $a_r$ . Достаточно заменить их на  $a_r$  и  $a_{r+1}$ . Стало быть, нет необходимости рассматривать алгоритм цепочкой.

д) 1. Алгоритм.

Найдем НОД для  $a_{2i+1}$  и  $a_{2i+2}$  и поместим его справа от цепочки. Процесс закончится, когда останется только одно число.

Алгоритм.

1)  $P := 0; Q := n$ .

2) Если  $P = Q - 1$ , то перейти к 4.

3)  $Q := Q + 1; a_Q = \text{НОД}(a_{P+1}, a_{P+2});$

$P := P + 2$ ; перейти к 2.

4) Результат :=  $a_Q$ .

2. Алгоритм.

Найдем НОД первого и последнего элементов цепочки. Исключим первый элемент, а последний заменим НОД. Вычисление заканчивается, когда остается только один элемент.

Алгоритм.

1)  $P := 1$ .

2) Если  $P := n$ , то перейти к 4.



$p$  метит простое число, из которого образуются кратные,  $q$  метит последовательные множители.  $q$  начинает двигаться от  $p$  и перемещается вправо; если  $q$  зачеркнуто, то не происходит ничего, если же  $q$  не зачеркнуто, то образуется произведение  $pq$  и зачеркивается. Движение  $q$  останавливается, когда  $pq$  выходит из цепочки;  $p$  сначала стоял на 2, а когда  $pq$  выходит из цепочки,  $p$  перемещается вправо до незачеркнутого числа. Процесс заканчивается, когда  $p$  выходит из цепочки.

**Задача 5.** Установление периодичности состоит в установлении периодичности остатков. Составим цепочку, содержащую остатки. Отметим первый остаток, который повторяется. Используем указатель  $P$ , отмечающий последний остаток, и указатель  $Q$ , пробегающий цепочку остатков.

**А л г о р и т м.**

1)  $P := 1; r_P := a.$

2)  $Q := 1.$

3) Если  $Q = P$ , то  $P := P + 1$ ; определить  $r_P$ , перейти к 2.

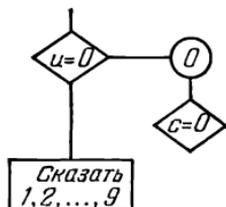
4) Если  $r_P = r_Q$ , то перейти к 6.

5)  $Q := Q + 1$ ; перейти к 3.

6) Результат  $:=$  период начинается с  $Q$  и кончается  $P - 1$ .

**Задача 6.** а) Блок-схема приводится ниже, триада записывается в виде  $cdu$ , где  $u$  — число единиц,  $d$  — число десятков,  $c$  — число сотен (см. блок-схему 4).

б) Часть, окруженная штриховой линией, исключается следующим образом:



Вместо «тысячи» говорят «миллион».

**Задача 7\*).** а) Чтение цифр.: 5 888 890 слов.

б) Чтение чисел с  $T_u$  триадами единиц и  $T_d$  триадами тысяч содержит: 1000 раз  $T_u$ , 1000 раз  $T_d$ , 999 000 — «тысяча» и название числа «нуль».

$T_u$  или  $T_d$  содержит 900 раз «сто», 800 раз числа сотен и 10 раз названия десятков единиц.

Имена десятков единиц содержат 3 раза по 4 слова. 25 раз по 3 слова, 50 раз по 2 слова, 21 раз по 1 слову.

Всего 8 559 001 слов.

**Задача 8.** а) Да. Достаточно уточнить, каково наибольшее положительное число.

б) Удобно иметь почти столько же положительных чисел, как и отрицательных.

\*) Ответ дан для случая французского языка.



$B := 10$ . Все числа, имеющие 0, 1, 2, 3, 4 в самой левой позиции, будут рассматриваться как положительные.

$B := 5$ . Приходим к тому, чтобы взять в качестве наибольшего положительного числа число, все цифры которого — двойки, что менее удобно.

З а д а ч а 9. а)  $10^8$ . б)  $9 \cdot 10^7$ . в)  $9 \cdot 10^7 + 10^8$ .

д) Место 10 может быть не определено, если мантисса оканчивается 0 или если порядок начинается с 0. Будем выбирать мантиссу, не оканчивающуюся 0.

Если символ 10 может быть опущен в конце числа, находим  $81 \cdot 10^6 + 6 \cdot 81 \cdot 10^5 + 9 \cdot 10^6 = 1386 \cdot 10^5$ .

З а д а ч а 10. а)  $TB \frac{B}{B-1}$ .

б) Основание 2, так как  $\frac{40}{9} T_2 > 2T_2$ .

в)  $\frac{B}{n(B-1)}$ .

З а д а ч а 11. а)  $T_{\bar{1}} = (pq + 1) T$ ,  $c_{\bar{1}} = (pq + 1) c$ ,

$$T_q = (p + q) T, \quad c_q = \left[ pq + \frac{p(p+1)}{2} \right] c.$$

б) Нет; в частности, можно переставить  $p$  и  $q$ , не изменяя  $T_q$ . Значит, всегда будем брать  $p \leq q$ .

в) Речь идет тогда о среднем времени

$$T_1 = \frac{B}{B-1} T, \quad c_1 = c \frac{B}{B-1},$$

$$T_q \approx \frac{B}{B-1}, \quad c_q \approx \frac{pB}{B-1} c.$$

Очевидно, интереса не представляет.

З а д а ч а 12. а)  $\frac{3}{2} qT + p(q+1) T_0$ .

б)  $\frac{3}{2} pqT + \frac{p(p+1)}{2} (q+1) T_0$ .

З а д а ч а 13. а) Можно разрешить, например: 0, 1, 2 положительные числа, 9, 8 отрицательные числа (то, что не симметрично).

б) Можно разрешить, например: 0, 1, 2 положительные числа, 9, 8, 7 отрицательные числа.

З а д а ч а 14. а) Мы используем таблицу сравнений с 10 символами. В греческом обозначении необходимо было бы использовать 3 таблицы по 9 символов.

б) Используем таблицу сложения  $10 \times 10$ . В греческом обозначении было бы необходимо использовать 3 таблицы  $9 \times 9$ .

в) Используем таблицу  $10 \times 10$ . В греческом обозначении необходимо было бы использовать одну таблицу  $27 \times 27$ .

З а д а ч а 15. а) Каждую цифру частного можно определить по двум цифрам делимого: по цифре  $n$ -го порядка и непосредственно следующего за ним более высокого порядка.

Пример.  $137: 37 \rightarrow 8; 13 \rightarrow 6$ , откуда 68.

б) Остаток может быть 0, 1, 2. В каждом случае можно найти соответствующую цифру частного и соответствующий остаток, так как последние цифры произведений числа 3 на правые десять цифр натурального ряда все разные. Завершение деления позволяет выбрать из трех операций ту, которая верна.

Пример.

137: остаток 0 дает 11 десятков  $+3 \times 9$ ;

11 десятков дает 1 сотню  $+3 \times 7$  десятков; невозможно;

137: остаток 1 дает 13 десятков  $+3 \times 2$ ;

13 десятков дает 1 сотню  $+3 \times 1$  десятков; невозможно;

137: остаток 2 дает 12 десятков  $+3 \times 5$ ;

12 десятков дает  $4 \times 3$ .

Отсюда

$$137 = 3 \times 45 + 2.$$

Задача 16. а) Можно осуществить деление раздельно, указывая для каждого оператора порядок цифры, которая предшествует его триаде.

б) Каждый оператор совершает три деления, чтобы учесть остатки 0, 1, 2, даваемые предыдущей триадой.

с) Если имеется  $p$  полос и мы пренебрегаем временем и стоимостью сбора результатов.

Деление на 2: время делится на  $p$ , стоимость та же.

Деление на 3: время делится на  $p/3$ , стоимость умножается на 3.

Задача 17. а). Каждая цифра по основанию 8 дает 3 двоичные цифры, которые служат для записи:

$$0 \rightarrow 000, 1 \rightarrow 001 \dots, 7 \rightarrow 111;$$

$$3417 \rightarrow 0,11100001111.$$

б) Запишем каждую десятичную цифру в двоично-пятеричной системе:

$$0 \rightarrow 00, 1 \rightarrow 01, 4 \rightarrow 04, 5 \rightarrow 10, 9 \rightarrow 14.$$

с) 03120204.

## 1. ОСНОВНЫЕ ПОНЯТИЯ

В главе I мы могли убедиться в важности понятия цепочки и поразмыслить над некоторыми вопросами работы с числами.

Эта глава познакомит нас с новым понятием — понятием синтаксиса — и со способами записи алгебраических выражений.

**1.1. Цепочки и выражения.** Алгебраическое (или какое-либо другое) выражение часто (но не всегда) представляет собой цепочку. Например:

$$ax + b$$

есть цепочка,

$$ax^2 + bx + c$$

становится цепочкой, если рассматривать  $^2$  как символ, отличный от цифры 2.

**У п р а ж н е н и е 1.** Можно ли рассматривать как цепочки следующие выражения?

$$\text{a) } \sin^2 x_1; \quad \text{b) } \int_a^b f(x) dx; \quad \text{c) } \frac{a+b}{c+d}; \quad \text{d) } \sqrt{a^2+b^2}.$$

**2.1. Алфавит.** Мы построим цепочки, состоящие из фиксированного числа символов. Эти символы будут составлять наш алфавит.

Например, целое число без знака может рассматриваться как цепочка, составленная из символов алфавита

$$0, 1, 2, \dots, 9.$$

Последовательность чисел

$$347 \ 519 \ 428 \ 1964$$

может рассматриваться как цепочка, имеющая в качестве алфавита множество чисел.

**2.2. Язык.** Вообще говоря, в языке используется только часть цепочек, которые могут быть составлены из символов заданного алфавита. Например, французский язык образуют те цепочки, составленные из букв латинского алфавита, которые являются словами французского языка.

**2.3. Синтаксис.** Существуют различные способы определения языка. Если язык содержит лишь конечное сравнительно небольшое количество цепочек, то их можно просто перечислить.

Удобнее язык определять при помощи правил, позволяющих:

— либо строить цепочки языка;

— либо распознавать, принадлежит цепочка языку или нет.

Набор таких правил образует *синтаксис* языка.

**З а м е ч а н и е.** Один и тот же язык может быть задан различными наборами правил.

Цепочка, удовлетворяющая правилам заданного синтаксиса  $S$ , будет называться *синтаксически корректной* (относительно  $S$ ). Это эквивалентно утверждению, что она принадлежит языку, определяемому  $S$ .

**2.4. Семантика.** Язык есть средство для выражения идей, фактов и различных свойств. Но одной синтаксической корректности для этого недостаточно. Цепочка языка, которая имеет (в системе мышления) какое-то значение, называется *семантически значимой*.

Итак, цепочка может быть:

— синтаксически некорректной;

— синтаксически корректной, но семантически не значимой;

— синтаксически корректной и семантически значимой.

**2.5. Примеры.** Рассмотрим русский язык \*), состоящий из множества слов и синтаксических правил грамматики.

Цепочка:

Дом находятся вчера

синтаксически некорректна (существительное в единственном числе, а глагол во множественном).

Фраза:

Дома находятся вчера

синтаксически корректна, но она семантически не значима.

---

\*) В оригинале французский язык. (*Прим. перев.*)

Напротив, фраза:

Дома находятся здесь

и синтаксически корректна, и семантически значима; при этом говорящий может лгать, но это уже другая проблема.

Возьмем теперь язык математических формул, например язык равенств алгебраических выражений с единственным правилом: равенство алгебраических выражений обязательно имеет вид

алгебраическое выражение = алгебраическое выражение

(при этом предполагается, что понятие алгебраического выражения было ранее определено).

Цепочка

$$a + ) = b$$

не является синтаксически корректной, поскольку  $a + )$  не есть алгебраическое выражение.

Напротив, равенство

$$a + bc = (a + b) (a + c)$$

синтаксически корректно. Оно не имеет смысла с точки зрения обычной алгебры, но является семантически значимым для специалистов, работающих с булевыми тождествами.

## II. ВЫРАЖЕНИЯ БЕЗ СКОБОК

**3.1. Ограничение объекта исследования.** Мы ограничимся изучением синтаксиса языков алгебраических выражений с двумя операциями, обозначаемыми  $+$  и  $\cdot$ . При этом мы увидим, что синтаксис даже таких языков в действительности очень сложен.

**3.2. Алфавит.** Мы будем использовать алфавит, содержащий строчные латинские буквы  $a, b, c, \dots, z$ , символы операций  $+$  и  $\cdot$  (последний может быть опущен) и скобки  $( )$ .

**4.1. Использование метаязыка.** Описать синтаксис языка при помощи самого языка очень трудно (представьте себе обучение китайскому языку при помощи единственного источника — китайской грамматики на китайском языке). Для этой цели мы воспользуемся метаязыком.

**4.2. Метаязык Бэкуса.** Мы будем пользоваться метаязыком Бэкуса и вначале разъясним его на примере (речь пойдет об определении записи целых чисел в десятичной системе счисления):

$\langle \text{цифра} \rangle ::= 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9$

$\langle \text{не ноль} \rangle ::= 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9$   $\langle \text{не ноль} \rangle \langle \text{цифра} \rangle$

$\langle \text{целое} \rangle ::= 0 | \langle \text{не ноль} \rangle$

Алфавит метаязыка состоит из символов:

$::=$  это символ, который означает, что то, что написано слева, по определению есть то, что написано справа;

| это разделитель, который помещается между различными вариантами одного и того же определения \*);

$\langle \text{цифра} \rangle$ ,  $\langle \text{не ноль} \rangle$  — промежуточные понятия,  $\langle \text{целое} \rangle$  — понятие, которое мы определяем;

$\langle \text{целое} \rangle$ ,  $\langle \text{не ноль} \rangle$  — цепочки;

$\langle \text{не ноль} \rangle \langle \text{цифра} \rangle$  означает, что в конце цепочки  $\langle \text{не ноль} \rangle$  помещается цифра.

У п р а ж н е н и е 2. а) Проверить, соответствует ли предыдущее определение наивному определению целого числа.

б) Что получится, если положить

$\langle \text{целое} \rangle ::= 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | \langle \text{целое} \rangle \langle \text{цифра} \rangle$  ?

с) Что получится, если определить целое как

$\langle \text{целое} \rangle ::= \langle \text{цифра} \rangle | \langle \text{целое} \rangle \langle \text{цифра} \rangle$  ?

**5.1. Язык выражений без скобок.** Рассмотрим следующий набор правил, записанных на метаязыке Бэкуса:

$\langle \text{буква} \rangle ::= a | b | \dots | z$

$\langle \text{знак} \rangle ::= + | \cdot$

$\langle \text{выражение} \rangle ::= \langle \text{буква} \rangle | \langle \text{выражение} \rangle \langle \text{знак} \rangle \langle \text{буква} \rangle$

Такому определению понятия  $\langle \text{выражение} \rangle$  удовлетворяет, например, цепочка

$$f + g \cdot a + x.$$

Этот набор правил называется *синтаксисом выражений без скобок*, а язык, который он определяет, называется *языком выражений без скобок*.

Относительно этого синтаксиса (а также относительно тех, которые мы будем изучать в дальнейшем) можно поставить следующие вопросы (на некоторые из них мы уже сейчас можем дать ответ):

а) построить все синтаксически корректные выражения;

б) выяснить, является ли заданное выражение синтаксически корректным;

\*) В метаязыке вместо слов «по определению» или «называется» используется символ  $::=$ , а вместо слова «или» используется вертикальная черта. (Прим. ред.)

с) разбить синтаксически корректное выражение на подвыражения так, чтобы суметь восстановить процесс построения всего выражения (этот вопрос ставится в связи с проблемой однозначности вычисления выражения).

**5.2. Необходимое и достаточное условие синтаксической корректности.** Ясно, что необходимым и достаточным условием синтаксической корректности выражения в языке выражений без скобок является следующее условие:

Выражение есть последовательность чередующихся букв и знаков. При этом первым и последним символами являются буквы.

**У п р а ж н е н и е 3.** Показать, что если вместо старого определения выражения использовать определение  $\langle \text{выражение} \rangle ::= \langle \text{буква} \rangle | \langle \text{выражение} \rangle \langle \text{знак} \rangle \langle \text{выражение} \rangle$ , то множество синтаксически корректных выражений не станет шире.

**6.1. Семантика слева направо.** С синтаксически корректными выражениями, приведенными выше, можно связать различные семантики \*). Рассмотрим одну из них (очень простую).

Выражение читается слева направо и каждый встречающийся знак означает операцию, первый операнд которой есть уже вычисленная часть выражения, а второй — буква, непосредственно следующая за этим знаком.

**П р и м е р.** Выражение  $a + b \cdot c + d$  интерпретируется следующим образом:

$$\alpha ::= a + b, \quad \beta ::= \alpha \cdot c, \quad \gamma ::= \beta + d;$$

это семантика слева направо.

\*) Например, выражение  $a + bi$  можно трактовать как форму записи комплексного числа (в этом случае знак  $+$  не воспринимается как знак операции), а можно трактовать как сумму действительного и мнимого чисел. Кроме того, выражение, вообще говоря, определяет процесс вычисления значения некоторой величины. Но как производить вычисления, в каком порядке выполнять операции в выражении, мы должны заранее условиться (т. е. оговорить семантику выражений). Например, выражение  $a \cdot b + c \cdot d$  можно вычислять двумя способами:

$$\begin{array}{ll} \text{I } \alpha := a \cdot b & \text{II } \alpha := a \cdot b \\ \beta := \alpha + c & \beta := c \cdot d \\ \gamma := \beta \cdot d & \gamma := \alpha + \beta \end{array}$$

Заметим, что во втором способе мы прежде всего выполняем операции умножения, т. е. эта операция встает в привилегированное положение, или, как говорят в таких случаях, получает приоритет. (Прим. ред.)

Эта семантика привлекательна своей простотой. Но, к сожалению, она слишком бедна. Например, она не позволяет получить значение выражения, которое записывается в обычной записи как

$$ab. + cd,$$

поскольку последняя буква должна быть множителем или слагаемым \*).

**У п р а ж н е н и е 4.** Показать, что выполнять выражения согласно правилам семантики слева направо может некоторая машина, которая имеет лишь одну ячейку внутренней памяти, один вход и одно устройство вычисления (сумматор), причем во внутреннюю ячейку памяти помещается первый операнд операции, а на вход подается второй операнд этой операции (результат помещается в ячейку внутренней памяти).

**У п р а ж н е н и е 5.** Как нужно изменить семантику слева направо, чтобы можно было вычислять выражения вида  $ab + cd$ ?

**7.1. Семантика с приоритетом умножения.** Обычно для выражений, определенных в 5.1, используется следующая семантика:

а) Сначала выражение разбивается на некоторое число подвыражений, в которых нет знаков  $+$ . Подвыражения вычисляются слева направо.

б) Подвыражения соединяются между собой знаками  $+$ . Сложение результатов вычисления подвыражений осуществляется слева направо.

В результате операция умножения получает приоритет перед операцией сложения.

Такая семантика называется *семантикой с приоритетом умножения*.

**П р и м е р.** Выражение  $a \cdot b + c \cdot d + e$  интерпретируется как

$$\alpha ::= a \cdot b, \quad \beta ::= c \cdot d, \quad \gamma ::= \alpha + \beta, \quad \delta ::= \gamma + e.$$

**У п р а ж н е н и е 6.** Показать, что в семантике с приоритетом умножения будет получен один и тот же результат, независимо от того, будем ли мы вычислять выражение справа налево или слева направо.

\*) Здесь автор не совсем точен. Мы можем вычислять это выражение согласно правилам семантики слева направо (это способ I в предыдущей сноске), но такой процесс вычисления будет отличаться от общепринятого (это способ II там же), а следовательно, и результаты вычисления будут разными. (Прим. ред.)

**7.2. Замечание о классической алгебраической записи.** Если взять «школьные» правила вычисления алгебраических выражений, то мы можем констатировать следующее!

Выражение, содержащее лишь знаки  $+$  и  $-$ , вычисляется слева направо.

**Пример.** а) Выражение  $a - b - c$  интерпретируется как

$$\alpha ::= a - b \quad \beta ::= \alpha - c.$$

б) Логично было бы использовать для вычисления выражений, содержащих операции умножения и деления, семантику слева направо.

На практике выражение  $a \cdot b \cdot c$  интерпретируется как

$$\alpha ::= a \cdot b, \quad \beta ::= \alpha \cdot c,$$

а выражение  $a \cdot b/c$  — как

$$\alpha ::= a \cdot b, \quad \beta ::= \alpha/c.$$

Следовало бы пойти дальше в этом направлении и интерпретировать  $a/b \cdot c$  как

$$\alpha ::= a/b, \quad \beta ::= \alpha \cdot c.$$

На самом же деле многие читают эту запись как

$$\gamma ::= b \cdot c, \quad \beta ::= a/\gamma.$$

Следовало бы также интерпретировать  $a/b/c$  как

$$\alpha ::= a/b, \quad \beta ::= \alpha/c.$$

На деле же многие рассматривают это выражение как неопределенное и отказываются его использовать.

**Упражнение 7.** а) Сравнить результаты вычислений выражения

$$a - b - c - d$$

при его вычислении слева направо и справа налево.

б) Сравнить ответ п. а) с результатом упражнения 6.

**7.3. Вычисление выражения с приоритетом умножения.** При вычислении выражения с приоритетом умножения сначала следует выполнить все умножения и запомнить полученные произведения. Таким образом, нам придется дважды просмотреть выражение.

Однако можно вычислять выражение, просматривая последовательно по 5 символов выражения. Если первые три символа имеют вид  $a \cdot b$ , то осуществляем умножение  $\alpha ::= a \cdot b$  и заменяем первые три символа на  $\alpha$ . Если

первые четыре символа имеют вид

$$a + b + \text{ или } a + b \text{ fin } *)_j$$

то осуществляем  $\beta ::= a + b$  и заменяем первые три символа на  $\beta$ . В случае, когда первые 5 символов имеют вид  $a + b \cdot c$ , выполняем  $\gamma ::= b \cdot c$  и заменяем эти 5 символов на  $a + \gamma$ .

**З а д а ч а 1.** а) Описать синтаксис записи целых чисел римскими цифрами.

б) Построить алгоритм, позволяющий распознавать, является ли запись числа римскими цифрами синтаксически корректной.

**З а д а ч а 2.** а) Описать синтаксис выражений без скобок, при условии, что знак  $\cdot$  опущен.

б) Найти необходимое и достаточное условие синтаксической корректности выражений без скобок.

### III. СИНТАКСИС ВЫРАЖЕНИЙ СО СКОБКАМИ

**8.1. Скобки.** Обычно при записи алгебраических выражений используются скобки. Мы будем предполагать, что читатель хорошо знаком с правилами использования скобок при записи алгебраических выражений. Однако напомним,

— что скобки всегда появляются парами (открывающая скобка предшествует закрывающей);

— что две пары скобок никогда не перекрещиваются, т. е. допускаются только конфигурации  $[( )]$  и  $( ) [ ]$ . Конфигурация  $[( ) [ ]$  не допускается.

Использование нескольких типов скобок не вносит ничего нового. В дальнейшем мы будем пользоваться только одним типом скобок.

**8.2. Язык скобок.** Извлечем из выражения скобки, которые оно содержит, абстрагируясь от всего остального. Опишем синтаксис полученного при этом языка:

$\langle \text{последовательность скобок} \rangle ::= | (\langle \text{последовательность скобок} \rangle) | \langle \text{последовательность скобок} \rangle$   
 $\langle \text{последовательность скобок} \rangle$ .

Заметим, что мы рассматриваем пустое множество как последовательность скобок.

\*) fin означает конец выражения. (Прим. ред.)

Заметим также, что при этом определении имеет место симметрия справа и слева.

**9.1. Необходимое и достаточное условие синтаксической корректности последовательности скобок.** Разрежем цепочку из скобок на две части, каждая из которых не пуста. Назовем эти части

- левой частью;
- правой частью.

Следующие условия необходимы для того, чтобы цепочка скобок была синтаксически правильной последовательностью скобок.

а) Цепочка содержит одинаковое число открывающих и закрывающих скобок. Справедливость этого утверждения следует непосредственно из определения.

б) Каково бы ни было сечение, левая часть, содержащая  $p$  открывающих скобок, содержит не более  $p$  закрывающих скобок.

Покажем, как можно доказать утверждение б) индукцией по числу скобок. Допустим, что оно выполняется для  $\langle$ последовательность скобок $\rangle_1$  и  $\langle$ последовательность скобок $\rangle_2$ . Оно, очевидно, верно и для любого из сечений  $\alpha, \beta, \gamma, \delta, \epsilon, \varphi$  следующих выражений:

$$\left( \left\langle \text{последовательность} \right|_{\alpha} \left. \text{скобок} \right\rangle_{\beta} \right)_{\gamma} \left\langle \text{последовательность} \right|_{\delta} \left. \text{скобок}_1 \right\rangle_{\epsilon} \left\langle \text{последовательность} \right|_{\varphi} \left. \text{скобок}_2 \right\rangle$$

**У п р а ж н е н и е 8.** Будут ли синтаксически корректны следующие выражения:

- а) (( )))(( )))((( ));
- б) (( ))(( )))

**9.2. Теорема существования.** Условия а) и б) из п. 9.1 являются достаточными для синтаксической корректности последовательности скобок.

Покажем это индукцией по числу  $n$  элементов цепочки. Для  $n := 0$  это утверждение верно.

Рассмотрим сечение (в предположении, что такое существует), которое оставляет слева столько же открывающих скобок, сколько и закрывающих. Обе части  $A$  и  $B$  удовлетворяют условиям а) и б) из п. 9.1. Значит, они синтаксически корректны. Вся цепочка тоже корректна, поскольку ее можно описать так:

$\langle$ последовательность скобок $\rangle$   $\langle$ последовательность скобок $\rangle$ .

Допустим теперь, что такого сечения не существует. Исключим самую левую скобку (которая обязательно открывающая) и самую правую (которая обязательно закрывающая). Оставшаяся часть снова удовлетворяет условиям а) и б) из п. 9.1 и, значит, она синтаксически корректна.

То же самое имеет место для исходной последовательности скобок, поскольку она записывается

(⟨последовательность скобок⟩).

**10.1. Пары скобок.** В определении последовательности скобок последние появляются парами. Если две скобки образуют пару

$$\left| \begin{array}{c} ( \\ \alpha \end{array} \right| \begin{array}{c} | \\ \gamma \end{array} \left| \begin{array}{c} ) \\ \beta \end{array} \right|$$

то часть выражения, заключенная между  $\alpha$  и  $\gamma$ , содержит больше открывающих скобок, чем закрывающих; часть между  $\gamma$  и  $\beta$  содержит больше закрывающих скобок, чем открывающих; часть между  $\alpha$  и  $\beta$  содержит столько же открывающих скобок, сколько и закрывающих.

Отсюда следует, что в синтаксически корректном выражении закрывающая (открывающая) скобка, образующая пару с открывающей (закрывающей) скобкой, определяется единственным образом при помощи следующего условия:

$$\left| \begin{array}{c} ( \\ \alpha \end{array} \right| \left( \quad \right) \left| \begin{array}{c} ) \\ \beta \end{array} \right|$$

Искомая закрывающая (открывающая) скобка такова, что часть цепочки, которую ограничивают заданная открывающая (закрывающая) скобка и искомая закрывающая (открывающая) скобка, имеет столько же открывающих скобок, сколько и закрывающих.

**У п р а ж н е н и е 9.** Показать, что две пары скобок не могут перекрываться:

$$( [ ] ).$$

**10.2. Использование стека для восстановления пар скобок.** Мы будем исходить из двух следующих замечаний:

а) Рассмотрим первую закрывающую скобку. Она образует пару с открывающей скобкой, которая ей предшествует.

б) Исключая парные скобки, мы не нарушаем ни синтаксической корректности выражения, ни парности других скобок.

Воспользуемся теперь для нахождения парных скобок следующим алгоритмом:

— Помещаем в стек открывающие скобки в том порядке, в каком они встречаются в цепочке;

— когда появляется закрывающая скобка, то в качестве парной мы берем скобку, находящуюся в вершине стека, и удаляем ее из стека.

П р и м е р.  $(1(2)3(4(5)6)7)8(9)_{10}$

Последовательные состояния стека и пар:

(1		
(1 (2		
(1	(2	)3
(1 (4		
(1 (4 (5		
(1 (4	(5	)6
(1	(4	)7
	(1	)8
(9		
	(9	)10

У п р а ж н е н и е 10. Можно ли использовать стек для распознавания синтаксической корректности последовательности скобок из п. 8.2?

**10.3. Теорема о разбиении.** Синтаксически корректная последовательность скобок может быть записана в виде

$$\underbrace{\langle \text{последовательность скобок} \rangle}_1 \underbrace{\langle \text{последовательность скобок} \rangle}_2 \underbrace{\langle \text{последовательность скобок} \rangle}_3 \dots \underbrace{\langle \text{последовательность скобок} \rangle}_4$$

одним и только одним способом (согласно замечанию из п. 8.2  $\langle \text{последовательность скобок} \rangle$  может быть пустым множеством).

Это свойство сразу следует из доказанного выше. В самом деле:

— скобка 2 образует пару со скобкой 1;

— за скобкой 2 следует открывающая скобка 3 (в противном случае головная часть выражения, заканчивающаяся скобкой 3, содержала бы меньше открывающих скобок, чем закрывающих);

— скобка 4 образует пару со скобкой 3 и т. д.

#### IV. СКОБОЧНЫЕ ВЫРАЖЕНИЯ

**11.1. Скобочные выражения.** Теперь мы определим язык скобочных выражений.

Таких языков существует несколько. Тот, который мы определим вначале, содержат все остальные.

##### 11.2. Общее скобочное выражение (ОСВ).

$\langle \text{буква} \rangle ::= a | b | \dots | z$

$\langle \text{знак} \rangle ::= + | \cdot$

Алфавит содержит также скобки (|)

$\langle \text{ОСВ} \rangle ::= \langle \text{буква} \rangle | (\langle \text{ОСВ} \rangle) | \langle \text{ОСВ} \rangle \langle \text{знак} \rangle \langle \text{ОСВ} \rangle$ .

Заметим, что здесь для скобок соблюдаются правила п. 8.2. Можно заметить также, что часть этого языка составляют выражения, определенные в п. 5.1. Выражения, принадлежащие этому языку, называются *общими скобочными выражениями*.

