И. С. БЕРЕЗИН и Н. П. ЖИДКОВ

МЕТОДЫ ВЫЧИСЛЕНИЙ

ТОМ ВТОРОЙ

Допущено
Министерством высшего образования СССР
в качестве учебного пособия
для высших учебных заведений

ГОСУДАРСТВЕННОЕ ИЗДАТЕЛЬСТВО
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ
МОСКВА 1959

АННОТАЦИЯ

Во втором томе книги рассмотрены численные методы решения систем линейных алгебраических уравнений, уравнений высших степеней и трансцендентных уравнений, численные методы отыскания собственных значений, приближенные методы решения обыкновенных дифференциальных уравнений, уравнений в частных производных и интегральных уравнений.

Книга предназначена в качестве учебного пособия для студентов механико-математических и физико-математических факультетов, специализирующихся по вычислительной математике, и лиц, интересующихся теорией и практикой численных методов.

Иван Семенович Березин и *Николай Петрович Жидков*

МЕТОДЫ ВЫЧИСЛЕНИЙ, т. II

Редакторы *Б. М. Будак* и *А. Д. Горбунов*.

Техн. редактор *Н. Я. Мурашова*. Корректор *А. С. Бакулова*.

Сдано в набор 16/VII 1959 г. Подписано к печати 3/XI 1959 г. Бумага $60\times92^1/_{16}$.
Физ. печ. л. 38,75. Условн. печ. л. 38,75 Уч.-изд. л. 41,96. Тираж 10 000.
Т-11038. Цена книги 14 р. 10 к. Заказ № 604

Государственное издательство физико-математической литературы.
Москва, В-71, Ленинский проспект, 15

Типография № 2 им. Евг. Соколовой УПП Ленсовнархоза.
Ленинград, Измайловский пр., 29.

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

В соответствии со сказанным в предисловии к книге, помещенном в первом томе, второй том содержит главы 6—10, что соответствует второй части курса «Методов вычислений», читаемой для студентов 4-го года обучения.

И. С. Березин, Н. П. Жидков

ГЛАВА 6

РЕШЕНИЕ СИСТЕМ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

§ 1. Классификация методов

Главой о численных методах решения систем линейных алгебраических уравнений мы начинаем второй раздел нашей книги, посвященный решению алгебраических, трансцендентных, дифференциальных и интегральных уравнений. Системы линейных алгебраических уравнений наиболее просты и в то же время к ним приводятся многие задачи численного анализа.

Известное из курса высшей алгебры правило Крамера для решения систем линейных алгебраических уравнений практически невыгодно, так как требует слишком большого количества арифметических операций и записей. Поэтому было предложено много различных способов, более пригодных для практики. И в настоящее время значительная часть литературы по вычислительной математике посвящена этому вопросу. Однако сейчас еще нельзя указать один или несколько способов наиболее эффективных в смысле быстроты получения решения с нужной точностью и минимального использования запоминающих устройств. Требуется большая и тщательная теоретическая и экспериментальная сравнительная оценка многочисленных известных способов с этой точки зрения. Мы дадим в этой главе лишь несколько уже давно испытанных методов и несколько методов, имеющих с нашей точки зрения перспективы для практики.

Используемые практически методы решения систем линейных алгебраических уравнений можно разделить на две большие группы: так называемые *точные методы* и *методы последовательных приближений*. Точные методы характеризуются тем, что с их помощью принципиально возможно, проделав конечное число операций, получить точные значения неизвестных. При этом, конечно, предполагается, что коэффициенты и правые части системы известны точно, а все вычисления производятся без округлений. Чаще всего они осуществляются в два этапа. На первом этапе преобразуют систему к тому или иному простому виду. На втором этапе решают упрощенную систему и получают значения неизвестных.

Методы последовательных приближений характеризуются тем, что с самого начала задаются какими-то приближенными значениями

неизвестных. Из этих приближенных значений тем или иным способом получают новые «улучшенные» приближенные значения. С новыми приближенными значениями поступают точно так же и т. д. При выполнении определенных условий можно придти, вообще говоря, после бесконечного числа шагов к точному решению.

Под нашу классификацию не подходят способы решения по методу Монте-Карло. Так названы методы, использующие случайные величины, математические ожидания которых дают решение системы. Пока методы Монте-Карло не могут соревноваться с другими методами, названными выше. Поэтому мы не будем ими здесь заниматься.

§ 2. Метод исключения

Мы начнем изучение численных методов решения систем линейных алгебраических уравнений с точных методов. Простейшим из таких методов является метод исключения.

С методом исключения мы сталкивались уже в обычном школьном курсе алгебры. Комбинируя каким-либо образом уравнения системы, добиваются того, что во всех уравнениях, кроме одного, будет исключено одно из неизвестных. Затем исключают другое неизвестное, третье и т. д. В результате получаем систему с треугольной или диагональной матрицей, решение которой не представляет труда. Метод исключений не вызывает каких-либо теоретических затруднений. Однако точность результата и затрачиваемое на его получение время будут во многом зависеть от организации вычислений. Этому вопросу мы и уделим основное внимание.

Рассмотрим ряд схем, осуществляющих метод исключения на примере системы четырех уравнений с четырьмя неизвестными:

$$\left.\begin{aligned} 1{,}1161\,x_1 + 0{,}1254\,x_2 + 0{,}1397\,x_3 + 0{,}1490\,x_4 &= 1{,}5471,\\ 0{,}1582\,x_1 + 1{,}1675\,x_2 + 0{,}1768\,x_3 + 0{,}1871\,x_4 &= 1{,}6471,\\ 0{,}1968\,x_1 + 0{,}2071\,x_2 + 1{,}2168\,x_3 + 0{,}2271\,x_4 &= 1{,}7471,\\ 0{,}2368\,x_1 + 0{,}2471\,x_2 + 0{,}2568\,x_3 + 1{,}2671\,x_4 &= 1{,}8471. \end{aligned}\right\} \quad (1)$$

Каждой схеме мы припишем то или иное название. Правда, эти названия не являются общепринятыми.

1. Схема Гаусса с выбором главного элемента. Если вычисления производятся не с помощью автоматических вычислительных машин, то удобно нашу систему записать в следующую схему:

№ п/п.	m_i	a_{i1}	a_{i2}	a_{i3}	a_{i4}	b_i	s_i
1		0,11610	0,12540	0,13970	0.14900	1,54710	3,07730
2		0,15820	1,16750	0,17680	0,18710	1,64710	3,33670
3		0,19680	0,20710	1,21680	0,22710	1,74710	3,59490
4		0,23680	0,24710	0,25680	1,26710	1,84710	3,85490

В первом столбце мы записываем номера уравнений. Значение второго столбца будет ясно дальше. 3-й, 4-й, 5-й и 6-й столбцы содержат коэффициенты уравнений, а 7-й столбец — свободные члены. Последний столбец содержит суммы коэффициентов и свободных членов данной строки.

Выбираем теперь наибольший по модулю элемент a_{ij}. Будем называть его *главным*. В нашем случае это $a_{44} = 1{,}26710$. Он в схеме подчеркнут. Делим все элементы столбца, в котором находится главный элемент (в нашем случае a_{i4}), на главный элемент и отношения с обратным знаком помещаем в столбце m_i в той же строке, где находится делимое:

№ п/п.	m_i	a_{i1}	a_{i2}	a_{i3}	a_{i4}	b_i	s_i
1	— 0,11759	1,11610	0,12540	0,13970	0,14900	1,54710	3,07730
2	— 0,14766	0,15820	1,16750	0,17680	0,18710	1,64710	3,33670
3	— 0,17923	0,19680	0,20710	1,21680	0,22710	1,74710	3,59490
4		0,23680	0,24710	0,25680	1,26710	1,84710	3,85490

Будем теперь прибавлять к каждой из строк схемы строку, содержащую главный элемент, умноженную на соответствующее m_i. Строку, содержащую главный элемент, в дальнейшем не выписываем. Не выписываем также и столбец, в котором содержится главный элемент, так как он состоит из нулей. В нашем случае получим:

№ п/п.	m_i	a_{i1}	a_{i2}	a_{i3}	a_{i4}	b_i	s_i
1	— 0,09353	1,08825	0,09634	0,10950		1,32990	2,62399
2	— 0,11862	0,12323	1,13101	0,13888		1,37436	2,76748
3		0,15436	0,16281	1,17077		1,41604	2,90398

Мы вписали сюда же значения m_i, которые получатся на следующем шаге. Производим проверку правильности вычислений. Для этого складываем столбцы a_{ij} и b_i и сравниваем со столбцом s_i. Если расхождения в пределах ошибок округления, то считаем вычисления правильными. Если расхождения слишком велики, то повторяем соответствующие вычисления.

В дальнейшем поступаем с нашей таблицей, как и с предыдущей. Выбираем главный элемент (в нашем случае это будет 1,17077) и делим на него элементы того же столбца. Результаты с обратными знаками записываются в столбец m_i (у нас это уже сделано). Затем последовательно умножаем строку, из которой взят главный элемент, на m_i и складываем с соответствующими строками. Производим

проверку и переходим к следующему шагу. Так продолжаем до тех пор, пока у нас не останется одна строка. В нашем примере будем иметь:

№	m_i	a_{i1}	a_{i2}	a_{i3}	a_{i4}	b_i	s_i
1	−0,07296	1,07381	0,08111			1,19746	2,35238
2		0,10492	1,11170			1,20639	2,42301
1		1,06616				1,10944	2,17560

Взяв уравнения, в которых выбирались главные элементы, получим новую систему, эквивалентную данной:

$$\left.\begin{aligned} &1{,}06616\,x_1 = 1{,}10944,\\ &0{,}10492\,x_1 + 1{,}11170\,x_2 = 1{,}20639,\\ &0{,}15436\,x_1 + 0{,}16281\,x_2 + 1{,}17077\,x_3 = 1{,}41604,\\ &0{,}23680\,x_1 + 0{,}24710\,x_2 + 0{,}25680\,x_3 + 1{,}26710\,x_4 = 1{,}84710. \end{aligned}\right\} \quad (2)$$

Матрица новой системы треугольная. Решение такой системы не встретит затруднений. Находим:

$$\left.\begin{aligned} x_1 &= \frac{1{,}10944}{1{,}06616} = 1{,}04059,\\ x_2 &= \frac{1{,}20639 - 0{,}10492 \cdot 1{,}04059}{1{,}11170} = \frac{1{,}09721}{1{,}11170} = 0{,}98697,\\ x_3 &= \frac{1{,}41604 - 0{,}15436 \cdot 1{,}04059 - 0{,}16281 \cdot 0{,}98697}{1{,}17077} = \frac{1{,}09473}{1{,}17077} = 0{,}93505,\\ x_4 &= \frac{1{,}84710 - 0{,}23680 \cdot 1{,}04059 - 0{,}24710 \cdot 0{,}98697 - 0{,}25680 \cdot 0{,}93505}{1{,}26710} =\\ &= \frac{1{,}11669}{1{,}26710} = 0{,}88130. \end{aligned}\right\} \quad (3)$$

Приведенную нами схему исключения неизвестных назовем *схемой исключения Гаусса с выбором главного элемента*. Сам процесс исключения называют *прямым ходом*, а решение системы с треугольной матрицей — *обратным ходом*. При практическом использовании схемы Гаусса не следует, конечно, разрывать отдельные этапы, как это сделано нами для облегчения объяснений. Не следует также выписывать и окончательную систему уравнений.

Контроль обратного хода осуществляется с помощью столбца s_i. Если в окончательной системе заменить b_i на s_i, то должны получить вместо x_i величины $x_i + 1$. Проверка полученных результатов может быть сделана также путем подстановки их в исходную систему уравнений. В нашем случае получим последовательно в левых частях равенства: 1,54711; 1,64712; 1,74710; 1,84711. Как мы видим, рас-

хождения между правыми и левыми частями не превосходят двух единиц пятого десятичного знака, что нужно считать удовлетворительным.

Смысл выбора главного элемента состоит в том, чтобы сделать возможно меньшими m_i и тем самым уменьшить вычислительную погрешность.

Как известно, при работе на цифровых машинах, да и при работе вручную, наибольшее количество времени затрачивается на производство действий умножения и деления. Поэтому важно знать, сколько действий умножения и деления потребуется для решения заданной системы. Если система имеет порядок n, то после выбора главного элемента нужно произвести $n-1$ делений для определения коэффициентов m_i. Затем нужно умножить строку, содержащую главный элемент, на каждый из этих множителей. Для этого потребуется $(n+1)(n-1)=n^2-1$ умножений. Таким образом, первый шаг работы по схеме Гаусса требует n^2+n-2 умножений и делений. Следующий шаг потребует $(n-1)^2+(n-1)-2$ таких операций и т. д. Всего до обратного хода нужно произвести

$$[n^2+n-2]+[(n-1)^2+(n-1)-2]+\ldots$$
$$\ldots+[1^2+1-2]=\frac{n(n+1)(2n+1)}{6}+\frac{n(n+1)}{2}-2n \qquad (4)$$

операций умножения и деления. Для обратного хода потребуется

$$1+2+3+\ldots+n=\frac{n(n+1)}{2} \qquad (5)$$

операций умножения и деления, если не производить контроля с столбцом s_i. Столько же операций потребуется при использовании такого контроля. Итак, всего на решение системы n уравнений по схеме Гаусса с выбором главного элемента и текущим контролем потребуется

$$\frac{n(n+1)(2n+1)}{6}+\frac{n(n+1)}{2}-2n+n(n+1)=\frac{n}{3}(n^2+6n-1) \qquad (6)$$

операций умножения и деления.

2. Компактная схема Гаусса. Придумано много различных видоизменений схемы Гаусса, дающих те или иные преимущества. Приведем одну такую схему. Рассмотрим сначала систему четырех уравнений общего вида:

$$\left.\begin{aligned}
a_{11}^{(0)}x_1+a_{12}^{(0)}x_2+a_{13}^{(0)}x_3+a_{14}^{(0)}x_4&=a_{15}^{(0)},\\
a_{21}^{(0)}x_1+a_{23}^{(0)}x_2+a_{23}^{(0)}x_3+a_{24}^{(0)}x_4&=a_{25}^{(0)},\\
a_{31}^{(0)}x_1+a_{32}^{(0)}x_2+a_{33}^{(0)}x_3+a_{34}^{(0)}x_4&=a_{35}^{(0)},\\
a_{41}^{(0)}x_1+a_{42}^{(0)}x_2+a_{43}^{(0)}x_3+a_{44}^{(0)}x_4&=a_{45}^{(0)}.
\end{aligned}\right\} \qquad (7)$$

Исходные данные, промежуточные и окончательные результаты будем записывать в следующую схему:

$a_{11}^{(0)}$	$a_{12}^{(0)}$	$a_{13}^{(0)}$	$a_{14}^{(0)}$	$a_{15}^{(0)}$
$a_{21}^{(0)}$	$a_{22}^{(0)}$	$a_{23}^{(0)}$	$a_{24}^{(0)}$	$a_{25}^{(0)}$
$a_{31}^{(0)}$	$a_{32}^{(0)}$	$a_{33}^{(0)}$	$a_{34}^{(0)}$	$a_{35}^{(0)}$
$a_{41}^{(0)}$	$a_{42}^{(0)}$	$a_{43}^{(0)}$	$a_{44}^{(0)}$	$a_{45}^{(0)}$
$b_{11}^{(1)}$	$c_{12}^{(2)}$	$c_{13}^{(2)}$	$c_{14}^{(2)}$	$c_{15}^{(2)}$
$b_{21}^{(1)}$	$b_{22}^{(3)}$	$c_{23}^{(4)}$	$c_{24}^{(4)}$	$c_{25}^{(4)}$
$b_{31}^{(1)}$	$b_{32}^{(3)}$	$b_{33}^{(5)}$	$c_{34}^{(6)}$	$c_{35}^{(6)}$
$b_{41}^{(1)}$	$b_{42}^{(3)}$	$b_{43}^{(5)}$	$b_{44}^{(7)}$	$c_{45}^{(8)}$
$x_1^{(12)}$	$x_2^{(11)}$	$x_3^{(10)}$	$x_4^{(9)}$	

Верхнюю половину схемы мы отводим для коэффициентов и свободных членов исходной системы, а в нижней половине будут помещаться промежуточные и окончательные результаты. Верхний индекс показывает порядок получения промежуточных и окончательных результатов.

Величины $b_{i1}^{(1)}$ просто совпадают с соответствующими величинами $a_{i1}^{(0)}$ и выписываются здесь лишь для удобства пользования схемой.

Величины $c_{1j}^{(2)}$ вычисляем по формулам

$$c_{1j}^{(2)} = \frac{a_{1j}^{(0)}}{b_{11}^{(1)}} \qquad (j = 2,\ 3,\ 4,\ 5). \tag{8}$$

При этом уравнение

$$x_1 + c_{12}^{(2)} x_2 + c_{13}^{(2)} x_3 + c_{14}^{(2)} x_4 = c_{15}^{(2)} \tag{9}$$

эквивалентно первому уравнению исходной системы.

После этого вычисляем величины $b_{i2}^{(3)}$ $(i > 1)$ по формулам

$$b_{i2}^{(3)} = a_{i2}^{(0)} - b_{i1}^{(1)} c_{12}^{(2)} \qquad (i = 2,\ 3,\ 4). \tag{10}$$

Таким образом, $b_{i2}^{(3)}$ будут являться коэффициентами при x_2 во втором, третьем и четвертом уравнениях системы после исключения в них неизвестного x_1 с помощью уравнения (9).

Следующим этапом будет являться получение коэффициентов и правой части второго уравнения после исключения из него указанным выше способом неизвестного x_1 и последующего деления на

коэффициент при x_2. Очевидно, эти величины $c_{2j}^{(4)}$ будут определяться по формулам:

$$c_{2j}^{(4)} = \frac{a_{2j}^{(0)} - b_{21}^{(1)} c_{1j}^{(2)}}{b_{22}^{(3)}} \qquad (j = 3,\ 4,\ 5). \tag{11}$$

Таким образом, после преобразования второе уравнение примет вид

$$x_2 + c_{23}^{(4)} x_3 + c_{24}^{(4)} x_4 = c_{25}^{(4)}. \tag{12}$$

Далее будем исключать неизвестное x_2 из третьего и четвертого уравнений. Опять сначала подсчитываем коэффициенты при x_3 в третьем и четвертом уравнениях. Они определятся по формулам:

$$b_{i3}^{(5)} = a_{i3}^{(0)} - b_{i1}^{(1)} c_{13}^{(2)} - b_{i2}^{(3)} c_{23}^{(4)} \qquad (i = 3,\ 4). \tag{13}$$

Первые два члена правой части этой формулы дают коэффициенты при x_3 третьего и четвертого уравнений после исключения x_1, а после вычитания последнего члена получим результат исключения x_2.

Коэффициенты и правая часть третьего уравнения после деления на коэффициент при x_3 примут вид

$$c_{3j}^{(6)} = \frac{a_{3j}^{(0)} - b_{31}^{(1)} c_{1j}^{(2)} - b_{32}^{(3)} c_{2j}^{(4)}}{b_{33}^{(5)}} \qquad (j = 4,\ 5), \tag{14}$$

а само это уравнение запишется в виде

$$x_3 + c_{34}^{(6)} x_4 = c_{35}^{(6)}. \tag{15}$$

Остается еще исключить x_3 из четвертого уравнения. При этом коэффициент при x_4 в нем примет вид

$$b_{44}^{(7)} = a_{44}^{(0)} - b_{41}^{(1)} c_{14}^{(2)} - b_{42}^{(3)} c_{24}^{(4)} - b_{43}^{(5)} c_{34}^{(6)}, \tag{16}$$

а свободный член после исключения x_3 и деления на коэффициент при x_4 будет

$$c_{45}^{(8)} = \frac{a_{45}^{(0)} - b_{41}^{(1)} c_{15}^{(2)} - b_{42}^{(3)} c_{25}^{(4)} - b_{43}^{(5)} c_{35}^{(6)}}{b_{44}^{(7)}}. \tag{17}$$

При этом четвертое уравнение запишется в виде

$$x_4 = c_{45}^{(8)}. \tag{18}$$

Нетрудно заметить, что для системы n уравнений при отыскании величин $b_{ij}^{(2j-1)}$, $c_{jl}^{(2j)}$ следует поочередно использовать формулы:

$$b_{ij}^{(2j-1)} = a_{ij}^{(0)} - \sum_{k=1}^{j-1} b_{ik}^{(2k-1)} c_{kj}^{(2k)} \qquad (i = j,\ j+1,\ \ldots,\ n), \tag{19}$$

$$c_{jl}^{(2j)} = \frac{a_{jl}^{(0)} - \sum_{k=1}^{j-1} b_{jk}^{(2k-1)} c_{kl}^{(2k)}}{b_{jj}^{(2j-1)}} \qquad (l = j+1,\ j+2,\ \ldots,\ n+1). \tag{20}$$

Если проследить ход вычислений по схеме, то легко обнаружить закон образования величин $b_{ij}^{(2j-1)}$ и $c_{ji}^{(2j)}$.

Неизвестные x_n, x_{n-1}, ..., x_1 находятся последовательно из системы уравнений

$$x_i + \sum_{k=i+1}^{n} c_{ik}^{(2i)} x_k = c_{i,n+1}^{(2i)} \qquad (i = n,\ n-1,\ \ldots,\ 1). \tag{21}$$

Будем называть эту схему *компактной схемой Гаусса*. При решении системы n уравнений по компактной схеме Гаусса требуется произвести столько же умножений и делений, как и в схеме главных элементов. Однако она требует меньше записей. Схема допускает такой же контроль, как и ранее.

Так как вычисления по компактной схеме Гаусса более систематизированы, чем по схеме главных элементов, то процесс вычислений легче программируется для автоматических машин. С другой стороны, вычисления по этой схеме могут привести к большой потере точности. Кроме того, для того чтобы процесс вычислений был осуществим, нужно требовать отличие от нуля всех $b_{ii}^{(2i-1)}$.

Приведем результаты вычислений при решении приведенной в начале параграфа системы (1) по компактной схеме Гаусса:

1,11610	0,12540	0,13970	0,14900	1,54710	3,07730
0,15820	1,16750	0,17680	0,18710	1,64710	3,33670
0,19680	0,20710	1,21680	0,22710	1,74710	3,59490
0,23680	0,24710	0,25680	1,26710	1,84710	3,85490
1,11610	0,11236	0,12517	0,13350	1,38617	2,75720
0,15820	1,14972	0,13655	0,14437	1,24187	2,52279
0,19680	0,18499	1,16691	0,14921	1,06655	2,21576
0,23680	0,22049	0,19705	1,17425	0,88130	1,88130
1,04058	0,98696	0,93505	0,88130		

Применяя компактную схему Гаусса, мы элементарными преобразованиями переводим матрицу A системы в верхнюю треугольную матрицу

$$C = \begin{pmatrix} 1 & c_{12}^{(2)} & c_{13}^{(2)} & c_{14}^{(2)} & \ldots & c_{1n}^{(2)} \\ 0 & 1 & c_{23}^{(4)} & c_{24}^{(4)} & \ldots & c_{2n}^{(4)} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & & 1 \end{pmatrix}. \tag{22}$$

Интересно отметить, что если рассмотреть еще матрицу

$$B = \begin{pmatrix} b_{11}^{(1)} & 0 & 0 & 0 & \ldots & 0 \\ b_{21}^{(1)} & b_{22}^{(3)} & 0 & 0 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ b_{n1}^{(1)} & b_{n2}^{(3)} & b_{n3}^{(5)} & b_{n4}^{(7)} & \ldots & b_{n,n}^{(2n-1)} \end{pmatrix}, \tag{23}$$

также получающуюся в процессе наших вычислений, то имеет место равенство

$$A = BC. \tag{24}$$

Это равенство просто получается, если использовать формулы (19) и (20). Так как B^{-1} снова треугольная матрица, то прямой ход при решении системы уравнений по компактной схеме Гаусса эквивалентен умножению системы на треугольную матрицу.

3. Обращение матрицы. Равенство (24) можно использовать для *обращения матриц*. Эта задача важна как сама по себе, так и в тех случаях, когда приходится решать много систем с одной и той же матрицей, но с различными правыми частями.

Перепишем (24) следующим образом:

$$A^{-1}B = C^{-1}, \quad CA^{-1} = B^{-1}. \tag{25}$$

Матрица C^{-1} будет верхней треугольной, и ее диагональные элементы равны единице. Поэтому если обозначить элементы матрицы A^{-1} через d_{ij}, то первое из равенств (25) даст $\frac{n(n+1)}{2}$ уравнений для определения d_{ij}:

$$\left.\begin{array}{rl} i = & 1 \quad 2 \quad 3 \ldots n, \\ \sum\limits_{j=1}^{n} d_{ij} b_{j1}^{(1)} = & 1 \quad 0 \quad 0 \ldots 0, \\ \sum\limits_{j=2}^{n} d_{ij} b_{j2}^{(3)} = & \quad 1 \quad 0 \ldots 0, \\ \ldots\ldots\ldots\ldots & \\ d_{nn} b_{nn}^{(2n-1)} = & \quad\quad\quad 1. \end{array}\right\} \tag{26}$$

Так как матрица B^{-1} нижняя треугольная, то второе из равенств (25) даст еще $\frac{n(n-1)}{2}$ уравнений для отыскания d_{ij}:

$$\left.\begin{array}{rl} j = & 2 \quad 3 \ldots n, \\ d_{1j} + \sum\limits_{i=2}^{n} d_{ij} c_{1i}^{(2)} = & 0 \quad 0 \ldots 0, \\ d_{2j} + \sum\limits_{i=3}^{n} d_{ij} c_{2i}^{(4)} = & \quad 0 \ldots 0, \\ \ldots\ldots\ldots\ldots & \\ d_{n-1,\,n} + d_{nn} c_{n-1,\,n}^{(2n-2)} = & \quad\quad 0. \end{array}\right\} \tag{27}$$

Таким образом, мы получили n^2 уравнений для определения неизвестных элементов d_{ij} обратной матрицы A^{-1}. Решение системы (26) — (27) не представляет труда. Полагаем в каждом из уравнений (26) $i=n$ и находим последовательно $d_{n,n}$, $d_{n,n-1}$, ..., $d_{n,1}$. Затем из уравнений (27) при $j=n$ находим $d_{n-1,n}$, $d_{n-2,n}$, ..., $d_{1,n}$. Потом снова возвращаемся к уравнениям (26) и, полагая в них $i=n-1$, находим $d_{n-1,n-1}$, $d_{n-1,n-2}$, ..., $d_{n-1,1}$. После этого из уравнений (27) при $j=n-1$ находим $d_{n-2,n-1}$, ..., $d_{1,n-1}$. Так, переходя поочередно от системы (26) к (27) и наоборот, мы в конце концов найдем все d_{ij}. Вычислительная схема остается прежней, только вместо одной строки для неизвестных x_i у нас будет n строк для матрицы A^{-1}.

Если воспользоваться результатами, полученными при решении примера (1) по компактной схеме Гаусса, то без труда найдем, что матрица, обратная к матрице данной системы, будет такова:

$$A^{-1}=\begin{pmatrix} 0{,}93794 & -0{,}06844 & -0{,}07961 & -0{,}08592 \\ -0{,}08852 & 0{,}90599 & -0{,}09919 & -0{,}10560 \\ -0{,}11135 & -0{,}11697 & 0{,}87842 & -0{,}12707 \\ -0{,}13546 & -0{,}14018 & -0{,}14381 & 0{,}85161 \end{pmatrix}. \tag{28}$$

Интересно заметить, что при этом оказывается

$$A^{-1}A=\begin{pmatrix} 0{,}99999 & 0{,}00000 & 0{,}00000 & 0{,}00000 \\ 0{,}00000 & 1{,}00001 & 0{,}00000 & -0{,}00001 \\ 0{,}00000 & 0{,}00000 & 0{,}99999 & 0{,}00000 \\ 0{,}00000 & 0{,}00000 & 0{,}00000 & 1{,}00000 \end{pmatrix}. \tag{29}$$

Точность вполне удовлетворительна. Если бы точность нас не удовлетворила, то для уточнения можно было бы применить какой-либо метод последовательного приближения, о чем будем говорить позже.

4. Вычисление определителей. Схемы Гаусса можно применить для вычисления определителей. При этом никаких новых трудностей не возникает, поэтому только кратко опишем для примера применение для этой цели схемы Гаусса с выбором главного элемента.

Анализируя схему Гаусса с выбором главного элемента для решения системы уравнений, легко убедиться, что при выполнении прямого хода мы совершаем такие преобразования матрицы коэффициентов исходной системы, при которых величина ее определителя не изменяется. После завершения прямого хода получается

система с такой матрицей, которую путем перестановки строк и столбцов можно преобразовать в треугольную матрицу, причем на главной диагонали будут стоять наши главные элементы. Следовательно, определитель матрицы исходной системы только знаком может отличаться от произведения главных элементов. С каким знаком нужно брать это произведение, легко сообразить, не выполняя преобразование матрицы к треугольному виду.

Эти рассуждения показывают, что для вычисления определителя по схеме Гаусса с выбором главного элемента нужно в точности повторить прямой ход для решения системы по этой схеме (не выполняя действий со столбцом свободных членов), а затем взять с соответствующим знаком произведение главных элементов.

5. Схема Жордана. Когда мы решали систему по схеме Гаусса, то на каждом шаге число уравнений уменьшалось на единицу. Будем теперь оставлять все уравнения, но при выборе главного элемента не будем учитывать коэффициенты тех уравнений, из которых уже выбирался главный элемент. Получим новую схему, которую будем называть *схемой Жордана*. Так как здесь нет по существу ничего нового, то мы лишь проиллюстрируем эту схему на том же самом примере:

№ п/п.	m_i	a_{i1}	a_{i2}	a_{i3}	a_{i4}	b_i	s_i
1	— 0,11759	1,11610	0,12540	0,13970	0,14900	1 54710	3,07730
2	— 0,14766	0,15820	1,16750	0,17680	0,18710	1,64710	3,33670
3	— 0,17923	0,19680	0,20710	1,21680	0,22710	1,74710	3,59490
4		0,23680	0,24710	0,25680	1,26710	1,84710	3,85490
1	— 0,09353	1,08825	0,09634	0,10950		1,32990	2,62399
2	— 0,11862	0,12323	1,13101	0,13888		1,37436	2,76748
3		0,15436	0,16281	1,17077		1,41604	2,90398
4	— 0,21934	0,23680	0,24710	0,25680	1,26710	1,84710	3,85490
1	— 0,07296	1,07381	0,08111			1,19746	2,35238
2		0,10492	1,11170			1,20639	2,42301
3	— 0,14645	0,15436	0,16281	1,17077		1,41604	2,90398
4	— 0,19015	0,20294	0,21139		1,26710	1,53651	3,21794
1		1,06616				1,10944	2,17560
2	— 0,09841	0,10492	1,11170			1,20639	2,42301
3	— 0,13037	0,13899		1,17077		1,23936	2,54912
4	— 0,17163	0,18299			1,26710	1,30711	2,75720
1		1,06616				1,10944	2,17560
2			1,11170			1,09721	2,20891
3				1,17077		1,09472	2,26549
4					1,26710	1,11670	2,38380

Отсюда без труда находим:

$$x_1 = 1{,}04059; \quad x_2 = 0{,}98697; \quad x_3 = 0{,}93504; \quad x_4 = 0{,}88130. \tag{30}$$

При решении по схеме Жордана система приводится к диагональному виду и обратный ход значительно облегчается. Всего для решения системы n уравнений с текущим контролем по схеме Жордана требуется произвести

$$\frac{n^3 + 4n^2 - n}{2} \tag{31}$$

операций умножения и деления.

Каждый шаг прямого хода по схеме Жордана эквивалентен умножению системы уравнений слева на матрицу вида

$$\begin{pmatrix} 1 & 0 & \dots & a_{1i} & 0 & \dots & 0 \\ 0 & 1 & \dots & a_{2i} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & a_{i+1,\,i} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_{n,\,i} & 0 & \dots & 1 \end{pmatrix}, \tag{32}$$

если не выбирать главный элемент, а идти последовательно, исключая x_1, x_2, ... с помощью первого, второго, третьего, ... уравнений.

6. Схема без обратного хода. Приведем еще одну схему исключения, при которой вообще не требуется обратного хода. Пусть дана система уравнений

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n &= b_2, \\ \dots\dots\dots\dots\dots\dots\dots\dots & \\ \dots\dots\dots\dots\dots\dots\dots\dots & \\ a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n &= b_n. \end{aligned}\right\} \tag{33}$$

Рассмотрим матрицу

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} & b_n \\ -1 & 0 & \dots & 0 & 0 \end{pmatrix}. \tag{34}$$

Подберем постоянные y_1, y_2, ..., y_n так, чтобы после умножения первой строки матрицы на y_1, второй на y_2, ..., n-й на y_n и сло-

жения их с последней строкой в этой последней все элементы, кроме крайнего правого, обратились бы в нули. Таким образом, $y_1, y_2, \ldots, y_n$ должны удовлетворять системе уравнений

$$\left.\begin{aligned} a_{11}y_1 + a_{21}y_2 + \ldots + a_{n1}y_n &= 1, \\ a_{12}y_1 + a_{22}y_2 + \ldots + a_{n2}y_2 &= 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{1n}y_1 + a_{2n}y_2 + \ldots + a_{nn}y_n &= 0. \end{aligned}\right\} \tag{35}$$

Умножая обе части последних равенств последовательно на $x_1, x_2, \ldots, x_n$, где $(x_1, x_2, \ldots, x_n)$ — решение исходной системы, и складывая их, получим:

$$\begin{aligned} &y_1(a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n) + \\ &+ y_2(a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n) + \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &+ y_n(a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n) = x_1. \end{aligned} \tag{36}$$

Но выражения в скобках равны последовательно $b_1, b_2, \ldots, b_n$. Таким образом,

$$b_1y_1 + b_2y_2 + \ldots + b_ny_n = x_1, \tag{37}$$

т. е. правый крайний элемент последней строки будет равен x_1. Совершенно аналогично, если бы мы взяли —1 во втором столбце последней строки, то крайний правый элемент последней строки стал бы равен x_2. Такие же рассуждения можно провести, беря —1 в любом из первых n столбцов. Рассмотрим теперь клеточную матрицу

$$\begin{pmatrix} A & \vdots & b \\ \cdot & \vdots & \cdot\cdot\cdot\cdot \\ -I & \vdots & 0 \end{pmatrix}, \tag{38}$$

где A — матрица коэффициентов системы, b — столбец свободных членов, I — единичная матрица и 0 — столбец нулей. Пусть нам удалось путем прибавления к последним n строкам этой матрицы линейных комбинаций n первых строк сделать минус единичную матрицу нулевой. Тогда в силу наших рассуждений нулевой столбец обратится в столбец неизвестных $x_1, x_2, x_3, \ldots, x_n$. Схема, которую мы сейчас проиллюстрируем на том же примере, и будет реализацией указанного процесса. За основу будет взята схема Гаусса с выбором главного элемента. Будем называть данную схему *схемой без обратного хода*.

№	m_i	a_{i1}	a_{i2}	a_{i3}	a_{i4}	b_i	s_i
1	— 0,11759	1,11610	0,12540	0,13970	0,14900	1,54710	3,07730
2	— 0,14766	0,15820	1,16750	0,17680	0,18710	1,64710	3,36670
3	— 0,17923	0,19680	0,20710	1,21680	0,22710	1,74710	3,59490
4		0,23680	0,24710	0,25680	1,26710	1,84710	3,85490
IV	0,78920				— 1		— 1
1	— 0,09353	1,08825	0,09634	0,10950		1,32990	2,62399
2	— 0,11862	0,12323	1,13101	0,13888		1,37436	2,76748
3		0,15436	0,16281	1,17077		1,41604	2,90398
IV	— 0,17311	0,18688	0,19501	0,20267		1,45773	2,04229
III	0,85414			— 1			— 1
1	— 0,07296	1,07381	0,08111			1,19746	2,35238
2		0,10492	1,11170			1,20639	2,42301
IV	— 0,15007	0,16016	0,16683			1,21260	1,53959
III	— 0,12509	0,13185	0,13906			1,20950	1,48041
II	0,89952		— 1				— 1
I		1,06616				1,10944	2,17560
IV	— 0,13545	0,14441				1,03156	1,17597
III	— 0,11136	0,11873				1,05859	1,17732
II	— 0,08852	0,09438				1,08517	1,17955
I	0,93795	— 1					— 1
IV						0,88129	0,88128
III						0,93504	0,93505
II						0,98696	0,98697
I						1,04060	1,04060

Значения неизвестных, содержащиеся в последних четырех строках, близки к найденным ранее. При решении системы уравнений по схеме без обратного хода с контрольным столбцом требуется произвести

$$\frac{n^3 + 5n^2}{2} \tag{39}$$

операций умножения и деления.

Приведенная схема допускает очевидные обобщения. Так, если вместо (38) взять

$$\left(\begin{array}{c|c} A & b \\ \hline -C & d \end{array}\right), \tag{40}$$

где C — произвольная квадратная матрица и d — произвольный столбец, то, действуя по схеме без обратного хода, мы получим на месте столбца d столбец $CA^{-1}b+d$.

Если же рассмотреть матрицу

$$\left(\begin{array}{c|c} A & B \\ \hline -C & 0 \end{array}\right), \tag{41}$$

где B — произвольная матрица и 0 — матрица нулей, то нашим процессом мы придем к матрице $CA^{-1}B$. В частности, если взять матрицу

$$\left(\begin{array}{c|c} A & I \\ \hline -I & 0 \end{array}\right), \tag{42}$$

то мы придем к матрице A^{-1}.

§ 3. Метод квадратного корня

В том случае, когда матрица A симметрическая, в приведенных ранее схемах можно сделать ряд упрощений. Мы не будем здесь останавливаться на этих довольно простых вопросах, а изложим вместо этого очень удобный для симметрических матриц *метод квадратного корня*.

Пусть данная нам система записана в виде

$$Ax=b, \tag{1}$$

где A — квадратная симметрическая матрица, b — вектор-столбец из правых частей системы и x — вектор-столбец неизвестных. Решение системы (1) будем осуществлять в два этапа. На первом этапе представим матрицу A в виде

$$A=LL', \tag{2}$$

где L — нижняя треугольная матрица и L' — транспонированная по отношению к L матрица. Такое представление всегда возможно. Чтобы не осложнять записей, ограничимся рассмотрением систем четвертого порядка. Будем разыскивать такие α_{ij}, что

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} = \begin{pmatrix} \alpha_{11} & 0 & 0 & 0 \\ \alpha_{21} & \alpha_{22} & 0 & 0 \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & 0 \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{pmatrix} \begin{pmatrix} \alpha_{11} & \alpha_{21} & \alpha_{31} & \alpha_{41} \\ 0 & \alpha_{22} & \alpha_{32} & \alpha_{42} \\ 0 & 0 & \alpha_{33} & \alpha_{43} \\ 0 & 0 & 0 & \alpha_{44} \end{pmatrix}. \tag{3}$$

Произведя умножение матриц в правой части и приравнивая затем соответствующие элементы правой и левой частей, получим следующие уравнения:

$$\left.\begin{aligned}
&\alpha_{11}^2 = a_{11}, \quad \alpha_{11}\alpha_{21} = a_{12}, \quad \alpha_{11}\alpha_{31} = a_{13}, \quad \alpha_{11}\alpha_{41} = a_{14},\\
&\alpha_{21}^2 + \alpha_{22}^2 = a_{22}, \quad \alpha_{21}\alpha_{31} + \alpha_{22}\alpha_{32} = a_{23}, \quad \alpha_{21}\alpha_{41} + \alpha_{22}\alpha_{42} = a_{24},\\
&\alpha_{31}^2 + \alpha_{32}^2 + \alpha_{33}^2 = a_{33}, \quad \alpha_{31}\alpha_{41} + \alpha_{32}\alpha_{42} + \alpha_{33}\alpha_{43} = a_{34},\\
&\alpha_{41}^2 + \alpha_{42}^2 + \alpha_{43}^2 + \alpha_{44}^2 = a_{44}.
\end{aligned}\right\} \tag{4}$$

Отсюда последовательно находим:

$$\left.\begin{aligned}
&\alpha_{11} = \sqrt{a_{11}}, \quad \alpha_{21} = \frac{a_{12}}{\sqrt{a_{11}}}, \quad \alpha_{31} = \frac{a_{13}}{\sqrt{a_{11}}}, \quad \alpha_{41} = \frac{a_{14}}{\sqrt{a_{11}}},\\
&\alpha_{22} = \sqrt{a_{22} - \alpha_{21}^2}, \quad \alpha_{32} = \frac{a_{23} - \alpha_{21}\alpha_{31}}{\alpha_{22}}, \quad \alpha_{42} = \frac{a_{24} - \alpha_{21}\alpha_{41}}{\alpha_{22}},\\
&\alpha_{33} = \sqrt{a_{33} - \alpha_{31}^2 - \alpha_{32}^2}, \quad \alpha_{43} = \frac{a_{34} - \alpha_{31}\alpha_{41} - \alpha_{32}\alpha_{42}}{\alpha_{33}},\\
&\alpha_{44} = \sqrt{a_{44} - \alpha_{41}^2 - \alpha_{42}^2 - \alpha_{43}^2}.
\end{aligned}\right\} \tag{5}$$

Нетрудно сообразить, как будут выражаться α_{ij} через a_{ij} в общем случае системы n-го порядка.

Нужно заметить, что при действительных a_{ij} могут получиться чисто мнимые значения α_{ij}. Но так как вычисления с чисто мнимыми величинами нисколько не труднее, чем с действительными, это не вызовет дополнительных трудностей. Если, кроме того, матрица A положительно определенная, то мнимых величин вообще не будет.

После того как матрица L найдена, переходят ко второму этапу. При этом сначала решают систему

$$Ly = b, \tag{6}$$

а затем находят x из системы

$$L'x = y. \tag{7}$$

Так как обе системы с треугольными матрицами, то они решаются без труда.

Схема квадратного корня очень удобна, требует небольшого количества операций умножения и деления и очень небольших записей. Всего при решении системы n уравнений придется n раз произвести извлечение корня и проделать

$$\frac{n^3 + 9n^2 + 2n}{6} \tag{8}$$

операций умножения и деления

Проиллюстрируем этот метод на примере системы шести уравнений с симметрической матрицей. Часть коэффициентов мы не выписывали, пользуясь симметрией.

6,1818	0,1818 7,1818 a_{ik}	0,3141 0,2141 8,2435	0,1415 0,1815 0,1214 9,3141	0,1516 0,1526 0,2516 0,3145 5,3116	0,2141 0,3114 0,2618 0,6843 0,8998 4,1313	7,1818 8,2435 9,3141 5,3116 4,1313 3,1816
2,486323	0,073120 2,678891 a_{ik}	0,126331 0,076473 2,867349	0,056911 0,066199 0,038066 3,050415	0,060974 0,055300 0,083585 0,099720 2,299543	0,086111 0,113892 0,084472 0,219198 0,373697 1,978909	2,888522 2,938364 3,041100 1,584361 1,468632 0,726854
1,040932	1,050668	1,026605	0,474071	0,578973	0,367300	y_i / x_i

Подставляя найденные значения в левые части системы, получим соответственно

7,181794; 8,243489; 9,314104; 5,311593; 4,131297; 3,181600. (9)

§ 4. Метод ортогонализации

Пусть дана система

$$A\bar{x} = \bar{b} \tag{1}$$

порядка n. Здесь мы, чтобы избежать в дальнейшем путаницы, над векторами поставили черточки. Решение системы будем разыскивать в виде

$$\bar{x} = \sum_{k=1}^{n} \alpha_k \bar{x}^{(k)}, \tag{2}$$

где $\bar{x}^{(1)}$, $\bar{x}^{(2)}$, ..., $\bar{x}^{(n)}$ — n векторов, удовлетворяющих условиям

$$(A\bar{x}^{(k)}, \bar{x}^{(l)}) = 0, \qquad \text{при } k > l \quad (k, l = 1, 2, \ldots, n). \tag{3}$$

Здесь рассматривается обычное скалярное произведение векторов в n-мерном векторном пространстве, т. е. если $\bar{x} = (x_1, x_2, \ldots, x_n)$ и $\bar{y} = (y_1, y_2, \ldots, y_n)$, то $(\bar{x}, \bar{y}) = \sum_{i=1}^{n} x_i y_i$. Пусть такие векторы найдены. Как это делается, будет показано ниже. Рассмотрим скалярное произведение обеих частей системы (1) с $\bar{x}^{(l)}$:

$$(A\bar{x}, \bar{x}^{(l)}) = (\bar{b}, \bar{x}^{(l)}) \qquad (l = 1, 2, \ldots, n). \tag{4}$$

Используя (2), получим:

$$(A\bar{x},\ \bar{x}^{(l)}) = \sum_{k=1}^{n} \alpha_k (A\bar{x}^{(k)},\ \bar{x}^{(l)}) = (\bar{b},\ \bar{x}^{(l)}) \qquad (l = 1,\ 2,\ \ldots,\ n) \qquad (5)$$

или, в силу выбора векторов $\bar{x}^{(i)}$,

$$\sum_{k=1}^{l} \alpha_k (A\bar{x}^{(k)},\ \bar{x}^{(l)}) = (\bar{b},\ \bar{x}^{(l)}) \qquad (l = 1,\ 2,\ \ldots,\ n). \qquad (6)$$

Итак, для определения коэффициентов α_k мы получили систему с треугольной матрицей. Определитель этой системы равен

$$(A\bar{x}^{(1)},\ \bar{x}^{(1)})\,(A\bar{x}^{(2)},\ \bar{x}^{(2)})\ \ldots\ (A\bar{x}^{(n)},\ \bar{x}^{(n)}). \qquad (7)$$

Следовательно, если $(A\bar{x}^{(k)},\ \bar{x}^{(k)}) \neq 0$ $(k = 1,\ 2.\ \ldots,\ n)$, то α_k возможно найти и находятся они без труда.

Особенно легко определятся α_k, если матрица A симметрическая. В этом случае, очевидно,

$$(A\bar{x}^{(k)},\ \bar{x}^{(l)}) = (\bar{x}^{(k)},\ A\bar{x}^{(l)}) = (A\bar{x}^{(l)},\ \bar{x}^{(k)}) \qquad (8)$$

и, следовательно,

$$(A\bar{x}^{(k)},\ \bar{x}^{(l)}) = 0 \qquad \text{при } l \neq k \quad (l,\ k = 1,\ 2,\ \ldots,\ n). \qquad (9)$$

Тогда система для определения α_k принимает вид

$$\alpha_k (A\bar{x}^{(k)},\ \bar{x}^{(k)}) = (\bar{b},\ \bar{x}^{(k)}) \qquad (10)$$

и

$$\alpha_k = \frac{(\bar{b},\ \bar{x}^{(k)})}{(A\bar{x}^{(k)},\ \bar{x}^{(k)})}\,. \qquad (11)$$

Метод можно обобщить. Пусть каким-то образом удалось найти систему $2n$ векторов $\bar{x}^{(1)},\ \bar{x}^{(2)},\ \ldots,\ \bar{x}^{(n)};\ \bar{y}^{(1)},\ \bar{y}^{(2)},\ \ldots,\ \bar{y}^{(n)}$ так, что

$$(A\bar{x}^{(k)},\ \bar{y}^{(l)}) = 0 \qquad \text{при } k > l \quad (k,\ l = 1,\ 2,\ \ldots,\ n). \qquad (12)$$

Умножая обе части равенства (1) на $\bar{y}^{(r)}$ и используя представление $\bar{x}$ через $\bar{x}^{(k)}$, как и ранее, получим:

$$\sum_{k=1}^{r} \alpha_k (A\bar{x}^{(k)},\ \bar{y}^{(r)}) = (\bar{b},\ \bar{y}^{(r)}) \qquad (r = 1,\ 2,\ \ldots,\ n). \qquad (13)$$

Опять получилась система линейных алгебраических уравнений с треугольной матрицей для определения α_k. Несколько усложнив вычисления, можно получить систему диагонального вида. Для этого построим три системы векторов $\bar{x}^{(1)},\ \bar{x}^{(2)},\ \ldots,\ \bar{x}^{(n)};\ \bar{y}^{(1)},\ \bar{y}^{(2)},\ \ldots,\ \bar{y}^{(n)};$

$\bar{z}^{(1)}, \bar{z}^{(2)}, \ldots, \bar{z}^{(n)}$, так что имеют место равенства:

$$\bar{x}^{(1)} = \bar{y}^{(1)}; \quad \bar{x}^{(k)} = \bar{y}^{(k)} + \sum_{i=1}^{k-1} c_i^{(k)} \bar{y}^{(i)} \qquad (k = 2, 3, \ldots, n), \tag{14}$$

$$\bar{z}^{(1)} = \bar{y}^{(1)}; \quad \bar{z}^{(k)} = \bar{y}^{(k)} + \sum_{i=1}^{k-1} d_i^{(k)} \bar{y}^{(i)} \qquad (k = 2, 3, \ldots, n), \tag{15}$$

$$(A\bar{x}^{(k)}, \bar{y}^{(r)}) = (A'\bar{z}^{(k)}, \bar{y}^{(r)}) = 0 \qquad (k > r). \tag{16}$$

Тогда

$$(\bar{z}^{(r)}, \bar{b}) = \sum_{i=1}^{r-1} \alpha_i (\bar{z}^{(r)}, A\bar{x}^{(i)}) + \alpha_r (\bar{z}^{(r)}, A\bar{x}^{(r)}) + \\ + \sum_{i=r+1}^{n} \alpha_i (\bar{z}^{(r)}, A\bar{x}^{(i)}) = \alpha_r (\bar{z}^{(r)}, A\bar{x}^{(r)}), \tag{17}$$

так как при $i < r$

$$(\bar{z}^{(r)}, A\bar{x}^{(i)}) = \left(\bar{z}^{(r)}, A\bar{y}^{(i)} + \sum_{j=1}^{i-1} c_j^{(i)} A\bar{y}^{(j)}\right) = \\ = (A'\bar{z}^{(r)}, \bar{y}^{(i)}) + \sum_{j=1}^{i-1} c_j^{(i)} (A'\bar{z}^{(r)}, \bar{y}^{(j)}) = 0 \tag{18}$$

и при $i > r$

$$(\bar{z}^{(r)}, A\bar{x}^{(i)}) = \left(\bar{y}^{(r)} + \sum_{j=1}^{r-1} d_j^{(r)} \bar{y}^{(j)}, A\bar{x}^{(i)}\right) = \\ = (A\bar{x}^{(i)}, \bar{y}^{(r)}) + \sum_{j=1}^{r-1} d_j^{(r)} (A\bar{x}^{(i)}, \bar{y}^{(j)}) = 0. \tag{19}$$

Таким образом,

$$\alpha_r = \frac{(\bar{z}^{(r)}, \bar{b})}{(\bar{z}^{(r)}, A\bar{x}^{(r)})}. \tag{20}$$

Остановимся подробнее на первом из описанных методов. Рассмотрим случай, когда матрица A симметрическая и положительно определенная. Последнее означает, что для любого вектора $\bar{x}$ квадратичная форма его компонент $(A\bar{x}, \bar{x})$ больше или равна нулю, причем равенство нулю возможно в том и только том случае, если вектор $\bar{x}$ нулевой. Как мы видели ранее, нужно построить систему векторов $\bar{x}^{(1)}, \bar{x}^{(2)}, \ldots, \bar{x}^{(n)}$, удовлетворяющих условиям

$$(A\bar{x}^{(k)}, \bar{x}^{(r)}) = 0 \qquad (k \neq r). \tag{21}$$

Это построение можно осуществить следующим образом. Исходим из какой-то системы линейно независимых векторов $\bar{y}^{(1)}, \bar{y}^{(2)}, \ldots, \bar{y}^{(n)}$,

например из системы единичных векторов, направленных по координатным осям:

$$\left.\begin{aligned} \bar{y}^{(1)} &= (1,\ 0,\ 0,\ \ldots,\ 0), \\ \bar{y}^{(2)} &= (0,\ 1,\ 0,\ \ldots,\ 0), \\ &\cdots\cdots\cdots\cdots \\ \bar{y}^{(n)} &= (0,\ 0,\ 0,\ \ldots,\ 1). \end{aligned}\right\} \tag{22}$$

Далее проводим «ортогонализацию» также, как это делалось в предыдущей главе. Принимаем $\bar{x}^{(1)} = \bar{y}^{(1)}$ и ищем $\bar{x}^{(2)}$ в виде

$$\bar{x}^{(2)} = \bar{y}^{(2)} + \lambda_1^{(2)}\bar{x}^{(1)}. \tag{23}$$

Из условия $(\bar{x}^{(2)},\ A\bar{x}^{(1)}) = 0$ находим:

$$\lambda_1^{(2)} = -\frac{(\bar{y}^{(2)},\ A\bar{x}^{(1)})}{(\bar{x}^{(1)},\ A\bar{x}^{(1)})}. \tag{24}$$

Ищем $\bar{x}^{(3)}$ в виде

$$\bar{x}^{(3)} = \bar{y}^{(3)} + \lambda_1^{(3)}\bar{x}^{(1)} + \lambda_2^{(3)}\bar{x}^{(2)}. \tag{25}$$

Условия $(\bar{x}^{(3)},\ A\bar{x}^{(1)}) = (\bar{x}^{(3)},\ A\bar{x}^{(2)}) = 0$ влекут за собой

$$\lambda_1^{(3)} = -\frac{(\bar{y}^{(3)},\ Ax^{(1)})}{(\bar{x}^{(1)},\ A\bar{x}^{(1)})}, \qquad \lambda_2^{(3)} = -\frac{(\bar{y}^{(3)},\ A\bar{x}^{(2)})}{(\bar{x}^{(2)},\ A\bar{x}^{(2)})}. \tag{26}$$

Далее поступаем так же.

Процесс будет осуществим, так как все $(\bar{x}^{(i)},\ A\bar{x}^{(i)}) \neq 0$. Это же обеспечит нам разрешимость системы для определения коэффициентов α_k. Заметим, что в нашем случае это будет процесс настоящей ортогонализации, как он проводился в предыдущей главе, если в пространстве векторов ввести новое скалярное произведение при помощи соотношения

$$[\bar{x},\ \bar{y}] = (\bar{x},\ A\bar{y}) = (A\bar{x},\ \bar{y}). \tag{27}$$

Нетрудно проверить, что введенное таким способом скалярное произведение будет удовлетворять всем требованиям, которые к нему предъявляются.

При решении системы n уравнений по настоящей схеме требуется произвести

$$\frac{n^3 + 3n^2 - n}{3} \tag{28}$$

операций умножения и деления.

Приведем пример на применение этого метода. Возьмем опять ту же систему, которую мы решали по методу квадратного корня. Верхняя часть схемы заполнена коэффициентами и правыми частями

системы. Мы не проверяем здесь положительную определенность матрицы A, так как это условие не является необходимым для проведения процесса. Вторая сверху часть схемы заполнена компонентами векторов $\bar{x}^{(i)}$ и коэффициентами λ. Они разделены ломаной линией. Содержание остальных частей не вызывает сомнений. В силу ошибок округления недиагональные элементы нижней части будут отличны от нуля. Их можно подсчитывать для контроля. Их можно также использовать в системе и решать последнюю методом последовательных приближений. В силу значительного преобладания диагональных элементов метод будет быстро сходиться.

	A						$\bar{b}$
	6,1818	0,1818 7,1818	0,3141 0,2141 8,2435	0,1415 0,1815 0,1214 9,3141	0,1516 0,1526 0,2516 0,3145 5,3116	0,2141 0,3141 0,2618 0,6843 0,8998 4,1313	7,1818 8,2435 9,3141 5,3116 4,1313 3,1816
$\bar{x}^{(1)}$	1	— 0,0294	— 0,0508	— 0,0229	— 0,0245	— 0,0346	λ_1
$\bar{x}^{(2)}$	— 0,0294	1	— 0,0286	— 0,0247	— 0,0206	— 0,0425	λ_2
$\bar{x}^{(3)}$	— 0,0500	— 0,0286	1	— 0,0133	— 0,0292	— 0,0295	λ_3
$\bar{x}^{(4)}$	— 0,0215	— 0,0243	— 0,0133	1	— 0,0327	— 0,0719	λ_4
$\bar{x}^{(5)}$	— 0,0217	— 0,0190	— 0,0288	— 0,0327	1	— 0,1625	λ_5
$\bar{x}^{(6)}$	— 0,0268	— 0,0368	— 0,0239	— 0,0666	— 0,1625	1	
$A\bar{x}^{(1)}$	6,1818	0,1818	0,3141	0,1415	0,1516	0,2141	
$A\bar{x}^{(2)}$		7,1765	0,2049	0,1773	0,1481	0,3051	
$A\bar{x}^{(3)}$			8,2217	0,1091	0,2397	0,2422	
$A\bar{x}^{(4)}$				9,3050	0,3042	0,6686	
$A\bar{x}^{(5)}$					5,2879	0,8593	
$A\bar{x}^{(6)}$						3,9161	$(\bar{b}, \bar{x}^{(i)})$
$(A\bar{x}^{(1)}, \bar{x}^{(1)})$	6,1818						7,1818
$(A\bar{x}^{(2)}, \bar{x}^{(2)})$		7,1765					8,0324
$(A\bar{x}^{(3)}, \bar{x}^{(3)})$			8,2217				8,7192
$(A\bar{x}^{(4)}, \bar{x}^{(4)})$				9,3050			4,8330
$(A\bar{x}^{(5)}, \bar{x}^{(5)})$					5,2879		3,3769
$(A\bar{x}^{(6)}, \bar{x}^{(6)})$						3,9161	1,4381

$\alpha_1 = 1{,}1618$; $\alpha_2 = 1{,}1193$; $\alpha_3 = 1{,}0605$; $\alpha_4 = 0{,}5194$; $\alpha_5 = 0{,}6386$; $\alpha_6 = 0{,}3672$;
$x_1 = 1{,}0410$; $x_2 = 1{,}0507$; $x_3 = 1{,}0264$; $x_4 = 0{,}4741$; $x_5 = 0{,}5789$; $x_6 = 0{,}3672$.

В случае несимметрической матрицы процесс ортогонализации проводится точно так же. Пусть векторы $\bar{x}^{(1)}, \bar{x}^{(2)}, \ldots, \bar{x}^{(k)}$ уже построены. Тогда $x^{(k+1)}$ ищется в виде

$$\bar{x}^{(k+1)} = \bar{y}^{(k+1)} + \sum_{i=1}^{k} \beta_i^{(k+1)} \bar{x}^{(i)}. \tag{29}$$

Коэффициенты $\beta_i^{(k+1)}$ определяются из системы

$$\sum_{i=1}^{k} \beta_i^{(k+1)} (A\bar{x}^{(i)}, \bar{x}^{(j)}) = -(A\bar{y}^{(k+1)}, \bar{x}^{(j)}) \qquad (j = 1, 2, \ldots, k). \tag{30}$$

Система в случае несимметрической матрицы будет треугольной.

Аналогично строится система «биортогональных» векторов, т. е. система $2n$ векторов, удовлетворяющих условию (12). При этом $\bar{y}^{(1)}, \bar{y}^{(2)}, \ldots, \bar{y}^{(n)}$ — произвольные n линейно независимых векторов, а векторы $\bar{x}^{(1)}, \bar{x}^{(2)}, \ldots, \bar{x}^{(n)}$ строятся последовательно в виде

$$\bar{x}^{(k+1)} = \bar{y}^{(k+1)} + \sum_{i=1}^{k} \gamma_i^{(k+1)} \bar{x}^{(i)}. \tag{31}$$

Коэффициенты $\gamma_i^{(k+1)}$ находятся из системы

$$\sum_{i=1}^{k} \gamma_i^{(k+1)} (A\bar{x}^{(i)}, \bar{y}^{(j)}) = -(A\bar{y}^{(k+1)}, \bar{y}^{(j)}) \qquad (j = 1, 2, \ldots, k). \tag{32}$$

Так же поступаем, отыскивая коэффициенты $c_i^{(k)}$ и $d_i^{(k)}$, при построении систем векторов (14) и (15), удовлетворяющих условиям (16). При этом получим две системы:

$$\left.\begin{aligned} &\sum_{i=1}^{k-1} c_i^{(k)} (A\bar{y}^{(i)}, \bar{y}^{(j)}) = -(A\bar{y}^{(k)}, \bar{y}^{(j)}), \\ &\sum_{i=1}^{k-1} d_i^{(k)} (A' y_i^{(i)}, \bar{y}^{(j)}) = -(A'\bar{y}^{(k)}, \bar{y}^{(j)}) \qquad (j = 1, 2, \ldots, k-1), \end{aligned}\right\} \tag{33}$$

из которых и определяем $c_i^{(k)}$ и $d_i^{(k)}$.

§ 5. Метод сопряженных градиентов

Пусть A — симметрическая положительно определенная матрица. Рассмотрим функцию

$$f(\bar{x}) = (\bar{x}, A\bar{x}) - 2(\bar{b}, \bar{x}). \tag{1}$$

Это — целая рациональная функция второй степени относительно компонент $(x_1, x_2, \ldots, x_n)$ вектора $\bar{x}$. Поверхности $f(\bar{x}) = \text{const}$ образуют в n-мерном пространстве $R^{(n)}$ семейство эллипсоидов с общим центром $\bar{x}^* = A^{-1}\bar{b}$.

Учитывая симметричность матрицы A, преобразуем разность $f(\bar{x}) - f(\bar{x}^*)$. Получим:

$$f(\bar{x}) - f(\bar{x}^*) = (\bar{x}, A\bar{x}) - 2(\bar{b}, \bar{x}) - (\bar{x}^*, A\bar{x}^*) + 2(\bar{b}, \bar{x}^*) =$$
$$= (\bar{x}, A\bar{x}) - 2(A\bar{x}, \bar{x}^*) + (A\bar{x}^*, \bar{x}^*) = (\bar{x} - \bar{x}^*, A(\bar{x} - \bar{x}^*)). \quad (2)$$

Так как A — положительно определенная матрица, то

$$f(\bar{x}) - f(\bar{x}^*) \geqslant 0, \quad (3)$$

причем знак равенства достигается лишь при $\bar{x} = \bar{x}^*$.

Таким образом, задача об отыскании решения уравнения $A\bar{x} = b$ эквивалентна задаче об отыскании вектора $\bar{x}$, обращающего в минимум функцию $f(\bar{x})$, определенную (1). Существует много методов решения последней задачи. Об одном из таких методов мы здесь и расскажем.

Берем произвольный начальный вектор $\bar{x}^{(0)}$. Будем смещаться, начиная с точки, определенной вектором $\bar{x}^{(0)}$, по нормали к эллипсоиду $f(\bar{x}) = f(\bar{x}^{(0)})$ до тех пор, пока эта нормаль не коснется какого-то эллипсоида семейства $f(\bar{x}) = \text{const}$. Получим новую точку. Вектор, проведенный в эту точку, обозначим через $\bar{x}^{(1)}$. Отыщем этот вектор. Введем обозначение

$$\bar{r}^{(0)} = \bar{b} - A\bar{x}^{(0)} \quad (4)$$

и покажем, что вектор $r^{(0)}$ имеет направление нормали к эллипсоиду $f(\bar{x}) = f(\bar{x}^{(0)})$ при $\bar{x} = \bar{x}^{(0)}$. Действительно, нормальное направление будет совпадать с направлением быстрейшего убывания функции $f(\bar{x})$ — направлением градиента. Таким образом, нам надо найти такой вектор $\bar{c}$, что

$$\left| \frac{d}{d\alpha} f(\bar{x}^{(0)} + \alpha\bar{c}) \right|_{\alpha=0} \quad (5)$$

принимает наибольшее значение при условии $(\bar{c}, \bar{c}) = \text{const}$. Функцию, стоящую под знаком производной, можно записать в виде

$$f(\bar{x}^{(0)} + \alpha\bar{c}) = (\bar{x}^{(0)} + \alpha\bar{c}, A\bar{x}^{(0)} + \alpha A\bar{c}) - 2(\bar{b}, \bar{x}^{(0)} + \alpha\bar{c}) =$$
$$= \alpha^2(\bar{c}, A\bar{c}) - 2\alpha(\bar{r}^{(0)}, \bar{c}) + f(\bar{x}^{(0)}). \quad (6)$$

Таким образом,

$$\frac{d}{d\alpha} f(\bar{x}^{(0)} + \alpha c)\Big|_{\alpha=0} = -2(\bar{r}^{(0)}, \bar{c}). \quad (7)$$

Используя неравенство Буняковского, получим:

$$(\bar{r}^{(0)}, \bar{c})^2 \leqslant (\bar{r}^{(0)}, \bar{r}^{(0)})\ (\bar{c}, \bar{c}), \quad (8)$$

причем при $\bar{c} = \bar{r}^{(0)}$ имеет место знак равенства. Итак, искомое направление будет определяться вектором $\bar{r}^{(0)}$.

Найдем теперь такое α, при котором $f(\bar{x}^{(0)} + \alpha\bar{r}^{(0)})$ принимает наименьшее значение. Используя (6) при $\bar{c} = \bar{r}^{(0)}$, найдем, что искомое α, которое мы обозначим через α_0, равно

$$\alpha_0 = \frac{(\bar{r}^{(0)}, \bar{r}^{(0)})}{(\bar{r}^{(0)}, A\bar{r}^{(0)})}. \tag{9}$$

Итак,

$$\bar{x}^{(1)} = \bar{x}^{(0)} + \alpha_0\bar{r}^{(0)}. \tag{10}$$

На рис. 1 показана геометрическая картина нашего построения при $n = 3$.

Можно было бы продолжить наш процесс и получить из $x^{(1)}$ вектор $\bar{x}^{(2)}$ так же, как мы получили из $\bar{x}^{(0)}$ вектор $\bar{x}^{(1)}$. Мы пришли бы тогда к методу последовательных приближений, носящему

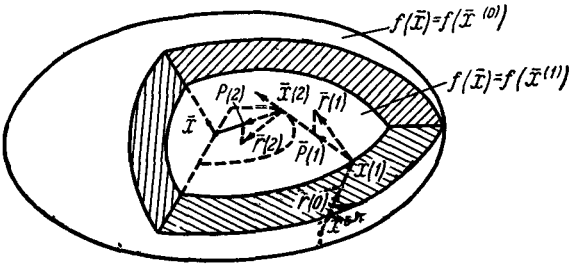

Рис. 1.

название *метода скорейшего спуска*. Рассмотрим его несколько позже. В настоящем параграфе будем получать $\bar{x}^{(2)}$ иначе.

Рассмотрим векторное уравнение

$$(A\bar{r}^{(0)}, \bar{x} - \bar{x}^{(1)}) = 0. \tag{11}$$

Оно определяет гиперплоскость $(n - 1)$-го измерения. Уравнению (11) удовлетворяет вектор $\bar{x} = \bar{x}^{(1)}$. Ему удовлетворяет и $\bar{x}^* = A^{-1}\bar{b}$. Действительно,

$$(A\bar{r}^{(0)}, A^{-1}\bar{b} - \bar{x}^{(1)}) = (\bar{r}^{(0)}, \bar{b} - A\bar{x}^{(1)}) = (\bar{r}^{(0)}, \bar{r}^{(1)}), \tag{12}$$

где через $\bar{r}^{(1)}$ обозначен вектор

$$\bar{r}^{(1)} = \bar{b} - A\bar{x}^{(1)} = \bar{r}^{(0)} + A(\bar{x}^{(0)} - \bar{x}^{(1)}) = \bar{r}^{(0)} - \alpha_0 A\bar{r}^{(0)}. \tag{13}$$

Как и раньше можно показать, что вектор $\bar{r}^{(1)}$ имеет направление нормали к эллипсоиду

$$f(\bar{x}) = f(\bar{x}^{(1)}) \tag{14}$$

при $\bar{x} = \bar{x}^{(1)}$. Вектор $\bar{r}^{(0)}$ касается этого эллипсоида в той же точке. Следовательно, $(\bar{r}^{(0)}, \bar{r}^{(1)}) = 0$.

Сечение гиперплоскостью (11) эллипсоидов $f(\bar{x}) = c$ дает диаметральные гиперплоскости $(n-1)$-го измерения, проходящие через точку, определяемую $\bar{x}^{(1)}$. Будем теперь проводить наши рассуждения лишь в этой гиперплоскости.

Нормаль к эллипсоиду, получающемуся в сечении $f(\bar{x}) = f(\bar{x}^{(1)})$ гиперплоскостью (11), можно получить, проектируя $\bar{r}^{(1)}$ на гиперплоскость. Обозначим эту проекцию через $\bar{p}^{(1)}$ и будем разыскивать ее в виде

$$\bar{p}^{(1)} = \bar{r}^{(1)} + \beta_0 \bar{r}^{(0)}. \tag{15}$$

Так как вектор $A\bar{r}^{(0)}$ ортогонален к гиперплоскости, то и $\bar{p}^{(1)}$ должен быть ортогонален к $A\bar{r}^{(0)}$. Это дает

$$\beta_0 = -\frac{(\bar{r}^{(1)}, A\bar{r}^{(0)})}{(\bar{r}^{(0)}, A\bar{r}^{(0)})}. \tag{16}$$

Будем теперь отыскивать такое α, что $f(\bar{x}^{(1)} + \alpha\bar{p}^{(1)})$ принимает наименьшее значение. Так как

$$f(\bar{x}^{(1)} + \alpha\bar{p}^{(1)}) = \alpha^2(\bar{p}^{(1)}, A\bar{p}^{(1)}) - 2\alpha(\bar{r}^{(1)}, \bar{p}^{(1)}) + f(\bar{x}^{(1)}), \tag{17}$$

то искомое значение α, которое мы будем обозначать через α_1 равно

$$\alpha_1 = \frac{(\bar{r}^{(1)}, \bar{p}^{(1)})}{(\bar{p}^{(1)}, A\bar{p}^{(1)})}. \tag{18}$$

Итак,

$$\bar{x}^{(2)} = \bar{x}^{(1)} + \alpha_1\bar{p}^{(1)}. \tag{19}$$

На следующем шаге мы перейдем к гиперплоскости $(n-2)$-х измерений, определенной уравнениями (11) и

$$(A\bar{p}^{(1)}, \bar{x} - \bar{x}^{(2)}) = 0. \tag{20}$$

В силу ортогональности $\bar{p}^{(1)}$ и

$$\bar{r}^{(2)} = \bar{b} - A\bar{x}^{(2)} = \bar{r}^{(1)} - \alpha_1 A\bar{p}^{(1)} \tag{21}$$

получим, что $\bar{x}^* = A^{-1}\bar{b}$ принадлежит нашей гиперплоскости. Таким образом, и эта гиперплоскость будет диаметральной. Отметим еще следующие свойства $\bar{r}^{(2)}$:

$$(\bar{r}^{(2)}, \bar{r}^{(0)}) = (\bar{r}^{(1)} - \alpha_1 A\bar{p}^{(1)}, \bar{r}^{(0)}) = 0, \tag{22}$$

$$(\bar{r}^{(2)}, \bar{r}^{(1)}) = (\bar{r}^{(1)} - \alpha_1 A\bar{p}^{(1)}, \bar{r}^{(1)}) = (\bar{r}^{(1)} - \alpha_1 A\bar{p}^{(1)}, \bar{p}^{(1)} - \beta_0\bar{r}^{(0)})$$

$$= (\bar{r}^{(1)}, \bar{p}^{(1)}) - \alpha_1(A\bar{p}^{(1)}, \bar{p}^{(1)}) = 0. \tag{23}$$

Проекцию нормального к эллипсоиду $f(\bar{x}) = f(\bar{x}^{(2)})$ при $\bar{x} = \bar{x}^{(2)}$ вектора $\bar{r}^{(2)}$ на гиперплоскость — вектор $\bar{p}^{(2)}$ — будем разыскивать в виде

$$\bar{p}^{(2)} = \bar{r}^{(2)} + \beta_1 \bar{p}^{(1)}. \tag{24}$$

Нужно потребовать ортогональность $\bar{p}^{(2)}$ к $A\bar{r}^{(0)}$ и $A\bar{p}^{(1)}$. Но

$$(\bar{p}^{(2)}, A\bar{r}^{(0)}) = (\bar{r}^{(2)} + \beta_1 \bar{p}^{(1)}, A\bar{r}^{(0)}) = (\bar{r}^{(2)}, A\bar{p}^{(0)}) = \frac{1}{\alpha_0}(\bar{r}^{(2)}, \bar{r}^{(0)} - \bar{r}^{(1)}) = 0. \tag{25}$$

Условие ортогональности $\bar{p}^{(2)}$ и $A\bar{p}^{(1)}$ записывается так:

$$\beta_1 = -\frac{(\bar{r}^{(2)}, A\bar{p}^{(1)})}{(\bar{p}^{(1)}, A\bar{p}^{(1)})}. \tag{26}$$

Теперь, как и ранее, отыскиваем значение α, для которого $f(\bar{x}^{(2)} + \alpha\bar{p}^{(2)})$ принимает наименьшее значение. Это значение α, которое мы обозначим через α_2, будет равно

$$\alpha_2 = \frac{(\bar{r}^{(2)}, \bar{p}^{(2)})}{(\bar{p}^{(2)}, A\bar{p}^{(2)})}. \tag{27}$$

Поэтому вектор $\bar{x}^{(3)}$ будет представляться в виде

$$\bar{x}^{(3)} = \bar{x}^{(2)} + \alpha_2 \bar{p}^{(2)}. \tag{28}$$

Так же продолжаем и дальше. Получим последовательность векторов $\{\bar{x}^{(k)}\}$, $\{\bar{r}^{(k)}\}$, $\{\bar{p}^{(k)}\}$, которые определятся рекуррентно следующим образом:

$$\bar{p}^{(0)} = \bar{r}^{(0)}, \tag{29}$$

$$\alpha_k = \frac{(\bar{r}^{(k)}, \bar{p}^{(k)})}{(\bar{p}^{(k)}, A\bar{p}^{(k)})}, \tag{30}$$

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} + \alpha_k \bar{p}^{(k)}, \tag{31}$$

$$\bar{r}^{(k+1)} = \bar{r}^{(k)} - \alpha_k A\bar{p}^{(k)}, \tag{32}$$

$$\beta_k = -\frac{(\bar{r}^{(k+1)}, A\bar{p}^{(k)})}{(\bar{p}^{(k)}, A\bar{p}^{(k)})}, \tag{33}$$

$$\bar{p}^{(k+1)} = \bar{r}^{(k+1)} + \beta_k \bar{p}^{(k)}. \tag{34}$$

Отметим следующие свойства членов этих последовательностей. В силу самого построения будем иметь:

$$(\bar{p}^{(i)}, A\bar{p}^{(j)}) = (A\bar{p}^{(i)}, \bar{p}^{(j)}) = 0 \quad \text{при} \quad i \neq j. \tag{35}$$

Далее, при $i > j$ получим:

$$(\bar{r}^{(i)}, \bar{p}^{(j)}) = (\bar{r}^{(i-1)} - \alpha_{i-1} A\bar{p}^{(i-1)}, \bar{p}^{(j)}) =$$

$$= (\bar{r}^{(i-1)}, \bar{p}^{(j)}) - \alpha_{i-1} (\bar{p}^{(j)}, A\bar{p}^{(i-1)}). \qquad (36)$$

Правая часть последнего равенства равна нулю при $i = j + 1$ в силу определения α_{i-1}. При $i > j + 1$ она будет равна $(\bar{r}^{(i-1)}, \bar{p}^{(j)})$, т. е. индекс у $\bar{r}^{(i)}$ понизился на единицу. Повторяя рассуждения достаточное количество раз, мы придем в конце концов к случаю, когда индекс у $\bar{r}$ будет на единицу больше индекса у $\bar{p}$. Следовательно, при $i > j$

$$(\bar{r}^{(i)}, \bar{p}^{(j)}) = 0. \qquad (37)$$

Наконец, рассмотрим $(\bar{r}^{(i)}, \bar{r}^{(j)})$ при $i \neq j$. Пусть, для определенности $i > j$. Тогда

$$(\bar{r}^{(i)}, \bar{r}^{(j)}) = (\bar{r}^{(i)}, \bar{p}^{(j)} - \beta_{j-1}\bar{p}^{(j-1)}) = (\bar{r}^{(i)}, \bar{p}^{(j)}) - \beta_{j-1}(\bar{r}^{(i)}, \bar{p}^{(j-1)}) = 0. \qquad (38)$$

Так как в n-мерном пространстве не может быть более n взаимно ортогональных векторов, то на некотором шаге $k \leqslant n$ мы получим $\bar{r}^{(k)} = 0$. При этом $\bar{x}^{(k)} = A^{-1}\bar{b}$. Таким образом, на некотором шаге мы придем к точному решению системы. Такой способ получения решения будем называть *методом сопряженных градиентов*.

Методу сопряженных градиентов можно дать простую алгебраическую трактовку. Будем таким же способом, как это делалось ранее, ортогонализировать систему векторов $\bar{r}^{(0)}, A\bar{r}^{(0)}, \ldots, A^k\bar{r}^{(0)}, \ldots$ Полученные при этом векторы $\bar{r}^{(0)}, \bar{r}^{(1)}, \ldots, \bar{r}^{(k)}, \ldots$ будут обладать следующими свойствами:

1. Вектор $\bar{r}^{(k)}$ является линейной комбинацией векторов $\bar{r}^{(0)}, A\bar{r}^{(0)}, \ldots, A^k\bar{r}^{(0)}$:

$$\bar{r}^{(k)} = c_0^{(k)}\bar{r}^{(0)} + c_1^{(k)}A\bar{r}^{(0)} + \ldots + c_k^{(k)}A^k\bar{r}^{(0)}. \qquad (39)$$

Это свойство очевидно, если вспомнить ход процесса ортогонализации, как он определен выше.

2. Вектор $\bar{r}^{(k+1)}$ ортогонален к наименьшему линейному многообразию, содержащему векторы $\bar{r}^{(0)}, \bar{r}^{(1)}, \ldots, \bar{r}^{(k)}$.

Действительно, если $\bar{r}^{(k+1)} = 0$, то утверждение тривиально. Если же $\bar{r}^{(k+1)} \neq 0$, то и каждый из векторов $\bar{r}^{(0)}, \bar{r}^{(1)}, \ldots, \bar{r}^{(k)}$ не равен нулю. Совокупность последних векторов образует базис наименьшего линейного многообразия, содержащего векторы $\bar{r}^{(0)}, A\bar{r}^{(0)}, \ldots, A^k\bar{r}^{(0)}$. Вектор $\bar{r}^{(k+1)}$, будучи ортогональным ко всем

векторам базиса, будет ортогональным и ко всему линейному многообразию.

3. Вектор $\bar{r}^{(k)} = 0$ тогда и только тогда, когда $\bar{r}^{(0)}, A\bar{r}^{(0)}, \ldots, A^k\bar{r}^{(0)}$ линейно зависимы.

Если $\bar{r}^{(k)} \neq 0$, то векторы $\bar{r}^{(0)}, \bar{r}^{(1)}, \ldots, \bar{r}^{(k)}$ как ненулевые взаимно ортогональные векторы линейно независимы. Это может быть только в том случае, если порождающие их векторы $\bar{r}^{(0)}, A\bar{r}^{(0)}, \ldots, A^k\bar{r}^{(0)}$ также линейно независимы. Если же $\bar{r}^{(k)} = 0$, то (39) дает линейную зависимость векторов $\bar{r}^{(0)}, A\bar{r}^{(0)}, \ldots, A^k\bar{r}^{(0)}$.

4. Если $\bar{r}^{(k)} \neq 0$, то коэффициент $c_k^{(k)}$ в (39) отличен от нуля.

Это свойство есть следствие второго, так как вектор, ортогональный к некоторому многообразию, не может принадлежать этому многообразию.

5. Каждый из векторов $A^m\bar{r}^{(0)}$ $(m = 0, 1, 2, \ldots, k)$ может быть представлен как линейная комбинация векторов $\bar{r}^{(0)}, \bar{r}^{(1)}, \ldots, \bar{r}^{(k)}$. Коэффициенты этой линейной комбинации определяются однозначно, если ни один из векторов $\bar{r}^{(0)}, \bar{r}^{(1)}, \ldots, \bar{r}^{(k)}$ не равен нулю.

Среди векторов $\bar{r}^{(0)}, A\bar{r}^{(0)}, \ldots, A^k\bar{r}^{(0)}$ имеется максимальное число линейно независимых. Таково же число ненулевых векторов среди $\bar{r}^{(0)}, \bar{r}^{(1)}, \ldots, \bar{r}^{(k)}$. Эти ненулевые векторы образуют базис наименьшего линейного многообразия, содержащего векторы $\bar{r}^{(0)}, A\bar{r}^{(0)}, \ldots, A^m\bar{r}^{(0)}$. Отсюда и следует утверждение.

Рассмотрим теперь наряду с обычным скалярным произведением скалярное произведение, определенное равенством (27) предыдущего параграфа. Последовательность векторов $\bar{r}^{(0)}, A\bar{r}^{(0)}, A^2\bar{r}^{(0)}, \ldots, A^k\bar{r}^{(0)}, \ldots$ можно ортогонализировать в смысле этого скалярного произведения. Получим новую последовательность векторов $\bar{p}^{(0)} = \bar{r}^{(0)}, \bar{p}^{(1)}, \ldots, \bar{p}^{(k)}, \ldots$ Для этих векторов будут справедливы высказанные нами утверждения 1—5. Кроме того, системы векторов $\{\bar{r}^{(i)}\}$ и $\{\bar{p}^{(i)}\}$ будут связаны некоторыми соотношениями. Векторы $\bar{p}^{(i)}$ при $i \leqslant k$ принадлежат линейному многообразию, порожденному векторами $\bar{r}^{(0)}, A\bar{r}^{(0)}, \ldots, A^k\bar{r}^{(0)}$, а вектор $\bar{r}^{(k+1)}$ ортогонален к этому многообразию. Таким образом, мы будем иметь:

$$(\bar{r}^{(i)}, \bar{p}^{(j)}) = 0 \quad \text{при} \quad i > j. \tag{40}$$

Аналогично показывается, что

$$[\bar{p}^{(i)}, \bar{r}^{(j)}] = (\bar{p}^{(i)}, A\bar{r}^{(j)}) = (A\bar{p}^{(i)}, \bar{r}^{(j)}) = 0 \quad \text{при} \quad i > j. \tag{41}$$

Воспользуемся равенствами (40) и (41) для установления формул, связывающих векторы $\bar{p}^{(k)}, \bar{r}^{(k)}, \bar{p}^{(k+1)}, \bar{r}^{(k+1)}, A\bar{p}^{(k)}$. Пусть векторы $\bar{r}^{(0)}$,

$\bar{r}^{(1)}, \ldots, \bar{r}^{(k)}$ отличны от нуля. Тогда и векторы $\bar{p}^{(0)}, \bar{p}^{(1)}, \ldots, \bar{p}^{(k)}$ отличны от нуля. Вектор $\bar{r}^{(k+1)}$ по (39) принадлежит линейному многообразию, порожденному векторами $\bar{r}^{(0)}, A\bar{r}^{(0)}, \ldots, A^{k+1}\bar{r}^{(0)}$. В этом линейном многообразии можно взять в качестве базиса векторы $\bar{r}^{(0)}$, $\bar{r}^{(1)}, \ldots, \bar{r}^{(k)}, A^{k+1}\bar{r}^{(0)}$. С другой стороны, $A^{k+1}\bar{r}^{(0)} = A(A^k\bar{r}^{(0)})$, а вектор $A^k\bar{r}^{(0)}$ может быть представлен как линейная комбинация векторов $\bar{p}^{(0)}, \bar{p}^{(1)}, \ldots, \bar{p}^{(k)}$. Таким образом, вектор $A^{k+1}\bar{r}^{(0)}$ может быть представлен как линейная комбинация векторов $A\bar{p}^{(0)}, A\bar{p}^{(1)}, \ldots$ $\ldots, A\bar{p}^{(k)}$. Векторы $A\bar{p}^{(0)}, A\bar{p}^{(1)}, \ldots, A\bar{p}^{(k-1)}$ принадлежат линейному многообразию, порожденному векторами $\bar{r}^{(0)}, A\bar{r}^{(0)}, \ldots, A^k\bar{r}^{(0)}$ и, следовательно, могут быть представлены как линейные комбинации векторов $\bar{r}^{(0)}, \bar{r}^{(1)}, \ldots, \bar{r}^{(k)}$. Поэтому вектор $\bar{r}^{(k+1)}$ можно записать в виде

$$\bar{r}^{(k+1)} = d_0^{(k+1)}\bar{r}^{(0)} + d_1^{(k+1)}\bar{r}^{(1)} + \ldots + d_k^{(k+1)}\bar{r}^{(k)} - \alpha_k A\bar{p}^{(k)}. \quad (42)$$

Помножим обе части равенства (42) скалярно на $\bar{r}^{(i)}$ ($i = 0, 1, \ldots$ $\ldots, k-1$). В силу ортогональности векторов $\bar{r}^{(i)}$ и $\bar{r}^{(j)}$ при $i \neq j$ и равенства (41) получим:

$$d_0^{(k+1)} = d_1^{(k+1)} = \ldots = d_{k-1}^{(k+1)} = 0. \quad (43)$$

Итак,

$$\bar{r}^{(k+1)} = d_k^{(k+1)}\bar{r}^{(k)} - \alpha_k A\bar{p}^{(k)}. \quad (44)$$

Умножим это равенство скалярно на $\bar{p}^{(k)}$. В силу (40) будем иметь:

$$0 = d_k^{(k+1)}(\bar{r}^{(k)}, \bar{p}^{(k)}) - \alpha_k(A\bar{p}^{(k)}, \bar{p}^{(k)}). \quad (45)$$

Коэффициент α_k отличен от нуля. Действительно, если $\bar{r}^{(k+1)} = 0$, то условие $\alpha_k = 0$ означает, что векторы $\bar{r}^{(0)}, \bar{r}^{(1)}, \ldots, \bar{r}^{(k)}$ линейно зависимы, если же $\bar{r}^{(k+1)} \neq 0$, то условие $\alpha_k = 0$ означает линейную зависимость векторов $\bar{r}^{(0)}, \bar{r}^{(1)}, \ldots, \bar{r}^{(k+1)}$. И то и другое невозможно. Скалярное произведение $(A\bar{p}^{(k)}, \bar{p}^{(k)})$ также отлично от нуля. Поэтому $d_k^{(k+1)} \neq 0$ и $(\bar{r}^{(k)}, \bar{p}^{(k)}) \neq 0$. Так как векторы $\bar{r}^{(k)}$ определяются с точностью до постоянного множителя, то мы всегда можем считать $d_k^{(k+1)} = 1$. Тогда из (45) получаем:

$$\alpha_k = \frac{(\bar{r}^{(k)}, \bar{p}^{(k)})}{(\bar{p}^{(k)}, A\bar{p}^{(k)})}, \quad (46)$$

$$\bar{r}^{(k+1)} = \bar{r}^{(k)} - \alpha_k A\bar{p}^{(k)}. \quad (47)$$

Обозначим через $\bar{x}^{(k)}$ решение уравнения

$$\bar{b} - A\bar{x} = \bar{r}^{(k)}. \tag{48}$$

Подставляя в (47) вместо $\bar{r}^{(i)}$ их выражения по (48), получим:

$$A(\bar{x}^{(k+1)} - \bar{x}^{(k)}) = \alpha_k A\bar{p}^{(k)}. \tag{49}$$

Итак,

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} + \alpha_k \bar{p}^{(k)}. \tag{50}$$

Рассмотрим теперь разность векторов $\bar{r}^{(k+1)} - \bar{p}^{(k+1)}$. Каждый из этих двух векторов представляется в виде линейной комбинации векторов $\bar{r}^{(0)}$, $A\bar{r}^{(0)}$, ..., $A^{k+1}\bar{r}^{(0)}$. Так как векторы $\bar{p}^{(i)}$ определяются с точностью до постоянного множителя, то мы всегда можем предполагать, что коэффициенты этих линейных комбинаций при $A^{k+1}\bar{r}^{(0)}$ равны. Тогда разность $\bar{r}^{(k+1)} - \bar{p}^{(k+1)}$ может быть представлена в виде линейной комбинации векторов $\bar{r}^{(0)}$, $A\bar{r}^{(0)}$, ..., $A^k\bar{r}^{(0)}$ или векторов $\bar{p}^{(0)}$, $\bar{p}^{(1)}$, ..., $\bar{p}^{(k)}$:

$$\bar{r}^{(k+1)} - \bar{p}^{(k+1)} = \beta_0 \bar{p}^{(0)} + \beta_1 \bar{p}^{(1)} + \ldots + \beta_k \bar{p}^{(k)}. \tag{51}$$

Умножим скалярно равенство (51) на $A\bar{p}^{(i)}$ ($i = 0, 1, \ldots, k-1$). При этом получим:

$$\beta_0 = \beta_1 = \ldots = \beta_{k-1} = 0. \tag{52}$$

После умножения на $A\bar{p}^{(k)}$ будем иметь:

$$-(\bar{r}^{(k+1)}, A\bar{p}^{(k)}) = \beta_k (\bar{p}^{(k)}, A\bar{p}^{(k)}), \tag{53}$$

или

$$\beta_k = -\frac{(\bar{r}^{(k+1)}, A\bar{p}^{(k)})}{(\bar{p}^{(k)}, A\bar{p}^{(k)})}. \tag{54}$$

Итак,

$$\bar{p}^{(k+1)} = \bar{r}^{(k+1)} + \beta_k \bar{p}^{(k)}, \tag{55}$$

где β_k определяются равенствами (54).

Нетрудно заметить, что коэффициенты α_k и β_k можно также вычислять по формулам:

$$\alpha_k = \frac{(\bar{r}^{(k)}, \bar{r}^{(k)})}{(\bar{p}^{(k)}, A\bar{p}^{(k)})}, \tag{56}$$

$$\beta_k = \frac{(\bar{r}^{(k+1)}, \bar{r}^{(k+1)})}{(\bar{r}^{(k)}, \bar{r}^{(k)})}. \tag{57}$$

Мы снова пришли к методу сопряженных градиентов. Рассуждая, как и прежде, мы придем к выводу, что процесс должен закон-

читься после $k \leqslant n$ шагов. При этом $\bar{r}^{(k)} = 0$ и $\bar{x}^{(k)} = A^{-1}\bar{b}$. Может оказаться, что вследствие ошибок округления $\bar{r}^{(n)}$ будет отлично от нуля. Тогда можно проделать еще несколько шагов до тех пор, пока не получим достаточно малое $\bar{r}^{(k)}$. С другой стороны, может оказаться, что уже после небольшого числа шагов $\bar{r}^{(k)}$ мало. Тогда можно и не продолжать процесса.

Метод сопряженных градиентов можно обобщить на случай произвольной неособенной матрицы A. Пусть B — некоторая положительная определенная симметрическая матрица. Выбираем произвольные векторы $\bar{p}^{(0)} = \bar{q}^{(0)}$ и строим последовательности векторов $\{\bar{p}^{(k)}\}$ и $\{\bar{q}^{(k)}\}$ при помощи рекуррентных формул:

$$\left.\begin{aligned} \alpha_k &= \frac{(\bar{q}^{(k)}, \bar{q}^{(k)})}{(\bar{p}^{(k)}, B\bar{p}^{(k)})}, \\ \bar{q}^{(k+1)} &= \bar{q}^{(k)} - \alpha_k B\bar{p}^{(k)}, \\ \beta_k &= \frac{(\bar{q}^{(k+1)}, \bar{q}^{(k+1)})}{(\bar{q}^{(k)}, \bar{q}^{(k)})}, \\ \bar{p}^{(k+1)} &= \bar{q}^{(k+1)} + \beta_k \bar{p}^{(k)}. \end{aligned}\right\} \tag{58}$$

Эти последовательности будут обладать указанными выше свойствами, конечно с заменой матрицы A на матрицу B. В частности, $\bar{q}^{(n)} = 0$. Следовательно,

$$\bar{q}^{(0)} = \sum_{k=0}^{n-1} \alpha_k B\bar{p}^{(k)}. \tag{59}$$

Пусть теперь

$$\bar{p}^{(0)} = \bar{q}^{(0)} = C(\bar{b} - A\bar{x}^{(0)}), \tag{60}$$

где $\bar{x}^{(0)}$ — произвольный начальный вектор и C — произвольная неособенная матрица. Тогда из (59) следует:

$$\bar{x}^* = A^{-1}\bar{b} = \bar{x}^{(0)} + \sum_{k=0}^{n-1} \alpha_k A^{-1}C^{-1}B\bar{p}^{(k)}. \tag{61}$$

Итак, если последовательно, начиная с выбранного $\bar{x}^{(0)}$, вычислять векторы

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} + \alpha_k A^{-1}C^{-1}B\bar{p}^{(k)}, \tag{62}$$

то получим $\bar{x}^{(n)} = \bar{x}^*$. Для векторов $\bar{r}^{(k)} = \bar{b} - A\bar{x}^{(k)}$ из (62) получим:

$$\bar{r}^{(k+1)} = \bar{r}^{(k)} - \alpha_k C^{-1}B\bar{p}^{(k)}. \tag{63}$$

Если вспомнить, что $\bar{q}^{(0)} = C\bar{r}^{(0)}$ и сравнить (63) и соответствующее равенство (58) для $\bar{q}^{(k+1)}$, то получим $\bar{q}^{(k)} = C\bar{r}^{(k)}$ при любом k.

Таким образом, (58) можно переписать в виде

$$\left.\begin{aligned} \bar{p}^{(0)} &= C\bar{r}^{(0)}, \\ \alpha_k &= \frac{(C\bar{r}^{(k)}, \bar{C}\bar{r}^{(k)})}{(\bar{p}^{(k)}, B\bar{p}^{(k)})}, \\ \bar{x}^{(k+1)} &= \bar{x}^{(k)} + \alpha_k A^{-1}C^{-1}B\bar{p}^{(k)}, \\ \bar{r}^{(k+1)} &= \bar{r}^{(k)} - \alpha_k C^{-1}B\bar{p}^{(k)}, \\ \beta_k &= \frac{(C\bar{r}^{(k+1)}, C\bar{r}^{(k+1)})}{(C\bar{r}^{(k)}, \bar{C}\bar{r}^{(k)})}, \\ \bar{p}^{(k+1)} &= C\bar{r}^{(k+1)} + \beta_k \bar{p}^{(k)}. \end{aligned}\right\} \tag{64}$$

Это и будут формулы, соответствующие (29) — (34) для несимметричного случая.

Рассмотрим два частных случая.

1-й случай. Выбираем $B = A'A$ и $C = A'$. При этом получим:

$$\left.\begin{aligned} \bar{p}^{(0)} &= A'\bar{r}^{(0)}, \\ \alpha_k &= \frac{(A'\bar{r}^{(k)}, A'\bar{r}^{(k)})}{(A\bar{p}^{(k)}, A\bar{p}^{(k)})}, \\ \bar{x}^{(k+1)} &= \bar{x}^{(k)} + \alpha_k \bar{p}^{(k)}, \\ \bar{r}^{(k+1)} &= \bar{r}^{(k)} - \alpha_k A\bar{p}^{(k)}, \\ \beta_k &= \frac{(A'\bar{r}^{(k+1)}, A'\bar{r}^{(k+1)})}{(A'\bar{r}^{(k)}, A'\bar{r}^{(k)})}, \\ \bar{p}^{(k+1)} &= A'\bar{r}^{(k+1)} + \beta_k \bar{p}^{(k)}. \end{aligned}\right\} \tag{65}$$

2-й случай. Выбираем $B = AA'$ и $C = I$. При этом получим:

$$\left.\begin{aligned} \bar{p}^{(0)} &= \bar{r}^{(0)}, \\ \alpha_k &= \frac{(\bar{r}^{(k)}, \bar{r}^{(k)})}{(A'\bar{p}^{(k)}, A'\bar{p}^{(k)})}, \\ \bar{x}^{(k+1)} &= \bar{x}^{(k)} + \alpha_k A'\bar{p}^{(k)}, \\ \bar{r}^{(k+1)} &= \bar{r}^{(k)} - \alpha_k AA'\bar{p}^{(k)}, \\ \beta_k &= \frac{(\bar{r}^{(k+1)}, \bar{r}^{(k+1)})}{(\bar{r}^{(k)}, \bar{r}^{(k)})}, \\ \bar{p}^{(k+1)} &= \bar{r}^{(k+1)} + \beta_k \bar{p}^{(k)}. \end{aligned}\right\} \tag{66}$$

Относительно применения этих формул можно сказать то же, что было сказано выше о применении формул (29) — (34).

§ 6. Метод разбиения на клетки

Обращение матрицы высокого порядка часто удается свести к обращению матриц низшего порядка, являющихся частью основной матрицы. Будем называть такие методы обращения матриц *методами разбиения на клетки.*

Пусть нам дана квадратная неособенная матрица A. Разобьем ее пунктирными линиями на частичные матрицы:

$$A=\left\|\begin{array}{ccccccc} a_{11} & a_{12} & \cdot & a_{1k} & a_{1,k+1} & \cdot\cdot & a_{1n} \\ a_{21} & a_{22} & \cdot\cdot & a_{2k} & a_{2,k+1} & \cdot & a_{2n} \\ \cdot & \cdot & & & & \cdot & \cdot \\ a_{i1} & a_{i2} & \cdot\cdot & a_{ik} & a_{i\,k+1} & \cdot\cdot & a_{in} \\ a_{i+1\,1} & a_{i+1\,2} & & a_{i+1\,k} & a_{i+1\,k+1} & & a_{i+1,\,n} \\ \cdot & \cdot & & \cdot & \cdot & & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nk} & a_{n,\,k+1} & \cdots & a_{n,n} \end{array}\right\|. \tag{1}$$

Это можно сокращенно записать так:

$$A=\begin{pmatrix} A_{11}^{(i,\,k)} & A_{12}^{(i,\,n-k)} \\ A_{21}^{(n-i,\,k)} & A_{22}^{(n-i,\,n-k)} \end{pmatrix}. \tag{2}$$

Здесь $A_{r,\,s}^{(l,\,m)}$ сами являются матрицами. Нижние индексы r, s показывают место частичной матрицы в полной матрице, так же как индексы элемента a_{ij} показывают его положение в матрице. Верхние индексы l, m показывают соответственно число строк и число столбцов частичной матрицы. Если такое разбиение осуществлено, то будем говорить, что матрица разбита на клетки или блоки.

Если имеется вторая матрица B с таким же разбиением

$$B=\begin{pmatrix} B_{11}^{(i,\,k)} & B_{12}^{(i,\,n-k)} \\ B_{21}^{(n-i,\,k)} & B_{22}^{(n-i,\,n-k)} \end{pmatrix}, \tag{3}$$

то

$$A+B=\begin{pmatrix} C_{11}^{(i,\,k)} & C_{12}^{(i,\,n-k)} \\ C_{21}^{(n-i,\,k)} & C_{22}^{(n-i,\,n-k)} \end{pmatrix}, \tag{4}$$

где

$$C_{r,\,s}^{(l,\,m)}=A_{r,\,s}^{(l,\,m)}+B_{r,\,s}^{(l,\,m)}. \tag{5}$$

Далее, если имеем матрицу

$$D=\begin{pmatrix} D_{11}^{(k,\,j)} & D_{12}^{(k,\,n-j)} \\ D_{21}^{(n-k,\,j)} & D_{22}^{(n-k,\,n-j)} \end{pmatrix}, \tag{6}$$

то

$$AD=\begin{pmatrix} E_{11}^{(i,\,j)} & E_{12}^{(i,\,n-j)} \\ E_{21}^{(n-i,\,j)} & E_{22}^{(n-i,\,n-j)} \end{pmatrix}, \tag{7}$$

где

$$\left.\begin{aligned} E_{11}^{(i,j)} &= A_{11}^{(i,k)}D_{11}^{(k,j)} + A_{12}^{(i,n-k)}D_{21}^{(n-k,j)};\\ E_{12}^{(i,n-j)} &= A_{11}^{(i,k)}D_{12}^{(k,n-j)} + A_{12}^{(i,n-k)}D_{22}^{(n-k,n-j)};\\ E_{21}^{(n-i,j)} &= A_{21}^{(n-i,k)}D_{11}^{(k,j)} + A_{22}^{(n-i,n-k)}D_{21}^{(n-k,j)};\\ E_{22}^{(n-i,n-j)} &= A_{21}^{(n-i,k)}D_{12}^{(k,n-j)} + A_{22}^{(n-i,n-k)}D_{22}^{(n-k,n-j)}. \end{aligned}\right\} \quad (8)$$

Обращаем внимание на то, что разбиение на клетки матрицы D должно быть согласовано с разбиением на клетки матрицы A для того, чтобы было возможно осуществить умножение клеток. В частности, если взять $i=j=k$, т. е. если

$$A=\begin{pmatrix} A_{11}^{(i,i)} & A_{12}^{(i,n-i)} \\ A_{21}^{(n-i,i)} & A_{22}^{(n-i,n-i)} \end{pmatrix}, \quad D=\begin{pmatrix} D_{11}^{(i,i)} & D_{12}^{(i,n-i)} \\ D_{21}^{(n-i,i)} & D_{22}^{(n-i,n-i)} \end{pmatrix}, \quad (9)$$

то клеточное умножение матриц возможно и произведение будет разбито на такие же клетки, что и каждый из сомножителей.

Нетрудно сообразить, как будут выглядеть правила действий с клеточными матрицами, если осуществлять разбиение на большее число клеток, и в том случае, если мы имеем дело с прямоугольными матрицами.

Пусть теперь матрица D в (9) равна A^{-1}. Тогда мы должны иметь $AD=I^{(n,n)}$ и

$$\left.\begin{aligned} A_{11}^{(i,i)}D_{11}^{(i,i)} + A_{12}^{(i,n-i)}D_{21}^{(n-i,i)} &= I_{11}^{(i,i)},\\ A_{11}^{(i,i)}D_{12}^{(i,n-i)} + A_{12}^{(i,n-i)}D_{22}^{(n-i,n-i)} &= O_{12}^{(i,n-i)},\\ A_{21}^{(n-i,i)}D_{11}^{(i,i)} + A_{22}^{(n-i,n-i)}D_{21}^{(n-i,i)} &= O_{21}^{(n-i,i)},\\ A_{21}^{(n-i,i)}D_{12}^{(i,n-i)} + A_{22}^{(n-i,n-i)}D_{22}^{(n-i,n-i)} &= I_{22}^{(n-i,n-i)}. \end{aligned}\right\} \quad (10)$$

Здесь через $I_{r,r}^{(l,l)}$ обозначены единичные матрицы соответствующих порядков, а через $O_{r,s}^{(l,m)}$ — матрицы, состоящие из сплошных нулей.

Таким образом, мы сумеем найти A^{-1}, если подберем матрицы $D_{r,s}^{(l,m)}$ так, чтобы были выполнены равенства (10). Непосредственной проверкой убеждаемся, что матрицы $D_{r,s}^{(l,m)}$ можно последовательно находить из равенств:

$$\left.\begin{aligned} D_{22}^{(n-i,n-i)} &= \left[A_{22}^{(n-i,n-i)} - A_{21}^{(n-i,i)}\left(A_{11}^{(i,i)}\right)^{-1}A_{12}^{(i,n-i)}\right]^{-1},\\ D_{12}^{(i,n-i)} &= -\left(A_{11}^{(i,i)}\right)^{-1}A_{12}^{(i,n-i)}D_{22}^{(n-i,n-i)},\\ D_{21}^{(n-i,i)} &= -D_{22}^{(n-i,n-i)}A_{21}^{(n-i,i)}\left(A_{11}^{(i,i)}\right)^{-1},\\ D_{11}^{(i,i)} &= \left(A_{11}^{(i,i)}\right)^{-1}\left(I_{11}^{(i,i)} - A_{12}^{(i,n-i)}D_{21}^{(n-i,i)}\right). \end{aligned}\right\} \quad (11)$$

Следовательно, для того чтобы обратить матрицу A порядка n, нам придется обратить две матрицы, одна из которых имеет порядок i, а другая порядок $n-i$.

Чаще всего берут i равным $n-1$. Тогда придется обращать всего лишь одну матрицу порядка $n-1$. Для ее обращения можно применять тот же прием. Это в свою очередь потребует обращения матрицы порядка $n-2$. Продолжая этот процесс дальше, мы в конце концов придем к матрице первого порядка. Таким образом, последовательно обращая матрицы

$$(a_{11}),\quad \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix},\quad \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix},\ \ldots, \tag{12}$$

мы придем к A^{-1}.

Можно дать другой подход к рассмотренному методу. Представляем матрицу A в виде

$$\begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \ldots & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix} = \begin{pmatrix} a_{11} & \ldots & a_{1,\,n-1} & 0 \\ a_{21} & \ldots & a_{2,\,n-1} & 0 \\ \cdot & \ldots & \cdot & \cdot \\ a_{n-1,\,1} & \ldots & a_{n-1,\,n-1} & 0 \\ 0 & \ldots & 0 & a_{nn} \end{pmatrix} +$$

$$+ \begin{pmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ 0 \\ 1 \end{pmatrix} (a_{n1}\ \ a_{n2}\ \ \ldots\ \ a_{n,\,n-1}\ \ 0) + \begin{pmatrix} a_{1n} \\ a_{2n} \\ \cdot \\ \cdot \\ a_{n-1,\,n} \\ 0 \end{pmatrix} (0\ \ 0\ \ \ldots\ \ 0\ \ 1). \tag{13}$$

Изложенный метод показывает, как, зная обратную матрицу для первого слагаемого правой части, получить обратную матрицу для суммы всех слагаемых правой части. Идя по такому пути, можно поставить следующую задачу. Матрица A представлена в виде

$$A = B + uv, \tag{14}$$

где B — некоторая квадратная матрица, для которой известна B^{-1}, u — матрица, состоящая из одного столбца и n строк, и v — матрица, состоящая из одной строки и n столбцов.

Требуется найти A^{-1}.

Решение поставленной задачи дает матрица

$$A^{-1} = B^{-1} - \frac{B^{-1}uvB^{-1}}{1 + vB^{-1}u}. \tag{15}$$

Действительно,

$$(B+uv)\left(B^{-1}-\frac{B^{-1}uvB^{-1}}{1+vB^{-1}u}\right)=I-\frac{uvB^{-1}}{1+vB^{-1}u}+uvB^{-1}-$$

$$-\frac{uvB^{-1}uvB^{-1}}{1+vB^{-1}u}=I+\frac{vB^{-1}u\cdot uvB^{-1}}{1+vB^{-1}u}-\frac{vB^{-1}u\cdot uvB^{-1}}{1+vB^{-1}u}=I. \qquad (16)$$

Используя формулу (15), можно получать обратные матрицы в широком классе случаев. Так, например, возьмем

$$B=\begin{pmatrix} a_{11} & 0 & 0 \dots 0 \\ 0 & a_{22} & 0 \dots 0 \\ \dots & \dots & \dots \\ 0 & 0 & 0 \dots a_{nn} \end{pmatrix}. \qquad (17)$$

Обратная матрица при этом находится без труда. За матрицы u и v примем

$$u=\begin{pmatrix} 0 \\ a_{21} \\ \cdot \\ \cdot \\ a_{n1} \end{pmatrix}, \quad v=(1,\ 0,\ \dots,\ 0). \qquad (18)$$

Тогда формула (15) даст нам обратную матрицу для

$$\begin{pmatrix} a_{11} & 0 & 0 \dots 0 \\ a_{21} & a_{22} & 0 \dots 0 \\ \dots & \dots & \dots \\ a_{n1} & 0 & 0 \dots a_{nn} \end{pmatrix}. \qquad (19)$$

Затем таким же образом можно исправить второй столбец, третий и т. д., пока не придем к матрице A. Формулу (15) можно обобщить, взяв вместо u и v матрицы U и V, имеющие соответственно несколько столбцов и несколько строк.

§ 7. Линейные операторы. Нормы операторов

Прежде чем переходить к изучению методов последовательных приближений, мы рассмотрим некоторые свойства операторов, которые нам при этом потребуются.

Пусть A — аддитивный оператор, заданный в линейном нормированном пространстве [1]) H. Этот оператор называется *ограниченным*, если существует такая постоянная C, что для любого элемента $x \in H$ выполнено неравенство

$$\|Ax\| \leqslant C\|x\|. \qquad (1)$$

[1]) См. § 1, гл. 4.

Ограниченный, аддитивный оператор будем в дальнейшем называть *линейным*. Наименьшее из чисел C будем называть *нормой оператора* и обозначать $\|A\|$. Таким образом, норма оператора A есть такое число $\|A\|$, что при любом $x \in H$ выполнено

$$\|Ax\| \leqslant \|A\| \|x\|, \tag{2}$$

и с другой стороны, для любого $\varepsilon > 0$ найдется такой элемент $x' \in H$, что

$$\|Ax'\| > (\|A\| - \varepsilon)\|x'\|. \tag{3}$$

Можно дать другое определение нормы оператора, а именно положить

$$\|A\| = \sup_{\|x\|=1} \|Ax\|. \tag{4}$$

Нетрудно показать, что оба определения эквивалентны. Действительно, неравенства (2) и (3) можно переписать в виде

$$\|Ay\| \leqslant \|A\|, \quad \|Ay'\| > \|A\| - \varepsilon, \tag{5}$$

где $y = x/\|x\|$ и $y' = x'/\|x'\|$ — элементы с единичной нормой. Поэтому норма оператора в смысле первого определения будет равна норме в смысле второго определения. Наоборот, если пользоваться равенством (4) для определения нормы, то будем иметь:

$$\left\|A\left(\frac{x}{\|x\|}\right)\right\| \leqslant \|A\|, \quad \|Ax\| \leqslant \|A\| \|x\| \tag{6}$$

для любого $x \in H$, т. е. имеем неравенство (2). С другой стороны, в силу определения верхней границы, для любого $\varepsilon > 0$ найдется такой элемент $x' \in H$, $\|x'\| = 1$, что

$$\|Ax'\| > \|A\| - \varepsilon = (\|A\| - \varepsilon)\|x'\|. \tag{7}$$

Поэтому норма оператора в смысле второго определения будет равна норме оператора в смысле первого определения.

Рассмотрим совокупность всевозможных линейных операторов, определенных на H. Эти операторы аддитивны, и поэтому для них определены операции сложения и умножения на число. Покажем, что эти операции не выводят за пределы множества линейных операторов. По определению

$$C = A + B, \tag{8}$$

если для любого $x \in H$ имеет место

$$Cx = Ax + Bx. \tag{9}$$

При этом

$$\sup_{\|x\|=1} \|Cx\| \leqslant \sup_{\|x\|=1} \|Ax\| + \sup_{\|x\|=1} \|Bx\| = \|A\| + \|B\|. \tag{10}$$

Итак, C — ограниченный оператор и его норма удовлетворяет неравенству

$$\|C\|=\|A+B\|\leqslant\|A\|+\|B\|. \tag{11}$$

Далее, пусть

$$B=cA, \tag{12}$$

где c — число. Это значит, что для любого $x\in H$

$$Bx=cAx. \tag{13}$$

При этом

$$\sup_{\|x\|=1}\|Bx\|=\sup_{\|x\|=1}\|cAx\|=|c|\sup_{\|x\|=1}\|Ax\|=|c|\|A\|. \tag{14}$$

Итак, оператор B ограничен и его норма удовлетворяет равенству

$$\|B\|=\|cA\|=|c|\|A\|. \tag{15}$$

Утверждение доказано. Заметим еще, что $\|A\|\geqslant 0$ и $\|A\|=0$ тогда и только тогда, когда $A=0$, т. е. когда оператор A переводит любой элемент в нулевой. Первая часть утверждения тривиальна. Если же $\|A\|=0$, то

$$\sup_{\|x\|=1}\|Ax\|=0, \tag{16}$$

т. е. оператор A переводит каждый элемент $x\in H$, имеющий единичную норму, в нулевой элемент. Это же будет верно и для остальных элементов H в силу аддитивности оператора A.

Мы доказали, что для нормы линейных операторов выполнены все свойства, которые требуются от нормы в линейном нормированном пространстве. Таким образом, *совокупность всех линейных операторов, заданных в линейном нормированном пространстве, образует в свою очередь линейное нормированное пространство.*

Но в множестве линейных операторов определена еще операция умножения оператора на оператор. По определению

$$C=AB, \tag{17}$$

если для любого $x\in H$

$$Cx=ABx. \tag{18}$$

При этом по первому определению нормы оператора будем иметь:

$$\|Cx\|=\|ABx\|\leqslant\|A\|\|Bx\|\leqslant\|A\|\|B\|\|x\|. \tag{19}$$

Итак, оператор C ограничен и

$$\|C\|=\|AB\|\leqslant\|A\|\|B\|. \tag{20}$$

1. Конечномерные линейные нормированные пространства. В настоящей главе мы будем рассматривать только конечномерные линейные нормированные пространства. При этом каждый элемент линейного пространства будет полностью определяться конечной

совокупностью чисел — компонент его по отношению к некоторому базису. Пусть размерность пространства равна n. Норма элемента $x \in H$ будет функцией его компонент

$$\|x\| = \varphi(x_1, x_2, \ldots, x_n). \tag{21}$$

При различных определениях нормы функции φ будут различны, но все они будут обладать рядом общих свойств.

Так как $\|cx\| = |c|\|x\|$, то

$$\varphi(cx_1, cx_2, \ldots, cx_n) = |c|\varphi(x_1, x_2, \ldots, x_n). \tag{22}$$

Функция $\varphi(x_1, x_2, \ldots, x_n)$ непрерывна. Действительно, если x и x' — два элемента H с компонентами $(x_1, x_2, \ldots, x_n)$ и $(x'_1, x'_2, \ldots, x'_n)$, то

$$|\varphi(x_1, x_2, \ldots, x_n) - \varphi(x'_1, x'_2, \ldots, x'_n)| = (\|x\| - \|x'\|) \leqslant \|x - x'\|. \tag{23}$$

Обозначим

$$\varphi(\underbrace{0, \ldots, 0}_{i-1 \text{ раз}}, 1, 0, \ldots, 0) = \alpha_i. \tag{24}$$

Тогда

$$\|x - x'\| \leqslant |x_1 - x'_1|\alpha_1 + |x_2 - x'_2|\alpha_2 + \ldots + |x_n - x'_n|\alpha_n \tag{25}$$

или

$$|\varphi(x_1, x_2, \ldots, x_n) - \varphi(x'_2, x'_2, \ldots, x'_n)| \leqslant$$
$$\leqslant |x_1 - x'_1|\alpha_1 + |x_2 - x'_2|\alpha_2 + \ldots + |x_n - x'_n|\alpha_n. \tag{26}$$

Отсюда и следует непрерывность φ.

Непрерывная функция $\varphi(x_1, x_2, \ldots, x_n)$ достигает на ограниченном замкнутом множестве

$$x_1^2 + x_2^2 + \ldots + x_n^2 = 1 \tag{27}$$

своего наибольшего значения $M < \infty$ и своего наименьшего значения $m > 0$. Так как при положительных c

$$\varphi(cx_1, cx_2, \ldots, cx_n) = c\varphi(x_1, x_2, \ldots, x_n), \tag{28}$$

то

$$\varphi(x_1, x_2, \ldots, x_n) < M \quad \text{при} \quad x_1^2 + x_2^2 + \ldots + x_n^2 < 1 \tag{29}$$

и

$$\varphi(x_1, x_2, \ldots, x_n) > m \quad \text{при} \quad x_1^2 + x_2^2 + \ldots + x_n^2 > 1. \tag{30}$$

Отсюда, в частности, следует, что множество элементов, удовлетворяющих условию $\|x\| = 1$, является замкнутым ограниченным множеством пространства $R^{(n)}$. Это множество не содержит нулевого элемента.

Пусть в нашем пространстве введены две нормы $\|x\|_1$ и $\|x\|_2$. Рассмотрим множество элементов $y \in H$, для которых $\|y\|_1 = 1$. В силу только что сделанного замечания $\|y\|_2$ на этом множестве достигает своего наибольшего значения L и своего наименьшего значения l, $l \neq 0$. Пусть теперь $x \neq 0$ — произвольный элемент H. Тогда

$$\|x\|_2 = \left\| \|x\|_1 \frac{x}{\|x\|_1} \right\|_2 = \|x\|_1 \left\| \frac{x}{\|x\|_1} \right\|_2, \tag{31}$$

и так как $\|x/\|x\|_1\|_1 = 1$, то

$$l\|x\|_1 \leqslant \|x\|_2 \leqslant L\|x\|_1. \tag{32}$$

Отсюда получаем, что для произвольного ненулевого элемента $x \in H$ выполнены неравенства

$$l\|x\|_1 \leqslant \|x\|_2 \leqslant L\|x\|_1; \quad \frac{1}{L}\|x\|_2 \leqslant \|x\|_1 \leqslant \frac{1}{l}\|x\|_2. \tag{33}$$

Они выполнены и для нулевого элемента.

Две нормы $\|x\|_1$ и $\|x\|_2$, для которых выполнено

$$m\|x\|_1 \leqslant \|x\|_2 \leqslant M\|x\|_1; \quad n\|x\|_2 \leqslant \|x\|_1 \leqslant N\|x\|_2, \tag{34}$$

где x — произвольный элемент H и m, M, n, $N > 0$, называются *эквивалентными*. Неравенства (33) говорят о том, что имеет место теорема. *В конечномерном линейном нормированном пространстве любые две нормы эквивалентны.*

На практике чаще всего используют следующие нормы:

$$\|x\|_1 = \max_i |x_i|, \tag{35}$$

$$\|x\|_2 = |x_1| + |x_2| + \ldots + |x_n|, \tag{36}$$

$$\|x\|_3 = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}. \tag{37}$$

Необходимо проверить выполнимость условий, налагаемых на норму. Условия $\|x\| \geqslant 0$ и $\|x\| \neq 0$, если $x \neq 0$, очевидно, выполнены во всех трех случаях. Так же очевидно, что выполнено условие $\|cx\| = |c|\,\|x\|$. Проверим выполнение условия

$$\|x + y\| \leqslant \|x\| + \|y\|.$$

В первом случае будем иметь:

$$\|x + y\|_1 = \max_i |x_i + y_i| \leqslant$$

$$\leqslant \max_i |x_i| + \max_i |y_i| = \|x\|_1 + \|y\|_1. \tag{38}$$

Во втором случае получим:

$$\|x + y\|_2 = \sum_{i=1}^{n} |x_i + y_i| \leqslant \sum_{i=1}^{n} |x_i| + \sum_{i=1}^{n} |y_i| = \|x\|_2 + \|y\|_2. \tag{39}$$

Наконец, из неравенства Буняковского следует:

$$\|x+y\|_3 = \sqrt{\sum_{i=1}^{n}(x_i+y_i)^2} \leqslant$$

$$\leqslant \sqrt{\sum_{i=1}^{n} x_i^2} + \sqrt{\sum_{i=1}^{n} y_i^2} = \|x\|_3 + \|y\|_3. \qquad (40)$$

Заметим еще следующий факт. Если $x^{(n)} \to x$ в смысле какой-то из этих трех норм, то $x_i^{(n)} \to x_i$. Так как у нас все нормы эквивалентны, то последнее заключение будет справедливо, если $x^{(n)} \to x$ по произвольной норме.

2. Линейные операторы в конечномерном линейном нормированном пространстве и их связь с матрицами. Пусть в n-мерном линейном нормированном пространстве задан аддитивный оператор A. Пусть этот оператор переводит базисные элементы $(0, 0, \ldots, 0, 1, 0, \ldots, 0)$ соответственно в элементы $(a_{1i}, a_{2i}, \ldots, a_{ni})$. Тогда в силу аддитивности он должен переводить элемент $(x_1, x_2, \ldots, x_n)$ в элемент $(y_1, y_2, \ldots, y_n)$, где

$$y_i = \sum_{j=1}^{n} a_{ij}x_j. \qquad (41)$$

Это можно записать в виде матричного равенства

$$\begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix} = \begin{pmatrix} a_{11}\ a_{12} \ldots a_{1n} \\ a_{21}\ a_{22} \ldots a_{2n} \\ \cdots\cdots\cdots \\ a_{n1}\ a_{n2} \ldots a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix}. \qquad (42)$$

Таким образом, *каждому аддитивному оператору в n-мерном линейном нормированном пространстве будет соответствовать квадратная матрица порядка n*. Обратно, каждая квадратная матрица порядка n определяет при помощи равенства (42) некоторое отображение n-мерного пространства самого на себя. Нетрудно проверить, что это будет аддитивный оператор. Поэтому в дальнейшем мы часто будем отождествлять аддитивные операторы в n-мерном пространстве с соответствующими им матрицами.

Так как множество элементов $x \in H$, для которых $\|x\| = 1$ образует замкнутое ограниченное множество $R^{(n)}$, то непрерывная функция $\|Ax\|$ компонент $(x_1, x_2, \ldots, x_n)$ будет на нем ограничена. Таким образом, имеет место теорема: *все аддитивные операторы на конечномерном линейном нормированном пространстве будут ограниченными*. Поэтому в нашем случае в соответствии с вышеизложенным мы сможем ввести понятие нормы оператора или

нормы матрицы. Различным способам введения нормы элемента будут соответствовать различные определения нормы оператора или матрицы. Рассмотрим, как будут определяться нормы операторов или матриц для норм элементов, определенных равенствами (35) — (37).

Покажем, что если норма элемента определяется равенством (35), то норма A будет определяться

$$\|A\|_1 = \max_i \sum_{k=1}^{n} |a_{ik}|. \tag{43}$$

Действительно,

$$\|Ax\|_1 = \max_i \left| \sum_{k=1}^{n} a_{ik} x_k \right| \leqslant \max_i \sum_{k=1}^{n} |a_{ik}|\,|x_k|, \tag{44}$$

и если $\|x\|_1 = 1$, то

$$\|Ax\|_1 \leqslant \max_i \sum_{k=1}^{n} |a_{ik}|. \tag{45}$$

Пусть $\max\limits_i \sum\limits_{k=1}^{n} |a_{ik}|$ достигается при $i = j$. Возьмем в качестве x элемент с компонентами $x_k = \dfrac{|a_{jk}|}{a_{jk}}$ при $a_{jk} \neq 0$ и $x_k = 1$, если $a_{jk} = 0$. Очевидно, $\|x\|_1 = 1$. При этом

$$\left| \sum_{k=1}^{n} a_{jk} x_k \right| = \sum_{k=1}^{n} |a_{jk}|. \tag{46}$$

Следовательно,

$$\|Ax\|_1 = \sum_{k=1}^{n} |a_{jk}| = \max_i \sum_{k=1}^{n} |a_{ik}|. \tag{47}$$

Утверждение доказано.

Для второй нормы будем иметь:

$$\|A\|_2 = \max_k \sum_{i=1}^{n} |a_{ik}|. \tag{48}$$

Пусть $\|x\|_2 = 1$. Тогда

$$\|Ax\|_2 = \sum_{i=1}^{n} \left| \sum_{k=1}^{n} a_{ik} x_k \right| \leqslant \sum_{i=1}^{n} \sum_{k=1}^{n} |a_{ik}|\,|x_k| \leqslant$$

$$\leqslant \sum_{k=1}^{n} |x_k| \max_k \sum_{i=1}^{n} |a_{ik}|. \tag{49}$$

Пусть $\max\limits_k \sum\limits_{i=1}^{n} |a_{ik}|$ достигается при $k=j$. Возьмем элемент x с компонентами $x_k=0$ при $k\neq j$ и $x_j=1$. Очевидно, $\|x\|_2=1$. При этом

$$\|Ax\|_2=\sum_{i=1}^{n}\left|\sum_{k=1}^{n}a_{ik}x_k\right|=\sum_{i=1}^{n}|a_{ij}|=\max_k\sum_{i=1}^{n}|a_{ik}|. \tag{50}$$

Для третьей нормы

$$\|A\|_3=\sqrt{\lambda_1}, \tag{51}$$

где λ_1 — наибольшее собственное значение матрицы $A'A$. Покажем это. Пусть $\|x\|_3=1$. Имеем:

$$\|Ax\|_3^2=(Ax,\ Ax)=(x,\ A'Ax). \tag{52}$$

$A'A$ является симметрической неотрицательной матрицей. (Это значит, что для любого x скалярное произведение $(Ax,\ x)\geqslant 0$. Известно, что все собственные значения такой матрицы, т. е. такие значения λ, для которых существуют ненулевые x со свойством $Ax=\lambda x$, — действительные неотрицательные числа [1].) Пусть $\lambda_1\geqslant\lambda_2\geqslant\ldots\geqslant\lambda_n\geqslant 0$ — собственные значения этой матрицы, а $x^{(1)}$, $x^{(2)}$, ..., $x^{(n)}$ — соответствующие им ортонормированные действительные собственные векторы. При этом

$$x=c_1x^{(1)}+c_2x^{(2)}+\ldots+c_nx^{(n)}, \tag{53}$$

где $c_1^2+c_2^2+\ldots+c_n^2=1$ и

$$\|Ax\|_3^2=(x,\ A'Ax)=(c_1x^{(1)}+c_2x^{(2)}+\ldots+c_nx^{(n)},$$

$$\lambda_1c_1x_1^{(1)}+\lambda_2c_2x^{(2)}+\ldots+\lambda_nc_nx^{(n)})=$$

$$=\lambda_1c_1^2+\lambda_2c_2^2+\ldots+\lambda_nc_n^2\leqslant\lambda_1. \tag{54}$$

Для $x=x^{(1)}$ будем иметь:

$$\|Ax^{(1)}\|_3^2=(x^{(1)},\ A'Ax^{(1)})=(x^{(1)},\ \lambda_1x^{(1)})=\lambda_1. \tag{55}$$

Таким образом, утверждение доказано.

3. Сходимость последовательностей матриц и матричных рядов. Перейдем теперь к вопросу о сходимости последовательности матриц. Будем говорить, что последовательность матриц A_1, A_2, ..., A_m, ... сходится к матрице A, если для всех i, k имеем $a_{ik}^{(m)}\to a_{ik}$. Это равносильно условию $\|A_n-A\|\to 0$. Из этого определения следует, что если $A_m\to A$ и $B_m\to B$, то

$$A_m+B_m\to A+B, \tag{56}$$

$$A_mB_m\to AB. \tag{57}$$

---

[1]) См. И. М. Гельфанд, Лекции по линейной алгебре, гл. II.

Далее, если T — некоторая неособенная постоянная матрица и $A_m \to A$, то

$$T^{-1}A_mT \to T^{-1}AT. \tag{58}$$

Рассмотрим степенной ряд

$$f(\lambda) = \alpha_0 + \alpha_1\lambda + \alpha_2\lambda^2 + \ldots + \alpha_m\lambda^m + \ldots \tag{59}$$

Ему можно поставить в соответствие матричный ряд

$$f(A) = \alpha_0 I + \alpha_1 A + \alpha_2 A^2 + \ldots + \alpha_m A^m + \ldots \tag{60}$$

Этот ряд будет называться сходящимся, если сходится последовательность его частичных сумм $f_m(A) = \alpha_0 I + \alpha_1 A + \alpha_2 A^2 + \ldots + \alpha_m A^m$. Найдем условия сходимости такого ряда. Имеет место теорема:

Для того чтобы степенной матричный ряд (60) *сходился, необходимо и достаточно, чтобы все собственные значения* λ_i *матрицы* A *лежали внутри круга сходимости степенного ряда* (59).

