Парадигма развития науки

Методологическое обеспечение

А.Е. Кононюк

ДИСКРЕТНО-НЕПРЕРЫВНАЯ МАТЕМАТИКА

Книга 4

Алгебры

(**четкие и нечеткие**)

Часть 2

Киев
Освіта України
2011


Структурная схема развивающейся панмедийной системы наук
I. Группа междисциплинарных наук
ПН Теория понятий
ЛГ Теория лингвистики
ПЗ Теория познания
СЗ Теория созидания
ПТ Общая теория понятоведения
II. Группа наук познания (фундаментальных наук)
ИН Теория информации
СТ Теория систем
РС Теория распознавания
КС Теория консалтинга (консультирования)
ПР Теория принятия решения
РР Теория реализации принятых решений
III. Группа наук созидания (прикладных наук)
СС Теория социальных систем
ЖО Теория жизнеобеспечения
КМ Теория коммуникаций и транспортирования
ОБ Теория образования (обучения)
СР Теория сохранения (защиты и охраны)
РЗ Теория развития
IV. Группа общедисциплинарных наук (обеспечивающих наук)
ДМ Теория дискретно-непрерывной математики (четкая и нечеткая)
НИ Теория научных исследований
ММ Теория моделей и моделирования
ОМ Теория оптимизации

УДК 51 (075.8)
ББК В161.я7
К 213

Рецензент: *Н.К.Печурин* - д-р техн. наук, проф. (Национальный авиационный университет).

Кононюк А.Е.
К65 Дискретно-непрерывная математика. Алгебры. К.4.Ч.2.
К.4:"Освіта України", 2011. - 668 с.
ISBN 978-966-7599-50-8

Многотомная работа содержит систематическое изложение математических дисциплин, исспользуемых при моделировании и исследованиях математических моделей систем.

В работе излагаются основы теории множеств, отношений, поверхностей, пространств, алгебраических систем, матриц, графов, математической логики, теории формальных грамматик и автоматов, теории алгоритмов, которые в совокупности образуют единную методолгически взамосвязанную математическую систему «Дискретно-непрерывная математика».

Для бакалавров, специалистов, магистров, аспирантов, докторантов и просто ученых и специалистов всех специальностей.

ББК В161.я7

ISBN 978-966-7599-50-8

Оглавление

Модуль 4.
Методы индукции и признаки делимости

Микромодуль 14.
Методы индукции

4.1. Ведение в методы индукции

Утверждения подразделяются на общие и частные.

Приведем примеры общих утверждений.

Все граждане Украины имеют право на образование.

Во всяком параллелограмме диагонали в точке пересечения делятся пополам.

Все числа оканчивающиеся нулем, делятся на 5.

Соответствующими примерами частных утверждений являются следующие:

- Петров имеет право на образование;
- в параллелограмме *ABCD* диагонали в точке пересечения делятся пополам
- 140 делится на 5.

Переход от общих утверждений к частным называется *дедукцией* Рассмотрим пример.

Все граждане Украины имеют право на образование. (1)

Петров — гражданин Украины. (2)

Петров имеет право на образование. (3)

Из общего утверждения (1) при помощи утверждения (2) получено частное утверждение (3).

Переход от частных утверждений к общим называется *индукцией.* Индукция может привести как к верным, так и к неверным выводам. Поясним это двумя примерами:

140 делится на 5. (4)

Все числа оканчивающиеся нулем, делятся на 5 (5)

Из частного утверждения (4) получено оообщее утверждение (5) Утверждение (5) верно.

140 делится на 5. (6)

Все трехзначные числа делятся на 5. (7)

Из частного утверждения (6) получено общее утверждение (7). Утверждение (7) неверно.

Спрашивается, как пользоваться в математике индукцией, чтобы получать толъко верные выводы? Ответ на этот вопрос и дается о этом микромодуле.

1. Рассмотрим сначала два примера индукции, недопустимой в математике.

Пример 1. Пусть

$$S_n = \frac{1}{1\cdot 2} + \frac{1}{2\cdot 3} + \frac{1}{3\cdot 4} + \ldots + \frac{1}{n(n+1)}.$$

Легко проверить, что

$$S_1 = \frac{1}{1\cdot 2} = \frac{1}{2},$$

$$S_2 = \frac{1}{1\cdot 2} + \frac{1}{2\cdot 3} = \frac{2}{3},$$

$$S_3 = \frac{1}{1\cdot 2} + \frac{1}{2\cdot 3} + \frac{1}{3\cdot 4} = \frac{3}{4},$$

$$S_4 = \frac{1}{1\cdot 2} + \frac{1}{2\cdot 3} + \frac{1}{3\cdot 4} + \frac{1}{4\cdot 5} = \frac{4}{5}.$$

На основании полеченных результатов утверждаем, что при всяком натуральном n

$$S_n = \frac{n}{n+1}.$$

Пример 2. Рассмотрим трехчлен $x^2 + x + 41$, на который обратил внимание еще Л. Эйлер. Подставим в этот трехчлен вместо x нуль, получим простое число 41. Подставим теперь в этот же трехчлен вместо x единицу, получим опять простое число 43. Продолжая подставлять в трехчлен вместо x последовательно 2, 3, 4, 5, 6, 7, 8, 9, 10, получаем всякий раз простое число 47, 53, 61, 71, 83, 97, 113, 131, 151. На основании полученных результатов утверждаем, чго при подстановке в трехчлен вместо x любого целого неотрицательного числа всегда в результате получается простое число.

Почему рассуждения, приведенные в этих примерах, недопустимы в математике? В чем порочность выводов, которые нами сделаны?

Дело в том, что в обоих этих рассуждениях мы высказали общее утверждение относительно любого (во втором примере относительно любого *x)* только на основании того, что это утверждение оказалось справедливым для некоторых значений n (или *x).*

Индукция широко применяется в математике, но применять ее надо умело. При легкомысленном же отношении к индукции можно получить неверные выводы.

Так, если в примере 1 сделанное нами общее утверждение случайно оказывается верным, как это доказано ниже в примере 4, то в примере 2 наше общее утверждение оказачось неверным.

В самом деле, при более внимательном изучении трехчлена x^2+x+41 обнаружили, что он равен простому числу при $x = 0, 1, 2, ..., 39$, но при $x = 40$ этот трехчлен равен 41^2, т. е. числу составному (и уж совсем сразу бросается в глаза, что при $x=41$ $x^2+ x + 41 =41^2 + 41 + 41$ делится на 41)).

2. В примере 2 мы встретились с утверждением, справедливым в 40 частных случаях и все же вообще оказавшимся несправедливым.

Приведем еще несколько примеров утверждений, которые справедливы в нескольких частных случаях, а вообще несправедливы.

Пример 3. Двучлен x^n—1, где n — натуральное число, представляет для математиков большой интерес. Достаточно сказать, что он тесно связан с геометрической задачей о делении окружности на n равных частей. Неудивительно поэтому, что двучлен этот всесторонне изучается в математике. Математиков, в частности, интересовал вопрос о разложении этого двучлена на множители с целыми коэффициентами.

Рассматривая эти разложения при многих частных значениях n, математики наблюдали, что все коэффициенты разложения по абсолютной величине своей не превосходят единицы. В самом деле,

$$x - 1 = x - 1,$$
$$x^2 - 1 = (x - 1)(x + 1),$$
$$x^3 - 1 = (x - 1)(x^2 + x + 1),$$
$$x^4 - 1 = (x - 1)(x + 1)(x^2 + 1),$$
$$x^5 - 1 = (x - 1)(x^4 + x^3 + x^2 + x + 1),$$
$$x^6 - 1 = (x - 1)(x + 1)(x^2 + x + 1)(x^2 - x + 1),$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

Были составлены таблицы, в пределах которых коэффициенты этим свойством обладали. Попытки доказать этот факт для всякое n успеха не имели.

В 1938 г. в журнале «Успехи математических наук» (вып. IV) была опубликована заметка выдающегося русского математика Н. Г. Чеботарева, в которой он предложил нашим математикам выяснить этот вопрос.

Эту задачу в 1941 г. решил В. Иванов. Оказалось, что указанным свойством обладают все двучлены $x^n - 1$, степень которых меньше 105. Одним же из множителей $x^{105} - 1$ является многочлен

$$x^{48} + x^{47} + x^{46} - x^{43} - x^{42} - 2x^{41} - x^{40} - x^{39} +$$
$$+ x^{36} + x^{35} + x^{34} + x^{33} + x^{32} + x^{31} - x^{28} - x^{26} - x^{24} -$$
$$- x^{22} - x^{20} + x^{17} + x^{16} + x^{15} + x^{14} + x^{13} +$$
$$+ x^{12} - x^9 - x^8 - 2x^7 - x^6 - x^5 + x^2 + x + 1,$$

уже не обладающий этим свойством.

Пример 4. Рассмотрим числа вида $2^{2^n}+1$. При n = 0, 1, 2, 3, 4 числа $2^{2^0}+1$ =3, $2^{2^1}+1$= 5. $2^{2^2}+1$= 17, $2^{2^3}+1$=257. $2^{2^4}+1$=65 537 — простые. Замечательный французский математик XVII в П. Ферма предполагал, что все числа такого вида — простые. Однако в XVIII в. Л. Эйлер нашел, что

$$2^{2^8}+1=4\ 294\ 967\ 297 = 641 \cdot 6\ 700\ 417$$

— составное число.

Пример 5. Знаменитый немецкий математик XVII в., один из создателей так называемой «высшей математики», Г. В. Лейбниц доказал, что при всяком целом положительном n число $n^3 - n$ делится на 3, число $n^5 - n$ делится на 5, число $n^7 - n$ делится на 7. На основании этого он предположил было, что при всяком нечетном k и любом натуральном n число $n^k - n$ делится на k, но скоро сам заметил, что $2^9 - 2 = 510$ не делится на 9.

Пример 6. В ошибку такого же рода впал однажды известный математик Д. Л. Граве, предположив, что для всех простых чисел p число $2^{p-1} - 1$ не делится на p^2. Непосредственная проверка подтвердила это предположение для всех простых чисел p, меньших тысячи. Вскоре, однако, было установлено, что $2^{1092} - 1$ делится на 1093^2 (1093— простое число), т. е. предположение Граве оказалось ошибочным.

Пример 7. На сколько частей даелят пространство n плоскостей, проходящих через одну точку, если никакие три из них не проходят через одну прямую?

Рассмотрим простейшие частные случаи этой задачи. Одна плоскость делит пространство на две части. Две плоскости, проходящие через одну точку, делят пространство на четыре части. Три плоскости, проходящие через одну точку, но не проходящие через одну прямую, делят пространство на восемь частей.

На первый взгляд может показаться, что с увеличением числа плоскостей на единицу количество частей, на которые разбивается пространство, увеличивается вдвое, и, таким образом, четыре плоскости разобьют пространство на 16 частей, пять — на 32 части, а вообще n плоскостей разобьют пространство на 2^n частей.

В действительности это не так, а именно: четыре плоскости разбивают пространство на 14 частей, пять плоскостей — на 22 части. Можно доказать, чго n плоскостей разбивают пространство на $n(n— 1) +2$ частей (решение буде приведено в примере 13).

Пример 8. Приведем еще один весьма убедительный пример. Подставляя в выражение $991n^2+1$ вместо n последовательно целые числа 1, 2, 3, ... мы никогда не получим числа, явтяющегося полным квадратом, сколько бы дней или даже лет мы ни посвятили этим вычислениям. Однако если мы сделаем отсюда вывод, что все числа такого вида не являются квадратами, то мы ошибемся. На самом деле оказывается, чго среди чисел вида $991n^2+1$ имеются и квадраты: только наименьшее значение n, при котором число $991n^2+1$ есть полный квадрат, очень велико. Вот это число:

$$n = 12\ 055\ 735\ 790\ 331\ 359\ 447\ 442\ 538\ 767.$$

Рассмотренные примеры позволяют сделать простой и в то же время важный вывод.

Утверждение может быть справедливым в целом ряде частных случаев и в то же время несправедливым вообще.

3. Теперь возникает такой вопрос. Имеется утверждение, справедчивое с нескольких частных случаях. Все частные случаи рассмотреть невозможно. Как же узнать, справедливо ли это утверждение вообще?

Этот вопрос иногда удается решить посредством применения особого метода рассуждений, называемого *методом математической индукции* (полной индукции, совершенной индукции).

В основе этого метода лежит принцип математической индукции, заключающийся в следующем:

Утверждение справедливо для всякого натурального n, если: 1) *оно справедливо для n*= 1 *и* 2) *из справедливости утверждения для какого-либо произвольного натурального n = k следует его справедливость для n = k* +1.

Доказательство. Предположим противное, т. е. предположим, что утверждение справедливо не для всякого натурального n. Тогда существует такое натуральное m, что 1) утверждение для $n = m$ несправедливо, 2) для всякого n, меньшего m, утверждение

справедливо (иными словами, *m* есть первое натуральное число, для которого утверждение несправедливо).

Очевидно, что $m>1$, так как для $n=1$ утверждение справедливо (условие 1). Следовательно, *m*—1— натуральное число. Выходит, что для натурального числа *m*—1 утверждение справедливо, а для следующего натурального числа *m* оно несправедливо. Это противоречит условию 2.

Конечно, при доказательстве принципа математической индукшш мы пользовались тем, *что в любой совокупности натуральных чисел содержится наименьшее число.* Легко видеть, что это свойство в свою очередь можно вывести как следствие из принципа математической индукции. Таким образом, оба эти предложения равносильны. Любое из них можно принять за одну из аксиом, определяющих натуральный ряд, тогда другое будет теоремой. Обычно за аксиому принимают как раз сам принцип математической индукции.

4. Доказательство, основанное на принципе математической индукции, называется доказательством *методом математической индукции.* Такое доказательство необходимо должно состоять из двух частей, из доказательства двух самостоятельных теорем:

Теорема 1. *Утверждение справедливо для* $n = 1$.

Теорема 2. *Утверждение справедливо для* $n=k+1$, *если оно справедливо для* $n = k$, *где* k *— какое-либо произвольное натуральное число.*

Если обе эти теоремы доказаны, то на основании принципа математической индукции утверждение справедливо для всякого натурального *n.*

Пример 9. Вычислить сумму (см, пример 1)

$$S_n = \frac{1}{1\cdot 2} + \frac{1}{2\cdot 3} + \frac{1}{3\cdot 4} + \cdots + \frac{1}{n(n+1)}.$$

Мы знаем, что

$$S_1 = \frac{1}{2},\quad S_2 = \frac{2}{3},\quad S_3 = \frac{3}{4},\quad S_4 = \frac{4}{5}.$$

Теперь мы не повторим ошибку, допущенную в примере 1, и не станем сразу утверждать, что при всяком натуральном *n*

$$S_n = \frac{n}{n+1}.$$

Будем осторожны и скажем, что рассмотрение сумм S_1, S_2, S_3, S_4 позволяет высказать гипотезу (предположение), что $S_n = \frac{n}{n+1}$ при всяком натуральном *n.* При этом мы знаем, что

гипотеза эта верна при n=1, 2, 3, 4. Для проверки гипотезы воспользуемся методом математической индукции.

Теорема 1. Для $n = 1$ гипотеза верна, так как

$$S_1 = \frac{1}{2}.$$

Теорема 2. Предположим, что гипотеза верна для $n = k$, т. е. что

$$S_k = \frac{1}{1\cdot 2} + \frac{1}{2\cdot 3} + \ldots + \frac{1}{k(k+1)} = \frac{k}{k+1},$$

где k — некоторое натуральное число. Докажем, что тогда гипотеза обязана быть верной и для $n = k+1$, т. е. чго

$$S_{k+1} = \frac{k+1}{k+2}.$$

Действительно,

$$S_{k+1} = S_k + \frac{1}{(k+1)(k+2)};$$

следовательно, по условию теоремы,

$$S_{k+1} = \frac{k}{k+1} + \frac{1}{(k+1)(k+2)} = \frac{k^2+2k+1}{(k+1)(k+2)} = \frac{k+1}{k+2}.$$

Обе теоремы доказаны. Теперь на основании принципа математической индукции мы утверждаем, что

$$S_n = \frac{n}{n+1}$$

при всяком натуральном n.

Замечание 1. Необходимо подчеркнуть, чго доказательство методом математической индукции безусловно требует доказательства обеих теорем, 1 и 2.

Мы уже видели, к чему привело пренебрежительное отношение к теореме 2 (пример 2).

Сейчас мы покажем, что нельзя опускать и теорему 1. Рассмотрим пример.

Пример 10. Теорема. *Всякое натуральное число равно следующему за ним натуральному числу.*

Доказательство проведем методом математической индукции. Предположим, чю

$$k=k+1. \qquad (8)$$

Докажем, что

$$k+1=k+2. \qquad (9)$$

Действительно, прибавив к каждой части равенсиза (8) по 1, получим равенство (9). Выхочит, что если утверждение справедливо для $n= k$, то оно справедливо и для $n=k+1$. Теорема доказана.

Следствие. Все натуральные числа равны.

Где же здесь ошибка? Ошибка заключается в том, чю первая теорема, необходимая для применения принципа математической индукции, не доказана и не верна, а доказана только одна вторая теорема.

Теоремы 1 и 2 имеют свое особое значение. Теорема 1 создает, так сказать, базу для проведения индукции. Теорема 2 дает право неограниченного автоматического расширения этой базы, право перехода от данного частного случая к следующему, от n к n+1.

Если не доказана теорема 1, а доказана теорема 2 (см. пример 6), то, следовательно, не создана база для проведения индукции, и тогда бессмысленно применять теорему 2, так как и расширять-то, собственно, нечего.

Если не доказана теорема 2, а доказана только теорема 1 (см. примеры 1 и 2), то, хотя база для проведения индукции и создана, право расширения этой базы отсутствует.

Замечание 2. Метод математической индукции разобран выше для простейшего случая. В более сложных случаях формулировки теорем 1 и 2 должны быть соответственно изменены.

Иногда вторая часть доказательства опирается на справедливость утверждения не только для $n = k$, по и для $n = k$ — 1. В эгом случае утверждение в первой части должно быть проверено для двух последовательных значений n.

Иногда также вторая часть доказательства состоит в установлении справедливости требуемого рассуждения для какого-то значения n в предположении справедливости его для всех натуральных чисел k, меньших n. Ниже читатель найдет примеры такого рода (см. пример 7 п. 4.2).

Иногда утверждение доказывается не для всякого натурального n, а для всякого целого n, превосходящего некоторое целое m (так, например, любое утверждение, касающееся свойств произвольных n-угольников, имеет смысл лишь при n>3). В этом случае в первой части доказательства утверждение проверяется для $n = m$+1, а если это требуется, то и для нескольких последующих значений n.

5. Вернемся еще раз к примеру 1 для выяснения одной существенной стороны метода математической индукции.

Изучая сумму

$$S_n = \frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \ldots + \frac{1}{n(n+1)}$$

при разных значениях n, мы подсчитали, что

$$S_1 = \frac{1}{2},\ S_2 = \frac{2}{3},\ S_3 = \frac{3}{4},\ S_4 = \frac{4}{5}, \ldots,$$

и это навело нас на гипотезу, что и при всяком n

$$S_n = \frac{n}{n+1}.$$

Для проверки гипотезы мы использовали метол математической индукции.

Нам повезло, мы высказали гипотезу, которая подтвердилась. Если бы мы высказали гипотезу неудачно, то порочность гипотезы обнаружилась бы при попытке доказательства теоремы 2.

Пример 11. Рассмотрим суммы

$$S_n = \frac{1}{1\cdot 2} + \frac{1}{2\cdot 3} + \ldots + \frac{1}{n(n+1)}.$$

Допустим, что, изучая S_n, мы высказали гипотезу

$$S_n = \frac{n+1}{3n+1}. \qquad (10)$$

При $n = 1$ формула (10) верна, так как $S_1 = \frac{1}{2}$. Предположим, что формула (10) верна при $n = k$, т. е.

$$S_k = \frac{k}{3k+1}$$

Попытаемся доказать, что формула (10) верна и при $n = k+1$, т. е. что

$$S_{k+1} = \frac{k+2}{3k+4}.$$

Имеем

$$S_{k+1} = S_k + \frac{1}{(k+1)(k+2)} =$$

$$= \frac{k+1}{3k+1} + \frac{1}{(k+1)(k+2)} = \frac{k^3+4k^2+8k+2}{(k+1)(k+2)(3k+1)},$$

т. е. результат получился иной.

Выходит, что из справедливости формулы (1) при $n = k$ не следует ее справедливость при $n=k+1$. Мы обнаружили, что формула (10) неверна.

Таким образом, *метод математической индукции позволяет в поисках общего закона испытывать возникающие при этом гипотезы, отбрасывать ложные и утверждать истинные.*

Для того чтобы научиться применять метод математической индукции, надо рассмотреть достаточное количество задач.

Чтобы не повторять без конца слова «Теорема 1» и «Теорема 2», мы условимся в дальнейшем помечать первую и вторую части доказательства по индукции (эти части и составляют содержание двух теорем, доказательство которых равносильно пользованию методом индукции) знаками 1° и 2°. Кроме того, мы будем различать примеры, снабженные подробными решениями, и задачи, предназначенные для самостоятельной работы читателя. В конце микромодуля будут приведены указания, относящиеся к решению всех приведенных в тексте задач. Иногда эти указания представляют собой лишь ссылку на иную, доступную читателям литературу; в других случаях они содержат полное решение задачи.

4.2. Доказательства тождеств для арифметических задач

Пример 1. Выпишем в порядке возрастания нечетные положительные числа 1, 3, 5, 7, ... Обозначим первое из них u_1, второе u_2, третье u_3 и т. д., т. е.

$$u_1 = 1, \quad u_2 = 3, \quad u_3 = 5, \quad u_4 = 7, \ldots$$

Поставим перед собой такую задачу: составить формулу, выражающую нечетное число u_n через его номер n.

Решение. Первое нечетное число u_1 можно записать так:

$$u_1 = 2 \cdot 1 - 1; \tag{1}$$

второе нечетное число u_2 можно записать так:

$$u_2 = 2 \cdot 2 - 1; \tag{2}$$

третье нечетное число u_3 можно записать так:

$$u_3 = 2 \cdot 3 - 1. \tag{3}$$

Внимательно рассматривая равенства (1), (2), (3), можно высказать гипотезу, чго для получения любого нечетного числа достаточно от удвоенного номера его отнять 1, т. е, для n-го нечетного числа имеем формулу

$$u_n = 2n - 1. \tag{4}$$

Докажем, что формула эта справедлива.

1°. Равенство (1) показывает, что для n — 1 формула (4) справедлива.

2°. Предположим, что формула (4) справедлива для $n = k$, т. е. k-е нечетное число имеет вид

$$u_k = 2k - 1.$$

Докажем, что тогда формула (4) обязана быть справедливой и для (k +1)-го нечетного числа, т. е. что (k +1)-е нечетное число имеет вид

$$u_{k+1} = 2(k + 1) - 1,$$

или, что все равно,

$$u_{k+1} = 2k + 1.$$

Для получения (k +1)-го нечетного числа достаточно к k-му нечетному числу прибавить 2, т. е.

$$u_{k+1} = u_k + 2.$$

Но, по условию, $u_k = 2k$—1. Значит,

$$u_{k+1} = (2k - 1) + 2 = 2k + 1.$$

что и требовалось доказать.

Ответ. $u_n = 2n$ — 1.

Пример 2. Вычислить сумму первых и нечетных чисел.

Решение. Обозначим искомую сумму S_n, т. е.

$$S_n = 1 + 3 + 5 + ... + (2n - 1).$$

Для решения таких задач в математике существуют готовые формулы. Нам интересно решить эту задачу, не прибегая к готовой формуле, а пользуясь методом математической индукции. Для этого прежде всего надо построить гипотезу, т. е. просто постараться угадать ответ.

Придаем n последовательно значения 1, 2, 3, ... до тех пор, пока у нас не накопится достаточно материала, чтобы на основе его построить более или менее надежную гипотезу. После этого останется только эту гипотезу проверить методом математической индукции.

Имеем

$$S_1 = 1, S_2 = 4, S_3 = 9, S_4 = 14, S_5 = 25, S_6 = 36.$$

Теперь все зависит от наблюдательности решающего задачу, от его способности по частным результатам угадать общий.

Полагаем, что в данном случае легко заметить, что

$$S_1 = 1^2, \ S_2 = 2^2, S_3 = 3^2, S_4 = 4^2.$$

На основе этого можно предположить, что вообще

$$S_n = n^2$$

Докажем, что гипотеза эта справедлива.

1°. При n=1 сумма представляется одним слагаемым, равным 1. Выражение n^2 при n = 1 также равно 1. Значит, при n=1 гипотеза верна.

2°. Допустим, что гипотеза верна для $n = k$, т. е. $S_k = k^2$. Докажем, что тогда гипотеза должна быть верна и для $n = k + 1$, т. е.

$$S_{k+1} = (k + 1)^2.$$

Действительно,

$$S_{k+1} = S_k + (2k + 1).$$

Но $S_k = k^2$ и потому

$$S_{k+1} = k^2 + (2k + 1) = (k + 1)^2,$$

что и требовалось доказать.

Ответ. $S_n = n^2$.

Пример 3. Доказать, чго сумма n первых чисел натурального ряда равна

$$\frac{n(n+1)}{2}.$$

Решение. Эта задача отличается от предыдущих тем, что гипотезу здесь строить не надо, она дана. Нужно только показать, что гипотеза верна.

Обозначим искомую сумму S_n, т. е.

$$S_n = 1 + 2 + 3 + \ldots + n.$$

1°. При $n = 1$ гипотеза верна.

2°. Пусть

$$S_k = 1 + 2 + 3 + \ldots + k = \frac{k(k+1)}{2}.$$

Покажем, что

$$S_{k+1} = \frac{(k+1)(k+2)}{2}.$$

В самом деле,

$$S_{k+1} = S_k + (k+1) = \frac{k(k+1)}{2} + (k+1) = \frac{(k+1)(k+2)}{2}.$$

Задача решена.

Пример 4. Доказать, что сумма квадратов n первых чисел натурального ряда равна $\frac{n(n+1)(2n+1)}{6}$.

Решение. Пусть $S_2(n) = 1^2 + 2^2 + 3^2 + \ldots + n^2$.

1°. $S_2(1) = 1^2 = \frac{1(1+1)(2\cdot 1+1)}{6}$.

2°. Предположим, что

$$S_2(n) = \frac{n(n+1)(2n+1)}{6}.$$

Тогда

$$S_2(n+1) = 1^2 + 2^2 + 3^2 + \ldots + n^2 + (n+1)^2 =$$

$$= \frac{n(n+1)(2n+1)}{6} + (n+1)^2$$

и окончательно

$$S_2(n+1) = \frac{(n+1)[(n+1)+1][2(n+1)+1]}{6}.$$

Пример 5. Доказать, что

$$S_n = 1 - 2^2 + 3^2 - 4^2 + \ldots + (-1)^{n-1} n^2 = (-1)^{n-1} \frac{n(n+1)}{2}.$$

Решение. 1°. При n=1 гипотеза, очевидно, верна ($(-1)^0$=1).

2°. Пусть

$$S_k = 1 - 2^2 + 3^2 - \ldots + (-1)^{k-1} k^2 = (-1)^{k-1} \frac{k(k+1)}{2}.$$

Докажем, что

$$S_{k+1} = 1 - 2^2 + 3^2 - \ldots + (-1)^{k-1} k^2 + (-1)^k (k+1)^2 = (-1)^k \frac{(k+1)(k+2)}{2}.$$

Действительно,

$$S_{k+1} = S_k + (-1)^k (k+1)^2 = (-1)^{k-1} \frac{k(k+1)}{2} + (-1)^k (k+1)^2 =$$

$$= (-1)^k \left[(k+1) - \frac{k}{2}\right](k+1) = (-1)^k \frac{(k+1)(k+2)}{2}.$$

Пример 6. Доказать, что

$$1 \cdot 2 + 2 \cdot 3 + 3 \cdot 4 + \ldots + (n-1)n = \frac{(n-1)n(n+1)}{3}.$$

Решение. 1°. $1 \cdot 2 = \frac{1 \cdot 2 \cdot 3}{3}$.

2°. Если

$$1 \cdot 2 + 2 \cdot 3 + 3 \cdot 4 + \ldots + (n-1)n = \frac{(n-1)n(n+1)}{3},$$

то

$$1 \cdot 2 + 2 \cdot 3 + 3 \cdot 4 + \ldots + (n-1)n + n(n+1) = \frac{(n-1)n(n+1)}{3} + n(n+1) = \frac{n(n+1)(n+2)}{3}.$$

Пример 6 можно также вывести из результатов примеров 3 и 4, если заметить, что

$$1 \cdot 2 + 2 \cdot 3 + 3 \cdot 4 + \ldots + (n-1)n =$$
$$= 1(1+1) + 2(2+1) + 3(3+1) + \ldots$$
$$\ldots + (n-1)[(n-1)+1] = [1^2 + 2^2 + \ldots + (n-1)^2] +$$
$$+ [1 + 2 + \ldots + (n-1)].$$

Пример 7. Доказать, что если $v_0 = 2$, $v_1 = 3$ и для всякого натурального k имеет место соотношение

$$v_{k+1} = 3v_k - 2v_{k-1},$$

то

$$v_n = 2^n + 1.$$

Решение. 1°. Для $n = 0$ и $n = 1$ утверждение справедливо по условию.

2°. Предположим, чго

$$v_{k-1} = 2^{k-1} + 1; \quad v_k = 2^k + 1.$$

Тогда

$$v_{k+1} = 3(2^k+1) - 2(2^k + 1) = 2^{k+1}+1.$$

Пример 8. Произведение $1 \cdot 2 \cdot 3 \ldots n$ обозначается знаком $n!$ и читается так: *«n факториал»*. Полезно запомнить, что 1!=1, 2! = 2, 3! = 6, 4! = 24. 5!= 120.

Вычислить

$$S_n = 1 \cdot 1! + 2 \cdot 2! + 3 \cdot 3! + \ldots + n \cdot n!$$

Решение.

$$S_1 = 1 \cdot 1! = 1, \quad S_2 = 1 \cdot 1! + 2 \cdot 2! = 5,$$
$$S_3 = 1 \cdot 1! + 2 \cdot 2! + 3 \cdot 3! = 23,$$
$$S_4 = 1 \cdot 1! + 2 \cdot 2! + 3 \cdot 3! + 4 \cdot 4! = 119.$$

Присматриваясь к этим результатам, можно заметить, что

$$S_1 = 2! - 1, \quad S_2 = 3! - 1, \quad S_3 = 4! - 1, \quad S_4 = 5! - 1.$$

Это даст возможность высказать гипотезу, что

$$S_n = (n+1)! - 1.$$

Проверим эту гипотезу.

1°. Для n—1 гипотеза верна, так как

$$S_1 = 1 \cdot 1! = 2! - 1.$$

2°. Пусть

$$S_k = 1 \cdot 1! + 2 \cdot 2! + \ldots + k \cdot k! = (k+1)! - 1.$$

Покажем, что

$$S_{k+1} = 1 \cdot 1! + 2 \cdot 2! + \ldots + k \cdot k! + (k+1)(k+1)! = \\ = (k+2)! - 1.$$

Действительно,

$$S_{k+1} = S_k + (k+1)(k+1)! = [(k+1)! - 1] + \\ + (k+1)(k+1)! = (k+1)![1 + (k+1)] - 1 = \\ = (k+1)!(k+2) - 1 = (k+2)! - 1.$$

Пример 9. Дано:

$$\alpha + \beta = m, \quad \alpha\beta = a, \quad A_2 = m - \frac{a}{m-1},$$

$$A_3 = m - \frac{a}{m - \dfrac{a}{m-1}}, \quad A_4 = m - \frac{a}{m - \dfrac{a}{m - \dfrac{a}{m-1}}} \text{ и т. д.,}$$

т е. для $k > 1$

$$A_{k+1} = m - \frac{a}{A_k} \quad (m \neq 1; \quad \alpha \neq \beta).$$

Доказать, что

$$A_n = \frac{(\alpha^{n-1} - \beta^{n-1}) - (\alpha^n - \beta^n)}{(\alpha^{n-1} - \beta^{n-1}) - (\alpha^{n-2} - \beta^{n-2})}. \qquad (1)$$

Решение. 1°. Докажем сначала, что формула (1) верна для $n = 2$. По условию,

$$A_2 = m - \frac{a}{m-1} = (\alpha + \beta) - \frac{\alpha\beta}{(\alpha+\beta)-1} = \frac{\alpha^2 + \beta^2 + \alpha\beta - \alpha - \beta}{\alpha + \beta - 1}.$$

По формуле (1)

$$A_2 = \frac{(\alpha^3 - \beta^3) - (\alpha^2 - \beta^2)}{(\alpha^2 - \beta^2) - (\alpha - \beta)}.$$

Сократив последнюю дробь на α–β, имеем

$$A_2 = \frac{\alpha^2 + \beta^2 + \alpha\beta - \alpha - \beta}{\alpha + \beta - 1},$$

что и требовалось доказать.

2°. Пусть формула (1) справедлива для $n = k$, т. е.

$$A_k = \frac{(\alpha^{k+1} - \beta^{k+1}) - (\alpha^k - \beta^k)}{(\alpha^k - \beta^k) - (\alpha^{k-1} - \beta^{k-1})}. \qquad (2)$$

Докажем, что тогда она должна быть справедлива и для $n=k+1$. т е

$$A_{k+1} = \frac{(\alpha^{k+2} - \beta^{k+2}) - (\alpha^{k+1} - \beta^{k+1})}{(\alpha^{k+1} - \beta^{k+1}) - (\alpha^{k} - \beta^{k})}.$$

Действительно,

$$A_{k+1} = m - \frac{\alpha}{A_k} \quad \text{или} \quad A_{k+1} = (\alpha + \beta) - \frac{\alpha\beta}{A_k}.$$

Пользуясь равенством (2), имеем

$$A_{k+1} = (\alpha + \beta) - \frac{\alpha\beta\,[(\alpha^{k} - \beta^{k}) - (\alpha^{k-1} - \beta^{k-1})]}{(\alpha^{k+1} - \beta^{k+1}) - (\alpha^{k} - \beta^{k})} =$$

$$= \frac{(\alpha^{k+2} - \beta^{k+2}) - (\alpha^{k+1} - \beta^{k+1})}{(\alpha^{k+1} - \beta^{k+1}) - (\alpha^{k} - \beta^{k})}.$$

Теорема доказана.

Пример 10. Доказать, что любое целое число рублей, большее 7, можно уплатить без сдачи денежными билетами, достоинством в 3 и 5 рублей.

Решение. 1°. Для 8 рублей утверждение справедливо (ибо 8 руб. = 3 руб. + 5 руб.).

2°. Пусть утверждение верно для k рублей, где k — целое число, большее или равное 8.

Возможны два случая: 1) k рублей уплачивается одними трехрублевыми билетами и 2) k рублей уплачивается денежными билетами, среди которых есть хоть один билет пятирублевого достоинства.

В первом случае трехрублевых билетов должно быть не менее трех, так как в этом случае $k > 8$. Для того чтобы уплатить k +1 рубль, заменим три трехрублевых билета двумя пятирублевыми.

Во втором случае для уплаты k + 1 рубля заменим один пятирублевый билет двумя трехрублевыми.

Пример 11. Доказать, чго сумма кубов трех последовательных натуральных чисел делится на 9.

Решение. 1°. Сумма $1^3+2^3+3^3$ делится на 9. Значит, утверждение справедливо, когда первым из трех последовательных натуральных чисел является 1.

2°. Пусть сумма $k^3+ (k + 1)^3+ (k + 2)^3$, где k — некоторое натуральное число, делится на 9. Сумма

$$(k+1)^3 + (k+2)^3 + (k+3)^3 =$$

$$= (k+1)^3 + (k+2)^3 + k^3 + 9k^2 + 27k + 27 =$$

$$= [k^3 + (k+1)^3 + (k+2)^2] + 9(k^2 + 3k + 3)$$

представляет собой сумму двух слагаемых, каждое из которых делится на 9, а потому тоже делится на 9.

Пример 12. Из $2n$ чисел 1, 2..... $2n$ произвольно выбрали $n+1$ число. Доказать, что среди выбранных чисел найдутся хотя бы два числа, из которых одно делится на другое.

Решение. 1°. Для двух чисел 1, 2 утверждение справедливо.

2°. Допустим, что из $2n$ чисел 1, 2.....$2n$, где $n \geq 2$,
удалось выбрать так $n + 1$ число, что ни одно из них не делится на другое. Совокупность всех этих чисел обозначим для краткости M_{n+1} Докажем, что тогда из $2n$ — 2 чисел 1, 2,.....,$2n$ — 2 можно выбрать n чисел таких, что опять ни одно из них не будет делиться на другое.

Возможны четыре случая:

1) M_{n+1} не содержит ни $2n$—1, ни $2n$.
2) M_{n+1} содержит $2n$—1 и не содержит $2n$.
3) M_{n+1} содержит $2n$ и не содержит $2n$ — 1.
4) M_{n+1} содержит и $2n$—1, и $2n$.

Случай 1. Исключим из M_{n+1} какое-нибудь число. Останется n чисел, каждое из которых не больше, чем $2n$ —2. Ни одно из этих чисел не делится на другое.

Случай 2. Исключим из M_{n+1} число $2n$—1. Останется n чисел, каждое из которых не больше, чем $2n$ — 2. Ни одно из этих n чисел не делится на другое.

Случай 3. Исключим из M_{n+1} число $2n$ и опять получим тот же результат.

Случай 4. Прежде всего заметим, что в M_{n+1} не содержится число n, так как иначе в M_{n+1} нашлось бы два числа ($2n$ и n), из которых одно делится на другое.

Исключим из M_{n+1} числа $2n$—1 и $2n$. Совокупность оставшихся n—1 чисел обозначим M_{n-1}. Присоединим к M_{n-1} число n. Получим n чисел, каждое из которых не превосходит $2n$ — 2. Остается показать, что среди этих n чисел ни одно не делится на другое.

В M_{n+1} не было двух чисел, из которых одно делится на другое. Значит, таких чисел не было и в M_{n-1}. Остается только убедиться в том, что таких чисел не появилось и тогда, когда мы к M_{n-1} присоединили число n.

Для этого достаточно убедиться в том, что: 1) ни одно число, входящее в M_{n-1}, не делится на n и 2) число n не делится ни на одно из чисел, входящих в M_{n-1}.

Первое вытекает из того, что все числа, входящие в M_{n-1}, не превосходят $2n$ — 2.

Второе вытекает из того, что число $2n$ не делится ни на одно из чисел, входящих в M_{n-1}.

Итак, если допустить, что утверждение неверно для $2n$ чисел 1, 2.....$2n$, то оно неверно и для $2(n—1)$ чисел1, 2, ... $2n — 2$. Значит, если утверждение верно для $2(n— 1)$ чисел 1, 2.....$2n — 2$, то оно верно и для $2n$ чисел 1, 2,..., $2n$.

Отсюда и из пункта 1° следует, что наше утверждение справедливо для $2n$ чисел 1, 2.....$2n$, где n — любое натуральное число.

Заметим, что эта задача имеет следующее простое решение. Выберем из $2n$ чисел 1, 2, ... $2n$ произвольное $n + 1$ число. Совокупность этих чисел обозначим M_{n+1}.

Каждое четное число, входящее в M_{n+1}, разделим на такую степень двойки, чтобы частное было нечетным. Совокупность этих частных и всех нечетных чисел, входящих в M_{n+1}, обозначим через M'_{n+1}. В M'_{n+1} содержится $n+1$ нечетное число, каждое из которых меньше $2n$.

Так как всех положительных нечетных чисел, меньших $2n$, имеется всего n, то в M'_{n+1} имеются хотя бы два равных числа. Каждое из этих чисел пусть равно k.

Полученный результат означает, что в M_{n+1} было два числа $2^s k$ и $2^t k$ (где одно из чисел s и t может равняться нулю). Но одно из чисел $2^s k$ и $2^t k$ делится на второе.

Пример 13. Доказать, что n плоскостей, проходящих через одну точку так, что никакие три из них не проходят через одну прямую, делят пространство на $A_n = n(n—1)+2$ частей.

Решение. 1°. Одна плоскость делит пространство на две части, и $A_1 = 2$. Для $n = 1$ утверждение справедливо.

2°. Предположим, чго утверждение справедливо для $n = k$, т. е. k плоскостей делят пространство на $k(k — 1)+2$ частей. Докажем, что тогда $k+1$ плоскостей делят пространство на $k(k+1) + 2$ частей.

Действительно, пусть P есть $(k+ 1)$-я плоскость. С каждой из первых k плоскостей плоскость P пересекается по некоторой прямой и, таким образом, плоскость P разбита на части посредством k различных прямых, проходящих через одну точку. На основании задачи 28 утверждаем, что плоскость P разбита на $2k$ частей, каждая из которых представляет собой плоский угол с вершиной в данной точке.

Первые k плоскостей делят пространство на некоторые многогранные углы. Некоторые из этих многогранных углов делятся посредством плоскости P на две части.

Общей гранью двух таких частей служит часть плоскости, ограниченная двумя лучами, по которым P пересекается с гранями

данного многогранного угла, т. е. один из $2k$ плоских углов, на которые плоскость P разбита.

Это означает, что число многогранных углов, разбиваемых на две части плоскостью P, не может быть больше, чем $2k$.

С другой стороны, каждая из $2k$ частей, на которые разбивается плоскость P, в результате пересечения ее с первыми k плоскостями, является общей гранью двух многогранных углов и таким образом делит многогранный угол, образованный первыми k плоскостями, на две части.

Это означает, что число многогранных углов, которые разбиваются на две части плоскостью P, не может быть меньше, чем $2k$.

Итак, плоскость P разбивает на две части точно $2k$ частей пространства, образованных первыми k плоскостями. Поэтому если k плоскостей разбивают пространство на $k(k - 1)+2$ частей, $k+1$ плоскость разбивает пространство на

$$[k(k-1)+2]+2k = k(k+1)+2$$

частей. Утверждение доказано.

4.3. Тригонометрические и алгебраические задачи

Пример 14. Доказать тождество

$$\cos\alpha\cos 2\alpha\cos 4\alpha\ldots\cos 2^n\alpha = \frac{\sin 2^{n+1}\alpha}{2^{n+1}\sin\alpha}.$$

Решение. 1°. При $n=0$ тождество справедливо, так как

$$\cos\alpha = \frac{\sin 2\alpha}{2\sin\alpha}.$$

2°. Пусть тождество справедливо при $n = k$, т. е.

$$\cos\alpha\cos 2\alpha\ldots\cos 2^k\alpha = \frac{\sin 2^{k+1}\alpha}{2^{k+1}\sin\alpha}.$$

Тогда оно справедливо и при $n = k+1$. Действительно,

$$\cos\alpha\cos 2\alpha\ldots\cos 2^k\alpha\cos 2^{k+1}\alpha = \frac{\sin 2^{k+1}\alpha\cos 2^{k+1}\alpha}{2^{k+1}\sin\alpha} = \frac{\sin 2^{k+2}\alpha}{2^{k+2}\sin\alpha}.$$

Пример 15. Доказать, что $A_n = \cos n\theta$, если известно, что $A_1 = \cos\theta$, $A_2 = \cos 2\theta$ и для всякого натурального $k > 2$ имеет место соотношение

$$A_k = 2\cos\theta A_{k-1} - A_{k-2}.$$

Решение. 1°. Утверждение справедливо при $n = 1$ и $n = 2$.

2°. Пусть

$$A_{k-2} = \cos(k-2)\theta, \quad A_{k-1} = \cos(k-1)\theta.$$

Тогда

$$A_k = 2\cos\theta\cos(k-1)\theta - \cos(k-2)\theta = \cos k\theta.$$

Пример 16. Доказать, что

$$\sin x + \sin 2x + \ldots + \sin nx = \frac{\sin\frac{n+1}{2}x}{\sin\frac{x}{2}}\sin\frac{nx}{2}.$$

Решение. 1°. При $n = 1$ утверждение справедливо.
2°. Пусть

$$\sin x + \sin 2x + \ldots + \sin kx = \frac{\sin\frac{k+1}{2}x}{\sin\frac{x}{2}}\sin\frac{kx}{2}.$$

Тогда

$$\sin x + \sin 2x + \ldots + \sin kx + \sin(k+1)x =$$

$$= \frac{\sin\frac{k+1}{2}x}{\sin\frac{x}{2}}\sin\frac{kx}{2} + \sin(k+1)x =$$

$$= \frac{\sin\frac{k+1}{2}x}{\sin\frac{x}{2}}\sin\frac{kx}{2} + 2\sin\frac{k+1}{2}x\cos\frac{k+1}{2}x =$$

$$= \frac{\sin\frac{k+2}{2}x}{\sin\frac{x}{2}}\sin\frac{k+1}{2}x,$$

ибо

$$2\cos\frac{k+1}{2}x\sin\frac{x}{2} = \sin\frac{k+2}{2}x - \sin\frac{kx}{2}.$$

Пример 17. Доказать, что

$$(1+i)^n = 2^{\frac{n}{2}}\left(\cos\frac{n\pi}{4} + i\sin\frac{n\pi}{4}\right).$$

Решение. 1°. При n=1 утверждение справедливо, так как

$$1 + i = 2^{\frac{1}{2}}\left(\cos\frac{\pi}{4} + i\sin\frac{\pi}{4}\right).$$

2°. Пусть

$$(1+i)^k = 2^{\frac{k}{2}}\left(\cos\frac{k\pi}{4} + i\sin\frac{k\pi}{4}\right).$$

Тогда

$$(1+i)^{k+1} = 2^{\frac{k}{2}}\left(\cos\frac{k\pi}{4} + i\sin\frac{k\pi}{4}\right) 2^{\frac{1}{2}}\left(\cos\frac{\pi}{4} + i\sin\frac{\pi}{4}\right) =$$
$$= 2^{\frac{k+1}{2}}\left(\cos\frac{(k+1)\pi}{4} + i\sin\frac{(k+1)\pi}{4}\right).$$

Пример 18. Доказать теорему.

Если в результате конечного числа рациональных действий (т. е. сложения, вычитания, умножения и деления) над комплексными числами $x_1, x_2, \ldots, x_n$ получается число $\overline{\overline{u}}$, то в результате тех же действий над сопряженными комплексными числами $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n$ получится число $\bar{u}$, сопряженное с $\overline{\overline{u}}$.

Решение. 1°. Прежде всего покажем, что утверждение верно для каждого из четырех действий над двумя комплексными числами. Пусть

$$x_1 = a + bi, \quad x_2 = c + di.$$

Тогда

$$x_1 + x_2 = (a+c) + (b+d)i = u,$$
$$\bar{x}_1 + \bar{x}_2 = (a - bi) + (c - di) = (a+c) - (b+d)i = \bar{u}.$$

Точно так же утверждение проверяется для вычитания, умножения и деления.

2°. Пусть теперь дано некоторое рациональное выражение от комплексных чисел $x_1, x_2, \ldots, x_n$. Вычисление такого выражения сводится, как известно, к последовательному выполнению одного из четырех действий над двумя комплексными числами, причем действия эти могут быть занумерованы.

Например, пусть

$$u = \frac{x_1x_2 + x_3x_4}{x_1 + x_2 - x_3}.$$

Для вычисления u достаточно произвести действия:

1) $x_1x_2 = u_1$, 4) $u_3 - x_3 = u_4$,
2) $x_3x_4 = u_2$, 5) $u_1 + u_2 = u_5$,
3) $x_1 + x_2 = u_3$, 6) $u_5 : u_4 = u$.

Предположим, что утверждение верно для всех выражений, которые для вычисления их требуют не более k «действий». Термин «действие» здесь означает сложение либо вычитание, либо умножение, либо деление двух комплексных чисел. Покажем, что тогда утверждение должно быть верно и для выражений, требующих k +1 «действий».

Действительно, последнее (k+1)-е «действие» мы выполняем над числами u_i и u_j, которые сами вычислялись посредством не более чем k «действий».

В результате замены чисел $x_1, x_2 ..... x_n$ сопряженными, числа u_i и u_j заменяются сопряженными $\overline{u}_i$ и $\overline{u}_j$, а тогда и результат (k+1)-го «действия» над ними, т. е. число u, также заменится сопряженным числом $\overline{u}$.

4.4. Задачи на доказательство неравенств

19. Доказать, что при любом натуральном $n > 1$

$$\frac{1}{n+1}+\frac{1}{n+2}+\ \dots\ +\frac{1}{2n}>\frac{13}{24}.$$

Решение. Обозначим левую часть неравенств через S_n.

1°. $S_2=\frac{7}{12}=\frac{14}{24}$, следовательно, при n=2 неравенство справедливо.

2°. Пусть $S_k>\frac{13}{24}$ при некотором k. Докажем, что тогда и $S_{k+1}>\frac{13}{24}$. Имеем

$$S_k=\frac{1}{k+1}+\frac{1}{k+2}+\ \dots\ +\frac{1}{2k},$$

$$S_{k+1}=\frac{1}{k+2}+\frac{1}{k+3}+\ \dots\ +\frac{1}{2k}+\frac{1}{2k+1}+\frac{1}{2k+2}.$$

Сравнивая S_k и S_{k+1} имеем

$$S_{k+1}-S_k=\frac{1}{2k+1}+\frac{1}{2k+2}-\frac{1}{k+1},$$

т. е.

$$S_{k+1}-S_k=\frac{1}{2(k+1)(2k+1)}.$$

При любом натуральном k правая часть последнего равенства положительна. Поэтому $S_{k+1} > S_k$. Но $S_k > \frac{13}{24}$, значит, и $S_{k+1} > \frac{13}{24}$.

Задача 24. Найти ошибку.

Утверждение. При любом натуральном n справедливо неравенство

$$2^n > 2n + 1.$$

Доказательство. Пусть неравенство справедливо при $n = k$, где k — некоторое натуральное число, т. е.

$$2^k > 2k + 1. \qquad (1)$$

Докажем, что тогда неравенство справедливо и при $n = k+ 1$. т. е.

$$2^{k+1} > 2(k + 1) + 1. \qquad (2)$$

Действительно, 2^k не меньше 2 при любом натуральном k. Прибавим к левой части неравенства (1) 2^k, а к правой 2. Получим справедливое неравенство

$$2^{k+1} + 2^k > 2k + 1 + 2,$$

или

$$2^{k+1} > 2(k + 1)+1.$$

Утверждение доказано.

Пример 20. При каких натуральных n справедливо неравенство

$$2^n > n^2?$$

Решение.

При $n = 1$ неравенство справедливо, так как $2^1 > 1^2$.
При $n = 2$ неравенство несправедливо, так как $2^2 = 2^2$.
При $n = 3$ неравенство несправедливо, так как $2^3 < 3^2$.
При $n = 4$ неравенство несправедливо, так как $2^1 = 4^2$.
При $n = 5$ неравенство справедливо, так как $2^5 > 5^2$.
При $n = 6$ неравенство справедливо, так как $2^8 > 6^2$.

По-видимому, неравенство справедливо при $n = 1$ и при любом $n > 4$. Докажем это.

1°. При $n = 5$ неравенство справедливо.

2°. Пусть

$$2^k > k^2. \qquad (3)$$

где k — некоторое натуральное число, большее 4.

Докажем, что

$$2^{k+1} > (k+1)^2. \qquad (4)$$

Мы знаем, что $2^k > 2k+1$ при $k > 4$ (задача 24). Поэтому если мы к левой части неравенства (3) прибавим 2^k, а к правой $2k+1$, получим справедливое неравенство (4).

Ответ. $2^n > n^2$, когда $n = 1$ и когда $n > 4$.

Пример 21. Доказать, что

$$(1 + \alpha)^n > 1+n\alpha.$$

где $\alpha > -1$, $\alpha \neq 0$, n - натуральное число, большее 1.

Решение. 1°. При $n = 2$ неравенство справедливо, так как $\alpha^2 > 0$.

2°. Пусть неравенство справедливо при $n=k$, где k — некоторое натуральное число, т. е

$$(1 + \alpha)^k > 1 + k\alpha \tag{5}$$

Покажем, что тогда неравенство и при $n= k +1$, т. е

$$(1 + \alpha)^{k+1} > 1 + (k + 1)\alpha \tag{6}$$

Действительно, по условию, $1+\alpha>0$, поэтому справедливо неравенство

$$(1 + \alpha)^{k+1} > (1 + k\alpha)(1 + \alpha), \tag{7}$$

полученное из неравенства (5) умножением каждой части его на $1+\alpha$.

Перепишем неравенство (7) так:

$$(1+\alpha)^{k+1} > 1+(k+1)\alpha+k\alpha^2$$

Отбросив в правой части последнего неравенства положительное слагаемое $k\alpha^2$, получим справедливое неравенство (6).

Пример 22. Доказать, что

$$2^{n-1}(a^n + b^n) > (a + b)^n. \tag{8}$$

где $a + b > 0$, $a \neq b$, n — натура чьиое число, большее 1.

Решение 1°. При $n = 2$ неравенство (8) принимает вид

$$2(a^2 + b^2) > (a + b)^2 \tag{9}$$

Так как $a \neq b$, то справедливо неравенство

$$(a - b)^2 > 0. \tag{10}$$

Прибавив к каждой части неравенства (10) по $(a+b)^2$. получим неравенство (9).

Этим доказано, что при $n = 2$ неравенство (8) справедливо.

2° Пусть неравенство (8) справедливо при $n = k$, где k — некоторое натуральное число, т. е.

$$2^{k-1}(a^k + b^k) > (a + b)^k. \tag{11}$$

Докажем, что тогда неравенство (8) должно быть справедливо и при $n = k + 1$, т. е.

$$2^k(a^{k+1} + b^{k+1}) > (a + b)^{k+1}. \tag{12}$$

Умножим обе части неравенства (11) на $a+b$. Так как, по условию, $a+b>0$, то получаем следующее справедливое неравенство:

$$2^{k-1}(a^k+b^k)(a+b)>(a+b)^{k+1}. \tag{13}$$

Для того чтобы доказать справедливость неравенства (12), достаточно показать, что

$$2^k(a^{k+1}+b^{k+1})\geq 2^{k-1}(a^k+b^k)(a+b), \tag{14}$$

или, что все равно,

$$a^{k+1}+b^{k+1}\geq a^kb+ab^k. \tag{15}$$

Неравенство (15) равносильно неравенству

$$(a^k-b^k)(a-b)\geq 0. \tag{16}$$

Если $a > b$, то $a^k > b^k$, и в левой части неравенства (16) имеем произведение двух положительных чисел. Если $a < b$, то $a^k< b^k$ и в левой части неравенства (16) имеем произведение двух отрицательных чисел. В обоих случаях неравенство (16) справедливо.

Этим доказано, что из справедливости неравенства (8) при $n=k$ следует его справедливость при $n = k+1$.

Пример 23. Доказать, что при любом $x > 0$ и при любом натуральном n справедливо неравенство

$$x^n+x^{n-2}+x^{n-4}+\ldots+\frac{1}{x^{n-4}}+\frac{1}{x^{n-2}}+\frac{1}{x^n}\geq n+1. \tag{17}$$

Решение. 1°. а) При $n=1$ неравенство (17) принимает вид

$$x+\frac{1}{x}\geq 2. \tag{18}$$

Неравенство (18) вытекает из очевидного неравенства

$$(x-1)^2\geq 0.$$

б) При $n = 2$ неравенство (17) принимает вид

$$x^2+1+\frac{1}{x^2}\geq 3. \tag{19}$$

Неравенство (18) справедливо при любом $x>0$; значит, оно справедливо и при замене x на x^2, т. е.

$$x^2+\frac{1}{x^2}\geq 2.$$

Прибавив к каждой части последнего неравенства по 1, получим неравенство (19).

2°. Предположим, что неравенство (17) справедливо при $n = k$, где k — некоторое натуральное число, т. е.

$$x^k + x^{k-2} + \ldots + \frac{1}{x^{k-2}} + \frac{1}{x^k} \geqslant k + 1. \qquad (20)$$

Докажем, что тогда неравенство (17) справедливо и при $n = k + 2$, т. е.

$$x^{k+2} + x^k + x^{k-2} + \ldots + \frac{1}{x^{k-2}} + \frac{1}{x^k} + \frac{1}{x^{k+2}} \geqslant k + 3. \qquad (21)$$

Заменив в неравенстве (18) x на x^{k+2}, получаем

$$x^{k+2} + \frac{1}{x^{k+2}} \geqslant 2. \qquad (22)$$

Сложив почленно неравенства (20) и (22), получим неравенство (21).

Подведем теперь итог.

В пп. 1° а) и б) мы доказали, что неравенство (17) справедливо при $n = 1$ и при $n = 2$.

В п. 2° мы доказали, что из справедливости неравенства (17) при $n = k$ вытекает его справедливость и при $n = k+1$. Иными словами, п. 2° дает нам право перехода от $n = k$ к $n = k + 2$.

Результаты пп. 1° а) и 2° дают нам право утверждать, что неравенство (17) справедливо при любом нечетном n. Точно так же резулыаты пп. 1° б) и 2° дают нам право утверждать, что неравенство (17) справедливо при любом четном n. В целом мы имеем право утверждать, что неравенство (17) справедливо при любом натуральном n.

Пример 24. Доказать теорему:

Среднее геометрическое нескольких положительных чисел не больше их среднего арифметического, т. е. если $a_1, a_2, \ldots, a_n$ положительны, то

$$\sqrt[n]{a_1 a_2 \ldots a_n} \leqslant \frac{a_1 + a_2 + \ldots + a_n}{n}. \qquad (23)$$

Решение. 1°. При $n = 2$ неравенство (23) принимает вид

$$\sqrt{a_1 a_2} \leqslant \frac{a_1 + a_2}{2}. \qquad (24)$$

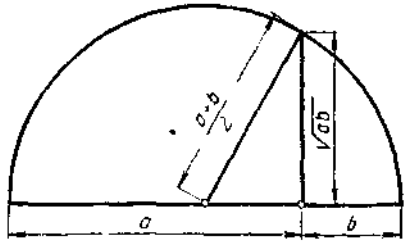
Это неравенство легко получить из справедливого при любых положительных a_1 и a_2 неравенства

$$(\sqrt{a_1} - \sqrt{a_2})^2 \geqslant 0.$$

Неравенство (24) имеет простой геометрический смысл. На прямой AB отложим последовательно отрезки a_1 и a_2. На сумме их как на диаметре опишем окружность. Тогда

$$\frac{a_1 + a_2}{2}$$

— радиус этой окружности, а $\sqrt{a_1 a_2}$ — половина хорды, перпендикулярной к диаметру в обшей точке a_1 и a_2 (см. рисунок), из чего и усматривается справедливость неравенства (24).



2°. Прелпоюжим, что неравенство (23) справедчиво при $n = k$.

Докажем, что тогда оно справедливо и при $n = 2k$. Действительно,

$$\sqrt{a_1 a_2 \dots a_{2k}} = \sqrt{\sqrt[k]{a_1 a_2 \dots a_k}\,\sqrt[k]{a_{k+1} \dots a_{2k}}} \leqslant$$

$$\leqslant \frac{\sqrt[k]{a_1 a_2 \dots a_k} + \sqrt[k]{a_{k+1} \dots a_{2k}}}{2} \leqslant$$

$$\leqslant \frac{\frac{a_1 + a_2 + \dots + a_k}{k} + \frac{a_{k+1} + \dots + a_{2k}}{k}}{2} =$$

$$= \frac{a_1 + a_2 + \dots + a_k + \dots + a_{2k}}{2k}.$$

Неравенство (23) проверено при n=2 и, таким образом, можем утверждать, что оно справедливо при n = 4, 8, 16 и т. д , т. е. вообще при $n = 2^s$, где s — натуральное число.

3°. Для того чтобы доказать справедливость неравенства (23) при всяком натуральном n, покажем, что из справедливости неравенства при $n = k$ следует его справедливость при $n = k$—1.

Итак, пусть $a_1, a_2, \dots, a_{k-1}$ — некоторые положительные числа. Пусть λ — некоторое, пока не определенное, положительное число Тогда

$$\sqrt[k]{a_1 a_2 \dots a_{k-1}\lambda} \leqslant \frac{a_1 + a_2 + \dots + a_{k-1} + \lambda}{k}.$$

Выберем λ так, чтобы

$$\frac{a_1 + a_2 + \dots + a_{k-1} + \lambda}{k} = \frac{a_1 + a_2 + \dots + a_{k-1}}{k-1},$$

т. е. положим

$$\lambda = \frac{a_1 + a_2 + \ldots + a_{k-1}}{k-1}.$$

Имеем

$$\sqrt[k]{\frac{a_1 a_2 \ldots a_{k-1}(a_1 + a_2 + \ldots + a_{k-1})}{k-1}} \leqslant \frac{a_1 + a_2 + \ldots + a_{k-1}}{k-1},$$

или

$$\sqrt[k-1]{a_1 a_2 \ldots a_{k-1}} \leqslant \frac{a_1 + a_2 + \ldots + a_{k-1}}{k-1}.$$

Пусть теперь m — произвольное натуральное число. Если $m = 2^s$, то согласно 2° для него неравенство справедливо. Если же $m \neq 2^s$, то найдем такое s, чтобы m. было меньше 2^s, и тогда на основании 2° и 3° утверждаем, что неравенство верно для $n = m$.

4.5. Доказательство некоторых теорем алгебры методом математической индукции

Теорема 1. *Квадрат многочлена равен сумме квад ратов всех его членов, сложенной со всевозможными их удвоенными попарными произведениями, т. е.*

$$(a_1 + a_2 + \ldots + a_n)^2 = a_1^2 + a_2^2 + \ldots + a_n^2 + 2(a_1 a_2 + a_1 a_3 + \ldots + a_{n-1} a_n). \quad (1)$$

1°. Для $n = 1$ формула (1) может быть доказана непосредственным умножением.

2°. Допустим, что формула (1) верна для $n = k$— 1, т. е.

$$(a_1 + a_2 + \ldots + a_{k-1})^2 = a_1^2 + a_2^2 + \ldots + a_{k-1}^2 + 2S.$$

где S — сумма всевозможных попарных произведений, составленных из a_1, a_2, ... a_{k-1}. Докажем, что

$$(a_1 + a_2 + \ldots + a_{k-1} + a_k)^2 = a_1^2 + a_2^2 + \ldots + a_{k-1}^2 + a_k^2 + 2S_1,$$

где S_1 — сумма всевозможных попарных произведений, составленных из a_1, a_2, ... a_{k-1}, a_k, т.е.

$$S_1 = S + (a_1 + a_2 + \ldots + a_{k-1}) a_k.$$

Действительно,

$$(a_1 + \ldots + a_{k-1} + a_k)^2 = [(a_1 + \ldots + a_{k-1}) + a_k]^2 =$$
$$= (a_1 + \ldots + a_{k-1})^2 + 2(a_1 + \ldots + a_{k-1})a_k + a_k^2 =$$
$$= a_1^2 + \ldots + a_{k-1}^2 + 2S + 2(a_1 + \ldots + a_{k-1})a_k + a_k^2 =$$
$$= a_1^2 + a_2^2 + \ldots + a_k^2 + 2S_1.$$

Теорема 2. *n*-й *член арифметической прогрессии может быть вычислен по формуле*

$$a_n = a_1 + d(n-1), \tag{2}$$

где a_1 — первый член, d — разность прогрессии.

1°. Для $n = 1$ формула (2) верна.

2°. Предположим, что формула (2) верна для $n = k$, т. е.

$$a_k = a_1 + d(k-1).$$

Тогда

$$a_{k+1} = a_k + d = a_1 + d(k-1) + d = a_1 + dk,$$

т. е. формула (2) оказывается справедливой и для $n = k+1$.

Теорема 3. n-й *член геометрической прогрессии может быть вычислен по формуле*

$$a_n = a_1 q^{n-1}, \tag{3}$$

где a_1 — первый член, q — знаменатель прогрессии.

1°. Для $n=1$ формула (3) верна.

2°. Пусть

$$a_k = a_1 q^{k-1}.$$

Тогда

$$a_{k+1} = a_k q = a_1 q^k.$$

Теорема 4. *Число перестановок из m элементов может быть вычислено по формуле*

$$P_m = m!. \tag{4}$$

1°. Прежде всего заметим, что $P_1 = 1$ и, таким образом, при $m=1$ формула (4) верна.

2°. Пусть $P_k = k!$. Докажем, что

$$P_{k+1} = (k+1)!.$$

Из данных $k+1$ элементов $a_1, a_2 \ldots .a_k, a_{k+1}$ возьмем только первые k и составим из них всевозможные перестановки. По условию, таких перестановок буает $k!$.

В каждой из этих перестановок поставим элемент a_{k+1} последовательно перед 1-м элементом, перед 2-м, ..., перед k -м, после k –го. Этим путем мы из одной перестановки из k элементов получим $k + 1$ перестановок из $k+1$ элементов. Всего имеем

$$k!\,(k+1) = (k+1)!$$

перестановок из $k+1$ элементов.

Необходимо выяснить:

1) нет ли среди этих $(k+1)!$ перестановок двух одинаковых,

2) все ли перестановки из $k+1$ элементов нами получены.

1) Допустим, что среди $(k+1)!$ перестановок имеются две одинаковые. Назовем их p_1 и p_2. Пусть в перестановке p_1 элемент a_{k+1} занимает s-е место, считая слева. Тогда и в p_2 элемент a_{k+1} занимает s-е место, считая слева.

Удалим из p_1 и p_2 элемент a_{k+1}. Получим две одинаковые перестановки из k элементов: $\overline{p}_1$ и $\overline{p}_2$.

Выходит, что для получения p_1 и p_2 в одну и ту же перестановку из элементов $a_1, a_2, \ldots, a_k$ два раза на одно и то же место был поставлен элемент a_{k+1}. Это противоречит правилу, по которому построены перестановки.

2) Допустим, что некоторая перестановка p из $k+1$ элементов нами не получена. Пусть в p элемент a_{k+1} занимает s-е место слева. Удалим из p элемент a_{k+1}. Получим перестановку p из первых k элементов. Значит, для получения p достаточно было взять перестановку $\overline{p}$ и поставить в нее элемент a_{k+1} так, чтобы он занял s-е место слева.

Мы не могли не взять перестановку $\overline{p}$, так как брали всевозможные перестановки из первых k элементов.

Мы не могли не поставить элемент a_{k+1} на указанное место, так как ставили его и первым, и вторым ..... и $(k+1)$-м слева

Итак, *составленные нами перестановки все различны и всякая перестановка из* $k+1$ *элементов нами получена.*

Из сказанного вытекает, что

$$P_{k+1} = (k+1)!.$$

Теорема 5. *Число размещений из m элементов по n может быть вычислено по формуле*

$$A_m^n = m(m-1)\ldots(m-n+1). \qquad (5)$$

1°. Прежде всего наметим, что $A_m = m$ и, таким образом, формула (5) верна при $n=1$.

2°. Предположим, что

$$A_m^k = m(m-1)\ldots(m-k+1),$$

где $k < m$. Докажем, чю

$$A_m^{k+1} = m(m-1)\ldots(m-k).$$

Для получения всех размещений из m элементов по k+1 элементу достаточно взять все размещения из m элементов по k и к каждому из них приписать в конце каждый из оставшихся m— k элементов. Нетрудно убедиться, что составленные таким образом размещения из m элементов по k +1 *все различны* и, кроме того, *всякое размещение из m элементов по k + 1 содержится среди полученных.*

Выходит, что

$$A_m^{k+1} = A_m^k (m - k) = m(m-1) \dots (m-k).$$

Теорема 6. *Число сочетаний из m элементов по n может быть вычислено по формуле*

$$C_m^n = \frac{m(m-1)\dots(m-n+1)}{1 \cdot 2 \dots n}. \qquad (6)$$

1°. Прежде всего заметим, что $C^1_m = m$ и, таким образом, при n=1 формула (6) верна.

2°. Допустим, что

$$C_m^k = \frac{m(m-1)\dots(m-k+1)}{1 \cdot 2 \dots k}.$$

Докажем, что

$$C_m^{k+1} = \frac{m(m-1)\dots(m-k+1)(m-k)}{1 \cdot 2 \dots k(k+1)}.$$

Для получения всех сочетаний из m элементов по k+1 выпишем все сочетания из m элементов по k и к каждому из них в качестве (k+1)-го элемента присоединим каждый из m — k оставшихся элементов.

Ясно, что таким путем будут получены все сочетания из m элементов по k + 1, но каждое из них получится k + 1 раз.

Действительно, сочетание $a_1, a_2, \dots, a_k, a_{k+1}$ получится, когда к сочетанию $a_2, a_3, \dots, a_k, a_{k+1}$ присоединится элемент a_1, когда к сочетанию $a_1, a_3, \dots, a_k, a_{k+1}$ присоединится элемент a_2 и т. д., когда, наконец, к сочетанию $a_1, a_2, \dots, a_k$, присоединится элемент a_{k+1}. Таким образом,

$$C_m^{k+1} = C_m^k \frac{m-k}{k+1} = \frac{m(m-1)\dots(m-k)}{1 \cdot 2 \dots k(k+1)}.$$

Теорема 7. *Каково бы ни была числа a и b и каково бы ни было натуральное число n. имеет место фо рмула*

$$(a+b)^n = a^n + C_n^1 a^{n-1} b + \dots + C_n^s a^{n-s} b^s + \dots + \dots + C_n^{n-1} a b^{n-1} + b^n \qquad (7)$$

(формула бинома Ньютона).

1°. При n = 1 имеем $a + b = b + a$ и, таким образом, для этого случая формула (7) верна.

2°. Пусть

$$(a+b)^k = a^k + C_k^1 a^{k-1}b + C_k^2 a^{k-2}b^2 + \ldots + b^k.$$

Тогда

$$\begin{aligned}(a+b)^{k+1} &= (a+b)^k (a+b) = \\ &= (a^k + C_k^1 a^{k-1}b + \ldots + b^k)(a+b) = \\ &= a^{k+1} + (1 + C_k^1) a^k b + (C_k^1 + C_k^2) a^{k-1}b^2 + \ldots \\ &\quad \ldots + (C_k^s + C_k^{s+1}) a^{k-s}b^{s+1} + \ldots + b^{k+1}.\end{aligned}$$

Имея в виду, что $C_k^s + C_k^{s+1} = C_{k+1}^{s+1}$, получаем

$$\begin{aligned}(a+b)^{k+1} &= a^{k+1} + C_{k+1}^1 a^k b + C_{k+1}^2 a^{k-1}b^2 + \ldots \\ &\quad \ldots + C_{k+1}^{s+1} a^{k-s}b^{s+1} + \ldots + b^{k+1}.\end{aligned}$$

Заключение

Напомним, что индукция (лат. inductio — наведение) — переход от частного к общему; дедукция (лат. deductio — вывод) — переход от общего к частному. Всем известна роль процессов обобщения результатов отдельных наблюдении и опыюв (т. е. индукции) для эмпирических, экспериментальных наук. Математика же издавна считалась классическим образцом осуществления чисто дедуктивных методов, поскольку явно или неявно всегда подразумевалось, что все математические предложения (кроме принятых за исходные — аксиом) доказываются, а конкретные применения этих предложений выводятся из доказательств, пригодных для общих случаев (дедукция).

Но вот мы отмечали: «Индукция широко применяется в математике, но применять ее надо умело»; «... как пользоваться в математике индукцией, чтобы получать только верные выводы?». Что же все это, собственно, значит? Не следует ли понимать дело так, что среди математических методов есть «достоверные», действующие, так сказать, безотказно (дедуктивные), и «не вполне надежные», дающие подчас, особенно в неумелых руках (как мы выражались, «при легкомысленном отношении»), осечку (индуктивные)? Если бы это было действительно так, то где же искать критерии надежности таких «индуктивных» методов? Как вернуть себе уверенность в непреложной обязательности математических выводов? Или это безнадежная затея, и достоверность математических заключений — той же природы, что и опытные обобщения экспериментальных наук, так что любой доказанный факт неплохо было бы еще «проверить»

(подобно тому как учащимся часто рекомендуется «проверять» правильность выполнения арифметических действий или решения уравнений по общей формуле)?

В действительности дело обстоит не так. Индукция т. е «наведение» (на мысль, на догадку, на гипотезу) играет в математике, безусловно, очень большую, но чисто эвристическую роль: она позволяет догадываться о том, каким, по всей видимости, должно быть решение. ***Устанавливаются же математические предложения только дедуктивно.*** Ни один математический результат не может претендовать на достоверность, истинность, коль скоро он не выведен из исходных посылок.

Ну, а как же «метод математической индукции»? Дело все в том, что «математическая индукция» есть дедуктивный метод. В самом деле, разберемся детальнее в структуре математических умозаключений, выглядящих как «переход от частного к общему». Легко убедиться, что так называемая математическая индукция на самом деле новее не есть индукция — это чисто дедуктивный метод рассуждения! Доказательство, проводимое этим методом, состоит из двух частей:

1) так называемый базис — доказательство (дедуктивное!) искомого предложения для одного (или нескольких) натурального числа (например, для 0 или 1; это то, что нами именуется «Теоремой 1»);

2) индукционный шаг («Теорема 2»), состоящей в доказательстве (опять-таки дедуктивном) общего утверждения, для всех n верно, что из того, что искомое утверждение справедливо для n, вытекает, что оно справедливо и для $n + 1$ «Принцип математической индукции» — точно формулируемое предложение (интуитивная убедительность которого признается многими математиками как неоспоримая, при аксиоматическом же построении арифметики он фигурирует в качестве аксиомы), позволяющее извлечь из базиса и индукционного шага чисто дедуктивное доказательство рассматриваемого предложения для всех натуральных чисел n. Таким образом, никаких «не учтенных в посылках» случаев, на которые затем («по индукции») надо было бы еще «распространять» заключение, не остается — теорема именно доказывается для всех натуральных чисел: из базиса, доказанного, скажем, для числа 0, мы получаем, по индукционному шагу, доказательство для числа 1, затем таким же образом для 2, затем для 3 ... — и так утверждение теоремы может быть обосновано для любого натурального числа.

Иначе говоря, название «математическая индукция» обусловлено тем, что этот метод просто ассоциируется в нашем сознании с традиционными «индуктивными» умозаключениями (ведь базис

действительно доказывается только для частного случая); индукционный шаг, в отличие от основанных на опыте критериев правдоподобности индуктивных умозаключений в естественных (и общественных) науках, есть общее утверждение, не нуждающееся ни в какой частной посылке и доказываемое по строгим канонам дедуктивных рассуждений. Потому-то и называют математическую «индукцию» «полной», или «совершенной», что она (в противоположность обычной, «несовершенной» индукции, не обеспечивающей нам полного знания) есть дедуктивный («стопроцентно надежный») метод доказательства.

Итак, в качестве метода доказательства индукция в математике не применяется, что, разумеется, никак не исключает широкого применения в ней дедуктивного метода «математической индукции».

Условившись отныне в таком понимании термина, мы можем, конечно, позволить себе теперь и вольные перефразировки вроде «индукции в геометрии» или «индукции в математике». Но при этом всегда надо помнить, что первое выражение, строго говоря, имеет совсем не тот смысл, что громоздкое (но точное!) выражение «употребление дедуктивного метода математической индукции для доказательства теорем геометрического содержания» (хотя, для облегчения речи, и употребляется как его синоним), а второе — отнюдь не то же самое (вопреки чисто грамматическим признакам), что «математическая индукция»; последний термин следует воспринимать целиком, а вовсе не в смысле «индукция в математике».

Метод математической индукции (в той форме, в какой он рассматривается нами) есть метод доказательства арифметических теорем, точнее, теорем, выражающих общие свойства натуральных чисел (0, 1, 2, ...; иногда, как в этой работе, натуральный ряд уславливаются начинать с единицы, что абсолютно не принципиально). И для арифметики натуральных чисел этот метод, в известном (достаточно разумном и сильном) смысле, является универсальным (а часто и единственным) орудием доказательств.

Чтобы это последнее утверждение не показалось читателю слишком сильным, он должен твердо уяснить себе, что при аксиоматическом (дедуктивном) построении арифметики все ее здание опирается на определения операций над натуральными числами по математической индукции (например, при определении сложения прежде всего определяется, — в качестве базиса индукции, — что значит прибавить единицу или нуль; затем — индукционный шаг определения— определение прибавления произвольною натурального числа сводится к определению прибавления предшествующего числа). И вполне

понятно потому, что «добираться» до общих свойств натуральных чисел, связанных, скажем, с операциями сложения или умножения, нам приходится (если уж мы хотим обосновывать их аксиоматически) по той же «лестнице» (на нижней «ступени» которой находится соответствующее свойство для наименьшего натурального числа), по которой мы «совершаем восхождение» к интересующему нас общему понятию; грубо говоря, иначе просто не видно, как за нужное нам доказательство «ухватиться»! И так обстоит дело с доказательством любого общего арифметического утверждения! И если это не видно мз начального курса арифметики и алгебры, то лишь потому, что он (совершенно резонно) опирается не столько на аксиоматический метод, сколько на опыт и интуицию. В тех же случаях, когда в начальном курсе доказываются какие-либо общие свойства натуральных чисел, то доказательство, если и не проводится по индукции, то лишь благодаря тому, что в качестве посылок (часто неявных) используются предложения, для строгого обоснования которых индукция все же необходима (подобно тому, как употребление постулата о параллельных в евклидовой геометрии можно «замаскировать», пользуясь вместо него каким-нибудь из его следствий).

В конце концов самый придирчивый и критически настроенный читатель часто довольствуется знанием того, что, скажем, дистрибутивность умножения относительно сложения м о ж н о доказать, и уже не требует самого доказательства. (Но такая, пусть вполне обоснованная, уверенность так же отличается от подлинного доказательства, как, скажем, газетная информация от подлинного знания очевидца, причем эта аналогия простирается весьма далеко.) Поэтому-то метод математической индукции и появляется в начальном курсе математики гораздо позже интуитивно прозрачных и легко постигаемых свойств арифметических действий, например, в связи с формулой бинома Ньютона, которая уже отнюдь не такова, чтобы справедливость ее «бросалась в глаза».

В той мере, в какой другие разделы математики опираются на арифметическую основу, они нуждаются в методе математической индукции. Потребность эта бывает двоякого рода. Прежде всего многие разделы математики просто строятся на базе арифметики натуральных чисел (скажем, теория рациональных чисел приводящая в свою очередь к теории действительных чисел), другие же могут быть интерпретированы в арифметических терминах (например, любой факт евклидовой геометрии можно выразить на «координатном языке» действительных чисел). В этих случаях утверждения,

предположим, геометрического содержания могу быть доказаны именно для такой арифметической интерпретации с помощью математической индукции. Можно сказать, что геометрическая или какая-либо иная «специфика» подобных предложений не более существенна для самого доказательства, чем, например, природа рассматриваемых объектов в задаче о сложении трех огурцов с пятью огурцами или трех пароходов с пятью пароходами.

Но бывает так, что базис индукции доказывается существенно неарифметическими методами. И в этм случае, однако, индукционный шаг (даже если он опирается на геометрические или какие-иибудь другие аксиомы) предсавляет собой некоторое общее утверждение о натуральных числах, поскольку в нем идет речь о выполнении некоторого свойства для любого натурального числа *n,* то есть сам переход «от *n* к *n*+1» доказывается для любого *n.*

Итак, математическая индукция по натуральным числам есть метод доказательства теорем, арифметических «по форме», по, быть может, геометрических или каких-нибудь других (скажем, механических) «по содержанию».

Отметим еще, что метод, оказавшийся столь плодотворным для проведения доказательств, следующих процессу построения натурального ряда 0, 1, 2..... может быть обобщен и на процессы совершенно другого вида Например, в исчислениях математической логики, оперирующих с формулами («высказываниями»), построенными из «элементарных формул» («элементарных высказываний») вида *A, B, C* ,.., с помощью, допустим, знаков *&* («и»), $\vee$ («или»), $\supset$ («если ..., то ...») и

метода математической индукции в математическом анализе, кстати, объясняется именно тем обстоятельством, что действительные числа, в отличие от натуральных, не являются продуктом такой развертывающейся четко очерченной конструкции, так что различного рода «индукции по действительным числам» далеко не обладают той универсальностью, как метод математической индукции в арифметике и его модификации в математической логике.)

Для разрешения тех вопросов общелогического и общематематическою характера, которые могли бы возникнуть теперь у читателя, отсылаем его к специальной литературе (См., например, Л. Г е н к и н, О математической индукции, *М.,* Физматгиз, 1962; И. В. Арнольд, Теоретическая арифметика, М., Учпедгиз, 1939, §13, 14, 17, 19; С. К. К л и н и, Введение в математику, М., ИЛ, 1957, § 7, 13, 21, 38 и др.; «Математическая индукция» — Философская энциклопедия, М,, 1964 (т, 3). Задачу же первоначального ознакомления с конкретными применениями метода математической индукции в элементарной математике может с успехом выполнить эта работа.

Микромодуль 14
Индивидуальные тестовые задания

Задача 1. Найти u_n, если известно, что $u_1 = 1$ и что при всяком натуральном $k > 1$

$$u_k = u_{k-1} + 3.$$

Указание, $u_1 = 3 \cdot 1 - 2$, $u_1 = 3 \cdot 2 - 2$.

Задача 2. Найти сумму

$$S_n = 1 + 2 + 2^2 + 2^3 + \ldots + 2^{n-1}.$$

Указание. $S_1 = 2 - 1$, $S_2 = 2^2 - 1$, $S_3 = 2^3 - 1$.

Задача 3. Доказать, что

$$1^2 + 3^2 + 5^2 + \ldots + (2n-1)^2 = \frac{n(2n-1)(2n+1)}{3}.$$

Задача 4. Доказать, что сумма кубов n первых чисел натурального ряда равна

$$\left[\frac{n(n+1)}{2}\right]^2.$$

Задача 5. Доказать, что

$$1 + x + x^2 + \ldots + x^n = \frac{x^{n+1} - 1}{x - 1} \quad (x \neq 1).$$

Задача 6. Доказать, что

$$1\cdot 2\cdot 3+2\cdot 3\cdot 4+3\cdot 4\cdot 5+\ \ldots\ +n(n+1)(n+2)=\\ =\frac{n(n+1)(n+2)(n+3)}{4}.$$

Задача 7. Доказать, что

$$\frac{1}{1\cdot 3}+\frac{1}{3\cdot 5}+\ \ldots\ +\frac{1}{(2n-1)(2n+1)}=\frac{n}{2n+1}.$$

Задача 8. Доказать, что

$$\frac{1^2}{1\cdot 3}+\frac{2^2}{3\cdot 5}+\ \ldots\ +\frac{n^2}{(2n-1)(2n+1)}=\frac{n(n+1)}{2(2n+1)}.$$

Задача 9. Доказать, чго

$$\frac{1}{1\cdot 4}+\frac{1}{4\cdot 7}+\frac{1}{7\cdot 10}+\ \ldots\ +\frac{1}{(3n-2)(3n+1)}=\frac{n}{3n+1}.$$

Задача 10. Доказать, что

$$\frac{1}{1\cdot 5}+\frac{1}{5\cdot 9}+\frac{1}{9\cdot 13}+\ \ldots\ +\frac{1}{(4n-3)(4n+1)}=\frac{1}{4n+1}.$$

Задача 11. Доказать, что

$$\frac{1}{a(a+1)}+\frac{1}{(a+1)(a+2)}+\ldots+\frac{1}{(a+n-1)(a+n)}=\frac{n}{a(a+n)}.$$

Задача 12. Доказать, что если

$$u_1=\frac{\alpha^2-\beta^2}{\alpha-\beta},\qquad u_2=\frac{\alpha^3-\beta^3}{\alpha-\beta}\quad(\alpha\neq\beta)$$

и для всякого натурального $k>2$ имеет место соотношение

$$u_k=(\alpha+\beta)u_{k-1}-\alpha\beta u_{k-2},$$

то

$$u_n=\frac{\alpha^{n+1}-\beta^{n+1}}{\alpha-\beta}.$$

Задача 13. Доказать тождество

$$\frac{1}{1+x}+\frac{2}{1+x^2}+\frac{4}{1+x^4}+\frac{8}{1+x^8}+\ \ldots\ +\frac{2^n}{1+x^{2^n}}=\\ =\frac{1}{x-1}+\frac{2^{n+1}}{1-x^{2^{n+1}}}.$$

Задача 14. Упростить многочлен

$$1-\frac{x}{1!}+\frac{x(x-1)}{2!}-\ \ldots\ +(-1)^n\frac{x(x-1)\ldots(x-n+1)}{n!}.$$

Ответ.

$$(-1)^n \frac{(x-1)(x-2)\ldots(x-n)}{n!}.$$

Задача 15. Доказать, что при целом $n \geq 0$

$$A_n = 11^{n+2} + 12^{2n+1}$$

делится на 133.

Задача 16. Доказать, что n различных прямых, проведенных на плоскости через одну точку, делят плоскость на $2n$ частей.

Задача 17. Доказать, что

$$\frac{1}{2} + \cos x + \cos 2x + \ldots + \cos nx = \frac{\sin \frac{2n+1}{2} x}{2 \sin \frac{x}{2}}.$$

Задача 18. Доказать, что

$$\sin x + 2 \sin 2x + 3 \sin 3x + \ldots + n \sin nx = \frac{(n+1) \sin nx - n \sin (n+1) x}{4 \sin^2 \frac{x}{2}}.$$

Задача 19. Доказать, что

$$\cos x + 2 \cos 2x + \ldots + n \cos nx = \frac{(n+1) \cos nx - n \cos (n+1) x - 1}{4 \sin^2 \frac{x}{2}}.$$

Задача 20. Доказать, что

$$\frac{1}{2} \operatorname{tg} \frac{x}{2} + \frac{1}{2^2} \operatorname{tg} \frac{x}{2^2} + \ldots + \frac{1}{2^n} \operatorname{tg} \frac{x}{2^n} = \frac{1}{2^n} \operatorname{ctg} \frac{x}{2^n} - \operatorname{ctg} x \quad (x \neq m\pi).$$

Задача 21. Доказать, что

$$\operatorname{arc\,ctg} 3 + \operatorname{arc\,ctg} 5 + \ldots + \operatorname{arc\,ctg} (2n+1) = \operatorname{arc\,tg} 2 + \operatorname{arc\,tg} \frac{3}{2} + \ldots + \operatorname{arc\,tg} \frac{n+1}{n} - n \operatorname{arc\,tg} 1.$$

Задача 22. Доказать, что

$$(\sqrt{3} - i)^n = 2^n \left(\cos \frac{n\pi}{6} - i \sin \frac{n\pi}{6} \right).$$

Задача 23. Доказать, что при любом натуральном n

$$(\cos x + i \sin x)^n = \cos nx + i \sin nx.$$

Задача 24. (содержание задачи см. в тексте п. 4.4)

Задача 25. При каких натуральных n справедливо неравенство

$$2^n > 2n+1 \ ?$$

Задача 26. Доказать, что при любом натуральном $n > 1$

$$\frac{1}{\sqrt{1}} + \frac{1}{\sqrt{2}} + \ \dots \ + \frac{1}{\sqrt{n}} > \sqrt{n}$$

Задача 27. Доказать, что при любом натуральном $n > 1$

$$\frac{4^n}{n+1} < \frac{(2n)!}{(n!)^2}.$$

Указания и решения приведенных выше задач

1. Гипотеза.

$$u_n = 3n - 2.$$

1°. Для $n = 1$ гипотеза верна.

2°. Пусть

$$u_k = 3k - 2.$$

Тогда

$$u_{k+1} = u_k + 3 = 3k - 2 + 3 = 3(k+1) - 2.$$

2. Гипотеза

$$S_n = 2^n - 1.$$

1°. Для $n = 1$ гипотеза верна.

2°. Пусть

$$S_k = 2^k - 1.$$

Тогда

$$S_{k+1} = S_k + 2^k = 2^{k+1} - 1.$$

[Можно также сразу образовать разность $2S_n = S_n$ и показать, что она равна 2^n— 1.]

3. 1°. При n= 1 утверждение справедливо.

2°. Пусть

$$1^2 + 3^2 + 5^2 + \ \dots \ + (2k-1)^2 = \frac{k(2k-1)(2k+1)}{3}.$$

Тогда

$$1^2 + 3^2 + \ \dots \ + (2k-1)^2 + (2k+1)^2 =$$

$$= \frac{k(2k-1)(2k+1)}{3} + (2k+1)^2 = \frac{(k+1)(2k+1)(2k+3)}{3}.$$

4. 1°. При n — 1 утверждение справедливо.

2°. Пусть

$$1^3 + 2^3 + \ \dots \ + k^3 = \left[\frac{k(k+1)}{2}\right]^2.$$

Тогда

$$1^3+2^3+\ldots+k^3+(k+1)^3=\frac{k^2(k+1)^2}{4}+(k+1)^3=\left[\frac{(k+1)(k+2)}{2}\right]^2.$$

5. 1°. При $n=1$ утверждение справедливо.
2°. Пусть

$$1+x+x^2+\ldots+x^k=\frac{x^{k+1}-1}{x-1}.$$

Тогда

$$1+x+x^2+\ldots+x^k+x^{k+1}=\frac{x^{k+1}-1}{x-1}+x^{k+1}=\frac{x^{k+2}-1}{x-1}.$$

6. 1°. При $n=1$ утверждение справедливо.
2°. Пусть

$$1\cdot2\cdot3+2\cdot3\cdot4+\ldots+k(k+1)(x+2)=\frac{k(k+1)(k+2)(k+3)}{4}.$$

Тогда

$$1\cdot2\cdot3+2\cdot3\cdot4+\ldots+k(k+1)(k+2)+(k+1)(k+2)(k+3)=$$
$$=\frac{k(k+1)(k+2)(k+3)}{4}+(k+1)(k+2)(k+3)=$$
$$=\frac{(k+1)(k+2)(k+3)(k+4)}{4}.$$

7. 1°. При $n=1$ утверждение справедливо
2°. Пусть

$$\frac{1}{1\cdot3}+\frac{1}{3\cdot5}+\ldots+\frac{1}{(2k-1)(2k+1)}=\frac{k}{2k+1}.$$

Тогда

$$\frac{1}{1\cdot3}+\frac{1}{3\cdot5}+\ldots+\frac{1}{(2k-1)(2k+1)}+\frac{1}{(2k+1)(2k+3)}=$$
$$=\frac{k}{2k+1}+\frac{1}{(2k+1)(2k+3)}=\frac{k+1}{2k+3}.$$

8. 1°. При $n=1$ утверждение справедливо.
2°. Пусть

$$\frac{1^2}{1\cdot3}+\frac{2^2}{3\cdot5}+\ldots+\frac{k^2}{(2k-1)(2k+1)}=\frac{k(k+1)}{2(2k+1)}.$$

Тогда

$$\frac{1^2}{1\cdot3}+\frac{2^2}{3\cdot5}+\ldots+\frac{k^2}{(2k-1)(2k+1)}+\frac{(k+1)^2}{(2k+1)(2k+3)}=$$
$$=\frac{k(k+1)}{2(2k+1)}+\frac{(k+1)^2}{(2k+1)(2k+3)}=(k+1)\,\frac{k(2k+3)+2(k+1)}{2(2k+1)(2k+3)}=$$
$$=\frac{(k+1)(2k^2+5k+2)}{2(2k+1)(2k+3)}=\frac{(k+1)(2k+1)(k+2)}{2(2k+1)(2k+3)}=\frac{(k+1)(k+2)}{2(2k+3)}.$$

9. 1°. При $n=1$ утверждение справедливо.

2°. Пусть

$$\frac{1}{1\cdot 4}+\frac{1}{4\cdot 7}+\ldots+\frac{1}{(3k-2)(3k+1)}=\frac{k}{3k+1}.$$

Тогда

$$\frac{1}{1\cdot 4}+\frac{1}{4\cdot 7}+\ldots+\frac{1}{(3k-2)(3k+1)}+\frac{1}{(3k+1)(3k+4)}=$$
$$=\frac{k}{3k+1}+\frac{1}{(3k+1)(3k+4)}=\frac{k+1}{3k+4}.$$

10. 1°. При $n=1$ утверждение справедливо.

2°. Пусть

$$\frac{1}{1\cdot 5}+\frac{1}{5\cdot 9}+\ldots+\frac{1}{(4k-3)(4k+1)}=\frac{k}{4k+1}.$$

Тогда

$$\frac{1}{1\cdot 5}+\frac{1}{5\cdot 9}+\ldots+\frac{1}{(4k-3)(4k+1)}+\frac{1}{(4k+1)(4k+5)}=$$
$$=\frac{k}{4k+1}+\frac{1}{(4k+1)(4k+5)}=\frac{k+1}{4k+5}.$$

11. 1°. При $n=1$ утверждение справедливо.

2°. Пусть

$$\frac{1}{a(a+1)}+\frac{1}{(a+1)(a+2)}+\ldots+\frac{1}{(a+k-1)(a+k)}=\frac{k}{a(a+k)}.$$

Тогда

$$\frac{1}{a(a+1)}+\frac{1}{(a+1)(a+2)}+\ldots$$
$$\ldots+\frac{1}{(a+k-1)(a+k)}+\frac{1}{(a+k)(a+k+1)}=$$
$$=\frac{k}{a(a+k)}+\frac{1}{(a+k)(a+k+1)}=\frac{k+1}{a(a+k+1)}.$$

12. 1°. При $n=1$ и $n=2$ утверждение справедливо.

2°. Пусть

$$u_{k-1}=\frac{\alpha^{k-1}-\beta^{k-1}}{\alpha-\beta},\quad u_k=\frac{\alpha^k-\beta^k}{\alpha-\beta}.$$

Тогда

$$u_k=(\alpha+\beta)\frac{\alpha^k-\beta^k}{\alpha-\beta}-\alpha\beta\frac{\alpha^{k-1}-\beta^{k-1}}{\alpha-\beta}=\frac{\alpha^{k+1}-\beta^{k+1}}{\alpha-\beta}.$$

13. 1°. При $n=0$ имеем

$$\frac{1}{1+x}=\frac{1}{x-1}+\frac{2}{1-x^2}.$$

Следовательно, утверждение справедливо.

2°. Пусть

$$\frac{1}{1+x}+\frac{2}{1+x^2}+\frac{4}{1+x^4}+\ldots+\frac{2^k}{1+x^{2^k}}=\frac{1}{x-1}+\frac{2^{k+1}}{1-x^{2^{k+1}}}.$$

Тогда

$$\frac{1}{1+x}+\frac{2}{1+x^2}+\frac{4}{1+x^4}+\ldots+\frac{2^k}{1+x^{2^k}}+\frac{2^{k+1}}{1+x^{2^{k+1}}}=$$
$$=\frac{1}{x-1}+\frac{2^{k+1}}{1-x^{2^{k+1}}}+\frac{2^{k+1}}{1+x^{2^{k+1}}}=\frac{1}{x-1}+\frac{2^{k+2}}{1-x^{2^{k+2}}}.$$

14. При $n = 1$ имеем

$$1-\frac{x}{1!}=-\frac{x-1}{1}.$$

При $n = 2$ имеем

$$1-\frac{x}{1!}+\frac{x(x-1)}{2!}=-\frac{x-1}{1}+\frac{x(x-1)}{2}=\frac{(x-1)(x-2)}{2!}.$$

При $n = 3$ имеем

$$1-\frac{x}{1!}+\frac{x(x-1)}{2!}-\frac{x(x-1)(x-2)}{3!}=$$
$$=\frac{(x-1)(x-2)}{2}-\frac{x(x-1)(x-2)}{6}=-\frac{(x-1)(x-2)(x-3)}{3!}.$$

Это наводит на гипотезу

$$1-\frac{x}{1!}+\frac{x(x-1)}{2!}-\ldots+(-1)^n\frac{x(x-1)\ldots(x-n+1)}{n!}=$$
$$=(-1)^n\frac{(x-1)(x-2)\ldots(x-n)}{n!}.$$

1°. При $n = 1$ гипотеза верна.

2°. Пусть

$$1-\frac{x}{1!}+\frac{x(x-1)}{2!}-\ldots+(-1)^k\frac{x(x-1)\ldots(x-k+1)}{k!}=$$
$$=(-1)^k\frac{(x-1)(x-2)\ldots(x-k)}{k!}.$$

Тогда

$$1-\frac{x}{1!}+\frac{x(x-1)}{2!}-\ldots+(-1)^k\frac{(x-1)(x-2)\ldots(x-k+1)}{k!}+$$
$$+(-1)^{k+1}\frac{x(x-1)\ldots(x-k)}{(k+1)!}=(-1)^k\frac{(x-1)(x-2)\ldots(x-k)}{k!}+$$
$$+(-1)^{k+1}\frac{x(x-1)\ldots(x-k)}{(k+1)!}=$$
$$=(-1)^{k+1}\frac{(x-1)(x-2)\ldots(x-k)}{k!}\left[\frac{x}{k+1}-1\right]=$$
$$=(-1)^{k+1}\frac{(x-1)(x-2)\ldots(x-k)(x-k-1)}{(k+1)!}.$$

15. 1°. При n =0 утверждение справедливо.
2°. Предположим, что утверждение справедливо при $n = k$, т. е. что

$$A_k = 11^{k+2} + 12^{2k+1}$$

делится на 133. Тогда

$$\begin{aligned} A_{k+1} &= 11^{k+3} + 12^{2(k+1)+1} = 11^{k+3} + 12^{2k+3} = \\ &= 11 \cdot 11^{k+2} + 144 \cdot 12^{2k+1} = \\ &= 11 \cdot 11^{k+2} + 133 \cdot 12^{2k+1} + 11 \cdot 12^{2k+1} = \\ &= 11(11^{k+2} + 12^{2k+1}) + 133 \cdot 12^{2k+1} = 11A_k + 133 \cdot 12^{2k+1}. \end{aligned}$$

Мы представили A_{k+1} в виде суммы двух слагаемых, каждое из которых делится на 133. Значит, A_{k+1} делится на 133.
16. 1°. При u = 1 утверждение задачи, очевидно, справедливо.
2°. Предположив, что при $n = k$ утверждение справедливо, т. е. k прямых делят плоскость на $2k$ углов, (k+1)-я прямая рассекает на части сразу два вертикальных угла, т. е. увеличивает число частей, на которые делится плоскость, на два. Поэтому (k+ 1)-я прямая делит плоскость на 2+ 2 =2(k+ 1) частей.
17. 1°. При n= 1 утверждение справедливо, так как

$$\frac{\sin\frac{3x}{2}}{2\sin\frac{x}{2}} = \frac{\sin\frac{x}{2} + \left(\sin\frac{3x}{2} - \sin\frac{x}{2}\right)}{2\sin\frac{x}{2}} = \frac{1}{2} + \cos x.$$

2°. Пусть

$$\frac{1}{2} + \cos x + \cos 2x + \ldots + \cos kx = \frac{\sin\frac{2k+1}{2}x}{2\sin\frac{x}{2}}.$$

Тогда

$$\frac{1}{2}+\cos x+\cos 2x+\ldots+\cos kx+\cos(k+1)x=$$
$$=\frac{\sin\frac{2k+1}{2}x}{2\sin\frac{x}{2}}+\cos(k+1)x=\frac{\sin\frac{2k+1}{2}x+2\sin\frac{x}{2}\cos(k+1)x}{2\sin\frac{x}{2}}=$$
$$=\frac{\sin\frac{2k+1}{2}x+\left(\sin\frac{2k+3}{2}x-\sin\frac{2k+1}{2}x\right)}{2\sin\frac{x}{2}}=\frac{\sin\frac{2k+3}{2}x}{2\sin\frac{x}{2}}.$$

18. 1°. При $n=1$ утверждение справедливо, так как

$$\frac{2\sin x-\sin 2x}{4\sin^2\frac{x}{2}}=\frac{2\sin x(1-\cos x)}{4\sin^2\frac{x}{2}}=\sin x.$$

2°. Пусть

$$\sin x+2\sin 2x+\ldots+k\sin kx=\frac{(k+1)\sin kx-k\sin(k+1)x}{4\sin^2\frac{x}{2}}.$$

Тогда

$$\sin x + 2\sin 2x + \ldots + k\sin kx + (k+1)\sin(k+1)x =$$
$$= \frac{(k+1)\sin kx - k\sin(k+1)x}{4\sin^2\frac{x}{2}} + (k+1)\sin(k+1)x =$$
$$= \frac{(k+1)\sin kx - k\sin(k+1)x + 2(k+1)\sin(k+1)x(1-\cos x)}{4\sin^2\frac{x}{2}} =$$
$$= \frac{(k+2)\sin(k+1)x + (k+1)\sin kx}{4\sin^2\frac{x}{2}} -$$
$$- \frac{2(k+1)\cos x\sin(k+1)x}{4\sin^2\frac{x}{2}} =$$
$$= \frac{(k+2)\sin(k+1)x + (k+1)\sin kx}{4\sin^2\frac{x}{2}} -$$
$$- \frac{(k+1)\{\sin(k+2)x + \sin kx\}}{4\sin^2\frac{x}{2}} =$$
$$= \frac{(k+2)\sin(k+1)x - (k+1)\sin(k+2)x}{4\sin^2\frac{x}{2}}.$$

19. 1°. При $n = 1$ утверждение справедлипо, так как

$$\frac{2\cos x - \cos 2x - 1}{4\sin^2\frac{x}{2}} = \frac{2\cos x - 2\cos^2 x}{4\sin^2\frac{x}{2}} = \frac{\cos x(1-\cos x)}{2\sin^2\frac{x}{2}} = \cos x.$$

2°. Пусть

$$\cos 2x + 2\cos 2x + \ldots + k\cos kx = \frac{(k+1)\cos kx - k\cos(k+1)x - 1}{4\sin^2\frac{x}{2}}.$$

Тогда

$$\cos x + 2\cos 2x + \ldots + k\cos kx + (k+1)\cos(k+1)x =$$
$$= \frac{(k+1)\cos kx - k\cos(k+1)x - 1}{4\sin^2\frac{x}{2}} + (k+1)\cos(k+1)x =$$

$$= \frac{(k+1)\cos kx - k\cos(k+1)x - 1}{4\sin^2\frac{x}{2}} +$$

$$+ \frac{2(k+1)\cos(k+1)x(1-\cos x)}{4\sin^2\frac{x}{2}} =$$

$$= \frac{(k+2)\cos(k+1)x + (k+1)\cos kx}{4\sin^2\frac{x}{2}} -$$

$$- \frac{2(k+1)\cos x\cos(k+1)x + 1}{4\sin^2\frac{x}{2}} =$$

$$= \frac{(k+2)\cos(k+1)x + (k+1)\cos kx}{4\sin^2\frac{x}{2}} -$$

$$- \frac{(k+1)[\cos(k+2)x + \cos kx] + 1}{4\sin^2\frac{x}{2}} =$$

$$= \frac{(k+2)\cos(k+1)x - (k+1)\cos(k+2)x - 1}{4\sin^2\frac{x}{2}}.$$

20. 1°. При $n = 1$ утверждение справедливо, так как

$$\frac{1}{2}\operatorname{ctg}\frac{x}{2} - \operatorname{ctg}x = \frac{1}{2}\operatorname{ctg}\frac{x}{2} - \frac{1-\operatorname{tg}^2\frac{x}{2}}{2\operatorname{tg}\frac{x}{2}} = \frac{\operatorname{tg}^2\frac{x}{2}}{2\operatorname{tg}\frac{x}{2}} = \frac{1}{2}\operatorname{tg}\frac{x}{2}.$$

2°. Пусть

$$\frac{1}{2}\operatorname{tg}\frac{x}{2} + \frac{1}{2^2}\operatorname{tg}\frac{x}{2^2} + \ldots + \frac{1}{2^k}\operatorname{tg}\frac{x}{2^k} = \frac{1}{2^k}\operatorname{ctg}\frac{x}{2^k} - \operatorname{ctg}x$$

Тогда

$$\frac{1}{2}\operatorname{tg}\frac{x}{2} + \frac{1}{2^2}\operatorname{tg}\frac{x}{2^2} + \ldots + \frac{1}{2^k}\operatorname{tg}\frac{x}{2^k} + \frac{1}{2^{k+1}}\operatorname{tg}\frac{x}{2^{k+1}} =$$

$$= \frac{1}{2^k}\operatorname{ctg}\frac{x}{2^k} - \operatorname{ctg}x + \frac{1}{2^{k+1}}\operatorname{tg}\frac{x}{2^{k+1}} = \frac{1}{2^{k+1}}\frac{\operatorname{ctg}^2\frac{x}{2^{k+1}} - 1}{\operatorname{ctg}\frac{x}{2^{k+1}}} +$$

$$+ \frac{1}{2^{k+1}\operatorname{ctg}\frac{x}{2^{k+1}}} - \operatorname{ctg}x = \frac{1}{2^{k+1}}\operatorname{ctg}\frac{x}{2^{k+1}} - \operatorname{ctg}x.$$

21. 1°. Имеем

$$\operatorname{tg}(\operatorname{arctg} 2 - \operatorname{arctg} 1) = \frac{2-1}{1+2\cdot 1} = \frac{1}{3}.$$

Поэтому

$$\operatorname{arctg} 2 - \operatorname{arctg} 1 = \operatorname{arctg}\frac{1}{3} = \operatorname{arcctg} 3.$$

Значит, при $n = 1$ утверждение справедливо.
2°. Покажем сначала, что

$$\operatorname{arcctg}(2k+3) = \operatorname{arctg}\frac{k+2}{k+1} - \operatorname{arctg} 1 \qquad (1)$$

Действительно

$$\operatorname{tg}\left(\operatorname{arctg}\frac{k+2}{k+1} - \operatorname{arctg} 1\right) = \frac{\frac{k+2}{k+1}-1}{1+\frac{k+2}{k+1}\cdot 1} = \frac{1}{2k+3}.$$

Значит,

$$\operatorname{arctg}\frac{1}{2k+3} = \operatorname{arcctg}(2k+3) = \operatorname{arctg}\frac{k+2}{k+1} - \operatorname{arctg} 1$$

Предположим, что утверждение справедливо при $n = k$, т. е

$$\operatorname{arcctg} 3 + \operatorname{arcctg} 5 + \ldots + \operatorname{arcctg}(2k+1) = \operatorname{arctg} 2 + \operatorname{arctg}\frac{3}{2} + \ldots + \operatorname{arctg}\frac{k+1}{k} - k\operatorname{arctg} 1 \qquad (2)$$

Докажем, что тогда оно справедливо и при $n = k + 1$, т. е

$$\operatorname{arcctg} 3 + \operatorname{arcctg} 5 + \ldots + \operatorname{arcctg}(2k+3) = \operatorname{arctg} 2 + \ldots + \operatorname{arctg}\frac{k+2}{k+1} - (k+1)\operatorname{arctg} 1 \qquad (3)$$

Сложив почленно равенства (1) и (2), получим равенство (3).
22. 1°. При $n = 1$ утверждение справедливо, так как

$$\sqrt{3} - i = 2\left(\cos\frac{\pi}{6} - i\sin\frac{\pi}{6}\right).$$

2°. Пусть

$$(\sqrt{3} - i)^k = 2^k\left(\cos\frac{k\pi}{6} - i\sin\frac{k\pi}{6}\right).$$

Тогда

$$(\sqrt{3} - i)^{k+1} = 2^k\left(\cos\frac{k\pi}{6} - i\sin\frac{k\pi}{6}\right)2\left(\cos\frac{\pi}{6} - i\sin\frac{\pi}{6}\right) = 2^{k+1}\left[\cos\frac{(k+1)\pi}{6} - i\sin\frac{(k+1)\pi}{6}\right].$$

23. 1°. При $n = 1$ утверждение справедливо.
2°. Пусть

$(\cos x + i \sin x)^k = \cos kx + i \sin kx.$

Тогда

$$(\cos x + i \sin x)^{k+1} = (\cos kx + i \sin kx)(\cos x + i \sin x) =$$
$$= (\cos kx \cos x - \sin kx \sin x) + i(\cos kx \sin x + \sin kx \cos x) =$$
$$= \cos(k+1)x + i \sin(k+1)x.$$

24. Ошибочна самая последняя фраза «Утверждение доказано». В действительности доказано что неравенство

$$2^n > 2n + 1$$

справедливо при $n=k+1$, если оно справедливо при $n = k$, где k — любое натуральное число.

Отсюда еще не следует, что неравенство это справедливо хотя бы при одном значении n и тем более при любом натуральном n

Короче говоря, ошибка заключается в том, что доказана только теорема 2, а теорема 1 не рассматривалась и база для индукции не создана

25. Легко видеть, что 3 — наименьшее натуральное значение n, при котором неравенство $2^n > 2n + 1$ справедливо.

Учитывая что из справедливости неравенства при $n = k$ следует его справедливость при $n=k+1$ (задача 23), утверждаем что неравенство справедливо при любом натуральном $n \geq 3$.

26. 1°. При $n = 2$ неравенство справедливо так как

$$1 + \frac{1}{\sqrt{2}} > 1 \quad \sqrt{2}$$

2° . Пусть

$$\frac{1}{1\sqrt{1}} + \frac{1}{\sqrt{2}} + \ldots + \frac{1}{1\sqrt{k}} > \sqrt{k} \qquad (1)$$

Докажем, что

$$\frac{1}{1\sqrt{1}} + \frac{1}{\sqrt{2}} + \ldots + \frac{1}{1\sqrt{k}} + \frac{1}{\sqrt{k+1}} > \sqrt{k+1} \qquad (2)$$

При любом $k \geq 0$ имеет место неравенство

$$\frac{1}{1\sqrt{k+1}} > 1\sqrt{k+1} - \sqrt{k} \qquad (3)$$

Действительно, неравенство (3) равносильно неравенству

$$1 + \sqrt{\frac{k}{k+1}} > 1$$

полученному из него умножением обеих частей на $\sqrt{k+1} + \sqrt{k}$. Сложив почленно неравенства (1) и (3), получим неравенство (2).

27. 1° При $n = 2$ неравенство принимает вид $\frac{16}{3} < 6$ и следовательно, справедливо.

2°. Пусть

$$\frac{4^k}{k+1} < \frac{(2k)!}{(k!)^2},$$

где $k \geq 2$. Нетрудно проверить, что при $k > 0$

$$\frac{4(k+1)}{k+2} < \frac{(2k+1)(2k+2)}{(k+1)^2}.$$

Поэтому

$$\frac{4^k}{k+1} \cdot \frac{4(k+1)}{k+2} < \frac{(2k)!}{(k!)^2} \cdot \frac{(2k+1)(2k+2)}{(k+1)^2},$$

т. е.

$$\frac{4^{k+1}}{k+2} < \frac{(2k+2)!}{[(k+1)!]^2}.$$

Микромодуль 15
Признаки делимости

Изложение основных фактов, относящихся к признакам делимости, является в ней поводом затронуть некоторые довольно абстрактные вопросы дискретной математики. К числу таких вопросов относятся, прежде всего, утверждения элементарной теории чисел, группирующиеся вокруг основной теоремы арифметики и анализа канонического разложения натурального числа на простые множители. Далее, сама делимость чисел рассматривается как отношение на множестве целых чисел, т. е. как реализация довольно общего и абстрактного понятия. Наконец, признаки делимости трактуются здесь как алгоритмы, перерабатывающие каждое число в ответ, делится ли оно на данное число или не делится. Мы сочли целесообразным среди признаков делимости особо выделить «признаки равноостаточности», перерабатывающие числа в остатки при их делении на данное число.

Для того чтобы оттенить разнообразные взаимосвязи между отдельными математическими фактами и возможности различных подходов к одному и тому же предмету, некоторые утверждения устанавливаются двумя различными путями.

4.5. Делимость чисел

1. Сумма, разность и произведение двух целых чисел — всегда целые числа. Этот факт иногда принято называть *замкнутостью* множества целых чисел по отношению к действиям сложения, вычитания и умножения.

По отношению же к действию деления множество всех целых чисел замкнутым не является: частное от деления одного целого числа на другое может, вообще говоря, и не быть целым.

Поэтому при изучении обстоятельств, связанных с делением целых чисел, одним из первых встает вопрос о выполнимости этого действия для данных двух чисел, т. е. о *делимости* этих чисел. При рассмотрении остальных арифметических действий над целыми числами подобный вопрос, очевидно, не возникает.

В дальнейшем мы будем считать известными основные свойства арифметических действий над целыми числами, а также простейшие свойства равенств и неравенств. Под «числом» всегда, если не оговорено противное, будет пониматься целое число.

Как обычно целые неотрицательные числа: 0, 1, 2, ... будут называться *натуральными.* Говоря о всех натуральных числах, мы будем пользоваться термином *множество всех натуральных чисел.*

Определение. Число *a делится* на число *b* (или, что то же самое, число *b делит* число *a),* если существует такое число *c*, что $a = bc$.

Этот факт называется *делимостью* числа *a* на число *b* и обозначается как $a \,\vdots\, b$.

Подчеркнем, что запись $a \,\vdots\, b$ означает не какое-то действие, которое надлежит произвести над числами *a* и *b,* а некоторое утверждение, касающееся этих чисел. В зависимости от того, каковы числа *a* и *b,* утверждение $a \,\vdots\, b$ может быть верным или неверным. Так, например, $4 \,\vdots\, 2$ верно, а $4 \,\vdots\, 3$ — нет.

Для выяснения того, является ли утверждение $a \,\vdots\, b$ верным или нет, т. е. для выяснения делимости числа *a* на число *b,* имеется довольно много разнообразных способов. Один из них состоит в непосредственном делении числа *a* на число *b.* Однако такое деление часто оказывается слишком долгим и утомительным занятием, и естественно появляется желание установить истинность интересующей нас делимости, не производя фактического деления. Не лишним представляется и такое соображение: пока нас интересует

только факт делимости числа a на число b; если же мы выполним деление, то мы попутно узнаем еще и частное от этого деления и остаток от него (если деление нацело «не получилось»); все эти числа, однако, для нас никакой ценности не представляют, так как мы в данный момент интересуемся только тем, будет ли остаток от деления равен нулю или нет. Значит, есть основания предполагать, что выполняя деление, мы какую-то (и по-видимому, немалую) часть работы потратили на получение «отходов производства». Можно надеяться, что более прямые способы выяснения делимости, чем «грубое» деление, которые не дадут нам столь обильных отходов, будут экономнее и позволят установить факт делимости более коротким путем. Эти надежды в действительности оправдываются, и такие способы выяснения делимости существуют. Они называются *признаками делимости.*

Некоторые признаки делимости, несомненно, известны читателю. Целью этой работы является рассмотрение различных признаков делимости, главным образом с принципиальной стороны.

Сущность всякого признака делимости на данное число b состоит в том, что при его помощи вопрос о делимости любого числа a на b сводится к вопросу о делимости на b некоторого числа, меньшего чем a, (Нетрудно видеть, что проверка делимости обычным делением также основана на этой идее.)

Таким образом, признак делимости является математическим объектом весьма распространенной, хотя и не бросающейся в глаза природы. Это не формула, не теорема, не определение, а некоторый процесс, совершенно такого же типа, что и процесс умножения чисел «столбиком» или, скажем, процесс вычисления одного за другим членов арифметической прогрессии.

Понятие признака делимости будет уточнено в следующем пункте.

2. В определении делимости чисел ничего не говорится о том, сколько различных значений может иметь частное от деления a на b. Выясним здесь этот вопрос до конца, чтобы в дальнейшем к нему больше не возвращаться.

Пусть

$$a = bc, \tag{1}$$

и вместе с тем

$$a = bc'.$$

Из этих равенств мы получаем

$$bc = bc',$$

или

$$b\,(c - c') = 0.$$

Если при этом $b \neq 0$, то $c - c' = 0$, т. е. $c=c'$. Если же $b = 0$, то, очевидно, и $a = 0$, а равенство (1) выполняется при любом c.

Таким образом, на нуль делится только нуль, а частное от такого деления неопределенно. Именно это и имеется в виду, когда говорят о невозможности деления на нуль. Если же делитель отличен от нуля и делимость имеет место, то частное имеет одно, вполне определенное значение.

Говоря о делении, мы всегда будем предполагать делитель отличным от нуля.

Установим несколько простейших свойств делимости.

Теорема 1. $a \vdots a$.

Это свойство делимости называется ее *рефлексивностью* (или возвратностью).

Доказательство.

Достаточно заметить, что $a = a \cdot 1$.

Теорема 2. *Если* $a \vdots b$ *и* $b \vdots c$, *то* $\vdots c$,

Это свойство делимости называетсяее *транзитивностью* (или переходностью).

Доказательство.

По условию, найдутся такие d_1 и d_2, что $a = bd_1$ и $b = cd_2$. Но тогда $a = cd_1d_2$, т. е. $a \mid c$.

Теорема 3. *Если* $a \vdots b$ *и* $b \vdots a$, *то либо* $a - b$, *либо* $a = -b$ (*антисимметричность* делимости).

Доказательство.

Мы имеем $a = bc_1$ и $b = ac_2$, откуда следует, что $a = ac_1c_2$, т. е. $c_1c_2 = 1$. Так как числа c_1 и c_2 по условию целые, то либо $c_1 = c_2 = 1$, либо $c_1 = c_2 = -1$. В первом из этих случаев $a = b$, а во втором $a = -b$.

Теорема 4. *Если,* $a \vdots b$ *и* $|b| > |a|$, *то* $a = 0$.

Доказательство.

Пусть $a = bc$. Если $|c| \geq 1$, то поскольку $|b| > |c|$, должно быть и $|bc| > |a|$, что, однако, противоречит предположенному. Значит, $|c| < 1$, а так как по условию число c целое, должно быть $c=0$, а потому и $a = 0$.

Следствие. *Если* $a \vdots b$ *и* $a \neq 0$, *то* $|a| \geq |b|$.

Теорема 5. *Для того чтобы* $a \vdots b$, *необходимо и достаточно, чтобы* $|a| \vdots |b|$.

Доказательство.

Очевидно, из $a = bc$ следует $|a| = |b||c|$, а из $|a| = |b||c|$ следует $a = b\,c$ или $a=b\,(-c)$, причем числа c, $-c$ и $|c|$ целые или нет одновременно.

На основании этой теоремы в дальнейшем достаточно ограничиваться рассмотрением случая, когда делитель есть положительное число. Равным образом делимость произвольных целых чисел сводится к делимости неотрицательных чисел.

Теорема 6. *Если* $a_1 \vdots b,\ a_2 \vdots b,\ \ldots,\ a_n \vdots b$, *то*

$$(a_1 + a_2 + \ldots + a_n) \vdots b.$$

Доказательство.

В самом деле, пусть

$$a_1 = bc_1,$$
$$a_2 = bc_2,$$
$$\cdot\ \cdot\ \cdot\ \cdot$$
$$a_n = bc_n,$$

где все числа $c_1, c_2, \ldots, c_n$ — целые. Сложив все эти равенства почленно, мы получим

$$a_1 + a_2 + \ldots + a_n = b(c_1 + c_2 + \ldots + c_n).$$

В скобках стоит целое число, что и доказывает требуемое.

Следствие. Если сумма двух чисел и одно из слагаемых делится на некоторое число *b*, то другое слагаемое также делится на *b*.

Не следует считать все эти теоремы очевидными и не нуждающимися в каком-либо особом доказательстве. Дело здесь даже не в том, что в математике доказательству подлежит всякое утверждение, кроме ***аксиом*** и ***определений***. Доказательства этих фактов (например того, что всякое число делится на себя) принципиально необходимы, так как они не могут быть получены только из определения делимости, а нуждаются в использовании свойств самих чисел.

Подробнее в этом разобраться нам поможет следующий пример.

Очевидно, сумма, разность и произведение четных чисел всегда четны. Вместе с тем деление одного четного числа на другое не всегда выполнимо, а если и выполнимо, то частное не обязательно четно. Поэтому можно ввести понятие четной делимости четных чисел.

Определение. Четное число *a четно делится* на четное число *b*, если существует такое четное число *c*, что $a = bc$.

Очевидно, для четной делимости теорема 1 неверна, так как, например, не существует такого четного числа *c,* для которого $a = ac$.

К вопросам, связанным четной делимостью четных чисел, мы еще будем несколько раз возвращаться. Пример четной делимости показывает, что можно строить различные теории делимости с различными свойствами, и теоремы, верные для одних таких теорий, могут оказаться неверными для других.

3. Уже при самом беглом знакомстве с конкретными фактами делимости бросается в глаза следующее обстоятельство: возможности делимости чисел практически не связаны с их величиной. С одной стороны, существуют маленькие числа, которые делятся на сравнительно большое количество чисел. Например, 12 делится на 1, 2, 3, 4, 6 и 12; число 60 имеет 32 делителей. Таким богатым делителями числам можно противопоставить весьма большие числа, которые имеют минимальное число делителей — 2 (согласно теореме 1 и задаче 2, каждое отличное от единицы число делится хотя бы на два различных числа). Хотя в действительности и известны некоторые закономерности, связывающие свойства делимости чисел с их величиной, но эти закономерности носят столь сложный и запутанный характер, что мы не будем их здесь касаться.

4. Тем более интересным оказывается тот факт, что сама делимость позволяет установить среди чисел некоторый порядок, отличающийся от их обычного порядка по величине, но имеющий с ним много общего.

В самом деле, вдумаемся, какой точный смысл вкладывается в слова о возможности упорядочить натуральные числа по их величине. Под этой возможностью, как нетрудно видеть, понимается то, что для некоторых пар чисел a и b имеет место отношение «больше или равно»:

$$a \geq b \text{ ,}$$

которое означает, что разность a — b неотрицательна (т. е. должно существовать такое натуральное число с, что $a = b+c$). Но ведь и явление делимости состоит в том, что некоторые пары чисел a и b подчиняются некоторому, вполне определенному условию (именно, существует такое целое c, что $a = bc$). Таким образом, отношение делимости и отношение «больше или равно» представляют собой понятия одной природы, и потому можно говорить об их общих свойствах или, наоборот, противопоставлять их друг другу.

В частности, подобно отношению делимости, отношение «больше или равно» между двумя натуральными числами является некоторым высказыванием об этих числах и может быть верным (например, $5 \geq 3$) или неверным (например, $3 \geq 5$).

Заметим сразу же, что отношение «больше или равно» имеет больше общих свойств с отношением делимости, чем отношение «больше». Это связано с тем, что отношение «больше или равно», подобно отношению делимости, рефлексивно (действительно, соотношение $a \geq a$ справедливо для любого a), а отношение «больше» рефлексивным

не является (неравенство $a > a$ не имеет места никогда). Именно поэтому здесь в качестве отношения порядка между натуральными числами рассматривается отношение «больше или равно», а не, казалось бы, более простое и естественное отношение «больше».

5. Отношение $\geq$ обладает следующими легко проверяемыми свойствами:

1° $a \geq a$ (рефлексивность).

2° Если $a \geq b$ и $b \geq a$, то $a = b$ (антисимметричность).

3° Если $a \geq b$ и $b \geq c$, то $a \geq c$ (транзитивность).

4° Во всякой последовательности натуральных чисел

$$a_1 \geq a_2 \geq a_3 \geq \ldots \geq a_n \geq \ldots,$$

все члены которой отличны друг от друга, найдется последнее число. Это свойство отношения иногда называется свойством *полной упорядоченности* множества натуральных чисел.

Свойство полной упорядоченности довольно сложно по формулировке и выглядит несколько искусственно. Однако оно вскрывает чрезвычайно важные черты в строении множества натуральных чисел, упорядоченных отношением $\geq$. Из него выводятся многие другие свойства этого отношения. Кроме того, мы увидим, что именно на нем основаны столь употребительные в разных вопросах математики рассуждения «по индукции».

В качестве полезного применения этого свойства отметим следующее: существует такое число a, что из $a \geq b$ следует a — b (здесь a и b — натуральные числа).

В самом деле, если бы такого числа не было, то мы могли бы по каждому a_n находить такое a_{n+1}, что $a_n \geq a_{n+1}$ и $a_n \neq a_{n+1}$. Начав с произвольного a_1, мы получили бы последовательность

$$a_1 \geq a_2 \geq a_3 \geq \ldots \geq a_n \geq a_{n+1} \geq \ldots$$

которая никогда не кончается. Но существование такой последовательности противоречит свойству полной упорядоченности множества натуральных чисел.

Таким образом, указанное число a действительно существует. Оно называется *первым,* или *минимальным,* числом (очевидно, это нуль). Заметим здесь же, что мы сейчас не установили единственности минимального числа. Эта единственность будет зафиксирована далее косвенным путем.

5° Каково бы ни было число a, существует отличное от a число b, для которого $b \geq a$.

Это свойство множества натуральных чисел называется его *неограниченностью* в смысле отношения $\geq$.

6° Каково бы ни было число a, не являющееся минимальным, существует такое b, что $a \geq b$, $a \neq b$, и для любого числа c из $a \geq c \geq b$ следует либо $c = a$, либо $c = b$. Это формальное утверждение в переводе на содержательный язык означает, что каждое натуральное число, кроме 0, имеет непосредственно предшествующее натуральное число. (Иначе это можно сформулировать так: среди всех чисел, меньших данного, есть наибольшее)

7° Либо $a \geq b$, либо $b \geq a$. Это свойство отношения называется его *дихотомичностью*. В математике термин дихотомичность обычно выражает обязательную реализацию одной из двух возможностей. Само это слово греческого происхождения и означает *разделение на две части*.

Подчеркнем, что 1°—7° являются свойствами самого отношения на множестве всех натуральных чисел, а не свойствами тех или иных чисел, связываемых этим отношением. Поэтому может оказаться, что для какого-нибудь другого отношения, связывающего числа в пары, но не по величине, а каким-либо иным способом, некоторые из утверждений 1°—7° могут оказаться и неверными.

6. Справедливость свойств отношения $\geq$ (как впрочем, и любого другого отношения) может быть установлена двояко. Во-первых, мы можем воспользоваться свойствами тех или иных чисел или известными особенностями строения множества всех натуральных чисел. Именно так проверялись нами свойства 1° — 7°. Во-вторых, мы можем, уже убедившись в справедливости свойств 1° —7°, отвлечься от того, что отношение $\geq$ связывает числа в пары и выводить дальнейшие свойства этого отношения только из его свойств 1°— 7°. Так были нами доказаны существование минимального числа и утверждения задачи 8

Второй подход к вопросу весьма употребителен в математике и носит название *аксиоматического* При таком подходе устанавливаются некоторые *аксиомы* (в нашем случае ими являются утверждения 1°—7°), которые отражают основные свойства изучаемых предметов и не подлежат доказательству, а из них чисто логическим путем, без повторного обращения к свойствам исследуемых предметов, выводятся все остальные утверждения, которые называются *теоремами*.

Быть может, некоторым из читателей рассмотрение свойств отношений в отрыве от связываемых этими отношениями объектов (например, чисел) покажется тем верхом математической абстракции, который в практической жизни совершенно не нужен. По этому поводу следует сделать два замечания.

Во-первых, с точки зрения математики все приводимые здесь рассуждения вовсе не являются «особенно абстрактными». Более того, математикам приходится рассматривать одновременно много отношений и даже (!) связывать пары различных отношений новыми отношениями (так сказать, отношениями «второго порядка»).

Изложенный до сих пор материал позволяет проиллюстрировать понятие отношения между отношениями примером.

Пусть α, β, ... — некоторый набор отношений, связывающих натуральные числа. Это значит, что для любой пары чисел *a* и *b* и любого отношения γ из нашего набора мы знаем, связывается ли пара *a, b* отношением γ или нет. Если *a* и *b* отношением γ связаны, будем писать *aγb*.

Будем говорить, что отношение α *сильнее* отношения ρ, и записывать это как $\alpha \supset \beta$, еели любая пара чисел, связанная отношением β, оказывается также связанной и отношением α, т. е. если из *aβb* следует *aαb*.

Так, например, обозначая отношение четной делимости через $\vdots_{ч}$, мы можем записать $\vdots \supset \vdots_{ч}$. Далее, очевидно, что $\geqq \supset >$.

Вместе с тем существуют естественные отношения на множестве натуральных чисел, относительно которых нельзя утверждать, что одно сильнее или слабее другого. Так, например, если для двух натуральных чисел *a* и *b* полагать *a* f *b,* если последняя цифра в десятичной записи числа *a* больше последней цифры числа *b,* то ни $> \supset \geqq$, ни $\geqq \supset >$.

Конечно, для свободного оперирования столь сложными понятиями как отношения между отношениями необходима специальная тренировка.

Во-вторых, такие и даже еще более отвлеченные рассуждения все чаще и чаще начинают встречаться в приложениях математики к экономике, биологии, лингвистике, военному делу.

7. С упорядоченностью множества натуральных чисел отношением ≥ тесно связана возможность применять метод *полной индукции* (называемый также методом *совершенной* индукции или методом *математической* индукции). Обычно этот метод применяется в следующей форме. Пусть *A(n)* — некоторое утверждение, касающееся произвольного натурального числа *n*. Это, по существу, означает, что мы имеем дело с бесконечной последовательностью утверждений

$$A(0),\ A(1),\ \ldots.\ A(n),\ \ldots$$

о каждом из натуральных чисел. Предположим, что

а) справедливо утверждение $A(0)$ («основание индукции») (часто за основание индукции принимают утверждение $A(1)$. Очевидно, это различие не является существенным. Важно лишь, что основание индукции касается первого из рассматриваемых нами чисел);

б) из справедливости утверждений $A(n)$ следует справедливость утверждения $A(n+1)$ («индуктивный переход»).

Принцип математической индукции утверждает, что в предположениях а) и б) $A(n)$ справедливо для любого натурального n.

Принцип математической индукции не является каким-то самостоятельным утверждением, а может быть выведен из свойств 1° — 7° упорядочения множества натуральных чисел отношением $\geq$.

Действительно, предположим, что условия а) и б) принципа индукции для утверждений $A(n)$ выполнены, но заключение этого принципа не имеет места. Последнее означает, что должны существовать такие числа m, для которых утверждение $A(m)$ неверно. Пусть m_1 — одно из таких чисел. Если для всех $n < m_1$ утверждение $A(n)$ верно, то m_1 — наименьшее из чисел, для которых $A(n)$ не имеет места. Если же $A(n)$ верно не для всех $n < m_1$, то должно существовать такое $m_2 < m_1$, что $A(m_2)$ неверно.

В итоге мы приходим к некоторой последовательности различных чисел

$$m_1 \geqq m_2 \geqq \ldots \geqq m_r \geqq, \ldots, \qquad (2)$$

для каждого из которых $A(m)$ не имеет места. По свойству полной упорядоченности 4° в последовательности (2) должен быть последний член m_r. Очевидно, число m_r является наименьшим из всех чисел, для которых $A(n)$ неверно.

Поскольку $A(0)$ верно по условию, $m_r \neq 0$, так что существует число m^*_r, непосредственно предшествующее m_r (в действительности этим числом является m_r — 1). Так как $m^*_r < m_r$ утверждение $A(m^*_r)$ должно быть верным. Но тогда по условию б) принципа математической индукции должно быть верным также и утверждение $A(m^*_r + 1)$, т.е. $A(m_r)$, и мы получили противоречие. Это противоречие показывает, что нет чисел m, для которых $A(m)$ не имело бы места (т е. не было бы справедливо).

Сделаем следующее замечание. Проведенные только что рассуждения не следует считать ни доказательством принципа индукции, ни его обоснованием. Они означают лишь возможность вывода одного математического утверждения (метода индукции) из других (из свойств отношения $\geq$). Сами же эти свойства принимались нами в качестве аксиом, и потому не доказывались, а лишь проверялись.

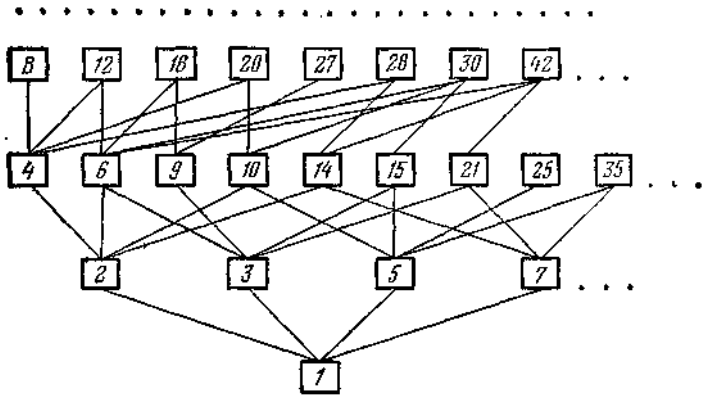
Всякая попытка их математического доказательства неизбежно натолкнулась бы на необходимость введения в качестве аксиом каких-то новых условий.

В частности, доказательства свойства полной упорядоченности должны использовать те же индуктивные рассуждения (читатель может в этом убедиться сам).

Подробно метод математической индукции в его различных вариантах рассмотрен нами в микромодуле 14 настоящей работы. На протяжении данного микромодуля этот метод также будет часто применяться.

8. Вернемся, однако, к отношению делимости. В случае положительных чисел теоремы 1, 2, 3 и задачи 3, 4 и 5 показывают, что в утверждениях 1°—6° мы можем заменить отношение $\geq$ отношением $\mathbb{N}$. Что же касается утверждения 7°, то в применении к делимости оно гласит: «из двух чисел хотя бы одно делится на другое».

Но это неверно. Таким образом, отношение делимости обладает всеми свойствами отношения порядка за исключением одного. В связи с этим отношение делимости упорядочивает натуральные числа не в виде линейной цепочки, а иным, более сложным образом (см. рисунок).



Заметим, что числа, близкие по величине, могуг оказаться довольно «далекими» друг от друга в смысле делимости. Наглядно демонстрируют это числа 4 и 5 или 7 и 8.

Попробуем от делимости целых положительных чисел перейти к делимости чисел натуральных, т. е. включим в рассмотрение нуль. Тогда схема на рисунке пополнится клеткой, лежащей выше всех остальных клеток схемы, ибо нуль делится на любое число и ни одно из чисел, отличных от нуля, на нуль не делится.

Читателю предоставляется самостоятельно переформулнровать и проверить утверждения 1° — 7° для этого случая.

9. Определение. Любое отношение $\mathfrak{f}$, подчиненное условиям:

1° рефлексивности ($a \mathfrak{f} a$)

2° антисимметричности (из $a \mathfrak{f} b$ и $b \mathfrak{f} a$ следует $a = b$);

3° транзитивности (из $a \mathfrak{f} b$ и $b \mathfrak{f} c$ следует $a \mathfrak{f} c$),

называется *частично упорядочивающим отношением.* Частично упорядочивающие отношения играют большую роль там, где «настоящее», линейное упорядочение не имеет места, например там, где каждый объект описывается или оценивается по нескольким различным, качественно несравнимым между собой показателям.

В качестве примера можно привести оценку результатов спортивных соревнований по нескольким различным видам спорта. Если одна из команд заняла по всем видам программы соревнований более высокие места, чем другая, то естественно считать, что первая команда добилась больших успехов. Если же эти более высокие места были заняты по всем видам программы, за исключением, скажем, игры в крокет (которая почему-то на этот раз оказалась включенной в программу соревнований), где вторая команда оказалась сильней, то вопрос об окончательном распределении мест между нашими командами оказывается уже не столь очевидными. Энтузиасты игры в крокет могут даже настаивать на более высоком месте для второй команды. Во всяком случае любое суммарное распределение мест должно быть связано с некоторыми условными пересчетами (например, с приписыванием очков).

10. Условия 1° — 3°, соблюдение которых делает отношение $\mathfrak{f}$ отношением частичного упорядочения, являются довольно свободными. Поэтому частично упорядоченными и притом упорядоченными весьма различными способами могут быть самые разнообразные объекты. В связи с этим о произвольном частично упорядочивающем отношении мало что можно сказать сверх того, что оно частично упорядочивающее. В частности, к объектам, для которых определено частично упорядочивающее отношение, нельзя, вообще говоря, применить метод математической индукции.

Дополним, однако, условия 1° — 3° следующими:

4° полная упорядоченность;

5° неограниченность;

6° каждый объект, отличный от минимального, имеет непосредственно предшествующий;

8° каждый объект имеет не более конечного числа предшествующих;

9° каковы бы ни были a и b f a ($b \neq a$), существует такое c, непосредственно предшествующее b, что c f a.

Оказывается, что на основе частичной упорядоченности множества натуральных чисел отношением, которое удовлетворяет условиям 1° — 6°, 8° и 9°, можно построить некоторое видоизменение метода индукции, состоящее в следующем.

Пусть снова $A(n)$ — утверждение, касающееся произвольного числа n. Предположим, что

а) Справедливо утверждение $A(a)$, где a есть минимальное число в смысле упорядочения f ;

б) Если n — некоторое число, и справедливость всех утверждений вида $A(m)$ для всех таких m, что n f m и $n \neq m$, установлена, то верно и утверждение $A(n)$.

Новая форма принципа индукции утверждает, что при соблюдении условий а) и б) утверждение $A(n)$ справедливо при любом n

Так как отношение делимости условиям 1[e] — 6°, 8° и 9" удовлетворяет (сформулируйте и проверьте для отношения делимости условия 8° и 9°), этот принцип индукции к отношению делимости применим.

В применении к делимости новый принцип индукции может быть сформулирован так: если некоторое утверждение $A(n)$ справедливо при $n=1$ и из справедливости его для всех делителей числа n, отличных от n, следует его справедливость для n, то оно имеет место для любого числа.

11. Деление целых чисел, как мы видели, выполнимо не всегда. Поэтому целесообразно наряду с действием деления рассматривать и другое, более общее действие, которое всегда выполнимо, а в случае выполнимости действия деления, по существу, совпадает с ним. Таким действием является *деление с остатком.*

Определение. *Разделить* число a на число b ($b > 0$) *с остатком* — значит представить число a в виде

$$a = bq + r,$$

где $0 \leq r \leq b$.

Число q при этом называется *неполным частным,* а число r — *остатком* от деления a на b. Очевидно, $r = 0$ тогда и только тогда, когда a N b. В этом случае q равно частному от деления a на b.

Покажем, что деление с остатком всегда выполнимо, а неполное частное и остаток вполне определяются делимым и делителем, т. е. единственны.

Пусть сначала $a \geq 0$. Будем выписывать одно за другим числа

$$a,\ a-b,\ a-2b,\ \ldots \qquad (3)$$

до тех пор, пока не появится отрицательное число (очевидно, рано или поздно такое число должно появиться (точнее говоря, это следует из полной упорядоченности множества натуральных чисел отношением $\geq$)). Пусть последним из неотрицательных членов последовательности (3), т. е. самым маленьким из них, окажется число $a - bq$. Обозначая его через r, мы имеем

$$a = bq + r. \qquad (4)$$

Очевидно, $r < b$ (иначе бы число $r - b$, т. е. $a - (q + 1)b$, было бы неотрицательным, а этого не может быть, так как r — наименьшее из неотрицательных чисел среди (3)). Таким образом, (4) и является искомым представлением числа a.

Пусть теперь $a<0$. Рассуждая аналогично предыдущему, будем выписывать последовательность чисел

$$a,\ a+b,\ a+2b,\ \ldots$$

до тех пор, пока не появится первое неотрицательное число r (легко проверить, что $r<b$). Пусть

$$r = a + bq'.$$

Тогда, обозначая $-q'$ через q, мы получаем

$$a = bq + r,$$

а это и требовалось.

Возможность деления с остатком доказана во всех случаях.

Докажем теперь однозначность этого деления, т. е., что из

$$a = bq + r \qquad (5)$$

и

$$a = bq_1 + r_1, \qquad (6)$$

следует

$$q = q_1 \quad \text{и} \quad r = r_1.$$

От такого доказательства единственности нельзя отмахнуться попросту, заявив, что так как, дескать, действие вычитания однозначно, последовательность (3) может быть построена единственным способом; последний ее неотрицательный член также вполне определен; пусть это будет наше r... и т. д. Такое рассуждение еще не избавляет нас от возможности получить другие значения q и r каким-нибудь совершенно иным путем.

Сопоставляя отношения (5) и (6), мы видим, что

$$bq + r = bq_1 + r_1$$

откуда

$$r - r_1 = b(q_1 - q),$$

т. е. $r - r_1$ делится на b. Но $|r - r_1| < b$, а по теореме 4 это возможно лишь при

$$r - r_1 = 0,$$

т. е. при $r = r_1$. Но тогда

$$b(q_1 - q) = 0,$$

и ввиду неравенства нулю числа b

$$q_1 - q = 0$$

т. е. $q_1 = q$. Однозначность деления с остатком доказана. Таким образом, нами доказана следующая теорема.

Теорема 7 (о делении с остатком). *Для произвольных чисел a и b ($b > 0$) существуют и единственны такие числа r и q, что*

$$a = bq + r,$$

причем $0 \le r < b$.

Заметим, что в частности при $b = 1$ должно быть $r = 0$, откуда $a = q$. Это соответствует утверждению задачи 2. Ясно вместе с тем, что если $b > 1$, то $a > q$.

12. Определение. Число p, не равное единице, называется *простым,* если оно делится только на себя и на единицу.

Простыми числами являются, например, числа 2, 3, 5, 7, 11, 13 и т. д.

Число, отличное от единицы и не являющееся простым, называется *составным.*

Теорема 8. *Простых чисел бесконечно много.*

Доказательство.

Доказательство ведется от противного. Предположим, что простых чисел конечное число, так что все они могут быть выписаны:

$$p_1, p_2, \ldots, p_n. \qquad (*)$$

Произведение всех этих чисел обозначим через P и рассмотрим разность $P - 1$. Эта разность больше каждого из простых чисел, перечисленных в списке (*), и потому не может быть простым числом. Следовательно, она делится хотя бы на одно простое число p_k. Но P также делится на p_k. Следовательно, на основании следствия теоремы 6, должно быть и $1 \mid p_k$, откуда следует, что $p_k = 1$, а это противоречит простоте числа p_k.

Приведенное доказательство бесконечности множества простых чисел было найдено Евклидом (IV век до н. э.).

Всякое число, делящее одновременно числа a и b, называется *общим делителем* этих чисел. Наибольший из общих делителей чисел a и b

называется их *наибольшим общим делителем* и обозначается обычно через *(a, b).*

Если наибольший общий делитель чисел *a* и *b* равен единице, то эти числа называются *взаимно простыми.*

Иначе говоря, числа *a* и *b* называются взаимно простыми, если они одновременно не делятся ни на какое число кроме единицы.

Теорема 9. *Если a* и *p* — *натуральные числа, причем число p простое, то либо a* ⋮ *p, либо числа a и p взаимно просты.*

Доказательство.

Если числа *a* и *p* взаимно просты, то теорема доказана. Если же эти числа не взаимно просты, то оба они делятся на одно и то же число, отличное от единицы. Ввиду простоты *p* таким числом может быть только само *p.* Значит, в этом случае *a* ⋮ *p,* а это и требовалось.

Всякое число, делящееся одновременно на числа *a* и *b,* называется *общим кратным* этих чисел. Наименьшее положительное общее кратное *a* и *b* называется *наименьшим общим кратным* этих чисел.

Теорема 10. Если *M — общее кратное a и b, а m — их наименьшее общее кратное, то M* ⋮ *m.*

Доказательство.

Разделив *M* на *m* с остатком, получим

$$M = mq + r,$$

где $0 \le r < m$. Так как *M* и *m* делятся на *a* и *b,* по следствию теоремы 6, число *r* также должно делиться и на *a,* и на *b* и тем самым быть общим кратным этих чисел. Но $r < m$, а *m* есть наименьшее положительное общее кратное *a* и *b.* Значит, *r* не может являться положительным числом, так что $r = 0$. Поэтому *M* ⋮ *m.*

Теорема 11. *Наименьшее общее кратное двух взаимно простых чисел равно их произведению.*

Доказательство.

Пусть числа *a* и *b* взаимно просты, и *m* — их наименьшее общее кратаое. Так как *ab* ⋮ *a* и *ab* ⋮ *b* по предыдущей теореме *ab* ⋮ *m.* Пусть *ab = mk.* Положим *m = ac.* Тогда *ab = ack,* т. е. *b = ck,* так что *b* ⋮ *k.* Точно так же убеждаемся в том, что и *a* ⋮ *k.* Так как числа *a* и *b* по условию взаимно простые, должно быть $k = 1$, а это и означает, что *m=ab.*

Следствие. Для того чтобы число *a* делилось на взаимно простые числа *b* и *c,* необходимо и достаточно, чтобы оно делилось на их произведение.

Теорема 12. *Если* $ab \vdots c$, *причем числа* b *и* c *взаимно простые, то* $a \vdots c$.

Доказательство.

Обозначим через m наименьшее общее кратное чисел b и c. По предыдущей теореме $m = bc$. Далее, по условию $ab \vdots c$, кроме того, очевидно, $ab \vdots b$. Значит, по теореме 10, $ab \vdots bc$, т. е. $ab = bck$ или, после сокращения на b, $a = ck$, а это и требовалось.

Теорема 13. *Если произведение нескольких сомножителей делится на простое число* p, *то хотя бы один из сомножителей делится на* p.

Доказательство.

Доказательство ведется индукцией по числу сомножителей. Если сомножитель один, то теорема тривиальна. Предположим, что теорема доказана для любого произведения n сомножителей. Пусть $a_1a_2 \ldots a_na_{n+1} \vdots p$. Обозначим $a_1a_2 \ldots a_n$ через A. Тогда $Aa_{n+1} \vdots p$. Если $a_{n+1} \vdots p$, то теорема доказана, а если нет, то по теореме 9 числа a_{n+1} и p взаимно просты. Но тогда по предыдущему $A \vdots p$. Так как A есть произведение n сомножителей, по индуктивному предположению один из них должен делиться на p. Теорема доказана.

Следствие. Вся дробь представляет собой целое число, т. е. ее числитель делится на знаменатель. Будем считать, что числитель является произведением двух сомножителей: p и $1 \cdot 2 \ldots (p-1)=(p-1)!$

Ни один из сомножителей знаменателя дроби не делится на р. Следовательно, по предыдущей теореме, на p не делится и весь знаменатель. Но тогда на основании теоремы 9 он взаимно прост с p. Поэтому на знаменатель должен делиться второй сомножитель числителя. Обозначая частное от этого деления через q, мы имеем $C^k_p = pq$, и требуемое доказано.

Следствие. Если p — простое и $0 < k < p$, то число

$$C_p^k = \frac{1 \cdot 2 \ldots (p-1)\,p}{1 \cdot 2 \ldots (k-1)\,k \cdot 1 \cdot 2 \ldots (p-k-1)\,(p-k)}$$

делится на p.

Теорема 14 (основная теорема арифметики). *Всякое целое положительное число, кроме единицы, может быть представлено в виде произведения простых сомножителей и притом единственным способом* (произведения, отличающиеся только порядком сомножителей, различными не считаются).

Доказательство.

Сначала докажем возможность разложения любого числа, отличного от единицы, на простые множители. Предположим, что все числа, меньшие N, могут быть так разложены. Если число N простое, то оно автоматически разлагается в произведение простых (именно, в произведение, состоящее только из одного сомножителя — самого числа N), и теорема доказана. Пусть теперь N составное, N_1 — некоторый делитель N, отличный как от N, так и от единицы, и N_2 — частное от деления N на N_1. Тогда $N = N_1N_2$, причем, как легко проверить, $1<N_2<N$. Так как N_1 и N_2 меньше N, то, по предположению, они разлагаются в произведения простых множителей. Пусть $N_1=p_1p_2 \dots p_k$ и $N_2=q_1q_2 \dots q_l$ — эти разложения. Тогда $p_1p_2 \dots p_k\, q_1q_2 \dots q_l$ является искомым разложением числа N. Возможность разложения, таким образом, доказана.

Переходим к доказательству единственности разложения. Пусть нам даны два разложения числа N на простые множители: $p_1p_2 \dots p_k$ и $q_1q_2 \dots q_l$. Очевидно,

$$p_1p_2 \dots p_k = q_1q_2 \dots q_l \qquad (**)$$

Так как $q_1q_2 \dots q_l$ делится на p_1, то по предыдущей теореме хотя бы одно из чисел $q_1, q_2, \dots, q_t$ делится на p_1. Пусть $q_1 \vdots p_1$ (то, что мы считаем именно первый сомножитель в (**) справа делящимся на p_1, никакого дополнительного предположения не означает, так как мы вправе переставлять сомножители местами и обозначить через q_1 именно тот из них, который делится на p_1). Так как число q_1 простое, это возможно лишь при $p_1 = q_1$. Сокращая равенство (**) на p_1 получаем

$$p_1p_2 \dots p_k = q_1q_2 \dots q_l. \qquad (***)$$

Аналогично предыдущему убеждаемся в том, что некоторое из чисел $q_2, q_3, \dots, q_l$ (например, q_2) делится на p_2, и потому $p_2 = q_2$. Сокращая равенство (***) на p_2, мы уменьшаем число сомножителей в его частях еще на единицу. Такой процесс сокращения, очевидно, можно продолжать до тех пор, пока мы не сократим одно из произведений полностью. Пусть первым сократится произведение, стоящее в (**) слева. Произведение, стоящее в (**) справа, тоже должно при этом сократиться нацело, так как в противном случае мы получили бы равенство вида

$$1 = q_{k+1} \dots q_l$$

которое невозможно, так как единица не делится ни на какое простое число. При этом мы получаем также, что

$$p_1 = q_1, \quad p_2 = q_2, \dots, p_k = q_k.$$

Теорема полностью доказана,

Основная теорема арифметики указывает на принципиальную возможность разложения любого числа на простые сомножители. Однако практическое осуществление такого разложения встречает большие трудности. Разложение больших чисел на множители или установление их простоты осуществляется на основе применения электронных вычислительных машин. Так, было обнаружено, что число $2^{19937} — 1$ является простым.

Пусть некоторое число a разложено в произведение простых сомножителей. Объединяя равные сомножители, мы получим формулу вида

$$a = p_1^{\alpha_1} p_2^{\alpha_2} \dots p_r^{\alpha_r}, \qquad (7)$$

где $p_1, p_2, \dots, p_r$ — различные простые числа, а $\alpha_1, \alpha_1, \dots, \alpha_r$ — некоторые целые положительные числа. Произведение, стоящее в правой части формулы (7),называется *каноническим разложением* числа α.

Теорема 15. *Для того чтобы числа a и b были взаимно простыми, необходимо и достаточно, чтобы ни один из простых сомножителей, входящих в каноническое разложение числа a, не входил в каноническое разложение числа b.*

Доказательство.

Пусть $p_1^{\alpha_1} p_2^{\alpha_2} \dots p_k^{\alpha_k}$ и $q_1^{\beta_1} q_2^{\beta_2} \dots q_t^{\beta_t}$ — соответственно канонические разложения чисел a и b, а d — некоторый общий делитель этих чисел. Если $d \neq 1$, то d делится на некоторое простое число p. Тогда по теореме 3 $a \vdots p$ и $b \vdots p$, так что p находится как среди чисел $p_1, p_2, \dots, p_k$, так и среди чисел $q_1, q_2, \dots, q_l$. Поэтому среди простых чисел, входящих в каноническое разложение α, существует хотя бы одно, входящее в каноническое разложение b.

Наоборот, если a и b взаимно просты и p входит в каноническое разложение a, то b не делится на p, так что p не может входить в каноническое разло жение b,

Теорема 16. *Пусть* (7)*—каноническое разложение числа a. Тогда для делимости* $b \vdots a$ *необходимо и достаточно, чтобы*

$$b \vdots p_1^{\alpha_1}, \quad b \vdots p_2^{\alpha_2}, \dots, b \vdots p_r^{\alpha_r}.$$

Доказательство.

Необходимость. Так как $a \vdots p_i^{\alpha_i}$ (i=1, 2, ..., k), мы из $b \vdots a$ получаем требуемое простой ссылкой на теорему 2.

Достаточность доказывается по индукции. Делимость $b \vdots p_1^{\alpha_1}$ мы имеем в числе условий. Предположим, что нами уже установлено, что

$$b \vdots p_1^{\alpha_1} \dots p_l^{\alpha_l} \quad (1 \leqq l < k).$$

Кроме того, в нашем распоряжении имеется делимость $b \vdots p_{l+1}^{\alpha_{l+1}}$. Так как числа $p_1^{\alpha_1} \dots p_l^{\alpha_l}$ и $p_{l+1}^{\alpha_{l+1}}$ по предыдущей теореме взаимно просты, мы можем применить следствие теоремы 11, которое дает нам

$$b \vdots p_1^{\alpha_1} \dots p_l^{\alpha_l} p_{l+1}^{\alpha_{l+1}}.$$

Этим индуктивный переход обоснован.

Из теорем 15 и 16 вытекает, что делимость на произведение нескольких взаимно простых чисел равносильна одновременной делимости на каждое из них.

Теорема 17. *Пусть*

$$a = p_1^{\alpha_1} p_2^{\alpha_2} \dots p_r^{\alpha_r}$$

— каноническое разложение числа a. Тогда для делимости a⋮b необходимо и достаточно, чтобы каноническое разложение b имело вид

$$b = p_1^{\beta_1} p_2^{\beta_2} \dots p_r^{\beta_r},$$

где

$$0 \leqq \beta_1 \leqq \alpha_1,$$
$$0 \leqq \beta_2 \leqq \alpha_2,$$
$$\dots\dots\dots$$
$$0 \leqq \beta_r \leqq \alpha_r.$$

Доказательство.

Необходимость. Пусть $a \vdots b$. Из теоремы 13 следует, что каждый простой делитель b является простым делителем a. Таким образом, b имеет вид

$$p_1^{\beta_1} p_2^{\beta_2} \dots p_k^{\beta_k},$$

где $0 \leqq \beta_1$, $0 \leqq \beta_2$, …, $0 \leqq \beta_k$. Предположим, что $\beta_1 > \alpha_1$. Так как

$$\frac{a}{b} = \frac{p_1^{\alpha_1} p_2^{\alpha_2} \dots p_k^{\alpha_k}}{p_1^{\beta_1} p_2^{\beta_2} \dots p_k^{\beta_k}} = \frac{p_2^{\alpha_2} \dots p_k^{\alpha_k}}{p_1^{\beta_1 - \alpha_1} p_2^{\beta_2} \dots p_k^{\beta_k}}$$

—целое число, числитель последней дроби должен делиться на знаменатель и тем более на число $p_1^{\beta_1 - \alpha_1}$. Но тогда по теореме 13 на

p_1 должно делиться хотя бы одно из чисел p_2, ..., p_k, чего не может быть. Значит, $\beta_1 \leqq \alpha_1$. Так как нумерация простых делителей a для нас безразлична, мы тем самым доказали, что и $\beta_2 \leqq \alpha_2, \ldots, \beta_k \leqq \alpha_k$. Необходимость доказана.

Для доказательства достаточности заметим, что если b имеет указанный вид, то

$$a = bp_1^{\alpha_1-\beta_1}p_2^{\alpha_2-\beta_2} \ldots p_k^{\alpha_k-\beta_k}.$$

Весьма важным для целей данного микромодуля оказывается следующий факт.

Теорема 18. *Пусть m и t — натуральные числа. Тогда m можно представить в виде такого произведения $m = m_1m_2$, что $(m_1,t)= 1$ и найдется такое k, для которого $t^k \vdots m_2$.*

Доказательство.

Напишем канонические разложения чисел m и t:

$$m = p_1^{\alpha_1} \ldots p_n^{\alpha_n}, \quad t = q_1^{\beta_1} \ldots q_l^{\beta_l}.$$

Отберем среди простых p_1, ..., p_n те, которые делят t, т. е. содержатся среди q_1, ..., q_l. Пусть для определенности это будут p_1, ..., p_r, равные соответственно числам q_1,..., q_r. Положим тогда

$$m_2 = p_1^{\alpha_1} \ldots p_r^{\alpha_r} \quad \text{и} \quad m_1 = p_{r+1}^{\alpha_{r+1}} \ldots p_n^{\alpha_n}.$$

Согласно теореме 15 будет $(m_1,t)=1$. Кроме того, возьмем натуральное число k, которое было бы не меньше каждого из отношений

$$\frac{\alpha_1}{\beta_1}, \ \ldots, \ \frac{\alpha_r}{\beta_r}.$$

Это значит, что $k\beta_i \geqq \alpha_i$ для $i= 1, ..., r$, откуда согласно теореме 17 $t^k \mathrel{\vdots} m_2$.

Этот факт имеет свои далеко идущие алгебраические аналогии, но мы их затрагивать не будем.

4.5. Делимость сумм и произведений

1. Во многих случаях при делении с остатком интересно найти именно остаток от деления числа a на число b, а величина неполного частного от деления не играет роли.

Пусть, например, мы хотим узнать, какой день недели был 1 января 2020 г. Легко справиться по календарю, что 1 января 2000 г.— вторник. Двадцать лет, разделяющие эти даты, состоят из 20·365 + 4 (последнее слагаемое — число високосных лет за это время), т. е. из

7305 дней. Эти дни составляют 1043 целых недель и еще 4 дня. По прошествии 1043 целых недель снова наступит вторник, так что еще через 4 дня, 1 января 2020 г., будет суббота. Очевидно, для решения поставленной нами сейчас перед собой задачи совершенно неважно знать, сколько именно целых недель прошло за 20 лет, а интересно только число дней, прошедших сверх этих недель.

С задачами такого рода приходится иногда сталкиваться историкам, особенно востоковедам, при сопоставлении дат, указанных по разным календарям,

Казалось бы, для нахождения остатка от деления одного числа на другое проще всего произвести деление с остатком непосредственно. Однако практически выполнить такое деление нередко представляется весьма затруднительным, особенно если подлежащее исследованию делимое задано в виде некоторого сложного выражения, вроде, скажем, $2^{1000} + 3^{1000}$. Вместе с тем львиная доля этой работы будет потрачена на нахождение неполного частного, которое нам само по себе не нужно. Необходимо поэтому попытаться выработать способ нахождения остатка непосредственно, минуя вычисление неполного частного.

Продемонстрируем один из таких приемов на только что решавшейся нами задаче о дате 1 января 2020 г. Мы можем рассуждать следующим образом. Каждый простой (невисокосный) год состоит из 365 дней, что составляет 52 полные недели и еще один день. Високосный же год составляет столько же недель и два дня. Значит, весь срок от 1 января 2000 г. до 1 января 2020 г. состоит из некоторого (совершенно неважно, какого) числа полных недель плюс число дней, равное числу содержащихся в этом сроке лет, причем каждый високосный год считается за два. Это число дней равно 20 + 5 = 25. Исключив из него 3 полных недели, получаем 4 дня, которые и следует отсчитывать от нашего вторника. Оказывается, такая «замена года днем» есть проявление весьма общего приема, изучением которого мы сейчас и займемся.

2. Другой пример, когда целью деления с остатком является получение именно остатка, а неполное частное рассматривается лишь как исходный материал для дальнейших операций, доставляет нам запись чисел в той или иной *позиционной системе счисления.* Напомним, что число A называется записанным в (позиционной) системе счисления с основанием t, или, короче, в t-ичной системе счисления (где t — целое положительное число, большее единицы), если оно представлено в виде

$$A = a_n t^n + a_{n-1} t^{n-1} + \ldots + a_1 t + a_0,$$

где

$$0 \leqq a_i < t \quad \text{при} \quad i = 0, 1, \ldots, n \tag{1}$$

числа $a_0, a_1, \ldots, a_n$ называются *t-ичными* цифрами числа A).

При t =10 мы получаем *десятичную* систему счисления. Запись числа в этой системе настолько привычна для нас, что говоря о числе, мы обычно только в этой форме его себе и представляем. В действительности, однако, если соображения привычности перестают играть роль, как это, например, имеет место при фиксации чисел в электронных вычислительных машинах, более удобными могут оказаться и другие системы счисления (двоичная, восьмеричная и т. д.).

Так как мы в этой работе не будем рассматривать непозиционных систем счисления (например, записей чисел «римскими» цифрами), мы далее указание на их позиционность будем, как правило, опускать.

Ясно, что из (1) следует

$$A = (a_n t^{n-1} + a_{n-1} t^{n-2} + \ldots + a_1) t + a_0,$$

т. е. последняя t-ичная цифра a_0 числа A является остатком от деления A на t с остатком. Неполное частное от такого деления стоит здесь в скобках. Разделив это неполное частное на t с остатком, мы получим

$$(a_n t^{n-2} + a_{n-1} t^{n-3} + \ldots + a_2) t + a_1.$$

Остатком оказывается предпоследняя t-ичная цифра числа A. Продолжая этот процесс повторного деления с остатком на t, мы будем последовательно получать все t-ичные цифры числа A, считая справа налево (т. е., от низших разрядов к высшим). Очевидно (а точнее — в силу полной упорядоченности множества натуральных чисел по величине), этот процесс последовательного деления с остатком должен рано или поздно оборваться. В результате мы получим все t-ичные цифры числа A, т. е. его запись в t-ичной системе счисления.

Так, в частности, осуществляется перевод чисел из одной системы счисления в другую. Например,

$$10\,000 = 6 \cdot 1\,666 + 4$$
$$1\,666 = 6 \cdot 277 + 4$$
$$277 = 6 \cdot 46 + 1$$
$$46 = 6 \cdot 7 + 4$$
$$7 = 6 \cdot 1 + 1$$
$$1 = 6 \cdot 0 + 1$$

Поэтому 10 000 в шестеричной системе счисления записывается как 114 144.

3. Определение. Назовем числа *a* и *b* *равноостаточными* при делении на *m,* если остатки от деления *a* и *b* на *m* равны. Установим несколько свойств равноостаточных чисел.

Теорема 19. *Для того чтобы числа a и b были равноостаточными при делении на m, необходимо и достаточно, чтобы (a- b)⋮m.*

Доказательство.

Необходимость. Пусть

$$a = mq_1 + r_1 \quad (0 \leqq r_1 < m), \qquad (*)$$
$$b = mq_2 + r_2 \quad (0 \leqq r_2 < m), \qquad (**)$$

Ввиду равноостаточности *a* и *b* должно быть $r_1 = r_2$. Значит,

$$a - b = m(q_1 - q_2),$$

т. е. $(a - b) \vdots m$.

Достаточность. Пусть *(a — b)⋮m.* Разделив *a* и *b* на *m* с остатком, мы получим (*) и (*). При этом

$$a - b = m(q_1 - q_2) + r_1 - r_2,$$

т. е.

$$(a - b) - m(q_1 - q_2) = r_1 - r_2.$$

По теореме 6 $(r_1 - r_2) \vdots m$. Но $|r_1 - r_2| < m$. Значит, по теореме $r_1 - r_2 = 0$ или $r_1 = r_2$, а это и требовалось.

Следствие. Если числа *a* и *b* равноостаточны при делении на *m,* и *m⋮d,* то *a* и *b* равноостаточны при делении на *d.*

Теорема 20. *Если при делении на m числа a_1, a_2, ..., a_n соответственно равноостаточны числам b_1, b_2, ..., b_n, то равноостаточными будут суммы $a_1+a_2+ \ldots+a_n$ и $b_1+ b_2+ \ldots+b_n$, а также произведения $a_1a_2 \ldots a_n$ и $b_1b_2 \ldots b_n$.*

Доказательство.

Из условия на основании теоремы 16 мы имеем

$$\left.\begin{aligned} a_1 &= b_1 + mq_1, \\ a_2 &= b_2 + mq_2, \\ &\cdots\cdots\cdots \\ a_n &= b_n + mq_n. \end{aligned}\right\} \qquad (***)$$

Сложив почленно эти равенства, мы после простых преобразований получаем

$$(a_1 + a_2 + \ldots + a_n) - (b_1 + b_2 + \ldots + b_n) = \\ = m(q_1 + q_2 + \ldots + q_n),$$

что по теореме 19 и означает равноостаточность сумм. Для доказательства равноостаточности произведений отметим следующее тождество:

$$(k + bm)(p + qm) = kp + (pq + lp + lqm)m.$$

Из него следует, что произведение двух чисел вида $a + bm$ снова является числом того же вида. Поэтому, рассуждая по индукции, мы убеждаемся в том, что произведение любого количества чисел вида $a + bm$ есть число этого же вида.

Перемножив теперь почленно все равенства (***) и применив к правой части только что проведенные рассуждения, мы получаем

$$a_1 a_2 \ldots a_n = b_1 b_2 \ldots b_n + mt,$$

где t — некоторое целое число. Равноостаточность произведений, таким образом, доказана.

Следствие. Если при делении на m числа a и b равноостаточны, то такими же являются и степени a^n и b^n при любом натуральном n.

Теорема 20 и ее следствие дают уже довольно богатые возможности для нахождения остатков от деления.

Приведем несколько примеров.

Пример 1. Найти остаток от деления на 3 числа

$$A=13^{16}\text{-}2^{23}\cdot 5^{15}.$$

Очевидно, при делении на 3 число 13 равноостаточно с 1, 2 равноостаточно с —1, а 5 тоже с —1. Значит, на основании доказанного число A при делении на 3 равноостаточно с числом

$$1^{16} — (—1)^{25}(—1)^{15} == 1 — 1 == 0;$$

т. е. искомый остаток равен нулю, а A делится на 3.

Пример 2. Найти остаток от деления того же числа A на 37.

Представим для этого A в следующем виде:

$$A = (13^2)^8\text{-}(2^5)^5\text{-}(5^3)^5.$$

Так как 13^2= 169 при делении на 37 равноостаточно с —16, 2^5 = 32 равноостаточно с —5, а 5^3=125 — с +14, то все число *A* равноостаточно с

$$(-16)^8\text{-}(-5)^5(+14)^5$$

или, что то же самое, с

$$(16^2)^4 + 70^5.$$

Но 16^2, т. е. 256, равноостаточно с —3, а 70 — с —4. Значит, *A* равноостаточно с

$$(—3)^4 + (—4)^5$$

или, что то же самое, с

$$81 — (2^5)^2,$$

а потому с

$$81\text{-}(-5)^2\text{= }81\text{-}25 = 56.$$

Наконец, 56 при делении на 37 равноостаточно с 19, которое неотрицательно и меньше 37 и потому является искомым остатком.

4. Равноостаточные при делении на *m* числа *a* и *b* называются также *сравнимыми по модулю m.* Это обозначается так:

$$a = b \pmod m,$$

а сама эта формула называется *сравнением.*

Сравнимость двух чисел по некоторому фиксированному модулю *m* или, что то же самое, их равноостаточность при делении на *m,* также является некоторым отношением, связывающим между собой целые числа.

Отметим несколько свойств отношения сравнимости по модулю.

1°. Рефлексивность: $a \equiv a \pmod m$.

Действительно, $a — a = 0 \vdots m$.

2°. Симметричность: если $a \equiv b \pmod m$, то $b \equiv a \pmod m$.

В самом деле, если $(a — b) \vdots m$, то (хотя бы по теореме 5) и $(b — a) \vdots m$

3°. Транзитивность: если $a \equiv b \pmod m$ и $b \equiv c \pmod m$, то $a \equiv c \pmod m$.

Для доказательства достаточно заметить, что из $(a—b) \vdots m$ и $(b— a) \vdots m$ по теореме 6 следует, что $(a — c) \vdots m$.

Если некоторое отношение (обозначим его через ~) обладает свойствами рефлексивности, симметричности и транзитивности, то оно называется отношением *эквивалентности* (или *эквивалентным отношением).* Простейшим примером отношения эквивалентности на множестве чисел является отношение равенства.

Так как отношение сравнимости по модулю *m* — отношение эквивалентности, оно также разбивает множество целых чисел на классы. Эти классы называют *классами вычетов* по модулю *m*.

4° Число классов вычетов по модулю m равно m.

В самом деле, два числа a и b принадлежат одному классу вычетов по модулю m тогда и только тогда, когда они при делении на m дают один и тот же остаток. Но остаток при делении на m может принимать ровно m значений: 0, 1, 2, ... m— 1. Следовательно, и число классов равно m

Отметим одно обстоятельство, являющееся уточнением следствия теоремы 19

Для того чтобы каждый класс вычетов по модулю m_1 содержался в некотором классе вычетов по модулю m_2, необходимо и достаточно, чтобы $m_1 \mid m_2$.

Действительно, рассмотрим класс вычетов K_1 по модулю m_1, содержащий число 0. Очевидно, класс K_1 состоит из всех чисел, дающих при делении на m_1 в остатке 0, т е. делящихся на m_1. В частности, он содержит число m_1. Класс вычетов по модулю m_2, содержащий K_1, также содержит 0 и потому состоит из всех чисел, делящихся на m_2. Так как в него входит число m_1, должно быть $m_1 \mid m_2$. Этим доказана необходимость, достаточность же очевидна.

Таким образом, отношение делимости можно определить через соотношения между классами вычетов. Этот прием позволяет определять делимость для объектов гораздо более общей и сложной природы, чем натуральные числа. Последовательное развитие этих идей приводит к теории групп.

Продолжим перечисление свойств сравнимости чисел. Из теоремы 20 немедленно следуют:

5°. Если $a \equiv b \pmod m$ и $c \equiv d \pmod m$, то

$$a + c \equiv b + d \pmod m$$

Следствие. Если $a \equiv b \pmod m$, то

$$a + r \equiv b + r \pmod m$$

для любого целого r.

6°. Если $a \equiv b \pmod m$ и $c \equiv d \pmod m$, то

$$ac \equiv bd \pmod m.$$

Свойства 5° и 6° показывают, что сравнения подобно равенствам можно почленно складывать и перемножать.

4.6. Признаки равноостаточности и признаки делимости

1. Весьма общий способ нахождения остатка от деления произвольного, но фиксированного натурального числа a на данное

натуральное число m заключается в следующем. Будем строить последовательность натуральных чисел

$$a = A_0, A_1, A_2, \ldots, \qquad (1)$$

равноостаточных при делении на m. Способ построения этой последовательности выберем такой, чтобы после всякого ее члена, большего или равного m, следовал еще хотя бы один член. Тогда, очевидно, всякий член последовательности (1), меньший чем m (если, конечно, такой существует), будет равен остатку от деления a на m. Таким членом может быть, например, последний член последовательности (опять-таки, если такой имеется).

Одним из простейших примеров последовательности (1) может служить последовательность (3) из п. 11 п. 4.3:

$$a, a - m, a - 2m, \ldots$$

В сущности, к построению последовательностей такого типа сводятся задачи нахождения остатков в примерах 1 и 2 из п. 4.6.

Всякий способ построения последовательности (1), обладающей последним членом, назовем *признаком равноостаточности при делении на m.*

Из только что приведенного примера следует, что одним из признаков равноостаточности при делении на m является процесс последовательного вычитаний числа m до получения первого числа, меньшего m.

2. Очевидно, для уверенности в безотказности работы признака равноостаточности при делении на m необходимо, чтобы он удовлетворял следующим трем требованиям:

1) Признак равноостаточности должен быть применим к любому натуральному числу a. Иными словами, каково бы ни было число a, конструируемая по нему последовательность (1) действительно должна обладать указанным выше свойством: после каждого ее члена, не меньшего чем m, должен следовать еще хотя бы один член. Это свойство признака называется его *массовостью*

2) Признак равноостаточности должен быть точно *определенным,* т. е. число a должно вполне определять все члены последовательности (1), не оставляя места какой-либо произвольности.

3) Наконец, мы должны иметь гарантию того, что в последовательности (1) хотя бы один член будет меньше чем m. Это требование будет выполнено, если строить последовательность (1) так, чтобы она обязательно имела лишь конечное число членов т. е. чтобы процесс ее построения не мог продолжаться неопределенно долго, а рано или поздно заканчивался бы появлением

остатка от деления a на m. Сформулированное свойство признака равноостаточности называется его *результативностью.*

3. Процессы, обладающие свойствами массовости, определенности и результативности, называются *алгоритмами*.

Разумеется, только что приведенная характеризация алгоритма как процесса, обладающего тремя перечисленными свойствами, не является его точным определением. Такое определение, хотя и выработано математикой, но сравнительно сложно и не может быть здесь сформулировано. Однако перечисленные предъявляемые к алгоритмам требования довольно полно отражают те условия, которым должны удовлетворять называемые алгоритмами математические процессы. Роль алгоритмов определяется тем, что они являются единообразными способами решения целого ряда однотипных задач. Так, каждый признак равноостаточности позволяет находить остатки от деления варьируемого числа a на некоторое фиксированное m.

Говоря несколько вольно, к алгоритмам сводятся все те математические задачи, решение которых можно автоматизировать. Поэтому не случайно развитие теории алгоритмов исторически совпало с появлением и распространением электронных вычислительных машин.

К алгоритмам сводятся не только вычислительные задачи в узком смысле этого слова, т. е. такие задачи, в которых по более или менее сложным правилам можно на основе исходных данных получить численный ответ. Можно также ставить вопрос о поисках алгоритма, позволяющего решать любую задачу из некоторой (разумеется, строго очерченной) области математики. Этот алгоритм должен уметь перерабатывать формулировки теорем в их доказательства. Такие алгоритмы разработаны для довольно широких областей математики. Вместе с тем для некоторых ее областей (например, для любой области, охватывающей всю арифметику) такие алгоритмы разработать весьма сложно

4. Уточним применительно к признакам равноостаточности содержание, а также последствия от соблюдения трех предъявляемых к алгоритмам требований.

Из массовости признака равноостаточности вытекает, что он должен перерабатывать различные числа, и результаты этой переработки также должны быть, вообще говоря, различными (ибо при делении на какое бы то ни было $m > 1$ не все числа равноостаточиы друг другу). Значит, необходимой составной частью этого процесса должно быть различение чисел (по их величине).

Свойство определенности признака равноостаточности означает, что уже выписанные числа $A_0, A_1, ..., A_n$ последовательности (1) должны быть настолько «опознаваемыми», чтобы на их основании можно было написать следующее число последовательности, A_{n+1}.

Наконец, свойство результативности влечет, кроме всего прочего, еще н необходимость неограниченных возможностей сравнивать (по величине) получаемое на каждом шаге нашего процесса число A_k с делителем m.

Таким образом, соблюдение каждого из трех требований алгоритмичности для признака равноостаточности упирается, прежде всего, в необходимость уметь сравнивать в произвольных парах числа по их величине и указывать, если они различны, какое из них больше, а какое — меньше.

5. Только что упомянутая «необходимость уметь», очевидно, также имеет алгоритмическую природу: сравнению по величине должны подлежать любые два натуральных числа (массовость), результатом сравнения может быть не более, чем один ответ: больше, меньше или равно (определенность), и этот ответ должен всегда достигаться (результативность). Значит, мы получаем основание говорить об алгоритмах сравнения двух чисел по величине. Построение такого алгоритма не является столь уж самоочевидным делом, как это могло бы показаться на первый взгляд. Например, вопрос о том, одинаковы или различны числа

$$2^{20} - 3 \cdot 5^2 \cdot 11 \cdot 31 \cdot 41 \quad \text{и} \quad 3^{10} - 2 \cdot 3 \cdot 13 \cdot 757, \qquad (2)$$

и если различны, то какое из них больше, требует для своего решения известных усилий, хотя в действительности первое из этих чисел есть всего-навсего 1, а второе 3.

Ясно, что сравнение по величине чисел из (2) затрудняется формой их записи. Следовательно, для построения признаков равноостаточности весьма важно иметь дело с представлением чисел в такой форме, которая обеспечивала бы возможность их сравнения по величине. Такие формы записи существуют.

Например, ими являются записи чисел в тех или иных (позиционных) системах счисления (см. п. 4.6 п. 2). Алгоритм сравнения двух чисел, записанных и одной и той же системе счисления, состоит в следующем:

I) Сначала в каждом из чисел зачеркиваются цифры по одной (начиная, скажем, справа); если посло того, как одно из чисел окажется зачеркнутым полностью, в другом еще останутся цифры, то второе число будет больше первого; если же запасы цифр в обоих

числах будут исчерпаны одновременно, то для сравнения чисел выполняется следующая процедура:

2) Записи сравниваемых чисел восстанавливаются, сравниваются их первые (слева) цифры. При этом большей цифре будет соответствовать большее число; если первые цифры оказываются одинаковыми, то сравниваются вторые цифры и т. д. до первого различия цифр. При этом опять-таки большая цифра будет указывать на большее число. Если все соответственные цифры чисел окажутся одинаковыми, то числа будут равны.

При проведении второй из указанных процедур предполагается, что сравнение по величине однозначных, т. е. меньших чем основание системы счисления, чисел мы производить умеем. Это значит, что в каждой системе счисления исходные значки-цифры заранее задаются в некотором фиксированном порядке; например, в общепринятой десятичной нумерации значок «2» предшествует значку «3» в том смысле, что значок «2» описывает меньшее количество, чем значок «3».

С точки зрения такого алгорифмического сравнения чисел по величине все системы счисления теоретически равноценны. Сравнение же в этом смысле систем счисления по их практическому удобству может служить примером неалгоритмической постановки вопроса (не выполняется условие определенности), и мы на нем останавливаться не будем. Обратим только внимание на то, что в этом вопросе сила привычки к десятичной системе счисления никаких особых преимуществ этой системе не дает.

6. Кроме алгоритмов сравнения чисел, записанных в одной и той же системе счисления, существуют и алгориимы выполнения арифметических действий над ними. Ими являются общеизвестные (и, очевидно, зависящие лишь несущественным образом от основания системы счисления) способы сложения, вычитания и умножения чисел «столбиком» и их деления «углом». Ясно, что в последнем случае было бы, пожалуй, уместнее говорить не просто о делении, а о делении с остатком.

В случае выполнения действий навыки в обращении с десятичной системой счисления приносят существенное облегчение. Например, выполнение в пятеричной системе счисления действия

$$\begin{array}{r|l} 13\,110 & 224 \\ \hline 12\,32 & 31 \\ \hline 240 & \\ 224 & \\ \hline 11 & \end{array}$$

требует известных умственных усилий.

Из алгоритмичности деления с остатком согласно сказанному в п. 2 п. 4.6 вытекает и алгоритмичность перевода записей чисел из одной системы счисления в другую. Следовательно, можно говорить также об алгоритмах сравнения чисел и действий над ними, если они записаны в различных системах счисления. Как дальнейшее следствие, отсюда получается, что алгоритмами являются всевозможные вычисления по арифметическим формулам, в которые вместо букв можно подставлять те или иные числа.

Обратим, наконец, внимание на то, что мы не говорим здесь об алгоритме самого процесса записи произвольно заданных натуральных чисел в той или иной системе счисления, ибо мало ли каким может оказаться это исходное задание.

7. В качестве иллюстративного примера рассмотрим следующее построение. Для каждого натурального числа n составим последовательность $a_0^{(n)}, a_1^{(n)}, a_2^{(n)}, \ldots$ чисел (цифр), являющихся цифрами бесконечного десятичного разложения числа $\sqrt{n}$ (если число n не является точным квадратом, то эта последовательность, очевидно, оказывается непериодической), и пусть $r_1^{(n)}, r_2^{(n)}, \ldots$ — номера всех тех цифр, которые равны нулю: $a_{r_i^{(n)}}^{(n)} = 0\ (i = 1, 2, \ldots)$. Если теперь число равных нулю цифр конечно (пусть последнее из них имеет номер $r_k^{(n)}$, так что $a_i^{(n)} > 0$ при $i > r_k^{(n)}$), то положим

$$f(n) = 10^{r_1^{(n)}} + 10^{r_2^{(n)}} + \ldots + 10^{r_k^{(n)}} + 1,$$

а если их число бесконечно, то положим, скажем, $f(n) = 0$. Каждое из чисел $f(n)$ является натуральным. Однако едва ли можно говорить об алгоритме, который перерабатывал бы число n в запись числа $f(n)$ в десятичной системе счисления,

Разумеется, вся неалгоритмичность этой конструкции состоит в требовании распознавать, будет ли в десятичном разложении $\sqrt{n}$ конечное или бесконечное число нулей. Между прочим, в известном

смысле (в каком именно — мы не станем здесь выяснять) естественно верить, что $f(n) = 0$ для любого натурального n

8. Одним из наиболее важных в математике алгоритмов является так называемый *алгоритм Евклида,* который состоит в следующем.

Пусть a и b — два натуральных числа, причем $b > 0$. Разделим a на b с остатком $a = bq_0 + r_1$, где $0 \leq r_1 < b$. Ести $r_1 \neq 0$, то мы имеем возможность разделить с остатком b на r_1: $b = r_1q_1 + r_2$, причем $0 \leq r_2 < r_1$. Продолжая эти последовательные деления с остатком на остаток от предыдущего деления, мы получим дальнейшие равенства:

$$r_1 = r_2q_3 + r_3,\ r_2 = r_3q_4 + r_4 \text{ и т. д.}$$

Покажем, что описаший процесс действительно является алгоритмом, т. е. обладает свойствами определенности, массовости и результативности.

Заметим, что рассматриваемый нами процесс состоит в последовательном выполнении действия деления с остатком.

Поэтому определенность и массовость этого процесса являются следствием неограниченной выполнимости и однозначности действия лечения с остатком. Результативность нашего процесса устанавливается также довольно просто. Число b и остатки от делений, составляющих наш процесс, образуют, очевидно, убывающую последовательность неотрицательных чисел:

$$b, r_1, r_2, \ldots \qquad (3)$$

Но число всех неотрицательных и не превосходящих b чисел равно $b + 1$. Поэтому и последовательность (3) не может насчитывать более чем b членов, так что наш процесс может состоять не более чем из b делений с остатком (на самом деле число этих делений не может превосходить числа $5\log b$. Это следует из рассмотрения чисел Фибоначчи). Таким образом, рассматриваемый процесс действительно является алгоритмом и вполне оправдывает свое название.

Выясним условия окончания процесса. Очевидно, последнее деление должно быть таким, чтобы дальнейшее деление на его остаток было уже невозможно. Но это может быть лишь в том случае, когда этот последний остаток равен нулю, т е. когда последнее деление совершилось нацело.

9. Применение алгоритмов (их, так сказать, «работа») может оказаться достаточно громоздким. В качестве примера рассмотрим процесс получения по числу n его канонического разложения (иными словами, алгоритм, перерабатывающий натуральное число в его каноническое разложение). Для оттенения алгоритмической сущности этого процесса включим его, как этап, в процесс последовательного нахождения канонических разложений

одного за другим всех натуральных чисел. Это дает нам основание вести рассуждения «по индукции» (см. п. 7 п. 4.5), Предположим, что для всех чисел, меньших n, канонические разложения уже выписаны. Из этого списка можно (вполне алгоритмично) усмотреть, какие из чисел, меньших n, являются простыми. Перечислив их все по возрастанию, будем делить n на каждое из них. Если n разделится на некоторое p, то будет $n = n_1 p$ и $n_1 < n$, а каноническое разложение n_1 в нашем перечне по предположенному уже имеется (ввиду результата задачи 13 нам достаточно произвести деления лишь на те p, которые меньше чем $\sqrt{n}$) и каноническое разложение получится из канонического разложения n_1 путем увеличения в нем показателя p на единицу.

10. Вернемся, однако, к признакам равноостаточностни. Алгоритмическое построение последовательности (1) может быть осуществлено весьма разнообразными путями. Наиболее естественный из них состоит в следующем.

Попробуем найти функцию $f(x)$\ подчиненную следующим условиям:

а) Значение $f(x)$ при $x \geq m$ есть натуральное число;

б) Значение $f(x)$ при $x < m$ не определено (т. е, не имеет смысла);

(Нет ничего удивительного в том, что та или иная функция теряет смысл при некоторых значениях аргумента. Например, не имеет смысла значение функции $\dfrac{1}{x(x-1)}$ при $x = 0$ или при $x=1$).

в) Если $x \geq m$, то $f(x) < x$;

г) Если $x \geq m$, то числа x и $f(x)$ равноостаточны при делении на m.

Такие функции существуют. Примером является функция $f_0(x)$:

$$f_0(x) = \begin{cases} x - m, & \text{если} \quad x \geqq m, \\ \text{не определена}, & \text{если} \quad x < m. \end{cases}$$

Именно эта функция и осуществляет построение последовательности (3) в п. 4.5.

Каждой функции $f(x)$, удовлетворяющей условиям а)—г), отвечает некоторый способ построения последовательности (1), т. е. некоторый признак равноостаточности при делении на m.

В самом деле, возьмем произвольное натуральное число a и будем строить последовательность чисел

$$A_0, A_1, A_2, \ldots, \tag{4}$$

где

$$A_0 = a \quad \text{и} \quad A_{k+1} = f(A_k) \quad \text{при} \quad k = 0, 1, \ldots \tag{5}$$

Если $A_k \geq m$, то значение функции $f(A_k)$ определено, и потому за A_k следует хотя бы один член. Если же $A_k < m$, то $f(A_k)$ не определена, и A_k является последним членом последовательности (4).

Итак, мы действительно имеем некоторый признак равноостаточности.

11. Покажем, что найденный признак равноостаточности обладает всеми тремя свойствами алгоритма.

Условие массовости здесь соблюдается потому, что любое число дает начало некоторой последовательности (4), обладающей свойством (5).

Условие определенности соблюдается ввиду того, что для вычисления значений $f(x)$ функции f достаточно уметь сравнивать по величине числа x и m и выполнять операцию вычитания (отнимания m от x). Обе эти процедуры (если мы имеем дело с числами, записанными в некоторой системе счисления), как было выяснено, являются алгоритмами и тем самым обладают свойством определенности.

Обратимся к условию результативности. По самому своему построению функция f выбрана так, что члены последовательности (4) положительны и убывают. Поэтому в ней найдется наименьший неотрицательный член. (Номер этого члена, как нетрудно проверить, не превосходит числа a.) Если бы этот член (обозначим его через α) был больше или хотя бы равен m, то существовало бы значение $f(a)$, по-прежнему неотрицательное, но меньшее α. Значит, член α, не был бы последним среди неотрицательных членов последовательности (4). Следовательно, последний неотрицательный член (4) должен быть меньше, чем m. Но тогда значение $f(\alpha)$ не имеет смысла, и α оказывается вообще последним членом нашей последовательности. Процесс построения последовательности, таким образом, заканчивается, и последний ее член является остатком от деления α на m.

В результате мы установили, что описанный нами признак равноостаточности действительно обладает требуемыми свойствами определенности, массовости и результативности, т. е. является алгоритмом.

12. Пользуясь изложенным в п. 9 приемом построения признаков равноостаточности, найдем несколько таких признаков. В соответствии со сказанным выше будем считать, что числа, остатки от деления которых требуется найти, записаны в позиционной системе счисления с некоторым основанием t. Признак равноостаточности при делении на некоторое m перерабатывает в остаток от деления на m

фактически не само число, а его запись в соответствующей системе счисления. Поэтому признак равноостаточности при делении на данное фиксированное число m будет, вообще говоря, зависеть от основания системы счисления. Вместе с тем буквальная формулировка признака равноостаточности при делении на данное m в t-ичной системе счисления вполне может подходить для признака равноостаточности при делении на другое m' в системе счисления с другим основанием t'. Соответствующие примеры будут получаться из содержания теорем 19, 20 и 21.

Во избежание возможных недоразумений условимся в дальнейшем как делитель m, так и основание системы счисления t записывать («называть») в десятичной системе счисления. Так, говоря о признаке равноостаточности при делении на 12 в семеричной системе счисления, мы будем под 12 понимать именно число 3·4, а не число 3·3 (как это было бы, если бы число 12 рассматривалось как запись в семеричной системе счисления).

В качестве первого примера найдем признак равноостаточности при делении на 5 в десятичной системе счисления.

Пусть A — натуральное число. Представим A в виде $10a+ b$ (b — последняя цифра числа A) и положим

$$f_1(A)=\begin{cases} b, & \text{если} \quad A \geqq 10, \\ b-5, & \text{если} \quad 5 \leqq A < 10, \\ \text{не определено}, & \text{если} \quad A < 5. \end{cases}$$

Читатель сам может проверить, что так определенная функция удовлетворяет условиям а)—г) п. 10.

Таким образом, для нахождения остатка от деления некоторого числа на 5 достаточно взять последнюю цифру этого числа. Если эта цифра меньше пяти, то она и будет искомым остатком; в противном случае от нее следует отнять 5. Заметим, что применение этого признака равноостатрчности к любому числу приводит к построению последовательности типа (4), состоящей не более чем из трех членов.

Разумеется, целью всех проведенных рассуждений является не обнаружение известного всем «признака делимости» на 5, а получение его тем единообразным приемом, который был описан в п. 10.

Теорема 21. *Представим произвольное натуральное число A в виде $at^k + b$, где $0 \leq b < t^k$, и положим*

$$f(A)=\begin{cases} b, & \textit{если} \quad A \geqq t^k, \\ b-m, & \textit{если} \quad m \leqq A < t^k, \\ \textit{не определено}, & \textit{если} \quad A < m. \end{cases}$$

Чтобы для данной функции f алгоритм построения последовательности (4) *по правилу* (5) *был признаком равноостаточности при делении на m, необходимо и достаточно, чтобы было* $t^k \equiv m$.

Доказательство.

Необходимость. Если описанный алгоритм является признаком равноостаточности при делении на *m,* то числа *A* и *b* при делении на *m* должны быть равноостаточными. В частности, это будет так, если $A = t^k + b$. Но это значит, что $A - b = t^k \equiv m$.

Достаточность. В наших обозначениях $A - b = at^k$, т. е. числа *A* и *b* равноостаточны при делении на *m.* Если $t^k \equiv m$, то по следствию теоремы 17 они равноостаточны и при делении на *m.* Поэтому конструируемая алгоритмом в этом случае последовательность $A_0, A_1, \ldots$ состоит из равноостаточных при делении на *m* чисел. Следовательно, процесс построения этой последовательности является признаком равноостаточности при делении на *m.*

13. В качестве второго примера рассмотрим признак равноостаточности при делении на 3 в десятичной системе счисления.

Запись натурального числа *A* в десятичной системе счисления имеет вид

$$a_n 10^n + a_{n-1} 10^{n-1} + \ldots + a_1 \cdot 10 + a_0,$$

где

$$0 \leqq a_i < 10 \quad \text{для} \quad i = 0, 1, \ldots, n.$$

Положим

$$f_2(A) = \begin{cases} a_0 + a_1 + \ldots + a_{n-1} + a_n, & \text{если } A \geqq 10, \\ \text{остатку от деления } A \text{ на } 3, & \text{если } 3 \leqq A < 10, \\ \text{не определена}, & \text{если } A < 3. \end{cases}$$

Теорема 22. *Представим произвольное натуральное число A в виде*

$$a_n t^{kn} + a_{n-1} t^{k(n-1)} + \ldots + a_1 t^k + a_0,$$

где

$$0 \leqq a_i < t^k \quad \textit{при} \quad i = 0, 1, \ldots, n,$$

и положим

$$f(A) = \begin{cases} a_0 + a_1 + \ldots + a_{n-1} + a_n, & \textit{если } A \geqq t^k, \\ A - m, & \textit{если } m \leqq A < t^k, \\ \textit{не определено}, & \textit{если } A < m. \end{cases}$$

Тогда для того чтобы, порождаемый функцией f алгоритм построения последовательности (4) *по правилу* (5) *был признаком равноостаточности при делении на m, необходимо и достаточно, чтобы было* $(t^k - 1) \vdots m$.

Доказательство.

Необходимость. Если описанный алгоритм действительно является признаком равноостаточности при делении на m, то он, в частности, должен быть применим и к числу $A = t^k + a_0$. Здесь $f(A) = a_0 + 1$, и равноостаточность чисел A и $f(A)$ при делении на m означает $(t^k - 1) \vdots m$.

Достаточность. Пусть $A \geq t^k$. Тогда из определения функции f следует, что

$$A - f(A) = a_n(t^{kn} - 1) + a_{n-1}(t^{k(n-1)} - 1) + \ldots + a_1(t^k - 1).$$

Здесь каждое слагаемое (см., например, задачу 22, п. д)) делится на $t^k - 1$. Значит, если $(t^k - 1) \vdots m$, то и $(A - f(A)) \vdots m$. Равноостаточность остальных членов последовательности (4), а также ее членов, если она начинается с числа $A < t^k$, вытекает из ее построения.

Теорема 23. *Пусть A — натуральное число, представленное в виде*

$$a_n t^{kn} + a_{n-1} t^{k(n-1)} + \ldots + a_1 t^k + a_0,$$

где

$$0 \leq a_i < t^k \quad при \quad i = 0, 1, \ldots, n.$$

Положим

$$f(A) = \begin{cases} a_0 - a_1 + a_2 - \ldots \pm a_n, & если\ A \geq t^k, \\ остатку\ от\ деления\ A\ на\ m, & если\ m \leq A, < t^k, \\ не\ определено, & если\ A < m. \end{cases}$$

Тогда для того чтобы порождаемый функцией f алгоритм построения последовательности (4) *по правилу* (5) *был признаком равноостаточности при делении на m, необходимо и достаточно, чтобы было* $(t^k + 1) \vdots m$.

Доказательство.

Необходимость. В случае $A = t^k + a_0$ равноостаточность чисел A и $f(A) = a_0 - 1$ при делении на m дает нам $(t^k + 1) \vdots m$.

Достаточность. Мы имеем в нашем случае при $A \geq t^k$

$$A - f(A) = a_n(t^{kn} \pm 1) + a_{n-1}(t^{k(n-1)} \mp 1) + \ldots$$
$$\ldots \mp a_1(t^k + 1) \quad (*)$$

(знак «плюс» стоит здесь в члене, соответствующем нечетному коэффициенту при k в показателе, а знай «минус» — в члене, соответствующем четному коэффициенту). Согласно пп. д) и е) задачи 22 при нечетном r выражение $i^{kr} + 1$ делится на t^k+1, а при четном r выражение t^{kr}—1 делится на t^k+1. Значит, если (t^k+1) $\vdots m$, то на m делится каждый член в (*) справа, а потому и вся разность A — $f(A)$, Тем самым числа A и $f(A)$ оказываются равноостаточными при делении на m. Равноостаточность остальных членов последовательности (4), а также членов этой последовательности, если она начинается с числа $A<t^k$, вытекает непосредственно из ее построения.

14. Во многих задачах несущественна не только величина неполного частного от деления одного числа на другое, но также и величина остатка от деления, а интересно только, обращается этот остаток в нуль или нет, т. е. делится или нет первое число на второе. После сказанного в п. 1 ясно, как подходить к задачам такого рода.

Назовем числа a и b *равноделимыми* при делении на m, если либо и a и b делятся на m, либо на m ни a, ни b не делятся.

15. Пусть нужно выяснить делимость на m числа A. Будем строить последовательность убывающих натуральных чисел

$$A = A_0, A_1, A_2, \ldots, \quad (6)$$

равноделимых с A при делении на m с остатком. Способ построения последовательности (6) выберем такой, чтобы за всяким членом этой последовательности, большим или равным m по абсолютной величине, следовал еще хотя бы один член. Если при этом последний член (6) будет равен нулю, то A делится на m, а если не равен нулю, то не делится.

Всякий способ построения последовательности (6) назовем *признаком делимости* на m.

Очевидно, признаки делимости должны быть алгоритмичными, т. е. удовлетворять таким же условиям определенности, массовости и результативности, что и признаки равноостаточности.

Нетрудно проверить (это предоставляется читателю), что при помощи всякой функции $f(x)$, удовлетворяющей условиям а)—в) из п. 10 и условию

г*) если $f(x)$ имеет смысл, то числа x и $f(x)$ равноделимы на m,

можно построить признак делимости на m точно таким же образом, как строился признак равноостаточности при делении на m по всякой функции, удовлетворяющей условиям а)—г).

Найдем несколько признаков делимости.

Согласно теореме 16 достаточно уметь определять делимость чисел на числа вида p^{α} (степени простого числа).

16. Признак делимости на 7 в десятичной системе счисления. Пусть A — натуральное число. Представим A в виде $10a+b$, где $0 \leq b < 10$, как это уже делалось раньше. Положим

$$f_3(A) = \begin{cases} |a - 2b|, & \text{если } A \geqq 19, \\ \text{остатку от деления } A \text{ на } 7, & \text{если } 7 \leqq A < 19, \\ \text{не определена}, & \text{если } A < 7. \end{cases}$$

Функция $f_3(A)$ дает нам известный признак делимости на 7: число $10a+b$ $(0 \leq b < 10)$ делится на 7 тогда и только тогда, когда на 7 делится число $a - 2b$; полученное число снова проверяется этим же способом на делимость на 7 и т. д.

17. Признак делимости на 13. Представим натуральное число A в виде $10a+b$ и положим

$$f_4(A) = \begin{cases} a + 4b, & \text{если } A \geqq 40, \\ \text{остатку от деления } A \text{ на } 13, & \text{если } 13 \leqq A < 40, \\ \text{не определена}, & \text{если } A < 13. \end{cases}$$

18. Признаки делимости того же типа имеются и для чисел, записанных в других, недесятичных системах счисления.

Признак делимости на 11 в шестеричной системе счисления. Представим натуральное число A в виде $6a+b$, где $0 \leq b < 6$ (в соответствии со сказанным выше все рассуждения ведутся с употреблением обозначений и названий чисел в десятичной системе счисления), и положим

$$f(A) = \begin{cases} a + 2b, & \text{если } A \geqq 11, \\ 0, & \text{если } A = 11, \\ \text{не определено}, & \text{если } A < 11. \end{cases}$$

19. В предыдущих пунктах этого параграфа мы познакомились с большим количеством самых разнообразных признаков равноостаточности и признаков делимости. Практической целью построения всех этих признаков является получение удобно работающих алгоритмов нахождения остатков при делении на не-

которые определенные числа (признаки равноостаточности) или алгоритмов, обнаруживающих, равны эти остатки нулю или нет (признаки делимости). Насколько же мы осуществили поставленную цель?

Некоторые признаки равноостаточности, такие, как при делении на 2, 3, 5, 10 в десятичной системе счисления (и вообще — на делитель степени основания системы счисления) действительно оказались весьма практичными и удобными. Применение других признаков связано с более или менее громоздкими вычислениями.

Естественно поэтому искать и применять такие признаки делимости и равноостаточности, использование которых приводит к цели по возможности более простым путем.

Одна из трудностей, с которой мы сталкиваемся при такого рода попытках, состоит в том, что мы должны уметь простоту (или, наоборот, сложность) применения того или иного признака оценивать некоторым числом. В качестве такой числовой характеристики можно, например, взять число арифметических действий над однозначными числами, которые необходимо произвести в процессе применения данного признака к тому или иному числу.

К сожалению, всякая такая характеристика объема вычислений в сильной мере зависит от индивидуальных свойств того числа, делимость которого мы хотим испытать.

Так, например, очень легко убедиться в том, что остаток от деления числа 31 025 на 8 есть 1. Для этого достаточно найти остаток от деления на 8 числа 25. Но для нахождения остатка от деления 30 525 на 8 следует разделить на 8 с остатком число 525, а это уже требует большего числа выкладок.

В качестве другого примера рассмотрим признак равноостаточности при делении на 37 (см. задачу 36). Остаток от деления на 37 числа 10014 023 находится сложением 10+14 + 23 и делением полученной суммы на 37. Остаток, как легко видеть, равен 10. Однако немногие смогут в уме применить этот признак равноостаточности к числу 782 639 485.

Поэтому, говоря об удобстве использования признаков делимости и равноостаточности, мы должны отвлекаться от сложностей индивидуальных испытаний чисел на делимость, а оценивать возможности каждого признака «в среднем». При таком подходе мы можем надеяться точно сформулировать меру сложности признака делимости или равноостаточности и даже найти наиболее экономный в этом смысле признак.

4.7. Общие признаки равноостаточности и делимости

1. Все построенные выше признаки равноостаточности, а также признаки делимости выглядят несколько искусственно, а на первый взгляд может показаться, что эти признаки или, во всяком случае, некоторые из них были найдены случайно или же в результате проб и испытаний. На самом деле это не так. Оказывается, существуют способы построения признаков делимости и равноостаточности на любое наперед заданное число. Они называются *общими признаками делимости* или соответственно *общими признаками равноостаточности.*

Общие признаки делимости являются способами получения конкретных признаков делимости. Поэтому конкретные признаки делимости можно считать теми результатами, к которым приводят общие признаки. С этой точки зрения общие признаки делимости относятся к конкретным совершенно так же, как конкретный признак делимости относится к результату своего применения к некоторому числу, т. е. к остатку от деления данного числа a на данное число m.

Общие признаки делимости и равноостаточности напоминают алгоритмы, и притом алгоритмы довольно своеобразные: их итогами, результатами должны быть снова алгоритмы, именно, конкретные признаки делимости или равноостаточности.

Однако для того чтобы говорить об общих признаках делимости и равноостаточности как об алгоритмах, мы должны убедиться в том, что они обладают нужными условиями определенности, массовости и результативности.

Говоря подробнее, указывая общий признак делимости (равно как и общий признак равноостаточности), мы должны проверить выполнение следующих условий. Во-первых, по всякому числу m он должен действительно давать признак делимости (равноостаточности) на это число. Он должен, так сказать, «перерабатывать» каждое натуральное число m в соответствующий признак. Именно в этом и состоит его результативность. Во-вторых, общий признак должен быть определенным, т. е., примененный к заданному числу m, он должен приводить вполне определенным способом к вполне определенному конкретному признаку делимости (равноостаточности) на это число. Наконец, в-третьих, признак должен быть массовым, т. е. действительно общим, и давать признаки

делимости или равноостаточности на любое наперед заданное натуральное число.

В этом смысле описанный в п. 6 п. 4.8 способ задания признака равноостаточности, а также описанный в п. 9 п. 4.8 способ нахождения признаков делимости не являются общими признаками. Действительно, указание функций, удовлетворяющих нужным условиям, является процессом, не удовлетворяющим пока ни одному из требований определенности, массовости и результативности.

В самом деле, эти способы не дают нам никакой гарантии в том, что нужная функция будет найдена; значит, они лишены результативности. Далее, если требуемая функция и существует, к ней можно прийти разными путями, не говоря уже о том, что таких функций может оказаться несколько. Значит, эти способы лишены определенности. Наконец, ему не хватает и массовости, так как, быть может, требуемых функций для тех или иных конкретных значений m нам найти не удастся. Сам способ нам, во всяком случае, ничего об этом не говорит. Таким образом, для того чтобы описанный процесс стал алгоритмом, он должен быть еще дополнен точными указаниями, гарантирующими построение вполне определенной функции f_m для каждого конкретного числа m.

Эта задача «алгоритмизации» построения признаков делимости может быть решена, и даже без особого труда, а общие признаки делимости известны уже давно.

Один такой общий признак равноостаточности фактически нами уже был построен в п. 11 п. 4.5 при выяснении вопроса о делении с остатком. Мы его можем сформулировать так: каждому целому положительному числу m ставится в соответствие процесс последовательного вычитания этого числа m до получения числа, меньшего чем m (см. последнюю фразу п. 1 п. 4.7). Ясно, что такое соответствие обладает необходимыми свойствами определенности (мы точно знаем, что ставится в соответствие числу m: процесс последовательного вычитания m), массовости (процесс последовательного вычитания можно пытаться применить к любому m) и результативности (такая попытка обязательно приведет к успеху). Однако практическая ценность описанного общего признака равноостаточности весьма невелика.

Некоторое усовершенствование общего признака равноостаточности, основанного на последовательном вычитании, приводит к известному процессу деления целых чисел «углом». Этот процесс деления тоже может рассматриваться как общий признак равноостаточности. Нелишним будет напомнить, что подавляющее

большинство людей пользуется при нахождении остатков от деления именно этим признаком. При этом рассуждение ведется по следующей схеме, которую мы воспроизведем в двух вариантах: на обычном «житейском языке» и на языке алгорифмов.

«На житейском языке»

1. Необходимо найти остаток от деления данного *a* на данное *m;*
2. Для этого будем делить на *m;*
3. Начинаем делить *a* на *m;*
4. .. делим и получаю остатка.

На языке алгоритмов

1. Общий признак равноостаточности начинает переработку числа *m*;
2. Общий признак «выдает» результат переработки числа *m:* конкретный признак равноостаточности при делении на *m,* заключающийся в непосредственном делении на *m* с остатком;
3. Полученный конкретный признак начинает переработку числа *a:* деление на *m* с остатком;
4. Конкретный признак приводит к цели: к остатку от деления *a* на *m*.

В этом рассуждении первые три шага уже очень просты, и поэтому не приходится удивляться, что четвертый шаг, состоящий в фактическом выполнении деления, оказывается таким громоздким. Цель создания общих признаков равноостаточности и делимости и состоит в разгрузке четвертого шага за счет усовершенствования второго. Именно это и имеют обычно в виду, когда говорят об общих признаках делимости и равноостаточности.

2. Истооически первым общим признаком делимости (точнее, даже признаком равноостаточности) является следующий, предложенный знаменитым французским математиком Паскалем еще в середине XVII столетия. Сущность этого признака такова.

Пусть *m* - натуральное число. Составим последовательность чисел

$$r_1, r_2, r_3, \ldots. \qquad (1)$$

полагая

r_1 равным остатку от деления 10 на m,
r_2 » » » $10r_1$ » m,
r_3 » » » $10r_2$ » m

и т. д.

Представим теперь произвольное натуральное число *A* в виде

$$10^n a_n + 10^{n-1} a_{n-1} + \ldots + 10a_1 + a_0$$

и определим функцию

$$F_m(a) =$$

$$= \begin{cases} a_0 + r_1 a_1 + r_2 a_2 + \ldots + r_n a_n, & \text{если } 10^n \geqq m, \\ \text{остатку от деления } A \text{ на } m, & \text{если } 10^n < m \leqq A, \\ \text{не определена,} & \text{если } A < m. \end{cases}$$

Итак, нами построен признак равноостаточности при делении на произвольное *m,* т. е. некоторый общий признак равноостаточности.

3. В п. 19 п. 4.7 мы говорили о сравнительных качествах признаков делимости (или равноостаточности) на данное число. Так как общий признак делимости должен давать нам признаки делимости на любое натуральное число, то неудивительно, что он может для различных чисел приводить к признакам делимости весьма различного качества.

Так, например, общий признак Паскаля наряду с вполне приемлемыми признаками равноостаточности при делении на 3 и 11 дает весьма громоздкий и неудобный к применению признак равноостаточности при делении на 7 (см. задачу 54, д)).

В связи с этим по поводу общих признаков делимости и равноостаточности можно высказать соображения, подобные тем, которые производились в п. 19 п. 4.7 при обсуждении качества конкретных признаков делимости. В этом смысле наилучшим общим признакком делимости (равноостаточности) должен считаться тот, который в применении к любому наперед заданному целому положительному *m* дает наилучший признак делимости (равноостаточности) на этот. Необходимо знать, что задача нахождения наилучшего общего признака делимости далека не только от своего решения, но даже от строгой постановки.

4.8. Делимость степеней

1. Начнем с описания процесса, который можно было бы назвать «очень общим признаком равноостаточности».

Пусть k — некоторое натуральное число и r есть остаток от деления t^k на m:

$$t^k = mq + r \quad (0 \leqq r < m).$$

По следствию теоремы 20 при любом n числа r^n и t^{kn} при делении на m также должны быть равноостаточными.