**У п р а ж н е н и е 11.** Являются ли следующие выражения синтаксически корректными общими скобочными выражениями:

$(a), ((a + (b))), a + (bc), ((d + e)?)$

**11.3. Необходимые условия синтаксической корректности.** Очевидно, что следующие условия необходимы:

- для букв и знаков условия из п. 5.2;
- для скобок — условия а) и б) из п. 9.1;
- не должны встречаться следующие выражения:

$(\langle \text{знак} \rangle ( ) \langle \text{буква} \rangle ($   
 $) \langle \text{буква} \rangle) (\langle \text{знак} \rangle).$

Справедливость этих условий доказывается индукцией по числу символов в выражении. Доказательство мы предоставляем читателю.

**11.4. Достаточный характер этих условий.** Эти условия достаточны для того, чтобы выражение было синтаксически корректным ОСВ.

Мы докажем это индукцией по числу символов в выражении.

Если выражение начинается

$\langle \text{буква} \rangle \langle \text{знак} \rangle,$

то оставшаяся часть снова удовлетворяет условиям п. 11.3 и значит, синтаксически корректна. Но выражение

$\langle \text{буква} \rangle \langle \text{знак} \rangle \langle \text{ОСВ} \rangle$

в свою очередь корректно по определению.

Если выражение начинается с (, то мы выделим ей парную скобку. Тогда выражение запишется в виде

$$(A) B,$$

где  $A$  удовлетворяет условиям из п. 11.3 и, значит, синтаксически корректно. Что касается  $B$ , то оно может быть пусто. В этом случае по предположению индукции \*) выражение синтаксически корректно. Если же  $B$  не пусто, то оно имеет вид

$$\langle \text{знак} \rangle C,$$

причем  $C$  синтаксически корректно. Но выражение  $(A) \langle \text{знак} \rangle C$  синтаксически корректно по определению.

**12.1. Семантика скобочных выражений.** Мы будем связывать со скобочными выражениями следующую семантику:

— если содержимое одной пары скобок не содержит других скобок, то его вычисляют, следуя семантике п. 6.1 или п. 7.1;

— часть выражения, ограниченная парой скобок и не содержащая внутри себя скобок, заменяется одной буквой;

— процесс повторяется;

— на последнем этапе может получиться выражение без скобок, которое вычисляется по одной из семантик: п. 6.1 или п. 7.1.

**У п р а ж н е н и е 12.** Показать, как можно использовать стек для вычисления скобочных выражений, следуя семантике слева направо для бесскобочных выражений.

**У п р а ж н е н и е 13.** То же, что в упражнении 12, но для семантики с приоритетом умножения.

**13.1. Расширенное скобочное выражение (РСВ).** Введем определение

$$\langle \text{РСВ} \rangle ::= \langle \text{буква} \rangle \mid \langle \langle \text{РСВ} \rangle \rangle \mid \langle \langle \text{РСВ} \rangle \langle \text{знак} \rangle \langle \text{РСВ} \rangle \rangle.$$

Тем самым, очевидно, определен некоторый подъязык общего скобочного языка. Выражения, принадлежащие этому языку, называются *расширенными скобочными выражениями*.

Заметим, что в этом языке всякое выражение, не сводящееся к одной букве при помощи алгоритма п. 12.1, заключено в пары скобок.

---

\*) Автор, по-видимому, забыл его указать. (Прим. ред.)

Из этого сразу следует, что расширенное скобочное выражение может быть записано следующим, и притом единственным, способом:  $\langle \langle \text{PCB} \rangle \langle \text{знак} \rangle \langle \text{PCB} \rangle \rangle$ .

Идея доказательства такова: если бы для одного и того же выражения существовали две записи  $(A * B), (C * D)$ , где  $A, B, C, D$  — расширенные скобочные выражения\*), то открывающая скобка, помещенная в начале  $A$  (и  $C$ ), должна была бы образовывать пару вместе с закрывающей скобкой, заканчивающей  $A$  (и с той, которая заканчивает  $C$ ), откуда получаем противоречие.

Индукцией по числу символов в выражении доказывается, что в семантической оценке такого выражения встречаются только выражения без скобок, не сводящиеся к одной букве, в форме  $\langle \text{буква} \rangle \langle \text{знак} \rangle \langle \text{буква} \rangle$ .

В этом очень простом частном случае семантики п. 6.1 и п. 7.1 совпадают.

**У п р а ж н е н и е 14.** Являются ли следующие выражения расширенными скобочными выражениями:

$$((a)), (a + b + c), (((a + b)) + c)?$$

**У п р а ж н е н и е 15.** Показать, что в расширенном скобочном выражении имеется по крайней мере столько же пар скобок, сколько и знаков.

**13.2. Строгое скобочное выражение (ССВ).** Положим

$$\langle \text{ССВ} \rangle ::= \langle \text{буква} \rangle | \langle \langle \text{ССВ} \rangle \langle \text{знак} \rangle \langle \text{ССВ} \rangle \rangle.$$

Определенный таким образом язык есть подъязык расширенного скобочного языка.

Выражения этого языка называются *строгими скобочными выражениями*.

Индукцией по числу символов в выражении доказывается, что число пар скобок равно числу знаков. Отсюда вытекает, что если добавить или убрать пару скобок в выражении, то оно перестанет принадлежать языку.

**У п р а ж н е н и е 16.** Являются ли следующие выражения строгими скобочными выражениями:

$$(a), (a + b), ((a + b)), (a + b + c), (a + (b + c))?$$

**13.3. Минимальное скобочное выражение (МСВ).** Рассмотрим следующий синтаксис:

$$\langle \text{буква} \rangle ::= a | b | \dots | z$$

\*) При этом предполагается, что  $A$  и  $C, B$  и  $D$  различны. (Прим. ред.)

$\langle \text{сумма} \rangle ::= \langle \text{произведение} \rangle + \langle \text{произведение} \rangle | \langle \text{сумма} \rangle + \langle \text{произведение} \rangle$

$\langle \text{произведение} \rangle ::= \langle \text{буква} \rangle | \langle \text{произведение} \rangle \cdot \langle \text{буква} \rangle | (\langle \text{сумма} \rangle) \cdot \langle \text{буква} \rangle | \langle \text{произведение} \rangle \cdot (\langle \text{сумма} \rangle) | (\langle \text{сумма} \rangle) \cdot (\langle \text{сумма} \rangle)$

$\langle \text{ОСВ} \rangle ::= \langle \text{сумма} \rangle | \langle \text{произведение} \rangle$ .

Легко видеть, что этот язык есть подязык общего скобочного языка. Выражения этого языка называются *минимальными скобочными выражениями*.

Очевидно, что содержимое скобок, которое не содержит никакой другой скобки, есть сумма произведений букв. Этот синтаксис хорошо соответствует использованию семантики с приоритетом умножения.

**У п р а ж н е н и е 17.** Являются ли следующие выражения минимальными скобочными выражениями:

$a + bc, (a + bc), (a + b)c, a + (bc), a + (b + c) \cdot (d + c)?$

**З а д а ч а 3.** Рассмотреть общие скобочные выражения и ответить на вопрос, являются ли они расширенными скобочными выражениями при следующих условиях.

а) Является ли следующее условие необходимым и достаточным для того, чтобы ОСВ было расширенным скобочным выражением:

— последовательности

$\langle \text{знак} \rangle \langle \text{буква} \rangle \langle \text{знак} \rangle$

начало  $\langle \text{буква} \rangle \langle \text{знак} \rangle \langle \text{буква} \rangle \langle \text{знак} \rangle$  конец \*) в выражении не встречаются.

б) Применить стек как в упражнении 12 со следующим способом сворачивания выражения, находящегося в стеке:

а) когда верхние три символа стека имеют вид  $(\alpha)$ , их исключают и помещают  $\alpha$  в вершину стека;

б) когда верхние пять символов стека имеют вид  $(\alpha * \beta)$ , выполняется операция  $\gamma ::= \alpha * \beta$ , исключаются пять символов и  $\gamma$  помещается в вершину стека.

Является ли наличие в стеке единственного элемента по окончании алгоритма необходимым и достаточным условием того, чтобы выражение, находящееся в стеке, было ОСВ?

с) *Индексом* знака назовем разность между числом открывающих и закрывающих скобок, предшествующих этому знаку. Будут ли следующие условия необходимыми и достаточными, чтобы выражение было ОСВ:

---

\*) Начало — начало последовательности; конец — конец последовательности. (Прим. ред.)

- α) никакой знак не имеет нулевого индекса;
- β) между двумя знаками одного и того же индекса имеется по крайней мере один знак меньшего индекса?

**Задача 4.** Рассмотреть скобочные выражения и постараться определить, являются ли они строгими скобочными выражениями.

а) Будет ли равенство числа пар скобок и знаков достаточным условием для этого?

б) Станет ли оно таковым, если добавить к этому, что выражение ОСВ?

с) Построить стек, при помощи которого можно было бы проверить, является ли данное выражение скобочным выражением.

**Задача 5.** а) Определить в форме Бэкуса язык последовательностей круглых и квадратных скобок (вспомните замечание из п. 8.1).

б) Что можно сказать о круглых скобках, содержащихся в паре квадратных скобок? Можно ли, исходя из этой идеи, сформулировать необходимое и достаточное условие синтаксической корректности?

с) Обозначим через  $i_a$  и  $j_a$  число круглых и число квадратных скобок, расположенных слева от  $a$ -го элемента цепочки.

Начертим в плоскости  $i, j$  путь, составленный из отрезков, параллельных осям и соединяющих последовательно точки  $(i_a, j_a)$ .

Показать, что этот путь сводится к нулю последовательным выбрасыванием отрезков, пробегаемых в одном направлении и сразу же затем в обратном направлении.

**Задача 6.** Назовем *высотой* части  $p$  выражения, не содержащей ни одной скобки, следующее число:

$h(p)$  = число открывающих скобок, лежащих слева от  $p$ , минус число закрывающих скобок, лежащих слева от  $p$ .

а) Рассмотрим разбиение выражения на связные части, все элементы которых имеют одинаковую высоту.

Показать, что необходимым и достаточным условием для того чтобы при вычислении такого куска формулы можно было бы отбрасывать скобки, является следующее условие:

его высота есть относительный максимум (т. е. только по левой части).

б) Как выражается посредством высот метод вычисления скобочных выражений при помощи стека из упражнения 12?

с) Удобно ли начинать вычисление частей с высотой, являющейся абсолютным максимумом?

**З а д а ч а 7.** Пусть требуется проверить синтаксическую корректность формул вида

*если A то B иначе C,*

где  $A, B, C$  пусты или являются скобочными выражениями.

а) Описать синтаксис такого языка в терминах языка Бэкуса.

б) Сформулировать необходимые и достаточные условия синтаксической корректности выражения в этом языке.

**З а д а ч а 8.** а) Воспользуемся общими скобочными выражениями с двумя операциями, в которых знак второй операции только подразумевается.

Показать, что его можно однозначно восстановить.

б) Воспользуемся общими скобочными выражениями, в которых знаки операций присутствуют, но открывающие и закрывающие скобки имеют одинаковый вид.

Показать, что открывающие и соответствующие им закрывающие скобки можно однозначно определить.

с) Показать, что если подразумевать знак  $\cdot$  и если представить тем же способом открывающие и закрывающие скобки, то получится неоднозначность.

**З а д а ч а 9.** В алгебраическом выражении занумеруем знаки операций в том порядке, в котором они будут выполняться, причем каждый знак операции приписывается множителям или частичным результатам, расположенным непосредственно слева или справа от знака. После выполнения операции исключаем сомножители и знак и ставим на их место результат операции.

**П р и м е р**

$$\begin{array}{ll}
 a \underset{2}{+} b \underset{4}{\cdot} c \underset{1}{+} d \underset{3}{\cdot} e & \alpha ::= c + d \\
 a \underset{2}{+} b \underset{4}{\cdot} \alpha \underset{3}{\cdot} e & \beta ::= a + b \\
 \beta \underset{4}{\cdot} \alpha \underset{3}{\cdot} e & \gamma ::= \alpha \cdot e \\
 \beta \underset{4}{\cdot} \gamma & \delta ::= \beta \cdot \gamma
 \end{array}$$

а) Показать, что от этой нотации можно перейти к нотации строгих скобочных выражений.

б) Всегда ли возможен обратный переход?

с) Имеет ли место единственность при просмотре слева направо? справа налево?

## У. ПРЕФИКСНАЯ И ПОСТФИКСНАЯ ПОТАЦИИ \*)

**14.1. Префиксная нотация.** В этой нотации выражение  $a + b$  записывается в форме  $+ ab$

Такая форма записи имеет различные преимущества:

— это есть функциональная нотация; в самом деле, сумма есть функция, переменными которой являются оба ее слагаемых;

— в этой форме можно записать сумму трех членов  $+ abc$  (при условии, что вначале уточняется, что она имеет три члена; впрочем, во избежание недоразумений можно писать  $+^3 abc$ ). Напротив, такая нотация как  $a + b + c$  не дает прямо сумму трех членов, а составляется из двух сумм по два члена.

Мы же будем использовать постфиксную нотацию  $ab +$

**14.2. Язык постфиксных выражений (ПВ).** Мы определим синтаксис языка постфиксных выражений посредством языка Бэкуса (мы ограничимся операциями с двумя членами):

$\langle \text{буква} \rangle ::= a \mid b \mid \dots \mid z$

$\langle \text{знак} \rangle ::= + \mid \cdot$

$\langle \text{ПВ} \rangle ::= \langle \text{буква} \rangle \mid \langle \text{ПВ} \rangle \langle \text{ПВ} \rangle \langle \text{знак} \rangle$

Например,

$$ab \cdot c +$$

есть корректное выражение в синтаксисе языка постфиксных выражений. Обычно оно записывается как

$$ab + c$$

**14.3. Связь со скобочным языком.** Мы сейчас докажем следующий результат: последовательности скобок, дополненные слева одной открывающей скобкой, взаимно однозначно соответствуют постфиксным выражениям посредством отображения

$$( \rightarrow \text{буква} \quad ) \rightarrow \text{знак}$$

В самом деле, переведем на язык скобок определение постфиксных выражений:

$$\langle X \rangle ::= ( \mid \langle X \rangle \langle X \rangle )$$

---

\*) Эти способы записи выражений были предложены польским математиком Лукасевичем. (Прим. ред.)

Из этого определения следует, что  $\langle X \rangle$  всегда начинается слева открывающей скобкой. Положим

$$\langle X \rangle ::= (\langle Y \rangle)$$

Определение принимает вид

$$\langle Y \rangle ::= | \langle Y \rangle (\langle Y \rangle)$$

Сравним его с определением из п. 8.2:

$\langle \text{последовательность скобок} \rangle ::= | (\langle \text{последовательность скобок} \rangle) | \langle \text{последовательность скобок} \rangle \langle \text{последовательность скобок} \rangle$

Ясно, что язык последовательностей скобок содержит язык выражений  $Y$ .

Докажем, что язык последовательностей скобок содержится в языке выражений  $Y$ . Мы будем проводить индукцию по числу символов и применим теорему о разбиении. Всякая последовательность скобок может быть записана в виде

$(\langle \text{последовательность скобок}_1 \rangle) (\langle \text{последовательность скобок}_2 \rangle) \dots (\langle \text{последовательность скобок}_p \rangle)$ .

Очевидно, что такие последовательности будут частью языка выражений  $\langle Y \rangle$ , как только каждая

$\langle \text{последовательность скобок}_i \rangle \quad i = 1, \dots, p,$   
составляет часть языка  $\langle Y \rangle$ .

**15.1. Теорема единственности.** Мы покажем, что в постфиксном выражении члены последней операции (знак которой есть последний символ цепочки) полностью определены. Это сводится к доказательству того, что последовательность скобок не может быть записана двумя различными способами:

$$Y_1 (Y_2) = Y_3 (Y_4).$$

Это свойство следует непосредственно из теоремы о разбиении из п. 10.3.

**У п р а ж н е н и е 18.** Являются ли следующие выражения постфиксными выражениями?

а)  $ab \cdot c \cdot \cdot$ , б)  $a \cdot b \cdot c \cdot \dot{d} \cdot$ , в)  $ab + cd \cdot +$  синтаксически корректными?

**У п р а ж н е н и е 19.** а) Имеются ли в постфиксном выражении операции, которые можно выполнять сразу?

б) Каково их максимальное число, если исходить из числа букв в выражении?

**З а д а ч а 10.** Осуществить непосредственное исследование постфиксных выражений.

а) Показать, что следующие условия необходимы для синтаксической корректности постфиксных выражений:

α) если выражение содержит  $n$  букв, то оно содержит  $n - 1$  знаков;

β) левая часть, содержащая  $p$  букв, содержит не более  $p - 1$  знаков.

б) Доказать, исходя из этих условий, теорему единственности.

с) Доказать, что эти условия достаточны для синтаксической корректности постфиксных выражений.

**З а д а ч а 11.** Построить стек для вычисления постфиксного выражения. Можно ли вместе с вычислением распознать и синтаксическую корректность?

**З а д а ч а 12.** а) Показать, что, заменяя

⟨буква⟩

на

⟨буква⟩ ⟨буква⟩ ⟨знак⟩,

мы не нарушаем синтаксическую корректность постфиксного выражения.

б) Вывести отсюда новый способ образования постфиксного выражения.

с) Показать, что процесс построения выражений согласен определению однозначен.

д) Можно ли это утверждать относительно процесса построения выражений, предлагаемого в б)?

**З а д а ч а 13.** Указать алгоритм, преобразующий постфиксное выражение в префиксное.

**З а д а ч а 14.** Сколько различных выражений можно составить из  $n$  букв, взятых по одному разу в алфавитном порядке и из двух операций:

а) в языке без скобок;

б) в постфиксном языке?

**З а д а ч а 15.** Рассмотрим относительно алгебры, имеющей операции:

• для двух членов;

\* для трех членов,

выражения в постфиксной нотации.

а) Описать синтаксис этих выражений в форме Бэкуса.

б) Указать необходимое и достаточное условие синтаксической корректности.

с) Можно ли сформулировать теорему единственности?

## РЕШЕНИЯ УПРАЖНЕНИЙ ГЛАВЫ II

- 1) а) Да. б) Нет.  
 с) Да, если записать  $(a + b)/(c + d)$ .  
 д) Да, если записать  $[a^2 + b^2]^{1/2}$ .  
 2) а) Очевидно.  
 б) Не хватает числа нуль.  
 с) Получим записи, которые могут иметь слева бесполезные нули.  
 3) Очевидно.

4) Машина комбинирует в своем логическом устройстве содержимое своей памяти и число, представленное на входе. Она помещает результат в свою память.

5)  $e ::= c \cdot d; a \cdot b + e$ .  
 6)  $(a + b) + c = a + (b + c)$  и то же самое для  $n$  членов. То же для умножения.

7) а)  $a - (b - (c - d)) = a - b + c - d$ .

б) Вычитание не коммутативно.

8) а) Нет: 2 открывающие, 3 закрывающие скобки.

б) Ничего нельзя сказать, поскольку задаваемые условия лишь необходимы (в действительности, как будет видно в п. 9.2, они и достаточны).

9) При построении выражения одна из пар была построена раньше другой.

10) Да. Выражение синтаксически корректно, если после просмотра всего выражения стек оказался пустой.

11) Да, для двух первых. Нет, для последних выражений.

12) Обозначим знаки через \*, не делая между ними различия. Поместим символы в стек.

Когда три верхних символа стека имеют вид  $a * b$ , то мы производим операцию, исключаем  $a * b$  из стека и помещаем результат операции в вершину стека.

Когда верхние три символа стека имеют вид  $(\alpha)$ , то мы исключаем их из стека и помещаем  $\alpha$  в вершину стека.

П р и м е р.

$$(((a + b)) + c + d) + (e + f).$$

Вот последовательные состояния стека:

(	$(\beta +$	
((	$(\beta + d$	$\gamma ::= \beta + d$
(((	$(\gamma$	
(((a	$(\gamma)$	
(((a + b	$\alpha ::= a + b$	$\gamma +$
(((\alpha	$\gamma + ($	
(((\alpha)	$\gamma + (e$	
((\alpha	$\gamma + (e +$	
((\alpha)	$\gamma + (e + f)$	$\varepsilon ::= e + f$
(\alpha	$\gamma + (\varepsilon)$	
(\alpha +	$\gamma + \varepsilon$	$\varphi ::= \gamma + \varepsilon$
(\alpha + c	$\beta ::= \alpha + c$	$\varphi$
(\beta		

13) Поместим символы в стек и произведем следующие сворачивания конечных последовательностей символов:

$a \cdot b$  заменяем 3 символа значением произведения

( $a$ ) заменяем на  $a$

$$\left. \begin{array}{l} a + b + \\ a + b) \\ a + b \text{ fin} \end{array} \right\} \text{заменяются на } \left\{ \begin{array}{l} a + \\ a) \\ a \end{array} \right.$$

при  $\alpha ::= a + b$ .

Пр и м е р.

$$(a + b \cdot c + x) + (d + e) \cdot f.$$

Вот последовательные состояния стека:

(		$\gamma +$	
( $a$		$\gamma + ($	
( $a +$		$\gamma + (d$	
( $a + b$		$\gamma + (d +$	
( $a + b \cdot$		$\gamma + (d + e$	
( $a + b \cdot c$	$\alpha ::= b \cdot c$	$\gamma + (d + e)$	$\delta ::= d + e$
( $a + \alpha$		$\gamma + (\delta)$	
( $a + \alpha +$	$\beta ::= a + \alpha$	$\gamma + \delta$	
( $\beta + x$		$\gamma + \delta \cdot f$	$\varepsilon ::= \delta \cdot f$
( $\beta + x$ )	$\gamma ::= \beta + x$	$\gamma + \varepsilon$	$\varphi ::= \gamma + \varepsilon$
( $\gamma$ )		$\varphi$	

14) Нет, для второго.

15) Каждый знак требует образования пары скобок.

16) Да, для второго и пятого.

17) Да, для первого, третьего и пятого.

18) а), б) Нет. в) Да.

19) а)  $ab \cdot cd \cdot +$

$ab \cdot$  и  $cd \cdot$  выполняются непосредственно (вообще, две буквы, сопровождаемые знаком, определяют операцию, выполняемую непосредственно).

б) Для  $2n$  и  $2n + 1$  букв максимум равен  $n$ .

## РЕШЕНИЯ ЗАДАЧ ГЛАВЫ II

Задача 1. а)

$\langle \text{цифра} \rangle ::= I|V|X|L|C|D|M$

Ясно, что синтаксис не фиксируется во всех деталях. Мы допустим, что обращения разрешаются только относительно главных символов (I, X, C), которые могут быть помещены перед одним из двух символов, стоящих выше него непосредственно по рангу (например C перед D или перед M):

$\langle \text{тысяча} \rangle ::= M|MM|MMM$

$\langle \text{сотня} \rangle ::= C|CC|CCC|CD|D|DC|DCC|DCCC|CM$

$\langle \text{десяток} \rangle ::= X|XX|XXX|XL|L|LX|LXX|LXXX|XC.$

$\langle \text{единица} \rangle ::= I|II|III|IV|V|VI|VII|VIII|IX$

⟨число ранга 2⟩ ::= ⟨десяток⟩⟨единица⟩|⟨десяток⟩ ⟨единица⟩  
⟨число ранга 3⟩ ::= ⟨сотня⟩|⟨число ранга 2⟩|⟨сотня⟩ ⟨число ранга 2⟩

⟨римское число⟩ ::= ⟨тысяча⟩|⟨число ранга 3⟩|⟨тысяча⟩ ⟨число ранга 3⟩

б) Читают слева направо, отождествляя по пути группы ⟨тысяча⟩, ⟨сотня⟩, ⟨десяток⟩, ⟨единица⟩, причем некоторые из этих групп могут отсутствовать.

З а д а ч а 2. а)

⟨буква⟩ ::=  $a|b| \dots |z$

⟨выражение⟩ ::= ⟨буква⟩|⟨выражение⟩ ⟨буква⟩|⟨выражение⟩ + ⟨буква⟩

б) Выражение начинается и заканчивается буквой.

Никогда нет двух рядом стоящих знаков +

З а д а ч а 3. а) Условие не является достаточным:

$$(a + b) + (c + d).$$

Легко видеть, что оно необходимо.

б) Индукцией показывается, что условие необходимо и достаточно.

с) Условие необходимо. Оно и достаточно (разрежем выражение на знаке с наименьшим индексом).

З а д а ч а 4. а) Нет. б) Да.

с) Взять снова стек из б) задачи 3, позволяя лишь операцию β).

З а д а ч а 5. а)

⟨буква⟩ ::=  $a|b| \dots |z$

⟨выражение⟩ ::= ( )| [ ]|⟨выражение⟩| [⟨выражение⟩]|

⟨выражение⟩ ⟨выражение⟩

Содержимое пары квадратных скобок есть корректное скобочное выражение; то же самое верно и для круглых скобок. Это условие необходимо и достаточно.

с) Можно свести выражение к пустому посредством последовательного вычеркивания [ ] или ( ).

З а д а ч а 6. а) Две скобки, которые ограничивают эту часть, имеют вид (и).

б) На каждом этапе исследуем самую левую часть, обладающую свойством а).

с) Нет, поскольку для нахождения такого максимума требовалось бы предварительно проанализировать все выражение.

З а д а ч а 7. Обозначим если через  $S$ , тогда — через  $A$ , иначе — через  $N$ .

а) ⟨выражение⟩ ::=  $S$  ⟨выражение⟩  $A$  ⟨выражение⟩  $N$  ⟨выражение⟩

б) Считаем  $S$  за 2, а  $A$  и  $N$  — за  $-1$ .

— Счет никогда не приводит к отрицательному результату и заканчивается 0;

— после  $S$ , для которого счет дает  $h$ , первая буква, для которой счет дает  $h - 1$ , не есть  $N$ ;

— после  $A$ , для которого счет дает  $h$ , первая буква, для которой счет дает  $h - 1$ , не есть  $A$ .

Эти условия необходимы и достаточны.

З а д а ч а 8. а) Между двумя последовательными буквами имеется знак операции. Он помещается после закрывающих скобок и перед открывающими скобками.

б) Скобки одинаковые. Они открывающие, если они предшествуют букве, и закрывающие, если они следуют за ней.

$$c) \quad a + b/c + d/e + f$$

может интерпретироваться как

$$(a + b) \cdot c + d \cdot (e + f) \text{ или } (a + b \cdot (c + d)) \cdot e + f$$

**Задача 9.** а) Помещаем открывающую скобку слева от левого сомножителя, а закрывающую скобку — справа от правого сомножителя каждой операции. Полученное таким образом выражение можно обобщенно считать строгим скобочным выражением.

б) Да, в силу единственности, доказанной в п. 13.1.

с) Единственность имеет место только при просмотре выражения слева направо. Выражение  $((a \cdot b) + (c \cdot d))$  может быть записано в виде

$$a \cdot b + c \cdot d \text{ или } a \cdot b + c \cdot d$$

1 3 2                      2 3 1

**Задача 10.** а) Доказывается индукцией по числу символов выражения.

б) Выражение не может быть правой частью выражения.

с) Рассекаем выражение посреди самой длинной левой части, которая также является выражением.

**Задача 11.** Читаем выражение слева направо и составляем стек из прочитанных элементов. Если три элемента вверху стека будут  $a b \langle \text{знак} \rangle$ , то осуществляем операцию, исключаем три элемента и помещаем результат операции в вершину стека.

**Пример.**  $abc + de \cdot + fg \cdot +$

Последовательные этапы стека:

$a$	$a\gamma$	$\delta := a\gamma$
$ab$	$\delta$	
$abc$	$\alpha := bc +$	$\delta f$
$ac$		$\delta fg$
$acd$		$\epsilon := fg \cdot$
$acde$	$\beta := de \cdot$	$\delta \epsilon$
		$\varphi := \delta \epsilon +$
$a\alpha\beta$	$\gamma := \alpha\beta +$	

Мы должны иметь в конце алгоритма выражение, полностью выметенное, и единственный элемент в стеке.

**Задача 12.** а) Проверить условия  $\alpha$ ) и  $\beta$ ) задачи 9.

б), с) Очевидно.

д) Можно получить несколько раз одно и то же выражение:

$$ab + a := cd \cdot cd \cdot b + b := ef \cdot cd \cdot ef \cdot +$$

или же

$$ab + b := ef \cdot aef \cdot + a := cd \cdot cd \cdot ef +$$

**Задача 13.** Читаем выражение справа налево. Составим новое выражение как цепочку справа налево.

Составим стек со знаками, наделяя их индексами. Когда читается знак, то он помещается в вершину стека с индексом 0.



Задача 14. а)  $2^{n-1}$ . Выбор относится только к знаку операции.

$$b) P_n = 2 \sum_p P_p P_{n-p};$$

$$P_1 = 1, P_2 = 2, P_3 = 8, P_4 = 40.$$

Задача 15. а)  $\langle \text{буква} \rangle ::= a | b | \dots | z$   
 $\langle \text{выражение} \rangle ::= \langle \text{буква} \rangle | \langle \text{выражение} \rangle \langle \text{выражение} \rangle \langle \text{выражение} \rangle * | \langle \text{выражение} \rangle \langle \text{выражение} \rangle$ .

б), с) Условия  $\alpha$ ) и  $\beta$ ) задачи 10, считая каждый знак  $+$  дважды (можно всюду заменить  $abc *$  на  $abc ++$ ).

## 1. ОБЩИЕ ПОНЯТИЯ

В произвольном интересующем нас множестве  $E$  выделяем подмножество  $F$ . Для элемента  $a$  из  $E$  указываем элемент  $\alpha$  из  $F$ , который «близок» к  $a$  и который будет называться *приближением* элемента  $a$ . Во всех практических расчетах  $a$  заменяется на  $\alpha$ . Основной вопрос состоит в том, чтобы узнать, будут ли «близкими» результаты, полученные с использованием  $\alpha$ , к тем, которые могли бы быть получены с использованием  $a$ .

**1.1. Погрешность.** Все обычно используемые множества (множество действительных чисел, множество комплексных чисел, множество числовых функций) наделены структурой коммутативной группы (по операции сложения).

*Погрешностью* пары элементов  $(a, \alpha)$  называется  $\alpha - a$ .

*Поправкой* этой пары  $(a, \alpha)$  называется  $a - \alpha$ .

Объясним, в чем здесь разница. Во всех теоретических исследованиях рассматривается значение  $a$ , которое изучается как основное. Значит, естественно сравнивать его с приближенным и, в соответствии с принятым у математиков, рассматривать разность

$$\alpha - a.$$

Напротив, если нам известно  $\alpha$ , то чтобы подправить это значение и получить  $a$ , надо прибавить к нему

$$a - \alpha,$$

ибо

$$\alpha + a - \alpha = a.$$

**1.2. Современное положение с определением этих понятий.** В настоящее время имеется неоднозначность в определении и обозначении этих понятий. Приведем несколько примеров, заимствованных из широко используемых книг (во Франции).