Доказательство. Приведем матрицу A к нормальной форме Жордана при помощи некоторой неособенной матрицы T: $T^{-1}AT = B$. Тогда

$$T^{-1}f_m(A)\,T = f_m(T^{-1}AT) = f_m(B). \tag{61}$$

Для того чтобы $f_m(A)$ имела предел, необходимо и достаточно, чтобы $f_m(B)$ имело предел. Предел $f_m(B)$ будет существовать тогда и только тогда, когда существуют все $\lim\limits_{m\to\infty} f_m(B_i)$, где B_i — «ящики» матрицы B. Пусть B_i соответствует элементарному делителю $(\lambda - \lambda_i)^{n_i}$. Тогда

$$B_i = \begin{Vmatrix} \lambda_i & 1 & 0 & \ldots & 0 \\ 0 & \lambda_i & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & \lambda_i \end{Vmatrix} \quad (n_i \text{ строк}), \tag{62}$$

а

$$B_i^2 = \begin{Vmatrix} \lambda_i & 1 & 0 & \ldots & 0 \\ 0 & \lambda_i & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & \lambda_i \end{Vmatrix} \begin{Vmatrix} \lambda_i & 1 & 0 & \ldots & 0 \\ 0 & \lambda_i & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & \lambda_i \end{Vmatrix} =$$

$$= \begin{pmatrix} \lambda_i^2 & 2\lambda_i & 1 & 0 & \ldots & 0 \\ 0 & \lambda_i^2 & 2\lambda_i & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & \lambda_i^2 \end{pmatrix}. \tag{63}$$

По индукции нетрудно доказать, что при $m > n_i$.

$$B_i^m = \begin{Vmatrix} \lambda_i^m & C_m^1\lambda_i^{m-1} & C_m^2\lambda_i^{m-2} & \dots & C_m^{n_i-1}\lambda_i^{m-n_i+1} \\ 0 & \lambda_i^m & C_m^1\lambda_i^{m-1} & \dots & C_m^{n_i-2}\lambda_i^{m-n_i+2} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \lambda_i^m \end{Vmatrix}. \tag{64}$$

Отсюда

$$f_m(B_i) = \begin{Vmatrix} f_m(\lambda_i) & f'_m(\lambda_i) & \frac{1}{2!}f''_m(\lambda_i) & \dots & \frac{1}{(n_i-1)!}f_m^{(n_i-1)}(\lambda_i) \\ 0 & f_m(\lambda_i) & f'_m(\lambda_i) & \dots & \frac{1}{(n_i-2)!}f_m^{(n_i-2)}(\lambda_i) \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & f_m(\lambda_i) \end{Vmatrix}. \tag{65}$$

Для того чтобы $f_m(B_i)$ была сходящейся последовательностью, необходимо и достаточно, чтобы $f_m(\lambda_i)$, $f'_m(\lambda_i)$ ... $f_m^{(n_i-1)}(\lambda_i)$ были сходящимися последовательностями. Последнее будет выполнено в том и только в том случае, если λ_i лежат внутри круга сходимости ряда $f(\lambda)$.

Возьмем, в частности, ряд

$$f(\lambda) = 1 + \lambda + \lambda^2 + \dots + \lambda^m + \dots \tag{66}$$

Его радиус сходимости равен 1, и ряд расходится в каждой точке границы круга сходимости. Следовательно, имеет место теорема. *Для того чтобы сходился матричный ряд*

$$f(A) = I + A + A^2 + \dots + A^m + \dots, \tag{67}$$

необходимо и достаточно, чтобы собственные значения матрицы A были по модулю меньше единицы.

Сделаем еще одно небольшое замечание. Пусть λ — произвольное собственное значение матрицы A. Тогда имеется ненулевой вектор x такой, что

$$Ax = \lambda x. \tag{68}$$

Следовательно,

$$\|Ax\| = |\lambda|\,\|x\|. \tag{69}$$

Но $\|Ax\| \leqslant \|A\|\,\|x\|$. Поэтому получаем $\|A\| \geqslant |\lambda|$, т. е. *любая норма матрицы больше или равна модулю произвольного собственного значения матрицы*, т. е.

$$\max_i |\lambda_i| \leqslant \|A\|.$$

§ 8. Разновидности методов последовательных приближений

В методах последовательных приближений решения системы $A\bar{x} = \bar{b}$ исходят из произвольного начального вектора $\bar{x}_0$ и получают векторы $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$ по рекуррентной формуле

$$\bar{x}_{k+1} = F_k(\bar{x}_0, \bar{x}_1, \ldots, \bar{x}_k), \tag{1}$$

где F_k — некоторая функция, зависящая, вообще говоря, от матрицы системы A, правой части $\bar{b}$, номера приближения k и предыдущих приближений $\bar{x}_0, \bar{x}_1, \ldots, \bar{x}_k$. Будем говорить, что *метод имеет первый порядок*, если F_k зависит только от $\bar{x}_k$ и не зависит от $\bar{x}_0, \bar{x}_1, \ldots, \bar{x}_{k-1}$. Будем называть метод *стационарным*, если F_k не зависит от k.

Простейшим случаем функций F_k будут линейные функции. Наиболее *общий линейный метод* последовательных приближений первого порядка должен иметь вид

$$\bar{x}_{k+1} = B_k\bar{x}_k + \bar{c}_k, \tag{2}$$

где B_k — квадратная матрица и $\bar{c}_k$ — вектор. Естественно требовать от методов последовательных приближений, что при подстановке в правую часть (1) и (2) вместо $\bar{x}_k$ точного решения системы $A^{-1}\bar{b}$ мы слева снова получили $A^{-1}\bar{b}$. В случае линейного метода первого порядка это приведет к равенству

$$A^{-1}\bar{b} = B_kA^{-1}\bar{b} + \bar{c}_k \tag{3}$$

или

$$\bar{c}_k = (I - B_k)\,A^{-1}\bar{b} = C_k\bar{b}. \tag{4}$$

В этом случае мы можем переписать (2) в виде

$$\bar{x}_{k+1} = B_k\bar{x}_k + C_k\bar{b}, \tag{5}$$

причем B_k и C_k — квадратные матрицы, не зависящие от $\bar{b}$ и такие, что

$$B_k + C_kA = I. \tag{6}$$

Выражение (5) можно также записать в виде

$$\bar{x}_{k+1} = \bar{x}_k - C_k(A\bar{x}_k - \bar{b}). \tag{7}$$

Наконец, если существует матрица C_k^{-1}, то выражение (5) можно записать в виде

$$D_k\bar{x}_{k+1} + E_k\bar{x}_k = \bar{b}. \tag{8}$$

При этом должно быть

$$D_k + E_k = A. \tag{9}$$

При пользовании формулой (8) мы получим $\bar{x}_{k+1}$ в неявной форме. Поэтому желательно, чтобы матрицу D_k было легко обратить. На практике обычно берут D_k диагональной или треугольной. В первом случае метод называют *полношаговым*, во втором — *одношаговым*.

Разнообразные схемы, осуществляющие линейные методы последовательных приближений первого порядка, можно считать реализацией формул (5), (7) или (8).

Много линейных и нелинейных методов последовательных приближений можно получить, используя *способ наименьших квадратов*. При этом минимизируется функция

$$f(\bar{x}) = \|A\bar{x} - \bar{b}\|^2 \tag{10}$$

или же $f(\bar{x})$, сложенная с некоторой постоянной. Разным способам минимизации и задания нормы будут соответствовать различные методы решения систем уравнений. Как мы видели в § 5, в случае симметрической положительно определенной матрицы A можно минимизировать функцию

$$f(\bar{x}) = (\bar{x},\ A\bar{x}) - 2(\bar{b},\ \bar{x}). \tag{11}$$

Один из способов такой минимизации там был рассмотрен. Этот способ можно обобщить. Выбираем какое-то направление, определяемое вектором $\bar{c}$. Так же как и в § 5, найдем, что $f(\bar{x} + \alpha\bar{c})$ принимает наименьшее значение при $\alpha = \alpha^*$, где

$$\alpha^* = \frac{(\bar{c},\ \bar{b} - A\bar{x})}{(\bar{c},\ A\bar{c})}. \tag{12}$$

Если возьмем вместо α^* величину $\alpha = \beta\alpha^*$, то будем иметь:

$$f(\bar{x}) - f(\bar{x} + \alpha\bar{c}) = (2\alpha\alpha^* - \alpha^2)(\bar{c},\ A\bar{c}) = \beta(2 - \beta)\alpha^{*2}(\bar{c},\ A\bar{c}). \tag{13}$$

Таким образом, если $(\bar{c},\ \bar{b} - A\bar{x}) \neq 0$, то $f(\bar{x}) > f(\bar{x} + \alpha\bar{c})$ при любом β $(0 < \beta < 2)$.

Идея многих способов минимизации $f(\bar{x})$, а следовательно и решения системы $A\bar{x} = \bar{b}$, заключается в следующем. Задаются начальным приближением $\bar{x}_0$. Определяют закон выбора векторов $\bar{c}_k$ (они могут зависеть от $\bar{x}_k$). Определяют закон выбора коэффициентов β_k (они могут зависеть от $\bar{x}_k$). На каждом шаге в качестве последующего вектора рассматривают

$$\bar{x}_{k+1} = \bar{x}_k + \beta_k\alpha_k^*\bar{c}_k, \tag{14}$$

где α_k^* было определено выше.

Если A не является симметрической положительно определенной матрицей, то можно вместо системы $A\bar{x} = \bar{b}$ рассматривать систему $A'RA\bar{x} = A'R\bar{b}$, где R — симметрическая положительно определенная

матрица. Часто в качестве R берут единичную матрицу I. Это, конечно, не означает, что во всех случаях придется делать предварительное преобразование системы. Бывает достаточно использовать преобразованную систему лишь для построения алгоритма.

§ 9. Линейные полношаговые методы первого порядка

1. Сходимость линейных полношаговых методов первого порядка. Простая итерация. Линейные полношаговые методы первого порядка определяются формулами (5) и (7) предыдущего параграфа. Исследуем сначала сходимость этих методов. Из формулы (5) следует:

$$\bar{x}_{k+1} - A^{-1}\bar{b} = B_k(\bar{x}_k - A^{-1}\bar{b}). \tag{1}$$

Применяя формулу (1) при $k = 0, 1, \ldots, m$, получим:

$$\bar{x}_{m+1} - A^{-1}\bar{b} = B_m B_{m-1} \ldots B_1 B_0 (\bar{x}_0 - A^{-1}\bar{b}). \tag{2}$$

Следовательно,

$$\|\bar{x}_{m+1} - A^{-1}\bar{b}\| \leqslant \|B_m\| \|B_{m-1}\| \ldots \|B_1\| \|B_0\| \|\bar{x}_0 - A^{-1}\bar{b}\|. \tag{3}$$

Если $\|B_m\| \|B_{m-1}\| \ldots \|B_1\| \|B_0\|$ стремится к нулю при $m \to \infty$, то $\|\bar{x}_{m+1} - A^{-1}\bar{b}\|$ стремится к нулю при любом начальном векторе $\bar{x}_0$. Как мы видели в § 7, из этого будет следовать, что все компоненты $\bar{x}_{m+1}$ будут стремиться к соответствующим компонентам $A^{-1}\bar{b}$. Для того чтобы произведение норм матриц, стоящее в правой части (3), стремилось к нулю, достаточно потребовать

$$\|B_k\| \leqslant \gamma < 1 \qquad (k = 0, 1, 2, \ldots). \tag{4}$$

В частности, если процесс стационарен, т. е. B_k не зависят от k, $B_k = B$; последнее условие означает, что какая-то из норм B меньше единицы. Однако для стационарного случая можно дать более точные условия, а именно докажем т е о р е м у:

Стационарный линейный процесс

$$\bar{x}_{k+1} = B\bar{x}_k + C\bar{b} \tag{5}$$

сходится при любом начальном векторе и любой правой части тогда и только тогда, когда все собственные значения матрицы B по модулю меньше единицы.

Действительно, легко проверить по индукции, что

$$\bar{x}_{k+1} = B^{k+1}\bar{x}_0 + (B^k + B^{k-1} + \ldots + B + I) C\bar{b}. \tag{6}$$

Очевидно, последовательность

$$(B^k + B^{k-1} + \ldots + B + I) C\bar{b} \qquad (k = 0, 1, 2, \ldots) \tag{7}$$

будет сходящейся для произвольного вектора $\bar{b}$ в том и только в том случае, если сходится матричный ряд

$$I + B + B^2 + \ldots + B^k + \ldots \tag{8}$$

Но при этом $B^{k+1} \to 0$. Следовательно, $\bar{x}_{k+1}$ образуют сходящуюся последовательность при произвольных векторах $\bar{x}_0$ и $\bar{b}$ в том и только в том случае, когда сходится ряд (8). Как мы видели, ряд (8) сходится в том и только в том случае, если все собственные значения матрицы B по модулю меньше единицы. Утверждение доказано.

Полученное нами условие хорошо при теоретических рассуждениях, так как точно отражает положение вещей. Однако оно неудобно для практических применений, ибо в большинстве случаев собственные значения нам неизвестны, а отыскание их представляет задачу более сложную, чем решение системы линейных алгебраических уравнений. В главе 8 мы дадим ряд способов оценки максимального по модулю собственного значения. Пока же будем использовать нормы матриц, данные в § 7, и неравенство $\max|\lambda_i| \leqslant \|B\|$. При этом получим следующие *три достаточных условия*:

$$\sum_{j=1}^{n} |b_{ij}| \leqslant \mu < 1 \qquad (i = 1, 2, \ldots, n), \tag{9}$$

$$\sum_{i=1}^{n} |b_{ij}| \leqslant \mu < 1 \qquad (j = 1, 2, \ldots, n), \tag{10}$$

$$\sum_{i,\,j=1}^{n} b_{ij}^2 \leqslant \mu < 1. \tag{11}$$

Пояснений требует только последнее условие. Оно обеспечивает, что норма $\|B\|_3$, равная наибольшему собственному значению λ_1 матрицы $B'B$, не превышает единицы. Действительно, все собственные значения матрицы $B'B$ неотрицательны. Поэтому $\lambda_1 \leqslant \lambda_1 + \lambda_2 + \ldots + \lambda_n$. Но последняя сумма равна следу матрицы, который и равен $\sum_{i,\,j=1}^{n} b_{ij}^2$.

Изложенный метод часто называют *простой итерацией*.

Для применения простой итерации необходимо предварительное преобразование системы, заданной в виде $A\bar{x} = \bar{b}$, к виду (5). Это можно сделать, например, так. Каждое из уравнений системы

$$a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n = b_i \qquad (i = 1, 2, \ldots, n) \tag{12}$$

делим на a_{ii} и переносим члены с $x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ в правую часть равенства. При этом i-е уравнение примет вид

$$x_i = \frac{b_i}{a_{ii}} - \frac{a_{i1}}{a_{ii}}x_1 - \frac{a_{i2}}{a_{ii}}x_2 - \ldots$$

$$\ldots - \frac{a_{i,\,i-1}}{a_{ii}}x_{i-1} - \frac{a_{i,\,i+1}}{a_{ii}}x_{i+1} - \ldots - \frac{a_{in}}{a_{ii}}x_n. \tag{13}$$

Для того чтобы этот прием был осуществим, коэффициенты a_{ii} должны быть отличны от нуля. Кроме того, для обеспечения

сходимости требуется значительное преобладание диагональных элементов над остальными коэффициентами. Так, неравенства (9) — (11) будут выполнены, если для коэффициентов будут выполнены следующие неравенства:

$$\sum_{\substack{j=1\\(j\neq i)}}^{n} |a_{ij}| < |a_{ii}| \qquad (i = 1,\ 2,\ \ldots,\ n), \tag{14}$$

$$\sum_{\substack{i=1\\(i\neq j)}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| < 1 \qquad (j = 1,\ 2,\ \ldots,\ n), \tag{15}$$

$$\sum_{j=1}^{n} \frac{1}{a_{jj}^2} \sum_{\substack{i=1\\(i\neq j)}}^{n} |a_{ij}|^2 < 1. \tag{16}$$

Приведем пример решения системы методом простой итерации. Опять будем решать ту же систему, которую мы уже использовали в § 2 и 3. В верхней части схемы стоят коэффициенты преобразованной системы. Далее идет начальное приближение, в качестве которого взяты свободные члены преобразованной системы, и идут последующие приближения:

	x_1	x_2	x_3	x_4	x_5	x_6
	0	— 0,029409	— 0,050810	— 0,022890	— 0,024524	— 0,034634
	— 0,025314	0	— 0,029811	— 0,025272	— 0,021248	— 0,043736
	— 0,038103	— 0,025972	0	— 0,014727	— 0,030521	— 0,031758
	— 0,015192	— 0,019487	— 0,013034	0	— 0,033765	— 0,073469
	— 0,028541	— 0,028730	— 0,047368	— 0,059210	0	— 0,169430
	— 0,051824	— 0,075376	— 0,063370	— 0,165638	— 0,217801	0
$\bar{x}^{(0)}$	1,161765	1,147832	1,129872	0,570275	0,777788	0,770121
$\bar{x}^{(1)}$	1,011799	1,020120	0,999199	0,432689	0,493906	0,287933
$\bar{x}^{(2)}$	1,049027	1,058410	1,034234	0,484170	0,598888	0,398231
$\bar{x}^{(3)}$	1,038527	1,048077	1,024356	0,470754	0,572321	0,359803
$\bar{x}^{(4)}$	1,041622	1,051212	1,027253	0,474964	0,580689	0,369760
$\bar{x}^{(5)}$	1,040736	1,050327	1,026420	0,473804	0,578438	0,366660
$\bar{x}^{(6)}$	1,040994	1,050587	1,026661	0,474149	0,579122	0,367508
$\bar{x}^{(7)}$	1,040920	1,051413	1,026592	0,474051	0,578931	0,367254
$\bar{x}^{(8)}$	1,040915	1,050535	1,026588	0,474062	0,578962	0,367257
$\bar{x}^{(9)}$	1,040940	1,050534	1,026610	0,474078	0,578986	0,367315
$\bar{x}^{(10)}$	1,040936	1,050529	1,026606	0,474073	0,578974	0,367305
$\bar{x}^{(11)}$	1,040937	1,050530	1,026607	0,474074	0,578976	0,367309
$\bar{x}^{(12)}$	1,040936	1,050530	1,026607	0,474074	0,578975	0,367308

Следует отметить, что ошибки, допущенные при вычислении, будут в дальнейшем исключены последующими приближениями. Однако при этом могут потребоваться лишние приближения.

Оценим теперь погрешность метода простой итерации. Для этого воспользуемся равенством

$$\bar{x}_{k+1} - A^{-1}\bar{b} = B\left(\bar{x}_k - A^{-1}\bar{b}\right). \tag{17}$$

Так как

$$A^{-1}\bar{b} = (I + B + B^2 + \ldots)C\bar{b}, \tag{18}$$

а

$$\bar{x}_k = (I + B + B^2 + \ldots + B^{k-1})C\bar{b} + B^k\bar{x}_0, \tag{19}$$

то

$$\bar{x}_{k+1} - A^{-1}\bar{b} = (B^{k+1} + B^{k+2} + \ldots)C\bar{b} + B^{k+1}\bar{x}_0. \tag{20}$$

Отсюда

$$\|\bar{x}_{k+1} - A^{-1}\bar{b}\| \leqslant \frac{\|B\|^{k+1}}{1 - \|B\|}\|C\bar{b}\| + \|B\|^{k+1}\|\bar{x}_0\|. \tag{21}$$

2. Метод Ричардсона. Рассмотрим еще один метод последовательных приближений первого порядка. Возьмем в формуле (7) предыдущего параграфа $-C_k = \beta_k I$. При этом получим:

$$\bar{x}_{k+1} = \bar{x}_k + \beta_k(A\bar{x}_k - \bar{b}). \tag{22}$$

Проведем подробное исследование этого метода для случая, когда матрица A симметрическая и положительно определенная. В этом случае имеется n положительных собственных значений $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n$ матрицы A и n соответствующих им взаимно ортогональных собственных векторов $\bar{u}_1, \bar{u}_2, \ldots, \bar{u}_n$. Разложим вектор $\bar{x}_0 - A^{-1}\bar{b}$ по этим собственным векторам

$$\bar{x}_0 - A^{-1}\bar{b} = \sum_{i=1}^{n} c_i\bar{u}_i \tag{23}$$

и преобразуем (22) к виду

$$\bar{x}_{k+1} - A^{-1}\bar{b} = \bar{x}_k - A^{-1}\bar{b} + \beta_k A\left(\bar{x}_k - A^{-1}\bar{b}\right) = (I + \beta_k A)\left(\bar{x}_k - A^{-1}\bar{b}\right). \tag{24}$$

Тогда

$$\left.\begin{aligned}
&\bar{x}_1 - A^{-1}\bar{b} = (I + \beta_0 A)\left(\bar{x}_0 - A^{-1}\bar{b}\right) = \sum_{i=1}^{n} c_i(1 + \beta_0\lambda_i)\bar{u}_i,\\
&\bar{x}_2 - A^{-1}\bar{b} = (I + \beta_1 A)\left(\bar{x}_1 - A^{-1}\bar{b}\right) = \sum_{i=1}^{n} c_i(1 + \beta_0\lambda_i)(1 + \beta_1\lambda_i)\bar{u}_i,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&\bar{x}_{k+1} - A^{-1}\bar{b} = (I + \beta_k A)\left(\bar{x}_k - A^{-1}\bar{b}\right) = \sum_{i=1}^{n} c_i \prod_{j=0}^{k}(1 + \beta_j\lambda_j)\bar{u}_i,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned}\right\} \tag{25}$$

Таким образом,

$$(\bar{x}_{k+1} - A^{-1}\bar{b},\ \bar{x}_{k+1} - A^{-1}\bar{b}) = \\ = \|\bar{x}_{k+1} - A^{-1}\bar{b}\|_0^2 \leqslant M_k \|\bar{x}_0 - A^{-1}\bar{b}\|_3^2, \tag{26}$$

где

$$M_k = \max_{0 \leqslant i \leqslant n} \left| \prod_{j=0}^{k} (1 + \beta_j \lambda_i) \right|. \tag{27}$$

Зафиксируем k и будем подбирать β_j так, чтобы M_k приняло возможно меньшее значение. При этом мы будем предполагать, что нам каким-то образом удалось найти такие a и b, что

$$a \leqslant \lambda_i \leqslant b \qquad (i = 1, 2, \ldots, n). \tag{28}$$

Заменим отрезок $[a, b]$ отрезком $[-1, +1]$, введя новое переменное

$$t = \frac{2\lambda}{b-a} - \frac{b+a}{b-a}. \tag{29}$$

При этом многочлен

$$P_{k+1}(\lambda) = \prod_{j=0}^{k} (1 + \beta_j \lambda) \tag{30}$$

перейдет в новый многочлен $Q_{k+1}(t)$. Так как $P_{k+1}(0) = 1$, то $Q_{k+1}\left(\frac{b+a}{a-b}\right) = 1$. Таким образом, перед нами возникает задача об отыскании многочлена степени $k+1$, равного 1 при $t = \frac{b+a}{a-b}$ и обладающего наименьшим максимумом модуля на отрезке $[-1, +1]$ среди всех многочленов, обладающих такими свойствами. Эту задачу решает многочлен (см. упражнения к гл. 4)

$$R_{k+1}(t) = \frac{T_{k+1}(t)}{T_{k+1}\left(\frac{b+a}{a-b}\right)}, \tag{31}$$

где $T_{k+1}(t)$ — многочлен Чебышева, наименее уклоняющийся от нуля. Корни многочлена $R_{k+1}(t)$ совпадают с корнями $T_{k+1}(t)$ и расположены в точках

$$t_i = \cos \frac{(2i-1)\pi}{2(k+1)}. \tag{32}$$

Корни же многочлена $P_{k+1}(\lambda)$ расположены в точках $-\beta_i^{-1}$. Следовательно,

$$\beta_i = 2\left[(a-b)t_i - (a+b)\right]^{-1}. \tag{33}$$

При этом

$$M_k = \max_{-1 \leqslant t \leqslant 1} |R_{k+1}(t)| = \left[T_{k+1}\left(\frac{b+a}{a-b}\right)\right]^{-1} < 1. \tag{34}$$

Если заранее не ясно, сколько шагов потребуется сделать для получения нужной точности, то целесообразно использовать β_i в ци-

клическом порядке. Задаемся каким-то k, подбираем соответствующие β_i по формуле (33) и производим вычисления по формуле (22), беря β_i в следующем порядке: $\beta_0, \beta_1, \ldots, \beta_k, \beta_0, \beta_2, \ldots, \beta_k, \ldots$ При $k=0$ процесс будет стационарным, при $k>0$ процесс будет нестационарным. Он всегда будет сходящимся в силу неравенств (26) и (34). Такой способ был впервые предложен Ричардсоном, и поэтому мы назовем его *методом Ричардсона.*

3. Обращение матриц методом последовательных приближений. Коснемся еще вопроса об обращении матриц методом последовательных приближений. Пусть нам удалось каким-то способом найти приближенное значение B_0 для матрицы A^{-1}. Предположим, далее, что некоторая норма матрицы

$$C_0 = I - AB_0 \tag{35}$$

меньше единицы, $\|C_0\| \leqslant q < 1$. Образуем тогда последовательности:

$$\left.\begin{array}{ll} B_1 = B_0(I + C_0), & C_1 = I - AB_1; \\ B_2 = B_1(I + C_1), & C_2 = I - AB_2; \\ \cdots\cdots\cdots & \cdots\cdots\cdots \\ B_{k+1} = B_k(I + C_k), & C_{k+1} = I - AB_{k+1}; \\ \cdots\cdots\cdots\cdots\cdots & \end{array}\right\} \tag{36}$$

при этом

$$C_k = I - AB_k = I - AB_{k-1}(I + C_{k-1}) = I - (I - C_{k-1})(I + C_{k-1}) = \\ = C_{k-1}^2 = C_{k-2}^4 = \ldots = C_0^{2^k}. \tag{37}$$

Таким образом, матрица C_k будет очень быстро стремиться к нулевой и, следовательно, матрица B_k к A^{-1}. Оценим норму разности $\|B_k - A^{-1}\|$. Имеем:

$$\|B_k - A^{-1}\| = \|A^{-1}(I - C_k) - A^{-1}\| = \|A^{-1}C_k\| = \\ = \|B_0(I - C_0)^{-1}C_0^{2^k}\| \leqslant \|B_0\| \cdot \|(I - C_0)^{-1}\| \cdot \|C_0\|^{2^k} \leqslant \|B_0\| \frac{q^{2^k}}{1-q}. \tag{38}$$

§ 10. Линейные одношаговые методы первого порядка

Перейдем теперь к изучению линейных одношаговых методов первого порядка. Большое количество таких методов можно получить следующим образом. Представляем матрицу A в виде суммы трех матриц:

$$A = B + C + D. \tag{1}$$

Здесь B — диагональная матрица, в матрице C равны нулю элементы, лежащие на и выше главной диагонали, а в матрице D равны

нулю элементы, лежащие на и под главной диагональю. Выбираем какое-то число $\omega \neq 0$ и осуществляем итерационный процесс по формуле

$$(\omega^{-1}B + C)\,\bar{x}^{(k+1)} + [(1-\omega^{-1})]\,B + D]\,\bar{x}^{(k)} = \bar{b}. \qquad (2)$$

Если разрешить (2) относительно $x_i^{(k+1)}$, то получим:

$$x_i^{(k+1)} = -\frac{\omega}{a_{ii}}\left[\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} + \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} - b_i\right] - (\omega - 1)\,x_i^{(k)}. \qquad (3)$$

Особенно часто используется случай $\omega = 1$. При этом получается так называемый *метод Зейделя*. Рассмотрим подробно этот метод.

1. Метод Зейделя. Метод Зейделя отличается от простой итерации тем, что, найдя какое-то приближение для компоненты, мы сразу же используем его для отыскания следующей компоненты. Вычисления ведутся по формуле

$$x_i^{(k+1)} = -\sum_{j=1}^{i-1}\frac{a_{ij}}{a_{ii}}x_j^{(k+1)} - \sum_{j=i+1}^{n}\frac{a_{ij}}{a_{ii}}x_j^{(k)} + \frac{b_i}{a_{ii}}. \qquad (4)$$

По начальному приближению $(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})$ находим $x_1^{(1)}$. Затем по $(x_1^{(1)}, x_2^{(0)}, \ldots, x_n^{(0)})$ находим $x_2^{(1)}$ и т. д. После того как будут найдены все $x_i^{(1)}$, таким же образом находим $x_i^{(2)}$, $x_i^{(3)}$, ..., пока не достигнем нужной точности.

Решим ту же систему, которая рассматривалась в предыдущем параграфе, методом Зейделя. Мы не будем повторять запись коэффициентов, а приведем лишь результаты вычислений:

	x_1	x_2	x_3	x_4	x_5	x_6
$\bar{x}^{(0)}$	1,161765	1,147832	1,129872	0,970275	0,777788	0,770121
$\bar{x}^{(1)}$	1,011799	1,020120	0,999199	0,432689	0,493906	0,287933
$\bar{x}^{(2)}$	1,049027	1,057467	1,031845	0,482485	0,591246	0,361969
$\bar{x}^{(3)}$	1,040158	1,049654	1,026331	0,474084	0,579940	0,367221
$\bar{x}^{(4)}$	1,040955	1,050021	0,026593	0,474057	0,579006	0,367343
$\bar{x}^{(5)}$	1,040950	1,050047	1,026618	0,474079	0,578982	0,367341
$\bar{x}^{(6)}$	1,040949	1,050528	1,026606	0,474071	0,578970	0,367341
$\bar{x}^{(7)}$	1,040937	1,050530	1,026607	0,474074	0,578975	0,367308
$\bar{x}^{(8)}$	1,040936	1,050530	1,026607	0,474074	0,578975	0,367308

Заметим, что здесь, так же как и при простой итерации, можно было бы сократить вычисления и записи. Прежде всего несколько

первых приближений можно было проводить с меньшим количеством знаков. Наоборот, в последних приближениях, когда старшие разряды уже установились, нет необходимости выписывать их вновь. Последовательные приближения продолжаются обычно до тех пор, пока два следующих друг за другом приближения не станут совпадать.

2. Сходимость метода Зейделя. Исследуем теперь сходимость метода Зейделя. Разрешая (2) при $\omega = 1$ относительно $\bar{x}^{(k+1)}$, получим:

$$\bar{x}^{(k+1)} = -(B+C)^{-1}D\bar{x}^{(k)} + (B+C)^{-1}\bar{b}. \tag{5}$$

Это означает, что метод Зейделя эквивалентен простой итерации с матрицей $-(B+C)^{-1}D$. Таким образом, чтобы метод сходился необходимо и достаточно, чтобы все собственные значения этой матрицы были по модулю меньше единицы. Таким образом, должны быть по модулю меньше единицы все значения λ, удовлетворяющие уравнению

$$|\lambda I + (B+C)^{-1}D| = 0. \tag{6}$$

Но корни этого уравнения будут совпадать с корнями уравнения

$$|\lambda(B+C) + D| = 0. \tag{7}$$

Итак мы доказали теорему. *Для сходимости метода Зейделя необходимости достаточно, чтобы все корни уравнения*

$$\begin{vmatrix} a_{11}\lambda & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21}\lambda & a_{22}\lambda & a_{23} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1}\lambda & a_{n2}\lambda & a_{n3}\lambda & \dots & a_{nn}\lambda \end{vmatrix} = 0 \tag{8}$$

были по модулю меньше единицы.

Области сходимости простой итерации и итерации по Зейделю лишь пересекаются. Это значит, что существуют такие матрицы, для которых метод Зейделя сходится, а метод простой итерации нет, и наоборот. Нетрудно показать, что первое и второе достаточные условия (9) и (10) § 9 для сходимости метода простой итерации будут одновременно достаточными условиями и для сходимости процесса Зейделя.

Иногда метод Зейделя дает более быструю сходимость, чем простая итерация. Так будет, например, если выполнено условие (14) предыдущего параграфа. Обозначим

$$\mu = \max_i \frac{\sum_{\substack{j=1 \\ (i \neq j)}}^{n} |a_{ij}|}{|a_{ii}|}. \tag{9}$$

Так как условие (14) выполнено, то $\mu < 1$. Обозначим также

$$-\frac{a_{ij}}{a_{ii}} = c_{ij}, \quad \frac{b_i}{a_{ii}} = d_i. \tag{10}$$

Тогда простой итерации будет соответствовать вычислительная схема

$$x_i^{(k+1)} = \sum_{\substack{j=1 \\ (i \neq j)}}^{n} c_{ij} x_j^{(k)} + d_i, \tag{11}$$

а методу Зейделя схема

$$x_i^{(k+1)} = \sum_{j=1}^{i-1} c_{ij} x_j^{(k+1)} + \sum_{j=i+1}^{n} c_{ij} x_j^{(k)} + d_i. \tag{12}$$

По (11) и (9) разность между точным решением $\bar{x}$ и $(k+1)$-м приближением, полученным простой итерацией, будет иметь оценку

$$\|\bar{x} - \bar{x}^{(k+1)}\|_1 \leqslant \mu \|\bar{x} - \bar{x}^{(k)}\|_1. \tag{13}$$

В то же время, если ввести обозначения

$$\sum_{j=1}^{i-1} |c_{ij}| = \beta_i; \quad \sum_{j=i+1}^{n} |c_{ij}| = \gamma_i; \quad \max_i \frac{\gamma_i}{1-\beta_i} = \mu', \tag{14}$$

то для разности между точным решением $\bar{x}$ и $(k-1)$-м приближением, полученным по методу Зейделя, получим оценку

$$|x_i - x_i^{(k+1)}| \leqslant \beta_i \|\bar{x} - \bar{x}^{(k+1)}\|_1 + \gamma_i \|\bar{x} - \bar{x}^{(k)}\|_1 \tag{15}$$

и

$$\|\bar{x} - \bar{x}^{(k+1)}\|_1 \leqslant \mu' \|\bar{x} - \bar{x}^{(k)}\|_1. \tag{16}$$

Но

$$\sum_{\substack{j=1 \\ (i \neq j)}}^{n} |c_{ij}| = \beta_i + \gamma_i \leqslant \mu < 1 \tag{17}$$

и

$$\beta_i + \gamma_i - \frac{\gamma_i}{1-\beta_i} = \frac{\beta_i(1-\beta_i-\gamma_i)}{1-\beta_i} \geqslant 0. \tag{18}$$

Отсюда

$$\mu = \max_i (\beta_i + \gamma_i) \geqslant \max_i \frac{\gamma_i}{1-\beta_i} = \mu', \tag{19}$$

что и требовалось доказать.

Теорема. *Если матрица A симметрическая и положительно определенная, а приведение системы $A\bar{x} = \bar{b}$ к виду $\bar{x} = C\bar{x} + \bar{d}$ осуществляется путем деления уравнений на диагональные элементы и последующего перенесения всех членов кроме x_i, где i — номер уравнения, направо, то метод Зейделя сходится.*

Проверим, что при выполнении наших условий все собственные значения матрицы

$$-(B+C)^{-1}D=-(B+C)^{-1}C' \tag{20}$$

по модулю меньше единицы. Пусть λ_i и λ_j — два каких-то собственных значения этой матрицы, а $\bar{z}_i$ и $\bar{z}_j$ — соответствующие им собственные векторы. Тогда мы можем записать:

$$\left.\begin{aligned}-(B+C)^{-1}C'\bar{z}_i&=\lambda_i\bar{z}_i,\\ -(B+C)^{-1}C'\bar{z}_j&=\lambda_j\bar{z}_j.\end{aligned}\right\} \tag{21}$$

Отсюда

$$\left.\begin{aligned}C'\bar{z}_i&=-\lambda_iB\bar{z}_i-\lambda_iC\bar{z}_i,\\ C'\bar{z}_j&=-\lambda_jB\bar{z}_j-\lambda_jC\bar{z}_j.\end{aligned}\right\} \tag{22}$$

Рассмотрим скалярные произведения $(C'\bar{z}_i,\ \bar{z}_j)$ и $(C'\bar{z}_j,\ \bar{z}_i)$. Они будут представляться в виде

$$\left.\begin{aligned}(C'\bar{z}_i,\ \bar{z}_j)&=-\lambda_i(B\bar{z}_i,\ \bar{z}_j)-\lambda_i(C\bar{z}_i,\ \bar{z}_j),\\ (C'\bar{z}_j,\ \bar{z}_i)&=-\lambda_j(B\bar{z}_j,\ \bar{z}_i)-\lambda_j(C\bar{z}_j,\ \bar{z}_i).\end{aligned}\right\} \tag{23}$$

Используя свойства скалярного произведения, получим:

$$\left.\begin{aligned}(C'\bar{z}_j,\ \bar{z}_i)&=(\bar{z}_j,\ C\bar{z}_i)=\overline{(C\bar{z}_i,\ \bar{z}_j)},\\ (C\bar{z}_j,\ \bar{z}_i)&=(\bar{z}_j,\ C'\bar{z}_i)=\overline{(C'\bar{z}_i,\ \bar{z}_j)},\\ (B\bar{z}_j,\ \bar{z}_i)&=(\bar{z}_j,\ B\bar{z}_i)=\overline{(B\bar{z}_i,\ \bar{z}_j)}.\end{aligned}\right\} \tag{24}$$

Поэтому второе из равенств (23) можно переписать в виде

$$\overline{(C\bar{z}_i,\ \bar{z}_j)}=-\lambda_j\overline{(B\bar{z}_i,\ \bar{z}_j)}-\lambda_j\overline{(C'\bar{z}_i,\ \bar{z}_j)}, \tag{25}$$

или

$$(C\bar{z}_i,\ \bar{z}_j)=-\bar{\lambda}_j(B\bar{z}_i,\ \bar{z}_j)-\bar{\lambda}_j(C'\bar{z}_i,\ \bar{z}_j). \tag{26}$$

Решая (26) и первое из равенств (23) относительно $(C\bar{z}_i,\ \bar{z}_j)$ и $(C'\bar{z}_i,\ \bar{z}_j)$, получим:

$$\left.\begin{aligned}(C\bar{z}_i,\ \bar{z}_j)&=\frac{\bar{\lambda}_j(\lambda_i-1)}{1-\lambda_i\bar{\lambda}_j}(B\bar{z}_i,\ \bar{z}_j),\\ (C'\bar{z}_i,\ \bar{z}_j)&=\frac{\lambda_i(\bar{\lambda}_j-1)}{1-\lambda_i\bar{\lambda}_j}(B\bar{z}_i,\ \bar{z}_j).\end{aligned}\right\} \tag{27}$$

Тогда

$$(A\bar{z}_i,\ \bar{z}_j)=(B\bar{z}_i,\ \bar{z}_j)+(C\bar{z}_i,\ \bar{z}_j)+(C'\bar{z}_i,\ \bar{z}_j)=$$

$$=\left[1+\frac{\bar{\lambda}_j(\lambda_i-1)}{1-\lambda_i\bar{\lambda}_j}+\frac{\lambda_i(\bar{\lambda}_j-1)}{1-\lambda_i\bar{\lambda}_j}\right](B\bar{z}_i,\ \bar{z}_j)=\frac{(1-\lambda_i)(1-\bar{\lambda}_j)}{1-\lambda_i\bar{\lambda}_j}(B\bar{z}_i,\ \bar{z}_j). \tag{28}$$

Положив здесь $i=j$, найдем:

$$(A\bar{z}_i,\ \bar{z}_i)=\frac{|1-\lambda_i|^2}{1-|\lambda_i|^2}(B\bar{z}_i,\ \bar{z}_i). \tag{29}$$

Отсюда

$$1-|\lambda_i|^2=|1-\lambda_i|^2\frac{(B\bar{z}_i,\ \bar{z}_i)}{(A\bar{z}_i,\ \bar{z}_i)}. \tag{30}$$

Так как A положительно определенная, то все ее диагональные элементы положительны. Поэтому

$$(A\bar{z}_i,\ \bar{z}_i)>0,\quad (B\bar{z}_i,\ \bar{z}_i)>0. \tag{31}$$

Заметим, что $\lambda=1$ не является собственным значением матрицы $-(B+C)^{-1}C'$. Действительно, если бы существовал такой вектор $\bar{z}$, что

$$-(B+C)^{-1}C'\bar{z}=\bar{z}, \tag{32}$$

то мы имели бы

$$-C'\bar{z}=(B+C)\bar{z} \tag{33}$$

или

$$(B+C+C')\bar{z}=A\bar{z}=0, \tag{34}$$

а это равенство возможно только при $z=0$, так как существует обратная матрица A^{-1}.

Итак, из (30) следует, что

$$1-|\lambda_i|^2>0, \tag{35}$$

и утверждение доказано.

Можно доказать также, что если матрица A симметрична и ее диагональные элементы положительны, то положительная определенность матрицы A необходима для сходимости метода Зейделя.

3. Релаксационный метод [1]). Метод Зейделя является разновидностью метода наименьших квадратов. При этом в качестве векторов $\bar{c}_k$, о которых говорилось в § 8, берутся в циклическом порядке единичные векторы, направленные по координатным осям.

---

[1]) См. обзорную статью М. В. Николаевой «О релаксационном методе Саусвелла», Труды математического института им. Стеклова, т. XXVIII, 1949 г.

Иногда бывает целесообразно для упрощения вычислений или для улучшения сходимости изменить порядок уравнений в заданной системе или же нумерацию неизвестных. Можно пойти и еще дальше, а именно при каждом цикле процесса последовательных *приближений* брать свой порядок. Так, например, поступают в *релаксационном методе*. Выбирают начальное приближение $\left(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}\right)$. Вычисляют так называемые невязки

$$\delta_i = a_{i1}x_1^{(0)} + a_{i2}x_2^{(0)} + \ldots + a_{in}x_n^{(0)} - b_i. \tag{36}$$

Находится $x_1^{(1)}$, удовлетворяющее равенству

$$a_{i1}x_1^{(1)} + a_{i2}x_2^{(0)} + \ldots + a_{in}x_n^{(0)} = b_i, \tag{37}$$

где i — номер уравнения с максимальной по модулю невязкой. Затем подсчитываем невязки

$$\delta_j = a_{j1}x_1^{(1)} + a_{j2}x_2^{(0)} + \ldots + a_{jn}x_n^{(0)} - b_j \quad (j \neq i) \tag{38}$$

и подбираем $x_2^{(1)}$, удовлетворяющее равенству

$$a_{j1}x_1^{(1)} + a_{j2}x_2^{(1)} + a_{j3}x_3^{(0)} + \ldots + a_{jn}x_n^{(0)} = b_j, \tag{39}$$

где j — номер уравнения с наибольшей по модулю невязкой. Так продолжаем и дальше, пока не используем все n уравнений. При этом будут найдены все $x_i^{(1)}$. Тогда начинаем второй цикл, который производится так же, как и первый, но вместо $\left(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}\right)$ используется $\left(x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}\right)$. Повторение циклов продолжают до тех пор, пока не достигнут требуемой точности. Иногда при выборе уравнения, из которого вычисляется «улучшенное» приближение, руководствуются не принципом максимальной по модулю невязки, а каким-либо другим. Во всех случаях стараются брать уравнения в таком порядке, чтобы в кратчайший срок получить нужное решение. Этот довольно-таки неопределенный принцип требует от вычислителя навыка и искусства. Поэтому релаксационный метод трудно осуществить на машинах.

Релаксационный метод является нестационарным.

§ 11. Метод скорейшего спуска

В качестве примера нелинейных методов рассмотрим в этом параграфе метод скорейшего спуска, о котором уже говорилось в § 4. В методе скорейшего спуска исходят из некоторого начального приближения $\bar{x}^{(0)} = \left(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}\right)$ и по нему находят следующее приближение $\bar{x}^{(1)} = \left(x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(n)}\right)$ по формуле

$$\bar{x}^{(1)} = \bar{x}^{(0)} + \alpha_0 \bar{r}^{(0)}, \tag{1}$$

где

$$\bar{r}^{(0)} = \left(r_1^{(0)}, r_2^{(0)}, \ldots, r_n^{(0)}\right), \tag{2}$$

$$r_i^{(0)} = b_i - \sum_{j=1}^{n} a_{ij} x_j^{(0)}, \tag{3}$$

$$\alpha_0 = \frac{\left(\sum_{j=1}^{n} r_j^{(0)\,2}\right)}{\left(\sum_{i,\,j=1}^{n} a_{ij} r_i^{(0)} r_j^{(0)}\right)}. \tag{4}$$

Если матрица A не является симметрической и положительно определенной, как это предполагалось при выводе формул (3) и (4), то можно рассмотреть систему $A'A\bar{x} = A'\bar{b}$, матрица которой симметрична и положительно определена. В этом случае вместо формул (3) и (4) мы получим:

$$r_i^{(0)} = -\sum_{j,\,k=1}^{n} a_{ij} a_{jk} x_k^{(0)} + \sum_{j=1}^{n} a_{ij} b_j, \tag{5}$$

$$\alpha_0 = \frac{\left(\sum_{j=1}^{n} r_j^{(0)\,2}\right)}{\sum_{i=1}^{n}\left(\sum_{j=1}^{n} a_{ij} r_j^{(0)}\right)^2}. \tag{6}$$

После того как будет найден вектор $\bar{x}^{(1)}$, по нему находят вектор $\bar{x}^{(2)}$ так же, как $\bar{x}^{(1)}$ находилось по $\bar{x}^{(0)}$. Процесс продолжают до тех пор, пока не достигнут требуемой точности.

Мы не будем входить в подробности вычислительных схем для метода скорейшего спуска, а рассмотрим вопросы сходимости. Ограничимся случаем симметрической положительно определенной матрицы A. Обозначим собственные значения матрицы A через

$$\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n > 0 \tag{7}$$

и соответствующие им ортонормированные собственные векторы через $\bar{z}_1, \bar{z}_2, \ldots, \bar{z}_n$. Тогда, если $\bar{x}$ — произвольный вектор

$$\bar{x} = \beta_1 \bar{z}_1 + \beta_2 \bar{z}_2 + \ldots + \beta_n \bar{z}_n, \tag{8}$$

то будем иметь:

$$(A\bar{x},\ \bar{x}) = \beta_1^2 \lambda_1 + \beta_2^2 \lambda_2 + \ldots + \beta_n^2 \lambda_n. \tag{9}$$

Отсюда

$$\lambda_n\left(\beta_1^2 + \beta_2^2 + \ldots + \beta_n^2\right) = \lambda_n(\bar{x},\ \bar{x}) \leqslant (A\bar{x},\ \bar{x}) \leqslant$$
$$\leqslant \lambda_1\left(\beta_1^2 + \beta_2^2 + \ldots + \beta_n^2\right) = \lambda_1(\bar{x},\ \bar{x}). \tag{10}$$

Следовательно, для нашей матрицы всегда можно найти две такие постоянные $m > 0$ и $M > 0$, что

$$m(\bar{x}, \bar{x}) \leqslant (A\bar{x}, \bar{x}) \leqslant M(\bar{x}, \bar{x}). \tag{11}$$

Рассмотрим разность $f(\bar{x}^{(1)}) - f(\bar{x}^{(0)})$, где $f(\bar{x})$ — функция, определенная формулой (1) § 5. Простые вычисления дают

$$f(\bar{x}^{(1)}) - f(\bar{x}^{(0)}) = \alpha_0^2(\bar{r}^{(0)}, A\bar{r}^{(0)}) - 2\alpha_0(\bar{r}^{(0)}, \bar{r}^{(0)}) = \frac{-(\bar{r}^{(0)}, \bar{r}^{(0)})^2}{(\bar{r}^{(0)}, A\bar{r}^{(0)})}. \tag{12}$$

Для разности $f(\bar{x}^{(0)}) - f(\bar{x}^*)$, где $\bar{x}^*$ — точное решение системы, получим:

$$f(\bar{x}^{(0)}) - f(\bar{x}^*) = (\bar{x}^{(0)} - \bar{x}^*, A(\bar{x}^{(0)} - \bar{x}^*)) = (A^{-1}\bar{r}^{(0)}, \bar{r}^{(0)}). \tag{13}$$

Таким образом,

$$\frac{f(\bar{x}^{(1)}) - f(\bar{x}^{(0)})}{f(\bar{x}^{(0)}) - f(\bar{x}^*)} = -\frac{(\bar{r}^{(0)}, \bar{r}^{(0)})^2}{(\bar{r}^{(0)}, A\bar{r}^{(0)})(A^{-1}\bar{r}^{(0)}, \bar{r}^{(0)})}. \tag{14}$$

Разложим вектор $\bar{r}^{(0)}$ по собственным векторам матрицы A:

$$\bar{r}^{(0)} = \gamma_1\bar{z}_1 + \gamma_2\bar{z}_2 + \ldots + \gamma_n\bar{z}_n. \tag{15}$$

Тогда

$$A\bar{r}^{(0)} = \gamma_1\lambda_1\bar{z}_1 + \gamma_2\lambda_2\bar{z}_2 + \ldots + \gamma_n\lambda_n\bar{z}_n, \tag{16}$$

$$A^{-1}\bar{r}^{(0)} = \gamma_1\lambda_1^{-1}\bar{z}_1 + \gamma_2\lambda_2^{-1}\bar{z}_2 + \ldots + \gamma_n\lambda_n^{-1}\bar{z}_n \tag{17}$$

и

$$(A\bar{r}^{(0)}, \bar{r}^{(0)}) = \gamma_1^2\lambda_1 + \gamma_2^2\lambda_2 + \ldots + \gamma_n^2\lambda_n, \tag{18}$$

$$(A^{-1}\bar{r}^{(0)}, \bar{r}^{(0)}) = \gamma_1^2\lambda_1^{-1} + \gamma_2^2\lambda_2^{-1} + \ldots + \gamma_n^2\lambda_n^{-1}. \tag{19}$$

Следовательно,

$$\frac{f(\bar{x}^{(0)}) - f(\bar{x}^{(1)})}{f(\bar{x}^{(0)}) - f(\bar{x}^*)} = \frac{(\gamma_1^2 + \gamma_2^2 + \ldots + \gamma_n^2)^2}{(\gamma_1^2\lambda_1 + \gamma_2^2\lambda_2 + \ldots + \gamma_n^2\lambda_n)(\gamma_1^2\lambda_1^{-1} + \gamma_2^2\lambda_2^{-1} + \ldots + \gamma_n^2\lambda_n^{-1})}. \tag{20}$$

Деля каждую из скобок в знаменателе на скобку, стоящую в числителе, и обозначая

$$\frac{\gamma_i^2}{\gamma_1^2 + \gamma_2^2 + \ldots + \gamma_n^2} = \delta_i; \quad \delta_i \geqslant 0; \quad \sum_{i=1}^{n} \delta_i = 1, \tag{21}$$

получим:

$$\frac{f(\bar{x}^{(0)}) - f(\bar{x}^{(1)})}{f(\bar{x}^{(0)}) - f(\bar{x}^*)} = \frac{1}{(\delta_1\lambda_1 + \delta_2\lambda_2 + \ldots + \delta_n\lambda_n)(\delta_1\lambda_1^{-1} + \delta_2\lambda_2^{-1} + \ldots + \delta_n\lambda_n^{-1})}. \tag{22}$$

Докажем следующее неравенство: если $0 < m < \lambda_i \leqslant M$ $(i = 1, 2, \ldots, n)$, то для любых действительных чисел $\delta_1, \delta_2, \ldots, \delta_n$; $\delta_i \geqslant 0$; $\sum_{i=1}^{n} \delta_i = 1$ и $\lambda_i > 0$ имеет место неравенство

$$\sum_{i=1}^{n} \delta_i \lambda_i \sum_{i=1}^{n} \delta_i \lambda_i^{-1} \leqslant \frac{1}{4}\left[\sqrt{\frac{M}{m}} + \sqrt{\frac{m}{M}}\right]^2. \tag{23}$$

Введем вместо λ_i новые числа λ_i', определяемые при помощи равенства

$$\lambda_i = \sqrt{mM}\,\lambda_i'. \tag{24}$$

При этом

$$\sqrt{\frac{m}{M}} \leqslant \lambda_i' \leqslant \sqrt{\frac{M}{m}}, \tag{25}$$

а левая часть доказываемого неравенства примет такую же форму, как и ранее:

$$\sum_{i=1}^{n} \delta_i \lambda_i \sum_{i=1}^{n} \delta_i \lambda_i^{-1} = \sum_{i=1}^{n} \delta_i \lambda_i' \sum_{i=1}^{n} \delta_i \lambda_i'^{-1}. \tag{26}$$

Применим к последнему выражению теорему о том, что среднее геометрическое меньше или равно среднему арифметическому:

$$\sum_{i=1}^{n} \delta_i \lambda_i' \sum_{i=1}^{n} \delta_i \lambda_i'^{-1} \leqslant \frac{1}{4}\left\{\sum_{i=1}^{n} \delta_i \left(\lambda_i' + \frac{1}{\lambda_i'}\right)\right\}^2. \tag{27}$$

Функция

$$\varphi(\lambda') = \lambda' + \frac{1}{\lambda'} \tag{28}$$

при $\lambda' > 0$ убывает в интервале $(0,1)$, возрастает в интервале $(1, \infty)$, имеет наименьшее значение, равное 2, при $\lambda' = 1$ и принимает наибольшее значение на отрезке $\left[\sqrt{\frac{m}{M}}, \sqrt{\frac{M}{m}}\right]$ при $\lambda' = \sqrt{\frac{m}{M}}$ и $\lambda' = \sqrt{\frac{M}{m}}$, равное

$$\sqrt{\frac{m}{M}} + \sqrt{\frac{M}{m}}. \tag{29}$$

Заменяя в (27) каждое из выражений $\lambda_i' + \frac{1}{\lambda_i'}$ его наибольшим значением (29), получим:

$$\sum_{i=1}^{n} \delta_i \lambda_i \sum_{i=1}^{n} \delta_i \lambda_i^{-1} \leqslant \frac{1}{4}\left\{\left(\sqrt{\frac{m}{M}} + \sqrt{\frac{M}{m}}\right)\sum_{i=1}^{n} \delta_i\right\}^2 = \frac{1}{4}\left[\sqrt{\frac{m}{M}} + \sqrt{\frac{M}{m}}\right]^2, \tag{30}$$

что и требовалось доказать.

Применяя доказанное неравенство к (22), найдем:

$$\frac{f(\bar{x}^{(0)})-f(\bar{x}^{(1)})}{f(\bar{x}^{(0)})-f(\bar{x}^{*})}\geqslant\frac{4}{\left[\sqrt{\frac{m}{M}}+\sqrt{\frac{M}{m}}\right]^2}=q, \tag{31}$$

причем $0<q<1$. Отсюда

$$f(\bar{x}^{(1)})-f(\bar{x}^{*})\leqslant(1-q)\,[f(\bar{x}^{(0)})-f(\bar{x}^{*})]=(1-q)\,c, \tag{32}$$

где через c обозначено выражение $[f(\bar{x}^{(0)})-f(\bar{x}^{*})]$. Таким образом, для любого k получаем:

$$f(\bar{x}^{(k)})-f(\bar{x}^{*})\leqslant c\,(1-q)^k. \tag{33}$$

Оценим теперь $\|\bar{x}^{(k)}-\bar{x}^{*}\|_3$. Имеем:

$$\|\bar{x}^{(k)}-\bar{x}^{*}\|_3^2=(\bar{x}^{(k)}-\bar{x}^{*},\ \bar{x}^{(k)}-\bar{x}^{*})\leqslant\frac{1}{m}(A\bar{x}^{(k)}-\bar{b},\ \bar{x}^{(k)}-\bar{x}^{*}). \tag{34}$$

Легко проверить, что

$$(A\bar{x}^{(k)}-\bar{b},\ \bar{x}^{(k)}-\bar{x}^{*})=f(\bar{x}^{(k)})-f(\bar{x}^{*}). \tag{35}$$

Таким образом,

$$\|\bar{x}^{(k)}-\bar{x}^{*}\|_3^2\leqslant\frac{1}{m}[f(\bar{x}^{(k)})-f(\bar{x}^{*})]\leqslant\frac{c}{m}(1-q)^k=\frac{c}{m}\left(\frac{M-m}{M+m}\right)^{2k}, \tag{36}$$

и мы доказали, что *метод скорейшего спуска сходится со скоростью геометрической прогрессии к точному решению.*

Можно усовершенствовать метод скорейшего спуска, осуществляя одновременно несколько шагов. Благодаря этому удается увеличить скорость сходимости.

Одновременное осуществление p шагов по методу скорейшего спуска эквивалентно вычислению $\bar{x}^{(k+1)}$ по формуле

$$\bar{x}^{(k+1)}=\bar{x}^{(k)}+\sum_{i=0}^{p-1}\alpha_i^{(k)}A^i\bar{r}^{(k)}, \tag{37}$$

где, как и ранее, $\bar{r}^{(k)}=\bar{b}-A\bar{x}^{(k)}$, а коэффициенты $\alpha_i^{(k)}$ подбираются так, чтобы $f(\bar{x}^{(k+1)})$ приняла наименьшее возможное значение. Это требование к $\alpha_i^{(k)}$ приводит к системе уравнений

$$\frac{\partial}{\partial\alpha_i^{(k)}}f\left(\bar{x}^{(k)}+\sum_{i=0}^{p-1}\alpha_i^{(k)}A^i\bar{r}^{(k)}\right)=0\qquad(l=0,\ 1,\ \ldots,\ p-1) \tag{38}$$

или

$$\sum_{j=0}^{p-1}\alpha_j^{(k)}(A^i\bar{r}^{(k)},\ A^{j+1}\bar{r}^{(k)})=(A^i\bar{r}^{(k)},\ \bar{r}^{(k)})\qquad(l=0,\ 1,\ 2,\ \ldots,\ p-1). \tag{39}$$

Исследуем сходимость этого метода. Подберем постоянные ε_i так, чтобы многочлен

$$\varphi(\lambda) = 1 - \sum_{i=0}^{p-1} \varepsilon_i \lambda^{i+1} \tag{40}$$

имел наименьшее возможное значение максимума модуля на отрезке $[m, M]$. При $\lambda = 0$ наш многочлен будет равен 1. Такое построение мы уже осуществили в § 8 и видели там, что $\varphi(\lambda)$ просто выражается через многочлены Чебышева, наименее уклоняющиеся от нуля. При этом $\sup_{\lambda \in [m, M]} |\varphi(\lambda)| = L < 1$.

Следовательно, собственные значения матрицы

$$B = I - \sum_{i=0}^{p-1} \varepsilon_i A^{i+1}, \tag{41}$$

которые равны $\varphi(\lambda_j)$ (λ_j — собственные значения матрицы A) также по модулю меньше единицы. Поэтому мы можем решать систему

$$\bar{y} = \bar{y} + \sum_{i=0}^{p-1} \varepsilon_i A^i (\bar{b} - A\bar{y}) \tag{42}$$

относительно $\bar{y}$ методом простой итерации:

$$\bar{y}^{(k+1)} = \bar{y}^{(k)} + \sum_{i=0}^{p-1} \varepsilon_i A^i (\bar{b} - A\bar{y}^{(k)}). \tag{43}$$

Так как простая итерация сходится, то система (42) имеет единственное решение. Таким решением может быть только $\bar{x}^* = A^{-1}\bar{b}$. Обозначим

$$\bar{u}^{(k)} = \bar{y}^{(k)} - \bar{x}^*. \tag{44}$$

В силу (43) имеем:

$$\bar{u}^{(k+1)} = B\bar{u}^{(k)}. \tag{45}$$

Таким образом,

$$(A\bar{u}^{(k+1)}, \bar{u}^{(k+1)}) = (AB\bar{u}^{(k)}, B\bar{u}^{(k)}) = (AB^2\bar{u}^{(k)}, \bar{u}^{(k)}). \tag{46}$$

Разложим вектор $\bar{u}^{(k)}$ по собственным ортонормированным векторам матрицы A:

$$\bar{u}^{(k)} = \gamma_1 \bar{z}_1 + \gamma_2 \bar{z}_2 + \ldots + \gamma_n \bar{z}_n. \tag{47}$$

Тогда

$$(AB^2\bar{u}^{(k)}, \bar{u}^{(k)}) = \left(\sum_{i=1}^{n} \gamma_i \lambda_i \varphi^2(\lambda_i) \bar{z}_i, \sum_{i=1}^{n} \gamma_i \bar{z}_i \right) =$$

$$= \sum_{i=1}^{n} \gamma_i^2 \varphi^2(\lambda_i) \lambda_i \leqslant \sup_{m \leqslant \lambda \leqslant M} \varphi^2(\lambda) \sum_{i=1}^{n} \gamma_i^2 \lambda_i = L^2 (A\bar{u}^{(k)}, \bar{u}^{(k)}). \tag{48}$$

Итак,

$$(A\bar{u}^{(k+1)}, \bar{u}^{(k+1)}) \leqslant L^2 (A\bar{u}^{(k)}, \bar{u}^{(k)}). \tag{49}$$

Возьмем теперь в качестве $\bar{y}^{(k)}$ вектор $\bar{x}^{(k)}$, полученный по методу скорейшего спуска. Обозначим

$$\bar{x}^{(i)} - \bar{x} = \bar{v}^{(i)} \qquad (i = 0, 1, 2, \ldots). \tag{50}$$

Так как

$$(A\bar{u}^{(k+1)},\ \bar{u}^{(k+1)}) = f(\bar{y}^{(k+1)}) - f(\bar{x}^{*}), \tag{51}$$

$$(A\bar{v}^{(k+1)},\ \bar{v}^{(k+1)}) = f(\bar{x}^{(k+1)}) - f(\bar{x}^{*}) \tag{52}$$

и $f(\bar{x}^{(k+1)}) \leqslant f(\bar{y}^{(k+1)})$, то

$$(A\bar{v}^{(k+1)},\ \bar{v}^{(k+1)}) \leqslant L^2 (A\bar{v}^{(k)},\ \bar{v}^{(k)}). \tag{53}$$

Отсюда, как и ранее, получаем:

$$f(\bar{x}^{(k)}) - f(\bar{x}^{*}) \leqslant L^{2k} [f(\bar{x}^{(0)}) - f(\bar{x}^{*})]. \tag{54}$$

Постоянная L определится в силу равенства (34) в § 9 выражением

$$L = \frac{2}{\left[\frac{M+m}{M-m} - \sqrt{\left(\frac{M+m}{M-m}\right)^2 - 1}\right]^p + \left[\frac{M+m}{M-m} + \sqrt{\left(\frac{M+m}{M-m}\right)^2 - 1}\right]^p}. \tag{55}$$

На этом мы пока прервем изучение численных методов решения систем линейных алгебраических уравнений. Вопросы, связанные с погрешностями методов и ускорением сходимости удобнее рассматривать после того, как будут известны способы отыскания собственных значений матриц. Поэтому мы вернемся к ним в главе 8.

УПРАЖНЕНИЯ

1. Решить всеми изложенными методами систему линейных алгебраических уравнений:

$$2{,}1546x_1 + 0{,}8431x_2 + 0{,}3146x_3 + 0{,}1615x_4 = 3{,}1826,$$

$$2{,}8431x_1 + 3{,}1415x_2 + 0{,}6241x_3 + 0{,}2131x_4 = 4{,}6123,$$

$$0{,}3146x_1 + 0{,}6241x_2 + 4{,}8216x_3 + 0{,}8245x_4 = 5{,}9681,$$

$$0{,}1615x_1 + 0{,}2131x_2 + 0{,}8245x_3 + 6{,}4131x_4 = 8{,}1418.$$

2. Показать, что если матрица A симметрическая и положительно определенная, то метод исключения Гаусса эквивалентен последовательной ортогонализации единичных координатных векторов в метрике A.

3. Показать, что компактная схема Гаусса применима в том случае, если все определители

$$|a_{11}|,\quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix},\quad \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix},\ \ldots$$

отличны от нуля.

4. Показать, что если в методе сопряженных градиентов для положительно определенной матрицы A все векторы $\bar{r}^{(0)}, \bar{r}^{(1)}, \ldots, \bar{r}^{(n-1)}$ отличны от нуля, то определитель матрицы A равен

$$\det A = (\alpha_0 \alpha_1 \ldots \alpha_{n-1})^{-1}.$$

5. Пусть для решения системы $A\bar{x} = \bar{b}$ используется итерационная формула

$$\bar{x}^{(k+1)} = (I + DA)\,\bar{x}^{(k)} - D\bar{b},$$

где D — некоторая неособенная матрица. Доказать, что метод сходится для любого начального вектора $\bar{x}^{(0)}$, если все характеристические корни матрицы DA расположены внутри круга с центром в точке $\lambda = -1$ и радиуса 1.

6. Доказать, что если DA предыдущей задачи есть cA, где c — отрицательная постоянная, удовлетворяющая одному из условий

$$|c| < \frac{8}{\left[\max\limits_i \sum\limits_{k=1}^{n} |a_{ik} + a_{ki}| + \max\limits_i \sum\limits_{k=1}^{n} |a_{ki} - a_{ik}|\right]^2},$$

или

$$|c| < \frac{2}{\left[\max\limits_i |a_{ij}| + \left(\sum\limits_{\substack{i,\,k=1 \\ (i \neq k)}}^{n} a_{ik}^2\right)^{\frac{1}{2}}\right]^2},$$

то метод последовательных приближений сходится.

7. Доказать, что метод последовательных приближений, приведенный в упражнении 5, сходится для всех симметрических положительно или отрицательно определенных матриц A, если взять $D = cI$, где

$$|c| < \frac{2}{\max\limits_i \sum\limits_{k=1}^{n} |a_{ik}|},$$

и $c < 0$ для положительно определенных A и $c > 0$ для отрицательно определенных A.

8. Пусть для решения системы $A\bar{x} = \bar{b}$ используется итерационная формула

$$\bar{x}^{(k+1)} = B^{-1}[\bar{b} - A(\bar{x}^{(0)} + \bar{x}^{(1)} + \ldots + \bar{x}^{(k)})],$$

где B — некоторая неособенная матрица. Доказать, что этот способ сходится и дает решение исходной системы, если максимальное собственное значение матрицы $C'C$, где $C = AB$ меньше, чем минимальное собственное значение матрицы $B'B$.

ЛИТЕРАТУРА

1. В. Н. Фаддеева, Вычислительные методы линейной алгебры, Гостехиздат, 1950.
2. Л. В. Канторович, Функциональный анализ и прикладная математика, УМН, т. 3, вып. 6, 1948.
3. Л. Фокс, Х. Д. Хаски, Дж. Х. Вилкинсон, Заметки о решении систем совместных линейных уравнений, УМН, т. 5, вып, 3, 1950.
4. М. Ш. Бирман, Некоторые оценки для метода наискорейшего спуска, УМН, т. 5, вып. 3, 1950.
5. М. Р. Шура-Бура, Оценка ошибок при численном обращении матриц высокого порядка, УМН, т. 6, вып. 4, 1951.

6. А. Тюринг, Ошибки округления в матричных процессах, УМН, т. 6, вып. 1, 1951.
7. М. А. Красносельский, С. Г. Крейн, Замечание о распределении ошибок при решении систем линейных уравнений при помощи итерационного процесса, УМН, т. 7, вып. 4, 1952.
8. Л. В. Канторович, Приближенное решение функциональных уравнений, УМН, т. 11, вып. 6, 1956.
9. Кертисс, Методы «Монте-Карло» для итерации линейных операторов, УМН, т. 12, вып. 5, 1957.
10. G. Forsythe, Решение линейных алгебраических уравнений может быть интересным, Bull. amer. Math, Soc., т. 59, № 64, 1953. (Приведена обширная библиография различных методов решения систем линейных алгебраических уравнений.)
11. Хаусхолдер, Основы численного анализа, ИЛ, 1956.
12. Милн, Численный анализ, ИЛ, 1951.

ГЛАВА 7

ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ ВЫСШИХ СТЕПЕНЕЙ И ТРАНСЦЕНДЕНТНЫХ УРАВНЕНИЙ

§ 1. Введение

В этой главе мы рассмотрим некоторые методы численного решения уравнений вида

$$f(z) = 0,$$

где $f(z)$ — заданная функция действительного или комплексного аргумента z; в частности, $f(z)$ может быть многочленом степени n от z.

При отыскании приближенных значений корней этого уравнения приходится решать две задачи:

1) *отделение корней*, т. е. отыскание достаточно малых областей, в каждой из которых заключен один и только один корень уравнения;

2) *вычисление корней с заданной точностью*.

Так как для алгебраических уравнений разработано больше методов отделения корней и методов их вычисления, то мы в дальнейшем будем более подробно останавливаться на алгебраических уравнениях.

§ 2. Отделение корней

1. Общие замечания. При решении уравнения $f(z) = 0$ прежде всего важно предварительно изучить расположение корней в комплексной плоскости z и заключить каждый корень в достаточно малую область, внутри которой не было бы других корней. Для этой цели иногда выгодно применять графические методы.

Если требуется найти только действительные корни уравнения, то для отыскания грубых значений корней можно построить график функции $y = f(x)$ и найти абсциссы точек пересечения графика с осью x (рис. 2).

Иногда удобней представить сначала уравнение в виде

$$\varphi(x) = \psi(x)$$

и затем, построив графики функций $y=\varphi(x)$ и $y=\psi(x)$, найти абсциссы их точек пересечения, которые и будут приближенными значениями корней. Например, если нужно найти корни уравнения

$$x\sin x=1,$$

то удобно представить его в виде

$$\sin x=\frac{1}{x}$$

и применить указанный способ (рис. 3).

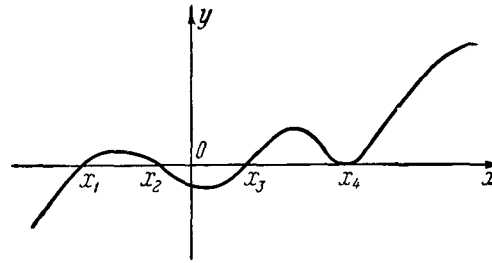

Рис. 2.

Корни уравнения симметричны относительно $x=0$. Поэтому мы можем рассматривать только положительные корни. Значения x_1, x_2 и, возможно, еще нескольких корней можно довольно точно определить графически. Однако на графике не будет x_n для больших значений n. Тем не менее по ходу графиков мы можем сказать, что значения x_n при больших n будут близки к πn. Эти значения можно

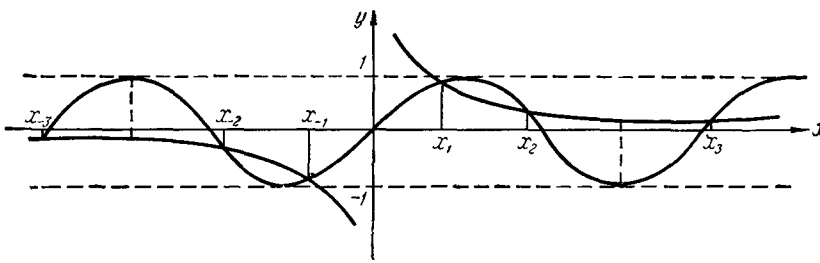

Рис. 3.

уточнить. Положим $x_n=\pi n+\varepsilon_n$, где ε_n — некоторые небольшие добавки. Тогда

$$\frac{1}{\pi n+\varepsilon_n}-\sin(\pi n+\varepsilon_n)=\frac{1}{\pi n+\varepsilon_n}-(-1)^n\sin\varepsilon_n=0.$$

В силу малости ε_n можно положить $\sin\varepsilon_n\approx\varepsilon_n$ и $\frac{1}{\pi n+\varepsilon_n}\approx\frac{1}{\pi n}$. Это дает

$$\varepsilon_n\approx\frac{(-1)^n}{\pi n},$$

и мы получили улучшенные значения корней

$$x_n\approx\pi n+\frac{(-1)^n}{\pi n}.$$

Если требуется, то можно продолжить уточнение корней, положив

$$x_n = \pi n + \frac{(-1)^n}{\pi n} + \varepsilon'_n.$$

Это дает уравнение для определения ε'_n:

$$\frac{1}{\pi n + \frac{(-1)^n}{\pi n} + \varepsilon'_n} - (-1)^n \sin\left[\frac{(-1)^n}{\pi n} + \varepsilon'_n\right] = 0.$$

Но

$$\frac{1}{\pi n + \frac{(-1)^n}{\pi n} + \varepsilon'_n} \approx \frac{1}{\pi n} - \frac{(-1)^n}{\pi^3 n^3},$$

$$\sin\left[\frac{(-1)^n}{\pi n} + \varepsilon'_n\right] \approx \varepsilon'_n + \frac{(-1)^n}{\pi n} - \frac{1}{3!}\left[\varepsilon'_n + \frac{(-1)^n}{\pi n}\right]^3 \approx$$

$$\approx \varepsilon'_n + \frac{(-1)^n}{\pi n} - \frac{(-1)^n}{6\pi^3 n^3}.$$

Таким образом,

$$\varepsilon'_n \approx -\frac{1}{\pi^3 n^3}\left[1 - \frac{(-1)^n}{6}\right]$$

и

$$x_n \approx \pi n + \frac{(-1)^n}{\pi n} - \frac{1}{\pi^3 n^3}\left[1 - \frac{(-1)^n}{6}\right].$$

При желании такой процесс можно продолжить.

Для отыскания комплексных корней уравнения $f(z) = 0$ можно, положив $z = x + iy$, представить уравнение в виде

$$\varphi(x, y) + i\psi(x, y) = 0,$$

где $\varphi(x, y)$ и $\psi(x, y)$ — действительные функции действительных переменных x и y. Это уравнение эквивалентно системе двух уравнений:

$$\varphi(x, y) = 0, \quad \psi(x, y) = 0.$$

Построив кривые $\varphi(x, y) = 0$ и $\psi(x, y) = 0$, мы получим действительные и мнимые части корней уравнения $f(z) = 0$ как соответственно абсциссы и ординаты их точек пересечения.

Имеется много специально разработанных способов графического решения уравнений, приспособленных для отдельных типов уравнений. Большое значение имеют номографические методы решения уравнений, но мы не будем останавливаться на этих методах.

Для выделения интервалов, в которых находятся действительные корни уравнения $f(x) = 0$, если $f(x)$ — непрерывная функция, можно воспользоваться следующими предложениями:

Если на концах некоторого отрезка непрерывная функция принимает значения разных знаков, то на этом отрезке уравнение $f(x) = 0$ имеет хотя бы один корень.

Если при этом $f(x)$ имеет первую производную, не меняющую знака, то корень единственный.

Пусть $f(x)$ есть аналитическая функция переменного x на отрезке $[a, b]$; если на концах отрезка $[a, b]$ она принимает значения разных знаков, то между a и b имеется нечетное число корней уравнения $f(x)=0$; если же на концах отрезка $[a, b]$ она принимает значения одинаковых знаков, то между a и b или нет корней этого уравнения, или их имеется четное число (учитывая и кратность корней).

2. Границы расположения корней алгебраического уравнения. Для алгебраического уравнения

$$f(z)=a_0z^n+a_1z^{n-1}+a_2z^{n-2}+\ldots+a_{n-1}z+a_n=0 \qquad (1)$$

задача отделения корней решается более просто и точно. Прежде чем отделять корни уравнения, естественно найти границы области, в которой расположены все корни уравнения, поэтому мы сначала приведем ряд способов отыскания этих границ.

Пусть

$$a=\max\{|a_1|, |a_2|, \ldots, |a_n|\}, \text{ а } a'=\max\{|a_0|, |a_1|, \ldots, |a_{n-1}|\}.$$

Теорема 1. *Все корни уравнения* (1) *расположены в кольце*

$$\frac{|a_n|}{a'+|a_n|}\leqslant|z|\leqslant 1+\frac{a}{|a_0|}. \qquad (2)$$

Доказательство. Действительно,

$$|f(z)|\geqslant||a_0z^n|-|a_1z^{n-1}+\ldots+a_{n-1}z+a_n||,$$

но при $|z|>1$ имеем:

$$|a_1z^{n-1}+\ldots+a_n|\leqslant a\{|z|^{n-1}+|z|^{n-2}+\ldots+|z|+1\}=$$
$$=a\frac{|z|^n-1}{|z|-1}<\frac{a|z|^n}{|z|-1}.$$

Следовательно, $|f(z)|>0$, как только

$$|a_0z^n|-a\frac{|z|^n}{|z|-1}\geqslant 0 \quad \text{или} \quad |a_0||z|-|a_0|-a\geqslant 0,$$

т. е. при $|z|\geqslant 1+\frac{a}{|a_0|}$. Таким образом, все корни уравнения находятся внутри круга радиуса $1+\frac{a}{|a_0|}$.

Далее, уравнение

$$a_0+a_1y+\ldots+a_ny^n=0 \qquad (3)$$

имеет корнями величины, обратные корням исходного уравнения. По доказанному все корни этого уравнения находятся внутри круга

радиуса $1+\frac{a'}{|a_n|}$, т. е. для любого корня z_i исходного уравнения имеет место неравенство

$$\frac{1}{|z_i|}<1+\frac{a'}{|a_n|} \quad \text{или} \quad |z_i|>\frac{|a_n|}{a'+|a_n|}.$$

Объединяя результаты, получим неравенство (2).

Предположим, что все коэффициенты уравнений действительные числа и $a_0>0$. Найдем границы действительных корней уравнения. Очевидно, достаточно иметь способы определения границ положительных корней, так как, заменяя x на $-x$, мы получим уравнение, корни которого отличаются от корней исходного уравнения знаком.

Теорема 2. *Обозначим через a максимум абсолютных величин отрицательных коэффициентов уравнения, и пусть первый отрицательный коэффициент в ряду $a_0, a_1, \ldots, a_n$ есть a_m. Тогда все положительные корни уравнения меньше* $1+\sqrt[m]{\frac{a}{a_0}}$.

(Если отрицательных коэффициентов нет, то нет и положительных корней.)

Доказательство. Заменим положительные коэффициенты $a_1, a_2, \ldots, a_{m-1}$ нулями, а все остальные коэффициенты на $-a$. Тогда при $x>1$ будем иметь:

$$\begin{aligned} f(x)&=a_0x^n+a_1x^{n-1}+\ldots+a_{n-1}x+a_n> \\ &>a_0x^n-a(x^{n-m}+x^{n-m-1}+\ldots+x+1)= \\ &=a_0x^n-a\frac{x^{n-m+1}-1}{x-1}>a_0x^n-a\frac{x^{n-m+1}}{x-1}= \\ &=\frac{x^{n-m+1}}{x-1}\{a_0x^{m-1}(x-1)-a\}. \end{aligned}$$

Отсюда при $x\geqslant 1+\sqrt[m]{\frac{a}{a_0}}$ имеем неравенство $f(x)>0$, так как

$$a_0x^{m-1}(x-1)-a\geqslant a_0\sqrt[m]{\frac{a}{a_0}}\left(1+\sqrt[m]{\frac{a}{a_0}}\right)^{m-1}-a>0,$$

а это и означает, что все положительные корни меньше $1+\sqrt[m]{\frac{a}{a_0}}$.

С помощью теоремы 2 можно найти границы действительных корней очень грубо. Иногда эти границы можно сузить, применив следующий простой прием.

Пусть в уравнении коэффициенты $a_0, a_1, \ldots, a_{m-1}$ неотрицательны, а $a_m, a_{m+1}, \ldots, a_n$ неположительны и $a_m<0$. Введем обозначения:

$$\begin{aligned} a_0x^n+a_1x^{n-1}+\ldots+a_{m-1}x^{n-m+1}&\equiv\varphi(x), \\ a_mx^{n-m}+a_{m+1}x^{n-m-1}+\ldots+a_{n-1}x+a_n&\equiv-\psi(x). \end{aligned}$$

Тогда

$$f(x)=\varphi(x)-\psi(x)=x^{n-m+1}\left\{\frac{\varphi(x)}{x^{n-m+1}}-\frac{\psi(x)}{x^{n-m+1}}\right\}.$$

Первое слагаемое в скобках содержит только положительные степени x, а второе только отрицательные. Следовательно, при $x>0$ первое слагаемое возрастает, а второе убывает с возрастанием x, т. е. при $x>0$ функция $f(x)$ возрастает вместе с x. Найдя какое-либо $x=\alpha>0$, для которого $f(\alpha)>0$, мы можем гарантировать, что все корни уравнения меньше α.

В общем случае представим $f(x)$ в виде

$$f(x)=F(x)+\Phi(x),$$

где $F(x)$ есть многочлен, содержащий все первые старшие по степени члены многочлена $f(x)$, имеющие положительные коэффициенты и все члены с отрицательными коэффициентами, а $\Phi(x)$ — многочлен, образованный всеми остальными членами исходного многочлена $f(x)$. Тогда, если мы найдем $x=\alpha>0$, для которого $F(\alpha)>0$, то $f(x)>0$ при всех $x\geqslant\alpha$, так как $\Phi(x)\geqslant 0$ при $x>0$ и все корни уравнения $f(x)=0$ будут меньше α.

Хороший способ отыскания верхней границы положительных корней указал Ньютон. Этот способ основан на утверждении: *если при* $x=a>0$ *имеют место неравенства*

$$f(a)>0,\quad f'(a)>0,\ \ldots,\ f^{(n)}(a)>0, \tag{4}$$

то уравнение $f(x)=0$ *не имеет корней, больших* a.

Действительно,

$$f(x)=f(a)+(x-a)f'(a)+\ \ldots\ +\frac{f^{(n)}(a)}{n!}(x-a)^n>0$$

при всех $x>a$.

Таким образом, способ Ньютона заключается в отыскании значения $a>0$, при котором многочлен $f(x)$ и все его производные имеют положительное значение. Тогда это значение будет верхней границей положительных корней.

Замечание. Нижняя граница положительных корней может быть найдена из уравнения $a_0+a_1y+\ \ldots\ +a_ny^n=0$ такими же приемами, так как если β есть верхняя граница положительных корней этого уравнения, то $\frac{1}{\beta}$ будет нижней границей положительных корней исходного уравнения.

Пример. Найти границы действительных корней уравнения

$$x^4-35x^3+380x^2-1350x+1000=0.$$

1-й способ (использование теоремы 2). В данном случае

$$a=1350,\quad a_0=1,\quad m=1.$$

Следовательно, все положительные корни уравнения меньше

$$1+\frac{1350}{1}=1351.$$

Для отыскания нижней границы положительных корней уравнения рассмотрим уравнение

$$1-35y+380y^2-1350y^3+1000y^4=0.$$

Так как $a_0'=1000$, $a'=1350$, $m=1$, то верхняя граница положительных корней этого уравнения будет:

$$1+\frac{1000}{1350}=\frac{2350}{1350},$$

а следовательно, нижняя граница корней исходного уравнения

$$\frac{1350}{2350}\approx 0{,}57.$$

Итак, все положительные корни уравнения находятся на отрезке [0,57; 1351].

Для отыскания границ отрицательных корней рассмотрим уравнение

$$z^4+35z^3+380z^2+1350z+1000=0,$$

получающееся заменой x на $-z$. Это уравнение, очевидно, не имеет положительных корней, а следовательно, исходное уравнение не имеет отрицательных корней.

2-й способ. Представим

$$f(x)=x^4-35x^3+380x^2-1350x+1000$$

в виде

$$f(x)=F(x)+\Phi(x),$$

где

$$F(x)=x^4-35x^3-1350x,\quad \Phi(x)=380x^2+1000.$$

При $x=40$ $F(40)=40^4-35\cdot 40^3-1350\cdot 40=266\,000>0$. Поэтому все корни уравнения меньше 40.

Для отыскания нижней границы снова заменим x на $\frac{1}{y}$. Получим:

$$f_1(y)=1000y^4-1350y^3+380y^2-35y+1=F_1(y)+\Phi_1(y),$$

где

$$F_1(y)=1000y^4-1350y^3-35y,\quad \Phi_1(y)=380y^2+1,$$

$F_1(1,5)=353{,}75>0$. Таким образом, положительные корни уравнения $f_1(y)=0$ меньше 1,5, а положительные корни исходного

уравнения больше $\frac{1}{1,5} \approx 0,66$. Следовательно, корни уравнения $f(x) = 0$ расположены на отрезке $[0,66; 40]$. Мы получили значительно лучший результат, чем в первом способе.

Способом Ньютона можно показать, что корни уравнения расположены на отрезке $[0,74; 22]$, т. е. удается еще улучшить результат.

3. Число действительных корней алгебраического уравнения. Оценку числа действительных корней алгебраического уравнения можно получить с помощью известного *правила Декарта*. Это правило мы получим как следствие более общей теоремы, которую назовем *обобщенным правилом Декарта*. Эта теорема позволяет находить числа действительных корней обобщенных многочленов. Прежде чем ее формулировать, введем некоторые определения.

Пусть дана конечная последовательность действительных чисел

$$a_0, a_1, a_2, \ldots, a_n.$$

Будем называть индекс m *местом перемены знака*, если $a_{m-k}a_m < 0$ и $a_{m-1} = a_{m-2} = \ldots = a_{m-k+1} = 0$. В этом случае говорят, что a_{m-k} и a_m образуют *перемену знака*.

Очевидны или легко доказываются следующие утверждения:

1) число перемен знака в последовательности не изменится, если члены, равные нулю, будут опущены, а оставшиеся члены сохранят свое расположение;

2) число перемен знака в последовательности не изменится, если вставить любое число членов, равных нулю, или рядом с членом последовательности вставить новый член того же знака;

3) при вычеркивании членов последовательности число перемен знака не увеличивается;

4) если a_j и a_k $(j < k)$ не равны нулю, то в последовательности $a_j, a_{j+1}, \ldots, a_{k-1}, a_k$ будет четное или нечетное число перемен знака, смотря по тому, будут ли a_j и a_k иметь одинаковые или разные знаки;

5) пусть m — место перемены знака в последовательности

$$a_0, a_1, \ldots, a_m, \ldots, a_n.$$

Тогда число перемен знака в этой последовательности на единицу больше числа перемен знака в последовательности

$$-a_0, -a_1, \ldots, -a_{m-1}, a_{m+1}, \ldots, a_n.$$

Пусть $\varphi_1(x), \varphi_2(x), \ldots, \varphi_n(x)$ — последовательность функций, непрерывных вместе со своими производными до порядка $n-1$

включительно на отрезке $[a, b]$. Рассмотрим определитель Вронского

$$W[\varphi_1, \varphi_2, \ldots, \varphi_n] = \begin{vmatrix} \varphi_1(x) & \varphi_2(x) & \ldots & \varphi_n(x), \\ \varphi_1'(x) & \varphi_2'(x) & \ldots & \varphi_n'(x), \\ \ldots & \ldots & \ldots & \ldots \\ \varphi_1^{(n-1)}(x) & \varphi_2^{(n-1)}(x) & \ldots & \varphi_n^{(n-1)}(x). \end{vmatrix} \quad (5)$$

Имеют место следующие свойства этого определителя:

1) $W[f\varphi_1, f\varphi_2, \ldots, f\varphi_n] = [f(x)]^n W[\varphi_1, \varphi_2, \ldots, \varphi_n]$. (6)

В самом деле, вторая строка определителя $W[f\varphi_1, f\varphi_2, \ldots, f\varphi_n]$ имеет вид

$$f\varphi_1' + f'\varphi_1, \quad f\varphi_2' + f'\varphi_2, \ldots, f\varphi_n' + f'\varphi_n.$$

Если из нее вычесть первую строку, умноженную на f'/f, то она превратится в

$$f\varphi_1', \quad f\varphi_2', \ldots, f\varphi_n'.$$

Третья строка имеет вид

$$f\varphi_1'' + 2f'\varphi_1' + f''\varphi_1, \quad f\varphi_2'' + 2f'\varphi_2' + f''\varphi_2, \ldots, \quad f\varphi_n'' + 2f'\varphi_n' + f''\varphi_n.$$

Если из нее вычесть первую, умноженную на f''/f, и преобразованную вторую, умноженную на $2f'/f$, то получим:

$$f\varphi_1'', \quad f\varphi_2'', \ldots, f\varphi_n''.$$

Продолжая процесс дальше и вынося затем из каждой строки f, получим требуемый результат.

2) $W[\varphi_1(f(x)), \varphi_2(f(x)), \ldots, \varphi_n(f(x))] =$

$$= [f'(x)]^{\frac{n(n-1)}{2}} W(\varphi_1(y), \varphi_2(y), \ldots, \varphi_n(y))_{y=f(x)} \quad (f'(x) \neq 0). \quad (7)$$

В самом деле, вторая строка определителя слева имеет вид

$$\varphi_{1y}' f', \quad \varphi_{2y}' f', \ldots, \quad \varphi_{ny}' f',$$

а третья строка

$$\varphi_{1y^2}'' f'^2 + \varphi_{1y}' f'', \quad \varphi_{2y^2}'' f'^2 + \varphi_{2y}' f'', \quad \ldots, \quad \varphi_{ny^2}'' f'^2 + \varphi_{ny}' f''.$$

Вычтем из нее вторую, умноженную на f''/f'. В результате получим:

$$\varphi_{1y^2}'' f'^2, \quad \varphi_{2y^2}'' f'^2, \ldots, \quad \varphi_{ny^2}'' f'^2.$$

Аналогично после вычитания из четвертой строки второй, умноженной на f'''/f', и преобразованной третьей, умноженной на $3f''/f'$, получим:

$$\varphi_{1y^3}''' f'^3, \quad \varphi_{2y^3}''' y'^3, \ldots, \quad \varphi_{ny^3}''' f'^3.$$

Продолжая упрощения далее и вынося за знак определителя f', $f'^2, \ldots, f'^{n-1}$, получим требуемый результат.

Теорема (*Обобщенное правило Декарта*). *Если на отрезке* $[a, b]$ *функции* $\varphi_1(x), \varphi_2(x), \ldots, \varphi_n(x)$ *непрерывны вместе с производными до порядка* $(n-1)$ *включительно и для любой последовательности* k_i $(1 \leqslant k_1 < k_2 < \ldots < k_m \leqslant n)$

$$W[\varphi_{k_1}, \varphi_{k_2}, \ldots, \varphi_{k_m}] > 0,$$

на $[a, b]$, *то число нулей комбинации*

$$a_1\varphi_1(x) + a_2\varphi_2(x) + \ldots + a_n\varphi_n(x)$$

с действительными коэффициентами a_i (не равными одновременно нулю) *на отрезке* $[a, b]$ *не превышает числа перемен знака в последовательности* $a_1, a_2, \ldots, a_n$.

Доказательство. При $n = 1$ утверждение тривиально, так как по условию теоремы все функции $\varphi_i(x)$ не обращаются на $[a, b]$ в нуль и имеют одинаковый знак. Предположим, что утверждение справедливо при $n \leqslant l-1$, и докажем, что оно останется справедливым при переходе к $n = l$. Пусть

$$\Phi(x) = a_1\varphi_1(x) + a_2\varphi_2(x) + \ldots + a_l\varphi_l(x) \qquad \left(\sum_{i=1}^{l} a_i^2 \neq 0\right)$$

и p — число нулей функции $\Phi(x)$ на $[a, b]$, а q — число перемен знака в последовательности $a_1, a_2, \ldots, a_l$. Если $q = 0$, то, очевидно, и $p = 0$, и утверждение верно. Пусть $q > 0$ и одна из перемен происходит на i-м месте. Введем обозначения:

$$\frac{d}{dx}\frac{\Phi(x)}{\varphi_i(x)} = \Phi^*(x), \quad \frac{d}{dx}\frac{\varphi_j(x)}{\varphi_i(x)} = -\varphi_j^*(x) \qquad (j = 1, 2, \ldots, i-1),$$

$$\frac{d}{dx}\frac{\varphi_j(x)}{\varphi_i(x)} = \varphi_j^*(x) \qquad (j = i+1, \ldots, l).$$

Тогда

$$\Phi^*(x) = -a_1\varphi_1^* - a_2\varphi_2^* - \ldots - a_{i-1}\varphi_{i-1}^* + a_{i+1}\varphi_{i+1}^* + \ldots + a_l\varphi_l^*.$$

Так как $\dfrac{\Phi(x)}{\varphi_i(x)}$ на $[a, b]$ имеет p нулей, то число нулей p^* функции $\Phi^*(x)$ по теореме Ролля не может быть меньше $p-1$, т. е. $p^* \geqslant p-1$. С другой стороны, число перемен знака q^* в последовательности

$$-a_1, \ -a_2, \ldots, \ -a_{i-1}, \ a_{i+1}, \ldots, \ a_l$$

равно $q^* = q - 1$. Так как

$$W[\varphi_{k_1}, \varphi_{k_2}, \ldots, \varphi_{k_j}, \varphi_i, \varphi_{k_{j+1}}, \ldots, \varphi_{k_s}] =$$

$$= \varphi_i^{s+1} W\left[\frac{\varphi_{k_1}}{\varphi_i}, \frac{\varphi_{k_2}}{\varphi_i}, \ldots, \frac{\varphi_{k_j}}{\varphi_i}, 1, \frac{\varphi_{k_{j+1}}}{\varphi_i}, \ldots, \frac{\varphi_{k_s}}{\varphi_i}\right] =$$

$$= \varphi_i^{s+1}(x) W\left[-\left(\frac{\varphi_{k_1}}{\varphi_i}\right)', -\left(\frac{\varphi_{k_2}}{\varphi_i}\right)', \ldots, -\left(\frac{\varphi_{k_j}}{\varphi_i}\right)',\right.$$

$$\left.\left(\frac{\varphi_{k_{j+1}}}{\varphi_i}\right)', \ldots, \left(\frac{\varphi_{k_s}}{\varphi_i}\right)'\right] = \varphi_i^{s+1}(x) W[\varphi_{k_1}^*, \varphi_{k_2}^*, \ldots, \varphi_{k_s}^*],$$

то определители Вронского для системы $\varphi_1^*, \varphi_2^*, \ldots, \varphi_{l-1}^*, \varphi_{l+1}^*, \ldots$ $\ldots, \varphi_l^*$ обладают тем же свойством, т. е. будут все положительны. Таким образом, в силу предположения $q^* \geqslant p^*$, т. е.

$$q = q^* + 1 \geqslant p^* + 1 \geqslant p.$$

В качестве примера рассмотрим совокупность функций

$$e^{\lambda_1 x}, e^{\lambda_2 x}, \ldots, e^{\lambda_n x},$$

где λ_i — действительные числа и $\lambda_1 < \lambda_2 < \ldots < \lambda_n$. В этом случае при всех ν

$$W[e^{\lambda_\alpha x}, e^{\lambda_\beta x}, \ldots, e^{\lambda_\nu x}] = e^{(\lambda_\alpha + \lambda_\beta + \ldots + \lambda_\nu)x} \begin{vmatrix} 1 & 1 & \ldots & 1 \\ \lambda_\alpha & \lambda_\beta & \ldots & \lambda_\nu \\ \ldots & \ldots & \ldots & \ldots \\ \lambda_\alpha^j & \lambda_\beta^j & \ldots & \lambda_\nu^j \end{vmatrix} > 0.$$

Следовательно, линейная комбинация

$$a_1 e^{\lambda_1 x} + a_2 e^{\lambda_2 x} + \ldots + a_n e^{\lambda_n x}$$

с не равными одновременно нулю коэффициентами a_i не может иметь нулей больше числа перемен знака в последовательности

$$a_1, a_2, \ldots, a_n.$$

Используя свойство (2) определителей Вронского и полагая $y = \ln x$, получим:

$$W[e^{\lambda_\alpha y}, e^{\lambda_\beta y}, \ldots, e^{\lambda_\nu y}] = W[x^{\lambda_\alpha}, x^{\lambda_\beta}, \ldots, x^{\lambda_\nu}] =$$

$$= x^{\frac{j(j-1)}{2}} W[e^{\lambda_\alpha y}, \ldots, e^{\lambda_\nu y}]_{y=\ln x}.$$

Таким образом, при $x > 0$

$$W[x^{\lambda_\alpha}, x^{\lambda_\beta}, \ldots, x^{\lambda_\nu}] > 0.$$

Следовательно, число положительных корней уравнения

$$a_1 x^{\lambda_1} + a_2 x^{\lambda_2} + \ldots + a_n x^{\lambda_n} = 0,$$

где a_i — действительные числа, $\sum_{i=1}^{n} a_i^2 \neq 0$, $\lambda_1 < \lambda_2 < \ldots < \lambda_n$ не превосходит числа перемен знака в последовательности $a_1, a_2, \ldots, a_n$.

В частности, правило Декарта имеет место для алгебраических уравнений

$$a_0 x^n + a_1 x^{n-1} + \ldots + a_n = 0$$

с действительными коэффициентами.

При $x \to \infty$ многочлен имеет знак, совпадающий со знаком a_0, а при $x = 0$ — знак, совпадающий со знаком a_n. Следовательно, многочлен имеет четное число положительных корней, если знаки a_0 и a_n совпадают, и нечетное число корней, если a_0 и a_n разных знаков. Но то же самое можно сказать и о числе перемен знака в последовательности $a_0, a_1, \ldots, a_n$. Таким образом, *разность между числом перемен знака в последовательности* $a_0, a_1, \ldots, a_n$ *и числом положительных корней есть число четное или нуль.*

Правило Декарта не дает точного числа корней на отрезке $[a, b]$, а лишь устанавливает их верхнюю границу, но зато оно очень просто, особенно в применении к обычным многочленам. Замена x на $-x$ позволяет получить также и верхнюю границу числа отрицательных корней уравнения.

Точное число действительных корней алгебраического уравнения, заключенных в данных пределах, может быть определено с помощью теоремы Штурма.

Теорема Штурма. *Пусть дано алгебраическое уравнение* $f(x) = 0$ *степени* n, *не имеющее кратных корней; найдем производную* $f'(x) = f_1(x)$ *и обозначим остаток от деления* $f(x)$ *на* $f'(x)$, *взятый с обратным знаком, через* $f_2(x)$; *остаток от деления* $f_1(x)$ *на* $f_2(x)$ *с обратным знаком — через* $f_3(x)$ *и т. д., до тех пор пока не придем к постоянной. Получим последовательность функций*

$$f(x), f_1(x), \ldots, f_n(x).$$

Число действительных корней уравнения $f(x) = 0$, *расположенных на отрезке* $[a, b]$, *равно разности между числом перемен знака нашей последовательности функций при* $x = a$ *и числом перемен знака последовательности при* $x = b$. (Доказательство теоремы можно найти, например, в книге А. Г. Куроша «Курс высшей алгебры».)

К теореме Штурма можно сделать следующие замечания:

1. Функции $f(x), f_1(x), \ldots, f_n(x)$ можно умножать на положительные числа.

2. Последовательность функций можно оборвать на такой функции, которая не обращается в нуль на отрезке $[a, b]$.

3. Если $f(x)=0$ имеет кратные корни, то, как и прежде, можно получить последовательность $f, f_1, \ldots, f_l, 0$; f_l не будет постоянным. Поделив все функции на f_l, получим новую последовательность. С помощью этой последовательности можно тем же способом получить число корней уравнения $f(x)=0$ на отрезке $[a, b]$, только без учета их кратности.

4. Последовательность $f, f_1, f_2, \ldots, f_n$, которую мы образовали, может быть заменена любой другой последовательностью $\varphi_0, \varphi_1, \ldots, \varphi_l$ функций, лишь бы они удовлетворяли следующим условиям:

а) последняя функция φ_l на $[a, b]$ не меняет знака;

б) две рядом стоящие функции не могут обращаться в нуль при одном и том же значении x;

в) если в последовательности $\varphi_0, \varphi_1, \ldots, \varphi_l$ какая-либо функция, за исключением первой, обращается в нуль при $x=\alpha$, то две соседние к ней функции в некоторой окрестности этого значения имеют различные знаки;

г) отношение φ_0/φ_1 при переходе через нуль меняет знак с отрицательного на положительный.

Такая последовательность функций называется *последовательностью Штурма.*

Теорема Штурма дает хороший в теоретическом отношении способ определения числа действительных корней, расположенных на данном отрезке $[a, b]$, но при практическом применении требует очень большой вычислительной работы. Менее совершенна в теоретическом отношении, но более удобна для практики *теорема Бюдана:*

Число действительных корней алгебраического уравнения $f(x)=0$ степени n, расположенных на отрезке $[a, b]$, не превышает числа потерянных перемен знака в последовательности

$$f(x), f'(x), \ldots, f^{(n)}(x)$$

при переходе от $x=a$ к $x=b$, и разность между ними есть число четное.

4. Отделение действительных корней алгебраического уравнения. Задача отделения действительных корней уравнения $f(x)=0$ заключается в том, чтобы каждый из корней заключить в интервал, не содержащий других корней уравнения. Обычно для решения этой задачи сначала находят нижнюю и верхнюю границы всех действительных корней уравнения, что может быть сделано одним из тех способов, о которых мы говорили ранее. Затем полученный отрезок разбивают на более мелкие, обычно равной длины, так, чтобы в каждом из них не могло содержаться больше одного корня. Для того чтобы определить длину этих частичных отрезков рас-

смотрим определитель

$$D=\begin{vmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \\ \dots & \dots & \dots & \dots \\ x_1^{n-1} & x_2^{n-1} & \dots & x_n^{n-1} \end{vmatrix}=\prod_{i>j}(x_i-x_j), \tag{8}$$

где $x_1, x_2, \dots, x_n$ — корни данного уравнения. Введем обозначение

$$s_k=x_1^k+x_2^k+\dots+x_n^k \tag{9}$$

и возведем D в квадрат, используя правило умножения определителей «строка на строку». Получим:

$$D^2=\begin{vmatrix} s_0 & s_1 & s_2 & \dots & s_{n-1} \\ s_1 & s_2 & s_3 & \dots & s_n \\ \dots & \dots & \dots & \dots & \dots \\ s_{n-1} & s_n & s_{n+1} & \dots & s_{2n-2} \end{vmatrix}=\prod_{i>j}(x_i-x_j)^2. \tag{10}$$

Зная величину D, можно оценить расстояние между корнями. Действительно,

$$\frac{D}{(x_\alpha-x_\beta)}=\prod_{i>j}{}'(x_i-x_j),$$

где штрих означает, что мы в правой части опускаем множитель $x_\alpha-x_\beta$. Если M — верхняя граница модулей корней уравнения, то

$$\frac{|D|}{|x_\alpha-x_\beta|}\leqslant(2M)^{\frac{n(n-1)}{2}-1} \quad \text{и} \quad |x_\alpha-x_\beta|\geqslant\frac{|D|}{(2M)^{\frac{n(n-1)}{2}-1}}. \tag{11}$$

Для отыскания D найдем D^2. Пусть уравнение имеет вид

$$f(x)=x^n+a_1x^{n-1}+\dots+a_{n-1}x+a_n=0.$$

Тогда

$$f'(x)=nx^{n-1}+(n-1)a_1x^{n-2}+\dots+a_{n-1}.$$

С другой стороны,

$$f'(x)=\frac{f(x)}{x-x_1}+\frac{f(x)}{x-x_2}+\dots+\frac{f(x_n)}{x-x_n}.$$

Производя деление, получим:

$$\frac{f(x)}{x-x_i}=x^{n-1}+(x_i+a_1)x^{n-2}+(x_i^2+a_1x_i+a_2)x^{n-3}+\dots+$$
$$+(x_i^{n-1}+a_1x_i^{n-2}+\dots+a_{n-1})$$

или

$$f'(x)=nx^{n-1}+(s_1+na_1)x^{n-2}+(s_2+a_1s_1+na_2)x^{n-3}+\dots+$$
$$+(s_{n-1}+a_1s_{n-2}+\dots+na_{n-1}).$$

Приравнивая коэффициенты при одинаковых степенях x, получим:

$$s_k + a_1 s_{k-1} + a_2 s_{k-2} + \ldots + a_{k-1} s_1 + k a_k = 0 \qquad (k = 1, 2, \ldots$$
$$\ldots, (n-1).\ s_0 = n).$$

Для отыскания соответствующих выражений при $k \geqslant n$ умножим наше уравнение на x^m и в уравнении

$$x^{n+m} + a_1 x^{n+m-1} + \ldots + a_{n-1} x^{m+1} + a_n x^m = 0$$

положим последовательно $x = x_1, x_2, \ldots, x_n$. Складывая полученные результаты, найдем:

$$s_{n+m} + a_1 s_{n+m-1} + \ldots + a_{n-1} s_{m+1} + a_n s_m = 0 \qquad (m = 0, 1, \ldots).$$

Эти формулы нам дают возможность последовательно найти $s_0, s_1, s_2, \ldots, s_{2n-2}$, а тем самым и D^2. Например, для кубичного уравнения $x^3 + a_1 x^2 + a_2 x + a_3 = 0$ будем иметь:

$$s_0 = 3,$$
$$s_1 + a_1 = 0, \quad s_1 = -a_1,$$
$$s_2 + a_1 s_1 + 2a_2 = 0, \quad s_2 = a_1^2 - 2a_2,$$
$$s_3 + a_1 s_2 + a_2 s_1 + 3a_3 = 0, \quad s_3 = -a_1^3 + 3a_1 a_2 - 3a_3,$$
$$s_4 + a_1 s_3 + a_2 s_2 + a_3 s_1 = 0, \quad s_4 = a_1^4 - 4a_1^2 a_2 + 4a_1 a_3 + 2a_2^2,$$
$$D^2 = a_1^2 a_2 - 4a_1^3 a_3 - 4a_2^3 - 27a_3^2 + 18 a_1 a_2 a_3.$$

Обычно этот способ дает очень заниженные значения для $\inf |x_i - x_j|$, и нужна очень большая вычислительная работа для последующих подстановок.

Рассмотрим *метод Фурье отделения корней*, основанный на теореме Бюдана. Но прежде чем излагать сам способ, докажем несколько вспомогательных утверждений.

1. *Если уравнение $f(x) = 0$ имеет на отрезке $[a, b]$ k корней, а при переходе от a к b в последовательности $f(x), f'(x), \ldots$ $\ldots, f^{(n)}(x)$ теряется $k + 2l$ перемен знака, то уравнение имеет еще по крайней мере $2l$ комплексных корней.*

Действительно, пусть между $-\infty$ и a имеется k_1 действительных корней, а между b и $+\infty$ k_2 действительных корней. По теореме Бюдана при переходе от $-\infty$ к a в последовательности $f(x), f'(x), \ldots, f^{(n)}(x)$ теряется $k_1 + 2l_1$ перемен знака, а при переходе от b к $+\infty$ теряется $k_2 + 2l_2$ перемен знака, где l_1 и l_2 — некоторые целые числа. Число перемен знака, теряющихся при переходе от $-\infty$ к $+\infty$, равно n. Таким образом, $n = k_1 + k_2 + k + 2l + 2l_1 + 2l_2$, а число комплексных корней $s = n - k_1 - k_2 - k$. Следовательно,

$$s = 2l + 2l_1 + 2l_2 \geqslant 2l.$$

2. *Если функция $f(x)$ на отрезке $[a, b]$ имеет две непрерывные производные и $f'(x)$ на этом отрезке не обращается в нуль, то функция $\varphi(x) = x - \frac{f(x)}{f'(x)}$ является возрастающей в тех точках, где $f(x)$ и $f''(x)$ имеют одинаковый знак, и убывающей в тех точках, где они имеют разные знаки.*

Действительно,

$$\varphi'(x) = 1 - \frac{[f'(x)]^2 - f''(x)f(x)}{[f'(x)]^2} = \frac{f(x)f''(x)}{[f'(x)]^2},$$

а это означает, что при $f(x)f''(x) > 0$ функция $\varphi(x)$ возрастает, а при $f(x)f'(x) < 0$ убывает.

Из второго утверждения можно получить два следствия.

Следствие 1. *Если на отрезке $[a, b]$ уравнение $f(x) = 0$ имеет два действительных корня, а $f''(x)$ не обращается в нуль на этом отрезке, то имеет место неравенство*

$$\frac{f(b)}{f'(b)} - \frac{f(a)}{f'(a)} < b - a.$$

Действительно, пусть α, β ($\alpha < \beta$) — корни уравнения $f(x) = 0$ на отрезке $[a, b]$; $f'(x)$ имеет на отрезке $[a, b]$ только один корень. Он расположен между $x = \alpha$ и $x = \beta$. Кривая $y = f(x)$ на отрезке $[a, b]$ или выпукла или вогнута в зависимости от знака $f''(x)$, поэтому $f(x)$ и $f''(x)$ на отрезках $[a, \alpha]$ и $[\beta, b]$ имеют одинаковые знаки (рис. 4 и 5). Следовательно, функция $\varphi(x) = x - \frac{f(x)}{f'(x)}$ на этих

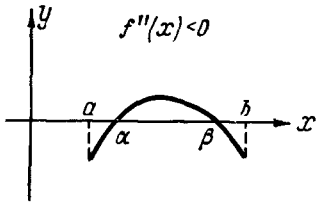

Рис. 4.

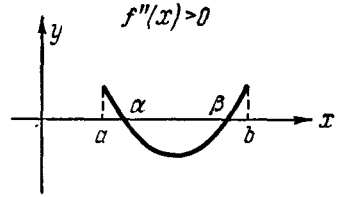

Рис. 5.

отрезках возрастает и $\varphi(a) < \varphi(\alpha)$, $\varphi(\beta) < \varphi(b)$. Но $\varphi(a) = \alpha$, а $\varphi(\beta) = \beta$, поэтому $\varphi(\alpha) < \varphi(\beta)$ и, тем более, $\varphi(\alpha) < \varphi(b)$, откуда и следует, что $\frac{f(b)}{f'(b)} - \frac{f(a)}{f'(a)} < b - a$.

Следствие 2. *Если на отрезке $[a, b]$ уравнение $f'(x) = 0$ имеет один корень x_1, а функции $f(x)$, $f''(x)$ не обращаются в нуль и $f(x)f''(x) > 0$ на $[a, b]$ и имеет место неравенство $\frac{f(b)}{f'(b)} - \frac{f(a)}{f'(a)} < b - a$, то его можно нарушить, увеличив a или уменьшив b на соответствующую величину.*

В самом деле, функция $\varphi(x)$ возрастает при изменении x от a до x_1 и при изменении x от x_1 до b, а функция $\varphi(a) - \varphi(x)$ будет

убывать на $[x_1, b]$ от $+\infty$ до $\varphi(a) - \varphi(b)$, а функция $\varphi(x) - \varphi(b)$ будет возрастать от $\varphi(a) - \varphi(b)$ до $+\infty$ на $[a, x_1]$. Если $a' \in [a, x_1]$ и $b' \in [x_1, b]$ и близки к x_1, то $\varphi(a) - \varphi(b') > 0$ и $\varphi(a') - \varphi(b) > 0$ и, таким образом,

$$\frac{f(b')}{f'(b')} - \frac{f(a)}{f'(a)} > b' - a, \quad \frac{f(b)}{f'(b)} - \frac{f(a')}{f'(a')} > b - a'.$$

Теперь можно изложить способ Фурье отделения действительных корней алгебраического уравнения $f(x) = 0$ степени n.

Пусть нам дано уравнение $f(x) = 0$ степени n. Образуем последовательность $f(x), f'(x), \ldots, f^{(n)}(x)$. На отрезке $[a, b]$, на котором могут быть действительные корни, возьмем ряд возрастающих чисел $a, \alpha_1, \alpha_2, \ldots, \alpha_n, b$. Если при переходе от α_i к α_{i+1} наша последовательность не теряет ни одной перемены знака, то уравнение не будет иметь ни одного корня на отрезке $[\alpha_i, \alpha_{i+1}]$. Если теряется только одна перемена знака, то имеется только один корень и он уже отделен. При потере двух перемен знака могут быть два случая: 1) на $[\alpha_i, \alpha_{i+1}]$ имеется два корня уравнения и 2) на $[\alpha_i, \alpha_{i+1}]$ нет корней. В последнем случае уравнение $f(x) = 0$ имеет обязательно пару комплексных корней. Если потеря перемен знака происходит в первых трех функциях последовательности, то эти два случая можно различить. В самом деле, $f''(x)$ не имеет корня в $[\alpha_i, \alpha_{i+1}]$. Поэтому, если окажется, что

$$\frac{f(\alpha_{i+1})}{f'(\alpha_{i+1})} - \frac{f(\alpha_i)}{f'(\alpha_i)} \geqslant \alpha_{i+1} - \alpha_i,$$

то по следствию второго утверждения $f(x)$ не может иметь двух корней в $[\alpha_i, \alpha_{i+1}]$, т. е. на отрезке нет ни одного корня. Если неравенство не выполняется, то нужно на отрезке $[\alpha_i, \alpha_{i+1}]$ выбрать точку β и проводить рассуждения с отрезками $[\alpha_i, \beta]$ и $[\beta, \alpha_{i+1}]$.

В общем случае, когда число потерянных перемен знака при переходе от α_i к α_{i+1} больше двух или равно двум, но потери происходят не за счет первых трех функций, вопрос может быть решен следующим образом. Обозначим через Δ_k число перемен знака потерянных при переходе от α_i к α_{i+1} в последовательности

$$f^{(k)}(x), \quad f^{(k+1)}(x), \ldots, f^{(n)}(x).$$

Общий случай соответствует $\Delta_0 \geqslant 2$. В последовательности $\Delta_0, \Delta_1, \ldots$ найдем первое число, равное 1. Пусть это будет Δ_m. Тогда Δ_{m-1} может иметь значение 0 или 2; первый случай невозможен, так как по условию $\Delta_0 \geqslant 2$, и поэтому можно было бы найти $\Delta_l = 1$ с $l < m$. Если при этом $\Delta_{m+1} \neq 0$, то всегда можно так уменьшить отрезок $[\alpha_i, \alpha_{i+1}]$, взяв $\alpha_i < \alpha_i' < \alpha_{i+1}' < \alpha_{i+1}$, что корень $f^{(m)}(x)$ принадлежит отрезку $[\alpha_i', \alpha_{i+1}']$, а $f^{(m+1)}(x)$ не имеет там корней. Тогда $f^{(m)}(x)$ не имеет корней на $[\alpha_i, \alpha_i']$ и $[\alpha_{i+1}', \alpha_{i+1}]$ и, следовательно,

Δ_m для них равно нулю. Поэтому первое из Δ, равное 1, перемещается для этих отрезков влево. Для отрезка $[\alpha'_i, \alpha'_{i+1}]$ будем иметь $\Delta_{m-2}=2$, $\Delta_m=1$, $\Delta_{m+1}=0$. При этом может оказаться, что Δ_m не будет первым из Δ, равным единице. Тогда мы так же переместимся влево и будем повторять наши рассуждения. Таким образом, нам нужно только рассмотреть случай, когда $\Delta_{m-1}=2$, $\Delta_m=1$, $\Delta_{m+1}=0$ и Δ_m — первое из Δ, равное 1.

В этом случае уравнение $f^{(m-1)}(x)$ либо имеет два корня на $[\alpha'_i, \alpha'_{i+1}]$, либо не имеет ни одного. Эти случаи можно различить так, как мы это делали при $\Delta_0=2$, $\Delta_1=1$, $\Delta_2=0$. Если имеется два корня, то мы отделяем их изложенным выше способом, при этом отрезок $[\alpha'_i, \alpha'_{i+1}]$ разобьется на два отрезка $[\alpha'_i, \beta]$ и $[\beta, \alpha'_{i+1}]$, для которых $\Delta_{m\ 1}=1$, т. е. первое из Δ, равное 1, переместится влево. Если окажется, что корни комплексные, то можно показать, что все уравнения

$$f^{(m-2)}(x)=0, \quad f^{(m-3)}(x)=0, \ \ldots, \quad f(x)=0$$

имеют пару комплексных корней. Тогда мы уменьшим все Δ_k при $k \leqslant m-1$ на 2 и получим $\Delta_{m-1}=0$. Таким образом, мы последовательно подвигаемся влево до тех пор, пока не придем к уже рассмотренным случаям.

Может случиться, что уравнение $f^{(n-1)}(x)$ имеет два равных корня в $[\alpha'_i, \alpha'_{i+1}]$. Тогда будет справедлив тот же вывод, если только само уравнение не имеет кратных корней. Если же двойной корень принадлежит уравнениям

$$f^{(n-2)}(x)=0, \quad f^{(n-3)}(x)=0, \ \ldots, \quad f(x)=0,$$

то исходное уравнение в данной точке имеет корень кратности $n-1$.

Пример. Отделить корни уравнения

$$x^5-4x^4+6x^3-3x^2+2x+1=0.$$

Последовательность $f(x)$, $f'(x)$, ..., $f^{(5)}(x)$ будет выглядеть так:

$$f(x)=x^5-4x^4+6x^3-3x^2+2x+1, \quad \frac{f'''(x)}{3!}=10x^2-16x+6,$$

$$\frac{f'(x)}{1!}=5x^4-16x^3+18x^2-6x+2, \quad \frac{f^{(4)}(x)}{4!}=5x-4,$$

$$\frac{f''(x)}{2!}=10x^3-24x^2+18x-3, \quad \frac{f^{(5)}(x)}{5!}=1.$$

По признаку Ньютона находим, что все корни уравнения заключены между -1 и $+1$. Составим таблицу знаков последовательности в точках -1, 0, $+1$:

x	-1	0	$+1$
f	$-$	$+$	$+$
f'	$+$	$+$	$+$
f''	$-$	$-$	$+$
f'''	$+$	$+$	0
$f^{(4)}$	$-$	$-$	$+$
$f^{(5)}$	$+$	$+$	$+$

На отрезке $[-1, 0]$ имеем $\Delta_0 = 1$, а следовательно, на этом отрезке имеется один действительный корень. На отрезке $[0, 1]$ нужно разобраться со знаком $\frac{f'''(x)}{3!} = 10(x-1)\left(x-\frac{3}{5}\right)$. Поэтому при значениях x, бо́льших 1, но близких к ней, все производные будут положительны. Таким образом, последовательность Δ будет иметь вид

$$4,\ 4,\ 3,\ 2,\ 1,\ 0.$$

Первое из Δ, равное 1, будет Δ_4. Исследуем, может ли уравнение $f'''(x) = 0$ иметь действительные корни на $[0, 1]$:

$$\frac{f'''(1)}{f^{(4)}(1)} - \frac{f'''(0)}{f^{(4)}(0)} = \frac{0}{1} - \frac{6}{(-4)} = \frac{3}{2} > 1 - 0.$$

Следовательно, мы должны уменьшить Δ_0, Δ_1, Δ_2, Δ_3 на 2. Получим 2, 2, 1, 0. Опять нужно исследовать, имеет ли $f'(x) = 0$ действительные корни на $[0, 1]$:

$$\frac{f'(1)}{f''(1)} - \frac{f'(0)}{f''(0)} = \frac{3}{1} - \frac{2}{(-3)} = \frac{11}{3} > 1 - 0,$$

а это означает, что действительных корней нет. Уменьшая еще раз все Δ на 2, получим $\Delta_0 = 0$. Следовательно, других действительных корней наше уравнение не имеет, т. е. имеет один действительный корень на отрезке $[-1, 0]$ и четыре комплексных корня.

5. Отделение комплексных корней алгебраических уравнений. Мы изложим способ отделения комплексных корней алгебраического уравнения

$$f(z) = z^n + a_1 z^{n-1} + \ldots + a_{n-1} z + a_n = 0, \qquad (12)$$

использующий понятие индекса многочлена $f(z)$ относительно некоторой заданной прямой в плоскости комплексного переменного z.

Пусть на прямой

$$\operatorname{Re} z = \alpha \quad \text{или} \quad z = \alpha + it \qquad (-\infty < t < +\infty) \qquad (13)$$

уравнение (12) не имеет корней. Значения многочлена $f(z)$ вдоль этой прямой являются функцией t, которую можно записать в таком виде:

$$f(\alpha+it)=P(t)+iQ(t)=R(t)\,e^{i\varphi(t)}, \tag{14}$$

где $P(t)$ и $Q(t)$ — действительные функции от t, $R(t)=\sqrt{P^2(t)+Q^2(t)}$, $\operatorname{tg}\varphi(t)=\frac{Q(t)}{P(t)}$. По условию $R(t)$ не обращается в нуль. $R(t)$ и $\varphi(t)$ — непрерывные функции от t при изменении t от $-\infty$ до $+\infty$.

Индекс многочлена $f(z)$ относительно прямой (13) определим так:

$$I_f(\operatorname{Re} z=\alpha)=\frac{1}{\pi}[\varphi(+\infty)-\varphi(-\infty)]. \tag{15}$$

Имеет место следующее свойство индекса I_f:

Если $f(z)=f_1(z)f_2(z)$, *где* $f_1(z)$ *и* $f_2(z)$ — *многочлены относительно* z, *то*

$$I_f(\operatorname{Re} z=\alpha)=I_{f_1}(\operatorname{Re} z=\alpha)+I_{f_2}(\operatorname{Re} z=\alpha). \tag{16}$$

В самом деле, если

$$f(\alpha+it)=R(t)\,e^{i\varphi(t)},\quad f_j(\alpha+it)=R_j(t)\,e^{i\varphi_j(t)}\qquad (j=1,\ 2),$$

то

$$R(t)=R_1(t)\cdot R_2(t),\quad \varphi(t)=\varphi_1(t)+\varphi_2(t).$$

Отсюда

$$I_f(\operatorname{Re} z=\alpha)=\frac{1}{\pi}[\varphi(+\infty)-\varphi(-\infty)]=$$

$$=\frac{1}{\pi}\{[\varphi_1(+\infty)-\varphi_1(-\infty)]+[\varphi_2(+\infty)-\varphi_2(-\infty)]\}=$$

$$=I_{f_1}(\operatorname{Re} z=\alpha)+I_{f_2}(\operatorname{Re} z=\alpha).$$

Рассмотрим связь индекса $I_f(\operatorname{Re} z=\alpha)$ с расположением корней уравнения (12) относительно прямой $\operatorname{Re} z=\alpha$. Пусть сначала $f(z)=z-z_0$, где $z_0=x_0+iy_0$. В этом случае

$$f(\alpha+it)=\alpha-x_0+i(t-y_0),\quad \operatorname{tg}\varphi(t)=\frac{t-y_0}{\alpha-x_0}.$$

Если $\alpha-x_0>0$, то при изменении t от $-\infty$ до $+\infty$ $\operatorname{tg}\varphi(t)$ изменяется от $-\infty$ до $+\infty$, а $\varphi(t)$ от $-\frac{\pi}{2}$ до $\frac{\pi}{2}$. (За $\varphi(-\infty)$ мы всегда будем принимать значение угла в пределах $\left(-\frac{\pi}{2},\ \frac{\pi}{2}\right)$, для которого тангенс равен $\lim\limits_{t\to-\infty}\frac{Q(t)}{P(t)}$.) Следовательно,

$$I_f(\operatorname{Re} z=\alpha)=\frac{1}{\pi}\left[\frac{\pi}{2}-\left(-\frac{\pi}{2}\right)\right]=1.$$

Если $\alpha - x_0 < 0$, то при изменении t от $-\infty$ до $+\infty$ $\operatorname{tg}\varphi(t)$ изменяется от $+\infty$ до $-\infty$, т. е. $\varphi(t)$ изменяется от $\frac{\pi}{2}$ до $-\frac{\pi}{2}$ и

$$I_f(\operatorname{Re} z = \alpha) = \frac{1}{\pi}\left[-\frac{\pi}{2} - \frac{\pi}{2}\right] = -1.$$

Таким образом,

$$I_{z-z_0}(\operatorname{Re} z = \alpha) = \begin{cases} 1, & \text{если } x_0 < \alpha, \\ -1, & \text{если } x_0 > \alpha. \end{cases} \tag{17}$$

Это означает, что для многочлена $f(z) = z - z_0$ индекс относительно прямой $\operatorname{Re} z = \alpha$ будет равен 1, если корень z_0 многочлена лежит слева от прямой $\operatorname{Re} z = \alpha$, и -1, если корень лежит справа от этой прямой.

Пусть теперь $f(z)$ — произвольный многочлен

$$f(z) = z^n + a_1 z^{n-1} + \ldots + a_{n-1} z + a_n = \\ = (z - z_1)(z - z_2) \ldots (z - z_n).$$

На основании свойства (16)

$$I_f(\operatorname{Re} z = \alpha) = \sum_{j=1}^{n} I_{(z-z_j)}(\operatorname{Re} z = \alpha).$$

Но $I_{(z-z_j)}(\operatorname{Re} z = \alpha)$ равен $+1$, если z_j лежит слева от прямой $\operatorname{Re} z = \alpha$ и -1, если z_j лежит справа от этой прямой. Таким образом,

$$I_f(\operatorname{Re} z = \alpha) = k - l, \tag{18}$$

где k — число корней уравнения $f(z) = 0$, лежащих слева от прямой $\operatorname{Re} z = \alpha$, а l — число корней уравнения, лежащих справа от этой прямой. Так как общее число корней уравнения $f(z) = 0$ равно $k + l = n$, то

$$k = \frac{n + I_f(\operatorname{Re} z = \alpha)}{2}, \quad l = \frac{n - I_f(\operatorname{Re} z = \alpha)}{2}. \tag{19}$$

Рассмотрим теперь прямую

$$\operatorname{Im} z = \beta, \quad \text{или} \quad z = t + i\beta \qquad (-\infty < t < \infty), \tag{20}$$

предполагая, что на ней нет корней уравнения (12). Пусть

$$f(t + i\beta) = P_1(t) + iQ_1(t) = R_1(t)\, e^{i\psi(t)} \qquad (R_1(t) \neq 0),$$

$$\operatorname{tg}\psi(t) = \frac{Q_1(t)}{P_1(t)}. \tag{21}$$

Определим индекс многочлена относительно этой прямой равенством

$$I_f(\operatorname{Im} z = \beta) = \frac{1}{\pi}[\psi(+\infty) - \psi(-\infty)]. \tag{22}$$

Снова изучим связь индекса $I_f(\operatorname{Im} z=\beta)$ с расположением корней относительно прямой $\operatorname{Im} z=\beta$, положив вначале, что $f(z)=z-z_0$ $(z_0=x_0+iy_0)$:

$$f(t+i\beta)=t-x_0+i(\beta-y_0),$$

$$\operatorname{tg}\psi(t)=\frac{\beta-y_0}{t-x_0}.$$

Если $\beta-y_0<0$, то $\operatorname{tg}\psi(t)$ при возрастании t от $-\infty$ до x возрастает от 0 до $+\infty$, а $\psi(t)$ возрастает от 0 до $\frac{\pi}{2}$; при возрастании t от x_0 до $+\infty$ $\frac{\beta-y_0}{t-x_0}$ возрастает от $-\infty$ до 0, а $\psi(t)$ от $\frac{\pi}{2}$ до π. Таким образом,

$$I_f(\operatorname{Im} z=\beta)=\frac{1}{\pi}[\pi-0]=1.$$

Если $\beta-y_0>0$, то $\operatorname{tg}\psi(t)$ при изменении t от $-\infty$ до x_0 убывает от 0 до $-\infty$, а $\psi(t)$ убывает от 0 до $-\frac{\pi}{2}$; при изменении t от x_0 до $+\infty$ $\operatorname{tg}\psi(t)$ убывает от $+\infty$ до 0, т. е. $\psi(t)$ убывает от $-\frac{\pi}{2}$ до $-\pi$. Следовательно,

$$I_f(\operatorname{Im} z=\beta)=\frac{1}{\pi}[-\pi-0]=-1.$$

Итак,

$$I_{z-z_0}(\operatorname{Im} z=\beta)=\begin{cases}+1, & \text{если } y_0>\beta,\\ -1, & \text{если } y_0<\beta,\end{cases} \tag{23}$$

т. е. если корень z_0 лежит выше прямой $\operatorname{Im} z=\beta$, то индекс равен $+1$; если же ниже прямой, то -1.

В общем случае многочлена

$$f(z)=z^n+a_1z^{n-1}+\ldots+a_{n-1}z+a_n=$$
$$=(z-z_1)(z-z_2)\ldots(z-z_n),$$

используя свойство (16), будем иметь:

$$I_f(\operatorname{Im} z=\beta)=\sum_{j=1}^{n}I_{z-z_j}(\operatorname{Im} z=\beta)=p-q, \tag{24}$$

где p — число корней уравнения $f(z)=0$, лежащих выше прямой $\operatorname{Im} z=\beta$, а q — число корней, лежащих ниже прямой. Так как

$$p+q=n,$$

то

$$p=\frac{n+I_f(\operatorname{Im} z=\beta)}{2},\quad q=\frac{n-I_f(\operatorname{Im} z=\beta)}{2}. \tag{25}$$

Теперь уже совершенно ясно, как можно выполнить отделение действительных и комплексных корней уравнения (12).

Для определения числа корней уравнения $f(z)=0$, расположенных в полосе $\alpha_0 < \operatorname{Re} z < \alpha_1$, вычисляем $I_f(\operatorname{Re} z=\alpha_0)$ и $I_f(\operatorname{Re} z=\alpha_1)$. Тогда число корней в этой полосе будет равно

$$k_1 - k_0 = \frac{1}{2}[I_f(\operatorname{Re} z=\alpha_1) - I_f(\operatorname{Re} z=\alpha_0)]. \tag{26}$$

Для определения числа корней в полосе $\beta_0 < \operatorname{Im} z < \beta_1$ вычисляем $I_f(\operatorname{Im} z=\beta_0)$ и $I_f(\operatorname{Im} z=\beta_1)$; тогда число корней в этой полосе будет

$$p_0 - p_1 = \frac{1}{2}[I_f(\operatorname{Im} z=\beta_0) - I_f(\operatorname{Im} z=\beta_1)]. \tag{27}$$

Зная границы области, в которой расположены все корни уравнения (12), и применяя этот метод, можно произвести отделение всех действительных и комплексных корней.