Составим теперь для произвольного числа A его разбиение на k-значные «грани» справа налево, т. е. представим его в виде

$$A = a_n t^{kn} + a_{n-1} t^{k(n-1)} + \ldots + a_1 t^k + a_0,$$

где

$$0 \leqq a_i < t^k \quad \text{при} \quad i = 0, 1, \ldots, n,$$

и положим

$$f(A) = \begin{cases} a_n r^n + a_{n-1} r^{n-1} + \ldots + a_1 r + a_0, & \text{если } A \geqq t^k, \\ \text{остатку от деления } A \text{ на } m, & \text{если } m \leqq A < t^k, \\ \text{не определено}, & \text{если } A < m. \end{cases}$$

Ясно, что, как и ранее в аналогичных случаях, процесс построения чисел

$$A_0 = A, \quad A_1 = f(A_0), \quad A_2 = f(A_1), \ldots$$

является признаком равноостаточности.

2. Говоря формально, при составлении в п. 1 общего признака равноостаточности мы пользовались свойствами степеней, относящимися к их делимости. Однако вопрос о делимости степеней по существу является вопросом о делимости некоторых произведений. Поэтому и решить этот вопрос в принципе удалось на основе результатов п. 4.6. Вместе с тем практическая реализация полученного признака равноостаточнссти для тех или иных комбинаций значений чисел t и m может приводить к крупным значениям k и r, так что вычисление значений функции f может потребовать выполнения значительных вычислений, возможно даже превосходящих по объему выкладки по непосредственному делению на m.

Ясно, что вычисление значений функции f оказывается тем проще, чем меньшими будут значения чисел k и r. Разумеется, наиболее удобным оказывается в этом отношении тот случай, когда $r = 1$. Тогда значение f получается в результате выполнения наименее трудоемкого действия: сложения.

Согласно теореме 22 этот случай ($r = 1$) имеет место тогда и только тогда, когда $(t^k - 1) \mathbin{\vdots} m$ или, иными словами, когда t^k при делении на m дает в остатке 1. Встает вопрос: найдется ли при данных t и m такое k, что $(t^k - 1) \mathbin{\vdots} m$?

Все сказанное приводит к необходимости заняться изучением делимости степеней несколько более подробно.

3. Расширим несколько наши познания в области теории чисел.

Теорема 24 (теорема Ферма). *Если число p простое, то разность a^p—a делится на p.*

Доказательство.

Доказательство ведется индукцией по *a*. При $a = 1$ имеем

$$a^p - a = 1 - 1 = 0$$

и $0 \vdots p$.

Предположим, что a^p — a делится на p, и докажем, что $(a+1)^p$—$(a+1)$ также делится на p. Действительно, разлагая $(a + 1)^p$ по формуле бинома Ньютона, имеем

$$\begin{aligned}(a+1)^p - (a+1) &= \\ &= a^p + C_p^1 a^{p-1} + C_p^2 a^{p-2} + \ldots + C_p^{p-1} a + 1 - a - 1 = \\ &= a^p - a + C_p^1 a^{p-1} + C_p^2 a^{p-2} + \ldots + C_p^{p-1} a. \end{aligned} \qquad (*)$$

a^p — a делится на p по предположению. По следствию теоремы 13 C^k_p (при $1 \le k \le p-1$) также делится на p. Следовательно, на p делится каждое слагаемое правой части соотношения (*), а потому (теорема 6) и вся сумма.

Индуктивный переход обоснован, и вся теорема доказана.

Следствие. По теореме Ферма

$$a^p - a = a(a^{p-1} - 1) \vdots p.$$

Если при эгом a не делится на p, то по теореме 13 на p должно делиться a^{p-1}— 1.

Не следует путать эту так называемую «малую теорему Ферма» с «великой теоремой Ферма». Последняя утверждает, что при целом $n>2$ не существует таких целых a, b и c, что $a^n + b^n = c^n$. Несмотря на многочисленные попытки, великая теорема Ферма до сих пор не доказана и не опровергнута.

Следствие. Если p — простое и a не делится на p, то a^{p-1} — 1 делится на p.

Пусть натуральное число m имеет каноническое разложение:

$$m = p_1^{\alpha} p_2^{\alpha_2} \ldots p_k^{\alpha_k}; \qquad (1)$$

положим

$$\varphi(m) = p_1^{\alpha_1 - 1}(p_1 - 1)\, p_2^{\alpha_2 - 1}(p_2 - 1) \ldots p_k^{\alpha_k - 1}(p_k - 1). \qquad (2)$$

Формулы (1) и (2) ставят в соответствие каждому натуральному числу m некоторое вполне определенное число $\varphi(m)$. Это значит, что мы можем говорить о функции φ от натурального аргумента.

Определение. Определенная выше функция φ называется *функцией Эйлера.*

Функция Эйлера играет исключительно важную роль во многих вопросах теории чисел. Далее будет указано несколько применений этой функции.

Теорема 25. *При взаимно простых m_1 и m_2 имеет место равенство*

$$\varphi(m_1 m_2) = \varphi(m_1)\,\varphi(m_2).$$

Доказательство.

Пусть

$$m_1 = p_1^{\alpha_1} \dots p_k^{\alpha_k} \quad \text{и} \quad m_2 = q_1^{\beta_1} \dots q_l^{\beta_l}.$$

По теореме 15 каждое из чисел $p_1,\dots,p_k$ отлично от каждого из чисел $q_1, \dots, q_l$. Значит, каноническим разложением $m_1 m_2$ будет $p_1^{\alpha_1} \dots p_k^{\alpha_k} q_1^{\beta_1} \dots q_l^{\beta_l}$. Поэтому

$$\varphi(m_1 m_2) = p_1^{\alpha_1 - 1}(p_1 - 1) \dots p_k^{\alpha_k - 1}(p_k - 1) \times \\ \times q_1^{\beta_1 - 1}(q_1 - 1) \dots q_l^{\beta_l - 1}(q_l - 1),$$

т. е.

$$\varphi(m_1 m_2) = \varphi(m_1)\,\varphi(m_2).$$

Теорема 26 (теорема Эйлера). *Если числа a и m взаимно просты, то $a^{\varphi(m)} - 1$ делится на m.*

Доказательство.

Докажем сначала индукцией по α, что $a^{p^\alpha(p-1)} - 1$ делится на p^α. При $\alpha = 1$ доказываемое утверждение является, очевидно, следствием теоремы Ферма, справедливость которого уже была установлена. Таким образом, основание индукции доказано.

Предположим теперь, что

$$\left(a^{p^{\alpha-1}(p-1)} - 1\right) \vdots\, p^\alpha,$$

и рассмотрим выражение $a^{p^\alpha(p-1)} - 1$. Мы должны доказать, что оно делится на $p^{\alpha+1}$. Но

$$a^{p^\alpha(p-1)} - 1 = \left(a^{p^{\alpha-1}(p-1)}\right)^p - 1.$$

Так как $a^{p^{\alpha-1}(p-1)} - 1$, по предположению, делится на p^α, число $a^{p^{\alpha-1}(p-1)}$ имеет вид $Np^\alpha + 1$. Значит,

$$a^{p^\alpha(p-1)} - 1 = (Np^\alpha + 1)^p - 1,$$

т. е. по формуле бинома

$$a^{p^{\alpha}(p-1)} - 1 =$$

$$= N^p p^{\alpha p} + C_p^1 N^{p-1} p^{\alpha(p-1)} + \ldots + C_p^{p-1} N p^{\alpha} + 1 - 1$$

В последней сумме первое слагаемое делится на $p^{\alpha+1}$, так как оно делится на $p^{\alpha p}$ и $\alpha p \geqq \alpha + 1$. В каждое из следующих p—1 слагаемых этой суммы входит p с показателем, не меньшим α, и, кроме того, биномиальный коэффициент, в силу следствия теоремы 13 делящийся на p. Значит, каждое из этих слагаемых также делится на $p^{\alpha+1}$. Наконец, разность 1 — 1=0 может быть отброшена. Поэтому по теореме 6

$$\left(a^{p^{\alpha}(p-1)} - 1\right) \vdots\ p^{\alpha+1}.$$

Случай, когда число m имеет только один простой делитель, таким образом, разобран.

Предположим теперь, что теорема Эйлера доказана для показателей m_1 и m_2, причем числа m_1 и m_2 взаимно простые. Докажем теорему Эйлера для показателя $m = m_1 m_2$. Если потом положить

$$m_1 = p_1^{\alpha_1} \ldots p_k^{\alpha_k} \quad \text{и} \quad m_2 = p_{k+1}^{\alpha_{k+1}},$$

то мы, очевидно, и получим индуктивный переход, необходимый нам для завершения доказательства теоремы. Итак, доказываем высказанное утверждение. Пусть числа a и m взаимно просты. Тогда a также взаимно просто с m_1. Значит, и $a^{\varphi(m_2)}$ взаимно просто с m_1. Поэтому, по предположению,

$$\left(a^{\varphi(m_2)}\right)^{\varphi(m_1)} - 1 = a^{\varphi(m_1)\varphi(m_2)} - 1 = a^{\varphi(m_1 m_2)} - 1 = a^{\varphi(m)} - 1$$

делится на m_1. Точно так же убеждаемся в том, что $a^{\varphi(m)} - 1$ делится и на m_2. А так как числа m_1 и m_2 взаимно простые, $a^{\varphi(m)} - 1$ делится и на их произведение, т. е. на m. Теорема Эйлера доказана.

Остатки при делении одного и того же делимого на различные делители связаны между собой достаточно сложным образом. Из теоремы Эйлера можно получить принципиально важную для нас зависимость остатков от деления на взаимно простые множители и от деления на их произведение.

4. На основании установленных фактов мы можем сформулировать общий признак равноостаточности для произвольного делителя m в произвольной системе счисления t в той явной и достаточно удобной форме, о которой говорилось в п. 1.

Напомним снова, что всякий признак равноостаточности есть алгоритм, т. е. некоторый процесс, и потому всякое его описание должно носить характер развивающегося повествования.

Итак, пусть нам даны числа m и t. Представим m в виде такого произведения $m = m_1 m_2$, что $(m_1, t) = 1$, и для некоторого показателя k имеет место делимость $t^k \mid m_2$. Согласно теореме 18 такое представление возможно. В силу задачи 66 вопрос о равноостаточносги при делении на $m_1 m_2$ может быть сведен к аналогичным вопросам для деления на m_1 и m_2. Но признак равноостаточности на m_2 содержится в теореме 21, а признак равноостаточности на m_1 — в теореме 22. После применения этих признаков равноостаточности следует воспользоваться результатом задачи 66.

Например, в случае нахождения признака равноостаточности при делении на 12 в десятичной системе счисления, очевидно, $m_1 = 3$, $m_2 = 4$ и $k = 2$.

Описанный процесс является общим признаком равноостаточности в том смысле, что по любому m он выдает некоторый конкретный признак равноостаточности. Это вытекает из алгоритмичности представления числа m в форме, указываемой в теореме 18, а сама эта алгоритмичность следует из алгоритмичности построения канонического разложения чисел (см. п. 9 п. 4.7).

Нам остается сформулировать в явном виде указанный признак равноостаточности при делении на m_1, пользуясь возможностью определить показатель k на основании теоремы Эйлера.

5. Применяя доказанные теоремы, построим несколько общих признаков делимости и равноостаточности.

Фиксируем натуральное m и представим число A в виде

$$A = a_0 + a_1 10^{\varphi(m)} + a_2 10^{2\varphi(m)} + \ldots + a_k 10^{k\varphi(m)},$$

где

$$0 \leqq a_0,\ a_1,\ a_2,\ \ldots,\ a_k \leqq 10^{\varphi(m)},$$

т. е. все a_i $(i = 0, 1, \ldots, k)$ являются $\varphi(m)$-значным числами.

Функция

$$F(A) = \begin{cases} a_0 + a_1 + \ldots + a_k, & \text{если } A \geqq 10^{\varphi(m)}, \\ \text{остатку от деления } A \text{ на } m, & \text{если } m \leqq A < 10^{\varphi(m)}, \\ \text{не определена}, & \text{если } A < m, \end{cases}$$

определяет, как нетрудно проверить, некоторый общий признак равноостаточности.

Теорема 27. *Если числа a и m взаимно просты, а числа* k_1 *и* k_2 *равноостаточны при делении на* $\varphi(m)$, *то числа* a^{k_1} *и* a^{k_2} *равноостаточны при делении на m.*

Доказательство.

Пусть

$$k_1 = \varphi(m)\, q_1 + r,$$
$$k_2 = \varphi(m)\, q_2 + r.$$

Тогда

$$a^{k_1} = a^{\varphi(m)\, q_1 + r} = \left(a^{\varphi(m)}\right)^{q_1} a^r.$$

На основании теоремы Эйлера и теоремы 20 число $a^{\varphi(m)\, q_1} a^r$ равноостаточно при делении на m с числом a^r. Аналогично устанавливается равноостаточность при этом делении чисел a^{k_2} и a^r. Значит, и числа a^{k_1} и a^{k_2} при делении на m равноостаточны.

6. Построенный общий признак равноостаточности не является во многих случаях, так сказать, «достаточно экономным», так как число $\varphi(m)$ может, вообще говоря, оказаться довольно большим.

Поэтому, с одной стороны, при пользовании этим признаком приходится складывать большие числа, а с другой стороны, $\varphi(m)$-значные числа при этом приходится делить на m непосредственно (или же пользоваться каким-нибудь другим признаком делимости и равноостаточности). Желательно поэтому попытаться взять вместо $\varphi(m)$ другой, меньший показатель. В ряде случаев это удается сделать. Например, при $m = 37$ вместо $\varphi(m)=36$ можно взять показатель 3, ибо 1000 при делении на 37 дает в остатке единицу; при $m= 11$ вместо $\varphi(m)= 10$ можно взять показатель 2 и т. д.

Определение. Наименьшее число δ, для которого a^δ при делении на m с остатком дает в остатке 1, называется *показателем, которому принадлежит число a* при делении на m с остатком.

Чаще это число принято называть *показателем, которому принадлежит число a по модулю m.*

Очевидно, каковы бы ни были взаимно простые числа a и m, показатель δ, которому принадлежит a при делении на m, не превосходит $\varphi(m)$. Этот показатель и можно взять вместо $\varphi(m)$ в формулировке общего признака равноостаточности из п. 5.

7. Показатель, которому принадлежит число a при делении на m, может, вообще говоря, быть и равным $\varphi(m)$. Например,

последовательностью остатков от деления степеней числа 2 на 11 будет

$$2, 4, 8, 5, 10, 9, 7, 3, 6, 1,$$

так что при делении на 11 число 2 принадлежит показателю 10. Значит, для применения признака равноостаточности из п. 5 в этом случае приходится брать $k = 10 = \varphi(11)$.

Однако во многих случаях удается обходиться показателем $\frac{1}{2}\varphi(m)$. Пусть, например, m есть степень простого числа: $m = p^\alpha$ и $p \neq 2$. Тогда $\varphi(m) = p^{\alpha-1}(p-1)$, и теорема Эйлера приобретает вид: для $(a, p) = 1$ должно быть $\left(a^{p^{\alpha-1}(p-1)} - 1\right) \vdots\ p^\alpha$. Так как число $p^{\alpha-1}(p-1)$ четное, последнее делимое есть разность квадратов, и мы имеем

$$\left(a^{\frac{1}{2}p^{\alpha-1}(p-1)} + 1\right)\left(a^{\frac{1}{2}p^{\alpha-1}(p-1)} - 1\right) \vdots\ p^\alpha.$$

Так как $p \neq 2$ оба сомножителя одновременно на p делиться не могут. Значит, на p^α делится либо $a^{\frac{1}{2}\varphi(m)} + 1$, либо $a^{\frac{1}{2}\varphi(m)} - 1$.

В первом случае мы оказываемся в условиях теоремы 23 с $k = \frac{1}{2}\varphi(m)$, а во втором — в условиях теоремы 22 с тем же $k = \frac{1}{2}\varphi(m)$.

8. Применения функции Эйлера и теоремы Эйлера не ограничиваются признаками делимости. Например, при их помощи можно решать уравнения в целых числах.

Теорема 28. *Если числа a и b взаимно просты, то уравнение*

$$ax + by = c \qquad (3)$$

всегда разрешимо в целых числах, и целыми его решениями будут все пары чисел (x_t, y_t), *где*

$$x_t = ca^{\varphi(b)-1} + bt,$$

$$y_t = c\frac{1 - a^{\varphi(b)}}{b} - at$$

(t — любое целое число).

Доказательство.

Найдем сначала хотя бы одно решение (x', y') этого уравнения. Очевидно, что для этого достаточно найти такое число x', что $(ax' - c) \vdots\ b$. По теореме Эйлера $(a^{\varphi(b)} - 1) \vdots\ b$. Значит, $(ca^{\varphi(b)} - c) \vdots\ b$, и в качестве x' можно взять число $ca^{\varphi(b)-1}$.

Пусть теперь *(x″, y″)* — какое-то другое решение уравнения $ax+by=c$. Покажем, что числа x' и x'' равноостаточны при делении на b. В самом деле, пусть

$$ax' + by' = c,$$
$$ax'' + by'' = c.$$

Вычитая почленно второе равенство из первого, получаем

$$a(x' - x'') - b(y' - y'') = 0,$$

откуда $a(x' - x'') \vdots b$. Так как a и b по условию взаимно просты, по теореме 12 $(x'-x'') \vdots b$, и нам остается сослаться на теорему 19.

Таким образом, все искомые значения x находятся среди чисел

$$x_t = ca^{\varphi(b)-1} + bt.$$

Но $(ax_t - c) \vdots b$, так что, полагая

$$y_t = \frac{-ax_t + c}{b} = c\frac{1 - a^{\varphi(b)}}{b} - at,$$

мы получаем, что все пары чисел x_t и y_t являются решениями нашего уравнения.

9. Теорема 29. *Пусть m взаимно просто с 10 и k равноостаточно с $10^{\varphi(m)-1}$ при делении на m. Тогда числа*

$$10a + b \quad и \quad a + kb$$

равноделимы на m.

Доказательство. Ввиду взаимной простоты m и 10, числа $10a + b$ и $(10a + b)10^{\varphi(m)-1}$ по теореме 15 равноделимы на m. Но

$$(10a + b)10^{\varphi(m)-1} = 10^{\varphi(m)}a + 10^{\varphi(m)-1}b,$$

так что по теореме Эйлера и теореме 20 число $10a + b$ равноделимо на m с числом $a + kb$.

Опираясь на эту теорему, можно построить следующий общий признак делимости. Обозначим через k остаток от деления $10^{\varphi(m)-1}$ на m с остатком, представим произвольное число A в виде

$$10a + b \qquad (0 \leqq b < 10)$$

и положим

$$F(A) = \begin{cases} a + kb, & \text{если } A > a + kb, \\ \text{остатку от деления } A \text{ на } m, & \text{если } m \leqq A < a + kb, \\ \text{не определена}, & \text{если } A < m. \end{cases}$$

Если k велико (близко к m), то вместо него в формулировке соответствующего признака целесообразно брать $k - m$.

Микромодуль 15
Индивидуальные тестовые задания

Задачи. Доказать следующие утверждения:

1. $0 \vdots a$.
2. $a \vdots 1$.
3. Если $1 \mid a$, то $a = 1$.
4. Каково бы ни было $a \neq 0$, существует такое отличное от a число b, что $b \mid a$.
5. Каково бы ни было число a, существует такое число b, что из $b \mid c$ и $c \mid a$ следует либо $c = b$, либо $c = a$.
6. Доказать теоремы, аналогичные теоремам 2, 3, 4 и 5 для четной делимости.
7. Построить такую теорию делимости, в которой теоремы 1, 3 и 4 были бы верными, а теоремы 2 и 6 — нет.

Задача 8. Опираясь только на свойства 1°—7° (см. п. 4.5) отношения $\geq$ и не пользуясь никакими свойствами самих чисел и действий над ними:

а) Доказать единственность минимального числа;

б) Доказать единственность непосредственно предшествующего числа;

в) Сформулировать определение числа, непосредственно следующего за данным числом a (т.е. числа $a + 1$), и доказать его существование и единственность.

Задача 9. Проверить, какие из утверждений 1°—7° остаются в силе для отношения «больше» (>).

Задача 10. Пусть пары объектов произвольной природы (ими могут быть числа, точки, функции, теоремы и т. д.) связываются некоторым орошением f, обладающим свойствами, аналогичными свойствам 1° — 7°. Доказать, что тогда эти объекты (элементы) можно перенумеровать (т. е. выписать их в некотором порядке) $A_1, A_2, A_3, \ldots$ так, что $A_i \leftarrow A_j$ тогда и только тогда, когда $i \geq j$.

Сказанное, по существу, означает, что отношение, обладающее свойствами 1°—7°, упорядочивает множество в линейную цепочку элементов:

$$A_1 \rightarrow A_2 \rightarrow A_3 \rightarrow \ldots$$

Задача 11. Вывести «новую форму» принципа индукции из ее «старой формы».

Задача 12. Сформулировать и доказать теорему о делении с остатком для четной делимости.

Задача 13. Оценить сверху наименьший простой делитель составного числа *a.*

Задача 14. Указать способ построения по каноническим разложениям двух чисел канонических разложений наименьшего общего кратного этих чисел и их наибольшего общего делителя.

Задача 15. Обозначим через $\tau(a)$ число различных делителей числа a (включая единицу и само число *a).* Показать, что для числа a с каноническимразложением $p_1^{\alpha_1}p_2^{\alpha_2}\dots p_r^{\alpha_r}$

$$\tau(a) = (\alpha_1 + 1)(\alpha_2 + 1)\dots(\alpha_r + 1).$$

Задача 16. Найти *a,* если известно, что $a \vdots 3$, $a \not\vdots 4$ и $\tau(a) = 14$.

Задача 17. Каноническое разложение числа a имеет вид $p_1^{\alpha_1}p_2^{\alpha_2}$ и $\tau(a^2) = 81$. Чему равно $\tau(a^3)$?

Задача 18. Чему равно *a,* если $a = 2\tau(a)$?

Задача 19. Доказать, что каково бы ни было $K > 0$, найдется такое натуральное число k, что для всякого числа *a,* имеющего k простых сомножителей, будет

$$\frac{\tau(a^2)}{\tau(a)} > K.$$

Задача 20. Верны ли для четной делимости аналоги теорем 11—14?

Задача 21. Найти остаток от деления:

а) $A = (116 + 17^{17})^{21}$ на 8;

б) $A = 14^{256}$ на 17.

Задача 22. Доказать, что при любом *n:*

а) $(n^3+11n) \vdots 6$;

б) $(4^n+15n - 1) \vdots 9$;

в) $(10^{3n} - 1) \vdots 3^{n+2}$;

г) При любом *a*

$$(a^{2n+1} + (a - 1)^{n+2}) \vdots (a^2 - a + 1);$$

д) При любом *k*

$$(n^k-1) \vdots (n-1);$$

е) При любом нечетном *k*

$$(n^k + 1) \vdots (n+1).$$

Задача 23. Эквивалентное отношение $\sim$ на множестве чисел разбивает это множество на такие классы (называемые классами эквивалентности), что любые два числа из одною класса связаны

отношением эквивалентности, а никакие два числа из разных классов этим отношением не связаны. (Доказать)

В этой задаче речь идет об отношении эквивалентности, связывающем числа. Однако это несущественно, и утверждение задачи справедливо для эквивалентных отношений, связывающих объекты совершенно произвольной природы

Задача 24. Если на множестве целых чисел задано эквивалентное отношение ~, разбивающее это множество на m классов, и такое, что из $a \sim b$ и $c \sim d$ следует $a + c \sim b + d$, то отношение ~ есть сравнимость по модулю m (т. е. $a \sim b$ тогда и только тогда, когда $a = b(mod\ m)$).

Задача 25. Сформулировать и доказать правила сокращения сравнений.

Задача 26. Если число p простое и a не делится на p, то никакие два числа из a, $2a$, $3a$.....$(p—1)a$ не сравнимы друг с другом по модулю p. Поэтому при делении на p чисел a, $2a$, $3a$.....$(p—1)a$ мы получим по одному разу все остатки, кроме нуля.

Задача 27 (теорема Вильсона). *Для того чтобы число p было простым, необходимо и достаточно, чтобы* $(p—1)!+1=1•2...(p— 1)+1$ *делилось на p.*

Задача 28. Сформулировать и доказать для равноостаточности теорему, аналогичную теореме 16.

Задача 29 а) Последний отличный от нуля остаток r_n в применении алюритма Евктида к числам a и b есть (a, b).

б) Каковы бы ни были натуральные a и b, существуют такие целые A и B ч го $aA + bB = (a, b)$

Задача 30 Вывести из результата б) задачи 29 теоремы 9, 12, 13 и 14. (Подчеркнем, что наши рассуждения, связанные с алгоритмом Евклида, были основаны только на возможности деления с остатком. Мы не пользовались в них ни теоремами 9 —14, ни какими-либо иными соображениями, опирающимися на основную теорему арифметики.)

Задача 31. Указать и проанализировать аналогичные признаки равноостаточности при делении на 2, 4, 8, 10, 16, 20 и 25 в десятичной системе счисления.

Задача 32. Указать и проанализировать аналогичные признаки равноостаточности при делении:

а) на 9 и 27 в троичной системе счисления;

б) на 8, 9, 16, 18, 24, 36, 48 и 72 в двенадцатеричной системе счисления.

Задача 33. Представим натуральное число A в виде

$$10^k a + b \quad (0 \leqq b < 10^k)$$

и положим

$$f(A) = \begin{cases} b, & \text{если } A \geqq 10^k, \\ \text{остатку от деления } A \text{ на } m, & \text{если } m \leqq A < 10^k, \\ \text{не определена}, & \text{если } A < m. \end{cases}$$

Для каких чисел m такой алгоритм при некотором k является признаком равноостаточности?

Задача 34. Проверить, что функция $f_2(x)$ удовлетворяет условиям а)—г) (см. п.10 п. 4.7) и определяет тем самым некоторый признак равноостаточности при делении на 3.

Задача 35. Применить построенный признак равноостаточности при делении на 3:

а) к числам 858 773 и 789 988;

б) к числу, десятичная запись которого состоит из 4444 четверок.

Задача 36. Указать и проанализировать аналогичные признаки равноостаточности при делении на 7, 9, 11, 13 и 37 в десятичной системе счисления.

Задача 37. Указать и проанализировать признаки равноостаточностп при делении:

а) на 2, 4 и 8 в троичной системе счисления;

б) на 2, 4 и 8 в семеричной системе счисления.

Задача 38. Указать подпадающие под теорему 22 признаки равноостаточности для чисел, записанных в шестеричной, семеричной, девятеричной и тринадцатеричной системах счисления.

Задача 39. Указать подпадающие под теорему 23 признаки равноостаточности для чисел, записанных в троичной, пятеричной, восьмеричной и десятичной системах счисления.

Задача 40. Каково бы ни было m, любые равноостаточные при делении на m числа являются равноделимыми на m. Показать на примере, что обратное неверно.

Задача 41. Для каких m из равноделимости двух чисел при деления на m следует их равноостаточность при этом делении?

Задача 42. Доказать, что отношение равноделимости при делении на данное число m является эквивалентным отношением и разбивает множество целых чисел на два класса.

Задача 43. Будет ли справедлива для равноделимых чисел теорема 20? Ее следствие?

Задача 44. Доказать, что всякий признак равноостаточности при делении на m является признаком делимости на m.

Задача 45. Проверить выполнение для функции $f_3(A)$ условий а)—в) и г*) (см. п.15 п. 4.7).

Задача 46. Доказать, что полученный признак делимости на 7 не является признаком равноостаточности при делении на 7 с остатком.

Задача 47. Проверить выполнение условий а)—в) и г*) (см. п.15 п. 4.7) для функции $f_4(x)$ и сформулировать полученный признак делимости на 13.

Задача 48. К каким последствиям приведет замена в определении функции f_4 числа 40 на меньшее?

Задача 49. По аналогии с построенными признаками делимости на 7 и 13 построить аналогичные признаки делимости на 17, 19, 23, 29 и 31.

Задача 50. Построить два признака делимости на 49.

Задача 51. Проверить соблюдение условий а)—в) и г*) (см. п.15 п. 4.7) для функции f и сформулировать полученный признак делимости.

Задача 52. По аналогии с только что построенным признаком делимости построить признаки делимости:

а) на 5 в семеричной системе счисления;

б) на 7 в одиннадцатеричной системе счисления;

в) на 17 в двенадцатеричной системе счисления.

Задача 53. Проверить, что функция F_m при любом m удовлетворяет условиям а)—г) из п. 10 п. 4.7.

Задача 54. Сформулировать получаемые из общего признака равноостаточности Паскаля признаки равноостаточности при делении:

а) на 2, 5 и 10;

б) на 4, 20 и 25;

в) на 3 и 9;

г) на 11;

д) на 7.

Задача 55. Пусть в последовательности (12)

r_1 есть остаток от деления 100 на m,

r_2 » » » $100\,r_1$ на m,

r_3 » » » $100\,r_2$ на m

и т. д.

Вывести отсюда общий признак равноостаточности, аналогичный общему признаку равноостаточности Паскаля.

Задача 56. Вывести общий признак равноостаточности в t-ичной системе счисления, аналогичный признаку Паскаля.

Задача 57. Убедиться в том, что процесс построения чисел

$$A_0=A,\ A_1=f(A_0),\ A_2=f(A_1)\ldots$$

является признаком равноостаточности.

Задача 58. Полагая $t = 10$ и $k = 2$, найти остаток от деления числа 1 048 576 на 7.

Задача 59. Убедиться в том, что описанный в задаче 57 признак равноостаточности является лишь более явной формой того обобщения признака Паскаля, которое было упомянуто в задаче 56.

Задача 60. Привести пример, показывающий, что как теорема 24, так и ее следствие для составного *p,* вообще говоря, неверны.

Задача 61. Доказать теорему Ферма, опираясь на результат задачи 26.

Задача 62. Вычислить φ(12), φ (120), φ (1000).

Задача 63. Определить все числа *m,* для которых:

а) φ (*m*)= 10;

б) φ (*m*)= 8.

Задача 64. Доказать, что не существует такого *m,* для которого φ (*m*)= 14.

Задача 65. Показать, что φ (*m*) равно числу натуральных чисел, взаимно простых с *m* и меньших *m.* Это свойство функции Эйлера является чрезвычайно важным. Его часто принимают за определение этой функции.

Задача 66. Пусть $(m_1 m_2) = 1$, а a_1 и a_2— числа, равноостаточные с A при делении соответственно на m_1 и m_2. Тогда равноостаточным с A при делении на $m_1 m_2$ будет число

$$(a_1 m_2 + a_2 m_1)(m_1 + m_2)^{\varphi(m_1 m_2)-1}.$$

Задача 67. Функция

$$F(A) = \begin{cases} a_0 + a_1 + \ldots + a_k, & \text{если } A \geqq 10^{\varphi(m)}, \\ \text{остатку от деления } A \text{ на } m, & \text{если } m \leqq A < 10^{\varphi(m)}, \\ \text{не определена}, & \text{если } A < m, \end{cases}$$

определяет, как нетрудно проверить, некоторый общий признак равноостаточности.

Проверить это обстоятельство.

Задача 68. Сформулировать получаемые на основе общего признака равноостаточности конкретные признаки равноостаточности при делении на 7, 11 и 13.

Задача 69. Сформулировать аналогичный общий признак равноостаточности для произвольной *t*-ичной системы счисления. Убедиться в том, что получаемый так общий признак

равноостаточности по своей формулировке не зависит от основания системы счисления *t*.

Задача 70. Доказать, что $(n^{13} — n)$ ⋮2730.

Задача 71. Модифицировать построенный общий признак делимости, используя вместо $\varphi(m)$ показатель, которому принадлежит 10 при делении на m с остатком.

Задача 72. То же для t-ичной системы счисления,

Задача 73. Доказать теорему, аналогичную теореме 28, не предполагая взаимной простоты чисел a и b.

Задача 74. Найти способ решения уравнений вида (3) в целых числах на основе результата задачи 29, б).

Задача 75. Решить в целых числах уравнения:

а) $5x + 7y = 9$;

б) $25x + 13y = 8$.

Задача 76. Проверить для функции F выполнение условий а)—в) из п. 10 п. 4.7 и г*) из п. 15 п. 4.7.

Задача 77. Па основании построенного общего признака делимости вывести признак делимости на числа 17, 19, 27, 29, 31 и 49.

Задача 78. Построить общий признак делимости, представляя произвольное натуральное число в виде

$$100a + b \quad (0 \leqq b < 100),$$

и вывести из него признаки делимости на 17, 43, 49, 67, 101, 199.

Задача 79. Построить аналогичный общий признак делимости в t-ичной системе счисления.

Задача 80. На основании построенного общего признака делимости вывести конкретные признаки делимости:

а) на число 21 в восьмеричной системе счисления;

б) на число 31 в двенадцатеричной системе счисления.

Решение тестовых задач

1. $0 = a \cdot 0$ при любом a.

2. $a = 1 \cdot a$, значит, a ⋮ 1.

3. Пусть 1 ⋮ a. Это значит, что $1 = ac$ при некотором целом c. Отсюда следует, что $|a| \leq 1$. А так как $a \neq 0$, должно быть $a = 1$.

4. Достаточно взять любое $c > 1$ и положить $b = ac$.

5. В качестве такого b можно взять, например, $2a$. Пусть при этом для некоторого c и $2a \nmid c$ и $c \nmid a$. Это значит, что найдутся такие d_1 и d_2, что $2a = d_1c$ и $c = d_2a$. Отсюда следует, что $2a = d_1d_2a$ или после сокращения на a,

$$2 = d_1d_2.$$

Но при целых d_1 и d_2 такое равенство возможно лишь в случае, когда одно из этих чисел равно 1, а другое 2. Если $d_1 = 1$, то $c = 2a = b$; если же $d_2 = 1$, то $c = a$.

6. Доказательства ничем не отличаются от доказательств в случае обычной делимости

7. Пусть n — некоторое фиксированное число, большее единицы. Положим $a \overset{\cdot}{\underset{n}{\vdots}} b$, если найдется такое целое c, что $a = bc$ и $c \leq n$ Справедливость теорем, аналогичных теоремам 1, 3 и 4, проверяется без труда. Однако если мы возьмем $a = nb$ и $b = nc$, то $a \overset{\cdot}{\underset{n}{\vdots}} b$ и $b \overset{\cdot}{\underset{n}{\vdots}} c$. В этом случае $a = n^2c$, а так как $n^2 > n$, делимость $a \overset{\cdot}{\underset{n}{\vdots}} c$ не имеет места Точно также не имеет места делимость

$$(a + a) \overset{\cdot}{\underset{n}{\vdots}} b.$$

8 а) Пусть имеются два минимальных числа a_1 и a_2. В силу дихотомичности либо $a_1 \geq a_2$, либо $a_2 \geq a_1$. Если $a_1 \geq a_2$, то из минимальности a_1 следует, что $a_1 = a_2$. Если же $a_2 \geq a_1$, то $a_1 = a_2$ следует из минимальности a_2.

б) Пусть a — некоторое число, а b_1 и b_2 — два непосредственно предшествующих ему числа По дихотомичносги должно быть либо $b_1 \geq b_2$, либо $b_2 \geq b_1$. Пусть, для определенности, $b_1 \geq b_2$. Мы имеем $a \geqq b_1 \geqq b_2$, а так как число b_2 непосредственно предшествует числу a, должно быть либо $b_1 = a$, либо, $b_1 = b_2$. Но по условию $b_1 \neq a$; значит, $b_1 = b_2$, и требуемая единственность доказана.

в) *Непосредственно следующим* за a числом называется такое b, что $b \geqq a$, $b \neq a$, и из $b \geqq c \geqq a$ следует либо $c = b$, либо $c = a$

Предположим, что некоторое a не имеет непосредственно следующего за ним числа. Это значит, что для любого $a_n \geq a$ и отличного от a найдется такое a_{n+1}, отличное как от a_n, так и от a, что $a_n \geq a_{n+1} \geq a$. Возьмем теперь произвольное $a_1 \geq a$ и отличное от a (в силу 2° это сделать можно) и, исходя из него построим бесконечную последовательность различных чисел

$$a_1 \geqq a_2 \geqq \ldots \geqq a_n \geqq a_{n+1} \geqq \ldots \geqq a.$$

Существование же этой последовательности противоречит 4° Следовательно, непосредственно следующее число существует.

Единетвенность его устанавливается при помощи дихотомичности подобно тому как это делалось в пп а) и б)

9. Остается в силе транзитивность (3°), неограниченности множества чисел (5°), свойство 4° и существование непосредственно предшествующего числа (6°) Дихотомичность заменяется *трихотомичностыо* (либо $a > b$, либо $b > a$, либо $a = b$.

Становится неверным свойство рефлексивности (1°), ибо $a > a$ всегда неверно.

Что же касается, наконец, утверждения 2°, то формально оно остается в силе (хотя, быть может, и выглядит несколько пародоксально)

В самом деле, говоря строго, это утверждение в нашем случае формулируется так: для любых натуральных чисел a и b из $a > b$ и $b > a$ следует $a = b$

Предположим, что это высказывание неверно. Тогда найдутся такие натуральные числа a и b, что одновременно будет и $a > b$ и $b > a$, и $a \neq b$, а этого не может быть Полученное противоречие доказывает истинность нашего утверждения

10. Пусть множество упорядочено отношением $\mathfrak{f}$, обладающим свойствами 1° — 7°. Как уже было установлено, оно обладает минимальным элементом. Обозначим этот элемент через a_0. Из результатов задачи 8 следует, что каждый элемент обладает непосредственно следующим. Обозначим непосредственно следующий за a_0 элемент через a_1, непосредственно следующий за a_1 через a_2 и т. д. В итоге мы получаем последовательность

$$a_0, a_1, a_2, \ldots, \qquad (1)$$

в которой $a_{n+1} \mathfrak{f} a_n$ при любом n. По рефлексивности и транзитнвности отношения $\mathfrak{f}$ отсюда следует, что $a_i \mathfrak{f} a_j$ тогда и только тогда, когда $i \geq j$. Нам остается показать, что последовательность (1) охватывает все рассматриваемые нами объекты. Это достигается довольно тонким рассуждением по индукции. Предположим, что b_0 не принадлежит последовательности (1). Получение этого b_0 будем считать первым шагом нашего индуктивного рассуждения. Пусть n его шагов уже проведены, в результате чего нами получен некоторый элемент b_{n-1}.

Если $b_{n-1} = a_0$, то наш процесс будем считать законченным; если же $b_{n-1} \neq a_0$, то элемент b_{n-1} имеет непосредственно предшествующий, который мы и возьмем в качестве b_n. В результате мы получаем последовательность различных элементов

$$b_0 \mathfrak{f} b_1 \mathfrak{f} b_2 \mathfrak{f} \ldots \mathfrak{f} b_n \mathfrak{f} \ldots$$

На основании 4° эта последовательность должна иметь последний член. Но по самому принципу построения этой последовательности ее последним членом может быть только a_0. Пусть для определенности $b_n = a_0$.

Нетрудно проверить, что если некоторое a непосредственно предшествует b, то b непосредственно следует за a. Значит,

$$b_{n-1} = a_1,\ b_{n-2} = a_2,\ \ldots,\ b_0 = a_n.$$

Последнее означает, что a_0 принадлежит последовательности (1), но это противоречит предположенному. Следовательно, последовательность (1) содержит все рассматриваемые нами объекты.

11. Пусть a — некоторое число. Всякую последовательность различных чисел $a_0 = a, a_1, a_2, \ldots, a_n$, для которых

$$a_0 \leftarrow a_1 \leftarrow a_2 \leftarrow \ldots \leftarrow a_n, \qquad (2)$$

где a_n минимально в смысле упорядочения f, назовем *цепью предшественников* a_0. Число n называется *длиной* этой цепи. Покажем сначала, что при тех условиях, которые мы наложили на упорядочение f, каждое конкретное число не может иметь сколь угодно длинных цепей предшественников. В самом деле, пусть a — некоторое число, а $b_1, b_2, \ldots, b_k$ — непосредственно предшествующие ему числа. Если a_1 не предшествует a_0 непосредственно, мы можем на основании 9° вставить в цепь (2) некоторое непосредственно предшествующее a число. Поэтому если имеются сколь угодно длинные цепи предшественников a, должны найтись и такие его сколь угодно длинные цепи предшественников, которые начинаются с чисел, непосредственно предшествующих a. Будем далее рассматривать только такие цепи.

Каждая цепь предшественников a ровно на единицу длиннее некоторой цепи предшественников одного из непосредственно предшествующих ему чисел. Если бы каждое из них имело цепи предшественников ограниченной длины, то само a не могло бы иметь сколь угодно длинных цепей предшественников.

Значит, при нашем предположении хотя бы одно из чисел, непосредственно предшествующих a_0, имеет сколь угодно длинные цепи предшественников. Обозначим это число через a_1 и повторим в применении к нему все только что проведенные рассуждечия. Это даст нам некоторое число a_2, непосредственно редшествующее a_1 и имеющее сколь угодно длинные цепи предшественников. Повторяя этот процесс, мы приходим к последовательности

$$a_0 \leftarrow a_1 \leftarrow a_2 \leftarrow \ldots,$$

которая в силу 4° должна рано или поздно оборваться. Это значит, что последовательность будет иметь такой член, к которому наши рассуждения уже будут неприменимы. Но применимость рассуждений к каждому последующему члену последовательюсти нами уже была установлена. Полученное противоречие показывает, что ни одно число не имеет сколь угодно длинных цепей предшественников.

Следовательно, для каждого числа a среди его цепей предшественников можно выбрать самую длинную. Обозначим ее длину через $n(a)$. Если b непосредственно предшествует a, то очевидно, $n(b)=n(a)$—1, а для всех минимальных a $n(a)=0$.

Пусть, наконец, $A(a)$ — высказывание, зависящее от a. Обозначим через $B(n)$ высказывание «$A(a)$ верно для всех чисел a, для которых $n(a) = n$». Тогда, как легко видеть, формулировка принципа индукции в новой форме для утверждений $A(a)$ совпадает с формулировкой этого принципа в старой форме для утверждений $B(n)$.

12. Каковы бы ни были четные числа a и b, существуют такие четные числа q и r, что

$$a = bq + r \qquad (0 \leqq r < 2b).$$

Такие числа q и r единственны.

Доказательство. Разделим a на $2b$ с остатком обычным образом:

$$a = 2bq + r \qquad (0 \leqq r < 2b). \tag{3}$$

При этом числа q и r определяются однозначно. Из четности a и $2bq$ следует четность их разности, т. е. числа r. Нам остается, положив $2q = q'$, переписать (3) в виде

$$a = q'b + r \qquad (0 \leqq r < 2b)$$

и заметить, что оба числа q' и r четные и определяются единственным образом.

13. Пусть p — наименьший простой делитель числа a. Отсюда следует, что $a = pb$. Всякий простой делитель q числа b является вместе с тем и делителем a. Поэтому $q \geq p$, значит, и $b \geq p$, так что

$a \geq p^2$ и, наконец, $p \leqq \sqrt{a}$.

14. Пусть $p_1, p_2, \ldots, p_k$ — полный список всех простых чисел, входящих хотя бы в одно из канонических разложений a и b. Положим

$$a = p_1^{\alpha_1} p_2^{\alpha_2} \ldots p_k^{\alpha_k},$$

$$b = p_1^{\beta_1} p_2^{\beta_2} \ldots p_k^{\beta_k}.$$

(Если a не делится на p_i, то $\alpha_i = 0$; если b не делится на p_i, то $\beta_i = 0$.) Пусть γ_i — наибольшее из чисел α_i и β_i для $i = 1, 2, ..., k$, а δ_i — наименьшее из них.

Тогда, на основании теоремы 17, наибольший общий делитель a и b есть $p_1^{\delta_1} p_2^{\delta_2} \dots p_k^{\delta_k}$, а их наименьшее общее кратное — $p_1^{\gamma_1} p_2^{\gamma_2} \dots p_k^{\gamma_k}$.

15. Как следует из теоремы 17, каждый делитель числа a с каноническим разложением $p_1^{\alpha_1} p_2^{\alpha_2} \dots p_k^{\alpha_k}$ должен иметь вид $p_1^{\beta_1}, \dots, p_k^{\beta_k}$, где β_1 принимает $\alpha_1 + 1$ значений: 0, 1, 2, ..., α_1, β_2 принимает $\alpha_2 + 1$ значений и т. д. Так как любые комбинации этих значений возможны и дают нам все делители a, причем; каждый по одному разу (если бы какой-нибудь делитель повторился несколько раз, то это означало бы наличие у него нескольких канонических разложений), число делителей a равно

$$(\alpha_1 + 1)(\alpha_2 + 1) \dots (\alpha_k + 1). \qquad (4)$$

16. Пусть каноническое разложение a есть $p_1^{\alpha_1} p_2^{\alpha_2} \dots p_k^{\alpha_k}$. Очевидно, можно положить $p_1 = 2$, $\alpha_1 \geqq 2$ и $p_2 = 3$, $\alpha_2 \geqq 1$.

Далее, согласно (4), мы имеем

$$(\alpha_1 + 1)(\alpha_2 + 1) \dots (\alpha_k + 1) = 14.$$

Отсюда $k = 2$, $\alpha_1 + 1 = 7$ и $\alpha_2 + 1 = 2$. Таким образом, $\alpha = 2^6 \cdot 3 = 192$.

17. Мы имеем

$$\tau(a^2) = \tau\left(p_1^{2\alpha_1} p_2^{2\alpha_2}\right) = (2\alpha_1 + 1)(2\alpha_2 + 1) = 81,$$

так что $(2\alpha_1 + 1)(2\alpha_2 + 1)$ есть разложение числа 81 на два множителя. Так как нумерация простых делителей a зависит от нас, ограничимся рассмотрением следующих возможностей:

$$2\alpha_1 + 1 = 1, \quad 2\alpha_2 + 1 = 81;$$
$$2\alpha_1 + 1 = 3, \quad 2\alpha_2 + 1 = 27;$$
$$2\alpha_1 + 1 = 9, \quad 2\alpha_2 + 1 = 9.$$

В первом из этих случаев $\alpha_1 = 0$, что противоречит предположенной положительности числа α_1. Оставшиеся случаи дают нам

$$\alpha_1 = 1, \quad \alpha_2 = 13;$$
$$\alpha_1 = 4, \quad \alpha_2 = 4.$$

Значит, либо

$$\tau(a^3) = \tau\left(p_1^{3\alpha_1} p_2^{3\alpha_1}\right) = \tau\left(p_1^3 p_2^{3\alpha}\right) = (3 + 1)(39 + 1) = 160,$$

либо

$$\tau(a^3) = \tau\left(p_1^{3\alpha_1} p_2^{3\alpha_\cdot}\right) = \tau\left(p_1^{12} p_2^{12}\right) = 13 \cdot 13 = 169.$$

18. Пусть $p_1^{\alpha_1} p_2^{\alpha_2} \dots p_k^{\alpha_k}$ — каноническое разложение числа a.
Условие задачи дает нам

$$p_1^{\alpha_1} p_2^{\alpha_2} \dots p_k^{\alpha_k} = 2(\alpha_1 + 1)(\alpha_2 + 1) \dots (\alpha_k + 1),$$

или

$$\frac{p_1^{\alpha_1}}{\alpha_1 + 1} \frac{p_2^{\alpha_2}}{\alpha_2 + 1} \dots \frac{p_k^{\alpha_k}}{\alpha_k + 1} = 2. \qquad (5)$$

Заметим, что

$$\frac{2^1}{1+1} = 1 < \frac{2^2}{2+1} = \frac{4}{3} < \frac{2^3}{3+1} = 2 < \frac{2^\alpha}{\alpha+1} \quad (\alpha \geqq 4),$$

$$1 < \frac{3^1}{1+1} < 2 < \frac{3^\alpha}{\alpha+1} \quad (\alpha \geqq 2),$$

$$2 < \frac{p^\alpha}{\alpha+1} \quad (p \geqq 5,\ \alpha \geqq 1).$$

Поэтому в (5) слева каждая дробь не меньше единицы и, следовательно, ни одна из дробей не может быть больше, чем 2. Значит, в левой части (5) могут стоять лишь дроби из следующего набора:

$$\frac{2^1}{1+1},\ \frac{2^2}{2+1},\ \frac{2^3}{3+1},\ \frac{3^1}{1+1},$$

причем их произведение есть 2. Но это может быть лишь в двух случаях: когда в (5) слева стоит только одна дробь $\frac{2^3}{3+1}$ или когда там стоят две дроби $\frac{2^2}{2+1}$ и $\frac{3^1}{1+1}$. Этим двум случаям соответствуют два ответа задачи: 8 и 12.

19. Напишем каноническое разложение числа a:

$$a = p_1^{\alpha_1} \dots p_k^{\alpha_k}.$$

Тогда

$$a^2 = p_1^{2\alpha_1} \dots p_k^{2\alpha_k},$$

и согласно (4) (задача 15)

$$\frac{\tau(a^2)}{\tau(a)} = \frac{(2\alpha_1 + 1) \dots (2\alpha_k + 1)}{(\alpha_1 + 1) \dots (\alpha_k + 1)}.$$

Легко видеть, что каждая дробь $(2\alpha_i + 1)/(\alpha_i + 1)$ с ростом α_i возрастает (приближаясь к 2), так что наименьшее значение этой дроби будет достигаться при $\alpha_i = 1$ и будет равно 3/2. Это значит, что

$$\frac{\tau(a^2)}{\tau(a)} \geqq \left(\frac{3}{2}\right)^k.$$

Ясно, что при достаточно большом k будет $(3/2)^k > K$. Для этого достаточно взять

$$k > \frac{\log K}{\log {}^3/_2}.$$

Например, при $K = 100$ достаточно взять $k > 2/0{,}18 = 11{,}1$; так как число k должно быть целым, можно взять $k = 12$.

20. Аналоги теорем 11—14 для четной делимости неверны. В самом деле, числа 30 и 42 четно простые. Их наименьшее четное кратное есть 420, а произведение — 1260.

Далее, $60 = 6 \cdot 10$ четно делится на четно простое число 30; 6 и 30 четно взаимно просты, а 10 четно на 30 не делится.

Наконец, $60 = 6 \cdot 10 = 30 \cdot 2$ — два различных разложения числа 60 на четно простые множители.

21. а) 116 при делении на 8 равноостаточно с 4, а 17 — с 1. Значит, A равноостаточно с $5^{21} = (5^2)^{10} \cdot 5$. Но $5^2 = 25$ при делении на 8 равноостаточно с единицей. Следовательно, A при делении на 8 дает в остатке 5.

б) 14 при делении на 17 равноостаточно с —3. Поэтому A равноостаточно с $(-3)^{256} = 3^{256} = (3^3)^{85} \cdot 3$. Но 3^3 мы можем заменить на 10: $10^{85} \cdot 3 = (10^2)^{42} \cdot 30$. Далее, 10^2 при делении на 17 равноостаточно с числом —2, а 2^4 с —1. Значит, A равноостаточно с $(-2)^{42} \cdot 30 = 2^{42} \cdot 30 = (2^4)^{10} \cdot 4 \cdot 30 = (-1)^{10} \cdot 4 \cdot 30 = 120$. Последнее же число при делении на 17 дает в остатке 1.

22. а) Пусть n_1 — остаток от деления n на 6. Тогда n_1 может принимать значения 0, 1, 2, 3, 4 и 5, а $n^3_1 + 11n_1$ при делении на 6 равноостаточно с $n^3 + 11n$. Значит, нам следует испытывать делимость на 6 чисел 0, 12, 30, 60, 108 и 180. Но все эти числа на 6 делятся.

Для получения того же результата можно воспользоваться и более частными соображениями. Число $n^3 + 11n$ равноостаточно при делении на 6 с числом $n^3 + 11n - 12n = n^3 - n = (n-1)n(n+1)$. Но из трех последовательных целых чисел $n-1$, n и $n+1$ хотя бы одно — четное (т. е. делится на 2) и ровно одно делится на 3. Значит (согласно следствию теоремы 11), произведение этих трех чисел делится на 6. Кстати, можно заметить, что

$$\frac{1}{6}(n-1)n(n+1) = C^3_{n+1}.$$

б) При $n \geq 2$ мы имеем (пользуясь формулой бинома)

$$4^n + 15n - 1 = (3+1)^n + 15n - 1 =$$
$$= 3^n - 3^{n-1}C_n^1 + \ldots + 3^2 C_n^{n-2} + 3C_n^{n-1} + 1 + 15n - 1 =$$
$$= 9\left(3^{n-2} + 3^{n-3}C_n^1 + \ldots + C_n^{n-2}\right) + 18n,$$

и оба слагаемых, очевидно, делятся на 9.

При n=1 наше выражение равно 4^1 +15•1— 1 =18.

в) Доказательство ведется по индукции. При $n = 0$

$$10^{3^0} - 1 = 10^1 - 1 = 9 \quad \text{и} \quad 3^{0+2} = 9.$$

Пусть теперь делимость

$$(10^{3^n} - 1) \vdots 3^{n+2}$$

имеет место. Тогда

$$10^{3^{n+1}} - 1 = (10^{3^n})^3 - 1^3 = (10^{3^n} - 1)(10^{2\cdot 3^n} + 10^{3^n} + 1).$$

Первый сомножитель справа делится на 3^{n+2} по индуктивному предположению. Во втором же сомножителе мы можем заменить десятки на равноостаточные им при делении на 3 единицы; полученное число 3 показывает, что второй сомножитель делится на 3. Следовательно, все произведение делится на $3^{n+3} = 3^{(n+1)+2}$, что и требовалось.

г) При делении на a^2— a+ 1, очевидно, a^2 равноостаточно с a—1. Значит, $a^{2n+1} + (a-1)^{n+2}$ " равноостаточно с

$$a^{2n+1} + (a^2)^{n+2} = a^{2n+1} + a^{2n+4} = a^{2n+1}(1 + a^3) =$$
$$= a^{2n+1}(1 + a)(1 - a + a^2),$$

чтои требовалось.

д) $(n^k - 1) = (n - 1)(n^{k-1} + n^{k-2} + \ldots + n + 1)$.

е) $(n^{2l+1} + 1) = (n + 1)(n^{2l} - n^{2l-1} + \ldots - n + 1)$.

23. Пусть ~ — эквивалентное отношение на множестве чисел. Возьмем произвольное число a и рассмотрим все числа, эквивалентные a. Все они ввиду транзитивности отношения ~ эквивалентны между собой. Обозначим через K класс всех этих чисел.

Рассмотрим теперь произвольное число b, не принадлежащее K. Если бы было $b \sim c$, где с — некоторое число из K, то было бы и $b \sim a$, чего, однако, не может быть по выбору b. Значит, ни одно из чисел, лежащих вне K, не эквивалентно ни одному из чисел K. Следовательно, K есть класс эквивалентности, содержащий a.

Так как число a было нами взято совершенно произвольно, проведенные рассуждения показывают, что каждое число принадлежит некоторому классу эквивалентности. Это и требовалось.

24. Очевидно, среди чисел 0, 1, .,., *m* найдутся два, принадлежащие одному классу. Пусть этими числами будуг *k* и *l*: *k* ~ *l*. Таких пар чисел из одного класса может оказаться, вообще говоря, и несколько. Выберем ту из них, для которой величина |*k* — *l*| будет наибольшей. Поскольку —*l* ~ —*l*, мы, по условию, получаем

$$k - l \sim l - l = 0.$$

Далее, мы находим, что и при любом целом *n*

$$n(k - l) \sim 0.$$

Наконец, при любом *r*

$$n(k - l) + r \sim r,$$

т.е. из *a* = *b(mod k* — *l)* следует a ~ *b*. Таким образом, классы отношения ~ содержат целиком классы вычетов по модулю *m*.

Для того чтобы классов ~ -эквивалентности было *m,* необходимо, чтобы каждый класс ~ -эквивалентности содержал не более одного класса вычетов и чтобы *k* — *l* = *m*.

25 а) Обе части сравнения и модуль можно разделить на одно и то же число (разумеется, отличное от нуля).

В самом деле,

$$ad \equiv bd \pmod{md}$$

означает, что

$$ad - bd = (a - b)\,d \vdots md,$$

т. е. (*a* — *b*) ⋮ *m,* откуда *a* = *b* (mod *m*).

б) Обе части сравнения можно разделить на число, взаимно простое с модулем.

Действительно, если *d* и *m* взаимно просты, то из

$$ad \equiv bd \pmod{m},$$

т.е. из *(a* — *b)d* ⋮ *m,* следует на основании теоремы 12, что (*a* — *b)* ⋮ *m,* что и требовалось.

26. Предположим, что

$$1 \leqq k < l \leqq p - 1, \quad ka \equiv la \pmod{p}.$$

Это значит, что *(l* — *k)a* ⋮ *p*. Поскольку *a* не делится на *p*, должно быть (*l* — *k)* ⋮ *p*. Но и этого не может быть, так как 0 < *l* — *k* < *p*.

27. Необходимость. Пусть число *p* простое. Возьмем 0 < *q* < *p*. Среди чисел $q, 2q, \ldots, (p-1)q$ найдется ровно одно, дающее при делении на *p* в остатке единицу. Пусть этим числом будет $\bar{q}q$:

$$\bar{q}q \equiv 1 \pmod{p}. \qquad (6)$$

С другой стороны, среди чисел $\bar{q}, 2\bar{q}, \ldots, (p-1)\bar{q}$ также может быть лишь одно, дающее при делении на p в остатке единицу. Это, как уже установлено, число $q\bar{q}$.

Выясним, в каких случаях $q = \bar{q}$. Во всех таких случаях сравнение (6) переписывается так:

$$q^2 \equiv 1 \pmod{p},$$

или, что то же самое,

$$q^2 - 1 \equiv 0 \pmod{p}.$$

Это значит, что

$$q^2 - 1 = (q+1)(q-1) \vdots p.$$

Ввиду того, что число p простое, по теореме 13 должно быть либо $(q+1) \not\vdots p$, либо $(q-1) \not\vdots p$. Так как число q заключено между нулем и p, первый из этих случаев возможен лишь при $q = p-1$, а второй — при $q = 1$. Таким образом, при $p = 2$ и $p = 3$ всегда $q = \bar{q}$, при $p \geq 5$ — лишь в случаях $q = 1$ и $q = p - 1$.

Следовательно, при $p \geq 5$ все оставшиеся числа 2, ..., $p-2$ можно объединить в такие $(p-3)/2$ пары, что произведение чисел, составляющих каждую из пар, при делении на p дает в остатке 1. Выпишем сравнения вида (6) для всех таких пар, добавив в этот список сравнение

$$p - 1 \equiv p - 1 \pmod{p},$$

и перемножим все $(p-1)/2$ полученных сравнений почленно.

В результате такого умножения мы получим слева произведение всех чисел от 2 до $p-1$, а справа $p-1$:

$$2 \cdot 3 \ldots (p-1) \equiv p - 1 \pmod{p},$$

или

$$1 \cdot 2 \cdot 3 \ldots (p-1) + 1 \equiv 0 \pmod{p}.$$

Последнее сравнение означает, что

$$(1 \cdot 2 \ldots (p-1) + 1) \vdots p,$$

а это и требовалось.

Остается проверить случаи $p = 2$ и $p = 3$. Но для них, очевидно, $(1+1) \not\vdots 2$ и $(2+1) \not\vdots 3$.

Достаточность. Если число p не простое, то оно может быть разложено в произведение двух меньших множителей: $p = p_1 p_2$.

Если $p_1 \neq p_2$, то и p_1 и p_2 входят сомножителями в произведение $1 \cdot 2 \ldots (p-1)$, которые тем самым делятся на $p_1 p_2$, т.е. на p. Пусть теперь $p_1 = p_2 = q$. Тогда $p = q^2$ (т.е. p есть квадрат простою числа). Если $q > 2$, то $p > 2q$, и в произведение $1 \cdot 2 \ldots (p-1)$ входят множителями q и $2q$,

так что в этом случае оно делится на q^2, т. е. на p. В обоих случаях $1 \cdot 2 \ldots (p-1)+1$ на p делиться не может. Наконец, если $p = 4$, то $1 \cdot 2 \cdot 3 - 1 = 5$, и на 4 не делится.

28. **Теорема**. *Пусть* $m = p_1^{\alpha_1} p_2^{\alpha_2} \ldots p_k^{\alpha_k}$ *— каноническое разложение m. Тогда для того чтобы числа A и B были равноостаточными при делении на m, необходимо и достаточно, чтобы они были равноостаточными при делении на* $p_1^{\alpha_1}$, *на* $p_2^{\alpha_2}$, ..., *на* $p_k^{\alpha_k}$.

Доказательство. *Необходимость*. Равноостаточность A и B при делении на m означает $(A-B) \vdots m$. Тем более $(A-B) \vdots p_i^{\alpha_i}$ $(i = 1, \ldots, k)$, и числа A и B оказываются равноостаточными при делении на все $p_i^{\alpha_i}$.

Достаточность. Пусть числа A и B при делении на каждое $p_i^{\alpha_i}$ равноостаточны. Обозначим через r_i остаток от деления A и B на $p_i^{\alpha_i}$ $(i = 1, 2, \ldots, k)$. Это значит, что

$$A \equiv r_i \left(\mathrm{mod}\, p_i^{\alpha_i}\right). \tag{7}$$

Положим, далее,

$$\frac{m}{p_i^{\alpha_i}} = m_i, \quad i = 1, \ldots, k,$$

и умножим в сравнении (7) обе части и модуль на m_i:

$$Am_i \equiv m_i r_i \,(\mathrm{mod}\, m). \tag{8}$$

Сложив все такие сравнения почленно, мы получим

$$A(m_1 + m_2 + \ldots + m_k) \equiv m_1 r_1 + m_2 r_2 + \ldots + m_k r_k \,(\mathrm{mod}\, m).$$

Ввиду равноостаточности A и B при деления на $p_1^{\alpha_1}$, $p_2^{\alpha_2}$, ..., $p_k^{\alpha_k}$ мы получаем также

$$B(m_1 + m_2 + \ldots + m_k) \equiv m_1 r_1 + m_2 r_2 + \ldots + m_k r_k \,(\mathrm{mod}\, m). \tag{9}$$

Вычтя почленно сравнение (9) из (8), мы имеем

$$(A-B)(m_1 + m_2 + \ldots + m_k) \equiv 0 \,(\mathrm{mod}\, m),$$

т.е. $(A-B)(m_1 + m_2 + \ldots + m_k) \vdots m$.

Но сумма $m_1 + m_2 + \ldots + m_k$ взаимно проста *с m. B* самом деле, если бы она имела с m некоторый общий простой делитель p, то он входил бы в каноническое разложение m, т. е. имел бы вид p_i. Но тогда на него делилась бы как вся сумма, так и каждое слагаемое, кроме одного, m_i, а этого быть не может.

Теперь мы можем применить теорему 12, которая даст, что $(A—B) \equiv m$, т. е. числа A и B при делении на m равноостаточны.

29. а) Выпишем систематически все равенства, описывающие деления с остатком, которые составляют применение алгоритма Евклида к числам a и b:

$$\left.\begin{aligned} a &= bq_0 + r_1, \\ b &= r_1q_1 + r_2, \\ r_1 &= r_2q_2 + r_3, \\ &\cdots\cdots\cdots \\ r_{n-2} &= r_{n-1}q_{n-1} + r_n, \\ r_{n-1} &= r_nq_n. \end{aligned}\right\} \qquad (10)$$

Мы имеем $r_{n-1} \vdots r_n$. Вместе с $r_{n-2} = r_{n-1}q_{n-1} + r_n$, это дает нам $r_{n-1} \equiv r_n$. Продвигаясь аналогичным образом вверх по системе равенств (10), мы получаем, наконец, что $a \vdots r_n$ и $b \vdots r_n$. Значит, r_n есть общий делитель a и b.

Пусть d — любой общий делитель a и b. Вместе с $a = bq_0 + r_1$ это дает нам $r_1 \equiv d$. Продвигаясь по системе равенств (10) вниз, мы будем последовательно получать, что $r_2 \vdots d$, $r_3 \vdots d$, ..., $r_n \vdots d$. Значит, r_n делится на любой общий делитель a и b, являясь тем самым наибольшим общим делителем этих чисел.

б) Доказательство ведется по индукции. Полагая $A_0=0$, $B_0= 1$, $A_1 = 1$, $B_1 = —q_0$, мы имеем $r_0 = b = aA_0 + bB_0$ и

$$r_1 = aA_1 + bB_1.$$

Пусть теперь

$$r_{k-1} = A_{k-1}a + B_{k-1}b,$$
$$r_k = A_ka + B_kb.$$

Но тогда

$$r_{k+1} = r_{k-1} - r_kq_{k+1} = (A_{k-1} - q_{k+1}A_k)\,a + (B_{k-1} - q_{k+1}B_k)\,b,$$

и нам остается положить

$$A_{k-1} - q_{k+1}A_k = A_{k+1},$$
$$B_{k-1} - q_{k+1}B_k = B_{k+1}.$$

Числа A_n и B_n окажутся искомыми A и B.

30. Если b и c взаимно просты, то по предыдущему можно найти такие целые B и C, что

$$bB + cC = 1,$$

или, после умножения на a,

$$abB + acC = a;$$

$ab \equiv c$ по условию; $ac \equiv c$ очевидным образом; значит, и $a \equiv c$.

31. Ограничимся рассмотрением признака равноостаточности при делении на 8.

Пусть произвольное натуральное A представлено в виде $1000a+ b$, где $0 \le b < 1000$ (т. е. b — трехзначноечисло, которым оканчивается A), и

$$f(A)=\begin{cases} b, & \text{если } A \geqq 1000, \\ \text{остатку от деления } A \text{ на } 8, & \text{если } 8 \leqq A < 1000, \\ \text{не определено}, & \text{если } A < 8. \end{cases}$$

32. Ограничимся рассмотрением признака равноостаточности при делении на 18 в двенадцатеричной системе счисления.

Пусть A представлено в виде $144a+ b$, где $0 \le b < 144$ (т. е. b — двузначное в двенадцатеричной системе счисления число, которым оканчивается записанное в этой системе число A), и

$$f(A)=\begin{cases} b, & \text{если } A \geqq 144, \\ \text{остатку от деления } A \text{ на } 18, & \text{если } 18 \leqq A < 144, \\ \text{не определено}, & \text{если } A < 18. \end{cases}$$

Проверка того, что процесс построения последовательности A, $f(A)$, $f(f(A))$, ... действительно является признаком равноостаточности, осуществляется стандартно.

33. Для тех m, у которых каноническое разложение имеет вид $2^{\alpha}\cdot 5^{\beta}$.

34. Условия а) и б) выполняются автоматически. Поскольку при делении на 3 числа 10 и 1 равноостаточны, равноостаточными же должны быть и числа A и $f(A)$. Наконец, то, что $f(A)<A$ при $A \ge 3$, устанавливается простым подсчетом.

35. а) $f(858\ 773) = 38$; $f(38)= 11$; $f(11)= 2$.

б) $f(A) = 4444\cdot 4= 17\ 776$; $f(17\ 776) = 28$; $f(28) = 10$; $f(10)= 1$.

36. Признак равноостаточности при делении на 9 аналогичен рассмотренному признаку равноостаточности при делении на 3.

Для получения признака равноостаточности при делении на 11 представим число A в виде

$$10^{2n}a_n + 10^{2n-2}a_{n-1} + \ldots + 10^2a_1 + a_0,$$

где $0 \le a_i < 100$. Очевидно, такое представление соответствует разбиению числа на двузначные «грани» (справа налево). Пусть

$$f(A)=$$
$$=\begin{cases} a_0+a_1+\ldots+a_n, & \text{если } A\geqq 100,\\ \text{остатку от деления } A \text{ на } 11, & \text{если } 11\leqq A<100,\\ \text{не определена,} & \text{если } A<11. \end{cases}$$

Нам остается указать, что числа A и $f(A)$ действительно равноостаточны при делении на 11 и, кроме того, $f(A)< A$.

Другой признак равноостаточности при делении на 11 получается на основе представления числа A в виде

$$A=10^n a_n+10^{n-1}a_{n-1}+\ldots+10a_1+a_0$$

и использования того, что 10 при делении на 11 равноостаточно с —1, а 100 — с 1. Поэтому A равноостаточно с числом $a_0-a_1+a_2-a_3+\ldots\pm a_n$, и формулировка соответствующего признака равноостаточности не составляет труда.

Наконец, можно, разбивая число A на трехзначные «грани», представить его в виде

$$10^{3n}a_n+10^{3n-3}a_{n-1}+\ldots+10^3a_1+a_0$$

($0\leq a_i<1000$). Тогда A при делении на 37 равноостаточно с суммой $a_0+a_1+\ldots+a_n$, а, при делении на 7, 11 и 13 — со знакопеременной суммой $a_0-a_1+a_2-\ldots\pm a_n$.

37. Для примера рассмотрим признак равноостаточности на 8 в троичной системе счисления. Представим для этого произвольное A в виде

$$a_n3^{2n}+a_{n-1}3^{2(n-1)}+\ldots+a_13^2+a_0, \quad \text{где} \quad 0\leqq a_i<9.$$

Здесь a_i суть двузначные грани, на которые разбивается число A, считая справа налево. Нам остается положить

$$f(A)=\begin{cases} a_0+a_1+\ldots+a_n, & \text{если } A\geqq 9,\\ 0, & \text{если } A=8,\\ \text{не определено,} & \text{если } A<8, \end{cases}$$

и провести стандартные рассуждения.

38. В шестеричной системе счисления: 5 (k = 1), 7 (k = 2), 43 (k = 3);

В семеричной системе счисления: 2, 3, 6 (k =1), 4, 6, 12, 16, 24 (k = 2), 171 (k = 3);

В девятеричной системе счисления: 2, 4, 8 (k = 1), 5, 10, 20, 40 (k = 2), 7, 13, 14, 26 и т. д. (k = 3);

В тринадцатеричной системе счисления: 2, 3, 4, 6 (k = 1), 7, 14, 21 и т. д. (k = 2).

39. В троичной системе счисления: 2, 4 (k =l), 8, 12, 24 (k =2), 13, 26 (k = 3), 41 (k = 4);
В пятеричной системе счисления: 2, 3, 6 (k=1), 8, 12, 24 (k =2), 31 (k = 3);
В восьмеричной системе счисления 3, 9 (k =l), 5, 13 (k = 2);
В десятичной системе счисления: 11 (k =l), 101 (k = 2), 7, 11, 13 (k = 3).

40. Если числа a и b равноостаточны, то $(a-b) \vdots m$. Поэтому в силу теоремы 6 числа a и b делятся или не делятся на m одновременно.

Числа 4 и 5 равноделимы, но не равноостаточны при делении на 3.

41. Пусть из равноделимости на m следует равноостаточность при делении на m. Это значит, что все не делящиеся на m числа имеют при делении на m один и тот же остаток. Значит, этот остаток должен быть равен единице, так что $m = 2$.

42. Отношение равноделимости на m, очевидно, является рефлексивным (всякое число равноделимо на m с самим собой), симметричным (если a равноделимо с b, то и b равноделимо с a) и транзитивным (если a равноделимо с b, а b равноделимо с c, то и a равноделимо с c).

Следовательно, это и есть отношение эквивалентности. При этом в один класс попадают все числа, делящиеся на m, а в другой — все не делящиеся на m.

43. Нетрудно проверить, что при $m > 2$ равноделимость сумм не следует из равноделимости слагаемых.

Для того чтобы равноделимость произведений вытекала из равноделимости их сомножителей, необходимо и достаточно, чтобы число m было простым.

В самом деле, если одно из произведений делится на простое p, то по теореме 13 на это p должен делиться хотя бы один из сомножителей этого произведения. Но тогда на p делится равноделимый ему сомножитель другого произведения, а потому и все произведение. Если же одно произведение на p не делится, то и другое на p делиться не может (ибо в противном случае, на основании только что установленного, на p делилось бы и первое произведение).

Наоборот, если число p составное, то произведения равноделимых сомножителей могут уже равноделимыми не быть. Достаточно положить $p = p_1p_2$ ($p_1 \neq 1$, $p_2 \neq 1$). Тогда числа 1 и p_1 а также числа 1 и p_2 равноделимы на p, а произведения $1 \cdot 1$ и $p_1 \cdot p_2$, очевидно, нет.

44. Непосредственное следствие задачи 36.

45. Выполнение условий а) и б) очевидно.

Если, далее, $a - 2b \geq 0$, то, очевидно, $f(A) < A$. Если же $a - 2b < 0$, то это неравенство может и нарушиться. При этом наибольшее значение модуля $|a - 2b|$ достигается при $a = 0$ и $b = 9$ и равно 18. Следовательно, при $A \geq 19$ должно быть $f(A) < A$. Справедливость этого неравенства при меньших значениях обеспечивается определением функции f.

Наконец, $10a + b$ равноделимо на 7 с $50a + 5b$ (ибо числа 5 и 7 взаимно простые) и тем самым с $50a + 5b - 7(7a + b) = a - 2b$.

46. Число 15 при делении на 7 дает в остатке 1, а $1 - 2 \cdot 5 = -9$ дает в остатке 5.

47. Условие в) $f(A) < A$ означает $a + 4b < 10a + b$, т. е. $3b < 9a$. Поэтому при $a \geq 4$ нужное условие выполняется.

Условие г) очевидно, $10a + b$ при делении на 13 равноделимо с $40a + 4b$, а последнее число равноостаточно с $a + 4b$.

48. Признак делимости утратит результативность, так как $f(39) = 39$.

49. Пусть нам нужно построить признак делимости на некоторое m. Постараемся подобрать такое s, взаимно простое с m и по возможности небольшое, что $(10s + 1) \vdots m$ (так было в случае $m = 7$; s оказалось равным 3) или же $(10s - 1) \vdots m$ (например, при $m = 13$, $s = 4$).

В первом из этих случаев $A = 10a + b$ равноделимо на m с

$$10as + bs = (10s + 1)a - a + bs,$$

т. е. с $a - bs$, а во втором — с

$$(10s - 1)a + a + bs,$$

т. е. с $a + bs$.

В связи со сказанным число $10a + b$

при делении на	17	равноделимо с	$a - 5b$,
»	19	»	$a + 2b$,
»	23	»	$a + 7b$,
»	29	»	$a - 3b$,
»	31	»	$a + 3b$.

Завершение точных формулировок этих признаков делимости предоставляется читателю.

50. а) Так как 100 при делении на 49 равноостаточно с 2, всякое число вида

$$10^{2n}a_n + 10^{2n-2}a_{n-1} + \ldots + 10^2a_1 + a_0 \qquad (0 \leqq a_i < 100)$$

при делении на 49 равноостаточно с

$$2^n a_n + 2^{n-1}a_{n-1} + \ldots + 2a_1 + a_0.$$

б) $10a+b$ при делении на 49 равноделимо с $a+5b$.

51. Очевидно, при $A \geq 6$ должно быть $f(A)<A$.

52. а) В семеричной системе счисления представление A в виде $7a+b$ дает, что при делении на 5 число A равноделимо с $a+3b$;

б) В одиннадцатиричной системе счисления представление A в виде $11a+b$ дает, что при делении на 7 число A равноделимо с $a+2b$;

в) В двенадцатиричной системе счисления, представляя A в виде $12a+b$, получаем, что при делении на 17 число равноделимо с $a-7b$.

53. Условия а) и б) выполняются автоматически. Условия в) и г) соблюдаются потому, что переход от A к $F(A)$ сводится к замене некоторых чисел на их остатки при делении на A (которые меньше самих чисел и равноостаточны с ними).

54. а) $r_2 = r_3 = \ldots = r_n = 0$, т. е. $r_k = 0$ $(k \geqq 2)$;

б) $r_3 = r_4 = \ldots = r_n = 0$, т. е. $r_k = 0$ $(k \geqq 3)$;

в) $r_1 = r_2 = \ldots = r_n = 1$, т. е. $r_k = 1$;

г) $r_1 = r_3 = \ldots = r_{2t-1} = -1$, $r_2 = r_4 = \ldots = r_{2t} = 1$, т. е. $r_k = (-1)^k$;

д) $r_{6t+1} = 3$, $r_{6t+2} = 2$, $r_{6t+3} = 6$, $r_{6t+4} = 4$, $r_{6t+5} = 5$, $r_{6t} = 1$.

55. Предоставляется читателю.

56. Возьмем произвольное m и положим

r_1 равным остатку от деления t на m,
r_2 » » » tr_1 на m

и т. д. Тогда число

$$a_n t^n + a_{n-1} t^{n-1} + \ldots + a_1 t + a_0$$

при делении на m равноостаточно с числом

$$a_n r_n + a_{n-1} r_{n-1} + \ldots + a_1 r_1 + a_0.$$

После этого построение требуемого признака не составляет труда.

57. Предоставляется читателю.

58. $10^2 = 7 \cdot 14 + 2$, так что $r = 2$, и мы имеем

$$A_0 = 1\,048\,576, \quad A_1 = 1 \cdot 2^3 + 4 \cdot 2^2 + 85 \cdot 2 + 76 = 270,$$
$$A_2 = 2 \cdot 2 + 70 = 74, \quad A_3 = 4.$$

59. Если t равноостаточно с $r=r_1$ при делении на m, то

tr_1 » $r^2=r_2$ » m,
tr_2 » $r^3=r_3$ » m,

и т. д.

60. Ни $2^4 - 2$, ни $2^3 - 1$ не делятся на 4.

61. Если $a \not\equiv p$, то $a^p \not\equiv p$, и теорема доказана. Если же a не делится на p, то a взаимно просто с p, и мы можем приведенное в условии теоремы сравнение сократить:

$$a^{p-1} \equiv 1 \pmod{p}.$$

Для доказательства последнего сравнения разделим каждое из чисел вида ta (t=l, 2, ..., p—1) на p с остатком:

$$ta = q_t p + r_t.$$

Это можно переписать так:

$$\left.\begin{aligned} a &\equiv r_1 \pmod{p}, \\ 2a &\equiv r_2 \pmod{p}, \\ &\cdots\cdots\cdots \\ (p-1)\,a &\equiv r_{p-1} \pmod{p}. \end{aligned}\right\} \qquad (11)$$

Из результата задачи 26 следует, что среди чисел r_t ровно по одному разу встретится каждое из чисел 1, 2, ..., p—1. Перемножая все сравнения, мы получаем

$$1 \cdot 2 \ldots (p-1)\,a^{p-1} \equiv 1 \cdot 2 \ldots (p-1) \pmod{p}.$$

Нам остается это сравнение сократить на $1 \cdot 2 \ldots (p\text{-}1)$.

62.
$$\varphi(12) = \varphi(2^2 \cdot 3) = 2^{2-1}(3-1) = 2 \cdot 2 = 4,$$
$$\varphi(120) = \varphi(2^3 \cdot 3 \cdot 5) = 2^{3-1}(3-1)(5-1) = 4 \cdot 2 \cdot 4 = 32,$$
$$\varphi(1000) = \varphi(2^3 \cdot 5^3) = 2^{3-1}5^{3-1}(5-1) = 4 \cdot 25 \cdot 4 = 400.$$

63. Будем искать m в виде $p_1^{\alpha_1} p_2^{\alpha_2} \ldots p_k^{\alpha_k}$. Тогда

$$\text{а)}\ p_1^{\alpha_1 - 1}(p_1 - 1)\,p_2^{\alpha_2 - 1}(p_2 - 1) \ldots p_k^{\alpha_k - 1}(p_k - 1) \equiv \dot{1}\dot{0}.$$

Стоящее слева произведение должно делиться на 5. Значит, либо одно из чисел $p_1, p_2, \ldots, p_k$ есть 5 (пусть для определенности $p_1 = 5$), либо на 5 делится одна из разностей $p_1 - 1$, p_2—1, …, $p_k - 1$ (пусть в этом случае $(p_1 - 1) \not\equiv 5$). В первом из этих случаев p_1 —1=4, чего не может быть, так как 10 на 4 не делится. Второй случай, поскольку p_1 должно быть простым числом и $10 \not\equiv (p_1 - 1)$, возможен лишь при $p_1 = 11$. Но тогда $\alpha_1 = 1$, и из теоремы 25 следует, что

$$\varphi\left(\frac{m}{11}\right) = 1,$$

т. е. либо $\frac{m}{11} = 1$, либо $\frac{m}{11} = 2$.

В итоге мы имеем $m_1 = 11$, $m_2 = 22$.

б) $p_1^{\alpha_1-1}(p_1-1)\,p_2^{\alpha_2-1}(p_2-1)\ldots p_k^{\alpha_k-1}(p_k-1)=8.$

Если m нечетное, то $\alpha_1=\alpha_2=\ldots=\alpha_k=1$ (ибо правая часть написанного равенства есть степень двойки):

$$(p_1-1)(p_2-1)\ldots(p_k-1)=8.$$

Это возможно лишь при $k=2$, $p_1=3$, $p_2=5$, т. е. при $m=15$.

Пусть теперь число m — четное. Положим для определенности $p_1=2$. Очевидно, по-прежнему $\alpha_2=\ldots=\alpha_k=1$, и мы имеем

$$2^{\alpha-1}(p_2-1)\ldots(p_k-1)=8.$$

Очевидно, $\alpha\leq 4$. Если $\alpha=1$, то случай подобен рассмотренному: написанное неравенство возможно также лишь при $k=3$, $p_2=3$, $p_3=5$, т. е. при $m=30$.

Если $\alpha=2$, то $k=2$, $p_3=5$ и $m=20$.

Если $\alpha=3$, то $k=2$, $p_2=3$ и $m=24$.

Если, наконец, $\alpha=4$, то $k=1$ и $m=16$.

Итак, решения нашей задачи: $m_1=15$, $m_2=30$, $m_3=20$, $m_4=24$, $m_5=16$.

64. Предположим, что

$$p_1^{\alpha_1-1}(p_1-1)\,p_2^{\alpha_2-1}(p_2-1)\ldots p_k^{\alpha_k-1}(p_k-1)=14.$$

Каждое из чисел вида p_i-1 есть либо единица, либо четное число, и потому не может быть семеркой. Будучи на единицу меньше простого числа, оно не может равняться 14. Значит, семеркой является одно из чисел $p_i^{\alpha_i-1}$. Но тогда $p_i-1=6$, а 14 на 6 не делится.

65. Пусть $m=p_1^{\alpha_1}p_2^{\alpha_2}\ldots p_k^{\alpha_k}$. Рассмотрим сначала случай, когда m есть степень простого числа: $m=p^{\alpha}$. Для того чтобы некоторое число было взаимно простым с m, необходимо и достаточно, чтобы оно не делилось на p. Но среди чисел 0, 1, 2, ..., m—1 имеется всего $\frac{m}{p}$ делящихся на p чисел. Следовательно, взаимно простых с p чисел в этом списке имеется столько:

$$m-\frac{m}{p}=m\left(1-\frac{1}{p}\right)=p^{\alpha}\left(1-\frac{1}{p}\right)=p^{\alpha-1}(p-1)=\varphi(m).$$

Заметим теперь, что для взаимной простоты a и m необходимо и достаточно, чтобы с a был взаимно прост остаток от деления a на m.

По только что установленному число остатков от деления на $p_i^{\alpha_i}$, взаимно простых с $p_i^{\alpha_i}$, равно $\varphi\left(p_i^{\alpha_i}\right)$. Но, как было выяснено в процессе решения задачи 28, из равноостаточности чисел при делении

на все $p_i^{\alpha_i}$ следует их равноостаточность при делении на m, и наоборот. Кроме того, для взаимной простоты некоторого числа с m необходимо и достаточно, чтобы оно было взаимно просто с каждым из чисел $p_i^{\alpha_i}$.

Следовательно, каждой комбинации остатков от деления на $p_1^{\alpha_1}$, $p_2^{\alpha_2}$, ..., $p_k^{\alpha_k}$, взаимно простых с соответствующими делителями, соответствует ровно один остаток от деления на m, взаимно простой с m. Нам остается заметить, что число таких комбинаций остатков равно

$$\varphi\left(p_1^{\alpha_1}\right)\varphi\left(p_2^{\alpha_2}\right)\ldots\varphi\left(p_k^{\alpha_k}\right)=\varphi(m).$$

66. Мы имеем

$$a_1 = A + q_1 m_1 \quad \text{и} \quad a_2 = A + q_2 m_2.$$

Поэтому

$$\begin{aligned}(a_1 m_2 + a_2 m_1)(m_1 + m_2)^{\varphi(m_1 m_2)-1} &= \\ = (A(m_1 + m_2) + (q_1 + q_2) m_1 m_2)(m_1 + m_2)^{\varphi(m_1 m_2)-1} &= \\ = A(m_1 + m_2)^{\varphi(m_1 m_2)} + (q_1 + q_2) m_1 m_2 (m_1 + m_2)^{\varphi(m_1 m_2)-1}.&\end{aligned}$$

Здесь по теореме Эйлера первое слагаемое при делении на $m_1 m_2$ равноостаточно с A, а второе делится на $m_1 m_2$. Значит, вся сумма при делении на $m_1 m_2$ равноостаточна с A.

67. Предоставляется читателю.

68. Предоставляется читателю.

69. Предоставляется читателю.

70. $n^{13} - n = n(n^{12} - 1)$. Но

$$n^{12} = n^{\varphi(13)} = n^{2\varphi(7)} = n^{3\varphi(5)} = n^{5\varphi(3)} = n^{12\varphi(2)}.$$

Поэтому либо $n \nmid p$, либо $(n^{12} - 1) \nmid p$ для $p = 2, 3, 5, 7, 13$. Остается сослаться на теорему 16.

71. Предоставляется читателю.

72. Предоставляется читателю.

73. Пусть наибольший общий делитель чисел a и b есть d. Если c не делится на d, то уравнение

$$ax + by = c$$

в целых числах неразрешимо. Если же c делится на d, то обе части уравнения можно сократить на d, и мы подходим к уже рассмотренному случаю.

74. Пусть A и B таковы, что

$$aA + bB = 1.$$

Положим

$$x_t = cA + bt,$$
$$y_t = c\frac{1 - aA}{b} - at.$$

Тогда

$$ax_t + by_t = a(cA + bt) + b\left(c\frac{1 - aA}{b} - at\right) =$$
$$= caA + abt + c(1 - aA) - abt = c,$$

и (x_t, y_t) действительно является решением нашего уравнения.

75. а) $x_t = 9 \cdot 5^5 + 7t = 28\,125 + 7t,$

$$y_t = 9\frac{1 - 5^6}{7} - 5t = -20\,088 - 5t.$$

Поскольку свободные члены и коэффициенты при t в выражениях для x_t и y_t, так сказать, «примерно пропорциональны», мы можем надеяться получить представления наших решений в меньших числах.

В самом деле, мы можем написать:

$$x_t = 6 + 7(t + 4017),$$
$$y_t = -3 - 5(t + 4017),$$

или, полагая

$$t + 4017 = t',$$

мы получаем

$$x_{t'} = 6 + 7t',$$
$$y_{t'} = -3 - 5t'.$$

Заметим, что способ решения уравнений в целых числах, приведенный в задаче 74, позволяет обходиться меньшими числами, хотя и требует несколько более сложных вычислений.

б) Воспользуемся тем, что 25 по модулю 13 принадлежит показателю 2. Мы можем написать:

$$x_t = 8 \cdot 25 + 13 = 200 + 13t,$$
$$y_t = 8\frac{1 - 25^2}{13} - 25t = -384 - 25t,$$

или, после упрощений,

$$x_{t'} = 5 + 13t',$$
$$y_{t'} = -9 - 25t'.$$

76. Условие в) обеспечивается автоматически, а условие г) следует из теоремы 25.

77.

m	17	19	27	29	31	49
k'	12 (или —5)	2	19	3	28 (или —3)	5

78. Предоставляется читателю.

79. Предоставляется читателю.

80. а) $8^{\varphi(21)-1} = 8^{11} = 64^5 \cdot 8$. При делении на 21 это число равноостаточно с 8. Значит, числа $8a + b$ и $a + 8b$ равноделимы на 21.

б) $12^{\varphi(31)-1} = 12^{29} = (12^2)^{14} \cdot 12 = 144^{14} \cdot 12$ при делении на 31 равноостаточно с $11^{14} \cdot 12 = 121^7 \cdot 12 = (-3)^7 \cdot 12 = -(3^3)^2 \cdot 3 \cdot 12 = -(31 - 4)^2(31 + 5)$, что равноостаточно с $-16 \cdot 5 = -80$. Последнее число очевидно равноостаточно с 13. Значит, числа $12a + b$ и $a + 13b$ равноделимы на 31.

Модуль 5
Элементы комбинаторики

Комбинаторика — один из разделов дискретной математики, которая имеет важное значение в связи с использованием ее в теории вероятностей, математической логике, теории чисел, вычислительной технике, кибернетике. Цель этого модуля — ознакомить читателей с основными понятиями комбинаторики и методами решения комбинаторных задач. При изучении комбинаторики мы считали целесообразным систематически использовать понятие множества и операций над множествами, поскольку большинство задач комбинаторики можно сформулировать как задачи теории конечных множеств.

При решении комбинаторных задач нужно особое внимание обратить на метод производящих функций и метод траекторий. Эти методы важны сами по себе, так как находят широкое применение не только в комбинаторике, но и во многих разделах современной математики.

Микромодуль 16
Основные принципы комбинаторики

5.1. Введение

Человеку часто приходится иметь дело с задачами, в которых нужно подсчитать число всех возможных способов расположения некоторых предметов или число всех возможных способов осуществления некоторого действия. Сколькими способами можно расположить 50 человек в очереди в кассу кино? Сколькими способами могут быть распределены золотая, серебряная и бронзовая медали на чемпионате мира по футболу? Задачи такого типа называются *комбинаторными.*

С комбинаторными вычислениями приходится иметь дело представителям многих специальностей: ученому-химику при рассмотрении различных возможных типов связей атомов в молекулах, биологу при изучении различных возможных последовательностей чередования аминокислот в белковых соединениях, конструктору вычислительных машин, агроному, который рассматривает различные возможные способы посевов на нескольких участках, диспетчеру при составлении графика движения. Комбинаторные соображения лежат в основе решения многих задач теории вероятностей — важного раздела современной математики, посвященного изучению случайных явлений. Усиленный интерес к комбинаторике обусловлен бурным развитием кибернетики, вычислительной техники.

Установим сначала важное правило, которое часто применяется при комбинаторных расчетах. Начнем с такой задачи.

З а д а ч а 1. Из Киева до Чернигова можно добраться пароходом, поездом, автобусом, самолетом; из Чернигова до Новгорода-Северского — пароходом и автобусом. Сколькими способами можно осуществить путешествие по маршруту Киев - Чернигов - Новгород-Северский?

Решение. Очевидно, число разных путей из Киева до Новгород-Северскго равно $4 \times 2 = 8$, так как, выбрав один из четырех возможных способов путешествия от Киева до Чернигова, имеем два возможных способа путешествия от Чернигова до Новгорода-Северского (рис. 5.1).

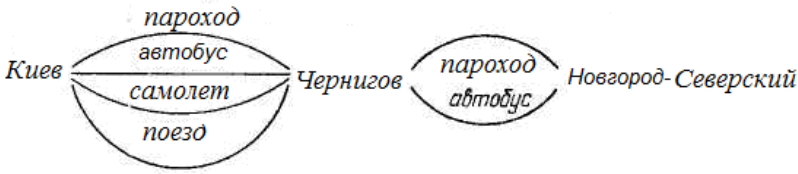


Рис. 5.1

Соображения, которые были приведены при решении задачи 1, доказывают справедливость следующего простого утверждения, которое будем называть *основным правилом комбинаторики.*

Если некоторый выбор A можно осуществить m различными способами, а для каждого из этих способов некоторый другой выбор B можно осуществить n способами, то выбор A и B (в указанном порядке) можно осуществить $m \times n$ способами.

Иначе говоря, если некоторое действие (например, выбор пути от Киева до Чернигова) можно осуществить m различными способами, после чего другое действие (выбор пути от Чернигова до Новгорода-Северского) можно осуществить n способами, то два действия вместе (выбор пути от Киева до Чернигова, выбор пути от Чернигова до Новгорода-Северского) можно осуществить $m \times n$ способами.

Задача 2. В розыгрыше первенства страны по футболу принимает участие 16 команд. Сколькими способами могут быть распределены золотая и серебряная медали?

Решение. Золотую медаль может получить одна из 16 команд. После того как определен владелец золотой медали, серебряную медаль может иметь одна из 15 команд. Следовательно, общее число способов, которыми могут быть распределены золотая и серебряная медали, равно $16 \times 15 = 240$.

Сформулируем теперь основное правило комбинаторики (правило умножения) в общем виде.

Пусть требуется выполнить одно за другим k действий. Если первое действие можно выполнить n_1 способами, другое действие — n_2 способами, третье действие — n_3 способами и так до k-го действия, которое можно выполнить n_k способами, то все k действий вместе могут быть выполнены

$$n_1 \times n_2 \times n_3 \times \ldots \times n_k$$

способами.

5.2. Конечные множества и операции над ними

В этом разделе напомним некоторые положения теории множеств, которые будут использоваться нами при изученные комбинаторики.

Всякая совокупность элементов произвольного рода образует множество. Можно рассматривать множество всех действительных чисел, множество натуральных чисел, множество всех студентов данного университета, множество столов в данной аудитории, множество всех жителей данного города и т.д. Множество считается определенным, если указаны все ее элементы. Эти элементы могут быть указаны с помощью некоторого общего признака или просто с помощью некоторого списка, где обозначены все элементы. Последний способ возможен лишь в том случае, если множество имеет конечное число элементов; такие множества будем называть *конечными. Комбинаторика есть теория конечных множеств.* Поэтому дальше мы будем иметь дело лишь с конечными множествами.

Основной характеристикой конечного множества является число егоэлементов.

Теория конечных множеств изучает правила: как, зная количество элементов некоторых множеств, вычислить количество элементов других множеств, которые составлены из первых с помощью некоторых операций. Операции над множествами мы введем несколько позднее.

Введем основные обозначения, которыми будем пользоваться в дальнейшем. Множества будем обозначать *большими* латинскими буквами, их элементы— *малыми:* $a \in A$*:* a есть элемент A, или a принадлежит A; $a \notin A$: a не есть элемент A, или a не принадлежит A. Количество элементов множества будем обозначать $N(A)$.

5.2.1. Операции над множествами.

Два множеств *равны* между собой, если элементы первого являются элементами второго и, наоборот, элементы второго являются элементами первого.

Если A и B - два множества, то множество C, которому принадлежат все те и только те элементы, которые входят либо в A, либо в B, называется *суммой* или *объединением множеств* A и B и обозначается $C = A \cup B$.

Эта операция «сложения» множеств удовлетворяет *коммутативному* и *ассоциативному* законам:

$$A \cup B = B \cup A, \quad A \cup (B \cup C) = (A \cup B) \cup C. \tag{5.1}$$

Первое из этих равенств вытекает из определения суммы. Второе есть следствие того, что и $A \cup (B \cup C)$ и $(A \cup B) \cup C$ есть совокупность элементов, которые входят хотя бы в одно из множеств A, либо B, либо C. Поэтому можно рассматривать сумму любого числа множеств $A_1 \cup A_2 \cup \ldots \cup A_n$, это будет множество, в которое входят элементы каждого из множеств $A_1, A_2, \ldots, A_n$ и только они (общие элементы считаются только по одному разу).

Однако это «сложение» отличается от обычного сложения. Чтобы объяснить это, рассмотрим числовые множества. Если $x_1, \ldots, x_n$ — некоторые числа, то через $\{x_1, \ldots, x_n\}$ обозначим множество, которое состоит из элементов $x_1, \ldots, x_n$. Предположим, что дано два множества $\{1, 2, 3\}$ и $\{2, 3, 4\}$. Тогда $\{1, 2, 3\} \cup \{2, 3, 4\} = \{1, 2, 3, 4\}$. Если множества A и B имеют общие элементы, то каждый из этих элементов входит в $A \cup B$ только один раз. Итак, число элементов в сумме множеств не обязательно равно сумме чисел элементов первого и второго множества, а может быть меньше ее. В частности, сложениие множеств приводит к такой необычной для чисел формуле:

$$A \cup A = A, \quad A \cup A \cup \ldots \cup A = A.$$

Множество C, которому принадлежат те и только те элементы, которые являются общими для множеств A и B (элементы, которые входят и оба эти множества), называется *пересечением* множеств A и B и обозначается $C = A \cap B$.

Например,

$$\{1,2,3\} \cap \{2,3,4\} = \{2,3\}.$$

Если $A_1, A_2, \ldots, A_n$ — некоторые множества, то $A_1 \cap A_2 \cap \ldots \cap A_n$ является множеством, состоящее из элементов, которые входят в каждое из множеств $A_1, A_2, \ldots, A_n$ (являются общими для этих множеств). Опять-таки пересечение множеств удовлетворяет *коммутативному* и *ассоциативному* законам:

$$A \cap B = B \cap A \quad A \cap (B \cap C) = (A \cap B) \cap C. \tag{5.2}$$

Отметим, что $A \cap A = A$, и потому $A \cap A \ldots \cap \ldots \cap A = A$.

Покажем, что операции объединения и пересечения множеств удовлетворяют также *дистрибутивному* закону:

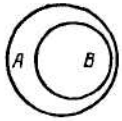
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C). \tag{5.3}$$

Действительно, множество $A \cap (B \cup C)$ содержит элементы, которые входят в множество A и в множество $B \cup C$, т. е. принадлежат A и либо множеству B, либо множеству C. Тогда они принадлежат либо A и B, либо A и C, т.е. либо $A \cap B$, либо $A \cap C$. Каждый элемент $A \cap (B \cup C)$ принадлежит множеству $(A \cap B) \cup (A \cap C)$. Наоборот, элементы $(A \cap B) \cup (A \cap C)$ принадлежат либо $A \cap B$, либо $A \cap C$, т.е. принадлежат A и либо B, либо C, т. е. $A \cap (B \cup C)$. Равенство (5.3) доказано.

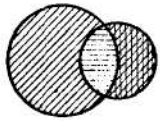
Примем во внимание то, что множество $A \cap B$ может быть неопределенным, если A и B не имеют общих элементов. Чтобы избежать этого, будем рассматривать еще *пустое* множество $\varnothing$, которое не содержит ни одного элемента. Тогда будем считать, что $A \cap B = \varnothing$, если A и B не имеют общих элементов. Пустое множество играет роль *нуля* в операциях над множествами:

$$A \cup \varnothing = A, \quad A \cap \varnothing = \varnothing.$$

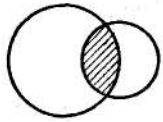
Наглядно операции над множествами можно иллюстрировать, изображая множества в виде кругов (иногда их называют кругами Эйлера) или других фигур на плоскости (рис. 5.2).



B подмножество *A*



Сумма множеств *A* и *B*



Пересечения множеств *A* и *B*

Рис. 5.2

5.2.2. Нахождение числа элементов суммы множеств.

Будем обозначать через $N(A)$ количество элементов множества A.

Основная формула, которой пользуются при нахождении числа элементов суммы двух множеств, такая;

$$N(A \cup B) = N(A) + N(B) - N(A \cap B). \qquad (5.4)$$

Действительно, $N(A)+N(B)$ есть число, которое мы получим, пересчитав все элементы множества A, а потом — все элементы множества B. Но в этом случае общие элементы (их число $N(A \cap B)$) будут перечислены дважды, т.е.

$$N(A) + N(B) = N(A \cup B) + N(A \cap B).$$

Отсюда и следует равенство (5.4). С помощью формулы (5.4) можем получить формулу для числа элементов суммы любого числа множеств. Например, для трех множеств имеем

$$\begin{aligned} N(A \cup B \cup C) &= N\{A \cup (B \cup C)\} = \\ &= N(A) + N(B \cup C) - N\{(A \cap B) \cup (A \cap C)\} = \\ &= N(A) + N(B) + N(C) - N(B \cap C) - \\ &- \{N(A \cap B) + N(A \cap C) - N((A \cap B) \cap (A \cap C))\} = \\ &= N(A) + N(B) + N(C) - N(A \cap B) - N(A \cap C) - \\ &\qquad - N(B \cap C) + N(A \cap B \cap C). \end{aligned}$$

Установим теперь общую формулу для нахождения числа элементов суммы нескольких множеств.

Теорема. *Если $A_1, \ldots, A_n$ — некоторые множества, то*

$$\begin{aligned} N(A_1 \cup \ldots \cup A_n) &= N(A_1) + \ldots + N(A_n) - \\ &- \{N(A_1 \cap A_2) + N(A_1 \cap A_3) + \ldots + N(A_{n-1} \cap A_n)\} + \\ &+ \{N(A_1 \cap A_2 \cap A_3) + N(A_1 \cap A_2 \cap A_4) + \ldots \\ &\ldots + N(A_{n-2} \cap A_{n-1} \cap A_n)\} - \ldots \\ &\ldots + (-1)^{n-1} N(A_1 \cap \ldots \cap A_n). \end{aligned} \tag{5.5}$$

Правая часть равенства (5.5) является суммой n слагаемых, k-е по порядку слагаемое имеет вид

$$(-1)^{k-1} S_k(A, \ldots, A_n).$$

где $S_k(A_1, \ldots, A_n)$ есть сумма чисел $N(A_{i_1} \cap \ldots \cap A_{i_k})$ по всем возможным пересечениям ровно k разных множеств из множеств $A_1, \ldots, A_n$.

Доказательство. Из формулы (5.4) следует, что формула (5.5) справедлива для двух множеств. Предположим, что она справедлива для n— 1 множества, и покажем, что она выполняется и для n множеств (т.е. проведем доазательство по индукции).

По предположению

$$N(A_1 \cup \ldots \cup A_n) = N(A_1) + N(A_2 \cup \ldots \cup A_n) -$$
$$- N\{(A_1 \cap A_2) \cup \ldots \cup (A_1 \cap A_n)\} =$$
$$= N(A_1) + \{S_1(A_2, \ldots, A_n) - S_2(A_2, \ldots, A_n) + \ldots$$
$$\ldots + (-1)^{n-2} S_{n-1}(A_2, \ldots, A_n)\} -$$
$$-\{S_1(A_1 \cap A_2, \ldots, A_1 \cap A_n) - S_2(A_1 \cap A_2, \ldots, A_1 \cap A_n) + \ldots$$
$$\ldots + (-1)^{n-2} S_{n-1}(A_1 \cap A_2, \ldots, A_1 \cap A_n)\}$$

Для того чтобы отсюда получить формулу (5.5), остается принять во внимание, что

$$N(A_1) + S_1(A_2, \ldots, A_n) = S_1(A_1, \ldots, A_n),$$
$$S_2(A_2, \ldots, A_n) + S_1(A_1 \cap A_2, \ldots, A_1 \cap A_n) =$$
$$= S_2(A_1, \ldots, A_n),$$
$$S_k(A_2, \ldots, A_n) + S_{k-1}(A_1 \cap A_2, \ldots, A_1 \cap A_n) =$$
$$= S_k(A_1, \ldots, A_n),$$
$$S_{n-1}(A_1 \cap A_2, \ldots, A_1 \cap A_n) = S_n(A_1, A_2, \ldots, A_n).$$

Теорема доказана.

5.3. Подмножества данного множества

5.3.1. Количество k-элементных подмножеств даного множества.

Если каждый элемент множества B принадлежит множеству A, то B называется *подмножеством* множества A. Это обозначается так: $A \supset B$, или $B \subset A$ (читается: A содержит B, B входит в A). Будем считать, что пустое множество является подмножеством любого множества: $\varnothing \subset A$. Для всякого множества A имеет место соотношения $A \subset A$. Если $A \subset B$ и $B \subset A$, то $A = B$. Подмножество B множества A называется *собственным*, если $B \neq A$ и $B \neq \varnothing$.

Если задано некоторое множество A, то можно рассматривать новое множество $M(A)$ - множество всех его подмножеств. Через $M_k(A)$ будем обозначать множество всех подмножеств A, которые имеют k элементов: $B \subset M_k(A)$, если $B \subset M(A)$ и $N(B) = k$.

Пример. Пусть $A = \{a, b, c\}$. Тогда

$M_1(A) = \{\{a\}, \{b\ \{c\}, \{a, b\}, \{a, c\}, \{b\ c\}, \{a, b\ c\}, 0\}$;
$M_2(A) = \{\{a,b\}, \{a, c\}, \{b\ c\}\}$.

Убеждаемся, что

$$N(M_1(A)) = 8 = 2^3, \quad N(M_2(A)) = 3.$$

Естественно задать вопрос сколько разных k-элементных подмножеств имеет множество из n элементов?

Будем обозначать символом $n!$ (читается n-факториал) произведение всех натуральных чисел от 1 до n включительно: $n! = 1 \cdot 2 \cdot ,,. \ n$. Удобно считать (далее ми убедимся, по каким имено причинам), что 0! =1

Теорема. *Число всех k-элементных подмножеств множества из n элементов равно*

$$N(M_k(A)) = \frac{n(n-1)\dots(n-k+1)}{1 \cdot 2 \dots k} = \frac{n!}{k!(n-k)!}. \tag{5.6}$$

Доказательство. Обозначим $N(M_k(A))=C^k_n$. Чтобы построить k-элементное подмножество множества A, нужно к $(k—1)$-элементному подмножеству присоединить один из $n — k + 1$ элементов, которые не входят в это подмножество. Поскольку $(k — 1)$-элементных подмножеств имеется C^{k-1}_n и каждое из них можно сделать k-элементным $n— k+1$ способами, то таким образом мы получим $(n — k + 1)C^{k-1}_n$ подмножеств. Но не все они будут разными, так как каждое k-элементное множество можно так построить k способами: присоединением каждого из k его элементов. Поэтому вычисленное нами число в k раз больше, чем число C^k_n k-элементных подмножеств. Следовательно,

$$kC^k_n = (n - k + 1)\, C^{k-1}_n.$$

Отсюда найдем

$$C^k_n = \frac{n-k+1}{k} C^{k-1}_n = \frac{(n-k+1)(n-k+2)}{k(k-1)} C^{k-2}_n = \dots$$

$$\dots = \frac{(n-k+1)\dots(n-1)}{k(k-1)\dots 2} C^1_n.$$

Но число одноэлементных подмножеств множества A равно количеству элементов, т.е. n. Подставив вместо C^k_n число n, получим (5.1).

Произвольное k-элементное подмножество n-элементнтного множества называется *сочетанием* из n элементов по k. Порядок элементов в подмножестве не имеет значения. Иногда вместо слова «*сочетание*» употребляется термин — *комбинация* из n элементов по k. Мы установили, что число сочетаний из n элементов по k равно

$$C_n^k = \frac{n!}{k!\,(n-k)!}.$$

Задача 1. Сколькими способами читатель может выбрать 3 книжки из 5?

Решение. Искомое число способов равно числу трехэлементных подмножеств множества из 5 элементов;

$$C_5^3 = \frac{5!}{3! \cdot 2!} = 10.$$

Задача 2. Сколькими способами из 7 человек можно выбрать комиссию, которая состоит из 3 человек?

Решение. Чтобы рассмотреть все возможные комиссии, нужно рассмотреть все возможные 3-элементные подмножества множества, которое состоит из 7 человек. Искомое число способов равно

$$C_7^3 = \frac{7 \cdot 6 \cdot 5}{1 \cdot 2 \cdot 3} = 35.$$

Следующая задача дает интересную геометрическую интерпретацию для чисел C^k_n.

Задача 3. Рассмотрим прямоугольную сетку квадратов размерами $m \times n$ («шахматный город», состоящий из $m \times n$ прямоугольных кварталов, разделенных n — 1 «горизонтальными» и m — 1 «вертикальными» улицами (рис. 5.3)).

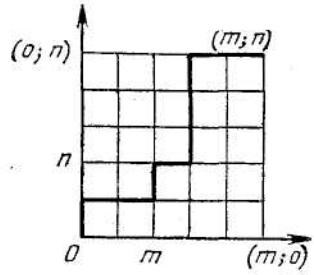


Рис. 5.3

Каково число различных кратчайших путей на этой сетке, ведущих из левого нижнего угла (точки (0; 0)) в правый верхний угол (точку *(m; n)*)?

Решение. Каждый кратчайший путь из точки (0; 0) в точку *(m; n)* состоит из $m+n$ отрезков, причем среди них есть m горизонтальных и n вертикальных отрезков. Разные пути отличаются лишь порядком чередования горизонтальных и вертикальных отрезков. Поэтому общее число путей равно числу способов, которыми из $m+n$ отрезков можно выбрать n вертикальных отрезков, т.е. $C^n_{m+n.}$

Можно было бы рассматривать число способов выбора не n вертикальных, а m горизонтальных отрезков, и мы получили бы тогда ответ $C^m_{m+n.}$ Итак, мы установили геометрически равенство $C^m_{m+n}=C^m_{m+n}$ в справедливости которого нетрудно убедиться и непосредственно, выражая число комбинаций через факториалы.

Итак, число кратчайших путей из точки (0; 0) в точку *(m; n)* равно $C^m_{m+n}=C^m_{m+n.}$

Теорема. *Имеет место равенство*

$$C_n^k = C_{n-1}^k + C_{n-1}^{k-1}. \tag{5.7}$$

Легко убедиться в справедливости равенства (5.7), используя формулу

$$C_n^k = \frac{n!}{k!\,(n-k)!}.$$

Советуем читателю сделать это самостоятельно. Приведем еще два других доказательства.

Доказательство 1. Рассмотрим некоторый элемент a множества A, состоящего из n элементов, и все k-элементные подмножества множества A (число таких подмножеств равно C^k_n). Все k-элементные подмножества множества A разделим на 2 группы: подмножества, в состав которых входит a, и подмножества, в состав которых a не входит. Число подмножеств в первой группе равно C^{k-1}_{n-1}, так как каждое такое подмножество получается присоединением к a некоторого $(k$ —1)-элементного подмножества множества A. Число подмножеств во второй группе равно C^k_{n-1}, так как каждое такое подмножество есть k -элементное подмножество множества A — $\{a\}$. Следовательно,

$$C_n^k = C_{n-1}^{k-1} + C_{n-1}^k.$$

Доказательство 2. Число кратчайших путей из точки $O(0; 0)$ в точку $A(k; n — k)$ равно $C^k_{k+(n-k)} = C^k_n$ (рис. 5.4).

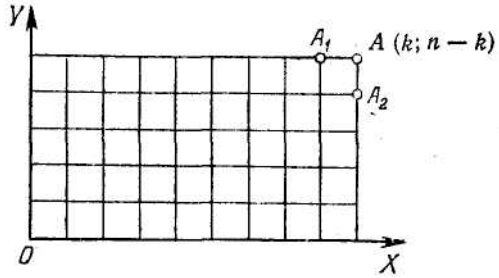


Рис. 5.4

Все такие пути можно разделить на 2 группы: пути, которые проходят через точку $A_1(k—1;\ n—k)$ (число их равно $C_{(k-1)+(n-k)}^{k-1} = C_{n-1}^{k-1}$), и пути, которые проходят через точку $A_2(k;\ n — k —1)$ (число их равно

$$C_{k+(n-k-1)}^{k} = C_{n-1}^{k}).$$

Следовательно,

$$C_n^k = C_{n-1}^{k-1} + C_{n-1}^k.$$

З а д а ч а 4. Доказать тождество

$$C_{2n}^n = (C_n^0)^2 + (C_n^1)^2 + \ \dots\ + (C_n^n)^2. \tag{4.8}$$

Решение. Число кратчайших путей из точки O (0; 0) в точку $A(n;n)$ равно $C^n_{2n.}$ Каждый тякой путь проходит через одну и только одну из точек $A_k(k;\ n — k)$, лежащих на диагонали BD (рис. 5.5).

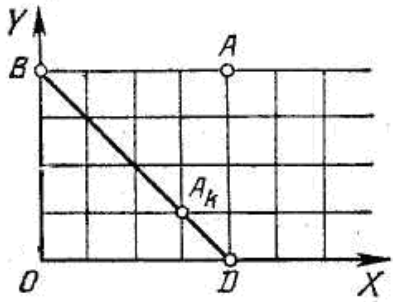


Рис. 5.5

Число путей из точки O в точку A_k равно $C^k_{k+(n-k)} = C^k_n$, а из точки A_k в точку A равно $C^k_{n-k+k} = C^k_n$, поэтому число путей из O в A, проходящих через A_k, равно $C^k_n \cdot C^k_n = (C^k_n)^2$ (правило умножения). Прибавив количество путей, которые проходят через кождую из точек A_k (k = 0, 1, ..., n), получим общее количество путей из O в A, т.е. C^n_{2n}. Это соображение и доказывает равенство (5.8).

5.3.2. Количество подмножеств данного множества.

Выясним теперь, сколько всего подмножеств имеет множество A, состоящее из n элементов (пустое множество также является подмножеством A).

Теорема. *Число всех подмножеств множества из n элементов равно* 2^n.

Приведем два различных доказательства.

Доказательство 1. Пусть M_a — множество всех подмножеств множества A, которые содержат элемент a. Очевидно, что каждое такое подмножество полностью определено, если указаны все его остальные (кроме a) элементы. Поэтому таких подмножеств будет столько, сколько будет подмножеств в множестве $A' = A$ — $\{a\}$, которое содержит все элементы A, кроме a. Это множество имеет n — 1 элементов. Поэтому, если q_n — число подмножеств множества из n элементов, то $N(M_a) = q_{n-1}$.

Если $\overline{M}_a$ — множество всех подмножеств множества A, не содержащих a, то $N(\overline{M}_a)$ также будет равно q_{n-1}. Поскольку $M(A) = M_a + \overline{M}_a$, то $N(M(A)) = 2q_{n-1}$. Отсюда находим $q_n = 2q_{n-1}$. Таким образом, $q_n = 2q_{n-1}$ — $2^2\, q_{n-2} = \ldots = 2^{n-1} q_1$. Множество, которое состоит из 1 элемента, имеет 2 подмножества (все множество и пустое множество). Поэтому $q_1 = 2$. Следовательно, $q_n = 2^n$.

Доказательство 2. Перенумеруем элементы множества A и для каждого подмножества множества A построим последовательность длиной n из нулей и единиц по следующему правилу: на k-м месте пишем 1, если элемент с номером k входит в подмножество, и 0, если элемент с номером k не входит в подмножество. Итак, каждому подмножеству соответствует своя последовательность нулей и единиц. Например, пустому множеству соответствует последовательность из одних нулей. Число всех возможных последовательностей длины n, составленных из нулей и единиц,

равно, согласно правилу умножения, $\underbrace{2 \cdot 2 \cdot \ldots \cdot 2}_{n \text{ разів}} = 2^n$. Следовательно, и число всех подмножеств множества A равно 2^n.

Как было указано выше, удобно считать 0!=1. При этом предположении формула

$$C_n^k = \frac{n!}{k!\,(n-k)!}$$

остается в силе и при $k = n$ и при $k = 0$.

Следствие. *Имеет место равенство*

$$\sum_{k=0}^{n} C_n^k = 2^n.$$

В самом деле, поскольку C^k_n — число k-элементных подмножеств множества из n элементов, то сумма в левой части есть число всех подмножеств.

5.4. Упорядоченные множества. Перестановки и размещения

5.4.1. Перестановки даного множества.

Множество называется *упорядоченным*, если каждому элементу этого множества поставлено в соответствие некоторое число (номер элемента) от 1 до n, где n — число элементов множества, так что различным элементам соответствуют различные числа. Всякое конечное множество можно сделать упорядоченным, если, например, переписать все элементы множества в некоторый список *(a, b, c, ...)*, а потом поставить в соответствие каждому элементу номер места, на котором он стоит в списке. Будем обозначать упорядоченное множество, которое получено из множества A, через $\tilde{A}$. Очвидно, что каждое множество, которое содержит больше одного элемента, можно упорядочить не единственным способом. Упорядоченные множества считаются различными, если они отличаются либо своими элементами, либо их порядком. Различные упорядоченные множества, которые отличаются лишь порядком элементов (т.е. могут быть получены из того же самого множества), называются *перестановками* этого множества.

Пример. Перестановки множества $A = \{a, b, c\}$ из 3 элементом имеют вид

(a, b, c), (a, c, b), (b, a, c),

(b, c, a), (c, a, b), (c, b, a).

Найдем число различных способов, которыми может быть упорядочено данное множество, т.е. число перестановок множества *A*. Пусть множество *A* имеет *n* элементов. Обозначим число ее перестановок через P_n.

Теорема.

$$P_n=n!.$$

Доказательство 1. Выберем некоторый элемент *a* из множества *A*. Рассмотрим все перестановки, в которых *a* имеет номер 1. Число таких перестановок будет равно числу перестановок из *n*—1 элементов множества *A,* которые остаются после исключения из множества элемента *a*. Поэтому число перестановок, для которых *a* имеет номер 1, равно P_{n-1}. Обозначим через *M* множество всех перестановок множества *A,* а через M_a — множество перестановок, в которых *a* имеет номер 1. Тогда

$$M = M_a \cup M_b \cup \ldots \cup M_f,$$

где *a, b, ..., f* — все элементы множества *A*. Поскольку никакие 2 множества из множеств M_a, M_b, ..., M_f не имеют общих элементов (напомним, что элементы этих множеств — перестановки, в различных множествах на первом месте стоят различные элементы, следовательно, и соответствующие перестановки будут различными), то

$$N(M) = N(M_a) + N(M_b) + \ldots + N(M_f).$$

Следовательно,

$$P_n=n\cdot P_{n-1}=n!.$$

Доказательство 2. Будем последовательно выбирать элементы множества *A* и размещать их в определенном порядке на *n* местах. На первое место можно поставить каждый из *n* элементов. После того как заполнено первое место, на второе место можно поставить любой из оставшихся *n* — 1 элементов и т.д. По правилу умножения все *n* мест можно заполнить

$$n(n-1)(n-2)\ldots 2\cdot 1 = n!$$

способами. Следовательно, множество *A* из *n* элементов можно упорядочить *n*! способами.

Задача 1. Сколькими способами можно разместить на полке 4 книги (обозначим их *A, B, C, D)?*

Решение. Искомое число способов равно числу способов упорядочения множества, которое состоит из 4 элементов, т.е.

$$P_4 = 1 \cdot 2 \cdot 3 \cdot 4 = 24.$$

Задача 2. Сколькими способами можно упорядочить множество $\{1, 2, \ldots, 2n\}$ так, чтобы каждое четное число имело четный номер?

Решение.. Четные числа можно расставить на местах с четными номерами (таких мест *n)* *n*! способами; каждому способа размещения четных чисел на местах с четными номерами соответствует *n*! способов размещения нечетных чисел на местах с нечетными номерами. Поэтому общее число перестановок указанного типа по правилу умножения равно $n!\cdot n!=(n!)^2$.

5.4.2. Упорядоченные подмножества данного множества (размещения).

Рассмотрим теперь упорядоченные подмножества данного множества *A.* Само множество *A* считаем неупорядоченным, поэтому каждое его подмножество может быть упорядочено каким-либо возможным способом. Число всех *k*-элементных подмножеств множества *A* равно C^k_n. Каждое такое подмножество можно упорядочить *k*! способами. Таким образом получим все упорядоченные *k*-элементные подмножества множества *A.* Следовательно, их число будет $k!\bullet C^k_n$.

Теорема. *Число упорядоченных k-элементных подмножеств множества, состоящего из n элементов, равно*

$$A^k_n = k!\bullet C^k_n = \frac{n!}{(n-k)!} = n\,(n-1)\,...(n-k+1).$$

Упорядоченные *k*-элементные подмножества множества из *n* элементов называются *размещениями* из *n* элементов по *k.* Различные размещения из *n* по *k* отличаются количеством элементов либо их порядком.

Следовательно, число различных размещений из *n* по *k* равно

$$A^k_n = n(n-1)\,...(n-k+1).$$

5.5. Перестановки с повторениями

Поставим такой вопрос. Сколькими способами можно разложить множество *A,* состоящее из *n* элементов, на сумму *m* подмножеств

$$A=B_1 \cup B_2 \cup \ldots \cup B_m$$

так, чтобы $N(B_1)=k_1$, $N(B_2)=k_2$, ..., $N(B_m)=k_m$, где $k_1, k_2, .,., k_m$ — данные числа ($k_i \geq 0$, $k_1+\ldots+k_m = n$)? Множества $B_1, B_2, \ldots, B_m$ не должны иметь общих элементов.

Все описанные выше разбиения множества *A* на *m* групп $B_1, B_2, \ldots, B_m$ можно получить так: возьмем произвольное k_1-элементное

подмножество B_1 множества A (это можно сделать $C^{k_1}_n$ способами); среди n — k_1 оставшихся элементов возьмем k_2-элементное подмножество B_2 (это можно сделать $C^{k_1}_{n-k_1}$ способами) и т.д. Общее число способов выбора различных множеств $B_1,..., B_m$ по правилу умножения равно

$$C_n^{k_1} \cdot C_{n-k_1}^{k_2} \cdot C_{n-k_1-k_2}^{k_3} \cdots C_{n-k_1-\ldots-k_{m-1}}^{k_m} =$$

$$= \frac{n!}{k_1!\,(n-k_1)!} \cdot \frac{(n-k_1)!}{k_2!\,(n-k_1-k_2)!} \cdot \frac{(n-k_1-k_2)!}{k_3!\,(n-k_1-k_2-k_3)!} \cdots$$

$$\cdots \frac{(n-k_1-\ldots-k_{m-1})!}{k_m!\,(n-k_1-\ldots-k_m)!} = \frac{n!}{k_1!\,k_2!\ldots k_m!}$$

(напомним, что 0! = 1).

Итак, мы доказали следующую теорему.

Теорема. *Пусть $k_1, k_2, ..., k_m$ — целые неотрицательные числа, причем $k_1+k_2+...+k_m=n$. Число способов, которыми можно представить множество A из n элементов в виде суммы m множеств $B_1, B_2, ..., B_m$, число элементов которых составляет соответственно $k_1,k_2, ..., k_m$, равно*

$$C_n(k_1, ..., k_m) = \frac{n!}{k_1!\ldots k_m!}$$

Числа $C_n(k_1, ..., k_m)$ называются *полиномиальными коэффициентами.* Они имеют еще одну очень важную комбинаторную интерпретацию.

Пусть имеется n букв: k_1 — букв a_1, k_2 — букв a_2, k_m — букв a_m $(k_1 + k_2 + ... + k_m = n)$. Определим, сколько различных слов можно сложить из этих букв. Перенумеруем места, на которых стоят буквы, числами 1, 2, ..., n. Каждое слово определяется множествами B_1 (номера мест, где стоит буква a_1), B_2 (номера мест, где стоит буква a_2), ..., B_m (номера мест, где стоит буква a_m). Следовательно, число различных слов равно числу способов, которыми можно представить множество $A = \{1, 2, ... n\}$ в виде суммы множеств $B_1, B_2, ..., B_m$, т.е.

$$C_n(k_1, ..., k_m) = \frac{n!}{k_1!\ldots k_m!}$$

Пример. Число различных слов, которое получим, переставляя буквы слова «математика», равно

$$\frac{10!}{2!\cdot 3!\cdot 2!}$$

Утверждение, установленное выше, можно сформулировать в виде следующей теоремы.

Теорема. *Число различных перестановок, которые можно составить из n элементов, среди которых имеется k_1 элементов первого типа, k_2 элементов второго типа, ..., k_m элементов m-го типа, равно*

$$C_n(k_1, ..., k_m) = \frac{n!}{k_1!...k_m!}$$

В связи с важностью теоремы приведем еще одно доказательство. Рассмотрим одну перестановку и заменим в ней все одинаковые элементы разными. Тогда число различных перестановок, которые можно составить из рассматриваемой нами перестановки, равно $k_1!\cdot k_2!$ ... $k_m!$. Проделав это для каждой перестановки, получим $n!$ перестановок. Следовательно,

$$C_n(k_1, ..., k_m)\cdot k_1! ... k_m! = n!,$$

что и доказывает утверждение теоремы.

5.6. Взаимно однозначное соответствие. Соединения с повторениями

5.6.1. Взаимно однозначное соответствие.

Предположим, что задано 2 множества A и B. Будем считать, что между этими множествами установлено соответствие, если каждому элементу a из A соответствует некоторый элемент b из B и, наоборот, для каждого элемента b из B существует такой элемент a из A, что b соответствует a. Это соответствие будет взаимно однозначным, если каждому элементу из A соответствует только один элемент из B и различным элементам множества A соответствуют различные элементы множества B.

Пример 1. A — множество студентов группы, B-множество столов. Каждому студенту соответствует стол за которым он сидит (это не взаимно однозначное соответствие).

Пример 2. A — множество всех городских жителей Украины, B — множество всех городов Украины. Каждому элементу A соответствует

город, в котором он живет (это также не взаимно однозначное соответствие).

Примером взаимно однозначного соответствия может быть соответствие между элементами упорядоченного множества A из n элементов и числами 1, 2.....n; каждому элементу соответствует его номер.

Определение. Множества, для которых существует взаимно однозначное соответствие, называются *эквивалентными.*

Теорема. *Для того чтобы два множеств были эквивалентными, необходимо и достаточно, чтобы они имели одинаковое число элементов.*

Доказательство. Если множества A и B имеют одинаковое число элементов n, то, упорядочивая каждое из них некоторым образом и ставя в соответствие k-му элементу множества $\dot{A}$ k-й элемент множества $\dot{B}$, получим взаимно однозначное соответствие между множествами A и B, т.е. множества A и B эквивалентны.

Предположим теперь, что A имеет n элементов и существует взаимно однозначное соответствие между A и B. Упорядочим множество A: пусть элементами A будут $a_1, a_2, ..., a_n$. Обозначим через b_k тот элемент B, который соответствует a_k. Поскольку каждому элементу из A соответствует элемент из B, различным элементам из A соответствуют различные элементы из B, и каждый элемент из B соответствует некоторому элементу из A, то B состоит из элементов $b_1, b_2 \ldots, b_n$, следовательно, B имеет n элементов.

Следствие. *Если два множества эквивалентны, то они имеют одинаковое число элементов.*

Это свойство эквивалентных множеств очень часто используют для вычисления количества элементов различных множеств.

Пример 3. Рассмотрим множество A последовательностей $x_1, x_2, \ldots, x_n$ из n чисел, где числа x, принимают только значение 0 и 1 и среди них ровно k единиц. Чтобы вычислить число элементов нашего множества, обращаем внимание, что оно эквивалентно множеству B всех k-элементных подмножеств множества $\{1, 2, ... n\}$: подмножество чисел $\{i_1, ..., i_k\}$ соответствует той последовательности $x_1, x_2, ..., x_n$, у которой $x_{i_1}=1$, $x_{i_2}=1$, ..., $x_{i_k}=1$. Следовательно,

$$N(A) = C^k_n.$$

Пример 4. Найти число размещений n одинаковых предметов в m урнах.

Перенумеруем урны. Поставим в соответствие каждому размещению предметов в урнах последовательность из нулей и единиц следующим образом: сначала последовательность имеет группу нулей, количество которых равно числу предметов в первой урне, потом записываем единицу и дальше пишем столько нулей, сколько предметов во второй урне, снова единицу, потом столько нулей, сколько предметов в третьей урне, и т.д. Заканчивается последовательность группой нулей, которая является числом предметов в последней урне. Следовательность, последовательность имеет n нулей и m — 1 единиц, всего $n + m$—1 чисел. Например, при n=10, m=4 последовательность 1011000000000 соответствует размещению: первая урна пустая, вторая урна имеет 1 предмет, третья урна пустая, четвертая урна имеет 9 предметов; а последовательность 0010010000001 соответствует размещению: первая урна имеет 2 предмета, вторая — 2 предмета, третья — 6 предметов, четвертая — пустая.

Используя предыдущий пример, получаем, что искомое число размещений будет C^{m-1}_{n+m-1}.

5.6.2. Сочетания с повторениями.

Сочетаниями из m элементов по n элементов с повторениями называются группы, которые содержат n элементов, причем каждый элемент принадлежит к одному з m типов.

Например, из трех элементов *a, b, c* можно составить такие сочетания по два с повторениями:

aa, acc, be, ab, bb, cc.

Теорема. *Число различных сочетаний из m элементов по n с повторениями равно*

$$\overline{C}^n_m = C^{m-1}_{n+m-1} = C^n_{m+n-1}$$

Доказательство. Каждое сочетание полностью определяется, если указать, сколько элементов каждого из m типов в него входит. Поставим в соответствие каждому сочетанию последовательность нулей и единиц, составленную по такому правилу: напишем подряд столько единиц, сколько элементов первого типа входит у сочетание, дальше поставим нуль и после него напишем столько единиц, сколько элементов второго типа содержит это сочетание и т.д. Например, написанным выше сочетанием из трех букв по две будут соответствовать такие последовательности:

1100, 1001, 0101, 1010, 0110, 0011.

Таким образом, каждому сочетанию из m по n соответствует последовательность из n единиц и m — 1 нулей, и наоборот, по каждой такой последовательности однозначно восстанавливается такое сочетание. Поэтому число сочетаний из m по n с повторениями равно числу последовательностей из n единиц и m — 1 нулей, т.е. равно C^{m-1}_{n+m-1},

Пример. Кости домино можно рассматривать как сочетания с повторениями по два из семы цифр: 0, 1,2, 3, 4, 5, 6. Число всех таких сочетаний равно

$$f^2_7 = \frac{8 \cdot 7}{2} = 28$$

5.7. Прямое произведение множеств

Предположим, что задано множества $A_1, \ldots, A_k$. Множество всех элементов вида $(a_1, \ldots, a_k)$, где $a_1 \in A_1$, $a_2 \in A_2$, ..., $a_k \in A_k$, называется *прямым произведением множеств* $A_1, \ldots, A_k$ и обозначается

$$A_1 \times A_2 \times \ldots \times A_k$$

Пример 1 . Если $A = \{a, b\}$, $B = \{c, d, e\}$, то

$$A \times B = \{(a, c), (a, d), (a, e), (b, c), (b, d), (b, e)\}.$$

Пример 2. Пусть имеется множество A из n элементов. Возьмем из A какой-нибудь элемент, обозначим его через a_1 и вернем снова в множество A. Далее возьмем из A некоторый элемент, обозначим его через a_2 (в частности, может случиться, что попадеться снова элемент a_1). Проделав эту операцию k раз, получим набор $(a_1, \ldots, a_k)$, который называют k-словом, составленным из элементов множества A. Множество всех k-слов, составленных из элементов A, является прямым произведением $A \times A \times \ldots\ldots \times A$ и кратко обозначается A^k. Например, если A — множество из двух букв $\{a, b\}$, то множество A^2 всех слов имеет вид $\{aa, ab, ba, bb\}$. С k-словами мы часто сталкиваемся в различных ситуациях. Все десятичные записи чисел являются словами, составленными из цифр 0, ..., 9; обычные слова - это слова, состоящие, например, из букв русского алфавита; фразы - это «слова», что состоящие из русских слов, и т.д.

Естественно задать вопросы: сколько различных k-слов можно составить из элементов множества A, имеющего n элементов?

Следующая теорема дает ответ на этот и на более общий вопрос: сколько элементов содержит прямое произведение $A_1 \times A_2 \times \ldots \times A_k$?

Теорема. $N(A_1 \times A_1 \times \ldots \times A_k) = N(A_1) \ldots N(A_k)$.

Доказательство. Покажем сначала, что

$$N(A_1 \times A_2) = N(A_1) \cdot N(A_2).$$

Обозначим через B_{c_1} подмножество множества $A_1 \times A_2$, которое состоит из элементов вида (c_1, a_2), где c_1 — фиксированный элемент из A_1, а a_2 произвольный элемент из A_2. Тогда $N(B_{c_1}) = N(A_2)$, так как B_{c_1} эквивалентно A_2 (элементу (c_1, a_2) соответствует a_2). Если $a_1, b_1, .., f_1$ — все элементы множества A_1, то

$$A_1 \times A_2 = B_{a_1} \cup B_{b_1} \cup \ldots \cup B_{f_1}.$$

Множества B_{a_1}, B_{b_1}, ..., B_{a_1}, попарно не имеют общих элементов. Поэтому

$$N(A_1 \times A_2) = N(B_{a_1}) + N(B_{b_1}) + \ldots + N(B_{f_1}) = \\ = N(A_1) \cdot N(A_2).$$

Примем теперь во внимание, что множества

$$A_1 \times (A_2 \times \ldots \times A_k) \text{ и } A_1 \times A_2 \ldots \times A_k$$

эквивалентны: элементу $[a_1 (a_2, \ldots, a_k)]$ соответствует элемент $(a_1, a_2 \ldots .a_k)$. Поэтому

$$N(A_1 \times \ldots \times A_k) = N(A_1) \cdot N(A_2 \times \ldots \times A_k) = \\ = N(A_1) \cdot N(A_2) \cdot N(A_3 \times \ldots \times A_k) = \\ = N(A_1) \cdot N(A_2) \ldots N(A_k),$$

что и требовалось доказать.

Пример 3. Число k-слов, составленных из элементов множества A, равно

$$N(A^{k)} = [N(A)]^k.$$

Пример 4. Дадим ответ на вопрос, сколькими способами можно распределить k разных предметов сред n лиц?

Пусть A — множество лиц, среди которых распределяют предметы. Перенумеруем предметы и поставим в соответствие каждому способу распределения символ $(a_1, a_2 \ldots .a_k)$, где a_1 — лицо, которое получило первый предмет, ..., a_k — лицо, которое получило k-й предмет. Очевидно, $(a_1, a_2 \ldots .a_k)$ - k-слово, составленное из элементов множества A. Установленное соотношение является взаимно однозначным, и поэтому число способов распределения k предметов сред n лиц равно числу k-слов, которые можно составить из элементов множества A, т.е., ровно n^k.

5.8. Бином Ньютона и полиномиальная формула

5.8.1. Бином Ньютона.

Известно, что

$$(a + b)^2 = a^2 + 2ab + b^2,$$
$$(a + b)^3 = a^3 + 3a^2b + 3ab^2 + b^3.$$

Как раскрыть скобки при вычислении выражения $(a + b)^n$? Ответ на этот вопрос дает следующая теорема

Теорема. *Имеет место равенство*

$$(a+b)^n = C_n^0 a^n b^0 + C_n^1 a^{n-1} b^1 + \ldots$$
$$\ldots + C_n^k a^{n-k} b^k + \ldots + C_n^n a^0 b^n, \qquad (5.9)$$

где

$$C_n^k = \frac{n!}{k!\,(n-k)!}.$$

Эту теорему иногда называют *биномиальной теоремой,* а числа C^k_n *— биномиальными коэффициентами.* Равенство (5.9) часто называют *биномом Ньютона,* хотя это название исторически не является справедливым, так как формулу для $(a + b)^n$ знали еще среднеазиатские математики Омар Хайям (1048— 1131), Гийас ад-Дин Джемшид ал-Каши (умер ок. 1430). В Европе до Ньютона ее знал Паскаль, (1623-1662). Заслуга Ньютона в том, что он обобщил формулу (5.9) для нецелого показателя n. Вид формулы бинома в этом случае рассмотрен в п. 5.9. Напомним, что C^k_n есть число k-элементных подмножеств множества из n элементов. Формулу (5.9) можно записать в виде

$$(a+b)^n = \sum_{k=0}^{n} C_n^k a^{n-k} b^k.$$

Доказательство. Перемножим последовательно $(a + b)$ n раз. Тогда получим сумму 2^n слагаемых вида $d_1 d_2 \ldots d_n$, где d_i $(i=1,\ldots,n)$ равно либо a, либо b. Разобьем все слагаемые на $n+1$ группу $B_0, B_1, \ldots, B_n$, относя к B_k все те произведения, в которых b встречается множителем k раз, а a - $(n — k)$ раз. Число произведений в B_k равно, очевидно, C^k_n (таким числом способов среди n множителей $d_1, \ldots, d_n$ можно выбрать k множителей, которые будут равны b), а каждое слагаемое в B_k равно $a^{n-k}b^k$. Поэтому

$$(a+b)^n = \sum_{k=0}^{n} C_n^k a^{n-k} b^k.$$

Теорема доказана.

Напомним следующее важное свойство биномиальных коэффициентов:

$$C_{n+1}^{k} = C_{n}^{k} + C_{n}^{k-1}. \tag{5.10}$$

Оно было установлено в п. 4 3.

Равенство (5.10) показывает, что биномиальные коэффициенты можно последовательно выписывать в виде треугольной таблицы, которая называется *треугольником Паскаля:*

$$\begin{array}{ccccccccccc} & & & & 1 & & 1 & & & & n=1 \\ & & & 1 & & 2 & & 1 & & & n=2 \\ & & 1 & & 3 & & 3 & & 1 & & n=3 \\ & 1 & & 4 & & 6 & & 4 & & 1 & n=4 \\ 1 & & 5 & & 10 & & 10 & & 5 & & 1 \quad n=5 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{array}$$

В n-й строке треугольника Паскаля стоят коэффициенты разложения $(a + b)^n$, причем каждый коэффициент, кроме крайних двух, которые равны 1, равен сумме соответствующих коэффициентов из предыдущей строки.

5.8.2. Полиномиальная теорема.

Рассмотрим вопрос о том, как раскрывать скобки при вычислении выражения вида

$$(a_1 + a_2 + \ldots + a_k)^n$$

Полиномиальная теорема. *Выражение*

$$(a_1 + a_2 + \ldots + a_k)^n$$

равно сумме всех возможных слагаемых вида

$$\frac{n!}{r_1! r_2! \ldots r_k!} a_1^{r_1} \cdot a_2^{r_2} \ldots a_k^{r_k},$$

где

$$r_1 + r_2 + \ldots + r_k = n,$$

т.е.

$$(a_1 + a_2 + \ldots + a_k)^n = \sum_{\substack{r_1 \geqslant 0, \ldots, r_k \geqslant 0 \\ r_1 + \ldots + r_k = n}} \frac{n!}{r_1! \ldots r_k!} a_1^{r_1} \ldots a_k^{r_k}, \tag{5.11}$$

Доказательство. Перемножим последовательно $a_1 + a_2 + \ldots + a_k$ n раз. Тогда получим k^n слагаемых вида $d_1 d_2 \ldots d_n$, где каждый

множитель d_i равен или a_1, или a_2, ..., или a_n. Обозначим через $B(r_1 \ldots r_k)$ совокупность всех тех слагаемых, где a_1 встречается множителем r_1 раз, a_2 — r_2 раз, ..., a_k — r_k раз.

Число таких слагаемых равно $C_n(r_1 \ldots, r_k)$ - числу способов представления множества из n элементов $\{1,2, \ldots, n\}$ в виде суммы k множеств $B_1, B_2, \ldots, B_k$ так, чтобы множество B_s имело r_s элементов ($r_s \geq 0$, $r_1 + \ldots + r_k = n$ множество B_s — это множество тех i, для которых $d_i = a_s$).

В п. 5.5 было показано, что

$$C_n(r_1, \ldots, r_k) = \frac{n!}{r_1! \ldots r_k!}.$$

Следовательно,

$$(a_1 + \ldots + a_k)^n = \sum_{\substack{r_1 \geqslant 0, \ldots, r_k \geqslant 0 \\ r_1 + \ldots + r_k = n}} \frac{n!}{r_1! \ldots r_k!} a_1^{r_1} \ldots a_k^{r_k}.$$

При $k = 2$ равенство (5.11) имеет вид

$$(a_1 + a_2)^n = \sum_{r=0}^{n} \frac{n!}{r!(n-r)!} a_1^{n-r} a_2^r.$$

Следовательно, мы получили формулу бинома Ньютона.

5.8.3. Биномиальные тождественности.

Числа C^k_n имеют ряд важных свойств. Укажем некоторые из них и установим ряд интересных тождеств, которым удовлетворяют биномиальные коэффициенты.

Напомним в первую очередь следующие равенства:

$$C^k_n = C^{n-k}_n, \quad (5.12)$$

$$C^k_{n+1} = C^k_n + C^{k-1}_n, \quad (5.13)$$

$$C^0_n + C^1_n + \ldots + C^n_n = 2^n, \quad (5.14)$$

$$C^0_n - C^1_n + C^2_n - \ldots + (-1)^n C^n_n = 0. \quad (5.15)$$

Равенство (5.12) легко проверяется вычислением. Оно следует также из того, что число k-элементных подмножеств множества из n элементов равно числу (n — k)-элементных подмножеств. Равенство (5.13) мы доказали в п. 5.3.1. Равенство (5.14) получим, взяв в формуле бинома $a = b = 1$. Оно следует также из комбинаторных соображений, которые были проведены в п. 5.3.1. Если в формуле бинома Ньютона положить $a = 1$, $b = -1$, получим равенство (5.15).

Иногда при доказательстве некоторых тождеств полезно иметь в виду геометрическую интерпретацию чисел C_n^k, которая была приведена в п. 5.3.1.

Докажем еще несколько важных биномиальных тождеств.

Задача 1. Доказать, что

$$C_n^r = C_{n-1}^{r-1} + C_{n-2}^{r-1} + \ldots + C_{r-1}^{r-1}. \tag{5.16}$$

Доказательство 1. Рассмотрим все r-элементные подмножества множества

$$A = \{a_1, \ldots, a_n\}.$$

Число их равно C_n^r. Разобьем все эти подмножества на классы $T_1, \ldots, T_{n-r+1}$ отнеся к классу T_k все те r-элементные подмножества множества A, в которых элемент с наименьшим индексом равен a_k.

Поскольку каждое подмножество из класса T_k может быть получено присоединениям к a_k некоторого $(r-1)$-элементного подмножества множества $\{a_{k+1}, a_{k+2}, \ldots a_n\}$, то класс T_k состоит из C_{n-k}^{r-1} подмножеств. Следовательно, получаем равенство (5.16).

Доказательство 2. Рассмотрим все кратчайшие ламаные, которые соединяет точку (0; 0) с точкой $(r; n-r)$.

Число всех таких ламаных равно C_n^r. Отнесем к классу B_k те ламаные, которые пересекают прямую $x = 1/2$ в точке $(1/2; k)$ ($k = 0, \ldots, n-1$). Очевидно, класс B_k состоит из C_{n-k-1}^{r-1} ломаных (рис. 5.6).

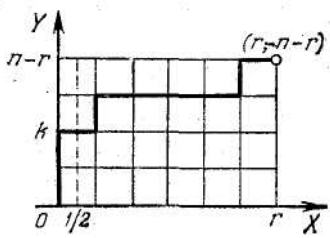


Рис. 5.6

Поэтому

$$C_n^r = \sum_{k=0}^{n-r} C_{n-k-1}^{r-1}.$$

Задача 2. Доказать, что

$$C_n^0 C_m^k + C_n^1 C_m^{k-1} + \ldots + C_n^k C_m^0 = C_{n+m}^k. \tag{5.17}$$

Доказательство 1. Все k-элементные подмножества множества

$$A = \{a_1 \ldots, a_n, a_{n+1}, \ldots, a_{n+m}\}$$

(число их равно C^k_{n+m}) разобьем на $k+1$ класс V_0, V_1, ..., V_k, отнеся к классу V_i все те подмножества, в состав которых входит ровно i элементов с индексами, не большими чем n. Каждое подмножество из класса V_i можно получить, присоединяя к некоторому i-элементному подмножеству множества $\{a_1, ..., a_n\}$ $(k — i)$-элементное подмножество множества $\{a_{n+1}, ..., a_{n+m}\}$. Поэтому в состав V_i входит $C^i_n C^{k-i}_m$ подмножеств. Следовательно,

$$C^k_{n+m} = \sum_{i=0}^{k} C^i_n C^{k-i}_m.$$

Считая в (5.17) $k = n = m$, имеем тождество

$$(C^0_n)^2 + (C^1_n)^2 + \ldots + (C^n_n)^2 = C^n_{2n}, \tag{5.18}$$

которое уже было доказано с помощью геометрических соображений в п. 5.3.1 (задача 6).

Доказательство 2. Воспользуемся следующим замечанием: если два многочлена $P(x)$ и $Q(x)$ равны при всех x, то коэффициенты этих многочленов при одинаковых степенях x равны. Действительно, если $P(x) \equiv Q(x)$ и при некоторой степени x коэффициенты не равны, то многочлен $P(x)—Q(x)$ будет многочленом с ненулевыми коэффициентами, имеющим бесчисленное множество действительных корней, что противоречит основной теореме алгебры.

Запишем равенство

$$(1+x)^n(1+x)^m = (1+x)^{m+n}. \tag{5.19}$$

Используя формулу бинома Ньютона, убеждаемся, что коэффициент при x^k в правой части равенства (5.19) равен C^k_{m+n}, а коэффициент при x^k в левой части этого равенства имеет вид

$$C^0_n C^k_m + C^1_n C^{k-1}_m + \ldots + C^k_n C^0_m.$$

Равенство (5.17) доказано.

Микромодуль16
Примеры решения типовых задач

Задача 1. Сколько четырехзначных чисел можно сложить из цифр 0, 1,2, 3, 4, 5, если:

а) ни одна из цифр не повторяется боле одного раза;

б) цифры могут повторяться;

в) числа должны быть нечетными (цифры могут повторяться)?

Решение. а) Первой цифрой числа может быть одна из 5 цифр 1, 2, 3, 4, 5 (0 не может быть первой цифрой, потому что в таком случае число не является четырехзначным); если первая цифра выбрана, то друга может быть выбранная 5 способами, третья - 4 способами, четвертая - 3 способами. Согласно правилу умножения общее число способов равно 5×5×4×3 = 300.

б) Первой цифрой может быть одна из цифр 1, 2, 3, 4, 5 (5 возможностей), для каждой из следующих цифр имеем 6 возможностей (0, 1, 2, 3, 4, 5). Следовательно, число искомых чисел равно $5\times6\times6\times6=5\times6^3=1080$.

в) Первой цифрой может быть одна из цифр 1, 2, 3, 4, 5, а последней - одна из цифр 1, 3, 5 (числа должны быть нечетными). Следовательно, общее количество чисел равно 5×6×6×3 = 540.

Задача 2. Каждый студент группы - либо девушка, либо блондин, либо любит математику. В группе 20 девушек, из них 12 блондинок, и одна блондинка любит математику. Всего в группе 24 студента-блондина, математику из них любят 12, а всего студентов (ребят и девушек), которые любят математику, 17, из них 6 девушек. Сколько студентов в данной группе?

Решение. Если A - множество девушек, B - блондинов, C — студентов, которые любят математику, то $N(A \cup B \cup C)$ - искомое число. $A \cap B$ — множество блондинок, $A \cap C$ — множество девушек, которые любят математику, $B \cap C$ — множество всех блондинов (ребят и девушек), которые любят математику, $A \cap B \cap C$ — множество блондинок, которые любят математику. Тогда

$$N(A \cup B \cup C) = N(A) + N(B) + N(C) - N(A \cap B) - N(A \cap C) - N(B \cap C) + N(A \cap B \cap C) = 20 + 24 + 17 - (12 + 6 + 12) + 1 = 32.$$

Задача 3. В турнире принимали участие n шахматистов, и каждые 2 шахматиста встретились 1 раз. Сколько партий было сыграно в турнире?

Решение. Партий было сыграно столько, сколько можно выделить 2-элементных подмножеств в множестве из n элементов, т.е.

$$C_n^2 = \frac{n(n-1)}{1 \cdot 2}.$$

Задача 4. В скольких точках пересекаются диагонали выпуклого n-угольника, если никакие 3 из них не пересекаются в одной точке?

Решение. Каждой точке пересечения двух диагоналей соответствует 4 вершины n-угольника, а каждым 4 вершинам n-угольника соответствует 1 точка пересечения (точка пересечения диагоналей четырехугольника с вершинами в данных 4 точках). Поэтому число всех точек пересечения равно числу способов, которыми среди n вершин можно выбрать 4 вершины

$$C_n^4 = \frac{n(n-1)(n-2)(n-3)}{1 \cdot 2 \cdot 3 \cdot 4} = \frac{n(n-1)(n-2)(n-3)}{24}.$$

Задача 5. Сколько можно составить перестановок из n элементов, в которых данные 2 элемента не стоят рядом?

Решение. Определим число перестановок, в которых данные 2 элемента a и b стоят рядом. Могут быть следующие случаи: a стоит на первом месте, a стоит на втором месте, ..., a стоит на (n—1)-в месте, а b стоит правее a; число таких случаев равно n— 1. Кроме того, a и b можно поменять местами, и, следовательно, существует 2(n—1) способов размещения a и b рядом. Каждому из этих способов соответствует (n — 2)! перестановок других элементов. Следовательно, число перестановок, в которых a и b стоят рядом, равно $2 \cdot (n - 1) \cdot (n - 2)! = 2 \cdot (n - 1)!$. Поэтому искомое число перестановок равно

$$n! - 2 \cdot (n - 1)! = (n - 1)! \cdot (n - 2).$$

Задача 6. Сколькими способами можно расположить на шахматной доске 8 ладей так, чтобы они не могли бить друг друга?

Решение. При указанном расположении ладей на каждой вертикали и каждой горизонтали стоит лишь одна ладья. Рассмотрим одно из таких расположений ладей. Пусть a_1 — номер вертикали, в которой стоит ладья из первой горизонтали, a_2 — номер вертикали, в которой стоит ладья из второй горизонтали, ... a_8 — номер вертикали, в которой стоит ладья из последней, восьмой, горизонтали. Тогда (a_1 ..., a_8) есть некоторая перестановка чисел 1, ... 8. Среди чисел a_1, ... a_8 нет ни одной пары равных, иначе 2 ладьи попали бы в одну вертикаль. Следовательно, каждому расположению ладей соответствует определенная перестановка чисел 1, ..., 8. Наоборот, каждой перестановке чисел 1, ..., 8 соответствует такое расположение ладей, при котором они не бьют друг друга. Следовательно, число искомых расположений ладей равно $P_8 = 8! = 40\,320$.

Задача 7. Студенту необходимо сдать 4 экзамена на протяжении 8 дней. Сколькими способами это можно сделать?

Решение. Искомое число способов равно числу 4-элементных упорядоченных подмножеств (дни сдачи экзаменов) множества из 8 элементов, т.е. $A^4_8 = = 8 \cdot 7 \cdot 6 \cdot 5 = 1680$ способов. Если известно, что

последний экзамен будет сдаваться на восьмой день, то число способов равно

$$4 \cdot A^3_7 = 7 \cdot 6\text{-}5 \cdot 4 = 840.$$

Задача 8. Сколькими способами можно рассадить 4 студента на 25 местах?

Решение. Искомое число способов равно числу размещений из 25 по 4:

$$A^4_{25} = 25 \cdot 24 \cdot 23 \cdot 22 = 303600.$$

Задача 9. Число слов, которые можно составить из 12 букв (4 буквы *а,* 4 буквы *б,* 2 буквы *в,* 2 буквы *г),* равно

$$\frac{12!}{4! \cdot 4! \cdot 2! \cdot 2!} = 207\,900.$$

Задача 10. Сколько целых неотрицательных решений имеет уравнение

$$x_1 + x_2 + \ \ldots \ + x_m = n?$$

Существует тесная связь между решениями указанного уравнения и сочетаниями из m элементов по n. Если имеем целые неотрицательные числа $x_1, \ldots \ldots, x_m$ такие, что $x_1 + \ldots + x_m = n$, то можем составить сочетание из m элементов по n, взяв x_1 элементов первого типа, x_2 — друго типа, ..., x_m — m-го типа. Наоборот, имея сочетание из m элементов по n, получим решение уравнения $x_1 + \ldots + x_m = n$ (x_1 — число элементов первого типа, x_2 — число элементов второго типа, ..., x_m — число элементов m-го типа) в целых неотрицательных числах. Следовательно, между множеством всех сочетаний из m элементов по n с повторениями и множеством всех целых неотрицательных решений уравнения $x_1 + \ldots + x_m = n$ устанавливается взаимно однозначное соответствие. Поэтому число решений равно

$$\overline{C}^n_m = C^n_{m+n-1}.$$

Задача 11. Пусть имеем множество $X = = \{x_1, \ldots . x_k\}$, состоящее из k элементов, и множество $Y = \{y_1, \ldots, y_n\}$, что состоящее из n элементов. Предположим, что каждому элементу множества X поставлен в соответствие некоторый элемент множества Y. Тогда говорят, что на *множестве X задана функция с областью значений Y*. Множество X называют *областью определение функции.* Естественно возникнет вопрос: сколько всего имеется различных функций с областью определения X и областью значений Y?

Каждую функцию можно задать таблицей значений:

x_1	x_2	. . .	x_s	. . .	x_k
y_{i_1}	y_{i_2}	. . .	y_{i_s}	. . .	y_{i_k}

где y_{i_s} — это тот элемент множества Y, который поставлен в соответствие x_s. Поскольку каждый из элементов y_{i_1}, ... y_{i_k} может быть одним из элементов множества Y, то всего есть n^k различных таблиц. (Можно рассждать еще и так: нижняя строка таблицы (y_{i_1} , ... y_{i_k}) есть k -слово, составленное из элементов множества Y, а, как известно, различных k-слов, составленных из элементов множества Y таково, что $N(Y)=n$, имеется n^k.) Следовательно, имеется n^k различных функций с областью определения X и областью значений Y.

Задача 12. Доказать, что ;

$$C_n^0 + C_n^m + C_n^{2m} + \ldots = \frac{2^n}{m} \sum_{k=1}^{m} \cos^n \frac{k\pi}{m} \cos \frac{nk\pi}{m} \tag{5.20}$$

(сумма в левой части вычисляется до тех пор, пока верхний индекс не станет больше нижнего).

Доказательство. Пусть $\varepsilon_1, \ldots, \varepsilon_m$ — корени уравнения $x^m - 1 = 0$, т. е.

$$\varepsilon_k = \cos \frac{2k\pi}{m} + i \sin \frac{2k\pi}{m} \qquad (k = 1, 2, \ldots, m). \tag{5.21}$$

Заметим, что для каждого натурального r

$$\varepsilon^r{}_1 + \ldots + \varepsilon^r{}_m = \begin{cases} m, \text{ если } r \text{ делится на } m, \\ 0, \text{ если } r \text{ не делится на } m. \end{cases} \tag{5.22}$$

Действительно, согласно (4.21), $\varepsilon_k = \varepsilon^k{}_1$, и поэтому

$$\varepsilon_1^r + \ldots + \varepsilon_m^r = \varepsilon_1^r + \varepsilon_1^{2r} + \ldots \varepsilon_1^{(m-1)r} + 1. \tag{5.23}$$

Если r делится на m, то все слагаемые в правой части равенства (5.23) равные 1. Если же r не делится на m ($r = qm+p$, $1 \le p \le m - 1$), то

$$\varepsilon_1^r = \varepsilon_1^{qm+p} = \varepsilon_1^p = \varepsilon_p \neq 1$$

и

$$1 + \varepsilon_1^r + \varepsilon_1^{2r} + \ldots + \varepsilon_1^{(m-1)r} = \frac{1 - (\varepsilon_1^r)^m}{1 - \varepsilon_1^r} = \frac{1 - \varepsilon_1^{rm}}{1 - \varepsilon_1^r} = 0,$$

поскольку ε^m_1==1. Равенство (5.22) доказано.

Из равенства (5.22) следует, что

$$(1+\varepsilon_1)^n + \ldots + (1+\varepsilon_m)^n = \sum_{r=0}^{n} C_n^r \left(\varepsilon_1^r + \ldots + \varepsilon_m^r\right) = \\ = m\left(C_n^0 + C_n^m + C_n^{2m} + \ldots + C_n^{m\left[\frac{n}{m}\right]}\right). \tag{5.24}$$

Поскольку

$$(1+\varepsilon_1)^n + \ldots + (1+\varepsilon_m)^n = \\ = \sum_{k=1}^{m} 2^n \cos^n \frac{k\pi}{m}\left(\cos\frac{nk\pi}{m} + i\sin\frac{nk\pi}{m}\right).$$

то

$$1+\varepsilon_k = 1 + \cos\frac{2\pi k}{m} + i\sin\frac{2\pi k}{m} = \\ = 2\cos\frac{k\pi}{m}\left(\cos\frac{k\pi}{m} + i\sin\frac{k\pi}{m}\right), \tag{5.25}$$

Из равенства (5.24) следует, что правая часть в (5.25) - действительное число. Поэтому

$$\sum_{k=1}^{m} 2^n \cos^n \frac{k\pi}{m} \sin\frac{nk\pi}{m} = 0. \tag{5.26}$$

Сравнивая (5.24) и (5.25), получим (5.20).

Задача 14. Доказать, что

$$C_n^m C_k^0 + C_{n-1}^{m-1} C_{k+1}^1 + \ldots + C_{n-m}^0 C_{k+m}^m = C_{n+k+1}^m. \tag{5.27}$$

Доказательство. Рассмотрим все кратчайшие пути, которые ведут из точки (0; 0) в точку *(n — m+k+1; m)*. Разобьем все такие пути на классы L_0, L_1, ..., L_m, отнеся к классу L_r все те пути, которые пересекают прямую $x = k + 0{,}5$ в точке $(k + 0{,}5; r)$ (рис. 5.7).

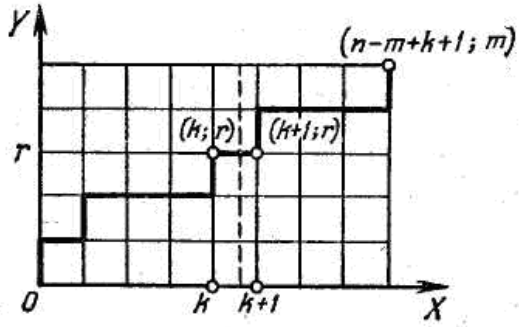


Рис. 5.7

Поскольку каждую ломаную из L_r можно разбить на 3 части (ломаную, соединяющую (0; 0) с *(k;r),* горизонтальный отрезок, соединяющий точки *(k; r)* и (*k*+1; *r),* и ломаную, соединяющую *(k*+1; *r*) с *(n — m + k* + 1; *m)*, то общее число ломаных, из которых состоит класс L_r, равно

$$C^r_{k+r}C^{m-r}_{n-r}.$$

Общее же число всех путей из точки (0; 0) в точку *(n — m*+*k*-+1; *m)* равно C^m_{n+k+1}. Поэтому имеет место (5.27).

Микромодуль 16
Основные тестовые задачи

Упражнение 1.

1. На вершину горы ведет 7 дорог. Сколькими способами турист может подняться на гору и спуститься с нее? Дайте ответ на тот же самый вопрос, если подъем и спуск осуществляются различными путями.
2. Сколько трехзначных чисел можно сложить из цифр 1, 2, 3, 4, 5?
3. Сколько трехзначных чисел можно составить из цифр 1, 2, 3, 4, 5, если каждую из этих цифр можно использовать не более одного раза?
4. Сколькими способами 7 человек могут разместиться в очереди в кассу?

5. В группе изучают 10 предметов. В понедельник 4 пары, причем все пары разные. Сколькими способами можно составить расписание на понедельник?
6. Сколько имеется пятизначных чисел, которые делятся на 5?
7. На одной из боковых сторон треугольника взято *n* точек, на другой — *m* точек. Каждая из вершин при основании треугольника соединена прямыми с точками, взятыми на противоположной стороне.
а) Сколько точек пересечения этих прямых образуется внутри треугольника? б) На сколько частей делят треугольник эти прямые?
8. Сколько есть двузначных чисел, у которых обе цифры четные?
9. Сколько есть пятизначных чисел, у которых все цифры нечетные?
10. Сколько четырехзначных чисел можно написать с помощью цифр 0, 1, 2, 3, 4, 5? Найти сумму всех этих чисел.
11. Сколько есть трехзначных чисел, которые записываются с помощью цифр 0, 1, 2, 3, 4, 5 и делятся на 3?
12. Сколько есть пятизначных чисел, которые одинаково читаются слева направо и справа налево (например, таких, как 67876, 17071)?
13. 5 ребят и 5 девушек садятся в ряд на 10 расположенных подряд стульев, причем ребята садятся на места с нечетными номерами, а девушки - на места с четными номерами. Сколькими способами это можно сделать?
14. Сколько разных слов можно составить перестановкой букв в слове «математика»?
15. Автомобильные номера состоят из одной, двух или трех букв и четырех цифр. Найти число таких номеров, используя 32 буквы русского алфавита,
16. В поселке живут 1500 жителей. Доказать, что по крайней мере двое из них имеют одинаковые инициалы.
17. а) Сколько разных делителей имеет число $3^5 \times 5^4$?
б) Пусть $p_1, ..., p_n$ - различные простые числа. Сколько делителей имеет число

$$m = p_1^{\alpha_1} \times ... \times p_n^{\alpha_n}$$

где $\alpha_1, ..., \alpha_r$ - некоторые натуральные числа?
18. От *A* до *B* 999 *км*. Вдоль дороги стоят столбы, на которых указанные расстояния до *A* и до *B*

0,999 ; 1,998; 2,997 ...; 999,0.

Сколько среди них таких, на которых есть только 2 различные цифры?

19. Пассажир оставил вещи в автоматической камере хранения, а когда пришел получать вещи, выяснилось, что он забыл номер. Он только помнит, что в номере были числа 23 и 37. Чтобы открыть камеру, нужно правильно набрать пятизначный номер. Какое наибольшее количество номеров нужно перебрать, чтобы открыть камеру?

20. В прямоугольной таблице из m строк и n столбцов записаны числа +1 и —1 так, что произведение чисел в каждой строке и каждом столбце равно 1. Сколькими способами это можно сделать?

Упражнение 2

1. *Разностью* множеств A и B (обозначается $A — B$) называется множество тех элементов A, которые не принадлежат B. Доказать соотношение

$$(A \cup B) - B = A - B,$$
$$A \cap (B \cup C) = A - (A - B) \cap (A - C),$$
$$(A - B) \cap C = (A \cap C) - (B \cap C),$$
$$A \cap B \cap C = A - (A - (B \cap C)),$$
$$(A \cap C) - B = (A \cap C) - (B \cap C),$$
$$(A - B) \cup (A - C) = A - (B \cap C).$$

2. Доказать, что $A — B = \varnothing$ тогда и только тогда, когда $A \cap B = A$

3. Пусть A — множество корней уравнения $x^2 — 3x + 2 = 0$, а $B = \{0, 2\}$. Найти $A \cap B$, $A \cup B$, $A — B$, $B — A$.

4. Пусть A — множество значений функции

$$y = \operatorname{sign} x = \begin{cases} 1 & \text{при } x > 0, \\ 0 & \text{при } x = 0, \\ -1 & \text{при } x < 0, \end{cases}$$

а B — множество корней уравнения $x(x—1)(x + 2)=0$.

Найти $A \cap B$, $A \cup B$, $A — B$, $B — A$.

5. Из 100 студентов английский язык знают 28 студентов, немецкий — 30, французский — 42, английский и немецкий — 8, английский и французский—10, немецкий и французский — 5, все три языки знают 3 студента. Сколько студентов не знают ни одного из трех языков?

Упражнение 3.

1. Сколькими способами из 30 студентов можно выбрать делегацию, которая состоит из 3 студентов?

2. В комнате n лампочек. Сколько всего разных способов освещения комнаты, при которых горит ровно k лампочек? Сколько всего может быть разных способов освещения комнаты?
3. Дано n точек, никакие 3 из которых не лежат на одной прямой. Сколько прямых можно провести, соединяя точки попарно?
4. На плоскости проведено n прямых так, что никакие 2 из них не параллельны и никакие 3 не пересекаются в одной точке. а) Найти количество точек пересечения этих прямых; б) Сколько треугольников образуют эти прямые? в) На сколько частей делят плоскость эти прямые? г) Сколько среди них ограниченных частей и сколько неограниченных?
5. Сколько есть четырехзначных чисел, в которых каждая следующая цифра большая предыдущей?
6. Сколько есть четырехзначных чисел, в которых каждая следующая цифра меньшая предыдущей?
7. Международная комиссия состоит из 9 человек. Материалы комиссии сохраняются в сейфе. Сколько замков должен иметь сейф, сколько ключей для них нужно изготовить и как их распределить между членами комиссии, чтобы доступ к сейфу был возможен тогда и только тогда, когда соберутся не менее 6 членов комиссии?
Рассмотреть задачу также в том случае, когда комиссия состоит из n членов комиссии, а сейф можно открыть при наличии m членов комиссии.
8. Имеется p белых и q черных шаров. Сколькими способами можно выложить в ряд все шары так, чтобы никакие 2 черных шара не лежали рядом?
9. В выпуклому n-угольнике проведены все диагонали. Известно, что никакие 3 из них не пересекаются в одной точке. На сколько частей разделится при этом многоугольник?

Упражнение 4.

1. Сколькими способами можно упорядочить множество $\{1, 2, ..., n\}$ так, чтобы числа 1, 2, 3 стояли рядом и в порядке возрастания?
2. Сколькими способами могут разместиться 5 покупателей в очереди в кассу?
3. Сколько существует перестановок из n элементов, в которых между двумя данными элементами стоит r элементов?
4. На собрании должны выступить 4 человек A, B, C, D. Сколькими способами их можно разместить в списке ораторов, если B не может выступать до того момента, пока не выступит A?

5. Сколькими способами можно рассадить n гостей за круглым столом?
6. Сколькими способами можно упорядочить множество {1,2, ..., n) так, чтобы каждое число, кратное 2, и каждое число, кратное 3, имело номер, кратный 2 и 3?
7. Если повернуть лист белой бумаги на 180°, то цифры 0, 1, 8 не изменяются, цифры 6 и 9 переходят одна в одну, а остальные цифры теряют смысл. Сколько существует семизначных чисел, величина которых не изменяется при повороте листа бумаги на 180°?
8. В розыгрыше первенства страны по футболу в высшей лиге принимает участие 10 команд. Команды, которые займут первое, второе и третье места, награждаются соответственно золотой, серебряной и бронзовой медалями, а команды, которые займут последние 2 места, покинут высшую лигу. Сколько разных результатов первенства может быть?

Упражнение 5.

1. Сколько различных слов можно составить, переставляя буквы слова «мама»? Напишите все эти слова.
2. Сколькими способнми можно разделить $m+ n + s$ предметов на 3 группы так, чтобы в одной группе было m предметов, в другой — n предметов, и третьей — s предметов?
3. Сколькими способами можно разделить $3n$ различных предметов между тремя людьми так, чтобы каждый человек получил n предметов?
4. Сколько пятибуквенных слов можно составить из букв *a, b, c,* если известно, что буква a встречается в слове не более двух раз, буква b — не более одного раза, буква c — не более трех раз?
5. Сколько различных слов можно составить, переставляя буквы слова «комбинаторика»?

Упражнение 6.

1. Напишите все соединения с повторениями из трех элементов *a, b, c* по 3.
2. Сколькими способами можно выбрать 6 одинаковых или разных пирожных в кондитерской, где есть 11 разных сортов пирожных?
3. Сколько можно сделать костей домино, используя числа 0, 1, ... r?
4. Сколько целых положительных решений имеет уравнение

$$x_1 + \quad ... \quad + m = n?$$

5. Сколько целых неотрицательных решений имеет неравенство

$$x_1 + \quad ... \quad + m \leq n?$$

Упражнение 7.

1. Пусть $A = \{a, b\}$, $B = \{a, b, c\}$. Указать все элементы множества $A \times B$.

2. Пусть $A = [0, 1)$, $B = (0, 1) \cup [2, 3]$. Изобразить в декартовой системе координат XOY множество $A \times B$.

Упражнение 8.

1. Доказать формулу бинома Ньютона, применяя метод математической индукции.

2. а) Доказать, что $C_n^{k+1} > C_n^k$ при $k < \frac{n-1}{2}$ и $C_n^{k+1} < C_n^k$ при $k > \frac{n-1}{2}$.

б) Указать наибольшее среди чисел C^k_n ($k = 0, 1, ..., n$).

3. Найти n, если известно, что в разложении $(1 + x)^n$ коэффициенты при x^5 и x^{12} равны.

4. Сколько рациональных членов содержит разложение

$$\left(\sqrt{2} + \sqrt[4]{3}\right)^{100}?$$

5. Пользуясь полиномиальной теоремой, вычислить $(x + y + z)^3$

6. Чему равен коэффициент при $x^2y^3z^2$ в выражении $(x + y + z)^7$

7. Найти коэффициент при $x^k y^r$ в разложении $(1 + x + y)^n$.

8. Найти коэффициенты при x^{17} и x^{18} в разложении $(1 + x^5 + x^7)^n$.

9. Доказать, что $C_{n+1}(i, l, k) = C_n(i\text{-}1, l, k) + C_n(i, l\text{-}1, k) + C_n(i, l, k\text{-}1)$.

10. Сколько членов содержит полиномиальное разложение (формула (4.11) в п. 4.8.2)?

11. Доказать, что сумма всех коэффициентов полиномиального разложения равна k^n.

12. Доказать, что числа $C^1_p, C^2_p, ..., C^{p-1}_p$ делятся на p, если p — простое число.

13. Доказать, что разность $a^p - a$ при любом целом a делится на p, если p — простое число *(малая теорема Ферма)*.

14. Доказать, что разность $[(2 + \sqrt{5})^p] - 2^{p+1}$ делится на n, если p — простое число ($p > 2$). (Символ $[x]$ обозначает целую часть x.)

15. Обозначим $a(a - h)(a - 2h)(a - 3h) \ldots (a - (n - 1)h) = a^{n/h}$ (в частности, $a^{n/0} = a^n$). Доказать, что

$$(a + b)^{n/h} = a^{n/h} + C_n^1 a^{(n-1)/h} b + \ldots + b^{n/h}.$$

Доказать тождества:

16. $C_n^0 + \frac{1}{2} C_n^1 + \ldots + \frac{1}{n+1} C_n^n = \frac{2^{n+1}-1}{n+1}$.

17. $C_n^1 + 2C_n^2 + \ldots + nC_n^n = n2^{n-1}$.

18. $C_n^1 - 2C_n^2 + \ldots + (-1)^{n-1} nC_n^n = 0$.

19. $C_n^0 + C_n^2 + \ldots + C_n^{2[n/2]} =$

$$= C_n^1 + C_n^3 + \ldots + C_n^{2[(n+1)/2]-1} = 2^{n-1}.$$

20. $C_{n-2}^{k-2} + 2C_{n-3}^{k-2} + \ldots + (n-k+1) C_{k-2}^{k-2} = C_n^k$.

21. $C_{k-3}^{k-3} C_{n-k+2}^2 + C_{k-2}^{k-3} C_{n-k+1}^2 + \ldots + C_{n-3}^{k-3} C_2^2 = C_n^k$.

22. $C_n^k + C_{n+1}^k + \ldots + C_{n+m}^k =$

$$= \begin{cases} C_{n+m+1}^{k+1} - C_n^{k+1} & \text{при } k \leqslant n-1, \\ C_{n+m+1}^{n+1} & \text{при } k = n. \end{cases}$$

23. $C_n^r + C_n^{r+m} + C_n^{r+2m} + \ldots =$

$$= \frac{2^n}{m} \sum_{k=1}^{n} \cos^n \frac{k\pi}{m} \cos \frac{(n-2r)\, k\pi}{m} \quad \text{при } m > 1.$$

Вычислить суммы:

24. $C_n^0 - C_n^1 + C_n^2 - \ldots + (-1)^m C_n^m$.

25. $C_{2n}^0 - C_{2n-1}^1 + C_{2n-2}^2 - \ldots + (-1)^n C_n^n$.

26. $(C_n^0)^2 - (C_n^1)^2 + (C_n^2)^2 - \ldots + (-1)^n (C_n^n)^2$.

27. $C_n^0 + C_n^2 + C_n^4 + \ldots$

28. $C_n^1 + C_n^3 + C_n^5 + \ldots$

29. $C_n^0 + C_n^4 + C_n^8 + \ldots$

30. $C_n^3 + C_n^7 + C_n^{11} + \ldots$

31. $C_n^2 + C_n^5 + C_n^8 + \ldots$

32. $1 + C_n^1 \cos\varphi + C_n^2 \cos 2\varphi + \ldots + C_n^n \cos n\varphi$.

33. $C_n^1 \sin\varphi + C_n^2 \sin 2\varphi + \ldots + C_n^n \sin n\varphi$.

34. Найти все корени уравнения

$$1 - \frac{x}{1!} + \frac{x(x-1)}{2!} - \ldots + (-1)^n \frac{x(x-1)\ldots(x-n+1)}{n!} = 0.$$

35. Пусть α, β, a — натуральные числа, $\alpha + \beta < m$, $\alpha + \beta < n$. Доказать, что

$$\sum_{k=0}^{\alpha-1} C_{a+m-k}^{m-k} C_{n-a-1+k}^{k} + \sum_{k=1}^{\beta} C_{m+a-\alpha+1}^{a+k} C_{n+\alpha-a-1}^{n-a-k} + \\ + \sum_{k=0}^{m-\alpha-\beta} C_{n+m-\alpha-\beta-1-k}^{m-k} C_{\alpha+\beta+k}^{k} = C_{m+n}^{m}.$$

36. В группе изучают $2n$ предметов. Все студенты учатся на 4 и 5. Никакие 2 из них не учатся одинаково, ни о каких двух из них нельзя сказать, что один из них учится лучше другого. Доказать, что число студентов в группе не превышает $C^{n}_{2n.}$

37. Пусть Q — некоторое множество, которое содержит n элементов, а A_1, ..., A_k — подмножества этого множества. Набор подмножеств A_1.....A_k будем называть *коллекцией Шпернера,* если ни одно из множеств A_1, ..., A_k не является частью другого.

а) Пусть $Q = \{a, b, c\}$. Какие из указанных ниже наборов являются коллекциями Шпернера:

$$K_1 = [\{a\}, \{b\}, \{c\}],$$
$$K_2 = [\{a, b\}, \{a, c\}, \{b, c\}],$$
$$K_3 = [\{a\}, \{a, c\}, \{b, c\}]?$$

б) Пусть $Q = \{a, b\}$. Указать все возможные коллекции Шпернера этого множества.

38. *(Теорема Шпернера.)* Пусть Q — множество, которое состоит из n элементов, A_1, ..., A_k — коллекция Шпернера этого множества. Тогда

$$k \leqslant C_n^{[n/2]}.$$

Доказать это.

39. Пусть Q — множество, которое состоит из n элементов, A_1, ..., A_k — коллекция Шпернера этого множества, i_1, ..., i_k — соответственно числа элементов множеств A_1, ..., A_k. Доказать, что

$$\frac{1}{C_n^{i_1}} + \frac{1}{C_n^{i_2}} + \ \dots \ + \frac{1}{C_n^{i_k}} \leqslant 1.$$

40. Получить из утверждение задачи 39 утверждение задачи 38.

41. Пусть A_n - число различных коллекций Шпернера для множества из n элементов. Доказать, что

$$2^{T_n} < A_n < C^{T_n}_{2^{T_n}}.$$

где

$$T_n = C_n^{[n/2]}.$$

42. Пусть x_1, ... , x_n - *действительные* числа, $|x_i| \geq 1$. Доказать, что в любом интервале длины 2 есть не более чем $C_n^{[n/2]}$ сумм вида $\sum \varepsilon_k x_k$, где $\varepsilon_k = \pm 1$.

43. Доказать тождества:

$$\text{а)}\ \sum_{k=0}^{n}(-1)^k C_n^k \frac{C_{2k}^k}{4^k} = \frac{C_{2n}^n}{4^n},$$

$$\text{б)}\ \sum_{k=0}^{n}(-1)^k \frac{C_n^k}{2k+1} = \frac{4^n}{(n+1)\,C_{2n+1}^n}.$$

Микромодуль 17
Методы комбинаторики

5.9. Метод рекурентних соотношений

Метод рекурентних соотношений заключается в том, что решение комбинаторной задачи с n предметами выражается через решение аналогичной задачи с меньшим числом предметов с помощью некоторого соотношения, которое называется *рекурентным* (повернутым). Пользуясь этим соотношением, искомую величину можно вычислить, исходя из того, что для небольшого количества предметов (одного, двух) решение задачи легко находится.

Проиллюстрируем метод рекурентних соотношений на примерах.

Пример 1. *(Соединение с повторениями.)* Соединение из n предметов по r с повторениями — это группы по r предметов, взятых из данных n предметов, причем каждый предмет может повторяться какое угодно число раз (порядок предметов в группе безразличен). Например, все сочетания из четырех чисел 1, 2, 3, 4 по два с повторениями имеют вид 11, 12, 13, 14, 22, 23, 24, 33, 34, 44. Таким образом, число сочетаний из 4 по 2 с повторениями равно 10.

Обозначим число ссочетаний из n предметов $\{a_1, \ldots, a_n\}$ по r с повторениями через $f^r{}_n$. Каждое сочетание из n по r или содержит, или не содержит a_1. Число сочетаний, которые не содержат a_1 равно $f^r{}_{n-1}$ (это сочетание из предметов $a_2, \ldots, a_n$ по r). Каждое сочетание, которое содержит a_1 может быть получено присоединениям к a_1 некоторого сочетания из n предметов по r— 1 (число таких сочетаний равно $f^{r-1}{}_n$). Следовательно,

$$f^r{}_n = f^r{}_{n-1} + f^{r-1}{}_n. \qquad (5.28)$$

Мы получили рекурентне соотношение, из которого можно найти $f^r{}_n$. Последовательно применяя (5.28), получим

$$f_n^r = f_n^{r-1} + f_{n-1}^r = f_n^{r-1} + (f_{n-1}^{r-1} + f_{n-2}^r) = \\ = f_n^{r-1} + f_{n-1}^{r-1} + \ldots f_2^{r-1} + f_1^r. \quad (5.29)$$

Очевидно,

$$f^1{}_n = n, \; f^r{}_1 = 1. \quad (5.30)$$

Считая в (5.29) $r = 2$, получим

$$f_n^2 = n + (n-1) + \ldots + 2 + 1 = \\ = \frac{n(n+1)}{2} = C_{n+1}^2. \quad (5.31)$$

При $r = 3$ из равенства (5.29) получим,

$$f_n^3 = C_{n+1}^2 + C_n^2 + \ldots + C_3^2 + C_2^2 = C_{n+2}^3$$

(мы использовали равенство (4.16) из п. 4.8.3). Повторяя эти же соображения далее, на (r- 1)-м шаге будем иметь

$$f_n^r = C_{n+r-1}^r. \quad (5.32)$$

Пример 2. Найдем число частей, на которые n окружностей делят плоскость, если каждые две окружности имеют общую хорду и никакие три окружности не пересекаются в одной точке. Пусть A_n — искомое число частей. На сколько увеличится число частей, если провести $(n + 1)$-ю окружность так, чтобы она пересекало все окружности и не проходила через точку пересечения каких-нибудь двух других окружностей? Поскольку $(n + 1)$-я окружность пересекается с каждой окружностью в двух точках, то она разделится точками пересечения на $2n$ дуг, каждая из которых делит пополам одну из частей, которая имеется в A_n. Следовательно, число частей увеличится на $2n$;

$$A_{n+1} = A_n + 2n. \quad (5.33)$$

Из этого рекуррентного соотношение получим

$$A_n = 2(n-1) + A_{n-1} = 2(n-1) + 2(n-2) + A_{n-2} = \\ = 2(n-1) + 2(n-2) + \ldots + 2 \cdot 1 + A_1.$$

Но $A_1 = 2$, и поэтому $A_n = n^2 — n + 2$.

5.10. Метод производящих функций

Метод производящих функций не является элементарным, так как при его использовании приходится иметь дело с некоторыми понятиями математического анализа. Остановимся коротко на этом

методе, поскольку он есть одним из наиболее эффективных методов решения комбинаторных задач.

Дальшее будем рассматривать суммы бесконечного числа слагаемых. В математическом анализе такие суммы называются *рядами.*

Пусть имеем бесконечную сумму

$$a_1+a_2+ \ldots + a_k + \ldots \qquad (5.34)$$

Как правило такие суммы записывают в виде

$$\sum_{k=1}^{\infty} a_k.$$

Приймем следующее определение. Пусть

$$s_n = a_1 +\ldots+a_n. \qquad (5.35)$$

Если существует $\lim\limits_{x\to\infty} s_n = s$, то ряд (5.34) называется *сходящимся* и число *s* называется *суммой* этого ряда:

$$a_1 +\ldots+a_k + \ldots =s.$$

Если же $\lim\limits_{x\to\infty} s_n$ не существует, то ряд называется *расходящимся*

5.11. Метод включения и исключение

Пусть *N(A)* - число элементов множества *A*. В 5.2 была установлена формула

$$\begin{aligned} N(A_1 \cup A_2 \cup \ldots \cup A_n) &= N(A_1) + \ldots + N(A_n) - \\ &- \{N(A_1 \cap A_2) + N(A_1 \cap A_3) + \ldots \\ &\qquad \ldots + N(A_1 \cap A_n) + \ldots + N(A_{n-1} \cap A_n)\} + \\ &+ \{N(A_1 \cap A_2 \cap A_3) + N(A_1 \cap A_2 \cap A_4) + \ldots \\ &\qquad \ldots + N(A_{n-2} \cap A_{n-1} \cap A_n)\} - \ldots \\ &\quad \ldots + (-1)^{n-1} N(A_1 \cap A_2 \cap \ldots \cap A_{n-1} \cap A_n). \end{aligned} \qquad (5.36)$$

Метод подсчета по формуле (5.36), состоящий в поочередном сложении и вычитании, называется *методом включения и исключения.* Равенство (5.36) можно записать в виде

$$N(A_1 \cup \ldots \cup A_n) =$$
$$= \sum_{1 \leqslant i_1 \leqslant n} N(A_{i_1}) - \sum_{1 \leqslant i_1 < i_2 \leqslant n} N(A_{i_1} \cap A_{i_2}) +$$
$$+ \sum_{1 \leqslant i_1 < i_2 < i_3 \leqslant n} N(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \ldots$$
$$\ldots + (-1)^{n-1} \sum_{1 \leqslant i_1 < i_2 < \ldots < i_n \leqslant n} N(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_n}),$$

где

$$\sum_{1 \leqslant i_1 < i_2 < \ldots < i_k \leqslant n} N(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k})$$

— сумма всех тех

$$N(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k})$$

у которых индексы i_1, ..., i_k удовлетворяют неравенству

$$1 \leqslant i_1 < i_2 < \ldots < i_k \leqslant n.$$

Например,

$$\sum_{1 \leqslant i_1 < i_2 \leqslant 4} N(A_{i_1} \cap A_{i_2}) = N(A_1 \cap A_2) + N(A_1 \cap A_3) +$$
$$+ N(A_1 \cap A_4) + N(A_2 \cap A_3) + N(A_2 \cap A_4) + N(A_3 \cap A_4).$$

Равенство (5.36) доказано в п. 5.2 с помощью метода математической индукции. Рассмотрим еще одно доказательство этого важного равенства.

Чтобы доказать (5.36), докататочно доказать, что каждый элемент из $A_1 \cup \ldots \cup A_n$ учитывается в правой части равенства (5.36) ровно один раз. Рассмотрим произвольный элемент a из $A_1 \cup \ldots \cup A_n$ и предположим, что a входит ровно в m множеств A_i. Тогда элемент a подсчитывается в правой части (5.36)

$$C_m^1 - C_m^2 + C_m^3 - \ldots + (-1)^{m-1} C_m^m$$

раз. Но

$$C_m^1 - C_m^2 + C_m^3 - \ldots + (-1)^{m-1} C_m^m =$$
$$= 1 - [1 - C_m^1 + C_m^2 - \ldots + (-1)^m C_m^m] =$$
$$= 1 - (1-1)^m = 1.$$

Следовательно, каждый элемент a из $A_1 \cup \ldots \cup A_n$ учитывается в правой части равенства (5.36) один раз. Это и доказывает равенство (5.36).

Пример 1. Рассматриваются все перестановки n чисел 1, 2, ..., n. Пусть D_n — число тех перестановок, в которых по крайней мере одно число стоит на своем месте. Найти D_n.

Обозначим через A_k совокупность тех перестановок, в которых на k-м месте стоит k. Тогда

$$D_n = N(A_1 \cup A_2 \cup \ldots \cup A_n).$$

Множество $A_{i_1} \cap \ldots \cap A_{i_k}$ содержит те перестановки, в которых на местах $i_1, \ldots, i_k$ стоят числа $i_1, \ldots, i_k$, а на других местах — другие $n - k$ чисел, которые упорядочены произвольно. Поэтому

$$N(A_{i_1} \cap \ldots \cap A_{i_k}) = (n-k)!,$$

а

$$\sum_{1 \leqslant i_1 < i_2 < \ldots < i_k \leqslant n} N(A_{i_1} \cap \ldots \cap A_{i_k}) = C_n^k (n-k)! = \frac{n!}{k!}.$$

Из равенства (5.36) следует, что

$$N(A_1 \cup \ldots \cup A_n) = n!\left(\frac{1}{1!} - \frac{1}{2!} + \frac{1}{3!} - \ldots + (-1)^{n-1}\frac{1}{n!}\right).$$

Пример 2. Пусть $a_1, \ldots, a_n$ — взаимно простые натуральные числа, а N — некоторое натуральное число. Найти число натуральных чисел, которые не превышают N и не делятся ни на одно из чисел $a_1, \ldots, a_n$.

Пусть A_i — множество натуральных чисел, которые не превышают N и делятся на a_1. Тогда количество чисел, которые делятся по крайней мере на одно из чисел $a_1, \ldots, a_n$, равно

$$N(A_1 \cup \ldots \cup A_n).$$

Очевидно,

$$N(A_i) = \left[\frac{N}{a_i}\right],$$

где $[x]$ — наибольшее целое число, не превосходящее x.
Множество

$$A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}$$

— это множество тех чисел, которые делятся на $a_{i_1}, \ldots, a_{i_k}$.

Поскольку числа $a_{i_1},...,a_{i_k}$ взаимно простые, то

$$N(A_{i_1} \cap \ldots \cap A_{i_k}) = \left[\frac{N}{a_{i_1} \ldots a_{i_k}}\right].$$

В силу равенства (5.36) количество чисел, которые не превышают N и делятся по крайней мере на одно из чисел $a_1, .., a_n$, равно

$$N(A_1 \cup \ldots \cup A_n) = \sum_{1 \leqslant i \leqslant n} \left[\frac{N}{a_i}\right] - \sum_{1 \leqslant i_1 < i_2 \leqslant n} \left[\frac{N}{a_{i_1} a_{i_2}}\right] +$$
$$+ \sum_{1 \leqslant i_1 < i_2 < i_3 \leqslant n} \left[\frac{N}{a_{i_1} a_{i_2} a_{i_3}}\right] - \ldots + (-1)^{n-1} \left[\frac{N}{a_1 \ldots a_n}\right].$$

Количество чисел, которые не превышают N и которые не делятся ни на одно из чисел $a_1,...,a_n$, равно

$$N - N(A_1 \cup \ldots \cup A_n) =$$
$$= N - \sum_{1 \leqslant i \leqslant n} \left[\frac{N}{a_i}\right] + \sum_{1 \leqslant i_1 < i_2 \leqslant n} \left[\frac{N}{a_{i_1} a_{i_2}}\right] -$$
$$- \sum_{1 \leqslant i_1 < i_2 < i_3 \leqslant n} \left[\frac{N}{a_{i_1} a_{i_2} a_{i_3}}\right] + \ldots + (-1)^n \left[\frac{N}{a_1 \ldots a_n}\right]. \tag{5.37}$$

Пример 3. Пусть n — натуральное число, разложение которого на простые множители имеет вид

$$n = p_1^{\alpha_1} \ldots p_k^{\alpha_k}$$

($p_1,...,p_k$— простые числа), а $\varphi(n)$ -число составленных натуральных чисел, которые не превышают n и взаимно простых с n. Доказать, что

$$\varphi(n) = n\left(1 - \frac{1}{p_1}\right) \ldots \left(1 - \frac{1}{p_k}\right),$$

(Функция $\varphi(n)$ называется *функцией Эйлера*).

Числа, взаимно простые с n, не делятся ни на одно из чисел $p_1,...,p_k$. Поэтому в силу (5.37)

$$\varphi(n) = n - \sum_{1 \leqslant i \leqslant k} \frac{n}{p_i} + \sum_{1 \leqslant i_1 < i_2 \leqslant k} \frac{n}{p_{i_1} p_{i_2}} - \ldots$$
$$\ldots + (-1)^k \frac{n}{p_1 \ldots p_k} =$$
$$= n\left(1 - \frac{1}{p_1}\right)\left(1 - \frac{1}{p_2}\right) \ldots \left(1 - \frac{1}{p_k}\right).$$

Рассмотрим теперь еще один способ применения метода включения и исключение.

Теорема. *Пусть* $N_{[m]}(A_1 \cup \ldots \cup A_n)$ *число элементов, входящих ровно в m множеств с* $A_1, \ldots, A_n$.
Тогда

$$N_{[m]}(A_1 \cup \ldots \cup A_n) = \\ = C_m^m \sum_{1 \leqslant i_1 < i_2 < \ldots < i_m \leqslant n} N(A_{i_1} \cap \ldots \cap A_{i_m}) - \\ - C_{m+1}^m \sum_{1 \leqslant i_1 < i_2 < \ldots < i_{m+1} \leqslant n} N(A_{i_1} \cap \ldots \cap A_{i_{m+1}}) + \ldots \\ \ldots + (-1)^{n-m} C_n^m \sum_{1 \leqslant i_1 < i_2 < \ldots < i_n \leqslant n} N(A_{i_1} \cap \ldots \cap A_{i_n}). \tag{5.38}$$

Доказательство. Пусть a — произвольный элемент, который входит в k множеств из множеств $A_1, \ldots, A_n$. Для доказательства теоремы достаточно показать, что элемент a учитывается в правой части равенства (5.38) один раз, если $k = m$, и не учитывается ни разу, если $k \neq m$. Если $k < m$, то a не учитывается в сумме (5.38) ни разу; если $k = m$, то a учитывается в (5.38) один раз, так как a входит лишь в одно из множеств вида

$$A_{i_1} \cap \ldots \cap A_{i_k}.$$

Пусть теперь $k > m$. Элемент a учитывается $C^m{}_k$ раз в сумме

$$\sum_{1 \leqslant i_1 < i_2 < \ldots < i_m \leqslant k} N(A_{i_1} \cap \ldots \cap A_{i_m}),$$

$C^{m+1}{}_k$ раз в сумме

$$\sum_{1 \leqslant i_1 < i_2 < \ldots < i_{m+1} \leqslant k} N(A_{i_1} \cap \ldots \cap A_{i_{m+1}}),$$

и т.д. C_k^k раз в сумме

$$\sum_{1 \leqslant i_1 < \ldots < i_k \leqslant k} N(A_{i_1} \cap \ldots \cap A_{i_k}).$$

В остальных суммах в правой части (5.38) элемент a не учитывается, так как он входит лишь в k множеств. Таким образом, a учитывается в (5.38),

$$C_m^m C_k^m - C_{m+1}^m C_k^{m+1} + C_{m+2}^m C_k^{m+2} - \ldots \\ \ldots + (-1)^{k-m} C_k^m C_k^k \tag{5.39}$$

раз. Остается доказать, что эта сумма равна нулю. Действительно, поскольку

$$C_r^m C_k^r = \frac{r!}{m!\,(r-m)!} \frac{k!}{r!\,(k-r)!} = C_k^m C_{k-m}^{k-r},$$

это сумма (5.39) при $k > m$ равно

$$C_k^m\left(C_{k-m}^{k-m}-C_{k-m}^{k-m-1}+\ \ldots\ +(-1)^{k-m}C_{k-m}^{0}\right)=$$
$$=C_k^m(1-1)^{k-m}=0.$$

5.12. Метод траекторий

Для многих комбинаторных задач можно указать такую геометрическую интерпретацию, которая сводит задачу к подсчету числа путей (траекторий), обладающих определенным свойством. В этом и составляется *метод траекторий.* В п. 5.3 мы пользовались методом траекторий при доказательства некоторых биномиальных тождеств. Преимуществом этого метода является чрезвычайная наглядность.

Задача 1. Возле кассы собралось $m+n$ человек, причем n из них имеют монеты стоимостью 50 коп., а другие m имеют лишь по рублю. Сначала в кассе нет денег, билет стоит 50 коп. Сколько всего меется испособов размещения $m + n$ покупателей в очереди так, чтобы ни один покупатель не ждал сдачи $(m \le n)$?

Допустим, что покупатели расположены в очереди некоторым образом. Пусть

$$\varepsilon_1=\begin{cases}1, \textit{ если } i-\textit{й покупатель имеет } 50 \textit{ коп.},\\ -1, \textit{ если } i-\textit{й покупатель имеет рубль.}\end{cases}$$

Рассмотрим

$$s_k=\varepsilon_1+\ \ldots\ +\varepsilon_k.$$

Вдумчивый читатель уже заметил, что s_k является разностью между количеством 50-копеечных монет и количеством рублей, которые поданы в кассу первыми k покупателями.

Рассмотрим теперь систему координат XOY. Построим в ней точки $A_k = (k; s_k)$ $(k=1, \ldots, m+n)$ и рассмотрим ламаную, которая соединяет точку $O =(0; 0)$ с точкой $A_{n+m} = (m + n; n — m)$ и которая проходит через точки $A_1, \ldots, A_{m+n-n}$. (рис. 5.8).

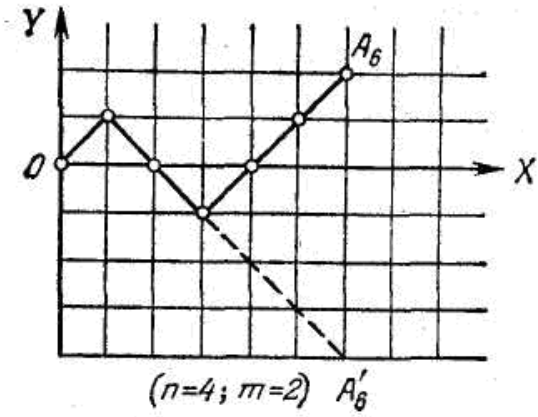


Рис. 5.8

Будем называть такую ламаную *траекторией,* соответствующей данному способу размещения покупателей в очереди. Каждая траектория состоит из $m + n$ отрезков, n из которых направлены вверх, а m направлены вниз. Если указать номера тех отрезков, которые направлены вверх, то тем самым траектория будет полностью определена. Следовательно, общее число траекторий равно $C_{n+m.}$

Траектории, соответствующие тем способам размещения покупателей, при которых ни один покупатель не ждет сдачи, не пересекают прямую y =—1. Действительно, если для некоторого k $s_{k-1} = 0$, s_k = —1, то это означает, что первые k—1 покупателей подали в кассу одинаковое количество полтинников и рублей, а k-й покупатель подал рубль и вынужден ожидать сдачу.

Определим число траекторий, которые пересекают прямую y=—1. Поставим в соответствие каждой траектории T, что пересекает прямую y = — 1 или имеющей с ней общую точку, новую траекторию T' по следующему правилу: до первой точки пересечения с прямой y=— 1 траектория T' совпадает с T, а далее T' является симметричным отображением траектории T относительно прямой y = — 1 (на рис. 4.8 траектория T' обозначена пунктирной линией). Все траектории T' заканчиваются в точке A'_{m+n} =$(m + n; m — n — 2)$, являющейся симметричным отображением точки A_{m+n} относительно прямой y= — 1. Установленное соответствие является взаимно однозначным, поэтому число траекторий, которые пересекают прямую y =—1, равно числу ламаных, которые соединяют точки O и $A'_{m+n.}$ Это число легко

подсчитать: если ламаная состоит из y отрезков, направленных вниз, и x отрезков, направленных вверх, то

$$x + y = m + n, \quad y - x = n + 2 - m,$$

откуда $y = n+1$. Таким образом, число траекторий, которые пересекают прямую $y = -1$, равно C^{n+1}_{m+n}. Искомое число траекторий равно

$$C^{n}_{m+n} - C^{n+1}_{m+n} = C^{m}_{m+n}\frac{n+1-m}{n+1}. \qquad (5.40)$$

Рассмотренная задача имеет важное значение в математической статистике, в частности в теории статистического контроля качества продукции. С ней также тесно связана так называемая задача о баллотировании, которую еще в 1887 г. рассматривал известный французский математик Бертран. Эта задача имеет интересные применения при изучении некоторых случайных процессов.

Задача 2. *(Задача о баллотировании).* Кандидат A собрал на выборах a голосов, кандидат B собрал b голосов $(a>b)$. Избиратели голосовали последовательно. Сколько существует таких способов подачи голосов, при которых A всегда будет впереди B по количеству поданных за него голосов?

Пусть $\varepsilon_i = +1$, если i-й голос подан за A, и $\varepsilon_i = = -1$, если i-й голос подан за B. Возьмем $s_k = \varepsilon_1 + \ldots + \varepsilon_k$ и рассмотрим в системе координат XOY ломаную, которая соединяет точки O, $(1;s_1),\ldots, (k;s_k), \ldots, (a + b;s_{a+b})$ (рис. 5.9).

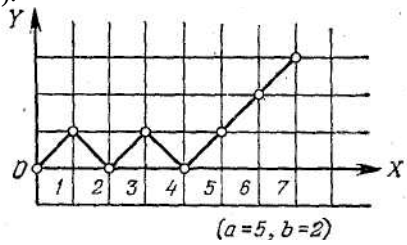


Рис. 5.9

Очевидно, $s_{a+b} = a - b$. Каждому способа подачи голосов соответствует определенная ломаная линия (траектория), соединяющая точки O и $(a + b; a - b)$. Траектория состоит из $a + b$ отрезков, причем a из них направлены вверх. Поэтому общее число траекторий равно C^{a}_{a+b}. Кандидат A всегда будет впереди B, если соответствующая траектория проходит через точку (1; 1) (первый голос должен быть подан за A) и не пересекает ось OX. Число таких траекторий может быть подсчитано по формуле (5.40), где следует

взять $n = a - 1$, $m = b$. Следовательно, искомое число способов подачи голосов равно

$$C_{a+b-1}^{a-1}\frac{a-1+1-b}{a-1+1}=\frac{a-b}{a+b}C_{a+b}^{a}. \tag{5.41}$$

Рассмотренные задачи показывают, насколько полезной может быть интерпретация задачи в терминах траекторий. Рассмотрим теперь некоторые общие утверждения, которые касаются подсчета числа траекторий.

Пусть $x>0$ и y— целые числа. *Траекторией* из начала координат в точку $(x; y)$ будем называть ламаную, которая соединяет точки O, $(1; s_1), \ldots, (k; s_k), \ldots, (x; s_x)$, где

$$s_i - s_{i-1} = \varepsilon_i = \begin{cases} +1, \\ -1, \end{cases} \quad s_x = y. \tag{5.42}$$

Пусть $N_{x, y}$ - число всех траекторий, которые соединяют точку (0; 0) с точкой $(x; y)$. Имеют место следующие теоремы.

Теорема 1,

$$N_{x, y}=\frac{x!}{\left(\frac{x+y}{2}\right)!\left(\frac{x-y}{2}\right)!},$$

если числа x и y - одинаковой четности, и

$$N_{x, y}=0,$$

если x и y — разной четности.

Доказательство. Предположим, что траектория состоит из p отрезков, направленных вверх, и q отрезков, направленных вниз (это означает, что среди чисел $\varepsilon_1 + \ldots + \varepsilon_x$ p чисел равны +1, а q чисел равны — 1). Тогда

$$p + q = x, \quad p - q = y,$$

откуда

$$p=\frac{x+y}{2}, \quad q=\frac{x-y}{2}$$

(поскольку p и q — целые числа, x и y должны быть числами одинаковой четности). Так как траектория полностью определяется, если указать, какие отрезки направлены вверх, общее число траекторий из точки O в точку $(x; y)$ равно

$$N_{x, y}=C_x^{(x+y)/2}=\frac{x!}{\left(\frac{x+y}{2}\right)!\left(\frac{x-y}{2}\right)!}.$$

Теорема 2. *(Принцип зеркального отображения). Пусть* $A = (a; \alpha)$,

B =(b; β) — *точки с челочисленными координатами, причем* $b > a \geq 0$, $\alpha > 0$, $\beta > 0$, *а* $A' = (a; -\alpha)$ — *точка, симметричная А относительно оси ОХ. Тогда число тех траекторий из А в В, которые пересекают ось ОХ или имеют с ней общую точку, равно числу траекторий из А′ в В.*

Доказательство. Каждой траектории *Т* из *А* в *В,* пересекающей ось *ОХ* или имеющей с ней общую точку, поставим в соответствие траекторию из *А*' в *В* по следующему правилу (рис. 5.10): берем участок траектории *Т* до первой точке встречи с осью *ОХ* и симметрично отображаем его относительно оси *ОХ,* а далее траектории *Т* и *Т′* совпадают. Таким образом, каждой траектории *Т* из *А* в *В,* пересекающей ось *ОХ,* соответствует определенная траектория *Т′* из *А′* в *В.*

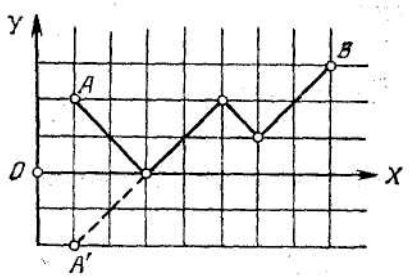


Рис. 5.10

Наоборот, каждой траектории из *А′* в *В* соответствует одна и только одна траектория из *А* в *В,* пересекающая ось *ОХ* (берем участок траектории из *А′* в *В* до первой встрече с осью *ОХ* и симметрично отображаем его относительно оси *ОХ).* Следовательно, между множеством траекторий из *А* в *В,* пересекающих ось *ОХ* или имеющих с ней общую точку, и множеством всех траекторий из *А′* в *В* установленно взаимно однозначное соответствие. Теорема доказана.

Теорема 3. *Пусть* $x > 0$, $y > 0$. *Тогда число траекторий из О в (х; у), не имеющих вершин на оси ОХ (кроме точки О), равно*

$$\frac{y}{x} N_{x,\, y}.$$

Доказательство. Все траектории, которые соединяют точку *О* с точкой *(х; у)* и не пересекающие ось *ОХ,* проходят через точку $A = (1;1)$ (рис. 5.11).

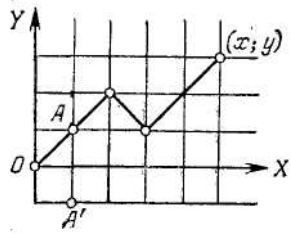


Рис. 5.11

Общее число траекторий, которые ведут из A в B, равно $N_{x-1,y-1}$. Общее число траекторий, которые ведут из A в B и пересекающих ось OX, равно, согласно теореме 2, числу траекторий, которые ведут из A' в B, т.е. $N_{x-1,y+1}$. Следовательно, искомое число траекторий равно

$$N_{x-1,\,y-1} - N_{x-1,\,y+1} = \\ = \frac{(x-1)!}{\left(\frac{x+y}{2}-1\right)!\left(\frac{x-y}{2}\right)!} - \frac{(x-1)!}{\left(\frac{x+y}{2}\right)!\left(\frac{x-y}{2}-1\right)!} = \\ = \frac{y}{x}\,\frac{x!}{\left(\frac{x+y}{2}\right)!\left(\frac{x-y}{2}\right)!} = \frac{y}{x}\,N_{x,\,y}.$$

Теорема доказана.

Установим теперь несколько свойств траекторий, которые соединяют точку O точкой $(2n; 0)$ на оси OX. Пусть

$$L_{2n} = \frac{1}{n+1}\,C_{2n}^{n}.$$

Теорема 4. *Среди C^{n}_{2n} траекторий, соединяющих точку O с точкой $(2n; 0)$, существует*

а) *ровно L_{2n-2} траекторий, лежащих выше оси OX и не имеющих общих точек с OX, кроме точек O и $(2n;0)$;*

б) *ровно L_{2n} траекторий, не имеющих вершин ниже оси OX.*

Доказательство. а) Все траектории, которые соединяют O с $(2n;0)$, лежащие выше оси OX и которые не имеют других общих точек с осью OX, обязательно проходят через точку $(2n— 1, 1)$. Согласно теореме 3 число траекторий, которые соединяют O с $(2n— 1; 1)$ и не пересекают ось OX, равно

$$\frac{1}{2n-1}\,N_{2n-1,\,1} = \frac{1}{2n-1}\,C_{2n-1}^{n} = \frac{1}{n}\,C_{2n-2}^{n-1} = L_{2n-2}.$$

б) Рассмотрим траекторию, которая соединяет O с $M(2n; 0)$ и не имеющую вершин ниже оси OX (рис. 5.12).

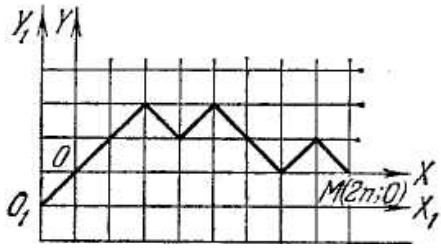


Рис. 5.12

Добавим еще один отрезок, который соединяет O с O_1(—1;—1). Примем O_1 за новое начало системы координат $X_1O_1Y_1$. В новой системе точка M имеет координаты (2n+1; 1), а точка O — координаты (1; 1). Число траекторий, которые соединяют точку O с точкой M и которые не имеют вершин ниже оси OX, равно числу траекторий, которые соединяют O_1 с M и которые лежат выше оси O_1X_1. Последнее число в силу теоремы 3 равно

$$\frac{1}{2n+1} \cdot N_{2n+1,\,1} = \frac{1}{2n+1} C_{2n+1}^{n} = \frac{1}{n+1} C_{2n}^{n} = L_{2n}.$$

Теорема доказана.

Рассмотрим пример применения доказанных выше теорем.

Задача 3. Решим задачу, рассмотренную выше (задача 1), допуская, что перед началом работы в кассе имеется p монет стоимостью 50 коп.

Очевидно, задача сводится к подсчету числа траекторий из точки O в точку $(m+n;\ n-m)$, которые не пересекают прямую $y = -(p+1)$. Согласно теореме 2 число тех траекторий, которые пересекают эту прямую, равно числу траекторий из точки $(0; -2(p+1))$ в точку $(n+m;\ n-m)$, т.е.

$$C_{m+n}^{p+n+1} = C_{m+n}^{m-p-1}.$$

Искомое число траекторий равно

$$C_{m+n}^{m} - C_{m+n}^{m-p-1}.$$

Микромодуль 17
Примеры решения типовых задач

Пример 1. Рассмотрим ряд

$$\frac{1}{1\cdot 2}+\frac{1}{2\cdot 3}+\ \dots\ +\frac{1}{k\,(k+1)}+\ \dots\ =\sum_{k=1}^{\infty}\frac{1}{k\,(k+1)}.$$

Имеем

$$s_n=\frac{1}{1\cdot 2}+\frac{1}{2\cdot 3}+\ \dots\ +\frac{1}{n\,(n+1)}=\left(1-\frac{1}{2}\right)+$$
$$+\left(\frac{1}{2}-\frac{1}{3}\right)+\ \dots\ +\left(\frac{1}{n}-\frac{1}{n+1}\right)=1-\frac{1}{n+1}.$$

Видим, что $\lim\limits_{x\to\infty} s_n=1$. Таким образом,

$$\frac{1}{1\cdot 2}+\frac{1}{2\cdot 3}+\ \dots\ +\frac{1}{k\,(k+1)}+\ \dots\ =1.$$

Пример 2. Рассмотрим ряд

$$1+\frac{1}{\sqrt{2}}+\frac{1}{\sqrt{3}}+\ \dots\ +\frac{1}{\sqrt{k}}+\ \dots\ =\sum_{k=1}^{\infty}\frac{1}{\sqrt{k}}. \qquad (1)$$

Имеем нерравенство

$$s_n=1+\frac{1}{\sqrt{2}}+\ \dots\ +\frac{1}{\sqrt{n}}>n\cdot\frac{1}{\sqrt{n}}=\sqrt{n},$$

которое показывает, что $\lim\limits_{x\to\infty} s_n=+\infty$. Таким образом, ряд (1) расходится.