## П о г р е ш н о с т ь.

Обозначение	Определение
$\Delta a$	$a - \alpha$
$\Delta a$	$a - \alpha$
нет обозначения	$ a - \alpha $
$\delta a$	$a - a$
$da$ (в книгах по физике).	

Как нетрудно усмотреть, некоторые авторы называют погрешностью то, что мы называем поправкой; другие вводят абсолютное значение, которое делает невозможным рассмотрение некоторых случаев (этот способ определения связан с понятием «абсолютная погрешность»). Физики используют обычно дифференциальные обозначения.

Кроме того, слово «погрешность» иногда используется в смысле ошибки. Мы вернемся к этому в главе V.

**1.3. Относительная погрешность.** Если изучаемый элемент множества есть именованная величина, то погрешность имеет ту же размерность; физики часто заменяют ее безразмерным числом, а именно — относительной погрешностью. Однако это определение содержит в себе некоторые трудности. В частности, на какую меру надо делить погрешность — точную (что кажется более естественным, но трудно осуществимым практически) или на приближенную? Кроме того, обычные формулировки приближенной погрешности произведения, частного не являются математически точными.

Мы обойдем эти трудности, определив относительную погрешность пары

$$(a, \alpha)$$

как погрешность пары  $(\ln a, \ln \alpha)$ , т. е.

$$\ln \alpha - \ln a.$$

Нетрудно заметить, что изменение единицы измерения (или умножение на  $k$ ) не изменит относительной погрешности.

Это определение имеет то преимущество, что оно позволяет распространять на относительную погрешность все, что получено или может быть получено для собственно погрешности.

**2.1. Информация о приближении.** Связь между элементами  $a$  и  $\alpha$ , вообще говоря, довольно неопределенная.

Исключена, например, возможность точно выписать погрешность (иначе зачем работать с  $\alpha$  вместо того, чтобы взять непосредственно

$$a = \alpha - (\alpha - a).$$

Однако это отсутствие связи между  $a$  и  $\alpha$  не является полным, иначе не было бы никакого основания работать с  $\alpha$ , а не с каким-нибудь другим элементом. Всегда располагают какой-нибудь информацией о приближении.

Эта информация может быть достаточно разнообразной. Приведем некоторые типы информации.

— Знак: например, 3 есть приближенное значение числа  $\pi$  с недостатком.

— Интервал:  $\pi$  заключено между 3,1 и 3,2.

— Верхняя грань абсолютного значения погрешности: 3,1 есть приближенное значение  $\pi$ , отличающееся от него меньше чем на 0,1 по абсолютному значению.

— Бесконечно малый порядок:  $x$  есть приближенное значение  $\sin x$  ( $\sin x \approx x$  при  $x \rightarrow 0$ ).

— Порядок величины: 9,81 есть приближенное значение ускорения свободного падения в Париже с точностью до 0,01. Это выражение обычно используется физиками, но ему не надо придавать такой уж строгий смысл (тогда 9,821 было бы приемлемым значением, когда на самом деле это не так).

— Вероятностная информация.

Считают, что эта ситуация реализуется для погрешностей измерения. Она используется также и для некоторых интерпретаций ошибок округления.

## II. ИНТЕРВАЛ ПРИБЛИЖЕНИЯ

3.1. Приближение с недостатком, с избытком. Будем говорить, что элемент  $\alpha$  из  $F$  есть *приближение с недостатком* элемента  $a$  из  $E$ , если

$$\alpha \leq a.$$

Будем говорить, что  $\alpha$  есть *приближение с избытком*, если

$$a \leq \alpha.$$

При этом определении  $a$  будет для самого себя приближением одновременно с недостатком и с избытком (впрочем, это единственный случай, когда оба понятия совпадают).

### 3.2. Теорема о монотонных функциях. Пусть функция

$$f(x_1, \dots, x_p, y_1, \dots, y_q)$$

возрастает относительно аргументов  $x_1, \dots, x_p$  и убывает относительно аргументов  $y_1, \dots, y_q$ .

Если  $\alpha_1, \dots, \alpha_p$  — приближения для  $a_1, \dots, a_p$  с недостатком (с избытком),  $\beta_1, \dots, \beta_q$  — приближения для  $b_1, \dots, b_q$  с избытком (с недостатком), то имеем  $f(\alpha_1, \dots, \alpha_p, \beta_1, \dots, \beta_q)$  — приближение для  $f(a_1, \dots, a_p, b_1, \dots, b_q)$  с недостатком (с избытком).

Мы предоставляем читателям самостоятельно доказать это почти очевидное свойство.

#### 4.1. Вычисления посредством интервала приближения.

Практическое применение изложенного состоит в нахождении для элемента  $a$  приближения  $\underline{a}$  с недостатком и приближения  $\bar{a}$  с избытком:

$$\underline{a} \leq a \leq \bar{a}.$$

В этом случае говорят, что имеется *интервал приближения* элемента  $a$ .

**У п р а ж н е н и е 1.** Найти интервал приближения элемента:

а)  $a + b$ ; б)  $a - b$ , зная интервал приближения для каждого из элементов  $a$  и  $b$  в отдельности.

**У п р а ж н е н и е 2.** Можно ли в теореме о монотонных функциях говорить о возрастании и убывании по совокупности аргументов?

**4.2. Случай двойного интервала.** Предположим, что мы располагаем двумя информациями об интервале приближения элемента  $a$ :  $\underline{a} \leq a \leq \bar{a}$  и  $\underline{a}' \leq a \leq \bar{a}'$ .

Можно (без потери информации) заменить эти два интервала приближения на один:

$$\max(\underline{a}, \underline{a}') \leq a \leq \min(\bar{a}, \bar{a}').$$

#### 4.3. Замена интервала приближения. Ясно, что если

$$\beta \leq \underline{a} \text{ и } \bar{a} \leq \bar{\beta},$$

то из интервала приближения  $\underline{a} \leq a \leq \bar{a}$  можно получить интервал приближения

$$\beta \leq a \leq \bar{\beta},$$

по этот новый интервал приближения несет заведомо меньше информации, чем исходный.

**Задача 1.** а) Рассмотрим функцию

$$f(x) = x(x^2 - 1)$$

и интервал приближения переменного:  $0 \leq x \leq 1$ .

Найти возможно более точный интервал приближения функции  $f(x)$ .

б) Может ли оказаться, что функция  $f(x)$  не возрастает, но одновременно выполняются два условия:

$$\underline{\alpha} \leq a \leq \bar{\alpha}, \quad f(\underline{\alpha}) \leq f(a) \leq f(\bar{\alpha})?$$

### III. ПРИМЕНЕНИЕ НОРМЫ

Предположим, что мы работаем в пространствах, наделенных нормой. Норма элемента  $x$  обозначается  $\|x\|$ .

Таковы, например, пространства  $\mathbf{R}$  (абсолютное значение есть норма),  $\mathbf{C}$  (модуль есть норма), векторное пространство  $\mathbf{R}^n$  \*).

**5.1. Точность.** Будем говорить, что пара  $(a, \alpha)$  наделена *точностью*  $\epsilon$ , если справедливо неравенство

$$\|a - \alpha\| \leq \epsilon.$$

В этом случае элемент  $\alpha$  называется *приближением* элемента  $a$  с *точностью*  $\epsilon$ .

**З а м е ч а н и е.** В некоторых (редких) случаях может представлять интерес использование неравенства

$$\|a - \alpha\| < \epsilon.$$

Тогда говорят о *приближении по крайней мере с точностью*  $\epsilon$ .

**З а м е ч а н и е.** В случае множества  $\mathbf{R}$  из неравенства для нормы  $|a - \alpha| \leq \epsilon$  получаем

$$\alpha - \epsilon \leq a \leq \alpha + \epsilon,$$

т. е. получаем интервал приближения. Однако желательно четко различать два понятия, и не только потому, что различны математические теории, но и потому, что практические вычисления не совсем одинаковы.

---

\*) *Норма* — понятие, близкое к понятию расстояния в его абстрактной форме.  $\mathbf{R}$  — пространство действительных чисел; норму в нем можно определить с помощью соотношения  $\|x\| = |x|$ .  $\mathbf{C}$  — пространство комплексных чисел;  $\|z\| = |z|$ . Векторное пространство  $\mathbf{R}^n$  —  $n$ -мерное пространство, которое является естественным обобщением наших представлений о реальном трехмерном пространстве  $\mathbf{R}^3$ . Подчеркнем, что в одном и том же пространстве норму можно ввести по-разному. С соответствующими примерами вы встретитесь в упражнении 3 и задаче 3.

У п р а ж н е н и е 3. Рассмотрим в  $\mathbb{C}$  норму  $|z|$  и норму

$$\|z\| = |\operatorname{Re} z| + |\operatorname{Im} z|.$$

а) Пусть пара  $(a, \alpha)$  наделена относительно первой нормы точностью  $\epsilon$ . Что можно сказать о ее точности  $\epsilon'$  относительно второй нормы?

б) Пусть пара  $(a, \alpha)$  наделена точностью  $\epsilon'$  относительно второй нормы. Что можно сказать о ее точности  $\tilde{\epsilon}$  относительно первой нормы?

с) Можно ли сформулировать какое-либо заключение?

**5.2. Точность таблицы.** Очень важным случаем понятия точности является понятие точности таблицы функций. В ней находят приближения

значений функции  $f$ :

$$f(x_1), f(x_2), \dots, f(x_n).$$

Пары  $(f(x_i), f_i)$ ,  $i = 1, \dots, n$ , вообще говоря, обладают одной и той же точностью, которая называется *точностью таблицы*.

**5.3. Точность линейной комбинации.** Пусть в векторном пространстве  $\mathbb{R}^n$  заданы элементы  $a_i$ , имеющие приближения  $\alpha_i$  с точностями  $\epsilon_i$ .

Рассмотрим элемент

$$a = \sum_i k_i a_i$$

с приближением

$$\alpha = \sum_i k_i \alpha_i.$$

Из определения нормы вытекает, что

$$\left\| \sum_i k_i \alpha_i - \sum_i k_i a_i \right\| \leq \sum_i |k_i| \|\alpha_i - a_i\|.$$

Отсюда следует, что пара  $(a, \alpha)$  наделена точностью

$$\sum_i |k_i| \epsilon_i.$$

У п р а ж н е н и е 4. Рассмотрим функцию

$$f(x, y, z)$$

с переменными и значениями из  $\mathbb{R}$ .

Предположим, что эта функция обладает первыми частными производными, удовлетворяющими условиям

$$|f'_x| \leq M_x, \quad |f'_y| \leq M_y, \quad |f'_z| \leq M_z,$$

где  $M_x, M_y, M_z$  — заданные константы.

Что можно утверждать о точности  $\varepsilon$  пары

$$(f(a, b, c), f(\alpha, \beta, \gamma)),$$

зная точности  $\varepsilon_x, \varepsilon_y, \varepsilon_z$  пар

$$(a, \alpha), (b, \beta), (c, \gamma)?$$

**6.1. Приближение и сходимость.** Пусть в  $\mathbb{R}^n$  задано всюду плотное множество  $F$ .

Допустим, что можно найти приближения  $\alpha_1, \alpha_2, \dots, \dots, \alpha_n, \dots$  элемента  $a$  с точностями соответственно  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n, \dots$ , стремящимися к нулю. Отсюда следует, что последовательность  $\alpha_1, \alpha_2, \dots, \alpha_n, \dots$  сходится к  $a$ .

В обратном направлении, знание сходимости последовательности

$$\alpha_1, \alpha_2, \dots, \alpha_n, \dots$$

к  $a$  если и позволяет утверждать, что точность пары

$$(a, \alpha_n)$$

может стремиться к нулю вместе с  $n$ , то оно не позволяет сделать никаких утверждений о точности конкретной пары, такой, как  $(a, \alpha_0)$ .

На самом деле теория приближения есть теория гораздо более сложная (и менее изученная), чем теория сходимости, и практическое использование понятия погрешности является гораздо более трудным, чем использование понятий классического анализа.

**З а д а ч а 2.** Рассмотрим в  $\mathbb{R}$  интервал приближения элемента  $a$ :

$$\underline{a} \leq a \leq \bar{a}.$$

а) Можно ли, принимая в качестве нормы абсолютное значение, найти приближение  $\alpha$  и точность  $\varepsilon$ , которые давали бы информацию, эквивалентную заданной?

б) Пусть  $\beta$  — приближение элемента  $a$ . Какая точность соответствует паре  $(a, \beta)$ , если в качестве информации имеется приведенный выше интервал приближения?

с) В каких из перечисленных случаев произошла потеря информации?

**Задача 3.** а) Рассмотрим в трехмерном пространстве нормы

$$N_1 = |x| + |y| + |z|,$$

$$N_2 = \sqrt{x^2 + y^2 + z^2},$$

$$N_3 = \max(|x|, |y|, |z|).$$

Знание точности  $\varepsilon$  относительно нормы  $N_i$ ,  $i = 1, 2, 3$ , дает знание точности  $K_{ij}\varepsilon$  относительно нормы  $N_j$ ,  $j = 1, 2, 3$ ,  $j \neq i$ .

Выписать  $K_{ij}$  в виде таблицы с двойным входом.

б) Пусть заданы приближение  $\alpha$  и точность  $\varepsilon$ , и пусть точка  $a$  лежит в некоторой области, имеющей объем  $V$ . Определить этот объем для каждой из норм.

#### IV. СИСТЕМАТИЧЕСКИЕ ДЕСЯТИЧНЫЕ ПРИБЛИЖЕНИЯ

На протяжении всего этого параграфа через  $E$  будет обозначаться множество действительных чисел, а через  $F$  — множество десятичных дробей.

Мы изучим систематические процедуры, позволяющие сопоставлять действительному числу приближения, которые будут десятичными дробями.

**7.1. Целая часть действительного числа.** *Целой частью* действительного числа  $a$ , обозначаемой

$$E(a),$$

называется наибольшее целое число, меньше или равное  $a$ .

**Пример.**

$$E(1) = 1, \quad E(0,99) = 0,$$

$$E(\sqrt{2}) = 1, \quad E(\pi) = 3,$$

$$E(-\pi) = -4, \quad E(1,01) = 1.$$

Представим на прямой значения, принимаемые функцией  $F(a)$  (рис. 1).

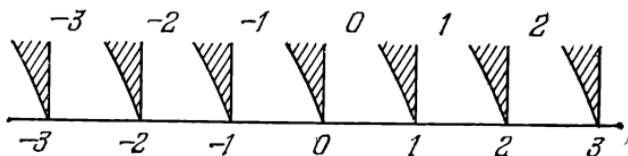


Рис. 1.

Ясно, что эта функция обладает свойством

$$E(a + 1) = 1 + E(a).$$

Соотношение между  $E(a)$  и  $E(-a)$  уже не будет таким простым:

$$E(-a) = \begin{cases} -1 - E(a), & \text{если } a \text{ — нецелое,} \\ -E(a), & \text{если } a \text{ — целое.} \end{cases}$$

Отметим еще свойства:

$$E(E(a)) = E(a);$$

из  $a \geq b$  следует  $E(a) \geq E(b)$  (возрастание в широком смысле).

**7.2. Запись отрицательных чисел без знака.** Понятие целой части числа хорошо приспособлено для такой записи. В самом деле, целая часть числа в записи без знака при помощи цифр получается путем замены всего, что стоит слева от запятой, на нули.

**Пример.** В емкости  $\boxed{\times \times, \times \times \times \times}$

$$E(\pi) = 03, 0000;$$

—  $\pi$  записывается в виде 96, 8584,

$$E(-\pi) = 96 \text{ (т. е. } -4\text{)};$$

— 4 записывается в виде 96, 0000,

$$E(-4) = -4.$$

Из сказанного выше следует, что если работают с ограниченной емкостью, то всякое число, имеющее некоторую запись, имеет также запись и для своей целой части

**У п р а ж н е н и е 5.** Предположим, что используется запись без знака в неограниченной емкости направо и принимается так называемая несобственная запись, например,

$$00,999999 \dots \text{ для } 1.$$

Останутся ли справедливыми предыдущие результаты?

**8.1. Десятичное приближение порядка  $n$  с недостатком.** Положим, по определению,

$$\alpha_n = 10^{-n} E(a \cdot 10^n).$$

$\alpha_n$  будем называть десятичным приближением порядка  $n$  с недостатком числа  $a$ .

В частности, запись без знака хорошо приспособлена для работы с этим понятием.

**8.2. Десятичное приближение порядка  $n$  с избытком.** Определим десятичное приближение порядка  $n$  с избытком числа  $a$ :

$$\bar{\alpha}_n = \underline{\alpha}_n + 10^{-n}.$$

Отсюда выводим

$$\underline{\alpha}_n \leq a < \bar{\alpha}_n.$$

Отметим отсутствие симметрии между этими определениями  $\bar{\alpha}_n$  и  $\underline{\alpha}_n$ : никакое число не равно своему десятичному приближению порядка  $n$  с избытком.

**9.1. Замена десятичного порядка.** Пусть  $a$  — некоторое число,  $\underline{\alpha}_n$  — его десятичное приближение с недостатком порядка  $n$  и  $p < n$ . Тогда

$$\underline{\alpha}_n = (\underline{\alpha}_p)_p.$$

В самом деле, неравенство  $\underline{\alpha}_n \leq a$  влечет неравенство  $(\underline{\alpha}_n)_p \leq \underline{\alpha}_p$ , где десятичное приближение  $\underline{\alpha}_p$  порядка  $p$  может быть лишь меньше или равно  $\underline{\alpha}_n$ . Но  $\underline{\alpha}_p \leq \underline{\alpha}_n$  влечет  $\underline{\alpha}_p \leq (\underline{\alpha}_n)_p$ , откуда и следует искомое равенство.

**У п р а ж н е н и е 6.** а) Как найти десятичное приближение порядка  $n$  с недостатком числа, записанного в записи без знака?

б) Как перейти от приближения порядка  $n$  к приближению порядка  $p$  ( $p < n$ )?

с) Применить к числу  $-6,5283$

— запись без знака в емкости  $\boxed{\times \times, \times \times \times}$ ,

— приближение порядка 3,

— приближение порядка 2.

д) Будет ли правило замены десятичного порядка столь же простым в записи при помощи абсолютной величины и знака?

**10.1. Натуральная целая часть числа  $a$ .** Модифицируем введенное выше понятие целой части, чтобы сделать его более применимым в случае записи при помощи абсолютной величины и знака.

Натуральная целая часть числа  $a$  есть число

$$[a],$$

определенное следующим образом:

$$[a] = \begin{cases} E(a), & \text{если } a \geq 0, \\ -E(-a), & \text{если } a \leq 0. \end{cases}$$

Пример:

$$[2, 4] = 2, [-2, 4] = -2,$$

тогда как

$$E(2, 4) = 2, E(-2, 4) = -3.$$

Легко представить значения функции  $[a]$  на прямой (рис. 2).

Эта функция удовлетворяет условию

$$[-a] = -[a].$$

Но симметрия натуральной целой части числа компенсируется некоторой аномалией, поскольку в окрестности 0 она равна нулю на интервале длины 2.

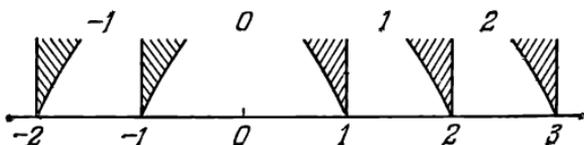


Рис. 2.

Эта функция удовлетворяет также условиям

$$[[a]] = [a],$$

из  $a \geq b$  следует  $[a] \geq [b]$ .

У п р а ж н е н и е 7. Выразить  $[a + 1]$  через  $[a]$  и  $a$ .

**10.2. Получение натуральной целой части в случае записи со знаком.** Получим натуральную целую часть конечной десятичной дроби  $a$ , исключая в ее записи со знаком цифры справа от запятой. Из предыдущего следует, что (в ограниченной емкости) если число имеет запись со знаком, то его натуральная целая часть тоже имеет запись со знаком.

**11.1. Натуральное десятичное приближение порядка  $n$ .** *Натуральным десятичным приближением порядка  $n$  числа  $a$  называется число*

$$\alpha_n = [a \cdot 10^n] 10^{-n}.$$

Ясно, что это приближение будет приближением с недостатком для положительных и с избытком для отрицательных чисел.

**11.2. Замена десятичного порядка.** Пусть  $\alpha_n$  — натуральное приближение порядка  $n$  числа  $a$  и пусть  $p < n$ .

Тогда

$$\alpha_p = (\alpha_n)_p.$$

В силу симметрии достаточно провести доказательство для  $a \geq 0$ ; тогда  $\alpha_n \geq 0$  и  $\alpha_p \geq 0$ . Неравенство  $\alpha_n \leq \alpha$  влечет неравенство  $(\alpha_n)_p \leq \alpha_p$ , где натуральное приближение  $\alpha_p$  порядка  $p$  может быть лишь меньше или равно  $\alpha_n$ . Но  $\alpha_p \leq \alpha_n$ ; следовательно,  $\alpha_p \leq (\alpha_n)_p$ , откуда и получаем искомое равенство.

У п р а ж н е н и е 8. а) Определить натуральное десятичное приближение порядка 3 для числа

$$3,141592.$$

б) Получить натуральное приближение порядка 1.

с) Сформулировать правило получения приближения порядка 1 исходя из приближения порядка 3.

**12.1. Наилучшее приближение порядка  $n$ .** Наиболее близкая по норме к числу  $a$  дробь  $n$ -го порядка называется *наилучшим приближением* числа  $a$ . Наилучшее приближение единственно для всех чисел, кроме чисел, допускающих конечную десятичную запись и таких, что самая правая отличная от нуля цифра есть 5 при десятичном порядке  $n + 1$ . Тогда имеются два решения.

Во всех случаях можно в качестве точности взять

$$\frac{1}{2} \cdot 10^{-n}.$$

П р и м е р. 2,15 имеет наилучшие приближения порядка 1:

$$2,1 \text{ и } 2,2.$$

**12.2. Автоматическое округление порядка  $n$ .** Мы приведем простой процесс нахождения наилучшего приближения порядка  $n$  числа  $a$ . Правда, этот процесс содержит неоднозначность, на которую указывалось в предыдущем пункте. Он зависит от того, какая принята запись — со знаком или без знака. Процесс состоит в том, что к десятичной цифре, стоящей на  $n + 1$ -ом месте после запятой, как к числу, прибавляется число 5, и в полученном результате все цифры после  $n$ -й отбрасываются.

П р и м е р.

Округление  $\pi$  порядка 3 дает 3,142,  
округление  $\pi$  порядка 2 дает 3,14.

Пусть теперь имеются отрицательные числа, записанные при помощи абсолютной величины и знака. Тогда округление  $-0,7354$  порядка 3 дает  $-0,735$ , округление  $-0,7354$  порядка 2 дает  $-0,74$ .

В записи без знака

число  $-0,7354$  принимает вид  $9,2646$ , округление порядка 3 дает  $9,265$  ( $-0,735$ ), округление порядка 2 дает  $9,26$  ( $-0,74$ ).

Наконец, возьмем

число  $-0,735$ , или, в записи без знака,  $9,265$ ; для округления порядка 2 находим:

в записи со знаком  $-0,74$ ,

в записи без знака  $9,27$  ( $-0,73$ ).

**З а м е ч а н и е.** Можно предложить другие правила для избежания неоднозначности, но они менее автоматичны. Отметим, например, правило Гаусса. В случае неоднозначности выбирают ту запись, которая оканчивается четной цифрой.

**У п р а ж н е н и е 9.** Показать, что если работают с ограниченной емкостью, то может оказаться, что число имеет запись в этой емкости, а его наилучшее приближение не имеет.

**У п р а ж н е н и е 10.** а) Показать, что наилучшее приближение порядка  $n$  не всегда допускает нахождение наилучшего приближения порядка  $p$  ( $p < n$ ).

б) Если взять наилучшее приближение порядка  $p$  наилучшего приближения порядка  $n$ , то какую точность нужно ему обеспечить?

**12.3. Свойства автоматического округления.** Мы будем обозначать автоматическое округление порядка  $n$  числа  $a$  через

$$\alpha_n^*.$$

Читатель без труда убедится в том, что

$$(\alpha_n^*)_n = \alpha_n^*;$$

из  $a \geq b$  следует  $\alpha_n^* \geq \beta_n^*$ .

**У п р а ж н е н и е 11.** Будут ли выполняться сформулированные выше свойства для округления, полученного по правилу Гаусса?

**13.1. Нахождение десятичного приближенного значения порядка  $n$  одного из трех приведенных типов, исходя из некоторого приближения.** Предположим, что нам изве-

стен для числа  $a$  интервал приближения

$$\underline{a} \leq a \leq \bar{a}.$$

Нахождение приближенного значения  $\underline{a}_n$  с недостатком невозможно, если существует десятичное число

$$\frac{N}{10^n},$$

удовлетворяющее условиям  $\underline{a}_n < \frac{N}{10^n} \leq \bar{a}$ , и возможно в противном случае.

Нахождение наилучшего приближенного значения порядка  $n$  невозможно, если существует такое  $N$ , что

$$\underline{a} < \frac{N + 1/2}{10^n} < \bar{a}.$$

Оно может привести к неоднозначности, если одно из неравенств становится равенством.

Нахождение десятичного приближения порядка  $n$  сопровождается потерей информации.

Мы оставим эти приближения для вполне определенных случаев, в частности, для представления определенных результатов (например, в числовых таблицах).

**14.1. Скользящее приближение порядка  $n$ .** Пусть  $a$  — отличное от нуля число. Мы ограничимся изучением наилучшего приближения для чисел, мантисса которых записывается при помощи абсолютного значения и знака.

Будем называть *скользящим наилучшим приближением порядка  $n$*  число, полученное следующим образом: прибавим к значащей части мантиссы  $5/10^{n+1}$  и исключим все цифры мантиссы после  $n$ -й.

Если полученное таким образом число  $r$  есть мантисса, то его сохраняют, так же как и его знак и степень. Если же число  $r$  уже не будет мантиссой, то его самая правая цифра будет 0; теперь заменяем  $r$  на  $r/10$ , ставим на место знак и прибавляем 1 к показателю.

**Пример.**  $n = 3$ ;  
 $0,9991 \cdot 10^{-3}$  дает  $0,999 \cdot 10^{-3}$ ,  
 $0,9998 \cdot 10^{-3}$  дает  $0,100 \cdot 10^{-2}$ .

**У п р а ж н е н и е 12.** Всегда ли возможно нахождение скользящего наилучшего приближения порядка  $n$ , если работать с ограниченной емкостью?

**З а д а ч а 4.** а) Что можно сказать относительно  $E(a + b)$ , если известны  $E(a)$  и  $E(b)$ ?

б) Тот же вопрос относительно  $E(a - b)$ .

с) Те же вопросы относительно  $[a + b]$ ,  $[a - b]$ .

**З а д а ч а 5.** Предположим, что действительное число записывается в виде конечной или бесконечной десятичной последовательности в записи при помощи абсолютного значения и знака.

Как можно найти его десятичное приближение порядка  $n$ ? Нужна ли предосторожность в частном случае десятичной дроби?

**З а д а ч а 6.** Предположим, что мы находим для  $K$  чисел наилучшие приближения порядка  $n$ , исходя из интервала приближения длины  $l$  ( $l < 10^{-n}$ ). Не имея никакой специальной информации о разбиении этих интервалов длины  $l$ , допустим, что число

$$\frac{N + 1/2}{10^n}$$

представимо с частотой  $l \cdot 10^n$ .

а) Каково число ожидаемых случаев невозможности нахождения наилучшего приближения порядка  $n$ ?

б) **П р и л о ж е н и е.** Таблица логарифмов с 5 десятичными знаками чисел от 1001 до 10000:

$$l = 10^{-7}, \quad l = 10^{-9}.$$

**З а д а ч а 7.** Предположим, что  $N$  чисел распределены случайным образом на отрезке  $[0,1]$  с постоянной плотностью распределения.

а) Что можно сказать о среднем этих чисел и о среднем их десятичных приближений порядка  $n$  с недостатком?

б) Та же задача для наилучшего приближения порядка  $n$ .

с) Тот же вопрос, что и в б), но с тем дополнением, что числа имеют вид

$$\frac{p + 1/2}{10^n}.$$

### РЕШЕНИЯ УПРАЖНЕНИЙ ГЛАВЫ III

1) а)  $\underline{\alpha} + \underline{\beta} \leq a + b \leq \bar{\alpha} + \bar{\beta}$ ; б)  $\underline{\alpha} - \bar{\beta} \leq a - b \leq \bar{\alpha} - \underline{\beta}$ .

2) Да.

3) а)  $\varepsilon' = \varepsilon \sqrt{2}$ ; б)  $\bar{\varepsilon} = \varepsilon'$ ; с) не следует без особой необходимости менять норму.

4) Можно взять  $\varepsilon = \varepsilon_x M_x + \varepsilon_y M_y + \varepsilon_z M_z$ .

5) Нет; необходимо запретить такую запись как 00,999... для 1 и 91,999999... для  $-8$ .

Достаточно положить за правило, что если число может быть записано посредством конечного числа отличных от нуля цифр, то мы обязаны использовать эту запись.

6) а) Исключим все цифры, соответствующие позициям справа от  $n$ -й.

б) Исключим в  $a_n$  все цифры, соответствующие позициям справа после  $p$ -й.

с) 93,4717; 93,471 (т. е.  $-6,529$ )  
93,47 (т. е.  $-6,53$ ).

д) Нет.

7) Для  $a \leq -1$  и  $a \geq 0$   $[a + 1] = [a] + 1$ .

Для  $-1 < a < 0$   $[a + 1] = [a]$ .

8) а) 3,141; б) 3,1; с) исключаем последние два десятичных знака.

9) Пример: емкость  $\boxed{\times\times, \times\times}$ .

Число 99,99 имеет в качестве наилучшего приближения порядка 1 число 100,0 которое не имеет записи в этой емкости.

10) а) 0,349; его наилучшее приближение порядка 2 равно 0,35.

Наилучшее приближение порядка 1 равно 0,3, тогда как для 0,35 оно двойственно.

б)  $(10^{-n} + 10^{-p})/2$ .

11) Да.

12) Нет, может оказаться переход показателя (так, для  $n = 3$   $0,9998 \cdot 10^{99}$  дает  $0,1000 \cdot 10^{100}$ ), что не может быть записано, если иметь только две позиции для показателя.