Для вычисления индекса многочлена $f(z)$ относительно прямой $\operatorname{Re} z=\alpha$ необходимо найти приращение $\varphi(t)=\arg f(\alpha+it)$ при возрастании t от $-\infty$ до $+\infty$, так как

$$I_f(\operatorname{Re} z=\alpha)=\frac{1}{\pi}[\varphi(+\infty)-\varphi(-\infty)],$$

где

$$\operatorname{tg}\varphi(t)=\frac{Q(t)}{P(t)}, \quad f(\alpha+it)=P(t)+iQ(t).$$

Функции $Q(t)$ и $P(t)$ являются многочленами относительно t, причем если степень n многочлена $f(z)$ четна, то $P(t)$ имеет степень n, а $Q(t)$ — не выше $n-1$; если же степень n нечетна, то $P(t)$ есть многочлен степени не выше $n-1$, а $Q(t)$ — в точности степени n. Так как предполагается, что на прямой $\operatorname{Re} z=\alpha$ нет корней уравнения $f(z)=0$, то $P(t)$ и $Q(t)$ не могут одновременно обращаться в нуль. Функция $\varphi(t)=\arg f(\alpha+it)$ при изменении t от $-\infty$ до $+\infty$ непрерывна, а $\operatorname{tg}\varphi(t)=\dfrac{Q(t)}{P(t)}$ терпит бесконечные разрывы в точках t_i, являющихся действительными корнями многочлена $P(t)$. Разрывы $\dfrac{Q(t)}{P(t)}$ могут иметь один из следующих видов:

$$\left.\begin{aligned} &1)\ \lim_{t\to t_i\mp 0}\frac{Q(t)}{P(t)}=+\infty; \quad 2)\ \lim_{t\to t_i\mp 0}\frac{Q(t)}{P(t)}=-\infty;\\ &3)\ \lim_{t\to t_i\mp 0}\frac{Q(t)}{P(t)}=\pm\infty; \quad 4)\ \lim_{t\to t_i\mp 0}\frac{Q(t)}{P(t)}=\mp\infty. \end{aligned}\right\} \tag{28}$$

Функция $\operatorname{tg}\varphi(t)$ имеет разрыв в точках $\varphi=\pm\frac{\pi}{2}, \pm\frac{3\pi}{2}, \pm\frac{5\pi}{2}, \ldots$, причем все эти разрывы вида 3).

Для того чтобы найти приращение $\varphi(t)$ при переходе по t от $-\infty$ до $+\infty$, нужно найти все действительные корни многочлена $P(t)$: $-\infty < t_1 < t_2 < \ldots < t_k < +\infty$, и рассмотреть последовательность интервалов: $(-\infty, t_1)$, (t_1, t_2), $\ldots$, (t_{k-1}, t_k), $(t_k, +\infty)$. На каждом из интервалов (t_i, t_{i+1}) $\operatorname{tg}\varphi(t)$ есть непрерывная функция, а значения $\varphi(t)$ находятся в интервале $\left(\frac{2k_i-1}{2}\pi, \frac{2k_i+1}{2}\pi\right)$ или $\left(-\frac{2k_i+1}{2}\pi, -\frac{2k_i-1}{2}\pi\right)$, где k_i — некоторое целое число. При переходе t через t_{i+1} в зависимости от типа разрыва $\frac{Q(t)}{P(t)}$ при $t=t_{i+1}$ значения $\varphi(t)$ остаются или в этом же интервале (в случае разрыва вида 1) или 2)) или выходят из этого интервала, причем в случае разрыва вида 3) $\varphi(t)$ уходит из этого интервала возрастая, а в случае разрыва вида 4) — убывая. Приписывая $\varphi(t)$ при $t=-\infty$ значение угла в пределах $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$, для которого тангенс равен $\lim\limits_{t\to-\infty}\frac{Q(t)}{P(t)}$, и следя за изменением $\varphi(t)$ при переходе через точки $t_1, t_2, \ldots, t_k$, мы и найдем приращение $\varphi(t)$ при возрастании t от $-\infty$ до $+\infty$. В случае четной степени n $\lim\limits_{t\to\mp\infty}\frac{Q(t)}{P(t)}=0$ и $\varphi(-\infty)=0$, а $\varphi(+\infty)=l\pi$, где l — целое число, равное разности числа разрывов $\frac{Q(t)}{P(t)}$ вида 3) и числа разрывов вида 4), т. е. в этом случае индекс равен числу разрывов вида 3) без числа разрывов вида 4). В случае нечетной степени n $\lim\limits_{t\to\mp\infty}\left|\frac{Q(t)}{P(t)}\right|=+\infty$. Если $\lim\limits_{t\to-\infty}\frac{Q(t)}{P(t)}=-\infty$, то $\varphi(-\infty)=-\frac{\pi}{2}$, а $\varphi(+\infty)=\frac{2l-1}{2}\pi$, где l равно разности числа разрывов вида 3) и числа разрывов вида 4) при $\lim\limits_{t\to+\infty}\frac{Q(t)}{P(t)}=-\infty$ и l равно разности числа разрывов вида 3) и числа разрывов вида 4) плюс единица при $\lim\limits_{t\to+\infty}\frac{Q(t)}{P(t)}=+\infty$. Если $\lim\limits_{t\to-\infty}\frac{Q(t)}{P(t)}=+\infty$, то $\varphi(-\infty)=\frac{\pi}{2}$, а $\varphi(+\infty)=\frac{2l+1}{2}\pi$, где l — число разрывов вида 3) без числа разрывов вида 4) при $\lim\limits_{t\to+\infty}\frac{Q(t)}{P(t)}=+\infty$ и l равно разности числа разрывов вида 3) без числа разрывов вида 4) минус единица при $\lim\limits_{t\to+\infty}\frac{Q(t)}{P(t)}=-\infty$. Это означает, что если рассматривать единую бесконечно удаленную точку и положить $+\infty=\infty-0$, $-\infty=\infty+0$ и в этой точке рассматривать типы разрыва так же, как и в (48), то можно и при четных и при нечетных n сформулировать одно и то же правило для вычисления индекса $I_f(\operatorname{Re} z=\alpha)$.

Индекс многочлена $f(z)$ относительно прямой $\operatorname{Re} z=\alpha$, на которой нет корней уравнения $f(z)=0$, равен разности числа

разрывов отношения $\frac{Q(t)}{P(t)}$ *вида* 3) *и числа разрывов вида* 4), *где* $P(t)$ *и* $Q(t)$ *— соответственно действительная и мнимая части функции* $f(\alpha+it)$.

Это правило, очевидно, справедливо и для вычисления индекса $I_f(\operatorname{Im} z=\beta)$, если на прямой $\operatorname{Im} z=\beta$ нет корней уравнения $f(z)=0$.

Заметим, что для вычисления индекса нет необходимости находить корни уравнения $P(t)=0$, а нужно лишь для каждого из них найти такой интервал, в котором нет других корней $P(t)$, а также корней многочлена $Q(t)$.

В заключение рассмотрим два примера.

Пример 1. Отделить корни уравнения

$$4z^3-2z^2-4z-3=0.$$

Для определения числа действительных корней и их отделения выписываем ряд Штурма

$$f_0(z)=4z^3-2z^2-4z-3,\quad f_1(z)=3z^2-z-1,\quad f_2(z)=26z+29,$$
$$f_3(z)=-1.$$

Таблица перемен знака в последовательности Штурма имеет вид:

z	$-\infty$	0	1	2	$+\infty$
$\operatorname{sign} f_0(z)$	$-$	$-$	$-$	$+$	$+$
$\operatorname{sign} f_1(z)$	$+$	$-$	$+$	$+$	$+$
$\operatorname{sign} f_2(z)$	$-$	$+$	$+$	$+$	$+$
$\operatorname{sign} f_3(z)$	$-$	$-$	$-$	$-$	$-$
Число перемен знака	2	2	2	1	1

Таким образом, действительный корень один и находится в интервале (1, 2).

Для отделения комплексных корней будем вычислять индексы многочлена относительно прямых $\operatorname{Re} z=\alpha$ и $\operatorname{Im} z=\beta$.

1) $\operatorname{Re} z=0$, $f(it)=2t^2-3-4i(t^3+1)$, $\frac{Q(t)}{P(t)}=-4\frac{t^3+1}{2t^2-3}$. Многочлен $2t^2-3$ имеет корни $t_1=-\sqrt{\frac{3}{2}}$, $t_2=+\sqrt{\frac{3}{2}}$; $\frac{Q(t)}{P(t)}$ имеет в этих точках разрывы вида: в t_1: $\pm\infty$, в t_2: $\mp\infty$. Далее,

$\lim\limits_{t\to-\infty}\frac{Q(t)}{P(t)}=+\infty$, а $\lim\limits_{t\to+\infty}\frac{Q(t)}{P(t)}=-\infty$, т. е. $\lim\limits_{t\to\mp\infty}\frac{Q(t)}{P(t)}=\pm\infty$.
Отсюда

$$I_f(\operatorname{Re} z=0)=1,\quad k-l=1,\quad k+l=3,\qquad k=2,\quad l=1.$$

Следовательно, оба комплексных корня лежат слева от прямой $\operatorname{Re} z=0$, т. е. имеют отрицательную действительную часть.

2) $\operatorname{Re} z=-1$, $f(-1+it)=14t^2-5+4i(3t-t^3)$, $\frac{Q(t)}{P(t)}=$
$=4\frac{3t-t^3}{14t^2-5}$.
Корни уравнения $14t^2-5=0$:

$$t_1\approx-0{,}6,\quad t_2\approx+0{,}6.$$

Типы разрывов t: ∞, t_1, t_2,

$$\mp\infty,\quad \mp\infty,\quad \mp\infty.$$

Отсюда $I_f(\operatorname{Re} z=-1)=-3$, $k-l=-3$, $k+l=3$, $k=0$. Таким образом, комплексные корни лежат вправо от прямой $\operatorname{Re} z=-1$, т. е. в полосе $-1<\operatorname{Re} z<0$.

3) $\operatorname{Im} z=1$, $f(t+i)=4t^3-2t^2-16t-1+4i(3t^2-t-2)$, $\frac{Q(t)}{P(t)}=4\frac{3t^2-t-2}{4t^3-2t^2-16t-1}$.
Корни знаменателя находятся в интервалах

$$t_1\in(-2;\,-1),\quad t_2\in\left(-\tfrac{2}{3};\,0\right),\quad t_3\in(2,\,3).$$

Типы разрывов: t_1: $\mp\infty$, t_2: $\mp\infty$, t_3: $\mp\infty$, ∞: 0, т. е. $I_f(\operatorname{Im} z=1)=-3$, $p-q=-3$, $p+q=3$, $p=0$. Следовательно, выше прямой $\operatorname{Im} z=1$ корней нет, а это означает, что мнимые части корней по модулю меньше 1.

Итак, уравнение $f(z)=0$ имеет три корня:

$$z_1=\alpha,\quad z_{2,3}=\beta\pm i\gamma,$$

где

$$1<\alpha<2,\quad -1<\beta<0,\quad 0<\gamma<1.$$

Пример 2. Отделить корни уравнения

$$f(z)=3z^4-z^3+6z^2-11z+3=0.$$

Метод Штурма отделения действительных корней дает такой результат: уравнение $f(z)=0$ имеет два действительных корня, которые расположены в интервалах (0, 1) и (1, 2). Для определения расположения пары комплексных корей применим изложенный метод.

1. Рассмотрим прямую $\operatorname{Re} z=0$, т. е. $z=it$:

$$f(it)=3t^4-6t^2+3+i(t^3-11t),\quad \frac{Q(t)}{P(t)}=\frac{1}{3}\frac{t^3-11t}{t^4-2t^2+1}.$$

Уравнение $P(t)=0$ имеет два двукратных корня: $t_1=-1$ и $t_2=1$. Типы разрывов t_1: $+\infty$, t_2: $-\infty$, ∞: 0, т. е. $I_f(\operatorname{Re} z=0)=0$, $k-l=0$, $k+l=4$, $k=2$. Следовательно, комплексные корни лежат влево от прямой $\operatorname{Re} z=0$, т. е. имеют отрицательную действительную часть.

2. Рассмотрим прямую $\operatorname{Re} z=-2$:

$$f(-2+it)=3t^4-84t^2+105+i\,[25t^3-143t],$$
$$\frac{Q(t)}{P(t)}=\frac{1}{3}\,\frac{25t^3-143t}{t^4-28t^2+35}.$$

Уравнение $t^4-28t^2+35=0$ имеет корни $t_1\approx-5{,}2$; $t_2\approx-1{,}1$; $t_3\approx 1{,}1$; $t_4\approx 5{,}2$. Типы разрывов:

$$t_1\colon \mp\infty,\quad t_2\colon \mp\infty,\quad t_3\colon \mp\infty,\quad t_4\colon \mp\infty,\quad \infty\colon 0.$$

Отсюда

$$I_f(\operatorname{Re} z=-2)=-4,\quad k-l=-4,\quad k+l=4,\quad k=0,$$

т. е. левее прямой $\operatorname{Re} z=-2$ корней нет, и комплексно-сопряженные корни находятся в полосе

$$-2<\operatorname{Re} z<0.$$

3. Рассмотрим прямую $\operatorname{Im} z=1$:

$$f(t+i)=3t^4-t^3-12t^2-8t+i\,(12t^3-3t^2-10),$$
$$\frac{Q(t)}{P(t)}=\frac{12t^3-3t^2-10}{3t^4-t^3-12t^2-8t}.$$

Корни уравнения $3t^4-t^3-12t^2-8t=0$: $t_1\approx-1{,}1$, $t_2\approx-1$, $t_3=0$, $t_4\approx 2{,}4$. Типы разрывов:

$$t_1\colon \mp\infty,\quad t_2\colon \pm\infty,\quad t_3\colon \mp\infty,\quad t_4\colon \mp\infty,\quad \infty\colon 0.$$

Отсюда

$$I_f(\operatorname{Im} z=1)=-2,\quad p-q=-2,\quad p+q=4,\quad p=1.$$

Следовательно, выше прямой $\operatorname{Im} z=1$ имеется один корень уравнения.

4. Рассмотрим прямую $\operatorname{Im} z=4$:

$$f(t+4i)=3t^4-t^3-282t^2+37t+675+i\,(48t^3-12t^2-720t+20),$$
$$\frac{Q(t)}{P(t)}=\frac{48t^3-12t^2-720t+20}{3t^4-t^3-282t^2+37t+675}.$$

Корни уравнения $3t^4-t^3-282t^2+37t+675=0$ находятся в интервалах $t_1\in(-10,\,-9)$; $t_2\in(-2,\,-1)$, $t_3\in(1,\,2)$, $t_4\in(9,\,10)$. Типы разрывов:

$$t_1\colon \mp\infty,\quad t_2\colon \mp\infty,\quad t_3\colon \mp\infty,\quad t_4\colon \mp\infty,\quad \infty\colon 0,$$

т. е. $I_f(\operatorname{Im} z = 4) = -4$, $p - q = -4$, $p + q = 4$, $p = 0$. Таким образом, выше прямой $\operatorname{Im} z = 4$ корней нет, а поэтому комплексные корни расположены по одному в прямоугольниках:

$$\{-2 < \operatorname{Re} z < 0,\ 1 < \operatorname{Im} z < 4\},\quad \{-2 < \operatorname{Re} z < 0,\ -4 < \operatorname{Im} z < -1\}.$$

§ 3. Метод Лобачевского решения алгебраических уравнений

Одним из широко распространенных методов решения алгебраических уравнений является метод Н. И. Лобачевского, предложенный им в 1834 г. Впоследствии он был несколько усовершенствован Греффе, Дандленом, и поэтому в литературе часто его связывают с их именами. Иногда его также называют методом квадратного корня.

1. Метод Лобачевского. Случай различных по абсолютной величине действительных корней. Пусть дано уравнение

$$a_0 x^n + a_1 x^{n-1} + \ldots + a_{n-1} x + a_n = 0, \tag{1}$$

о корнях которого известно, что все они действительны и удовлетворяют условию

$$|x_1| \gg |x_2| \gg \ldots \gg |x_n|. \tag{2}$$

Смысл знака $\gg$ (значительно больше) мы уточним несколько позднее. Воспользуемся связью между корнями и коэффициентами нашего уравнения

$$\left.\begin{aligned} x_1 + x_2 + \ldots + x_n &= -\frac{a_1}{a_0}, \\ x_1x_2 + x_1x_3 + \ldots + x_{n-1}x_n &= \frac{a_2}{a_0}, \\ x_1x_2x_3 + x_1x_2x_4 + \ldots + x_{n-2}x_{n-1}x_n &= -\frac{a_3}{a_0}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ x_1x_2x_3 \ldots x_n &= (-1)^n \frac{a_n}{a_0}. \end{aligned}\right\} \tag{3}$$

Вынесем в первом из этих равенств за скобки x_1. Получим:

$$x_1\left[1 + \frac{x_2}{x_1} + \ldots + \frac{x_n}{x_1}\right] = -\frac{a_1}{a_0}.$$

Будем считать, что x_1 настолько больше остальных корней по абсолютной величине, что можно отбросить в скобке все отношения $\frac{x_i}{x_1}$ и в пределах нашей точности равенство не нарушится. Тогда

$$x_1 \approx -\frac{a_1}{a_0}.$$

Аналогично, вынося из второго равенства за скобки x_1x_2, получим:

$$x_1x_2\left[1+\frac{x_1x_3}{x_1x_2}+\ \ldots\ +\frac{x_{n-1}x_n}{x_1x_2}\right]=\frac{a_2}{a_0}.$$

Предполагая, что отношениями $\frac{x_ix_j}{x_1x_2}$, стоящими в скобке, можно пренебречь по сравнению с единицей, будем иметь:

$$x_1x_2\approx\frac{a_2}{a_0}$$

или

$$x_2\approx-\frac{a_2}{a_1}.$$

Продолжая эти рассуждения дальше, найдем:

$$x_i\approx-\frac{a_i}{a_{i-1}}\qquad(i=1,\ 2,\ \ldots,\ n). \tag{4}$$

Таким образом, в нашем случае мы сумеем найти приближенные значения всех корней уравнения.

Н. И. Лобачевский предложил способ получения из данного уравнения (1) нового уравнения, корни которого равны квадратам корней исходного уравнения. Если исходное уравнение имело только различные по абсолютной величине действительные корни, то, применяя достаточное число раз процесс, предложенный Лобачевским, — *квадрирование*, получим новое уравнение, корни которого удовлетворяют условию (2). Таким образом, мы сможем найти корни последнего уравнения, а затем и корни исходного уравнения. Изложим процесс квадрирования. Запишем уравнение (1) в виде

$$a_0(x-x_1)(x-x_2)\ \ldots\ (x-x_n)=0.$$

Уравнение, корни которого противоположны по знаку корням уравнения (1), будет иметь вид

$$a_0(x+x_1)(x+x_2)\ \ldots\ (x+x_n)=0.$$

Перемножая эти два уравнения, получим:

$$a_0^2(x^2-x_1^2)(x^2-x_2^2)\ \ldots\ (x^2-x_n^2)=0.$$

Если теперь положить $z=-x^2$, то получим новое уравнение относительно z, корни которого равны соответственно

$$-x_1^2,\ -x_2^2,\ \ldots,\ -x_n^2.$$

Отсюда, для того чтобы получить нужное нам уравнение, мы должны перемножить уравнение (1) и уравнение, получающееся из него заменой x на $-x$, и положить затем $-x^2=z$. Найдем выражения для коэффициентов нового уравнения через старые коэффициенты. Нам нужно перемножить

$$a_0x^n+a_1x^{n-1}+a_2x^{n-2}+\ \ldots\ +a_{n-1}x+a_n=0 \tag{5}$$

и

$$a_0x^n - a_1x^{n-1} + \ldots + (-1)^{n-1}a_{n-1}x + (-1)^n a_n = 0. \quad (5')$$

Произведение будет иметь вид

$$a_0^2x^{2n} - (a_1^2 - 2a_0a_2)\,x^{2n-2} + (a_2^2 - 2a_1a_3 + 2a_0a_4)\,x^{2n-4} + \ldots$$
$$\ldots + (-1)^n a_n^2 = 0. \quad (6)$$

После замены $-x^2$ на z получим:

$$a_0^2z^n + (a_1^2 - 2a_0a_2)\,z^{n-1} +$$
$$+ (a_2^2 - 2a_1a_3 + 2a_0a_4)\,z^{n-2} + \ldots + a_n^2 = 0. \quad (7)$$

Коэффициент b_k при z^{n-k} в этом уравнении получается из коэффициентов исходного уравнения следующим образом: из квадрата a_k вычитается удвоенное произведение двух соседних с ним симметрично расположенных коэффициентов, прибавляется удвоенное произведение двух следующих за ними симметрично расположенных коэффициентов и т. д., до тех пор пока не придем к a_0 или a_n, т. е.

$$b_k = a_k^2 - 2a_{k-1}a_{k+1} + 2a_{k-2}a_{k+2} - 2a_{k-3}a_{k+3} + \ldots \quad (8)$$

Возникает вопрос: как узнать, что процесс квадрирования проведен достаточное число раз? Для того чтобы ответить на него, мы рассмотрим два уравнения:

$$b_0y^n + b_1y^{n-1} + b_2y^{n-2} + \ldots + b_n = 0,$$
$$c_0z^n + c_1z^{n-1} + c_2z^{n-2} + \ldots + c_n = 0,$$

получающиеся в процессе квадрирования, причем второе получается квадрированием первого. Если бы для первого уравнения условие (2) было выполнено, то тем более оно будет выполнено для второго. Таким образом,

$$y_i \approx -\frac{b_i}{b_{i-1}}, \quad z_i \approx -\frac{c_i}{c_{i-1}} \quad (i = 1, 2, \ldots, n).$$

Но так как $z_i = -y_i^2$, то мы получим:

$$\frac{c_i}{c_{i-1}} \approx \frac{b_i^2}{b_{i-1}^2} \qquad (i = 1, 2, \ldots, n)$$

и, следовательно, в силу того, что $c_0 = b_0^2$, будем иметь:

$$c_i \approx b_i^2, \quad (9)$$

т. е. все коэффициенты c_i являются примерно квадратами b_i. Можно показать, что при этом разделение корней уже имеет место, если исходное уравнение удовлетворяет нашим требованиям.

m		a_0	a_1	a_2	a_3	a_4
0		1	— 35	380	— 1 350	10^3
	a_i^2 $-2a_{i-1}a_{i+1}$ $2a_{i-2}a_{i+2}$	1	1 225 — 760	144 400 — 94 500 2 000	1 822 500 — 760 000	10^6
1		1	465	51 900	1 062 500	10^6
	a_i^2 $-2a_{i-1}a_{i+1}$ $2a_{i-2}a_{i+2}$	1	216 225 — 103 800	$269\,361 \cdot 10^4$ $-98\,813 \cdot 10^4$ $200 \cdot 10^4$	$112\,891 \cdot 10^7$ $-10\,380 \cdot 10^7$	10^{12}
2		1	112 425	$170\,748 \cdot 10^4$	$102\,511 \cdot 10^7$	10^{12}
	a_i^2 $-2a_{i-1}a_{i+1}$ $2a_{i-2}a_{i+2}$	1	$1\,263\,938 \cdot 10^4$ $-341\,496 \cdot 10^4$	$291\,549 \cdot 10^{13}$ $-23\,050 \cdot 10^{13}$	$105\,085 \cdot 10^{19}$ $-341 \cdot 10^{19}$	10^{24}
3		1	$922\,442 \cdot 10^4$	$268\,499 \cdot 10^{13}$	$104\,744 \cdot 10^{19}$	10^{24}
	a_i^2 $-2a_{i-1}a_{i+1}$ $2a_{i-2}a_{i+2}$	1	$850\,899 \cdot 10^{14}$ $-53\,700 \cdot 10^{14}$	$790\,917 \cdot 10^{31}$ $-1\,932 \cdot 10^{31}$	$109\,713 \cdot 10^{43}$	10^{48}
4		1	$797\,199 \cdot 10^{14}$	$718\,985 \cdot 10^{31}$	$109\,713 \cdot 10^{43}$	10^{48}
	a_i^2 $-2a_{i-1}a_{i+1}$ $2a_{i-2}a_{i+2}$	1	$635\,526 \cdot 10^{34}$ $-1438 \cdot 10^{34}$	$516\,940 \cdot 10^{68}$ $-2 \cdot 10^{68}$	$120\,369 \cdot 10^{91}$	10^{96}
5		1	$634\,088 \cdot 10^{34}$	$516\,938 \cdot 10^{68}$	$120\,369 \cdot 10^{91}$	10^{96}
	a_i^2 $-2a_{i-1}a_{i+1}$ $2a_{i-2}a_{i+2}$	1	$402\,068 \cdot 10^{74}$ $-1 \cdot 10^{74}$	$267\,225 \cdot 10^{142}$	$144\,887 \cdot 10^{187}$	10^{192}
6		1	$402\,067 \cdot 10^{74}$	$267\,225 \cdot 10^{142}$	$144\,887 \cdot 10^{187}$	10^{192}

Рассмотрим в качестве примера решение методом Лобачевского уравнения (промежуточные вычисления приведены в схеме на стр. 106)

$$x^4 - 35x^3 + 380x^2 - 1350x + 1000 = 0.$$

$$z_1 = 402\,067 \cdot 10^{74}; \quad \lg z_1 = 79{,}6042985;$$

$$\frac{1}{64}\lg z_1 = 1{,}2438172;$$

$$x_1 = 17{,}5314; \quad f(x_1) = -0{,}02;$$

$$z_2 = \frac{267\,225}{402\,067} \cdot 10^{68} = 664\,628 \cdot 10^{62};$$

$$\lg z_2 = 67{,}8225786; \quad \frac{1}{64}\lg z_2 = 1{,}0597278;$$

$$x_2 = 11{,}4743;$$

$$f(x_2) = 0{,}03; \quad z_3 = \frac{144\,887}{267\,225} \cdot 10^{45} = 0{,}542191 \cdot 10^{45};$$

$$\lg z_3 = 44{,}7341523;$$

$$\frac{1}{64}\lg z_3 = 0{,}6989711; \quad x_3 = 5{,}00001; \quad f(x_3) = 0{,}003;$$

$$z_4 = \frac{1}{144\,887} \cdot 10^5 = 0{,}690193; \quad \lg z_4 = \bar{1}{,}8389706;$$

$$\frac{1}{64}\lg z_4 = \bar{1}{,}9974839; \quad x_4 = 0{,}994223; \quad f(x_4) = -0{,}03.$$

2. Метод Лобачевского. Случай комплексных корней. Рассмотрим теперь случай, когда уравнение имеет комплексные корни. Процесс квадрирования можно будет производить, как и раньше, только выводы о способе определения корней будут уже недействительны. Предположим пока, что мы имеем только пару комплексно-сопряженных корней:

$$x_2 = \rho(\cos\varphi + i\sin\varphi), \quad x_3 = \rho(\cos\varphi - i\sin\varphi). \tag{10}$$

Тогда после m квадрирований получим уравнение, которое будет иметь пару комплексно-сопряженных корней:

$$\rho^{2^m}(\cos 2^m\varphi + i\sin 2^m\varphi),$$
$$\rho^{2^m}(\cos 2^m\varphi - i\sin 2^m\varphi).$$

Пусть, например, остальные корни будут расположены так:

$$|x_1| > |x_2| = |x_3| > |x_4| > \ldots > |x_n|. \tag{11}$$

Воспользуемся опять зависимостью между корнями и коэффициентами уравнения, полученного после m квадрирований:

$$\left.\begin{aligned} x_1^k + x_2^k + x_3^k + \ldots + x_n^k &= \frac{c_1}{c_0}, \\ x_1^k x_2^k + x_1^k x_3^k + \ldots + x_{n-1}^k x_n^k &= \frac{c_2}{c_0} \quad (k = 2^m), \\ x_1^k x_2^k x_3^k + x_1^k x_2^k x_4^k + \ldots + x_{n-2}^k x_{n-1}^k x_n^k &= \frac{c_3}{c_0}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ x_1^k x_2^k \ldots x_n^k &= \frac{c_n}{c_0}. \end{aligned}\right\} \tag{12}$$

Если при этом

$$|x_1^k| \gg |\rho^k| \gg |x_4^k| \gg \ldots \gg |x_n^k|,$$

то

$$\left.\begin{aligned} x_1^k &\approx \frac{c_1}{c_0}, \\ 2x_1^k\rho^k\left[\cos k\varphi + \frac{x_1^k x_4^k}{2x_1^k\rho^k} + \ldots\right] &= \frac{c_2}{c_0}, \\ x_1^k\rho^{2k} &\approx \frac{c_3}{c_0}, \\ \ldots\ldots\ldots & \\ x_1^k\rho^{2k}x_4^k \ldots x_n^k &= \frac{c_n}{c_0}, \end{aligned}\right\} \tag{13}$$

и мы можем найти все действительные корни и модуль пары комплексно-сопряженных корней. При этом все коэффициенты, кроме второго, ведут себя, как и раньше, т. е. стремятся к квадратам коэффициентов предыдущего уравнения, а c_2 ведет себя неправильно, возможно даже меняя свой знак. Последнее и служит признаком наличия пары комплексно-сопряженных корней. Легко заметить, что если модуль пары комплексных сопряженных корней заключен между x_i и x_{i+3}, то ведет себя неправильно коэффициент c_{i+1}, и мы точно так же можем найти все действительные корни и модуль пары комплексно-сопряженных корней. Вообще, если уравнение имеет несколько пар комплексно-сопряженных корней и их модули различны и отличны от модулей действительных корней, то можно по тем же признакам определить, где они расположены, и найти действительные корни и модули всех комплексных корней.

Если мы имеем пару или две пары комплексно-сопряженных корней, то аргументы этих корней находятся также без труда. Дей-

ствительно, если имеется только пара комплексно-сопряженных корней x_2, x_3, то из исходного уравнения имеем:

$$x_1 + 2\rho\cos\varphi + x_4 + \ldots + x_n = -\frac{a_1}{a_0}. \qquad (14)$$

В равенстве (14) нам не известен только $\cos\varphi$, и он может быть из этого равенства найден. Если комплексных корней имеется две пары: x_2, x_3 и x_4, x_5, то в уравнении следует заменить неизвестное на $1/y$. Тогда уравнение относительно y примет вид

$$a_n y^n + a_{n-1} y^{n-1} + \ldots + a_1 y + a_0 = 0$$

и будет иметь корнями:

$$\frac{1}{x_1}, \quad \frac{1}{x_2} = \frac{1}{\rho_1}(\cos\varphi - i\sin\varphi), \quad \frac{1}{x_3} = \frac{1}{\rho_1}(\cos\varphi + i\sin\varphi),$$

$$\frac{1}{x_4} = \frac{1}{\rho_2}(\cos\psi - i\sin\psi), \quad \frac{1}{x_5} = \frac{1}{\rho_2}(\cos\psi + i\sin\psi), \quad \frac{1}{x_6}, \ldots, \frac{1}{x_n}.$$

Следовательно,

$$\frac{1}{x_1} + \frac{2}{\rho_1}\cos\varphi + \frac{2}{\rho_2}\cos\psi + \frac{1}{x_6} + \ldots + \frac{1}{x_n} = -\frac{a_{n-1}}{a_n}, \qquad (15)$$

что вместе с уравнением

$$x_1 + 2\rho_1\cos\varphi + 2\rho_2\cos\psi + x_6 + \ldots + x_n = -\frac{a_1}{a_0} \qquad (16)$$

определит $\cos\varphi$ и $\cos\psi$. Для трех пар комплексных корней можно использовать наряду с исходным уравнением и уравнением, полученным заменой x на $1/y$ еще уравнение, полученное после первого квадрирования. Для случая трех и более пар комплексных корней, так же как и для случаев, которые мы уже рассмотрели, можно воспользоваться следующим приемом. Поскольку мы знаем модуль пары комплексно-сопряженных корней x_i и x_{i+1}, то в трехчлене $(x - x_i)(x - x_{i+1}) = x^2 - 2\rho x\cos\varphi + \rho^2$, соответствующем этой паре, нам будет неизвестно только $p = 2\rho\cos\varphi$. Но его можно найти из условия, что уравнение делится нацело на $x^2 - px + \rho^2$. Проще всего это сделать путем деления $f(x)$ на $x^2 - px + \rho^2$. Получится остаток вида $P(p)x + Q(p)$, где $P(p)$ — многочлен степени $n-1$ относительно p, а $Q(p)$ — многочлен степени $n-2$. Они должны обращаться в нуль при одном и том же значении p. Следовательно, они должны иметь общий делитель, который также можно найти путем деления.

Другой *способ* отыскания аргументов комплексных корней предложил *Энке*. Всегда можно считать, что уравнение имеет четную

степень, так как в противном случае его можно умножить на x. Запишем его в виде

$$x^{2k}+a_1x^{2k-1}+a_2x^{2k-2}+\ldots+a_{2k-1}x+a_{2k}=0, \tag{17}$$

и пусть модуль ρ пары комплексных корней найден, так что корни будут:

$$x_i=\rho(\cos\varphi+i\sin\varphi),\quad x_{i+1}=\rho(\cos\varphi-i\sin\varphi). \tag{18}$$

После подстановки этих корней в уравнение и приравнивания действительной и мнимой частей нулю получим уравнение относительно φ:

$$\left.\begin{aligned}\rho^{2k}\cos 2k\varphi+a_1\rho^{2k-1}\cos(2k-1)\varphi+\ldots+a_{2k-1}\rho\cos\varphi+a_{2k}=0,\\ \rho^{2k}\sin 2k\varphi+a_1\rho^{2k-1}\sin(2k-1)\varphi+\ldots+a_{2k-1}\rho\sin\varphi=0.\end{aligned}\right\} \tag{19}$$

Умножим первое из уравнений (19) на $\cos k\varphi$, второе на $\sin k\varphi$ и сложим их почленно. Затем умножим первое из уравнений (19) на $\sin k\varphi$, второе на $\cos k\varphi$ и вычтем первое из второго. В результате получим:

$$\left.\begin{aligned}&\beta_0\cos k\varphi+\frac{\beta_1}{\rho}\cos(k-1)\varphi+\frac{\beta_2}{\rho^2}\cos(k-2)\varphi+\ldots\\&\qquad\ldots+\frac{\beta_{k-1}}{\rho^{k-1}}\cos\varphi+\frac{\beta_k}{2\rho^k}=0,\\&\gamma_0\sin k\varphi+\frac{\gamma_1}{\rho}\sin(k-1)\varphi+\frac{\gamma_2}{\rho^2}\sin(k-2)\varphi+\ldots\\&\qquad\ldots+\frac{\gamma_{k-1}}{\rho^{k-1}}\sin\varphi=0,\end{aligned}\right\} \tag{20}$$

где

$$\left.\begin{aligned}1+a_{2k}\rho^{-2k}&=\beta_0, & 1-a_{2k}\rho^{-2k}&=\gamma_0,\\ a_1+a_{2k-1}\rho^{-2k+2}&=\beta_1, & a_1-a_{2k-1}\rho^{-2k+2}&=\gamma_1,\\ \ldots\ldots\ldots & & \ldots\ldots\ldots &\\ a_{k-1}+a_{k+1}\rho^{-2}&=\beta_{k-1}, & a_{k-1}-a_{k+1}\rho^{-2}&=\gamma_{k-1}.\\ 2a_k&=\beta_k, & &\end{aligned}\right\} \tag{21}$$

Для того чтобы найти значения, удовлетворяющие обоим уравнениям (20), используем выражения для $\cos n\varphi$ и $\frac{\sin n\varphi}{\sin\varphi}$ через $\cos\varphi$:

$$\cos n\varphi=(2\cos\varphi)^n-\frac{n}{1}(2\cos\varphi)^{n-2}+\frac{n(n-3)}{2!}(2\cos\varphi)^{n-4}-\\ -\frac{n(n-4)(n-5)}{3!}(2\cos\varphi)^{n-6}+\frac{n(n-5)(n-6)(n-7)}{4!}\times\\ \times(2\cos\varphi)^{n-8}-\ldots,$$

$$\frac{\sin n\varphi}{\sin\varphi} = (2\cos\varphi)^{n-1} - \frac{n-2}{1}(2\cos\varphi)^{n-3} + \frac{(n-3)(n-4)}{2!}\times$$

$$\times (2\cos\varphi)^{n-5} - \frac{(n-4)(n-5)(n-6)}{3!}(2\cos\varphi)^{n-7} +$$

$$+ \frac{(n-5)(n-6)(n-7)(n-8)}{4!}(2\cos\varphi)^{n-9} - \ldots$$

Полагая $-2\rho\cos\varphi = p$, получим следующие уравнения для определения p:

$$\left.\begin{aligned}
&\beta_0 p^k - \beta_1 p^{k-1} + \beta_2 p^{k-2} - \ldots + (-1)^k \beta_k - \rho^2 [k\beta_0 p^{k-2} - \\
&\quad - (k-1)\beta_1 p^{k-3} + (k-2)\beta_2 p^{k-4} - \ldots] + \\
&\quad + \rho^4 \left[\frac{k(k-3)}{2!}\beta_0 p^{k-4} - \frac{(k-1)(k-4)}{2!}\beta_1 p^{k-5} + \right. \\
&\left. + \frac{(k-2)(k-5)}{2!}\beta_2 p^{k-6} - \ldots\right] - \rho^6 \left[\frac{k(k-4)(k-5)}{3!}\beta_0 p^{k-6} - \right. \\
&\quad \left. - \frac{(k-1)(k-5)(k-6)}{3!}\beta_1 p^{k-7} + \ldots\right] + \rho^8 \times \\
&\quad \times \left[\frac{k(k-5)(k-6)(k-7)}{4!}\beta_0 p^{k-8} - \ldots\right] - \ldots = 0, \\
&\gamma_0 p^{k-1} - \gamma_1 p^{k-2} + \gamma_2 p^{k-3} - \ldots + (-1)^{k-1}\gamma_{k-1} - \rho^2 \times \\
&\times [(k-2)\gamma_0 p^{k-3} - (k-3)\gamma_1 p^{k-4} + (k-4)\gamma_2 p^{k-5} - \ldots] + \\
&+ \rho^4 \left[\frac{(k-3)(k-4)}{2!}\gamma_0 p^{k-5} - \frac{(k-4)(k-5)}{2!}\gamma_1 p^{k-6} + \ldots\right] - \\
&- \rho^6 \left[\frac{(k-4)(k-5)(k-6)}{3!}\gamma_0 p^{k-7} - \frac{(k-5)(k-6)(k-7)}{3!}\times\right. \\
&\left. \times \gamma_1 p^{k-8} + \ldots\right] + \rho^8 \left[\frac{(k-5)(k-6)(k-7)(k-8)}{4!}\times\right. \\
&\left. \times \gamma_0 p^{k-9} - \ldots\right] - \ldots = 0.
\end{aligned}\right\} \quad (22)$$

Общий корень p этих двух уравнений также находят простым делением.

Оба приведенных способа отыскания аргументов комплексных корней требуют громоздких вычислений.

Рассмотрим пример.

Решить методом Лобачевского уравнение

$$x^5 - 4x^4 + 6x^3 - 3x^2 + 2x + 1 = 0.$$

Схема вычислений и результаты выглядят следующим образом.

n		a_0	a_1	a_2	a_3	a_4	a_5
0		1	— 4	6	— 3	2	1
	a_i^2	1	16	36	9	4	1
	$-2a_{i-1}a_{i+1}$		— 12	— 24	— 24	6	
	$2a_{i-2}a_{i+2}$			4	— 8		
1		1	4	16	— 23	10	1
	a_i^2	1	16	256	529	100	1
	$-2a_{i-1}a_{i+1}$		— 32	184	— 320	46	
	$2a_{i-2}a_{i+2}$			20	8		
2		1	— 16	460	217	146	1
	a_i^2	1	256	211 600	47 089	21 316	1
	$-2a_{i-1}a_{i+1}$		— 920	6 944	— 134 320	— 434	1
	$2a_{i-2}a_{i+2}$			292	— 32		
3		1	— 664	218 836	— 87 263	20 882	1
	a_i^2	1	440 896	$4{,}788919 \cdot 10^{10}$	$7{,}614831 \cdot 10^{9}$	$4{,}360579 \cdot 10^{8}$	1
	$-2a_{i-1}a_{i+1}$		— 437 672	$-0{,}011588 \cdot 10^{10}$	$-9{,}139467 \cdot 10^{9}$	$0{,}001745 \cdot 10^{8}$	1
	$2a_{i-2}a_{i+2}$			$0{,}000004 \cdot 10^{10}$	$-0{,}000001 \cdot 10^{9}$		
4		1	3 224	$4{,}77734 \cdot 10^{10}$	$-1{,}52464 \cdot 10^{9}$	$4{,}36232 \cdot 10^{8}$	1

	a_i^2 $-2a_{i-1}a_{i+1}$ $2a_{i-2}a_{i+2}$	1	$0{,}001039\cdot10^{10}$ $-9{,}55468\cdot10^{10}$	$2{,}28230\cdot10^{21}$	$0{,}232453\cdot10^{19}$ $-4{,}168057\cdot10^{19}$	$1{,}90298\cdot10^{17}$	1
5		1	$-9{,}55364\cdot10^{10}$	$2{,}28230\cdot10^{21}$	$-3{,}93560\cdot10^{19}$	$1{,}90298\cdot10^{17}$	1
	a_i^2 $-2a_{i-1}a_{i+1}$ $2a_{i-2}a_{i+2}$	1	$9{,}12720\cdot10^{21}$ $-4{,}56460\cdot10^{21}$	$5{,}20889\cdot10^{42}$	$15{,}48895\cdot10^{38}$ $-8{,}68634\cdot10^{38}$	$3{,}62133\cdot10^{34}$	1
6		1	$4{,}56260\cdot10^{21}$	$5{,}20889\cdot10^{42}$	$6{,}80261\cdot10^{38}$	$3{,}62133\cdot10^{34}$	1
	a_i^2 $-2a_{i-1}a_{i+2}$ $2a_{i-2}a_{i+2}$	1	$2{,}08173\cdot10^{43}$ $-1{,}04178\cdot10^{43}$	$2{,}71325\cdot10^{85}$	$4{,}62755\cdot10^{77}$ $-3{,}77262\cdot10^{77}$	$1{,}31140\cdot10^{69}$	1
7		1	$1{,}03995\cdot10^{43}$	$2{,}71325\cdot10^{85}$	$0{,}85493\cdot10^{77}$	$1{,}31140\cdot10^{69}$	1
	a_i^2 $-2a_{i-1}a_{i+1}$ $2a_{i-2}a_{i+2}$	1	$1{,}08150\cdot10^{86}$ $-0{,}54265\cdot10^{86}$	$7{,}36173\cdot10^{170}$	$0{,}73090\cdot10^{154}$ $-7{,}11631\cdot10^{154}$	$1{,}71977\cdot10^{138}$	1
8		1	$0{,}53885\cdot10^{86}$	$7{,}36173\cdot10^{170}$	$-6{,}38541\cdot10^{154}$	$1{,}71977\cdot10^{138}$	1

Продолжение

n		a_0	a_1	a_2	a_3	a_4	a_5
	a_i^2 $-2a_{i-1}a_{i+1}$ $2a_{i-2}a_{i+2}$	1	$2{,}90359 \cdot 10^{171}$ $-1{,}47235 \cdot 10^{171}$	$5{,}41951 \cdot 10^{341}$	$4{,}07735 \cdot 10^{309}$ $-2{,}53210 \cdot 10^{309}$	$2{,}95761 \cdot 10^{276}$	1
9		1	$1{,}43124 \cdot 10^{171}$	$5{,}41951 \cdot 10^{341}$	$1{,}54525 \cdot 10^{309}$	$2{,}95761 \cdot 10^{276}$	1

$$x_5^{512} = \frac{1}{2{,}95761} \cdot 10^{-276}; \qquad 512 \lg |x_5| = -276{,}47004; \qquad \lg |x_5| = -0{,}53998; \qquad x_5 = -0{,}28841;$$

$$\rho_1^{1024} = 5{,}41951 \cdot 10^{341}; \qquad 1024 \lg \rho_1 = 341{,}73396; \qquad \lg \rho_1 = 0{,}33372; \qquad \rho_1 = 2{,}15637;$$

$$\rho_2^{1024} = \frac{2{,}95761}{5{,}41951} \cdot 10^{-65}; \qquad 1024 \lg \rho_2 = -65{,}263019; \qquad \lg \rho_2 = \bar{1}{,}93627; \qquad \rho_2 = 0{,}86351;$$

$$-0{,}28841 + 4{,}31274 \cos \varphi_1 + 1{,}72702 \cos \varphi_2 = 4;$$

$$\frac{-1}{0{,}28841} + \frac{2 \cos \varphi_1}{2{,}15637} + \frac{2 \cos \varphi_2}{0{,}86351} = -2;$$

$$\cos \varphi_1 = 0{,}88213; \qquad \cos \varphi_2 = 0{,}28027; \qquad \sin \varphi_1 = 0{,}47101; \qquad \sin \varphi_2 = 0{,}95992; \qquad x_{1,2} = 1{,}90219 \pm 1{,}01568i;$$

$$x_{3,4} = 0{,}24201 \pm 0{,}82890i.$$

3. Метод Лобачевского. Случай близких или равных корней. Если в уравнении имеется пара близких по абсолютной величине корней, то наш метод не дает возможности найти каждый из них в отдельности, так как если, например, $x_2 \approx x_3$, то в равенстве

$$(x_1x_2)^k + (x_1x_3)^k + \ldots + (x_{n-1}x_n)^k = \frac{c_2}{c_0}$$

два слагаемых будут близки друг к другу. Но тогда мы сумеем найти из равенства

$$(x_1x_2x_3)^k + \ldots = \frac{c_3}{c_0}$$

произведение x_2x_3, а дальше будем отыскивать трехчлен $x^2 - px + x_2x_3$, являющийся делителем $f(x)$, что позволит нам найти $p = x_2 + x_3$. На этот случай может быть перенесен также метод Энке.

Если имеется несколько пар равных по модулю комплексно-сопряженных корней или равных по модулю действительных корней, то может получиться, что процесс квадрирования не приведет к цели. Так, например, если заданное уравнение будет

$$x^4 + x^3 + x^2 + x + 1 = 0,$$

то все квадрированные уравнения имеют вид

$$x^4 - x^3 + x^2 - x + 1 = 0.$$

В этом случае целесообразно сделать замену неизвестного на $x - h$, что приведет к разделению корней с равными модулями.

Если корни уравнения необходимо найти с большой точностью, то после вычисления их приближенных значений по методу Лобачевского целесообразно произвести их уточнение, используя методы последовательных приближений, о которых мы будем говорить позже. Применение этих методов для уточнения требует меньшего объема вычислений и позволяет избежать трудности работы с очень большими числами, с которой приходится встречаться в методе Лобачевского.

4. Погрешность метода Лобачевского. Погрешности в значениях корней, полученных по методу Лобачевского, могут происходить по трем причинам:

1) в силу неточности коэффициентов исходного уравнения точные значения корней не могут быть найдены: это — *неустранимая погрешность*, не зависящая от способа получения корней, и мы на ней останавливаться не будем;

2) процесс квадрирования уменьшает величины отношений $\frac{x_i^k}{x_{i-1}^k}$, и чем большее количество раз проведен процесс квадрирования, тем точнее будут приближенные равенства для определения корней, но при любом конечном числе квадрирований эти равенства будут

оставаться приближенными и, следовательно, корни могут быть найдены только приближенно; погрешность в значениях корней, происходящая по этой причине, есть *погрешность* самого метода;

3) в процессе квадрирования получаются числа с очень большим количеством знаков и их приходится округлять, что вносит дополнительную погрешность в значения корней, — это *погрешность округления*. В дальнейшем мы и рассмотрим погрешность метода и погрешность округления.

Пусть дано уравнение

$$x^n + a_1 x^{n-1} + a_2 x^{n-2} + \ldots + a_{n-1} x + a_n = 0, \tag{23}$$

корни которого пока будем предполагать действительными, различными по абсолютной величине и упорядоченными следующим образом:

$$|x_1| > |x_2| > \ldots > |x_n|.$$

После s квадрирований мы получим новое уравнение:

$$x^n + b_1 x^{n-1} + b_2 x^{n-2} + \ldots + b_{n-1} x + b_n = 0, \tag{24}$$

причем, если положить $2^s = m$, то

$$b_k = \sum_{1 \leqslant i_1 < i_2 < \ldots < i_k \leqslant n} x_{i_1}^m x_{i_2}^m \ldots x_{i_k}^m. \tag{25}$$

Для сокращения записей примем обозначение $x_i^m = y_i$. Тогда

$$b_k = \sum_{1 \leqslant i_1 < i_2 < \ldots < i_k \leqslant n} y_{i_1} y_{i_2} \ldots y_{i_k}. \tag{26}$$

Рассмотрим отношение

$$\frac{b_k b_{k-2}}{b_{k-1}^2} =$$

$$= \frac{\sum\limits_{1 \leqslant i_1 < i_2 < \ldots < i_k \leqslant n} y_{i_1} y_{i_2} \ldots y_{i_k} \sum\limits_{1 \leqslant i_1 < i_2 < \ldots < i_{k-2} \leqslant n} y_{i_1} y_{i_2} \ldots y_{i_{k-2}}}{\left(\sum\limits_{1 \leqslant i_1 < i_2 < \ldots < i_{k-1} \leqslant n} y_{i_1} y_{i_2} \ldots y_{i_{k-1}}\right)^2}.$$

Вынесем из каждой суммы наибольшее слагаемое

$$\frac{b_k b_{k-2}}{b_{k-1}^2} = \frac{y_1 y_2 \ldots y_k \left[1 + \frac{y_{k+1}}{y_k} + \ldots\right] y_1 y_2 \ldots y_{k-2} \left[1 + \frac{y_{k-1}}{y_{k-2}} + \ldots\right]}{y_1^2 y_2^2 \ldots y_{k-1}^2 \left[1 + \frac{y_k}{y_{k-1}} + \ldots\right]^2} =$$

$$= \frac{y_k}{y_{k-1}} \frac{\left[1 + \frac{y_{k+1}}{y_k} + \ldots\right]\left[1 + \frac{y_{k-1}}{y_{k-2}} + \ldots\right]}{\left[1 + \frac{y_k}{y_{k-1}} + \ldots\right]^2},$$

где выписаны только наибольшие слагаемые. Так как $\frac{y_k}{y_{k-1}} = \left|\frac{x_k}{x_{k-1}}\right|^m$, то при достаточно большом m $\alpha_k = \frac{b_k b_{k-2}}{b_{k-1}^2}$ — достаточно малая величина. Далее, заменяя в числителе все отношения нулями, а в знаменателе наибольшей величиной $\frac{y_k}{y_{k-1}}$, получим неравенство

$$\alpha_k = \frac{b_k b_{k-2}}{b_{k-1}^2} \geqslant \frac{y_k}{y_{k-1}} \left[1 + (C_n^{k-1} - 1)\frac{y_k}{y_{k-1}}\right]^{-2},$$

откуда

$$\frac{y_k}{y_{k-1}} \leqslant \alpha_k \left[1 + (C_n^{k-1} - 1)\frac{y_k}{y_{k-1}}\right]^2 = \alpha_k \left[1 + \beta_k \frac{y_k}{y_{k-1}}\right]^2$$
$$(\beta_k = C_n^{k-1} - 1). \tag{27}$$

Если m настолько велико, что $4\alpha_k\beta_k < 1$, то из неравенства (27) следует:

$$\frac{y_k}{y_{k-1}} \leqslant 2\alpha_k\gamma_k, \text{ где } \gamma_k = \frac{1}{1 - 2\alpha_k\beta_k + (1 - 4\alpha_k\beta_k)^{\frac{1}{2}}}\ ^{1}). \tag{28}$$

Более простой, но несколько более грубой оценкой будет

$$\frac{y_k}{y_{k-1}} \leqslant \frac{2\alpha_k}{1 - 2\alpha_k\beta_k}. \tag{29}$$

Рассмотрим теперь отношение

$$\frac{b_k}{b_{k-1}} = \frac{\sum\limits_{1 \leqslant i_1 < i_2 < \ldots < i_k \leqslant n} y_{i_1} y_{i_2} \cdots y_{i_k}}{\sum\limits_{1 \leqslant i_1 < i_2 < \ldots < i_{k-1} \leqslant n} y_{i_1} y_{i_2} \cdots y_{i_{k-1}}} = y_k \frac{1 + \frac{y_{k+1}}{y_k} + \ldots}{1 + \frac{y_k}{y_{k-1}} + \ldots}.$$

Заменяя в числителе все слагаемые, кроме первого, нулями, а в знаменателе наибольшим значением $\left(\frac{y_k}{y_{k-1}}\right)$, мы уменьшим правую часть, а заменив все отношения в числителе наибольшим отношением $\frac{y_{k+1}}{y_k}$,

---

1) Напомним, что $\frac{y_k}{y_{k-1}} < 1$, поэтому $\frac{y_k}{y_{k-1}} \geqslant \frac{2\alpha_k}{1 - 2\alpha_k\beta_k - (1 - 4\alpha_k\beta_k)^{\frac{1}{2}}}$ при достаточно больших m невозможно, так как $\alpha_k \to 0$ при $m \to +\infty$, а β_k не зависит от m, и следовательно $\frac{2\alpha_k}{1 - 2\alpha_k\beta_k - (1 - 4\alpha_k\beta_k)^{\frac{1}{2}}} \to +\infty$ при $m \to \infty$.

а все отношения в знаменателе нулями, мы увеличим правую часть. Следовательно,

$$y_k\left[1+\frac{y_k}{y_{k-1}}\left(C_n^{k-1}-1\right)\right]^{-1}\leqslant\frac{b_k}{b_{k-1}}\leqslant y_k\left[1+\frac{y_{k+1}}{y_k}\left(C_n^k-1\right)\right]$$

или

$$y_k\left[1+\beta_k\frac{y_k}{y_{k-1}}\right]^{-1}\leqslant\frac{b_k}{b_{k-1}}\leqslant y_k\left[1+\beta_{k+1}\frac{y_{k+1}}{y_k}\right],$$

откуда

$$\frac{b_k}{b_{k-1}}\left[1+\beta_{k+1}\frac{y_{k+1}}{y_k}\right]^{-1}\leqslant y_k\leqslant\frac{b_k}{b_{k-1}}\left[1+\beta_k\frac{y_k}{y_{k-1}}\right]. \qquad (30)$$

Используя неравенство (29), получим:

$$\frac{b_k}{b_{k-1}}\left[1+\frac{2\alpha_{k+1}\beta_{k+1}}{1-2\alpha_{k+1}\beta_{k+1}}\right]^{-1}\leqslant y_k\leqslant\frac{b_k}{b_{k-1}}\left[1+\frac{2\alpha_k\beta_k}{1-2\alpha_k\beta_k}\right]$$

или

$$\frac{b_k}{b_{k-1}}\left(1-2\alpha_{k+1}\beta_{k+1}\right)\leqslant y_k\leqslant\frac{b_k}{b_{k-1}}\left(1-2\alpha_k\beta_k\right)^{-1}. \qquad (31)$$

Так как $y_k=x_k^m$, то

$$\left(\frac{b_k}{b_{k-1}}\right)^{\frac{1}{m}}\left(1-2\alpha_{k+1}\beta_{k+1}\right)^{\frac{1}{m}}\leqslant|x_k|\leqslant\left(\frac{b_k}{b_{k-1}}\right)^{\frac{1}{m}}\left(1-2\alpha_k\beta_k\right)^{-\frac{1}{m}}. \qquad (32)$$

Если x_k^* — приближенное значение x_k, полученное после m квадрирований, то $x_k^{*m}=\frac{b_k}{b_{k-1}}$. Следовательно, (32) можно записать в таком виде:

$$|x_k^*|\left(1-2\alpha_{k+1}\beta_{k+1}\right)^{\frac{1}{m}}\leqslant|x_k|\leqslant|x_k^*|\left(1-2\alpha_k\beta_k\right)^{-\frac{1}{m}}. \qquad (33)$$

Оценка (33) неудобна тем, что с помощью ее нельзя заранее предсказать число квадрирований, необходимых для получения нужной точности. Этот недостаток можно исправить, наложив новые ограничения на уравнение.

Допустим, что существует такое число $0<r<1$, что $|x_{i+1}|\leqslant r|x_i|$ $(i=1, 2, \ldots, n-1)$. Тогда $|x_j|\leqslant r^{j-i}|x_i|$ при $j>i$. Положив $t=r^m$, будем иметь $y_j\leqslant t^{j-i}y_i$. Пусть $y_{i_1}y_{i_2}\ldots y_{i_k}$ — некоторое слагаемое в сумме $b_k=\sum\limits_{1\leqslant i_1<i_2<\ldots<i_k\leqslant n}y_{i_1}y_{i_2}\ldots y_{i_k}$. Тогда

$$\frac{y_{i_1}y_{i_2}\ldots y_{i_k}}{y_1y_2\ldots y_k}\leqslant t^{(i_1-1)+(i_2-2)+\ldots+(i_k-k)}=t^{i_1+i_2+\ldots+i_k-\frac{k(k+1)}{2}},$$

а следовательно,

$$b_k \leqslant y_1 y_2 \dots y_k \sum_{1 \leqslant i_1 < i_2 < \dots < i_k \leqslant n} t^{i_1+i_2+\dots+i_k-\frac{k(k+1)}{2}} = \\ = y_1 y_2 \dots y_k [1 + Q_k(t)]. \tag{34}$$

Число слагаемых в сумме, стоящей в правой части неравенства (34), имеющих одну и ту же степень p, равно числу целочисленных решений уравнения

$$i_1 + i_2 + \dots + i_k = p + \frac{k(k+1)}{2},$$

удовлетворяющих условию $1 \leqslant i_1 < i_2 < \dots < i_k \leqslant n$.

Так как коэффициенты $Q_k(t)$ — целые положительные числа, то

$$y_1 y_2 \dots y_k \leqslant b_k \leqslant y_1 y_2 \dots y_k [1 + Q_k(t)]. \tag{35}$$

Аналогично

$$y_1 y_2 \dots y_{k-1} \leqslant b_{k-1} \leqslant y_1 y_2 \dots y_{k-1} [1 + Q_{k-1}(t)],$$

откуда

$$y_k [1 + Q_{k-1}(t)]^{-1} \leqslant \frac{b_k}{b_{k-1}} \leqslant y_k [1 + Q_k(t)].$$

Следовательно,

$$\frac{b_k}{b_{k-1}} [1 + Q_k(t)]^{-1} \leqslant y_k \leqslant \frac{b_k}{b_{k-1}} [1 + Q_{k-1}(t)] \tag{36}$$

или

$$\left(\frac{b_k}{b_{k-1}}\right)^{\frac{1}{m}} [1 + Q_k(t)]^{-\frac{1}{m}} \leqslant |x_k| \leqslant \left(\frac{b_k}{b_{k-1}}\right)^{\frac{1}{m}} [1 + Q_{k-1}(t)]^{\frac{1}{m}}, \tag{37}$$

или

$$[1 + Q_k(t)]^{-\frac{1}{m}} - 1 \leqslant \frac{|x_k| - |x_k^*|}{|x_k^*|} \leqslant [1 + Q_{k-1}(t)]^{\frac{1}{m}} - 1. \tag{38}$$

Таким образом, относительная погрешность зависит только от t, n, m и k. Если известно r или хотя бы его нижняя граница, то можно использовать эту оценку для определения числа квадрирований s ($2^s = m$), необходимых для получения корней с нужной относительной погрешностью.

Найдем теперь величину $Q_k(t)$. Известно, что число целочисленных решений уравнения

$$i_1 + i_2 + \dots + i_k = p + \frac{k(k+1)}{2} \qquad (1 \leqslant i_1 < i_2 < \dots < i_k \leqslant n)$$

равно коэффициенту при z^p в многочлене

$$P_{n,k}(z) = \frac{(1-z^n)(1-z^{n-1}) \dots (1-z^{n-k+1})}{(1-z)(1-z^2) \dots (1-z^k)} \tag{39}$$

Таким образом,

$$1+Q_k(t)=P_{n,k}(t).$$

Вычисление значений $P_{n,k}(t)$ для всех $k=1, 2, \ldots, n$ требует большой вычислительной работы, поэтому целесообразно найти k, при котором $P_{n,k}(t)$, где t — фиксированное положительное число, не равное единице, имеет максимальное значение. Так как

$$P_{n,k+1}(t)=P_{n,k}(t)\frac{1-t^{n-k}}{1-t^{k+1}},$$

а

$$\frac{1-t^{n-k}}{1-t^{k+1}}=\frac{1+t+t^2+\ldots+t^{n-k-1}}{1+t+t^2+\ldots+t^k}=\lambda,$$

где $\lambda\geqslant 1$, если $n-k-1\geqslant k$, и $\lambda<1$; если $n-k-1<k$, то

$$P_{n,k+1}(t)\geqslant P_{n,k}(t) \quad \text{при} \quad k\leqslant\frac{n-1}{2},$$

$$P_{n,k+1}(t)<P_{n,k}(t) \quad \text{при} \quad k>\frac{n-1}{2}.$$

Это означает, что

$$\max_{k=1,2,\ldots,n} P_{n,k}(t)=P_{n,\left[\frac{n}{2}\right]}(t). \tag{40}$$

Если в (36) и (37) заменить $1+Q_k(t)$ и $1+Q_{k-1}(t)$ через $P_{n,\left[\frac{n}{2}\right]}(t)$, то получим оценки, пригодные для всех корней.

В качестве иллюстрации допустим, что нам дано $n=10$, $r=0{,}9$ и произведено четыре квадрирования. Тогда $m=2^4=16$, $t=(0{,}9)^{16}=$ $=0{,}01853$, $k=\left[\frac{n}{2}\right]=5$.

$$1+Q_k(t)=\frac{(1-t^6)(1-t^7)\ldots(1-t^{10})}{(1-t)(1-t^2)\ldots(1-t^5)}=$$

$$=(1+t^5)(1+t^4)(1+t^3+t^6)(1+t^2+t^4)(1+t+t^2+\ldots+t^6).$$

Простые вычисления показывают, что

$$1+Q_k(t)\leqslant 1{,}01922, \quad \text{а} \quad [1+Q_k(t)]^{\frac{1}{16}}\leqslant 1{,}0012.$$

Отсюда

$$-0{,}0012<\frac{1}{1{,}0012}-1\leqslant\frac{|x_k|-|x_k^*|}{|x_k^*|}\leqslant 1{,}0012-1=0{,}0012.$$

Таким образом, относительная ошибка корней после четырех квадрирований не будет превышать 0,0012.

Допустим теперь, что уравнение имеет комплексные корни, но такие, что соотношение $|x_k|\leqslant r|x_{k-1}|$ сохраняется для корней с разными модулями, и если x_k — действительный корень, а x_{k+1}

и x_{k+2} — пара комплексно-сопряженных корней, то

$$|x_{k+2}| \leqslant r^2 |x_k|.$$

Тогда будут применимы все предыдущие рассуждения лишь с очень небольшими изменениями. Вместо неравенства (35) будет справедливо неравенство

$$|y_1||y_2| \ldots |y_k| \leqslant |b_k| \leqslant |y_1||y_2| \ldots |y_k|[1 + Q_k(t)]. \quad (41)$$

При $Q_k(t) < 1$ левая часть этого неравенства может быть заменена на $|b_k| \geqslant |y_1||y_2| \ldots |y_k|[1 - Q_k(t)]$, и следовательно,

$$|y_1||y_2| \ldots |y_k|[1 - Q_k(t)] \leqslant |b_k| \leqslant |y_1||y_2| \ldots |y_k|[1 + Q_k(t)]. \quad (42)$$

Заменив $Q_k(t)$ максимальной величиной $P_{n, \left[\frac{n}{2}\right]}(t) - 1 = Q(t)$, получим:

$$\left|\frac{b_k}{b_{k-1}}\right|^{\frac{1}{m}} \left[\frac{1+Q(t)}{1-Q(t)}\right]^{-\frac{1}{m}} \leqslant |x_k| \leqslant \left|\frac{b_k}{b_{k-1}}\right|^{\frac{1}{m}} \left[\frac{1+Q(t)}{1-Q(t)}\right]^{\frac{1}{m}}. \quad (43)$$

Оценим теперь скорость сходимости процесса квадрирования. Пусть мы провели дополнительное квадрирование. Тогда

$$\left|\frac{b'_k}{b'_{k-1}}\right|^{\frac{1}{2m}} \left[\frac{1+Q(t^2)}{1-Q(t^2)}\right]^{-\frac{1}{2m}} \leqslant |x_k| \leqslant \left|\frac{b'_k}{b'_{k-1}}\right|^{\frac{1}{2m}} \left[\frac{1+Q(t^2)}{1-Q(t^2)}\right]^{\frac{1}{2m}}, \quad (44)$$

где b'_k — коэффициенты нового уравнения. Допустим, что проведено достаточное число квадрирований, так что t настолько мало, что можно пренебречь его высшими степенями. При этом

$$\left[\frac{1+Q(t)}{1-Q(t)}\right]^{\frac{1}{m}} \approx \left[\frac{1+ct}{1-ct}\right]^{\frac{1}{m}} \approx 1 + \frac{2ct}{m},$$

в то время как

$$\left[\frac{1+Q(t^2)}{1-Q(t^2)}\right]^{\frac{1}{2m}} \approx \left[\frac{1+ct^2}{1-ct^2}\right]^{\frac{1}{2m}} \approx 1 + \frac{ct^2}{m},$$

где c коэффициент при t в $Q(t)$. Это показывает, что относительная погрешность будет убывать в отношении $t/2$ при выполнении одного квадрирования.

Все предыдущие рассуждения проводились в предположении, что при реализации метода Лобачевского все вычисления проводились точно. На практике же приходится ограничиваться лишь конечным числом разрядов, т. е. проводить округление результатов операций. Поэтому даже если коэффициенты исходного уравнения были точными числами, то после некоторого числа квадрирований мы получим уравнение с приближенными коэффициентами и эта погрешность в дальнейшем будет сказываться на коэффициентах, а следовательно и на корнях уравнений, получающихся при последующих

квадрированиях. Дать какие-либо точные оценки влияния этих ошибок округления в коэффициентах квадрированных уравнений очень трудно. Поэтому мы приведем лишь очень грубые рассуждения, которые будут указывать на опасность резкой потери точности при применении метода Лобачевского в том или ином конкретном случае.

Рассмотрим сначала случай уравнения с различными по абсолютной величине действительными корнями. Пусть после s квадрирований мы получили уравнение

$$y^n + b_1 y^{n-1} + b_2 y^{n-2} + \ldots + b_{n-1} y + b_n = 0,$$

корни которого $y_i = -x_i^{2^s}$. Предположим, что b_k имеют m верных знаков. Выполняя еще одно квадрирование, получим уравнение

$$z^n + b_1' z^{n-1} + b_2' z^{n-2} + \ldots + b_{n-1}' z + b_n' = 0,$$

корнями которого будут числа $-y_1^2, -y_2^2, \ldots, -y_n^2$. Так как

$$b_k = \sum_{1 \leqslant i_1 < i_2 < \ldots < i_k \leqslant n} y_{i_1} y_{i_2} \ldots y_{i_k},$$

а

$$b_k' = \sum_{1 \leqslant i_1 < i_2 < \ldots < i_k \leqslant n} y_{i_1}^2 y_{i_2}^2 \ldots y_{i_k}^2,$$

то $b_k' < b_k^2$. С другой стороны, по самому способу построения квадрированного уравнения

$$b_k' = b_k^2 - 2b_{k-1}b_{k+1} + 2b_{k-2}b_{k+2} - 2b_{k-3}b_{k+3} + \ldots$$

Можно считать, что b_k^2 и $b_{k-j}b_{k+j}$ имеют по m верных знаков. Если b_k' имеет порядок b_k^2, то это будет означать (при небольших n), что b_k' будут также иметь m верных знаков. Но если b_k' значительно меньше, чем b_k^2, что будет в том случае, когда из b_k^2 вычитается сумма $2(b_{k-1}b_{k+1} - b_{k-2}b_{k+2} + \ldots)$, близкая к b_k^2, то будем иметь резкую потерю верных знаков в b_k'. Если $\dfrac{b_k^2}{b_k'}$ имеет порядок 10^h, то в b_k' мы теряем примерно h верных знаков. Но

$$\frac{b_k^2}{b_k'} = \frac{y_1^2 y_2^2 \ldots y_k^2 \left[\sum\limits_{1 \leqslant i_1 < i_2 < \ldots < i_k \leqslant n} \dfrac{y_{i_1} y_{i_2} \ldots y_{i_k}}{y_1 y_2 \ldots y_k}\right]^2}{\sum\limits_{1 \leqslant i_1 < i_2 < \ldots < i_k \leqslant n} \dfrac{y_{i_1}^2 y_{i_2}^2 \ldots y_{i_k}^2}{y_1^2 y_2^2 \ldots y_k^2}} \leqslant [1 + Q_k(t)]^2,$$

так как $\sum\limits_{1 \leqslant i_1 < i_2 < \ldots < i_k \leqslant n} \dfrac{y_{i_1}^2 y_{i_2}^2 \ldots y_{i_k}^2}{y_1^2 y_2^2 \ldots y_k^2} > 1$. Таким образом, по величине $[1 + Q_k(t)]^2$ можно судить о потере значащих цифр в b_k'.

Если рассмотреть крайний случай, когда все корни уравнения равны, т. е. $y_1 = y_2 = \ldots = y_n = y$, то $b_k^2 = (C_n^k y)^2$, а $b_k' = C_n^k y^2$. Отсюда $\frac{b_k^2}{b_k'} = C_n^k$, и так как C_n^k может быть велико, то мы будем иметь большую потерю точности. Например, если $n = 10$, а $k = 5$, то $\frac{b_5^2}{b_5'} = C_{10}^5 = 252$ и, следовательно, можно ожидать потери двух знаков при каждом квадрировании. В случае близких по модулю корней мы тоже будем иметь большую потерю точности и корни по методу Лобачевского будут найдены с большой относительной погрешностью. Точность будет тем лучше, чем лучше разделены корни.

В случае наличия комплексных корней, но при ограничениях на модули корней, при которых имеет место неравенство (42), заменяя $Q_k(t)$ через $Q(t) = Q_{\left[\frac{n}{2}\right]}(t)$, получим:

$$\frac{|b_k|^2}{|b_k'|} \leqslant \frac{[1 + Q(t)]^2}{1 - Q(t^2)},$$

т. е. и здесь по величине отношения $\frac{[1 + Q(t)]^2}{1 - Q(t^2)}$ можно судить о потере точности в результате округлений.

5. Видоизменение Лемера метода Лобачевского. Наряду с исходным многочленом

$$P_0(x) = a_0^{(0)} x^n + a_1^{(0)} x^{n-1} + \ldots + a_{n-1}^{(0)} x + a_n^{(0)} \tag{45}$$

рассмотрим многочлен

$$\begin{aligned} Q_0(x) = P_0(x - h) = P_0(x) - hP_0'(x) + \ldots = \\ = \left[a_0^{(0)} x^n + a_1^{(0)} x^{n-1} + \ldots + a_{n-1}^{(0)} x + a_n^{(0)}\right] + \\ + h\left[b_1^{(0)} x^{n-1} + b_2^{(0)} x^{n-2} + \ldots + b_{n-1}^{(0)} x + b_n^{(0)}\right] + \ldots, \end{aligned} \tag{46}$$

где h — произвольное действительное или комплексное число. В равенстве (46), как и всюду в дальнейшем, не выписываются явно члены с h^2 и более высокими степенями h. Очевидно,

$$b_i^{(0)} = -(n - i + 1)\, a_{i-1}^{(0)} \qquad (i = 1, 2, \ldots, n). \tag{47}$$

В целях упрощения записи формул мы условимся в дальнейшем считать все $a_i^{(k)}$, $b_i^{(k)}$, где i отрицательно или больше n, а также $b_0^{(k)}$ равными нулю. Будем квадрировать многочлен $Q_0(x)$. После

первого квадрирования получим многочлен $Q_1(x)$, коэффициенты которого имеют вид

$$\left(a_i^{(0)}+hb_i^{(0)}+\ldots\right)^2+2\sum_k(-1)^k\left(a_{i-k}^{(0)}+hb_{i-k}^{(0)}+\ldots\right)\times$$

$$\times\left(a_{i+k}^{(0)}+hb_{i+k}^{(0)}+\ldots\right)=\left(a_i^{(0)}+hb_i^{(0)}+\ldots\right)^2+$$

$$+2\sum_{j=0}^{i-1}(-1)^{i+j}\left(a_j^{(0)}+hb_j^{(0)}+\ldots\right)\left(a_{2i-j}^{(0)}+hb_{2i-j}^{(0)}+\ldots\right)=$$

$$=\left[a_i^{(0)\,2}+2\sum_{j=0}^{i-1}(-1)^{i+j}a_j^{(0)}a_{2i-j}^{(0)}\right]+$$

$$+2h\left[a_i^{(0)}b_i^{(0)}+\sum_{j=0}^{i-1}(-1)^{i+j}\left(a_j^{(0)}b_{2i-j}^{(0)}+a_{2i-j}^{(0)}b_j^{(0)}\right)\right]+\ldots$$

Таким образом, $Q_1(x)$ примет вид

$$Q_1(x)=\left[a_0^{(1)}x^n+a_1^{(1)}x^{n-1}+\ldots+a_{n-1}^{(1)}x+a_n^{(1)}\right]+$$

$$+2h\left[b_1^{(1)}x^{n-1}+b_2^{(1)}x^{n-2}+\ldots+b_{n-1}^{(1)}x+b_n^{(1)}\right]+\ldots, \quad (48)$$

где

$$a_i^{(1)}=a_i^{(0)\,2}+2\sum_{j=0}^{i-1}(-1)^{i+j}a_j^{(0)}a_{2i-j}^{(0)},$$

$$b_i^{(1)}=a_i^{(0)}b_i^{(0)}+\sum_{j=0}^{i-1}(-1)^{i+j}\left(a_j^{(0)}b_{2i-j}^{(0)}+a_{2i-j}^{(0)}b_j^{(0)}\right)=$$

$$=\sum_{j=0}^{2i-1}(-1)^{i+j}a_j^{(0)}b_{2i-j}^{(0)}. \quad (49)$$

При последующих квадрированиях мы будем получать многочлены $Q_2(x)$, $Q_3(x)$, ..., коэффициенты которых находятся последовательным применением формул (49), т. е.

$$Q_k(x)=\left[a_0^{(k)}x^n+a_1^{(k)}x^{n-1}+\ldots+a_{n-1}^{(k)}x+a_n^{(k)}\right]+$$

$$+2^kh\left[b_1^{(k)}x^{n-1}+b_2^{(k)}x^{n-2}+\ldots+b_{n-1}^{(k)}x+b_n^{(k)}\right]+\ldots, \quad (50)$$

где

$$\left.\begin{aligned}a_i^{(k)}&=a_i^{(k-1)\,2}+2\sum_{j=0}^{i-1}(-1)^{i+j}a_j^{(k-1)}a_{2i-j}^{(k-1)},\\ b_i^{(k)}&=\sum_{j=0}^{2i-1}(-1)^{i+j}a_j^{(k-1)}b_{2i-j}^{(k-1)}.\end{aligned}\right\} \quad (51)$$

Обозначим корни уравнения $P_0(x)=0$ через $x_1, x_2, \ldots, x_n$, причем будем предполагать, что

$$|x_1|\geqslant|x_2|\geqslant\ldots\geqslant|x_n|.$$

Корнями уравнения $Q_0(x)=0$ будут: x_1+h, x_2+h, ..., x_n+h. Если обозначить 2^k через m, то корнями уравнения $Q_k(x)=0$ будут $-(x_1+h)^m$, $-(x_2+h)^m$, ..., $-(x_n+h)^m$. Не уменьшая общности, можно предположить, что $a_0^{(0)}=1$, тогда и все $a_0^{(k)}=1$. Так как

$$a_i^{(k)}+2^k h b_i^{(k)}+\ldots=(x_1+h)^m(x_2+h)^m\ldots(x_i+h)^m+$$

$$+(x_1+h)^m(x_2+h)^m\ldots(x_{i-1}+h)^m(x_{i+1}+h)^m+\ldots$$

$$\ldots+(x_{n-i+1}+h)^m(x_{n-i+2}+h)^m\ldots(x_n+h)^m=x_1^m x_2^m\ldots x_i^m+$$

$$+x_1^m x_2^m\ldots x_{i-1}^m x_{i+1}^m+\ldots+x_{n-i+1}^m x_{n-i+2}^m\ldots x_n^m+$$

$$+2^k h\left[x_1^{m-1}x_2^m\ldots x_i^m+x_1^m x_2^{m-1}x_3^m\ldots x_i^m+\ldots\right.$$

$$\left.\ldots+x_{n-i+1}^m x_{n-i+2}^m\ldots x_{n-1}^m x_n^{m-1}\right]+\ldots$$

и h — произвольное число, то

$$a_i^{(k)}=x_1^m x_2^m\ldots x_i^m+x_1^m x_2^m\ldots x_{i-1}^m x_{i+1}^m+\ldots$$

$$\ldots+x_{n-i+1}^m x_{n-i+2}^m\ldots x_n^m,$$

$$b_i^{(k)}=x_1^{m-1}x_2^m\ldots x_i^m+x_1^m x_2^{m-1}x_3^m\ldots x_i^m+\ldots$$

$$\ldots+x_{n-i+1}^m x_{n-i+2}^m\ldots x_{n-1}^m x_n^{m-1}.$$

Полагая здесь $i=1, 2, \ldots, n$, получим:

$$\left.\begin{aligned}
a_1^{(k)}&=x_1^m+x_2^m+x_3^m+\ldots+x_n^m,\\
a_2^{(k)}&=x_1^m x_2^m+x_1^m x_3^m+\ldots+x_{n-1}^m x_n^m,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
a_n^{(k)}&=x_1^m x_2^m\ldots x_n^m,
\end{aligned}\right\}\qquad(52)$$

$$\left.\begin{aligned}
b_1^{(k)}&=x_1^{m-1}+x_2^{m-1}+\ldots+x_n^{m-1},\\
b_2^{(k)}&=x_1^{m-1}x_2^m+x_1^m x_3^{m-1}+x_1^{m-1}x_3^m+\ldots+x_{n-1}^{m-1}x_n^m+\\
&\qquad+x_{n-1}^m x_n^{m-1},\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
b_n^{(k)}&=x_1^{m-1}x_2^m\ldots x_n^m+x_1^m x_2^{m-1}x_3^m\ldots x_n^m+\ldots\\
&\qquad\ldots+x_1^m x_2^m\ldots x_{n-1}^m x_n^{m-1}.
\end{aligned}\right\}\qquad(53)$$

Предположим сначала, что

$$|x_1| > |x_2| > \ldots > |x_n|.$$

Тогда

$$\left.\begin{aligned}
&a_1^{(k)} \approx x_1^m, && b_1^{(k)} \approx x_1^{m-1},\\
&a_2^{(k)} \approx x_1^m x_2^m, && b_2^{(k)} \approx x_1^{m-1} x_2^m + x_1^m x_2^{m-1},\\
&a_3^{(k)} \approx x_1^m x_2^m x_3^m, && b_3^{(k)} \approx x_1^{m-1} x_2^m x_3^m + x_1^m x_2^{m-1} x_3^m + \\
& && \qquad + x_1^m x_2^m x_3^{m-1},\\
&\ldots\ldots\ldots && \ldots\ldots\ldots\ldots\ldots\ldots\\
&a_n^{(k)} \approx x_1^m x_2^m \ldots x_n^m, && b_n^{(k)} \approx x_1^{m-1} x_2^m x_3^m \ldots x_n^m + \\
& && \qquad + x_1^m x_2^m x_3^m \ldots x_n^m + \ldots \\
& && \qquad \ldots + x_1^m x_2^m \ldots x_{n-1}^m x_n^{m-1}.
\end{aligned}\right\} \qquad (54)$$

Полагая $c_i^{(k)} = \dfrac{b_i^{(k)}}{a_i^{(k)}}$, получим:

$$\begin{aligned}
c_1^{(k)} &\approx \frac{1}{x_1},\\
c_2^{(k)} &\approx \frac{1}{x_1} + \frac{1}{x_2},\\
c_3^{(k)} &\approx \frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3},\\
&\ldots\ldots\ldots\ldots\\
c_n^{(k)} &\approx \frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}
\end{aligned}$$

или

$$\left.\begin{aligned}
x_1 &\approx \frac{1}{c_1^{(k)}},\\
x_2 &\approx \frac{1}{c_2^{(k)} - c_1^{(k)}},\\
x_3 &\approx \frac{1}{c_3^{(k)} - c_2^{(k)}},\\
&\ldots\ldots\\
x_n &\approx \frac{1}{c_n^{(k)} - c_{n-1}^{(k)}}.
\end{aligned}\right\} \qquad (55)$$

Если корни действительны, но

$$|x_1| = |x_2| = \ldots = |x_r| > |x_{r+1}| > \ldots > |x_n|,$$

то из (52) и (53) имеем:

$$\left.\begin{aligned}
&a_1^{(k)} \approx r x_1^m, && b_1^{(k)} \approx r x_1^{m-1},\\
&a_2^{(k)} \approx C_r^2 x_1^{2m}, && b_2^{(k)} \approx 2C_r^2 x_1^{2m-1},\\
&\dots\dots && \dots\dots\dots\\
&a_r^{(k)} \approx C_r^r x_1^{rm}, && b_r^{(k)} \approx r C_r^r x_1^{rm-1},\\
&a_{r+1}^{(k)} \approx x_1^{rm} x_{r+1}^m, && b_{r+1}^{(k)} \approx x_1^{rm} x_{r+1}^{m-1} + r x_1^{rm-1} x_{r+1}^m,\\
&\dots\dots && \dots\dots\dots\dots\dots\dots\\
&a_n^{(k)} \approx x_1^{rm} x_{r+1}^m \dots x_n^m, && b_n^{(k)} \approx x_1^{rm}\left(x_{r+1}^{m-1} x_{r+2}^m \dots x_n^m + \dots\right.\\
& && \left. \dots + x_{r+1}^m \dots x_{n-1}^m x_n^{m-1}\right) +\\
& && + r x_1^{rm-1} x_{r+1}^m \dots x_n^m,
\end{aligned}\right\} \qquad (56)$$

откуда

$$c_1^{(k)} \approx \frac{1}{x_1}, \quad c_2^{(k)} \approx \frac{2}{x_1}, \quad \dots, \quad c_r^{(k)} \approx \frac{r}{x_1}, \quad c_{r+1}^{(k)} \approx \frac{1}{x_{r+1}} + \frac{r}{x_1},$$

$$c_{r+2}^{(k)} \approx \frac{1}{x_{r+1}} + \frac{1}{x_{r+2}} + \frac{r}{x_1}, \quad \dots, \quad c_n^{(k)} \approx \frac{1}{x_{r+1}} + \frac{1}{x_{r+2}} + \dots + \frac{1}{x_n} + \frac{r}{x_1}.$$

Следовательно,

$$\left.\begin{aligned}
&x_1 \approx \frac{1}{c_1^{(k)}}, \quad x_{r+1} \approx \frac{1}{c_{r+1}^{(k)} - c_r^{(k)}},\\
&x_{r+2} \approx \frac{1}{c_{r+2}^{(k)} - c_{r+1}^{(k)}}, \quad \dots, \quad x_n \approx \frac{1}{c_n^{(k)} - c_{n-1}^{(k)}}.
\end{aligned}\right\} \qquad (57)$$

Рассмотрим теперь случай, когда имеется пара комплексно-сопряженных корней. Пусть это будут корни x_1, x_2:

$$x_1 = \rho(\cos\varphi + i\sin\varphi), \quad x_2 = \rho(\cos\varphi - i\sin\varphi).$$

Относительно остальных корней предположим, что они удовлетворяют условию

$$|x_1| = |x_2| > |x_3| > \dots > |x_n|.$$

Тогда из (52) и (53) следует, что

$$a_1^{(k)} = 2\rho^m \cos m\varphi + x_3^m + \ldots, \qquad b_1^{(k)} = 2\rho^{m-1}\cos(m-1)\varphi + x_3^{m-1} + \ldots,$$

$$a_2^{(k)} = \rho^{2m} + \rho^m x_3^m \cos m\varphi + \ldots, \qquad b_2^{(k)} = 2\rho^{2m-1}\cos\varphi + 2\rho^{m-1}x_3^m \cos(m-1)\varphi + \ldots,$$

$$a_3^{(k)} = \rho^{2m} x_3^m + \ldots, \qquad b_3^{(k)} = \rho^{2m} x_3^{m-1} + 2\rho^{2m-1} x_3^m \cos\varphi + \ldots,$$

$$\ldots\ldots\ldots \qquad \ldots\ldots\ldots\ldots\ldots$$

$$a_n^{(k)} = \rho^{2m} x_3^m \ldots x_n^m, \qquad b_n^{(k)} = 2\rho^{2m-1} x_3^m \ldots x_n^m \cos\varphi + \rho^{2m}\left(x_3^{m-1}x_4^m \ldots x_n^m + \ldots + x_3^m x_4^m \ldots x_{n-1}^m x_n^{m-1}\right),$$

или

$$\rho^{2m} \approx a_2^{(k)}, \quad \rho \approx \sqrt[2m]{a_2^{(k)}}, \quad b_2^{(k)} \approx 2\rho^{2m-1}\cos\varphi, \tag{58}$$

т. е. мы найдем ρ и $\cos\varphi$, а следовательно и x_1, и x_2. Далее,

$$a_3^{(k)} \approx \rho^{2m} x_3^m, \quad b_3^{(k)} \approx \rho^{2m} x_3^{m-1} + 2\rho^{2m-1} x_3^m \cos\varphi, \quad x_3 \approx \frac{1}{c_3^{(k)} - c_2^{(k)}}. \tag{59}$$

Остальные корни находятся так же, как и в предыдущих случаях, по формулам (55). Можно рассмотреть аналогично случаи, когда кратные или комплексные корни имеют промежуточные модули. Наличие комплексных корней и их место обнаруживаются так же, как и в методе Лобачевского.

Видоизменение Лемера метода Лобачевского имеет несомненные преимущества по сравнению с обычным методом Лобачевского, так как при его применении не приходится извлекать корней высоких степеней, а также при наличии различных корней с одинаковыми модулями нет необходимости разъединять эти корни, рассматривая уравнение $P(x-h)=0$ и применяя к нему повторно метод Лобачевского.

§ 4. Итерационные методы решения алгебраических и трансцендентных уравнений

Для решения алгебраических и трансцендентных уравнений вида $f(x)=0$ разработано много различных итерационных методов. Сущность этих методов заключается в следующем. Пусть известна достаточно малая область, в которой содержится единственный корень $x=\alpha$ уравнения

$$f(x)=0. \tag{1}$$

В этой области выбирается точка x_0 — начальное приближение корня, — достаточно близкая к искомому корню $x=\alpha$, и с помощью некоторого рекуррентного соотношения

$$x_k = \varphi_k(x_0, x_1, \ldots, x_{k-1}) \tag{2}$$

строится последовательность точек $x_1, x_2, \ldots, x_n, \ldots$, сходящаяся к $x=\alpha$. Сходимость последовательности обеспечивается соответствующим выбором функции φ_k и начального приближения x_0. Выбирая различными способами функцию φ_k, которая зависит от $f(x)$ и в общем случае от номера k, можно получить различные итерационные методы.

Чаще всего по функции $f(x)$ строят функцию $\varphi(x)$ такую, что искомый корень $x=\alpha$ уравнения (1) является и корнем уравнения

$$x = \varphi(x), \tag{3}$$

и затем строят последовательность $\{x_n\}$ с помощью соотношения

$$x_n = \varphi(x_{n-1}) \qquad (n = 1, 2, \ldots), \tag{4}$$

исходя из некоторого начального приближения x_0. В этом случае функция φ не зависит от номера k, и методы такого типа мы будем называть *стационарными*.

Ниже мы исследуем ряд стационарных и нестационарных методов.

1. Принцип сжатых отображений и его применение к доказательству сходимости итерационных методов. Для исследования сходимости итерационных методов, а также для доказательства существования корня уравнения широко применяется принцип сжатых отображений, который мы сформулируем и докажем в общем виде и в форме, удобной для наших целей.

Пусть R — некоторое метрическое пространство, а Ax — некоторый оператор, определенный на этом пространстве. Будем говорить, что этот оператор осуществляет *сжатое* отображение R в себя, если существует такое положительное число α, меньшее единицы, что для любых двух элементов $x, y \in R$ имеет место неравенство

$$\rho(Ax, Ay) \leqslant \alpha\rho(x, y), \tag{5}$$

т. е. оператор A сближает элементы.

Теорема (*Принцип сжатых отображений*). *Если R — полное метрическое пространство, а оператор Ax осуществляет сжатое отображение R в себя, то существует одна и только одна неподвижная точка этого отображения, т. е. уравнение*

$$Ax = x \tag{6}$$

имеет одно и только одно решение. Решение этого уравнения может быть получено как предел последовательности

$$x_k = Ax_{k-1} \qquad (k = 1, 2, \ldots), \tag{7}$$

где x_0 *— произвольный элемент из* R.

Доказательство. Пусть x_0 — произвольный элемент из R. Образуем последовательность (7) и докажем, что она является фундаментальной последовательностью. В самом деле, если $n > m$, то

$$\rho(x_n, x_m) = \rho(Ax_{n-1}, Ax_{m-1}) \leqslant \alpha\rho(x_{n-1}, x_{m-1}),$$
$$\rho(x_{n-1}, x_{m-1}) \leqslant \alpha\rho(x_{n-2}, x_{m-2}),$$
$$\cdots\cdots\cdots\cdots\cdots$$
$$\rho(x_{n-m+1}, x_1) \leqslant \alpha\rho(x_{n-m}, x_0).$$

Следовательно,

$$\rho(x_n, x_m) \leqslant \alpha^m \rho(x_{n-m}, x_0).$$

Но

$$\rho(x_{n-m}, x_0) \leqslant \rho(x_0, x_1) + \rho(x_1, x_2) + \ldots + \rho(x_{n-m-1}, x_{n-m}) \leqslant$$
$$\leqslant \rho(x_0, x_1)[1 + \alpha + \alpha^2 + \ldots + \alpha^{n-m-1}] \leqslant \frac{1}{1-\alpha}\rho(x_0, x_1).$$

Таким образом,

$$\rho(x_n, x_m) \leqslant \frac{\alpha^m}{1-\alpha}\rho(x_0, x_1),$$

что и доказывает фундаментальность последовательности $\{x_n\}$, так как при достаточно большом m и любом $n > m$ правая часть последнего неравенства может быть сделана меньше любого заданного положительного числа.

Из полноты пространства R следует существование элемента $x \in R$, к которому сходится последовательность $\{x_n\}$, т. е.

$$x = \lim_{n\to\infty} x_n.$$

Далее, так как

$$\rho(Ax_n, Ax) \leqslant \alpha\rho(x_n, x)$$

и $\rho(x_n, x) \to 0$ при $n \to \infty$, то $\lim\limits_{n\to\infty} Ax_n = Ax$. Но $Ax_n = x_{n+1} \to x$. Следовательно,

$$x = Ax.$$

Для доказательства единственности допустим, что существуют два неподвижных элемента x и y, т. е.

$$x = Ax, \quad y = Ay, \quad \rho(x, y) \neq 0.$$

Но оператор Ax осуществляет сжатое отображение, поэтому

$$\rho(Ax, Ay) = \rho(x, y) \leqslant \alpha\rho(x, y),$$

что невозможно, так как $0 < \alpha < 1$.

Приведем еще одну теорему, которая уточняет доказанный выше принцип сжатых отображений.

Теорема. *Пусть в полном метрическом пространстве R или на его части, содержащей окрестность S элемента y_0: $S=\{\rho(x, y_0)\leqslant r\}$, определен оператор Ax. Пусть для любых $x, y\in S$*

$$\rho(Ax, Ay)\leqslant\alpha\rho(x, y) \tag{5'}$$

и

$$\rho(Ay_0, y_0)\leqslant(1-\alpha)r, \tag{5''}$$

где α — некоторое фиксированное положительное число, меньшее единицы. Тогда в S существует одно и только одно решение уравнения (6), которое может быть получено как предел последовательности (7), где x_0 — произвольный элемент из S.

Эта теорема непосредственно следует из принципа сжатых отображений, так как если $x\in S$, то

$$\rho(Ax, y_0)\leqslant\rho(Ax, Ay_0)+\rho(Ay_0, y_0)\leqslant\alpha\rho(x, y_0)+(1-\alpha)r\leqslant r,$$

т. е. $Ax\in S$. Это означает, что оператор Ax осуществляет отображение S в себя. Из (5') следует, что это отображение сжатое. Совокупность всех элементов S с тем же самым понятием расстояния $\rho(x, y)$ можно рассматривать как полное метрическое пространство, так как если $\{y_n\}$ — любая фундаментальная последовательность элементов из S, то она сходится в R, т. е. существует элемент $y\in R$ такой, что

$$y=\lim_{n\to\infty}y_n.$$

Но $\rho(y_n, y_0)\leqslant r$, отсюда $\lim\limits_{n\to\infty}\rho(y_n, y_0)=\rho(\lim\limits_{n\to\infty}y_n, y_0)=\rho(y, y_0)\leqslant r$, т. е. $y\in S$. Теперь уже утверждение теоремы прямо следует из принципа сжатых отображений.

Рассмотрим применение этого принципа к исследованию сходимости итерационного метода решения уравнения

$$x=\varphi(x) \tag{3}$$

при некоторых ограничениях на функцию $\varphi(x)$.

Предположим, что уравнение (3) имеет корень $x=\alpha$ и в круге $R\{|x-\alpha|\leqslant r\}$ функция $\varphi(x)$ удовлетворяет условию Липшица

$$|\varphi(x')-\varphi(x'')|\leqslant K|x'-x''| \tag{8}$$

для любых точек $x', x''\in R$.

Теорема. *Каково бы ни было $x_0\in R$, последовательность*

$$x_k=\varphi(x_{k-1})\qquad(k=1, 2, 3, \ldots) \tag{4}$$

сходится к α, если только $\varphi(x)$ в круге R удовлетворяет условию Липшица с константой $K<1$, причем скорость сходимости характеризуется неравенством

$$|x_n-\alpha|\leqslant K^n|x_0-\alpha|. \tag{9}$$

Доказательство. Совокупность точек круга R, если определить расстояние между точками x и y соотношением $\rho(x, y)=|x-y|$, образует полное метрическое пространство. Если $x\in R$, то $y=\varphi(x)$ также принадлежит R, ибо

$$\rho(y, \alpha)=|y-\alpha|=|\varphi(x)-\varphi(\alpha)|\leqslant K|x-\alpha|<r.$$

Отображение, определяемое функцией $\varphi(x)$, есть сжатое отображение R в себя, так как для любых x, $y\in R$

$$\rho[\varphi(x), \varphi(y)]=|\varphi(x)-\varphi(y)|\leqslant K|x-y|=K\rho(x, y) \quad \text{и} \quad K<1.$$

Поэтому по принципу сжатых отображений в R существует одна и только одна неподвижная точка, но эта точка $x=\alpha$. Эту точку можно получить как предел последовательности

$$x_k=\varphi(x_{k-1}) \qquad (k=1, 2, \ldots)$$

при любом $x_0\in R$.

Используя условие Липшица (8), имеем:

$$|x_n-\alpha|=|\varphi(x_{n-1})-\varphi(\alpha)|\leqslant K|x_{n-1}-\alpha|,$$

т. е.

$$|x_n-\alpha|\leqslant K|x_{n-1}-\alpha|\leqslant K^2|x_{n-2}-\alpha|\leqslant \ldots \leqslant K^n|x_0-\alpha|.$$

Таким образом, $\{x_n\}$ сходится к α со скоростью геометрической прогрессии со знаменателем K.

В доказанной теореме мы предполагали существование корня уравнения (3). Принцип сжатых отображений может быть использован и для доказательства существования корня.

Теорема. *Если функция $\varphi(x)$* в некотором круге $R\{|x-x_0|\leqslant r\}$ *удовлетворяет условию Липшица* (8) *с константой $K<1$ и в точке x_0 имеет место неравенство*

$$|x_0-\varphi(x_0)|\leqslant(1-K)r, \tag{10}$$

то в R уравнение (3) *имеет единственный корень $x=\alpha$, который может быть найден как предел последовательности*

$$y_k=\varphi(y_{k-1}) \qquad (k=1, 2, \ldots),$$

где y_0 — любая точка из круга R.

Доказательство. Пусть $x\in R$. Тогда

$$|\varphi(x)-x_0|=|\varphi(x)-\varphi(x_0)+\varphi(x_0)-x_0|\leqslant$$
$$\leqslant|\varphi(x)-\varphi(x_0)|+|\varphi(x_0)-x_0|\leqslant K|x-x_0|+(1-K)r\leqslant r,$$

т. е. функция $\varphi(x)$ осуществляет отображение R в себя. Это — сжатое отображение, так как для любых $x,\ y \in R$

$$\rho(\varphi(x),\ \varphi(y)) = |\varphi(x) - \varphi(y)| \leqslant K|y - x| = K\rho(x,\ y).$$

Следовательно, имеет место принцип сжатых отображений, из которого прямо следует теорема.

Мы требовали от функции $\varphi(x)$ выполнения условия Липшица с константой $K < 1$ в некоторой окрестности корня $x = \alpha$ уравнения (3). Чаще всего функция $\varphi(x)$ имеет производные, и условие Липшица с константой $K < 1$ будет иметь место, если в некоторой окрестности корня $x = \alpha$ функция $\varphi(x)$ имеет производную $\varphi'(x)$, по модулю меньшую некоторого числа, меньшего единицы, т. е.

$$|\varphi'(x)| \leqslant K < 1. \tag{11}$$

Если производная $\varphi'(x)$ непрерывна, то из условия $|\varphi'(\alpha)| < K < 1$ следует выполнение этого неравенства в некоторой окрестности корня $x = \alpha$, и для отыскания этого корня можно применить метод итераций, выбирая начальное приближение x_0 из этой окрестности. Скорость сходимости будет тем больше, чем меньше K.

Введем *понятие порядка итерации* (4) для решения уравнения (3). Будем говорить, что итерация (4) имеет порядок m, если

$$\varphi'(\alpha) = \varphi''(\alpha) = \ldots = \varphi^{(m-1)}(\alpha) = 0, \quad \varphi^{(m)}(\alpha) \neq 0. \tag{12}$$

Если $\varphi(x)$ имеет в окрестности корня m непрерывных производных, то по формуле Тейлора

$$\varphi(x) - \alpha = (x - \alpha)\varphi'(\alpha) + \frac{(x-\alpha)^2}{2!}\varphi''(\alpha) + \ldots$$

$$\ldots + \frac{(x-\alpha)^{m-1}}{(m-1)!}\varphi^{(m-1)}(\alpha) + \frac{(x-\alpha)^m}{m!}\varphi^{(m)}(\xi).$$

В случае итерации порядка m имеем:

$$\varphi(x) - \alpha = \frac{(x-\alpha)^m}{m!}\varphi^{(m)}(\xi),$$

откуда

$$x_k - \alpha = \frac{(x_{k-1} - \alpha)^m}{m!}\varphi^{(m)}(\xi_k). \tag{13}$$

Обозначая через M_m максимум модуля $\varphi^{(m)}(x)$ в некоторой фиксированной окрестности корня $x = \alpha$, получим неравенство

$$|x_k - \alpha| \leqslant \frac{M_m}{m!}|x_{k-1} - \alpha|^m, \tag{14}$$

из которого следует:

$$|x_k-\alpha|\leqslant\left(\frac{M_m}{m!}\right)^{1+m+m^2+\ldots+m^{k-1}}|x_0-\alpha|^{m^k}=$$

$$=\left(\frac{M_m}{m!}\right)^{\frac{m^k-1}{m-1}}|x_0-\alpha|^{m^k}=$$

$$=\left(\frac{M_m}{m!}|x_0-\alpha|\right)^{\frac{m^k-1}{m-1}}|x_0-\alpha|^{\frac{m^{k+1}-2m^k+1}{m-1}}. \qquad (15)$$

Таким образом, если $|x_0-\alpha|<1$ и $\frac{M_m}{m!}|x_0-\alpha|=\omega<1$, то

$$|x_k-\alpha|\leqslant\omega^{\frac{m^k-1}{m-1}}, \qquad (16)$$

что дает очень быструю сходимость x_k к α.

Для случая, когда $\varphi(x)$ — действительная функция переменного x и $x=\alpha$ — действительный корень уравнения (3), метод итерации (4) имеет хорошую геометрическую интерпретацию. На рис. 6 и 7

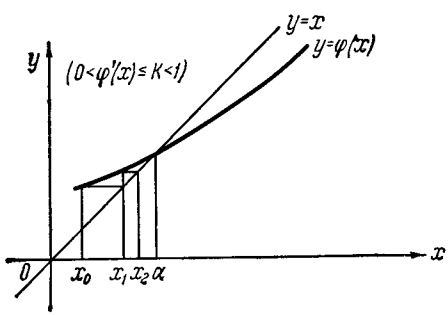

Рис. 6.

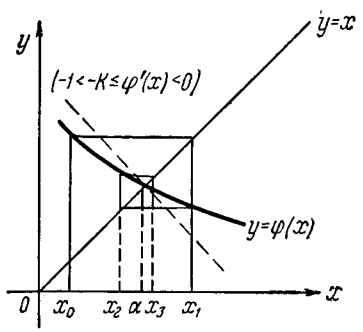

Рис. 7.

изображена геометрическая картина метода итераций (4) для случаев, когда $0<\varphi'(x)\leqslant K<1$ и $-1<-K\leqslant\varphi'(x)<0$.

Геометрически видно, что если в окрестности корня $x=\alpha$ имеет место неравенство $0<\varphi'(x)\leqslant K<1$, то последовательность $\{x_n\}$ монотонно сходится к корню, причем с той стороны, с которой расположено начальное приближение x_0, а в случае $-1<-K\leqslant\varphi'(x)<0$ последовательные приближения расположены поочередно с разных сторон от точки $x=\alpha$. В последнем случае по двум последовательным приближениям корня можно судить о достигнутой точности на каждом шаге, так как отклонение x_n от α не больше $|x_n-x_{n-1}|$. Эти утверждения можно доказать и аналитически, но мы не будем останавливаться на этом доказательстве.

2. Простейшие итерационные методы: метод секущих и метод Ньютона. Если уравнение $f(x)=0$ имеет корень $x=\alpha$, а функция $\psi(x)$ непрерывна в окрестности $x=\alpha$, то уравнение

$$x=\varphi(x)\equiv x-\psi(x)f(x) \tag{17}$$

также имеет корень $x=\alpha$. Функцию $\psi(x)$ можно подобрать так, что итерационный процесс для уравнения (17) будет сходящимся.

Рассмотрим два классических метода, которые можно получить этим способом.

Пусть $f(x)$ — действительная функция действительного переменного x, а $x=\alpha$ — действительный корень уравнения $f(x)=0$. Предположим, что в некоторой окрестности точки $x=\alpha$ функция $f(x)$ вместе с $f'(x)$ и $f''(x)$ непрерывна, а $f'(x)$ и $f''(x)$ в этой окрестности не меняют знака. Это означает, что при переходе через $x=\alpha$ функция $f(x)$ меняет знак и имеет точку $x=\alpha$ простым корнем. Пусть x_0 — точка рассматриваемой окрестности, в которой $f(x_0)f''(x_0)>0$. В (17) в качестве функции $\psi(x)$ возьмем функцию

$$\psi(x)\equiv\frac{x-x_0}{f(x)-f(x_0)}.$$

Тогда уравнение

$$x=\varphi(x)\equiv x-\frac{x-x_0}{f(x)-f(x_0)}f(x)=\frac{x_0f(x)-xf(x_0)}{f(x)-f(x_0)} \tag{18}$$

также имеет корнем $x=\alpha$. За начальное приближение примем любую, достаточно близкую к α точку x_1 рассматриваемой окрестности, в которой $f(x_1)$ имеет знак, противоположный знаку $f(x_0)$, а последующие приближения будем строить обычным способом:

$$x_n=\frac{x_0f(x_{n-1})-x_{n-1}f(x_0)}{f(x_{n-1})-f(x_0)}\qquad(n=2,\ 3,\ \ldots). \tag{19}$$

Так как, с одной стороны,

$$\varphi'(\alpha)=\frac{[x_0f'(\alpha)-f(x_0)][f(\alpha)-f(x_0)]-f'(\alpha)[x_0f(\alpha)-\alpha f(x_0)]}{[f(\alpha)-f(x_0)]^2}=$$
$$=\frac{f(x_0)+(\alpha-x_0)f'(\alpha)}{f(x_0)},$$

а с другой стороны, по формуле Тейлора

$$f(x)=f(\alpha)+(x-\alpha)f'(\alpha)+\frac{(x-\alpha)^2}{2!}f''(\xi),$$

где ξ заключено между α и x, то, полагая $x=x_0$, получим:

$$f(x_0)+(\alpha-x_0)f'(\alpha)=\frac{(x_0-\alpha)^2}{2!}f''(\xi).$$

Следовательно,

$$\varphi'(\alpha)=\frac{(x_0-\alpha)^2}{2}\frac{f''(\xi)}{f(x_0)}.$$

При x_0, достаточно близком к $x=\alpha$, $\varphi'(\alpha)$ — малое число, и поэтому существует такая окрестность точки α, в которой будет иметь место неравенство

$$|\varphi'(x)| \leqslant K < 1,$$

и если x_1 взято из этой окрестности, то последовательность (19) будет сходиться к $x=\alpha$.

Так как $f(x_n) = f(x_n) - f(\alpha) = f'(\xi)(x_n - \alpha)$, то, положив $m = \min\limits_{x \in [x_0,\ x_1]} |f'(x)|$, будем иметь:

$$|x_n - \alpha| \leqslant \frac{|f(x_n)|}{m}, \tag{20}$$

что позволяет на каждом шаге по значениям $f(x_n)$ следить за достигнутой точностью.

Геометрически этот метод состоит в том, что значение x_{n+1} есть абсцисса точки пересечения прямой, проходящей через точки $(x_0, f(x_0))$ и $(x_n, f(x_n))$, с осью x (рис. 8). Поэтому этот метод

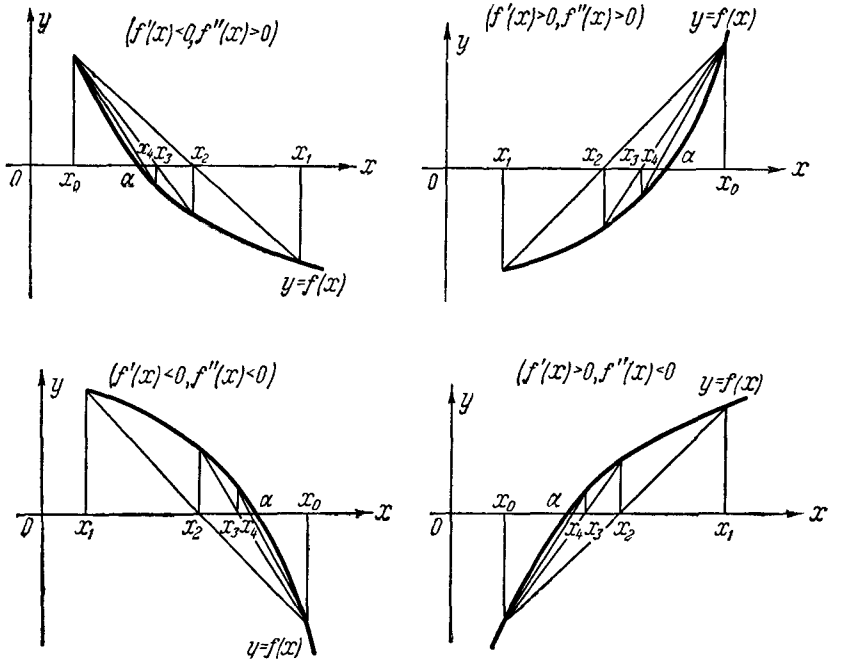

Рис. 8.

называют *методом секущих* или *методом линейной интерполяции*, так как на каждом шаге за приближенное значение корня x_{n+1} принимается корень интерполяционного многочлена первой степени, построенного по значениям $f(x)$ в точках x_0 и x_n.

Метод секущих является итерационным методом первого порядка.

Второй классический метод решения уравнения $f(x)=0$ — *метод Ньютона* — получим, если положить в (17)

$$\psi(x) \equiv \frac{1}{f'(x)},$$

т. е. свести отыскание корня $x=\alpha$ уравнения $f(x)=0$ к отысканию корня уравнения

$$x = x - \frac{f(x)}{f'(x)} \equiv \varphi(x). \tag{21}$$

Будем предполагать, что на отрезке $[a, b]$, содержащем единственный корень $x=\alpha$ уравнения $f(x)=0$, функция $f(x)$ имеет непрерывные производные $f'(x)$ и $f''(x)$, не обращающиеся в нуль на этом отрезке. В этом случае

$$\varphi'(x) = 1 - \frac{f'^2(x) - f(x) f''(x)}{f'^2(x)}$$

и $\varphi'(\alpha)=0$. Это означает, что существует такая окрестность точки $x=\alpha$, что если начальное приближение $x=x_0$ взято из этой окрестности, то последовательность

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})} \qquad (n=1, 2, \ldots) \tag{22}$$

будет сходиться к $x=\alpha$. Начальное приближение x_0 целесообразно выбирать так, чтобы было

$$f(x_0) f''(x_0) > 0. \tag{23}$$

Метод Ньютона применим не только для отыскания действительных корней уравнения $f(x)=0$, но и комплексных корней, только нужно иметь в виду, что при отыскании комплексного корня в случае действительной функции $f(x)$ начальное приближение x_0 нужно брать комплексным числом, а не действительным.

В случае, если $x=\alpha$ является действительным корнем уравнения $f(x)=0$, этот метод имеет простую геометрическую интерпретацию. Значение x_{n+1} есть абсцисса точки пересечения касательной к кривой $y=f(x)$ в точке $x=x_n$ с осью x (рис. 9). Поэтому метод Ньютона часто называют *методом касательных*.

Как видно из рис. 9 последовательные приближения к действительному корню в методе Ньютона сходятся к нему монотонно, приближаясь со стороны x_0.

Если за начальное приближение в методе Ньютона взять точку x_0, где $f(x_0) f''(x_0) < 0$, то, как видно из рис. 10, мы можем не прийти к корню $x=\alpha$, если только начальное приближение не очень хорошее.

Так как в методе Ньютона $\varphi'(\alpha)=0$, а $\varphi''(\alpha)$, вообще говоря, не равна нулю, то метод Ньютона является итерационным методом второго порядка.

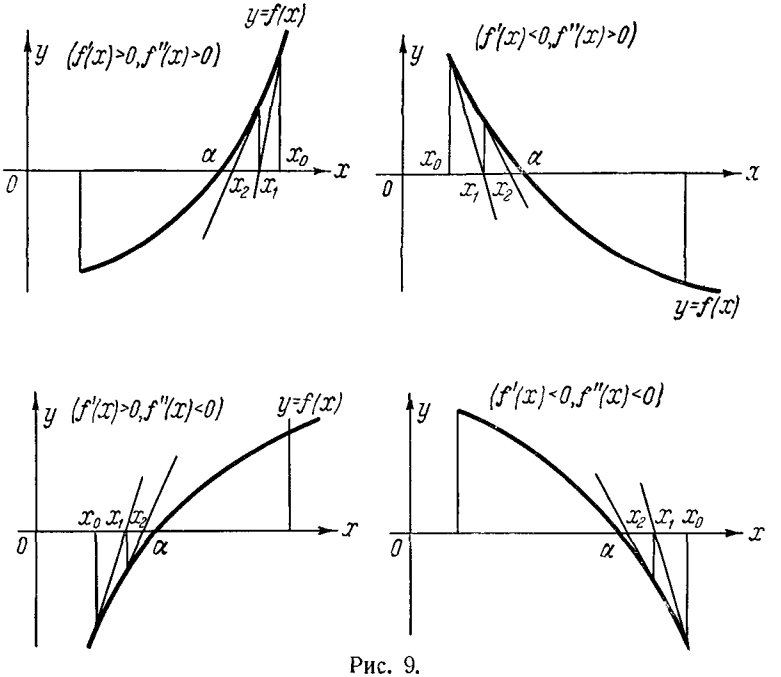

Рис. 9.

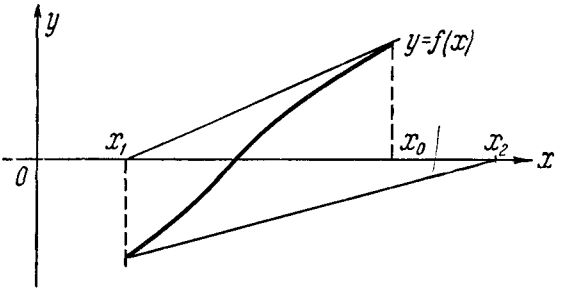

Рис. 10.

Скорость сходимости метода Ньютона можно оценить следующим образом. По формуле Тейлора

$$0 = f(\alpha) = f(x_n) + f'(x_n)(\alpha - x_n) + \frac{1}{2} f''(\xi)(\alpha - x_n)^2,$$

где ξ заключено между α и x_n. Отсюда

$$\frac{f(x_n)}{f'(x_n)} = x_n - \alpha - \frac{1}{2}\frac{f''(\xi)}{f'(x_n)}(\alpha - x_n)^2.$$

Следовательно,

$$x_{n+1} - \alpha = x_n - \alpha - \frac{f(x_n)}{f'(x_n)} = \frac{1}{2}\frac{f''(\xi)}{f'(x_n)}(\alpha - x_n)^2.$$

Если $m_1 = \min\limits_{[a,\,b]} |f'(x)|$, а $M_2 = \max\limits_{[a,\,b]} |f''(x)|$, где $[a, b]$ — отрезок, содержащий x_0 и α, на котором не меняют знака $f'(x)$ и $f''(x)$, то

$$|x_{n+1} - \alpha| \leqslant \frac{M_2}{2m_1} |x_n - \alpha|^2. \tag{24}$$

Это указывает на быструю сходимость метода Ньютона.

Комбинируя метод секущих и метод Ньютона, можно получить нестационарный метод отыскания действительных корней уравнения $f(x) = 0$, преимущество которого заключается в том, что при прежних предположениях относительно $f'(x)$ и $f''(x)$ последовательные приближения x_n и x_{n+1} лежат по разные стороны от корня, и поэтому можно следить в процессе вычислений за достигнутой точностью, и в то же время он сходится значительно быстрей метода секущих.

Пусть на отрезке $[a, b]$ содержится единственный корень уравнения $f(x) = 0$, а $f'(x)$ и $f''(x)$ на этом отрезке не меняют знаков. Если $f(a) f''(a) > 0$, то находим x_0 и x_1 по формулам:

$$x_0 = a - \frac{f(a)}{f'(a)}, \quad x_1 = \frac{af(b) - bf(a)}{f(b) - f(a)}, \tag{25}$$

а следующие приближения находим по формулам:

$$x_{2n} = x_{2n-2} - \frac{f(x_{2n-2})}{f'(x_{2n-2})}, \quad x_{2n+1} = \frac{x_{2n-2} f(x_{2n-1}) - x_{2n-1} f(x_{2n-2})}{f(x_{2n-1}) - f(x_{2n-2})}. \tag{26}$$

Если же $f(b) f''(b) > 0$, то x_0 и x_1 находим по формулам:

$$x_0 = b - \frac{f(b)}{f'(b)}, \quad x_1 = \frac{af(b) - bf(a)}{f(b) - f(a)}, \tag{25'}$$

а следующие приближения — по тем же формулам (26). Как видно из рис. 11, последовательные приближения x_{2n} и x_{2n+1} всегда расположены по разные стороны от $x=\alpha$ и первые совпадающие знаки x_{2n} и x_{2n+1} и будут верными знаками для $x=\alpha$.

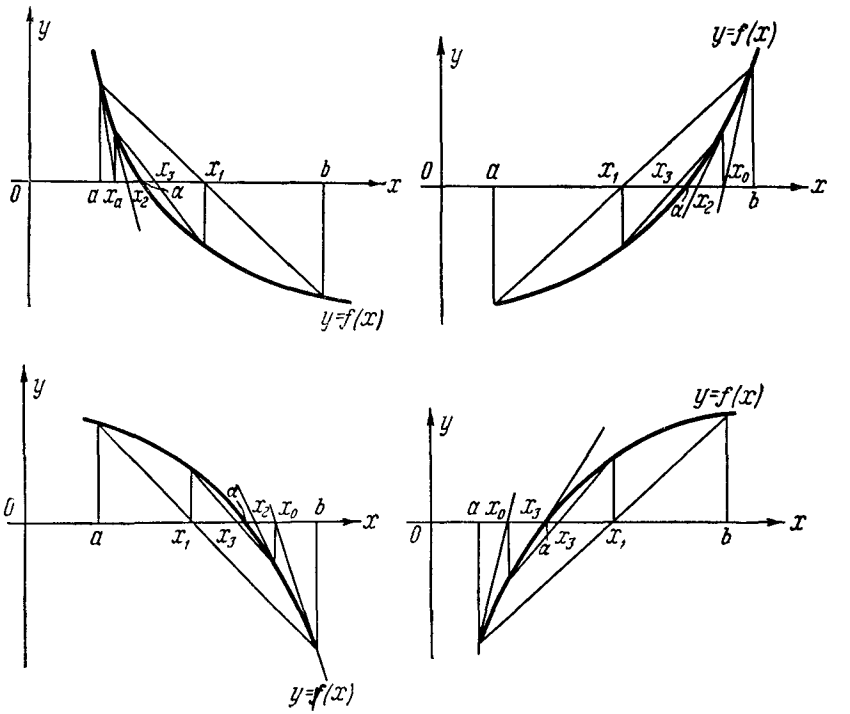

Рис. 11.

3. Метод Чебышева построения итераций высших порядков. В 1838 г. П. Л. Чебышев предложил метод отыскания действительных корней уравнения $f(x)=0$, частными случаями которого явились многие, разработанные до него методы. В основе метода Чебышева лежит представление функции, обратной к функции $f(x)$, по формуле Тейлора.

Пусть уравнение $f(x)=0$ на отрезке $[a, b]$ имеет корень $x=\alpha$. Относительно функции $f(x)$ предположим, что она непрерывна на отрезке $[a, b]$ вместе с производными достаточно высокого порядка и $f'(x)\neq 0$ на $[a, b]$. При этих предположениях функция $y=f(x)$ имеет обратную функцию $x=F(y)$, определенную на отрезке $[c, d]$, являющемся областью значений $f(x)$ при $x\in[a, b]$. Функция $F(y)$ имеет столько же непрерывных производных, сколько имеет и $f(x)$. Так как

$$x\equiv F[f(x)] \quad (x\in[a, b]), \quad y\equiv f[F(y)] \quad (y\in[c, d]), \qquad (27)$$

то

$$\alpha = F(0). \tag{28}$$

При $y \in [c, d]$ формула Тейлора дает

$$\alpha = F(0) = F(y) + \sum_{k=1}^{r} (-1)^k \frac{F^{(k)}(y)}{k!} y^k + R_{r+1}, \tag{29}$$

где остаточный член может быть записан в виде

$$R_{r+1} = (-1)^{r+1} \frac{F^{(r+1)}(\eta)}{(r+1)!} y^{r+1} \tag{30}$$

(η заключено между 0 и y), или

$$\alpha = x + \sum_{k=1}^{r} (-1)^k \frac{F^{(k)}[f(x)]}{k!} [f(x)]^k + (-1)^{r+1} \frac{F^{(r+1)}(\eta)}{(r+1)!} [f(x)]^{r+1}. \tag{31}$$

Для упрощения записи положим

$$F^{(k)}[f(x)] \equiv a_k(x), \quad \varphi_r(x) \equiv x + \sum_{k=1}^{r} (-1)^k \frac{a_k(x)}{k!} [f(x)]^k. \tag{32}$$

Уравнение

$$x = \varphi_r(x) \tag{33}$$

имеет корень $x = \alpha$, так как $\varphi_r(\alpha) = \alpha - \sum_{k=1}^{r} (-1)^k \frac{a_k(\alpha)}{k!} [f(\alpha)]^k = \alpha$.

Положив

$$x_{n+1} = \varphi_r(x_n) \qquad (n = 0, 1, \ldots, x_0 \in [a, b]), \tag{34}$$

получим итерационный метод $(r+1)$-го порядка, так как

$$\varphi_r^{(l)}(\alpha) = 0 \qquad (l = 1, 2, \ldots, r).$$

Если x_0 взято достаточно близко к α, то последовательность $\{x_n\}$ сходится к α, ибо существует такая окрестность точки α, в которой

$$|\varphi'(x)| \leqslant K < 1,$$

и для сходимости $\{x_n\}$ нужно только потребовать, чтобы x_0 принадлежала этой окрестности.

Функцию $\varphi_r(x)$ можно найти в явном виде через $f(x)$ и ее производные, так как из тождества (27) имеем:

$$F'[f(x)] \cdot f'(x) = 1,$$

$$F''[f(x)] f'^2(x) + F'[f(x)] f''(x) = 0,$$

$$F''[f(x)] f'^3(x) + 3F''[f(x)] f'(x) f''(x) + F'[f(x)] f'''(x) = 0$$

. . . . . .

или

$$\left.\begin{aligned} &a_1(x)f'(x)=1,\\ &a_2(x)f'^2(x)+a_1(x)f''(x)=0,\\ &a_3(x)f'^3(x)+3a_2(x)f'(x)f''(x)+a_1(x)f'''(x)=0,\\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\end{aligned}\right\} \tag{35}$$

т. е. можно последовательно найти $a_1(x)$, $a_2(x)$, ..., а следовательно, и $\varphi_r(x)$. При $r=1$

$$\varphi_1(x)=x-\frac{f(x)}{f'(x)} \quad \text{и} \quad x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)}, \tag{36}$$

т. е. мы снова получаем метод Ньютона. При $r=2$

$$\left.\begin{aligned} &\varphi_2(x)=x-\frac{f(x)}{f'(x)}-\frac{f''(x)f^2(x)}{2f'^3(x)}\\ &\text{и}\\ &x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)}-\frac{f''(x_n)f^2(x_n)}{2f'^3(x_n)}.\end{aligned}\right\} \tag{37}$$

Оценка погрешности и скорость сходимости легко получаются из равенства (31). Полагая в нем $x=x_n$ и учитывая (34), получим:

$$\alpha-x_{n+1}=(-1)^{r+1}\frac{F^{(r+1)}[f(\xi)]}{(r+1)!}[f(x_n)]^{r+1}, \tag{38}$$

где ξ лежит между α и x_n. Если положить $L=\max\limits_{x\in[a,\,b]}|f'(x)|$, $M_{r+1}=\max\limits_{x\in[a,\,b]}|F^{(r+1)}[f(x)]|$ и учесть, что

$$|f(x_n)|=|f(x_n)-f(\alpha)|=|f'(\eta)||x_n-\alpha|\leqslant L|x_n-\alpha|,$$

то из (38) имеем:

$$|\alpha-x_{n+1}|\leqslant\frac{M_{r+1}L^{r+1}}{(r+1)!}|x_n-\alpha|^{r+1}=q|x_n-\alpha|^{r+1}\left(q=\frac{M_{r+1}L^{r+1}}{(r+1)!}\right). \tag{39}$$

Отсюда следует, что

$$\begin{aligned}|\alpha-x_m|&\leqslant q^{1+(r+1)+(r+1)^2+\ldots+(r+1)^{m-1}}|x_0-\alpha|^{(r+1)^m}=\\&=(q|x_0-\alpha|)^{\frac{(r+1)^m-1}{r}}|x_0-\alpha|^{\frac{(r+1)^m(r-1)+1}{r}}.\end{aligned}$$

Таким образом, если $|x_0-\alpha|<1$ и $q|x_0-\alpha|=\omega<1$, то

$$|\alpha-x_m|\leqslant\omega^{\frac{(r+1)^m-1}{r}}, \tag{40}$$

что указывает на очень быструю сходимость метода. Так, если $\omega < 0{,}1$ и $|x_0 - \alpha| < 1$ то при $r = 1$ (метод Ньютона)

$$|\alpha - x_1| \leqslant 10^{-1}, \quad |\alpha - x_2| \leqslant 10^{-3}, \quad |\alpha - x_3| \leqslant 10^{-7},$$
$$|\alpha - x_4| \leqslant 10^{-15}, \ldots;$$

при $r = 2$

$$|\alpha - x_1| \leqslant 10^{-1}, \quad |\alpha - x_2| \leqslant 10^{-4}, \quad |\alpha - x_3| \leqslant 10^{-13},$$
$$|\alpha - x_4| \leqslant {}^{-40}, \ldots,$$

т. е. количество верных десятичных знаков быстро возрастает.

4. Построение итераций высших порядков с помощью теоремы Кёнига. А. *Теорема Кёнига.* Прежде чем излагать указанный метод построения итераций, докажем *теорему Кёнига.*

Теорема. *Если $f(z)$ и $\varphi(z)$ — аналитические функции в области $|z| < r$, содержащей единственный корень $z = \alpha$ уравнения $f(z) = 0$ кратности единица и $\varphi(\alpha) \neq 0$, то*

$$\alpha = \lim_{n \to \infty} \frac{c_n}{c_{n+1}}, \tag{41}$$

где c_n — коэффициент при z^n в разложении $\frac{\varphi(z)}{f(z)}$ по степеням z.

Доказательство. Так как для функции $\frac{\varphi(z)}{f(z)}$ точка $z = \alpha$ является ближайшей к началу координат особой точкой, то ряд

$$\frac{\varphi(z)}{f(z)} = \sum_{n=0}^{\infty} c_n z^n \tag{42}$$

сходится при всех $|z| < \alpha$, а ряд

$$(z - \alpha)\frac{\varphi(z)}{f(z)} = \sum_{n=0}^{\infty} d_n z^n \tag{43}$$

сходится при всех z, для которых $|z| < r$. Отсюда

$$\sum_{n=0}^{\infty} c_n z^n (z - \alpha) = \sum_{n=0}^{\infty} d_n z^n$$

и

$$-\alpha c_0 = d_0, \quad c_{n-1} - \alpha c_n = d_n \qquad (n = 1, 2, \ldots). \tag{44}$$

Таким образом,

$$\sum_{k=0}^{m} d_k \alpha^k = -\alpha c_0 + \sum_{k=1}^{m} (c_{k-1} - \alpha c_k)\alpha^k = -c_m \alpha^{m+1}. \tag{45}$$

Если положить

$$S_m(z)=\sum_{k=0}^{m} d_k z^k, \tag{46}$$

то из (45) имеем:

$$c_m\alpha^{m+1}=-S_m(\alpha) \tag{47}$$

$$\frac{c_m}{c_{m+1}}=\alpha\frac{S_m(\alpha)}{S_{m+1}(\alpha)}. \tag{48}$$

Если ρ — некоторое число, удовлетворяющее неравенству $|\alpha|<\rho<r$, то ряд (43) сходится при $z=\rho$ и $|d_n\rho^n|\to 0$. Пусть $A=\max\limits_n|d_n\rho^n|$. Тогда

$$|d_n|\leqslant\frac{A}{\rho^n}. \tag{49}$$

Так как

$$\alpha-\frac{c_m}{c_{m+1}}=\alpha\left(1-\frac{S_m(\alpha)}{S_{m+1}(\alpha)}\right)=\frac{d_{m+1}\alpha^{m+2}}{S_{m+1}(\alpha)}, \tag{50}$$

то при $0<\delta<B$, где $B=|S(\alpha)|$, существует такое m_0, что при $m>m_0$ и $|S_{m+1}(\alpha)|\geqslant B-\delta$

$$\left|\alpha-\frac{c_m}{c_{m+1}}\right|\leqslant\frac{A}{\rho^{m+1}}\frac{|\alpha|^{m+2}}{B-\delta}=\frac{A\alpha}{B-\delta}\left(\left|\frac{\alpha}{\rho}\right|^{m+1}\right). \tag{51}$$

Но $\left|\frac{\alpha}{\rho}\right|<1$, поэтому правая часть в (51) стремится к нулю при $m\to\infty$, т. е.

$$\lim_{m\to\infty}\frac{c_m}{c_{m+1}}=\alpha.$$

Если функция $f(z)$ имеет единственный простой корень $z=\alpha$ в области $|z-x|<r$, где x — некоторое число, а $f(z)$ и $\varphi(z)$ — аналитические функции в этой области, причем $\varphi(\alpha)\neq 0$, то из теоремы Кёнига следует, что

$$\alpha-x=\lim_{n\to\infty}\frac{c_n(x)}{c_{n+1}(x)}, \tag{52}$$

где $c_n(x)$ — коэффициент при $(z-x)^n$ в разложении $\frac{\varphi(z)}{f(z)}$ в ряд по степеням $z-x$.