Пример 3. Рассмотрим ряд

$$\sum_{k=1}^{\infty}(-1)^{k-1}=1+(-1)+1+(-1)+\dots \qquad (2)$$

Нетрудно убедиться в том, что

$$s_n=\begin{cases}0, \textit{ если } n-\textit{ четное число},\\ 1, \textit{ если } n-\textit{ нечетное число}.\end{cases}$$

Следовательно, $\lim\limits_{x\to\infty} s_n$ не существует, и ряд (2) расходится.

Пример 4. *(Сумма бесконечно убывающей геометрической прогрессии.)* В курсе алгебры при изучении геометрической прогрессии рассматривается ряд

$$1 + x + x^2 + \ldots + x^k + \ldots = \sum_{k=1}^{\infty} x^{k-1}. \qquad (3)$$

Выясним, при каких x этот ряд сходится, и вычислим его сумму. По формуле суммы членов геометрической прогрессии

$$s_n = 1 + x + x^2 + \ldots + x^{n-1} = \frac{1 - x^n}{1 - x}$$

(при $x \neq 1$). Если $|x| < 1$, то $\lim\limits_{x\to\infty} x^n = 0$ и $\lim\limits_{x\to\infty} s_n = \dfrac{1}{1-x}$. Если же $|x| > 1$, то $\lim\limits_{x\to\infty} x^n = \infty$, $\lim\limits_{x\to\infty} s_n$ не существует, и ряд (3) расходится. Ряд (3) расходится также при $x = -1$ (пример 3) и при $x — 1$ (в последнем случае $s_n = n$ и $\lim\limits_{x\to\infty} s_n = +\infty$).

Следовательно, ряд (3) сходится при $|x| < 1$, и в этом случае

$$1 + x + x^2 + \ldots + x^k + \ldots = \frac{1}{1 - x}. \qquad (4)$$

Пример 5. *(Биномиальный ряд Ньютона.)* В курсах математического анализа с помощью методов дифференциального исчисления выводится следующая формула, которая была открыта впервые Ньютоном:

$$(1 + x)^{\alpha} = 1 + \alpha x + \frac{\alpha(\alpha - 1)}{2} x^2 + \ldots + \frac{\alpha(\alpha - 1)\ldots(\alpha - k + 1)}{k!} x^k + \ldots \qquad (5)$$

Формула имеет место при $|x| < 1$. Если $a = n$ — натуральное число, то все слагаемые в правой части равенства (5), начиная с $(n + 2)$-го, равные 0, так как все они содержат множитель $(n — n)$. Формула (5) в этом случае обращается в биномиальную формулу, которую мы установили выше. При $\alpha = -1$ формула (5) принимает вид

$$\frac{1}{1 + x} = 1 - x + x^2 - \ldots + (-1)^k x^k + \ldots \qquad (6)$$

Эта формула вытекает с (4) (довольно заменить в (4) x на $-x$). Доказывать формулу (5) в общем случае не будем, а установим ее ниже лишь для целых отрицательных α.

Пример 6. *(Степенные ряды.)* Ряд вида

$$a_0 + a_x x + a_2 x^2 + \ldots + a_k x^k + \ldots \qquad (7)$$

называется *степенным*. Ряды (4.), (5), (6) являются примерами степенных рядов. В курсах математического анализа доказываются следующие важные свойства степенных рядов:

1) Областью сходимости (т.е. множеством тех x, при которых ряд сходится) является множество вида $\{x: |x| < c\}$, к которому могут иногда принадлежать точки $x = -c$ и $x = c$ или одна из этих точек.
2) Степенные ряды можно перемножать, собирая коэффициенты при одинаковых степенях x так, как это делается при умножении многочленов.
3) Если два степенных ряда имеют одинаковую сумму при всех x из области сходимости этих рядов, то коэффициенты при соответствующих степенях x этих рядов равны.

Пример 7. Выведем формулу

$$\frac{1}{(1-x)^n} = 1 + C_n^1 x + C_{n+1}^2 x^2 + \ldots + C_{n+k-1}^k x^k + \ldots \tag{8}$$

(при $|x| < 1$), которая является частным случаем биномиальной формулы (8).

Воспользуемся методом полной математической индукции. При n=1 формула (8) имеет место (формула (4)). Предположим, что (8) имеет место. Тогда

$$\frac{1}{(1-x)^{n+1}} = \frac{1}{(1-x)^n}\,\frac{1}{1-x} =$$
$$= \left(1 + C_n^1 x + C_{n+1}^2 x^2 + \ldots + C_{n+k-1}^k x^k + \ldots\right) \cdot$$
$$\cdot (1 + x + x^2 + \ldots + x^k + \ldots).$$

Перемножим эти ряды и соберем коэффициенты при одинаковых степенях x. Коэффициент при x^k равен

$$C_{n-1}^0 + C_n^1 + C_{n+1}^2 + \ldots + C_{n+k-1}^k.$$

Используя равенство

$$C_{n-1}^0 + C_n^1 + C_{n+1}^2 + \ldots + C_{n+k-1}^k = C_{n+k}^k,$$

которое вытекает из равенства (5.16) в п. 5.8.3, имеем

$$\frac{1}{(1-x)^{n+1}} = 1 + C_{n+1}^1 x + C_{n+2}^2 x^2 + \ldots + C_{n+k}^k x^k + \ldots,$$

что и доказывает формулу (8).

Определение. *Производящей функцией* последовательности $\{a_n\}$ называется сумма степенного ряда

$$A(s) = \sum_{n=0}^{\infty} a_n s^n.$$

Пример 8. Производящей функцией последовательности $a_n = a^n$

(n = 0, 1, ...) является $A(s)\dfrac{1}{1-as}$. Действительно,

$$\sum_{n=0}^{\infty} a^n s^n = \frac{1}{1-as}$$

(сумма бесконечно убывающей геометрической прогрессии; ряд в левой части сходится при $|as| < 1$).

Пример 9. Производящая функция последовательности $a_k = C_n^k$ (k = 0, 1, ..., n) равна

$$A(s) = \sum_{k=0}^{n} C_n^k s^k = (1+s)^n.$$

Пример 10. Производящая функция последовательности $a_k = C_{n+k}^k$ (k = 0, 1, ..., n) равна

$$A(s) = \frac{1}{(1-s)^{n+1}}.$$

Это следует из равенства (8).

Пример 11. Производящая функция последовательности

$$a_n = \begin{cases} 0 \text{ при } 0 \leqslant n < k, \\ C_n^k \text{ при } n \geqslant k \end{cases}$$

равна

$$A(s) = \frac{s^k}{(1-s)^{k+1}},$$

Действительно,

$$A(s) = \sum_{n=k}^{\infty} C_n^k s^n.$$

Считая в этой сумме $n = k + i$, где i = 0, 1, ..., будем иметь, принимая во внимание пример 10,

$$A(s) = s^k \sum_{i=0}^{\infty} C_{k+i}^k s^i = s^k \sum_{i=0}^{\infty} C_{k+i}^i s^i = \frac{s^k}{(1-s)^{k+1}}. \quad (9)$$

Идея применения метода производящих функций такая: нужно вычислить все члены некоторой последовательности (a_n). С помощью рекурентного соотношения для a_n или выходя непосредственно из комбинаторных соображений вычисляют производящую функцию

$$A(s) = \sum_{n=0}^{\infty} a_n s^n.$$

Раскладывая потом $A(s)$ в ряд и находя коэффициент при s^n, тем самым находят a_n.

Пример 12. *(Сочетания с повторениями.)* Пусть f^r_n — число сочетаний из n предметов по r с повторениями. Мы установили (см. пример 1 в п. 5.9), что

$$f^r_n = f^{r-1}_n + f^r_{n-1}. \tag{10}$$

При этом $f^1_n = n$; для того чтобы (10) имело место и при $r = 1$, достаточно считать, что $f^0_n = 1$. Пусть

$$A_n(s) = \sum_{r=0}^{\infty} f^r_n s^r. \tag{11}$$

Умножим обе части равенств (10) на s^r ($r = 1, 2, \ldots$) и сложим почленно все равенства; тогда получим

$$\sum_{r=1}^{\infty} f^r_n s^r = s \sum_{r=1}^{\infty} f^{r-1}_n s^{r-1} + \sum_{r=1}^{\infty} f^r_{n-1} s^r. \tag{12}$$

Но

$$\sum_{r=1}^{\infty} f^r_n s^r = A_n(s) - 1,$$

$$\sum_{r=1}^{\infty} f^{r-1}_n s^{r-1} = \sum_{t=0}^{\infty} f^t_n s^t = A_n(s) \qquad (\text{где } t = r-1),$$

$$\sum_{r=1}^{\infty} f^r_{n-1} s^r = A_{n-1}(s) - 1,$$

и, подставляя эти суммы в (12), будем иметь

$$A_n(s) = \frac{1}{1-s} A_{n-1}(s). \tag{13}$$

Отсюда

$$A_n(s) = \frac{1}{(1-s)^2} A_{n-2}(s) = \frac{1}{(1-s)^3} A_{n-3}(s) = \frac{A_1(s)}{(1-s)^{n-1}}.$$

Отметим теперь, что $f^r_1 = 1$ при всех r, и поэтому

$$A_1(s) = \sum_{r=0}^{\infty} f^r_1 s^r = \sum_{r=0}^{\infty} s^r = \frac{1}{1-s}.$$

Следовательно,

$$A_n(s) = \frac{1}{(1-s)^n}. \tag{14}$$

Из равенства (8) следует, что

$$f_n^r = C_{n-1+r}^r.$$

Пример 13. Найдем все члены последовательности Фибоначчи, которая задается по закону

$$B_0 = 1, \quad B_1, =2, \tag{15}$$

$$B_n = B_{n-1} + B_{n-2} \qquad (\text{при } n \geq 2). \tag{16}$$

Рассмотрим функцию

$$B(s) = \sum_{n=0}^{\infty} B_n s^n.$$

Умножив обе части равенства (16) на s^n и сложив потом все полученные равенства, будем иметь

$$\sum_{n=2}^{\infty} B_n s^n = s \sum_{n=2}^{\infty} B_{n-1} s^{n-1} + s^2 \sum_{n=2}^{\infty} B_{n-2} s^{n-2}. \tag{17}$$

Принимая во внимание то, что

$$\sum_{n=2}^{\infty} B_n s^n = B(s) - 1 - 2s,$$

$$\sum_{n=2}^{\infty} B_{n-1} s^{n-1} = \sum_{m=1}^{\infty} B_m s^m = B(s) - 1 \qquad (\text{где } m = n-1),$$

$$\sum_{n=2}^{\infty} B_{n-2} s^{n-2} = \sum_{p=0}^{\infty} B_p s^p = B(s) \qquad (\text{где } p = n-2),$$

получим из равенства (17)

$$B(s) - 1 - 2s = s[B(s) - 1] + s^2 B(s),$$

откуда

$$B(s) = \frac{s+1}{1-s-s^2}. \tag{18}$$

Вычислив коэффициент при s^n в разложении функции $B(s)$ в ряд, найдем B_n.

Разложим в ряд функцию

$$\frac{1}{1-s-s^2} = -\frac{1}{(s-s_1)(s-s_2)},$$

где

$$s_1 = \frac{-1-\sqrt{5}}{2}, \quad s_2 = \frac{-1+\sqrt{5}}{2}.$$

Имеем

$$-\frac{1}{(s-s_1)(s-s_2)}=\left(\frac{1}{s-s_1}-\frac{1}{s-s_2}\right)\frac{1}{s_2-s_1}=$$
$$=\frac{1}{\sqrt{5}}\left(-\frac{1}{s_1}\cdot\frac{1}{1-\frac{s}{s_1}}+\frac{1}{s_2}\cdot\frac{1}{1-\frac{s}{s_2}}\right).$$

Используя равенство

$$\frac{1}{1-x}=\sum_{n=0}^{\infty}x^n,$$

будем иметь

$$\frac{1}{1-s-s^2}=\frac{1}{\sqrt{5}}\sum_{n=0}^{\infty}\left(\frac{1}{s_2^{n+1}}-\frac{1}{s_1^{n+1}}\right)s^n,$$

Поэтому коефициент при s^n в разложении функции (18) равен

$$B_n=\frac{1}{\sqrt{5}}\left(\frac{1}{s_2^{n+1}}-\frac{1}{s_1^{n+1}}\right)+\frac{1}{\sqrt{5}}\left(\frac{1}{s_2^n}-\frac{1}{s_1^n}\right).$$

После упрощений получим формулу

$$B_n=\frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^{n+2}-\left(\frac{1-\sqrt{5}}{2}\right)^{n+2}\right).$$

Замечание. Во многих задачах производящая функция является рациональной функцией, т.е. функцией вида

$$A(s)=\frac{P(s)}{Q(s)},$$

где *P(s)* и *Q(s)*— многочлены относительно *s*. В том случае, если многочлен *Q(s)* имеет лишь простые корни, можно указать простой метод разложения такой функции в ряд по степеням *s*.

Предположим, что *Q*(*s*) - многочлен степени *m,* а степень многочлена *P(s)* меньше, чем степень *Q(s)* (этого всегда можно достигнуть, разделив *P(s)* на *Q(s)).* Тогда

$$\frac{P(s)}{Q(s)}=\sum_{k=1}\frac{A_k}{s-s_k}. \qquad (19)$$

Приведя к общему знаменателю дроби в правой части, получим в знаменателе многочлен *Q(s)* (допустим, что коэффициент при старшем члене *Q(s)* равен 1; это допущение не является ограничением), а в числителе — некоторый многочлен. Приравнивая коэффициенты этого многочлена соответствующим коэффициентам многочлена *P(s),* можно подобрать A_k так, чтобы равенство (19) выполнялось при всех тех *s,* при которых оно имеет смысл.

Иногда производящую функцию можно найти, не пользуясь рекурентними соотношениями, а исходя из некоторых комбинаторных соображений.

Пример 14. *(Производящая функция для числа сочетаний.)* Пусть a_1, a_2, a_3 — некоторые числа. Рассмотрим произведение

$$(1 + a_1 s)(1 + a_2 s)(1 + a_3 s).$$

Перемножив и собрав коэффициенты при одинаковых степенях s, будем иметь

$$\begin{aligned}(1 + a_1 s)(1 + a_2 s)(1 + a_3 s) &= \\ &= 1 + s(a_1 + a_2 + a_3) + s^2(a_1 a_2 + a_1 a_3 + a_2 a_3) + s^3 a_1 a_2 a_3 = \\ &= 1 + A_1 s + A_2 s^2 + A_3 s^3. \end{aligned} \qquad (20)$$

Отметим, что число слагаемых в каждом коэффициенте A_r равно числу сочетаний из 3 по r. Например, выписав соответствующие индексы при a_i в каждом слагаемом A_3 ((1,2), (1,3), (2,3)), получим все сочетания из трех чисел 1, 2, 3 по два.

Будем считать, что в выражении (20) $a_1=a_2=a_3=1$; тогда коэффициент при s^r будет равен числу сочетаний из 3 по r. Следовательно, $(1 + s)^3$ является производящей функцией для числа сочетаний из трех предметов.

Пример 15. Рассмотрим сочетание из трех предметов 1, 2, 3, причем 1, 2 могут встречаться не более двух раз, а 3 - не более одного раза.

Рассмотрим произведение

$$\begin{aligned}\left(1 + a_1 s + a_1^2 s^2\right)\left(1 + a_2 s + a_2^2 s^2\right)(1 + a_3 s) &= \\ = 1 + s(a_1 + a_2 + a_3) + s^2\left(a_1^2 + a_1 a_2 + a_1 a_3 + a_2 a_3 + a_2^2\right) + \\ + s^3\left(a_1^2 a_2 + a_1^2 a_3 + a_1 a_2^2 + a_1 a_2 a_3 + a_2^2 a_3\right) + \\ + s^4\left(a_1^2 a_2^2 + a_1^2 a_2 a_3 + a_1 a_2^2 a_3\right) + s^5 a_1^2 a_2^2 a_3 = \\ = 1 + A_1 s + A_2 s^2 + A_3 s^3 + A_4 s^4 + A_5 s^5.\end{aligned}$$

Опять же число слагаемых в каждом A_r равно числу всех возможных комбинаций с 3 предметов по r при сделанных выше ограничениях. Например, выписав соответствующие индексы в каждом слагаемом A_3, получим все возможные сочетания из трех чисел 1, 2, 3 по три: 112, 113, 122, 123, 223. Поэтому при $a_1= a_2 = a_3 = 1$ каждый коэффициент A_r будет равен числу соответствующих сочетаний. Следовательно, производящая функция числа всех сочетаний с указанными ограничениями равна

$$(1 + s + s^2)(1 + s + s^2)(1 + s) = (1 + s + s^2)^2 (1 + s).$$

Рассмотренные примеры дают представление о том, как написать производящую функцию для числа сочетаний при наличии других ограничений.

Так, производящая функция для числа сочетаний из n предметов с повторениями при условии, что каждый предмет встречается любое число раз, равна

$$(1 + s + s^2 + \ldots + s^k + \ldots)^n = \frac{1}{(1-s)^n}. \qquad (21)$$

Коэффициент при s^r в разложении функции (21) равен

$$f_n^r = C_{n+r-1}^r$$

(это следует из равенства (8)). Снова получили формулу для числа сочетаний с повторениями.

Производящая функция для числа сочетаний из n предметов по r при условии, что каждый предмет встречается по крайней мере 1 раз, равна

$$(s + s^2 + s^3 + \ldots)^n = \frac{s^n}{(1-s)^n}.$$

Используя равенство (8), получим

$$\frac{s^n}{(1-s)^n} = s^n \sum_{i=0}^{\infty} C_{n+i-1}^i s^i = \sum_{i=0}^{\infty} C_{n+i-1}^i s^{n+i}.$$

Положив $n+i = r$, будем иметь

$$\frac{s^n}{(1-s)^n} = \sum_{r=n}^{\infty} C_{r-1}^{r-n} s^r = \sum_{r=n}^{\infty} C_{r-1}^{n-1} s^r.$$

Коэффициент при s^r и есть искомое число сочетаний. Следовательно, число сочетаний из n предметов по r при условии, что каждый предмет встречается по крайней мере 1 раз, равно 0, если $r < n$, и равно C^{n-1}_{r-1}, если $r \geq n$.

Пример 16. Пусть A_n — число целых неотрицательных решений уравнения

$$k^1 x_1 + k_2 x_2 + \ldots + k_r x_r = n, \qquad (22)$$

где $k_1, k_2, \ldots, k_r$ - данные натуральные числа.

Обозначим

$$A(s) = \sum_{n=0}^{\infty} A_n s^n.$$

Легко видеть, что

$$A(s) = \frac{1}{(1-s^{k_1})(1-s^{k_2})\dots(1-s^{k_r})}.$$

Запишем равенства

$$\frac{1}{1-s^{k_1}} = 1 + s^{k_1} + s^{2k_1} + \dots,$$
$$\frac{1}{1-s^{k_2}} = 1 + s^{k_2} + s^{2k_2} + \dots,$$
$$\dots\dots\dots\dots\dots\dots$$
$$\frac{1}{1-s^{k_r}} = 1 + s^{k_r} + s^{2k_r} + \dots$$

Если перемножить их почленно, то показатели при степенях s будут иметь вид

$$s^{k_1x_1+k_2x_2+\dots+k_rx_r},$$

где все x_i получают натуральные значения (такой показатель получится, если в правой части каждого из написанных равенств выбрать соответственно

$$s^{k_1x_1},\ s^{k_2x_2},\ \dots,\ s^{k_rx_r}$$

и перемножить их). Число всех возможных представлений n в выражении (22) как раз и будет коэффициентом при s^n в том разложении, которое получим после перемножения всех равенств.

Рассмотрим следующую теорему о производящих функциях, которая является полезной при решении некоторых задач.

Определение. *Сверткой* двух последовательностей $\{a_n\}$ и $\{b_n\}$ называется последовательность $\{c_n\}$, общий член которой имеет вид

$$c_n = a_0b_n + a_1b_{n-1} + \dots + a_kb_{n-k} + \dots + a_nb_0.$$

Теорема. *Производящая функция свертки двух последовательностей равна произведению производящих функций этих последовательностей.*

Доказательство. Пусть $\{c_n\}$ — свертка последовательностей $\{a_n\}$ и $\{b_n\}$, а

$$A(s) = \sum_{n=0}^{\infty} a_ns^n, \quad B(s) = \sum_{n=0}^{\infty} b_ns^n, \quad C(s) = \sum_{n=0}^{\infty} c_ns^n$$

- производящие функции последовательностей $\{a_n\}$, $\{b_n\}$, $\{c_n\}$ сответственно. Требуется доказать, что

$$C(s) = A(s)\,B(s).$$

Перемножим два ряда

$$A(s) = \sum_{n=0}^{\infty} a_ns^n \quad \text{и} \quad B(s) = \sum_{n=0}^{\infty} b_ns^n.$$

Коэффициент при s^n в полученном произведении равен

$$a_0b_n + a_1b_{n-1} + \ldots + a_kb_{n-k} + \ldots + a_nb_0.$$

Следовательно,

$$A(s)\,B(s)=C(s).$$

Для иллюстрации теоремы рассмотрим следующий пример.

Пример 17. На окружности взято $2n$ точек. Сколькими способами можно соединить попарно эти точки n хордами, не пересекающимися внутри окружности?

Обозначим через a_n число способов, которыми можно разбить на пары $2n$ точек, взятых на окружности, так, чтобы хорды, которые соединяют эти пары точек, не пересекались. Обозначим точки буквами в таком порядке, в котором они размещены на окружности: $A_1, A_2, \ldots, A_{2n}$. Точку A_1 можно соединить лишь с одной из точек $A_2, A_4, \ldots, A_{2n}$. в противном случае по каждую сторону от хорды, выходящей из A_1, будет размещено нечетное число точек и, значит, при попарном соединении точек между собой по крайней мере одна хорда пересечет хорду, которая выходит из A_1.

Определим, сколько существует различных способов соединения точек, при которых A_1 соединенная с A_{2k}. По одну сторону от A_1A_{2k} размещено $2k$ — 2 точек; их можно соединить попарно a_{k-1} способами. По другую сторону от A_1A_{2k} размещено $2(n - k)$ точек; их можно соединить попарно a_{n-k} способами. Комбинируя каждый из a_{k-1} способов соединения точек $A_2, \ldots, A_{2k-1}$ с каждым из a_{n-k} способов соединения точек $A_{2k+1}, \ldots, A_{2n}$, получим, что число таких способов попарного соединения, при которых A_1 соединено с A_{2k}, равно $a_{k-1}\,a_{n-k}$. Но k может принимать значения 1, 2, ..., n, и поэтому

$$a_n = a_{n-1} + a_{n-2}a_1 + \ldots + a_{n-k}a_{k-1} + \ldots + a_1a_{n-2} + a_{n-1}.$$

Возьмем $a_0 = 1$, и пусть

$$A(s) = \sum_{n=0}^{\infty} a_n s^n$$

— производящая функция последовательности $\{a_n\}$. Тогда

$$a_n = a_0a_{n-1} + a_1a_{n-2} + \ldots + a_ka_{n-k} + \ldots + a_{n-1}a_0,$$

где $n \leq 1$. Умножим обе части предыдущего равенства на s^n и просуммируем по n. Тогда

$$\sum_{n=1}^{\infty} a_n s^n = s \sum_{n=1}^{\infty} s^{n-1}\,(a_0a_{n-1} + a_1a_{n-2} + \ldots + a_{n-1}a_0).$$

Поскольку

$$\sum_{n=1}^{\infty} a_n s^n = A(s) - 1,$$

это, применяя теорему о свертке, имеем

$$A(s) - 1 = sA^2(s). \qquad (23)$$

Решая квадратное уравнение относительно $A(s)$, получим

$$A(s) = \frac{1 \pm \sqrt{1-4s}}{2s}. \qquad (24)$$

Поскольку $A(0)= 1$, так как $a_0 = 1$, то в формуле (24) нужно выбрать знак минус. Возьмем во внимание, что

$$\lim_{s\to 0} \frac{1-\sqrt{1-4s}}{2s} = 1 \quad \text{и} \quad \lim_{s\to 0} \frac{1+\sqrt{1-4s}}{2s} = \infty.$$

Следовательно, производящая функция последовательности $\{a^n\}$ равна

$$A(s) = \frac{1-\sqrt{1-4s}}{2s}. \qquad (25)$$

Теперь остается найти коэффициент при s^n в разложении функции (25). Для этого воспользуемся формулой (5). Согласно (5) будем иметь

$$(1-4s)^{1/2} = 1 + \sum_{k=1}^{\infty} \frac{\frac{1}{2}\left(\frac{1}{2}-1\right)\dots\left(\frac{1}{2}-k+1\right)}{k!}(-1)^k 4^k s^k = 1 - \sum_{k=1}^{\infty} C_{2k}^k \frac{1}{2k-1} s^k. \qquad (26)$$

Отсюда, согласно (25),

$$A(s) = \frac{1-\sqrt{1-4s}}{2s} = \frac{1}{2}\sum_{k=1}^{\infty} C_{2k}^k \frac{1}{2k-1} s^{k-1} = \frac{1}{2}\sum_{n=0}^{\infty} C_{2n+2}^{n+1} \frac{1}{2n+1} s^n \qquad (\text{где } n = k-1).$$

Следовательно,

$$a_n = \frac{1}{2} C_{2n+2}^{n+1} \frac{1}{2n+1} = \frac{1}{n+1} C_{2n}^n.$$

Пример 18. Найти число всех перестановок чисел 1, 2, ..., n, в которых m чисел стоят на своих местах.

Пусть A_1 — множество тех перестановок, в которых на i-м месте стоит i. Тогда имеем

$$N_{[m]}(A_1 \cup \ldots \cup A_n) = C_m^m C_n^m (n-m)! - C_{m+1}^m C_n^{m+1}(n-m-1)! + \ldots + (-1)^{n-m} C_n^m C_n^n = \frac{n!}{m!}\left[1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \ldots + (-1)^{n-m}\frac{1}{(n-m)!}\right] = \frac{n!}{m!}\left[\frac{1}{2!} - \frac{1}{3!} + \ldots + (-1)^{n-m}\frac{1}{(n-m)!}\right].$$

Микромодуль 17

Индивидуальные тестовые задачи

1. На какое наибольшее число частей могут разделить плоскость n прямых?
2. На какое наибольшее число частей могут разделить пространство n плоскостей?
3. На какое наибольшее число частей могут разделить пространство n сфер?
4. Сколькими способами r различных предметов можно разместить в n ящиках?
5. Сколькими способами r пассажиров могут разместиться в n вагонах поезда?
6. Пусть $A(r, n)$ число таких способов размещения r различных предметов в n ящиках ($A(r, n) = 0$, если $r < n$), при которых нет пустых ящиков. Доказать, что

$$A(r, n+1) = \sum_{k=1}^{r} C_r^k A(r-k, n).$$

Как следствие установить, что

$$A(r, n) = \sum_{i=0}^{n} (-1)^i C_n^i (n-i)^r.$$

7. Доказать, что число способов размещения r различных предметов в n ящиках, при которых ровно m ящиков пустые, равно

$$C_n^m A(r, n-m) = C_n^m \sum_{i=0}^{n-m} (-1)^i C_{n-m}^i (n-m-i)^r.$$

8. Сколько существует способов размещения r пассажиров в n вагонах поезда, при которых ровно m вагонов будут свободны?
9. Используя результат задачи 6, вычислить сумму

$$1^k C_n^1 - 2^k C_n^2 + 3^k C_n^3 - \ldots + (-1)^{n-1} n^k C_n^n.$$

Вычислить производящие функции последовательностей:

10. $$a_n = \begin{cases} 1 & \text{при} \quad 0 \leqslant n \leqslant N, \\ 0 & \text{при} \quad n > N. \end{cases}$$

11.

$$a_n = \begin{cases} q^n & \text{при} \quad 0 \leqslant n \leqslant N, \\ 0 & \text{при} \quad n > N. \end{cases}$$

Производящая функция последовательности $\{a_n\}$ равна $A(s)$. Вычислить производящие функции последовательностей:

12.

$$b_n = \begin{cases} 0 & \text{при} \quad n < r, \\ a_{n-r} & \text{при} \quad n \geqslant r. \end{cases}$$

13. $b_n = ca_n$.

14. $b_n = a_n + b$.

15. $b_n = a_n + a_{n-1} + \ldots + a_0 \ldots$

16. $b_n = a_n + a_{n-1}a + a_{n-2}a^2 + \ldots + a_0 a^n \ldots$

17. $b_n = a_{2n}$.

18. Сколькими способами выпуклый n-угольник можно разбить на треугольники диагоналями, которые не пересекаются внутри n-угольника?

19. Вычислить а) $\varphi(100)$; б) $\varphi(1000)$; б) $\varphi(p)$, где p — простое число.

20. Каждый читатель библиотеки прочитал по крайней мере одну книгу из этой библиотеки. О любых k книг из библиотеки ($1 \leq k \leq n$, n — число книг в библиотеке) можно сказать, сколько читателей прочитали все эти книги. Как по этим данным установить, сколько читателей в библиотеке?

21. Используя метод включения и исключение, найти число способов размещения r различных предметов в n ящиках, при которых нет пустых ящиков.

22. Используя метод включения и исключение, найти число способов размещения r различных предметов в n ящиках, при которых ровно m ящиков являются пустыми.

23. Пусть $N_m(A_1, \ldots, A_n)$ — число тех элементов, которые входят по крайней мере в m из множеств $A_1, \ldots, An$. Доеазать, что

$$N_m(A_1, \ldots, A_n) = \sum_{1 \leqslant i_1 < i_2 < \ldots < i_m \leqslant n} N(A_{i_1} \cap \ldots \cap A_{i_m}) -$$
$$- C_m^{m-1} \sum_{1 \leqslant i_1 < i_2 < \ldots < i_{m+1} \leqslant n} N(A_{i_1} \cap \ldots \cap A_{i_{m+1}}) +$$
$$+ C_{m+1}^{m-1} \sum_{1 \leqslant i_1 < i_2 < \ldots < i_{m+2} \leqslant n} N(A_{i_1} \cap \ldots \cap A_{i_{m+2}}) - \ldots$$
$$\ldots (-1)^{n-m} C_{n-1}^{m-1} \sum_{1 \leqslant i_1 < i_2 < \ldots < i_n \leqslant n} N(A_{i_1} \cap \ldots \cap A_{i_n}).$$

24. Найти число способов размещения 8 ладей на шахматной доске так, чтобы они не могли бить друг друга и чтобы ни одна из них не стояла на белой главной диагонали.

25. Будем называть траекторией из точки (0; 0) в точку *(x; y)* ломаную, соединяющую точки (0;0), (1, s_1), .., *(k,*s_k*)*, где

$$s_i - s_{i-1} = \begin{cases} 0, \\ 1, \\ -1. \end{cases} \qquad s_x = y.$$

Пусть $T_{x,\,y}$ — число траекторий из (0; 0) в *(x; y)*. Доказать, что

а)

$$T_{x,\,y} = \sum_{k=0}^{\left[\frac{x-|y|}{2}\right]} \frac{x!}{\left(\frac{y+|y|}{2}+k\right)!\left(\frac{|y|-y}{2}+k\right)!\,(x-|y|-2k)!}.$$

б) Пусть $A=(a;\alpha)$, $B = (b; (\beta)$— точки с целочисленными координатами, причем $b>a\geq 0$, $\alpha>0$, $\beta>0$, а $A'=(a; -\alpha)$ — точка, симметричная A относительно оси OX. Доказать, что число траекторий из A в B равно числу траекторий из A' в B.

в) Пусть $x > 0$, $y > 0$. Доказать, что число всех траекторий из (0; 0) в *(x; y),* не имеющих вершин на оси OX, равно

$$T_{x-1,\,y-1} - T_{x-1,\,y+1}.$$

г) Доказать, что среди $T_{n,0}$ траекторий, которые соединяют (0; 0) с (n; 0), существует:

1) ровно $L_{n-1} = T_{n-2,\,0} - T_{n-2,\,2}$ траекторий, не имеющих вершин на оси OX, кроме крайних точек (0; 0) и *(n;* 0);

2) ровно $L_{n+1} = T_{n,\,0} - T_{n,\,2}$ траекторий, которые не пересекают ось OX.

26. а) Доказать, что среди любых шести человек обнаружится или трое знакомых или трое незнакомых.

б) Если 17 ученых состоят в переписке друг с другом по трем различных научным темам, причем каждая пары ученых ведет

переписку лишь по одной теме, то всегда найдутся трое ученых, переписывающиеся по одной и той же теме. Доказать это.

в) Пусть a_n - последовательность натуральных чисел, которая образована по следующему закону:

$$a_1 = 2 \quad a_{n+1} = (n+1)\,a_n + 1.$$

Предположим, что есть a_n+1 точек, никакие три из которых не лежат на одной прямой. Каждые две точки соединены отрезком одного из n цветов. Доказать, что существует треугольник с вершинами в данных точках, стороны которого имеют один и тот же цвет.

г) Доказать, что

$$a_n \leqslant [e \cdot n!].$$

где e — основа натуральных логарифмов.

27. Рассмотрим на плоскости множество точек, никакие три из которых не лежат на одной прямой; соединим каждые две точки отрезком одного из цветов — красного или черного. Пусть $l(k,m)$ - наименьшее натуральное число такое, что среди любых $l(k,m)$ точек найдется или k точек, соединенных красными отрезками, или m точек, соединенных черными отрезками. Доказать,

а) $l(k, m) \leqslant l(k-1, m) + l(k, m-1)$;

б) $l(k, m) \leqslant C_{k+m-2}^{k-1}$.

Микромодуль 18
Алгоритмы комбинаторики

5.13. Расписания - как примеры комбинаторных задач

1. Календарное планирование. Вся человеческая деятельность планируется во времени: без этого невозможна координированная работа предприятий и производственных участков, по графику идет строительство большой электростанции и детской спортивной площадки, строго во времени расписанны исследования, проведимые космонавтами на борту орбитальной станции, даже время выхода из дома на работу и рабочие, и служащие, и ученые, и инженеры определяют, руководствуясь графиком работы городского транспорта или движения пригородных поездов.

Расписание - это синоним организованности, одно из важнейших средств эффективного выполнения любого рода деятельности, любого рода работ. Чем лучше составлено расписание, тем выше

производительность работы, тем меньше затраты (особенно «нервной энергии»), которые связаны с той или иной деятельностью, тем лучше и сами достигаемые результаты, и условия их достижения.

Так что расписания необходимо уметь составлять, да и притом возможно лучше — *оптимально,* как говорят математики.

Расписания — и часто неплохие — люди научились составлять давно, но составление оптимального расписания стало возможным с тех пор, как к этой работе были привлечении электронные вычислительные машины. Правда, помощь ЭВМ требуется в составлении расписаний для сравнительно сложных комплексов работ, оптимальные расписания для многих простых случаев можно составить и, как говорят, вручную. Однако для этого прежде всего необходимо поставить задачу составления оптимального расписания как математическую задачу, и хотя это сделать довольно просто, но на первый взгляд все же неочевидно.

2. Формулирование математических задач. Начиная с первого класса, люди привыкают думать о *задаче* как о требовании найти некоторое неизвестное поначалу число, которое удовлетворяет *условиям задачи.* Выраженные словами условия задачи необходимо перевести на язык алгебраических формул, которые связывают обозначаемое через *x неизвестное с исходными данными* некоторой «цепочкой отношений» (уравнений, преобразований). Проложив такую «цепочку отношений» от исходных данных к неизвестному, не так уже вычислить вычислить и само неизвестное.

Пример. Поезд шел со скоростью 36 *км/ч,* затем на перегоне, равном 1,5 *км,* поезд шел равноускоренно с ускорением 0,1 *м/сек*^{2}. Найти время, в течение которого поезд прошел перегон.

Очевидно, что легко связать слово «равноускоренно» с формулой

$$S = v_0 t + \frac{at^2}{2}$$

и обнаружить, что S, v_0, a задыны условиями задачи, составить уравнение

$$\frac{at^2}{2} + v_0 t - S = 0$$

и найти t, правильно применяя формулу решения квадратного уравнения к нашему случаю

$$t = \frac{-v_0 \pm \sqrt{v^2{}_0 + 2aS}}{a} \qquad (5.43)$$

учтя при этом, что $t \geq 0$.

Решения, подобные (5.43), где представлен способ решения задачи (связывающая с неизвестным исходные данные, замененные буквами, «цепочка отношений») без выполнения самых вычислений, называют, *решениями в общем виде, формульными решениями.*

Решение некоторой задачи в общем виде, которое составлено в форме четкой инструкции, точно следуя которой получают ***однозначный правильный ответ*** *в каждом конкретном случае (при подстановке исходных данных), называют* ***<u>алгоритмом</u>.***

В начальной математике, где приходится решать сравнительно простые задачи, не предъявляется сколько-нибудь серьезных требований к строгости, единообразию и обозримости представления способа решения задачи — разговор заходит разве что об аккуратности записи отыскания конечного результата.

Алгоритмы решения задач будем записывать с помощью *блок-схем* (образец блок-схемы представлен на рис. 5.13).

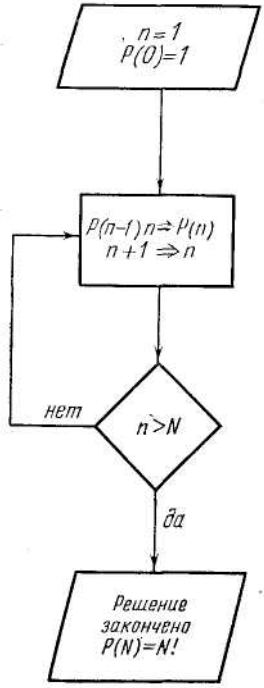


Рис. 5.13. Блок-схема алгоритма вычислении произведения первых N натуральных чисел.

В так называемых *«задачах на доказательство»* — обычно некоторого тождества – требовалось проложить «цепочки отношений», связывающие элементарными преобразованиями левые и правые части равенств. Это те же задачи, но как бы с известными ответами. В *задачах на построение* в геометрии нужно было найти некоторый неизвестный графический объект, определенной условиями задачи конфигурации (окружность, треугольник, угол и т.п.). Здесь объект поиска - не число, более того, здесь обычно еще требуется выполнить построение с помощью некоторых инструментов - например, одного циркуля или одной линейки.

Общим для этих рассмотренных типов задач является то, что с самого начала известно, к какому классу принадлежат искомые объекты (первый - к числам, второй - геометрическим фигурам), с помощью каких операций (арифметических вычислений или вычерчиваний) можно найти эти объекты. Расписания мы тоже поначалу будем формально представлять, а потом уже - строить.

3. Представление расписаний. Существуют различные способы представления расписаний - достаточно вспомнить университетские расписания, футбольный календарь или программу телевидения.

На рис. 5.14 и рис. 5.15 изображенны два способа графического представления расписаний: в первом случае - в виде временной диаграммы, во второму - в виде сетевого графика.

На *временных диаграммах* наглядно отображается выполнение некоторой задачи во времени, при этом все выполнение задачи разбивается на отдельные этапы — *работы,* каждая отдельная работа представляется отрезком, по длине равным продолжительности выполнения работы в выбранном масштабе времени. На диаграмме начало отрезка соответствует моменту времени начала работы, конец отрезка - моменту окончания работы.

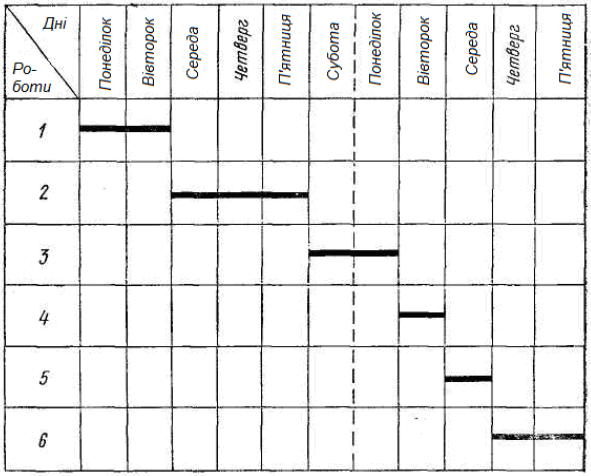


Рис. 5.14. Примерная временная диаграмма подготовки произведения по украинской литературе.

1. Продумывание темы, основных высказываемых мыслей, наметка плана произведения. 2. Просмотр литературы, подбор материалов, эпиграфа, цитируемых высказываний. 3. Составление подробного огдавления, плана - по абзацам. 4. Окончательный подбор материалов, высказываний, цитат согласно плану, шлифовка плана произведения по абзацам. 5. Печатание черновика. 6. Перепечатка набело, проверка произведения.

Сетевым графиком (или стрелочной диаграммой) обычно представляют логическую взаимосвязь во времени отдельных работ, на которые разбивается выполнение некоторой задачи. На рис. 5.15 сетевой график представлен в так называемой *форме «вершины-работы».* На рисунке кружочками, называемыми *вершинами* сетевого графика, обозначены отдельные работы, на которые расчленяется общий комплекс работ, *стрелочки* информируют о взаимозависимости очередности выполнения работ. Так из рис. 5.15 видно, что приступить к выполнению работы 5 следует (разумно, до́лжно) только после того, как выполнены работы 2 и 3.

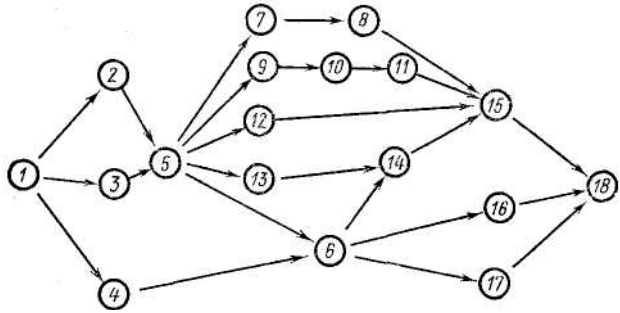


Рис. 4.15. Сетевой график подготовки вечера, посвященного творчеству драматурга.

1. Предварительное обсуждение программы вечера, выделение ответственных по разделам. 2. Определение индивидуальных исполнителей, продумывание их выступлений. 3. Обсуждение плана подготовки сцен из пьес драматурга, подбор исполнителей. 4. Выяснение вопроса о возможности приглашения на вечер известных поэтов, артистов, демонстрации документального кинофильма. 5. Составление плана обеспечения вечера (оформление зала, реквизит для спектакля). 6. Окончательное обсуждение и согласования программы и сроков проведения вечера, списка приглашений. 7. Подготовка индивидуальных выступлений. 8. Репетиции индивидуальных выступлений. 9. Подготовка ролей в спектаклях. 10. Первая репетиция сцен. 11. Вторая репетиция сцен. 12. Подготовка костюмов. 13. Подготовка декораций и стендов для оформления зала. 14. Оформление зала. 15. Генеральная репетиция. 16. Доставка кинофильма. 17. Приглашение на вечер. 18. Проведение вечера.

Логическая взаимосвязь очередности работ, представленная сетевым графиком, облегчает прежде всего распараллеливание комплекса робот, распределение их между исполнителями, позволяет объективно определить общую продолжительность выполнения задачи, оценив предварительно время выполнения отдельных этапов.

На рис. 5.16 представлено тот же сетевой график, что и на рис. 5.15, только каждой вершине поставлено в соответствие число — оценка времени выполнения отдельной работы в днях.

Жирными стрелками обозначен так называемый *критический путь:* если сложить продолжительности работ, составляющих этот путь в нашем примере, то станет ясно, что от принятия решения о проведении вечера до дня проведения вечера должно пройти не менее 16 дней (при данных оценках времени выполнения отдельных работ).

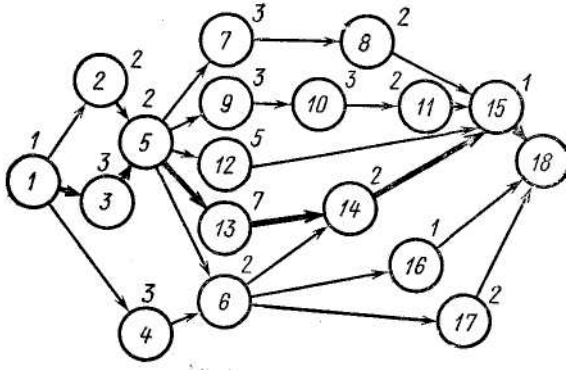


Рис. 5.16. Сетевой график с оценками продолжительностей отдельных работ и указанием критического пути.

Любая другая последовательность взаимосвязанных во времени работ - другой путь в сетевом графике – не оказывает на срок проведения вечера столь решающего влияния, как критический путь: в отличие от работ, лежащих на критическом пути, можно даже несколько (в допустимых пределах) повременить с выполнением таких работ - окончательный срок проведения вечера при этом не изменится.

Составив предварительно сетевой график и отыскав на нем критический путь, организаторы вечера могут уже более обоснованно подойти к установлению срока проведения вечера, уверенно определить сроки выполнения отдельных работ, усилить контроль и подготовку мероприятий, которые лежат на критическом пути, так как именно здесь срыв срока выполнения отдельной работы наиболее угрожает срокам проведения вечера.

4. Оптимальные расписания. Представьте теперь, что наряду с сетевым графиком подготовки вечера, посвященного драматургу, составлено и сетевой график подготовки выставки «История района, в котором я живую», и в том и в другом мероприятии заняты одни и те же «исполнители». Вам надо составить общий график (допустим, в виде временной диаграммы) выполнения намеченных дел, обосновано назначить сроки проведения мероприятий с учетом того, что, возможно, подготовка выставки приурочена к какой-то дате, что в это время должны пройти соревнования по легкой атлетике и что все это, разумеется, не должно мешать занятиям.

Теперь вы, наверное, убедились и в сложности задачи составления расписаний, и в необходимости составления расписаний: очень часто,

когда кто-то «разрывается» между многими делами и запустил обучение, причина - в неумении «раскроить» время, упорядочить дела и поручения во времени.

Значительно более сложно, разумеется, спланировать во времени разработку и производство новейшего космического корабля или создание современного промышленного комбината так, чтобы выполнить все работы в кратчайшие сроки или в заданные сроки, но с минимальными затратами. Непросто составить график движения железнодорожных составов, максимально загружающих ветвь данной пропускной способностью.

Трудности в решении задач теории расписаний (календарного планирования) происходят из-за их *вариантности*: иногда можно менять очередность выполнения работ, иногда сроки, иногда - подбирать других исполнителей.

Ограничение на возможности изменения очередности, выбор исполнителей обусловливаются различными (в том числе и экономическими) требованиями.

Взаимозаменяемость и ограниченность увеличивают эту вариантность, делают труднее задачу составления наилучшего — *оптимального* расписания.

И в «жизненных» ситуациях тоже возникают задачи составления оптимальных расписаний (хотя мы часто и не осознаем этого). С таких простых примеров мы и начнем рассмотрение раздела математики, изучающего вопросы *упорядочения* (во времени) объектов различной природы.

5.14. Экстремальные перестановки

1. Перестановки. Последовательное расположение в любом порядке чисел от 1 до n образует, как мы знаем, их некоторую *перестановку* или *n-перестановку.* Можно говорить о перестановках любых предметов (объектов), а так как эти предметы можно перенумеровать, то с формальной точки зрения изучение таких перестановок можно свести к изучению n-перестановок.

Произвольную n-перестановоку будем обозначать символом

$$\sigma_n = \langle\, i_1, i_2, \ldots, i_k, \ldots, i_n \,\rangle\,; \qquad (5.44)$$

таким образом, i_k (или $i_k(\sigma_n)$) - число, стоящее в перестановке σ_n на k-м (от начала) месте; $k(i)$ (или $k(i/\sigma)$) - номер места, которое занимает в перестановке σ число i. В перестановке $\langle 2, 4, 1, 3 \rangle$, например, $i_3 = 1$, $k(3)=4$.

Число P(n) всех возможных различных n-перестановок равно произведению чисел от 1 *до n, называемому n-факториалом и обозначаемому* $n! = 1 \cdot 2 \cdot 3 \ldots n$.

Действительно, легко доказать справедливость следующего рекурентного соотношения:

$$P(n) = P(n-1) \cdot n, \quad P(1) = 1, \tag{5.45}$$

(*Рекурентними соотношениями* называют формулы для общего члена u_n последовательности, если u_n можно выразить как функцию некоторого числа l предыдущих членов последовательности $u_n = f(u_{n-1}, u_{n-2}, \ldots, u_{n-l})$. Такие последовательности называют также *обратными*) откуда и следует, что

$$P(n) = n! \tag{5.46}$$

Доказательство строится по индукции. Справедливость $P(1)= 1$ очевидна. Различных n-перестановок с зафиксированным i_1, очевидно, столько, сколько можно построить различных перестановок остальных n— 1 элементов в σ_n, т.е. $P(n$— 1). Кроме того, очевидно, что i_1 может принимать n значений. Отсюда и следует (5.45).

Число $P(n)$ быстро растет с возрастанием n (табл. 5.1) - быстрее, чем даже показательная функция a^n при любом наперед заданном a (так что $n! > a^n$ начиная с какого-то n).

Порядок роста функции $P(n)$ устанавливает следующая формула Стирлинга

$$n!\ ;\ \sqrt{2\pi n} \cdot n^n e^{-n}. \tag{5.47}$$

Таблица 5.1

Значения $n!$, вычисленые непосредственно (т.е. по формуле (5.46) с последующим округлением) и по формуле Стирлинга ($n! = S(n)$)

n	$n!$	$S(n)$	n	$n!$	$S(n)$
1	1	0,9221	8	40 320	39 901
2	2	1,9189	9	362 880	359 536
3	6	5,8362	10	3 628 800	3 598 690
4	24	23,5052	15	$1{,}308 \cdot 10^{12}$	$1{,}3005 \cdot 10^{12}$
5	120	118,017	20	$2{,}433 \cdot 10^{18}$	$2{,}423 \cdot 10^{18}$
6	720	710,06	25	$1{,}511 \cdot 10^{25}$	$1{,}546 \cdot 10^{25}$
7	5 040	4980,4	50	$3{,}041 \cdot 10^{64}$	$3{,}037 \cdot 10^{64}$
			100	$9{,}333 \cdot 10^{157}$	$9{,}327 \cdot 10^{157}$

Многие практические задачи сводятся к определению некоторой перестановки. Например, как n-перестановка (5.44) может быть представлено распределение n претендентов по n местам: i_k здесь — номер претендента, распределенного на k-е место. Как n-перестановка может рассматриваться также решение задачи об очередности последовательного выполнения n (перенумерованных предварительно) работ; i_k здесь— номер той работы, которая, согласно решению, будет выполнена k-й по порядку.

2. Задача директора. В приемной в ожидании личной встречи с директором собралось n посетителей. Предварительный опрос позволил выяснить, сколько времени должен уделить директор рассмотрению вопроса каждого посетителя: для i-го посетителя (согласно тому, как мы их перенумеровали) это время обозначим через T_i. Директор, зная, что хотя общее (суммарное) время, которе он уделит всем посетителям, одно и тот же, $T = \sum_i T_i$ (независимо от очередности их приема), хотел бы так организовать прием, чтобы посетители находились в приемной в целом как можно меньше. Иными слонами, он хотел бы *минимизировать время ожидания* посетителей в приемной.

Решением задачи директора, очевидно, будет некоторая n-перестановка чисел

$$\sigma_n = \langle\, i_1, i_2, \ldots, i_k, \ldots, i_n \,\rangle, \tag{5.48}$$

соответствующая очередности приема посетителей.

Обозначим через $\tau_k(\sigma_n)$ время ожидания в приемной посетителя i_k при очередности приема, который задается перестановкой σ_n.

Очевидно, что

$$\tau_k(\sigma_n) = \tau_{k-1}(\sigma_n) + T_{i_{k-1}}, \quad \tau_1(\sigma_n) = 0, \tag{5.49}$$

причем, естественно, предполагается, что директор не делает заметных (не равных 0) перерывов между приемами двух посетителей.

Директор ставит задачу: *найти такую перестановку* σ_n, *на которой величина*

$$F(\sigma_n) = \sum_k \tau_k(\sigma_n) \tag{5.50}$$

принимает наименьшее значение.

Другими словами, требуется среди всех возможных перестановок σ_n найти такую, которая минимизирует значение функции $F(\sigma_n)$. Последняя фраза — уже почти «математическая» постановка задачи

директора: здесь определено множество объектов, среди которых надо искать решение задачи — перестановки (о посетителях и директоре мы уже на время — до решения задачи — забыли). Здесь же указано свойство искомого решения (для него значение $F(\sigma_n)$ является минимальным); ранее было определено, как вычислить $F(\sigma_n)$ для каждой перестановки.

3. Экстремальные перестановки и задачи очередности. Задача директора - простейшая задача очередности. *Задачей очередности* называют такую задачу составления расписаний, которая сводится к строгому упорядочению выполнения некоторых *работ* — в нашем случае упорядочению приема директором посетителей. После соответствующей формализации такая задача сводится к поиску экстремальной перестановки.

В *задачах поиска экстремальной перестановки* требуется найти n-перестановку σ_n, на которой некоторая функция $F(\sigma_n)$ достигает *экстремума* (минимума или максимума в зависимости от постановки задачи, в нашем примере — минимума) и задан способ вычисления $F(\sigma_n)$ на каждой перестановке σ_n.

Функцию, екстремум которой находится в некоторой задаче, называют *функцией оптимизации* или *функцией-критерием.*

Выбор функции-критерия является важным звеном в формализации задачи. Вид (способ вычисления значений) функции-критерия должен соответствовать требованиям, предъявляемым к варианту, который может быть признанный «наилучшим». Различные функции-критерии считаются *эквивалентными,* если приводят к одному и тому же решение (в этом случае говорят также, что *задачи* — постановки — *эквивалентны)* .

То, что выбор функции-критерия - далеко не очевидная вещь - показывает следующая задача.

4. Задача о назначениях, или как рассадить группу за столами. Эта задача больше известная как *задача о женихах и невестах.* На некотором одиноком острове живут n женихов и n невест. Вопрос заключается в том, как лучшим образом всех их переженить. Решение становится невозможным, если принять во внимание, которое

«Оля любит Сашу,
И Зоя любит Сашу,
А Саша любит Лилю,
Лиля же любит Андрея,
Но Андрей никого не любит,
Но зато вот Сергей
По-прежнему любит Олю...».

Отношения между «женихами и невестами», соответствующие приведенному утверждению, схематично отображены на графе рис. 5.17.

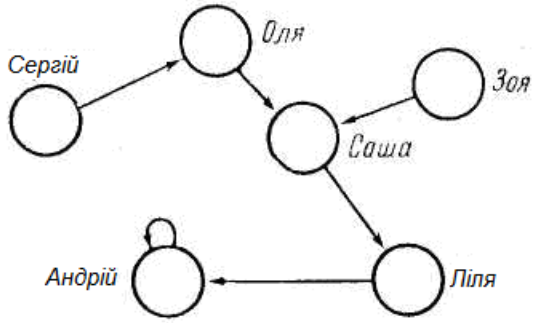


Рис. 5.17. Граф взаимоотношений в одном коллективе, стрелка соответствует отношению «любит» (тот, от кого стрелка направлена, того, на кого направлена).

Жители острова уважают чувство молодых людей, но и проявляют заботу об интересах острова в целом. Противоречивый узел отношений молодых людей ими разрубается следующим образом. Считается, что не исключена возможность заключення любого брака, но одни браки предпочтительнее других. Эта предпочтительность выражена численно: для брака между i-м женихом и k-й невестой она получает значение a_{ik}. Это означает, что в общую суммарную «пользу» (для островного государства) каждая свадьба вносит свое слагаемое, равное a_{ik}. В общей постановке задачи о назначении требуется n претендентов назначить на n должностей — на каждую должность по одному. Общая польза рассматривается как выгода, суммируемая по всем назначениям, а выгода от назначения i-го претендента на k-ю должность оценивается в a_{ik}, независимо от других назначений. Как уже говорилось, всякая перестановка σ_n представляет возможное решение задачи о назначениях. Требуется найти такую перестановку σ_n, которая максимизирует $F(\sigma_n)$:

$$F(\sigma_n) = \sum_k a_{i_k k}. \tag{5.51}$$

Определять выгоду a_{ik} от назначения i-го претендента на k-ю должность принято с помощью так называемых *экспертных оценок,* для чего собирают некоторый представительный консилиум специалистов-экспертов, и они выставляют свои оценки — например,

в шестибальной системе — выгодности каждого возможного назначения (аналогично тому, как выставляют оценки за выполнение программы в фигурном катании на коньках). Крайние оценки (наименьшая и наибольшая), как правило, отбрасываются, оставшиеся усередняются (суммируются и делятся на их число), эта усредненная оценка и принимается за значение a_{ik}. Повторяем, такая оценка должна быть заранее поставлена любому возможному назначению - любой паре *(i, k).* Явная нежелательность какого-то назначения фиксируется в наиболее низкой оценке для этого назначения.

5.15. Метод перебора и схема конструирования вариантов

1. Метод перебора. Объект, на котором функция-критерий принимает экстремальное — минимальное или максимальное значения, — называют *экстремальным элементом* или *оптимальным решением.* Множество элементов (объектов), среди которых ищется экстремальный, называют *множеством возможных решений, допустимых вариантов.* Если это множество *конечно,* т.е. состоит из конечного числа вариантов и этих вариантов относительно немного, то можно последовательно рассмотреть все возможные решения, на каждом допустимом варианте вычислить значение функции-критерия и, сравнивая эти значения, выбрать оптимальное решение. Собственно, в этом состоит *метод перебора* (в современной математической терминологии - «метод проб и ошибок»), *применимый, повторяем, в случае конечного множества допустимых решений* и притом сравнительно немногочисленного, чтобы можно было провести все вычисления и сравнение в приемлемое время (или, как еще говорят, в реальном масштабе времени). Табл. 5.3 демонстрирует применение метода перебора в решении задачи, которая задана упражнением 5 (см. «Типовые тестовые задачи» данного микромодуля).

Таблица 5.3

Возможные решения	Значения критерия	Наилучшее?	Возможные решения	Значения критерия	Наилучшее?
⟨1, 2, 3⟩	10	—	⟨2, 3, 1⟩	12	Да
⟨1, 3, 2⟩	10	—	⟨3, 1, 2⟩	6	—
⟨2, 1, 3⟩	9	—	⟨3, 2, 1⟩	9	—

Здесь нетрудно выписать все возможные решения (первый столбец табл. 5.3), потом вычислить значения функции-критерия для каждого варианта по формуле 5.51 (второй столбец табл. 5.3), указать (после просмотра этого столбца) наилучшее значение (2, 3, 1).

В некотором смысле метод перебора - самый примитивный метод решения задачи, который «чистые» математики вряд ли даже назовут математическим. Однако математики-прикладники не так уже редко используют этот метод для решения практических задач, особенно с появлением ЭВМ.

Часто метод перебора применяют на заключительных стадиях решения задач, когда другими методами отобрано сравнительно мало вариантов, среди которых лежат оптимальные.

Пример. Рассмотрим задачу определения экстремума функции $F(x) = 3x^4 + 4x^3 — 12x^2$, заданной на отрезке [—2, 2]. Экстремум такой функции достигается или в крайних точках отрезка, или внутри отрезка **в точках, где производная функции обращается в нуль**. Таким образом, вместо бесконечного множества точек отрезка [-2, 2] экстремум функции следует искать среди значений $f(x)$ для точек —2, 0, 1, 2, (—2, 0, 1 — корни уравнения $f'(x) = 0$, производная же нашей функции есть $f'(x) = 12x^3 + 12x^2 — 24x = 12x(x^2 + x — 2)$). Максимум $f(x)$ достигается, как это нетрудно проверить теперь методом перебора, в точке $x = 2$, $f(x) = 32$, минимум достигается в точке $x = —2$, $f(x) = —32$.

Вообще говоря, установив некоторые свойства оптимального варианта, можно значительно сузить множество допустимых решений, вплоть до получения конечного множества, где часто оказывается возможным применить метод перебора. Неоценимое значение метода перебора заключается в том, что он принципиально всегда «под рукой». Для конечных множеств допустимых решений это означает, следовательно, что существует конечный алгоритм решения задачи, т.е. задача будет разрешима за конечное время. Плохо то, что для метода перебора это «конечное» время оказывается неприемлемо большим уже даже в простых случаях. Так представим себе, что в задаче поиску экстремальной перестановки, в случае всего 10 элементов, на построение одной перестановки и вычисление значения функции-критерия мы затрачиваем один минуту. Тогда нетрудно подсчитать: при восьмичасовом рабочему дне методом перебора такую задачу пришлось бы решать ... больше двух лет. В случае же 20 элементов даже с помощью ЭВМ такая задача методом перебора решалась бы дестилетиями! Значит, чтобы построить метод точного

решения такого рода задач, надо изобрести что-то лучшее, чем примитивный перебор всех возможных вариантов.

И все же отдадим должное методу перебора: мы еще не раз возвратимся к нему, во-первых, для решения сравнительно простых задач, во-вторых, хотя бы для оценки того, насколько другой предложенный нами метод решения задачи лучше (эффективнее) метода перебора - такое сравнение делается довольно часто. В-третьих, многие эффективные методы решения *дискретных задач оптимизации* (т.е. задач с конечным множеством вариантов) «изобретаются» вроде бы как некоторое «улучшение» метода перебора — это мы проиллюстрируем в следующих разделах модуля.

2. Алгоритмы построения *n*-перестановок. Однако, чтобы «улучшать» метод перебора, нужно, прежде всего, уметь им пользоваться - для задач поиску экстремальных перестановок это означает уметь строить все возможные *n*-перестановки, иначе говоря, надо знать алгоритм построения всех *n*-перестановок.

Нетрудно после некоторых попыток «нащупать» элементарный регулярный прием получения последовательности всех *n*-перестановок (чем мы уже неявно воспользовались при формировании табл. 5.3 из предыдущего пункта), начиная с начального упорядочения чисел 1, 2, .., *n* по возрастанию (пусть $n = 5$):

1, 2, 3, 4, 5
1, 2, 3, 5, 4
1, 2, 4, 3, 5
1, 2, 4, 5, 3
1, 2, 5, 3, 4
1, 2, 5, 4, 3
1, 3, 2, 4, 5

Чтобы попроще описать найденный прием, введем некоторые понятия. Пара соседних чисел (в перестановке) назовем *упорядоченной*, если первое число в паре меньше второго.

Рассмотрим некоторую перестановку σ_n. Найдем первую с конца перестановки упорядоченную пару. Так в перестановке σ_n=(1,3, 5, 4, 2) первая с конца упорядоченная пара есть пара (3, 5). Первое число такой пары назовем *обрывающим*. *Перестановочный хвост* в σ_n образует последовательность чисел, начиная с обрывающего.

Реупорядочить перестановочний хвост означает:

1) заменить обрывающее число на наименьшее из перестановочного хвоста число, которое превосходит обрывающее;

2) все другие числа из перестановочного хвоста (вместе с обрывающим) расположить в порядке возрастания.

Так в нашей перестановке $\sigma_n = (1, 3, 5, 4, 2)$ обрывающее число есть 3, перестановочный хвост есть последовательность (3, 5, 4, 2).

Заметим, обрывающего числа не найдется только в перестановке, в которой все числа расположены в порядке убывания. В нашем алгоритме это сигнал того, что решение закончено.

Введение понятий «обрывающего числа», «перестановочного хвоста», «реупорядочения» позволяет упростить описание алгоритма построения всех *n*-перестановок. Этот алгоритм — назовем его *Алгоритмом*-1 - представлен блок-схемой на рис. 5.18.

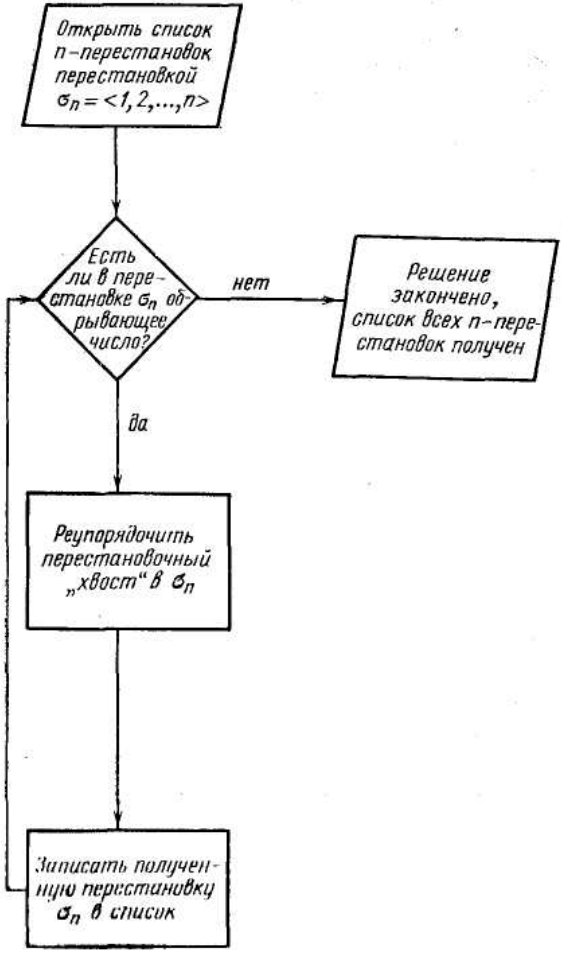


Рис. 5.18. Блок-схема Алгоритма-1 получение всех *n*-перестановок.

Получение первых нескольких перестановок по этому алгоритму отображено в табл. 5.4.

Таблица 5.4

Первые 6 перестановок, которые получены согласно Алгоритму-1

№	Перестановка	Обрывающее число	Перестановочный хвост и его реупорядочение
1	⟨1, 2, 3, 4, 5⟩	4	(4, 5) ⟶ (5, 4)
2	⟨1, 2, 3, 5, 4⟩	3	(3, 5, 4) ⟶ (4, 3, 5)
3	⟨1, 2, 4, 3, 5⟩	3	(3, 5) ⟶ (5, 3)
4	⟨1, 2, 4, 5, 3⟩	4	(4, 5, 3) ⟶ (5, 3, 4)
5	⟨1, 2, 5, 3, 4⟩	3	(3, 4) ⟶ (4, 3)
6	⟨1, 2, 5, 4, 3⟩	2	(2, 5, 4, 3) ⟶ (3, 2, 4, 5)

Нетрудно убедиться в том, что Алгоритм-1 действительно решает поставленную задачу. Этот факт, очевиден для n=1, можно проверить и для n =2. Пусть это верно для $(n-1)$, т.е. алгоритм действительно получает все различные перестановки в случае $n-1$ элементов. Но если применить этот алгоритм для n элементов, то цифра 1, стоящая на первом месте в исходной перестановке, будет заменена на 2, только тогда, когда она станет обрывающим числом, т.е. когда будут получены все $(n-1)!$ различных перестановок других чисел. Точно так же цифра 2 на первом месте в перестановках будет заменена на 3 только после получения всех $(n-1)!$ различных перестановок других элементов и т.д. Это и означает, что алгоритм получает все $n\cdot(n-1)!$ перестановок, при этом среди них не будет совпадающих.

Другой алгоритм — *Алгоритм*-2 — получение всех n-перестановок представлены блок-схемой на рис. 5.19.

Только один термин к блок-схеме рис. 5.19 нуждается в пояснении.

Назовем *«вращением»* некоторой последовательности A чисел замену ее другой последовательностью B, где число, стоящее в A на первом месте, оказывается в B на последнем месте, взаимное расположение других чисел не меняется. Так вращение (1, 2, 3) приводит к (2, 3, 1).

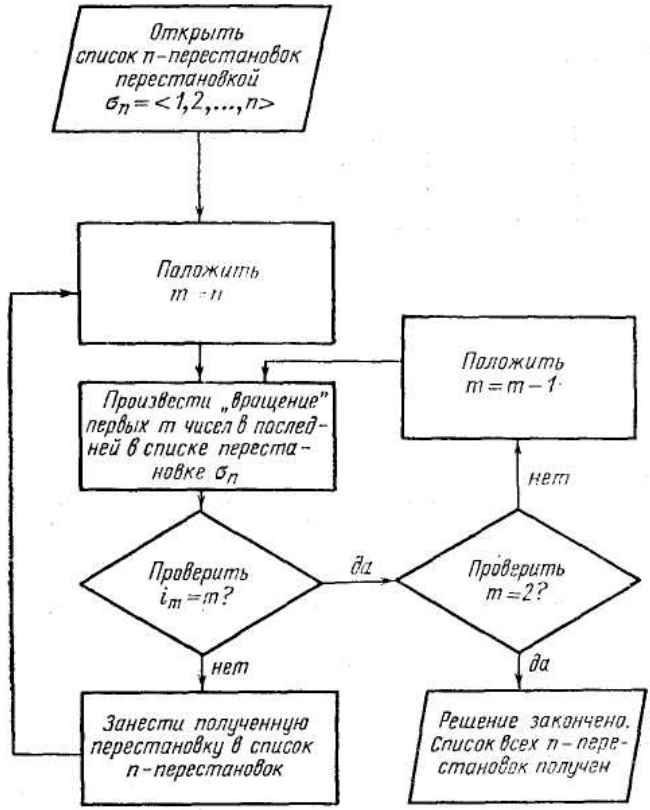


Рис. 5.19. Блок-схема Алгоритма-2 получение всех *n*-перестановок.

Табл. 5.5 поясняет ход решения по этому алгоритму при получении первых нескольких перестановок.

Таблица 5.5

Первые перестановки, полученные6 согласно Алгоритму-2

№	Перестановка	Вращаемая часть	Результат вращения
1	⟨1, 2, 3, 4, 5⟩	$m=5:\langle 1,2,3,4,5\rangle$	⟨2, 3, 4, 5, 1⟩
2	⟨2, 3, 4, 5, 1⟩	$m=5:\langle 2,3,4,5,1\rangle$	⟨3, 4, 5, 1, 2⟩
3	⟨3, 4, 5, 1, 2⟩	$m=5:\langle 3,4,5,1,2\rangle$	⟨4, 5, 1, 2, 3⟩
4	⟨4, 5, 1, 2, 3⟩	$m=5:\langle 4,5,1,2,3\rangle$	⟨5, 1, 2, 3, 4⟩
5	⟨5, 1, 2, 3, 5⟩	$m=5:\langle 5,1,2,3,4\rangle$	⟨1, 2, 3, 4, 5⟩
		$m=4:\langle 1,2,3,4\rangle$	⟨2, 3, 4, 1⟩
6	⟨2, 3, 4, 1, 5⟩	$m=5:\langle 2,3,4,1,5\rangle$	⟨3, 4, 1, 5, 2⟩

3. Схема конструирования вариантов. Порфириан. Как видно из предыдущего пункта, множество всех возможных вариантов можно получать различными способами даже в одной задаче - поиска экстремальной перестановки.

И все же существует единообразный прием представления последовательного построения всех возможных вариантов в самих разнообразных задачах дискретной оптимизации. Его мы и продемонстрируем сейчас на примере образования всех возможных n-перестановок. Изобразим графически кружочком множество всех возможных n-перестановок. Разобьем это множество на n подмножеств, отнеся к одному подмножеству все те перестановки, у которых на первом месте стоит одно и то же число i_1. В первое подмножество попадут все перестановки, у которых на первом месте расположена 1 ($i_1 = 1$), во второе подмножество попадут σ_n с $i_1 = 2$ и т.д. Изобразим эти подмножества графически также кружочками, внутри кружочков запишем значения i_1, соединим стрелками кружочки со знаком множества всех возможных n-перестановок, как это показано на рис. 5.20.

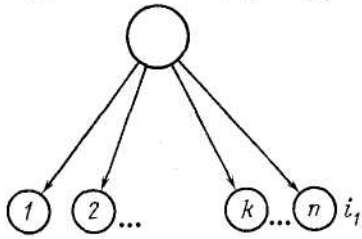


Рис. 5.20

В свою очередь каждую из этих подмножеств можно еще разделить на непересекающиеся подмножества в зависимости от того, какое число размещено в перестановке на втором месте (рис. 5.21), вновь образованные подмножества можно разбить на части в зависимости от того, какое число расположено на третьем месте и т.д.

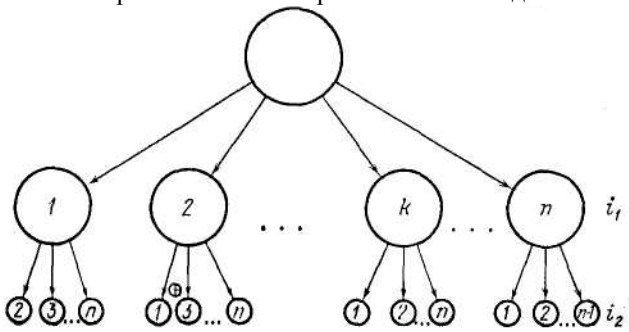


Рис. 5.21

Процесс такой последовательной разбиения множеств оборвется тогда, когда мы дойдем до отдельных, единичных перестановок. На рис. 5.22 представлен результат такой последовательной разбиения множества всех перестановок чисел 1, 2, 3.

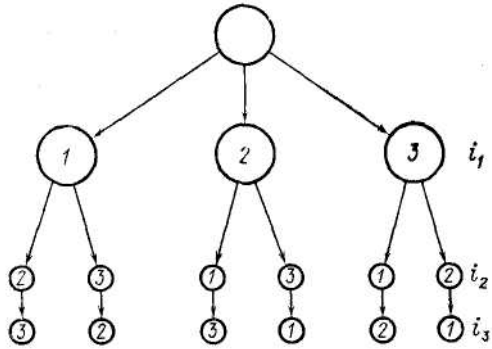


Рис. 5.22

Полученный граф (такие графы в математике за их внешний вид называют *деревьями),* в сущности, представляет реализацию алгоритма получения всех перестановок (в нашем случае Алгоритма-1). В дальнейшем такое графическое изображение последовательного

разбиения множества всех вариантов мы будем называть *порфирианом*.

Каждая вершина в порфириане располагается, как мы будем говорить, *на некотором уровне.* В нашем примере вершины, соответствующие фиксации i_1, расположены на первому уровне, фиксации i_2 — на втором уровне. Вершине, которая расположенной на k-м уровне, соответствуют *отрезки длины k n*-перестановок

$$\sigma_k = \langle i_1, i_2, \ldots, i_k \rangle$$

Вершинам, расположенным на n-м уровне, соответствуют «полные» перестановки $\sigma_{г.}$

Построение вершин $(k + 1)$-го уровня порфириана, исходя из некоторой вершины k-го уровня, будем в дальнейшем называть *операцией разветвления,* или *ветвлением.*

Умение правильно определить операцию ветвления равносильно знанию способа конструирования всего порфириана, т.е. знанию алгоритма построения всех возможных вариантов решения задачи (возможными вариантами считаются те перестановки, которые удовлетворяют всем условиям задачи, за исключением условия приводить к экстремуму функции-критерии).

На порфириан можно также смотреть как на представление последовательной классификации множества всех возможных вариантов, т.е. процесса разбиения его на все более мелкие подмножества, вплоть до отдельных вариантов.

Для нас важнее, что порфириан задает способ построения (точнее, им определяется этот способ) - конструирования - всех возможных вариантов, которые удовлетворяют условиям задачи (за исключением, еще раз подчеркнем, условия приводить к экстремуму функции-критерия).

В этом смысле каждая вершина порфириана (кроме начальной) замыкает собой как бы некоторый фиксированный фрагмент в представлении некоторого возможного решения. Так отмеченная на рис. 5.21 знаком + вершина соответствует, если проследить путь от исходной вершины до отмеченной, последовательному закреплению сначала $i_1 = 2$, потом $i_2 = 1$. Таким образом, будем считать, что отмеченная вершина представляет собой фрагмент $\sigma_2 = \langle 2, 1 \rangle$ (В аспекте классификации это означает, что отмеченная вершина представляет то подмножество вариантов, которое начинается с чисел 2 и 1 ($i_1 = 2$, $i_2 = 1$), и $\langle 2, 1 \rangle$ есть классификационный код, введенное классификацией имя этого множества).

В данном пункте мы сознательно говорили о некоторых фрагментах решения, чтобы подчеркнуть, что понятие порфириана не связано с какой-нибудь конкретной формой представления решения, а только с возможностью последовательно записать это решение по частям. В дальнейшем мы уточним и углубим содержание этого высказывания.

5.16. Общая схема построения порфириана

1. Путешествие бродячего торговца по плоскости и на графе. В самой общей постановке в задаче бродячего торговца не предполагается, что расстояния между городами измерены циркулем на плоской карте. Задача значительно упрощается, если считать, что город задан точками на плоскости, расстояние между двумя городами задается длиной отрезка, их соединяющего, и требуется соединить все точки замкнутым контуром минимальной длины — это и есть *задача бродячего торговца на плоскости.*

Для такой постановки оказываются справедливыми следующие утверждения.

Утверждение 1. *Путь бродячего торговца* (*в задаче на плоскости) не имеет самопересечений.*

Назовем отрезок пути $\sigma_k = \langle i_1, i_2, \ldots, i_k \rangle$ *разделяющим,* если

1) он разбивает множество всех городов — точек на плоскости, отличных от $i_1, i_2, \ldots, i_k$, на две части такие, что

2) не существует отрезка, соединяющего точку из одной части множества и точку из другой части множества и не пересекает при этом σ_k.

Утверждение 2. *Оптимальный маршрут не содержит разделяющих отрезков* σ_k $(k < n)$.

Утверждение 3. *Кратчайший путь бродячего торговца по всем точкам, являющимся вершинами некоторого выпуклого многоугольника, есть граница этого многоугольника.*

Так для совокупности пунктов на плоскости, представленных рис. 5.23, где внутри выпуклого многоугольника расположен всего один пункт, заведомо ясно, что решением будет контур, близко подходящий к границе многоугольника (согласно утверждению 3).

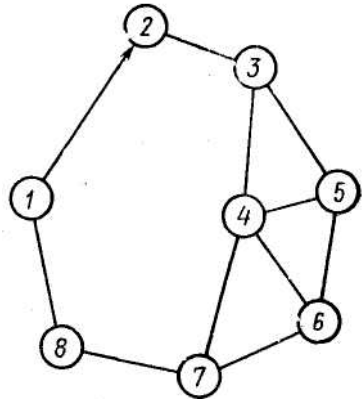


Рис. 5.23

Чисто визуально оценив, что в пункт 4 можно зайти из пунктов 3, 5 или 6, а из других пунктов явно нецелесообразно, легко получить множество возможных решений в таком случае - усього три варианта, представленных порфирианом на рис. 5.24:

$$\langle 1,2,\ 3,4,\ \ 5,6,7,8 \rangle ,$$

$$\langle 1,2,3,5,\ \ 4,6,7,8 \rangle ,$$

$$\langle 1,2,3,5,\ \ 6,4,7,8 \rangle .$$

(Допустимо и прохождение маршрута в обратном порядке, в «плоской» постановке такие маршруты можно считать эквивалентными).

В задаче, подобной приведенной, можно сказать, что мы используем такую схему построения порфириана, как в случае поиска множества всех перестановок, однако в операции разветвления при переходе от σ_k к σ_{k+1} проверяем прежде всего допустимость такого перехода — в нашем примере, сообразуется ли путь бродячего торговца σ_{k+1} с дорогами, нанесенными на карту рис. 5.23.

На рис. 5.24 крестиком отмеченные варианты, которые оказываются допустимыми при существующих дорогах, но неприемлемыми, так как нарушается условие, сформулированное утверждениям 2.

В нашем примере нанесение на карту ограничивающих передвижение «дорог» явилось результатом теоретического исследования свойств оптимального решения. Во многих практических постановках такие ограничивающие передвижения бродячего торговца дороги бывают указаны условиями задачи и

представляются обычно в форме графа вида рис. 5.25. Такие постановки получили название *задач бродячего торговца на графе.*

Фиксация возможных переходов на графе резко суживает множество допустимых решений.

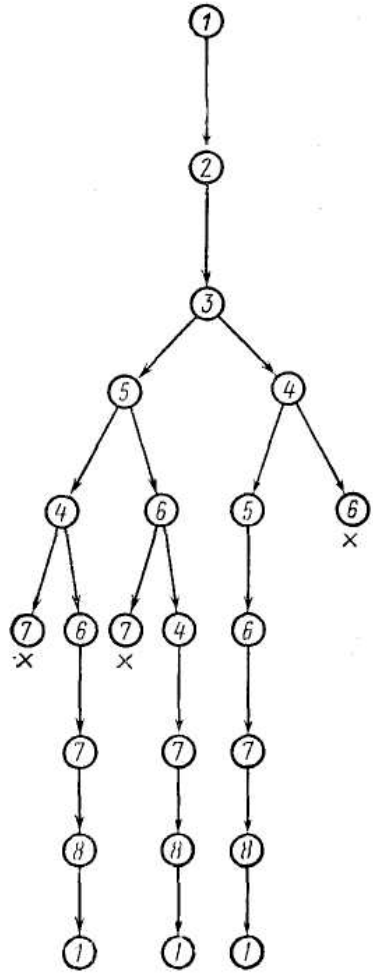


Рис. 5.24

Рассмотрим поиск пути бродячего торговца на графе, приведенного на рис. 5.25.

Операция разветвления при построении порфириана в этом случае при переходе от $\sigma_k = \langle i_1, i_2, \ldots, i_k \rangle$ к $\sigma_{k+1} = (\sigma_k, j)$ требует строить вершины $(k + 1)$-го уровня (обозначая их j), если в графе рис. 5.25 имеется стрелка, которая соединяет i_k с j.

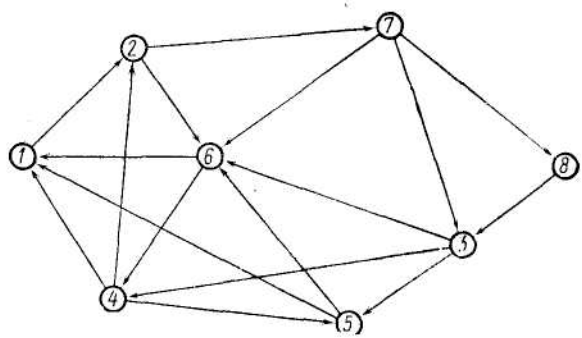


Рис. 5.25.

В нашем примере мы тоже получим всего три возможных решения:

(1, 2, 7, 8, 3, 4, 5, 6),
(1, 2, 7, 8, 3, 5, 6, 4), (5.52)
(1, 2, 7, 8, 3, 6, 4, 5).

2. Расстановка оборудования вдоль кругового конвейера. Круговой конвейер движется в одном направлении. Вдоль конвейера требуется расставить n станков. Обработанные на станке i детали передаются на другой станок j в количестве a_{ij} по конвейеру.

Требуется расставить станки в таком порядке, чтобы минимизировать общее время передачи деталей от станка к станку. Последнее означает, что если t_{ij} — время передачи детали от станка i к станку j, то требуется минимизировать

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} t_{ij}.$$

Будем считать, что время t_{ij} пропорционально расстоянию ρ_{ij} между станками вдоль конвейера, а это расстояние полностью определяется тем, «как далеко» отстоят i и j в перестановке σ_n.

Пусть k (i) и $k(j)$ - соответственно места, которые занимают i и j в перестановке σ_n. Конвейер движется вдоль станков в направлении от начала σ_n к концу. Это означает, что

$$\rho_{ij}(\sigma_n) = \begin{cases} k(j) - k(i), & \text{если} \quad k(i) < k(j), \\ n + k(j) - k(i), & \text{если} \quad k(i) > k(j). \end{cases} \tag{5.53}$$

Таким образом, математически задача расстановки станков вдоль кругового конвейера сведена нами к поиску перестановки σ_n, на которой минимизируется функция

$$F(\sigma_n) = \sum_{i=1}^{n} a_{ij}\rho_{ij}(\sigma_n), \tag{5.54}$$

где $\rho_{ij}(\sigma_n)$ определяется по формуле (5.53).

Предложен довольно простой прием, позволяющий свести эту задачу к поиску пути бродячего торговца на графе.

Зададимся вопросом: что произойдет, если мы в перестановке σ_n поменяем местами два каких-нибудь соседних станка i и j? Ясно, что все то, что направлялось к первому станку i (вдоль перестановки), будет теперь идти на единицу времени дольше, за исключением тех деталей, которые шли к i от второго станка j, для них время пути сократилось на n — 2 единиц. Все, что шло к второму станку j рассматриваемой пары, будет находиться в пути на единицу времени меньше, зато все, что шло от этого станка к другим, будет находиться в пути на единицу времени дольше. То, что шло к нему от соседнего станка i, будет находиться в пути на n — 2 единиц времени дольше (см. схему рис. 5.26).

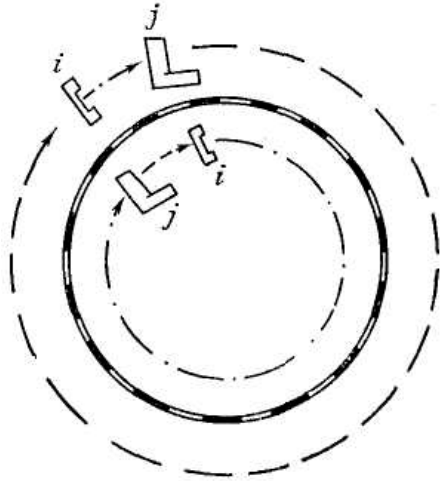


Рис. 5.26

Окончательный эффект от этой перестановки будет таким:

$$\varphi_{ij} = \sum_{l \neq j} a_{il} - \sum_{l \neq j} a_{il} + a_{ij}(n-2) + \\ + \sum_{l \neq i} a_{jl} - \sum_{l \neq i} a_{lj} - a_{ji}(n-2). \quad (5.55)$$

Очевидно, что если $\varphi_{ij}<0$, значит, выгодно поменять местами соседние станки i и j в перестановке σ_n — общее число «передач» при этом уменьшится, если $\varphi_{ij} \geq 0$, значит, этого не следует делать.

Рассмотрим пример, в котором исходные данные представлены табл. 5.6.

Таблица 5.6

Матрица a_{ij}

i \ j	1	2	3	4	5
1		30		100	40
2	10		20	20	50
3		40		20	30
4		10	100		
5	30	20	40	10	

Для каждой пары станков проведем вычисление того, в каком порядке удобно в перестановке σ_n располагать станки i и j при условии, что они соседние.

Вычисления рекомендуем сделать с помощью специальной расчетной таблицы (РТ) (табл. 5.7).

Таблица 5.7

Расчетная таблица (РТ)

+	−
I	II
x	y

Запишем согласно формуле (5.55):

1) все элементы строки i матрицы a_{ij} из табл. 5.6 (за исключением стоящего в столбце j) в графу II РТ;

2) все элементы, которые стоят в столбце с номером i (за исключением стоящего в строке j), в графу I РТ;

3) все элементы, которые стоят в строке j (за исключением i-го столбца), в графу I;

4) все элементы, которые стоят в столбце j (за исключением i-й строки), в графу II;

5) $(n — 2)\ a_{ij}$ в графу I;

6) $(n — 2)\ a_{ji}$ в графу II.

Суммируем все числа, выписанные в графе I, и полагаем x равным этой сумме; затем суммируем все числа, выписанные в графе II, и полагаем y равным этой сумме.

Если $x > y$ (что соответствует $\varphi_{ij} > 0$), в таблице S_{ij} в клетке i-й строки, j-м столбце пишем +, в клетке j-й строки, i-го столбца пишем —.

Если $x \le y$ (что соответствует $\varphi_{ij} \le 0$), в таблице - S_{ij} в клетке i-й строки, j-м столбце пишем —, в клетке j-й строки i-го столбца пишем +.

Знак + в таблице 5.9 означает, что переход от i к j при построении порфириана допустим. Знак -, что недопустим.

Описанные правила вычислений иллюстрирует табл. 5.8 для $i = 1$, $j = 2$.

Таблица. 5.8

+	—
30	100
	40
20	40
20	10
50	20
4 · 30	4 · 10
240	260

Согласно описанному правилу в клетку $i = 1$, $j = 2$ таблицы заносим —, а в клетку $j = 1$, $i = 2$ заносим +

Окончательный результат вычислений представлен табл. 5.9.

Таблица 5.9

i \ l	1	2	3	4	5
1		−	−	+	−
2	+		−	+	+
3	+	+		−	−
4	−	−	+		−
5	+	−	+	+	

Рис. 5.27 определяет граф, соответствующий таблице 5.9. Этот граф задает допустимые пути поиска перестановки - как и в задаче бродячего торговца как бы «направляя» построение порфириана.

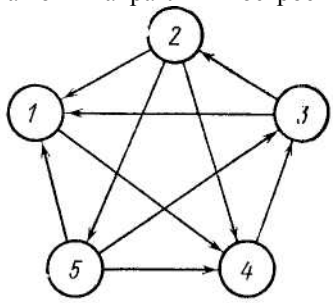


Рис. 5.27.

Нетрудно убедиться, что σ_n = (1, 4, 3, 2, 5) есть единственное решение нашей задачи (рис. 5.28).

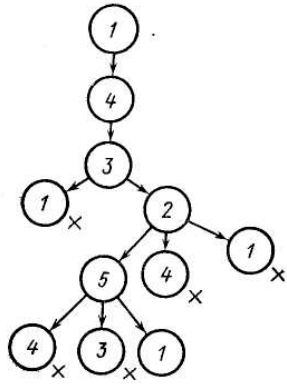


Рис. 5.28

При построении порфириана, соответствующего графу рис. 5.27, поиск маршрута удобно начинать с пункта i_1=1, в этом случае i_2 = 4 — единственное i_2 в порфириане.

3. Загадка маленькой мушки. Маленькая мушка - дрозофила - любимый объект экспериментирования генетиков: за одну неделю можно наблюдать несколько поколений мушки.

В некоторых экспериментах есть все основания полагать, что в результате эксперимента у мушки оказывается измененным целый участок хромосомы, что вызывает целую цепочку мутаций - изменения наследственных признаков.

В каждом опыте удается установить, какие признаки одновременно оказались захваченными мутациями.

Перенумеруем наблюдаемые признаки и составим таблицу результатов проведенных опытов, где будем отмечать на пересечении i -й строки и j -го столбца

$$\delta_{ij} = \begin{cases} 1, \textit{ если признаки } i \textit{ и } j \textit{ одновременно} \\ \quad \textit{мутировали в некотором опыте,} \\ 0, \textit{ если такого не наблюдалось.} \end{cases} \qquad (5.56)$$

По результатам проведенных экспериментов можно теперь чисто математически строить *генетическую карту* мушки-дрозофилы, т.е. определить, в какой последовательности расположено в хромосоме участки, ведающие изучаемыми признаками организма.

Математическая постановка задачи хорошо иллюстрируется примером, представленным на рис. 5.29, *а, б*.

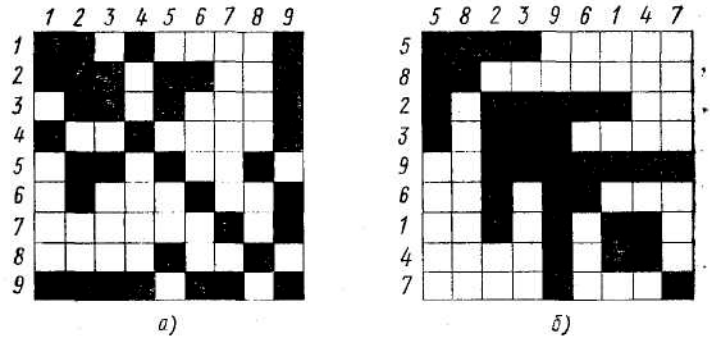


Рис. 5.29

Рис. 5.29, *а* представляет результаты эксперимента — закрашенные в черный цвет клетки соответствуют δ_{ij} = 1. Задача состоит в том,

чтобы перестановкой i (т.е. одновременно строк и столбцов) привести таблицу к виду, соответствующему рис. 5.29,*б*, где закрашенные клетки некоторым образом «прилегают» к главной диагонали таблицы.

Это свойство «прилегания» к главной диагонали можно выразить и аналитически:

$$\delta_{ij} \geq \delta_{ij+1} \text{ для всех } j \geq i. \qquad (5.57)$$

Изложим здесь метод решения нашей *задачи диагонализации* таблицы результатов эксперимента (матрицы δ_{ij}) путем построения порфириана (отметим, что существует и более эффективный метод решения задачи, точнее, лучше «отбраковывающий» варианты в операции разветвления).

Исследуем для начала некоторые свойства искомого решения.

Назовем «хвостом» (не путайте с «перестановочним хвостом») i для σ_k все те j, которые не вошли в σ_k и для которых $\delta_{ij} = 1$.

Утверждение 4. *Если* σ_k *и* $\sigma_{k+1} = \langle \sigma_k, j \rangle$ *— фрагменты искомого решения, то* j *входит во все (непустые) «хвосты»* σ_k.

Утверждение 5. *Пусть для* σ_k *выбраны два каких-то «хвоста». Если каждый из выбранных «хвостов» содержит по одном* j*, не входящему в другой «хвост», то полученный* σ_k *не может быть фрагментом искомого решения.*

Четвертое утверждение доказывается способом от обратного. Пусть j не входит в «хвост» некоторого i, принадлежащего σ_k, и пусть

$$\Sigma_n = \langle \sigma_k, j, \ldots, l, \ldots \rangle$$

-искомое решение, l — первый индекс из «хвоста» i.

Для σ_n

$$\delta_{ij} = 0, \quad \text{а} \quad \delta_{il} = 1,$$

в то время как $k(j) < k(l)$, что противоречит условию (5.57).

Докажем и пятое утверждение.

Действительно, пусть «хвост» i_1 содержит j_1, но не содержит j_2, а «хвост» i_2 не содержит j_1, но содержит j_2. Тогда в перестановке

$$\sigma^1{}_n = \langle \sigma_k, \ldots, j_1, \ldots, j_2, \ldots \rangle$$

т.е. при $k(j_1) < k(j_2)$,

$$\delta_{i_2 j_1} = 0, \text{ а } \delta_{i_1 j_1} = 1,$$

иначе говоря, нарушается свойство (5.57), в перестановке же

$$\sigma^2{}_n = \langle \sigma_k, \ldots, j_2, \ldots, j_1, \ldots \rangle$$

т.е. при $k(j_1) > k(j_2)$,

$$\delta_{i_1 j_2} = 0, \text{ а } \delta_{i_1 j_1} = 1,$$

т.е. снова же нарушается свойство (5.57).

Так в нашем примере при $\sigma_2 = \langle 1, 2 \rangle$ «хвост» 1 содержит $j = 4$, «хвост» 2 содержит $j = 5$. Если мы образуем перестановку

$$\sigma_k = \langle 1, 2, 4, 5, \ldots \rangle,$$

в ней $\delta_{24} = 0$, а $\delta_{25} = 1$. Если мы образуем перестановку

$$\sigma_n = \langle 1, 2, 4, 5, \ldots \rangle,$$

в ней $\delta_{15} = 0$, а $\delta_{14} = 1$.

Воспользуемся общей схемой построения порфириана, с той только разницей, что в операции «ветвления»

1) при переходе от σ_k к σ_{k+1} к σ_k можно присоединить только j, содержащееся во всех непустых «хвостах» для i, вошедших в σ_k; если все «хвосты» пустые — любую j (не из σ_k, конечно).

2) σ_k, удовлетворяющее условию утверждении 5, отмечаем (крестиком) как недопустимое, дальнейшему «ветвлению» неподлежащее.

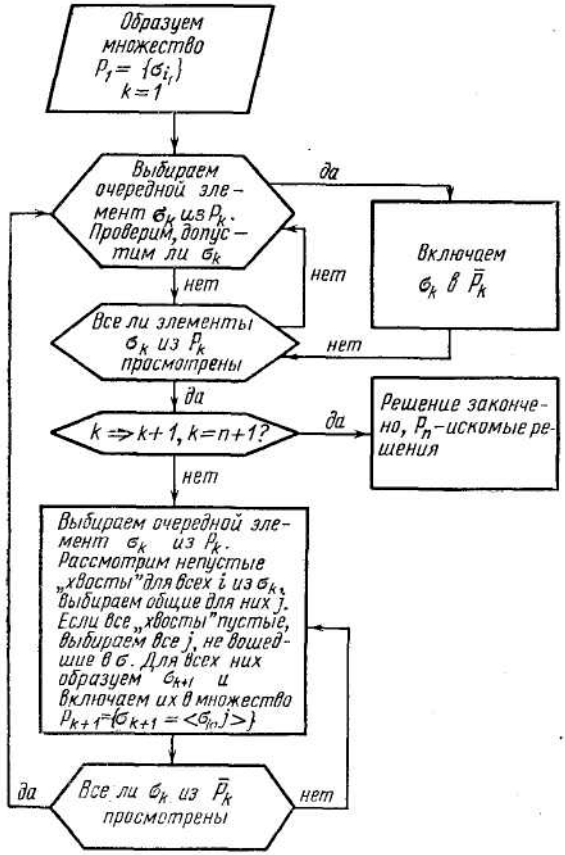


Рис. 5.30

Рис. 5.30 представляет общую схему алгоритма построения порфириана в нашем примере, табл. 5.10 - в немного своеобразной (приспособленной для записи) форме результат решения задачи.

Таблица 5.10

№	σ_k	Решение ?	№	σ_k	Решение ?
1	1, 2 ×	—	22	6, 2, 9 ×	—
2	1, 4, 9, 2 ×	—	23	6, 9, 2 ×	—
3	1, 9, 2 ×	—	24	7, 9, 1, 2 ×	—
4	1, 9, 4, 2 ×	—	25	7, 9, 1, 4, 2 ×	—
5	2, 1, ×	—	26	7, 9, 2 ×	—
6	2, 3, 5 ×	—	27	7, 9, 4, 1, 2 ×	—
7	2, 3, 9 ×	—	28	7, 9, 6, 2 ×	—
8	2, 5 ×	—	29	8, 5, 2, 3, 9, 1 ×	—
9	2, 6, 9 ×	—	30	8, 5, 2, 3, 8, 6, 1, 4, 7	да
10	2, 9 ×	—	31	8, 5, 3, 2 ×	—
11	3, 2, 5 ×	—	32	9, 1, 2 ×	—
12	3, 2, 9 ×	—	33	9, 1, 4, 2 ×	—
13	3, 5 ×	—	34	9, 2 ×	—
14	3, 9 ×	—	35	9, 4, 1, 2 ×	—
15	4, 1, 9, 2 ×	—	36	9, 6, 2 ×	—
16	4, 9, 1, 2 ×	—	37	9, 7, 1, 2 ×	—
17	5, 2 ×	—	38	9, 7, 1, 4, 2 ×	—
18	5, 3 ×	—	39	9, 7, 2 ×	—
19	5, 8, 2, 3, 9, 1 ×	—	40	9, 7, 4, 1, 2 ×	—
20	5, 8, 2, 3, 8, 6, 1, 4, 7	да	41	9, 7, 6, 2 ×	—
21	5, 8, 3, 2 ×	—			

Обратите внимание: вместо нескольких сотен тысяч перестановок (если пользоваться методом перебора) в нашем методе оказалось достаточным построить всего 41 фрагмент σ_k.

5.17. Перестановочный метод

1. Решение задачи директора. Хороший директор интуитивно давно определил, в какой очередности следует принимать посетителей: прежде всего необходимо переговорить с теми, кто пришел с короткими вопросами, оставив «на потом» вопрос посложнее.

Это правило может быть строго доказано.

Предположим

$$\sigma^1_n = \langle i_1 ..., i, j ..... i_n \rangle$$

- оптимальное решение. Напомним, что

$$F(\sigma) = \sum_i \underline{t}_i(\sigma). \tag{5.58}$$

Здесь $\underline{t}_i(\sigma)$ — время начала приема i-го посетителя. Поменяем порядок приема каких-то двух посетителей, принимаемыз в σ^1_n один непосредственно за другим:

$$\sigma^2_n = \langle i_1,..., j, i,.....,i_n \rangle .$$

Можно записать, что

$$\begin{aligned} F(\sigma^1_n) &= A + \underline{t}_i(\sigma^1_n) + \underline{t}_j(\sigma^1_n), \\ F(\sigma^2_n) &= A + \underline{t}_i(\sigma^2_n) + \underline{t}_j(\sigma^2_n); \end{aligned} \tag{5.59}$$

A — общая неизменная часть $F(\sigma)$, так как времена ожидания в приемной посетителей, имеющих номер, отличный от i и j, не изменились в σ^2_n по сравнению с σ^1_n.

Так как σ^1_n — по предположению оптимальное решение, то

$$F(\sigma^1_n) \leqslant F(\sigma^2_n). \tag{5.60}$$

Пусть τ — момент окончания приема предыдущего посетителя для i-го в σ^1_n, j-го в σ^2_n.

Тогда для σ^1_n

$$\underline{t}_i(\sigma^1_n) = \tau, \quad \underline{t}_j(\sigma^1_n) = \tau + T_i,$$

для σ^2_n

$$\underline{t}_j(\sigma^2_n) = \tau, \quad \underline{t}_i(\sigma^2_n) = \tau + T_j.$$

Согласно (4.87) и (4.88)

$$A + \tau + \tau + T_i \leq A + \tau + \tau + T_j$$

или

$$T_i \leq T_j. \tag{5.61}$$

Последнее как раз и означает, что решение оптимально тогда, когда принята очередность согласно правилу: «*предыдущая не длительней последующей*». Это правило носит также название *правила кратчайшей операции:* первым в оптимальном решении выбирается i с наименьшей длительностью T_i.

2. Задача одного станка. Правила, которые позволяют найти решение без какого-либо перебора вариантов, носят название *решающих правил.* Такие решающие правила могут быть найдены для многих практических постановок. Это позволяет избежать построения порфириана и выполнения в связи с этим множества сложных логических и вычислительных операций.

Небольшое осложнение задачи директора приводит к формулировке так называемой *задачи одного станка,* которая нередко встречается в приложениях.

Директор должен рассмотреть n вопросов; длительность рассмотрения i-го вопроса равна T_i; по вопросу i к директору заходит α_i посетителей; допустим, что нет посетителя, который пришел одновременно по двум вопросам.

Задача и в этом случае сводится к установлению такой очередности рассмотрения вопросов, чтобы общее время ожидания посетителями приема было наименьшим, т.е. к поиску перестановки

$$\sigma_г = \langle i_1, i_2, \ldots . i_n \rangle \ ,$$

с функцией-критерием

$$F(\sigma_n) = \sum_i \alpha_i t_i . \tag{5.62}$$

Если считать, что α_i — не только целые числа, мы получим задачу одного станка: установить очередность обработки n различных деталей, если время обработки i-й детали T_i известно заранее; по детали i за каждую минуту ожидания обработки взимается «штраф» α_i

Нетрудно повторить все рассуждения предыдущего пункта, чтобы обнаружить важное свойство оптимального варианта.

Пусть

$$\sigma^1{}_n = \langle i_1, \ldots, i, j, \ldots . , i_n \rangle$$

- оптимальное решение, а

$$\sigma^2{}_n = \langle i_1, \ldots, j, i, \ldots . , i_n \rangle$$

— вариант, который отличается перестановкой только двух каких-то соседних элементов в $\sigma^1{}_n$.

И в этом случае, так как $\sigma^1{}_n$ - по предположению оптимально,

$$F(\sigma_n^1) \leqslant F(\sigma_n^2). \tag{5.63}$$

Обозначив $t_i(\sigma^1{}_n)$ через τ, как и в предыдущем пункте, получим, что неравенство (5.63) эквивалентно

$$\alpha_i \tau + \alpha_j (\tau + T_i) \leqslant \alpha_j \tau + \alpha_i (\tau + T_j)$$

или

$$\alpha_j T_i \leq \alpha_i T_j$$

а так как все $T_i > 0$, то

$$\frac{\alpha_i}{T_i} \geqslant \frac{\alpha_j}{T_j}. \qquad (5.64)$$

Последнее выражение и устанавливает свойство оптимального решения, которое легко позволяет это решение построить.

Утверждение (решающее правило*)*.

Перестановка

$$\sigma^1{}_n = \langle i_1, i_2, \ldots, i_n \rangle$$

тогда есть решение задачи одного станка по критерию (5.62), *когда*

$$\frac{\alpha_{i_1}}{T_{i_1}} \geqslant \frac{\alpha_{i_2}}{T_{i_2}} \geqslant \ldots \geqslant \frac{\alpha_{i_n}}{T_{i_n}}. \qquad (5.65)$$

Прием исследования, который мы успешно применили и в задаче директора, и в задаче одного станка, получил название *перестановочного приема.*

3. Задача двух станков. Еще одно эффективное применение перестановочного приема — *задача двух станков.*

В этой задаче требуется за минимальное время закончить обработку n деталей. Каждая деталь i обрабатывается сначала на первом станке (первая операция), продолжительность обработки равна a_i, потом на втором станке (вторая операция), продолжительность обработки b_i.

Табл. 5.11 представляет исходные данные простого иллюстративного примера задачи двух станков, рис. 5.31 — временнỳю диаграмму для этого примера $\sigma_1 = (1, 2, 3, 4, 5, 6)$.

Таблица 5.11

i	1	2	3	4	5	6
a_i	3	2	4	1	4	2
b_i	1	4	3	2	2	5

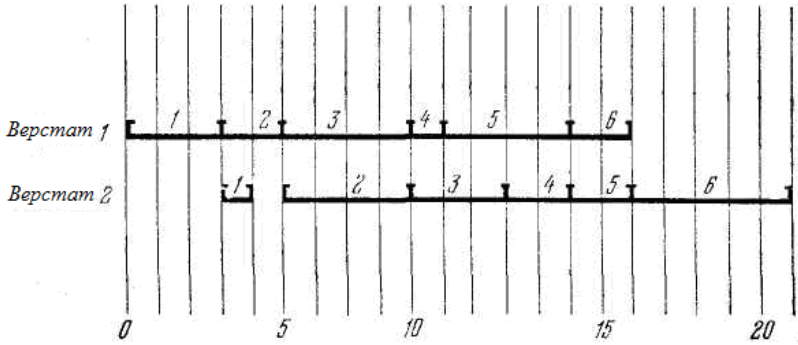


Рис. 5.31

Важно отметить, что для задачи двух станков вторая операция не может начать выполняться, пока не закончилась предыдущая, а также пока станок еще занят выполнениям предыдущей операции.

Покажем, что и задача двух станков сводится к поиску экстремальной перестановки.

Для этого, прежде всего, надо установить, что в оптимальном решении не может быть такого, чтобы для двух каких-то деталей *i, j* на первом станке операции выполнялись в очередности *i* p *j* (p — знак *«предшествует»),* а на втором станке *j* p *i.* Такой случай демонстрируется несколько схематично на рис. 5.32.

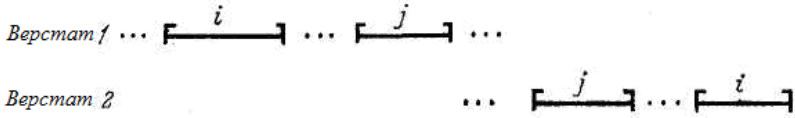


Рис. 5.32

Рис. 5.32 помогает увидеть, что в графике, отображенном на рисунке, можно деталь *i* на первом станке начать обрабатывать непосредственно после детали *j*, «подвинув» деталь *j* (вместе с операциями, которые занимали прежде место между *i* и *j*) влево на отрезок, равный продолжительности выполнения первой операции *i.* Такое преобразование графика допустимо, так как сдвиг первых операций влево может только уменьшить сроки начала выполнения вторых операций. Благодаря этому, время окончания выполнения всех операций также может только уменьшиться, а вместе с ним и срок окончания обработки всех деталей.

Итак, мы можем считать, что в оптимальном решении порядок обработки деталей на первом станке и на втором один и тот же и задается перестановкой

$$\Sigma_n = \langle i_1, i_2, \ldots, i_n \rangle \tag{5.66}$$

Рассмотрим некоторый фрагмент графика $\sigma_k = \langle i_1, i_2, \ldots, i_k \rangle$, $k < n$. Введем обозначение: $A(\sigma_k)$— время окончания выполнения обработки деталей графика σ_k первым станком, $B(\sigma_k)$ - вторым станком.

Очевидно, что

$$\begin{aligned} A(\sigma_{k+1}) &= A(\sigma_k) + a_i, \\ B(\sigma_{k+1}) &= \max [A(\sigma_{k+1}),\ B(\sigma_k)] + b_i, \end{aligned} \tag{5.67}$$

где $\sigma_{k+1} = \langle \sigma_k, i \rangle$ (рис. 5.33). Критерий оптимальности, следовательно, — минимизация $B(\sigma_n)$

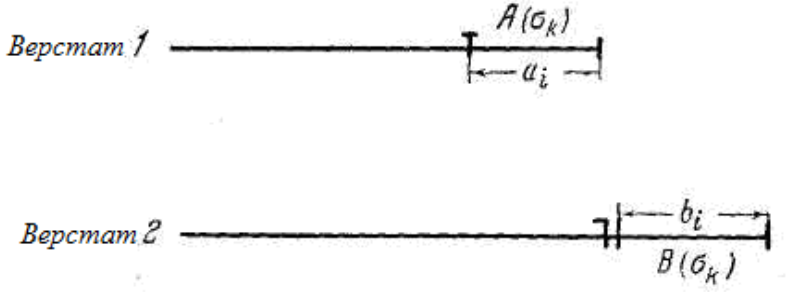


Рис. 5.33

Пусть теперь

$$\sigma^1_n = \langle i_1, \ldots, i_k, i, j, \ldots, i_n \rangle$$

- некоторое решение. Построим

$$\sigma^2_n = \langle i_1, \ldots, i_k, j, i, \ldots, i_n \rangle$$

Обозначим

$$A(\sigma_k) = \tau, \qquad B(\sigma_k) = \tau + \Delta.$$

Тогда в σ^1_n

$$\begin{aligned} A(\sigma^1_{k+1}) &= \tau + a_i, \\ B(\sigma^1_{k+1}) &= \max(\tau + a_i;\ \tau + \Delta) + b_i, \end{aligned}$$

далее,

$$\begin{aligned} A(\sigma^1_{k+2}) &= \tau + a_i + a_j, \\ B(\sigma^1_{k+2}) &= \max(\tau + a_i + a_j;\ \max(\tau + a_i,\ \tau + \Delta) + b_i) + b_j \end{aligned}$$

или

$$B(\sigma^1_{k+2}) = \tau + b_j + \max(a_i + a_j,\ a_i + b_i,\ \Delta + b_i). \tag{5.68}$$

Аналогично для σ^2_n

$$A(\sigma^2_{k+2}) = \tau + a_j + a_i,$$
$$B(\sigma^2_{k+2}) = \tau + b_i + \max(a_j + a_i,\ a_j + b_j,\ \Delta + b_j).$$

Нетрудно сообразить, что если

$$B(\sigma^1_{k+2}) \leqslant B(\sigma^2_{k+2}), \tag{5.69}$$

это первое решение лучше второго, так как второй станок высвобождается раньше, следовательно, и $B(\sigma^1_n)$ может быть только меньше, чем во втором случае.

Очевидно, неравенство (5.69) эквивалентно следующему:

$$\tau + b_j + \max(a_i + a_j,\ a_i + b_i,\ \Delta + b_i) \leqslant$$
$$\leqslant \tau + b_i + \max(a_j + a_i,\ a_j + b_j,\ \Delta + b_j)$$

или

$$\tau + b_j + b_i + a_i + a_j + \max(-b_i,\ -a_j,\ \Delta - a_i - a_j) \leqslant$$
$$\leqslant \tau + b_i + b_j + a_j + a_i + \max(-b_j,\ -a_i,\ \Delta - a_i - a_j),$$

или

$$\max(-b_i,\ -a_j,\ \Delta - a_i - a_j) \leqslant$$
$$\leqslant \max(-b_j,\ -a_i,\ \Delta - a_i - a_j). \tag{5.70}$$

Неравенство (5.70) выполняется, если выполняется

$$\max(-b_i,\ -a_j) \leqslant \max(-b_j,\ -a_i). \tag{5.71}$$

Неравенство (5.71) эквивалентно:

$$\min(a_i,\ b_j) \leqslant \min(a_j,\ b_i). \tag{5.72}$$

Итак, если соотношение (5.72) имеет место, то $i \rho j$ (i предшествует j) в σ_n. Нетрудно обнаружить, что (5.72) выполняется, если

$$\begin{array}{ll} \text{а)} & a_i \leq b_i,\ \ a_j \leq b_j \ \ a_i \leq a_j; \\ \text{б)} & a_i \leq b_i,\ \ \ a_j > b_j; \\ \text{в)} & a_i \geq b_i,\ \ a_j \geq b_j,\ \ \ a_j > b_i. \end{array} \tag{5.73}$$

Последнее утверждение эквивалентно следующему.

Утверждение (решающее правило). *Пусть имеется k таких i, что* $a_i \leq b_i$.

Тогда, если $a_{il} \leq b_{il}$ *для* $l \leq k$ *и, кроме того,* $a_{il} \leq a_{i,i+1}$ *при* $l < k$; $a_{il} > b_{il}$ *для* $l > k$ *и, кроме того,* $b_{il} \geq b_{i,j+1}$ *при* $l > k$, *то*

$$\sigma_n = \langle i_1, \ldots, i_k;\ i_{k+1}, \ldots, i_n \rangle \tag{5.74}$$

оптимально.

Согласно утверждению сначала выбираются детали i, в которых первая операция a_i короче второй операции b_i. Эти детали обрабатываются в порядке возрастания a_i. Остальные детали обрабатываются в порядке убывания b_i.

Алгоритм, который представлено на рис. 5.34, воплощает процедурно приведенное решающее правило, а рис. 5.35 демонстрирует оптимальное решение (4, 2, 6, 3, 5, 1) нашего примера.

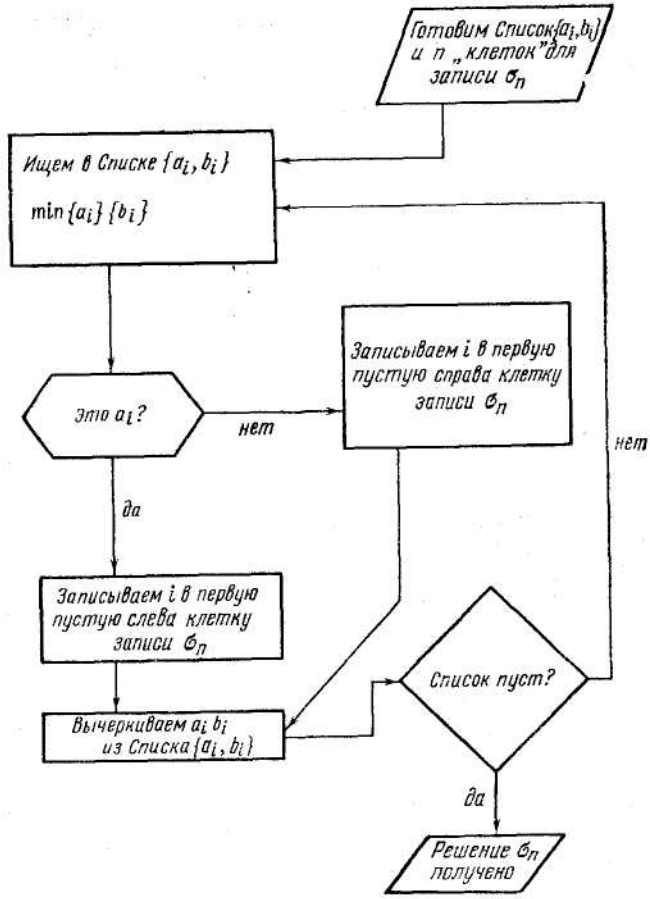


Рис. 5.34

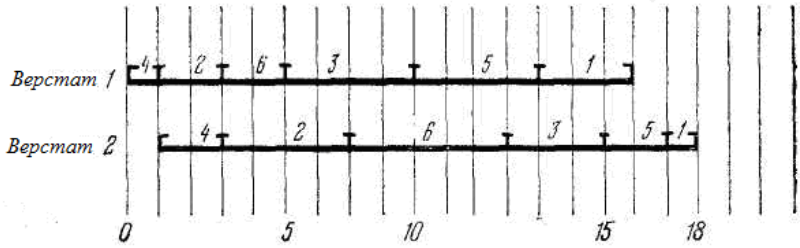


Рис. 5.35

4. Интервалы очередности. Если в задаче одного станка $\varphi_i(\underline{t}) = \alpha_i \underline{t}^2$, то применение перестановочного приема приведет нас к сравнению двух функций от τ:

$$2\alpha_j \tau T_i + \alpha_j T_i^2 \quad \text{и} \quad 2\alpha_i \tau T_j + \alpha_i T_j^2. \tag{5.75}$$

Сравнение (5.76) эквивалентно выяснению знака следующего выражения:

$$2(\alpha_j T_i - \alpha_i T_j)\tau + \alpha_j T_i^2 - \alpha_i T_j^2. \tag{5.76}$$

Если

$$\alpha_j T_i - \alpha_i T_j = 0, \text{ т. е. } \frac{\alpha_i}{T_i} = \frac{\alpha_i}{T_i},$$

то

$$\text{при } \alpha_j T_i^2 - \alpha_i T_j^2 \leqslant 0 \quad i \to j,$$
$$\text{при } \alpha_j T_i^2 - \alpha_i T_j^2 > 0 \quad j \to i.$$

Если $\alpha_j T_i - \alpha_i T_j \neq 0$, то определим

$$\tau_{ij}^0 = -\frac{\alpha_j T_i^2 - \alpha_i T_j^2}{2(\alpha_j T_i - \alpha_i T_j)}. \tag{5.77}$$

Теперь из (5.77) следует, что (пусть $\frac{\alpha_i}{T_i} \geq \frac{\alpha_j}{T_j}$)

$$\text{при } \tau \geqslant \tau_{ij}^0 \quad i \to j,$$
$$\text{при } \tau < \tau_{ij}^0 \quad j \to i. \tag{5.78}$$

Там, где $\tau^0_{ij} \le 0$ или $\tau^0_{ij} \ge \sum_i T_i$, можно заранее установить согласно (5.78) $i \rhd j$ или $j \rhd i$. Если τ^0_{ij} попадает в интервал $(0, \sum_i T_i)$, вопрос $i \rhd j$ или $j \rhd i$ решается по-другому.

Общая схема решения в таком случае — построение порфириана. В операции ветвления конкретно определяется в зависимости от $\overline{t}(\sigma_k)$, какие продолжения $\sigma_{k+2} = \langle \sigma_k, i, j \rangle$ согласно (5.78) могут быть разрешены, какие запрещены.

Решим пример задачи одного станка, представленный табл. 5.13, где нужно минимизировать $F(\sigma_n) = \sum \alpha_i t_i(\sigma)$.

В табл. 5.13 операции перенумерованы в порядке убывания $\frac{\alpha_i}{T_i}$.

Таблица 5.13

i	1	2	3	4	5
α_i	6	1	2/9	1	1/6
T_i	3	1	1/3	2	1/2

Вычислим по формуле (5.77) значение τ_{ij} и результаты вычислений представим в табл. 5.14, где в клетках (i, j) указаны соответствующие *интервалы очередности:* в этих интервалах $[a, b)$ $i \rhd j$, если только $a \le \tau < b$.

Так, для $i = 1$, $j = 2$ по формуле (5.73) имеем

$$\tau^0_{12} = -\frac{1 \cdot 3^2 - 6 \cdot 1^2}{2(1 \cdot 3 - 6 \cdot 1)} = \frac{1}{2}.$$

Таблица 5.14

i \ j	1	2	3	4	5
1	—	$\left[\frac{1}{2}, \infty\right)$	$\left[\frac{1}{2}, \infty\right)$	$[0, \infty)$	$[0, \infty)$
2	$\left[0, \frac{1}{2}\right)$	—	$\left[\frac{1}{2}, \infty\right)$	$[0, \infty)$	$[0, \infty)$
3	$\left[0, \frac{1}{2}\right)$	$\left[0, \frac{1}{2}\right)$	—	$[0, \infty)$	$[0, \infty)$
4	—	—	—	—	$\left[\frac{5}{4}, \infty\right)$
5	—	—	—	$\left[0, \frac{5}{4}\right)$	—

Таким образом, $1 \rho 2$ для всех $\tau \geq 1/2$, что и отмечено в табл. 5.14 в клетке (1, 2) интервалом очередности $[1/2, \infty)$.

Отсюда следует, что $2 \rho 1$ для всех τ при $-\infty < \tau < 1/2$, что и отмечено в клетке (2, 1) интервалом очередности $[0, 1/2)$ (так как в наших условиях $\tau \geq 0$).

Для $i = 1$, $j = 4$ вычисление по формуле (4.101) дают $\tau^0_{14} = -(5/6) < 0$. Отсюда $1 \rho 4$ при всех $\tau \geq 0$, так что $4 \rho 1$ не может ни при каком интересующем нас τ.

Построение порфириана приведет нас к трем возможным решениям:

$$\langle 2, 1, 3, 4, 5 \rangle ,$$
$$\langle 3, 1, 2, 4, 5 \rangle , \qquad (5.79)$$
$$\langle 3, 2, 1, 4, 5 \rangle .$$

Метод перебора, применяемый к полученным тремх возможным решениям (6.79), позволяет определить, что оптимальное решение нашего примера есть

$$\sigma_n = \langle 2, 1, 3, 4, 5 \rangle , \quad F(\sigma_n) = 35.$$

Микромодуль16.
Индивидуальные тестовые задачи

Упражнение 1. Составьте временнỳю диаграмму выполнения работ для сетевого графика, представленного на рис. 5.15 и рис. 5.16, исходя из целей подготовки вечера в кратчайший срок. Можно ли сдвинуть во времени (передвинуть вправо на один день отрезок во временнòй диаграмме) работу 12? 14? Насколько можно сдвигать сроки выполнения работ, лежащих и не лежащих на критическом пути?

Упражнение 2. Попробуйте на примерах отыскать критический путь и написать общую инструкцию определения критического пути в сетевом графике.

Упражнение 3. Ответить, в чем заключаются трудности составления наилучшего расписания? наилучшего футбольного календаря? и что означает в этих случаях «наилучшее расписание»?

Упражнение 4. Покажите, что задачи поиску очередности приема посетителей по критерию минимизации общего времени ожидания в приемной, по критерию минимизации среднего времени ожидания в приемной

$$F(\sigma_n) = \frac{\sum_k \tau_k(\sigma_n)}{n}$$

и по критерию минимизации общего времени нахождения в приемной, включая и время приема

$$F(\sigma_n) = \sum_k \left(\tau_k(\sigma_n) + T_{i_k}\right),$$

эквивалентны. Более общо: умножение функции-критерия на константу или прибавление к ней константы приводит к эквивалентной постановке.

Упражнение 5. Попытайтесь для начала рассадить наилучшим образом студентов, перечисленных на рис. 5.17, за тремя столами так, чтобы за каждым столом сидели студент и студентка, а суммарная польза от размещения за столами складывалась из отдельных выгод a_{ik}, согласно табл. 5.2 (составленной из экспертных оценок).

Таблица 5.2

i \ k	Оля	Зоя	Лиля
Саша	5	4	3
Андрей	2	2	4
Сережа	4	1	3

Упражнение 6. Постройте и запишите строгий алгоритм решения задачи о назначениях и докажите, что он всегда приводит к наилучшему решению?

Упражнение 7. В группе не поровну юношей и девушек. Можно ли и в этом случае задачу о размещении студентов за столами свести к решению задачи о назначениях? Точнее: если вы уже знаете, как решать задачу о назначениях, как бы вы решали задачу о размещении студентов за столами, если в группе не поровну юношей и девушек?

Упражнение 8. *Задача бродячего торговца или коммивояжера* формулируется следующим образом: некий коммивояжер должен объехать n городов, чтобы в каждом побывать только один раз, и совершить это, сделав минимальный путь (расстояния a_{ik} между любыми двумя городами i и k заданы). Покажите, что задачу бродячего торговца можно свести к поиску экстремальной перестановки. Что вы можете сказать о функции-критерии в этой задаче? В чем существенное отличие формальных постановок задачи бродячего торговца от задачи о назначениях? Изменилась бы постановка задачи, если бы вместо расстояний указывалось время путешествия между городами и требовалось совершить объезд городов за минимальное время? (В такой постановке задача бродячего торговца может рассматриваться как задача составления расписания).

Сформулируйте задачу разнесения приглашений на вечер как задачу бродячего торговца. Проверьте, рационально ли вы обходите квартиру, убирая ее пылесосом.

Упражнение 9. Улучшите разъяснение Алгоритма-1. Докажите, что по Алгоритму-2 действительно получают все n-перестановки?

Упражнение 10. Предложить алгоритм получения всех n-перестановок, отличный от изложенных? Докажите, что по алгоритму-2 возможно получить действительно все перестановки (а возможно – нет).

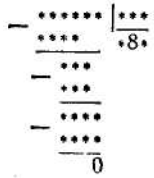
Упражнение 11. Докажите утверждения 1, 2, 3 п. 5.16.

Упражнение 12. Постройте порфириан для примера рис. 5.25 и покажите, что только варианты (5.52) удовлетворяют условиям задачи, отмечая крестиком как недопустимый вариант с повторным попаданием в город.

Упражнение 13. Вы получили записку, где написано: абвллеююя. Попробуйте расшифровать таинственную запись, построив всевозможные осмысленные буквосочетания с помощью порфириана («осмысленность» σ_k может быть проверена по словарю).

Упражнение 14. Найдите решение известной задачи о волке, козе и капусте с помощью порфириана. Напомним задачу. Нужно переправить через реку волка, козу и капусту, в лодке же можно поместить только один предмет. Если оставить козу с капустой, коза съест капусту, если оставить волка с козой, волк съест козу.

Упражнение 15. Решите с помощью порфириана следующую задачу, где требуется восстановить стершиеся цифры:



Пользуясь перестановочным приемом, выполните следующие упражнения.

Упражнение 16. Введем в задаче одного станка критерий

$$F(\sigma_n) = \sum_i \varphi_i(t_i), \qquad (*)$$

назовем его *суммой штрафов,* $\varphi_i(t_i)$ - функция штрафа. Найдите решающее правило, если

$$\varphi_i(t) = \alpha_i + \beta_i t,$$

где $\alpha_i > 0$, $\beta_i > 0$ заданы.

Упражнение 17. Найдите решающее правило в задаче одного станка, если $\varphi_i(t) = \alpha_i t^{\beta_i t}$.

Упражнение 18. Заданы последовательности $\{\alpha_i\}$ и $\{\beta_i\}$, $i = 1, 2.....n$, при этом

$$\alpha_1 \leqslant \alpha_2 \leqslant \ldots \leqslant \alpha_n.$$

Покажите, что

$$\sum_i \alpha_i \beta_i$$

максимальна, если

$$\beta_1 \leqslant \beta_2 \leqslant \ldots \leqslant \beta_n,$$

и минимальна, если

$$\beta_1 \geqslant \beta_2 \geqslant \ldots \geqslant \beta_n.$$

Упражнение 19. В задаче директора и в задаче одного станка с самого начала мы молчаливо предполагали, что оптимальное решение *плотно,* т.е. что не делается перерывов между приемами посетителей (в обработке деталей). Покажите, что свойство быть плотным оптимального расписания в задаче одного станка связано с *монотонностью* изменения (возрастания, например) функций φ_i в (*).

Упражнение 20. Покажите, что если

$$\varphi_i(t) = \begin{cases} \alpha_i(t-\tau_i) & \text{при } t \geqslant \tau_i, \\ \beta_i(\tau_i - t) & \text{при } t \leqslant \tau_i, \end{cases}$$

то оптимальное расписание, вообще говоря, не обязательно плотно. (К такой постановке, например, сводится задача управления взлетами и посадками в аэропорту).

Упражнение 21. В какой очередности следует выполнять сегодня домашние задачи, если оценить для каждого предмета i влияние утомляемости через $\alpha_i \underline{t}$, где $\underline{t}$ — время начала работы над i-м предметом.

Упражнение 22. Покажите, что если $\alpha_i \geq \alpha_j$ $T_i \leq T_j$, то $i \rho j$ при $\tau \geq 0$.

Упражнение 23. Воспользовавшись результатом упражнения 22, покажите, что в примере, заданном табл. 5.12,

$$\sigma_n = \langle 5, 4, 3, 2, 1 \rangle$$

- оптимальное решение при

$$\varphi_i(\underline{t}) = \alpha_i \underline{t}_i^2.$$

Таблица 5.12

t	1	2	3	4	5
α_i	1	2	2	3	3
T_i	3	3	2	2	1

Упражнение 24. Постройте порфириан, соотетствующий табл. 5.14.

Указание. Например, $i_1 = 4$ недопустимо, так как в момент $\tau = 0$ $i = 4$ согласно табл. 5.14 не может предшествовать никакому j. Сочетание $\sigma_2 = \langle 1, 4 \rangle$ в момент $\tau=0$ допустимо, более того, представляется, что

допустимо и $\sigma_3 = \langle 1, 4, 5 \rangle$, так как при $\tau = \overline{t_1} = 3$, согласно табл. 5.14, 4p 5. Однако дальнейшее исследование показывает, что $\sigma_3 = \langle 1, 4, 5 \rangle$ недопустимо, так как в момент $\tau = \overline{t_4} = 5$ $i=5$ не может предшествовать никакому j.

Микромодуль 19
Методы отсеивания вариантов

5.18. Последовательное отсеивание вариантов. Доминирование.

1. Отсеивание по правилам доминирования. В предыдущих разделах мы продемонстрировали, как выявление отдельных свойств оптимального решения может быть использовано для сокращение множества рассматриваемых при построении порфириана вариантов (вплоть до ликвидации вообще всякой вариантности – п. 5.17).

Характерным для методов п. 5.17 было, что эти свойства устанавливались заранее и проверялись с целью (полного) «запрещения» ветвления в процессе построения порфириана. В п. 5.16 и п. 5.17 такое запрещение достигалось только частично.

Ниже мы рассмотрим другие способы последовательного отсеивания вариантов при построении порфириана путем выявления некоторых свойств решений в процессе ветвления.

Весьма плодотворной для такого отсеивания вариантов появилась очевидная идея перенесения процедуры метода перебора вариантов с n-го, последнего уровня порфириана на более высокие его уровни.

Однако если при последовательном просмотре вариантов σ_n в методе перебора надо было сравнивать их по значению $F(\sigma_n)$, чтобы установить, какой из них лучший (в конце перебора — оптимальный), для фрагментов решений σ_k, этого недостаточно.

Фрагменты σ^1_k и σ^2_k, по существу, как это уже говорилось, в порфириане представляют собой множества вариантов σ_n, которые начинаются с фиксированных σ^1_k и σ^2_k. Сказать, что σ^1_k лучше σ^2_k — это значит установить, что среди всех продолжений σ^1_k (т.е. элементов σ_n, которые начинаются с фиксированного σ^1_k) найдется элемент σ^1_k с $F(\sigma^1_k)$, меньший (для задач минимизации), чем $F(\sigma^2_k)$ любого продолжения σ^2_k.

Мы будем называть *правилами доминирования* процедуры, позволяющие относительно некоторых σ^1_k и σ^2_k устанавливать, что какое-то из них лучше в указанном выше смысле.

В предыдущей фразе акцентировано внимание на том, что правила доминирования устанавливаются только для некоторых σ^1_k и σ^2_k . Обычно сравнивать можно только те σ^1_k и σ^2_k , которые приводят *в одно и тот же состояние*, в том понимании, что как результат какого-то процесса их можно считать *идентичными*.

Так в задаче коммивояжера в одно и тот же состояние приводят два пути σ^1_k и σ^2_k , если

а) они начинаются в одном и том же пункте, т.е.

$$i_1(\sigma^1_k) = i_1(\sigma^2_k);$$

б) они заканчиваются в одном и том же пункте, т.е.

$$i_k(\sigma^1_k) = i_k(\sigma^2_k);$$

в) эти пути проходят через одно и то же множество пунктов (только в разном порядке).

Последнее мы будем обозначать символом

$$M(\sigma^1_k) = M(\sigma^2_k);$$

$M(\sigma_k)$ -это множество пунктов i, через которые проходит σ_k (σ_k, таким образом, — это некоторое упорядочение $M(\sigma_k)$). Такие варианты σ^1_k и σ^2_k мы назовем *сравнимыми.*

Так, пути $\sigma^1_5 = \langle 2, 1, 3, 5, 4 \rangle$ и $\sigma^2_5 = \langle 2, 5, 3, 1, 4 \rangle$ приводят в одинаковые состояния, пути $\sigma^1_5 = \langle 2, 1, 3, 4, 5 \rangle$ и $\sigma^2_5 = \langle 2, 3, 1, 5, 4 \rangle$ -в разные.

В дальнейшем (упражнение 1) станет ясно, что в некоторых случаях варианты $\sigma^1_4 = \langle 2, 1, 3, 4 \rangle$ и $\sigma^2_5 = \langle 2, 3, 1,5,4 \rangle$ также могут стать сравнимыми.

Пусть в задаче бродячего торговца

$$F(\sigma_k) = \sum_{l=1}^{k-1} a_{i_l i_{l+1}}.$$

Утверждение (правило доминирования). *В задаче бродячего торговца, если σ^1_k и σ^2_k сравнимы и при этом $F(\sigma^1_k) \le F(\sigma^2_k)$, то σ^1_k лучше σ^2_k в том смысле, что среди продолжений σ^1_k найдется σ^1_n, для которого $F(\sigma^1_n) < F(\sigma^2_n)$ для всех σ^2_n, являющихся продолжениями σ^2_k.*

Тем самым дальнейшие ветвления σ^2_k при построении порфириана явно нецелесообразны — в этом и состоит *последовательное отсеивание* вариантов в процессе построения порфириана *согласно правилам доминирования.*

Пусть

$$\sigma_n^2 = \langle i_1^2,\ i_2^2,\ \ldots,\ i_k^2;\ i_{k+1}^2,\ \ldots,\ i_n^2 \rangle$$

— лучшее решение среди продолжений $\sigma^2{}_k = \langle \overline{i}_1^{\,2}, \ldots, \overline{i}_k^{\,2} \rangle$,

$$F(\sigma_n^2) = F(\sigma_k^1) + \varphi(\sigma_k^2) \leqslant F(\sigma_n^2),$$

где

$$\varphi(\sigma_k^2) = \sum_{l=k+1}^{n-1} a_{i_l^2 i_{l+1}^2} + a_{i_n^2 i_1^2}.$$

Построим теперь решение

$$\sigma_n^1 = \langle i_1^1, \ldots,\ i_k^1,\ i_{k+1}^1, \ldots,\ i_n^1 \rangle,$$

продолжение $\sigma^1{}_k = \langle\ i^1{}_1\ \ldots..i^1{}_k \rangle$ которого совпадает с продолжением $\sigma^2{}_k$, т.е. $i^1{}_l = i^2{}_l$ для всех $l > k$.

Тогда нетрудно видеть, что так как

$$F(\sigma_k^1) \leqslant F(\sigma_k^2),$$

то

$$F(\sigma_n^2) = F(\sigma_k^1) + \varphi(\sigma_k^2) \leqslant F(\sigma_n^2),$$

что и доказывает наше утверждение.

2. Решение задачи бродячего торговца. Рассмотрим пример, заданный табл. 5.15, представляющий матрицу (a_{ij}) расстояний между пунктами i и j.

Таблица 5.15

Матрица a_{ij}

i \ j	1	2	3	4	5	6
1		7	4	7	8	2
2	6		7	6	4	8
3	5	8		6	7	5
4	8	7	6		7	6
5	7	4	8	6		8
6	2	7	7	7	7	

Рис. 5.36 иллюстрирует *схему решения задачи* (упрощенно изложенный алгоритм) бродячего торговца методом отсеивания вариантов в процессе построения порфириана согласно правилам доминирования.

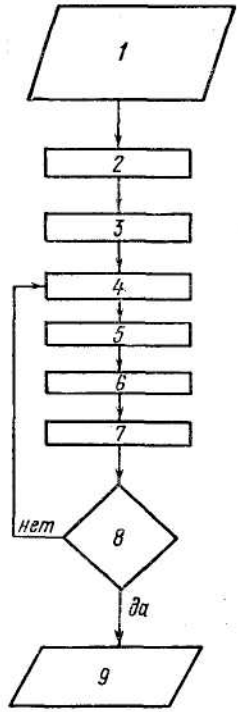


Рис. 5.36. Схема решения задачи бродячего торговца последовательным отсеиванием вариантов.

1. Подготовка матрицы a_{ij} к решению. 2. Разветвление $\sigma_1 = \langle 1 \rangle$, построение множества P_2 всех σ_2. 3. k=2. 4. Операция ветвления; каждому σ_k из A_k ставятся в соответствие всевозможные $\sigma_{k+1} = \langle \sigma_k, j \rangle$, j не принадлежит $M\langle \sigma_k \rangle$, тем самым образуется множество P_{k+1} таких σ_{k+1}. 5. Группировка *сравнимых* σ_{k+1}. 6. Отбор «лучших» σ_{k+1} согласно правилам доминирования, создание P_{k+1} всех «перспективных» σ_{k+1}. 7. $k + 1 \Rightarrow k$. 8. Проверка: $k > n$? 9. Решение закончено: $\overline{P}_n$ - решение.

Объясним некоторые блоки этой схемы.

В блоке *«группировка сравнимых* σ_k» собирают вместе все σ_k, приводящие в одно и то же состояние.

Нетрудно показать, что число таких различных подмножеств — группировок — будет равно числу сочетаний из n элементов по k — 2 элементов:

$$C_n^{k-2} = \frac{n!}{(n-k+2)!\,(k-2)!},$$

Так как у всех сравнимых вариантов σ_k фиксированы начальный и конечный элементы i_1 и i_k. (Напомним, что *сочетанием из n элементов по k* называется число всех различных подмножеств, которые состоят из k элементов, взятых из фиксированного множества, которое состоит из n элементов). В блоке *«отбора лучших* σ_k» выполняется перебор всех элементов, попавших в одну и ту же группировку сравнимых вариантов, и в множестве $\overline{P}_k$ оставляется по одному лучшему элементу из каждой такой группировки. Ветвление не попавших в $\overline{P}_k$ вариантов σ_k в дальнейшем запрещается. В блоке *«операция ветвления»* происходит последовательный просмотр всех σ_k k-го уровня, не запрещенных для ветвления, от каждого такого σ_k (из $\overline{P}_k$) выполняется построение всевозможных $\sigma_{k+1} = \langle \sigma_k, i \rangle$ — как и при построении порфириана. Процесс решения приведенного в табл. 5.15 примера по схеме, представленной на рис. 5.36, воспроизводится в табл. 5.16. В табл. 5.16, как и ранее, крестиком × отмечены «бесперспективные» согласно правилам доминирования варианты, т.е. отсеиваемые (запрещаемые для ветвления) блоком «отбора лучших вариантов». Плюс (+) отмечает «перспективные» варианты.

Таблица 5.16

Этапы решения примера (при равенстве $F(\sigma_k)$ лучшим считается первый вариант)

σ_2

σ_k	$F(\sigma_k)$
⟨1, 2⟩	7
⟨1, 3⟩	4
⟨1, 4⟩	7
⟨1, 5⟩	8
⟨1, 6⟩	2

Таблица 5.16 (продолжение)

σ_3

σ_k	$F(\sigma_k)$	σ_k	$F(\sigma_k)$
⟨1, 2, 3⟩	14	⟨1, 4, 5⟩	14
⟨1, 2, 4⟩	13	⟨1, 4, 6⟩	13
⟨1, 2, 5⟩	11	⟨1, 5, 2⟩	12
⟨1, 2, 6⟩	15	⟨1, 5, 3⟩	16
⟨1, 3, 2⟩	12	⟨1, 5, 4⟩	14
⟨1, 3, 4⟩	10	⟨1, 5, 6⟩	16
⟨1, 3, 5⟩	11	⟨1, 6, 2⟩	9
⟨1, 3, 6⟩	12	⟨1, 6, 3⟩	9
⟨1, 4, 2⟩	14	⟨1, 6, 4⟩	9
⟨1, 4, 3⟩	13	⟨1, 6, 5⟩	9

σ_4 (группирование и отбор лучших — отмечены знакомый +)

σ_k	$F(\sigma_k)$	?	σ_k	$F(\sigma_k)$	?	σ_k	$F(\sigma_k)$	?
⟨1, 2, 3, 4⟩ ⟨1, 3, 2, 4⟩	20 18	× +	⟨1, 2, 6, 4⟩ ⟨1, 6, 2, 4⟩	22 15	× +	⟨1, 3, 6, 5⟩ ⟨1, 6, 3, 5⟩	19 16	× +
⟨1, 2, 3, 5⟩ ⟨1, 3, 2, 5⟩	21 16	× +	⟨1, 2, 6, 5⟩ ⟨1, 6, 2, 5⟩	22 13	× +	⟨1, 4, 5, 2⟩ ⟨1, 5, 4, 2⟩	18 21	+ ×
⟨1, 2, 3, 6⟩ ⟨1, 3, 2, 6⟩	19 20	+ ×	⟨1, 3, 4, 2⟩ ⟨1, 4, 3, 2⟩	17 21	+ ×	⟨1, 4, 5, 3⟩ ⟨1, 5, 4, 3⟩	22 17	× +
⟨1, 2, 4, 3⟩ ⟨1, 4, 2, 3⟩	19 21	+ ×	⟨1, 3, 4, 5⟩ ⟨1, 4, 3, 5⟩	17 20	+ ×	⟨1, 4, 5, 6⟩ ⟨1, 5, 4, 6⟩	22 20	× +
⟨1, 2, 4, 5⟩ ⟨1, 4, 2, 5⟩	20 18	× +	⟨1, 3, 4, 6⟩ ⟨1, 4, 3, 6⟩	16 18	+ ×	⟨1, 4, 6, 2⟩ ⟨1, 6, 4, 2⟩	20 16	× +
⟨1, 2, 4, 6⟩ ⟨1, 4, 2, 6⟩	19 22	+ ×	⟨1, 3, 5, 2⟩ ⟨1, 5, 3, 2⟩	15 24	+ ×	⟨1, 4, 6, 3⟩ ⟨1, 6, 4, 3⟩	20 15	× +
⟨1, 2, 5, 3⟩ ⟨1, 5, 2, 3⟩	17 18	+ ×	⟨1, 3, 5, 4⟩ ⟨1, 5, 3, 4⟩	17 22	+ ×	⟨1, 4, 6, 5⟩ ⟨1, 6, 4, 5⟩	20 16	× +
⟨1, 2, 5, 4⟩ ⟨1, 5, 2, 4⟩	15 18	+ ×	⟨1, 3, 5, 6⟩ ⟨1, 5, 3, 6⟩	19 21	+ ×	⟨1, 5, 6, 2⟩ ⟨1, 6, 5, 2⟩	23 13	× +
⟨1, 2, 5, 6⟩ ⟨1, 5, 2, 6⟩	19 20	+ ×	⟨1, 3, 6, 2⟩ ⟨1, 6, 3, 2⟩	19 17	× +	⟨1, 5, 6, 3⟩ ⟨1, 6, 5, 3⟩	23 17	× +
⟨1, 2, 6, 3⟩ ⟨1, 6, 2, 3⟩	22 16	× +	⟨1, 3, 6, 4⟩ ⟨1, 6, 3, 4⟩	19 15	× +	⟨1, 5, 6, 4⟩ ⟨1, 6, 5, 4⟩	28 15	× +

Таблица 5.16 (продолжение)

σ_5

σ_k	$F(\sigma_k)$	?	σ_k	$F(\sigma_k)$	?	σ_k	$F(\sigma_k)$	?
⟨1, 3, 2, 4, 5⟩	25	×	⟨1, 4, 5, 2, 3⟩	25	×	⟨1, 3, 4, 5, 2⟩	21	+
⟨1, 2, 4, 3, 5⟩	26	×	⟨1, 4, 2, 5, 3⟩	26	×	⟨1, 3, 5, 4, 2⟩	24	×
⟨1, 3, 4, 2, 5⟩	21	+	⟨1, 2, 5, 4, 3⟩	23	+	⟨1, 5, 4, 3, 2⟩	24	×
⟨1, 3, 2, 4, 6⟩	24	+	⟨1, 4, 5, 2, 6⟩	26	×	⟨1, 3, 4, 5, 6⟩	25	×
⟨1, 2, 4, 3, 6⟩	24	×	⟨1, 4, 2, 5, 6⟩	26	×	⟨1, 3, 5, 4, 6⟩	23	×
⟨1, 3, 4, 2, 6⟩	25	×	⟨1, 2, 5, 4, 6⟩	23	+	⟨1, 5, 4, 3, 6⟩	22	+
⟨1, 3, 2, 5, 4⟩	22	×	⟨1, 2, 4, 6, 3⟩	26	×	⟨1, 3, 4, 6, 2⟩	23	×
⟨1, 2, 5, 3, 4⟩	25	×	⟨1, 6, 2, 4, 3⟩	21	+	⟨1, 6, 3, 4, 2⟩	22	+
⟨1, 3, 5, 2, 4⟩	21	+	⟨1, 6, 4, 2, 3⟩	23	×	⟨1, 6, 4, 3, 2⟩	23	×
⟨1, 3, 2, 5, 6⟩	24	×	⟨1, 2, 4, 6, 5⟩	26	×	⟨1, 3, 4, 6, 5⟩	23	×
⟨1, 2, 5, 3, 6⟩	24	×	⟨1, 6, 2, 4, 5⟩	22	×	⟨1, 6, 3, 4, 5⟩	22	+
⟨1, 3, 5, 2, 6⟩	23	+	⟨1, 6, 4, 2, 5⟩	20	+	⟨1, 6, 4, 3, 5⟩	22	×
⟨1, 2, 3, 6, 4⟩	26	×	⟨1, 2, 5, 6, 3⟩	26	×	⟨1, 3, 5, 6, 4⟩	26	×
⟨1, 6, 2, 3, 4⟩	22	+	⟨1, 6, 2, 5, 3⟩	21	×	⟨1, 6, 3, 5, 4⟩	22	+
⟨1, 6, 3, 2, 4⟩	23	×	⟨1, 6, 5, 2, 3⟩	20	+	⟨1, 6, 5, 3, 4⟩	25	×
⟨1, 2, 3, 6, 5⟩	26	×	⟨1, 2, 5, 6, 4⟩	26	×	⟨1, 6, 5, 4, 2⟩	22	×
⟨1, 6, 2, 3, 5⟩	23	×	⟨1, 6, 2, 5, 4⟩	19	×	⟨1, 5, 4, 6, 2⟩	27	×
⟨1, 6, 3, 2, 5⟩	21	+	⟨1, 6, 5, 2, 4⟩	19	+	⟨1, 6, 4, 5, 2⟩	20	+

Таблица 5.16 (продолжение)

σ_5

σ_k	$F(\sigma_k)$	?	σ_k	$F(\sigma_k)$	?
⟨1, 5, 4, 3, 6, 2⟩	29	×	⟨1, 3, 2, 4, 6, 5⟩	31	×
⟨1, 6, 3, 4, 5, 2⟩	26	+	⟨1, 6, 2, 3, 4, 5⟩	29	×
⟨1, 6, 3, 5, 4, 2⟩	29	×	⟨1, 6, 2, 4, 3, 5⟩	28	×
⟨1, 6, 4, 5, 3, 2⟩	29	×	⟨1, 6, 3, 4, 2, 5⟩	26	+
⟨1, 2, 5, 4, 6, 3⟩	30	×	⟨1, 3, 4, 2, 5, 6⟩	29	×
⟨1, 6, 4, 2, 5, 3⟩	28	×	⟨1, 3, 5, 2, 4, 6⟩	27	+
⟨1, 6, 5, 2, 4, 3⟩	25	+	⟨1, 2, 5, 4, 3, 6⟩	28	×
⟨1, 6, 4, 5, 2, 3⟩	27	×	⟨1, 3, 4, 5, 2, 6⟩	29	×
⟨1, 3, 5, 2, 6, 4⟩	30	×			
⟨1, 6, 3, 2, 5, 4⟩	27	×			
⟨1, 6, 5, 2, 3, 4⟩	26	+			
⟨1, 6, 3, 5, 2, 4⟩	26	×			

$\langle \sigma_5, 1 \rangle$

σ_k	$F(\sigma_k)$	?
⟨1, 6, 3, 4, 5, 2, 1⟩	32	×
⟨1, 6, 5, 2, 4, 3, 1⟩	30	×
⟨1, 6, 5, 2, 3, 4, 1⟩	34	×
⟨1, 6, 3, 4, 2, 5, 1⟩	33	×
⟨1, 3, 5, 2, 4, 6, 1⟩	29	+

3. Поиск «критического пути». Определение «критического пути» в сетевом графике п. 5.13 можно также провести с помощью построения порфириана.

По-прежнему путь в сетевом графике есть произвольная последовательность работ

$$\sigma_k = \langle i_1, i_2, \ldots, i_k \rangle$$

если только любые две соседние работы в σ_k соединены на сетевом графике стрелками.

Для наглядности (образности) можно считать, что работы i_k в сетевом графике есть некоторые «города», и интерпретировать задачу

поиску критического пути на сетевом графике как требование определить самый длинный путь, который может пройти турист, между исходным пунктом (начальной работой) и конечным (заключительной работой).

Будем обозначать через $F(\sigma_k)$ общую длину пути σ_k, т.е.

$$\sum_i \tau_i ,$$

если путь проходит через i (в сетевом графике τ_i — продолжительность работы i, в нашей интерпретации — время пребывания туриста в пункте i); здесь перемещениями между пунктами мы пренебрегаем (турист считает только то время, в течение которого он знакомится с достопримечательностями).

Два пути σ_k и σ_l, которые приводят в один и тот же пункт (в наших обозначениях $i_k(\sigma_k) = i_l(\sigma_l)$), *сравнимы.* Правило доминирования естественно вытекает из того, что можно говорить о критическом пути для каждого пункта, так как можно построить для каждого пункта самый длинный путь, который соединяет его с начальным пунктом. Заметим, что *каждая часть критического пути есть критический путь* в том смысле, что лежащие на критическом пути пункты i и j также соединены «наиболее длинно» и не существует другого пути, который соединяет эти пункты «более длинным образом».

Во многих задачах поска «наилучших» траекторий путь обладает *свойством оптимальности* (удовлетворяет принципу оптимальности). В таких случаях правила доминирования находятся легко и естественно

Утверждение (правило доминирования). *Из двух сравнимых путей* σ_k и σ_l *«лучше» тот, у которого больше длина пути.*

Это утверждение следует из предыдущего замечания.

В задачах подобного типа — поиска некоторого одного пути на графе обычно обходятся без построения порфириана. Такая возможность обеспечивается приемом расстановки *меток* на графе, т.е. прокладыванием путей σ_k на самом сетевом графике (маршруте туриста) — например, красным карандашом сверху по плану, вычерченному предварительно черной тушью.

Соединим предварительно с исходным пунктом те вершины сетевого графика, в которые приводят пути от исходного пункта и только от него. *Отметим* эти пункты j, т.е. наведем дополнительно на графике пути, которые соединяют их с исходным пунктом, и проставим возле них *метку* $r(j)$ — пройденное расстояние, длину пути.

Пусть теперь отмечены какие-то пункты. Найдем пункт *i*, в который приводят стрелки только из отмеченных метками *r(j)* пунктов *j*. Построим теперь

$$r(j) + \tau(i) \qquad (5.80)$$

и выберем среди них максимальное. Тот пункт *j*, который дает максимальное (5.80), соединим с пунктом *i*, т.е. наведем стрелку, соединяющую *j* с *i*, и отметим *i* величиной (5.80).

Решение закончится, когда конечный пункт станет отмеченным. Рис. 5.37 приводит только что описанный алгоритм.

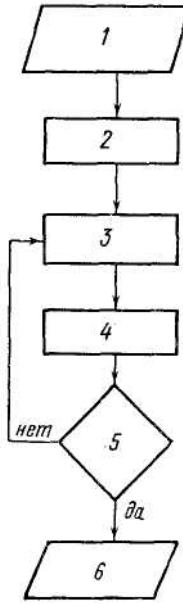


Рис. 5.37. Схема поиска критического пути методом расстановки меток.

1. Подготовка информации о графе и значениях *r(i)*. 2. «Отмечаем» исходную вершину: полагаем для нее *r*(*j*)=0, ставим «метку» *i* =0. 3. Ищем вершину *j*, в которую приходят стрелки только от отмеченных вершин. 4. Вычисляем для этой вершины $r(i) + \tau_i$, находим i_0, для которого эта сумма максимальна, отмечаем вершину *j*: полагаем для нее $r(j)= r(i_0)+ \tau_{ij}$ и ставим «метку» i_0. 5. Проверяем: конечная вершина отмечена? 6. Решение найдено, остается по меткам восстановить критический путь.

5.19. Отсеивание вариантов. Ветви и границы

1. Схема «ветвей и границ». Представьте, что каким-то образом нам удается оценить для каждого σ_k, что среди продолжений σ_k можно получить в лучшем случае σ^2_n с некоторым $F(\sigma^2_n)$. Такую оценку $F(\sigma^2_n)$ мы будем называть *«обещанием»* и обозначать через $b(\sigma_k)$.

Оставим пока в стороне вопрос, как получено такое «обещание» (насколько можно ему верить).

Рассмотрим следующую идею отсеивания вариантов в процессе построения порфириана.

Будем подвергать ветвлению σ_k с наиболее удобным (т.е. наибольшим в случае задачи максимизации, наименьшим при минимизации) «обещанием».

Это правило ветвления будем использовать до тех пор, пока не получим некоторое (окончательное) σ^0_n. Вычислим $F(\sigma^0_n)$. Теперь можно отсеять (запретить ветвление) все те σ_k, для которых $b(\sigma_k) \geq F(\sigma^0_n)$ — в задаче минимизации (или $b(\sigma_k) \leq F(\sigma^0_n)$ в задаче минимизации).

Поиск σ_n (улучшающих σ^0_n) по описанной схеме продолжается до тех пор, пока не останется ни одного σ_k, который бы «предлагал лучшее» (чем у уже найденных σ_n) «обещание».

Таким образом, в схеме «ветвей и границ» (после умения строить порфириан) главное — это умение оценивать $b(\sigma_k)$ для каждого σ_k.

Очевидно, при этом для каждого σ_k необходимо сформулировать некоторую вспомогательную задачу (вычисление $b(\sigma_k)$) и, конечно, более простую, чем исходная (поиск наилучшего σ_n среди продолжений σ_k).

Чаще всего вспомогательная задача формулируется как некоторое (естественное) упрощение исходной задачи.

Примером продуктивности этой рекомендации может служить следующая задача.

2. Задача о рюкзаке. Собираясь в поход, вы хотели бы взять n предметов, каждый предмет i весом a_i и «ценностью» в походе c_i.

Вы можете взять предметов с собой по весу не больше A (предполагается, что $\sum_i a_i > A$— иначе не было бы задачи!). Какие предметы взять?

Решение задачи можно представить последовательностью

$$\Sigma_n = \langle x_1, x_2, \ldots, x_n \rangle ,$$

где

$$x_i = \begin{cases} 1, \textit{если мы берем с собой предмет } i, \\ 0, \textit{если не берем.} \end{cases}$$

Таким образом, нам надо найти такую последовательность σ_n, что на ней

$$\sum x_i a_i \leqslant A$$

и достигается максимум $\sum_i x_i c_i$.

Задача о рюкзаке, как задача теории расписаний, возникает, когда нужно определить перечень операций, которые следует включить в план выполнения в некоторый временно́й интервал (пятидневку, неделю) .

Построение вспомогательной задачи в этом случае осуществляется довольно просто и поучительно.

В исходной задаче о рюкзаке x_i может принимать значение только 0 и 1. Отказ от этого предположения и приводит к вспомогательной, легко решаемой задаче:

$$0 \leq x_i \leq 1.$$

Физически это означает, что грузы наши «сыпучие» и мы можем взять любую часть груза.

Решение задачи в случае сыпучих грузов очевидно.

Найдем величину $\frac{c_i}{a_i}$, и пусть грузы благоустроены по убыванию $\frac{c_i}{a_i}$, т.е. по величине «ценности» единицы веса груза,

$$\frac{c_1}{a_1} \geqslant \frac{c_2}{a_2} \geqslant \ldots \geqslant \frac{c_n}{a_n}.$$

Тогда необходимо брать груз, начиная с первого, до тех пор, пока мы не «засипем» рюкзак до веса *A*. Другими словами:

а) если $a_1 \leq A$, то $x_1 = 1$;

б) если

$$\sum_{i=1}^{k} a_i \leqslant A,$$

то $x_k = 1$;

в) если

$$\sum_{i=1}^{k} a_i < A, \quad \sum_{i=1}^{k+1} a_i > A,$$

то

$$x_{k+1} = \frac{A - \sum_{i=1}^{k} a_i}{a_{k+1}}$$

и $x_i = 0$, $i > k+1$.
Решение задачи

$$C(\sigma_n) = \sum_{i=1}^{n} c_i x_i$$

в случае «сыпучих» грузов и может служить искомой оценкой — «обещанием» — значений $F(\sigma_n)$, ибо , как нетрудно видеть,

$$F(\sigma_n) \le C(\sigma_n).$$

Построение порфириана достигается очевидным способом. Множество всех возможных σ_n разбивается на два: первое с $x_1 = 0$, вторая с $x_1 = 1$. Множества второго уровня σ_2 есть разбиения полученных множеств в зависимости от того, $x_2 = 0$ или $x_2 = 1$ и т.п.
На рис. 5.38 представленная общая схема решения задачи о рюкзаке.

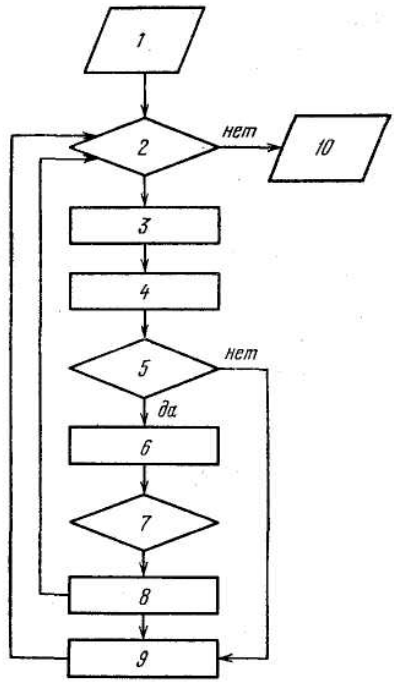


Рис. 5.38. Схема решения задачи о рюкзаке.

Множество возможных решений обозначим через σ_0. Положим рекорд f^* равным нулю. (Смысл рекорда f устанавливается в блоке 8.) Формулируем вспомогательную задачу ($0 \le \chi_i \le 1$), находим $\sigma_0 \Delta(\sigma_0)$. 2. Проверяем, есть ли в построенном фрагменте «висячие> вершины $\sigma_k \Delta(\sigma_k)$. («Висячие» - те, в которых нет еще «продолжений», они же - мажоранты порфириана.) 3. Операция выбора: выбираем «висячую» вершину σ_k с $\Delta(\sigma_k) > f$ и при этом с наибольшим $\Delta(\sigma_k)$. Можно выбирать σ_k с $\Delta(\sigma_k) > f$ (при этом с наибольшим k) - это определит более эффективный «челночный» вариант схемы решения. 4. Операция ветвления: добавляем к порфириану две вершины: $\sigma'_{k+1} = \langle \sigma_k, 0 \rangle$ и $\sigma''_{k+1} = \langle \sigma_k, 0 \rangle$, вычеркиваем σ''_{k+1}, если ею определяется решение, недопустимое неравенством $\sum a_i \gamma_i < A$ (нарушает это условие). 5. Проверяем, σ'_{k+1} и σ''_{k+1} – не являются ли они полными решениями, т.е. $k + 1 = n$? 6. Вычисляем для σ'_{k+1} и σ''_{k+1} величину $k + 1 = n$. 7. Проверяем, найдется ли $f(\sigma_{k+1}) > f$. 8. Полагаем новое f равным максимальному из таких $f(\sigma_n)$ — это и есть рекорд. Фиксируем σ_n - *рекордный* план. 9. Для найденных σ'_{k+1} и σ''_{k+1} (если они допустимы) формулируем вспомогательную задачу ($0 \le \chi_i \le 1$, с учетом уже найденных χ_i, $i \le k+1$, уже определенных σ_{k+1}), находим $\Delta(\sigma_{k+1})$. 10. Решение закончено: рекордный план и есть оптимальный вариант загрузки рюкзака, а рекорд f - значение функции-критерия для этого плана.

Блок *«формулирование вспомогательной задачи для* σ_k*»* означает следующее.

Обозначим для σ_k

$$A(\sigma_k) = \sum_{l=1}^{k} a_{t_l} x_{t_l}, \qquad C(\sigma_k) = \sum_{l=1}^{k} c_{t_l} x_{t_l}.$$

Вспомогательная задача для σ_k . Найти $x_i (i > k)$ такие, что $0 \le x_i \le 1$,

$$\sum_{i=k+1}^{n} a_i x_i \leqslant A - A(\sigma_k),$$

и при этом

$$C(x) = \sum_{i=k+1}^{n} c x_i + C(\sigma_k) \tag{5.81}$$

достигает максимума.

Рис. 5.39 представляет конечный результат построения порфириана по схеме решения задачи о рюкзаке (см. рис. 5.38) и данных, представленных табл. 5.17.

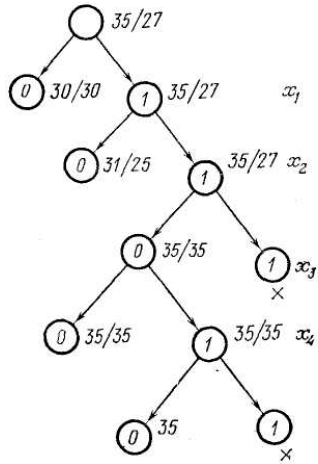


Рис. 5.39. К решению задачи о рюкзаке.

Возле вершины записаны $c(x)/c'(x)$, где $c'(x)$ вычисляется по формуле (5.81), но x_i выбираются из x только целые.

Таблица 5.17

К задаче о рюкзаке

i	1	2	3	4	5
a_i	5	4	5	4	4
c_i	15	12	10	8	8
$\frac{c_i}{a_i}$	3	3	2	2	2

3. Решение задачи бродячего торговца методом ветвей и границ. Отметим: вспомогательная задача в методе ветвей и границ формулируется путем отказа от некоторых ограничений или послаблением каких-то ограничений (в задаче о рюкзаке — отказом от дискретности x_i).

В задаче бродячего торговца вспомогательная задача может быть сформулирована так.

Рассмотрим несколько другое, чем ранее, представление решения задачи бродячего торговца.

Пусть

$$x_{ij} = \begin{cases} 1, \textit{ если с пункта } i \textit{ торговец} \\ \textit{направляется в пункт } j, \\ 0, \textit{ если нет.} \end{cases}$$

Такое представление упрощает представление способа вычисления значения функции-критерия для решения

$$F(\sigma) = \sum_{i} x_{ij} a_{ij}, \tag{5.82}$$

где a_{ij} — расстояние между пунктами i и j (заданное матрицей — таблицей расстояний).

Это же представление усложняет формулирование условия, что торговец не должен побывать в одном и том же городе дважды, за исключением исходного.

Вот от этого-это ограничения мы теперь и откажемся.

Нетрудно видеть, что в оговоренном случае задача превращается в задачу о назначении, где нужно найти максимум (4.110) при условиях, что для любого i найдется такое j, что x_i=1; для каждого x_{ij} найдется такое i, что x_{ij} = 1. Это можно записать и иначе:

$$\sum_{j=1}^{n} x_{ij} = \sum_{i=1}^{n} x_{ij} = 1 \tag{5.83}$$

для любых i и j.

Задача о назначении дает соединения городов (в случае задачи бродячего торговца), вообще говоря, отдельными несвязанными круговыми маршрутами.

Таким образом, формальное решение задачи бродячего торговца по схеме «ветвей и границ» можно свести к решению задачи о назначении.

Так что, прежде всего, желательно научиться решать задачу о назначениях.

Задача о назначениях в свою очередь также может быть решена по схеме «ветвей и границ».

Для этого выполним сначала следующие преобразования.

Утверждение 1. *Решение задачи о назначениях не изменится, если каждое a_{ij} в i-й строке уменьшить на то же число Δ_i.*

Утверждение 2. *Решение задачи о назначениях не изменится, если каждое a_{ij} в j-м столбце уменьшить на то же число δ_j.*

Вычтем теперь из матрицы a_{ij} величины Δ_i и δ_j так, чтобы в каждой строке и каждом столбце образовались так называемые *нулевые элементы* или просто «нули».

Для этого возьмем *i*-ю строку, найдем в ней максимальное a_{ij} и положим Δ_i - равным этому максимальному a_{ij}. Затем во вновь образованной матрице $\overline{a}_{ij} = a_{ij} - \Delta_i$ возьмем *j*-й столбец и положим δ_j - равным максимальному элементу $\overline{a}_{ij}$ в столбце. Перейдем теперь к матрице $\overline{\overline{a}}_{ij} = a_{ij} - \Delta_i - \delta_j$. В ней в каждых строке и столбце теперь есть $\overline{\overline{a}}_{ij} = 0$ — это и есть *«нулевые элементы»,* а все $a_{ij} \leq 0$. (Мы рассматривали задачу о назначении как задачу максимизации, так ее и решаем, хотя, как вспомогательная для задачи бродячего торговца, задача о назначении есть задача минимизации - там выбирается минимальное a_{ij}).

Легко видеть, что, например, табл. 5.2 преобразуется к такому виду (табл. 5.18) Δ_1=5, $\Delta_2 = \Delta_3$= 4, $\delta_1 = \delta_3 = 0$, $\delta_2 = -1$.

Таблица 5.18

	Оля	Зоя	Лиля
Саша	0	0	−2
Андрей	−2	−1	0
Сережа	0	−2	−1

Будем строить теперь порфириан для задачи о назначениях (рис. 5.40 воспроизводит решение для условий, заданных табл. 5.18).

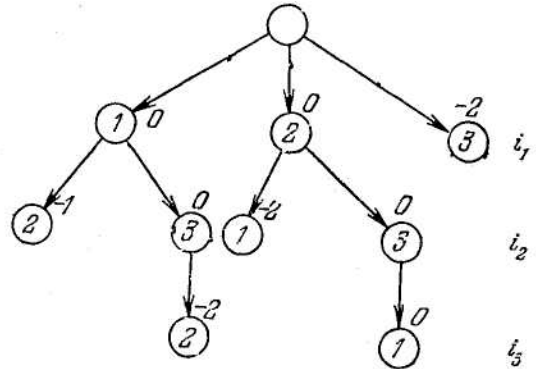


Рис. 5.40. Решение задачи о назначении методом «ветвей и границ»

Множество возможных решений сначала разобьем (в зависимости от того, чему равно i_1). Возле каждой вершины i_1 запишем соответствующее $b(i_1)= \overline{\overline{a}}_{i1}$ ($u = i_1$). Ветвление будем производить для той вершины i_k, для которой $b\ (i_k)$ максимально.

Вообще, пусть полученыа вершины, соответствующие σ_k. Возможны три случая:

а) среди них нет ни одного σ_n;

б) получены σ_n, но найдется некоторый σ_k $(k < n)$, что $b(\sigma_k) > F(\sigma_n)$ для всех σ_n;

в) нет такого σ_k.

В случае в) решение закончено и максимальное $F(\sigma_n)$ есть экстремум, в случаях а) и б) разветвлению подвергается элемент с максимальным $b(\sigma_k)$, $b(\sigma_{k+1})= b(\sigma_k) + a_{i_{k+1}k+1}$. (Если он не один, то самый «нижний» и самый левый среди «нижних».)

Упражнение. Постройте блок-схему описанного алгоритма.

Нетрудно понять, что решение задачи бродячего торговца может быть выполнено по совершенно аналогической схемой с той только разницей, что в этом случае будет просто другое построение порфириана (со слежением за тем, чтобы не получилось замкнутых путей) - все остальные процедуры алгоритма решения задачи о назначениях могут быть перенесены в алгоритм решения задачи бродячего торговца без изменения.

Иными словами, не надо каждый раз решать вспомогательную задачу о назначениях, необходимо просто решать и в задаче бродячего торговца ту же вспомогательную задачу, что и в задаче о назначениях.

Нетрудно показать справедливость утверждений 1 и 2 и для задачи бродячего торговца. Табл. 5.19 представляет матрицу $\overline{\overline{a}}_{ij}$ табл. 5.15 после переобразований вида $\overline{\overline{a}}_{ij} = a_{ij} — \Delta_i — \delta_j$, где $\Delta_1 = \Delta_6 = 2$, $\Delta_2 = \Delta_5 = 4$, $\Delta_3 = 5$, $\Delta_4 = 6$, $\delta_1 = \delta_2 = \delta_3 = \delta_5 = \delta_6 = 0$, $\delta_4 = 1$.

Таблица 5.19

Матрица a_{ij}

i \ j	1	2	3	4	5	6
1		5	2	4	6	0
2	2		3	1	0	4
3	0	3		0	2	0
4	2	1	0		1	0
5	3	0	4	1		4
6	0	5	5	4	5	

Будем строить решение задачи бродячего торговца как порфириан, производя ветвления и отсеивание вариантов по следующим правилам. По-прежнему считаем, что путешествие начинается с пункта 1, так что все множество возможных вариантов отождествляется с множеством, где $\sigma_1 = \langle 1 \rangle$; примем, что $b(\sigma_1) = 0$.

Пусть построена какая-то часть порфириана, все *конечные* вершины которого, как известно, можно отождествить с некоторыми маршрутами $\sigma_k = \langle i_1, i_2, \ldots, i_k \rangle$ — множество этих маршрутов обозначим через *P*.

Ветвление некоторого σ_k означает построение $\sigma_{k+1}=(\sigma_k, j)$, где j не входит в $M(\sigma_k)$, и соответствующих вершин в порфириане. При этом

$$b(\sigma_{k+1}) = b(\sigma_k) + \overline{\overline{a}}_{i_k j}.$$

Выбор σ_k для ветвления или остановов решения определяется следующим. Если в *P* содержатся σ_n, вычислим для них

$$\varphi(\sigma_n) = b(\sigma_n) + \overline{\overline{a}}_{i_n 1}.$$

Нетрудно понять, что вследствие выполненного преобразования матрицы a_{ij} в матрицу $\overline{\overline{a}}_{ij}$, длина $F(\sigma_n)$ полного пути $\langle 1, i_1, \ldots, i_n, 1 \rangle$, т.е. $(\sigma_n, 1)$ будет равна

$$F(\sigma_n) = \varphi(\sigma_n) + \varepsilon,$$

где

$$\varepsilon = \sum_i \Delta_i + \sum_j \delta_j.$$

Назовем наименьшее в нашем случае $\varphi(\sigma_n)$ *рекордом* на *P,* а соответствующее σ_n — рекордным. Если рекорд $\varphi(\sigma_n)$ не больше всех $b(\sigma_k)$ на *P,* то, как легко видеть, соответствующее рекордное σ_n представляет оптимальное решение задачи бродячего торговца.

Если это не так, ветвлению подлежит σ_k с минимальным $b(\sigma_k)$, а если таких σ_k несколько, то крайнее левое из них в построенной части порфириана.

Рис. 5.41 иллюстрирует решение задачи бродячего торговца, представленной табл. 5.15, по схеме ветвей и границ.

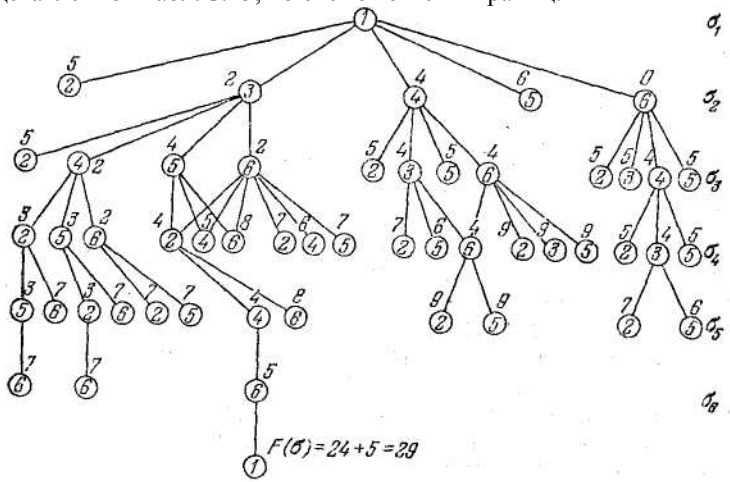


Рис. 5.41. Решение задачи бродячего торговца методом «ветвей и границ».

3. Обобщенная задача о рюкзаке. Мы будем называть *обобщенной задачей о рюкзаке* следующую постановку.

Максимизировать

$$\sum_i c_i x_i$$

при условиях

$$\sum a_{ij} x_i \le A_j, \quad j=1, \ldots, m, \qquad (5.84)$$

$$x_i = 0 \quad \text{или} \quad 1. \qquad (5.85)$$

Отличие ее от «обычной» задачи о рюкзаке в том, что решение должно удовлетворять больше чем одному ограничению (не одному, а m — (5.84)). Как и обычная задача о рюкзаке, обобщенная также может быть интерпретирована как задача теории расписаний (см. п. 2).

Построение порфириана в этой задаче выполняется, как и в задаче о рюкзаке. Весь вопрос применения схемы «ветвей и границ» — в вычислении «обещаний» на каждом фрагменте решения σ_k.

Заметим, что если, как и в задаче о рюкзаке (п. 2), отказаться от дискретности x_i, т.е. считать, что «грузы сыпучие», $0 \le x_i \le 1$, то мы придем к формулировке вспомогательной задачи как задачи линейного программирования, которую можно решить стандартными методами. Однако можно еще больше упростить решение. Вспомните рекомендации из формулированию вспомогательной задачи (п. 1). Самое грубое «естественное» упрощение — это отбросить m—1 неравенство и свести задачу к обычной задаче о рюкзаке. Но какое неравенство оставить? Первое? А если второе? А что будет, если мы решим m задач о рюкзаке, по одной для каждого неравенства? Наверное, мы получим m «обещаний» по числу ограничений (5.84), каждое из которых как бы отражает значение «своего» неравенства при построении решения.

Рассмотрим пример. Максимизировать

$$\sum c_i x_i$$

при условиях

$$\sum a_i x_i \le A, \quad \sum b_i x_i \le B, \quad x_i = 0 \text{ или } 1.$$

Значение c_i, a_i, b_i, A, B представлены табл. 5.20.

Таблица 5.20

t	1	2	3	4	5	
c_t	15	12	12	15	10	
a_t	3	3	4	5	5	$A=13$
c_t/a_t	5	4	3	3	2	
b_t	5	4	3	3	2	$B=11$
c_t/b_t	3	3	4	5	5	

Начнем, как обычно, построение порфириана с начальной вершины, отождествляемой с множеством всех возможных вариантов. Найдем прогноз значения функции-критерия обещания, в наших определениях (п. 1), сформулировав вспомогательную задачу как задачу о рюкзаке и выбрав первое ограничение:

Максимизировать

$$C_1(x) = \sum c_i x_i$$

при условиях

$$\sum a_i x_i \le A,\ 0 \le x_i \le 1.$$

С помощью табл. 5.20 легко найти это решение: $x_1 = x_2 = x_3 = 1$, $x_4 = 3/5$, $x_5 = 0$, $C_1(x) = 48$.

Сформулируем другу вспомогательную задачу, выбрав второе ограничение:

Максимизировать

$$C_2(x) = \sum c_i x_i$$

при условиях

$$\sum b_i x_i \le B,\ 0 \le x_i \le 1.$$

Легко получить, что $C_2(x) = 46$, $x_1 = 0$, $x_2 = 3/4$, $x_3 = x_4 = x_5 = 1$.

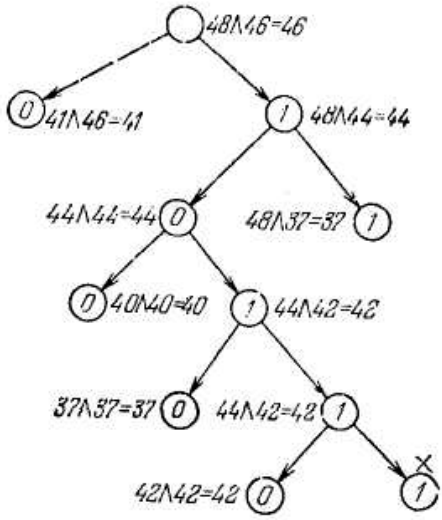


Рис. 5.42. Решение обобщенной задачи о рюкзаке по схеме «ветвей и границ» (крестиком отмечен недопустимый ограничениями вариант).

Ясно, что мы должны верить меньшему обещанию, так как легко доказать, что для всех σ_n одновременно

$$F(\sigma_n) \le C_1(x) \text{ и } F(\sigma_n) \le C_2(x).$$

Рис. 5.42 иллюстрирует ход решения в рассмотренном примере по схеме «ветвей и границ». Возле каждой вершины выписан символ вида

$x \wedge y = z(z(\sigma_k))$, где x — обещание по первому ограничению, y — по второму, z (или $z(\sigma_k)$) — наименьшее из x и y. Ветвлению подвергается σ_k с наибольшим $z(\sigma_k)$.

5.20. Задача трех станков

1. Постановка задачи и свойства оптимального решения.

Задача трех станков формулируется аналогично залаче двух станков (п. 5.17, п. 3). Дано n деталей, каждая из которых обрабатывается последовательно на трех станках. Продолжительность первой операции на первом станке для i-й детали обозначим через a_i, второй операции на втором станке b_i, третьей на третьем станке c_i. Как и прежде, предполагаем, что операция не может начаться, пока не закончилась предыдущая.

Задача трех станков также сводится к поиску экстремальной перестановки. Это доказывается аналогично тому, как и в задаче двух станков - сначала для первых двух станков, потом для последних двух.

Таким образом, каждое расписание на трех станках может быть представлено как n-перестановка

$$\sigma_n = \langle i_1, i_2, \ldots, i_k, \ldots, i_n \rangle \ldots$$

Обозначим через $A(\sigma_k)$ время окончания обработки k первых операций в σ_n на первом станке, через $B(\sigma_k)$ - время окончания обработки первых k операций в σ_n на втором станке, через $C(\sigma_k)$ - время окончания обработки первых k операций в σ_n на третьем станке.

Тогда по условиям задачи (считая, что в момент $t = 0$ все станка готовые к выполнению работ)

$$\begin{aligned} A(\sigma_1) &= a_{i_1}, \\ B(\sigma_1) &= a_{i_1} + b_{i_1} = A(\sigma_1) + b_{i_1}, \\ C(\sigma_1) &= a_{i_1} + b_{i_1} + c_{i_1} = B(\sigma_1) + c_{i_1}, \end{aligned} \qquad (5.86)$$

и, дальше,

$$A(\sigma_k) = A(\sigma_{k-1}) + a_{i_k},$$
$$B(\sigma_k) = \max(A(\sigma_k), B(\sigma_{k-1})) + b_{i_k}, \qquad (5.87)$$
$$C(\sigma_{k\,t}) = \max(B(\sigma_k), C(\sigma_{k-1})) + c_{i_k}.$$

Утверждение 1 (правила доминирования). *Фрагменты решений* $\sigma_k^{\,1}$ *и* $\sigma_k^{\,2}$ *сравнимые, если, как и в (*п. 5.18, п. 1),

$$M(\sigma_k^{\,1}) = M(\sigma_k^{\,2}).$$

Вариант $\sigma_k^{\,1}$ *лучше варианта* $\sigma_k^{\,2}$, *если одновременно*

$$B(\sigma_k^{\,1}) \le B(\sigma_k^{\,2}), \quad C(\sigma_k^{\,1}) \le C(\sigma_k^{\,2}). \qquad (5.88)$$

Обратите внимание: если раньше в задачах мы применяли, пользуясь правилами доминирования, при проверке только одну неровность (это довольно сильно суживало множество возможных продолжений при разветвлении), то здесь приходится проверять две неровности, и может случиться так, что одновременно они не имеют места.

В частности, если

$$a_i = 4, \quad b_i = 1, \quad c_1 = 3, \quad a_i = 3, \quad b_i = 2, \quad c_i = 1,$$

это в $\sigma_2^{\,1} = \langle i, j \rangle$

$$B(\sigma_{2)}) = 9, \quad C(\sigma_{2)}) = 10,$$

а в $\sigma_2^{\,2} = \langle j, i \rangle$

$$B(\sigma_{2)}) = 8, \quad C(\sigma_{2)}) = 11.$$

Заметим, что относительно пары *(i, j)* не очень-это просто ответить на вопрос, *i* передует *j* в оптимальном решении, или *j* передует *i*, даже если

$$B(\sigma_2^{\,1}) \le B(\sigma_2^{\,2}), \quad C(\sigma_2^{\,1}) \le C(\sigma_2^{\,2})$$

— довольно нетривиальная здесь зависимость от $B(\sigma_k)$ и $C(\sigma_k)$; довольно внимательно следует проводить изложение, если воспользоваться перестановочним приемом.

В некоторых простых случаях ответ на эти вопросы удается получить довольно очевидным образом.

Так, пусть детали перенумерованы таким образом, что оказалось

$$a_1 \le a_2 \le \ldots \le a_n,$$
$$b_1 \le b_2 \le \ldots \le b_n, \qquad (5.89)$$
$$c_1 \le c_2 \le \ldots \le c_n,$$

и, кроме того,

$$a_i \le b_i \le c_i \quad \text{для всех } 1 \le i \le n. \tag{5.90}$$

Утверждение 2. *При условиях* (5.89) *и* (5.90)

$$\sigma_n = \langle 1, 2, ..., n \rangle$$

есть оптимальная очередность обработки деталей.

Доказательство этого утверждения очевидно: во-первых, согласно (5.89) и (п. 5.17, п. 3) обработка деталей на двух станках — втором и третьем — оптимальна, и, кроме того, $B(\sigma_0) = a_i$ для такой очередности минимальна (a_1) из всех возможных.

Рассмотрим другой случай.

Пусть

$$\min_j a_j \ge \max_i b_i . \tag{5.91}$$

Тогда, применяя перестановочный прием, несложно показать, что i предшествует j в оптимальном решении как только

$$\min (a_i + b_i, b_j + c_j) \le \min (a_j + b_j, b_i + c_i). \tag{5.92}$$

Задача, таким образом, решается аналогично задаче двух станков с помощью решающего правила.

2. Решение методом последовательного отсеивания вариантов. Рассмотрим пример, заданный табл. 5.21.

Таблица 5.21

i	1	2	3	4	5	6
a_i	1	3	2	2	1	3
b_i	2	2	2	4	5	2
c_i	3	3	2	2	3	1

Воспользуемся утверждением 1 (правилом доминирования) и построим блок-схему метода отсеивания вариантов для задачи трех станков (рис. 5.43).

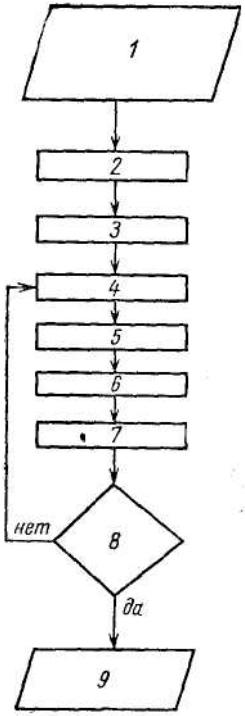


Рис. 5.43. Схема решения задачи трех станков последовательным отсеиванием вариантов.

1. Подготовка таблицы a_i, b_i, c_i к решению. 2. Построение $\sigma_1 = \langle i_1 \rangle$, вычисление $A(\sigma_1)$, $B(\sigma_1)$, $C(\sigma_1)$. Образуем P_1 — множество (список) всех σ_1. 3. k=l. 4. *Операция ветвления:* каждая σ_k из P_k порождает $\sigma_{k+1} = \langle \sigma_k, j \rangle$, где j не входит в $M(\sigma_k)$. 5. Группирование сравнимых σ_{k+1}, вычисление $B(\sigma_{k+1})$ и $C(\sigma_{k+1})$. 6. Отбор «лучших» σ_{k+1} согласно правилам доминирования, образование P_{k+1} из «лучших» σ_{k+1}. 7. $k+1 = k$. 8. $k = n$? 9. P_n - решение.

Следующий графический прием упрощает вычисления.
Пусть мы построили

$$\sigma_3 = \langle 1, 2, 3 \rangle$$

Поставим в соответствие σ_3 ее сетевой график и проведем вычисление $B(\sigma_k)$ и $C(\sigma_k)$ как критических путей в этом сетевом графике. Рис. 5.44 подробно иллюстрирует способ вычислений,

впрочем совсем идентичный формулам (5.87). Решение примера представлено таблицей 5.22.

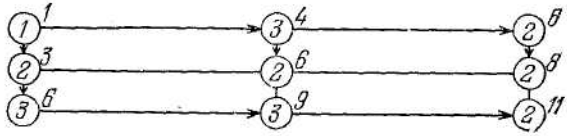


Рис. 5.44. К вычислению $A(\sigma_k)$, $B(\sigma_k)$, $C(\sigma_k)$ в задаче трех станков.

Таблица 5.22

Решение задачи трех станков последовательным отсеиванием вариантов

σ_1

σ_1	ABC	σ_1	ABC	σ_1	ABC
⟨1⟩ ⟨2⟩	1, 3, 6 3, 5, 8	⟨3⟩ ⟨4⟩	2, 4, 6 2, 6, 8	⟨5⟩ ⟨6⟩	1, 6, 9 3, 5, 2

σ_2

σ_2	ABC	?	σ_2	ABC	?	σ_2	ABC	?
⟨1, 2⟩ ⟨2, 1⟩	4, 6, 9 4, 7, 11	+ −	⟨2, 3⟩ ⟨3, 2⟩	5, 7, 10 5, 7, 10	+ −	⟨3, 6⟩ ⟨6, 3⟩	5, 7, 8 5, 7, 9	+ −
⟨1, 3⟩ ⟨3, 1⟩	3, 5, 8 3, 6, 9	+ −	⟨2, 4⟩ ⟨4, 2⟩	5, 9, 11 5, 8, 11	− +	⟨4, 5⟩ ⟨5, 4⟩	3, 11, 14 3, 10, 12	− +
⟨1, 4⟩ ⟨4, 1⟩	3, 7, 9 3, 8, 11	+ −	⟨2, 5⟩ ⟨5, 2⟩	4, 10, 13 4, 8, 12	− +	⟨4, 6⟩ ⟨6, 4⟩	5, 8, 9 5, 9, 11	+ −
⟨1, 5⟩ ⟨5, 1⟩	2, 8, 11 2, 8, 12	+ −	⟨2, 6⟩ ⟨6, 2⟩	6, 8, 9 6, 8, 11	+ −	⟨5, 6⟩ ⟨6, 5⟩	4, 8, 9 4, 10, 13	+ −
⟨1, 6⟩ ⟨6, 1⟩	4, 8, 10 4, 8, 10	− +	⟨3, 4⟩ ⟨4, 3⟩	3, 9, 12 3, 8, 11	+ +	⟨3, 5⟩ ⟨5, 3⟩	4, 6, 7 3, 8, 11	+ −

Таблица 5.22 (продолжение)

σ_3

σ_3	ABC	?	σ_3	ABC	?	σ_3	ABC	?
⟨1, 2, 3⟩	6, 8, 11	+	⟨1, 2, 6⟩	7, 9, 10	+	⟨1, 3, 6⟩	6, 8, 9	+
⟨1, 3, 2⟩	6, 8, 11	−	⟨1, 6, 2⟩	7, 9, 12	−	⟨1, 6, 3⟩	6, 8, 10	−
⟨2, 3, 1⟩	6, 9, 13	−	⟨6, 2, 1⟩	7, 10, 14	−	⟨3, 6, 1⟩	6, 9, 12	−
⟨1, 2, 4⟩	6, 10, 12	−	⟨1, 3, 4⟩	5, 9, 11	+	⟨1, 4, 5⟩	4, 12, 15	−
⟨1, 4, 2⟩	6, 9, 12	+	⟨1, 4, 3⟩	5, 9, 11	−	⟨1, 5, 4⟩	4, 12, 14	+
⟨4, 2, 1⟩	6, 10, 14	−	⟨3, 4, 1⟩	5, 10, 13	−	⟨5, 4, 1⟩	4, 12, 15	−
⟨1, 2, 5⟩	5, 11, 14	−	⟨1, 3, 5⟩	4, 10, 13	+	⟨1, 4, 6⟩	6, 9, 10	+
⟨1, 5, 2⟩	5, 10, 14	+	⟨1, 5, 3⟩	4, 10, 13	−	⟨1, 6, 4⟩	6, 10, 12	−
⟨5, 2, 1⟩	5, 10, 15	−	⟨5, 3, 1⟩	4, 10, 14	−	⟨4, 6, 1⟩	6, 10, 13	−

σ_3

σ_3	ABC	?	σ_3	ABC	?	σ_3	ABC	?
⟨1, 5, 6⟩	5, 10, 12	+	⟨4, 2, 5⟩	6, 13, 16	−	⟨3, 4, 6⟩	7, 10, 11	+
⟨1, 6, 5⟩	5, 11, 14	−	⟨5, 2, 4⟩	6, 12, 14	+	⟨3, 6, 4⟩	7, 11, 13	−
⟨5, 6, 1⟩	5, 10, 13	−	⟨5, 4, 2⟩	6, 12, 15	−	⟨4, 6, 3⟩	7, 10, 12	−
⟨2, 3, 4⟩	7, 11, 13	−	⟨4, 2, 6⟩	8, 10, 12	+	⟨3, 5, 6⟩	6, 11, 13	−
⟨4, 2, 3⟩	7, 10, 13	+	⟨6, 2, 4⟩	8, 12, 14	−	⟨3, 6, 5⟩	6, 12, 15	−
⟨3, 4, 2⟩	7, 10, 13	−	⟨4, 6, 2⟩	8, 10, 13	−	⟨5, 6, 3⟩	6, 10, 12	+
⟨2, 3, 5⟩	6, 12, 15	−	⟨5, 2, 6⟩	7, 10, 13	+	⟨5, 4, 6⟩	6, 12, 13	+
⟨5, 2, 3⟩	6, 10, 14	+	⟨6, 2, 5⟩	7, 13, 16	−	⟨4, 6, 5⟩	6, 13, 16	−
⟨5, 3, 2⟩	6, 10, 14	−	⟨5, 6, 2⟩	7, 10, 13	−	⟨5, 6, 4⟩	6, 12, 14	−
⟨2, 3, 6⟩	8, 10, 11	+	⟨3, 4, 5⟩	5, 13, 16	−			
⟨6, 2, 3⟩	8, 10, 13	−	⟨5, 3, 4⟩	5, 12, 14	+			
⟨3, 6, 2⟩	8, 10, 13	−	⟨5, 4, 3⟩	5, 12, 14	−			

Таблица 5.22 (продолжение)

σ_4

σ_4	ABC	?	σ_4	ABC	?	σ_4	ABC	?
⟨1, 2, 3, 4⟩	8, 12, 14	—	⟨1, 5, 2, 6⟩	8, 12, 15	+	⟨4, 2, 3, 5⟩	8, 13, 16	+
⟨1, 4, 2, 3⟩	8, 11, 14	+	⟨1, 2, 6, 5⟩	8, 14, 17	—	⟨5, 2, 3, 4⟩	8, 14, 16	—
⟨1, 3, 4, 2⟩	8, 11, 14	—	⟨1, 5, 6, 2⟩	8, 12, 15	—	⟨5, 2, 4, 3⟩	8, 14, 16	—
⟨4, 2, 3, 1⟩	8, 12, 16	—	⟨5, 2, 6, 1⟩	8, 12, 16	—	⟨5, 3, 4, 2⟩	8, 14, 17	—
⟨1, 2, 3, 5⟩	7, 13, 16	—	⟨1, 3, 4, 5⟩	6, 14, 17	—	⟨4, 2, 3, 6⟩	10, 12, 14	+
⟨1, 5, 2, 3⟩	7, 12, 16	+	⟨1, 3, 5, 4⟩	6, 14, 16	+	⟨2, 3, 6, 4⟩	10, 14, 16	—
⟨1, 3, 5, 2⟩	7, 12, 16	—	⟨1, 5, 4, 3⟩	6, 14, 16	—	⟨4, 2, 6, 3⟩	10, 14, 16	—
⟨5, 2, 3, 1⟩	7, 12, 17	—	⟨5, 3, 4, 1⟩	6, 14, 17		⟨3, 4, 6, 2⟩	10, 12, 15	—
⟨1, 2, 3, 6⟩	9, 11, 12	+	⟨1, 3, 4, 6⟩	8, 11, 12	+	⟨5, 2, 3, 6⟩	9, 12, 15	+
⟨1, 2, 6, 3⟩	9, 11, 13	—	⟨1, 3, 6, 4⟩	8, 12, 14	—	⟨2, 3, 6, 5⟩	9, 15, 18	—
⟨1, 3, 6, 2⟩	9, 11, 14	—	⟨1, 4, 6, 3⟩	8, 11, 13	—	⟨5, 2, 6, 3⟩	9, 12, 15	—
⟨2, 3, 6, 1⟩	9, 12, 15	—	⟨3, 4, 6, 1⟩	8, 12, 15	—	⟨3, 5, 6, 2⟩	9, 13, 16	—
⟨1, 4, 2, 5⟩	7, 14, 17	—	⟨1, 3, 5, 6⟩	7, 12, 14	+	⟨5, 2, 4, 6⟩	9, 14, 15	+
⟨1, 5, 2, 4⟩	7, 14, 16	+	⟨1, 3, 6, 5⟩	7, 13, 16	—	⟨4, 2, 6, 5⟩	9, 15, 18	—
⟨1, 5, 4, 2⟩	7, 14, 17	—	⟨1, 5, 6, 3⟩	7, 12, 14	—	⟨5, 2, 6, 4⟩	9, 14, 16	—
⟨5, 2, 4, 1⟩	7, 14, 18	—	⟨3, 5, 6, 1⟩	7, 13, 16	—	⟨5, 4, 6, 2⟩	9, 14, 17	—
⟨1, 4, 2, 6⟩	9, 11, 13	+	⟨1, 5, 4, 6⟩	7, 14, 15	+	⟨5, 3, 4, 6⟩	8, 14, 15	+
⟨1, 2, 6, 4⟩	9, 13, 15	—	⟨1, 4, 6, 5⟩	7, 14, 17	—	⟨3, 4, 6, 5⟩	8, 15, 18	—
⟨1, 4, 6, 2⟩	9, 11, 14	—	⟨1, 5, 6, 4⟩	7, 14, 16	—	⟨5, 6, 3, 4⟩	8, 14, 16	—
⟨4, 2, 6, 1⟩	9, 12, 15	—	⟨5, 4, 6, 1⟩	7, 14, 17	—	⟨5, 4, 6, 3⟩	8, 14, 15	—

Таблица 5.22 (продолжение)

σ_5

σ_3	ABC	?	σ_5	ABC	?
⟨1, 4, 2, 3, 6⟩	9, 16, 19	—	⟨1, 5, 2, 4, 6⟩	10, 16, 18	+
⟨1, 5, 2, 3, 4⟩	9, 16, 18	+	⟨1, 4, 2, 6, 5⟩	10, 16, 19	—
⟨1, 5, 2, 4, 3⟩	9, 16, 18	—	⟨1, 5, 2, 6, 4⟩	10, 16, 18	—
⟨1, 3, 5, 4, 2⟩	9, 16, 19	—	⟨1, 5, 4, 6, 2⟩	10, 16, 19	—
⟨4, 2, 3, 5, 1⟩	9, 15, 19	+	⟨5, 2, 4, 6, 1⟩	10, 16, 19	—
⟨1, 4, 2, 3, 6⟩	11, 13, 15	+	⟨1, 3, 5, 4, 6⟩	9, 16, 17	+
⟨1, 2, 3, 6, 4⟩	11, 15, 17	—	⟨1, 3, 4, 6, 5⟩	9, 16, 19	—
⟨1, 4, 2, 6, 3⟩	11, 13, 15	—	⟨1, 3, 5, 6, 4⟩	9, 16, 18	—
⟨1, 3, 4, 6, 2⟩	11, 13, 16	—	⟨1, 5, 4, 6, 3⟩	9, 16, 18	—
⟨4, 2, 3, 6, 1⟩	11, 14, 17	—	⟨5, 3, 4, 6, 1⟩	9, 16, 19	—
⟨1, 5, 2, 3, 6⟩	10, 14, 17	+	⟨4, 2, 3, 5, 6⟩	11, 15, 17	+
⟨1, 2, 3, 6, 5⟩	10, 16, 18	—	⟨4, 2, 3, 6, 5⟩	11, 17, 20	—
⟨1, 5, 2, 6, 3⟩	10, 14, 17	—	⟨5, 2, 3, 6, 4⟩	11, 16, 18	—
⟨1, 3, 5, 6, 2⟩	10, 14, 17	—	⟨5, 2, 4, 6, 3⟩	11, 16, 17	—
⟨5, 2, 3, 6, 1⟩	10, 14, 18	—	⟨5, 3, 4, 6, 2⟩	11, 16, 19	—

σ_6

σ_6	ABC	?
⟨1, 5, 2, 3, 4, 6⟩	12, 18, 20	+
⟨4, 2, 3, 5, 1, 6⟩	12, 17, 20	—
⟨1, 4, 2, 3, 6, 5⟩	12, 18, 21	—
⟨1, 5, 2, 3, 6, 4⟩	12, 18, 20	—
⟨1, 5, 2, 4, 6, 3⟩	12, 18, 20	—
⟨1, 3, 5, 4, 6, 2⟩	12, 18, 20	—
⟨4, 2, 3, 6, 1, 5⟩	12, 17, 20	—

Замечание. Мы ищем только одно оптимальное решение

3. Решение по схеме «ветвей и границ». Вспомогательная задача для решения задачи трех станков по схеме «ветвей и границ» может быть сформулирована различными способами.

Пусть у нас построен некоторый фрагмент решения σ_k. Вспомогательную задачу можно сформулировать как некоторую простую задачу двух станков или же даже одного станка, если пренебречь значениями времен обработки каких-то операций.

Так, если принимать во внимание только времена обработки на третьем станке для всех j, не попавших в $M(\sigma_k)$ - обычно это делается тогда, когда вообще

$$\sum c_i \gg \sum a_i \quad \text{и} \quad \sum_i c_i \gg \sum a_i ,$$

не обращая внимания на значения a_i и b_i (т.е. считая, что $a_i = b_i = 0$), то мы получим совершенно тривиально решаемую вспомогательную задачу (рис. 5.45):

$$\Delta(\sigma_k) = C(\sigma_k) + \sum_{\overline{M}(\sigma_k)} c_j. \tag{5.93}$$

Знак $\overline{M}(\sigma_k)$ здесь следует понимать как то, что сумма длительностей c_j считается только по тем j, которые не входят в $M(\sigma_k)$.

Совсем аналогично можно считать существенным только продолжительности вторых операций - особенно если

$$\sum_i b_i \gg \sum_i c_i, \quad \sum_i b_i \gg \sum_i a_i.$$

В этом случае прогноз-обещание можно вычислять по формуле

$$\Delta(\sigma_k) = B(\sigma_k) + \sum_{\overline{M}(\sigma_k)} b_j + \min_{\overline{M}(\sigma_k)} c_j. \tag{5.94}$$

Здесь знак $\overline{M}(\sigma_k)$ означает, что минимум выбирается среди c_i, которые не входят в $M(\sigma_k)$.

Точно так же, считая существенными только длительности первых операций — особенно если

$$\Sigma a_i \gg \Sigma b_i$$

и

$$\sum_i a_i > \sum_i c_i$$

- можно получить прогноз

$$\Delta(\sigma_k) = \Delta(\sigma_k) + \sum_{\overline{M}(\sigma_k)} a_j + \min_{M(\sigma_k)} (b_j + c_j). \tag{5.95}$$

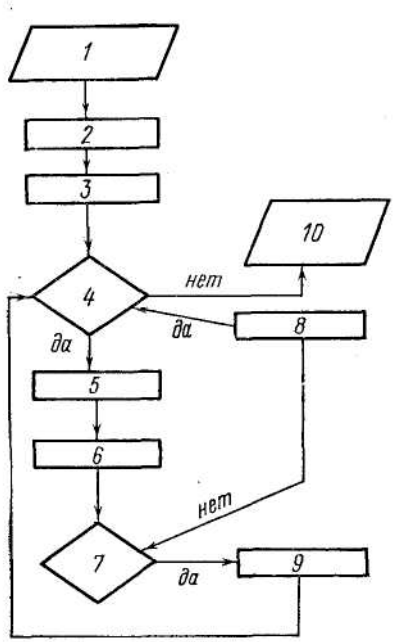


Рис. 5.45. Схема решения задачи трех станков методом ветвей и границ.

1. Подготовка информации к решению. 2. Формируем вспомогательные задачи, вычисляем $\Delta(\sigma_0)$ — для всего множества возможных решений. 3. Строим всевозможные $\sigma_1=\langle i_1 \rangle$, $F^0=\infty$. 4. Существует σ_l такой, что $\Delta(\sigma_l)<F^0(\sigma_n)$? 5. Выберем (σ_l) с наименьшим $\Delta(\sigma_l)$. Применяем операцию ветвления к нашему $\sigma_l=\sigma_k$, т.е. образуем $\sigma_{k+1}=\langle \sigma_k , j \rangle$, j не приналежить (σ_k). 6. $l=k$. 7. $k+1=n$?
8. Формируем для вновь построенных σ_{k+1} вспомогательные задачи, вычисляем $\Delta(\sigma_k)$, фиксируем $\overline{\Delta}(\sigma_k)$, максимальное из всех $\Delta(\sigma_k)$ при фиксированном σ_k. 9. Вычисляем $F(\sigma_n)$ для вновь образованных σ_n, наименьшее $F(\sigma_n)$ и σ_n — запоминаем. 10. Решение закончено, $F^0(\sigma_n)$ — экстремум.

Прогнозы, вычисляемые по формулам (5.95), (5.94), (5.93), есть прогнозы со вспомогательными задачами как бы задачами одного станка. Обозначим их соответственно через Δ_1, Δ_2, Δ_3.

Прогнозы-обещания можно подсчитывать, если в качестве вспомогательной решать (п. 5.17, п. 3) задачи двух станков. Так, пренебрегая длительностями выполнения первых операций, можно для всех j, не входящих в $M(\sigma_k)$, решить задачу двух станков и вычислить оптимальное значение функции-критерия для такой вспомогательной задачи. Обозначим его $B'(b_i, c_j, \sigma_k)$. Тогда $\Delta_4(\sigma_k) = B(\sigma_k) + B'(b_j, c_j, \sigma_k)$ может быть принята за прогноз значений $F(\sigma_n)$ при данном σ_k.

Так, пусть в нашем примере $\sigma_k = \langle 1, 2, 3 \rangle$.

Согласно рис. 5.44 $B(\sigma_k)$=8. Упорядочим оптимально детали которые остались, а именно, 4, 5, 6, считая, что они обрабатываются только на втором и третьем станках. Согласно алгоритму, приведенному на рис. 5.34, эти детали надо обрабатывать в порядке убывания продолжительности третьих операций (так как они короче вторых во всех деталях), т.е. в последовательности $\langle 5,4,6 \rangle$.

Легко подсчитать по правилам расчета $B(\sigma_k)$ для задачи двух станков (формула (5.67), что в нашем случае $B'(b_i, c_i, \sigma_k)$= 12, т.е. $\Delta_4(\sigma_k) = 20$ (Δ_4 в этом случае совпадает с Δ_2).

На рис. 5.46 приведено решение примера, заданного табл. 5.21, по схеме «ветвей и границ» на рисунке рядом с σ_k выписаны $\Delta_1, \Delta_2, \Delta_3, \Delta_4$, Δ_5, где Δ_5 максимум среди $\Delta_1, \Delta_2, \Delta_3, \Delta_4$.

4. Задача четырех станков. Нетрудно понять, что, хотя метод отсеивания вариантов по правилам доминирования и является методом точного решения задачи, уже при малых n трудности вычислений становятся непреодолимыми. Нельзя заранее предсказать и число ветвлений в решении по схеме «ветвей и границ», решающие же правила удается сформулировать только в сравнительно простых случаях.

Еще хуже обстоит дело с более сложными постановками задач теории расписаний. Например, *задачу четырех станков,* где деталь последовательно проходит четыре стадии обработки на разных станках, вообще говоря, нельзя уже свести к поиску экстремальной перестановки.

Действительно, пусть заданы детали, которые последовательно обрабатываются на четырех станках, a_i, b_i, c_i, d_i — продолжительности обработки детали i на 1, 2, 3, 4 станке.

Как и в случае задачи двух станков, можно доказать, что в оптимальном расписании обработки деталей очередность обработки деталей на первых двух станках одинакова, одинакова она и на двух последних станках.

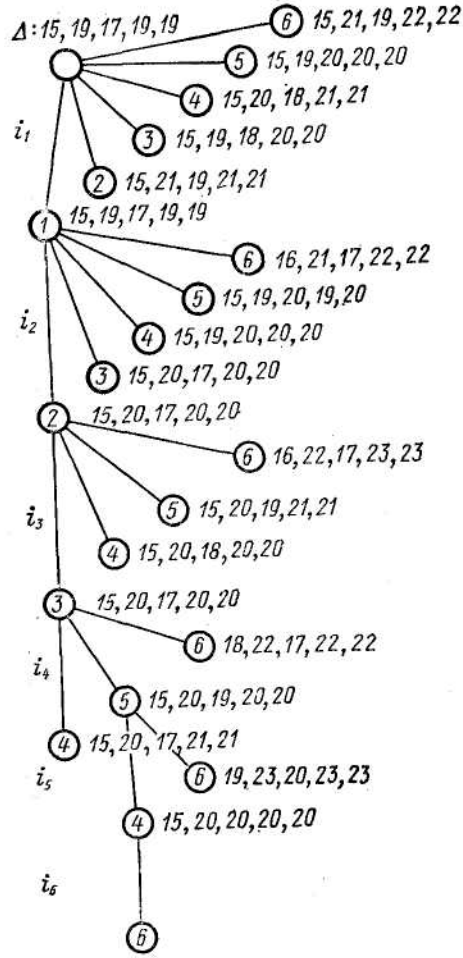


Рис. 5.46. К решению задачи трех станков методом ветвей и границ (пример).