### РЕШЕНИЯ ЗАДАЧ ГЛАВЫ III

Задача 1. а)  $\frac{-2}{3\sqrt{3}} \leq f(x) \leq 0$ .

б) Да, предыдущая функция с интервалом приближения  $-2 \leq x \leq 2$ .

Задача 2. а)  $\alpha = \frac{\bar{\alpha} + \alpha}{2}$ ,  $\varepsilon = \frac{\bar{\alpha} - \alpha}{2}$ .

б)  $\max(|\beta - \alpha|, |\beta - \bar{\alpha}|)$ .

с) В случае б).

Задача 3. а)

	$N_1$	$N_2$	$N_3$
$N_1$	1	1	1
$N_2$	$\sqrt{3}$	1	1
$N_3$	3	$\sqrt{3}$	1

б)  $V_1 = \frac{4}{8} \pi e^3$ ,  $V_2 = \frac{4}{8} \pi e^3$ ,  $V_3 = 8e^3$ .

Задача 4. а)  $E(a) + E(b) \leq E(a + b) \leq E(a) + E(b) + 1$ .

б)  $E(a) - E(b) - 1 \leq E(a - b) \leq E(a) - E(b)$ .

При доказательстве этих двух соотношений можно свести все к случаю  $E(a) = E(b) = 0$ .

с) В силу симметрии можно исследовать только случай

$$a \geq b \geq 0,$$

$$[a] + [b] \leq [a + b] \leq [a] + [b] + 1,$$

$$0 \leq [a - b] \leq [a] - [b].$$

З а д а ч а 5. Необходимо запретить такие записи, как

0,1999999 ... для числа 0,2

(т. е. надо потребовать, что если число имеет конечную десятичную запись, то она и должна использоваться). Тогда правило состоит в исключении всех десятичных знаков справа от  $n$ -го.

З а д а ч а 6. а)  $Kl \cdot 10^n$ . б)  $K = 9000$ ,  $n = 5$ .

Для  $l = 10^{-7}$  следует ожидать 90 случаев невозможности.

Для  $l = 10^{-9}$  следует ожидать около 1 случая.

З а д а ч а 7. а) Среднее чисел равно 0,5. Среднее приближений равно  $0,5 \cdot (1 - 10^{-n})$ .

б) Оба средних равны 0,5.

с) Среднее приближений равно  $0,5 \cdot (1 + 10^{-n})$ .

## I. СИСТЕМА ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

**1.1 Напоминание результатов.** Система линейных алгебраических уравнений может быть записана в матричной форме

$$AX = B;$$

$A$  — матрица коэффициентов при неизвестных,

$X$  — столбец неизвестных,

$B$  — столбец правых частей.

Мы ограничимся случаем  $n$  уравнений с  $n$  неизвестными.

Математическое решение этой задачи хорошо известно:

Если  $\det A \neq 0$ , то система имеет, и притом единственное, решение

$$X = A^{-1}B.$$

Если  $\det A = 0$ , то система несовместна или неопределенна. Но мы оставим в стороне этот случай.

В том случае, когда решение определено, каждое неизвестное можно выписать явно в виде частного двух определителей порядка  $n$ . Если коэффициенты рациональны, то неизвестные тоже рациональны.

**1.2. Численный аспект решения.** На самом деле эти результаты создают довольно ложное впечатление о реальности решения систем, начиная с достаточно большого  $n$  (на практике часто встречаются системы 100 уравнений со 100 неизвестными).

— С одной стороны, в случае рациональных коэффициентов запись в виде дроби от нескольких неизвестных требует привлечения чрезвычайно больших целых чисел.

— С другой стороны, вычисление определителя порядка  $n$ , который содержит  $n!$  различных членов, для больших  $n$  становится делом безнадежным. Как мы увидим ниже, существуют методы, для которых порядок числа операций будет гораздо меньше, чем  $n!$ .

**2.1. Принципы решения.** Согласно п. 1.1 найти решение просто во всех случаях, когда легко найти обратную матрицу  $A^{-1}$ . Существуют различные категории матриц, которые обладают этим свойством.

а) **Д и а г о н а л ь н а я м а т р и ц а.** Пусть  $\Delta$  — матрица, все элементы которой, лежащие вне главной диагонали, равны нулю. Обратная к ней есть диагональная матрица  $\Delta^{-1}$ .

Система, матрица которой диагональна, в действительности состоит из  $n$  независимых уравнений первой степени.

б) **О р т о г о н а л ь н а я м а т р и ц а о т н о с и т е л ь н о с т о л б ц о в.** Это матрица  $A$ , удовлетворяющая условию

$$A^T A = \Delta,$$

где  $\Delta$  — диагональная матрица,  $A^T$  — транспонированная к  $A$ .

Система  $A\dot{X} = B$  может быть записана в виде

$$\Delta X = A^T B,$$

и все сводится к предыдущему случаю а).

с) **В е р х н е т р е у г о л ь н а я м а т р и ц а.** Так называют квадратную матрицу, все элементы которой, расположенные ниже главной диагонали, равны нулю. Система, если принять  $n = 4$ , записывается в виде

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 &= b_1, \\ a_{22}x_2 + a_{23}x_3 + a_{24}x_4 &= b_2, \\ a_{33}x_3 + a_{34}x_4 &= b_3, \\ a_{44}x_4 &= b_4. \end{aligned}$$

Определитель системы равен

$$a_{11}a_{22}a_{33}a_{44}.$$

Поскольку мы предполагаем, что он отличен от нуля, то

$$a_{11} \neq 0, \quad a_{22} \neq 0, \quad a_{33} \neq 0, \quad a_{44} \neq 0.$$

Ясно, что система разрешима. Из последнего уравнения определяется  $x_4$ . Тогда третье уравнение определяет  $x_3$ , далее  $x_2$ , затем  $x_1$ .

В общем случае матрица системы уравнений очень редко обладает указанными свойствами. Но мы можем пытаться от плохой матрицы преобразованием системы перейти к матрице с указанными свойствами. Именно это

мы и будем сейчас делать для третьего из приведенных случаев, т. е. будем приходить к треугольной матрице.

**3.1. Метод преобразования Гаусса.** Попробуем получить нули в первом столбце системы ниже главной диагонали. Для этого вычитаем первое уравнение из 2-го, 3-го ... уравнений с соответствующими множителями.

Пусть

$$\begin{aligned}x + 2y + 3z + 4t &= 5, \\3x + 5y - 4z - 2t &= 0, \\2x - y + 5z - t &= 2, \\-x + 7y - z &= 1.\end{aligned}$$

Получаем

$$\begin{aligned}x + 2y + 3z + 4t &= 5, \\-y - 13z - 14t &= -15, \\-5y - z - 9t &= -8, \\9y + 2z + 4t &= 6.\end{aligned}$$

Затем, не трогая первого уравнения, получим нули во втором столбце ниже диагонали, вычитая (с соответствующими множителями) из уравнений 3-го, 4-го, ... второе.

В нашем примере получим

$$\begin{aligned}x + 2y + 3z + 4t &= 5, \\y - 13z - 14t &= -15, \\64z + 61t &= 67, \\-115z - 122t &= -129.\end{aligned}$$

Действуя далее аналогично, закончим этот пример!

$$\begin{aligned}x + 2y + 3z + 4t &= 5, \\-y - 13z - 14t &= -15, \\64z + 61t &= 67, \\-\frac{793t}{64} &= -\frac{551}{64}.\end{aligned}$$

Отсюда получаем  $t = 551/793$ , затем  $z = 305/793$ , затем  $y = 216/793$ , затем  $x = 414/793$ .

**3.2. Число операций.** Подсчитаем, сколько операций содержит это решение.

а) Подбор множителей:

$$(n-1) + (n-2) + \dots + 1 = \frac{n(n-1)}{2} \text{ делений.}$$

б) Комбинация уравнений:

$$n(n-1) + (n-1)(n-2) + \dots + 2 \cdot 1,$$

а именно:

$$\frac{n(n+1)(n-1)}{3} \text{ умножений и сложений.}$$

с) Окончательное решение:

$n$  делений,

$$(n-1) + (n-2) + \dots + 1 = \frac{n(n-1)}{2}$$

умножений и сложений.

Стало быть, всего

$$\frac{n(n+1)}{2} \text{ делений,}$$

$$n(n-1) \left( \frac{n+1}{5} + \frac{1}{2} \right) \text{ умножений и сложений.}$$

Легко видеть, что общее число умножений и сложений имеет порядок  $n^3/3$ .

**У п р а ж н е н и е 1.** а) Сколько нужно операций, чтобы решить систему двух уравнений?

б) Проверить полученный результат при помощи прямой перенумерации.

**У п р а ж н е н и е 2.** Сколько времени понадобится для решения на машине системы со 100 неизвестными, если известно, что каждая операция длится около  $10^{-5}$  секунд и что только половина времени тратится на операции?

**4.1. Выбор диагонали.** В описанном методе предполагается, что члены, которые становятся членами главной диагонали треугольной матрицы, все отличны от нуля. Это условие может оказаться не выполненным. Может случиться, например, что  $a_{11} = 0$ .

Но поскольку мы предположили, что  $A \neq 0$ , то в первом столбце обязательно найдется отличный от нуля элемент, допустим,  $a_{i1} \neq 0$ .

Переставляя  $i$ -ю строку с 1-й, приводим систему к виду, в котором можно получить желаемые нули в первом столбце. Тот же процесс, в случае необходимости, может быть произведен и с другими столбцами.

У п р а ж н е н и е 3. а) Применить этот метод к системе, матрица которой имеет вид

$$\begin{bmatrix} 0 & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & \times & \times \\ 0 & 0 & 0 & 0 & \times \\ \times & \times & \times & \times & \times \end{bmatrix}.$$

Крестами обозначаются отличные от нуля элементы.

б) Можно ли предложить лучший метод для решения этой системы?

4.2. Плохо обусловленная система. Решение системы уравнений первой степени может в некоторых случаях привести к трудностям, которые мы продемонстрируем на примере.

Пусть имеется система

$$\begin{aligned} 3,0000 x - 7,0001 y &= 0,9999, \\ 3,0000 x - 7,0000 y &= 1. \end{aligned} \quad (1)$$

Правые части выбраны так, чтобы система имела решение

$$x = 5, y = 2.$$

Изучим теперь систему, близкую к исходной:

$$\begin{aligned} 3,0000 x - 7,0001 y &= 1, \\ 3,0000 x - 7,0000 y &= 1. \end{aligned}$$

Эта система имеет очевидное решение

$$x = 1/3, y = 0.$$

Изменение на  $10^{-4}$  одной из правых частей очень сильно изменило решение системы.

Рассмотрим снова систему, очень близкую к первой:

$$\begin{aligned} 3,0000 x - 7,0000 y &= 0,9999, \\ 3,0000 x - 7,0000 y &= 1. \end{aligned}$$

Легко видеть, что эта система несовместна.

Системы такого типа называются *плохо обусловленными*.

Они довольно часто встречаются в приложениях.

З а д а ч а 1. Сколько потребуется дополнительных операций, если вместо того чтобы решать единственную систему уравнений, мы будем решать одновременно вто-

рую систему, которая отличается от первой лишь правыми частями?

**Задача 2.** а) Найти матрицу, обратную матрице левой части системы (1).

б) Может ли полученный результат объяснить, почему эта система очень чувствительна к малому изменению правых частей?

с) Можно ли сформулировать обратное утверждение?

## II. МНОГОЧЛЕНЫ. ИНТЕРПОЛЯЦИЯ

В главе III мы видели, что богатство множества действительных чисел обязывает ввести понятие численного приближения. Понятие функции приводит нас к множествам еще более широким. Кроме того, такие операции над функциями как дифференцирование, интегрирование требуют перехода к пределу, т. е. бесконечную последовательность этапов, которые, разумеется, нельзя пробежать реально.

Способ обойти эту трудность заключается в том, чтобы обратиться к приближению многочленами. Операции дифференцирования и интегрирования для многочленов являются алгебраическими операциями. Более того, для непрерывной на отрезке  $[0, 1]$  функции можно найти сколь угодно хорошие приближения многочленами.

Приступим к действиям с многочленами.

**5.1. Многочлен степени не более  $n$ , принимающий заданные значения для  $n + 1$  заданных значений переменного.** Одна из наиболее важных проблем состоит в том, чтобы уметь записать многочлен  $P_n(x) = f(x)$  степени не выше  $n$ , обладающий тем свойством, что

$$P_n(a_i) = f_i, \quad i = 0, 1, 2, \dots, n.$$

Такой многочлен, очевидно, единственный. В самом деле, если существуют два таких многочлена  $P_n(x)$ ,  $Q_n(x)$ , то уравнение

$$P_n(x) - Q_n(x) = 0$$

степени не выше  $n$  имеет  $n + 1$  корней  $a_i$ , что невозможно.

Существование такого многочлена будет вытекать из самой конструкции многочленов.

**У п р а ж н е н и е 4.** Показать при помощи записи условий, определяющих коэффициенты этого многочлена,

а) что если многочлен никогда не является неопределенным, то он существует и единствен;

б) что многочлен существует и единствен (не принимая во внимание приведенное выше доказательство единственности).

**5.2. Коэффициенты Лагранжа.** Коэффициентами Лагранжа называют многочлены

$$L_i(x) = \frac{(x - a_0) \dots (x - a_{i-1})(x - a_{i+1}) \dots (x - a_n)}{(a_i - a_0) \dots (a_i - a_{i-1})(a_i - a_{i+1}) \dots (a_i - a_n)},$$
$$i = 0, 1, \dots, n.$$

Эти многочлены обладают следующими свойствами:

— они имеют степень  $n$ ,

—  $L_i(a_i) = 1$ ,

—  $L_i(a_j) = 0$  для  $j \neq i$ .

**5.3. Выражение для многочлена, принимающего заданные значения.** Ясно, что многочлен

$$P_n(x) = \sum_i f_i L_i(x)$$

— имеет степень не выше  $n$ ,

— удовлетворяет условиям  $P_n(a_i) = f_i$ .

Будем называть выражение для  $P_n(x)$  *формой Лагранжа*.

**У п р а ж н е н и е 5.** Выписать коэффициенты Лагранжа и  $P_n(x)$  для

а)  $n = 1$ ,

б)  $n = 2$ .

**У п р а ж н е н и е 6.** Интерпретировать

$$L_i(a_j) = \begin{cases} 1 & \text{для } i = j, \\ 0 & \text{для } i \neq j, \end{cases}$$

как произведение двух матриц.

**6.1. Барицентрическая формула.** В форме, приведенной выше, выражение для  $L_i(x)$  и выражение для  $P_n(x)$  требуют очень большого числа умножений и стало быть очень неудобны. Можно преобразовать выражение для  $P_n(x)$  следующим образом. Прежде всего очевидно, что если взять  $f_i$ , равные 1, то  $P_n(x) = 1$ .

Следовательно,

$$1 = \sum_i L_i(x),$$

и

$$P_n(x) = \frac{f_0 L_0(x) + f_1 L_1(x) + \dots + f_n L_n(x)}{L_0(x) + L_1(x) + \dots + L_n(x)}.$$

Разделим числитель и знаменатель на

$$(x - x_0)(x - x_1) \dots (x - x_n).$$

Получим

$$P_n(x) = \frac{\sum_i \frac{f_i A_i}{x - x_i}}{\sum_i \frac{A_i}{x - x_i}},$$

где

$$A_i = \frac{1}{(a_i - a_0) \dots (a_i - a_{i-1})(a_i - a_{i+1}) \dots (a_i - a_n)}.$$

$A_i$  имеют сложный вид, но зависят только от  $a_i$ . Значит, их можно вычислить один раз для всех задаваемых точек. Эта формула называется *барицентрической формулой*.

**У п р а ж н е н и е 7.** Выписать барицентрическую формулу для

$$n = 1, a_0 = 0, a_1 = 1.$$

**7.1. Формула Ньютона.** Мы приведем другое выражение для  $P_n(x)$ .

Очевидно, можно записать

$$P_n(x) = f_n + (x - a_n) Q_{n-1}(x),$$

где  $Q_{n-1}(x)$  представляет собой многочлен степени  $n - 1$ , удовлетворяющий условию

$$Q_{n-1}(a_i) = \frac{f_i - f_n}{a_i - a_n}, \quad i = 0, 1, \dots, n - 1.$$

В самом деле, нетрудно убедиться в том, что

$$P_n(a_n) = f_n, \quad P_n(a_i) = f_i, \quad i = 0, 1, \dots, n - 1.$$

Таким образом, нахождение  $Q_{n-1}(x)$  сводится к той же задаче, что и в п. 5.1, но на одну точку меньше, и  $n$  заменяется на  $n - 1$ . Продолжая таким образом, придем к  $n = 0$ , т. е. к случаю, когда решение очевидно.

Итак, получаем для многочлена  $P_n(x)$  выражение, которое мы запишем в виде

$$f_n + (x - a_n) g_{n-1} + (x - a_n)(x - a_{n-1}) g_{n-2} + \dots \\ \dots + (x - a_n)(x - a_{n-1}) \dots (x - a_1) g_0.$$

Мы не выписываем в явном виде значения коэффициентов  $g_i$ . Формула, которая должна получиться, если написать в явном виде  $g_i$ , называется *формулой Ньютона*.

**7.2. Изменение числа точек.** Легко видеть, что член

$$(x - a_n)(x - a_{n-1}) \dots (x - a_1) g_0$$

обращается в нуль при

$$x = a_i, \quad i = 1, 2, \dots, n.$$

Следовательно, многочлен

$$f_n + (x - a_n) g_{n-1} + \dots + (x - a_n) \dots (x - a_2) g_1$$

степени не выше  $n - 1$  принимает при  $x = a_i$  значения  $f_i$ ,  $i = 1, 2, \dots, n$ .

Стало быть формула Ньютона обладает тем свойством (удобным для использования), что она позволяет легко находить формулы, используя больше или меньше точек (однако убирать точки можно лишь в строго определенном порядке).

**У п р а ж н е н и е 8.** Выписать формулу Ньютона для

$$a_0 = 0, \quad a_1 = 1, \quad a_2 = 2,$$

$$f_0 = 1, \quad f_1 = 2, \quad f_2 = 0.$$

**У п р а ж н е н и е 9.** Какой член следует прибавить к многочлену  $P_1(x) = 3x + 2$ , чтобы не изменить ни  $P_1(0)$ , ни  $P_1(1)$ , и получить  $P_2(2) = 7$ ?

**7.3. Использование формулы Ньютона для вычислений.**

Формула Ньютона очень хорошо приспособлена для численных оценок, когда известны  $g_i$ . В самом деле, можно записать

$$P_n(x) = (\dots ((g_0(x - a_1) + g_1)(x - a_2) + g_2) \\ (x - a_3) + \dots + g_{n-1})(x - a_n) + f_n$$

Это есть обобщение схемы Горнера и вычисление  $P_n(x)$  требует  $n$  умножений и  $2n$  сложений или вычитаний.

**8.1. Случай равноотстоящих точек.** Мы не будем касаться вычисления коэффициентов формулы Ньютона в общем случае. Мы сделаем это для случая равноотстоящих точек.

**8.2. Восходящие разности.** Рассмотрим значения  $f_0, f_1, \dots, f_n$ . Положим

$$\Delta f_i = f_{i+1} - f_i, \quad i = 0, \dots, n-1,$$

$$\Delta^2 f_i = \Delta f_{i+1} - \Delta f_i, \quad i = 0, \dots, n-2,$$

$$\dots \dots \dots \dots \dots \dots \dots \dots \dots \dots$$

$$\Delta^p f_i = \Delta^{p-1} f_{i+1} - \Delta^{p-1} f_i, \quad i = 0, \dots, n-p.$$

Расположим эти значения в таблицу.

Пример.

$i$	$f$	$\Delta f$	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$
0	5	-2	1	+2	-5
1	3	-1	3	-3	
2	2	2	0		
3	4	2			
4	6				

У п р а ж н е н и е 10. До какого порядка можно вычислять разности, если располагать  $n + 1$  значениями?

8.3. Формула Ньютона — Грегори. Многочлен  $P_n(x)$ , соответствующий точкам

$$a_i = a_0 + ih, \quad i = 0, \dots, n,$$

записывается в виде формулы Ньютона — Грегори:

$$f_0 + \frac{x - a_0}{1!h} \Delta f_0 + \frac{(x - a_0)(x - a_1)}{2!h^2} \Delta^2 f_0 + \dots$$

$$\dots + \frac{(x - a_0)(x - a_1) \dots (x - a_{n-1})}{n!h^n} \Delta^n f_0.$$

Доказательство этого утверждения проведем индукцией по  $n$ . При  $n = 0$  формула очевидна. Допустим, что она доказана для  $n - 1$ , и докажем, что она верна для  $n$ . Формула справедлива для  $x = a_0, \dots, x = a_{n-1}$ , поскольку для этих значений последний член обращается в нуль и многочлен, лишенный этого последнего члена, есть многочлен степени  $n - 1$ , принимающий значения

$$P_{n-1}(a_i) = f_i, \quad i = 0, 1, \dots, n-1.$$

Остается, таким образом, исследовать случай  $x = a_n$ .

Нетрудно убедиться в том, что

$$f_n = f_0 + \frac{n}{1!} \Delta f_0 + \frac{n(n-1)}{2!} \Delta^2 f_0 + \dots + \Delta^n f_0,$$

то есть

$$f_n = \sum_{i=0}^n C_n^i \Delta^i f_0.$$

Но

$$f_{n-1} = \sum_{i=0}^{n-1} C_{n-1}^i \Delta^i f_0, \quad \Delta f_{n-1} = \sum_{j=0}^{n-1} C_n^j \Delta^{j+1} f_0.$$

Отсюда

$$f_n = f_{n-1} + \Delta f_{n-1} = \sum_{i=0}^n (C_{n-1}^i + C_{n-1}^{i-1}) \Delta^i f_0 = \sum_{i=0}^n C_n^i \Delta^i f_0.$$

У п р а ж н е н и е 11. а) Решить снова упражнение 8, используя разности.

б) Можно ли сделать так, чтобы получить в точности результат упражнения 8?

**9.1. Интерполяция.** Рассмотрим функцию  $f(x)$ , удовлетворяющую условиям

$$f(a_i) = f_i, \quad i = 0, 1, \dots, n.$$

Можно ожидать, что многочлен  $P_n(x)$ , удовлетворяющий условию  $P_n(a_i) = f_i$ , будет хорошим приближением для  $f(x)$ , по крайней мере если не слишком удаляться от интервала, содержащего все  $a_i$ .

Будем называть этот многочлен *интерполяционным многочленом* функции  $f(x)$  в точках  $a_i$ .

**9.2. Поправка интерполяции.** Мы хотим вычислить поправку

$$\varphi_f(a) = f(x) - P_n(x).$$

Функция

$$g(x) = f(x) - P_n(x) - K(x - a_0)(x - a_1) \dots (x - a_n),$$

где  $K$  выбрано так, чтобы  $g(x_0) = 0$ , обращается в нуль в  $a_0, a_1, \dots, a_n, x_0$ , и значит, ее  $(n+1)$ -я производная обращается в нуль для некоторого значения  $\xi$  из наименьшего интервала, содержащего  $a_i$  и  $x_0$ :

$$f^{(n+1)}(\xi) = K(n+1)!,$$

и следовательно,

$$f(x_0) = P_n(x_0) + \frac{f^{(n+1)}(\xi)}{(n+1)!} (x_0 - a_0) \dots (x_0 - a_n),$$

$$\varphi_f(x_0) = (x_0 - a_0) \dots (x_0 - a_n) \frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

**9.3. Производные и разности.** Мы покажем, что для функции,  $n$  раз дифференцируемой,

$$\Delta^n f_0 = h^n f^{(n)}(\xi), \quad \xi \in [a_0, a_n].$$

Докажем это утверждение индукцией по  $n$ . Свойство выполняется для  $n = 1$ . Допустим, что оно верно для  $n$ . Тогда можно записать

$$\Delta^n (f(x+h) - f(x))_0 = h^n (f^{(n)}(\xi+h) - f^{(n)}(\xi)),$$

$$\xi \in [a_0, a_n].$$

Но левая часть записывается в виде

$$\Delta^{n+1} f_1 - \Delta^n f_0 = \Delta^{n+1} f_0,$$

а правая — в виде

$$h^{n+1} f^{(n+1)}(\xi'), \quad \xi' \in [a_0, a_{n+1}],$$

что и доказывает наше утверждение.

**10.1. Формула Тейлора с остаточным членом в интегральной форме.** Мы приведем другое выражение для  $\gamma_f(x)$ . Нам понадобится формула Тейлора с остаточным членом в интегральной форме:

$$f(x) = f(a) + \frac{x-a}{1!} f'(a) + \dots$$

$$\dots + \frac{(x-a)^n}{n!} f^{(n)}(a) + \int_a^x \frac{(x-t)^n}{n!} f^{(n+1)}(t) dt.$$

Эта классическая формула доказывается индукцией по  $n$ . Для  $n = 0$  она верна. Допустим, что она верна для  $n$ . Интегрированием по частям можно получить соотношения

$$\int_a^x \frac{(x-t)^n}{n!} f^{(n+1)}(t) dt =$$

$$= \left[ -\frac{(x-t)^{n+1}}{(n+1)!} f^{(n+1)}(t) \right]_a^x + \int_a^x \frac{(x-t)^{n+1}}{(n+1)!} f^{(n+2)}(t) dt =$$

$$= \frac{(x-a)^{n+1}}{(n+1)!} f^{(n+1)}(a) + \int_a^x \frac{(x-t)^{n+1}}{(n+1)!} f^{(n+2)}(t) dt,$$

которое и доказывает формулу для  $n+1$ .

Запишем эту формулу в несколько ином виде:

$$f(x) = f(a) + \frac{(x-a)}{1!} f'(a) + \dots + \frac{(x-a)^n}{n!} f^{(n)}(a) +$$

$$+ \int_{-\infty}^{+\infty} H_n(x, t, a) f^{(n+1)}(t) dt,$$

где

$$H_n(x, t, a) = \frac{(x-t)^n}{n!} (U(x, t) - U(a, t)),$$

и

$$U(\lambda, t) = \begin{cases} 1 & \text{для } t \leq \lambda, \\ 0 & \text{для } t > \lambda. \end{cases}$$

**10.2. Интегральное выражение поправки интерполяции.** Интерполяционные формулы линейно зависят от функции, поэтому ясно, что

$$\gamma_{(f_1+f_2)}(x) = \gamma_{f_1}(x) + \gamma_{f_2}(x).$$

Для многочлена  $Q$  степени не выше  $n$  имеем

$$\gamma_Q(x) = 0.$$

Следовательно, заметив, что

$$\int_{-\infty}^{+\infty} \frac{(x-t)^n}{n!} U(a, t) f^{(n+1)}(t) dt$$

есть многочлен от  $x$  степени  $n$ , и положив

$$\varphi(x) = \int_{-\infty}^{+\infty} \frac{(x-t)^n}{n!} U(a, t) f^{(n+1)}(t) dt,$$

можем написать:

$$\gamma_f(x) = \gamma_\varphi(x).$$

Положим еще

$$K_n(x, t) = \gamma_{\frac{(x-t)^n}{n!} U(x, t)}(x).$$

$K_n(x, t)$  называется *ядром интерполяционной формулы*.

Легко видеть, что ядро есть функция от  $t$ , обращающаяся в нуль вне наименьшего интервала, содержащего точки  $a_i$  и точку  $x$ .

Применяя формулу Лагранжа, получаем

$$\begin{aligned} \gamma_\varphi(x) &= \sum_{i=0}^n L_i(x) \varphi(a_i) - \varphi(x) = \\ &= \int_{-\infty}^{+\infty} \left[ \sum_{i=0}^n L_i(x) \frac{(a_i-t)^n}{n!} U(a_i, t) - \frac{(x-t)^n}{n!} U(x, t) \right] \times \\ &\quad \times f^{(n+1)}(t) dt. \end{aligned}$$

Выражение в скобках есть не что иное, как

$$K_n(x, t).$$

Отсюда

$$\gamma_f(x) = \int_{-\infty}^{+\infty} K_n(x, t) f^{(n+1)}(t) dt.$$

**З а д а ч а 3.** а) Проверить, что  $K_n(x, t)$  обращается в нуль вне наименьшего интервала, содержащего  $a_i$  и  $x$ .

б) Указать верхнюю грань ошибки линейной интерполяции.

с) Применить к таблице десятичных логарифмов с 8 десятичными знаками для  $x \geq 1000$  и с шагом длины 1.

### III. КВАДРАТУРНЫЕ ФОРМУЛЫ

**11.1. Основной принцип методов.** Для приближенного вычисления интеграла

$$I = \int_{\alpha}^{\beta} a(t) f(t) dt$$

рассматривается приближающий многочлен  $P_n(t)$ . В качестве приближенного значения интеграла  $I$  берется интеграл

$$\int_{\alpha}^{\beta} a(t) P_n(t) dt.$$

Часто будет удобно выбирать в качестве  $P_n(t)$  интерполяционный многочлен.

**11.2. Использование формы Лагранжа.** Возьмем многочлен

$$P_n(t) = \sum_{i=0}^n L_i(t) f_i;$$

тогда

$$\int_{\alpha}^{\beta} a(t) P_n(t) dt = \sum_{i=0}^n f_i \int_{\alpha}^{\beta} a(t) L_i(t) dt = \sum_{i=0}^n f_i A_i.$$

Коэффициенты  $A_i$  называются *весами* квадратурной формулы. Они зависят от функции  $a(t)$  и от узлов  $\alpha$ ,  $\beta$ ,  $a_i$ , но не зависят от функции  $f(t)$ . Когда эти коэффициенты

известны, приближенное вычисление интеграла становится очень простым.

**11.3. Практическое нахождение весов.** Коэффициенты  $A_i$  можно находить при помощи метода неопределенных коэффициентов, записав, что формула является точной для многочленов

$$1, x, \dots, x^n.$$

Пусть, например, требуется найти приближенную квадратуру для  $\int_{-1}^1 f(x) dx$ , используя узлы интерполяции  $-1, 0, 2$ .

Имеем:

$$\text{для } 1 \quad A_{-1} + A_0 + A_2 = 2,$$

$$\text{для } x \quad -A_{-1} + 2A_2 = 0,$$

$$\text{для } x^2 \quad A_{-1} + 4A_2 = 2/3.$$

Отсюда сразу же получаем  $A_{-1} = 2A_2$ ,  $6A_2 = 2/3$ ,  $A_2 = 1/9$ ,  $A_{-1} = 2/9$ ,  $A_0 = 15/9$ .

**У п р а ж н е н и е 13.** Найти приближенную квадратурную формулу для

$$\int_{-1}^1 xf(x) dx$$

с узлами  $-1, 0, 1$ .