Рассмотрим уравнение

$$x=\varphi_p(x)\equiv x+\frac{c_p(x)}{c_{p+1}(x)}\qquad(p\geqslant m_0) \tag{53}$$

и итерацию

$$x_{n+1}=\varphi_p(x_n)\qquad(n=0,\ 1,\ 2,\ \ldots). \tag{54}$$

Докажем, что последовательность $\{x_n\}$ сходится к α, если x_0 достаточно близко к α, и итерация (54) имеет порядок не ниже $p+2$. В самом деле, если

$$\frac{\varphi(z)}{f(z)} = \sum_{n=0}^{\infty} c_n(x)(z-x)^n, \quad (z-\alpha)\frac{\varphi(z)}{f(z)} = \sum_{n=0}^{\infty} d_n(x)(z-x)^n,$$

то совершенно аналогично равенству (50) имеет место равенство

$$\alpha - x - \frac{c_p(x)}{c_{p+1}(x)} = \frac{(\alpha-x)^{p+2} d_{p+1}(x)}{S_{p+1}(\alpha, x)}, \tag{55}$$

где

$$S_{p+1}(\alpha, x) = \sum_{k=0}^{p+1} d_k(x)\alpha^k.$$

Отсюда видно, что $\alpha - \varphi_p(x)$ имеет множитель $(\alpha-x)^{p+2}$, поэтому $\varphi_p(\alpha)=\alpha$ и $\varphi_p^{(l)}(\alpha)=0$ при $l=1, 2, \ldots, p+1$. Это и означает, что итерация (54) имеет порядок не ниже $p+2$ и будет сходиться к α, если x_0 взято из окрестности $|z-\alpha| \leqslant \rho$, в которой

$$|\varphi_p'(z)| \leqslant K < 1.$$

Б. *Построение итераций высших порядков.* Функцию $\varphi_p(x)$ можно находить разными путями. Прежде всего, так как $c_p(x)$ есть коэффициент разложения $\frac{\varphi(z)}{f(z)}$ по степеням $z-x$, то

$$c_p(x) = \frac{1}{p!}\left[\frac{\varphi(z)}{f(z)}\right]^{(p)}_{z=x} \tag{56}$$

и

$$\varphi_p(x) = x + (p+1)\frac{\left[\frac{\varphi(z)}{f(z)}\right]^{(p)}_{z=x}}{\left[\frac{\varphi(z)}{f(z)}\right]^{(p+1)}_{z=x}}. \tag{57}$$

Если известны разложения функций $f(z)$ и $\varphi(z)$ по степеням $z-x$:

$$f(z) = \sum_{k=0}^{\infty} a_k(x)(z-x)^k; \quad \varphi(z) = \sum_{k=0}^{\infty} b_k(x)(z-x)^k, \tag{58}$$

то

$$\sum_{k=0}^{\infty} b_k(x)(z-x)^k = \sum_{k=0}^{\infty} a_k(x)(z-x)^k \cdot \sum_{k=0}^{\infty} c_k(x)(z-x)^k,$$

откуда

$$\left.\begin{aligned}
b_0(x) &= a_0(x)\,c_0(x),\\
b_1(x) &= a_1(x)\,c_0(x) + a_0(x)\,c_1(x),\\
b_2(x) &= a_2(x)\,c_0(x) + a_1(x)\,c_1(x) + a_0(x)\,c_2(x),\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
b_p(x) &= a_p(x)\,c_0(x) + a_{p-1}(x)\,c_1(x) + \ldots + a_0(x)\,c_p(x),
\end{aligned}\right\} \tag{59}$$

и $c_k(x)$ можно найти, используя эти рекуррентные соотношения. Рассматривая (59) как систему линейных уравнений относительно $c_0(x)$, $c_1(x)$, ..., $c_p(x)$ и решая ее по правилу Крамера, получим:

$$c_p(x) = \frac{(-1)^p}{[a_0(x)]^{p+1}} \begin{vmatrix} b_0(x) & a_0(x) & 0 & 0 & \ldots & 0\\ b_1(x) & a_1(x) & a_0(x) & 0 & \ldots & 0\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\ b_p(x) & a_p(x) & a_{p-1}(x) & a_{p-2}(x) & \ldots & a_1(x) \end{vmatrix} = \frac{(-1)^p \Delta_p(x)}{a_0^{p+1}(x)} \tag{60}$$

и

$$\varphi_p(x) = x - \frac{a_0(x)\,\Delta_p(x)}{\Delta_{p+1}(x)}. \tag{61}$$

Заметим, что при $\varphi(z) \equiv 1$ и $p = 0$ мы снова приходим к методу Ньютона.

5. Метод Эйткена построения итераций высших порядков. Эйткен предложил способ получения из данной итерации или из двух данных итераций одного и того же порядка итерации более высокого порядка.

Пусть имеются итерации

$$x_n^{(1)} = \varphi_1\left(x_{n-1}^{(1)}\right), \quad x_n^{(2)} = \varphi_2\left(x_{n-1}^{(2)}\right) \qquad (n = 1,\ 2,\ \ldots) \tag{62}$$

порядка r, сходящиеся к $x = \alpha$. С помощью функций $\varphi_1(x)$ и $\varphi_2(x)$ построим фунцию $\Phi(x)$:

$$\Phi(x) \equiv \frac{x\varphi_1[\varphi_2(x)] - \varphi_1(x)\,\varphi_2(x)}{x - \varphi_1(x) - \varphi_2(x) + \varphi_1[\varphi_2(x)]}. \tag{63}$$

Тогда итерация

$$x_n = \Phi(x_{n-1}) \qquad (n = 1,\ 2,\ \ldots) \tag{64}$$

имеет порядок выше r, если только выполнено условие

$$[\varphi_1'(\alpha) - 1][\varphi_2'(\alpha) - 1] \neq 0. \tag{65}$$

Для доказательства этого утверждения заметим, что последовательность $y_n^{(i)} = x_n^{(i)} - \beta$ $(i = 1,\ 2)$ сходится к $\alpha - \beta$ и получается при решении уравнения

$$y = \varphi_i(y + \beta) - \beta = \psi_i(y) \qquad (i = 1,\ 2) \tag{66}$$

методом итераций, так как

$$x_n^{(i)} = y_n^{(i)} + \beta = \varphi_i\left(x_{n-1}^{(i)}\right) = \varphi_i\left(y_n^{(i)} + \beta\right).$$

При $\beta = \alpha$ члены последовательности $\eta_n^{(i)} = x_n^{(i)} - \alpha$ представляют отклонения $x_n^{(i)}$ от точного значения корня α и получаются применением метода итераций к уравнению

$$\eta = \omega_i(\eta) \equiv \varphi_i(\eta + \alpha) - \alpha. \tag{67}$$

Так как $\varphi_i(x)$ определяет итерацию порядка r, то имеет место разложение

$$\omega_i(\eta) = \alpha_r^{(i)}\eta^r + \alpha_{r+1}^{(i)}\eta^{r+1} + \ldots \tag{68}$$

Далее,

$$\Phi(\eta+\alpha) - \alpha = \frac{(\eta+\alpha)\,\varphi_1[\varphi_2(\eta+\alpha)] - \varphi_1(\eta+\alpha)\,\varphi_2(\eta+\alpha)}{\eta+\alpha-\varphi_1(\eta+\alpha)-\varphi_2(\eta+\alpha)+\varphi_1[\varphi_2(\eta+\alpha)]} - \alpha =$$

$$= \frac{(\eta+\alpha)\,\varphi_1[\omega_2(\eta)+\alpha] - [\omega_1(\eta)+\alpha][\omega_2(\eta)+\alpha]}{\eta+\alpha-[\omega_1(\eta)+\alpha]-[\omega_2(\eta)+\alpha]+\varphi_1[\omega_2(\eta)+\alpha]} - \alpha =$$

$$= \frac{(\eta+\alpha)\{\omega_1[\omega_2(\eta)]+\alpha\} - [\omega_1(\eta)+\alpha][\omega_2(\eta)+\alpha]}{\eta-\omega_1(\eta)-\omega_2(\eta)-\alpha+\omega_1[\omega_2(\eta)]+\alpha} - \alpha =$$

$$= \frac{\eta\omega_1[\omega_2(\eta)] - \omega_1(\eta)\,\omega_2(\eta) + \alpha\{\omega_1[\omega_2(\eta)] - \omega_1(\eta) - \omega_2(\eta) + \eta\}}{\eta-\omega_1(\eta)-\omega_2(\eta)+\omega_1[\omega_2(\eta)]} - \alpha =$$

$$= \frac{\eta\omega_1[\omega_2(\eta)] - \omega_1(\eta)\,\omega_2(\eta)}{\eta-\omega_1(\eta)-\omega_2(\eta)+\omega_1[\omega_2(\eta)]}. \tag{69}$$

Подставляя разложения (68) в (69), получим:

$$\Phi(\eta+\alpha)-\alpha = \frac{\eta\left[\alpha_r^{(1)}\left(\alpha_r^{(2)}\eta^r+\ldots\right)^r+\ldots\right] - \left[\alpha_r^{(1)}\eta^r+\ldots\right]\left[\alpha_r^{(2)}\eta^r+\ldots\right]}{\eta-\left[\alpha_r^{(1)}\eta^r+\ldots\right]-\left[\alpha_r^{(2)}\eta_i^r+\ldots\right]+\left[\alpha_r^{(1)}\left(\alpha_r^{(2)}\eta_i^r+\ldots\right)^r+\ldots\right]} =$$

$$= \frac{\left[\alpha_r^{(1)}\alpha_r^{(2)r}\eta_i^{r^2+1}+\ldots\right] - \left[\alpha_r^{(1)}\alpha_r^{(2)}\eta_i^{2r}+\ldots\right]}{\eta-\left[\alpha_r^{(1)}+\alpha_r^{(2)}\right]\eta_i^r+\alpha_r^{(1)}\alpha_r^{(2)r}\eta_i^{r^2}+\ldots}.$$

При $r > 1$ наименьшая степень η в числителе $2r$, а в знаменателе 1, следовательно, разложение $\Phi(\eta+\alpha)-\alpha$ по степеням η начинается с η^{2r-1}, т. е. $\Phi^{(l)}(\alpha) = 0$ при $l = 1, 2, \ldots, 2r-2$, что означает, что итерация (64) имеет порядок не ниже $2r-1$. При $r = 1$ наименьшая степень η в числителе не меньше трех, так как члены со вторыми степенями взаимно уничтожаются, а в знаменателе η входит с коэффициентом

$$1 - \alpha_r^{(1)} - \alpha_r^{(2)} + \alpha_r^{(1)}\alpha_r^{(2)} = \left(1-\alpha_r^{(1)}\right)\left(1-\alpha_r^{(2)}\right) = \left[1-\varphi_1'(\alpha)\right]\left[1-\varphi_2'(\alpha)\right].$$

Если это произведение не равно нулю, то первая степень в разложении знаменателя присутствует и разложение $\Phi(\alpha+\eta)-\alpha$ по

степеням η начинается по крайней мере с η^2. Следовательно, $\Phi'(\alpha)=0$ и итерация (64) имеет порядок не меньше двух.

В частности, можно положить $\varphi(x)=\varphi_1(x)=\varphi_2(x)$; тогда

$$\Phi(x)=\frac{x\varphi[\varphi(x)]-\varphi^2(x)}{x-2\varphi(x)+\varphi[\varphi(x)]} \tag{70}$$

определяет итерацию не ниже второго порядка, если $\varphi(x)$ определяет итерацию первого, и не ниже $(2r-1)$-го порядка, если $\varphi(x)$ определяет итерацию порядка r.

Заметим, что если итерация, определяемая функцией $\varphi(x)$, не сходится, как бы близко к α мы ни выбирали начальное приближение x_0 (что, например, будет при $|\varphi'(\alpha)|>1$), итерация, определяемая функцией, построенной по формуле (70), будет сходящейся при выборе начального приближения, достаточно близкого к α, так как $\Phi'(\alpha)=0$ и существует окрестность $x=\alpha$, в которой $|\Phi'(x)|\leqslant K<1$, а это является достаточным условием сходимости итерации, если только x_0 взято из этой окрестности.

При построении итерации

$$x_n=\Phi(x_{n-1}) \qquad (n=1,\ 2,\ \ldots),$$

где $\Phi(x)$ определена равенством (70), нет необходимости в явном виде находить $\Phi(x)$, а можно поступать следующим образом. Исходя из x_0, находим:

$$x_1=\varphi(x_0) \quad \text{и} \quad x_2=\varphi(x_1).$$

Затем определяем x_3 с помощью соотношения

$$x_3=\frac{x_0x_2-x_1^2}{x_0-2x_1+x_2}=x_0-\frac{(\Delta x_0)^2}{\Delta^2x_0},$$

где положено

$$\Delta x_i=x_{i+1}-x_i; \quad \Delta^2x_i=x_{i+2}-2x_{i+1}+x_i.$$

Далее, находим:

$$x_4=\varphi(x_3); \quad x_5=\varphi(x_4)$$

и

$$x_6=\frac{x_3x_5-x_4^2}{x_3-2x_4+x_5}$$

и т. д. Получим нестационарной итерационный процесс:

$$\left.\begin{aligned} x_{3i+1}&=\varphi(x_{3i}),\\ x_{3i+2}&=\varphi(x_{3i+1}), \qquad (i=0,\ 1,\ 2,\ \ldots).\\ x_{3(i+1)}&=x_{3i}-\frac{(\Delta x_{3i})^2}{\Delta^2x_{3i}} \end{aligned}\right\} \tag{71}$$

Точно так же как по φ строилась итерация Φ более высокого порядка, можно, исходя из Φ, построить итерацию еще более высокого порядка и т. д.

6. Пример. Проиллюстрируем некоторые из рассмотренных в этом параграфе итерационные методы на примере отыскания корня уравнения

$$f(x) = x^3 + 5x^2 - 15x - 7 = 0,$$

расположенного между $x = 2{,}4$ и $x = 2{,}5$.

1. *Метод секущих.*

$$x_n = \frac{x_0 f(x_{n-1}) - x_{n-1} f(x_0)}{f(x_{n-1}) - f(x_0)} \qquad (n = 2,\ 3,\ \ldots),$$

$x_0 = 2{,}5$, $f(x_0) = 2{,}375$. Начальное приближение $x_1 = 2{,}4$, $f(x_1) = -0{,}376$.

n	x_n	$f(x_n)$	$x_0 f(x_n) - x_n f(x_0)$	$f(x_n) - f(x_0)$	$x_{n+1} = \frac{x_0 f(x_n) - x_n f(x_0)}{f(x_n) - f(x_0)}$
1	2,4	—0,376	—6,6400	—2,751	2,4136677
2	2,4136677	—0,0145312	—5,7687888	—2,3895312	2,4141927
3	2,4141927	—0,0005555	—5,7350965	—2,3755555	2,4142128
4	2,4142128	—0,0000205	—5,7338049	—2,3750205	2,4142128

$$m = \min_{|2{,}4;\ 2{,}5|} |f'(x)| = 28{,}28, \quad |x_4 - \alpha| \leqslant \frac{|f(x_4)|}{m} < 8 \cdot 10^{-7}.$$

2. *Метод Ньютона.*

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})} \qquad (n = 1,\ 2,\ \ldots), \quad x_0 = 2{,}5.$$

n	$f(x_n)$	$f'(x_n)$	$\frac{f(x_n)}{f'(x_n)}$	$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$
0	2,375	28,75	0,0826089	2,4173931
1	0,0843686	27,3043304	0,0031035	2,4142878
2	0,0019769	26,6292347	0,0000742	2,4142136
3	0,0000010	26,6274179	0,0000000	2,4142136

3. *Метод Чебышева третьего порядка.*

$$x_n = x_{n-1} - \frac{f(x_n)}{f'(x_n)} - \frac{f''(x_{n-1}) f^2(x_{n-1})}{2f'^3(x_{n-1})}, \quad x_0 = 2{,}4.$$

n	$f(x_n)$	$f'(x_n)$	$f''(x_n)$	$\frac{f(x_n)}{f'(x_n)}$	$\frac{f(x_n) f^2(x_n)}{2f'^3(x_n)}$	x_{n+1}
0	—0,376	26,28	24,4	—0,0143075	0,0000190	2,4142885
1	0,0019956	26,6292520	24,4857310	0,0000749	0,0000000	2,4142136
2	0,0000012	26,6274179	24,4852816	0,0000000	0,0000000	2,4142136

4. *Метод Эйткена.* Используя первые два приближения, полученные по методу секущих, найдем следующее приближение с помощью равенства:

$$x_4 = x_1 - \frac{(\Delta x_1)^2}{\Delta^2 x_1}.$$

В этом случае $x_1 = 2{,}4$, $x_2 = 2{,}4136677$, $x_3 = 2{,}4141927$,

$$x_4 = 2{,}4 + \frac{(0{,}0136677)^2}{0{,}0131427} = 2{,}4142137.$$

Если использовать первые три приближения, полученных по методу секущих, то

$$x_5 = 2{,}4136677 + \frac{(0{,}0005250)^2}{0{,}0005049} = 2{,}4142136.$$

Значение искомого корня с восемью верными знаками

$$\alpha = 2{,}4142136.$$

Таким образом, все восемь верных знаков мы получили по методу Ньютона после трех шагов, по методу Чебышева третьего порядка — после двух шагов и после трех шагов по методу секущих и последующего уточнения по методу Эйткена.

§ 5. Решение систем уравнений

Решение системы уравнений

$$f_i(x_1, x_2, \ldots, x_n) = 0 \qquad (i = 1, 2, \ldots, n) \tag{1}$$

представляет значительно более сложную задачу, чем решение одного уравнения. Мы опишем только наиболее распространенные методы решения систем.

1. Метод итераций решения систем специального вида. Пусть известно, что система

$$x_i = \varphi_i(x_1, x_2, \ldots, x_n) \qquad (i = 1, 2, \ldots, n) \tag{2}$$

в некоторой области пространства $x_1, x_2, \ldots, x_n$ имеет единственное решение $x_i = \alpha_i$ $(i = 1, 2, \ldots, n)$, а $x_i^{(0)}$ — числа, соответственно близкие к α_i $(i = 1, 2, \ldots, n)$. При некоторых ограничениях на функции $\varphi_i(x_1, x_2, \ldots, x_n)$, исходя из этих приближенных значений, можно найти приближенные значения α_i с наперед заданной точностью. Это уточнение может быть выполнено с помощью *метода итераций*, заключающегося в том, что по $x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}$ находится следующее приближение по формулам:

$$x_i^{(1)} = \varphi_i\left(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}\right) \qquad (i = 1, 2, \ldots, n).$$

По полученным значениям находятся

$$x_i^{(2)} = \varphi_i\left(x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}\right) \qquad (i = 1, 2, \ldots, n)$$

и т. д. Если найдено k-е приближение $x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}$, то $(k+1)$-е приближение находится по формулам:

$$x_i^{(k+1)} = \varphi_i\left(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}\right) \qquad (i = 1, 2, \ldots, n). \tag{3}$$

Если при $k \to \infty$ $x_i^{(k)} \to \alpha_i$ $(i = 1, 2, \ldots, n)$, то говорят, что метод итераций сходится к искомому решению. Для того чтобы получить решение с нужной точностью, практически продолжают процесс до тех пор, пока два последовательных приближения будут совпадать с заданной точностью.

Прежде чем формулировать условия сходимости метода, для удобства записи соотношений и формулировок введем некоторые понятия и обозначения.

Будем рассматривать $x_1, x_2, \ldots, x_n$ как компоненты n-мерного вектора $x = (x_1, x_2, \ldots, x_n)$. Определив норму вектора x равенством

$$\|x\|_1 = \max_{i=1, 2, \ldots, n} |x_i|, \tag{4}$$

можно ввести понятие расстояния между векторами x' и x'', положив

$$\rho(x', x'') = \|x' - x''\|_1 = \max_i |x_i' - x_i''| \tag{5}$$

Определим оператор

$$y = Ax, \tag{6}$$

где $y = (y_1, y_2, \ldots, y_n)$, а $y_i = \varphi_i(x_1, x_2, \ldots, x_n)$.

Если положить $x^{(k)} = \left(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}\right)$, то вместо n равенств (3) можно писать одно векторное равенство:

$$x^{(k+1)} = Ax^{(k)}. \tag{7}$$

Будем говорить, что система функций $\varphi_i(x_1, x_2, \ldots, x_n)$ $(i = 1, 2, \ldots, n)$ удовлетворяет условию Липшица с константой K в некоторой области G, если для любых двух векторов x', $x'' \in G$ имеют место неравенства

$$|\varphi_i(x') - \varphi_i(x'')| \leqslant K\rho(x', x'') \qquad (i = 1, 2, \ldots, n). \tag{8}$$

Достаточное условие сходимости метода итераций для решения системы (2) дает следующая *теорема*:

Если на множестве R всех векторов x, для которых $\rho(x, \alpha) \leqslant r$ $[\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)]$, система функций $\varphi_i(x_1, x_2, \ldots, x_n)$ $(i = 1, 2, \ldots, n)$ удовлетворяет условию Липшица с константой K, меньшей единицы, то при любом начальном векторе $x^{(0)} \in R$ последовательность

$$x^{(k+1)} = Ax^{(k)} \qquad (k = 0, 1, 2, \ldots) \tag{7}$$

сходится к α, *причем*

$$\rho(x^{(k)}, \alpha) \leqslant K^k \rho(x^{(0)}, \alpha). \tag{9}$$

Справедливость этой теоремы легко следует из принципа сжатых отображений. В самом деле, оператор Ax осуществляет сжатое отображение R в себя, так как если $x \in R$ и $y = Ax$, то

$$\rho(y, \alpha) = \rho(Ax, A\alpha) = \max_i |\varphi_i(x) - \varphi_i(\alpha)| \leqslant K\rho(x, \alpha) \leqslant r,$$

т. е. $y \in R$. Далее, если x', $x'' \in R$, то

$$\rho(Ax', Ax'') = \max_i |\varphi_i(x') - \varphi_i(x'')| \leqslant K\rho(x', x''),$$

и так как $K < 1$, то отображение Ax — сжатое отображение R в себя. Множество векторов R является полным метрическим пространством, следовательно, по принципу сжатых отображений в R существует одна и только одна неподвижная точка, т. е. одно и только одно решение уравнения

$$x = Ax,$$

которое будет пределом последовательности (7) при любом векторе $x^{(0)} \in R$. Но так как $\alpha \in R$ и $\alpha = A\alpha$, то α и есть неподвижная точка, т. е.

$$\alpha = \lim_{k \to \infty} x^{(k)}. \tag{10}$$

Далее,

$$\rho(x^{(k)}, \alpha) = \rho(Ax^{(k-1)}, A\alpha) \leqslant K\rho(x^{(k-1)}, \alpha) \tag{11}$$

или

$$\rho(x^{(k)}, \alpha) \leqslant K\rho(x^{(k-1)}, \alpha) \leqslant K^2\rho(x^{(k-2)}, \alpha) \leqslant \ldots \leqslant K^k\rho(x^{(0)}, \alpha),$$

что и доказывает неравенство (9).

Если мы будем *понимать под* R *совокупность векторов* x, *для которых* $\rho(x, y_0) \leqslant r$ (y_0 — фиксированный вектор) *и в* R *система функций* $\varphi_i(x)$ ($i = 1, 2, \ldots, n$) *удовлетворяет условию Липшица с константой* $K < 1$, *а*

$$\rho[Ay_0, y_0] < (1 - K)r, \tag{12}$$

то из теоремы § 3, уточняющей принцип сжатых отображений, следует, что *система* (2) *имеет в* R *единственное решение, которое можно получить методом итераций, исходя из произвольного* $x^{(0)} \in R$.

Предположим теперь, что в некоторой выпуклой области G пространства $x_1, x_2, \ldots, x_n$ функции $\varphi_i(x)$ имеют непрерывные первые производные $\frac{\partial \varphi_i(x)}{\partial x_j}$, $M_{ij} = \max_G \left|\frac{\partial \varphi_i}{\partial x_j}\right|$ и в области G система (2) имеет единственное решение $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$. Предположим далее,

что при некотором начальном приближении $x^{(0)} = \left(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}\right)$ все следующие приближения $x^{(k)} = \left(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}\right)$:

$$x^{(k)} = Ax^{(k-1)} \qquad (k = 1, 2, \ldots) \tag{13}$$

не выходят из области G. Тогда

$$x_i^{(k)} - \alpha_i = \varphi_i\left(x^{(k-1)}\right) - \varphi_i(\alpha) =$$

$$= \sum_{j=1}^{n} \frac{\partial \varphi_i\left(p_i^{(k-1)}\right)}{\partial x_j}\left(x_j^{(k-1)} - \alpha_j\right) \qquad (i = 1, 2, \ldots, n), \tag{14}$$

где $p_i^{(k-1)}$ — некоторая точка отрезка прямой, соединяющей точки $\left(x_1^{(k-1)}, x_2^{(k-1)}, \ldots, x_n^{(k-1)}\right)$ и $(\alpha_1, \alpha_2, \ldots, \alpha_n)$. Обозначим через M_k матрицу

$$\begin{pmatrix} \dfrac{\partial \varphi_1\left(p_1^{(k)}\right)}{\partial x_1} & \dfrac{\partial \varphi_1\left(p_1^{(k)}\right)}{\partial x_2} & \ldots & \dfrac{\partial \varphi_1\left(p_1^{(k)}\right)}{\partial x_n} \\ \dfrac{\partial \varphi_2\left(p_2^{(k)}\right)}{\partial x_1} & \dfrac{\partial \varphi_2\left(p_2^{(k)}\right)}{\partial x_2} & \ldots & \dfrac{\partial \varphi_2\left(p_2^{(k)}\right)}{\partial x_n} \\ \ldots\ldots & \ldots\ldots & \ldots & \ldots\ldots \\ \dfrac{\partial \varphi_n\left(p_n^{(k)}\right)}{\partial x_1} & \dfrac{\partial \varphi_n\left(p_n^{(k)}\right)}{\partial x_2} & \ldots & \dfrac{\partial \varphi_n\left(p_n^{(k)}\right)}{\partial x_n} \end{pmatrix}. \tag{15}$$

Тогда равенства (14) можно коротко записать в виде одного векторного равенства

$$x^{(k)} - \alpha = M_{k-1}\left(x^{(k-1)} - \alpha\right), \tag{16}$$

из которого имеем:

$$x^{(k)} - \alpha = M_{k-1}M_{k-2} \ldots M_1M_0\left(x^{(0)} - \alpha\right). \tag{17}$$

Для того чтобы $x^{(k)} \to \alpha$ при $k \to \infty$, достаточно выполнения условия

$$M_{k-1}M_{k-2} \ldots M_1M_0 \to 0 \quad \text{при} \quad k \to \infty. \tag{18}$$

Но это условие будет выполнено, если $M^k \to 0$ при $k \to \infty$, где

$$M = \begin{pmatrix} M_{11} & M_{12} & \ldots & M_{1n} \\ M_{21} & M_{22} & \ldots & M_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ M_{n1} & M_{n2} & \ldots & M_{nn} \end{pmatrix}, \tag{19}$$

так как элементы матрицы M_i по абсолютной величине не больше соответствующих элементов матрицы M, а отсюда уже следует, что элементы матрицы $M_{k-1}M_{k-2} \ldots M_1M_0$ по абсолютной величине не больше соответствующих элементов матрицы M^k. Но для того чтобы $M^k \to 0$, необходимо и достаточно, чтобы все собственные числа матрицы были по модулю меньше единицы, достаточным же

условием является условие, что какая-нибудь норма матрицы меньше единицы. Если собственные значения матрицы M по модулю меньше единицы, то $x^{(k)} \to \alpha$, а это означает, что если начальное приближение выбрано достаточно близким к α, то все $x^{(k)}$ не будут выходить из области G, и мы будем иметь *теорему*:

Если функции $\varphi_i(x_1, x_2, \ldots, x_n)$ $(i = 1, 2, \ldots, n)$ *в некоторой выпуклой области* G, *содержащей решение* $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ *системы* (2), *непрерывны и имеют непрерывные первые производные, то для сходимости метода итераций достаточно, чтобы у матрицы* (19), *где* $M_{ij} = \max\limits_G \left|\frac{\partial \varphi_i}{\partial x_j}\right|$, *все собственные значения были по модулю меньше единицы, а начальное приближение* $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})$ *достаточно близко к решению* $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$. *В частности, это условие будет выполнено, если какая-нибудь из норм матрицы* M *меньше единицы.*

Из этой теоремы вытекают следующие практически более удобные достаточные признаки сходимости метода итераций.

Для сходимости метода итераций решения системы (2) *достаточно выполнения одного из следующих трех условий*:

$$\sum_{j=1}^{n} M_{ij} < 1 \qquad (i = 1, 2, \ldots, n), \tag{20}$$

$$\sum_{i=1}^{n} M_{ij} < 1 \qquad (j = 1, 2, \ldots, n), \tag{21}$$

$$\sum_{i,\, j=1}^{n} M_{ij}^2 < 1, \tag{22}$$

причем в этих случаях за $x^{(0)}$ *можно принимать любой вектор* $x^{(0)}$ *из окрестности* $\rho(x, \alpha) = \|x - \alpha\|_l \leqslant r$ ($l = 1$ или 2, или 3, в зависимости от того, рассматриваем ли условия (20) или (21), или (22)), *а*

$$\|x - \alpha\|_1 = \max_i |x_i - \alpha_i|; \quad \|x - \alpha\|_2 = \sum_{i=1}^{n} |x_i - \alpha_i|;$$

$$\|x - \alpha\|_3 \leqslant \sqrt{\sum_{i=1}^{n} (x_i - \alpha_i)^2}, \tag{23}$$

если только эта окрестность целиком принадлежит G.

Неравенства (20) — (22) соответственно означают, что первая, вторая или третья нормы матрицы M меньше единицы.

2. Метод Ньютона. Рассмотрим систему n уравнений с n неизвестными

$$f_i(x_1, x_2, \ldots, x_n) = 0 \qquad (i = 1, 2, \ldots, n). \tag{24}$$

Относительно функций $f_i(x)$ предположим, что в некоторой выпуклой области G, содержащей решение $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ системы (24), они имеют непрерывные производные первого порядка и в некоторой окрестности решения α матрица

$$f_x'(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \cdots\cdots\cdots\cdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{pmatrix} \tag{25}$$

не вырождена. При этих условиях решение α системы (24) можно найти следующим образом. Пусть

$$f_x'^{-1}(x) = \begin{pmatrix} g_{11} & g_{12} & \cdots & g_{1n} \\ g_{21} & g_{22} & \cdots & g_{2n} \\ \cdots\cdots\cdots\cdots \\ g_{n1} & g_{n2} & \cdots & g_{nn} \end{pmatrix} \tag{26}$$

— матрица, обратная к $f_x'(x)$. Рассмотрим систему уравнений

$$x_i = \varphi_i(x_1, x_2, \ldots, x_n) \equiv x_i - \sum_{j=1}^{n} g_{ij}(x_1, \ldots, x_n) f_j(x_1, x_2, \ldots, x_n)$$
$$(i = 1, 2, \ldots, n), \tag{27}$$

или коротко в векторной форме

$$x = \varphi(x) \equiv x - f_x'^{-1}(x) f(x). \tag{27'}$$

Решение $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ системы (24) является и решением системы (27), которая имеет специальный вид, рассмотренный в п. 1. Покажем, что α можно найти методом итераций, описанным в п. 1, примененным к системе (27), т. е. покажем, что последовательность

$$x^{(m)} = \varphi(x^{(m-1)}) \equiv x^{(m-1)} - f_x'^{-1}(x^{(m-1)}) f(x^{(m-1)})$$
$$(m = 1, 2, \ldots) \tag{28}$$

сходится к α, если только начальное приближение $x^{(0)}$ взято достаточно близко к α. Для этого воспользуемся очевидным равенством

$$f_i(z_1, z_2, \ldots, z_n) - f_i(y_1, y_2, \ldots, y_n) =$$
$$= \int_0^1 \frac{d}{dt} f_i[y_1 + t(z_1 - y_1), \ldots, y_n + t(z_n - y_n)]\, dt =$$
$$= \int_0^1 \sum_{j=1}^{n} \frac{\partial}{\partial x_j} f_i[y_1 + t(z_1 - y_1), \ldots, y_n + t(z_n - y_n)](z_j - y_j)\, dt. \tag{29}$$

Введем обозначения:

$$F_{ij}(y,\ z)=\int_0^1 \frac{\partial}{\partial x_j} f_i[y_1+t(z_1-y_1),\ \ldots,\ y_n+t(z_n-y_n)]\,dt,$$

$$F(y,\ z)=\begin{pmatrix} F_{11}(y,\ z) & F_{12}(y,\ z) & \ldots & F_{1n}(y,\ z)\\ F_{21}(y,\ z) & F_{22}(y,\ z) & \ldots & F_{2n}(y,\ z)\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ F_{n1}(y,\ z) & F_{n2}(y,\ z) & \ldots & F_{nn}(y,\ z)\end{pmatrix}. \tag{30}$$

Очевидно, что при $y=z=x$

$$F_{ij}(x,\ x)=\frac{\partial f_i(x)}{\partial x_j},\quad F(x,\ x)=f_x(x), \tag{31}$$

$$f(z)-f(y)=F(y,\ z)(z-y). \tag{32}$$

Так как

$$x^{(m)}=\varphi(x^{(m-1)}),\quad \alpha=\varphi(\alpha), \tag{33}$$

то

$$x^{(m)}-\alpha=\varphi(x^{(m-1)})-\varphi(\alpha)=x^{(m-1)}-\alpha-f_x^{-1}(x^{(m-1)})\,f(x^{(m-1)})=$$

$$=x^{(m-1)}-\alpha-f_x^{-1}(x^{(m-1)})[f(x^{(m-1)})-f(\alpha)], \tag{34}$$

ибо

$$f(\alpha)=(f_1(\alpha_1,\ \alpha_2,\ \ldots,\ \alpha_n),\ \ldots,\ f_n(\alpha_1,\ \alpha_2,\ \ldots,\ \alpha_n))=0.$$

Используя (32), можно записать (34) в виде

$$x^{(m)}-\alpha=x^{(m-1)}-\alpha-f_x^{-1}(x^{(m-1)})F(x^{(m-1)},\ \alpha)(x^{(m-1)}-\alpha)=$$

$$=[I-f_x^{-1}(x^{(m-1)})F(x^{(m-1)},\ \alpha)](x^{(m-1)}-\alpha), \tag{35}$$

где I — единичная матрица. Матрица $F(y,\ z)$ является непрерывной функцией своих аргументов, и при $x=y=z$ $F(x,\ x)=f_x(x)$. Если $x^{(0)}$ достаточно близко к α, то $F(x^{(0)},\ x^{(0)})=f_x(x^{(0)})\neq 0$ и $f_x^{-1}(x^0)F(x^{(0)},\ x^{(0)})=I$, а матрица $f_x^{-1}(x^{(0)})F(x^{(0)},\ \alpha)$ достаточно близка к единичной матрице I, а $I-f_x^{-1}(x^0)F(x^0,\ \alpha)$ близка к нулевой матрице. Как бы мы ни определили норму вектора x и соответственно норму матрицы, имеет место неравенство

$$\|x^{(1)}-\alpha\|=\|[I-f_x^{-1}(x^{(0)})F(x^{(0)},\ \alpha)](x^{(0)}-\alpha)\|\leqslant$$

$$\leqslant\|I-f_x^{-1}(x^{(0)})F(x^{(0)},\ \alpha)\|\cdot\|x^{(0)}-\alpha\|. \tag{36}$$

Так как нулевая матрица имеет норму, равную нулю, а

$$I-f_x^{-1}(x^{(0)})F(x^{(0)},\ \alpha)$$

близка к нулевой, то существует такое $\delta > 0$, что если $\rho(x, \alpha) \leqslant \delta$, то

$$\|I - f_x^{-1}(x) F(x, \alpha)\| \leqslant K < 1. \tag{37}$$

Если $\rho(x^{(0)}, \alpha) \leqslant \delta$, то

$$\rho(x^{(1)}, \alpha) = \|x^{(1)} - \alpha\| \leqslant K \|x^{(0)} - \alpha\| = K\rho(x^{(0)}, \alpha) \leqslant K\delta < \delta.$$

Следовательно, и

$$\|I - f_x^{-1}(x^{(1)}, \alpha) F(x^{(1)}, \alpha)\| \leqslant K.$$

Предположим, что мы уже доказали, что

$$\left.\begin{aligned} &\rho(x^{(m)}, \alpha) \leqslant K\rho(x^{(m-1)}, \alpha) < \delta, \\ &\|I - f_x^{-1}(x^{(m)}) F(x^{(m)}, \alpha)\| \leqslant K. \end{aligned}\right\} \tag{38}$$

Тогда

$$\rho(x^{(m+1)}, \alpha) = \|x^{(m+1)} - \alpha\| = \|[I - f_x^{-1}(x^{(m)}) F(x^{(m)}, \alpha)](x^{(m)} - \alpha)\| \leqslant$$

$$\leqslant \|I - f_x^{-1}(x^{(m)}) F(x^{(m)}, \alpha)\| \cdot \|x^{(m)} - \alpha\| \leqslant K\rho(x^{(m)}, \alpha) < \delta,$$

т. е. $\rho(x^{(m+1)}, \alpha) < \delta$ и по (37)

$$\|I - f_x^{-1}(x^{(m+1)}) F(x^{(m+1)}, \alpha)\| \leqslant K.$$

Таким образом, неравенства (38) справедливы при всех m. Последовательно применяя их, получим неравенство

$$\rho(x^{(m)}, \alpha) \leqslant K^m \rho(x^{(0)}, \alpha). \tag{39}$$

Так как $K < 1$, то при $m \to \infty$ $K^m \to 0$ и

$$\lim_{m \to \infty} \rho(x^{(m)}, \alpha) = 0, \tag{40}$$

что и доказывает сходимость процесса итераций к решению системы (24). На начальное приближение $x^{(0)}$ накладывается условие, что оно должно принадлежать окрестности $\rho(x, \alpha) \leqslant \delta$, в которой имеет место неравенство (37).

При более жестких ограничениях на функции $f_i(x_1, x_2, \ldots, x_n)$ можно доказать более сильную теорему, доказательство которой можно найти в статье Л. В. Канторовича «Функциональный анализ и прикладная математика» (УМН, т. 3, вып. 6, 1948):

Теорема. *Если в области G функции $f_i(x_1, x_2, \ldots, x_n)$ имеют вторые производные, не превосходящие по абсолютной величине числа L, в точке $x^{(0)} \in G$ матрица $f_x(x)$ не вырождена и выполнено условие*

$$h = M^2 L \delta n^2 \leqslant \frac{1}{2}, \tag{41}$$

где

$$\left.\begin{aligned} &\left|f_i\left(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}\right)\right| \leqslant \delta \qquad (i=1, 2, \ldots, n), \\ &\|f_x^{-1}(x_0)\|_1 = \max_i \sum_{j=1}^{n} |g_{ij}(x^{(0)})| \leqslant M, \end{aligned}\right\} \tag{42}$$

то система (24) *имеет решение* $x=\alpha$, *которое находится в области*

$$\|x - x^{(0)}\|_1 = \max_i |x_1 - x_i^{(0)}| \leqslant \frac{1-\sqrt{1-2h}}{h}\,\delta \tag{43}$$

и может быть получено как предел последовательности

$$x^{(m)} = x^{(m-1)} - f_x^{-1}(x^{(m-1)})\, f(x^{(m-1)}) \tag{28}$$

и быстрота сходимости оценивается неравенством

$$\|x^{(m)} - \alpha\|_1 \leqslant \frac{1}{2^{m-1}}(2h)^{2^m-1}\delta. \tag{44}$$

Векторное равенство (28) эквивалентно системе линейных алгебраических уравнений

$$\sum_{j=1}^{n} \frac{\partial}{\partial x_j} f_i(x^{(m-1)})\left(x_j^{(m)} - x_j^{(m-1)}\right) = -f_i(x^{(m-1)}) \qquad (i=1, 2, \ldots, n). \tag{45}$$

Вычисление последовательных приближений с помощью равенства (28) или путем решения системы (45) связано с большой вычислительной работой, так как на каждом шаге нужно решать систему со своей матрицей или на каждом шаге находить матрицу $f_x^{-1}(x^{(m)})$. В связи с этим вместо рассмотренного метода решения системы (24), который носит название *метода Ньютона*, иногда применяют следующий более простой метод.

Вместо системы (27) рассматривают систему

$$x_i = \psi_i(x_1, x_2, \ldots, x_n) \equiv x_i - \sum_{j=1}^{n} g_{ij}\left(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}\right) \times \\ \times f_j(x_1, x_2, \ldots, x_n) \qquad (i=1, 2, \ldots, n), \tag{46}$$

или в векторной записи

$$x = \psi(x) \equiv x - f_x^{-1}(x^{(0)})\, f(x), \tag{46'}$$

где $x^{(0)}$ достаточно близок к решению α системы (24), а следовательно и (46), которую также решают методом итераций, описанным в п. 1.

От функций $f_i(x_1, x_2, \ldots, x_n)$ будем требовать выполнения тех же условий, которые накладывались в начале этого пункта. Используя (32), можно записать:

$$\psi(z) - \psi(y) = z - y - f_x^{-1}(x^{(0)})[f(z) - f(y)] = z - y - f_x^{-1}(x^{(0)}) \times \\ \times F(z, y)(z - y) = [I - f_x^{-1}(x^{(0)}) F(z, y)](z - y). \quad (47)$$

Если y и z близки к $x^{(0)}$, то матрица $F(z, y)$ близка к матрице $f_x(x^{(0)})$, а $I - f_x^{-1}(x^{(0)}) F(z, y)$ близка к нулевой матрице, т. е. для некоторого $K < 1$ можно найти такое $r > 0$, что при $\rho(y, x^{(0)}) = \|y - x^{(0)}\|_1 \leqslant r$ и $\rho(z, x^{(0)}) \leqslant r$ будет иметь место неравенство

$$\|I - f_x^{-1}(x^{(0)}) F(z, y)\|_1 \leqslant K, \quad (48)$$

а из (47)

$$\|\psi(z) - \psi(y)\|_1 \leqslant \|I - f_x^{-1}(x^{(0)}) F(z, y)\|_1 \|z - y\|_1 \leqslant K \|y - z\|_1, \quad (49)$$

т. е. будет иметь место условие Липшица с константой $K < 1$. Далее, из (46′)

$$\|\psi(x^{(0)}) - x^{(0)}\|_1 \leqslant \|f_x^{-1}(x^{(0)})\|_1 \|f(x^{(0)})\|_1 \quad (50)$$

и выбирая $x^{(0)}$ достаточно близким к α [$\rho(x^{(0)}, \alpha) \leqslant r$], можно добиться выполнения неравенства

$$\|\psi(x^{(0)}) - x^{(0)}\|_1 < (1 - K) r, \quad (51)$$

а выполнение неравенств (49) и (51), как это было показано в п. 1, гарантирует, что в окрестности $\rho(x, x^{(0)}) = \|x - x^{(0)}\|_1 \leqslant r$ существует единственное решение системы (46′), которое может быть получено как предел последовательности

$$x^{(m)} = \psi(x^{(m-1)}) \equiv x^{(m-1)} - f_x^{-1}(x^{(0)}) f(x^{(m-1)}) \qquad (m = 2, 3, \ldots), \quad (52)$$

где за начальное приближение $x^{(1)}$ можно взять любую точку указанной окрестности. Но этим решением и будет $x = \alpha$, т. е.

$$\lim_{m \to \infty} x^{(m)} = \alpha, \quad (53)$$

и сходимость метода доказана.

В цитированной выше статье Л. В. Канторовича показано, что в условиях теоремы, которую мы приводили выше, быстрота сходимости последовательности $x^{(m)}$ к α определяется неравенством

$$\|x^{(m)} - \alpha\|_1 \leqslant q^{m-1} \|x^{(1)} - \alpha\|_1 \qquad (q = 1 - \sqrt{1 - 2h} < 1). \quad (54)$$

Этот процесс, который можно рассматривать как видоизменение метода Ньютона, хотя сходится и медленней, чем процесс Ньютона,

имеет то преимущество, что последовательные приближения по (52) находятся значительно проще, так как достаточно один раз найти матрицу $f_x^{-1}(x^{(0)})$.

Хотя мы и обосновали теоретически сходимость метода Ньютона и его видоизменения, при их фактическом применении неизбежные в процессе счета ошибки округления все же могут привести к тому, что решение с заданной наперед точностью не удастся. Вопросы влияния ошибок округления на точность результата в общем случае еще не изучены.

Пример. Уточнить по методу Ньютона приближенные значения решения $x_0 = 0{,}4$, $y_0 = 0{,}9$ системы

$$f_1(x, y) \equiv 4x^2 + y^2 + 2xy - y - 2 = 0,$$

$$f_2(x, y) \equiv 2x^2 + 3xy + y^2 - 3 = 0.$$

Для этого будем последовательно решать системы:

$$f'_{1x}(x_{m-1}, y_{m-1})\,\Delta x_{m-1} + f'_{1y}(x_{m-1}, y_{m-1})\,\Delta y_{m-1} = -f_1(x_{m-1}, y_{m-1}),$$

$$f'_{2x}(x_{m-1}, y_{m-1})\,\Delta x_{m-1} + f'_{2y}(x_{m-1}, y_{m-1})\,\Delta y_{m-1} = -f_2(x_{m-1}, y_{m-1}),$$

где

$$f'_{1x} = 8x + 2y, \quad f'_{1y} = 2x + 2y - 1,$$
$$f'_{2x} = 4x + 3y, \quad f'_{2y} = 3x + 2y;$$

$$f_1(x_0, y_0) = -0{,}73, \quad f'_{1x}(x_0, y_0) = 5{,}0, \quad f'_{1y}(x_0, y_0) = 1{,}6,$$
$$f_2(x_0, y_0) = -0{,}79, \quad f'_{2x}(x_0, y_0) = 4{,}3, \quad f'_{2y}(x_0, y_0) = 3{,}0;$$

$$5{,}0\,\Delta x_0 + 1{,}6\,\Delta y_0 = 0{,}73,$$
$$4{,}3\,\Delta x_0 + 3{,}0\,\Delta y_0 = 0{,}79;$$

$$\Delta x_0 = 0{,}114, \quad x_1 = x_0 + \Delta x_0 = 0{,}514,$$
$$\Delta y_0 = 0{,}100, \quad y_1 = y_0 + \Delta y_0 = 1{,}000;$$

$$f_1(x_1, y_1) = 0{,}084784, \quad f'_{1x}(x_1, y_1) = 6{,}112, \quad f'_{1y}(x_1, y_1) = 2{,}028,$$
$$f_2(x_1, y_1) = 0{,}070392, \quad f'_{2x}(x_1, y_1) = 5{,}056, \quad f'_{2y}(x_1, y_1) = 3{,}542;$$

$$6{,}112\,\Delta x_1 + 2{,}028\,\Delta y_1 = -0{,}084784,$$
$$5{,}056\,\Delta x_1 + 3{,}542\,\Delta y_1 = -0{,}70392;$$

$$\Delta x_1 = -0{,}013826, \quad x_2 = 0{,}500174,$$
$$\Delta y_1 = -0{,}000138, \quad y_2 = 0{,}999862;$$

$$f_1(x_2, y_2) = 0{,}000768,\quad f'_{1x}(x_2, y_2) = 6{,}001116,\quad f'_{1y}(x_2, y_2) = 2{,}000072,$$
$$f_2(x_2, y_2) = 0{,}000387,\quad f'_{2x}(x_2, y_2) = 5{,}000282,\quad f'_{2y}(x_2, y_2) = 3{,}500246;$$
$$6{,}001116\,\Delta x_2 + 2{,}000072\,\Delta y_2 = -0{,}000768,$$
$$5{,}000282\,\Delta x_2 + 3{,}500246\,\Delta y_2 = -0{,}000387;$$
$$\Delta x_2 = -0{,}000174,\quad x_3 = 0{,}500000,$$
$$\Delta y_2 = \quad 0{,}000138,\quad y_3 = 1{,}000000;$$
$$f_1(x_3, y_3) = 0,\quad f_2(x_3, y_3) = 0.$$

Таким образом, мы получили точное решение.

3. Метод скорейшего спуска. Кратко остановимся на методе скорейшего спуска. В этом методе решение системы (24) сводится к задаче отыскания минимумов функции $\Phi(x_1, x_2, \ldots, x_n)$, которую можно построить различными способами, положив, например,

$$\Phi(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} f_i^2(x_1, x_2, \ldots, x_n), \tag{55}$$

или

$$\Phi(x_1, x_2, \ldots, x_n) = \sum_{i,\,j=1}^{n} a_{ij} f_i(x_1, x_2, \ldots, x_n)\, f_j(x_1, x_2, \ldots, x_n), \tag{56}$$

где a_{ij} — элементы некоторой положительно определенной матрицы.

Если $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ есть некоторое решение системы (24), то $f_i(\alpha) = 0$ $(i = 1, 2, \ldots, n)$ и $\Phi(\alpha) = 0$. В других точках x $\Phi(x) > 0$. Таким образом, каждый нулевой минимум функции $\Phi(x)$ даст решение системы (24) и отыскание решений системы (24) сводится к отысканию нулевых минимумов вспомогательной функции $\Phi(x)$. Метод скорейшего спуска отыскания последних заключается в следующем. Если известно примерное расположение нулевого минимума, то выбираем вектор $x^{(0)} = (x_1^{(0)}, \ldots, x_n^{(0)})$, близкий к α, вычисляем производные $\dfrac{\partial \Phi(x^{(0)})}{\partial x_i}$ и в направлении вектора

$$\Phi_x(x^{(0)}) = \operatorname{grad} \Phi(x^{(0)}) = \left(\frac{\partial \Phi(x^{(0)})}{\partial x_1}, \frac{\partial \Phi(x^{(0)})}{\partial x_2}, \ldots, \frac{\partial \Phi(x^{(0)})}{\partial x_n}\right) \tag{57}$$

проводим прямую

$$x = x^{(0)} - \lambda \Phi_x(x^{(0)}), \tag{58}$$

проходящую через точку $x^{(0)}$ в направлении вектора $\Phi_x(x^{(0)})$, ортогонального к поверхности $\Phi(x) = \Phi(x^{(0)})$. Определяем λ_0 и $x^{(1)} = x^{(0)} - \lambda_0 \Phi_x(x^{(0)})$ из условия минимума функции

$$\psi_0(\lambda) = \Phi(x^{(0)} - \lambda \Phi_x(x^{(0)})).$$

Если $\Phi(x^{(1)}) \neq 0$, то продолжаем процесс, исходя из $x^{(1)}$ и двигаясь в направлении вектора $\operatorname{grad} \Phi(x^{(1)}) = \Phi_x(x^{(1)})$, и снова на прямой $x = x^{(1)} - \lambda\Phi_x(x^{(1)})$ отыскиваем точку, в которой

$$\psi_1(\lambda) = \Phi(x^{(1)} - \lambda\Phi_x(x^{(1)}))$$

имеет минимальное значение. Если уже найдено k-е приближение $x^{(k)}$, то λ_k находим из условия минимума функции

$$\psi_k(\lambda) = \Phi(x^{(k)} - \lambda\Phi_x(x^{(k)})) \tag{59}$$

и полагаем

$$x^{(k+1)} = x^{(k)} - \lambda_k\Phi_x(x^{(k)}). \tag{60}$$

На каждом шаге придется решать уравнение

$$\psi_k'(\lambda) = 0 \tag{61}$$

с одним неизвестным λ, что может быть выполнено одним из описанных выше методов.

Реализуя этот метод, мы на каждом шаге движемся в направлении быстрейшего убывания функции Φ. Если начальное приближение выбрано достаточно хорошо и в окрестности искомого решения α нет других минимумов, то этот процесс быстро даст искомое решение с заданной точностью. Если в окрестности α имеются другие минимумы, то при неудачном выборе $x^{(0)}$ процесс сойдется, но не приведет к искомому решению.

Применение метода скорейшего спуска на каждом шаге требует выполнения большой вычислительной работы. Поэтому, вместо того чтобы двигаться в направлении градиента $\Phi(x)$, можно двигаться из $x^{(0)}$ в направлении какого-либо другого вектора, не касательного к поверхности $\Phi(x) = \Phi(x^{(0)})$. Проще всего брать векторы в направлении координатных осей. Так, в *релаксационном методе*, имея начальное приближение $x^{(0)}$, вычисляют производные $\frac{\partial\Phi(x^{(0)})}{\partial x_j}$, и если $\frac{\partial\Phi(x^{(0)})}{\partial x_i}$ — наибольшая из них, то $x^{(1)}$ находят из условия

$$\frac{\partial}{\partial x_i}\Phi(x^{(0)} - \lambda l_i) = 0, \tag{62}$$

где l_i — вектор в направлении i-й координатной оси. Это равносильно уточнению i-й неизвестной при неизменных остальных.

§ 6. Отыскание корней алгебраических уравнений методом выделения множителей

Известно, что многочлен

$$f(x) = x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n \tag{1}$$

с действительными коэффициентами может быть представлен в виде произведения многочленов степени не выше двух тоже с действи-

тельными коэффициентами. Линейные множители в этом разложении соответствуют действительным корням уравнения

$$f(x) = x^n + a_1 x^{n-1} + \ldots + a_{n-1}x + a_n = 0, \tag{2}$$

а квадратичные — парам комплексно-сопряженных корней. Таким образом, имея способы разложения многочлена на множители, мы сведем задачу отыскания корней уравнения (2) к решению совсем простых уравнений. В связи с этим разработаны методы выделения действительных множителей многочлена $f(x)$.

При применении методов выделения множителей приходится выполнять многократно деление многочлена на многочлен, т. е. находить частное и остаток. Если делитель имеет первую степень, то это удобно выполнять по *схеме Горнера*. Пусть требуется найти частное и остаток от деления многочлена

$$f(x) = a_0 x^n + a_1 x^{n-1} + \ldots + a_{n-1}x + a_n \tag{3}$$

на многочлен $x - r$. Обозначим остаток от деления через R, а частным пусть будет многочлен

$$g(x) = b_0 x^{n-1} + b_1 x^{n-2} + \ldots + b_{n-2}x + b_{n-1}. \tag{4}$$

Тогда

$$f(x) = g(x)(x - r) + R. \tag{5}$$

Сравнение коэффициентов при одинаковых степенях x в правой и левой частях тождества (5) дает

$$a_0 = b_0, \quad a_k = b_k - b_{k-1}r \quad (k = 1, 2, \ldots, n-1), \quad a_n = R - b_{n-1}r, \tag{6}$$

откуда

$$b_0 = a_0, \quad b_k = a_k + b_{k-1}r \quad (k = 1, 2, \ldots, n-1), \quad R = a_n + b_{n-1}r. \tag{7}$$

Вычисления удобно располагать по следующей схеме — *схеме Горнера*:

$$\begin{array}{cccccccc} a_0 & a_1 & a_2 & a_3 & \ldots & a_{n-1} & a_n & \big| r \\ & b_0 r & b_1 r & b_2 r & \ldots & b_{n-2} r & b_{n-1} r & \\ \hline b_0 & b_1 & b_2 & b_3 & \ldots & b_{n-1} & R & \end{array}$$

где $b_0 = a_0$, а каждое число нижней строки равно сумме двух чисел, стоящих над ним.

Так как из (5) видно, что $R = f(r)$, то схему Горнера удобно применять для отыскания значений многочлена $f(x)$ при $x = r$.

Для отыскания частного

$$g(x) = b_0 x^{n-2} + b_1 x^{n-3} + \ldots + b_{n-3}x + b_{n-2}$$

и остатка

$$R(x) = c_0 x + c_1$$

от деления многочлена (3) на множитель x^2+px+q можно использовать следующую схему:

$$\begin{array}{c|ccccccc} & a_0 & a_1 & a_2 & a_3 & \dots & a_{n-2} & a_{n-1} & a_n \\ -p & & -pb_0 & -pb_1 & -pb_2 & \dots & -pb_{n-3} & -pb_{n-2} & \\ -q & & & -qb_0 & -qb_1 & \dots & -qb_{n-4} & -qb_{n-3} & -qb_{n-2} \\ \hline & b_0 & b_1 & b_2 & b_3 & \dots & b_{n-2} & c_0 & c_1 \end{array}$$

где последняя строка получается как сумма первых трех строк.

Эта схема просто получается из тождества

$$f(x)=(x^2+px+q)\,g(x)+R(x) \tag{8}$$

сравнением коэффициентов при одинаковых степенях x. Это дает

$$a_0=b_0,\quad a_1=b_1+pb_0,\quad a_k=b_k+pb_{k-1}+qb_{k-2}\quad (k=2,\ 3,\ \dots,\ n-2,$$
$$a_{n-1}=c_0+pb_{n-2}+qb_{n-3},\quad a_n=qb_{n-2}+c_1,$$

откуда следует:

$$\left.\begin{array}{c} b_0=a_0,\quad b_1=a_1-pb_0, \\ b_k=a_k-pb_{k-1}-qb_{k-2}\qquad (k=2,\ 3,\ \dots,\ n-2), \\ c_0=a_{n-1}-pb_{n-2}-qb_{n-3},\quad c_1=a_n-qb_{n-2}, \end{array}\right\} \tag{9}$$

что и реализовано в схеме.

Нетрудно сообразить, как будет выглядеть схема для определения коэффициентов частного и остатка при делении многочлена (3) на многочлен $x^m+d_1x^{m-1}+\dots+d_{m-1}x+d_m\qquad (m<n)$.

1. Метод Лина выделения множителей. Метод Лина, или метод предпоследнего остатка, выделения множителя $g_m(x)$ степени m из многочлена

$$f_n(x)=x^n+a_1x^{n-1}+\dots+a_{n-1}x+a_n \tag{10}$$

состоит в следующем. За начальное приближение $g_m(x)$ берется некоторый многочлен степени m

$$g_{m,\,1}(x)=x^m+b_1^{(1)}x^{m-1}+\dots+b_{m-1}^{(1)}x+b_m^{(1)} \tag{11}$$

и производится деление $f_n(x)$ на $g_{m,\,1}(x)$ до тех пор, пока в остатке получится многочлен степени m — предпоследний остаток. За следующее приближение $g_{m,\,2}(x)$ берется этот остаток, деленный на коэффициент при x^m — приведенный предпоследний остаток, далее процесс повторяется, т. е. если уже найдено k-е приближение к $g_m(x)$:

$$g_{m,\,k}(x)=x^m+b_1^{(k)}x^{m-1}+\dots+b_{m-1}^{(k)}x+b_m^{(k)}, \tag{12}$$

то $g_{m,\,k+1}(x)$ определяется как приведенный предпоследний остаток от деления многочлена $f_n(x)$ на $g_{m,\,k}(x)$. Процесс продолжают до

тех пор, пока коэффициенты двух последовательных приближений будут совпадать в пределах заданной точности.

Практически приемлемых критериев сходимости этого метода в общем случае нет. Приведем без доказательства один результат, практическая ценность которого невелика.

Если $\alpha_1, \alpha_2, \ldots, \alpha_m$ *— корни выделяемого множителя* $g_m(x)$, *а* $P_{n-m}(x) = x^{n-m} + c_1 x^{n-m-1} + \ldots + c_{n-m-1}x + c_{n-m}$ *— частное от деления* $f_n(x)$ *на* $g_m(x)$, т. е. $f_n(x) = g_m(x) P_{n-m}(x)$, *то сходимость метода Лина будет иметь место в случае, когда*

$$\rho_i = 1 - \frac{P_{n-m}(\alpha_i)}{c_{n-m}} \qquad (i = 1, 2, \ldots, m) \tag{13}$$

по модулю меньше единицы и начальное приближение $g_{m,1}(x)$ *выбрано достаточно близко к* $g_m(x)$. В случае, если наибольшее из них по модулю есть действительное число, то $b_j^{(k+1)} - b_j^{(k)}$ сходятся к нулю со скоростью геометрической прогрессии со знаменателем, равным модулю этого числа.

При практическом применении метода Лина чаще всего выделяют линейные и квадратичные множители, так как только эти множители заранее можно определить достаточно точно и применять процесс Лина для их уточнения. Но только в случае выделения линейных множителей, соответствующих действительным корням уравнения $f_n(x) = 0$, можно дать простые критерии сходимости метода.

Пусть выделяется множитель $x - \alpha$ и $x - \alpha^{(k)}$ — его k-е приближение, тогда имеет место тождество

$$f_n(x) = x^n + a_1 x^{n-1} + \ldots + a_{n-1}x + a_n =$$

$$= \left(x - \alpha^{(k)}\right)\left(x^{n-1} + c_1^{(k)} x^{n-2} + \ldots + c_{n-2}^{(k)} x\right) + d_0^{(k)}\left(x - \alpha^{(k+1)}\right), \tag{14}$$

где $x - \alpha^{(k+1)}$ есть $k+1$-е приближение. Из этого тождества имеем:

$$f_n\left(\alpha^{(k)}\right) = d_0^{(k)}\left(\alpha^{(k)} - \alpha^{(k+1)}\right); \quad f_n(0) = -d_0^{(k)}\alpha^{(k+1)}$$

или

$$\frac{\alpha^{(k+1)} - \alpha^{(k)}}{\alpha^{(k+1)}} = \frac{f_n\left(\alpha^{(k)}\right)}{f_n(0)},$$

откуда

$$\alpha^{(k+1)} = \frac{f_n(0)\,\alpha^{(k)}}{f_n(0) - f_n\left(\alpha^{(k)}\right)}. \tag{15}$$

Это равенство можно рассматривать как итерацию для отыскания корня α уравнения

$$x = \varphi(x) \equiv \frac{x f_n(0)}{f_n(0) - f_n(x)}. \tag{16}$$

Так как

$$\varphi'(\alpha) = 1 + \alpha \frac{f_n'(\alpha)}{f_n(0)}, \tag{17}$$

то при выполнении условия $|\varphi'(\alpha)|<1$ найдется некоторая окрестность корня $x=\alpha$, в которой будет иметь место неравенство

$$|\varphi(x)|\leqslant K<1$$

и итерация (15) будет сходиться, если начальное приближение $\alpha^{(1)}$ взято из этой окрестности. Условие $|\varphi'(\alpha)|<1$ будет иметь место, если α — наименьший по модулю корень уравнения $f_n(x)=0$, при условии, что все корни уравнения действительны и одного знака. В самом деле, если в этом случае расположить все корни в порядке возрастания их модулей, т. е.

$$|\alpha_1|<|\alpha_2|\leqslant\ldots\leqslant|\alpha_n|,$$

то

$$f_n(0)=(-1)^n\alpha_1\alpha_2\ldots\alpha_n;\quad f_n'(\alpha_1)=(\alpha_2-\alpha_1)(\alpha_3-\alpha_1)\ldots(\alpha_n-\alpha_1)$$

и

$$\varphi'(\alpha_1)=1+\frac{\alpha_1 f_n'(\alpha_1)}{f_n(0)}=1-\left(1-\frac{\alpha_1}{\alpha_2}\right)\left(1-\frac{\alpha_1}{\alpha_3}\right)\ldots\left(1-\frac{\alpha_1}{\alpha_n}\right),\quad(18)$$

и так как все отношения $\frac{\alpha_1}{\alpha_i}$ $(i=2,\ 3,\ \ldots,\ n)$ положительны и меньше единицы, то

$$0<1+\frac{\alpha_1 f_n'(\alpha_1)}{f_n(0)}<1.\quad(19)$$

Пример. Выделить множитель второй степени из многочлена

$$x^4+7x^3+24x^2+25x-15.$$

В качестве начального приближения возьмем многочлен

$$g_{2,1}(x)=x^2.$$

Не приводя промежуточных вычислений, которые сводятся к делению многочлена $f_n(x)=x^n+a_1x^{n-1}+\ldots+a_{n-1}x+a_n$ на трехчлен $x^2+b_1^{(k)}x+b_2^{(k)}$ до получения предпоследнего остатка, что можно выполнить по схеме:

	1	a_1	a_2	a_3	$\ldots$	a_{n-3}	a_{n-2}	a_{n-1}	a_n
$-b_1^{(k)}$		$-b_1^{(k)}$	$-b_1^{(k)}c_1^{(k)}$	$-b_1^{(k)}c_2^{(k)}$	$\ldots$	$-b_1^{(k)}c_{n-4}^{(k)}$	$-b_1^{(k)}c_{n-3}^{(k)}$		
$-b_2^{(k)}$			$-b_2^{(k)}$	$-b_2^{(k)}c_1^{(k)}$	$\ldots$	$-b_2^{(k)}c_{n-5}^{(k)}$	$-b_2^{(k)}c_{n-4}^{(k)}$	$-b_2^{(k)}c_{n-3}^{(k)}$	
	1	$c_1^{(k)}$	$c_2^{(k)}$	$c_3^{(k)}$	$\ldots$	$c_{n-3}^{(k)}$	$d_0^{(k)}$	$d_1^{(k)}$	$d_2^{(k)}$,

где последняя строка есть сумма первых трех, при этом $x^{n-2}+c_1^{(k)}x^{n-3}+\ldots+c_{n-3}^{(k)}x$ — предпоследнее частное, а $d_0^{(k)}x^2+d_1^{(k)}x+d_2^{(k)}$ — предпоследний остаток, приведем результаты вычислений для данного примера.

k	Приведенный предпоследний остаток			Приведенное частное		Предпоследний остаток		
		$b_1^{(k)}$	$b_2^{(k)}$		$c_1^{(k)}$	$a_0^{(k)}$	$a_1^{(k)}$	$a_2^{(k)}$
1	1	1,042	— 0,625	1	7	24	25	— 15
2	1	1,5597	— 0,8145	1	5,957	18,4168	27,7238	— 15
3	1	1,8024	— 0,9186	1	5,4403	16,3293	29,4311	— 15
4	1	1,9147	— 0,9646	1	5,1976	15,5504	29,7745	— 15
5	1	1,9639	— 0,9850	1	5,0853	15,2278	29,9053	— 15
6	1	1,9849	— 0,9937	1	5,0361	15,0946	29,9606	— 15
7	1	1,9937	— 0,9974	1	5,0151	15,0392	29,9835	— 15
8	1	1,9974	— 0,9989	1	5,0063	15,0163	29,9933	— 15
9	1	1,9989	— 0,9996	1	5,0026	15,0067	29,9971	— 15
10	1	1,9996	— 0,9998	1	5,0011	15,0029	29,9991	— 15
11	1	1,9998	— 0,9999	1	5,0004	15,0010	29,9994	— 15
12	1	1,9999	— 0,9999	1	5,0002	15,0005	29,9997	— 15

Последнее частное в общем случае имеет вид

$$x^{n-2} + c_1^{(k)} x^{n-3} + \ldots + c_{n-3}^{(k)} x + d_0^{(k)}$$

и будет являться приближенным представлением второго множителя, дополняющего искомый множитель $g_m(x)$ до $f_n(x)$. В нашем примере после 12 шагов имеем:

$$x^4 + 7x^3 + 24x^2 + 25x - 15 \approx$$
$$\approx (x^2 + 1{,}9999x - 0{,}9999)(x^2 + 5{,}0002x + 15{,}0005).$$

Точное разложение имеет вид

$$x^4 + 7x^3 + 24x^2 + 25x - 15 = (x^2 + 2x - 1)(x^2 + 5x + 15).$$

Получили хорошее приближение, хотя начальный множитель x^2 далек от истинного.

Если использовать полученное разложение для отыскания корней уравнения

$$x^4 + 7x^3 + 24x^2 - 15 = 0,$$

то получим для корней следующие значения:

$$\tilde{x}_1 = -2{,}4141, \quad \tilde{x}_2 = 0{,}4142, \quad \tilde{x}_{3,4} = -2{,}5001 \pm 2{,}9581i.$$

Истинные же значения корней с четырьмя верными десятичными знаками:

$$x_1 = -2{,}4142, \quad x_2 = 0{,}4142, \quad x_{3,4} = -2{,}5000 \pm 2{,}9581i.$$

2. Метод Фридмана. Метод Фридмана выделения множителя $g_m(x)$ многочлена $f_n(x)$ не обладает таким единообразием, как метод Лина. В методе Фридмана, если $g_{m,k}(x)$ есть k-е приближение искомого

множителя $g_m(x)$, для получения $(k+1)$-го приближения поступают следующим образом: делят многочлен $f_n(x)$ на $g_{m,k}(x)$ так же, как и в методе Лина, но только до получения последнего остатка; полученное частное располагают по возрастающим степеням x и делят на него многочлен $f_n(x)$, расположенный также по возрастающим степеням x, до тех пор, пока в частном получится многочлен степени m; это частное, деленное на коэффициент при x^m, и принимают за $g_{m,k+1}(x)$.

Отыскание каждого приближения по методу Фридмана требует более чем в два раза больше операций, чем в методе Лина, но в некоторых случаях метод Фридмана имеет значительно лучшую сходимость, чем метод Лина. Это можно проиллюстрировать на примере выделения линейного множителя, соответствующего наименьшему по абсолютной величине корню уравнения $f_n(x)=0$, если это уравнение имеет только действительные корни и одного знака. В самом деле, если искомый линейный множитель имеет вид $x-\alpha$, а k-е приближение его есть

$$g_{1,k}(x)=x-\alpha^{(k)}, \tag{20}$$

то в результате деления $f_n(x)$ на $g_{1,k}(x)$ получим:

$$\frac{f_n(x)}{x-\alpha^{(k)}}=x^{n-1}+b_1^{(k)}x^{n-2}+\ldots+b_{n-2}^{(k)}x+b_{n-1}^{(k)}+\frac{r_0^{(k)}}{x-\alpha^{(k)}}, \tag{21}$$

откуда

$$\left.\begin{aligned} &r_0^{(k)}=f_n(\alpha^{(k)}); \quad b_{n-1}^{(k)}=\frac{f_n(\alpha^{(k)})-f_n(0)}{\alpha^{(k)}};\\ &b_{n-2}^{(k)}=\lim_{x\to 0}\frac{f_n(x)-r_0^{(k)}-b_{n-1}^{(k)}(x-\alpha^{(k)})}{x(x-\alpha^{(k)})}=\frac{b_{n-1}^{(k)}-f_n'(0)}{\alpha^{(k)}}=\\ &\qquad=\frac{f_n(\alpha^{(k)})-f_n(0)-\alpha^{(k)}f_n'(0)}{\alpha^{(k)\,2}}. \end{aligned}\right\} \tag{22}$$

При делении многочлена $f_n(x)$, расположенного по возрастающим степеням, на многочлен $q_k(x)=b_{n-1}^{(k)}+b_{n-2}^{(k)}x+\ldots+b_1^{(k)}x^{n-2}+x^{n-1}$ до получения частного первой степени находим следующее частное:

$$\frac{a_n}{b_{n-1}^{(k)}}+\frac{1}{b_{n-1}^{(k)}}\left(a_{k-1}-\frac{a_n b_{n-2}^{(k)}}{b_{n-1}^{(k)}}\right)x,$$

которое после деления на коэффициент при x приобретет вид

$$g_{1,k+1}(x)=x-\alpha^{(k+1)}\equiv x-\frac{a_n b_{n-1}^{(k)}}{a_n b_{n-2}^{(k)}-a_{n-1}b_{n-1}^{(k)}}, \tag{23}$$

откуда, принимая во внимание равенства (22), будем иметь:

$$\alpha^{(k+1)} = \frac{f_n(0)\,[f_n(\alpha^{(k)}) - f_n(0)]\,\alpha^{(k)}}{f_n(0)\,[f_n(\alpha^{(k)}) - f_n(0)] - f_n'(0)\,f_n(\alpha^{(k)})\,\alpha^{(k)}}. \tag{24}$$

Это — итерация для решения уравнения

$$x = \psi(x) \equiv \frac{f_n(0)\,[f_n(x) - f_n(0)]\,x}{f_n(0)\,[f_n(x) - f_n(0)] - f_n'(0)\,f_n(x)\,x}. \tag{25}$$

Для того чтобы итерация (24) сходилась к α, достаточно, чтобы в некоторой окрестности α имело место неравенство

$$|\psi'(x)| \leqslant K < 1 \tag{26}$$

и $\alpha^{(1)}$ было взято из этой окрестности. Но (26) будет иметь место, если

$$|\psi'(\alpha)| = \left|1 - \frac{f_n'(0)\,f_n'(\alpha)\,\alpha^2}{f_n^2(0)}\right| < 1. \tag{27}$$

Если все корни $\alpha_1, \alpha_2, \ldots, \alpha_n$ действительны, одного знака и

$$|\alpha_1| < |\alpha_i| \qquad (i = 2, 3, \ldots, n), \tag{28}$$

то

$$\psi'(\alpha_1) = 1 - \frac{f_n'(0)\,f_n'(\alpha_1)\,\alpha_1^2}{f_n^2(0)} =$$

$$= 1 - \left(1 + \frac{\alpha_1}{\alpha_2} + \frac{\alpha_1}{\alpha_3} + \ldots + \frac{\alpha_1}{\alpha_n}\right)\left(1 - \frac{\alpha_1}{\alpha_2}\right)\left(1 - \frac{\alpha_1}{\alpha_3}\right)\ldots\left(1 - \frac{\alpha_1}{\alpha_n}\right), \tag{29}$$

так как

$$f_n(0) = (-1)^n \alpha_1\alpha_2 \ldots \alpha_n,$$

$$f_n'(0) = (-1)^{n-1}[\alpha_2\alpha_3 \ldots \alpha_n + \alpha_1\alpha_3\alpha_4 \ldots \alpha_n + \alpha_1\alpha_2 \ldots \alpha_{n-1}],$$

$$f_n'(\alpha_1) = (-1)^{n-1}(\alpha_2 - \alpha_1)(\alpha_3 - \alpha_1)\ldots(\alpha_n - \alpha_1).$$

Методом индукции докажем, что

$$0 < \left(1 + \frac{\alpha_1}{\alpha_2} + \frac{\alpha_1}{\alpha_3} + \ldots + \frac{\alpha_1}{\alpha_n}\right)\left(1 - \frac{\alpha_1}{\alpha_2}\right) \times$$

$$\times \left(1 - \frac{\alpha_1}{\alpha_3}\right)\ldots\left(1 - \frac{\alpha_1}{\alpha_n}\right) < 1. \tag{30}$$

Так как при $i \geqslant 2$ $0 < \frac{\alpha_1}{\alpha_i} < 1$, то при $n = 2$

$$0 < \left(1 + \frac{\alpha_1}{\alpha_2}\right)\left(1 - \frac{\alpha_1}{\alpha_2}\right) = 1 - \frac{\alpha_1^2}{\alpha_2^2} < 1,$$

и неравенство (30) справедливо. Пусть оно справедливо при $n=m$. Тогда при $n=m+1$

$$0<\left(1+\frac{\alpha_1}{\alpha_2}+\ldots+\frac{\alpha_1}{\alpha_{m+1}}\right)\left(1-\frac{\alpha_1}{\alpha_2}\right)\left(1-\frac{\alpha_1}{\alpha_3}\right)\ldots\left(1-\frac{\alpha_1}{\alpha_{m+1}}\right)=$$
$$=\left[\left(1+\frac{\alpha_1}{\alpha_2}+\ldots+\frac{\alpha_1}{\alpha_m}\right)\left(1-\frac{\alpha_1}{\alpha_2}\right)\left(1-\frac{\alpha_1}{\alpha_3}\right)\ldots\left(1-\frac{\alpha_1}{\alpha_m}\right)\right]\times$$
$$\times\left(1-\frac{\alpha_1}{\alpha_{m+1}}\right)+\frac{\alpha_1}{\alpha_{m+1}}\left[\left(1-\frac{\alpha_1}{\alpha_2}\right)\left(1-\frac{\alpha_1}{\alpha_3}\right)\ldots\left(1-\frac{\alpha_1}{\alpha_{m+1}}\right)\right]<$$
$$<1-\frac{\alpha_1}{\alpha_{m+1}}+\frac{\alpha_1}{\alpha_{m+1}}=1,$$

так как каждая квадратная скобка положительна и меньше единицы.

Таким образом, неравенство (30) справедливо при всех n. Из (29) и (30) имеем:

$$0<\psi'(\alpha_1)=1-\left(1+\frac{\alpha_1}{\alpha_2}+\frac{\alpha_1}{\alpha_3}+\ldots+\frac{\alpha_1}{\alpha_n}\right)\times$$
$$\times\left(1-\frac{\alpha_1}{\alpha_2}\right)\left(1-\frac{\alpha_1}{\alpha_3}\right)\ldots\left(1-\frac{\alpha_1}{\alpha_n}\right)<1. \quad (31)$$

Сравнивая (18) и (31), мы видим, что

$$0<\psi'(\alpha_1)<\varphi'(\alpha_1)<1,$$

а это означает, что в рассматриваемом случае метод Фридмана сходится быстрей метода Лина.

В общем случае области сходимости метода Фридмана и метода Лина не совпадают.

Проиллюстрируем метод Фридмана на примере, который использовался для иллюстрации метода Лина, т. е. выделим квадратный множитель многочлена

$$x^4+7x^3+24x^2+25x-15.$$

За начальное приближение снова возьмем $g_{2,1}(x)=x^2$. Результаты вычислений приведены в таблице:

k	$g_{2,\,k-1}(x)$			Первое частное			Второе частное		
	x^2	x^1	x^0	x^2	x^1	x^0	x^2	x^1	x^0
1	1	0	0	1	7	24	16,03	29,44	—15
2	1	1,83	—0,93	1	5,17	15,47	14,9396	30,0128	—15
3	1	2,009	—1,004	1	4,991	14,9771	15,0048	29,9986	—15
4	1	1,9993	—0,9997	1	5,0007	15,0018			

Таким образом, после трех приближений мы получили результат почти с такой же точностью, как и после 12 приближений по методу Лина.

Если по полученному разложению найти корни многочлена, то получим:

$$\bar{x}_1 = -2{,}4135, \quad \bar{x}_2 = 0{,}4142, \quad \bar{x}_{3,4} = -2{,}5004 \pm 2{,}9580i.$$

В случае выделения квадратного множителя, если известно k-е приближение искомого множителя $g_{2,k}(x) = x^2 + b_1^{(k)}x + b_2^{(k)}$, первое частное $x^{n-2} + c_1^{(k)}x^{n-3} + \ldots + c_{n-3}^{(k)}x + c_{n-2}^{(k)}$ и остаток $r_0^{(k)}x + r_1^{(k)}$ находятся по схеме, приведенной в начале параграфа, причем остаток можно использовать для оценки точности достигнутого приближения. Отыскание второго частного $d_0^{(k)}x^2 + d_1^{(k)}x + d_2^{(k)}$, деля которое на $d_0^{(k)}$, мы получаем следующее приближение, сводится к вычислениям по формулам:

$$\left.\begin{aligned} d_2^{(k)} &= \frac{a_n}{c_{n-2}^{(k)}}; \quad d_1^{(k)} = \frac{1}{c_{n-2}^{(k)}}\left(a_{n-1} - c_{n-3}^{(k)}d_2^{(k)}\right), \\ d_0^{(k)} &= \frac{1}{c_{n-2}^{(k)}}\left(a_{n-2} - c_{n-3}^{(k)}d_1^{(k)} - c_{n-1}^{(k)}d_2^{(k)}\right). \end{aligned}\right\} \tag{32}$$

3. Метод Хичкока выделения квадратного множителя. Произведем деление заданного многочлена

$$f_n(x) = x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n \tag{33}$$

на трехчлен

$$g_2(x) = x^2 + px + q \tag{34}$$

с неопределенными коэффициентами p и q. Обозначив через $L(x)$ частное от деления, получим тождество

$$f_n(x) \equiv (x^2 + px + q)L(x) + xP(p, q) + Q(p, q), \tag{35}$$

где $P(p, q)$ и $Q(p, q)$ — многочлены от p и q. Для того чтобы при некоторых значениях p, q трехчлен (34) был делителем $f_n(x)$, необходимо и достаточно обращения в нуль многочленов $P(p, q)$ и $Q(p, q)$. Таким образом, для отыскания коэффициентов квадратичного делителя (34) многочлена $f_n(x)$ нужно найти решение системы

$$P(p, q) = 0; \quad Q(p, q) = 0. \tag{36}$$

Хичкок предложил для решения этой системы метод, который по существу является методом Ньютона, но только в методе Хичкока не используется явный вид многочленов P и Q, а их значения и значения производных, нужные в методе Ньютона, находятся путем двукратного деления $f_n(x)$ на приближенное выражение $g_2(x)$.

Покажем, как можно на этом пути получить производные от P и Q. Разделим многочлен $L(x)$, входящий в (35), снова на x^2+px+q и запишем тождество

$$L(x)\equiv(x^2+px+q)L_1(x)+xR(p,\ q)+S(p.\ q). \tag{37}$$

Подставляя $L(x)$ в (35), будем иметь:

$$f_n(x)\equiv(x^2+px+q)^2L_1(x)+(x^2+px+q)[xR(p,\ q)+S(p,\ q)]+ \\ +xP(p,\ q)+Q(p,\ q). \tag{38}$$

Продифференцируем последнее тождество по p и q и в результат дифференцирования подставим вместо x один из корней $\alpha_i\,(i=1,\ 2)$ трехчлена (34). В результате получим:

$$\left.\begin{aligned}\alpha_i^2R(p,\ q)+\alpha_iS(p,\ q)+\alpha_iP'_p(p,\ q)+Q'_p(p,\ q)=0,\\ \alpha_iR(p,\ q)+S(p,\ q)+\alpha_iP'_q(p,\ q)+Q'_q(p,\ q)=0\\ (i=1,\ 2).\end{aligned}\right\} \tag{39}$$

Учитывая, что

$$\alpha_i^2=-p\alpha_i-q \qquad (i=1,\ 2), \tag{40}$$

равенства (39) можно записать в таком виде:

$$\left.\begin{aligned}\alpha_i[P'_p(p,\ q)+S(p,\ q)-pR(p,\ q)]+[Q'_p(p,\ q)-qR(p,\ q)]=0,\\ \alpha_i[P'_q(p,\ q)+R(p,\ q)]+[Q'_q(p,\ q)+S(p,\ q)]=0\\ (i=1,\ 2).\end{aligned}\right\} \tag{41}$$

Если трехчлен x^2+px+q имеет различные корни, т. е. $\alpha_1\neq\alpha_2$, то из (41) следует равенство нулю каждой из квадратных скобок, поэтому

$$\left.\begin{aligned}&P'_p(p,\ q)=pR(p,\ q)-S(p,\ q); \quad P'_q(p,\ q)=-R(p,\ q),\\ &Q'_p(p,\ q)=qR(p,\ q); \qquad\qquad Q'_q(p,\ q)=-S(p,\ q).\end{aligned}\right\} \tag{42}$$

Таким образом двукратное деление $f_n(x)$ на x^2+px+q позволяет получить и частные производные от $P(p,\ q)$, $Q(p,\ q)$ по p и q, и систему (36) можно решать по методу Ньютона. Если уже известно k-е приближение $p^{(k)}$ и $q^{(k)}$ — коэффициентов искомого множителя, то двукратным делением на трехчлен $x^2+p^{(k)}x+q^{(k)}$ находим $P(p^{(k)},\ q^{(k)})$, $Q(p^{(k)},\ q^{(k)})$, $R(p^{(k)},\ q^{(k)})$, $S(p^{(k)},\ q^{(k)})$ и по (42) $P'_p(p^{(k)},\ q^{(k)})$, $P'_q(p^{(k)},\ q^{(k)})$, $Q'_p(p^{(k)},\ q^{(k)})$, $Q'_q(p^{(k)},\ q^{(k)})$; в соответствии с методом Ньютона находим следующее приближение $p^{(k+1)}$, $q^{(k+1)}$, решая систему

$$\left.\begin{aligned}&P'_p(p^{(k)},\ q^{(k)})\,\Delta p^{(k)}+P'_q(p^{(k)},\ q^{(k)})\,\Delta q^{(k)}=-P(p^{(k)},\ q^{(k)}),\\ &Q'_p(p^{(k)},\ q^{(k)})\,\Delta p^{(k)}+Q'_q(p^{(k)},\ q^{(k)})\,\Delta q^{(k)}=-Q(p^{(k)},\ q^{(k)}),\end{aligned}\right\} \tag{43}$$

где

$$\Delta p^{(k)}=p^{(k+1)}-p^{(k)}; \quad \Delta q^{(k)}=q^{(k+1)}-q^{(k)}. \tag{44}$$

Если начальное приближение $p^{(0)}$, $q^{(0)}$ выбрано достаточно хорошо, то сходимость не вызывает никаких сомнений.

Пример. Снова будем разыскивать квадратичный множитель многочлена

$$f_4(x) = x^4 + 7x^3 + 24x^2 + 25x - 15.$$

Приняв за начальное приближение $p^{(0)} = 0$, $q^{(0)} = 0$, двукратным делением $f_4(x)$ на x^2 находим:

$$P_0 = 25; \quad Q_0 = -15; \quad R_0 = 7; \quad S_0 = 24.$$

(Для сокращения записи полагаем $P_k = P(p^{(k)}, q^{(k)})$ и т. д.)

Система (43) примет вид

$$-24\,\Delta p^{(0)} - 7\,\Delta q^{(0)} = -25,$$
$$-25\,\Delta q^{(0)} = 15,$$

откуда

$$p^{(1)} = \Delta p^{(0)} = -0{,}625, \quad q^{(1)} = \Delta q^{(0)} = 1{,}224.$$

Дальнейшие вычисления понятны без пояснений:

	1	7,0000	24,0000	25,0000	−15,0000
−1,224		−1,2240	−7,0698	−21,4876	
0,625			0,6250	3,6100	10,9720
	1	5,7760	17,5552	$P_1 = 7{,}1224$;	$Q_1 = -4{,}0280$
−1,224		−1,2240			
0,625			0,6250		
	1	$R_1 = 4{,}5520$	$S_1 = 18{,}1802$		

$$12{,}6086\,\Delta p^{(1)} + 4{,}5520\,\Delta q^{(1)} = 7{,}1224;$$
$$2{,}8450\,\Delta p^{(1)} + 18{,}1802\,\Delta q^{(1)} = -4{,}0280;$$
$$\Delta p^{(1)} = 0{,}6835; \qquad \Delta q^{(1)} = -0{,}3285;$$
$$p^{(2)} = 1{,}9075; \qquad q^{(2)} = -0{,}9535;$$

	1	7,0000	24,0000	25,0000	−15,0000
−1,9075		−1,9075	−9,7139	−29,0695	
0,9535			0,9535	4,8557	14,5310
	1	5,0925	15,2396	$P_2 = 0{,}7862$;	$Q_2 = -0{,}4690$
−1,9075		−1,9075			
0,9535			0,9535		
	1	$R_2 = 3{,}1850$	$S_2 = 16{,}1931$		

$$10{,}1177\,\Delta p^{(2)} + 3{,}1850\,\Delta q^{(2)} = 0{,}7862;$$
$$3{,}0369\,\Delta p^{(2)} + 16{,}1931\,\Delta q^{(2)} = -0{,}4690;$$
$$\Delta p^{(2)} = 0{,}0923; \qquad \Delta q^{(2)} = -0{,}0463;$$
$$p^{(3)} = 1{,}9998; \qquad q^{(3)} = -0{,}9998.$$

Отклонение третьего приближения от точных значений $p=2$, $q=-1$ равно 0,0002.

УПРАЖНЕНИЯ

1. Показать, что нули многочлена

$$z^n + a_1 z^{n-1} + \ldots + a_{n-1} z + a_n$$

по модулю не превосходят единственного положительного нуля многочлена

$$z^n - |a_1| z^{n-1} - \ldots - |a_{n-1}| z - |a_n|.$$

2. Показать, что нули многочлена

$$z^n + a_1 z^{n-1} + \ldots + a_{n-1} z + a_n$$

при $a_n \neq 0$ по модулю не меньше единственного положительного нуля многочлена

$$z^n + |a_1| z^{n-1} + |a_2| z^{n-2} + \ldots + |a_{n-1}| z - |a_n|.$$

3. Пусть $d_0, d_1, \ldots, d_n$ — положительные числа и

$$d_n \geqslant |a_1| d_{n-1} + |a_2| d_{n-2} + \ldots + |a_n| d_0.$$

Показать, что нули многочлена

$$z^n + a_1 z^{n-1} + \ldots + a_n$$

не превосходят по модулю наибольшего из чисел

$$\frac{d_n}{d_{n-1}}, \quad \sqrt{\frac{d_n}{d_{n-2}}}, \quad \sqrt[3]{\frac{d_n}{d_{n-3}}}, \quad \ldots, \quad \sqrt[n]{\frac{d_n}{d_0}}.$$

4. Показать, что корни уравнения

$$z^n + a_1 z^{n-1} + \ldots + a_{n-1} z + a_n = 0$$

не превосходят по модулю наибольшего из чисел

$$n|a_1|, \quad \sqrt{n|a_2|}, \quad \sqrt[3]{n|a_3|}, \quad \ldots, \quad \sqrt[n]{n|a_n|},$$

а также наибольшего из чисел

$$\sqrt[k]{\frac{2^n-1}{C_n^k}|a_k|} \qquad (k=1, 2, \ldots, n).$$

5. Пусть $p_0 > p_1 > p_2 > \ldots > p_n > 0$. Показать, что в единичном круге $|z| \leqslant 1$ не содержится ни одного нуля многочлена

$$p_0 + p_1 z + \ldots + p_{n-1} z^{n-1} + p_n z^n.$$

6. Доказать, что если все коэффициенты $p_0, p_1, \ldots, p_n$ многочлена

$$p_0 z^n + p_1 z^{n-1} + \ldots + p_{n-1} z + p_n$$

положительны, то нули его лежат в круговом кольце $\alpha \leqslant |z| \leqslant \beta$, где α — наименьшее, а β — наибольшее из чисел $\frac{p_1}{p_0}, \frac{p_2}{p_1}, \ldots, \frac{p_n}{p_{n-1}}$.

7. Используя теорему Ролля, найти условие действительности корней уравнения $x^m + px^n + q = 0$.

8. Для уравнения

$$pa_0 x^n + a_1 x^{n-1} + \ldots + a_{n-1} x + a_n = 0$$

построена последовательность функций

$$\begin{aligned} f(x) &= a_0 x^n + a_1 x^{n-1} + \ldots + a_{n-1} x + a_n, \\ f_1(x) &= a_0 x^{n-1} + a_1 x^{n-2} + \ldots + a_{n-1}, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ f_{n-1}(x) &= a_0 x + a_1, \\ f_n(x) &= a_0. \end{aligned}$$

Доказать, что число действительных корней уравнения $f(x) = 0$, превышающих данное положительное число a, не превышает числа перемен знака в нашей последовательности при $x = a$ и разность между этими числами всегда четное число (теорема Лагерра).

9. Найти положительные корни уравнения

$$x = \operatorname{tg} x$$

с точностью до 10^{-4}.

10. Используя видоизменение Лемера метода Лобачевского, найти все корни уравнения

$$x^5 + 0{,}5x^4 - 3x^3 + 27x^2 + 13{,}5x - 81 = 0.$$

11. Показать, что если внутри некоторого круга с центром в начале координат уравнение $f(x) = 0$, где

$$f(x) = a_0 + a_1 x + a_2 x^2 + \ldots,$$

имеет единственный корень $x = \alpha$, то

$$\alpha = -\frac{a_0}{a_1} - \frac{a_0^2 a_2}{a_1 \begin{vmatrix} a_1 & a_2 \\ a_0 & a_1 \end{vmatrix}} - \frac{a_0^3 \begin{vmatrix} a_2 & a_3 \\ a_1 & a_2 \end{vmatrix}}{\begin{vmatrix} a_1 & a_2 \\ a_0 & a_1 \end{vmatrix} \begin{vmatrix} a_1 & a_2 & a_3 \\ a_0 & a_1 & a_2 \\ 0 & a_0 & a_1 \end{vmatrix}} - \ldots$$

(Воспользоваться теоремой Кёнига с $\varphi(z) = a_0$.)

12. Используя метод Ньютона для решения уравнения $x^2 - N = 0$, построить итерацию второго порядка для вычисления $\sqrt{N}$.

13. Показать, что функция

$$\varphi(x) \equiv x \frac{(m-1)x^m + (m+1)N}{(m+1)x^m + (m-1)N}$$

определяет итерацию

$$x_{n+1} = \varphi(x_n)$$

третьего порядка для отыскания $\sqrt[m]{N}$ (Бейли). (Воспользоваться теоремой Кёнига с $f(x) \equiv x^m - N$ и $\varphi(x) \equiv 1$.)

14. Обобщить метод Хичкока на случай выделения множителей третьей и четвертой степеней.

15. Методами выделения множителей найти корни уравнения

$$x^4 + 16x^3 + 71x^2 + 122x + 120 = 0$$

с четырьмя верными десятичными знаками.

16. Вычислить с четырьмя знаками два наименьших положительных корня уравнения

$$\cos x \operatorname{ch} x = 1.$$

17. Вычислить с четырьмя знаками корни следующего уравнения:

$$x^4 - 0{,}41x^3 + 1{,}632x^2 - 9{,}146x + 7{,}260 = 0.$$

18. Найти с тремя знаками корни уравнения

$$x^5 - 20{,}2x^4 + 132{,}18x^3 - 60{,}592x^2 - 72{,}693x - 14{,}525 = 0.$$

19. Найти с пятью десятичными знаками решение системы уравнений

$$\sin x = y + 1{,}32.$$
$$\cos y = x - 0{,}85,$$

20. Найти с пятью десятичными знаками решение системы

$$x^7 - 5x^2y^4 + 1510 = 0,$$
$$y^5 - 3x^4y - 105 = 0.$$

Если известно, что $x = 2$, $y = 3$ есть приближенное решение.

ЛИТЕРАТУРА

1. Хаусхолдер, Основы численного анализа, ИЛ, 1956.
2. Милн, Численный анализ, ИЛ, 1951.
3. Л. В. Канторович, Функциональный анализ и прикладная математика, УМН, т. 3, вып. 6, 1948.
4. А. Эйткен, О разложении многочленов на множители итерационными методами, УМН, т. 8, вып. 6, 1953.
5. Е. Я. Ремез, О знакопеременных рядах, которые могут быть связаны с двумя алгорифмами М. В. Остроградского для приближения иррациональных чисел, УМН, т. 6, вып. 5, 1951.
6. Г. С. Салехов, М. А. Мертвецова, О сходимости некоторых итерационных процессов, Изв. Казанского филиала АН СССР, сер. физ-матем. и техн. наук, 5, 1954.

ГЛАВА 8

ВЫЧИСЛЕНИЕ СОБСТВЕННЫХ ЗНАЧЕНИЙ И СОБСТВЕННЫХ ВЕКТОРОВ МАТРИЦ

§ 1. Введение

Большое число задач механики и физики требует отыскания собственных значений и собственных векторов матриц, т. е. отыскания таких значений λ, для которых существуют нетривиальные решения однородной системы линейных алгебраических уравнений

$$A\overline{x} = \lambda\overline{x}, \tag{1}$$

и отыскания этих нетривиальных решений. Здесь A — квадратная матрица порядка n с элементами a_{ik} и $\overline{x}$ — вектор с компонентами $x_1, x_2, \ldots, x_n$. Чтобы найти λ, нужно определить корни уравнения

$$D(\lambda) = |A - \lambda I| = 0, \tag{2}$$

где I — единичная матрица. На первый взгляд задача кажется очень простой. Достаточно раскрыть определитель (2) и решить полученное уравнение n-й степени одним из тех методов, о которых говорилось в предыдущей главе. Однако при больших значениях n раскрытие определителя (2) обычными методами высшей алгебры связано с громоздкой и утомительной работой. Основное затруднение вызвано тем, что λ входит в каждый столбец и каждую строку определителя.

В вычислительной математике выработано много различных приемов и методов, облегчающих труд по раскрытию определителя (2). Существуют также методы, позволяющие отыскивать собственные значения и собственные векторы без раскрытия определителя (2). В этой главе мы и займемся этими вопросами. Здесь мы также коснемся задачи об отыскании значений λ, удовлетворяющих уравнению

$$|A_0\lambda^m + A_1\lambda^{m-1} + \ldots + A_{m-1}\lambda + A_m| = 0, \tag{3}$$

где A_i — квадратные матрицы порядка n.

§ 2. Метод А. Н. Крылова

1. Отыскание собственных значений матрицы. Академик А. Н. Крылов одним из первых предложил довольно удобный метод раскрытия определителя (2) § 1.

Суть метода А. Н. Крылова состоит в преобразовании определителя $D(\lambda)$ к виду

$$D_1(\lambda) = \begin{vmatrix} b_{11}-\lambda & b_{12} & \dots & b_{1n} \\ b_{21}-\lambda^2 & b_{22} & \dots & b_{2n} \\ \dots & \dots & \dots & \dots \\ b_{n1}-\lambda^n & b_{n2} & \dots & b_{nn} \end{vmatrix}, \tag{1}$$

причем при некоторых условиях уравнения $D(\lambda)=0$ и $D_1(\lambda)=0$ имеют одни и те же корни. Раскрыть определитель $D_1(\lambda)$ значительно проще, чем $D(\lambda)$, так как λ содержится только в первом столбце. Как мы увидим позже, нам даже не придется вычислять миноры $D_1(\lambda)$.

Преобразование $D(\lambda)$ к виду (1) будем осуществлять следующим образом. Возьмем первое из уравнений (1) § 1:

$$a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = \lambda x_1 \tag{2}$$

и умножим его на λ. Появившиеся в левой части выражения λx_i заменим на равные им в силу системы (1) § 1 выражения:

$$a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n \qquad (i=1,\ 2,\ \dots,\ n). \tag{3}$$

Получим новое уравнение:

$$b_{21}x_1 + b_{22}x_2 + \dots + b_{2n}x_n = \lambda^2 x_1, \tag{4}$$

где

$$b_{2i} = \sum_{k=1}^{n} a_{1k}a_{ki} \qquad (i=1,\ 2,\ \dots,\ n). \tag{5}$$

С уравнением (4) поступаем так же, как и с уравнением (2). При этом получим новое уравнение:

$$b_{31}x_1 + b_{32}x_2 + \dots + b_{3n}x_n = \lambda^3 x_1, \tag{6}$$

где

$$b_{3i} = \sum_{k=1}^{n} b_{2k}a_{ki} \qquad (i=1,\ 2,\ \dots,\ n). \tag{7}$$

Этот процесс продолжаем до тех пор, пока не придем к уравнению

$$b_{n1}x_1 + b_{n2}x_2 + \dots + b_{nn}x_n = \lambda^n x_1. \tag{8}$$

Для единообразия обозначений условимся считать $b_{1i} = a_{1i}$.

Будем рассматривать вместо системы (1) § 1 систему

$$\left.\begin{aligned} b_{11}x_1 + b_{12}x_2 + \ldots + b_{1n}x_n &= \lambda x_1, \\ b_{21}x_1 + b_{22}x_2 + \ldots + b_{2n}x_n &= \lambda^2 x_1, \\ \ldots\ldots\ldots\ldots\ldots\ldots & \\ b_{n1}x_1 + b_{n2}x_2 + \ldots + b_{nn}x_n &= \lambda^n x_1. \end{aligned}\right\} \tag{9}$$

Определитель этой системы как раз и равен $D_1(\lambda)$. Покажем, что

$$\frac{D_i(\lambda)}{D(\lambda)} = \begin{vmatrix} 1 & 0 & \ldots & 0 \\ b_{11} & b_{12} & \ldots & b_{1n} \\ \ldots\ldots\ldots\ldots & & & \\ b_{n-1,1} & b_{n-1,2} & \ldots & b_{n-1,n} \end{vmatrix} = C = \text{const}. \tag{10}$$

Действительно, произведение $CD(\lambda)$, если воспользоваться формулой

$$b_{ij} = \sum_{k=1}^{n} b_{i-1,\,k}a_{kj} \tag{11}$$

и условием $b_{1k} = a_{1k}$, может быть записано в виде

$$\begin{vmatrix} b_{11}-\lambda & b_{12} & \ldots & b_{1n} \\ b_{21}-\lambda b_{11} & b_{22}-\lambda b_{12} & \ldots & b_{2n}-\lambda b_{1n} \\ b_{31}-\lambda b_{21} & b_{32}-\lambda b_{22} & \ldots & b_{3n}-\lambda b_{2n} \\ \ldots\ldots\ldots\ldots\ldots\ldots & & & \\ b_{n1}-\lambda b_{n-1,1} & b_{n2}-\lambda b_{n-1,2} & \ldots & b_{nn}-\lambda b_{n-1,n} \end{vmatrix}. \tag{12}$$

Прибавим к второй строке полученного определителя первую, умноженную на λ. При этом первые две строки нового определителя будут такими же, как и у $D_1(\lambda)$. Затем прибавим к третьей строке вторую, умноженную на λ. Тогда и третья строка станет такой же, как и у $D_1(\lambda)$. Продолжая этот процесс, мы придем в конце концов к определителю $D_1(\lambda)$. Утверждение доказано.

Если C отлично от нуля, то уравнения $D(\lambda) = 0$ и $D_1(\lambda) = 0$ эквивалентны. Если же $C = 0$, то наше преобразование не приведет к цели. Посмотрим теперь, как иногда можно обойти этот случай.

Процесс получения коэффициентов b_{ij} можно интерпретировать как последовательное получение из вектора

$$\bar{b}_0 = (1,\ 0,\ 0,\ \ldots,\ 0) \tag{13}$$

векторов

$$\bar{b}_i = (b_{i1},\ b_{i2},\ \ldots,\ b_{in}) \tag{14}$$

при помощи соотношения

$$\bar{b}_i = A'\bar{b}_{i-1}, \tag{15}$$

где A' — транспонированная по отношению к A матрица. При любом i

$$\bar{b}_i = A'^i\bar{b}_0. \tag{16}$$

Естественно напрашивается следующее обобщение. Возьмем вместо того частного вектора $\bar{b}_0$ вида (13) произвольный вектор

$$\bar{b}_0 = (b_{01}, b_{02}, \ldots, b_{0n}) \tag{17}$$

и получим с помощью его по формулам (16) векторы $\bar{b}_i$ ($i = 1, 2, \ldots, n$). Это соответствует преобразованию уравнения

$$u = b_{01}x_1 + b_{02}x_2 + \ldots + b_{0n}x_n \tag{18}$$

в систему

$$\left.\begin{aligned} u &= b_{01}x_1 + b_{02}x_2 + \ldots + b_{0n}x_n, \\ \lambda u &= b_{11}x_1 + b_{12}x_2 + \ldots + b_{1n}x_n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ \lambda^n u &= b_{n1}x_1 + b_{n2}x_2 + \ldots + b_{nn}x_n \end{aligned}\right\} \tag{19}$$

таким же процессом, которым преобразовывалось уравнение (2). Определитель системы (19) имеет вид

$$D_2(\lambda) = \begin{vmatrix} 1 & b_{01} & b_{02} & \ldots & b_{0n} \\ \lambda & b_{11} & b_{12} & \ldots & b_{1n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \lambda^n & b_{n1} & b_{n2} & \ldots & b_{nn} \end{vmatrix}. \tag{20}$$

Так же как и ранее, нетрудно убедиться, что

$$\frac{D_2(\lambda)}{D(\lambda)} = \begin{vmatrix} b_{01} & b_{02} & \ldots & b_{0n} \\ b_{11} & b_{12} & \ldots & b_{1n} \\ \ldots & \ldots & \ldots & \ldots \\ b_{n-1,\,1} & b_{n-1,\,2} & \ldots & b_{n-1,\,n} \end{vmatrix} = C_1 = \text{const}. \tag{21}$$

Может оказаться, что $C = 0$, но $C_1 \neq 0$.

Чтобы представить определители $D_1(\lambda)$ или $D_2(\lambda)$ в виде многочленов по степеням λ, нет необходимости подсчитывать их миноры. Воспользуемся теоремой Гамильтона — Кэли, утверждающей, что

$$A'^n + p_1 A'^{n-1} + \ldots + p_{n-1}A' + p_n I = 0, \tag{22}$$

где $p_1, p_2, \ldots, p_n$ — коэффициенты характеристического многочлена

$$D(\lambda) = (-1)^n [\lambda^n + p_1\lambda^{n-1} + \ldots + p_{n-1}\lambda + p_n]. \tag{23}$$

Умножая равенство (22) на $\bar{b}_0$, получим:

$$\bar{b}_n + p_1\bar{b}_{n-1} + p_2 b_{n-2} + \ldots + p_{n-1}\bar{b}_1 + p_n\bar{b}_0 = 0. \tag{24}$$

Это векторное равенство дает систему n линейных алгебраических уравнений для определения n коэффициентов характеристического уравнения. Определитель этой системы равен C_1. Решать систему можно любым из приведенных в главе 6 методов. Заметим, что мы

могли бы с таким же успехом использовать теорему Гамильтона — Кэли для матрицы A и пришли бы тогда к системе

$$\bar{c}_n + p_1\bar{c}_{n-1} + p_2\bar{c}_{n-2} + \ldots + p_{n-1}\bar{c}_1 + p_n\bar{c}_0 = 0, \tag{25}$$

где $\bar{c}_i = A^i\bar{c}_0$ и $\bar{c}_0$ — произвольный начальный вектор.

Рассмотрим пример. Пусть матрица A имеет вид

$$A = \begin{pmatrix} 1 & 3 & 1 & 4 \\ 2 & 4 & 1 & 1 \\ 3 & 5 & 4 & 2 \\ 4 & 3 & 1 & 2 \end{pmatrix}. \tag{26}$$

В качестве вектора $\bar{c}_0$ возьмем

$$\bar{c}_0 = (1,\ 0,\ 0,\ 0). \tag{27}$$

Тогда:

$\bar{c}_0$	$A\bar{c}_0$	$A^2\bar{c}_0$	$A^3\bar{c}_0$	$A^4\bar{c}_0$
1	1	26	194	1973
0	2	17	174	1651
0	3	33	337	3260
0	4	21	230	2095

(28)

и система для определения коэффициентов характеристического многочлена будет такова:

$$\left.\begin{aligned} -p_4 - p_3 - 26p_2 - 194p_1 &= 1973, \\ -2p_3 - 17p_2 - 174p_1 &= 1651, \\ -3p_3 - 33p_2 - 337p_1 &= 3260, \\ -4p_3 - 21p_2 - 230p_1 &= 2095. \end{aligned}\right\} \tag{29}$$

Решая систему (29), найдем:

$$p_1 = -11,\quad p_2 = 7,\quad p_3 = 72,\quad p_4 = -93. \tag{30}$$

Таким образом, характеристический многочлен примет вид

$$D(\lambda) = \lambda^4 - 11\lambda^3 + 7\lambda^2 + 72\lambda - 93. \tag{31}$$

В приведенном примере мы пришли к характеристическому многочлену. Это оказалось возможным благодаря тому, что $C_1 \neq 0$. Исследуем теперь подробнее причины, по которым C_1 может оказаться равным нулю. Как уже упоминалось, матрица A (и транспонированная матрица A') удовлетворяет своему характеристическому уравнению $D(A) = 0$. Но может оказаться, что существуют

многочлены $\varphi(\lambda)$ степени меньше n, для которых также выполнены равенства

$$\varphi(A) = \varphi(A') = 0. \tag{32}$$

Среди таких многочленов имеется единственный многочлен $\psi(\lambda)$ со старшим коэффициентом 1, имеющий наименьшую степень. Этот многочлен в линейной алгебре принято называть *минимальным многочленом*. Напомним некоторые свойства минимальных многочленов. Произвольный многочлен $\varphi(\lambda)$, для которого $\varphi(A) = 0$, делится на минимальный многочлен. В частности, характеристический многочлен $D(\lambda)$ делится на $\psi(\lambda)$. При этом

$$D(\lambda) = \psi(\lambda) D_{n-1}(\lambda), \tag{33}$$

где $D_{n-1}(\lambda)$ — общий наибольший делитель всех миноров $(n-1)$-го порядка матрицы $A - \lambda I$. Корнями $\psi(\lambda)$ служат все различные корни $D(\lambda)$. Так, если

$$D(\lambda) = (-1)^n (\lambda - \lambda_1)^{\alpha_1} (\lambda - \lambda_2)^{\alpha_2} \dots (\lambda - \lambda_k)^{\alpha_k}$$
$$(\lambda_i \neq \lambda_j,\ \alpha_1 + \alpha_2 + \dots + \alpha_k = n), \tag{34}$$

то

$$\psi(\lambda) = (\lambda - \lambda_1)^{\beta_1} (\lambda - \lambda_2)^{\beta_2} \dots (\lambda - \lambda_k)^{\beta_k} \qquad (1 \leqslant \beta_i \leqslant \alpha_i). \tag{35}$$

Возьмем теперь произвольные векторы $\bar{b}_0$ и $\bar{c}_0$. В каждой из последовательностей векторов

$$\bar{c}_0,\ A\bar{c}_0,\ \dots,\ A^n\bar{c}_0 \tag{36}$$

и

$$\bar{b}_0,\ A'\bar{b}_0,\ \dots,\ A'^n\bar{b}_0 \tag{37}$$

не может быть более n независимых. Более того, для любых векторов $\bar{b}_0$ и $\bar{c}_0$ наверняка имеет место линейная зависимость

$$\psi(A')\bar{b}_0 = \psi(A)\bar{c}_0 = 0. \tag{38}$$

Таким образом, если степень $\psi(\lambda)$ меньше n, то как бы мы ни выбирали $\bar{b}_0$ и $\bar{c}_0$, системы (24) и (25) будут иметь нулевые определители.

Посмотрим теперь, что может дать метод А. Н. Крылова в этом случае. Все рассуждения будем проводить с матрицей A. Для произвольного вектора $\bar{c}_0$ найдется многочлен $\psi_{\bar{c}_0}(\lambda)$ минимальной степени со старшим коэффициентом 1 такой, что

$$\psi_{\bar{c}_0}(A)\bar{c}_0 = 0. \tag{39}$$

Назовем его *минимальным многочленом вектора* $\bar{c}_0$. Любой другой многочлен $\psi(\lambda)$, обладающий свойством

$$\varphi(A)\bar{c}_0 = 0, \tag{40}$$

должен делиться на $\psi_{\bar{c}_0}(\lambda)$. Следовательно, минимальный многочлен $\psi(\lambda)$ матрицы A, для которого равенство (40) выполнено при любом векторе $\bar{c}_0$, должен делиться на минимальный многочлен любого вектора $\bar{c}_0$.

Итак, мы приходим к выводу, что если в последовательности (36) m является наибольшим индексом, для которого векторы

$$\bar{c}_0,\ A\bar{c}_0,\ A^2\bar{c}_0,\ \ldots,\ A^{m-1}\bar{c}_0 \tag{41}$$

линейно независимы, а вектор $A^m\bar{c}_0$ линейно зависит от них:

$$A^m\bar{c}_0 = -\alpha_0\bar{c}_0 - \alpha_1 A\bar{c}_0 - \ \ldots\ - \alpha_{m-1}A^{m-1}\bar{c}_0, \tag{42}$$

то многочлен

$$\lambda^m + \alpha_{m-1}\lambda^{m-1} + \ \ldots\ + \alpha_1\lambda + \alpha_0 \tag{43}$$

будет являться или минимальным многочленом матрицы A или его делителем. Этот многочлен мы и получим по методу А. Н. Крылова.

Проиллюстрируем последний случай на следующем примере. В качестве матрицы A возьмем

$$A = \begin{pmatrix} 5 & 2 & -1 & -1 \\ 3 & 3 & 0 & 0 \\ 1 & -2 & 4 & 1 \\ 3 & 0 & 0 & 3 \end{pmatrix}. \tag{44}$$

Если снова взять $\bar{c}_0 = (1,\ 0,\ 0,\ 0)$, то получим:

$$\left.\begin{aligned} A\bar{c}_0 &= (5,\ 3,\ 1,\ 3), \\ A^2\bar{c}_0 &= (27,\ 24,\ 6,\ 24), \\ A^3\bar{c}_0 &= (153,\ 153,\ 27,\ 153). \end{aligned}\right\} \tag{45}$$

Вычисления дают

$$A^3\bar{c}_0 - 12A^2\bar{c}_0 + 45A\bar{c}_0 - 54\bar{c}_0 = 0. \tag{46}$$

Таким образом, многочлен

$$\lambda^3 - 12\lambda^2 + 45\lambda - 54 \tag{47}$$

будет минимальным для вектора $\bar{c}_0$.

На этом примере мы убеждаемся не только в том, что векторы $\bar{c}_0,\ A\bar{c}_0,\ \ldots,\ A^m\bar{c}_0$ могут оказаться линейно зависимыми при $m < n$, но и в трудностях, которые возникают при обнаружении такой линейной зависимости. В связи с этим остановимся на методах решения системы (25).

Рассмотрим матрицу

$$\begin{pmatrix} c_{01} & c_{02} & \dots & c_{0n} & 1 \\ c_{11} & c_{12} & \dots & c_{1n} & \lambda \\ \dots & \dots & \dots & \dots & \dots \\ c_{m1} & c_{m2} & \dots & c_{mn} & \lambda^m \end{pmatrix}, \tag{48}$$

где

$$A^t\bar{c}_0 = (c_{i1}, c_{i2}, \dots, c_{in}). \tag{49}$$

Задача об отыскании коэффициентов α_i в (42) эквивалентна задаче об отыскании такой линейной комбинации первых m строк матрицы (48), которая в сумме с $(m+1)$-й строкой обращает первые n элементов последней в нули. При этом в последнем столбце получится минимальный многочлен вектора $\bar{c}_0$. Поэтому целесообразно для получения нужной линейной комбинации применить метод исключения Гаусса. Процесс исключения можно осуществлять и до того, как будет получена полная матрица (48). Так, получив $(c_{11}, c_{12}, \dots, c_{1n})$, мы можем обратить в нуль какой-либо из элементов c_{1i}, прибавив ко второй строке (48) первую, умноженную на подходящий множитель m_{11}. При этом в последнем столбце второй строки будет стоять вместо λ выражение $\lambda + m_{11}$. После такого преобразования вторая строка (48) примет вид

$$c^*_{11}, c^*_{12}, \dots, c^*_{1,\,i-1}, 0, c^*_{1,\,i+1}, \dots, c^*_{1n}, \lambda + m_{11}. \tag{50}$$

Теперь будем умножать на A вектор $(c^*_{11}, \dots, c^*_{1,\,i-1}, 0, c^*_{1,\,i+1}, \dots, c^*_{1n})$. Получим вместо прежней третьей строки (48) новую строку:

$$c^*_{21}, c^*_{22}, \dots, c^*_{2n}, \lambda^2 + m_{11}\lambda. \tag{51}$$

Подберем постоянные m_{21} и m_{22} так, чтобы при прибавлении первой строки (48), умноженной на m_{21}, и строки (50), умноженной на m_{22}, к строке (51) в последней обратились в нули элементы, стоящие на i-м и j-м месте, где $i \neq j$ и $j < n+1$. Продолжая этот процесс, мы придем в конце концов к строке, все элементы которой, кроме крайнего правого, равны нулю. Это показывает нам, что мы уже получили линейно зависимые векторы. Правый крайний элемент последней строки и даст нам минимальный многочлен вектора $\bar{c}_0$. Преимущество такого метода исключения на каждом шаге состоит в том, что на A будут умножаться векторы, имеющие все больше и больше нулевых компонент. Благодаря этому сокращается число необходимых операций умножения.

Проиллюстрируем это на примере той же матрицы A (44). Возьмем прежний вектор $\bar{c}_0$. Получив $A\bar{c}_0$ такое же, как и в (45), исключим c_{11}. При этом строка (50) примет вид

$$0, 3, 1, 3, \lambda - 5. \tag{52}$$

После следующего умножения на A получим строку

$$2,\ 9,\ 1,\ 9,\ \lambda^2 - 5\lambda. \tag{53}$$

Вычтем отсюда первую строку (48), умноженную на 2, и (52), умноженную на 3. При этом получим:

$$0,\ 0,\ -2,\ 0,\ \ \lambda^2 - 8\lambda + 13. \tag{54}$$

Следующее умножение на A даст

$$2,\ 0,\ -8,\ 0,\ \lambda^3 - 8\lambda^2 + 13\lambda. \tag{55}$$

Вычтем отсюда первую строку (48), умноженную на 2, и (54), умноженную на 4. Это даст

$$0,\ 0,\ 0,\ 0,\ \lambda^3 - 12\lambda^2 + 45\lambda - 54. \tag{56}$$

Отсюда мы заключаем, что минимальным многочленом вектора $\bar{c}_0$ является

$$\lambda^3 - 12\lambda^2 + 45\lambda - 54. \tag{57}$$

Этот же многочлен мы получили и ранее.

Многочлен (57) имеет третью степень. Следовательно, он не совпадает с характеристическим многочленом, имеющим четвертую степень. В данном случае мы можем найти недостающий корень характеристического уравнения. Действительно, след матрицы, т. е. сумма ее диагональных элементов, должен равняться сумме корней характеристического многочлена матрицы. В нашем случае он равен 15. С другой стороны, сумма корней многочлена (57) равна 12. Следовательно, недостающее собственное значение равно 3. Нетрудно видеть, что многочлен (57) можно записать в виде

$$(\lambda - 3)^2(\lambda - 6). \tag{58}$$

Таким образом, собственные значения матрицы A равны

$$\lambda_1 = \lambda_2 = \lambda_3 = 3; \quad \lambda_4 = 6. \tag{59}$$

В главе 6 мы уже обращали внимание на важность подсчета числа операций умножения и деления, необходимых для решения задачи. Произведем такой подсчет и для метода Крылова в его последнем варианте.

Для образования $A\bar{c}_0$ потребуется n^2 операций умножения (предполагается, что все компоненты $\bar{c}_0$ отличны от нуля). Исключение одного из c_{1i} потребует n умножений и делений. Последующее умножение полученного вектора на A потребует $n(n-1)$ операций умножения, а исключение двух элементов потребует $2n - 1$

умножений и делений. Продолжая эти подсчеты и дальше, мы обнаружим, что всего потребуется

$$n^2+n(n-1)+n(n-2)+\ldots+n\cdot 1+n+[n+(n-1)]+$$
$$+[n+(n-1)+(n-2)]+\ldots$$
$$\ldots+[n+(n-1)+\ldots+1]=\frac{5n^3+6n^2+n}{6} \qquad (60)$$

операций умножения и деления.

При этом подсчете мы не учитывали действий с последним столбцом матрицы (48). Для этих действий потребуется дополнительно

$$1+(1+2)+(1+2+3)+\ldots$$
$$\ldots+(1+2+3+\ldots+(n-2))=\frac{n^3-3n^2+2n}{6} \qquad (61)$$

операций умножения.

Таким образом, если все n шагов осуществимы, то метод Крылова раскрытия векового определителя потребует

$$\frac{2n^3+n^2+n}{2} \qquad (62)$$

операций умножения и деления.

2. Отыскание собственных векторов матрицы. Рассмотрим теперь вопрос об отыскании собственных векторов. Пусть λ_i является корнем минимального многочлена, соответствующего вектору c_0. Если степень этого минимального многочлена равна m, то будем разыскивать собственный вектор $\bar{x}_i$ в виде

$$\bar{x}_i=\gamma_1\bar{c}_0+\gamma_2A\bar{c}_0+\ldots+\gamma_mA^{m-1}\bar{c}_0. \qquad (63)$$

Из

$$A\bar{x}_i=\lambda_i\bar{x}_i \qquad (64)$$

следует:

$$\gamma_1A\bar{c}_0+\gamma_2A^2\bar{c}_0+\ldots+\gamma_mA^m\bar{c}_0=$$
$$=\lambda_i(\gamma_1\bar{c}_0+\gamma_2A\bar{c}_0+\ldots+\gamma_mA^{m-1}\bar{c}_0) \qquad (65)$$

или в силу (42)

$$\gamma_1A\bar{c}_0+\gamma_2A^2\bar{c}_0+\ldots+\gamma_{m-1}A^{m-1}\bar{c}_0-\gamma_m\times$$
$$\times(\alpha_0\bar{c}_0+\alpha_1A\bar{c}_0+\ldots+\alpha_{m-1}A^{m-1}\bar{c}_0)=$$
$$=\lambda_i(\gamma_1\bar{c}_0+\gamma_2A\bar{c}_0+\ldots+\gamma_mA^{m-1}\bar{c}_0). \qquad (66)$$

Итак,

$$(\lambda_i\gamma_1+\alpha_0\gamma_m)\bar{c}_0+(\lambda_i\gamma_2+\alpha_1\gamma_m-\gamma_1)A\bar{c}_0+\ldots$$
$$\ldots+(\lambda_i\gamma_m+\alpha_{m-1}\gamma_m-\gamma_{m-1})A^{m-1}\bar{c}_0=0. \qquad (67)$$

Векторы $\bar{c}_0$, $A\bar{c}_0$, ..., $A^{m-1}\bar{c}_0$ линейно независимы. Поэтому из (67) следует:

$$\left.\begin{aligned} \lambda_i\gamma_1 + \alpha_0\gamma_m &= 0, \\ \lambda_i\gamma_2 + \alpha_1\gamma_m - \gamma_1 &= 0, \\ \cdots\cdots\cdots\cdots\cdots & \\ \lambda_i\gamma_m + \alpha_{m-1}\gamma_m - \gamma_{m-1} &= 0. \end{aligned}\right\} \tag{68}$$

Отсюда получаем:

$$\left.\begin{aligned} \gamma_{m-1} &= \lambda_i\gamma_m + \alpha_{m-1}\gamma_m, \\ \gamma_{m-2} &= \lambda_i\gamma_{m-1} + \alpha_{m-2}\gamma_m, \\ \cdots\cdots\cdots\cdots & \\ \gamma_1 &= \lambda_1\gamma_2 + \alpha_i\gamma_m. \end{aligned}\right\} \tag{69}$$

Мы не выписали повторно первое из равенств (68), так как оно является следствием остальных и условия

$$\lambda_i^m + \alpha_{m-1}\lambda_i^{m-1} + \ldots + \alpha_1\lambda_i + \alpha_0 = 0. \tag{70}$$

Коэффициент γ_m должен быть отличным от нуля, так как в противном случае все остальные коэффициенты γ_i были бы равны нулю, что невозможно, ибо $\bar{x}_i$ является собственным вектором. Задаваясь каким-либо значением $\gamma_m \neq 0$ (например, 1), мы при помощи равенств (69) последовательно найдем все остальные γ_i, а следовательно, и вектор $\bar{x}_i$. Так, для матрицы (44) мы получим, если положим $\lambda_i = 3$ и $\gamma_3 = 1$:

$$\gamma_2 = -9, \quad \gamma_1 = 18 \tag{71}$$

и

$$\bar{x}_i = 18\bar{c}_0 - 9A\bar{c}_0 + A^2\bar{c}_0 = -3(0,\ 1,\ 1,\ 1). \tag{72}$$

Нетрудно проверить, что это действительно собственный вектор матрицы (44), соответствующий собственному значению $\lambda_i = 3$.

Если выбрать другое значение $\gamma_m \neq 0$, то новый собственный вектор будет отличаться от предыдущего лишь постоянным множителем. Таким образом, если степень минимального многочлена вектора $\bar{c}_0$ равна m и λ_i — произвольный корень этого минимального многочлена, то в линейном множестве, порожденном векторами $\bar{c}_0$, $A\bar{c}_0$, ..., $A^{m-1}\bar{c}_0$, найдется единственный, с точностью до множителя, собственный вектор $\bar{x}_i$, соответствующий собственному значению λ_i. Для того чтобы найти другие собственные векторы, соответствующие собственному значению λ_i, если такие существуют, и собственные векторы, соответствующие собственным значениям λ_j, не являющимся корнями минимального многочлена вектора $\bar{c}_0$, следует выбрать новый начальный вектор $\bar{c}_0'$, не являющийся линейной комбинацией векторов $\bar{c}_0$, $A\bar{c}_0$, ..., $A^{m-1}\bar{c}_0$, и повторить с ним тот же процесс, который мы проделали с вектором $\bar{c}_0$. Иногда придется повторить это несколько раз.

§ 3. Метод Ланцоша

1. Отыскание собственных значений. Решение систем (25) или (42) предыдущего параграфа для определения коэффициентов характеристического или минимального многочлена можно осуществлять методами ортогонализации, изложенными в § 4 главы 6. Процесс ортогонализации целесообразно проводить после каждого умножения на матрицу A. В настоящем параграфе мы и рассмотрим возникающие при этом алгорифмы.

Выбираем произвольный начальный вектор $\bar{c}_0 \neq 0$ и находим $A\bar{c}_0$. Подберем теперь коэффициент α_{10} так, чтобы вектор

$$\bar{c}_1 = A\bar{c}_0 - \alpha_{10}\bar{c}_0 \tag{1}$$

был ортогонален к вектору $\bar{c}_0$. Это всегда возможно и условие ортогональности дает

$$\alpha_{10} = \frac{(A\bar{c}_0, \bar{c}_0)}{(\bar{c}_0, \bar{c}_0)}. \tag{2}$$

Может оказаться, что $\bar{c}_1 = 0$. В этом случае векторы $\bar{c}_0$ и $A\bar{c}_0$ линейно зависимы и $P_1(\lambda) = \lambda - \alpha_{10}$ будет делителем минимального многочлена матрицы A. Тогда дальнейшие действия с вектором $\bar{c}_0$ прекращаются. Если же $\bar{c}_1 \neq 0$, то образуем вектор $A\bar{c}_1$ и подбираем коэффициенты α_{21} и α_{20} так, чтобы вектор

$$\bar{c}_2 = A\bar{c}_1 - \alpha_{21}\bar{c}_1 - \alpha_{20}\bar{c}_0 \tag{3}$$

был ортогонален к векторам $\bar{c}_0$ и $\bar{c}_1$. Это также всегда возможно. При этом

$$\alpha_{21} = \frac{(A\bar{c}_1, \bar{c}_1)}{(\bar{c}_1, \bar{c}_1)}; \quad \alpha_{20} = \frac{(A\bar{c}_1, \bar{c}_0)}{(\bar{c}_0, \bar{c}_0)}. \tag{4}$$

Если окажется, что $\bar{c}_2 = 0$, то

$$A(A\bar{c}_0 - \alpha_{10}\bar{c}_0) - \alpha_{21}(A\bar{c}_0 - \alpha_{10}\bar{c}_0) - \alpha_{20}\bar{c}_0 = 0 \tag{5}$$

даст линейную зависимость между векторами $\bar{c}_0$, и $A\bar{c}_0$ и $A^2\bar{c}_0$, а многочлен

$$P_2(\lambda) = (\lambda - \alpha_{21})(\lambda - \alpha_{10}) - \alpha_{20} = (\lambda - \alpha_{21})P_1(\lambda) - \alpha_{20} \tag{6}$$

будет делителем минимального многочлена матрицы A. Если же $\bar{c}_2 \neq 0$, то продолжаем процесс ортогонализации.

Пусть нами уже найдены векторы $\bar{c}_0, \bar{c}_1, \ldots, \bar{c}_{m-1}$, удовлетворяющие условиям:

$$\bar{c}_{k+1} = A\bar{c}_k - \alpha_{k+1,\,k}\bar{c}_k - \alpha_{k+1,\,k-1}\bar{c}_{k-1} - \ldots - \alpha_{k+1,\,0}\bar{c}_0 \quad (k = 1, 2, \ldots, m-1), \tag{7}$$

$$(\bar{c}_i, \bar{c}_j) = 0 \text{ при } i \neq j, \quad \bar{c}_k \neq 0 \quad (i, j, k = 0, 1, 2, \ldots, m-1). \tag{8}$$

Тогда подбираем коэффициенты $\alpha_{m,\,m-1},\ \alpha_{m,\,m-2},\ \ldots,\ \alpha_{m,\,0}$ так, чтобы вектор

$$\bar{c}_m = A\bar{c}_{m-1} - \alpha_{m,\,m-1}\bar{c}_{m-1} - \alpha_{m,\,m-2}\bar{c}_{m-2} - \ldots - \alpha_{m0}\bar{c}_0 \qquad (9)$$

был ортогонален к каждому из векторов $\bar{c}_0,\ \bar{c}_1,\ \ldots,\ \bar{c}_{m-1}$. Это возможно. Коэффициенты α_{mi} должны быть определены по формулам

$$\alpha_{mi} = \frac{(A\bar{c}_{m-1},\ \bar{c}_i)}{(\bar{c}_i,\ \bar{c}_i)}. \qquad (10)$$

Параллельно с построением системы взаимно ортогональных векторов $\bar{c}_0,\ \bar{c}_1,\ \ldots,\ \bar{c}_m,\ \ldots$ строим последовательность многочленов

$$P_0(\lambda) = 1;\ P_m(\lambda) = (\lambda - \alpha_{m,\,m-1}) \times$$
$$\times P_{m-1}(\lambda) - \alpha_{m,\,m-2}P_{m-2}(\lambda) - \ldots - \alpha_{m0}P_0(\lambda). \qquad (11)$$

Так как в нашем пространстве имеется не более n взаимно ортогональных векторов, то на каком-то шаге будем иметь $\bar{c}_m = 0$. При этом

$$A\bar{c}_{m-1} - \alpha_{m,\,m-1}\bar{c}_{m-1} - \alpha_{m,\,m-2}\bar{c}_{m-2} - \ldots - \alpha_{m,\,0}\bar{c}_0 = 0 \qquad (12)$$

даст линейную зависимость векторов $\bar{c}_0,\ A\bar{c}_0,\ \ldots,\ A^m\bar{c}_0$ и, следовательно, многочлен

$$P_m(\lambda) = (\lambda - \alpha_{m,\,m-1})P_{m-1}(\lambda) - \sum_{k=0}^{m-2}\alpha_{m,\,k}P_k(\lambda) \qquad (13)$$

будет делителем минимального многочлена матрицы A. При $m = n$ $P_m(\lambda)$ будет являться характеристическим многочленом матрицы A. Если $m < n$, то выбираем новый начальный вектор $\bar{c}'_0$, ортогональный к векторам $\bar{c}_0,\ \bar{c}_1,\ \ldots,\ \bar{c}_{m-1}$, и повторяем с ним тот же процесс. Если этого окажется недостаточно, т. е. общее количество ортогональных векторов все еще будет меньше n, то проводим наши рассуждения с новым вектором $\bar{c}''_0$, ортогональным ко всем предыдущим, и т. д.