Для первого случая надо в точности повторить доказательство п. 5.17, п. 3, для второго случая в это доказательство надо внести очевидное изменение: деталь i на последнем станке можно поставить

непосредственно перед деталью j, «пододвинув» вправо деталь j (вместе с операциями, которые занимали место перед i и j) на отрезок, равный продолжительности выполнения последней операции детали i. Такое преобразование допустимо, значение функции-критерия не изменится (оно может только уменьшиться, если теперь попробовать сделать график «плотным».

Как показывает простой пример, представленный на рис. 5.47, последовательность выполнения первых двух и последних двух операций в оптимальном решении может не совпадать между собой. На рисунке представлены все возможные соединения очередности выполнения двух деталей ($a_1 = b_1 = c_1 = d_1 = 3$, $a_2 = d_2 = 3$, $b_2 = c_2 = 1$), допустимые с учетом сказанного в предыдущем абзаце.

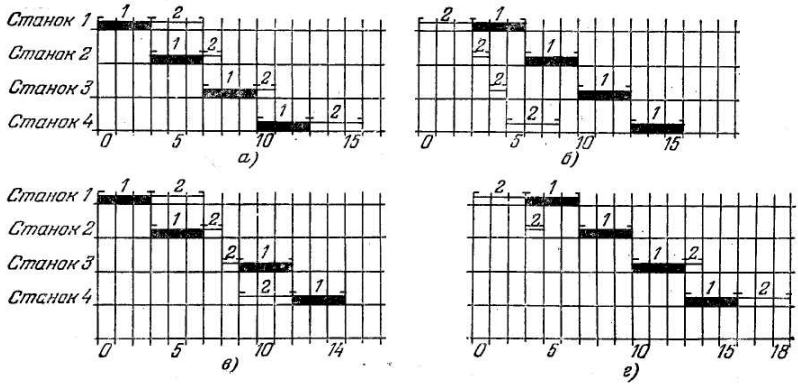


Рис. 5.47. Пример расписания к задаче четырех станков.

Отсюда и следует, что решение задачи четырех станков нельзя представить одной перестановкой, но можно представлять двумя перестановками, соответствующими порядку обработки деталей на первых двух станках и на последних двух.

В реальных условиях решают и более сложные задачи; в них, например, число станков m велико, и последовательность прохождения (в процессе обработки) станков деталями не одинакова.

Пользоваться для решения таких задач точными методами, наподобие описанных, практически безнадежно.

5. Приоритеты. Где нельзя воспользоваться точными методами, обычно пользуются приближенными. Впрочем, не везде удается воспользоваться и приближенными методами, от которых требуется «подойти» к исходному решению как угодно близко, т.е. получить

решение, отстоящее от оптимального на любую заданную заранее величину.

В схеме «ветвей и границ» мы на первых шагах решения уже можем определить, в каких пределах заключается оптимум, в процессе решения нас может удовлетворить достигнутая точность, однако нельзя знать заранее, сколько потребуется ветвлений, чтобы достичь заранее заданную точность.

И все-таки схема «ветвей и границ» — один из самых удобных способов решения дискретных задач оптимизации. Если очень много различных решений, близких к оптимуму, — как говорят, «оптимум размыт», схема «ветвей и границ» позволяет установить, что мы имеем дело с «размытым оптимумом». В этом случае имеет смысл остановить решение, оценив, как далеко может отстоять оптимум от найденного рекордного значения $F(\sigma_n)$. Когда оптимум не «размыт», а «ярко выражен», схема «ветвей и границ» позволяет его обычно быстро и получить,

В практических постановках (с большим числом деталей, станков, с разным прохождением станков деталями в процессе обработки) и точные, и приближенные методы приводят к неприемлемо продолжительным вычислениям. В таких случаях обычно прибегают к помощи эвристических методов, т.е. методов рациональных, однако не показано, что они действительно могут привести к оптимуму или оценить, насколько близко за данное число шагов можно подойти к этому оптимуму.

В эвристических схемах обычно указывается некоторый *приоритет,* в соответствии с которомым и рекомендуется отбирать операции к назначению из списка готовых к выполнению операций. Таким образом, *приоритет* - *это как бы решающее правило* (п. 5.17, п. 2), относительно которого мы, правда, в точности не знаем, ведет ли оно к оптимуму. У нас просто есть какие-то основания быть уверенными в этом.

Так, например, *решающее* правило кратчайшей операции в задаче директора приводит, как мы знаем, к оптимальному решению (п. 5.17, п. 1). Правило кратчайшей операции поэтому очень часто используется в эвристических методах решения произвольных задач составления расписаний: приоритет дается менее продолжительной операции.

Схема ветвления с приоритетами имеет очень простую структуру. На каждом σ_k вычисляется некоторое $\varphi(\sigma_k)$ — его *приоритет.*

Ветвлению подлежит то σ_k, у которого приоритет наибольший (в правиле кратчайшей операции $\varphi(\sigma_k) = \frac{1}{T_{i_k}}$).

6. Случайные ветвления. Схема ветвления с приоритетами благодаря своей простоте часто применяется для решения практических задач календарного планирования. Предложены десятки простых и сложных приоритетов, каждый из которых хорош, вообще говоря, для *своего класса задач*,- мы видели, что в частном случае задачи трех станков (п. 5.20, п. 1) приоритет $\left(\frac{1}{a_i}\right)$ приводит к оптимальному решению.

Предложены и используются на практике и так называемые *случайные ветвления,* когда σ_k выбирается для ветвления случайно, в соответствии с некоторой заданной или вычисляемой вероятностью выбора σ_k. Такие схемы ветвления является естественным применением идеи случайного поиска решений к отысканию оптимума.

При случайных ветвлениях всегда есть *положительная вероятность* получения оптимума, для одних способов задания (вычисления) вероятностей выбора σ_k для ветвления бо̀льшая, для других — меньшая.

Разработаны методы, в которых вероятность получения оптимума увеличивается по мере решения задачи — такие методы получили название *методов адаптации.*

Показано, что лучшая адаптация достигается при так называемом *человек-машинном решении задач* в режиме диалога.

Микромодуль19.
Индивидуальные тестовые задачи

Упражнение 1. Показать, что σ^1_k лучше σ^2_k, если

$$t_1(\sigma^1_k) = t_1(\sigma^2_l), \qquad t_k(\sigma^1_k) = t_l(\sigma^2_l);$$

$M(\sigma^1_k)$ содержит все $M(\sigma^2_k)$, а $F(\sigma^1_l) \le F(\sigma^1_l)$.

Упражнение 2. Как мы отмечали, каждая часть критического пути есть критический путь в том смысле, что лежащие на критическом пути пункты i и j также соединены «наиболее длинно» и не

существует другого пути, который соединяет эти пункты «более долгим образом». Докажите это.

Упражнение 3. Найдите критический путь на сетевом графике рис. 5.16 по алгоритму, приведенному на рис. 5.37.

Упражнение 4.

Утверждение 1. *Решение задачи о назначениях не изменится, если каждое a_{ij} в i-й строке уменьшить на одно и то же число Δ_i.*

Утверждение 2. *Решение задачи о назначениях не изменится, если каждое a_{ij} в j-м столбце уменьшить на одно и то же число δ_j.*

Доказать эти утверждения, воспользовавшись формулировкой: показать, что значение функции-критерия при этом уменьшается на

$$\sum_i \Delta_i + \sum_j \delta_j.$$

Упражнение 5. Число ветвлений можно сократить, если лучше «прогнозировать» значение $F(\sigma_n)$, вычисляя $b(\sigma_k)$. Для этого, получив σ_k, например, можно вычеркнуть все строки i, где i берется из $M(\sigma_{k-1})$, а матрицу, которая осталась, преобразовать согласно утверждениям 1 и 2. Продумайте это предложение и постарайтесь использовать при решении задачи.

Упражнение 6. Постройте блок-схему описанного алгоритма и проверьте по ней правильность решения примера. От чего зависит количество ветвлений в решении? Оцените предложение: пока не получено σ_n, ветвлению подвергается σ_k с максимальным k, а среди таких — уже с максимальным в $b(\sigma_k)$.

Упражнение 7. Продумайте следующее практическое замечание: «лучше вычислять $b(\sigma_k)$ проще, чем точнее».

Упражнение 8. Прокомментируйте еще раз упражнение 7. Составьте блок-схему алгоритма решения обобщенной задачи о рюкзаке.

Упражнение 9. Задача трех станков сводится к поиску экстремальной перестановки. Это доказывается аналогично тому, как и в задаче двух станков - сначала для первых двух станков, потом для последних двух. Докажите это.

Упражнение 10. Постройте что-то аналогичное «интервалам очередности» для задачи трех станков.

Упражнение 11. Исходя из задачи двух станков, сформулируйте и решите другие частные случаи задачи трех станков? Например, рассмотрев условие, аналогичное (5.92), по второму и третьему станку.

Упражнение 12. Объясните, откуда в формуле (5.94) взялось последнее слагаемое.

Упражнение 13. Постройте схему «ветвей и границ» для задачи четырех станков? Могли бы с помощью схемы «ветвей и границ» показать, что решение

$$\sigma = \begin{pmatrix} 1, 2, 3, 4 \\ 1, 2, 3, 4 \end{pmatrix}$$

оптимально для условий задачи, заданных табл. 5.23,

Таблица 5.23

i	1	2	3	4
a_i	1	1	1	2
b_i	2	2	3	3
c_i	3	3	2	4
d_i	4	4	2	3

Микромодуль 20
Комбинаторика и нечеткие структуры

5.21. Введение в теорию трансверсалей (представителей)

$\forall t \in T$ выполняется условие $t \in S_{\phi(t)}$.

Другими словами, для элементов трансверсали существует такая нумерация, при которой $t_i \in S_i, i = 1, 2, \ldots, m$.

Поскольку *T* — множество, то все его элементы различны, отсюда и название «система *различных* представителей».

Если E — непустое конечное множество и $\varphi = (S_1, \ldots, S_m)$ — семейство (не обязательно различных) непустых его подмножеств, трансверсалью (или системой различных представлений) для φ называется подмножество множества E, состоящее из m элементов: по одному из каждого множества S_i.

Общие трансверсали. Если E — непустое конечное множество, а $\varphi = (S_1, \ldots, S_m)$ и $\tau = (T_1, \ldots, T_m)$ — два семейства его непустых подмножеств, то интересно знать, когда существует общая трансверсаль для φ и τ, то есть множество, состоящее из m различных элементов множества E и являющееся трансверсалью и для φ, и для τ.

Рассмотрим пример. Предположим, что $E = \{1, 2, 3, 4, 5, 6\}$, а $S_1 = S_2 = \{1, 2\}, S_3 = S_4 = \{2, 3\}$, $S_5 = \{1, 4, 5, 6\}$. Подсемейство $\varphi' = (S_1, S_2, S_3, S_5)$ имеет трансверсаль, например $\{1, 2, 3, 4\}$. Трансверсаль произвольного подсемейства семейства φ будем называть частичной трансверсалью для φ; в нашем примере семейство φ имеет несколько частичных трансверсалей (например, $\{1, 2, 3, 6\}$, $\{2, 3, 6\}$, $\{1, 5\}$, $\emptyset$ и т.д.). Ясно, что любое подмножество частичной трансверсали само является частичной трансверсалью.

Естественно спросить: при каких условиях данное семейство подмножеств некоторого множества имеет трансверсаль? Легко

увидеть связь между этой задачей и задачей о свадьбах, если взять за E множество девушек, а за S_i — множество девушек, знакомых юноше $b_i (1 \leqslant i \leqslant m)$; трансверсалью в этом случае является множество из m девушек, такое, что каждому юноше соответствует ровно одна (знакомая ему) девушка. Следовательно, теорема Холла дает необходимое и достаточное условие существования трансверсали для данного семейства множеств. Сформулируем теорему Холла для этого случая и дадим другое ее доказательство, принадлежащее **Р.Радо**.

Теорема Пусть E — непустое конечное множество и $\varphi = (S_1, \ldots, S_m)$ — семейство непустых его подмножеств; тогда φ имеет трансверсаль в том и только в том случае, если для любых k подмножеств S_i их объединение содержит, по меньшей мере, k элементов $(1 \leqslant k \leqslant m)$.

Доказательство Необходимость этого условия очевидна. Для доказательства достаточности установим, что если одно из подмножеств (скажем, S_1) содержит более одного элемента, то можно удалить один элемент из S_1, не нарушив условия теоремы. Повторением этой процедуры мы добьемся сведения задачи к тому случаю, когда каждое подмножество содержит только один элемент. Тогда утверждение станет очевидным.

Осталось обосновать законность этой "процедуры сведения". Предположим, что S_1 содержит элементы x и y, удаление каждого из которых нарушает условие теоремы. Тогда существуют подмножества A и B множества $\{2, 3, \ldots, m\}$, обладающие тем свойством, что

$$\left|\bigcup_{j\in A} S_j \bigcup (S_1 - \{x\})\right| \leqslant |A|$$

$$\left|\bigcup_{j\in B} S_j \bigcup (S_1 - \{y\})\right| \leqslant |B|.$$

Но эти два неравенства приводят к противоречию, поскольку

$$|A| + |B| + 1 = \left|A \bigcup B\right| + \left|A \bigcap B\right| + 1 \leqslant$$

$$\leqslant \left|\bigcup_{jA \bigcup B} S_j \bigcup S_1\right| + \left|\bigcup_{j\in A \bigcap B} S_j\right| \leqslant \quad \text{(по условию)}$$

$$\leqslant \left|\bigcup_{j\in A} S_j \bigcup (S_1 - \{x\})\right| + \left|\bigcup_{j\in B} S_j \bigcup S_1 - \{y\})\right| \leqslant$$

$$\leqslant \quad \text{(так как } |S_1| \geqslant 2\text{)}$$

$$\leqslant |A| + |B| \quad \text{(по предположению).}$$

Преимущество этого доказательства в том, что оно проводится, по существу, лишь в один шаг, в отличие от доказательства Халмоша-Вогена, которое предполагает исследование двух отдельных случаев. (Однако доказательство Радо труднее перевести на весьма наглядный матримониальный язык!).

Следствие. В тех же обозначениях, что и выше, φ имеет частичную трансверсаль мощности t тогда и только тогда, если для любых k подмножеств S_i их объединение содержит, по меньшей мере, $k + t - m$ элементов.

Доказательство Требуемый результат можно получить, применив теорему Холла в лексике трансверсалей к семейству

$\varphi' = S_1 \bigcup D, \ldots, S_m \bigcup D$, где D — произвольное множество, не пересекающееся с E и состоящее из $m - t$ элементов. Заметим, что φ имеет частичную трансверсаль мощности t тогда и только тогда, если φ' имеет трансверсаль.

Следствие Если E и φ такие же, как и прежде, а X — любое подмножество из E, то X содержит частичную трансверсаль мощности t для φ тогда и только тогда, если для каждого подмножества A множества $\{1, \ldots, m\}$

$$\left| (\bigcup_{j \in A} S_j) \bigcap X \right| \geqslant |A| + t - m.$$

Доказательство Достаточно применить предыдущее следствие к семейству $\varphi_x = (S_1 \bigcap X, \ldots, S_m \bigcup X)$.

Приложение теории трансверсалей

Используются понятия трансверсалей, на основании чего доказывается теорема о модификации латинского прямоугольника. Вводятся определения $(0,1)$-матрицы, формулируются и доказываются теоремы Кенига-Эгервари и об общей трансверсали.

Теорема Пусть M латинский $m \times n$-прямоугольник, причем, $m < n$; тогда M можно расширить до латинского квадрата добавлением $n - m$ новых строк.

Доказательство Докажем, что M можно расширить до латинского $(m+1) \times n$-прямоугольника; повторяя эту процедуру, мы придем к латинскому квадрату.

Пусть $E = \{1, 2, \ldots, n\}$ и $\varphi = (S_1, \ldots, S_n)$, где через S_i обозначено множество, состоящее из тех элементов множества E, которые **не встречаются** в i-м столбце матрицы M. Если мы сможем доказать, что φ имеет трансверсаль, то тем самым мы докажем теорему, поскольку элементы этой трансверсали и образуют дополнительную строку. По теореме Холла достаточно доказать, что объединение любых k множеств S_i содержит по меньшей мере k различных элементов. А это очевидно, ибо любое такое объединение содержит $(n - m) \times k$ элементов (включая повторения), значит, по крайней мере, один из них повторялся бы более чем $n - m$ раз, что невозможно.

Определение (0,1) матрицы или матрицы инциденций. Другой подход к изучению трансверсалей семейства $\varphi = (S_1, \ldots, S_m)$ непустых подмножеств множества $E = \{e_1, \ldots, e_n\}$ состоит в исследовании $(m \times n)$-матрицы $A = (a_{ij})$, в которой $a_{ij} = 1$, если $e_j \in S_i$, и $a_{ij} = 0$ в противном случае. (Любую такую матрицу, все элементы которой равны 0 или 1, мы называем $(0,1)$-матрицей) этого семейства.

Определение словарного ранга. Назовем словарным рангом матрицы A наибольшее число единиц в A, никакие две из которых не лежат в одной и той же строке или в одном и том же столбце. Тогда φ имеет трансверсаль в том и только в том случае, если словарный ранг матрицы A равен m. Более того, словарный ранг матрицы A равен в точности числу элементов частичной трансверсали, обладающей наибольшей возможной мощностью. В качестве второго приложения теоремы Холла рассмотрим известный результат о $(0,1)$-матрицах, называемой теоремой Кенига-Эгервари.

Теорема (Кенига-Эгервари, 1931). Словарный ранг $(0,1)$-матрицы A равен минимальному числу μ строк и столбцов, которые в совокупности содержат все единицы из A.

Замечание В качестве иллюстрации этой теоремы рассмотрим матрицу

$$\begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \end{array} \begin{array}{c} \begin{array}{cccccc} e_1 & e_2 & e_3 & e_4 & e_5 & e_6 \end{array} \\ \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \end{array},$$

которая является матрицей семейства $\varphi = (S_1, \ldots, S_5)$. Ясно, что и ее словарный ранг, и число μ равны четырем.

Доказательство. Очевидно, что словарный ранг не может превосходить числа μ. Чтобы доказать равенство, можно без потери общности предположить, что все единицы из A содержатся в r строках и s столбцах (где $r + s = \mu$) и что строки и столбцы расположены в таком порядке, что в нижнем левом углу матрицы А находится $(m-r) \times (n-s)$-подматрица, полностью состоящая из нулей.

Если $i \leqslant r$, то определим S_i как множество целых чисел $j \leqslant n - s$, таких, что $a_{ij} = 1$. Нетрудно проверить, что объединение любых k множеств S_i содержит по меньшей мере k целых чисел; поэтому семейство $\varphi = (S_1, \ldots, S_r)$ имеет трансверсаль. Отсюда следует, что подматрица M из A содержит множество из r единиц, никакие две из которых не принадлежат одной и той же строке или одному и тому же столбцу. Аналогично,

матрица N содержит множество из s единиц, обладающих тем же свойством. Таким образом, матрица A содержит множество из $r+s$ единиц, никакие две из которых не принадлежат одной и той же строке или одному и тому же столбцу. Тем самым показано, что μ не превосходит словарного ранга.

Только что была доказана теорема Кенига-Эгервари с помощью теоремы Холла, а доказательство теоремы Холла с помощью теоремы Кенига-Эгервари значительно проще. Следовательно, эти две теоремы в некотором смысле эквивалентны.

Общие трансверсали. Если E — непустое конечное множество, а $\varphi=(S_1, \ldots, S_m)$ и $\tau=(T_1, \ldots, T_m)$ — два семейства его непустых подмножеств, то интересно знать, когда существует общая трансверсаль для φ и τ, то есть множество, состоящее из m различных элементов множества E и являющееся трансверсалью и для φ, и для ι.

Сформулируем необходимое и достаточное условие для того, чтобы два семейства φ и τ имели общую трансверсаль; заметим, что эта теорема сводится к теореме Холла, если положить T_j-E для $1 \leqslant j \leqslant m$.

Теорема Пусть E — непустое конечное множество, а $\varphi=(S_1, \ldots, S_m)$ и $\tau=(T_1, \ldots, T_m)$ — два семейства его непустых подмножеств. Тогда φ и τ имеют общую трансверсаль в том и только в том случае, если для всех подмножеств A и B множества $\{1, \ldots, m\}$

$$\left|\left(\bigcup_{i \in A} S_i\right) \bigcap \left(\bigcup_{j \in B} T_j\right)\right| \geqslant |A|+|B|-m.$$

Набросок доказательства. Рассмотрим семейство $U = \{U_i\}$ подмножеств множества $E \bigcup \{1, \ldots, m\}$ (считаем, что E и $\{1, \ldots, m\}$ не пересекаются), где множеством индексов также является $E \bigcup \{1, \ldots, m\}$ и где $U_i = S_i$, если $i \in \{1, \ldots, m\}$, и $U_i = \{i\} \bigcup \{i\} \bigcup \{j : i \in T_j\}$, если $i \in E$. Нетрудно проверить, что φ и τ имеют общую трансверсаль тогда и только тогда, если семейство U имеет трансверсаль. Применяя затем теорему Холла к семейству U, получим нужный результат.

Условия, при которых существует общая трансверсаль для трех семейств непустых подмножеств некоторого множества, пока что не известны, и задача нахождения таких условий кажется очень трудной. Многие попытки решения этой задачи используют теорию матроидов; и действительно, некоторые задачи теории трансверсалей становятся почти тривиальными, если рассматривать их с точки зрения теории матроидов.

5.22. Системы различных представителей семейств множеств, которые принадлежат различным классам по отношению эквивалентности

Одной из тенденций, которые сложились при обработке больших объемов информации, является разбиения множества обрабатываемых данных на классы толерантности по некоторому рефлексивному и симметричному отношению. Часто таким отношением служит «отношение соседства» на некоторой метрике. Следующим шагом является выделения системы представителей для классов толерантности и формирование системы «эталонных представителей». Реализация данного подхода приводит к необходимости решения *комбинаторных* задач, связанных с определением существования системы различных представителей для семейства множеств, и определение числа систем различных представителей, которые

удовлетворяют различным ограничениям. Среди ограничений такого рода обычно используются ограничения на разграничение представителей. Разграничение представителей может вводиться как принадлежность различным классам по некоторому отношению эквивалентности, или через значение весовой функции.

Другим источником задач о системах представителей являются вопросы алгоритмического построения латинских квадратов. Среди таких алгоритмов важное место занимают алгоритмы, основанные на построении латинского квадрата по строкам путем нахождения системы различных представителей семейства множеств элементов, отсутствующих в столбцах.

Для случая обычных систем множеств сформулированные вопросы рассматриваются *в теории систем представителей (теории трансверсалей), которая, как мы видели, представляет важную часть современной комбинаторики*. Основу представляют уже упоминавшиеся классические результаты Холла, Фробениуса, Кенига и др. по существованию и перечислению систем различных представителей.

В данном микромодуле задачи о существовании и числе систем различных представителей рассматриваются для систем нечетких множеств с двумя видами ограничений: представители должны принадлежать различных классам эквивалентности или значение весовой функции ограниченно снизу. Из описанных результатов следует, что основные результаты теории трансверсалей для обычных семейств множеств переносятся с небольшими изменениями на случай взвешенных семейств множеств. Это позволяет использовать для решения поставленных вопросов многочисленные результаты теории перманентов. В последних разделах микромодуля рассмотрен вопрос об автоморфизме нечетких множеств, в частности об установлении нетривиальности группы автоморфизмов нечеткого множества.

Пусть дано множество X, на котором заданы отношения эквивалентности R. Пусть есть семейство подмножеств X_1,....., X_n множества X.

Определение. Систему элементов a_1,....., a_n будем называть трансверсалью семейства X_1,....., X_n и отношения *R*, если выполнены условия:

1. a_i Î X_i, $i = 1, ..., n$,
2. a_i не сравнимо с a_j *(mod R)* при $i \neq j$.

Вопрос существования трансверсали семейства X_1,....., X_n и отношения *R* решается следующим утверждением.

Теорема 1. *Для семейства множеств* X_1,....., X_n *и отношения эквивалентности* R *существует трансверсаль в том и только в том случае, когда выполнены условия:*

Множество X_{i1} *Ė.....* $\dot{E}$ X_{ik}, *имеет не менее* k *классов эквивалентности относительно* R *для всех* k = 1,....., n *и всех* $1 \in i_1 < ..... < i_{ik} \in n$.

Доказательство. Если для семейства X_1,....., X_n и отношения эквивалентности R существует трансверсаль, то тогда для каждого k = 1, ..., n и любых $1 \in i_1 < ..... < i_{ik} \in n$ число классов множества X_{i1} *Ė.....* $\dot{E}$ X_{ik} по отношению R будет не меньше, чем k. Поэтому условия теоремы необходимы.

Пусть теперь условия теоремы выполнены. Докажем достаточность индукцией по n. При n = 1 утверждение очевидно. Пусть утверждение справедливо для любого семейства из n -1 подмножеств. Пусть дано семейство из n подмножеств X_1,.....X_n. Возможны два случая.

а) Семейство X_1,....., X_n такое, что для каждого k, $1 \in k < n$ и всех $1 \in i_1 < ..... < i_k \in n$ множество X_{i1} *Ė.....* $\dot{E}$ X_{ik} имеет не меньше k + 1 классов эквивалентности по R.

По условию X_1[1]0. Выберем произвольный элемент x_1 Î X_1, и исключим элемент x_1 и все эквивалентные с ним элементы из всех подмножеств X_2,....., X_n. Получим семейство X'_2,.....X'_n, которое состоит из n - 1 подмножеств и удовлетворяющее условию теоремы. По предположению индукции существует трансверсаль семейства X'_2,.....,X'_n и отношения эквивалентности R. Пусть это будет набор x_2,....., x_n. Тогда набор x_1, x_2,....., x_n будет трансверсалью семейства X_1,....., X_n и отношения эквивалентности R.

б) Семейство множеств X_1,.....,X_n такое, что существуют k, $1 \in k < n$ и набор $1 \in i_1 < i_2 < ... < i_k \in n$ такие, что ножина X_{i1} *Ė.....* $\dot{E}$ X_{ik} имеет точно k классов эквивалентности по R.

Не нарушая общности, считаем, что i_1 = 1, i_2 = 2, ..., i_k = k. Поскольку $k \in n$ - 1, то по предположению индукции существует трансверсаль семейства X_1, ..., X_k и отношения эквивалентности R. Пусть это будет набор x_1, ..., x_k. Исключим теперь элементы x_1, ..., x_k, а также все элементы, эквивалентные им, из множеств X_{k+1},..... , X_n. Получим
семейство множеств X'_{k+1},..... , X'_n Покажем, что для полученного семейства выполнены условия теоремы.

Предположим противное, и пусть множество X'_{k+t1}*Ė.....* $\dot{E}$ X'_{k+ts} имеет меньше, чем s классов эквивалентности относительно R для некоторых индексов s ,1 $\in s \in$ n- k, 1 $\in$ £t_1 < .... < $t_s \in n$-k. Тогда множество

$$X_1 \dot{E}\ldots\ldots \dot{E}\, X_k \dot{E}\, X'_{k+t1} \dot{E}\ldots\ldots \dot{E}\, X'_{k+ts},$$

так же, как и множество

$$X_1 \dot{E}\ldots\ldots \dot{E}\, X_k \dot{E}\, X'_{k+t1} \dot{E}\ldots\ldots \dot{E}\, X'_{k+ts},$$

имеет меньше, чем $k + s$ классов эквивалентности относительно R, что противоречит условию теоремы. Значит по предположению индукции для множеств $X'_{k+1},\ldots\ldots,X'_n$ и отношения эквивалентности R существует трансверсаль. Объединяя ее с трансверсалью для множеств $X_1,\ldots\ldots,X_k$ и отношения R, получим необходимую трансверсаль. Теорема доказана.

Обобщим теперь приведенное высше определение трансверсаля семейства множеств и отношения эквивалентности.

Определение. Частной трансверсалью семейства множеств $X_1,\ldots\ldots,X_n$ и отношения эквивалентности R назовем трансверсаль некоторого подсемейства множеств данного семейства и отношения R.

Сведением к предыдущей теореме может быть доказана следующая теорема.

Теорема 2. *Семейство множеств $X_1,\ldots\ldots,X_n$ и отношения эквивалентности R имеют частную трансверсаль мощности t тогда и только тогда, когда для любых k подмножеств из $X_1,\ldots\ldots,X_n$ их объединение содержит не менее $k + t$-п классов эквивалентности по R.*

Рассмотрим теперь вопрос о числе трансверсалей семейства множеств $X_1,\ldots\ldots,X_n$ и отношения эквивалентности R.

Определим матрицу инциденций M семейства $X_1,\ldots\ldots,X_n$ и отношения эквивалентности R. Строки матрицы M соответствуют множествам $X_1,\ldots\ldots,X_n$, а столбцы - классам эквивалентности множества $X = X_1 \dot{E}\ldots\ldots \dot{E}\, X_n$ по отношению R. Пусть это классы $C_1,\ldots\ldots, C_q$. Множеству X_i и классу C_j ставим в соответствие число t_{ij} элементов множества X_i, которые принадлежат классу C_j (если таких элементов нет - ставим нуль). Справедливая следующая теорема.

Теорема 3. *Число трансверсалей семейства множеств $X_1,\ldots\ldots,X_n$ и отношений эквивалентности R равно перманенту соответствующей матрицы инциденций M.*

Доказательство. Пусть $t_{1i1} \cdot\ldots\cdot t_{nin}$ - произвольный элемент перманента матрицы инциденций M. Если $t_{1i1} \cdot\ldots\cdot t_{nin} = 0$ для некоторого набора $i_1,\ldots\ldots, i_n$, то значит $t_{pip} = 0$ для некоторого $p \hat{I} \{1, \ldots\ldots n\}$ и тогда нет элементов множества X_p в классе C_{ip}, поэтому трансверсали $x_1,\ldots\ldots,x_n$ с условием

$$x_1 \hat{I} X_1,\ x_2 \hat{I} X_2,\ \ldots\ldots\ldots x_n \hat{I} X_n$$
$$x_1 \hat{I} C_{i1},\ x_2 \hat{I} C_{i2},\ldots\ldots\ldots x_n \hat{I} C_{in}$$

не существует. Если $t_{1u1}.....t_{nin}$ ^{1} 0, то существет t_{1u1} элементов X_1 в классе C_{1u1}, t_{2u2} элементов X_2 в классе C_{2u2} и т.д. Таким образом, имеем $t_{1u1}.....t_{nin}$ трансверсалей семейства множеств $X_1,.....,X_n$ и отношений эквивалентности R. Значит, каждому члену перманента $t_{1u1}.....t_{nin}$ соответствует $t_{1u1}.....t_{nin}$ трансверсалей. Обратно, пусть $x_1, ...,x_n$ -
трансверсаль семейства множеств $X_1,.....,X_n$ и отношения эквивалентности R. Определим перестановку $i_1,... ,i_n$, где $x_1 \hat{I} C_{i1}$, $x_2 \hat{I} C_{i2},.....x_n \hat{I} C_{in}$. Тогда трансверсали $x_1 , ...,x_n$ отвечают $t_{1u1}.....t_{nin}$ трансверсалей $y_1, ...,y_n$, таких, что

$$x_1 \circ y_1 \ (mod\ R),....., x_n \circ y_n (mod\ R).$$

Эти $t_{1u1}.....t_{nin}$ трансверсалей отвечают члену перманента $t_{1u1}.....t_{nin}$. Следовательно, с одной стороны,

$$per\ M = \mathring{a}\ t_{1u1}.....t_{nin},$$
$$(i_1.....i_n)$$

а с другой стороны - это число трансверсалей семейства множеств $X_1, ...,X_N$ и отношения эквивалентности R. Теорема доказана.

Замечание. Если R - отношение равенства, то теорема 1 превращается в известную теорему Хола, а теоремы 2 и 3 - в стандартные комбинаторные утверждения.

5.23. Системы различных представителей семейства нечетких множеств

Пусть $X_1,.....X_n$ - семейство нечетких подмножеств конечного множества X. Для каждого i= 1.....n нечеткое множество X_i, определяется весовой функцией $m_i : X \to R^+$, где R^+ - множество неотрицательных действительных чисел, при этом полагаем

$m_i (x) = 0$, если $x \hat{I} X_i$,

$m_i (x)$ - вес элемента $x \hat{I} X_i$, $0 \in m_i (x)$.

Определение. Набор элементов $(a_1 , ..., a_n)$ множества X будем называть системой раздичных представителей семейства нечетких подмножеств $X_1,.....,X_n$, если выполнены условия:

1. $m_i(a_i) > 0$ для всех $i = 1, ..., n$,
2. $a_i \ ^1 \ a_j$ при $i \ ^1 \ j$

Определим вес весовой функции $m : X \to R^+$ как число элементов X, на которых функция m принимает ненулевое значение. Обозначим через $||m(x)||$ вес функции $m(x)$. Пусть X_1, X_2 - два нечеткие множества, которые определяются весовыми функциями $m_1(x)$ и $m_2(x)$ соответственно. Тогда множество $X_1 \dot{E} X_2$ имеет весовую функцию

$m(x) = max\ (m_1\ (\ x),\ m_2\ (x))$. Теперь можно указать условия, при которых существует система различных представителей для семейства нечетких множеств. Справедлива следующая теорема.

Теорема 4. *Семейство нечетких множеств* $X_1, \ldots, X_n$ *имеет систему различных представителей в том и только в том случае, когда выполнены условия:*

$$\| my\ (x) \|^1\ k \text{ для } Y = X_{i1}\ \dot{E} \ldots\ldots\ \dot{E}\ X_{ik}$$

для всех $k = 1, \ldots\ldots, n$ *и всех* $1 \in i_i, < \ldots < i_k \in n$.

Доказательство идейно повторяет доказательство теоремы 1.

Рассмотрим теперь вопрос о числе систем различных представителей для семейств нечетких множеств. Пусть $X_1, \ldots, X_n$ - семейство нечетких множеств. Уровнем системы различных представителей $(a_1,\ldots\ldots,a_n)$ семейства назовем число

$$a = min\ (m_1\ (a_1),\ldots\ldots,\ m_n\ (a_n)).$$

Определим матрицу инциденций семейства нечетких множеств $X_1, \ldots, X_n$ как матрицу $A = (a_{ij})$ размера $n \times m$, где

$$a_{ij} = m_i(x_j),\ X = \{x_1,\ldots\ldots,\ x_m\};\ i = 1,\ldots\ldots,\ n;\ j = 1,\ldots\ldots,\ m.$$

Для матрицы A определим скелетную матрицу $\bar{A} = (b_{ij})$, где $b_{ij} = 0$ при $a_{ij} = 0$, $b_{ij} = 1$ при $a_i > 0$.

Для произвольного значения $a \hat{I} R^+$ определим матрицу инциденций уровня a , где $A_a = (c_{ij})$, причем $c_{ij} = 0$ при $a_{ij} < a$, $c_{ij} = a_{ij}$ при $a_{ij}^3 a$ и, соответственно, определим скелетную матрицу инциденций уровня a : $\bar{A}_a$. Справедлива следующая теорема.

Теорема 5. *Для произвольного уровня* $a\ \hat{I}\ R^+$ *число систем различных представителей семейства нечетких множеств* $X_1,\ldots.\ X_n$ *уровня, не меньшего, чем a, равно перманенту скелетной матрицы инциденций уровня a.*

Доказательство проводится по той же схеме, что и доказательство теоремы 3.

5.24. Системы различных представителей семейства нечетких множеств которые принадлежат различным классам по отношению эквивалентности

Объединим теперь оба подхода для характеристики систем различных представителей. Пусть дано множество X, на котором заданы отношения эквивалентности R. Пусть имеется n нечетких подмножеств $X_1,\ldots.\ X_n$, которые определены весовыми функциями $m_1\ldots\ldots m_n$, где $m_i: X \to R^+$.

Определение. Набор элементов a_1, ..., a_n будем называть трансверсалью относительно отношения R , если выполнено условия:

1. $m_i(a_i) > 0$ для всех $i = 1, ..., n,$
2. $a_i\, a_j (mod\ R)$ при $i \neq j$

Вопрос о существовании трансверсали семейства нечетких множеств относительно отношения R решает следующее утверждение.

Теорема 6. *Для семейства нечетких множеств X_1,...., X_n и отношения эквивалетности R существует трансверсаль в том и только в том случае, когда выполнены условия: функция $m_y(x)$ для $Y = X_{i1} \dot{E} ... \dot{E} X_{ik}$ отличная от нуля не менее, чем на k классах Y/R для всех $k = 1,....., n$, и всех $1 \in i_1 <.....< i_k \in n$.*

Доказательство проводится по той же схеме, что и доказательство теоремы 1.

Перейдем теперь к вопросу о числе систем различных представителей семейства нечетких множеств и отношения эквивалентности R.

Определим матрицу инциденций P семейства X_1 ,.... X_n и отношения эквивалентности R . Строки матрицы P соответствуют множествам X_1 .... X_n , т.е. функциям m_1 , ... , m_n, а столбцы - классам эквивалентности множества $X = X_1 \dot{E} ... \dot{E} X_n$ по отношению R . Пусть это классы C_1, ..., C_q. Множеству X_i и классу C_j ставим в соответствие число t_{ij} элементов множества X_i, которые лежат в классе C_j и для которых выполнено $m_i(x_j)>0$, $x_j \hat{I} C_j$ (если таких элементов нет - ставим нуль). Справедлива следующая теорема.

Теорема 7. Число *трансверсалій семейства нечетких множеств X_1 .... X_n и отношение эквивалентности R равно перманенту соответствующей матрицы P инциденций системы множеств и отношений R.*

Доказательство аналогично доказательству теоремы 3.

Аналогично предыдущему можно рассмотреть вопрос об учете уровня системы различных представителей и определить число систем различных представителей уровня, не ниже заданной величины $a \hat{I} R^+$. Ясно, что данный вопрос также сводится к вычислению некоторых перманентов неотрицательных матриц.

Таким образом, вопрос о существовании и перечислении систем различных представителей заданного уровня для семейств нечетких множеств также сводится к вычислению перманентов некоторых матриц.

В теории перманентов есть большое количество фактов, которое касается их эффективного вычисления или оценивание.

5.25. Об автоморфизме нечеткого множества

Пусть *(X, m)* - произвольное нечеткое множество, на котором задана группа преобразований *G*. Пусть A_G *(m)* - группа автоморфизмов множества *(X, m)* в группе *G* , т.е.

$$A_G(m) = \{q \in G^{1/2} m(qx) = m(x)\} \text{ для всех } x \in X.$$

Автоморфизмы нечеткого множества сохраняют системы различных представителей произвольных семейств его подмножеств.

Определение. Собственная нетривиальная подгруппа *H* группы *G* называется плотно вложенной, если *H* имеет нетривиальное пересечение с каждой нетривиальной циклической подгруппой группы *G* .

Теорема 8. *Группа автоморфизмов множества (X, m) в группе G нетривиальна в том и только в том случае, если она нетривиальна в нее плотно вложенной подгруппе H.*

Действительно, пусть $A_G(m)^1 e$. Это значит, что существует $q \in G$, $q^1 e$ такое, что $qm = m$. Рассмотрим циклическую группу $< q >$, ($< q >^1 e$ по условию). Очевидно, что $< q > \in A_G(m)$. Поскольку *H* плотно вложена, то $H_G < q >^1 e$. Значит, найдется такое целое *k*, что $q^k \in H$, $q^{k\,1} e$. Очевидно, что $q^k = m$, и тогда имеем $A_H(m)^1 e$. Обратное утверждение очевидно.

Таким образом, в случае, если группа *G* имеет плотно вложенную подгруппу *H,* то при решении вопроса о тривиальности группы автоморфизмов нечеткого множества можно заменить группу *G* группой *H*, которая имеет меньший порядок и, вероятно, более простое строение. Вопрос об описании и нахождение плотно вложенных подгрупп произвольной группы довольно сложен. Мы ограничимся рассмотрением циклических групп.

Теорема 9. *Пусть* $G = < a, a^n = e >$ *- циклическая группа порядка n. Тогда, если n свободно от квадратов простых чисел, то G не имеет плотно вложенных подгрупп. Если n не свободно от квадратов, то группа n* $H = < a^{p1,\ldots,ps} >$ *порядка* $p_1,.., p_s$, *где* $n = p^{a1}{}_1,..., p^{as}{}_s$ *- разложение n на простые множители, будет плотно вложенной подгруппой группы G* .

Если *n* свободно от квадратов простых чисел, то *G* - прямое произведение циклических подгрупп простых порядков. Если *H* плотно вложенная, то *H* содержит элементы простых порядков, а значит и группу, порожденную ими. Значит, *G = H.*

Если *n* не свободно от квадратов, то пусть *b* - элемент простого порядка p_1 , т.е. $b^{p1} = e$. Тогда $b = a^k$ для какого-то *k*, $k < n$, т.е.

выполнен $a^{\kappa}p^{1} = e$ или $k \times p_1$°0 *(mod n).* Отсюда $kp_1 = nq$, q- целое. Следовательно , имеем

$$b = (a^{p}{}_{1})^{q}=(a^{p}{}_{1},..., {}^{p}{}_{s})^{q \times p2..ps}\hat{I}H,$$

т.е. элементы простых порядков входят в группу *H.* Любая нетривиальная подгруппа содержит циклическую подгруппу простого порядка и поэтому *H* имеет не пустое пересечение с любой циклической подгруппой, т.е. *H*— плотно вложенная.

5.26. О свойствах нечеткого единичного куба

Нечетким единичным кубом размерности *n* будем называть пары (E_n, f), где E_n - множество наборов длины *n* из элементов 0,1, а *f* - функция принадлежности элемента $(x_1 , ..., x_n)$ множеству E_n , т.е. $f: E_n$® R^{+}. В сущности, нечеткий единичный куб определяется псевдобулевой функцией $f(x_1,....., x_n)$.

Для задания нечеткого множества (E_n, f) могут быть использованы различные способы представления функции *f.*

Ниже будут использоваться представления псевдобулевих функций в виде действительных многочленов и системы весов. Напомним некоторые определения.

Представление псевдобулевой функции $f(xi,....., x_n)$ действительным многочленом определим индуктивно.

При $n = 1$ функцию $f(x)$ представляет многочлен

$$P_f(x) = (f(1)-f(0))x+f(0).$$

Если для n-1 представляющий многочлен определен, то для n функцию $f(x_1,....., x_n)$ представляет многочлен

$$P_f(x_1 ....., x_n) = (P_f(x_1 , ..., x_{n-1}) - P_f(x_1 , ...,x_{n-1},_0))x_n + P_f(x_1, ..., x_{n-1},_0).$$

Определим вес псевдобулевой функции $f(x_1,....., x_n)$ как действительную сумму

$$||f|| = \mathring{a}f(x_1,.....,x_n)$$
$$(x_1,....,x_n)\hat{I}E_n$$

Для произвольной псевдобулевой функции $f(x_1,....., x_n)$ образуем 2^{n}-1 функций следующим образом:

$$f \times x_1 , f \times x_2 , ..., f x_n , \ f \times x_1 \times x_2 , ... , f \times x_1.... x_n$$

Это легко доказывается.

Теорема 10. *Система* 2^{n} *весов* $\| f \|, \| f \times x_1 \| , ..., \| f \times x_1.... x_n \|$ *однозначно определяет псевдобулеву функцию f.*

Пусть (E_n, f) - нечеткий единичный куб. Пусть $q : E_n$® E_n- некоторая биекция. Рассмотрим вопрос об условиях, когда биекция *q* является

автоморфизмом множества (E_n, f). Ясно, что в этом случае q представляется семейством булевых функций $q\ (q_1,\ldots..,\ q_n)$.

Теорема 11. *Биекция $(q_1,\ldots..,\ q_n)$ является автоморфизмом нечеткого множества (E_n, f) тогда и только тогда, когда выполнено $2^n - 1$ условий.*

$\| f \times q_1,\ldots..,\ q_n\| = \|fx_1.....x_n\|$

$\| f \times q_1,\ldots..,\ q_{n-1}\| = \|fx_1.....x_{n-1}\|$

$\|fq_{i1},\ldots..,\ q_{ik}\| = \|fx_{i1}.....x_{ik}\|$ (*)

$\| f\ q_i\| = \| fx_i\|,\ u = 1,...,n$

Доказательство заключается в непосредственной проверке эквивалентности условий (*) и условия q - автоморфизм поэлементно.

Так, первое равенство в (*) дает

$$f(1..... 1) = f(x_1,.....,x_n),$$

где

$$q_1(x_1,\ldots..,\ x_n) = 1,$$
$$q_n(x_1,\ldots..,\ x_n) = 1.$$

Из следующего равенства получаем

$$f(1..... 1,0)+f(1,.....1,1)=f(x)+f(y),$$

где

$$q_1(x_1.....,\ x_n) = 1,\ q_1\ (y_1,\ .....\ y_n) = 1,$$
$$q_n(x_1,\ldots..,\ x_n) = 1,\ q_n(y_1.....,\ y_n) = 0.$$

Продолжая таким образом, получаем

$$f\ (q^{-1}\ (x)) = f\ (x)\ ''_x\ \hat{I}\ E_n,$$

т.е. - автоморфизм нечеткого множества.

Ясно, что справедливо и обратное утверждение.

Модуль 6.
Алгебра структурных чисел

Микромодуль 21
Введение в структурные числа

6.1. Основные понятия

6.1.1. Определение структурного числа

Пусть X — подмножество абстрактного пространства P. Элементы множества X обозначим

$$\alpha_i,\ \beta_i,\ \gamma_i,\ \ldots \in \mathcal{X}.$$

Рассмотрим систему элементов в виде таблицы

$$A = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ \alpha_{m1} & \alpha_{m2} & \ldots & \alpha_{mn} \end{bmatrix}. \qquad (1.1)$$

Будем рассматривать эту систему как совокупность столбцов a_k, т. е.

$$A = \{a_1,\ a_2,\ \ldots,\ a_n\},\quad a_i \neq a_j\quad (i \neq j). \qquad (1.2)$$

Столбцы a_k в свою очередь представляют собой неупорядоченные множества элементов α_{i_k}.

$$a_k = \{\alpha_{1k},\ \alpha_{2k},\ \ldots,\ \alpha_{m_k k}\},\quad \alpha_{ik} \neq \alpha_{j_k}\quad (i \neq j). \qquad (1.3)$$

Столбцы будем считать равными, если они содержат одинаковые элементы. Положим по определению, что система (1.1) не содержит одинаковых столбцов. Систему типа (1.1) будем рассматривать как элемент новой алгебры — ***алгебры структурных чисел***. Согласно определениям абстрактной алгебры, алгебру структурных чисел можно отнести к категории операторных алгебр, т. е. ее можно характеризовать упорядоченной тройкой

$$\langle E,\ \Omega,\ e\rangle,$$

где E — носитель алгебры (в нашем случае семейство множеств); Ω, — двухэлементное множество операторов ω_1, ω_2, определяющих

сумму и произведение; *e* — результат, т. е. функция, которая выражению $A\omega B$ ставит в соответствие элемент $C \in E$, являющийся результатом действия.

Введем вспомогательное понятие, которое используем при определении структурного числа.

Рассмотрим последовательность элементов x_i, необязательно различных:

$$\langle x_1,\ x_2,\ \ldots,\ x_i,\ \ldots,\ x_n \rangle. \qquad (1.4)$$

Обозначим через $r(x_k)$ — число одинаковых элементов последовательности (1.4).

Структурным числом называется система элементов α_{i_k} вида (1.1) [с учетом (1.2) и (1.3)], удовлетворяющая следующим определениям.

Определение 1.1. Два структурных числа считаются равными $(A = B)$ тогда и только тогда, когда $(a \in A) \Leftrightarrow (a \in B)$ или

$$A = B \Leftrightarrow \bigwedge_a (a \in A \Leftrightarrow a \in B). \qquad (1.5)$$

Определение 1.2. Суммой структурных чисел A и B называется структурное число

$$C = \{x \mid (x \in A) \vee (x \in B),\ x \notin A \cap B\} = A \triangleq B; \qquad (1.6)$$

в этом случае можно написать $C = A + B$.

Выражение $A \triangleq B$ в формуле (1.6) означает симметричную разность множеств A и B.

Определение 1.3. Произведением структурных чисел A и B называется структурное число

$$C = \{a \cup b \mid a \cap b = \phi,\ r(a \cup b) \in \{1, 3, \ldots,\}\ a \in A,\ b \in B\}, \qquad (1.7)$$

которое записывается в виде $C = AB$.

В соответствии с определением суммы при сложении структурных чисел опускаются столбцы, одновременно присутствующие в обоих числах A и B, а в соответствии с определением произведения при умножении структурных чисел A и B опускаются те столбцы $a \cup b$, в которых какой-либо элемент повторяется, т. е. для которых $a \cap b \neq \phi$, а также опускается четное число идентичных столбцов.

Можно заметить, что равенство структурных чисел представляет собой отношение эквивалентности, т. е. является рефлексивным, симметричным и транзитивным.

Далее будут приведены примеры действий со структурными числами, элементы которых $\alpha_{i_k} \in X$ представляют собой натуральные числа (этот случай имеет большое значение для применения алгебры структурных чисел), а также даны словесные формулировки действий со структурными числами, которые менее точны, чем вышеприведенные, однако более понятны для читателей, не имеющих достаточной математической подготовки.

Пример 1.1. Равенство структурных чисел:

$$\begin{bmatrix} 1 & 1 & 2 \\ 3 & 2 & 5 \end{bmatrix} = \begin{bmatrix} 5 & 1 & 1 \\ 2 & 3 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 2 & 3 \\ 1 & 5 & 1 \end{bmatrix}.$$

Два структурных числа равны, если содержат идентичные столбцы, независимо от порядка элементов в столбцах и порядка столбцов.

Пример 1.2. Сложение структурных чисел:

$$\begin{bmatrix} 2 & 3 & 4 \\ 7 & 5 & 7 \end{bmatrix} + \begin{bmatrix} 3 & 2 & 5 \\ & 7 & 4 \end{bmatrix} = \begin{bmatrix} 2 & 3 & 4 & 3 & 2 & 5 \\ 7 & 5 & 7 & & 7 & 4 \end{bmatrix} = \begin{bmatrix} 3 & 4 & 3 & 5 \\ 5 & 7 & & 4 \end{bmatrix}.$$

Суммой двух структурных чисел *A* и *B* называется структурное число, содержащее все столбцы чисел *A* и *B,* за исключением идентичных столбцов, и не содержащее других столбцов.

Пример 1.3. Умножение структурных чисел:

$$\begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 3 \end{bmatrix} \begin{bmatrix} 3 & 2 & 2 \\ & & 3 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 2 \\ 2 & 3 & 1 \\ 3 & 4 & 3 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}.$$

Произведением двух структурных чисел *A* и *B* называется структурное число, столбцы которого представляют собой суммы (согласно понятиям теории множеств) всех возможных комбинаций столбцов *A* и *B,* за исключением наибольшего четного числа идентичных столбцов и таких столбцов, в которых какой-либо элемент повторяется (произведение других столбцов не содержит).

Из определения суммы и произведения структурных чисел следует, что эти операции всегда можно выполнить на множестве этих чисел. Из тех же определений можно сделать вывод, что сложение и умножение структурных чисел коммутативны и ассоциативны, а умножение дистрибутивно относительно сложения.

Для трех произвольных структурных чисел имеют место следующие соотношения, подобные тем, которые справедливы для элементарной алгебры:

$$A + B = B + A,$$
$$AB = BA,$$
$$A(BC) = (AB)C,$$
$$A(B + C) = AB + AC. \qquad (1.8)$$

Следует различать структурное число $[\phi]$, содержащее один столбец, который является пустым множеством $\varnothing$, и структурное число $[\ ]$, не содержащее ни одного столбца.

Заметим, что число [ ] служит модулем суммирования и для произвольного структурного числа A выполняется равенство

$$A + [\ ] = A,$$

поэтому число [ ] будем обозначать символом 0, записывая его в виде

$$[\ ] = 0. \qquad (1.9)$$

Число $[\phi]$ в свою очередь есть модуль умножения, так как

$$A[\phi] = A, \qquad (1.10)$$

поэтому число $[\phi]$ обозначим символом 1, записав

$$[\phi] = 1. \qquad (1.11)$$

Для любого A имеет место соотношение

$$A[\ ] = [\ ].$$

Рассмотрим структурное число вида

$$A = \{\phi, a_1, a_2, \ldots, a_\lambda\}, \qquad (1.12)$$

т. е. число, содержащее один пустой столбец.

Легко заметить, что для такого числа справедливо равенство

$$AA = 1.$$

Для структурных чисел, не содержащих пустого столбца,

$$AA = 0.$$

Если множество структурных чисел вида (1.12) обозначить как A, а множество всех остальных структурных чисел — как B, то можно написать

$$(A \in \mathscr{A}) \Rightarrow AA = 1,$$
$$(A \in \mathscr{B}) \Rightarrow AA = 0. \qquad (1.13)$$

Следовательно, легко заметить, что для произвольного структурного числа

$$\underbrace{A + A + \ldots + A}_{n \text{ раз}} = \begin{cases} A \text{ при нечетном числе слагаемых } n, \\ 0 \text{ при четном } n. \end{cases} \tag{1.13а}$$

Таким образом,

$$A + A = 0. \tag{1.13б}$$

Из соотношения (1.13) вытекает, что равенство

$$AB = 0$$

но требует в общем случае равенств $A = 0$ или $B=0$, т. е. множество структурных чисел содержит делители нуля. Пару структурных чисел, для которой выполняется равенство $AB = 0$, назовем *особой парой.*

Пример 1.4. Особую пару представляют собой следующие структурные числа A и B:

$$A = \begin{bmatrix} 1 & 3 \\ 2 & 2 \end{bmatrix}; \quad B = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \text{ так как } \begin{bmatrix} 1 & 3 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = 0.$$

Естественно, число $[\ ] = 0$ в сочетании с любым структурным числом дает особую пару.

Обобщая изложенные свойства структурных чисел, можно сформулировать следующие теоремы.

Теорема 1.1. Множество структурных чисел, на котором определены операции сложения и умножения, образует коммутативное кольцо. Это кольцо обычно содержит делители нуля. Из определения суммы и произведения следуют соотношения, справедливые для любого структурного числа:

$$\begin{bmatrix} \alpha_{11} & \ldots & \alpha_{1n} \\ \vdots & & \vdots \\ \alpha_{m1} & \ldots & \alpha_{m_n n} \end{bmatrix} = \sum_{k=1}^{n} \begin{bmatrix} \alpha_{1k} \\ \vdots \\ \alpha_{m_k k} \end{bmatrix},$$

$$\begin{bmatrix} \alpha_{1k} \\ \vdots \\ \alpha_{m_k k} \end{bmatrix} = \prod_{i=1}^{m_k} [\alpha_{ik}] = \prod_{i=1}^{m_k} s_{\alpha_{ik}}, \tag{1.14}$$

где $s_{i_k} = \left[\alpha_{i_k} \right]$ — одноэлементное структурное число.

Теорема 1.2. Структурное число A всегда можно представить и миде

$$A = \sum_{k} \prod_{i} s_{\alpha_{ik}}, \qquad (1.15)$$

где $s_{\alpha_{ik}} = \left[\alpha_{i_k} \right]$

Следует отметить, что выражение (1.15) в алгебре структурных чисел играет роль, аналогичную выражению $z = a + ib$ в теории функций комплексного переменного, с помощью которой можно записать любое комплексное число $z = \langle a, b \rangle$.

Структурное число

$$s_{\alpha_{ik}} = [\alpha_{i_k}]$$

называется *структурной единицей*, которая служит аналогом действительной или мнимой единицы в области комплексных чисел.

6.1.2. Вычитание структурных чисел

Рассмотрим два произвольных структурных числа A и B. Из определения равенства и суммы структурных чисел следует, что существует только одно структурное число, удовлетворяющее равенству

$$B + X = A, \qquad (1.16)$$

которое вследствие коммутативности суммирования структурных чисел можно переписать как

$$X + B = A, \qquad (1.16а)$$

Структурное число X, удовлетворяющее равенствам (1.16) и (1.16а), называется *разностью* структурных чисел A и B:

$$X = A - B.$$

Действие нахождения разности структурных чисел называется вычитанием. Легко заметить, что разность чисел A и B есть число $X = A+B$. Действительно, подставляя в выражение (1.16) $X = A + B$, получаем уравнение

$$B + (A + B) = A,$$

которое в соответствии с (1.13*б*) представляет собой тождество. Таким образом, получаем обоснованное соотношение

$$A - B = A + B, \qquad (1.17)$$

которое в случае $A = 0$ записывается в виде

$$-B = B. \qquad (1.18)$$

Из сказанного следует, что на множестве структурных чисел вычитание всегда можно заменить сложением. Вычитание, следо-

вательно, определено однозначно и всегда, выполнимо, поэтому множество структурных чисел замкнуто по отношению к суммированию и вычитанию.

Подводя итог рассмотренным свойствам структурных чисел, можно заключить, что кольцо структурных чисел 1) не содержит степеней и 2) не содержит коэффициентов (кроме 0 и 1); а 3) сложение идентично вычитанию.

6.2. Свойства структурных чисел

6.2.1. Делители нуля

Пусть **A*** — множество структурных чисел X, удовлетворяющее уравнению

$$AX = 0,$$

где A — некоторое структурное число, и пусть A^* — элементы этого множества. Тогда

$$AX = 0 \Rightarrow X = A_i^* \in \mathbf{A}^*. \qquad (1.19)$$

Числа $A^* \in \mathbf{A}^*$, удовлетворяющие уравнению $AX = 0$, называются сопряженными по отношению к A или *делителями нуля.*

Следствие. Если два структурных числа X_1 и X_2 удовлетворяют равенству $AX = 0$, то такому же равенству удовлетворяет их линейная комбинация $C_1X_1 + C_2X_2$, а также произведение CX_1X_2, где C_1, C_2, C — произвольные структурные числа, включая 0 и 1. Тогда

$$X_1,\ X_2 \in \mathbf{A}^* \Rightarrow C_1X_1 + C_2X_2 \in \mathbf{A}^*;\ CX_1X_2 \in \mathbf{A}^*. \qquad (1.20)$$

Обоснование этого положения элементарно и предлагается выполнить читателю.

Полагая в выражении (1.20) $C_1 = C_2 = C = 1$, приходим к выводу, что к **A*** относятся сумма $X_1 + X_2$ и произведение X_1X_2 структурных чисел X_1 и X_2, удовлетворяющих уравнению $AX = 0$.

Множество **A*** решений уравнения $AX = 0$ можно в общем случае определить с помощью выражения, записанного в символах математической логики:

$$(AX = 0) \Leftrightarrow \bigwedge_{\substack{a \in A \\ x \in X}} \{[a \cap x \neq \phi] \vee [r(a \cup x) \in \{0, 2, \ldots\}\}. \qquad (1.21)$$

Свойство (1.21) следует непосредственно из определения произведения структурных чисел.

6.2.2. Делимость структурных чисел

Если для двух структурных чисел A и B существует такое число X, что

$$A = XB, \tag{1.22}$$

то A делится на B, или B — делитель числа A, т. е.

$$B \mid A \quad \text{и} \quad A \neq 0. \tag{1.23}$$

Очевидно, каждое структурное число $A \neq 1$ и $A \neq 0$ имеет самое малое два делителя, а именно 1 и A; число 1, в свою очередь, имеет лишь один кратный делитель.

Структурные числа $A \ni \not\varnothing$, содержащие только один делитель A, называются простыми числами; любое другое структурное число называется сложным. Каждый делитель, представляющий собой однострочное структурное число, называется основным делителем.

Теорема 1.3. Структурное число B представляет собой делитель структурного числа A тогда и только тогда, когда оно удовлетворяет следующим условиям:

1) $AB = 0$;

2) все столбцы числа A являются подмножествами некоторых столбцов числа B.

Доказательство. Если число B есть делитель числа A, то существует такое X, что

$$BX = A.$$

Но это уравнение имеет решение тогда и только тогда, когда все столбцы числа A представляют собой подмножества некоторых столбцов числа B, а B — элемент, сопряженный с A. Следовательно,

$$B \mid A \Leftrightarrow (AB = 0) \wedge \Big[\bigwedge_{b_k \in B} \bigvee_{a_k \in A} (a_k \not\supset b_k)\Big]. \tag{1.24}$$

Следует заметить, что деление, определенное на множестве структурных чисел, обладает свойством

$$A \mid B \quad \text{и} \quad B \mid C \Rightarrow A \mid C. \tag{1.25}$$

Деление также представляет собой слабо симметричное отношение, т. е. $(A \mid B$ и $B \mid A) \Rightarrow A = B$, что вытекает из следующей теоремы.

Теорема 1.4. Если структурное число B — делитель структурного числа A, а число A — делитель B, то $A = B$.

Доказательство. Положим, что одновременно имеет место

$$B \mid A \quad \text{и} \quad A \mid B.$$

Из теоремы 1.3 следует, что тогда могут быть одновременно выполнены условия

$$\Big\{\bigwedge_{a_k\in A}\bigvee_{b_k\in B}(a_k \supset b_k)\Big\}\wedge\Big\{\bigwedge_{a_k\in A}\bigvee_{b_k\in B}(b_k \supset a_k)\Big\},$$

что может иметь место только при $A = B$.

Для структурных чисел имеет место правило сокращения, т. е. если $CA = CB$, то $A = B$. Это положение можно обосновать.

Теорема 1.5. Уравнение $AB=AX$ имеет общее решение на множестве структурных чисел

$$X = B + A^*,$$

где A^* — произвольный сопряженный элемент A. Тогда

$$(AB = AX) \Leftrightarrow (X = B + A^*;\ A^* \in \mathbf{A^*}). \quad (1.26)$$

Доказательство. Из уравнения $AB = AX$ следует, что $A\,(B + X) = 0$ и $B + X$ — число, сопряженное с A, а соответственно и $X = B + A^*$, где A^* — произвольный элемент множества решений уравнения $AX = 0$. Подставляя число $X — B + A^*$ в уравнение $AB = AX$, убеждаемся, что это число действительно удовлетворяет данному уравнению.

Теорема 1.6. Каждое сложное структурное число имеет по крайней мере один делитель, представляющий собой простое число, не равное единице.

Доказательство. В соответствии с определением сложное число A имеет делители, отличные от 1 и A. Положим в таком случае, что B — один из этих делителей, т. е.

$$A = X_0B, \quad X_0 \neq 1, \quad B \neq 1. \quad (1.27)$$

Если B — непростое число, то его можно представить как $B = X_1B_1$. При этом получим

$$A = B_lX_0X_1X_2 \ldots X_l. \quad (1.28)$$

Но для $A \neq 0$ должно выполняться очевидное неравенство

$$l \leqslant m_A, \quad (1.29)$$

где m_A — число элементов в столбце числа A, содержащем наименьшее количество элементов. Множество натуральных чисел $\{1, 2, \ldots, l\}$ имеет наибольший элемент $l_{макс}$, поэтому $l_{макс}$ есть простой делитель числа A. Для $A = 0$ неравенство (1.29) не должно выполняться, но, согласно изложенному, из произвольного делителя числа A (сложного, ненулевого) можно извлечь простой делитель, что и доказывает теорему 1.6.

Теорема 1.7. Каждое структурное число представляет собой простое число или произведение простых чисел.

Правильность этого положения следует из теоремы 1.6. Действительно, если структурное число A можно представить в виде (1.28) с простым делителем l, то на простые делители можно разложить каждый дополнительный делитель X_0, X_1, X_2, ... X_l. Тогда структурное число можно всегда представить в виде произведения простых чисел

$$A = P_1 P_2 \ldots P_r. \tag{1.30}$$

В случае, когда A само будет простым числом (не равным единице), произведение сводится к одному сомножителю. Разложение числа A на простые числа запишем тогда в следующем виде:

$$\begin{matrix} P_1 \\ P_2 \\ \vdots \\ P_r \end{matrix} \Bigg| A \tag{1.31}$$

Нетрудно заметить, что структурные числа имеют следующие свойства:

1. Любое структурное число, состоящее из разных (неповторяющихся) элементов и содержащее более одного столбца, есть простое число.
2. Каждое структурное число, состоящее из одной строки, простое.
3. Каждое структурное число, состоящее из одного столбца, сложное ($n > 1$).
4. Сумма простых чисел может быть сложным числом, сумма сложных чисел может быть простым числом.

Пример 1.5.

$$\begin{bmatrix} \alpha & \alpha & \alpha_1 \\ \beta_1 & \beta_2 & \end{bmatrix} + [\alpha_1] = \begin{bmatrix} \alpha & \alpha \\ \beta_1 & \beta_2 \end{bmatrix} = [\alpha]\,[\beta_1 \ \beta_2],$$

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} \alpha_1 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \alpha & \alpha_1 \\ \beta & \beta_1 \end{bmatrix}.$$

В данном примере, суммируя вначале два простых структурных числа, получаем сложное число; затем, суммируя сложное число, получаем простое число. Следствием второго свойства структурных чисел является то, что множество простых структурных чисел бесконечно, если бесконечно множество X, из которого взяты элементы структурных чисел. Оказывается, разложение структурных чисел на простые имеет специфические особенности, отличные,

например, от особенностей разложения в области натуральных чисел. Одна из этих особенностей рассматривается в следующей теореме.

Теорема 1.8. Каждое сложное число имеет бесконечное множество способов разложения на простые числа.

Доказательство. Положим, что $A = P_1 P_2 \ldots P_r$. Легко заметить, что величина этого произведения не изменится, если любое из чисел P_i дополнить столбцами, содержащими некоторые элементы всех столбцов одного из оставшихся сомножителей.

Так как число таких возможных дополнений бесконечно, то каждое сложное структурное число можно разложить на простые бесконечно большим числом способов.

Пример 1.6.

$$[1]\,[2] = [1]\,[2\ \ 1] = [1]\begin{bmatrix} 2 & 1 \\ & 2 \end{bmatrix} = [1]\begin{bmatrix} 2 & 1 \\ & 3 \end{bmatrix} = \ldots .$$

Рассмотрим свойства структурных чисел с одинаковым числом элементов в строках и одинаковым числом элементов в столбцах, которые представляют наибольший интерес для применения алгебры структурных чисел.

Определение 1.4. Разложение структурного числа с одинаковым числом элементов в строках на простые числа также с одинаковым числом элементов в строках, содержащих только элементы числа *A,* называется *каноническим разложением.*

Очевидно, что каждое структурное сложное число с равным числом элементов в строках имеет конечное число канонических разложений. Это имеет большое практическое значение, например, в случае применения алгебры структурных чисел к синтезу электрических цепей.

Теорема 1.9. Если структурное число с одинаковым числом элементов в строках $A \neq 0$ (с m строками) имеет каноническое разложение

$$A = \prod_{i=1}^{m} P_i, \tag{1.32}$$

то все остальные канонические разложения числа A на простые однострочные числа имеют вид

$$A = \prod_{j=1}^{n} \sum_{i=1}^{m} \varepsilon_{ij} P_i, \tag{1.33}$$

где числа ε_{ij} принимают только значения 0 или 1.

Доказательство. Положим, существует разложение

$$A = \prod_{j=1}^{m} P'_j,$$

отличное от (1.32). Тогда, перемножая A и любое P'_j, получим

$$AP'_j = P'_j \prod^{m} P_i = 0.$$

Далее, $P'_j = (P_1 P_2 \ldots P_m)^*_j$, т. е. P'_j — сопряженный элемент по отношению к $\prod_{i=1}^{m} P_i$. Однако можно заметить, что в классе однострочных структурных чисел P_i справедливо соотношение

$$(P_1 P_2 \ldots P_m)^*_j = \sum_{i=1}^{m} \varepsilon_{ij} P_i, \qquad \varepsilon_{ij} = 0, 1. \qquad (1.34)$$

Тогда, действительно,

$$A = \prod_{j=1}^{m} \sum_{i=1}^{m} \varepsilon_{ij} P_i, \qquad \varepsilon_{ij} = \begin{cases} 0 \\ 1. \end{cases}$$

Непосредственно из теоремы 1.9 следует, что двустрочное сложное структурное число имеет лишь три возможных канонических разложения на однострочные числа

$$A = P_1 P_2 = P_1 (P_1 + P_2) = P_2 (P_1 + P_2). \qquad (1.35)$$

Несмотря на то что известен общий вид канонического разложения, определение общего числа возможных канонических разложений m-строчного сложного структурного числа — довольно трудная комбинационная задача.

Теорема 1.10. Число возможных канонических разложений отличного от нуля сложного структурного числа с m строками на однострочные сомножители удовлетворяет неравенству

$$R_m < 1 + k + \frac{k(k-1)}{2!} + \frac{k(k-1)(k-2)}{3!} + \ldots + \frac{k(k-1)(k-2)\ldots m}{(m-1)^2!}, \qquad (1.36)$$

где $k = m(m-1)$.

Доказательство. Если структурное m-строчное число не имеет вовсе разложения на однострочные сомножители, неравенство (1.36) полностью удовлетворяется. В случае когда имеется возможное разложение, оно имеет вид (1.33).

На основе этого разложения можно выделить $(m-1)^2 + 1$ множеств разложений числа A, ставя в соответствие каждому отдельному множеству те разложения, которые имеют одинаковое число нулевых чисел ε_{ij}. В общем в выражении (1.33) имеем m^2 чисел ε_{ij}, причем

нулевое значение может одновременно принимать не более чем $k = m^2 - m = m(m-1)$ чисел ε_{ij}. С другой стороны, нулевое значение должно иметь минимум $m - 1$ чисел ε_{ij}, так как в противном случае в разложении (1.33) имелись бы идентичные сомножители и $A = 0$. Обозначая через e число нулевых значений ε_{ij}, имеем неравенство

$$m - 1 \leqslant e \leqslant m(m-1) = k.$$

Таким образом, Δe выделенных множеств разложений числа A равно

$$\Delta e = (m-1)m - (m-1) + 1 = (m-1)^2 + 1.$$

Если через R^e_m обозначить число разложений с e нулевыми числами ε_{ij}, то с помощью простых комбинаций можно определить следующие зависимости:

$$R_m^k = 1, \quad R_m^{k-1} = k,$$
$$R_m^{k-2} = \frac{k(k-2)}{2!} < \frac{k(k-1)}{2!},$$
$$R_m^{k-3} < \frac{k(k-1)(k-2)}{3!},$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$R_m^{m-1} = R^{k-(m-1)^2} < \frac{k(k-1)\ldots m}{(m-1)^2!},$$

а так как $R_m = R_m^k + R_m^{k-1} + \ldots + R_m^{m-1}$, то неравенство (1.36) действительно выполняется. Рассмотрев зависимость

$$1 + k + \frac{k(k-1)}{2!} + \ldots + \frac{k(k-1)(k-2)\ldots 1}{k!} = 2^k, \quad (1.37)$$

нетрудно заметить, что можно оценить число возможных разложений m-строчного структурного числа на однострочные множители по простой формуле

$$R_m < 2^{m(m-1)}, \quad m > 1. \quad (1.38)$$

Однако эта оценка дает худшие результаты по сравнению с неравенством (1.36).

6.3. Геометрическое изображение структурного числа

До сих пор мы рассматривали структурные числа как элементы переменного кольца и его общие свойства, исходя из определения

действий сложения, умножения и т. д. Попробуем дать геометрическую интерпретацию структурного числа. Следует отметить, что геометрическая интерпретация встречается также и в других случаях, например в случае комплексных чисел, которым ставятся в соответствие некоторые точки плоскости Гаусса.

Геометризация структурного числа имеет значение прежде всего для его применения при анализе и синтезе электрических цепей.

Определение 2.5. Если столбцы структурного числа A взаимно однозначно соответствуют деревьям графа Γ так, что каждый столбец представляет собой множество значений описывающей функции соответствующего дерева, то граф Γ называется геометрическим изображением числа A и записывается в виде

$$\Gamma = \mathrm{ob}\,(A). \qquad (1.39)$$

Следовательно, геометрическим изображением структурного числа A служит любой ***детерминированный граф***, удовлетворяющий условию (1.39), или класс графов подобных структур. Из принятого определения следует, что геометрическое изображение структурного числа — не однозначное понятие, так как структурному числу может соответствовать многоэлементное семейство графов, составляющих класс с подобной структурой. Однако это в известном смысле является достоинством метода, так как становится возможным, например в задачах синтеза цепей, нахождение не одного, а множества вариантов цепи, удовлетворящей заданным условиям.

Не каждое структурное число изображается ***связным графом — топологической цепью***. Определение условий, при которых существует изображение структурного числа в виде связного графа, имеет принципиальное значение для применения метода структурных чисел. Эти условия будут сформулированы в теореме 1.12.

Теорема 1.11. Структурное число A с одинаковым числом элементов в строках, геометрическим изображением которого служит связный граф с вершинами $p_1, p_2, \ldots, p_n$, равно произведению n — 1 простых однострочных сомножителей

$$A = P_1 P_2 \ldots P_{n-1}, \qquad (1.40)$$

причем сомножители состоят из значений описывающей функции ребер, инцидентных произвольно выбранной вершине p_i ($p_i \neq p_j$, если $i \neq j$) графа Γ.

Доказательство. Равенство (1.40), очевидно, справедливо в случае графов с одной и двумя вершинами. Рассмотрим произвольный связный граф с n вершинами. Соединим в нем две произвольные вершины ребром α_k. Положим, что выражение (1.40) справедливо для

образованного таким образом графа Γ^*, т. е. что структурное число равно

$$A^* = (P_1 + P_2)\, P_3 \ldots P_{n-1} = (P_1' + P_2')\, A^{**},$$

где

$$P_1 = P_1' + [\alpha_k], \quad P_2 = P_2' + [\alpha_k].$$

Найдем число

$$A' = P_1 P_2 A^{**} = (P_1' + P_2')\, A^{**}\,[\alpha_k] + P_1' P_2' A^{**} = A^*\,[\alpha_k] + A''. \quad (1)$$

Структурное число A графа Γ с несоединенными вершинами и висячим ребром α_k можно представить в виде

$$A = A^*\,[\alpha_k] + A^0, \quad (2)$$

поскольку множество деревьев графа Γ^* (с замкнутым ребром α_k), дополненное элементом α_k, представляет собой множество всех деревьев графа Γ с ребром α_k. В выражении (2) символ A^0 обозначает структурное число, составленное из всех столбцов числа A, не содержащих α_k. Просуммировав равенства (1) и (2), получим

$$A + A' = A^0 + A''. \quad (3)$$

Так как левая часть этого равенства не зависит от выбора ребра α_k, а правая часть не содержит α_k, значит, для каждого из следующих уравнений имеем

$$A + A_1' = A_1^0 + A_1'' \not\ni \alpha_1,$$
$$A + A_2' = A_2^0 + A_2'' \not\ni \alpha_2,$$
$$\ldots\ldots\ldots\ldots$$
$$A + A_n' = A_n^0 + A_n'' \not\ni \alpha_n.$$

Отсюда левая часть равенства (3) не содержит обозначений ребра графа Γ и равна нулю. Тогда

$$A = A' = P_1 P_2 \ldots P_{n-1},$$

что и требовалось доказать.

Сформулируем теорему об условиях, при которых структурное число имеет связное изображение.

Теорема 1.12. Необходимые и достаточные условия существования геометрического изображения структурного числа в виде связного графа состоят в том, чтобы структурное число A имело разложение на простые однострочные сомножители

$$A = P_1 P_2 \ldots P_m, \quad (1.41)$$

причем произвольный элемент α_{ik}, должен встречаться самое большее в двух простых числах P_i, P_j.

Доказательство. Разложение (1.41) непосредственно следует из теоремы 1.11 и не требует специального обоснования. Условие того, что элемент α_{ik} встречается максимум в двух числах P_i, P_j, тоже очевидно, так как в графе имеют место лишь ребра с двумя концами (одномерные симплексы).

В задачах синтеза при определении алгоритма образования структурных чисел на цифровой машине удобно добавить к приведенным условиям следующие дополнительные условия, подтверждающие отличие структурного числа от нуля:

1) в произведении $A = P_1P_2 \ldots P_m$ не может быть одинаковых сомножителей, т. е.

$$P_i \neq P_j, \quad i, j = 1, 2, \ldots, m \quad (i \neq j); \qquad (1.42)$$

2) любой сомножитель P_k произведения (1.41) не может быть равен сумме произвольного числа остальных сомножителей, т. е.

$$P_i \neq \sum_k P_k, \qquad k = 1, 2, \ldots, m \qquad (k \neq i). \qquad (1.43)$$

Из теоремы 1.12 следует, что структурное число, у которого число элементов в строках различно, не имеет связного геометрического изображения. Условие имеет не только теоретическое значение. Оно однозначно условию физического соответствия матрицы полных проводимостей и пассивной электрической цепи.

По сравнению с другими способами определения условий реализации матрицы полных проводимостей определение, основанное на теории структурных чисел, особенно просто и логично.

6.4. Дополнительное структурное число и геометрическое обратное изображение

Определение 1.6. Дополнительным структурным числом для данного структурного числа A называется структурное число A^d, столбцы которого представляют собой дополнения столбцов числа A до множества элементов α_{ik}, из которых состоит структурное число A.

Если обозначить множество элементов α_{ik}, из которых состоит число A, через L, то столбцы C^d_i числа A^d определим как разность (в смысле понятий алгебры множеств)

$$C^d_1 = L - C_2, \; C^d_2 = L - C_2, \; \ldots, \; C^d_n = L - C_n, \qquad (1.44)$$

где $C_1, C_2, \ldots, C_n$ — столбцы числа A.

Дополнительное структурное число можно в таком случае записать в виде

$$A^d = \{b_k \mid (b_k = L - a_k) \Delta (a_k \in A)\} \qquad (1.44а)$$

или иначе

$$A^d = \{\{\alpha_{ik}\} - a_{pk} \mid \alpha_{ik} \in L, \quad a_k \in A\}. \qquad (1.44б)$$

Следует отметить справедливость такого свойства

$$(A + B)^d = A^d + B^d, \quad L = L_A \cup L_B, \qquad (1.45)$$

которое означает, что дополнение — операция аддитивная. Дополнительное структурное число можно также определить по отношению к другому множеству L^*, такому, что $L \subset L^*$, и тогда

$$A^d_{L^*} = \{L^* - a_k \mid a_k \in A\}. \qquad (1.44в)$$

Способ получения дополнительного структурного числа иллюстрирует следующий пример.

Пример 1.7. Определить структурное число A^d по отношению к структурному числу

$$A = \begin{bmatrix} 2 & 6 & 9 \\ 3 & 2 & 8 \\ 5 & 8 & 3 \end{bmatrix}.$$

Множество элементов числа L таково:

$$L = \{2, 3, 5, 6, 8, 9\}.$$

Дополнительное структурное число равно

$$A^d = \begin{bmatrix} 6 & 3 & 2 \\ 8 & 5 & 5 \\ 9 & 9 & 6 \end{bmatrix}.$$

Оказывается, что для структурного числа удобно иметь дуальное геометрическое изображение, поэтому введем понятие обратного изображения геометрического структурного числа.

Определение 1.7. Граф Γ называется обратным изображением структурного числа A, если столбцы числа A взаимно однозначно соответствуют дополнениям деревьев графа Γ так, что столбец числа A представляет собой множество значений описывающей функции соответствующего дополнения дерева. Тогда напишем

$$\Gamma = \text{cob}\ (A). \qquad (1.46)$$

Нетрудно заметить, что обратное изображение дополнительного числа A^d одновременно служит изображением числа A и наоборот.

Обратное изображение — это граф дуальной структуры в понимании Кауэра по отношению к геометрическому изображению данного структурного числа. Связное обратное изображение существует для любого структурного числа, имеющего связное изображение.

Таким образом, структурному числу ставится в соответствие пара графов дуальной структуры. Один из них служит геометрическим изображением, другой — обратным изображением. Примеры изображений простейших структурных чисел приведены в конце настоящего модуля.

Для обратного изображения имеет место следующая теорема.

Теорема 1.13. Структурное число *A* с одинаковым числом элементов в строках, геометрическое обратное изображение которого суть связный граф *Г,* характеризующийся цикломатическим числом *m,* равняется произведению *m* простых однострочных сомножителей

$$A = P_1 P_2 \ldots P_m,$$

соответствующих линейно независимым контурам графа *G.*

Доказательство. Докажем эту теорему методом индукции.

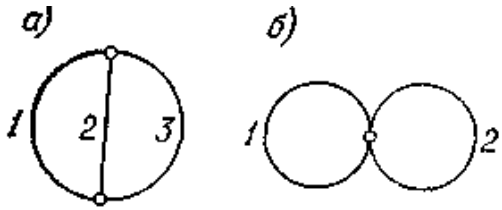


Рис. 1.1. Граф с двумя циклами.

Теорема справедлива для графа с одним и двумя контурами. Действительно, такой граф всегда может быть упрощен и приведен к виду, показанному на рис. 1.1, *а* или 1.1, *б,* где ребра *1, 2, 3* — суммы соответствующих ребер графа с двумя контурами.

Для графа (рис. 1.1, *а)* имеем

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 3 & 3 \end{bmatrix},$$

т. е. действительно $A = [1\ 2]\ [1\ 3]$.

Для случая, изображенного на рис. 1.1, *б,* теорема также справедлива, так как

$$A = \begin{bmatrix} 1 \\ 2 \end{bmatrix} = [1]\ [2].$$

Можно доказать, что теорема справедлива и тогда, когда ребра *1, 2, 3* заменены последовательным соединением произвольного числа ребер.

Положим, что теорема справедлива для графа с цикломатическим числом m — 1. Тогда можно доказать, что она справедлива и для графа с числом контуров m.

Таким образом, теорема справедлива для графов с произвольным числом независимых контуров и произвольной структурой.

Теоремы 1.11 и 1.13 особенно важны для применения метода структурных чисел к анализу электрических цепей. Они служат основой расчета структурных чисел, соответствующих заданным графам, представляющим структуру рассматриваемой цепи.

6.5. Алгебраическая производная и обратная производная структурного числа

На множестве структурных чисел можно определить различные операции; одна из них — операция алгебраической производной.

Определение 1.8. Алгебраической производной структурного числа называется число $\partial A/\partial\alpha$, определенное как

$$\frac{\partial A}{\partial \alpha} = A \,\Big|\, \begin{matrix}\text{столбцы, не содержащие}\\ \text{элемент } \alpha \text{ исключены.}\end{matrix} \tag{1.47}$$

Если структурное число представить как совокупность множеств, то производная

$$\frac{\partial A}{\partial \alpha} = \{b_k \mid b_k = a_k - \{\alpha\},\ \alpha \in a_k,\ a_k \in A\}. \tag{1.47а}$$

Легко доказать правильность следующих зависимостей, аналогичных «обычной» производной:

$$\begin{aligned}\frac{\partial}{\partial \alpha}(A_1 + A_2) &= \frac{\partial A_1}{\partial \alpha} + \frac{\partial A_2}{\partial \alpha},\\ \frac{\partial}{\partial \alpha}(A_1 A_2) &= \frac{\partial A_1}{\partial \alpha} A_2 + \frac{\partial A_2}{\partial \alpha} A_1.\end{aligned} \tag{1.48}$$

Алгебраическую производную обозначим как A_α, т. е.

$$\frac{\partial A}{\partial \alpha} = A_\alpha. \tag{1.49}$$

Следует заметить, что для одноэлементного структурного числа

$$\frac{\partial}{\partial \alpha}[\alpha] = 1. \tag{1.50}$$

Пример 1.8. Нахождение алгебраической производной структурного числа:

$$A=\begin{bmatrix}1&1&5&4\\2&3&3&3\\4&2&7&1\end{bmatrix},\quad \frac{\partial A}{\partial 1}=\begin{bmatrix}2&3&4\\4&2&3\end{bmatrix},\quad \frac{\partial A}{\partial 4}\begin{bmatrix}1&3\\2&1\end{bmatrix}.$$

По аналогии с математическим анализом нахождение производной будем называть дифференцированием.

Дифференцирование структурного числа имеет весьма про-. стую геометрическую интерпретацию, сформулированную ниже. *Свойство 1.* Геометрическое изображение структурного числа $\partial A/\partial\alpha$ представляет собой геометрическое изображение структурного числа A с замкнутым ребром α.

Свойство 1 обосновано теоремами 1.11 и 1.13. Действительно, если положить, что опорным узлом служит любой узел цепи, неинцидентный с ребром α, то элемент α будет встречаться в двух простых сомножителях P_1 и P_2, т. е.

$$A = P_1(\alpha)\, P_2(\alpha)\, P_3 \ldots P_{n-1},$$

где n — число вершин графа. Отсюда

$$\frac{\partial A}{\partial \alpha}=\frac{\partial P_1}{\partial \alpha}P_2\ldots P_{n-1}+\frac{\partial P_2}{\partial \alpha}P_1P_3\ldots P_{n-1}.$$

Так как для однострочных простых чисел P_1 и P_2 справедливо, что $\partial P_1/\partial\alpha = \partial P_2/\partial\alpha = 1$, то

$$\frac{\partial A}{\partial \alpha}=(P_1+P_2)P_3\ldots P_{n-1}. \qquad (1.51)$$

Это означает замыкание ребра α в геометрическом изображении или отключение (или однополюсное отключение) ребра α в обратном геометрическом изображении (тогда $n - 1 = m$ — цикло-матическое число графа). Поскольку величина структурного числа не зависит от выбора опорного узла, полученный результат носит общий характер.

Кроме алгебраической производной, сформулируем для структурных чисел еще одно понятие (в известном смысле дуальное по отношению к производной) — понятие обратной алгебраической производной.

Алгебраической обратной производной структурного числа называется структурное число $\delta A/\delta\alpha$, равное

$$\frac{\delta A}{\delta \alpha}=A\,\Big|\,\begin{matrix}\text{столбцы, содержащие}\\ \text{элемент } \alpha, \text{ опущены.}\end{matrix} \qquad (1.52)$$

Воспользовавшись способом записи структурного числа в виде семейства множеств, можно записать обратную производную как

$$\frac{\delta A}{\delta \alpha} = \{a_k \mid \alpha \notin a_k,\ a_k \in A\}. \qquad (1.52а)$$

Для обратной алгебраической производной имеют место соотношения

$$\frac{\delta}{\delta \alpha}(A_1 + A_2) = \frac{\delta A_1}{\delta \alpha} + \frac{\delta A_2}{\delta \alpha},$$
$$\frac{\delta}{\delta \alpha}(A_1 A_2) = \frac{\delta A_1}{\delta \alpha} A_2 + \frac{\delta A_2}{\delta \alpha} A_1 + A_1 A_2, \qquad (1.53)$$

справедливые для произвольных чисел A_1 и A_2.

Кроме того,

$$\frac{\delta}{\delta \alpha}(A_1 A_2) = \frac{\delta A_1}{\delta \alpha} \frac{\partial A_2}{\partial \alpha}. \qquad (1.53а)$$

Для одноэлементного структурного числа имеем

$$\frac{\delta}{\delta \alpha}[\alpha] = 0. \qquad (1.54)$$

Соотношение алгебраических производной и обратной производной можно записать следующим образом:

$$\frac{\partial A}{\partial \alpha}(A[\alpha]) = \frac{\delta A}{\delta \alpha}. \qquad (1.53б)$$

Алгебраическую обратную будем обозначать как

$$\frac{\delta A}{\delta \alpha} = A^{\alpha}. \qquad (1.55)$$

Пример 1.9. Расчет алгебраической обратной производной:

$$A = \begin{bmatrix} 1 & 2 & 1 & 5 \\ 3 & 4 & 2 & 4 \\ 5 & 7 & 3 & 8 \end{bmatrix}, \quad \frac{\delta A}{\delta 1} = \begin{bmatrix} 2 & 5 \\ 4 & 4 \\ 7 & 8 \end{bmatrix}, \quad \frac{\delta A}{\delta 2} \begin{bmatrix} 1 & 5 \\ 3 & 4 \\ 5 & 8 \end{bmatrix}.$$

Алгебраическая обратная производная имеет простую геометрическую интерпретацию.

Свойство 2. Геометрическое изображение структурного числа $\delta A/\delta\alpha$ представляет собой геометрическое изображение числа *A,* в котором ребро отключено в одной вершине и замкнуто в петлю. Обратное геометрическое изображение структурного числа $\delta A/\delta\alpha$ представляет собой обратное изображение геометрического числа *A* с замкнутым ребром α. Правильность этого свойства следует из определений изображения, обратного изображения структурного числа и обратной производной.

Вследствие простых соотношений между алгебраическими действиями, выраженными через операции производной и обратной производной, и действиями на графе, который является геометрической интерпретацией структурного числа, эти операции особенно важны в применениях алгебры структурных чисел, например, к анализу электрических цепей.

Отметим, что для структурного числа A всегда имеет место соотношение

$$A = \frac{\delta A}{\delta \alpha} + [\alpha] \frac{\partial A}{\partial \alpha}, \qquad (2.56)$$

где α — элемент числа A.

6.6. Детерминантная функция структурного числа

Аналогично с матричным исчислением на множестве структурных чисел можно определить различные функции, например детерминантную функцию.

Определение 1.9. Детерминантной функцией структурного числа A называется функция

$$\det_{Z} A = \det_{Z} \begin{bmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2n} \\ \alpha_{m1} & \alpha_{m2} & \dots & \alpha_{mn} \end{bmatrix} = \sum_{k=1}^{n} \prod_{i=1}^{m_k} z_{\alpha_{ik}}, \qquad (1.57)$$

где Z — заданное множество комплексных чисел $z_{\alpha_{ik}}$, т.е. $z_{\alpha_{ik}} \in Z$.

Определение этой функции весьма просто. Нужно перемножить комплексные числа, поставленные в соответствие индексам столбцов, и просуммировать полученные выражения, соответствующие столбцам.

Эта функция может быть кратко названа определителем или детерминантом структурного числа.

По аналогии с теорией матриц для ее обозначения используем также символ $\det_{Z} A$

$$|A| \quad \text{или} \quad \begin{vmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2n} \\ \dots & \dots & \dots & \dots \\ \alpha_{m11} & \alpha_{m22} & \dots & \alpha_{mnn} \end{vmatrix}. \qquad (1.58)$$

Пример 1.10. Нахождение определителя.

Вычислить определитель числа $A = \begin{bmatrix} 1 & 2 & 1 & 2 \\ 3 & 4 & 4 & 3 \\ 7 & 5 & 8 & 4 \end{bmatrix}$ по отношению к комплексным числам $z_1, z_2, z_3, z_4, z_5, z_6, z_7, z_8 \in Z$.

$$\det_z A = z_1 z_3 z_7 + z_2 z_4 z_5 + z_1 z_4 z_8 + z_2 z_3 z_4.$$

Очевидно, что раскрытие определителя матрицы немного сложнее, чем раскрытие определителя структурного числа. Определитель структурного числа имеет следующие свойства:

$$(A_1 = A_2) \Rightarrow (\det_z A_2 = \det_z A_1),$$

$$\det_z \frac{\partial A}{\partial \alpha} = \frac{\partial}{\partial z_\alpha} [\det_z A].$$

6.7. Функция совпадения структурного числа

Кроме ранее введенных операций сложения и умножения структурных чисел, определим еще одну операцию — конъюнкцию.

Определение 1.10. Конъюнкцией $A \cap B$ структурных чисел A и B называется структурное число, содержащее общие столбцы чисел A и B и не содержащее других столбцов.

Пример 1.11.

$$A = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}, \quad B = \begin{bmatrix} 5 & 2 & 4 \\ 3 & 1 & 1 \end{bmatrix}, \quad A \cap B = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

Определим на множестве структурных чисел еще одну функцию, важную для применения алгебры структурных чисел,— функцию совпадения и обозначим

$$\underset{z}{\mathrm{Sim}}\,(A, B)^{\varphi\psi}, \quad z_{\alpha_{ik}} \in Z.$$

Функция совпадения равна

$$\underset{z}{\mathrm{Sim}}\,(A, B)^{\varphi\psi} = \begin{cases} \det_z (A \cap B) & \text{при } \varphi^-, \\ -\det_z (A \cap B) & \text{при } \psi. \end{cases} \tag{1.59}$$

Очевидно, имеется в виду случай $A \cap B \neq 0$.

Формула (1.59) дает общее определение функции совпадения, однако в прикладном значении этой функции наиболее важна частная форма записи функции совпадения: эта функция относится к структурному числу *A,* геометрическое обратное изображение которого содержит два ориентированных ребра α и β.

Определение 1.11. Функцией совпадения

$$\underset{Z}{\mathrm{Sim}}\left(\frac{\partial A}{\partial \alpha},\ \frac{\partial A}{\partial \beta}\right),\quad z_{\alpha_{ik}},\ z_{\alpha},\ z_{\beta}\in Z \tag{1.60}$$

структурного числа *A,* обратное геометрическое изображение которого имеет два ориентированных ребра α и β, называется функция, обладающая следующими свойствами:

1) функция (1.60) — линейная комбинация выражений, имеющихся в определителях

$$\det_{Z}\left(\partial A/\partial \alpha\right) \quad \text{и} \quad \det_{Z}\left(\partial A/\partial \beta\right);$$

2) если исключить из обратного изображения ребра, определенные данным выражением, получим цикл, в котором ребра α и β ориентированы согласно или встречно, то слагаемое имеет соответственно коэффициент +1 или —1 (рис. 1.2).

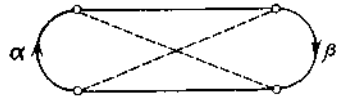


Рис. 1.2. Пояснение определения функции совпадения.

Функцию совпадения (1.60) можно в таком случае записать

$$\underset{Z}{\mathrm{Sim}}\left(\frac{\partial A}{\partial \alpha},\ \frac{\partial A}{\partial \beta}\right)=\begin{cases}\det\limits_{Z}\left(\dfrac{\partial A}{\partial \alpha}\cap\dfrac{\partial A}{\partial \beta}\right), & \text{когда ребра } \alpha \text{ и } \beta \text{ ориентированы согласно,}\\ -\det\limits_{Z}\left(\dfrac{\partial A}{\partial \alpha}\cap\dfrac{\partial A}{\partial \beta}\right), & \text{когда ребра } \alpha \text{ и } \beta \text{ ориентированы встречно.}\end{cases} \tag{1.61}$$

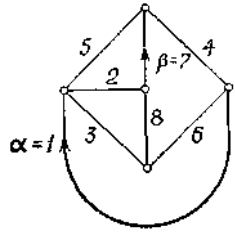


Рис. 1.3.

Пример 1.12. Определить функции совпадения структурного числа *A,* обратное изображение которого есть граф

$$\underset{Z}{\mathrm{Sim}}\left(\frac{\partial A}{\partial \alpha}, \frac{\partial A}{\partial \beta}\right),$$

изображенный на рис. 1.3 ($\alpha = 1$, $\beta = 7$).

Решение. Структурное число определим по выражению (1.13) как произведение простых чисел, соответствующих всем независимым контурам обратного изображения.

Имеем

$$\begin{array}{l} 2\ 5\ 7 \\ 2\ 3\ 8 \\ \times\ 4\ 7\ 8\ 6 \\ 1\ 3\ 6 \end{array}$$

$$\left[\begin{array}{cccccccccccccccccccccccccccccccc} 2&2&2&2&2&2&2&2&2&2&2&2&2&2&5&5&5&5&5&5&5&5&5&5&5&5&5&5&5&5&5&5 \\ 3&3&3&3&3&3&3&8&8&8&8&8&8&8&8&2&2&2&2&2&2&2&2&2&2&2&3&3&3&3&3&3 \\ 4&4&7&7&8&8&4&4&4&7&7&7&6&6&4&4&4&7&7&7&8&8&8&6&6&4&4&7&7&8 \\ 1&6&1&6&1&6&1&3&6&1&3&6&1&3&1&3&6&1&3&6&1&3&6&1&3&1&6&1&6&1 \\ 5&5&5&5&5&5&5&5&5&5&7&7&7&7&7&7&7&7&7&7&7&7&7&7&7&7&7&7&7&7 \\ 3&3&8&8&8&8&8&8&8&8&2&2&2&2&2&2&2&2&3&3&3&3&3&8&8&8&8&8&8 \\ 8&6&4&4&4&7&7&7&6&6&4&4&4&8&8&8&6&6&4&4&8&8&6&4&4&4&6&6 \\ 6&1&1&3&6&1&3&6&1&3&1&3&6&1&3&6&1&3&1&6&1&6&1&1&3&6&1&3 \end{array}\right].$$

Далее определим алгебраические производные

$$\frac{\partial A}{\partial 1} = \left[\begin{array}{cccccccccccccccccccccccc} \boxed{2}&2&2&2&2&5&5&5&\boxed{5}&5&5&\boxed{5}&5&5&5&\boxed{5}&7&7&7&7&7&7&7 \\ \boxed{3}&3&3&3&8&8&2&2&2&\boxed{2}&3&3&\boxed{3}&3&8&8&\boxed{8}&2&2&3&3&3&8&8 \\ \boxed{4}&7&8&4&6&4&7&8&\boxed{6}&4&7&\boxed{8}&6&4&7&\boxed{6}&4&6&4&8&6&4&6 \end{array}\right],$$

$$\frac{\partial A}{\partial 7} = \left[\begin{array}{cccccccccccccccccccccccc} 2&5&5&\boxed{5}&5&\boxed{5}&5&\boxed{5}&\boxed{5}&2&\boxed{2}&2&2&2&3&3&3&3&3&8&8&8&8&8 \\ 3&2&2&\boxed{2}&3&\boxed{3}&8&\boxed{8}&\boxed{8}&4&\boxed{4}&4&6&6&4&4&8&8&6&4&4&4&6&6 \\ 1&1&3&\boxed{6}&1&\boxed{6}&1&\boxed{3}&\boxed{6}&1&\boxed{3}&6&1&3&1&6&1&6&1&1&3&6&1&3 \end{array}\right].$$

Столбцы, общие для *dA/d1* и *дA/д7,* взяты в рамки.

При рассмотрении графа на рис. 1.3 легко заметить, что при исключении ребер 5 2 6, 5 3 6, 5 8 3, 5 8 6 граф сводится к графу с одним контуром, в котором ребра $\alpha = 1$ и $\beta = 7$ ориентированы согласно. С другой стороны, при исключении из графа ребер 2 3 4 в полученном контуре ребра 1 и 7 ориентированы встречно.

Поэтому выражения $z_5z_2z_6$, $z_5z_3z_6$, $z_3z_8z_3$. $z_5z_8z_6$ имеют знак плюс, $z_2z_3z_4$ — знак минус.

Окончательно получим

$$\operatorname*{Sim}_{Z}\left(\frac{\partial A}{\partial 1},\ \frac{\partial A}{\partial 7}\right)=z_5z_2z_6+z_5z_3z_6+z_5z_8z_3+z_5z_8z_6-z_2z_3z_4.$$

Можно также обосновать свойство, согласно которому при исключении из обратного изображения структурного числа ребер, определенных столбцами $\partial A/\partial\alpha \cap \partial A/\partial\beta$, граф всегда сводится к такому графу, у которого цикломатическое число $m = 1$. Определение ориентации ребер α и β по отношению друг к другу не встречает трудностей. Не каждый граф отображает электрическую цепь, в которой не могут присутствовать лишние элементы (обесточенные или на которых нет напряжения). Для определения класса графов, с которыми имют дело при анализе электрических цепей, введем общее определение соответственного или сильно связного графа.

Определение 1.12. Граф называется соответственным, если каждые две его вершины принадлежат хотя бы одному элементарному контуру.

Для соответственного графа справедливо следующее свойство.

Свойство 3. Граф (мультиграф) будет соответственным тогда и только тогда, когда он служит обратным изображением структурного числа *A,* удовлетворяющего условию

$$\bigvee_{\alpha\in A}\bigwedge_{\beta\in A}\left[\frac{\partial A}{\partial\alpha}\cap\frac{\partial A}{\partial\beta}\neq 0\right]. \tag{1.62}$$

Справедливость этого свойства следует из определения функции совпадения, согласно которомустолбцы

$$\frac{\partial A}{\partial\alpha}\cap\frac{\partial A}{\partial\beta}$$

соответствуют ребрам, исключение которых приводит к упрощению графа обратного изображения к одному циклу с ребрами α и β.

На рис. 1.4 показано несколько графов, из которых только один граф соответственный.Если применить условие (1.62), например к графу, показанному на рис. 1.4, б, получим

$$A=[1\ \ 2]\,[3\ \ 4]=\begin{bmatrix}1&1&2&2\\3&4&3&4\end{bmatrix},$$

$$\frac{\partial A}{\partial\alpha}=[3\ \ 4],\ \frac{\partial A}{\partial 2}=[3\ \ 4],\ \frac{\partial A}{3}=[1\ \ 2],\ \frac{\partial A}{\partial 4}=[1\ \ 2].$$

Тогда

$$\frac{\partial A}{\partial \alpha} \cap \frac{\partial A}{\partial 2} = [3\ 4] \neq 0,$$

а также

$$\frac{\partial A}{\partial \alpha} \cap \frac{\partial A}{\partial 3} = \frac{\partial A}{\partial \alpha} \cap \frac{\partial A}{\partial 4} = 0,$$

т. е. условие (1.62) не выполняется для графа (рис. 1.4, *б)* и этот граф не соответственный.

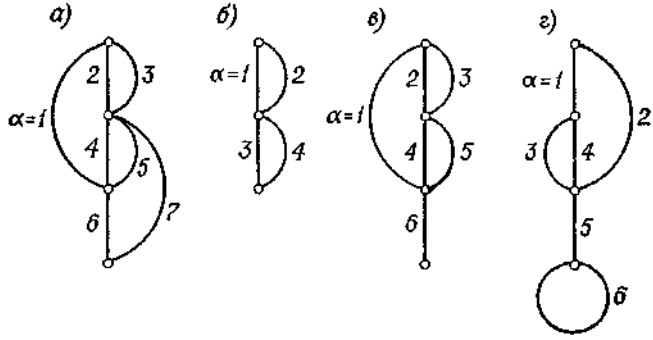


Рис. 1.4. Примеры графов: *а*) соответственный; *б, в, г)* несоответственные.

Очевидно, что применение условия (1.62) для определения характера графа излишне, если известна его структура. Из рассмотрения контуров графа можно непосредственно сделать вывод о том, выполняется ли условие (1.62). Однако это условие весьма ценно, если известно только структурное число, не разложенное на первичные сомножители, а также для использования при синтезе электрических цепей с помощью структурных чисел на ЭВМ.

6.8. Понятие ряда и последовательности структурных чисел

Если натуральным числам поставить в соответствие структурные числа, то можно сказать, что таким образом определена последовательность структурных чисел, записываемая в виде

$$\langle A_n \rangle = A_1,\ A_2,\ A_3,\ \ldots,\ A_n,\ \ldots .$$

Понятия сходимости и границы последовательности структурных чисел основываются на понятии метрики. Положим, дано структурное число

$$A = \{a_k \mid \alpha_{ij} \in a_k,\ i, j, k = 1, 2, 3, \ldots\},$$

где мощность множеств a и A конечная, а α_{ij} — элементы нормированного пространства.

Введем сначала понятие нормы множества a, которую обозначим как $\| a \|$. Примем определение

$$a = \{\alpha_1, \alpha_2, \ldots, \alpha_\nu\} \Rightarrow \| a \| = \sqrt{\| \alpha_1 \|^2 + \| \alpha_2 \| + \ldots + \| \alpha_\nu \|^2},\ \| \phi \| = 0,$$

где $\| \alpha_i \|$ —норма элемента α_i.

Норму структурного числа определим как

$$\| A \| = \sqrt{\sum_{\lambda} \| a_\lambda \|^2},$$

где λ проходят все столбцы, имеющиеся в числе A. Метрику на множестве структурных чисел определим как

$$\rho\ (A, B) = \| A \stackrel{\Delta}{=} B \|,$$

где $\stackrel{\Delta}{=}$ — означает симметричную разность множеств. Из этого определения следует, что метрика $\rho(A, B)$ удовлетворяет следующим основным условиям:

$$\rho\ (A, B) = 0 \Leftrightarrow A = B,$$

$$\rho\ (A. B) = \rho\ (B, A),$$

$$\rho\ (A, B) + \rho\ (B, C) \geqslant \rho\ (A, C).$$

Для двух произвольных структурных чисел справедливо также неравенство

$$\| AB \| \leqslant \| A \| \| B \|,$$

которое следует из неравенства Буняковского — Шварца.

Если для последовательности структурных чисел A_n существует структурное число A, удовлетворяющее равенству

$$\lim_{n \to \infty} \rho\ (A_n, A) = 0,$$

то структурное число A называется границей последовательности структурных чисел A_n и записывается в виде

$$A = \lim_{n \to \infty} A_n.$$

Последовательность A_n называется сходящейся, если имеет границу, и, наоборот, расходящейся, если таковая отсутствует.

Кроме сходимости по отношению к метрике, введем и другие понятия сходимости последовательности структурных чисел, которые обозначаются

$$\operatorname*{Lim}_{n\to\infty} \mathrm{ob}\, A_n \quad \text{и} \quad \operatorname*{Lim}_{n\to\infty} \mathrm{cob}\, A_n$$

и определяются с помощью изображения и обратного изображения структурного числа:

$$(\operatorname*{Lim}_{n\to\infty} \mathrm{ob}\, A_n = A) \Leftrightarrow [\operatorname*{Lim}_{n\to\infty} \mathrm{ob}\,(A_n) = \sigma] \wedge]\sigma = \mathrm{ob}\,(A)],$$

$$(\operatorname*{Lim}_{n\to\infty} \mathrm{cob}\, A_n = A) \Leftrightarrow [\operatorname*{Lim}_{n\to\infty} \mathrm{cob}\,(A_n) = \sigma] \wedge [\sigma = \mathrm{cob}\,(A)].$$

Примером сходимости последовательности A_n по отношению к обратному изображению может служить цепь, метрический граф которой имеет ступенчатую структуру с равномерно распределенными на отрезке [0, 1] вершинами (рис. 1.5).

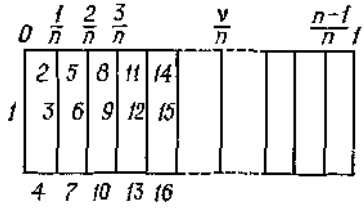


Рис. 1.5. Лестничный граф с равномерно распределенными вершинами.

Если увеличить число делений отрезка [0, 1], то при $n \to \infty$ граф преобразуется в структуру с густым (однако четным) множеством ребер. Нумеруя грани графа, например, как показано на рис. 1.5, можно определить следующую последовательность структурных чисел:

$$A_1 = [1\ 2\ 3\ 4],$$
$$A_2 = [1\ 2\ 3\ 4]\ [3\ 5\ 6\ 7],$$
$$A_3 = [1\ 2\ 3\ 4]\ [3\ 5\ 6\ 7]\ [6\ 8\ 9\ 10],$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

обратным изображением которой и служит наш граф.

Эта последовательность сходится к обратному изображению структурного числа

$$A = \text{cob}^{-1}\,\sigma, \quad \sigma = \lim_{n\to\infty} S_n,$$

где S_n — последовательность цепей (метрических графов) вида изображенных на рис. 1.5.

Рядом структурных чисел называется выражение

$$A_1 + A_2 + A_3 + \ldots + A_n + \ldots = \sum_{n=1}^{\infty} A_n.$$

Структурные числа $A_1, A_2, A_3, \ldots$ называются составляющими ряда, числа же

$$\begin{aligned} S_1 &= A_1, \\ S_2 &= A_1 + A_2, \\ S_3 &= A_1 + A_2 + A_3, \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ S_n &= A_1 + A_2 + A_3 + \ldots + A_n \end{aligned}$$

есть частные суммы ряда. Бесконечный ряд структурных чисел называется сходящимся, если последовательность частичных сумм сходится. Предел последовательности частичных сумм называется суммой ряда структурных чисел.

Если

$$\lim_{n\to\infty} S_n = S,$$

то

$$S = \sum_{n=1}^{\infty} A_n.$$

Микромодуль 22
Структурные числа высшей категории

Физические системы состоят из элементов, взаимодействующих друг с другом различным образом. Например, электрические цепи могут состоять из многополюсных элементов (многополюсников); их также можно рассматривать как системы, состоящие из блоков или подцепей. Топологические модели таких систем представим в виде

графов второй категории, построенных из двумерных континуумов (блоков) с выделенными точками, называемыми полюсами. Блоки соответствуют ребрам линейных графов первой категории. Структурные числа блок-графов назовем структурными числами высших категорий — второй, третьей и т. д. Эти числа, подобно матрицам, состоящим из подматриц, представляют собой семейства структурных чисел низшей категории. Основываясь на определении операций над числами первой категории, определим в соответствии с теорией множеств и элементами математической логики операции над структурными числами высшей категории.

6.9. Определение структурного числа второй категории

В общем определении структурных чисел не уточнялись характерные черты множеств элементов, из которых состоит это число, поэтому можно рассмотреть случаи, когда эти элементы также являются структурными числами. В связи с этим введем понятие структурного числа ${}^2A$ второй категории следующим образом.

Определение 2.1. Структурное число второй категории ${}^2A$ есть семейство множеств ${}^2a_j$

$$ {}^2A = \{{}^2a_j\}_{j=1,\,2,\,\ldots,\,n}, \tag{2.1} $$

где

$$ {}^2a_j = \{A_{ij}\}_{i=1,\,2\,\ldots,\,m}, $$

$\bigwedge\limits_i \bigwedge\limits_j A_{ij}$ — структурное число первой категории.

Структурное число второй категории можно также записать в виде

$$ {}^2A = \begin{bmatrix} A_{11} & \ldots & A_{1n} \\ A_{21} & \ldots & A_{2n} \\ \ldots & \ldots & \ldots \\ A_{m1} & \ldots & A_{mn} \end{bmatrix}, \tag{2.2} $$

или

$$ {}^2A = [A_{ij}]_{\substack{i=1,\,2,\,\ldots,\,m \\ j=1,\,2,\,\ldots,\,n}}, \tag{2.3} $$

где элементы A_{ij} — структурные числа первой категории.

Введем понятие замещающего числа для структурного числа второй категории.

Определение 2.2. Замещающим числом для структурного числа второй категории ${}^2A$ называется структурное число первой категории A, полученное применением операций алгебры структурных чисел над элементами A_{ij} числа ${}^2A$:

$$A = \sum_{j=1}^{n} \prod_{i=1}^{m} A_{ij}, \quad {}^2A = [A_{ij}]_{\substack{i=1,2,\ldots,m \\ j=1,2,\ldots,n}}. \tag{2.4}$$

Обозначим соотношение соответствия замещающего числа A числу ${}^2A$ через

$$A \overset{e}{=} {}^2A \quad \text{или} \quad {}^2A \overset{e}{=} A. \tag{2.5}$$

Поясним способ нахождения замещающего числа следующим примером:

$${}^2A = \begin{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 3 \end{bmatrix} & \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \\ [1\ 2\ 4] \begin{bmatrix} 1 & 5 \\ 2 & 6 \end{bmatrix} & \\ \begin{bmatrix} 6 \\ 5 \end{bmatrix} &]\phi] \end{bmatrix} \overset{e}{=} \begin{bmatrix} 1 & 1 \\ 2 & 3 \end{bmatrix} [1\ 2\ 4] \begin{bmatrix} 6 \\ 5 \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \begin{bmatrix} 1 & 5 \\ 2 & 6 \end{bmatrix} [\phi] =$$

$$= \begin{bmatrix} 1 & 1 \\ 2 & 3 \\ 4 & 4 \\ 5 & 5 \\ 6 & 6 \end{bmatrix} = A, \quad {}^2A \overset{e}{=} A.$$

Определим для структурных чисел второй категории понятие равенства, а также операции сложения и умножения:

$$({}^2A = {}^2B) \overset{df}{\Leftrightarrow} (A = B), \tag{2.6}$$

$${}^2A + {}^2B \overset{e}{=} A + B, \tag{2.7}$$

$${}^2A \cdot {}^2B \overset{e}{=} AB, \tag{2.8}$$

где $A \overset{e}{=} {}^2A$, $B \overset{e}{=} {}^2B$.

Таким образом, соотношение $\overset{e}{=}$ представляет собой гомеоморфизм.

Для структурных чисел второй категории справедливы следующие соотношения:

$$\{({}^2A \subset {}^2B) \wedge ({}^2B \subset {}^2A)\} \Rightarrow ({}^2A = {}^2B), \tag{2.9}$$

$$\{{}^2C = ({}^2A \triangleq {}^2B)\} \Rightarrow ({}^2C = {}^2A + {}^2B), \tag{2.10}$$

$$[{}^2C = \{{}^2c \mid ({}^2c = {}^2a \cup {}^2b) \wedge ({}^2a \in {}^2A) \wedge ({}^2b \in {}^2B) \wedge$$
$$\wedge ({}^2a \wedge {}^2b = \phi) \wedge (r({}^2a \cup {}^2b) = 2k - 1)\}] \Rightarrow ({}^2C = {}^2A \cdot {}^2B), \tag{2.11}$$

где $\triangleq$ означает симметричную разность, r — функция повторений, а k — натуральное число.

Легко заметить, что равенство структурных чисел второй категории рефлексивно, симметрично и транзитивно, операции сложения и умножения коммутативны и ассоциативны, а умножение дистрибутивно по отношению к сложению.

Заметим, что модуль сложения ${}^2[\ ] = 0$ структурных чисел второй категории есть всякое структурное число второй категории, замещающее число которого служит модулем сложения $[\ ]$ структурных чисел первой категории; а модуль умножения ${}^2[\phi] = 1$ структурных чисел второй категории представляет собой всякое структурное число второй категории, замещающее число которого служит модулем умножения $[\phi]$ структурных чисел первой категории. При этом справедливы следующие соотношения:

1. $$[\underset{1\ 2\ \ldots\ n}{AA\ \ldots\ A}] = \begin{cases} \overset{e}{=} A \text{ при нечетном } n, \\ = 0 \text{ при четном } n; \end{cases} \tag{2.12}$$

2. $$\left[\underset{n \text{ раз}}{\begin{Bmatrix} A \\ A \\ \vdots \\ A \end{Bmatrix}}\right] \begin{cases} = 0 \text{ при } A \in \mathfrak{U}, \\ = 1 \text{ при } A \in \mathfrak{U},\ n \text{ четное}, \\ \overset{e}{=} A \text{ при } A \in \mathfrak{U},\ n \text{ нечетное}, \end{cases} \tag{2.13}$$

где $\mathfrak{U}$ — множество структурных чисел вида

$$A = \{[\phi],\ a_1,\ a_2,\ \ldots\}$$

3. $$([A_1 \ldots A_j \ldots A_n] = 0) \Leftrightarrow$$
$$\Leftrightarrow \{(A_j = [A_1 \ldots A_{j-1} A_{j+1} \ldots A_n]) \vee$$
$$\vee (A_1 = A_2 = \ldots = A_j = \ldots = A_n = 0)\},$$
$$j = 1,\ 2,\ \ldots,\ n. \tag{2.14}$$

4. $$\left(\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = 0\right) \not\Rightarrow (A_1 = 0 \vee A_2 = 0).$$

(2.15)

5. Равенство

$$\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = 0 \qquad (2.16)$$

не имеет единственного решения для A_1 и A_2.

6. $$\left(\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} A_1 \\ A_3 \end{bmatrix}\right) /\Rightarrow (A_2 = A_3). \qquad (2.17)$$

Структурное число второй категории ${}^2A^d$, элементы которого — дополнительные числа A^d_{ij}, назовем дополнительным числом второй категории. Для дополнительных чисел второй категории применяем те же самые операции, что и для структурных чисел второй категории, поэтому приведенные выше определения и соотношения справедливы также и для дополнительных чисел второй категории ${}^2A^a$.

Если в числе ${}^2A$ все элементы A_{ij} заменить на их дополнительные числа A^d_{ij}, то получим дополнительное число второй категории $({}^2A)^{d*}$, замещающее число которого A^{d*} в общем случае не равно дополнению A^d замещающего структурного числа A для числа ${}^2A$, т. е.

$$A^{d*} \neq A^d, \quad A^{d*} \overset{e}{=} ({}^2A)^{d*}, \quad A \overset{e}{=} {}^2A,$$

а следовательно,

$${}^2A^d \neq ({}^2A)^{d*}.$$

6.9.1. Алгебраическая производная и обратная производная структурного числа второй категории

Алгебраическую производную и обратную производную определим на основе понятий производной и обратной производной замещающего структурного числа.

Определение 2.3. Алгебраической (обратной) производной $\partial\,({}^2A)/\partial\alpha\cdot[\delta\,({}^2A)/\delta\alpha]$ структурного числа второй категории ${}^2A$ по элементу α называется всякое структурное число второй категории, замещающее число которого $\partial A/\partial\alpha\cdot(\delta A/\delta\alpha)$ есть производная (обратная производная) замещающего числа A для числа ${}^2A$ по элементу α. Это определение соответственно можно представить в виде соотношений

$$\frac{\partial\,(^2A)^{-1}}{\partial\alpha} = \frac{\partial A}{\partial\alpha}, \quad \frac{\delta\,(^2A)^{-1}}{\delta\alpha} = \frac{\delta A}{\delta\alpha}, \quad {}^2A = A. \tag{2.18}$$

На основании правил для производной и обратной производной суммы и произведения структурных чисел первой категории можно написать следующие соотношения для структурных чисел второй категории:

$$\frac{\partial}{\partial\alpha}[A_1 A_2] = \left[\frac{\partial A_1}{\partial\alpha}\ \frac{\partial A_2}{\partial\alpha}\right], \tag{2.19}$$

$$\frac{\partial}{\partial\alpha}\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} \frac{\partial A_1}{\partial\alpha} & A_1 \\ A_2 & \frac{\partial A_2}{\partial\alpha} \end{bmatrix}, \tag{2.20}$$

$$\frac{\delta}{\delta\alpha}[A_1 A_2] = \left[\frac{\delta A_1}{\delta\alpha}\ \frac{\delta A_2}{\delta\alpha}\right], \tag{2.21}$$

$$\frac{\delta}{\delta\alpha}\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} \frac{\delta A_1}{\delta\alpha} \\ \frac{\delta A_2}{\delta\alpha} \end{bmatrix}. \tag{2.22}$$

Следовательно, обратная производная

$$\delta\,(^2A)/\delta\alpha$$

— операция аддитивная и мультипликативная.

На основе этих соотношений можно определить алгебраические производную и обратную производную по элементу α любого структурного числа второй категории.

Пример 2.1.

$$\frac{\partial}{\partial\alpha}\begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix} = \begin{bmatrix} \frac{\partial A_1}{\partial\alpha} & A_1 & \frac{\partial A_2}{\partial\alpha} & A_2 \\ A_3 & \frac{\partial A_3}{\partial\alpha} & A_4 & \frac{\partial A_4}{\partial\alpha} \end{bmatrix}.$$

Если, например, структурные числа A_1 и A_2 не содержат элемента α, то

$$\frac{\partial}{\partial\alpha}\begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix} = \begin{bmatrix} A_1 & A_2 \\ \frac{\partial A_3}{\partial\alpha} & \frac{\partial A_4}{\partial\alpha} \end{bmatrix}.$$

По этим же правилам находятся алгебраическая производная и обратная производная дополняющих чисел второй категории.

Для определения алгебраической производной структурного числа второй категории ${}^2A$ по элементу первой или второй категории принимаем следующие правила:

$$\frac{\partial A}{\partial [\alpha_1 \alpha_2]} \overset{\text{df}}{=} \left[\frac{\partial A}{\partial \alpha_1} \frac{\partial A}{\partial \alpha_2} \right], \quad \frac{\partial A}{\partial \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}} \overset{\text{df}}{=} \frac{\partial^2 A}{\partial \alpha_1 \partial \alpha_2}, \tag{2.23}$$

из которых следует

$$\frac{\partial ({}^2A)}{\partial [A_1 A_2]} = \frac{\partial ({}^2A)}{\partial (A_1 + A_2)} = \frac{\partial ({}^2A)}{\partial A_1} + \frac{\partial ({}^2A)}{\partial A_2},$$

$$\frac{\partial ({}^2A)}{\partial \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}} = \frac{\partial ({}^2A)}{\partial (A_1 A_2)} = \frac{\partial^2 ({}^2A)}{\partial A_1 \partial A_2} \tag{2.24}$$

и

$$\frac{\partial ({}^2A)}{\partial ({}^2B)} = \frac{\partial ({}^2A)}{\partial B} \overset{\text{e}}{=} \frac{\partial A}{\partial B} = (A^d B)^d, \; A \overset{\text{e}}{=} {}^2A, \; B \overset{\text{e}}{=} {}^2B, \tag{2.25}$$

как естественное обобщение соотношения $\partial A / \partial \alpha = (A^d [\alpha])^d$, а также

$$\frac{\partial ({}^2A)}{\partial A} \overset{\text{e}}{=} \frac{\partial A}{\partial A} = \begin{cases} 1 \text{ для нечетного количества столбцов числа } A, \\ 0 \text{ для четного количества столбцов числа } A, \end{cases} \tag{2.26}$$

где $A \overset{\text{e}}{=} {}^2A$.

Пример 2.2. Найдем алгебраическую производную структурного числа второй категории

$${}^2A = \begin{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} 2 & 5 \\ 4 & 6 \end{bmatrix} \\ [4] \quad [3\ 6] \end{bmatrix}$$

по структурномучислу первой категории $A = \begin{bmatrix} 1 & 1 \\ 2 & 3 \end{bmatrix}$.

Решение

$$\frac{[\partial ({}^2A)}{\partial A} = \frac{\partial^2 ({}^2A)}{\partial 1 \partial 2} + \frac{\partial^2 ({}^2A)}{\partial 1 \partial 3} = \frac{\partial}{\partial 2} \left[\frac{\partial}{\partial 1} \begin{bmatrix} 1 & 1 \\ 2 & 3 \end{bmatrix} \right] +]4]$$

$$+ \frac{\partial}{\partial 3} \left[\frac{\partial}{\partial 1} \begin{bmatrix} 1 & 1 \\ 2 & 3 \end{bmatrix} \right] \overset{\text{e}}{=} [4] +]4] = 0.$$

6.10. Структурные числа *k*-й категории

Структурные числа *k*-й категории построим из структурных чисел *(k* — 1)-й категории подобно тому, как были построены структурные числа 2-й категории. В связи с этим введем следующее определение.

Определение 2.4. Структурным числом *k*-й категории ${}^{k}A$ называется семейство множеств ${}^{k}a_j$

$$ {}^{k}A = \{{}^{k}a_j\}_{j=1, 2, \ldots, n}, \tag{2.27} $$

где

$$ {}^{k}a_j = \{{}^{k-1}A_{ij}\}_{i=1, 2, \ldots, m}, $$

$\bigwedge_i \bigwedge_j {}^{k-1}A_{ij}$ — структурное число *(k* — 1)-й категории. Структурное число *k* -й категории можно также записать в виде

$$ {}^{k}A = \begin{bmatrix} {}^{k-1}A_{11} & \ldots & {}^{k-1}A_{1n} \\ {}^{k-1}A_{21} & \ldots & {}^{k-1}A_{2n} \\ \ldots & \ldots & \ldots \\ {}^{k-1}A_{m1} & \ldots & {}^{k-1}A_{mn} \end{bmatrix} \tag{2.28} $$

или

$$ {}^{k}A = [{}^{k-1}A_{ij}]_{\substack{i=1, 2, \ldots, m \\ j=1, 2, \ldots, n}}, \tag{2.29} $$

где элементы ${}^{k-1}A_{ij}$ — структурные числа *(k* — 1)-й категории.

Введем понятие замещающего числа для структурного числа *k*-й категории.

Определение 2.5. Замещающим числом для структурного числа *k*-й категории ${}^{k}A$ называется структурное число первой категории *A,* полученное применением операций алгебры структурных чисел над замещающими числами для структурных чисел *(k* — 1)-й категории, являющимися элементами числа ${}^{k}A$*:*

$$ A = \sum_{j=1}^{n} \prod_{i=1}^{m} A_{ij}, \quad {}^{k}A = [{}^{k-1}A_{ij}]_{\substack{i=1, 2, \ldots, m \\ j=1, 2, \ldots, n}} \tag{2.30} $$

где A_{ij} — замещающее число для ${}^{k-1}A_{ij}$. Обозначим соотношение соответствия замещающего числа A числу ${}^{k}A$ через

$$ A \overset{c}{=} {}^{k}A \quad \text{или} \quad {}^{k}A \overset{c}{=} A. \tag{2.31} $$

Для структурного числа *k*-й категории определим понятие равенства, а также операции сложения и умножения

$$\left.\begin{array}{l}({}^kA = {}^kB) \overset{\text{df}}{\Leftrightarrow} (A = B),\\ {}^kA + {}^kB \overset{e}{=} A + B,\\ {}^kA \cdot {}^kB \overset{e}{=} AB,\end{array}\right\} \left\{A \overset{e}{=} {}^kA, B \overset{e}{=} {}^kB.\right. \qquad (2.32)$$

Таким образом, соотношение $\overset{e}{=}$ представляет собой гомеоморфизм.

Для структурных чисел *k*-й категории справедливы следующие соотношения:

$$({}^kA \subset {}^kB) \wedge ({}^kB \subset {}^kA) \Rightarrow ({}^kA = {}^kB), \qquad (2.33)$$

$$\{{}^kC = ({}^kA \overset{\triangle}{=} {}^kB)\} \Rightarrow ({}^kC = {}^kA + {}^kB), \qquad (2.34)$$

$$[{}^kC = \{{}^kc \mid ({}^kc = {}^ka \cup {}^kb) \wedge ({}^ka \in {}^kA) \wedge$$
$$\wedge ({}^kb \in {}^kB) \wedge ({}^ka \cap {}^kb = \phi) \wedge$$
$$\wedge (r({}^ka \cup {}^kb) = 2n - 1)\}] \Rightarrow ({}^kC = {}^kA \cdot {}^kB), \qquad (2.35)$$

где $\overset{\triangle}{=}$ есть симметричная разность, r — функция повторений, а n — натуральное число.

Формулы (2.32) можно обобщить на структурные числа разных категорий ${}^kA$ и ${}^mB$ $(k > 1,\ m > 1)$:

$$\left.\begin{array}{l}({}^kA = {}^mB) \Leftrightarrow (A = B),\\ {}^kA + {}^mB \overset{e}{=} A + B,\\ {}^kA \cdot {}^mB \overset{e}{=} AB.\end{array}\right\} A \overset{e}{=} {}^kA,\ B \overset{e}{=} {}^mB. \qquad (2.36)$$

Подобно структурным числам второй категории, равенство чисел *k*-й категории рефлексивно, симметрично и транзитивно, операции сложения и умножения коммутативны и ассоциативны, а умножение дистрибутивно относительно сложения. Поэтому можно написать следующие соотношения:

$$[{}^kA_1 [{}^kA_2 {}^kA_3]] = [[{}^kA_1 {}^kA_2] {}^kA_3] = [{}^kA_1 {}^kA_2 {}^kA_3],$$

$$\left[\begin{array}{c}{}^kA_1\\ \left[\begin{array}{c}{}^kA_2\\ {}^kA_3\end{array}\right]\end{array}\right] = \left[\left[\begin{array}{c}{}^kA_1\\ {}^kA_2\end{array}\right]\begin{array}{c}\\ {}^kA_3\end{array}\right] = \left[\begin{array}{c}{}^kA_1\\ {}^kA_2\\ {}^kA_3\end{array}\right] \qquad \left[\begin{array}{c}{}^kA_1\\ [{}^kA_2 {}^kA_3]\end{array}\right] = \left[\begin{array}{c}{}^kA_1 {}^kA_1\\ {}^kA_2 {}^kA_3\end{array}\right].$$

Модуль сложения ${}^k[\ ] = 0$ структурных чисел k-й категории есть всякое структурное число k-й категории, замещающее число которого служит модулем сложения [ ] структурных чисел первой категории.

Модуль умножения ${}^k[\phi] = 1$ структурных чисел k-й категории есть всякое структурное число k-й категории, замещающее число которого служит модулем умножения $[\phi]$ структурных чисел первой категории.

Из этого вытекают следующие соотношения и соответствия:

1. $$[{}^kA\,{}^kA \ldots {}^kA] = \begin{cases} \overset{e}{=} {}^kA \text{ при нечетном } n, \\ = 0 \text{ при четном } n. \end{cases} \qquad (2.37)$$

2. $$\left[n \text{ раз} \begin{Bmatrix} {}^kA \\ {}^kA \\ \vdots \\ {}^kA \end{Bmatrix} \right] = \begin{cases} 0 \text{ при } {}^kA \notin \mathfrak{A}, \\ 1 \text{ при } {}^kA \in \mathfrak{A},\ n \text{ четное}, \\ {}^kA \text{ при } {}^kA \in \mathfrak{A},\ n \text{ нечетное}, \end{cases} \qquad (2.38)$$

где ${}^k\mathfrak{A}$ — множество структурных чисел k-й категории

$${}^kA = \{{}^{k-1}[\phi],\ {}^ka_1,\ {}^ka_2,\ \ldots,\ {}^ka_n\}.$$

3. $$([{}^kA_1 \ldots {}^kA_j \ldots {}^kA_n] = 0) \Leftrightarrow$$

$$\Leftrightarrow \{({}^kA_j \overset{e}{=} [{}^kA_1 \ldots {}^kA_{j-1}\,{}^kA_{j+1} \ldots {}^kA_n]) \vee$$
$$\vee ({}^kA_1 = {}^kA_2 = \ldots = {}^kA_j = \ldots = {}^kA_n = 0)\},\ j = 1, 2, \ldots, n. \qquad (2.39)$$

4. $$\left(\begin{bmatrix} {}^kA_1 \\ {}^kA_2 \end{bmatrix} = 0 \right) /\!\Rightarrow ({}^kA_1 = 0 \vee {}^kA_2 = 0). \qquad (2.40)$$

5. Равенство $$\begin{bmatrix} {}^kA_1 \\ {}^kA_2 \end{bmatrix} = 0 \qquad (2.41)$$

не имеет единственного решения.

6. $$\left(\begin{bmatrix} {}^kA_1 \\ {}^kA_2 \end{bmatrix} = \begin{bmatrix} {}^kA_1 \\ {}^kA_3 \end{bmatrix} \right) /\!\Rightarrow ({}^kA_2 = {}^kA_3). \qquad (2.42)$$

Если элементы структурного числа k-й категории ${}^kA^d$ являются дополнительными числами $(k-1)$-й категории ${}^{k-1}A_i^d$

$${}^kA^d = \{{}^ka_j^d\}_{j=1, 2, \ldots, n},\ {}^ka_j^d = \{{}^{k-1}A_i^d\}_{i=1, 2, \ldots, m}, \qquad (2.43)$$

то число A^d называется дополнительным структурным числом k-й категории или, короче, дополнительным числом k-й категории.

Для дополнительных чисел *k*-й категории справедливы те же операции, что и для структурных чисел *k*-й категории.

6.10.1. Алгебраическая производная и обратная производная структурного числа *k*-й категории

Подобно структурным числам второй категории, определим алгебраическую производную и обратную производную структурного числа *k*-й категории с помощью понятий производной и обратной производной замещающего числа.

Определение 2.6. Алгебраической производной (обратной производной) $[\partial ({}^{k}A)/\partial\alpha]$ $[\delta ({}^{k}A)/\delta\alpha]$ структурного числа *k*-й категории ${}^{k}A$ по элементу α называется всякое структурное число *k*-й категории, замещающее число которого $(\partial A/\partial\alpha)$ $(\delta A/\delta\alpha)$ есть алгебраическая производная (обратная производная) замещающего числа *A* для структурного числа ${}^{k}A$ по элементу α. Это определение можно представить в виде следующих соотношений:

$$\frac{\partial ({}^{k}A)}{\partial\alpha} \overset{e}{=} \frac{\partial A}{\partial\alpha}, \quad \frac{\delta ({}^{k}A)}{\delta\alpha} \overset{e}{=} \frac{\delta A}{\delta\alpha}, \quad {}^{k}A \overset{e}{=} A. \qquad (2.44)$$

На основании правил для производной и обратной производной суммы и произведения структурных чисел первой категории можно записать следующие соотношения для структурных чисел *k*-й категории:

$$\frac{\partial}{\partial\alpha}\left[{}^{k-1}A_1 {}^{k-1}A_2\right] = \left[\frac{\partial ({}^{k-1}A_1)}{\partial\alpha}\ \frac{\partial ({}^{k-1}A_2)}{\partial\alpha}\right], \qquad (2.45)$$

$$\frac{\partial}{\partial\alpha}\begin{bmatrix} {}^{k-1}A_1 \\ {}^{k-1}A_2 \end{bmatrix} = \begin{bmatrix} \dfrac{\partial ({}^{k-1}A_1)}{\partial\alpha} & {}^{k-1}A_1 \\ {}^{k-1}A_2 & \dfrac{\partial ({}^{k-1}A_2)}{\partial\alpha} \end{bmatrix}, \qquad (2.46)$$

$$\frac{\delta}{\delta\alpha}\left[{}^{k-1}A_1 {}^{k-1}A_2\right] = \left[\frac{\delta ({}^{k-1}A_1)}{\delta\alpha}\ \frac{\delta ({}^{k-1}A_2)}{\delta\alpha}\right], \qquad (2.47)$$

$$\frac{\delta}{\delta\alpha}\begin{bmatrix} {}^{k-1}A_1 \\ {}^{k-1}A_2 \end{bmatrix} = \begin{bmatrix} \dfrac{\delta ({}^{k-1}A_1)}{\delta\alpha} \\ \dfrac{\delta ({}^{k-1}A_2)}{\delta\alpha} \end{bmatrix}. \qquad (2.48)$$

Из правил (2.23) для алгебраической производной структурных чисел по сумме и произведению одноэлементных структурных чисел имеем

$$\frac{\partial ({}^{k_1}A)}{\partial ({}^{k_2}A_1 + {}^{k_2}A_2)} = \left[\frac{\partial ({}^{k_1}A)}{\partial ({}^{k_2}A_1)}\ \frac{\partial ({}^{k_1}A)}{\partial ({}^{k_2}A_2)}\right], \qquad (2.49)$$

$$\frac{\partial\,({}^{k_1}A)}{\partial\,({}^{k_2}A_1\,{}^{k_2}A_2)} = \frac{\partial^2\,({}^{k_1}A)}{\partial\,({}^{k_2}A_1)\,\partial\,({}^{k_2}A_2)}, \tag{2.50}$$

$$\frac{\partial\,({}^{k}A)}{\partial\,({}^{r}B)} \overset{c}{=} \frac{\partial A}{\partial B} = (A^d B)^d, \quad A \overset{e}{=} {}^{k}A, \quad B \overset{e}{=} {}^{r}B. \tag{2.51}$$

Эти зависимости представляют собой обобщения формул (2.24) и (2.25). Обобщения формул (2.19) и (2.20) в виде

$$\frac{\partial\,[{}^{k_1}A_1\,{}^{k_1}A_2]}{\partial\,({}^{k_2}A)} = \left[\frac{\partial\,({}^{k_1}A_1)}{\partial\,({}^{k_2}A)}\ \frac{\partial\,({}^{k_1}A_2)}{\partial\,({}^{k_2}A)}\right], \tag{2.52}$$

$$\frac{\partial}{\partial\,({}^{k_2}A)}\begin{bmatrix} {}^{k_1}A_1 \\ {}^{k_1}A_2 \end{bmatrix} \begin{bmatrix} \dfrac{\partial\,({}^{k_1}A_1)}{\partial\,({}^{k_2}A)} & {}^{k_1}A_1 \\ {}^{k_1}A_2 & \dfrac{\partial\,({}^{k_1}A_2)}{\partial\,({}^{k_2}A)} \end{bmatrix} \tag{2.53}$$

также справедливы.

Пример 2 3.

$$\frac{\partial \begin{bmatrix} {}^{k}A_1 \\ {}^{k}A_2 \end{bmatrix}}{\partial \begin{bmatrix} {}^{r}A_{11} & {}^{r}A_{12} \\ {}^{r}A_{21} & {}^{r}A_{22} \end{bmatrix}} = \frac{\partial}{\partial\,({}^{r}A_{21})}\begin{bmatrix} \dfrac{\partial\,({}^{k}A_1)}{\partial\,({}^{r}A_{11})} & {}^{k}A_1 \\ {}^{k}A_2 & \dfrac{\partial\,({}^{k}A_2)}{\partial\,({}^{r}A_{11})} \end{bmatrix} + \frac{\partial}{\partial\,({}^{r}A_{22})}\begin{bmatrix} \dfrac{\partial\,({}^{k}A_1)}{\partial\,({}^{r}A_{12})} & {}^{k}A_1 \\ {}^{k}A_2 & \dfrac{\partial\,({}^{k}A_2)}{\partial\,({}^{r}A_{12})} \end{bmatrix} =$$

$$= \begin{bmatrix} \dfrac{\partial^2\,({}^{k}A_1)}{\partial\,({}^{r}A_{11})\,\partial\,({}^{r}A_{21})} & \dfrac{\partial\,({}^{k}A_1)}{\partial\,({}^{r}A_{11})} & \dfrac{\partial\,({}^{k}A_1)}{\partial\,({}^{r}A_{21})} & {}^{k}A_1 & \dfrac{\partial^2\,({}^{k}A_1)}{\partial\,({}^{r}A_{12})\,\partial\,({}^{r}A_{22})} & \dfrac{\partial\,({}^{k}A_1)}{\partial\,({}^{r}A_{12})} \\ {}^{k}A_2 & \dfrac{\partial\,({}^{k}A_2)}{\partial\,({}^{r}A_{21})} & \dfrac{\partial\,({}^{k}A_2)}{\partial\,({}^{r}A_{11})} & \dfrac{\partial^2\,({}^{k}A_2)}{\partial\,({}^{r}A_{11})\,\partial\,({}^{r}A_{11})} & {}^{k}A_2 & \dfrac{\partial\,({}^{k}A_2)}{\partial\,({}^{r}A_{22})} \\ & & & \dfrac{\partial\,({}^{k}A_1)}{\partial\,({}^{r}A_{22})} & {}^{k}A_1 & \\ & & & \dfrac{\partial\,({}^{k}A_2)}{\partial\,({}^{r}A_{12})} & \dfrac{\partial^2\,({}^{k}A_2)}{\partial\,({}^{r}A_{12})\,\partial\,({}^{r}A_{22})} & \end{bmatrix}.$$

6.10.2. Геометрическое изображение структурного числа *k*-й категории

Обозначим $f^{\%}$ класс подобных графов, представляющих собой геометрическое изображение структурного числа первой категории A, определенного на конечном множестве элементов α_{ij}, и $f^{\%}$ — функцию гомеоморфного преобразования

$$\tilde{f}\colon \tilde{\Gamma} \to A. \tag{2.54}$$

Пусть **A** — множество всех равных структурных чисел категории $k>1$.

$$\mathbf{A} = \{{}^{k}A \mid {}^{k}A \overset{e}{=} A\}, \quad k = 2, 3, \ldots . \qquad (2.55)$$

Преобразование $\tilde{\varphi}$ класса $\tilde{F}$ подобных графов в множество **A** структурных чисел определим как

$$(\tilde{\varphi}: \tilde{F} \to \mathbf{A}) \equiv (\tilde{f}: \tilde{F} \to A), \quad A \overset{e}{=} {}^{k}A. \qquad (2.56)$$

Это означает, что геометрическое изображение структурного числа k-й категории ${}^{k}A$ есть граф, представляющий собой изображение замещающего структурного числа A $(A \overset{e}{=} {}^{k}A)$, а также что данный граф служит изображением и всех других чисел категории $k>1$, которые имеют то же самое замещающее число.

Известно, что преобразование пространства конечных структурных чисел первой категории в пространство конечных графов не является непрерывной функцией, поэтому не каждое структурное число k-й. категории имеет геометрическое изображение. Очевидно, для существования геометрического изображения структурного числа k-й категории необходимо и достаточно существование геометрического изображения его замещающего числа

Для существования геометрического изображения структурного числа недостаточно существования изображений его элементов. Например, структурное число второй категории

$${}^{2}A = \begin{bmatrix} \{1\} & \{4\ 5\} \\ \begin{bmatrix} 1 & 2 & 1 \\ 2 & 3 & 3 \end{bmatrix} & \begin{bmatrix} 2 \\ 5 \end{bmatrix} \end{bmatrix}$$

не имеет геометрического изображения, несмотря на то что все его элементы обладают такими изображениями, так как его замещающее число

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \\ 3 & 5 \end{bmatrix}$$

не имеет изображения. Наоборот, структурное число второй категории

$${}^{2}A = \left[\begin{bmatrix} 1 & 2 \\ 2 & 4 \\ 3 & 5 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 4 & 5 \end{bmatrix} \right]$$

обладает геометрическим изображением, так как его замещающее число

$$A = \begin{bmatrix} 1 & 1 \\ 2 & 3 \\ 3 & 4 \end{bmatrix}$$

имеет изображение, хотя оба его элемента таких изображений не имеют. С практической точки зрения интересны структурные числа k-й категории ${}^{k}A$, имеющие геометрическое изображение и построенные из структурных чисел, которые также имеют геометрическое изображение. В этом случае геометрическое изображение числа ${}^{k}A$ может рассматриваться как иерархическое изображение, состоящее из подызображений, которые в свою очередь тоже могут быть иерархическими изображениями.

6.11. Полные структурные числа

Структурные числа второй или высшей категорий можно привести к замещающим структурным числам первой категории, применяя операции алгебры структурных чисел над их элементами. При этом должны быть известны структурные числа блоков графа, т. е. их структура. Однако в практических применениях, например при анализе или синтезе электрических схем, не всегда возможно или необходимо знать структуру блоков графа. Иногда удобно рассматривать блок-схему цепи, не углубляясь в структуру отдельных ее блоков, называемых многополюсниками. Например, эти многополюсники могут представлять собой «черные ящики» с выделенными полюсами (зажимами), в которых могут находиться неизвестные электрические цепи с индуктивными или емкостными связями, с распределенными или сосредоточенными параметрами и т. д. В таком случае достаточно измерить входные величины многополюсника, например напряжения на его зажимах или входные импедансы, или задать эти величины при исследовании системы. При определении структурного числа блок-графа по правилам рассматриваемой ранее алгебры структурных чисел необходимо иметь сведения о структуре отдельных блоков графа. Эти сведения не нужны, если применить видоизмененную алгебру, основанную на операциях, подобных операциям обычной алгебры, элементы которой

назовем *полными структурными числами.* Принципы этой алгебры будут изложены ниже.

Неупорядоченный набор элементов x, не обязательно различных, удовлетворяющих некоторой функции $\Phi(x)$, обозначим

$$X = \langle x \mid \Phi(x) \rangle. \tag{2.57}$$

Через $r_x(x)$ обозначим число, определяющее, сколько раз элемент x встречается в системе X. Рассмотрим две системы X и Y:

$$X = \langle x \mid \Phi(x) \rangle, \quad Y = \langle y \mid \Phi(y) \rangle. \tag{а}$$

Обозначим

$$(X \subset_r Y) \Leftrightarrow \bigwedge_{x \in X} \bigvee_{y \in Y} \{[r_X(x) \leqslant r_Y(y)] \wedge (x = y)\}. \tag{2.58}$$

Две системы вида (а) будем считать равными, если

$$(X = Y) \Leftrightarrow (X \subset_r Y) \wedge (Y \subset_r X). \tag{2.59}$$

Суммой систем X и Y вида (а) назовем систему, содержащую все элементы обеих этих систем:

$$S = (X \cup_r Y) \stackrel{\text{df}}{=} \langle s \mid (s \in X) \vee (s \in Y) \wedge \\ \wedge (r_S(s) = r_X(s) + r_Y(s) \rangle. \tag{2.60}$$

Пересечением систем X и Y вида (а) назовем систему, определяемую следующим образом:

$$Z = (X \cap_r Y) \stackrel{\text{df}}{=} \langle z \mid (z \in X) \wedge (z \in Y) \wedge (r_Z(z) = \\ = \min [r_X(z),\ r_Y(z)]) \rangle. \tag{2.61}$$

Введем определение полного структурного числа.

Определение. Полным структурным числом A называется неупорядоченная система наборов вида (а)

$$\mathcal{A} = \langle a_k \mid a_k = \langle \alpha_{ik} \mid \alpha_{ik} \in N \rangle_{\substack{i=1,2,\cdots,m \\ k=1,2,\cdots,n}}, \tag{2.62}$$

где N — множество натуральных чисел при следующих условиях:

$$\begin{aligned} &1.\ (\mathcal{A} = \mathcal{B}) \Leftrightarrow (\mathcal{A} \subset_r \mathcal{B}) \wedge (\mathcal{B} \subset_r \mathcal{A}), \\ &2.\ \mathcal{A} + \mathcal{B} = \mathcal{A} \cup_r \mathcal{B}, \\ &3.\ \mathcal{A}\mathcal{B} = \langle a \cup_r b \mid (a \in \mathcal{A}) \wedge (b \in \mathcal{B}) \rangle. \end{aligned} \tag{2.63}$$

Из этого определения следуют известные соотношения элементарной алгебры.

$$\left.\begin{aligned} \mathcal{A}+\mathcal{B}&=\mathcal{B}+\mathcal{A},\\ \mathcal{A}\mathcal{B}&=\mathcal{B}\mathcal{A}\end{aligned}\right\}\ \text{(коммутативность)},$$

$$\left.\begin{aligned} \mathcal{A}+(\mathcal{B}+\mathcal{C})&=(\mathcal{A}+\mathcal{B})+\mathcal{C},\\ \mathcal{A}(\mathcal{B}\mathcal{C})&=(\mathcal{A}\mathcal{B})\mathcal{C}\end{aligned}\right\}\ \text{(ассоциативность)}. \tag{2.64}$$

$$\mathcal{A}(\mathcal{B}+\mathcal{C})=\mathcal{A}\mathcal{B}+\mathcal{A}\mathcal{C}\quad\text{(дистрибутивность)}.$$

Нетрудно заметить, что *модулем сложения полных структурных чисел* служит число $\langle\ \rangle$, которое представляет собой пустую систему, а *модулем умножения* — число $\langle\phi\rangle$, содержащее одно и только одно пустое число *а.* Для произвольного числа A

$$\left.\begin{aligned} \mathcal{A}+\langle\ \rangle&=\mathcal{A},\\ \mathcal{A}\langle\phi\rangle&=\mathcal{A}.\end{aligned}\right\} \tag{2.65}$$

В соответствии с этим число $\langle\ \rangle$ обозначим через 0, а число $\langle\phi\rangle$— через 1:

$$\langle\ \rangle=0,\qquad \langle\phi\rangle=1. \tag{2.66}$$

Уравнение

$$\mathcal{A}+\mathcal{B}=0 \tag{2.67}$$

имеет решение

$$\mathcal{A}=-\mathcal{B}. \tag{2,68}$$

Число — $\mathcal{B}$ будем называть *отрицательным полным структурным числом.*

Разность чисел $\mathcal{A}-\mathcal{B}$ определяем как операцию, обратную по отношению к сложению, т. е.

$$\mathcal{C}+\mathcal{B}=\mathcal{A}\Rightarrow\mathcal{C}=\mathcal{A}-\mathcal{B}. \tag{2.69}$$

Эту разность можно отыскать, если исключить из числа А все системы *а,* равные системам *b* числа $\mathcal{B}$.

Следствие. Разность полных структурных чисел $\mathcal{A}-\mathcal{B}$ существует тогда и только тогда, когда

$$\mathcal{B}\subset_r\mathcal{A}.$$

В противном случае разность двух структурных чисел может быть лишь упрощена путем вычитания из обоих чисел разности их пересечения $\mathcal{A}\cap_r\mathcal{B}$.

Выражение $\mathcal{A}-\mathcal{B}$ в общем случае имеет следующие свойства:

$$(\mathcal{A}-\mathcal{B}=\mathcal{C}-\mathcal{D}) \Leftrightarrow (\mathcal{A}+\mathcal{D}=\mathcal{C}+\mathcal{B}),$$
$$(\mathcal{A}-\mathcal{B})+(\mathcal{C}-\mathcal{D})=(\mathcal{A}+\mathcal{C})-(\mathcal{B}+\mathcal{D}), \qquad (2,70)$$
$$(\mathcal{A}-\mathcal{B})(\mathcal{C}-\mathcal{D})=(\mathcal{A}\mathcal{C}+\mathcal{B}\mathcal{D})-(\mathcal{A}\mathcal{D}+\mathcal{B}\mathcal{C}).$$

Если обозначить

$$\underbrace{\mathcal{A}+\mathcal{A}+\ldots+\mathcal{A}}_{N}=N\mathcal{A}, \qquad (2.71)$$

$$\underbrace{\mathcal{A}\mathcal{A}\ldots\mathcal{A}}_{N}=\mathcal{A}^{N}, \qquad (2.72)$$

где N — произвольное натуральное число, легко заметить, что для двух произвольных натуральных чисел N и M справедливы следующие соотношения:

$$N\mathcal{A}+M\mathcal{A}=(N+M)\mathcal{A},$$
$$M(N\mathcal{A})=(MN)\mathcal{A},$$
$$\mathcal{A}^{N}\mathcal{A}^{M}=\mathcal{A}^{N+M}, \qquad (2.73)$$
$$(\mathcal{A}^{N})^{M}=\mathcal{A}^{(NM)},$$

а также

$$(\mathcal{A}+\mathcal{B})^{2}=\mathcal{A}^{2}+2\mathcal{A}\mathcal{B}+\mathcal{B}^{2},$$
$$(\mathcal{A}+\mathcal{B})(\mathcal{A}-\mathcal{B})=\mathcal{A}^{2}-\mathcal{B}^{2}. \qquad (2.74)$$

Допустим, что полное структурное число $\mathcal{A}$ можно представить в виде произведения

$$\mathcal{A}=\mathcal{P}_1\mathcal{P}_2\ldots\mathcal{P}_n. \qquad (2.75)$$

Числа $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_n$ назовем *делителями* числа $\mathcal{A}$. Полное структурное число P имеющее только два делителя 1 и P назовем *простым полным структурным числом.* Заметим, что каждое полное структурное число P, представляющее собой систему одноэлементных наборов, имеет только два делителя: 1 и P , т. е. будет простым полным структурным числом.

Делители, представляющие собой простые полные структурные числа, будем называть *простыми сомножителями числа* $\mathcal{A}$.

Если для двух данных полных структурных чисел $\mathcal{A}$ и P существует такое полное структурное число X, что

$$\mathcal{B}\mathcal{X}=\mathcal{A}, \qquad (2.76.)$$

то можно написать

$$\mathcal{X} = \frac{\mathcal{A}}{\mathcal{B}} \quad \text{или} \quad \mathcal{X} = \mathcal{A} : \mathcal{B} \tag{2.77}$$

т. е.

$$(\mathcal{B}\mathcal{X} = \mathcal{A}) \Leftrightarrow \left(\mathcal{X} = \frac{\mathcal{A}}{\mathcal{B}}\right), \quad \mathcal{B} \neq 0. \tag{2.78}$$

Условием существования *частного* X, кроме, $\mathcal{B} \neq 0$, будет требование, чтобы множество простых сомножителей числа P было подмножеством простых сомножителей числа $\mathcal{A}$, т. е.

$$\{\mathcal{P}'_1, \mathcal{P}'_2, \ldots, \mathcal{P}'_m\} \subset \{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_n\}, \tag{2.79}$$

где

$$\mathcal{A} = \mathcal{P}_1\mathcal{P}_2 \ldots \mathcal{P}_n, \quad \mathcal{B} = \mathcal{P}'_1\mathcal{P}'_2 \ldots \mathcal{P}'_m.$$

Введем следующее обобщение.

Пару чисел $(\mathcal{A}, \mathcal{B})$ будем записывать в виде $\mathcal{A}/\mathcal{B}$, если

$$\left.\begin{aligned} &\left(\frac{\mathcal{A}}{\mathcal{B}} = \frac{\mathcal{C}}{\mathcal{D}}\right) \Leftrightarrow (\mathcal{A}\mathcal{D} = \mathcal{B}\mathcal{C}), \\ &\frac{\mathcal{A}}{\mathcal{B}} + \frac{\mathcal{C}}{\mathcal{D}} = \frac{\mathcal{A}\mathcal{D} + \mathcal{B}\mathcal{C}}{\mathcal{B}\mathcal{D}}, \\ &\frac{\mathcal{A}}{\mathcal{B}}\,\frac{\mathcal{C}}{\mathcal{D}} = \frac{\mathcal{A}\mathcal{C}}{\mathcal{B}\mathcal{D}}. \end{aligned}\right\} \mathcal{B} \neq 0, \ \mathcal{D} \neq 0. \tag{2.80}$$

Для полных структурных чисел справедливы тождества

$$\frac{\mathcal{A}^M}{\mathcal{A}^N} = \mathcal{A}^{M-N}, \quad \mathcal{A}^0 = 1. \tag{2.81}$$

В соответствии с определением (2.62) полное структурное число $\mathcal{A}$ можно также записать в виде таблицы

$$\mathcal{A} = \left\{\begin{matrix} \alpha_{11}\alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21}\alpha_{22} & \ldots & \alpha_{2n} \\ \ldots & \ldots & \ldots \\ \alpha_{m1}\alpha_{m2} & \ldots & \alpha_{mn} \end{matrix}\right\} \tag{2.82}$$

и рассматривать как неупорядоченную систему столбцов a_k (не обязательно различных), состоящих из неупорядоченных элементов (также не обязательно различных)

где

$$\left.\begin{aligned} \mathcal{A} &= \langle a_1 a_2 \ \dots \ a_n \rangle, \\ a_k &= \begin{Bmatrix} \alpha_{1k} \\ \alpha_{2k} \\ \vdots \\ \alpha_{mk} \end{Bmatrix}. \end{aligned}\right\} \tag{2.83}$$

Равенство, а также способ получения суммы и произведения полных структурных чисел, записанных в виде таблицы, иллюстрируют следующие примеры':

$$\begin{Bmatrix} 1 & 2 & 2 & 2 \\ 2 & 3 & 3 & 4 \\ 3 & 4 & 4 & 5 \end{Bmatrix} = \begin{Bmatrix} 1 & 3 & 4 & 2 \\ 2 & 2 & 5 & 3 \\ 3 & 4 & 2 & 4 \end{Bmatrix},$$

$$\begin{Bmatrix} 1 & 3 \\ 2 & 4 \end{Bmatrix} + \begin{Bmatrix} 1 & 4 \\ 2 & 5 \end{Bmatrix} = \begin{Bmatrix} 1 & 3 & 1 & 4 \\ 2 & 4 & 2 & 5 \end{Bmatrix},$$

$$\begin{Bmatrix} 1 & 3 \\ 2 & 4 \end{Bmatrix} \cdot \begin{Bmatrix} 1 & 4 \\ 2 & 5 \end{Bmatrix} = \begin{Bmatrix} 1 & 1 & 4 & 4 \\ 2 & 2 & 5 & 5 \\ 1 & 3 & 1 & 3 \\ 2 & 4 & 2 & 4 \end{Bmatrix}.$$

Заметим, что для полных структурных чисел справедлива следующая теорема.

Теорема. *Полное структурное число можно всегда записать в канонической форме*

$$\mathcal{A} = \sum_{k=1}^{n} \coprod_{i=1}^{m} \langle \alpha_{ik} \rangle. \tag{2.84}$$

Действительно, согласно принятым определениям операций, имеем

$$\begin{Bmatrix} \alpha_{11}\alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21}\alpha_{22} & \dots & \alpha_{2n} \\ \dots & \dots & \dots \\ \alpha_{m1}\alpha_{m2} & \dots & \alpha_{mn} \end{Bmatrix} = \sum_{k=1}^{n} \begin{Bmatrix} \alpha_{1k} \\ \alpha_{2k} \\ \vdots \\ \alpha_{mk} \end{Bmatrix} = \sum_{k=1}^{n} \prod_{i=1}^{m} \langle \alpha_{ik} \rangle.$$

Обозначая

$$\langle \alpha_{i_k} \rangle = a_{i_k}, \tag{2.85}$$

формулу (2.84) можно записать в виде

$$\mathcal{A} = \sum_{k=1}^{n} \prod_{i=1}^{m} a_{ik}. \qquad (2.86)$$

Выражение a_{ik} в формуле (2.85) назовем *полной структурной единицей.*

Если полное структурное число A имеет все столбцы, такие же, как структурное число *A,* то можно написать

$$\mathcal{A} \overset{s}{=\!=} A. \qquad (2.87)$$

Это равенство симметрично, т. е.

$$(\mathcal{A} \overset{s}{=\!=} A) \Leftrightarrow (A \overset{s}{=\!=} \mathcal{A}) \qquad (2.88)$$

и, кроме того, имеют место следующие зависимости:

$$\left.\begin{aligned} &(\mathcal{A} \overset{s}{=\!=} A_1) \wedge (A_1 = A_2) \Rightarrow (\mathcal{A} \overset{s}{=\!=} A_2), \\ &(A \overset{s}{=\!=} \mathcal{A}_1) \wedge (\mathcal{A}_1 = \mathcal{A}_2) \Rightarrow (A \overset{s}{=\!=} \mathcal{A}_2). \end{aligned}\right\}, \qquad (2.89)$$

Применяя соотношение

$$(k_A \overset{s}{=\!=} \mathcal{A}) \Leftrightarrow \{(A \overset{s}{=\!=} \mathcal{A}) \wedge (A \overset{e}{=\!=} k_A)\}$$

и предполагая, что $A \overset{s}{=} \mathcal{A}$, запишем в случае необходимости структурное число *A* графа *Г* в виде полного структурного числа A.

Определим алгебраическую производную полного структурного числа A по элементу α.

Определение. Алгебраическая производная полного структурного числа

$$\mathcal{A} = \langle a_k \rangle_{k=1,\,2,\,\ldots,\,n}, \quad a_k = \langle \alpha_{ik} \rangle_{i=1,\,2,\,\ldots,\,m}$$

по элементу α представляет собой полное структурное число $\partial\mathcal{A}/\partial\alpha$, определенное следующим образом:

$$\frac{\partial \mathcal{A}}{\partial \alpha_{ik}} = \langle a'_k \,|\, (\langle a'_k \rangle = \langle a_k \rangle - \langle \alpha_{ik} \rangle) \wedge [r_{\partial\mathcal{A}/\partial\alpha_{ik}}(a'_k) = r_{\mathcal{A}}(a_k) \cdot r_{a_k}(\alpha_{ik})] \rangle_{k=1,\,2,\,\ldots,\,n}. \qquad (2.90)$$

Алгебраическая производная полного структурного числа по элементу α находится по правилам дифференцирования алгебраических многочленов. Эту производную можно также рассчитать путем дифференцирования полного структурного числа, записанного в канонической форме (2.84) или (2.86).

Пример.

$$\frac{\partial}{\partial 2}\begin{Bmatrix}1&1&2&1&1&1\\2&2&2&2&2&1\\3&2&2&2&2&4\end{Bmatrix}=\frac{\partial}{\partial 2}\{\langle 1\rangle\langle 2\rangle\langle 3\rangle+3\langle 1\rangle\langle 2\rangle^2+\langle 2\rangle^3+\langle 1\rangle^2\langle 4\rangle\}=$$

$$=\langle 1\rangle\langle 3\rangle+6\langle 1\rangle\langle 2\rangle+3\langle 2\rangle^2=\begin{Bmatrix}1&1&1&1&1&1&1&2&2&2\\3&2&2&2&2&2&2&2&2&2\end{Bmatrix}.$$

Очевидно, если $\mathscr{A}\overset{s}{=}A$, то

$$\frac{\partial\mathscr{A}}{\partial\alpha}\overset{s}{=}\frac{\partial A}{\partial\alpha}. \tag{2.91}$$

Для алгебраической производной суммы, произведения и частного полных структурных чисел справедливы те же соотношения, что и при дифференцировании алгебраических многочленов, а именно

$$\left.\begin{aligned}&\frac{\partial}{\partial\alpha}(\mathscr{A}+\mathscr{B})=\frac{\partial\mathscr{A}}{\partial\alpha}+\frac{\partial\mathscr{B}}{\partial\alpha},\\&\frac{\partial}{\partial\alpha}(\mathscr{A}\mathscr{B})=\mathscr{A}\frac{\partial\mathscr{B}}{\partial\alpha}+\mathscr{B}\frac{\partial\mathscr{A}}{\partial\alpha},\\&\frac{\partial}{\partial\alpha}\left(\frac{\mathscr{A}}{\mathscr{B}}\right)=\frac{\mathscr{B}\dfrac{\partial\mathscr{A}}{\partial\alpha}-\mathscr{A}\dfrac{\partial\mathscr{B}}{\partial\alpha}}{\mathscr{B}^2},\quad \mathscr{B}\neq 0.\end{aligned}\right\} \tag{2.92}$$

С целью упрощения формы записи полных структурных чисел будем применять *операцию переноса нижних индексов* в соответствии со следующими правилами:

$$\left.\begin{aligned}&\frac{\partial\mathscr{A}}{\partial\alpha}=\mathscr{A}_\alpha,\\&\frac{\partial\mathscr{A}_k}{\partial\alpha_k}=\mathscr{A}_{k\alpha},\\&A^d[\alpha]\overset{s}{=}\mathscr{A}^d_\alpha,\quad A^d\overset{s}{=}\mathscr{A}^d,\\&A^d_k[\alpha_k]\overset{s}{=}\mathscr{A}^d_{k\alpha},\quad A^d_k\overset{s}{=}\mathscr{A}^d_k.\end{aligned}\right\} \tag{2.93}$$

Определим обратную алгебраическую производную полного структурного числа.

Определение. Обратной производной полного структурного числа

$$\mathscr{A}=\langle a_k\rangle_{k=1,\,2,\,\ldots,\,n},\quad a_k=\langle\alpha_{ik}\rangle_{i=1,\,2,\,\ldots,\,m}$$

по элементу α называется полное структурное число $\delta\mathscr{A}/\delta\alpha$, определенное как

$$\frac{\delta \mathcal{A}}{\delta \alpha_{ik}} \stackrel{\mathrm{df}}{=} \mathcal{A} - \langle a_k \mid \alpha_{ik} \in a_k \rangle_{k=1,\,2,\,\ldots,\,n}. \qquad (2.94)$$

Это означает, что обратную производную $\delta \mathcal{A}/\delta\alpha$ составляют лишь те столбцы числа A , которые не содержат элемента α.

Пример.

$$\frac{\delta}{\delta 2}\begin{Bmatrix} 1\,1\,2\,1\,1\,1 \\ 2\,2\,2\,2\,2\,1 \\ 3\,2\,2\,2\,2\,4 \end{Bmatrix} = \begin{Bmatrix} 1 \\ 1 \\ 4 \end{Bmatrix}.$$

Очевидно, если $\mathcal{A} \stackrel{s}{=} A$, то

$$\frac{\delta \mathcal{A}}{\delta \alpha} \stackrel{s}{=} \frac{\delta A}{\delta \alpha}. \qquad (2.95)$$

Заметим, что для обратной производной суммы, произведения и частного полных структурных чисел справедливы следующие соотношения:

$$\left.\begin{aligned} \frac{\delta}{\delta\alpha}(\mathcal{A}+\mathcal{B}) &= \frac{\delta \mathcal{A}}{\delta\alpha} + \frac{\delta \mathcal{B}}{\delta\alpha}, \\ \frac{\delta}{\delta\alpha}(\mathcal{A}\mathcal{B}) &= \frac{\delta \mathcal{A}}{\delta\alpha}\frac{\delta \mathcal{B}}{\delta\alpha}, \\ \frac{\delta}{\delta\alpha}\left(\frac{\mathcal{A}}{\mathcal{B}}\right) &= \frac{\delta\mathcal{A}/\delta\alpha}{\delta\mathcal{B}/\delta\alpha}, \quad \mathcal{B} \neq 0, \quad \frac{\delta \mathcal{B}}{\delta\alpha} \neq 0. \end{aligned}\right\} \qquad (2.96)$$

Легко доказать, что между действиями над структурными числами и действиями над полными структурными числами имеют место следующие соотношения:

1. $A + B \stackrel{s}{=} \mathcal{A} + \mathcal{B} - 2(\mathcal{A} \cap_r \mathcal{B}),\ A \stackrel{s}{=} \mathcal{A},\ B \stackrel{s}{=} \mathcal{B}.$ (2.97)

Выражение $2(\mathcal{A} \cap_r \mathcal{B})$ называется *дефектом суммы* структурных чисел *A* и *B*. Если

$$\mathcal{A} \cap_r \mathcal{B} = 0$$

(если $A \stackrel{s}{=} \mathcal{A}$ и $B \stackrel{s}{=} \mathcal{B}$, то пересечение $\mathcal{A} \cap_r \mathcal{B}$ запишем в виде $\mathcal{A} \cap \mathcal{B}$),

т. е.

$$A \cap B = 0,$$

то говорим, что сумма чисел *A* и *B* не обладает дефектом суммы, т. е. $A + B \stackrel{s}{=} \mathcal{A} + \mathcal{B}$.

2. $AB \stackrel{s}{=} \mathcal{A}\mathcal{B} - \mathcal{D},$ (2.98)

где

$$\mathscr{D} = \sum \langle a_0 \rangle + \sum_i \left(\left\{ k_i - \frac{1}{2} [1 + (-1)^{k_i - 1}] \right\} \langle a_i \rangle \right); \tag{2.99}$$

$a_i \neq a_0$ — столбцы произведения $\mathscr{A}\mathscr{B}$, имеющие по крайней мере два идентичных элемента $r_{a_0}(\alpha) \geqslant 2$;

k_i — число идентичных столбцов a_i в произведении $\mathscr{A}\mathscr{B}$.

Полное структурное число D в выражении (2.98) называется *дефектом произведения* структурных чисел *A* и *B*. Заметим, что

$$\mathscr{D} \overset{s}{=} [\ ] = 0 \tag{2.100}$$

при условии

$$\mathscr{D} = \langle a \mid \left(\bigvee_{\alpha \in a} r_a(\alpha) > 1 \right) \vee (r_{\mathscr{D}}(a) = 2k) \rangle, \tag{2.101}$$

где k — натуральное число.

Если $\mathscr{D} = 0$, то произведение структурных чисел *A* и *B* не обладает дефектом, т. е. $AB \overset{s}{=} \mathscr{A}\mathscr{B}$.

Легко заметить, что произведение структурных чисел $A_1, A_2, \ldots, A_q$ блоков блок-графа не имеет дефекта, т. е.

$$\coprod_{i=1}^{g} A_i \overset{e}{=} \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_g \end{bmatrix} \overset{s}{=} \coprod_{i=1}^{g} \mathscr{A}_i, \quad \bigwedge_i (A_i \overset{s}{=} \mathscr{A}_i), \tag{2.102}$$

так как эти числа не имеют общих элементов.

Модуль 7.
Введение в интервальную алгебру

Введение

Излагаемый в этом и последующих модулях материал относится к актуальному направлению вычислительной математики, получившему название «интервальный анализ» или даже «интервальная математика». Интерес к этому направлению обусловлен, в первую очередь, широким применением ЭВМ для всевозможных расчетов. Если эти расчеты проводятся по традиционной схеме, то часто очень трудно, а иногда просто невозможно дать математически строгий ответ на естественный вопрос о соотношении числа, напечатанного машиной, и истинного значения вычисляемой величины. Интервальный анализ дает возможность получить такой ответ ценой увеличения времени счета.

Основная идея интервального анализа чрезвычайно проста. Вещественное число представляется в памяти ЭВМ не одним, а двумя машинными числами — оценкой снизу и оценкой сверху, образующими интервальное число. Арифметические операции над этими числами выполняются так, что если $[a_1, a_2] = [b_1, b_2] \circ [c_1, c_2]$, $b \in [b_1, b_2]$, $c \in [c_1, c_2]$, то $b \circ c_1 \in [a_1, a_2]$, где $\circ \in \{+, —, \times, /\}$. Таким образом, интервальный анализ дает возможность автоматически учитывать погрешности в задании исходных данных и погрешности, вызываемые машинными округлениями. Это создает основу для аккуратного учета погрешностей, вызываемых используемым приближенным методом вычислений.

Первой монографией по интервальному анализу была вышедшая в 1966 г. книга Р. Е. Мура, которая во многом способствовала становлению этого направления. В настоящее время общее число опубликованных работ в этой области составляет несколько тысяч. Такое обилие публикаций объясняется, в частности, тем, что практическая реализация описанной выше основной идеи интервального анализа сталкивается с большими трудностями. Например, оказалось, что алгоритм Гаусса может стать неприменимым к системе линейных уравнений с интервальными коэффициентами из-за возникающих делений на интервалы, содержащие нуль. В других случаях, когда традиционный численный метод переносился на интервальные числа, в результате вычислений получались интервалы, гарантированно содержащие истинное

значение, но столь широкие, что найденные двусторонние оценки были практически бесполезными. Выяснилось, что для успешного применения интервального анализа нужно пересмотреть весь арсенал численных методов.

Такой пересмотр и делается в настоящей работе, которая базируется на книге Г. Алефельда, Ю. Херцберга «Введение в интервальые вычисления». При этом Г. Алефельд, Ю. Херцберг не ограничиваются только описанием различных алгоритмов, но и проводят их сравнение как по достигаемой точности, так и по вычислительной сложности. Большая часть излагаемого материала посвящена задачам линейной алгебры, что неудивительно ввиду той базисной роли, которую играет линейная алгебра в численных методах. Вместе с тем вне рамок настоящей работы остались многие другие важные для приложений разделы математики, в которых интервальный анализ успешно применяется, например, обыкновенные дифференциальные уравнения. Это объясняется, прежде всего, содержанием раздела математики, который рассматривается в настоящей книге, а именно, - «Дискретная математика. Алгебры »

Интервальный анализ вышел за рамки чисто теоретического исследования и достаточно широко применяется на практике с помощью соответствующего программного обеспечения. В приложениях В и С приведены программы для интервальных вычислений на Алголе 60. За прошедшее время, с одной стороны, появились реализации интервальной арифметики на более мощных языках. С другой стороны, был разработан стандарт ANSI/IEEE на машинное представление чисел и правила выполнения операций над ними. Этот стандарт реализован как программно, так и аппаратно в микропроцессорах. Это в значительной степени облегчает программную реализацию интервальных вычислений, которые становятся ненамного более медленными, чем традиционные.

Следует обратить внимание на терминологию, которая еще полностьюне устоялась ни в мире, ни на русском языке, что надо иметь в виду при чтении работ других авторов. На ***интервальные числа*** можно смотреть двояко: ***как на способ задания вещественных чисел, которые мы знаем лишь с некоторой погрешностью, и как на самостоятельные объекты*** Это различие почти нигде не ощущается, пожалуй, его единственное проявление — это определение равенства интервалов. При первом подходе равенство $[a_1, a_2] = [b_1, b_2]$ выполняется тогда и только тогда, когда $a_1 = a_2 = b_1 = b_2$, при втором, принятом в этой книге, оно справедливо тогда и только тогда, когда $a_1 = b_1$ и $a_2 = b_2$.

Микромодуль 23
Вещественная интервальная арифметика

7.1. Основные понятия и определения

В этом и последующих модулях поле вещественных чисел будет обозначаться через $\mathbb{R}$, а строчные буквы *a, b, c, ..., x, y, z* будут использоваться для обозначения его элементов. Подмножество *A* множества $\mathbb{R}$, такое что

$$A = [a_1, a_2] = \{I \mid a_1 \leqslant I \leqslant a_2,\ a_1, a_2 \in \mathbb{R}\},$$

будет называться замкнутым вещественным интервалом, или просто интервалом, если это не сможет вызвать недоразумения. Впоследствии в некоторых случаях, чтобы избежать путаницы, мы будем обозначать границы интервала *A* через

$$i(A) = a_1 \quad \text{и} \quad s(A) = a_2.$$

Множество всех замкнутых вещественных интервалов обозначим через $I(\mathbb{R})$, а прописные буквы $A, B, C, \ldots, X, Y, Z$ зарезервируем для обозначения его элементов. Всякое вещественное число *x* из $\mathbb{R}$ может считаться особым элементом из $I(\mathbb{R})$, имеющим вид $[x, x]$; чаще всего мы будем называть его точечным интервалом.

Определение 1. Два интервала $A = [a_1, a_2]$ и $B = [b_1, b_2]$ называются равными (записывается: *A* =*B*), если они равны в теоретико-множественном смысле.

Из этого определения непосредственно следует, что

$$A = B \Leftrightarrow a_1 = b_1, \quad a_2 = b_2.$$

Отношение равенства между двумя элементами из $I(\mathbb{R})$ рефлексивно, симметрично и транзитивно.

Теперь мы можем обобщить арифметику вещественных чисел, введя операции над элементами из $I(\mathbb{R})$.

Определение 2. Пусть $* \in \{+, -, \cdot, :\}$ — бинарная операция на множестве вещественных чисел. Если $A, B \in I(\mathbb{R})$, то

$$A * B = \{z = a * b \mid a \in A,\ b \in B\} \qquad (1)$$

определяет бинарную операцию на $I(\mathbb{R})$.

В определении предполагается, что в случае деления $0 \notin B$, и в дальнейшем это явно указываться не будет. Заметим также, что символы операций на множествах *I*(R) и R совпадают. Это не должно

вызывать затруднений, поскольку из контексга всегда ясно, к чему применяется операция: к вещественным числам или интервалам.

Результат операции над интервалами $A = [a_1, a_2]$ и $B = [b_1, b_2]$ может быть получен явно с помощью формул

$$\begin{cases} A + B = [a_1 + b_1,\ a_2 + b_2], \\ A - A = [a_1 - b_2,\ a_2 - b_1] = A + [-1,\ -1] \cdot B, \\ A \cdot B = [\min\{a_1b_1,\ a_1b_2,\ a_2b_1,\ a_2b_2\}, \\ \qquad\qquad \max\{a_1b_1,\ a_1b_2,\ a_2b_1,\ a_2b_2\}], \\ A : B = [a_1a_2] \cdot [1/b_2,\ 1/b_1]. \end{cases} \tag{2}$$

Их обоснованием служит тот факт, что $z = f(x,\ y) = x * y$, где $* \in \{+,\ -,\ \cdot,\ :\}$ — непрерывная функция на компактном множестве. Следовательно, *f(x, y)* принимает как наименьшее и наибольшее значения, так и все прочие значения между ними. Таким образом, $A * B$— также замкнутый вещественный интервал. Теперь понятно, что (2) — это формулы для вычисления наименьшего и наибольшего значений *f(x, y).* Из сказанного следует замкнутость множества $I(\mathbb{R})$ относительно введенных таким образом операций, а также изоморфизм между вещественными числами *x, y, ...* и интервалами $[x,\ x]$, $[y,\ y]$, ... Поэтому всюду далее операция $[x,\ x] * A$, в которой участвуют точечный интервал $[x,\ x]$ и произвольный интервал *A,* будет записываться в упрощенной форме $x * A$. Кроме того, мы часто будем опускать знак умножения.

Набор операций вида (1) может быть дополнен другими традиционными, в основном унарными операциями над интервалами.

Определение 3. Если *r(x)*— непрерывная унарная операция на R, то

$$r(X) = [\min_{x \in X} r(x),\ \max_{x \in X} r(x)]$$

определяет соответствующую ей операцию на $I(\mathbb{R})$.

Примерами таких унарных операций могут служить

$$X^k (k \in \mathbb{R}),\quad e^X,\quad \ln X,\quad \sin X,\quad \cos X \text{ и т. д.}$$

Соберем теперь вместе наиболее важные свойства операций на $I(\mathbb{R})$.

Теорема 4. *Пусть* $A,\ B,\ C \in I(\mathbb{R})$. *Тогда*

$$A + B = B + A,\quad A \cdot B = B \cdot A \ (\textit{коммутативность});$$
$$(A + B) + C = A + (B + C), \tag{3}$$

$$(A \cdot B) \cdot C = A \cdot (B \cdot C) \text{ (ассоциативность);} \tag{4}$$

$X = [0, 0]$ *и* $Y = [1, 1]$ *—единственные нейтральные элементы соответственно сложения и умножения, т. е.*

$$A = X + A = A + X \quad \text{для всех} \quad A \in I(\mathbb{R}) \Leftrightarrow X = [0, 0],$$
$$A = Y \cdot A = A \cdot Y \quad \text{для всех} \quad A \in I(\mathbb{R}) \Leftrightarrow Y = [1, 1]; \tag{5}$$

I(R) *не имеет делителей нуля;* (6)

произвольный элемент $A = [a_1, a_2] \in I(\mathbb{R})$, *у которого* $a_1 \neq a_2$, *не имеет обратного ни по сложению, ни по умножению. Тем не менее,*

$$0 \in A - A \quad \text{и} \quad 1 \in A : A; \tag{7}$$

$$A(B + C) \subseteq AB + AC \text{ (субдистрибутивность),}$$

$$a(B + C) = aB + aC, \quad \text{где} \quad a \in \mathbb{R}, \tag{8}$$

$$A(B + C) = AB + AC, \quad \text{где} \quad bc \geqslant 0 \quad \text{для всех} \quad b \in B \quad \text{и} \quad c \in C.$$

Доказательство. (3): Пусть $* \in \{+, \cdot\}$. Тогда

$$A * B = \{z = a * b \mid a \in A, \ b \in B\}$$
$$= \{z = b * a \mid b \in B, \ a \in A\} = B * A$$

(4): Пусть $* \in \{+, \cdot\}$. Тогда

$$(A * B) * C = \{z = y * c \mid y \in A * B, \ c \in C\}$$
$$= \{z = (a * b) * c \mid a \in A, \ b \in B, \ c \in C\}$$
$$= \{z = a * (b * c) \mid a \in A, \ b \in B, \ c \in C\}$$
$$= \{z = a * x \mid a \in A, \ x \in B * C\} = A * (B * C).$$

(5): Необходимость доказывается тривиально. Если N и $\bar{N}$ -два нейтральных элемента сложения, то

$$N + \bar{N} = \bar{N} \quad \text{и} \quad \bar{N} + N = N.$$

Из свойства коммутативности (3) следует, что $N = \bar{N}$.

Единственность $Y = [1, 1]$, нейтрального элемента умножения, может быть показана подобным же образом.

(6): Пусть $A \cdot B = 0$, т. е.

$$A \cdot B = \{z = a \cdot b \mid a \in A, \ b \in B\} = [0, 0].$$

Из этого следует, что по крайней мере один из интервалов A и B, принадлежащих $I(\mathbb{R})$, должен быть равен [0, 0].

(7): Утверждения, которые нужно доказать, эквивалентны следующим:

$$A - B = [0,\ 0] \Rightarrow A = [a,\ a] = B,$$
$$A \cdot B = [1,\ 1] \Rightarrow A = [a,\ a], \quad B = [1/a,\ 1/a].$$

Пусть

$$A - B = \{z = a - b \mid a \in A,\ b \in B\} = [0,\ 0].$$

Отсюда следует, что $z=a - b =0$ для всех $a \in A,\ b \in B.$ Фиксируя b из B, получаем, что $a=b$ для всех a из A, т. е. $A=[b,\ b]$. Соответственно можно заключить, что $B = [a,\ a]$ и, следовательно, $a=b$. Второе утверждение доказывается подобным же образом.

Так как для a из A

$$0 = a - a \in \{z = x - y \mid x \in A,\ y \in A\},$$

то очевидно, что $0 \in A - A$. Аналогично, $1 \in A : A$, где $0 \notin A$.

(8): $$A(B + C) = \{z = a(b + c) \mid a \in A,\ b \in B,\ c \in C\}$$
$$\subseteq \{y = ab + \tilde{a}c \mid a,\ \tilde{a} \in A,\ b \in B,\ c \in C\}$$
$$= AB + AC.$$

Для того чтобы показать невыполнение равенства в общем случае, приведем один пример:

$$A = [0,\ 1], \quad B = [1,\ 1], \quad C = [-1,\ -1],$$
$$A(B + C) = [0,\ 0] \subset [-1,\ 1] = AB + AC.$$

Далее имеем

$$a(B + C) = \{z = a(b + c) \mid b \in B,\ c \in C\}$$
$$= \{z = ab + ac \mid b \in B,\ c \in C\}$$
$$= \{x = ab \mid b \in B\} + \{y = ac \mid c \in C\}$$
$$= aB + aC.$$

Доказывая последнее равенство, будем считать b_1 и c_1 неотрицательными, что не приведет к потере общности. Если $a_1 \geq 0$, то

$$A(B + C) = [a_1(b_1 + c_1),\ a_2(b_2 + c_2)]$$

и

$$AB + AC = [a_1b_1,\ a_2b_2] + [a_1c_1,\ a_2c_2] = [a_1(b_1 + c_1),\ a_2(b_2 + c_2)],$$

т. е. для этого случая утверждение доказано,

Случай $a_2 \leq 0$ может быть сведен к $a_1 \geq 0$ путем замены A на $-A$. Если $a_1a_2 < 0$, то получаем

$$A(B + C) = [a_1(b_2 + c_2),\ a_2(b_2 + c_2)],$$

а также

$$AB + AC = [a_1b_2,\ a_2b_2] + [a_1c_2,\ a_2c_2] = [a_1(b_2 + c_2),\ a_2(b_2 + c_2)],$$

что доказывает утверждение (8) и для этого случая.

Теперь мы хотим остановиться на вопросе разрешимости уравнения

$$AX = B,$$

Где $A \neq [0, 0]$ и $X \in I(\mathbb{R})$. Для того чтобы ответить на этот вопрос, введем вспомогательную функцию χ:

$$\chi(A) = \begin{cases} a_1/a_2, & \text{если } |a_1| \leqslant |a_2|, \\ a_2/a_1 & \text{в остальных случаях} \end{cases}$$

(эту функцию предложил Ратшек).

Справедливо следующее утверждение: уравнение $AX=B$ разрешимо относительно X из $I(\mathrm{R})$ тогда и только тогда, когда

$$\chi(A) \geqslant \chi(B).$$

Решение не единственно лишь в случае

$$\chi(A) = \chi(B) \leqslant 0.$$

Проиллюстрируем приведенное утверждение примером. Пусть дано уравнение

$$[1, 2]X = [-1, 3].$$

Равенство выполняется лишь при X = [—1/2, 3/2], поскольку

$$\chi[1, 2] = \frac{1}{2} > \chi[-1, 3] = -\frac{1}{3}.$$

С другой стороны, если рассмотреть множество решений всех уравнений вида

$$ax = b,$$

у которых

$$a \in [1, 2], \quad b \in [-1, 3],$$

то получим

$$\{x = b/a \mid a \in [1, 2], b \in [-1, 3]\} = [-1, 3]/[1, 2] = [-1, 3] \supset X.$$

Это множество решений существенно отличается от интервала X, удовлетворяющего равенству $AX=B$. По этой причине мы не называем X решением уравнения $AX=B$, а предпочитаем говорить об «алгебраическом» решении.

Вообще, можно доказать следующее утверждение. Пусть уравнению $AX = B$, где $0 \notin A$, удовлетворяет некоторое X из $I(\mathrm{R})$. Тогда

$$X \subseteq B : A.$$

Действительно,

$$x \in X \Rightarrow \text{существуют } a \in A,\ b \in B, \text{ для которых}$$
$$ax = b \Rightarrow x = b/a \in B : A.$$

Заметим также, что равенство $AX = B$ может быть выполнено, даже если $B : A$ не определено. Примером служит уравнение

$$\left[-\frac{1}{3},\ 1\right]X = [-1,\ 2],$$

для которого единственным решением является X = [—1, 2], причем $\chi\left[-\frac{1}{3},\ 1\right] > \chi[-1,\ 2]$.

Основное свойство интервальных вычислений — монотонность включения. Следующая теорема разъясняет это свойство.

Теорема 5. *Пусть*

$$A^{(k)},\ B^{(k)} \in I(\mathbb{R}),\ k = 1,\ 2,$$

и предполагается,что

$$A^{(k)} \subseteq B^{(k)},\quad k = 1,\ 2.$$

Тогда для операции $*$ *из* $\{+,\ -,\ \cdot,\ :\}$ *имеем*

$$A^{(1)} * A^{(2)} \subseteq B^{(1)} * B^{(2)}. \tag{9}$$

Доказательство. Так как $A^{(k)} \subseteq B^{(k)},\ k = 1,\ 2,$ то

$$A^{(1)} * A^{(2)} = \{z = x * y \mid x \in A^{(1)},\ y \in A^{(2)}\}$$
$$\subseteq \{w = u * v \mid u \in B^{(1)},\ v \in B^{(2)}\} = B^{(1)} * B^{(2)}.$$

Приведем частный случай теоремы 5

Следствие 6. *Пусть* $A,\ B \in I(\mathbb{R})$ *и* $a \in A,\ b \in B.$ *Тогда*

$$a * b \in A * B,$$

где $* \in \{+,\ -,\ \cdot,\ :\}$.

Унарные операции *r(X)* из определения 3 обладают сходными свойствами:

$$X \subseteq Y \Rightarrow r(X) \subseteq r(Y),$$
$$x \in X \Rightarrow r(x) \in r(X). \tag{10}$$

Непосредственное обобщение этих соотношений на случай интервальных выражений дано в теореме 3 п. 7.3.

Замечания. Это элементарное введение в вещественную интервальную арифметику соответствует описанию, данному Муром. Большинство унарных операций из определения 3 легко задаются в виде функций от левой и правой границ интервального аргумента. К примеру, это можно без труда проделать для монотонных функций x^k и $\sqrt{x}$.

Четыре основные операции (+, —, • и :) на точечных множествах общего вида ввел Янг. Им были получены и некоторые элементарные соотношения, например (3), (4) и (8).

Кулиш исследовал, какие свойства операций, заданных на множестве *M,* переносятся на множество всех его подмножеств *P(M).* Интервальные операции вида (1) получаются при этом как частный случай в числе прочих результатов.

Представление интервалов, которое применил Сунага, соответствует круговым комплексным интервалам, описываемым в последующих микромодулях. При этом способе записи пара чисел *(a, r)* обозначает интервал $[a — r, a + r]$. Данное представление было использовано им для явного описания и последующего применения интервальных операций вида (1).

Ортольф отождествил интервалы $[a_1, a_2]$ с точками (a_1, a_2) из R×R. На этой основе ему удалось построить определение операций над всеми элементами R × R. При $a_1 \le a_2$ его определение сводится к операциям вида (2). Подобным же образом над точками из $R \times R$ вводятся отрицание (аддитивная инверсия) и, если $0 \notin [a_1, a_2]$, обращение (мультипликативная инверсия).

Кахан предложил обобщение интервальных операций вида (2). Наряду с обычными вещественными числами аргументами обобщенных операций могут быть +∞ и —∞. В некоторых случаях результатом операции оказывается «интервал» Ω, включающий все вещественные числа. Кроме Ω допускаются также интервалы $[a_1, a_2[$, $]a_1, a_2[$, $]a_1, a_2]$, причем разрешены $a_1 = \pm\infty$ и $a_2 = \pm\infty$. Более того, a_2 может быть меньше, чем a_1 (Например, запись [3,2] заменяет выражение $[-\infty, 2] \cup [3, +\infty]$ Подобные объекты могут возникать в результате разрешенного в этой арифметике деления на интервал, содержащий нуль). Для интерпретации такого представления интервалов используется ориентированная окружность, на которой располагаются вещественные числа. Введенные подобным образом интервалы могут содержать $\infty \equiv +\infty \equiv -\infty$, быть открытыми и полуоткрытыми. Их арифметика определяется в соответствии с (1).

В общем виде интервальные вычисления в частично упорядоченных пространствах описал Апостолатос**.** И на этот раз $I(R)$ возникает как частный случай.

Клауа разработал трехзначную теорию множеств. Он вводит так называемые частичные множества и частичные кардинальные числа Получающаяся в результате арифметика кардинальных чисел для конечного случая в точности соответствует интервальной. Таким

образом, аналогом интервальной арифметики на $I(\mathrm{R})$ служат операции над трехзначными числами. Наряду с отношением = из определения 1 применяется более слабое отношение $=_{\#}$, которое для $A=[a_1, a_2]$ и $B = [b_1, b_2]$ задается следующим образом:

$$A =_{\#} B \Leftrightarrow A \cap B \neq \varnothing \Leftrightarrow \max\{a_1, b_1\} \leqslant \min\{a_2, b_2\}.$$

Отношение $=_{\#}$ рефлексивно и симметрично; кроме того,

$$A = B \Rightarrow A =_{\#} B.$$

Это означает, что если $A \neq_{\#} B$, то для всех a и b, таких что $a \in A$ и $b \in B$, мы всегда имеем $a \neq b$. Соответственно

$$A \neq_{\#} B \Rightarrow A \neq B.$$

Имея в виду отношение $=_{\#}$, можно рассмотреть $I(\mathrm{R})$ как разновидность обобщенного поля и, например, доказать следующие свойства:

$$X - X =_{\#} 0 \qquad \text{для} \quad X \in I(\mathbb{R});$$

$$AX =_{\#} B \Leftrightarrow X =_{\#} B : A \qquad \text{для} \quad A,\ B,\ X \in I(\mathbb{R}),\ \text{причем}\ A \neq_{\#} 0;$$

$$X(Y + Z) =_{\#} XY + XZ \qquad \text{для} \quad X,\ Y,\ Z \in I(\mathbb{R}).$$

Каухер предложил расширенное множество $\overline{I(\mathbb{R})}$, получив его как результат дополнения $I(\mathrm{R})$ так называемыми нерегулярными интервалами, т. е. интервалами отрицательной ширины. В этом случае точечные интервалы $[a, a]$ больше не являются минимальными элементами в смысле порядка, задаваемого отношением $\subseteq$. Все структуры $I(\mathrm{R})$ переносятся на $I(\mathbb{R}) \cup \overline{I(\mathbb{R})}$, и с помощью несобственных элементов p и $-p$ достигается замкнутость. Подобным образом можно определить деление на интервал $[a_1, a_2]$, у которого $a_1 \leqslant 0 \leqslant a_2$ и $a_1 \neq a_2$.

7.2. Свойства интервальной арифметики

Введем теперь понятие расстояния на множестве вещественных интервалов.

Определение 1. Расстояние $q(A, B)$ между двумя интервалами A и B, такими что $A = [a_1, a_2],\ B = [b_1, b_2] \in I(\mathbb{R})$, определяется равенством

$$q(A, B) = \max\{|a_1 - b_1|, |a_2 - b_2|\}.$$

Легко показать, что отображение q задает на $I(\mathrm{R})$ метрику. Действительно, q обладает следующими свойствами:

$$q(A, B) \geqslant 0 \quad \text{и} \quad q(A, B) = 0 \Leftrightarrow A = B,$$

$$q(A, B) \leqslant q(A, C) + q(B, C) \text{ (неравенство треугольника).}$$

Выполнение неравенства треугольника проверяется следующим образом:

$$\begin{aligned} q(A, C) + q(B, C) &= \max\{|a_1 - c_1|, |a_2 - c_2|\} \\ &\quad + \max\{|b_1 - c_1|, |b_2 - c_2|\} \\ &\geqslant \max\{|a_1 - c_1| + |b_1 - c_1|, |a_2 - c_2| + |b_2 - c_2|\} \\ &\geqslant \max\{|a_1 - b_1|, |a_2 - b_2|\} = q(A, B). \end{aligned}$$

Если применить введенное таким способом расстояние к точечным интервалам, то оно сведется к обычному расстоянию между вещественными числами. Иначе говоря,

$$q([a, a], [b, b]) = |a - b|.$$

Предложенная здесь метрика является для $I(\mathrm{R})$ хаусдорфовой. Хаусдорфова метрика обобщает понятие расстояния между двумя точками в метрическом пространстве (у нас таким пространством является R с $q(x, y) = |x - y|$) на случай пространства всех компактных непустых подмножеств данного пространства. Если U и V — непустые компактные множества вещественных чисел, то хаусдорфово расстояние определяется как

$$q(U, V) = \max\{\sup_{v \in V} \inf_{u \in U} q(u, v), \sup_{u \in U} \inf_{v \in V} q(u, v)\}.$$

Существуют другие полезные определения хаусдорфовой метрики. Легко убедиться в том, что для вещественных интервалов A и B хаусдорфова метрика задается выражением из определения 1.

Вводя на множестве $I(\mathrm{R})$ метрику, мы делаем его топологическим пространством. При этом понятия сходимости и непрерывности могут использоваться обычным образом, как и в случае метрического пространства. В этой связи мы получаем, что последовательность интервалов $\{A^{(k)}\}_{k=0}^{\infty}$ сходится к интервалу $A = [a_1, a_2]$ тогда и только тогда, когда последовательность границ отдельных членов последовательности сходится к его соответствующим границам. Следовательно, мы можем записать

$$\lim_{k \to \infty} A^{(k)} = A \Leftrightarrow \left(\lim_{k \to \infty} a_1^{(k)} = a_1 \text{ и } \lim_{k \to \infty} a_2^{(k)} = a_2\right). \qquad (1)$$

Доказательство этого утверждения мы опускаем, так как его легко получить непосредственно из определения расстояния между двумя интервалами.

Введенная нами метрика используется в следующей теореме

Теорема 2. *Метрическое пространство* ($I(\mathrm{R})$, q) *с метрикой из определения* 1 *является замкнутым метрическим пространством.*

(Это означает, что любая интервальная последовательность Коши сходится к интервалу.)

В теореме 3 рассматривается характер сходимости широко используемого класса интервальных последовательностей.

Теорема 3. *Каждая последовательность интервалов* $\{A^{(k)}\}_{k=0}^{\infty}$, *для которой справедливо соотношение*

$$A^{(0)} \supseteq A^{(1)} \supseteq A^{(2)} \supseteq \ldots,$$

сходится к интервалу $A = \bigcap_{k=0}^{\infty} A^{(k)}$.

Доказательство. Пусть имеется последовательность границ, такая что

$$a_1^{(0)} \leqslant a_1^{(1)} \leqslant a_1^{(2)} \leqslant a_1^{(3)} \leqslant \ldots \leqslant a_2^{(3)} \leqslant a_2^{(2)} \leqslant a_2^{(1)} \leqslant a_2^{(0)}.$$

Тогда последовательность нижних границ интервалов из $\{A^{(k)}\}_{k=0}^{\infty}$, является монотонной неубывающей последовательностью вещественных чисел, ограниченной сверху величиной $a_2^{(0)}$. Эта последовательность сходится к вещественному числу a_1. Аналогично, монотонная невозрастающая последовательность вещественных чисел $\{a_2^{(k)}\}_{k=0}^{\infty}$ сходится к вещественному числу a_2, причем $a_1 \leq a_2$. Равенство

$$A = \bigcap_{k=0}^{\infty} A^{(k)}$$

проверяется столь же простым способом.

Как видно из доказательства, каждая последовательность $\{A^{(k)}\}_{k=0}^{\infty}$, для которой

$$A^{(0)} \supseteq A^{(1)} \supseteq A^{(2)} \supseteq A^{(3)} \supseteq \ldots \supseteq B,$$

сходится к такому интервалу A, что $A \supseteq B$.

Для арифметических, а также других определенных выше операций справедлива

Теорема 4. *Введенные ранее операции сложения, вычитания, умножения и деления интервалов непрерывны.*

Доказательство. Мы приводим доказательство только для операции сложения. Пусть $\{A^{(k)}\}_{k=0}^{\infty}$ и $\{B^{(k)}\}_{k=0}^{\infty}$ — две последовательности интервалов, причем $\lim\limits_{k\to\infty} A^{(k)} = A$ и $\lim\limits_{k\to\infty} B^{(k)} = B$. Из (1) вытекает, что последовательность интервальных сумм $\{A^{(k)} + B^{(k)}\}_{k=0}^{\infty}$ имеет предел

$$\begin{aligned}\lim_{k\to\infty}(A^{(k)} + B^{(k)}) &= \lim_{k\to\infty}\left[a_1^{(k)} + b_1^{(k)},\ a_2^{(k)} + b_2^{(k)}\right] \\ &= \left[\lim_{k\to\infty}\left(a_1^{(k)} + b_1^{(k)}\right),\ \lim_{k\to\infty}\left(a_2^{(k)} + b_2^{(k)}\right)\right] \\ &= [a_1 + b_1,\ a_2 + b_2] = A + B.\end{aligned}$$

Доказательство непрерывности остальных операций может быть проведено аналогичным способом.

Обобщением теоремы 4 служит (см. определение 3 п.7.1)

Следствие 5 . *Пусть r — непрерывная функция и*

$$r(X) = \left[\min_{x\in X} r(x),\ \max_{x\in X} r(x)\right].$$

Тогда r(X) — непрерывное интервальное выражение.

Доказательство этого следствия основывается непосредственно на факте непрерывности функции r и поэтому здесь будет опущено. Следствие 5 гарантирует непрерывность выражений, подобных X^k, $\sin X$ и e^X.

Определение 6. Пусть $A = [a_1, a_2] \in I(\mathbb{R})$. Абсолютной величиной этого интервала будем называть величину

$$|A| = q(A, [0, 0]) = \max\{|a_1|, |a_2|\}.$$

Абсолютную величину интервала можно записать и в виде

$$|A| = \max_{a\in A}|a|. \tag{2}$$

Очевидно, что если $A,\ B \in I(\mathbb{R})$, то

$$A \subseteq B \Rightarrow |A| \leqslant |B|. \tag{3}$$

Докажем теперь некоторые свойства, связанные с метрикой на $I(\mathbb{R})$.

Теорема 7. *Пусть*

$$A = [a_1, a_2],\ B = [b_1, b_2],\ C = [c_1, c_2],\ D = [d_1, d_2] \in I(\mathbb{R}).$$

Тогда

$$q(A + B,\ A + C) = q(B,\ C), \tag{4}$$

$$q(A + B,\ C + D) \leqslant q(A,\ C) + q(B,\ D), \tag{5}$$

$$q(aB,\ aC) = |a|\,q(B,\ C), \quad a \in \mathbb{R}, \tag{6}$$

$$q(AB,\ AC) \leqslant |A|\,q(B,\ C). \tag{7}$$

Доказательство. (4): Из определения метрики q следует, что

$$q(A+B,\ A+C) = \max\{|a_1+b_1-(a_1+c_1)|,\ |a_2+b_2-(a_2+c_2)|\}$$
$$= \max\{|b_1-c_1|,\ |b_2-c_2|\} = q(B,\ C).$$

(5): Из неравенства треугольника, предыдущего свойства (4) и симметричности q вытекает, что

$$q(A+B,\ C+D) \leqslant q(A+B,\ B+C) + q(C+D,\ B+C)$$
$$= q(A,\ C) + q(B,\ D).$$

(6): $q(aB,\ aC) = \max\{|ab_1-ac_1|,\ |ab_2-ac_2|\} = |a|\,q(B,\ C).$

(7): Пусть $A = [a_1,\ a_2]$. Для краткости будем использовать обозначения $i(A) = a_1$ и $s(A) = a_2$. Тогда утверждение (7) можно записать в виде

$$\max\{|i(AB)-i(AC)|,\ |s(AB)-s(AC)|\} \leqslant |A|\,q(B,\ C).$$

Докажем, что

$$|i(AB)-i(AC)| \leqslant |A|\,q(B,\ C).$$

Неравенство

$$|s(AB)-s(AC)| \leqslant |A|\,q(B,\ C)$$

доказывается аналогично.

Перепишем предыдущее соотношение (6):

$$\max\{|i(aB)-i(aC)|,\ |s(aB)-s(aC)|\} = |a|\,q(B,\ C).$$

Теперь без потери общности можно предположить, что

$$i(AB) \geqslant i(AC).$$

(Случай $i(AB) < i(AC)$ рассматривается точно так же.)

Поскольку

$$AC = \{ac \mid a \in A,\ c \in C\},$$

существует такое a из A, что

$$i(AC) = i(aC).$$

Из свойства монотонности включения следует, что

$$aB \subseteq AB \text{ и } i(aB) \geqslant i(AB),$$

откуда видно, что

$$i(aB) - i(aC) \geqslant i(AB) - i(AC) \geqslant 0.$$

Итак,

$$|i(AB)-i(AC)|=i(AB)-i(AC)\leqslant i(aB)-i(aC)$$
$$=|i(aB)-i(aC)|\leqslant|a|q(B,\ C)$$
$$\leqslant|A|q(B,\ C).$$

Отождествляя $|A|$ с $q(A, 0)$, получаем следующие легко проверяемые свойства абсолютного значения:

$$\begin{aligned}&|A|\geqslant 0 \text{ и } |A|=0\Leftrightarrow A=[0,\ 0],\\&|A+B|\leqslant|A|+|B|,\\&|xA|=|x|\,|A| \text{ для } x\in\mathbb{R},\\&|AB|=|A||B|.\end{aligned}\tag{8}$$

Вот доказательство последнего равенства:

$$|AB|=\max_{c\in AB}|c|=\max_{a\in A,\ b\in B}|ab|=\max_{a\in A,\ b\in B}(|a||b|)$$
$$=\max_{a\in A}|a|\max_{b\in B}|b|=|A||B|.$$

Остальные соотношения доказываются подобным же образом.

Определение 8. Шириной интервала $A=[a_1, a_2]$ будем называть

$$d(A)=a_2-a_1\geqslant 0.$$

Множество точечных интервалов можно теперь описать как

$$\{A\in I(\mathbb{R})\mid d(A)=0\}.$$

Из определения 8 сразу же получаем свойства

$$A\subseteq B\Rightarrow d(A)\leqslant d(B),\tag{9}$$

$$d(A\pm B)=d(A)+d(B).\tag{10}$$

Утверждение (9) доказывается тривиально — достаточно определение 8 переписать в виде

$$d(A)=\max_{a,\ b\in A}|a-b|.\tag{11}$$

Проверим свойство (10) для операции сложения:

$$d(A+B)=d([a_1+b_1,\ a_2+b_2])$$
$$=a_2+b_2-(a_1+b_1)$$
$$=a_2-a_1+b_2-b_1=d(A)+d(B).$$

Вычитание проверяется точно так же Кроме того, имеет место теорема.

Теорема 9. *Пусть A и B — вещественные интервалы из $I(\mathbb{R})$ Тогда*

$$d(AB)\leqslant d(A)|B|+|A|d(B),\tag{12}$$

$$d(AB)\geqslant\max\{|A|d(B),\ |B|d(A)\},\tag{13}$$

$$d(aB) = |a| d(B),\ a \in \mathbb{R}, \tag{14}$$

$$d(A^n) \leqslant n|A|^{n-1} d(A),\ n = 1, 2, \ldots, \tag{15}$$

$$(A^n := A \cdot A \cdot \ldots \cdot A,\ n \text{ раз}),$$

$$d((X-x)^n) \leqslant 2(d(X))^n, \text{ где } x \in X,\ n = 1, 2, \ldots, \tag{16}$$

$$((X-x)^n := (X-x)(X-x) \ldots (X-x),\ n \text{ раз}).$$

Если $C \in l(\mathbb{R})$ *и* $0 \in C$, то

$$|C| \leqslant d(C) \leqslant 2|C|. \tag{17}$$

Доказательство (12): Используя тождество (11), получаем

$$\begin{aligned} d(AB) &= \max_{a,\, a' \in A,\ b,\ b' \in B} |ab - a'b'| \\ &= \max_{a,\, a' \in A,\ b,\ b' \in B} |ab - ab' + ab' - a'b'| \\ &\leqslant \max_{a,\, a' \in A,\ b,\ b' \in B} \{|a(b-b')| + |(a-a')b'|\} \\ &\leqslant \max_{a \in A\ \ b,\, b' \in B} |a||b-b'| + \max_{a,\, a' \in A,\ b' \in B} |a-a'||b'| \\ &= \left(\max_{a \in A} |a|\right)\left(\max_{b,\ b' \in B} |b-b'|\right) + \left(\max_{a\ \ a' \in A} |a-a'|\right)\left(\max_{b' \in B} |b'|\right) \\ &= |A| d(B) + d(A)|B|. \end{aligned}$$

(13): Сначала докажем, что

$$\begin{aligned} d(AB) &= \max_{a,\, a' \in A\ \ b\ \ b' \in B} |ab - a'b'| \geqslant \max_{a \in A,\ b,\ b' \in B} |ab - ab'| \\ &= \max_{a \in A,\ b\ \ b' \in B} |a||b-b'| = |A| d(B). \end{aligned}$$

Подобным образом можно показать, что

$$d(AB) \geqslant |B| d(A),$$

откуда сразу же вытекает (13).

(14): $$\begin{aligned} d(aB) &= \max_{b,\ b' \in B} |ab - ab'| = \max_{b,\ b' \in B} \{|a||b-b'|\} \\ &= |a| \max_{b,\ b' \in B} |b-b'| = |a| d(B). \end{aligned}$$

(15): При n=1 имеет место равенство. Если неравенство выполняется для некоторого $n \geq 1$, то, используя (12) и последнее соотношение из (8), имеем

$$\begin{aligned} d(A^{n+1}) &= d(A^n A) \leqslant d(A^n)|A| + |A|^n d(A) \\ &\leqslant n|A|^{n-1} d(A)|A| + |A|^n d(A) \\ &= (n+1)|A|^n d(A). \end{aligned}$$

(16): Поскольку $x \in X$, из (9) и свойства монотонности включения получаем

$$d((X-x)^n) \leqslant d((X-X)^n) = d([-d(X),\ d(X)]^n)$$
$$= d([-(d(X))^n,\ (d(X))^n]) = 2(d(X))^n.$$

(17): Так как $0 \in C = [c_1,\ c_2]$, то $c_1 \leqslant 0 \leqslant c_2$, откуда имеем

$$d(C) = c_2 - c_1 = |c_2| + |c_1| \geqslant \max\{|c_1|,\ |c_2|\} = |C|.$$

Итак,

$$d(C) = |c_1| + |c_2| \leqslant 2\max\{|c_1|,\ |c_2|\} = 2|C|.$$

Теперь докажем следующую теорему.

Теорема 10. *Пусть* $A,\ B \in I(\mathbb{R})$, *причем* A *— симметричный интервал, т. е.* $A = -A$. *Тогда имеют место следующие свойства:*

$$AB = |B|A, \tag{18}$$

$$d(AB) = |B|d(A). \tag{19}$$

Если $b_1 \geqslant 0$ или $b_2 \leqslant 0$, то второе свойство выполняется и в случае, когда $0 \notin A$.