**12.1. Сдвиг для  $a(t) = 1$ .** Ясно, что если формула

$$\int_{\alpha}^{\beta} P_n(x) dx = \sum_{i=0}^n A_i P_n(a_i)$$

является точной для любого многочлена степени  $n$ , то это будет так и для формулы

$$\int_{\alpha}^{\beta} P_n(x + \lambda) dx = \sum_{i=0}^n A_i P_n(a_i + \lambda),$$

поскольку  $Q_n(x) = P_n(x + \lambda)$  есть снова многочлен степени  $n$ .

Положим  $y = x + \lambda$ ; тогда

$$\int_{\alpha+\lambda}^{\beta+\lambda} P_n(y) dy = \sum_{i=0}^n A_i P_n(a_i + \lambda).$$

Таким образом, веса инвариантны относительно общего сдвига узлов  $\alpha, \beta, a_i$ .

**12.2. Гомотетия для  $a(t) = 1$ .** Ясно, что если формула

$$\int_{\alpha}^{\beta} P_n(x) dx = \sum_{i=0}^n A_i P_n(a_i)$$

является точной для любого многочлена степени  $n$ , то такова будет и формула

$$\int_{\alpha}^{\beta} P_n(\lambda x) dx = \sum_{i=0}^n A_i P_n(\lambda a_i),$$

поскольку  $Q_n(x) = P_n(\lambda x)$  снова есть многочлен степени  $n$ .

Положим  $y = \lambda x$ ; получим

$$\frac{1}{\lambda} \int_{\lambda\alpha}^{\lambda\beta} P_n(y) dy = \sum_{i=0}^n A_i P_n(\lambda a_i).$$

Следовательно, гомотетия с коэффициентом  $\lambda$ , примененная к  $\alpha, \beta, a_i$ , умножает веса на  $\lambda$ .

**12.3. Симметрия для  $a(t) = 1$**  Рассмотрим, в частности, для  $a(t) = 1$  такую формулу, чтобы  $\alpha = -\beta$ :

$$a_i = -a_{n-i}.$$

Применяя результат предыдущего пункта с  $\lambda = -1$ , получаем

$$A_i = A_{n-i}.$$

**13.1. Симметричная формула, имеющая нечетное число точек для  $a(t) = 1$ .** Пусть имеется симметричная формула с  $2n + 1$  точками. Она является точной для любого многочлена степени  $2n$ . Но

$$\int_{-\alpha}^{\alpha} x^{2n+1} dx = 0,$$

$$\sum_{i=0}^n A_i x_i^{2n+1} = 0,$$

и значит, формула является точной для многочленов степени  $2n + 1$  (хотя она и имеет всего  $2n + 1$  точек).

**14.1. Выражение для поправки квадратурной формулы.**  
 Поправка квадратурной формулы имеет вид

$$\int_{\alpha}^{\beta} a(t) f(t) dt - \int_{\alpha}^{\beta} a(t) P_n(t) dt = \int_{\alpha}^{\beta} a(t) \gamma_f(t) dt = \\ = \int_{\alpha}^{\beta} a(t) \left( \int_{-\infty}^{+\infty} K(t, u) f^{(n+1)}(u) du \right) dt.$$

Это можно, меняя порядок интегрирования и полагая

$$Q(u) = \int_{\alpha}^{\beta} K(t, u) a(t) dt,$$

записать еще и в другом виде:

$$\int_{-\infty}^{+\infty} f^{(n+1)}(u) \left( \int_{\alpha}^{\beta} K(t, u) a(t) dt \right) du = \int_{-\infty}^{+\infty} f^{(n+1)}(u) Q(u) du;$$

$Q(u)$  есть *ядро квадратурной формулы*.

**У п р а ж н е н и е 14.** Показать, что ядро обращается в нуль вне наименьшего интервала, содержащего  $a_i$ ,  $\alpha$  и  $\beta$ .

**14.2. Случай определенных формул.** Когда  $Q(u)$  имеет постоянный знак, формула называется *определенной*. Этот случай встречается довольно часто.

В этом случае можно оценить сверху абсолютное значение погрешности квадратуры посредством

$$M_{n+1} K,$$

где  $M_{n+1}$  — верхняя грань  $|f^{(n+1)}(\xi)|$  на наименьшем интервале, содержащем  $a_i$ ,  $\alpha$  и  $\beta$ ,

$$K = \left| \int_{-\infty}^{+\infty} Q(u) du \right|.$$

При этом  $K$  очень легко найти; достаточно взять абсолютное значение относительной погрешности для

$$\frac{x^{n+1}}{(n+1)!},$$

т. е. функцию,  $(n+1)$ -я производная которой равна 1.

Теперь мы приведем несколько конкретных формул. Мы в каждом случае будем брать интервал, дающий самую простую формулу.

**15.1. Формулы нулевой степени.** Взяв многочлен степени 0, находим в качестве приближения интеграла

$$\int_{\alpha}^{\beta} f(t) du$$

выражение

$$f(a_0)(\beta - \alpha).$$

Для  $\alpha \leq a_0 \leq \beta$  погрешность квадратурной формулы записывается в виде

$$\int_{\alpha}^{\beta} Q(u) f'(u) du,$$

где

$$Q(u) = \begin{cases} u - \alpha & \text{для } \alpha < u < a_0, \\ u - \beta & \text{для } a_0 < u < \beta. \end{cases}$$

**У п р а ж н е н и е 15.** Доказать эту формулу, преобразуя

$$\int_{\alpha}^{\beta} Q(u) f'(u) du.$$

**15.2. Формула Понселе.** Возьмем снова рассмотренный выше пример с

$$\alpha = -\frac{1}{2}, \quad \beta = \frac{1}{2}, \quad a_0 = \frac{\alpha + \beta}{2}, \quad \text{т. е. } a_0 = 0.$$

Формула симметрична в нечетном числе точек. Значит, она справедлива для многочленов первой степени. Мы получим ее другим способом. Заметим, что в этом случае

$$\int_{-1/2}^{1/2} Q(u) du = 0.$$

Стало быть, полагая

$$P(u) = - \int_{-1/2}^u Q(v) dv,$$

можно преобразовать выражение погрешности

$$\begin{aligned} \int_{-1/2}^{1/2} Q'(u) f'(u) du &= [-Q(u) f'(u)]_{-1/2}^{1/2} + \\ &+ \int_{-1/2}^{1/2} P(u) f''(u) du = \int_{-1/2}^{1/2} P(u) f''(u) du. \end{aligned}$$

Следовательно, получаем *формулу Понселе*, которая является точной для многочленов степени 1:  $f(0)$ .

Легко видеть, что она определена относительно второй производной.

Находим  $K$ , применяя формулу к  $x^2/2$ , т. е.

$$K = \int_{-1/2}^{1/2} \frac{x^2}{2} dx = \frac{1}{24}.$$

**15.3. Формула трапеций.** Возьмем

$$\alpha = 0, \quad \beta = 1, \quad a_0 = 0, \quad a_1 = 1.$$

Получаем формулу, справедливую для многочленов степени 1:

$$\frac{f(0) + f(1)}{2}.$$

Она называется *формулой трапеций*. Легко видеть, что она определена. Применяя эту формулу, получим  $K = -1/12$ .

Если измерять точность формулы значением  $K$ , то можно утверждать, что она менее точная, чем формула Понселе (примененная к тому же интервалу). Если  $f''(a)$  сохраняет на интервале постоянный знак, то погрешности будут сохранять противоположные знаки (значит, получим интервал приближения для точного значения).

У п р а ж н е н и е 16. Вычислить

$$\int_0^1 e^x dx$$

- a) при помощи формулы Понселе;
- b) при помощи формулы трапеций;
- c) точно;
- d) что можно сказать об этих трех значениях?

**16.1. Формула Симпсона.** Рассмотрим

$$\alpha = -1, \quad \beta = 1, \quad a_0 = -1, \quad a_1 = 0, \quad a_2 = 1.$$

Найдем веса путем отождествления правых и левых частей. Для  $f = 1$  должно быть

$$A_{-1} + A_0 + A_1 = 2;$$

для  $f = x$  должно быть  $A_{-1} = A_1$ ;

для  $f = x^2$  должно быть  $A_{-1} + A_1 = 2/3$ .

Отсюда  $A_{-1} = A_1 = 1/3$ ,  $A_0 = 4/3$ , и приближенное значение интеграла равно

$$\frac{1}{3}(f(-1) + 4f(0) + f(1)).$$

Это *формула Симпсона*. Она симметрична в нечетном числе точек, и значит, справедлива для многочленов вплоть до 3-го порядка.

Можно показать, что она определена относительно 4-й производной. Коэффициент  $K$  можно получить, если взять функцию  $x^4/24$ . Находим

$$K = \frac{2}{3 \cdot 24} - \frac{2}{120} = \frac{1}{90}.$$

**У п р а ж н е н и е 17.** Получить вновь формулу Симпсона, комбинируя формулу Понселе и формулу трапеций.

**17.1. Практическое применение формул.** Как правило, приведенные выше формулы не применяются непосредственно на всем отрезке интегрирования  $[a, b]$ . Чтобы вычислить приближенное значение интеграла

$$\int_a^b f(x) dx,$$

разделим отрезок  $[a, b]$  на  $n$  частичных интервалов промежуточными точками  $a_1, \dots, a_{n-1}$ , и применим к каждому из интервалов одну из приведенных выше формул.

**17.2. Случай формулы трапеций.** Запишем формулу трапеций, полагая  $l = b - a$  и принимая частичные интервалы равными по длине  $h = (b - a)/n$ :

$$h \left[ \frac{f(a)}{2} + f(a_1) + \dots + f(a_{n-1}) + \frac{f(b)}{2} \right].$$

Абсолютное значение погрешности оценивается сверху посредством формулы

$$\frac{nh^3 M_2}{12} = \frac{lh^2}{12} M_2,$$

где  $M_2$  — верхняя грань  $|f''|$  на  $[a, b]$ .

**17.3. Случай формулы Симпсона.** Разделим отрезок  $[a, b]$  на  $2n$  частичных равных интервалов точками

$$a_0 = a, \quad a_1, a_2, \dots, a_{2n-1}, \quad a_{2n} = b.$$

Формула Симпсона записывается в виде

$$\frac{h}{3} \left[ \frac{f(a_0)}{2} + 2f(a_1) + f(a_2) + 2f(a_3) + \dots \right. \\ \left. \dots + f(a_{2n-2}) + 2f(a_{2n-1}) + \frac{f_1(a_{2n})}{2} \right].$$

Погрешность оценивается посредством формулы

$$\frac{nh^5}{90} M_4 = \frac{lh^4}{180} M_4,$$

где  $M_4$  — верхняя грань  $|f^{(4)}|$  на отрезке  $[a, b]$ .

**З а д а ч а 4.** а) Убедиться в том, что погрешность формулы трапеций на отрезке  $[0, 1]$  может быть записана в виде

$$\int_0^1 Q(u) f''(u) du, \quad Q(x) = \frac{x(1-x)}{2}.$$

б) Вывести отсюда, что формула является определенной.

с) Вновь найти значение  $K$ .

д) Доказать формулу погрешности из п. 17.2.

**З а д а ч а 5.** а) Показать, что формула Симпсона на отрезке  $[-1, +1]$  имеет в качестве ядра относительно  $f^{(4)}$

$$\frac{(x+1)^3(1-3x)}{72} \quad \text{для} \quad -1 \leq x \leq 0,$$

$$\frac{(1-x)^3(1+3x)}{72} \quad \text{для} \quad 0 \leq x \leq 1.$$

б) Показать, что ядро имеет фиксированный знак.

с) Найти вновь значение  $K$ .

д) Вновь получить формулу погрешности из п. 17.3.

#### IV. ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ С НАЧАЛЬНЫМИ УСЛОВИЯМИ

**18.1. Задача с начальными условиями.** Рассмотрим дифференциальное уравнение

$$y' = Y(y, t). \quad (1)$$

С учетом условий регулярности, которые мы предполагаем выполненными, оно имеет на отрезке  $[t_0, t_0 + H]$  единственное решение, удовлетворяющее условию

$$y(t_0) = y_0. \quad (2)$$

Задача с начальными условиями состоит в отыскании функций, удовлетворяющих одновременно условиям (1) и (2).

Мы определим приближенные решения этой задачи.

**19.1. Метод касательной\*).** Разделим отрезок  $[t_0, t_0 + H]$  на частичные интервалы длины  $h$ .

На каждом из них заменим

$$y(t+h) \text{ на } y(t) + hy'(t).$$

Таким образом, получаем алгоритм

$$y_{i+1} = y_i + hY_i, \quad Y_i = Y(y_i, t_i).$$

Это — метод касательной.

**19.2. Сходимость.** Можно соединить точки  $(t_{i-1}, y_{i-1})$ ,  $(t_i, y_i)$  отрезком прямой и показать, что (при очень широких условиях) предел этой ломаной при  $h$ , стремящемся к нулю, будет решением задачи с начальными условиями. В таком случае говорят, что метод является *сходящимся*.

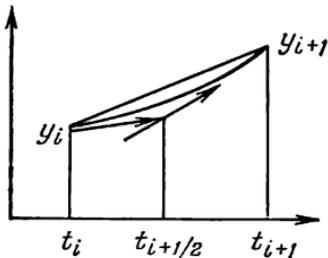


Рис. 3.

**19.3. Улучшенный метод касательной.** Метод касательной мало эффективен.

Мы получим лучшие результаты, если заменим  $Y_i$  на  $Y_{i+1/2}$  (рис. 3). Чтобы получить это

значение, вычислим

$$y_{i+1/2} = y_i + \frac{h}{2} Y_i$$

(т. е. применим метод касательной с шагом  $h/2$ , а затем положим

$$Y_{i+1/2} = Y(y_{i+1/2}, t_i + h/2)$$

и наконец,

$$y_{i+1} = y_i + hY_{i+1/2}.$$

Это есть *улучшенный метод касательной*.

**У п р а ж н е н и е 18.** а) Применить метод касательной к задаче  $y' = y$ ,  $y(0) = 1$ , взяв  $h = 1/n$ .

\*) Этот метод называется чаще *методом ломаных Эйлера*. (Прим. ред.)

б) Чему равны: значение, полученное после  $n$ -го шага, его предел, когда  $n$  стремится к бесконечности?

с) Провести то же исследование, используя улучшенный метод касательной.

д) Сравнить результаты этих двух методов при равной работе (считая, что один шаг улучшенного метода стоит в два раза дороже, чем один шаг в методе касательной).

**20.1. Неявный метод Адамса порядка 1.** Для решения той же задачи с начальными условиями мы можем выписать строго

$$y(t_{i+1}) = y(t_i) + \int_{t_i}^{t_{i+1}} Y(y, t) dt.$$

Интеграл, фигурирующий в этой формуле, можно вычислить методом приближенной квадратуры. Например, взяв в качестве приближения для

$$\int_{t_i}^{t_{i+h}} Y(y, t) dt \text{ значение } hY(y_i, t_i),$$

придем к формуле касательной.

Взяв теперь формулу трапеций, получим

$$y_{i+1} - y_i = \frac{h}{2} (Y_{i+1} + Y_i).$$

Эта формула, к сожалению, является неявной, так как  $y_{i+1}$  фигурирует в  $Y_{i+1}$ .

Она называется *неявной формулой Адамса порядка 1*.

**У п р а ж н е н и е 19.** а) Показать, что применение неявной формулы Адамса удобно для линейного уравнения.

б) Исследовать уравнение  $y' = y$  этим методом, следуя тем же путем, что и в упражнении 18.

**20.2. Явный метод Адамса порядка 1.** Ни один из приведенных выше способов вычисления интеграла

$$\int_{t_i}^{t_{i+1}} Y(y, t) dt$$

не представляет особого интереса (кроме как для уравнений специального вида).

Мы получим лучшие результаты, если обратимся к формуле вычисления этого интеграла при помощи переменных

$$t_{i-1} \text{ и } t_i.$$

Находим

$$y_{i+1} - y_i = \frac{h}{2}(3Y_i - Y_{i-1}).$$

Отсюда

$$y_{i+1} = y_i + \frac{h}{2}(3Y_i - Y_{i-1}).$$

Это явная формула Адамса порядка 1.

Заметим, что этот метод требует знания  $y_0$  и  $y_1$ , чтобы иметь возможность применять формулу. Значение  $y_1$  находится, например, разложением в ряд.

У п р а ж н е н и е 20. Доказать приведенную выше квадратурную формулу.

21.1. Метод Нистрема. Если записать

$$y(t_{i+1}) = y(t_{i-1}) + \int_{t_{i-1}}^{t_{i+1}} Y(y, t) dt$$

и вычислить интеграл по формуле Понселе, то придем к алгоритму

$$y_{i+1} = y_{i-1} + 2hY_i.$$

Он известен под названием *метода Нистрема*.

21.2. Неустойчивость. Применим метод Нистрема к уравнению

$$y' = -y, \quad y(0) = 1$$

с  $h = 0,2$ ,  $y(0,2) = 0,8$ .

Находим

$y(0,4) = 0,68$	$y(1,8) = 0,1077$
$y(0,6) = 0,5280$	$y(2,0) = 0,2065$
$y(0,8) = 0,4688$	$y(2,2) = 0,0251$
$y(1,0) = 0,3405$	$y(2,4) = 0,1965$
$y(1,2) = 0,3326$	$y(2,6) = -0,0535$
$y(1,4) = 0,2075$	$y(2,8) = 0,2179$
$y(1,6) = 0,2496$	$y(3,0) = -0,1407$

Как нетрудно видеть, результаты вычислений становятся все более и более неправильными с расходящимся колебанием, период которого равен двум шагам.

Говорят, что в этом случае имеет место *неустойчивость*.

З а д а ч а 6. а) Применить явный метод Адамса к уравнению  $y' = y$ .

б) Показать, что соотношение между  $y_{i+1}$ ,  $y_i$ ,  $y_{i-1}$  обладает общим решением вида

$$y_i = Ar_1^i + Br_2^i.$$

Найти  $r_1$ ,  $r_2$ ,  $A$ ,  $B$  для  $y_1 = 1 + h$ .

с) Показать, что на любом интервале имеет место сходимость.

д) Провести то же исследование для уравнения  $y' = -y$  и метода Нистрема.

е) Как можно объяснить неустойчивость?

ф) Имеется ли сходимость на любом интервале?

#### РЕШЕНИЯ УПРАЖНЕНИЙ ГЛАВЫ IV

1) а), б) 3 сложения, 3 умножения, 3 деления.

2)  $4 \frac{10^6}{3} 10^{-5} \approx 14$  секунд.

3) а) Уравнения, следующие в порядке 1, 2, 3, 4, 5, приводятся последовательно к порядкам 5, 2, 3, 4, 1 | 5, 1, 3, 4, 2 | 5, 1, 2, 4, 3 | 5, 1, 2, 3, 4. Заметим, что не имеется никакой арифметической операции для перехода к треугольной форме.

б) Приспособить непосредственно этот порядок.

4) а) Линейная система с произвольными правыми частями является либо системой Крамера, либо, в зависимости от правых частей, несоместной или неопределенной.

б) Определитель отличен от нуля (определитель Вандермонда).

5) а)  $\frac{x - a_1}{a_0 - a_1}, \frac{x - a_0}{a_1 - a_0},$

$$P_n(x) = \frac{f_0(x - a_1)}{a_0 - a_1} + \frac{f_1(x - a_0)}{a_1 - a_0} = \frac{x(f_1 - f_0) + f_0a_1 - f_1a_0}{a_1 - a_0}.$$

б)  $\frac{(x - a_1)(x - a_2)}{(a_0 - a_1)(a_0 - a_2)}, \frac{(x - a_0)(x - a_1)}{(a_1 - a_0)(a_1 - a_2)}, \frac{(x - a_0)(x - a_1)}{(a_2 - a_0)(a_2 - a_1)},$

$$P_n(x) = f_0 \frac{(x - a_1)(x - a_2)}{(a_0 - a_1)(a_0 - a_2)} + f_1 \frac{(x - a_0)(x - a_2)}{(a_1 - a_0)(a_1 - a_2)} + f_2 \frac{(x - a_0)(x - a_1)}{(a_2 - a_0)(a_2 - a_1)}.$$

6) Произведение матрицы коэффициентов  $L_i(x)$  на матрицу из  $a_i^j = 1$  равно 1. Значит, одна из матриц обратна другой.

$$7) \frac{\frac{f_0}{x - x_0} - \frac{f_1}{x - x_1}}{\frac{1}{x - x_0} - \frac{1}{x - x_1}}.$$

$$8) -2(x-2) = \frac{3}{2}(x-2)(x-1).$$

$$9) kx(x-1) + P_1(x) = P_2(x), k = -1/2.$$

10) Порядок  $n$ .

11) а)

$$\begin{array}{ccc} f & \Delta & \Delta^2 \\ 1 & 1 & -3 \\ 2 & -2 & \\ 0 & & \end{array} \quad P_2 = 1 + \frac{x}{1!} - 3 \frac{x(x-1)}{2!}.$$

б)

$$x = 2 - y, \quad \begin{array}{ccc} f & \Delta & \Delta^2 \\ 0 & 2 & -3 \\ 2 & -1 & \\ 1 & & \end{array},$$

$$P_2 = 2y - \frac{3}{2}y(y-1) = 2(2-x) - \frac{3}{2}(2-x)(1-x).$$

$$12) (x-a_0)(x-a_1) \dots (x-a_n).$$

$$13) A_{-1} + A_0 + A_1 = 0, \quad -A_{-1} + A_1 = 2/3, \quad A_{-1} + A_1 = 0,$$

откуда  $A_1 = 1/3, A_{-1} = -1/3, A_0 = 0$ .

14) Для  $u$ , внешнего к этому интервалу,  $K(t, u)$  равно нулю.

15)

$$\begin{aligned} \int_{\alpha}^{\beta} Q(u) f'(u) du &= \int_{\alpha}^{\alpha_0} (u-a) f'(u) du + \int_{\alpha_0}^{\beta} (u-\beta) f'(u) du = \\ &= [(u-a) f(u)]_{\alpha}^{\alpha_0} - \int_{\alpha}^{\alpha_0} f(u) du + [(u-\beta) f(u)]_{\alpha_0}^{\beta} - \\ &= \int_{\alpha_0}^{\beta} f(u) du = (\beta - a) f(\alpha_0) - \int_{\alpha}^{\beta} f(u) du. \end{aligned}$$

16) а) 1,648, б) 1,859, в) 1,718.

д) Первый почти в два раза ближе к точному значению, чем второй.

17) Комбинируем так, чтобы заставить исчезнуть главные члены погрешности:

метод Понселе между  $-1$  и  $+1$ :  $2f(0)$ ,

метод трапеций:  $[f(-1) + f(1)]$ , откуда получаем

$$\frac{1}{3} [f(-1) + 4f(0) + f(1)].$$

18) а)  $y_i = (1+h)^i$ .

б)  $(1+1/n)^n$ , предел которого равен  $e$ .

в)  $y_i = (1+h+h^2/2)^i$ , что снова дает тот же предел  $e$  при  $i = n, n \rightarrow \infty$ .

д) Нужно взять шаг  $1/n$  для метода касательной и  $2/n$  для улучшенного метода касательной, откуда получаем

$$\frac{-e}{2n} \text{ и } \frac{-2e}{n^2}.$$

Второй шаг лучше для  $n > 4$ .

19) а)  $y' = a(t)y + b(t)$ ,

$$y_{i+1} = y_i + \frac{h}{2} (a(t_{i+1})y_{i+1} + a(t_i)y_i + b(t_{i+1}) + b(t_i)),$$

г. е. уравнение первой степени относительно  $y_{i+1}$ .

б)  $y_{i+1} = \left( \frac{1 + h/2}{1 - h/2} \right)^i$ .

Находим для абсциссы 1 главную часть погрешности  $\frac{2e}{3n^2}$ .

20) Пишем, что формула будет точной для  $Y = 1, Y = t$ .

#### РЕШЕНИЯ ЗАДАЧ ГЛАВЫ IV

З а д а ч а 1. Пункт а) остается без изменений. Пункт б) насчитывает  $n(n-1)/2$  сложений и умножений. Пункт с) остается тем же. Итак, всего будет  $n$  делений,  $n(n-1)$  умножений и  $n(n-1)$  сложений.

З а д а ч а 2. а) 
$$\begin{bmatrix} \frac{-70\,000}{3} & \frac{70\,001}{3} \\ -10\,000 & +10\,000 \end{bmatrix}$$

б) Члены обратной матрицы очень велики.

с) Как только обратная матрица начинает содержать очень большие члены, решение становится очень чувствительным к малым изменениям правых частей.

З а д а ч а 3. а) Справа все  $U$  равны нулю, слева — все равны 1

$$\frac{(x-t)^n}{n!} = \sum_{i=0}^n L_i(x) \frac{(a_i-t)^n}{n!}.$$

б)  $\left| \frac{(x-a_0)(a_1-x)}{2!} f''(\xi) \right|;$

$(a_1 - a_0)^2 M_2/2!$ , где  $M_2$  — верхняя грань  $|f''(\xi)|$  на  $[a_0, a_1]$ .

с)  $\frac{0,44}{2x^2}$  для шага 1. Для  $x \geq 1000$  точность меньше, чем  $0,22 \cdot 10^{-6}$ .

З а д а ч а 4. а)

$$\begin{aligned} \int_0^1 \frac{x(1-x)}{2} f''(x) dx &= \left[ \frac{x(1-x)}{2} f'(x) \right]_0^1 + \frac{1}{2} \int_0^1 (2x-1) f'(x) dx = \\ &= \frac{1}{2} [(2x-1) f(x)]_0^1 - \int_0^1 f(x) dx. \end{aligned}$$

б) Очевидно.

$$c) K = \int_0^1 \frac{x(1-x)}{2} dx = \left[ \frac{x^2}{4} - \frac{x^3}{6} \right]_0^1 = \frac{1}{12}.$$

д) Когда от шага 1 переходим к шагу  $h$ ,  $K$  переходит от  $\int_0^1 \frac{x^2}{2} dx$  к  $\int_0^h \frac{x^2}{2} dx$ , т. е. умножается на  $h^3$ .

З а д а ч а 5. а) Интегрируем по частям ядро до получения формулы Симпсона.

б) Ядро, очевидно, положительно.

с) Находим  $1/90$ .

д) Формула из п. 16.1 относительно  $h = 1$ ,  $l = 2$ . Когда переходят от шага 1 к шагу  $h$ , член погрешности умножается на  $h^5$ .

З а д а ч а 6. а)  $y_{i+1} - y_i \left(1 + \frac{3h}{2}\right) + \frac{h}{2} y_{i-1} = 0$ .

$$b) r^2 - \left(1 + 3\frac{h}{2}\right)r + \frac{h}{2} = 0, \quad r_1 \approx 1 + h, \quad r_2 \approx \frac{h}{2},$$

$$A + B = 1, \quad Ar_1 + Br_2 = 1 + h.$$

с) Когда  $h \rightarrow 0$ ,  $A \rightarrow 1$ ,  $B \rightarrow 0$ .

д)  $r^2 + 2rh - 1 = 0$ ,  $r_1 \approx 1 - h$ ,  $r_2 \approx -1 - h$ .

е) Неустойчивость сводится к тому, что  $|-1 - h| > 1$ .

ф) Когда  $h \rightarrow 0$ ,

$$A \rightarrow 1, \quad B \rightarrow 0, \quad r_1^{x/h} \rightarrow e^{-x}, \quad r_2^{x/h} \rightarrow e^x,$$

причем сходимость имеет место на любом интервале (несмотря на неустойчивость).

## 1. КЛАССИФИКАЦИЯ ПОГРЕШНОСТЕЙ

**1.1. Четыре категории погрешностей.** Мы различаем следующие категории погрешностей:

- погрешности методов;
- погрешности округления — вычислительные погрешности;
- погрешности из-за инструментов;
- погрешности, возникающие из-за неточности исходных данных, — неустранимые погрешности.

Мы сразу же подробнее рассмотрим эти типы погрешностей.

**1.2. Погрешности методов.** Эти погрешности возникают из-за замены одного понятия, не поддающегося численной обработке, другим понятием, более податливым для вычислений.

**П р и м е р ы.** Замена

$$\int_{\alpha}^{\beta} f(x) dx \text{ на } \int_{\alpha}^{\beta} P_n(x) dx.$$

Замена

$$y' = Y(y, t) \text{ на } y_{i+1} = y_i + hY_i.$$

Вообще говоря, имеется информация о порядке величины погрешности методов, а в благоприятных случаях имеется интервал приближения или точность приближения.

**П р и м е р.** Для вычисления  $\int_a^b f(x) dx$  можно применить квадратурную формулу трапеций и оценить погрешность метода, если известно  $M_2$  — верхняя грань  $|f''|$  на  $[a, b]$ . В противном случае можно вычислить  $f''(x)$  для нескольких точек и взять наибольшее

из найденных или же определить  $\Delta^2$  и вывести отсюда порядок величины  $f''(x)$ .

Погрешность зависит в общем случае довольно просто от некоторых параметров метода. Можно, изменяя эти параметры, сделать эту зависимость более слабой, ценой, вообще говоря, возрастания объема вычислений.

**1.3. Погрешности округления.** В любом вычислении мы приходим к ограничению числа цифр для записи чисел, с которыми мы работаем, некоторым значением  $n$ . Может оказаться, что некоторые данные задачи не удовлетворяют этому условию. Так, в результате умножения или деления друг на друга двух чисел из  $n$  цифр получаем обычно числа, имеющие более  $n$  цифр. Это приводит к пренебрежению цифр низших разрядов.

Погрешность округления появляется также в результате различных операций (в частности, может играть роль порядок, в котором производятся операции). Оценка этой погрешности очень трудоемка и в достаточно длинных вычислениях часто трудно реализуема. Информация о такой погрешности является неопределенной.

Погрешность округления можно сделать сколь угодно малой, сохраняя достаточное количество цифр в каждом из промежуточных результатов. Это удлиняет продолжительность каждой операции.

Если работают с материалом определенной емкости, то иногда возрастание числа цифр может быть очень дорогостоящим. Чтобы обрабатывать числа с 7 цифрами на машине емкости 6, необходимо разрезать их на 2 части и комбинировать различные части между собой. (Было бы немногим дороже работать с числами из 12 цифр.)

**У п р а ж н е н и е 1.** Мы хотим вычислить выражение

$$\frac{9,81 \times 1,41}{3,14},$$

сохраняя каждый раз два десятичных знака. Осуществить операции во всех возможных порядках и сравнить.