Для симметрической матрицы A равенства (7) упрощаются. Действительно, в этом случае

$$\alpha_{k+1,\,i} = \frac{(A\bar{c}_k,\ \bar{c}_i)}{(\bar{c}_i,\ \bar{c}_i)} = \frac{(\bar{c}_k,\ A\bar{c}_i)}{(\bar{c}_i,\ \bar{c}_i)} = \frac{(\bar{c}_k,\ \bar{c}_{i+1} + \alpha_{i+1,\,i}\bar{c}_i + \ldots + \alpha_{i+1,\,0}\bar{c}_0)}{(\bar{c}_i,\ \bar{c}_i)}$$
$$(i = 0,\ 1,\ 2,\ \ldots,\ k), \qquad (14)$$

и если $i < k - 1$, то $\alpha_{k+1,\,i} = 0$. Таким образом, если матрица A симметрическая, то вместо (7) будем иметь:

$$\bar{c}_{k+1} = A\bar{c}_k - \alpha_{k+1,\,k}\bar{c}_k - \alpha_{k+1,\,k-1}\bar{c}_{k-1}. \qquad (15)$$

Аналогичное упрощение можно получить и для несимметрической матрицы, заменив процесс ортогонализации процессом биортогонализации, подобно тому как это сделано в § 4 главы 6.

Будем исходить из двух начальных векторов $\bar{c}_0$ и $\bar{b}_0$. Найдем по ним векторы $A\bar{c}_0$ и $A'\bar{b}_0$ и образуем линейные комбинации

$$\bar{c}_1 = A\bar{c}_0 - \alpha_{10}\bar{c}_0; \quad \bar{b}_1 = A'\bar{b}_0 - \beta_{10}\bar{b}_0. \tag{16}$$

Коэффициенты α_{10} и β_{10} подберем так, чтобы оказалось $(\bar{c}_1, \bar{b}_0) = (\bar{b}_1, \bar{c}_0) = 0$. Это возможно, если начальные векторы $\bar{c}_0$ и $\bar{b}_0$ не были ортогональны, так как

$$\alpha_{10} = \beta_{10} = \frac{(A\bar{c}_0, \bar{b}_0)}{(\bar{c}_0, \bar{b}_0)} = \frac{(\bar{c}_0, A'\bar{b}_0)}{(\bar{c}_0, \bar{b}_0)}. \tag{17}$$

В дальнейшем будем предполагать, что $(\bar{c}_0, \bar{b}_0) \neq 0$. Тогда по найденным $\bar{c}_1$ и $\bar{b}_1$ строим векторы $A\bar{c}_1$ и $A'\bar{b}_1$ и образуем линейные комбинации

$$\bar{c}_2 = A\bar{c}_1 - \alpha_{21}\bar{c}_1 - \alpha_{20}\bar{c}_0, \quad \bar{b}_2 = A'\bar{b}_1 - \beta_{21}\bar{b}_1 - \beta_{20}\bar{b}_0 \tag{18}$$

так, чтобы оказалось

$$(\bar{c}_2, \bar{b}_1) = (\bar{c}_2, \bar{b}_0) = (\bar{b}_2, \bar{c}_1) = (\bar{b}_2, \bar{c}_0) = 0. \tag{19}$$

При этом будем иметь:

$$\left.\begin{aligned} \alpha_{21} = \beta_{21} &= \frac{(A\bar{c}_1, \bar{b}_1)}{(\bar{c}_1, \bar{b}_1)} = \frac{(\bar{c}_1, A'\bar{b}_1)}{(\bar{c}_1, \bar{b}_1)}; \\ \alpha_{20} = \beta_{20} &= \frac{(A\bar{c}_1, \bar{b}_0)}{(\bar{c}_0, \bar{b}_0)} = \frac{(\bar{c}_0, A'\bar{b}_1)}{(\bar{c}_0, \bar{b}_0)} = \frac{(\bar{c}_1, \bar{b}_1)}{(\bar{c}_0, \bar{b}_0)}, \end{aligned}\right\} \tag{20}$$

и наше построение возможно, если $(\bar{c}_1, \bar{b}_1) \neq 0$, $(\bar{c}_0, \bar{b}_0) \neq 0$. Будем предполагать, что эти условия выполнены, и продолжим построение дальше. Пусть у нас уже построены векторы

$$\bar{c}_0, \bar{c}_1, \ldots, \bar{c}_k, \tag{21}$$

$$\bar{b}_0, \bar{b}_1, \ldots, \bar{b}_k, \tag{22}$$

причем эти две системы биортогональны, т. е.

$$(\bar{b}_i, \bar{c}_j) = 0 \quad \text{при} \quad i \neq j \qquad (i, j = 0, 1, \ldots, k). \tag{23}$$

Тогда строим векторы

$$\left.\begin{aligned} \bar{c}_{k+1} &= A\bar{c}_k - \alpha_{k+1,\,k}\bar{c}_k - \alpha_{k+1,\,k-1}\bar{c}_{k-1} - \ldots - \alpha_{k+1,\,0}\bar{c}_0, \\ \bar{b}_{k+1} &= A'\bar{b}_k - \beta_{k+1,\,k}\bar{b}_k - \beta_{k+1,\,k-1}\bar{b}_{k-1} - \ldots - \beta_{k+1,\,0}\bar{b}_0 \end{aligned}\right\} \tag{24}$$

так, чтобы

$$(\bar{c}_{k+1},\ \bar{b}_i) = (\bar{b}_{k+1},\ \bar{c}_i) = 0 \qquad (i = 0,\ 1,\ 2,\ \ldots,\ k). \tag{25}$$

Условия (25) дают

$$\alpha_{k+1,\,i} = \beta_{k+1,\,i} = \frac{(A\bar{c}_k,\ b_i)}{(\bar{c}_i,\ \bar{b}_i)} = \frac{(A'\bar{b}_k,\ \bar{c}_i)}{(\bar{c}_i,\ \bar{b}_i)} \qquad (i = 0,\ 1,\ \ldots,\ k). \tag{26}$$

При этом, если $i < k - 1$, то

$$\alpha_{k+1,\,i} = \beta_{k+1,\,i} = \frac{(A\bar{c}_k,\ \bar{b}_i)}{(\bar{c}_i,\ \bar{b}_i)} = \frac{(\bar{c}_k,\ A'\bar{b}_i)}{(\bar{c}_i,\ \bar{b}_i)} = \frac{(\bar{c}_k,\ \bar{b}_{i+1})}{(\bar{c}_i,\ \bar{b}_i)} = 0. \tag{27}$$

Таким образом соотношения (24) примут вид

$$\left.\begin{aligned} \bar{c}_{k+1} &= A\bar{c}_k - \alpha_{k+1,\,k}\bar{c}_k - \alpha_{k+1,\,k-1}\bar{c}_{k-1}, \\ \bar{b}_{k+1} &= A'\bar{b}_k - \alpha_{k+1,\,k}\bar{b}_k - \alpha_{k+1,\,k-1}\bar{b}_{k-1}. \end{aligned}\right\} \tag{28}$$

Наши построения будут возможны до тех пор, пока $(\bar{c}_k,\ \bar{b}_k) \neq 0$. Это условие может нарушаться в следующих трех случаях:

$$\left.\begin{aligned} &\text{а) } \bar{c}_k = 0 \text{ и } \bar{b}_k = 0; \\ &\text{б) либо } \bar{c}_k = 0, \text{ либо } \bar{b}_k = 0; \\ &\text{в) } \bar{c}_k \neq 0,\ \bar{b}_k \neq 0, \text{ но } \bar{c}_k \perp \bar{b}_k. \end{aligned}\right\} \tag{29}$$

Все эти три случая могут встречаться фактически. Продемонстрируем это на примере матрицы (44) § 2.

а) Возьмем сначала

$$\bar{b}_0 = \bar{c}_0 = (0,\ 1,\ 0,\ 0). \tag{30}$$

Тогда

$$A\bar{c}_0 = (2,\ 3,\ -2,\ 0); \quad A'\bar{b}_0 = (3,\ 3,\ 0,\ 0) \tag{31}$$

и

$$\alpha_{10} = \frac{(A\bar{c}_0,\ b_0)}{(\bar{c}_0,\ \bar{b}_0)} = 3. \tag{32}$$

Следовательно,

$$\bar{c}_1 = (2,\ 0,\ -2,\ 0); \quad \bar{b}_1 = (3,\ 0,\ 0,\ 0). \tag{33}$$

Далее,

$$A\bar{c}_1 = (12,\ 6,\ -6,\ 6), \quad A'\bar{b}_1 = (15,\ 6,\ -3,\ -3). \tag{34}$$

Отсюда

$$\alpha_{20} = \frac{(A\bar{c}_1,\ \bar{b}_0)}{(\bar{b}_0,\ \bar{c}_0)} = 6; \quad \alpha_{21} = \frac{(A\bar{c}_1,\ \bar{b}_1)}{(\bar{c}_1,\ \bar{b}_1)} = 6 \tag{35}$$

и

$$\bar{c}_2 = (0,\ 0,\ 6,\ 6); \quad \bar{b}_2 = (-3,\ 0,\ -3,\ -3). \tag{36}$$

Продолжая процесс, найдем:

$$\left.\begin{array}{ll} A\bar{c}_2 = (-12, 0, 30, 18), & A'\bar{b}_2 = (-27, 0, -9, -9), \\ \alpha_{31} = \dfrac{(A\bar{c}_2, \bar{b}_1)}{(\bar{c}_1, \bar{b}_1)} = -6, & \alpha_{32} = \dfrac{(A\bar{c}_2, \bar{b}_2)}{(\bar{c}_2, \bar{b}_2)} = 3, \\ \bar{c}_3 = (0, 0, 0, 0), & \bar{b}_3 = (0, 0, 0, 0). \end{array}\right\} \quad (37)$$

б) Возьмем теперь

$$\bar{b}_0 = \bar{c}_0 = (1, 0, 0, 0). \quad (38)$$

При этом

$$\left.\begin{array}{ll} A\bar{c}_0 = (5, 3, 1, 3), & A'\bar{b}_0 = (5, 2, -1, -1), \\ & \alpha_{10} = \dfrac{(A\bar{c}_0, \bar{b}_0)}{(\bar{c}_0, \bar{b}_0)} = 5, \\ \bar{c}_1 = (0, 3, 1, 3), & \bar{b}_1 = (0, 2, -1, -1), \\ A\bar{c}_1 = (2, 9, 1, 9), & A'\bar{b}_1 = (2, 8, -4, -4), \\ \alpha_{20} = \dfrac{(A\bar{c}_1, \bar{b}_0)}{(\bar{c}_0, \bar{b}_0)} = 2, & \alpha_{21} = \dfrac{(A\bar{c}_1, \bar{b}_1)}{(\bar{c}_1, \bar{b}_1)} = 4, \\ \bar{c}_2 = (0, -3, -3, -3), & \bar{b}_2 = (0, 0, 0, 0). \end{array}\right\} \quad (39)$$

в) Наконец, если взять

$$\bar{c}_0 = (1, 0, 0, 0), \quad \bar{b}_0 = \left(1, \frac{\sqrt{3}-1}{3}, 0, 0\right), \quad (40)$$

то

$$\left.\begin{array}{l} A\bar{c}_0 = (5, 3, 1, 3), \quad A'\bar{b}_0 = \left(4+\sqrt{3}, 1+\sqrt{3}, -1, -1\right), \\ \alpha_{10} = 4+\sqrt{3}, \\ \bar{c}_1 = \left(1-\sqrt{3}, 3, 1, 3\right), \quad \bar{b}_1 = \left(0, \frac{4}{3}, -1, -1\right) \end{array}\right\} \quad (41)$$

и

$$(\bar{c}_1, \bar{b}_1) = 0. \quad (42)$$

Если минимальный многочлен матрицы A имеет степень m, то векторы $\bar{c}_0$, $A\bar{c}_0$, ..., $A^m\bar{c}_0$ и $\bar{b}_0$, $A'\bar{b}_0$, ..., $A'^m\bar{b}_0$ линейно зависимы. В силу этого наш процесс обязательно закончится не позже чем через $k \leqslant m$ шагов. При этом в случаях а) и б) мы найдем линейную зависимость между векторами $\bar{c}_0$, $A\bar{c}_0$, ..., $A^k\bar{c}_0$ или $\bar{b}_0$, $A'\bar{b}_0$, ..., $A'^k\bar{b}_0$, а следовательно, и минимальный многочлен матрицы A или его делитель. Случай в) может встретиться лишь как исключение при неудачном выборе начальных векторов $\bar{c}_0$ и $\bar{b}_0$ и его всегда можно избежать, выбрав другие начальные векторы.

Предполагая, что мы получили $c_k = 0$, или $\bar{b}_k = 0$, последовательно находим минимальный многочлен A или его делитель по формулам:

$$\left.\begin{aligned} &P_0(\lambda) = 1, \\ &P_1(\lambda) = (\lambda - \alpha_{10}) P_0(\lambda), \\ &P_2(\lambda) = (\lambda - \alpha_{21}) P_1(\lambda) - \alpha_{20} P_0(\lambda), \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &P_{k-1}(\lambda) = (\lambda - \alpha_{k-1,\,k-2}) P_{k-2}(\lambda) - \alpha_{k-1,\,k-3} P_{k-3}(\lambda), \\ &P_k(\lambda) = (\lambda - \alpha_{k,\,k-1}) P_{k-1}(\lambda) - \alpha_{k,\,k-2} P_{k-2}(\lambda). \end{aligned}\right\} \tag{43}$$

В частности, в рассмотренных нами примерах будем иметь: в случае (30)

$$\left.\begin{aligned} &P_0(\lambda) = 1, \\ &P_1(\lambda) = \lambda - 3, \\ &P_2(\lambda) = (\lambda - 6)(\lambda - 3) - 6 = \lambda^2 - 9\lambda + 12, \\ &P_3(\lambda) = (\lambda - 3)(\lambda^2 - 9\lambda + 12) + 6(\lambda - 3) = \lambda^3 - 12\lambda^2 + 45\lambda - 54, \end{aligned}\right\} \tag{44}$$

в случае (38)

$$\left.\begin{aligned} &P_0(\lambda) = 1, \\ &P_1(\lambda) = \lambda - 5, \\ &P_2(\lambda) = (\lambda - 4)(\lambda - 5) - 2 = \lambda^2 - 9\lambda + 18. \end{aligned}\right\} \tag{45}$$

Такой способ получения минимального многочлена или его делителя будем называть *методом Ланцоша*.

Пусть процесс, осуществляемый по методу Ланцоша, продолжается до $k = n - 1$. Рассмотрим матрицы

$$C = \begin{pmatrix} c_{01} & c_{11} & \cdots & c_{n-1,\,1} \\ c_{02} & c_{12} & \cdots & c_{n-1,\,2} \\ \cdots & \cdots & \cdots & \cdots \\ c_{0n} & c_{1n} & \cdots & c_{n-1,\,n} \end{pmatrix}, \quad B = \begin{pmatrix} b_{01} & b_{11} & \cdots & b_{n-1,\,1} \\ b_{02} & b_{12} & \cdots & b_{n-1,\,2} \\ \cdots & \cdots & \cdots & \cdots \\ b_{0n} & b_{1n} & \cdots & b_{n-1,\,n} \end{pmatrix}, \tag{46}$$

где c_{ij} и b_{ij} — соответственно компоненты векторов $\bar{c}_i$ и $\bar{b}_i$. В силу биортогональности будем иметь:

$$B'C = \begin{pmatrix} \delta_0 & 0 & \cdots & 0 \\ 0 & \delta_1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \delta_{n-1} \end{pmatrix} = D, \tag{47}$$

где $\delta_i = (\bar{c}_i, \bar{b}_i)$. Следовательно, $C^{-1} = D^{-1}B'$ и $B^{-1} = D^{-1}C'$. Далее, так как

$$\left.\begin{aligned} &(\bar{b}_i, A\bar{c}_i) = \alpha_{i+1,\,i}(\bar{c}_i, \bar{b}_i) = \alpha_{i+1,\,i}\delta_i, \\ &(\bar{b}_i, A\bar{c}_{i-1}) = (\bar{b}_{i-1}, A\bar{c}_i) = (\bar{c}_i, \bar{b}_i) = \delta_i, \\ &(\bar{b}_i, A\bar{c}_j) = 0 \quad \text{при} \quad |i - j| > 1, \end{aligned}\right\} \tag{48}$$

то

$$B'AC=\begin{pmatrix} \alpha_{10}\delta_0 & \delta_1 & 0 & \dots & 0 & 0 \\ \delta_1 & \alpha_{21}\delta_1 & \delta_2 & \dots & 0 & 0 \\ 0 & \delta_2 & \alpha_{32}\delta_2 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \delta_{n-1} & \alpha_{n,n-1}\delta_{n-1} \end{pmatrix}. \tag{49}$$

Поэтому

$$C^{-1}AC=D^{-1}B'AC\begin{pmatrix} \alpha_{10} & \alpha_{20} & 0 & \dots & 0 & 0 \\ 1 & \alpha_{21} & \alpha_{31} & \dots & 0 & 0 \\ 0 & 1 & \alpha_{32} & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & \alpha_{n,\,n-1} \end{pmatrix}. \tag{50}$$

Таким образом, в рассматриваемом случае наш процесс эквивалентен приведению матрицы A к тридиагональной форме. Если производить процесс Ланцоша без указанных упрощений, то он будет эквивалентен приведению матрицы A к верхней треугольной форме. В симметрическом случае всегда получим тридиагональную форму.

Случай $k<n-1$ потребует выбора новых начальных векторов, как это указывалось в начале параграфа. При этом матрица приводится к

$$\begin{pmatrix} A_1 & & & 0 \\ & A_2 & & \\ & & \ddots & \\ 0 & & & A_k \end{pmatrix}, \tag{51}$$

где клетки A_i имеют вид, указанный выше.

Вернемся еще раз к вопросу о применении метода Ланцоша в случае симметрической матрицы A. Этот случай особенно выгоден для этого метода.

Прежде всего отметим, что для симметрической матрицы

$$\alpha_{k+1,\,k-1}=\frac{(A\bar{c}_k,\ \bar{c}_{k-1})}{(\bar{c}_{k-1},\ \bar{c}_{k-1})}=\frac{(\bar{c}_k,\ A\bar{c}_{k-1})}{(\bar{c}_{k-1},\ \bar{c}_{k-1})}=\frac{(\bar{c}_k,\ \bar{c}_k)}{(\bar{c}_{k-1},\ \bar{c}_{k-1})}\geqslant 0, \tag{52}$$

причем знак равенства будет достигаться лишь при $\bar{c}_k=0$. Далее, из условия

$$P_{k+1}(\lambda)=(\lambda-\alpha_{k+1,\,k})P_k(\lambda)-\alpha_{k+1,\,k-1}P_{k-1}(\lambda) \tag{53}$$

следует, что никакие два многочлена $P_k(\lambda)$ и $P_{k+1}(\lambda)$ ($k=0, 1, \dots, m-1$) не могут обращаться в нуль при одном и том же значении $\lambda=\lambda'$. Действительно, если бы $P_{k+1}(\lambda')=P_k(\lambda')=0$, то из (52) и (53) следовало бы $P_{k-1}(\lambda')=0$. Повторяя эти рассуждения для многочленов $P_k(\lambda)$ и $P_{k-1}(\lambda)$, мы пришли бы к выводу, что $P_{k-2}(\lambda')=0$. Продолжая дальше, мы пришли бы в конце концов к заключению, что $P_0(\lambda')=0$. Но это невозможно, так как $P_0(\lambda)\equiv 1$.

Изучим теперь взаимное расположение корней $P_k(\lambda)$. Многочлен $P_1(\lambda)$ имеет единственный корень $\lambda_1^{(1)} = \alpha_{10}$. При этом $P_2(\lambda_1^{(1)}) = -\alpha_{20} P_0 < 0$. Следовательно, квадратный многочлен $P_2(\lambda)$, положительный для достаточно больших по абсолютной величине значений λ, будет обращаться в нуль в точках $\lambda_1^{(2)}$ и $\lambda_2^{(2)}$, $\lambda_1^{(2)} < \lambda_1^{(1)} < \lambda_2^{(2)}$. Для $P_3(\lambda)$ будем иметь: $P_3(\lambda_1^{(2)}) = -\alpha_{31} P_1(\lambda_1^{(2)}) > 0$, $P_3(\lambda_2^{(2)}) = -\alpha_{31} P_1(\lambda_2^{(2)}) < 0$. Так как $P_3(\lambda)$ отрицателен при достаточно больших по абсолютной величине, но отрицательных λ и положителен при достаточно больших положительных λ, то имеются три корня $P_3(\lambda)$: $\lambda_1^{(3)}$, $\lambda_2^{(3)}$, $\lambda_3^{(3)}$, таких, что $\lambda_1^{(3)} < \lambda_1^{(2)} < \lambda_2^{(3)} < \lambda_2^{(2)} < \lambda_3^{(3)}$. Теперь нетрудно по индукции показать, что все корни многочленов $P_k(\lambda)$: $\lambda_1^{(k)}, \lambda_2^{(k)}, \ldots, \lambda_k^{(k)}$, действительны, различны и удовлетворяют условиям:

$$\lambda_1^{(k)} < \lambda_1^{(k-1)} < \lambda_2^{(k)} < \lambda_2^{(k-1)} < \ldots < \lambda_{k-1}^{(k)} < \lambda_{k-1}^{(k-1)} < \lambda_k^{(k)}. \tag{54}$$

Пусть это выполнено для $k = 1, 2, 3, \ldots, l$. Тогда в силу (53)

$$P_{l+1}(\lambda_i^{(l)}) = -\alpha_{l+1,\, l-1} P_{l-1}(\lambda_i^{(l)}) \tag{55}$$

и, следовательно, в силу (52)

$$P_{l+1}(\lambda_i^{(l)}) P_{l-1}(\lambda_i^{(l)}) < 0. \tag{56}$$

По предположению индукции $P_{l-1}(\lambda)$ не имеет кратных корней, поэтому $P_{l-1}(\lambda_i^{(l)})$ меняет знак при переходе от i к $i+1$. Следовательно, в силу (56) это верно и для $P_{l+1}(\lambda_i^{(l)})$. Так как степени $P_{l+1}(\lambda)$ и $P_{l-1}(\lambda)$ имеют одинаковую четность, то эти многочлены должны иметь одинаковые знаки левее и правее множества всех их действительных корней. $\lambda_1^{(l)}$ лежит левее всех нулей $P_{l-1}(\lambda)$, а $\lambda_l^{(l)}$ — правее всех нулей $P_{l-1}(\lambda)$. Поэтому $P_{l+1}(\lambda)$ должен обратиться в нуль левее $\lambda_1^{(l)}$ и правее $\lambda_l^{(l)}$. Так как он должен обратиться в нуль между каждыми двумя соседними $\lambda_i^{(l)}$, то утверждение доказано.

Из доказанного следует, что совокупность многочленов $P_k(\lambda)$, $k = 0, 1, 2, \ldots, m$ образует систему Штурма (см. главу 7, § 1). Поэтому мы можем использовать ее для отделения корней многочлена $P_m(\lambda)$, как это указано в предыдущей главе.

Подсчитаем число операций умножения и деления, необходимых для получения характеристического многочлена симметрической матрицы A порядка n по способу Ланцоша. Пусть процесс, начинающийся с вектора $\bar{c}_0$, может быть продолжен вплоть до $\bar{c}_n$. Конечно, $\bar{c}_n = 0$. Вычисление $A\bar{c}_0$ потребует n^2 умножений. Получение $\alpha_{10} = (A\bar{c}_0, \bar{c}_0)/(\bar{c}_0, \bar{c}_0)$ потребует $2n+1$ умножений и делений. Поэтому переход от $\bar{c}_0$ к $\bar{c}_1 = A\bar{c}_0 - \alpha_{10}\bar{c}_0$ потребует всего $n^2 + 3n + 1$ операций умножения и деления. На следующем шаге, при переходе от $\bar{c}_1$ к $\bar{c}_2$, придется затратить n^2 операций умножения для получения $A\bar{c}_1$, $3n+2$ операций умножения и деления для получения

коэффициентов $\alpha_{21} = (A\bar{c}_1, \bar{c}_1)/(\bar{c}_1, \bar{c}_1)$ и $\alpha_{20} = (A\bar{c}_1, \bar{c}_0)/(\bar{c}_0, \bar{c}_0)$, $2n$ операций умножения для получения $\bar{c}_2 = A\bar{c}_1 - \alpha_{21}\bar{c}_1 - \alpha_{20}\bar{c}_0$. Таким образом, всего второй шаг потребует $n^2 + 5n + 2$ операций умножения и деления. Столько же потребует и каждый следующий шаг. Всего для получения последовательности векторов $\bar{c}_0, \bar{c}_1, \ldots, \bar{c}_{n-1}$ потребуется

$$(n^2 + 3n + 1) + (n - 2)(n^2 + 5n + 2) = n^3 + 4n^2 - 5n - 3 \quad (57)$$

операций умножения и деления. Нам еще придется подсчитать $\alpha_{n, n-1}$ и $\alpha_{n, n-2}$, для чего нужно $3n + 2$ операций умножения и деления. Наконец, получение многочлена $P_n(\lambda)$ требует

$$1 + 3 + 5 + \ldots + (2n - 3) = (n - 1)^2 \quad (58)$$

операций умножения. Таким образом, всего для получения характеристического многочлена матрицы A потребуется

$$n^3 + 4n^2 - 5n - 3 + 3n + 2 + (n - 1)^2 = n^3 + 5n^2 - 4n \quad (59)$$

операций умножения и деления.

2. Отыскание собственных векторов. В заключение остановимся на отыскании собственных векторов, Предположим, что, применяя метод Ланцоша, мы получили $\bar{c}_m = 0$. Пусть λ_i — какой-нибудь корень минимального многочлена вектора $\bar{c}_0$. Тогда будем разыскивать собственный вектор, соответствующий этому собственному значению в виде

$$\bar{x}_i = \gamma_0\bar{c}_0 + \gamma_1\bar{c}_1 + \ldots + \gamma_{m-1}\bar{c}_{m-1}. \quad (60)$$

Условие $A\bar{x}_i = \lambda_i\bar{x}_i$ дает

$$\begin{aligned}
&\gamma_0(\bar{c}_1 + \alpha_{10}\bar{c}_0) + \gamma_1(c_2 + \alpha_{21}\bar{c}_1 + \alpha_{20}\bar{c}_0) + \ldots \\
&\quad \ldots + \gamma_{m-2}(\bar{c}_{m-1} + \alpha_{m-1,\, m-2}\bar{c}_{m-2} + \alpha_{m-1,\, m-3}\bar{c}_{m-2}) + \\
&\quad + \gamma_{m-1}(\alpha_{m,\, m-1}\bar{c}_{m-1} + \alpha_{m,\, m-2}\bar{c}_{m-2}) = \\
&\quad = \lambda_i\gamma_0\bar{c}_0 + \lambda_i\gamma_1\bar{c}_1 + \ldots + \lambda_i\gamma_{m-1}\bar{c}_{m-1}.
\end{aligned} \quad (61)$$

В силу линейной независимости векторов $\bar{c}_0, \bar{c}_1, \ldots, \bar{c}_{m-1}$ из (61) следует:

$$\left.\begin{aligned}
&\gamma_0(\lambda_i - \alpha_{10}) - \alpha_{20}\gamma_1 = 0, \\
&\gamma_1(\lambda_i - \alpha_{21}) - \alpha_{31}\gamma_2 - \gamma_0 = 0, \\
&\gamma_2(\lambda_i - \alpha_{32}) - \alpha_{42}\gamma_3 - \gamma_1 = 0, \\
&\quad \ldots\ldots\ldots\ldots\ldots \\
&\gamma_{m-2}(\lambda_i - \alpha_{m-1,\, m-2}) - \alpha_{m,\, m-2}\gamma_{m-1} - \gamma_{m-3} = 0, \\
&\gamma_{m-1}(\lambda_i - \alpha_{m,\, m-1}) - \gamma_{m-2} = 0.
\end{aligned}\right\} \quad (62)$$

Коэффициент γ_{m-1} должен быть отлучен от нуля, так как в противном случае и все остальные коэффициенты γ_i были бы равны нулю. Положим, например, $\gamma_{m-1}=1$. Тогда остальные коэффициенты γ_i последовательно находятся из равенств:

$$\left.\begin{aligned}
\gamma_{m-2} &= (\lambda_i - \alpha_{m,\,m-1}),\\
\gamma_{m-3} &= \gamma_{m-2}(\lambda_i - \alpha_{m-1,\,m-2}) - \alpha_{m,\,m-2},\\
&\cdots\cdots\cdots\cdots\cdots\cdots\\
\gamma_0 &= \gamma_1(\lambda_i - \alpha_{21}) - \alpha_{31}\gamma_2.
\end{aligned}\right\} \tag{63}$$

Как и для метода Крылова, первое из равенств (62) будет следствием остальных и условия, что λ_i является корнем минимального многочлена вектора $\bar{c}_0$.

Найдем собственные векторы матрицы (44) § 2. Если за вектор $\bar{c}_0$ принять (30), то вычисления при $\lambda_i=3$ дают

$$\left.\begin{aligned}
\gamma_2 &= 1,\\
\gamma_1 &= \lambda_i - \alpha_{32} = 0,\\
\gamma_0 &= \gamma_1(\lambda_i - \alpha_{21}) - \alpha_{31} = 6
\end{aligned}\right\} \tag{64}$$

и

$$\bar{x}_i = (0,\ 6,\ 0,\ 0) + (0,\ 0,\ 6,\ 6) = (0,\ 6,\ 6,\ 6). \tag{65}$$

При $\lambda_i=6$ соответственно получим:

$$\left.\begin{aligned}
\gamma_2 &= 1,\\
\gamma_1 &= \lambda_i - \alpha_{32} = 3,\\
\gamma_0 &= \gamma_1(\lambda_i - \alpha_{21}) - \alpha_{31} = 6
\end{aligned}\right\} \tag{66}$$

и

$$\bar{x}_i = (0,\ 6,\ 0,\ 0) + (6,\ 0,\ -6,\ 0) + (0,\ 0,\ 6,\ 6) = (6,\ 6,\ 0,\ 6). \tag{67}$$

Можно ввести в рассмотрение многочлены:

$$\left.\begin{aligned}
q_0(\lambda) &= 1,\\
q_1(\lambda) &= (\lambda - \alpha_{m,\,m-1})\,q_0(\lambda),\\
q_2(\lambda) &= (\lambda - \alpha_{m-1,\,m-2})\,q_1(\lambda) - \alpha_{m,\,m-2}q_0(\lambda),\\
&\cdots\cdots\cdots\cdots\cdots\cdots\\
q_{m-1}(\lambda) &= (\lambda - \alpha_{2,\,1})\,q_{m-2}(\lambda) - \alpha_{31}q_{m-3}(\lambda),\\
q_m(\lambda) &= (\lambda - \alpha_{10})\,q_{m-1}(\lambda) - \alpha_{20}q_{m-2}(\lambda).
\end{aligned}\right\} \tag{68}$$

Тогда собственный вектор $\bar{x}_i$, соответствующий собственному значению λ_i, можно записать в виде

$$\bar{x}_i = q_{m-1}(\lambda_i)\,\bar{c}_0 + q_{m-2}(\lambda_i)\,\bar{c}_1 + \ldots + q_1(\lambda_i)\,\bar{c}_{m-2} + q_0(\lambda_i)\,\bar{c}_{m-1}. \tag{69}$$

Многочлен $q_m(\lambda)$ в (68) совпадает с многочленом $P_m(\lambda)$.

§ 4. Метод Данилевского

Довольно простой и изящный способ получения характеристического многочлена дал А. М. Данилевский. Суть его метода состоит в преобразовании уравнения

$$D(\lambda) = |A - \lambda I| = 0 \tag{1}$$

к виду

$$\begin{vmatrix} p_1 - \lambda & p_2 & p_3 & \dots & p_{n-1} & p_n \\ 1 & -\lambda & 0 & \dots & 0 & 0 \\ 0 & 1 & -\lambda & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & -\lambda \end{vmatrix} = 0. \tag{2}$$

При этом определитель (2) легко раскрывается, и мы получим:

$$D(\lambda) = (-1)^n \left[\lambda^n - p_1\lambda^{n-1} - p_2\lambda^{n-2} - \dots - p_n\right]. \tag{3}$$

Проиллюстрируем ход вычислений по методу Данилевского на примере матрицы четвертого порядка

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix}. \tag{4}$$

Эта матрица должна быть преобразована к виду

$$\begin{pmatrix} p_1 & p_2 & p_3 & p_4 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \tag{5}$$

(нормальная форма Фробениуса) преобразованиями подобия.

Делим все элементы третьего столбца на a_{43}. Если $a_{43} = 0$, то предварительно находим среди элементов a_{41} и a_{42} отличный от нуля. Пусть это оказался a_{42}. Тогда меняем местами вторую и третью строки и второй и третий столбцы. Если $a_{41} = a_{42} = a_{43} = 0$, то характеристический многочлен матрицы (4) примет вид

$$(a_{44} - \lambda) \begin{vmatrix} a_{11} - \lambda & a_{12} & a_{13} \\ a_{21} & a_{22} - \lambda & a_{23} \\ a_{31} & a_{32} & a_{33} - \lambda \end{vmatrix}, \tag{6}$$

и наша задача сведется к отысканию характеристического многочлена матрицы третьего порядка.

Вычтем теперь из i-го столбца ($i = 1, 2, 4$) полученной матрицы

$$\begin{Vmatrix} a_{11} & a_{12} & \frac{a_{13}}{a_{43}} & a_{14} \\ a_{21} & a_{22} & \frac{a_{23}}{a_{43}} & a_{24} \\ a_{31} & a_{32} & \frac{a_{33}}{a_{43}} & a_{34} \\ a_{41} & a_{42} & 1 & a_{44} \end{Vmatrix} \tag{7}$$

третий столбец, умноженный на a_{4i}. При этом матрица примет вид

$$\begin{Vmatrix} a_{11} - \frac{a_{13}a_{41}}{a_{43}} & a_{12} - \frac{a_{13}a_{42}}{a_{43}} & \frac{a_{13}}{a_{43}} & a_{14} - \frac{a_{13}a_{44}}{a_{43}} \\ a_{21} - \frac{a_{23}a_{41}}{a_{43}} & a_{22} - \frac{a_{23}a_{42}}{a_{43}} & \frac{a_{23}}{a_{43}} & a_{24} - \frac{a_{23}a_{44}}{a_{43}} \\ a_{31} - \frac{a_{33}a_{41}}{a_{43}} & a_{32} - \frac{a_{33}a_{42}}{a_{43}} & \frac{a_{33}}{a_{43}} & a_{34} - \frac{a_{33}a_{44}}{a_{43}} \\ 0 & 0 & 1 & 0 \end{Vmatrix}. \tag{8}$$

Наш процесс эквивалентен умножению матрицы (4) справа на матрицу

$$B_1 = \begin{Vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -\frac{a_{41}}{a_{43}} & -\frac{a_{42}}{a_{43}} & \frac{1}{a_{43}} & -\frac{a_{44}}{a_{43}} \\ 0 & 0 & 0 & 1 \end{Vmatrix}. \tag{9}$$

Чтобы получить матрицу, подобную исходной, мы должны умножить (8) слева на матрицу B_1^{-1}, равную

$$B_1^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ a_{41} & a_{42} & a_{43} & a_{44} \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{10}$$

При этом получим:

$$B_1^{-1}AB_1 = \begin{Vmatrix} a_{11} - \frac{a_{13}a_{41}}{a_{43}} & a_{12} - \frac{a_{13}a_{42}}{a_{43}} & \frac{a_{13}}{a_{43}} & a_{14} - \frac{a_{13}a_{44}}{a_{43}} \\ a_{21} - \frac{a_{23}a_{41}}{a_{43}} & a_{22} - \frac{a_{23}a_{42}}{a_{43}} & \frac{a_{23}}{a_{43}} & a_{24} - \frac{a_{23}a_{44}}{a_{43}} \\ b_{31} & b_{32} & b_{33} & b_{34} \\ 0 & 0 & 1 & 0 \end{Vmatrix}. \tag{11}$$

Умножение на B_1^{-1} не изменяет первой, второй и четвертой строк. Третья же строка получается путем сложения строк (8), умноженных соответственно на a_{41}, a_{42}, a_{43} и a_{44}.

На следующем этапе преобразуем третью строку. Это достигается путем умножения матрицы (10) справа на матрицу

$$B_2=\begin{pmatrix} 1 & 0 & 0 & 0 \\ -\frac{b_{31}}{b_{32}} & \frac{1}{b_{32}} & -\frac{b_{33}}{b_{32}} & -\frac{b_{34}}{b_{32}} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{12}$$

и последующего умножения слева на матрицу B_2^{-1}, равную

$$B_2^{-1}=\begin{pmatrix} 1 & 0 & 0 & 0 \\ b_{31} & b_{32} & b_{33} & b_{34} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{13}$$

Ход вычислений подобен предыдущему. Матрица A будет приведена к виду

$$\begin{pmatrix} c_{11} & c_{12} & c_{13} & c_{14} \\ c_{21} & c_{22} & c_{23} & c_{24} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \tag{14}$$

Наконец, умножая (14) справа на

$$B_3=\begin{pmatrix} \frac{1}{c_{21}} & -\frac{c_{22}}{c_{21}} & -\frac{c_{23}}{c_{21}} & -\frac{c_{24}}{c_{21}} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{15}$$

и слева на

$$B_3^{-1}=\begin{pmatrix} c_{21} & c_{22} & c_{23} & c_{24} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{16}$$

мы придем к нормальной форме Фробениуса.

Приведем числовой пример. Возьмем уже использованную в главе 6 матрицу A:

$$A=\begin{pmatrix} 6{,}1818 & 0{,}1818 & 0{,}3141 & 0{,}1415 & 0{,}1516 & 0{,}2141 \\ 0{,}1818 & 7{,}1818 & 0{,}2141 & 0{,}1815 & 0{,}1526 & 0{,}3141 \\ 0{,}3141 & 0{,}2141 & 8{,}2435 & 0{,}1214 & 0{,}2516 & 0{,}2618 \\ 0{,}1415 & 0{,}1815 & 0{,}1214 & 0{,}3141 & 0{,}3145 & 0{,}6843 \\ 0{,}1516 & 0{,}1526 & 0{,}2516 & 0{,}3145 & 5{,}3116 & 0{,}8998 \\ 0{,}2141 & 0{,}3114 & 0{,}2613 & 0{,}6343 & 0{,}8998 & 4{,}1313 \end{pmatrix}. \tag{17}$$

Вычисления сведены в следующую таблицу:

№	$(B_i^{-1})'$	1	2	3	4	5	6	S	S'
1		6,1818	0,1818	0,3141	0,1415	0,1516	0,2141	7,1849	
2		0,1818	7,1818	0,2141	0,1815	0,1526	0,3141	8,2232	
3		0,3141	0,2141	8,2435	0,1214	0,2516	0,2618	9,4065	
4		0,1415	0,1815	0,1214	9,3141	0,3145	0,6843	10,7573	
5		0,1516	0,1526	0,2516	0,3145	5,3116	0,8998	7,0817	
6		0,2141	0,3114	0,2618	0,6843	0,8998	4,1313	6,5027	
7 (6′)	B_1^{-1}	0,2141	0,3114	0,2618	0,6843	0,8998	4,1313	6,5027	
8	B_1	—0,2379	—0,3461	—0,2910	—0,7605	(—1) 1,1114	—4,5914	7,2268	
9	0,2141	6,1457	0,1293	0,2700	0,0262	0,1685	—0,4820	6,2577	6,0893
10	0,3114	0,1455	7,1290	0,1697	0,0654	0,1696	—0,3865	7,2927	7,1231
11	0,2618	0,2542	0,1270	8,1703	—0,0699	0,2796	—0,8934	7,8678	7,5882
12	0,6843	0,0667	0,0727	0,0299	9,0749	0,3495	—0,7597	8,8340	8,4844
13	0,8998	—1,1120	—1,6857	—1,2941	—3,7250	5,9031	—23,4879	—25,4016	—31,3047
14	4,1313	0	0	0	0	1	0	1	
15 (13′)	B_2^{-1}	0,4727	0,8139	1,1057	2,8659	9,8442	—22,1117	—7,0093	
16	B_2	—0,1649	—0,2840	—0,3858	(—1) 0,3489	—3,4349	7,7154	2,4458	
17	0,4727	6,1414	0,1219	0,2599	0,0091	0,0785	—0,2799	6,3309	6,3218
18	0,8139	0,1347	7,1104	0,1445	0,0228	—0,0550	0,1181	7,4755	7,4527
19	1,1057	0,2657	0,1469	8,1973	—0,0244	0,5197	—1,4327	7,6725	7,6968
20	2,8659	—1,4298	—2,5046	—3,4712	3,1665	—30,8219	69,2568	34,1958	31,0294
21	9,8442	0	0	0	1	0	0	1	
22	—22,1117	0	0	0	0	1	0	1	
23 (20′)	B_3^{-1}	—0,7912	—1,1707	—0,6439	18,9150	—109,8772	196,8627	103,2947	
24	B_3	—1,2288	—1,8181	(—1) —1,5530	29,3757	—170,6432	305,7348	160,4204	

Продолжение

№	$(B_i^{-})'$	1	2	3	4	5	6	S	S'
25	—0,7912	5,8220	0,3506	—0,4036	7,6438	—44,2717	79,1806	47,6205	4848,0242
26	—1,1707	—0,0429	6,8477	—0,2244	4,2676	—24,7128	44,2968	30,4319	30,6562
27	—0,6439	—9,8071	—14,7566	—12,7307	240,7770	—1398,2938	2504,7672	1309,9560	1322,6860
28	18,9150	0	0	1	0	0	0	1	
29	—109,8772	0	0	0	1	0	0	1	
30	196,8627	0	0	0	0	1	0	1	
31 (27′)	B_4^{-1}	1,7586	1,7626	27,6943	—275,9574	1161,1832	—1727,3256	—810,8843	
32	B_4	—0,9977	(—1) 0,5673	—15,7122	156,5627	—658,7900	979,9873	460,0501	
33	1,7586	6,1718	—0,1989	5,1051	—47,2471	186,7001	—264,4029	—113,8719	—113,6731
34	1,7626	—6,8749	3,8850	—107,8168	1076,3620	—4535,9092	6754,9558	3184,6019	3180,7170
35	27,6943	0	1	0	0	0	0	1	
36	—275,9574	0	0	1	0	0	0	1	
37	1161,1832	0	0	0	1	0	0	1	
38	—1727,3256	0	0	0	0	1	0	1	
39 (34′)	B_5^{-1}	—1,2640	34,1922	—457,0175	2975,2901	—9393,9883	11441,3062	4598,5187	
40	B_5	(—1) —0,7911	27,0508	—361,5645	2353,8688	—7431,9528	9051,6663	3638,0686	
41	—1,2640	—4,8828	166,7532	—2226,3987	14480,3604	—45681,8262	55600,6712	22334,6771	22339,5599
42	34,1922	1	0	0	0	0	0	1	
43	—457,0175	0	1	0	0	0	0	1	
44	2975,2901	0	0	1	0	0	0	1	
45	—9393,9883	0	0	0	1	0	0	1	
46	11441,3062	0	0	0	0	1	0	1	
		40,3641	—667,7935	5789,4581	—27697,1638	69183,1345	—70279,2484	—23631,2492	

Теперь поясним схему. В первом столбце помещена нумерация строк. Столбцы, обозначенные цифрами 1—6, предназначены для элементов исходной матрицы A, матриц B_i и B_i^{-1}, промежуточных матриц $B_i^{-1} \ldots B_2^{-1}B_1^{-1}AB_1B_2 \ldots B_i$ и окончательной матрицы, получающихся в процессе вычислений по методу Данилевского. Последние два столбца — контрольные.

Первые шесть строк (1—6) в столбцах 1—6 использованы для элементов исходной матрицы A. В столбце S стоят суммы элементов A по строкам.

Строки 7 и 8 предназначены для элементов B_1 и B_1^{-1}. Ввиду особенностей строения этих матриц, достаточно записать только их пятые строки. В столбцах 1—6 седьмой строки записаны элементы пятой строки B_1^{-1}, которые просто равны элементам шестой строки матрицы A. В столбцах 1—6 восьмой строки помещены элементы пятой строки матрицы B_1, равные соответственно $b_{51} = -\frac{a_{61}}{a_{65}}$, $b_{52} = -\frac{a_{62}}{a_{65}}$, $b_{53} = -\frac{a_{63}}{a_{65}}$, $b_{54} = -\frac{a_{64}}{a_{65}}$, $b_{55} = \frac{1}{a_{65}}$, $b_{56} = -\frac{a_{66}}{a_{65}}$. Элемент a_{65}, на который производится деление, в схеме подчеркнут. В столбце S восьмой строки стоит результат деления S, стоящего в шестой строке, на a_{65} с обратным знаком. Поэтому он должен равняться сумме остальных элементов этой строки, если заменить b_{55} на —1 (в схеме —1 показана в скобках).

Следующим этапом будет умножение A на B_1. Соответствующие элементы c_{ik} помещены в строках 9—14 и столбцах 1—6. При этом в столбце 5 будут помещаться результаты деления элементов a_{i5} на a_{65} или, что то же самое, результаты умножения a_{i5} на b_{55}. Остальные элементы c_{ik} вычисляются по формулам

$$c_{ik} = a_{ik} + b_{5k}a_{i5}. \tag{18}$$

После этого шестая строка примет нужный нам вид. В столбце S, как и всегда, помещаем суммы элементов строк, а в столбце S' — результаты вычислений со столбцом строк 1—6 по формулам, аналогичным (18). Контроль будет заключаться в том, что сумма столбца S' и столбца 5 должна давать столбец S.

После этого производим умножение на B_1^{-1}. При этом изменяется только пятая строка определителя AB_1, т. е. строка 13 нашей схемы. Пятая строка определителя $B_1^{-1}AB_1$ равна сумме произведений строк AB_1 на соответствующие элементы пятой строки B_1^{-1}. Для удобства элементы B_1^{-1}, на которые умножаются строки AB_1, выписаны во втором столбце схемы против соответствующих им строк. Результат вычислений помещен в строке 15. Эта строка будет одновременно являться четвертой строкой матрицы B_2^{-1}.

Дальнейшие вычисления происходят аналогично. Роль a_{65} теперь играет элемент строки 15 и столбца 4. Четвертая строка B_2

помещается в строке 16 схемы. Матрица $B_2^{-1}B_1^{-1}AB_1B_2$ помещается в строках 17—23 и т. д.

В данном случае искомый характеристический многочлен будет иметь вид

$$D(\lambda) = \lambda^6 - 40{,}3641\lambda^5 + 667{,}7935\lambda^4 - 5789{,}4581\lambda^3 + \\ + 27697{,}1638\lambda^2 - 69183{,}1345\lambda + 70279{,}2484. \quad (19)$$

Интересно отметить, что след матрицы A, который должен равняться коэффициенту при λ^5 с обратным знаком, в нашем случае равен 40,3641. Совпадение полное.

Приведенная нами схема не является лучшей. Однако она довольно удобна и проста для объяснений. Она применима для матриц любого порядка.

Произведем подсчет числа операций умножения и деления, необходимых для получения характеристического многочлена матрицы порядка n по приведенной выше схеме, учитывая применение контроля.

Получение B_1 потребует $n+1$ операций деления. Вычисление $(n-1)$-го столбца AB_1 потребует $n-1$ умножений. Получение остальной части AB_1 (с включением контрольного столбца) потребует $n(n-1)$ умножений. Наконец, умножение AB_1 на B_1^{-1} потребует $(n-1)(n+1)$ умножений. На следующем шаге, при переходе от $B_1^{-1}AB_1$ и $B_2^{-1}B_1^{-1}AB_1B_2$, нам потребуется соответственно $n+1$ делений, $n-2$ умножений, $n(n-2)$ умножений и $(n-2)(n+1)$ умножений. Так же производится подсчет и дальше. Таким образом, всего будет нужно

$$(n-1)(n+1) + [(n-1) + \\ + (n-2) + \ldots + 2 + 1](2n+2) = (n^2-1)(n+1) \quad (20)$$

операций умножения и деления.

1. Видоизменение метода Данилевского. Рассмотрим теперь одно видоизменение метода Данилевского. Опять для иллюстрации воспользуемся матрицей четвертого порядка (4). Вместо B_1 и B_1^{-1} возьмем теперь матрицы

$$C_1 = \begin{pmatrix} 0 & 0 & 0 & a_{14} \\ 1 & 0 & 0 & a_{24} \\ 0 & 1 & 0 & a_{34} \\ 0 & 0 & 1 & a_{44} \end{pmatrix} = \left(\begin{array}{c|c} 0 & t \\ \hline I & \tau \end{array}\right) \quad (21)$$

и

$$C_1^{-1} = \begin{pmatrix} -a_{24}/a_{14} & 1 & 0 & 0 \\ -a_{34}/a_{14} & 0 & 1 & 0 \\ -a_{44}/a_{14} & 0 & 0 & 1 \\ 1/a_{14} & 0 & 0 & 0 \end{pmatrix} = \left(\begin{array}{c|c} -\frac{\tau}{t} & I \\ \hline \frac{1}{t} & 0 \end{array}\right). \quad (22)$$

Произведение $C_1^{-1}A$ равно

$$C_1^{-1}A = \begin{pmatrix} a'_{11} & a'_{12} & a'_{13} & 0 \\ a'_{21} & a'_{22} & a'_{23} & 0 \\ a'_{31} & a'_{32} & a'_{33} & 0 \\ a'_{41} & a'_{42} & a'_{43} & 1 \end{pmatrix}, \tag{23}$$

где

$$\left.\begin{aligned} a'_{ik} &= a_{i+1,\,k} - a_{ik}\frac{a_{i+1,\,4}}{a_{14}} \qquad (i,\ k = 1,\ 2,\ 3); \\ a'_{4k} &= \frac{a_{1k}}{a_{14}} \qquad (k = 1,\ 2,\ 3,\ 4). \end{aligned}\right\} \tag{24}$$

Умножение $C_1^{-1}A$ справа на C_1 даст

$$C_1^{-1}AC_1 = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & 0 & a_{14}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & 0 & a_{24}^{(1)} \\ a_{31}^{(1)} & a_{32}^{(1)} & 0 & a_{34}^{(1)} \\ a_{41}^{(1)} & a_{42}^{(1)} & 1 & a_{44}^{(1)} \end{pmatrix}, \tag{25}$$

где

$$\left.\begin{aligned} a_{ik}^{(1)} &= a'_{ik+1} \qquad (i = 1,\ 2,\ 3,\ 4;\ k = 1,\ 2), \\ a_{i4}^{(1)} &= \sum_{k=1}^{3} a'_{ik}a_{1k} \qquad (i = 1,\ 2,\ 3,\ 4). \end{aligned}\right\} \tag{26}$$

Следующим шагом будет являться умножение матрицы (25) на матрицу C_2^{-1}, равную

$$C_2^{-1} = \begin{pmatrix} -a_{24}^{(1)}/a_{14}^{(1)} & 1 & 0 & 0 \\ -a_{34}^{(1)}/a_{14}^{(1)} & 0 & 1 & 0 \\ -a_{44}^{(1)}/a_{14}^{(1)} & 0 & 0 & 1 \\ 1/a_{14}^{(1)} & 0 & 0 & 0 \end{pmatrix} = \left(\begin{array}{c|c} -\dfrac{\tau^{(1)}}{t^{(1)}} & I \\ \hline \dfrac{1}{t^{(1)}} & 0 \end{array}\right) \tag{27}$$

и справа на матрицу

$$C_2 = \begin{pmatrix} 0 & 0 & 0 & a_{14}^{(1)} \\ 1 & 0 & 0 & a_{24}^{(1)} \\ 0 & 1 & 0 & a_{34}^{(1)} \\ 0 & 0 & 1 & a_{44}^{(1)} \end{pmatrix} = \left(\begin{array}{c|c} 0 & t^{(1)} \\ \hline I & \tau^{(1)} \end{array}\right). \tag{28}$$

Первое умножение даст нам

$$C_2^{-1}C_1^{-1}AC_1=\begin{pmatrix} a_{11}^{(1)'} & a_{12}^{(1)'} & 0 & 0\\ a_{21}^{(1)'} & a_{22}^{(1)'} & 0 & 0\\ a_{31}^{(1)'} & a_{32}^{(1)'} & 1 & 0\\ a_{41}^{(1)'} & a_{42}^{(1)'} & 0 & 1 \end{pmatrix}, \tag{29}$$

где $a_{ik}^{(1)'}$ находятся по формулам, аналогичным (24), а второе умножение приведет к

$$C_2^{-1}C_1^{-1}AC_1C_2=\begin{pmatrix} a_{11}^{(2)} & 0 & 0 & a_{14}^{(2)}\\ a_{21}^{(2)} & 0 & 0 & a_{24}^{(2)}\\ a_{31}^{(2)} & 1 & 0 & a_{34}^{(2)}\\ a_{41}^{(2)} & 0 & 1 & a_{44}^{(2)} \end{pmatrix}, \tag{30}$$

где $a_{ik}^{(2)}$ находятся по формулам, аналогичным (26). Если еще раз повторить этот процесс и использовать матрицы

$$C_3=\begin{pmatrix} 0 & 0 & 0 & a_{14}^{(2)}\\ 1 & 0 & 0 & a_{24}^{(2)}\\ 0 & 1 & 0 & a_{34}^{(2)}\\ 0 & 0 & 1 & a_{44}^{(2)} \end{pmatrix}, \quad C_3^{-1}=\begin{pmatrix} -a_{24}^{(2)}/a_{14}^{(2)} & 1 & 0 & 0\\ -a_{34}^{(2)}/a_{14}^{(2)} & 0 & 1 & 0\\ -a_{44}^{(2)}/a_{14}^{(2)} & 0 & 0 & 1\\ 1/a_{14}^{(2)} & 0 & 0 & 0 \end{pmatrix}, \tag{31}$$

то мы придем к

$$C_3^{-1}C_2^{-1}C_1^{-1}AC_1C_2C_3=\begin{pmatrix} 0 & 0 & 0 & a_{14}^{(3)}\\ 1 & 0 & 0 & a_{24}^{(3)}\\ 0 & 1 & 0 & a_{34}^{(3)}\\ 0 & 0 & 1 & a_{44}^{(3)} \end{pmatrix}. \tag{32}$$

В общем случае матрицы n-го порядка мы после $n-1$ шагов придем к матрице

$$\begin{pmatrix} 0 & 0 & 0 & \dots & 0 & a_{1n}^{(n-1)}\\ 1 & 0 & 0 & \dots & 0 & a_{2n}^{(n-1)}\\ 0 & 1 & 0 & \dots & 0 & a_{3n}^{(n-1)}\\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\ 0 & 0 & 0 & \dots & 1 & a_{nn}^{(n-1)} \end{pmatrix}. \tag{33}$$

Характеристический многочлен матрицы (33) равен

$$\begin{vmatrix} -\lambda & 0 & 0 \ldots 0 & a_{1n}^{(n-1)} \\ 1 & -\lambda & 0 \ldots 0 & a_{2n}^{(n-1)} \\ 0 & 1 & -\lambda \ldots 0 & a_{3n}^{(n-1)} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 \ldots 1 & a_{nn}^{(n-1)} - \lambda \end{vmatrix} = (-1)^n \left[\lambda^n - \sum_{i=0}^{n-1} a_{i+1,n}^{(n-1)} \lambda^i \right]. \quad (34)$$

Заметим, что в качестве C_1 можно взять произвольную матрицу вида

$$\begin{pmatrix} 0 & t \\ \cdots & \cdots \\ I & \tau \end{pmatrix} \quad (35)$$

с $t \neq 0$. Это вызвало бы лишь добавление одного шага. В то же время удачным выбором последнего столбца (35) возможно уменьшить вычислительную погрешность. Можно показать, что выгодно брать в качестве последнего столбца (35) компоненты вектора, близкого к собственному вектору A, соответствующему наименьшему по модулю собственному значению. Описанное нами видоизменение метода Данилевского соответствует выбору последнего столбца (35) в виде

$$\begin{pmatrix} 1 \\ 0 \\ 0 \\ \cdot \\ \cdot \\ 0 \end{pmatrix}. \quad (36)$$

Обозначим через $\bar{t}_0$ вектор, компоненты которого равны элементам последнего столбца матрицы (35), и через S_n — матрицу $C_1 C_2 \ldots C_n$. Тогда если $A^{(n)} = C_n^{-1} C_{n-1}^{-1} \ldots C_1^{-1} A C_1 \ldots C_{n-1} C_n$, то имеем:

$$AS_n = S_n A^{(n)}. \quad (37)$$

Так как $A^{(n)}$ имеет вид (33), то из (37) следует, что каждый столбец S_n получается из предыдущего путем умножения его на матрицу A. Первый столбец S_n состоит из компонент вектора $\bar{t}_0$. Следовательно, столбцы матрицы S_n состоят соответственно из компонент векторов

$$\bar{t}_0, \ A\bar{t}_0, \ A^2\bar{t}_0, \ \ldots, \ A^{n-1}\bar{t}_0. \quad (38)$$

Этим можно воспользоваться для отыскания собственных векторов матрицы A, как это было сделано в методе Крылова.

Как в исходном методе Данилевского, так и в данном нами видоизменении можно увеличить точность, выбирая в качестве элемента t, на который производится деление, наибольший по модулю

элемент соответствующей строки или столбца. Так, в видоизмененном методе Данилевского можно взять

$$C_i = \left(\begin{array}{c:c:c} I_1 & 0 & \tau_1 \\ \hdashline 0 & 0 & t \\ \hdashline 0 & I_2 & \tau_2 \end{array}\right), \quad C_i^{-1} = \left(\begin{array}{c:c:c} I_1 & -\frac{\tau_1}{t} & 0 \\ \hdashline 0 & -\frac{\tau_2}{t} & I_2 \\ \hdashline 0 & \frac{1}{t} & 0 \end{array}\right), \tag{39}$$

где последний столбец C_i совпадает с последним столбцом преобразуемой на данном этапе матрицы, а t — наибольший по модулю элемент этого столбца. I_1 и I_2 — единичные матрицы. Порядок I_2 должен быть больше, чем число столбцов преобразуемой матрицы, уже приведенных к нормальной форме Фробениуса.

В видоизмененном методе Данилевского, так же как и в исходном, процесс может привести к матрице вида

$$\left(\begin{array}{c:c} A_1 & 0 \\ \hdashline A_2 & F \end{array}\right), \tag{40}$$

где F — матрица, имеющая нормальную форму Фробениуса. Это не вызовет никаких дополнительных затруднений, так как характеристический многочлен F является делителем характеристического многочлена матрицы A, а остальную часть последнего характеристического многочлена можно получить, продолжая преобразования матрицы A_1.

Возможен следующий контроль метода Данилевского. Вместо A будем преобразовывать матрицу

$$\left(\begin{array}{c:c} 1 & b \\ \hdashline 0 & A \end{array}\right), \tag{41}$$

где строка b выбрана так, чтобы сумма элементов каждого столбца матрицы (41) была равна 1. Тогда если выбрать $\bar{t}_0$ так, что сумма его компонент также равна 1, то сумма элементов каждого столбца всех преобразованных матриц будет равна 1.

Видоизмененный способ Данилевского несколько удобнее для вычисления на автоматических вычислительных машинах.

§ 5. Обзор других способов получения характеристического многочлена

В настоящее время известно большое число других способов получения характеристического многочлена. Нет никакой возможности изложить их подробно в нашей книге. Поэтому мы в настоящем параграфе ограничимся лишь кратким обзором некоторых методов, не изложенных ранее.

1. Метод Леверрье. Обозначим корни характеристического многочлена через $\lambda_1, \lambda_2, \ldots, \lambda_n$. Тогда если

$$|A-\lambda I| = (-1)^n[\lambda^n + p_1\lambda^{n-1} + \ldots + p_{n-1}\lambda + p_n], \tag{1}$$

то

$$p_k = (-1)^k \sum_{1 \leqslant i_1 < i_2 < \ldots < i_k \leqslant n} \lambda_{i_1}\lambda_{i_2} \ldots \lambda_{i_k}. \tag{2}$$

Выражения, стоящие в правой части равенства (2), являются симметрическими функциями корней характеристического уравнения. Рассмотрим еще следующие симметрические функции корней:

$$S_k = \lambda_1^k + \lambda_2^k + \ldots + \lambda_n^k. \tag{3}$$

Из курса высшей алгебры известно, что S_k и p_k связаны соотношениями (формулы Ньютона)

$$S_k + S_{k-1}p_1 + S_{k-2}p_2 + \ldots + S_1p_{k-1} + kp_k = 0 \qquad (k \leqslant n). \tag{4}$$

Отсюда получаем:

$$\left.\begin{aligned} &p_1 = -S_1, \\ &p_2 = -\frac{1}{2}(S_1p_1 + S_2), \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &p_k = -\frac{1}{k}(S_1p_{k-1} + S_2p_{k-2} + \ldots + S_{k-1}p_1 + S_k), \end{aligned}\right\} \tag{5}$$

и мы можем найти все p_k, если будут известны S_k. Но S_k равняется следу матрицы A^k. Поэтому, вычислив A^k $(k = 1, 2, \ldots, n)$ и найдя их следы, мы можем найти характеристический многочлен. Процесс довольно трудоемкий, так как приходится производить большое количество умножений матриц. Д. К. Фаддеев предложил усовершенствование метода Леверрье. Оно состоит в следующем.

Рассмотрим матрицу $C(\lambda)$, присоединенную для матрицы $A - \lambda I$. При этом

$$C(\lambda)(A - \lambda I) = D(\lambda)I, \tag{6}$$

где $D(\lambda)$ — характеристический многочлен матрицы A (1). Матрицу $C(\lambda)$ можно записать в виде

$$C(\lambda) = C_0\lambda^{n-1} + C_1\lambda^{n-2} + \ldots + C_{n-2}\lambda + C_{n-1}, \tag{7}$$

где C_i — квадратные матрицы порядка n, не зависящие от λ. Сравнивая коэффициенты при одинаковых степенях λ в правой и левой

частях (6) и учитывая (7), получим:

$$\left.\begin{aligned} -C_0 &= (-1)^n I, \\ C_1 - C_0 A &= (-1)^{n-1} p_1 I, \\ C_2 - C_1 A &= (-1)^{n-1} p_2 I, \\ \cdots\cdots\cdots\cdots\cdots\cdots & \\ C_{n-1} - C_{n-2} A &= (-1)^{n-1} p_{n-1} I, \\ C_{n-1} A &= (-1)^n p_n I. \end{aligned}\right\} \tag{8}$$

Отсюда получим:

$$C_0 = (-1)^{n-1} I, \tag{9}$$

$$C_1 = C_0 A + (-1)^{n-1} p_1 I = (-1)^{n-1} [A + p_1 I]. \tag{10}$$

Так как p_1 определяется следом матрицы A и, следовательно, известно, то мы сможем найти C_1. Умножая равенство (10) на A и беря след от обеих частей равенства (мы будем обозначать след матрицы B через $\mathrm{Sp}(B)$), найдем в силу второго из равенств (5):

$$\mathrm{Sp}(C_1 A) = (-1)^{n-1} [\mathrm{Sp}(A^2) + p_1 \mathrm{Sp}(A)] =$$

$$= (-1)^{n-1} [S_2 + p_1 S_1] = (-1)^n 2p_2. \tag{11}$$

Это нам позволит найти p_2. Затем мы находим C_2 при помощи третьего из равенств (8) и умножая C_2 на A, найдем:

$$\mathrm{Sp}(C_2 A) = \mathrm{Sp}(C_1 A^2) + (-1)^{n-1} p_2 \mathrm{Sp}(A) = (-1)^{n-1} [\mathrm{Sp}(A^3) +$$

$$+ p_1 \mathrm{Sp}(A^2) + p_2 \mathrm{Sp}(A)] = (-1)^{n-1} [S_3 + p_1 S_2 + p_2 S_1] = (-1)^n 3p_3. \tag{12}$$

Продолжая эти рассуждения, мы придем в конце концов к равенству

$$\mathrm{Sp}(C_{n-1} A) = (-1)^n n p_n. \tag{13}$$

Последнее из уравнений (8) будет служить для контроля правильности вычислений.

В процессе вычислений мы найдем определитель матрицы A, равный $(-1)^n p_n$, присоединенную к A матрицу, равную $(-1)^n C_{n-1}$, а следовательно и обратную матрицу A^{-1}. Если λ_i — корень характеристического многочлена A и $C(\lambda_i) \neq 0$, то столбцы $C(\lambda_i)$ являются собственными векторами A, так как

$$(A - \lambda_i I) C(\lambda_i) = D(\lambda_i) I = 0. \tag{14}$$

Это наверняка произойдет, если λ_i — простой корень $D(\lambda)$. В случае, если λ_i — кратный корень $D(\lambda)$, для получения собственных векторов может потребоваться переход от $C(\lambda)$ к производным ее по λ.

Метод Фаддеева также требует большого числа операций, но зато он дает возможность кроме характеристического многочлена находить еще ряд величин.

2. Метод окаймления. Если записать матрицу A в виде

$$A = A_n = \begin{pmatrix} A_{n-1} & b_{n-1} \\ C_{n-1} & d_{n-1} \end{pmatrix}, \tag{15}$$

где A_{n-1} — квадратная матрица, состоящая из элементов первых $n-1$ строк и столбцов A, то матрица $C(\lambda)$, о которой говорилось ранее, может быть представлена так:

$$C(\lambda) = C_n(\lambda) = \begin{pmatrix} f_{n-1}(\lambda) & g_{n-1}(\lambda) \\ h_{n-1}(\lambda) & D_{n-1}(\lambda) \end{pmatrix}, \tag{16}$$

где многочлен $D_{n-1}(\lambda)$ является характеристическим для A_{n-1} и разбиение (16) на клетки соответствует разбиению (15). В силу равенства

$$(A_n - \lambda I_n) C(\lambda) = D(\lambda) I \tag{17}$$

будем иметь:

$$\left.\begin{aligned} (A_{n-1} - \lambda I_{n-1}) g_{n-1}(\lambda) + b_{n-1} D_{n-1}(\lambda) = 0, \\ C_{n-1} g_{n-1}(\lambda) + (d_{n-1} - \lambda) D_{n-1}(\lambda) = D(\lambda). \end{aligned}\right\} \tag{18}$$

Таким образом, если известен многочлен $D_{n-1}(\lambda)$, первое из равенств (18) даст нам возможность найти $g_{n-1}(\lambda)$, а второе из равенств даст возможность найти $D(\lambda)$. Этими рассуждениями можно воспользоваться для отыскания характеристического многочлена $D(\lambda)$, если, начиная с $D_2(\lambda)$, последовательно находить все $D_i(\lambda)$.

3. Эскалаторный метод. Приведем еще один метод, позволяющий использовать собственные значения и собственные векторы матрицы A_{n-1}, для получения собственных значений и собственных векторов матрицы (15). Чтобы избежать разбора возможных исключительных случаев и не слишком усложнять изложение, предположим, что матрица A_{n-1} симметрическая и все ее собственные значения различны. Обозначим собственные значения A_{n-1} через λ_i:

$$\lambda_1 < \lambda_2 < \lambda_3 < \ldots < \lambda_{n-1},$$

и соответствующие им ортонормированные собственные векторы — через

$$\bar{x}_i = (x_{1i}, x_{2i}, \ldots, x_{n-1, i}) \qquad (i = 1, 2, \ldots, n-1).$$

При этом

$$A_{n-1}X = X\Lambda, \tag{19}$$

где

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1,\,n-1} \\ x_{21} & x_{22} & \dots & x_{2,\,n-1} \\ \dots & \dots & \dots & \dots \\ x_{n-1,\,1} & x_{n-1,\,2} & \dots & x_{n-1,\,n-1} \end{pmatrix}, \qquad \Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \lambda_{n-1} \end{pmatrix}. \tag{20}$$

Будем предполагать, что в (15) C_{n-1} получено транспонированием b_{n-1}. Собственный вектор (15) ищем в виде

$$y = \begin{pmatrix} Xz \\ t \end{pmatrix}, \tag{21}$$

где z — некоторый $(n-1)$-мерный вектор-столбец, а t — число. Из равенства

$$Ay = \lambda y \tag{22}$$

следует:

$$\left.\begin{aligned} A_{n-1}Xz + tb_{n-1} &= \lambda Xz, \\ C_{n-1}Xz + d_{n-1}t &= \lambda t. \end{aligned}\right\} \tag{23}$$

Первое равенство (23) можно записать в виде

$$X\Lambda z + tb_{n-1} = \lambda Xz. \tag{24}$$

Так как

$$XX' = I, \tag{25}$$

то из равенства (24) следует:

$$\Lambda z + X'b_{n-1}t = \lambda z \tag{26}$$

и

$$z = (\lambda I - \Lambda)^{-1} X'b_{n-1}t. \tag{27}$$

Собственный вектор A определяется с точностью до постоянного множителя. Поэтому мы можем выбрать t произвольным числом. Следовательно, (27) дает возможность найти z, если известно λ.

Значение λ можно найти, воспользовавшись вторым из равенств (23). Подставляя туда вместо z его значение по (27), получим:

$$C_{n-1}X(\lambda I - \Lambda)^{-1} X'b_{n-1} = \lambda - d_{n-1} \tag{28}$$

или

$$\sum_{i=1}^{n-1} \frac{(\bar{c}_{n-1},\ \bar{x}_i)^2}{\lambda - \lambda_i} = \lambda - d_{n-1}. \tag{29}$$

Формула (29) показывает, что имеется ровно n собственных значений A. Эти собственные значения расположены следующим образом: одно из них меньше λ_1, $n-2$ расположены между λ_i и λ_{i+1} и одно

больше чем λ_{n-1}. Так как собственные значения A разделены собственными значениями A_{n-1}, то последующее их вычисление по тем или иным формулам предыдущей главы не вызывает затруднений.

Такой метод получения собственных значений матрицы A и ее собственных векторов называют *эскалаторным*.

4. Метод Самуэльсона. Запишем матрицу A в виде

$$A=\left(\begin{array}{c|ccc} a_{11} & a_{12} & \dots & a_{1n} \\ \hline a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{array}\right)=\begin{pmatrix} a_{11} & R \\ S & M \end{pmatrix} \tag{30}$$

и пусть характеристические многочлены матриц A и M имеют вид

$$f(\lambda)=(-1)^n\left[\lambda^n+p_1\lambda^{n-1}+\dots+p_{n-1}\lambda+p_n\right], \tag{31}$$

$$\varphi(\lambda)=(-1)^{n-1}\left[\lambda^{n-1}+q_1\lambda^{n-2}+\dots+q_{n-2}\lambda+q_{n-1}\right]. \tag{32}$$

Между коэффициентами p_i и q_i имеют место следующие соотношения:

$$\left.\begin{aligned} p_1&=-a_{11}+q_1, \\ p_2&=-RS-q_1a_{11}+q_2, \\ p_3&=-RMS-q_1RS-q_2a_{11}+q_3, \\ &\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ p_{n-1}&=-RM^{n-3}S-q_1RM^{n-4}S-\dots-q_{n-2}a_{11}+q_{n-1}, \\ p_n&=-RM^{n-2}S-q_1RM^{n-3}S-\dots-q_{n-2}RS-q_{n-1}a_{11}. \end{aligned}\right\} \tag{33}$$

Эти соотношения можно получить, например, следующим образом. Запишем матрицу $C(\lambda)$, присоединенную к $A-\lambda I_n$[1]), в виде

$$C(\lambda)=\begin{pmatrix} \varphi(\lambda) & g_{n-1}(\lambda) \\ h_{n-1}(\lambda) & F_{n-1}(\lambda) \end{pmatrix}. \tag{34}$$

Условие

$$C(\lambda)\begin{pmatrix} a_{11}-\lambda & R \\ S & M-\lambda I_{n-1} \end{pmatrix}=f(\lambda)\,I_n \tag{35}$$

даст

$$\left.\begin{aligned} \varphi(\lambda)(a_{11}-\lambda)+g_{n-1}(\lambda)S&=f(\lambda), \\ \varphi(\lambda)R+g_{n-1}(\lambda)(M-\lambda I_{n-1})&=0. \end{aligned}\right\} \tag{36}$$

Отсюда

$$g_{n-1}(\lambda)=-\varphi(\lambda)R(M-\lambda I_{n-1})^{-1} \tag{37}$$

и

$$\varphi(\lambda)(a_{11}-\lambda)-\varphi(\lambda)R(M-\lambda I_{n-1})^{-1}S=f(\lambda). \tag{38}$$

Из (38) и следуют (33).

---

[1]) I_n — единичная матрица порядка n.

Применим теперь для отыскания коэффициентов q_i метод Крылова, взяв за начальный вектор R'. Получим систему уравнений

$$M'^{n-1}R' + q_1 M'^{n-2}R' + \ldots + q_{n-1}R' = 0. \tag{39}$$

Так как коэффициенты p_i являются линейными комбинациями q_i, то их можно определить (не находя q_i) по схеме Гаусса без обратного хода (см. § 2 главы 6). Это и осуществляет метод Самуэльсона.

5. Интерполяционный метод. Для интерполяционного метода специальный вид определителя, дающего характеристический многочлен, не имеет значения. Поэтому мы будем рассматривать произвольный определитель, элементы которого являются многочленами от λ:

$$f(\lambda = \begin{vmatrix} f_{11}(\lambda) & f_{12}(\lambda) & \ldots & f_{1n}(\lambda) \\ f_{21}(\lambda) & f_{22}(\lambda) & \ldots & f_{2n}(\lambda) \\ \ldots & \ldots & \ldots & \ldots \\ f_{n1}(\lambda) & f_{n2}(\lambda) & \ldots & f_{nn}(\lambda) \end{vmatrix}. \tag{40}$$

Пусть степень многочлена $f(\lambda)$ равна m. Вычислим определитель (40) при каких-либо $m+1$ различных значениях λ_i и построим соответствующий интерполяционный многочлен. Этот интерполяционный многочлен будет совпадать с $f(\lambda)$. Теоретически метод совершенно прост. Практически он может потребовать выполнения большого числа операций. Так как вопросы интерполирования разобраны нами очень подробно, то в детали входить не будем.

На этом мы и закончим рассмотрение способов приведения характеристического определителя к многочленному виду.

§ 6. Определение границ собственных значений

Для того чтобы решить полученное тем или иным способом характеристическое уравнение, желательно иметь представление о расположении его корней. В предыдущей главе мы уже рассматривали подобные вопросы. Однако характеристический многочлен тесно связан с породившей его матрицей A, и поэтому можно указать ряд методов, более приспособленных к рассматриваемому случаю. Кроме того, часто возникает необходимость знать границы собственных значений и совершенно не требуется характеристический многочлен. В связи с этим в данном параграфе будут рассмотрены методы определения границ собственных значений матрицы, не использующие ее характеристического многочлена в явном виде. Эти методы будут пригодны и для получения границ корней произвольного многочлена, если записать последний в виде характеристического многочлена некоторой матрицы. Для этого можно использовать, например, нормальную форму Фробениуса.

1. Случай симметрической матрицы. Рассмотрим уравнение

$$A\bar{x} = \lambda\bar{x}, \tag{1}$$

где A — действительная симметрическая матрица, $\bar{x}$ — вектор-столбец и λ — действительное или комплексное число. Известно, что в этом случае имеется n действительных собственных значений

$$\lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_n \tag{2}$$

и n соответствующих им собственных векторов $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n$, которые можно считать действительными и ортонормированными. Тогда, если $\bar{x} \neq 0$ — произвольный вектор-столбец:

$$\bar{x} = c_1\bar{x}_1 + c_2\bar{x}_2 + \ldots + c_n\bar{x}_n, \tag{3}$$

то

$$(A\bar{x},\ \bar{x}) = \lambda_1 c_1^2 + \lambda_2 c_2^2 + \ldots + \lambda_n c_n^2. \tag{4}$$

Таким образом,

$$\lambda_1 (c_1^2 + c_2^2 + \ldots + c_n^2) = \lambda_1 (\bar{x},\ \bar{x}) \leqslant (A\bar{x},\ \bar{x}) \leqslant$$
$$\leqslant \lambda_n (\bar{x},\ \bar{x}) = \lambda_n (c_1^2 + c_2^2 + \ldots + c_n^2), \tag{5}$$

или

$$\lambda_1 \leqslant \frac{(A\bar{x},\ \bar{x})}{(\bar{x},\ \bar{x})} \leqslant \lambda_n. \tag{6}$$

Эти неравенства, которые иногда называют *принципом Релея*, дают некоторое представление о собственных значениях. Если взять в неравенстве (6) в качестве $\bar{x}$ собственные векторы, соответствующие собственным значениям λ_1 и λ_n, то будут выполнены знаки равенства. Таким образом, наибольшее собственное значение A равно

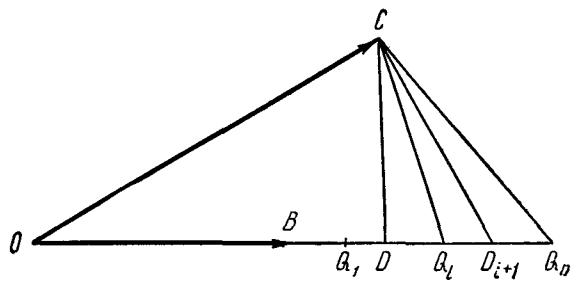

Рис. 12.

верхней границе, а наименьшее собственное значение A — нижней границе отношений, стоящих в (6).

Дадим принципу Релея геометрическую интерпретацию. Пусть отрезок OB изображает нормированный вектор $\bar{x}$ и OC — вектор $A\bar{x}$. На прямой OB отложим точки Q_i так, что $OQ_i = \lambda_i$ (рис. 12).

Имеем:

$$(\overrightarrow{CQ_i},\ \overrightarrow{CQ_{i+1}}) = CQ_i \cdot CQ_{i+1}\cos\alpha = (\lambda_i\bar{x} - A\bar{x},\ \lambda_{i+1}\bar{x} - A\bar{x}) =$$

$$= \sum_{k=1}^{n} (\lambda_i - \lambda_k)(\lambda_{i+1} - \lambda_k)\, c_k^2 \geqslant 0. \quad (7)$$

Отсюда

$$\alpha \leqslant \frac{\pi}{2}. \quad (8)$$

Таким образом, угол между двумя векторами $\overrightarrow{CQ_i}$ и $\overrightarrow{CQ_{i+1}}$ не может превышать $\frac{\pi}{2}$. Проведем перпендикуляр CD и заметим, что

$$(A\bar{x},\ \bar{x}) = OD. \quad (9)$$

Тогда из (6) следует, что на отрезках $(-\infty, D)$ и (D, ∞) также имеется по крайней мере одно собственное значение. Объединяя (8) и последнее замечание, мы приходим к выводу, что если провести

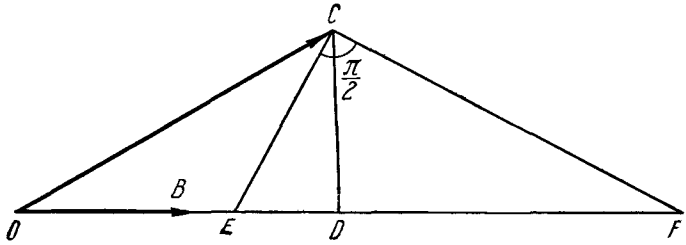

Рис. 13.

через точку C произвольные прямые CE и CF, пересекающие прямую OB соответственно в точках E и F, и если угол $\angle ECF \geqslant \frac{\pi}{2}$, то на отрезке EF имеется по крайней мере одно собственное значение. Это и есть геометрическая интерпретация принципа Релея.

Дадим некоторые приложения этой геометрической интерпретации. Пусть опять $\bar{x}$ — произвольный вектор, $(\bar{x},\ \bar{x}) = 1$. Обозначим

$$\alpha = (A\bar{x},\ \bar{x}), \quad \beta = (A\bar{x} - \alpha\bar{x},\ A\bar{x} - \alpha\bar{x}). \quad (10)$$

Докажем, что если $a < \alpha$ и

$$b = \alpha + \frac{\beta}{\alpha - a}, \quad (11)$$

то на отрезке $[a, b]$ имеется по крайней мере одно собственное значение A. Действительно, $\alpha = OD$, $\beta = (CD)^2$. Пусть $OE = a < OD$ и $\angle ECF = \frac{\pi}{2}$. Тогда (рис. 13)

$$\left.\begin{aligned} ED \cdot DF &= (CD)^2 = \beta, \\ OF &= OD + DF = \alpha + \frac{\beta}{\alpha - a} = b. \end{aligned}\right\} \quad (12)$$

Отсюда и следует утверждение.

Докажем еще, что если a — произвольное действительное число, $\bar{x}$ — произвольный нормированный вектор, $\alpha = (A\bar{x}, \bar{x})$, $\gamma = (A^2\bar{x}, \bar{x})$, то на отрезке

$$\left[a - (\gamma - 2a\alpha + a^2)^{\frac{1}{2}},\ a + (\gamma - 2a\alpha + a^2)^{\frac{1}{2}}\right] \tag{13}$$

найдется по крайней мере одно собственное значение A. Рассмотрим геометрическую картину (рис. 14).

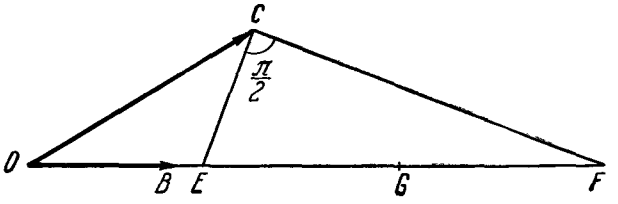

Рис. 14.

Отрезок OG берем равным a. Точки E и F строим так, что

$$CG = EG = FG. \tag{14}$$

При этом

$$EG = FG = (a^2 + \gamma - 2\alpha a)^{\frac{1}{2}}. \tag{15}$$

Так как угол $\angle ECF$ прямой, то утверждение доказано.

Дадим небольшое обобщение принципа Релея. Возьмем два многочлена

$$\left.\begin{aligned} \varphi(\lambda) &= a_0 + a_1\lambda + \ldots + a_k\lambda^k, \\ \psi(\lambda) &= b_0 + b_1\lambda + \ldots + b_l\lambda^l \end{aligned}\right\} \tag{16}$$

и рассмотрим

$$f(\lambda) = \frac{\varphi(\lambda)}{\psi(\lambda)}. \tag{17}$$

Как известно, собственные значения $\bar{\lambda}_i$: $\bar{\lambda}_1 \leqslant \bar{\lambda}_2 \leqslant \ldots \leqslant \bar{\lambda}_n$, матрицы $f(A)$ равны $f(\lambda_i)$, $i = 1, 2, \ldots, n$. Как и ранее, для произвольного $\bar{x} \neq 0$ будем иметь:

$$\bar{\lambda}_1 \leqslant \frac{(\bar{x}, f(A)\bar{x})}{(\bar{x}, \bar{x})} \leqslant \bar{\lambda}_n. \tag{18}$$

Вследствие этого, построив графики функций $y = f(x)$ и $y = \dfrac{(\bar{x}, f(A)\bar{x})}{(\bar{x}, \bar{x})}$, мы сможем судить о собственных значениях A. Так, например, если

расположение этих графиков имеет вид, показанный на рис. 15, то мы можем утверждать, что по крайней мере одно собственное значение находится на отрезке $[a, b]$ и по крайней мере одно вне этого отрезка.

Вычисление матрицы $[\psi(A)]^{-1}$, входящей в определение $f(A)$, связано с большими затруднениями. Можно несколько видоизменять (18) с целью упрощения вычислений. Предположим сначала, что все

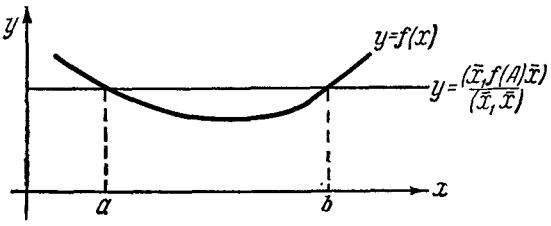

Рис. 15.

собственные значения $\psi(A)$ положительны. Тогда существует матрица, обозначаемая $[\psi(A)]^{\frac{1}{2}}$, такая, что

$$[\psi(A)]^{\frac{1}{2}}[\psi(A)]^{\frac{1}{2}} = \psi(A). \tag{19}$$

Обозначим

$$[\psi(A)]^{-\frac{1}{2}}\bar{x} = \bar{y}. \tag{20}$$

При этом

$$\frac{(\bar{x}, f(A)\bar{x})}{(\bar{x}, \bar{x})} = \frac{(\bar{x}, [\psi(A)]^{-1}\varphi(A)\bar{x})}{(\bar{x}, \bar{x})} = \frac{\left([\psi(A)]^{-\frac{1}{2}}\bar{x}, \varphi(A)[\psi(A)]^{-\frac{1}{2}}\bar{x}\right)}{(\bar{x}, \bar{x})} =$$

$$= \frac{(\bar{y}, \varphi(A)\bar{y})}{(\bar{y}, \psi(A)\bar{y})}, \tag{21}$$

и мы можем вместо (18) рассматривать правую часть (21) при произвольном ненулевом векторе $\bar{y}$. Если условие о положительности собственных значений $\psi(A)$ не выполнено, то можно взять

$$f(\lambda) = \frac{\psi(\lambda)\varphi(\lambda)}{[\psi(\lambda)]^2}. \tag{22}$$

Тогда правая часть (21) перейдет в

$$\frac{(\psi(A)\bar{y}, \varphi(A)\bar{y})}{(\psi(A)\bar{y}, \psi(A)\bar{y})}. \tag{23}$$

Рассмотрим теперь некоторые частные случаи $f(\lambda)$. Пусть $f(\lambda)=\frac{1}{\lambda-\alpha}$, где α — произвольное действительное число. В этом случае можно утверждать, что в последовательности

$$\frac{1}{\lambda_1-\alpha},\ \frac{1}{\lambda_2-\alpha},\ \ldots,\ \frac{1}{\lambda_n-\alpha} \tag{24}$$

имеются числа большие и меньшие

$$\xi=\frac{(\bar{x},\ (A-\alpha I)\,\bar{x})}{((A-\alpha I)\,\bar{x},\ (A-\alpha I)\,\bar{x})} \tag{25}$$

(рис. 16).

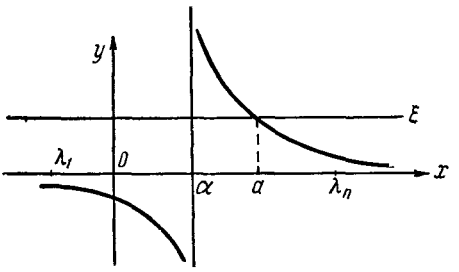

Рис. 16.

Возьмем теперь $f(\lambda)=(\lambda-\alpha)^2$. В этом случае

$$\xi=\frac{(\bar{x},\ (A-\alpha I)^2\,\bar{x})}{(\bar{x},\ \bar{x})}=\frac{((A-\alpha I)\,\bar{x},\ (A-\alpha I)\,\bar{x})}{(\bar{x},\ \bar{x})} \tag{26}$$

и найдутся две точки пересечения $y=f(x)$ и $y=\xi$ (рис. 17). Если обозначить абсциссы точек пересечения через a и b, то можно утверждать, что имеется по крайней мере одно собственное значение на отрезке $[a,\ b]$ и по крайней мере одно вне его. Вычисления приводят к отрезку (13). Пусть, далее, $f(\lambda)=\frac{\lambda}{\alpha}+\frac{\alpha}{\lambda}$, $\alpha>0$. Тогда

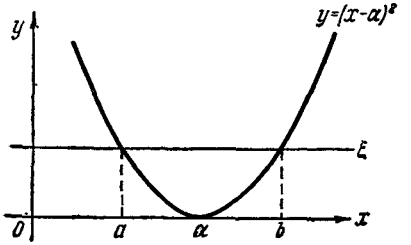

Рис. 17.

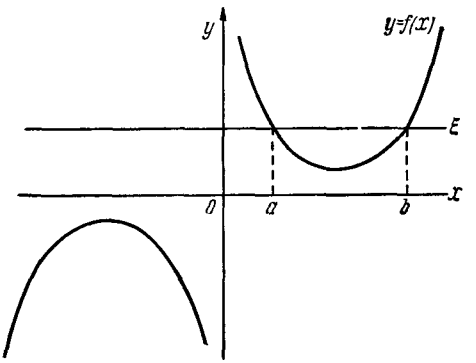

Рис. 18.

$$\xi=\frac{(\bar{x},\ (\alpha^{-1}A+\alpha A^{-1})\,\bar{x})}{(\bar{x},\ \bar{x})}=\frac{(A\bar{y},\ A\bar{y})+\alpha^2(\bar{y},\ \bar{y})}{\alpha\,(\bar{y},\ A\bar{y})},\quad \bar{x}=A^{\frac{1}{2}}\bar{y}. \tag{27}$$

Снова получим две точки пересечения графиков $y=f(x)$ и $y=\xi$ (рис. 18). Если все собственные значения A положительны, то на отрезке $[a,\ b]$ имеется по крайней мере одно собственное значение.

В данном случае a и b определяются при помощи равенств:

$$\left.\begin{aligned} a &= (1+\delta-\sqrt{\delta^2+2\delta})\,\alpha, \\ b &= (1+\delta+\sqrt{\delta^2+2\delta})\,\alpha, \end{aligned}\quad \delta=\frac{\xi}{2}-1.\right\} \tag{28}$$

Возьмем еще один случай. Пусть $f(\lambda)=\left(\frac{\lambda}{\alpha}+\frac{\alpha}{\lambda}\right)^{-1}$. При этом

$$\xi=\frac{\alpha(\bar{x}, A\bar{x})}{\alpha^2(\bar{x}, \bar{x})+(A\bar{x}, A\bar{x})}. \tag{29}$$

Опять будет две точки пересечения графиков $y=f(x)$ и $y=\xi$.

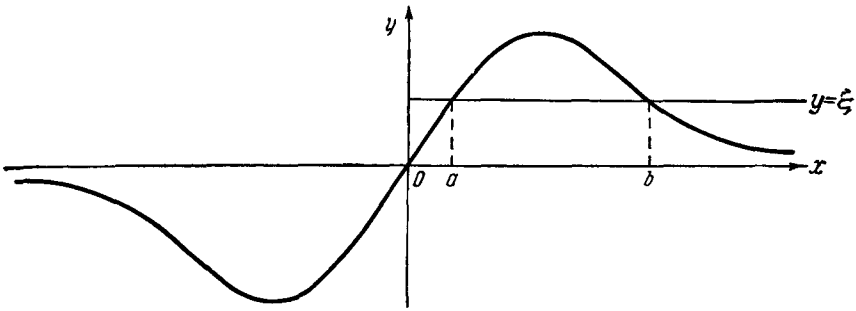

Рис. 19.

Абсциссы этих точек определяются следующим образом:

$$\left.\begin{aligned} a &= (1+\varepsilon-\sqrt{\varepsilon^2+2\varepsilon})\,\alpha, \\ b &= (1+\varepsilon+\sqrt{\varepsilon^2+2\varepsilon})\,\alpha, \end{aligned}\quad \varepsilon=\frac{\operatorname{sign}\xi}{2\xi}-1.\right\} \tag{30}$$

По крайней мере одно собственное значение A лежит на отрезке $[a, b]$ (рис. 19).

Наконец, возьмем $f(\lambda)=\left(\frac{\lambda}{\alpha}+\frac{\alpha}{\lambda}\right)^2$. При этом

$$\xi=\frac{[(A^2+\alpha^2 I)\,\bar{x},\ (A^2+\alpha^2 I)\,\bar{x}]}{\alpha^2(A\bar{x}, A\bar{x})}. \tag{31}$$

По крайней мере одно собственное значение A удовлетворяет неравенствам

$$a \leqslant \left|\frac{\lambda_i}{\alpha}\right| \leqslant b, \tag{32}$$

где

$$\left.\begin{aligned} a &= 1+\delta-\sqrt{\delta^2+2\delta}, \\ b &= 1+\delta+\sqrt{\delta^2+2\delta}, \end{aligned}\quad \xi=4(1+\delta)^2\right\} \tag{33}$$

(рис. 20).

На этом мы закончим рассмотрение различных случаев использования обобщенного принципа Релея.

Покажем на примере, как можно использовать полученные формулы. Возьмем снова матрицу A § 4 (см. (17) § 4). Будем выбирать различные векторы $\bar{x}$ и применять наши формулы. Сначала выберем $\bar{x}=(1, 0, 0, 0, 0, 0)$. При этом $\alpha=(A\bar{x}, \bar{x})=6{,}1818$ и $\gamma=(A\bar{x}, A\bar{x})=38{,}4352$, и если взять $a=\alpha=6{,}1818$, то формула (13) даст нам, что на отрезке [5,7121; 6,6515] имеется по крайней мере одно собственное значение матрицы A. Возьмем, далее, $\bar{x}=(0, 1, 0, 0, 0, 0)$. При этом $\alpha=(A\bar{x}, \bar{x})=7{,}1818$ и $\gamma=(A\bar{x}, A\bar{x})=51{,}8120$. Снова выбрав $a=\alpha=7{,}1818$, по формуле (13) получим, что на отрезке [6,6982; 7,6654] находится по крайней мере одно собственное значение матрицы A. Теперь выберем $\bar{x}=(0, 0, 1, 0, 0, 0)$. Это даст $\alpha=(A\bar{x}, \bar{x})=8{,}2435$, $\gamma=(A\bar{x}, A\bar{x})=68{,}2464$, и если взять $a=\alpha=8{,}2435$, то по формуле (13) найдем отрезок [7,7039; 8,7831]. При $\bar{x}=(0, 0, 0, 1, 0, 0)$ $\alpha=(A\bar{x}, \bar{x})=9{,}3141$, $\gamma=(A\bar{x}, A\bar{x})=87{,}3873$. Возьмем здесь $a=10$. Тогда формула (13) даст отрезок [8,9486; 11,0514]. Наконец, возьмем $\bar{x}=(0, 0, 0, 0, -1, 2)\cdot\frac{1}{\sqrt{5}}$. Это даст $\alpha=(A\bar{x}, \bar{x})=3{,}6475$.

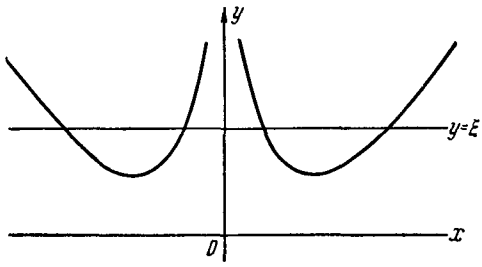

Рис. 20.

На основании принципа Релея заключаем, что имеется собственное значение меньшее найденного α. Нетрудно проверить, что все собственные значения матрицы A положительны. Полученная далее формула (43) показывает, что все собственные значения матрицы A больше 2. Таким образом, нам удалось найти пять непересекающихся отрезков: [2; 3,6475]; [5,7121; 6,6515]; [6,6982; 7,6654]; [7,7039; 8,7831]; [8,9486; 11,0514], в каждом из которых содержится по крайней мере одно собственное значение A. Корнями многочлена (19) § 4, получившегося путем раскрытия векового определителя матрицы A по способу Данилевского, являются числа: 3,5922564; 5,62574641; 6,05652110; 7,28694071; 8,24088053; 9,56175808.

Как мы видим, нам удалось довольно простыми средствами отделить пять из шести корней многочлена. Не все полученные формулы имеют практическое значение. Однако примеры дают возможность уяснить пути использования принципа Релея для отде-

ления собственных значений. Сформулируем теперь некоторое уточнение принципа Релея.

Рассмотрим наряду с симметрической действительной матрицей A еще положительно определенную действительную матрицу B. Будем обозначать собственные значения матрицы B через μ_i:

$$0 < \mu_1 \leqslant \mu_2 \leqslant \ldots \leqslant \mu_n. \tag{34}$$

Собственные значения ν_i, $i = 1, 2, \ldots, n$, матрицы $B^{-1}A$ также действительны и их можно определить при помощи соотношения

$$\nu_i = \min_{c_{ik}} \max_{\bar{x}} \frac{(A\bar{x}, \bar{x})}{(B\bar{x}, \bar{x})}. \tag{35}$$

Здесь ищется минимум при всевозможных значениях c_{ik} максимумов записанного отношения для всевозможных ненулевых векторов $\bar{x} = (x_1, x_2, \ldots, x_n)$, компоненты которых связаны $i - 1$ соотношениями:

$$c_{j1}x_1 + c_{j2}x_2 + \ldots + c_{jn}x_n = 0 \qquad (j = 1, 2, \ldots, i-1). \tag{36}$$

Мы не будем приводить доказательства высказанного утверждения, так как оно довольно громоздко. Желающих ознакомиться с ними мы отсылаем к специальной литературе по теории матриц (см., например, Гантмахер Ф. Р., Теория матриц, Гостехиздат, 1953, глава X, § 7).

Из равенства (35), в частности, следует:

$$\mu_n \nu_i \geqslant \lambda_i \geqslant \mu_1 \nu_i. \tag{37}$$

Неравенствами (37) можно воспользоваться, если известны собственные значения B и $B^{-1}A$. Рассмотрим, например, такой случай. Матрица C получается из матрицы A путем деления строк A последовательно на положительные числа $p_1, p_2, \ldots, p_n$. Этот процесс можно записать как умножение A справа на B^{-1}, где

$$B = \begin{pmatrix} p_1 & 0 & \ldots & 0 \\ 0 & p_2 & \ldots & 0 \\ 0 & 0 & \ldots & p_n \end{pmatrix}. \tag{38}$$

Поэтому, если ν_i — собственные значения C, то

$$\nu_i \max_k (p_k) \geqslant \lambda_i \geqslant \nu_i \min_k (p_k). \tag{39}$$

Дадим теперь ряд оценок максимальных и минимальных собственных значений действительной положительно определенной матрицы A. Обозначим через d определитель A и через S — след матрицы A. Рассмотрим

$$\frac{S - \lambda_i}{n-1} = \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_{i-1} + \lambda_{i+1} + \ldots + \lambda_n}{n-1} \tag{40}$$

и

$$\sqrt[n-1]{\frac{d}{\lambda_i}} = \sqrt[n-1]{\lambda_1\lambda_2 \ldots \lambda_{i-1}\lambda_{i+1} \ldots \lambda_n}. \tag{41}$$

Так как среднее арифметическое положительных чисел не меньше чем их среднее геометрическое, то будем иметь:

$$\frac{d}{\lambda_i} \leqslant \left(\frac{S-\lambda_i}{n-1}\right)^{n-1}. \tag{42}$$

Неравенство можно усилить, если заменить в левой части неравенства λ_i на λ_1, а в правой части λ_i на нуль. Тогда получим:

$$\lambda_1 \geqslant d\left(\frac{n-1}{S}\right)^{n-1}. \tag{43}$$

Неравенство (43) дает оценку минимального собственного значения A снизу. Воспользуемся теперь (42) для того, чтобы получить оценку наибольшего собственного значения сверху.

Преобразовывая (42), получим:

$$\lambda_i \leqslant S - (n-1)\sqrt[n-1]{\frac{d}{\lambda_i}}. \tag{44}$$

Если подставить в правую часть вместо λ_i любое число μ, превышающее собственные значения A, то получим оценку сверху для собственных чисел A:

$$\lambda_n \leqslant S - (n-1)\sqrt[n-1]{\frac{d}{\mu}}. \tag{45}$$

В качестве μ можно взять, например, S. Эту оценку можно уточнить, воспользовавшись рекуррентной формулой

$$\mu_{k+1} = S - (n-1)\sqrt[n-1]{\frac{d}{\mu_k}}. \tag{46}$$

Получим убывающую последовательность чисел

$$S = \mu_1 \geqslant \mu_2 \geqslant \mu_3 \geqslant \ldots, \tag{47}$$

для которой

$$\lim_{k\to\infty} \mu_k = \mu_0 \geqslant \lambda_n. \tag{48}$$

Можно показать, что $\mu_0 = x_0^{n-1}$, где x_0 является наибольшим положительным корнем уравнения

$$x^n - Sx + (n-1)\sqrt[n-1]{d} = 0. \tag{49}$$

Неравенство (43) также можно уточнить. При $x \geqslant y \geqslant 0$ имеем:

$$x^n - y^n \geqslant n(x-y)y^{n-1}. \tag{50}$$

Положим здесь $S=x$, $y=S-\lambda_i$ и применим (42). При этом получим:

$$\lambda_1 \geqslant S - \sqrt[n]{S^n - nd(n-1)^{n-1}}. \tag{51}$$

Эта оценка лучше, чем (43).

Последние неравенства можно использовать не только для положительно определенных матриц. Пусть A произвольная действительная неособая матрица. Матрица $A'A$ будет симметрической и положительно определенной. Обозначим собственные значения $A'A$ через μ_i:

$$0 < \mu_1 \leqslant \mu_2 \leqslant \ldots \leqslant \mu_n. \tag{52}$$

При этом, если λ_i — собственное значение A и $\bar{x}_i$ — соответствующий собственный вектор, то из

$$A\bar{x}_i = \lambda_i \bar{x}_i \tag{53}$$

следует

$$(A\bar{x}_i,\ A\bar{x}_i) = \lambda_i \bar{\lambda}_i (\bar{x}_i,\ \bar{x}_i) \tag{54}$$

или

$$\frac{(\bar{x}_i,\ A'A\bar{x}_i)}{(\bar{x}_i,\ \bar{x}_i)} = |\lambda_i|^2. \tag{55}$$

В силу принципа Релея будем иметь:

$$\mu_1 \leqslant |\lambda_i|^2 \leqslant \mu_n. \tag{56}$$

Таким образом,

$$\left.\begin{aligned} \min_i |\lambda_i| &\geqslant \sqrt{\mu_1} \geqslant |d| \sqrt[n]{\left(\frac{n-1}{S}\right)^{n-1}}, \\ \max_i |\lambda_i| &\leqslant \sqrt{\mu_n} \leqslant \sqrt{S - \sqrt[n]{S^n - nd^2(n-1)^{n-1}}}, \end{aligned}\right\} \tag{57}$$

где d — определитель A и S — след AA', т. е. $S=\sum\limits_{i,\,j=1}^{n} a_{ij}^2$.

Полученные нами для симметрических матриц результаты переносятся и на *эрмитовы матрицы*. Матрица A называется эрмитовой, если $A=\overline{A'}$ (здесь черта над A означает комплексную сопряженность). Все собственные значения эрмитовой матрицы — действительные числа. Она имеет n взаимно ортогональных собственных векторов. Принцип Релея применим к эрмитовым матрицам, только вместо $(\bar{x},\ \bar{x})$ и $(\bar{x},\ A\bar{x})$ нужно брать $(\bar{y},\ \bar{x})$ и $(\bar{y},\ A\bar{x})$, где $\bar{y}$—вектор, компоненты которого комплексно-сопряженные по отношению к компонентам $\bar{x}$ числа. Следовательно, все оценки, основанные на принципе Релея, сохранятся с соответствующими видоизменениями. Так же можно перенести и остальные оценки.

2. Случай несимметрической матрицы. В главе 6 мы доказали, что любая норма матрицы больше модулей ее собственных значений. Этим можно воспользоваться для оценки модулей собственных значений сверху. В частности, мы получим:

$$\left.\begin{aligned} &\max_i \sum_{k=1}^{n} |a_{ik}| \geqslant |\lambda_j|, \\ &\max_k \sum_{i=1}^{n} |a_{ik}| \geqslant |\lambda_j| \end{aligned} \qquad (j = 1, 2, \ldots, n). \right\} \tag{58}$$

Если использовать третью норму матрицы, то получим уже известное неравенство (56). Неравенства (58) можно также записать в виде

$$|\lambda_j| \leqslant \min\left\{ \max_i \sum_{k=1}^{n} |a_{ik}|, \ \max_k \sum_{i=1}^{n} |a_{ik}| \right\}. \tag{59}$$

Рассмотрим наряду с матрицей A еще матрицы

$$B = \frac{1}{2}(A + \bar{A}'), \quad C = \frac{1}{2i}(A - \bar{A}'). \tag{60}$$

Это — эрмитовы матрицы. Пусть $\bar{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jn})$ — нормированный собственный вектор A, соответствующий собственному значению $\lambda_j = \alpha_j + i\beta_j$. Тогда

$$\left.\begin{aligned} &\sum_{r,\,s=1}^{n} a_{rs} x_{js} \bar{x}_{jr} = \alpha_j + i\beta_j, \\ &\sum_{r,\,s=1}^{n} \bar{a}_{rs} x_{js} \bar{x}_{jr} = \alpha_j - i\beta_j. \end{aligned} \right\} \tag{61}$$

Отсюда

$$\left.\begin{aligned} \alpha_j &= \sum_{r,\,s=1}^{n} b_{rs} x_{js} \bar{x}_{jr}, \\ \beta_j &= \sum_{r,\,s=1}^{n} c_{rs} x_{js} \bar{x}_{jr}, \end{aligned} \right\} \tag{62}$$

где b_{rs} и c_{rs} — соответственно элементы B и C. Из (62) получаем:

$$|\alpha_j| \leqslant \sum_{r,\,s=1}^{n} |b_{rs}||x_{js}||\bar{x}_{jr}| \leqslant \frac{1}{2} \sum_{r,\,s=1}^{n} |b_{rs}|(|x_{js}|^2 + |x_{jr}|^2) =$$

$$= \frac{1}{2} \sum_{r=1}^{n} |b_{rs}| + \frac{1}{2} \sum_{s=1}^{n} |b_{rs}|. \tag{63}$$

Таким образом,

$$|\alpha_j| \leqslant \max_s \sum_{r=1}^{n} |b_{rs}| = \max_r \sum_{s=1}^{n} |b_{rs}|. \tag{64}$$

Аналогично найдем:

$$|\beta_j| \leqslant \max_s \sum_{r=1}^{n} |c_{rs}| = \max_r \sum_{s=1}^{n} |c_{rs}|. \tag{65}$$

Мы получили оценки действительной и мнимой частей собственных значений.

Уточним несколько последний результат. Если обозначить собственные значения эрмитовой матрицы B через $\mu_1 \leqslant \mu_2 \leqslant \ldots \leqslant \mu_n$, а собственные значения эрмитовой матрицы C — через $\nu_1 \leqslant \nu_2 \leqslant \ldots \leqslant \nu_n$, то в силу принципа Релея из (62) следует:

$$\left.\begin{aligned} \mu_1 \leqslant \alpha_j \leqslant \mu_n, \\ \nu_i \leqslant \beta_j \leqslant \nu_n. \end{aligned}\right\} \tag{66}$$

Это — более точные оценки, чем (64) и (65), хотя их использование связано с получением собственных значений матриц B и C.

Пусть матрица A действительная. Тогда матрица B будет симметрической и для получения μ_1 и μ_n можно использовать приведенные ранее оценки. Матрица $D = iC = \frac{1}{2}(A - A')$ будет действительной кососимметрической. Такая матрица может иметь только чисто мнимые или нулевые собственные значения. Действительно, если λ — ненулевое собственное значение D и $\bar{x} \neq 0$ — соответствующий собственный вектор, то будем иметь:

$$\lambda(\bar{x}, \bar{x}) = (D\bar{x}, \bar{x}) = (\bar{x}, D'\bar{x}) = \overline{(D'\bar{x}, \bar{x})} = -\bar{\lambda}(\bar{x}, \bar{x}) \tag{67}$$

или

$$\lambda = -\bar{\lambda}. \tag{68}$$

Так как элементы D — действительные числа, то собственные значения будут попарно сопряжены. Обозначим ненулевые собственные значения D через

$$\pm i\mu_1, \pm i\mu_2, \ldots, \pm i\mu_r \qquad (2r \leqslant n). \tag{69}$$

Собственные значения C будут отличаться от собственных значений D только лишь делителем i. Поэтому $|\beta_j| \leqslant \max\limits_k |\mu_k|$. Оценим $\max\limits_k |\mu_k|$. Сумма попарных произведений собственных значений D будет равна

$$\mu_1^2 + \mu_2^2 + \ldots + \mu_r^2, \tag{70}$$

и в то же время она равна коэффициенту при λ^{n-2} в разложении определителя $|\lambda I - D|$ по степеням λ. Последний коэффициент равен сумме всех главных миноров второго порядка матрицы D. (Глав-

ными минорами порядка k матрицы D с элементами d_{ik} называются миноры вида

$$\begin{vmatrix} d_{i_1 i_1} & d_{i_1 i_2} & \dots & d_{i_1 i_k} \\ d_{i_2 i_1} & d_{i_2 i_2} & \dots & d_{i_2 i_k} \\ \dots & \dots & \dots & \dots \\ d_{i_k i_1} & d_{i_k i_2} & \dots & d_{i_k i_k} \end{vmatrix}, \tag{71}$$

где $1 \leqslant i_1 < i_2 < \dots < i_k \leqslant n$.) В нашем случае будем иметь:

$$\mu_1^2 + \mu_2^2 + \dots + \mu_r^2 = \sum_{r<s} \begin{vmatrix} 0 & d_{rs} \\ -d_{rs} & 0 \end{vmatrix} = \sum_{r<s} d_{rs}^2. \tag{72}$$

Обозначим $\max |d_{rs}| = \max \frac{|a_{rs} - a_{sr}|}{2}$ через g. При этом из (72) следует:

$$\max_i |\mu_i|^2 \leqslant \sum_{i=1}^{r} |\mu_i|^2 = \sum_{r<s} d_{rs}^2 \leqslant \frac{n(n-1)}{2} g^2. \tag{73}$$

Отсюда

$$|\beta_j| \leqslant g \sqrt{\frac{n(n-1)}{2}}. \tag{74}$$

Эта оценка менее точна, но осуществляется проще.

В заключение этого параграфа покажем, что если обозначить

$$\sum_{\substack{s=1 \\ (s \neq r)}}^{n} |a_{rs}| = P_r; \qquad \sum_{\substack{r=1 \\ (r \neq s)}}^{n} |a_{rs}| = Q_s, \tag{75}$$

то каждое собственное значение λ матрицы A лежит по крайней мере в одном из кругов

$$|z - a_{rr}| \leqslant P_r \tag{76}$$

и по крайней мере в одном из кругов

$$|z - a_{ss}| \leqslant Q_s. \tag{77}$$

Пусть λ — собственное значение A и $\bar{x} = (x_1, x_2, \dots, x_n)$ — соответствующий собственный вектор. Тогда

$$\sum_{s=1}^{n} a_{rs} x_s = \lambda x_r \qquad (r = 1, 2, \dots, n). \tag{78}$$

Пусть $\max_i |x_i| = |x_r|$. При этом из

$$\lambda x_r - a_{rr} x_r = \sum_{\substack{s=1 \\ (s \neq r)}}^{n} a_{rs} x_s \tag{79}$$

следует:

$$|\lambda - a_{rr}||x_r| \leqslant P_r|x_r|, \tag{80}$$

и так как $|x_r| \neq 0$, то получаем (76). Если провести эти же рассуждения для транспонированной матрицы, имеющей те же самые собственные значения, то получим (77). Утверждение доказано.

Из (76) и (77) снова следует (59). Кроме того, если

$$d = \min_r(|a_{rr}| - P_r) > 0, \tag{81}$$

то

$$|\lambda| \geqslant d.$$

§ 7. Итерационные методы отыскания собственных значений и собственных векторов матриц

1. Отыскание наибольшего по модулю действительного собственного значения матрицы простой структуры. Случай симметрической матрицы. Будем предполагать сначала, что матрица A имеет *простую структуру*. Это значит, что если порядок матрицы A равен n, то имеется n линейно независимых собственных векторов A. Обозначим собственные значения матрицы A через λ_i: $|\lambda_1| \geqslant |\lambda_2| \geqslant \ldots \ldots \geqslant |\lambda_n|$, и соответствующие им линейно независимые собственные векторы — через $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n$.

Произвольный вектор $\bar{v}$ может быть записан в виде

$$\bar{v} = \alpha_1\bar{x}_1 + \alpha_2\bar{x}_2 + \ldots + \alpha_n\bar{x}_n. \tag{1}$$

Образуем последовательность векторов:

$$\left.\begin{aligned} A\bar{v} &= \alpha_1\lambda_1\bar{x}_1 + \alpha_2\lambda_2\bar{x}_2 + \ldots + \alpha_n\lambda_n\bar{x}_n, \\ A^2\bar{v} &= \alpha_1\lambda_1^2\bar{x}_1 + \alpha_2\lambda_2^2\bar{x}_2 + \ldots + \alpha_n\lambda_n^2\bar{x}_n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ A^k\bar{v} &= \alpha_1\lambda_1^k\bar{x}_1 + \alpha_2\lambda_2^k\bar{x}_2 + \ldots + \alpha_n\lambda_n^k\bar{x}_n. \end{aligned}\right\} \tag{2}$$

Вынося в последнем выражении λ_1^k за скобки, получим:

$$A^k\bar{v} = \lambda_1^k\left[\alpha_1\bar{x}_1 + \alpha_2\left(\frac{\lambda_2}{\lambda_1}\right)^k\bar{x}_2 + \ldots + \alpha_n\left(\frac{\lambda_n}{\lambda_1}\right)^k\bar{x}_n\right]. \tag{3}$$

При этом если $\alpha_1 \neq 0$ и $|\lambda_1| > |\lambda_i|$ для $i \geqslant 2$, то

$$A^k\bar{v} = \lambda_1^k\alpha_1\bar{x}_1 + 0\left[\left(\frac{\lambda_2}{\lambda_1}\right)^k\right]. \tag{4}$$

Таким образом, для достаточно больших k вектор $A^k\bar{v}$ будет близок к собственному вектору матрицы A, соответствующему наибольшему по модулю собственному значению. Может случайно оказаться, что $\alpha_1 = 0$, т. е. компонента $\bar{v}$ по направлению вектора $\bar{x}_1$ равна

нулю. Однако это не изменит принципиально наших выводов. В процессе вычислений в связи с ошибками округления мы неизбежно введем соответствующую компоненту. Введенная компонента в конце концов забьет все остальные. Может оказаться лишь, что потребуется значительное число итераций.

Далее, будем обозначать i-ю компоненту некоторого вектора $\bar{u}$ по отношению к произвольным координатным векторам $\bar{e}_1, \bar{e}_2, \ldots, \bar{e}_n$ через $(\bar{u})_i$. Тогда из (2) следует:

$$\frac{(A^{k+1}\bar{v})_i}{(A^k\bar{v})_i} = \lambda_1 \frac{\left(\alpha_1\bar{x}_1 + \alpha_2\left(\frac{\lambda_2}{\lambda_1}\right)^{k+1}\bar{x}_2 + \ldots + \alpha_n\left(\frac{\lambda_n}{\lambda_1}\right)^{k+1}\bar{x}_n\right)_i}{\left(\alpha_1\bar{x}_1 + \alpha_2\left(\frac{\lambda_2}{\lambda_1}\right)^{k}\bar{x}_2 + \ldots + \alpha_n\left(\frac{\lambda_n}{\lambda_1}\right)^{k}\bar{x}_n\right)_i} =$$

$$= \lambda_1 + 0\left[\left(\frac{\lambda_2}{\lambda_1}\right)^{k+1}\right]. \tag{5}$$

Таким образом, при достаточно больших k величина

$$\lambda_1^* = \frac{(A^{k+1}\bar{v})_i}{(A^k\bar{v})_i} \tag{6}$$

будет близка к наибольшему по модулю собственному значению. При практических вычислениях показателем того, что мы достаточно хорошо приблизились к $\bar{x}_1$ и к λ_1, будет постоянство отношений (с требуемой точностью) соответствующих компонент $A^{k+1}\bar{v}$ и $A^k\bar{v}$.

Пусть, в частности, матрица A симметрическая. Тогда векторы $\{\bar{x}_i\}$ можно считать ортонормированными и за приближенное значение собственного вектора $\bar{x}_i$ будем брать

$$\bar{x}_1^* = \frac{A^k\bar{v}}{\|A^k\bar{v}\|}. \tag{7}$$

(Здесь везде берется третья норма главы 6.) Так как $\bar{x}_i$ ортонормированы, то для λ_1 можно получить более точное приближение. Обозначим $(A^k\bar{v}, A^k\bar{v}) = c_k$. В силу (4) будем иметь $c_k = \alpha_1^2\lambda_1^{2k} + \alpha_2^2\lambda_2^{2k} + \ldots + \alpha_n^2\lambda_n^{2k}$ и, следовательно, $c_{k+1}/c_k = \lambda_1^2 + 0\left[\left(\frac{\lambda_2}{\lambda_1}\right)^{2k}\right]$. Покажем еще, как можно уточнить значение λ_1, если A есть симметрическая матрица. Положим

$$\bar{x}_1^* = \bar{x}_1 + \bar{\varepsilon}, \quad \lambda_1^* = \lambda_1 + \delta. \tag{8}$$

Тогда, так как

$$\left.\begin{aligned} A\bar{x}_1 - \lambda_1\bar{x}_1 &= 0, \\ A\bar{x}_1^* - \lambda_1^*\bar{x}_1^* &= \bar{\eta}, \end{aligned}\right\} \tag{9}$$

где $\bar{\eta}$, вообще говоря, отличен от нуля, то

$$A\bar{\varepsilon} - \lambda_1\bar{\varepsilon} - \delta\bar{x}_1 - \delta\bar{\varepsilon} = \bar{\eta}. \tag{10}$$

Считая $\bar{\delta}$ и $\bar{\varepsilon}$ малыми, отбросим последний член левой части как малую величину более высокого порядка. При этом равенство (10) перейдет в

$$A\bar{\varepsilon} - \lambda_1\bar{\varepsilon} - \delta\bar{x}_1 = \bar{\eta}. \tag{11}$$

Умножим скалярно обе части равенства (11) на $\bar{x}_1$. Получим:

$$(A\bar{\varepsilon},\ \bar{x}_1) - \lambda_1(\bar{\varepsilon},\ \bar{x}_1) - \delta(\bar{x}_1,\ \bar{x}_1) = (\bar{\eta},\ \bar{x}_1). \tag{12}$$

Но

$$(A\bar{\varepsilon},\ \bar{x}_1) = (\bar{\varepsilon},\ A\bar{x}_1) = \lambda_1(\bar{\varepsilon},\ \bar{x}_1) \tag{13}$$

и $(\bar{x}_1,\ \bar{x}_1) = 1$. Следовательно, равенство (12) даст

$$\delta = -(\bar{\eta},\ \bar{x}_1). \tag{14}$$

Полученное δ можно использовать для уточнения найденного приближенного значения λ_1^*, если заменить в правой части $\bar{x}_1$ на $\bar{x}_1^*$.

Умножим теперь обе части равенства (11) на $\bar{x}_i$ ($i = 2, 3, \ldots, n$). Получим:

$$(A\bar{\varepsilon},\ \bar{x}_i) - \lambda_1(\bar{\varepsilon},\ \bar{x}_i) = (\bar{\eta},\ \bar{x}_i). \tag{15}$$

Но

$$(A\bar{\varepsilon},\ \bar{x}_i) = (\bar{\varepsilon},\ A\bar{x}_i) = \lambda_i(\bar{\varepsilon},\ \bar{x}_i) \tag{16}$$

и, следовательно,

$$(\lambda_i - \lambda_1)(\bar{\varepsilon},\ \bar{x}_i) = (\bar{\eta},\ \bar{x}_i), \tag{17}$$

$$(\bar{\varepsilon},\ \bar{x}_i) = \frac{(\bar{\eta},\ \bar{x}_i)}{\lambda_i - \lambda_1} \qquad (i = 2,\ 3,\ \ldots\ n). \tag{18}$$

Найдем еще $(\bar{\varepsilon},\ \bar{x}_1)$. Так как

$$\|\bar{x}_1^*\| = \|\bar{x}_1 + \bar{\varepsilon}\| = 1, \tag{19}$$

то

$$1 = (\bar{x}_1 + \bar{\varepsilon},\ \bar{x}_1 + \bar{\varepsilon}) = 1 + 2(\bar{x}_1,\ \bar{\varepsilon}) + (\bar{\varepsilon},\ \bar{\varepsilon}) \tag{20}$$

и

$$(\bar{x}_1,\ \bar{\varepsilon}) = -\frac{1}{2}(\bar{\varepsilon},\ \bar{\varepsilon}). \tag{21}$$

Выражения (18) и (21) позволяют оценить норму вектора $\bar{\varepsilon}$. Действительно,

$$\|\bar{\varepsilon}\|^2 = (\bar{\varepsilon},\ \bar{\varepsilon}) = \left(\sum_{i=1}^{n}(\bar{\varepsilon},\ \bar{x}_i)\,\bar{x}_i,\ \sum_{i=1}^{n}(\bar{\varepsilon},\ \bar{x}_i)\,\bar{x}_i\right) =$$

$$= \sum_{i=1}^{n}(\bar{\varepsilon},\ \bar{x}_i)^2 = \frac{1}{4}\|\bar{\varepsilon}\|^2 + \sum_{i=2}^{n}(\bar{\varepsilon},\ \bar{x}_i)^2. \tag{22}$$

Таким образом,

$$\|\bar{\varepsilon}\|^2 = \frac{4}{3}\sum_{i=2}^{n}(\bar{\varepsilon},\ \bar{x}_i)^2. \tag{23}$$

Перепишем (18) в виде

$$|(\bar{\varepsilon},\ \bar{x}_i)| = \frac{|(\bar{\eta},\ \bar{x}_i)|}{|\lambda_i - \lambda_1|} = \frac{|(\bar{\eta},\ \bar{x}_i)|}{|\lambda_1|}\,\frac{1}{\left|1-\frac{\lambda_i}{\lambda_1}\right|} \qquad (i=2,\ 3,\ \ldots,\ n). \tag{24}$$

Так как

$$\left|1-\frac{\lambda_i}{\lambda_1}\right| \geqslant 1-\left|\frac{\lambda_i}{\lambda_1}\right|, \tag{25}$$

то

$$|(\bar{\varepsilon},\ \bar{x}_i)| \leqslant \frac{|(\bar{\eta},\ \bar{x}_i)|}{|\lambda_1|}\,\frac{1}{1-\left|\frac{\lambda_i}{\lambda_1}\right|} \qquad (i=2,\ 3,\ \ldots,\ n). \tag{26}$$

В силу (23) будем иметь:

$$\|\bar{\varepsilon}\|^2 \leqslant \frac{4}{3}\sum_{i=2}^{n}\frac{|(\bar{\eta},\ \bar{x}_i)|^2}{\lambda_1^2}\,\frac{1}{\left(1-\left|\frac{\lambda_i}{\lambda_1}\right|\right)^2}, \tag{27}$$

и так как $\frac{|\lambda_i|}{|\lambda_1|}$ $(i=3,\ 4,\ \ldots,\ n)$ меньше чем $\frac{|\lambda_2|}{|\lambda_1|}$, то

$$\|\bar{\varepsilon}\| \leqslant \frac{2}{\sqrt{3}}\,\frac{\|\bar{\eta}\|}{|\lambda_1|\left(1-\left|\frac{\lambda_2}{\lambda_1}\right|\right)}. \tag{28}$$

Это выражение может служить оценкой для $\|\bar{\varepsilon}\|$, если использовать вместо λ_1 и λ_2 какие-то их приближенные значения.

2. Отыскание других собственных значений и соответствующих им собственных векторов для симметрических матриц. Используем теперь векторы $A^k\bar{v}$ для отыскания других собственных значений. Обозначим i-ю компоненту вектора $A^k\bar{v}$ через v_{ki}. Если единичные векторы $\bar{e}_i$, направленные по осям координат, представить в виде

$$\bar{e}_i = e_{i1}\bar{x}_1 + e_{i2}\bar{x}_2 + \ldots + e_{in}\bar{x}_n, \tag{29}$$

то будем иметь:

$$v_{ki} = (A^k\bar{v},\ \bar{e}_i) = \left(\sum_{j=1}^{n}\alpha_j\lambda_j^k\bar{x}_j,\ \sum_{j=1}^{n}e_{ij}\bar{x}_j\right) = \sum_{j=1}^{n}\lambda_j^k\alpha_j e_{ij} \tag{30}$$

или

$$v_{ki} = b_{i1}\lambda_1^k + b_{i2}\lambda_2^k + \ldots + b_{in}\lambda_n^k, \tag{31}$$

где обозначено

$$b_{ij} = \alpha_j e_{ij}. \tag{32}$$

Составим определитель

$$\beta_{rs}^{(k)} = \begin{vmatrix} v_{kr} & v_{ks} \\ v_{k+1,\,r} & v_{k+1,\,s} \end{vmatrix}. \tag{33}$$

Воспользовавшись равенствами (31), получим:

$$\beta_{rs}^{(k)} = (b_{r2}b_{s1} - b_{s2}b_{r1})\lambda_1^{k+1}\lambda_2^k + (b_{r1}b_{s2} - b_{s1}b_{r2})\lambda_1^k\lambda_2^{k+1} + \\ + (b_{r3}b_{s1} - b_{r1}b_{s3})\lambda_1^{k+1}\lambda_3^k + (b_{r1}b_{s3} - b_{s1}b_{r3})\lambda_1^k\lambda_3^{k+1} + \ldots, \tag{34}$$

где многоточием обозначены остальные члены, содержащие λ_3, λ_4, ... Таким образом,

$$\beta_{rs}^{(k)} = \lambda_1^k\lambda_2^k\left\{(b_{r2}b_{s1} - b_{s2}b_{r1})(\lambda_1 - \lambda_2) + \right. \\ \left. + (b_{r3}b_{s1} - b_{r1}b_{s3})\frac{\lambda_3^k(\lambda_1 - \lambda_3)}{\lambda_2^k} + \ldots\right\}. \tag{35}$$

Аналогично найдем:

$$\beta_{rs}^{(k+1)} = \lambda_1^{k+1}\lambda_2^{k+1}\left\{(b_{r2}b_{s1} - b_{s2}b_{r1})(\lambda_1 - \lambda_2) + \right. \\ \left. + (b_{r3}b_{s1} - b_{r1}\lambda_{s3})\frac{\lambda_3^{k+1}(\lambda_1 - \lambda_3)}{\lambda_2^{k+1}} + \ldots\right\}. \tag{36}$$

Отсюда

$$\frac{\beta_{rs}^{(k+1)}}{\beta_{rs}^{(k)}} = \lambda_1\lambda_2\frac{(b_{r2}b_{s1} - b_{s2}b_{r1})(\lambda_1 - \lambda_2) + (b_{r3}b_{s1} - b_{r1}b_{s3})\frac{\lambda_3^{k+1}(\lambda_1 - \lambda_3)}{\lambda_2^{k+1}} + \ldots}{(b_{r2}b_{s1} - b_{s2}b_{r1})(\lambda_1 - \lambda_2) + (b_{r3}b_{s1} - b_{r1}b_{s3})\frac{\lambda_3^k(\lambda_1 - \lambda_3)}{\lambda_2^k} + \ldots}. \tag{37}$$

Пусть λ_1 и λ_2 превышают по модулю остальные значения $|\lambda_i|$ и $(b_{r2}b_{s1} - b_{s2}b_{r1}) \neq 0$. Тогда

$$\frac{\beta_{rs}^{(k+1)}}{\beta_{rs}^{(k)}} = \lambda_1\lambda_2 + 0\left[\left(\frac{\lambda_3}{\lambda_2}\right)^k\right]. \tag{38}$$

Аналогично можно получить произведение трех и большего числа собственных значений. Так, если λ_1, λ_2 и λ_3 превышают по модулю остальные собственные значения, то

$$\frac{\gamma_{rst}^{(k+1)}}{\gamma_{rst}^{(k)}} = \lambda_1\lambda_2\lambda_3 + 0\left[\left(\frac{\lambda_4}{\lambda_3}\right)^k\right], \tag{39}$$

где

$$\gamma_{rst}^{(k)} = \begin{vmatrix} v_{kr} & v_{ks} & v_{kt} \\ v_{k+1,\,r} & v_{k+1,\,s} & v_{k+1,\,t} \\ v_{k+2,\,r} & v_{k+2,\,s} & v_{k+2,\,t} \end{vmatrix}. \tag{40}$$

Нужно отметить, что при больших k строки написанных выше определителей почти пропорциональны. Следовательно, сами определители будут близки к нулю. Это приведет к большой вычислительной ошибке. Если же ограничиваться небольшими значениями k, то получим большую ошибку метода. Вследствие этих причин λ_2, λ_3 и т. д. можно найти с небольшой точностью.