Доказательство. Предположим, что $A = -A$, или, что то же самое, $a_2 = a = -a_1$. Тогда

$$AB = [\min\{ab_1,\ ab_2,\ -ab_1,\ -ab_2\},\ \max\{ab_1,\ ab_2,\ -ab_1,\ -ab_2\}]$$
$$= [a\min\{b_1,\ -b_1,\ b_2,\ -b_2\},\ a\max\{b_1,\ -b_1,\ b_2,\ -b_2\}]$$
$$= [a(-|B|),\ a|B|] = [-a,\ a]|B| = |B|A.$$

Опираясь на равенство (14), получаем (19) Остальные случаи могут быть доказаны аналогичным образом.

Теорема 11. *Для интервалов* A *и* B *из* $I(\mathrm{R})$ *справедливы следующие свойства:*

$$d(A) = |A - A|, \tag{20}$$

$$A \subseteq B \Rightarrow \frac{1}{2}(d(B) - d(A)) \leqslant q(A,\ B) \leqslant d(B) - d(A). \tag{21}$$

Доказательство.

(20): $$d(A) = a_2 - a_1 = |A - A|.$$

(21): Пусть $A \subseteq B$. Тогда $b_1 \leqslant a_1 \leqslant a_2 \leqslant b_2$ и, следовательно,

$$q(A,\ B) = \max\{|a_1 - b_1|,\ |a_2 - b_2|\} = \max\{a_1 - b_1,\ b_2 - a_2\}$$
$$\leqslant b_2 - a_2 + a_1 - b_1 = b_2 - b_1 - (a_2 - a_1) = d(B) - d(A),$$

откуда

$$q(A,\ B) = \max\{a_1 - b_1,\ a_2 - b_2\} \geqslant \frac{1}{2}(a_1 - b_1 + b_2 - a_2)$$
$$= \frac{1}{2}(d(B) - d(A)).$$

Введем теперь на $I(\mathbb{R})$ еще одну бинарную операцию. Пусть $A,\ B \in I(\mathbb{R})$. Тогда отношение

$$A \cap B = \{c \mid c \in A,\ c \in B\} \tag{22}$$

представляет собой теоретико-множественное пересечение двух интервалов. Результат этой операции принадлежит $I(\mathrm{R})$ тогда и только тогда, когда пересечение не пусто. В этом случае

$$A \cap B = [\max\{a_1,\ b_1\},\ \min\{a_2,\ b_2\}]. \tag{23}$$

В приводимом ниже следствии собраны важные свойства операции пересечения.

Следствие 12. *Пусть* $A,\ B,\ C,\ D \in I(\mathbb{R})$. *Тогда*

$$A \subset C,\quad B \subseteq D \Rightarrow A \cap B \subseteq C \cap D \ (\textit{монотонность включения}). \tag{24}$$

Пока операция пересечения не выводит из $I(\mathrm{R})$, *она непрерывна.*

Доказательство. Монотонность включения (24) вытекает из определения (22). Доказательство непрерывности может быть получено с помощью (23).

Замечания. Мур использовал хаусдорфову метрику на $I(\mathrm{R})$, соответствующую определению 1. Некоторые из правил для вычисления абсолютного значения $|A|$ и ширины $d(A)$ имеются в работах Мура и Кулиша. Важное в приложениях неравенство (7) впервые доказал Майер.

Иногда абсолютное значение вводится следующим способом, основанным на определении 3 п. 7.1:

$$\mathrm{abs}(A) = \left[\min_{a \in A} |a|,\ \max_{a \in A} |a|\right].$$

Поскольку такое определение редко применяется в приложениях, мы также не будем его использовать.

Согласно С. М. Румпу (частное сообщение), в выражении (16) можно обойтись без сомножителя 2, если уточнить соответствующую оценку. Для x, принадлежащего X, $X - x = [a,\ b]$, где $a \leqslant 0,\ b \geqslant 0$. Предположим, что $b \geqslant -a = |a|$ (если это не так, будем иметь дело с x— X). Тогда

$$(X - x)(X - x) = [ab,\ b^2],$$

и с помощью полной индукции получаем

$$(X - x)^n = \left[ab^{n-1},\ b^n\right].$$

Следовательно,

$$d\left((X - x)^n\right) = b^n - ab^{n-1} = b^{n-1}(b - a).$$

Теперь $b - a = d(X - x) = d(X)$, и, поскольку $a \leqslant 0,\ b \geqslant 0$, имеем $b \leqslant d(X - x) = d(X)$. Итак,

$$d((X - x)^n) \leqslant d(X)^n.$$

Микромодуль 24
Интервальное оценивание

В этом микромодуле мы обсуждаем непрерывные вещественные функции. Пусть *f* относится к их числу. Аналитическое выражение для *f=f(x)* представляет собой запись вычислительной процедуры, выдающей значение функции *f* для произвольного аргумента *x* Примем при этом, что все выражения, с которыми мы будем иметь дело, составлены из операций и операндов, число которых конечно, одновременно мы предполагаем, что если эти выражения вычисляются в интервальной арифметике, то составляющие их операции трактуются в соответствии с определениями 2 п. 7.1 и 3 п.7.1. Выражение, содержащее константы $a^{(0)}$, ..., $a^{(m)}$, будет для наглядности записываться в виде $f(x;\ a^{(0)},\ ...,\ a^{(m)})$. Чтобы упростить изложение, в дальнейшем мы всегда будем предполагать, что каждая из констант $a^{(k)}$, $0 \le k \le m$, встречается в аналитическом выражении функции только один раз Этого всегда можно добиться, вводя новые константы, равные константам, встречающимся неоднократно.

Пример. Двумя аналитическими выражениями функции *g* являются

$$g^{(1)}(x;\ a) = \frac{ax}{1 - x},\quad x \neq 1,\quad x \neq 0,$$

и

$$g^{(2)}(x;\ a) = \frac{a}{1/x - 1},\quad x \neq 1,\quad x \neq 0.$$

Запись

$$W(f,\ X;\ A^{(0)},\ \ldots,\ A^{(m)})$$
$$= \{f(x;\ a^{(0)},\ \ldots,\ a^{(m)}) \mid x \in X,\ a^{(k)} \in A^{(k)},\ 0 \leqslant k \leqslant m\}$$
$$= \left[\min_{\substack{x \in X \\ a^{(k)} \in A^{(k)}, \\ 0 \leqslant k \leqslant m}} f(x;\ a^{(0)},\ \ldots,\ a^{(m)}),\quad \max_{\substack{x \in X \\ a^{(k)} \in A^{(k)}, \\ 0 \leqslant k \leqslant m}} f(x;\ a^{(0)},\ \ldots,\ a^{(m)})\right]$$

будет в дальнейшем обозначать диапазон изменения функции *f*, причем предполагается, что x из X и $a^{(k)}$ из $A^{(k)}$, $0 \leqslant k \leqslant m$, не зависят друг от друга. Согласно этому определению, интервал $W(f,\ X;\ A^{(0)},\ \ldots,\ A^{(m)})$ будет одним и тем же при любом аналитическом выражении для *f*.

Пример. Возьмем *g* из предыдущего примера. Для

$$A = [0,\ 1] \text{ и } X = [2,\ 3]$$

получаем

$$W(g,\ [2,\ 3];\ [0,\ 1]) = \left\{\frac{ax}{1-x} \,\middle|\, 2 \leqslant x \leqslant 3,\ 0 \leqslant a \leqslant 1\right\} = [-2,\ 0].$$

Введем теперь понятие интервального оценивания вещественной функции *f*. Пусть для *f* имеется аналитическое выражение. Заменив в этом выражении все вещественные операнды и операции над ними на интервальные операнды и операции, получим выражение $f(X;\ A^{(0)},\ \ldots,\ A^{(m)})$. Если все операнды попадают в области, на которых заданы операции из определений 2 п. 7.1 и 3 п.7.1, то $f(X;\ A^{(0)},\ \ldots,\ A^{(m)})$ называется *интервальной оценивающей функцией,* или, для краткости, *оценкой f,* а получение ее значения— *вычислением,* или *оцениванием, f в интервальной арифметике.*

Для функций, рассматриваемых нами, замена описанного типа возможна всегда. Константы $a^{(0)},\ \ldots,\ a^{(m)}$, как и переменная *x,* превращаются в интервалы. Очевидно, что результат оценивания функции *f* зависит от выбора для нее аналитического выражения. Впоследствии мы будем использовать этот факт. А сейчас приведем простой пример.

Пример. Пусть *g* — функция из предыдущих двух примеров. Для $A = [0,\ 1]$ и $X = [2,\ 3]$ получим две различные оценки:

$$g^{(1)}([2,\ 3];\ [0,\ 1]) = \frac{[0,\ 1][2,\ 3]}{1-[2,\ 3]} = [-3,\ 0],$$

$$g^{(2)}([2,\ 3];\ [0,\ 1]) = \frac{[0,\ 1]}{1/[2,\ 3]-1} = [-2,\ 0] \neq g^{(1)}([2,\ 3];\ [0,\ 1]).$$

Введенные выше обозначения можно распространить на функции от нескольких переменных. В этом случае множеством значений

$f(x^{(1)}, \ldots, x^{(n)}; a^{(0)}, \ldots, a^{(m)})$ при независимых $x^{(k)}$ из $X^{(k)}$, $1 \leqslant k \leqslant n$, и $a^{(j)}$ из $A^{(j)}$, $0 \leqslant j \leqslant m$, становится $W(f, X^{(1)}, \ldots, X^{(n)}; A^{(0)}, \ldots, A^{(m)})$. Подобным же образом обобщается понятиеинтервальной оценки — теперь она обозначается через

$$f(X^{(1)}, \ldots, X^{(n)}; A^{(0)}, \ldots, A^{(m)})$$

(Множество $W(f, X^{(1)}, \ldots, X^{(n)}; A^{(0)}, \ldots, A^{(m)})$ Мур и другие авторы называют объединенным интервальным расширением, а функцию $f(X^{(1)}, \ldots, X^{(n)}; A^{(0)}, \ldots, A^{(m)})$ — естественным интервальным расширением).

Приведем теперь пример выражения, из которого не удается получить его всюду определенный интервальный аналог путем простой замены операций и операндов. Вещественная функция

$$f(x) = 1 \Big/ \left(x^2 + \frac{1}{2}\right)$$

определена для всех x из R. Представим f в виде

$$\tilde{f}(x) = 1 \Big/ \left(x \cdot x + \frac{1}{2}\right).$$

Заменим теперь независимую переменную x на интервал X=[—1, 1], содержащийся в области определения f. Замена всех операций на соответствующие им интервальные приводит к интервальному выражению

$$\tilde{f}([-1, 1]) = \frac{1}{[-1, 1][-1, 1] + \frac{1}{2}} = \frac{1}{[-1, 1] + \frac{1}{2}} = \frac{1}{\left[-\frac{1}{2}, \frac{3}{2}\right]},$$

которое не определено.

Познакомимся с рядом свойств интервального оценивания. Два свойства, используемые в последующих утверждениях, легко выводятся из теоремы 5 п.7.1 и следствия 6 п.7.1.

Теорема 1. *Пусть f — непрерывная вещественная функция, $f(x^{(1)}, \ldots, x^{(n)}; a^{(0)}, \ldots, a^{(m)})$ — аналитическое выражение для f. Предположим также, что для интервалов $Y^{(1)}, \ldots, Y^{(n)}, B^{(0)}, \ldots, B^{(m)}$ имеется оценка $f(Y^{(1)}, \ldots, Y^{(n)}; B^{(0)}, \ldots, B^{(m)})$. Тогда*

а) *для всех*

$$X^{(k)} \subseteq Y^{(k)}, \quad A^{(j)} \subseteq B^{(j)}, \quad 1 \leqslant k \leqslant n, \quad 0 \leqslant j \leqslant m$$

справедливо свойство включения

$$W(f, X^{(1)}, \ldots, X^{(n)}; A^{(0)}, \ldots, A^{(m)}) \subseteq f(X^{(1)}, \ldots, X^{(n)}; A^{(0)}, \ldots, A^{(m)}); \tag{1}$$

б) *для всех*

$$X^{(k)} \subseteq Z^{(k)} \subseteq Y^{(k)}, \quad A^{(j)} \subseteq C^{(j)} \subseteq B^{(j)}, \quad 1 \leqslant k \leqslant n, \quad 0 \leqslant j \leqslant m$$

имеет место монотонность включения

$$f(X^{(1)}, \ldots, X^{(n)}; A^{(0)}, \ldots, A^{(m)}) \subseteq f(Z^{(1)}, \ldots, Z^{(n)}; C^{(0)}, \ldots, C^{(m)}). \qquad (1')$$

Пример. *Функция f задана выражением*

$$f(x; a) = a - x/(1 + x), \quad x \neq -1.$$

Для

$$X = \left[-\tfrac{1}{2}, 1\right], \quad Z = \left[-\tfrac{1}{2}, 2\right], \quad A = C = [2, 3]$$

получаем

$$W\left(f, \left[-\tfrac{1}{2}, 1\right]; [2, 3]\right) = \left[\tfrac{3}{2}, 4\right] \subset f\left(\left[-\tfrac{1}{2}, 1\right]; [2, 3]\right) = [0, 4],$$

$$f\left(\left[-\tfrac{1}{2}, 1\right]; [2, 3]\right) = [0, 4] \subset f\left(\left[-\tfrac{1}{2}, 2\right]; [2, 3]\right) = [-2, 4].$$

Свойство включения (1) позволяет соотнести множество значений функции с ее интервальной оценкой. Позднее мы дадим формулы для их качественного сравнения.

Можно привести примеры, когда в (1) достигается равенство. Очевидно, что к их числу относится случай однократного вхождения каждой из величин $x^{(1)}, \ldots, x^{(n)}; a^{(0)}, \ldots, a^{(m)}$ в выражение $f(x^{(1)}, \ldots, x^{(n)}; a^{(0)}, \ldots, a^{(m)})$.

Теорема 2. *Пусть p — многочлен от вещественной переменной x, определяемый выражением*

$$p(x; a^{(0)}, \ldots, a^{(m)}) = (\ldots((a^{(m)}x + a^{(m-1)})^{n_{m-1}} + a^{(m-2)})^{n_{m-2}} + \ldots + a^{(1)})^{n_1} + a^{(0)},$$

где $n_\nu \geqslant 2$, $1 \leqslant \nu \leqslant m - 1$. *Если встречающиеся в этом выражении степени вычисляются по формуле*

$$X^k = \left[\min_{x \in X} x^k, \max_{x \in X} x^k\right]$$

(см. определение 3 п. 7.1), *то*

$$W(p, X; a^{(0)}, \ldots, a^{(m)}) = p(X; a^{(0)}, \ldots, a^{(m)}).$$

Доказательство. Для $m = 2$ истинность теоремы очевидна:

$$p(x; a^{(0)}, a^{(1)}, a^{(2)}) = (a^{(2)}x + a^{(1)})^{n_1} + a^{(0)}.$$

Остальную часть доказательства получаем полной индукцией.

Не всякий многочлен можно привести к форме, требуемой теоремой 2. Однако многочлен второй степени

$$p(x;\ b^{(0)},\ b^{(1)}) = x^2 + b_{(1)}x + b^{(0)}$$

может быть преобразован к виду

$$p(x;\ a^{(0)},\ a^{(i)}) = (x + a^{(1)})^2 + a^{(0)},$$

где

$$a^{(1)} = b^{(1)}/2, \quad a^{(0)} = b^{(0)} - (b^{(1)})^2/4.$$

Наряду с носящей общий характер теоремой 1 и разобранными выше частными случаями, представляет также интерес качественное утверждение о приближении множества значений функции f с помощью ее интервальной оценки. В случае функции одной вещественной переменной справедлива

Теорема 3. *Пусть f — вещественная функция от вещественной переменной x. $f(x;\ a^{(0)},\ \ldots,\ a^{(m)})$ — ее аналитическое выражение. Обозначим через $\bar{f}(x^{(1)},\ \ldots,\ x^{(n)};\ a^{(0)},\ \ldots,\ a^{(m)})$ выражение, полученное заменой в $f(x,\ a^{(0)},\ \ldots,\ a^{(m)})$ каждого вхождения x на новую переменную $x^{(k)}$, $1 \le k \le n$. Пусть определена оценивающая функция $f(Y;\ A^{(0)},\ \ldots,\ A^{(m)})$, где $Y,\ A^{(0)},\ \ldots,\ A^{(m)} \in I(\mathbb{R})$.*

Кроме того, предположим, что для каждой переменной $x^{(k)}$, $1 \le k \le n$, из интервала Y и произвольно выбранных $x^{(j)}$ из Y,

$$1 \leqslant j \leqslant n,\ j \neq k,\ \text{и}\ a^{(j)}\ \text{из}\ A^{(j)},\ 0 \leqslant j \leqslant m,$$

выражение $\bar{f}(x^{(1)},\ \ldots,\ x^{(n)};\ a^{(0)},\ \ldots,\ a^{(m)})$ удовлетворяет условию Липшица. В остальном обозначения имеют тот же смысл, что и в теореме 1. При этих предположениях для $X \subseteq Y$ имеем

$$q(W(\bar{f},\ X;\ A^{(0)},\ \ldots,\ A^{(m)}),\ f(X;\ A^{(0)},\ \ldots,\ A^{(m)})) \leqslant \gamma d(X), \quad (2)$$
$$\gamma \geqslant 0.$$

Доказательство. Во-первых, заметим, что

$$\bar{f}(x,\ \ldots,\ x;\ a^{(0)},\ \ldots,\ a^{(m)}) = f(x;\ a^{(0)},\ \ldots,\ a^{(m)}), \quad x \in Y.$$

Теперь мы можем получить интервальную оценку для f в виде

$$f(X;\ A^{(0)},\ \ldots,\ A^{(m)}) = W(\bar{f},\ X,\ \ldots,\ X;\ A^{(0)},\ \ldots,\ A^{(m)}), \quad X \subseteq Y.$$

Остается показать, что

$$q(W(\bar{f},\ X;\ A^{(0)},\ \ldots,\ A^{(m)}),\ W(\bar{f},\ X,\ \ldots,\ X;\ A^{(0)},\ \ldots,\ A^{(m)})) \leqslant \gamma d(X), \quad X \subseteq Y.$$

Если теперь для $X \subseteq Y$ записать

$$W(f, X; A^{(0)}, \ldots, A^{(m)}) = [f(u; a^{(0)}, \ldots, a^{(m)}), f(v; b^{(0)}, \ldots, b^{(m)})],$$
$$u, v \in X, \quad a^{(j)}, b^{(j)} \in A^{(j)}, \quad 0 \leqslant j \leqslant m,$$
$$W(\tilde{f}, X, \ldots, X; A^{(0)}, \ldots, A^{(m)})$$
$$= [\tilde{f}(x^{(1)}, \ldots, x^{(n)}; c^{(0)}, \ldots, c^{(m)}),$$
$$\tilde{f}(y^{(1)}, \ldots, y^{(n)}; e^{(0)}, \ldots, e^{(m)})],$$
$$x^{(k)}, y^{(k)} \in X, \quad 1 \leqslant k \leqslant n, \quad c^{(j)}, e^{(j)} \in A^{(j)}, \quad 0 \leqslant j \leqslant m,$$

и принять во внимание соотношение

$$W(f, X; A^{(0)}, \ldots, A^{(m)}) \subseteq W(\tilde{f}, X, \ldots, X; A^{(0)}, \ldots, A^{(m)}),$$

то получим

$$|f(u; a^{(0)}, \ldots, a^{(m)}) - \tilde{f}(x^{(1)}, \ldots, x^{(n)}; c^{(0)}, \ldots, c^{(m)})|$$
$$= f(u; a^{(0)}, \ldots, a^{(m)}) - \tilde{f}(x^{(1)}, \ldots, x^{(n)}; c^{(0)}, \ldots, c^{(m)})$$
$$\leqslant f(u; c^{(0)}, \ldots, c^{(m)}) - \tilde{f}(x^{(1)}, \ldots, x^{(n)}; c^{(0)}, \ldots, c^{(m)})$$
$$= \tilde{f}(u, \ldots, u; c^{(0)}, \ldots, c^{(m)}) - \tilde{f}(x^{(1)}, \ldots, x^{(n)}; c^{(0)}, \ldots, c^{(m)})$$
$$\leqslant \gamma \max_{1 \leqslant k \leqslant n} |u - x^{(k)}| \leqslant \gamma d(X).$$

Разность верхних границ может быть оценена аналогичным образом. Получение этих двух оценок доказывает утверждение теоремы.

Теорема 3, как видно из ее доказательства, легко обобщается на случай функции от нескольких переменных $x^{(1)}, \ldots, x^{(n)}$. Вместо $\gamma d(X)$ имеем величину

$$\sum_{k=1}^{n} \gamma^{(k)} d(X^{(k)}) \quad \left(\leqslant \gamma \max_{1 \leqslant k \leqslant n} d(X^{(k)})\right).$$

Следующий пример иллюстрирует тот факт, что степень близости множества значений функции f и ее интервальной оценки зависит от выбора аналитического выражения $f(x; a^{(0)}, \ldots, a^{(m)})$.

Пример. Пусть $f(x) = x - x^2$ и $X = [0, 1]$. Тогда

$$W(f, [0, 1]) = \{x - x^2 \mid 0 \leqslant x \leqslant 1\} = \left[0, \frac{1}{4}\right].$$

Различные аналитические выражения для f дают следующие результаты:

$$f^{(0)}(x) = x - x^2 \Rightarrow f^{(0)}([0,\ 1]) = [0,\ 1] - [0,\ 1] = [-1,\ 1],$$
$$f^{(1)}(x) = x(1-x) \Rightarrow f^{(1)}([0,\ 1]) = [0,\ 1](1 - [0,\ 1]) = [0,\ 1],$$
$$f^{(2)}(x) = \frac{1}{4} - \left(x - \frac{1}{2}\right)\left(x - \frac{1}{2}\right) \Rightarrow$$
$$f^{(2)}([0,\ 1]) = \frac{1}{4} - \left([0,\ 1] - \frac{1}{2}\right)\left([0,\ 1] - \frac{1}{2}\right) = \left[0,\ \frac{1}{2}\right],$$
$$f^{(3)}(x) = \frac{1}{4} - \left(x - \frac{1}{2}\right)^2 \Rightarrow$$
$$f^{(3)}([0,\ 1]) = \frac{1}{4} - \left([0,\ 1] - \frac{1}{2}\right)^2 = \left[0,\ \frac{1}{4}\right] = W(f,\ [0,\ 1]).$$

Для некоторых классов аналитических выражений можно доказать более сильные утверждения, нежели то, которое приведено в теореме 3. К их числу относится так называемая центрированная форма записи функции. Центрированная форма представляет собой специальное выражение, предназначенное для оценки функции f на интервале X. Ограничим наше рассмотрение случаем одной вещественной переменной. Выберем в X произвольную точку z и представим $f(x)$ в виде

$$f(x) = f(z) + (x - z)h(x - z), \tag{3}$$

где сомножитель $h(x - z)$ зависит от новой переменной z, равной $x - z$. Будем называть (3) формой $f(x)$, центрированной относительно z. Применительно к многочленам центрированная форма есть не что иное, как обычное тейлоровское разложение $f(x)$ в окрестности точки z, записанное с сомножителем $x - z$, имеющимся у всех членов, отличных от постоянного.

Рациональная функция $f(x) = p(x)/q(x)$ может быть, согласно Ратшеку, приведена к центрированной форме следующим образом. Пусть n — максимум из степеней многочленов $p(x)$ и $q(x)$. Для z из X определим

$$\gamma_\nu := p^{(\nu)}(z) - f(z)q^{(\nu)}(z), \quad 1 \leqslant \nu \leqslant n.$$

Функция

$$h(y) = \sum_{\nu=1}^{n} \gamma_\nu \frac{y^{\nu-1}}{\nu!} \Bigg/ \sum_{\nu=0}^{s} q^{(\nu)}(z) \frac{y^\nu}{\nu!}$$

является решением функционального уравнения

$$f(x) = f(z) + (x - z)h(x - z).$$

Теорема 4. *Пусть f — вещественная функция от вещественного аргумента x и*

$$f(x) = f(z) + (x - z)h(x - z)$$

— аналитическое выражение для f в центрированной форме. Кроме того, пусть имеется выражение $\tilde{h}(x^{(1)}-z,\ \ldots,\ x^{(n)}-z)$, *аналогичное соответствующему выражению из теоремы* 3. *Допустим, что для некоторого Y из I(R) существует интервальная оценка f(Y) и для каждой своей переменной* $\tilde{h}(x^{(1)}-z,\ \ldots,\ x^{(n)}-z)$ *удовлетворяет условию Липшица, подобно тому как это было в теореме* 3. *Тогда для* $X \subseteq Y$ *выполняется соотношение*

$$q(W(f,\ X)f(x)) \leqslant c(d(X))^2, \quad c \geqslant 0. \tag{4}$$

Доказательство. Поскольку

$$\tilde{h}(x-z,\ \ldots,\ x-z) = h(x-z)$$

и

$$\tilde{f}(x^{(0)},\ \ldots,\ x^{(n)}) = f(z) + (x^{(0)}-z)\,\tilde{h}(x^{(1)}-z,\ \ldots,\ x^{(n)}-z),$$

то

$$\begin{aligned}\tilde{f}(x,\ \ldots,\ x) &= f(z) + (x-z)\,\tilde{h}(x-z,\ \ldots,\ x-z)\\ &= f(z) + (x-z)h(x-z) = f(x).\end{aligned}$$

Теперь можно получить интервальную оценку дли *f*, записанной в центрированной форме:

$$f(X) = W(\tilde{f},\ X,\ \ldots,\ X).$$

Этот результат позволяет переписать (4) в виде

$$q(W(f,\ X),\ W(\tilde{f},\ X,\ \ldots,\ X)) \leqslant c(d(X))^2, \quad c \geqslant 0.$$

Пусть

$$\begin{aligned}W(\tilde{f},\ X,\ \ldots,\ X) = [&f(z) + (x^{(0)}-z)\,\tilde{h}(x^{(1)}-z,\ \ldots,\ x^{(n)}-z),\\ &f(z) + (y^{(0)}-z)\,\tilde{h}(y^{(1)}-z,\ \ldots,\ y^{(n)}-z)],\\ &x^{(k)},\ y^{(k)} \in X, \quad 0 \leqslant k \leqslant n.\end{aligned}$$

Заметим, что

$$W(f,\ X) \subseteq W(\tilde{f},\ X,\ \ldots,\ X).$$

Из (21 п. 7.2) вытекает

$$q(W(f,\ X),\ W(\tilde{f},\ X,\ \ldots,\ X)) \leqslant d(W(\tilde{f},\ X,\ \ldots,\ X)) - d(W(f,\ X)).$$

Теперь положим

$$\min_{x \in X} |h(x-z)| = |h(w-z)|.$$

Легко

$$f(z) + (X-z)h(w-z) \subseteq f(z) + \{(x-z)h(x-z) \mid x \in X\} = W(f,\ X),$$

убедиться в истинности соотношения

если проанализировать два случая, связанных со знаком $h(w—z)$. Далее, исходя из (9 п. 7.2) и (14 п. 7.2), получаем

$$d(W(f, X)) \geqslant d((X - z)h(w - z)) = d(X)|h(w - z)|, \quad w \in X.$$

Наконец,

$$\begin{aligned}
&q(W(f, X), W(\tilde{f}, X, \ldots, X)) \\
&\leqslant (y^{(0)} - z)\tilde{h}(y^{(1)} - z, \ldots, y^{(n)} - z) \\
&\quad - (x^{(0)} - z)\tilde{h}(x^{(1)} - z, \ldots, x^{(n)} - z) - d(X)|h(w - z)| \\
&= (y^{(0)} - z)\tilde{h}(y^{(1)} - z, \ldots, y^{(n)} - z) \\
&\quad - (y^{(0)} - z)\tilde{h}(x^{(1)} - z, \ldots, x^{(n)} - z) \\
&\quad + (y^{(0)} - z)\tilde{h}(x^{(1)} - z, \ldots, x^{(n)} - z) \\
&\quad - (x^{(0)} - z)\tilde{h}(x^{(1)} - z, \ldots, x^{(n)} - z) - d(X)|h(w - z)| \\
&= (y^{(0)} - z)(\tilde{h}(y^{(1)} - z, \ldots, y^{(n)} - z) \\
&\quad - \tilde{h}(x^{(1)} - z, \ldots, x^{(n)} - z)) \\
&\quad + (y^{(0)} - x^{(0)})\tilde{h}(x^{(1)} - z, \ldots, x^{(n)} - z) \\
&\quad - d(X)|\tilde{h}(w - z, \ldots, w - z)| \\
&\leqslant |y^{(0)} - z|\,|\tilde{h}(y^{(1)} - z, \ldots, y^{(n)} - z) \\
&\quad - \tilde{h}(x^{(1)} - z, \ldots, x^{(n)} - z)| + |y^{(0)} - x^{(0)}| \\
&\quad \times |\tilde{h}(x^{(1)} - z, \ldots, x^{(n)} - z)| - d(X)|\tilde{h}(w - z, \ldots, w - z)| \\
&\leqslant d(X)(|\tilde{h}(y^{(1)} - z, \ldots, y^{(n)} - z) - \tilde{h}(x^{(1)} - z, \ldots, x^{(n)} - z)| \\
&\quad + \|\tilde{h}(x^{(1)} - z, \ldots, x^{(n)} - z)| - |\tilde{h}(w - z, \ldots, w - z)|\|) \\
&\leqslant d(X)\Big(c^{(1)} \max_{1 \leqslant k \leqslant n} |y^{(k)} - x^{(k)}| + c^{(2)} \max_{1 \leqslant k \leqslant n} |x^{(k)} - w|\Big) \\
&\leqslant d(X)(c^{(1)} + c^{(2)})d(X) = c(d(X))^2.
\end{aligned}$$

Выполнение использованного в этом построении условия Липшица было оговорено для $\overset{0}{H}$ и как следствие для $[\overset{0}{H}]$.

Утверждение теоремы 4 также можно распространить на случай функции нескольких переменных. Обобщенное таким образом соотношение (4) было дано Хансеном; его же, но с применением другой техники получили Шуба и Миллер.

Как следствие из теоремы 3 возникает

Теорема 5. *Пусть f — вещественная функция от вещественного аргумента x, f(x) — аналитическое выражение для f. Будем считать,*

что выполнены все предположения теоремы 3. *Тогда для* $X \subseteq Y$ *имеет место неравенство*

$$d(f(X)) \leqslant cd(X), \qquad c \geqslant 0. \tag{5}$$

Доказательство. Опираясь на теорему 3 и соотношение (21 п.7.2), получаем

$$\begin{aligned} d(f(X)) &\leqslant 2q(f(X),\ W(f,\ X)) + d(W(f,\ X)) \\ &\leqslant 2c^{(1)}d(X) + d(W(f,\ X)), \qquad c^{(1)} \geqslant 0. \end{aligned}$$

Исходя из условия Липшица для функции f, можно записать неравенство

$$d(W(f,\ X)) = f(x) - f(y) \leqslant c^{(2)} |x - y|, \quad \text{где} \quad x,\ y \in X, \quad c^{(2)} \geqslant 0,$$

из которого следует

$$d(f(X)) \leqslant 2c^{(1)}d(X) + c^{(2)}d(X) = cd(X),$$

что и требовалось доказать.

Соответствующее обобщение на случай нескольких переменных выглядит так:

$$d(f(X^{(1)},\ X^{(2)},\ \ldots,\ X^{(n)})) \leqslant \sum_{k=1}^{n} c^{(k)}d(X^{(k)}) \leqslant c \max_{1 \leqslant k \leqslant n} d(X^{(k)}). \tag{5'}$$

Теперь докажем теорему о вхождении множества *W(f, X)* в другое множество, появляющееся в результате вычисления интервального выражения на основе теоремы о среднем значении.

Теорема 6. *Пусть* f *— вещественная функция от вещественного аргумента* x, *дифференцируемая на интервале* $X = [x_1,\ x_2]$, *и пусть* $f'(x)$ *— аналитическое выражение для* f' *такое, что интервальное выражение для* $f'(X)$ *определено. Тогда если для* f' *справедливы предположения теоремы* 5, *то*

$$W(f,\ X) \subseteq f(y) + f'(X)(X - y), \tag{а}$$

$$q(W(f,\ X),\ f(y) + f'(X)(X - y)) \leqslant \tilde{c}(d(X))^2, \tag{б}$$

где $y \in X$ *и константа* $\tilde{c} \geqslant 0$.

Доказательство. (а) Из теоремы о среднем, примененной к x и y из X, получаем

$$f(x) = f(y) + f'(y + \theta(x - y))(x - y), \qquad 0 < \theta < 1.$$

Из

$$y + \theta(x - y) \in y + [0,\ 1](X - y) = X$$

с учетом монотонности включения вытекает

$$f(x) \in f(y) + f'(X)(X - y),$$

что доказывает утверждение (а), (б) Пусть

$$W(f, X) = [f(u), f(v)], \qquad u, v \in X.$$

Тогда из теоремы о среднем следует, что

$$d(W(f, X)) = f(v) - f(u) = |f(v) - f(u)|$$
$$\geqslant |f(x_1) - f(x_2)| = |f'(\xi)|\, d(X), \qquad \xi \in X.$$

Формулы (12 п.7.2), (3 п.7.2) и (20 п.7.2) дают

$$d(f'(X)(X - y)) \leqslant |f'(X)|\, d(X) + d(f'(X))\, |X - y|$$
$$\leqslant |f'(X)|\, d(X) + d(f'(X))\, d(X).$$

Так как $f'(\xi) \in f'(X)$, то, принимая во внимание (21 п.7.2), получаем

$$q(f'(X), f'(\xi)) \leqslant d(f'(X)).$$

Теперь используем неравенство

$$|f'(X)| - |f'(\xi)| \leqslant q(f'(X), f'(\xi)),$$

которое следует из (4 п.7.2), (5 п.7.2) и определения 6 п.7.2. Применяя к *f(X)* соотношения (а), (21 п.7.2) и теорему 5, получаем требуемый результат:

$$q(W(f, X), f(y) + f'(X)(X - y))$$
$$\leqslant d(f(y) + f'(X)(X - y)) - d(W(f, X))$$
$$\leqslant d(f'(X))\, d(X) + (|f'(X)| - |f'(\xi)|)\, d(X)$$
$$\leqslant d(f'(X))\, d(X) + q(f'(X), f'(\xi))\, d(X)$$
$$\leqslant 2c\,(d(X))^2 = \bar{c}\,(d(X))^2.$$

Из теоремы 6 вытекает качественный результат теоремы 4 для записанного в центрированной форме выражения

$$f(y) + f'(X)(X - y), \qquad y \in X.$$

Это важный факт, поскольку уже для многочленов получение центрированной формы требует применения полной схемы Горнера Теорема 6 также может быть обобщена на случай нескольких переменных. Детали этого мы опускаем.

Рассмотрим теперь рациональную функцию

$$f(x) = p(x)/q(x),$$

где $p(x) = \sum_{\nu=0} a_\nu x^\nu$ и $q(x) = \sum_{\nu=0} b_\nu x^\nu$.

Связав *p(x)* и *q(x)* некоторыми условиями, можно указать выражения, для которых сохраняет силу свойство

$$q(W(f, X), f(X)) \leqslant c d'(X)^2, \qquad c \geqslant 0.$$

и более простые, нежели центрированная форма или использованное в теореме 6 представление, основанное на теореме о среднем.

Пусть даны $c = m(x)$— середина интервала и тейлоровские разложения $p(x) = \sum_{\nu=0}^{r} a'_\nu (x-c)^\nu$ и $q(x) = \sum_{\nu=0}^{s} b'_\nu (x-c)^\nu$. Без потери общности допустим, что $b'_0 = 1$ и $0 \notin q(X)$, где $q(X) = 1 + \sum_{\nu=1}^{s} b'_\nu (X-c)^\nu$. Если теперь

$$\operatorname{sign}(a'_1)\operatorname{sign}(b'_1 \cdot a'_0) \leqslant 0 \qquad (7)$$

и предполагается, что $p(x)$ и $q(x)$ удовлетворяют неравенствам

$$d(p(X)) \leqslant c_1 d(X),$$
$$d(q(X)) \leqslant c_2 d(X),$$

то для интервального выражения

$$f(X) = \sum_{\nu=0}^{r} a'_\nu (X-c)^\nu \Big/ \left(1 + \sum_{\nu=1}^{s} b'_\nu (X-c)^\nu\right)$$

выполнено свойство (6). Это утверждение справедливо для обоих вышеприведенных выражений независимо от того, вычисляются они с помощью степеней $X - c$ либо по схеме Горнера. Если мы по-прежнему находимся в условиях предположения (7) и

$$0 \notin 1 + (X-c)\,q'(X),$$

то в (6) можно подставить выражение

$$f(X) = \frac{a'_0 + (X-c)\,p'(X)}{1 + (X-c)\,q'(X)}.$$

Сомножитель $p'(X)$ представляет собой интервальную оценку для первой производной функции $p(x)$, удовлетворяющую соотношению $d(p'(X)) \leqslant \alpha d(X)$. Аналогично, $q'(X)$ — интервальная оценка для первой производной $q(x)$, удовлетворяющая неравенству $d(q'(X)) \leqslant \beta d(X)$.

В дальнейшем мы рассмотрим методы локализации нулей, использующие включения на участках монотонности функции.

Ниже мы дадим ряд возможных включений, применимых к отношению разностей. Эти включения будут частично упорядочены Оказывается, чго оптимальное включение может быть описано просто и систематично и что вычисления с соответствующими итерациями могут быть выполнены с теми же вычислительными затратами, что и интервальное оценивание производной. Эти включения выводятся другими способами, значительно отличающимися от использовавшихся у Хансена, решавшего ту же задачу

Включения для примеров, приведенных Хансеном, в точности соответствуют оптимальным включениям для этих же примеров, полученных методами, которые изложены в данной работе. Далее в рассуждениях следуем работе Алефельда.

Пусть имеется многочлен

$$p(x) = \sum_{\nu=0}^{n} a_\nu x^\nu.$$

Справедливость двух нижеследующих равенств очевидна:

$$p(x) - p(y) = \sum_{i=0}^{n} a_i (x^i - y^i) = \left(\sum_{i=1}^{n} a_i \sum_{j=1}^{i} x^{i-j} y^{j-1} \right)(x - y) \qquad (8)$$

$$= \left(\sum_{i=1}^{n} \left(\sum_{j=i}^{n} a_j y^{j-i} \right) x^{i-1} \right)(x - y),$$

$$p(x) - p(y) = \sum_{i=0}^{n} a_i (x^i - y^i) = \left(\sum_{i=1}^{n} a_i \sum_{j=1}^{i} y^{i-j} x^{j-1} \right)(x - y) \qquad (9)$$

$$= \left(\sum_{i=1}^{n} \left(\sum_{j=i}^{n} a_j x^{j-i} \right) y^{i-1} \right)(x - y).$$

Для фиксированного y и произвольного x из X с помощью (8) и свойства монотонности включения получаем

$$\frac{p(x) - p(y)}{x - y} \in \left(\sum_{i=1}^{n} c_{i-1} X^{i-1} \right)_{\mathrm{H}} =: J_1 \subseteq J_2 := \sum_{i=1}^{n} c_{i-1} X^{i-1},$$

где

$$c_{i-1} = \sum_{j=i}^{n} a_j y^{j-i}, \qquad 1 \leqslant i \leqslant n.$$

Здесь и далее буква H обозначает, что выражение, которое ею помечено, вычисляется по схеме Гопнепа. В многочлене J_2 степени X' вычисляются по правилу $X^0 = 1$, $X^r = X^{r-1} X$ при $r \geq 1$.

Включение $J_1 \subseteq J_2$ следует из закона субдистрибутивности. Для вещественного числа y и интервалов A_j, $0 \leqslant j \leqslant n-1$, всегда

$$\sum_{i=1}^{n} A_{i-1} y^{i-1} = \left(\sum_{i=1}^{n} A_{i-1} y^{i-1} \right)_{\mathrm{H}}.$$

При фиксированном y и произвольном x из X, $x \neq y$, используя субдистрибутивность, данное равенство и формулу (9), получаем

$$\frac{p(x)-p(y)}{x-y} \in \sum_{i=1}^{n} (C_{i-1})_{\Pi}\, y^{i-1} = \left(\sum_{i=1}^{n} (C_{i-1})_{\Pi}\, y^{i-1}\right)_{\Pi} =: J_3$$

$$\subseteq J_4 := \sum_{i=1}^{n} C_{i-1} y^{i-1} = \left(\sum_{i=1}^{n} C_{i-1} y^{i-1}\right)_{\Pi}.$$

где

$$(C_{i-1})_{\Pi} = \left(\sum_{j=i}^{n} a_j X^{j-i}\right)_{\Pi}, \qquad 1 \leqslant i \leqslant n,$$

и

$$C_{i-1} = \sum_{j=i}^{n} a_j X^{j-i}, \qquad 1 \leqslant i \leqslant n.$$

Докажем теперь еще одну теорему.

Теорема 7. *Введенные выше выражения удовлетворяют соотношениям*

$$J_1 \subseteq J_2 \subseteq J_4, \tag{a}$$

$$J_1 \subseteq J_3 \subseteq J_4, \tag{b}$$

$$J_4 \subseteq p'(X) = \sum_{\nu=1}^{n} \nu a_\nu X^{\nu-1}. \tag{c}$$

Доказательство. Для простоты ограничимся случаем многочлена четвертой степени (n = 4). Общий случай может быть рассмотрен аналогичным образом.

(a) и (c).Нам достаточно показать, что $J_2 \subseteq J_4 \subseteq p'(X)$. Из свойства монотонности включения и соотношения (8 п. 7.1) получаем

$$\begin{aligned}
J_2 &= \sum_{i=1}^{n} c_{i-1} X^{i-1} \\
&= (a_1 + a_2 y + a_3 y^2 + a_4 y^3) X^0 + (a_2 + a_3 y + a_4 y^2) X \\
&\quad + (a_3 + a_4 y) X^2 + a_4 X^3 \\
&\subseteq a_1 + a_2 X + a_3 X^2 + a_4 X^3 + a_2 y + a_3 y X + a_4 y X^2 \\
&\quad + a_3 y^2 + a_4 y^2 X + a_4 y^3 \\
&= a_1 + a_2 X + a_3 X^2 + a_4 X^3 + (a_2 + a_3 X + a_4 X^2) y \\
&\quad + (a_3 + a_4 X) y^2 + a_4 y^3 = J_4 \\
&\subseteq a_1 + a_2 X + a_3 X^2 + a_4 X^3 + a_2 X + a_3 X^2 + a_4 X^3 \\
&\quad + a_3 X^2 + a_4 X^3 + a_4 X^3 = p'(X).
\end{aligned}$$

Достаточно показать, что

$$J_1 \subseteq J_3. \qquad \text{(b)}$$

$$\begin{aligned}
J_1 &= ((c_3X + c_2)X + c_1)X + c_0 \\
&= ((a_4X + (a_3 + a_4y))X + a_2 + a_3y + a_4y^2)X + a_1 + a_2y + a_3y^2 + a_4y^3 \\
&\subseteq ((a_4X + a_3)X + a_4yX + a_2 + a_3y + a_4y^2)X + a_1 + a_2y + a_3y^2 + a_4y^3 \\
&= (((a_4X + a_3)X + a_2) + a_4yX + a_3y + a_4y^2)X + a_1 + a_2y + a_3y^2 + a_4y^3 \\
&= (((a_4X + a_3)X + a_2) + (a_4X + a_3)y + a_4y^2)X + a_1 + a_2y + a_3y^2 + a_4y^3 \\
&\subseteq ((a_4X + a_3)X + a_2)X + (a_4X + a_3)yX + a_4y^2X + a_1 + a_2y + a_3y^2 + a_4y^3 \\
&= (((a_4X + a_3)X + a_2)X + a_1)y^0 + ((a_4X + a_3)X + a_2)y \\
&\quad + (a_4X + a_3)y^2 + a_4y^3 = J_3.
\end{aligned}$$

Итак, теорема доказана.

Нельзя ответить в общем случае на вопрос о том, какое из выражений: J_2 или J_3 — дает лучшее включение. Возможно и $J_2 \subseteq J_3$, и $J_3 \subseteq J_2$. Пусть, например,

$$p(x) = x^3 - x^2, \quad X = [-1, 2], \quad y = 1.$$

Тогда имеем

$$J_2 = (a_1 + a_2y + a_3y^2)X^0 + (a_2 + a_3y)X + a_3X^2 = X^2 = [-2, 4]$$

и

$$\begin{aligned}
J_3 &= ((a_3X + a_2)X + a_1)y^0 + (a_3X + a_2)y + a_3y^2 \\
&= (X - 1)X + (X - 1) + 1 = [-5, 4].
\end{aligned}$$

Здесь $J_2 \subset J_3$.

Если, с другой стороны, $y = 0$ и, следовательно, $c_{i-1} = a_i$, $1 \leqslant i \leqslant n$, то получаем

$$J_2 = \sum_{i=1}^{n} a_iX^{i-1} \quad \text{и} \quad J_3 = \left(\sum_{i=1}^{n} a_iX^{i-1}\right)_{\text{и}}.$$

Теперь $J_3 \subseteq J_2$.

Рассмотрим снова пример с $p(x) = x^3 - x^2$ при $y = 0$ и $X = [0, 2]$. В этом случае

$$J_2 = X^2 - X = [-2, 4] \quad \text{и} \quad J_3 = (X - 1)X = [-2, 2],$$

откуда имеем $J_3 \subset J_2$.

Получение интервалов J_1 и J_2 с помощью теоремы 7 требует предварительного нахождения $c_{i-1} = \sum_{j=i}^{n} a_jy^{j-i}$, $1 \leqslant i \leqslant n$. Если вычисляется значение многочлена $p(x)$ в точке y, что встречается, например, в итерационных методах , которые будут пассмотрены дальше, то нахождение c_{i-1} не требует выполнения каких-либо

дополнительных арифметических операций. Значения c_{i-1} могут быть найдены в процессе вычисления $p(y)$. Пусть, как и ранее,

$$p(x) = \sum_{i=0}^{n} a_i x_i.$$

Воспользуемся схемой Горнера

$$p_n := a_n,$$

и для $i = n, n-1, \ldots, 1$

$$p_{i-1} := p_i y + a_{i-1},$$

откуда $p_0 = p(y)$. По определению

$$\begin{aligned} c_{n-1} &= a_n && (= p_n), \\ c_{n-2} &= a_n y + a_{n-1} && (= p_{n-1}), \\ &\vdots && \vdots \\ c_0 &= c_1 y + a_1 && (= p_1). \end{aligned}$$

Следовательно, $c_{i-1} = p_i$, $1 \leqslant i \leqslant n$.

Примеры.

$$p(x) = x^4 - 1, \quad x = [0.5, 3.5], \quad y = 2 \qquad \text{(a)}$$

Имеем

$$J_1 = J_2 = J_3 = J_4 = [10.625, 89.375],$$
$$p'(X) = (p'(X))_{\text{H}} = [0.5, 171.5].$$

Оценка J_1 совпадает с той, которую получил на этих же данных Хансен.

$$p(x) = x^3 + 4x - 16, \quad X = [-1, 3], \quad y = 1. \qquad \text{(b)}$$

Получаем

$$J_1 = J_2 = J_3 = J_4 = [1, 17],$$
$$p'(X) = (p'(X))_{\text{H}} = [-5, 31],$$

что совпадает с результатом Хансена.

$$p(x) = \sum_{i=0}^{n} a_i x^i, \quad 0 \in X, \quad y = 0. \qquad \text{(c)}$$

В этом случае

$$c_0 = a_1, \quad c_1 = a_2, \ \ldots, \ c_{n-1} = a_n$$

и

$$J_1 = \left(\sum_{i=1}^{n} c_{i-1}X^{i-1}\right)_H = \left(\sum_{i=1}^{n} a_i X^{i-1}\right)_H.$$

$$p(x) = x^3 - x^2, \quad X = [1, 3], \quad y = 2. \tag{d}$$

$$J_1 = J_2 = J_3 = [4, 14] \subset [2, 16] = J_4 \subset (p'(X))_H$$

$$= [1, 21] \subset [-3, 25] = p'(X),$$

что опять совпадает со значением, вычисленным Хансеном.

Однако при $X = [-1, 2]$ и $y = 1$

$$J_1 = J_2 = [-2, 4], \quad J_3 = [-5, 4], \quad J_4 = [-5, 7],$$

$$(p'(X))_H = [-10, 8], \quad p'(X) = [-10, 14].$$

Пусть $x_0 \in X$ и $f \in C^{n+1}(X)$. (e)

Используя тейлоровское разложение, получаем

$$f(x) = p(x) + \varphi(x),$$

где

$$\varphi(x) = \int_{x_0}^{x} \frac{(x-t)^n}{n!} f^{(n+1)}(t)\, dt$$

и

$$p(x) = \sum_{k=0}^{n} \frac{(x-x_0)^k}{k!} f^{(k)}(x_0).$$

Функция φ дифференцируема и

$$\varphi'(x) = \int_{x_0}^{x} \frac{(x-t)^{n-1}}{(n-1)!} f^{(n+1)}(t)\, dt.$$

Интегральная теорема о среднем дает формулу

$$\varphi'(x) = f^{(n+1)}(\eta) \int_{x_0}^{x} \frac{(x-t)^{n-1}}{(n-1)!}\, dt = \frac{(x-x_0)^n}{n!} f^{(n+1)}(\eta),$$

где η лежит между x и x_0. Применяя к φ теорему о среднем, получаем

$$f(x) - f(y) = p(x) - p(y) + \varphi(x) - \varphi(y)$$

$$= \left\{\sum_{k=1}^{n} c_{k-1}(x-x_0)^{k-1} + \varphi'(\xi)\right\}(x-y),$$

где

$$c_{k-1} = \sum_{i=k}^{n} (y-x_0)^{i-k} \frac{f^{(k)}(x_0)}{k!}, \quad 1 \leqslant k \leqslant n,$$

и

$$\varphi'(\xi) = \frac{(\xi - x_0)^n}{n!} f^{(n+1)}(\eta),$$

причем ξ лежит между x и y, а η — между x_0 и ξ. Полагая $y = x_0$, имеем

$$c_0 = f'(x_0)/1!, \ \ldots, \ c_{n-1} = f^{(n)}(x_0)/n!.$$

Если для $(n+1)$-й производной существует вычислимое интервальное выражение, то для $y = x_0$

$$\frac{f(x) - f(y)}{x - y} \in \sum_{k=1}^{n} \frac{f^{(k)}(x_0)}{k!} (X - x_0)^{k-1} + f^{(n+1)}(X) \frac{(X - x_0)^n}{n!},$$

поскольку $\eta, \xi \in X$. Эта оценка снова совпадает с данной Хансеном.

$$p(x) = x^7 + 3x^6 - 4x^5 - 12x^4 - x^3 - 3x^2 + 4x + 12,$$
$$X = [1.8, 3], \quad y = 2. \qquad \text{(f)}$$

Теперь

$$J_1 = [173.2362, 2400], \qquad J_2 = [161.4762, 2411.76],$$
$$J_3 = [24.72, 2400], \qquad J_4 = [-870.2933, 3443.5296],$$
$$(p'(X))_H = [71.799808, 6520], \qquad p'(X) = [-2378.791292, 8970.592].$$

Сделанные утверждения могут быть распространены на многомерный случай.

Замечания. В этом микромодуле было рассмотрено интервальное оценивание вещественных функций. Мы сознательно не говорили о произвольных отображениях из $I(\mathrm{R})$ в $I(\mathrm{R})$. Приложения в последующих микромодулях требуют использования многих свойств, которые возможно доказать только для интервальных оценок. Если разрешить на $I(\mathrm{R})$ отображения более общего вида, то для каждого приложения необходимо описать множество условий. Следующий пример показывает, сколь велик класс отображений из $I(\mathrm{R})$ в $I(\mathrm{R})$. Единственное ограничение здесь состоит в том, что если область определения сузить до R, то множество значений также будет принадлежать R. Итак, пусть f — вещественная функция, $f(x)$ — ее аналитическое выражение, $f(X)$ — оценивающая функция для $f(x)$. Тогда для произвольной $\varphi(x)$ такой, что $\varphi(0)=0$,

$$\Psi(X) = f(X) + \varphi(d(X))[-1, 1]$$

определяет отображение из $I(\mathrm{R})$ в $I(\mathrm{R})$. Очевидно, что $\Psi([x, x]) \in \mathbb{R}$. Если $\varphi(x) \geqslant 0$ при $x \geqslant 0$, то $W(f, X) \subseteq \Psi(X)$, и если $\varphi(x)$ монотонна и не убывает при $x \geq 0$, то $\Psi(X)$ обладает свойством монотонности включения в форме (1'). Этот пример

показывает, что, соответствующим образом выбирая $\varphi(x)$, можно строить отображения Ψ, имеющие различные свойства. Если же потребовать от отображений из $I(\mathrm{R})$ в $I(\mathrm{R})$ выполнения всех свойств интервальных оценок, то не найдется никаких других полезных отображений, кроме в точности этих оценок.

Покажем, как можно сократить доказательство теоремы 4, используя теорему 5 примерно так же, как и при доказательстве теоремы 6. Записанная в центрированной форме интервальная оценка

$$f(X) = f(z) + (X - z)h(X - z)$$

удовлетворяет соотношению

$$W(f, X) \subseteq f(X).$$

В соответствии с (21 п.7.2) получаем

$$q(W(f, X), f(X)) \leqslant d(f(X)) - d(W(f, X)).$$

Пусть теперь

$$\min_{x \in X} |h(x - z)| = |h(w - z)|.$$

Тогда

$$f(z) + (X - z)h(w - z) \subseteq f(z) + \{(x - z)h(x - z) \mid x \in X\} = W(f, X).$$

С учетом (9 п.7.2), (14 п.7.2) и приведенного выше включения имеем

$$d(W(f, X)) \geqslant d((X - z)h(w - z)) = d(X)|h(w - z)|, \quad w \in X.$$

Исходя из (10 п.7.2), (12 п.7.2), (3 п.7.2) и (20 п.7.2), получаем

$$\begin{aligned} d(f(X)) &= d(f(z) + (X - z)h(X - z)) = d((X - z)h(X - z)) \\ &\leqslant |X - z| d(h(X - z)) + d(X)|h(X - z)| \\ &\leqslant d(X) d(h(X - z)) + d(X)|h(X - z)|. \end{aligned}$$

Поскольку $h(w - z) \in h(X - z)$, из (21 п.7.2) следует, что

$$q(h(X - z), h(w - z)) \leqslant d(h(X - z)).$$

Определение 6 п.7.2 и соотношения (4 п.7.2), (5 п.7.2) дают неравенство

$$|h(X - z)| - |h(w - z)| \leqslant q(h(X - z), h(w - z)).$$

На основе приведенных выше неравенств получаем

$$\begin{aligned} q(W(f, X), f(X)) &\leqslant d(f(z) + (X - z)h(X - z)) - d(W(f, X)) \\ &\leqslant d(X)d(h(X - z)) + d(X)\,|h(X - z)| - d(X)|h(w - z)| \\ &= d(X)d(h(X - z)) + (|h(X - z)| - |h(w - z)|)d(X) \\ &\leqslant d(X)d(h(X - z)) + q(h(X - z), h(w - z))d(X) \\ &\leqslant d(X) \cdot 2 \cdot d(h(X - z)). \end{aligned}$$

И наконец, после применения теоремы 5 к выражению $h(X — z)$ оказывается, что

$$q(W(f, X), f(X)) \leqslant d(X) \cdot 2 \cdot \tilde{c} \cdot d(X) = c(d(X))^2.$$

При выполнении некоторых условий дифференцируемости для аналитических выражений можно определить общие условия, при которых соотношение (4) справедливо. При этом теоремы 4 и 6 получаются как частные случаи. У Херцбергера можно найти простую интерполяционную формулу, вычисляющую множество значений для семейства многочленов с заданными коэффициентами Им используется тот вытекающий из теоремы 2 факт, что интервальное оценивание позволяет точно вычислить множество значений функции, когда все переменные и параметры входят в аналитическое выражение лишь по одному разу.

Корнелиус и Лонер предложили выражение *f(X),* для которого $q(W(f, X), f(X)) \leqslant cd(X)^{s+1}$, $s \geqslant 1$. При s = 1, 2, 3 *f(X)* вычисляется с помощью весьма простого алгоритма.

Микромодуль 25
Машинная и комплексная интервальная арифметика

7.3. Машинная интервальная арифметика

Теперь мы остановимся на вопросах реализации интервальных операций на цифровой вычислительной машине Обшеизвестно, что в машине может быть представлено лишь конечное множество чисел. Чаще всего они записываются в полулогарифмической форме, а точнее — в форме с плавающей точкой:

$$x = m \cdot b^e.$$

Здесь m — мантисса, b — основание степени, e — порядок. Как правило, для внутримашинного представления выбирается основание b, равное 2, а мантисса нормализуется, т. е. ее абсолютное значение

помещается в интервал [1/2, 1). Целое e принадлежит интервалу $[e_{\min}, e_{\max}]$.

Множество машинных чисел описанного типа обозначим через R_M, и всюду далее будем предполагать, что оно симметрично относительно нуля, т. е. $R_M = — R_M$. Для аппроксимации вещественных чисел, лежащих в интервале $[\min\limits_{y \in R_M} y;\ \max\limits_{y \in R_M} y]$, можно с успехом использовать машинные числа $\{\tilde{x} \mid \tilde{x} \in \mathbb{R}_M\}$. Аппроксимация достигается применением отображения

$$fl\colon \mathbb{R} \ni x \to \tilde{x} = fl(x) \in \mathbb{R}_M. \tag{1}$$

Это отображение называется округлением, если выполнено свойство

$$x \leqslant y \Rightarrow fl(x) \leqslant fl(y) \quad \text{(монотонность)}. \tag{2}$$

Округление, которое отображает R_M в R_M так, что

$$x \in \mathbb{R}_M \Rightarrow fl(x) = x, \tag{3}$$

называется оптимальным (приведенное определение не является стандартным. Обычно под оптимальным понимают такое округление, которое отображает округляемое число x в $\mathscr{Y}$, ближайшее в некотором смысле к x.)

Особый интерес представляют так называемые направленные округления. Если для округления $\downarrow$ справедлива импликация

$$x \in \mathbb{R} \Rightarrow \downarrow x \leqslant x, \tag{4}$$

то говорят об округлении вниз. Аналогично,

$$\uparrow x := -(\downarrow(-x)), \qquad x \leqslant \mathbb{R}, \tag{5}$$

определяет округление вверх. Техника выполнения этих округлений для различных способов представления чисел неоднократно освещалась в литературе. Подобно тому как вещественные числа приближаются с помощью машинных, можно вещественные интервалы приближать машинными интервалами. В этом случае интервал X из $I(R)$, для которого справедливо соотношение $X \subseteq [\min\limits_{y \in R_M} y,\ \max\limits_{y \in R_M} y]$, заменяется соответствующим машинным интервалом из множества

$$I(\mathbb{R}_M) = \{[x_1, x_2] \mid x_1, x_2 \in \mathbb{R}_M,\ x_1 \leqslant x_2\} \subset I(\mathbb{R}).$$

Для того чтобы основные свойства интервальных операций выполнялись и для их машинных аналогов, применяется округление интервалов (интервальное округление)

$$\updownarrow : I(\mathbb{R}) \ni X \to \updownarrow X \in I(\mathbb{R}_M),$$

причем

$$X \in I(\mathbb{R}) \Rightarrow X \subseteq \updownarrow X \tag{6}$$

и

$$X, Y \in I(\mathbb{R}), \quad X \subseteq Y \Rightarrow \updownarrow X \subseteq \updownarrow Y. \tag{7}$$

Если рассмотреть переход от интервала $X = [x_1, x_2]$ из $I(\mathrm{R})$ к его машинному представлению $\tilde{X} = [\tilde{x}_1, \tilde{x}_2]$, то окажется, что (7) означает, что необходимо осуществить этот переход путем округления каждой из границ X. Из (6) следует, что границы должны быть округлены направленно. Таким образом, округление интервала X состоит в нахождении $\updownarrow X$ по правилу

$$\updownarrow X = \updownarrow [x_1, x_2] = [\downarrow x_1, \uparrow x_2]. \tag{8}$$

Проведенное обсуждение показывает, что для того, чтобы округлить интервал, достаточно иметь $\downarrow$ — направленное округление вниз. С другой стороны, $\uparrow$ и $\downarrow$ не обязательно должны быть связаны соотношением (5).

Если над двумя машинными числами x и y из $\mathrm{R_M}$ производится машинная операция *, где $* \in \{+, -, \cdot, :\}$, то ее результатом оказывается новое число z из $\mathrm{R_M}$. Проигнорировав возможность выхода за пределы допустимого диапазона (переполнение и антипереполнение), можно, используя соответствующее округление *fl*, представить z в виде

$$z = fl(x * y). \tag{9}$$

Таким же образом мы можем определить результат машинной операции над интервалами.

Определение 1. Пусть $A, B \in I(\mathbb{R}_M)$, $* \in \{+, -, \cdot, :\}$, $\updownarrow$ — интервальное округление. Тогда результат операции *, выполненной над A и B с применением $\updownarrow$ есть

$$C = \updownarrow (A * B) \in I(\mathbb{R}_M). \tag{10}$$

Теперь мы покажем, что основные свойства интервальной арифметики при использовании этого определения сохраняются.

Теорема 2. *Для машинных интервальных операций, задаваемых определением* 1, *справедливо следующее утверждение:*

$$A^{(k)}, B^{(k)} \in I(\mathbb{R}_M), * \in \{+, -, \cdot, :\}, \quad A^{(k)} \subseteq B^{(k)}, \quad k = 1, 2, \tag{11}$$
$$\Rightarrow C^{(1)} = \updownarrow (A^{(1)} * A^{(2)}) \subseteq C^{(2)} = \updownarrow (B^{(1)} * B^{(2)}).$$

Доказательство теоремы 2 следует непосредственно из свойства интервальных округлений (7).

Утверждение (11) отражает не что иное, как свойство монотонности включения (9 п.7.1) применительно к машинным интервальным операциям.

Очередная теорема представляет интерес с точки зрения оценки погрешностей округлений.

Теорема 3. *Пусть* $\updownarrow$ *— интервальное округление, сводящееся с помощью* (8) *к направленным округлениям* $\downarrow$ *и* $\uparrow$, *и пусть* $* \in \{+, -, \cdot, :\}$. *Тогда*

$$A,\ B \in I(\mathbb{R}_M) \Rightarrow A * B \subseteq C = \updownarrow (A * B) \in I(\mathbb{R}_M),$$
$$a \in A,\ b \in B \Rightarrow a * b \in C = \updownarrow (A * B) \in I(\mathbb{R}_M). \qquad (12)$$

Если имеется округление fl, применение которого приводит к выполнению неравенства

$$\downarrow a \leqslant fl(a) \leqslant \uparrow a, \quad a \in \mathbb{R},$$

то для x, y, z из $\mathrm{R_M}$ *справедливо*

$$z = fl(x * y) \in Z = \updownarrow ([x, x] * [y, y]) \in I(\mathbb{R}_M). \qquad (13)$$

Доказательство свойств (12) и (13) мы опускаем, поскольку оно элементарно и следует непосредственно из соответствующих определений.

Интервальное оценивание аналитического выражения функции, проведенное с использованием операций из определения 1, дает интервалы, объемлющие значения оценивающей функции. Среди этих интервалов находятся и оценки множества значений функции. Более того, при выполнении подобных вычислений сохраняется свойство монотонности включения.

На практике машинные интервальные операции реализуются с помощью соответствующих программно-аппаратных средств Эти средства могут служить поддержкой языка программирования высокого уровня. Один из вариантов реализации — набор подпрограмм, написанных, скажем, на Алголе. Рассмотрим вкратце последнюю возможность. В большинстве случаев, в частности у Криста, такой набор содержит средство, с помощью которого выполняется округление $\downarrow$. Это средство может быть, например, оформлено в виде процедуры-функции LOW; через нее определяются стандартные операции интервальной арифметики — ADD, SUB, MUL и DIV, а также элементарные функции. На деталях реализации подобных подпрограмм мы остановимся в приложении В.

Посмотрим теперь на алгоритмы, описанные в терминах вещественных чисел. К их числу можно отнести, скажем, схему Горнера или метод Гаусса. Если такой алгоритм реализуется на компьютере, т.

е. с использованием машинной арифметики, то даже исходные данные в общем случае не могут быть представлены точно. Возникающие при этом трудности преодолеваются применением машинной интервальной арифметики. Исходные данные просто заключаются в интервалы, имеющие своими границами машинные числа. Если теперь представить себе, что алгоритм будет выполняться без учета погрешностей округления, то, как показано в микромодулю 24, ширина реализующего интервала станет, вообще говоря, возрастать в большей степени, чем это обусловлено исходными данными. При наличии погрешностей округления описанное свойство проявляется еще сильнее.

Обсудим следующий вопрос: на какой рост точности результата можно рассчитывать, если перейти от алгоритма, использующего машинную интервальную арифметику с t_1 цифрами в мантиссе, к алгоритму, использующему t_2-значную арифметику, где $t_2>t_1$? Предполагается, что при таком переходе диапазон возможных значений порядка остается неизменным. Таким образом, все числа, представимые с t_1 цифрами, столь же точно записываются с t_2 цифрами.

Пусть $x \in K$, $x \neq 0$ и

$$x = \left(\sum_{\nu=-1}^{-\infty} a_\nu b^\nu \right) b^e, \quad 1 \leqslant a_{-1} \leqslant b-1, \quad 0 \leqslant a_\nu \leqslant b-1, \quad \nu \leqslant -2.$$

Чтобы гарантировать единственность представления x, мы предполагаем, что не существует ν_0 такого, что при $\nu \leq \nu_0$ все a_ν равны b — 1. Будем также считать, что число x не представимо в t_1-значной системе с плавающей точкой. (Если бы последнее допущение отсутствовало, то следующее рассуждение было бы совершенно излишним.) Пусть, кроме того, интервальное округление (8) осуществляется через оптимальное округление границ. В соответствии с (8) при $x > 0$ получаем

$$\updownarrow x = \updownarrow [x, x] = [\downarrow x, \uparrow x],$$

где

$$\downarrow x = \left(\sum_{\nu=-1}^{-t} a_\nu b^\nu \right) b^e, \quad \uparrow x = \left(\sum_{\nu=-1}^{-t_1} a_\nu b^\nu \right) b^e + b^{-t_1+e}.$$

Очевидно, что ширина $\updownarrow x$ есть

$$d(\updownarrow x) = b^{-t_1+e}.$$

Точно такой же результат получается для ширины $\updownarrow x$ при $x < 0$. В дальнейшем зависимость результата от длины мантиссы будет отражаться с помощью записи $fl_1(x)$ (соответственно $fl_2(x)$). Следовательно, под fl мы будем понимать интервальное округление

вещественного числа (а впоследствии и вещественного интервала). Предыдущее равенство теперь может быть переписано в виде

$$d(fl_1(x)) = b^{-t_1+e}.$$

Аналогично, для мантиссы длины $t_2 = t_1 + l$ получаем

$$d(fl_2(x)) \leqslant b^{-t_1+e-l}.$$

Неравенство выполняется как строгое в случае, когда x представляется точно с t_2-разрядной мантиссой. Как следствие

$$d(fl_2(x)) \leqslant b^{-l} d(fl_1(x)). \tag{14}$$

Из предположений, сформулированных для интервальных округлений, следует, что для двух машинных интервалов A и B

$$\updownarrow(A * B) = fl_1(A * B) = [(1 - \varepsilon_1)(A * B)_1, (1 + \varepsilon_2)(A * B)_2]$$

(см. определение 1). С помощью $(A*B)_1$ и $(A*B)_2$ вычисляются границы точного значения результата, причем

$$-\varepsilon_1(A * B)_1 \leqslant 0, \quad \varepsilon_2(A * B)_2 \geqslant 0$$

и

$$|\varepsilon_1|, |\varepsilon_2| \leqslant b^{1-t_1}.$$

Следовательно, можно записать

$$fl_1(A * B) = A * B + [-\varepsilon_1(A * B)_1, \varepsilon_2(A * B)_2]. \tag{15a}$$

Оценкой ширины результата служит

$$d(fl_1(A * B)) \leqslant d(A * B) + 2b^{1-t_1}|A * B|. \tag{15b}$$

Эта оценка показывает, что когда используется мантисса фиксированной длины, то рост ширины $d(fl_1(A * B))$ определяется величиной $|A * B|$. Пусть мы знаем, что x принадлежит интервалу X из $I(\mathrm{R})$. Естественно выбрать некоторое x из X в качестве приближенного значения x. Оценим абсолютную и относительную погрешность такого приближения:

$$|x - \tilde{x}| \leqslant d(X) =: \Delta(X), \tag{16}$$

и если $0 \notin X$, $x \neq 0$, то

$$\left|\frac{x - \tilde{x}}{\tilde{x}}\right| \leqslant \frac{d(X)}{\min\{|x| \mid x \in X\}} =: \rho(X). \tag{17}$$

Теорема 4. *Пусть A, B, C и D — машинные интервалы, причем*

$$A \subseteq C, \quad B \subseteq D, \tag{18}$$

и

$$\begin{aligned} d(C) &\leqslant s_1, & d(D) &\leqslant s_2, \\ d(A) &\leqslant b^{-t}s_1, & d(B) &\leqslant b^{-t}s_2. \end{aligned} \tag{19}$$

Предположим, что * *— одна из арифметических операций над вещественными интервалами и* $0 \notin fl_1(C * D)$. *Тогда границы для* $\Delta(fl_2(A * B))$ *(соответственно* $\rho(fl_2(A * B))$*) оказываются в* b^l *раз меньше, чем границы для* $\Delta(fl_1(C * D))$ *(соответственно* $\rho(fl_1(C * D))$*).*

Доказательство. Используя (15b), (10 п.7.2), (12 п.7.2), неравенство

$$d(1/X) \leqslant |1/X|^2 d(X) \quad (0 \notin X),$$

а также первую строку из (19), сразу же получаем

$$d(fl_1(C * D)) \leqslant d(C * D) + 2b^{1-t_1}|C * D| \leqslant \left\{ \begin{array}{ll} s_1 + s_2, & * = +, - \\ |C| s_2 + s_1 |D|, & * = \cdot \\ |C|\,|1/D|^2 s_2 + |1/D| s_1, & * = : \end{array} \right\} + 2b^{1-t_1}|C * D|.$$

На основе (18) и (19) аналогичным способом доказывается, что

$$d(A * B) \leqslant b^{-t} \left\{ \begin{array}{ll} s_1 + s_2, & * = +, - \\ |C| s_2 + s_1 |D|, & * = \cdot \\ |C|\,|1/D|^2 s_2 + |1/D| s_1, & * = : \end{array} \right\}. \quad (20)$$

Из (18) и теоремы 2 следует включение

$$fl_2(A * B) \subseteq fl_2(C * D) \subseteq fl_1(C * D).$$

Оно справедливо, поскольку мы предположили, что интервальное округление задано через оптимальные округления границ. Таким образом,

$$\min\{|x| \mid x \in fl_2(A * B)\} \geqslant \min\{|x| \mid x \in fl_1(C * D)\}. \quad (21)$$

Из (15b), (20) и неравенства $|A * B| \leqslant |C * D|$ вытекает

$$d(fl_2(A * B)) \leqslant d(A * B) + 2b^{1-t_1-t}|C * D| \leqslant \left\{ \begin{array}{ll} s_1 + s_2, & * = +, - \\ |C| s_2 + s_1 |D|, & * = \cdot \\ |C|\,|1/D|^2 s_2 + |1/D| s_1, & * = : \end{array} \right\} + 2b^{1-t_1-t}|C * D|.$$

Это доказывает утверждение теоремы для верхних границ абсолютной погрешности. Для верхних границ относительной погрешности требуемый результат получается непосредственно из (21).

Простое, но важное следствие из теоремы 4 содержит

Теорема 5. *Допустим, что справедливы все приведенные выше предположения, касающиеся машинной интервальной арифметики. Кроме того, имеется заданный в поле вещественных чисел алгоритм, который выполняется в машинной интервальной арифметике с мантиссой длины* t_1. *Если затем этот алгоритм выполнить в*

арифметике с мантиссой длины t_2, где $t_2 = t_1 + l$, $l > 0$, то границы абсолютной и относительной погрешностей уменьшатся в b^l раз. (Под алгоритмом здесь понимается однозначно определенная последовательность арифметических операций вместе с конкретными входными данными.)

Доказательство. Из (14) следует, что в нашем случае интервальное округление входных данных удовлетворяет важному предположению (19) теоремы 4. Свойства интервальной арифметики обеспечивают справедливость (18). Окончательно доказательство получаем из теоремы 4 применением полной индукции. Теорема 5 указывает способ получения результата с наперед заданной абсолютной или относительной точностью. Пусть, например, d_1 — наибольшая ширина, которую имеют результирующие интервалы, вычисленные с помощью t_1-значной мантиссы, а e — требуемая абсолютная точность. Если $d_1 \leqslant \varepsilon$, то цель достигнута. В противном случае число цифр в мантиссе увеличивается на l, где l удовлетворяет неравенству

$$b^{-l} d_1 \leqslant \varepsilon.$$

(Такой выбор не гарантирует, что абсолютная погрешность уменьшится в b^l раз. В соответствии с теоремой 5 эта оценка верна лишь для верхней границы абсолютной погрешности.)

Факты, обсужденные и доказанные в теореме 5, были изучены Румпом; он же проиллюстрировал их числовыми примерами. Один из этих примеров — решение системы уравнений, задаваемой матрицей Гильберта размерности 7×7, причем в правой части каждого уравнения стоит 1. Результаты применения к этой системе алгоритма Гаусса, реализованного в машинной интервальной арифметике с 15, 20, 25, 30 и 35 цифрами в мантиссе, воспроизведены в табл. 1.

Таблица 1

Верхняя граница $\rho(X_i)$ относительной погрешности в алгоритме Гаусса

Число цифр в мантиссе, i	15	20	25	30	35
1	$> 1^a$	0.11×10^{-3}	0.11×10^{-8}	0.11×10^{-13}	0.11×10^{-18}
2	0.34×10^{0}	0.29×10^{-5}	0.29×10^{-10}	0.29×10^{-15}	0.29×10^{-20}
3	0.18×10^{-1}	0.17×10^{-6}	0.17×10^{-11}	0.17×10^{-16}	0.17×10^{-21}
4	0.16×10^{-2}	0.16×10^{-7}	0.16×10^{-12}	0.16×10^{-17}	0.16×10^{-22}
5	0.26×10^{-3}	0.25×10^{-8}	0.25×10^{-13}	0.25×10^{-18}	0.25×10^{-23}
6	0.64×10^{-4}	0.64×10^{-9}	0.64×10^{-14}	0.64×10^{-19}	0.64×10^{-24}
7	0.58×10^{-4}	0.58×10^{-9}	0.58×10^{-14}	0.58×10^{-19}	0.58×10^{-24}

a) $\rho(X_1) > 1$ означает здесь, что интервал X_i содержит 0.

Следует иметь в виду, что в ней дана лишь верхняя граница $\rho(X_i)$ относительной погрешности для каждой компоненты вектора результата.

Рассмотрим следующую задачу. Пусть имеются машинные интервалы (т. е. вещественные интервалы, границами которых служат машинные числа)

$$C_0,\ A_0,\ B_0,\ D_0,\ A_1,\ B_1,\ D_1,\ \ldots,\ A_{n-1},\ B_{n-1},\ D_{n-1},$$

а также машинное число a_n. Требуется вычислить выражение

$$R_n = (1/a_n)\{C_0 - A_0(B_0 - D_0) - A_1(B_1 - D_1) - \ldots - A_{n-1}(B_{n-1} - D_{n-1})\}.$$

Теоретически можно воспользоваться таким алгоритмом:

$$\text{(S)}\quad \begin{aligned} &S_0 := C_0,\\ &S_i := S_{i-1} - A_{i-1}(B_{i-1} - D_{i-1}), \quad 1 \leqslant i \leqslant n,\\ &R_n := S_n/a_n. \end{aligned}$$

На практике, однако, этот алгоритм выполняется в виде

$$(\overline{S})\quad \begin{aligned} &\overline{S}_0 := S_0 := C_0,\\ &\overline{S}_i := fl(\overline{S}_{i-1} - fl(A_{i-1} fl(B_{i-1} - D_{i-1}))), \quad 1 \leqslant i \leqslant n,\\ &\overline{R}_n := fl(\overline{S}_n/a_n). \end{aligned}$$

Начав с (15а), установим $\mathrm{eps} := \frac{1}{2} b^{1-t}$ и для произвольных интервалов A и B получим, что

$$fl(A * B) \subseteq A * B + [-\varepsilon, \varepsilon]\, A * B. \tag{22}$$

где $\max\{|\varepsilon_1|, |\varepsilon_2|\} \leqslant \varepsilon,\ \varepsilon = 2\ \mathrm{eps}$.

Предположим на время, что

$$\overline{S}_0 = S_0 = C_0, \quad \overline{S}_1, \ \ldots, \ \overline{S}_{n-1}$$

уже вычислено. Тогда из (22) следует

$$\begin{aligned} &fl(B_{n-1} - D_{n-1}) \subseteq B_{n-1} - D_{n-1} + |B_{n-1} - D_{n-1}|[-\varepsilon, \varepsilon],\\ &fl(A_{n-1} fl(B_{n-1} - D_{n-1}))\\ &\subseteq A_{n-1}(B_{n-1} - D_{n-1} + |B_{n-1} - D_{n-1}|[-\varepsilon, \varepsilon])\\ &\quad + |A_{n-1}(B_{n-1} - D_{n-1} + |B_{n-1} - D_{n-1}|[-\varepsilon, \varepsilon])|[-\varepsilon, \varepsilon]\\ &\subseteq A_{n-1}(B_{n-1} - D_{n-1}) + |A_{n-1}||B_{n-1} - D_{n-1}|[-2\varepsilon - \varepsilon^2, 2\varepsilon + \varepsilon^2], \end{aligned}$$

а значит,

$$\begin{aligned}\bar{S}_n &\subseteq \bar{S}_{n-1} - A_{n-1}(B_{n-1} - D_{n-1}) \\ &\quad - |A_{n-1}||B_{n-1} - D_{n-1}|[-2\varepsilon - \varepsilon^2,\ 2\varepsilon + \varepsilon^2] \\ &\quad + |\bar{S}_{n-1} - A_{n-1}(B_{n-1} - D_{n-1}) \\ &\quad - |A_{n-1}||B_{n-1} - D_{n-1}|[-2\varepsilon - \varepsilon^2,\ 2\varepsilon + \varepsilon^2]|[-\varepsilon,\ \varepsilon] \\ &\subseteq \bar{S}_{n-1} - A_{n-1}(B_{n-1} - D_{n-1}) + |\bar{S}_{n-1}|[-\varepsilon,\ \varepsilon] \\ &\quad + |A_{n-1}||B_{n-1} - D_{n-1}|[-3\varepsilon - 3\varepsilon^2 - \varepsilon^3,\ 3\varepsilon + 3\varepsilon^2 + \varepsilon^3].\end{aligned} \tag{23}$$

При помощи математической индукции покажем, что

$$\begin{aligned}\bar{S}_n \subseteq S_n + [-\varepsilon,\ \varepsilon]\sum_{i=0}^{n-1}|\bar{S}_i| + [-3\varepsilon - 3\varepsilon^2 - \varepsilon^3,\ 3\varepsilon + 3\varepsilon^2 + \varepsilon^3] \\ \times \sum_{i=0}^{n-1}|A_i||B_i - D_i|.\end{aligned} \tag{24}$$

Для n=1 из (23) с учетом гого, что $\bar{S}_0 = S_0 = C_0$, получаем

$$\begin{aligned}\bar{S}_1 &\subseteq \bar{S}_0 - A_0(B_0 - D_0) + |\bar{S}_0|[-\varepsilon,\ \varepsilon] \\ &\quad + |A_0||B_0 - D_0|[-3\varepsilon - 3\varepsilon^2 - \varepsilon^3,\ 3\varepsilon + 3\varepsilon^2 + \varepsilon^3] \\ &= S_1 + [-\varepsilon,\ \varepsilon]|\bar{S}_0| + [-3\varepsilon - 3\varepsilon^2 - \varepsilon^3,\ 3\varepsilon + 3\varepsilon^2 + \varepsilon^3]|A_0||B_0 - D_0|,\end{aligned}$$

откуда видна справедливость нашего предположения при n= 1. Если для некоторого $n \geq 1$ выполнено (24), тогда замена n на n+ 1 в (23), а также использование (5) дают

$$\begin{aligned}\bar{S}_{n+1} &\subseteq \bar{S}_n - A_n(B_n - D_n) + [-\varepsilon,\ \varepsilon]|\bar{S}_n| \\ &\quad + [-3\varepsilon - 3\varepsilon^2 - \varepsilon^3,\ 3\varepsilon + 3\varepsilon^2 + \varepsilon^3]|A_n||B_n - D_n| \\ &\subseteq S_{n+1} + [-\varepsilon,\ \varepsilon]\sum_{i=0}^{n}|\bar{S}_i| + [-3\varepsilon - 3\varepsilon^2 - \varepsilon^3,\ 3\varepsilon + 3\varepsilon^2 + \varepsilon^3] \\ &\quad \times \sum_{i=0}^{n}|A_i||B_i - D_i|.\end{aligned}$$

Последнее выражение тождественно (24), у которого n заменено на n+1. Повторное применение (22) приводит к окончательному результату

$$\bar{R}_n \subseteq \bar{S}_n/a_n + (|\bar{S}_n|/|a_n|)[-\varepsilon,\ \varepsilon]. \tag{25}$$

Неравенства (24) и (25) будут использованы дальше.

Замечания. Понятие округления в том виде, как оно встречается в этом микромодуле, подробно рассмотрено Миранкером и Кулишем. Неравенство (15b), явившееся исходным пунктом при обсуждении влияния погрешносгей округления, можно найти у Валлиша и Грюцманна. Оценка (24) была доказана Алефельдом и Рокном; мы еще вернемся к ней дальше. Теорема 5 была доказана Муром.

7.4. Комплексная интервальная арифметика

Теперь определим так называемую комплексную интервальную арифметику. Будет показано, что многие из свойств и результатов, полученных для вещественной интервальной арифметики, можно перенести на случай комплексной. Чтобы это проделать, определим множества комплексных чисел, которые будут использоваться в качестве комплексных интервалов. Имеются два предпочтительных подхода, к рассмотрению которых и перейдем.

А. Прямоугольники в качестве комплексных интервалов

Определение 1. Пусть A_1 и A_2 — произвольные элементы из $I(\mathrm{R})$. Тогда множество комплексных чисел

$$A = \{a = a_1 + ia_2 \mid a_1 \in A_1,\ a_2 \in A_2\} \qquad (i = \sqrt{-1})$$

называется комплексным интервалом.

Определенные таким образом множества комплексных чисел могут быть изображены на комплексной плоскости в виде прямоугольников со сторонами, параллельными осям координат. Множество всех таких комплексных интервалов обозначим через R(C), а прописные буквы *A, B, C, ..., X, Y, Z* будем использовать для обозначения его элементов. Всякое A из R(C) можно записать в виде

$$A = A_1 + iA_2, \text{где } A_1,\ A_2 \in I(\mathbb{R}).$$

Комплексное число $a = a_1 + ia_2$ можно рассматривать как точечный комплексный интервал:

$$A = [a_1,\ a_1] + i[a_2,\ a_2] \in R(\mathbb{C}),$$

а каждый элемент A_1 из $I(\mathrm{R})$ — как сумму $A = A_1 + i[0,\ 0] \in \mathrm{R(C)}$, откуда видно, что $I(\mathbb{R}) \subset R(\mathbb{C})$.

Определение 2. Пусть $A = A_1 + iA_2$ и $B = B_1 + iB_2$ —два элемента из R(C). Тогда A и B считаются равными (запись: $A = B$), если

$$A_1 = B_1 \quad \text{и} \quad A_2 = B_2$$

см. также определение 1. п.7.1).

Определенное здесь отношение равенства рефлективно, симметрично и транзитивно.

Теперь мы обобщим арифметику комплексных чисел на случай R(C).

Определение 3. Пусть * из $\{+, -, \cdot, :\}$— бинарная операция над элементами из I(R) (как в определении 2 п.7.1). Тогда если

$$A = A_1 + iA_2, \qquad B = B_1 + iB_2 \in R(\mathbb{C}),$$

то мы полагаем

$$A \pm B = A_1 \pm B_1 + i(A_2 \pm B_2),$$
$$A \cdot B = A_1B_1 - A_2B_2 + i(A_1B_2 + A_2B_1),$$
$$A : B = (A_1B_1 + A_2B_2) : (B_1^2 + B_2^2) + i(A_2B_1 - A_1B_2) : (B_1^2 + B_2^2).$$

Считается, что в случае деления $0 \notin B_1^2 + B_2^2$. При вычислении степеней $B_1^2 = B_1B_1$ и $B_2^2 = B_2B_2$ это требование может оказаться невыполненными, даже если $0 \notin B_1 + iB_2$ в соответствии с определением 2 п.7.1. Если при этом оказывается, что $0 \in B_1^2 + B_2^2$, то деление не определено.

Чтобы проиллюстрировать сказанное, рассмотрим следующий пример.

Пример. Пусть

$$B = [-1, 1] + i[1, 3].$$

Тогда

$$0 \in [0, 10] = [-1, 1] + [1, 9] = B_1B_1 + B_2B_2.$$

Поэтому мы оговариваем, что в определении 3, если производится деление двух элементов из R(C), выражение $B_1^2 + B_2^2$ следует вычислять по правилу

$$B_1^2 + B_2^2 = \{b_1^2 \mid b_1 \in B_1\} + \{b_2^2 \mid b_2 \in B_2\}$$

(см. также определение 3 п.7.1).

Тогда в приведенном выше примере

$$B_1^2 + B_2^2 = [0, 1] + [1, 9] = [1, 10].$$

Теперь рассмотрим более внимательно свойства введенной выше комплексной интервальной арифметики.

Сразу же видно, что если $A, B \in R(\mathbb{C})$, то равенство

$$A \pm B = \{a \pm b \mid a \in A, b \in B\}$$

справедливо для сложения (соответственно вычитания) на множестве R(C). Аналогичное равенство для умножения и деления, вообще говоря, не выполняется. Это можно увидеть из следующего простого примера.

Пример. Пусть

$$A = [2,\ 4] + i[0,\ 0], \qquad B = [1,\ 1] + i[1,\ 1].$$

Из определения 3 получаем

$$AB = [2,\ 4] + i[2,\ 4].$$

С другой стороны,

$$\{ab \mid a \in A,\ b \in B\} = \{s(1+i) \mid s \in \mathbb{R},\ 2 \leqslant s \leqslant 4\} \subset AB.$$

Справедлива, однако, следующая теорема.

Теорема 4. *Операции, введенные определением* 3, *удовлетворяют соотношению*

$$\{a * b \mid a \in A,\ b \in B\} \subseteq A * B.$$

Для сложения и вычитания включение может быть заменено на равенство. Для умножения

$$AB = \inf \{X \in R(\mathbb{C}) \mid \{ab \mid a \in A,\ b \in B\} \subseteq X\},$$

где точная нижняя грань берется в смысле частичного порядка на R(C), *определяемого теоретико-множественным включением.*

Доказательство. Случай сложения и вычитания был рассмотрен ранее. Пусть теперь $a \in A$ и $b \in B$. Используя монотонность включения вещественных интервалов, для $a = a_1 + ia_2$ и $b = b_1 + ib_2$ имеем

$$\begin{aligned} ab &= a_1b_1 - a_2b_2 + i(a_1b_2 + a_2b_1) \\ &\in A_1B_1 - A_2B_2 + i(A_1B_2 + A_2B_1) = AB. \end{aligned}$$

Поскольку каждая переменная входит в выражение $a_1b_1 - a_2b_2$ лишь один раз, то получаем, что

$$\{a_1b_1 - a_2b_2 \mid a_k \in A_k,\ b_k \in B_k,\ k = 1,\ 2\} = A_1B_1 - A_2B_2.$$

По той же причине

$$\{a_1b_2 + a_2b_1 \mid a_k \in A_k,\ b_k \in B_k,\ k = 1,\ 2\} = A_1B_2 + A_2B_1.$$

Последние два соотношения показывают, что для каждого вещественного числа c_1 такого, что

$$c_1 = a_1b_1 - a_2b_2 \in A_1B_1 - A_2B_2,$$

где $a_k \in A_k,\ b_k \in B_k,\ k = 1,\ 2,$ можно найти другое вещественное число c_2 такое, что

$c_2 = a_2b_1 + a_1b_2 \in A_2B_1 + A_1B_2$, где $a_k \in A_k$, $b_k \in B_k$, $k = 1,\ 2$ и $c_1 + ic_2 \in AB$. Это и требовалось доказать.

Соотношение

$$\{a : b \mid a \in A,\ b \in B\} \subseteq A : B$$

также следует из монотонности включения.

Утверждение теоремы 4, касающееся умножения, не допускает, вообще говоря, распространения на случай деления. Тем не менее можно получить «уточнение» (в смысле сужения объемлющего интервала), если определить, что

$$A : B = A \cdot \frac{1}{B},$$

а затем вычислять $1/B$ по формуле

$$1/B = \inf \{X \in R(\mathbb{C}) \mid \{1/b \mid b \in B\} \subseteq X\}.$$

Данная возможность была предложена Рокном и Ланкастером в виде набора формул, требующих значительной вычислительной работы.

В. Круги в качестве комплексных интервалов

Определение 5. Пусть a из С — комплексное число, и пусть $r \geq 0$. Мы называем множество

$$Z = \{z \in \mathbb{C} \mid |z - a| \leqslant r\}$$

кругом или круговым интервалом (или просто комплексным интервалом, когда не опасаемся спутать его с прямоугольным интервалом).

Множество всех кругов обозначим через К(С), а прописные буквы *A, B, C, ..., X, Y,* Z используем для обозначения его элементов. Круг Z с центром a и радиусом r будем записывать в виде

$$Z = \langle a, r \rangle.$$

Комплексные числа можно рассматривать как специальные элементы из К(С), имеющие вид $\langle a, 0 \rangle$. Ясно, что $\mathrm{C} \subset \mathrm{K}(\mathrm{C})$.

Определение 6. Два круга $A = \langle a, r_1 \rangle$ и $B = \langle b, r_2 \rangle$ называются равными (обозначение: *A = B),* если они равны в теоретико-множесгвенном смысле. В этом случае $a = b$ и $r_1 = r_2$.

Это отношение равенства также рефлексивно, симметрично и транзитивно.

Операции на К(С) вводятся как обобщения операций над вещественными числами следующим образом.

Определение 7. Пусть * из $\{+, -, \cdot, /\}$ — бинарная операция над комплексными числами. Тогда если $A = \langle a, r_1 \rangle$ и $B = \langle b, r_2 \rangle$, то

$$A \pm B = \langle a \pm b,\ r_1 + r_2 \rangle,$$

$$A \cdot B = \langle ab,\ |a| r_2 + |b| r_1 + r_1 r_2 \rangle,$$

$$\frac{1}{B} = \left\langle \frac{\bar{b}}{b\bar{b} - r_2^2},\ \frac{r_2}{b\bar{b} - r_2^2} \right\rangle \quad \text{при условии, что} \quad 0 \notin B,$$

$$A : B = A \cdot \frac{1}{B} \quad \text{при условии, что} \quad 0 \notin B.$$

Здесь $|a| = \sqrt{a_1^2 + a_2^2}$ обозначает евклидову норму комплексного числа $a = a + ia_2$, а $\bar{b} = b_1 + ib_2$ — сопряженное с $b = b_1 + ib_2$.

Очевидно, что для сложения и вычитания кругов выполнено равенство

$$A \pm B = \{a \pm b \mid a \in A,\ b \in B\}.$$

То же справедливо для операции обращения круга: если мы применим теорию конформных отображений к отображению вида $w = 1/z$ для не содержащего нуля круга, то получим другой круг. Иными словами,

$$1/B = \{1/b \mid b \in B\}.$$

Элементарными преобразованиями легко проверить формулы определения 7 для центра и радиуса области $\{1/b \mid b \in B\}$.

Для умножения (а следовательно, и для деления) двух элементов из K (C) по правилам определения 7 верно, вообще говоря, только то, что

$$\{z_1 z_2 \mid z_1 \in A,\ z_2 \in B\} \subseteq AB.$$

Это вытекает из следующих неравенств:

$$\begin{aligned} |z_1 z_2 - ab| &= |a(z_2 - b) + b(z_1 - a) + (z_1 - a)(z_2 - b)| \\ &\leqslant |a||z_2 - b| + |b||z_1 - a| + |z_1 - a||z_2 - b| \\ &\leqslant |a| r_2 + |b| r_1 + r_1 r_2. \end{aligned}$$

Аналогично теореме 4 п.7.1, соберем теперь вместе наиболее важные свойства операций на R(C) и K (C). Если не оговорено противное, то I(C) можно понимать и как обозначение множества R(C) с операциями из определения 3, и как обозначение множества K (C) с операциями из определения 7.

Теорема 8. *Пусть* $A,\ B,\ C \in I(\mathbb{C})$. *Тогда*

$$A + B = B + A, \quad AB = BA \quad (\textit{коммутативность}); \tag{1}$$

$$(A + B) + C = A + (B + C), \tag{2}$$

$$(AB)C = A(BC) \quad \textit{для } A,\ B,\ C \textit{ из } K(\mathbb{C}) \ (\textit{ассоциативность});$$

$$[0,\ 0]+i[0,\ 0]\in R(\mathbb{C})\ (\textit{соответственно}\ \langle 0,\ 0\rangle\in K(\mathbb{C}))\ \textit{и}\qquad (3)$$
$$[1,\ 1]+i[0,\ 0]\in R(\mathbb{C})\ (\textit{соответственно}\ \langle 1,\ 0\rangle\in K(\mathbb{C}))$$

— определенные единственным образом нейтральные элементы сложения (нуль) и умножения (единица);

$I(\mathbb{C})$ *не имеет делителей нуля;* (4)

элемент Z *множества* $I(\mathbb{C})$ *имеет противоположный и обратный элементы, если и только если* $Z\in C$ u (5)

в случае мультипликативных операций $Z\neq 0$.

Однако $0\in A-A$ *и* $1\in A:A$;

$$A(B+C)\subseteq AB+AC\quad (\textit{субдистрибутивность}),\qquad (6)$$
$$a(B+C)=aB+aC\quad \textit{для}\quad a\in\mathbb{C}.$$

Доказательство. Доказательство этих утверждений следует из определений операций 3 и 7. В качестве примера докажем (6) для K(C). Если $A=\langle a,\ r_1\rangle$, $B=\langle b,\ r_2\rangle$, $C=\langle c,\ r_3\rangle\in K(\mathbb{C})$, то

$$\begin{aligned}A(B+C)&=\langle a,\ r_1\rangle\langle b+c,\ r_2+r_3\rangle\\&=\langle a(b+c),\ |a|(r_2+r_3)+|b+c|r_1+r_1(r_2+r_3)\rangle\\&\subseteq\langle ab+ac,\ |a|r_2+|a|r_3+|b|r_1+|c|r_1+r_1r_2+r_1r_3\rangle\\&=\langle ab,\ |a|r_2+|b|r_1+r_1r_2\rangle+\langle ac,\ |a|r_3+|c|r_1+r_1r_3\rangle\\&=AB+AC.\end{aligned}$$

В случае $A=\langle a,\ 0\rangle$, т. е. равенства нулю r_1, доказательство приводит к соотношению

$$a(B+C)=aB+aC.$$

Необходимо особо отметить, что ассоциативный закон (2) в общем случае не выполняется при перемножении элементов R(C). Это можно увидеть из следующего примера. Пример.

$$A=[2,\ 4]+i[0,\ 0],\quad B=[1,\ 1]+i[1,\ 1],\quad C=[1,\ 1]+i[1,\ 1],$$
$$(AB)C=([2,\ 4]+i[2,\ 4])([1,\ 1]+i[1,\ 1])=[-2,\ 2]+i[4,\ 8],$$
$$A(BC)=([2,\ 4]+i[0,\ 0])([0,\ 0]+i[2,\ 2])=[0,\ 0]+i[4,\ 8].$$

$I(\mathbb{C})$ обладает также монотонностью включения.

Теорема 9. *Пусть* $A^{(k)},\ B^{(k)}\in I(\mathbb{C}),\ k=1,\ 2,$ *таковы, что*

$$A^{(k)}\subseteq B^{(k)},\quad k=1,\ 2.$$

Тогда соотношение

$$A^{(1)}*A^{(2)}\subseteq B^{(1)}*B^{(2)}$$

выполняется для операций * *из* $\{+,\ -,\ \cdot,\ :\}$.

Доказательство. Это верно для $R(\mathbb{C})$, поскольку монотонность включения выполняется для элементов $I(\mathbb{R})$ (см. теорему 5 п.7.1).
В случае сложения и вычитания элементов $K(\mathbb{C})$ имеем

$$A^{(1)} \pm A^{(2)} = \{z = x \pm y \mid x \in A^{(1)},\ y \in A^{(2)}\}$$
$$\subseteq \{w = u \pm v \mid u \in B^{(1)},\ v \in B^{(2)}\} = B^{(1)} \pm B^{(2)}.$$

Рассмотрим теперь умножение на множестве $K(\mathbb{C})$. Пусть

$$A^{(k)} = \langle a^{(k)},\ r^{(k)} \rangle, \qquad B^{(k)} = \langle b^{(k)},\ s^{(k)} \rangle, \qquad k = 1,\ 2.$$

Предположение $A^{(k)} \subseteq B^{(k)}$, $k = 1,\ 2$, эквивалентно тому, что

$$|a^{(k)} - b^{(k)}| \leqslant s^{(k)} - r^{(k)}, \qquad k = 1,\ 2.$$

Далее,

$$A^{(1)}A^{(2)} = \langle a^{(1)}a^{(2)},\ |a^{(1)}|r^{(2)} + |a^{(2)}|r^{(1)} + r^{(1)}r^{(2)} \rangle,$$
$$B^{(1)}B^{(2)} = \langle b^{(1)}b^{(2)},\ |b^{(1)}|s^{(2)} + |b^{(2)}|s^{(1)} + s^{(1)}s^{(2)} \rangle.$$

Требуется доказать, что

$$|a^{(1)}a^{(2)} - b^{(1)}b^{(2)}| \leqslant |b^{(1)}|s^{(2)} + |b^{(2)}|s^{(1)} + s^{(1)}s^{(2)}$$
$$-(|a^{(1)}|r^{(2)} + |a^{(2)}|r_1 + r^{(1)}r^{(2)}).$$

Из неравенства треугольника получаем

$$-|b^{(2)}| \leqslant -|a^{(2)}| + |a^{(2)} - b^{(2)}|,$$
$$-|b^{(1)}| \leqslant -|a^{(1)}| + |a^{(1)} - b^{(1)}|,$$

а поскольку

$$|a^{(k)} - b^{(k)}| \leqslant s^{(k)} - r^{(k)}, \qquad k = 1,\ 2,$$

имеем

$$-|b^{(2)}|r^{(1)} \leqslant -|a^{(2)}|r^{(1)} + r^{(1)}(s^{(2)} - r^{(2)})$$
$$= -|a^{(2)}|r^{(1)} + r^{(1)}s^{(2)} - r^{(1)}r^{(2)},$$
$$-|b^{(1)}|r^{(2)} \leqslant -|a^{(1)}|r^{(2)} + r^{(2)}(s^{(1)} - r^{(1)})$$
$$= -|a^{(1)}|r^{(2)} + r^{(2)}s^{(1)} - r^{(1)}r^{(2)}.$$

Отсюда

$$|a^{(1)}a^{(2)} - b^{(1)}b^{(2)}| \leqslant |b^{(2)}||a^{(1)} - b^{(1)}| + |b^{(1)}||a^{(2)} - b^{(2)}|$$
$$+ |a^{(1)} - b^{(1)}||a^{(2)} - b^{(2)}|$$
$$\leqslant |b^{(2)}|(s^{(1)} - r^{(1)}) + |b^{(1)}|(s^{(2)} - r^{(2)})$$
$$+ (s^{(1)} - r^{(1)})(s^{(2)} - r^{(2)})$$
$$\leqslant |b^{(2)}|s^{(1)} + |b^{(1)}|s^{(2)} + s^{(1)}s^{(2)}$$
$$-(|a^{(2)}|r^{(1)} + |a^{(1)}|r^{(2)} + r^{(1)}r^{(2)}),$$

что доказывает теорему для случая умножения. Из того, что

$$1/A^{(2)} = \{z = 1/x \mid x \in A^{(2)}\} \subseteq \{w = 1/u \mid u \in B^{(2)}\} = 1/B^{(2)},$$

следует, что

$$A^{(1)} : A^{(2)} = A^{(1)} \cdot \frac{1}{A^{(2)}} \subseteq B^{(1)} \cdot \frac{1}{B^{(2)}} = B^{(1)} : B^{(2)}.$$

Теорема доказана.

Частным случаем теоремы 9 является

Следствие 10. *Пусть* $A,\ B \in I(\mathbb{C})$ *и* $a \in A,\ b \in B$. *Тогда*

$$a * b \in A * B,$$

где $* \in \{+,\ -,\ \cdot,\ :\}$.

Замечания. Круги в качестве комплексных интервалов впервые ввели в систематическое употребление Гаргантиии и Энричи. Рассмотренная в этом микромодуле арифметика кругов была предложена ими и была использована для одновременной локализации корней многочлена. Другие приложения арифметики кругов будут изложены дальше. Важное свойство монотонности включения в том виде, как оно сформулировано в теореме 9, впервые доказано Г. Алефельдом и Ю. Херцбергом. Как уже было отмечено, решена проблема такого определения умножения кругов, которое приводит к получению меньших множеств. Арифметика на *R*(C), сводящаяся при выполнении некоторых условий к вещественной, предложена Алефельдом. Свойства этой арифметики исследовали также Бош, Рокн и Ланкастер.

Как уже неоднократно указывалось, в основе почти всех приложений интервальной арифметики лежит монотонность включения (теорема 9). Умножение, предложенное Криером, не обладает этим свойством, что было показано там же на примере.

Приближенная реализация арифметики прямоугольников на цифровой вычислительной машине не вызывает проблем, поскольку операции на *R*(C) сводятся к операциям на *I*(R). Ранее было показано, как сделать возможной реализацию приближенных операций над элементами *I*(R) на ЦВМ без потери наиболее важных свойств арифметики, что позволяет проделать то же самое для *R*(C).

7.5. Метрика, абсолютная величина и ширина в *I*(C)

В этом микромодуле *q* будет обозначать метрику на *I*(R), задаваемую определением 1 п.7.2. Введем теперь метрику на *R*(C).

Определение 1. Пусть $A = A_1 + iA_2$ и $B = B_1 + iB_2$ принадлежат $R(C)$. Тогда расстояние p между A и B определяется формулой

$$p(A, B) = q(A_1, B_1) + q(A_2, B_2).$$

Если p используется в пространстве $I(\mathrm{R})$, то оно принимает те же самые значения, что и q из определения 1 п.7.2. Поэтому в дальнейшем расстояние на $R(\mathrm{C})$ будет обозначаться через q, так что

$$q(A, B) = q(A_1, B_1) + q(A_2, B_2).$$

Поскольку q является метрикой на $I(\mathrm{R})$, легко доказать, что q останется ею и при переходе к $R(\mathrm{C})$. Введение на $R(\mathrm{C})$ метрики q делает его топологическим пространством. Если теперь обычным для метрических пространств способом определить сходимость, то окажется, что последовательность $\{A^{(k)}\}_{k=0}^{\infty}$, где $A^{(k)} = A_1^{(k)} + iA_2^{(k)} \in R(\mathbb{C})$, сходится к $A = A_1 + iA_2$ — элементу $R(\mathrm{C})$ тогда и только тогда, когда

$$\lim_{k \to \infty} A_1^{(k)} = A_1 \quad \text{и} \quad \lim_{k \to \infty} A_2^{(k)} = A_2. \tag{1}$$

Из факта замкнутости метрического пространства $(I(\mathbb{R}), q)$ вытекает, что $R(\mathrm{C})$ с заданной на нем метрикой q также образует замкнутое метрическое пространство.

Определение 2. Пусть $A = A_1 + iA_2 \in R(\mathbb{C})$. Тогда

$$|A| = q(A, 0) = |A_1| + |A_2| = q(A_1, 0) + q(A_2, 0)$$

называется абсолютной величиной A.

Если, в частности, $A = [a_1, a_1] + i[a_2, a_2] = a_1 + ia_2 = a$, то

$$|A| = |a| = |a_1| + |a_2|. \tag{2}$$

Итак, абсолютная величина из определения 2 не совпадает с евклидовой нормой комплексного числа. В дальнейшем из контексга всегда будет ясно, какая из них используется в конкретном случае. Кроме того, заметим, что из (2) следует справедливость соотношения

$$|A| = \max_{a \in A} |a|.$$

Пусть d в соответствии с определением 8 п.7.2 служит обозначением ширины вещественного интервала. Тогда имеем

Определение 3. Если $A = A_1 + iA_2 \in R(\mathbb{C})$, то шириной A будем называть

$$d(A) = d(A_1) + d(A_2).$$

Теперь перейдем к множеству $K(\mathrm{C})$.

Определение 4. Пусть

$$A = \langle a,\ r_1\rangle,\ B = \langle b,\ r_2\rangle \in K(\mathbb{C}).$$

Тогда назовем

(a) $q(A, B) = |a - b| + |r_1 - r_2|$ расстоянием между A и B,
(b) $|A| = |a| + r_1$ абсолютной величиной A и
(c) $d(A) = 2r_1$ шириной A.

Для того чтобы определить расстояние между двумя круговыми интервалами на комплексной плоскости, в определении 4 используется евклидова метрика. Когда абсолютная величина кругового интервала применяется к обычным комплексным числам, она совпадает с евклидовой нормой. Обратим внимание, что и на этот раз выполняется соотношение

$$|A| = \max_{a \in A} |a|.$$

Если сходимость последовательности, состоящей из элементов $K(\mathbb{C})$, определена, как обычно, с помощью соответствующей метрики, то можно легко убедиться в замкнутости метрического пространства $(K(\mathbb{C}), q)$. Опираясь на такого рода определение, получаем, что

$$\lim_{k\to\infty} A^{(k)} = A \Leftrightarrow \lim_{k\to\infty} a^{(k)} = a, \quad \lim_{k\to\infty} r^{(k)} = r, \tag{3}$$

где $\{A^{(k)}\}_{k=0}^{\infty} = \{\langle a^{(k)},\ r^{(k)}\rangle\}_{k=0}^{\infty},\ A = \langle a,\ r\rangle$.

В следующей теореме собраны наиболее важные свойства метрики, абсолютной величины и ширины на множествах $R(\mathbb{C})$ и $K(\mathbb{C})$.

Теорема 5. *Пусть имеются A, B, C, D из* $I(\mathbb{R})$. *Тогда*

$$q(A + B,\ A + C) = q(B,\ C), \tag{4}$$

$$q(A + B,\ C + D) \leqslant q(A,\ C) + q(B.\ D), \tag{5}$$

$$q(aB,\ aC) \leqslant |a|\, q(B,\ C), \quad a \in \mathbb{C}. \tag{6}$$

Если B и C принадлежат $K(\mathbb{C})$, *то в* (6) *всегда имеет место равенство.*

$$q(AB,\ AC) \leqslant |A|\, q(B,\ C), \tag{7}$$

$$|A| \geqslant 0, \quad |A| = 0 \Leftrightarrow A = 0, \tag{8}$$

$$|A + B| \leqslant |A| + |B|, \tag{9}$$

$$|aB| \leqslant |a|\,|B|, \quad a \in \mathbb{C}. \tag{10}$$

Если $B \in K(\mathbb{C})$, *то* (10) *превращается в равенство.*

$$|AB| \leqslant |A|\,|B|, \tag{11}$$

$$d(aB) = |a|\, d(B), \quad a \in \mathbb{C}. \tag{12}$$

$$d(AB) \leqslant |A|\, d(B) + |B|\, d(A). \tag{13}$$

$$d(A) = |A - A|, \tag{14}$$

$$d(AB) \geqslant |A|\, d(B), \tag{15}$$

$$d(A \pm B) = d(A) + d(B), \tag{16}$$

$$A \subseteq B \Rightarrow \frac{1}{2}(d(B) - d(A)) \leqslant q(A, B) \leqslant d(B) - d(A). \tag{17}$$

Доказательство. Сначала докажем перечисленные свойства из $R(\mathbb{C})$. Справедливость (4) — (7) следует непосредственно из соответствующих свойств (4 п.7.2) — (7 п.7.2), сформулированных в теореме 7 п.7.2 применительно к вещественным интервалам.

Пусть

$$A = A_1 + iA_2, \quad B = B_1 + iB_2,$$
$$C = C_1 + iC_2, \quad D = D_1 + iD_2 \in R(\mathbb{C}).$$

(4): $r(A + B, A + C) = q(A_1 + B_1 + i(A_2 + B_2), A_1 + C_1 + i(A_2 + C_2))$
$= q(A_1 + B_1, A_1 + C_1) + q(A_2 + B_2, A_2 + C_2)$
$= q(B_1, C_1) + q(B_2, C_2) = q(B, C).$

(5): $q(A + B, C + D) = q(A_1 + B_1, C_1 + D_1) + q(A_2 + B_2, C_2 + D_2)$
$\leqslant q(A_1, C_1) + q(B_1, D_1) + q(A_2, C_2) + q(B_2, D_2)$
$= q(A, C) + q(B, D).$

(6), (7):

$q(AB, AC) = q(A_1B_1 - A_2B_2, A_1C_1 - A_2C_2)$
$+ q(A_1B_2 + A_2B_1, A_1C_2 + A_2C_1)$
$\leqslant |A_1|\, q(B_1, C_1) + |A_2|\, q(B_2, C_2) + |A_1|\, q(B_2, C_2) + |A_2|\, q(B_1, C_1)$
$= (|A_1| + |A_2|)\, q(B, C) = |A|\, q(B, C).$

Свойства (8)—(11) доказываются на основании определения $|A|$.

(8): $|A| = q(A, 0) = q(A_1, 0) + q(A_2, 0) = |A_1| + |A_2| \geqslant 0,$
$|A| = 0 \Leftrightarrow |A_1| = |A_2| = 0 \Leftrightarrow A = 0.$

(9). $|A + B| = q(A + B, 0) \leqslant q(A, 0) + q(B, 0) = |A| + |B|$
(в соответствии с 5)).

(10), (11):

$$|AB| = q(AB, 0) = q(AB, A \cdot 0) \leqslant |A|\, q(B, 0) = |A|\,|B|$$

(из 6) и (7)).

(12): Пусть $a = a_1 + ia_2 \in \mathbb{C}$. Из определения 5.3 п. 7.4 следует

$$aB = a_1B_1 - a_2B_2 + i(a_1B_2 + a_2B_1),$$

и с помощью (2) получаем

$$d(aB) = d(a_1B_1 - a_2B_2) + d(a_1B_2 + a_2B_1)$$
$$= d(a_1B_1) + d(a_2B_2) + d(a_1B_2) + d(a_2B_1)$$
$$= |a_1|d(B_1) + |a_2|d(B_2) + |a_1|d(B_2) + |a_2|d(B_1)$$
$$= (|a_1| + |a_2|)(d(B_1) + d(B_2)) = |a|d(B).$$

(13): $d(AB) = d(A_1B_1 - A_2B_2) + d(A_1B_2 + A_2B_1)$
$$= d(A_1B_1) + d(A_2B_2) + d(A_1B_2) + d(A_2B_1)$$
$$\leqslant |A_1|d(B_1) + |B_1|d(A_1) + |A_2|d(B_2) + |B_2|d(A_2)$$
$$+ |A_1|d(B_2) + |B_2|d(A_1) + |A_2|d(B_1) + |B_1|d(A_2)$$
$$= (|A_1| + |A_2|)(d(B_1) + d(B_2))$$
$$+ (|B_1| + |B_2|)(d(A_1) + d(A_2))$$
$$= |A|d(B) + |B|d(A).$$

(14): $d(A) = d(A_1) + d(A_2) = |A_1 - A_1| + |A_2 - A_2| = |A - A|.$

(15): $d(AB) = d(A_1B_1 - A_2B_2) + d(A_1B_2 + A_2B_1)$
$$\geqslant |A_1|d(B_1) + |A_2|d(B_2) + |A_1|d(B_2) + |A_2|d(B_1)$$
$$= (|A_1| + |A_2|)(d(B_1) + d(B_2)) = |A|d(B).$$

(16): $d(A \pm B) = d(A_1 \pm B_1) + d(A_2 \pm B_2)$
$$= d(A_1) + d(A_2) + d(B_1) + d(B_2)$$
$$= d(A) + d(B).$$

(17): Это свойство является прямым следствием (21 п.7.2.). Теперь пусть

$$A = \langle a, r_1\rangle, \quad B = \langle b, r_2\rangle,$$
$$C = \langle c, r_3\rangle, \quad D = \langle d, r_4\rangle \in K(\mathbb{C}).$$

(4): $q(A + B, A + C) = |a + b - (a + c)| + |r_1 + r_2 - (r_1 + r_3)|$
$$= |b - c| + |r_2 - r_3| = q(B, C).$$

(5): $q(A + B, C + D) = |a + b - (c + d)| + |r_1 + r_2 - (r_3 + r_4)|$
$$\leqslant |a - c| + |r_1 - r_3| + |b - d| + |r_2 - r_4|$$
$$= q(A, C) + q(B, D).$$

(6): $q(aB, aC) = |ab - ac| + ||a|r_2 - |a|r_3|$
$$= |a|\{|b - c| + |r_2 - r_3|\} = |a|q(B, C).$$

(7): $q(AB, AC) = |ab - ac| + ||a|r_2 + |b|r_1 + r_1r_2$
$$- (|a|r_3 + |c|r_1 + r_1r_3)|$$
$$\leqslant |a||b - c| + |a||r_2 - r_3| + r_1||b| - |c|| + r_1|r_2 - r_3|$$
$$\leqslant (|a| + r_1)(|b - c| + |r_2 - r_3|) = |A|q(B, C).$$

(8): $|A| = |a| + r_1 \geqslant 0, \quad |A| = 0 \Leftrightarrow (a = 0, r_1 = 0).$

(9): $|A+B|=|a+b|+|r_1+r_2|\leqslant|a|+r_1+|b|+r_2=|A|+|B|$.

(10): $$|aB|=|ab|+|a|r_2=|a||B|.$$

(11): $|AB|=q(AB,\ 0)=q(AB,\ A\cdot 0)\leqslant|A|q(B,\ 0)=|A||B|$ (из (7)).

(12): $$d(aB)=2|a|r_2=|a|d(B).$$

(13): $$\begin{aligned} d(AB)&=2\{|a|r_2+|b|r_1+r_1r_2\}=2\{(|a|+r_1)r_2+|b|r_1\}\\ &\leqslant 2\{(|a|+r_1)r_2+(|b|+r_2)r_1\}=|A|d(B)+|B|d(A).\end{aligned}$$

(14): $$d(A)=2r_1=|\langle 0,\ 2r_1\rangle|=|A-A|.$$

(15): $$\begin{aligned} d(AB)&=2\{|a|r_2+|b|r_1+r_1r_2\}\\ &=2\{(|a|+r_1)r_2+|b|r_1\}\\ &\geqslant 2(|a|+r_1)r_2=|A|d(B).\end{aligned}$$

(16): $d(A\pm B)=d(\langle a\pm b,\ r_1+r_2\rangle)=2(r_1+r_2)=d(A)+d(B)$.

(17): $A\subseteq B\Leftrightarrow|a-b|\leqslant r_2-r_1$. Следовательно,

$$\begin{aligned}\frac{1}{2}(d(B)-d(A))&=|r_2|-|r_1|\leqslant|r_2-r_1|\leqslant|a-b|+|r_1-r_2|\\ &=q(A,\ B)\leqslant r_2-r_1+|r_2-r_1|=d(B)-d(A).\end{aligned}$$

С теоремой 4 п.7.1 связана

Теорема 6. *Операции* $\{+,\ -,\ \cdot,\ :\}$, *заданные на* $R(\mathbb{C})$ *определением* 3 п.7.4, *а на* $K(\mathbb{C})$ *— определением* 7 п.7.4, *непрерывны.*

Доказательство. Пусть $\{A^{(k)}\}_{k=0}^{\infty}$, $\{B^{(k)}\}_{k=0}^{\infty}$ — последовательности, у которых

$$A^{(k)}=A_1^{(k)}+iA_2^{(k)},\qquad B^{(k)}=B_1^{(k)}+iB_2^{(k)}\in R(\mathbb{C})$$

и

$$\lim_{k\to\infty}A^{(k)}=A=A_1+iA_2,\qquad \lim_{k\to\infty}B^{(k)}=B=B_1+iB_2.$$

Докажем непрерывность умножения:

$$\begin{aligned}\lim_{k\to\infty}A^{(k)}B^{(k)}&=\lim_{k\to\infty}\{A_1^{(k)}B_1^{(k)}-A_2^{(k)}B_2^{(k)}+i(A_1^{(k)}B_2^{(k)}+A_2^{(k)}B_1^{(k)})\}\\ &=\lim_{k\to\infty}(A_1^{(k)}B_1^{(k)}-A_2^{(k)}B_2^{(k)})+i\lim_{k\to\infty}(A_1^{(k)}B_2^{(k)}+A_2^{(k)}B_1^{(k)})\\ &=A_1B_1-A_2B_2+i(A_1B_2+A_2B_1)=AB,\end{aligned}$$

поскольку операции отделения вещественной и мнимой частей комплексного числа непрерывны на $I(\mathbb{C})$.

Подобное доказательство можно провести и для остальных операций на $R(\mathbb{C})$ и $K(\mathbb{C})$.

Аналогично (22 п.7.1) вводится еще одна бинарная операция на $R(\mathbb{C})$. Теоретико-множественное пересечение F и B из $R(\mathbb{C})$, задаваемое формулой

$$A \cap B = \{c \mid c \in A,\ c \in B\}, \qquad (18)$$

называется пересечением A и B. Это пересечение принадлежит $R(\mathbb{C})$, если оно не пусто. При $A = A_1 + iA_2$ и $B = B_1 + iB_2$ имеем

$$A \cap B = A_1 \cap B_1 + i(A_2 \cap B_2), \qquad (19)$$

где $A_i \cap B_i$ находится по формуле (23 п.7.1).

Со следствием 12 п.7.1 связано

Следствие 7. *Пусть* $A,\ B,\ C,\ D \in R(\mathbb{C})$. *Тогда имеем*

$A \subseteq C,\ B \subseteq D \Rightarrow A \cap B \subseteq C \cap D$ *(монотонность включения).* (20)

При этом операция пересечения непрерывна, если ее результат остается в R(C).

Данное следствие можно доказать с помощью следствия 12 п.7.1, если последовательно применить последнее к вещественной и мнимой частям.

Модуль 8
Методы локализации

Микромодуль 26
Локализация нулей функций одной вещественной переменной

В этом микромодуле мы рассмотрим методы локализации нулей вещественной функции f одной вещественной переменной x. Эти методы позволят найти множество интервалов наименьшей возможной ширины, таких что каждый интервал содержит один или несколько нулей функции f из заданного интервала $X^{(0)} \in I(\mathrm{R})$. Особый интерес представляет случай одного изолированного нуля в $X^{(0)}$. При разработке таких методов будет обращено особое внимание на два обстоятельства. С одной стороны, методы должны быть применимы к широким классам функций при легко проверяемых условиях. С другой стороны, должна быть гарантирована локализация нулей и в том случае, когда рассматриваемые методы реализуются на вычислительной машине, где вместо обычной интервальной арифметики возникает машинная интервальная арифметика,

описанная ранее. Поэтому такие методы радикально отличаются от методов для конкретных классов функций и от других процедур общего назначения.

Простые реализации таких методов задаются с помощью так называемых методов деления. Это — интервальные варианты метода двоичного поиска или других методов поиска. Мы кратко опишем такую процедуру. Для нее требуется лишь существование интервального вычисления функции f в интервале $X^{(0)}$. Чтобы улучшить локализацию нулей в $X^{(0)}$\ мы делим $X^{(0)}$ пополам точкой

$$m(X^{(0)}) = \frac{1}{2}\left(x_1^{(0)} + x_2^{(0)}\right)$$

на интервалы $U^{(0)}$ и $V^{(0)}$, такие что

$$X^{(0)} = U^{(0)} \cup V^{(0)} = \left[x_1^{(0)},\ m(X^{(0)})\right] \cup \left[m(X^{(0)}),\ x_2^{(0)}\right].$$

Если $0 \in f(U^{(0)})$, то $U^{(0)}$ может содержать нуль функции f, и потому мы повторяем ту же процедуру для $U^{(0)}$. Если $0 \in f(V^{(0)})$, то мы аналогичным образом повторяем процедуру для $V^{(0)}$. Если же мы имеем $0 \notin f(U^{(0)})$ или $0 \notin f(V^{(0)})$, то игнорируем соответствующий подынтервал, так как ввиду (1 микромодуля 24) он не может содержать нуля функции f. Поэтому такой подынтервал исключается из дальнейших вычислений. Описанный итерационный процесс порождает последовательность подынтервалов, содержащихся в $X^{(0)}$ и «подозрительных на наличие нуля функции f». Ширина этих интервалов стремится к нулю, так как она уменьшается вдвое на каждом шаге. Ввиду (5 микромодуля 24) эти постепенно вычисляемые интервалы сходятся к нулям функции f в интервале $X^{(0)}$.

Чтобы предотвратить слишком сильный рост количества «подозрительных» интервалов, мы можем ввести следующую модификацию. На каждом шаге мы исследуем либо только правую половину интервала, либо только левую. Если на некотором шаге мы имеем $0 \notin f(Y)$ для этого полуинтервала Y, то процедура повторяется снова, начиная с интервала $\left[x_1^{(0)},\ y_1\right] \subset X^{(0)}$ (соответственно $\left[y_2,\ x_2^{(0)}\right] \subset X^{(0)}$). Таким образом, мы последовательно вычисляем отдельные нули функции f в порядке справа налево (соответственно слева направо) и не сталкиваемся с проблемой хранения большого количества «подозрительных» интервалов.

8.1. А. Методы ньютоновского типа

В этом и следующем разделах мы исследуем интервальные модификации метода Ньютона. Для этого рассмотрим непрерывную функцию f, имеющуюнуль в данном интервале $X^{(0)} = [x_1^{(0)}, x_2^{(0)}]$:

$$f(\xi) = 0$$

для некоторого $\xi \in X^{(0)}$. Пусть

$$f(x_1^{(0)}) < 0 \quad \text{и} \quad f(x_2^{(0)}) > 0 \tag{1}$$

в граничных точках интервала $X^{(0)}$. Пусть, далее, m_1, m_2 — границы разностных отношений

$$0 < m_1 \leqslant \frac{f(x) - f(\xi)}{x - \xi} = \frac{f(x)}{x - \xi} \leqslant m_2 < \infty, \quad \xi \neq x \in X^{(0)}. \tag{2}$$

Эти границы определяют интервал $M = [m_1, m_2] \in I(\mathbb{R})$. (Аналогичные соображения справедливы, если предположить, что $f(x_1^{(0)}) > 0$, $f(x_2^{(0)}) < 0$ и $m_2 < 0$.) Очевидно, что при сделанных предположениях функция f не имеет других корней в $X^{(0)}$. Начав с исходного локализующего интервала $X^{(0)} \ni \xi$, мы вычисляем итерационно новые интервалы $X^{(k)}$, $k \geqslant 1$, согласно следующей процедуре:

$$X^{(k+1)} = \{m(X^{(k)}) - f(m(X^{(k)}))/M\} \cap X^{(k)}, \quad k \geqslant 0, \tag{3}$$

где

$$m(X^{(k)}) \in X^{(k)}.$$

Этот шаг изображен на рис. 1.

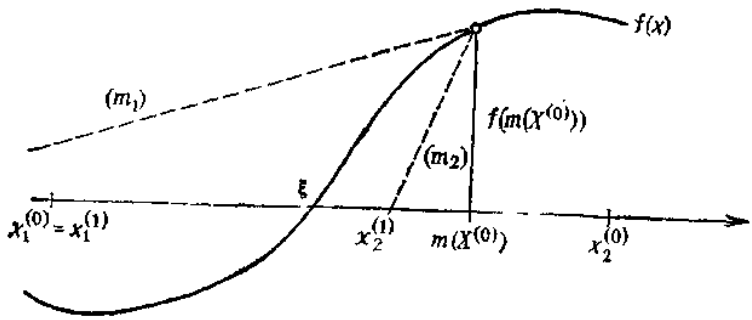


Рис. 1

Не пользуясь интервальными операциями, можно записать итерации (3) в виде

$$x_1^{(k+1)} = \begin{cases} \max\{x_1^{(k)}, m(X^{(k)}) - f(m(X^{(k)}))/m_1\}, & \text{если } f(m(X^{(k)})) \geqslant 0, \\ m(X^{(k)}) - f(m(X^{(k)}))/m_2, & \text{если } f(m(X^{(k)})) \leqslant 0, \end{cases}$$

$$x_2^{(k+1)} = \begin{cases} m(X^{(k)}) - f(m(X^{(k)}))/m_2, & \text{если } f(m(X^{(k)})) \geqslant 0, \\ \min\{x_2^{(k)}, m(X^{(k)}) - f(m(X^{(k)}))/m_1\}, & \text{если } f(m(X^{(k)})) \leqslant 0. \end{cases} \quad (3')$$

В обеих формулировках (3) и (3')

$$m: I(\mathbb{R}) \to \mathbb{R}$$

обозначает некоторую процедуру выбора вещественного числа m из данного интервала. Часто используют его середину

$$m(X) = \frac{1}{2}(x_1 + x_2). \quad (4)$$

Теперь мы перечислим важнейшие свойства последовательности итераций $\{X^{(k)}\}_{k=0}^{\infty}$.

Теорема 1. *Пусть f — непрерывная функция, ξ — нуль функции f в интервале $X^{(0)}$, причем имеют место* (1) *и* (2) *для интервала* $M = [m_1, m_2]$, $m_1 > 0$. *Тогда последовательность* $\{X^{(k)}\}_{k=0}^{\infty}$, *вычисленная по формулам* (3), *обладает следующими свойствами:*

$$\xi \in X^{(k)}, \quad k \geqslant 0, \quad (5)$$

$$X^{(0)} \supset X^{(1)} \supset X^{(2)} \supset \ldots, \quad \text{где} \quad \lim_{k\to\infty} X^{(k)} = \xi, \quad (6)$$

либо эта последовательность стабилизируется через конечное число шагов на точке $[\xi, \xi]$;

$$d(X^{(k+1)}) \leqslant (1 - m_1/m_2)\, d(X^{(k)}). \quad (7)$$

Доказательство. (5): Из (2) и следствия 1.п.7.5 получаем

$$\xi = m(X^{(0)}) - \frac{f(m(X^{(0)}))}{f(m(X^{(0)}))/(m(x^{(0)}) - \xi)} \in \{m(X^{(0)}) - f(m(X^{(0)}))/M\} \cap X^{(0)} = X^{(1)}.$$

Для $k > 1$ применяем метод математической индукции.

(6), (7): Предположим, что $f(m(X^{(k)})) > 0$. Теперь если имеет место $f(m(X^{(k)})) \geqslant (m(X^{(k)}) - x_1^{(k)})\, m_1$, то с помощью (3') мы получаем

$$\begin{aligned} d(X^{(k+1)}) &= x_2^{(k+1)} - x_1^{(k+1)} = m(X^{(k)}) - f(m(X^{(k)}))/m_2 - x_1^{(k)} \\ &\leqslant (m(X^{(k)}) - x_1^{(k)}) - (m(X^{(k)}) - x_1^{(k)})\, m_1/m_2 \\ &= (m(X^{(k)}) - x_1^{(k)})(1 - m_1/m_2) \leqslant d(X^{(k)})(1 - m_1/m_2). \end{aligned}$$

Если же имеет место $f(m(X^{(k)})) \leqslant (m(X^{(k)}) - x_1^{(k)})\, m_1$, то из (3′) получаем

$$\begin{aligned} d(X^{(k+1)}) &= x_2^{(k+1)} - x_1^{(k+1)} \\ &= m(X^{(k)} - f(m(X^{(k)}))/m_2 - m(X^{(k)}) + f(m(X^{(k)}))/m_1 \\ &= \frac{f(m(X^{(k)}))}{m_1}(1 - m_1/m_2) \leqslant (m(X^{(k)}) - x_1^{(k)})(1 - m_1/m_2) \\ &\leqslant d(X^{(k)})(1 - m_1/m_2). \end{aligned}$$

Случай $f(m(X^{(k)})) < 0$ доказывается аналогично. Если, однако, $f(m(X^{(k)})) = 0$, то $m(X^{(k)}) = \xi$, и потому $d(X^{(k+1)}) = 0$ и $X^{(k+1)} = \xi$, $i \geqslant 1$. Это доказывает (7). Ввиду $m_1 \leq m_2$

$$d(X^{(k+1)}) \leqslant \gamma^{k+1} d(X^{(0)}), \qquad 0 \leqslant \gamma = (1 - m_1/m_2) < 1,$$

откуда следует

$$\lim_{k \to \infty} d(X^{(k+1)}) = 0.$$

Отсюда и из (5) следует $\lim\limits_{k \to \infty} X^{(k)} = \xi$, если только элементы последовательности не вырождаются в точку: $X^{(k_0+1)} = \xi$, $i \geqslant 1$, для некоторого k_0. Первая часть соотношения (6) следует из формул (3).

Таким образом, теорема 1 гарантирует, что в ее предположениях последовательные приближения $X^{(k)}$, $k \geqslant 0$, сходятся к нулю ξ функции f, причем каждый из этих интервалов содержит искомый нуль. Если же мы применяем (3), начав с интервала $X^{(0)}$, такого что $\xi \notin X^{(0)}$, то найдется индекс k_0, для которого пересечение в (3) пусто. Это легко доказать от противного, используя (7) и предположение, что пересечение непусто. Этот итерационный метод в общей постановке был глубоко исследован. Он связан также с итерацией монотонных функций для нахождения неподвижной точки. Аналогичные процедуры для многочленов были использованы Бауэром еще в 1917 г.

Рассмотрим теперь два уточнения формул (3), возникающие при конкретном выборе точки m. Взяв в качестве m середину интервала, мы получаем следующее утверждение.

Следствие 2. *Если в предположениях теоремы* 1 *сделан выбор*

$$m(X^{(k)}) = \frac{1}{2}(x_1^{(k)} + x_2^{(k)}), \qquad k \geqslant 0,$$

то для последовательности приближений $\{X^{(k)}\}_{k=0}^{\infty}$ *верно неравенство*

$$d(X^{(k+1)}) \leqslant \frac{1}{2}(1 - m_1/m_2)\, d(X^{(k)}), \tag{8}$$

уточняющее (7).

Доказательство. В доказательстве соотношения (7) из теоремы 1 мы имеем при нашем конкретном выборе точки $m(X^w)$, что

$$m(X^{(k)}) - x_1^{(k)} = \frac{1}{2} d(X^{(k)}).$$

Отсюда получаем (8).

Таким образом, при выборе середины интервала в качестве $m(X^{(k)})$ нам гарантировано уменьшение ширины локализующего интервала по крайней мере вдвое.

Рассматривались и другие выборы $m(X^{(k)})$. Например, используется

$$m(X^{(k)}) = m(X^{(k-1)}) - f(m(X^{(k-1)}))/m_0 \quad \text{для} \quad m_0 \in M$$

и

$$m(X^{(k)}) \in \{x_1^{(k)},\ x_2^{(k)}\},$$

если значение $m(X^{ik})$ из предыдущей формулы не принадлежит интервалу $X^{(k)}$. $k \geqslant 1$.

Интервал M, содержащий разностное отношение (2), требуется и в теореме 1, и в следствии 2. Если функция f непрерывно дифференцируема и $f'(x) \neq 0$ для

$$x \in X^{(0)},$$

то в силу теоремы о среднем можно положить

$$M = [\inf_{y \in X^{(0)}} f'(y),\ \sup_{y \in X^{(0)}} f'(y)].$$

В общем случае можно лишь локализовать этот интервал, например, путем интервального оценивания f', т. е. полагая

$$M = f'(X^{(0)}).$$

Условие $m_1 > 0$ может быть гарантировано в случае необходимости с помощью априорной оценки величины

$$\inf_{y \in X^{(0)}} f'(y).$$

8.2. В. Определение оптимального метода

В методе итераций (3), рассмотренном в разд. А, имеется определенная степень свободы при выборе $m(X^{(k)}) \in X^{(k)}$. В зависимости от выбора точек $m(X^{(k)})$ в интервалах $X^{(k)}$ мы получаем различные последовательности локализующих интервалов $\{X^{(k)}\}_{k=0}^{\infty}$. Эти последовательности в общем случае несравнимы почленно в смысле включения интервалов. Поэтому, очевидно, необходимо попытаться найти методы выбора $m(X^{(k)}) \in X^{(k)}$, порождающие последовательности $\{X^{(k)}\}_{k=0}^{\infty}$, в которых ширина отдельных элементов наименьшая возможная. Уточним это требование. Обозначим через $\varphi[X]$ класс функций f, обладающих следующими свойствами для данного интервала

$$X = [x_1, x_2]:$$

(α) $f(x_1) < 0$ и $f(x_2) > 0$;

(β) для интервала $M = [m_1, m_2]$, такого что $m_1 > 0$, имеет место

$$m_1 \leqslant (f(x) - f(y))/(x - y) \leqslant m_2 \text{ для } x \neq y,\ x,\ y \in X.$$

Очевидно, что любая функция $f \in \varphi[X]$ имеет один и только один корень ξ в интервале X. Поэтому выполнены все условия применимости метода итераций (3) и справедливы все утверждения теоремы 1.

Процесс выбора подходящего $m(X^{(k)}) \in X^{(k)}$ мы разобьем на шаги. Обозначим последовательные приближения (3) через $\{X^{(k)}\}_{k=0}^{\infty}$. Для вычисления нового приближения $X^{(k+1)}$ нам нужны величины $m(X^{(k)})$ и $f(m(X^{(k)}))$. Если мы зафиксируем величину $m(X^{(k)}) = x \in X^{(k)}$ из $X^{(k)}$, то $X^{(k+1)}$ будет зависеть только от $f(m(X^{(k)}))$. Но это значение функции f может меняться лишь между $y_1^{(k)}$ и $y_2^{(k)}$, ибо $f \in \varphi[X]$ и значения $f(m(X^{(i)}))$, $0 \leqslant i < k$, зафиксированы. Это позволяет нам определить наибольшую возможную ширину

$$\max \{d(X^{(k+1)}) \mid m(X^{(k)}) = x,\ y_1^{(k)} \leqslant f(m(X^{(k)})) \leqslant y_2^{(k)}\}.$$

Это «наихудший» случай для функции $f \in \varphi[X]$.

Теперь мы определим $\tilde{x} = m(X^{(k)}) \in X^{(k)}$ таким образом, чтобы минимизировать эту наибольшую ширину. Иными словами, вычисляется величина

$$\min_{x \in X^{(k)}} \{\max d(X^{(k+1)}) \mid m(X^{(k)}) = x,\ y_1^{(k)} \leqslant f(m(X^{(k)})) \leqslant y_2^{(k)}\}$$

и соответствующее значение x выбирается в качестве $m(X^{(k)})$. Таким образом, определение $m(X^{(k)})$ производится путем минимизации «наихудшего» случая.

Теперь опишем эту процедуру подробно. Не умаляя общности, рассмотрим случай $f(m(X^{(k-1)})) > 0$. На рис. 2 отмечена область, куда могут попасть значения $f(m(X^{(k)}))$ в предположении, что $f \in \varphi[X]$, причем $f(m(X^{(k-1)})) > 0$.

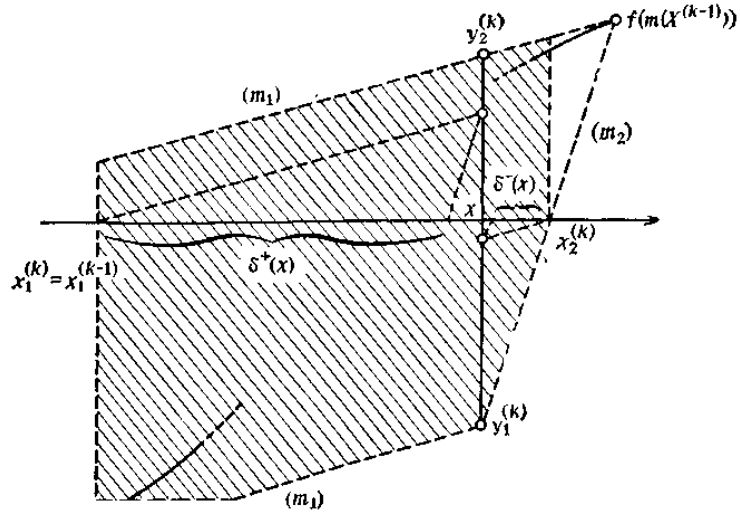


Рис. 2

Нижняя ограничивающая прямая (m_1) может отсутствовать на рис. 2, например, в случае, когда $f(m(X^{(i)})) > 0,\ 0 \leqslant i < k-1$.

Остальные значения $f(m(X^{(i)}))$, $0 \leqslant i < k-1$, не налагают никаких дополнительных ограничений на эту область.

Теперь мы оценим возможные значения $d(X^{(k+1)})$ при заданном $m(X^{(k)}) = x \in X^{(k)}$. Пусть сначала $f(m(X^{(k)})) \geqslant 0$. Для всех значений

$$0 \leqslant f(x) \leqslant (x - x_1^{(k)})\, m_1$$

мы получаем из (3'), что

$$d(X^{(k+1)}) = x - \frac{f(x)}{m_2} - x + \frac{f(x)}{m_1} = f(x)\left(\frac{1}{m_1} - \frac{1}{m_2}\right).$$

Аналогично, для всех значений

$$\left(x - x_1^{(k)}\right) m_1 \leqslant f(x) \leqslant y_2^{(k)}$$

(3') дает

$$d(X^{(k+1)}) = x - f(x)/m_2 - x_1^{(k)}.$$

(Отметим, что ввиду $x_1^{(k)} = \max\{x^{(k-1)}, \quad m(X^{(k-1)}) - f(m \times (X^{(k-1)}))/m_1\}$ мы всегда имеем $y_2^{(k)} \geqslant \left(x - x_1^{(k)}\right) m_1$.) В первом случае $d(X^{(k+1)})$ — монотонно возрастающая функция от $f(x)$, во втором — монотонно убывающая. Для $f(x) = \left(x - x_1^{(k)}\right) m_1$ мы имеем максимум

$$\delta^+(x) = \left(x - x_1^{(k)}\right)(1 - m_1/m_2).$$

Оставшиеся случаи $f(m(X^{(k)})) \leqslant 0$ рассматриваются аналогично и дают максимум $d(X^{(k+1)})$, равный

$$\delta^-(x) = \left(x_2^{(k)} - x\right)(1 - m_1/m_2).$$

На рис. 2 показаны два варианта вычисления $X^{(k+1)}$, приводящие к максимальной ширине $\delta^+(x)$ (и соответственно $\delta^-(x)$).

Теперь мы определим минимум

$$\min_{x \in X^{(k)}} \max\{\delta^+(x), \ \delta^-(x)\}.$$

Выражения $\delta^+(x)$ и $\delta^-(x)$ удовлетворяют условию

$$\delta^+\left(\frac{1}{2}\left(x_1^{(k)} + x_2^{(k)}\right) - t\right) = \delta^-\left(\frac{1}{2}\left(x_1^{(k)} + x_2^{(k)}\right) + t\right)$$

для $0 \leqslant |t| \leqslant \frac{1}{2}\left(x_2^{(k)} - x_1^{(k)}\right)$. Поэтому минимум равен

$$d(X^{(k+1)}) = \frac{1}{2} d(X^{(k)})(1 - m_1/m_2)$$

и достигается в точке

$$\bar{x} = \frac{1}{2}\left(x_1^{(k)} + x_2^{(k)}\right).$$

(Ср. со следствием 2.)

Распространим теперь принцип оптимизации, который мы применили к вычислению $X^{(k+1)}$, на определение величин $m(X^{(i)})$, $0 \leq i < k$. Мы хотим определять значения $m(X^{(0)}) = x^{(0)}$, $\ldots$, $m(X^{(k)}) = x^k$ таким образом, чтобы при этом достигались величины

$$\min_{x^{(0)} \in X^{(0)}} \max_{y_1^{(0)} \leqslant f(x^{(0)}) \leqslant y_2^{(0)}} \cdots \min_{x^{(k)} \in X^{(k)}} \max_{y_1^{(k)} \leqslant f(x^{(k)}) \leqslant y_2^{(k)}} d(X^{(k+1)}).$$

Оказывается, что это просто, так как оптимальная величина $d(X^{(k+1)})$ пропорциональна величине $d(X^{(k)})$ при фиксированном

$m(X^{(k-1)})$. Область допустимых значений функции $f(m(X^{(k-1)}))$ определяется лишь значением $f(m(X^{(k-2)}))$. Поэтому можно провести для $m(X^{(k-1)})$ те же рассуждения, что и для $m(X^{(k)})$, и получить оптимальное значение

$$m(X^{(k-1)}) = \frac{1}{2}(x_1^{(k-1)} + x_2^{(k-1)}).$$

Соответствующим образом мы получаем результаты для

$$m(X^{(i)}), \quad i = k-2,\ k-3,\ \ldots,\ 0$$

в этом порядке.

Теорема 3. *Пусть метод итерации* (3) *применяется к функции* $f \in \varphi[X]$. *Если спользуется правило*

$$m(X^{(k)}) = \frac{1}{2}(x_1^{(k)} + x_2^{(k)}), \quad 0 \leqslant k \leqslant i, \quad i \geqslant 0,$$

то максимальная ширина $d(X^{(i+1)})$ *для функций* $f \in \varphi[X]$ *меньше, чем для любых других выборов точки* $m(X^{(k)})$. *Если* $f \in \varphi[X]$, *то мы имеем*

$$d(X^{(i+1)}) \leqslant \frac{1}{2^{i+1}}(1 - m_1/m_2)^{i+1}\, d(X^{(0)}).$$

Существует $g \in \varphi[X]$, *для которой в последнем соотношении выполняется равенство.*

Доказательство этой теоремы содержится в только что проведенном рассуждении. Следует отметить, что $g \in \varphi[X]$ можно выбрать в виде кусочно-линейной функции, проходящей через точки $(m(X^{(k)}),\ f(m(X^{(k)}))),\ 0 \leqslant k \leqslant i$.

8.3. С. Квадратично сходящиеся методы

В методе (3) мы используем фиксированную пару m_1, m_2 границ для разностных отношений функции f. Эта процедура соответствует интервальному варианту упрощенного метода Ньютона. Если мы предположим, что f непрерывно дифференцируема и для производной f имеется интервальная оценка $f'(X)$, то мы можем определить интервальный вариант и для обычного метода Ньютона. Эта новая процедура получается, если мы модифицируем метод (3), заменяя интервал M на интервал

$$M^{(k)} = f'(X^{(k)}) \tag{9}$$

на k-м шаге итерации. Если известны априорные оценки

$$0 < l_1 \leqslant f'(x) \leqslant l_2. \quad x \in X^{(0)},$$

то можно гарантировать оценку $m_1 > 0$ и использовать выражение

$$M^{(k)} = [m_1^{(k)},\ m_2^{(k)}] = f'(X^{(k)}) \cap L, \quad L = [l_1,\ l_2]. \tag{10}$$

Таким образом, мы получаем

$$X^{(k+1)} = \{in(X^{(k)}) - f(m(X^{(k)}))/M^{(k)}\} \cap X^{(k)}, \quad k \geqslant 0, \tag{11}$$

где $m(X^{(k)}) \in X^{(k}$.

С помощью (11) порождается последовательность интервалов $\{X^{(k)}\}_{k=0}^{\infty}$, для которой можно доказать утверждение, аналогичное теореме 1.

Теорема 4. *Пусть f — непрерывно дифференцируемая функция и f удовлетворяет в интервале $X^{(0)}$ условиям теоремы* 3 п.7.4. *Пусть, далее, для $X^{(0)}$ выполнено* (1), *нуль функции f в $X^{(0)}$ обозначен через ξ и интервалы $M^{(k)}$ определены формулами* (9) *или* (10). *Тогда последовательность $\{X^{(k)}\}_{k=0}^{\infty}$ либо удовлетворяет соотношениям*

(5) $\xi \in X^{(k)},\ k \geqslant 0,$

(6) $X^{(0)} \supset X^{(1)} \supset X^{(2)} \supset \ldots \quad и \quad \lim\limits_{k \to \infty} X^{(k)} = \xi,$

либо стабилизируется на значении $[\xi,\ \xi]$ через конечное число шагов

$$d(X^{(k+1)}) \leqslant (1 - m_1^{(k)}/m_2^{(k)})\, d(X^{(k)}) \leqslant \beta (d(X^{(k)}))^2, \quad \beta \geqslant 0, \tag{12}$$

т. е. R — порядок метода итераций (11) *(см. приложение* А) *удовлетворяет условию*

$$O_R((11), \xi) \geqslant 2.$$

Доказательство. Для $x \in X^{(k)}$ верно

$$\frac{f(x)}{x - \xi} = \frac{f(x) - f(\xi)}{x - \xi} = f'(\eta) \in M^{(k)}, \quad \eta = x + \theta(\xi - x), \quad 0 < \theta < 1.$$

Поэтому аналогичное утверждение для $M^{(k)}$ можно теперь доказать так же, как в теореме 1.

Остается установить (12). Как и в доказательстве теоремы 1, получаем

$$d(X^{(k+1)}) \leqslant \left(1 - \frac{m_1^{(k)}}{m_2^{(k)}}\right) d(X^{(k)}) = \frac{m_2^{(k)} - m_1^{(k)}}{m_2^{(k)}}\, d(X^{(k)}),$$

откуда с помощью 9 п. 7.2 и теоремы 5 микромодуля 24 имеем

$$d(X^{(k+1)}) \leqslant \frac{d(M^{(k)})}{m_1^{(0)}}\, d(X^{(k)}) \leqslant \frac{d(f'(X^{(k)}))}{m_1^{(0)}}\, d(X^{(k)})$$

$$\leqslant (c/m_1^{(0)})\, (d(X^{(k)}))^2,\ c/m_1^{(0)} \geqslant 0.$$

Метод итераций (11) и теорема 4 соответствуют известным формулировкам. Продолжим дальше модифицикаю этого метода Заметим, что если $f(m(X^{(k)}))>0$ (соответственно $f(m(X^{(k)}))<0$. то искомый нуль ξ должен лежать в интервале $[x_1^{(k)},\ m(X^{(k)})]$ (соответственно $[m(X^{(k)}),\ x_2^{(k)}]$. Если $f(m(X^{(k)}))=0$, то $m(X^{(k)})=\xi$, и процесс итерации заканчивается. Поэтому в (11) достаточно положить

$$M^{(k)}=f(Y^{(k)})\cap L,\quad L \text{ взято из (10)},$$

где

$$Y^{(k)}=\begin{cases}[x_1^{(k)},\ m(X^{(k)})], & \text{если } f(m(X^{(k)}))>0,\\ [m(X^{(k)}),\ x_2^{(k)}], & \text{если } f(m(X^{(k)}))<0,\\ X^{(k)} & \text{в противном случае.}\end{cases} \qquad (13)$$

Тогда имеем $f'(Y^{(k)})\subseteq f'(X^{(k)})$ и $d(Y^{(k)})\leqslant d(X^{(k)})$, и, таким образом, легче выполнить условие $m_1^{(k)}>0$. Теорема (4) справедлива и для метода (13)

По поводу выбора точек $m(X^{(k)})\in X^{(k)}$ для метода (11) имеем утверждение, аналогичное следствию 2, и можем провести те же рассуждения, что и разд. В. Мы не будем углубляться в это.

Теперь поясним интервальный метод Ньютона на числовых примерах.

Примеры, (а) Функция

$$f(x)=x^2\left(\frac{1}{3}x^2+\sqrt{2}\sin x\right)-\sqrt{3}/19$$

имеет нуль ξ в интервале $X^{(0)}=[0.1,\ 1]$. Производную

$$f'(x)=x\left(\frac{4}{3}x^2+\sqrt{2}(2\sin x+x\cos x)\right)$$

можно оценить в $X^{(0)}$:

$$l_1=0{,}0013\leqslant f'(x)\leqslant l_2,\qquad x\in X^{(0)}.$$

Локализующие интервалы

$$X^{(k)},\quad k\geqslant 0,\quad \text{по формулам (13)},$$
$$Y^{(k)},\quad k\geqslant 0,\quad \text{по формулам (10)}$$

вычислялись на компьютере в соответствии с (11) до тех пор, пока не переставали происходить изменения. Результаты приведены в табл. 1.

(β) Многочлен

$$p(x)=x(x^9-1)-1$$

имеет единственный нуль β в интервале $X^{(0)}$=[1, 1.5]. Интервальное вычисление $p'(x)$ его производной $p'(x)=10x^9-1$

Таблица 1

$X^{(k)}$	
[1.000000000000, 1.153909281002]	
[1.074525733152, 1.075772270022]	
[1.075764355129, 1.075767749943]	
[1.075766066086, 1.075766066088]	
$Y^{(k)}$	$d(X^{(k)})/d(Y^{(k)})$
[1.000000000000, 1.231579011696]	0.665
[1.018539065305, 1.102153489956]	0.015
[1.071809768336, 1.084762444669]	3×10^{-4}
[1.075647094319, 1.075931180877]	6×10^{-9}
[1.075766039501, 1.075766097327]	...
[1.075766066085, 1.075766066090]	...
[1.075766066085, 1.075766066088]	...

Дает $0 \notin p'(x)$ для $X \subseteq X^{(0)}$. Итерированные локализующие интервалы

$$X^{(k)}, \quad k \geqslant 0, \quad \text{по формулам (13) с } L = p'(X^{(0)}),$$

$$Y^{(k)}, \quad k \geqslant 0, \quad \text{по формулам (9)}$$

вычислялись в соответствии с (11) с использованием (4). Полученные значения, приведены в табл. 2.

Таблица 2

$X^{(k)}$	
[0.09999999999990, 0.4384388546433]	
[0.3382030708107, 0.4384388546433]	
[0.3915056049954, 0.3924484948316]	
[0.3923789206719, 0.3923799504692]	
[0.3923795071350, 0.3923795071378]	
$Y^{(k)}$	$d(X^{(k)})/d(Y^{(k)})$
[0.09999999999990, 0.5181776715881]	0.809
[0.3455588336928, 0.5181776715881]	0.581
[0.3739864679691, 0.4075613703040]	0.028
[0.3922481030413, 0.3925441306206]	0.004
[0.3923794945039, 0.3923795211850]	0.001
[0.3923795071350, 0.3923795071378]	–

На практике определяющим условием для итераций (11) является $m_1 > 0$. Показано, что можно добиться его выполнения с помощью (10), используя известную нижнюю оценку l_1 производной $f'(x)$ в интервале $X^{(0)}$. Если такая оценка l_1 неизвестна или если $0 \in f'(X^{(0)})$, то процедуру (11) нельзя даже начать. Поэтому перед началом итераций следует выполнить несколько шагов метода разбиения интервалов, описанного во введении к этому модулю. Таким образом, мы найдем интервал $Y^{(0)} \subset X^{(0)}$, для которого верно $0 \notin f'(Y^{(0)})$.

Имеется еще одна модификация метода Ньютона, применимая даже в случае $0 \in f'(X^{(0)})$. Рассмотрим этот метод, работающий даже при наличии нескольких нулей функции f в $X^{(0)}$. Если $0 \notin f'(X^{(0)})$, то этот метод совпадает с методом (11). Предположим поэтому, что $0 \in f(X^{(0)})$. Разобьем $X^{(0)}$ на подынтервалы

$$U^{(1)} = [x_1^{(0)},\ m(X^{(0)}) - |f(m(X^{(0)}))|/m_2^{(0)}],$$

$$V^{(1)} = [m(X^{(0)}) + |f(m(X^{(0)}))|/m_2^{(0)},\ x_2^{(0)}],$$

предполагая, что $f(m(X^{(0)})) \neq 0$. Все нули функции f в $X^{(0)}$ должны лежать также и в $U^{(1)} \cup V^{(1)}$. Действительно, нуль $\xi \in X^{(0)}$ должен удовлетворять неравенству

$$\left| \frac{f(m(X^{(0)}))}{\xi - m(X^{(0)})} \right| \leqslant m_2^{(0)},$$

откуда следует

$$|f(m(X^{(0)}))|/m_2^{(0)} \leqslant |\xi - m(X^{(0)})|$$

и

$$\xi \geqslant m(X^{(0)}) + \frac{|f(m(X^{(0)}))|}{m_2^{(0)}}$$

или

$$\xi \leqslant m(X^{(0)}) - \frac{|f(m(X^{(0)}))|}{m_2^{(0)}}$$

Из последнего неравенства, однако, следует, что $\xi \in U^{(1)} \cup V^{(1)}$. Кроме того, при условии $f(m(X^{(0)})) \neq 0$ имеет место

$$d(U^{(1)} \cup V^{(1)}) = d(X^{(0)}) - 2|f(m(X^{(0)}))|/m_2^{(0)} < d(X^{(0)}).$$

Теперь эту процедуру можно повторить для подынтервалов $U^{(1)}$ и $V^{(1)}$ и т. д. Суммарная ширина этих интервалов стремится к нулю. Если f имеет в $X^{(0)}$ только простые нули, то после некоторого шага итерации

все они окажутся в непересекающихся подынтервалах. Далее, после некоторого шага k процедура превращается в итерацию вида (11). После этого либо подынтервалы стремятся к интервалу, содержащему нуль, либо в какой-то момент получается пустое пересечение.

Вместо того чтобы использовать $M^{(k)} := f'(X^{(k)})$ в (11), в соответствии с (3) можно для многочленов использовать интервалы J_1, J_2, J_3 и J_4 из теоремы 7 микромодуля 24 с $y := m(X^{(k)})$ и $X := X^{(k)}$ для локализации производной. Все утверждения теоремы 4 остаются справедливыми. Так как в теореме 7 микромодуля 24 было показано, что J_1 — оптимальная локализация, то для получения наилучшей локализации на каждом шаге разумно использовать именно этот интервал для локализации производной.

В этой связи рассмотрим следующий многочлен.

Пример. Пусть

$$p(x) = x^7 + 3x^6 - 4x^5 - 12x^4 - x^3 - 3x^2 + 4x + 12.$$

Этот многочлен имеет корень ξ в $X^{(0)} = [1.8, 2.4]$. Используя формулы (11), мы находим локализующие интервалы для корня, вычисляя $M^{(k)} := p'(X^{(k)})$ по схеме Горнера. Таблица 3 содержит вычисленные интервалы.

Таблица 3

k	$X^{(k)}$
0	[1.8, 2.4]
1	[1.8, 2 0727618077482]
2	[1.9742900052812, 2 0727618077842]
3	[1.9948757147483, 2 0059215482353]
4	[1.9999888234200, 2 0000115390070]
5	[1.9999999999894, 2 0000000000107]
6	[2.0, 2 0]

Таблица 4

k	$X^{(k)}$
0	[1.8, 2 4]
1	[1.9419538108826, 2 0566964050488]
2	[1.9999999975872, 2 0001112993369]
3	[1 9999999975872, 2 0000000029595]
4	[2 0, 2 0]

Если $p'(X^{(k)})$ заменено на J_1, мы аналогичным образом получаем табл. 4. В табл. 5 мы приводим для каждого шага итерации частное $d_1^{(k)}/d_2^{(k)}$ от деления ширины первого итерированного интервала на ширину второго итерированного интервала.

Таблица 5

k	0	1	2	3
$d_1^{(k)}/d_2^{(k)}$	1	2.37	492 35	3.7×10^6

8.4. D. Методы более высоких порядков

Рассмотрим методы более высоких порядков для нахождения в интервале $X^{(0)} = [x_1^{(0)}, x_2^{(0)}]$ нуля ξ строго монотонно возрастающей или монотонно убывающей вещественной функции, обладающей непрерывными производными достаточно высокого порядка. Эти методы всегда сходятся. Идея описываемого построения принадлежит Эрманну. С помощью этой идеи и методов интервального анализа можно разработать методы, для которых обязательно имеет место сходимость. Как и в разд. A, B, C, предположим, не умаляя общности, что

$$f(x_1^{(0)}) < 0 \quad \text{и} \quad f(x_2^{(0)}) > 0. \tag{1}$$

Пусть, далее, m_1 и m_2 снова обозначают границы разностных отношений

$$0 < m_1 \leqslant \frac{f(x) - f(\xi)}{x - \xi} = \frac{f(x)}{x - \xi} \leqslant m_2 < \infty, \quad \xi \neq x \in X^{(0)}. \tag{2}$$

Границы m_1 и m_2 определяют интервал $M = [m_1, m_2]$. Предполагаем, что функция f $(p+1)$-кратно непрерывно дифференцируема и для $F_i \in I(\mathbb{R})$, $2 \leqslant i \leqslant p+1$, имеет место

$$f^{(i)}(x) \in F_i, \quad x \in X^{(0)}. \tag{14}$$

Интервалы F_i могут быть найдены, например, с помощью интервального оценивания производных функций f на интервале $X^{(0)}$. Если интервальные выражения для производных не определены (например, из-за деления на X при $0 \in X$), то можно подразбить интервал $X^{(0)}$, а затем построить интервалы F_i, взяв объединение оценок, полученных для частичных интервалов.

Рассмотрим теперь следующий метод итераций:

$$
\begin{cases}
x^{(k)} = m(X^{(k)}) \in X^{(k)}, \\
X^{(k+1,\,0)} = \{x^{(k)} - f(x^{(k)})/M\} \cap X^{(k)}, \\
X^{(k+1,\,i)} = \left\{ x^{(k)} - \dfrac{1}{f'(x^{(k)})} \left[f(x^{(k)}) \right.\right. \\
\qquad + \sum_{\nu=2}^{i} \dfrac{f^{(\nu)}(x^{(k)})}{\nu!} (X^{(k+1,\,i-1)} - x^{(k)})^{\nu} \\
\qquad \left.\left. + \dfrac{1}{(i+1)!} F_{i+1}(X^{(k+1,\,i-1)} - x^{(k)})^{i+1} \right] \right\} \cap X^{(k+1,\,i-1)}, \\
X^{(k+1)} = X^{(k+1,\,p)}, \qquad \begin{matrix} 1 \leqslant i \leqslant p, \\ k \geqslant 0. \end{matrix}
\end{cases}
\tag{15}
$$

Как и в разд. А этого модуля, *m(X)* означает произвольный выбор вещественного числа в интервале *X*. Формулы, приведенные выше, требуют на каждом шаге вычисления значений $f(x^{(k)})$, $f'(x^{(k)})$, ..., $f^{(p)}(x^{(k)})$ и обладают следующими свойствами.

Теорема 5. *Пусть функция f имеет (p* + 1) *непрерывную производную, p*≥ 1, *и пусть для* $X^{(0)}$ *верно соотношение* (1). *Пусть* ξ — *нуль функции f в* $X^{(0)}$, *интервал M* = $[m_1, m_2]$ *определен соотношением* (2) *и имеет место* (14). *Тогда для итераций* (15) *верно, что*

$$\xi \in X^{(k)}, \quad k \geqslant 0, \tag{16}$$

и либо

$$X^{(0)} \supset X^{(1)} \supset X^{(2)} \supset \ldots \quad и \quad \lim_{k \to \infty} X^{(k)} = \xi, \tag{17}$$

либо эти итерации стабилизируются через конечное число шагов на точке $[\xi, \xi]$;

$$d(X^{(k+1)}) \leqslant \gamma (d(X^{(k)}))^{p+1} \tag{18}$$

для некоторого $\gamma \geqslant 0$. *Таким образом, согласно теореме* 2 *из приложения* А, *R-порядок последовательности* $\{X^{(k)}\}_{k=0}^{\infty}$ *не меньше, чем p* +1.

Доказательство. (16): Допустим, что $\xi \in X^{(k)}$ для некоторого *k*≥ 0. Это верно для *k* = 0 в силу условия теоремы. Как и в теореме 1, показываем, что

$$\xi \in X^{(k+1,\,0)}.$$

Пусть $\xi \in X^{(k+1,\,i)}$ для некоторого *i*≥ 0. Это верно для *i* =0,

Теперь имеем

$$\xi - x^{(k)} \in X^{(k+1,\,i)} - x^{(k)}.$$

Из формулы Тейлора получаем

$$0 = f(\xi) = f(x^{(k)}) + f'(x^{(k)})(\xi - x^{(k)}) + \ldots + \frac{1}{(i+1)!} f^{(i+1)}(x^{(k)})(\xi - x^{(k)})^{i+1} + \frac{1}{(i+2)!} f^{(i+2)}(\eta_{i+2})(\xi - x^{(k)})^{i+2}$$

для некоторого η_{i+2}, лежащего между $x^{(k)}$ и ξ. Вместе с монотонностью это дает нам соотношение

$$\xi = x^{(k)} - \frac{1}{f'(x^{(k)})}\left[f(x^{(k)}) + \sum_{\nu=2}^{i+1} \frac{f^{(\nu)}(x^{(k)})}{\nu!}(\xi - x^{(k)})^{\nu} + \frac{f^{(i+2)}(\eta_{i+2})}{(i+2)!}(\xi - x^{(k)})^{i+2}\right]$$

$$\in \left\{x^{(k)} - \frac{1}{f'(x^{(k)})}\left[f(x^{(k)}) + \sum_{\nu=2}^{i+1} \frac{f^{(\nu)}(x^{(k)})}{\nu!}(X^{(k+1,\,i)} - x^{(k)})^{\nu} + \frac{F_{i+2}}{(i+2)!}(X^{(k+1,\,i)} - x^{(k)})^{i+2}\right]\right\} \cap X^{(k+1,\,i)} = X^{(k+1,\,i+1)}.$$

Поэтому имеем $\xi \in X^{(k+1,\,i)}$, $0 \leqslant i \leqslant p$, и $\xi \in X^{(k+1)} = X^{(k+1,\,p)}$.

(17): Тем же методом, что и в доказательстве теоремы 1, можно показать, что $X^{(k)} \supset X^{(k+1,\,0)}$, откуда следует $X^{(k)} \supset X^{(k+1)}$, $k \geqslant 0$, так как в формулах (15) берутся пересечения. Затем, как и в теореме 1, можно показать, что

$$d(X^{(k+1,\,0)}) \leqslant (1 - m_1/m_2)\, d(X^{(k)}).$$

Так как в формулах (15) берутся пересечения, получаем

$$d(X^{(k+1)}) \leqslant (1 - m_1/m_2)\, d(X^{(k)}), \qquad k \geqslant 0.$$

Тем же методом, что и в теореме 1, получаем сходимость $\lim\limits_{k\to\infty} X^{(k)} = \xi$. Остающаяся часть утверждения (17) доказывается так же, как в теореме 1.

18): Мы имеем $d(X^{(k+1,\,0)}) \leqslant d(X^{(k)})$, откуда

$$d(X^{(k+1,\,1)}) \leqslant d\left(x^{(k)} - \frac{1}{f'(x^{(k)})}\left(f(x^{(k)}) + \frac{1}{2}F_2(X^{(k+1,\,0)} - x^{(k)})^2\right)\right)$$

$$\leqslant \frac{1}{2}\,d\left(\frac{F_2}{f'(x^{(k)})}\,(X^{(k)} - X^{(k)})^2\right)$$

$$\leqslant \frac{1}{2}\,d\left(\frac{F_2}{M}\,[-(d(X^{(k)}))^2,\ (d(X^{(k)}))^2]\right)$$

$$= |F_2/M|(d(X^{(k)}))^2 = \gamma_1 (d(X^{(k)}))^2$$

с константой $\gamma_1 = |F_2/M|$, не зависящей от k.

Предположим, что для некоторого $i \geq 1$ имеем

$$d(X^{(k+1,\,i)}) \leqslant \gamma_i (d(X^{(k)}))^{i+1},$$

где γ_i не зависит от k. Это только что доказано для $i = 1$ Для $i > 1$ имеем из формул (15) с помощью правил из п.7.2 для вычисления ширины, что

$$
\begin{aligned}
d(X^{(k+1,\,i+1)}) &\leqslant d\left(\sum_{\nu=2}^{i+1}\frac{f^{(\nu)}(x^{(k)})}{\nu!\,f'(x^{(k)})}(X^{(k+1,\,i)}-x^{(k)})^{\nu}\right.\\
&\quad\left.+\frac{1}{(i+2)!}\,\frac{F_{i+2}}{f'(x^{(k)})}(X^{(k+1,\,i)}-x^{(k)})^{i+2}\right)\\
&\leqslant\sum_{\nu=2}^{i+1}\frac{1}{\nu!}\left|\frac{f^{(\nu)}(x^{(k)})}{f'(x^{(k)})}\right|d\left((X^{(k+1,\,i)}-x^{(k)})^{\nu}\right)\\
&\quad+\frac{1}{(i+2)!}\,d\left(\frac{F_{i+2}}{f'(x^{(k)})}(X^{(k+1,\,i)}-x^{(k)})^{i+2}\right)\\
&\leqslant\sum_{\nu=2}^{i+1}\frac{1}{\nu!}\left|\frac{F_{\nu}}{M}\right|\nu\,|X^{(k+1,\,i)}-x^{(k)}|^{\nu-1}\,d(X^{(k+1,\,i)}-x^{(k)})\\
&\quad+\frac{1}{(i+2)!}\,d\left(\frac{F_{i+2}}{f'(x^{(k)})}(X^{(k+1,\,i)}-x^{(k)})^{i+2}\right)\\
&\leqslant\sum_{\nu=2}^{i+1}\frac{1}{(\nu-1)!}\left|\frac{F_{\nu}}{M}\right||X^{(k)}-x^{(k)}|^{\nu-1}\,d(X^{(k+1,\,i)})\\
&\quad+\frac{1}{(i+2)!}\,d\left(\frac{F_{i+2}}{M}(X^{(k)}-x^{(k)})^{i+2}\right)\\
&\leqslant\sum_{\nu=2}^{i+1}\frac{1}{(\nu-1)!}\left|\frac{F_{\nu}}{M}\right|(d(X^{(k)}))^{\nu-1}\gamma_i(d(X^{(k)}))^{i+1}\\
&\quad+\frac{1}{(i+2)!}\,d\left(\frac{F_{i+2}}{M}[-(d(X^{(k)}))^{i+2},\,(d(X^{(k)}))^{i+2}]\right)\\
&=(d(X^{(k)}))^{i+2}\sum_{\nu=2}^{i+1}\frac{1}{(\nu-1)!}\left|\frac{F_{\nu}}{M}\right|\gamma_i(d(X^{(k)}))^{\nu-2}\\
&\quad+\frac{2}{(i+2)!}\left|\frac{F_{i+2}}{M}\right|(d(X^{(k)}))^{i+2}\\
&\leqslant\left(\sum_{\nu=2}^{i+1}\frac{1}{(\nu-1)!}\left|\frac{F_{\nu}}{M}\right|\gamma_i(d(X^{(0)}))^{\nu-2}\right.\\
&\quad\left.+\frac{2}{(i+2)!}\left|\frac{F_{i+2}}{M}\right|\right)(d(X^{(k)}))^{i+2}\\
&=\gamma_{i+1}(d(X^{(k)}))^{i+2}
\end{aligned}
$$

с константой

$$\gamma_{i+1} = \gamma_i \sum_{\nu=2}^{i+1} \frac{1}{(\nu-1)!} \left| \frac{F_\nu}{M} \right| (d(X^{(0)}))^{\nu-1} + \frac{2}{(i+2)!} \left| \frac{F_{i+2}}{M} \right|,$$

не зависящей от k. Поэтому соотношение

$$d(X^{(k+1,\,i)}) \leqslant \gamma_i (d(X^{(k)}))^{i+1}$$

справедливо для $1 \leqslant i \leqslant p$.

Далее

$$d(X^{(k+1)}) = d(X^{(k+1,\,p)}) \leqslant \gamma_p (d(X^{(k)}))^{p+1}$$

с константой γ_p, не зависящей от k. Это доказывает формулу (18) для $\gamma = \gamma_p$, что и завершает доказательство теоремы.

Теперь исследуем случай $p = 1$ несколько подробнее. Для $p = 1$ формулы (15) можно переписать в виде

$$\begin{cases} x^{(k)} = m(X^{(k)}) \in X^{(k)}, \\ X^{(k+1,\,0)} = \{x^{(k)} - f(x^{(k)})/M\} \cap X^{(k)}, \\ X^{(k+1,\,1)} = \left\{x^{(k)} - (1/f'(x^{(k)}))(f(x^{(k)}) + \frac{1}{2} F_2 (X^{(k+1,\,0)} - x^{(k)})^2)\right\} \cap X^{(k+1,\,0)}, \\ X^{(k+1)} = X^{(k+1,\,1)}, \quad k \geqslant 0. \end{cases}$$

Этот метод имеет те же свойства, что и метод Мура, приведенный в разд. С. Если не считать некоторых дополнительных арифметических операций, он менее трудоемок, чем метод Мура, так как значение функции и производной приходится вычислять в одной и той же точке $x^{(k)}$. Для метода Мура производную приходится вычислять с использованием интервала $X^{(k)}$. Это в общем случае требует больше затрат, чем вычисление значений в одной точке $x^{(k)}$. Если интервал F_2 вычисляется легко, то метод (15) с $p = 1$ предпочтительнее метода Мура. Эти результаты полностью справедливы лишь в теории, когда вычисления считаются точными. Если мы хотим гарантировать локализацию нулей при реализации метода на компьютере, то должны учесть влияние погрешностей округления. Это делается путем реализации всех действий в виде машинных интервальных операций. В частности, нам требуется вычислять $f'(x^{(k)})$, используя машинную интервальную арифметику. В этом случае метод (15) при $p = 1$ требует, если не считать нескольких арифметических операций, того же объема вычислений, что и метод Мура. Так как приходится вычислять еще и интервал F_2, следует, видимо, предпочесть метод Мура, если нужно учитывать погрешности округления.

Следует упомянуть, что существует следующий метод:

$$\begin{cases} x^{(k)} = m(X^{(k)}) \in X^{(k)}, \\ X^{(k+1)} = \left\{ x^{(k)} - (1/f'(x^{(k)}))(f(x^{(k)}) \right. \\ \left. + \frac{1}{2} f''(X^{(k)})(X^{(k)} - x^{(k)})^2) \right\} \cap X^{(k)}, \quad k \geqslant 0, \end{cases}$$

в предположении, что f дважды дифференцируема. Мы имеем $\xi \in X^{(k)}$, $k \geqslant 0$. Условия сходимости $\lim\limits_{k \to \infty} X^{(k)} = \xi$ не приводятся. Если метод сходится, то последовательность $d(X^{(k)})$ сходится к нулю квадратично, если $f'(\xi) \neq 0$. По сравнению с методом (15) для p=1 мы должны на каждом шаге производить интервальное оценивание второй производной $f''(X^{(k)})$. Это уменьшает постоянную сходимость, но не улучшает порядок сходимости. (То же верно для метода (15) при p= 1, если на каждом шаге заменить постоянный интервал F_2 на $f''(X^{(k)})$.) Использование этого метода на практике, т. е. с учетом погрешностей округления при вычислении $f(x^{(k)})$ и $f'(x^{(k)})$, увеличивает объем вычислений еще примерно на треть. Так как сходимость не гарантирована, этот метод несколько менее привлекателен. Применяя метод (15), мы должны выбрать конкретный порядок p. Отметим еще, что при определенных предположениях метод (15) оптимален при p = 2, т. е. он является методом третьего порядка.

8.5. Е. Интерполяционные методы

Как и в предыдущем разделе мы рассматриваем сходящиеся методы высшего порядка. На этот раз в основу положен хорошо известный интерполяционный метод нахождения нулей функции. Изменим его с помощью приемов из интервального анализа таким образом, чтобы всегда получать монотонную локализацию корней. Так же, как и в разд. D, нам нужны интервальные оценки старших производных функции f.

Различные методы определяются с помощью (n+1)-элементного множества неотрицательных параметров

$$m_0, m_1, \ldots, m_n.$$

Положим

$$r = \sum_{i=0}^{n} m_i$$

и допустим, что

$$m_0 m_n > 0,$$

откуда следует, что $r > 0$ Мы хотим найти нуль $\xi \in X^{(0)} = [x_1^{(0)}, x_2^{(0)}]$ функции f, которая предполагается дифференцируемой нужное число раз в $X^{(0)}$.

Теперь находятся интервалы H и K, такие что имеет место

$$f'(x) \in H, \ x \in X^{(0)}$$
$$f^{(r)}(x) \in K, \ x \in X^{(0)}. \quad \text{и} \quad 0 \notin H = [h_1, h_2],$$

Чтобы описать очередной $(k + 1)$-й шаг итерации, допустим, что мы уже имеем n+1 попарно различных приближений к нулю ξ:

$$x^{(k)}, \ x^{(k-1)}, \ \ldots, \ x^{(k-n)} \text{ в } X^{(0)}$$

и что последний из найденных локализующих интервалов $X^{(k)}$ имеет вид

$$X^{(k)} = [x^{(k)} - \varepsilon^{(k)}, \ x^{(k)} + \varepsilon^{(k)}]$$

для некоторого $\varepsilon^{(k)} > 0$. Допустим еще, что

$$\xi \neq x_1^{(0)} \text{ и } \xi \neq x_2^{(0)}$$

После исполнения описываемых ниже шагов (S1)—(S5) определяется улучшенный локализующий интервал — новая аппроксимация $X^{(k+1)}$

(S1) *Нахождение единственного интерполяционного многочлена Эрмита*

$$p_k(x) = p_{(m_0, m_1, \ldots, m_n)}(x; x^{(k)}, \ldots, x^{(k-n)}),$$

удовлетворяющего интерполяционным условиям

$$p_k^{(j)}(x^{(k-i)}) = f^{(j)}(x^{(k-i)}), \quad 0 \leqslant i \leqslant n, \quad 0 \leqslant j \leqslant m_i - 1.$$

(Мы полагаем $f^{(0)} = f$, и если $m_i = 0$, то условия в точке $x^{(k-i)}$ отсутствуют.) Определяется интервал $Z^{(k)} \subset X^{(k)}$ по формулам

$$Z^{(k)} = \begin{cases} [x^{(k)} - \varepsilon^{(k)}, \ x^{(k)}], & \text{если} \quad f(x^{(k)}) h_1 > 0, \\ [x^{(k)}, \ x^{(k)} + \varepsilon^{(k)}], & \text{если} \quad f(x^{(k)}) h_1 < 0. \end{cases}$$

(S2) *Нахождение вещественного корня y^k многочлена $p^k(x)$ в интервале*

$$[x^{(k)} - 2\varepsilon^{(k)}, \ x^{(k)} + 2\varepsilon^{(k)}] \cap X^{(0)}.$$

Если такого корня нет, то переходим непосредственно к шагу (S5), *положив*

$$\tilde{X}^{(k+1)} := [\tilde{x}_1^{(k+1)}, \ \tilde{x}_1^{(k+1)}] = Z^{(k)}.$$

(S3) *Вычисление интервала $F^{(k)}$, локализующего значение*

$f(y^{(k)})$, *с помощью выражения*

$$F^{(k)} = \frac{K}{r!} \prod_{i=0}^{n} (y^{(k)} - x^{(k-i)})^{m_i}.$$

(S4) *Вычисление улучшенного локализующего интервала по формуле*

$$\tilde{X}^{(k+1)} = \{y^{(k)} - F^{(k)}/H\} \cap Z^{(k)}.$$

(S5) *Нахождение нового приближения*

$$x^{(k+1)} = (\tilde{x}_1^{(k+1)} + \tilde{x}_2^{(k+1)})/2,$$

нового значения

$$\varepsilon^{(k+1)} = (\tilde{x}_2^{(k+1)} - \tilde{x}_1^{(k+1)})/2$$

и нового интервала

$$X^{(k+1)} = [x^{(k+1)} - \varepsilon^{(k+1)},\ x^{(k+1)} + \varepsilon^{(k+1)}] = \tilde{X}^{(k+1)}.$$

(Если оказывается, что некоторые из точек $x^{(k+1)}$, $x^{(k)}$, $\ldots$, $x^{(k-n+1)}$ *совпали между собой, то можно взять* $x^{(k+1)} \in \tilde{X}^{(k+1)}$ *по формулам*

$$x^{(k+1)} = \begin{cases} \frac{1}{2}(\tilde{x}_1^{(k+1)} + \tilde{x}_2^{(k+1)}) + \tilde{\varepsilon}^{(k)}, & \text{если } f(x^{(k)})\, h_1 > 0, \\ \frac{1}{2}(\tilde{x}_1^{(k+1)} + \tilde{x}_2^{(k+1)}) - \tilde{\varepsilon}^{(k)}, & \text{если } f(x^{(k)})\, h_1 < 0, \end{cases}$$

с подходящим $\tilde{\varepsilon}^{(k)}$, *гарантирующим несовпадение этих точек и соотношение* $x^{(k+1)} \in \tilde{X}^{(k+1)}$. *Такой выбор возможен всегда, когда* $x^{(k-i)} \neq \xi$, $i = 0, 1, \ldots, n$. *Наконец, новое значение* $\varepsilon^{(k+1)}$ *выбирается по формуле*

$$\varepsilon^{(k+1)} = \max\{\tilde{x}_2^{(k+1)} - x^{(k+1)},\ x^{(k+1)} - \tilde{x}_1^{(k+1)}\},$$

что дает

$$X^{(k+1)} = [x^{(k+1)} - \varepsilon^{(k+1)},\ x^{(k+1)} + \varepsilon^{(k+1)}] \supset \tilde{X}^{(k+1)}.)$$

Следует отметить, что определение локализующих интервалов $\{X^{(k)}\}$ не использует перемен знака рассматриваемой функции. Это означает, что новый локализующий интервал будет вычислен всегда, даже если несколько последовательно вычисленных значений функции имеют один и тот же знак. Будет также показано, что шаги (S3) и (S4) могут быть пропущены лишь конечное число раз. Не требуется, чтобы локализующий интервал $X^{(i)}$ содержал какое либо из предыдущих приближений, кроме $x^{(i)}$. Свойства определенного выше алгоритма собраны в следующей теореме.

Теорема 6. *Пусть f — вещественная функция, имеющая нуль* ξ, *для которого задан локализующий интервал*

$$X^{(0)} = \{x \mid | x^{(0)} - x | \leqslant \varepsilon^{(0)}\} \ni \xi,$$

такой что

$$\xi \neq x_1^{(0)}, \quad \xi \neq x_2^{(0)} \quad \left(X^{(0)} = \left[x_1^{(0)}, x_2^{(0)}\right]\right).$$

Пусть далее для ξ *заданы попарно различные приближения*

$$x^{(-n)}, \ x^{(-n+1)}, \ \ldots, \ x^{(0)} \in X^{(0)},$$

и производные функции f удовлетворяют условиям

$$f'(x) \in H, \ \text{где} \ 0 \notin H, \ x \in X^{(0)}$$

и

$$f^{(r)}(x) \in K, \quad x \in X^{(0)}$$

для интервалов H, K и всех x в интервале $X^{(0)}$ *при заданных неотрицательных параметрах*

$$m_0, \ m_1 \ \ldots, \ m_n, \ \text{где} \ r = \sum_{i=0}^{n} m_i, \quad m_0 m_n > 0.$$

Тогда для итераций, заданных шагами (S1—S5), *верны следующие утверждения:*

$$\xi \in X^{(k)}, \quad k \geqslant 0, \tag{19}$$

$$X^{(0)} \supset X^{(1)} \supset X^{(2)} \supset \ldots \ u \lim_{k \to \infty} X^{(k)} = \xi \tag{20}$$

или после конечного числа шагов последовательность стабилизируется на точке $[\xi, \xi]$.

R-порядок итераций (S1) — (S5) *(см. приложение* А) (21)

равен $O_R((S1) - (S5), \xi) \geqslant s$, *где s — единственный положительный корень многочлена*

$$p(s) = s^{n+1} - \sum_{i=0}^{n} m_i s^{n-i}.$$

Доказательство (19): В силу (S1) мы имеем $\xi \in Z^{(k)}$. Остаточный член интерполяционной формулы Эрмита дает

$$f(x) = p_k(x) + \frac{f^{(r)}(\eta)}{r!} \prod_{i=0}^{n} (x - x^{(k-i)})^{m_i},$$

где число η лежит в интервале, образованном точками

$$x, \ x^{(k)}, \ \ldots, \ x^{(k-n)},$$

лежащими в $X^{(k-n)}$. Так как $p_k(y^{(k)}) = 0$, получаем, используя свойство включения интервальной арифметики, что

$$f(y^{(k)}) \in \frac{K}{r!} \prod_{j=0}^{n} (y^{(k)} - x^{(k-j)})^{m_j} = F^{(k)}.$$

Так как (S4)—шаг упрощенного метода Ньютона, мы имеем $\xi \in \tilde{X}^{(k+1)}$, и потому $\xi \in X^{(k+1)}$.

(20): Непосредственно следует из определения шага (S1) вместе с (19) и определением шага (S5), где $X^{(k+1)}$ всегда можно выбрать таким образом, что

$$d(X^{(k+1)}) \leqslant c \cdot d(X^{(k)}),$$

где $(1/2) \leq c < 1$.

(21): Можно показать, что всегда имеется такой индекс k^1, что для всех шагов с номерами $i \geq k_1$ имеется вещественный корень в интервале $[x^{(k)} - 2\varepsilon^{(k)},\ x^{(k)} + 2\varepsilon^{(k)}]$. Доказательство этого утверждения основано на оценке, показывающей, что если $\varepsilon^{(k)}$ достаточно мало, то $p_k(x)$ меняет знак либо между $x^{(k)}$ и $x^{(k)} + 2\varepsilon^{(k)}$, либо между $x^{(k)}$ и $x^{(k)} - 2\varepsilon^{(k)}$. Таким образом, при оценке порядка сходимости мы всегда можем считать, что шаги (S3) и (S4) выполняются. Из (S4) получаем оценку ширины

$$d(\tilde{X}^{(k+1)}) \leqslant d(y^{(k)} - F^{(k)}/H) = d(F^{(k)}/H).$$

Используя (S3) и эту оценку, мы получаем

$$d(\tilde{X}^{(k+1)}) \leqslant \frac{1}{r!} \prod_{j=0}^{n} |y^{(k)} - x^{(k-j)}|^{m_j} d\left(\frac{K}{H}\right).$$

Так как имеет место соотношение

$$y^{(k)} \in [x^{(k-j)} - 2\varepsilon^{(k-j)},\ x^{(k-j)} + 2\varepsilon^{(k-j)}], \quad 0 \leqslant j \leqslant n,$$

получаем, наконец, что

$$d(\tilde{X}^{(k+1)}) \leqslant \tilde{c} \prod_{j=0}^{n} d(X^{(k-j)})^{m_j}$$

и

$$d(X^{(k+1)}) \leqslant c \prod_{j=0}^{n} d(X^{(k-j)})^{m_j}.$$

Окончательный результат получаем, применив теорему 3 из приложения А.

Обсудим теперь еще одно модифицированное семейство методов локализации. Оно получается путем модификации шага (S2) в случае, когда не существует нужного корня $y^{(k)}$. Чтобы сделать возможным

выполнение шага (S4), необходимо тогда определить $y^{(k)}$ иначе. Это делается так.

(S2') *Нахождение вещественного корня $y^{(k)}$ многочлена $p_k(x)$ в интервале $[x^{(k)} - 2\varepsilon^{(k)}, x^{(k)} + 2\varepsilon^{(k)}] \cap X^{(0)}$. Если такого корня не существует, то полагаем*

$$y^{(k)} = \begin{cases} x^{(k)} - \varepsilon^{(k)}, & \text{если} \quad f(x^{(k)})\, h_1 > 0, \\ x^{(k)} + \varepsilon^{(k)}, & \text{если} \quad f(x^{(k)})\, h_1 < 0. \end{cases}$$

(S3') *Вычисление локализующего интервала $F^{(k)}$ для величины $f(y^{(k)})$ с помощью выражения*

$$F^{(k)} = \begin{cases} \dfrac{K}{r!} \prod_{j=0}^{n} (y^{(k)} - x^{(k-j)})^{m_j}, & \text{если} \quad p_k(y^{(k)}) = 0, \\ p_k(y^{(k)}) + \dfrac{K}{r!} \prod_{j=0}^{n} (y^{(k)} - x^{(k-j)})^{m_j} & \text{в противном случае.} \end{cases}$$

Непосредственно очевидно, что теорема 6 истинна без каких-либо изменений и для итераций (SI), (S2'), (S3'), (S4), (S5). Дальнейшие предложенные модификацииотносятся к шагу (S1).

(S1') *Построение единственного интерполяционного многочлена Эрмита*

$$p_k(x) = p_{(m_0, m_1, \ldots, m_n)}(x;\ x^{(k)}, \ldots, x^{(k-n)}),$$

удовлетворяющего условиям

$$p_k^{(j)}(x^{(k-i)}) = f^{(j)}(x^{(k-i)}), \quad 0 \leqslant i \leqslant n, \quad 0 \leqslant j \leqslant m_i - 1.$$

(Полагаем $f^{(0)}=f$ и считаем, что при $m_i=0$ условия в соответствующей точке отсутствуют.) Теперь вычисляем интервал $Z^{(k)} \subset X^{(k)}$ по формуле

$$Z^{(k)} = \{x^{(k)} - f(x^{(k)}) / H\} \cap X^{(k)}.$$

Очевидно, что теорема 6 верна без изменений также и для итераций (S1'), (S2'), (S3'), (S4), (S5).

Среди рассмотренных выше методов вычисления нулей содержался при $n=1$ и $m_0 = m_1 = 1$ интервальный вариант метода секущих. В случае $n = 2$ и $m_0 = m_1 = m_2 = 1$ получаем интервальный вариант метода Мюллера.

Рассмотрим пример применения интервального метода секущих.

Пример. Рассмотрим функцию

$$f(x) = 2xe^{-5} + 1 - 2e^{-5x}$$

на интервале $X^{(0)} = [0, 1]$. Интервальный метод секущих в применении к этой формуле порождает приближения, приведенные в табл. 6.

Таблица 6

$X^{(k)}$
[0.0000000000000, 1.0000000000000]
[0.0000000000000, 0.2703167011351]
[0.1351583505675, 0.1417860581860]
[0.1380542457667, 0.1382588014849]
[0.1382505159257, 0.1382572086567]
[0.1382571542348, 0.1382572086567]
[0.1382571550288, 0.1382571550581]
[0.1382571550553, 0.1382571550581]

Кроме описанного выше интервального метода секущих, существует так называемый интервальный метод regula falsi (ложного основания) Этот метод предполагает примерно те же свойства функции *f,* что и интервальный метод секущих. Он также оказывается интерполяционным методом. В противоположность описанным ранее методам он использует разделенные разности Ньютона. Опишем его кратко.

Предположим, что функция f дважды непрерывно дифференцируема в интервале X и имеет в X единственный и притом простой нуль ξ Далее, мы имеем интервалы H, K, удовлетворяющие условиям

$$f'(x) \in H, \quad x \in H, \quad \text{где} \quad 0 \notin H,$$
$$f''(x) \in K, \quad x \in X.$$

Интервальный *метод regula falsi,* сокращенно RF, описывается теперь так:

$$X^{(0)} = X, \; x^{(0)} = m(X^{(0)}) \;\; (\text{середина интервала } X^{(0)}),$$
$$X^{(1)} = \{x^{(0)} - f(x^{(0)})/H\} \cap X^{(0)},$$
$$\text{RF} \begin{cases} x^{(k)} = m(X^{(k)}) \;\; (\text{середина интервала } X^{(k)}), \\ Z^{(k+1)} = \{x^{(k)} - f(x^{(k)})/H\} \cap X^{(k)}, \\ X^{(k+1)} = \begin{cases} \left\{x^{(k)} - \dfrac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})}\left(f(x^{(k)}) + \dfrac{1}{2}K(Z^{(k+1)} - x^{(k)})\right.\right. \\ \left.\left. \times (Z^{(k+1)} - x^{(k-1)})\right)\right\} \cap Z^{(k+1)}, \quad \text{если} \quad f(x^{(k)}) \neq 0, \\ Z^{(k+1)} \qquad\qquad \text{в противном случае} \end{cases} \end{cases}$$

при $k \geq 1$.

Свойства алгоритма RF резюмированы в следующем утверждении.

Теорема 7. *Пусть дважды дифференцируемая функция f имеет простой нуль ξ в интервале X. Допустим далее, что выполнены условия*

$$f'(x) \in H, \quad x \in X \quad u \quad 0 \notin H,$$
$$f''(x) \in K, \quad x \in X.$$

Тогда последовательность $\{X^{(k)}\}$, *вычисленная согласно процедуре* RF, *либо удовлетворяет условиям*

$$\xi \in X^{(k)}, \quad k \geqslant 0, \tag{22}$$

$$X^{(0)} \supset X^{(1)} \supset X^{(2)} \supset \ldots \; u \lim_{k \to \infty} X^{(k)} = \xi, \tag{23}$$

либо стабилизируется на точке $[\xi, \xi]$ *через конечное число шагов.*

Для некоторого $\gamma \geqslant 0$ *имеет место соотношение*

$$d(X^{(k+1)}) \leqslant \gamma d(X^{(k)}) d(X^{(k-1)}), \tag{24}$$

т. е. (см. приложение А)

$$O_R((\mathrm{RF}), \xi) \geqslant \frac{1}{2}(1 + \sqrt{5}).$$

Доказательство. (22): Доказывается с помощью математической индукции по *k*. Для *k* = 1 утверждение $\xi \in X^{(1)}$ очевидно. Допустим теперь, что для фиксированного *k* мы имеем $x^{(k)} = m(X^{(k)}) \neq \xi$ и $\xi \in X^{(k)}$. Тогда имеет место $\xi \in Z^{(k+1)}$ и $x \neq x_{k-1}$. Рассмотрение интерполяционной формулы Ньютона показывает, что

$$f(\xi) = f(x^{(k)}) + \frac{f(x^{(k)}) - f(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}(\xi - x^{(k)}) + \frac{1}{2} f''(\eta)(\xi - x^{(k)})(\xi - x^{(k-1)}),$$

где η находится в интервале, образованном точками $x^{(k)}$, $x^{(k-1)}$ и ξ. Из $f(\xi) = 0$ следует, что

$$\xi = x^{(k)} - \frac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})}\left(f(x^{(k)}) + \frac{f''(\eta)}{2}(\xi - x^{(k)})(\xi - x^{(k-1)})\right).$$

Из предположения $f''(\eta) \in K$ и того, что $\xi \in Z^{(k+1)}$, а также свойства включения для интервальной арифметики следует, что

$$\xi \in x^{(k)} - \frac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})}\left(f(x^{(k)}) + \frac{K}{2}(Z^{(k+1)} - x^{(k)})(Z^{(k+1)} - x^{(k-1)})\right),$$

откуда в силу $\xi \in Z^{(k+1)}$ мы получаем $\xi \in X^{(k+1)}$.

(23): Следует из построения $Z^{(k)}$ и того, что $X^{(k)} \subset Z^{(k)}$.

(24): Пусть $x^{(k)} \neq \xi$. Тогда мы получаем

$$d(X^{(k+1)}) \leqslant d\left(x^{(k)} - \frac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})}\left(f(x^{(k)}) + \frac{1}{2}K(Z^{(k+1)} - x^{(k)})(Z^{(k+1)} - x^{(k-1)})\right)\right)$$
$$= d\left(-\frac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})} \cdot \frac{1}{2}K(Z^{(k+1)} - x^{(k)})(Z^{(k+1)} - x^{(k-1)})\right).$$

Из

$$\frac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})} = (f'(\tau))^{-1} \in \frac{1}{H}, \quad \tau \in X$$

следует, что

$$d(X^{(k+1)}) \leqslant \frac{1}{2}d((K/H)(Z^{(k+1)} - x^{(k)})(Z^{(k+1)} - x^{(k-1)})).$$

Ввиду

$$X^{(k-1)} \supset X^{(k)} \supset Z^{(k+1)} \supseteq X^{(k+1)},$$

получаем, наконец,

$$d(X^{(k+1)}) \leqslant \frac{1}{2}d((K/H)(X^{(k)} - x^{(k)})(X^{(k-1)} - x^{(k-1)})).$$

Интервал $A \in I(\mathbb{R})$ с центром $m(A)$ удовлетворяет условию

$$A - m(A) = [-d(A)/2,\ d(A)/2],$$

откуда следует соотношение

$$(X^{(k)} - x^{(k)})(X^{(k-1)} - x^{(k-1)}) = [-d(X^{(k)})d(X^{(k-1)})/4,\ d(X^{(k)})d(X^{(k-1)})/4].$$

В применении к предыдущему неравенству оно дает

$$d(X^{(k+1)}) \leqslant \frac{1}{2}d((K/H)[-d(X^{(k)})d(X^{(k-1)})/4,\ d(X^{(k)})d(X^{(k-1)})/4]),$$

откуда получается

$$d(X^{(k+1)}) \leqslant \frac{1}{2}|K/H| \cdot 2 \cdot \frac{1}{4}d(X^{(k)})d(X^{(k-1)})$$
$$= \gamma\, d(X^{(k)})d(X^{(k-1)}),$$

где $\gamma = \frac{1}{4}|K/H|$. Из теоремы 3 (приложение А) мы получаем затем требуемое неравенство для R-порядка.

Используя для данной функции f разделенные разности более высокого порядка, мы можем построить новые методы, порядок сходимости которых тоже лежит между 1 и 2 и которые требуют лишь по одному новому значению функции на каждый шаг итерации.

Можно также построить интервальные варианты некоторых методов, используя интервальный метод regula falsi (RF). Эти методы имеют порядок сходимости выше первого, хотя они используют значения

только самой функции f. Мы опишем здесь такой метод, представляющий собой прямое обобщение интервального метода regula falsi.

Снова предполагается, что функция f имеет простой нуль ξ в интервале X, а интервалы H, K удовлетворяют условиям теоремы 7. Задан параметр p — целое число ≥ 1. Теперь *параметрический метод* regula falsi, короче p-RF, формулируется следующим образом:

$$
p\text{-RF}\left\{\begin{array}{l}
X^{(0)} = X, \quad x^{(0)} = m(X^{(0)}), \\
X^{(1)} = \{x^{(0)} - f(x^{(0)})/H\} \cap X^{(0)}. \\
\text{Для } k \geqslant 1 \text{ вычисляем приближения по следующим формулам:} \\
x^{(k)} = m(X^{(k)}) \text{ (середина интервала } X^{(k)}), \\
X^{(k+1,\,0)} = \{x^{(k)} - f(x^{(k)})/H\} \cap X^{(k)}, \\
X^{(k+1,\,1)} = \left\{\begin{array}{l}
\left\{x^{(k)} - \dfrac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})}\left(f(x^{(k)})\right.\right. \\
\left.\left. + \dfrac{1}{2}K(X^{(k+1,\,0)} - x^{(k)})(X^{(k+1,\,0)} - x^{(k-1)})\right)\right\} \\
\cap X^{(k+1,\,0)}, \quad \text{если} \quad f(x^{(k)}) \neq 0, \\
X^{(k+1,\,0)} \qquad \text{в противном случае.}
\end{array}\right. \\
\text{После этого производим вычисления для } i = 2, 3, \ldots, p \text{ (только для } p > 1). \\
z^{(i)} = m(X^{(k+1,\,i-1)}), \\
X^{(k+1,\,i)} = \left\{\begin{array}{l}
\left\{z^{(i)} - \dfrac{z^{(i)} - x^{(k)}}{f(z^{(i)}) - f(x^{(k)})}\left(f(z^{(i)})\right.\right. \\
\left.\left. + \dfrac{1}{2}K(X^{(k+1,\,i-1)} - z^{(i)})(X^{(k+1,\,i-1)} - x^{(k)})\right)\right\} \\
\cap X^{(k+1,\,i-1)}, \text{ если } f(x^{(k)}) \neq 0, \\
X^{(k+1,\,i-1)} \quad \text{в противном случае,}
\end{array}\right. \\
X^{(k+1)} = X^{(k+1,\,p)}.
\end{array}\right.
$$

Метод *p-RF* обладает следующими свойствами.

Теорема 8. *Пусть функция f дважды непрерывно дифференцируема в интервале X и имеет там нуль* ξ. *Пусть выполнены условия*

$$f'(x) \in H, \quad x \in H, \quad \text{где} \quad 0 \notin H,$$
$$f''(x) \in K, \quad x \in X.$$

Тогда последовательность $\{X^{(k)}\}$, *вычисленная по формулам* p-RF, *удовлетворяет для* $p \geq 1$ *условиям*

$$\xi \in X^{(k)}, \quad k \geqslant 0, \quad (25)$$

$$X^{(0)} \supset X^{(1)} \supset X^{(2)} \supset \ldots, \text{ где } \lim_{k\to\infty} X^{(k)} = \xi, \quad (26)$$

или стабилизируется через конечное число шагов на точке $[\xi, \xi]$.

Для некоторого $\gamma \geq 0$ *справедлива оценка* (27)

$$d(X^{(k+1)}) \leqslant \gamma d(X^{(k)})^p d(X^{(k-1)})$$

и

$$O_R((p\text{-RF}), \xi) \geqslant \frac{1}{2}(p + \sqrt{p^2+4}).$$

Доказательство. При $p = 1$ теорема 8 сводится к теореме 7, поэтому мы предполагаем далее, что $p \geq 2$.

Установим соотношение (25). Мы наметим доказательство методом математической индукции. Доказываемое утверждение очевидно для $k = 0, 1$. Мы допустим, что оно верно для фиксированного k и докажем его для $k+1$. Из нашего индукционного предположения следует, что

$$\xi \in X^{(k+1,\,0)} \text{ и } X^{(k+1,\,0)} = [\xi, \xi] \text{ для } x^{(k)} = \xi.$$

В случае когда $x^{(k)} = \xi$, мы имеем $f(x^{(k)}) = 0$, откуда следует, что

$$X^{(k+1,\,i)} = [\xi, \xi]$$

для $i = 2, 3, \ldots, p$. Поэтому имеем

$$\xi \in X^{(k+1)} = X^{(k+1,\,p)} = [\xi, \xi].$$

Если теперь $x^{(k)} \neq \xi$, то $f(x^{(k)}) \neq 0$ и получаем соотношение

$$0 = f(\xi) = f(x^{(k)}) + \frac{f(x^{(k)}) - f(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}(\xi - x^{(k)}) + \frac{1}{2}f''(\eta)(\xi - x^{(k)})(\xi - x^{(k-1)}),$$

где $x^{(k)} \neq x^{(k-1)}$, $x^{(k-1)} \neq \xi$.

Тем же методом, что и в доказательстве теоремы 7, отсюда получаем, что

$$\xi \in x^{(k)} - \frac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})}\left(f(x^{(k)}) + \frac{1}{2}K(X^{(k+1,\,0)} - x^{(k)}) \times (X^{(k+1,\,0)} - x^{(k-1)})\right)$$

и окончательно $\xi \in X^{(k+1,1)}$. Остается показать что $\xi \in X^{(k+1,i)}$, $i=2,3,\ldots,p$.

Для $x^{(k)} \neq \xi$ и $z^{(i)} \neq \xi$ снова показываем, используя остаточный

член интерполяционной формулы Ньютона, что из $\xi \in X^{(k+1,i-1)}$ всегда следует $\xi \in X^{(k+1,i)}$.

Это снова очевидно для $x^{(k)} \neq \xi$ и $z^{(i)} = \xi$. Так как $\xi \in X^{(k+1,\,i)}$, мы получаем

$$\xi \in X^{(k+1,\,i)}, \quad i = 1,\ 2,\ \ldots,\ p,$$

а потому и

$$\xi \in X^{(k+1)} = X^{(k+1,\,p)}.$$

Соотношение (26) немедленно следует из формул, по которым вычисляется $X^{(k+1,0)}$. Мы не приводим здесь элементарного обоснования этого факта.

(27): Не умаляя общности, предположим, что $x^{(k)} \neq \xi$. Тогда из определения процедуры p-RF немедленно следует, что

$$d(X^{(k+1,\,0)}) < \frac{1}{2}\, d(X^{(k)}).$$

Аналогично тому, как доказано утверждение (24) в теореме (7), получаем, что

$$d(X^{(k+1,\,1)}) \leqslant \frac{1}{4}\,|K/H|\, d(X^{(k)})\, d(X^{(k-1)})$$

и аналогичным образом

$$d(X^{(k+1,\,i)} \leqslant \frac{1}{4}\,|K/H|\, d(X^{(k+1,\,i-1)})\, d(X^{(k)}), \quad i = 2,\ 3,\ \ldots,\ p$$

Эта простая рекурсия дает соотношение

$$d(X^{(k+1,\,i)}) \leqslant \beta_i\, d(X^{(k)})^i\, d(X^{(k-1)}),$$

$$\beta_i = \left(\frac{1}{4}\,|K/H|\right)^i, \quad i = 1,\ 2,\ \ldots,\ p.$$

Используя равенство $X^{(k+1)} = X^{(k+1,\,p)}$, получаем соотношение

$$d(X^{(k+1)}) \leqslant \beta_p\, d(X^{(k)})^p\, d(X^{(k-1)}),$$

из которого снова получаем требуемое неравенство для R-порядка с помощью теоремы 3 из приложения А.

Замечания. Были предложены и исследовались различные модификации методов (3) и (11). Вариант, приведенный в [407], был исследован в [268]. Процедура, аналогичная (3'), используется для уточнения оценок всех нулей в предписанном интервале $[x_1^{(0)},\ x_2^{(0)}]$. Уточнения нижней границы $x_1^{(k+1)}$ вычисляются отдельно от уточнений верхней границы $x_2^{(k+1)}$. Для этой процедуры требуются как значение функции $f(x_1^{(k)})$ (соответственно $f(x_2^{(k)})$), так и оценка $m=|\mathrm{M}|$ величины $|f(x)|$ для $x \in [x_1^{(k)},\ x_2^{(k)}]$. Тогда последовательность нижних границ $x_1^{(0)},\ x_1^{(1)}, \ldots$ сходится к

наименьшему нулю, принадлежащему интервалу $\left[x_1^{(0)}, x_2^{(0)}\right]$. Соответственно последовательность верхних границ сходится к наибольшему нулю функции *f(x)* в $\left[x_1^{(0)}, x_2^{(0)}\right]$. Существует аналогичный метод для многочленов.

Используется неявное представление метода (3) (соответственно (11)). Функция *f(x)* локализуется интервальным выражением

$$f(x) = f\left(x^{(k)}\right) + \left(x - x^{(k)}\right) M^{(k)}, \quad x^{(k)} \in \left[x_1^{(k)}, x_2^{(k)}\right] = X^{(k)},$$

где $M^{(k)}$ определяется аналогично (9). Тогда интервалы $\left[x_1^{(k+1)}, x_2^{(k+1)}\right]$ вычисляются с помощью требования

$$\left[x_1^{(k+1)}, x_2^{(k+1)}\right] := \left\{x \mid x \in X^{(k)}, \ 0 \in f(x)\right\} \cap X^{(k)}.$$

При этом возникают различные случаи в зависимости от того, верно ли, что $0 \in M^{(k)}$. Этот метод сходится при определенных условиях, и если имеется несколько нулей, то возникает несколько подпоследовательностей. Найдена рекурсивная процедура для этого случая. Этот метод был эскизно описан в конце разд. С. Он был применен к нахождению нулей производной дважды непрерывно дифференцируемой функции на интервале, что используется при нахождении глобального минимума функции.

Приведенные в теореме 1 (соответственно в теореме 4) результаты о методе (3) (соответственно о методе (11)) можно обобщить следующим образом. Предполагается, что для функции *f* на интервале $X^{(0)}$ имеется интервал *M,* такой что $0 \notin M$ и

$$\frac{f(x) - f(y)}{x - y} \in M \quad \text{для} \quad x, \ y \in X^{(0)}, \quad x \neq y. \tag{28}$$

Если

$$X^{(1)} = m\left(X^{(0)}\right) - f\left(m\left(X^{(0)}\right)\right)/M \in X^{(0)},$$

то найдется $\xi \in X^{(1)}$, такое что $f(\xi) = 0$. Это будет в том случае, если мы предположим, не умаляя общности, что $f\left(m\left(X^{(0)}\right)\right) > 0$, $m_1 > 0$, а затем $f\left(x_1^{(1)}\right) > 0$. Из этого предположения получается противоречие, если рассмотреть

$$x_1^{(1)} > m\left(X^{(0)}\right) - \frac{f\left(m\left(X^{(0)}\right)\right)}{\left(f\left(m\left(X^{(0)}\right)\right) - f\left(x_1^{(1)}\right)\right)/\left(m\left(X^{(0)}\right) - x_1^{(1)}\right)} \geqslant x_1^{(1)}.$$

Поэтому мы должны иметь $f\left(x_1^{(1)}\right) \leqslant 0$, откуда следует, что $X^{(0)}$, а тогда в силу теоремы 1 также и $X^{(1)}$ содержит нуль ξ.

Мы хотим теперь, используя средства интервальной арифметики, дать короткое доказательство того, что уравнение $f(x)= 0$ имеет корень в интервале $X^{(0)} = [x^{(0)} - r,\ x^{(0)} + r]$. Предположим, что f дважды непрерывно дифференцируема и что

$$|f''(x)| \leqslant \gamma, \quad x \in X^{(0)}.$$

Допустим еще, что $f'(x^{(0)}) \neq 0$, и положим

$$\left|\frac{f(x^{(0)})}{f'(x^{(0)})}\right| = \eta, \quad \left|\frac{1}{f'(x^{(0)})}\right| = \beta.$$

Тогда для любого $y \in X^{(0)}$, $x^{(0)} \neq y$ и некоторого θ, лежащего между $x^{(0)}$ и y, имеем

$$\frac{f(x^{(0)}) - f(y)}{x^{(0)} - y} = f'(x^{(0)}) + \frac{1}{2} f''(\theta)(y - x^{(0)})$$

$$\in f'(x^{(0)}) + \frac{1}{2}\gamma[-r, r] =: M^{(0)}.$$

Если теперь $0 \notin M^{(0)}$ и

$$X^{(1)} = x^{(0)} - f(x^{(0)})/M^{(0)} \subset X^{(0)},$$

то мы показываем тем же методом, что и раньше, что имеется нуль $\xi \in X^{(1)} \subseteq X^{(0)}$. Требование $X^{(1)} \subseteq X^{(0)}$ выполнено тогда и только тогда, когда

$$\frac{1}{2}\beta\gamma r^2 - r + \eta \leqslant 0.$$

Это эквивалентно неравенству

$$\beta\gamma\eta \leqslant \frac{1}{2} \tag{29}$$

вместе с

$$(1 - \sqrt{1 - 2\beta\gamma\eta})/\beta\gamma \leqslant r \leqslant (1 + \sqrt{1 - 2\beta\gamma\eta})/\beta\gamma. \tag{30}$$

Неравенства (29) и (30) гарантируют выполнение условий теоремы Канторовича о существовании нуля в интервале $X^{(0)}$.