**1.4. Погрешности из-за инструментов.** Когда вычисления производятся с использованием инструментов, то появляются погрешности, причиной которых является несовершенство инструментов.

Эта погрешность логически могла бы быть приближена к погрешности метода. В действительности же она отличается от таковой!

— прежде всего очень мало известно о ней с точки зрения ее причин и эффектов;

— ее очень трудно оценить, и, в частности, она может быть нерегулярной;

— ее очень трудно уменьшить. А ниже некоторого уровня это невозможно.

Информация о такой погрешности бывает следующего типа: порядок величины точности или информация вероятностного типа.

**1.5. Погрешности, возникающие из-за неточности исходных данных, — неустраняемые погрешности.** Две очень близкие величины неразличимы. Необходимо всякий раз, как изучается некоторая величина, изучать и близкие ей величины. Отсюда следует, что всякая выкладка, в которую входят приближенные величины, имеет некоторый небольшой изъян неопределенности.

Информация о такой погрешности бывает типа интервала приближения или точности, или же вероятностного типа.

Эта погрешность имеет ту особенность, что она не может снижаться в результате изменений счета, кроме тех случаев, когда обращаются к вероятностным задачам.

**У п р а ж н е н и е 2.** Какие типы погрешностей появляются:

а) при решении алгебраической системы уравнений;

б) при осуществлении тройного правила на счетной логарифмической линейке.

**З а д а ч а 1.** Допустим, что мы хотим вычислить методом трапеций некоторый интеграл на отрезке  $[0, 1]$ , разделив этот отрезок на  $n$  равных частей. Для второй производной подынтегральной функции на этом отрезке имеется интервал приближения

$$0,3 \leq f''(x) \leq 0,8.$$

Мы вычисляем интеграл при  $n = 10$  и  $n = 20$  с погрешностью вычисления, имеющей точность  $10^{-5}$ .

Находим  $I_{10} = 3,61482$ ,  $I_{20} = 3,61453$ .

а) Указать один интервал приближения для значения интеграла, используя  $I_{10}$ , и другой интервал — используя  $I_{20}$ .

б) Можно ли указать для  $I$  приближенное значение так, чтобы величина точности была минимально возможной?

## II. РАСПРОСТРАНЕНИЕ ПОГРЕШНОСТЕЙ

**2.1. Обычный способ вычисления погрешностей.** Рекомендуемый обычно способ для вычисления погрешности состоит в том, чтобы выяснять во время каждой операции информацию (выбранного типа) о результатах операции, принимая во внимание:

- информацию о погрешностях членов операции;
- информацию о погрешностях собственно операции.

Это, вообще говоря, очень тяжело.

Например, для вычисления при помощи интервала приближения на возрастающих функциях каждая операция алгоритма будет в действительности заменяться двумя операциями:

— одной — над нижними границами интервала приближения;

— другой — над верхними границами.

При вычислениях при помощи некоторой точности будет наблюдаться мало заметное возрастание этого числа, причем каждый переход от погрешности к точности сопровождается некоторой потерей информации.

**2.2. Пример.** Рассмотрим рекуррентную формулу

$$u_{n+1} = 3u_n - 2u_{n-1}.$$

Допустим, что члены  $u_0$  и  $u_1$  заданы с точностью  $b$  и на каждом шаге погрешность вычисления имеет точность  $b$ .

Можно задать для  $u_2$  точность  $6b$ .

Для  $u_i$  находим точность

$$k_i b,$$

где

$$k_{i+1} = 3k_i + 2k_{i-1} + 1, \quad i = \overline{1, 6}.$$

Таким образом, для  $u_7$  получаем

$$3441b.$$

Мы увидим, что, действуя другим образом, можно получить гораздо более слабую точность. Для этого мы введем несколько общих понятий.

**3.1. Источник погрешности.** Для изучения погрешностей в вычислениях удобно их разделять на типы.

Пусть, например, нужно вычислить

$$\pi \sqrt{2}.$$

Появляются следующие три погрешности

— в результате использования приближенного значения для  $\pi$ , например, 3,142;

— использования приближенного значения для  $\sqrt{2}$ , например, 1,414;

— опускания последних десятичных знаков в произведении, например, последних двух.

Точное вычисление, в результате которого появляется каждая из этих погрешностей, называется *источником* этой погрешности.

**3.2. Погрешности, исходящие из одного и того же источника.** Толкование в качестве независимых двух погрешностей, исходящих из одного источника, не представляет никакого интереса. Покажем это на примере.

Пусть требуется вычислить  $a(6 - a)$ , зная, что  $3 \leq a \leq 3,1$ .

Легко показать, что вычисляемая функция убывает на рассматриваемом интервале приближения. Значит, искомое значение заключено между  $3 \cdot (6 - 3)$  и  $3,1 \cdot 2,9$ , т. е. между 9 и 8,99.

Если мы теперь запишем эту функцию в виде  $6a + a^2$ , то первый член будет заключен между 18 и 18,6, второй — между  $-9,61$  и  $-9$ .

Отсюда выводим, что результат должен быть заключен между 8,39 и 9,6.

Вернемся к записи  $a(6 - a)$  и отметим, что  $a$  заключено между 3 и 3,1, а  $(6 - a)$  — между 2,9 и 3; значит, результат заключен между 8,7 и 9,3.

Каждый из этих двух интервалов приближения гораздо грубее, чем тот, с которым мы имели дело вначале.

**4.1. Распространенная погрешность.** Возьмем вновь пример из п. 3.1.

Теперь мы для изучения каждой из погрешностей введем некоторое фиктивное вычисление, при котором она появляется единственный раз.

Введем следующие вычисления:

$C_1$  — точное вычисление  $\pi\sqrt{2}$ ;

$C_2$  — точное вычисление  $3,142 \cdot \sqrt{2}$ ;

$C_3$  — точное вычисление  $3,142 \cdot 1,414$ ;

$C_4$  — вычисление с тремя десятичными знаками  $3,142 \cdot 1,414$ .

Каждое из этих вычислений эффективно отличается от предыдущего введением источника погрешностей.

Каждое из этих вычислений имеет по отношению к пре-

дыдущему погрешность, которая называется *распространенной погрешностью* соответствующего источника.

Общая погрешность — единственная, которая нас интересует, — является просто *алгебраической суммой* распространенных погрешностей.

Этот способ подсчета погрешностей, когда он применим, имеет следующие преимущества:

— он отчетливо выделяет в общей погрешности ответственность каждого источника погрешности;

— он устраняет переход от погрешности к информации о погрешности, что снижает потерю информации.

**4.2. Первый пример подсчета общей погрешности.** Возьмем вновь вычисление из п. 4.1.

Заменяем сначала  $\pi$  на 3,142:

$$0 < 3,142\sqrt{2} - \pi\sqrt{2} < 8 \cdot 10^{-4} \quad (1)$$

(мажорируем  $5 \cdot 10^{-4}\sqrt{2}$  посредством  $8 \cdot 10^{-4}$ ).

Теперь заменим  $\sqrt{2}$  на 1,414:

$$-16 \cdot 10^{-4} < 3,142 \cdot 1,414 - 3,142 < 0 \quad (2)$$

(заменяем  $5 \cdot 10^{-4} \cdot 3,142$  на  $16 \cdot 10^{-4}$ ).

Осуществляем умножение:

$$3,142 \cdot 1,414 = 4,442788$$

и результат заменяем на 4,443; имеем

$$0 < 4,443 - 3,142 \cdot 1,414 < 3 \cdot 10^{-4}. \quad (3)$$

Сложив почленно неравенства (1), (2), (3), получим:

$$-16 \cdot 10^{-4} < 4,443 - \pi\sqrt{2} < 11 \cdot 10^{-4}.$$

**У п р а ж н е н и е 3.** Сравнить это вычисление (перед окончательным округлением) с точки зрения стоимости (число цифр) и точности (интервалы приближения преобразовать в точности) со следующими вычислениями:

— интервал приближения, исходя из

$$3,141 < \pi < 3,142 \text{ и } 1,414 < \sqrt{2} < 1,4145,$$

— точность, исходя из

$$3,142 \text{ и } 1,414 \quad \text{точность } \frac{1}{2} \cdot 10^{-3}$$

$$3,14175 \text{ и } 1,41425 \quad \text{точность } \frac{1}{4} \cdot 10^{-3}.$$

**4.3. Второй пример.** Вернемся к примеру из п. 2.2. Формула

$$u_{n+1} = 3u_n - 2u_{n-1}$$

линейна. Стало быть источники погрешностей независимы один от другого и каждая погрешность распространяется по самой рекуррентной формуле.

В результате погрешность в задании  $u_0$  и  $u_1$ , а также погрешность  $b$ , которая дополнительно вносится при каждом вычислении по рекуррентной формуле  $u_i, i = 2, \dots, 7$ , дают следующий вклад в общую погрешность для  $u_7$ .

Погрешность для  $u_0$  дает  $126b$

$u_1$	—	$127b$	$u_5$	—	$7b$
$u_2$	—	$63b$	$u_6$	—	$3b$
$u_3$	—	$31b$	$u_7$	—	$b$
$u_4$	—	$15b$	всего	—	$373b$

Это почти в 10 раз меньше, чем то, что мы находим в п. 2.2.

**5.1. Перенос погрешности.** Часто оказывается, что погрешность, совершенная на некотором шаге, отражается на конечном результате так же, как и некоторая другая, фиктивная, погрешность, которая была бы совершена на некотором другом шаге вычисления.

Замена одной погрешности другой будет называться *переносом погрешности*.

**П р и м е р ы.** Погрешности, совершаемые на каждом шаге приближенного решения дифференциального уравнения методом конечных шагов, могут быть истолкованы как переход от одной интегральной кривой к другой. Их можно интерпретировать как погрешности, возникающие из-за начальных условий.

Погрешности, которые появляются в процессе приведения матрицы к треугольному виду, могут быть истолкованы как погрешности значений членов исходной матрицы.

**6.1. Нейтральность погрешностей.** Рассмотрим вновь пример из п. 4.1. Различные промежуточные вычисления проводятся, исходя из значений, из которых одни являются точными, другие — приближенными.

Будем говорить, что погрешность с источником  $s$  *нейтральна* относительно погрешности с предшествующим

источником  $s_0$ , если исключение  $s_0$  незначительно изменяет распространенную погрешность относительно  $s$

**П р и м е р.** При вычислении в п. 4.1 погрешность, возникающая из-за замены  $\sqrt{2}$  на 1,414, нейтральна относительно погрешности приближения  $\pi$ , поскольку 3,142 ( $1,414 - \sqrt{2}$ ) и  $\pi$  ( $1,414 - \sqrt{2}$ ) отличаются между собой менее, чем на 0,02%.

Напротив, окончательная округленная погрешность не будет нейтральной относительно замены  $\pi$  на 3,14.

Теперь пусть вычисляется

$$\frac{1}{\pi - \sqrt{2} - \sqrt{3}}$$

с двумя десятичными знаками для  $\pi$ ,  $\sqrt{2}$ ,  $\sqrt{3}$ . Погрешность, которую мы получаем в результате из-за приближения  $\sqrt{2}$ , не будет нейтральной относительно приближения  $\sqrt{3}$ , поскольку  $\pi$  не берется точным.

В самом деле,

$$\frac{\frac{1}{\pi - 1,41 - \sqrt{3}} - \frac{1}{\pi - \sqrt{2} - \sqrt{3}}}{\frac{1}{\pi - 1,41 - 1,73} - \frac{1}{\pi - \sqrt{2} - 1,73}} \approx -2,2,$$

т. е. получаем число, не являющееся близким к 1.

Когда мы замечаем, что некоторые погрешности не являются нейтральными, следует быть особо внимательным

- к порядку операций;
- к вычислению погрешностей.

**З а м е ч а н и е.** Большинство погрешностей становятся нейтральными, когда различные погрешности источника берутся достаточно малыми.

**6.2. Пример упрощенного вычисления погрешностей в случае нейтральности.** Пусть требуется вычислить погрешность произведения  $abc$ , зная приближенные значения  $\alpha$ ,  $\beta$ ,  $\gamma$  в случае, когда имеется уверенность в нейтральности; вместо того, чтобы искать правильное выражение погрешности, мы будем довольствоваться ее «главной частью». Погрешность, происходящая из-за замены  $a$  на  $\alpha$ , равна  $(\alpha - a) \beta \gamma$ .

Точно то же для других членов.

Отсюда получаем полную погрешность

$$(\alpha - a) \beta \gamma + (\beta - b) \alpha \gamma + (\gamma - c) \alpha \beta.$$

Это можно переписать так:

$$\alpha \beta \gamma \left( \frac{\alpha - a}{\alpha} + \frac{\beta - b}{\beta} + \frac{\gamma - c}{\gamma} \right).$$

Получаем обычную формулу погрешности относительно произведения.

У п р а ж н е н и е 4. Мы хотим вычислить произведение

$$c = ab \text{ при } \left. \begin{array}{l} a = 1 \\ b = 1 \end{array} \right\} \text{ с точностью } i.$$

Какой точностью нужно наделить значение 1, рассматриваемое как приближение  $c$ ?

а) Сначала провести точное вычисление.

б) Затем провести приближенное вычисление, предполагая погрешности нейтральными.

с) До какого значения  $i$  применимо это упрощенное вычисление?

З а д а ч а 2. Проинтегрируем уравнение  $y' = y$ , исходя из условия  $y(0) = 1$ , методом касательной с  $h = 0,1$ . Погрешность на каждом шаге имеет в качестве главной части  $-\frac{h^2}{2}y$ .

Показать, что можно интерпретировать метод касательной с шагом  $h$  как решение дифференциального уравнения

$$y' = y - \frac{h}{2}y,$$

с погрешностью на одном шаге, пренебрежимой сравнительно с  $h^2$ .

Вывести отсюда погрешность значения, равного 1 для  $h = 1/n$ .

Сравнить с результатами упражнения 18 главы IV.

З а д а ч а 3. Пусть требуется вычислить приближение для  $f(a, b, c)$ , зная приближения  $\alpha, \beta, \gamma$  для  $a, b, c$ .

Предположим, что функция  $f$  обладает непрерывными первыми производными.

а) К чему сводится в этом случае условие нейтральности?

б) Показать, что оно выполняется, если  $\alpha, \beta, \gamma$  достаточно близки к  $a, b, c$ .

**З а д а ч а 4.** Пусть требуется решить методом Гаусса систему

$$ax + by = e, \quad cx + dy = f.$$

Полагаем  $\alpha = c/a, \quad g = d - b\alpha, \quad h = f - e\alpha$ . Вычисление этих выражений вводит погрешности источника  $\delta\alpha, \delta g, \delta h$ .

а) Найти погрешности  $\varepsilon_g$  и  $\varepsilon_h$  для  $g$  и  $h$ .

б) Найти систему, точная триангуляризация которой дает

$$ax + by = e, \quad (g + \varepsilon_g) y = h + \varepsilon_h.$$

с) Показать, что если  $\delta\alpha, \delta g, \delta h$  достаточно малы, то погрешности источника нейтральны. (При этом не рассматриваются погрешности, появившиеся при решении треугольной системы.)

д) Будет ли выполняться это условие для системы из гл. IV, п. 4.2, с погрешностями порядка  $10^{-4}$ ?

е) Тот же вопрос для погрешностей порядка  $10^{-8}$ .

### III. ОБЩИЕ ПРОБЛЕМЫ, ОТНОСЯЩИЕСЯ К ПОГРЕШНОСТЯМ

**7.1. Невозможность априорного изучения погрешностей.** За исключением нескольких очень простых случаев, изучение погрешностей требует информации о некоторых промежуточных или конечных результатах (точных или приближенных). Стало бы априорное изучение погрешностей, вообще говоря, невозможно.

**7.2. Основная и вторичная задачи.** Трудно построить стройную теорию, основанную на информации, которая относится частично к точным результатам, а частично к приближенным. На самом деле мы будем различать два типа задач, относящихся к погрешностям:

— *основные задачи*, в которых информация относится к приближенным результатам;

— *вторичные задачи*, в которых информация относится к точным результатам.

Выбор термина *основной* для задач первого типа объясняется тем, что это наиболее реальная задача, которая на самом деле интересует вычислителя.

**8.1. Апостериорная основная задача.** Эта задача состоит в вычислении погрешности после того как все вычис-

ления проведены. Значит, мы имеем всю желаемую информацию о приближенных результатах. Это наиболее употребительный случай в вычислительной практике. Он относительно благоприятен.

Элементы для решения этой задачи были приведены в двух первых параграфах настоящей главы.

**8.2. Независимое вычисление погрешности.** Необходимо упомянуть здесь один частный случай. Может случиться, что можно вычислить погрешность некоторого результата при помощи процесса, не зависящего от самого вычисления. Допустим, например, что нам известно для некоторого корня  $a$  уравнения  $f(x) = 0$  приближение  $\alpha$  и что на некотором интервале, содержащем  $a$  и  $\alpha$ ,

$$f'(x) \approx k, \quad k \neq 0.$$

Тогда можно считать  $f(\alpha) - f(a) \approx k(\alpha - a)$ . Отсюда

$$\alpha - a \approx \frac{f(\alpha)}{k}.$$

**У п р а ж н е н и е 5.** Рассмотрим линейную алгебраическую систему  $AX = B$ . Допустим, что известны матрица  $A^{-1}$  и приближенное решение  $X_0$ .

Можно ли вычислить погрешность, порожденную  $X_0$ , посредством прямой оценки?

**8.3. Установление плана вычисления, когда на погрешность налагаются условия.** Установление плана вычисления, когда на погрешность налагаются условия, должно было бы требовать вычисления погрешности априори, что, вообще говоря, невозможно. Стараются получить самую необходимую информацию при помощи одного из следующих процессов:

— проведение предварительного грубого вычисления, предназначенного только для того, чтобы получить самую необходимую информацию о погрешности;

— рассмотрение произвольного решения. Однако это решение может иметь серьезную мотивировку. Например, если имеется необходимость в порядке величины некоторого промежуточного этапа вычисления, то мы смотрим, имеет ли она физическое значение. Информация, полученная таким способом, является информацией относительно точных величин, но ее можно переносить и на приближенные величины. В других случаях, руководствуясь вычислительным опытом, обращаются к аналогичным случаям.

Как только вычисление проведено, получают апостериорную оценку. При этом можно ожидать таких исходов:

— величина оценки значительно ниже той, которая требуется. В этом нет ничего удивительного, поскольку информация, которой мы располагаем, гораздо полнее той, которая имелась в нашем распоряжении в момент построения плана вычисления;

— если было принято неудачное решение, то апостериорная оценка неприемлема. Тогда можно попытаться взять более точную погрешность. Если это не удастся, вычисление плохо проведено, и его надо проделать заново.

**9.1. Экспериментальное сравнение методов с точки зрения погрешности.** Может представлять интерес сопоставление нескольких методов, примененных к одной и той же задаче, точное решение которой известно.

Первый шаг состоит в эффективном исследовании различных приближенных вычислений. Тем самым получают различными методами значение погрешности.

При этом, однако, необходимо заметить, что:

— этот шаг очень долгий и дорогостоящий, поскольку приходится выполнять большой счет;

— результаты относятся к единственной изучаемой задаче и не имеют, стало быть, никакого общего значения. Чтобы придать им более значимый характер, требуется рассмотреть много задач; впрочем, при этом всегда имеется опасность оставить в стороне важные случаи, которые могли бы повести к различным заключениям;

— полученные погрешности могут быть по-разному интерпретированы. Вообще говоря, интерес представляют погрешности метода, но в вычисление могут вмешиваться почти произвольно погрешности округления;

— нет никакого «понимания» того, что происходит.

В защиту этого экспериментального метода заметим, что он позволяет составить мнение не только о погрешности, но и о плане вычисления (его продолжительности, практических трудностях, тонких местах и т. д.).

**9.2. Вторичная задача.** Другой способ, пригодный для сопоставления различных методов для одной задачи, теоретическое решение которой известно, состоит в оценке погрешности, исходя из этого теоретического решения. Тогда возникает то, что мы называем *вторичной* задачей.

Этот способ действий является быстрым, он дает средство изучения только желаемой погрешности (например,

погрешности метода, за исключением погрешностей округления), но имеет то неудобство, что он касается не самой погрешности, а информации относительно погрешности. Может оказаться, что различные типы информации имеют столь различную природу, что их невозможно сравнить. Например, формула трапеций для приближенных квадратур дает погрешность порядка  $f''(\xi)$ , а метод Симпсона — погрешность порядка  $f^{IV}(\xi)$ . С другой стороны, сравнение информации некоторых типов не позволяет сделать окончательного заключения. Один метод может дать погрешность меньше 0,1, а другой — погрешность меньше 0,01. Совершенно случайным образом, не более того, можно вывести отсюда, что второй метод лучше. Ведь могло оказаться, что оценка погрешности в первом случае была гораздо более точной.

Этот метод рекомендуется особенно для случая, когда о погрешности имеется информация следующих типов:

- порядок величины погрешности;
- вероятностное распределение погрешности.

**9.3. Изучение погрешности для множества вспомогательных вычислений.** Результаты, полученные для второй задачи, могут, вообще говоря, быть сформулированы в таком виде, что они будут применимы не только к единственной задаче, но и ко всей категории аналогичных задач, различающихся, например, исходными данными. Тогда можно изучать общие свойства погрешности.

Удобным оказывается рассматривать семейство функций, наделенное вероятностным распределением.

**Задача 5.** а) Сравнить погрешность на одном шаге в явном методе Адамса и в улучшенном методе касательной для уравнения  $y' = y$ , когда шаг в обоих случаях один и тот же.

б) Сравнить сами эти погрешности при равной работе. (Взять в первом случае шаг  $h$ , а во втором — шаг  $2h$ .)

#### РЕШЕНИЯ УПРАЖНЕНИЙ ГЛАВЫ V

$$1) 1^\circ. a \times b = 13,83, \frac{a \times b}{c} = 4,40;$$

$$2^\circ. \frac{a}{c} = 3,12, \frac{a}{c} \times b = 4,40;$$

$$3^\circ. \frac{b}{c} = 0,44, \frac{b}{c} \times a = 4,41.$$

Наиболее точен третий способ.

- 2) а) Погрешности округления.  
 б) Погрешности из-за инструмента.  
 3) Интервал приближения — 79 цифр точность  $12 \cdot 10^{-4}$ .

Точность  $\frac{1}{2} 10^{-3}$  — 33 цифры, точность  $23 \cdot 10^{-4}$ .

Точность  $\frac{1}{2} 10^{-3}$  — 62 цифры, точность  $12 \cdot 10^{-4}$ .

Рекомендуемый метод — 33 цифры, точность  $12 \cdot 10^{-4}$ .

4) а)  $i_c = i^2 + 2i$ . б)  $i'_c = 2i$ .

с) Для  $c = 0,2$  находим  $i_c = 0,44$ ,  $i'_c = 0,4$ . Упрощение снова применимо.

5)  $A(X_0 - X) = AX_0 - B$ , откуда  $X_0 - X = A^{-1}(AX_0 - B)$ .

## РЕШЕНИЯ ЗАДАЧ ГЛАВЫ V

Задача 1.

а)  $3,61482 - 68 \cdot 10^{-5} \leq I \leq 3,61482 - 24 \cdot 10^{-5}$ ;  
 $3,61453 - 23 \cdot 10^{-5} \leq I \leq 3,61453 - 5 \cdot 10^{-5}$ .

б)  $[3,61439 - I] \leq 9 \cdot 10^{-5}$ .

Задача 2. Решение уравнения  $y' = y(1 - h/2)$  имеет вид  $e^{(1-h/2)x}$ . Для  $x = h$  находим

$$e^{h-h^2/2} = 1 + \left(h - \frac{h^2}{2}\right) + \frac{1}{2} \left(h - \frac{h^2}{2}\right)^2 + \dots = 1 + h + \lambda h^3 + \dots$$

Пренебрегая всеми членами, начиная с содержащих  $h^3$ , посредством приближенного метода с абсциссой 1, получаем  $e^{1-1/(2n)}$ .

Непосредственно вычислим

$$\left(1 + \frac{1}{n}\right)^n = e^{n \ln(1+1/n)},$$

начало этого разложения имеет вид  $e^{1-1/(2n)}$ .

Задача 3. Запишем условно

$$f(\alpha, \beta, \gamma) - f(a, b, c) = (\alpha - a) f'_a(a_1, b_1, c_1) + \\ + (\beta - b) f'_b(a_2, b_2, c_2) + (\gamma - c) f'_c(a_3, b_3, c_3).$$

а) Нейтральность сводится к утверждению, что  $f'_a, f'_b, f'_c$  мало меняются на прямоугольниках, на которых  $a, b, c, \alpha, \beta, \gamma$  являются двумя противоположными вершинами.

б) Это следует из непрерывности производных.

Задача 4. а)  $\varepsilon_g = \delta g - b\delta\alpha, \varepsilon_h = \delta h - e\delta\alpha$ .

б)  $ax + by = e, (c + a\delta\alpha)x + (d - \delta g)y = f + \delta h$ .

с) Можно применить формулу конечных приращений.

д) Нет, так как в этом случае определитель может обратиться в нуль.

е) Да, так как определитель, хотя и очень мал, очень мало меняется по относительному значению.

**Задача 5. а) Метод Адамса:**

$$\frac{h}{2} \left( 3 - 1 + h - \frac{h^2}{2} \right) = h - \frac{h^2}{2} - \frac{h^3}{4}.$$

(Предполагаем  $y_{-1}$  точным вплоть до 2-го порядка.) Погрешность:  $-5h^3/12$ .

Улучшенный метод касательной:  $h + h^2/2$ . Погрешность:  $-h^3/6$ .

Второй метод лучше.

б) На втором шаге.

Метод Адамса:  $-5h^3/6$ .

Улучшенный метод касательной:  $-8h^3/6$ .

Первый метод лучше.

**1.1. Ошибка.** В обычной речи не делается различия между словами погрешность и ошибка, поэтому мы кратко разъясним различие в их смысле применительно к вычислениям.

До сих пор мы изучали понятие *погрешности*. При этом мы придерживались того, что, кроме очень редких случаев, всякий результат есть только приближенный результат, содержащий некоторую погрешность.

*Ошибка* же происходит от неправильного использования математического понятия или орудия счета. Могут быть ошибки рассуждения, ошибки записи, ошибки счета, ошибки программирования.

**1.2. Тактика поведения по отношению к ошибкам.** Одно из первых неприятных соображений, возникающих у тех, кто хочет практически применять математику, может быть сформулировано следующим образом: «Вычисление, достаточно длинное и не проверенное, почти наверняка неправильно».

Перед этой ситуацией не следует сокрушаться или отчаиваться, а следует:

— искать способ совершать как можно меньше ошибок;

— отыскать и исправить уже сделанные.

Первый пункт осуществляется тренировкой и знанием самого себя. Второй пункт осуществляется при помощи технических приемов, к изложению которых мы и переходим.

## I. ПРОВЕРКА

**2.1. Общие понятия.** Под *проверкой* мы понимаем множество манипуляций, при помощи которых в процессе счета или по его завершении можно получить представление о его точности.

Проверки могут быть в самой различной обстановке, в зависимости от того, проводится счет одним оператором

(вручную или на настольной машине), или же идет счет на машине с программой.

Во втором случае ошибки могут быть довольно редки, как только программа создана (и значит, проверка будет достаточно быстрой). Напротив, при создании программы могут появиться логические ошибки.

В первом же случае, наоборот, следует считать нормальным наличие довольно большого числа ошибок. Следовательно, результат должен рассматриваться как пригодный лишь после очень серьезных проверок. Мы будем рассматривать именно этот случай.

Вообще говоря, не существует общего метода, который позволял бы проверить вычисление, а имеются только много частных случаев, когда можно указать некоторый способ проверки. И очень редко случается, чтобы нельзя было применить тот или другой из этих способов.

Мы различаем

- операционные проверки;
- проверки на основе различных соображений;
- проверки повтором вычисления;
- проверки замыканием.

Перейдем к рассмотрению этих типов проверок.

**3.1. Операционные проверки.** Мы будем называть *операционными проверками* те, которые основаны на математических свойствах осуществляемой операции и будем различать:

- арифметические проверки;
- проверки функциональных тождеств;
- линейные проверки.

**3.2. Арифметические проверки.** Они носят характер проверки при помощи деления на 9 и применяются лишь при операциях, выполняемых точно.

Приведем несколько примеров:

- проверка значения определителя с целыми членами посредством деления на 9;
- проверка численного значения многочлена посредством деления на 9;
- проверка значения факториала при помощи отыскания степени числа 3, которое он содержит.

**У п р а ж н е н и е 1.** Вычислено значение многочлена

$$5x^6 + 8x^5 - x^4 + 7x^3 - 2x^2 + 1,$$

при а)  $x = 2$ , б)  $x = 3$  и найдены соответственно значения 305 и 5680.

Проверить эти значения надлежащим способом.

**3.3. Проверка функциональных тождеств.** Такое тождество можно проверить, задавая переменному (или переменным) простые частные значения. Например, проверяем произведение многочленов, задавая  $x$  значение 1.

**У п р а ж н е н и е 2.** Пусть имеется разложение

$$\frac{2x^4 + x^3 - 27x^2 + 27x - 243}{(x^2 - 9)^3} = \\ = \frac{1}{(x-3)^2} + \frac{1}{(x+3)^2} + \frac{2}{(x+3)^3} - \frac{1}{(x-3)^3}.$$

Проверить это разложение.

**3.4. Линейные проверки.** Когда вычисление (или часть вычисления) линейно, то его можно проверить, используя тот факт, что линейное преобразование сохраняет сумму.

Например, произведение двух матриц можно проверить, прибавив к первой дополнительно одну строку, являющуюся суммой строк матрицы. В произведении появится дополнительная строка, которая тоже будет суммой остальных строк. Проверка такого типа является классической при решении систем уравнений первой степени.

Такая проверка применяется также, например, к линейному рекуррентному соотношению. Пусть  $u_{n+1} = au_n + bu_{n-1}$ . Тогда

$$\sum_2^{12} u_i = a \sum_1^{11} u_i + b \sum_0^{10} u_i.$$

Этот способ проверки применяется очень часто, поскольку очень многие вычисления содержат линейные части.

**У п р а ж н е н и е 3.** Указать в методе Гаусса проверки, базирующиеся на линейности.

**4.1. Проверки посредством изучения результата.** Мы будем различать:

— проверки, исходящие из порядка величины или знака;

— эмпирические констатации;

— проверки на основе физических условий;

— проверки семейства близких вычислений;

— проверки на основе внутренней связи.