В связи с этим рассмотрим еще один метод отыскания промежуточных собственных значений в случае, когда A есть симметрическая матрица. После того как λ_1 и $\bar{x}_1$ найдены, возьмем вместо A новую матрицу:

$$A_1 = A - \lambda_1 \bar{x}_1 \bar{x}_1', \tag{41}$$

где под $\bar{x}_1\bar{x}_1'$ понимается произведение вектора-столбца с компонентами $\bar{x}_1$ и вектора-строки с теми же компонентами по правилу умножения матриц. A_1 — также симметрическая матрица. При этом

$$A_1\bar{x}_1 = A\bar{x}_1 - \lambda_1\bar{x}_1\bar{x}_1'\bar{x}_1 = A\bar{x}_1 - \lambda_1\bar{x}_1 = 0, \tag{42}$$

так как

$$\bar{x}_1'\bar{x}_1 = (\bar{x}_1,\ \bar{x}_1) = ||\bar{x}_1||^2 = 1. \tag{43}$$

В то же время

$$A_1\bar{x}_i = A\bar{x}_i - \lambda_i\bar{x}_1\bar{x}_1'\bar{x}_i = A\bar{x}_i = \lambda_i\bar{x}_i \quad (i = 2,\ 3,\ 4,\ \ldots,\ n), \tag{44}$$

так как векторы $\bar{x}_i$ и $\bar{x}_j$ ортогональны при $i \neq j$. Таким образом, матрица A_1 будет иметь те же собственные значения и собственные векторы, что и A, за исключением λ_1. Вместо λ_1 матрица A_1 будет иметь собственное значение 0, и ему соответствует собственный вектор $\bar{x}_1$. После того как будут найдены λ_2 и $\bar{x}_2$, можно дальше повторить этот процесс, и мы найдем, в конце концов, все λ_i и $\bar{x}_i$. При этом можно каждый раз понижать порядок матрицы. Пусть $\bar{x}_1 = (x_{11},\ x_{21},\ \ldots,\ x_{n1})$. Обозначим

$$\alpha = \begin{pmatrix} x_{21} \\ x_{31} \\ \cdot \\ \cdot \\ x_{n1} \end{pmatrix} \tag{45}$$

и рассмотрим матрицу

$$B = \begin{pmatrix} x_{11} & -\alpha' \\ \alpha & I_{n-1} - \mu\alpha\alpha' \end{pmatrix}, \tag{46}$$

где μ — некоторая постоянная.

Произведение BB' будет равно

$$BB' = \begin{pmatrix} x_{11}^2 + \alpha'\alpha & x_{11}\alpha' - \alpha' + \mu\alpha'\alpha\alpha' \\ x_{11}\alpha - \alpha + \mu\alpha\alpha'\alpha & \alpha\alpha' + I_{n-1} - 2\mu\alpha\alpha' + \mu^2\alpha\alpha'\alpha\alpha' \end{pmatrix}. \tag{47}$$

Но

$$x_{11}^2 + \alpha'\alpha = \|x_1\|^2 = 1 \tag{48}$$

и

$$x_{11}\alpha' - \alpha' + \mu\alpha'\alpha\alpha' = \left[x_{11} - 1 + \mu\left(1 - x_{11}^2\right)\right]\alpha'. \tag{49}$$

Таким образом, если взять $\mu = \frac{1}{(1 + x_{11})}$, то все элементы первой строки (47) обратятся в нуль, кроме левого крайнего, который будет равен единице. В силу симметрии матрицы BB' все элементы первого столбца (47), кроме верхнего, будут нулями. Далее, при $\mu = \frac{1}{(1 + x_{11})}$ имеем:

$$\alpha\alpha' - 2\mu\alpha\alpha' + \mu^2\alpha\alpha'\alpha\alpha' = \frac{x_{11} - 1}{x_{11} + 1}\alpha\alpha' + \frac{1}{(1 + x_{11})^2}\left(1 - x_{11}^2\right)\alpha\alpha' = 0. \tag{50}$$

Итак, $BB' = I$. Следовательно, $B'AB$ имеет такие же собственные значения, что и A. Возьмем матрицу

$$\beta = \begin{pmatrix} 1 \\ 0 \\ \cdot \\ \cdot \\ 0 \end{pmatrix}. \tag{51}$$

Произведение $B'AB\beta$ равно $\lambda_1\beta$. Действительно, $B\beta = \bar{x}_1$, $A\bar{x}_1 = \lambda_1\bar{x}_1$ и $\lambda_1 B'\bar{x}_1 = \lambda_1\beta$. Таким образом, β является собственным вектором матрицы $B'AB$, соответствующим собственному значению λ_1. Это может быть только в том случае, если $B'AB$ имеет вид

$$B'AB = \begin{pmatrix} \lambda_1 & 0 \\ 0 & A_1 \end{pmatrix}. \tag{52}$$

Дальнейшие вычисления можно производить с симметрической матрицей A_1, имеющей собственные значения $\lambda_2, \lambda_3, \ldots, \lambda_n$. Если $\bar{y}_i$ — собственный вектор A_1, соответствующий собственному значению λ_i, то из

$$A_1\bar{y}_i = \lambda_i\bar{y}_i \tag{53}$$

следует:

$$B'AB\begin{pmatrix}0\\ \bar{y}_i\end{pmatrix}=\begin{pmatrix}\lambda_1 & 0\\ 0 & A_1\end{pmatrix}\begin{pmatrix}0\\ \bar{y}_i\end{pmatrix}=\begin{pmatrix}0\\ \lambda_i\bar{y}_i\end{pmatrix}=\lambda_i\begin{pmatrix}0\\ y_i\end{pmatrix} \tag{54}$$

и

$$AB\begin{pmatrix}0\\ \bar{y}_i\end{pmatrix}=\lambda_i B\begin{pmatrix}0\\ \bar{y}_i\end{pmatrix}. \tag{55}$$

Таким образом, $B\begin{pmatrix}0\\ \bar{y}_i\end{pmatrix}$ будет собственным вектором A, соответствующим собственному значению λ_i. В дальнейшем мы можем поступить с A_1 так же, как мы поступили с A.

Рассмотрим еще один способ получения $\lambda_2, \lambda_3, \ldots, \lambda_n$. Пусть λ_1 и $\bar{x}_1$ уже найдены. Рассмотрим произвольный вектор $\bar{x}$ и образуем

$$\bar{y}=\bar{x}-(\bar{x}, \bar{x}_1)\bar{x}_1. \tag{56}$$

Этот вектор ортогонален $\bar{x}_1$. Действительно,

$$(\bar{x}_1, \bar{y})=(\bar{x}, \bar{x}_1)-(\bar{x}, \bar{x}_1)(\bar{x}_1, \bar{x}_1)=0, \tag{57}$$

так как $(\bar{x}_1, \bar{x}_1)=\|\bar{x}_1\|^2=1$. Поэтому разложение $\bar{y}$ по векторам $\bar{x}_i$ имеет вид

$$\bar{y}=\alpha_2\bar{x}_2+\alpha_3\bar{x}_3+\ldots+\alpha_n\bar{x}_n \tag{58}$$

и

$$A^k\bar{y}=\alpha_2\lambda_2^k\bar{x}_2+\alpha_3\lambda_3^k\bar{x}_3+\ldots+\alpha_n\lambda_n^k\bar{x}_n. \tag{59}$$

Таким образом, если $|\lambda_2|>|\lambda_i|$ при $i\geqslant 3$, то

$$A^k\bar{y}\approx\alpha_2\lambda_2^k\bar{x}_2, \quad \frac{(A^{k+1}\bar{y})_i}{(A^k\bar{y})_i}\approx\lambda_2. \tag{60}$$

Найдя λ_2 и $\bar{x}_2$, можно искать следующее собственное значение и следующий собственный вектор, взяв за начальный вектор:

$$\bar{z}=\bar{x}-(\bar{x}, \bar{x}_1)\bar{x}_1-(\bar{x}, \bar{x}_2)\bar{x}_2 \tag{61}$$

и т. д. Нужно только помнить, что при этом в связи с ошибками округления при итерациях будут появляться компоненты собственных векторов, соответствующих наибольшим по модулю собственным значениям матрицы A. Поэтому время от времени необходимо исключать такие компоненты по формулам (56), (61) или им подобным.

Промежуточные собственные значения можно также определить, если рассмотреть вместо матрицы A матрицу $A-\mu I$ или $\mu I-A$, где μ — соответствующим образом подобранное число. В частности, если $\mu>\lambda_1$ и A — положительно определенная матрица, то итерация с матрицей $\mu I-A$ даст наименьшее собственное значение. Можно использовать и многочлены более высокой степени относительно матрицы A.

Пример. Проиллюстрируем теперь наши рассуждения простым примером. Пусть матрица A имеет вид

$$A=\begin{Vmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{Vmatrix}. \tag{62}$$

За начальный вектор примем $\bar{x}=(1, -1, 1, -1)$. Итерации дадут:

№	1	2	3	4
	2 −1 0 0	−1 2 −1 0	0 −1 2 −1	0 0 −1 2
0	1	−1	1	−1
1	3	−4	4	−3
2	10	−15	15	−10
3	35	−55	55	−35
4	125	−200	200	−125
5	450	−725	725	−450
6	1 625	−2 625	2 625	−1 625
7	5 875	−9 500	9 500	−5 875
8	21 250	−34 375	34 375	−21 250

Отношения соответствующих компонент восьмой и седьмой итераций последовательно равны

3,61702; 3,61842; 3,61842; 3,61702.

Они уже достаточно близки. Более точное приближение для собственного значения мы получим, если поделим скалярный квадрат вектора, полученного при восьмой итерации, на скалярное произведение векторов, полученных при седьмой и восьмой итерациях. В результате получим:

$$\lambda_1^* = 3{,}61804. \tag{63}$$

Соответствующий нормированный собственный вектор будет иметь вид

$$\bar{x}_1^* = (0{,}37182;\ -0{,}60147;\ 0{,}60147;\ -0{,}37182). \tag{64}$$

При этом оказывается, что

$$A\bar{x}_1^* - \lambda_1^*\bar{x}_1^* = (-0{,}000147;\ -0{,}000098;\ 0{,}000098;\ 0{,}000147), \tag{65}$$

и поправка δ по формуле (14) настоящего параграфа примет вид

$$\delta = -(\bar{\eta}, \bar{x}_1^*) = -0{,}000008. \tag{66}$$

Составим теперь определители (33) для нашего случая. Получим:

$$\beta_{12}^{(7)} = \begin{vmatrix} 5875 & -9500 \\ 21\,250 & -34\,375 \end{vmatrix} = -78\,125; \quad \beta_{12}^{(6)} = \begin{vmatrix} 1625 & -2625 \\ 5875 & -9500 \end{vmatrix} = -15\,625. \tag{67}$$

Остальные определители ничего нового не добавят. Отношение $\beta_{12}^{(7)}/\beta_{12}^{(6)}$ в данном случае равно

$$\frac{\beta_{12}^{(7)}}{\beta_{12}^{(6)}} = 5 \quad \text{и} \quad \lambda_2^* = 1{,}38196. \tag{68}$$

Определители $\gamma_{rst}^{(6)}$ в данном случае все равны нулю. Поэтому для отыскания следующих собственных значений применим преобразование матрицы A с помощью матрицы B (46). Вычисления дают:

$$B = \begin{pmatrix} 0{,}37182 & 0{,}60147 & -0{,}60147 & 0{,}37182 \\ -0{,}60147 & 0{,}73628 & 0{,}26372 & -0{,}16302 \\ 0{,}60147 & 0{,}26372 & 0{,}73628 & 0{,}16302 \\ -0{,}37182 & -0{,}16302 & -0{,}16302 & 0{,}89922 \end{pmatrix}$$

и

$$B'AB = \begin{pmatrix} 3{,}61815 & 0{,}00007 & 0{,}00008 & 0{,}00012 \\ 0{,}00007 & 0{,}81194 & -0{,}25039 & -0{,}46329 \\ 0{,}00008 & -0{,}25039 & 1{,}68884 & -0{,}80775 \\ 0{,}00012 & -0{,}46329 & -0{,}80775 & 1{,}88119 \end{pmatrix}.$$

Расхождения с теоретическими результатами получились за счет ошибок округления.

Теперь мы должны отыскивать собственные значения матрицы третьего порядка:

$$A_1 = \begin{pmatrix} 0{,}81194 & -0{,}25039 & -0{,}46329 \\ -0{,}25039 & 1{,}68884 & -0{,}80775 \\ -0{,}46329 & -0{,}80775 & 1{,}88119 \end{pmatrix}. \tag{69}$$

Получим их итерационным способом, начиная с вектора $\bar{x} = (0, 0, 1)$. Итерированные векторы будут иметь следующие компоненты:

0	0	0	1
1	—0,46329	—0,80775	1,88119
2	—1,04545	—0,05552	4,40597
3	—2,87618	—3,39092	8,81766
4	—5,57137	—12,12902	20,65291
5	—11,05492	—35,77135	51,23042
6	—23,75378	—99,02542	130,3901
7	—54,90002	—266,1301	333,2812

Отношения соответствующих компонент последних двух векторов равны последовательно 2,311; 2,687; 2,556. Приближенное значение λ_2 будем находить, используя скалярные произведения так же, как и в предыдущем случае. При этом получим:

$$\lambda_2^* = 2{,}6003. \tag{70}$$

Обращаем внимание на то, что у нас получилось другое значение λ_2^*, чем в первом случае. В данном случае это произошло благодаря тому, что начальный вектор при итерации с помощью матрицы A оказался ортогональным к собственному вектору $\bar{x}_2$. Таким образом, фактически (68) дает не λ_2^*, а λ_3^*. Точные значения собственных значений матрицы A с шестью десятичными знаками таковы:

$$\lambda_1 = 3{,}618034, \quad \lambda_2 = 2{,}618034, \quad \lambda_3 = 1{,}381966, \quad \lambda_4 = 0{,}381966.$$

3. Отыскание собственных значений и собственных векторов несимметрических матриц, имеющих простую структуру. Обобщим теперь полученные для симметрических матриц результаты на более общий случай. Предварительно напомним некоторые факты из линейной алгебры. Пусть A — произвольная матрица с действительными или комплексными элементами a_{ik}. Матрица A^*, элементы которой a_{ik}^* удовлетворяют условиям $a_{ik}^* = \bar{a}_{ki}$ (транспонирование и комплексная сопряженность), называется *сопряженной* по отношению к A. Если $A = A^*$, то матрица A называется *эрмитовой*. Эрмитова матрица имеет простую структуру и все ее собственные значения действительны. Систему собственных векторов эрмитовой матрицы можно считать ортонормированной в соответствующем комплексном векторном n-мерном пространстве R. Вследствие этого на эрмитовы матрицы можно перенести с необходимыми видоизменениями результаты, полученные для симметрических матриц.

Рассмотрим связь между инвариантными многообразиями матрицы A и сопряженной матрицы A^*. Пусть некоторое линейное многообразие M инвариантно относительно A, т. е. из $\bar{x} \in M$ следует $A\bar{x} \in M$. Рассмотрим совокупность N векторов $\bar{y} \in R$, ортогональных к каждому вектору $\bar{x} \in M$. Если размерность M меньше n, то множество N не пусто. Очевидно, N в свою очередь является линейным многообразием. Покажем, что это многообразие инвариантно относительно A^*. Действительно, если $\bar{x}$ — произвольный вектор из M, то для любого вектора $\bar{y} \in N$ имеет место $(A\bar{x}, \bar{y}) = 0$. Но $(A\bar{x}, \bar{y}) = (\bar{x}, A^*\bar{y})$ и, следовательно, $A^*\bar{y} \in N$, что и требовалось доказать.

Пусть матрица A имеет простую структуру и $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n$ — ее линейно независимые собственные векторы, соответствующие собственным значениям $\lambda_1, \lambda_2, \ldots, \lambda_n$. Линейное многообразие, построенное на векторах $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{k-1}, \bar{x}_{k+1}, \ldots, \bar{x}_n$, будет иметь раз-

мерность $n-1$. Следовательно, в R имеется вектор $\bar{y}_k$, ортогональный этому многообразию. Как явствует из предыдущего, $\bar{y}_k$ будет являться собственным вектором A^*. При этом $\bar{y}_k$ не может быть ортогональным $\bar{x}_k$. Следовательно, умножив его на подходящий множитель, можно достигнуть того, что $(\bar{x}_k, \bar{y}_k) = 1$. Проведя эти рассуждения для всех $k = 1, 2, \ldots, n$, мы придем к выводу, что A^* имеет простую структуру и собственные векторы A и A^* можно выбрать так, что они будут образовывать биортонормированную систему, т. е. $(\bar{x}_i, \bar{y}_j) = \delta_{ij}$.

Отметим еще, что если A и A^* имеют общий собственный вектор $\bar{x} \neq 0$, то собственные значения, которым соответствует этот собственный вектор, будут комплексно сопряжены. Действительно, если $A\bar{x} = \lambda\bar{x}$ и $A^*\bar{x} = \mu\bar{x}$, то

$$\lambda(\bar{x}, \bar{x}) = (A\bar{x}, \bar{x}) = (\bar{x}, A^*\bar{x}) = \bar{\mu}(\bar{x}, \bar{x}), \tag{71}$$

откуда и следует утверждение.

Матрицу A называют *нормальной*, если она перестановочна с своей сопряженной $AA^* = A^*A$. Перестановочные матрицы A и B всегда имеют общий собственный вектор. Действительно, если $A\bar{x} = \lambda\bar{x}$, $\bar{x} \neq 0$, то $AB^k\bar{x} = \lambda B^k\bar{x}$. Начиная с некоторого p, линейное многообразие M, построенное на векторах $\bar{x}$, $B\bar{x}$, ..., $B^{p-1}\bar{x}$, будет инвариантно относительно B. Следовательно, в этом многообразии будет существовать собственный вектор B: $B\bar{y} = \mu\bar{y}$. Но любой вектор этого многообразия является собственным для A. Тем самым утверждение доказано. В частности, если A — нормальная матрица, то для A и A^* будет иметься общий собственный вектор. Как следует из предыдущего, собственные значения, которым соответствует этот общий собственный вектор, будут комплексно сопряжены. Обозначим найденный таким образом общий собственный вектор A и A^* через $\bar{x}_1$ и пусть $A\bar{x}_1 = \lambda_1\bar{x}_1$. Тогда $A^*\bar{x}_1 = \bar{\lambda}_1\bar{x}_1$. Рассмотрим линейное многообразие M векторов, ортогональных к $\bar{x}_1$. Как следует из предыдущего, это линейное многообразие будет инвариантно как относительно A, так и относительно A^*. Такими же рассуждениями, как и ранее, придем к выводу, что в M будет существовать общий собственный вектор $\bar{x}_2$ для A и A^*. При этом, если $A\bar{x}_2 = \lambda_2\bar{x}_2$, то $A^*\bar{x}_2 = \bar{\lambda}_2\bar{x}_2$. Очевидно, $(\bar{x}_1, \bar{x}_2) = 0$. Затем можно провести такие же рассуждения в линейном многообразии векторов, ортогональных к $\bar{x}_1$ и $\bar{x}_2$, и т. д. В конце концов, мы придем к заключению, что *нормальная матрица имеет полную ортонормированную систему собственных векторов*. Эти векторы являются также собственными векторами A^*, причем соответствующие собственные значения комплексно сопряжены.

Можно показать, что наличие полной ортонормированной системы собственных векторов является необходимым и достаточным условием нормальности матрицы. Нормальная матрица будет являться эрмитовой тогда и только тогда, когда все ее собственные значения действительны.

Из изложенного видно, что и для нормальных матриц можно провести рассуждения, аналогичные тем, которые были проведены для симметрических матриц. При отыскании модулей собственных значений матрицы простой структуры, не являющейся нормальной, может оказаться целесообразным наряду с векторами $A^k\bar{x}$ образовывать векторы $A^{*k}\bar{y}$ и рассматривать их скалярное произведение. При этом если

$$\left.\begin{aligned}\bar{x} &= \alpha_1\bar{x}_1 + \alpha_2\bar{x}_2 + \ \dots\ + \alpha_n\bar{x}_n,\\ \bar{y} &= \beta_1\bar{y}_1 + \beta_2\bar{y}_2 + \ \dots\ + \beta_n\bar{y}_n,\end{aligned}\right\} \tag{72}$$

где $\{\bar{x}_k\}$ и $\{\bar{y}_k\}$ образуют биортогональную систему векторов, то

$$(A^k\bar{x},\ A^{*k}\bar{y}) = \alpha_1\beta_1|\lambda_1|^{2k} + \alpha_2\beta_2|\lambda_2|^{2k} + \ \dots\ + \alpha_n\beta_n|\lambda_n|^{2k}. \tag{73}$$

Если матрица A имеет несколько равных или близких корней, то это не вызывает никаких принципиальных затруднений. Так, если для матрицы A простой структуры $\lambda_1 = \lambda_2$, то разложение (2) примет вид

$$A^k\bar{v} = \alpha_1\lambda_1^k\bar{x}_1 + \alpha_2\lambda_1^k\bar{x}_2 + \alpha_3\lambda_3^k\bar{x}_3 + \ \dots\ + \alpha_n\lambda_n^k\bar{x}_n. \tag{74}$$

Таким образом, при $|\lambda_1| > |\lambda_i|$ $(i = 3,\ 4,\ \dots,\ n)$ снова будем иметь:

$$\lambda_1^* \approx \frac{(A^{k+1}\bar{v})_i}{(A^k\bar{v})_i}. \tag{75}$$

При этом один из собственных векторов, соответствующих собственному значению λ_1, будет приближенно равен $A^k\bar{v}$ при достаточно большом k. Второй собственный вектор, соответствующий собственному значению λ_1, можно получить, если взять другой начальный вектор. Однако неизбежные ошибки округления будут изменять компоненты $A^k\bar{v}$ по направлению векторов $\bar{x}_1$ и $\bar{x}_2$ и тем самым замедлять сходимость. Медленная сходимость будет наблюдаться также и при наличии близких корней. Если матрица имеет два равных по модулю, но различных корня, превышающих по абсолютной величине все остальные корни, то сходимости вообще наблюдаться не будет.

Во всех трех случаях мы сможем при больших k написать приближенные равенства:

$$\left.\begin{aligned}A^k\bar{v} &\approx \alpha_1\lambda_1^k\bar{x}_1 + \alpha_2\lambda_2^k\bar{x}_2,\\ A^{k+1}\bar{v} &\approx \alpha_1\lambda_1^{k+1}\bar{x}_1 + \alpha_2\lambda_2^{k+1}\bar{x}_2,\\ A^{k+2}\bar{v} &\approx \alpha_1\lambda_1^{k+2}\bar{x}_1 + \alpha_2\lambda_2^{k+2}\bar{x}_2.\end{aligned}\right\} \tag{76}$$

Таким образом, между тремя векторами $A^k\bar{v}$, $A^{k+1}\bar{v}$ и $A^{k+2}\bar{v}$ будет иметь место приближенная линейная зависимость. Нетрудно проверить, что эта линейная зависимость может быть записана в виде

$$A^{k+2}\bar{v} - (\lambda_1 + \lambda_2) A^{k+1}\bar{v} + \lambda_1\lambda_2 A^k\bar{v} = 0. \tag{77}$$

Следовательно, если в процессе вычислений мы обнаружим, что векторы $A^k\bar{v}$, $A^{k+1}\bar{v}$ и $A^{k+2}\bar{v}$ связаны некоторым линейным соотношением вида

$$A^{k+2}\bar{v} + pA^{k+1}\bar{v} + qA^k\bar{v} = 0, \tag{78}$$

то λ_1 и λ_2 будут удовлетворять квадратному уравнению

$$z^2 + pz + q = 0. \tag{79}$$

Фактически квадратное уравнение можно получить, если рассмотреть определители

$$\begin{vmatrix} 1 & v_{k,\,r} & v_{k,\,s} \\ z & v_{k+1,\,r} & v_{k+1,\,s} \\ z^2 & v_{k+2,\,r} & v_{k+2,\,s} \end{vmatrix} = 0 \qquad (r,\ s = 1,\ 2,\ \ldots,\ n;\ r \neq s;\ v_{ki} = (A^k\bar{v})_i). \tag{80}$$

Рассмотрим пример на применение этого метода. Ниже приведены матрица четвертого порядка, для которой ищутся собственные значения и собственные векторы, и результаты итерации с ее помощью вектора (—1, 1, 0, 0):

№	1	2	3	4
	—2	1	1	1
	—7	—5	—2	—4
	0	—1	—3	—2
	—1	0	—1	0
0	—1	1	0	0
1	3	2	—1	1
2	—4	—33	—1	—2
3	—28	203	40	5
4	304	—919	—333	—12
5	—1 872	3 181	1 942	29
6	8 896	—6 081	—9 065	—70
7	—33 728	—9 857	34 136	169
8	91 904	216 433	—92 889	—408
9	—60 672	—1 538 083	63 050	985

Как видно из этой таблицы, отношения соответствующих компонент последовательных итераций ведут себя довольно беспорядочно,

имеют место перемены знаков. Это указывает на наличие комплексных корней. Уравнение для определения λ_1 и λ_2 будет иметь вид

$$\begin{vmatrix} 1 & -33\,728 & -9\,857 \\ z & 91\,904 & 216\,433 \\ z^2 & -60\,672 & -1\,538\,083 \end{vmatrix} = 0. \tag{81}$$

Отсюда

$$z^2 + 8{,}02z + 20{,}05 = 0 \tag{82}$$

и

$$z = -4{,}01 \pm 1{,}99i. \tag{83}$$

Точные значения корней в данном случае будут

$$z = -4 \pm 2i. \tag{84}$$

У нас получились довольно грубые значения. Это объясняется тем, что мы провели недостаточное число итераций и, вообще говоря, корни определять было еще рано.

После того как будет решено уравнение (79), можно найти и собственные векторы. Из равенств (76) получаем:

$$\left.\begin{aligned} A^{k+1}\bar{v} - \lambda_1 A^k \bar{v} &= \alpha_2 \lambda_2^k (\lambda_2 - \lambda_1)\bar{x}_1, \\ A^{k+1}\bar{v} - \lambda_2 A^k \bar{v} &= \alpha_1 \lambda_1^k (\lambda_1 - \lambda_2)\bar{x}_2. \end{aligned}\right\} \tag{85}$$

Эти результаты можно обобщить на случай, когда имеется более двух равных по модулю или близких по модулю собственных значений.

4. Некоторые замечания об отыскании собственных значений и собственных векторов матриц общей структуры. Рассмотрим теперь кратко случай, когда матрица не является матрицей простой структуры. Если элементарные делители $A - \lambda I$ имеют вид

$$(\lambda - \lambda_1)^{k_1}, \quad (\lambda - \lambda_2)^{k_2}, \quad \ldots, \quad (\lambda - \lambda_m)^{k_m} \tag{86}$$

(некоторые λ_i могут совпадать), то найдутся такие векторы:

$$\bar{e}_1^{(i)}, \ \bar{e}_2^{(i)}, \ \ldots, \ \bar{e}_{k_i}^{(i)} \qquad (i = 1, 2, \ldots, m), \tag{87}$$

что

$$A\bar{e}_1^{(i)} = \lambda_i \bar{e}_1^{(i)}; \ A\bar{e}_2^{(i)} = \lambda_i \bar{e}_2^{(i)} + \bar{e}_1^{(i)}; \ \ldots; \ A\bar{e}_{k_i}^{(i)} = \lambda_i \bar{e}_{k_i}^{(i)} + \bar{e}_{k_i - 1}. \tag{88}$$

Общее количество векторов $\bar{e}_j^{(i)}$ равно порядку матрицы n, и их можно принять за базис n-мерного векторного пространства.

Запишем произвольный вектор $\bar{v}$ в виде

$$\begin{aligned} \bar{v} = \alpha_1^{(1)}\bar{e}_1^{(1)} + \alpha_2^{(1)}\bar{e}_2^{(1)} + \ \ldots \ + \alpha_{k_1}^{(1)}\bar{e}_{k_1}^{(1)} + \alpha_1^{(2)}\bar{e}_1^{(2)} + \alpha_2^{(2)}\bar{e}_2^{(2)} + \ \ldots \\ \ldots + \alpha_{k_2}^{(2)}\bar{e}_{k_2}^{(2)} + \ \ldots \ + \alpha_1^{(m)}\bar{e}_1^{(m)} + \alpha_2^{(m)}\bar{e}_2^{(m)} + \ \ldots \\ \ldots + \alpha_{k_m}^{(m)}\bar{e}_{k_m}^{(m)} = \bar{v}_1 + \bar{v}_2 + \ \ldots \ + \bar{v}_m. \end{aligned} \tag{89}$$

Здесь $\bar{v}_i$ принадлежат инвариантному для матрицы A линейному многообразию, построенному на векторах $\bar{e}_1^{(i)}, \bar{e}_2^{(i)}, \ldots, \bar{e}_k^{(i)}$.

В силу равенств (88) при $p > k_i$ и $k_i \geqslant j > 1$ будем иметь:

$$A^p\bar{e}_j^{(i)} = A^{p-1}\left(\lambda_i\bar{e}_j^{(i)} + \bar{e}_{j-1}^{(i)}\right) = A^{p-2}\left(\lambda_i^2\bar{e}_j^{(i)} + 2\lambda_i\bar{e}_{j-1}^{(i)} + \bar{e}_{j-2}^{(i)}\right) =$$
$$= A^{p-3}\left(\lambda_i^3\bar{e}_j^{(i)} + 3\lambda_i^2\bar{e}_{j-1}^{(i)} + 3\lambda_i\bar{e}_{j-2}^{(i)} + \bar{e}_{j-3}^{(i)}\right) = \ldots$$
$$\ldots = \lambda_i^p\bar{e}_j^{(i)} + \lambda_i^{p-1}C_p^1\bar{e}_{j-1}^{(i)} + \lambda_i^{p-2}C_p^2\bar{e}_{j-2}^{(i)} + \ldots + C_p^{j-1}\lambda_i^{p-j+1}e_1^{(j)}. \quad (90)$$

При $j = 1$ получим:

$$A^p\bar{e}_1^{(i)} = \lambda_i^p\bar{e}_1^{(i)}. \quad (91)$$

Таким образом,

$$A^p\bar{v}_i = \lambda_i^p\left(\alpha_1^{(i)}\bar{e}_1^{(i)} + \alpha_2^{(i)}\bar{e}_2^{(i)} + \ldots + \alpha_{k_i}^{(i)}\bar{e}_{k_i}^{(i)}\right) + \lambda_i^{p-1}C_p^1\left(\alpha_2^{(i)}\bar{e}_1^{(i)} +\right.$$
$$+ \alpha_3^{(i)}\bar{e}_2^{(i)} + \ldots + \alpha_{k_i}^{(i)}\bar{e}_{k_i-1}^{(i)}) + \lambda_1^{p-2}C_p^2\left(\alpha_3^{(i)}\bar{e}_1^{(i)} + \alpha_4^{(i)}\bar{e}_2^{(i)} + \ldots\right.$$
$$\left.\ldots + \alpha_{k_i}^{(i)}\bar{e}_{k_i-2}^{(i)}\right) + \ldots + \lambda_i^{p-k_i+1}C_p^{k_i-1}e_1^{-(i)}. \quad (92)$$

За норму вектора $\bar{v}$ примем величину

$$\|\bar{v}\| = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{k_i}|\alpha_j^{(i)}|^2}. \quad (93)$$

Тогда из (92) следует:

$$\lim_{p\to\infty}\frac{\left\|A^p\bar{v}_i - \lambda_i^{p-k_i+1}C_p^{k_i-1}\alpha_{k_i}^{(i)}e_1^{(i)}\right\|}{\|A^p\bar{v}_i\|} = 0. \quad (94)$$

Таким образом, с увеличением p вектор $A^p\bar{v}_i$ будет все больше и больше приближаться по направлению к вектору $\bar{e}_1^{(i)}$, если только $\alpha_{k_i}^{(i)} \neq 0$. Аналогичная картина будет наблюдаться и с остальными компонентами $\bar{v}_j$ вектора $\bar{v}$.

Из только что полученного результата можно сделать следующие заключения. При возрастании p вектор $A^p\bar{v}$ будет неограниченно приближаться к инвариантному многообразию, порожденному векторами (87), соответствующими наибольшим по модулю собственным значениям. Если λ_1 — единственное наибольшее по модулю собственное значение кратности единица, то будут справедливы те же выводы, которые мы делали для матриц простой структуры. Если λ_1 — единственное наибольшее по модулю собственное значение и кратность его больше единицы, то поведение вектора $A^p\bar{v}$ будет определяться элементарными делителями матрицы $A - \lambda I$, соответствующими собственному значению λ_1. Если все они простые, то исследование проводится так же, как и для матрицы простой структуры. Если имеется

элементарный делитель $(\lambda - \lambda_1)^k$ с наибольшим k, то при больших p вектор $A^p\bar{v}$ будет находиться в инвариантном многообразии, порожденном некоторыми векторами $e_1^{(s)}, e_2^{(s)}, \ldots, e_k^{(s)}$ системы (87), соответствующими λ_1. Все векторы $\bar{u}$ этого многообразия будут обладать свойством $(A - \lambda_1 I)^k \bar{u} = 0$ или

$$A^k\bar{u} - C_k^1\lambda_1 A^{k-1}\bar{u} + C_k^2\lambda_1^2 A^{k-2}\bar{u} + \ldots + (-1)^k\lambda_1^k\bar{u} = 0. \quad (95)$$

Поэтому при достаточно больших p векторы $A^p\bar{v}$, $A^{p+1}\bar{v}$, ..., $A^{p+k}\bar{v}$ будут связаны линейной зависимостью вида (95). При этом λ_1 находится из уравнения

$$\lambda^k - C_k^1\lambda_1\lambda^{k-1} + C_k^2\lambda_1^2\lambda^{k-2} - \ldots + (-1)^k\lambda_1^k = (\lambda - \lambda_1)^k = 0. \quad (96)$$

Если имеется несколько элементарных делителей наивысшей степени, то в характере поведения векторов $A^p\bar{v}$ изменения не произойдет, но вместо уравнения (96) получим уравнение $(\lambda - \lambda_1)^{kl} = 0$, где l — число соответствующих элементарных делителей.

Аналогично исследуется случай, когда имеется несколько различных максимальных по модулю собственных значений. В каждом таком случае при достаточно большом p векторы $A^p\bar{v}$, $A^{p+1}\bar{v}$, ..., $A^{p+r}\bar{v}$ будут связаны некоторой линейной зависимостью вида

$$A^{p+r}\bar{v} + \beta_1 A^{p+r-1}\bar{v} + \ldots + \beta_{r-1}A^{p+1}\bar{v} + \beta_r A^p\bar{v} = 0. \quad (97)$$

При этом максимальные по модулю значения λ_i находятся из уравнения

$$\lambda^r + \beta_1\lambda^{r-1} + \ldots + \beta_{r-1}\lambda + \beta_r = 0. \quad (98)$$

Так как подробный разбор всех возможных случаев потребовал бы длинных рассуждений и так как матрицы, не имеющие простой структуры, встречаются сравнительно редко в вычислительной практике, мы этим заниматься не будем. Желающих изучить эти вопросы подробнее отсылаем к специальной литературе (см., например, К. А. Семендяев, О нахождении собственных значений и инвариантных многообразий матриц посредством итераций, ПММ, т. 7, 3, 1943, стр. 193—222).

§ 8. Ускорение сходимости итерационных процессов при решении задач линейной алгебры

В шестой главе были рассмотрены итерационные методы решения систем линейных алгебраических уравнений и в предыдущем параграфе рассмотрены итерационные методы отыскания собственных значений матриц. На практике часто случается, что при применении этих методов нужно произвести очень много итераций для получения

результата с требуемой точностью. В связи с этим возникает необходимость применять те или иные методы ускорения сходимости. В этом параграфе мы и рассмотрим некоторые из методов ускорения сходимости для указанных выше задач.

1. Ускорение сходимости итерационного метода решения систем линейных алгебраических уравнений. Общие замечания. Стационарный итерационный метод решения системы линейных алгебраических уравнений

$$A\bar{x} = \bar{b} \tag{1}$$

можно записать в виде

$$\bar{x}_{k+1} = B\bar{x}_k + C\bar{b}, \tag{2}$$

где B и C — такие матрицы, что

$$B + CA = I \tag{3}$$

(см. главу 6). Из (2) следует

$$\bar{x}_{k+1} - \bar{x} = B(\bar{x}_k - \bar{x}) \tag{4}$$

или

$$\bar{x}_{k+1} - \bar{x} = B^{k+1}(\bar{x}_0 - \bar{x}). \tag{5}$$

Обозначим максимум модуля собственных значений матрицы B через λ. Пусть p — наивысшая степень элементарного делителя, соответствующего наибольшему по модулю собственному значению. Тогда из формулы (94) предыдущего параграфа следует, что при больших значениях k будет иметь место асимптотическое равенство

$$\|\bar{x}_{k+1} - \bar{x}\| \sim C_{k+1}^{p-1}\lambda^{k+2-p}\|\bar{x}_0 - \bar{x}\|. \tag{6}$$

В связи с этим будем численно характеризовать скорость сходимости итерационного процесса (2) величиной

$$R(B) = -\log\lambda. \tag{7}$$

При этом скорость сходимости будет примерно обратно пропорциональна числу итераций, необходимых для получения решения с заданной точностью.

Систему (1) можно бесчисленным множеством способов привести к виду (2). В силу изложенного это нужно делать так, чтобы максимум модуля собственных значений матрицы B был возможно меньшим. Рассмотрим случай, когда $B = f(A)$ и $C = g(A)$, где f и g — некоторые многочлены. Условие (3) в данном случае примет вид

$$f(A) + g(A)A = I. \tag{8}$$

Если обозначить собственные значения матрицы A через λ_i $(i = 1, 2, \ldots, n)$, то для сходимости итерационного процесса (2) нужно потребовать

$$|f(\lambda_i)| = |1 - g(\lambda_i)\lambda_i| < 1. \tag{9}$$

Предположим теперь, что все собственные значения матрицы A действительны и расположены на отрезке $[m, M]$, где $m > 0$. Постараемся использовать эту информацию для наилучшего выбора многочлена $f(\lambda)$. Прежде всего в силу (8) для $f(\lambda)$ должно быть выполнено условие

$$f(0) = 1. \tag{10}$$

Многочлен $f(\lambda)$ должен иметь наименьший максимум модуля на отрезке $[m, M]$ среди всех многочленов данной степени s, удовлетворяющих условию (10). Мы уже решили такую задачу в главе 6. Искомый многочлен будет равен

$$f(\lambda) = \frac{T_s\left(\frac{M+m-2\lambda}{M-m}\right)}{T_s\left(\frac{M+m}{M-m}\right)}, \tag{11}$$

где $T_s(x)$ — многочлены Чебышева, наименее уклоняющиеся от нуля. При этом

$$\max_{\lambda \in [m, M]} |f(\lambda)| = \frac{1}{\left|T_s\left(\frac{M+m}{M-m}\right)\right|} < 1 \tag{12}$$

и сходимость обязательно будет иметь место. В простейшем случае можно взять в качестве $f(\lambda)$ многочлен первой степени. Он должен иметь вид

$$f(\lambda) = 1 - C\lambda. \tag{13}$$

При этом

$$C = g(\lambda) = \frac{2}{M+m}. \tag{14}$$

2. Метод М. К. Гавурина. Можно подойти к вопросу об ускорении сходимости и несколько иначе. Пусть итерационная формула (2) уже построена. Начиная с некоторого $\bar{x}_0$, построим по формуле (2) векторы $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{s+1}$. Поставим следующую задачу: подобрать коэффициенты α_i так, чтобы линейная комбинация

$$\bar{y} = \bar{x}_0 + \alpha_0(\bar{x}_1 - \bar{x}_0) + \ldots + \alpha_s(\bar{x}_{s+1} - \bar{x}_s) \tag{15}$$

возможно лучше приближала точное решение системы (1). Будем предполагать, что матрица B имеет простую структуру и что все ее собственные значения действительны. Обозначим собственные векторы B через $\bar{z}_1, \bar{z}_2, \ldots, \bar{z}_n$ и запишем разложение $\bar{x}_1 - \bar{x}_0$ по векторам $\bar{z}_i$ в виде

$$\bar{x}_1 - \bar{x}_0 = c_1\bar{z}_1 + c_2\bar{z}_2 + \ldots + c_n\bar{z}_n. \tag{16}$$

При этом

$$\bar{x}_{k+1} - \bar{x}_k = c_1\lambda_1^k\bar{z}_1 + c_2\lambda_2^k\bar{z}_2 + \ldots + c_n\lambda_n^k\bar{z}_n. \tag{17}$$

Точное решение $\bar{x}$ может быть записано в виде

$$\bar{x} = \bar{x}_0 + \sum_{k=0} (\bar{x}_{k+1} - \bar{x}_k) = \bar{x}_0 + \sum_{k=0}^{\infty} (c_1 \lambda_1^k \bar{z}_1 + c_2 \lambda_2^k \bar{z}_2 + \ldots + c_n \lambda_n^k \bar{z}_n) =$$

$$= \bar{x}_0 + \frac{c_1}{1-\lambda_1} \bar{z}_1 + \frac{c_2}{1-\lambda_2} \bar{z}_2 + \ldots + \frac{c_n}{1-\lambda_n} \bar{z}_n. \quad (18)$$

Таким образом, разность между точным решением (18) и приближенным (15) представится так:

$$\bar{x} - \bar{y} = \sum_{i=1}^{n} c_i \left[\frac{1}{1-\lambda_i} - P(\lambda_i) \right] \bar{z}_i, \quad (19)$$

где через $P(\lambda)$ обозначен многочлен

$$P(\lambda) = \alpha_0 + \alpha_1 \lambda + \ldots + \alpha_s \lambda^s. \quad (20)$$

Очевидно, наша задача сводится к тому, чтобы сделать величины, стоящие в квадратных скобках (19), возможно меньшими. Пусть нам известна верхняя граница $M < 1$ модулей собственных значений B. Тогда надо выбрать многочлен $P(\lambda)$ так, чтобы он на отрезке $[-M, M]$ наилучшим образом аппроксимировал функцию $\frac{1}{1-\lambda}$. Решение этой задачи также сводится к многочленам Чебышева, наименее уклоняющимся от нуля. Можно показать (см. упражнения к главе 4), что

$$P(\lambda) = \frac{2\alpha^{s+1}}{(1-\alpha^2)^2} \frac{1}{\lambda - 1} \left[T_{s+1}\left(\frac{\lambda}{M}\right) - 2\alpha T_s\left(\frac{\lambda}{M}\right) + \alpha^2 T_{s-1}\left(\frac{\lambda}{M}\right) \right] + \frac{1}{\lambda - 1}, \quad (21)$$

где

$$\alpha = \frac{1}{M} - \sqrt{\frac{1}{M^2} - 1}. \quad (22)$$

Применяя этот метод, предложенный М. К. Гавуриным, мы можем производить несколько итераций по формуле (2) и затем улучшать полученные результаты, используя многочлен $P(\lambda)$.

3. Метод Л. А. Люстерника. Пусть теперь нам известно (приближенно) наибольшее по модулю собственное значение λ_1^* матрицы B. Будем сначала предполагать, что оно действительно, единственно и что матрица B имеет простую структуру. Разложим $\bar{x}_1 - \bar{x}_0$ по собственным векторам матрицы B:

$$\bar{x}_1 - \bar{x}_0 = d_1 \bar{z}_1 + d_2 \bar{z}_2 + \ldots + d_n \bar{z}_n. \quad (23)$$

Тогда

$$\bar{x}_{m+1} - \bar{x}_m = B^m (\bar{x}_1 - \bar{x}_0) = d_1 \lambda_1^m \bar{z}_1 + d_2 \lambda_2^m \bar{z}_2 + \ldots + d_n \lambda_n^m \bar{z}_n \quad (24)$$

и

$$\bar{x} - \bar{x}_m = \sum_{k=0}^{\infty} (\bar{x}_{m+k+1} - \bar{x}_{m+k}) =$$

$$= \frac{\lambda_1^m}{1-\lambda_1} d_1 \bar{z}_1 + \frac{\lambda_2^m}{1-\lambda_2} d_2 \bar{z}_2 + \ldots + \frac{\lambda_n^m}{1-\lambda_n} d_n \bar{z}_n. \quad (25)$$

При достаточно большом m будем иметь приближенные равенства

$$\bar{x} - \bar{x}_m \approx \frac{\lambda_1^m}{1-\lambda_1} d_1 \bar{z}_1, \quad (26)$$

$$\bar{x}_{m+1} - \bar{x}_m \approx \lambda_1^m d_1 \bar{z}_1. \quad (27)$$

Таким образом, можно ожидать, что вектор

$$\bar{x}' = \bar{x}_m + \frac{1}{1-\lambda_1^*} (\bar{x}_{m+1} - \bar{x}_m) \quad (28)$$

будет ближе к точному решению $\bar{x}$, чем $\bar{x}_m$ и $\bar{x}_{m+1}$. Оценим порядок ошибки $\bar{x} - \bar{x}'$, предполагая, что $\lambda_1 = \lambda_1^* + \varepsilon$, где $\varepsilon = 0\left(\left|\frac{\lambda_2}{\lambda_1}\right|^m\right)$, а λ_2 — следующее за λ_1 по величине модуля собственное значение B. Если ввести матрицу

$$B_1 = \frac{1}{1-\lambda_1^*} [B - \lambda_1^* I], \quad (29)$$

то, с одной стороны, $B_1(\bar{x} - \bar{x}_m) = \bar{x} - \bar{x}'$, а с другой стороны, используя (25), имеем:

$$B_1(\bar{x} - \bar{x}_m) =$$

$$= \frac{1}{1-\lambda_1^*} \left[\frac{\lambda_1^m \varepsilon}{1-\lambda_1} d_1 \bar{z}_1 + \frac{\lambda_2^m (\lambda_2 - \lambda_1^*)}{1-\lambda_2} d_2 \bar{z}_2 + \ldots + \frac{\lambda_n^m (\lambda_n - \lambda_1^*)}{1-\lambda_n} d_n \bar{z}_n \right],$$

откуда получим:

$$\bar{x} - \bar{x}' =$$

$$= \frac{1}{1-\lambda_1^*} \left[\frac{\lambda_1^m \varepsilon}{1-\lambda_1} d_1 \bar{z}_1 + \frac{\lambda_2^m (\lambda_2 - \lambda_1^*)}{1-\lambda_2} d_2 \bar{z}_2 + \ldots + \frac{\lambda_n^m (\lambda_n - \lambda_1^*)}{1-\lambda_n} d_n \bar{z}_n \right]. \quad (30)$$

Поэтому

$$\bar{x} - \bar{x}' = 0(\lambda_2^m). \quad (31)$$

Так как по (27) порядок $\bar{x} - \bar{x}_m$ равен $0(\lambda_1^m)$, то улучшение сходимости будет тем больше, чем меньше отношение $\left|\frac{\lambda_2}{\lambda_1}\right|$. Если

λ_1 близко к 1, то целесообразно вместо формулы (28) использовать

$$\bar{x}' = \bar{x}_m + \frac{1}{1-\lambda_1^{*p}}(\bar{x}_{m+p} - \bar{x}_m) \tag{32}$$

Вывод этой формулы аналогичен выводу формулы (28). Метод применим и в том случае, когда λ_1 — кратное собственное значение. Пусть теперь λ_1 — комплексное число. Если матрица B действительна, то существует и комплексно-сопряженное собственное значение $\bar{\lambda}_1$. Обозначим соответствующие этим собственным значениям комплексно-сопряженные собственные векторы через $\bar{z}_1$ и $\bar{z}_2$. Тогда при больших значениях m будем иметь приближенное равенство

$$\bar{x} - \bar{x}_m \approx \frac{\lambda_1^m}{1-\lambda_1} d_1\bar{z}_1 + \frac{\bar{\lambda}_1^m}{1-\bar{\lambda}_1} d_2\bar{z}_2. \tag{33}$$

Далее, из

$$\left.\begin{aligned} \bar{x}_{m+1} - \bar{x}_m &\approx \lambda_1^m d_1\bar{z}_1 + \bar{\lambda}_1^m d_2\bar{z}_2, \\ \bar{x}_{m+2} - \bar{x}_{m+1} &\approx \lambda_1^{m+1} d_1\bar{z}_1 + \bar{\lambda}_1^{m+1} d_2\bar{z}_2 \end{aligned}\right\} \tag{34}$$

находим:

$$(\lambda_1 + \bar{\lambda}_1)(\bar{x}_{m+1} - \bar{x}_m) - (\bar{x}_{m+2} - \bar{x}_{m+1}) \approx \lambda_1^m\bar{\lambda}_1 d_1\bar{z}_1 + \bar{\lambda}_1^m\lambda_1 d_2\bar{z}_2. \tag{35}$$

Таким образом, можно взять

$$\bar{x}' = \bar{x}_m + \frac{(\bar{x}_{m+1} - \bar{x}_m)(1 - \lambda_1 - \bar{\lambda}_1) + (\bar{x}_{m+2} - \bar{x}_{m+1})}{1 - (\lambda_1 + \bar{\lambda}_1) + \lambda_1\bar{\lambda}_1}. \tag{36}$$

Величины $\lambda_1 + \bar{\lambda}_1$ и $\lambda_1\bar{\lambda}_1$ можно, например, найти как коэффициенты квадратного уравнения, о котором говорилось в предыдущем параграфе. Можно использовать и два наибольших по модулю собственных значения. При этом будет применима формула, аналогичная (36). Приведенный метод был предложен Л. А. Люстерником.

4. δ^2-процесс Эйткена. Можно использовать равенство (26) и иначе. Запишем его для компонент $x_i - x_i^{(m)}$ вектора ошибки $\bar{x} - \bar{x}_m$ в виде

$$x_i - x_i^{(m)} = \alpha_1\lambda_1^m + \alpha_2\lambda_2^m + \ldots + \alpha_n\lambda_n^m. \tag{37}$$

Опять предположим, что λ_1 преобладает по модулю над остальными собственными значениями. Тогда при большом m можно приближенно считать:

$$\left.\begin{aligned} x_i - x_i^{(m)} &= \alpha_1\lambda_1^m, \\ x_i - x_i^{(m+1)} &= \alpha_1\lambda_1^{m+1}, \\ x_i - x_i^{(m+2)} &= \alpha_1\lambda_1^{m+2}. \end{aligned}\right\} \tag{38}$$

Исключим из этих трех равенств величины α_1 и λ_1. Обозначая

$$\left.\begin{aligned} x_i^{(m+1)} - x_i^{(m)} &= \Delta x_i^{(m)}, \\ \Delta x_i^{(m+1)} - \Delta x_i^{(m)} &= \Delta^2 x_i^{(m)}, \\ \cdots\cdots\cdots\cdots &, \end{aligned}\right\} \tag{39}$$

получим:

$$\left.\begin{aligned} \Delta x_i^{(m)} &= \alpha_1 \lambda_1^m (1 - \lambda_1), \\ \Delta x_i^{(m+1)} &= \alpha_1 \lambda_1^{m+1} (1 - \lambda_1), \\ \Delta^2 x_i^{(m)} &= -\alpha_1 \lambda_1^m (1 - 2\lambda_1 + \lambda_1^2). \end{aligned}\right\} \tag{40}$$

Следовательно,

$$\alpha_1 \lambda_1^m = -\frac{\left(\Delta x_i^{(m)}\right)^2}{\Delta^2 x_i^{(m)}} \tag{41}$$

и

$$x_i = x_i^{(m)} - \frac{\left(\Delta x_i^{(m)}\right)^2}{\Delta^2 x_i^{(m)}} = \frac{x_i^{(m)} x_i^{(m+2)} - \left(x_i^{(m+1)}\right)^2}{x_i^{(m+2)} - 2x_i^{(m+1)} + x_i^{(m)}}. \tag{42}$$

Получили δ^2-*процесс Эйткена*, о котором говорилось в предыдущей главе. Указанный способ можно обобщить на случай, когда преобладающими считаются два или больше собственных значения. Так, пусть

$$x_i - x_i^{(m)} = \alpha_1 \lambda_1^m + \alpha_2 \lambda_2^m. \tag{43}$$

Тогда

$$\left.\begin{aligned} \Delta x_i^{(m)} &= \alpha_1 \lambda_1^m (1 - \lambda_1) + \alpha_2 \lambda_2^m (1 - \lambda_2), \\ \Delta^2 x_i^{(m)} &= -\alpha_1 \lambda_1^m (1 - \lambda_1)^2 - \alpha_2 \lambda_2^m (1 - \lambda_2)^2, \\ \Delta^3 x_i^{(m)} &= \alpha_1 \lambda_1^m (1 - \lambda_1)^3 + \alpha_2 \lambda_2^m (1 - \lambda_2)^3, \\ \Delta^4 x_i^{(m)} &= -\alpha_1 \lambda_1^m (1 - \lambda_1)^4 - \alpha_2 \lambda_2^m (1 - \lambda_2)^4. \end{aligned}\right\} \tag{44}$$

Используя выражения (43) и (44), нетрудно показать, что

$$\begin{vmatrix} x_i - x_i^{(m)} & \Delta x_i^{(m)} & \Delta^2 x_i^{(m)} \\ \Delta x_i^{(m)} & \Delta^2 x_i^{(m)} & \Delta^3 x_i^{(m)} \\ \Delta^2 x_i^{(m)} & \Delta^3 x_i^{(m)} & \Delta^4 x_i^{(m)} \end{vmatrix} = 0. \tag{45}$$

Таким образом,

$$x_i = \frac{\begin{vmatrix} x_i^{(m)} & \Delta x_i^{(m)} & \Delta^2 x_i^{(m)} \\ \Delta x_i^{(m)} & \Delta^2 x_i^{(m)} & \Delta^3 x_i^{(m)} \\ \Delta^2 x_i^{(m)} & \Delta^3 x_i^{(m)} & \Delta^4 x_i^{(m)} \end{vmatrix}}{\begin{vmatrix} \Delta^2 x_i^{(m)} & \Delta^3 x_i^{(m)} \\ \Delta^3 x_i^{(m)} & \Delta^4 x_i^{(m)} \end{vmatrix}}. \tag{46}$$

Аналогичные формулы мы получим, если будем считать преобладающим 3, 4 и т. д. собственные значения.

Формулы (42) и (46) находят применение для улучшения сходимости самых различных процессов.

5. Улучшение сходимости итерационных процессов для отыскания собственных значений матриц. Кратко коснемся вопроса об улучшении сходимости итерационных процессов для отыскания собственных значений матриц. И в этом случае мы находим некоторую последовательность чисел вида

$$x_k = \alpha_1\lambda_1^k + \alpha_2\lambda_2^k + \ldots + \alpha_n\lambda_n^k \qquad (k = 0,\ 1,\ 2,\ \ldots). \tag{47}$$

Преобладающее по модулю собственное значение, определяемое с помощью этого равенства, будет найдено тем точнее, чем меньше отношение $|\lambda_2/\lambda_1|$. Поэтому если нам известны приближенно λ_1^* и λ_2^*, то можно вместо последовательности (47) рассматривать последовательность

$$y_p = \sum_{k=0}^{p} c_k x_k = \sum_{i=1}^{n} \alpha_i \sum_{k=0}^{p} c_k \lambda_i^k = \sum_{i=1}^{n} \alpha_i P_p(\lambda_i), \tag{48}$$

где

$$P_p(\lambda) = \frac{1}{T_p\left(\dfrac{\lambda_1^*}{\lambda_2^*}\right)} T_p\left(\frac{\lambda}{\lambda_2^*}\right). \tag{49}$$

Если известно грубо положение всех собственных значений $\lambda_1^*, \lambda_2^*, \ldots, \lambda_n^*$, то для уточнения собственного значения λ_i^* можно использовать итерацию с матрицей $P_i(A)$, где

$$P_i(A) = (A - \lambda_1^* I)(A - \lambda_2^* I) \ldots (A - \lambda_{i-1}^* I)(A - \lambda_{i+1}^* I) \ldots (A - \lambda_n^* I). \tag{50}$$

§ 9. Неустранимая погрешность при численном решении систем линейных алгебраических уравнений

Пусть требуется решить систему линейных алгебраических уравнений

$$A\bar{x} = \bar{b}. \tag{1}$$

На практике часто бывает так, что мы не знаем точно ни матрицы A, ни вектора $\bar{b}$. Предположим, что вместо матрицы A нам дана матрица A^* и вместо вектора $\bar{b}$ дан вектор $\bar{b}^*$. Тогда вместо вектора $\bar{x}$ мы сумеем найти только вектор $\bar{x}^*$, удовлетворяющий системе уравнений

$$A^*\bar{x}^* = \bar{b}^*. \tag{2}$$

При этом возникает задача об оценке отклонения вектора x^* от вектора $\bar{x}$, т. е. об оценке $\|\bar{x}-\bar{x}^*\|$, где норма понимается в том или ином смысле.

Аналогичная задача может возникнуть и в том случае, когда матрица A и вектор $\bar{b}$ известны точно, но в процессе решения системы (1) были допущены те или иные погрешности. Тогда вместо точного вектора решения $\bar{x}$ мы получим некоторый приближенный вектор $\bar{x}^*$. Вычисляя $A\bar{x}^*$, мы получим не $\bar{b}$, а $\bar{b}^*$. Опять возникает необходимость оценить $\|\bar{x}-\bar{x}^*\|$ по известной разности $\bar{b}-\bar{b}^*$. Так как вторая задача есть частный случай первой, то мы ограничимся первым случаем.

Пусть

$$A^*=A-\varepsilon;\quad \bar{b}^*=\bar{b}-\bar{\delta}. \tag{3}$$

Тогда

$$(A^*+\varepsilon)(\bar{x}-\bar{x}^*)=(A^*+\varepsilon)\bar{x}-A^*\bar{x}^*-\varepsilon\bar{x}^*=\bar{\delta}-\varepsilon\bar{x}^* \tag{4}$$

или

$$\|(A^*+\varepsilon)(\bar{x}-\bar{x}^*)\|\leqslant\|\bar{\delta}\|+\|\varepsilon\|\,\|\bar{x}^*\|. \tag{5}$$

Определяя различным образом нормы векторов, мы получим различные оценки. Воспользуемся сначала третьей нормой векторов, введенной в шестой главе:

$$\|\bar{r}\|_3=\sqrt{(\bar{r},\bar{r})}. \tag{6}$$

Тогда согласованная с ней норма матрицы A будет равна

$$\|A\|_3=\sqrt{\lambda_{\max}^{A'A}}, \tag{7}$$

где через $\lambda_{\max}^{A'A}$ обозначено наибольшее собственное значение симметрической матрицы $A'A$. Из равенства (2) получаем

$$\|A^*\bar{x}^*\|_3=\|\bar{b}^*\|_3. \tag{8}$$

Но

$$\|A^*\bar{x}^*\|_3^2=(A^*\bar{x}^*,\ A^*\bar{x}^*)=(\bar{x}^*,\ A^{*\prime}A^*\bar{x}^*). \tag{9}$$

Обозначим собственные значения матрицы $A^{*\prime}A^*$ через $\lambda_i^{A^{*\prime}A^*}$. Так как мы предположили, что матрица A^* невырожденная, а матрица $A^{*\prime}A^*$ является симметрической, то все $\lambda_i^{A^{*\prime}A^*}$ положительны. Им соответствует ортонормированная система собственных векторов $\bar{u}_1,\ \bar{u}_2,\ \ldots,\ \bar{u}_n$, образующая базис нашего пространства. Пусть

$$\bar{x}^*=c_1\bar{u}_1+c_2\bar{u}_2+\ \ldots\ +c_n\bar{u}_n. \tag{10}$$

Тогда из (9) и (10) получаем:

$$\|A^*\bar{x}^*\|_3^2=c_1^2\lambda_1^{A^{*\prime}A^*}+c_2^2\lambda_2^{A^{*\prime}A^*}+\ \ldots\ +c_n^2\lambda_n^{A^{*\prime}A^*}. \tag{11}$$

Обозначим минимальное из $\lambda_i^{A^{*\prime}A^*}$ через $\lambda_{\min}^{A^{*\prime}A^*}$. Из (11) следует:

$$\|A^*\bar{x}^*\|_3^2 \geqslant \lambda_{\min}^{A^{*\prime}A^*}\left(c_1^2+c_2^2+\ldots+c_n^2\right)=\lambda_{\min}^{A^{*\prime}A^*}\|\bar{x}^*\|_3^2. \tag{12}$$

Таким образом,

$$\|A^*\bar{x}^*\|_3 \geqslant \sqrt{\lambda_{\min}^{A^{*\prime}A^*}}\,\|\bar{x}^*\|_3. \tag{13}$$

Из (8) и (13) находим:

$$\sqrt{\lambda_{\min}^{A^{*\prime}A^*}}\,\|\bar{x}^*\|_3 \leqslant \|\bar{b}^*\|_3 \tag{14}$$

или

$$\|\bar{x}^*\|_3 \leqslant \frac{\|\bar{b}^*\|_3}{\left(\lambda_{\min}^{A^{*\prime}A^*}\right)^{\frac{1}{2}}}. \tag{15}$$

Поэтому второе слагаемое правой части неравенства (5) может быть оценено следующим образом:

$$\|\varepsilon\|_3\,\|\bar{x}^*\|_3 \leqslant \sqrt{\frac{\lambda_{\max}^{\varepsilon'\varepsilon}}{\lambda_{\min}^{A^{*\prime}A^*}}}\,\|\bar{b}^*\|_3. \tag{16}$$

Очевидно,

$$\|(A^*+\varepsilon)(\bar{x}-\bar{x}^*)\|_3 \geqslant \|A^*(\bar{x}-\bar{x}^*)\|_3 - \|\varepsilon(\bar{x}-\bar{x}^*)\|_3. \tag{17}$$

Уменьшаемое в правой части неравенства (17) оценивается так же, как мы оценивали (8). При этом получим:

$$\|A^*(\bar{x}-\bar{x}^*)\|_3 \geqslant \sqrt{\lambda_{\min}^{A^{*\prime}A^*}}\,\|\bar{x}-\bar{x}^*\|_3. \tag{18}$$

Вычитаемое в правой части (17) может быть оценено так:

$$\|\varepsilon(\bar{x}-\bar{x}^*)\|_3 \leqslant \|\varepsilon\|_3\,\|\bar{x}-\bar{x}^*\|_3 = \sqrt{\lambda_{\max}^{\varepsilon'\varepsilon}}\,\|\bar{x}-\bar{x}^*\|_3. \tag{19}$$

Таким образом,

$$\|(A^*+\varepsilon)(\bar{x}-\bar{x}^*)\|_3 \geqslant \left(\sqrt{\lambda_{\min}^{A^{*\prime}A^*}}-\sqrt{\lambda_{\max}^{\varepsilon'\varepsilon}}\right)\|\bar{x}-\bar{x}^*\|_3. \tag{20}$$

Если $\left(\lambda_{\min}^{A^{*\prime}A^*}\right)^{\frac{1}{2}}-\left(\lambda_{\max}^{\varepsilon'\varepsilon}\right)^{\frac{1}{2}}>0$, то (5) и (20) дают

$$\|\bar{x}-\bar{x}^*\|_3 \leqslant \frac{\|\bar{\delta}\|_3+\left(\lambda_{\max}^{\varepsilon'\varepsilon}/\lambda_{\min}^{A^{*\prime}A^*}\right)^{\frac{1}{2}}\|\bar{b}^*\|_3}{\left(\lambda_{\min}^{A^{*\prime}A^*}\right)^{\frac{1}{2}}-\left(\lambda_{\max}^{\varepsilon'\varepsilon}\right)^{\frac{1}{2}}}. \tag{21}$$

Это и есть искомая оценка.

Приведем пример на применение полученной формулы. Пусть дана система:

$$\begin{aligned}
26{,}7x_1 - 5{,}1x_2 + 1{,}3x_3 - 0{,}3x_4 &= 43{,}888,\\
-5{,}1x_1 - 28{,}6x_2 + 6{,}2x_3 + 1{,}2x_4 &= -58{,}203,\\
1{,}3x_1 + 6{,}2x_2 + 29{,}1x_3 - 4{,}3x_4 &= 88{,}390,\\
-0{,}3x_1 + 1{,}2x_2 - 4{,}3x_3 + 27{,}1x_4 &= 98{,}664.
\end{aligned}$$

Будем предполагать, что коэффициенты и правые части системы заданы точно, но в процессе решения системы вследствие ошибок округления мы получили не точное решение, а приближенное:

$$\bar{x}_1^* = (2;\ 2{,}5;\ 3;\ 4).$$

В этом случае вектор $\bar{\delta}$ будет выглядеть так:

$$\bar{\delta} = (0{,}538;\ -0{,}03;\ 0{,}190;\ 0{,}764).$$

Итак, $\|\bar{\delta}\|_3 = 0{,}95$. Матрица ε в нашем случае нулевая. Поэтому

$$\|\bar{x} - \bar{x}^*\|_3 \leqslant \frac{\|\bar{\delta}\|_3}{\left(\lambda_{\min}^{A^{*\prime}A^*}\right)^{\frac{1}{2}}}.$$

Так как матрица A симметрическая, то $\left(\lambda_{\min}^{A^{*\prime}A^*}\right)^{\frac{1}{2}} = |\lambda^A|_{\min}$. Для $|\lambda^A|_{\min}$ нетрудно получить грубую оценку $|\lambda^A|_{\min} \geqslant 16{,}1$. Таким образом,

$$\|\bar{x} - \bar{x}^*\|_3 \leqslant \frac{0{,}95}{16{,}1} \approx 0{,}06.$$

Возьмем теперь в качестве нормы вектора $\bar{r} = (r_1,\ r_2,\ \ldots,\ r_n)$ величину

$$\|\bar{r}\|_2 = \sum_{i=1}^{n} |r_i|. \tag{22}$$

Тогда согласованная с ней норма матрицы A с элементами a_{ij} будет равна

$$\|A\|_2 = \max_j \sum_{i=1}^{n} |a_{ij}|. \tag{23}$$

Перепишем равенство (4) в виде

$$\bar{\xi} = A^{*-1}\bar{\delta} - A^{*-1}\varepsilon\bar{x}^* - A^{*-1}\varepsilon\bar{\xi}, \tag{24}$$

где $\xi = \bar{x} - \bar{x}^*$. Отсюда следует:

$$\|\bar{\xi}\|_2 \leqslant \|A^{*-1}\bar{\delta}\|_2 + \|A^{*-1}\varepsilon\|_2\|\bar{x}^*\|_2 + \|A^{*-1}\varepsilon\|_2\|\bar{\xi}\|_2. \tag{25}$$

Будем предполагать, что погрешности ε_{ik} настолько малы, что $\|A^{*-1}\varepsilon\|_2 < 1$. Тогда

$$\|\bar{\xi}\|_2 \leqslant \frac{\|A^{*-1}\bar{\delta}\|_2 + \|A^{*-1}\varepsilon\|_2\|\bar{x}^*\|_2}{1 - \|A^{*-1}\varepsilon\|_2} \leqslant$$

$$\leqslant \frac{\|A^{*-1}\|_2\|\bar{\delta}\|_2 + \|A^{*-1}\varepsilon\|_2\|\bar{x}^*\|_2}{1 - \|A^{*-1}\varepsilon\|_2} \leqslant \frac{\|\bar{\delta}\|_2 + \|\varepsilon\|_2\|\bar{x}^*\|_2}{\|A^{*-1}\|_2^{-1} - \|\varepsilon\|_2}. \tag{26}$$

Последнее справедливо в том случае, если $\|A^{*-1}\|_2\|\varepsilon\|_2 < 1$. Далее, заметим, что

$$\|A^{*-1}\varepsilon\|_2 = \max_j \sum_{i=1}^{n}\left|\sum_{k=1}^{n} b_{ik}\varepsilon_{kj}\right| \leqslant \max_j \sum_{i=1}^{n}\sum_{k=1}^{n}|b_{ik}||\varepsilon_{kj}| \leqslant E\sum_{i=1}^{n}\sum_{k=1}^{n}|b_{ik}|, \tag{27}$$

где через E обозначено

$$E = \max_{i,\,k}|\varepsilon_{ik}| \tag{28}$$

и через b_{ik} обозначены элементы матрицы A^{*-1}:

$$b_{ik} = \frac{A^*_{ik}}{\Delta^*}. \tag{29}$$

Здесь A^*_{ik} — алгебраическое дополнение элемента a^*_{ik} матрицы A^* и Δ^* — определитель матрицы A^*. Таким образом, оценка для $\|\bar{\xi}\|_2$ при $E < 1\Big/\left(\sum_{i=1}^{n}\sum_{k=1}^{n}|b_{ik}|\right)$ примет вид

$$\|\bar{\xi}\|_2 \leqslant \frac{\sum_{i=1}^{n}\sum_{k=1}^{n}|A^*_{ik}|}{|\Delta^*|}\cdot\frac{\|\bar{\delta}\|_2 + E\|\bar{x}^*\|}{1-\dfrac{E}{|\Delta^*|}\sum_{i=1}^{n}\sum_{k=1}^{n}|A^*_{ik}|}. \tag{30}$$

Такая же оценка справедлива и для каждой компоненты вектора $\bar{\xi}$. Однако для компонент можно получить и более точную оценку. По (24) имеем:

$$|\xi_j| \leqslant \sum_{i=1}^{n}|b_{ji}||\delta_i| + \sum_{i=1}^{n}\sum_{k=1}^{n}|b_{ji}||\varepsilon_{ik}||x^*_k| + \sum_{i=1}^{n}\sum_{k=1}^{n}|b_{ji}||\varepsilon_{ik}||\xi_k|. \tag{31}$$

Отсюда

$$|\xi_j| \leqslant \max_i|\delta_i|\sum_{i=1}^{n}|b_{ji}| + E\|\bar{x}^*\|_2\sum_{i=1}^{n}|b_{ji}| + E\|\bar{\xi}\|_2\sum_{i=1}^{n}|b_{ji}|. \tag{32}$$

Подставляя сюда вместо $\|\xi\|_2$ выражение (30) и используя (29), получим:

$$|\xi_j| \leqslant \frac{\sum_{i=1}^{n}|A^*_{ji}|}{|\Delta^*|}\cdot\frac{\|\bar{\delta}\|_2 + E\|\bar{x}^*\|_2}{1-\dfrac{E}{|\Delta^*|}\sum_{i=1}^{n}\sum_{k=1}^{n}|A^*_{ik}|}. \tag{33}$$

Неудобство этой оценки состоит в том, что приходится вычислять определители высокого порядка.

Приведенные рассуждения показывают, что в некоторых случаях влияние неточности коэффициентов и правых частей на решение может оказаться значительным. Так, например, если величина

$$q = \frac{1}{|\Delta^*|} \sum_{i=1}^{n} \sum_{k=1}^{n} |A^*_{ik}| \tag{34}$$

велика, так что qE будет близко к единице, то правая часть (30) или (33) будет велика. В связи с этим вводилось ряд характеристик, аналогичных (34), так называемой «обусловленности» систем. При этом система считалась хорошо обусловленной, если небольшие изменения в коэффициентах и правой части мало изменяют решение системы, и плохо обусловленной, если небольшие изменения в коэффициентах и правой части вызывают большие изменения в решении системы. Пока эти вопросы еще не исследованы до конца.

В этом параграфе мы рассматривали только неустранимые погрешности. В процессе решения системы будут возникать также и ошибки округления. Однако, для того чтобы изучить влияние ошибок округления, нужно детальнее учесть алгоритм и вычислительные средства. Пока такие исследования проведены лишь для отдельных алгоритмов. Рассуждения, которые при этом приводятся, очень громоздки. Поэтому мы здесь этих вопросов касаться не будем.

УПРАЖНЕНИЯ

1. Найти все собственные значения матрицы

$$\begin{pmatrix} 4{,}1313 & 1{,}1617 & 2{,}1426 & 1{,}1718 \\ 0{,}1041 & 0{,}8144 & 3{,}1518 & 2{,}1413 \\ 1{,}1617 & 0{,}9864 & 0{,}8114 & 1{,}4656 \\ 3{,}1415 & 4{,}1617 & 2{,}3114 & 1{,}1681 \end{pmatrix}.$$

2. Показать, что в методе Ланцоша можно подбирать векторы $\bar{c}_k$ и $\bar{b}_k$, исходя из условий:

а) $\bar{c}_k$ — линейная комбинация $\bar{c}_0$, $A\bar{c}_0$, ..., $A^{k-1}\bar{c}_0$; $\bar{b}_k$ — линейная комбинация $\bar{b}_0$, $A'\bar{b}_0$, ..., $A'^{k-1}\bar{b}_0$;

б) скалярное произведение $(\bar{c}_k, \bar{b}_k)$ минимально.

3. Показать, что если $\bar{c}_0$ в методе Ланцоша ортогонален к одному из собственных векторов-строк матрицы A, скажем $\bar{v}_i$, то множитель $\lambda - \lambda_i$ выпадает из $\psi_m(\lambda)$; аналогично, если $\bar{b}_0$ ортогонален к одному из векторов-столбцов $\bar{u}_i$.

4. Рассмотрим произвольную одноколонную матрицу $x_0 = \begin{pmatrix} c_1 \\ c_2 \\ \cdot \\ \cdot \\ \cdot \\ c_n \end{pmatrix}$ и матрицу, состоящую из одной строки $y_0 = (d_1, d_2, \ldots, d_n)$. Обозначим

$$x_k = A^k x_0, \quad y_k = y_0 A^k, \quad c_k = y_i x_j \qquad (i + j = k).$$

Доказать, что

$$\begin{pmatrix} c_0 & c_1 & \dots & c_{n-1} \\ c_1 & c_2 & \dots & c_n \\ \dots & \dots & \dots & \dots \\ c_n & c_{n+1} & \dots & c_{2n-1} \end{pmatrix} \begin{pmatrix} p_n \\ p_{n-1} \\ \cdot \\ \cdot \\ p_1 \end{pmatrix} = \begin{pmatrix} c_n \\ c_{n+1} \\ \cdot \\ \cdot \\ c_{2n} \end{pmatrix},$$

где p_i — коэффициенты характеристического многочлена матрицы A:

$$D(\lambda) = (-1)^n \{\lambda^n - p_1\lambda^{n-1} - p_2\lambda^{n-2} - \dots - p_{n-1}\lambda - p_n\}.$$

5. Доказать, что если матрица A имеет вид

$$A = \begin{pmatrix} a_{11} & a_{12} & 0 & 0 & \dots & 0 & 0 \\ a_{21} & a_{22} & a_{23} & 0 & \dots & 0 & 0 \\ 0 & a_{32} & a_{33} & a_{34} & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & a_{n,\,n-1} & a_{nn} \end{pmatrix},$$

причем все $a_{k,k-1}a_{k,k+1} > 0$ $(k = 1, 2, \dots, n-1)$, то все собственные значения A действительны.

6. Используя результаты последней задачи, доказать, что все корни многочленов Лежандра действительны.

7. Показать, что все корни матрицы D § 6 удовлетворяют неравенству $|\mu_i| < g \operatorname{ctg} \frac{\pi}{2n}$, где g определено там же.

8. Показать, что уравнение

$$|A_0\lambda^3 + A_1\lambda^2 + A_2\lambda + A_3| = 0$$

при $|A_0| \neq 0$ эквивалентно уравнению

$$\begin{vmatrix} A_0^{-1}A_1 + \lambda I_n & A_0^{-1}A_2 & A_0^{-1}A_3 \\ -I_n & \lambda I_n & 0 \\ 0 & -I_n & \lambda I_n \end{vmatrix} = 0.$$

9. Пусть собственные значения A удовлетворяют условию

$$\lambda_1 > \lambda_2 > \dots > \lambda_{n-1} > \lambda_n > 0$$

и мы приближенно знаем λ_1, λ_2, λ_{n-1}, λ_n. Доказать, что тогда для уточнения λ_1 итерационным способом лучше всего взять матрицу $A - pI$ где $p = \frac{1}{2}(\lambda_2 + \lambda_n)$, а для уточнения λ_n — матрицу $A - qI$, где $q = \frac{1}{2}(\lambda_1 + \lambda_{n-1})$.

10. Показать, что собственные значения матрицы n-го порядка вида

$$\begin{pmatrix} 2 & -1 & 0 & 0 & \dots & 0 & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & -1 & 2 \end{pmatrix}$$

равны $\lambda_k = 2\left(1 + \cos \frac{\pi k}{n+1}\right)$ $(k = 1, 2, \dots, n)$.

11. Исследовать, при каких условиях итерация

$$\bar{x}_{k+1} = (2A^2 - I)\,\bar{x}_k + 2(A - I)\,\bar{b}$$

будет давать лучшую сходимость, чем итерация

$$\bar{x}_{k+1} = A\bar{x}_k + \bar{b}.$$

12. Получить следующую оценку для нормы разности $\|\bar{x} - \bar{x}^*\|_3$ между точным $\bar{x}$ и приближенным $\bar{x}^*$ решениями системы $A\bar{x} = b$:

$$\|\bar{x} - \bar{x}^*\|_3 \leqslant \frac{\|\bar{\delta}\|_3 \left(\sum\limits_{i,k=1}^{n} a_{ik}^2\right)^{n-1}}{|A|^2 (n-1)^{n-1}},$$

где $\bar{\delta} = A\bar{x} - A\bar{x}^*$, $|A|$ — определитель A.

13. Используя обозначения § 9, получить следующую оценку:

$$|\xi_j| \leqslant \left(\sum_{i=1}^{n} b_{ij}^2\right)^{\frac{1}{2}} \frac{(\bar{\delta}, \bar{\delta})^{\frac{1}{2}} + \left(\sum\limits_{i=1}^{n}\sum\limits_{k=1}^{n} \varepsilon_{ik}^2\right)^{\frac{1}{2}} (\bar{x}^*, \bar{x}^*)^{\frac{1}{2}}}{1 - \left(\sum\limits_{i=1}^{n}\sum\limits_{k=1}^{n} b_{ik}^2\right)^{\frac{1}{2}} \left(\sum\limits_{i=1}^{n}\sum\limits_{k=1}^{n} \varepsilon_{ik}^2\right)^{\frac{1}{2}}}.$$

14. Провести геометрический анализ влияния изменения коэффициентов системы на ее решение.

ЛИТЕРАТУРА

1. В. Н. Фаддеева, Вычислительные методы линейной алгебры, Гостехиздат, 1951.
2. Вейланд, Представление векового уравнения в виде многочлена, УМН, т. 2, вып. 4, 1947.
3. Хаусхолдер, Основы численного анализа, ИЛ, 1956.
4. М. Л. Бродский, Вероятностные оценки погрешностей при определении собственных значений и собственных векторов варьирующейся матрицы, УМН, т. 7, вып. 5, 1952.
5. И. М. Стесин, Вычисление собственных значений при помощи непрерывных дробей, УМН, т. 9, вып. 2, 1954.
6. М. А. Красносельский, О некоторых приемах приближенного вычисления собственных значений и собственных векторов положительно определенной матрицы, УМН, т. 11, вып. 3, 1956.
7. G. Forsythe, Решение линейных алгебраических уравнений может быть интересным, Bull. Amer. Math. Soc., т. 59, № 64 (1953).
8. К. А. Семендяев, О нахождении собственных значений и инвариантных многообразий матриц посредством итераций, ПММ, 7, 1943.

ГЛАВА 9

ПРИБЛИЖЕННЫЕ МЕТОДЫ РЕШЕНИЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

§ 1. Введение

Обыкновенные дифференциальные уравнения встречаются довольно часто в различных прикладных вопросах. При этом во многих случаях имеют дело с уравнениями, общее решение которых не выражается в квадратурах. Поэтому возникает необходимость применять те или иные методы, дающие приближенное решение задачи. С некоторыми из таких методов встречаются уже при изучении общей теории обыкновенных дифференциальных уравнений. Так, например, в общей теории обыкновенных дифференциальных уравнений изучается вопрос о возможности представления решения уравнения в виде ряда

$$y(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \ldots + a_n(x - x_0)^n + \ldots \quad (1)$$

или же более общего ряда

$$y(x) = (x - x_0)^\sigma [a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \ldots + a_n(x - x_0)^n + \ldots], \quad (2)$$

где σ — некоторое число, не обязательно целое и положительное. Если решение можно представить в виде (1) или (2) и если удается фактически найти достаточно большое количество коэффициентов $a_0, a_1, \ldots, a_n$ так, что абсолютная величина суммы остальных членов, т. е.

$$\sum_{k=n+1}^{+\infty} a_k (x - x_0)^k$$

или соответственно

$$|x - x_0|^\sigma \left| \sum_{k=n+1}^{\infty} a_k (x - x_0)^k \right|, \quad (3)$$

меньше, чем заданная нам допустимая погрешность, то соответствующий отрезок ряда может служить приближенным представлением искомого решения. Таким же образом можно использовать тригонометрические ряды или ряды по другим ортогональным функциям.

При доказательстве существования решения дифференциального уравнения

$$y' = f(x, y) \tag{4}$$

с начальным условием $y(x_0) = y_0$ (или же системы таких уравнений) часто используют метод последовательных приближений Пикара. При этом точное решение получается как предел последовательности

$$y_0(x),\ y_1(x),\ \ldots,\ y_n(x),\ \ldots, \tag{5}$$

где

$$y_n(x) = y_0 + \int_{x_0}^{x} f(x, y_{n-1}(x))\,dx \qquad (n = 1, 2, \ldots). \tag{6}$$

Процесс последовательных приближений Пикара сходится при выполнении таких условий:

1. Функция $f(x, y)$ непрерывна в области $R = \{|x - x_0| \leqslant a, |y - y_0| \leqslant b\}$.

2. Функция $f(x, y)$ удовлетворяет в R условию Липшица по y:

$$|f(x, \bar{y}) - f(x, \bar{\bar{y}})| \leqslant L|\bar{y} - \bar{\bar{y}}|.$$

Здесь L — постоянная, не зависящая от x, $\bar{y}$ и $\bar{\bar{y}}$, а точки $(x, \bar{y})$ и $(x, \bar{\bar{y}})$ — произвольные точки области R.

При выполнении этих условий $y_n(x)$ равномерно сходится к функции $y(x)$ на $[x_0 - h, x_0 + h]$, где $h = \min\left(a, \frac{b}{M}\right)$ и $M = \sup\limits_R |f(x, y)|$, и функция $y(x)$ удовлетворяет дифференциальному уравнению (4) и начальному условию.

Если метод последовательных приближений сходится и может быть фактически осуществлен для достаточно больших n, так что $|y_n(x) - y(x)|$ не превышает заданной нам допустимой погрешности, то мы можем принять $y_n(x)$ за приближенное решение задачи.

Мы не будем здесь подробно останавливаться на этих методах, достаточно хорошо изложенных в общих курсах.

Методы, которые мы будем изучать, можно разделить на две большие группы. Одни из них дают приближенное решение в виде аналитического выражения, другие — в виде таблицы. Будем называть первую группу методов *аналитическими*, вторую — *численными*.

Мы начнем изложение с аналитического метода, предложенного академиком Сергеем Алексеевичем Чаплыгиным.

§ 2. Метод С. А. Чаплыгина

1. Теоремы о дифференциальных неравенствах. Метод Чаплыгина основан на его теореме, которую он назвал теоремой о дифференциальных неравенствах. Приведем доказательство этой теоремы в формулировке, несколько более уточненной по сравнению с данной самим С. А. Чаплыгиным.

Теорема 1. *Пусть функции* $f(x, y)$ *и* $F(x, y)$ *непрерывны в области*

$$R = \{x_0 \leqslant x \leqslant x_0 + a;\ |y - y_0| \leqslant b\} \qquad (a > 0,\ b > 0) \tag{1}$$

и удовлетворяют условию

$$f(x, y) \leqslant F(x, y). \tag{2}$$

Пусть, далее, $y = y(x)$ *и* $y = U(x)$ *— решения дифференциальных уравнений*

$$y' = f(x, y), \quad U' = F(x, U), \tag{3}$$

проходящие через точку (x_0, y_0), *определенные при* $x_0 \leqslant x \leqslant x_0 + a$ *и лежащие между* $y_0 - b$ *и* $y_0 + b$. *Тогда если* $f(x, y)$ *удовлетворяет в нашей области условию Липшица, то при* $x_0 \leqslant x \leqslant x_0 + a$ *имеет место неравенство* $U(x) \geqslant y(x)$. *Более того, если в некоторой точке* $x_1 > x_0$ *имеет место неравенство* $U(x_1) > y(x_1)$, *то* $U(x) > y(x)$ *при всех* $x \in [x_1, x_0 + a]$.

Для доказательства рассмотрим

$$U' - y' = F(x, U) - f(x, y) \tag{4}$$

и обозначим $U - y = z$. Тогда для z получим следующее дифференциальное уравнение:

$$z' = F(x, U) - f(x, U - z). \tag{5}$$

Рассматривая здесь U как известную функцию x и обозначая

$$F(x, U) - f(x, U - z) = g(x, z), \tag{6}$$

получим, что z удовлетворяет дифференциальному уравнению

$$z' = g(x, z) \tag{7}$$

и начальному условию $z(x_0) = 0$. Запишем уравнение (7) в виде

$$z' - g(x, 0) = g(x, z) - g(x, 0). \tag{8}$$

На основании (2) имеем:

$$g(x, 0) = F(x, U) - f(x, U) \geqslant 0. \tag{9}$$

Так как $f(x, y)$ удовлетворяет условию Липшица, то

$$|g(x, z) - g(x, 0)| = |F(x, U) - f(x, U - z) - F(x, U) + f(x, U)| = |f(x, U) - f(x, U - z)| \leqslant L|z|. \tag{10}$$

Таким образом,

$$-L|z| \leqslant z' - g(x, 0) \leqslant L|z|. \tag{11}$$

Предположим, что вопреки утверждению теоремы в некоторой точке x_1 отрезка $[x_0, x_0 + a]$ $U(x_1) < y(x_1)$. Так как $U(x_0) = y(x_0)$, то на отрезке $[x_0, x_1]$ найдутся точки, где $U(x) = y(x)$. Не уменьшая

общности, мы можем предположить, что эта точка единственная и что она совпадает с x_0. Тогда $U(x_0) = y(x_0)$ и $U(x) < y(x)$ для всех точек $(x_0, x_1]$. На отрезке $[x_0, x_1]$ неравенство (11) можно записать в виде

$$Lz \leqslant z' - g(x, 0) \leqslant -Lz. \tag{12}$$

Отсюда

$$z' - Lz = \varphi(x) \geqslant g(x, 0) \geqslant 0 \tag{13}$$

и для всех точек $[x_0, x_1]$

$$z = \int_{x_0}^{x} e^{L(x-t)}\varphi(t)\,dt \geqslant \int_{x_0}^{x} e^{L(x-t)} g(t, 0)\,dt \geqslant 0, \tag{14}$$

что противоречит нашему предположению. Итак, мы доказали, что на всем отрезке $[x_0, x_0 + a]$ $U(x) \geqslant y(x)$ или $z \geqslant 0$.

Может оказаться, что $z \equiv 0$ при $x \in [x_0, x_0 + a]$. Тогда $U(x) \equiv y(x)$, $g(x, 0) \equiv 0$ и $F(x, U) \equiv f(x, U)$. Это может быть даже тогда, когда $f(x, y)$ совпадает с $F(x, y)$ не всюду в области R. Так, если

$$F(x, y) \equiv 1, \quad f(x, y) = 1 - (y - x)^2, \tag{15}$$

то

$$F(x, y) > f(x, y) \tag{16}$$

всюду, за исключением точек, лежащих на прямой $y = x$. Но дифференциальные уравнения

$$U' = F(x, U), \quad y' = f(x, y) \tag{17}$$

имеют совпадающее решение $y = U = x$, удовлетворяющее начальному условию $y(0) = U(0) = 0$.

Предположим теперь, что имеется такая точка $x_1 \in [x_0, x_0 + a]$, в которой $U(x_1) > y(x_1)$. Покажем тогда, что на всем отрезке $[x_1, x_0 + a]$ будет иметь место неравенство $U(x) > y(x)$. Если бы это было не так, то нашлась бы такая точка $x_2 \in [x_1, x_0 + a]$, что $U(x_2) = y(x_2)$ и $U(x) > y(x)$ на (x_1, x_2). Обозначим через x_0' ближайшую к x_1 точку отрезка $[x_0, x_1]$, в которой $U(x_0') = y(x_0')$. На интервале (x_0', x_2) неравенство (11) может быть записано в виде

$$-Lz \leqslant z' - g(x, 0) \leqslant Lz. \tag{18}$$

Отсюда

$$z' + Lz = \varphi(x) \geqslant g(x, 0) \geqslant 0 \tag{19}$$

или

$$z = \int_{x_0'}^{x} e^{-L(x-t)}\varphi(t)\,dt \geqslant \int_{x_0'}^{x} e^{-L(x-t)} g(t, 0)\,dt \geqslant 0. \tag{20}$$

В силу нашего предположения, что $z(x_2)=0$, будем иметь:

$$\int_{x_0'}^{x_2} e^{-L(x-t)}g(t,0)\,dt=0 \tag{21}$$

или $g(x,0)\equiv 0$ при $x\in[x_0', x_2]$, так как $e^{-L(x-t)}>0$. Но тогда правая часть неравенства (18) даст

$$z'-Lz=\psi(x)\leqslant g(x,0)\equiv 0,\quad x\in[x_0', x_2], \tag{22}$$

или

$$z=\int_{x_0'}^{x} e^{L(x-t)}\psi(t)\,dt\leqslant\int_{x_0'}^{x} e^{L(x-t)}g(t,0)\,dt\equiv 0, \tag{23}$$

а это противоречит нашему предположению, что $z(x_1)>0$. Теорема доказана полностью.

Сделаем несколько замечаний к доказанной теореме.

1. Если вместо (2) предположить, что

$$F(x,y)>f(x,y) \tag{24}$$

в рассматриваемой области (что делал С. А. Чаплыгин), то при всех $x>x_0$ будет $U>y$. Действительно, в силу неравенства (24) будем иметь $U'(x_0)>y'(x_0)$. Отсюда $U(x)>y(x)$ при всех $x>x_0$ и достаточно близких к x_0. В силу теоремы 1 это будет справедливо и при всех $x\in(x_0, x_0+a]$.

2. Если бы вместо $F(x,y)$ взять непрерывную функцию $\varphi(x,y)$ такую, что

$$f(x,y)\geqslant\varphi(x,y) \tag{25}$$

в рассматриваемой области, то, повторяя рассуждения доказательства теоремы 1, мы доказали бы, что решения уравнений

$$u'=\varphi(x,u),\quad y'=f(x,y), \tag{26}$$

удовлетворяющие начальным условиям $u(x_0)=y(x_0)=y_0$ при $x\in[x_0, x_0+a]$, удовлетворяют неравенству

$$u(x)\leqslant y(x), \tag{27}$$

причем, если в какой-то точке $x_1\in[x_0, x_0+a]$, $u(x)<y(x)$, то $u(x)<y(x)$ при всех $x\in[x_1, x_0+a]$.

Теорема о дифференциальных неравенствах Чаплыгина позволяет в некоторых случаях отыскать границы $U(x)$ и $u(x)$, в которых заключено точное решение $y(x)$.

На практике чаще всего для отыскания U и u подбирают функции $F(x,y)$ и $\varphi(x,y)$ так, чтобы имели место неравенства

$F(x, y) \geqslant f(x, y) \geqslant \varphi(x, y)$, а уравнения

$$U' = F(x, U), \quad u' = \varphi(x, u) \tag{28}$$

легко интегрировались бы в квадратурах.

Рассмотрим такой пример. Пусть дано уравнение

$$y' = x^2 + y^2$$

и нам требуется найти решение его при $0 \leqslant x \leqslant 1$, удовлетворяющее начальному условию $y(0) = 0$. В качестве функций $F(x, y)$ и $\varphi(x, y)$ можно взять

$$F(x, y) = 1 + y^2; \quad \varphi(x, y) = x^2.$$

В результате получим следующие функции U и u:

$$U = \operatorname{tg} x; \quad u = \frac{x^3}{3}.$$

Иногда удается подобрать так функции U и u, что $u(x_0) = U(x_0) = y_0$ и

$$\frac{dU}{dx} - f(x, U) \geqslant 0, \quad \frac{du}{dx} - f(x, u) \leqslant 0. \tag{29}$$

Тогда, очевидно, будем иметь $U \geqslant y \geqslant u$. Такими функциями, например, будут:

$$\left.\begin{aligned} U &= y_0 + \int_{x_0}^{x} e^{L(x-t)} |f(t, y_0)|\, dt = y_0 + Y, \\ u &= y_0 - \int_{x_0}^{x} e^{L(x-t)} |f(t, y_0)|\, dt = y_0 - Y. \end{aligned}\right\} \tag{30}$$

Действительно,

$$\begin{aligned} U' - f(x, U) &= |f(x, y_0)| + LY - f(x, y_0 + Y) = \\ &= |f(x, y_0)| - f(x, y_0) + LY - \{f(x, y_0 + Y) - f(x, y_0)\}. \end{aligned} \tag{31}$$

В силу условия Липшица, фигурная скобка по модулю не больше LY. Таким образом,

$$U' - f(x, U) \geqslant 0. \tag{32}$$

Аналогично доказывается, что

$$u' - f(x, u) \leqslant 0. \tag{33}$$

2. Способ Чаплыгина построения улучшенных приближений. Не всегда удается такими способами получить сразу достаточно тесные границы для $y(x)$. Поэтому возникает задача об улучшении этих границ. С. А. Чаплыгин предложил способ по найденным U и u получать улучшенные приближения U_1 и u_1. При этом приходится накладывать довольно сильные ограничения на функцию $f(x, y)$,

а именно предполагать, что $\frac{\partial^2 f}{\partial y^2}$ сохраняет свой знак в области, ограниченной кривыми $y=U(x)$ и $y=u(x)$ и прямыми $x=x_0$ и $x=x_0+a$. Геометрически это означает, что сечения поверхности

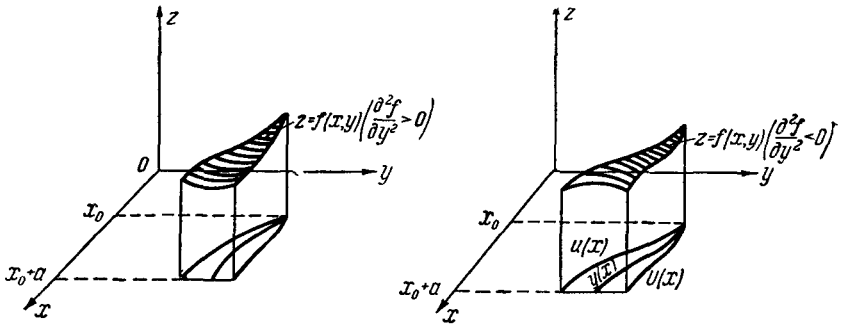

Рис. 21.

$z=f(x, y)$ плоскостями $x=\text{const}$ будут либо все время выпуклы, либо все время вогнуты (рис. 21).

Рассмотрим некоторое сечение поверхности $z=f(x, y)$ плоскостью $x=\text{const}$ (рис. 22). Через u, y, U на рис. 21 обозначены соответственно точки пересечения кривых $y=u(x)$, $y=y(x)$, $y=U(x)$

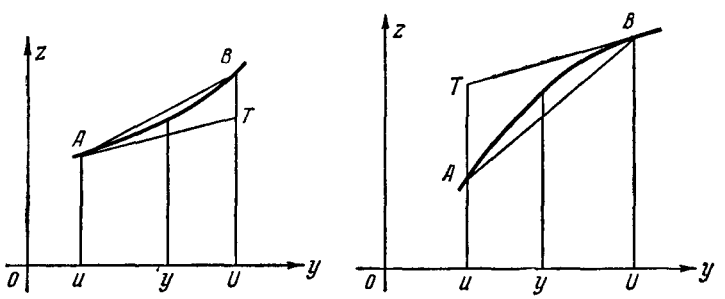

Рис. 22.

плоскостью сечения, а через A и B — проекции u и U на сечение поверхности $z=f(x, y)$.

Проведем секущую AB и касательную AT, если $\frac{\partial^2 f}{\partial y^2}>0$, и BT, если $\frac{\partial^2 f}{\partial y^2}<0$. Произведя такие построения при каждом x рассматриваемой области, получим две поверхности, одна из которых образована секущими, другая — касательными. Одна из поверхностей, будем обозначать ее $z=F(x, y)$, будет расположена над

поверхностью $z=f(x, y)$, другая $z=\varphi(x, y)$ будет расположена под поверхностью $z=f(x, y)$. Таким образом,

$$F(x, y) \geqslant f(x, y) \geqslant \varphi(x, y). \tag{34}$$

Заметим, что функции $F(x, y)$ и $\varphi(x, y)$ линейны относительно y. Следовательно, уравнения

$$\frac{dU_1}{dx}=F(x, U_1); \quad \frac{du_1}{dx}=\varphi(x, u_1) \tag{35}$$

интегрируются в квадратурах. Пусть U_1 и u_1 соответственно решения этих уравнений, удовлетворяющие тем же начальным условиям, что и прежде. Покажем, что эти функции удовлетворяют поставленным условиям. В самом деле, вдоль кривой $y=U(x)$ имеем:

$$\frac{dU}{dx}-f(x, U) \geqslant 0 \tag{36}$$

и, кроме того, $f(x, U)=F(x, U)$. Таким образом, $\frac{dU}{dx}-F(x, U) \geqslant 0$ и $U \geqslant U_1$, а из неравенства $f(x, y) \leqslant F(x, y)$ следует $U_1 \geqslant y$. Аналогично проводится доказательство и для u.

Приведем аналитическое изложение этого метода. Будем предполагать, что $\frac{\partial^2 f}{\partial y^2}>0$. Имеем:

$$\left.\begin{aligned} &\frac{dy}{dx}=f(x, y), \\ &\frac{du}{dx}=f(x, u)-\psi, \quad (\psi \geqslant 0). \end{aligned}\right\} \tag{37}$$

Обозначим $y-u=z$. Вычитая из верхнего равенства нижнее и применяя к разности $f(x, y)-f(x, u)$ формулу Тейлора, получим:

$$\frac{dz}{dx}=z\frac{\partial f(x, u)}{\partial y}+\frac{z^2}{2}\frac{\partial^2 f(x, \xi)}{\partial y^2}+\psi(x), \tag{38}$$

где $u \leqslant \xi \leqslant y$. Решение этого уравнения при начальных данных $z(x_0)=0$ дает поправку, которую нужно прибавить к $u(x)$, чтобы получить $y(x)$. Отбросим член с z^2. Тогда функция $Z(x)$, удовлетворяющая уравнению

$$\frac{dZ}{dx}=Z\frac{\partial f(x, u)}{\partial y}+\psi(x) \tag{39}$$

и начальным условиям $Z(x_0)=0$, будет удовлетворять неравенству $Z \leqslant z$. С другой стороны, если положить $\psi(x) \equiv 0$, то решением уравнения (39) с нулевым начальным значением будет тождественный нуль (по теореме единственности). Таким образом, $Z \geqslant 0$,

и если принять $u_1 = u + Z$, то будем иметь $u \leqslant u_1 \leqslant y$, что нам и требовалось:

$$\frac{du_1}{dx} = \frac{du}{dx} + \frac{dZ}{dx} = f(x, u) = \psi + Z\frac{\partial f(x, u)}{dy} + \psi(x) =$$
$$= f(x, u) + (u_1 - u)\frac{\partial f(x, u)}{dy} = \varphi(x, u_1).$$

Для улучшения оценки сверху рассмотрим уравнения

$$\left.\begin{aligned} &\frac{dU}{dx} = f(x, U) + \theta(x) \quad (\theta(x) \geqslant 0), \\ &\frac{dy}{dx} = f(x, y). \end{aligned}\right\} \tag{40}$$

Положим $U - y = t$ и вычтем из верхнего равенства нижнее. Получим:

$$\frac{dt}{dx} = f(x, U) - f(x, y) + \theta(x). \tag{41}$$

Введем обозначение

$$\frac{f(x, U) - f(x, u)}{U - u} = F(x). \tag{42}$$

Тогда

$$f(x, U) - f(x, y) = tF(x) + \beta(x), \tag{43}$$

где

$$\beta(x) = f(x, U) - f(x, y) - (U - y)\frac{f(x, U) - f(x, u)}{U - u} =$$
$$= (U - y)\left[\frac{f(x, U) - f(x, y)}{U - y} - \frac{f(x, U) - f(x, u)}{U - u}\right]. \tag{44}$$

Обозначим через $\varphi(\tau)$ функцию

$$\varphi(\tau) = \frac{f(x, U) - f(x, \tau)}{U - \tau} \qquad (\tau \leqslant U). \tag{45}$$

Производная этой функции равна

$$\varphi'(\tau) = \frac{-\frac{\partial f(x, \tau)}{\partial \tau}(U - \tau) + f(x, U) - f(x, \tau)}{(U - \tau)^2} =$$
$$= \left[-\frac{\partial f(x, \tau)}{\partial \tau} + \frac{\partial f(x, \xi)}{\partial \tau}\right] : (U - \tau) =$$
$$= \frac{\xi - \tau}{U - \tau}\frac{\partial^2 f(x, \eta)}{\partial \tau^2} \qquad (\tau \leqslant \xi,\ \eta \leqslant U). \tag{46}$$

Так как по предположению $\frac{\partial^2 f(x, y)}{\partial y^2} \geqslant 0$, то $\varphi'(\tau) \geqslant 0$. Поэтому $\varphi(\tau)$ — неубывающая функция и $\varphi(y) > \varphi(u)$. Отсюда $\beta(x) \geqslant 0$. Таким образом, решение уравнения

$$\frac{dT}{dx} = TF(x) + \theta(x), \tag{47}$$

обращающееся в нуль при $x = x_0$, будет при $x > x_0$ удовлетворять неравенству $T \leqslant t$. Так же как и для Z, мы получим $T \geqslant 0$. Поэтому, если принять $U_1 = U - T$, то получим улучшенную верхнюю функцию

$$\frac{dU_1}{dx} = \frac{dU}{dx} - \frac{dT}{dx} = f(x, U) + \theta(x) - TF(x) - \theta(x) =$$

$$= f(x, U) + (U_1 - U)\frac{f(x, U) - f(x, u)}{U - u} = F(x, U_1).$$

При $\frac{\partial^2 f}{\partial y^2} < 0$ порядок действий будет обратным.

В приведенном выше примере имеем $\frac{\partial^2 f}{\partial y^2} = 2 > 0$. Следовательно, мы можем применить наши рассуждения. Уравнение для определения Z будет

$$\frac{dZ}{dx} = Z\frac{2x^3}{3} + \frac{x^6}{9}.$$

Его решение, обращающееся в нуль при $x = 0$, имеет вид

$$Z = e^{\frac{x^4}{6}} \int_0^x e^{-\frac{x^4}{6}} \frac{x^6}{9}\, dx.$$

Таким образом,

$$u_1 = \frac{x^3}{3} + \frac{1}{9} e^{\frac{x^4}{6}} \int_0^x e^{-\frac{x^4}{6}} x^6\, dx.$$

Уравнение для T примет вид

$$\frac{dT}{dx} = \left(\operatorname{tg} x + \frac{x^3}{3}\right) T + (1 - x^2)$$

и

$$T = \frac{1}{\cos x} e^{\frac{x^4}{12}} \int_0^x (1 - x^2) \cos x e^{-\frac{x^4}{12}}\, dx,$$

а

$$U_1 = \operatorname{tg} x - \frac{1}{\cos x} e^{\frac{x^4}{12}} \int_0^x (1 - x^2) \cos x e^{-\frac{x^4}{12}}\, dx.$$

Процесс уточнения границ можно повторять неограниченно, если только квадратуры выполнимы. Как было показано акад. Н. Н. Лузиным, последовательность $y - u_n$, если $\frac{\partial^2 f}{\partial y^2} \geqslant 0$, или $U_n - y$, если $\frac{\partial^2 f}{\partial y^2} \leqslant 0$, будет стремиться к нулю как $\frac{C}{2^{2^n}}$ в предположении, что начальное приближение взято достаточно близким к y. Это очень быстрая сходимость, такая же, как и у метода Ньютона для решения алгебраических и трансцендентных уравнений. Общность между

методом Ньютона и методом Чаплыгина не заканчивается на этом. Можно показать, что метод Чаплыгина является обобщением метода Ньютона на нелинейные функциональные уравнения.

К сожалению, обычно метод Чаплыгина приводит к очень сложным квадратурам, не выражающимся в элементарных функциях.

3. Второй способ построения улучшенных приближений. Улучшенные приближения u_1 и U_1 можно находить по способу Чаплыгина только в том случае, когда $\frac{\partial^2 f}{\partial y^2}$ сохраняет свой знак в рассматриваемой области. Дадим еще один способ получения улучшенных приближений, свободный от этого недостатка.

Теорема 2. *Пусть функции* $f(x, y)$, $F(x, y)$, $y(x)$ *и* $U(x)$ *определены как и в теореме* 1. *Тогда если положить*

$$h(x) = U'(x) - f(x, U(x)), \tag{48}$$

то функция

$$U_1(x) = U(x) - \int_{x_0}^{x} e^{-L(x-t)} h(t)\, dt, \tag{49}$$

где L *— константа Липшица для функции* $f(x, y)$, *является верхней функцией на отрезке* $[x_0, x_0 + a]$ *и на этом отрезке имеет место неравенство*

$$y(x) \leqslant U_1(x) \leqslant U(x). \tag{50}$$

Очевидно, $U_1(x) \leqslant U(x)$. Далее,

$$U_1'(x) - f(x, U_1(x)) = U'(x) - h(x) + L \int_{x_0}^{x} e^{-L(x-t)} h(t)\, dt - f(x, U_1(x)) =$$

$$= f(x, U_1(x)) - f(x, U_1(x)) + L[U(x) - U_1(x)] \geqslant 0. \tag{51}$$

Следовательно, $y(x) \leqslant U_1(x)$, что и требовалось доказать.

Совершенно аналогично доказывается, что если $u(x)$ является нижней функцией и мы образуем

$$h_1(x) = u'(x) - f(x, u(x)), \tag{52}$$

то

$$u_1(x) = u(x) - \int_{x_0}^{x} e^{-L(x-t)} h_1(t)\, dt \tag{53}$$

также будет нижней функцией и будет удовлетворять неравенствам $u(x) \leqslant u_1(x) \leqslant y(x)$.

И в этом случае мы можем, по крайней мере теоретически, неограниченно продолжать процесс получения последовательных приближений. В связи с этим докажем следующую теорему:

Теорема 3. *Пусть* $f(x, y)$, $F(x, y)$, $y(x)$, $U(x)$ *определены как и в теореме* 1. *Положим*

$$h_0(x) = U'(x) - f(x, U(x)) \tag{54}$$

и образуем последовательности

$$U_n(x) = U_{n-1}(x) - \int_{x_0}^{x} e^{-L(x-t)} h_{n-1}(t)\, dt \qquad (n = 1, 2, \ldots), \tag{55}$$

$$h_n(x) = U_n'(x) - f[x, U_n(x)] \qquad (n = 1, 2, \ldots). \tag{56}$$

Тогда последовательность $\{U_n(x)\}$ *равномерно на* $[x_0, x_0 + a]$ *сходится к* $y(x)$.

Прежде всего отметим, что все функции $U_n(x)$ определены в каждой точке отрезка $[x_0, x_0 + a]$, так как они заключены между $U(x)$ и $y(x)$ и не могут выйти за пределы рассматриваемой области до пересечения с прямой $x = x_0 + a$.

Далее, рассмотрим ряд

$$U_0(x) + [U_1(x) - U_0(x)] + \ldots + [U_n(x) - U_{n-1}(x)] + \ldots \tag{57}$$

Если подставить сюда вместо разностей $U_n(x) - U_{n-1}(x)$ их выражения из (55), то получим:

$$U_0(x) - \int_{x_0}^{x} e^{-L(x-t)} [h_0(t) + h_1(t) + \ldots]\, dt. \tag{58}$$

Отсюда следует, что ряд (57) равномерно сходится, если равномерно сходится ряд

$$h_0(x) + h_1(x) + \ldots + h_n(x) + \ldots \tag{59}$$

Дифференцируя (55), получим:

$$U_n'(x) = U_{n-1}'(x) - h_{n-1}(x) + L \int_{x_0}^{x} e^{-L(x-t)} h_{n-1}(t)\, dt. \tag{60}$$

Отсюда находим, подставляя вместо $h_{n-1}(x)$ перед интегралом его выражение через U_{n-1} и f:

$$U_n'(x) = f(x, U_{n-1}) + L \int_{x_0}^{x} e^{-L(x-t)} h_{n-1}(t)\, dt. \tag{61}$$

Таким образом,

$$h_n(x) = f(x, U_{n-1}) - f(x, U_n) + L \int_{x_0}^{x} e^{-L(x-t)} h_{n-1}(t)\, dt. \tag{62}$$

В силу условия Липшица получим:

$$h_n(x) \leqslant 2L \int_{x_0}^{x} e^{-L(x-t)} h_{n-1}(t)\, dt. \tag{63}$$

Определим теперь последовательность $\{H_n(x)\}$ при помощи рекуррентного соотношения

$$H_n(x) = 2L \int_{x_0}^{x} e^{-L(x-t)} H_{n-1}(t)\, dt, \tag{64}$$

причем в качестве $H_0(x)$ возьмем $h_0(x)$. Очевидно, что $h_n(x) \leqslant H_n(x)$. Произведем оценку $H_n(x)$. Для этого выразим все $H_n(x)$ через $H_0(x)$. Подставляя в $H_2(x)$ выражение $H_1(x)$ через $H_0(x)$ и меняя порядок интегрирования, получим:

$$H_2(x) = 2L \int_{x_0}^{x} e^{-L(x-t)} H_1(t)\, dt = (2L)^2 \int_{x_0}^{x} e^{-L(x-t)} \left\{ \int_{x_0}^{t} e^{-L(t-\tau)} H_0(\tau)\, d\tau \right\} dt =$$

$$= (2L)^2 \int_{x_0}^{x} d\tau \int_{\tau}^{x} e^{-L(x-\tau)} H_0(\tau)\, dt = (2L)^2 \int_{x_0}^{x} (x-\tau) e^{-L(x-\tau)} H_0(\tau)\, d\tau. \tag{65}$$

Аналогично для $H_3(x)$ будем иметь:

$$H_3(x) = 2L \int_{x_0}^{x} e^{-L(x-t)} H_2(t)\, dt =$$

$$= (2L)^3 \int_{x_0}^{x} e^{-L(x-t)} \left\{ \int_{x_0}^{t} (t-\tau)\, e^{-L(t-\tau)} H_0(\tau)\, d\tau \right\} dt =$$

$$= (2L)^3 \int_{x_0}^{x} d\tau \int_{\tau}^{x} (t-\tau)\, e^{-L(x-\tau)} H_0(\tau)\, dt =$$

$$= (2L)^3 \int_{x_0}^{x} \frac{(x-\tau)^2}{2} e^{-L(x-t)} H_0(\tau)\, d\tau. \tag{66}$$

Покажем по индукции, что

$$H_n(x) = (2L)^n \int_{x_0}^{x} \frac{(x-t)^{n-1}}{(n-1)!} e^{-L(x-t)} H_0(t)\, dt. \tag{67}$$

Действительно,

$$H_{n+1}(x) = (2L)\int_{x_0}^{x} e^{-L(x-t)}H_n(t)\,dt =$$

$$= (2L)^{n+1}\int_{x_0}^{x} e^{-L(x-t)}\left\{\int_{x_0}^{t}\frac{(t-\tau)^{n-1}}{(n-1)!}e^{-L(t-\tau)}H_0(\tau)\,d\tau\right\}dt =$$

$$= (2L)^{n+1}\int_{x_0}^{x} d\tau\int_{\tau}^{x} e^{-L(x-\tau)}\frac{(t-\tau)^{n-1}}{(n-1)!}H_0(\tau)\,dt =$$

$$= (2L)^{n+1}\int_{x_0}^{x}\frac{(x-\tau)^n}{n!}e^{-L(x-\tau)}H_0(\tau)\,d\tau, \qquad (68)$$

что и требовалось доказать.

Обозначим через M_0 верхнюю границу $h_0(x)$ на $[x_0,\ x_0+a]$. Тогда

$$H_n(x) \leqslant \frac{(2L)^n}{(n-1)!}M_0\int_{x_0}^{x}(x-t)^{n-1}e^{-L(x-t)}\,dt \leqslant$$

$$\leqslant \frac{(2L)^n}{(n-1)!}M_0\int_{x_0}^{x}(x-t)^{n-1}\,dt \leqslant \frac{(2La)^n}{n!}M_0. \qquad (69)$$

Отсюда следует абсолютная и равномерная сходимость ряда $\sum_{n=0}^{\infty}H_n(x)$.

Поэтому ряд $\sum_{n=0}^{\infty}h_n(x)$ также абсолютно и равномерно сходится и, следовательно, $\{U_n(x)\}$ стремится равномерно к некоторой непрерывной функции $Y(x)$. Покажем, что $Y(x)$ будет являться решением дифференциального уравнения

$$Y'(x) = f(x,\ Y). \qquad (70)$$

Для этого рассмотрим

$$Y - U_n + \int_{x_0}^{x}[f(t,\ U_n) - f(t,\ Y)]\,dt + \int_{x_0}^{x}h_n(t)\,dt = r_n(x). \qquad (71)$$

Так как $U_n(x)$ равномерно стремится к $Y(x)$, а $h_n(x)$ к нулю, то $r_n(x)$ равномерно стремится к нулю. Интегрируя (56) и используя равенства $U_n(x_0) = y_0$, можно (71) записать в виде

$$Y = y_0 + \int_{x_0}^{x} f(t,\ Y)\,dt + r_n(x). \qquad (72)$$

Отсюда

$$Y = y_0 + \int_{x_0}^{x} f(t, Y)\, dt, \tag{73}$$

что и требовалось доказать.

Так как при любом n $U_n(x_0) = y_0$, то и $Y(x_0) = y_0$. Тем самым теорема доказана полностью. В силу теоремы единственности решения дифференциального уравнения $Y(x) \equiv y(x)$.

4. Метод Чаплыгина приближенного решения линейных дифференциальных уравнений второго порядка. На этом мы закончим рассмотрение метода Чаплыгина для уравнений первого порядка. Рассмотрим теперь некоторые факты, связанные с применением метода Чаплыгина к линейным уравнениям второго порядка. Будем рассматривать частное решение уравнения

$$y''(x) + p(x)\, y'(x) + q(x)\, y(x) = r(x), \tag{74}$$

удовлетворяющее начальным условиям $y(x_0) = y_0$, $y'(x_0) = y_0'$. Относительно коэффициентов уравнения будем предполагать, что $p(x)$, $q(x)$ и $r(x)$ — непрерывные функции на некотором отрезке $[a, b]$, содержащем точку x_0, а $p(x)$ непрерывно дифференцируема на этом отрезке. Рассмотрим, кроме того, дважды непрерывно дифференцируемую на $[a, b]$ функцию $v(x)$, удовлетворяющую при $x \geqslant x_0$ неравенству

$$v''(x) + p(x)\, v'(x) + q(x)\, v(x) > r(x) \tag{75}$$

и начальным условиям $v(x_0) = y_0$, $v'(x_0) = y_0'$. Применяя теорему Тейлора, будем иметь:

$$y(x) = y_0 + (x - x_0)\, y_0' + \frac{(x - x_0)^2}{2}\, y''(x_0) + o[(x - x_0)^2], \tag{76}$$

$$v(x) = y_0 + (x - x_0)\, y_0' + \frac{(x - x_0)^2}{2}\, v''(x_0) + o[(x - x_0)^2]. \tag{77}$$

Так как $v''(x_0) > y''(x_0)$, то при достаточно малом $|x - x_0|$ и $x - x_0 > 0$ имеет место неравенство $v(x) > y(x)$. Но в отличие от того, что мы имели раньше для уравнения первого порядка, мы уже не можем утверждать, что $v(x) > y(x)$ на всем отрезке, где выполнено условие (75). Так, рассмотрим уравнение

$$y'' + y = 0.$$

Частное решение его, удовлетворяющее начальным данным $y(0) = y'(0) = 0$, будет, очевидно, $y(x) \equiv 0$. Если же взять уравнение

$$v'' + v = \varphi(x),$$

то решение его, удовлетворяющее условиям $v(0) = v'(0) = 0$, будет

$$v = \int_0^x \varphi(t) \sin(x - t)\, dt,$$

что легко проверяется простыми вычислениями. Функция v может принимать отрицательные значения при $x > x_0$, даже если $\varphi(x)$ всюду положительна. Возьмем, например, $\varphi(x) = e^{-ax}$, где a — некоторое положительное число. Тогда

$$v(x) = \int_0^x e^{-at} \sin(x - t)\, dt = \frac{a \sin x - \cos x}{1 + a^2} + \frac{e^{-ax}}{1 + a^2}.$$

Возьмем $x = \pi + \varepsilon$, где $\varepsilon > 0$ — фиксированное, как угодно малое число. При этом

$$v(\pi + \varepsilon) = \frac{\cos \varepsilon - a \sin \varepsilon}{1 + a^2} + \frac{e^{-a(\pi + \varepsilon)}}{1 + a^2}.$$

При достаточно большом a $v(x + \varepsilon)$ будет отрицательным, так как отрицательный член

$$-\frac{a \sin \varepsilon}{1 + a^2}$$

будет преобладать над остальными. Геометрически дело здесь заключается в следующем. Синус, стоящий под знаком интеграла, меняет знак, причем положительный множитель при нем имеет относительно большие значения там, где синус отрицателен (рис. 23).

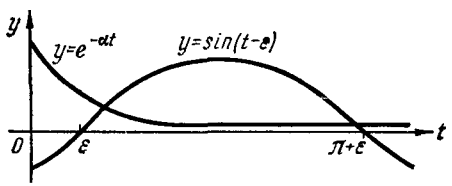

Рис. 23.

Исследуем теперь общий случай. Введем функцию $z(x) = v(x) - y(x)$. Она будет удовлетворять неравенству

$$z'' + p(x) z' + q(x) z > 0 \tag{78}$$

и начальным условиям $z(x_0) = z'(x_0) = 0$. Наряду с этим неравенством рассмотрим уравнение

$$t'' - p(x) t' - p'(x) t + q(x) t = 0 \tag{79}$$

и пусть $t(x)$ — некоторое его решение, положительное на некотором отрезке $[x_0, x_1] \in [a, b]$. Умножим неравенство (78) на $t(x)$, а уравнение (79) на $z(x)$ и вычтем из первого второе. Получим:

$$(z''t - t''z) + p(z't + zt') + p'(tz) > 0. \tag{80}$$

Но

$$z''t - t''z = (z't - t'z)', \tag{81}$$

а

$$p(z't + zt') + p'(tz) = (pzt)'. \tag{82}$$

Поэтому неравенство (80) можно записать в виде

$$(z't - zt')' + (pzt)' > 0. \tag{83}$$

Интегрируя в пределах (x_0, x), получим:

$$(z't - t'z)\Big|_{x_0}^{x} + pzt\Big|_{x_0}^{x} > 0 \tag{84}$$

или

$$z'(x)t(x) - t'(x)z(x) + p(x)z(x)t(x) > 0. \tag{85}$$

Поделив на $t(x)$, получим:

$$z'(x) - z(x)\left[\frac{t'(x)}{t(x)} - p(x)\right] > 0, \tag{86}$$

а отсюда на основании теоремы Чаплыгина о дифференциальных неравенствах для уравнений первого порядка следует, что $z(x) > 0$. Таким образом, если нам удастся найти некоторое положительное решение $t(x)$ уравнения (79), то на соответствующем отрезке $v(x) > y(x)$.

Пусть $T(x)$ — решение уравнения (79), удовлетворяющее начальным условиям $T(x_0) = 0$, $T'(x_0) = 1$. Может оказаться, что оно не пересечет нигде ось x на $[x_0, b]$. В этом случае $v(x) > y(x)$ на всем отрезке $[x_0, b]$. Пусть теперь найдется такая точка $x_1 \in [x_0, b]$, где $T(x)$ пересекает ось x. Покажем, что при этом найдется такая функция $z(x)$, что для нее будет выполнено неравенство (78) и соответствующие начальные условия, но она становится отрицательной в некоторых точках отрезка $[x_1, x_1 + \varepsilon]$, где $\varepsilon > 0$ произвольно.

Для доказательства рассмотрим решение $T_1(x)$ уравнения (79), удовлетворяющее начальным данным $T_1(x_1 + \varepsilon) = 0$, $T_1'(x_1 + \varepsilon) = -1$. В силу теоремы о разделении нулей решений линейного дифференциального уравнения второго порядка на $[x_0, x_1]$ найдется такая точка $x = \alpha$, что $T_1(\alpha) = 0$. Без уменьшения общности можно считать, что других нулей $T_1(x)$ между α и $x_2 = x_1 + \varepsilon$ нет, так как иначе мы могли бы взять за α ближайший слева от x_2 нуль $T_1(x)$ и за x_0 ближайший слева от α нуль $T(x)$. Графики $T(x)$ и $T_1(x)$ показаны на рис. 24.

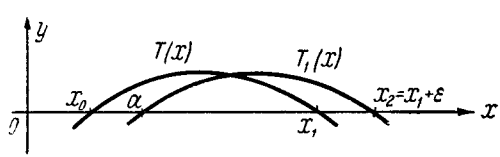

Рис. 24.

Обозначим через $h(x)$ функцию

$$z''(x) + p(x)z'(x) + q(x)z(x) = h(x). \tag{87}$$

В силу наших предположений $h(x) > 0$. Для $T_1(x)$ будем иметь:

$$T_1''(x) - p(x)T_1'(x) - p'(x)T_1(x) + q(x)T_1(x) = 0. \tag{88}$$

Как и ранее, умножим (87) на $T_1(x)$, а (88) на $z(x)$ и произведем вычитание. Получим:

$$\left(z'T_1 - zT_1'\right)' + (pzT_1)' = hT_1. \tag{89}$$

Интегрируя в пределах (x_0, x_2), получим:

$$\left(z'T_1 - zT_1'\right)\Big|_{x_0}^{x_2} + (pzT_1)\Big|_{x_0}^{x_2} = \int_{x_0}^{x_2} h(x)\, T_1(x)\, dx \tag{90}$$

или

$$z(x_2) = \int_{x_0}^{x_2} h(x)\, T_1(x)\, dx. \tag{91}$$

Подбирая в качестве $h(x)$ функцию с такими же свойствами, что и в приведенном выше примере, мы сумеем достичь того, что

$$\int_{x_0}^{x_2} h(x)\, T_1(x)\, dx < 0. \tag{92}$$

Таким образом, мы получили, что какова бы ни была функция $v(x)$, удовлетворяющая неравенству (75) и начальным условиям $v(x_0) = y_0$, $v'(x_0) = y_0'$, на интервале (x_0, x_1) всегда $v(x) > y(x)$. С другой стороны, какое бы $\varepsilon > 0$ мы ни взяли, найдется такая функция $v(x)$, для которой выполнены неравенство (75) и начальные условия $v(x_0) = y_0$; $v'(x_0) = y_0'$, но $v(x_1 + \varepsilon) < y(x_1 + \varepsilon)$.

Точка x_1 называется *порогом применимости теоремы Чаплыгина*. Порог применимости можно находить и иначе. Для этого в уравнении (79) произведем замену

$$\sigma = \frac{t'}{t} \tag{93}$$

или

$$t = \sigma t. \tag{94}$$

Такая замена законна и σ непрерывна, если $t \neq 0$. При этом

$$t'' = t(\sigma' + \sigma^2), \tag{95}$$

и уравнение (79) перейдет в

$$\sigma' + \sigma^2 - p\sigma - p' + q = 0. \tag{96}$$

Если использовать для отыскания порога применимости теоремы Чаплыгина уравнение (96) вместо (79), то мы должны найти наибольший отрезок $[x_0, x_1]$, на котором существует непрерывное решение (96).

Так как уравнение Риккати, вообще говоря, в квадратурах не интегрируется, то применение метода Чаплыгина к линейным урав-

нениям второго порядка встречает большие затруднения. Еще большие трудности возникают при использовании метода Чаплыгина для линейных уравнений высших порядков, для нелинейных уравнений высших порядков и систем уравнений. Поэтому этих вопросов мы здесь касаться не будем.

§ 3. Метод малого параметра

В ряде механических и физических задач приходится сталкиваться с дифференциальными уравнениями, содержащими некоторые параметры. При этом иногда удается найти частное решение дифференциального уравнения для некоторых фиксированных значений этих параметров, удовлетворяющее начальным условиям. Тогда пытаются разыскивать решение дифференциального уравнения в виде ряда по степеням $\lambda_1 - \lambda_1^{(0)}$, $\lambda_2 - \lambda_2^{(0)}$, ..., $\lambda_k - \lambda_k^{(0)}$, где λ_1, λ_2, ..., λ_k — входящие в уравнение параметры и $\lambda_1^{(0)}$, $\lambda_2^{(0)}$, ..., $\lambda_k^{(0)}$ — их частные значения, при которых найдено частное решение. Возможность такого разложения при некоторых предположениях о дифференциальных уравнениях была впервые изучена Пуанкаре в его книге «Новые методы небесной механики» в 1892 г. Сейчас мы приведем его теорему. Для сокращения записей ограничимся случаем системы двух уравнений и одного параметра.