Если $X^{(1)} \cap X^{(0)} = \varnothing$, то f не имеет нулей в $X^{(0)}$. Условие $X^{(1)} \cap X^{(0)} = \varnothing$ выполнено тогда и только тогда, когда

$$\frac{1}{2}\beta\gamma r^2 + r - \eta < 0,$$

т. е. при $\eta \neq 0$ тогда и только тогда, когда

$$0 \leqslant r < (-1 + \sqrt{1 + 2\beta\gamma\eta})/\beta\gamma.$$

Это уверждение может быть аналогичным образом использовано для доказательства теорем об исключении в банаховых пространствах.

Микромодуль 27

Методы одновременной локализации вещественных корней многочленов

В этом микромодуле мы рассмотрим интервальные методы ньютоновского типа для вычисления интервалов, локализующих все вещественные корни вещественного многочлена. Сначала рассматривается случай, когда все корни вещественны. Комплексные корни рассматриваются в в микромодуле 29. Для случая, когда все корни вещественные и простые, строим короткошаговый метод, сходящийся быстрее, чем квадратично. В качестве приложения используем этот метод для вычисления всех собственных значений симметрической трехдиагональной матрицы.

Дан вещественный многочлен

$$p(x) = a^{(n)}x^n + a^{(n-1)}x^{n-1} + \ldots + a^{(0)} \qquad (1)$$

и мы предполагаем в дальнейшем, что

$$a^{(n)} = 1.$$

Далее предполагается, что этот многочлен имеет n вещественных корней $\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(n)}$, которые собраны в вектор $(\xi^{(i)})$, причем кратные корни выписаны столько раз, какова их кратность Предполагается, что для всех корней известны локализующие интервалы

$$\xi^{(j)} \in X^{(0,\,j)} = \left[x_1^{(0,\,j)},\ x_2^{(0,\,j)}\right], \quad 1 \leqslant j \leqslant n.$$

Предположим сначала, что все эти локализующие интервалы попарно не пересекаются, т. е.

$$X^{(0,\,j)} \cap X^{(0,\,k)} = \varnothing, \qquad 1 \leqslant j < k \leqslant n. \qquad (2)$$

Многочлен $p(x)$ можно записать в виде

$$p(x) = \prod_{j=1}^{n} (x - \xi^{(j)})$$

или

$$p(x) = (x - \xi^{(i)}) \prod_{j=1,\ j \neq i}^{n} (x - \xi^{(j)}),$$

откуда следует

$$\xi^{(i)} = x - p(x) \Big/ \prod_{j=1,\ j \neq i}^{n} (x - \xi^{(j)}).$$

Если мы выберем $x = x^{(0,\,i)} \in X^{(0,\,i)}$, то получим, что

$$0 \notin \prod_{j=1,\ j \neq i}^{n} (x^{(0,\,i)} - X^{(0,\,j)}),$$

и из (9 п.7.1) следует, что

$$\xi^{(i)} \in X^{(1,\,i)} = \left\{ x^{(0,\,i)} - p(x^{(0,\,i)}) \Big/ \prod_{j=1,\ j \neq i}^{n} (x^{(0,\,i)} - X^{(0,\,j)}) \right\} \cap X^{(0,\,i)}.$$

Таким образом, интервальное выражение, стоящее в правой части этого равенства, задает новый локализующий интервал $X^{(1,i)}$, для которого верно

$$\xi^{(i)} \in X^{(1,\,i)} \subseteq X^{(0,\,i)}.$$

Это соотношение порождает следующую итерационную схему:

$$X^{(k+1,\,i)} = \left\{ x^{(k,\,i)} - p(x^{(k,\,i)}) \Big/ \prod_{j=1,\ j \neq i}^{n} (x^{(k,\,i)} - X^{(k,\,j)}) \right\} \cap X^{(k,\,i)}, \quad (3)$$

где

$$x^{(k,\,i)} \in X^{(k,\,i)}, \qquad 1 \leqslant i \leqslant n, \qquad k \geqslant 0.$$

Для интервального выражения в знаменателе вводится сокращение

$$Q^{(k,\,i)} = \prod_{j=1,\ j \neq i}^{n} (x^{(k,\,i)} - X^{(k,\,j)}).$$

Итерационная схема (3) задает полношаговый метод локализации корней многочлена $\xi^{(i)}$, $1 \leqslant i \leqslant n$. Если в $Q^{(k,\,i)}$ мы всегда используем последнее вычисленное значение локализующих интервалов, то получаем

$$R^{(k,\,i)} = \prod_{j=1}^{i-1} (x^{(k,\,i)} - X^{(k+1,\,j)}) \prod_{j=i+1}^{n} (x^{(k,\,i)} - X^{(k,\,j)}),$$

что приводит к соответствующей короткошаговой итерации. Теперь мы хотим провести для этой короткошаговой итерации такие же рассуждения, как для метода (13 микромодуль 26). Локализующий интервал $X^{(k+1,\,i)}$ урезается до $Y^{(k+1,\,i)}$ в зависимости от знаков выражений $p(x^{(k+1,\,i)})$ и $R^{(k,\,i)}$. Функция sign для интервалов определяется соотношением

$$\operatorname{sign}(X) = \begin{cases} 1, & \text{если} \quad x_1 > 0, \\ -1, & \text{если} \quad x_2 < 0, \\ 0 & \text{в противном случае.} \end{cases} \quad (4)$$

Интервалы $Y^{(k+1,\,i)}$, также содержащие корни $\xi^{(i)}$, определяются тогда следующим образом:

$$Y^{(k+1,\ i)} = \begin{cases} [x_1^{(k+1,\ i)},\ x^{(k+1,\ i)}], & \text{если } \operatorname{sign}(R^{(k,\ i)})\operatorname{sign}(p(x^{(k+1,\ i)})) > 0, \\ [x^{(k+1,\ i)},\ x_2^{(k+1,\ i)}], & \text{если } \operatorname{sign}(R^{(k,\ i)})\operatorname{sign}(p(x^{(k+1,\ i)})) < 0, \\ X^{(k+1,\ i)} & \text{в противном случае.} \end{cases}$$

Заметим, что всегда имеет место соотношение

$$\operatorname{sign}(R^{(0,\ i)}) = \operatorname{sign}(R^{(1,\ i)}) = \ldots, \qquad 1 \leqslant i \leqslant n.$$

Используя только что введенные новые локализующие интервалы, вычисляем теперь новое значение знаменателя с помощью выражения

$$S^{(k+1,\ i)} = \prod_{j=1}^{i-1} (x^{(k+1,\ i)} - Y^{(k+2,\ j)}) \cdot \prod_{j=i+1}^{n} (x^{(k+1,\ i)} - Y^{(k+1,\ j)}).$$

Применяя его, приходим к следующему модифицированному короткошаговому методу:

$$\begin{cases} Y^{(0,\ i)} = X^{(0,\ i)},\ x^{(0,\ i)} \in X^{(0,\ i)}, \\ X^{(k+1,\ i)} = \{x^{(k,\ i)} - p(x^{(k,\ i)})/S^{(k,\ i)}\} \cap X^{(k,\ i)}, \quad \text{где} \\ S^{(k,\ i)} = \prod_{j=1}^{i-1} (x^{(k,\ i)} - Y^{(k+1,\ j)}) \prod_{j=i+1}^{n} (x^{(k,\ i)} - Y^{(k,\ j)}), \\ x^{(k+1,\ i)} \in X^{(k+1,\ i)}, \\ Y^{(k+1,\ i)} = \begin{cases} [x_1^{(k+1,\ i)},\ x^{(k+1,\ i)}], \\ \qquad \text{если } \operatorname{sign}(S^{(k,\ i)})\operatorname{sign}(p(x^{(k+1,\ i)})) > 0, \\ [x^{(k+1,\ i)},\ x_2^{(k+1,\ i)}], \\ \qquad \text{если } \operatorname{sign}(S^{(k,\ i)})\operatorname{sign}(p(x^{(k+1,\ i)})) < 0, \\ X^{(k+1,\ i)} \quad \text{в противном случае.} \end{cases} \end{cases} \tag{5}$$

Оба метода (3), (5) можно считать интервальными вариантами известных методов одновременного нахождения корней вещественных многочленов. Преимущество интервальных вариантов этих методов состоит в том, что они не только дают локализующие интервалы для корней, но и всегда сходятся при сделанных выше допущениях. Это устанавливается в следующей теореме.

Теорема 1. *Пусть дан многочлен* (1), *имеющий n простых корней* $\xi^{(i)}$, $1 \leqslant i \leqslant n$, *причем известны локализующие интервалы* $X^{(0,\ i)} \ni \xi^{(i)}$, *для которых верно* (2). *Тогда последовательность приближений* $\{X^{(k,\ i)}\}_{k=0}^{\infty}$, $1 \leqslant i \leqslant n$, *вычисленная по формулам* (3) *(соответственно* (5)*), либо удовлетворяет условиям*

$$\xi^{(i)} \in X^{(k,\ i)}, \qquad k \geqslant 0,$$

и

$$X^{(0,\,i)} \supset X^{(1,\,i)} \supset X^{(2,\,i)} \supset \ldots, \quad \text{где} \quad \lim_{k\to\infty} X^{(k,\,i)} = \xi^{(i)},$$

либо стабилизируется за конечное число шагов на точке $[\xi^{(i)}, \xi^{(i)}]$.

Теорема 1 получается тем же методом, что и соответствующее утверждение в теореме 1 микромодуль 26, так как каждое из интервальных выражений $Q^{(k,\,i)}$, $S^{(k,\,i)}$ обладает либо свойством (2 микромодуль 26), либо соответствующим свойством в интервале $X^{(k,\,i)}$ для $m_2 < 0$.

Выбирая в обоих методах

$$x^{(k,\,i)} = \frac{1}{2}\left(x_1^{(k,\,i)} + x_2^{(k,\,i)}\right)$$

и рассматривая структуру выражений (3) и (5), немедленно получаем, что ширина локализующего интервала для каждого из корней уменьшается на каждом шаге итерации по крайней мере вдвое.

Теорема 1 частично сохраняется и в случае, когда многочлен имеет кратные корни. Если выпишем эти кратные корни вместе

$$\xi^{(m)}, \ \xi^{(m+1)}, \ \ldots, \ \xi^{(n)},$$

то оба метода (3), (5) нужно изменить таким образом, чтобы вычисление локализующих интервалов производилось только для индексов $1 \leqslant i < m$. Если эти интервалы для простых корней перевычисляются на каждом шаге, а остальные интервалы остаются неизменными, то теорема 1 сохраняется для простых корней.

Можно обобщить метод (3) таким образом, что в теореме 1 удастся заменить предположение (2), касающееся локализующих интервалов $X^{(0,\,i)}$, $1 \leqslant i \leqslant n$, на более слабое условие. Для этого следует полностью использовать возможность варьирования значения $x^{(k,\,i)} \in X^{(k,\,i)}$, а не применять систематическую процедуру выбора, где это значение полагают равным, например, среднему арифметическому границ интервала.

Теперь рассмотрим подробнее поведение последовательности

$$\{d(X^{(k,\,i)})\}_{k=0}^{\infty}, \qquad 1 \leqslant i \leqslant n.$$

Для метода (3) получаем оценку

$$d(X^{(k+1,\,i)}) \leqslant d(\{x^{(k,\,i)} - p(x^{(k,\,i)})/Q^{(k,\,i)}\})$$
$$= d(p(x^{(k,\,i)})/Q^{(k,\,i)}) = |p(x^{(k,\,i)})|\, d(1/Q^{(k,\,i)})$$

с помощью (9 п.7.2), (10 п.7.2) и (14 п.7.2). Так как

$$|p(x^{(k,\,i)})| = |p(x^{(k,\,i)}) - p(\xi^{(i)})| = |(x^{(k,\,i)} - \xi^{(i)})\,p'(\tilde{\eta}^{(k,\,i)})|$$
$$\leqslant d(X^{(k,\,i)})\,|p'(\tilde{\eta}^{(k,\,i)})| \leqslant d(X^{(k,\,i)})\,|p'(X^{(0,\,i)})|,$$

отсюда следует, что

$$d(X^{(k+1,\,i)}) \leqslant d(X^{(k,\,i)})\,|p'(X^{(0,\,i)})|\,d(1/Q^{(k,\,i)}).$$

Применяя теорему 5 микромодуль 24, мы получаем оценку

$$d(1/Q^{(k,\,i)}) \leqslant \gamma^{(k,\,i)}\,d(Q^{(k,\,i)})$$

и, так как

$$Q^{(k,\,i)} \subseteq \prod_{j=1,\ j\neq i}^{n} (X^{(0,\,i)} - X^{(0,\,j)}),$$

получаем далее

$$d\left(\frac{1}{Q^{(k,\,i)}}\right) \leqslant \gamma^{(i)}\,d(Q^{(k,\,i)}) = \gamma^{(i)}\,d\left(\prod_{j=1,\ j\neq i}^{n} (x^{(k,\,i)} - X^{(k,\,j)})\right)$$

с константами $\gamma^{(i)}$, зависящими только от $X^{(0,\,j)}$, $1 \leqslant j \leqslant n$. Многократно применяя (12 п.7.2), мы получаем далее

$$d\left(\frac{1}{Q^{(k,\,i)}}\right) \leqslant \gamma^{(i)} \sum_{j=1,\ j\neq i}^{n} \eta^{(i,\,j)}\,d(X^{(k,\,j)})$$

с подходящими константами $\eta^{(i,j)}$, зависящими только от $X^{(0,j)}$, $1 \leqslant j \leqslant n$, так как $X^{(k,\,j)} \subseteq X^{(0,\,j)}$ и применимы (3 п.7.2) и (9 п.7.2). Резюмируя все это, получаем следующее неравенство:

$$d(X^{(k+1,\,i)}) \leqslant |p'(X^{(0,\,i)})|\,\gamma^{(i)}\,d(X^{(k,\,i)}) \sum_{j=1,\ j\neq i}^{n} \eta^{(i,\,j)}\,d(X^{(k,\,j)}), \quad (6)$$
$$1 \leqslant i \leqslant n.$$

То же самое рассуждение можно провести для метода (5); единственное дополнительное обстоятельство, которое нужно учесть — это соотношение $Y^{(k,\,i)} \subseteq X^{(k,\,i)}$. Это приводит к неравенству

$$d(X^{(k+1,\,i)}) \leqslant |p'(X^{(0,\,i)})|\,\gamma^{(i)}\,d(X^{(k,\,i)}) \left(\sum_{j=1}^{i-1} \eta^{(i,\,j)}\,d(X^{(k+1,\,j)}) \right. \quad (7)$$
$$\left. + \sum_{j=i+1}^{n} \eta^{(i,\,j)}\,d(X^{(k,\,j)})\right), \qquad 1 \leqslant i \leqslant n.$$

В следующем утверждении оценивается R-порядок методов (3) и (5).

Теорема 2. *В предположениях и обозначениях теоремы* 1 *для R-порядка методов* (3) *и* (5) *имеет место*

$$Q_R((3),\,(\xi^{(i)})) \geqslant 2 \quad (8)$$

и

$$Q_R((5), (\xi^{(i)})) \geqslant 1 + \sigma^{(n)}, \tag{9}$$

где $\sigma^{(n)} > 1$ — *единственный положительный корень многочлена*

$$\tilde{q}^{(n)}(y) = y^n - y - 1.$$

Доказательство. (8): Из (6) немедленно получаем, что

$$d(X^{(k+1,\,i)}) \leqslant |p'(X^{(0,\,i)})|\gamma^{(i)}\left(\sum_{l=1,\,l\neq i}^{n}\eta^{(i,\,l)}\right)(d^{(k)})^2$$

$$\leqslant \max_{1\leqslant i\leqslant n}\left\{|p'(X^{(0,\,i)})|\gamma^{(i)}\left(\sum_{l=1,\,l\neq i}^{n}\eta^{(i,\,l)}\right)\right\}(d^{(k)})^2$$

$$\leqslant \gamma(d^{(k)})^2, \quad 1\leqslant i\leqslant n,$$

где

$$d^{(k)} = \max_{1\leqslant i\leqslant n}\{d(X^{(k,\,i)})\}.$$

Отсюда следует, что

$$d^{(k+1)} = \max_{1\leqslant i\leqslant n}\{d(X^{(k+1,\,i)})\} \leqslant \gamma(d^{(k)})^2.$$

Тогда из теоремы 2 приложения А следует, что

$$O_R((3), (\xi^{(i)})) \geqslant 2$$

(9): Доказательство соотношения (9) требует больших усилий.

Пусть

$$\gamma = \max_{1\leqslant i,\,l\leqslant n}\{\eta^{(i,\,l)}|p'(X^{(0,\,i)})|\gamma^{(i)}\}.$$

Тогда мы можем переписать (7) в виде

$$d(X^{(k+1,\,i)}) \leqslant \gamma d(X^{(k,\,i)})\left(\sum_{j=1}^{i-1} d(X^{(k+1,\,j)}) + \sum_{l=i+1}^{n} d(X^{(k,\,l)})\right).$$

Применяя подстановку

$$d(X^{(k,\,i)}) = \frac{1}{(n-1)\gamma}h^{(k,\,i)}, \quad 1\leqslant i\leqslant n, \quad \hat{\varepsilon} = \frac{1}{n-1}$$

это соотношение можно далее переписать в виде

$$h^{(k+1,\,i)} \leqslant \hat{\varepsilon}h^{(k,\,i)}\left(\sum_{l=1}^{i-1} h^{(k+1,\,l)} + \sum_{l=i+1}^{n} h^{(k,\,l)}\right).$$

Предположим без потери общности, что

$$h^{(0,\,i)} \leqslant h < 1, \quad 1\leqslant i\leqslant n.$$

Тогда мы имеем

$$h^{(k+1,\,i)} \leqslant h^{w(k+1,\,i)}, \quad 1\leqslant i\leqslant n, \quad k\geqslant 0.$$

Вектор $u_{\mathrm{p}}^{(k+1)} = (u^{(k+1,\,1)}, \ldots, u^{(k+1,\,n)})^{\mathrm{T}}$, компоненты которого — целые числа, может быть вычислен по правилу

$$u_{\mathrm{p}}^{(k+1)} = \mathscr{A}_{\mathrm{p}} u_{\mathrm{p}}^{(k)}, \qquad k \geqslant 0,$$

с помощью начального вектора $u_{\mathrm{p}}^{(0)} = (1, 1, \ldots, 1)^{\mathrm{T}}$. (Здесь векторы с вещественными компонентами, называемые точечными векторами, обозначаются буквами $a_{\mathrm{p}}, b_{\mathrm{p}}, \ldots$ с индексами p (point), чтобы отличить их от интервальных векторов. То же делается для «точечных матриц».

Тогда матрица $\mathscr{A}_{\mathrm{p}}$ имеет вид

$$\mathscr{A}_{\mathrm{p}} = \begin{pmatrix} 1 & 1 & & & & & \\ & 1 & 1 & & 0 & & \\ & & 1 & 1 & & & \\ & & & & & & \\ & 0 & & & & & \\ & & & & & 1 & 1 \\ 1 & 1 & 0 & & \cdot\cdot & 0 & 1 \end{pmatrix}$$

Доказательство этого соотношения может быть проведено методом математической индукции, и мы не приводим его здесь. Матрица $\mathscr{A}_{\mathrm{p}}$ неотрицательна, и ее направленный граф очевидным образом является сильно связным. Отсюда следует, что $\mathscr{A}_{\mathrm{p}}$ неприводима. Тогда из теоремы Перрона — Фробениуса следует, что $\mathscr{A}_{\mathrm{p}}$ имеет собственное число $\lambda^{(1)}$, равное ее спектральному радиусу $\rho(\mathscr{A}_{\mathrm{p}})$. Матрица $\mathscr{A}_{\mathrm{p}}$ примитивна. Поэтому остальные ее собственные числа удовлетворяют соотношению

$$\lambda^{(1)} = \rho(\mathscr{A}_{\mathrm{p}}) > |\lambda^{(2)}| \geqslant \ldots \geqslant |\lambda^{(n)}|.$$

Так как матрица $\mathscr{A}_{\mathrm{p}}$ примитивна, мы имеем для некотопого натурального числа $k^{(0)}$, что

$$\mathscr{A}_{\mathrm{p}}^{k} = (a_{ij}^{(k)}) > O_{\mathrm{p}}, \qquad k \geqslant k^{(0)}.$$

Для матриц, обладающих двумя последними свойствами, имеем

$$\lim_{k \to \infty} (a_{ij}^{(k+1)} / a_{ij}^{(k)}) = \lambda^{(1)}.$$

Поэтому для данного $\varepsilon > 0$ верно, что

$$a_{ij}^{(k+1)} / a_{ij}^{(k)} \geqslant \rho(\mathscr{A}_{\mathrm{p}}) - \varepsilon, \qquad k \geqslant k(\varepsilon) \geqslant k^{(0)},$$

или

$$a_{ij}^{(k+1)} \geqslant \alpha(\rho(\mathscr{A}_{\mathrm{p}}) - \varepsilon), \qquad 1 \leqslant i, \quad j \leqslant n$$

где

$$\alpha = \min_{1 \leqslant i, j \leqslant n} a_{ij}^{(k)} > 0.$$

Отсюда следует, что

$$a_{ij}^{(k+2)} \geqslant a_{ij}^{(k+1)} \left(\rho\left(\mathscr{A}_{\mathrm{p}}\right) - \varepsilon\right) \geqslant \alpha \left(\rho\left(\mathscr{A}_{\mathrm{p}}\right) - \varepsilon\right),$$

что дает

$$a_{ij}^{(k+r)} \geqslant \alpha \left(\rho\left(\mathscr{A}_{\mathrm{r}}\right) - \varepsilon\right)^{r}, \quad 1 \leqslant i, \quad j \leqslant n, \quad r \geqslant 0.$$

Используя правило вычисления векторов

$$u_{\mathrm{p}}^{(k)},$$

мы получаем тогда

$$u_{\mathrm{p}}^{(k+r)} = \mathscr{A}_{\mathrm{p}}^{k+r} u_{\mathrm{p}}^{(0)} = \left(\sum_{i=1}^{n} a_{ij}^{(k+r)}\right) \geqslant \left(n\alpha \left(\rho\left(\mathscr{A}_{\mathrm{p}}\right) - \varepsilon\right)^{r}\right) e_{\mathrm{p}},$$

где $e_{\mathrm{p}} = (1, 1, \ldots, 1)'$. Поэтому получаем

$$h^{(k+r)} \leqslant h^{a^{(k+r)}} \leqslant h^{n\alpha \left(\rho\left(\mathscr{A}_{\mathrm{p}}\right) - \varepsilon\right)^{r}},$$

$$1 \leqslant i \leqslant n, \quad r \geqslant 0, \quad k \geqslant k(\varepsilon) \geqslant k^{(0)}.$$

Это легко переформулировать в виде

$$d\left(X^{(k+r, i)}\right) \leqslant (\hat{\varepsilon}/\gamma) h^{n\alpha \left(\rho\left(\mathscr{A}_{\mathrm{p}}\right) - \varepsilon\right)^{r}}.$$

Пусть теперь

$$d^{(k)} = \max_{1 \leqslant i \leqslant n} \left\{d\left(X^{(k, i)}\right)\right\}.$$

Тогда получаем

$$d^{(k+r)} \leqslant (\hat{\varepsilon}/\gamma) h^{\alpha n \left(\rho\left(\mathscr{A}_{\mathrm{p}}\right) - \varepsilon\right)^{r}}.$$

Поэтому мы можем заключить, что R-фактор удовлетворяет соотношению

$$R_{\rho\left(\mathscr{A}_{\mathrm{p}}\right) - \varepsilon}\left\{d^{(k)}\right\} = \limsup_{r \to \infty} \left(d^{(k+r)}\right)^{\left[1/\left(\rho\left(\mathscr{A}_{\mathrm{p}}\right) - \varepsilon\right)^{r}\right]}$$

$$\leqslant \limsup_{r \to \infty} \left(\frac{\hat{\varepsilon}}{\gamma} h^{\alpha n \left(\rho\left(\mathscr{A}_{\mathrm{p}}\right) - \varepsilon\right)^{r}}\right)^{\left[1/\left(\rho\left(\mathscr{A}_{\mathrm{p}}\right) - \varepsilon\right)^{r}\right]} = h^{\alpha n} < 1.$$

Отсюда следует, что

$$O_R((5), (\xi^{(i)})) \geqslant \rho\left(\mathscr{A}_{\mathrm{p}}\right) - \varepsilon$$

для всех $\varepsilon > 0$, а потому

$$O_R((5), (\xi^{(i)})) \geqslant \rho(\mathscr{A}_p).$$

Рассмотрим теперь характеристический многочлен

$$q^{(n)}(\lambda)$$

матрицы $\mathscr{A}_p$:

$$q^{(n)}(\lambda) = (\lambda - 1)^n - (\lambda - 1) - 1.$$

Полагая $\tau = \lambda - 1$, мы можем написать его в виде

$$\tilde{q}^{(n)}(\tau) = \tau^n - \tau - 1.$$

Так как

$$\tilde{q}^{(n)}(1) = -1 < 0 \quad \text{и} \quad \tilde{q}^{(n)}(2) = 2^n - 3 \geqslant 1 > 0$$

для $n \geq 2$, то по правилу Декарта многочлен $\tilde{q}^{(n)}(\tau)$ имеет ровно один положительный корень $\sigma^{(n)}$, для которого

$$1 < \sigma^{(n)} < 2.$$

Поэтому спектральный радиус матрицы $\mathscr{A}_p$ удовлетворяет соотношению

$$\rho(\mathscr{A}_p) = 1 + \sigma^{(n)} > 2,$$

откуда получается

$$O_R((5), (\xi^{(i)})) \geqslant 1 + \sigma^{(n)}.$$

Мы опишем теперь одно приложение метода (5). Пусть дана вещественная симметрическая матрица $\mathscr{A}'_p = (a_{ij})$ размерности $n \times n$. Нужно определить собственные числа этой матрицы, т. е. числа λ, для которых выполнено равенство

$$\mathscr{A}'_p x_p = \lambda x_p \quad \text{при} \quad x_p \neq o_p.$$

Чтобы сделать это, применяем конечное число раз ортогональные преобразования подобия

$$\mathscr{A}_p = \mathscr{U}_p^T \mathscr{A}_p \mathscr{U}_p,$$

преобразующие, вообще говоря, неразреженную матрицу $\mathscr{A}'_p$ в матрицу $\mathscr{A}_p$, имеющую вид

$$\mathscr{A}_p = \begin{pmatrix} a^{(1)} & b^{(1)} & & & \\ b^{(1)} & a^{(2)} & b^{(2)} & & 0 \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \\ & 0 & & b^{(n-1)} & a^{(n)} \end{pmatrix}$$

Затем собственные числа матрицы $\mathscr{A}_p$ (а тем самым и матрицы $\mathscr{A}'_p$) вычисляются как корни характеристического многочлена

$$p(\lambda) = \det(\lambda \mathscr{J}_p - \mathscr{A}_p)$$

матрицы $\mathscr{A}_p$. Значение $p(\lambda)$ может быть найдено с помощью следующей рекуррентной формулы:

$$\begin{cases} f^{(0)}(\lambda) = 1, \quad f^{(1)}(\lambda) = \lambda - a^{(1)}, \\ f^{(k)}(\lambda) = (\lambda - a^{(k)}) f^{(k-1)}(\lambda) - (b^{(k-1)})^2 f^{(k-2)}(\lambda), \qquad 2 \leqslant k \leqslant n, \\ p(\lambda) = f^{(n)}(\lambda). \end{cases} \quad (10)$$

Если теперь $\mathscr{A}_p$ имеет только простые собственные числа и для них найдены непересекающиеся локализующие интервалы (например, с помощью теоремы Гершгорина), то мы можем применить (5). Это иллюстрирует следующий пример.

Пример, (а) Рассмотрим матрицу

$$\mathscr{A}_p = \begin{pmatrix} 15 & 1 & & & & & & & \\ 1 & 10 & 1 & & & & & 0 & \\ & 1 & 7 & 1 & & & & & \\ & & 1 & 4 & 1 & & & & \\ & & & 1 & 0 & 1 & & & \\ & & & & 1 & -4 & 1 & & \\ & & & & & 1 & -7 & 1 & \\ & 0 & & & & & 1 & -10 & 1 \\ & & & & & & & 1 & -15 \end{pmatrix}$$

Применяя теорему Гершгорина, мы получаем для собственных чисел матрицы $\mathscr{A}_p$ следующие локализующие интервалы:

$$X^{(0,1)} = [+13.99999999995, +16.00000000005],$$
$$X^{(0,2)} = [+7.999999999974, +12.00000000005],$$
$$X^{(0,3)} = [+4.999999999981, +9.000000000015],$$
$$X^{(0,4)} = [+1.999999999992, +6.000000000022],$$
$$X^{(0,5)} = [-2.000000000008, +2.000000000008],$$
$$X^{(0,6)} = [-6.000000000022, -1.999999999992],$$
$$X^{(0,7)} = [-9.000000000015, -4.999999999981],$$
$$X^{(0,8)} = [-12.00000000005, -7.999999999974],$$
$$X^{(0,9)} = [-17.00000000005, -12.99999999995].$$

Исходя из этих интервалов, строим итерации по формулам (5) и получаем следующие результаты:

$$X^{(5,1)} = [+\underline{15.19709300868}, \quad +\underline{15.19709300872}],$$
$$X^{(4,2)} = [+\underline{10.13174515464}, \quad +\underline{10.13174515471}],$$
$$X^{(4,3)} = [+\underline{7.001927580904}, \quad +\underline{7.001927580960}],$$
$$X^{(4,4)} = [+\underline{3.920346203678}, \quad +\underline{3.920346203715}],$$
$$X^{(5,5)} = [-0.1096791595101 \times 10^{-10}, \quad +0.1096791595101 \times 10^{-10}],$$
$$X^{(4,6)} = [-\underline{3.920346203719}, \quad -\underline{3.920346203674}],$$
$$X^{(4,7)} = [-\underline{7.001927580969}, \quad -\underline{7.001927580895}],$$
$$X^{(3,8)} = [-\underline{10.13174515473}, \quad -\underline{10.13174515463}],$$
$$X^{(3,9)} = [-\underline{15.19709300876}, \quad -\underline{15.19709300866}].$$

Эти интервалы невозможно улучшить, используя имеющуюся программу (см. приложение С). Подчеркнуты знаки, совпадающие в верхней и нижней границах.

(β) Теперь рассмотрим матрицу

$$\mathscr{A}_p = \begin{pmatrix} 12 & 1 & & & 0 \\ 1 & 9 & 1 & & \\ & 1 & 6 & 1 & \\ & & 1 & 3 & 1 \\ 0 & & & 1 & 0 \end{pmatrix}$$

Снова применяя теорему Гершгорина, мы находим для собственных чисел матрицы $\mathscr{A}_p$ следующие локализуюшае интервалы:

$$X^{(0,1)} = [+10.99999999998, +13.00000000003],$$
$$X^{(0,2)} = [+6.999999999970, +11.00000000003],$$
$$X^{(0,3)} = [+3.999999999989, +8.000000000021],$$
$$X^{(0,4)} = [+0.9999999999945, +5.000000000019],$$
$$X^{(0,5)} = [-1.000000000004, +1.000000000004].$$

Следующие уточненные интервалы были вычислены с помощью итерационного метода (5) (ср. с замечаниями, сделанными после теоремы 1):

$$X^{(1,1)} = [+ 12.11013986010, + 12.55506993010],$$
$$X^{(1,2)} = [+ 9.006328989416, + 9.048379503166],$$
$$X^{(1,3)} = [+ 5.999999999958, + 6.000000000041],$$
$$X^{(1,4)} = [+ 2.979804773200, + 2.987022580008],$$
$$X^{(1,5)} = [- 0.3230758693540, - 0.3162523763767].$$

$$X^{(2,1)} = [+ 12.31617201370, + 12.31774922532],$$
$$X^{(2,2)} = [+ 9.016110401580, + 9.016149094187],$$
$$X^{(2,3)} = [+ 5.999999999958, + 6.000000000013],$$
$$X^{(2,4)} = [+ 2.983860239266, + 2.983864788268],$$
$$X^{(2,5)} = [- 0.3168759526293, - 0.3168759526051],$$

$$X^{(3,1)} = [+ 12.31687595112, + 12.31687595546].$$
$$X^{(3,2)} = [+ 9.016136303134, + 9.016136303198],$$
$$X^{(3,3)} = X^{(2,3)}$$
$$X^{(3,4)} = [+ 2.983863696823, + 2.983863696853],$$
$$X^{(3,5)} = [- 0.3168759526293, - 0.3168759526051],$$
$$X^{(4,1)} = [+ 12.31687595258, + 12.31687595266],$$
$$X^{(4,2)} = [+ 9.016136303134, + 9.016136303181],$$
$$X^{(4,3)} = X^{(3,3)}$$
$$X^{(4,4)} = X^{(3,4)}$$
$$X^{(4,5)} = [- 0.3168759526284, - 0.3168759526051].$$

Микромодуль 28
Методы одновременной локализации комплексных корней многочленов

В этом микромодуле обсудим метод одновременной локализации комплексных (в общем случае) корней многочленов, предложенный Гаргантини и Энричи. Пусть задан многочлен

$$p(z) = a^{(n)}z^n + a^{(n-1)}z^{n-1} + \ldots + a^{(1)}z + a^{(0)}, \qquad (1)$$

где $a^{(i)} \in \mathbb{C}$, $0 \leqslant i \leqslant n$, $n \geqslant 2$. Предположим далее, что заданы n интервалов

$$W^{(0,\,i)} = \langle z^{(0,\,i)},\ r^{(0,\,i)} \rangle \in K(\mathbb{C}),$$

для которых

$$\zeta^{(i)} \in W^{(0,\,i)}, \quad p(\zeta^{(i)}) = 0, \quad 1 \leqslant i \leqslant n, \qquad (2)$$

$$W^{(0,\,i)} \cap W^{(0,\,j)} = \varnothing, \quad 1 \leqslant i < j \leqslant n. \qquad (3)$$

Элемент $Z \in K(\mathbb{C})$ представляется в дальнейшем **в** виде

$$Z = \langle m(Z),\ r(Z) \rangle.$$

Рассмотрим следующий метод итерации:

$$\begin{cases} z^{(k,\,i)} = m(W^{(k,\,i)}), \\ C^{(k,\,i)} = \sum\limits_{j=1,\ j \neq i}^{n} \dfrac{1}{z^{(k,\,i)} - W^{(k,\,j)}}, \\ q(z^{(k,\,i)}) = \dfrac{p'(z^{(k,\,i)})}{p(z^{(k,\,i)})} \quad \text{для} \quad p(z^{(k,\,i)}) \neq 0, \\ W^{(k+1,\,i)} = \langle z^{(k+1,\,i)},\ r^{(k+1,\,i)} \rangle = -\dfrac{1}{q(z^{(k,\,i)}) - C^{(k,\,i)}}, \\ 1 \leqslant i \leqslant n, \quad k \geqslant 0, \end{cases} \qquad (4)$$

и пусть

$$r^{(k)} = \max_{1 \leqslant i \leqslant n} \{r^{(k,\,i)}\}, \qquad (5)$$

$$\rho^{(k)} = \min_{1 \leqslant i < j \leqslant n} \{\min\{|z| \mid z \in z^{(k,\,i)} - W^{(k,\,i)}\}\}. \qquad (6)$$

Для $i \neq j$ из (3) следует, что

$$\min\{|z| \mid z \in z^{(0,\,i)} - W^{(0,\,i)}\} = |z^{(0,\,i)} - z^{(0,\,j)}| - r^{(0,\,j)} \geqslant \rho^{(0)}. \qquad (7)$$

Определим еще величины $\eta^{(k)}$ соотношением

$$\rho^{(k)} = (n-1)\eta^{(k)}. \qquad (8)$$

Тогда для итерационной схемы (4) верно следующее.

Теорема 1. *Пусть p(z) есть многочлен* (1) *и его корни* $\xi^{(i)}$, $1 \le i \le n$, *удовлетворяют условиям* (2) *и* (3). *В обозначениях* (5), (6), (8) *пусть*

$$6r^{(0)} \leqslant \eta^{(0)}. \tag{9}$$

Тогда

(a) *итерация* (4) *всегда осуществима, причем*

$$z^{(i)} \in W^{(k,\ i)}, \quad 1 \leqslant i \leqslant n, \quad k \geqslant 0;$$

(b) *имеет место неравенство*

$$r^{(k+1)} \leqslant \frac{1}{\rho^{(0)}\left(\eta^{(0)} - 4r^{(0)}\right)} \left(r^{(k)}\right)^3 \leqslant \frac{1}{12\,(n-1)}\, r^{(k)}, \quad k \geqslant 0.$$

Замечание. Из (6) следует, что $\lim\limits_{k \to \infty} r^{(k)} = 0$. Поэтому в силу (a) получаем, что

$$\lim_{k \to \infty} W^{(k,\ i)} = \xi^{(i)}, \quad 1 \leqslant i \leqslant n.$$

Из (6) следует с помощью теоремы 2 из приложения А, что *R*-порядок итераций (4) удовлетворяет условию $O_R((4), (\xi^{(i)})) \geqslant 3$.

Доказательство, (a): Из

$$\begin{aligned} &| z^{(0,\ i)} - \xi^{(i)} | \leqslant r^{(0,\ i)} \leqslant r^{(0)}, \\ &| z^{(0,\ i)} - \xi^{(j)} | \geqslant | z^{(0,\ i)} - z^{(0,\ j)} | - | z^{(0,\ j)} - \xi^{(j)} | \\ &\qquad \geqslant | z^{(0,\ i)} - z^{(0,\ j)} | - r^{(0,\ j)} \geqslant \rho^{(0)} \end{aligned}$$

следует, что

$$\begin{aligned} | q(z^{(0,\ i)}) | &= \left| \sum_{j=1}^{n} \frac{1}{z^{(0,\ i)} - \xi^{(j)}} \right| \\ &\geqslant \left| \frac{1}{z^{(0,\ i)} - \xi^{(i)}} \right| - \sum_{j=1,\ j \neq i}^{n} \left| \frac{1}{z^{(0,\ i)} - \xi^{(j)}} \right| \\ &\geqslant \frac{1}{r^{(0)}} - \frac{1}{\eta^{(0)}} \quad \text{для} \quad z^{(0,\ i)} \neq \xi^{(i)}. \end{aligned}$$

Из

$$| z^{(0,\ i)} - z^{(0,\ j)} | - r^{(0,\ j)} \geqslant \rho^{(0)} > 0 \tag{10}$$

имеем

$$0 \notin z^{(0,\ i)} - W^{(0,\ j)},$$

а также

$$\frac{1}{z^{(0,\,i)}-W^{(0,\,i)}}\subset\left\langle 0,\ \frac{1}{\rho^{(0)}}\right\rangle.$$

$$C^{(0,\,i)}=\sum_{j=1,\ j\neq i}^{n}\frac{1}{z^{(0,\,i)}-W^{(0,\,i)}}\subset\sum_{j=1,\ j\neq i}^{n}\left\langle 0,\ \frac{1}{\rho^{(0)}}\right\rangle=\left\langle 0,\ \frac{1}{\eta^{(0)}}\right\rangle,$$

$$q(z^{(0,\,i)})-C^{(0,\,i)}\subset\langle q(z^{(0,\,i)}),\ 1/\eta^{(0)}\rangle. \tag{11}$$

Так как

$$|q(z^{(0,\,i)})|-1/\eta^{(0)}\geqslant 1/r^{(0)}-2/\eta^{(0)}>0,$$

то ясно, что

$$0\notin q(z^{(0,\,i)})-C^{(0,\,i)},$$

и потому определены

$$W^{(1,\,i)},\qquad 1\leqslant i\leqslant n.$$

Ввиду

$$\frac{p'(z^{(0,\,i)})}{p(z^{(0,\,i)})}=\sum_{j=1}^{n}\frac{1}{z^{(0,\,i)}-\zeta^{(j)}}$$

из (2) и монотонности включения следует, что

$$\zeta^{(i)}=z^{(0,\,i)}-p(z^{(0,\,i)})\Big/\left[p'(z^{(0,\,i)})-p(z^{(0,\,i)})\sum_{j=1,\ j\neq i}^{n}\frac{1}{z^{(0,\,i)}-\zeta^{(j)}}\right]$$

$$\in z^{(0,\,i)}-\frac{1}{q(z^{(0,\,i)})-C^{(0,\,i)}}=W^{(1,\,i)},\qquad 1\leqslant i\leqslant n.$$

Это доказывает (a) для $k=1$.

(b): Из

$$|z^{(0,\,i)}-z^{(0,\,j)}|^2-(r^{(0,\,j)})^2\geqslant(\rho^{(0)}+r^{(0,\,j)})^2-(r^{(0,\,j)})^2\geqslant(\rho^{(0)})^2$$

получаем, что

$$r\left(\frac{1}{z^{(0,\,i)}-W^{(0,\,j)}}\right)=\frac{r^{(0,\,j)}}{|z^{(0,\,i)}-z^{(0,\,j)}|^2-(r^{(0,\,j)})^2}\leqslant\frac{r^{(0)}}{(\rho^{(0)})^2},$$

а потому

$$r(C^{(0,\,i)})\leqslant\frac{n-1}{\rho^{(0)}}\cdot\frac{r^{(0)}}{\rho^{(0)}}=\frac{r^{(0)}}{\eta^{(0)}\rho^{(0)}}.$$

Используя это неравенство и (11), получаем теперь

$$r(q(z^{(0,\,i)})-C^{(0,\,i)})=r(C^{(0,\,i)}),$$

$$|m(q(z^{(0,\,i)})-C^{(0,\,i)})|\geqslant 1/r^{(0)}-2/\eta^{(0)}+r(q(z^{(0,\,i)})-C^{(0,\,i)})$$

$$=1/r^{(0)}-2/\eta^{(0)}+r(C^{(0,\,i)}),$$

а отсюда и неравенство

$$r(W^{(1,\,i)}) = r\left(\frac{1}{q(z^{(0,\,i)}) - C^{(0,\,i)}}\right)$$
$$= \frac{r(q(z^{(0,\,i)}) - C^{(0,\,i)})}{|m(q(z^{(0,\,i)}) - C^{(0,\,i)})|^2 - (r(q(z^{(0,\,i)} - C^{(0,\,i)}))^2}$$
$$\leqslant \frac{(r^{(0)})^3}{\rho^{(0)}(\eta^{(0)} - 4r^{(0)})},$$

т. е.

$$r^{(1)} \leqslant \frac{(r^{(0)})^3}{\rho^{(0)}(\eta^{(0)} - 4r^{(0)})}. \tag{12}$$

Применяя (9), мы получаем из предыдущей оценки неравенство

$$r^{(1)} \leqslant \frac{1}{12(n-1)} r^{(0)}.$$

Пусть

$$\delta^{(0)} = \max_{1 \leqslant i \leqslant n} \{|z^{(0,\,i)} - z^{(1,\,i)}|\}.$$

Тогда, применяя (6), мы получаем

$$\rho^{(1)} \geqslant \rho^{(0)} - \delta^{(0)} - 2r^{(1)}. \tag{13}$$

Чтобы оценить $\delta^{(0)}$, используем (10), (11) и соотношение

$$z^{(1,\,i)} - z^{(0,\,i)} \in \frac{1}{q(z^{(0,\,i)}) - C^{(0,\,i)}},$$

чтобы получить

$$|z^{(1,\,i)} - z^{(0,\,i)}| \leqslant \left|\frac{1}{\langle q(z^{(0,\,i)}),\, 1/\eta^{(0)}\rangle}\right| = \frac{1}{|q(z^{(0,\,i)})| - 1/\eta^{(0)}} \leqslant \frac{r^{(0)}\eta^{(0)}}{\eta^{(0)} - 2r^{(0)}},$$

что дает, наконец,

$$\delta^{(0)} \leqslant \frac{r^{(0)}\eta^{(0)}}{\eta^{(0)} - 2r^{(0)}}. \tag{14}$$

Используя (12), (13) и (14), получаем из (9), что

$$\eta^{(1)} - 6r^{(1)} = \rho^{(1)}/(n-1) - 6r^{(1)} \tag{15}$$
$$\geqslant \eta^{(0)} - r^{(0)}\left(\frac{\eta^{(0)}}{\eta^{(0)} - 2r^{(0)}} + \frac{8(r^{(0)})^2}{\rho^{(0)}(\eta^{(0)} - 4r^{(0)})}\right)$$
$$\geqslant \eta^{(0)} - 3r^{(0)} \geqslant 0,$$

т. е.

$$\eta^{(1)} \geqslant 6r^{(1)}.$$

Отсюда можно так же, как и раньше, получить, что

$$r^{(2)} \leqslant \frac{1}{\rho^{(1)}(\eta^{(1)} - 4r^{(1)})}(r^{(1)})^3 \leqslant \frac{1}{12(n-1)} r^{(1)}.$$

Из (13) тем же способом, каким было получено (15), выводим, что

$$\eta^{(1)} - 4r^{(1)} \geqslant \eta^{(0)} - r^{(0)}\left(\frac{\eta^{(0)}}{\eta^{(0)} - 2r^{(0)}} + \frac{6(r^{(0)})^2}{\rho^{(0)}(\eta^{(0)} - 4r^{(0)})}\right) \geqslant 0 \quad (16)$$

и

$$\eta^{(1)} \geqslant \eta^{(0)} - r^{(0)}\left(\frac{\eta^{(0)}}{\eta^{(0)} - 2r^{(0)}} + \frac{2(r^{(0)})^2}{\rho^{(0)}(\eta^{(0)} - 4r^{(0)})}\right) \geqslant 0. \quad (17)$$

Используя оба последних неравенства, мы получаем из (9)

$$\eta^{(1)}(\eta^{(1)} - 4r^{(1)}) \geqslant (\eta^{(0)})^2 - \eta^{(0)} r^{(0)}\left(\frac{2\eta^{(0)}}{\eta^{(0)} - 2r^{(0)}} + \frac{8(r^{(0)})^2}{\rho^{(0)}(\eta^{(0)} - 4r^{(0)})}\right)$$
$$\geqslant \eta^{(0)}(\eta^{(0)} - 4r^{(0)}),$$

а потому

$$r^{(2)} \leqslant \frac{1}{\rho^{(0)}(\eta^{(0)} - 4r^{(0)})}(r^{(1)})^3.$$

Оставшаяся часть доказательства получается методом математической индукции.

Теперь проиллюстрируем итерационный метод (4). Для этого рассмотрим задачу вычисления собственных чисел гессен-берговой матрицы с помощью последовательности локализаций. Нужные при этом значения характеристического многочлена и его производных можно вычислить с помощью метода Хаймана. В качестве конкретного примера рассмотрим матрицу

$$\mathcal{H}_p = \begin{pmatrix} 12+16i & 1 & 0 & 0 \\ 0 & 9+12i & 1 & 0 \\ 0 & 0 & 6+8i & 1 \\ 1 & 0 & 0 & 3+4i \end{pmatrix},$$

Где $i = \sqrt{-1}$. Применяя теорему Гершгорина, мы получаем, что каждый из кругов

$$W^{(0,1)} = \langle 12+16i,\ 1\rangle, \qquad W^{(0,2)} = \langle 9+12i,\ 1\rangle,$$
$$W^{(0,3)} = \langle 6+8i,\ 1\rangle, \qquad W^{(0,4)} = \langle 3+4i,\ 1\rangle$$

содержит в точности одно собственное число матрицы H_P. *С* помощью (4) мы получаем уточненные локализующие множества $W^{(k,i)}$ для собственных чисел матрицы H_P. Ниже в табл. 1 используется представление

$$W^{(k,\ i)} = \langle m(W^{(k,\ i)}),\ r(W^{(k,\ i)})\rangle,$$

где

$$m(W^{(k,\ i)}) = \mathrm{Re}(m(W^{(k\ \ i)})) + i\,\mathrm{Im}(m(W^{(k,\ i)})).$$

Таблица 1

k	i	Re	Im	r
1	1	+ 11.99875131516	+ 15.99953080496	0.1001255×10^{-6}
	2	+ 9.003742419628	+ 12.00140833328	0.1494005×10^{-5}
	3	+ 5.996257580383	+ 7.998591666711	0.1493969×10^{-5}
	4	+ 3.001248654837	+ 4.000469195035	0.1000782×10^{-6}
2	1	+ 11.99875136181	+ 15.99953080159	0.1019500×10^{-9}
	2	+ 9.003742437190	+ 12.00140832752	$0.8760740 \times 10^{-10}$
	3	+ 5.996257562811	+ 7.998591672458	$0.3665239 \times 10^{-10}$
	4	+ 3.001248638204	+ 4.000469198423	$0.2555951 \times 10^{-10}$
3	1	+ 11.99875136181	+ 15 99953080159	0.1019496×10^{-9}
	2	+ 9.003742437190	+ 12 00140832752	$0.8760740 \times 10^{-10}$
	3	+ 5.996257562811	+ 7.998591672458	$0.3665353 \times 10^{-10}$
	4	+ 3.001248638204	+ 4 000469198423	$0.2556093 \times 10^{-10}$

Замечания. Итерационный метод, исследованный в теореме 1, можно назвать полношаговым методом. Если на каждом шаге использовать только что уточненные значения приближений, то получим метод, который можно назвать короткошаговым. Можно показать, что этот метод имеет более чем кубический порядок сходимости к нулю радиуса аппроксимирующего круга.

Микромодуль 29
Операции над интервальными матрицами

Множество вещественных матриц размерности $m \times n$ обозначается через $M_{mn}(\mathbb{R})$, а множество комплексных матриц размерности $m \times n$ — через $M_{mn}(C)$. Элементы множества $M_{mn}(\mathbb{R})$, $M_{mn}(\mathbb{C})$ обозначаются через $\mathcal{A}_p$, $\mathcal{B}_p$, $\mathcal{C}_p$, ..., $\mathcal{X}_p$, $\mathcal{Y}_p$, $\mathcal{Z}_p$. Матрицы-столбцы, т. е. вещественные или комплексные векторы, обозначаются через a_p, b_p, c_p, ..., x_p, y_p, z_p. Множество вещественных n-мерных векторов обозначается через $V_n(\mathbb{R})$, множество комплексных векторов — через $V_n(C)$. Аналогичным образом обозначаем через $M_{mn}(I(\mathbb{R}))$ множество матриц, элементами которых являются вещественные интервалы, а через $M_{mn}(I(\mathbb{C}))$—

множество матриц, элементами которых являются комплексные интервалы; здесь $I(\mathbb{C})$ может обозначать как $R(\mathbb{C})$, так и $K(\mathbb{C})$. Элементы множества $M_{mn}(I(\mathbb{R}))$ (соответственно $M_{mn}(I(\mathbb{C}))$ обозначаются через $\mathcal{A}, \mathcal{B}, \mathcal{C}, \ldots, \mathcal{X}, \mathcal{Y}, \mathcal{Z}$, и мы называем их вещественными (соответственно комплексными) интервальными матрицами. Интервальные матрицы-столбцы, т. е. вещественные или комплексные интервальные векторы, обозначаются через $a, b, c, \ldots, x, y, z$. Множество вещественно-интервальных векторов-столбцов обозначается через $V_n(I(\mathbb{R}))$, а множество комплексно-интервальных векторов-столбцов — через $V_n(I(\mathbb{C}))$. Интервальные векторы и матрицы записываются, как обычно, в виде $\mathcal{A}=(A_{ij})$ в случае матриц и $a=(A_i)$ в случае векторов. Интервальная матрица, все компоненты которой являются точечными интервалами, называется *точечной матрицей. Точечные векторы* определяются аналогично. Упомянем очевидные соотношения

$$M_{mn}(I(\mathbb{R}))\subset M_{mn}(R(\mathbb{C}))$$

и

$$V_{mn}(I(\mathbb{R}))\subset V_n(R(\mathbb{C})).$$

Определение 1. Две интервальные матрицы $\mathcal{A}=(A_{ij})$ и $\mathcal{B}=(B_{ij})$ размерности $m\times n$ равны (это записывается, как обычно, в виде $\mathcal{A}=\mathcal{B}$), если равны их соответствующие компоненты. Иными словами, $\mathcal{A}=\mathcal{B}\leftrightarrow A_{ij}=B_{ij}$, $1\leqslant i\leqslant m$, $1\leqslant j\leqslant n$.

Введем частичный порядок на множестве интервальных матриц.

Определение 2. Пусть $\mathcal{A}=(A_{ij})$ и $\mathcal{B}=(B_{ij})$ — интервальные матрицы размерности $m\times n$. Тогда полагаем

$$\mathcal{A}\subseteq\mathcal{B}\Leftrightarrow A_{ij}\subseteq B_{ij},\quad 1\leqslant i\leqslant m,\quad 1\leqslant j\leqslant m.\ \blacksquare$$

Отношение $\mathcal{A}\subset\mathcal{B}$ вводится аналогичным поэлементным определением. Если при этом $\mathcal{A}_n=(a_{ij})$ — точечная матрица, то пишем также $\mathcal{A}_n\in\mathcal{B}$. Каждую интервальную матрицу можно рассматривать как множество точечных матриц. Отношения $\subseteq$ и $\subset$ между множествами точечных матриц понимаются в обычном теоретико-множественном смысле.

Следующая цель — определить операции над интервальными матрицами, формально соответствующие операциям над точечными матрицами.

Определение 3. (а) Пусть $\mathcal{A}=(A_{ij})$, $\mathcal{B}=(B_{ij})$ — две интервальные матрицы размерности $m\times n$. Тогда соотношения

$$\mathscr{A} \pm \mathscr{B} := (A_{ij} \pm B_{ij})$$

определяют соответственно сложение и вычитание интервальных матриц

(b) Пусть $\mathscr{A} = (A_{ij})$ — интервальная матрица размерности $m \times r$ и $\mathscr{B} = (B_{ij})$ — интервальная матрица размерности $r \times n$. Тогда соотношение

$$\mathscr{A}\mathscr{B} := \left(\sum_{\nu=1}^{r} A_{i\nu} B_{\nu j} \right)$$

определяет умножение интервальных матриц В частности, для интервальной матрицы $\mathscr{A} = (A_{ij})$ размерности $n \times r$ и интервального вектора $u = (U_i)$ размерности r мы имеем

$$\mathscr{A}u = \left(\sum_{\nu=1}^{r} A_{i\nu} U_{\nu} \right).$$

(c) Пусть $\mathscr{A} = (A_{ij})$ — интервальная матрица и X — интервал. Тогда полагаем

$$X\mathscr{A} = \mathscr{A}X := (XA_{ij}).$$

В дальнейшем предполагается, что интервальные матрицы, участвующие в интервальной операции, имеют нужное для этой операции число строк и столбцов, и это обстоятельство не будет специально оговариваться. Далее предполагается, что интервальные операнды (т. е. аргументы операций) имеют подходящие элементы Если, например, мы имеем $\mathscr{A} \in M_{mn}(K(\mathbb{C}))$, то произведение AB определено, только если $\mathscr{B} \in M_{nr}(K(\mathbb{C}))$.

Операции над интервальными матрицами и векторами были формально введены в определения 3. Для вещественных интервальных операций мы имеем простое определение 2 п.7.1. Для интервальных матриц аналогичное определение невозможно, однако в общем случае имеет место

$$\{\mathscr{A}_p \mathscr{B}_p \mid \mathscr{A}_p \in \mathscr{A},\ \mathscr{B}_p \in \mathscr{B}\} \subseteq \{\mathscr{C}_p \mid \mathscr{C}_p \in \mathscr{A}\mathscr{B}\}.$$

Доказательство получается с помощью монотонности отношения включения для интервальных операций. Следующий пример показывает, что равенство не имеет места в общем случае. Пусть

$$\mathscr{A}_p = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}, \qquad u_p = \begin{pmatrix} [0, & 1] \\ [0, & 1] \end{pmatrix}$$

Тогда мы имеем

$$\mathscr{A}_p u = \begin{pmatrix} [& 0, & 2] \\ [-1, & & 1] \end{pmatrix}$$

и если возьмем

$$x_p = \begin{pmatrix} 2 \\ -1 \end{pmatrix} \in \mathcal{A}_p u,$$

то видим, чш не найдется $y_p \in u$, такого что $\mathcal{A}_p y_p = x_p$.

Пусть $\mathcal{A}$, $\mathcal{B} \in M_{mn}(I(\mathbb{R}))$ и $c_p \in V_n(\mathbb{R})$. Тогда справедливы соотношения

$$\left\{ \begin{array}{c} \{\mathcal{A}_p \pm \mathcal{B}_p \mid \mathcal{A}_p \in \mathcal{A},\ \mathcal{B}_p \in \mathcal{B}\} = \mathcal{A} \pm \mathcal{B}, \\ \{\mathcal{A}_p c_p \mid \mathcal{A}_p \in \mathcal{A}\} = \mathcal{A} c_p, \end{array} \right\} \qquad (1)$$

которыми будем пользоваться в дальнейшем.

Множество интервальных матриц замкнуто относительно операций из определения 3. Множество вещественных или комплексных матриц изоморфно соответствующему множеству точечных матриц. Именно по этой причине в определении 3 использованы те же символы операций, что и для соответствующих вещественных и комплексных операций.

Теперь сформулируем некоторые свойства введенных операций.

Теорема 4. *Пусть* $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ — *интервальные матрицы. Тогда*

$$\mathcal{A} + \mathcal{B} = \mathcal{B} + \mathcal{A}, \qquad (2)$$

$$\mathcal{A} + (\mathcal{B} + \mathcal{C}) = (\mathcal{A} + \mathcal{B}) + \mathcal{C}, \qquad (3)$$

$$\mathcal{A} + O_p = O_p + \mathcal{A} = \mathcal{A}, \text{ где } O_p - \text{нулевая матрица}, \qquad (4)$$

$$\mathcal{A} \mathcal{I}_p = \mathcal{I}_p \mathcal{A} = \mathcal{A}, \text{ где } \mathcal{I}_p - \text{единичная матрица}, \qquad (5)$$

$$\left\{ \begin{array}{l} (\mathcal{A} + \mathcal{B})\mathcal{C} \subseteq \mathcal{A}\mathcal{C} + \mathcal{B}\mathcal{C} \\ \mathcal{C}(\mathcal{A} + \mathcal{B}) \subseteq \mathcal{C}\mathcal{A} + \mathcal{C}\mathcal{B} \end{array} \right\} \quad (\text{субдистрибутивность}), \qquad (6)$$

$$(\mathcal{A} + \mathcal{B})\mathcal{C}_p = \mathcal{A}\mathcal{C}_p + \mathcal{B}\mathcal{C}_p, \qquad (7)$$

$$\mathcal{C}_p(\mathcal{A} + \mathcal{B}) = \mathcal{C}_p\mathcal{A} + \mathcal{C}_p\mathcal{B}, \qquad (8)$$

$$\left\{ \begin{array}{c} \mathcal{A}(\mathcal{B}_p\mathcal{C}_p) \subseteq (\mathcal{A}\mathcal{B}_p)\mathcal{C}_p, \\ (\mathcal{A}_p\mathcal{B})\mathcal{C} \subseteq \mathcal{A}_p(\mathcal{B}\mathcal{C}) \text{ для } \mathcal{C} = -\mathcal{C}, \\ \mathcal{A}_p(\mathcal{B}\mathcal{C}_p) = (\mathcal{A}_p\mathcal{B})\mathcal{C}_p, \\ \mathcal{A}(\mathcal{B}\mathcal{C}) = (\mathcal{A}\mathcal{B})\mathcal{C} \text{ для } \mathcal{A},\ \mathcal{B},\ \mathcal{C} \in M_{mn}(I(\mathbb{R})) \\ \text{и } \mathcal{B} = -\mathcal{B},\ \mathcal{C} = -\mathcal{C}. \end{array} \right. \qquad (9)$$

Доказательство. Соотношения (2)—(8) доказываются поэлементно с использованием формул из теорем 4 п.7.1 и 8 п.7.4. Докажем (9) для квадратных матриц. Из дистрибутивности (8 п.7.1) и формулы (6 п.7.4) получаем

$$\mathscr{A}(\mathscr{B}_{\mathrm{p}}\mathscr{C}_{\mathrm{p}}) = \sum_{l=1}^{n} A_{il}\left(\sum_{t=1}^{n} b_{lt}c_{tk}\right)\Big) \subseteq \left(\sum_{l=1}^{n}\sum_{t=1}^{n} A_{il}b_{lt}c_{tk}\right)$$
$$= \left(\sum_{l=1}^{n}\left(\sum_{j=1}^{n} A_{ij}b_{jl}\right)c_{lk}\right) = (\mathscr{A}\mathscr{B}_{\mathrm{p}})\mathscr{C}_{\mathrm{p}},$$

что доказывает первое из соотношений (9). Равенства

$$(\mathscr{A}_{\mathrm{p}}\mathscr{B})\mathscr{C} = \left(\sum_{l=1}^{n}\left(\sum_{k=1}^{n} a_{ik}B_{kl}\right)C_{lj}\right) = \left(\sum_{l=1}^{n}\left|\sum_{k=1}^{n} a_{ik}B_{kl}\right|C_{lj}\right)$$
$$\subseteq \left(\sum_{l=1}^{n}\left(\sum_{k=1}^{n}|a_{ik}||B_{kl}|\right)C_{lj}\right) \subseteq \left(\sum_{l=1}^{n}\left(\sum_{k=1}^{n}|a_{ik}||B_{kl}|C_{lj}\right)\right)$$
$$= \left(\sum_{k=1}^{n}|a_{ik}|\left(\sum_{l=1}^{n}|B_{kl}|C_{lj}\right)\right)$$
$$= \left(\sum_{k=1}^{n} a_{ik}\left(\sum_{l=1}^{n} B_{kl}C_{lj}\right)\right) = \mathscr{A}_{\mathrm{p}}(\mathscr{B}\mathscr{C}).$$

дают второе соотношение. Из равенств

$$(\mathscr{A}_{\mathrm{p}}\mathscr{B})\mathscr{C}_{\mathrm{p}} = \left(\sum_{l=1}^{n}\left(\sum_{k=1}^{n} a_{ik}B_{kl}\right)c_{lj}\right) = \left(\sum_{l=1}^{n}\left(\sum_{k=1}^{n} a_{ik}B_{kl}c_{lj}\right)\right)$$
$$= \left(\sum_{k=1}^{n} a_{ik}\left(\sum_{l=1}^{n} B_{kl}c_{lj}\right)\right)$$

мы получаем третье соотношение.

Последнее соотношение получается следующим образом с помощью третьей из формул (8 п.71.):

$$\mathcal{A}(\mathcal{B}\mathcal{C}) = \left(\sum_{k=1}^{n} A_{ik}\left(\sum_{l=1}^{n} B_{kl}C_{lj}\right)\right)$$
$$= \left(\sum_{k=1}^{n} A_{ik}\left(\sum_{l=1}^{n} |B_{kl}|C_{lj}\right)\right)$$
$$= \left(\sum_{k=1}^{n} |A_{ik}|\left(\sum_{l=1}^{n} |B_{kl}|C_{lj}\right)\right)$$
$$= \left(\sum_{k=1}^{n} \left(\sum_{l=1}^{n} |A_{ik}||B_{kl}|C_{lj}\right)\right)$$
$$= \left(\sum_{l=1}^{n} \left(\sum_{k=1}^{n} |A_{ik}||B_{kl}|C_{lj}\right)\right)$$
$$= \left(\sum_{l=1}^{n} \left(\sum_{k=1}^{n} |A_{ik}||B_{kl}|\right)C_{lj}\right)$$
$$= \left(\sum_{l=1}^{n} \left(\sum_{k=1}^{n} |A_{ik}|B_{kl}\right)C_{lj}\right)$$
$$= \left(\sum_{l=1}^{n} \left(\sum_{k=1}^{n} A_{ik}B_{kl}\right)\right)C_{lj} = (\mathcal{A}\mathcal{B})\mathcal{C}.$$

В общем случае ассоциативный закон не имеет места для интервальных матриц. Это показывает следующий

Пример.

$$\begin{bmatrix} [-1, 1] & 1 \\ -1 & [0, 1] \end{bmatrix}\left\{\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} -1 & 0 \\ 1 & -1 \end{bmatrix}\right\} = \begin{bmatrix} 1 & [-2, 0] \\ [0, 1] & [0, 1] \end{bmatrix},$$

$$\left\{\begin{bmatrix} [-1, 1] & 1 \\ -1 & [0, 1] \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}\right\}\begin{bmatrix} -1 & 0 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} [-1, 3] & [-2, 0] \\ [0, 1] & [0, 1] \end{bmatrix}.$$

Основное свойство монотонности включения справедливо и для интервальных матричных операций.

Теорема 5. *Пусть* $\mathcal{A}^{(k)}, \mathcal{B}^{(k)}, k = 1, 2,$ — *интервальные матрицы.*

Далее, пусть X, Y—интервалы и

$$\mathcal{A}^{(k)} \subseteq \mathcal{B}^{(k)}, \quad k = 1, 2 \quad и \quad X \subseteq Y$$

Тогда соотношения

$$\begin{cases} \mathcal{A}^{(1)} * \mathcal{A}^{(2)} \subseteq \mathcal{B}^{(1)} * \mathcal{B}^{(2)}, \\ X\mathcal{A}^{(1)} \subseteq Y\mathcal{B}^{(1)} \end{cases} \tag{10}$$

имеют место для $* \in \{+, -, \cdot\}$.

Доказательство соотношений (10) проводится покомпонентно с использованием (9 п.7.1) и теоремы 9 п.7.4. Имеет место частный случай соотношений (10):

$$\mathcal{A}_\mathrm{p} \in \mathcal{A}, \quad \mathcal{B}_\mathrm{p} \in \mathcal{B} \Rightarrow \mathcal{A}_\mathrm{p} * \mathcal{B}_\mathrm{p} \in \mathcal{A} * \mathcal{B}, \ * \in \{+, -, \cdot\},$$
$$x \in X, \quad \mathcal{A}_\mathrm{p} \in \mathcal{A} \Rightarrow x\mathcal{A}_\mathrm{p} \in X\mathcal{A}.$$

Введем теперь понятия ширины и абсолютной величины интервальных матриц.

Определение 6. Пусть $\mathsf{A} = (A_{ij})$ — интервальная матрица. Тогда

(a) вещественная неотрицательная матрица

$$d(\mathcal{A}) := (d(A_{ij}))$$

называется шириной матрицы A*;*

(b) вещественная неотрицательная матрица

$$|\mathcal{A}| := (|A_{ij}|)$$

называется матрицей абсолютных величин или абсолютной величиной матрицы A .

Соберем теперь в одном месте некоторые свойства ширины и абсолютной величины интервальных матриц. Частичный порядок

$$\mathcal{X}_\mathrm{p} \leqslant \mathcal{Y}_\mathrm{p} \Leftrightarrow x_{ij} \leqslant y_{ij}, \quad 1 \leqslant i \leqslant m, \quad 1 \leqslant j \leqslant n$$

используется здесь для вещественных интервальных матриц X_p и $\mathcal{Y}_\mathrm{p}$ размерности $m \times n$. Перечислим эти свойства.

$$\mathcal{A} \subseteq \mathcal{B} \Rightarrow d(\mathcal{A}) \leqslant d(\mathcal{B}), \tag{11}$$

$$d(\mathcal{A} \pm \mathcal{B}) = d(\mathcal{A}) + d(\mathcal{B}), \tag{12}$$

$$d(\mathcal{A}) = \sup_{\mathcal{A}'_\mathrm{p}, \mathcal{A}_\mathrm{p} \in \mathcal{A}} |\mathcal{A}_\mathrm{p} - \mathcal{A}'_\mathrm{p}|, \tag{13}$$

$$|\mathcal{A}| = \sup_{\mathcal{A}_\mathrm{p} \in \mathcal{A}} |\mathcal{A}_\mathrm{p}|, \tag{14}$$

$$\mathcal{A} \subseteq \mathcal{B} \Rightarrow |\mathcal{A}| \leqslant |\mathcal{B}|, \tag{15}$$

$$|\mathcal{A}| \geqslant \mathcal{O}_\mathrm{p} \quad \text{и} \quad |\mathcal{A}| = \mathcal{O}_\mathrm{p} \Leftrightarrow \mathcal{A} = \mathcal{O}_\mathrm{p},$$
$$|\mathcal{A} + \mathcal{B}| \leqslant |\mathcal{A}| + |\mathcal{B}|,$$
$$|x\mathcal{A}| = |\mathcal{A}x| = |x||\mathcal{A}|, \quad x \in \mathbb{C}, \tag{16}$$
$$\mathcal{A} \in M_{mn}(I(\mathbb{R})) \quad \text{или} \quad \mathcal{A} \in M_{mn}(K(\mathbb{C})),$$
$$|\mathcal{A}\mathcal{B}| \leqslant |\mathcal{A}||\mathcal{B}|,$$

$$d(\mathcal{A}\mathcal{B}) \leqslant d(\mathcal{A})|\mathcal{B}| + |\mathcal{A}|d(\mathcal{B}), \tag{17}$$

$$d(\mathcal{A}\mathcal{B}) \geqslant |\mathcal{A}|d(\mathcal{B}), \ d(\mathcal{A}\mathcal{B}) \geqslant d(\mathcal{A})|\mathcal{B}| \tag{18}$$

$$\begin{cases} d(a\mathscr{B}) = |a|\,d(\mathscr{B}),\ a \in \mathbb{C}, \\ d(\mathscr{A}_p\mathscr{B}) = |\mathscr{A}_p|\,d(\mathscr{B}), \quad d(\mathscr{B}\mathscr{A}_p) = d(\mathscr{B})|\mathscr{A}_p|. \end{cases} \tag{19}$$

Для вещественных интервальных матриц A, B имеем

$$O_p \in \mathscr{A} \Rightarrow |\mathscr{A}| \leqslant d(\mathscr{A}) \leqslant 2|\mathscr{A}|, \tag{20}$$

$$\mathscr{A} = (-1)\mathscr{A} \Rightarrow \mathscr{A}\mathscr{B} = \mathscr{A}|\mathscr{B}|, \tag{21}$$

$$O_p \in \mathscr{A}, \quad 0 \notin \mathscr{B}_{ij} \quad \text{для} \quad \mathscr{B} = (B_{ij}) \Rightarrow d(\mathscr{A}\mathscr{B}) = d(\mathscr{A})|\mathscr{B}|. \tag{22}$$

Доказательство этих свойств проводится покомпонентно с использованием свойств *I*(R) из п.7.2 и свойств *I*(C) из п.7.5.

Мы заметим также, что соотношения (20)—(22) неверны для комплексных интервальных матриц. Рассмотрим, например, (21) для матрицы размерности 1×1 с элементами из *R(C),* т. е. интервал $A = A_1 + iA_2 \in R(\mathbb{C})$. Утверждение $A=(-1)A$ эквивалентно равенствам

$$A_1 = (-1)A_1, \quad A_2 = (-1)A_2$$

для $A = A_1 + iA_2$. Используя (18 п.7.2), мы получаем для $B = -B_1 + iB_2$, что

$$AB = (A_1B_1 - A_2B_2) + i(A_1B_2 + A_2B_1)$$
$$= A_1|B_1| + A_2|B_2| + i(A_1|B_2| + A_2|B_1|).$$

С другой стороны, мы имеем

$$A|B| = A_1(|B_1| + |B_2|) + iA_2(|B_1| + |B_2|).$$

Эти два интервала различны в общем случае, например при B_1=0. Так как соотношения (20)—(22) не потребуются нам для комплексных интервалов, не будем рассматривать случаев, в которых они справедливы.

Теперь введем понятие матрицы расстояний для пары интервальных матриц.

Определение 7. Пусть $\mathscr{A} = (A_{ij})$ и $\mathscr{B} = (B_{ij})$ — интервальные матрицы. Тогда вещественная неотрицательная матрица

$$q(\mathscr{A}, \mathscr{B}) := (q(A_{ij}, B_{ij}))$$

называется матрицей расстояний или расстоянием между матрицами A и B.

Соотношения

$$q(\mathscr{A}, \mathscr{B}) = O_p \Leftrightarrow \mathscr{A} = \mathscr{B},$$
$$q(\mathscr{A}, \mathscr{B}) \leqslant q(\mathscr{A}, \mathscr{C}) + q(\mathscr{B}, \mathscr{C})$$

очевидным образом справедливы для расстояний между интервальными матрицами вместе с соотношениями

$$q(\mathcal{A}+\mathcal{C},\ \mathcal{B}+\mathcal{C}) = q(\mathcal{A},\ \mathcal{B}), \tag{23}$$

$$q(\mathcal{A}+\mathcal{B},\ \mathcal{C}+\mathcal{D}) \leqslant q(\mathcal{A},\ \mathcal{C}) + q(\mathcal{B},\ \mathcal{D}), \tag{24}$$

$$q(\mathcal{A}\mathcal{B},\ \mathcal{A}\mathcal{C}) \leqslant |\mathcal{A}|\, q(\mathcal{B},\ \mathcal{C}). \tag{25}$$

Доказательства последних свойств проводятся покомпонентно с использованием соответствующих свойств $I(\mathrm{R})$ (см. п.7.2) или $I(\mathrm{C})$ (см. п.7.5). С помощью понятия расстояния между интервальными матрицами, введенного в определении 7 можно, используя монотонную норму матриц $\|\cdot\|$, определить метрику на множестве интервальных матриц как $\|q(\mathcal{A},\mathcal{B})\|$. Множество всех интервальных матриц размерности $m \times n$ можно также рассматривать как $m \cdot n$-кратное произведение полного метрического пространства $I(\mathrm{C})$ на себя Известные теоремы топологии показывают, что это произведение снова является полным метрическим пространством. Сходимость в произведении пространств эквивалентна сходимости отдельных компонент. Поэтому справедливы следующие утверждения.

Сходимость последовательности $\{\mathcal{A}^{(k)}\}_{k=0}^{\infty}$ интервальных матриц размерности $m \times n$ к матрице A, т. е. $\lim\limits_{k\to\infty} \mathcal{A}^{(k)} = \mathcal{A}$, эквивалентна (26)

$$\lim_{k\to\infty} A_{ij}^{(k)} = A_{ij}, \quad 1 \leqslant i \leqslant m, \quad 1 \leqslant j \leqslant n.$$

Следствие 8. *Любая последовательность интервальных матриц* $\{\mathcal{A}^{(k)}\}_{k=0}^{\infty}$ *размерности* $m \times n$, *для которой имеет место*

$$\mathcal{A}^{(0)} \supseteq \mathcal{A}^{(1)} \supseteq \mathcal{A}^{(2)} \supseteq \ldots,$$

сходится к интервальной матрице $\mathrm{A} = (A_{ij})$, *где*

$$A_{ij} = \bigcap_{k=0}^{\infty} A_{ij}^{(k)}, \quad 1 \leqslant i \leqslant m, \quad 1 \leqslant j \leqslant n.$$

Это утверждение следует из (26) определения 2 и утверждения об интервалах, аналогичного следствию (8).

Следствие 9. *Операции, введенные в определении* 3, *непрерывны.*

Доказательство получается из того, что непрерывность операций на элементах влечет за собой непрерывность операций в целом. В силу определения 3 и теорем 6 п.7.2, 6 п.7.5 элементы результата операции непрерывно зависят от операндов.

Следующее соотношение справедливо ввиду (21 п.7.2) и (17. п.7.4).

$$\mathcal{X} \subseteq \mathcal{Y} \Rightarrow \frac{1}{2}(d(\mathcal{Y}) - d(\mathcal{X})) \leqslant q(\mathcal{X},\ \mathcal{Y}) \leqslant d(\mathcal{Y}) - d(\mathcal{X}). \tag{27}$$

Теперь введем операцию пересечения на $M_{mn}(I(\mathbb{R}))$ и $M_{mn}(I(\mathbb{C}))$ таким же образом, как это было сделано в п.7.2 для элементов множества *I*(R) и в п.7.5 для элементов множества *R*(C). Так как $M_{mn}(I(\mathbb{R})) \subset M_{mn}(R(\mathbb{C}))$, достаточно определить эту операцию на $M_{mn}(R(\mathbb{C}))$. Пусть

$$\mathscr{A} = (A_{ij}),\ \mathscr{B} = (B_{ij}) \in M_{mn}(R(\mathbb{C})).$$

Тогда определяем пересечение A и B как теоретико-множественное пересечение

$$\mathscr{A} \cap \mathscr{B} = \{\mathscr{C}_p \mid \mathscr{C}_p \in \mathscr{A},\ \mathscr{C}_p \in \mathscr{B}\}.$$

Пересечение двух интервальных матриц A и B принадлежит множеству $M_{mn}(R(\mathbb{C}))$ тогда и только тогда, когда это теоретико-множественное пересечение непусто. В этом случае мы имеем

$$\mathscr{A} \cap \mathscr{B} = (A_{ij} \cap B_{ij}),$$

где $A_{ij} \cap B_{ij}$, $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant m$, строится согласно (23 п.7.2) (соответственно (19 п.7.5)).

Аналогично следствиям 12 п.7.2 и 7 п.7.5 получаем

Следствие 10. *Пусть* $\mathscr{A}, \mathscr{B}, \mathscr{C}, \mathscr{D} \in M_{mn}(R(\mathbb{C}))$. *Тогда имеем*

$\mathscr{A} \subseteq \mathscr{C},\ \mathscr{B} \subseteq \mathscr{D} \Rightarrow \mathscr{A} \cap \mathscr{B} \subseteq \mathscr{C} \cap \mathscr{D}$ *(монотонность включения)*

и пересечение, если оно не выводит за пределы $M_{mn}(R(\mathbb{C}))$, *является непрерывной операцией.*

Как и в случае следствий 12 п.7.2 и 7 п.7.5 , утверждение следует из того, что непрерывность операций на элементах влечет за собой непрерывность операций в целом.

Чтобы обобщить понятие билинейности на операторы из $V_n(\mathbb{C}) \times V_n(\mathbb{C})$ в $V_n(C)$, мы рассмотрим теперь трехмерные массивы с интервальными элементами. Множество всех таких массивов обозначается через $M_{n^3}(I(\mathbb{C}))$. Имеем

$$\mathscr{B} = (B_{ijk}) \in M_{n^3}(I(\mathbb{C})),$$

где

$$B_{ijk} \in I(\mathbb{C}), \quad 1 \leqslant i,\ j,\ \quad k \leqslant n$$

Равенство, включение и сложение определяются поэлементно, т. е. так же, как для интервальных матриц. Аналогично определяются ширина, расстояние и абсолютная величина. Например, определение

$$|\mathscr{B}| := (|\mathscr{B}_{ijk}|)$$

вводит билинейный оператор из $V_n(\mathbb{R}) \times V_n(\mathbb{R})$ в $V_n(\mathbb{R})$. Множество всех билинейных операторов из $V_n(\mathbb{R}) \times V_n(\mathbb{R})$ в $V_n(\mathbb{R})$ обозначается через $M_{n^3}(\text{Ў})$.

Определение 11. Пусть $\mathscr{B} = (B_{ijk}) \in M_{n^3}(I(\mathbb{C}))$. $x = (X_i)$ $y = (Y_i) \in V_n(I(\mathbb{C}))$ и $\mathscr{A} \in M_{nn}(I(\mathbb{C}))$.

Тогда полагаем (a) $\mathscr{C} := \mathscr{B}x \in M_{nn}(I(\mathbb{C}))$, где

$$C_{ij} = \sum_{k=1}^{n} B_{ijk} X_k, \quad 1 \leqslant i, j \leqslant n,$$

(b) $z := (\mathscr{B}x)\, y \in V_n(I(\mathbb{C}))$, где

$$Z_i = \sum_{j=1}^{n} C_{ij} Y_j = \sum_{j=1}^{n} \left(\sum_{k=1}^{n} B_{ijk} X_k \right) Y_j, \quad 1 \leqslant i < n.$$

Вместо $\mathscr{B}(x)\, y$ будем иногда писать $\mathscr{B}xy$.

(c) Далее полагаем

$$\mathscr{C} := \mathscr{A}\mathscr{B} \in M_{n^3}(I(\mathbb{C})),$$

где

$$C_{ijk} = \sum_{\nu=1}^{n} A_{i\nu} B_{\nu jk}, \quad 1 \leqslant i, j, \quad k \leqslant n.$$

Следующая теорема содержит некоторые соотношения, нужные в дальнейшем.

Теорема 12. *Пусть*

$$\mathscr{A}_{\mathrm{p}} = (a_{ij}) \in M_{nn}(\mathbb{C}), \quad \mathscr{B} = (B_{ijk}) \in M_{n^3}(I(\mathbb{C})),$$
$$x = (X_i), \quad y = (Y_i) \in V_n(I(\mathbb{C})).$$

Тогда

(a) $$(\mathscr{A}_{\mathrm{p}}\mathscr{B})\, xy \subseteq \mathscr{A}_{\mathrm{p}}(\mathscr{B}xy).$$

Если $x = -x$, то

(b) $$\mathscr{B}xx = \tfrac{1}{2} |\mathscr{B}|\, d(x)\, x$$

и

(c) $$d(\mathscr{B}xx) = \frac{1}{2} |\mathscr{B}|\, d(x)\, d(x).$$

Доказательство.

(a) Пусть $\mathscr{C} = \mathscr{A}_{\mathrm{p}}\mathscr{B} = (C_{i,k})$ и $z = (Z_i) = \mathscr{B}xy$. Тогда с помощью определения 11 и субдистрибутивности получаем

$$
\begin{aligned}
(\mathcal{A}_p\mathcal{B})\,xy &= \left(\sum_{j=1}^{n}\left(\sum_{k=1}^{n} C_{ijk}X_k\right)Y_j\right) = \left(\sum_{j=1}^{n}\left(\sum_{k=1}^{n}\left(\sum_{\nu=1}^{n} a_{i\nu}B_{\nu jk}\right)X_k\right)Y_j\right)\\
&\subseteq \left(\sum_{j=1}^{n}\left(\sum_{k=1}^{n}\left(\sum_{\nu=1}^{n} a_{i\nu}B_{\nu jk}X_k\right)\right)Y_j\right)\\
&= \left(\sum_{j=1}^{n}\left(\sum_{\nu=1}^{n}\left(\sum_{k=1}^{n} a_{i\nu}B_{\nu jk}X_k\right)\right)Y_j\right)\\
&= \left(\sum_{j=1}^{n}\left(\sum_{\nu=1}^{n} a_{i\nu}\left(\sum_{k=1}^{n} B_{\nu jk}X_k\right)\right)Y_j\right)\\
&\subseteq \left(\sum_{j=1}^{n}\left(\sum_{\nu=1}^{n} a_{i\nu}\left(\sum_{k=1}^{n} B_{\nu jk}X_k\right)Y_j\right)\right)\\
&= \left(\sum_{\nu=1}^{n}\left(\sum_{j=1}^{n} a_{i\nu}\left(\sum_{k=1}^{n} B_{\nu jk}X_k\right)Y_j\right)\right)\\
&= \left(\sum_{\nu=1}^{n} a_{i\nu}\left(\sum_{j=1}^{n}\left(\sum_{k=1}^{n} B_{\nu jk}X_k\right)Y_j\right)\right)\\
&= \left(\sum_{\nu=1}^{n} a_{i\nu}Z_\nu\right) = \mathcal{A}_p(\mathcal{B}xy).
\end{aligned}
$$

(b) Полагая $z = \mathcal{B}xx$ и используя симметричность по x, получаем в обозначениях определения 11 (b), что

$$
\begin{aligned}
Z_i &= \sum_{j=1}^{n}\left(\sum_{k=1}^{n} B_{ijk}X_k\right)X_j = \sum_{j=1}^{n}\left|\sum_{k=1}^{n} B_{ijk}X_k\right|X_j\\
&= \sum_{j=1}^{n}\left|\left(\sum_{k=1}^{n}|B_{ijk}|X_k\right) = \sum_{j=1}^{n}\left(\sum_{k=1}^{n}\frac{1}{2}|B_{ijk}|d(X_k)\right)X_j.\right.
\end{aligned}
$$

Поэтому $\mathcal{B}xx = \frac{1}{2}|\mathcal{B}|d(x)x$.

(c) Используя соотношения из п. (b), получаем

$$
d(Z_i) = \sum_{j=1}^{n}\left(\sum_{k=1}^{n}\frac{1}{2}|B_{ijk}|d(X_k)\right)d(X_j),
$$

т. е.

$$
d(\mathcal{B}xx) = \frac{1}{2}|\mathcal{B}|d(x)d(x).
$$

Модуль 9
Интервальная арифметика для решения систем уравнений

Микромодуль 30
Итерационная локализация неподвижной точки для систем нелинейных уравнений

Рассмотоим теперь функцию $f(x_p)$ от векторной переменной $x_p = (x_1, \ldots, x_n)^T \in V_n(\mathbb{C})$, принимающую значения в С. Будем предполагать, что функция f построена при помощи основных арифметических операций, а также стандартных функций синус, косинус и т. д. Это означает, что ее можно вычислять и как интервальную функцию Предположим еще, что функция f зависит от m параметров $a_1, a_2, \ldots, a_m \in \mathbb{C}$. Таким образом, мы можем записать f в виде

$$f(x_p) = f(x_1, x_2, \ldots, x_n; a_1, a_2, \ldots, a_m).$$

Пусть теперь заданы n функции такого вида

$$f_i(x_p), \quad 1 \leqslant i \leqslant n.$$

Тогда соотношение

$$y_p = \mathcal{f}_p(x_p) = (f_i(x_p))$$

определяет отображение из $V_n(\mathbb{C})$ в $V_n(\mathbb{C})$, а соотношение

$$y = \mathcal{f}_p(x) = (f_i(x))$$

определяет отображение на множестве n-компонентных интервальных векторов

$$\mathcal{f}_p: V_n(I(\mathbb{C})) \to V_n(I(\mathbb{C})), \quad \text{где} \quad \mathcal{f}_p(x) = (f_i(x)).$$

Ниже будет использоваться интервальная арифметика для локализации решений системы уравнений

$$\mathcal{f}_p(x_p) = o_p,$$

где

$$\mathcal{f}_p(x_p) = (f_i(x_p))$$

и

$$f_i(x_p) = f_i(x_1, \ldots, x_n; a_{i1}, \ldots, a_{im_i}) \qquad 1 \leqslant i \leqslant n,$$

в предположении, что параметры a_{ij} независимо изменяются в некоторых комплексных интервалах.

Много возможностей дня решения этой задачи дает метод итераций. Заметим, что данное уравнение всегда может быть преобразовано к виду

$$x_p = f_p(x_p).$$

Вычисление правой части этого уравнения для произвольного интервального вектора $x^{(0)}$ дает интервальный вектор $x^{(1)}$. Продолжая в том же духе, получаем метод итераций

$$x^{(k+1)} = f_p(x^{(k)}), \qquad k \geq 0.$$

Возникают следующие вопросы, (a) Когда существует последовательность $\{x^{(k)}\}_{k=0}^{\infty}$? (b) Когда эта последоватетьность сходится? (c) Когда предел x^* единствен? (d) Какое отношение имеет предел x^* к решению сформулированной выше задачи?

Сначала докажем теорему о неподвижной точке, опирающуюся на моноточность интервального оценивания функций относительно включения.

Теорема 1. *Пусть дано отображение*

$$f_p: V_n(\mathbb{C}) \to V_n(\mathbb{C}), \quad \text{где} \quad f_p(x_p) = (f_i(x_p)),$$

причем функции $f_i(x_p)$ *имеют указанный выше вид. Рассмотрим метод итераций в* $V_n(I(\mathbb{C}))$, *заданный соотношением*

$$x^{(k+1)} = f_p(x^{(k)}), \qquad k \geq 0,$$

и удовлетворяющий условию

$$x^{(1)} \subseteq x^{(0)}.$$

Тогда имеет место следующее.

(1) *Последовательность результатов итерации* $\{x^{(k)}\}_{k=0}^{\infty}$ *сходится к пределу x, такому что* $x = f_p(x)$.

(2) *Любой вектор удовлетворяющий* $x_p \in x^{(0)}$, *уравнению* $x_p = f_p(x_p)$, *содержится в x, т. е имеет место*

$$\{x_p \mid x_p \in x^{(0)}, x_p = f_p(x)\} \subseteq x.$$

Доказательство. (1) По предположению имеем $x^{(1)} \subseteq x^{(0)}$. Так как интервальные вычисления монотонны относительно включения, мы получаем

$$x^{(2)} = f_p(x^{(1)}) \subseteq f_p(x^{(0)}) = x^{(1)} \subseteq x^{(0)}.$$

С помощью математической индукции можно показать, что

$$\ldots \subseteq x^{(3)} \subseteq x^{(2)} \subseteq x^{(1)} \subseteq x^{(0)}.$$

Из следствия 8 микромодуля 29 вытекает, что эта последовательность сходится к некоторому элементу $x \in V_n(I(\mathbb{C}))$. Из непрерывности интервальных оценок следует, что

$$x = \lim_{k\to\infty} x^{(k)} = \lim_{k\to\infty} f_p(x^{(k)}) = f_p(x).$$

(2). Пусть $x_p \in x^{(0)}$ и $x_p = f_p(x_p)$.

Из монотонности интервальных операций относительно включения снова следует, что

$$x_p = f_p(x_p) \in f_p(x^{(0)}) = x^{(1)},$$

и по индукции получаем

$$x \in x^{(k)}, \qquad k \geqslant 0,$$

откуда следует, что $x_p \in x$.

Условия теоремы 1 гарантируют существование неподвижной точки, но не ее единственность. Это показывает следующий пример.

Пример. Рассмотрим уравнение

$$X = X \cdot X \cdot X$$

в $R(\mathbb{C})$.Очевидно, что этому уравнению удовлетворяют следующие элементы $R(\mathbb{C})$:

$$X = [-1, 1], [1, 1], [-1, -1], [0, 1], [-1, 0], [0, 0], i[-1, 1].$$

Докажем теперь другую теорему о неподвижной точке, имеющую несколько иные условия и использующую несколько иной итерационный процесс.

Теорема 2. *Пусть задано отображение*

$$f_p:\ V_n(\mathbb{C}) \to V_n(\mathbb{C}), \quad \text{где} \quad f_p(x_p) = (f_i(x_p))$$

с функциями $f_i(x_p)$ *указанного выше вида.*

Рассмотрим итерационный процесс

$$x^{(k+1)} = f_p(x^{(k)}) \cap x^{(k)}, \qquad k \geqslant 0,$$

в $V_n(R(\mathbb{C}))$. *Предположим, что существует элемент* $\tilde{x}_p \in x^{(0)}$, *удовлетворяющий уравнению* $\tilde{x}_p = f_p(\tilde{x}_p)$. *Тогда верно следующее.*

(3) *Последовательные приближения* $\{x^{(k)}\}_{k=0}^{\infty}$ *удовлетворяют условию* $\lim_{k\to\infty} x^{(k)} = x$, *причем* $x = f_p(x) \cap x$.

(4) *Любой вектор* $x_p \in x^{(0)}$, *удовлетворяющий уравнению* $x_p = f_p(x_p)$, *содержитсяв x, т, е.*

$$\{x_p \mid x_p \in x^{(0)},\ x_p = f_p(x_p)\} \subseteq x.$$

Доказательство. Рассуждением, аналогичным доказательству теоремы 1, из $\tilde{x}_p \in x^{(0)}$ мы получаем соотношени $x_p \in f_p(x^{(0)}) \cap x^{(0)} = x^{(1)}$. Применяя индукцию, получаем $\tilde{x}_p \in x^{(k)}$ $k \geq 0$. Так как пересечение здесь всегда непусто, мы получаем последовательность интервалов

$$x^{(0)} \supseteq x^{(1)} \supseteq \ldots,$$

которая в силу следствия 8 микромодуль 29 сходится к некоторому пределу x. Ввиду непрерывности интервальных вычислений и пересечений мы имеем $x = f_p(x) \cap x$ для этого предела, а также $\hat{x}_p \in x$ для всех $\hat{x}_p = f_p(\hat{x}_p) \in x^{(0)}$.

По поводу единственности неподвижной точки в теореме 2 можно сказать то же самое, что и в случае теоремы 1. Приведем теперь две теоремы о неподвижной точке, которые будут применены в дальнейшем к двум конкретным итерационным процедурам. В отличие от теорем 1 и 2, имеющих довольно общий характер, эти новые теоремы обеспечивают единственность неподвижной точки. Сначала введем одно понятие.

Определение 3. Пусть

$$f_p\colon V_n(\mathbb{C}) \to V_n(\mathbb{C}), \qquad f_p(x_p) = (f_i(x_p)),$$

— отображение указанного выше вида. f_p называется $\mathcal{P}_p$-сжатием, если существует неотрицательная матрица $\mathcal{P}_p$, такая что

$$q(f_p(x), f_p(y)) \leqslant \mathcal{P}_p q(x, y) \quad \text{для всех} \quad x, y \in V_n(I(\mathbb{C})),$$

где

$$\rho(\mathcal{P}_p) < 1.$$

Здесь ρ обозначает спектральный радиус матрицы $\mathcal{P}_p$, а q обозначает расстояние между двумя интервальными векторами, определенное в микромодуле 29. Докажем следующее утверждение.

Теорема 4. *Если* $f_p\colon V_n(\mathbb{C}) \to V_n(\mathbb{C})$ *есть* $\mathcal{P}_p$*-сжатие, то уравнение* $x = f_p(x)$ *имеет единственную неподвижную точку* $x^* \in V_n(I(\mathbb{C}))$. *При этом для любого* $x^{(0)} \in V_n(I(\mathbb{C}))$ *итерации сходятся к* x^*.

Доказательство Из того, что $\rho(\mathcal{P}_p) < 1$ и $\mathcal{P}_p \geqslant O_p$, следует существование матрицы $(\mathcal{I}_p - \mathcal{P}_p)^{-1}$ и соотношение

$$(\mathcal{I}_p - \mathcal{P}_p)^{-1} = \sum_{i=1}^{\infty} \mathcal{P}_i^i \geqslant \sum_{i=0}^{m-1} \mathcal{P}_p^i \geqslant O_p.$$

Тогда мы получаем для любого k и $m \geq 1$

$$q(x^{(k+m)}, x^{(k)}) \leqslant \sum_{j=0}^{m-1} \mathscr{P}_p^j q(x^{(k+1)}, x^{(k)}) \leqslant (\mathscr{I}_p - \mathscr{P}_p)^{-1} \mathscr{P}^k q(x^1, x^{(0)}).$$

Так как $\lim\limits_{k\to\infty} \mathscr{P}_p^k = \mathscr{O}_p$, каждая компонента последовательности $\{x^{(k)}\}_{k=0}^{\infty}$, а значит, и сама эта последовательность удовлетворяют условию сходимости Коши. Так как пространство $V_n(I(\mathbb{C}))$ полно, а отображение $\mathscr{P}_p$ является сжатием и потому непрерывно, мы получаем

$$\lim_{k\to\infty} x^{(k)} = x^* \quad \text{и} \quad x^* = f_p(x^*).$$

Единственность неподвижной точки следует из соотношений

$$q(x^*, y^*) = q(f_p(x^*), f_p(y^*)) \leqslant \mathscr{P}_p q(x^*, y^*)$$

и $(\mathscr{I}_p - \mathscr{P}_p)^{-1} \geqslant \mathscr{O}_p$.

Эта теорема — частный случай более общего результата, доказанного Шредером.

Вот еще одна теорема о неподвижной точке, которая будет использована в дальнейшем.

Теорема 5. *Пусть*

$$f_p: V_n(\mathbb{C}) \to V_n(\mathbb{C})$$

и

$$g_p: V_n(\mathbb{C}) \times V_n(\mathbb{C}) \to V_n(\mathbb{C}),$$

где f_p *и* g_p *имеют описанный выше вид, причем*

$$g_p(x, x) = f_p(x) \quad \textit{для всех} \quad x \in V_n(I(\mathbb{C})),$$

$$q(g_p(x, z), g_p(y, z)) \leqslant \mathscr{Q}_p q(x, y),$$

$$q(g_p(z, x), g_p(z, y)) \leqslant \mathscr{R}_p q(x, y) \quad \textit{для всех} \quad x, y, z \in V_n(I(\mathbb{C}))$$

и

$$\mathscr{Q}_p \geqslant \mathscr{O}_p, \quad \mathscr{R}_p \geqslant \mathscr{O}_p, \quad \rho(\mathscr{Q}_p) < 1, \quad \rho((\mathscr{I}_p - \mathscr{Q}_p)^{-1} \mathscr{R}_p) < 1$$

Тогда уравнение $f_p(x) = x$ *имеет единственную неподвижную точку* $x^* \in V_n(I(\mathbb{C}))$, *и для любого* $x^{(0)}$ *существует единственная последовательность* $\{x^{(k)}\}_{k=0}^{\infty}$, *удовлетворяющая уравнению*

$$x^{(k+1)} = g_p(x^{k+1}, x^k), \quad k \geqslant 0$$

При этом

$$\lim_{k\to\infty} x^{(k)} = x^*.$$

Доказательство. Матрицу $(\mathscr{I} - \mathscr{Q})^{-1}\mathscr{R}_p$ можно рассматри-

вать как итерационную матрицу, соответствующую матрице $\mathscr{T}_p - \mathscr{Q}_p = \mathscr{R}_p$. Из $\mathscr{P}_p = \mathscr{Q}_p + \mathscr{R}_p \geqslant O_p$ и того, что $\rho(\mathscr{P}_p) < 1$, следует $\rho((\mathscr{T}_p - \mathscr{Q}_p)^{-1}\mathscr{R}_p) < 1$ Отсюда с помощью (23 микромодуль 29) и (10 микромодуль 29) получаем для произвольных $x, y \in V_n(I_n(\mathbb{C}))$, что

$$q(f_p(x), f_p(y)) \leqslant q(g_p(x, x), g_p(x, y)) + q(g_p(x, y), g_p(y, y)) \leqslant (\mathscr{R}_p + \mathscr{Q}_p) q(x, y)$$

т. е. что f_P есть P_p-сжатие. В силу предыдущей теоремы уравнение $f_p(x) = x$ имеет единственную неподвижную точку $x^* \in V_n(I_n(\mathbb{C}))$. Из наших предположений следует, что отображение $g(\cdot, z)$ для фиксированного $z \in V_n(I_n(\mathbb{C}))$ является Q_p-сжатием. Применяя предыдущую теорему, получаем, что $g(\cdot, x^{(k)})$ имеет единственную неподвижную точку $x^{(k+1)}$. Таким образом, существование последовательности $\{x^{(k)}\}_{k=0}^{\infty}$ установлено для произвольного $x^{(0)} \in V_n(I(\mathbb{C}))$. Из того, что

$$q(x^{(k+1)}, x^*) \leqslant q(g_p(x^{(k+1)}, x^{(k)}), g(x^*, x^{(k)})) + q(g(x^*, x^{(k)}), g(x^*, x^*)) \leqslant \mathscr{Q}_p q(x^{(k+1)}, x^*) + \mathscr{R}_p q(x^{(k)}, x^*)$$

или

$$q(x^{(k+1)}, x^*) \leqslant (\mathscr{T}_p - \mathscr{Q}_p)^{-1} \mathscr{R}_p q(x^{(k)}, x^*) \leqslant \left((\mathscr{T}_p - \mathscr{Q}_p)^{-1} \mathscr{R}\right)^{k+1} q(x^{(0)}, x^*),$$

следует, что

$$\lim_{k \to \infty} x^{(k)} = x^*,$$

так как $\rho\left((\mathscr{T}_p - \mathscr{Q}_p)^{-1} \mathscr{R}_p\right) < 1$.

Это завершает доказательство теоремы.

Результат из теоремы 5 был сначала получен для отображений из $V_n(\mathbb{R})$ в $V_n(\mathbb{R})$.

В связи с двумя последними теоремами продемонстрируем соотношение между единственной неподвижной точкой и потенциальными решениями уравнения

$$x_p = f_p(x_p), \qquad x_p \in V_n(\mathbb{C}).$$

Следствие 6. *Пусть задано отображение*

$$f: V_n(\mathbb{C}) \to V_n(\mathbb{C}), \qquad f_p(x_p) = (f_i(x_p)),$$

причем

$$f_i(x) = f_i(X_1, X_2, \ldots, X_n; A_{i1}, \ldots, A_{im_i}), \quad 1 \leqslant i \leqslant n.$$

Пусть выполнены условия одной из теорем 4, 5, *и пусть* x^* — *единственная неподвижная точка уравнения* $x = f_p(x)$, *существование которой доказано там. Тогда*

$$\{x_p \mid x_p = f_p(x_p), \quad a_{ij} \in A_{ij}, \quad 1 \leqslant i \leqslant n, \quad 1 \leqslant j \leqslant m_i\} \subseteq x^*$$

Доказательство. Рассмотрим уравнение

$$x_p = f_p(x_p), \qquad x_p \in V_n(\mathbb{C}),$$

при фиксированном выборе элементов $a_{ij} \in A_{ij}$, $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant m_i$, и допустим, что ему удовлетворяет элемент $x_p^* \in V_n(\mathbb{C})$. Мы можем тогда начать итерации в теореме 4 или 5 со значения $x^{(0)} = x_p^*$ и в пределе получим неподвижную точку x^*. Так же, как это было сделано в доказательстве п. (2) теоремы 1, мы можем использовать монотонность включения, чтобы показать, что всегда имеет место

$$x_p^* \in x^{(k)}, \qquad k \geqslant 0.$$

Отсюда следует $x_p^* \in x^*$.

Теперь рассмотрим практическое нахождение констант Липшица для интервальных вычислений. Мы увидим, что константы Липшица для интервальных вычислений будут мажорировать константы Липшица для соответствующих точечных функций. Это означает, что каждая из систем $x_p = f(x_p)$, рассмотренных в следствии 6, удовлетворяет условиям теорем 4 и 5 при ограничении на множество $V_n(\mathbb{C})$, а потому имеет единственное решение x_p^*.

С помощью найденных констант Липшица мы сможем проверить выполнены ли в конкретных условиях предположения теоремы 4 или 5. Для простоты мы ограничимся пространством $V_n(I(\mathbb{R}))$ Аналогичные формулы для $V_n(I(\mathbb{C}))$ могут быть получены без труда. Для подготовки докажем одно свойство метрики q, которое следует из того, что она является метрикой Хаусдорфа.

Лемма 7. *Пусть* $Y, Z \in I(\mathbb{R})$ *и* $\alpha \geqslant 0$. *Тогда*

$$q(Y, Z) \leqslant \alpha \Leftrightarrow \{\text{для любого } y \in Y \text{ существует } z \in Z \text{ со свойством } |y - z| \leqslant \alpha, \text{ и для любого } z \in Z \text{ существует } y \in Y \text{ со свойством } |z - y| \leqslant \alpha\}$$

Доказательство. Если $Y = [y_1, y_2]$ и $Z = [z_1, z_2]$, то $q(Y, Z) = \max\{|y_1 - z_1|, |y_2 - z_2|\}$. Докажем сначала импликацию $\Rightarrow$. Пусть $q(Y, Z) \leqslant \alpha$ и зафиксирован $y \in Y$. Если $z \notin Z$, то можно

взять $z=y$, что дает первую половину правой части $\Rightarrow$. Если же $y \notin Z$, то при $z_1 > y$ мы получаем для $z = z_1$, что

$$\alpha \geqslant |z_1 - y_1| = z_1 - y_1 \geqslant z_1 - y = |z_1 - y|,$$

а в случае $z_2 < y$ получаем для $z = z_2$, что

$$\alpha \geqslant |z_2 - y_2| = y_2 - z_2 \geqslant y - z_2 = |y - z_2|.$$

Так как в этом рассуждении y и z равноправны, оно дает и вторую половину правой части $\Rightarrow$ для фиксированного $z \in Z$,

Докажем обратную импликацию $\Leftarrow$. Пусть сначала $y_1 \leqslant z_1$. Зафиксируем $y = y_1$. Тогда найдется $z \geq z_1$, такое что

$$\alpha \geqslant |z - y_1| = z - y_1 \geqslant z_1 - y_1 = |z_1 - y_1|.$$

Если же $z_1 < y_1$, то зафиксируем $z = z_1$. Тогда найдется $y \geqslant y_1$, для которого

$$\alpha \geqslant |y - z_1| = y - z_1 \geqslant y - z_1 = |y_1 - z_1|.$$

Таким образом, всегда получается $|y_1 - z_1| \leqslant \alpha$. Соотношение $|y_2 - z_2| \leqslant \alpha$ доказывается тем же методом. Это и дает

$$q(Y, Z) \leqslant \alpha.$$

Теперь используем эту лемму, чтобы получить утверждение о константах Липшица для интервальных вычислений

Теорема 8. *Пусть f — вещественная функция вещественной переменной x. Из выражения $f(x; a_1, \ldots, a_m)$, принадлежащего функции f, построим выражение $f(x_1, x_2, \ldots, x_n; a_1, \ldots, a_m)$, заменяя каждое вхождение переменной x на новую переменную x_i, $1 \leqslant i \leqslant n$.*

Допустим, что это новое выражение удовлетворяет условию Липшица

$$|\tilde{f}(x_1, \ldots, y_i, \ldots, x_n; a_1, \ldots, a_m) - \tilde{f}(x_1, \ldots, z_i, \ldots, x_n; a_1, \ldots, a_m)| \leqslant l_i |y_i - z_i|$$

при каждом i, $1 \leqslant i \leqslant n$ и фиксированных

$$x_k \in X, \quad 1 \leqslant k \leqslant n, \quad i \neq k \quad и \quad a_k \in A, \quad 1 \leqslant k \leqslant m.$$

Если существует интервальная оценка $f(X; A_1, \ldots, A_m)$, то для любых интервалов $Y \subseteq X$ и $Z \subseteq X$ выполнено следующее условие Липшица:

$$q(f(Y; A_1, \ldots, A_m), f(Z; A_1, \ldots, A_m)) \leqslant \left(\sum_{i=1}^{n} l_i\right) q(Y, Z). \quad (5)$$

Доказательство. Из способа построения выражения $\tilde{f}(x_1, \ldots, x_n; a_1, \ldots, a_m)$ следует, что

$$f(X; A_1, \ldots, A_m) = \tilde{f}(X, \ldots, X; A_1, \ldots, A_m) = W(\tilde{f}, X, \ldots, X; A_1, \ldots, A_m).$$

Если произвольным образом выбрано

$$u \in f(Y; A_1, \ldots, A_m),$$

то мы имеем

$$u = \tilde{f}(y_1, \ldots, y_n; a_1, \ldots, a_m) \quad \text{для} \quad y_i \in Y, \quad 1 \leqslant i \leqslant n, \quad a_k \in A_k, \quad 1 \leqslant k \leqslant m.$$

Из леммы 7 следует, что для произвольного y_i существует $z_i \in Z$, для которого

$$|y_i - z_i| \leqslant q(Y, Z).$$

Рассматривая значение функции

$$\tilde{f}(z_1, \ldots, z_n; a_1, \ldots, a_m) \in f(Z; A_1, \ldots, A_m),$$

получаем с помощью сделанных предположений и многократного использования неравенства треугольника, что

$$\begin{aligned} |\tilde{f}(y_1, \ldots, y_n; a_1, \ldots, a_m) - \tilde{f}(z_1, \ldots, z_n; a_1, \ldots, a_m)| \\ \leqslant |\tilde{f}(y_1, \ldots, y_n; a_1, \ldots, a_m) \\ - \tilde{f}(z_1, y_2, \ldots, y_n; a_1, \ldots, a_m)| + \ldots \\ + |\tilde{f}(z_1, \ldots, z_{n-1}, y_n; a_1, \ldots, a_m) \\ - \tilde{f}(z_1, \ldots, z_n; a_1, \ldots, a_m)| \\ \leqslant \sum_{i=1}^{n} l_i |y_i - z_i| \leqslant \left(\sum_{i=1}^{n} l_i \right) q(Y, Z) = \alpha. \end{aligned}$$

Аналогичное неравенство можно установить для произвольного $v \in f(Z, A_1, \ldots, A_m)$. Применяя лемму (7), получаем неравенство (5).

Условия теоремы 8 на практике почти всегда выполнены. Рассматриваемые здесь функциональные выражения составлены из основных арифметических операций и стандартных функций. Поэтому они почти всегда дифференцируемы на своей области определения.

В качестве приложения доказанной теоремы приведем несколько конкретных примеров.

Примеры. (a) $f(x;\ a) = ax$. Интервальная оценка этой функции для $A \in I(\mathbb{R})$ и произвольных $Y,\ Z \in I(\mathbb{R})$ удовлетворяет в силу (5) неравенству

$$q(AY,\ AZ) \leqslant |A|\, q(Y,\ Z).$$

(b) $$f(x;\ a_0,\ \ldots,\ a_n) = \sum_{k=0}^{n} a_k x^k.$$

В силу (5) интервальная оценка этой функции удовлетворяет для $A_k \in I(\mathbb{R})$, $0 \leqslant k \leqslant n$, и произвольных $Y,\ Z \in I(\mathbb{R})$ неравенству

$$q\left(\sum_{k=0}^{n} A_k Y^k,\ \sum_{k=0}^{n} A_k Z^k\right) \leqslant \left(\sum_{k=1}^{n} k\,|A_k|\,|X|^{k-1}\right) q(Y,\ Z).$$

Здесь X— наименьший интервал, содержащий $Y \cup Z$. Полученное неравенство верно независимо от того, как вычисляется X^k — как произведение $X \cdot \ldots \cdot X$ или как одноместная операция согласно определению 3 п.7.1.

Читателю предоставляется в качестве простого упражнения получить аналогичную формулу для рациональных выражений.

(c) $$f(x;\ a) = x/2 - 2e^{ax}.$$

В силу (5) интервальная оценка этой функции удовлетворяет для $A \in I(\mathbb{R})$ и произвольных $Y,\ Z \in I(\mathbb{R})$ неравенству

$$q(Y/2 - 2e^{AY},\ Z/2 - 2e^{AZ}) \leqslant \left(\frac{1}{2} + 2|A|e^{|A||X|}\right) q(Y,\ Z).$$

Здесь X снова обозначает наименьший интервал, содержащий $Y \cup Z$.

Теорема 8 без труда переносится на функции нескольких переменных. Таким образом, мы получаем формулы для интервальных вычислений отображений f_{P}, описанных в начале этого микромодуля. Нужно только применить эти неравенства покомпонентно. Это без труда следует из уже доказанных результатов, и мы опускаем подробности. Такие формулы в свою очередь позволяют вычислять практически матрицу P_p, участвующую в определении 3. Приведем простой пример.

Пример. Дано n функций

$$f_i(x_1,\ \ldots,\ x_n;\ a_{i1},\ \ldots,\ a_{in}) = \sum_{\nu=1}^{n} \sin(a_{i\nu} x_\nu), \quad 1 \leqslant i \leqslant n.$$

Интервальная оценка существует для произвольных интервалов $X_\nu \in I(\mathbb{R})$, $1 \leqslant \nu \leqslant n$ и $A_{ij} \in I(\mathbb{R})$, $1 \leqslant i,\ j \leqslant n$. Кроме того, каждая из данных функций удовлетворяет условиям теоремы 8 для данных интервалов A_{ij}, $1 \leqslant i,\ j \leqslant n$, а именно

$$|f_i(x_1, \ldots, x_\nu, \ldots, x_n; a_{i1}, \ldots, a_{in}) - f_i(x_1, \ldots, y_\nu, \ldots, x_n; a_{i1}, \ldots, a_{in})| \leqslant |A_{i\nu}||x_\nu - y_\nu|, \quad x_j \in X_j, \quad 1 \leqslant j \leqslant n, \quad j \neq \nu, \quad x_\nu, \quad y_\nu \in X_\nu.$$

Если мы теперь применим теорему 8 к каждой компоненте, то получим

$$\mathcal{f}_\mathrm{p}(x) = (f_i(x)) = \left(\sum_{\nu=1}^{n} \sin(A_{i\nu}X_\nu)\right),$$

$$q(\mathcal{f}_\mathrm{p}(x), \mathcal{f}(y)) \leqslant \mathcal{P}_\mathrm{p} q(x, y), \quad \text{где} \quad \mathcal{P}_\mathrm{p} = (|A_{if}|).$$

Если теперь $\rho(\mathcal{P}_\mathrm{p}) < 1$, то, согласно теореме 4, уравнение

$$x = \mathcal{f}_\mathrm{p}(x) = (f_i(x))$$

имеет единственную неподвижную точку, которая может быть найдена методом итераций. Применяя теорему 5, можно показать, что короткошаговый метод также сходится.

В теореме 4 мы предположим для простоты, что отображение f_P определено на всем пространстве $V_n(\mathbb{C})$, что f_P можно интервально оценить для всех элементов множества $V_n(I(\mathbb{C}))$ и что f_P есть P_p -сжатие. Простые примеры [в частности, пример (b)] показывают, однако, что матрица P_p в определении 3 может зависеть от x и y и что условие $\rho(\mathcal{P}_\mathrm{p}) < 1$ может нарушаться для некоторых x, y. В этом случае может помочь следующее утверждение.

Теорема 9. *Пусть* $\mathcal{f}_\mathrm{p}$: $\theta \subseteq V_n(\mathbb{C}) \to V_n(\mathbb{C})$ *есть* P_p *-сжатие для всех* x, y *из замкнутого множества* $I(\theta) = \{z \in V_n(I(\mathbb{C})) \mid z \subseteq \theta\}$. *Если* $\mathcal{f}_\mathrm{p}(x) \in I(\theta)$ *для всех* $x \in I(\theta)$, *то уравнение* $x = \mathcal{f}_\mathrm{p}(x)$ *имеет единственную неподвижную точку* $x^* \in f(\theta)$, *причем, итерации* $x^{(k+1)} = \mathcal{f}_\mathrm{p}(x^{(k)})$, $k \geqslant 0$, *сходятся к этой неподвижной точке* x^* при любом начальном векторе $x^{(0)} \in I(\theta)$.

Доказательство можно провести аналогично доказательству теоремы 4.

Заметим, что из неравенства

$$q(x^{(k+m)}, x^{(k)}) \leqslant \sum_{j=1}^{m} \mathcal{P}_\mathrm{p}^j q(x^{(k)}, x^{(k-1)}) \leqslant (\mathcal{J}_\mathrm{p} - \mathcal{P}_\mathrm{p})^{-1} \mathcal{P}_\mathrm{p} q(x^{(k)}, x^{(k-1)}),$$

которое следует из теоремы 4, мы получаем оценку погрешности

$$q(x^{(k)}, x^*) \leqslant (\mathcal{J}_\mathrm{p} - \mathcal{P}_\mathrm{p})^{-1} \mathcal{P}_\mathrm{p} q(x^{(k)}, x^{(k-1)})$$

для теорем 4 и 9, переходя к пределу при $m \to \infty$ в предыдущем неравенстве.

Сформулируем еще одну лемму, которая будет использована позднее.

Лемма 10. *Пусть* f_p: $x \subseteq \theta \subseteq V_n(\mathbb{R}) \to V_n(\mathbb{R})$ *—непрерывное отображение.*

Пусть отображение $\mathcal{Y}_p$: $x \subseteq \theta \subseteq V_n(\mathbb{R}) \to V_n(\mathbb{R})$ *также непрерывно и для всех* $x_p \in x$ *существует обращение* $\mathcal{Y}_p(x_p)^{-1}$. *Если* $f_p(x_p) \in x$ *при всех* $x_p \in x$ *верно для отображения* f_p: $x \to V_n(\mathbb{R})$, *определенного формулой*

$$f_p(x_p) := x_p - \mathcal{Y}_p(x_p) f_p(x_p),$$

то f_p *имеет нуль в интервале* x.

Простое доказательство этого утверждения использует теорему Брауэра о неподвижной точке. Действительно, f_p отображает выпуклое компактное множество $\{x_p \mid x_p \in x\} \subset V_n(\mathbb{R})$ в себя и поэтому имеет неподвижную точку. Так как $\mathcal{Y}_p(x_p)$ несингулярно, отображение f_p имеет нуль в x.

Замечания. Лемма 10 сформулирована и доказана Алефельдом. Она обобщает утверждение для постоянного $\mathcal{Y}_p$, доказанное Муром.

Эта лемма важна, так как мы можем проверить выполнение ее условий с помощью конечного числа арифметических операций над интервальным вектором, найдя мажоранту для отображения $\mathcal{Y}_p(x_p)$ в виде интервального выражения, зависящего от интервального вектора x.

Подробности такой проверки описаны в последующих микромодулях,.

Прямая проверка условий теоремы Брауэра о неподвижной точке представляет собой, напротив, очень трудную задачу.

Микромодуль 31

Системы линейных уравнений, поддающиеся методу итерации

Мы предполагаем, что рассматриваемая здесь система линейных уравнений уже имеет вид

$$x_p = \mathcal{A}_p x_p + b_p, \tag{1}$$

где $\mathcal{A}_p = (a_{ij})$ и $b_p = (b_i)$.

Пусть известно, что элементы a_{ij} матрицы $\mathcal{A}_p$ лежат в интервалах A_{ij}, а компоненты b_i вектора b_p лежат в интервалах B_i. Нас интересует множество решений, получающихся, когда входные данные

изменяются в данных интервалах. Поэтому мы вводим интервальную матрицу $\mathcal{A} = (A_{ij})$ размерности $n \times n$, содержащую интервальные коэффициенты системы, и интервальный вектор $\mathcal{b} = (B_i)$, содержащий интервальные правые части. Рассмотрим теперь отображение

$$f_p:\ V_n(\mathbb{C}) \to V_n(\mathbb{C}),$$

определяемое формулой

$$f_p(x_p) = \mathcal{A}_p x_p + \mathcal{b}_p.$$

Прежде всего мы имеем следующее утверждение.

Теорема 1. *Метод последовательных приближений*

$$(x)^{(k+1)} = \mathcal{A}x^{(k)} + \mathcal{b}, \qquad k \geqslant 0, \tag{2}$$

сходится к единственной неподвижной точке x^* *уравнения*

$$x = \mathcal{A}x + \mathcal{b}$$

Для любого $x^{(0)} \in V_n(I(\mathbb{C}))$ *тогда и только тогда, когда*

$$\rho(|\mathcal{A}|) < 1.$$

Доказательство. Покажем, что f_P есть $|\mathcal{A}|$-сжатие. Пусть $x,\ y \in V_n(I(\mathbb{C}))$. Применяя (23, микромодуль 29) и (25, микромодуль 29), получаем

$$q(f_p(x),\ f_p(y)) = q(\mathcal{A}x + \mathcal{b},\ \mathcal{A}y + \mathcal{b}) \leqslant |\mathcal{A}|\, q(x,\ y).$$

Из теоремы 4 микромодуля 30 следует, что условие $\rho(|\mathcal{A}|) < 1$ достаточно для сходимости метода и единственности неподвижной точки. Для доказательства обратного утверждения допустим, что последовательные приближения $x^{(k+1)} = \mathcal{A}x^{(k)} + \mathcal{b},\ k \geqslant 0$, сходятся для каждого $x^{(0)} \in V_n(I(\mathbb{C}))$ к неподвижной точке x^*. Мы должны показать, что $\rho(|\mathcal{A}|) < 1$. Из теоремы Перрона и Фробениуса следует, что вещественная неотрицательная матрица $|\mathcal{A}|$ имеет неотрицательный собственный вектор, соответствующий собственному числу $\lambda = \rho(|\mathcal{A}|)$. Из сходимости приближений $x^{(k+1)} = \mathcal{A}x^{(k)} + \mathcal{b},\ k \geqslant 0$, к x^* при любом начальном векторе $x^{(0)}$ следует, что последовательность $\{d(x^{(k)})\}_{k=0}^{\infty}$ сходится к $d(x^*)$. Выберем теперь $x^{(0)}$ таким образом, чтобы $d(x^{(0)})$ был собственным вектором, соответствующим собственному числу $\lambda = \rho(|\mathcal{A}|)$ матрицы $|\mathcal{A}|$, причем хотя бы одна компонента вектора $d(x^{(0)})$ была строго больше, чем соответствующая компонента

вектора $d(x^*)$. Тогда из (2) следует в силу (12, микромодуль 29) и (18, микромодуль 29), что в предположении $\lambda = \rho(|\mathscr{A}|) \geqslant 1$ верно

$$d(x^{(1)}) = d(\mathscr{A}x^{(0)} + \mathscr{b}) = d(\mathscr{A}x^{(0)}) + d(\mathscr{b})$$
$$\geqslant d(\mathscr{A}x^{(0)}) \geqslant |\mathscr{A}|\, d(x^{(0)}) = \lambda d(x^{(0)}) \geqslant d(x^{(0)}),$$
$$d(x^{(2)}) \geqslant |\mathscr{A}|\, d(x^{(1)}) \geqslant \lambda |\mathscr{A}|\, d(x^{(0)}) = \lambda^2 d(x^{(0)}) \geqslant d(x^{(0)}).$$

Для произвольного k получаем

$$d(x^{(k+1)}) \geqslant |\mathscr{A}|\, d(x^{(k)}) \geqslant \ldots \geqslant \lambda^{(k+1)} d(x^{(0)}) \geqslant d(x^{(0)}).$$

Переходя к пределу при $k \to \infty$, получаем

$$d(x^*) \geqslant d(x^{(0)}),$$

что противоречит выбору $x^{(0)}$. Поэтому верно $\rho(|\mathscr{A}|) < 1$.

Установим связь этой теоремы с задачей, сформулированной во введении к этому микромодулю. (См. также следствие 6, микромодуль 30)

Теорема 2. *Пусть $\mathscr{A}$ — интервальная матрица, такая что $\rho(|\mathscr{A}|) < 1$. Тогда для неподвижной точки x^* уравнения $x^* = \mathscr{A}x^* + \mathscr{b}$ (которая существует и единственна в силу теоремы* 1) *верно соотношение*

$$\{y_p = (\mathscr{I}_p - \mathscr{A}_p)^{-1} \mathscr{b}_p \mid \mathscr{A}_p \in \mathscr{A},\ \mathscr{b}_p \in \mathscr{b}\} \subseteq \{x_p \mid x_p \in x^*\}.$$

Если $\mathscr{A} = (A_{ij}) \in M_{nn}(I(\mathbb{R}))$, $\mathscr{b} \in V_n(I(\mathbb{R}))$ и неравенство $i(A_{ij}) \geqslant 0$ справедливо для $A_{ij} = [i(A_{ij}), s(A_{ij})]$, то x^* оптимальна в следующем смысле. Не существует интервального вектора $x \in V_n(I(\mathbb{R}))$ такого, что $x \subseteq x^*$, $x \neq x^*$, но

$$\{y_p = (\mathscr{I}_p - \mathscr{A}_p)^{-1} \mathscr{b}_p \mid \mathscr{A}_p \in \mathscr{A},\ \mathscr{b}_p \in \mathscr{b}\} \subseteq \{x_p \mid x_p \in x\}.$$

Доказательство. Покажем сначала, что система линейных уравнений

$$y_p = \mathscr{A}_p y_p + \mathscr{b}_p$$

имеет решение

$$y_p = (\mathscr{I}_p - \mathscr{A}_p)^{-1} \mathscr{b}_p$$

для $\mathscr{A}_p \in \mathscr{A}$ и $\mathscr{b}_p \in \mathscr{b}$. Так как верно $\mathscr{A}_p \in \mathscr{A}$, имеем $|\mathscr{A}_p| \leqslant |\mathscr{A}|$ и из теоремы Перрона и Фробениуса получаем

$$\rho(\mathscr{A}_p) \leqslant \rho(|\mathscr{A}_p|) \leqslant \rho(|\mathscr{A}|) < 1.$$

Отсюда следует, что матрица $\mathscr{I}_p - \mathscr{A}_p$ несингулярна, что и требовалось.

Рассмотрим теперь последовательные приближения

$$x^{(k+1)} = \mathcal{A}x^{(k)} + \mathcal{b}, \quad k \geqslant 0,$$

где $x^{(0)} = y_{\mathrm{p}} = \mathcal{A}_{\mathrm{p}} y_{\mathrm{p}} + \mathcal{b}_{\mathrm{p}}$. Из монотонности включения следует, что

$$y_{\mathrm{p}} = \mathcal{A}_{\mathrm{p}} y_{\mathrm{p}} + \mathcal{b}_{\mathrm{p}} \in \mathcal{A}x^{(0)} + \mathcal{b} = x^{(1)},$$

и для произвольного k

$$y_{\mathrm{p}} = \mathcal{A}_{\mathrm{p}} y_{\mathrm{p}} + \mathcal{b}_{\mathrm{p}} \in \mathcal{A}x^{(k)} + \mathcal{b} = x^{(k+1)}.$$

Из $\rho(|\mathcal{A}|) < 1$ следует, что $\lim_{k \to \infty} x^{(k)} = x^*$, а потому и $y_{\mathrm{p}} \in x^*$.

Так как x^* не зависит от начального вектора, мы получаем первую часть теоремы.

Для доказательства второй части теоремы построим вектор $u_{\mathrm{p}} \in V_n(\mathbb{R})$ из n нижних границ компонент вектора x^*. Из верхних границ аналогичным образом строится вектор $v_{\mathrm{p}} \in V_n(\mathbb{R})$. Тогда из $x^* = \mathcal{A}x^* + \mathcal{b}$ следует по правилам интервальной арифметики, что

$$u_{\mathrm{p}} = \mathcal{A}^*_{\mathrm{p}} u_{\mathrm{p}} + r_{\mathrm{p}} \quad \text{и} \quad v_{\mathrm{p}} = \mathcal{A}^{**}_{\mathrm{p}} v_{\mathrm{p}} + s_{\mathrm{p}},$$

где $u_{\mathrm{p}i} = (u_i)$, $v_{\mathrm{p}i} = (v_i)$ и

$$\mathcal{A}^*_{\mathrm{p}} = (a^*_{ij}), \quad a^*_{ij} = \begin{cases} i(A_{ij}), & u_j > 0, \\ s(A_{ij}), & u_j \leqslant 0, \end{cases}$$

$$\mathcal{A}^{**}_{\mathrm{p}} = (a^{**}_{ij}), \quad a^{**}_{ij} = \begin{cases} s(A_{ij}), & v_j < 0 \\ i(A_{ij}), & v_j \leqslant 0, \end{cases}$$

$$r_{\mathrm{p}} = (i(B_i)), \quad s_{\mathrm{p}} = (s(B_i)).$$

Из этих равенств следует, что u_{P} и v_{P} являются членами множества

$$\{y_{\mathrm{p}} = (\mathcal{I}_{\mathrm{p}} - \mathcal{A}_{\mathrm{p}})^{-1} \mathcal{b}_{\mathrm{p}} \mid \mathcal{A}_{\mathrm{p}} \in \mathcal{A}, \ \mathcal{b}_{\mathrm{p}} \in \mathcal{b}\},$$

что и завершает доказательство.

Метод итерации, рассмотренный в теореме 1, можно назвать полношаговым (T) по аналогии с соответствующим методом для «точечной системы уравнений». Аналогичный короткошаговый метод (S) получается разложением интервальной матрицы A в сумму

$$\mathcal{A} = \mathcal{L} + \mathcal{D} + \mathcal{U},$$

где $\mathcal{L}$—строго нижняя треугольная матрица, $\mathcal{D}$ — диагональная матрица и $\mathcal{U}$— строго верхняя треугольная матрица. Тогда короткошаговый итерационный метод определяется формулами

$$x^{(k+1)} = \mathcal{L}x^{(k+1)} + (\mathcal{D} + \mathcal{U})x^{(k)} + \mathcal{b}, \quad k \geqslant 0. \qquad (3)$$

Следующее утверждение касается сходимости этого короткошагового метода.

Теорема 3. *Итерационный метод*

$$x^{(k+1)} = \mathscr{L} x^{(k+1)} + (\mathscr{D} + \mathscr{U}) x^{(k)} + b, \quad k \geqslant 0,$$

с произвольным начальным вектором $x^{(0)} \in V_n(I(\mathbb{C}))$ *сходится к единственной неподвижной точке* x^* *тогда и только тогда, когда*

$$\rho((\mathscr{I}_p - |\mathscr{L}|)^{-1}(|\mathscr{D}| + |\mathscr{U}|)) < 1.$$

Доказательство. Мы собираемся применить теорему 5 из микромодуля 30 и с этой целью полагаем

$$f\colon V_n(\mathbb{C}) \to V_n(\mathbb{C}),$$

где

$$f_p(x_p) = \mathscr{L}_p x_p + (\mathscr{D}_p + \mathscr{U}_p) x_p + b_p$$

и

$$g_p\colon V_n(\mathbb{C}) \times V_n(\mathbb{C}) \to V_n(\mathbb{C}),$$

где

$$g_p(x_p, y_p) = \mathscr{L}_p x_p + (\mathscr{D}_p + \mathscr{U}_p) y_p + b_p.$$

Мы имеем тогда

$$g_p(x, x) = f_p(x_p) \quad \textit{для всех} \quad x \in V_n(I(\mathbb{C})),$$

и из (23, микромодуль 29) и (25, микромодуль 29) следует, что

$$q(g_p(x, z), g_p(y, z)) = q(\mathscr{L} x + (\mathscr{D} + \mathscr{U}) z + b, \mathscr{L} y + (\mathscr{D} + \mathscr{U}) z + b) \leqslant |\mathscr{L}| q(x, y),$$

$$q(g_p(z, x), g_p(z, y)) = q(\mathscr{L} z + (\mathscr{D} + \mathscr{U}) x + b, \mathscr{L} z + (\mathscr{D} + \mathscr{U}) y + b) \leqslant (|\mathscr{D}| + |\mathscr{U}|) q(x, y)$$

для всех x, y_p, $z \in V_n(I(\mathbb{C}))$. Мы имеем $\rho(|\mathscr{L}|) = 0$, так как $|\mathscr{L}|$ — строго нижняя треугольная матрица. Поэтому, полагая $Q_p := |\mathscr{L}|$, $\mathscr{R}_p := |\mathscr{D}| + |\mathscr{U}|$, мы оказываемся в условиях теоремы 5, из микромодуля 30, так что условие

$$\rho((\mathscr{I}_p - |\mathscr{L}|)^{-1}(|\mathscr{D}| + |\mathscr{U}|)) < 1$$

достаточное. Доказательство необходимости этого условия для сходимости при любом начальном векторе проводится так же, как соответствующее доказательство для полношагового метода в теореме 1.

Теорема Штейна и Розенберга, а также ее обобщение утверждают, что для $\mathscr{A} = \mathscr{L} + \mathscr{D} + \mathscr{U}$ верна эквивалентность $\rho(|\mathscr{A}|) < 1$ тогда и только тогда, когда

$$\rho((\mathcal{I}_p - |\mathcal{L}|)^{-1}(|\mathcal{D}| + |\mathcal{U}|)) < 1.$$

Так как условия сходимости полношагового и короткошагового методов необходимы и достаточны, получаем следующее утверждение.

Теорема 4. *Полношаговый метод* (2) *сходится для любого начального значени я* $x^{(0)} \in V_n(I(\mathbb{C}))$ *к единственной неподвижной точке тогда и только тогда, когда короткошаговый метод* (3) *сходится к единственной неподвижной точке для любого начального значения* $x^{(0)} \in V_n(I(\mathbb{C}))$.

Этот результат существенно отличается от соответствующего результата для точечных систем уравнений, где сходимость или расходимость полношагового метода не обязательно означает сходимость или расходимость короткошагового метода.

Поскольку умножение интервальных матриц на интервальные векторы в общем случае не дистрибутивно, то даже для случая $\rho(|\mathcal{A}|) < 1$ не очевидно, что неподвижная точка полношаговой итерации, удовлетворяющая равенству

$$x^* = \mathcal{A}x^* + b,$$

совпадает с неподвижной точкой короткошаговой итерации, удовлетворяющей равенству

$$\tilde{x}^* = \mathcal{L}\tilde{x}^* + (\mathcal{D} + \mathcal{U})\tilde{x}^* + b.$$

Однако, используя специальный вид матриц $\mathcal{L}$, $\mathcal{D}$ и $\mathcal{U}$, мы получаем с помощью определения действий над интервальными матрицами и векторами, что

$$\tilde{x}^* = \mathcal{L}x^* + (\mathcal{D} + \mathcal{U})\tilde{x}^* + b = (\mathcal{L} + \mathcal{D} + \mathcal{U})\tilde{x}^* + b = \mathcal{A}\tilde{x}^* + b.$$

Отсюда получается, что $x^* = x^*$, и приходим к следующему утверждению.

Следствие 5. *Если* $\rho(|\mathcal{A}|) < 1$, *то полношаговый и короткошаговый методы сходятся к неподвижной точке* x^* *уравнения*

$$x = \mathcal{A}x + b.$$

Рассмотрим теперь симметрический короткошаговый метод (SS), в котором матрица A раскладывается в сумму

$$\mathcal{A} = \mathcal{L} + \mathcal{U},$$

где $\mathcal{L}$— строго нижняя, а $\mathcal{U}$ — строго верхняя треугольные матрицы. Метод (SS) определяется соотношениями

$$\begin{cases} x^{(k+1/2)} = \mathscr{L}x^{(k+1/2)} + \mathscr{U}x^{(k)} + b, \\ x^{(k+1)} = \mathscr{L}x^{(k+1/2)} + \mathscr{U}x^{(k+1)} + b, \quad k \geqslant 0. \end{cases} \tag{SS}$$

Если не все диагональные элементы матрицы A обращаются в нуль, то вместо этого мы должны рассмотреть итерационный метод

$$\begin{cases} x^{(k+1/2)} = \mathscr{L}x^{(k+1/2)} + \mathscr{D}x^{(k)} + \mathscr{U}x^{(k)} + b, \\ x^{(k+1)} = \mathscr{L}x^{(k+1/2)} + \mathscr{D}x^{(k+1/2)} + \mathscr{U}x^{(k+1)} + b, \quad k \geqslant 0. \end{cases} \tag{SS'}$$

Для этого итерационного метода можно доказать утверждения, аналогичные тем, которые будут доказаны ниже для метода (SS).

Сходимость метода (SS) будет получена из следующего общего результата.

Теорема 6. *Пусть интервальная матрица* $\mathscr{A} \in M_{nn}(I(\mathbb{C}))$ *разложена в сумму* $\mathscr{A} = \mathscr{M} + \mathscr{N}$ *двух интервальных матриц* $\mathscr{M}$ *и* $\mathscr{N}$, *для которых верно* $\rho(|\mathscr{M}|) < 1$ *и* $\rho(|\mathscr{N}|) < 1$. *Тогда для произвольного вектора* $b \in V_n(I(\mathbb{C}))$ *верно следующее:*

(a) *Для любого интервального вектора* $x^{(0)} \in V_n(I(\mathbb{C}))$ *существует последовательность* $\{x^{(k)}\}_{k=0}^{\infty}$, *которая идовлетворяет итерационным формулам*

$$\text{(V)} \quad \begin{cases} x^{(k+1/2)} = \mathscr{M}x^{(k+1/2)} + \mathscr{N}x^{(k)} + b, \\ x^{(k+1)} = \mathscr{M}x^{(k+1/2)} + \mathscr{N}x^{(k+1)} + b, \quad k = 0, 1, 2, \ldots. \end{cases}$$

(b) *Если* $\rho\left((\mathscr{I}_p - |\mathscr{N}|)^{-1}|\mathscr{M}|(\mathscr{I}_p - |\mathscr{M}|)^{-1}|\mathscr{N}|\right) < 1$, *то уравнение* $x = \mathscr{A}x + b$ *имеет единственную неподвижную точку* x^*. *Если сверх того*

$$\mathscr{A}x^* = (\mathscr{M} + \mathscr{N})x^* = \mathscr{M}x^* + \mathscr{N}x^*,$$

то последовательность, вычисленная по формулам (V), *сходится к* x^* *для любого начального вектора* $x^{(0)}$. *(Как мы уже видели, для интервальных матриц не выполнен дистрибутивный закон. См. микромодуль* 29 *формулы* (6).)

(c) *Обратно, если уравнение* $x = \mathscr{A}x + b$ *имеет единственную неподвижную точку* x^* *и последовательность* (V) *сходится к* x^* *для любого начального приближения* $x^{(0)}$, *то*

$$\mathscr{A}x^* = (\mathscr{M} + \mathscr{N})x^* = \mathscr{M}x^* + \mathscr{N}x^*$$

и

$$\rho\left((\mathscr{I}_p - |\mathscr{N}|)^{-1}|\mathscr{M}|(\mathscr{I}_p - |\mathscr{M}|)^{-1}|\mathscr{N}|\right) < 1$$

Доказательство. (a) Для произвольного интервального вектора z из (23, микромодуль 29) и (25, микромодуль 29) следует, что для любых векторов x, y имеет место

$$q(\mathcal{M}x + \mathcal{N}z + b,\ \mathcal{M}y + \mathcal{N}z + b) = q(\mathcal{M}x,\ \mathcal{M}y) \leqslant |\mathcal{M}|\, q(x, y).$$

Ввиду $\rho(|\mathcal{M}|) < 1$ мы получаем по теореме 1, что для любого k уравнение

$$x^{(k+1/2)} = \mathcal{M}x^{(k+1/2)} + \mathcal{N}x^{(k)} + b$$

имеет единственную неподвижную точку $x^{(k+1/2)}$. Аналогично можно показать, что для любого k уравнение

$$x^{(k+1)} = \mathcal{M}x^{(k+1/2)} + \mathcal{N}x^{(k+1)} + b$$

имеет единственную неподвижную точку $x^{(k+1)}$. Тем самым доказаны существование и единственность последовательности $\{x^{(k)}\}_{k=0}^{\infty}$ при данном начальном векторе $x^{(0)}$.

(b) Сначала покажем, что $\rho(|\mathcal{A}|) < 1$. Так как $\rho(|\mathcal{M}|) < 1$, $\rho(|\mathcal{N}|) < 1$, то обратные матрицы

$$(\mathcal{I}_p - |\mathcal{M}|)^{-1} \text{ и } (\mathcal{I}_p - |\mathcal{N}|)^{-1},$$

как известно, существуют и неотрицательны. Поэтому вещественная матрица

$$(\mathcal{I}_p - |\mathcal{N}|)^{-1} |\mathcal{M}| (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{N}| = (\mathcal{I} - |\mathcal{N}|)^{-1} \times (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{M}|\,|\mathcal{N}|$$

также неотрицательна. Применяя известную теорему, получаем

$$\mathcal{O}_p \leqslant (\mathcal{I}_p - (\mathcal{I}_p - |\mathcal{N}|)^{-1} (\mathcal{I}_p - |\mathcal{M}|)^{-1} (|\mathcal{M}|\,|\mathcal{N}|)^{-1} = (\mathcal{I}_p - (|\mathcal{M}| + |\mathcal{N}|))^{-1} (\mathcal{I}_p - |\mathcal{M}|)(\mathcal{I}_p - |\mathcal{N}|),$$

а используя

$$(\mathcal{I}_p - |\mathcal{M}|)^{-1} \geqslant \mathcal{O}_p, \quad (\mathcal{I}_p - |\mathcal{N}|)^{-1} \geqslant \mathcal{O}_p,$$

имеем, наконец,

$$(\mathcal{I}_p - (|\mathcal{M}| + |\mathcal{N}|))^{-1} \geqslant \mathcal{O}_p.$$

Отсюда по известной теореме следует неравенство.

$$\rho(|\mathcal{M}| + |\mathcal{N}|) < 1.$$

Из соотношения

$$|\mathcal{A}| = |\mathcal{M} + \mathcal{N}| \leqslant |\mathcal{M}| + |\mathcal{N}|$$

ввиду теоремы Перрона и Фробениуса следует, что

$$\rho(|\mathcal{A}|) \leqslant \rho(|\mathcal{M}| + |\mathcal{N}|) < 1.$$

Поэтому ввиду теоремы 1 уравнение $x = \mathcal{A}x + b$ имеет единственную неподвижную точку x^*. Из равенства

$$x^* = \mathcal{A}x^* + b = \mathcal{M}x^* + \mathcal{N}x^* + b$$

с помощью (23, микромодуль 29)—(25, микромодуль 29) и (V) следует, что

$$\begin{aligned} q(x^{(k+1)}, x^*) &= q(\mathcal{M}x^{(k+1/2)} + \mathcal{N}x^{(k+1)} + b,\ \mathcal{M}x^* + \mathcal{N}x^* + b) \\ &\leqslant q(\mathcal{M}x^{(k+1/2)} + \mathcal{N}x^{(k+1)},\ \mathcal{M}x^{(k+1/2)} + \mathcal{N}x^*) \\ &\quad + q(\mathcal{M}x^{(k+1/2)} + \mathcal{N}x^*,\ \mathcal{M}x^* + \mathcal{N}x^*) \\ &\leqslant |\mathcal{N}|\, q(x^{(k+1)}, x^*) + |\mathcal{M}|\, q(x^{(k+1/2)}, x^*), \end{aligned}$$

а так как $(\mathcal{I}_p - |\mathcal{N}|)^{-1} \geqslant \mathcal{Q}_p$, это дает

$$q(x^{(k+1)}, x^*) \leqslant (\mathcal{I}_p - |\mathcal{N}|)^{-1} |\mathcal{M}|\, q(x^{(k+1/2)}, x^*).$$

Аналогичным образом получаем

$$q(x^{(k+1/2)}, x^*) \leqslant (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{N}|\, q(x^{(k)}, x^*);$$

откуда, наконец,

$$\begin{aligned} q(x^{(k+1)}, x^*) &\leqslant (\mathcal{I}_p - |\mathcal{N}|)^{-1} |\mathcal{M}| (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{N}|\, q(x^k, x^*) \\ &\leqslant \{(\mathcal{I}_p - |\mathcal{N}|)^{-1} |\mathcal{M}| (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{N}|\}^{k+1} q(x^{(0)}, x^*). \end{aligned}$$

Из того, что спектральный радиус выражения в фигурных скобках меньше 1, получаем, что $\lim_{k\to\infty} x^{(k)} = x^*$.

(c): Пусть уравнение $x = \mathcal{A}x + b$ имеет единственную неподвижную точку x^*. Из неравенства

$$q(x^{(k+1/2)}, x^*) \leqslant (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{N}|\, q(x^{(k)}, x^*),$$

которое выводится так же, как в доказательстве п. (b), следует, что последовательность $\{x^{(k+1/2)}\}_{k=0}^{\infty}$ сходится к x^* для любого $x^{(0)}$. Отсюда и из верхнего равенства (V) следует при $k \to \infty$, что

$$x^* = \mathcal{M}x^* + \mathcal{N}x^* + b,$$

т. е.

$$o_p = q(x^*, x^*) = q(\mathcal{M}x^* + \mathcal{N}x^* + b,\ \mathcal{A}x^* + b) = q(\mathcal{M}x^* + \mathcal{N}x^*,\ \mathcal{A}x^*)$$

или

$$\mathcal{A}x^* = (\mathcal{M} + \mathcal{N})x^* = \mathcal{M}x^* + \mathcal{N}x^*.$$

Мы должны еще доказать, что

$$\rho((\mathcal{I}_p - |\mathcal{N}|)^{-1} |\mathcal{M}| (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{N}|) < 1.$$

Чтобы сделать это, поступаем так же, как в доказательстве теоремы 1.

Из теоремы Перрона и Фробениуса известно, что матрица

$$(\mathcal{I}_p - |\mathcal{N}|)^{-1} |\mathcal{M}| (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{N}|$$

имеет неотрицательный собственный вектор, соответствующий неотрицательному собственному числу

$$\bar{\lambda} = \rho((\mathcal{I}_p - |\mathcal{N}|)^{-1} |\mathcal{M}| (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{N}|).$$

Теперь мы выбираем $x^{(0)}$ так, чтобы $d(x^{(0)})$ был собственным вектором, соответствующим собственному числу $\tilde{\lambda}$, и при этом хотя бы одна компонента вектора $d(x^{(0)})$ была больше, чем соответствующая компонента вектора $d(x^*)$. Тогда из (V) с помощью (12, микромодуль 29) и (18, микромодуль 29) следует, что

$$d(x^{(k+1/2)}) \geqslant |\mathcal{M}| d(x^{(k+1/2)}) + |\mathcal{N}| d(x^{(k)})$$

или

$$d(x^{(k+1/2)}) \geqslant (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{N}| d(x^{(k)}),$$

а также

$$d(x^{(k+1/2)}) \geqslant (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{N}| d(x^{(k)})$$

или

$$d(x^{(k+1)}) \geqslant (\mathcal{I}_p - |\mathcal{N}|)^{-1} |\mathcal{M}| d(x^{(k+1/2)}).$$

Наконец, мы получаем

$$\begin{aligned} d(x^{(k+1)}) &\geqslant (\mathcal{I}_p - |\mathcal{N}|)^{-1} |\mathcal{M}| (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{N}| d(x^{(k)}) \\ &\geqslant \{(\mathcal{I}_p - |\mathcal{N}|)^{-1} |\mathcal{M}| (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{N}|\}^{(k+1)} d(x^{(0)}) \\ &= \bar{\lambda}^{(k+1)} d(x^{(0)}). \end{aligned}$$

Из сходимости итераций (V) к x^* следует сходимость последовательности $\{d(x^{(k)})\}_{k=0}^{\infty}$ к $d(x^*)$. Предположение $\tilde{\lambda} \geqslant 1$ приводит к неравенству

$$d(x^{(k+1)}) \geqslant \bar{\lambda}^{(k+1)} d(x^{(0)}) \geqslant d(x^{(0)}), \quad k \geqslant 0,$$

что в пределе при $k \to \infty$ дает $d(x^*) \geqslant d(x^{(0)})$, а это противоречит выбору вектора $x^{(0)}$. Поэтому $\tilde{\lambda} < 1$, и теорема доказана.

Теперь возьмем в теореме 6

$$\mathcal{M} := \mathcal{L}, \quad \mathcal{N} := \mathcal{U}.$$

Тогда имеем $\rho(|\mathcal{M}|) = \rho(|\mathcal{N}|) = 0$. Ввиду специального выбора матриц $\mathcal{L}$ и $\mathcal{U}$ равенство

$$\mathcal{A}x = \mathcal{L}x + \mathcal{U}x$$

справедливо для всех интервальных векторов. Это дает приводимое ниже следствие теоремы 6.

Следствие 7. *Симметрический короткошаговый метод* (SS) *сходится к единственной неподвижной точке* x^* *уравнения* $x = \mathcal{A}x + b$ *для любого начального вектора* $x^{(0)}$ *тогда и только тогда, когда спектральный радиус матрицы*

$$(\mathcal{I}_p - |\mathcal{U}|)^{-1} |\mathcal{L}| (\mathcal{I}_p - |\mathcal{L}|)^{-1} |\mathcal{U}|$$

меньше единицы.

Мы уже установили в доказательстве теоремы 6 (b), что неравенство

$$\rho((\mathcal{I}_p - |\mathcal{N}|)^{-1} |\mathcal{M}| (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{N}|) < 1$$

влечет за собой $\rho(|\mathcal{A}|) < 1$. Теперь мы хотим показать, что в предположении

$$|\mathcal{A}| = |\mathcal{M}| + |\mathcal{N}|$$

верна и обратная импликация, если выполнены условия $\rho(|\mathcal{M}|) < 1$ и $\rho(|\mathcal{N}|) < 1$ из теоремы 6. Итак, предположим, что

$$\rho(|\mathcal{A}|) = \rho(|\mathcal{M}| + |\mathcal{N}|) < 1.$$

В силу известной теоремы матрица, обратная к $\mathcal{I}_p - |\mathcal{A}|$, существует, и мы имеем $(\mathcal{I}_p - |\mathcal{A}|)^{-1} \geqslant \mathcal{O}_p$. Рассмотрим следующее разложение:

$$\mathcal{I}_p - |\mathcal{A}| = (\mathcal{I}_p - |\mathcal{M}|)(\mathcal{I}_p - |\mathcal{N}|) - |\mathcal{M}||\mathcal{N}|.$$

Из $\rho(|\mathcal{M}|) < 1$, $\rho(|\mathcal{N}|) < 1$ следует, что обращения матриц $\mathcal{I}_p - |\mathcal{M}|$ и $\mathcal{I}_p - |\mathcal{N}|$ существуют и неотрицательны. Поэтому неотрицательно и произведение $(\mathcal{I}_p - |\mathcal{N}|)^{-1} \times (\mathcal{I}_p - |\mathcal{M}|)^{-1}$. Тем самым рассматриваемое разложение матрицы $\mathcal{I}_p - |\mathcal{A}|$ регулярно, так как матрица $|\mathcal{M}||\mathcal{N}|$ неотрицательна. Отсюда с помощью известной теоремы мы получаем соотношение

$$\rho((\mathcal{I}_p - |\mathcal{N}|)^{-1} (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{M}||\mathcal{N}|) = \rho((\mathcal{I}_p - |\mathcal{N}|)^{-1} \times |\mathcal{M}| (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\Lambda| < 1.$$

Собирая все вместе и применяя теорему 6 (b), получаем следующее утверждение.

Теорема 8. *Пусть интервальная матрица* A *разложена в сумму* $\mathcal{A} = \mathcal{M} + \mathcal{N}$ *двух интервальных матриц, для которых выполнено*

$$|\mathcal{A}| = |\mathcal{M}| + |\mathcal{N}| \text{ и } \rho(|\mathcal{M}|) < 1, \ \rho(|\mathcal{N}|) < 1.$$

Тогда неравенство

$$\rho((\mathcal{I}_p - |\mathcal{N}|)^{-1} |\mathcal{M}| (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{N}|) < 1$$

эквивалентно неравенству

$$\rho(|\mathscr{A}|) < 1.$$

Равенство $|\mathscr{A}| = |\mathscr{M}| + |\mathscr{N}|$ будет выполнено, например, в случае, когда $\mathscr{A} = (A_{ij})$ разложена в сумму $\mathscr{A} = \mathscr{M} + \mathscr{N}$ с $\mathscr{M} = (M_{ij})$, $\mathscr{N} = (N_{ij})$ таким образом, что для любых $1 \leq i, j \leq n$ по крайней мере одна из компонент M_{ij} и N_{ij} равна нулю.

Это последнее условие выполнено, например, для разложения, с которого начинается описание симметрического короткошагового метода (SS). Теорема 8 немедленно дает приводимое ниже утверждение.

Следствие 9. *Симметрический короткошаговый метод сходится к единственной неподвижной точке x^* уравнения $x = \mathscr{A}x + \mathscr{b}$ для любого начального вектора $x^{(0)} \in V_n(I(\mathbb{C}))$ тогда и только тогда, когда $\rho(|\mathscr{A}|) < 1$, т.е. когда полношаговый метод (а потому и короткошаговый метод) сходится к x^* для любого $x^{(0)} \in V_n(I(\mathbb{C}))$*

Теперь мы рассмотрим скорость, с которой сходится к x^* последовательность $\{x^{(k)}\}_{k=0}^{\infty}$ интервальных векторов, порожденных итерационной процедурой

$$x^{(k+1)} = f_p(x^{(k)}), \quad k \geqslant 0. \tag{4}$$

Определение 10. *Пусть $x^* = f_p(x^*)$ и пусть $\mathfrak{C}$ —множество всех последовательностей $\{x^{(k)}\}_{k=0}^{\infty}$, вычисленных по формуле* (4) *и удовлетворяющих условию $\lim_{k\to\infty} x^{(k)} = x^*$. Тогда величина*

$$\alpha = \sup\left\{\lim_{k\to\infty} \sup \|q(x^{(k)}, x^*)\|^{1/k} \{x^{(k)}\}_{k=0}^{\infty} \in \mathfrak{C}\right\}$$

называется асимптотическим фактором сходимости итерации (4) *к точке x^*.*

Пусть $\{x^{(k)}\}_{k=0}^{\infty}$ — последовательность, сходящаяся к x^*. Положим

$$\beta = \lim_{k\to\infty} \sup \|q(x^{(k)}, x^*)\|^{1/k}.$$

Так как $\lim_{k\to\infty} \|q(x^{(k)}, x^*)\| = 0$, мы получаем, что $0 \leqslant \beta < 1$, а значит, и $0 \leqslant \alpha \leqslant 1$. Из определения β следует, что для любого $\varepsilon > 0$ найдется k_0, такое что

$$\|q(x^{(k)}, x^*)\| \leqslant (\beta + \varepsilon)^k, \quad k \geqslant k_0. \tag{5}$$

Если $\beta < 1$, то можно выбрать $\varepsilon > 0$ таким образом, что $\beta + \varepsilon < 1$. Тогда неравенство (5) показывает, что последовательность $\|q(x^{(k)}, x^*)\|$ асимптотически сходится к нулю не хуже,

чем геометрическая прогрессия со знаменателем $\beta + \varepsilon$. Точная верхняя грань по всем последовательностям из $\mathfrak{C}$ характеризует асимптотически наихудший выбор вектора $x^{(0)}$. Определение 10 — это непосредственное обобщение определения фактора асимптотической сходимости для последовательностей точечных векторов.

Для дальнейшего важно, что α не зависит от нормы. Чтобы убедиться в этом, достаточно показать, что β не зависит от нормы Пусть $\|\cdot\|$ и $\|\cdot\|'$ — две нормы на векторах. Из теоремы об эквивалентности норм следует, что существуют вещественные числа $d \geqslant c > 0$, такие, что $c\|x_p\| \leqslant \|x_p\|' \leqslant d\|x_p\|$ для любых точечных векторов. Отсюда мы получаем

$$\begin{aligned}\limsup_{k\to\infty} \|q(x^{(k)}, x^*)\|^{1/k} &\leqslant \lim_{k\to\infty} \left(\frac{1}{c}\right)^{1/k} \limsup_{k\to\infty} \|q(x^{(k)}, x^*)\|'^{1/k} \\ &\leqslant \lim_{k\to\infty} \left(\frac{d}{c}\right)^{1/k} \limsup_{k\to\infty} \|q(x^{(k)}, x^*)\|^{1/k} \\ &= \limsup_{k\to\infty} \|q(x^{(k)}, x^*)\|^{1/k}.\end{aligned}$$

Перед тем как применить определение 10 к методам, рассмотренным в этом микромодуле, докажем одну теорему о точечных матрицах.

Рассмотрим положительный точечный вектор $h_p = (h_i)$, $h_i > 0$ и диагональную точечную матрицу $\mathscr{H}_p = \operatorname{diag}(1/h_i)$. Тогда неравенство

$$\|x_p\| = \max_{1 \leqslant i \leqslant n} (|x_i|/h_i)$$

определяет монотонную норму на векторах (короче, векторную норму), т .е. из

$$|x_p| \leqslant |y_p|$$

следует, что

$$\|x_p\| \leqslant \|y_p\|$$

(в частности, из $o_p \leqslant x_p \leqslant y_p$ следует, что справедливо неравенство $\|x_p\| \leqslant \|y_p\|$).

Полагая

$$\|\mathscr{A}_p\| = \sup_{\|x_p\|=1} \|\mathscr{A}_p x_p\|,$$

мы получаем норму на матрицах, подчиненную векторной норме $\|x_p\|$, причем

$$\|\mathscr{A}_p\| = \max_{1 \leqslant i \leqslant n} \left(\frac{1}{h_i} \sum_{j=1}^{n} |a_{ij}| h_j\right).$$

Теперь докажем следующее утверждение.

Теорема 11. *Для любой точечной матрицы* $\mathscr{A}_0 \geqslant O_0$ *и любого* $\varepsilon > 0$ *существует монотонная векторная норма* $\|\cdot\|$, *такая что имеет место*

$$\|\mathscr{A}_p\| = \sup_{\|x_p\|=1} \|\mathscr{A}_p x_p\| \leqslant \rho(\mathscr{A}_p) + \varepsilon.$$

Доказательство. Пусть матрица $\mathscr{A}_p \geqslant O_p$ неприводима. Тогда $\mathscr{A}_p$ имеет положительный собственный вектор $c_p = (c_i)$, $c > 0$, соответствующий собственному числу $\lambda = \rho(\mathscr{A}_p)$. Из равенства $\mathscr{A}_p c_p = \lambda c_p$ следует, что

$$\rho(\mathscr{A}_p) = \lambda = \frac{1}{c_i} \sum_{j=1}^{n} a_{ij} c_j = \|\mathscr{A}_p\| = \sup_{\|x_p\|=1} \|\mathscr{A}_p x_p\|,$$

где

$$\|x_p\| = \max_{1 \leqslant i \leqslant n} \frac{|x_i|}{c_i},$$

т. е.

$$\|\mathscr{A}_p\| \leqslant \rho(\mathscr{A}_p) + \varepsilon.$$

Если $\mathscr{A}_p \geqslant O_p$ приводима, то определим неприводимую матрицу $\tilde{\mathscr{A}}_p \geqslant O_p$ равенствами

$$\tilde{\mathscr{A}}_p = (\tilde{a}_{ij}), \quad \text{где } \tilde{a}_{ij} = \begin{cases} a_{ij}, & \text{если } a_{ij} > 0 \\ a > 0, & \text{если } a_{ij} = 0. \end{cases}$$

Очевидно, что $\tilde{\mathscr{A}}_p \geqslant \mathscr{A}_p \geqslant O_p$. Если теперь $\tilde{c}_p = (\tilde{c}_i)$ — положительный собственный вектор матрицы $\tilde{\mathscr{A}}_p$, соответствующий собственному числу $\lambda = \rho(\mathscr{A}_p)$, то из неравенства $|\lambda| \leqslant \|\tilde{\mathscr{A}}_p\|$ (которое верно для любой матричной нормы) следует, что

$$\rho(\tilde{\mathscr{A}}_p) = \frac{1}{\tilde{c}_i} \sum_{j=1}^{n} \tilde{a}_{ij} \tilde{c}_j = \|\tilde{\mathscr{A}}_p\| \geqslant \max_{1 \leqslant i \leqslant n} \frac{1}{c_i} \sum_{j=1}^{n} a_{ij} \tilde{c}_j = \|\mathscr{A}_p\| \geqslant \rho(\mathscr{A}_p),$$

где $\|x_p\| = \max_{1 \leqslant i \leqslant n} |x_i|/\tilde{c}_i$ и $\|\mathscr{A}_p\| = \sup_{\|x_p\|=1} \|\mathscr{A}_p x_p\|$. Так как спектральный радиус $\rho(\mathscr{A}_p)$ — непрерывная функция от элементов матрицы $\mathscr{A}_p$, получаем, что для каждого $\varepsilon > 0$ существует $\delta = \delta(\varepsilon) > 0$, такое что

$$\rho(\tilde{\mathscr{A}}_p) - \rho(\mathscr{A}_p) \leqslant \varepsilon$$

имеет место для всех $a \leqslant \delta(\varepsilon)$. Так как

$$\rho(\tilde{\mathscr{A}}_p) \geqslant \|\mathscr{A}_p\|,$$

получаем отсюда, что

$$\|\mathscr{A}_p\| - \rho(\mathscr{A}_p) \leqslant (\tilde{\mathscr{A}}_p) - \rho(\mathscr{A}_p) \leqslant \varepsilon$$

или

$$\|\mathscr{A}_p\| \leqslant \rho(\mathscr{A}_p) + \varepsilon.$$

После этой подготовки мы легко докажем следующее утверждение.

Теорема 12. *Пусть дано уравнение*

$$x = \mathscr{A}x + b,$$

где b *— интервальный вектор и* $\mathscr{A}$ *— интервальная матрица, для которой* $\rho(|\mathscr{A}|) < 1$. *Тогда асимптотический фактор сходимости* α_T *для полношагового метода удовлетворяет неравенству*

$$\alpha_T \leqslant \rho(|\mathscr{A}|),$$

фактор α_S *для короткошагового метода удовлетворяет неравенству*

$$\alpha_S \leqslant \rho((\mathscr{I}_p - |\mathscr{L}|)^{-1}(|\mathscr{D}| + |\mathscr{U}|)),$$

а фактор α_{SS} *для симметрического короткошагового метода удовлетворяет неравенству*

$$\alpha_{SS} \leqslant \rho((\mathscr{I}_p - |\mathscr{U}|)^{-1}|\mathscr{L}|(\mathscr{I}_p - |\mathscr{L}|)^{-1}|\mathscr{U}|).$$

Доказательство. Проведем доказательство для полношагового метода. Если $\rho(|\mathscr{A}|) < 1$, то, согласно теореме 1, полношаговый метод сходится для любого начального вектора $x^{(0)}$ к единственной неподвижной точке x^* уравнения $x = \mathscr{A}x + b$. Выбрав $x^{(0)}$ произвольным образом, мы получаем из свойств расстояния q, что

$$q(x^{(1)}, x^*) \leqslant |\mathscr{A}| q(x^{(0)}, x^*),$$

$$q(x^{(2)}, x^*) \leqslant |\mathscr{A}|^2 q(x^{(0)}, x^*),$$

и для произвольного $k \geq 1$

$$q(x^{(k)}, x^*) \leqslant |\mathscr{A}|^k q(x^{(0)}, x^*).$$

Используя монотонную векторную норму, которая существует по теореме 11, и подчиненную ей матричную норму, имеем

$$\|q(x^{(k)}, x^*)\| \leqslant (\rho(|\mathscr{A}|) + \varepsilon^k \|q(x^{(0)}, x^*)\|,$$

т. е.

$$\limsup_{k \to \infty} \|q(x^{(k)}, x^*)\|^{1/k} \leqslant \rho(|\mathscr{A}|) + \varepsilon.$$

Так как $\varepsilon > 0$ было произвольным и α_T не зависит от нормы, получаем неравенство $\alpha_T \leqslant \rho(|\mathscr{A}|)$, т. е. утверждение теоремы для этого случая. Остальные случаи рассматриваются аналогично.

Неизвестно, можно ли поставить знак равенства в наших оценках для $\alpha_T, \alpha_S, \alpha_{SS}$. Чтобы доказать, что это верно, скажем, для полношагового метода, нужно было бы указать начальный вектор $x^{(0)}$, для которого имеет место

$$\lim_{k\to\infty}\sup \|q(x^{(k)}, x^*)\| = \rho(|\mathcal{A}|).$$

Это можно довольно легко сделать в конкретных случаях, но доказательства для общего случая нет.

Замечания. Непосредственное применение интервального анализа к итерационному решению систем уравнений было впервые рассмотрено и показано, что полношаговый метод (2) для вещественной интервальной матрицы $\mathcal{A}$ и вещественного интервального вектора $\mathcal{b}$ сходится к единственной неподвижной точке, если $\||\mathcal{A}|\| < 1$. Необходимые и достаточные условия из теорем 1 и 3 были найдены для вещественных интервальных матриц и для интервальных матриц с элементами из $R(\mathbb{C})$. Заметим, не вдаваясь в подробности, что более общую задачу о неподвижной точке

$$\mathcal{X} = \mathcal{A}\mathcal{X} + \mathcal{B}, \quad \mathcal{A}, \mathcal{B}, \mathcal{X} \in M_{nn}(I(\mathbb{C}))$$

можно исследовать тем же методом, который был использован в этом микромодуле. Например, если $\rho(\mathcal{A}) < 1$, то для итерации

$$\mathcal{X}^{(k+1)} = \mathcal{A}\mathcal{X}^{(k)} + \mathcal{B}, \quad k \geqslant 0,$$

имеется единственная неподвижная точка $\mathcal{X}^*$ при любой начальной матрице $\mathcal{X}^{(0)} \in M_{nn}(I(\mathbb{C}))$. Мы имеем в этом случае

$$\{\mathcal{Y}_p = (\mathcal{I}_p - \mathcal{A}_p)^{-1}\mathcal{B}_p \,|\, \mathcal{A}_p \in \mathcal{A}, \mathcal{B}_p \in \mathcal{B}\} \subseteq \mathcal{X}^*.$$

Следует также сделать некоторые замечания по поводу условия $\rho(|\mathcal{A}|) < 1$, которое в силу теоремы 1 необходимо и достаточно для сходимости полношагового метода. В частном случае, когда $\mathcal{A}$ — точечная матрица и $\mathcal{b}$ — точечный вектор, мы получаем, что последовательные приближения

$$x^{(k+1)} = \mathcal{A}_p x^{(k)} + \mathcal{b}_p, \quad k \geqslant 0, \tag{6}$$

сходятся к решению

$$x_p = (\mathcal{I}_p - \mathcal{A}_p)^{-1}\mathcal{b}_p$$

для любого интервального вектора $x^{(0)}$ тогда и только тогда, когда $\rho(|\mathcal{A}_p|) < 1$. С другой стороны, последовательные приближения

$$x_p^{(k+1)} = \mathcal{A}_p x_p^{(k)} + \mathcal{b}_p, \quad k \geqslant 0, \tag{7}$$

сходятся для всех $x_p^{(0)} \in V_n(\mathbb{C})$к

$$x_{\rm p} = (\mathcal{J}_{\rm p} - \mathcal{A}_{\rm p})^{-1}\, b,$$

если $\rho(\mathcal{A}_{\rm p}) < 1$.

Если мы проводим итерацию (6) в $V_n(K(\mathbb{C}))$, то в обозначениях $\mathcal{A}_{\rm p} = (a_{rs})$, $a_{rs} = a_{rs}^{(1)} + ia_{rs}^{(2)}$, $1 \leqslant r,\ s \leqslant n$, мы имеем:

$$|\mathcal{A}_{\rm p}| = |\mathcal{A}_{\rm p}|_2 := \left(\sqrt{(a_{rs}^{(1)})^2 + (a_{rs}^{(2)})^2}\right).$$

Если, с другой стороны, мы проводим итерацию (6) в $V_n(R(\mathbb{C}))$, то имеем

$$|\mathcal{A}_{\rm p}| = |\mathcal{A}_{\rm p}|_1 := \left(|a_{rs}^{(1)}| + |a_{rs}^{(2)}|\right).$$

В силу неравенства

$$\sqrt{a_1^2 + a_2^2} \leqslant |a_1| + |a_2|$$

имеем

$$|\mathcal{A}_{\rm p}|_2 \leqslant |\mathcal{A}_{\rm p}|_1.$$

Из

$$\mathcal{O}_{\rm p} \leqslant |\mathcal{A}_{\rm p}|_2 \leqslant |\mathcal{A}_{\rm p}|_1$$

следует, что

$$\rho(\mathcal{A}_{\rm p}) \leqslant \rho(|\mathcal{A}_{\rm p}|_2) \leqslant \rho(|\mathcal{A}_{\rm p}|_1). \qquad (8)$$

Последнее неравенство показывает, что (6) требует в общем случае более сильных предположений, чем (7). Это происходит потому, что необходимое и достаточное условие для (6) гарантирует сходимость для любых интервальных векторов, а множество всех точечных векторов, которые допускаются в (7) в качестве начальных, является собственным подмножеством множества всех интервальных векторов.

Вторая часть неравенства (8) показывает, что при реализации итерации (6) в $V_n(K(\mathbb{C}))$ от $\mathcal{A}_{\rm p}$ требуется выполнение не менее сильных условий, чем в случае $V_n(R(\mathbb{C}))$. Это верно и для систем уравнений, коэффициенты которых — невырожденные интервалы. С этой точки зрения арифметические операции в $K(\mathbb{C})$ имеют некоторые преимущества перед операциями в $R(\mathbb{C})$.

Микромодуль 32
Методы релаксации

Мы уже рассмотрели полношаговый, короткошаговый и симметрический короткошаговый методы. Существует много других методов решения линейных систем точечных уравнений вида

$$x_p = \mathcal{A}_p x_p + b_p,$$

для которых можно уменьшить асимптотический фактор сходимости путем введения одного или нескольких параметров. Большую часть этих приемов можно перенести на итерационные методы, использующие интервальные векторы. В качестве примера мы рассмотрим метод релаксации для короткошаговою случая.

Как и в короткошаговом методе, разложим матрицу $\mathcal{A}$ в сумму $\mathcal{A} = \mathcal{L} + \mathcal{D} + \mathcal{U}$, где $\mathcal{L}$— нижняя строго треугольная матрица, $\mathcal{U}$— верхняя строго треугольная матрица и $\mathcal{D}$—диагональная матрица. Затем строим последовательные приближения

$$\tilde{X}_i^{(k+1)} = \sum_{j=1}^{i-1} A_{ij} X_j^{(k+1)} + \sum_{j=i}^{n} A_{ij} X_j^{(k)} + B_i,$$

$$X_i^{(k+1)} = (1-\omega) X_i^{(k)} + \omega \tilde{X}_i^{(k+1)}, \quad 1 \leqslant i \leqslant n, \quad k \geqslant 0,$$

начиная с произвольного интервального вектора $x^{(0)}$. С помощью векторных обозначений эти формулы можно записать в виде

$$x^{(k+1)} = (1-\omega) x^{(k)} + \omega \{\mathcal{L} x^{(k+1)} + (\mathcal{D} + \mathcal{U}) x^{(k)} + b\}, \quad k \geqslant 0,$$

где ω>0 — параметр. Так же, как это делалось для полношагового или короткошагового метода, можно показать, что

$$\rho((\mathcal{I}_p - \omega|\mathcal{L}|)^{-1} \{|1-\omega| \mathcal{I}_p + \omega(|\mathcal{D}| + |\mathcal{U}|)\}) < 1$$

является необходимым и достаточным условием сходимости метода к единственной неподвижной точке для произвольного начального вектора.

Можно далее показать, что при $\rho(|\mathcal{A}|) < 1$ это условие выполнено для всех значений ω, которые удовлетворяют неравенству

$$0 < \omega < 2/(1 + \rho(|\mathcal{A}|)).$$

Если полношаговый метод сходится и ω удовлетворяет приведенному неравенству, то метод релаксации тоже сходится к неподвижной точке $\tilde{x}^*$, удовлетворяющей уравнению

$$\tilde{x}^* = (1-\omega)\tilde{x}^* + \omega(\mathcal{A}\tilde{x}^* + b).$$

Эта неподвижная точка, вообще говоря, отлична от неподвижной точки x^* уравнения $x = \mathcal{A}x + b$. Чтобы показать это, заметим сначала, что для всех вещественных чисел *a, b* и невырожденных интервалов *Z* (т. е. таких Z, что $d(Z) > 0$) при *ab*>0 выполнено соотношение

$$(a+b)Z = aZ + bZ.$$

Пусть $0 < \omega < 1$. Тогда имеем

$$(1-\omega)x^* + \omega(\mathcal{A}x^* + b) = (1-\omega)x^* + \omega x^* = (1-\omega+\omega)x^* = x^*;$$

т. е. $x^* = \tilde{x}^*$, так как неподвижная точка $\tilde{x}^*$, вычисленная методом релаксации, единственна.

Если же, напротив, мы имеем ω > 1, то

$$\begin{aligned}\mathcal{A}x^* + b = x^* &= (1-\omega+\omega)x^* \subseteq (1-\omega)x^* + \omega x^* \\ &= (1-\omega)x^* + \omega(\mathcal{A}x^* + b) \\ &= (1-\omega)x^* + \omega\{\mathcal{L}x^* + (\mathcal{D}+\mathcal{U})x^* + b\}.\end{aligned}$$

Если мы возьмем начальное приближение $\tilde{x}^{(0)} = x^*$ для метода релаксации, то из этого включения и монотонности включения следует, что

$$x^* \subseteq (1-\omega)\tilde{x}^{(0)} + \omega\{\mathcal{L}\tilde{x}^{(1)} + (\mathcal{D}+\mathcal{U})\tilde{x}^{(0)} + b\} =: \tilde{x}^{(1)}.$$

С помощью математической индукции можно показать, что

$$x^* \subseteq \tilde{x}^{(k)}, \quad k \geqslant 0, \text{ т. е. } x^* \subseteq \tilde{x}^*.$$

Из простых примеров видно, что включение здесь собственное. Поэтому при ω > 1 мы должны учитывать, что применение метода релаксации «увеличивает» неподвижную точку уравнения $x^* = \mathcal{A}x^* + b$. Это нежелательный эффект, так как задача состоит в нахождении интервального вектора, который содержит множество

$$\{y_p = (\mathcal{I}_p - \mathcal{A}_p)^{-1} b_p \mid \mathcal{A}_p \in \mathcal{A},\ b_p \in b\}$$

и дает достаточно хорошую локализацию.

В частном случае, когда $\mathcal{A}$ — точечная матрица и b — точечный вектор, мы всегда имеем $x^* = x^*$. Множество точечных векторов — это подмножество множества интервальных векторов. Если мы выберем начальный вектор точечным, то все последовательные приближения и неподвижная точка также будут точечными векторами. Поэтому мы имеем для полношагового метода равенство

$$x_p^* = \mathcal{A}_p x_p^* + b_p,$$

а для метода релаксации — равенство

$$\hat{x}_p^* = (1-\omega)\hat{x}_p^* + \omega(\mathcal{A}_p\hat{x}_p^* + b_p),$$

откуда следует, что $\hat{x}_p^* = x_p^*$. Таким образом, для случая точечной матрицы и точечного вектора оба метода сходятся к решению

$$x_p^* = (\mathcal{I}_p - \mathcal{A}_p)^{-1} b_p$$

уравнения $x_p = \mathcal{A}_p x_p + b_p$.

Мы хотим теперь исследовать этот случай несколько подробнее. Последовательность $\{d(x^{(k)})\}_{k=0}^{\infty}$, полученная вычислением ширины из последовательности $\{x^{(k)}\}_{k=0}^{\infty}$, сходится к нулевому вектору, так как $\lim_{k\to\infty} x^{(k)} = x_p^* = (\mathcal{I}_p - \mathcal{A}_p)^{-1} b_p$. Поэтому кажется, что естественно характеризовать скорость сходимости величиной $\mathcal{R}$, вводимой следующим определением.

Определение 1. Пусть $x_p^* = f_p(x_p^*)$, и пусть 6 обозначает множество всех последовательностей $\{x^{(k)}\}_{k=0}^{\infty}$, которые получаются применением итерационного метода

$$x^{(k+1)} = f_p(x^{(k)}), \quad k \geqslant 0,$$

и для которых $\lim_{k\to\infty} x^{(k)} = x_p^*$. Тогда величина

$$\tilde{\alpha} = \sup\left\{\limsup_{k\to\infty} \left\| d\left(x_p^{(k)}\right)\right\|^{1/k} \mid \{x^{(k)}\}_{k=0}^{\infty} \in \mathfrak{G}\right\}$$

называется асимптотическим фактором сходимости этого итерационного метода в точке x_p^*.

По аналогии с величиной α (определение 10, микромодуль 31) мы можем сказать, что $\mathcal{R}$ характеризует асимптотически наихудший случай при произвольном выборе $x^{(0)}$. Точно так же, как для α, можно показать, что $\mathcal{R}$ не зависит от используемой нормы.

Докажем теперь следующее утверждение.

Теорема 2. *Пусть задано уравнение*

$$x_p = \mathcal{A}_p x_p + b_p$$

для точечной матрицы $\mathcal{A}_p$, *такой что*

$$\rho(|\mathcal{A}_p|) < 1,$$

и точечного вектора b_p. *Тогда асимптотический фактор сходимости* $\tilde{\alpha}_T$ *для полношагового метода* (2, микромодуль 31) *удовлетворяет равенству*

$$\tilde{\alpha}_T = \rho(|\mathscr{A}_p|),$$

а асимптотический фактор сходимости $\mathscr{D}_R$ *для метода релаксации — равенству*

$$\tilde{\alpha}_R = \rho((\mathscr{I}_p - \omega|\mathscr{L}_p|)^{-1}\{|1-\omega|\mathscr{I}_p + \omega(|\mathscr{D}_p| + |\mathscr{U}_p|)\})$$

для

$$0 < \omega < 2/(1+\rho(|\mathscr{A}_p|)).$$

Доказательство. Проведем рассуждение для метода релаксации. Начав с произвольного интервального вектора $x^{(0)}$, мы с помощью (12, микромодуль 29) и (19, микромодуль 29) получаем из нашей итерационной формулы

$$x^{(k+1)} = (1-\omega)x^{(k)} + \omega\{\mathscr{L}_p x^{(k+1)} + (\mathscr{D}_p + \mathscr{U}_p)x^{(k)} + f_p\}, \quad k \geqslant 0$$

что

$$d(x^{(k+1)}) = |1-\omega|d(x^{(k)}) + \omega|\mathscr{L}_p|d(x^{(k+1)}) + \omega(|\mathscr{D}_p| + |\mathscr{U}_p|)d(x^{(k)})$$

или

$$d(x^{(k+1)}) = (\mathscr{I}_p - \omega|\mathscr{L}_p|)^{-1}\{|1-\omega|\mathscr{I}_p + \omega(|\mathscr{D}_p| + |\mathscr{U}_p|)\}d(x^{(k)})$$
$$= \{(\mathscr{I}_p - \omega|\mathscr{L}_p|)^{-1}(|1-\omega|\mathscr{I}_p + \omega(|\mathscr{D}_p| + |\mathscr{U}_p|))\}^{k+1}d(x^{(0)}).$$

Отсюда непосредственно получаем, что

$$\tilde{\alpha}_R \leqslant \rho((\mathscr{I}_p - \omega|\mathscr{L}_p|)^{-1}\{|1-\omega|\mathscr{I}_p + \omega(|\mathscr{D}_p| + |\mathscr{U}_p|)\}).$$

Если теперь выбрать конкретный вектор $x^{(0)}$ так, что $d(x^{(0)})$ — собственный вектор неотрицательной матрицы $(\mathscr{I}_p - \omega|\mathscr{L}_p|)^{-1} \times \times\{|1-\omega|\mathscr{I}_p + \omega(|\mathscr{D}_p| + |\mathscr{U}_p|)\}$, соответствующий собственному числу λ, равному спектральному радиусу этой матрицы, то из уравнения для $d(x^{(k+1)})$ следует, что

$$d(x^{(k+1)}) = \lambda^{k+1}d(x^{(0)}),$$

откуда получается нужное утверждение. Доказательство для полношагового метода можно провести аналогично.

Только что доказанная теорема позволяет сформулировать утверждение об асимптотически наискорейшем (в смысле определения 1) методе.

Теорема 3. *В условиях теоремы 2*

$$\min\{\tilde{\alpha}_R | 0 < \omega < 2/(1+\rho(|\mathscr{A}_p|))\} = \tilde{\alpha}_R|_{\omega=1}$$
$$= \rho((\mathscr{I}_p - |\mathscr{L}_p|)^{-1}(|\mathscr{D}_p| + |\mathscr{U}_p|)) = \tilde{\alpha}_S.$$

а также

$$\bar{a}_S \leqslant \bar{a}_T.$$

Доказательство. Рассмотрим вещественную точечную матрицу

$$\mathscr{P}_p = ((1 - |1 - \omega|)/\omega)\,\mathscr{I}_p - |\mathscr{A}_p|$$

и ее разложение $\mathscr{P}_p = \mathscr{M}_{p\omega} - \mathscr{N}_{p\omega}$, где

$$\mathscr{M}_{p\omega} = (1/\omega)(\mathscr{I}_p - \omega|\mathscr{L}_p|), \quad \mathscr{N}_{p\omega} = (1/\omega)\{|1 - \omega|\mathscr{I}_p + \omega(|\mathscr{D}_p| + |\mathscr{U}_p|)\}.$$

Это разложение регулярно, так как $\mathscr{M}_{p\omega}^{-1}$ существует и $\mathscr{M}_{p\omega}^{-1} \geqslant \mathscr{O}_p$, $\mathscr{N}_{p\omega} \geqslant \mathscr{O}_p$. Если ω удовлетворяет неравенству

$$0 < \omega < 2/(1 + \rho|\mathscr{A}_p|), \quad \text{то} \quad \mathscr{P}_p^{-1} \geqslant \mathscr{O}_p,$$

и в силу неравенства $\mathscr{N}_{p\omega} \geqslant \mathscr{N}_{p1}$ мы получаем , что

$$\rho(\mathscr{M}_{p1}^{-1}\mathscr{N}_{p1}) = \rho((\mathscr{I}_p - |\mathscr{L}_p|)^{-1}(|\mathscr{D}_p| + |\mathscr{U}_p|)) \leqslant \rho(\mathscr{M}_{p\omega}^{-1}\mathscr{N}_{p\omega}) < 1.$$

Этим доказана первая часть теоремы. Ее вторая часть следует из теоремы Штейна и Розенберга и ее обобщения, утверждающего, что из $\rho(|\mathscr{A}_p|) < 1$ следует

$$\rho((\mathscr{I}_p - |\mathscr{L}_p|)^{-1}(|\mathscr{D}_p| + |\mathscr{U}_p|)) \leqslant \rho(|\mathscr{A}_p|).$$

В предыдущей теореме утверждается, что в случае системы точечных уравнений невозможно асимптотически (в смысле определения 1) ускорить сходимость короткошагового метода, используя метод релаксации. Кроме того, короткошаговый метод сходится не медленнее, чем полношаговый.

Теперь рассмотрим связь между описанным выше итерационным методом в пространстве интервальных векторов и принципом локализации решений, который получен иным способом.

Рассмотрим систему линейных уравнений

$$x_p = \mathscr{A}_p x_p + b_p,$$

где $\mathscr{A}_p$ — вещественная точечная матрица и b_p — вещественный точечный вектор. Мы вводим естественный (т. е. покомпонентный) порядок на точечных матрицах и точечных векторах. Точечная матрица $\mathscr{P}_p$ называется изотонной (соответственно антитонной), если из $x_p \geqslant o_p$ следует $\mathscr{P}_p x_p \geqslant o_p$ (соответственно o_p). Теперь матрица $\mathscr{A}_p$ раскладывается в сумму изотонной и антитонной точечных матриц:

$$\mathscr{A}_p = \mathscr{T}_{p1} + \mathscr{T}_{p2}.$$

Начиная с пары точечных векторов $v_p^{(0)}$ и $w_p^{(0)}$, для которых верно $v_p^{(0)} \leqslant w_p^{(0)}$, наш итерационный метод вычисляет две последовательности $\{v_p^{(k)}\}_{k=0}^{\infty}$ и $\{w_p^{(k)}\}_{k=0}^{\infty}$ по формулам

$$\begin{cases} v_p^{(k+1)} = \mathcal{T}_{p1} v_p^{(k)} + \mathcal{T}_{p2} w_p^{(k)} + b_p, \\ w_p^{(k+1)} = \mathcal{T}_{p1} w_p^{(k)} + \mathcal{T}_{p2} v_p^{(k)} + b_p, \quad k \geqslant 0. \end{cases} \tag{1}$$

Если теперь $v_p^{(0)} \leqslant v_p^{(1)} \leqslant w_p^{(1)} \leqslant w_p^{(0)}$, то с помощью математической индукции можно показать, что

$$v_p^{(0)} \leqslant v_p^{(1)} \leqslant \ldots \leqslant v_p^{(k)} \leqslant v_p^{(k+1)} \leqslant w_p^{(k+1)} \leqslant w_p^{(k)} \leqslant \ldots \leqslant w_p^{(1)} \leqslant w_p^{(0)}.$$

Поэтому последовательности $\{v_p^{(k)}\}_{k=0}^{\infty}$ и $\{w_p^{(k)}\}_{k=0}^{\infty}$ сходятся, и простые рассуждения показывают, что решение x_p^* уравнения $x_p = \mathcal{A}_p x_p + b_p$ существует и находится между граничными точками. Если $\rho(|\mathcal{A}_p|) = \rho(\mathcal{T}_{p1} - \mathcal{T}_{p2}) < 1$, то $v_p^* = w_p^* = x_p^*$.

Рассматривая вместе с методом итераций для

$$\{v_p^{(k)}\}_{k=0}^{\infty}$$

и $\{w_p^{(k)}\}_{k=0}^{\infty}$ также итерационный метод

$$x^{(k+1)} = \mathcal{A}_p x^{(k)} + b_p, \quad k \geqslant 0, \tag{2}$$

с интервалом $x^{(0)} = ([v_i^{(0)}, w_i^{(0)}])$ в качестве начального приближения и учитывая правила умножения интервальных матриц на интервальные векторы, мы сразу видим, что границы компонент последовательности интервальных векторов $\{x^{(k)}\}_{k=0}^{\infty}$ совпадают с компонентами последовательностей $\{v_p^{(k)}\}_{k=0}^{\infty}$ и $\{w_p^{(k)}\}_{k=0}^{\infty}$.

Сказанное только что об итерационном методе (2) показывает, что в методе (1) можно избавиться от предположения $v_p^{(0)} \leqslant v_p^{(1)} \leqslant w_p^{(1)} \leqslant w_p^{(0)}$, если выполнено условие $v_p^{(0)} \leqslant x_p^* \leqslant w_p^{(0)}$, которое гарантирует локализацию решения x_p^*. Последовательности $\{v_p^{(k)}\}_{k=0}^{\infty}$ и $\{w_p^{(k)}\}_{k=0}^{\infty}$ в этом случае уже не обязательно сходятся монотонно. Монотонность можно восстановить, если брать пересечение после каждого шага.

Замечания. Утверждения, аналогичные тем, которые мы доказали для метода релаксации, справедливы и для симметрического метода релаксации (SR)

$$x^{(k+1/2)} = (1-\omega)x^{(k)} + \omega\{\mathscr{L}x^{(k+1/2)} + \mathscr{U}x^{(k)} + b\},$$
$$x^{(k+1)} = (1-\omega)x^{(k+1/2)} + \omega\{\mathscr{L}x^{(k+1/2)} + \mathscr{U}x^{(k+1)} + b\}, \quad k \geqslant 0,$$

который при $\omega = 1$ сводится к симметрическому короткошаговому методу (SS), описанному в микромодуле 31. Необходимым и достаточным условием сходимости этого метода к единственной неподвижной точке при произвольном начальном интервальном векторе является

$$\rho((\mathscr{I}_p - \omega|\mathscr{U}|)^{-1}(|1-\omega|\mathscr{I}_p + \omega|\mathscr{L}|)(\mathscr{I}_p - \omega|\mathscr{L}|)^{-1} \times (|1-\omega|\mathscr{I}_p + \omega|\mathscr{U}|)) < 1.$$

Если $\rho(|\mathscr{A}|) < 1$, то предыдущее условие выполняется при

$$0 < \omega < 2/(1 + \rho(|\mathscr{A}|)).$$

Доказательство можно провести так же, как для метода релаксации.

Если $\mathscr{A}$ — точечная матрица, то по аналогии с рассуждением из теоремы 2 можно показать, что асимптотический фактор сходимости, введенный в определении 1, равен

$$\tilde{\alpha}_{SR} = \rho((\mathscr{I}_p - \omega|\mathscr{U}_p|)^{-1}(|1-\omega|\mathscr{I}_p + \omega|\mathscr{L}_p|) \times (\mathscr{I}_p - \omega|\mathscr{L}_p|)^{-1}(|1-\omega|\mathscr{I}_p + \omega|\mathscr{U}_p|)),$$

и по аналогии с теоремой 3 имеем

$$\tilde{\alpha}_{SS} \leqslant \tilde{\alpha}_{SR}$$

для $0 < \omega < 2/(1 + \rho(|\mathscr{A}_p|))$.

Покажем, что имеет место $\tilde{\alpha}_{SS} \leqslant \tilde{\alpha}_S$. Матрица

$$(\mathscr{I}_p - |\mathscr{L}_p|)^{-1}|\mathscr{U}_p|$$

неотрицательна и всегда приводима, так как ее первый столбец состоит из нулей. Если добавить положительную матрицу $\Delta\mathscr{U}_p$ (вообще говоря, не являющуюся верхней треугольной), то можно сделать матрицу $(\mathscr{I}_p - |\mathscr{L}_p|)^{-1}(|\mathscr{U}_p| + \Delta\mathscr{U}_p)$ неприводимой. Используя теорему Перрона и Фробениуса, мы получаем

$$(\mathscr{I}_p - |\mathscr{L}_p|)^{-1}(|\mathscr{U}_p| + \Delta\mathscr{U}_p)x_p = \lambda x_p,$$

где λ — спектральный радиус матрицы из левой части и вектор x_p имеет только положительные компоненты. Простое преобразование (корректное при $0 < \lambda < 1$ и достаточно малой $\Delta\mathscr{U}_p$) дает

$$x_{\mathrm{p}} = (\mathcal{I}_{\mathrm{p}} - (1/\lambda)(|\mathcal{U}_{\mathrm{p}}| + \Delta\mathcal{U}_{\mathrm{p}}))^{-1} |\mathcal{L}_{\mathrm{p}}| x_{\mathrm{p}}$$
$$\geqslant (\mathcal{I}_{\mathrm{p}} - (|\mathcal{U}_{\mathrm{p}}| + \Delta\mathcal{U}_{\mathrm{p}}))^{-1} |\mathcal{L}_{\mathrm{p}}| x_{\mathrm{p}}.$$

Окончательно получаем

$$(\mathcal{I}_{\mathrm{p}} - (|\mathcal{U}_{\mathrm{p}}| + \Delta\mathcal{U}_{\mathrm{p}}))^{-1} |\mathcal{L}_{\mathrm{p}}| (\mathcal{I}_{\mathrm{p}} - |\mathcal{L}_{\mathrm{p}}|)^{-1} (|\mathcal{U}_{\mathrm{p}}| + \Delta\mathcal{U}_{\mathrm{p}}) x_{\mathrm{p}} \leqslant \lambda x_{\mathrm{p}}.$$

Это неравенство верно и при $X = 0$. Применяя теперь известную теорему, получаем, что спектральный радиус матрицы из левой части не превосходит λ. Так как это верно для всех матриц $\Delta\mathcal{U}_{\mathrm{p}}$, делающих матрицу $(\mathcal{I}_{\mathrm{p}} - |\mathcal{L}_{\mathrm{p}}|)^{-1}(|\mathcal{U}_{\mathrm{p}}| + \Delta\mathcal{U}_{\mathrm{p}})$ неприводимой, то мы получаем нужное утверждение, устремляя $\Delta\mathcal{U}_{\mathrm{p}}$ к нулю, так как собственные числа непрерывно зависят от элементов матрицы.

Микромодуль 33

Оптимальность симметрического короткошагового метода со взятием пересечения на каждом шаге

В этом микромодуле предполагается, что все интервальные матрицы взяты из пространства $M_{nn}(R(\mathbb{C}))$, а все интервальные векторы — из $V_n(R(\mathbb{C}))$.

Мы собираемся теперь исследовать некоторые модификации полношагового, короткошагового и симметрического корочкошагового методов. Если полношаговый метод

$$x^{(k+1)} = \mathcal{A}x^{(k)} + \mathcal{b}$$

имеет начальный вектор $x^{(0)}$ для которого $x^* \subseteq x^{(0)}$, то из монотонности включениясследует, что

$$x^* = \mathcal{A}x^* + \mathcal{b} \subseteq \mathcal{A}x^{(0)} + \mathcal{b} = x^{(1)}.$$

Это показывает, что и $x^{(1)}$ содержит неподвижную точку и вектор $x^{(0)}$, а значит, и их пересечение $x^{(0)} \cap x^{(1)}$. Поэтому естественно продолжать итерацию, используя это новое включение. Это приводит к итерационной процедуре

$$x^{(k+1)} = \{\mathcal{A}x^{(k)} + \mathcal{b}\} \cap x^{(k)}, \quad k \geqslant 0,$$

которую мы будем называть полношаговым методом со взятием пересечения на каждом шаге (TI).

Если провести те же рассмотрения для короткошагового метода, то получится итерационная процедура

$$x^{(k+1)} = \{\mathscr{L} x^{(k+1)} + (\mathscr{D} + \mathscr{U}) x^{(k)} + b\} \cap x^{(k)}, \quad k \geqslant 0,$$

которую мы назовем короткошаговым методом со взятием пересечения на каждом шаге (SI).

В случае короткошагового метода имеется еще одна возможность:

$$X_i^{(k+1)} = \left\{ \sum_{j=1}^{i-1} A_{ij} X_j^{(k+1)} + \sum_{j=i}^{n} A_{ij} X_j^{(k)} + B_i \right\} \cap X_i^{(k)}, \quad 1 \leqslant i \leqslant n, \quad k \geqslant 0.$$

После того как вычислена первая компонента, образуется пересечение со старым приближением, дающее новое приближение. Это новое приближение используется для вычисления нового приближения для второй компоненты и т. д. Эта модификация называется короткошаговым методом со взятием пересечения после каждой компоненты (SIC).

Наконец, для случая, когда все диагональные элементы матрицы $\mathscr{A}$ обращаются в нуль, рассмотрим (снова в предположении $x^* \subseteq x^{(0)}$) итерационную процедуру

$$X_i^{(k+1/2)} = \left\{ \sum_{j=1}^{i-1} A_{ij} X_j^{(k+1/2)} + \sum_{j=i+1}^{n} A_{ij} X_j^{(k)} + B_i \right\} \cap X_i^{(k)}, \quad 1 \leqslant i \leqslant n,$$

$$X_i^{(k+1)} = \left\{ \sum_{j=1}^{i-1} A_{ij} X_j^{(k+1/2)} + \sum_{j=i+1}^{n} A_{ij} X_j^{(k+1)} + B_i \right\} \cap X_i^{(k+1/2)},$$

$$1 \leqslant i \leqslant n, \quad k \geqslant 0.$$

Мы будем называть эту процедуру симметрическим короткошаговым методом со взятием пересечения после каждой компоненты (SSIC).

Метод (SSIC) можно выполнять таким образом, что он будет требовать на каждом шаге (за исключением самого первого) того же количества интервальных умножений, что и метод (SIC). В обоих случаях требуется $n^2 - n$ умножений (в предположении, что диагональные элементы обращаются в 0). При этом используются следующие соображения. Допустим, что для некоторого $k>0$ известны суммы

$$\sum_{j=i+1}^{n} A_{ij} X_j^{(k)}, \quad 1 \leqslant i \leqslant n-1.$$

Вычисление векторов $x^{(k+1/2)}$ по методу (SSIC) требует $\frac{1}{2}(n^2 - n)$ умножений. Если запоминается n — 1 сумма

$$\sum_{j=1}^{i-1} A_{ij} X_j^{(k+1/2)}, \quad 2 \leqslant i \leqslant n,$$

то вычисление $x^{(k+1)}$ из $x^{(k+1/2)}$ требует еще $\frac{1}{2}(n^2 - n)$ умножений.

Таким образом, всего для вычисления приближения $x^{(k+1)}$, исходя из $x^{(k)}$ нам требуется $n^2 - n$ умножений, как и в методе (SIC). Если при вычислении $x^{(k+1)}$ из $x^{(k)}$ запоминаются суммы

$$\sum_{j=i+1}^{n} A_{ij} X_j^{(k+1)},$$

то все эти рассуждения проходят для индекса, увеличенного на 1.

Следующее утверждение показывает, что итерационные методы (TI), (SI), (SIC) и (SSIC) сходятся к неподвижной точке x^*.

Теорема 1. *Пусть* $\mathcal{A}$ *—интервальная матрица и* $\rho(|\mathcal{A}|) < 1$. *Если* x^* *— неподвижная точка уравнения* $x = \mathcal{A}x + b$ *и* $x^{(0)} \supseteq x^*$, *то итерационные методы* (TI), (SI), (SIC) *и* (SSIC) *сходятся к* x^*.

Доказательство. Докажем теорему для метода (TI). Так как мы берем пересечения, полученная последовательность приближений $\{x^{(k)}\}_{k=0}^{\infty}$ удовлетворяет условию

$$x^{(0)} \supseteq x^{(1)} \supseteq \ldots \supseteq x^{(k)} \supseteq x^{(k+1)} \supseteq \ldots .$$

В силу следствия 8 из микромодуля 29 эта последовательность сходится к некоторому пределу $\tilde{x}^*$, и этот предел удовлетворяет условию $x^* \subseteq \tilde{x}^*$. Операция взятия пересечения также непрерывна (если пересечение непусто), поэтому при $k \to \infty$ мы имеем

$$\tilde{x}^* = \{\mathcal{A}\tilde{x}^* + b\} \cap \tilde{x}^*,$$

откуда следует, что

$$\mathcal{A}\tilde{x}^* + b \supseteq \tilde{x}^*.$$

Мы рассматриваем полношаговый метод

$$y^{(k+1)} = \mathcal{A}y^{(k)} + b, \quad k \geqslant 0,$$

где $y^{(0)} = \tilde{x}^*$. Из сказанного следует, что

$$y^{(1)} = \mathcal{A}y^{(0)} + b = \mathcal{A}\tilde{x}^* + b \supseteq \tilde{x}^* \supseteq x^*,$$
$$y^{(2)} = \mathcal{A}y^{(1)} + b \supseteq \mathcal{A}\tilde{x}^* + b \supseteq \tilde{x}^* \supseteq x^*$$

и вообще

$$y^{(k+1)} = \mathcal{A}y^{(k)} + b \supseteq \mathcal{A}\tilde{x}^* + b \supseteq \tilde{x}^* \supseteq x^*.$$

Последовательность $\{y^{(k)}\}_{k=0}^{\infty}$, вычисленная с помощью рассматриваемого итерационного метода, сходится к x^* в силу теоремы 1 из микромодуля 31 . Последнее из доказанных включений дает при $k \to \infty$ соотношение

$$x^* \supseteq \tilde{x}^* \supseteq x^*,$$

т. е. $\tilde{x}^* = x^*$. Для остальных методов доказательство аналогично.

Сравним методы (T), (S), (TI), (SI), (SIC) и (SSIC) по скорости сходимости для случая, когда итерации начинаются с интервального вектора, содержащего неподвижную точку x^*.

В первой теореме метод (T) сравнивается с (TI), а мегод (S) - с (SI).

Теорема 2. *Пусть последовательности* $\{x^{(k)}\}_{k=0}^{\infty}$ *и* $\{\tilde{x}^{(k)}\}_{k=0}^{\infty}$ *вычислены согласно итерационным методам* (T) *и* (TI) *в предположении, что* $x^{(0)} \supseteq \tilde{x}^{(0)} \supseteq x^*$. *Тогда имеет место*

$$x^{(k)} \supseteq \tilde{x}^{(k)} \supseteq x^* \quad \textit{для всех} \quad k \geqslant 0.$$

Такое же утверждение справедливо для последовательностей, вычисленных согласно итерационным методам (S) *и* (SI).

Доказательство. Докажем теорему для методов (T) и (TI). Соотношение $\tilde{x}^{(k)} \supseteq x^*$, $k \geqslant 0$, уже было доказано в предположении $\tilde{x}^{(0)} \supseteq x^*$ при получении формул для итерационного метода (TI). Допустим, что для некоторого $k \geqslant 0$ уже доказано, что

$$x^{(k)} \supseteq \tilde{x}^{(k)}.$$

Для $k = 0$ это верно ввиду нашего исходного допущения. Используя монотонность включения, получаем

$$x^{(k+1)} = \mathcal{A}x^{(k)} + b \supseteq \mathcal{A}\tilde{x}^{(k)} + b \supseteq \{\mathcal{A}\tilde{x}^{(k)} + b\} \cap \tilde{x}^{(k)} = \tilde{x}^{(k+1)},$$

что завершает доказательство по индукции.

Доказательство для методов (S) и (SI) можно провести аналогичным образом.

Теорема 3. *Пусть последовательности* $\{x^{(k)}\}_{k=0}^{\infty}$ *и* $\{\tilde{x}^{(k)}\}_{k=0}^{\infty}$ *вычислены согласно итерационным методам* (TI) *и* (SIC) *в предположении, что* $x^{(0)} \supseteq \tilde{x}^{(0)} \supseteq x^*$. *Тогда имеет место*

$$x^{(k)} \supseteq \tilde{x}^{(k)} \supseteq x^* \quad \textit{для всех} \quad k \geqslant 0.$$

Такое же утверждение справедливо для последовательностей, вычисленных согласно итерационным методам (S) *и* (SIC).

Доказательство. Нам нужно доказать лишь соотношение $x^{(k)} \supseteq \tilde{x}^{(k)}$. Мы проведем доказательство для последовательностей, вычисленных согласно итерационным методам (TI) и (SIC). Допустим, что для некоторого $k \geqslant 0$ верно

$$x^{(k)} \supseteq \tilde{x}^{(k)}.$$

Для k =0 это верно в силу нашего исходного допущения.

Тогда в обозначениях

$$x^{(k)} = (X_i^{(k)}),\ \tilde{x}^{(k)} = (\tilde{X}_i^{(k)}),\ \mathcal{A} = (A_{ij}),\ \mathcal{b} = (B_i)$$

мы имеем

$$X_1^{(k+1)} = \left(\sum_{j=1}^{n} A_{1j} X_j^{(k)} + B_1\right) \cap X_1^{(k)},$$

$$\tilde{X}_1^{(k+1)} = \left(\sum_{j=1}^{n} A_{1j} \tilde{X}_j^{(k)} + B_1\right) \cap \tilde{X}_1^{(k)}.$$

Из $\tilde{X}_i^{(k)} \subseteq X_i^{(k)}$, $1 \leqslant i \leqslant n$, и монотонности включения следует, что

$$\sum_{j=1}^{n} A_{1j} \tilde{X}_j^{(k)} + B_1 \subseteq \sum_{j=1}^{n} A_{1j} X_j^{(k)} + B_1,$$

а потому

$$\tilde{X}_1^{(k+1)} \subseteq X_1^{(k+1)}.$$

Ввиду $\tilde{X}_1^{(k+1)} \subseteq \tilde{X}_1^{(k)} \subseteq X_1^{(k)}$ отсюда следует, что

$$A_{21} X_1^{(k+1)} + \sum_{j=2}^{n} A_{2j} \tilde{X}_j^{(k)} + B_2 \subseteq \sum_{j=1}^{n} A_{2j} X_j^{(k)} + B_2,$$

а ввиду

$$X_2^{(k+1)} = \left(\sum_{j=1}^{n} A_{2j} X_j^{(k)} + B_2\right) \cap X_2^{(k)},$$

$$\tilde{X}_2^{(k+1)} = \left(A_{21} \tilde{X}_1^{(k+1)} + \sum_{j=2}^{n} A_{2j} \tilde{X}_j^{(k)} + B_2\right) \cap \tilde{X}_2^{(k)}$$

получаем

$$\tilde{X}_2^{(k+1)} \subseteq X_2^{(k+1)}.$$

Таким же образом мы показываем, что $\tilde{X}_i^{(k+1)} \subseteq X_i^{(k+1)}$, $3 \leqslant i \leqslant n$, т. е. $\tilde{x}^{(k+1)} \subseteq x^{(k+1)}$, что завершает доказательство по

индукции. Доказательство для последовательностей, вычисленных согласно итерационным методам (SI) и (SIC), можно провести аналогичным образом.

Теорема 4. *Пусть последовательности* $\{z^{(k)}\}_{k=0}^{\infty}$ *и* $\{x^{(k)}\}_{k=0}^{\infty}$ *вычислены согласно итерационным методам* (SIC) *и* (SSIC) *в предположении* $z^{(0)} \supseteq x^{(0)} \supseteq x^*$. *Тогда имеет место*

$$z^{(k)} \supseteq x^{(k)} \supseteq x^* \quad \text{для всех} \quad k \geqslant 0.$$

Доказательство. Допустим, что для некоторого $k \geq 0$ верно $z^{(k)} \supseteq x^{(k)} \supseteq x^*$. Для $k = 0$ это верно в силу нашего исходного допущения. Из формул для (SSIC) и первой формулы для (SIC) получаем с помощью математической индукции по индексам компонент, что

$$z^{(k+1)} \supseteq x^{(k+1/2)} \supseteq x^*.$$

С помощью второй формулы для (SSIC) получаем, еще раз применяя индукцию по индексам компонент, что

$$x^{(k+1/2)} \supseteq x^{(k+1)} \supseteq x^*.$$

Сочетание этих включений дает

$$z^{(k+1)} \supseteq x^{(k+1)} \supseteq x^*.$$

Используя доказанные выше утверждения, мы можем теперь указать оптимальный метод. Пусть (M) обозначает любой метод из множества

$$\{(T), (S), (TI), (SI), (SIC), (SSIC)\}.$$

Мы допускаем в качестве начального вектора любой интервальный вектор $x^{(0)}$, такой что $x^* \subseteq x^{(0)}$, где x^* — неподвижная точка, т. е. решение уравнения $x = \mathcal{A}x + b$. Введем частичный порядок на рассматриваемом множестве итерационных методов, полагая $(M) \leqslant (M)$, если $x^{(k)} \subseteq \tilde{x}^{(k)}$ для всех $k \geqslant 0$. Здесь $\{x^{(k)}\}_{k=0}^{\infty}$ и $\{\tilde{x}^{(k)}\}_{k=0}^{\infty}$ обозначают последовательности, вычисленные согласно методам (M) и ($\tilde{M}$). Из теоремы 2 имеем

$$(TI) \leqslant (T) \quad \text{и} \quad (SI) \leqslant (S).$$

Аналогично из теоремы 3 имеем

$$(SIC) \leqslant (TI) \quad \text{и} \quad (SIC) \leqslant (SI)$$

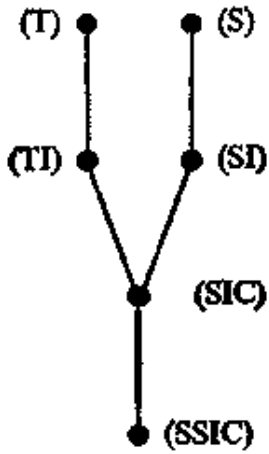
Наконец, из теоремы 4 имеем, что

$$(SSIC) \leqslant (M)$$

для любого из рассматриваемых итерационных методов (M). Объединяем эти результаты в следующее утверждение.

Теорема 5. *Пусть* $\mathcal{A}$ *— интервальная матрица, такая что* $\rho(|\mathcal{A}|) < 1$, *и* $\mathcal{b}$ *— интервальный вектор. Если начать вычисление по одному из методов* (1), (S), (TI), (SI), (SIC) *и* (SSIC) *с вектора* $x^{(0)}$, *удовлетворяющего соотношениию* $x^{(0)} \supseteq x^* = \mathcal{A}x^* + \mathcal{b}$, *то метод* (SSIC) *всегда даст наименьшую (в смысле включения) локализацию последовательности для* x^*.

Следующая диаграмма наглядно выражает содержание теоремы 5:



Чтобы проиллюстрировать теорему 5, были просчитаны различные примеры на ЭВМ (где мантисса содержит 8 десятичных цифр).