**4.2. Проверка на основе порядка величины или знака.**

Часто оказывается, что порядок величины числа или его

знак могут быть установлены в результате математических заключений.

Например, в итерации часто можно показать, что последовательные значения приближаются к своему пределу как частичные суммы геометрической прогрессии.

В интерполяции можно предвидеть, вообще говоря, порядок величины и знак членов, соответствующих первым разностям, вторым и т. д.

**У п р а ж н е н и е 4.** Интерполируя в таблице

$x$	$f(x)$
0,32	0,43195
0,33	0,45258

находим  $f(0,326) = 0,44932$ .

Является ли этот результат удовлетворительным?

**5.1. Эмпирические констатации.** Мы относим к этому случай, когда в процессе вычисления устанавливается, что последовательные частичные результаты либо все положительны, либо имеют чередующиеся знаки, либо убывают.

Здесь речь идет не о результатах, доказанных заранее, а об эмпирическом установлении. Наблюдение за такими частными утверждениями позволяет обнаружить аномалию, которая если не показывает наличие ошибки, то по крайней мере служит основанием для более тщательного пересмотра этой части вычисления.

**5.2. Проверка на основе физических условий.** Этот тип проверки состоит в том, чтобы убедиться, что найденное решение хорошо удовлетворяет требуемым условиям физического явления: порядок величины, знак, сравнение с результатами, полученными другими приемами.

Если эта проверка не дает удовлетворительных результатов, то это можно относить:

— либо за счет ошибки,

— либо за счет недостаточно хорошо составленного уравнения.

И наконец, следует учитывать, что грубые ошибки могут быть не подвержены такому способу проверки. Например, инженер или физик может утверждать, что некоторый коэффициент при вычислении

— положителен;

— близок к 1.

Тогда коэффициент 1, 2 будет приемлемым, чего нельзя сказать с точностью лучше 30%.

У п р а ж н е н и е 5. Вычисления, относящиеся к физическим процессам, привели к следующим кривым (рис. 4). Приемлемы ли эти результаты?

6.1. Проверка семейства близких вычислений. В часто встречающихся случаях семейств близких вычислений информация составляется из результатов различных вы-

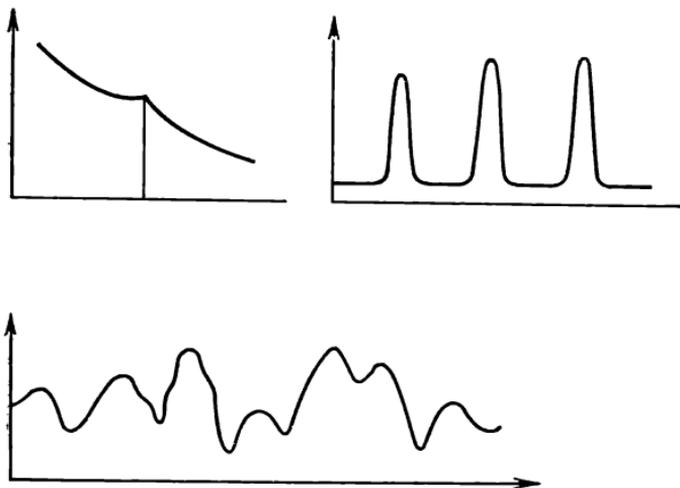


Рис. 4.

числений. В том случае, когда все эти вычисления, будучи близкими, отличаются большим числом исходных данных, довольствуются изучением вопроса, будет ли разница между результатами сравнима с разницей между исходными данными.

6.2. Случай, когда вычисления отличаются значениями параметра. Очень часто вычисления отличаются значениями параметра; тогда можно составить график, представляющий различные результаты как функцию параметра, и изучить, будет ли этот график носить достаточно правильный характер.

6.3. Применение разностей. Использование графика приводит лишь к грубой проверке, порядка  $1/100$  максимального значения, участвующего в вычислениях. Точность метода можно повысить, используя разности, если значения параметра распределены регулярно.

Нерегулярности гораздо более чувствительны к разностям. Однако по этому пути нельзя пройти очень далеко (см. задачу 2).

**6.4. Внутренняя связь результатов.** В некоторых задачах, например, при решении дифференциальных уравнений или при решении дифференциальных уравнений в частных производных, имеются многочисленные результаты, которые можно сравнивать между собой.

Так, для дифференциальной задачи с начальными условиями составляются разности последовательных значений решения. Для уравнения в частных производных составляется график полученных результатов на основе изменения только одного переменного.

Такие проверки возможны в очень многих случаях, они в высшей степени эффективны.

У п р а ж н е н и е 6. След приближенного решения дифференциальной задачи с начальными условиями имеет вид кривой на рис. 5.

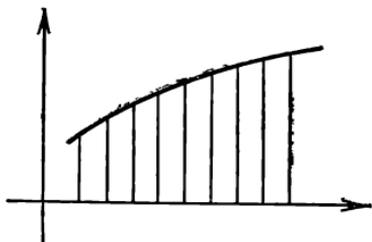


Рис. 5.

Что можно сказать об этом?

**6.5. Случай итерации.** Мы еще будем обращаться в этом параграфе к проверке итерации. Последовательные результаты являются близкими и в общем случае известен способ, при помощи которого они стремятся к точному результату.

Методы итерации — наиболее автокорректируемые. Отдельная ошибка, вообще говоря, не мешает последовательности итераций стремиться к точному результату.

У п р а ж н е н и е 7. Вычисление привело к последовательности комплексных чисел:

$$1 + i + \left(\frac{2 - 3i}{5}\right)^n.$$

Как расположены эти числа?

**7.1. Проверка посредством повтора вычисления.** В этом случае можно иметь в виду два варианта:

- полное повторение вычисления;
- выполнение вычисления по-другому.

**7.2. Проверка посредством полного повторения.** Часто бывает полезно проверить вычисление, проведя его дважды. Это длинный процесс, поскольку он требует вдвое больше времени. Впрочем, он и не очень эффекти-

вен. В самом деле, если человек плохо прочитал цифры или плохо интерпретировал требуемую работу, то он рискует повториться, особенно если вычисление проводится снова непосредственно после первого.

Таким образом, рекомендуется либо сделать достаточно продолжительной перерыв перед повторением вычисления, либо доверить второе вычисление другому лицу.

**7.3. Проверка путем другого вычисления.** Предпочтительно, когда это возможно, проверить вычисление, проведя другое вычисление, отличное от первого. Различие может быть в методе и в исполнении.

Например, одно и то же дифференциальное уравнение можно решать при помощи двух различных методов, при помощи одного метода, но с разным шагом.

**7.4. Взаимосвязь с оценкой погрешности.** Этот способ проверки является довольно дорогим с точки зрения времени.

Например, если повторить интегрирование дифференциального уравнения с двойным шагом, то время вычисления возрастет на 50%. Но этот способ действий очень рентабелен, поскольку сравнение двух результатов позволяет в то же время получить оценку погрешности.

**З а м е ч а н и е 1.** Повторение вычисления с тем же шагом удваивает время вычисления и ничего не дает для погрешности.

**З а м е ч а н и е 2.** Применение этого способа зависит от интуиции вычислителя. Если, например, два вычисления (второе несколько менее точное) дают 1,41 и 1,39, то следует интерпретировать эту разницу в 0,02 либо как погрешность, либо как ошибку. Необходимо иметь представление о разумной разнице между этими вычислениями. Но даже с этой информацией могут остаться сомнения, например, если ожидаемая разница составляет около 0,01.

**У п р а ж н е н и е 8.** Указать правило для оценки погрешности, исходя из двух вычислений площади методом трапеций, один — с шагом  $h$ , а другой — с шагом  $2h$ .

**8.1. Проверка замыканием.** Замыканием называют всякое проверочное вычисление, которое, исходя из результатов, приводит к заданным заранее числам.

Существует очень много способов проверки замыканием. Приведем несколько примеров.

**С ч и т ы в а н и е.** В этом случае копируются одно или несколько чисел и читается копия, которая сравнивается

с оригиналом. Считывание производится двумя людьми, первый из которых читает копию, а второй следит за ним по оригиналу. Немалое время можно выиграть при этом, если вместо чисел читать последовательно цифры или тройки цифр.

Повторение тотализатора на клавиатуре в настольной машинке. Когда пользуются настольной машинкой, то очень часто можно заставить перейти число с тотализатора на клавиатуру способом, отличным от исходного.

Работа проверяется сведением тотализатора к нулю при вычитании числа, записанного на клавиатуре.

Вычисление разностей. Проверяем вычисления разностей  $\Delta = b - a$ , составляя  $\Delta + a = b$ .

Во всех замыканиях проверка должна производиться точно. Когда проверка бывает лишь приближенной, то вообще говоря, имеется указание на погрешность.

Например, получив решение системы уравнений первой степени, можно проверить его подстановкой в систему. То же самое вычисление может служить для оценки погрешностей. (По этому вопросу см. задачу 2.)

**9.1. Эффективность проверки.** Некоторые проверки малоэффективны и их воздерживаются применять. Например, деление на 2 или на 5 оставляет в каждом числе лишь цифру меньшего достоинства. Подстановка  $x = 0$  в многочлен оставляет лишь свободный член.

**У п р а ж н е н и е 9.** а) Что можно сказать об эффективности двух первых проверок, предложенных в упражнении 2?

б) Как можно классифицировать с точки зрения эффективности проверку умножения двух многочленов

- делением на 9;
- численной подстановкой?

**9.2. План проверки.** В принципе, план проверки составляется заранее (что не должно мешать уделять внимание<sup>1</sup> любым особенностям).

Этот план будет выбираться в зависимости от наиболее частых ошибок, присущих тому, кто осуществляет вычисление, в установлении такого порядка, чтобы никакая часть вычисления не ускользнула от проверки.

Имеется тенденция считать, что вычисление, с успехом выдержавшее первую проверку, полностью правильно. Однако, например, связь результатов между собой не исключает систематической погрешности, осуществляе-

мой регулярным образом в различных результатах. И напротив, следует избегать различных проверок там, где достаточно одной. Например, если решается линейная задача, то бесполезно проводить вычисление дважды, поскольку проверка посредством сумм весьма эффективна.

При ручном счете или при счете на настольных машинах на проверку тратится от 20% до 30% общего времени.

**У п р а ж н е н и е 10.** Пусть хорошо обусловленная система уравнений первой степени решена методом Гаусса.

а) Что мы проверим, подставив значение неизвестных во второе уравнение?

б) Что мы проверим, подставив эти значения в последнее уравнение? Будет ли эта проверка эффективной?

**10.1. Уверенность в вычислении.** Логическая уверенность есть понятие из области интуиции, вычисление же есть конкретное действие. В этой области нужно довольствоваться практической уверенностью. Автор хорошо проведенного и хорошо проверенного вычисления уверен в отсутствии ошибки.

Можно заметить, что математик, представляющий доказательство, находится точно в той же ситуации, ибо он не способен доказать себе, что его доказательство верно.

**11.1. Локализация ошибок.** Проверки позволяют нам обнаружить случайные ошибки. Остается их локализовать.

Эта работа оказывается связанной с механической или электрической системой, к ней приходят в результате размышления о структуре вычисления и ошибки. В результате же надлежащих испытаний можно утверждать, что та или иная часть вычисления является точной или, наоборот, содержит ошибку.

**11.2. Размышления о структуре вычисления и ошибки.** Случается, что некоторые ошибки по своей природе могут произойти лишь во вполне определенной части вычисления.

Например, если вся серия аналогичных результатов ошибочна, то это происходит, вероятно, из-за элементов, общих для всей серии. Если же в серии ошибок всего один элемент, то, напротив, вероятно, что общие элементы являются точными.

Если результаты вычисления связаны между собой, но не совпадают с опытом, то исследуют использовавшиеся формулы и основные документы. Исследуют, не было ли

оплошностей с обозначениями. (Например, для функции  $E_1(x)$ , для некоторых функций Бесселя имеется несколько не совпадающих определений.)

Если ошибка возрастает по величине, то можно заранее исключить все части вычисления, которые вносят лишь незначительные изменения (например, интерполяции).

**11.3. Отыскание ошибок при помощи разбиения вычисления.** В том случае, когда ни один из описанных выше способов не позволяет локализовать ошибку (что случается часто), производят сечение, т. е. вычисление разбивается на две части и пытаются выяснить относительно каждой части, содержит ли она ошибки.

Уже проведенные проверки часто позволяют сразу вынести заключение; в противном случае производится новая проверка.

Следует придерживаться следующих правил:

— не забывать, что ошибка может быть в самой констатации расхождения, если это расхождение произошло из-за проверки, то прежде всего надо тщательно проверить проверку (повторив ее), поскольку это естественно короче, чем повторять все вычисления;

— не бросать изучаемую часть вычисления, не узнав с уверенностью, содержит она ошибки или нет.

**11.4. Замечания о природе ошибок.** Самые частые ошибки в то же время и самые простые (запятая, знак, плохо написанная или плохо прочитанная цифра).

**З а д а ч а 1.** Вычисляется

$$\begin{vmatrix} 1 & 5 & -2 & 4 \\ 6 & 0 & 3 & -2 \\ 9 & 13 & 8 & 4 \\ 1 & 7 & 2 & -5 \end{vmatrix} = -20 \cdot 129.$$

а) Проверить при помощи надлежащих делений.

б) Достаточны ли деления на 4 и 5?

**З а д а ч а 2.** Возьмем плохо обусловленную систему из главы IV, п. 4.2.

а) Простая подстановка  $n$  чисел в уравнения может ли:

α) привести к появлению ошибки;

β) дать гарантию отсутствия ошибки?

б) При исследовании различных возможных ошибок отыскать те, которые будут обнаружены подстановкой в уравнения.

## II. КОНТРОЛЬ

**12.1. Контроль.** Термином *контроль* мы обозначаем те логические операции, которые совершаются над результатами вычислений, чтобы:

- убедиться в отсутствии или наличии ошибки (чтобы в случае необходимости локализовать или исправить их);
- характеризовать погрешность.

Контроль отличается от проверки тем, что он имеет дело только с конечными результатами. Промежуточные результаты, сопровождающие метод, бывают известны, вообще говоря, лишь частично.

В частности, различие между погрешностью и ошибкой затушевывается, если неизвестен способ, при помощи которого производится вычисление. Имеется только понятие отклонения между полученным результатом и теоретическим значением.

**13.1. Предварительная информация.** Следует пытаться получить предварительно всю возможную информацию относительно изучаемого материала, например:

- самые последние публикации, если их было несколько;

- последовательные введения;

- степень распространенности материала;

- статьи, излагающие условия вычисления;

- критические статьи по этому поводу;

- документы, содержащие сравнимые результаты.

Из этих документов пытаются вывести предварительное мнение о том, насколько материал заслуживает доверия. Например, весьма вероятно, что классические таблицы логарифмов с пятью десятичными знаками не имеют ошибок. Напротив, можно относиться с подозрением к работе некоторых авторов, неопытных или имеющих репутацию небрежных.

**14.1. Полный контроль. Частичный контроль.** Будем называть *полным* контроль, исход которого позволяет утверждать, что отсутствуют ошибки в результатах контроля, и дает существенную информацию о погрешности. Например, подстановка различных корней в уравнение осуществляет в этом уравнении полный контроль самих корней. То же самое происходит при образовании элементарных симметричных функций корней.

*Частичный* контроль не дает столь утвердительных результатов. Из него в случае успеха вытекают лишь некоторая вероятность отсутствия ошибок и частичная ин-

формация о погрешности. В случае неуспеха (и после проверки отсутствия ошибки в контроле) из него делается вывод, что результаты ложны.

**14.2. Подготовка программы контроля.** Выбор испытаний контроля зависит от информации, которую хотят получить.

Например, в инженерных расчетах часто допускают, что результаты могут иметь грубую точность и хотят знать порядок величины погрешности.

В таблице требования бывают гораздо более сильными. Они могут доходить до требования, чтобы все заданные значения были наилучшими приближенными значениями с  $n$  десятичными знаками.

Выбор испытаний зависит также от информации, которая имеется и, в частности, от способа, которым эти результаты были получены.

Например, нельзя осуществлять контроль таблицы одним и тем же способом, если она получена последовательным табулированием или же ее значения определялись независимо друг от друга.

При выборе частичного контроля следует иметь в виду, что его испытания должны быть по возможности отличными от методов вычисления и проверки.

Пусть, например, требуется осуществить контроль корней алгебраического уравнения, отыскивавшихся последовательно, причем нахождение каждого корня сопровождалось делением на множитель, таким образом отделяемый.

Допустим, что найден корень  $\alpha$ ; упростим процесс при помощи множителя

$$x - \alpha,$$

не проверяя, обращается ли в нуль остаток. Даже если  $\alpha$  ошибочно, полученное уравнение будет таким, что сумма всех корней (включая  $\alpha$ ) будет правильной.

В этом случае наиболее эффективной проверкой было бы умножение.

**У п р а ж н е н и е 11.** Таблица значений тригонометрических функций вычислена следующим образом:

— вычисление  $\sin n$  ( $n$  — целое число) разложением в ряд;

— подтабулирование (систематическая интерполяция возрастающей степени) для значений от минуты к минуте.

Представить себе контроль этой таблицы.

**14.3. Сравнение с независимыми результатами.** Если располагают результатами, независимыми от контролируемых результатов, то можно произвести совмещение. Этот случай очень часто употребляется для таблиц. В случае неприемлемой расходимости останется решить, какое из двух значений неверно.

**15.1. Коррекция найденных ошибок.** Когда результаты четко признаны ошибочными, их часто можно подправить, не проводя заново самого вычисления.

Например, если таблица содержит изолированное значение, явно ошибочное, его можно исправить, находя его снова путем интерполяции между соседними с ним значениями или при помощи надлежащего изучения разностей.

У п р а ж н е н и е 12. Рассмотрим следующую таблицу:

0,37	0,5541
0,38	0,5855
0,39	0,6248
0,40	0,6419
0,41	0,6669
0,42	0,6897
0,43	0,7103

а) Осуществить контроль посредством разности.

б) Возможно, имеется ошибка. Если да, то исправить.

с) Можно ли после коррекции еще улучшить, чтобы сделать вторые разности более правильными?

**15.2. Отыскание источника обнаруженной ошибки.** Часто, когда ошибка обнаружена и исправлена, можно восстановить ее механизм. Речь идет, разумеется, лишь о предположениях (которые, однако, могут быть весьма вероятными).

Так, если в таблице замечено, что некоторое значение ошибочно в одной цифре, но не в последней (например,  $\ln 2$  есть 0,31103 вместо 0,30103), то можно с большой вероятностью считать, что речь идет об ошибке.

Точно так же, если установлено, что при интегрировании дифференциального уравнения две последовательные точки имеют необычное расстояние, то можно полагать, что это происходит от ошибки в длине шага. Тогда повторяют вычисление с ошибочной длиной, чтобы убедиться в справедливости этого предположения. Если вновь получают ошибочный результат, то предположение подтверждается.

## РЕШЕНИЯ УПРАЖНЕНИЙ ГЛАВЫ VI

1) а) Для  $x = 2$  делением на 9 находим 6 вместо 8, результат ложный, точное значение 609.

б) Делением на 4 получаем верный результат. Делением на 5 получаем верный результат. Рассматриваемый результат точен. Деление на 9 в этом случае не эффективно, так как  $x^2$  кратно 9.

2) Подставить

$$= \begin{cases} 0 & \text{— верно,} \\ \infty & \text{— верно,} \\ 1 & \text{— верно.} \end{cases}$$

3) В триангуляризации прибавляем справа к правой части дополнительный столбец так, чтобы сумма всех столбцов была равна нулю. Это свойство должно сохраняться на протяжении всего процесса триангуляризации.

В процессе решения составляем сумму всех уравнений. Подставляем значения неизвестных в это уравнение и смотрим, удовлетворяется ли оно.

4) Явно слишком близко к  $f(0,33)$ .

5) Встречаются такие кривые: изменение состояния; спектр лучей.

Третья кривая имеет много больше шансов соответствовать ложному вычислению или некорректному методу.

6) Забыли 5-ю точку и сдвинули следующие или взяли длину 5-го шага в два раза больше.

7) На спирали, крутящейся против часовой стрелки примерно с 7-ю точками на витке.

8) Погрешности:  $\varepsilon_1$  для шага  $h$  и  $\varepsilon_2$  для шага  $2h$ . Приближенные значения:  $I_1, I_2$ . Имеем

$$\varepsilon_1 \sim Ah^2, \quad \varepsilon_2 \sim A4h^2, \quad \varepsilon_1 \sim \frac{\varepsilon_2 - \varepsilon_1}{3} = \frac{I_1 - I_2}{3}.$$

9) а) Заставить участвовать только члены более высокой и более низкой степени, чем числитель.

б) Если имеется единственная ошибка, то вторая проверка обнаружит ее обязательно, а первая может ее не обнаружить.

10) а) Проверяется лишь правильность комбинации двух первых уравнений.

б) Проверка вносится в множество вычислений.

11) Вычислить снова непосредственно разложением в ряд значения тригонометрических дуг  $n$  градусов 30 минут.

12)

	$\Delta$	$\Delta^2$
5541	314	79
5855	393	—221
6248	471	79
6419	250	—22
6669	228	—22
6897	206	
7103		

- а) Разности явно неправильны вначале;  
 б) неверное значение 3-е; 6148 дало бы

5541	314	—21
5855	.293	—22
6148	271	—21
6419	250	—22
6669	228	—22
6897	206	
7103		

- с) Нет, так как они не могут быть более регулярными (правильными), чем являются сейчас.

## РЕШЕНИЯ ЗАДАЧ ГЛАВЫ VI

**Задача 1.** а) Деление на 4 дает верный результат. Деление на 5 дает верный результат. Деление на 9 дает неверный результат (3 и 6). Деление на 7 дает неверный результат (3 и 4). Точный результат  $+20 \times 129$ .

б) Деление на 4 и 5 не вскрывает ошибку в знаке.

**Задача 2.** а) Да для  $\alpha$ , нет для  $\beta$ .

б) Два уравнения почти пропорциональны. Всякая система значений, удовлетворяющая первому уравнению, будет также почти удовлетворять второму.

Единственными ошибками, обнаруживаемыми подстановкой, являются те, которые происходят от конечного определения  $x$ .

## I. СВЕДЕНИЯ О СРЕДСТВАХ ВЫЧИСЛЕНИЙ

Знание о средствах вычислений необходимо для достижения хорошей производительности в вычислениях; здесь мы приведем только некоторые сведения о наиболее распространенных средствах счета.

**1.1. Ручной счет.** Вычисления вручную очень медленны и довольно утомительны. Для ответственного счета они уже не используются. Заметим, что продолжительность одного сложения пропорциональна длине наименьшего из двух членов, а время одного умножения пропорционально произведению длин двух множителей.

Некоторые этапы вычислений возможно проводить вручную. Основной этап — это сложение  $n$  положительных чисел единственной операцией. (Это единственная возможность проведения такого вычисления.)

При ручном счете очень часто возникают ошибки. Допускается, что тренированный вычислитель совершает одну ошибку на 500 операций.

К наиболее распространенным ошибкам относятся недостаточное внимание к расстановке запятых, ошибки при записи, в частности, перестановка цифр (85 вместо 58) и неправильное повторение одной цифры (667 вместо 677).

Это все очень важно, так как даже при автоматизированном вычислении остаются процессы ручного воспроизведения. Эти процессы и подвержены именно тем ошибкам, которые указаны выше.

---

\*) Существенный недостаток этой главы заключается в том, что автор ничего не говорит в ней о средствах вычислений, которые играют в настоящее время наиболее важную роль, — об электронно-вычислительных машинах. Это тем более досадно потому, что предыдущие главы давали материал для такого разговора. Отмеченный недостаток нельзя исправить примечаниями и редактированием. (Прим. ред.)

**1.2. Вычисления на счетной линейке.** Счет на линейке требует большого внимания при прочтении чисел, а следовательно, является очень утомительным. Его трудно рассматривать иначе, чем как эпизодическую возможность. Счетная линейка — это полный и очень гибкий инструмент, она может при умелом использовании оказать очень большую помощь. (Например, она позволяет единой операцией находить корни уравнения второй степени.)

Приведем приблизительное время, затрачиваемое на операции (включая запись результата): умножение — 25 секунд, тройное правило — 35 секунд.

**У п р а ж н е н и е 1.** Удобно ли вычислять на счетной линейке значение многочлена по схеме Горнера?

**2.1. Вычисления на настольном арифмометре.** Машины, о которых идет речь, являются машинами на четыре операции без распечатки.

Существуют арифмометры ручные, а также, более распространенные, электромеханические. Не так давно стали появляться электронные арифмометры.

Арифмометры, как ручные, так и электрические, надежны и экономичны (мало поломок, минимальный уход, длительная служба).

Эти машины вследствие своего принципа действия позволяют производить интересные цепочки вычислений, такие, как

$$\frac{aa' + bb' + cc' + \dots}{d} \cdot$$

Некоторые из них имеют усовершенствования, такие, как перенос тотализатора на клавиатуру, возможность находить значения квадратного корня. Но польза этих усовершенствований может оказаться иллюзорной, если учитывать тот факт, что стоят они могут довольно дорого.

Ошибки, не происходящие по вине оператора, чрезвычайно редки.

Время операций фиксировано при сложении, пропорционально числу цифр множителя при умножении, пропорционально числу цифр частного при делении.

Можно допустить, что для достаточно совершенной электрической машины темпы могут быть следующими:  $n_i$  чисел по  $i$  цифр каждое.

$$\begin{aligned} \text{Сложение} & - n_3 + n'_3 - 10 \text{ секунд,} \\ & - n_{10} + n'_{10} - 10 \text{ секунд,} \end{aligned}$$

$$\begin{aligned} \text{умножение} & - n_3 \times n_3 - 12 \text{ секунд,} \\ & - n_{10} \times n'_{10} - 30 \text{ секунд,} \\ \text{тройное правило} & - \frac{n_3 \times n'_3}{n_3} - 24 \text{ секунды,} \\ & - \frac{n_{10} \times n'_{10}}{n_{10}} - 50 \text{ секунд.} \end{aligned}$$

**2.2. Замечание о десятичных знаках.** При ручном счете лишние десятичные знаки могут оказаться весьма дорогостоящими. Пусть, например, работа содержит сильную пропорцию умножений и делений. Переход от 5-и цифр к 6-и увеличивает время вычисления вручную в 36/25 раза, т. е. возрастает на 44%.

Имеется довольно распространенное мнение, что при переходе к счету на арифмометре лишние десятичные знаки ничего не стоят. В действительности, при тех же условиях, которые были приведены выше, время счета на арифмометре при переходе от 5-го к 6-му знакам возрастает на 20%.

**У п р а ж н е н и е 2.** Можно ли модифицировать метод Гаусса (решения системы алгебраических уравнений первой степени) так, чтобы воспользоваться особенностями арифмометра?

**3.1. Таблицы.** Таблицы часто встречающихся функций являются необходимым вспомогательным материалом при вычислениях вручную или на арифмометре. Никким образом не может стоять вопрос о повторении каждый раз вычисления значений тригонометрических дуг или степеней, в которых возникла необходимость.

Для большинства часто используемых величин имеются отличные таблицы. Мы вернемся к этому вопросу в следующей главе.

**4.1. Машина с программой.** Машины, о которых говорилось выше, осуществляют только сами операции, а последовательность операций проводится оператором. Машины с программой принимают на себя совокупность работ.

Но зато они требуют обучения одному или нескольким языкам, записи и введения программ, операторов и перфораторщиков, задержки различных ответов в соответствии со способом эксплуатации машины.

Если мы должны воспользоваться при расчетах табличными значениями функций, то часто экономичнее за-

ставить машину их вычислить, чем вводить эти значения в машину или сохранять их в памяти машины, так как продолжительность операций с учетом емкости машины не зависит от числа используемых цифр.

Ошибки в вычислениях на электронных машинах существенным образом являются ошибками человека: это ошибки программирования, ошибки перфораций, ошибки данных.

**5.1. Графические и аналоговые методы.** Графические и аналоговые методы имеют много общих черт, которые можно сгруппировать следующим образом:

— слабая точность и недостаточно четкий порядок величины погрешности;

— невозможность исследовать одновременно величины сильно отличающихся друг от друга порядков;

— возможность решать очень элегантно некоторые точные задачи, и напротив, весьма затруднительно другие, близкие задачи (иными словами, отсутствие универсальности.) Например, аналоговые методы легко решают задачу

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0,$$

и не могут решить задачу

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y).$$

Графические методы отличаются от аналоговых (по крайней мере те, которые мы имеем в виду) ролью оператора. В графических методах оператор вмешивается, чтобы самому удостовериться в исполнении и в связи между отдельными частями работы. Это связано с вычислением на арифмометре. В аналоговых методах оператор вмешивается лишь для того, чтобы констатировать конечный результат. Это связано с вычислением на машине с программой.

**5.2. Базовые операции в графических методах.** В графических методах оперируют с точками и кривыми. Базовая операция есть пересечение, численный поиск которого является очень тяжелой задачей. Этим объясняется простота некоторых графических решений.

**П р и м е р.** Решение системы

$$f(x_1, y) = 0, \quad g(x_1, y) = 0.$$

**5.3. Иррациональный характер различных графических операций.** Некоторые операции, носящие описательный уровень, очень трудны для аналитического исполнения, а подчас и не имеют никакого точного смысла. Например:

— Найти центр малого треугольника, образованного тремя почти совпадающими прямыми. Найти точку в треугольнике, вообще говоря, будет легко, например, его центр тяжести. Но это уже очень тяжело, если треугольник задан своими сторонами, а не вершинами.

— Пересечь правильную кривую, проходя наилучшим образом между  $n$  точками. Можно предложить различные способы, чтобы придать смысл этой задаче. Все они будут очень трудными.

**5.4. Неспособность графических методов действовать с объектами более двух измерений.** Невозможность использовать в графических методах геометрические объекты более двух измерений бросается в глаза. Однако небесполезно привести здесь некоторые иллюстрации этого факта.

Существуют графические методы решения линейных систем. Эти решения далеки от того, чтобы быть интуитивными. Графическое интегрирование уравнения

$$y' = Y(y, t)$$

является очень красивой задачей. А графическое интегрирование системы

$$y' = Y(y, z, t), z' = Z(y, z, t)$$

практически невозможно.

**5.5. Необратимый характер графических методов.** Графические методы состоят в начертании на одном листе бумаги некоторого числа линий. Новые линии переплетаются со старыми, делая трудным возврат к прежней стадии работы, и например, к исправлению ошибки.