Пусть нам дана система

$$\left.\begin{aligned} \frac{dY}{dx} &= F_1(x, Y, Z, \lambda), \\ \frac{dZ}{dx} &= F_2(x, Y, Z, \lambda) \end{aligned}\right\} \tag{1}$$

и требуется найти ее решение, удовлетворяющее условиям: при $x = x_0$, $Y = Y_0$, $Z = Z_0$. Предположим, что при $\lambda = \lambda_0$ мы умеем находить такое частное решение и оно будет

$$Y = \varphi(x); \quad Z = \psi(x). \tag{2}$$

Введем новые переменные и параметр следующим образом:

$$y = Y - \varphi(x); \quad z = Z - \psi(x); \quad \mu = \lambda - \lambda_0. \tag{3}$$

После подстановки в систему получим:

$$\left.\begin{aligned} \frac{dy}{dx} + \frac{d\varphi}{dx} &= F_1[x, y + \varphi(x), z + \psi(x), \mu + \lambda_0], \\ \frac{dz}{dx} + \frac{d\psi}{dx} &= F_2[x, y + \varphi(x), z + \psi(x), \mu + \lambda_0] \end{aligned}\right\} \tag{4}$$

или

$$\left.\begin{aligned} \frac{dy}{dx} &= F_1[x, y + \varphi(x), z + \psi(x), \mu + \lambda_0] - \frac{d\varphi}{dx}, \\ \frac{dz}{dx} &= F_2[x, y + \varphi(x), z + \psi(x), \mu + \lambda_0] - \frac{d\psi}{dx}. \end{aligned}\right\} \tag{5}$$

Обозначим правые части (5) соответственно через $f_1(x, y, z, \mu)$ и $f_2(x, y, z, \mu)$. Теперь нам нужно найти решение системы

$$\left.\begin{aligned} \frac{dy}{dx} &= f_1(x, y, z, \mu), \\ \frac{dz}{dx} &= f_2(x, y, z, \mu), \end{aligned}\right\} \tag{6}$$

обращающееся в нуль при $x = x_0$. При $\mu = 0$ таким решением будет $y \equiv 0$, $z \equiv 0$. Следовательно, $f_1(x, 0, 0, 0) \equiv 0$, $f_2(x, 0, 0, 0) \equiv 0$. Будем предполагать, что f_1 и f_2 являются аналитическими функциями y, z и μ:

$$\left.\begin{aligned} f_1(x, y, z, \mu) &= \sum_{i, j, k=0}^{\infty} a_{ijk} y^i z^i \mu^k, \\ f_2(x, y, z, \mu) &= \sum_{i, j, k=0}^{\infty} b_{ijk} y^i z^i \mu^k, \end{aligned}\right\} \tag{7}$$

где a_{ijk} и b_{ijk} — непрерывные функции x в некотором отрезке $[\alpha, \beta]$, содержащем точку x_0, и ряды сходятся при $|y| \leqslant \rho$, $|z| \leqslant \rho$, $|\mu| \leqslant \rho$ и любом x, принадлежащем $[\alpha, \beta]$. Так как $f_1(x, 0, 0, 0) = 0$ и $f_2(x, 0, 0, 0) = 0$, то $a_{0,0,0} = 0$, $b_{0,0,0} = 0$. Будем разыскивать решение системы (6), обращающееся в нуль при $x = x_0$, в виде

$$\left.\begin{aligned} y(x) &= \mu y_1(x) + \mu^2 y_2(x) + \ldots + \mu^n y_n(x) + \ldots, \\ z(x) &= \mu z_1(x) + \mu^2 z_2(x) + \ldots + \mu^n z_n(x) + \ldots \end{aligned}\right\} \tag{8}$$

Сначала определяем функции $y_n(x)$ и $z_n(x)$ так, чтобы ряды (8) формально удовлетворяли системе. Подставляя (8) в (6), получим:

$$\left.\begin{aligned} \sum_{n=1}^{\infty} \mu^n \frac{dy_n}{dx} &= \sum_{i, j, k=0}^{\infty} a_{ijk} \left(\sum_{l=1}^{\infty} \mu^l y_l\right)^i \left(\sum_{m=1}^{\infty} \mu^m z_m\right)^j \mu^k, \\ \sum_{n=1}^{\infty} \mu^n \frac{dz_n}{dx} &= \sum_{i, j, k=0}^{\infty} b_{ijk} \left(\sum_{l=1}^{\infty} \mu^l y_l\right)^i \left(\sum_{m=1}^{\infty} \mu^m z_m\right)^j \mu^k. \end{aligned}\right\} \tag{9}$$

Приравнивая коэффициенты при μ в левой и правой частях, находим:

$$\left.\begin{aligned} \frac{dy_1}{dx} &= a_{100} y_1 + a_{010} z_1 + a_{001}, \\ \frac{dz_1}{dx} &= b_{100} y_1 + b_{010} z_1 + b_{001}. \end{aligned}\right\} \tag{10}$$

Отыскиваем решение этой системы линейных дифференциальных уравнений, удовлетворяющее условиям $y_1(x_0) = z_1(x_0) = 0$. Затем

приравниваем коэффициенты при μ^2 левой и правой частей:

$$\left.\begin{aligned}\frac{dy_2}{dx} &= a_{100}y_2 + a_{010}z_2 + a_{002} + a_{200}y_1^2 + a_{020}z_1^2 + \\ &\quad + a_{110}y_1z_1 + a_{101}y_1 + a_{011}z_1, \\ \frac{dz_2}{dx} &= b_{100}y_2 + b_{010}z_2 + b_{002} + b_{200}y_1^2 + b_{020}z_1^2 + \\ &\quad + b_{110}y_1z_1 + b_{101}y_1 + b_{011}z_1.\end{aligned}\right\} \tag{11}$$

При этом последние слагаемые

$$\left.\begin{aligned}u_2(x) &= a_{002} + a_{200}y_1^2 + a_{020}z_1^2 + a_{110}y_1z_1 + a_{101}y_1 + a_{011}z_1, \\ v_2(x) &= b_{002} + b_{200}y_1^2 + b_{020}z_1^2 + b_{110}y_1z_1 + b_{101}y_1 + b_{011}z_1,\end{aligned}\right\} \tag{12}$$

являются уже известными функциями x. Таким образом, нам нужно найти решение системы

$$\left.\begin{aligned}\frac{dy_2}{dx} &= a_{100}y_2 + a_{010}z_2 + u_2(x), \\ \frac{dz_2}{dx} &= b_{100}y_2 + b_{010}z_2 + v_2(x),\end{aligned}\right\} \tag{13}$$

удовлетворяющее начальным условиям $y_2(x_0) = 0$, $z_2(x_0) = 0$. Эта система отличается от предыдущей только свободными членами. Вообще для определения функций $y_n(x)$ и $z_n(x)$ нам придется отыскивать решение системы

$$\left.\begin{aligned}\frac{dy_n}{dx} &= a_{100}y_n + a_{010}z_n + u_n(x), \\ \frac{dz_n}{dx} &= b_{100}y_n + b_{010}z_n + v_n(x),\end{aligned}\right\} \tag{14}$$

удовлетворяющее начальным условиям $y_n(x_0) = 0$, $z_n(x_0) = 0$, где u_n и v_n являются многочленами от $y_1, y_2, \ldots, y_{n-1}$ и $z_1, z_2, \ldots, z_{n-1}$ с коэффициентами a_{ijk} и b_{ijk}. Если нам известна фундаментальная система решений соответствующей однородной системы, то все y_n и z_n могут быть найдены с помощью квадратур.

Установим теперь сходимость полученных таким образом рядов. Обозначим через M верхнюю границу модулей функций f_1 и f_2 при $x \in [\alpha, \beta]$, $|y| \leqslant \rho$, $|z| \leqslant \rho$, $|\mu| \leqslant \rho$. Найдем оценку для модулей коэффициентов a_{ijk} и b_{ijk}. Произведем эту оценку хотя бы для a_{ijk}. Обозначим $y = r_1e^{i\varphi}$, $z = r_2e^{i\psi}$, $\mu = r_3e^{i\theta}$. Здесь $0 \leqslant r_1, r_2, r_3 < \rho$. При этом

$$f_1(x, y, z, \mu) = \sum_{k,l,m=0}^{\infty} a_{klm}r_1^k r_2^l r_3^m e^{ik\varphi}e^{il\psi}e^{im\theta}, \tag{15}$$

а

$$\bar{f}_1(x, y, z, \mu) = \sum_{k,l,m=0}^{\infty} \bar{a}_{klm}r_1^k r_2^l r_3^m e^{-ik\varphi}e^{-il\psi}e^{-im\theta}. \tag{16}$$

Перемножим (15) и (16) и проинтегрируем произведение по φ, ψ и θ в пределах от 0 до 2π. Получим:

$$\int_0^{2\pi} d\varphi \int_0^{2\pi} d\psi \int_0^{2\pi} |f|^2 d\theta =$$

$$= \sum_{k, l, m, k', l', m'=0}^{\infty} a_{klm}\bar{a}_{k'l'm'} r_1^{k+k'} r_2^{l+l'} r_3^{m+m'} \int_0^{2\pi} e^{i(k-k')\varphi} d\varphi \times$$

$$\times \int_0^{2\pi} e^{i(l-l')\psi} d\psi \int_0^{2\pi} e^{i(m-m')\theta} d\theta. \qquad (17)$$

Но

$$\int_0^{2\pi} e^{i(k-k')\varphi} d\varphi = \begin{cases} 0, & k \neq k', \\ 2\pi, & k = k'. \end{cases} \qquad (18)$$

Аналогичные результаты получатся и для остальных интегралов. Таким образом,

$$\int_0^{2\pi} d\varphi \int_0^{2\pi} d\psi \int_0^{2\pi} |f|^2 d\theta = 8\pi^3 \sum_{k, l, m=0}^{\infty} |a_{klm}|^2 r_1^{2k} r_2^{2l} r_3^{2m}. \qquad (19)$$

Заменяя $|f|^2$ на M^2, будем иметь:

$$M^2 \geqslant \sum_{k, l, m=0}^{\infty} |a_{klm}|^2 r_1^{2k} r_2^{2l} r_3^{2m}. \qquad (20)$$

Отсюда следует, что

$$M^2 \geqslant |a_{klm}|^2 r_1^{2k} r_2^{2l} r_3^{2m} \qquad (21)$$

или

$$|a_{klm}| \leqslant \frac{M}{r_1^k r_2^l r_3^m}. \qquad (22)$$

Здесь r_i могут принимать любые значения от 0 до ρ. Для доказательства сходимости мы используем мажорантные ряды. Ряд

$$\sum_{k, l, m=0}^{\infty} A_{klm} y^k z^l \mu^m \qquad (23)$$

будет называться мажорантным по отношению к ряду

$$\sum_{k, l, m=0}^{\infty} a_{klm} y^k z^l \mu^m, \qquad (24)$$

если при любых k, l, m

$$A_{klm} \geqslant |a_{klm}|. \qquad (25)$$

Рассмотрим функцию

$$\frac{M(y+z+\mu)}{\rho\left[1-\frac{y+z+\mu}{\rho}\right]}. \tag{26}$$

Коэффициент при $y^k z^l \mu^m$ в разложении этой функции по степеням y, z, μ будет положителен и больше чем

$$\frac{M}{\rho^{k+l+m}}. \tag{27}$$

В силу полученной нами оценки для коэффициентов a_{klm}, b_{klm} этот ряд будет мажорировать ряды для f_1 и f_2. Тем более, будет мажорировать их ряд для функции

$$\frac{M(y+z+\mu)\left(1+\frac{y+z+\mu}{\rho}\right)}{\rho\left[1-\frac{y+z+\mu}{\rho}\right]}. \tag{28}$$

Рассмотрим тогда вспомогательную систему уравнений

$$\frac{dY}{dx}=\frac{dZ}{dx}=\frac{M(Y+Z+\mu)\left[1+\frac{Y+Z+\mu}{\rho}\right]}{\rho\left[1-\frac{Y+Z+\mu}{\rho}\right]} \tag{29}$$

и найдем ее решение, обращающееся в нуль при $x=x_0$. Для этого применим к ней тот же метод, что и для исходной. Функции Y_1 и Z_1 должны удовлетворять системе уравнений

$$\left.\begin{aligned}\frac{dY_1}{dx}&=A_{100}Y_1+A_{010}Z_1+A_{001},\\ \frac{dZ_1}{dx}&=B_{100}Y_1+B_{010}Z_1+B_{001}\end{aligned}\right\} \tag{30}$$

и начальным данным $Y_1(x_0)=Z_1(x_0)=0$. Применяя метод последовательных приближений к системам (10) и (30), нетрудно убедиться, что

$$Y_1\geqslant|y_1|;\quad Z_1\geqslant|z_1|. \tag{31}$$

Те же рассуждения дадут

$$Y_n\geqslant|y_n|;\quad Z_n\geqslant|z_n| \tag{32}$$

при любом n. Итак, если мы докажем сходимость рядов

$$\sum_{n=1}^{\infty}\mu^n Y_n,\ \sum_{n=1}^{\infty}\mu^n Z_n, \tag{33}$$

то докажем и сходимость рядов (8). Для того чтобы убедиться в сходимости рядов (33), достаточно доказать, что решение

системы (29), обращающееся в нуль при $x = x_0$, может быть представлено в виде рядов по степеням μ.

Положим $Y + Z + \mu = \rho\Phi$. Тогда система (29) перейдет в

$$\frac{d\Phi}{dx} = \frac{2M\Phi(1+\Phi)}{\rho(1-\Phi)}. \tag{34}$$

Функция Φ при $x = x_0$ равна $\frac{\mu}{\rho}$. Решим уравнение (34) разделением переменных. Получим:

$$\frac{(1-\Phi)\,d\Phi}{\Phi(1+\Phi)} = \frac{2M\,dx}{\rho} \tag{35}$$

или

$$\left(\frac{1}{\Phi} - \frac{2}{1+\Phi}\right)d\Phi = \frac{2M}{\rho}\,dx. \tag{36}$$

Отсюда

$$\ln\Phi - 2\ln(1+\Phi) = \frac{2M}{\rho}(x - x_0) + \ln C \tag{37}$$

или

$$\frac{\Phi}{(1+\Phi)^2} = Ce^{\frac{2M}{\rho}(x-x_0)}. \tag{38}$$

Подберем C так, чтобы было выполнено начальное условие. Получим:

$$C = \left.\frac{\Phi e^{-\frac{2M}{\rho}(x-x_0)}}{(1+\Phi)^2}\right|_{x=x_0} = \frac{\mu\rho}{(\mu+\rho)^2}. \tag{39}$$

Итак,

$$\frac{\Phi}{(1+\Phi)^2} = \frac{\mu\rho}{(\mu+\rho)^2}e^{\frac{2M}{\rho}(x-x_0)}. \tag{40}$$

Обозначим

$$\frac{\Phi}{(1+\Phi)^2} = \frac{\mu\rho}{(\mu+\rho)^2}e^{\frac{2M}{\rho}(x-x_0)} = \alpha. \tag{41}$$

Тогда

$$\Phi^2 + \left(2 - \frac{1}{\alpha}\right)\Phi + 1 = 0, \tag{42}$$

или

$$\Phi = -1 + \frac{1}{2\alpha} \pm \frac{\sqrt{1-4\alpha}}{2\alpha}. \tag{43}$$

Если α мало, то

$$\Phi = -1 + \frac{1}{2\alpha} \pm \frac{1 - 2\alpha - 2\alpha^2 - \ldots}{2\alpha} \tag{44}$$

Взяв здесь знак минус (чтобы Φ удовлетворяла начальным условиям) перед второй дробью, получим:

$$\Phi = +\alpha + \alpha^2 C_1 + \ldots \tag{45}$$

Таким образом, Ф, а следовательно, и Y и Z представимы в виде ряда по степеням α, если только $|\alpha| < \frac{1}{4}$, т. е. если

$$\left| e^{\frac{2M(x-x_0)}{\rho}} \frac{\mu\rho}{(\mu+\rho)^2} \right| < \frac{1}{4}, \tag{46}$$

а этого всегда можно достигнуть выбором достаточно малого μ. В свою очередь α может быть представлено в виде ряда по степеням μ. Тем самым доказана сходимость рядов (33), а следовательно и рядов (8). Уравнения (29) показывают, что будут сходиться ряды и для производных. Этим мы и закончим доказательство теоремы Пуанкаре.

Изложенный метод применим и к уравнениям высших порядков, содержащим параметр.

Иногда, если само уравнение не содержит параметра, его можно ввести искусственно. При этом обычно параметр вставляют так, чтобы при нулевом значении его решение получалось без труда. Тогда находят решение в виде ряда по степеням параметра и затем дают параметру такое значение, при котором получается исходное уравнение.

В качестве примера рассмотрим уравнение

$$\frac{d^2y}{dx^2} + m^2y = f(x)\,y. \tag{47}$$

Вместо уравнения (47) возьмем

$$\frac{d^2Y}{dx^2} + m^2Y = \varepsilon f(x)\,Y. \tag{48}$$

Очевидно, уравнение (47) получится из уравнения (48) при $\varepsilon = 1$. Будем искать решение уравнения (48) в виде

$$Y = y_0 + \varepsilon y_1 + \varepsilon^2 y_2 + \ldots \tag{49}$$

Подстановка (49) в (48) даст

$$y_0'' + \varepsilon y_1'' + \varepsilon^2 y_2'' + \ldots + m^2 y_0 + m^2\varepsilon y_1 + m^2\varepsilon^2 y_2 + \ldots$$
$$\ldots = \varepsilon f(x)\,[y_0 + \varepsilon y_1 + \varepsilon^2 y_2 + \ldots]. \tag{50}$$

Для определения y_0 получим уравнение

$$y_0'' + m^2 y_0 = 0. \tag{51}$$

Его общее решение имеет вид

$$y_0 = A\cos mx + B\sin mx. \tag{52}$$

Уравнением для определения y_n будет

$$y_n'' + m^2 y_n = f(x)\,y_{n-1}. \tag{53}$$

Таким образом, мы можем последовательно находить все y_n при помощи квадратур. Так, если начальные данные будут заданы при $x=0$, то функции y_n могут быть последовательно определены по формулам

$$y_n = \frac{1}{m}\int_0^x f(t)\, y_{n-1}(t) \sin m(x-t)\, dt. \tag{54}$$

Получив решение Y в виде ряда (49), мы затем полагаем $\varepsilon = 1$. Конечно, мы должны предварительно убедиться в том, что ряд (49) сходится при $\varepsilon = 1$.

Фактически используют не полные ряды, а лишь их отрезки. Поэтому нужна дополнительная оценка величины отброшенных членов.

Метод малого параметра часто используется в теории нелинейных колебаний. Не останавливаясь здесь на теоретических вопросах, связанных с существованием периодического решения и возможностью его представления в виде ряда по степеням параметра, рассмотрим один пример. Пусть дано уравнение

$$\ddot{x} + \alpha x + \beta x^2 = F\cos\omega t. \tag{55}$$

Будем разыскивать периодическое решение этого уравнения, имеющее ту же частоту, что и $F\cos\omega t$. Произведем замену переменных $\theta = \omega t$. При этом уравнение (55) перейдет в

$$\omega^2 \frac{d^2x}{d\theta^2} + (\alpha x + \beta x^2) = F\cos\theta. \tag{56}$$

Потребуем, чтобы функция $x(\theta)$ удовлетворяла следующим условиям:

$$\left.\begin{aligned} x(\theta + 2\pi) &= x(\theta), \\ x(0) &= A, \\ x'(0) &= 0. \end{aligned}\right\} \tag{57}$$

Здесь штрихом отмечена производная по θ.

В качестве параметра выберем β и будем разыскивать $x(\theta)$ и ω в виде рядов по степеням β:

$$\left.\begin{aligned} x(\theta) &= x_0(\theta) + \beta x_1(\theta) + \beta^2 x_2(\theta) + \ldots, \\ \omega &= \omega_0 + \beta\omega_1 + \beta^2\omega_2 + \ldots \end{aligned}\right\} \tag{58}$$

От функций $x_i(\theta)$ потребуем

$$\left.\begin{aligned} &x_i(\theta + 2\pi) = x_i(\theta), \\ &x_0(0) = A;\ x_i(0) = 0, \quad (i > 0). \\ &x_0'(0) = 0; \quad x_i'(0) = 0 \end{aligned}\right\} \tag{59}$$

Для упрощения полагаем также, что амплитуда F мала:

$$F = \beta F_0. \tag{60}$$

Подставляя (58) и (60) в (56), получим:

$$(\omega_0^2 + 2\beta\omega_0\omega_1 + \ldots)(x_0'' + \beta x_1'' + \ldots) + \alpha(x_0 + \beta x_1 + \ldots) + \\ + \beta(x_0^2 + 2\beta x_0 x_1 + \ldots) - \beta F_0 \cos\theta = 0. \tag{61}$$

Приравнивая нулю члены, не содержащие β, находим:

$$\omega_0^2 x_0'' + \alpha x_0 = 0. \tag{62}$$

Общее решение этого уравнения имеет вид

$$x_0 = A_0 \cos\frac{\sqrt{\alpha}}{\omega_0}\theta + B_0 \sin\frac{\sqrt{\alpha}}{\omega_0}\theta. \tag{63}$$

Из условий (59) получаем:

$$\omega_0 = \sqrt{\alpha}; \quad A_0 = A; \quad B_0 = 0. \tag{64}$$

Итак,

$$x_0 = A\cos\theta; \quad \omega_0 = \sqrt{\alpha}. \tag{65}$$

Приравняем теперь нулю члены, содержащие β в первой степени. Получим:

$$\omega_0^2 x_1'' + \alpha x_1 = -2\omega_0\omega_1 x_0'' - x_0^2 + F_0\cos\theta, \tag{66}$$

или после подстановки вместо x_0 его значения

$$\omega_0^2 x_1'' + \alpha x_1 = \left(2\omega_0\omega_1 A - \frac{1}{2}A^2 + F_0\right)\cos\theta - \frac{1}{2}A^2\cos 2\theta. \tag{67}$$

Так как x_1 должна быть периодической функцией, то член с $\cos\theta$ в правой части должен обратиться в нуль. Поэтому

$$2\omega_0\omega_1 A - \frac{1}{2}A^2 + F_0 = 0 \tag{68}$$

или

$$\omega_1 = \frac{1}{2\sqrt{\alpha}}\left(\frac{1}{2}A - \frac{F_0}{A}\right). \tag{69}$$

Таким образом, общее решение уравнения (67) примет вид

$$x_1 = A_1\cos\theta + B_1\sin\theta + \frac{A^2}{6\alpha}\cos 2\theta. \tag{70}$$

Условия (59) дадут

$$A_1 = -\frac{A^2}{6\alpha}; \quad B_1 = 0. \tag{71}$$

Итак,

$$x_1 = \frac{A^2}{6\alpha}(-\cos\theta + \cos 2\theta). \tag{72}$$

Продолжая процесс, найдем:

$$\left.\begin{aligned} x &= A\cos\theta + \frac{\beta A^2}{6\alpha}(-\cos\theta + \cos 2\theta) + \ldots, \\ \omega &= \sqrt{\alpha} + \beta\frac{1}{2\sqrt{\alpha}}\left(\frac{1}{2}A - \frac{F_0}{A}\right) + \ldots \end{aligned}\right\} \quad (73)$$

§ 4. Метод Рунге — Кутта

Методы, рассмотренные в предыдущих параграфах, давали приближенное представление решения в аналитической форме. Как мы видели, их применение связано с выполнением большого числа интегрирований, а это не всегда может быть осуществлено практически. Перейдем теперь к изучению численных методов, позволяющих получить таблицу значений решения. Мы начнем с метода, предложенного Рунге и усовершенствованного Кутта и другими математиками.

1. Метод Рунге—Кутта решения дифференциальных уравнений первого порядка. Пусть нам требуется найти решение дифференциального уравнения

$$y' = f(x, y), \quad (1)$$

удовлетворяющее начальному условию $y = y_0$ при $x = x_0$. Будем предполагать, что в рассматриваемой области $f(x, y)$ имеет непрерывные частные производные до некоторого порядка n. Тогда искомое решение будет иметь непрерывные производные до порядка $n+1$, и мы можем записать:

$$\Delta y_0 = y(x) - y_0 = (x - x_0)y_0' + \frac{(x-x_0)^2}{2!}y_0'' + \ldots$$

$$\ldots + \frac{(x-x_0)^{n+1}}{(n+1)!}y_0^{(n+1)} + o\left(|x - x_0|^{n+1}\right). \quad (2)$$

Обозначим $x - x_0 = h$. При достаточно малом h мы можем отбросить в (2) член $o\left(|x - x_0|^{n+1}\right)$ и приближенно считать

$$\Delta y_0 = y(x_0 + h) - y_0 = hy_0' + \frac{h^2}{2}y_0'' + \ldots + \frac{h^{n+1}}{(n+1)!}y_0^{(n+1)}. \quad (3)$$

Может оказаться, что для получения Δy_0 с нужной нам точностью не требуется использовать все члены (3).

Производные, входящие в правую часть (3), могут быть фактически найдены. Так,

$$y_0' = f(x_0, y_0) = f_0. \quad (4)$$

Далее,

$$y_0'' = \frac{\partial f_0}{\partial x} + f_0\frac{\partial f_0}{\partial y}. \quad (5)$$

Чтобы запись последующих производных была менее громоздкой, введем операторы

$$A_m(u) = \left(\frac{\partial}{\partial x} + f\frac{\partial}{\partial y}\right)^m u = \sum_{k=0}^{m} C_m^k f^k \frac{\partial^m u}{\partial x^{m-k}\,\partial y^k}. \tag{6}$$

Для этих операторов будут справедливы следующие равенства:

$$\left.\begin{aligned} A_m(u+v) &= A_m(u) + A_m(v), \\ A_1(uv) &= A_1(u)\,v + uA_1(v). \end{aligned}\right\} \tag{7}$$

Заметим, далее, что

$$A_1[A_m(u)] = A_{m+1}(u) + mA_1(f)\,A_{m-1}\left(\frac{\partial u}{\partial y}\right). \tag{8}$$

Действительно,

$$A_1[A_m(u)] = A_1\left[\sum_{k=0}^{m} C_m^k f^k \frac{\partial^m u}{\partial x^{m-k}\,\partial y^k}\right] = \sum_{k=0}^{m} C_m^k A_1\left[f^k \frac{\partial^m u}{\partial x^{m-k}\,\partial y^k}\right] =$$

$$= \sum_{k=1}^{m} C_m^k \frac{\partial^m u}{\partial x^{m-k}\,\partial y^k} A_1(f^k) + \sum_{k=0}^{m} C_m^k f^k A_1\left[\frac{\partial^m u}{\partial x^{m-k}\,\partial y^k}\right]. \tag{9}$$

Но

$$\sum_{k=1}^{m} C_m^k \frac{\partial^m u}{\partial x^{m-k}\,\partial y^k} A_1(f^k) = A_1(f) \sum_{k=1}^{m} C_m^k k f^{k-1} \frac{\partial^m u}{\partial x^{m-k}\,\partial y^k} =$$

$$= mA_1(f) \sum_{k=0}^{m-1} C_{m-1}^k f^k \frac{\partial^{m-1} u}{\partial x^{m-1-k}\partial y^k}\left(\frac{\partial u}{\partial y}\right) = m \cdot A_1(f) \cdot A_{m-1}\left(\frac{\partial u}{\partial y}\right) \tag{10}$$

и

$$\sum_{k=0}^{m} C_m^k f^k A_1\left[\frac{\partial^m u}{\partial x^{m-k}\,\partial y^k}\right] =$$

$$= \sum_{k=0}^{m} C_m^k f^k \left[\frac{\partial^{m+1} u}{\partial x^{m+1-k}\,\partial y^k} + f\,\frac{\partial^{m+1} u}{\partial x^{m-k}\,\partial y^{k+1}}\right] =$$

$$= \sum_{k=0}^{m} C_m^k f^k \frac{\partial^{m+1} u}{\partial x^{m+1-k}\,\partial y^k} + \sum_{k=0}^{m} C_m^k f^{k+1} \frac{\partial^{m+1} u}{\partial x^{m-k}\,\partial y^{k+1}} =$$

$$= \sum_{k=0}^{m+1} \left(C_m^k + C_m^{k-1}\right) f^k \frac{\partial^{m+1} u}{\partial x^{m+1-k}\,\partial y^k} = A_{m+1}(u), \tag{11}$$

что и требовалось доказать. Применение оператора A_1 к функции $u(x, y)$ эквивалентно дифференцированию этой функции по x в предположении, что y является решением дифференциального уравнения (1). Таким образом, последовательно дифференцируя (1), мы получим:

$$\left.\begin{aligned}
&y'' = A_1(f),\\
&y''' = A_1[A_1(f)] = A_2(f) + A_1(f)\frac{\partial f}{\partial y},\\
&y^{(IV)} = A_1\left[A_2(f) + A_1(f)\frac{\partial f}{\partial y}\right] =\\
&\quad = A_3(f) + 3A_1(f)\,A_1\left(\frac{\partial f}{\partial y}\right) + A_2(f)\frac{\partial f}{\partial y} + A_1(f)\left(\frac{\partial f}{\partial y}\right)^2,\\
&y^{(V)} = A_4(f) + A_3(f)\frac{\partial f}{\partial y} + 4A_2(f)\cdot A_1\left(\frac{\partial f}{\partial y}\right) + 6A_1(f)\,A_2\left(\frac{\partial f}{\partial y}\right) +\\
&\quad + 7A_1(f)\cdot A_1\left(\frac{\partial f}{\partial y}\right)\frac{\partial f}{\partial y} + 3[A_1(f)]^2\frac{\partial^2 f}{\partial y^2} +\\
&\quad + A_2(f)\left(\frac{\partial f}{\partial y}\right)^2 + A_1(f)\left(\frac{\partial f}{\partial y}\right)^3,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned}\right\} \qquad (12)$$

С увеличением порядка выражения для производных становится все более и более громоздкими, даже при операторной записи. Рассмотрим еще операторы

$$\bar{A}_m(u) = \left(\frac{\partial}{\partial x} + f_0\frac{\partial}{\partial y}\right)^m u = \sum_{k=0}^{m} C_m^k f_0^k \frac{\partial^m u}{\partial x^{m-k}\,\partial y^k} \quad (f_0 = \text{const}). \qquad (13)$$

Для них также выполнены равенства (7), а (8) переходит в

$$\bar{A}_1[\bar{A}_m(u)] = \bar{A}_{m+1}(u). \qquad (14)$$

Таким образом,

$$\bar{A}_m = [\bar{A}_1]^m. \qquad (15)$$

Производные $y_0^{(i)}$ мы получим, если в (12) заменим f на f_0, а A_j на $\bar{A}_j$.

Мы убедились, что производные $y_0^{(j)}$, входящие в (3), могут быть фактически вычислены. Но в связи с тем, что формулы (12) очень громоздки, их непосредственное использование в (3) для вычисления Δy_0 вряд ли может оказаться полезным на практике.

Рунге предложил вместо этого составлять линейную комбинацию

$$p_{r1}k_1(h)+p_{r2}k_2(h)+\ldots+p_{rr}k_r(h) \tag{16}$$

с постоянными коэффициентами p_{ri} некоторых функций

$$k_i(h)=hf(\xi_i, \eta_i) \qquad (i=1, 2, \ldots, r), \tag{17}$$

где

$$\xi_i=x_0+\alpha_i h; \qquad \eta_i=y_0+\beta_{i1}k_1(h)+\beta_{i2}k_2(h)+\ldots+\beta_{i,\,i-1}k_{i-1}(h) \tag{18}$$

и α_i, β_{ij} — постоянные, $\alpha_1=0$. Таким образом,

$$\left.\begin{aligned} &k_1(h)=hf(x_0, y_0),\\ &k_2(h)=hf(x_0+\alpha_2 h, y_0+\beta_{21}k_1),\\ &k_3(h)=hf(x_0+\alpha_3 h, y_0+\beta_{31}k_1+\beta_{32}k_2),\\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ &k_r(h)=hf(x_0+\alpha_r h, y_0+\beta_{r1}k_1+\beta_{r2}k_2+\ldots+\beta_{r,\,r-1}k_{r-1}). \end{aligned}\right\} \tag{19}$$

Зная α_i, β_{ij} и выбрав шаг h, мы можем последовательно вычислить все $k_i(h)$.

Выбор постоянных p_{ri}, α_i, β_{ij} производится так, чтобы разложение (3) и разложение (16) по степеням h совпадали до возможно более высоких степеней h при произвольной функции $f(x, y)$ и произвольном шаге h. При этом функция

$$\varphi_r(h)=y(x_0+h)-y_0-[p_{r1}k_1(h)+p_{r2}k_2(h)+\ldots+p_{rr}k_r(h)] \tag{20}$$

будет обладать свойствами

$$\varphi_r(0)=\varphi_r'(0)=\varphi_r''(0)=\ldots=\varphi_r^{(s)}(0)=0, \quad \varphi_r^{(s+1)}(0)\neq 0, \tag{21}$$

и мы должны подобрать p_{ri}, α_i, β_{ij} так, чтобы s было возможно большим для произвольных h и $f(x, y)$.

Разность между точным значением Δy_0 при продвижении на один шаг длиной h и приближенным значением его, вычисленным по формуле

$$\Delta y_0=p_{r1}k_1+p_{r2}k_2+\ldots+p_{rr}k_r, \tag{22}$$

т. е. погрешность метода на одном шаге, будет равна

$$R_r(h)=\frac{h^{s+1}\varphi_r^{(s+1)}(\xi)}{(s+1)!}, \tag{23}$$

где $0\leqslant\xi\leqslant h$.

Условие $\varphi_r(0)=0$ в нашем случае будет выполнено всегда. Условие $\varphi_r^{(l)}(0)=0$ означает

$$y_0^{(l)}=p_{r1}k_1^{(l)}(0)+p_{r2}k_2^{(l)}(0)+\ \ldots\ +p_{rr}k_r^{(l)}(0). \tag{24}$$

Производная, стоящая в левой части равенства, может быть всегда вычислена указанным ранее способом. Займемся сейчас вычислением производных $k_i^{(l)}(0)$.

Для $k_1(h)$ будем иметь:

$$k_1'(h)=f(x_0,\ y_0),\qquad k_1^{(l)}(h)\equiv 0\quad\text{при}\quad l\geqslant 2. \tag{25}$$

Для $k_i(h)$ при $i>1$ получим:

$$\begin{aligned}k_i'(h)&=f(\xi_i,\ \eta_i)+h\left[\alpha_i\frac{\partial}{\partial x}+\eta_i'(h)\frac{\partial}{\partial y}\right]f(\xi_i,\ \eta_i)=\\&=f(\xi_i,\ \eta_i)+h\left[\alpha_i\frac{\partial}{\partial x}+(\beta_{i1}k_1'+\beta_{i2}k_2'+\ \ldots\right.\\&\qquad\left.\ldots+\beta_{i,\,i-1}k_{i-1}')\frac{\partial}{\partial y}\right]f(\xi_i,\ \eta_i).\end{aligned} \tag{26}$$

Чтобы сделать выражения менее громоздкими, снова введем операторную запись. Обозначим

$$\begin{aligned}B_m^{(i)}[\varphi(\xi_i,\ \eta_i)]&=\left[\alpha_i\frac{\partial}{\partial x}+\eta_i'\frac{\partial}{\partial y}\right]^m\varphi(\xi_i,\ \eta_i)=\\&=\sum_{k=0}^{m}C_m^k\alpha_i^{m-k}\eta_i'^{k}\frac{\partial^m\varphi(\xi_i,\ \eta_i)}{\partial x^{m-k}\,\partial y^k}.\end{aligned} \tag{27}$$

Операторы $B_m^{(i)}$ обладают следующими свойствами:

$$\left.\begin{aligned}B_m^{(i)}[\varphi(\xi_i,\ \eta_i)+\psi(\xi_i,\ \eta_i)]&=B_m^{(i)}[\varphi(\xi_i,\ \eta_i)]+B_m^{(i)}[\psi(\xi_i,\ \eta_i)],\\B_1^{(i)}[\varphi(\xi_i,\ \eta_i)\,\psi(\xi_i,\ \eta_i)]&=B_1^{(i)}[\varphi(\xi_i,\ \eta_i)]\,\psi(\xi_i,\ \eta_i)+\\&\quad+\varphi(\xi_i,\ \eta_i)\,B_1^{(i)}[\psi(\xi_i,\ \eta_i)].\end{aligned}\right\} \tag{28}$$

Применение оператора $B_1^{(i)}$ к некоторой функции от ξ_i и η_i эквивалентно дифференцированию этой функции по h. Заметим, что

$$\begin{aligned}\frac{d}{dh}B_m^{(i)}[\varphi(\xi_i,\ \eta_i)]&=\sum_{k=0}^{m}C_m^k\alpha_i^{m-k}\frac{d}{dh}\left[\eta_i'^{k}\frac{\partial^m\varphi(\xi_i,\ \eta_i)}{\partial x^{m-k}\,\partial y^k}\right]=\\&=\sum_{k=1}^{m}C_m^k\alpha_i^{m-k}k\eta_i'^{k-1}\eta_i''\frac{\partial^m\varphi(\xi_i,\ \eta_i)}{\partial x^{m-k}\,\partial y^k}+B_{m+1}^{(i)}[\varphi(\xi_i,\ \eta_i)]=\\&=B_{m+1}^{(i)}[\varphi(\xi_i,\ \eta_i)]+m\eta_i''B_{m-1}^{(i)}\left[\frac{\partial\varphi(\xi_i,\ \eta_i)}{\partial y}\right].\end{aligned} \tag{29}$$

Применяя к (26) правила (28) и (29), последовательно получим:

$$k_i'(h) = f(\xi_i,\ \eta_i) + hB_1^{(i)}[f(\xi_i,\ \eta_i)], \tag{30}$$

$$k_i''(h) = 2B_1^{(i)}[f(\xi_i,\ \eta_i)] + hB_2^{(i)}[f(\xi_i,\ \eta_i)] + h\eta_i'' \frac{\partial f(\xi_i,\ \eta_i)}{\partial y}, \tag{31}$$

$$k_i'''(h) = 3B_2^{(i)}[f(\xi_i,\ \eta_i)] + 3\eta_i'' \frac{\partial f(\xi_i,\ \eta_i)}{\partial y} + hB_3^{(i)}[f(\xi_i,\ \eta_i)] + \\ + 3h\eta_i'' B_1^{(i)}\left[\frac{\partial f(\xi_i,\ \eta_i)}{\partial y}\right] + h\eta_i''' \frac{\partial f(\xi_i,\ \eta_i)}{\partial y}, \tag{32}$$

$$k_i^{(\mathrm{IV})}(h) = 4B_3^{(i)}[f(\xi_i,\ \eta_i)] + 12\eta_i'' B_1^{(i)}\left[\frac{\partial f(\xi_i,\ \eta_i)}{\partial y}\right] + 4\eta_i''' \frac{\partial f(\xi_i,\ \eta_i)}{\partial y} + \\ + hB_4^{(i)}[f(\xi_i,\ \eta_i)] + 6h\eta_i'' B_2^{(i)}\left[\frac{\partial f(\xi_i,\ \eta_i)}{\partial y}\right] + 4h\eta_i''' B_1^{(i)}\left[\frac{\partial f(\xi_i,\ \eta_i)}{\partial y}\right] + \\ + 3h(\eta_i'')^2 \frac{\partial^2 f(\xi_i,\ \eta_i)}{\partial y^2} + h\eta_i^{(\mathrm{IV})} \frac{\partial f(\xi_i,\ \eta_i)}{\partial y}, \tag{33}$$

$$k_i^{(\mathrm{V})}(h) = 5B_4^{(i)}[f(\xi_i,\ \eta_i)] + 30\eta_i'' B_2^{(i)}\left[\frac{\partial f(\xi_i,\ \eta_i)}{\partial y}\right] + \\ + 20\eta_i''' B_1^{(i)}\left[\frac{\partial f(\xi_i,\ \eta_i)}{\partial y}\right] + 15(\eta_i'')^2 \frac{\partial^2 f(\xi_i,\ \eta_i)}{\partial y^2} + 5\eta_i^{(\mathrm{IV})} \frac{\partial f(\xi_i,\ \eta_i)}{\partial y} + \\ + hB_5^{(i)}[f(\xi_i,\ \eta_i)] + 10h\eta_i'' B_3^{(i)}\left[\frac{\partial f(\xi_i,\ \eta_i)}{\partial y}\right] + 10h\eta_i''' B_2^{(i)}\left[\frac{\partial f(\xi_i,\ \eta_i)}{\partial y}\right] + \\ + 15h(\eta_i'')^2 B_1^{(i)}\left[\frac{\partial^2 f(\xi_i,\ \eta_i)}{\partial y^2}\right] + 5h\eta_i^{(\mathrm{IV})} B_1^{(i)}\left[\frac{\partial f(\xi_i,\ \eta_i)}{\partial y}\right] + \\ + 10h\eta_i''\eta_i''' \frac{\partial^2 f(\xi_i,\ \eta_i)}{\partial y^2} + h\eta_i^{(\mathrm{V})} \frac{\partial f(\xi_i,\ \eta_i)}{\partial y}, \tag{34}$$

. . . . . .

И в этом случае наряду с операторами $B_m^{(i)}$ введем операторы

$$\overline{B}_m^{(i)}[\varphi(x,\ y)] = \sum_{k=0}^{m} C_m^k \alpha_i^{m-k} [\eta_i'(0)]^k \frac{\partial^m \varphi(x,\ y)}{\partial x^{m-k}\, \partial y^k}. \tag{35}$$

Для операторов $\overline{B}_m^{(i)}$ будут справедливы соотношения (28), а равенство (29) заменится на

$$\overline{B}_1^{(i)}\left\{\overline{B}_m^{(i)}[\varphi(x,\ y)]\right\} = \overline{B}_{m+1}^{(i)}[\varphi(x,\ y)]. \tag{36}$$

Таким образом,

$$\overline{B}_n^{(i)} = [\overline{B}_1^{(i)}]^n. \tag{37}$$

Итак, при $h=0$ мы получим:

$$\left.\begin{aligned}
&k_i'(0)=f(x_0,\ y_0)=f_0,\\
&k_i''(0)=2\overline{B}_1^{(i)}[f_0],\\
&k_i'''(0)=3\overline{B}_2^{(i)}[f_0]+3\eta_i''(0)\frac{\partial f_0}{\partial y},\\
&k_i^{(\mathrm{IV})}(0)=4\overline{B}_3^{(i)}[f_0]+12\eta_i''(0)\overline{B}_1^{(i)}\left[\frac{\partial f_0}{\partial y}\right]+4\eta_i'''(0)\frac{\partial f_0}{\partial y},\\
&k_i^{(\mathrm{V})}(0)=5\overline{B}_4^{(i)}[f_0]+30\eta_i''(0)\overline{B}_2^{(i)}\left[\frac{\partial f_0}{\partial y}\right]+20\eta_i'''(0)\overline{B}_1^{(i)}\left[\frac{\partial f_0}{\partial y}\right]+\\
&\qquad\qquad+15[\eta_i''(0)]^2\frac{\partial^2 f_0}{\partial y^2}+5\eta_i^{(\mathrm{IV})}(0)\frac{\partial f_0}{\partial y},\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned}\right\}\qquad(38)$$

Рассмотрим теперь некоторые частные случаи.

1-й случай: $r=1$. При этом

$$\varphi_1(h)=y(x_0+h)-y(x_0)-p_{11}hf(x_0,\ y_0),\qquad(39)$$

$$\varphi_1'(h)=y'(x_0+h)-p_{11}f(x_0,\ y_0).\qquad(40)$$

При $h=0$ получим:

$$\varphi_1'(0)=y_0'-p_{11}f(x_0,\ y_0),\qquad(41)$$

и $\varphi_1'(0)=0$ для произвольной f тогда и только тогда, когда $p_{11}=1$. Далее,

$$\varphi_1''(0)=y_0''\qquad(42)$$

и, вообще говоря, в нуль не обращается. Таким образом, приближенная формула

$$\Delta y_0\approx hf(x_0,\ y_0)\qquad(43)$$

имеет ошибку метода на одном шаге, равную

$$R_1(h)=\frac{h^2}{2}y''(\xi)=\frac{h^2}{2}A_1(f)\Big|_{x=\xi}\qquad(x_0\leqslant\xi\leqslant x_0+h).\qquad(44)$$

Говорят, что в этом случае погрешность метода на одном шаге имеет порядок h^2. Интегрирование дифференциального уравнения по формуле (43) называют иногда *способом Эйлера*.

2-й случай: $r=2$. При этом

$$\varphi_2(h)=y(x_0+h)-y_0-[p_{21}k_1(h)+p_{22}k_2(h)],\qquad(45)$$

$$\varphi_2'(0)=y_0'-[p_{21}k_1'(0)+p_{22}k_2'(0)]=f_0-[p_{21}f_0+p_{22}f_0].\qquad(46)$$

Таким образом, $\varphi_2'(0) = 0$ для произвольной f в том и только в том случае, если

$$p_{21} + p_{22} = 1. \tag{47}$$

Далее,

$$\varphi_2''(0) = y_0'' - [p_{21}k_1''(0) + p_{22}k_2''(0)] = \overline{A}_1(f_0) - p_{22}2\overline{B}_1^{(2)}(f_0). \tag{48}$$

Необходимым и достаточным условием обращения $\varphi_2''(0)$ в нуль будет

$$\overline{A}_1 = 2p_{22}\overline{B}_1^{(2)}, \tag{49}$$

т. е.

$$1 = 2p_{22}\alpha_2, \qquad 1 = 2p_{22}\beta_{21}. \tag{50}$$

Третья производная $\varphi'''(0)$ будет равна

$$\varphi'''(0) = \overline{A}_2(f_0) + \overline{A}_1(f_0)\frac{\partial f_0}{\partial y} - 3p_{22}\overline{B}_2^{(2)}(f_0) \tag{51}$$

и, вообще говоря, в нуль не обращается. Таким образом, беря p_{21}, p_{22}, α_2, β_{21}, удовлетворяющие условиям

$$\left.\begin{aligned} p_{21} + p_{22} &= 1, \\ p_{22}\alpha_2 &= \frac{1}{2}, \\ p_{22}\beta_{21} &= \frac{1}{2}, \end{aligned}\right\} \tag{52}$$

мы получим формулы, имеющие порядок ошибки h^3. Из (52) следует, что $p_{22} \neq 0$, $\alpha_2 \neq 0$, $\beta_{21} \neq 0$ и $\alpha_2 = \beta_{21}$. Равенства (52) являются системой трех уравнений относительно четырех неизвестных. Эта система имеет бесчисленное множество решений. Каждое решение даст формулу, имеющую порядок ошибки h^3. На практике следует выбирать такие решения (52), которые дают удобные для вычислений формулы. Можно, например, взять $\alpha_2 = \beta_{21} = 1$. Тогда $p_{22} = p_{21} = \frac{1}{2}$, и получаем приближенную формулу

$$\Delta y_0 \approx \frac{1}{2}(k_1 + k_2); \quad k_1 = hf(x_0, y_0); \quad k_2 = hf(x_0 + h, y_0 + k_1). \tag{53}$$

Возьмем еще вариант: $\alpha_2 = \beta_{21} = \frac{1}{2}$. Тогда $p_{22} = 1$, $p_{21} = 0$, и получим формулу

$$\Delta y_0 \approx k_2; \quad k_1 = hf(x_0, y_0); \quad k_2 = hf\left(x_0 + \frac{h}{2}, \quad y_0 + \frac{k_1}{2}\right). \tag{54}$$

Можно также подбирать p_{21}, p_{22}, α_2, β_{21} так, чтобы в (51) сократилась часть членов. Так, например, если потребовать, чтобы

$$\overline{A}_2(f_0) = 3p_{22}\overline{B}_2^{(2)}(f_0), \tag{55}$$

то, используя (15), (36) и (49), найдем $4p_{22}^2 = 3p_{22}$ или $p_{22} = \frac{3}{4}$. При этом $\alpha_2 = \beta_{21} = \frac{2}{3}$ и $p_{21} = \frac{1}{4}$. Получаем формулу

$$\left.\begin{aligned} \Delta y_0 \approx \tfrac{1}{4} k_1 + \tfrac{3}{4} k_2; \quad & k_1 = hf(x_0, y_0); \\ k_2 = hf\left(x_0 + \tfrac{2}{3} h, \quad & y_0 + \tfrac{2}{3} k_1\right). \end{aligned}\right\} \tag{56}$$

Остаточный член при выбранных p_{21}, p_{22}, α_2, β_{21}, удовлетворяющих системе (52), оценивается по общей формуле (23), причем в нашем случае

$$\varphi_2'''(h) = y'''(x_0 + h) - [p_{21}k_1'''(h) + p_{22}k_2'''(h)] =$$
$$= A_2\{f(x_0 + h,\ y(x_0 + h))\} + A_1\{f(x_0 + h,\ y(x_0 + h))\} \times$$
$$\times \frac{\partial f(x_0 + h,\ y(x_0 + h))}{\partial y} - p_{22}\{3\overline{B}_2^{(2)}[f(\xi_2,\ \eta_2)] + h\overline{B}_3^{(2)}[f(\xi_2,\ \eta_2)]\}. \tag{57}$$

3-й случай: $r = 3$. В этом случае

$$\varphi_3'(0) = y_0' - [p_{31}k_1'(0) + p_{32}k_2'(0) + p_{33}k_3'(0)] =$$
$$= y_0' - [p_{31} + p_{32} + p_{33}]f_0. \tag{58}$$

Отсюда

$$p_{31} + p_{32} + p_{33} = 1. \tag{59}$$

Далее,

$$\varphi_3''(0) = y_0'' - [p_{31}k_1''(0) + p_{32}k_2''(0) + p_{33}k_3''(0)] =$$
$$= \overline{A}_1(f_0) - [2p_{32}\overline{B}_1^{(2)}(f_0) + 2p_{33}\overline{B}_1^{(3)}(f_0)]. \tag{60}$$

Чтобы $\varphi_3''(0)$ обращалась в нуль при произвольной f, необходимо и достаточно, чтобы

$$\overline{A}_1 = 2p_{32}\overline{B}_1^{(2)} + 2p_{33}\overline{B}_1^{(3)} \tag{61}$$

или

$$\left.\begin{aligned} 1 &= 2p_{32}\alpha_2 + 2p_{33}\alpha_3, \\ 1 &= 2p_{32}\beta_{21} + 2p_{33}(\beta_{31} + \beta_{32}). \end{aligned}\right\} \tag{62}$$

Равенство нулю третьей производной даст

$$y_0''' - [p_{31}k_1'''(0) + p_{32}k_2'''(0) + p_{33}k_3'''(0)] =$$
$$= \overline{A}_2(f_0) + \overline{A}_1(f_0)\frac{\partial f_0}{\partial y} - \left[3p_{32}\overline{B}_2(f_0) + 3p_{32}\eta_2''(0)\frac{\partial f_0}{\partial y} + \right.$$
$$\left. + 3p_{33}\overline{B}_2^{(3)}(f_0) + 3p_{33}\eta_3''(0)\frac{\partial f_0}{\partial y}\right] = 0. \tag{63}$$

Но $\eta_2(h) = y_0 + \beta_{21} k_1(h)$ и $\eta_2''(h) \equiv 0$. Далее, $\eta_3(h) = y_0 + \beta_{31} k_1(h) + \beta_{32} k_2(h)$ и, следовательно,

$$\eta_3''(0) = \beta_{32} k_2''(0) = 2\beta_{32} \overline{B}_1^{(2)}(f_0). \tag{64}$$

Поэтому равенство (63) можно записать в виде

$$\overline{A}_2(f_0) + \overline{A}_1(f_0) \frac{\partial f_0}{\partial y} =$$

$$= 3p_{32} \overline{B}_2^{(2)}(f_0) + 3p_{33} \overline{B}_2^{(3)}(f_0) + 6p_{33}\beta_{32} \overline{B}_1^{(2)}(f_0) \frac{\partial f_0}{\partial y}. \tag{65}$$

Оно может выполняться для произвольных f только в том случае, когда

$$\overline{A}_1 = 6p_{33}\beta_{32} \overline{B}_1^{(2)}, \tag{66}$$

$$\overline{A}_2 = 3p_{32} \overline{B}_2^{(2)} + 3p_{33} \overline{B}_2^{(3)}. \tag{67}$$

Из равенства (66) следует, что операторы $\overline{A}_1$ и $\overline{B}_1^{(2)}$ могут отличаться только постоянным множителем. Поэтому

$$\alpha_2 = \beta_{21}; \quad \overline{B}_1^{(2)} = \alpha_2 \overline{A}_1, \tag{68}$$

и равенство (66) переходит в

$$p_{33}\beta_{32}\alpha_2 = \frac{1}{6}. \tag{69}$$

Равенства (68) и (61) показывают, что операторы $\overline{A}_1$ и $\overline{B}_1^{(3)}$ также отличаются лишь постоянным множителем. Это возможно лишь при условии

$$\alpha_3 = \beta_{31} + \beta_{32}. \tag{70}$$

При этом

$$\overline{B}_1^{(3)} = \alpha_3 \overline{A}_1. \tag{71}$$

В силу (71), (68), (37) и (15) мы получим из (67):

$$\overline{A}_2 = (3p_{32}\alpha_3^2 + 3p_{33}\alpha_3^2)\, \overline{A}_2 \tag{72}$$

или

$$p_{32}\alpha_2^2 + p_{33}\alpha_3^2 = \frac{1}{3}. \tag{73}$$

Приравняв четвертую производную нулю и использовав (68) и (71), мы получим равенство

$$\overline{A}_3(f_0) + 3\overline{A}_1(f_0)\, \overline{A}_1\left(\frac{\partial f_0}{\partial y}\right) + \overline{A}_2(f_0) \frac{\partial f_0}{\partial y} + \overline{A}_1(f_0)\left(\frac{\partial f_0}{\partial y}\right)^2 =$$

$$= 4p_{32}\alpha_2 \overline{A}_3(f_0) + 4p_{33}\alpha_3^3 \overline{A}_2(f_0) + 24p_{33}\beta_{32}\alpha_2 \overline{A}_1(f_0)\, \alpha_3 \overline{A}_1\left(\frac{\partial f_0}{\partial y}\right) +$$

$$+ 12p_{33}\beta_{32}\alpha_2^2 \overline{A}_2(f_0) \frac{\partial f_0}{\partial y}. \tag{74}$$

Его, вообще говоря, удовлетворить не удастся, так как в левой части содержится член $\overline{A}_1(f_0)\left(\frac{\partial f_0}{\partial y}\right)^2$, а в правой такого члена нет.

Итак, если подобрать величины p_{31}, p_{32}, p_{33}, α_2, α_3, β_{21}, β_{31}, β_{32} удовлетворяющими системе

$$\left.\begin{aligned} \alpha_2 &= \beta_{21}, \\ \alpha_3 &= \beta_{31} + \beta_{32}, \\ p_{31} + p_{32} + p_{33} &= 1, \\ p_{32}\alpha_2 + p_{33}\alpha_3 &= \frac{1}{2}, \\ p_{32}\alpha_2^2 + p_{33}\alpha_3^2 &= \frac{1}{3}, \\ p_{33}\beta_{32}\alpha_2 &= \frac{1}{6}, \end{aligned}\right\} \tag{75}$$

то получим формулу Рунге — Кутта, с порядком погрешности на одном шаге h^4.

Последние три равенства (75) при выбранных α_2 и α_3 можно рассматривать как систему линейных алгебраических уравнений относительно p_{32} и p_{33}. Для совместности этой системы необходимо и достаточно обращение в нуль определителя

$$\begin{vmatrix} \alpha_2 & \alpha_3 & \frac{1}{2} \\ \alpha_2^2 & \alpha_3^2 & \frac{1}{3} \\ 0 & \beta_{32}\alpha_2 & \frac{1}{6} \end{vmatrix}. \tag{76}$$

Отсюда

$$\frac{1}{6}\begin{vmatrix} \alpha_2 & \alpha_3 \\ \alpha_2^2 & \alpha_3^2 \end{vmatrix} - \beta_{32}\alpha_2 \begin{vmatrix} \alpha_2 & \frac{1}{2} \\ \alpha_2^2 & \frac{1}{3} \end{vmatrix} = 0, \tag{77}$$

или

$$\alpha_2\alpha_3(\alpha_3 - \alpha_2) - \beta_{32}\alpha_2^2(2 - 3\alpha_2) = 0. \tag{78}$$

Последнее из уравнений (75) показывает, что $p_{33} \neq 0$, $\alpha_2 \neq 0$, $\beta_{32} \neq 0$. Поэтому мы можем поделить (78) на α_2. Это даст

$$\alpha_3(\alpha_3 - \alpha_2) - \beta_{32}\alpha_2(2 - 3\alpha_2) = 0. \tag{79}$$

Итак, мы сначала должны подобрать α_2, α_3, β_{21}, β_{31}, β_{32} так, чтобы они удовлетворяли системе

$$\left.\begin{aligned} \alpha_2 &= \beta_{21}, \\ \alpha_3 &= \beta_{31} + \beta_{32}, \\ \alpha_3(\alpha_3 - \alpha_2) - \beta_{32}\alpha_2(2 - 3\alpha_2) &= 0, \end{aligned}\right\} \tag{80}$$

и затем найти p_{33}, p_{32}, p_{31} последовательным исключением из системы

$$\left.\begin{aligned} p_{33}\beta_{32}\alpha_2 &= \frac{1}{6}, \\ p_{32}\alpha_2 + p_{33}\alpha_3 &= \frac{1}{2}, \\ p_{31} + p_{32} + p_{33} &= 1. \end{aligned}\right\} \tag{81}$$

Рассмотрим некоторые варианты.

а) Возьмем $\alpha_2 = \beta_{21} = \frac{1}{2}$, $\alpha_3 = 1$. Тогда третье из уравнений (80) даст $\beta_{32} = 2$. Второе уравнение (80) дает $\beta_{31} = -1$. Затем последовательно находим $p_{33} = \frac{1}{6}$, $p_{32} = \frac{2}{3}$, $p_{31} = \frac{1}{6}$. Получаем приближенную формулу

$$\Delta y_0 \approx \frac{1}{6}[k_1 + 4k_2 + k_3], \tag{82}$$

где

$$\left.\begin{aligned} k_1 = hf(x_0, y_0); \quad k_2 = hf\left(x_0 + \frac{1}{2}h, \; y_0 + \frac{1}{2}k_1\right); \\ k_3 = hf(x_0 + h, \; y_0 - k_1 + 2k_2). \end{aligned}\right\} \tag{83}$$

б) Возьмем $\alpha_2 = \beta_{21} = \frac{1}{3}$, $\alpha_3 = \frac{2}{3}$. При этом $\beta_{32} = \frac{2}{3}$, $\beta_{31} = 0$, $p_{33} = \frac{3}{4}$, $p_{32} = 0$, $p_{31} = \frac{1}{4}$. Получаем вторую формулу

$$\Delta y_0 \approx \frac{1}{4}k_1 + \frac{3}{4}k_3, \tag{84}$$

где

$$k_1 = hf(x_0, y_0), \quad k_2 = hf\left(x_0 + \frac{1}{3}h, \; y_0 + \frac{1}{3}k_1\right);$$

$$k_3 = hf\left(x_0 + \frac{2}{3}h, \; y_0 + \frac{2}{3}k_2\right). \tag{85}$$

в) Возьмем $\alpha_2 = \beta_{21} = \frac{1}{2}$, $\alpha_3 = \frac{3}{4}$. При этом $\beta_{32} = \frac{3}{4}$, $\beta_{31} = 0$, $p_{33} = \frac{4}{9}$, $p_{32} = \frac{1}{3}$, $p_{31} = \frac{2}{9}$. Отсюда

$$\Delta y_0 \approx \frac{1}{9}[4k_3 + 3k_2 + 2k_1], \tag{86}$$

где

$$k_1 = hf(x_0, y_0); \quad k_2 = hf\left(x_0 + \frac{h}{2}, \; y_0 + \frac{k_1}{2}\right),$$

$$k_3 = hf\left(x_0 + \frac{3}{4}h, \; y_0 + \frac{3}{4}k_2\right). \tag{87}$$

В последнем случае левая часть (74) будет отличаться от правой лишь на член $\bar{A}_1(f_0)\left(\frac{\partial f_0}{\partial y}\right)^2$.

Остаточный член выражается по общей формуле

$$R_3(h) = \frac{h^4}{24}\varphi_3^{(IV)}(\xi). \tag{88}$$

Мы не будем здесь выписывать выражения для $\varphi_3^{(IV)}(h)$ ввиду его громоздкости.

4-й случай: $r = 4$. Имеем:

$$\varphi_4'(0) = y_0' - [p_{41} + p_{42} + p_{43} + p_{44}] f_0. \tag{89}$$

Отсюда $\varphi_4'(0) = 0$ тогда и только тогда, когда

$$p_{41} + p_{42} + p_{43} + p_{44} = 1. \tag{90}$$

Далее

$$\varphi_4''(0) = y_0'' - [p_{42}k_2''(0) + p_{43}k_3''(0) + p_{44}k_4''(0)] =$$
$$= \overline{A}_1(f_0) - 2[p_{42}\overline{B}_1^{(2)}(f_0) + p_{43}\overline{B}_1^{(3)}(f_0) + p_{44}\overline{B}_1^{(4)}(f_0)]. \tag{91}$$

Отсюда

$$\frac{1}{2}\overline{A}_1 = p_{42}\overline{B}_1^{(2)} + p_{43}\overline{B}_1^{(3)} + p_{44}\overline{B}_1^{(4)}. \tag{92}$$

Приравняв нулю третью производную, получим:

$$\overline{A}_2(f_0) + \overline{A}_1(f_0)\frac{\partial f_0}{\partial y} = 3p_{42}\overline{B}_2^{(2)}(f_0) + 3p_{43}\overline{B}_2^{(3)}(f_0) +$$
$$+ 6p_{43}\beta_{32}\overline{B}_1^{(2)}(f_0)\frac{\partial f_0}{\partial y} + 3p_{44}\overline{B}_2^{(4)}(f_0) +$$
$$+ 6p_{44}\beta_{42}\overline{B}_1^{(2)}(f_0)\frac{\partial f_0}{\partial y} + 6p_{44}\beta_{43}\overline{B}_1^{(3)}(f_0)\frac{\partial f_0}{\partial y}. \tag{93}$$

Отсюда

$$\overline{A}_2 = 3p_{42}\overline{B}_2^{(2)} + 3p_{43}\overline{B}_2^{(3)} + 3p_{44}\overline{B}_2^{(4)}, \tag{94}$$

$$\overline{A}_1 = 6p_{43}\beta_{32}\overline{B}_1^{(2)} + 6p_{44}\beta_{42}\overline{B}_1^{(2)} + 6p_{44}\beta_{43}\overline{B}_1^{(3)}. \tag{95}$$

Наконец, приравнивание нулю четвертой производной даст

$$\overline{A}_3(f_0) + 3\overline{A}_1(f_0)\overline{A}_1\left(\frac{\partial f_0}{\partial y}\right) + \overline{A}_2(f_0)\frac{\partial f_0}{\partial y} + \overline{A}_1(f_0)\left(\frac{\partial f_0}{\partial y}\right)^2 =$$
$$= 4p_{42}\overline{B}_3^{(2)}(f_0) + 4p_{43}\overline{B}_3^{(3)}(f_0) + 4p_{44}\overline{B}_3^{(4)}(f_0) +$$
$$+ 24p_{43}\beta_{32}\overline{B}_1^{(2)}(f_0)\overline{B}_1^{(3)}\left(\frac{\partial f_0}{\partial y}\right) + 24p_{41}\beta_{42}\overline{B}_1^{(2)}(f_0)\overline{B}_1^{(4)}\left(\frac{\partial f_0}{\partial y}\right) +$$
$$+ 24p_{44}\beta_{43}\overline{B}_1^{(3)}(f_0)\overline{B}_1^{(4)}\left(\frac{\partial f_0}{\partial y}\right) + 12p_{43}\beta_{32}\overline{B}_2^{(2)}(f_0)\frac{\partial f_0}{\partial y} +$$
$$+ 12p_{44}\beta_{42}\overline{B}_2^{(2)}(f_0)\frac{\partial f_0}{\partial y} + 12p_{44}\beta_{43}\overline{B}_2^{(3)}(f_0)\frac{\partial f_0}{\partial y} +$$
$$+ 24p_{44}\beta_{43}\beta_{32}\overline{B}_1^{(2)}(f_0)\left(\frac{\partial f_0}{\partial y}\right)^2. \tag{96}$$

В (96) $\left(\frac{\partial f_0}{\partial y}\right)^2$ входит множителем только в последние члены слева и справа. Поэтому

$$\overline{A}_1 = 24 p_{44}\beta_{43}\beta_{32}\overline{B}_1^{(2)}. \tag{97}$$

Равенство (97) показывает, что $\overline{A}_1$ и $\overline{B}_1^{(2)}$ отличаются лишь постоянным множителем. Это возможно лишь в том случае, когда

$$\alpha_2 = \beta_{21}. \tag{98}$$

Тогда

$$\overline{B}_1^{(2)} = \alpha_2 \overline{A}_1, \tag{99}$$

а (97) переходит в

$$p_{44}\beta_{43}\beta_{32}\alpha_2 = \frac{1}{24}. \tag{100}$$

Из (99) и (95) следует, что $\overline{A}_1$ и $\overline{B}_1^{(3)}$ также отличаются лишь постоянным множителем. При этом должно быть

$$\alpha_3 = \beta_{31} + \beta_{32} \tag{101}$$

и

$$\overline{B}_1^{(3)} = \alpha_3 \overline{A}_1. \tag{102}$$

Поэтому (95) переходит в

$$p_{43}\beta_{32}\alpha_2 + p_{44}\beta_{42}\alpha_2 + p_{44}\beta_{43}\alpha_3 = \frac{1}{6}. \tag{103}$$

Аналогично из (92) получим:

$$\alpha_4 = \beta_{41} + \beta_{42} + \beta_{43}, \tag{104}$$

$$\overline{B}_1^{(4)} = \alpha_4 \overline{A}_1 \tag{105}$$

и

$$p_{42}\alpha_2 + p_{43}\alpha_3 + p_{44}\alpha_4 = \frac{1}{2}. \tag{106}$$

Равенство (94) даст

$$p_{42}\alpha_2^2 + p_{43}\alpha_3^2 + p_{44}\alpha_4^2 = \frac{1}{3}. \tag{107}$$

Сравнивая члены с производными третьего порядка в (96), получим:

$$p_{42}\alpha_2^3 + p_{43}\alpha_3^3 + p_{44}\alpha_4^3 = \frac{1}{4}. \tag{108}$$

Из равенства (96) мы получаем еще следующие два соотношения:

$$p_{43}\beta_{32}\alpha_2\alpha_3 + p_{44}\beta_{42}\alpha_2\alpha_4 + p_{44}\beta_{43}\alpha_3\alpha_4 = \frac{1}{8}, \tag{109}$$

$$p_{43}\beta_{32}\alpha_2^2 + p_{44}\beta_{42}\alpha_2^2 + p_{44}\beta_{43}\alpha_3^2 = \frac{1}{12}. \tag{110}$$

Обращение в нуль пятой производной мы обеспечить в нашем случае не можем.

Собирая все соотношения, связывающие величины p_{4i}, α_i, β_{ij}, мы получим следующую систему уравнений:

$$\left.\begin{aligned}
&\alpha_2 = \beta_{21},\\
&\alpha_3 = \beta_{31} + \beta_{32},\\
&\alpha_4 = \beta_{41} + \beta_{42} + \beta_{43},\\
&p_{41} + p_{42} + p_{43} + p_{44} = 1,\\
&p_{42}\alpha_2 + p_{43}\alpha_3 + p_{44}\alpha_4 = \frac{1}{2},\\
&p_{42}\alpha_2^2 + p_{43}\alpha_3^2 + p_{44}\alpha_4^2 = \frac{1}{3},\\
&p_{42}\alpha_2^3 + p_{43}\alpha_3^3 + p_{44}\alpha_4^3 = \frac{1}{4},\\
&p_{43}\beta_{32}\alpha_2 + p_{44}\beta_{42}\alpha_2 + p_{44}\beta_{43}\alpha_3 = \frac{1}{6},\\
&p_{43}\beta_{32}\alpha_2\alpha_3 + p_{44}\beta_{42}\alpha_2\alpha_4 + p_{44}\beta_{43}\alpha_3\alpha_4 = \frac{1}{8},\\
&p_{43}\beta_{32}\alpha_2^2 + p_{44}\beta_{42}\alpha_2^2 + p_{44}\beta_{43}\alpha_3^2 = \frac{1}{12},\\
&p_{44}\beta_{43}\beta_{32}\alpha_2 = \frac{1}{24}.
\end{aligned}\right\} \tag{111}$$

Из последнего уравнения системы следует, что $p_{44} \neq 0$, $\beta_{43} \neq 0$, $\beta_{32} \neq 0$, $\alpha_2 \neq 0$. Поэтому, если нам каким-то образом удалось найти величины α_i и β_{ij}, то p_{4i} могут быть найдены последовательным исключением из следующей системы:

$$\left.\begin{aligned}
&p_{44}\beta_{43}\beta_{32}\alpha_2 = \frac{1}{24},\\
&p_{43}\beta_{32}\alpha_2 + p_{44}\beta_{42}\alpha_2 + p_{44}\beta_{43}\alpha_3 = \frac{1}{6},\\
&p_{42}\alpha_2 + p_{43}\alpha_3 + p_{44}\alpha_4 = \frac{1}{2},\\
&p_{41} + p_{42} + p_{43} + p_{44} = 1.
\end{aligned}\right\} \tag{112}$$

Найдем теперь соотношения, связывающие α_i и β_{ij}. Возьмем восьмое, десятое и одиннадцатое уравнения системы (111) и будем рассматривать их как систему линейных алгебраических уравнений относительно p_{43} и p_{44}. Для совместности этой системы требуется:

$$\begin{vmatrix}
\beta_{32}\alpha_2 & \beta_{42}\alpha_2 + \beta_{43}\alpha_3\dfrac{1}{6}\\
\beta_{32}\alpha_2^2 & \beta_{42}\alpha_2^2 + \beta_{43}\alpha_3^2\dfrac{1}{12}\\
0 & \beta_{43}\beta_{32}\alpha_2 \quad \dfrac{1}{24}
\end{vmatrix} = 0. \tag{113}$$

Отсюда

$$\begin{vmatrix} 1 & \alpha_3 & 4 \\ \alpha_2 & \alpha_3^2 & 2 \\ 0 & \beta_{32}\alpha_2 & 1 \end{vmatrix} = 0 \tag{114}$$

или

$$-4\beta_{32}\alpha_2^2 + 2\beta_{32}\alpha_2 - \alpha_3^2 + \alpha_2\alpha_3 = 0. \tag{115}$$

Соотношение (115) позволит нам выразить β_{32} через α_2 и α_3, если $\alpha_2 \neq \frac{1}{2}$, и даст равенство $\alpha_2 = \alpha_3$, если $\alpha_2 = \frac{1}{2}$.

Умножая восьмое уравнение на α_4 и вычтя из него девятое, получим:

$$\frac{1}{6}\alpha_4 - \frac{1}{8} = p_{43}\beta_{32}\alpha_2(\alpha_4 - \alpha_3). \tag{116}$$

Проделаем тоже самое с пятым и шестым уравнениями, а затем с шестым и седьмым. Это даст два равенства:

$$\frac{1}{2}\alpha_4 - \frac{1}{3} = p_{42}\alpha_2(\alpha_4 - \alpha_2) + p_{43}\alpha_3(\alpha_4 - \alpha_3), \tag{117}$$

$$\frac{1}{3}\alpha_4 - \frac{1}{4} = p_{42}\alpha_2^2(\alpha_4 - \alpha_2) + p_{43}\alpha_3^2(\alpha_4 - \alpha_3). \tag{118}$$

Умножим (117) на α_2 и вычтем из (118). Будем иметь:

$$\left(\frac{1}{3}\alpha_4 - \frac{1}{4}\right) - \alpha_2\left(\frac{1}{2}\alpha_4 - \frac{1}{3}\right) = p_{43}\alpha_3(\alpha_3 - \alpha_2)(\alpha_4 - \alpha_3). \tag{119}$$

Подставим сюда вместо $p_{43}(\alpha_4 - \alpha_3)$ его выражение из (116); получим:

$$\beta_{32}\alpha_2\left[\left(\frac{1}{3}\alpha_4 - \frac{1}{4}\right) - \alpha_2\left(\frac{1}{2}\alpha_4 - \frac{1}{3}\right)\right] = \left(\frac{1}{6}\alpha_4 - \frac{1}{8}\right)\alpha_3(\alpha_3 - \alpha_2) \tag{120}$$

или

$$8\beta_{32}\alpha_2^2 - 6\beta_{32}\alpha_2 + 3\alpha_3^2 - 3\alpha_2\alpha_3 = \left[12\beta_{32}\alpha_2^2 - 8\beta_{32}\alpha_2 + 4\alpha_3^2 - 4\alpha_2\alpha_3\right]\alpha_4. \tag{121}$$

Прибавим теперь к левой части (121) утроенную левую часть (115), а к скобке в правой части учетверенную левую часть (115). При этом равенство (121) перейдет в

$$-4\beta_{32}\alpha_2^2 = -4\beta_{32}\alpha_2^2\alpha_4. \tag{122}$$

Так как $\beta_{32} \neq 0$, $\alpha_2 \neq 0$, то

$$\alpha_4 = 1. \tag{123}$$

Потребуем теперь совместности пятого, шестого, седьмого и одиннадцатого уравнений системы (111) относительно p_{42}, p_{43}, p_{44} с учетом (123). Это возможно только, если

$$\begin{vmatrix} \alpha_2 & \alpha_3 & 1 & \frac{1}{2} \\ \alpha_2^2 & \alpha_3^2 & 1 & \frac{1}{3} \\ \alpha_2^3 & \alpha_3^3 & 1 & \frac{1}{4} \\ 0 & 0 & \beta_{43}\beta_{32}\alpha_2 & \frac{1}{24} \end{vmatrix} = 0. \tag{124}$$

Отсюда

$$\begin{vmatrix} 1 & 1 & 1 & 12 \\ \alpha_2 & \alpha_3 & 1 & 8 \\ \alpha_2^2 & \alpha_3^2 & 1 & 6 \\ 0 & 0 & \beta_{43}\beta_{32}\alpha_2 & 1 \end{vmatrix} = 0. \tag{125}$$

Раскрывая определитель и производя некоторые сокращения, найдем еще одно соотношение, связывающее α_i и β_{ij}:

$$\beta_{43}\beta_{32}\alpha_2 [12\alpha_2\alpha_3 - 8\alpha_2 - 8\alpha_3 + 6] = (1 - \alpha_2)(1 - \alpha_3). \tag{126}$$

Наконец, последнее соотношение мы получим, если потребуем совместности восьмого, девятого и одиннадцатого уравнений (121) относительно p_{43} и p_{44}. Это даст

$$\begin{vmatrix} \beta_{32}\alpha_2 & \beta_{42}\alpha_2 + \beta_{43}\alpha_3\frac{1}{6} \\ \beta_{32}\alpha_2\alpha_3 & \beta_{42}\alpha_2 + \beta_{43}\alpha_3\frac{1}{8} \\ 0 & \beta_{43}\beta_{32}\alpha_2 \quad \frac{1}{24} \end{vmatrix} = 0. \tag{127}$$

Отсюда

$$\begin{vmatrix} 1 & \beta_{42}\alpha_2 + \beta_{43}\alpha_3\frac{1}{6} \\ \alpha_3 & \beta_{42}\alpha_2 + \beta_{43}\alpha_3\frac{1}{8} \\ 0 & \beta_{43}\beta_{32}\alpha_2 \quad \frac{1}{24} \end{vmatrix} = 0 \tag{128}$$

и окончательно

$$\beta_{42}\alpha_2(1 - \alpha_3) = \beta_{43}\beta_{32}\alpha_2(3 - 4\alpha_3) + \beta_{43}\alpha_3(\alpha_3 - 1). \tag{129}$$

Итак, для удовлетворения системе (111) мы сначала подбираем α_i и β_{ij}, удовлетворяющие условиям:

$$\left.\begin{aligned}
&\alpha_2 = \beta_{21},\\
&\alpha_3 = \beta_{31} + \beta_{32},\\
&\alpha_4 = \beta_{41} + \beta_{42} + \beta_{43}, \quad \alpha_4 = 1,\\
&4\beta_{32}\alpha_2^2 - 2\beta_{32}\alpha_2 + \alpha_3^2 - \alpha_2\alpha_3 = 0,\\
&\beta_{43}\beta_{32}\alpha_2 [12\alpha_2\alpha_3 - 8\alpha_2 - 8\alpha_3 + 6] = (1-\alpha_2)(1-\alpha_3),\\
&\beta_{42}\alpha_2(1-\alpha_3) = \beta_{43}\beta_{32}\alpha_2(3-4\alpha_3) + \beta_{43}\alpha_3(\alpha_3 - 1),
\end{aligned}\right\} \tag{130}$$

а затем находим p_{4i}, решая систему (112). Рассмотрим некоторые частные варианты формул Рунге — Кутта, имеющих порядок погрешности на одном шаге h^5.

а) $\alpha_2 = \frac{1}{2}$, $\alpha_3 = \frac{1}{2}$, $\beta_{32} = \frac{1}{2}$. При этом $\beta_{21} = \frac{1}{2}$, $\beta_{31} = 0$, $\alpha_4 = 1$, $\beta_{43} = 1$, $\beta_{42} = 0$, $\beta_{41} = 0$, $p_{44} = \frac{1}{6}$, $p_{43} = \frac{1}{3}$, $p_{42} = \frac{1}{3}$, $p_{41} = \frac{1}{6}$.

Получаем формулу

$$\Delta y_0 \approx \frac{1}{6}[k_1 + 2k_2 + 2k_3 + k_4], \tag{131}$$

где

$$k_1 = hf(x_0, y_0); \quad k_2 = hf\left(x_0 + \frac{h}{2}, y_0 + \frac{k_1}{2}\right);$$

$$k_3 = hf\left(x_0 + \frac{h}{2}, y_0 + \frac{k_2}{2}\right), \quad k_4 = hf(x_0 + h, y_0 + k_3). \tag{132}$$

Это одна из наиболее распространенных формул Рунге — Кутта.

б) Полагаем $\alpha_2 = \frac{1}{3}$, $\alpha_3 = \frac{2}{3}$. Тогда $\beta_{21} = \frac{1}{3}$, $\alpha_4 = 1$, $\beta_{32} = 1$, $\beta_{31} = -\frac{1}{3}$, $\beta_{43} = 1$, $\beta_{42} = -1$, $\beta_{41} = 1$, $p_{44} = \frac{1}{8}$, $p_{43} = \frac{3}{8}$, $p_{42} = \frac{3}{8}$, $p_{41} = \frac{1}{8}$. Получаем следующую формулу:

$$\Delta y_0 \approx \frac{1}{8}[k_1 + 3k_2 + 3k_3 + k_4], \tag{133}$$

где

$$k_1 = hf(x_0, y_0); \quad k_2 = hf\left(x_0 + \frac{h}{3}, y_0 + \frac{k_1}{3}\right);$$

$$k_3 = hf\left(x_0 + \frac{2}{3}h, \quad y_0 - \frac{k_1}{3} + k_2\right);$$

$$k_4 = hf(x_0 + h, y_0 + k_1 - k_2 + k_3). \tag{134}$$

в) Полагаем $\alpha_2=\beta_{21}=\frac{1}{4}$, $\alpha_3=\frac{1}{2}$. Тогда $\beta_{32}=\frac{1}{2}$, $\beta_{31}=0$, $\alpha_4=1$, $\beta_{43}=2$, $\beta_{42}=-2$, $\beta_{41}=1$, $p_{44}=\frac{1}{6}$, $p_{43}=\frac{2}{3}$, $p_{42}=0$, $p_{41}=\frac{1}{6}$. При этом получаем:

$$\Delta y_0 \approx \frac{1}{6}[k_1+4k_3+k_4], \tag{135}$$

где

$$\left.\begin{aligned} k_1&=hf(x_0,\ y_0);\\ k_2&=hf\left(x_0+\frac{1}{4}h,\ y_0+\frac{1}{4}k_1\right),\\ k_3&=hf\left(x_0+\frac{h}{2},\ y_0+\frac{k_2}{2}\right);\\ k_4&=hf(x_0+h,\ y_0+k_1-2k_2+k_3). \end{aligned}\right\} \tag{136}$$

При желании набор формул можно увеличить. Погрешность каждой из этих формул равна

$$R_4(h)=\frac{h^5\varphi_4^{(5)}(\xi)}{120}, \tag{137}$$

где

$$\varphi_4(h)=y(x_0+h)-y(x_0)-[p_{41}k_1(h)+p_{42}k_2(h)+ \\ +p_{43}k_3(h)+p_{44}k_4(h)]. \tag{138}$$

Вычисления показывают, что при $r=5$ мы не достигаем увеличения порядка точности. Поэтому эти формулы применения не находят.

Можно получить формулы, имеющие порядок ошибки h^6, но при этом придется брать $r\geqslant 6$. Получаются очень громоздкие формулы, неудобные для практики. Поэтому о них мы говорить не будем.

Применяя ту или иную формулу Рунге — Кутта, мы найдем приближенное значение Δy_0, а следовательно и $y_1=y(x_0+h)$. Затем можно взять за начальную точку $x_1=x_0+h$ и за начальное значение $y_1=y(x_0+h)$ и продвинуться еще на один шаг такой же или другой длины. Повторяя этот процесс, мы получим таблицу значений искомого решения в некоторых точках.

Приведем примеры на применение формул Рунге — Кутта. Будем искать решение уравнения $y'=y$, удовлетворяющее начальному условию $y(0)=1$ на отрезке $[0,\ 1]$. Шаг возьмем равным 0,1.

Точным решением будет $y = e^x$ и его значения будут даны позже. Ход вычислений будет виден из таблицы. Применяется формула

$$\Delta y_i = \frac{1}{6}[k_1 + 4k_2 + k_3],$$

$$k_1 = hf(x_i, y_i);\ k_2 = hf\left(x_i + \frac{h}{2},\ y_i + \frac{k_1}{2}\right);\quad k_3 = hf(x_i + h,\ y_i + 2k_2 - k_1).$$

x	y	$f(x, y)$	hf		
x_i $x_i + \frac{h}{2}$ $x_i + h$	y_i $y_i + \frac{k_1}{2}$ $y_i + 2k_2 - k_1$	$f(x_i, y_i)$ $f\left(x_i + \frac{h}{2}, y_i + \frac{k_1}{2}\right)$ $f(x_i + h, y_i + 2k_2 - k_1)$	k_1 k_2 k_3	k_1 $4k_2$ k_3	$2k_2$ $2k_2 - k_1$
0,0	1,0000			$6\Delta y_i$	Δy_i
0,00 0,05 0,10	1,0000 1,0500 1,1100	1,0000 1,0500 1,1100	0,1000 0,1050 0,1110	0,1000 0,4200 0,1110	0,2100 0,1100
0,1	1,1052			0,6310	0,1052
0,10 0,15 0,20	1,1052 1,1605 1,2267	1,1052 1,1605 1,2267	0,1105 0,1160 0,1227	0,1105 0,4640 0,1227	0,2320 0,1215
0,2	1,2214			0,6972	0,1162
0,20 0,25 0,30	1,2214 1,2825 1,3557	1,2214 1,2825 1,3557	0,1221 0,1282 0,1356	0,1221 0,5128 0,1356	0,2564 0,1343
0,3	1,3498			0,7705	0,1284
0,30 0,35 0,40	1,3498 1,4173 1,4982	1,3498 1,4173 1,4982	0,1350 0,1417 0,1498	0,1350 0,5668 0,1498	0,2834 0,1484
0,4	1,4917			0,8516	0,1419

Продолжение

x	y	$f(x, y)$	hf		
0,40	1,4917	1,4917	0,1492	0,1492	
0,45	1,5663	1,5663	0,1566	0,6264	0,3132
0,50	1,6567	1,6567	0,1657	0,1657	0,1640
0,5	1,6486			0,9413	0,1569
0,50	1,6486	1,6486	0,1649	0,1649	
0,55	1,7310	1,7310	0,1731	0,6924	0,3462
0,60	1,8299	1,8299	0,1830	0,1830	0,1813
0,6	1,8220			1,0403	0,1734
0,60	1,8220	1,8220	0,1822	0,1822	
0,65	1,9131	1,9131	0,1913	0,7652	0,3826
0,70	2,0224	2,0224	0,2022	0,2022	0,2004
0,7	2,0136			1,1496	0,1916
0,70	2,0136	2,0136	0,2014	0,2014	
0,75	2,1143	2,1143	0,2114	0,8456	0,4228
0,80	2,2350	2,2350	0,2235	0,2235	0,2214
0,8	2,2254			1,2705	0,2118
0,80	2,2254	2,2254	0,2225	0,2225	
0,85	2,3366	2,3366	0,2337	0,9348	0,4674
0,90	2,4703	2,4703	0,2470	0,2470	0,2449
0,9	2,4595			1,4043	0,2341
0,90	2,4595	2,4595	0,2460	0,2460	
0,95	2,5825	2,5825	0,2582	1,0328	0,5164
1,00	2,7299	2,7299	0,2730	0,2730	0,2704
1,0	2,7181			1,5518	0,2586

$$\Delta y_i = \frac{1}{4}(k_1 + 3k_3)$$

x_i	y_i	$f(x_i, y_i)$	$hf_i = k_i$		
x_i $x_i + \frac{h}{3}$ $x_i + \frac{2h}{3}$	y_i $y_i + \frac{k_1}{3}$ $y_i + \frac{2k_2}{3}$	$f(x_i, y_i)$ $f\left(x_i + \frac{h}{3}, y_i + \frac{k_1}{3}\right)$ $f\left(x_i + \frac{2h}{3}, y_i + \frac{2k_2}{3}\right)$	$hf_1 = k_1$ $hf_2 = k_2$ $hf_3 = k_3$	$hf_1 = k_1$ $3k_3$	
0,0	1,0000			$4\Delta y_i$	Δy_i
0,0 0,0333 0,0667	1,0000 1,0333 1,0689	1,0000 1,0333 1,0689	0,1000 0,1033 0,1069	0,1000 0,3207	
0,1	1,1052			0,4207	0,1052
0,1000 0,1333 0,1667	1,1052 1,1420 1,1814	1,1052 1,1420 1,1814	0,1105 0,1142 0,1181	0,1105 0,3543	
0,2	1,2214			0,4648	0,1162
0,2000 0,2333 0,2667	1,2214 1,2621 1,3055	1,2214 1,2621 1,3055	0,1221 0,1262 0,1305	0,1221 0,3917	
0,3	1,3498			0,5138	0,1284
0,3000 0,3333 0,3667	1,3498 1,3948 1,4428	1,3498 1,3948 1,4428	0,1350 0,1395 0,1443	0,1350 0,4329	
0,4	1,4918			0,5679	0,1420

Продолжение

x_i	y_i	$f(x_i, y_i)$	$hf_i = k_i$		
0,4000	1,4918	1,4918	0,1492	0,1492	
0,4333	1,5415	1,5415	0,1541		
0,4667	1,5946	1,5946	0,1595	0,4784	
0,5	1,6487			0,6276	0,1569
0,500	1,6487	1,6487	0,1649	0,1649	
0,5333	1,7037	1,7037	0,1704		
0,5667	1,7623	1,7623	0,1762	0,5286	
0,6	1,8221			0,6935	0,1734
0,6000	1,8221	1,8221	0,1822	0,1822	
0,6333	1,8828	1,8828	0,1883		
0,6667	1,9476	1,9476	0,1948	0,5843	
0,7	2,0137			0,7665	0,1918
0,7000	2,0137	2,0137	0,2014	0,2014	
0,7333	2,0808	2,0808	0,2081		
0,7667	2,1524	2,1524	0,2152	0,6456	
0,8	2,2254			0,8470	0,2117
0,8000	2,2254	2,2254	0,2225	0,2225	
0,8333	2,2996	2,2996	0,2300		
0,8667	2,3787	2,3787	0,2379	0,7136	
0,9	2,4594			0,9361	0,2340
0,9000	2,4594	2,4594	0,2459	0,2459	
0,9333	2,5414	2,5414	0,2541		
0,9667	2,6288	2,6288	0,2629	0,7887	
1,0	2,7180			1,0346	0,2586

$$\Delta y_i = \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

x_i	y_i	$f(x_i, y_i)$	k_i		
x_i	y_i	$f(x_i, y_i)$	$hf_1 = k_1$	k_1	
$x_i + \frac{h}{2}$	$y_i + \frac{1}{2} k_1$	$f\left(x_i + \frac{h}{2}, y_i + \frac{k_1}{2}\right)$	$hf_2 = k_2$	$2k_2$	
$x_i + \frac{h}{2}$	$y_i + \frac{1}{2} k_2$	$f\left(x_i + \frac{h}{2}, y_i + \frac{k_2}{2}\right)$	$hf_3 = k_3$	$2k_3$	
$x_i + h$	$y_i + k_3$	$f(x_i + h, y_i + k_3)$	$hf_4 = k_4$	k_4	
0,0	1,000000			$6\Delta y_i$	Δy_i
0,00	1,000000	1,000000	0,100000	0,100000	
0,05	1,050000	1,050000	0,105000	0,210000	
0,05	1,052500	1,052500	0,105250	0,210500	
0,10	1,105250	1,105250	0,110525	0,110525	
0,1	1,10517			0,631025	0,105174
0,10	1,105170	1,105170	0,105517	0,110517	
0,15	1,160429	1,160429	0,116043	0,232086	
0,15	1,163191	1,163191	0,116319	0,232638	
0,20	1,221489	1,221489	0,122149	0,122149	
0,2	1,22140			0,697390	0,116231
0,20	1,221400	1,221400	0,122140	0,122140	
0,25	1,282470	1,282470	0,128247	0,256494	
0,25	1,285524	1,285524	0,128552	0,257104	
0,30	1,349952	1,349952	0,134995	0,134995	
0,3	1,34986			0,770733	0,128456
0,30	1,349860	1,349860	0,134986	0,134986	
0,35	1,417353	1,417353	0,141735	0,283470	
0,35	1,420727	1,420727	0,142073	0,284146	
0,40	1,491933	1,491933	0,149193	0,149193	
0,4	1,49183			0,851795	0,141966

Продолжение

x_i	y_i	$f(x_i, y_i)$	k_i		
0,40	1,491830	1,491830	0,149183	0,149183	
0,45	1,566421	1,566421	0,156642	0,313284	
0,45	1,570151	1,570151	0,157015	0,314030	
0,50	1,648845	1,648845	0,164884	0,164884	
0,5	1,64873			0,941381	0,156897
0,50	1,648730	1,648730	0,164873	0,164873	
0,55	1,731166	1,731166	0,173117	0,346234	
0,55	1,735288	1,735288	0,173529	0,347058	
0,60	1,822259	1,822259	0,182226	0,182226	
0,6	1,82213			1,040391	0,173398
0,60	1,822130	1,822130	0,182213	0,182213	
0,65	1,913236	1,913236	0,191324	0,382648	
0,65	1,917792	1,917792	0,191779	0,383558	
0,70	2,013909	2,013909	0,201391	0,201391	
0,7	2,01377			1,149810	0,191635
0,70	2,013770	2,013770	0,201377	0,201377	
0,75	2,114458	2,114458	0,211446	0,422892	
0,75	2,119493	2,119493	0,211949	0,423898	
0,80	2,225719	2,225719	0,222572	0,222572	
0,8	2,22556			1,270739	0,211790
0,80	2,225560	2,225560	0,222556	0,222556	
0,85	2,336838	2,336838	0,233684	0,467368	
0,85	2,342402	2,342402	0,234240	0,468480	
0,90	2,459800	2,459800	0,245980	0,245980	
0,9	2,45962			1,404384	0,234064
0,90	2,459620	2,459620	0,245962	0,245962	
0,95	2,582601	2,582601	0,258260	0,516520	
0,95	2,588750	2,588750	0,258875	0,517750	
1,00	2,718495	2,718495	0,271850	0,271850	
1,0	2,71830			1,552082	0,258680

Приведем таблицу значений e^x с пятью верными десятичными знаками:

x	0,1	0,2	0,3	0,4	0,5	0,6
e^x	1,10517	1,22140	1,34986	1,49182	1,64872	1,82212

x	0,7	0,8	0,9	1,0
e^x	2,01375	2,22554	2,45960	2,71828

Как мы видим, результаты получились довольно хорошие.

2. Метод Рунге — Кутта решения систем дифференциальных уравнений первого порядка. Метод Рунге — Кутта без труда переносится на системы обыкновенных дифференциальных уравнений. Для сокращения записей ограничимся системой двух уравнений:

$$\frac{dy}{dx} = f(x, y, z); \quad \frac{dz}{dx} = g(x, y, z), \tag{139}$$

и будем разыскивать ее решение, удовлетворяющее начальным условиям $y(x_0) = y_0$, $z(x_0) = z_0$. Как и ранее, образуем функции

$$k_i(h) = hf(\xi_i, \eta_i, \zeta_i); \quad l_i(h) = hg(\xi_i, \eta_i, \zeta_i), \tag{140}$$

где

$$\left.\begin{aligned}
&\xi_i = x_0 + \alpha_i h, \quad \alpha_1 = 0;\\
&\bar{\xi}_i = \bar{x}_0 + \bar{\alpha}_i h, \quad \bar{\alpha}_1 = 0;\\
&\eta_i = y_0 + \beta_{i1}k_1 + \beta_{i2}k_2 + \ldots + \beta_{i,\,i-1}k_{i-1};\\
&\bar{\eta}_i = y_0 + \bar{\beta}_{i1}k_1 + \bar{\beta}_{i2}k_2 + \ldots + \bar{\beta}_{i,\,i-1}k_{i-1};\\
&\zeta_i = z_0 + \gamma_{i1}l_1 + \gamma_{i2}l_2 + \ldots + \gamma_{i,\,i-1}l_{i-1};\\
&\bar{\zeta}_i = z_0 + \bar{\gamma}_{i1}l_1 + \bar{\gamma}_{i2}k_2 + \ldots + \bar{\gamma}_{i,\,i-1}l_{i-1}.
\end{aligned}\right\} \tag{141}$$

Задача опять будет заключаться в подборе постоянных α_i, β_{ij}, γ_{ij}, $\bar{\alpha}_i$, $\bar{\beta}_{ij}$, $\bar{\gamma}_{ij}$, p_{ri}, q_{ri} так, чтобы разложения функций

$$\Delta y_0 - [p_{r1}k_1(h) + p_{r2}k_2(h) + \ldots + p_{rr}k_r(h)], \tag{142}$$

$$\Delta z_0 - [q_{r1}l_1(h) + q_{r2}l_2(h) + \ldots + q_{rr}l_r(h)] \tag{143}$$

по степеням h начинались с возможно более высоких степеней h. Введем операторы

$$A_m(u) = \left(\frac{\partial}{\partial x} + f\frac{\partial}{\partial y} + g\frac{\partial}{\partial z}\right)^m u =$$

$$= \sum_{r+s+t=m} \frac{m!}{r!\,s!\,t!} f^s g^t \frac{\partial^m u}{\partial x^r \partial y^s \partial z^t}, \tag{144}$$

$$\bar{A}_m(u) = \left(\frac{\partial}{\partial x} + f_0\frac{\partial}{\partial y} + g_0\frac{\partial}{\partial z}\right)^m u =$$

$$= \sum_{r+s+t=m} \frac{m!}{r!\,s!\,t!} f_0^s g_0^t \frac{\partial^m u}{\partial x^r \partial y^s \partial z^t}. \tag{145}$$

Как и ранее, получим:

$$A_1[A_m(u)] = A_{m+1}(u) + mA_1(f)\,A_{m-1}\left(\frac{\partial u}{\partial y}\right) + mA_1(g)\,A_{m-1}\left(\frac{\partial u}{\partial z}\right), \tag{146}$$

$$\bar{A}_1[\bar{A}_m(u)] = \bar{A}_{m+1}(u). \tag{147}$$

Используя (146) и формулы, аналогичные (7), получим:

$$\left.\begin{aligned}
y_0'' &= \bar{A}_1(f_0),\\
y_0''' &= \bar{A}_2(f_0) + \bar{A}_1(f_0)\frac{\partial f_0}{\partial y} + \bar{A}_1(g_0)\frac{\partial f_0}{\partial z},\\
y_0^{(\mathrm{IV})} &= \bar{A}_3(f_0) + 3\bar{A}_1(f_0)\,\bar{A}_1\left(\frac{\partial f_0}{\partial y}\right) + 3\bar{A}_1(g_0)\,\bar{A}_1\left(\frac{\partial f_0}{\partial z}\right) +\\
&\quad + \bar{A}_2(f_0)\frac{\partial f_0}{\partial y} + \bar{A}_1(f_0)\left(\frac{\partial f_0}{\partial y}\right)^2 + \bar{A}_1(g_0)\frac{\partial f_0}{\partial z}\frac{\partial f_0}{\partial y} +\\
&\quad + \bar{A}_2(g_0)\frac{\partial f_0}{\partial z} + \bar{A}_1(f_0)\frac{\partial g_0}{\partial y}\frac{\partial f_0}{\partial z} + \bar{A}_1(g_0)\frac{\partial g_0}{\partial z}\frac{\partial f_0}{\partial z},
\end{aligned}\right\} \tag{148}$$

$$
\left.\begin{aligned}
z_0'' &= \bar{A}_1(g_0),\\
z_0''' &= \bar{A}_2(g_0) + \bar{A}_1(f_0)\frac{\partial g_0}{\partial y} + \bar{A}_1(g_0)\frac{\partial g_0}{\partial z},\\
z_0^{(\mathrm{IV})} &= \bar{A}_3(g_0) + 3\bar{A}_1(f_0)\,\bar{A}_1\left(\frac{\partial g_0}{\partial y}\right) + 3\bar{A}_1(g_0)\,\bar{A}_1\left(\frac{\partial g_0}{\partial z}\right) +\\
&\quad + \bar{A}_2(f_0)\frac{\partial g_0}{\partial y} + \bar{A}_1(f_0)\frac{\partial f_0}{\partial y}\frac{\partial g_0}{\partial y} + \bar{A}_1(g_0)\frac{\partial f_0}{\partial z}\frac{\partial g_0}{\partial z} +\\
&\quad + \bar{A}_2(g_0)\frac{\partial g_0}{\partial z} + \bar{A}_1(f_0)\frac{\partial g_0}{\partial y}\frac{\partial g_0}{\partial z} + \bar{A}_1(g_0)\left(\frac{\partial g_0}{\partial z}\right)^2.
\end{aligned}\right\} \tag{149}
$$

Далее, вводим операторы

$$
\left.\begin{aligned}
B_m^{(i)}(u) &= \left[\alpha_i\frac{\partial}{\partial x} + \eta_i'(h)\frac{\partial}{\partial y} + \zeta_i'(h)\frac{\partial}{\partial z}\right]^m u =\\
&= \sum_{r+s+t=m} \frac{m!}{r!\,s!\,t!}\,\alpha_i^r\left[\eta_i'(h)\right]^s\left[\zeta_i'(h)\right]^t\frac{\partial^m u}{\partial x^r\,\partial y^s\,\partial z^t},\\
C_m^{(i)}(u) &= \left[\bar{\alpha}_i\frac{\partial}{\partial x} + \bar{\eta}_i'(h)\frac{\partial}{\partial y} + \bar{\zeta}_i'(h)\frac{\partial}{\partial z}\right]^m u =\\
&= \sum_{r+s+t=m} \frac{m!}{r!\,s!\,t!}\,\bar{\alpha}_i^r\left[\bar{\eta}_i'(h)\right]^s\left[\bar{\zeta}_i'(h)\right]^t\frac{\partial^m u}{\partial x^r\,\partial y^s\,\partial z^t}.
\end{aligned}\right\} \tag{150}
$$

Очевидно,

$$
\left.\begin{aligned}
B_1^{(i)}\left[B_m^{(i)}(u)\right] &= B_{m+1}^{(i)}(u),\\
C_1^{(i)}\left[C_m^{(i)}(u)\right] &= C_{m+1}^{(i)}(u).
\end{aligned}\right\} \tag{151}
$$

Заметим также, что

$$
\frac{d}{dh}B_m^{(i)}\left[f(\xi_i,\ \eta_i,\ \zeta_i)\right] = B_{m+1}^{(i)}\left[f(\xi_i,\ \eta_i,\ \zeta_i)\right] +
$$

$$
+ m\eta_i''(h)\,B_{m-1}^{(i)}\left[\frac{\partial f(\xi_i,\ \eta_i,\ \zeta_i)}{\partial y}\right] + m\zeta_i''(h)\,B_{m-1}^{(i)}\left[\frac{\partial f(\xi_i,\ \eta_i,\ \zeta_i)}{\partial z}\right], \tag{152}
$$

$$
\frac{d}{dh}C_m^{(i)}\left[f(\bar{\xi}_i,\ \bar{\eta}_i,\ \bar{\zeta}_i)\right] = C_{m+1}^{(i)}\left[f(\bar{\xi}_i,\ \bar{\eta}_i,\ \bar{\zeta}_i)\right] +
$$

$$
+ m\bar{\eta}_i''(h)\,C_{m-1}^{(i)}\left[\frac{\partial f(\bar{\xi}_i,\ \bar{\eta}_i,\ \bar{\zeta}_i)}{\partial y}\right] + m\bar{\zeta}_i''(h)\,C_{m-1}^{(i)}\left[\frac{\partial f(\bar{\xi}_i,\ \bar{\eta}_i,\ \bar{\zeta}_i)}{\partial z}\right]. \tag{153}
$$

Операторы $B_m^{(i)}$ и $C_m^{(i)}$ зависят от параметра h. Их значение при $h=0$ будем отмечать чертой сверху.

Используя свойства операторов $B_m^{(i)}$ и $C_m^{(i)}$, находим (при $i > 1$):

$$\left.\begin{aligned}
k_i'(h) &= f(\xi_i, \eta_i, \zeta_i) + hB_1^{(i)}[f(\xi_i, \eta_i, \zeta_i)],\\
k_i'(h) &= 2B_1^{(i)}[f(\xi_i, \eta_i, \zeta_i)] + hB_2^{(i)}[f(\xi_i, \eta_i, \zeta_i)] +\\
&\quad + h\eta_i''(h)\frac{\partial f(\xi_i, \eta_i, \zeta_i)}{\partial y} + h\zeta_i''(h)\frac{\partial f(\xi_i, \eta_i, \zeta_i)}{\partial z},\\
k_i'''(h) &= 3B_2^{(i)}[f(\xi_i, \eta_i, \zeta_i)] + 3\eta_i''(h)\frac{\partial f(\xi_i, \eta_i, \zeta_i)}{\partial y} +\\
&\quad + 3\zeta_i''(h)\frac{\partial f(\xi_i, \eta_i, \zeta_i)}{\partial z} + hB_3^{(i)}[f(\xi_i, \eta_i, \zeta_i)] +\\
&\quad + 3h\eta_i''(h)B_1^{(i)}\left[\frac{\partial f(\xi_i, \eta_i, \zeta_i)}{\partial y}\right] + 3h\zeta_i''(h)B_1^{(i)}\left[\frac{\partial f(\xi_i, \eta_i, \zeta_i)}{\partial z}\right] +\\
&\quad + h\eta_i'''(h)\frac{\partial f(\xi_i, \eta_i, \zeta_i)}{\partial y} + h\zeta_i'''(h)\frac{\partial f(\xi_i, \eta_i, \zeta_i)}{\partial z},\\
k_i^{(IV)}(h) &= 4B_3^{(i)}[f(\xi_i, \eta_i, \zeta_i)] + 12\eta_i'''(h)B_1^{(i)}\left[\frac{\partial f(\xi_i, \eta_i, \zeta_i)}{\partial y}\right] +\\
&\quad + 12\zeta_i''(h)B_1^{(i)}\left[\frac{\partial f(\xi_i, \eta_i, \zeta_i)}{\partial z}\right] + 4\eta_i'''(h)\frac{\partial f(\xi_i, \eta_i, \zeta_i)}{\partial y} +\\
&\quad + 4\zeta_i'''(h)\frac{\partial f(\xi_i, \eta_i, \zeta_i)}{\partial z} + h\{\ldots\}.
\end{aligned}\right\} \quad (154)$$

Члены в фигурных скобках для нас сейчас значения иметь не будут. Таким образом,

$$\left.\begin{aligned}
k_i'(0) &= f_0,\\
k_i''(0) &= 2\bar{B}_1^{(i)}(f_0),\\
k_i'''(0) &= 3\bar{B}_2^{(i)}(f_0) + 3\eta_i''(0)\frac{\partial f_0}{\partial y} + 3\zeta_i''(0)\frac{\partial f_0}{\partial z},\\
k_i^{(IV)}(0) &= 4\bar{B}_3^{(i)}(f_0) + 12\eta_i''(0)\bar{B}_1^{(i)}\left(\frac{\partial f_0}{\partial y}\right) + 12\zeta_i''(0)\bar{B}_1^{(i)}\left(\frac{\partial f_0}{\partial z}\right) +\\
&\quad + 4\eta_i'''(0)\frac{\partial f_0}{\partial y} + 4\zeta_i'''(0)\frac{\partial f_0}{\partial z}.
\end{aligned}\right\} \quad (155)$$

Аналогично получим:

$$\left.\begin{aligned}
l_i'(0) &= g_0,\\
l_i''(0) &= 2\bar{C}_1^{(i)}(g_0),\\
l_i'''(0) &= 3\bar{C}_2^{(i)}(g_0) + 3\bar{\eta}_i''(0)\frac{\partial g_0}{\partial y} + 3\bar{\zeta}_i''(0)\frac{\partial g_0}{\partial z},\\
l_i^{(IV)}(0) &= 4\bar{C}_3^{(i)}(g_0) + 12\bar{\eta}_i''(0)\bar{C}_1^{(i)}\left(\frac{\partial g_0}{\partial y}\right) + 12\bar{\zeta}_i''(0)\bar{C}_1^{(i)}\left(\frac{\partial g_0}{\partial z}\right) +\\
&\quad + 4\bar{\eta}_i'''(0)\frac{\partial g_0}{\partial y} + 4\bar{\zeta}_i'''(0)\frac{\partial g_0}{\partial z}.
\end{aligned}\right\} \quad (156)$$

При $i=1$ будем иметь:

$$k_1'(h)=f_0; \quad k_1^{(j)}(h)\equiv 0 \quad (j>1); \quad l_1'(h)=g_0; \quad l_1^{(j)}(h)\equiv 0 \quad (j>1).$$

Будем получать формулы Рунге — Кутта, имеющие порядок погрешности на одном шаге h^5. Для этого необходимо, чтобы функции

$$\left.\begin{aligned}\varphi(h)&=y(x_0+h)-y(x_0)-[p_{41}k_1(h)+p_{42}k_2(h)+\\&\qquad+p_{43}k_3(h)+p_{44}k_4(h)],\\ \psi(h)&=z(x_0+h)-z(x_0)-[q_{41}l_1(h)+q_{42}l_2(h)+\\&\qquad+q_{43}l_3(h)+q_{44}l_4(h)]\end{aligned}\right\} \tag{157}$$

обладали свойством

$$\left.\begin{aligned}\varphi(0)=\varphi'(0)=\varphi''(0)=\varphi'''(0)=\varphi^{(\mathrm{IV})}(0)=0;\\ \psi(0)=\psi'(0)=\psi''(0)=\psi'''(0)=\psi^{(\mathrm{IV})}(0)=0.\end{aligned}\right\} \tag{158}$$

Равенство нулю первых производных даст

$$\left.\begin{aligned}p_{41}+p_{42}+p_{43}+p_{44}=1,\\ q_{41}+q_{42}+q_{43}+q_{44}=1.\end{aligned}\right\} \tag{159}$$

Из $\varphi''(0)=\psi''(0)=0$ следует:

$$\left.\begin{aligned}\bar{A}_1=2p_{42}\bar{B}_1^{(2)}+3p_{43}\bar{B}_1^{(3)}+2p_{44}\bar{B}_1^{(4)},\\ \bar{A}_1=2q_{42}\bar{C}_1^{(2)}+2q_{43}\bar{C}_1^{(3)}+2q_{44}\bar{C}_1^{(4)}.\end{aligned}\right\} \tag{160}$$

Приравнивая нулю третьи производные, получим:

$$\left.\begin{aligned}\bar{A}_2&=3p_{42}\bar{B}_2^{(2)}+3p_{43}\bar{B}_2^{(3)}+3p_{44}\bar{B}_2^{(4)},\\ \bar{A}_1&=6p_{43}\beta_{32}\bar{B}_1^{(3)}+6p_{44}\beta_{42}\bar{B}_1^{(2)}+6p_{44}\beta_{43}\bar{B}_1^{(3)},\\ \bar{A}_1&=6p_{43}\gamma_{32}\bar{C}_1^{(2)}+6p_{44}\gamma_{42}\bar{C}_1^{(2)}+6p_{44}\gamma_{43}\bar{C}_1^{(3)},\\ \bar{A}_2&=3q_{42}\bar{C}_2^{(2)}+3q_{43}\bar{C}_2^{(3)}+3q_{44}\bar{C}_2^{(4)},\\ \bar{A}_1&=6q_{43}\bar{\beta}_{32}\bar{B}_1^{(2)}+6q_{44}\bar{\beta}_{42}\bar{B}_1^{(2)}+6q_{44}\bar{\beta}_{43}\bar{B}_1^{(3)},\\ \bar{A}_1&=6q_{43}\bar{\gamma}_{32}\bar{C}_1^{(2)}+6q_{44}\bar{\gamma}_{42}\bar{C}_1^{(2)}+6q_{44}\bar{\gamma}_{43}\bar{C}_1^{(3)}.\end{aligned}\right\} \tag{161}$$

Наконец, равенство нулю четвертых производных даст

$$\begin{aligned}&\bar{A}_3(f_0)+3\bar{A}_1(f_0)\bar{A}_1\left(\frac{\partial f_0}{\partial y}\right)+3\bar{A}_1(g_0)\bar{A}_1\left(\frac{\partial f_0}{\partial z}\right)+\bar{A}_2(f_0)\frac{\partial f_0}{\partial y}+\\&+\bar{A}_1(f_0)\left(\frac{\partial f_0}{\partial y}\right)^2+\bar{A}_1(g_0)\frac{\partial f_0}{\partial y}\frac{\partial f_0}{\partial z}+\bar{A}_2(g_0)\frac{\partial f_0}{\partial z}+\bar{A}_1(f_0)\frac{\partial g_0}{\partial y}\frac{\partial f_0}{\partial z}+\\&+\bar{A}_1(g_0)\frac{\partial g_0}{\partial z}\frac{\partial f_0}{\partial z}=4p_{42}\bar{B}_3^{(2)}(f_0)+4p_{43}\bar{B}_3^{(3)}(f_0)+4p_{44}\bar{B}_3^{(4)}(f_0)+\\&+24p_{43}\gamma_{32}\bar{C}_1^{(2)}(g_0)\bar{B}_1^{(3)}\left(\frac{\partial f_0}{\partial z}\right)+24p_{44}\gamma_{42}\bar{C}_1^{(2)}(g_0)\bar{B}_1^{(4)}\left(\frac{\partial f_0}{\partial z}\right)+\end{aligned}$$

$$+24p_{44}\gamma_{43}\bar{C}_1^{(3)}(g_0)\bar{B}_1^{(4)}\left(\frac{\partial f_0}{\partial z}\right)+12p_{43}\beta_{32}\bar{B}_2^{(2)}(f_0)\frac{\partial f_0}{\partial y}+$$

$$+12p_{44}\beta_{42}\bar{B}_2^{(2)}(f_0)\frac{\partial f_0}{\partial y}+12p_{44}\beta_{43}\bar{B}_2^{(3)}(f_0)\frac{\partial f_0}{\partial y}+$$

$$+24p_{44}\beta_{43}\beta_{32}\bar{B}_1^{(2)}(f_0)\left(\frac{\partial f_0}{\partial y}\right)^2+24p_{43}\beta_{32}\bar{B}_1^{(2)}(f_0)\bar{B}_1^{(3)}\left(\frac{\partial f_0}{\partial y}\right)+$$

$$+24p_{44}\beta_{42}\bar{B}_1^{(2)}(f_0)\bar{B}_1^{(4)}\left(\frac{\partial f_0}{\partial y}\right)+24p_{44}\beta_{43}\bar{B}_1^{(3)}(f_0)\bar{B}_1^{(4)}\left(\frac{\partial f_0}{\partial y}\right)+$$

$$+24p_{44}\beta_{43}\gamma_{32}\bar{C}_1^{(2)}(g_0)\frac{\partial f_0}{\partial z}\frac{\partial f_0}{\partial y}+12p_{43}\gamma_{32}\bar{C}_2^{(2)}(g_0)\frac{\partial f_0}{\partial z}+$$

$$+12p_{44}\gamma_{42}\bar{C}_2^{(2)}(g_0)\frac{\partial f_0}{\partial z}+12p_{44}\gamma_{43}\bar{C}_2^{(3)}(g_0)\left(\frac{\partial f_0}{\partial z}\right)+$$

$$+24p_{44}\gamma_{43}\bar{\beta}_{32}\bar{B}_1^{(2)}(f_0)\frac{\partial g_0}{\partial y}\frac{\partial f_0}{\partial z}+24p_{44}\gamma_{43}\bar{\gamma}_{32}\bar{C}_1^{(2)}(g_0)\frac{\partial g_0}{\partial z}\frac{\partial f_0}{\partial z}. \qquad (162)$$

$$\bar{A}_3(g_0)+3\bar{A}_1(f_0)\bar{A}_1\left(\frac{\partial g_0}{\partial y}\right)+3\bar{A}_1(g_0)\bar{A}_1\left(\frac{\partial g_0}{\partial z}\right)+\bar{A}_2(f_0)\frac{\partial g_0}{\partial y}+$$

$$+\bar{A}_1(f_0)\frac{\partial f_0}{\partial y}\frac{\partial g_0}{\partial y}+\bar{A}_1(g_0)\frac{\partial f_0}{\partial z}\frac{\partial g_0}{\partial y}+\bar{A}_2(g_0)\frac{\partial g_0}{\partial z}+\bar{A}_1(f_0)\frac{\partial g_0}{\partial y}\frac{\partial g_0}{\partial z}+$$

$$+\bar{A}_1(g_0)\left(\frac{\partial g_0}{\partial z}\right)^2=4q_{42}\bar{C}_3^{(2)}(g_0)+4q_{43}\bar{C}_3^{(3)}(g_0)+4q_{44}\bar{C}_3^{(4)}(g_0)+$$

$$+24q_{43}\bar{\beta}_{32}\bar{B}_1^{(2)}(f_0)\bar{C}_1^{(3)}\left(\frac{\partial g_0}{\partial y}\right)+24q_{44}\bar{\beta}_{42}\bar{B}_1^{(2)}(f_0)\bar{C}_1^{(4)}\left(\frac{\partial g_0}{\partial y}\right)+$$

$$+24q_{44}\bar{\beta}_{43}\bar{B}_1^{(3)}(f_0)\bar{C}_1^{(4)}\left(\frac{\partial g_0}{\partial y}\right)+24q_{43}\bar{\gamma}_{32}\bar{C}_1^{(2)}(g_0)\bar{C}_1^{(3)}\left(\frac{\partial g_0}{\partial z}\right)+$$

$$+24q_{44}\bar{\gamma}_{42}\bar{C}_1^{(2)}(g_0)\bar{C}_1^{(4)}\left(\frac{\partial g_0}{\partial z}\right)+24q_{44}\bar{\gamma}_{43}\bar{C}_1^{(3)}(g_0)\bar{C}_1^{(4)}\left(\frac{\partial g_0}{\partial z}\right)+$$

$$+12q_{43}\bar{\beta}_{32}\bar{B}_2^{(2)}(f_0)\frac{\partial g_0}{\partial y}+12q_{44}\bar{\beta}_{42}\bar{B}_2^{(2)}(f_0)\frac{\partial g_0}{\partial y}+12q_{44}\bar{\beta}_{43}\bar{B}_2^{(3)}(f_0)\frac{\partial g_0}{\partial y}+$$

$$+24q_{44}\bar{\beta}_{43}\beta_{32}\bar{B}_1^{(2)}(f_0)\frac{\partial f_0}{\partial y}\frac{\partial g_0}{\partial y}+24q_{44}\bar{\beta}_{43}\gamma_{32}\bar{C}_1^{(2)}(g_0)\frac{\partial f_0}{\partial z}\frac{\partial g_0}{\partial y}+$$

$$+12q_{43}\bar{\gamma}_{32}\bar{C}_2^{(2)}(g_0)\frac{\partial g_0}{\partial z}+12q_{44}\bar{\gamma}_{42}\bar{C}_2^{(2)}(g_0)\frac{\partial g_0}{\partial z}+12q_{44}\bar{\gamma}_{43}\bar{C}_2^{(3)}(g_0)\frac{\partial g_0}{\partial z}+$$

$$+24q_{44}\bar{\gamma}_{43}\bar{\beta}_{32}\bar{B}_1^{(2)}(f_0)\frac{\partial g_0}{\partial y}\frac{\partial g_0}{\partial z}+24q_{44}\bar{\gamma}_{43}\bar{\gamma}_{32}\bar{C}_1^{(2)}(g_0)\left(\frac{\partial g_0}{\partial z}\right)^2. \qquad (163)$$

Из полученных равенств теми же рассуждениями, которые применялись для одного уравнения, покажем, что наши постоянные удовлетворяют следующей системе уравнений:

$$
\begin{array}{ll}
1.\ \alpha_2=\beta_{21}=\gamma_{21}, & \bar\alpha_2=\bar\beta_{21}=\bar\gamma_{21},\\
2.\ \alpha_3=\beta_{31}+\beta_{32}=\gamma_{31}+\gamma_{32}, & \bar\alpha_3=\bar\beta_{31}+\bar\beta_{32}=\bar\gamma_{31}+\bar\gamma_{32},\\
3.\ \alpha_4=\beta_{41}+\beta_{42}+\beta_{43}= & \bar\alpha_4=\bar\beta_{41}+\bar\beta_{42}+\bar\beta_{43}=\\
\quad=\gamma_{41}+\gamma_{42}+\gamma_{43}, & \quad=\bar\gamma_{41}+\bar\gamma_{42}+\bar\gamma_{43},\\
4.\ p_{41}+p_{42}+p_{43}+p_{44}=1, & q_{41}+q_{42}+q_{43}+q_{44}=1,\\
5.\ p_{42}\alpha_2+p_{43}\alpha_3+p_{44}\alpha_4=\frac{1}{2}, & q_{42}\bar\alpha_2+q_{43}\bar\alpha_3+q_{44}\bar\alpha_4=\frac{1}{2},\\
6.\ p_{42}\alpha_2^2+p_{43}\alpha_3^2+p_{44}\alpha_4^2=\frac{1}{3}, & q_{42}\bar\alpha_2^2+q_{43}\bar\alpha_3^2+q_{44}\bar\alpha_4^2=\frac{1}{3},\\
7.\ p_{42}\alpha_2^3+p_{43}\alpha_3^3+p_{44}\alpha_4^3=\frac{1}{4}, & q_{42}\bar\alpha_2^3+q_{43}\bar\alpha_3^3+q_{44}\bar\alpha_4^3=\frac{1}{4},\\
8.\ p_{43}\beta_{32}\alpha_2+p_{44}\beta_{42}\alpha_2+ & q_{43}\bar\gamma_{32}\bar\alpha_2+q_{44}\bar\gamma_{42}\bar\alpha_2+\\
\quad+p_{44}\beta_{43}\alpha_3=\frac{1}{6}, & \quad+q_{44}\bar\gamma_{43}\bar\alpha_3=\frac{1}{6},\\
9.\ p_{43}\beta_{32}\alpha_2\alpha_3+p_{44}\beta_{42}\alpha_2\alpha_4+ & q_{43}\bar\gamma_{32}\bar\alpha_2\bar\alpha_3+q_{44}\bar\gamma_{42}\bar\alpha_2\bar\alpha_4+\\
\quad+p_{44}\beta_{43}\alpha_3\alpha_4=\frac{1}{8}, & \quad+q_{44}\bar\gamma_{43}\bar\alpha_3\bar\alpha_4=\frac{1}{8},\\
10.\ p_{43}\beta_{32}\alpha_2^2+p_{44}\beta_{42}\alpha_2^2+ & q_{43}\bar\gamma_{32}\bar\alpha_2^2+q_{44}\bar\gamma_{42}\bar\alpha_2^2+\\
\quad+p_{44}\beta_{43}\alpha_3^2=\frac{1}{12}, & \quad+q_{44}\bar\gamma_{43}\bar\alpha_3^2=\frac{1}{12},\\
11.\ p_{44}\beta_{43}\beta_{32}\alpha_2=\frac{1}{24}, & q_{44}\bar\gamma_{43}\bar\gamma_{32}\bar\alpha_2=\frac{1}{24},\\
12.\ p_{43}\gamma_{32}\bar\alpha_2+p_{44}\gamma_{42}\bar\alpha_2+ & q_{43}\bar\beta_{32}\alpha_2+q_{44}\bar\beta_{42}\alpha_2+\\
\quad+p_{44}\gamma_{43}\bar\alpha_3=\frac{1}{6}, & \quad+q_{44}\bar\beta_{43}\alpha_3=\frac{1}{6},\\
13.\ p_{43}\gamma_{32}\alpha_3\bar\alpha_2+p_{44}\gamma_{42}\alpha_4\bar\alpha_2+ & q_{43}\bar\beta_{32}\alpha_2\bar\alpha_3+q_{44}\bar\beta_{42}\alpha_2\bar\alpha_4+\\
\quad+p_{44}\gamma_{43}\alpha_4\bar\alpha_3=\frac{1}{8}, & \quad+q_{44}\bar\beta_{43}\alpha_3\bar\alpha_4=\frac{1}{8},\\
14.\ p_{43}\gamma_{32}\bar\alpha_2^2+p_{44}\gamma_{42}\bar\alpha_2^2+ & q_{43}\bar\beta_{32}\alpha_2^2+q_{44}\bar\beta_{42}\alpha_2^2+\\
\quad+p_{44}\gamma_{43}\bar\alpha_3^2=\frac{1}{12}, & \quad+q_{44}\bar\beta_{43}\alpha_3^2=\frac{1}{12},\\
15.\ p_{44}\gamma_{43}\bar\gamma_{32}\bar\alpha_2=\frac{1}{24}, & q_{44}\bar\beta_{43}\beta_{32}\alpha_2=\frac{1}{24},\\
16.\ p_{44}\beta_{43}\gamma_{32}\bar\alpha_2=\frac{1}{24}, & q_{44}\bar\gamma_{43}\bar\beta_{32}\alpha_2=\frac{1}{24},\\
17.\ p_{44}\gamma_{43}\bar\beta_{32}\alpha_2=\frac{1}{24}, & q_{44}\bar\beta_{43}\gamma_{32}\bar\alpha_2=\frac{1}{24}.
\end{array}
\qquad (164)
$$

Проанализируем теперь систему (164). Возьмем 11, 15, 16 и 17-е уравнения системы в левом столбце. Легко видеть, что эти уравнения эквивалентны системе уравнений

$$\left.\begin{aligned} p_{44}\beta_{43}\beta_{32}\alpha_2 &= \frac{1}{24}, \\ \frac{\beta_{43}}{\gamma_{43}} = \frac{\bar\beta_{32}}{\bar\beta_{32}}, \quad \frac{\bar\alpha_3}{\alpha_2} &= \frac{\beta_{32}}{\gamma_{32}} = \frac{\bar\beta_{32}}{\bar\gamma_{32}}. \end{aligned}\right\} \tag{165}$$

Четыре соответствующих уравнения правого столбца будут эквивалентны системе

$$\left.\begin{aligned} q_{44}\bar\gamma_{43}\bar\gamma_{32}\bar\alpha_2 &= \frac{1}{24}, \\ \frac{\bar\beta_{43}}{\bar\gamma_{43}} = \frac{\bar\gamma_{32}}{\bar\gamma_{32}}, \quad \frac{\bar\alpha_3}{\alpha_2} &= \frac{\beta_{32}}{\gamma_{32}} = \frac{\bar\beta_{32}}{\bar\gamma_{32}}. \end{aligned}\right\} \tag{166}$$

Таким образом, вместо восьми исходных уравнений можно взять шесть следующих:

$$p_{44}\beta_{43}\beta_{32}\alpha_2 = q_{44}\bar\gamma_{43}\bar\gamma_{32}\bar\alpha_2 = \frac{1}{24}, \tag{167}$$

$$\frac{\beta_{43}}{\gamma_{43}} = \frac{\bar\beta_{32}}{\beta_{32}}, \quad \frac{\bar\beta_{43}}{\bar\gamma_{43}} = \frac{\bar\gamma_{32}}{\gamma_{32}}; \quad \frac{\bar\alpha_3}{\alpha_3} = \frac{\beta_{32}}{\gamma_{32}} = \frac{\bar\beta_{32}}{\bar\gamma_{32}}. \tag{168}$$

Используя равенства (168), мы получим из 8-х и 9-х уравнений (164):

$$\left.\begin{aligned} \beta_{42}\alpha_2 + \beta_{43}\alpha_3 &= \gamma_{42}\bar\alpha_2 + \gamma_{43}\bar\alpha_3, \\ \bar\beta_{42}\alpha_2 + \bar\beta_{43}\alpha_3 &= \bar\gamma_{42}\bar\alpha_2 + \bar\gamma_{43}\bar\alpha_3. \end{aligned}\right\} \tag{169}$$

Тринадцатые уравнения (164) будут являться следствиями 9-х уравнений и (169). Исключая p_{ij} и q_{ij} из 10, 11 и 14-го уравнений и используя (169), найдем:

$$\left.\begin{aligned} \beta_{43}\bar\alpha_2\alpha_3(\alpha_2 - \alpha_3) - \gamma_{43}\alpha_2\bar\alpha_3(\bar\alpha_2 - \bar\alpha_3) &= 2(\alpha_2 - \bar\alpha_2)\beta_{43}\beta_{32}\alpha_2, \\ \bar\beta_{43}\bar\alpha_2\alpha_3(\alpha_2 - \alpha_3) - \bar\gamma_{43}\alpha_2\bar\alpha_3(\bar\alpha_2 - \bar\alpha_3) &= 2(\alpha_2 - \bar\alpha_2)\bar\gamma_{43}\bar\gamma_{32}\bar\alpha_2. \end{aligned}\right\} \tag{170}$$

Уравнения (170) не являются независимыми, так как по (168)

$$\frac{\bar\beta_{32}}{\beta_{32}} = \frac{\bar\gamma_{32}}{\gamma_{32}} \tag{171}$$

и, следовательно, опять по (168)

$$\frac{\beta_{43}}{\gamma_{43}} = \frac{\bar{\beta}_{43}}{\bar{\gamma}_{43}} = \frac{\bar{\beta}_{32}}{\beta_{32}} = \frac{\bar{\gamma}_{32}}{\gamma_{32}}; \quad \frac{\beta_{43}}{\bar{\beta}_{43}} = \frac{\gamma_{43}}{\bar{\gamma}_{43}}. \tag{172}$$

Кроме того,

$$\frac{\beta_{43}\beta_{32}\alpha_2}{\bar{\gamma}_{43}\bar{\gamma}_{32}\bar{\alpha}_2} = \frac{\beta_{43}}{\bar{\beta}_{43}}. \tag{173}$$

Таким образом, достаточно взять одно из уравнений (170), например первое. Оказывается и оно является следствием предыдущих. Действительно, если рассматривать первые 11 уравнений (164) как систему для определения α_i, β_{ij}, p_{ri}, то должно быть выполнено четвертое из соотношений (130). Точно так же, если рассматривать эти уравнения как систему для определения $\bar{\alpha}_i$, $\bar{\gamma}_{ij}$, q_{ri}, то должно быть выполнено соответствующее соотношение

$$\bar{\alpha}_3(\bar{\alpha}_3 - \bar{\alpha}_2) = 2\bar{\gamma}_{32}\bar{\alpha}_2 - 4\bar{\gamma}_{32}\bar{\alpha}_2^2. \tag{174}$$

Последовательно упрощая первое из уравнений (170) и используя при этом (168) и указанные соотношения, мы придем в конце концов к тождеству

$$\alpha_2 - \bar{\alpha}_2 = \alpha_2 - \bar{\alpha}_2. \tag{175}$$

Итак, искомые постоянные должны удовлетворять 1—11 уравнениям (164) и, кроме того, (168) и (169). Проще всего решать полученную систему следующим образом. Находим α_i, β_{ij}, p_{ri} так, как это указывалось ранее. Таким же образом находим $\bar{\alpha}_i$, $\bar{\gamma}_{ij}$, q_{ri}. Затем последовательно получаем γ_{ij} и $\bar{\beta}_{ij}$, используя первые три строки (164), (168) и (169). Это даст

$$\left.\begin{aligned}
&\gamma_{21} = \alpha_2, && \bar{\beta}_{21} = \bar{\alpha}_2,\\
&\gamma_{32} = \frac{\beta_{32}\alpha_2}{\bar{\alpha}_2}, && \bar{\beta}_{32} = \frac{\bar{\gamma}_{32}\bar{\alpha}_2}{\alpha_2},\\
&\gamma_{31} = \alpha_3 - \gamma_{32}, && \bar{\beta}_{31} = \bar{\alpha}_3 - \bar{\beta}_{32},\\
&\gamma_{43} = \frac{\beta_{43}\beta_{32}}{\bar{\beta}_{32}}, && \bar{\beta}_{43} = \frac{\bar{\gamma}_{43}\bar{\gamma}_{32}}{\gamma_{32}},\\
&\gamma_{42} = \beta_{42}\frac{\alpha_2}{\bar{\alpha}_2} + \beta_{43}\frac{\alpha_3}{\bar{\alpha}_2} - \gamma_{43}\frac{\bar{\alpha}_3}{\bar{\alpha}_2}, && \bar{\beta}_{42} = \bar{\gamma}_{42}\frac{\bar{\alpha}_2}{\alpha_2} + \bar{\gamma}_{43}\frac{\bar{\alpha}_3}{\alpha_2} - \bar{\beta}_{43}\frac{\alpha_3}{\alpha_2},\\
&\gamma_{41} = 1 - \gamma_{42} - \gamma_{43}, && \beta_{41} = 1 - \bar{\beta}_{42} - \bar{\beta}_{43}.
\end{aligned}\right\} \tag{176}$$

Проще всего просто положить $\alpha_i = \bar{\alpha}_i$, $\beta_{ij} = \bar{\beta}_{ij}$, $\gamma_{ij} = \bar{\gamma}_{ij}$, $p_{ri} = q_{ri}$. Легко видеть, что это не противоречит нашей системе. Так, например, можно пользоваться формулами:

$$\left.\begin{aligned} \Delta y_0 &= \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4), \\ \Delta z_0 &= \frac{1}{6}(l_1 + 2l_2 + 2l_3 + l_4), \end{aligned}\right\} \tag{177}$$

где

$$\left.\begin{aligned} k_1 &= hf_0, & l_1 &= hg_0, \\ k_2 &= hf\left(x_0 + \frac{h}{2},\ y_0 + \frac{k_1}{2},\ z_0 + \frac{l_1}{2}\right), & & \\ & & l_2 &= hg\left(x_0 + \frac{h}{2},\ y_0 + \frac{k_1}{2},\ z_0 + \frac{l_1}{2}\right), \\ k_3 &= hf\left(x_0 + \frac{h}{2},\ y_0 + \frac{k_2}{2},\ z_0 + \frac{l_2}{2}\right), & & \\ & & l_3 &= hg\left(x_0 + \frac{h}{2},\ y_0 + \frac{k_2}{2},\ z_0 + \frac{l_2}{2}\right), \\ k_4 &= hf(x_0 + h,\ y_0 + k_3,\ z_0 + l_3), & & \\ & & l_4 &= hg(x_0 + h,\ y_0 + k_3,\ z_0 + l_3). \end{aligned}\right\} \tag{178}$$

Последний вывод можно применить для произвольных систем обыкновенных дифференциальных уравнений первого порядка.