Для каждого примера приводим начальный вектор $x^{(0)} = (X_i^{(0)})$ и число итераций k^*, после которого процедура стабилизируется. Примеры показывают, что метод (SIC) требует примерно на 25% больше шагов итерации, чем метод (SSIC).

В первых двух примерах и исходные данные, и неподвижные точки — невырожденные интервалы. В этом случае приводим также и вектор $x^{(k^*)}$. В остальных примерах приводим только наибольшую ширину компоненты вектора $x^{(k)}$, т. е. величину $d^{(k)} = \max\limits_{1 \leqslant i \leqslant n} \{d(X_i^{(k)})\}$. Все примеры были приведены к виду $x = \mathcal{A}x + \mathcal{b}$ таким образом, что диагональные элементы матрицы $\mathcal{A}$ равны нулю.

I. Пример.

$$x^{(0)} = \begin{pmatrix} [0.9, & 1.2] \\ [0.4, & 0.7] \\ [0, & 0.2] \\ [-0.4, & -0.1] \end{pmatrix},$$

$$(\mathrm{SIC}): k^* = 37, \quad (\mathrm{SSIC}): k^* = 31,$$

$$x^{(k^*)} = \begin{pmatrix} [1.0328601, & 1.0597579] \\ [0.55075440, & 0.57481398] \\ [0.099483623, & 0.12251379] \\ [-0.24354582, & -0.21269841] \end{pmatrix}.$$

II. Пример.

$$x^{(0)} = \begin{pmatrix} [0.8, & 1.0] \\ [0.65, & 0.85] \\ [0.55, & 0.7] \end{pmatrix}, \qquad x^{(k^*)} = \begin{pmatrix} [0.89636817, & 0.89647991] \\ [0.76595755, & 0.76520225] \\ [0.61424734, & 0.61452184] \end{pmatrix},$$

$$(\mathrm{SIC}): k^* = 18, \quad (\mathrm{SSIC}): k^* = 15.$$

III. Пример.

$$x^{(0)} = (X_i^{(0)}), \quad \text{где} \quad X_i^{(0)} = [0, 1], \quad i = 1(1)8.$$

$$(\mathrm{SIC}): k^* = 27, \quad (\mathrm{SSIC}): k^* = 22.$$

$d^{(k)}$	k				
	0	5	10	15	20
(SIC)	1	4.0×10^{-2}	8.4×10^{-4}	1.8×10^{-5}	3.8×10^{-7}
(SSIC)	1	7.0×10^{-3}	5.4×10^{-5}	4.1×10^{-7}	3.7×10^{-9}

IV. Пример.

$$x^{(0)} = (X_i^{(0)}), \quad \text{где} \quad X_i^{(0)} = [0, 0.5], \quad 1 \leqslant i \leqslant 4.$$

$$(\mathrm{SIC}): k^* = 56, \quad (\mathrm{SSIC}): k^* = 44.$$

$d^{(k)}$	k					
	0	5	10	20	30	40
(SIC)	5.0×10^{-1}	1.1×10^{-1}	2.9×10^{-2}	6.7×10^{-4}	2.7×10^{-5}	1.5×10^{-6}
(SSIC)	5.0×10^{-1}	3.7×10^{-2}	4.7×10^{-3}	7.4×10^{-5}	1.2×10^{-6}	1.8×10^{-8}

V. Пример.

$$x^{(0)} = (X_i^{(0)}), \quad \text{где} \quad X_i^{(0)} = [-0.036016,\ 0.674056], \quad 1 \leqslant i \leqslant 3.$$

$$(\mathrm{SIC}): k^* = 52, \quad (\mathrm{SSIC}): k^* = 42.$$

$d^{(k)}$	k				
	0	5	10	20	40
(SIC)	7.1×10^{-1}	1.2×10^{-1}	2.0×10^{-2}	5.3×10^{-4}	3.7×10^{-7}
(SSIC)	7.1×10^{-1}	5.5×10^{-2}	5.7×10^{-3}	6.2×10^{-5}	3.7×10^{-9}

VI. Пример. Возьмем в примере V

$$X_i^{(0)} = [0.059459,\ 0.643243], \quad 1 \leqslant i \leqslant 3.$$

$$(\mathrm{SIC}): k^* = 51, \quad (\mathrm{SSIC}): k^* = 41$$

$d^{(k)}$	k				
	0	5	10	20	40
(SIC)	5.8×10^{-1}	1.0×10^{-1}	1.7×10^{-2}	4.3×10^{-4}	3.0×10^{-7}
(SSIC)	5.8×10^{-1}	4.5×10^{-2}	4.7×10^{-3}	5.1×10^{-5}	3.7×10^{-9}

Использование локализующих множеств важно при реализации итерационных вычислений на ЭВМ. Если в этом случае мы начинаем вычисления с вектора, представимого в машине и содержащего неподвижную точку, т. е. решение уравнения $x^* = \mathcal{A}x^* + b$, то все следующие приближения снова содержат эту неподвижную точку. Так как вычисление новых приближений искажается погрешностями округления, мы можем в действительности в какой-то момент «потерять» неподвижную точку: некоторый вновь вычисленный интервал уже не будет содержать ее. Если все операции выполняются в машинной интервальной арифметике, то свойство содержать неподвижную точку не будет потеряно. Если мы применяем метод, где берутся пересечения, то получается последовательность

$$\tilde{x}^{(0)} \supseteq \tilde{x}^{(1)} \supseteq \ldots \tilde{x}^{(k-1)} \supseteq \tilde{x}^{(k)} = \tilde{x}^{(k+1)} = \ldots.$$

Последовательность приближений, вычисленных на машине, стабилизируется, начиная с некоторого номера k^*. Это следует из того факта, что на цифровой машине представимо лишь конечное количество вещественных чисел.

Покажем, что для методов (TI), (SI), (SIC) и (SSIC) после конечного числа шагов не нужно брать пересечений. Сформулируем и докажем такую теорему для метода итераций (TI).

Теорема 6. *Пусть* $\mathcal{A}$ *— интервальная матрица, для которой* $\rho(|\mathcal{A}|) < 1$. *Пусть вектор* $x^{(0)} = (X_i^{(0)}) = (\lfloor i(X_i^{(0)}),\ s(X_i^{(0)}) \rfloor)$, *выбран таким образом, что для неподвижной точки* $x^* = (\lfloor i(X_i^*),\ s(X_i^*) \rfloor)$, *т. е. решения уравнения* $x = \mathcal{A}x + \mathcal{b}$ *выполнено включение* $x^{(0)} \supset x^*$, *которое вводится соотношениями* $i(X_i^{(0)}) < i(X_i^*) \leqslant s(X_i^*) < s(X_i^{(0)}),\ i = 1, 2, \ldots, n.$ *Тогда существует* $\tilde{k} \geqslant 0$, *такое что при всех* $k \geqslant \tilde{k}$ *для итерационного метода*

$$x^{(k+1)} = (\mathcal{A}x^{(k)} + \mathcal{b}) \cap x^{(k)}, \quad k \geqslant 0,$$

выполнено равенство

$$x^{(k+1)} = (\mathcal{A}x^{(k)} + \mathcal{b}) \cap x^{(k)} = \mathcal{A}x^{(k)} + \mathcal{b}.$$

т. е. верно включение

$$\mathcal{A}x^{(k)} + \mathcal{b} \subseteq x^{(k)}.$$

Доказательство. Мы ограничимся случаем, когда все элементы матрицы $\mathcal{A}$ и векторов $\mathcal{b}$, $x^{(0)}$ принадлежат $I(\mathbb{R})$. Случай, когда разрешены элементы из $R(\mathbb{C})$, может быть рассмотрен аналогичным образом. Сначала мы покажем, что из включения $x^{(0)} \supset x^*$ следует, что не может выполняться соотношение

$$x^{(0)} \subseteq \mathcal{A}x^{(0)} + \mathcal{b}.$$

Действительно, если бы оно выполнялось, то из формул

$$z^{(k+1)} = \mathcal{A}z^{(k)} + \mathcal{b}, \quad k \geqslant 0,$$

определяющих наш метод итераций, следовало бы при $z^{(0)}=x^{(0)}$, что

$$z^{(1)} = \mathcal{A}z^{(0)} + \mathcal{b} = \mathcal{A}x^{(0)} + \mathcal{b} \supseteq x^{(0)} \supset x^*,$$
$$z^{(2)} = \mathcal{A}z^{(1)} + \mathcal{b} \supseteq \mathcal{A}x^{(0)} + \mathcal{b} \supseteq x^{(0)} \supset x^*$$

и вообще

$$z^{(k+1)} = \mathcal{A}z^{(k)} + \mathcal{b} \supseteq \mathcal{A}x^{(0)} + \mathcal{b} \supseteq x^{(0)} \supset x^*, \quad k \geqslant 0.$$

Так как $\rho(|\mathcal{A}|) < 1$, мы имели бы тогда

$$\lim_{k \to \infty} z^{(k)} = z^*,$$

где $z^* = \mathcal{A}z^* + \mathcal{b}$.

Из последнего соотношения следует, что $z^* \supseteq x^{(0)} \supset x^*$. Это противоречит единственности неподвижной точки, т. е. решения уравнения $x = \mathcal{A}x + \mathcal{b}$. Теперь полагаем

$$x^{(k)} = (X_i^{(k)}), \qquad \mathcal{A} = (A_{ij}), \qquad \mathcal{b} = (B_i), \qquad y^{(k)} = (Y_i^{(k)}),$$

где

$$Y_i^{(k+1)} = \sum_{j=1}^{n} A_{ij} X_j^{(k)} + B_i, \qquad k \geqslant 0, \qquad 1 \leqslant i \leqslant n.$$

Из только что установленного факта следует, что найдется номер i, $1 \leqslant i \leqslant n$, такой что имеет место в точности одна из следующих двух возможностей:

(a) $Y_i^{(1)} \subset X_i^{(0)}$, т. е. $X_i^{(1)} = Y_i^{(1)} \cap X_i^{(0)} = Y_i^{(1)}$;

(b) $Y_i^{(1)} \not\subset X_i^{(0)}$ и $X_i^{(1)} = Y_i^{(1)} \cap X_i^{(0)} \subset X_i^{(0)}$.

В случае (a) мы получаем ввиду $x^{(0)} \supseteq x^{(1)}$ и монотонности включения, что

$$Y_i^{(2)} = \sum_{j=1}^{n} A_{ij} X_j^{(1)} + B_i \subseteq \sum_{j=1}^{n} A_{ij} X_j^{(0)} + B_i = Y_i^{(1)} = X_i^{(1)},$$
$$X_i^{(2)} = Y_i^{(2)} \cap X_i^{(1)} = Y_i^{(2)},$$

а в общем случае ввиду $x^{(k)} \supseteq x^{(k+1)}$ методом математической индукции получаем

$$Y_i^{(k+1)} = \sum_{j=1}^{n} A_{ij} X_j^{(k)} + B_i \subseteq \sum_{j=1}^{n} A_{ij} X_j^{(k-1)} + B_i = Y_i^{(k)} = X_i^{(k)},$$
$$X_i^{(k+1)} = Y_i^{(k+1)} \cap X_i^{(k)} = Y_i^{(k+1)}.$$

(b). Полагая $i(A) = a_1, s(A) = a_2$ для $A = [a_1, a_2] \in I(\mathbb{R})$, мы можем, не умаляя общности, считать, что

$$i(Y_i^{(1)}) < i(X_i^{(0)}) \leqslant s(Y_i^{(1)}) < s(X_i^{(0)}),$$

т. е.

$$X_i^{(1)} = [i(X_i^{(0)}),\ s(Y_i^{(1)})].$$

(Возможен еще случай

$$i(X_i^{(0)}) < i(Y_i^{(1)}) \leqslant s(X_i^{(0)}) < s(Y_i^{(1)}),$$

т. е.

$$X_i^{(1)} = [i(Y_i^{(1)}),\ s(X_i^{(0)})],$$

но он рассматривается аналогично.) Так как $x^{(0)} \supseteq x^{(1)}$, мы

имеем

$$Y_i^{(2)} = \sum_{j=1}^{n} A_{ij} X_j^{(1)} + B_i \subseteq \sum_{j=1}^{n} A_{ij} X_j^{(0)} + B_i = Y_i^{(1)},$$

т. е.

$$i\left(Y_i^{(1)}\right) \leqslant i\left(Y_i^{(2)}\right),\ s\left(Y_i^{(2)}\right) \leqslant s\left(Y_i^{(1)}\right) = s\left(X_i^{(1)}\right),$$

а также

$$X_i^{(2)} = \left[\max\left\{i\left(Y_i^{(2)}\right),\ i\left(X_i^{(0)}\right)\right\},\ s\left(Y_i^{(2)}\right)\right].$$

Так как $x^{(k)} \supseteq x^{(k+1)}$, то можно показать методом математической индукции, что

$$i\left(Y_i^{(k)}\right) \leqslant i\left(Y_i^{(k+1)}\right),\ s\left(Y_i^{(k+1)}\right) \leqslant s\left(Y_i^{(k)}\right) = s\left(X_i^{(k)}\right), \qquad k \geqslant 1,$$

т. е.

$$X_i^{(k+1)} = \left[\max\left\{i\left(Y_i^{(k+1)}\right),\ i\left(X_i^{(0)}\right)\right\},\ s\left(Y_i^{(k+1)}\right)\right].$$

По предположению мы имеем $i\left(X^{(0)}\right) < i\left(X_i^*\right)$, а по теореме 1 $\lim\limits_{k\to\infty} i\left(X_i^{(k)}\right) = i\left(X_i^*\right)$. Поэтому найдется $k_0 \geq 1$, такое что имеет место

$$\max\left\{i\left(Y_i^{(k_0+1)}\right),\ i\left(X_i^{(0)}\right)\right\} = i\left(Y_i^{(k_0+1)}\right)$$

или

$$X_i^{(k_0+1)} = \left[i\left(Y_i^{(k_0+1)}\right),\ s\left(Y_i^{(k_0+1)}\right)\right],$$

т. е.

$$X_i^{(k_0+1)} = Y_i^{(k_0+1)} \cap X_i^{(k_0)} = Y_i^{(k_0+1)}.$$

Метод математической индукции позволяет теперь установить, что

$$X_i^{(k_0+\nu)} = Y_i^{(k_0+\nu)} \cap X_i^{(k_0+\nu-1)} = Y_i^{(k_0+\nu)}, \qquad \nu \geqslant 1,$$

так как $x^{(k)} \supseteq x^{(k+1)}$. Ввиду соотношений

$$\lim_{k\to\infty} x^{(k)} = x^* \text{ и } x^{(0)} \supset x^*$$

получаем, что для любого i, $1 \leqslant i \leqslant n$, не удовлетворяющего ни одному из условий (a), (b), хотя бы одно из этих условий выполнится после нескольких следующих шагов итерации. Наконец, получаем

$$x^{(k+1)} = (\mathcal{A}x^{(k)} + b) \cap x^{(k)} = \mathcal{A}x^{(k)} + b, \qquad k \geqslant \tilde{k} \geqslant 0.$$

Замечания.

Теоремы этого микромодуля без всяких изменений переносятся на соответствующие итерационные методы нахождения неподвижной точки, т. е. решения нелинейного уравнения

$$x = f_p(x),$$

где

$$f_p(x_p) = (f_i(x_1, \ldots, x_n; a_{i1}, \ldots, a_{im_i}))$$

есть $\mathcal{P}_p$-сжатие.

Самое существенное для доказательства этих теорем свойство — монотонность включения.

Микромодуль 34

О применимости метода Гаусса к системам уравнений с интервальными коэффициентами

Пусть $\mathcal{A}$ — интервальная матрица, b — интервальный вектор. Будем предполагать, что обращение $\mathcal{A}_p^{-1}$ существует для всех $\mathcal{A}_p \in \mathcal{A}$. Мы хотим найти множество

$$\Omega = \{x_p \mid \mathcal{A}_p x_p = b_p,\ \mathcal{A}_p \in \mathcal{A},\ b_p \in b\}.$$

Это множество в общем случае не имеет простого описания. Поэтому мы ограничимся его локализацией с помощью интервального вектора. Очевидный способ нахождения такого интервального вектора — применение непосредственного обобщения метода Гаусса на системы с интервальными коэффициентами. Иными словами, пусть нам дана таблица коэффициентов

$$\begin{array}{ccc} A_{11} & \ldots & A_{1n} \quad B_1 \\ \cdot & & \cdot \quad \cdot \\ \cdot & & \cdot \quad \cdot \\ \cdot & & \cdot \quad \cdot \\ A_{n1} & \ldots & A_{nn} \quad B_n. \end{array}$$

Применяя формулы

$$A'_{1j} = A_{1j}, \qquad 1 \leqslant j \leqslant n, \qquad B'_1 = B_1,$$
$$A'_{ij} = A_{ij} - A_{1j}(A_{i1}/A_{11}), \qquad 2 \leqslant i,\ j \leqslant n,$$
$$B'_i = B_i - B_1(A_{i1}/A_{11}), \qquad 2 \leqslant i \leqslant n,$$
$$A'_{i1} = 0, \qquad 2 \leqslant i \leqslant n,$$

в предположении $0 \notin A_{11}$, мы вычисляем новую таблицу коэффициентов

$$\begin{matrix} A'_{11} & A'_{12} & \dots & A'_{1n} & B'_1 \\ 0 & A'_{22} & \dots & A'_{2n} & B'_2 \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ 0 & A'_{n2} & \dots & A'_{nn} & B'_n. \end{matrix}$$

Покажем теперь, что имеет место

$$\{x_p \mid \mathscr{A}_p x_p = \mathscr{b}_p,\ \mathscr{A}_p \in \mathscr{A},\ \mathscr{b}_p \in \mathscr{b}\} \subseteq \{y_p \mid \mathscr{A}'_p y_p = \mathscr{b}'_p,$$
$$\mathscr{A}'_p \in \mathscr{A}',\ \mathscr{b}'_p \in \mathscr{b}'\}.$$

Допустим, что $\mathscr{A}_p = \mathscr{A}$ и $\mathscr{b}_p \in \mathscr{b}$, и рассмотрим систему линейных уравнений

$$\mathscr{A}_p x_p = \mathscr{b}_p.$$

Строим матрицу $\mathscr{A}'_p = (a'_{ij})$ и вектор $\mathscr{b}'_p = (b'_i)$, где

$$a'_{1j} = a_{1j}, \qquad 1 \leqslant j \leqslant n, \qquad b'_1 = b_1,$$

и

$$a'_{ij} = a_{ij} - a_{1j}(a_{i1}/a_{11}), \qquad 2 \leqslant i,\ j \leqslant n,$$
$$b'_i = b_i - b_1(a_{i1}/a_{11}), \qquad 2 \leqslant i \leqslant n,$$
$$a'_{i1} = 0, \qquad 2 \leqslant i \leqslant n.$$

Из теории линейных уравнений хорошо известно, что система уравнений $\mathscr{A}'_p y_p = \mathscr{b}'_p$ имеет те же решения, что и $\mathscr{A}_p x_p = \mathscr{b}_p$. Из монотонности включения следует $\mathscr{A}'_p \in \mathscr{A}'$ и $\mathscr{b}'_p \in \mathscr{b}'$, что и доказывает наше утверждение. Если описанный выше шаг проведен n—1 раз, то исходная таблица коэффициентов превращается в верхнюю треугольную матрицу

$$\begin{matrix} \tilde{A}_{11} & \tilde{A}_{12} & \dots & \tilde{A}_{1n} & \tilde{B}_1 \\ & \tilde{A}_{22} & & \cdot & \cdot \\ & & \cdot & \cdot & \cdot \\ & & & \cdot & \cdot \\ & & & \tilde{A}_{nn} & \tilde{B}_n \end{matrix},$$

для которой имеет место

$$\{x_p \mid \mathscr{A}_p x_p = \mathscr{b}_p,\ \mathscr{A}_p \in \mathscr{A},\ \mathscr{b}_p \in \mathscr{b}\}$$
$$\subseteq \{\tilde{x}_p \mid \tilde{\mathscr{A}}_p \tilde{x}_p = \tilde{\mathscr{b}}_p, \tilde{\mathscr{A}}_p \in \tilde{\mathscr{A}},\ \tilde{\mathscr{b}}_p \in \tilde{\mathscr{b}}\}.$$

Используя формулы

$$X_n = \tilde{B}_n / \tilde{A}_{nn},$$

$$X_i = \left(\tilde{B}_i - \sum_{l=i+1}^{n} \tilde{A}_{il} X_l \right) / \tilde{A}_{ii}, \quad 1 \leqslant i \leqslant n-1,$$

получаем тогда интервальный вектор $x = (X_i)$, удовлетворяющий условию

$$\{x_\mathrm{p} \mid \mathscr{A}_\mathrm{p} x_\mathrm{p} = b_\mathrm{p},\ \mathscr{A}_\mathrm{p} \in \mathscr{A},\ b_\mathrm{p} \in b\} \subseteq x.$$

В частности, если $\mathscr{A}_\mathrm{p} = (a_{ij})$— невырожденная точечная матрица, то метод Гаусса применим, когда в правой части стоит произвольный интервальный вектор. При этом в процессе исключения по Гауссу может потребоваться перестановка столбцов. Это эквивалентно умножению матрицы $\mathscr{A}_\mathrm{p}$ слева на матрицу перестановки перед началом процесса исключения.

Теперь определим отображение

$$g_\mathrm{p}:\ M_{nn}(\mathbb{C}) \times V_n(\mathbb{C}) \to V_n(\mathbb{C})$$

для невырожденной матрицы $\mathscr{A}_\mathrm{p}$ и точечного вектора b_p. Это отображение представляет собой применение метода Гаусса к системе линейных уравнений

$$\mathscr{A}_\mathrm{p} x_\mathrm{p} = b_\mathrm{p},$$

дающее результат

$$x_\mathrm{p} = g_\mathrm{p}(\mathscr{A}_\mathrm{p},\ b_\mathrm{p}).$$

Отображение g_p единственно, но, как обычно, для g_p имеются различные выражения. Например, мы имеем $\mathscr{A}_\mathrm{p}^{-1} b_\mathrm{p} = g_\mathrm{p}(\mathscr{A}_\mathrm{p},\ b_\mathrm{p})$. Кроме того, метод Гаусса дает различные выражения для g_p в зависимости от выбора главных элементов.

Следующие свойства не зависят от выбора главных элементов. Интервальное выражение для g_p обозначается через $g_\mathrm{p}(\mathscr{A},\ b)$. Поэтому интервальный вектор x, получаемый после выполнения описанного выше метода Гаусса, можно задать равенством $x = g_\mathrm{p}(\mathscr{A},\ b)$.

Имеем следующие свойства:

$$\mathscr{A},\ \mathscr{B} \in M_{nn}(I(\mathbb{C})), \quad a,\ b \in V_n(I(\mathbb{C})), \tag{1}$$
$$\mathscr{A} \subseteq \mathscr{B}, \quad a \subseteq b.$$

Отсюда следует, что

$$g_\mathrm{p}(\mathscr{A},\ a) \subseteq g_\mathrm{p}(\mathscr{B},\ b).$$
$$\mathscr{A}_\mathrm{p} \in M_{nn}(\mathbb{C}), \quad b = u + v \in V_n(I(\mathbb{C})); \tag{2}$$

Отсюда следует, что

$$g_p(\mathcal{A}_p, \mathcal{b}) = g_p(\mathcal{A}_p, u) + g_p(\mathcal{A}_p, v).$$
$$\mathcal{A}_p \in M_{nn}(\mathbb{R}), \quad \mathcal{b} \in V_n(I(\mathbb{R})). \tag{3}$$

Отсюда следует, что

$$\mathcal{A}_p^{-1}\mathcal{b} \subseteq g_p(\mathcal{A}_p, \mathcal{b}).$$
$$\mathcal{A}_p \in M_{nn}(\mathbb{C}), \quad a, \ \mathcal{b} \in V_n(I(\mathbb{C})), \quad d(a) \leqslant \alpha d(\mathcal{b}) \tag{4}$$
$$\text{для некоторого } \alpha \geqslant 0.$$

Поэтому для ширины имеет место

$$d(g_p(\mathcal{A}_p, a)) \leqslant \alpha d(g_p(\mathcal{A}_p, \mathcal{b})).$$

По поводу доказательства этих свойств заметим, что (1) сразу следует из монотонности включения, а (2) — из соотношения $a(B+C) = aB + aC$ и формул, определяющих метод Гаусса. Чтобы доказать (3), используем следующий факт.

Если имеются два рациональных выражения f_1 и f_2 для одной и той же функции $f\colon \mathbb{R} \to \mathbb{R}$, причем f_1 содержит переменную x ровно один раз, а f_2 содержит эту переменную m раз, то для вычислений f_1 и f_2 в интервальной арифметике имеет место $f_1(X) \subseteq f_2(X)$. Аналогичное утверждение верно и для функций от нескольких переменных. Рассмотрим теперь i-е, $1 \leqslant i \leqslant n$, компоненты векторов $\mathcal{A}_p^{-1}\mathcal{b}$ и $g_p(\mathcal{A}_p, \mathcal{b})$. Для точечных векторов $\mathcal{b}_p$ имеет место $(\mathcal{A}_p^{-1}\mathcal{b}_p)_i = (g_p(\mathcal{A}_p, \mathcal{b}_p))_i$. Из формул, определяющих метод Гаусса, видно, что компоненты вектора $\mathcal{b}_p$ могут входить несколько раз в выражение $(g_p(\mathcal{A}_p, \mathcal{b}_p))_i$. Так как они входят всего один раз в $(\mathcal{A}_p^{-1}\mathcal{b}_p)_i$, мы получаем (3).

Наконец, (4) получается путем использования формул, описывающих метод Гаусса, правил (10 п.7.2), (14 п.7.2), (12 п.7.5) и (16 п.7.5), а также предположения $d(a_p) \leqslant \alpha d(\mathcal{b}_p)$.

Формулы, описывающие метод Гаусса, применимы только к таким интервальным матрицам, для которых условие $0 \notin A_{ii}$ выполняется на всех шагах приведения к верхней треугольной форме. Если это не так при некотором i, $1 \leqslant i \leqslant n-1$, т. е. $0 \in A_{ii}$, то все еще возможно, что подходящая перестановка столбцов позволит избежать соотношения $0 \in A_{ii}$. Однако это не всегда возможно, даже если предположить, что сначала $\mathcal{A}_p^{-1}$ существует для всех $\mathcal{A}_p \in \mathcal{A}$. Причина заключается в следующем.

Очевидно, что в предположении существования $\mathscr{A}_p^{-1}$ для всех $\mathscr{A}_p \in \mathscr{A}$ мы всегда можем выполнить первый шаг метода Гаусса. Если бы мы не смогли сделать этот шаг из-за того, что все элементы первого столбца содержат нуль, то исходная интервальная матрица $\mathscr{A}$ содержала бы вырожденную точечную матрицу $\mathscr{A}_p$, а это противоречит нашему предположению. Рассмотрим теперь матрицу $\mathscr{A}^{(1)} = (A_{ij}^{(1)})$ размерности $(n—1)\times((n—1)$, для которой $A_{ij}^{(1)} = A'_{ij}$, $2 \leqslant i,\ j \leqslant n$. Пусть $\mathscr{U}$ обозначает интервальную матрицу

$$\mathscr{U} = \begin{pmatrix} -A_{21}/A_{11} & 1 & & 0 \\ -A_{31}/A_{11} & 0 & 1 & \\ \vdots & & & \ddots \\ -A_{n1}/A_{11} & 0 & \cdots & 0 \quad 1 \end{pmatrix},$$

размерности $(n— 1)\times n$, а $\mathscr{V}$—интервальную матрицу

$$\mathscr{V} = \begin{pmatrix} A_{12} & \cdots & A_{1n} \\ \vdots & & \vdots \\ A_{n2} & \cdots & A_{nn} \end{pmatrix}.$$

размерности $n\times (n—1)$. Мы имеем $\mathscr{A}^{(1)} = \mathscr{U}\mathscr{V}$. Аналогичным образом положим

$$\mathscr{R}_p = \begin{pmatrix} -a_{21}/a_{11} & 1 & & 0 \\ -a_{31}/a_{11} & 0 & 1 & \\ \vdots & & & \ddots \\ -a_{n1}/a_{11} & 0 & \cdots & 0 \quad 1 \end{pmatrix},$$

$$\mathscr{S}_p = \begin{pmatrix} a_{12} & \cdots & a_{1n} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ a_{n2} & \cdots & a_{nn} \end{pmatrix},$$

где $a_{ij} \in A_{ij}$, $1 \leqslant i,\ j \leqslant n$. Мы получаем, что матрица $\mathscr{Z}_p = \mathscr{R}_p\mathscr{S}_p$ невырожденная, так как $\mathscr{A}_p \in \mathscr{A}$ невырожденная. Поэтому в первом столбце матрицы $\mathscr{Z}_p$ найдется хотя бы один ненулевой элемент. Из монотонности включения следует, что

$$\{\mathscr{Z}_p = \mathscr{R}_p\mathscr{S}_p \mid \mathscr{A}_p \in \mathscr{A}\} \subseteq \{\mathscr{A}_p^{(1)} \mid \mathscr{A}_p^{(1)} \in \mathscr{A}^{(1)} = \mathscr{U}\mathscr{V}\},$$

и в общем случае здесь нет равенства, как показывает простой пример из микромодуля 29. Поэтому нет гарантии, что в первом столбце матрицы $\mathcal{A}^{(1)}$ будет хотя бы один элемент, не содержащий нуля.

Те же рассуждения проходят и для следующих шагов метода. Покажем это с помощью следующего простого примера.

Пример. Рассмотрим интервальную матрицу

$$\mathcal{A} = \begin{pmatrix} [1, 5] + i[-1, 1] & 1 \\ 25 & [-1, 1] + i[-1, 1] \end{pmatrix}.$$

Покажем сначала, что $\mathcal{A}$ не содержит вырожденной точечной матрицы. Так как

$$\det \mathcal{A}_p = a_{11}a_{22} - a_{21}a_{12} \in A_{11}A_{22} - A_{21}A_{12},$$

получаем для любой точечной матрицы $\mathcal{A}_p \in \mathcal{A}$, что

$$\det \mathcal{A}_p \in [-31, -19] + i[-6, 6].$$

Из того что интервал в правой части не содержит нуля, следует, что $\mathcal{A}_p$ невырожденная.

Теперь шаг исключения согласно методу Гаусса порождает (в силу определения 3 п.7.4 и следующего за ним замечания) интервал

$$A'_{22} = \left[-126, \frac{1}{26}\right] + i[-26, 26].$$

Поэтому невозможно продолжить применение метода Гаусса. Даже если мы отделим в матрице $\mathcal{A}$ вещественную и мнимую части, нам все равно не удастся решить получившуюся систему из 4 интервальных уравнений с 4 неизвестными с помощью метода Гаусса, хотя исходная интервальная матрица снова не содержит вырожденных точечных матриц.

Покажем теперь, что метод Гаусса всегда можно выполнить при $n \leq 2$, если коэффициенты — вещественные интервалы и все точечные матрицы $\mathcal{A}_p \in \mathcal{A}$ невырожденные.

Теорема 1. *Пусть* $1 \leq n \leq 2$ *интервальная матрица* $\mathcal{A}_p = (A_{ij})$ *размерности* $n \times n$ *не содержит невырожденных матриц* $\mathcal{A}_p$. *Тогда можно выполнить метод Гаусса.*

Доказательство. При $n = 1$ наше предположение означает, что $0 \notin A_{11}$, и в этом случае наша теорема доказана для $\mathcal{A} = (A_{11})$. При $n = 2$ хотя бы один из интервалов A_{11}, A_{21} не содержит нуля. Если бы это было не так, то существовала бы вырожденная матрица $\mathcal{A}_p \in \mathcal{A}$, что противоречит условию теоремы. Не умаляя общности, считаем, что $0 \notin A_{11}$; если это не так, переставим строки матрицы $\mathcal{A}$. Теперь метод Гаусса дает

$$A'_{22} = A_{22} - (1/A_{11})\, A_{21} A_{12}.$$

Мы можем рассматривать A'_{22} как оценивание рациональной функции a'_{22} от четырех переменных a_{11}, a_{21}, a_{12} и a_{22} в интервальной арифметике, определяемое формулой

$$a'_{22}(a_{11}, a_{12}, a_{21}, a_{22}) = a_{22} - (1/a_{11})\, a_{21} a_{12}.$$

По условию теоремы мы имеем для любой $\mathscr{A}_p \in \mathscr{A}$ соотношение

$$\det(\mathscr{A}_p) = a_{11}a_{22} - a_{21}a_{12} \neq 0,$$

откуда

$$a'_{22}(a_{11}, a_{12}, a_{21}, a_{22}) = (1/a_{11}) \det(\mathscr{A}_p) \neq 0.$$

Приведенное интервальное оценивание даст точную локализацию, если заменить a_{11} на A_{11}, a_{12} на A_{12}, a_2^1 на A_2^1 и a_{22} на A_{22}, так как каждая переменная входит в выражение для a'_{22} всего один раз. Поэтому мы имеем $0 \notin A'_{22}$, что и означает возможность выполнения метода Гаусса.

Данное выше доказательство не обобщается на случай $n \geq 3$. Даже для $n = 2$ теорема будет неверна, если элементы интервальных матриц берутся из $R(C)$ или из $K(C)$.

Рассмотрим теперь один конкретный класс интервальных матриц, для которого метод Гаусса всегда может быть выполнен. В этом классе можно даже не переставлять строки или столбцы. В дальнейшем мы ограничимся системами уравнений, где элементы матрицы коэффициентов и правой части принадлежат множеству вещественных интервалов или комплексных круговых интервалов. Системы уравнений, элементами которых являются комплексные прямоугольные интервалы, можно свести к первому из упомянутых случаев, отделяя вещественную и мнимую части.

Чтобы объединить приводимые ниже доказательства для вещественных интервалов и комплексных круговых интервалов, мы заметим, что вещественный интервал вида $A = [a_1, a_2]$ можно представить в виде

$$A = [a - r, a + r],$$

где

$$a = \tfrac{1}{2}(a_1 + a_2), \qquad r = \tfrac{1}{2}\, d(A) = \tfrac{1}{2}(a_2 - a_1).$$

Здесь a — центр интервала, а r — радиус, т. е. половина ширины. Мы вводим обозначение

$$A = [a - r, a + r] =: \langle a, r \rangle,$$

чтобы подчеркнуть аналогию с комплексными круговыми интервалами. Арифметические операции на вещественных интервалах также можно определить через центр и ширину. Пусть $A=\langle a, r\rangle$, $B=\langle b, s\rangle$. Тогда сложение и вычитание интервалов A и B можно записать в виде

$$A \pm B=\langle a \pm b,\ r+s\rangle.$$

Формально это соответствует сложению и вычитанию комплексных круговых интервалов. Для умножения нам нужно только равенство

$$[-r,\ r][-s,\ s]=\langle 0,\ r\rangle\langle 0,\ s\rangle=\langle 0,\ rs\rangle.$$

Если мы заметим теперь, что при $0 \notin A=[a_1,\ a_2]$ имеет место представление

$$\frac{1}{A}=\left[\frac{1}{a+r},\ \frac{1}{a-r}\right]=\left[\frac{a}{a^2-r^2}-\frac{r}{a^2-r^2},\ \frac{a}{a^2-r^2}+\frac{r}{a^2-r^2}\right],$$

то получим

$$\frac{1}{A}=\left\langle \frac{a}{a^2-r^2},\ \frac{r}{a^2-r^2}\right\rangle.$$

Формально это соответствует обращению круговою комплексного интервала. Для представления вещественных интервалов в виде $A=\langle a, r\rangle$ верно также

$$|A|=\max\{|a_1|, |a_2|\}=|a|+r.$$

Кроме того, имеем

$$0 \notin A \Leftrightarrow |a|-r>0.$$

Наконец, имеем

$$A=\langle a,\ r\rangle \subseteq \langle 0,\ |A|\rangle=\langle 0,\ |a|+r\rangle.$$

Докажем сначала следующее утверждение.

Лемма 2. *Пусть*

$$A=\langle a, r_1\rangle,\ B=\langle b, r_2\rangle,\ C=\langle c, r_3\rangle \text{ и } D=\langle d, r_4\rangle$$

— вещественные интервалы или круговые комплексные интервалы, причем $0 \notin D$. *Тогда для*

$$Z=\langle z,\ r_5\rangle=A-(1/D)BC$$

справедливо неравенство

$$|a|-r_1-|B||C|/(|d|-r_4) \leqslant |z|-r_5.$$

Доказательство. Из монотонности включения следует, что

$$Z = \langle z,\ r_5\rangle = A - BC\frac{1}{D} \subseteq A - \langle 0,\ |B|\rangle\langle 0,\ |C|\rangle\left\langle \frac{\bar{d}}{d\bar{d} - r_4^2},\ \frac{r_4}{d\bar{d} - r_4^2}\right\rangle$$

$$= \langle a,\ r_1\rangle - \left\langle 0,\ |B||C|\frac{|d|}{|d\bar{d} - r_4^2|} + |B||C|\frac{r_4}{d\bar{d} - r_4^2}\right\rangle$$

$$= \langle a,\ r_1\rangle - \left\langle 0,\ |B||C|\frac{1}{|d| - r_4}\right\rangle$$

$$= \left\langle a,\ r_1 + |B||C|\frac{1}{|d| - r_4}\right\rangle =: \langle a,\ r_6\rangle.$$

Отсюда и из $Z \subseteq \langle a,\ r_6\rangle$ следует, что

$$|a| - |z| \leqslant |a - z| \leqslant r_6 - r_5$$

или

$$|z| - r_5 \geqslant |a| - r_1 - |B||C|(1/(|d| - r_4)).$$

Чтобы сформулировать следующее утверждение, нам нужно понятие *M*-матрицы. Мы применим здесь эквивалентное определение. Вещественная матрица $\mathscr{B}_p = (b_{ij})$ называется *M*-матрицей, если выполнены условия

$$b_{ij} \leqslant 0, \quad i \neq j, \tag{1}$$

$$\mathscr{B}_p^{-1} \geqslant O_p. \tag{2}$$

В (2) подразумевается покомпонентный частичный порядок. Условие (2) можно заменить следующим условием:

Существует вещественный вектор $u_p = (u_i)$, такой что $u_i > 0,\ 1 \leqslant i \leqslant n$ и $\mathscr{B}_p u_p > o_p$. (2')

Этот факт, а также то обстоятельство, что диагональные элементы *M*-матрицы положительны, используется в следующем утверждении.

Теорема 3. *Пусть* $\mathscr{A} = (A_{ij})$ *— интервальная матрица, причем* $A_{ij} = \langle a_{ij},\ r_{ij}\rangle,\ 1 \leqslant i,\ j \leqslant n,$ *и пусть*

$$\mathscr{B}_p = (b_{ij})$$

— вещественная матрица, определенная соотношением

$$b_{ij} = \begin{cases} |a_{ij}| - r_{ij}, & i = j, \\ -|A_{ij}| & \text{в противном случае.} \end{cases}$$

Если $\mathscr{B}_p$ *является M-матрицей, то для матрицы* $\mathscr{A}_p$ *можно выполнить алгоритм Гаусса без перестановки строк или столбцов.*

Доказательство. Предположение, что $\mathscr{B}_p$ является *M*-матрицей, означает существование вектора $u_p = (u_i)$ с положительными элементами, такого что $\mathscr{B}_p u_p > o_p$, т. е. верно

$$(|a_{ii}|-r_{ii})u_i > \sum_{j=1,\ j\neq i}^{n} |A_{ij}|u_j, \qquad 1\leqslant i\leqslant n.$$

Ввиду неравенства $|a_{11}|-r_{11}>0$ выполнено условие $0\notin A_{11}$ и применимы формулы из первой части этого микромодуля. Теперь мы покажем, что условия теоремы выполнены и для интервальной матрицы $\tilde{A}'=(\tilde{A}'_{ij})$ размерности $(n-1)\times(n-1)$, где

$$\tilde{A}'_{ij}=A'_{ij}=\langle a'_{ij},\ r'_{ij}\rangle,\ 2\leqslant i,\ j\leqslant n.$$

Это позволит немедленно завершить доказательство теоремы с помощью математической индукции.

Для $i\geq 2$ имеет место

$$\sum_{j=2,\ j\neq i}^{n} |A'_{ij}|u_j = \sum_{j=2,\ j\neq i}^{n} \left|A_{ij}-A_{i1}\frac{A_{1j}}{A_{11}}\right|u_j$$

$$\leqslant \sum_{j=2,\ j\neq i}^{n} |A_{ij}|u_j + |A_{i1}|\left|\frac{1}{A_{11}}\right| \sum_{j=2,\ j\neq i}^{n} |A_{1j}|u_j.$$

С помощью неравенства

$$\sum_{j=2,\ j\neq i}^{n} |A_{1j}|u_j < (|a_{11}|-r_{11})u_1 - |A_{i1}|u_i,$$

справедливого в силу условия теоремы, можно получить оценки

$$\sum_{j=2,\ j\neq i}^{n} |A'_{ij}|u_j \leqslant \sum_{j=2,\ j\neq i}^{n} |A_{ij}|u_j$$

$$+|A_{i1}|\frac{1}{|a_{11}|-r_{11}}\{(|a_{11}|-r_{11})u_1-|A_{i1}|u_i\}$$

$$=\sum_{j=1,\ j\neq i}^{n} |A_{ij}|u_j - \frac{|A_{i1}||A_{1i}|}{|a_{11}|-r_{11}}u_i$$

$$<u_i\left(|a_{ii}|-r_{ii}-\frac{|A_{i1}||A_{1i}|}{|a_{11}|-r_{11}}\right)\leqslant(|a'_{ii}|-r'_{ii})u_i,$$

где последнее неравенство получается по лемме 2. Это завершает доказательство.

Следующее определение вводит важный класс интервальных матриц, удовлетворяющих условиям теоремы 3.

Определение 4. Говорят, что интервальная матрица $\mathcal{A}=(A_{ij})$, компоненты которой $A_{ij}=\langle a_{ij},\ r_{ij}\rangle$ являются вещественными интервалами или круговыми комплексными интервалами, имеет

сильно доминирующую диагональ (или что ее диагональ сильно доминирует), если

$$|a_{ii}| - r_{ii} > \sum_{j=1,\ j\neq i}^{n} |A_{ij}|, \qquad 1 \leqslant i \leqslant n.$$

Очевидно, что элементы сильно доминирующей диагонали не содержат нулей и что для любой точечной матрицы $\hat{\mathcal{A}}_p = (\hat{a}_{ij}) \in \mathcal{A}$ выполнено соотношение

$$|\hat{a}_{ii}| > \sum_{j=1,\ j\neq i}^{n} |\hat{a}_{ij}|, \qquad 1 \leqslant i \leqslant n.$$

Поэтому любая точечная матрица $\hat{\mathcal{A}}_p \in \mathcal{A}$ имеет сильно доминирующую диагональ в обычном смысле и потому невырождена.

Для матрицы с сильно доминирующей диагональю можно выполнить условия теоремы 3, если взять вектор $u_p = (u_i)$, такой что $1 \leq i \leq n$. Мы доказали следующее утверждение.

Следствие 5. *Пусть интервальная матрица $\mathcal{A}$ имеет сильно доминипиющую диагональ Тогда метод Гаусса можно выполнить для $\mathcal{A}$ без перестановки строк или столбцов.*

Требование строгого доминирования диагонали можно ослабить, сохранив применимость метода Гаусса, если данная интервальная матрица имеет вид

$$\mathcal{A} = \begin{pmatrix} A_1 & C_1 & & 0 \\ B_2 & A_2 & C_2 & \\ & \ddots & \ddots & \ddots \\ 0 & & B_n & A_n \end{pmatrix};$$

т. е. является трехдиагональной интервальной матрицей. Мы предположим еще, что $C_i \neq 0$, $1 \leqslant i \leqslant n-1$ и $B_i \neq 0$, $2 \leqslant i \leqslant n$, так как в противном случае задача распадается на меньшие задачи, для которых эти условия выполнены.

Теорема 6. *Пусть $\mathcal{A}$ — трехдиагональная интервальная матрица, такая что*

$$A_i = \langle a_i, r_i \rangle, \qquad 1 \leqslant i \leqslant n,$$
$$B_i = \langle b_i, s_i \rangle \neq 0, \qquad 2 \leqslant i \leqslant n,$$
$$C_i = \langle c_i, t_i \rangle \neq 0, \qquad 1 \leqslant i \leqslant n-1.$$

Предположим далее, что

$$|a_1| - r_1 > |C_1|,$$
$$|a_i| - r_i \geqslant |B_i| + |C_i|, \quad 2 \leqslant i \leqslant n-1,$$
$$|a_n| - r_n > |B_n|.$$

Тогда метод Гаусса может быть выполнен без перестановки строк или столбцов.

Замечание. В случае трехдиагональной матрицы $\mathscr{A}$ условия из определения 4 выполнены только для первой и последней строк.

Доказательство теоремы 6. Первый шаг метода Гаусса состоит в порождении трехдиагональной матрицы $\mathscr{A}'$, для которой

$$A'_1 = A_1, \quad C'_1 = C_1,$$
$$B'_2 = 0, \quad B'_i = B_i, \qquad 3 \leqslant i \leqslant n,$$
$$A'_2 = A_2 - C_1 B_2 (1/A_1), \quad A'_i = A_i, \quad 3 \leqslant i \leqslant n,$$
$$C'_i = C_i, \qquad 2 \leqslant i \leqslant n-1.$$

Покажем, что в матрице $\mathscr{A}'$ сильный критерий суммы по строкам выполнен не только для первой, но и для второй строки, т. е. верно

$$|a'_2| - r'_2 > |C_2| = |C'_2|.$$

Имеем

$$A'_2 = A_2 - C_1 (B_2/A_1) \subseteq A_2 - |1/A_1| \langle 0, |C_1| \rangle \langle 0, |B_2| \rangle$$
$$= \langle a_2, r_2 + |C_1||B_2|/(|a_1| - r_1) \rangle,$$

т. е.

$$|a'_2| - r'_2 \geqslant |a_2| - \left(r_2 + \frac{|C_1||B_2|}{|a_1| - r_1} \right).$$

Так как

$$|a_1| - r_1 > |C_1| > 0, \quad |B_2| > 0$$

и

$$-|L_2| - r_2 + |a_2| \geqslant |C_2|,$$

получаем, что

$$|a'_2| - r'_2 \geqslant |a_2| - \left(r_2 + \frac{|C_1||B_2|}{|a_1| - r_1} \right)$$
$$> -|B_2| - r_2 + |a_2| \geqslant |C_2| = |C'_2|.$$

После n—1 шага такого типа приходим к матрице $\hat{\mathscr{A}}$, имеющей ненулевые элементы только для главной диагонали и супердиагонали (расположенной непосредственно над главной), причем никакой элемент главной диагонали не содержит нуля. Третье

предположение $|a_n| - r_n > |B_n|$ применяется на $(n-1)$-м шаге. Заметим, что доказательство теоремы 6 можно несколько сократить. Из условий теоремы 6 следует, что матрица B_p, введенная в теореме 3, имеет неприводимо сильно доминирующую главную диагональ, а потому является M-матрицей. Это означает, что наше утверждение — просто частный случай теоремы 3.

Рассмотрим теперь систему, имеющую более подходящий для итерации вид

$$x = \mathscr{C}x + c,$$

где $\mathscr{C} = (C_{ij})$, $C_{ij} = \langle c_{ij}, r_{ij} \rangle$, $1 \leqslant i, j \leqslant n$ — вещественная интервальная матрица и c — интервальный вектор. Согласно теореме 1, из микромодуля 31, итерационный метод

$$x^{(k+1)} = \mathscr{C}x^{(k)} + c, \quad k = 0, 1, 2, \ldots$$

сходится для любого начального интервального вектора $x^{(0)}$ к единственной неподвижной точке, т. е. решению уравнения $x = \mathscr{C}x + x$, тогда и только тогда, когда спектральный радиус матрицы $|\mathscr{C}| = (|C_{ij}|)$ меньше единицы. Мы хотим показать, что при этом условия теоремы 3 всегда выполнены для матрицы $\mathscr{A} = \mathscr{I}_p - \mathscr{C} = (A_{ij})$, где $\mathscr{I}_p$ — единичная матрица.

Имеем

$$A_{ij} = \begin{cases} \langle 1 - c_{ii}, r_{ii} \rangle, & i = j, \\ -C_{ij} & \text{в противном случае.} \end{cases}$$

Из $\rho(|\mathscr{C}|) < 1$ следует, что $|C_{ii}| = |c_{ii}| + r_{ii} < 1$, $1 \leqslant i \leqslant n$.

Матрица $\mathscr{B}_p = (b_{ij})$, введенная в теореме 3 имеет для $\mathscr{A} = \mathscr{I}_p - \mathscr{C}$ элементы

$$b_{ij} = \begin{cases} |1 - c_{ii}| - r_{ii}, & i = j, \\ -|C_{ij}| & \text{в противном случае.} \end{cases}$$

Мы рассмотрим теперь вещественную матрицу $\mathscr{B}_{pl} = \mathscr{I}_p - |\mathscr{C}|$. Ввиду $\rho(|\mathscr{C}|) < 1$ обратная матрица $\mathscr{B}_{pl}^{-1}$ существует в силу известной теоремы, причем $\mathscr{B}_{pl}^{-1} \geqslant O_p$. Рассмотрим теперь представление матрицы $\mathscr{B}_{pl}$ в виде

$$\mathscr{B}_{pl} = \mathscr{M}_{pl} - \mathscr{N}_{pl},$$

где

$$\mathscr{M}_{pl} = \operatorname{diag}(1 - |C_{ii}|), \quad \mathscr{N}_{pl} = -(\mathscr{B}_{pl} - \mathscr{M}_{pl}) = \mathscr{M}_{pl} - \mathscr{B}_{pl}.$$

Имеем $\mathcal{M}_{\mathrm{pl}}^{-1} \geqslant O_{\mathrm{p}}$, $\mathcal{N}_{\mathrm{pl}} \geqslant O_{\mathrm{p}}$. Отсюда в силу известной теоремы следует, что $\rho\left(\mathcal{M}_{\mathrm{pl}}^{-1} \mathcal{N}_{\mathrm{pl}}\right) < 1$.

Рассмотрим, наконец, представление матрицы $\mathcal{B}_{\mathrm{p}}$ в виде

$$\mathcal{B}_{\mathrm{p}} = \mathcal{M}_{\mathrm{p}} - \mathcal{N}_{\mathrm{p}},$$

где

$$\mathcal{M}_{\mathrm{p}} = \operatorname{diag}(|1 - c_{ii}| - r_{ii}), \quad \mathcal{N}_{\mathrm{p}} = -(\mathcal{B}_{\mathrm{p}} - \mathcal{M}_{\mathrm{p}}) = \mathcal{M}_{\mathrm{p}} - \mathcal{B}_{\mathrm{p}}.$$

Имеем

$$|1 - c_{ii}| - r_{ii} \geqslant 1 - |c_{ii}| - r_{ii} = 1 - |C_{ii}| > 0, \quad 1 \leqslant i \leqslant n,$$

откуда

$$\mathcal{M}_{\mathrm{p}} \geqslant \mathcal{M}_{\mathrm{pl}}, \text{ т. е. } \mathcal{M}_{\mathrm{pl}}^{-1} \geqslant \mathcal{M}_{\mathrm{p}}^{-1}.$$

Отсюда и из $\mathcal{N}_{\mathrm{p}} = \mathcal{N}_{\mathrm{pl}}$ следует, что $\mathcal{M}_{\mathrm{p}}^{-1} \mathcal{N}_{\mathrm{p}} \leqslant \mathcal{M}_{\mathrm{pl}}^{-1} \mathcal{N}_{\mathrm{pl}}$.

Теперь теорема Перрона-Фробениуса дает

$$\rho\left(\mathcal{M}_{\mathrm{p}}^{-1} \mathcal{N}_{\mathrm{p}}\right) \leqslant \rho\left(\mathcal{M}_{\mathrm{pl}}^{-1}, \mathcal{N}_{\mathrm{pl}}\right), \text{ т. е. } \rho\left(\mathcal{M}_{\mathrm{p}}^{-1} \mathcal{N}_{\mathrm{p}}\right) < 1.$$

Применяя известную теорему, мы получаем, наконец, что $\mathcal{B}_{\mathrm{p}}$ является *M*-матрицей.

Замечания. Утверждения о применимости метода Гаусса (теорема 3) можно найти в литературе. Доказательство теоремы 3 — это непосредственное обобщение доказательства таких же утверждений для точечных матриц. Пусть матрица $\mathcal{A}^{\mathrm{T}}$— транспонированная для интервальной матрицы $\mathcal{A}$. Мы имеем утверждение, соответствующее следствию 5. Если $\mathcal{A}^{\mathrm{T}}$ имеет сильно доминирующую диагональ (в смысле определения 4), то метод Гаусса может быть выполнен для интервальной матрицы $\mathcal{A}$ без перестановок строк. Доказательство получается ссылкой на теорему 3, так как введенная там матрица $\mathcal{B}_{\mathrm{p}}$ снова оказывается *M*-матрицей. Аналогичное утверждение справедливо для трехдиагональной матрицы $\mathcal{A}$, если для $\mathcal{A}^{\mathrm{T}}$ выполнены условия теоремы 6.

Только что доказанные утверждения будут далее использованы для решения систем нелинейных точечных уравнений.

Вопрос о применимости метода Гаусса не был пока что достаточно исследован удовлетворительным образом. В литературе было показано, что этот метод применим для частного класса уравнений.

Микромодуль 35
Метод и процедура Хансена

1. Метод Хансена

Если в системе линейных интервальных уравнений диагональ не является сильно доминирующей (определение 4 из микромодуля 34), то ее можно решать с помощью преобразования, предложенного Хансеном. Цель этого преобразования — сделать данную систему интервальных уравнений системой с сильно доминирующей диагональю. Пусть дана интервальная матрица $\mathcal{A} = (A_{ij})$, где $A_{ij} = \langle a_{ij}, r_{ij} \rangle$ — элементы из $I(\mathrm{R})$ или $K(\mathrm{C})$. Будем предполагать, что существуют обращения всех точечных матриц $\mathcal{A}_\mathrm{p} \in \mathcal{A}$. Берем обращение точечной матрицы $m(\mathcal{A}) := (\dot{a}_{ij})$ и с помощью $m(\mathcal{A})^{-1}$ строим интервальную матрицу

$$\tilde{\mathcal{A}} = m(\mathcal{A})^{-1}\mathcal{A}$$

и интервальный вектор

$$\tilde{b} = m(\mathcal{A})^{-1} b.$$

Имеем

$$\{x_\mathrm{p} \mid \mathcal{A}_\mathrm{p} x_\mathrm{p} = b_\mathrm{p} \mathcal{A}_\mathrm{p} \in \mathcal{A},\ b_\mathrm{p} \in b\} \subseteq \{y_\mathrm{p} \mid \mathcal{A}_\mathrm{p} y_\mathrm{p} = \tilde{b}_\mathrm{p} \tilde{\mathcal{A}}_\mathrm{p} \subset \tilde{\mathcal{A}},\ \tilde{b}_\mathrm{p} \in \tilde{b}\}.$$

Чтобы показать это, предположим, что x_p принадлежит множеству из левой части, т. е

$$\mathcal{A}_\mathrm{p} x_\mathrm{p} = b_\mathrm{p}, \text{ где } \mathcal{A}_\mathrm{p} \in \mathcal{A},\ b_\mathrm{p} \in b.$$

Тогда

$$m(\mathcal{A})^{-1}\mathcal{A}_\mathrm{p} x_\mathrm{p} = m(\mathcal{A})^{-1} b_\mathrm{p}$$

и наше утверждение следует из

$$m(\mathcal{A})^{-1}\mathcal{A}_\mathrm{p} \in \tilde{\mathcal{A}}, \quad m(\mathcal{A})^{-1} b_\mathrm{p} \in \tilde{b}.$$

Идея рассматриваемого преобразования состоит в том, что если элементы матрицы $\mathcal{A}$ имеют не слишком большую ширину, то диагональ матрицы $\tilde{\mathcal{A}}$ будет сильно доминирующей. Тогда можно применить метод Гаусса. В пределе мы имеем $d(\mathcal{A}) = \mathcal{O}_\mathrm{p}$, т. е. $\tilde{\mathcal{A}} = \mathcal{I}_\mathrm{p}$, так что в этом случае матрица $\tilde{\mathcal{A}}$ наверняка имеет сильно

доминирующую диагональ. Если же ширина компонент матрицы $\mathscr{A}$ невелика, то $\tilde{\mathscr{A}}$ несильно отличается от $\mathscr{I}_p$.

Ясно, что сильное доминирование диагонали для матрицы $\tilde{\mathscr{A}} = m(\mathscr{A})^{-1}\mathscr{A}$ зависит не только от ширины компонент матрицы $\mathscr{A}$. Действительно, если мы представим компоненты этой интервальной матрицы (которые могут быть круговыми комплексными или вещественными интервалами) в виде

$$A_{ij} = \langle a_{ij},\ r_{ij}\rangle, \quad 1 \leqslant i,\ j \leqslant n,$$

то, вводя еще матрицу

$$\mathscr{D} = (D_{ij}),\ D_{ij} = \langle 0,\ r_{ij}\rangle,\ 1 \leqslant i,\ j \leqslant n,$$

получим, что

$$\begin{aligned}\tilde{\mathscr{A}} = m(\mathscr{A})^{-1}\mathscr{A} &= m(\mathscr{A})^{-1}(m(\mathscr{A}) + \mathscr{D}) \\ &= \mathscr{I}_p + m(\mathscr{A})^{-1}\mathscr{D} = \mathscr{I}_p + |m(\mathscr{A})^{-1}|\mathscr{D} \\ &= \mathscr{I}_p + \mathscr{H},\end{aligned}$$

где

$$\mathscr{H} = |m(\mathscr{A})^{-1}|\mathscr{D}.$$

Так как

$$\begin{aligned}\|\,|\mathscr{H}|\,\| &\leqslant \|\,|m(\mathscr{A})^{-1}|\,\| \cdot \|\,|\mathscr{D}|\,\| = \frac{1}{2}\|\,|m(\mathscr{A})^{-1}|\,\|\,\|d(\mathscr{A})\| \\ &\leqslant \frac{1}{2}\|\,|m(A)^{-1}|\,\| \cdot \|d(\mathscr{A})\|\|m(\mathscr{A})\|,\end{aligned}$$

то мы видим, что при данной точечной матрице $m(\mathscr{A})$ диагональ матрицы $\mathscr{A}$ будет сильно доминировать с тем большей вероятностью, чем меньше число

$$\hat{\varkappa} = \|\,|m(\mathscr{A})^{-1}|\,\|\,\|m(\mathscr{A})\|$$

Если $m(\mathscr{A})$ — вещественная точечная матрица и мы используем монотонную матричную норму, то

$$\bar{\varkappa} = \|m(\mathscr{A})^{-1}\|\,\|m(\mathscr{A})\|,$$

где $\bar{\varkappa}$ —хорошо известная величина, обусловленность матрицы $m(\mathscr{A})$. Поэтому применимость метода Хансена к вещественной интервальной матрице зависит не только от величины $d(\mathscr{A})$, но (даже более существенным образом) и от обусловленности матрицы $m(\mathscr{A})$.

В предыдущем микромодуле мы описали метод преобразования множества линейных интервальных уравнений к верхней треугольной

форме. Другие методы, разработанные в теории вещественных систем линейных уравнений, также могут быть обобщены на линейные системы интервальных уравнений. Мы упомянем метод Гаусса — Жордана, которым данная матрица приводится к диагональному виду. После этого вычисление решения исходной системы требует еще n добавочных делений. Подробное описание этого варианта алгоритма Гаусса можно найти в литературе. Тем же методом, что и в теореме 3 из микромодуля 34, можно показать, чго этот метод всегда применим к данной системе линейных интервальных уравнений, если ее матрица имеет сильно доминирующую диагональ.

Пусть дана вещественная интервальная матрица $\mathcal{A} = (A_{ij})$. Будем предполагать, что обратная матрица $\mathcal{A}_p^{-1}$ существует для любой $\mathcal{A}_p \in \mathcal{A}$. Пусть далее $\mathcal{b} = (B_i)$ — вещественный интервальный вектор.

Запишем в виде

$$\mathcal{l} = (L_1,\ L_2,\ \ldots,\ L_n)^T$$

вещественный интервальный вектор с компонентами L_i, $1 \leqslant i \leqslant n$, которые получаются из множества

$$\mathfrak{L} = \{x_p \mid \mathcal{A}_p x_p = \mathcal{b}_p,\ \mathcal{A}_p \in \mathcal{A},\ \mathcal{b}_p \in \mathcal{b}\}$$

проектированием на соответствующие оси координат. Иными словами,

$$L_i = L_i(\mathcal{A},\ \mathcal{b}) = \{l_i \mid (l_1,\ \ldots,\ l_i,\ \ldots,\ l_n)^T \in \mathfrak{L}\},\ 1 \leqslant i \leqslant n.$$

$\mathcal{l}$ — интервальный вектор наименьшей ширины, содержащий множество $\mathfrak{L}$.

Мы хотим теперь исследовать вопрос о том, насколько хорошо интервальные векторы, вычисляемые по методу Хансена, аппроксимируют вектор $\mathcal{l}$. Для «решения» преобразованной системы линейных уравнений мы используем тогда метод Гаусса — Жордана. Будет показано, что разность между шириной полученного вектора и шириной вектора $\mathcal{l}$ стремится к нулю при стремлении к нулю ширины векторов $\mathcal{A}$ и $\mathcal{b}$. Это показывает, что метод Хансена дает вектор, достаточно близкий к вектору $\mathcal{l}$, если ширина исходных данных не слишком велика. Мы снова представляем вещественный интервал $A = [a_1, a_2]$ с помощью его середины $a = \frac{1}{2}(a_1 + a_2)$ и полуширины $r = \frac{1}{2} d(A) = \frac{1}{2}(a_2 - a_1)$:

$$A = \langle a,\ r \rangle.$$

Лемма 1. *Пусть* $\mathscr{A}_p v_p = \mathscr{b}_p$, *где* $\mathscr{A}_p = (a_{ij})$— *вещественная невырожденная матрица, имеющая обратную* $\mathscr{A}_p^{-1} = \mathscr{B}_p = (b_{ij})$, *и пусть* $x_p = (x_j)$, $\mathscr{b}_p = (b_i)$— *точечные векторы. Пусть далее* $\mathscr{A} = (A_{ij})$ — *интервальная матрица с элементами* $A_{ij} = \langle a_{ij}, r_{ij} \rangle$, $1 \leqslant i, j \leqslant n$, *а* $\mathscr{b} = (B_i)$ — *интервальный вектор с элементами* $B_i = \langle b_i, r_i \rangle$, $1 \leqslant i \leqslant n$. *Тогда для k-й компоненты* L_k *введенного выше интервального вектора* $\ell = (L_i)$ *справедливо соотношение*

$$\frac{1}{2} d(L_k) = \sum_{i=1}^{n} \sum_{j=1}^{n} | b_{ki} x_j r_{ij} | + \sum_{i=1}^{n} | b_{ki} r_i | + O(d^2).$$

Здесь

$$d = \max \left\{ \max_{1 \leqslant i, j \leqslant n} \{r_{ij}\}, \max_{1 \leqslant i \leqslant n} \{r_i\} \right\}.$$

Запись $O(d^2)$ *обозначает любую вещественную функцию f от d для которой*

$$| f/d^2 | \leqslant \gamma \quad при \quad d \leqslant d_0,$$

где $\gamma \geqslant 0$, $d_0 > 0$ — *константы.*

Доказательство В силу правила Крамера множество L_k является образом $(n^2 + n)$-мерного гиперкуба при отображении

$$x_k = x_k(\mathscr{A}_p, \mathscr{b}_p).$$

Из теоремы о среднем значении мы получаем, что

$$x_k(\hat{\mathscr{A}}_p, \hat{\mathscr{b}}_p) = x_k(\mathscr{A}_p, \mathscr{b}_p) + \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial x_k}{\partial a_{ij}} (\hat{a}_{ij} - a_{ij}) + \sum_{i=1}^{n} \frac{\partial x_k}{\partial b_i} (\hat{b}_i - b_i) + \frac{1}{2} x_k''(u_p + t(v_p - u_p))(v_p - u_p) \times (v_p - u_p).$$

Здесь $t \in (0, 1)$, через x_k'' обозначен гессиан отображения x_k, а u_p, v_p — это векторы из $V_{n^2+n}(\mathbb{R})$, причем значениями компонент вектора u_p (соответственно вектора v_p) являются элементы матрицы $\mathscr{A}_p$ и вектор $\mathscr{b}_p$ (соответственно элементы матрицы $\hat{\mathscr{A}}_p$ и вектор $\hat{\mathscr{b}}_p$). Если теперь продифференцировать n уравнений

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \quad 1 \leqslant i \leqslant n$$

по a_{ij}, то, используя соотношение

$$\mathscr{A}_{\mathrm{p}} \frac{\partial}{\partial a_{ij}} x_{\mathrm{p}} = -x_j e_{\mathrm{p}}^{i},$$

мы получим уравнение

$$\frac{\partial}{\partial a_{ij}} x_{\mathrm{p}} = \left(\frac{\partial x_1}{\partial a_{ij}}, \frac{\partial x_2}{\partial a_{ij}}, \ldots, \frac{\partial x_n}{\partial a_{ij}} \right)^{\mathrm{T}},$$

где e_{p}^{i} есть i-й единичный вектор. Отсюда мы получаем

$$\frac{\partial}{\partial a_{ij}} x_{\mathrm{p}} = -x_j \mathscr{A}_{\mathrm{p}}^{-1} e_{\mathrm{p}}^{i} \text{ или } \frac{\partial x_k}{\partial a_{ij}} = -b_{ki} x_j.$$

Из равенства

$$x_{\mathrm{p}} = \mathscr{B}_{\mathrm{p}} b_{\mathrm{p}}$$

мы получаем

$$\frac{\partial x_k}{\partial b_i} = b_{ki}.$$

Если использовать эти формулы для производных в теореме о среднем значении, то мы получим

$$x_k(\hat{\mathscr{A}}_{\mathrm{p}}, \hat{b}_{\mathrm{p}}) \in x_k(\mathscr{A}_{\mathrm{p}}, b_{\mathrm{p}}) + \sum_{i=1}^{n} \sum_{j=1}^{n} |b_{ki} x_j| \langle 0, r_{ij} \rangle + \sum_{i=1}^{n} |b_{ki}| \langle 0, r_i \rangle + \ldots,$$

так как

$$\mathscr{A}_{\mathrm{p}} \in \mathscr{A}, \quad b_{\mathrm{p}} \in b.$$

Иными словами,

$$\frac{1}{2} d(L_k) = \sum_{i=1}^{n} \sum_{j=1}^{n} |b_{ki} x_j r_{ij}| + \sum_{i=1}^{n} |b_{ki} r_i| + O(d^2).$$

Лемма 2. *Пусть $\mathscr{A}_{\mathrm{p}} = (a_{ij})$ — вещественная невырожденная матрица, для которой*

$$\mathscr{A}_{\mathrm{p}}^{-1} = \mathscr{B}_{\mathrm{p}} = (b_{ij}),$$

и пусть — $b_{\mathrm{p}} = (b_i)$ вещественный вектор. Пусть далее $\mathscr{A} = (A_{ij})$ — вещественная интервальная матрица с элементами $A_{ij} = \langle a_{ij}, r_{ij} \rangle$, $1 \leqslant i, j \leqslant n$, и $b = (B_i)$ — вещественный интервальный вектор с элементами $B_i = \langle b_i, r_i \rangle$, $1 \leqslant i \leqslant n$. Запишем в виде $\tilde{\mathscr{A}} = (A_{ij})$ интервальную матрицу $\mathscr{A}_{\mathrm{p}}^{-1} \mathscr{A} = \mathscr{B}_{\mathrm{p}} \mathscr{A}$, а в виде $b = (B_i)$ — интервальный вектор $\mathscr{A}_{\mathrm{p}}^{-1} b = \mathscr{B}_{\mathrm{p}} b$. Предположим, что все точечные матрицы, принадлежащие $\tilde{\mathscr{A}}$, невырожденные. Тогда интервальный вектор

$$\tilde{l} = (\tilde{L}_1,\ \tilde{L}_2,\ \ldots,\ \tilde{L}_n)^{\mathrm{T}},$$

где

$$L_i = L_i(\tilde{\mathcal{A}},\ \tilde{b}) = \{\tilde{l}_i \mid (\tilde{l}_1,\ \ldots,\ \tilde{l}_i,\ \ldots,\ \tilde{l}_n)^{\mathrm{T}} \in \mathfrak{L}\},\quad 1 \leqslant i \leqslant n,$$

$$\mathfrak{L} = \{\tilde{x}_{\mathrm{p}} \mid \tilde{\mathcal{A}}_{\mathrm{p}} = \tilde{b}_{\mathrm{p}},\ \tilde{\mathcal{A}}_{\mathrm{p}} \in \tilde{\mathcal{A}},\ \tilde{b}_{\mathrm{p}} \in \tilde{b}\},$$

удовлетворяет соотношению

$$d(\tilde{L}_k) = d(L_k) + O(d^2),\quad 1 \leqslant k \leqslant n.$$

Доказательство. Пусть $\mathcal{A}_{\mathrm{p}} x_p = b_{\mathrm{p}}$. Из леммы 1 следует, что

$$\frac{1}{2} d(L_k) = \sum_{i=1}^{n} \sum_{j=1}^{n} |b_{ki} x_j r_{ij}| + \sum_{i=1}^{n} |b_{ki} r_i| + O(d^2).$$

Элементы интервальной матрицы $\tilde{\mathcal{A}} = (\tilde{A}_{ij})$ имеют вид

$$\tilde{A}_{ij} = \left\langle \delta_{ij},\ \sum_{m=1}^{n} |b_{im} r_{im}| \right\rangle,\quad 1 \leqslant i,\ j \leqslant n,$$

где δ_{ij} — символ Кронекера. Элементы интервального вектора $\tilde{b} = (\tilde{B}_i)$ имеют вид

$$\tilde{B}_i = \left\langle x_i,\ \sum_{m=1}^{n} |b_{im} r_m| \right\rangle,\quad 1 \leqslant i \leqslant n.$$

Применяя лемму 1 еще раз, мы получаем

$$\begin{aligned}
\frac{1}{2} d(\tilde{L}_k) &= \sum_{i=1}^{n} \sum_{j=1}^{n} \left| \delta_{ki} x_j \left(\sum_{m=1}^{n} |b_{im} r_{im}| \right) \right| \\
&\quad + \sum_{i=1}^{n} \left| \delta_{kl} \left(\sum_{m=1}^{n} |b_{im} r_m| \right) \right| + O(\tilde{d}^2) \\
&= \sum_{j=1}^{n} \left| x_j \left(\sum_{m=1}^{n} |b_{km} r_{mj}| \right) \right| + \sum_{m=1}^{n} |b_{km} r_m| + O(\tilde{d}^2) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} |b_{ki} x_j r_{ij}| + \sum_{i=1}^{n} |b_{kl} r_i| + O(\tilde{d}^2),
\end{aligned}$$

где

$$\tilde{d} = \max \left\{ \max_{1 \leqslant i,\ j \leqslant n} \{\tilde{r}_{ij}\},\ \max_{1 \leqslant i \leqslant n} \{\tilde{r}_i\} \right\}.$$

Из формул для элементов матрицы $\tilde{\mathcal{A}}$ и вектора $\tilde{b}$ мы сразу усматриваем, что

$$\tilde{d} = O(d).$$

Сравнение с выражением для $\frac{1}{2}\, d(L_k)$ дает

$$d(\tilde{L}_k) = d(L_k) + O(d^2),$$

что и доказывает лемму.

Теперь метод Гаусса — Жордана за конечное число шагов порождает по данной интервальной матрице $\mathscr{A}$ некоторую диагональную матрицу. В результате каждого шага в новой матрице появляется хотя бы один новый нуль вне диагонали. Предположим, что даны интервальная матрица $\mathscr{H} = (H_{ij})$ и интервальный вектор $h = (H_i)$, такие что

$$H_{ij} = \langle h_{ij},\ e_{ij} \rangle, \quad 1 \leqslant i,\ j \leqslant n,$$
$$H_i = \langle h_i,\ e_i \rangle, \quad 1 \leqslant i \leqslant n,$$

где $h_{ij} = \delta_{ij},\ h_i = x_i$
и

$$\max \left\{ \max_{1 \leqslant i,\ j \leqslant n} \{e_{ij}\},\ \max_{1 \leqslant i \leqslant n} \{e_i\} \right\} = O(d).$$

Пусть, кроме того, $\mathscr{A} = (A_{ij}),\ b = (b_{ij})$, где

$$A_{ij} = \langle a_{ij},\ r_{ij} \rangle, \quad 1 \leqslant i,\ j \leqslant n,$$
$$B_i = \langle b_i,\ r_i \rangle, \quad 1 \leqslant i \leqslant n.$$

Положим по определению

$$d = \max \left\{ \max_{1 \leqslant i,\ j \leqslant n} \{r_{ij}\},\ \max_{1 \leqslant i \leqslant n} \{r_i\} \right\}.$$

Теперь мы можем доказать по индукции, что наши предположения о матрице $\mathscr{H}$ и векторе h верны на любом шаге алгоритма. Как показано в лемме 2, эти предположения справедливы для матрицы $\mathscr{H} := \mathscr{A} = \mathscr{A}_0^{-1}\mathscr{A}$ и интервального вектора $h := b = \mathscr{A}_0^{-1} b$. Кроме того, из леммы 2 следует, что

$$d(\tilde{L}_k) = d(L_k(\mathscr{H},\ h)) = d(L_k) + O(d^2), \quad 1 \leqslant k \leqslant n.$$

Чтобы получить нуль для пары $(r,\ s)$ индексов, (где $r \neq s$) по матрице $\mathscr{H}'$ и вектору h' строятся матрица $\mathscr{H}$ и вектор h согласно следующим формулам:

$$H'_{ij} = H_{ij}, \qquad i \neq r,$$
$$H'_{rj} = H_{rj} - H_{sj} H_{rs}/H_{ss}, \quad j \neq s,$$
$$H'_{rs} = 0,$$
$$H'_i = H_i, \qquad i \neq r,$$
$$H_r = H_r - H_s H_{rs}/H_{ss}.$$

Ввиду $h_{ij} = \delta_{ij}$ отсюда следует, что

$$\begin{aligned} H'_{rj} &= \langle h_{rj},\ e_{rj}\rangle - \langle h_{sj},\ e_{sj}\rangle \langle h_{rs},\ e_{rs}\rangle / \langle h_{ss},\ e_{ss}\rangle \\ &= \langle \delta_{rj},\ e_{rj}\rangle - \langle 0,\ e_{sj}\rangle \langle 0,\ e_{rs}\rangle / \langle h_{ss},\ e_{ss}\rangle \\ &= \langle \delta_{rj},\ e_{rj}\rangle - \langle 0,\ e_{sj}e_{rs}\rangle (1/|(-e_{ss})) \\ &= \langle \delta_{rj},\ e_{rj} + [e_{sj}e_{rs}/(1 - e_{ss})]\rangle \end{aligned}$$

и

$$\begin{aligned} H'_r &= \langle x_r,\ e\rangle - \langle x_s,\ e_s\rangle \langle 0,\ e_{rs}\rangle / \langle h_{ss},\ e_{ss}\rangle \\ &= \langle x_r,\ e_r + ((|\, x_s\, |e_{rs} + e_{rs}e_s)/(1 - e_{ss}))\rangle. \end{aligned}$$

Поэтому представление

$$H'_{ij} = \langle \delta_{ij},\ e'_{ij}\rangle, \quad H'_i = \langle x_i,\ e'_i\rangle$$

справедливо также для матрицы $\mathscr{H}' = (H'_{ij})$ и вектора $k' = (H'_i)$.

Из леммы 1 следует, что

$$\begin{aligned} \frac{1}{2} d(L_k(\mathscr{H},\ k)) &= \sum_{i=1}^{n} \sum_{j=1}^{n} |\, \delta_{ki} x_j e_{ij}\, | + \sum_{i=1}^{n} |\, \delta_{ki} e_i\, | + O(d^2) \\ &= \sum_{j=1}^{n} |\, x_j e_{kj}\, | + e_k + O(d^2). \end{aligned}$$

Аналогично мы получаем

$$\frac{1}{2} d(L_k(\mathscr{H}',\ k')) = \sum_{j=1}^{n} |\, x_j e'_{kj}\, | + e'_k + O(d^2).$$

Для $k \neq r$ имеем $e'_{kj} = e_{kj}$ и $e'_k = e_k$. Из допущения

$$d(L_k(\mathscr{H},\ k)) = d(L_k) + O(d^2)$$

получаем поэтому

$$d(L_k(\mathscr{H}',\ k')) = d(L_k) + O(d^2), \quad k \neq r.$$

Для $k = r$ имеем

$$\sum_{j=1}^{n} |\, x_k e'_{kj}\, | = \sum_{j=1,\ j\neq s}^{n} |\, x_j e_{kj}\, | + O(d^2)$$

и

$$e'_k = e_k + |\, x_s e_{ks}\, | + O(d^2),$$

откуда

$$\frac{1}{2} d(L_k(\mathscr{H}',\ k')) = \sum_{j=1}^{n} |\, x_j e_{kj}\, | + e_k + O(d^2),$$

т. е.

$$d(L_k(\mathscr{H}',\ \hbar')) = d(L_k) + O(d^2) \text{ для } k = r.$$

Этим завершается доказательство по индукции. Решение диагональной системы уравнений не увеличивает ширину элементов, поэтому соотношение

$$d(L_k(\widehat{\mathscr{H}},\ \widehat{\hbar})) = d(L_k(\mathscr{A},\ \ell)) + O(d^2)$$

справедливо для заключительной диагональной матрицы $\mathscr{H}$ и соответствующего интервального вектора $\hbar$.

2. Прцедура Купермана и Хансена

Пусть $\mathscr{A}$ — интервальная матрица, такая, что $\mathscr{A}_\text{р}$— невырожденная для всех $\mathscr{A}_\text{р} \in \mathscr{A}$, и пусть

$$\mathfrak{L} = \{x_\text{р} \mid \mathscr{A}_\text{р} x_\text{р} = \ell_\text{р},\ \mathscr{A}_\text{р} \in \mathscr{A},\ \ell_\text{р} = \ell\}$$

— множество всех решений для данного интервального вектора ℓ Даже простые примеры показывают, что метод Хансена, описанный в предыдущем пункте, вычисляет в общем случае лишь некоторое подмножество векгора ℓ, имеющего компоненты

$$L_k = \{l_k \mid (l_1,\ \ldots,\ l_k,\ \ldots,\ l_n)^\text{T} \in \mathfrak{L}\}.$$

Куперман описал неинтервальную процедуру, которая в некоторых случаях дает лучшую локализацию множества $\mathfrak{L}$. Впоследствии Хансен обобщил эту процедуру до интервального метода.

Рассмотрим множество линейных уравнений

$$\mathscr{A}_\text{р} x_\text{р} = \ell_\text{р},\quad x_\text{р} = (x_i),$$

где $\mathscr{A}_\text{р}$ — неособенная точечная матрица и $\ell_\text{р}$— вещественный вектор. Если частная производная неотрицательна

$$\frac{\partial x_k}{\partial a_{ij}} \geqslant 0,$$

то x_k— неубывающая функция от a_{ij}. Если теперь a_{ij} может изменяться в вещественном интервале

$$A_{ij} = [a_{ij}^1,\ a_{ij}^2],$$

то величина x_k, рассматриваемая как функция от a_{ij} на интервале $A_{ij} = [a_{ij}^1,\ a_{ij}^2]$, принимает свое наименьшее значение при $a_{ij} = a_{ij}^1$, а наибольшее значение при $a_{ij} = a_{ij}^2$. То же самое можно сказать и о зависимости компонент x_k от ℓ_i.

Пусть теперь $\mathscr{A}=(A_{ij})$ — данная вещественная интервальная матрица и $\mathscr{b}=(B_i)$— данный интервальный вектор. Чтобы локализовать интервал

$$L_k=[l_k^1,\ l_k^2],\quad 1\leqslant k\leqslant n,$$

поступаем следующим образом.

Начинаем с вещественной интервальной матрицы $\mathscr{A}=(A_{ij})$, $A_{ij}=[a_{ij}^1,\ a_{ij}^2]$, $1\leqslant i,\ j\leqslant n$, и вещественного интервального вектора $\mathscr{b}=(B_i)$, $B_i=[b_i^1,\ b_i^2]$, $1\leqslant i\leqslant n$. Затем строим интервальные матрицы

$$\tilde{\mathscr{A}}=(\tilde{A}_{ij})\ \text{и}\ \hat{\mathscr{A}}=(\hat{A}_{ij})$$

и интервальные векторы

$$\tilde{\mathscr{b}}=(\tilde{B}_i)\ \text{и}\ \hat{\mathscr{b}}=(\hat{B}_i)$$

согласно следующим правилам:

$$\tilde{A}_{ij}=\begin{cases}[a_{ij}^1,\ a_{ij}^1], & \text{если } \partial x_k/\partial a_{ij}\geqslant 0 \text{ для всех } \mathscr{A}_p\in\mathscr{A} \text{ и } \mathscr{b}_p\in\mathscr{b},\\ [a_{ij}^2,\ a_{ij}^2], & \text{если } \partial x_k/\partial a_{ij}\leqslant 0 \text{ для всех } \mathscr{A}_p\in\mathscr{A} \text{ и } \mathscr{b}_p\in\mathscr{b},\\ A_{ij} & \text{в противном случае.}\end{cases}$$

$$\hat{A}_{ij}=\begin{cases}[a_{ij}^2,\ a_{ij}^2], & \text{если } \partial x_k/\partial a_{ij}\geqslant 0 \text{ для всех } \mathscr{A}_p\in\mathscr{A} \text{ и } \mathscr{b}_p\in\mathscr{b},\\ [a_{ij}^1,\ a_{ij}^1], & \text{если } \partial x_k/\partial a_{ij}\leqslant 0 \text{ для всех } \mathscr{A}_p\in\mathscr{A} \text{ и } \mathscr{b}_p\in\mathscr{b},\\ A_{ij} & \text{в противном случае.}\end{cases}$$

$$\tilde{B}=\begin{cases}[b_i^1,\ b_i^1], & \text{если } \partial x_k/\partial b_i\geqslant 0 \text{ для всех } \mathscr{A}_p\in\mathscr{A} \text{ и } \mathscr{b}_p\in\mathscr{b},\\ [b_i^2,\ b_i^2], & \text{если } \partial x_k/\partial b_i\leqslant 0 \text{ для всех } \mathscr{A}_p\in\mathscr{A} \text{ и } \mathscr{b}_p\in\mathscr{b},\\ B_i & \text{в противном случае.}\end{cases}$$

$$\hat{B}_i=\begin{cases}[b_i^2,\ b_i^2], & \text{если } \partial x_k/\partial b_i\geqslant 0 \text{ для всех } \mathscr{A}_p\in\mathscr{A} \text{ и } \mathscr{b}_p\in\mathscr{b},\\ [b_i^1,\ b_i^1], & \text{если } \partial x_k/\partial b_i\leqslant 0 \text{ для всех } \mathscr{A}_p\in\mathscr{A} \text{ и } \mathscr{b}_p\in\mathscr{b},\\ B_i & \text{в противном случае.}\end{cases}$$

Теперь вычисляем локализующие интервалы

$$X_k=[x_k^1,\ x_k^2]\quad\text{и}\quad Y_k=[y_k^1,\ y_k^2]$$

для интервалов

$$L_k(\tilde{\mathscr{A}},\ \tilde{\mathscr{b}})=[\tilde{l}_k^1,\ \tilde{l}_k^2]\quad\text{и}\quad L_k(\hat{\mathscr{A}},\ \hat{\mathscr{b}})=[\hat{l}_k^1,\ \hat{l}_k^2],$$

используя, например, метод Хансена, описанный в предыдущем пункте. Предыдущие рассуждения показывают, что для $L_k(\mathscr{A},\ \mathscr{b})=[l_k^1,\ l_k^2]$ имеет место

$$l_k^1\geqslant\tilde{l}_k^1\ \text{и}\ l_k^2\leqslant\hat{l}_k^2.$$

Отсюда следует, что верно

$$[\hat{l}_k^1,\ \hat{l}_k^2] \supseteq L_k(\mathscr{A},\ \mathscr{b}).$$

Поэтому имеем также

$$[x_k^1,\ y_k^2] \supseteq [\hat{l}_k^1,\ \hat{l}_k^2] \supseteq L_k(\mathscr{A},\ \mathscr{b}).$$

Для построения интервальных матриц $\tilde{\mathscr{A}}$, $\hat{\mathscr{A}}$ и интервальных векторов $\tilde{\mathscr{b}}$, $\hat{\mathscr{b}}$ при фиксированном k нужны частные производные $\partial x_k/\partial a_{ij}$ и $\partial x_k/\partial b_i$. Формулы для них уже были получены в предыдущем пункте: полагая $\mathscr{A}_\mathrm{p}^{-1} = (\tilde{a}_{ij})$, имеем

$$\frac{\partial x_k}{\partial a_{ij}} = -\tilde{a}_{ki}x_j, \quad \frac{\partial x_k}{\partial b_i} = \tilde{a}_{ki}.$$

Чтобы выяснить распределение знаков этих производных, мы вычисляем, используя, например, метод Хансена, интервальный вектор $\bar{\mathscr{l}} = (\bar{L}_i)$, содержащий множество $\mathfrak{L}$, и интервальную матрицу $\mathscr{A} = (\bar{A}_{ij})$, содержащую обращения всех $\mathscr{A}_\mathrm{p} \in \mathscr{A}$. Если теперь нижняя граница интервала $\bar{A}_{ki}\bar{L}_i$ неотрицательна для некоторого фиксированного k, то $\partial x_k/\partial a_{ij} \leqslant 0$ для всех $\mathscr{A}_\mathrm{p} \in \mathscr{A}$ и $\mathscr{b}_\mathrm{p} \in \mathscr{b}$. Аналогично, если верхняя граница интервала $\bar{A}_{ki}\bar{L}_i$ неположительна, то $\partial x_k/\partial a_{ij} \geqslant 0$. Соответствующее утверждение верно и для $\partial x_k/\partial b_i$. Если $0 \in \bar{A}_{ki}\bar{L}_i$ (соответственно $0 \in \bar{A}_{ki}$), то мы все еще можем иметь $\partial x_k/\partial a_{ij} \geqslant 0$ или ≤ 0 (соответственно $\partial x_k/\partial b_i \geqslant 0$ или ≤ 0), так как $\mathscr{l}$ только содержит множество $\mathfrak{L}$, а $\mathscr{A}$ только локализует множество обращений матриц $\mathscr{A}_\mathrm{p} \in \mathscr{A}$. При построении матриц $\tilde{\mathscr{A}}$, $\hat{\mathscr{A}}$ и векторов $\tilde{\mathscr{b}}$, $\hat{\mathscr{b}}$ элементы матрицы $\mathscr{A}$ и вектора $\mathscr{b}$ преобразуются в этом случае так, как будто $\partial x_k/\partial a_{ij}$ (соответственно $\partial x_k/\partial b_i$) меняет знак, потому, что вычисленные значения не позволяют принять иное решение. Этот метод дает, вообще говоря, гораздо лучшую локализацию, чем метод Хансена из п.1. Его недостаток —большой объем вычислений. В общем случае приходится не только вычислять интервальную матрицу $\bar{\mathscr{A}}$, содержащую обращения всех матриц $\mathscr{A}_\mathrm{p} \in \mathscr{A}$, но и решать две системы интервальных уравнений для каждой компоненты $L_k(\mathscr{A},\ \mathscr{b})$.

Замечания. Метод Купермана и Хансена применим и к итерационным методам. Можно искать метод, основанный на решении $2n$ систем уравнений методом Гаусса. Однако использование метода Хансена всегда дает лучшую локализацию.

Микромодуль 36
Итерационные методы для локализации обратной матрицы и разложения на треугольные

Пусть даны невырожденная матрица $\mathscr{A}_{\mathrm{p}} \in M_{nn}(\mathbb{R})$ размерности $n \times n$ и интервальная матрица $\mathscr{X}^{(0)} \in M_{nn}(I(\mathbb{R}))$, такая, что $\mathscr{X}^{(0)}$ локализует матрицу, обратную к $\mathscr{A}_{\mathrm{p}}$, т. е. $\mathscr{A}_{\mathrm{p}}^{-1} \in \mathscr{X}^{(0)}$.

Рассмотрим здесь процедуры, которые итерационно улучшают локализующую матрицу $\mathscr{X}^{(0)}$.

При этом будем использовать преобразование m, отображающее множество интервальных матриц размерности $n \times n$ во множестве вещественных точечных матриц размерности $n \times n$. Оно переводит каждую интервальную матрицу в такую, элементами которой являются середины соответствующих элементов исходной матрицы. Иными словами, в обозначениях

$$i(X) = x_1, \quad s(X) = x_2 \quad X = [x_1, \ x_2] \in I(\mathbb{R})$$

вводим отображение.

$$m: \ M_{nn}(I(\mathbb{R})) \to M_{nn}(\mathbb{R}) \tag{1}$$

равенствами $m(\mathscr{X}) = \frac{1}{2}(i(X_{ij}) + s(X_{ij}))$.

Это срединное отображение интервальных матриц очевидным образом непрерывно. Оно обладает следующими свойствами:

$$m(\mathscr{X} \pm \mathscr{Y}) = m(\mathscr{X}) \pm m(\mathscr{Y}), \quad \mathscr{X}, \ \mathscr{Y} \in M_{nn}(I(\mathbb{R})), \tag{2}$$

$$m(\mathscr{B}_{\mathrm{p}}\mathscr{X}) = \mathscr{B}_{\mathrm{p}} m(\mathscr{X}), \ m(\mathscr{X}\mathscr{B}_{\mathrm{p}}) = m(\mathscr{X})\mathscr{B}_{\mathrm{p}}, \tag{3}$$

$$\mathscr{B}_{\mathrm{p}} \in M_{nn}(\mathbb{R}), \ \mathscr{X} \in M_{nn}(I(\mathbb{R})),$$

$$m(\mathscr{B}_{\mathrm{p}}) = \mathscr{B}_{\mathrm{p}}, \quad \mathscr{B}_{\mathrm{p}} \in M_{nn}(\mathbb{R}). \tag{4}$$

Вот краткое доказательство соотношения (3):

$$m(\mathscr{B}_{\mathrm{p}}\mathscr{X}) = m\left(\sum_{k=1}^{n} b_{lk} X_{kj}\right) = \left(\sum_{k=1}^{n} m(b_{lk}[i(X_{kj}), \ s(X_{kj})])\right)$$

$$= \left(\sum_{k=1}^{n} b_{lk} \frac{1}{2}(i(X_{kj}) + s(X_{kj}))\right) = \left(\sum_{k=1}^{n} b_{lk} m(X_{kj})\right)$$

$$= \mathscr{B}_{\mathrm{p}} m(\mathscr{X}).$$

Утверждения (2) и (4) могут быть доказаны аналогично. Теперь сформулируем первый метод, позволяющий вычислять

последовательность локализаций для обратной матрицы $\mathscr{A}_{\mathrm{p}}^{-1}$. Пусть $r > 1$ — фиксированное натуральное число.

Для произвольной точечной матрицы $\mathscr{B}_{\mathrm{p}}$ положим

$$\mathscr{B}_{\mathrm{p}}^{(0)} = \mathscr{I}_{\mathrm{p}} \text{ (где } \mathscr{I}_{\mathrm{p}} - \text{единичная матрица).}$$

Рассмотрим итерационную процедуру

$$\mathscr{X}^{(k+1)} = m\left(\mathscr{X}^{(k)}\right) \sum_{\nu=0}^{r-2} \left(\mathscr{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m\left(\mathscr{X}^{(k)}\right)\right)^{\nu} + \mathscr{X}^{(k)} \left(\mathscr{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m\left(\mathscr{X}^{(k)}\right)\right)^{r-1}, \quad k \geqslant 0. \tag{5}$$

В случае $r = 2$ получаем формулу

$$\mathscr{X}^{(k+1)} = m\left(\mathscr{X}^{(k)}\right) + \mathscr{X}^{(k)} \left(\mathscr{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m\left(\mathscr{X}^{(k)}\right)\right), \quad k \geqslant 0,$$

которую можно считать интервальным вариантом метода Шульца для вычисления обратной матрицы.

Свойства итерационного метода (5) собраны в следующем утверждении.

Теорема 1. *Пусть* $\mathscr{A}_{\mathrm{p}}$ *— невырожденная матрица размерности* $n \times n$ *и* $\mathscr{X}^{(0)}$ *— интервальная матрица той же размерности, такая что* $\mathscr{A}_{\mathrm{p}}^{-1} \in \mathscr{X}^{(0)}$. *Пусть последовательность* $\left\{\mathscr{X}^{(k)}\right\}_{k=0}^{\infty}$ *интервальных матриц вычисляется по формулам* (5). *Тогда*

каждое приближение $\mathscr{X}^{(k)}$, $k \geqslant 0$, *содержит* $\mathscr{A}_{\mathrm{p}}^{-1}$; (6)

последовательность $\left\{\mathscr{X}^{(k)}\right\}_{k=0}^{\infty}$ *сходится к* $\mathscr{A}_{\mathrm{p}}^{-1}$ *тогда и* (7)

только тогда, когда спектральный радиус $\rho\left(\mathscr{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m\left(\mathscr{X}^{(0)}\right)\right)$ *меньше* 1;

для матричной нормы $\|\cdot\|$ *последовательность* $\left\{d\left(\mathscr{X}^{k}\right)\right\}_{k=0}^{\infty}$ (8)

удовлетворяет условию $\left\|d\left(\mathscr{X}^{(k+1)}\right)\right\| \leqslant \gamma \left\|d\left(\mathscr{X}^{(k)}\right)\right\|^{r}, \quad \gamma \geqslant 0,$

т. е. R-порядок метода (5) *удовлетворяет неравенству* $O_R\left((5), \mathscr{A}^{-1}\right) \geqslant r$ *(см. приложение А, теорема* 2).

Доказательство. (6): Для произвольной матрицы $m\left(\mathscr{X}^{(k)}\right) \in M_{nn}(\mathbb{R})$ легко проверить соотношение

$$m\left(\mathscr{X}^{(k)}\right) \sum_{\nu=0}^{r-2} \left(\mathscr{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m\left(\mathscr{X}^{(k)}\right)\right)^{\nu} = \mathscr{A}_{\mathrm{p}}^{-1} - \mathscr{A}_{\mathrm{p}}^{-1} \left(\mathscr{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m\left(\mathscr{X}^{(k)}\right)\right)^{r-1}.$$

Для $k = 0$ утверждение (6) верно в силу условия теоремы.

Допустим теперь, что $\mathscr{A}_{\mathrm{p}}^{-1} \in \mathscr{X}^{(k)}$. Используя только что полученное равенство и соотношения (10' из микромодуля 29), получаем

$$\mathscr{A}_p^{-1} = m(\mathscr{X}^{(k)}) \sum_{\nu=0}^{r-2} (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))^\nu + \mathscr{A}^{-1} (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))^{r-1}$$

$$\in m(\mathscr{X}^{(k)}) \sum_{\nu=0}^{r-2} (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))^\nu + \mathscr{X}^{(k)} (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))^{r-1}$$

$$= \mathscr{X}^{(k+1)}.$$

Этим завершается доказательство соотношения (6) методом математической индукции.

(7): Используя равенства (2)—(4) для срединного отображения, участвующего в формулах (5), получаем для последовательности $\{m(\mathscr{X}^{(k)})\}_{k=0}^{\infty}$ следующую рекуррентную формулу:

$$m(\mathscr{X}^{(k+1)}) = m(\mathscr{X}^{(k)}) \sum_{\nu=0}^{r-1} (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))^\nu.$$

Это — обобщение итерационной процедуры Шульца. Умножая обе части этого равенства на $\mathscr{A}_p$, получаем

$$\mathscr{A}_p m(\mathscr{X}^{(k+1)}) = (\mathscr{I}_p - (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))) \sum_{\nu=0}^{r-1} (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))^\nu$$

$$= \mathscr{I}_p - (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))^r$$

или

$$\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k+1)}) = (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))^r$$

$$= (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(0)}))^{r^{(k+1)}}.$$

Отсюда следует, что

$$\lim_{k\to\infty} m(\mathscr{X}^{(k)}) = \mathscr{A}_p^{-1} \Leftrightarrow \lim_{k\to\infty} (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(0)}))^k = \mathscr{O}_p$$

$$\Leftrightarrow \rho(\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(0)})) < 1.$$

Теперь мы покажем, что последовательность $\{\mathscr{X}^{(k)}\}_{k=0}^{\infty}$ сходится к $\mathscr{A}_p^{-1}$ тогда и только тогда, когда последовательность $\{m(\mathscr{X}^{(k)})\}_{k=0}^{\infty}$ срединных матриц сходится к $\mathscr{A}_p^{-1}$. Действительно, рассмотрим последовательность $\{d(\mathscr{X}^{(k)})\}_{k=0}^{\infty}$, которая в силу (12 из микромодуля 29) и (19 из микромодуля 29) удовлетворяет рекуррентному соотношению

$$d(\mathscr{X}^{(k+1)}) = d(\mathscr{X}^{(k)}) \left| (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))^{r-1} \right|.$$

Если мы имеем теперь $\lim\limits_{k\to\infty} m(\mathscr{X}^{(k)}) = \mathscr{A}_p^{-1}$, то из последнего соотношения следует, что

$$\lim_{k\to\infty} d(\mathscr{X}^{(k)}) = \mathscr{O}_p.$$

С другой стороны, из непрерывности отображенияm и равенства (4) сразу получается, что

$$\lim_{k\to\infty} \mathscr{X}^{(k)} = \mathscr{A}_p^{-1}$$

влечет за собой $\lim\limits_{k\to\infty} m(\mathscr{X}^{(k)}) = \mathscr{A}_p^{-1}$.

Так как выше уже было показано, что условие

$$\rho(\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(0)})) < 1$$

необходимо и достаточно для сходимости последовательности $\{m(\mathscr{X}^{(k)})\}_{k=0}^{\infty}$, мы получаем (7).

(8): Имеем

$$\begin{aligned} d(\mathscr{X}^{(k+1)}) &= d(\mathscr{X}^{(k)}) \left| (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))^{r-1} \right| \\ &= d(\mathscr{X}^{(k)}) \left| (\mathscr{A}_p \mathscr{A}_p^{-1} - \mathscr{A}_p m(\mathscr{X}^{(k)}))^{r-1} \right| \\ &\leqslant d(\mathscr{X}^{(k)}) \left(\|\mathscr{A}_p\| \left\| \mathscr{A}_p^{-1} - m(\mathscr{X}^{(k)}) \right\| \right)^{r-1} \\ &\leqslant d(\mathscr{X}^{(k)}) 2^{-(r-1)} (\|\mathscr{A}_p\| d(\mathscr{X}^{(k)}))^{r-1}. \end{aligned}$$

Мы используем монотонную и мультипликативную матричную норму $\|\cdot\|'$, поэтому из только что доказанного соотношения следует

$$\|d(\mathscr{X}^{(k+1)})\|' \leqslant 2^{-(r-1)} \|\mathscr{A}_p\|^{(r-1)} \|d(\mathscr{X}^{(k)})\|'^{r}.$$

Неравенство

$$\|\mathscr{B}_p\| \gamma_1 \leqslant \|\mathscr{B}_p\|' \leqslant \gamma_2 \|\mathscr{B}_p\|, \quad \gamma_1 > 0, \ \gamma_2 > 0,$$

верно для любой матричной нормы $\|\cdot\|$. Из этого неравенства следует

$$\|d(\mathscr{X}^{(k+1)})\| \gamma_1 \leqslant 2^{-(r-1)} \gamma_2^{-(r-1)} \|\mathscr{A}_p\|^{r-1} \gamma_2^{n} \|d(\mathscr{X}^{(k)})\|^{r},$$

что и доказывает (8).

Из доказательства видно, что сходимость имеет место для произвольной матрицы $\mathscr{X}^{(0)}$, не обязательно содержащей $\mathscr{A}_p^{-1}$.

В этом случае, однако, последовательные приближения не обязаны содержать $\mathscr{A}_p^{-1}$. Отметим, что критерий (7) зависел не от всей локализующей матрицы $\mathscr{X}_p^{(0)}$, а только от ее срединной матрицы $m(\mathscr{X}^{(0)})$. При этом ширина $a(\mathscr{X}^{(0)})$ может быть произвольной. Это

значит, что имея подходящую аппроксимацию $m(\mathscr{X}^{(0)})$ матрицы $\mathscr{A}_{\mathrm{p}}^{-1}$, удовлетворяющую неравенству $\rho(\mathscr{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m(\mathscr{X}^{(0)})) < 1$, с помощью определенных оценок понорме всегда можно построить интервальную матрицу $\mathscr{X}^{(0)}$, такую что $\mathscr{A}_{\mathrm{p}}^{-1} \in \mathscr{X}^{(0)}$. Тогда последовательные приближения, полученные согласно (5), сходятся к $\mathscr{A}_{\mathrm{p}}^{-1}$ по теореме 1.

Так как последовательные приближения из (5) всегда содержат $\mathscr{A}_{\mathrm{p}}^{-1}$ в силу (6), кажется естественным брать пересечение следующего приближения с предыдущим и продолжать итерационный процесс с этим новым потенциально улучшенным приближением. Это приводит к следующей итерационной процедуре:

$$\begin{cases} \mathscr{Y}^{(k+1)} = m(\mathscr{X}^{(k)}) \sum\limits_{\nu=0}^{r-2} (\mathscr{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m(\mathscr{X}^{(k)}))^{\nu} \\ \qquad + \mathscr{X}^{(k)} (\mathscr{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m(\mathscr{X}^{(k)}))^{r-1}, \\ \mathscr{X}^{(k+1)} = \mathscr{Y}^{(k+1)} \cap \mathscr{X}^{(k)}, \quad k \geqslant 0. \end{cases} \tag{9}$$

Применяя эту итерационную процедуру, получаем монотонную последовательность $\mathscr{X}^{(0)} \supseteq \mathscr{X}^{(1)} \supseteq \mathscr{X}^{(2)} \supseteq \ldots$ локализаций для матрицы $\mathscr{A}_{\mathrm{p}}^{-1}$. Следующий численный пример показывает, что в этом случае критерий (7), вообще говоря, не достаточен для сходимости.

Возьмем $r = 2$ и положим

$$\mathscr{A}_{\mathrm{p}} = \begin{pmatrix} 0.4 & 0.6 \\ -0.6 & 0.4 \end{pmatrix}, \quad \mathscr{X}^{(0)} = \begin{pmatrix} [-2, 4] & [-3, 3] \\ [-3, 3] & [-2, 4] \end{pmatrix},$$

откуда следует, что $m(\mathscr{X}^{(0)}) = \mathscr{I}_{\mathrm{p}}$. Мы получаем

$$\mathscr{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m(\mathscr{X}^{(0)}) = \begin{pmatrix} 0.6 & -0.6 \\ 0.6 & 0.6 \end{pmatrix},$$

откуда

$$\rho(\mathscr{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m(\mathscr{X}^{(0)})) < 1.$$

Поэтому процедура (5) с этим начальным приближением сходится к $\mathscr{A}_{\mathrm{p}}^{-1}$. Используя (9), получаем

$$\mathscr{Y}^{(1)} = m(\mathscr{X}^{(0)}) + \mathscr{X}^{(0)} (\mathscr{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m(\mathscr{X}^{(0)})) = \begin{pmatrix} [-2, 5.2] & [-4, 2.3] \\ [-3, 4.2] & [-2, 5.2] \end{pmatrix},$$

откуда следует, что $\mathscr{X}^{(1)} = \mathscr{X}^{(0)}$. Поэтому последовательность приближений, вычисленная по формулам (9), не сходится к $\mathscr{A}_{\mathrm{p}}^{-1}$

в противоположность последовательности, вычисленной по формулам (5). Условие сходимости для итерации (9) содержится в следующей теореме.

Теорема 2. *Пусть* $\mathcal{A}_{\mathrm{p}}$ *— невырожденная матрица размерности* $n \times n$, *а* $\mathcal{X}^{(0)}$ *— интервальная матрица размерности* $n \times n$, *такая что* $\mathcal{A}_{\mathrm{p}}^{-1} \in \mathcal{X}^{(0)}$. *Тогда*

каждое приближение $\mathcal{X}^{(k)}$, $k \geqslant 0$, *содержит* $\mathcal{A}_{\mathrm{p}}^{-1}$; (6')

если неравенство $\rho(|\mathcal{I}_{\mathrm{p}} - \mathcal{A}_{\mathrm{p}}\mathcal{X}|) < 1$ *выполнено для* (10)
всех $\mathcal{X}_{\mathrm{p}} \in \mathcal{X}^{(0)}$, *то последовательность* $\{\mathcal{X}^{k}\}_{k=0}^{\infty}$
сходится к $\mathcal{A}_{\mathrm{p}}^{-1}$;

последовательность $\{d(\mathcal{X}^{(k)})\}_{k=0}^{\infty}$ *может быть следую-* (8′)
щим образом ограничена в матричной норме $\|\cdot\|$:

$$\|d(\mathcal{X}^{(k+1)})\| \leqslant \gamma^{r}\|d(\mathcal{X}^{(k)})\|^{r}, \quad \gamma \geqslant 0,$$

т. е. R-порядок итерационной процедуры (9) *удовлетворяет неравенству* $O_R((9), \mathcal{A}_{\mathrm{p}}^{-1}) \geqslant r$ *(см. приложение* А, *теорема* 2).

Доказательство. (6'): Как и в доказательстве утверждения (6), мы устанавливаем сначала, что $\mathcal{A}_{\mathrm{p}}^{-1} \in \mathcal{Y}^{(k+1)}$, откуда ввиду $\mathcal{A}_{\mathrm{p}}^{-1} \in \mathcal{X}^{(k)}$ немедленно следует $\mathcal{A}_{\mathrm{p}}^{-1} \in \mathcal{X}^{(k+1)}$.

(10): В силу следствия 8 из микромодуля 29 последовательные приближения

$$\mathcal{X}^{(0)} \supseteq \mathcal{X}^{(1)} \supseteq \mathcal{X}^{(2)} \supseteq \ldots,$$

всегда сходятся к некоторой интервальной матрице $\mathcal{X}$. Теперь покажем, что в условиях нашей теоремы выполнено равен-ство $d(\mathcal{X}) = \mathcal{O}_{\mathrm{p}}$. Положив

$$\mathcal{Y} = m(\mathcal{X}) \sum_{\nu=0}^{r-2} (\mathcal{I}_{\mathrm{p}} - \mathcal{A}_{\mathrm{p}} m(\mathcal{X}))^{\nu} + \mathcal{X}(\mathcal{I}_{\mathrm{p}} - \mathcal{A}_{\mathrm{p}} m(\mathcal{X}))^{r-1},$$

получаем $\mathcal{X} = (X_{ij} \cap Y_{ij}) \subseteq \mathcal{Y}$ в силу (9). Используя (11 из микромодуля 29), получим $d(\mathcal{X}) \leqslant d(\mathcal{Y})$. Для $d(\mathcal{X})$ имеем из (9) соотношение

$$d(\mathcal{X})|\mathcal{I}_{\mathrm{p}} - \mathcal{A}_{\mathrm{p}} m(\mathcal{X})|^{r-1} \geqslant d(\mathcal{X})|(\mathcal{I}_{\mathrm{p}} - \mathcal{A}_{\mathrm{p}} m(\mathcal{X}))^{r-1}| = d(\mathcal{Y}) \geqslant d(\mathcal{X}),$$

откуда следует, что

$$d(\mathcal{X})(\mathcal{I}_{\mathrm{p}} - |\mathcal{I}_{\mathrm{p}} - \mathcal{A}_{\mathrm{p}} m(\mathcal{X})|^{r-1}) \leqslant \mathcal{O}_{\mathrm{p}}.$$

Из условия $\rho(|\mathcal{I}_{\mathrm{p}} - \mathcal{A}_{\mathrm{p}} m(\mathcal{X})|) < 1$ следует существование

матрицы $(\mathcal{I}_p - |\mathcal{I}_p - \mathcal{A}_p m(\mathcal{X})|^{r-1})^{-1}$. Эта обратная матрица также неотрицательна. Отсюда следует, что $d(\mathcal{X}) \leqslant \mathcal{O}_p$, т. е. $d(\mathcal{X}) = \mathcal{O}_p$. Ввиду (6') мы имеем поэтому $\mathcal{X} = \mathcal{A}_p^{-1}$.

(8'): Как и в доказательстве утверждения (8), мы показываем сначала, что для монотонной и мультипликативной матричной нормы $\|\cdot\|'$ имеет место неравенство

$$\|d(\mathcal{Y}^{(k+1)})\|' \leqslant \gamma \|d(\mathcal{X}^{(k)})\|'^{r}.$$

Отсюда с помощью (11 из микромодуля 29), монотонности нормы $\|\cdot\|'$ и включения $\mathcal{X}^{(k+1)} \subseteq \mathcal{Y}^{(k+1)}$ следует неравенство

$$\|d(\mathcal{X}^{k+1})\|' \leqslant \|d(\mathcal{Y}^{(k+1)})\|' \leqslant \gamma \|d(\mathcal{X}^{(k)})\|'^{r}.$$

Так же как и в доказательстве утверждения (8), мы используем теперь теорему об эквивалентности норм для доказательства утверждения (8).

В отличие от критерия (7) условие сходимости (10) зависит от ширины матрицы $\mathcal{X}^{(0)}$, локализующей $\mathcal{A}_p^{-1}$. Эту зависимость легко охарактеризовать формулами. Если, например, матрица $\mathcal{X}^{(0)}$ удовлетворяет для монотонной мультипликативной нормы $\|\cdot\|$ неравенству $\|\mathcal{I}_p - \mathcal{A}_p m(\mathcal{X}^{(0)})\| < 1$, то условие

$$\|d(\mathcal{X}^{(0)})\| < 2(1 - \|\mathcal{I}_p - \mathcal{A}_p m(\mathcal{X}^{(0)})\|)/\|\mathcal{A}_p\| \qquad (11)$$

достаточно для того, чтобы $\|\mathcal{I}_p - \mathcal{A}_p \mathcal{X}\| < 1$ было верно для всех $\mathcal{X}_p \in \mathcal{X}^{(0)}$. Рассмотрим теперь вкратце вопрос о нахождении подходящей интервальной матрицы $\mathcal{X}^{(0)}$. Допустим, что $\mathcal{A}_p$ можно представить в виде

$$\mathcal{A}_p = \mathcal{I}_p - \mathcal{B}_p, \text{ где } \|\mathcal{B}_p\| < 1.$$

При $m(\mathcal{X}^{(0)}) := \mathcal{I}_p$ мы имеем

$$\|\mathcal{I}_p - \mathcal{A}_p m(\mathcal{X}^{(0)})\| = \|\mathcal{B}_p\| < 1,$$

так что последовательность (5) сходится в силу критерия (7) для любой интервальной матрицы $\mathcal{X}^{(0)}$, для которой $m(\mathcal{X}^{(0)}) = \mathcal{I}_p$. Чтобы обеспечить соотношение $\mathcal{A}_p^{-1} \in \mathcal{X}^{(1)}$, рассмотрим равенство

$$\mathcal{A}_p \mathcal{X}_p = (\mathcal{I}_p - \mathcal{B}_p)\mathcal{X}_p = \mathcal{I}_p,$$

или

$$\mathcal{X}_p = \mathcal{B}_p \mathcal{X}_p + \mathcal{I}_p.$$

Из него следует в силу мультипликативности матричной нормы $\|\cdot\|$, что

$$\|\mathscr{X}_p\| \leqslant a := 1/(1 - \|\mathscr{B}_p\|).$$

Если теперь мы используем норму, задаваемую суммами по столбцам или суммами по строкам, то получим

$$-a \leqslant x_{ij} \leqslant a, \quad 1 \leqslant i, \ j \leqslant n,$$

для элементов матрицы $\mathscr{X}_p = (x_{ij})$. Для матрицы $\mathscr{X}^{(0)}$ с элементами

$$X_{ij}^{(0)} = \begin{cases} [-a, a] & \text{для } i \neq j, \\ [-a, 2 + a] & \text{для } i = j \end{cases}$$

имеем $\mathscr{A}_p^{-1} \in \mathscr{X}^{(0)}$ и $m(\mathscr{X}^{(0)}) = \mathscr{T}_p$. Поэтому итерационный метод (5) сходится к $\mathscr{A}_p^{-1}$ в силу теоремы 1. Если теперь неравенство (11) выполняется после некоторого шага итерации, то процесс можно продолжать дальше по формулам (9).

При практическом выполнении алгоритма (5) встречающиеся выражения вычисляются по аналогии со схемой Горнера.

Это дает формулу

$$\mathscr{X}^{(k+1)} = \ldots (\mathscr{X}^{(k)} \mathscr{T}_p^{(k)} + m(\mathscr{X}^{(k)})) \mathscr{T}_p^{(k)} + m(\mathscr{X}^{(k)}) \mathscr{T}_p^{(k)} \ldots) \mathscr{T}_p^{(k)} + m(\mathscr{X}^{(k)}), \tag{5'}$$

где

$$\mathscr{T}_p^{(k)} = \mathscr{T}_p - \mathscr{A}_p m(\mathscr{X}_p^{(k)}).$$

Так как умножение матриц становится неассоциативным при появлении интервальных матриц, мы имеем в общем случае

$$\mathscr{X}^{(k)} (\mathscr{T}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))^{r-1} \neq (\ldots (\mathscr{X}^{(k)} (\mathscr{T}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))) \ldots) \times (\mathscr{T}_p - \mathscr{A}_p m(\mathscr{X}^{(k)})).$$

Даже при одной и той же начальной матрице формулы (5) и (5') порождают в общем случае разные последовательности. Однако теорема 1 все же верна и для итераций (5'). Рассмотрим теперь объем вычислений, нужных на каждом шаге (5').

Если $\mathscr{A}_p$ — матрица размерности $n \times n$, то (5') требует на каждом шаге

$$rn^3 \text{ умножений и } rn^3 - n^2 + n \text{ сложений.}$$

Даже те члены в (5'), которые, как $\mathscr{T}_p^{(k)}$, не содержат ничего интервального, приходится вычислять в интервальной арифметике,

чтобы обеспечить локализацию матрицы $\mathscr{A}_p^{-1}$. Если пренебречь более низкими степенями n, то мы увидим, что объем вычислений по алгоритму (5') пропорционален r.

Теперь мы хотим оценить число k шагов итерации, которые требуются, чтобы, исходя из данного $\mathscr{X}^{(0)}$, достичь величины

$$\|d(\mathscr{X}^{(k)})\|,$$

меньшей, чем заранее предписанная погрешность.

Так же, как в доказательстве теоремы 1, мы получаем для (5') следующее соотношение:

$$d(\mathscr{X}^{(k+1)} = d(\mathscr{X}^{(k)}) | (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)})) |^{r-1}$$
$$\leqslant d(\mathscr{X}^{(k)}) | (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(0)})) |^{r^k(r-1)}$$

т. е.

$$d(\mathscr{X}^{(k+1)}) \leqslant d(\mathscr{X}^{(0)}) \prod_{\nu=0}^{k} | (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(0)})) |^{r^\nu(r-1)}.$$

Если по-прежнему мы используем монотонную и мультипликативную матричную норму $\|\cdot\|$ и допустим, что $\|\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(0)})\| < 1$, то получим, что

$$\|d(\mathscr{X}^{(k+1)})\| \leqslant \|d(\mathscr{X}^{(0)})\| \left(\prod_{\nu=0}^{k} \|\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(0)})\|^{r^\nu} \right)^{r-1}$$
$$= \|d(\mathscr{X}^{(0)})\| \|\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(0)})\|^{r^{k+1}-1}.$$

Это выражение позволяет нам оценить $\bar{k}$ при сделанных предположениях

Исходя из соотношения

$$\|d(\mathscr{X}^{(k)})\| \leqslant \|d(\mathscr{X}^{(0)})\| \|\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(0)})\|^{r^k-1},$$

мы определим, при каком значении r итерационный метод требует наименьшего объема вычислений для достижения заданной точности для $\|d(\mathscr{X}^{(k)})\|$. Согласно предшествующим рассуждениям, этот объем вычислений можно считать пропорциональным величине r. Пусть теперь даны $r^{(1)}>1$ и $r^{(2)}>1$, причем $r^{(1)} \neq r^{(2)}$. После $p^{(1)}$ (соответственно $p^{(2)}$) шагов итерации (5') со значением $r = r^{(1)}$ (соответственно $r = r^{(2)}$) при одном и том же начальном значении $\mathscr{X}^{(0)}$ мы выполним один и тот же объем вычислений. Иными словами, имеем

$$r^{(1)}p^{(1)} = r^{(2)}p^{(2)}.$$

Точность, достигнутую при использовании этих методов, можно оценить величиной $(r^{(1)})^{p^{(1)}} - 1$ (соответственно $(r^{(2)})^{p^{(2)}} - 1$).

Мы требуем от «оптимальной» итерационной процедуры $r = r^{(1)}$ чтобы для всех других значений $r^{(2)}$ и количества шагов $p^{(1)}$, $p^{(2)}$ мы имели бы

$$(r^{(1)})^{p^{(1)}} > (r^{(2)})^{p^{(2)}},$$

что ввиду $p^{(2)} = r^{(1)}p^{(1)}/r^{(2)}$ эквивалентно неравенству

$$(r^{(1)})^{1/r^{(1)}} > (r^{(2)})^{1/r^{(2)}}.$$

Так как функция $x^{1/x}$ для начальных x имеет максимум при x=3, получаем, что итерация (5') оптимальна в описанном смысле при этом значении.

Заметим еще, что метод (5) можно применять и для комплексных матриц, используя арифметику в $R(\mathbb{C})$. При этом будет верна теорема 1. Мы упомянем в этой связи более общие исследования. Рассмотрим теперь методы монотонной локализации обратной матрицы, обладающие свойствами, похожими на свойства метода (9). Основные вычисления в них вообще не используют интервальной арифметики. Верхняя и нижняя границы вычисляются по отдельным формулам. Этот метод, однако, применим лишь в случае, когда $\mathcal{A}_p^{-1} \geqslant \mathcal{O}_p$.

Итак, пусть $\mathcal{A}_p$ —невырожденная матрица и $r \geq 2$ — натуральное число. Рассмотрим итерационную процедуру

$$\begin{cases} \mathcal{X}_p^{(k+1)} = \mathcal{X}_p^{(k)} + (\mathcal{I}_p - \mathcal{X}_p^{(k)}\mathcal{A}_p)\sum\limits_{\nu=0}^{r-2}(\mathcal{I}_p - \mathcal{X}_p^{(k)}\mathcal{A}_p)^\nu \mathcal{X}_p^{(k)}, \\ \mathcal{Y}_p^{(k+1)} = \mathcal{Y}_p^{(k)} + (\mathcal{I}_p - \mathcal{Y}_p^{(k)}\mathcal{A}_p)\sum\limits_{\nu=0}^{r-2}(\mathcal{I}_p - \mathcal{X}_p^{(k)}\mathcal{A}_p)^\nu \mathcal{X}_p^{(k)}, \\ k \geqslant 0. \end{cases} \tag{12}$$

с заданными $\mathcal{X}_p^{(0)}$, $\mathcal{Y}_p^{(0)}$.

Исполняя эту процедуру, мы получим две последовательности точечных матриц, для которых верно следующее утверждение.

Теорема 3. *Пусть $\mathcal{A}_p$ — невырожденная матрица размерности $n \times n$, причем $\mathcal{A}_p^{-1} \geqslant \mathcal{O}_p$. Пусть далее $\mathcal{X}_p^{(0)}$, $\mathcal{Y}_p^{(0)}$ — две матрицы размерности $n \times n$, для которых верно*

$$\mathcal{X}_p^{(0)} \geqslant \mathcal{O}_p \text{ и } \mathcal{X}_p^{(0)}\mathcal{A}_p \leqslant \mathcal{I}_p \leqslant \mathcal{Y}_p^{(0)}\mathcal{A}_p.$$

Пусть последовательности $\{\mathcal{X}_p^{(k)}\}_{k=0}^{\infty}$ и $\{\mathcal{Y}_p^{(k)}\}_{k=0}^{\infty}$ вычислены по формулам (12). *Тогда верны следующие утверждения:*

$$O_p \leqslant \mathscr{X}_p^{(0)} \leqslant \ldots \leqslant \mathscr{X}_p^{(k)} \leqslant \mathscr{X}_p^{(k+1)} \leqslant \ldots \leqslant \mathscr{A}_p^{-1} \leqslant \ldots \leqslant \mathscr{Y}_p^{(k+1)} \leqslant \mathscr{Y}_p^{(k)} \leqslant \ldots \leqslant \mathscr{Y}_p^{(0)}. \quad (13)$$

(7') *Обе последовательности*

$$\{\mathscr{X}_p^{(k)}\}_{k=0}^{\infty}, \quad \{\mathscr{Y}^{(k)}\}_{k=0}^{\infty}$$

сходятся к $\mathscr{A}_p^{-1}$ *тогда и только тогда, когда спектральный радиус* $\rho(\mathscr{I}_p - \mathscr{X}_p^{(0)}\mathscr{A}_p)$ *меньше* 1.

Если процедура сходится, то величины

$$d^{(k)} = \|\mathscr{Y}_p^{(k)} - \mathscr{X}_p^{(k)}\|$$

удовлетворяют соотношению (14) $d^{(k+1)} \leqslant \gamma (d^{(k)})^r, \quad \gamma \geqslant 0.$

Поэтому, если понимать (12) *как метод итераций для вычисления интервальных матриц* $([x_{ij}^{(k)}, y_{ij}^{(k)}])$, *то верно*

$$O_R((12), \mathscr{A}_p^{-1}) \geqslant r$$

(см. приложение А, *теорема* 2).

Доказательство. (13): Докажем соотношение

$$O_p \leqslant \mathscr{X}_p^{(0)} \leqslant \ldots \leqslant \mathscr{X}_p^{(k-1)} \leqslant \mathscr{X}_p^{(k)},$$
$$\mathscr{Y}_p^{(k)} \leqslant \mathscr{Y}_p^{(k-1)} \leqslant \ldots \leqslant \mathscr{Y}_p^{(0)},$$
$$\mathscr{X}_p^{(k)}\mathscr{A}_p \leqslant \mathscr{I}_p \leqslant \mathscr{Y}_p^{(k)}\mathscr{A}_p$$

математической индукцией по $k \geq 0$. Эти неравенства выполнены для $k = 0$ по условию теоремы. Из

$$\mathscr{I}_p - \mathscr{Y}_p^{(k)}\mathscr{A}_p \leqslant O_p \leqslant \mathscr{I}_p - \mathscr{X}_p^{(k)}\mathscr{A}_p, \quad \mathscr{X}_p^{(k)} \geqslant O_p$$

следует, что

$$(\mathscr{I}_p - \mathscr{Y}_p^{(k)}\mathscr{A}_p) \sum_{\nu=0}^{r-2} (\mathscr{I}_p - \mathscr{X}_p^{(k)}\mathscr{A}_p)^\nu \mathscr{X}_p^{(k)} \leqslant O_p \leqslant (\mathscr{I}_p - \mathscr{X}_p^{(k)}\mathscr{A}_p) \sum_{\nu=0}^{r-2} (\mathscr{I}_p - \mathscr{X}_p^{(k)}\mathscr{A}_p)^\nu \mathscr{X}_p^{(k)},$$

т. е

$$\mathscr{X}_p^{(k)} \leqslant \mathscr{X}_p^{(k+1)}, \quad \mathscr{Y}_p^{(k+1)} \leqslant \mathscr{Y}_p^{(k)}.$$

Из

$$\mathcal{I}_p - \mathcal{X}_p^{(k+1)}\mathcal{A}_p = \mathcal{I}_p - \mathcal{X}_p^{(k)}\mathcal{A}_p - (\mathcal{I}_p - \mathcal{X}_p^{(k)}\mathcal{A}_p)$$
$$\times \sum_{\nu=0}^{r-2} (\mathcal{I}_p - \mathcal{X}_p^{(k)}\mathcal{A}_p)^\nu \mathcal{X}_p^{(k)}\mathcal{A}_p = (\mathcal{I}_p - \mathcal{X}_p^{(k)}\mathcal{A}_p)^r \geqslant \mathcal{O}_p$$

и

$$\mathcal{Y}_p^{(k+1)}\mathcal{A}_p - \mathcal{I}_p = \mathcal{Y}_p^{(k)}\mathcal{A}_p - \mathcal{I}_p - (\mathcal{Y}_p^{(k)}\mathcal{A}_p - \mathcal{I}_p)$$
$$\times \sum_{\nu=0}^{r-2} (\mathcal{I}_p - \mathcal{X}_p^{(k)}\mathcal{A}_p)^\nu \mathcal{X}_p^{(k)}\mathcal{A}_p = (\mathcal{Y}_p^{(k)}\mathcal{A}_p - \mathcal{I}_p)(\mathcal{I}_p - \mathcal{X}_p^{(k)}\mathcal{A}_p)^{r-1} \geqslant \mathcal{O}_p$$

следует, что

$$\mathcal{X}_p^{(k+1)}\mathcal{A}_p \leqslant \mathcal{I}_p \leqslant \mathcal{Y}_p^{(k+1)}\mathcal{A}_p.$$

Ввиду $\mathcal{A}_p^{-1} \geqslant \mathcal{O}_p$ получаем

$$\mathcal{X}_p^{(k+1)} \leqslant \mathcal{A}_p^{-1} \leqslant \mathcal{Y}_p^{(k+1)}.$$

(7'): Используя соотношения

$$\mathcal{I}_p - \mathcal{X}_p^{(k+1)}\mathcal{A}_p = (\mathcal{I}_p - \mathcal{X}_p^{(k)}\mathcal{A}_p)^r,$$
$$\mathcal{Y}_p^{(k+1)}\mathcal{A}_p - \mathcal{I}_p = (\mathcal{Y}_p^{(k)}\mathcal{A}_p - \mathcal{I}_p)(\mathcal{I}_p - \mathcal{X}_p^{(k)}\mathcal{A}_p)^{r-1},$$

установленные при доказательстве неравенства (13), можем показать по индукции, что

$$\mathcal{I}_p - \mathcal{X}_p^{(k)}\mathcal{A}_p = (\mathcal{I}_p - \mathcal{X}_p^{(0)}\mathcal{A}_p)^{r^k},$$
$$\mathcal{Y}_p^{(k)}\mathcal{A}_p - \mathcal{I}_p = (\mathcal{Y}_p^{(0)}\mathcal{A}_p - \mathcal{I}_p)(\mathcal{I}_p - \mathcal{X}_p^{(0)}\mathcal{A}_p)^{r^k-1},$$

откуда и следует нужное утверждение.

(14): Снова используя соотношения, установленные при доказательстве (13), получаем

$$\|\mathcal{A}_p^{-1} - \mathcal{X}_p^{(k+1)}\| \leqslant \|\mathcal{A}_p\|^{r-1}\|\mathcal{A}_p^{-1} - \mathcal{X}_p^{(k)}\|^r,$$
$$\|\mathcal{Y}_p^{(k+1)} - \mathcal{A}_p^{-1}\| \leqslant \|\mathcal{A}_p\|^{r-1}\|\mathcal{Y}_p^{(k)} - \mathcal{A}_p^{-1}\|\,\|\mathcal{A}_p^{-1} - \mathcal{X}_p^{(k)}\|^{r-1}.$$

С помощью монотонной матричной нормы эту оценку можно продолжить следующим образом:

$$\|\mathcal{Y}_p^{(k+1)} - \mathcal{X}_p^{(k+1)}\| \leqslant \|\mathcal{A}_p^{-1} - \mathcal{X}_p^{(k+1)}\| + \|\mathcal{Y}_p^{(k+1)} - \mathcal{A}_p^{-1}\|$$
$$\leqslant 2\|\mathcal{A}_p\|^{r-1}\|\mathcal{Y}_p^{(k)} - \mathcal{X}_p^{(k)}\|^r,$$

т. е. $d^{(k+1)} \leqslant \gamma(d^{(k)})^r$, $\gamma = 2\|\mathcal{A}_p\|^{r-1}$. Тепень нужное соотношение следует из теоремы 2 приложения А.

В методе (12) интервальные операции не используются. Несмотря на это, он порождает, подобно методу (9), монотонную последовательность границ для матрицы $\mathcal{A}_p^{-1}$. Был дан также

необходимый и достаточный критерий сходимости, аналогичный критерию для метода (5). Однако применимость метода (12) ограничивается требованием $\mathcal{A}_p^{-1} \geqslant \mathcal{O}_p$.

Для определенных классов матриц можно указать вполне общие начальные значения $\mathcal{X}_p^{(0)}$, $\mathcal{Y}_p^{(0)}$, при которых (12) всегда будет сходиться. Если матрица $\mathcal{A}_p = (a_{ij})$ удовлетворяет условиям

$$a_{ii} > 0, \quad 1 \leqslant i \leqslant n,$$

$$a_{ij} \leqslant 0, \quad 1 \leqslant i, \ j \leqslant n, \quad i \neq j,$$

$$\sum_{i=1}^{n} a_{ij} > 0, \quad 1 \leqslant j \leqslant n,$$

то $\mathcal{A}_p$ является M-матрицей. Начальные матрицы $\mathcal{X}_p^{(0)} = (x_{ij})$ и $\mathcal{Y}_p^{(0)} = (y_{ij})$, такие что

$$x_{ij} = \begin{cases} 1/a_{ii} & \text{для } i = j \\ 0 & \text{в противном случае,} \end{cases} \quad \text{и} \quad y_{ij} = 1 \Big/ \sum_{\nu=1}^{n} a_{\nu i}, \ 1 \leqslant i, \ j \leqslant n,$$

удовлетворяют условиям теоремы 3.

Рассмотрим теперь невырожденную вещественную матрицу $\mathcal{A}_p = (a_{ij})$ размерности $n \times n$, строки которой были переставлены, чтобы стало возможным разложение

$$\mathcal{A}_p = (\mathcal{J}_p + \mathcal{L}_p^*)\,\mathcal{U}_p^*, \tag{15}$$

где $\mathcal{J}_p$ — единичная матрица, $\mathcal{L}_p^*$ — строго нижняя треугольная матрица, $\mathcal{U}_p^*$ — верхняя треугольная матрица. Как известно, это разложение можно найти с помощью метода Гаусса. Мы хотим описать итерационную структуру, постепенно улучшающую границы, между которыми заключены элементы матриц $\mathcal{L}_p^*$ и $\mathcal{U}_p^*$. Такие границы, включающие все ошибки округления, можно вычислить, если использовать при исполнении метода Гаусса для точечной матрицы $\mathcal{A}_p$ машинную интервальную арифметику с округлением наружу. Допустим поэтому, что $\mathcal{L}^{(0)}$ — строго нижняя треугольная интервальная матрица и $\mathcal{U}^{(0)}$ — верхняя треугольная матрица, для которых

$$\mathcal{L}_p^* \in \mathcal{L}^{(0)}, \quad \mathcal{U}_p^* \in \mathcal{U}^{(0)}. \tag{16}$$

Предположим сначала, что $\mathcal{L}_p$ и $\mathcal{U}_p$ — произвольные, но фиксированные треугольные матрицы, причем $\mathcal{L}_p$ — строго нижняя, а $\mathcal{U}_p$ — верхняя. Тогда мы имеем из (15)

$$\mathcal{R}_p = (r_{ik}) := \mathcal{A}_p - (\mathcal{I}_p + \mathcal{L}_p)\,\mathcal{U}_p = (\mathcal{I}_p + \mathcal{L}_p^*)\,(\mathcal{U}_p^* - \mathcal{U}_p) + (\mathcal{L}_p^* - \mathcal{L}_p)\,\mathcal{U}_p.$$

Будем считать матрицы $\mathcal{U}_p^* - \mathcal{U}_p$ и $\mathcal{L}_p^* - \mathcal{L}_p$ неизвестными, а множитель $\mathcal{I}_p + \mathcal{L}_p^*$ перед $\mathcal{U}_p^* - \mathcal{U}_p$ известным и будем затем решать это уравнение путем поочередного нахождения сначала строки матрицы $\mathcal{U}_p^* - \mathcal{U}_p$, а затем столбца матрицы $\mathcal{L}_p^* - \mathcal{L}_p$. В предложении $u_{ij} = 0$ для

$$1 \leqslant i \leqslant n$$

получаем тогда

$$\begin{cases} u_{ik}^* - u_{ik} = r_{ik} - \sum\limits_{j=1}^{i-1} l_{ij}^* (u_{jk}^* - u_{jk}) - \sum\limits_{j=1}^{i-1} (l_{ij}^* - l_{ij})\, u_{jk}, \quad i \leqslant k \leqslant n, \\ l_{ki}^* - l_{ki} = \dfrac{1}{u_{ii}} \left\{ r_{ki} - \sum\limits_{j=1}^{i} l_{kj}^* (u_{ji}^* - u_{ji}) - \sum\limits_{j=1}^{i-1} (l_{kj}^* - l_{kj})\, u_{ji} \right\}, \quad i < k \leqslant n, \\ 1 \leqslant i \leqslant n. \end{cases}$$

Полагая

$$\left.\begin{aligned} r_{ik} &= a_{ik} - \sum_{j=1}^{i-1} l_{ij} u_{jk}, - u_{ik}, \quad i \leqslant k \leqslant n, \\ r_{ki} &= a_{ki} - \sum_{j=1}^{i} l_{kj} u_{ji}, \qquad\qquad i+1 \leqslant k \leqslant n, \end{aligned}\right\} \; 1 \leqslant i \leqslant n,$$

получаем тогда соотношения

$$\begin{cases} u_{ik}^* = a_{ik} - \sum\limits_{j=1}^{i-1} l_{ij}^* u_{jk} - \sum\limits_{j=1}^{i-1} l_{ij}^* (u_{jk}^* - u_{jk}), \quad i \leqslant k \leqslant n, \\ l_{ki}^* = \dfrac{1}{u_{ii}} \left\{ a_{ki} - \sum\limits_{j=1}^{i-1} l_{kj}^* u_{ji} - \sum\limits_{j=1}^{i} l_{kj}^* (u_{ji}^* - u_{ji}) \right\}, \quad i < k < n, \\ 1 \leqslant i \leqslant n. \end{cases}$$

Выделенные жирным шрифтом выражения происходят из множителя $\mathcal{I}_p + \mathcal{L}_p^*$, который мы предположили известным. Если воспользоваться предположением, что $\mathcal{L}_p^* \in \mathcal{L}^{(0)}$ для этих величин и выбрать $\mathcal{U}_p := \mathcal{U}_p^0 \in \mathcal{U}^{(0)}$, где $u_{ii}^{(0)} \neq 0$, $1 \leqslant i \leqslant n$, то получим для элементов матриц $\mathcal{L}_p^*$ и $\mathcal{U}_p^*$ следующие локализующие интервалы:

$$\begin{cases} u_{ik}^{*} \in U_{ik}^{(1)} = \left\{ a_{ik} - \sum\limits_{j=1}^{i-1} L_{ij}^{(1)} u_{jk}^{(0)} - \sum\limits_{j=1}^{i-1} L_{ij}^{(0)} \left(U_{jk}^{(1)} - u_{jk}^{(0)} \right) \right\} \cap U_{ik}^{(0)}, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad 1 \leqslant k \leqslant n, \\ l_{ki}^{*} \in L_{ki}^{(1)} = \left\{ \frac{1}{u_{ii}^{(0)}} \left(a_{ki} - \sum\limits_{j=1}^{i-1} L_{kj}^{(1)} u_{ji}^{(0)} - \sum\limits_{j=1}^{i} L_{kj}^{(0)} \left(U_{ji}^{(1)} - u_{ji}^{(0)} \right) \right) \right\} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \cap L_{ki}^{(0)}, \quad i < k \leqslant n, \end{cases}$$

$$1 \leqslant i \leqslant n.$$

Систематическое повторение этой процедуры приводит к следующему итерационному методу:

$$\begin{cases} U_{ik}^{(m+1)} = \left\{ a_{ik} - \sum\limits_{j=1}^{i-1} L_{ij}^{(m+1)} u_{jk}^{(m)} - \sum\limits_{j=1}^{i-1} L_{ij}^{(m)} \left(U_{jk}^{(m+1)} - u_{jk}^{(m)} \right) \right\} \cap U_{ik}^{(m)}, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i \leqslant k \leqslant n, \\ L_{ki}^{(m+1)} = \left\{ \frac{1}{u_{ii}^{(m)}} \left(a_{ki} - \sum\limits_{j=1}^{i-1} L_{kj}^{(m+1)} u_{ji}^{(m)} - \sum\limits_{j=1}^{i} L_{kj}^{(m)} \left(U_{ji}^{(m+1)} - u_{ji}^{(m)} \right) \right) \right\} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \cap L_{kj}^{(m)}, \quad i < k \leqslant n, \end{cases}$$

$$1 \leqslant i \leqslant n,$$

$$m \geqslant 0.$$

С помощью рассуждений, аналогичных проведенным при описании первого шага этого метода, можно показать, что в общем случае верно следующее утверждение.

Теорема 4. *Пусть матрица* $\mathcal{A}_{\mathrm{p}}$ *имеет разложение* $\mathcal{A}_{\mathrm{p}} = (\mathcal{I}_{\mathrm{p}} + \mathcal{L}_{\mathrm{p}}^{*}) \mathcal{U}_{\mathrm{p}}^{*}$. *Пусть* $\mathcal{L}_{\mathrm{p}}^{*} \in \mathcal{L}^{(0)}$, $\mathcal{U}_{\mathrm{p}}^{*} \in \mathcal{U}^{(0)}$. *Если* $\mathcal{L}_{\mathrm{p}}^{(m)} \in \mathcal{L}^{(m)}$, $\mathcal{U}_{\mathrm{p}}^{(m)} \in \mathcal{U}^{(m)}$, *то для всех* $m \geq 0$ *верно*

$$\mathcal{L}_{\mathrm{p}}^{*} \in \mathcal{L}^{(m+1)}, \ \mathcal{U}_{\mathrm{p}}^{*} \in \mathcal{U}^{(m+1)}. \qquad (18)$$

Мы отметим, что метод (17) может быть выполнен при начальных интервалах произвольной, но конечной ширины, если выполнено предположение (16). Единственное деление встречается при вычислении $L_{ki}^{(m+1)}$. Так как $0 \neq u_{ii}^{*} \in U_{ii}^{(m)}$, мы всегда можем выбрать $u_{ii}^{(m)} \in U_{ii}^{(m)}$ так, чтобы $u_{ii}^{(m)} \neq 0$.

Покажем теперь, что метод (17) при отсутствии ошибок округления дает точное треугольное разложение за конечное число шагов.

Теорема 5. *В условиях предыдущей теоремы метод* (17) *вычисляет точное разложение матрицы* $\mathcal{A}_0$ *размерности* $n \times n$ *самое большее за* $2n-1$ *шагов.*

Доказательство. При $m = 0$ мы получаем из (17) для i=l, что

$$U_{1k}^{(1)} = a_{1k} \cap U_{1k}^{(0)} = a_{1k} = u_{1k}^{*}, \quad 1 \leqslant k \leqslant n.$$

Это значит, что для произвольных начальных матриц, удовлетворяющих условиям локализации (16), матрица $\mathcal{U}^{(1)}$ имеет в первой строке значения, нужные для треугольного разложения. Матрица $\mathcal{U}^{(2)}$ имеет нужные значения в первой строке точно так же, как $\mathcal{U}^{(1)}$. Поэтому для первою столбца матрицы $\mathcal{L}^{(2)}$ мы получаем из (17) при m= 1, i= 1, что

$$L_{k1}^{(2)} = \left(\frac{a_{k1}}{u_{11}}\right) \cap L_{k1}^{(1)} = \frac{a_{k1}}{a_{11}} = l_{k1}^{*}, \quad 2 \leqslant k \leqslant n.$$

Поэтому после второго шага итерации нужные значения имеют и первый столбец матрицы $\mathcal{L}^{(2)}$, и первая строка матрицы $\mathcal{U}^{(2)}$. Покажем теперь, что если первые i строк матрицы $\mathcal{U}^{(m)}$ и первые i столбцов матрицы $\mathcal{L}^{(m)}$ уже имеют нужные значения, то не менее чем i+1 строк матрицы $\mathcal{U}^{(m+1)}$ (и i столбцов матрицы $\mathcal{L}^{(m+1)}$) имеют нужные значения. Предыдущее рассуждение показывает, что это верно при $i = 0$. Для i>0 применим математическую индукцию.

Заметим прежде всего, что $\mathcal{L}^{(m+1)}$ и $\mathcal{U}^{(m+1)}$ имеют те же элементы, что $\mathcal{L}^{(m)}$ и $\mathcal{U}^{(m)}$ в первых i строках и столбцах. Таким образом эти элементы все еще имеют нужные значения. Это непосредственно следует из (17). Из (17) мы получаем также следующее соотношение для $(i + 1)$-й строки матрицы $\mathcal{U}^{(m+1)}$:

$$U_{i+1,k}^{(m+1)} = \left\{a_{i+1,k} - \sum_{j=1}^{i-1} l_{i+1,j}^{*} u_{jk}^{*}\right\} \cap U_{i+1,k}^{(m)} = u_{i+1,k}^{*}, \quad i+1 \leqslant k \leqslant n.$$

Чтобы завершить доказательство по индукции, мы должны показать, что если для некоторого $m \geq 0$ первые i строк матрицы $\mathcal{U}^{(m)}$ и первые i—1 столбцов матрицы $\mathcal{L}^{(m)}$ имеют нужные значения, то $\mathcal{L}^{(m+1)}$ имеют нужные значения в первых столбцах (а $\mathcal{U}^{(m+1)}$ — в первых i строках).

Это было доказано выше для i= 1. Для $i > 1$ заметим, что в силу (17) матрица $\mathcal{U}^{(m+1)}$ имеет нужные значения по крайней мере в тех же строках, что и $\mathcal{U}^{(m)}$, а матрица $\mathcal{L}^{(m+1)}$ — по крайней мере в тех же

строках, что $\mathscr{L}^{(m)}$. Тогда для i-го столбца матрицы $\mathscr{L}^{(m+1)}$ получаем из (17), что

$$L_{ki}^{(m+1)} = \left\{ \frac{1}{u_{ii}^{*}} \left(a_{ki} - \sum_{j=1}^{i-1} l_{kj}^{*} u_{ji}^{*} \right) \right\} \cap L_{ki}^{(m)} = l_{ki}^{*}, \quad i+1 \leqslant k \leqslant n.$$

Таким образом, самое большее через $2n$— 1 шаг мы получим точное решение.

Следующая теорема показывает, что данный метод обладает так называемым «квадратичным» свойством сходимости, хорошо знакомым по методу Ньютона — Рафсона.

Теорема 6. *Пусть*

$$d^{(m)} := \max_{1 \leqslant i,j \leqslant n} \{ \max \{ d(L_{ij}^{(m)}),\ d(U_{ij}^{(m)}) \} \}.$$

Тогда для метода (17) *имеет место*

$$d^{(m+1)} \leqslant \alpha (d^{(m)})^2,$$

где α *— неотрицательное вещественное число, не зависящее от* m, *т. е. на каждом шаге ширина интервала примерно возводится в квадрат.*

Доказательство (методом математической индукции). Из (17) мы получаем

$$\begin{cases} d(U_{ik}^{(m+1)}) \leqslant \sum_{j=1}^{i-1} d(L_{ij}^{(m+1)}) |u_{jk}^{(m)}| + \sum_{j=1}^{i-1} d(L_{ij}^{(m)}) |U_{jk}^{(m+1)} - u_{jk}^{(m)}| \\ \qquad + \sum_{j=1}^{i-1} |L_{ij}^{(m)}| d(U_{jk}^{(m+1)}), \quad i \leqslant k \leqslant n, \\ d(L_{ki}^{(m+1)}) \leqslant \frac{1}{|u_{ii}^{(m)}|} \left\{ \sum_{j=1}^{i-1} d(L_{kj}^{(m+1)}) |u_{ji}^{(m)}| \right. \\ \qquad + \sum_{j=1}^{i} d(L_{kj}^{(m)}) |U_{ji}^{(m+1)} - u_{ji}^{(m)}| \\ \qquad \left. + \sum_{j=1}^{i} |L_{kj}^{(m)}| d(U_{ji}^{(m+1)}) \right\}, \quad i < k \leqslant n, \\ 1 \leqslant i \leqslant n. \end{cases} \tag{17'}$$

Положим теперь при $0 \notin U_{ii}^{(0)}, \quad 1 \leqslant i \leqslant n,$

$$\begin{cases} \alpha_{ik} = \begin{cases} \sum\limits_{j=1}^{i-1} (\beta_{ij} |U_{jk}^{(0)}| + 1 + |L_{ij}^{(0)}| \alpha_{jk}), & i \leqslant k \leqslant n, \\ 0 \qquad\qquad \text{в противном случае,} \end{cases} \\ \beta_{kl} = \begin{cases} \left| \frac{1}{U_{ii}^{(0)}} \right| \left\{ \sum\limits_{j=1}^{i-1} \beta_{kj} |U_{ji}^{(0)}| + \sum\limits_{j=1}^{i} (1 + |L_{kj}^{(0)}| \alpha_{ji}) \right\}, & i < k \leqslant n, \\ 0 \qquad\qquad \text{в противном случае,} \end{cases} \end{cases} \tag{17''}$$

и наконец

$$\alpha = \max_{1 \leqslant i,\ k \leqslant n} \{\max \{\alpha_{ik}, \beta_{ki}\}\}.$$

Используя определение (17''), мы немедленно получаем из (17') при $i = 1$, что

$$d\left(U_k^{(m+1)}\right) \leqslant \alpha_{1k} \left(d^{(m)}\right)^2, \quad 1 \leqslant k \leqslant n,$$

и

$$d\left(L_{k1}^{(m+1)}\right) \leqslant \beta_{k1} \left(d^{(m)}\right)^2, \quad 1 < k \leqslant n.$$

Допустим теперь, что для первых $i \neq 1$ строк и столбцов имеет место при $(1 \leqslant l \leqslant i-1)$

$$\begin{cases} d\left(U_{lk}^{(m+1)}\right) \leqslant \alpha_{lk} \left(d^{(m)}\right)^2, & l \leqslant k \leqslant n, \\ d\left(L_{kl}^{(m+1)}\right) \leqslant \beta_{kl} \left(d^{(m)}\right)^2, & l \leqslant k \leqslant n. \end{cases} \tag{17'''}$$

Это очевидно при $i = 1$. Теперь мы получаем из (17') и (17''), что

$$d\left(U_{ik}^{(m+1)}\right) \leqslant \sum_{j=1}^{i-1} \beta_{ij} |U_{jk}^{(0)}| \left(d^{(m)}\right)^2 + \sum_{j=1}^{i-1} \left(d^{(m)}\right)^2 + \sum_{j=1}^{i-1} |L_{ij}^{(0)}| \alpha_{jk} \left(d^{(m)}\right)^2$$
$$= \alpha_{ik} \left(d^{(m)}\right)^2, \quad i \leqslant k \leqslant n.$$

Аналогично

$$d\left(L_{ki}^{(m+1)}\right) \leqslant \left| \frac{1}{U_{ii}^{(0)}} \right| \left\{ \sum_{j=1}^{i-1} \beta_{kj} |U_{ji}^{(0)}| \left(d^{(m)}\right)^2 + \sum_{j=1}^{i} \left(d^{(m)}\right)^2 \right.$$
$$\left. + \sum_{j=1}^{i} |L_{kj}^{(0)}| \alpha_{ji} \left(d^{(m)}\right)^2 \right\} = \beta_{ki} \left(d^{(m)}\right)^2, \quad i < k \leqslant n.$$

Из этих соотношений следует доказываемое утверждение

$$d^{(m+1)} \leqslant \alpha \left(d^{(m)}\right)^2.$$

Если исполнять (17) на вычислительной машине, применяя машинную интервальную арифметику, то в противоположность теореме 5 мы, вообще говоря, не получим за конечное число шагов

точного треугольного разложения матрицы $\mathscr{A}_p$. Рассмотрим, какой окончательной точности здесь можно достичь. В этих рассуждениях примем те же допущения, которые привели нас к формулам (4 15а) и (4 15Ь), а тем самым и к (4 22), (4.23). Последние две формулы были использованы при доказательстве формул (4 24) и (4.25). Теперь мы применим эти две формулы к (17). Это дает ширину вычисленных элементов

$$\begin{cases} d\left(\bar{U}_{ik}^{(m+1)}\right) \leqslant d\left(U_{ik}^{(m+1)}\right) + 2\varepsilon \sum_{j=1}^{2i-2} |\bar{S}_j| + 2\varepsilon\left(3 + 3\varepsilon + \varepsilon^2\right) \\ \qquad \times \sum_{j=1}^{i-1} \left(\left|L_{ij}^{(m)}\right| \left|\bar{U}_{jk}^{(m+1)} - u_{jk}^{(m)}\right| + \left|\bar{L}_{ij}^{(m+1)}\right| \left|u_{jk}^{(m)}\right|\right), \\ \qquad\qquad i \leqslant k \leqslant n, \\ d\left(\bar{L}_{ki}^{(m+1)}\right) \leqslant d\left(L_{ki}^{(m+1)}\right) + \left|\frac{1}{u_{ii}^{(m)}}\right| \left\{ 2\varepsilon \sum_{j=1}^{2i-1} |\bar{T}_j| + 2\varepsilon_{2i} |\bar{T}_{2i}| \right. \\ \qquad + 2\varepsilon\left(3 + 3\varepsilon + \varepsilon^2\right) \left(\sum_{j=1}^{i-1} \left|\bar{L}_{kj}^{(m+1)}\right| \left|u_{ji}^{(m)}\right| \right. \\ \left. \left. \qquad + \sum_{j=1}^{i} \left|L_{kj}^{(m)}\right| \left|\bar{U}_{ji}^{(m+1)} - u_{ji}^{(m)}\right| \right) \right\}, \quad i < k \leqslant n, \\ 1 \leqslant i \leqslant n. \end{cases}$$

В этих неравенствах S_j и $\bar{T}_j$, обозначают фактические результаты промежуточных вычислений. Эти неравенства можно интерпретировать следующим образом. Допустим, что все элементы матрицы $\mathscr{L}_p$ не превосходят 1 по абсолютной величине. Это предположение выполняется хотя бы приближенно, если строки матрицы $\mathscr{A}_p$ упорядочены таким образом, что не приходится переставлять строки в процессе исключения по Гауссу с выбором главных элементов по столбцам. Если ширина интервалов $L_{ij}^{(m)}$ не слишком велика, то $\left|L_{ij}^{(m)}\right|$ не намного больше 1, а в силу того, что в (17) берутся пересечения, это верно и для $\left|L_{ij}^{(m+1)}\right|$. При тех же предположениях те же рассуждения показывают, что $\left|\bar{U}_{jk}^{(m+1)}\right| - \left|u_{jk}^{(m)}\right|$ и $\left|U_{ji}^{(m+1)}\right| - \left|u_{ji}^{(m)}\right|$ малы. Поэтому делаем вывод, что при малой ширине элементов на m-м шаге разность между $d\left(\bar{U}_{ik}^{(m+1)}\right)$ и $d\left(U_{ik}^{(m+1)}\right)$ существенно зависит от $\left|u_{jk}^{(m)}\right|$ и от величины промежуточных результатов. То же верно и для разности

между $d\left(\overline{L}_{kl}^{(m+1)}\right)$ и $d\left(L_{kl}^{(m+1)}\right)$, если добавить еще, что малые величины $\left|u_{ii}^{(m)}\right|$ могут ухудшить эту разность.

Так как в силу теоремы 6 первые члены приведенных выше неравенств приближенно равны квадратам таких же членов на предыдущем шаге, мы получим небольшую ширину, если выполнены следующие условия:

(a) Элементы матрицы $\mathcal{U}_{\mathrm{p}}^{*}$ по абсолютной величине не намного больше единицы.

(b) Диагональные элементы матрицы $\mathcal{U}_{\mathrm{p}}^{*}$ по абсолютной величине не намного меньше единицы.

(c) Элементы матрицы $\mathcal{L}_{\mathrm{p}}^{*}$ по абсолютной величине не больше единицы.

(d) Вычисленные промежуточные результаты в (17) не слишком велики по абсолютной величине.

Модуль 10
Методы Ньютоновского типа

Микромодуль 37
Методы Ньютоновского типа для системы нелинейных уравнений

Рассмотрим методы итерационной локализации решений для систем нелинейных уравнений. Дано множество функций $f_i(x_{\mathrm{p}}),\ 1 \leqslant i \leqslant n$ векторной переменной $x_{\mathrm{p}} = (x_1, \ldots, x_n)^{\mathrm{T}}$, которые мы объединим в векторную функцию $f_{\mathrm{p}}(x_{\mathrm{p}}) = (f_1(x_{\mathrm{p}}), \ldots$ $\ldots, f_n(x_{\mathrm{p}}))^{\mathrm{T}}$. Допустим, что производная Фреше $f'_{\mathrm{p}}(x_{\mathrm{p}})$ функции $f_{\mathrm{p}}(x_{\mathrm{p}})$ существует на множестве $\mathfrak{B} \subseteq V_n(\mathbb{R})$ и что $x^{(0)} = (X_1^{(0)}, \ldots, X_n^{(0)})^{\mathrm{T}} \subseteq \mathfrak{B}$. Мы предположим также,что для производной Фреше имеется интервальная оценка на $x^{(0)}$. Пусть теперь дан вектор

$$y_{\mathrm{p}} = (y_1, \ldots, y_n)^{\mathrm{t}} \in x^{(0)},$$

удовлетворяющий уравнению

$$f_{\mathrm{p}}(y_{\mathrm{p}}) = o_{\mathrm{p}}. \qquad (1)$$

При сделанных предположениях мы можем линейно аппроксимировать $f_i(x_p)$ в точке $y_p \in x^{(0)}$ с помощью формулы Тейлора

$$f_i(x_p) = f_i(y_p) + \sum_{r=1}^{n} \frac{\partial}{\partial x_r} f_i(y_p + \theta_i(x_p - y_p))(x_r - y_r), \quad (2)$$

$$0 < \theta_i < 1, \quad 1 \leqslant i \leqslant n, \quad x_p \in x^{(0)}.$$

Из (1) следует, что вектор y_p удовлетворяет уравнению

$$f_p(x_p) = \mathcal{J}_p(x_p)(x_p - y_p),$$

где матрица $\mathcal{J}_p(x_p)$ определяется равенством

$$\mathcal{J}_p(x_p) := \left(\frac{\partial}{\partial x_r} f_i(z_p)\Big|_{z_p = y_p + \theta_i(x_p - y_p)}\right). \quad (3)$$

Вместо $\mathcal{J}_p(x_p)$ будем иногда писать

$$\mathcal{J}_p(x_p, y_p, \theta_1, \ldots, \theta_n),$$

так как эта матрица зависит от всех величин, входящих в ее определение. Так как величины θ_i лежат в открытом интервале (0, 1), имеем

$$y_p + \theta_i(x_p - y_p) \in x^{(0)}.$$

Пусть интервальная матрица $f'_p(x^{(0)})$ обозначает интервальное оценивание производной Фреше на интервале $x^{(0)}$. Тогда среди решений z_p множества линейных уравнений

$$f_p(x_p) = \mathcal{F}_p \cdot (x_p - z_p), \quad \text{где} \quad \mathcal{F}_p \in f'_p(x^{(0)}), \quad (4)$$

имеется и вектор y_p, так как $\mathcal{J}_p(x_p) \in f'_p(x^0)$. Теперь возникла задача — вычислить вектор $x^{(1)}$, локализующий множество решений уравнения (4), а затем использовать этот интервальный вектор как потенциально лучшую локализацию вектора-решения y_p. Перед тем как браться за эту задачу, мы рассмотрим еще одну возможность построения множества уравнений, аналогичного (4). В этом случае применяется линейная аппроксимация значения $f_p(x_p)$ в точке y_p, несколько отличная от (2).

Рассмотрим $f_i(x_1, \ldots, x_n)$ как функцию только от x_1 и разложим ее в точке y_1:

$$f_i(x_1, \ldots, x_n) = f_i(y_1, x_2, \ldots, x_n) + (x_1 - y_1)\frac{\partial}{\partial x_1} f_i(y_1 + \theta_{i1}(x_1 - y_1), x_2, \ldots, x_n),$$

где $\theta_{i1} \in (0, 1)$.

Затем мы разложим $f_i(y_1, x_2, \ldots, x_n)$ как функцию от x_2 в точке y_2 и получим

$$f_i(y_1, x_2, \ldots, x_n) = f_i(y_1, y_2, x_3, \ldots, x_n) + (x_2 - y_2)\frac{\partial}{\partial x_2} f_i(y_1, y_2 + \theta_{i2}(x_2 - y_2), x_3, \ldots, x_n),$$

где $\theta_{i2} \in (0, 1)$.

Теперь снова разложим $f_i(y_1, y_2, x_3, \ldots, x_n)$ как функцию от x_3 в точке y_3 и т. д. и, наконец, разложим $f_i(y_1, y_2, \ldots, y_{n-1}, x_n)$ в точке y_n. Собирая эти разложения, мы получим

$$f_i(x_1, \ldots, x_n) = f_i(y_1, \ldots, y_n) + \sum_{j=1}^{n} (x_j - y_j)\frac{\partial}{\partial x_j} f_i(y_1, \ldots, y_{j-1}, y_j + \theta_{ij}(x_j - y_j), x_{j+1}, \ldots, x_n),$$

где $\theta_{ij} \in (0, 1)$, $1 \leqslant i \leqslant n$.

Таким образом, мы имеем соотношение

$$f_p(x_p) = \tilde{J}_p(x_p)(x_p - y_p),$$

где матрица $\tilde{J}_p(x_p)$ определена равенствами

$$\tilde{J}_p(x_p) = \left(\frac{\partial}{\partial x_j} f_i(y_1, \ldots, y_{j-1}, z_j, x_{j+1}, \ldots, x_n)\Big|_{z_j = y_j + \theta_{ij}(x_j - y_j)}\right). \tag{5}$$

Мы можем взять в качестве y_p решение множества линейных уравнений

$$f_p(x_p) = \tilde{F}_p(x_p - z_p), \quad \text{где} \quad \tilde{F}_p \in \tilde{J}_p(x^{(0)}), \tag{6}$$

так как $y_j \in X^{(0)}$ и $y_j + \theta_{ij}(x_j - y_j) \in X_j^{(0)}$ для $1 \leqslant j \leqslant n$. Упомянутая здесь интервальная матрица $\tilde{J}_p(x^{(0)})$ определена равенствами

$$\tilde{J}_p(x^{(0)}) = \left(\frac{\partial}{\partial x_j} f_i(X_1^{(0)}, \ldots, X_{j-1}^{(0)}, X_j^{(0)}, x_{j+1}, \ldots, x_n)\right),$$

и легко видеть, что для интервальной матрицы имеем

$$\tilde{J}_p(x^{(0)}) \subseteq f'_p(x^{(0)}).$$

Поэтому множество (6) содержит не больше систем линейных уравнений, чем множество из (4). В случае когда $n = 1$ или когда частные производные $\partial f_i/\partial x_i$ зависят только от x_j, интервальные матрицы $f'_n(x^{(0)})$ и $\mathcal{F}_n(x^{(0)})$ совпадают.

Вернемся теперь к первоначальной задаче — как улучшить локализующий вектор $x^{(0)}$ для вектора решений y_p. В соответствии с только что проведенными рассуждениями мы вычислим интервальный вектор, содержащий множество решений систем (4) (соответственно (6)), а потому и вектор y_p. Рассмотрим сначала (4). Предположим, что y_p — простой корень функции $f_p(x_p)$ в $x^{(0)}$ и что все матрицы $\mathcal{F}_p \in f'_p(x^{(0)})$ невырождены. Пусть $\mathcal{V}$ — интервальная матрица, содержащая все обращения $\mathcal{F}_p^{-}$ матриц $\mathcal{F}_p \in f'_p(x^{(0)})$. Мы уже продемонстрировали ранее практические методы вычисления такой интервальной матрицы $\mathcal{V}$. Вычислим теперь с помощью $\mathcal{V}$ интервальный вектор

$$y^{(1)} = m(x^{(0)}) - \mathcal{V} \cdot f_p(m(x^{(0)})),$$

где $m(x^{(0)})$ определяется согласно (1 из микромодуля 36). Таким образом, получим локализацию для множества решений уравнений (4). Здесь мы использовали соотношение $\mathcal{J}_p(m(x^{(0)})) \in f'_p(x^{(0)})$. Аналогичным образом можно использовать уравнение (6). Ввиду $y_p \in x^{(0)}$ мыможем построить интервальный вектор

$$x^{(1)} = \{m(x^{(0)}) - \mathcal{V} \cdot f_p(m(x^{(0)}))\} \cap x^{(0)} \subseteq x^{(0)}$$

и получить новую локализацию для y_p. Повторение этого вычисления приводит к итерационному методу

$$x^{(k+1)} = \{m(x^{(k)} - \mathcal{V} \cdot f_p(m(x^{(k)}))\} \cap x^{(k)}, \qquad k \geqslant 0, \qquad (7)$$

где $m(x^{(k)})$ —срединное отображение из (1 микромодуля 36).

Эта итерация порождает монотонную последовательность

$$x^{(0)} \supseteq x^{(1)} \supseteq x^{(2)} \supseteq \ldots$$

интервальных векторов, для которой докажем следующее утверждение.

Теорема 1. *Пусть $x^{(0)}$ —интервальный вектор и $y_p \in x^{(0)}$ —корень функции $f_p(x_p)$. Пусть $\mathcal{V}$ — интервальная матрица, содержащая*

обратные матрицы $\mathscr{J}_p(x_p)^{-1}$ *для* $x_p \in x^{(0)}$. *Тогда последовательность* $\{x^{(k)}\}_{k=0}^{\infty}$ *интервальных векторов, вычисленная согласно* (7), *удовлетворяет следующим условиям.*

Каждый интервальный вектор $x^{(k)}$, $k \geqslant 0$, *содержит корень* y_p. (8)

Если все матрицы $\mathscr{Y}_p \in \mathscr{Y}$ *невырождены, то* $\lim\limits_{k\to\infty} x^{(k)} = y_p$. (9)

Доказательство. Рассмотрим (8). По условию теоремы мы имеем $y_p \in x^{(0)}$, а ввиду $m(x^{(0)}) \in x^{(0)}$ также и

$$\mathscr{J}_p(m(x^{(0)}))^{-1} \in \mathscr{Y}.$$

Поэтому из (10' микромодуля 29) следует, что

$$y_p = m(x^{(0)}) - \mathscr{J}_p(m(x^{(0)}))^{-1} f_p(m(x^{(0)})) \in m(x^{(0)}) - \mathscr{Y} \cdot f_p(m(x^{(0)})),$$

а потому

$$y_p \in \{m(x^{(0)}) - \mathscr{Y} \cdot f_p(m(x^{(0)}))\} \cap x^{(0)} = x^{(1)}.$$

Это рассуждение позволяет доказать (8) методом математической индукции.

(9): В силу следствия 8 из микромодуля 29 монотонная локализующая последовательность $x^{(0)} \supseteq x^{(1)} \supseteq x^{(2)} \ldots$ сходится к некоторому интервальному вектору n. Мы покажем, что $d(x) = o_p$, а вместе с (8) это даст $\lim\limits_{k\to\infty} x^{(k)} = y_p$. Допустим для приведения к противоречию, что $d(x) \neq o_p$. Из непрерывности отображения (7) следует, что x удовлетворяет уравнению

$$x = \{m(x) - \mathscr{Y} \cdot f_p(m(x))\} \cap x.$$

Рассмотрев $m(x) \in x$, мы видим, что $m(x) \neq y_p$, так как $d(x) \neq o$. С другой стороны, мы имеем

$$m(x) \in m(x) - \mathscr{Y} \cdot f_p(m(x)).$$

Кроме того, из (1 микромодуля 29) следует соотношение

$$\mathscr{A}x_p = \{\mathscr{A}_p x_p \mid \mathscr{A}_p \in \mathscr{A}\}$$

для интервальной матрицы $\mathscr{A}$ и вещественного вектора x_p. Отсюда получается представление

$$m(x) = m(x) - \mathscr{Y}_p \cdot f_p(m(x))$$

для некоторой матрицы $\mathscr{Y}_p \in \mathscr{Y}$. Поэтому имеем

$$o_p = \mathscr{Y}_p \cdot f_p(m(x)),$$

что дает противоречие, так как $m(x) \neq y_p$ и матрица $\mathcal{Y}_p$ невырождена.

Достаточное условие сходимости (9) для итерационного метода (7) использовало более слабое условие

$$\{\mathcal{J}_p(x_p)^{-1} \mid x_p \in x^{(0)}\} \subseteq \mathcal{Y}$$

на $\mathcal{Y}$, чем условие, использованное при выводе формул (7). Первоначально мы потребовали, чтобы выполнялось условие

$$\{\mathcal{F}_p^{-1} \mid \mathcal{F}_p \in f'_p(x^{(0)})\} \subseteq \mathcal{Y},$$

которое легче поддается проверке.

Доказательство соотношения (9) можно модифицировать так, чтобы показать, что из $m(x^{(k)}) \neq y_p$ следует

$$m(x^{(k)}) \notin x^{(k+1)}, \quad k \geqslant 0.$$

Это означает, что в методе (7) по крайней мере одна компонента ширины $d(x^{(k)})$ уменьшается на $(k+1)$-м шаге более чем вдвое. Метод (7) можно улучшить, определяя на каждом шаге новую матрицу $\mathcal{Y}^{(k)}$. При выводе формул (7) было видно, что для соотношения (8) нужно лишь

$$\{\mathcal{J}_p(x_p)^{-1} \mid x_p \in x^{(k)}\} \subseteq \mathcal{Y}^{(k)}$$

и

$$\{\mathcal{F}_p^{-1} \mid \mathcal{F}_p \in f'_p(x^{(k)})\} \subseteq \mathcal{Y}^{(k)}.$$

Иными словами, на $(k+1)$-м шаге находим локализующее множество для множества уравнений

$$f_p(m(x^{(k)})) = \mathcal{F}_p \cdot (m(x^{(k)}) - x_p), \quad \text{где} \quad \mathcal{F}_p \in f'_p(x^{(k)}).$$

Теперь покажем, как определить подходящую последовательность $\{\mathcal{Y}^{(k)}\}_{k=0}^{\infty}$. Сначала найдем локализующее множество для

$$\{\mathcal{F}_p^{-1} \mid \mathcal{F}_p \in f'_p(x^{(k)})\}$$

с помощью одного из итеративных методов, описанных ранее. При этом используем сокращение $\mathcal{F}^{(k)}$ для интервальной матрицы $f'_p(x^{(k)})$ и сокращение для $\tilde{\mathcal{F}}^{(k)}$ интервальной матрицы $\tilde{\mathcal{J}}_p(x^{(k)})$.

Начинаем с интервальной матрицы $\mathcal{W}^{(0)}$, которая была найдена, например, с помощью обычных оценок по норме и удовлетворяет условию

$$\{\mathcal{F}_p^{-1} \mid \mathcal{F}_p \in \mathcal{F}^{(k)}\} \subseteq \mathcal{W}^{(0)}.$$

Затем исполняем итерационную процедуру

$$\mathcal{W}^{(\nu+1)} = \left\{\left(\mathcal{I}_p - (m(\mathcal{F}^{(k)}))^{-1}\mathcal{F}^{(k)}\right)\mathcal{W}^{(\nu)} + (m(\mathcal{F}^{(k)}))^{-1}\right\} \cap \mathcal{W}^{(\nu)}, \quad \nu \geqslant 0. \tag{10}$$

Последовательность

$$\mathcal{W}^{(0)} \supseteq \mathcal{W}^{(1)} \supseteq \mathcal{W}^{(2)} \supseteq \ldots$$

всегда будет сходиться к некоторой однозначно определенной интервальной матрице $\mathcal{W} \subseteq \mathcal{W}^{(0)}$, если спектральный радиус удовлетворяет условию

$$\rho\left(\left|\mathcal{I}_p - (m(\mathcal{F}^{(k)}))^{-1}\mathcal{F}^{(k)}\right|\right) < 1$$

Матрица $\mathcal{I}_p - (m(\mathcal{F}^{(k)}))^{-1}\mathcal{F}^{(k)}$ в (10) симметрична относительно центра $\mathcal{O}_p$. Это означает, что процедуру (10) можно разбить на две одинаковые вещественные итерационные процедуры соответственно для матриц верхних и нижних границ для $\mathcal{W}$. Поэтому для вычисления $\mathcal{W}$ нужно решить всего одну систему вещественных уравнений. Положим теперь

$$\mathcal{Y}^{(k)} := \mathcal{W} \supseteq \left\{\mathcal{F}_p^{-1} \mid \mathcal{F}_p \in \mathcal{F}^{(k)}\right\}. \tag{11}$$

С помощьюэтой матрицы $\mathcal{Y}^{(k)}$ мы делаем шаг итерации

$$x^{(k+1)} = \left\{m(x^{(k)}) - \mathcal{Y}^{(k)}f_p(m(x^{(k)}))\right\} \cap x^{(k)}.$$

Вычисляется новая матрица $\mathcal{F}^{(k+1)} = f_p'(x^{(k+1)})$ и затем запускается процедура (10). Из $x^{(k+1)} \subseteq x^{(k)}$ легко следует, что $\mathcal{F}^{(k+1)} \subseteq \mathcal{F}^{(k)}$ и $\left\{\mathcal{F}_p^{-1} \mid \mathcal{F}_p \in \mathcal{F}^{(k+1)}\right\} \subseteq \mathcal{Y}^{(k)}$.

Поэтому матрицу $\mathcal{Y}^{(k)}$ можно использовать в качестве нового начального элемента для (10). Мы получаем таким образом следующую итерационную процедуру, в которой $\mathcal{Y}^{(0)}$ содержит множество $\left\{\mathcal{F}_p^{-1} \mid \mathcal{F}_p \in \mathcal{F}^{(0)}\right\}$:

$$x^{(k+1)} = \left\{m(x^{(k)}) - \mathcal{Y}^{(k)}f_p(m(x^{(k)}))\right\} \cap x^{(k)}, \quad k \geqslant 0, \tag{12}$$

где $\mathcal{Y}^{(k)}$—матрица, обладающая свойством (11) и вычисленная с помощью процедуры (10); причем $\mathcal{Y}^{(0)}$ используется в качестве начальной матрицы в (10) при k= 1.

Докажем следующее утверждение о методе (12).

Теорема 2. *Пусть $x^{(0)}$ — интервальный вектор, а $y_p \in x^{(0)}$ — корень функции $f_p(x_p)$. Пусть $\mathcal{Y}^{(0)}$ — интервальная матрица, содер-*

жащая все обращения $\mathscr{F}_p^{-1}$ *матриц* $\mathscr{F}_p \in \mathscr{F}^{(0)} = \mathscr{f}'_p(x^{(0)})$. *Допустим еще, что производная Фреше* $\mathscr{f}'_p(x_p)$ *для значения аргумента* $x^{(0)}$ *удовлетворяет условиям теоремы* 5 п. 7.3 *для каждого элемента. Тогда последовательность интервальных векторов, вычисленная по формулам* (12), *обладает следующими свойствами.*

(8) *Каждое из приближений* $x^{(k)}$, $k \geqslant 0$, *содержит корень* y_p;

(9) *если любая матрица* $\mathscr{Y}_p \in \mathscr{Y}^{(0)}$ *неособенная, то мы имеем*

$$\lim_{k\to\infty} x^{(k)} = y_p;$$

если последовательность $\{x^{(k)}\}_{k=0}^{\infty}$ *сходится к* y_p, *то для некоторой нормы* $\|\cdot\|$ *верно*

$$\|d(x^{(k+1)})\| \leqslant \gamma \|d(x^{(k)})\|^2, \quad \gamma \geqslant 0, \tag{13}$$

т. е. R-порядок метода (12) *удовлетворяет неравенству*

$$O_R((12), y_p) \geqslant 2$$

(см. приложение А, *теорема* 2).

Доказательство. Проверка соотношений (8) и (9) проводится совершенно так же, как в доказательстве теоремы (1). Нужно лишь рассмотреть

$$\{\mathscr{F}_p^{-1} \mid \mathscr{F}_p \in \mathscr{F}^{(k)}\} \subseteq \mathscr{Y}^{(k)} \subseteq \mathscr{Y}^{(0)}, \quad k \geqslant 0.$$

(13): Из того что $x^{(k+1)} \subseteq m(x^{(k)}) - \mathscr{Y}^{(k)}\mathscr{f}_p(m(x^{(k)}))$, следует, что

$$\begin{aligned} d(x^{(k+1)}) &\leqslant d(m(x^{(k)}) - \mathscr{Y}^{(k)}\mathscr{f}_p(m(x^{(k)}))) \\ &= d(\mathscr{Y}^{(k)} \mid \mathscr{f}_p(m(x^{(k)}))) = d(\mathscr{Y}^{(k)} \mid \mathscr{f}_p(m(x^{(k)})) - \mathscr{f}_p(y_p)) \\ &= d(\mathscr{Y}^{(k)} \mid \mathscr{J}_p(m(x^{(k)}))(m(x^{(k)}) - y_p)) \\ &\leqslant d(\mathscr{Y}^{(k)} \mid \mathscr{J}_p(m(x^{(k)}))) \left| \frac{1}{2} d(x^{(k)}) \leqslant d(\mathscr{Y}^{(k)} \mid \mathscr{F}^{(0)}) \right| \frac{1}{2} d(x^{(k)}). \end{aligned}$$

Теперь, применяя монотонную векторную норму и совместную с ней монотонную матричную норму, получим

$$\|d(x^{(k+1)})\| \leqslant \|d(\mathscr{Y}^{(k)})\| \frac{1}{2} \|\mathscr{F}^{(0)}\| \|d(x^{(k)})\| = c \|d(\mathscr{Y}^{(k)})\| \|d(x^{(k)})\|.$$

Наконец, нужно оценить $\|d(\mathscr{Y}^{(k)})\|$. Так как $\mathscr{Y}^{(k)}$ удовлетворяет равенству

$$\mathscr{Y}^{(k)} = \{(\mathscr{J}_p - (m(\mathscr{F}^{(k)}))^{-1}\mathscr{F}^{(k)})\mathscr{Y}^{(k)} + (m(\mathscr{F}^{(k)}))^{-1}\} \cap \mathscr{Y}^{(k)},$$

получаем из (21 микромодуля 29) и (29 микромодуля 29), что

$$d(\mathcal{Y}^{(k)}) \leqslant d\left(\left(\mathcal{I}_p - (m(\mathcal{F}^{(k)}))^{-1}\mathcal{F}^{(k)}\right)\mathcal{Y}^{(k)} + (m(\mathcal{F}^{(k)}))^{-1}\right)$$
$$= d\left(\left(\mathcal{I}_p - (m(\mathcal{F}^{(k)}))^{-1}\mathcal{F}^{(k)}\right)\mathcal{Y}^{(k)}\right)$$
$$= d\left(\left((m(\mathcal{F}^{(k)}))^{-1}(m(\mathcal{F}^{(k)}) - \mathcal{F}^{(k)})\right)\mathcal{Y}^{(k)}\right)$$
$$= \left|m(\mathcal{F}^{(k)})^{-1}\right| d(\mathcal{F}^{(k)})\,|\mathcal{Y}^{(k)}|.$$

Применяя упомянутую выше монотонную матричную норму, а также соотношения

$$|\mathcal{Y}^{(k)}| \leqslant |\mathcal{Y}^{(0)}| \quad \text{и} \quad \left|(m(\mathcal{F}^{(k)}))^{-1}\right| \leqslant |\mathcal{Y}^{(0)}|,$$

получаем, что

$$\|d(\mathcal{Y}^{(k)})\| \leqslant \tilde{\gamma}\,\|d(\mathcal{F}^{(k)})\|,$$

где $\tilde{\gamma}$— константа, не зависящая от k.

Отдельные элементы матрицы $d(\mathcal{F}^{(k)})$ имеют вид

$$d\left(\frac{\partial}{\partial x_j} f_i(X_1^{(k)}, \ldots, X_n^{(k)})\right),$$

и с помощью соотношения (5' п.7.3) для них можно получить оценку

$$d\left(\frac{\partial}{\partial x_j} f_i(X_1^{(k)}, \ldots, X_n^{(k)})\right) \leqslant \sum_{r=1}^{n} \gamma_{ij/r}^{(k)} d(X_r^{(k)}) \leqslant \sum_{r=1}^{n} \gamma_{ij/r}^{(0)} d(X_r^{(k)}).$$

Следовательно, мы имеем

$$d(\mathcal{F}^{(k)}) \leqslant \mathcal{C}_p\, d(x^{(k)}),$$

где $\mathcal{C}_p = (\gamma_{ij/l}^{(0)})$— билинейный оператор из $V_n(\mathbb{R}) \times V_n(\mathbb{R})$ в $V_n(\mathbb{R})$. Если определить норму для $\mathcal{C}_p$ обычным равенством

$$\|\mathcal{C}_p\| = \sup_{\|x_p\|=1} \sup_{\|y_p\|=1} \|\mathcal{C}_p x_p y_p\|,$$

то получается, что

$$\|d(\mathcal{F}^{(k)})\| \leqslant \|\mathcal{C}_p\|\,\|d(x^{(k)})\|,$$

т. е.

$$\|d(\mathcal{Y}^{(k)})\| \leqslant \bar{c}\,\|d(x^{(k)})\|, \quad \text{где} \quad \bar{c} = \|\mathcal{C}_p\|\,\tilde{\gamma}.$$

Подставляя это в полученную выше оценку для $d(x^{(k+1)})$, получаем

$$\|d(x^{(k+1)})\| \leqslant \bar{c}\,\|d(x^{(k)})\|^2, \quad \text{где} \quad \tilde{c} = c\bar{c}$$

для монотонной векторной нормы. Применяя понятие эквивалентности норм так же, как при доказательстве соотношения (8 из микромодуля 36) в теореме 1 из микромодуля 36, мы убеждаемся, что

могли использовать произвольную норму, и имеем соотношение (13) с константой γ, не зависящей от k.

Теперь мы хотим рассмотреть другой способ, позволяющий вычислять решение y_p уравнения $f_p(x_p) = o_p$ путем последовательных локализаций. Этот метод подходит для определенных классов нелинейных систем уравнений.

Исходим из системы (4), полагая $x_p := m(x^{(0)})$. Приводимые ниже соображения можно реализовать и исходя из (6). Если, например, предположить, что $\mathcal{F}^{(0)} = f'_p(x^{(0)})$ удовлетворяет условиям теоремы 3 из микромодуля 33, то к интервальной матрице $\mathcal{F}^{(0)}$ и вектору $f_p(m(x^{(0)}))$ можно применить метод Гаусса. Если $w^{(0)}$ — интервальный вектор, полученный в результате, то вектор $m(x^{(0)}) - w^{(0)}$, а значит, и вектор

$$\tilde{x}^{(0)} = \{m(x^{(0)}) - w^{(0)}\} \cap x^{(0)},$$

содержат решение y_p системы нелинейных уравнений.

Напишем уравнение

$$f_p(x_p) = \mathcal{J}_p(x_p)(x_p - y_p)$$

по образцу (2). Представим матрицу $\mathcal{J}_p(x_p)$, входящую в это уравнение, в виде

$$\mathcal{J}_p(x_p) = \mathcal{D}_p(x_p) - \mathcal{B}_p(x_p),$$

где $\mathcal{D}_p(x_p)$ — содержит диагональ матрицы $\mathcal{J}_p(x_p)$. Полагая $x_p = m(x^{(0)})$, получаем из монотонности включения, что

$$\begin{aligned} y_p &= m(x^{(0)}) - \mathcal{D}_p(m(x^{(0)}))^{-1}\{\mathcal{B}_p(m(x^{(0)}))(m(x^{(0)}) - y_p) + f_p(m(x^{(0)}))\} \\ &\in \{m(x^{(0)}) - \mathcal{D}_p(x^{(0)})^{-1}\{\mathcal{B}_p(x^{(0)})(m(x^{(0)}) - \tilde{x}^{(0)}) \\ &\quad + f_p(m(x^{(0)}))\}\} \cap \tilde{x}^{(0)} =: x^{(1)}. \end{aligned}$$

Здесь мы использовали такое же представление матрицы $\mathcal{F}^{(0)} = f'_p(x^{(0)})$ в виде

$$\mathcal{F}^{(0)} = \mathcal{D}_p(x^{(0)}) - \mathcal{B}_p(x^{(0)}),$$

как и выше, для $\mathcal{J}_p(x_p)$. Используя равенство

$$\mathcal{D}_p(x^{(0)}) = \operatorname{diag}(D_i(x^{(0)})),$$

полагаем $\mathcal{D}_p(x^{(0)})^{-1} = \operatorname{diag}(1/D_i(x^{(0)}))$. Если матрица $\mathcal{F}^{(1)} = f'_n(x^{(1)})$ снова удовлетворяет условиям теоремы 3 из микромодуля 33, то мы можем, исходя из $x^{(1)}$, вычислить тем же

способом, что и раньше, новую локализацию $x^{(2)}$ для y_p, и т. д. Повторение этого шага приводит к следующему итерационному методу.

(14)

(a) $\mathscr{F}^{(k)} = f'_p(x^{(k)}) = \mathscr{D}_p(x^{(k)}) - \mathscr{B}_p(x^{(k)})$.

(b) Применение метода Гаусса к $(\mathscr{F}^{(k)}, f_p(m(x^{(k)})))$ дает $w^{(k)}$; здесь $m(x^{(k)})$ вычисляется согласно (18.1).

(c) $\tilde{x}^{(k)} = \{m(x^{(k)}) - w^{(k)}\} \cap x^{(k)}$.

(d) $x^{(k+1)} = \{m(x^{(k)}) - \mathscr{D}_p(x^{(k)})^{-1}\{\mathscr{B}_p(x^{(k)})(m(x^{(k)}) - \tilde{x}^{(k)}) + f_p(m(x^{(k)}))\}\} \cap \tilde{x}^{(k)}, \quad k \geqslant 0.$

Шаг (d) этого алгоритма требует $\sim n^2$ операций (интервальных умножений и делений) для неразреженной матрицы. Из того что шаг (b) требует $\sim n^3/3$ операций, следует, что объем вычислений на шаге (d) несуществен при больших n.

Докажем несколько лемм, чтобы исследовать условия, при которых для метода (14) выполнено

$$\lim_{k\to\infty} x^{(k)} = y_p,$$

и оценить R-порядок сходимости.

Лемма 3. *Пусть $\mathscr{A} = (A_{ij})$— вещественная интервальная матрица, удовлетворяющая условиям теоремы* 3 из микромодуля 33. *Если $\mathscr{A}$ представлена в виде $\mathscr{A} = \mathscr{D} - \mathscr{B}$, где $\mathscr{D}$ содержит диагональ матрицы $\mathscr{A}$, то матрица $\mathscr{D} = \operatorname{diag}(1/A_{ii})$ удовлетворяет неравенству*

$$\rho(|\mathscr{D}||\mathscr{B}|) < 1.$$

Доказательство Мы снова представим элементы матрицы $\mathscr{A}$ через середину и полуширину в виде $A_{ij} = \langle a_{ij}, r_{ij}\rangle$, $1 \leqslant i, j \leqslant n$. Учтем теперь определение матрицы $\mathscr{B}_p$ в теореме 3 из микромодуля 33 и равенство

$$|1/A_{ii}| = 1/(a_{ii} - r_{ii}).$$

Сразу видно, что вещественная матрица $|\mathscr{D}||\mathscr{B}|$ является полношаговой матрицей для полношагового метода, соответствующего матрице $\mathscr{B}_p$. По известной теореме получаем, что

$$\rho(|\mathscr{D}||\mathscr{B}|) < 1.$$

Пусть теперь $\mathscr{A} = (A_{ij})$— вещественная интервальная матрица, x— вещественный интервальный вектор и выполнено

$$d(A_{ij}) \leqslant \alpha_{ij}\|d(x)\|, \quad \alpha_{ij} \geqslant 0, \quad 1 \leqslant i, \quad j \leqslant n. \qquad (15)$$

Если $0 \notin A_{ij}$, то из (15) следует соотношение

$$d(1/A_{ij}) \leqslant \hat{\alpha}_{ij} \| d(x) \|, \quad \hat{\alpha}_{ij} \geqslant 0, \quad 1 \leqslant i, \quad j \leqslant n. \tag{16}$$

Пусть $\mathcal{K} = (H_i)$ — вещественный интервальный вектор, удовлетворяющий условиям

$$|H_i| \leqslant \beta_i \| d(x) \|, \quad \beta_i \geqslant 0, \quad 1 \leqslant i \leqslant n, \tag{17}$$

$$d(H_i) \leqslant \gamma_i \| d(x) \|, \quad \gamma_i \geqslant 0, \quad 1 \leqslant i \leqslant n. \tag{18}$$

Лемма 4. *Пусть для* $\mathcal{A} = (A_{ij})$ *и* $\mathcal{K} = (H_i)$ *выполнены неравенства* (15)—(18). *Если для интервальной матрицы* $\mathcal{A}$ *и интервального вектора* $\mathcal{K}$ *можно выполнить метод Гаусса, то результирующая верхняя треугольная матрица* $\tilde{\mathcal{A}} = (\tilde{A}_{ij})$ *и соответствующий вектор* $\tilde{\mathcal{K}} = (\tilde{H}_i)$ *удовлетворяют условиям* (15)—(18) *с подходящими константами* $\tilde{\alpha}_{ij} \geqslant 0$, $\tilde{\beta}_i \geqslant 0$, $\tilde{\gamma}_i \geqslant 0$.

Доказательство. Применим математическую индукцию по шагам метода Гаусса (всего их n—1). Для каждого шага покажем, что если рассматриваемые условия выполнены для пары вектор — матрица на предыдущем шаге, то они выполнены и для пары, полученной на текущем шаге. Мы проведем доказательство лишь для первого шага. На этом шаге по данной матрице $\mathcal{A}$ и вектору $\mathcal{K}$ вычисляются новая матрица $\mathcal{A}'$ и новый вектор $\mathcal{K}'$ по формулам

$$\begin{aligned}
&A'_{1j} = A_{1j}, && 1 \leqslant j \leqslant n, \quad H'_1 = H_1, \\
&A'_{i1} = 0, && 2 \leqslant i \leqslant n, \\
&A'_{ij} = A_{ij} - (A_{i1}/A_{11})\, A_{1j}, && 2 \leqslant i, \ j \leqslant n, \\
&H'_i = H_i - (A_{i1}/A_{11})\, H_1, && 2 \leqslant i \leqslant n.
\end{aligned}$$

Тогда для $2 \leqslant i, \ j \leqslant n$ получаем с помощью (15)—(18) и (12 п. 7.2), что

$$\begin{aligned}
d(A'_{ij}) &= d(A_{ij}) + d\left((A_{i1}/A_{11})\, A_{1j}\right) \\
&\leqslant \alpha_{ij} \| d(x) \| + \left|\frac{A_{i1}}{A_{11}}\right| d(A_{1j}) + \left(d(A_{i1}) \left|\frac{1}{A_{11}}\right| + d\left(\frac{1}{A_{11}}\right) |A_{i1}|\right) |A_{1j}| \\
&\leqslant \tilde{\alpha}_{ij} \| d(x) \|, \quad \tilde{\alpha}_{ij} \geqslant 0.
\end{aligned}$$

Аналогичным образом получаем для $2 \leqslant i \leqslant n$, что

$$|H'_i| \leqslant |H_i| + |A_{i1}/A_{11}|\, |H_1| \leqslant \tilde{\beta}_i \| d(x) \|, \quad \tilde{\beta}_i \geqslant 0,$$

а также

$$d(H_i') = d(H_i) + |H_1|\left(d(A_{i1})\left|\frac{1}{A_{11}}\right| + d\left(\frac{1}{A_{11}}\right)|A_{i1}|\right)$$
$$+ d(H_1)\left|\frac{A_{i1}}{A_{11}}\right| \leqslant \tilde{\gamma}_i \|d(x)\|^2, \quad \tilde{\gamma}_i \geqslant 0.$$

Для остальных элементов $\mathscr{A}'$ и h' неравенства (15)—(18) выполнены тривиальным образом.

Лемма 5. *Пусть* $\tilde{\mathscr{A}} = (\tilde{A}_{ij})$ *— верхняя треугольная матрица, такая что* $0 \notin \tilde{A}_{ii}$, $1 \leqslant i \leqslant n$, *причем она и вектор* $\tilde{h} = (\tilde{H}_i)$ *удовлетворяют условиям* (15)—(18). *Тогда интервальный вектор* $= (Y_i)$, *вычисленный по формулам*

$$Y_n = \frac{\tilde{H}_n}{\tilde{A}_{nn}}, \quad Y_i = \frac{1}{\tilde{A}_{ii}}\left(\tilde{H}_i - \sum_{j=i+1}^{n} \tilde{A}_{ij} Y_j\right), \quad 1 \leqslant i \leqslant n-1,$$

удовлетворяет неравенству

$$\|d(y)\| \leqslant c\|d(x)\|^2, \quad c \geqslant 0.$$

Доказательство. Из (12 п. 7.2) следует, что

$$d(Y_n) \leqslant d(\tilde{H}_n)\left|\frac{1}{\tilde{A}_{nn}}\right| + d\left(\frac{1}{\tilde{A}_{nn}}\right)|\tilde{H}_n| \leqslant \delta_n \|d(x)\|^2, \quad \delta_n \geqslant 0,$$

и что

$$|Y_n| = |1/\tilde{A}_{nn}||\tilde{H}_n| \leqslant \varkappa_n \|d(x)\|.$$

Если теперь верно, что

$$d(Y_i) \leqslant \delta_i \|d(x)\|^2, \quad |Y_i| \leqslant \varkappa_i \|d(x)\|, \quad 1 < i_0 + 1 \leqslant i \leqslant n,$$

то мы получаем

$$d(Y_{i_0}) \leqslant d\left(\frac{1}{\tilde{A}_{i_0 i_0}}\right)\left|\tilde{H}_{i_0} - \sum_{j=i_0+1}^{n} \tilde{A}_{i_0 j} Y_j\right|$$
$$+ \left|\frac{1}{\tilde{A}_{i_0 i_0}}\right| d\left(\tilde{H}_{i_0} - \sum_{j=i_0+1}^{n} \tilde{A}_{i_0 j} Y_j\right) \leqslant \delta_{i_0} \|d(x)\|^2,$$

$$|Y_{i_0}| \leqslant \left|\frac{1}{\tilde{A}_{i_0 i_0}}\right|\left(|\tilde{H}_{i_0}| + \sum_{j=i_0+1}^{n} |\tilde{A}_{i_0 j}||Y_j|\right) \leqslant \varkappa_{i_0} \|d(x)\|.$$

Отсюда следует, что

$$d(Y_i) \leqslant \delta_i \|d(x)\|^2, \quad 1 \leqslant i \leqslant n,$$
$$\max_{1 \leqslant i \leqslant r} \{d(Y_i)\} \leqslant \max_{1 \leqslant i \leqslant n} \{\delta_i\} \|d(x)\|^2.$$

Теперь нужное утверждение получается из теоремы об эквивалентности норм.

Соединяя леммы 4 и 5, имеем следующее утверждение.

Лемма 6. *Пуаь неравенства* (15)—(18) *выполнены для* $\mathcal{A} = (A_{ij})$ *и* $h = (H_i)$. *Если для* $\mathcal{A}$ и h *можно выполнить метод Гаусса, то результирующий интервальный вектор* y *удовлетворяет неравенству*

$$\| d(y) \| \leqslant c \| d(x) \|^2, \quad c \geqslant 0.$$

После этих подготовительных лемм мы докажем следующее утверждение о методе (14).

Теорема 7. *Пусть* $x^{(0)}$ *—интервальный вектор и* $y_p \in x^{(0)}$ *- корень функции* $f_p(x_p)$. *Пусть интервальное вычисление* $f'_p(x)$ *производной Фреше* $f'_p(x_p)$ *удовлетворяет условиям теоремы* 3 из микромодуля 33 *для всех* $x \subseteq x^{(0)}$. *Пусть, кроме того, неравенство* (5' п.7.3) *покомпонентно выполнено для* $f'_p(x^{(0)})$. *Тогда последовательность* $\{x^{(k)}\}_{k=0}^{\infty}$, *вычисленная согласно* (14), *удовлетворяет следующим условиям.*

Каждое приближение $x^{(k)}$, $k \geqslant 0$, *содержит корень* y_p. (18)

Итерационный метод (14) *может быть выполнен для любого* $k \geq 0$, *и имеет место* (19)

$$\lim_{k \to \infty} x^{(k)} = y_p.$$

Имеет место неравенство (20)

$$\| d(x^{(k+1)}) \| \leqslant c \| d(x^{(k)}) \|^2, \quad c \geqslant 0,$$

т. е. R-порядок сходимости метода (14) *удовлетворяет неравенству*

$$O_R((14), y) \geqslant 2$$

(см. приложение А, *теорема* 2).

Доказательство. Проверка соотношения (8) сделана для $k = 1$ при выводе формул (14). В общем случае доказательство можно провести методом математической индукции. (19): Теорема 3 из микромодуля 33 гарантирует возможность выполнения метода Гаусса. В условиях нашей теоремы диагональ матрицы $f'(x^{(k)})$ не может содержать нулей, так что п. (d) в (14) всегда может быть выполнен. Интервальные векторы, вычисленные согласно (14), удовлетворяют соотношениям

$$x^{(0)} \supseteq x^{(1)} \supseteq \ldots \supseteq x^{(k)} \supseteq x^{(k+1)} \supseteq \ldots .$$

В силу следствия 8 из микромодуля 29 эта последовательность сходится к некоторому предельному значению x. Отсюда следует, что

$$\lim_{k\to\infty} m(x^{(k)}) = m(x),\ \lim_{k\to\infty} f_p(m(x^{(k)})) = f_p(m(x)),\ \lim_{k\to\infty} f'_p(x^{(k)}) = f'_p(x).$$

Так как $w^{(k)}$ непрерывно зависит от $f_p(m(x^{(k)}))$ и $f'_p(x^{(k)})$, отсюда следует, что $\lim_{k\to\infty} w^{(k)} = w$, а также

$$\lim_{k\to\infty} \tilde{x}^{(k)} = \tilde{x} = \{m(x) - w\} \cap x,$$

$$\lim_{k\to\infty} x^{(k)} = x = \{m(x) - \mathscr{D}_p(x)^{-1}\{\mathscr{B}_p(x)(m(x) - \tilde{x}) + f_p(m(x))\}\} \cap \tilde{x}.$$

Из этих равенств следует, что

$$\tilde{x} \subseteq x,$$

$$x \subseteq m(x) - \mathscr{D}_p(x)^{-1}\{\mathscr{B}_p(x)(m(x) - \tilde{x}) + f_p(m(x))\},$$

т. е.

$$x \subseteq m(x) - \mathscr{D}_p(x)^{-1}\{\mathscr{B}_p(x)(m(x) - x) + f_p(m(x))\}. \quad (21)$$

В частности, тогда имеет место

$$m(x) \in m(x) - \mathscr{D}_p(x)^{-1}\{\mathscr{B}_p(x)(m(x) - x) + f_p(m(x))\},$$

Откуда

$$o_p \in \mathscr{D}_p(x)^{-1}\{\mathscr{B}_p(x)(m(x) - x) + f_p(m(x))\}.$$

Отсюда следует, что

$$o_p \in \mathscr{B}_p(x)(m(x) - x) + f_p(m(x)).$$

Применяя (19 п. 7.2) к (21) покомпонентно, получаем, что

$$\begin{aligned} d(x) &\leqslant |\mathscr{D}_p(x)^{-1}|\, d(\mathscr{B}_p(x)(m(x) - x) + f_p(m(x))) \\ &= |\mathscr{D}_p(x)^{-1}|\,|\mathscr{B}_p(x)|\, d(x) \\ &\leqslant |\mathscr{D}_p(x^{(0)})^{-1}|\,|\mathscr{B}_p(x^{(0)})|\, d(x). \end{aligned}$$

В силу леммы 3 имеем $\rho(|\mathscr{D}_p(x^{(0)})^{-1}|\,|\mathscr{B}_p(x^{(0)})|) < 1$; значит, $d(x) = o_p$. Ввиду

$$y_p \in x^{(k)},\ k \geqslant 0,$$

отсюда следует с помощью (27 из микромодуля 29), что $\lim_{k\to\infty} x^{(k)} = y_p$, а это и требовалось показать.

(20): Покажем, что неравенства (15)—(18) с заменой ω на $\omega^{(0)}$ выполнены для

$$\mathcal{A} := f'_p(x^{(0)}) \text{ и } k := f_p(m(x^{(0)})) = (f_i(m(x^{(0)}))).$$

По условию теоремы каждый элемент матрицы $f'_p(x^{(0)})$ удовлетворяет неравенству (5' п. 7.3), откуда следует в силу эквивалентности норм, что

$$d(A_{ij}) \leqslant a_{ij} \| d(x^{(0)}) \|, \quad 1 \leqslant i, j \leqslant n,$$

т. е. (15). Из теоремы о среднем значении получаем для некоторых $0 < \theta_i < 1$ соотношение

$$|H_i| = |f_i(m(x^{(0)}))| \leqslant |f'_i(y_p + \theta_i(y_p - m(x^{(0)})))| |m(x^{(0)}) - y_p| \leqslant |f'_i(x^{(0)})| |m(x^{(0)}) - y_p|.$$

Ввиду $m(x^{(0)}),\ y_p \in x^{(0)}$ и теоремы об эквивалентности норм отсюда следует, что

$$|H_i| \leqslant \beta_i \| d(x^{(0)}) \|, \quad 1 \leqslant i \leqslant n,$$

т. е. верно (17). Из $d(f_p(m(x^{(0)}))) = o_p$ тривиально следует, что верно (18). Теперь из леммы 6 следует, что

$$\| d(w^{(0)}) \| \leqslant c \| d(x^{(0)}) \|^2,$$

а п. (c), (d) из (14) дают тогда

$$\| d(x^{(1)}) \| \leqslant c \| d(x^{(0)}) \|^2.$$

Эти рассуждения можно повторить для любого $k \geq 0$, заменяя соответствующие константы на константы первого шага, так как $x^{(k)} \supseteq x^{(k+1)},\ k \geqslant 0$. Тем самым доказано неравенство

$$\| d(x^{(k+1)}) \| \leqslant c \| d(x^{(k)}) \|^2, \quad k \geqslant 0.$$

Теперь наша теорема получается из теоремы 2 приложения А.

Опишем теперь класс задач, для которых выполнены условия теоремы 7. Для этого рассмотрим краевую задачу

$$y'' = f(t, y), \quad y(0) = \bar{a}, \quad y(1) = b, \quad f_y(t, y) \geqslant 0, \quad t \in (0, 1).$$

Применение к этой задаче обычного разностного метода приводит к системе нелинейных уравнений, удовлетворяющих всем условиям теоремы 7. (Далее эта система будет выписана явно.) Мы выбираем $n = 25$ точек в интервале (0, 1) и приводим в табл. 1 результаты последовательных приближений к значению $y\left(\frac{1}{2}\right)$ для задачи

$$f(t, y) = y + \sin(y),\ y(0) = 0,\ y(1) = 1.$$

Таблица 1

k	Нижняя граница	Верхняя граница
0	−0.5	0.5
1	0.3966601342385	0.4046693140542
2	0.3986876258099	0.3986882685536
3	0.3986880255411	0.3986880255471
4	0.3986880255420	0.3986880255466

Интервальный вектор $x^{(0)}$ вычислен по методу из микромодуля 39. В вычислениях были учтены все ошибки округления.

Вернемся теперь к лемме 10 из микромодуля 30 и используем ее для доказательства некоторых теорем существования решений систем нелинейных уравнений.

Теорема 8. *Пусть отображение* $f_p: \mathfrak{B} \subseteq V_n(\mathbb{R}) \to V_n(\mathbb{R})$ *непрерывно дифференцируемо в* $\mathfrak{B}$ *и для некоторого* $\lambda \subseteq \mathfrak{B}$ *существует вычисление производной в интервальной арифметике. Пусть также метод Гаусса применим к* $f'_p(x)$ *и его применение к* $f'_p(x)$ *и правой части* $f_p(y_p)$ *для фиксированного* $y_p \in x$ *дает в результате интервальный вектор* z. *Тогда из*

$$y_p - z \subseteq x$$

следует, что f_p *имеет корень в* x, *а из* $x \cap \{y_p - z\} = \varnothing$ *следует, что* f_p *не имеет корня в* x.

Доказательство. Мы начнем с уравнения (2), записанного в матричном виде

$$f_p(x_p) - f_p(y_p) = J_p(x_p)(x_p - y_p).$$

Из того что для матрицы $f'_p(x)$ можно выполнить метод Гаусса, следует, что все точечные матрицы из $f'_p(x)$, в частности матрица $J_p(\tau_p)$, являются невырожденными. Рассмотрим отображение

$$r_p : x \subset V_n(\mathbb{R}) \to V_n(\mathbb{R}),$$

такое что

$$r_p(x_p) = x_p - J_p(x_p)^{-1} f_p(x_p).$$

Из этой формулы следует, что

$$f_p(x_p) = x_p - \mathcal{J}_p(x_p)^{-1} f_p(y_p) + \mathcal{J}_p(x_p)^{-1}(f_p(y_p) - f_p(x_p))$$
$$= y_p - \mathcal{J}_p(x_p)^{-1} f_p(y_p) \in y_p - z_p.$$

Поэтому из $y_p - z \subseteq x$ следует, что $f_p(x_p) \in x$ для $x_p \in x$, и из леммы 10 микромодуля 30 следует, что f_p имеет корень в x.

Чтобы доказать вторую половину теоремы, мы предположим, что f_p имеет корень x_p^* в x. Тогда, полагая

$$x_p = x_p^*$$

в формуле, определяющей отображение f_p, получаем

$$f_p(x_p^*) = x_p^* \in y_p - z.$$

Так как $x_p^* \in x$, мы имеем $x \cap \{y_p - z\} \neq \varnothing$, что дает противоречие.

Из доказательства предыдущей теоремы видно, что можно избежать применения метода Гаусса, если известна интервальная матрица $\mathcal{Y}$, такая что

$$\mathcal{J}_p(x_p)^{-1} \in \mathcal{Y} \quad \text{для} \quad x_p \in x.$$

Тогда с помощью отображения f_p, введенного в доказательстве предшествующей теоремы, мы получим соотношение

$$f_p(x_p) = x_p - \mathcal{J}_p(x_p)^{-1} f_p(x_p) \in y_p - \mathcal{Y} f_p(y_p).$$

Итак, мы имеем следующее утверждение.

Теорема 9. *Пусть отображение* $f_p: \mathfrak{B} \subseteq V_n(\mathbb{R}) \to V_n(\mathbb{R})$ *непрерывно дифференцируемо в* $\mathfrak{B}$, *причем производная имеет вычисление в интервальной арифметике для некоторого* $x \subseteq \mathfrak{B}$. *Пусть интервальная матрица* $\mathcal{Y}$ *такова, что* $\mathcal{J}_p(x_p)^{-1} \in \mathcal{Y}$ *для*

$$x_p \in x.$$

Если теперь

$$y_p - \mathcal{Y} f_p(y_p) \subseteq x$$

для фиксированного $y_p \in x$, *то в* f_p *имеется нуль функции* x.

Если же

$$\{y_p - \mathcal{Y} f_p(y_p)\} \cap x = \varnothing,$$

то в x *нет нулей функции* f_p.

В обшем случае мы будем вычислять матрицу $\mathscr{Y}$ со свойствами, нужными в этой теореме, применяя метод Гаусса n раз к интервальной матрице $f'_p(x)$со столбцами единичной матрицы в качестве правых частей.

Теорема 10. *Пусть отображение* f_p: $\mathfrak{B} \subseteq V_n(\mathbb{R}) \to V_n(\mathbb{R})$ *непрерывно дифференцируемо в* $\mathfrak{B}$, *причем производная имеет вычисление в интервальной арифметике для некоторого* $x \subseteq \mathfrak{B}$. *Если включение*

$$k_p(x) = y_p - \mathscr{Y}_p \cdot f_p(y_p) + (\mathscr{I}_p - \mathscr{Y}_p \cdot f'_p(x))(x - y_p) \subseteq x$$

имеет место для некоторого $y_n \in x$, *невырожденной матрицы* $\mathscr{Y}_p \in M_{nn}(\mathbb{R})$ *и единичной матрицы* $\mathscr{I}_p$, *то* f_p *имеет корень в* x. *Если же*

$$\{y_p - \mathscr{Y}_p \cdot f_p(y_p) + (\mathscr{I}_p - \mathscr{Y}_p \cdot f'_p(x))(x - y_p)\} \cap x = \varnothing,$$

то f_p *не имеет корней в* x.

Очевидным кандидатом на роль $\mathscr{Y}_p$ является $m(f'_p(x))^{-1}$.

Нахождение этой матрицы требует обращения точечной матрицы, а также некоторого числа умножений матриц на матрицы и некоторого числа умножений матриц на векторы для проверки включений. Существенную часть этих умножений можно сэкономить, если применить метод Гаусса к некоторой системе линейных уравнений, у которой матрица коэффициентов точечная, а правая часть — интервальный вектор. Это замечание уточняется в следующем утверждении.

Теорема 11. *Пусть отображение* $f_p: \mathfrak{B} \subseteq V_n(\mathbb{R}) \to V_n(\mathbb{R})$ *непрерывно дифференцируемо в* $\mathfrak{B}$, *причем производная имеет вычисление в интервальной арифметике для некоторого* $x \subseteq \mathfrak{B}$. *Предположим еще, что* $\mathscr{Y}_p$ — *невырожденная вещественная точечная матрица и что метод Гаусса в применении к точечной матрице* $\tilde{\mathscr{Y}}_p$ *и правой части* $f_p(y_p) - (\tilde{\mathscr{Y}}_p - f'_p(x))(x - y_p)$ *для некоторого* $y_p \in x$ *дает интервальный вектор* z. *Если теперь выполнено* $y_p - z \subseteq x$, *то* f_p *имеет нуль в* x. *Если же* $x \cap \{y_p - z\} = \varnothing$, *то* f_p *не имеет нулей в* x.