## II. СОВЕТЫ К ВЫПОЛНЕНИЮ ВЫЧИСЛЕНИЙ

**6.1. Общие советы.** Математика — это работа, а не развлечение или спорт. Чтобы проделывать вычисления, надо прежде всего удобно расположиться. Если чувствуете себя усталыми, надо остановиться или, если это возможно, удвоить внимание. И ни под каким предлогом не надо нервничать.

**6.2. Полезный эффект.** Хорошая производительность должна быть существенным стремлением, даже в школьной среде.

Этого можно достигнуть прочными знаниями используемого теоретического материала и материала технического, хорошей организацией труда.

**7.1. Несколько практических советов.** Работать следует на достаточно больших листах бумаги, пронумерованных, заполняя их только с одной стороны (другую сторону можно потом использовать для другой работы). Это дает возможность иметь перед глазами в любой момент общий результат уже проделанной работы.

Наносить на листы указания так, чтобы их можно было легко найти (даже после достаточно долгого перерыва).

Использовать разумные обозначения (в частности, те, которые использовались в исходных данных).

Располагать данные, вычисления и результаты по возможности правильным образом. Приведем несколько примеров.

Во многих вопросах косинус идет впереди синуса. Значит, всегда надо стараться начинать с него, даже если для этого придется отказаться от определенных привычек.

В системе линейных алгебраических уравнений нужно стремиться записывать неизвестные в определенном порядке; строка означает уравнение, столбец содержит неизвестное.

Следует избегать, например, такой записи:

$$x + y = a,$$

$$y + z = b,$$

$$x + z = c;$$

надо записывать

$$x + y = a_x,$$

$$y + z = b_x,$$

$$x + z = c.$$

Следует избегать бесполезного переписывания.

Часто сложение положительных чисел записывается в строку. Не надо производить бесполезные операции — достаточно сложить их, исходя из записи в строку.

Таким же образом можно производить вычитание, умножение или деление на число, состоящее из одной цифры.

У п р а ж н е н и е 3. а) Сложить в строку числа с основанием 2:

$$11011 + 101 + 100011.$$

б) Произвести деление чисел с основанием 10:

$$148315 : 7.$$

**8.1. Стандартные алгоритмы.** Некоторые операции, которые часто производятся, полезно представить в стандартной форме. Это позволяет улучшить эффективность и уменьшить возможность возникновения ошибок.

Рассмотрим, например, действие с многочленами. Вместо того, чтобы записывать члены суммы многочленов в беспорядке, можно расположить их в строки и столбцы: строки — многочлены, столбцы — степени.

Чтобы при помощи этой техники произвести умножение многочленов, нет необходимости выполнять все промежуточные операции.

У п р а ж н е н и е 4. а) Выполнить умножение многочлена

$$7x^3 + 8x^2 - 5x + 4 \text{ на } 3x^2 + 2x + 6,$$

не производя всех обычных операций.

б) Проверить полученный результат.

**9.1. Организация вычислений.** Если нужно провести важное вычисление, и если оно, в частности, содержит повторяющиеся фазы, следует прежде чем начинать, организовать его, в частности, установить порядок, в котором будут производиться операции, зафиксировать число цифр, которые будут учитываться, места, где будут отмечаться промежуточные результаты, проверки, которые будут производиться в процессе работы. Для этого часто проводят модельное вычисление, т. е. подробно изучают некоторые типичные случаи.

В течение этого модельного вычисления не стремятся к эффективности и записывают все промежуточные результаты. При этом обдумывают, какой порядок величин возможен, а также различные обстоятельства, которые могут возникнуть при различных данных, которые будут использоваться.

**9.2. Пример.** Пусть требуется вычислить таблицу с двойным входом (около 20 значений  $u$  и 20 значений  $w$ ) для

Таблица 1

$4\pi^2 w^2$	$u^2$					
	0	0,01	0,04	0,09	0,16	0,25
0	0	0,01	0,04	0,09	0,16	0,25
1,57914	1,57914	1,58914	1,61914	1,66914	1,73914	1,82914
6,31655	6,31655	6,32655	6,35655	6,40655	6,47655	6,56655
14,21223	14,21223	14,22223	14,25223	14,30223	14,37223	14,46223
25,26618	25,26618	25,27618	25,30618	25,35618	25,42618	25,51618
39,47841	39,47841	39,48841	39,51841	39,56841	39,63841	39,72841

Таблица 2

	$u^2$
$4\pi^2 w^2$	$D_1$

Таблица 3

	$\operatorname{ch} 2u$
$\cos 4\pi w$	$D_1 D_2$

Таблица 4

	$u \operatorname{sh} 2u$
$2\pi w \sin 4\pi w$	$\frac{N_1}{D_1 D_2}$

Таблица 5

	$u$
$4\pi w$	$N_2$

Таблица 6

	$\frac{\operatorname{sh} 2u}{2u}$
$\frac{\sin 4\pi w}{4\pi w}$	$\frac{N_1 N_2}{D_1 D_2}$

выражения

$$\frac{u \operatorname{sh} 2u - 2\pi w \sin 4\pi w}{(u^2 + 4\pi^2 w^2)(\operatorname{ch} 2u + \cos 4\pi w)} + \frac{i4\pi w \left( \frac{\sin 4\pi w}{4w} + \frac{\operatorname{sh} 2u}{2u} \right)}{(u^2 + 4\pi^2 w^2)(\operatorname{ch} 2u + \cos 4\pi w)},$$

которое мы запишем более компактно:

$$\frac{N_1}{D_1 D_2} + i \frac{N_2 N_3}{D_1 D_2}.$$

(Заметим, что замена тригонометрических или гиперболических функций степенями переменных  $u$  и  $w$  ничего не дает.)

Мы располагаем необходимыми таблицами и настольным арифмометром с 4-мя действиями. Нам нужно в конечном результате иметь три цифры. Для этого в промежут-

точных вычислениях мы должны брать 5 цифр ( $u = 0,1, 0,2, \dots, 0,5$ ;  $w = 0,1, 0,2, \dots, 0,5$ ).

Приведем удобное расположение счета в виде 5 таблиц. Первая таблица дает  $D_1 = u^2 + 4\pi^2 w^2$ . Она имеет следующий вид (табл. 1), который мы схематизируем посредством табл. 2. Остальные таблицы имеют схематизированный вид, изображенный табл. 3—6 (в которых по одному разу используются: в табл. 3 — табл. 2, в табл. 4 — табл. 3, в табл. 6 — табл. 3 и 5).

## РЕШЕНИЯ УПРАЖНЕНИЙ ГЛАВЫ VII

1) Производим умножения на линейке и складываем результаты, начиная со старших членов. Располагаем  $x$  на шкале линейки.

2) Находим коэффициенты линейных комбинаций, располагая нули в следующем порядке: нуль на 2-й строке, нуль на 3-й строке и т. д.

В процессе решения используется формула, которая дает явно каждое неизвестное как функцию неизвестных, найденных перед ней.

3) а) 1 000 101; б) 21 187, остаток 6.

З а м е ч а н и е. Для этих операций удобно разделение на триады.

$$\begin{array}{r}
 4) \text{ а) } 21x^5 + 24x^4 - 15x^3 + 12x^2 \\
 \quad \quad \quad 14x^4 + 16x^3 - 10x^2 + 8 \\
 \quad \quad \quad \quad \quad 42x^3 + 48x^2 - 30x + 24 \\
 \hline
 21x^5 + 38x^4 + 43x^3 + 50x^2 - 22x + 24
 \end{array}$$

б) Положить  $x = 1$ .

Мы предполагаем привести здесь обзор практических аспектов таблиц.

**1.1. Общий характер табличных вычислений.** Среди работ по численному счету построение таблиц отличается своим универсальным характером. Вычисление сопротивления материалов может быть вычислено только в некоторой определенной точке, гидродинамическое вычисление может быть проведено для течения лишь в некотором канале. Напротив, таблицы бесселевых функций могут использоваться всюду.

Следствием универсальности является высокий требуемый стандарт качества. В самом деле, в вычислении, предназначенном для вполне определенного использования, всегда возможно, при помощи диалога между вычислителем и пользователем, оценивать и улучшать результаты. Для таблиц же, очевидно, это невозможно.

Высокий стандарт качества влечет за собой длительность жизни таблиц. Хорошо созданные таблицы могут оставаться полезными 50 или даже 100 лет. Какой другой научный труд может рассчитывать на такую долгую жизнь?

**1.2. Древний характер таблиц.** Уже в XVI веке мы встречаем важные таблицы. Например, Ретикус (1514—1576) вычислил тригонометрические функции с 10-ю десятичными знаками с интервалом в 10 секунд. Его труд был опубликован в 1596 году под названием «Opus Palatinum de Triangulis».

В 1613 г. Питикус опубликовал работу «Thesaurus Mathematicus», дополнив вычисления Ретикуса, содержащие синусы с интервалом в 10 секунд с 15 десятичными знаками.

Эти первые таблицы отвечали очевидному стремлению: осуществить, раз и навсегда, вычисления, в высшей степени трудоемкие, и предложить результаты в распоряжение ученых.

**1.3. Историческая эволюция.** С развитием приложений науки появилась необходимость в практическом использовании таблиц. Таблицы стали более разнообразными. Например, для тригонометрических функций стали появляться таблицы относительно градусов, минут и секунд, градусов с десятичным делением, в радианах.

Появляются таблицы, предназначенные для облегчения проведения операций (квадратов, кубов) (Барлоу, 1814).

Появляются также таблицы, приспособленные для использования в конкретных отраслях (навигация, финансовое дело и т. д.)

**2.1. Современное состояние и тенденции.** Появление настольных арифмометров, перфорационных машин и машин с программами привело к глубоким изменениям.

По мере того как распространялись арифмометры, таблицы умножения, таблицы обратных величин, квадратов, логарифмов переходили в разряд устаревших инструментов.

Появление машин с программами привело вначале к бурному расцвету процесса составления таблиц.

Но в следующем периоде пришли к выводу, что для машины проще вычислить заново необходимые значения, чем обращаться к таблицам, заложенным в памяти машины, и интерполировать их.

Третий период, скорее предполагаемый, чем жизненный, состоит в том, чтобы считать, что большие машины с программным обеспечением могут полностью заменить использование таблиц благодаря системе диалога. Можно было бы спросить машину, чему равно значение  $\sin 1,3427$ . Она должна была бы остановить текущую работу, чтобы вычислить это значение и выдать ответ на этот вопрос.

Системы-диалоги были введены под названием терминалов, но их использование с указанной выше целью не является рентабельным, так что таблицы сохраняют, по крайней мере в настоящее время, свое значение.

**3.1. Целесообразность составления таблицы.** В каком случае полезно составить таблицу? Мы различаем несколько аспектов.

а) Результаты таблицы не опубликованы. Например, могут потребоваться: 100 000 первых десятичных знаков числа, или простые числа до  $2 \cdot 10^7$ , или делители чисел вида  $1 + n^4$  для  $n \leq 1000$ .

Ответ на эти вопросы может быть дан только после выполнения самого вычисления. Почти в такой же ситуации мы окажемся, если в процессе работы получаем новую функцию. Чтобы ее эффективно использовать, нужно знать ее значения, что, естественно, приводит к вычислению таблицы. Если эта функция представляет общий интерес, то не вызывает сомнения, что эта таблица будет полезна многим, и что даже пользователи больших машин воспользуются ею для более близкого ознакомления с особенностями этой функции.

Результаты таблицы уже известны или легко могут быть получены, но мы хотим иметь более удобное их представление. Например, пользуясь тригонометрическими таблицами углов, выраженных в градусах, можно найти синус угла, выраженного в радианах. Но, разумеется, хотелось бы иметь таблицу для углов в радианной мере. Точно так же инженеры, моряки, авиаторы предпочитают иметь в удобном для них формате таблицы численных данных, которые им необходимы, с надлежащей точностью, а не работать со всей библиотекой. Этим объясняется широкое распространение малых таблиц элементарных функций.

**3.2. Пределы табулирования.** Основной преградой для табулирования служит использование функций, содержащих много параметров. Если возможно табулировать вполне определенную функцию одного переменного, то уже значительно более затруднительно табулировать функцию, зависящую от параметра или функцию двух переменных, или же функцию комплексного переменного. И почти невозможно табулировать функции, содержащие два или более параметров.

**3.3. Рентабельность таблицы.** Публикация небольшой таблицы, предназначенной для использования в определенной сфере деятельности, часто очень рентабельна. Ситуация резко меняется в отношении вычисления и издания многих важных таблиц.

Что касается их вычисления, вопрос не является столь уж тяжелым. Большинство важных таблиц было вычислено либо отдельными специалистами, либо большими научными коллективами, располагающими соответствующими возможностями.

Рентабельность издания таблицы есть проблема неразрешимая. Имеется много очень серьезных неизданных таблиц (а значит, и не используемых).

Знаменитым примером служат таблицы Прони значений тригонометрических функций углов в градусной мере. Выполненные в связи с введением метрической системы в сотрудничестве с самыми знаменитыми учеными той эпохи, они содержат значения тригонометрических функций с 22-мя десятичными знаками с шагом  $10^{-4}$  градуса. Эти таблицы были созданы в 2 или 3 года, но никогда не были опубликованы. В настоящее время распространение некоторых из этих документов возможно при помощи микрофильмирования.

У п р а ж н е н и е 1. Сколько стоило бы издание таблиц Прони (синусов и косинусов), если считать на одной странице 50 строк текста (и две функции)? Печатная страница стоит 200 франков.

**4.1. Фиксирование характеристик таблицы.** При построении таблицы мы действуем так же, как и при любом важном вычислении. Сначала мы изучаем, и насколько это возможно, используем уже имеющиеся таблицы той же функции и литературу на эту тему (в частности, опечатки).

Далее мы исследуем вопрос о том, как может быть использована информация, полученная из этих документов (и в частности, вопрос о том, насколько она заслуживает доверия).

Принимая во внимание конечную цель и средства работы, а также способ печати, которым мы располагаем, мы приходим к уточнению следующих вопросов:

- табулируемые величины (основные и вспомогательные);
- шаг таблицы;
- число сохраняемых десятичных знаков;
- способ вычисления (прямой или получаемый исходя из других таблиц).

Следует воздерживаться от желания сделать слишком много, ибо это может послужить причиной того, что замысел погибнет прежде его завершения.

**4.2. Выбор табулируемых величин.** Выбор тех величин, которые мы хотим табулировать, должен быть предметом серьезных размышлений. Не следует забывать, что подчас изменение простой константы может сделать использование таблицы гораздо менее удобной.

Например, в статистике используются таблицы величин  $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ . Ясно, что таблицы величин  $e^{-x^2}$  теоретически эквивалентны, но значительно менее пригодны.

Точно так же имеются таблицы натуральных логарифмов и таблиц десятичных логарифмов, однако эти числа связаны соотношением  $\ln x = \ln 10 \lg x$ . Кроме того, пользователи обладают некоторыми привычными навыками, которые не следует нарушать, кроме как в случае крайней необходимости. Так, для малых углов пишут:

$$\lg \sin x = \lg x + \lg \frac{\sin x}{x}.$$

Необходимы очень серьезные основания, чтобы предложить пользоваться записью  $\lg \frac{x}{\sin x}$ .

**4.3. Выбор шага.** Этот выбор в высшей степени важен, так как именно им определяются:

— распространенность таблицы; разделив шаг пополам, мы автоматически удваиваем число печатных страниц, т. е. цену таблицы, а также число страниц для переворачивания в процессе их использования;

— простота интерполирования — чем мельче шаг, тем проще интерполирование.

Различаются следующие типы таблиц:

— первоначальные таблицы, в которых содержатся значения, очень удаленные друг от друга, и откуда можно вывести другие значения посредством интерполяции очень высокого порядка;

— таблицы с интерполяцией среднего порядка;

— таблицы с линейной интерполяцией;

— критические таблицы, для которых нет необходимости ни в какой интерполяции, поскольку в них функция не меняется более чем на одну единицу последнего записанного порядка.

Последний случай встречается редко. Точно так же линейная интерполяция редко встречается для таблиц с большим числом десятичных знаков.

**У п р а ж н е н и е 2.** а) Сколько входов должна содержать критическая таблица логарифмов с 5-ю десятичными знаками?

б) Сравнить с обычными таблицами.

с) Разумен ли проект такой таблицы?

**5.1. Вспомогательные величины к табулированию.** Таблица, требующая интерполирования, должна, насколько это возможно, содержать средства для этого (разности различных порядков). Для некоторых функций

могут оказаться полезными специальные процессы интерполяции.

Например, для функции  $e^x$  можно записать

$$e^{nh+\theta} = e^{nh} e^{\theta},$$

и дать очень сжатую таблицу функции  $e^{\theta}$  для  $0 \leq \theta \leq h$  ( $h$  — шаг таблицы).

Можно задаться вопросом, удобно ли пользоваться этими специальными процессами? В самом деле, кроме того случая, когда они предназначены для использования специалистами, таблицы должны быть просты в употреблении и информативны настолько, насколько это возможно.

**6.1. Использование наиболее распространенных таблиц.** Таблицы обычных функций редко пересчитываются. Гораздо более удобно использовать (и даже часто полностью копировать) прежние документы.

Использование специализированной литературы показывает, что авторы таблиц очень доверчивы к их источникам, и часто даже повторяют их ошибки.

В использовании старых таблиц мы различаем несколько случаев:

1) В нашем распоряжении имеется таблица одновременно с минимальным шагом и максимальным числом десятичных знаков. Тогда достаточно извлечь из нее некоторые значения и надлежащим образом их округлить.

2) Мы имеем таблицу с максимальным шагом и максимальным числом десятичных знаков. Тогда надо провести интерполирование или лучше подтабулирование.

Недостатком этого способа действий является тот факт, что при этом, естественно, повторяются все ошибки исходной таблицы. Стало быть, необходимо подвергнуть таблицы, которые взяты за исходные, очень серьезному контролю, и в частности, проверить все обнаруженные ошибки.

**6.2. Непосредственное вычисление.** При отсутствии документов, которыми можно воспользоваться, производится непосредственное вычисление.

Эффективно используемые для вычисления процессы — совсем не те, которыми пользуются аналитики. Например, чтобы вычислить таблицу тригонометрических функций в градусах, минутах и секундах, можно тщательно

вычислить  $\sin 1^\circ$ ,  $\cos 1^\circ$  с большим числом десятичных знаков, а уже отсюда по индукции получить

$$\begin{aligned}\sin(n+1)^\circ &= \sin n^\circ \cos 1^\circ + \sin 1^\circ \cos n^\circ, \\ \cos(n+1)^\circ &= \cos n^\circ \cos 1^\circ - \sin 1^\circ \sin n^\circ.\end{aligned}$$

Дополнить посредством подтабулирования (и даже часто — двух последовательных подтабулирований).

**6.3. Проверки.** Учитывая требования высокого качества, проверки должны быть частыми и строгими.

В частности, требуется сличить результаты с известными ранее результатами (или по крайней мере с некоторыми из них, если их слишком много). Найденные отклонения должны обсуждаться, а обнаруженные ошибки должны быть опубликованы.

**7.1. Указания относительно точности.** В идеале было бы желательно задавать для каждого значения наилучшее десятичное приближение порядка  $n$  (т. е. с минимальной погрешностью  $(1/2) \cdot 10^{-n}$ ). Это условие выполнено для очень распространенных таблиц (логарифмы с 5-ю десятичными знаками). Для других, менее используемых таблиц, оно выполняется редко.

Попытка к этому приблизиться приводит к необходимости:

— проводить вычисления с гораздо большим числом знаков, чем нужно для опубликования, например, вычисляют 13 десятичных знаков с минимальной погрешностью  $2 \cdot 10^{-12}$  (значит, 13-й знак не нужен, но его сохраняют для оценки погрешностей вычисления);

— производить строгую оценку погрешности. В самом деле, граница погрешности равна  $(1/2)10^{-n}$  и конечное округление дает погрешность, которая может в точности достигать этого значения. Впрочем, чем меньше граница совершенной погрешности, тем меньше шансов ее найти в ситуации, когда невозможно определить наилучшее приближение порядка  $n$  (гл. III, задача 6).

Всякая серьезная таблица должна была бы быть снабжена указанием интервала приближения, на котором она осуществлялась. Но это редко делается.

**У п р а ж н е н и е 3.** Мы располагаем таблицей с 7-ю десятичными знаками, из которой предполагаем извлечь (без интерполяции) таблицу с 5-ю десятичными знаками. Таблица имеет 10 000 входов, все значения с 7-ю десятичными знаками представляют собой наилучшие приближения порядка 7.

а) Во скольких случаях можно ожидать, что не будет наилучшего приближения 5-го порядка?

б) Как поступать в сомнительных случаях?

**8.1. Материальное представление.** После того как вычисления закончены, остается еще большое число деталей, которые нужно урегулировать, связанных с материальным представлением: выбор формата, качество бумаги, содержимое страницы, расположение на странице, выбор шрифтов.

Все эти вопросы на самом деле очень важны.

Совершенно необходима короткая информация о том, какие таблицы были использованы, какой способ вычисления был принят, какой интервал приближения, а также как пользоваться таблицами.

Вспомогательные величины, необходимые для интерполирования, должны быть поданы в удобной форме. Так, необходимые пропорциональные части должны быть организованы так, чтобы их можно было использовать на любой странице.

**8.2. Процесс воспроизведения.** Огромную трудность при издании таблиц представляет процесс воспроизведения готовой рукописи. Набор опасен возможностью внесения в высшей степени неприятных ошибок, ведь каждая из цифр может оказаться неверной, и она должна быть проверена. Ошибки эти абсолютно непредвиденны.

Если таблица вычислена на машине с программой, то наилучшим решением является фотографическое воспроизведение конечного результата, хотя эти цифры не так удобны для чтения, как хорошие типографские шрифты.

**З а д а ч а 1.** Предполагается создать таблицу синусов и косинусов для углов, измеряемых в часах и минутах, с 5-ю десятичными знаками.

Составить проект такой таблицы:

- а) число используемых значений переменного;
- б) расположение на страницах;
- с) интерполирование;
- д) процесс вычисления.

**З а д а ч а 2.** Требуется протабулировать функцию

$$\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

с 5-ю десятичными знаками.

- а) Где необходимо оборвать таблицу?  
 б) Какой шаг нужно ей задать, чтобы иметь возможность линейно интерполировать (с минимальной погрешностью интерполяции  $10^{-6}$ )?

### РЕШЕНИЯ УПРАЖНЕНИЙ ГЛАВЫ VIII

- 1) 10000 страниц по 200 франков составляют 2 млн франков.  
 2) а)  $10^5$ . б) В 11 раз больше входов.  
 с) Нет.  
 3) а) Для всех чисел, запись которых с 7-ю десятичными знаками оканчивается на 50, т. е. примерно для 100 чисел.  
 б) Округляем, следуя одному из обычных правил (правило Гаусса, автоматическое округление). Интервал приближения, вместо того, чтобы быть половиной единицы последнего порядка, равен  $\frac{1}{2} \cdot 1,01$  единицы.

### РЕШЕНИЯ ЗАДАЧ ГЛАВЫ VIII

З а д а ч а 1. а) 180 значений (достаточно исходить от 0 до 3 часов).

б) 6 страниц по 30 значений, или 3 страницы по 60 значений (второе решение требует специального формата).

с) Шаг в радианах равен  $\pi/720$ . Погрешность линейной интерполяции мажорируется посредством  $\frac{\pi^2}{(720)^2 8} \approx \frac{10^{-5}}{4,105}$ .

д) Значения можно извлечь из обычных таблиц. Одна минута времени равна 15 минутам дуги.

З а д а ч а 2. а)  $\frac{1}{\sqrt{2\pi}} e^{-x^2/2} = \frac{1}{2} 10^{-5}$ ,  $x \approx 4,5$ .

б)  $|f''| \leq \frac{1}{\sqrt{2\pi}}$ ,  $h \approx 4,5 \cdot 10^{-3}$ .

Практически берется  $h = 5 \cdot 10^{-3}$ .

- Автоматическое округление 75  
 Адамса формула порядка 1 неявная 103  
 — — — — явная 104  
 Алгоритм 12  
 Алфавит 38  
 Аналоговые методы 142  
 Апостериорная основная задача 118  
 Арифмометр 140  
  
 Баричесентрическая формула 88  
 Бинарное умножение 24  
 Бистек 12  
 Блок-схема 13  
 Бэкуса метаязык 40  
  
 Верхнетреугольная матрица 82  
 Веса квадратурной формулы 94  
 Восходящее сравнение чисел 20  
 Восходящие разности 89  
 Вторичная задача 118  
 Выражение без скобок 41  
 — скобочное 49  
 — — минимальное 52  
 — — общее 49  
 — — расширенное 50  
 — — строгое 51  
 Вычисления близкие 128  
  
 Гаусса метод преобразования 83  
 Графические методы 142  
  
 Декодирование 26  
 Десятичное приближение порядка  $n$  72  
 — — — — натуральное 74  
 — — — — с избытком 73  
 — — — — с недостатком 72  
  
 Емкость 23  
  
 Задача вторичная 118.  
 — основная 118  
 Замыкание 130  
 Запись без знака 16  
 — с плавающей запятой 17  
 Значимость семантическая 39  
  
 Индекс знака 52  
 Интервал приближения 67  
 Интерполяционный многочлен 91  
 Интерполяция 91  
 Информация о приближении 65  
 Источник погрешности 113  
  
 Касательной метод 102  
 — — улучшенный 102  
 Кодирование 26  
 Контроль 134  
 — полный 134  
 — частичный 134  
 Корректность синтаксическая 39  
 Коррекция ошибок 136  
  
 Лагранжа коэффициенты 87  
 — форма 87  
 Линейная проверка 126  
 Локализация ошибок 132  
  
 Мантисса 17  
 Математические таблицы 148  
 Матрица верхнетреугольная 82  
 — диагональная 82  
 — ортогональная 82  
 Машина с программой 141  
 Метаязык Бэкуса 40  
 Метод Адамса неявный 103  
 — — явный 104  
 — касательной 102  
 — — улучшенный 102

- Метод ломаных Эйлера 102  
 — Нистрема 104  
 — трапеций 99  
 Минимальное скобочное выражение 51  
 Многочлен интерполяции 91  
  
 Наилучшее приближение порядка  $n$  75  
 — — — скользящее 77  
 Натуральная целая часть числа 73  
 Натуральное десятичное приближение 74  
 Независимое вычисление погрешности 119  
 Нейтральная погрешность 115  
 Нейтральность 116  
 Неустойчивость 104  
 Нистрема метод 104  
 Нисходящее сравнение чисел 20  
 Норма 68  
 Нормализованная запись 18  
 — — с плавающей запятой 18  
 Нотация постфиксная 55  
 — префиксная 55  
 Ньютона формула 89  
 Ньютона—Грегори формула 90  
  
 Обратный указатель 10  
 Общие скобочные выражения 49  
 Округление автоматическое 75  
 Операционная проверка 125  
 Определенная формула 97  
 Организация вычислений 145  
 Ортогональная матрица 82  
 Основная задача 118  
 Относительная погрешность 65  
 Ошибка 124  
  
 Пары скобок 47  
 Перенос погрешности 115  
 Переполнение емкости 23  
 Плавающая запись 17  
 — — нормализованная 18  
 План вычисления 119  
 — проверки 131  
 Плохо обусловленная система 85  
  
 Погрешность 64  
 — из-за инструментов 110  
 — неустраняемая 111  
 — округления 110  
 — относительная 65  
 — распространенная 114  
 Полный контроль 134  
 Понселе формула 99  
 Поправка 64  
 — интерполяции 91  
 — квадратурной формулы 97  
 Порядок 17  
 Последовательности скобок 45  
 Постфиксная нотация 55  
 Постфиксное выражение 55  
 Префиксная нотация 55  
 Приближение 64  
 — десятичное порядка  $n$  с избытком 73  
 — — — — с недостатком 72  
 — наилучшее 75  
 — с избытком 66  
 — скользящее порядка  $n$  77  
 — с недостатком 66  
 — с точностью  $\epsilon$  68  
 Приоритет умножения 43  
 Проверка 124  
 — операционная 125  
 — повтором вычисления 129  
 Программа контроля 135  
 Прямой указатель 10  
  
 Разделитель 9  
 Разности 89, 91, 128  
 Разрядная сетка 23  
 Распространенная погрешность 114  
 Расширенные скобочные выражения 50  
 Рентабельность таблицы 150  
  
 Семантика 39  
 — слева направо 42  
 — с приоритетом умножения 43  
 Семантически значимая цепочка 39  
 Симпсона формула 100  
 Синтаксис 39  
 — выражений без скобок 41  
 — языка 39  
 Синтаксически корректная цепочка 39

- Система алгебраических уравнений 81  
— плохо обусловленная 85  
Скобки 45  
Скобочные выражения 49  
— — минимальные 52  
— — общие 49  
— — расширенные 50  
— — строгие 52  
Скользящее приближение 77  
Сравнение 19  
— ускоренное 21  
Стандартный алгоритм 145  
Стек 12  
Строгие скобочные выражения 51  
Схема Горнера 14
- Таблицы 141, 148  
Тейлора формула 92  
Точность 68  
— таблицы 69  
Трапеций формула 99  
Триада 18
- Указатель 10  
Ускоренное сравнение чисел 21
- Ускоренное умножение 24
- Форма Лагранжа 87  
Формула барицентрическая 88  
— Ньютона 89  
— Ньютона—Грегори 90  
— Понселе 99  
— Симпсона 100  
— Тейлора 92  
— трапеций 99
- Целая часть числа 71  
— — — натуральная 73  
Цепочка 9
- Частичный контроль 134
- Эйлера метод ломаных 102
- Ядро интерполяционной формулы 93  
— квадратурной формулы 97  
Язык 39  
— выражений без скобок 41  
— скобок 45

**Ж. Кунцман**  
**ЧИСЛЕННЫЕ МЕТОДЫ**

---

М., 1979 г., 160 стр. с илл.

Редакторы *И. В. Викторенкова, Р. Л. Смелянский*,  
Техн. редактор *С. Я. Шкляр*,  
Корректор *Л. С. Сомова*.

**ИБ № 11303**

---

Сдано в набор 26.06.79.  
Подписано к печати 12.09.79.  
Бумага 84×108<sup>1</sup>/<sub>32</sub> тип. № 3.  
Обыкновенная гарнитура. **Высокая печать.**  
Условн. печ. л. 8,4. Уч.-изд. л. 8,22.  
Тираж 80 000 экз. Заказ № 2018 Цена книги 40 коп.

---

Издательство «Наука»  
Главная редакция физико-математической  
литературы  
117071, Москва, В-71, Ленинский проспект, 15  
2-я типография издательства «Наука»,  
Москва, Шубинский пер., 10