Доказательство. Рассмотрим отображение

$$p_p: x \subseteq \mathfrak{B} \subseteq V_n(\mathbb{R}) \to V_n(\mathbb{R}),$$

определяемое формулой

$$p_p(x_p) = x_p - \tilde{\mathscr{Y}}_p^{-1} f_p(x_p).$$

Мы видим, что верно равенство

$$f_p(x_p) = y_p - \tilde{\mathcal{Y}}_p^{-1} f_p(y_p) + x_p - y_p - \tilde{\mathcal{Y}}_p^{-1}(f_p(x_p) - f_p(y_p))$$
$$= y_p - \tilde{\mathcal{Y}}_p^{-1} f_p(y_p) + \tilde{\mathcal{Y}}_p^{-1}(\tilde{\mathcal{Y}}_p - \mathcal{J}_p(x_p))(x_p - y_p),$$

т. е. выполнено равенство

$$\tilde{\mathcal{Y}}_p(f_p(x_p) - y_p) = -f_p(y_p) + (\tilde{\mathcal{Y}}_p - \mathcal{J}_p(x_p))(x_p - y_p).$$

Ввиду $x_p \in x$ и $\mathcal{J}_p(x_p) \in f'_p(x)$ получаем, что

$$-f_p(y_p) + (\tilde{\mathcal{Y}}_p - \mathcal{J}_p(x_p))(x_p - y_p) \in$$
$$-f_p(y_p) + (\tilde{\mathcal{Y}}_p - f'_p(x))(x - y_p).$$

Если теперь применить метод Гаусса к системе линейных уравнений, заданной матрицей $\tilde{\mathcal{Y}}_p$ и правой частью $f_p(y_p) - (\tilde{\mathcal{Y}}_p - f'_p(x))(x - y_p)$, то получим интервальный вектор z, такой что имеет место $f_p(x_p) - y_p \in -z$ для $x_p \in x$ Если теперь

$$y_p - z \subseteq x,$$

то отсюда следует, что $f_p(x) \in x$ для $x_p \in x$. Теперь можно завершить доказательство так же, как в предыдущей теореме.

Основное достоинство этой теоремы заключается не в экономии вычислений по сравнению с процедурой из теоремы 10. По сравнению с аналогичным результатом из теоремы 8 главное достижение состоит в том, что метод Гаусса всегда применим к невырожденной точечной матрице. С другой стороны, условие $k_p(x) \subseteq x$ слабее, чем требование $y_p = m(x)$ в теореме 11, по крайней мере при выборе $y_p = m(x)$ (середина вектора *x)* и $\mathcal{Y}_p^{-1} = m(f'(x)) = \mathcal{Y}$. В этом можно убедиться следующим образом. Используя введенное ранее обозначение z для вектора, вычисленного методом Гаусса, мы имеем для произвольной неособенной матрицы $\mathcal{Y}_p$

$$z = g_p(\mathcal{Y}_p, f_p(y_p) - (\mathcal{Y}_p - f'_p(x))(x - y_p)).$$

Применяя доказанное в микромодуле 33 свойство (3) алгоритма Гаусса, мы получим

$$\mathcal{Y}_p^{-1}(f_p(y_p) - (\mathcal{Y}_p - f'(x))(x - y_p)) \subseteq$$
$$g_p(\mathcal{Y}_p, f_p(y_p) - (\mathcal{Y}_p - f'_p(x))(x - y_p)),$$

т. е. мы имеем включение

$$y_p - \mathcal{Y}_p^{-1}\left(f_p(y_p) - \left(\mathcal{Y}_p - f_p'\right)(x)\right)(x - y_p)) \subseteq y_p - z.$$

Покажем теперь, что при $y_p = m(x)$, $\mathcal{Y}_p = m\left(f_p'(x)\right)$ левая часть этого соотношения равна значению $\varepsilon_p(x)$ из теоремы 10.

Используя симметричность выражений $m\left(f_p'(x)\right) - f'(x)$ и $x - m(x)$, а также (8 из микромодуля 29) и последнюю формулу из (9 микромодуля 29), получим

$$m(x) - m\left(f_p'(x)\right)^{-1}\left(f_p(m(x)) - \left(m\left(f_p'(x)\right) - \left(f_p'(x)\right)(x - m(x))\right)\right)$$
$$= m(x) - m\left(f_p'(x)\right)^{-1} f_p(m(x)) - \left(\mathcal{I}_p - m\left(f_p'(x)\right)^{-1} f_p'(x)\right)$$
$$\times (x - m(x)) = \varepsilon_p(x).$$

Рассмотрим теперь некоторые теоремы существования, в которых участвует вычисление в интервальной арифметике второй производной. Для этого предположим, что отображение $f_p\colon \mathfrak{B} \subseteq V_n(\mathbb{R}) \to V_n(\mathbb{R})$ дважды непрерывно дифференцируемо и что вторая производная $f_p''(x_p) \in M_n(\mathbb{R})$ имеет вычисление в интервальной арифметике. Для некоторого $x \subseteq \mathfrak{B}$ и произвольного вначале вектора $m(x) \in x$ мы рассмотрим следующие интервальные векторы.

$$\varepsilon_{p1}(x) := m(x) - \mathcal{Y}_p \cdot f_p(m(x)) - \mathcal{Y}_p\left(f_p''(x)(x - m(x))(x - m(x))\right), \quad (22)$$

где $\mathcal{Y}_p = f_p'(m(x))^{-1}$,

$$\varepsilon_{p2}(x) := m(x) - \mathcal{Y}_p \cdot f_p(m(x)) - \frac{1}{2}\left(\mathcal{Y}_p \cdot f_p''(x)\right)(x - m(x)) \quad (23)$$
$$\times (x - m(x))$$

где $\mathcal{Y}_p = f_p'(m(x))^{-1}$,

$$\varepsilon_{p3}(x) := m(x) - \mathcal{Y}_p \cdot f_p(m(x)) \quad (24)$$
$$+ \left\{\mathcal{I}_p - \mathcal{Y}_p \cdot f_p'(m(x)) - \frac{1}{2}\left(\mathcal{Y}_p \cdot f_p''(x)\right)(x - m(x))\right\}$$
$$\times (x - m(x))$$

с произвольной невырожденной матрицей $\mathcal{Y}_p$.

Мы докажем сначала следующее утверждение, аналогичное предыдущей теореме.

Теорема 12. *Пусть отображение* $f_p\colon \mathfrak{B} \subseteq V_n(\mathbb{R}) \to V_n(\mathbb{R})$ *дважды непрерывно дифференцируемо в* $\mathfrak{B}$. *Если для некоторого* i, $1 \le i \le 3$ *и* $x \subseteq \mathfrak{B}$ *верно*

$$\varepsilon_{pi}(x) \subseteq x,$$

то f_p *имеет нуль в* x. *Если же* $k_{pi}(x) \cap x = \varnothing$ *для некоторого* i, $1 \le i \le 3$, *то* f_p *не имеет нулей в* x.

Доказательство. (22): Это случай $i=1$. Снова введем отображение

$$r_p(x_p) = x_p - \gamma_p \cdot f_p(x_p)$$

и воспользуемся представлением

$$r_p(x_p) = m(x) - \gamma_p \cdot f_p(m(x)) + (x_p - m(x)) - \gamma_p (f_p(x_p) - f_p(m(x)))$$

для $x_p \in x$. Применяя теорему о среднем значении, получим

$$f_p(x_p) - f_p(m(x)) = J_p(x_p, m(x))(x_p - m(x)).$$

Здесь i-я строка матрицы $J_p(x_p, m(x))$ равна $f_i'(z_p^{(i)})$, где

$$z_p^{(i)} = m(x) + \theta_i (x_p - m(x)) \text{ и } 0 < \theta_i < 1, \quad 1 \le i \le n.$$

Далее мы имеем, что

$$f_i'(z_p^{(i)}) = f_i'(m(x)) + \int_0^1 f_i''(m(x) + t(z_p^{(i)} - m(x)))(z_p^{(i)} - m(x))\,dt,$$

т. е.

$$J_p(x_p, m(x)) = f_p'(m(x)) + B_p,$$

где i-я строка матрицы B_P равна

$$\int_0^1 f_i''(m(x) + t(z_p^{(i)} - m(x)))(z_p^{(i)} - m(x))\,dt.$$

Отсюда следует, что

$$r_p(x_p) = m(x) - \gamma_p \cdot f_p(m(x)) + \{J_p - \gamma_p (f_p'(m(x)) - B_p)\}(x_p - m(x)).$$

Полагая

$$\gamma_p = f_p'(m(x))^{-1},$$

получим окончательно

$$r_p(x_p) = m(x) - \gamma_p \cdot f_p(m(x)) - \gamma_p B_p (x_p - m(x)).$$

Так как $t \in [0, 1]$ и $z_p^{(i)} \in x$, имеем $B_p \in f_p''(x)(x - m(x))$ и потому получаем

$$r_p(x_p) \in m(x) - \gamma_p \cdot f_p(m(x)) - \gamma_p (f_p''(x)(x - m(x))(x - m(x))) = k_{p1}(x).$$

(23) и (24): Для $k_{p2}(x)$ и $k_{p3}(x)$ можно рассуждать аналогично тому, как было сделано для $k_{p1}(x)$. Для $1 \le i \le n$ мы имеем

$$f_i(x_p) - f_i(m(x)) = f_i'(m(x))(x_p - m(x))$$
$$+ \frac{1}{2} f_i''(m(x) + \theta_i(x_p - m(x)))(x_p - m(x))(x_p - m(x)),$$
$$0 < \theta_i < 1,$$

т. е.

$$f_p(x_p) - f_p(m(x)) = f_p'(m(x))(x_p - m(x))$$
$$+ \frac{1}{2} \mathcal{B}_p(x - m(x))_p(x_p - m(x))$$

с билинейным оператором $\mathcal{B}_p$. Далее отсюда следует, что

$$\mathcal{Y}_p(f_p(x_p) - f_p(m(x))) = \mathcal{Y}_p \cdot f_p'(m(x))(x - m(x))$$
$$+ \frac{1}{2} \mathcal{Y}_p \mathcal{B}_p(x_p - m(x))(x_p - m(x)),$$

и потомуввиду $\mathcal{B}_p \in f_p''(x)$ мы получаем

$$f_p(x_p) = m(x) - \mathcal{Y}_p \cdot f_p(m(x)) + \{\mathcal{I}_p - \mathcal{Y}_p \cdot f_p'(m(x))$$
$$- \frac{1}{2}(\mathcal{Y}_p \mathcal{B}_p)(x_p - m(x))\}(x_p - m(x))$$
$$\in m(x) - \mathcal{Y}_p \cdot f_p(m(x)) + \left\{\mathcal{I}_p - \mathcal{Y}_p \cdot f_p'(m(x))\right.$$
$$\left. - \frac{1}{2}(\mathcal{Y}_p \cdot f_p'(x))(x - m(x))\right\}(x - m(x))$$
$$= k_{p3}(x).$$

Полагая $\mathcal{Y}_p = f_p'(m(x))^{-1}$, мы получаем $f_p(x_p) \in k_{p2}(x)$.

Теперь первая часть нашей теоремы доказывается с помощью теоремы Брауэра о неподвижной точке. Вторая часть устанавливается так же, как вторая часть теоремы 8.

Теперь мы сравним некоторые из этих теорем существования с некоторыми результатами, имеющимися в литературе. Определим замкнутыйшар с центром y_p и радиусом r:

$$\zeta(y_p, r) := \{x_p \in V_n(\mathbb{R}) \mid \| x_p - y_p \| \leqslant r\}.$$

Под нормой *всегда* понимаем ∞-нормы. Поэтому можно рассматривать $\zeta(y_p, r)$ как интервальный вектор, все компоненты которого имеют ширину $2r$. Нормы линейных и билинейных операторов определяются естественным образом исходя из векторной ∞-нормы. Первое из приводимых ниже уравнений содержит теорему существования Ньютона — Канторовича.

Теорема 13. *Пусть* $x = \zeta(m(x), r) \in V_n(I)(\mathbb{R}))$ *отображение* $f_0 : V_n(\mathbb{R}) \to V_n(\mathbb{R})$ *дважды непрерывно дифференцируемо и существует* $\mathcal{Y}_p := f'_p(m(x))^{-1}$.

Если выполнены условия

$$\| \mathcal{Y}_p \cdot f_p(m(x)) \| \leqslant \eta, \tag{25}$$

$$\| \mathcal{Y}_p \cdot f'_p(x_p) \| \leqslant k \quad \text{для} \quad x_p \in \zeta(m(x), r), \tag{26}$$

$$h := k\eta \leqslant \frac{1}{2}, \tag{27}$$

$$r \geqslant ((1 - \sqrt{1 - 2h})/h)\,\eta, \tag{28}$$

то f_p *имеет нуль в* $\zeta(m(x), r)$.

В следующей теореме изучается соотношение между теоремами 12 и 13. В теореме 12 в качестве 12$m(x)$ всегда берется центр.

Теорема 14. (a) *Пусть* $x = \zeta(m(x), r)$, *где* $r := \frac{1}{2}\| d(x) \|$, *и для* $r_i := \frac{1}{2}\| d(k_{pi}(x)) \|$ *имеет место*

$$\zeta(m(k_{pi}(x)), r_i) \subseteq \zeta(m(x), r) = x$$

для i=l, *или 2. Тогда выполнены условия* (25)—(28) *теоремы* 13.

(b) *Для* $i = 2$ *верно и обратное утверждение. Если выполнены условия* (25)—(28) *теоремы* 14 *и дополнительное ограничение*

$$r \leqslant ((1 + \sqrt{1 - 2h_2})/h_2)\,\eta,$$

то $\zeta(m(k_{p2}(x)), r_2) \subseteq \zeta(m(x), r) = x$. *В частности,* $k_{p2}(x) \subseteq x$.

Доказательство. (a) Включение $\zeta(m(k_{pi}(x)), r_i) \subseteq \zeta(m(x), r)$ выполнено тогда и только тогда, когда

$$\| m(k_{pi}(x)) - m(x) \| + r_i \leqslant r.$$

Из того, что

$$m(k_{pi}(x)) = m(x) - \mathcal{Y}_p \cdot f_p(m(x))$$

следует

$$\eta := \| \mathcal{Y}_p \cdot f_p(m(x)) \| = \| m(k_{pi}(x)) - m(x) \|.$$

В силу теоремы 12(b) из микромодуля 29, (c) ширина векторов $k_{p1}(x))$ и $k_{p2}(x)$ удовлетворяет неравенствам

$$d(k_{p1}(x)) = \frac{1}{2} | \mathcal{Y}_p | | f''_p(x) | d(x) d(x),$$

$$d(k_{p2}(x)) = \frac{1}{4} | \mathcal{Y}_p \cdot f''_p(x) | d(x) d(x).$$

Полагая

$$\mathscr{C}_p = |\mathscr{Y}_p||f''_p(x)| = (c_{ijk}), \quad k_1 = \max_{1 \leqslant i \leqslant n} \sum_{j,k=1}^{n} c_{ijk}$$

и

$$\mathscr{D}_p = |\mathscr{Y}_p \cdot f''_p(x)| = (d_{ijk}), \quad k_2 = \max_{1 \leqslant i \leqslant n} \sum_{j,k=1}^{n} d_{ijk},$$

получим

$$\|d(\mathscr{k}_{p1}(x))\| = 2k_1 r^2 \quad \text{и} \quad \|d(\mathscr{k}_{p2}(x))\| = k_2 r^2,$$

так как $d(x) = 2r(1, 1, \ldots, 1)^T$. Поэтому имеем

$$r_1 = \frac{1}{2}\|d(\mathscr{k}_p(x))\| = k_1 r^2, \quad r_2 = \frac{1}{2}\|d(\mathscr{k}_{p2}(x))\| = \frac{1}{2} k_2 r^2.$$

Наконец, условие теоремы $\zeta(m(\mathscr{k}_{pi}(x)), r_i) \subseteq \zeta(m(x), r)$ дает

$$\eta + k_1 r^2 \leqslant r \quad \text{для} \quad i = 1$$

и

$$\eta + \frac{1}{2} k_2 r^2 \leqslant r \quad \text{для} \quad i = 2.$$

Так как

$$\eta + \frac{1}{2} k_1 r^2 \leqslant \eta + k_1 r^2,$$

получаем

$$\eta + \frac{1}{2} k_i r^2 \leqslant r, \quad i = 1 \text{ или } 2.$$

Отсюда следует, что

$$h_i = k_i \eta \leqslant \frac{1}{2}, \quad i = 1 \text{ или } 2,$$

и

$$\frac{1 - \sqrt{1 - 2h_i}}{h_i} \eta \leqslant r \leqslant \frac{1 + \sqrt{1 - 2h_i}}{h_i} \eta, \quad i = 1 \text{ или } 2.$$

Это показывает, что выполнены условия (25), (27) и (28) теоремы 13. Остается проверить (26). Из $x_p \in \zeta(m(x), r) = \omega$ следует, что

$$\mathscr{Y}_p \cdot f''_p(x_p) \in \mathscr{Y}_p \cdot f''_p(x),$$

а потому и

$$|\mathscr{Y}_p \cdot f''_p(x_p)| \leqslant |\mathscr{Y}_p \cdot f''_p(x)| \leqslant |\mathscr{Y}_p||f''_p(x)|.$$

Отсюда следует, что

$$\|\mathscr{Y}_p \cdot f_p''(x_p)\| = \|\,|\mathscr{Y}_p \cdot f_p''(x_p)|\,\| \leqslant \|\,|\mathscr{Y}_p \cdot f_p''(x)|\,\| = \|\mathscr{D}_p\|$$

$$\leqslant \max_{1 \leqslant i \leqslant n} \sum_{j,k=1}^{n} d_{ijk}$$

и

$$\|\mathscr{Y}_p \cdot f_p''(x_p)\| \leqslant \|\mathscr{Y}_p\|\,\|f_p''(x_p)|\,\| = \|\mathscr{C}_p\|$$

$$\leqslant \max_{1 \leqslant i \leqslant n} \sum_{j,k=1}^{n} c_{ijk}.$$

Этим завершается доказательство (a).

(b) Заметим, что из условий теоремы следует неравенство

$$\eta + \frac{1}{2} k_2 r^2 \leqslant r,$$

эквивалентное включению

$$\zeta(m(\mathscr{k}_{p2}(x)),\ r_2) \subseteq \zeta(m(x),\ r) = x.$$

Ввиду $\mathscr{k}_{p2}(x) \subseteq \zeta(m(\mathscr{k}_{p2}(x)),\ r_2)$ отсюда следует, что

$$\mathscr{k}_{p2}(x) \equiv x.$$

Условие $\mathscr{k}_p(x) \subseteq x$ приводит в случае

$$x = \zeta(m(x),\ r)$$

к

$$\left((1 - \sqrt{1 - 4h})/2h\right)\eta \leqslant r \leqslant \left((1 + \sqrt{1 - 4h})/2h\right)\eta.$$

Здесь предполагается, что $h := b\lambda\eta \leqslant \frac{1}{4}$, где λ — некоторая константа Липшица для первой производной, и

$$\|\mathscr{Y}_p\| \leqslant b, \quad \|\mathscr{Y}_p \cdot f_p(m(x))\| \leqslant \eta\,(\mathscr{Y}_p = f_p'(m(x))^{-1}).$$

Вопрос о том, когда для $\mathscr{k}_{p1}(x) \subseteq x$ верно включение $x = \zeta(m(x),\ r)$, приводит по существу к тем же условиям, $\mathscr{k}_{p1}(x) \equiv x$ наверняка выполнено, если имеет место $\zeta(m(\mathscr{k}_{p1}(x)),\ r_1) \equiv \zeta(m(x),\ r)$. Как и в доказательстве п. (a) предыдущей теоремы, это эквивалентно неравенству

$$\eta + k_1 r^2 \leqslant r,$$

что приводит к оценке

$$h_1 = k_1\eta \leqslant \frac{1}{4}$$

и

$$((1-\sqrt{1-4h_1})/2h_1)\,\eta \leqslant r \leqslant ((1+\sqrt{1-4h_1})/2h_1)\,\eta.$$

Если теперь заметить, что λb является по существу верхней границей для k_1, то для включения $\mathcal{k}_{p1}(x) \subseteq x$ получатся те же достаточные условия, что и для $\mathcal{k}_p(x) \subseteq x$.

Отметим, наконец, что в случае одного уравнения с одним неизвестным условие $\mathcal{k}_{p2}(x) \subseteq x$ выполнено тогда и только тогда, когда верно

$$\zeta(m(\mathcal{k}_{p2}(x)), r_2) \subseteq \zeta(m(x), r).$$

Вместе с предыдущей теоремой это дает следующее утверждение.

Следствие 15. В *случае одного уравнения с одним неизвестным условия теоремы Ньютона — Канторовича выполнены тогда и только тогда, когда* $\mathcal{k}_{p2}(x) \subseteq x = \zeta(m(x), r)$.

Воспроизведем теперь один результат, чтобы сравнить его впоследствии с теоремой 12 для случая $\mathcal{k}_{p3}(x)$.

Теорема 16. *Пусть* $x = \zeta(m(x), r) \in V_n(I(\mathbb{R}))$ *и отображение* $\mathcal{f}_p\colon V_n(\mathbb{R}) \to V_n(\mathbb{R})$ *дважды дифференцируемо. Допустим, что для* $\mathcal{Y}_p \in M_{nn}(\mathbb{R})$ *выполнены следующие условия*

$$\|\mathcal{Y}_p \cdot \mathcal{f}_p(m(x))\| \leqslant \bar{\eta}, \tag{29}$$

$$\|\mathcal{I}_p - \mathcal{Y}_p \cdot \mathcal{f}_p'(m(x))\| \leqslant \delta < 1, \tag{30}$$

$$\|\mathcal{Y}_p \cdot \mathcal{f}_p''(x_p)\| \leqslant \bar{k} \text{ для всех } x_p \in \zeta(m(x), r), \tag{31}$$

$$h := \bar{k}\bar{\eta}/(1-\delta)^2 \leqslant \frac{1}{2}, \tag{32}$$

$$r \geqslant \bar{r}_0 = ((1-\sqrt{1-2h})/h)\,(\bar{\eta}/(1-\delta)). \tag{33}$$

Тогда функция $\mathcal{f}_p$ *имеет нуль в* $\zeta(m(x), r)$.

В формулировке следующей теоремы используется возможность записать $\mathcal{k}_{p3}(x)$ в виде

$$\mathcal{k}_{p3}(x) = m(x) - \mathcal{Y}_p \cdot \mathcal{f}_p(m(x)) + (\mathcal{I}_p - \mathcal{Y}_p \cdot \mathcal{f}_p'(m(x)))(x - m(x)) + \tfrac{1}{2}(\mathcal{Y}_p \cdot \mathcal{f}_p''(x))(x - m(x))(x - m(x)),$$

где $m(x)$ — центр x. Чтобы показать это, заметим, что величина $x - m(x)$ симметрична, так что исходное определение $\mathcal{k}_{p3}(x)$ можно переписать в виде

$$\mathcal{k}_{p3}(x) = m(x) - \mathcal{Y}_p \cdot \mathcal{f}_p(m(x)) + |\mathcal{I}_p - \mathcal{Y}_p \cdot \mathcal{f}_p'(m(x)) - \tfrac{1}{2}(\mathcal{Y}_p \cdot \mathcal{f}_p''(x))(x - m(x))|(x - m(x)).$$

Далее мы имеем

$$\begin{aligned}&\left|\mathscr{I}_p - \mathscr{Y}_p \cdot \mathscr{f}'_p(m(x)) - \frac{1}{2}\left(\mathscr{Y}_p \cdot \mathscr{f}''_p(x)\right)(x - m(x))\right| \\ &\quad = \left|\mathscr{I}_p - \mathscr{Y}_p \cdot \mathscr{f}'_p(m(x))\right| + \frac{1}{4} d\left(\left(\mathscr{Y}_p \cdot \mathscr{f}''_p(x)\right)(x - m(x))\right) \\ &\quad = \left|\mathscr{I}_p - \mathscr{Y}_p \cdot \mathscr{f}'_p(m(x))\right| + \frac{1}{4}\left|\mathscr{Y}_p \cdot \mathscr{f}''_p(x)\right| d(x).\end{aligned}$$

Так как обе матрицы, входящие в эту сумму, неотрицательны, мы можем записать $\mathscr{k}_{p3}(x)$ в виде

$$\begin{aligned}\mathscr{k}_{p3}(x) = m(x) - \mathscr{Y}_p \cdot \mathscr{f}_p(m(x)) + \left|\mathscr{I}_p - \mathscr{Y}_p \cdot \mathscr{f}'_p(m(x))\right|(x - m(x)) \\ + \frac{1}{4}\left|\mathscr{Y}_p \cdot \mathscr{f}''_p(x)\right| d(x)(x - m(x)),\end{aligned}$$

что и дает равенство

$$\begin{aligned}\mathscr{k}_{p3} = m(x) - \mathscr{Y}_p \cdot \mathscr{f}_p(m(x)) + \left(\mathscr{I}_p - \mathscr{Y}_p \cdot \mathscr{f}'_p(m(x))\right)(x - m(x)) \\ + \frac{1}{2}\left(\mathscr{Y}_p \cdot \mathscr{f}''_p(x)\right)(x - m(x))(x - m(x)).\end{aligned}$$

Теперь мы можем сформулировать следующее утверждение.

Теорема 17. *Положим*

$$\begin{aligned}r' &= \frac{1}{2}\left\| d\left(m(x) - \mathscr{Y}_p \cdot \mathscr{f}_p(m(x)) + \frac{1}{2}\left(\mathscr{Y}_p \cdot \mathscr{f}''_p(x)\right)\right.\right. \\ &\qquad \left.\left. \times (x - m(x))(x - m(x))\right)\right\|, \\ r'' &= \frac{1}{2}\left\| d\left(\left(\mathscr{I}_p - \mathscr{Y}_p \cdot \mathscr{f}'_p(m(x))\right)(x - m(x))\right)\right\|, \\ r_3 &= r' + r'', \\ r &= \frac{1}{2}\| d(x)\|, \\ \delta &= \left\|\mathscr{I}_p - \mathscr{Y}_p \cdot \mathscr{f}'_p(m(x))\right\| < 1\end{aligned}$$

и допустим, что имеет место

$$\zeta(m(\mathscr{k}_{p3}(x)), r_3) \subseteq \zeta(m(x), r) = \omega.$$

Тогда выполнены условия (29)—(33) *теоремы* 16 (а *значит, и условия теоремы* (13)). *Заметим, что ввиду*

$$\frac{1}{2}\| d(\mathscr{k}_{p3}(x))\| \leqslant r' + r''$$

наше предположение сильнее, чем наше предположение сильнее, чем требования на $\mathscr{k}_{p1}(x)$ *или* $\mathscr{k}_{p2}(x)$ *в теореме* 14.

Доказательство. Мы имеем

$$d(m(x) - \mathcal{Y}_p \cdot f_p(m(x)) + \frac{1}{2}(\mathcal{Y}_p \cdot f_p''(x))(x - m(x))(x - m(x)))$$
$$= \frac{1}{4}|\mathcal{Y}_p \cdot f_p''(x)|\, d(x)\, d(x).$$

Из того что $d(x) = 2r(1, 1, \ldots, 1)^T$, следует

$$r' = \frac{1}{2}\tilde{k}r^2,$$

где использованы обозначения

$$\mathcal{C}_p := |\mathcal{Y}_p \cdot f_p''(x)|$$

и

$$\tilde{k} = \max_{1 \leqslant i \leqslant n} \sum_{l,\, k} c_{ilk}.$$

Далеемы имеем

$$d((\mathcal{I}_p - \mathcal{Y}_p \cdot f_p'(m(x)))(x - m(x)))$$
$$= |\mathcal{I}_p - \mathcal{Y}_p \cdot f_p'(m(x))|\, d(x),$$

откуда следует

$$r'' = \delta r.$$

Условие $\zeta(m(\xi_{p3}(x)), r_3) \subseteq \zeta(m(x), r)$ выполнено тогда и только тогда, когда

$$\|m(\xi_{p3}(x)) - m(x)\| + r_3 \leqslant r.$$

Полагая $\tilde{\eta} = \|\mathcal{Y}_p \cdot f_p(m(x))\|$, получаем, что оно эквивалентно условию

$$\tilde{\eta} + r' + r'' \leqslant r$$

ввиду равенства

$$m(\xi_{p3}(x)) = m(x) - \mathcal{Y}_p \cdot f_p(m(x)).$$

Иными словами, оно эквивалентно неравенству

$$\tilde{\eta} + \frac{1}{2}\tilde{k}r^2 \leqslant r(1 - \delta).$$

Отсюда следует, что

$$h := \tilde{k}\tilde{\eta}/(1 - \delta)^2 \leqslant \frac{1}{2}$$

и

$$\frac{1 - \sqrt{1 - 2h}}{h} \frac{\tilde{\eta}}{1 - \delta} \leqslant r \leqslant \frac{1 + \sqrt{1 - 2h}}{h} \frac{\tilde{\eta}}{1 - \delta}.$$

Таким образом, мы проверили выполнение условий (29), (30) и (32), (33) теоремы 16. Остается доказать, что

$$\|\mathcal{Y}_p \cdot f_p''(x_p)\| \leqslant \tilde{k} \text{ для } x_p \in x.$$

Из $x_k \in x$ следует, что

$$|\mathcal{Y}_p \cdot f_p''(x_p)| \leqslant |\mathcal{Y}_p \cdot f_p''(x)|,$$

а потому

$$\|\mathcal{Y}_p \cdot f_p''(x_p)\| \leqslant \|\mathcal{Y}_p \cdot f_p''(x)\| = \|\mathcal{C}_p\| \leqslant \max_{1 \leqslant i \leqslant n} \sum_{j,k} c_{ijk}.$$

Как уже отмечена, из $k_{pl}(x) \subseteq x$ еще не следует, что выполнены условия теоремы 13. Поэтому рассмотрим следующий простой пример.

Пусть $f_p(x_p) = \begin{pmatrix} x^2 - 1 \\ y^2 - 0.01 \end{pmatrix}$, $x_p = \begin{pmatrix} x \\ y \end{pmatrix}$ и $x = \begin{pmatrix} [0.98, & 1.18] \\ [0, & 0.2] \end{pmatrix}$.

Тогда f_p имеет в x корень $x_p^* = \begin{pmatrix} 1 \\ 0.1 \end{pmatrix}$. Ввиду $m(x) = \begin{pmatrix} 1.08 \\ 0.1 \end{pmatrix}$

мы имеем $f_p(m(x)) = \begin{pmatrix} 0.1664 \\ 0 \end{pmatrix}$. Далее получаем $f_p'(x) =$

$$= \begin{pmatrix} 2x & 0 \\ 0 & 2y \end{pmatrix}, \quad f_p'(m(x)) = \begin{pmatrix} 2.16 & 0 \\ 0 & 0.2 \end{pmatrix}, \quad f_p'(m(x))^{-1} = \begin{pmatrix} 1/2.16 & 0 \\ 0 & 5 \end{pmatrix},$$

$$f_p'(m(x))^{-1} f_p(m(x)) = \begin{pmatrix} 0.1664/2.16 \\ 0 \end{pmatrix}, \text{ т. е. } \eta = 0.1664/2.16.$$

Вторая производная равна

$$f_p''(x_p) = \left(\begin{array}{cc|cc} 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 \end{array}\right),$$

откуда следует

$$f_p'(m(x))^{-1} f_p''(x_p) = \left(\begin{array}{cc|cc} 2/2.16 & 0 & 0 & 0 \\ 0 & 0 & 0 & 10 \end{array}\right)$$

для $x_p \in x$. Простое вычисление дает $\|f_p'(m(x))^{-1} f_p''(x_p)\|_\infty = 10 =: k$. Отсюда следует $h = k\eta = 1.664/2.16 > \frac{1}{2}$, т. е. теорема 13 не применима. С другой стороны, для величины $k_{p1}(x)$ из теоремы 12 мы, полагая $\mathcal{Y}_p = m(f_p'(x))^{-1}$, получаем, что

$$k_{p1}(x) = m(x) - \mathcal{Y}_p \cdot f_p(m(x_j) - \mathcal{Y}_p \cdot (f_p''(x)(x - m(x))(x - m(x)))$$

$$\subseteq \begin{pmatrix} [0.9937, & 1.0123] \\ [0, & 0.2] \end{pmatrix} \subseteq x.$$

Это означает, что применима теорема 12 при i=1 (равенство (22)). Аналогичное вычисление показывает, что в нашем примере выполнено $k_{p2}(x) \subseteq x$ и $k_p(x) \subseteq x$ (см. теорему 10) при

$$\mathcal{Y}_p = m\left(f'_p(x)\right)^{-1}.$$

Если сравнивать достоинства и недостатки теорем 12 и 13, то теорема 12 кажется предпочтительнее, так как она применима к произвольным интервальным векторам, а не только к шарам, определяемым ∞-нормой. С другой стороны, результат теоремы 13 справедлив для произвольных норм. Эти замечания показывают, что теоремы существования, сформулированные в этом микромодуле, можно считать существенным дополнением к теореме Ньютона — Канторовича.

Это верно и для утверждений о существовании, содержащихся в теоремах 8—11. Их важное преимущество по сравнению с классическим утверждением из теоремы 13 состоит в том, что не нужны ни вторая производная, ни константа Липшица.

Замечания. Рокн и Ланкастер также занимались задачей о локализации решений системы уравнений $f_p(x_p) = o_p$, похожей на задачу с одним уравнением от одной переменной. Для системы делается один шаг метода Ньютона и с помощью оценок для f_p, f'_p и f''_p находится локализация. В действительности это применение теоремы Канторовича. Задача о собственных числах сводится известным способом к решению системы нелинейных уравнений. Существует модификация интервального метода Ньютона (7). При этой модификации нет необходимости вычислять локализации для обращений всех матриц, содержащихся в данной интервальной матрице. Для итерационной процедуры

$$\begin{cases} x^{(k+1)} = \left\{m\left(x^{(k)}\right) - \mathcal{Y}_p \cdot f_p\left(m\left(x^{(k)}\right)\right) + \left(\mathcal{I}_p - \mathcal{Y}_p \cdot f'_p\left(x^{(k)}\right)\right)\right. \\ \qquad \left. \times \left(x^{(k)} - m\left(x^{(k)}\right)\right)\right\} \cap x^{(k)}, \\ m\left(x^{(k)}\right) \in x^{(k)} \end{cases} \tag{34}$$

требуется только матрица $\mathcal{Y}_p$. Если выполнено неравенство

$$\left\| \left| \mathcal{I}_p - \mathcal{Y}_p \cdot f'_p\left(x^{(k)}\right) \right| \right\| < \frac{1}{2},$$

то доказывается сходимость локализаций к нулю. Если $\mathcal{Y}_p$ замена на $\mathcal{Y}_p^{(k)} = f'_p\left(m\left(x^{(k)}\right)\right)^{-1}$, то доказывается суперлинейная сходимость. Эта процедура используется для локализации собственных чисел вещественных матриц. Выражение в левой части пересечения, определяющего $x^{(k)}$ в формулах (34),— это $k_p(x)$ из теоремы 10.

В ряде работ исследуется сходимость метода (34). Докажем здесь обобщение этих результатов. В действительности мы хотим показать, что если соотношение

$$k_p(x^{(0)}) = m(x^{(0)}) \mathcal{Y}_p \cdot f_p(m(x^{(0)})) + (\mathcal{J}_p - \mathcal{Y}_p \cdot f_p'(x^{(0)}))(x^{(0)} - m(x^{(0)})) \subseteq x^{(0)}$$

имеет место для некоторой матрицы $\mathcal{Y}_p \in M_{nn}(\mathbb{R})$ и интервального вектора $x^{(0)}$ с центром $m(x^{(0)})$, то итерационный метод (34), где $x^{(k)}$ — центр вектора $m(x^{(k)})$ при $k \geq 0$, сходится к единственному нулю y_p функции f_p в интервале $x^{(0)}$.

Отношение $k_p(x^{(0)}) \subseteq x^{(0)}$ определяется аналогично тому, как это сделано в теореме 14.6. Сформулированное выше условие $k_p(x^{(0)}) \subseteq x^{(0)}$ влечет за собой $d(x^{(0)}) > o_p$. Мы имеем $k_p(x^{(0)}) \subset x^{(0)}$ тогда и только тогда, когда расстояние от $k_p(x^{(0)})$ до $m(x^{(0)})$ удовлетворяет неравенству

$$q(k_p(x^{(0)}), \quad m(x^{(0)})) < \frac{1}{2} d(x^{(0)}),$$

которое эквивалентно неравенству

$$|k_p(x^{(0)}) - m(x^{(0)})| < \frac{1}{2} d(x^{(0)}).$$

Ввиду равенства

$$|k_p(x^{(0)}) - m(x^{(0)})| = |\mathcal{Y}_p \cdot f_p(m(x^{(0)}))| + \frac{1}{2} |\mathcal{J}_p - \mathcal{Y}_p \cdot f(x^{(0)})| d(x^{(0)})$$

это дает

$$|\mathcal{J}_p - \mathcal{Y}_p \cdot f_p'(x^{(0)})| d(x^{(0)}) < d(x^{(0)}).$$

Отсюда следует, что спектральный радиус неотрицательной матрицы $|\mathcal{J}_p - \mathcal{Y}_p \cdot f_p'(x^{(0)})|$ меньше 1. (Все скалярные произведения строк на положительный вектор $d(x^{(0)})$ меньше 1.) Предельное значение x^*, полученное согласно (34), удовлетворяет равенству

$$x^* = \{m(x^*) - \mathcal{Y}_p \cdot f_p(m(x^*)) + (\mathcal{J}_p - \mathcal{Y}_p \cdot f_p'(x^*))(x^* - m(x^*))\} \cap x^*,$$

т. е.

$$d(x^*) \leqslant |\mathcal{J}_p - \mathcal{Y}_p \cdot f_p'(x^*)| d(x^*).$$

Из монотонности включения следует, что $f_p'(x^*) \subseteq f_p'(x^{(0)})$, и поэтому мы имеем

$$|\mathcal{J}_p - \mathcal{Y}_p \cdot f_p'(x^*)| \leqslant |\mathcal{J}_p - \mathcal{Y}_p \cdot f_p'(x^{(0)})|.$$

Из теоремы Перрона и Фробениуса о монотонности спектрального радиуса неотрицательных матриц следует, что выполняется и неравенство $\rho\left(\left|\mathscr{I}_p - \mathscr{Y}_p \cdot f'_p(x^*)\right|\right) < 1$.

Теперь мы имеем $d(x^*) = \varepsilon_p$ и $\lim_{k\to\infty} x^{(k)} = y_p$. Результат Мура, упомянутый выше, содержит частный случай интервалов, для которых $d\left(X_1^{(0)}\right) = \ldots = d\left(X_n^{(0)}\right)$ («гиперкубов»).

По аналогии с методом (34), основанным на теореме существования 10, можем, исходя из теоремы 11, определить следующий итерационный метод:

$$\begin{cases} \tilde{x}^{(k+1)} = m(x^{(k)}) - g_p(\mathscr{Y}_p, f_p(m(x^{(k)})) \\ \quad - (\mathscr{Y}_p - f'_p(x^{(k)}))(x^{(k)} - m(x^{(k)})), \\ \qquad \text{где } m(x^{(k)}) \in x^{(k)} - \text{центр } x^{(k)}, \\ x^{(k+1)} = \tilde{x}^{(k+1)} \cap x^{(k)}, \quad k \geqslant 0. \end{cases} \tag{35}$$

Объем вычислений на каждом шаге этого метода значительно меньше, чем для метода (34). (Здесь через $g_p(\mathscr{A}, \mathscr{b})$ снова обозначен интервальный вектор, полученный применением метода Гаусса к интервальной матрице $\mathscr{A}$ и правой части $\mathscr{b}$.

Используем сокращение

$$\tilde{k}_p(x) = m(x) - g_p\left(\mathscr{Y}_p, f_p(m(x)) - (\mathscr{Y}_p - f'_p(x))(x - m(x))\right).$$

Аналогично утверждению, доказанному для метода (34), покажем следующее. Если для некоторой матрицы $\mathscr{Y}_p \in M_{nn}(\mathbb{R})$ и интервального вектора $x^{(0)}$ с центром $m(x^{(0)})$ верно соотношение $\tilde{k}_p(x^{(0)}) \subset x^{(0)}$, то итерационный метод (35), в котором $m(x^{(k)})$ для $k \geq 0$ означает центр $x^{(k)}$, сходится к единственному корню $\lim_{k\to\infty} x^{(k)} = x_p^*$.

Чтобы доказать это утверждение, заметим, что по теореме 11 из $\tilde{k}_p(x^{(0)}) \subset x^{(0)}$ следует существование корня x_p^* функции f_p в $x^{(0)}$. Используя приведенные ранее соотношения для метода Гаусса, можем теперь записать

$$
\begin{aligned}
x_p^* &= x_p^* - \mathcal{Y}_p^{-1} f_p(x_p^*) \\
&= m(x^{(0)}) - \mathcal{Y}_p^{-1} f_p(m(x^{(0)})) + x_p^* - m(x^{(0)}) \\
&+ \mathcal{Y}_p^{-1}(f_p(x^{(0)})) - f_p(x_p^*)) \\
&\in m(x^{(0)}) - \mathcal{Y}_p^{-1} f_p(m(x^{(0)})) \\
&- (\mathcal{Y}_p - f_p'(x^{(0)})(x^{(0)} - m(x^{(0)}))) \\
&\subseteq m(x^{(0)}) - g_p(\mathcal{Y}_p, f_p(m(x^{(0)})) \\
&- (\mathcal{Y}_p - f_p'(x^{(0)})(x^{(0)} - m(x^{(0)}))) = \tilde{k}_p(x^{(0)}).
\end{aligned}
$$

Поэтому $x^{(1)}$ существует и имеет место $x_p^* \in x^{(1)}$. С помощью математической индукции получаем, что $x_p^* \in x^{(k)}$, $k \geq 0$.

Условие $\tilde{k}_p(x^{(0)}) \subset x^{(0)}$ эквивалентно покомпонентному собственному включению. Иными словами, найдется вещественное число $0 \leqslant \alpha < 1$, для которого

$$d(\tilde{k}_p(x^{(0)})) \leqslant \alpha d(x^{(0)}).$$

Из

$$\tilde{x}^{(1)} = \tilde{k}_p(x^{(0)}) \subset x^{(0)}$$

следует, что

$$x^{(1)} = \tilde{k}_p(x^{(0)}) \cap x^{(0)} = \tilde{k}_p(x^{(0)}) = \tilde{x}^{(1)},$$

а потому

$$d(\tilde{x}^{(1)}) = d(x^{(1)}) \leqslant \alpha d(x^{(0)}).$$

Из $x^{(1)} \subset x^{(0)}$ и монотонности включения следует, что

$$|\mathcal{Y}_p - f_p'(x^{(1)})| \leqslant |\mathcal{Y}_p - f_p'(x^{(0)})|,$$

а потому

$$
\begin{aligned}
& d(f_p(m(x^{(1)})) - (\mathcal{Y}_p - f_p'(x^{(1)}))(x^{(1)} - m(x^{(1)}))) \\
&= |\mathcal{Y}_p - f_p'(x^{(1)})| d(x^{(1)}) \\
&\leqslant \alpha |\mathcal{Y}_p - f_p'(x^{(0)})| d(x^{(0)}) \\
&= \alpha d(f_p(m(x^{(0)})) - (\mathcal{Y}_p - f_p'(x^{(0)}))(x^{(0)} - m(x^{(0)}))).
\end{aligned}
$$

Ранее было доказано, что отсюда следует

$$d(\tilde{x}^{(2)}) \leqslant \alpha d(\tilde{x}^{(1)}),$$

т. е.

$$d(\tilde{x}^{(2)}) \leqslant \alpha^2 d(x^{(0)}).$$

Из

$$d(x^{(2)}) \leqslant \inf\{d(\tilde{x}^{(2)}),\ d(x^{(1)})\}$$

следует, что

$$d(x^{(2)}) \leqslant \alpha^2 d(x^{(0)}).$$

Методом математической индукции доказываем, что

$$d(x^{(k+1)}) \leqslant \alpha^{k+1} d(x^{(0)}), \quad k \geqslant 0.$$

Отсюда следует, что $\lim_{k\to\infty} d(x^{(k)}) = o_p$, а ввиду $x_p^* \in x^{(k)}$ это дает $\lim_{k\to\infty} x^{(k)} = x_p^*$, что и завершает доказательство.

Теперь мы хотим сделать еще несколько замечаний о методе (35). (Те же замечания справедливы и для метода (34).) Если применять метод (35), начиная с произвольной невырожденной матрицы $\mathcal{Y}_p$ и произвольного интервального вектора $x^{(0)}$, не обязательно удовлетворяющего условию $\tilde{k}_p(x^{(0)}) \subset x^{(0)}$, то могут представиться следующие два случая.

(i) Для некоторого $k_0 \geqslant 0$ верно $\tilde{x}^{(k_0+1)} \cap x^{(k_0)} = \phi$. Тогда f_p не имеет нулей в $x^{(0)}$. Действительно, если бы f_p имела нуль в $x^{(0)}$, то, как и в случае $\tilde{k}_p(x^{(0)}) \subset x^{(0)}$, мы могли бы показать, что пересечение непусто для всех k.

(ii) Метод (35) определен для всех $k \geq 0$. Тогда верно равенство $\lim_{k\to\infty} x^{(k)} = x^*$. Если теперь $d(x^*) \neq o_p$, то невозможно сказать что-либо о наличии или отсутствии нуля. Если же $d(x^*) = o_p$, т. е. $\lim_{k\to\infty} x^{(k)} = x_p^*$, то x_p^* — нуль функции f_p. Это получается из (35) переходом к пределу при $k \to \infty$.

Условие $d(\tilde{k}_p(x^{(0)})) \leqslant \alpha d(x^{(0)}), \quad 0 \leqslant \alpha < 1$, оказывается достаточным для того, чтобы имело место (i) или (ii) с $d(x^*) = o_p$. Если это условие выполнено, то те же рассуждения, что и в случае более сильного условия $\tilde{k}_p(x^{(0)}) \subset x^{(0)}$, показывают, что для всех $k \geq 0$, для которых $\tilde{x}^{(k+1)} \cap x^{(k)} \neq \phi$, верно неравенство $d(x^{(k+1)}) \leqslant \alpha^{k+1} d(x^{(0)})$. Если f_p не имеет корней в $x^{(0)}$, то пересечение станет пустым для некоторого $k \geq 0$, так как иначе равенство $\lim_{k\to\infty} x^{(k)} = x_p^*$ противоречило бы равенству $f_p(x_p^*) = o_p$. Если же f_p имеет корень x^* в $x^{(0)}$, то так же, как при

условии $\tilde{k}_p(x^{(0)}) \subset x^{(0)}$, мы покажем, что пересечение непусто при любом $k \geq 0$, а потому верно

$$\lim_{k \to \infty} x^{(k)} = x_p^* \quad \text{и} \quad f_p(x_p^*) = 0_p.$$

Локализация решений системы линейных уравнений, например системы вида (6), используется в методах (7) и (12). Такую локализацию можно в принципе найти с помощью метода Гаусса. Такой метод наверняка будет сходиться, если ширина вектора $x^{(0)}$ достаточно мала. Пока не найдено проверяемых условий, гарантирующих сходимость. Было бы чрезвычайно полезно иметь такие критерии, так как это позволило бы избежать обращения матриц $m(\mathcal{F}^{(k)})$ и применения итерации (10). Тогда можно было бы проводить итерационную локализацию решения y_p с разумными вычислительными затратами. Хотя в (14) приходится применять метод Гаусса и делать шаг (d), эта процедура все же намного дешевле в смысле объема вычислений, чем другие методы, рассмотренные в этом микромодуле. Важно и то, что в теореме 7 нет ограничений на величину или ширину вектора $x^{(0)}$. Поэтому утверждение о сходимости в теореме 7 дает практически глобальную сходимость. Численный пример применения метода (14) показывает важность этого утверждения. Хорошо известно, что решение дискретизированной граничной задачи методом Ньютона не представляет проблем, когда функция *f(t, y)* выпукла по *y*. Однако в невыпуклом случае сходимость удается доказать лишь для начальных значений, достаточно близких к решению. Для метода (14) не нужно никаких условий выпуклости.

В микромодуле 39 рассматривается метод, который также решает упомянутую дискретизированную граничную задачу, но не использует метода Гаусса. Однако там для достижения квадратичной сходимости требуется на каждом шаге не только интервальное вычисление производной, но и два вычисления значений самой функции.

Применение метода Гаусса для локализации множества решений системы (4) или (6) приводит к алгоритму

$$x^{(k+1)} = \left\{ m(x^{(k)}) - g_p\left(f_p'(x_p^{(k)}),\ f_p(m(x^{(k)}))\right) \right\} \cap x^{(k)}. \quad (36)$$

Здесь $g_p(\cdot, \cdot)$ означает отображение, определяемое методом Гаусса.

Микромодуль 38

Методы Ньютоновского типа не использующие обращения матриц

В предыдущем мукромодуле было показано, что итерация (12 из микромодуля 37) сходится по ширине последовательных приближений не хуже, чем квадратично. На каждом шаге приходится обращать вещественную матрицу $m(\mathcal{F}^{(k)})$ и производить вычисления по формулам (10 из микромодуля 37), пока там не будет достигнута неподвижная точка. Теперь мы изменим (12 из микромодуля 37) так, что будет выполняться всего один шаг итерационного нахождения локализации $\nabla^{(k)}$ для множества $\{\mathcal{F}_p^{-1} \mid \mathcal{F}_p \in \mathcal{F}^{(k)}\}$. Напомним, что в (12 из микромодуля 37) приходится проводить итерацию (10 из микромодуля 37). Мы выберем новый метод итерационной локализации обращения интервальной матрицы. Он будет обобщением на интервальные матрицы алгоритма (9 из микромодуля 36) при $r = 2$.

Теорема 1. *Пусть даны вещественная невырожденная матрица* $\mathcal{A}_p$ *и последовательность интервальных матриц* $\{\mathcal{A}^{(k)}\}_{k=0}^{\infty}$, *для которой* $\mathcal{A}_p \in \mathcal{A}^{(k)}$, $k \geqslant 0$. *Пусть* $\lim_{k\to\infty} \mathcal{A}^{(k)} = \mathcal{A}_0$ *и* $\mathcal{W}^{(0)}$ — *интервальная матрица, для которой* $\mathcal{A}_p^{-1} \in \mathcal{W}^{(0)}$. *Для итерационного метода*

$$\mathcal{W}^{(k+1)} = \{m(\mathcal{W}^{(k)}) + \mathcal{W}^{(k)}(\mathcal{I}_p - \mathcal{A}^{(k)} m(\mathcal{W}^{(k)}))\} \cap \mathcal{W}^{(k)}, \tag{1}$$

иде $m(\mathcal{W}^{(k)})$ *для* $k \geqslant 0$ *определено в* (1 из микромодуля 37), *верно следующее;*

$$\mathcal{A}_p^{-1} \in \mathcal{W}^{(k)}, \quad k \geqslant 0. \tag{2a}$$

Если любая матрица $\mathcal{W}_p \in \mathcal{W}^{(0)}$ *невырожденная, то*

$$\lim_{k\to\infty} \mathcal{W}^{(k)} = \mathcal{A}_p^{-1}. \tag{2b}$$

Доказательство. (2a): По условию имеем

$$\mathcal{A}_p^{-1} \in \mathcal{W}^{(0)} \quad \text{и} \quad \mathcal{A}_p \in \mathcal{A}^{(0)}.$$

Отсюда и из (10' из микромодуля 29) следует, что

$$\mathcal{A}_p^{-1} = m(\mathcal{W}^{(0)}) + \mathcal{A}_p^{-1}(\mathcal{I}_p - \mathcal{A}_p \cdot m(\mathcal{W}^{(0)}))$$
$$\in \{m(\mathcal{W}^{(0)}) + \mathcal{W}^{(0)}(\mathcal{I}_p - \mathcal{A}^{(0)} m(\mathcal{W}^{(0)}))\} \cap \mathcal{W}^{(0)}.$$

Доказательство завершается методом математической индукции. (2b): Монотонно убывающая последовательность матриц

$$\mathscr{W}^{(0)} \supseteq \mathscr{W}^{(1)} \supseteq \mathscr{W}^{(2)} \supseteq \ldots$$

сходится к некоторой интервальной матрице $\mathscr{W}$. В силу непрерывной зависимости центра $m(\mathscr{W}^{(k)})$ от $\mathscr{W}^{(k)}$ имеем

$$\lim_{k\to\infty} m(\mathscr{W}^{(k)}) = m(\mathscr{W}).$$

Отсюда и из непрерывности операций, входящих в (1), следует, что

$$\mathscr{W} = \lim_{k\to\infty} (\{m(\mathscr{W}^{(k)}) + \mathscr{W}^{(k)}(\mathscr{I}_p - \mathscr{A}^{(k)} m(\mathscr{W}^{(k)}))\} \cap \mathscr{W}^{(k)})$$

$$= \{m(\mathscr{W}) + \mathscr{W}(\mathscr{I}_p - \mathscr{A}_p \cdot m(\mathscr{W}))\} \cap \mathscr{W},$$

и потому

$$m(\mathscr{W}) \in m(\mathscr{W}) + \mathscr{W}(\mathscr{I}_p - \mathscr{A}_p \cdot m(\mathscr{W})).$$

Обозначим j-й столбец матрицы $m(\mathscr{W})$ через

$$(m(\mathscr{W}))_j.$$

В силу (1 из микромодуля 29) мы получаем представление

$$(m(\mathscr{W}))_j = (m(\mathscr{W}))_j + \mathscr{W}_p^{(j)}\left((\mathscr{I}_p - \mathscr{A}_p \cdot m(\mathscr{W}))\right)_j$$

с матрицей $\mathscr{W}_p^{(j)} \in \mathscr{W} \subseteq \mathscr{W}^{(0)}$. Отсюда следует, что

$$\mathscr{W}_p^{(j)}\left((\mathscr{I}_p - \mathscr{A}_p \cdot m(\mathscr{W}))\right)_j = o_p.$$

Матрицы $\mathscr{A}_p$ и $\mathscr{W}_p^{(j)}$ невырожденные по условию, поэтому, используя

$$(\mathscr{A}_p m(\mathscr{W}))_j = \mathscr{A}_p (m(\mathscr{W}))_j,$$

мы получаем

$$(m(\mathscr{W}))_j = (\mathscr{A}_p^{-1})_j,$$

что доказывает равенство $m(\mathscr{W}) = \mathscr{A}_p^{-1}$. Но отсюда следует равенство $\mathscr{W} = \mathscr{A}_p^{-1}$, что и завершает доказательство теоремы.

Итерационный метод (9 из микромодуля 36) при $r = 2$ получается теперь как частный случай ппи $\mathscr{A}^{(k)} = \mathscr{A}_p$. Если мы предположим, что последовательность $\{\mathscr{A}^{(k)}\}_{k=0}^{\infty}$ в (1) удовлетворяет условию

$$\mathscr{A}^{(0)} \supseteq \mathscr{A}^{(1)} \supseteq \mathscr{A}^{(2)} \supseteq \ldots, \text{ где } \lim_{k\to\infty} \mathscr{A}^{(k)} = \mathscr{A},$$

то аналогично (2a) получим, что из включения

$$\{\mathscr{A}_p^{-1} \mid \mathscr{A}_p \in \mathscr{A}\} \subseteq \mathscr{W}^{(0)}$$

всегда следует

$$\{\mathscr{A}_p^{-1} \mid \mathscr{A}_p \in \mathscr{A}\} \subseteq \mathscr{W}^{(k)}, \quad k \geqslant 0.$$

Если теперь выбрать $\mathscr{A}^{(k)} = \mathscr{A}$, то мы получим из (1) метод итерационной локализации обратной матрицы. Аналогичный итерационный метод был рассмотрен в (10 из микромодуля 37).

Теперь используем (1), чтобы построить метод, аналогичный методу (12 из микромодуля 37). Пусть $x^{(0)}$ — интервальный вектор, содержащий нуль функции $f_p(x_p)$. Допустим еще, что $\mathscr{Y}^{(0)}$ — интервальная матрица, такая что $J_p(x_p)^{-1} \in \mathscr{Y}^{(0)}$ для $x_p \in x^{(0)}$, где $J_p(x_p)$ определенаформулами (3 из микромодуля 36). Положим, как и раньше, $\mathscr{F}^{(k)} = f'_p(x^{(k)})$.

Рассмотрим теперь итерацию

$$\begin{cases} x^{(k+1)} = \{m(x^{(k)}) - \mathscr{Y}^{(k)} f_p(m(x^{(k)}))\} \cap x^{(k)}, \\ \mathscr{Y}^{(k+1)} = \{m(\mathscr{Y}^{(k)}) + \mathscr{Y}^{(k)}(J_p - \mathscr{F}^{(k+1)} m(\mathscr{Y}^{(k)}))\} \cap \mathscr{Y}^{(k)}, \quad k \geqslant 0. \end{cases} \quad (3)$$

Метод (3) заключается в одновременном исполнении (7 из микромодуля 36) и (1). Его главное отличие от итерации (12 из микромодуля 36) состоит в вычислении новой интервальной матрицы $\mathscr{Y}^{(k+1)}$. Здесь исполняется только один шаг подходящего метода.

В методе (3) вычисляются последовательности интервальных матриц и интервальных векторов

$$x^{(0)} \supseteq x^{(1)} \supseteq x^{(2)} \supseteq \ldots \quad \text{и} \quad \mathscr{Y}^{(0)} \supseteq \mathscr{Y}^{(1)} \supseteq \mathscr{Y}^{(2)} \supseteq \ldots$$

Докажем некоторые свойства этих последовательностей.

Теорема 2. *Пусть $x^{(0)}$ — интервальный вектор, а $y_p \in x^{(0)}$ — нуль функции $f_p(x_p)$. Пусть производная функции $f_p(x_p)$ удовлетворяет условию Липшица при значении аргумента $x^{(0)}$, а $\mathscr{Y}^{(0)}$ — интервальная матрица, содержащая матрицы $J_p(x)^{-1}$ для всех $x_p \in x^{(0)}$. Пусть, наконец, условие*

$$\|d(f'_p(x))\|' \leqslant c\,\|d(x)\|'$$

выполнено при $x \subseteq x^{(0)}$ для вычисления в интервальной арифметике $f'_p(x)$ производной Фреше $f'_p(x_p)$. (Выполнение этого *условия обеспечено, если любой элемент матрицы удовлетворяет условию* (5' п. 7.3).)

Тогда интервальные векторы $\{x^{(k)}\}_{k=}^{\infty}$ и интервальные матрицы $\{\mathscr{Y}^{(k)}\}_{k=0}^{\infty}$, вычисленные по формулам (3), *удовлетворяют следующим условиям.*

Каждый интервальный вектор $x^{(k)}$, $k \geqslant 0$, содержит корень y_p. (4)

Если все матрицы $\mathcal{Y}_p \in \mathcal{Y}^{(0)}$ *невырождены, то* (5)

$$\lim_{k\to\infty} x^{(k)} = y_p \quad и \quad \lim_{k\to\infty} \mathcal{Y}^{(k)} = f'_p(y_p)^{-1}.$$

Пусть $\mathcal{H}^{(k)} \in M_{n,\,n+1}(I(\mathbb{R}))$, $k \geqslant 0$, *— интервальная матрица, первый столбец которой совпадает с интервальным вектором* $x^{(k)}$, *а остальные являются столбцами матрицы* $\mathcal{Y}^{(k)}$, *т. е.* $\mathcal{H}^{(k)} = (x^{(k)}, \mathcal{Y}^{(k)})$. *Тогда*

$$O_R\left((3), \left(y_p, f'_p(y_p)^{-1}\right)\right) \geqslant 2 \qquad (6)$$

(см. приложение А, *теорема* 2).

Доказательство. Снова положим $\mathcal{F}^{(k)} = f'_p(x^{(k)})$.

(4): Точно так же, как в доказательстве соответствующего утверждения (8 из микромодуля 36) теоремы 1 из микромодуля 36, мы показываем, что $y_p \in x^{(1)}$. В силу $x^{(1)} \subseteq x^{(0)}$ отсюда следует, что для $x_p \in x^{(1)}$ верно

$$\mathcal{J}_p(x_p)^{-1} \in \mathcal{Y}^{(0)} \quad \text{и} \quad \mathcal{J}_p(x_p) \in \mathcal{F}^{(1)}.$$

Поэтому с помощью (10 из микромодуля 29) получаем

$$\begin{aligned}\mathcal{J}_p(x_p)^{-1} &= m(\mathcal{Y}^{(0)}) + \mathcal{J}_p(x_p)^{-1}(\mathcal{J}_p - \mathcal{J}_p(x_p)m(\mathcal{Y}^{(0)})) \\ &\in \{m(\mathcal{Y}^{(0)} + \mathcal{Y}^{(0)}(\mathcal{J}_p - \mathcal{F}^{(1)}m(\mathcal{Y}^{(0)}))\} \cap \mathcal{Y}^{(0)} = \mathcal{Y}^{(1)}.\end{aligned}$$

Таким образом, мы имеем

$$\{\mathcal{J}_p(x_p)^{-1} \mid x_p \in x^{(1)}\} \subseteq \mathcal{Y}^{(1)}.$$

Используя этот результат, мы показываем, что

$$y_p \in x^{(2)},$$

и (4) получается методом математической индукции.

(5): Ввиду $\mathcal{Y}^{(k)} \subseteq \mathcal{Y}^{(0)}$ мы можем доказать равенство $\lim x^{(k)} = y_p$ тем же методом, что и (9 из микромодуля 36).

Второе утверждение о сходимости следует из теоремы 1, так как $f_p(y_p)^{-} \in \mathcal{F}^{(k)}$, $k \geqslant 0$.

(6): Используем сокращение $\hat{\mathcal{J}}_p := f'_p(y_p)$. Из (3) с помощью (6 из микромодуля 29), (7 из микромодуля 29), (9 из микромодуля 29) следует, что

$$x^{(k+1)} - y_p \subseteq -\gamma^{(k)}\{f_p(m(x^{(k)})) - f_p(y_p) - f'_p(y_p)(m(x^{(k)}) - y_p)\}$$
$$- \gamma^{(k)}(\hat{\mathcal{J}}_p \cdot (m(x^{(k)}) - y_p)) + m(x^{(k)}) - y_p$$
$$= -\gamma^{(k)}\{f_p(m(x^{(k)}) - f_p(y_p)) - f'_p(y_p)(m(x^{(k)}) - y_p)\}$$
$$+ (\hat{\mathcal{J}}_p^{-1} - \gamma^{(k)}) \cdot (\hat{\mathcal{J}}_p \cdot (m(x^{(k)}) - y_p)).$$

Теперь возьмем абсолютные величины (определение 6 из микромодуля 29) и применим (15 из микромодуля 29), (27 из микромодуля 29) и соотношения

$$q(x^{(k)}, y_p) = q(x^{(k)} - y_p, o_p) = |x^{(k)} - y_p|,$$
$$q(\gamma^{(k)}, \mathcal{J}_p^{-1}) = q(\gamma^{(k)} - \hat{\mathcal{J}}_p^{-1}, O_p) = |\gamma^{(k)} - \hat{\mathcal{J}}_p^{-1}|.$$

Теперь, используя монотонную векторную норму, подчиненную ей монотонную матричную норму, теорему об эквивалентности норм и соотношение $m(x^{(k)}) \in x^{(k)}$, получим

$$\|q(x^{(k+1)}, y_p)\| \leqslant \|\,|\gamma^{(k)}|\,\| c_1 \|q(x^{(k)}, y_p)\|^2$$
$$+ \|q(\gamma^{(k)}, \hat{\mathcal{J}}_p^{-1})\| \cdot \|\,|\hat{\mathcal{J}}_p|\,\| \cdot \|q(x^{(k)}, y_p)\|.$$

Из второго равенства в (3) аналогично получим, что

$$\gamma^{(k+1)} - \hat{\mathcal{J}}_p^{-1} \subseteq -((\hat{\mathcal{J}}_p^{-1} - \gamma^{(k)}) \cdot \hat{\mathcal{J}}_p) \cdot (\hat{\mathcal{J}}_p^{-1} - m(\gamma^{(k)}))$$
$$- \gamma^{(k)}((\mathcal{F}^{(k+1)} - \hat{\mathcal{J}}_p) m(\gamma^{(k)})).$$

Снова возьмем абсолютные величины с учетом соотношений $m(\gamma^{(k)}) \in \gamma^{(k)}$, (15 из микромодуля 29), (27 из микромодуля 29) и

$$|\mathcal{F}^{(k+1)} - \hat{\mathcal{J}}_p| = q(\mathcal{F}^{(k+1)}, \hat{\mathcal{J}}_p) \leqslant d(\mathcal{F}^{(k+1)}),$$
$$\|d(\mathcal{F}^{(k+1)})\|' \leqslant c\,\|d(x^{(k+1)})\|'.$$

Используя монотонную матричную норму и теорему об эквивалентности норм, получим, что

$$\|q(\gamma^{(k+1)}, \hat{\mathcal{J}}_p^{-1})\| \leqslant c_3 \|\,|\hat{\mathcal{J}}_p|\,\| \cdot \|q(\gamma^{(k)}, \hat{\mathcal{J}}_p^{-1})\|^2$$
$$+ \|\,|\gamma^{(k)}|\,\|^2 c_4 \|q(x^{(k+1)}, y_p)\|.$$

Для $\mathcal{A}_p = (\mathcal{C}_p, \mathcal{B}_p) \in M_{n,\,n+1}(\mathbb{R})$ введем норму $\|\mathcal{A}_p\| = \max\{\|\mathcal{C}_p\|, \|\mathcal{B}_p\|\}$. Полагая $r^{(k)} = \|(q(x^{(k)}, y_p), q(\gamma^{(k)}, \hat{\mathcal{J}}_p^{-1}))\|$, замечая, что $\lim\limits_{k\to\infty} \gamma^{(k)} = \mathcal{J}_p^{-1}$, $\|\,|\gamma^{(k)}|\,\| \leqslant s$ и полагая $\|\,|\hat{\mathcal{J}}_p|\,\| = p$, получаем из установленных выше неравенств, что

$$\|q(x^{(k+1)}, y_p)\| \leqslant (sc_1 + pc_2)(r^{(k)})^2,$$
$$\|q(\gamma^{(k+1)}, \hat{\mathcal{J}}_p^{-1})\| \leqslant (c_3 p + s^2 c_4(sc_1 + pc_2))(r^{(k)})^2.$$

Если мы теперь положим

$$\gamma = \max\{sc_1 + pc_2,\ c_3 p + s^2 c_4 (sc_1 + pc_2)\}.$$

то получим, что

$$r^{(k+1)} \leqslant \gamma (r^{(k)})^2, \qquad k \geqslant 0.$$

Теперь из теоремы 2 приложения А следует, что мы имеем не менее чем квадратичную сходимость последовательности

$$\{(x^{(k)},\ y^{(k)})\}_{k=0}^{\infty} \quad \text{к} \quad \left(y_p,\ f'_p(y_p)^{-1}\right).$$

Микромодуль 39
Методы Ньютоновского типа для частных типов систем несинейных уравнений

Рассмотрим теперь ту же постановку задачи, что в двух предшествующих микромодулях, введя дополнительные предположения о рассматриваемых системах

$$f_p(x_p) = o_p.$$

Важно отметить, что не делается никаких предположений о выпуклости f_p. Производная Фреше функции f_p в точке x_p обозначается через $f'_p(x_p) = (\partial f_i / \partial x_j)$. Используя естественный частичный порядок на множестве $V_n(\mathbb{R})$ и на множестве $M_{nn}(\mathbb{R})$, состоящем из всехвещественных точечных матриц размерности $n \times n$. сначала докажем общую теорему, а затем покажем, как ее можно применить к конкретным системам.

Теорема 1. *Пусть отображение* f_n: $\mathfrak{D} \subseteq V_n(\mathbb{R}) \to V_n(\mathbb{R})$. *имеет производную Фреше, а множество* $\{x_p \mid x_p^{(0)} \leqslant x_p \leqslant y_p^{(0)}\}$ *содержится в* $\mathfrak{D}$. *Пусть выполнено*

$$f_p(x_p^{(0)}) \leqslant o_p \leqslant f_p(y_p^{(0)}). \tag{0}$$

Задано отображение

$$\mathcal{A}_p\colon \{x_p \mid x_p^{(0)} \leqslant x_p \leqslant y_p^{(0)}\} \times \{x_p \mid x_p^{(0)} \leqslant x_p \leqslant y_p^{(0)}\} \to M_{nn}(\mathbb{R}),$$
$$\mathcal{A}_p = \mathcal{A}_p(x_p,\ y_p),$$

для которого верно

$$f_p(x_p) - f_p(y_p) = \mathcal{A}_p(x_p,\ y_p)(x_p - y_p) \text{ для } x_p^{(0)} \leqslant x_p \leqslant y_p \leqslant y_p^{(0)}. \tag{1}$$

Задано непрерывное отображение

$$\mathscr{B}_p\colon \{x_p \mid x_p^{(0)} \leqslant x_p \leqslant y_p^{(0)}\} \times \{x_p \mid x_p^{(0)} \leqslant x_p \leqslant y_p^{(0)}\} \to M_{nn}(\mathbb{R}), \quad \mathscr{B}_p = \mathscr{B}_p(x_p, y_p),$$

для которого верно

$$\mathscr{B}_p(x_p, y_p) \leqslant \mathscr{B}_p(x_p, y_p) \quad \text{для} \tag{2}$$

$$x_p^{(0)} \leqslant x_p \leqslant \bar{x}_p \leqslant \bar{y}_p \leqslant y_p \leqslant y_p^{(0)},$$

$$\mathscr{A}_p(x_p, y_p) \leqslant \mathscr{B}_p(x_p, y_p) \quad \text{для} \quad x_p^{(0)} \leqslant x_p \leqslant y_p \leqslant y^{(0)}, \tag{3}$$

существует $\mathscr{B}_p(x_p, y_p)^{-1}$ *и* $\mathscr{B}_p(x_p, y_p)^{-1} \geqslant \mathscr{O}_p$ *для* (4)

$$x_p^{(0)} \leqslant x_p \leqslant y_p \leqslant y_p^{(0)}.$$

Матрица $\mathscr{P}_p^{(0)} \in M_{nn}(\mathbb{R})$ *невырождена, и имеет место*

$$\mathscr{B}_p(x_p^{(0)}, y_p^{(0)})\,\mathscr{P}_p^{(0)} \leqslant \mathscr{I}_p \quad (\mathscr{I}_p - \text{единичная матрица}), \tag{5}$$

$$\mathscr{P}_p^{(0)}\mathscr{B}_p(x_p^{(0)}, y_p^{(0)}) \leqslant \mathscr{I}_p, \tag{6}$$

$$\mathscr{P}_p^{(0)} \geqslant \mathscr{O}_p. \tag{7}$$

Тогда итерация

$$\begin{cases} y_p^{(k+1)} = y_p^{(k)} - \mathscr{P}_p^{(k)} f_p(y_p^{(k)}), \\ x_p^{(k+1)} = x_p^{(k)} - \mathscr{P}_p^{(k)} f_p(x_p^{(k)}), \\ \mathscr{P}_p^{(k+1)} = \mathscr{P}_p^{(k)} - \mathscr{P}_p^{(k)}\left(\mathscr{B}_p(x_p^{(k+1)}, y_p^{(k+1)})\,\mathscr{P}_p^{(k)} - \mathscr{I}_p\right), \quad k \geqslant 0, \end{cases} \tag{8}$$

может быть выполнена без ограничений и имеет место

$$x_p^{(0)} \leqslant x_p^{(1)} \leqslant \ldots \leqslant x_p^{(k)} \leqslant x_p^{(k+1)} \leqslant y_p^{(k+1)} \leqslant y_p^{(k)} \leqslant \ldots \leqslant y_p^{(1)} \leqslant y_p^{(0)}, \tag{9}$$

$$f_p(x_p^{(k)}) \leqslant o_p \leqslant f_p(y_p^{(k)}), \tag{10}$$

$$\mathscr{P}_p^{(0)} \leqslant \mathscr{P}_p^{(1)} \leqslant \ldots \leqslant \mathscr{P}_p^{(k)} \leqslant \mathscr{P}_p^{(k+1)} \leqslant \ldots, \tag{11}$$

$$\mathscr{P}_p^{(k)} \text{ невырождена}, \tag{12}$$

$$\mathscr{P}_p^{(k)}\mathscr{B}_p(x_p^{(k)}, y_p^{(k)}) \leqslant \mathscr{I}_p, \tag{13}$$

$$\mathscr{B}_p(x_p^{(k)}, y_p^{(k)})\,\mathscr{P}_p^{(k)} \leqslant \mathscr{I}_p, \tag{14}$$

$$\lim_{k\to\infty} x_p^{(k)} = x_p^* \leqslant y_p^* = \lim_{k\to\infty} y_p^{(k)}, \tag{15}$$

$$\mathscr{P}_p^* := \lim_{k\to\infty} \mathscr{P}_p^{(k)} = \mathscr{B}_p(x_p^*, y_p^*)^{-1}. \tag{16}$$

Если f_p *непрерывна, то* $f_p(x_p^*) = f_p(y_p^*) = o_p$. *Все решения* x_p^* *системы* $f_p(x_p) = o_p$, *принадлежащие множеству*

$$\{x_p \mid x_p^{(0)} \leqslant x_p \leqslant y_p^{(0)}\}$$

удовлетворяют условию $x_p^* \leqslant z_p^* \leqslant y_p^*$.

Если матрица $\mathscr{A}_p(x_p, y_p)$ *невырожденная при*

$$x_p^{(0)} \leqslant x_p \leqslant y_p \leqslant y_p^{(0)}, \text{ то } x_p^* = y_p^*.$$

Если сверх того производная Фреше $f'_p(x_p)$ *удовлетворяет условию Липшица вида*

$$\|f'_p(x_p) - f'_p(y_p)\| \leqslant c_1 \|x_p - y_p\| \quad \text{для} \quad x_p^{(0)} \leqslant x_p, \quad y_p \leqslant y_p^{(0)} \tag{17}$$

и отображение $\mathscr{B}_p(x_p, y_p)$ *удовлетворяет условию*

$$\|f'_p(z_p) - \mathscr{B}_p(x_p, y_p)\| \leqslant c_2 \|x_p - y_p\| \quad \text{для} \tag{18}$$
$$x_p^{(0)} \leqslant x_p \leqslant z_p \leqslant y_p \leqslant y_p^{(0)},$$

то

$$\lim_{k \to \infty} \mathscr{P}_p^{(k)} = f'_p(x_p^*)^{-1} \tag{19}$$

и последовательность

$$\{([x_i^{(k)}, y_i^{(k)}]), ([p_{ij}^{(k)}, p_{ij}^{(k)}])\}_{k=0}^{\infty} \in M_{n,\, n+1}(I(\mathbb{R}))$$

сходится к $(x_p^*, \mathscr{P}_p^*)$, *причем R-порядок сходимости не меньше* 2.

Доказательство. (9)—(14): Доказываются методом математической индукции.

Пусть $k \geqslant 0$ и $x_p^{(k)} \leqslant x_p \leqslant y_p^{(k)}$.

Из (2) и (3) следует, что

$$\mathscr{A}_p(x_p, y_p^{(k)}) \leqslant \mathscr{B}_p(x_p, y_p^{(k)}) \leqslant \mathscr{B}_p(x_p^{(k)}, y_p^{(k)}).$$

Ввиду $\mathscr{P}_p^{(k)} \geqslant O_p$ из (11) следует, что

$$\mathscr{P}_p^{(k)} \mathscr{B}_p(x_p^{(k)}, y_p^{(k)}) \geqslant \mathscr{P}_p^{(k)} \mathscr{A}_p(x_p, y_p^{(k)}),$$

а потому с помощью (13) мы получаем

$$\mathscr{I}_p - \mathscr{P}_p^{(k)} \mathscr{A}_p(x_p, y_p^{(k)}) \geqslant \mathscr{I}_p - \mathscr{P}_p^{(k)} \mathscr{B}_p(x_p^{(k)}, y_p^{(k)}) \geqslant O_p.$$

Отсюда, используя (1) и тот факт, что $x_p - y_p^{(k)} \leqslant o_p$, получаем неравенства

$$\begin{aligned} x_p - \mathscr{P}_p^{(k)} f_p(x_p) &= y_p^{(k+1)} + (x_p - y_p^{(k)}) - \mathscr{P}_p^{(k)}(f_p(x_p) - f_p(y_p^{(k)})) \\ &= y_p^{(k+1)} + (x_p - y_p^{(k)}) - \mathscr{P}_p^{(k)} \mathscr{A}_p(x_p, y_p^{(k)})(x_p - y_p^{(k)}) \\ &= y_p^{(k+1)} + (\mathscr{I}_p - \mathscr{P}_p^{(k)} \mathscr{A}_p(x_p, y_p^{(k)}))(x_p - y_p^{(k)}) \leqslant y_p^{(k+1)}. \end{aligned} \tag{20}$$

Полагая $x_p = x_p^{(k)}$, используя (10) и тот факт, что $\mathcal{P}_p^{(k)} \geqslant O_p$, получаем неравенство

$$x_p^{(k)} \leqslant x_p^{(k)} - \mathcal{P}_p^{(k)} f_p\left(x_p^{(k)}\right) = x_p^{(k+1)} \leqslant y_p^{(k+1)} = y_p^{(k)} - \mathcal{P}_p^{(k)} f_p\left(y_p^{(k)}\right),$$

которое доказывает (9).

Из (2), (3) и того факта, что $x_p^{(k)} \leqslant y_k^{(k+1)} \leqslant y_p^{(k)}$, следует неравенство

$$\mathcal{A}_p\left(y_p^{(k+1)}, y_p^{(k)}\right) \leqslant \mathcal{B}_p\left(y_p^{(k+1)}, y_p^{(k)}\right) \leqslant \mathcal{B}_p\left(x_p^{(k)}, y_p^{(k)}\right).$$

Используя (14) и неравенство $\mathcal{P}_p^{(k)} \geqslant O_p$, получаем, что

$$\mathcal{I}_p - \mathcal{A}_p\left(y_p^{(k+1)}, y_p^{(k)}\right) \mathcal{P}_p^{(k)} \geqslant \mathcal{I}_p - \mathcal{B}_p\left(x_p^{(k)}, y_p^{(k)}\right) \mathcal{P}_p^{(k)} \geqslant O_p.$$

Отсюда с помощью (1) и неравенства

$$f_p\left(y_p^{(k)}\right) \geqslant o_p$$

следует, что

$$\begin{aligned} f_p\left(y_p^{(k+1)}\right) &= f_p\left(y_p^{(k)}\right) + \mathcal{A}_p\left(y_p^{(p+1)}, y_p^{(k)}\right)\left(y_p^{(k+1)} - y_p^{(k)}\right) \\ &= \left(\mathcal{I}_p - \mathcal{A}_p\left(y_p^{(k+1)}, y_p^{(k)}\right) \mathcal{P}_p^{(k)}\right) f_p\left(y_p^{(k)}\right) \geqslant o_p. \end{aligned}$$

Аналогичным образом показываем, что $f_p\left(x_p^{(k+1)}\right) \leqslant o_p$, откуда и следует (10).

Из

$$x_p^{(k)} \leqslant x_p^{(k+1)} \leqslant y_p^{(k+1)} \leqslant y_{p+1}^{(k)}$$

следует с помощью (2), что

$$\mathcal{B}_p\left(x_p^{(k+1)}, y_p^{(k+1)}\right) \leqslant \mathcal{B}_p\left(x_p^{(k)}, y_p^{(k)}\right).$$

С помощью (13) и (14) получаем в силу $\mathcal{P}_p^{(k)} \geqslant O_p$, что

$$\begin{cases} \mathcal{B}_p\left(x_p^{(k+1)}, y_p^{(k+1)}\right) \mathcal{P}_p^{(k)} \leqslant \mathcal{B}_p\left(x_p^{(k)}, y_p^{(k)}\right) \mathcal{P}_p^{(k)} \leqslant \mathcal{I}_p, \\ \mathcal{P}_p^{(k)} \mathcal{B}_p\left(x_p^{(k+1)}, y_p^{(k+1)}\right) \leqslant \mathcal{P}_p^{(k)} \mathcal{B}_p\left(x_p^{(k)}, y_p^{(k)}\right) \leqslant \mathcal{I}_p. \end{cases} \tag{21}$$

С помощью (8) получаем, что верны равенства

$$\mathcal{I}_p - \mathcal{B}_p\left(x_p^{(k+1)}, y_p^{(k+1)}\right) \mathcal{P}_p^{(k+1)} = \left(\mathcal{I}_p - \mathcal{B}_p\left(x_p^{(k+1)}, y_p^{(k+1)}\right) \mathcal{P}_p^{(k)}\right)^2 \geqslant O_p$$

и

$$\mathcal{I}_p - \mathcal{P}_p^{(k+1)} \mathcal{B}_p\left(x_p^{(k+1)}, y_k^{(k+1)}\right) = \left(\mathcal{I}_p - \mathcal{P}_p^{(k)} \mathcal{B}_p\left(x_p^{(k+1)}, y_p^{(k+1)}\right)\right)^2 \geqslant O_p, \tag{22}$$

а также

$$\mathscr{P}_{\rm p}^{(k+1)} = \mathscr{P}_{\rm p}^{(k)} - \mathscr{P}_{\rm p}^{(k)}\left(\mathscr{B}_{\rm p}\left(x_{\rm p}^{(k+1)},\ y_{\rm p}^{(k+1)}\right)\mathscr{P}_{\rm p}^{(k)} - \mathscr{I}_{\rm p}\right) \geqslant \mathscr{P}_{\rm p}^{(k)} \geqslant \mathscr{O}_{\rm p}.$$

Отсюда получаются (11), (13) и (14). Чтобы доказать (12), мы должны показать, что матрица $\mathscr{P}_{\rm p}^{(k+1)}$ неособенная. Так как $\mathscr{P}_{\rm p}^{(k)}$ неособенная, мы имеем

$$\mathscr{B}_{\rm p}\left(x_{\rm p}^{(k+1)},\ y_{\rm p}^{(k+1)}\right) = \mathscr{M}_{\rm p} - \mathscr{N}_{\rm p},$$

где $\mathscr{M}_{\rm p} = \left(\mathscr{P}^{(k)}\right)^{-1}$, $\mathscr{N}_{\rm p} = \left(\mathscr{P}^{(k)}\right)^{-1} - \mathscr{B}_{\rm p}\left(x_{\rm p}^{(k+1)},\ y_{\rm p}^{(k+1)}\right)$.

Условие теоремы показывает, что верно

$$\mathscr{M}_{\rm p}^{-1} = \mathscr{P}_{\rm p}^{(k)} \geqslant \mathscr{O}_{\rm p},$$

откуда с помощью (21) следует, что

$$\mathscr{M}_{\rm p}^{-1}\mathscr{N}_{\rm p} = \mathscr{I}_{\rm p} - \mathscr{P}_{\rm p}^{(k)}\mathscr{B}_{\rm p}\left(x_{\rm p}^{(k+1)},\ y_{\rm p}^{(k+1)}\right) \geqslant \mathscr{O}_{\rm p},$$

$$\mathscr{N}_{\rm p}\mathscr{M}_{\rm p}^{-1} = \mathscr{I}_{\rm p} - \mathscr{B}_{\rm p}\left(x_{\rm p}^{(k+1)},\ y_{\rm p}^{(k+1)}\right)\mathscr{P}_{\rm p}^{(k)} \geqslant \mathscr{O}_{\rm p},$$

т. е. $\mathscr{M}_{\rm p} - \mathscr{N}_{\rm p}$ является слабо регулярным разбиением матрицы $\mathscr{B}_{\rm p}\left(x_{\rm p}^{(k+1)},\ y_{\rm p}^{(k+1)}\right)$.

В силу (4) имеем $\mathscr{B}_{\rm p}\left(x_{\rm p}^{(k+1)},\ y_{\rm p}^{(k+1)}\right)^{-1} \geqslant \mathscr{O}_{\rm p}$. Из известной теоремы следует неравенство

$$\rho\left(\mathscr{I}_{\rm p} - \mathscr{P}_{\rm p}^{(k)}\mathscr{B}_{\rm p}\left(x_{\rm p}^{(k+1)},\ y_{\rm p}^{(k+1)}\right)\right) < 1,$$

где ρ обозначает спектральный радиус. Это означает, что матрица

$$\mathscr{I}_{\rm p} - \left(\mathscr{I}_{\rm p} - \mathscr{P}_{\rm p}^{(k)}\mathscr{B}_{\rm p}\left(x_{\rm p}^{(k+1)},\ y_{\rm p}^{(k+1)}\right)\right)^2$$

обратима. В силу (22) отсюда следует, что $\mathscr{P}_{\rm p}^{(k+1)}$ обратима. Соотношение (15) выводится из (9) с помощью стандартных рассуждений.

Из (13) или (14) следует с помощью (4), что

$$\mathscr{P}_{\rm p}^{(k)} \leqslant \mathscr{B}_{\rm p}\left(x_{\rm p}^{(k)},\ y_{\rm p}^{(k)}\right)^{-1}.$$

Из непрерывности отображения $\mathscr{B}_{\rm p}(x_{\rm p},\ y_{\rm p})$ следует, что

$$\lim_{k\to\infty} \mathscr{B}_{\rm p}\left(x_{\rm p}^{(k)},\ y_{\rm p}^{(k)}\right) = \mathscr{B}_{\rm p}\left(x_{\rm p}^{*},\ y_{\rm p}^{*}\right),$$

и ввиду $x_{\rm p}^{(k)} \leqslant x_{\rm p}^{*} \leqslant y_{\rm p}^{*} \leqslant y_{\rm p}^{(k)}$ из (2) следует, что

$$\mathscr{B}_{\rm p}\left(x_{\rm p}^{*},\ y_{\rm p}^{*}\right) \leqslant \mathscr{B}_{\rm p}\left(x_{\rm p}^{(k)},\ y_{\rm p}^{(k)}\right).$$

Отсюда с помощью (4) следует, что

$$\mathscr{B}_{\rm p}\left(x_{\rm p}^{(k)},\ y_{\rm p}^{(k)}\right)^{-1} \leqslant \mathscr{B}_{\rm p}\left(x_{\rm p}^{*},\ y_{\rm p}^{*}\right)^{-1},$$

т. е. верно

$$\mathscr{P}_p^{(k)} \leqslant \mathscr{B}_p\left(x_p^*,\ y_p^*\right)^{-1}.$$

Поэтому последовательность (11) ограничена сверху, а значит она сходится

Из формул (8), описывающих итерацию, следует с помощью (21) и (11), что

$$\mathscr{P}_p^{(k)} - \mathscr{P}_p^{(k+1)} = \mathscr{P}_p^{(k)}\left(\mathscr{B}_p\left(x_p^{(k+1)},\ y_p^{(k+1)}\right)\mathscr{P}_p^{(k)} - \mathscr{I}_p\right)$$
$$\leqslant \mathscr{P}_p^{(0)}\left(\mathscr{B}_p\left(x_p^{(k+1)},\ y_p^{(k+1)}\right)\mathscr{P}_p^{(k)} - \mathscr{I}_p\right) \leqslant \mathscr{O}_p,$$

откуда получается, что

$$\mathscr{O}_p = \lim_{k\to\infty}\left(\mathscr{P}_p^{(k)} - \mathscr{P}_p^{(k+1)}\right) \leqslant \lim_{k\to\infty}\mathscr{P}_p^{(0)}\left[\mathscr{B}_p\left(x_p^{(k+1)},\ y_p^{(k+1)}\right)\mathscr{P}_p^{(k)} - \mathscr{I}_p\right]$$
$$= \mathscr{P}_p^{(0)}\left(\mathscr{B}_p\left(x_p^*,\ y_p^*\right)\mathscr{P}_p^* - \mathscr{I}_p\right) \leqslant \mathscr{O}_p.$$

Так как матрица $\mathscr{P}_p^{(0)}$ невырожденная, мы обосновали (16). Пусть для некоторого $k \geq 0$ верно $x_p^{(k)} \leqslant z_p \leqslant y_p^{(k)}$, где $\mathscr{f}_p(z_p) = o_p$. Тогдамы получаем из (20), что

$$z_p = z_p - \mathscr{P}_p^{(k)}\mathscr{f}_p\left(z_p\right) \leqslant y_p^{(k+1)}.$$

Из соотношения

$$x_p - \mathscr{P}_p^{(k)}\mathscr{f}_p\left(x_p\right) \geqslant x_p^{(k+1)},$$

которое справедливо для $x_p^{(k)} \leqslant x_p \leqslant y_p^{(k)}$ и может быть доказано аналогично (20), получаем, полагая $x_p := z_p$, что

$$z_p = z_p - \mathscr{P}_p^{(k)}\mathscr{f}_p\left(z_p\right) \geqslant x_p^{(k+1)},$$

т. е.

$$x_p^{(k)} \leqslant x_p^{(k+1)} \leqslant z_p \leqslant y_p^{(k+1)} \leqslant y_p^{(k)}$$

для всех $k \geq 0$. Переходя в этих неравенствах к пределу, мы видим, что решения уравнения $\mathscr{f}_p\left(x_p\right) = o_p$, локализованные между $x_p^{(0)}$ и $y_p^{(0)}$, локализованы также и между x_p^* и y_p^*.

Из формул (8), описывающих итерацию, следует, что для непрерывной $\mathscr{f}_p$ имеет место

$$o_p = \lim_{k\to\infty}\left(y_p^{(k)} - y_p^{(k+1)}\right) = \lim_{k\to\infty}\mathscr{P}_p^{(k)}\mathscr{f}_p\left(y_p^{(k)}\right) = \mathscr{P}_p^*\mathscr{f}_p\left(y_p^*\right).$$

Из того что

$$\mathscr{P}_p^* = \mathscr{B}_p\left(x_p^*,\ y_p\right)^{-1}$$

невырожденная, следует, что $f_p(y_p^*) = o_p$. Аналогично устанавливается, что $f_p(x_p^*) = o_p$.

Если матрица $A_p(x_p, y_p)$ для $x_p^{(0)} \leqslant x_p \leqslant y_p \leqslant y_p^{(0)}$ невырожденная, то из (1) следует, что

$$o_p = f_p(x_p^*) - f_p(y_p^*) = A_p(x_p^*, y_p^*)(x_p^* - y_p^*),$$

т. е. $x_p^* = y_p^*$.

Если предположим теперь, что выполнено (18), то $\lim\limits_{k\to\infty} x_p^{(k)} = x_p^{(k)} = y_p^* = \lim\limits_{k\to\infty} y_p^{(k)}$ влечет за собой

$$\lim_{k\to\infty} B_p(x_p^{(k)}, y_p^{(k)}) = f_p'(x_p^*)$$

С помощью (16) мы устанавливаем (19).

Введем сокращения

$$B_p^{(k)} := B_p(x_p^{(k)}, y_p^{(k)}), \quad p := \|f_p'(x_p^*)\|, \quad P_p^* = f_p'(x_p^*)^{-1}.$$

Ввиду $\lim\limits_{k\to\infty} P_p^{(k)} = P_p^* = f_p'(x_p^*)^{-1}$ существует константа s, такая что верно $\|P_p^{(k)}\| \leqslant s$, $k \geqslant 0$.

Из формул (8) следует, с помощью (17), что

$$\begin{aligned}\|y_p^{(k+1)} - x_p^*\| &= \|y_p^{(k)} - x_p^* - P_p^{(k)} f_p(y_p^{(k)})\| \\ &= \|-P_p^{(k)}(f_p(y_p^{(k)}) - f_p(x_p^*) - f_p'(x_p^*)(y_p^{(k)} - x_p^*)) \\ &\quad + (f_p'(x_p^*)^{-1} - P_p^{(k)}) f_p'(x_p^*)(y_p^{(k)} - x_p^*)\| \\ &\leqslant (c_1/2)\|P_p^{(k)}\|\|y_p^{(k)} - x_p^*\|^2 + p\|P_p^* - P_p^{(k)}\|\|y_p^{(k)} - x_p^*\|.\end{aligned} \tag{23}$$

Аналогично устанавливается, что

$$\begin{aligned}\|x_p^{(k+1)} - x_p^*\| &\leqslant (c_1/2)\|P_p^{(k)}\|\|x_p^{(k)} - x_p^*\|^2 \\ &\quad + p\|P_p^* - P_p^{(k)}\|\|x_p^{(k)} - x_p^*\|.\end{aligned} \tag{24}$$

Используя монотонную норму и теорему об эквивалентности норм, мы можем записать (23) и (24) в виде

$$\|y_p^{(k+1)} - x_p^*\| \leqslant \gamma_1\|y_p^{(k)} - x_p^{(k)}\|^2 + \gamma_2\|P_p^* - P_p^{(k)}\|\|y_p^{(k)} - x_p^{(k)}\|, \tag{23'}$$

$$\|x_p^{(k+1)} - x_p^*\| \leqslant \gamma_3\|y_p^{(k)} - x_p^{(k)}\|^2 + \gamma_4\|P_p^* - P_p^*\|\|y_p^{(k)} - x_p^{(k)}\|. \tag{24'}$$

Из (8) с помощью равенства $J_p = f_p'(x_p^*) P_p^*$ получаем

$$\begin{aligned}P_p^{(k+1)} - P_p^* &= P_p^{(k)} - P_p^* - P_p^{(k)}(B_p^{(k+1)} P_p^{(k)} - J_p) \\ &= (P_p^* - P_p^{(k)}) f_p'(x_p^*)(P_p^{(k)} - P_p^*) + P_p^{(k)}(f_p'(x_p^*) - B_p^{(k+1)}) P_p^{(k)}.\end{aligned}$$

С помощью (18) получаем теперь

$$\left\|\mathscr{P}_{\mathrm{p}}^{(k+1)}-\mathscr{P}_{\mathrm{p}}^{*}\right\| \leqslant p\left\|\mathscr{P}_{\mathrm{p}}^{(k)}-\mathscr{P}_{\mathrm{p}}^{*}\right\|^{2}+c_{2}\left\|\mathscr{P}_{\mathrm{p}}^{(k)}\right\|^{2}\left\|x_{\mathrm{p}}^{(k+1)}-y_{\mathrm{p}}^{(k+1)}\right\| \quad (25)$$

$$\leqslant \gamma_{5}\left\|\mathscr{P}_{\mathrm{p}}^{(k)}-\mathscr{P}_{\mathrm{p}}^{*}\right\|^{2}+\gamma_{6}\left\|y_{\mathrm{p}}^{(k+1)}-x_{\mathrm{p}}^{(k+1)}\right\|.$$

Введем норму для матриц $\mathscr{A}_{\mathrm{p}}=(\mathscr{C}_{\mathrm{p}}, \mathscr{B}_{\mathrm{p}}) \in M_{n, n+1}(\mathbb{R})$ с помощью равенства

$$\|\mathscr{A}_{\mathrm{p}}\|=\max\left\{\|\mathscr{C}_{\mathrm{p}}\|, \|\mathscr{B}_{\mathrm{p}}\|\right\}$$

и положим

$$d^{(k)}=\left\|\left(y_{\mathrm{p}}^{(k)}-x_{\mathrm{p}}^{(k)}, \mathscr{P}_{\mathrm{p}}^{*}-\mathscr{P}_{\mathrm{p}}^{(k)}\right)\right\|.$$

Из (23'), (24') и (25) получаем теперь

$$\left\|y_{\mathrm{p}}^{(k+1)}-x_{\mathrm{p}}^{(n+1)}\right\| \leqslant\left(\sum_{i=1}^{4} \gamma_{i}\right)\left(d^{(k)}\right)^{2},$$

$$\left\|\mathscr{P}_{\mathrm{p}}^{*}-\mathscr{P}_{\mathrm{p}}^{(k+1)}\right\| \leqslant\left(\gamma_{5}+\gamma_{6}\left(\sum_{i=1}^{4} \gamma_{i}\right)\right)\left(d^{(k)}\right)^{2},$$

т. е.

$$d^{(k+1)} \leqslant \gamma\left(d^{(k)}\right)^{2},$$

где

$$\gamma=\max\left\{\sum_{i=1}^{4} \gamma_{i}, \gamma_{5}+\gamma_{6} \sum_{i=1}^{4} \gamma_{i}\right\}.$$

Теперь получаем нашу теорему, сославшись на теорему 2 из приложения А. I

Содержание теоремы 1 легко перенести на методы с более высокой скоростью сходимости. Для этого мы сохраняем матрицу $\mathscr{P}_{\mathrm{p}}^{(k)}$ неизменной в течение нескольких шагов. Приближенное обращение матрицы $\mathscr{P}_{\mathrm{p}}\left(x_{\mathrm{p}}^{(k+1)}, y_{\mathrm{p}}^{(k+1)}\right)$ делается с помощью метода высокого порядка. В результате это дает следующую итерацию:

$$\begin{cases} y_p^{(k,0)} = y_p^{(k)}, \\ y_p^{(k,r+1)} = y_p^{(k,r)} - \mathscr{P}_p^{(k)} \mathscr{f}_p\left(y_p^{(k,r)}\right), \quad 0 \leqslant r \leqslant m, \\ y_p^{(k+1)} = y_p^{(k,m+1)}, \\ x_p^{(k,0)} = x_p^{(k)}, \\ x_p^{(k,r+1)} = x_p^{(k,r)} - \mathscr{P}_p^{(k)} \mathscr{f}_p\left(x_k^{(k,r)}\right), \quad 0 \leqslant r \leqslant m, \\ x_p^{(k+1)} = x_p^{(k,m+1)}, \\ \mathscr{P}_p^{(k+1)} = \mathscr{P}_p^{(k)} + \mathscr{P}_p^{(k)} \sum\limits_{\mu=1}^{m+1} (-1)^{\mu} \left[\mathscr{B}_p\left(x_p^{(k+1)}, y_p^{(k+1)}\right) \mathscr{P}_p^{(k)} - \mathscr{I}_p\right]^{\mu}, \\ k \geqslant 0. \end{cases}$$

Этот метод сходится в условиях теоремы 1, и порядок сходимости равен $m + 2$.

Следующее утверждение содержит частный случай теоремы 1, который получается, когда обращение матрицы $\mathscr{B}_p\left(x_p^{(k)}, y_p^{(k)}\right)$ производится точно, а не приближенно.

Теорема 2. *Пусть в теореме* 1

$$\mathscr{P}_p^{(k)} = \mathscr{B}_p\left(x_p^{(k)}, y_p^{(k)}\right)^{-1},$$

а итерация (8) *заменена на*

$$\begin{cases} y_p^{(k+1)} = y_p^{(k)} - \mathscr{P}_p^{(k)} \mathscr{f}_p\left(y_p^{(k)}\right), \\ x_p^{(k+1)} = x_p^{(k)} - \mathscr{P}_p^{(k)} \mathscr{f}_p\left(x_p^{(k)}\right), \quad k \geqslant 0. \end{cases} \tag{8'}$$

Тогда верны все утверждения теоремы 1.

Доказательство. (5), (6), (7), а также (12), (13) и (14) выполнены тривиальным образом ввиду (4). Неравенство (11) также следует из (4). Остальную часть доказательства можно повторить без изменений.

Следует заметить, что при численном исполнении метода (8') треугольное разложение матрицы

$$\mathscr{B}_p\left(x_p^{(k)}, y_p^{(k)}\right)$$

по методу Гаусса следует производить только один раз.

Рассмотрим теперь класс систем нелинейных уравнений, к которому можно применить предыдущие теоремы.

Пусть

$$\mathscr{f}_p\colon \mathscr{D}_p \subseteq V_n(\mathbb{R}) \to V_n(\mathbb{R}), \quad \mathscr{f}_p(x_p) = \mathscr{H}_p x_p + \varepsilon \mathscr{h}_p(x_p) + \mathscr{c}_p, \tag{26}$$

где $\varepsilon \geqslant 0$, $\mathscr{c}_p \in V_n(\mathbb{R})$ и $\mathscr{H}_p$ является M-матрицей, т. е. $h_{ij} \leq 0$ для $i \neq j$ и $\mathscr{H}_p^{-1} \geqslant \mathscr{O}_p$. Отображение

$$\mathcal{h}_p\colon \mathfrak{D} \subseteq V_n(\mathbb{R}) \to V_n(\mathbb{R}), \quad \mathcal{h}_p(x_p) = (\mathcal{h}_i(x_p))$$

непрерывно дифференцируемо и

$$h_i(x_p) = \sum_{j=1}^{n} a_{ij} g_j(x_j), \quad 1 \leqslant i \leqslant n,$$

где $g_i\colon \mathfrak{D}_i \subseteq \mathbb{R} \to \mathbb{R}$, $g_i = g_i(s)$, $1 \leqslant i \leqslant n$,

$$\frac{d}{ds} g_i(s) \geqslant 0, \quad 1 \leqslant i \leqslant n,$$

и выполнено

$$a_{ij} \geqslant 0, \quad 1 \leqslant i, \ j \leqslant n$$

Полагая

$$g_p(x_p) = (g_i(x_i)) \quad \text{и} \quad \mathcal{U}_p = (a_{ij}),$$

можем записать $\mathcal{f}_p$ в виде

$$\mathcal{f}_p(x_p) = \mathcal{H}_p x_p + \varepsilon \mathcal{U}_p g_p(x_p) + \mathcal{b}_p.$$

Пусть $\mathcal{A}_p(x_p, y_p)$ — матрица, i-я строка которой равна

$$f_i'(x_p + \theta_i(y_p - x_p)) = \Big(h_{i1} + \varepsilon a_{i1} \frac{d}{dx_1} g_1(x_1 + \theta_i(y_1 - x_1)), \ \ldots,$$
$$h_{in} + \varepsilon a_{in} \frac{d}{dx_n} g_n(x_n + \theta_i(y_n - x_n))\Big)$$

для $\theta_i \in (0, 1)$, $1 \leqslant i \leqslant n$. Если $\mathfrak{D}$ выпукла, то из теоремы о среднем значении для отображений, определенных на $\mathfrak{D} \subseteq V_n(\mathbb{R}) \to \mathbb{R}$, следует, что

$$\mathcal{f}_p(x_p) - \mathcal{f}_p(y_p) = \mathcal{A}_p(x_p, y_p)(x_p - y_p)$$

для $x_p, y_p \in \mathfrak{D}$.

Пусть $v_i(x_i, y_i)$ — верхняя граница в интервальном вычислении производной dg_i/dx_i для интервала $[x_i, y_i]$ с $x_i \leqslant y_i$.

Из монотонности включения следует, что

$$\frac{d}{dx_i} g_i(\bar{x}_i + \bar{\theta}_i(\bar{y}_i - \bar{x}_i)) \in \frac{d}{dx_i} g_i([\bar{x}_i, \bar{y}_i]) \subseteq \frac{d}{dx_i} g_i([x_i, y_i])$$

для $[\bar{x}_i, \bar{y}_i] \subseteq [x_i, y_i]$, т. е.

$$v_i(x_i, y_i) \geqslant \frac{d}{dx_i} g_i(\bar{x}_i + \bar{\theta}_i(\bar{y}_i - \bar{x}_i)) \geqslant 0.$$

Если теперь положим

$$\overline{\mathcal{U}}_p = (\bar{u}_{ij}), \text{ где } \bar{u}_{ij} = a_{ij} v_i(x_i, y_i), \quad 1 \leqslant i, \ j \leqslant n,$$

то в обозначениях получим,

$$\mathscr{B}_p(x_p, y_p) = (h_{ij} + \varepsilon a_{ij} v_j(x_j, y_j))$$

что

$$\mathscr{B}_p(\bar{x}_p, \bar{y}_p) \leqslant \mathscr{B}_p(x_p, y_p) \quad \text{для} \quad x_p \leqslant \bar{x}_p \leqslant \bar{y}_p \leqslant y_p,$$

а также

$$\mathscr{A}_p(x_p, y_p) \leqslant \mathscr{B}_p(x_p, y_p).$$

Теперь выполнены условия (1), (2), (3) теоремы 1. Так как $\mathscr{H}_p$ — M-матрица, то из известной теоремы и того факта, что $\varepsilon \geq 0$, следует при достаточно малых $\mathscr{U}_p \geqslant \mathscr{O}_p$, что $\mathscr{B}_p(x_p, y_p)$ является M-матрицей. Это означает, что выполнено (4) (при условии, что ε удовлетворяет неравенству

$$h_{ij} + \varepsilon a_{ij} v_i\left(x_j^{(0)}, y_j^{(0)}\right) \leqslant 0, \quad i \neq j, \ 1 \leqslant i, \ j \leqslant n).$$

Пусть $\mathscr{D}_p$ обозначает диагональную часть матрицы $\mathscr{B}_p\left(x_p^{(0)}, y_p^{(0)}\right)$. Ввиду

$\mathscr{D}_p^{-1} \geqslant \mathscr{O}_p$ и $\mathscr{B}_p\left(x_p^{(0)}, y_p^{(0)}\right) = \mathscr{D}_p - \left(\mathscr{D}_p - \mathscr{B}_p\left(x_p^{(0)}, y_p^{(0)}\right)\right)$ мы,

полагая $\mathscr{P}_p^{(0)} = \mathscr{D}_p^{-1}$, получаем, что

$$\mathscr{B}_p\left(x_p^{(0)}, y_p^{(0)}\right) \mathscr{P}_p^{(0)} = \mathscr{I}_p - \left(\mathscr{D}_p - \mathscr{B}_p\left(x_p^{(0)}, y_p^{(0)}\right)\right) \mathscr{D}_p^{-1} \leqslant \mathscr{I}_p,$$

$$\mathscr{P}_p^{(0)} \mathscr{B}_p\left(x_p^{(0)}, y_p^{(0)}\right) = \mathscr{I}_p - \mathscr{D}_p^{-1}\left(\mathscr{D}_p - \mathscr{B}_p\left(x_p^{(0)}, y_p^{(0)}\right)\right) \leqslant \mathscr{I}_p,$$

так как $\mathscr{D}_p - \mathscr{B}_p\left(x_p^{(0)}, y_p^{(0)}\right) \geqslant \mathscr{O}_p$.

Наш выбор матрицы $\mathscr{P}_p^{(0)}$ обеспечивает выполнение условий (5), (6) и (7) теоремы 1.

Если матрица $\mathscr{B}_p\left(x_p^{(0)}, y_p^{(0)}\right)$ представлена в виде $\mathscr{D}_p - \mathscr{L}_p - \mathscr{R}_p$, где $\mathscr{L}_p$ — строго нижняя треугольная матрица, а $\mathscr{R}_p$ — строго верхняя треугольная матрица, то

$$\mathscr{P}_p^{(0)} = (\mathscr{D}_p - \mathscr{L}_p)^{-1} \geqslant \mathscr{O}_p,$$

а ввиду $\mathscr{R}_p \geqslant \mathscr{O}_p$ следует, что

$$\mathscr{B}_p\left(x_p^{(0)}, y_p^{(0)}\right) \mathscr{P}_p^{(0)} = \mathscr{I}_p - \mathscr{R}_p\left(\mathscr{D}_p - \mathscr{L}_p\right)^{-1} \leqslant \mathscr{I}_p,$$

$$\mathscr{P}_p^{(0)} \mathscr{B}_p\left(x_p^{(0)}, y_p^{(0)}\right) = \mathscr{I}_p - \left(\mathscr{D}_p - \mathscr{L}_p\right)^{-1} \mathscr{R}_p \leqslant \mathscr{I}_p.$$

Это указывает еще одного возможного кандидата на роль $\mathscr{P}_p^{(0)}$.

Условие (17) будет выполнено, если производные отображений g_i удовлетворяют условию Липшица. Тогда выполнено и условие (18). Действительно, для

$$\mathscr{f}'_{\mathrm{p}}(z_{\mathrm{p}}) = \left(h_{ij} + \varepsilon\alpha_{ij}\frac{d}{dx_j}g_j(z_j)\right)$$

и

$$\mathscr{B}_{\mathrm{p}}(x_{\mathrm{p}}, y_{\mathrm{p}}) = (h_{ij} + \varepsilon\alpha_{ij}v_j(x_j, y_j))$$

очевидным образом выполнено соотношение

$$\mathscr{f}'_{\mathrm{p}}(z_{\mathrm{p}}) - \mathscr{B}_{\mathrm{p}}(x_{\mathrm{p}}, y_{\mathrm{p}}) = \left(\varepsilon\alpha_{ij}\left(\frac{d}{dx_j}g_j(z_j) - v_j(x_j, y_j)\right)\right).$$

Если положим теперь

$$[u_j(x_j, y_j), v_j(x_j, y_j)] := \frac{d}{dx_j}g_j([x_j, y_j]),$$

то из монотонности включения следует, что

$$\frac{d}{dx_j}g_j(z_j) \in [u_j(x_j, y_j), v_j(x_j, y_j)].$$

Теперь из теоремы 5 п.7.3 следует, что

$$v_j(x_j, y_j) - \frac{d}{dx_j}g_j(z_j) \leqslant v_j(x_j, y_j) - u_j(x_j, y_j) \leqslant c_j(y_j - x_j).$$

Поэтому получаем

$$\mathscr{B}_{\mathrm{p}}(x_{\mathrm{p}}, y_{\mathrm{p}}) - \mathscr{f}'_{\mathrm{p}}(z_{\mathrm{p}}) \leqslant \max_{1 \leqslant i \leqslant n}\{y_i - x_i\}\mathscr{C}_{\mathrm{p}}$$

для матрицы $\mathscr{C}_{\mathrm{p}} = (c_{ij})$, $c_{ij} = c_i\varepsilon\alpha_{ij} \geqslant 0$, $1 \leqslant i, j \leqslant n$, откуда следует (18).

Системы нелинейных уравнений вида (26) получаются при численном решении граничных задач вида

$$y'' = f(t, y), \quad y(0) = \bar{a}, \quad y(1) = \bar{b}.$$

Применение метода конечных разностей к такой системе дает n уравнений

$$x_{i-2} - 2x_i + x_{i+1} - h^2\{\alpha f(t_{i-1}, x_{i-1}) + \beta f(t_i, x_i) + \gamma f(t_{i+1}, x_{i+1})\} = 0, \quad 1 \leqslant i \leqslant n,$$

где $t_i = ih$, $0 \leqslant i \leqslant n+1$, $(n+1)h = 1$ и $x_0 = \bar{a}$, $x_{n+1} = \bar{b}$. Числа x_i — приближения к $y(t_i)$. При $\alpha = \gamma = 0$, $\beta = 1$ получается стандартный метод конечных разностей. При

$$\alpha = \gamma = \frac{1}{2}, \quad \beta = \frac{10}{12}$$

получается эрмитовский метод. Вводя обозначения

$$x_p = \begin{Vmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{Vmatrix}, \quad g_p(x_p) = \begin{Vmatrix} f(t_1, x_1) \\ f(t_2, x_2) \\ \vdots \\ f(t_n, x_n) \end{Vmatrix}, \quad f_p = \begin{Vmatrix} -\bar{a} + ah^2 f(0, \bar{a}) \\ 0 \\ \vdots \\ -\bar{b} + \gamma h^2 f(1, \bar{b}) \end{Vmatrix},$$

$$\mathcal{H}_p = \begin{Vmatrix} 2 & -1 & & & 0 \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ 0 & & & -1 & 2 \end{Vmatrix}, \quad \mathcal{U}_p = \begin{Vmatrix} \beta & \gamma & & & 0 \\ \alpha & \beta & \gamma & & \\ & \ddots & \ddots & \ddots & \\ & & \alpha & \beta & \gamma \\ 0 & & & \alpha & \beta \end{Vmatrix}$$

и полагая $\varepsilon = h^2$, можно записать эти n уравнений в виде

$$f_p(x_p) = \mathcal{H}_p x_p + \varepsilon \mathcal{U}_p g_p(x_p) - b_p.$$

Из предположения $\partial f/\partial y \geqslant 0$ для $t \in (0, 1)$ следует, что f_p удовлетворяет всем условиям, наложенным на системы вида (26). В случае эрмитовского метода нужно еще выбрать h достаточно малым. Мы рассмотрим в качестве конкретного примера граничную задачу

$$y'' = \sin y + y, \quad y(0) = 1, \quad y(1) = 1.$$

Так как $\partial^2 f/\partial y^2 = -\sin y$ меняет знак, то очевидно, что дискретизированная задача не является выпуклой на всем пространстве $V_n(\mathbb{R})$. Последовательные приближения к величине $y\left(\frac{1}{2}\right)$ для n=5, 25, 51 и 101 приведены в табл. 1 для обычного метода конечных разностей и в табл. 2 для эрмитовского метода. При этих значениях для эрмитовского метода имеет место

$$f_p(x_p^{(0)}) \leqslant o_p \leqslant f_p(y_p^{(0)}).$$

Таблица 1

Обычный метод конечных разностей

k		
	$n = 5$	
0	−0.5	0.5
1	0.3940299983760	0.4000335866235
2	0.3989339526997	0.3989347005095
3	0.3989344659822	0.3989344659822
	$n = 25$	
0	−0.5	0.5
1	0.3935413781128	0.3997788906381
2	0.3986874733555	0.3986882610402
3	0.3986880255447	0.3986880255452
4	0.3986880255452	0.3986880255452
k	$n = 51$	
0	−0.5	0.5
1	0.3935206369679	0.3997680696930
2	0.3986771185469	0.3986779080124
3	0.3986776725106	0.3986776725169
4	0.3986776725137	0.3986776725153
k	$n = 101$	
0	−0.5	0.5
1	0.3935155168238	0.3997653993461
2	0.3986745645657	0.3986753545055
3	0.3986751189564	0.3986751189564

Таблица 2

Эрмитовский метод

k		
	$n = 5$	
0	−0.5	0.5
1	0.3938048950831	0.3997635541509
2	0.3986758325059	0.3986765352929
3	0.3986763144021	0.3986763144021
k	$n = 25$	
0	−0.5	0.5
1	0.3935292233327	0.3997644587939
2	0.3986736779243	0.3986744630652
3	0.3986742283178	0.3986742283191
k	$n = 51$	
0	−0.5	0.5
1	0.3935175960669	0.3997644612118
2	0.3986736691791	0.3986744580066
3	0.3986742226748	0.3986742226762
4	0.3986742226762	0.3986742226762
k	$n = 101$	
0	−0.5	0.5
1	0.3935147300836	0.3997644611468
2	0.3986736680718	0.3986744577855
3	0.3986742223087	0.3986742223447
4	0.3986742223155	0.3986742223174

Микромодуль 40

Полношаговые и короткошаговые методы Ньютоновского типа

В этом микромодуле рассмотрим итерационные методы локализации решений вещественных систем нелинейных уравнений, где не требуется обращать матрицы ни точно, ни приближенно. Мы снова исходим из системы нелинейных уравнений

$$f_p(x_p) = o_p$$

и предполагаем, что f_p непрерывно дифференцируема над данным интервальным вектором $x^{(0)}$. Применяя теорему о среднем значении и полагая $f_p(x_p) = (f_i(x_p))$, получаем тогда

$$f_i(x_p) = f_i(y_p) + \sum_{j=1}^{n} \frac{\partial}{\partial x_j} f_i(y_p + \theta_i(x_p - y_p))(x_j - y_j),$$

$$0 < \theta_i < 1, \quad 1 \leqslant i \leqslant n, \quad y_p,\ x_p \in x^{(0)},$$

и, полагая

$$\mathcal{J}_p(x_p) = \left(\frac{\partial}{\partial x_j} f_i(y_p + \theta_i(x_p - y_p))\right).$$

получаем уравнение

$$f_p(x_p) - f_p(y_p) = \mathcal{J}_p(x_p)(x_p - y_p).$$

Если y_p — решение исходной системы уравнений, принадлежащее $x_p^{(0)}$, и мы представили матрицу $\mathcal{J}_p(x_p)$ в виде

$$\mathcal{J}_p(x_p) = \mathcal{D}_p(x_p) - \mathcal{L}_p(x_p) - \mathcal{U}_p(x_p),$$

где $\mathcal{D}_p(x_p)$ — диагональная матрица, $\mathcal{L}_p(x_p)$ — строго нижняя треугольная матрица, $\mathcal{U}_p(x_p)$ — строго верхняя треугольная матрица, то из последнего уравнения следует для невырожденной матрицы $\mathcal{D}_p(x_p)$ соотношение

$$y_p = x_p - \mathcal{D}_p(x_p)^{-1}\{\mathcal{B}_p(x_p)(x_p - y_p) + f_p(x_p)\},$$

где $\mathcal{B}_p(x_p) = \mathcal{L}_p(x_p) + \mathcal{U}_p(x_p)$.

Ввиду $y_p + \theta_i(x_p - y_p) \in x^{(0)}$, $1 \leqslant i \leqslant n$, из монотонности включения следует, что

$$\frac{\partial}{\partial x_j} f_i(y_p + \theta_i(x_p - y_p)) \in \frac{\partial}{\partial x_j} f_i(x^{(0)}), \quad 1 \leqslant i,\ j \leqslant n.$$

Представим теперь интервальную матрицу

$$f'_p(x^{(0)}) = \left(\frac{\partial}{\partial x_j} f_i(x^{(0)})\right)$$

в том же виде, что и матрицу $\mathcal{J}_p(x_p)$:

$$f'_p(x^{(0)}) = \mathcal{D}_p(x^{(0)}) - \mathcal{L}_p(x^{(0)}) - \mathcal{U}_p(x^{(0)}).$$

Положим еще $\mathcal{B}_p(x^{(0)}) := \mathcal{L}_p(x^{(0)}) + \mathcal{U}_p(x^{(0)})$ и $x_p := m(x^{(0)})$.

Если никакой диагональный элемент матрицы $f'_p(x^{(0)})$ не содержит нуля, то с помощью монотонности включения получим

$$y_p \in m(x^{(0)}) - \mathcal{D}_p(x^{(0)})^{-1}\{\mathcal{B}_p(x^{(0)})(m(x^{(0)}) - y_p) + f_p(m(x^{(0)}))\}.$$

Здесь подразумевается, что $\mathcal{D}_p(x^{(0)})^{-1}$ — это диагональная матрица, полученная из диагональной матрицы

$$\mathcal{D}_p(x^{(0)}) = \operatorname{diag}\left(\frac{d}{dx_i} f_i(x^{(0)})\right)$$

обращением диагональных элементов.

Если теперь выбран интервальный вектор $x^{(0)}$, содержащий y_p, то с помощью математической индукции по номеру шага итерации в итерационном методе

$$z^{(\nu+1)} = \{m(x^{(0)}) - \mathcal{D}_p(x^{(0)})^{-1}\{\mathcal{B}_p(x^{(0)}(m(x^{(0)}) - z^{(\nu)}) + \mathcal{f}_p(m(x^{(0)}))\}\} \cap z^{(\nu)} \quad (1)$$

можно показать, что имеет место $y_p \in z^{(\nu)}$, $\nu \geqslant 0$. Так как на каждом шаге берется пересечение, последовательность $\{z^{(\nu)}\}_{\nu=0}^{\infty}$ сходится к некоторому пределу z, и имеем $y_p \in z$. Исходя из $x^{(1)} := z$, вычисляем интервальную матрицу $\mathcal{f}_1'(x^{(1)})$, а с ее помощью — новую локализацию $x^{(2)}$ решения y_p и т. д. Однако мы выберем сейчас другой путь В итерационном методе (1) положим $z^{(0)} := x^{(0)}$. Из предшествующих рассуждений следует, что

$$y_p \in z^{(1)} = \{m(x^{(0)}) - \mathcal{D}_p(x^{(0)})^{-1}\{\mathcal{B}_p(x^{(0)})(m(x^{(0)}) - x^{(0)}) + \mathcal{f}_p(m(x^{(0)}))\}\} \cap x^{(0)}$$

и $z^{(1)} \subseteq x^{(0)}$. Кажется естественным взять $x^{(1)} := z^{(1)}$ и вычислить новую интервальную матрицу $\mathcal{f}_p'(x^{(1)})$ с помощью этого интервального вектора, который локализует y_p не хуже, чем $x^{(0)}$. В итерационном методе (1) для вычисления новой локализации нужен ровно один шаг. Соединяя все это, мы окончательно получаем следующую итерационную процедуру:

$$\begin{cases} u^{(k+1)} = m(x^{(k)}) - \mathcal{D}_p(x^{(k)})^{-1}\{\mathcal{B}_p(x^{(k)})(m(x^{(k)}) - x^{(k)}) + \mathcal{f}_p(m(x^{(k)}))\} \\ x^{(k+1)} = x^{(k)} \cap u^{(k+1)}, \quad k \geqslant 0. \end{cases} \quad (2)$$

Таким образом, ***данная система <u>нелинейных</u> уравнений преобразуется, как и в методе Ньютона, в систему <u>линейных</u> уравнений с <u>интервальными коэффициентами</u>.*** По этой линейной системе новая локализация для y_p вычисляется с помощью метода, похожего на полношаговый. Поэтому мы называем его полношаговым методом ньютоновского типа.

По аналогии с предыдущими рассуждениями мы устанавливаем, что (в обозначениях $u^{(k)} = (U_i^{(k)})$, $x^{(k)} = (X_i^{(k)})$) итерационный метод

$$\begin{cases} U_i^{(k+1)} = m\left(X_i^{(k)}\right) - \dfrac{1}{\left(\partial/\partial x_i\right) f_i\left(x^{(k)}\right)} \left\{ \sum\limits_{j=1}^{i-1} \dfrac{\partial}{\partial x_j} f_i\left(x^{(k)}\right)\left(m\left(X_j^{(k)}\right) - X_j^{(k+1)}\right) \right. \\ \qquad \left. + \sum\limits_{j=i+1}^{n} \dfrac{\partial}{\partial x_j} f_i\left(x^{(k)}\right)\left(m\left(X_j^{(k)}\right) - X_j^{(k)}\right) + f_i\left(m\left(x^{(k)}\right)\right) \right\}, \\ X_i^{(k+1)} = U_i^{(k+1)} \cap X_i^{(k)}, \quad 1 \leqslant i \leqslant n, \quad k \geqslant 0, \end{cases} \tag{3}$$

порождает векторы $x^{(k)}$, содержащие решение y_p. Назовем его короткошаговым методом ньютоновского типа со взятием покомпонентных пересечений.

Теперь исследуем для этих двух методов условия сходимости к решению $y_p \in x^{(0)}$. Так как мы берем пересечение, каждый из этих методов порождает последовательность

$$x^{(0)} \supseteq x^{(1)} \supseteq \ldots \supseteq x^{(k)} \supseteq x^{(k+1)} \supseteq \ldots,$$

которая в силу следствия 8 из микромодуля 29 сходится к некоторому интервальному вектору x. Далее имеем

$$y_p \in x^{(k)}, \quad k \geqslant 0 \quad \text{и} \quad y_p \in x.$$

Без дополнительных предположений имеем в общем случае равенства $\lim\limits_{k \to \infty} x^{(k)} = y_p$. Следующая теорема дает достаточные условия равенства $d(x) = o_p$, из которого ввиду $y_p \in x$ следует $\lim\limits_{k \to \infty} x^{(k)} = y_p$.

Теорема 1. *Пусть система нелинейных уравнений* $f_p(x_p) = o_p$ *имеет решение* y_p, *принадлежащее интервальному вектору* $x^{(0)}$. *Представим интервальную матрицу*

$$f'_p\left(x^{(0)}\right) = \left(\frac{\partial}{\partial x_j} f_i\left(x^{(0)}\right)\right)$$

в виде

$$f'_p\left(x^{(0)}\right) = \mathcal{D}_p\left(x^{(0)}\right) - \mathcal{L}_p\left(x^{(0)}\right) - \mathcal{U}_p\left(x^{(0)}\right)$$

и предположим, что

$$0 \notin \frac{\partial}{\partial x_i} f_i\left(x^{(0)}\right), \quad 1 \leqslant i \leqslant n.$$

Пусть $\mathcal{B}_p\left(x^{(0)}\right) = \mathcal{L}_p\left(x^{(0)}\right) + \mathcal{U}_p\left(x^{(0)}\right)$. *Если выполнено условие*

$$\rho\left(\left|\mathcal{D}_p\left(x^{(0)}\right)^{-1}\right| \left|\mathcal{B}_p\left(x^{(0)}\right)\right|\right) < 1,$$

где ρ *обозначает спектральный радиус, то последовательность* $\left\{x^{(k)}\right\}_{k=0}^{\infty}$ *для методов* (2) *и* (3) *сходится к решению* y_p.

Доказательство. Докажем это утверждение для полношагового метода (2). При $k \to \infty$ из формул (2) следует, что

$$u = m(x) - \mathscr{D}_p(x)^{-1}\{\mathscr{L}_p(x)(m(x) - x) + \mathscr{U}_p(x)(m(x) - x) + f_p(m(x))\},$$

$$x = x \cap u.$$

Из $x = x \cap u$ следует, что $x \subseteq u$, в частности

$$m(x) \subseteq m(x) - \mathscr{D}_p(x)^{-1}\{\mathscr{L}_p(x)(m(x) - x) + \mathscr{U}_p(x)(m(x) - x) + f_p(m(x))\},$$

т. е.

$$o_p \in \mathscr{D}_p(x)^{-1}\{\mathscr{L}_p(x)(m(x) - x) + \mathscr{U}_p(x)(m(x) - x) + f_p(m(x))\}.$$

Из того, что диагональные элементы матрицы $\mathscr{D}_p(x)^{-1}$ не содержат нуля, следует

$$o_p \in \mathscr{L}_p(x)(m(x) - x) + \mathscr{U}_p(x)(m(x) - x) + f_p(m(x)).$$

С помощью (19 п. 7.2) получаем теперь, что

$$d(x) \leqslant d(u) = |\mathscr{D}_p(x)^{-1}|(|\mathscr{L}_p(x)|d(x) + |\mathscr{U}_p(x)|d(x)) \leqslant |\mathscr{D}_p(x)^{-1}|(|\mathscr{L}_p(x)|d(u) + |\mathscr{U}_p(x)|d(u)) = |\mathscr{D}_p(x)^{-1}||\mathscr{B}_p(x)|d(u).$$

Из того что $\rho(|\mathscr{D}_p(x)^{-1}||\mathscr{B}_p(x)|) \leqslant \rho(|\mathscr{D}_p(x^{(0)})^{-1}||\mathscr{B}_p(x^{(0)})|) < 1$, следует, что $d(u) = o_p$, т. е. $d(x) = o_p$. Тем самым доказано, что последовательность $\{x^{(k)}\}_{k=0}^{\infty}$ сходится к y_p. Доказательство для короткошагового метода аналогично.

Из условий теоремы 1 следует, что система уравнений $f_p(x_p) = o_p$ не может иметь отличных от y_p решений, содержащихся в $x^{(0)}$.

Описанные выше методы можно модифицировать разными способами. Опишем эти модификации для полношагового метода ньютоновского типа. В случае когда вычисление $f'_p(x^{(k)})$ очень трудоемко, имеет смысл делать в методе (1) более одного шага вычисления локализации для y_p. Количество r шагов, которые выполняются после вычисления $f'_p(x^{(k)})$, тоже может зависеть от k: $r = r_k$. Это приводит к следующей итерации:

$$\begin{cases} x^{(0,\,0)} = x^{(0)}, \\ u^{(k+1,\,i)} = m\left(x^{(k)}\right) - \mathscr{D}_{\mathrm{p}}\left(x^{(k)}\right)^{-1}\left\{\mathscr{B}_{\mathrm{p}}\left(x^{(k)}\right)\left(m\left(x^{(k)}\right)\right.\right. \\ \qquad \left.\left. - x^{(k,\,i-1)}\right) + \mathscr{f}_{\mathrm{p}}\left(m\left(x^{(k)}\right)\right)\right\}, \\ x^{(k,\,i)} = u^{(k+1,\,i)} \cap x^{(k,\,i-1)}, \quad 1 \leqslant i \leqslant r_k, \\ x^{(k+1)} = x^{(k,\,r_k)}, \\ x^{(k+1,\,0)} = x^{(k+1)}, \qquad k \geqslant 0. \end{cases} \tag{4}$$

Аналогичным образом можно модифицировать и второй из описанных выше методов — короткошаговый метод (3) ньютоновского типа со взятием покомпонентных пересечений. Полученный метод сходится к y_{p} в условиях теоремы 1.

Модифицированные методы интересны еще и потому, что сходятся суперлинейно при подходящих условиях на последовательность $\{r_k\}_{k=0}^{\infty}$. Сформулируем и докажем это только для полношагового метода ньютоновскоого типа. Формулировка и доказательство следующего утверждения сохраняются и для короткошагового метода.

Теорема 2. *Пусть выполнены условия теоремы* 1. *Если модифицированный полношаговый метод* (4) *удовлетворяет условию* $r_k \to \infty$ *при* $k \to \infty$, *то*

$$\| d\left(x^{(k+1)}\right)\| \leqslant c_k \| d\left(x^{(k)}\right)\|, \quad k \geqslant 0,$$

причем

$$\lim_{k \to \infty} c_k = 0,$$

если для элементов матрицы $\mathscr{f}'_{\mathrm{p}}\left(x^{(k)}\right)$ *выполнено* (5' п. 7.3).

Доказательство. Перед тем как переходить к подробному доказательству, заметим, что в условиях теоремы 1 метод (4) сходится к y_{p}. Это можно показать так же, как в доказательстве теоремы 1.

Покажем теперь, что имеет место

$$d\left(x^{(k+1)}\right) = d\left(x^{(k,\,r_k)}\right) \leqslant \left(\mathscr{M}_{\mathrm{p}}^{r_k} + \mathscr{R}_{\mathrm{p}r_k}\right) d\left(x^{(k)}\right),$$

где

$$\mathscr{M}_{\mathrm{p}} = \left|\mathscr{D}_{\mathrm{p}}\left(x^{(0)}\right)^{-1}\right| \left|\mathscr{B}_{\mathrm{p}}\left(x^{(0)}\right)\right|$$

и где $\mathscr{R}_{\mathrm{p}r_k}$ — вещественная матрица, зависящая от $x^{(k)}$ и такая, что $\mathscr{R}_{\mathrm{p}r_k} \to O_{\mathrm{p}}$ при $k \to \infty$. После этого утверждения теоремы сразу получатся из $\rho\left(\mathscr{M}_{\mathrm{p}}\right) < 1$. В доказательстве нашего неравенства воспользуемся тем, что из (5' п.7.3) следует

$$d\left(\mathscr{D}_{\mathrm{p}}\left(x^{(k)}\right)^{-1}\mathscr{B}_{\mathrm{p}}\left(x^{(k)}\right)\right) \leqslant \mathscr{C}_{\mathrm{p}} d\left(x^{(k)}\right),$$

а также

$$d\left(\mathscr{D}_{\mathrm{p}}(x^{(k)})^{-1}\right)\left|f'_{\mathrm{p}}(x^{(k)})\right| \leqslant d\left(\mathscr{D}_{\mathrm{p}}(x^{(k)})^{-1}\left|f'(x^{(0)})\right|\right) \leqslant \mathscr{E}_{\mathrm{p}} d(x^{(k)}).$$

Здесь $\mathscr{C}_{\mathrm{p}}$ и $\mathscr{E}_{\mathrm{p}}$ — симметрические билинейные операторы, не зависящие от k.

Мы используем также включение $\mathscr{A}(\mathscr{B}c_{\mathrm{p}}) \subseteq (\mathscr{A}\mathscr{B})\, c_{\mathrm{p}}$ для произвольных интервальных матриц $\mathscr{A}$, $\mathscr{B}$ и точечного вектора c_{p}. (Его доказательство основано на определении 3 из микромодуля 29 и законе дистрибутивности $(A+B)c = Ac + Bc$ для $A,\ B \in I(R),\ c \in \mathbb{R}$.)

Ввиду $x^{(k,\,0)} = x^{(k)}$ получаем из (5),что

$$\begin{aligned} d(x^{(k,\,i)}) &\leqslant d(x^{(k+1,\,i)}) \\ &\leqslant d\left(\mathscr{D}_{\mathrm{p}}(x^{(k)})^{-1}\left\{\mathscr{B}_{\mathrm{p}}(x^{(k)})(m(x^{(k)}) - x^{(k,\,0)})\right\}\right) \\ &\quad + d\left(\mathscr{D}_{\mathrm{p}}(x^{(k)})^{-1} f_{\mathrm{p}}(m(x^{(k)}))\right) \\ &< (\mathscr{M}_{\mathrm{p}} + \mathscr{R}_{\mathrm{p}1})\, d(x^{(k)}), \end{aligned}$$

где

$$\mathscr{R}_{\mathrm{p}1} = d\left(\mathscr{D}_{\mathrm{p}}(a^{(k)})^{-1}\right)\left|f'_{\mathrm{p}}(x^{(k)})\right|$$

и $\mathscr{R}_{\mathrm{p}1} \to O_{\mathrm{p}}$ при $k \to \infty$.

Допустим теперь, что для некоторого $i \geq 1$ имеем

$$d(x^{(k,\,i)}) \leqslant (\mathscr{M}_{\mathrm{p}}^{i} + \mathscr{R}_{\mathrm{p}i})\, d(x^{(k)})$$

и $\mathscr{R}_{\mathrm{p}i} \to O_{\mathrm{p}}$ при $k \to \infty$.Тогда получаем из (5), что

$$\begin{aligned} d(x^{(k,\,i+1)}) &\leqslant d(x^{(k+1,\,i+1)}) \\ &\leqslant d\left(\mathscr{D}_{\mathrm{p}}(x^{(k)})^{-1}\left\{\mathscr{B}_{\mathrm{p}}(x^{(k)})(m(x^{(k)}) - m(x^{(k,\,i)})\right.\right. \\ &\quad \left.\left. + m(x^{(k,\,i)}) - x^{(k,\,i)})\right\}\right) + d\left(\mathscr{D}_{\mathrm{p}}(x^{(k)})^{-1}\left|f'_{\mathrm{p}}(x^{(k)})\right|\right| m(x^{(k)}) \\ &\quad - m(x^{(k,\,i)}) + m(x^{(k,\,i)}) - y_{\mathrm{p}}| \\ &\leqslant \mathscr{C}_{\mathrm{p}} d(x^{(k)})\,|m(x^{(k)}) - m(x^{(k,\,i)})| \\ &\quad + |\mathscr{D}_{\mathrm{p}}(x^{(k)})^{-1}|\,|\mathscr{B}_{\mathrm{p}}(x^{(k)})|\, d(x^{(k,\,i)}) \\ &\quad + \mathscr{E}_{\mathrm{p}} d(x^{k})\,|m(x^{(k)}) - m(x^{(k,\,i)})| + \mathscr{R}_{\mathrm{p}1} d(x^{(k,\,i)}) \\ &\leqslant ((\mathscr{E}_{\mathrm{p}} + \mathscr{C}_{\mathrm{p}})\,|m(x^{(k)}) - m(x^{(k,\,i)})| \\ &\quad + \mathscr{M}_{\mathrm{p}}(\mathscr{M}_{\mathrm{p}}^{i} + \mathscr{R}_{\mathrm{p}i}))\, d(x^{(k)}) + \mathscr{R}_{\mathrm{p}1}(\mathscr{M}_{\mathrm{p}}^{i} + \mathscr{R}_{\mathrm{p}i})\, d(x^{(k)}) \\ &= (\mathscr{M}_{\mathrm{p}}^{i+1} + \mathscr{R}_{\mathrm{p}i+1})\, d(x^{(k)}), \end{aligned}$$

где

$$\mathscr{R}_{\mathrm{p}i+1} = (\mathscr{E}_{\mathrm{p}} + \mathscr{C}_{\mathrm{p}})\,|m(x^{(k)}) - m(x^{(k,\,i)})| + \mathscr{M}_{\mathrm{p}}\mathscr{R}_{\mathrm{p}i} + \mathscr{R}_{\mathrm{p}1}(\mathscr{M}_{\mathrm{p}}^{i} + \mathscr{R}_{\mathrm{p}i}).$$

Ввиду того что

$$|m(x^{(k)}) - m(x^{(k,\,l)})| \to o_p \quad \text{при} \quad k \to \infty,$$
$$\mathscr{R}_{pl} \to \mathcal{O}_p \quad \text{при} \quad k \to \infty$$

в

$$\mathscr{R}_{pl} \to \mathcal{O}_p \quad \text{при} \quad k \to \infty,$$

получаем, что

$$\mathscr{R}_{pl+1} \to \mathcal{O}_p \quad \text{при} \quad k \to \infty.$$

Отсюда можно вывести утверждение теоремы. I

Теперь исследуем один класс систем нелинейных уравнений, к которому можно применить описанные методы. Ищется решение нелинейной краевой задачи

$$\Delta u \equiv u_{ss} + u_{tt} = f(s,\ t,\ u), \quad (s,\ t) \in \Omega$$

при $u(s,\ t) = \varphi(s,\ t)$ для $(s,\ t) \in \dot{\Omega}$. Здесь Ω — односвязная ограниченная открытая область плокости (*s, t), а* $\dot{\Omega}$ обозначает границу области Ω. Предположим для простоты, что $\Omega = (0,\ 1) \times (0,\ 1)$. Если $f\colon \Omega \times \mathbb{R} \to \mathbb{R}$ непрерывно дифференцируема и $f_u(s,\ t,\ u) \geqslant 0$, $(s,\ t) \in \Omega$, $u \in \mathbb{R}$, то при относительно слабых условиях на φ рассматриваемая краевая задача имеет единственное решение. ***Для нахождения численных приближений к этому решению мы можем, например, преобразовать краевую задачу в систему нелинейных уравнений с помощью разностного метода.*** Если выбран один и тот же шаг по *s* и по *t*, то замена частных производных подходящими разностными отношениями приводит к системе из n^2 нелинейных уравнений

$$4x_{ij} - x_{i-1,\,j} - x_{i+1,\,j} - x_{i,\,j+1} - x_{i,\,j-1} + h^2 f(ih,\ jh,\ x_{ij}) = 0,$$
$$1 \leqslant i,\ j \leqslant n.$$

Здесь $h = 1/(n+1)$ и x_{ij} — приближенное значение величины *u (ih, jh).* Для *i* = 0, *n* + 1 и *j* = 0, *n* + 1 значения x_{ij} определяются функцией φ(*s, t).*

Наши n^2 уравнений можно записать в виде

$$f_p(x_p) = \mathscr{A}_p x_p + \varphi_p(x_p) - b_p = o_p, \quad (5)$$

где $x_p = (x_{ij}) \in V_{n^2}(\mathbb{R})$. Хорошо известно, что $\mathscr{A}_p$— блочно треугольная матрица и $\varphi_{ij}(x_p) = g_{ij}(x_{ij})$. В предположении $f_u(s, t, u) \geq 0$ нетрудно показать, что условия теоремы 1 выполнены независимо от ширины $d(x^{(0)})$ данной локализации $x^{(0)}$ единственного решения системы (5).

В качестве конкретного примера мы рассмотрим уравнение

$$u_{ss} + u_{tt} = u^3/(s^2 + t^2 + 1), \quad (s, t) \in (0, 1) \times (0, 1),$$
$$u(s, 0) = 1, \qquad s \in [0, 1],$$
$$u(0, t) = 1, \qquad t \in [0, 1],$$
$$u(1, t) = 2 - e^t, \quad t \in [0, 1],$$
$$u(s, 1) = 2 - e^s, \quad s \in [0, 1]$$

при $n = 5$. Легко убедиться, что векторы

$$y_p = (y_{ij}), \quad z_p = (z_{ij}) \in V_{n^2}(\mathbb{R}),$$

такие, что $y_{ij} = -1$, $z_{ij} = 2$, $1 \leqslant i, j \leqslant n$, удовлетворяют неравенству

$$f_p(y_p) \leqslant o_p \leqslant f_p(z_p)$$

для функции f_p из (5). Так как f_p является M-функцией, отсюда следует, что интервальный вектор $x^{(0)} = (X_{ij}^{(0)}) \in V_{n^2}(I(\mathbb{R}))$ для $X_{ij}^{(0)} = [-1, 2]$ содержит решение системы (5).

Приведем теперь результаты вычисления приближений к значению $u\left(\frac{1}{2}, \frac{1}{2}\right)$ с помощью такой же модификации короткошагового метода ньютоновского типа со взятием покомпонентных пересечений, как модификация полношагового метода, рассмотренная в теореме 2.

Выбираем $r_1 = 1$, $r_{k+1} = r_k + 1$, $k \geqslant 1$ и приводим результаты в табл. 1.

Таблица 1

k	*Нижняя граница*	*Верхняя граница*
0	[−1	2]
1	[− 0.8683495081969,	1.932431741809]
2	[− 0.4334368762048,	1.662891859166]
3	[0.1747052174969,	1.110238786145]
4	[0.4964693372903,	0.7889353835863]
5	[0.6094825964828,	0.6756608548531]
6	[0.6370682656242,	0.6480110082876]
7	[0.6418668139331,	0.6432014513712]
8	[0.6424732262048,	0.6425936520955]
9	[0.6425293531665,	0.6425373968087]
10	[0.6425331717500,	0.6425335694949]
11	[0.6425333631204,	0.6425333776870]
12	[0.6425333701936,	0.6425333705975]
13	[0.6425333703873,	0.6425333704038]

Итерация прекращалась, как только половинная ширина всех локализующих интервалов становилась меньше, чем 10^{-10}. С другой стороны, методу (3) потребовался 81 шаг для достижения той же точности.

Простые соображения показывают, что в этом примере количество арифметических операций на один шаг итерации примерно одно и то же. Однако в методе (3) потребовалось 81 раз производить вычисление функции и интервальное вычисление производной по сравнению с 13 вычислениями для модифицированного метода. Это отразилось и на времени вычисления для нашего примера; для метода (3) оно оказалось в 6—7 раз больше.

Замечания. Если на каждом шаге производить итерацию (1) полностью, то получится последовательность локализаций для y_p, которая, как можно показать, сходится к y_p в условиях теоремы 1. При этом ширина сходится к нулю квадратично. Таким образом, для систем, удовлетворяющих условиям теоремы 1, этот метод дает большую экономию вычислений по сравнению с системами, для которых приходится использовать метод, исследованный в теореме 2 из микромодуля 37. Ведь там приходится на каждом шаге обращать точечную матрицу и производить итеративную локализацию матрицы, обратной к некоторой интервальной матрице.

Приложение А

ПОРЯДОК СХОДИМОСТИ ИТЕРАЦИОННЫХ МЕТОДОВ в $V_n(I(C))$ и $M_{mn}(C)$)

Пусть I — итерационный метод, порождающий последовательности $\{x^{(k)}\}$, членами которых являются комплексные интервальные векторы $x^{(k)}$ из $V_n(I(\mathbb{C}))$, $k \geqslant 0$. Кроме того, предположим, что каждая такая последовательность сходится к пределу x^*, принадлежащему $V_n(I(\mathbb{C}))$. Поскольку большинство реальных итерационных методов используется для локализации конкретных решений, мы будем предполагать, что

$$x^* \subseteq x^{(k)}, \quad k \geqslant 0. \tag{1}$$

Имеются два способа, позволяющих измерить отклонение элемента $x^{(k)}$ от предела x^*:

$$\varepsilon_p^{(k)} = q(x^{(k)}, x^*) \tag{2}$$

и

$$h_p^{(k)} = d(x^{(k)}) - d(x^*). \tag{3}$$

Обе величины представляют собой неотрицательные вещественные векторы $(e_p^{(k)},\ h_p^{(k)} \geqslant o_p)$, удовлетворяющие соотношению

$$e_p^{(k)} = o_p \Leftrightarrow h_p^{(k)} = o_p \Leftrightarrow x^{(k)} = x^*. \tag{4}$$

Кроме того, $e_p^{(k)}$ и $h_p^{(k)}$ являются монотонными по включению отображениями:

$$x^{(m)} \subseteq x^{(k)} \Rightarrow e_p^{(m)} \leqslant e_p^{(k)}, \quad h_p^{(m)} \leqslant h_p^{(k)}. \tag{5}$$

Справедливость (5) непосредственно вытекает из следующего свойства метрики q:

$$A \subseteq B \subseteq C \Rightarrow q(B, A) \leqslant q(C, A), \quad A, B, C \in I(\mathbb{C}),$$

а также из свойства (9 п. 7.2), истинность которого на $I(\mathbb{C})$ очевидна. Таким образом, меньшая локализация x^* означает и меньшую величину отклонения от x^*.

Часто проще бывает выразить отклонение $x^{(k)}$ от x^* через неотрицательное вещественное число. Для этого можно использовать монотонную векторную норму $\|\cdot\|$, подставив в нее $e_p^{(k)}$ (соответственно $h_p^{(k)}$):

$$e^{(k)} = \|q(x^{(k)}, x^*)\|$$

(соответственно $h^{(k)} = \|d(x^{(k)}) - d(x^*)\|$).

Снова имеем

$$e^{(k)} = 0 \Leftrightarrow h^{(k)} = 0 \Leftrightarrow x^{(k)} = x^*,$$

а также свойство монотонности

$$x^{(m)} \subseteq x^{(k)} \Rightarrow e^{(m)} \leqslant e^{(k)}, \quad h^{(m)} \leqslant h^{(k)}.$$

Если $X = [x^*, x^*]$, то в интервальном пространстве $I(\mathbb{R})$ введенные величины получают простую интерпретацию. В этом случае

$$h^{(k)} = d(X^{(k)})$$

представляет собой расстояние между включающими границами x^*, а

$$e^{(k)} = q(X^{(k)}, x^*) = \max\{|x_1^{(k)} - x^*|, |x_2^{(k)} - x^*|\}$$

— максимальное отклонение элемента x из $X^{(k)}$ от x^*.

Теперь предположим, что последовательности $\{x^{(k)}\}$, порожденные итерационной процедурой I, сходятся к x^* и удовлетворяют (1). Тогда соответствующие последовательности $\{e_p^{(k)}\}$ и $\{h_p^{(k)}\}$ сходятся

к нулю. К ним можно применить хорошо известное по работе Ортеги и Рейнбольдта понятие порядка сходимости.

Предварительно определим так называемый *R*-фактор:

$$R_t\left\{u_p^{(k)}\right\}=\begin{cases}\lim\limits_{k\to\infty}\sup\left\|u_p^{(k)}\right\|^{1/k}, & t=1,\\ \lim\limits_{k\to\infty}\sup\left\|u_p^{(k)}\right\|^{1/t^k}, & t>1,\end{cases}$$

где $\left\{u_p^{(k)}\right\}$ обозначает одну из последовательностей $\left\{e_p^{(k)}\right\}$ или $\left\{f_p^{(k)}\right\}$. Ортега и Рейнбольдт, опираясь на теорему об эквивалентностинорм, доказали, что значение R_t не зависит от того, какая норма используется в качестве $\|\cdot\|$. Справедлива также следующая лемма.

Лемма 1. *Пусть* $\{x^{(k)}\}$ *— последовательность, порожденная процедурой I. Предположим, что эта последовательность сходится к* x^* *и обладает свойством* (1). *Тогда для соответствующих последовательностей* $\left\{e_p^{(k)}\right\}$ *и* $\left\{f_p^{(k)}\right\}$ *выполнено равенство*

$$R_t\left\{e_p^{(k)}\right\}=R_t\left\{f_p^{(k)}\right\},\quad t\geqslant 1.$$

Доказательство. Исходя из (27 из микромодуля 29) для $k\geq 0$, получаем

$$\frac{1}{2}f_p^{(k)}\leqslant e_p^{(k)}\leqslant f_p^{(k)}.$$

При использовании монотонной векторной нормы имеем

$$\frac{1}{2}\left\|f_p^{(k)}\right\|\leqslant\left\|e_p^{(k)}\right\|\leqslant\left\|f_p^{(k)}\right\|.$$

Так как очевидно, что

$$\lim_{k\to\infty}\left(\frac{1}{2}\right)^{1/t^k}=1,\quad t>1$$

то

$$\begin{aligned}\lim_{k\to\infty}\sup\left\|f_p^{(k)}\right\|^{1/t^k}&\geqslant\lim_{k\to\infty}\sup\left\|e_p^{(k)}\right\|^{1/t^k}\\&\geqslant\lim_{k\to\infty}\sup\left(\frac{1}{2}\right)^{1/t^k}\left\|f_p^{(k)}\right\|^{1/t^k}\\&=\lim_{k\to\infty}\sup\left\|f_p^{(k)}\right\|^{1/t^k}.\end{aligned}$$

Поскольку доказательство для $t=1$ можно провести аналогичным образом, оно здесь опущено.

Из леммы 1 следует, что в дальнейших рассуждениях можно не указывать, какая именно из последовательностей — $\left\{e_p^{(k)}\right\}$ или

$\{h_p^{(k)}\}$ — имеется в виду. Это не так, если (1) не выполнено. Проиллюстрируем сказанное на примере.

Пример. Последовательность

$$\left\{[0,\ 1] + \left(\frac{1}{2}\right)^k\right\}_{k=0}^{\infty}$$

имеет предел [0, 1]. Очевидно, что

$$e^{(k)} = \left(\frac{1}{2}\right)^k, \quad k \geqslant 0,$$

и, следовательно, при $t = 1$

$$R_1\{e^{(k)}\} = \lim_{k\to\infty} \sup \left(\left(\frac{1}{2}\right)^k\right)^{1/k} = \frac{1}{2}.$$

С другой стороны, $h^{(k)} = 0$ при всех $k \geq 0$, откуда следует, что

$$R_1\{h^{(k)}\} = 0.$$

Обозначим через $\mathfrak{G}(I, x^*)$ множество порожденных итерационной процедурой I последовательностей $\{x^{(k)}\}$, для которых справедливо

$$\lim_{k\to\infty} x^{(k)} = x^* \text{ и } x^* \subseteq x^{(k)}, \quad k \geqslant 0.$$

Под R-фактором процедуры I будем понимать

$$R_t(I, x^*) = \sup\{R_t\{e_p^{(k)}\} \mid \{x^{(k)}\} \in \mathfrak{G}(I, x^*)\}, \quad t \geqslant 1.$$

Из доказанных выше утверждений следует, что этот R-фактор не зависит от того, какая именно норма была использована в его определении; кроме того, с равным успехом он мог быть определен через последовательность $\{h_p^{(k)}\}$.

Теперь мы можем определить R-порядок итерационной процедуры I как

$$O_R(I, x^*) = \begin{cases} +\infty, \text{ если } R_t(I, x^*) = 0 \text{ и } t \geqslant 1, \\ \inf\{t \mid t \in [1, \infty),\ R_t(I, x^*) = 1\} \\ \qquad \text{в остальных случаях.} \end{cases} \tag{6}$$

Это определение применительно к последовательностям векторов из $V_n(\mathbb{R})$ предложили Ортега и Рейнбольдт.

Приведем несколько утверждений, касающихся вычисления (и оценивания) R-порядка.

Теорема 2. *Пусть 1 — итерационная процедура с пределом* x^*, *и пусть* $\mathfrak{G}(I, x^*)$ *— множество всех порожденных процедурой I последовательностей, для которых* $\lim_{k\to\infty} x^{(k)} = x^*$ *и* $x^* \subseteq x^{(k)}$,

$k \geq 0$. *Если существует* $t \geq 1$ *и константа* γ_2 *такая что для всех* $\{x^{(k)}\}$ *из* $\mathfrak{G}(I, x^*)$, *и нормы* $\|\cdot\|$ *выполнено*

$$\left\| e_p^{(k+1)} \right\| \leqslant \gamma_2 \left\| e_p^{(k)} \right\|^t, \quad k \geqslant k(\{x^{(k)}\}),$$

либо

$$\left\| \varepsilon_p^{(k+1)} \right\| \leqslant \gamma_2 \left\| \varepsilon_p^{(k)} \right\|^t, \quad k \geqslant k(\{x^{(k)}\}),$$

то R-порядок процедуры I удовлетворяет неравенству

$$O_R(I, x^*) \geqslant t.$$

С другой стороны, если существуют положительная константа γ_1 *и последовательность* $\{x^{(k)}\}$ *из* $\mathfrak{G}(I, x^*)$, *такие что*

$$\left\| e_p^{(k+1)} \right\| \geqslant \gamma_1 \left\| e_p^{(k)} \right\|^t > 0, \quad k \geqslant k_0,$$

либо

$$\left\| \varepsilon_p^{(k+1)} \right\| \geqslant \gamma_1 \left\| \varepsilon_p^{(k)} \right\|^t > 0, \quad k \geqslant k_0,$$

то R-порядок процедуры I удовлетворяет неравенству

$$O_R(I, x^*) \leqslant t.$$

Если с некоторыми константами γ_1 *и* γ_2 *выполнены оба указанных условия, то*

$$O_R(I, x^*) = t.$$

Доказательство этой теоремы строится аналогично соответствующему доказательству у Ортеги и Рейнбольдта.

Теорема 3. *Пусть I — итерационная процедура с пределом* x^*, *и пусть* $\mathfrak{G}(I, x^*)$ *— множество всех порожденных I последовательностей, для которых*

$$\lim_{k \to \infty} x^{(k)} = x^* \quad u \quad x^* \subseteq x^{(k)}, \quad k \geqslant 0.$$

Пусть

$r = \sum_{i=0}^{n} m_i > 0$, *где все* m_i *— целые неотрицательные числа. Тогда, если существует положительная константа* γ, *такая что для всех* $\{x^{(k)}\}$ *из* $\mathfrak{G}(I, x^*)$ *выполнено одно из неравенств*

$$\left\| e_p^{(k+1)} \right\| \leqslant \gamma \prod_{i=0}^{n} \left\| e_p^{(k-i)} \right\|^{m_i}, \quad k \geqslant k(\{x^{(k)}\}),$$

либо

$$\left\| \varepsilon_p^{(k+1)} \right\| \leqslant \gamma \prod_{i=0}^{n} \left\| \varepsilon_p^{(k-i)} \right\|^{m_i}, \quad k \geqslant k(\{x^{(k)}\}),$$

то R-порядок процедуры I удовлетворяет неравенству

$$O_R(I, x^*) \geqslant s,$$

где s — единственное положительное решение уравнения

$$s^{n+1} - \sum_{i=0}^{n} m_i s^{n-i} = 0.$$

Часто оказывается возможным проверить какое-либо из приведенных в теоремах 2 и 3 неравенств и тем самым получить оценку R -порядка на данной итерации.

Сделанные в этом приложении утверждения могут быть перенесены на случай $M_{mn}(I(\mathbb{C}))$. В частности, останутся справедливыми лемма 1 и теорема 2 и 3.

Приложение В

РЕАЛИЗАЦИЯ МАШИННОЙ ИНТЕРВАЛЬНОЙ АРИФМЕТИКИ НА АЛГОЛЕ 60

Теперь мы хотим дополнить материал микромодуля 25, более внимательно рассмотрев вопросы реализации машинной интервальной арифметики. Наше изложение будет в основном следовать работе.

Формулы (2 п.7.1) и определение 1из микромодуля 25 задают рамки, в которых реализуются машинные интервальные операции: логическую часть, определяющую порядок вычисления границ результата по формулам (2 п.7.1), и арифметическую часть, в которой пара границ результирующего интервала вычисляется при помощи арифметических операций с направленными округлениями (см. например формулы (8 из микромодуля 25) и (9 из микромодуля 25)). Первая часть может быть без труда описана на Алголе 60 и не нуждается поэтому в более подробном объяснении. Вторая часть, однако, требует больших усилий, поскольку в реализациях Алгола 60 направленные округления обычно не доступны. Поэтому займемся моделированием машинной арифметики, целиком основанной на использовании направленного округления вниз (↓). Соответствующая процедура LOW будет определена через обычные машинные операции. При употреблении знака * для обозначения машинных операций предполагается выполнение установленных в микромодуле 25 соглашений. В частности, считается, что

(а) множество машинных чисел R_M симметрично относительно нуля и состоит из нормализованных чисел с плавающей точкой, т. е имеющих вид

$$x = mb^e,$$

где m — мантисса, b — основание степени, e — порядок;

(б) все арифметические операции с плавающей точкой оптимальны, т. е. результат вещественной операции округляется с помощью отображения *fl*, обладающего свойством (3 из микромодуля 25) (см (9 из микромодуля 25)).

Как было указано в микромодуле 25, направленное округление вверх ↑ получается из ↓ по формуле (5 из микромодуля 25). Эта формула сохраняет силу и для машинных операций:

$$\uparrow(x * y) = \begin{cases} -\downarrow((-x)*(-y)), & \text{где } * \in \{+, -\}, \\ -\downarrow((-x)*y), & \text{где } * \in \{\cdot, /\}. \end{cases}$$

Здесь ↓ обозначает результат выполнения процедуры LOW, а *— машинную операцию

Запишем теперь процедуру LOW на Алголе 60. Эта процедура позволяет вычислить нижнюю границу вещественного результата операции над числами с плавающей точкой. Ее заголовок выглядит так:

```
'REAL''PROCEDURE'LOW(X);
        'VALUE'X; 'REAL'X;
```

Каким будет тело процедуры, зависит от некоторых особенностей машинной арифметики. К их числу относятся

(I) тип округления *fl*, используемого для перехода от вещественных операций к операциям с плавающей точкой. Мы различаем

(1) округление к ближайшему со стороны нуля машинному числу, получаемое отсечением младших разрядов мантиссы, и

(2) округление к ближайшему машинному числу;

(II) поведение вблизи нуля Если результат вещественной операции оказывается меньше по абсолютной величине, чем минимальное положительное машинное число, то мы различаем

(3) установление результата в нуль и

(4) установление результата равным минимальному по абсолютной величине машинному числу со знаком, совпадающим со знаком вещественного результата,

(III) поведение в случае выхода за диапазон машинных чисел. Осмысленное расширение машинных интервальных операций возможно лишь в исключительных случаях. Поэтому в подобной

ситуации будем всегда останавливать вычисления. В теле процедуры LOW можно применить следующую конструкцию, вызывающую останов:

'IF'X'NOTGREATER' ⟨ lowerbound ⟩ THEN'

'BEGIN' X:=Ø, X:=l/X 'END'

Здесь ⟨ lowerbound ⟩ обозначает наименьшее число с плавающей точкой, т. е. $\min \{x \mid x \in \mathbb{R}_M\}$.

Пусть k — количество цифр в мантиссе, основание степени b равно 2 или 10. В тексте процедуры LOW будут использоваться следующие обозначения:

$$\langle \text{minpos} \rangle := \lim \{x \mid x \in \mathbb{R}_M,\ x > 0\},$$

$$\langle \text{factor } A \rangle := 1 - b^{1-k},$$

$$\langle \text{factor } B \rangle := 1 - \frac{1}{2} b^{1-k}.$$

Две последние константы на двоичной машине должны быть порождены как частное от деления двух целых чисел.

$$\langle \text{factor } A \rangle := (2^{k-1} - 1)/2^{k-1},$$

$$\langle \text{factor } B \rangle := (2^k - 1)/2^k.$$

Пункты (I) и (П) дают четыре возможные комбинации их подпунктов (1) —(3), (1) —(4), (2) —(3) и (2) —(4). Поэтому мы приводим четыре варианта для тела процедуры LOW.

У Криста можно найти доказательство того, что приведенные тела процедур действительно реализуют округление ↓ в случае, когда X получен в результате выполнения обычной машинной операции

Теперь, используя процедуру-функцию LOW и определяемую через нее машинную арифметику с округлениями, мы можем без труда реализовать машинные интервальные операции. Интервал $A=[a_1, a_2]$ представляется в виде 'ARRAY'1, где $A[1]$ равно a_1, A [2] равно a_2. Заголовки процедур содержат как описание интервалов-операндов A и B, так и описание результирующего интервала C (Текст вышеупомянутых процедур приведен ниже):

Арифметика (1), (3)

```
'BEGIN'
    'IF' X 'NOTGREATER' ϕ 'THEN'
      'BEGIN'
        X:= X/⟨      A⟩;
        'IF' X 'GREATER' – ⟨minpos⟩ 'THEN' X:= – ⟨minpos
      'END';
    LOW:= X
'END'
```

Арифмтика (1), (4)

```
'BEGIN'
    'IF' X 'NOTLESS' ϕ 'THEN
      'BEGIN'
        'IF' X 'NOTGREATER' ⟨minpos⟩ 'THEN' X:= ϕ
      'END'
     ELSE'
      'BEGIN'
        X:= X/⟨factor A⟩;
        'IF' X 'GREATER' – ⟨minpos⟩ 'THEN' X:= – ⟨minpos⟩
      'END';
    LOW:= X
'END'
```

Арифметика (2), (3)

```
'BEGIN'
    'IF' X 'GREATER' ϕ 'THEN'
      'BEGIN'
        X:= X × ⟨factor B⟩;
        'IF' X 'NOTGREATER' ⟨minpos⟩ 'THEN' X:= ϕ
      'END'
    'ELSE'
      'BEGIN'
        X:= X/⟨factor B⟩;
        'IF' X 'GREATER' – ⟨minpos⟩ 'THEN' X:= – ⟨minpos⟩
      'END';
    LOW:= X
'END'
```

Арифметика (2), (4)

```
'BEGIN'
    'IF' X 'NOTLESS' ϕ 'THEN'
      'BEGIN'
        X := X × ⟨factor B⟩;
        'IF' X 'NOTGREATER' ⟨minpos⟩ 'THEN' X := ϕ
      'END'
    'ELSE'
      'BEGIN'
        X := X/⟨factor B⟩;
        'IF' X 'GREATER' – ⟨minpos⟩ 'THEN' X := – ⟨minpos⟩
      'END';
    LOW := X
'END'
```

Сложение $C = A + B$:

```
'PROCEDURE'ADD(A, B, C); 'ARRAY'A, B, C;
  'BEGIN'C[1] := LOW(A[1] + B[1]);
         C[2] := – LOW(– A[2] – B[2]) 'END'
```

Вычитание $C = A - B$:

```
'PROCEDURE'SUB(A, B, C); 'ARRAY'A, B, C,
  'BEGIN'C[1] := LOW(A[1] – B[2]);
         C[2] := – LOW(B[1] – A[2]) 'END'
```

Умножение $C = A \times B$:

```
'PROCEDURE'MUL(A, B, C); 'ARRAY'A, B, C,
    'BEGIN' 'REAL'A1, A2, B1, B2, C1, C2, P; 'BOOLEAN'BA, BB;
    A2 := – A[2]; B2 := – B[2];
    BA := A2 'LESS' ∅; BB := B2 'LESS' ∅;
    'IF'BA  'THEN'A1 := A[1]
            'ELSE' 'BEGIN'A1 := A2; A2 := A[1] 'END';
    'IF'BB  'THEN'B1 := B[1]
            'ELSE' 'BEGIN'B1 := B2; B2 := B[1] 'END';
    C2 := LOW(– A2 × B2);
    'IF'B1  'LESS' ∅
    'THEN' 'BEGIN'C1 := LOW(– A2 × B1);
                'IF'A1  'LESS' ∅
                'THEN' 'BEGIN'P := LOW(– A1 × B2);
                              'IF'P 'LESS'C1 'THEN'C1 := P;
                              P := LOW(– A1 × B1);
                              'IF'P 'LESS'C2 'THEN'C2 := P
                       'END'
           'END'
    'ELSE'C1 := 'IF'A1 'LESS' ∅ 'THEN' LOW(– A1 × B2)
                               'ELSE' LOW(A1 × B1)
    'IF'BA 'EQUIV'BB
    'THEN' 'BEGIN'C[1] := C1; C[2] := – C2 'END
    'ELSE' 'BEGIN'C[1] := C2; C[2] := – C1 'END
    'END'
```

Деление $C = A/B$:

```
'PROCEDURE'DIV(A, B, C); 'ARRAY'A, B, C;
   'BEGIN' 'REAL'A1, A2, B1, B2, C1, C2; 'BOOLEAN'BA, BB;
    A2 := – A[2]; B2 := – B[2];
    BA := A2 'LESS' ∅; BB := B2 'LESS' ∅;
    'IF'BA  'THEN'A1 := A[1]
            'ELSE' 'BEGIN'A1 := A2; A2 := A[1] 'END',
    'IF'BB 'THEN'B1 := B[1]
            'ELSE' 'BEGIN'B1 := B2; B2 := B[1] 'END';

    'IF'B1'NOTGREATER' ∅
    'THEN' 'BEGIN' 'COMMENT'ОСТАНОВ ПРИ ДЕЛЕНИИ;
                              C1 := ∅; C1 := 1/C1 'END'
    'ELSE' 'BEGIN'C2 := LOW(A2/B1);
                C1 := 'IF'A1 'LESS' ∅ 'THEN' LOW(A1/B1)
                                      'ELSE'  LOW(– A1/B2)
           'END';
    'IF'BA 'EQUIV'BB
    'THEN' 'BEGIN'C[1] := C1; C[2] := – C2 'END'
    'ELSE' 'BEGIN'C[1] := C2; C[2] := – C1 'END'
    'END'
```

Вкратце обсудим, как осуществить ввод интервалов. Трудность здесь состоит в том, что при чтении десятичных дробных чисел может появиться погрешность при их преобразовании к внутреннему двоичному представлению. Величина этой погрешности, вообще говоря, неизвестна. Поэтому рекомендуется до ввода умножить границы на величины 10^n с тем, чтобы сделать их целыми числами. Эти целые числа считаются без погрешности. После ввода считанные границы делятся на 10^n, а применение процедуры LOW гарантирует корректное округление. Ниже приводится текст процедуры ввода, в котором <10^n> обозначает выбранную соответствующим образом степень десяти.

Ввод *A*

```
'PROCEDURE' READIN(A); 'ARRAY'A,
   'BEGIN' 'REAL'A1, A2;
           READ(A1, A2); A[1]:= LOW(A1/<10^n>);
                         A[2]:= – LOW(– A2/<10^n>)
   'END'
```

Рассмотрим небольшой пример программирования интервальных операций. Для простоты пример оформлен таким образом, что не возникает необходимости во вводе и выводе интервалов.

Пример. Имеется многочлен *p(x),* записанный в виде

$$p(x) = (((\ldots(a_{10}x + a_9)\,x + a_8)\ \ldots\ + a_1)\,x + a_0.$$

Это выражение должно быть вычислено в интервальной арифметике. Все коэффициенты и переменная представлены интервалами A_{10}=10, A_9 = 9, ... A_1 = 1, A_0= [—0.02, —001], а X = [—0.1, —0.02]. Мы хотим определить: $0 \in p(X)$ или нет?

(см. ниже).

```
'BEGIN'
'REAL' 'PROCEDURE'LOW(A);'VALUE'A; 'REAL'A;
'BEGIN'
'IF' A 'NOTLESS' Ø
'THEN' 'BEGIN'A:= A × Ø.999999999999; 'IF' A 'NOTGREATER' 10 − 6Ø4
  'THEN'A:= Ø 'END'
'ELSE' 'BEGIN'A:= A/Ø.999999999999; 'IF' A 'NOTLESS' − 10 − 604
  'THEN'A:= − 10 − 6Ø4 'END';
LOW:= A;
'IF'A'NOTGREATER' − 10625'THEN''BEGIN'WRITE("OUTOFRANGE"); A:= 0;
  A:= 1/A 'END'
'END';
'PROCEDURE'ADD(A, B, C); 'ARRAY' A, B, C;
'BEGIN'C[1]:= LOW(A[1] + B[1]); C[2]:= − LOW(− A[2] − B[2]) 'END';
'PROCEDURE'MUL(A, B, C); 'ARRAY'A, B, C;
  'BEGIN' 'REAL'A1, A2, B1, B2, C1, C2, P; 'BOOLEAN'BA, BB;
A2:= − A[2]; B2:= − B[2]; BA:= A2 'LESS' Ø; BB:= B2 'LESS' Ø;
'IF'BA'THEN'A1:= A[1] 'ELSE' 'BEGIN' A1:= A2; A2:= A[1] 'END';
'IF'BB 'THEN'B1:= B[1] 'ELSE' 'BEGIN' B1:= B2; B2:= B[1] 'END';
C2:= LOW(− A2 × B2);
'IF' B1 'LESS' Ø
'THEN' 'BEGIN' C1:= LOW(− A2 × B1);
                'IF' A1 'LESS' Ø
                'THEN' 'BEGIN' P:= LOW(− A1 × B2); 'IF' P 'LESS' C1
                                'THEN'C1:= P;
                               P:= LOW(− A1 × B1); 'IF' P 'LESS' C2
                                'THEN'C2:= P
                        'END'
        'END'
'ELSE'C1:= 'IF' A1 'LESS' Ø 'THEN'LOW(− A1 × B2) 'ELSE'LOW(A1 × B1);
'IF' BA 'EQUIV'BB
'THEN' 'BEGIN'C[1]:= C1; C[2]:= − C2 'END'
'ELSE' 'BEGIN'C[1]:= C2; C[2]:= − C1 'END'
'END';
'INTEGER'I; 'BOOLEAN'B; 'ARRAY'A[Ø:1Ø, 1:2], X, F, H[1:2];
'FOR'I:= 1 'STEP' 1 'UNTIL' 1Ø 'DO'A[I, 1]:= A[I, 2]:= I;
A[Ø, 1]:= LOW(− 2/1ØØ); A[Ø, 2]:= − LOW(1/1ØØ);
X[1]:= LOW(− 1/1Ø); X[2]:= − LOW(1/5Ø);
F[1]:= A[1Ø, 1]; F[2]:= A[1Ø, 2];
'FOR'I:= 9 'STEP' − 1 'UNTIL' Ø 'DO'
        'BEGIN'
            MUL(F, X, F);
            H[1]:= A[I, 1]; H[2]:= A[I, 2];
            ADD(F, H, F)
        'END',
B:= F[1] 'NOTGREATER' Ø 'AND' F[2] 'NOTLESS' Ø;
PRINT(B)
'END'
```

Приложение С
ПРОЦЕДУРЫ НА АЛГОЛЕ

А. Локализация собственных значений

Приведем теперь текст программы, представляющей собой алгол-процедуру для итерационного уточнения границ, в которых лежат собственные значения симметричных трехдиагональных вещественных матриц. Процедура реализует метод одновременной локализации в форме (5 из микромодуля 27). Постановка задачи приведена в конце микромодуля 27. Как и в микромодуле 27, собственные значения симметричной трехдиагональной матрицы $\mathscr{A}_p$ мы будем обозначать через $\lambda_1, \ldots, \lambda_n$, а характеристический многочлен для $\mathscr{A}_p$ — через $p(x)$. Его значения могут быть вычислены с помощью (10 из микромодуля 27).

При реализации метода (5 из микромодуля 27) необходимо позаботиться о том, чтобы все встречающиеся в его описании операции были заменены на соответствующие машинные интервальные операции. Замену можно осуществить с помощью алгол-процедур из приложения В.

Необходимо также заметить, что все вещественные операции в (5 из микромодуля 27) должны быть выполнены как операции над точечными интервалами. Лишь при соблюдении этого требования можно быть уверенными в том, что локализация собственных значений будет гарантирована. Так, например, если $p([x^{(k,i)}, x^{(k,i)}])$ вычисляется как интервальное выражение, то результат вычисления объемлет вещественное значение $p(x^{(k,i)})$. Поэтому при реализации (5 из микромодуля 27) следует различать случаи, используя функцию sign в виде

$$\operatorname{sign}(p([x^{(k,i)}, x^{(k,i)}]));$$

функция эта определена в микромодуле 27. Встречающиеся в процедуре элементы матрицы представлены как интервалы по той же самой причине.

Если метод (5 из микромодуля 27) выполняется так, как это указано выше, то мы получаем последовательность интервалов

$$X^{(0,i)} \supset X^{(1,i)} \supset X^{(2,i)} \supset \ldots \supset X^{(k(i),i)} = X^{(k(i)+1,i)} = \ldots,$$
$$1 \leqslant i \leqslant n,$$

которые перестают изменяться после конечного числа шагов и которые на каждом шаге содержат собственное значение λ_i. В

процедуре EIGIMP, реализующей этот метод, в качестве критерия остановки выбрано выполнение равенства

$$X^{(k,\,i)} = X^{(k+1,\,i)}.$$

Перечислим формальные параметры процедуры EIGIMP

N Целое число, задающее размерность матрицы.

A Диагональ матрицы. Компоненты $A[I,\ 1]$ и $A[I,\ 2]$ служат для представления границ, лежащих на диагонали интервалов $[a_{i1},\ a_{i2}]$, $1 \leqslant i \leqslant n$.

B Вектор элементов матрицы, лежащих над ее диагональю. Компоненты $B[I,\ 1]$ и $B[I,\ 2]$ служат для представления границ составляющих этот вектор интервалов $[b_{i1},\ b_{i2}]$, $1 \leqslant i \leqslant n-1$.

X Вектор, состоящий из пар $X[I,\ 1]$, $X[I,\ 2]$. Служит для представления интервалов $[x_{i,\,1},\ x_{i,\,2}]$, локализующих собственные значения λ_i, $1 \leqslant i \leqslant n$. В момент вызова процедуры должен содержать первоначальные приближения $X^{(0,\,i)}$, $1 \leqslant i \leqslant n$. После выполнения процедуры содержит локализации, уточненные согласно (5 из микромодуля 27).

BO Булевский вектор размерности *n,* у которого *BO*[*I*] имеет значение 'TRUE', если на следующей итерации требуется построить новую локализацию для λ_i. В противном случае *BO*[*I*] = 'FALSE'. Если очередная локализация оказалась равной предыдущей, то соответствующая компонента *BO* устанавливается в 'FALSE'. Перед вызовом EIGIMP компоненты *BO,* соответствующие уточняемым компонентам *X,* должны быть установлены в 'TRUE', а остальные — в 'FALSE'. После выхода из EIGIMP *BO* [*I*] = 'FALSE' для всех *I*.

EPC Вещественная неотрицательная компонента ε, такая что если ширина локализации $d(X^{(k,\,i)})$ удовлетворяет соотношению

$$d(X^{(k,\,i)}) < \varepsilon \max\{|s(X^{(k,\,i)})|,\ |i(X^{(k,\,i)})|\},$$

где $d(X^{(k,\,i)}) = s(X^{(k,\,i)}) - i(X^{(k,\,i)})$, то *BO*[*I*] устанавливается в 'FALSE' (т. е. дальнейшее уточнение не производится). Если достигнуть выполнения данного неравенства не удается (например, в случае, когда ε оказалось равным 0), тогда применяется критерий остановки, упомянутый при описании *BO*.

EMPTY Это метка, на которую передается управление, когда при некотором *i* попытка взять пересечение, согласно (5 из микромодуля 27), приводит к получению пустого множества. Необходимость в этом возникает, когда $\lambda_i \notin X^{(0,\,i)}$.

EXIT На эту метку передается управление, если какая-либо из интервальных компонент входных массивов *A*, *B* или *X* имеет нижнюю границу больше верхней. То же самое происходит, если начальные локализации не являются попарно непересекающимися, как того требует метод (5 из микромодуля 27). Предполагается, что собственные значения пронумерованы в порядке их роста. Если по каким-либо причинам более удобен другой порядок, то необходимо изменить следующий фрагмент:

```
'FOR'I:=1 'STEP' 1 'UNTIL' N-1 'DO'
'IF'(BO[I] 'OR'BO[I + 1]) 'AND'X[I, 2] 'NOTLESS'X[I + 1, 1]
'THEN' 'GOTO'EXIT;
```

Глобальные имена. Процедура EIGIMP использует процедуры LOW и MUL как глобальные.

В. Короткошаговый метод

Теперь дадим описание алгол-процедуры ITERATION. Эта процедура реализует короткошаговый метод с взятием пересечения после вычисления каждой компоненты (SIC). Метод предназначен для нахождения неподвижной точки x^* уравнения $x = \mathcal{A}x + \mathcal{b}$ и удовлетворяет построчному или постолбцовому критерию (обсуждение метода см. в гл. 14). Элементы $\mathcal{A}$ и $\mathcal{b}$ — вещественные интервалы. Как показал Алефельд, в этом случае легко находится вектор $x^{(0)}$, содержащий x^*.

Перейдем к перечислению формальных параметров.

N Целое число, задающее размерность системы уравнений.

MAT Соответствует интервальной матрице $\mathcal{A}$, имеющей размерность $n \times n$. Компонента *MAT*[*I, J,* 1] (соответственно *MAT*[*I, J,* 2]) представляет нижнюю (соответственно верхнюю) границу A_{ij} — элемента матрицы $\mathcal{A}$.

VEC Соответствует *n*-мерному интервальному вектору $\mathcal{b}$. Его компоненты *VEC*[*I,* 1] и *VEC*[*I,* 2] представляют границы b_{i1} и b_{i2} компоненты B_i вектора *b*.

EXIT Метка, на которую передается управление, если $\mathcal{A}$ не удовлетворяет построчному или постолбцовому критерию.

X Возвращаемое значение неподвижной точки x^*.

Процедура ITERATION использует процедуры LOW, ADD и MUL как глобальные (см. ниже).

```
'PROCEDURE'EIGIMP (N,A,B,X,BO,EPS,EMPTY,EXIT);
                'VALUE'N,EPS;
                'INTEGER'N;
                'REAL'EPS;
                'ARRAY'A,B,X;
                'BOOLEAN' 'ARRAY'BO;
                'LABEL'EMPTY,EXIT;
'BEGIN'
     'INTEGER'I,J;
     'REAL'N1,N2,U,V,Y,Z;
     'BOOLEAN'BIT;
     'ARRAY'BB,F[1:N,1:2],XM,D[1:N],H,L,M,F1,F2[1:2];

'COMMENT'ВХОДНЫЕ ДАННЫЕ;
'FOR'I:= 1 'STEP' 1 'UNTIL' N 'DO'
'BEGIN'
     'IF'A[I,2] 'LESS'A[I,1] 'OR'B[I,2] 'LESS'B[I,1]
     'THEN' 'GOTO'EXIT;
     H[1]:= B[I,1];
     H[2]:= B[I,2];
     MUL(H,H,H);
     BB[I,1]:= H[1];
     BB[I,2]:= H[2]
'END'I;

'FOR'I:= 1 'STEP' 1 'UNTIL' N - 1 'DO'
'IF'(BO[I] 'OR'BO[I+1]) 'AND'X[I,2] 'NOTLESS' X[I+1,1]
'THEN' 'GOTO'EXIT;
```

```
'COMMENT'НАЧАЛЬНЫЕ ЗНАЧЕНИЯ;
'FOR'I:=1 'STEP' 1 'UNTIL' N 'DO'
'IF'BO[I] 'THEN'
'BEGIN'
      XM[I]:=(X[I,1]+X[I,2])/2;
      D[I]:=-LOW(X[I,1]-X[I,2]),
      F1[1]:=F1[2]:=1;
      F2[1]:=LOW(XM[I]-A[1,2]);
      F2[2]:=-LOW(A[1,1]-XM[I]);
      'FOR'J:=2 'STEP' 1 'UNTIL' N 'DO'
      'BEGIN'
            L[1]:=BB[J-1,1];
            L[2]:=BB[J-1,2];
            MUL(L,F1,M);
            L[1]:=LOW(XM[I]-A[J,2]);
            L[2]:=-LOW(A[J,1]-XM[I]);
            MUL(L,F2,L);
            F1[1]:=F2[1];
            F1[2]:=F2[2];
            F2[1]:=LOW(L[1]-M[2]);
            F2[2]:=-LOW(M[1]-L[2])
      'END'J;
      F[I,1]:=F2[1];
      F[I,2]:=F2[2]
'END'I,

'COMMENT'ИТЕРАЦИЯ;
REPEAT:
BIT:='FALSE';
'FOR' I:=1 'STEP' 1 'UNTIL' N 'DO'
'IF'BO[I] 'THEN'
'BEGIN'
      'COMMENT'ЗНАМЕНАТЕЛЬ;
      H[1]:=H[2]:=1;
      'FOR'J:=1 'STEP' 1 'UNTIL'I-1,I+1 'STEP' 1 'UNTIL' N 'DO'
      'BEGIN'
            L[1]:=LOW(XM[I]-X[J,2]);
            L[2]:=-LOW(X[J,1]-XM[I]);
            MUL(H,L,H)
      'END'J,
```

```
'COMMENT'НЬЮТОН;
N1:= 'IF'H[1] 'GREATER' ∅
     'THEN'
     ('IF'F[I,1] 'LESS' ∅ 'THEN'LOW(F[I,1]/H[1])
                          'ELSE'LOW(F[I,1]/H[2]))
     'ELSE'
     'IF'F[I,2] 'GREATER' ∅ 'THEN'LOW(F[I,2]/H[2])
                            'ELSE'LOW(F[I,2]/H[1]);

N2:= 'IF'H[1] 'GREATER' ∅
     'THEN'
     ('IF'F[I,2] 'GREATER' ∅ 'THEN'-LOW(-F[I,2]/H[1])
                             'ELSE'-LOW(-F[I,2]/H[2]))
     'ELSE'
     'IF'F[I,1] 'LESS' ∅ 'THEN'-LOW(-F[I,1]/H[2])
                         'ELSE'-LOW(-F[I,1]/H[1]);
Y.= N1;
N1:= LOW(XM[1]-N2);
N2:= -LOW(Y-XM[1]);
'IF'N1 'LESS'X[I,1] 'THEN'N1:= X[I,1];
'IF'N2 'GREATER'X[I,2] 'THEN'N2:= X[I,2];
'IF'N2 'LESS' N1 'THEN' 'GOTO'EMPTY;
```

```
'COMMENT'ДЕЛЕНИЕ ПОПОЛАМ'
XM[I]:= (N1+N2)/2;
F1[1]:= F1[2]:= 1;
F2[1]:= LOW(XM[I]-A[1,2]);
F2[2]:= -LOW(A[1,1]-XM[I]);
'FOR'J:= 2 'STEP' 1 'UNTIL' N 'DO'
'BEGIN'
      L[1]:= BB[J-1,1];
      L[2]:= BB[J-1,2];
      MUL(L,F1,M);
      L[1]:= LOW(XM[I]-A[J,2]);
      L[2]:= -LOW(A[J,1]-XM[I]);
      MUL(L,F2,L);
      F1[1]:= F2[1];
      F1[2]:= F2[2];
      F2[1]:= LOW(L[1]-M[2]);
      F2[2]:= -LOW(M[1]-L[2])
'END'J;
F[I,1]:= F2[1];
F[I,2]:= F2[2];
'IF'F[I,1] 'GREATER' ∅
'THEN'
      'BEGIN'
            'IF'H[1] 'GREATER' ∅
            'THEN'
                  'BEGIN'
                        X[I,2]:= XM[I];
                        X[I,1]:= N1
                  'END'
            'ELSE'
                  'BEGIN'
                        X[I,1]:= XM[I];
                        X[I,2]:= N2
                  'END'
      'END'
 'ELSE'
 'IF'F[I,2] 'LESS' ∅
 'THEN'
```

```
            'BEGIN'
                'IF'H[I] 'GREATER' ∅
                'THEN'
                    'BEGIN'
                        X[I,1]:= XM[I];
                        X[I,2]:= N2
                    'END'
                'ELSE'
                    'BEGIN'
                        X[I,2]:= XM[I];
                        X[I,1]:= N1
                    'END'
            'END'
        'ELSE'
            'BEGIN'
                X[I,1]:= N
                X[I,2]:= N2
            'END',

        'COMMENT'ПРОВЕРКА;
        Z:= -LOW(X[I,1]-X[I,2]);
        U:= ABS(X[I,1]);
        V:= ABS(X[I,2]);
        'IF' V 'LESS' U 'THEN' V:= U;
        'IF' Z 'NOTLESS'D[I] 'OR'Z 'NOTGREATER'EPS × V
        'THEN'BO[I]:= 'FALSE' 'ELSE'BIT:= 'TRUE';
        D[I]:= Z
    'END'КОНЕЦ ИТЕРАЦИИ;
    'IF'BIT 'THEN' 'GOTO'REPEAT
    'END'КОНЕЦ ПРОЦЕДУРЫ
```

С. Обращение матрицы

Перейдем теперь к описанию алгол-процедуры, которая реализует изложенный в микромодуле 36 итерационный метод для локализации матрицы, обратной к $\mathscr{A}_p$. Чтобы сохранить гарантированность включений при вычислениях в машинной интервальной арифметике, нужно заменить вещественные операции на операции с точечными интервалами. Только в этом случае мы сможем учесть все погрешности, возникающие из-за округлений. По этой причине, а также из-за того, что при вводе тоже нельзя избежать появления погрешностей, все вещественные константы и матрицы заменяются на интервальные. Для получения направленных округлений при выполнении интервальных операций мы будем пользоваться процедурами, описанными в приложении В. Сначала ознакомимся с

процедурой MATINV, которая имеет следующие формальные параметры (см. выше):

N Целое число, задающее размерность матрицы

A Служит для представления обращаемой интервальной матрицы $\mathcal{A}$ размерности $n \times n$. Компоненты $A[I, J, 1]$ и $A[I, J, 2]$ соответствуют нижней и верхней границам элемента A_{ij} матрицы $\mathcal{A}$. Если A_{ij} представим в виде точечного интервала, то, конечно, эти границы равны.

X Служит для представления интервальной матрицы размерности $n \times n$. Компоненты $X[I, J, 1]$ и $X[I, J, 2]$ соответствуют нижней и верхней границам *I*-го элемента матрицы. Параметр используется и как входной, и как выходной. При вызове процедуры этот формальный параметр заменяется фактическим параметром $\mathcal{X}^{(0)}$, таким что

$$\{\mathcal{A}_p^{-1} \mid \mathcal{A}_p \in \mathcal{A}\} \in \mathcal{X}^{(0)} \quad \text{и} \quad \| \mathcal{I}_p - \mathcal{A}m(\mathcal{X}^{(0)}) \| < 1,$$

где ||·||— норма матрицы. После подстановки $\mathcal{X}$ итерация (5 из микромодуля 36) повторяется (с $r = 2$) до тех пор, пока не будет сформирована матрица, удовлетворяющая неравенству (11 из микромодуля 36).

В качестве нормы используется

$$\| \mathcal{X} \| = n \max_{1 \leqslant i, j \leqslant n} | x_{ij} |.$$

Когда выполнено (11 из микромодуля 36), тогда при r=2 итерации продолжаются в соответствии с (9 из микромодуля 36), пока две следующие подряд локализации не окажутся равными друг другу. Тогда итерации останавливаются, а последняя интервальная матрица сохраняется в *X*, служащем выходным параметром. Для нахождения матрицы $\mathcal{X}^{(0)}$ применяется алгол-процедура INCLUSIONSET.

Процедуры LOW, ADD и MUL из приложения В используются нами как глобальные.

```
'PROCEDURE'ITERATION(MAT,VEC,N,EXIT)RESULT'(X);
'VALUE'MAT,VEC,N; 'ARRAY'MAT,VEC,X; 'INTEGER'N; 'LABEL'EXIT;
'BEGIN' 'INTEGER'I,J; 'REAL'MAX,HMAX,NORM,HNORM; 'BOOLEAN'BV;
    'ARRAY'Y[1:N],IK,HMAT,HX[1:2];
    'REAL' 'PROCEDURE'UP(A); 'VALUE'A; 'REAL'A; UP:= -LOW(-A);
    'REAL' 'PROCEDURE'ABSVAL(LBD,VBD); 'REAL'LBD,UBD;
        'IF'ABS(LBD) 'LESS'ABS(UBD) 'THEN'ABSVAL := ABS(UBD) 'ELSE'
        ABSVAL:= ABS(LBD);
    MAX:= 0;
    'COMMENT'ОЦЕНИВАНИЕ С ПОМОЩЬЮ НОРМЫ, ВЫЧИСЛЯЕМОЙ
        СУММИРОВАНИЕМ ПО СТРОКАМ;
    'FOR'I:= 1 'STEP' 1 'UNTIL' N 'DO'
        'BEGIN'Y[I]:= 0;
            'FOR'J:= 1 'STEP' 1 'UNTIL' N 'DO'
            Y[I]:= UP(Y[I] + ABSVAL (MAT[I,J,1],MAT[I,J,2]));
            'IF'Y[I] 'NOTLESS' 1 'THEN' 'GOTO'M1
        'END';
    'COMMENT'НАХОЖДЕНИЕ НАЧАЛЬНОГО ВЕКТОРА, ЕСЛИ
        УДОВЛЕТВОРЕН КРИТЕРИЙ СУММЫ ПО СТРОКАМ;
    'FOR'I:= 1 'STEP' 1 'UNTIL' N 'DO'
    'BEGIN'HMAX := 0;
        'FOR'J:= 1 'STEP' 1 'UNTIL' N 'DO'
        HMAX:= UP(HMAX+UP(ABSVAL(MAT[I,J,1],MAT[I,J,2]) ×
        ABSVAL(VEC[J,1],VEC[J,2])));
        HMAX:= UP(HMAX/LOW(1-Y[I]));
        'IF'MAX 'LESS'HMAX 'THEN'MAX:= HMAX
    'END', 'GOTO'M2;
    'COMMENT'ОЦЕНИВАНИЕ С ПОМОЩЬЮ НОРМЫ, ВЫЧИСЛЯЕМОЙ
        СУММИРОВАНИЕМ ПО СТОЛБЦАМ;
    M1 : NORM:= 0;
    'FOR'J:= 1 'STEP' 1 'UNTIL' N 'DO'
    'BEGIN'HNORM:= 0;
        'FOR'I:= 1 'STEP' 1 'UNTIL' N 'DO'
        HNORM:= UP(HNORM + ABSVAL(MAT[I,J,1],MAT[I,J,2]));
        'IF'HNORM 'NOTLESS' 1 'THEN' 'GOTO'EXIT;
        'IF'HNORM 'GREATER'NORM 'THEN'NORM:= HNORM
    'END';
```

```
'COMMENT'НАХОЖДЕНИЕ НАЧАЧАЛЬНОГО ВЕКТОРА,ЕСЛИ
    УДОВЛЕТВОРЕН КРИТЕРИЙ СУММЫ ПО СТОЛБЦАМ;
'FOR'I:= 1 'STEP' 1 'UNTIL' N 'DO'
'FOR'J:= 1 'STEP' 1 'UNTIL' N 'DO'
MAX:= UP(MAX+UP(ABSVAL(MAT[I,J,1],MAT[I,J,2]) ×
 ABSVAL(VEC[J,1],VEC[J,2])));
MAX:= UP(MAX/LOW(1-NORM));
M2 . 'FOR'I:= 1 'STEP' 1 'UNTIL' N 'DO'
    'BEGIN'X[I,1]:= LOW(VEC[I,1]-MAX);
    X[I,2]:= UP(VEC[I,2]+MAX) 'END';
'COMMENT'НАЧАЛО ОЧЕРЕДНОЙ ИТЕРАЦИИ;
M3 : BV:='TRUE';
'FOR'I:= 1 'STEP' 1 'UNTIL' N 'DO'
'BEGIN'IK[1]:= VEC[I,1]; IK[2]:= VEC[I,2];
    'FOR'J:= 1 'STEP' 1 'UNTIL' N 'DO'
    'BEGIN'HMAT[1]:= MAT[I,J,1]; HMAT[2]:= MAT[I,J,2];
        HX[1]:= X[J,1]; HX[2]:= X[J,2];
        MUL(HMAT,HX,HX); ADD(HX,IK,IK)
    'END';
    HX[1]:= X[I,1]; HX[2]:= X[I,2];
    'IF'IK[1] 'LESS'HX[1] 'THEN'IK[1]:= HX[1];
    'IF'IK[2] 'GREATER'HX[2] 'THEN'IK[2]:= HX[2];
    'IF'BV 'THEN' 'BEGIN' 'IF'IK[1] 'NOTEQUAL'X[I,1] 'OR'IK[2]
      'NOTEQUAL'
    X[I,2] 'THEN'BV = 'FALSE' 'END',
    X[I,1] = IK[1], X[I,2]:= IK[2]
'END';
'IF' 'NOT'BV 'THEN' 'GOTO'M3
'END'
```

Последней мы рассмотрим алгол-процедуру INCLUSIONSET (см. ниже), которая вычисляет начальную матрицу $\mathscr{R}^{(0)}$ для последующей передачи ее в процедуру MATINV.

```
'PROCEDURE' INCLUSIONSET(B,N,X,L);
        'VALUE'N;
        'INTEGER'N,
        'ARRAY'B,X;
        'LABEL'L;
'BEGIN' 'INTEGER'I,J;
         'REAL'NOR,A1,A2,HI;
         NOR:=0;

'COMMENT' ПОСТРОЧНАЯ НОРМА;
'FOR' I:=1 'STEP' 1 'UNTIL' N 'DO'
'BEGIN'HI:=0;
     'FOR'J:=1 'STEP' 1 'UNTIL' N 'DO'
     'BEGIN'A1:=ABS(B[I,J,1]); A2:=ABS(B[I,J,2]);
            HI:=-LOW(-HI-('IF'A1 'GREATER'A2 'THEN'A1 'ELSE'A2))
     'END'J;
     'IF'HI 'GREATER'NOR 'THEN'NOR:=HI;
     'IF'NOR 'NOTLESS' 1 'THEN' 'GOTO'COLNORM
'END'I,КОНЕЦ ПОСТРОЧНОЙ НОРМЫ;
'GOTO'INCL;
COLNORM : NOR:=0;
'FOR'J:=1 'STEP' 1 'UNTIL' N 'DO'
'BEGIN'HI:=0;
     'FOR'I:=1 'STEP' 1 'UNTIL' N 'DO'
     'BEGIN'A1:=ABS(B[I,J,1]); A2:=ABS(B[I,J,2]);
            HI:=-LOW(-HI-('IF'A1 'GREATER'A2 'THEN'A1 'ELSE'A2))
     'END'I;
     'IF'HI 'GREATER'NOR 'THEN'NOR:=HI;
     'IF'NOR 'NOTLESS' 1 'THEN' 'GOTO'L
'END'J,КОНЕЦ ПОСТОЛБЦОВОЙ НОРМЫ;

INCL :
A1:=LOW(1-NOR);
A1:=-LOW(-1/A1);
'FOR'I:=1 'STEP' 1 'UNTIL' N 'DO'
'BEGIN' 'FOR'J:=1 'STEP' 1 'UNTIL'I-1,I+1 'STEP' 1 'UNTIL' N 'DO'
     BEGIN'X[I,J,1]:=-A1;
           X[I,J,2]:=A1
     'END'J;
     X[I,I,1]:=-A1;
     X[I,I,2].=-LOW(-2-A1)
'END'I
'END'
```

Процедура имеет следующие формальные параметры:
N Задает размерность обращаемой матрицы.
Соответствует интервальной матрице $\mathcal{B}$ размерности $n \times n$. Компонента $B[I, J, 1]$ (соответственно $B[I, J, 2]$) представляет нижнюю (соответственно верхнюю) границу B_{ij} — элемента

матрицы $\mathscr{B}$. Фактическим параметром для *B* служит машинное представление обращаемой матрицы $\mathscr{A}_p$.

X Соответствует интервальной матрице $\mathscr{X}$ размерности $n \times n$. Компонента $X[1, 1, 1]$ (соответственно $X[1, 1, 2]$) представляет нижнюю (соответственно верхнюю) границу X_{ij} — элемента матрицы $\mathscr{X}$. После выполнения процедуры *X* соответствует интервальной матрице $\mathscr{X}$, обладающей свойством

$$\{\mathscr{B}^{-1} \mid \mathscr{B}_p \in \mathscr{B}\} \subseteq \mathscr{X}.$$

Вычисление $\mathscr{X}$ производится с помощью формул, следующих из теоремы 2 микромодуля 36. В качестве $|\mathscr{B}|$ используются как построчная, так и постолбцовая нормы.

L Это метка, на которую передается управление, если ни построчная, ни постолбцовая нормы $|\mathscr{B}|$ не меньше 1 (в этом случае матрицу $\mathscr{X}$ невозможно вычислить с помощью упомянутых выше формул).

Процедура LOW из прилож. В используется как глобальная.

```
'PROCEDURE'MATINV(A,N,X);
'VALUE'N;
'INTEGER'N;
'ARRAY'A,X;
'BEGIN' 'INTEGER'I,J,K;
        'REAL'U,V,W,NA,NDX,NDR,H;
        'BOOLEAN'B1,B2;
        'ARRAY'S,H1,H2[1:2],Y[1:N,1:2],M[1:N,1:N],
        Z,R[1:N,1:N,1:2];
U:=LOW(1/N);
B1:=B2:='FALSE';

'COMMENT'НОРМА A;
NA:=0;
'FOR'I:=1 'STEP' 1 'UNTIL' N 'DO'
'FOR'J:=1 'STEP' 1 'UNTIL' N 'DO'
'BEGIN'V:=ABS(A[I,J,1]); W:=ABS(A[I,J,2]):
    'IF'W 'LESS'V 'THEN'W:=V;
    'IF'W 'GREATER'NA 'THEN'NA:=W;
    Z[I,J,1]:=X[I,J,1]; Z[I,J,2]:=X[I,J,2]
'END'КОНЕЦ НОРМЫ A;
L1:B1:='FALSE';
```

```
'COMMENT'ВЫЧИСЛЕНИЕ СРЕДНЕЙ ТОЧКИ;
'FOR'I:= 1 'STEP' 1 'UNTIL' N 'DO'
'FOR'J:= 1 'STEP' 1 'UNTIL' N 'DO'
M[I,J]:= (X[I,J,1] + X[I,J,2])/2;

'COMMENT'РАЗНОСТЬ;
'FOR'I:= 1 'STEP' 1 'UNTIL' N 'DO'
'BEGIN' 'FOR'J:= 1 'STEP' 1 'UNTIL' N 'DO'
    'BEGIN'S[1]:= S[2]:= 0;
        'FOR'K:= 1 'STEP' 1 'UNTIL' N 'DO'
        'BEGIN'H1[1]:= A[I,K,1]; H1[2]:= A[I,K,2];
                H2[1]:= H2[2]:= M[K,J];
                MUL(H1,H2,H1); ADD(S,H1,S)
        'END'K;
        R[I,J,1]:= -S[2]; R[I,J,2]:= -S[1]
    'END'J;
    R[I,I,1]:= LOW(1+R[I,I,1]);
    R[I,I,2]:= -LOW(-1-R[I,I,2])
'END' КОНЕЦ РАЗНОСТИ;

'COMMENT'ОЧЕРЕДНАЯ ИТЕРАЦИЯ;
'IF'B2 'THEN' 'FOR'I:= 1 'STEP' 1 'UNTIL' N 'DO'
            'FOR'J:= 1 'STEP' 1 'UNTIL' N 'DO'
            'BEGIN'Z[I,J,1]:= X[I,J,1];
                Z[I,J,2]:= X[I,J,2]
            'END';
'FOR'I:= 1 'STEP' 1 'UNTIL' N 'DO'
'BEGIN' 'FOR'J:= 1 'STEP' 1 'UNTIL' N 'DO'
    'BEGIN'S[1]:= S[2]:= 0;
        'FOR'K:= 1 'STEP' 1 'UNTIL' N 'DO'
        'BEGIN'H1[1]:= Z[I,K,1]; H1[2]:= Z[I,K,2];
                H2[1]:= R[K,J,1]; H2[2]:= R[K,J,2];
                MUL(H1,H2,H1); ADD(S,H1,S)
        'END'K;
        H1[1]:= H1[2]:= M[I,J]; ADD(S,H1,S);
        Y[J,1]:= S[1]; Y[J,2]:= S[2]
    'END'J;
    'FOR'J:= 1 'STEP' 1 'UNTIL' N 'DO'
    'BEGIN'Z[I,J,1]:= Y[J,1];
        Z[I,J,2]:= Y[J,2]
    'END'J
'END'I;

'COMMENT'ПЕРЕСЕЧЕНИЕ;
'IF'B2 'THEN'
L2 : 'BEGIN' 'FOR'I:= 1 'STEP' 1 'UNTIL' N 'DO'
            'FOR'J:= 1 'STEP' 1 'UNTIL' N 'DO'
            'BEGIN' 'IF'X[I,J,1] 'LESS'Z[I,J,1] 'THEN'
                'BEGIN'X[I,J,1]:= Z[I,J,1]; B1:= 'TRUE' 'END';
                'IF'X[I,J,2] 'GREATER'Z[I,J,2] 'THEN'
                'BEGIN'X[I,J,2]:= Z[I,J,2]; B1:= 'TRUE' 'END'
            'END'I,J;
            'GOTO' 'IF'B1 'THEN'L1 'ELSE'L3
```

```
'PROCEDURE'ITERATION(MAT,VEC,N,EXIT)RESULT'(X);
'VALUE'MAT,VEC,N; 'ARRAY'MAT,VEC,X; 'INTEGER'N; 'LABEL'EXIT;
'BEGIN' 'INTEGER'I,J; 'REAL'MAX,HMAX,NORM,HNORM; 'BOOLEAN'BV;
    'ARRAY'Y[1:N],IK,HMAT,HX[1:2];
    'REAL' 'PROCEDURE'UP(A); 'VALUE'A; 'REAL'A; UP:= -LOW(-A);
    'REAL' 'PROCEDURE'ABSVAL(LBD,VBD); 'REAL'LBD,UBD;
        'IF'ABS(LBD) 'LESS'ABS(UBD) 'THEN'ABSVAL:= ABS(UBD) 'ELSE'
        ABSVAL:= ABS(LBD);
    MAX:= 0;
    'COMMENT'ОЦЕНИВАНИЕ С ПОМОЩЬЮ НОРМЫ, ВЫЧИСЛЯЕМОЙ
        СУММИРОВАНИЕМ ПО СТРОКАМ;
    'FOR'I:= 1 'STEP' 1 'UNTIL' N 'DO'
        'BEGIN'Y[I]:= 0;
            'FOR'J:= 1 'STEP' 1 'UNTIL' N 'DO'
            Y[I]:= UP(Y[I] + ABSVAL (MAT[I,J,1],MAT[I,J,2]));
            'IF'Y[I] 'NOTLESS' 1 'THEN' 'GOTO'M1
        'END';
    'COMMENT'НАХОЖДЕНИЕ НАЧАЛЬНОГО ВЕКТОРА, ЕСЛИ
        УДОВЛЕТВОРЕН КРИТЕРИЙ СУММЫ ПО СТРОКАМ;
    'FOR'I:= 1 'STEP' 1 'UNTIL' N 'DO'
    'BEGIN'HMAX:= 0;
        'FOR'J:= 1 'STEP' 1 'UNTIL' N 'DO'
        HMAX:= UP(HMAX+UP(ABSVAL(MAT[I,J,1],MAT[I,J,2]) ×
        ABSVAL(VEC[J,1],VEC[J,2])));
        HMAX:= UP(HMAX/LOW(1-Y[I]));
        'IF'MAX 'LESS'HMAX 'THEN'MAX:= HMAX
    'END', 'GOTO'M2;
    'COMMENT'ОЦЕНИВАНИЕ С ПОМОЩЬЮ НОРМЫ, ВЫЧИСЛЯЕМОЙ
        СУММИРОВАНИЕМ ПО СТОЛБЦАМ;
    M1 : NORM:= 0,
    'FOR'J:= 1 'STEP' 1 'UNTIL' N 'DO'
    'BEGIN'HNORM:= 0;
        'FOR'I:= 1 'STEP' 1 'UNTIL' N 'DO'
        HNORM:= UP(HNORM + ABSVAL(MAT[I,J,1],MAT[I,J,2]));
        'IF'HNORM 'NOTLESS' 1 'THEN' 'GOTO'EXIT;
        'IF'HNORM 'GREATER'NORM 'THEN'NORM:= HNORM
    'END';
```

Список литературы

1. Горбатов В.А. Основы дискретной математики. - М.: Высшая школа, 1986.
2. Коршунов Ю.М. Математические основы кибернетики. - М.:Энергия, 1980.
3. Кузнецов О.П., Адельсон-Вельский Г.М. Дискретная математика для инженера. - М.:Энергия, 1980.
4. Кук Д., Бейз Г. Компьютерная математика.- М.: Наука, 1990.
5. Сигорский В.П. Математический аппарат инженера.-К.: Техника, 1977.
6. Кузичев А.С. Диаграммы Венная.- М.: Наука, 1968.
7. Кононюк А.Ю. Вища математика. В 2 ч. Ч.1,2 - К:КТМ, 2007.
7. Кононюк А.Е. Дискретная математика. В 2 ч. Ч.1,2 - К:Освіта України, 2010.
8. Аверкин А.Н., Батыршин И.З. и др. Нечеткие множества в моделях управления и икусственного интелекта.- М.: Наука, 1986.
9. Кофман А. Введение в теорию нечетких множеств.- М.: Радио и связь, 1982.
10. Кофман А. Введение в прикладную комбинаторику.- М.: Наука, 1975.
11. Згуровский М.З. Интегрированные системы оптимального управления и проектирования.- К.: Высшая школа, 1990.
13. Минский М. Фреймы для представления знаемый. - М.:Энергия, 1979.
14. Вильсон А. Дж. Энтропийные методы моделирования сложных систем.- М.: Наука, 1978.
15. В.Ю. Юрков, О.В. Лукина /Прикладная геометрия, вып. 8, N 18 (2006), стр. 9-36.
16. И.И. Ежев, А.В. Скороход, М.И. Ядренко. Элементы комбинаторики.-М.: Наука, 1977.
17. Алгебраическая теория автоматов, языков и полугрупп. Под ред.
М.А. Арбиба. - М.: Статистика, 1975.
18. Минк Х. Перманенты. М.: Мир, 1982.
19. Холл. М. М.: Мир, 1970.
20. Алефельд Г., Херцбергер Ю. Введение в интервальные вычисления. М.: Мир, 1987.

21. Борисов А. Н., Алексеев А. В. Нечеткие алгоритмы в ситуационных моделях управления организационными системами. — В кн.: Методика построения систем ситуационного управления /Науч. совет АН СССР по комплексной проблеме «Кибернетика». — М., 1978, с. 3—10.

22. Борисов А. Н., Аппен Е. П. Оценка возможных характеристик при анализе альтернатив. — В кн.: Методы принятия решений в условиях неопределенности. — Рига: РПИ, 1980, с. 94—100.

23. Борисов А. Н., Голендер В. Е. Оптимальное разделение размытых образов — Методы и средства технической кибернетики. — Рига: РПИ, 1969, вып. 5, с. 32—38.

24. Борисов А. Н., Корнеева Г. В. Лингвистический подход к построению моделей принятия решений в условиях неопределенности. — В кн.: Методы принятия решений в условиях неопределенности. — Рига: РПИ, 1980, с. 4—12.

24. Борисов А. Н., Крумберг О. А. Анализ решений при выборе технологических объектов. — Там же, с. 127—134.

Научно-практическое издание

Кононюк Анатолий Ефимович

Дискретная математика

Книга 4
Алгебры
Часть 2

Авторская редакция

Подписано в печать 21.01.2011 г.
Формат 60x84/16.
Усл. печ. л. 16,5. Тираж 300 экз.

Издатель и изготовитель:
Издательство «Освита Украины»
04214, г. Киев, ул. Героев Днепра, 63, к. 40
Свидетельство о внесении в Государственный реестр
издателей ДК №1957 от 23.04.2009 г.
Тел./факс (044) 411-4397; 237-5992
E-mail: osvita2005@ukr.net, www.rambook.ru