3. Метод Рунге—Кутта решения уравнений второго порядка. Уравнения высших порядков могут быть сведены к системе уравнений первого порядка и, следовательно, к ним также будет применим метод Рунге — Кутта. Но так как в этом случае правые части будут иметь очень простой вид, то можно получить более простые схемы для их решения.

Рассмотрим уравнение второго порядка

$$y'' = f(x, y, y') \tag{179}$$

и будем отыскивать его решение, удовлетворяющее начальным данным: $y(x_0) = y_0$, $y'(x_0) = y'_0$. Это уравнение может быть сведено к системе

$$\left.\begin{aligned} y' &= z, \\ z' &= f(x, y, z). \end{aligned}\right\} \tag{180}$$

Чтобы сократить записи, будем подробно разбирать только формулы, имеющие порядок погрешности на одном шаге h^4. При этом

уравнения для определения α_i, β_{ij}, γ_{ij}, $\bar{\alpha}_i$, $\bar{\beta}_{ij}$, $\bar{\gamma}_{ij}$, p_{ri}, q_{ri} примут вид:

$$\left.\begin{array}{ll}
1.\ \alpha_2=\beta_{21}=\gamma_{21}, & \bar{\alpha}_2=\bar{\beta}_{21}=\bar{\gamma}_{21},\\
2.\ \alpha_3=\beta_{31}+\beta_{32}=\gamma_{31}+\gamma_{32}, & \bar{\alpha}_3=\bar{\beta}_{31}+\bar{\beta}_{32}=\bar{\gamma}_{31}+\bar{\gamma}_{32},\\
3.\ p_{31}+p_{32}+p_{33}=1, & q_{31}+q_{32}+q_{33}=1,\\
4.\ p_{32}\alpha_2+p_{33}\alpha_3=\frac{1}{2}, & q_{32}\bar{\alpha}_2+q_{33}\bar{\alpha}_3=\frac{1}{2},\\
5.\ p_{32}\alpha_2^2+p_{33}\alpha_3^2=\frac{1}{3}, & q_{32}\bar{\alpha}_2^2+q_{33}\alpha_3^2=\frac{1}{3},\\
6.\ p_{33}\beta_{32}\alpha_2=\frac{1}{6}, & q_{33}\bar{\gamma}_{32}\bar{\alpha}_2=\frac{1}{6},\\
\multicolumn{2}{c}{\dfrac{\bar{\alpha}_2}{\alpha_2}=\dfrac{\beta_{32}}{\gamma_{32}}=\dfrac{\bar{\beta}_{32}}{\bar{\gamma}_{32}}.}
\end{array}\right\}\quad(181)$$

Для системы (180) получим:

$$\left.\begin{array}{l}
k_1(h)=hy_0', \qquad l_1(h)=hf_0,\\
k_2(h)=h\left(y_0'+\gamma_{21}hf_0\right),\\
\qquad l_2(h)=hf\left(x_0+\bar{\alpha}_2h,\ y_0+\bar{\beta}_{21}hy_0',\ y_0'+\bar{\gamma}_{21}hf_0\right),\\
k_3(h)=h\left[y_0'+\gamma_{31}hf_0+\right.\\
\qquad \left.+\gamma_{32}hf\left(x_0+\bar{\alpha}_2h,\ y_0+\bar{\beta}_{21}hy_0',\ y_0'+\bar{\gamma}_{21}hf_0\right)\right],\\
\qquad l_3(h)=hf\left[x_0+\bar{\alpha}_3h,\ y_0+\bar{\alpha}_3hy_0'+\bar{\beta}_{32}\gamma_{21}h^2f_0,\ y_0'+\right.\\
\left.+\bar{\gamma}_{31}hf_0+\bar{\gamma}_{32}hf\left(x_0+\bar{\alpha}_2h,\ y_0+\bar{\beta}_{21}h\bar{y}_0',\ y_0'+\bar{\gamma}_{21}hf_0\right)\right],\\
\Delta y_0=p_{31}k_1+p_{32}k_2+p_{33}k_3=hy_0'+h^2\left\{\left(p_{32}\gamma_{21}+p_{33}\gamma_{31}\right)f_0+\right.\\
\qquad \left.+p_{33}\gamma_{32}f\left(x_0+\bar{\alpha}_2h,\ y_0+\bar{\beta}_{21}hy_0',\ y_0'+\bar{\gamma}_{21}hf_0\right)\right\},\\
\Delta z_0=q_{31}l_1+q_{32}l_2+q_{33}l_3.
\end{array}\right\}\quad(182)$$

Наиболее простые формулы получатся, если

$$p_{32}\gamma_{21}+p_{33}\gamma_{31}=0. \qquad (183)$$

В этом случае в силу второго и четвертого уравнений (181) $p_{33}\gamma_{32}=\frac{1}{2}$. Такому условию удовлетворяет, если положить $\alpha_i=\bar{\alpha}_i$;

$\beta_{ij}=\bar{\beta}_{ij}=\gamma_{ij}=\bar{\gamma}_{ij}$, вариант б) формул, имеющих порядок ошибки h^4, рассмотренной нами ранее. При этом

$$\left.\begin{aligned}
&\alpha_2=\bar{\alpha}_2=\beta_{21}=\bar{\beta}_{21}=\gamma_{21}=\bar{\gamma}_{21}=\frac{1}{3},\\
&\alpha_3=\bar{\alpha}_3=\beta_{32}=\bar{\beta}_{32}=\gamma_{32}=\bar{\gamma}_{32}=\frac{2}{3},\\
&\beta_{31}=\bar{\beta}_{31}=\gamma_{31}=\bar{\gamma}_{31}=0,\\
&p_{33}=q_{33}=\frac{3}{4};\quad p_{32}=q_{32}=0;\quad p_{31}=q_{31}=\frac{1}{4}
\end{aligned}\right\} \tag{184}$$

и

$$\left.\begin{aligned}
&\Delta y_0=hy_0'+\frac{1}{2}h^2f\left(x_0+\frac{1}{3}h,\ y_0+\frac{1}{3}hy_0',\ y_0'+\frac{1}{3}hf_0\right),\\
&\Delta z_0=y_1'-y_0'=\frac{1}{4}hf_0+\frac{3}{4}hf\left[x_0+\frac{2}{3}h,\ y_0+\frac{2}{3}hy_0'+\right.\\
&\left.+\frac{2}{9}h^2f_0,\ y_0'+\frac{2}{3}hf\left(x_0+\frac{1}{3}h,\ y_0+\frac{h}{3}y_0',\ y_0'+\frac{1}{3}hf_0\right)\right].
\end{aligned}\right\} \tag{185}$$

Проиллюстрируем применение этой формулы на примере уравнения

$$y''=y. \tag{186}$$

Найдем два его частные решения. Одно из них должно удовлетворять начальным условиям $y(0)=1$, $y'(0)=1$, другое $y(0)=1$, $y'(0)=-1$. Точными решениями в этом случае будут e^x и e^{-x}. Правый столбец отведем для значений точного решения. Вычисления в нашем случае упростятся благодаря тому, что в правой части отсутствует y'. Ход вычислений будет виден из первой таблицы. (См. стр. 323).

Как мы видим, результаты получились неплохие.

Аналогичные схемы можно получить и для других вариантов значений α_i, β_{ij}, p_{ri}. Приведем три готовые схемы, не входя в подробности их получения (см. стр. 326).

Мы уже говорили о том, как можно производить оценку погрешности формул Рунге — Кутта для одного уравнения первого порядка. Аналогичные рассуждения годятся и для систем уравнений и уравнений высших порядков. Но оценки эти будут очень грубыми, так как они получатся в результате сложения большого числа отдельно оцениваемых выражений. Не будем здесь проводить всех выкладок, так как они очень громоздки, а практически ценность результата незначительна.

Для варианта а) формул Рунге — Кутта, имеющих порядок погрешности на одном шаге h^5, Бибербахом была получена следующая оценка:

$$|y(x_1)-y_1|<\frac{6MN|x_1-x_0|^5|N^5-1|}{|N-1|}. \tag{187}$$

$y'' = y$; $x_0 = 0$; $y_0 = 1$; $y_0' = 1$; $h = 0{,}1$; $x \in [0, 1]$

№	x	y	k, Δy	$\bar{f} = h^2 f$	e^x
	x_0 $x_{01} = x_0 + \frac{1}{3} h$ $x_{02} = x_0 + \frac{2}{3} h$	y_0 $y_{01} = y_0 + \frac{1}{3} k_1$ $y_{02} = y_0 + \frac{2}{3} k_2$	$k_1 = h y_0'$ $k_2 = k_1 + \frac{1}{3} \bar{f}_0$ $\Delta y_0 = k_1 + \frac{1}{2} \bar{f}_{01}$	$\bar{f}_0 = h^2 f_0$ $\bar{f}_{01} = h^2 f(x_{01}, y_{01})$ $\bar{f}_{02} = h^2 f(x_{02}, y_{02})$	
0	0 0,0333 0,0667	1 1,0333 1,0669	0,1 0,1003 0,1052	0,01 0,0103 0,0107	1
1	0,1 0,1333 0,1667	1,1052 1,1420 1,1813	0,1105 0,1142 0,1162	0,0111 0,0114 0,0118	1,1052
2	0,2 0,2333 0,2667	1,2214 1,2621 1,3055	0,1221 0,1262 0,1284	0,0122 0,0126 0,0131	1,2214
3	0,3 0,3333 0,3667	1,3498 1,3948 1,4428	0,1350 0,1395 0,1420	0,0135 0,0139 0,0144	1,3499
4	0,4 0,4333 0,4667	1,4918 1,5415 1,5946	0,1492 0,1542 0,1569	0,0149 0,0154 0,0159	1,4918
5	0,5 0,5333 0,5667	1,6487 1,7036 1,7622	0,1648 0,1703 0,1733	0,0165 0,0170 0,0176	1,6487
6	0,6 0,6333 0,6667	1,8220 1,8827 1,9475	1,1821 0,1882 0,1915	0,0182 0,0188 0,0195	1,8221
7	0,7 0,7333 0,7667	2,0135 2,0806 2,1522	0,2013 0,2080 0,2117	0,0201 0,0208 0,0215	2,0138

Продолжение

№	x	y	k, Δy	$F = h^2 f$	e^x
8	0,8 0,8333 0,8667	2,2252 2,2994 2,3785	0,2225 0,2299 0,2340	0,0223 0,0230 0,0238	2,2255
9	0,9 0,9333 0,9667	2,4592 2,5412 2,6286	0,2459 0,2541 0,2586	0,0246 0,0254 0,0263	2,4596
10	1,0	2,7178			2,7183

$y'' = y; \quad x_0 = 0, \quad y_0 = 1; \quad y_0' = -1; \quad h = 0{,}1; \quad x \in [0,1]$

№	x	y	k, Δy	$\bar{f} = h^2 f$	e^{-x}
0	0 0,0333 0,0667	1 0,9667 0,9355	— 0,1 — 0,0967 — 0,0952	0,01 0,0097 0,0094	1
1	0,1 0,1333 0,1667	0,9048 0,8747 0,8465	— 0,0904 — 0,0874 — 0,0860	0,0090 0,0087 0,0085	0,9048
2	0,2 0,2333 0,2667	0,8188 0,7915 0,7671	— 0,0818 — 0,0791 — 0,0778	0,0082 0,0079 0,0077	0,8187
3	0,3 0,3333 0,3667	0,7410 0,7163 0,6933	— 0,0740 — 0,0715 — 0,0704	0,0074 0,0072 0,0069	0,7408
4	0,4 0,4333 0,4667	0,6706 0,6483 0,6059	— 0,0670 — 0,0648 — 0,0638	0,0067 0,0065 0,0061	0,6703
5	0,5 0,5333 0,5667	0,6068 0,5865 0,5676	— 0,0608 — 0,0588 — 0,0578	0,0061 0,0059 0,0057	0,6065

Продолжение

№	x	y	k, Δy	$\bar{f} = h^2 f$	e^{-x}
6	0,6 0,6333 0,6667	0,5490 0,5307 0,5135	— 0,0550 — 0,0532 — 0,0524	0,0055 0,0053 0,0051	0,5488
7	0,7 0,7333 0,7667	0,4966 0,4800 0,4645	— 0,0498 — 0,0481 — 0,0474	0,0050 0,0048 0,0046	0,4966
8	0,8 0,8333 0,8667	0,4492 0,4342 0,4201	— 0,0451 — 0,0436 — 0,0429	0,0045 0,0043 0,0042	0,4493
9	0,9 0,9333 0,9667	0,4063 0,3927 0,3800	— 0,0408 — 0,0394 — 0,0388	0,0041 0,0039 0,0038	0,4066
10	1,0	0,3675			0,3679

$y'' = f(x, y)$, порядок ошибки h^4

№	x	y	k; Δy	$\bar{f} = h^2 f$
0	x_0 $x_{01} = x_0 + \frac{1}{2} h$ $x_{02} = x_0 + \frac{3}{4} h$	y_0 $y_{01} = y_0 + \frac{1}{2} h y'_0$ $y_{02} = y_0 + \frac{3}{4} k_0$	$k_1 = h y'_0$ $k_2 = k_1 + \frac{1}{2} \bar{f}_0$ $\Delta y_0 = k_1 +$ $+ \frac{1}{6} (\bar{f}_0 + 2\bar{f}_{01})$	$\bar{f}_0 = h^2 f_0$ $\bar{f}_{01} = h^2 f(x_{01}, y_{01})$ $\bar{f}_{02} = h^2 f(x_{02}, y_{02})$
1	$x_1 = x_0 + h$	$y_1 = y_0 + \Delta y_0$	$k'_1 = \Delta y_0 +$ $+ \frac{1}{18} (\bar{f}_0 + 8\bar{f}_{02})$	$\bar{f}_1 = h^2 f(x_1, y_1)$
. .	. . . . .	. . . . .	. . . . .	. . . . .

$y'' = f(x, y, y')$, порядок ошибки h^5

№	x	y	$v = ky'$	$k = \frac{h^2}{2} f\left(x, y, \frac{v}{h}\right)$	Добавки
0	x_0	y_0	v_{10}	k_1	$k = \frac{1}{3}(k_1 + k_2 + k_3)$
	$x_0 + \frac{h}{2}$	$y_0 + \frac{1}{2} v_{10} + \frac{1}{4} k_1$	$v_{10} + k_1$	k_2	
	$x_0 + \frac{h}{2}$	$y_0 + \frac{1}{2} v_{10} + \frac{1}{4} k_1$	$v_{10} + k_2$	k_3	$k' = \frac{1}{3}(k_1 + 2k_2 + 2k_3 + k_4)$
	$x_0 + h$	$y_0 + v_{10} + k_3$	$v_{10} + 2k_3$	k_4	
1	$x_1 = x_0 + h$	$y_1 = y_0 + v_{10} + k$	$v_{11} = v_{10} + k'$		
...	. . . .	. . . . .	. . . .	. . . . .	. . . . .

$y''' = f(x, y, y', y'')$, порядок ошибки h^5

№	x	y	$hy' = v_1$	$\frac{h^2}{2} y'' = v_2$	$k = \frac{h^3}{6} f\left(x, y, \frac{v_1}{h}, \frac{2v_2}{h^2}\right)$	Добавки
0	x_0	y_0	v_{10}	v_{20}	k_1	$k = \frac{1}{20}(9k_1 + 6k_2 +$
	$x_0 + \frac{h}{2}$	$y_0 + \frac{1}{2} v_{10} + \frac{1}{4} v_{20} + \frac{1}{8} k_1$	$v_{10} + v_{20} + \frac{3}{4} k_1$	$v_{20} + \frac{3}{2} k_1$	k_2	$+ 6k_3 - k_4)$
	$x_0 + \frac{h}{2}$	$y_0 + \frac{1}{2} v_{10} + \frac{1}{4} v_{20} + \frac{1}{8} k_1$	$v_{10} + v_{20} + \frac{3}{4} k_1$	$v_{20} + \frac{3}{2} k_2$	k_3	$k' = k_1 + k_2 + k_3$
	$x_0 + h$	$y_0 + v_{10} + v_{20} + k_3$	$v_{10} + 2v_{20} + 3k_3$	$v_{20} + 3k_3$	k_4	$k'' = \frac{1}{2}(k_1 + 2k_2 +$
						$+ 2k_3 + k_4)$
1	$x_1 = x_0 + h$	$y_1 = y_0 + v_{10} + v_{20} + k$	$v_{11} = v_{10} + 2v_{20} + k'$	$v_{21} = v_{20} + k''$		
...	. . . .	. . . . .	. . . . .	. . . .	. . . . .	. . . . .

Здесь $y(x_1)$ — точное решение в точке x_1, а y_1 — соответствующее приближенное значение. Предполагается, что в области $|x-x_0|<a$, $|y-y_0|<b$ $f(x, y)$ и ее производные до четвертого порядка включительно удовлетворяют условиям:

$$|f(x, y)|<M, \quad \left|\frac{\partial^{i+k} f}{\partial x^i \partial y^k}\right|<\frac{N}{M^{k-1}}, \tag{188}$$

$$|x-x_0|N<1, \quad aM<b. \tag{189}$$

Беря вместо (188) условия

$$|f(x, y)|<M, \quad \left|\frac{\partial^{i+k} f}{\partial x^i \partial y^k}\right|<\frac{L^{i+k}}{M^{k-1}}, \tag{190}$$

Лоткин получил другую оценку:

$$|y(x_1)-y_1|\leqslant\frac{73}{720}ML^4h^5, \quad h=x_1-x_0, \tag{191}$$

которая иногда выгоднее (187). И та и другая оценки очень грубы и практически малопригодны.

На практике пользуются приемом Рунге, производя вычисления дважды: один раз с шагом h, а втором — с шагом $2h$, или $\frac{h}{2}$. Об этом мы уже говорили в главе 3, когда обсуждали вопрос о практической пригодности формул остаточных членов при численном интегрировании.

§ 5. Разностные методы решения обыкновенных дифференциальных уравнений первого порядка

Большим недостатком метода Руиге — Кутта является то, что для получения одного нового значения решения дифференциального уравнения приходится подсчитывать правую часть уравнения в нескольких точках. Если правая часть сложна, это связано с большой вычислительной работой. Сейчас мы перейдем к разностным методам решения обыкновенных дифференциальных уравнений, применение которых требует только однократного вычисления правой части на каждом шаге. Ограничимся пока случаем одного уравнения первого порядка.

Пусть требуется найти решение уравнения

$$y'=f(x, y), \tag{1}$$

удовлетворяющее начальному условию $y(x_0)=y_0$. Предположим, что нам удалось каким-то образом найти приближенные значения $y(x)$ в точках x_m, x_{m-1}, ..., x_{m-k} ($x_{m-i}=x_m-ih$, h — шаг). Обозначим $y(x_i)=y_i$ и $f(x_i, y_i)=f_i$. По f_m, f_{m-1}, ..., f_{m-k} можно

построить то или иное приближенное представление $f[x, y(x)]$ в виде функции, которая легко интегрируется. Пусть это будет $\varphi(x)$. Тогда можно приближенно считать

$$y_{m+1} = y_{m-j} + \int_{x_{m-j}}^{x_{m+1}} \varphi(x)\, dx. \tag{2}$$

Таким образом, мы продвинем таблицу значений $y(x)$ на один шаг. Затем можно снова применить такой же прием и продвинуться еще на один шаг и т. д.

Для того чтобы было возможно начать этот процесс, нам необходимо знать, кроме начального значения y_0, значения $y(x)$ в точках $x_1, x_2, \ldots, x_k$. Их можно отыскивать либо методом Рунге — Кутта, либо одним из тех аналитических методов, о которых говорилось выше. Сами разностные методы дают итерационные способы для отыскания $y_1, y_2, \ldots, y_k$. Как это делается, мы укажем несколько позже.

Наиболее простой способ приближенного представления $f[x, y(x)]$ дает интерполяция алгебраическими многочленами. Только его мы и будем рассматривать. Обозначим через $L_{m,k}(x)$ алгебраический интерполяционный многочлен, принимающий в точках $x_m, x_{m-1}, \ldots, x_{m-k}$ соответственно значения $f_m, f_{m-1}, \ldots, f_{m-k}$. При этом

$$L_{m,k}(x) = \sum_{i=0}^{k} f_{m-i} P_i(x), \tag{3}$$

где $P_i(x)$ не зависят от f_j. Введем новое временное t, положив $x - x_m = th$. Тогда $P_i(x)$ перейдут в многочлены $Q_i(t)$ степени k, не зависящие ни от шага h, ни от m. Таким образом,

$$y_{m+1} = y_{m-j} + \int_{x_{m-j}}^{x_{m+1}} L_{m,k}(x)\, dx = y_{m-j} + h \sum_{i=0}^{k} \beta_i f_{m-i}, \tag{4}$$

где

$$\beta_i = \int_{-j}^{1} Q_i(t)\, dt \tag{5}$$

— постоянные, не зависящие ни от f, ни от m, ни от h.

Беря различные значения j и различные формы интерполяционного многочлена, получим различные разностные формулы численного интегрирования дифференциальных уравнений. Эти формулы носят название *экстраполяционных* в связи с тем, что они полу-

чены путем интегрирования интерполяционного многочлена, проэкстраполированного на отрезок $[x_m, x_{m+1}]$.

При построении интерполяционного многочлена можно использовать, кроме $f_m, f_{m-1}, \ldots, f_{m-k}$, еще неизвестное нам значение f_{m+1}. Повторяя предыдущие рассуждения, мы придем к формуле

$$y_{m+1} = y_{m-j} + h \sum_{i=-1}^{k} \gamma_i f_{m-i}, \tag{6}$$

при этом и в правую и в левую части входит неизвестное значение y_{m+1}. Поэтому для отыскания y_{m+1} нужно решить алгебраическое или трансцендентное уравнение. Чаще всего это уравнение решают методом последовательных приближений. Для сходимости его придется потребовать выполнения условия $h|\gamma_{-1}|M_1 < 1$, где $M_1 = \sup\left|\frac{\partial f}{\partial y}\right|$ в рассматриваемой области. Формулы типа (6) называются *интерполяционными*. Из второй главы известно, что точность экстраполирования обычно бывает меньше точности интерполирования. Поэтому следует ожидать, что формулы типа (6) будут давать лучшую точность, чем формулы типа (4). В справедливости этого утверждения мы убедимся при исследовании остаточных членов.

1. Некоторые экстраполяционные формулы для интегрирования дифференциальных уравнений первого порядка. Рассмотрим некоторые частные случаи разностных формул. Возьмем $j = 0$ и запишем $L_{m,k}(x)$ в виде интерполяционного многочлена Ньютона для интерполирования назад:

$$L_{m,k}(x) = f_m + t f^1_{m-\frac{1}{2}} + \frac{t(t+1)}{2} f^2_{m-1} + \ldots$$
$$\ldots + \frac{t(t+1)\ldots(t+k-1)}{k!} f^k_{m-\frac{k}{2}}. \tag{7}$$

Тогда

$$\Delta y_m = y_{m+1} - y_m = \int_{x_m}^{x_{m+1}} L_{m,k}(x)\,dx = h\int_0^1 L_{m,k}(x_m + th)\,dt =$$
$$= h\int_0^1 \left[f_m + t f^1_{m-\frac{1}{2}} + \frac{t(t+1)}{2} f^2_{m-1} + \ldots \right.$$
$$\left. \ldots + \frac{t(t+1)\ldots(t+k-1)}{k!} f^k_{m-\frac{k}{2}} \right] dt =$$
$$= h\left[f_m + a_1 f^1_{m-\frac{1}{2}} + a_2 f^2_{m-1} + \ldots + a_k f^k_{m-\frac{k}{2}} \right]. \tag{8}$$

Здесь

$$\left.\begin{aligned} a_1 &= \int_0^1 t\,dt = \frac{1}{2}, \\ a_2 &= \int_0^1 \frac{t(t+1)}{2}\,dt = \frac{5}{12}, \\ a_3 &= \int_0^1 \frac{t(t+1)(t+2)}{6}\,dt = \frac{3}{8}, \\ a_4 &= \int_0^1 \frac{t(t+1)(t+2)(t+3)}{24}\,dt = \frac{251}{720}, \\ a_5 &= \int_0^1 \frac{t(t+1)(t+2)(t+3)(t+4)}{120}\,dt = \frac{95}{288}. \end{aligned}\right\} \tag{9}$$

Дальнейшие значения для коэффициентов a_i будут

$$a_6 = \frac{19\,087}{60\,480}, \quad a_7 = \frac{5275}{17\,280}, \quad a_8 = \frac{1\,070\,017}{3\,628\,800}, \quad a_9 = \frac{1\,082\,753}{7\,257\,600}. \tag{10}$$

Таким образом, формула (8) примет вид

$$\Delta y_m = h\left[f_m + \frac{1}{2} f^1_{m-\frac{1}{2}} + \frac{5}{12} f^2_{m-1} + \frac{3}{8} f^3_{m-\frac{3}{2}} + \right.$$
$$\left. + \frac{251}{720} f^4_{m-2} + \frac{95}{288} f^5_{m-\frac{5}{2}} + \dots\right]. \tag{11}$$

Целесообразно ввести в рассмотрение величины $q_i = hf(x_i, y_i)$. Тогда формулу (11) можно переписать в виде

$$\Delta y_m = q_m + \frac{1}{2} q^1_{m-\frac{1}{2}} + \frac{5}{12} q^2_{m-1} + \frac{3}{8} q^3_{m-\frac{3}{2}} +$$
$$+ \frac{251}{720} q^4_{m-2} + \frac{95}{288} q^5_{m-\frac{5}{2}} + \dots \tag{12}$$

Эта формула носит название *экстраполяционной формулы Адамса*. Схема для вычислений Δy_m и y_{m+1} по экстраполяционной формуле Адамса будет выглядеть так:

x	y	Δy	$q = hf$	q^1	q^2	q^3
. . .	. . .	. . .	. . .	. . .	. . .	. . .
x_{m-3}	y_{m-3}		q_{m-3}			
		Δy_{m-3}		$q^1_{m-\frac{5}{2}}$		
x_{m-2}	y_{m-2}		q_{m-2}		q^2_{m-2}	
		Δy_{m-2}		$q^1_{m-\frac{3}{2}}$		$q^3_{m-\frac{3}{2}}$
x_{m-1}	y_{m-1}		q_{m-1}		q^2_{m-1}	. . .
		Δy_{m-1}		$q^1_{m-\frac{1}{2}}$	. . .	
x_m	y_m		q_m			
. . .	. . .	. . .	. . .	. . .		

Предполагая, что третьи разности почти постоянны, можно ограничиться первыми четырьмя членами формулы (12). По известным значениям y_{m-3}, y_{m-2}, y_{m-1}, y_m находим q_{m-3}, q_{m-2}, q_{m-1}, q_m. Затем по формуле Адамса находим значение Δy_m и, прибавляя его к y_m, находим y_{m+1}. Это позволит нам продвинуться в таблице значений q и ее разностей на один шаг вниз и получить по формуле Адамса еще одно значение Δy и т. д.

Иногда бывает целесообразно выражать значения Δy непосредственно через $y'_i = f_i$. Для этого выразим разности, входящие в формулу (12), через значения y'_j. Получим:

$$y_{m+1} - y_m = hy'_m + \frac{h}{2}(y'_m - y'_{m-1}) + \frac{5h}{12}(y'_m - 2y'_{m-1} + y'_{m-2}) + \\ + \frac{3h}{8}(y'_m - 3y'_{m-1} + 3y'_{m-2} - y'_{m-3}) + \ldots \qquad (13)$$

Если ограничиться одним членом правой части, то придем снова к *формуле Эйлера*

$$y_{m+1} = y_m + hy'_m. \qquad (14)$$

Два члена правой части дадут

$$y_{m+1} = y_m + \frac{h}{2}(3y'_m - y'_{m-1}). \qquad (15)$$

Три члена правой части дадут формулу

$$y_{m+1} = y_m + \frac{h}{12}(23y'_m - 16y'_{m-1} + 5y'_{m-2}). \qquad (16)$$

Четыре члена правой части приведут к

$$y_{m+1} = y_m + \frac{h}{24}(55y'_m - 59y'_{m-1} + 37y'_{m-2} - 9y'_{m-3}). \qquad (17)$$

Возьмем теперь $j = 1$ и в качестве $L_{m,k}(x)$ снова интерполяционный многочлен Ньютона для интерполирования назад (7). При этом

получим:

$$y_{m+1}-y_{m-1}=\int_{x_{m-1}}^{x_{m+1}}\left[f_m+tf^1_{m-\frac12}+\frac{t(t+1)}{2}f^2_{m-1}+\dots\right.$$

$$\left.\dots+\frac{t(t+1)\dots(t+k-1)}{k!}f^k_{m-\frac{k}{2}}\right]dx=h\int_{-1}^{+1}\left[f_m+tf^1_{m-\frac12}+\right.$$

$$\left.+\frac{t(t+1)}{2}f^2_{m-1}+\dots+\frac{t(t+1)\dots(t+k-1)}{k!}f^k_{m-\frac{k}{2}}\right]dt=$$

$$=a_0q_m+a_1q^1_{m-\frac12}+a_2q^2_{m-1}+\dots+a_kq^k_{m-\frac{k}{2}}. \tag{18}$$

Здесь

$$\left.\begin{aligned}
a_0&=\int_{-1}^{+1}dt=2,\\
a_1&=\int_{-1}^{+1}t\,dt=0,\\
a_2&=\int_{-1}^{+1}\frac{t(t+1)}{2}dt=\frac13,\\
a_3&=\int_{-1}^{+1}\frac{t(t+1)(t+2)}{6}dt=\frac13,\\
a_4&=\int_{-1}^{+1}\frac{t(t+1)(t+2)(t+3)}{24}dt=\frac{29}{90},\\
a_5&=\frac{28}{90},\quad a_6=\frac{18\,229}{60\,480},\quad a_7=\frac{35\,424}{120\,960},\\
a_8&=\frac{1\,036\,064}{3\,628\,800},\quad a_9=\frac{2\,025\,472}{7\,257\,600}.
\end{aligned}\right\} \tag{19}$$

Если ограничиться разностями до четвертого порядка и взять в качестве коэффициента при четвертой разности вместо $\frac{29}{90}$ число $\frac{30}{90}=\frac13$, т. е. изменить этот коэффициент всего лишь на $\frac{1}{90}$, то получим особенно простую для вычислений формулу:

$$y_{m+1}-y_m=2q_m+\frac13q^2_{m-1}+\frac13q^3_{m-\frac32}+\frac13q^4_{m-2}. \tag{20}$$

2. Примеры интерполяционных формул. Рассмотрим теперь примеры интерполяционных формул. Возьмем в качестве интерполяционного многочлена опять формулу Ньютона для интерполиро-

вания назад, но за начальную точку примем не x_m, а x_{m+1}:

$$L_{m,k} = f_{m+1} + tf^1_{m+\frac{1}{2}} + \frac{t(t+1)}{2!} f^2_m + \ldots$$
$$\ldots + \frac{t(t+1)\ldots(t+k-1)}{k!} f^k_{m+1-\frac{k}{2}}. \qquad (21)$$

При $j = 0$ интегрирование по t придется проводить по отрезку $[-1, 0]$. Получим:

$$\Delta y_m = y_{m+1} - y_m = \int_{x_m}^{x_{m+1}} \left[f_{m+1} + tf^1_{m+\frac{1}{2}} + \frac{t(t+1)}{2} f^2_m + \ldots \right.$$
$$\left. \ldots + \frac{t(t+1)\ldots(t+k-1)}{k!} f^k_{m+1-\frac{k}{2}} \right] dx = h \int_{-1}^{0} \left[f_{m+1} + tf^1_{m+\frac{1}{2}} + \right.$$
$$\left. + \frac{t(t+1)}{2} f^2_m + \ldots + \frac{t(t+1)\ldots(t+k-1)}{k!} f^k_{m+1-\frac{k}{2}} \right] dt =$$
$$= a_0 q_{m+1} + a_1 q^1_{m+\frac{1}{2}} + a_2 q^2_m + \ldots + a_k q^k_{m+1-\frac{k}{2}}. \qquad (22)$$

В этом случае

$$\left.\begin{aligned}
a_0 &= \int_{-1}^{0} dt = 1, \\
a_1 &= \int_{-1}^{0} t\,dt = -\frac{1}{2}, \\
a_2 &= \int_{-1}^{0} \frac{t(t+1)}{2}\,dt = -\frac{1}{12}, \\
a_3 &= \int_{-1}^{0} \frac{t(t+1)(t+2)}{6}\,dt = -\frac{1}{24}, \\
a_4 &= \int_{-1}^{0} \frac{t(t+1)(t+2)(t+3)}{24}\,dt = -\frac{19}{720}, \\
a_5 &= \int_{-1}^{0} \frac{t(t+1)(t+2)(t+3)(t+4)}{120}\,dt = -\frac{9}{160}, \\
a_6 &= -\frac{863}{60\,480}, \qquad a_7 = -\frac{275}{24\,195}, \\
a_8 &= \frac{33\,953}{3\,628\,800}, \qquad a_9 = -\frac{57\,281}{7\,257\,600}.
\end{aligned}\right\} \qquad (23)$$

Мы получили *интерполяционную формулу Адамса*:

$$\Delta y_m = q_{m+1} - \frac{1}{2} q^1_{m+\frac{1}{2}} - \frac{1}{12} q^2_m - \frac{1}{24} q^3_{m-\frac{1}{2}} - \frac{19}{720} q^4_{m-1} - \frac{9}{160} q^5_{m-\frac{3}{2}} - \dots \quad (24)$$

Эта формула может быть использована для контроля вычислений, произведенных по экстраполяционной формуле. Ее можно использовать и для уточнения приближенных значений для Δy_m, полученных другими способами. При этом можно пользоваться той же схемой, что и для экстраполяционной формулы Адамса. По приближенному значению y_{m+1} находим q_{m+1}, $q^1_{m+\frac{1}{2}}$, q^2_m, $q^3_{m-\frac{1}{2}}$. Затем используем интерполяционную формулу Адамса и уточняем значение Δy_m. Исправляем q_{m+1}, $q^1_{m+\frac{1}{2}}$, q^2_m, $q^3_{m-\frac{1}{2}}$. Снова по интерполяционной формуле Адамса уточняем Δy_m и затем q_{m+1}, $q^1_{m+\frac{1}{2}}$, q^2_m, $q^3_{m-\frac{1}{2}}$. Это продолжаем до тех пор, пока исправляемые величины не будут изменяться при той точности, с которой производятся вычисления. Первое приближение для Δy_m можно получить также путем экстраполяции высшей разности $q^3_{m-\frac{1}{2}}$ на один шаг и последующего вычисления разностей низшего порядка. Можно также предварительно представить Δy_m в виде линейной комбинации y_i и y'_i, а уж затем производить последовательные приближения. При этом если ограничиться первым членом формулы, то получим:

$$y_{m+1} = y_m + h y'_{m+1}. \quad (25)$$

Если взять два члена, то будем иметь:

$$y_{m+1} = y_m + \frac{h}{2}(y'_{m+1} + y'_m). \quad (26)$$

При трех членах получим:

$$y_{m+1} = y_m + \frac{h}{12}(5y'_{m+1} + 8y'_m - y'_{m-1}), \quad (27)$$

при четырех

$$y_{m+1} = y_m + \frac{h}{24}(9y'_{m+1} + 19y'_m - 5y'_{m-1} + y'_{m-2}), \quad (28)$$

при пяти

$$y_{m+1} = y_m + \frac{h}{720}(251y'_{m+1} + 646y'_m - 264y'_{m-1} + 106y'_{m-2} - 19y'_{m-3}). \quad (29)$$

Рассмотрим еще один случай. Пусть $j=1$, а в качестве $L_{m,k}$ возьмем интерполяционную формулу Стирлинга:

$$L_{m,k}(x)=f_m+tf_m^1+\frac{t^2}{2}f_m^2+\frac{t(t^2-1)}{6}f_m^3+\frac{t^2(t^2-1)}{24}f_m^4+\ldots \tag{30}$$

В результате интегрирования получим:

$$y_{m+1}-y_{m-1}=$$

$$=\int_{x_{m-1}}^{x_{m+1}}\left[f_m+tf_m^1+\frac{t^2}{2}f_m^2+\frac{t(t^2-1)}{6}f_m^3+\frac{t^2(t^2-1)}{24}f_m^4+\ldots\right]dx=$$

$$=h\int_{-1}^{+1}\left[f_m+tf_m^1+\frac{t^2}{2!}f_m^2+\frac{t(t^2-1)}{6}f_m^3+\frac{t^2(t^2-1)}{24}f_m^4+\ldots\right]dt=$$

$$=h\left[a_0f_m+a_1f_m^1+a_2f_m^2+a_3f_m^3+a_4f_m^4+\ldots\right]. \tag{31}$$

При этом

$$\left.\begin{aligned}
&a_0=2,\\
&a_1=\int_{-1}^{+1}t\,dt=0,\\
&a_2=\int_{-1}^{+1}\frac{t^2}{2}\,dt=\frac{1}{3},\\
&a_3=\int_{-1}^{+1}\frac{t(t^2-1)}{6}\,dt=0,\\
&a_4=\int_{-1}^{+1}\frac{t^2(t^2-1)}{24}\,dt=-\frac{1}{90},\\
&\ldots\ldots\ldots\ldots\ldots
\end{aligned}\right\} \tag{32}$$

Таким образом, наша формула примет вид

$$y_{m+1}-y_{m-1}=h\left[2f_m+\frac{1}{3}f_m^2-\frac{1}{90}f_m^4+\ldots\right]. \tag{33}$$

Если выразить здесь разности через y_i', то получим, принимая во внимание разности первого, второго и третьего порядка:

$$y_{m+1}-y_{m-1}=2hy_m', \tag{34}$$

$$y_{m+1}-y_{m-1}=\frac{h}{3}\left[y_{m+1}'+4y_m'+y_{m-1}'\right], \tag{35}$$

$$y_{m+1}-y_{m-1}=\frac{h}{90}\left[-y_{m+2}'+34y_{m+1}'+114y_m'+26y_{m-1}'-y_{m-2}'\right]. \tag{36}$$

Последняя формула содержит y'_{m+2} и вряд ли будет полезна на практике.

При желании набор формул можно было бы значительно увеличить. Можно получать новые формулы беря линейные комбинации формул, полученных для различных j и различных $L_{m,k}(x)$. Все эти формулы можно записать в виде

$$\alpha_k y_{m+k} + \alpha_{k-1} y_{m+k-1} + \ldots + \alpha_0 y_m = \\ = h\left[\beta_k y'_{m+k} + \beta_{k-1} y'_{m+k-1} + \ldots + \beta_0 y'_m\right], \qquad (37)$$

где $y'_i = f(x_i, y_i)$, α_i и β_i — постоянные, не зависящие ни от $y(x)$, ни от h, $\alpha_k \neq 0$. Характерным для полученных ранее формул было то, что они давали точное значение для $y(x)$, если $y(x)$ является алгебраическим многочленом степени не выше некоторого p.

3. Метод неопределенных коэффициентов вывода разностных формул. Можно и не прибегать к помощи интерполирования для получения формул (37), а воспользоваться методом неопределенных коэффициентов. Разложим $y_{m+\nu}$ и $y'_{m+\nu}$ по формуле Тейлора до членов с производными порядка $p+1$. Получим:

$$y_{m+\nu} = y_m + \nu h y'_m + \frac{\nu^2 h^2}{2} y''_m + \ldots + \frac{\nu^p h^p}{p!} y^{(p)}_m + \\ + \frac{\nu^{p+1} h^{p+1}}{(p+1)!} y^{(p+1)}_m + o\left(h^{p+1}\right), \qquad (38)$$

$$y'_{m+\nu} = y'_m + \nu h y''_m + \frac{\nu^2 h^2}{2} y'''_m + \ldots + \frac{\nu^{p-1} h^{p-1}}{(p-1)!} y^{(p)}_m + \\ + \frac{\nu^p h^p}{p!} y^{(p+1)}_m + o\left(h^p\right). \qquad (39)$$

Потребуем, чтобы после подстановки (38) и (39) в (37) коэффициенты при h^0, h^1, h^2, ..., h^p в правой и левой частях полученного равенства были бы равны для произвольной функции $y(x)$. Это даст нам следующие равенства:

$$\sum_{\nu=0}^{k} \alpha_\nu = 0, \qquad (40)$$

$$\sum_{\nu=1}^{k} \left[\alpha_\nu \nu^s - s\beta_\nu \nu^{s-1}\right] = 0 \quad (s = 2, \ldots, p); \qquad \sum_{\nu=1}^{k} \alpha_\nu \nu - \sum_{\nu=0}^{k} \beta_\nu = 0. \qquad (41)$$

Всего имеем $p+1$ однородных линейных алгебраических уравнений относительно $2k+2$ неизвестных α_i и β_i. Таким образом, максимальное значение p равно $2k$. Ошибка метода на одном шаге или локальная ошибка метода будет определяться разностью между левой и правой частями. Член с наименьшей степенью h в этой

разности будет равен

$$\frac{h^{p+1}}{(p+1)!}\sum_{\nu=1}^{k}[\alpha_\nu \nu^{p+1}-(p+1)\beta_\nu \nu^p]\,y_m^{(p+1)}=C_{p+1}h^{p+1}y_m^{(p+1)}. \qquad (42)$$

Можно было бы ожидать, что чем больше будет p, тем точнее будет формула (37). Однако это не всегда так в связи с возникающими при вычислениях по этой формуле погрешностями округления.

Выведенные нами ранее формулы будут являться частными случаями формулы (37). Приведем таблицу значений p и $C_{p+1}h^{p+1}$ для этих формул:

№	Формулы	k	p	$C_{p+1}h^{p+1}$
1	$y_{m+1}-y_m=hy'_m$	1	1	$\frac{1}{2}h^2$
2	$y_{m+2}-y_{m+1}=\frac{1}{2}h\left[3y'_{m+1}-y'_m\right]$	2	2	$\frac{5}{12}h^3$
3	$y_{m+3}-y_{m+2}=\frac{1}{12}h\left[23y'_{m+2}-16y'_{m+1}+5y'_m\right]$	3	3	$\frac{3}{8}h^4$
4	$y_{m+4}-y_{m+3}=\frac{1}{24}h\left[55y'_{m+3}-59y'_{m+2}+37y'_{m+1}-9y'_m\right]$	4	4	$\frac{251}{720}h^5$
5	$y_{m+2}-y_m=2hy'_{m+1}$	2	2	$\frac{1}{3}h^3$
6	$y_{m+3}-y_{m+1}=\frac{1}{3}h\left[7y'_{m+2}-2y'_{m+1}+y'_m\right]$	3	3	$\frac{1}{3}h^4$
7	$y_{m+4}-y_{m+2}=\frac{1}{3}h\left[8y'_{m+3}-5y'_{m+2}+4y'_{m+1}-y'_m\right]$	4	4	$-\frac{29}{30}h^5$
8	$y_{m+1}-y_m=hy'_{m+1}$	1	1	$-\frac{1}{2}h^2$
9	$y_{m+1}-y_m=\frac{1}{2}h\left[y'_{m+1}+y'_m\right]$	1	2	$-\frac{1}{12}h^3$
10	$y_{m+2}-y_{m+1}=\frac{h}{12}\left[5y'_{m+2}+8y'_{m+1}-y'_m\right]$	2	3	$-\frac{1}{24}h^4$
11	$y_{m+3}-y_{m+2}=\frac{h}{24}\left[9y'_{m+3}+19y'_{m+2}-5y'_{m+1}+y'_m\right]$	3	4	$-\frac{19}{720}h^5$
12	$y_{m+4}-y_{m+3}=\frac{1}{720}h\left[251y'_{m+4}+646y'_{m+3}-264y'_{m+2}+106y'_{m+1}-19y'_m\right]$	4	5	$-\frac{3}{180}h^6$
13	$y_{m+2}-y_m=\frac{1}{3}h\left[y'_{m+2}+4y'_{m+1}+y'_m\right]$	2	4	$-\frac{1}{90}h^5$

Уже этот обзор формул дает основания ожидать, что интерполяционные формулы дают лучшую точность, чем экстраполяционные.

Найдем теперь несколько формул типа (37) методом неопределенных коэффициентов. Условимся называть формулу (37) *разностным уравнением* или *уравнением в конечных разностях*. Число k назовем *порядком* этого уравнения, а число p — *степенью*. Попробуем найти формулу второго порядка экстраполяционного типа, имеющую степень 3. Тогда $\beta_2 = 0$. Коэффициент α_2 без уменьшения общности можно считать равным единице. При этом уравнения (40) и (41) примут вид

$$\left.\begin{aligned} &\alpha_0 + \alpha_1 + 1 = 0, \\ &\alpha_1 + 2 - (\beta_0 + \beta_1) = 0; \end{aligned}\right\} \tag{43}$$

$$\left.\begin{aligned} &\alpha_1 + 4 - 2\beta_1 = 0, \\ &\alpha_1 + 8 - 3\beta_1 = 0. \end{aligned}\right\} \tag{44}$$

Уравнения системы (44) дадут $\beta_1 = 4$, $\alpha_1 = 4$. Далее, получим $\beta_0 = 2$, $\alpha_0 = -5$. Таким образом, искомая формула примет вид

$$y_{m+2} + 4y_{m+1} - 5y_m = h\left[4y'_{m+1} + 2y'_m\right]. \tag{45}$$

Найдем теперь при $k = 2$ формулу интерполяционного типа, имеющую $p = 4$. Уравнения (40) и (41) примут вид (опять полагаем $\alpha_2 = 1$)

$$\left.\begin{aligned} &\alpha_0 + \alpha_1 + 1 = 0, \\ &\alpha_1 + 2 - (\beta_0 + \beta_1 + \beta_2) = 0, \\ &\alpha_1 + 4 - 2(\beta_1 + 2\beta_2) = 0, \\ &\alpha_1 + 8 - 3(\beta_1 + 4\beta_2) = 0, \\ &\alpha_1 + 16 - 4(\beta_1 + 8\beta_2) = 0. \end{aligned}\right\} \tag{46}$$

В этом случае последние три уравнения дадут $\beta_2 = \frac{1}{3}$, $\beta_1 = \frac{4}{3}$ и $\alpha_1 = 0$. Из остальных уравнений получаем $\beta_0 = \frac{1}{3}$, $\alpha_0 = -1$. Мы снова пришли к формуле (35).

Можно рассматривать и более сложные разностные уравнения:

$$\sum_{\nu=0}^{k} \alpha_\nu y_{m+\nu} + h \sum_{\nu=0}^{k} \beta_\nu y'_{m+\nu} + h^2 \sum_{\nu=0}^{k} \gamma_\nu y''_{m+\nu} = 0. \tag{47}$$

К формулам такого типа мы придем, например, если в качестве $L_{m,k}(x)$ будем брать интерполяционный многочлен Эрмита. При-

менять формулы типа (47) удобно лишь в тех случаях, когда производные от $f(x, y)$ легко находятся и легко вычисляются. Можно еще усложнить (47), включив туда производные более высоких порядков.

4. Метод Крылова отыскания начальных значений решения. Коснемся теперь немного способов вычисления начальных значений $y_1, y_2, \ldots, y_k$, основанных на разностных формулах. Приведем один такой способ, предложенный академиком А. Н. Крыловым.

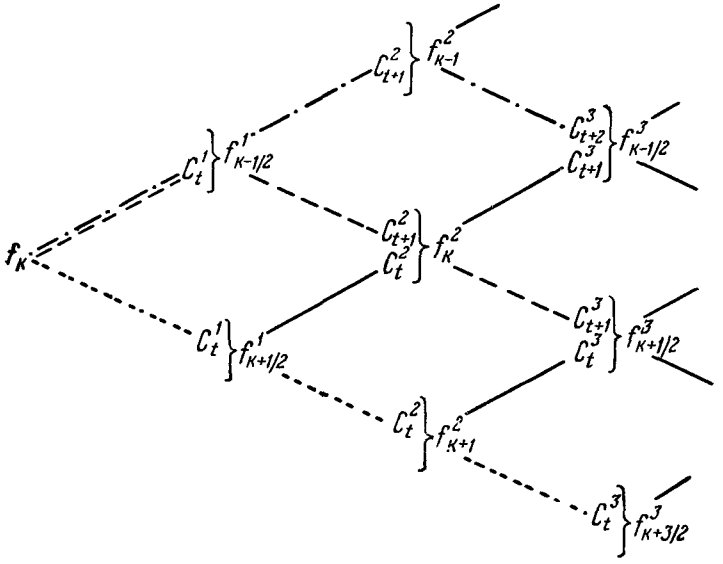

Рис. 25.

Ограничимся случаем $k = 3$. Возьмем кусок диаграммы Фрезера и запишем три интерполяционные формулы, пути для которых указаны пунктиром, тире и тире с пунктиром. Эти формулы имеют вид

$$\left.\begin{aligned} f_t &= f_k + t f^1_{k+\frac{1}{2}} + \frac{t(t-1)}{2} f^2_{k+1} + \frac{t(t-1)(t-2)}{6} f^3_{k+\frac{3}{2}}, \\ f_t &= f_k + t f^1_{k-\frac{1}{2}} + \frac{t(t+1)}{2} f^2_k + \frac{t(t^2-1)}{6} f^3_{k+\frac{1}{2}}, \\ f_t &= f_k + t f^1_{k-\frac{1}{2}} + \frac{t(t+1)}{2} f^2_{k-1} + \frac{t(t+2)(t+1)}{6} f^3_{k-\frac{1}{2}}. \end{aligned}\right\} \quad (48)$$

Интегрируя их в пределах от 0 до 1, получим:

$$\left.\begin{aligned}
\Delta y_k &= q_k + \frac{1}{2} q^1_{k+\frac{1}{2}} - \frac{1}{12} q^2_{k+1} + \frac{1}{24} q^3_{k+\frac{3}{2}},\\
\Delta y_k &= q_k + \frac{1}{2} q^1_{k-\frac{1}{2}} + \frac{5}{12} q^2_k - \frac{1}{24} q^3_{k+\frac{1}{2}},\\
\Delta y_k &= q_k + \frac{1}{2} q^1_{k-\frac{1}{2}} + \frac{5}{12} q^2_{k-1} + \frac{3}{8} q^3_{k-\frac{1}{2}}.
\end{aligned}\right\} \tag{49}$$

Положив в первом из равенств (49) $k=0$, во втором $k=1$ и в третьем $k=2$, будем иметь:

$$\left.\begin{aligned}
\Delta y_0 &= q_0 + \frac{1}{2} q^1_{\frac{1}{2}} - \frac{1}{12} q^2_1 + \frac{1}{24} q^3_{\frac{3}{2}},\\
\Delta y_1 &= q_1 + \frac{1}{2} q^1_{\frac{1}{2}} + \frac{5}{12} q^2_1 - \frac{1}{24} q^3_{\frac{3}{2}},\\
\Delta y_2 &= q_2 + \frac{1}{2} q^1_{\frac{3}{2}} + \frac{5}{12} q^2_1 + \frac{3}{8} q^3_{\frac{3}{2}}.
\end{aligned}\right\} \tag{50}$$

Выражения в правых и левых частях зависят от y_0, y_1, y_2, y_3, и мы имеем систему уравнений для определения y_1, y_2, y_3. Ход вычислений по этим формулам лучше всего показать на примере. Рассмотрим уравнение $y'=y$ и будем отыскивать его решение, удовлетворяющее начальному условию $y(0)=1$. При этом если шаг равен 0,1, то $q_0=0{,}1$. Принимаем приближенно $\Delta y_0 = q_0$. Отсюда находим первое приближение для y_1, которое мы будем обозначать через y_{11}. В нашем случае $y_{11}=1{,}01$. Теперь мы имеем возможность найти первое приближение для $q^1_{\frac{1}{2}} = q_1 - q_0$. В нашем случае $q_1=0{,}101$ и $q^1_{\frac{1}{2}}=0{,}001$. Вычисляем второе приближение для $y_1 - y_{12}$ и находим

$$y_{12} = y_0 + q_0 + \frac{1}{2} q^1_{\frac{1}{2}} = 1{,}1005. \tag{51}$$

Поэтому исправленное значение для q_1 будет $q_1=0{,}11005$ и для $q^1_{\frac{1}{2}}$ будет $q^1_{\frac{1}{2}}=0{,}0100$. Найдем теперь первое приближение для $y_2 - y_{21}$ по формуле

$$y_{21} = y_{12} + q_1 + \frac{1}{2} q^1_{\frac{1}{2}} = 1{,}1005 + 0{,}1100 + 0{,}0050 = 1{,}2155. \tag{52}$$

Теперь мы можем найти q_2, $q^1_{\frac{3}{2}}$ и q^2_1:

$$q_2 = 0{,}1216; \quad q^1_{\frac{3}{2}} = 0{,}1216 - 0{,}1100 = 0{,}0116; \quad q^2_1 = 0{,}0016. \tag{53}$$

Тогда

$$y_{31} = y_{21} + q_2 + \frac{1}{2} q^1_{\frac{1}{2}} + \frac{5}{12} q^2_1 = 1{,}2155 + 0{,}1216 + \\ + 0{,}0050 + 0{,}0007 = 1{,}3425 \tag{54}$$

и

$$\left.\begin{aligned} q_3 = 0{,}1342; \quad q^1_{\frac{3}{2}} = 0{,}1342 - 0{,}1216 = 0{,}0126; \\ q^2_2 = 0{,}0010; \quad q^3_{\frac{3}{2}} = -0{,}0006. \end{aligned}\right\} \tag{55}$$

Пересчитывая по формулам (50) значения y_1, y_2, y_3, находим:

$$\left.\begin{aligned} y_{13} = y_0 + q_0 + \frac{1}{2} q^1_{\frac{1}{2}} - \frac{1}{12} q^2_1 + \frac{1}{24} q^3_{\frac{3}{2}} = 1{,}1049, \\ y_{22} = y_1 + q_1 + \frac{1}{2} q^1_{\frac{1}{2}} + \frac{5}{12} q^2_1 - \frac{1}{24} q^3_{\frac{3}{2}} = 1{,}2685, \\ y_{32} = y_2 + q_2 + \frac{1}{2} q^1_{\frac{1}{2}} + \frac{5}{12} q^2_1 + \frac{3}{8} q^3_{\frac{3}{2}} = 1{,}4482. \end{aligned}\right\} \tag{56}$$

Вычисляем заново значения разностей

$$\left.\begin{aligned} q_1 = 0{,}11049; \quad q_2 = 0{,}1268; \quad q_3 = 0{,}1448, \\ q^1_{\frac{1}{2}} = q_1 - q_0 = 0{,}0105, \\ q^2_1 = q_2 - 2q_1 + q_0 = 0{,}0059, \\ q^3_{\frac{3}{2}} = q_3 - 3q_2 + 3q_1 - q_0 = -0{,}0042. \end{aligned}\right\} \tag{57}$$

После нового пересчета y_1, y_2, y_3 находим:

$$y_{14} = 1{,}1049, \quad y_{23} = 1{,}2233, \quad y_{33} = 1{,}3562. \tag{58}$$

Так как расхождения с предыдущими приближениями еще очень велики, то проделываем вычисления еще раз. Получим:

$$\left.\begin{aligned} q_1 = 0{,}11049, \quad q_2 = 0{,}12233, \quad q_3 = 0{,}13562, \\ q^1_{\frac{1}{2}} = 0{,}0105, \quad q^2_1 = 0{,}0013, \quad q^3_{\frac{3}{2}} = 0{,}0001, \\ y_{15} = 1{,}1051, \quad y_{24} = 1{,}2213, \quad y_{34} = 1{,}3493. \end{aligned}\right\} \tag{59}$$

Новый пересчет даст

$$\left.\begin{array}{lll} q_1 = 0{,}11051, & q_2 = 0{,}12213, & q_3 = 0{,}13493, \\ q^1_{\frac{1}{2}} = 0{,}0105, & q^2_1 = 0{,}0011, & q^3_{\frac{3}{2}} = 0{,}0001, \\ y_{16} = 1{,}1051, & y_{25} = 1{,}2213, & y_{35} = 1{,}3499. \end{array}\right\} \quad (60)$$

В дальнейшем изменений происходить не будет.

Как мы видели ранее, для сходимости процесса последовательных приближений достаточно, чтобы собственные значения некоторой матрицы, составленной из частных производных правых частей, были по модулю меньше 1. В нашем случае решается методом последовательных приближений система

$$\left.\begin{array}{l} y_1 = y_0 + \frac{h}{24}[9f(x_0, y_0) + 19f(x_1, y_1) - 5f(x_2, y_2) + f(x_3, y_3)], \\ y_2 = y_1 + \frac{h}{24}[-f(x_0, y_0) + 13f(x_1, y_1) + 13f(x_2, y_2) - f(x_3, y_3)], \\ y_3 = y_2 + \frac{h}{24}[-11f(x_0, y_0) + 19f(x_1, y_1) + 7f(x_2, y_2) + 9f(x_3, y_3)]. \end{array}\right\} \quad (61)$$

Соответствующее вековое уравнение примет вид

$$\begin{vmatrix} \frac{19M_1h}{24} - \lambda & \frac{5M_1h}{24} & \frac{M_1h}{24} \\ 1 + \frac{13M_1h}{24} & \frac{13M_1h}{24} - \lambda & \frac{M_1h}{24} \\ \frac{19M_1h}{24} & 1 + \frac{7M_1h}{24} & \frac{9M_1h}{24} - \lambda \end{vmatrix} = 0. \quad (62)$$

Раскрывая определитель, получим:

$$\lambda^3 - \frac{41}{24} M_1h\lambda^2 + \left(\frac{37}{38} M_1^2h^2 - \frac{1}{4} M_1h\right)\lambda - \left(\frac{361}{3465} M_1^3h^3 - \frac{11}{144} M_1^2h^2 + \frac{1}{24} M_1h\right) = 0. \quad (63)$$

При достаточно малом h все корни этого уравнения по модулю меньше единицы.

5. Примеры. Проиллюстрируем теперь ход вычислений по разностным формулам при решении дифференциального уравнения первого порядка $y' = y$ с начальным условием $y(0) = 1$ и шагом 0,1. Начальные приближения находились с помощью рядов с пятью верными десятичными знаками. Все разностные формулы брались не более чем до третьих разностей. Для интерполяционных способов при отыскании Δy методом последовательных приближений за начальные приближения брались результаты, полученные по экстраполяционным способам. Промежуточные результаты вычислялись с шестью десятичными знаками. Окончательные результаты округлялись до пяти десятичных знаков.

$$\Delta y_m = q_m + \frac{1}{2} q^1_{m-1/2} + \frac{5}{12} q^2_{m-1} + \frac{3}{8} q^3_{m-3/2}$$

x	y	Δy	q	q^1	q^2	q^3
0,0	1,00000		0,100000			
				10 517		
0,1	1,10517		0,110517		1106	
				11 623		117
0,2	1,22140		0,122140		1223	
				12 846		127
0,3	1,34986		0,134986		1350	
		0,14196		14 196		143
0,4	1,49182		0,149182		1493	
		0,15689		15 689		157
0,5	1,64871		0,164871		1650	
		0,17339		17 339		174
0,6	1,82210		0,182210		1824	
		0,19163		19 163		191
0,7	2,01373		0,201373		2015	
		0,21178		21 178		212
0,8	2,22551		0,222551		2227	
		0,23405		23 405		
0,9	2,45956		0,245956			
		0,25866				
1,0	2,71822					

$$\Delta y_m = q_{m+1} - \frac{1}{2} q^1_{m+1/2} - \frac{1}{12} q^2_m - \frac{1}{24} q^3_{m-1/2}$$

x	y	Δy	q	q^1	q^2	q^3
0,0	1,00000		0,100000			
				10 517		
0,1	1,10517		0,110517		1106	
				11 623		117
0,2	1,22140		0,122140		1223	
				12 846		128
0,3	1,34986		0,134986		1351	
		0,14197		14 197		142
0,4	1,49183		0,149183		1493	
		0,15690		15 690		157
0,5	1,64873		0,164873		1650	
		0,17340		17 340		173
0,6	1,82213		0,182213		1823	
		0,19163		19 163		193
0,7	2,01376		0,201376		2016	
		0,21179		21 179		211
0,8	2,22555		0,222555		2227	
		0,23406		23 406		235
0,9	2,45961		0,245961		2462	
		0,25868		25 868		
1,0	2,71829		0,271829			

$$y_{m+1} - y_{m-1} = 2q_m + \frac{1}{3}q^2_{m-1} + \frac{1}{3}q^3_{m-3/2}$$

x	y	q	q^1	q^2	q^3
0,0	1,00000	0,100000			
			10 517		
0,1	1,10517	0,110517		1106	
			11 623		117
0,2	1,22140	0,122140		1223	
			12 846		127
0,3	1,34986	0,134986		1350	
			14 196		144
0,4	1,49182	0,149182		1494	
			15 690		155
0,5	1,64872	0,164872		1649	
			17 339		175
0,6	1,82211	0,182211		1824	
			19 163		191
0,7	2,01374	0,201374		2015	
			21 178		213
0,8	2,22552	0,222552		2228	
			23 406		
0,9	2,45958	0,245958			
1,0	2,71825				

$$y_{m+1} - y_{m-1} = 2q_m + \frac{1}{3}q^2_m$$

x	y	q	q^1	q^2
0,0	1,00000	0,100000		
			10 517	
0,1	1,10517	0,110517		1106
			11 623	
0,2	1,22140	0,122140		1223
			12 846	
0,3	1,34986	0,134986		1350
			14 196	
0,4	1,49182	0,149182		1494
			15 690	
0,5	1,64872	0,164872		1649
			17 339	
0,6	1,82211	0,182211		1825
			19 164	
0,7	2,01375	0,201375		2014
			21 178	
0,8	2,22553	0,222553		2229
			23 407	
0,9	2,45960	0,245960		2460
			25 867	
1,0	2,71827	0,271827		

Сравнение полученных результатов со значениями e^x дает следующую таблицу ошибок (в единицах пятого десятичного знака):

x	1-й способ	2-й способ	3-й способ	4-й способ
0,0	0	0	0	0
0,1	0	0	0	0
0,2	0	0	0	0
0,3	0	0	0	0
0,4	0	+1	+1	+1
0,5	−1	+1	0	0
0,6	−2	+1	−1	−1
0,7	−2	+1	−1	0
0,8	−3	+1	−2	−1
0,9	−4	+1	−2	0
1,0	−6	+1	−3	−1

Как видно из этой таблицы, интерполяционные способы имеют заметное преимущество в точности по сравнению с экстраполяционными. Формулы, при помощи которых мы решали уравнение в 3-м и 4-м случае, связывают значения y_{k+1} не с y_k, а с y_{k-1}, и поэтому может случиться, что значения y_k с четными индексами и нечетными индексами будут слабо связаны друг с другом. Так это и произошло в четвертом случае. Тогда прибегают к сглаживанию или самих значений функции или их разностей.

Полученные нами способы без каких-либо дополнительных рассуждений переносятся на системы обыкновенных дифференциальных уравнений первого порядка. При этом, конечно, придется параллельно решать все уравнения системы.

§ 6. Разностные методы решения обыкновенных дифференциальных уравнений высших порядков

Уравнения высших порядков могут быть сведены к системе уравнений первого порядка. Поэтому на них переносятся все те методы, о которых говорилось в предыдущем параграфе. Однако при этом получаются системы очень специфичного вида, для которых возможны упрощения общих методов. К этому вопросу мы сейчас и перейдем.

Пусть задано дифференциальное уравнение

$$y'' = f(x, y, y') \tag{1}$$

и требуется найти его решение, удовлетворяющее начальным условиям: $y(x_0) = y_0$, $y'(x_0) = y'_0$. Воспользуемся тем же способом,

который мы применяли для уравнений первого порядка. Предполагаем, что нам известны значения y и y' в точках x_m, x_{m-1}, ..., x_{m-k}. Находим тогда значения f_m, f_{m-1}, ..., f_{m-k} и строим интерполяционный многочлен Ньютона для интерполирования назад:

$$f_t = f_m + tf^1_{m-1/2} + \frac{t(t+1)}{2} f^2_{m-1} + \ldots$$
$$\ldots + \frac{t(t+1)\ldots(t+k-1)}{k!} f^k_{m-k/2}. \qquad (2)$$

Интегрируя его по t в пределах $[0, \xi]$, получим:

$$y'(x_0+\xi h) = y'_m + h\int_0^{\xi}\left[f_m + tf^1_{m-1/2} + \frac{t(t+1)}{2!} f^2_{m+1} + \ldots\right.$$
$$\left.\ldots + \frac{t(t+1)\ldots(t+k-1)}{k!} f^k_{m-k/2}\right]dt. \qquad (3)$$

При $\xi = 1$ будем иметь:

$$y'_{m+1} = y'_m + a_0 q_m + a_1 q^1_{m-1/2} + \ldots + a_k q^k_{m-k/2}, \qquad (4)$$

где a_i — коэффициенты экстраполяционной формулы Адамса. По (4) мы можем вычислить y'_{m+1}. Интегрируя (3) по ξ в пределах $[0, 1]$, получим:

$$y_{m+1} = y_m + hy'_m + h^2\int_0^1 d\xi\int_0^{\xi}\left[f_m + tf^1_{m-1/2} + \ldots\right.$$
$$\left.\ldots + \frac{t(t+1)\ldots(t+k-1)}{k!} f^k_{m-k/2}\right]dt =$$
$$= y_m + hy'_m + h^2\left[\alpha_0 f_m + \alpha_1 f^1_{m-1/2} + \ldots + \alpha_k f^k_{m-k/2}\right], \qquad (5)$$

где

$$\alpha_i = \int_0^1 d\xi\int_0^{\xi}\frac{t(t+1)\ldots(t+i-1)}{i!}dt. \qquad (6)$$

Вычислив коэффициенты α_i, мы сможем найти y_{m+1} и тем самым продвинуться на один шаг.

На практике приходится сталкиваться с уравнениями, в которых $f(x, y, y')$ не зависит от y'. Тогда нет необходимости на каждом шаге вычислять y'. Поэтому целесообразно совершенно исключить y'_m из (5). Для этого проинтегрируем (3) по ξ в пределах $[-1, 0]$.

Получим:

$$y_m = y_{m-1} + hy'_m + h^2 \int_{-1}^{0} d\xi \int_{0}^{\xi} \left[f_m + t f^1_{m-1/2} + \frac{t(t+1)}{2} f^2_{m-1} + \dots \right.$$

$$\left. \dots + \frac{t(t+1)\dots(t+k-1)}{k!} f^k_{m-k/2} \right] dt = y_{m-1} + hy'_m +$$

$$+ h^2 \left[\beta_0 f_m + \beta_1 f^1_{m-1/2} + \beta_2 f^2_{m-1} + \dots + \beta_k f^k_{m-k/2} \right], \qquad (7)$$

где

$$\beta_i = \int_{-1}^{0} d\xi \int_{0}^{\xi} \frac{t(t+1)\dots(t+i-1)}{i!}\, dt. \qquad (8)$$

Вычитая (8) из (5), найдем:

$$\Delta^2 y_{m-1} = y_{m+1} - 2y_m + y_{m-1} =$$

$$= h^2 \left[(\alpha_0 - \beta_0) f_m + (\alpha_1 - \beta_1) f^1_{m-1/2} + \dots + (\alpha_k - \beta_k) f^k_{m-k/2} \right]. \qquad (9)$$

При этом

$$\alpha_i - \beta_i = \int_{0}^{1} d\xi \int_{0}^{\xi} \frac{t(t+1)\dots(t+i-1)}{i!}\, dt -$$

$$- \int_{-1}^{0} d\xi \int_{0}^{\xi} \frac{t(t+1)\dots(t+i-1)}{i!}\, dt =$$

$$= \frac{1}{i!} \int_{0}^{1} d\xi \int_{-\xi}^{\xi} t(t+1)\dots(t+i-1)\, dt. \qquad (10)$$

В частности,

$$\left.\begin{aligned}
&\alpha_0 - \beta_0 = \int_0^1 d\xi \int_{-\xi}^{\xi} 1\, dt = 1, \quad \alpha_1 - \beta_1 = \int_0^1 d\xi \int_{-\xi}^{\xi} t\, dt = 0,\\
&\alpha_2 - \beta_2 = \int_0^1 d\xi \int_{-\xi}^{+\xi} \frac{t(t+1)}{2}\, dt = \frac{1}{12},\\
&\alpha_3 - \beta_3 = \int_0^1 d\xi \int_{-\xi}^{+\xi} \frac{t(t+1)(t+2)}{6}\, dt = \frac{1}{12},\\
&\alpha_4 - \beta_4 = \frac{19}{240}; \quad \alpha_5 - \beta_5 = \frac{9}{120}; \quad \alpha_6 - \beta_6 = \frac{863}{12 \cdot 1008}.
\end{aligned}\right\} \qquad (11)$$

Таким образом, мы получим:

$$\Delta^2 y_{m-1} = h^2 f_m + \frac{1}{12} h^2 \left[f^2_{m-1} + f^3_{m-3/2} + \frac{19}{20} f^4_{m-2} + \right.$$
$$\left. + \frac{9}{10} f^5_{m-5/2} + \frac{863}{1008} f^6_{m-3} + \ldots \right]. \quad (12)$$

Эта *формула* носит имя *Штёрмера.* Процесс вычислений по формуле Штёрмера идет так же, как и для уравнений первого порядка, только вместо $q = hf(x, y)$ берут $r = h^2 f(x, y)$.

В качестве примера рассмотрим уравнение

$$y'' = y \quad (13)$$

и будем отыскивать его решение, удовлетворяющее начальным данным: $y(0) = 1$, $y'(0) = 1$. Будем использовать формулу (12) до третьих разностей. Начальные четыре значения возьмем из таблицы, полученной по формуле Рунге — Кутта.

Получим таблицу:

x	y	Δy	$\Delta^2 y$	$r = h^2 f$	r^1	r^2	r^3
0	1			0,010000			
					1052		
0,1	1,1052			0,011052		110	
					1162		13
0,2	1,2214			0,012214		123	
		1285			1285		12
0,3	1,3499		135	0,013499		135	
		1420			1420		14
0,4	1,4919		149	0,014919		149	
		1569			1569		16
0,5	1,6488		165	0,016488		165	
		1734			1734		17
0,6	1,8222		182	0,018222		182	
		1916			1916		20
0,7	2,0138		202	0,020138		202	
		2118			2118		21
0,8	2,2256		223	0,022256		223	
		2341			2341		
0,9	2,4597		246	0,024597			
		2587					
1,0	2,7184						

Как мы видим, вычисления оказались очень несложными — производятся в уме. Записей очень мало. Результаты довольно точные.

Рассмотрим еще пример интерполяционной разностной формулы для решения дифференциальных уравнений второго порядка. На

этот раз применим интерполяционную формулу Стирлинга:

$$f(x_m+th)=f_m+tf_m^1+\frac{t^2}{2}f_m^2+$$

$$+\frac{t(t^2-1)}{6}f_m^3+\frac{t^2(t^2-1)}{24}f_m^4+\ldots \quad (14)$$

Как и ранее, получим:

$$y'(x_m+\xi h)=y'_m+h\int_0^{\xi}\left[f_m+tf_m^1+\frac{t^2}{2}f_m^2+\right.$$

$$\left.+\frac{t(t^2-1)}{6}f_m^3+\frac{t^2(t^2-1)}{24}f_m^4+\ldots\right]dt. \quad (15)$$

В частности, при $\xi=1$ будем иметь:

$$y'_{m+1}=y'_m+h\int_0^1\left[f_m+tf_m^1+\frac{t^2}{2}f_m^2+\right.$$

$$\left.+\frac{t(t^2-1)}{6}f_m^3+\frac{t^2(t^2-1)}{24}f_m^4+\ldots\right]dt=$$

$$=y'_m+h\left[\alpha_0 f_m+\alpha_1 f_m^1+\alpha_2 f_m^2+\alpha_3 f_m^3+\alpha_4 f_m^4+\ldots\right], \quad (16)$$

где

$$\left.\begin{aligned}\alpha_{2i+2}&=\int_0^1\frac{t^2(t^2-1)\ldots(t^2-i^2)}{(2i+2)!}dt,\\ \alpha_{2i+1}&=\int_0^1\frac{t(t^2-1)\ldots(t^2-i^2)}{(2i+1)!}dt.\end{aligned}\right\} \quad (17)$$

Интегрируем еще раз (15) по ξ в пределах [0, 1]. Будем иметь:

$$y_{m+1}=y_m+hy'_m+h^2\int_0^1 d\xi\int_0^{\xi}\left[f_m+tf_m^1+\frac{t^2}{2}f_m^2+\right.$$

$$\left.+\frac{t(t^2-1)}{6}f_m^3+\frac{t^2(t^2-1)}{24}f_m^4+\ldots\right]dt=$$

$$=y_m+hy'_m+h^2\left[\beta_0 f_m+\beta_1 f_m^1+\beta_2 f_m^2+\beta_3 f_m^3+\beta_4 f_m^4+\ldots\right], \quad (18)$$

где

$$\left.\begin{aligned}\beta_{2i+1}&=\int_0^1 d\xi\int_0^{\xi}\frac{t(t^2-1)\dots(t^2-i^2)}{(2i+1)!}\,dt,\\ \beta_{2i+2}&=\int_0^1 d\xi\int_0^{\xi}\frac{t^2(t^2-1)\dots(t^2-i^2)}{(2i+2)!}\,dt.\end{aligned}\right\}\tag{19}$$

Интегрирование (15) по ξ в пределах $[-1, 0]$ дает

$$\begin{aligned}y_m=y_{m-1}+hy'_m+h^2\int_{-1}^{0}d\xi\int_0^{\xi}\Big[f_m+tf^1_m+\frac{t^2}{2}f^2_m+\\ +\frac{t(t^2-1)}{6}f^3_m+\frac{t^2(t^2-1)}{24}f^4_m+\dots\Big]dt=\\ =y_{m-1}+hy'_m+h^2[\gamma_0f_m+\gamma_1f^1_m+\gamma_2f^2_m+\\ +\gamma_3f^3_m+\gamma_4f^4_m+\dots],\end{aligned}\tag{20}$$

где

$$\left.\begin{aligned}\gamma_{2i+1}&=\int_{-1}^{0}d\xi\int_0^{\xi}\frac{t(t^2-1)\dots(t^2-i^2)}{(2i+1)!}\,dt,\\ \gamma_{2i+2}&=\int_{-1}^{0}d\xi\int_0^{\xi}\frac{t^2(t^2-1)\dots(t^2-i^2)}{(2i+2)!}\,dt.\end{aligned}\right\}\tag{21}$$

Вычитая (20) из (18), получим:

$$\begin{aligned}y_{m+1}-2y_m+y_{m-1}=h^2[\delta_0f_m+\delta_1f^1_m+\\ +\delta_2f^2_m+\delta_3f^3_m+\delta_4f^4_m+\dots],\end{aligned}\tag{22}$$

где

$$\delta_{2i+1}=\beta_{2i+1}-\gamma_{2i+1}=\int_0^1 d\xi\int_{-\xi}^{+\xi}\frac{t(t^2-1)\dots(t^2-i^2)}{(2i+1)!}\,dt=0,\tag{23}$$

так как под знаком внутреннего интеграла стоит нечетная функция t, а

$$\delta_{2i+2}=\beta_{2i+2}-\gamma_{2i+2}=\int_0^1 d\xi\int_{-\xi}^{\xi}\frac{t^2(t^2-1)\dots(t^2-i^2)}{(2i+2)!}\,dt.\tag{24}$$

В частности,

$$\left.\begin{aligned}
\delta_0 &= \int_0^1 d\xi \int_{-\xi}^{\xi} dt = 1,\\
\delta_2 &= \int_0^1 d\xi \int_{-\xi}^{\xi} \frac{t^2}{2}\, dt = \frac{1}{12},\\
\delta_4 &= \int_0^1 d\xi \int_{-\xi}^{+\xi} \frac{t^2(t^2-1)}{24}\, dt = -\frac{1}{240},\\
\delta_6 &= \frac{31}{60\,480}; \quad \delta_8 = -\frac{289}{3\,628\,800}.
\end{aligned}\right\} \tag{25}$$

Подберем теперь такую функцию φ_m, для которой бы столбец f_k являлся столбцом вторых разностей. Это можно сделать бесчисленным множеством способов. Можно, например, произвольным образом задать $\varphi^1_{-1/2}$ и последовательным сложением с f_0, f_1, f_2, ... заполнить столбец первых разностей искомой функции, а затем произвольным образом задать φ_{-1} и последовательным сложением с $\varphi^1_{-1/2}$, $\varphi^1_{1/2}$, $\varphi^1_{3/2}$, ... получить таблицу значений φ_i. Будем обозначать $\varphi^1_{i+1/2}$ через $f^{-1}_{i+1/2}$, а φ_i — через f_i^{-2}.

Левая часть (22) является второй разностью для значений y_i, взятой в точке m. Правая часть (22) является второй разностью от

$$h^2\left[\delta_0 f_i^{-2} + \delta_2 f_i + \delta_4 f_i^2 + \delta_6 f_i^4 + \delta_8 f_i^6 + \ldots\right], \tag{26}$$

взятой также в точке m. Поэтому первые разности этих двух табличных функций могут отличаться только на постоянную. Таким образом,

$$y_{m+1} - y_m = h^2\left[\delta_0 f^{-1}_{m+1/2} + \delta_2 f^1_{m+1/2} + \delta_4 f^3_{m+1/2} + \right.$$
$$\left. + \delta_6 f^5_{m+1/2} + \delta_8 f^7_{m+1/2} + \ldots\right] + C. \tag{27}$$

Эта постоянная C будет равна нулю, если при $m = 0$ она нуль, т. е. если

$$y_1 - y_0 = h^2\left[\delta_0 f^{-1}_{1/2} + \delta_2 f^1_{1/2} + \delta_4 f^3_{1/2} + \delta_6 f^5_{1/2} + \delta_8 f^7_{1/2} + \ldots\right]. \tag{28}$$

Последнее равенство будет выполнено, если подходящим образом выбрать $f^{-1}_{-1/2}$. Будем предполагать, что это сделано и C в (27) равно нулю. Тогда, еще раз применяя такие же рассуждения, получим:

$$y_m = h^2\left[\delta_0 f_m^{-2} + \delta_2 f_m + \delta_4 f_m^2 + \delta_6 f_m^4 + \delta_8 f_m^6 + \ldots\right] + C_1. \tag{29}$$

Опять если выбрать f_{-1}^2 так, что будет выполнено равенство

$$y_0 = h^2\left[\delta_0 f_0^{-2} + \delta_2 f_0 + \delta_4 f_0^2 + \delta_6 f_0^4 + \delta_8 f_0^6 + \ldots\right], \tag{30}$$

то (29) перейдет в

$$y_m = h^2\left[\delta_0 f_m^{-2} + \delta_2 f_m + \delta_4 f_m^2 + \delta_6 f_m^4 + \delta_8 f_m^6 + \ldots\right] \tag{31}$$

или, если использовать (25),

$$y_m = h^2\left[f_m^{-2} + \frac{1}{12} f_m - \frac{1}{240} f_m^2 + \frac{31}{60\,480} f_m^4 - \right.$$
$$\left. - \frac{289}{3\,628\,800} f_m^6 + \ldots\right]. \tag{32}$$

При этом $f_{-1/2}^{-1}$ и f_{-1}^{-2} нужно выбрать так, чтобы были выполнены равенства:

$$\left.\begin{aligned} y_1 - y_0 &= h^2\left[f_{1/2}^{-1} + \frac{1}{12} f_{1/2}^1 - \frac{1}{240} f_{1/2}^3 + \frac{31}{60\,480} f_{1/2}^5 - \right.\right. \\ &\qquad \left. - \frac{289}{3\,628\,800} f_{1/2}^7 + \ldots\right], \\ y_0 &= h^2\left[f_0^{-2} + \frac{1}{12} f_0 - \frac{1}{240} f_0^2 + \frac{31}{60\,480} f_0^4 - \right. \\ &\qquad \left.\left. - \frac{289}{3\,628\,800} f_0^6 + \ldots\right]. \right| \end{aligned}\right\} \tag{33}$$

Равенство (32) называют *формулой суммирования Гаусса.*

В качестве примера на применение формулы суммирования Гаусса рассмотрим задачу о колебании математического маятника. Дифференциальное уравнение этих колебаний

$$\frac{d^2\varphi}{dt^2} = -\frac{g}{l}\sin\varphi \tag{34}$$

путем замены независимой переменной

$$x = t\sqrt{\frac{g}{l}} \tag{35}$$

приводим к виду

$$\frac{d^2\varphi}{dx^2} = -\sin\varphi. \tag{36}$$

Пусть начальные условия будут: $\varphi(0) = 60° = \frac{\pi}{3} = 1{,}04720$, $\varphi'(0) = 0$. Шаг h возьмем равным 0,2. Вычисления будем производить по формуле

$$\varphi_m = r_m^{-2} + \frac{1}{12} r_m - \frac{1}{240} r_m^2 = r_m^{-2} + \tau_m, \tag{37}$$

где $r_m = h^2 f_m = -0{,}04 \sin\varphi_m$.

Предварительно вычислим значения φ для $x = \pm 0{,}2$ и $x = \pm 0{,}4$, пользуясь формулами:

$$\left.\begin{aligned} &\varphi_{n+1} = \varphi_n + 0{,}2\varphi'_n + 0{,}02\varphi''_n, \quad \varphi_{-(n+1)} = \varphi_{-n} - 0{,}2\varphi'_{-n} + 0{,}02\varphi''_{-n}, \\ &\varphi'_{n+1} = \varphi'_n + 0{,}2\varphi''_n, \quad \varphi'_{-(n+1)} = \varphi'_{-n} - 0{,}2\varphi''_{-n}, \\ &\varphi''_n = -\sin\varphi_n. \end{aligned}\right\} \tag{38}$$

Вычисления дадут

$$\left.\begin{aligned} \varphi_2 &= \varphi_{-2} = 0{,}9782, \\ \varphi_1 &= \varphi_{-1} = 1{,}0299. \end{aligned}\right\} \tag{39}$$

Теперь уточним эти значения по формуле (37). Для этого вычисляем $r_1 = r_{-1}$ и $r_2 = r_{-2}$ и разности r_j^i, которые можно найти для пяти известных значений r_m. Далее, воспользовавшись тем, что $\frac{r_n^2}{240}$ в этих строках не достигает 0,5 единицы последнего разряда, находим $\tau_{-2} = \tau_2$ и $\tau_{-1} = \tau_1$ по формуле $\tau_n = \frac{1}{12} r_n$. Находим $r_{1/2}^{-1}$:

$$r_{1/2}^{-1} = (\varphi_1 - \varphi_0) - \frac{1}{12} r_{1/2}^1 = -0{,}01732 \tag{40}$$

и r_0^{-2}:

$$r_0^{-2} = \varphi_0 - \tau_0 = 1{,}05009. \tag{41}$$

Заполняем таблицу $r_{i+1/2}^{-1}$ и r_i^{-2} и исправляем φ_i и φ_2 по формуле (37). Получим:

$$\varphi_1 = r_1^{-2} + \tau_1 = 1{,}02991,$$

$$\varphi_2 = r_2^{-2} + \tau_2 = 0{,}97840.$$

Исправленные значения вставляем в таблицу. При этом значение r_1 и r_2 не изменяются.

Теперь можно продвигаться дальше. Проще всего это можно осуществить путем экстраполяции значений τ_n. Для этого составляем таблицу разностей τ_n, экстраполируем вторую разность на один шаг и находим $\tau_3 = -259$. Отсюда $\varphi_3 = r_3^{-2} + \tau_3 = 0{,}8938$. Это дает $r_3 = -0{,}03118$ и $\tau_3 = -260$. Исправление не влияет на полученные значения φ_3 и r_3. Поэтому после исправления разностей можно продвигаться на следующий шаг. Результаты вычислений приведены в данной ниже таблице.

x_n	φ_n	r_n^{-2}	r_n^{-1}	r_n	r_n^1	r_n^2	r_n^3	r_n^4	τ_n	τ_n^1	τ_n^2
									—276		
—0,4	0,9784	0,98116		—3 318						— 10	
			5 161		—111						
—0,2	1,0299	1,03277		—3 429		+76			—286		7
			1 732		—35		—6			—3	
0,0	1,0472	1,05009		—3 464		70		+12	—289		6
			—1 732		+35		+6			+3	
0,2	1,0299	1,03277		—3 429		76		7	—286		7
			—5 161		+111		13			10	
0,4	0,9784	0,98116		—3 318		89		8	—276		6
			—8 479		200		21			16	
0,6	0,8938	0,98637		—3 118		110		—4	—260		10
			—11 597		310		17			26	
0,8	0,7781	0,78040		—2 808		127		—12	—234		10
			—14 405		437		5			36	
1,0	0,6344	0,63635		—2 371		132		—16	—198		12
			—16 776		569		—11			48	
1,2	0,4671	0,46859		—1 802		121		—30	—150		9
			—18 578		690		—41			57	
1,4	0,2819	0,28281		—1 112		80		—12	—93		7
			—19 690		770		—53			64	
1,6	0,0856	0,0591		—342		27		—9	—29		3
			—20 032		797		—62			67	
1,8	—0,1140	—0,11441		+456		—35		+8	+38		—4
			—19 577		762		—54			63	
2,0	—0,3092	—0,31018		+1 217		—89			+101		—7
			—18 360		673					56	
2,2	—0,4922	—0,49378		+1 890					+157		

Интерполируя, легко находим, что $\varphi = 0$ при $x = 1{,}6857$. Точное значение x с четырьмя десятичными знаками равно 1,6858.

Можно было бы получить еще ряд формул. Но мы этим ограничимся, так как уверены, что читатель или сам сумеет получить удобные в данных конкретных условиях формулы, или найдет их в многочисленных руководствах, посвященных этому вопросу. Заметим только, что и в этом случае могут найти применение формулы типа (47) предыдущего параграфа.

§ 7. Оценка погрешности, сходимость и устойчивость разностных методов решения обыкновенных дифференциальных уравнений

1. Линейные разностные уравнения. Прежде чем переходить к вопросам, указанным в заголовке данного параграфа, рассмотрим некоторые вопросы теории линейных разностных уравнений. Эта теория близка к теории линейных обыкновенных дифференциальных уравнений. Поэтому доказательства здесь приводиться не будут.

Линейным разностным уравнением k-го порядка называют выражение вида

$$a_k(n)\,y_{n+k} + a_{k-1}(n)\,y_{n+k-1} + \ldots + a_0(n)\,y_n = b(n), \qquad (1)$$

где $a_i(n)$, $b(n)$ — заданные функции целочисленного аргумента n, $a_k(n) \neq 0$, $a_0(n) \neq 0$. Если $b(n) \equiv 0$, т. е. (1) принимает вид

$$a_k(n)\, y_{n+k} + a_{k-1}(n)\, y_{n+k-1} + \ldots + a_0(n)\, y_n = 0, \qquad (2)$$

то уравнение называется однородным. В противном случае уравнение называется неоднородным. Решением уравнения (1) или (2) называется всякая функция целочисленного аргумента n, после подстановки которой в уравнение последнее обращается в тождество по n. Ясно, что если задать $y_0, y_1, \ldots, y_{k-1}$ (начальные значения), то мы можем по (1) или (2) последовательно вычислить все y_n.

Имеют место следующие утверждения:

1. Если $y_n^{(1)}, y_n^{(2)}, \ldots, y_n^{(k)}$ являются частными решениями однородного линейного разностного уравнения (2), то и любая их линейная комбинация

$$z_n = C_1 y_n^{(1)} + C_2 y_n^{(2)} + \ldots + C_k y_n^{(k)} \qquad (3)$$

с постоянными коэффициентами C_i будет являться решением уравнения (2).

2. Если, кроме того, определитель

$$\begin{vmatrix} y_1^{(1)} & y_2^{(1)} & \ldots & y_k^{(1)} \\ y_1^{(2)} & y_2^{(2)} & \ldots & y_k^{(2)} \\ \ldots & \ldots & \ldots & \ldots \\ y_1^{(k)} & y_2^{(k)} & \ldots & y_k^{(k)} \end{vmatrix} \qquad (4)$$

отличен от нуля, то любое частное решение (2) может быть представлено в виде (3), т. е. (3) дает общее решение уравнения (2). Частные решения $y_n^{(1)}, y_n^{(2)}, \ldots, y_n^{(k)}$ в этом случае называют линейно независимыми.

3. Общее решение уравнения (1) можно представить как сумму какого-то частного его решения и общего решения однородного уравнения (2).

4. Если $y_n^{(1)}, y_n^{(2)}, \ldots, y_n^{(k)}$ являются частными решениями (2), для которых определитель (4) отличен от нуля, то функция

$$z_n = \sum_{i=0}^{n-1} \frac{\begin{vmatrix} y_{i+1}^{(1)} & y_{i+1}^{(2)} & \ldots & y_{i+1}^{(k)} \\ y_{i+2}^{(1)} & y_{i+2}^{(2)} & \ldots & y_{i+2}^{(k)} \\ \ldots & \ldots & \ldots & \ldots \\ y_{i+k-1}^{(1)} & y_{i+k-1}^{(2)} & \ldots & y_{i+k-1}^{(k)} \\ y_n^{(1)} & y_n^{(2)} & \ldots & y_n^{(k)} \end{vmatrix}}{\begin{vmatrix} y_{i+1}^{(1)} & y_{i+1}^{(2)} & \ldots & y_{i+1}^{(k)} \\ y_{i+2}^{(1)} & y_{i+2}^{(2)} & \ldots & y_{i+2}^{(k)} \\ \ldots & \ldots & \ldots & \ldots \\ y_{i+k}^{(1)} & y_{i+k}^{(2)} & \ldots & y_{i+k}^{(k)} \end{vmatrix}} \cdot b(i) \qquad (5)$$

является частным решением неоднородного уравнения (1) (аналог метода вариации постоянных).

Если коэффициенты $a_i(n)$ уравнения (1) или (2) не зависят от n (тогда мы их будем обозначать просто a_i), то будем говорить, что (1) или (2) являются уравнениями с постоянными коэффициентами. Начиная с этого момента, мы будем предполагать, что $a_i(n)$ не зависят от n.

Будем тогда разыскивать решение уравнения (2) в виде

$$y_n = z^n. \tag{6}$$

При этом для определения неизвестных значений z получим алгебраическое уравнение

$$a_k z^k + a_{k-1} z^{k-1} + \ldots + a_1 z + a_0 = 0. \tag{7}$$

Это уравнение называется *характеристическим* для (2). Если все корни $z_1, z_2, \ldots, z_k$ характеристического уравнения простые, то

$$z_1^n, \; z_2^n, \; \ldots, \; z_k^n \tag{8}$$

образуют линейно независимую систему решений (2). Часть корней z_i или все корни могут оказаться комплексными. Мы ограничимся случаем, когда все коэффициенты a_i действительны. Поэтому комплексные корни будут попарно сопряжены. Пусть, например,

$$z_k = \rho(\cos\varphi + i\sin\varphi); \quad z_j = \rho(\cos\varphi - i\sin\varphi). \tag{9}$$

Тогда этой паре комплексно-сопряженных корней будут соответствовать два действительных частных решения уравнения (2):

$$\rho^n \cos n\varphi, \quad \rho^n \sin n\varphi. \tag{10}$$

Если произвести такую замену для каждой пары комплексно-сопряженных корней, то полученные решения совместно с решениями, соответствующими действительным корням, снова образуют k линейно независимых решения (2).

Если z_i является корнем (7) кратности r, то частными решениями (2) будут:

$$z_i^n, \quad nz_i^n, \; n^2 z_i^n, \; \ldots, \; n^{r-1} z_i^n. \tag{11}$$

И в этом случае комплексные корни можно заменить действительными. Во всех случаях мы сможем получить k линейно независимых действительных решения уравнения (2).

2. Разностное уравнение для погрешности приближенного решения. Перейдем теперь к вопросу об оценке погрешности разностных способов решения обыкновенных дифференциальных уравнений. Ограничимся случаем одного уравнения первого порядка

$$\frac{dy}{dx} = f(x, y). \tag{12}$$

Будем предполагать, что $f(x, y)$ обладает непрерывными производными до некоторого порядка $p \geqslant 1$ в области G[1]):

$$|y - y(x)| \leqslant r \quad (x_0 \leqslant x \leqslant x_0 + a), \tag{13}$$

где $y(x)$ — решение (12), удовлетворяющее начальному условию $y(x_0) = y_0$, а r таково, что приближенное решение, выходящее за пределы области G, становится неинтересным. При этом $y(x)$ будет обладать в G непрерывными производными до порядка $p+1$.

Численное решение (12) будем отыскивать по формуле

$$\sum_{i=0}^{k} \alpha_i y_{n+1} - h \sum_{i=0}^{k} \beta_i f_{n+i} = 0, \tag{14}$$

предполагая, что начальные значения $y_0, y_1, \ldots, y_{k-1}$ нам заданы.

Прежде чем переходить к основным вопросам этого параграфа, наложим некоторые ограничения на коэффициенты формулы (14). При этом будет удобно ввести в рассмотрение многочлены

$$\left.\begin{aligned} \rho(z) &= \alpha_k z^k + \alpha_{k-1} z^{k-1} + \ldots + \alpha_1 z + \alpha_0, \\ \sigma(z) &= \beta_k z^k + \beta_{k-1} z^{k-1} + \ldots + \beta_1 z + \beta_0 \end{aligned}\right\} \tag{15}$$

и оператор E, определенный при помощи равенства

$$E y_n = y_{n+1} \quad \text{или} \quad E y(x) = y(x+h). \tag{16}$$

Тогда (14) можно записать в виде

$$\rho(E)\, y_n - h\sigma(E)\, f_n = 0. \tag{17}$$

Естественно считать, что *все коэффициенты* α_i *и* β_i *действительны и что* $\alpha_k \neq 0$. Предположим, далее, что $\rho(z)$ *и* $\sigma(z)$ *не имеют общих множителей*. Это связано со следующими соображениями. Если бы нашелся такой многочлен $\varphi(z)$, что

$$\rho(z) = \varphi(z)\, \rho_1(z) \quad \text{и} \quad \sigma(z) = \varphi(z)\, \sigma_1(z), \tag{18}$$

где $\rho_1(z)$ и $\sigma_1(z)$ — многочлены, то (17) можно было бы записать в виде

$$\varphi(E)\, [\rho_1(E)\, y_n - h\sigma_1(E)\, f_n] = 0. \tag{19}$$

---

[1]) По поводу сходимости и оценки погрешности разностного метода в случае разрывной правой части см. Б. М. Будак и А. Д. Горбунов, О разностном методе решения задачи Коши для уравнения $y' = f(x, y)$ и для системы уравнений $x_i' = X_i(t, x_1, \ldots, x_n)$, $i = 1, \ldots, n$ с разрывными правыми частями, Вестник МГУ, сер. матем. № 5, 1958, стр. 7—11, или А. Д. Горбунов и Б. М. Будак, О разностном методе решения задачи Коши для системы уравнений $x_i' = X_i(t, x_1, \ldots, x_n)$ $(i = 1, \ldots, n)$ с разрывными правыми частями, Научные доклады Высшей школы, физ.-мат. науки, № 6, 1958, стр. 25—29.

Если обозначить

$$\rho_1(E)\,y_n - h\sigma_1(E)\,f_n = \psi_n, \tag{20}$$

то (19) перейдет в

$$\varphi(E)\,\psi_n = 0. \tag{21}$$

Начальных условий $y_0, y_1, \ldots, y_{k-1}$ будет достаточно для того, чтобы найти сначала решение (21), а затем (20). Так как разностное уравнение (21) линейно и имеет постоянные коэффициенты, то ψ_n находятся без труда. Но тогда y_n будут находиться из уравнения (20), имеющего порядок ниже, чем уравнение (17).

Нас будут интересовать только такие формулы (14), которые обеспечивают равномерную сходимость приближенного решения к точному при $h \to 0$. (Предполагается, конечно, что начальные значения задаются точно и все вычисления производятся точно.) Поэтому потребуем, чтобы для любого $\varepsilon > 0$ существовало такое $\delta > 0$, что как только $h < \delta$, то

$$-\varepsilon < y_{n+i} - y(x) < \varepsilon \qquad (i = 0,\ 1,\ 2,\ \ldots,\ k). \tag{22}$$

Здесь $nh = x - x_0$. Вследствие (22) имеем:

$$\sum_{i=0}^{k} \alpha_i y(x) = \sum_{i=0}^{k} \alpha_i (y_{k+i} + \eta_i),$$

где $|\eta_i| < \varepsilon$. Таким образом, по (14)

$$|y(x)| \left| \sum_{i=0}^{k} \alpha_i \right| \leqslant \varepsilon \sum_{i=0}^{k} |\alpha_i| + O(h). \tag{23}$$

Так как, вообще говоря, $y(x) \not\equiv 0$, то $\sum\limits_{i=0}^{k} \alpha_i = 0$ или

$$\rho(1) = 0. \tag{24}$$

Это третье ограничение на (14).

Положим

$$\rho(z) = (z-1)\,\rho_1(z). \tag{25}$$

При этом (14) перейдет в

$$\rho_1(E)\,(y_{i+1} - y_i) - h\sigma(E)\,f_i = 0. \tag{26}$$

Суммируя равенства (26) по i от 0 до n, получим:

$$\rho_1(E)\,(y_{n+1} - y_0) = \sigma(E) \sum_{i=0}^{n} h f_i. \tag{27}$$

Переходя в (27) к пределу при $n \to \infty$, получим:

$$\rho_1(1)\,[y(x) - y_0] = \sigma(1) \int_{x_0}^{x} f\,[t,\ y(t)]\,dt. \tag{28}$$

Так как $y(x)$ должно удовлетворять уравнению (12), а $\sigma(1) \neq 0$ в силу второго предположения, то

$$\rho_1(1) = \rho'(1) = \sigma(1). \tag{29}$$

Это четвертое ограничение на (14). Равенства (24) и (29) необходимы для равномерной сходимости точного решения разностного уравнения к точному решению дифференциального.

Будем теперь под y_n понимать значение точного решения $y(x)$ дифференциального уравнения (12), удовлетворяющего начальному условию $y(x_0) = y_0$, в точке $x_n = x_0 + nh$ и под f_n — значение $f(x_n, y_n)$. Эти величины не будут удовлетворять уравнению (14), но будут удовлетворять некоторому другому разностному уравнению

$$\rho(E)\, y_n - h\sigma(E)\, f_n + l_n = 0, \tag{30}$$

где l_n можно выразить по формуле (42) § 5 и в некоторых случаях можно оценить.

Обозначим через $\tilde{y}_n$ численное решение, фактически полученное по формуле (14), и через ε_n — разность

$$\tilde{y}_n - y_n = \varepsilon_n. \tag{31}$$

Так как формула (14) применима лишь в том случае, если заданы первые k значений $\tilde{y}_i : \tilde{y}_0, \tilde{y}_1, \ldots, \tilde{y}_{k+1}$, то будем предполагать, что известны оценки для первых k значений ε_i. Задача заключается в том, чтобы получить оценки для всех ε_i.

При численном решении мы получим $\tilde{y}_n$, удовлетворяющие не уравнению (14), а некоторому другому уравнению:

$$\rho(E)\, \tilde{y}_n = h\sigma(E)\, \tilde{f}_n + \eta_n. \tag{32}$$

Здесь $\tilde{f}_n = f(x_n, \tilde{y}_n)$ и η_n — некоторая величина, появление которой вызывается следующими причинами:

1. Как правило, мы можем находить не точные значения $\tilde{f}_n$, а лишь приближенные f_n^*.

2. Вычисления по формуле (14) ведутся с округлениями.

3. Если (14) — интерполяционная формула, то уравнение для определения $\tilde{y}_{n+k}$ может быть решено только приближенно.

Таким образом, η_n определяется ошибками, возникающими на данном шаге при вычислении $\tilde{y}_{n+k}$ по формуле (14) без учета ошибок начальных значений и ошибки метода. Величину η_k иногда также удается оценить.

Вычитая (30) из (32), получим:

$$\rho(E)(\tilde{y}_n - y_n) = h\sigma(E)(\tilde{f}_n - f_n) + \eta_n + l_n. \tag{33}$$

Применив формулу Лагранжа

$$\tilde{f}_n - f_n = f(x_n, \tilde{y}_n) - f(x_n, y_n) = (\tilde{y}_n - y_n) f'_y(x_n, \zeta_n)$$
$$(\zeta_n \in (\tilde{y}_n, y_n)) \qquad (34)$$

и обозначив $f'_y(x_n, \zeta_n) = m_n$, мы можем переписать (33) в виде

$$\rho(E)\varepsilon_n = h\sigma(E) m_n \varepsilon_n + \eta_n + l_n. \qquad (35)$$

Отсюда мы можем получить последовательно оценки для всех ε_i, если известны оценки для $\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_{k-1}$, l_n, η_n, m_n.

3. Оценки погрешности решений, получаемых по формулам Адамса. Чтобы проиллюстрировать предыдущие рассуждения и показать, как их можно применять в практических случаях, рассмотрим подробнее экстраполяционную и интерполяционную формулы Адамса. При этом мы будем считать, что $\eta_n \equiv 0$.

Экстраполяционная формула Адамса может быть записана в виде

$$\tilde{y}_{k+1} = \tilde{y}_k + h \sum_{i=0}^{r} \alpha_i \tilde{f}_{k-1}, \qquad (36)$$

где

$$\alpha_i = (-1)^i \sum_{\nu=i}^{r} C_\nu^i a_\nu \qquad (37)$$

и

$$a_\nu = \frac{1}{\nu!} \int_0^1 t(t+1)\ldots(t+\nu-1)\,dt. \qquad (38)$$

При этом

$$\varepsilon_{k+1} = \tilde{y}_{k+1} - y_{k+1} = \tilde{y}_{k+1} - y_k - \int_{x_k}^{x_{k+1}} f[\xi, y(\xi)]\,d\xi. \qquad (39)$$

Последнее выражение можно записать в виде

$$\varepsilon_{k+1} = \varepsilon_k + h \sum_{i=0}^{r} \alpha_i [\tilde{f}_{k-i} - f_{k-i}] + l_k. \qquad (40)$$

Здесь через l_k обозначено следующее выражение:

$$l_k = h \sum_{i=0}^{r} \alpha_i f_{k-i} - \int_{x_k}^{x_{k+1}} f[\xi, y(\xi)]\,d\xi. \qquad (41)$$

Величину l_k нетрудно оценить. Для этого заменим подынтегральное выражение в (41) интерполяционным многочленом Ньютона для

интерполирования назад с остаточным членом

$$f[\xi, y(\xi)] = f_k + tf^1_{k-1/2} + \ldots + \frac{t(t+1)\ldots(t+r-1)}{r!} f^r_{k-r/2} + \\ + \frac{h^{r+1} t(t+1)\ldots(t+r)}{(r+1)!} \frac{d^{r+1} f[\xi, y(\xi)]}{d\xi^{r+1}}\Big|_{\xi=\eta}. \quad (42)$$

Проинтегрировав интерполяционный многочлен, мы получим первую сумму в левой части (41). Таким образом,

$$l_k = -h^{r+2} \int_0^1 \frac{t(t+1)\ldots(t+r)}{(r+1)!} \frac{d^{r+1} f[\xi, y(\xi)]}{d\xi^{r+1}}\Big|_{\xi=\eta} dt. \quad (43)$$

Используя теорему о среднем, получим:

$$l_k = -h^{r+2} \frac{d^{r+1} f[\xi, y(\xi)]}{d\xi^{r+1}}\Big|_{\xi=\theta} \int_0^1 \frac{t(t+1)\ldots(t+r)}{(r+1)!} dt \quad (44)$$

и, обозначая через M_{r+1} максимум модуля $\frac{d^{r+1} f[\xi, y(\xi)]}{d\xi^{r+1}}$ в рассматриваемой области, найдем:

$$|l_k| \leqslant l = h^{r+2} a_{r+1} M_{r+1}. \quad (45)$$

Переходя к абсолютным величинам и обозначая через L постоянную Липшица для функции $f(x, y)$, будем иметь:

$$|\varepsilon_{k+1}| \leqslant |\varepsilon_k| + hL \sum_{i=0}^{r} |\alpha_i| |\varepsilon_{k-i}| + l. \quad (46)$$

Это неравенство можно получить прямо из формулы (35) при $\eta_n = 0$, используя оценки для l_n и m_n.

Покажем теперь, как можно перейти от этой рекуррентной оценки к оценке $|\varepsilon_n|$ через данные в начале счета величины. Рассмотрим разностное уравнение

$$E_{k+1} = E_k + hL \sum_{i=0}^{r} |\alpha_i| E_{k-1} + l. \quad (47)$$

Если в качестве начальных значений для E_k выбрать положительные величины, большие чем модули соответствующих ε_k, то при любом k будем иметь $E_k \geqslant |\varepsilon_k|$.

Характеристическое уравнение для (47) будет иметь вид

$$\varphi(z) = z^{r+1} - z^r - hL \sum_{i=0}^{r} |\alpha_i| z^{r-i} = 0. \quad (48)$$

Если обозначить его корни через $z_1, z_2, \ldots, z_{r+1}$ и частное решение уравнения (47) через D, то общее решение (47) может быть записано в виде

$$E_k = C_1 z_1^k + C_2 z_2^k + \ldots + C_{r+1} z_{r+1}^k + D. \tag{49}$$

Так как $\varphi(1) < 0$ и

$$\varphi(1+hLA) = (1+hLA)^r\, hLA - hL \sum_{i=0}^{r} |\alpha_i|\, (1+hLA)^{r-i} > 0, \tag{50}$$

где

$$A = \sum_{i=0}^{r} |\alpha_i|, \tag{51}$$

то уравнение (48) имеет корень z_1, заключенный в пределах

$$1 < z_1 < 1 + hLA. \tag{52}$$

Поэтому, полагая $C_2 = C_3 = \ldots = C_{r+1} = 0$ в (49), мы получим такое решение $E_k = C_1 z_1^k + D$ уравнения (47), которое при достаточно большом положительном C_1 будет удовлетворять неравенству $E_k \geqslant |\varepsilon_k|$ при всех $k = 0, 1, \ldots, r-1$, а следовательно и при всех $k \geqslant 0$. Выбором C_1 займемся после определения константы D, представляющей собой частное решение уравнения (47). Подставляя в (47) D вместо E_k, получим:

$$l + hL \sum_{i=0}^{r} |\alpha_i|\, D = 0, \tag{53}$$

откуда

$$D = -\frac{l}{hL \sum_{i=0}^{r} |\alpha_i|} = -\frac{l}{hLA}. \tag{54}$$

Поэтому

$$E_k = C_1 z_1^k - \frac{l}{hLA}. \tag{55}$$

Подберем теперь постоянную C_1. Пусть начальные r значений $\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_{r-1}$ нам известны и все они не превышают по абсолютной величине $\varepsilon > 0$. Тогда, для того чтобы все E_k ($k = 0, 1, \ldots, r-1$) были не меньше ε, достаточно взять

$$C_1 = \varepsilon + \frac{l}{hLA}. \tag{56}$$

Таким образом, мы находим:

$$|\tilde{y}_k - y_k| \leqslant E_k = \varepsilon z_1^k + \frac{l}{hLA}(z_1^k - 1). \tag{57}$$

Если воспользоваться правой частью неравенства (51), то (57) можно заменить более грубым неравенством:

$$|\tilde{y}_k - y_k| \leqslant \varepsilon (1 + hLA)^k + \frac{l}{hLA}[(1 + hLA)^k - 1]. \tag{58}$$

Воспользуемся теперь неравенством

$$\left(1 + \frac{\alpha}{i}\right)^i < e^{\alpha} \tag{59}$$

и заменим h на $\frac{x_k - x_0}{k}$. Тогда

$$|\tilde{y}_k - y_k| \leqslant \varepsilon e^{AL(x_k - x_0)} + \frac{l}{hLA}\left[e^{AL(x_k - x_0)} - 1\right]. \tag{60}$$

Если подставить сюда выражение для l и положить $\varepsilon = 0$, то увидим, что при $h \to 0$ правая часть стремится к нулю при фиксированном положении x_k. Таким образом, *приближенное решение, полученное по методу Адамса, сходится к точному решению, если начальные значения задавать точно и все вычисления производить точно.*

Для интерполяционного метода Адамса рассуждения будут аналогичны. Формулу записываем в виде

$$\tilde{y}_{k+1} = \tilde{y}_k + h\sum_{i=0}^{r} \beta_i \tilde{f}_{k-i+1}, \tag{61}$$

где

$$\beta_i = (-1)^i \sum_{j=i}^{r} C_j^i b_j \tag{62}$$

и

$$b_i = \int_{-1}^{0} \frac{t(t+1)\ldots(t+i-1)}{i!}\, dt. \tag{63}$$

Как и ранее, находим:

$$|\varepsilon_{k+1}| \leqslant |\varepsilon_k| + Lh\sum_{i=0}^{r} |\beta_i||\varepsilon_{k+1-i}| + h^{r+2}|b_{r+1}|M_{r+1}. \tag{64}$$

Мажорирующее уравнение примет вид

$$E_{k+1} = E_k + Lh\sum_{i=0}^{r} |\beta_i| E_{k+1-i} + \bar{l}, \quad \bar{l} = h^{r+2}|b_{r+2}|M_{r+1}. \tag{65}$$

Его частное решение равно

$$D = -\frac{\bar{l}}{LhB}, \qquad B = \sum_{i=0}^{r} |\beta_i|. \tag{66}$$

Характеристическое уравнение в данном случае имеет вид

$$z^r = z^{r-1} + Lh \sum_{i=0}^{r} |\beta_i| z^{r-1}. \tag{67}$$

Оно имеет единственный корень, заключенный между 1 и $+\infty$, если только $Lh|\beta_0| < 1$. Обозначим его через $\bar{z}_0$. Тогда получаем следующую оценку:

$$|\varepsilon_k| < \varepsilon \bar{z}_0^k + \frac{\bar{l}}{LhB}(\bar{z}_0^k - 1), \tag{68}$$

где через ε, как и раньше, обозначено наибольшее из значений $|\varepsilon_0|, |\varepsilon_1|, \ldots, |\varepsilon_{r-1}|$.

При этом мы считали, что уравнение относительно y_{k+1} на каждом шаге может быть решено точно.

Для сравнения двух формул Адамса в смысле точности приведем уравнения для определения z_0 и коэффициенты при $h^{r+1}M_{r+1}/L$ в том и другом случае.

Экстраполяционная формула Адамса

Уравнение	z_0	Коэффициент
$z = 1 + Lh$		0,5
$z^2 = z + \frac{Lh}{2}(3z+1)$	$z_0 = \left(\frac{1}{2} + \frac{3}{4}Lh\right) + \sqrt{\left(\frac{1}{2} + \frac{3}{4}Lh\right)^2 + \frac{Lh}{2}}$	0,2083
$z^3 = z^2 + \frac{Lh}{12}(23z^2 + 16z + 5)$	$1 < z_0 < 1 + Lh \cdot \frac{11}{3}$	0,1023
$z^4 = z^3 + \frac{Lh}{24}(55z^3 + 59z^2 + 37z + 9)$	$1 < z_0 < 1 + \frac{20}{3}Lh$	0,0523

Интерполяционная формула Адамса

Уравнение	$\bar{z}_0$	Коэффициент
$z = 1 + Lhz$	$\frac{1}{1 - Lh}$	0,5
$z = 1 + \frac{Lh}{2}(z+1)$	$\frac{1 + \frac{Lh}{2}}{1 - \frac{Lh}{2}}$	0,0833
$z^2 = z + \frac{Lh}{12}(5z^2 + 8z + 1)$		0,0357
$z^3 = z^2 + \frac{Lh}{24}(9z^3 + 19z^2 + 5z + 1)$		0,0186
$z^4 = z^3 + \frac{Lh}{720}(251z^4 + 646z^3 + 264z^2 + 106z + 19)$		0,0105

Как мы видим из этой таблицы, оценки для интерполяционной формулы Адамса получились лучшими, чем для экстраполяционной. Это подтверждает, до некоторой степени, факт, обнаруженный нами при практических вычислениях.

Совершенно аналогично можно получить оценки и для других формул, которые нами введены.

4. Устойчивость разностных методов решения дифференциальных уравнений. Вернемся снова к общей формуле (14). Нам будет полезно иметь выражение ε_{n+k} через все предыдущие ε_i. Введем обозначения:

$$\lambda = h\beta_k\alpha_k^{-1}; \quad u_\nu = y_\nu - \lambda f_\nu; \quad \tilde{u}_\nu = \tilde{y}_\nu - \lambda \tilde{f}_\nu. \tag{69}$$

При этом (33) можно записать в виде

$$\rho(E)(\tilde{u}_n - u_n) = q_n, \tag{70}$$

где

$$q_n = \sum_{j=0}^{k-1} (h\beta_j - \lambda\alpha_j) E^j (\tilde{f}_n - f_n) + \eta_n + l_n. \tag{71}$$

Обращаем внимание на тот факт, что в правой части суммирование происходит до $j = k-1$, а не до $j = k$. В этом и смысл указанного преобразования.

Будем разыскивать решение уравнения (70) в виде

$$\tilde{u}_n - u_n = z_n + \theta_n, \tag{72}$$

где z_n является решением неоднородного уравнения (70) и удовлетворяет нулевым начальным условиям, а θ_n удовлетворяет уравнению (70) при $q_n \equiv 0$ и тем же начальным условиям, что и $\tilde{u}_n - u_n$. Таким образом,

$$\rho(E) z_n = q_n, \quad z_0 = z_1 = \ldots = z_{k-1} = 0; \tag{73}$$

$$\rho(E)\theta_n = 0, \quad \theta_\nu = \tilde{u}_\nu - u_\nu \qquad (\nu = 0, 1, 2, \ldots, k-1). \tag{74}$$

Для отыскания z_n введем величины g_n (аналог функции Грина), удовлетворяющие следующим условиям:

$$g_i = \begin{cases} 0 & \text{при } i < k, \\ \alpha_k^{-1} & \text{при } i = k. \end{cases} \tag{75}$$

$$\rho(E) g_n = \begin{cases} 1 & \text{при } n = 0, \\ 0 & \text{при } n > 0. \end{cases} \tag{76}$$

При этом

$$\sum_{\nu=0}^{n+k} g_{n+k-\nu} q_\nu = \sum_{\nu=0}^{n+k} g_{n+k-\nu} \rho(E) z_\nu = \sum_{\nu=0}^{n+k} g_{n+k-\nu} \sum_{i=0}^{k} \alpha_i z_{\nu+i} =$$

$$= \sum_{i=0}^{k} \alpha_i \sum_{\nu=0}^{n+k} g_{n+k-\nu} z_{\nu+i}. \tag{77}$$

Введем вместо ν новый индекс суммирования, положив $\nu + i = j$. Тогда

$$\sum_{\nu=0}^{n+k} g_{n+k-\nu} q_\nu = \sum_{i=0}^{k} \alpha_i \sum_{j=i}^{n+k+i} g_{n+k-j+i} z_j. \tag{78}$$

Так как $z_j = 0$ при $j < k$, то нижний индекс во внутренней сумме (78) можно считать нулем. Верхний предел суммирования по j можно положить равным $n+k$ в силу начальных условий для g_n. Поэтому

$$\sum_{\nu=0}^{n+k} g_{n+k-\nu} q_\nu = \sum_{i=0}^{k} \alpha_i \sum_{j=0}^{n+k} g_{n+k-j+i} z_j = \sum_{j=0}^{n+k} z_j \sum_{i=0}^{k} \alpha_i g_{n+k-j+i} =$$

$$= \sum_{j=0}^{n+k} z_j \rho(E)\, g_{n+k-j}. \tag{79}$$

Отсюда, в силу (76), получим:

$$z_{n+k} = \sum_{\nu=0}^{n+k} g_{n+k-\nu} q_\nu, \tag{80}$$

или по (75)

$$z_{n+k} = \sum_{\nu=0}^{n} g_{n+k-\nu} q_\nu. \tag{81}$$

Следовательно,

$$\tilde{u}_{n+k} - u_{n+k} = \sum_{\nu=0}^{n} g_{n+k-\nu} q_\nu + \theta_{n+k}. \tag{82}$$

Отметим тот факт, что для отыскания g_n и θ_n нам придется находить решения одного и того же уравнения

$$\rho(E)\, v_n = 0. \tag{83}$$

Поведение решений этого уравнения при возрастании n будет определяться расположением корней характеристического уравнения

$$\rho(z) = 0. \tag{84}$$

Если все корни уравнения (84) расположены внутри или на границе единичного круга, причем корни, лежащие на границе, не являются кратными, то при любых начальных значениях v_n будут оставаться ограниченными. Если же уравнение (84) имеет корни, расположенные вне единичной окружности или же кратные на этой окружности, то имеется бесчисленное множество решений уравнения (83), неограниченно возрастающих по абсолютной величине по показательному или степенному закону при $n \to \infty$. Эти выводы основываются на представлении общего решения линейного однородного разностного уравнения с постоянными коэффициентами, данном в начале параграфа.

Таким образом, при наличии кратных корней уравнения (84) на единичной окружности или при наличии корней, расположенных вне единичной окружности, мы столкнемся, вообще говоря, с неограниченным возрастанием $|\theta_n|$ и $|g_n|$. Это вызовет по (82) неограниченное возрастание $|\tilde{u}_{n+k} - u_{n+k}|$, а тем самым и $|\tilde{y}_{n+k} - y_{n+k}|$. Такое явление чрезвычайно невыгодно для вычислительной практики, так как связано с быстрым накоплением ошибок.

Проиллюстрируем это примером. Будем искать решение дифференциального уравнения $y' = y$, удовлетворяющее начальному условию $y(0) = 1$. Решение будет разыскиваться на отрезке $[0, 1]$ с шагом $h = 0{,}1$ по разностной формуле

$$y_{n+2} + 4y_{n+1} - 5y_n = h[4f_{n+1} + 2f_n], \tag{85}$$

полученной в § 5. Это — экстраполяционная формула второго порядка, имеющая наивысший относительно h порядок ошибки на шаге, равный 4. В нашем случае (85) перейдет в

$$y_{n+2} + 3{,}6y_{n+1} - 5{,}2y_n = 0; \quad y_0 = 1. \tag{86}$$

Точное решение уравнения (86) имеет вид

$$y_n = (1 - C)\, z_1^n + Cz_2^n, \tag{87}$$

где $z_1 \approx 1{,}105168$ и $z_2 \approx -4{,}705168$ являются корнями квадратного уравнения

$$z^2 + 3{,}6z - 5{,}2 = 0 \tag{88}$$

и C — постоянная, зависящая от выбора y_1. Точным решением дифференциального уравнения будет e^x. Возьмем в качестве y_1 значение $e^{0{,}1}$ с шестью верными десятичными знаками: $y_1 = 1{,}105171$, и будем последовательно находить y_n по (85) также с шестью десятичными знаками. Результаты вычислений приведены в таблице:

n	$\tilde{y}_n$	$(y_n - \tilde{y}_n) \cdot 10^6$
0	1,000000	0
1	1,105171	0
2	1,221384	19
3	1,349907	— 48
4	1,491532	293
5	1,650001	— 1 280
6	1,815963	6 156
7	2,042538	— 28 785
8	2,089871	135 670
9	3,097662	— 638 059
10	— 0,284254	3 002 536

Как мы видим, ошибки чрезвычайно быстро растут. Хотя было сделано мало шагов, последние значения совершенно не удовлетво-

рительны. Взяв любую экстраполяционную формулу второго порядка с худшей ошибкой метода, например формулу Адамса, мы получили бы значительно лучшие результаты. В данном случае z_1^n дает неплохое представление $e^{0,1n}$ на отрезке $[0, 1]$. Так при $n = 10$ разность $e^{0,1n} - z_1^n$ достигает лишь 77 единиц шестого десятичного знака. Поэтому казалось бы, что, взяв более грубое значение y_1, при котором C в формуле (87) обратится в нуль, мы можем существенно улучшить результаты. Однако, в силу ошибок округления, второе слагаемое в (87) рано или поздно начнет сказываться и исказит результаты. Вычисления по (85) при $y_1 = 1{,}105168$ дают:

n	y_n	$(e^{0,1n} - y_n) \cdot 10^6$
0	1,000000	0
1	1,105168	3
2	1,221395	8
3	1,349852	7
4	1,491787	38
5	1,648797	— 76
6	1,821623	496
7	2,015192	— 2 149
8	2,215192	10 349
9	2,507999	— 48 396
10	2,490202	228 080

Сначала погрешности умеренны, но затем они все быстрее и быстрее растут. Если бы продолжить наши вычисления дальше, то погрешности превысили бы значения y_n.

В связи с отмеченным явлением будем называть формулу численного интегрирования (14) *устойчивой*, если все корни уравнения (84) лежат или внутри или на границе единичного круга, причем последние не являются кратными. В противном случае назовем формулу (14) *неустойчивой*. Вообще говоря, неустойчивые формулы дают неудовлетворительные численные результаты и должны применяться с большой осторожностью.

5. Оценка погрешности и сходимость устойчивых разностных методов решения дифференциальных уравнений. Дальнейшую оценку будем производить для того случая, когда формула численного интегрирования (14) устойчива.

Нам необходимо принять некоторые меры предосторожности для того, чтобы исследуемые функции не выходили из области G, в которой проводятся все рассуждения. Это потребует некоторых дополнительных предположений. Прежде всего предположим, что h

настолько мало, что для постоянной Липшица L функции $f(x, y)$ выполнено неравенство

$$|\lambda| L = h\left|\frac{\beta_k}{\alpha_k}\right| L < 1. \tag{89}$$

На основании принципа сжатых отображений можно утверждать, что если

$$u(x) = y(x) - \lambda f(x, y) \tag{90}$$

и $\tilde{u}(x)$ — заданная непрерывная функция, удовлетворяющая неравенству

$$|\tilde{u}(x) - u(x)| < r(1 - |\lambda| L), \tag{91}$$

где r было определено ранее (см. начало п. 2), то уравнение

$$\tilde{u}(x) = \tilde{y}(x) - \lambda f(x, \tilde{y}) \tag{92}$$

имеет в G одно и только одно решение $\tilde{y}(x)$. При этом

$$|\tilde{y} - y| \leqslant |\tilde{u} - u| + |\lambda| |f(x, \tilde{y}) - f(x, y)| \leqslant |\tilde{u} - u| + \lambda L |\tilde{y} - y| \tag{93}$$

или

$$|\tilde{y} - y| \leqslant (1 - |\lambda| L)^{-1} |\tilde{u} - u| = L_1 |\tilde{u} - u|. \tag{94}$$

Необходимо также наложить некоторые ограничения на погрешности η_n, θ_0, θ_1, ..., θ_{k-1}, l_n. Будем предполагать, что

$$\sum_{\nu < \frac{a}{h}} |\eta_\nu| + \max_{j \leqslant k-1} |\theta_j| < \varepsilon \tag{95}$$

и что для любой $p+1$ раз дифференцируемой функции $y(x)$

$$|\rho(E) y - h\sigma(E) y'| \leqslant M_{p+1} C' h^{p+1}, \tag{96}$$

где $M_{p+1} = \sup |y^{(p+1)}(x)|$ на $[x_0, x_0 + a]$ и C' — постоянная, не зависящая от $y(x)$.

Обозначим $\max\limits_{j \leqslant k-1} |\tilde{u}_j - u_j| = \theta$. Так как формула (14) предполагается устойчивой, то можно найти такую постоянную g, что при любых n

$$|g_n| \leqslant g; \quad |\theta_n| \leqslant g\theta. \tag{97}$$

Сначала получим искомую оценку, не исследуя вопрос о том, находятся ли $\tilde{u}_n$ и $\tilde{y}_n$ в области G или нет. Затем мы наложим еще одно ограничение на ε и h, которое обеспечит нам невыхождение $\tilde{u}_n$ и $\tilde{y}_n$ из G.

Начнем с оценки $|q_n|$. По (71) и (94) имеем:

$$|q_\nu| \leqslant h \sum_{j=0}^{k-1} |\beta_j - \beta_k \alpha_k^{-1} \alpha_j| |E^j| |\tilde{f}_\nu - f_\nu| + |\eta_\nu| + |l_\nu| \leqslant$$

$$\leqslant h \sum_{j=0}^{k-1} |\beta_j - \beta_k \alpha_k^{-1} \alpha_j| |E^j| LL_1 |\tilde{u}_\nu - u_\nu| + |\eta_\nu| + |l_\nu| \leqslant$$

$$\leqslant hLL_1 \max_{\mu \leqslant \nu+k-1} |\tilde{u}_\mu - u_\mu| \sum_{j=0}^{k-1} |\beta_j - \beta_k \alpha_k^{-1} \alpha_j| + |\eta_\nu| + |l_\nu|. \quad (98)$$

или если обозначить

$$LL_1 \sum_{j=0}^{k-1} |\beta_j - \beta_k \alpha_k^{-1} \alpha_j| = L_2, \quad (99)$$

то

$$|q_\nu| \leqslant hL_2 \max_{\mu \leqslant \nu+k-1} |\tilde{u}_\mu - u_\mu| + |\eta_\nu| + |l_\nu|. \quad (100)$$

Поэтому формула (82) даст

$$|\tilde{u}_{n+k} - u_{n+k}| \leqslant g \sum_{\nu=0}^{n} \left\{ hL_2 \max_{\mu \leqslant \nu+k-1} |\tilde{u}_\mu - u_\mu| + |\eta_\nu| + |l_\nu| \right\} + g\theta. \quad (101)$$

Но по (95)

$$g \sum_{\nu=0}^{n} |\eta_\nu| + g\theta \leqslant g\varepsilon \quad (102)$$

и по (96)

$$\sum_{\nu=0}^{n} |l_\nu| \leqslant M_{p+1} C' n h^{p+1} \leqslant M_{p+1} C' a h^p. \quad (103)$$

Следовательно,

$$|\tilde{u}_{n+k} - u_{n+k}| \leqslant hL_2 g \sum_{\nu=0}^{n} \max_{\mu \leqslant \nu+k-1} |\tilde{u}_\mu - u_\mu| + g(\varepsilon + M_{p+1} C' a h^p). \quad (104)$$

Наряду с величинами $|\tilde{u}_n - u_n|$ рассмотрим последовательность величин w_n, определенную рекуррентным соотношением:

$$w_{n+k} = hL_2 g \sum_{\nu=k-1}^{n+k-1} w_\nu + g(\varepsilon + M_{p+1} C' a h^p) \quad (n \geqslant 0). \quad (105)$$

Возьмем

$$w_{k-1} = g(\varepsilon + M_{p+1} C' a h^p). \quad (106)$$

При этом, очевидно,

$$w_{k-1} \geqslant \max_{\mu \leqslant k-1} |\tilde{u}_\mu - u_\mu|. \quad (107)$$

Взяв в (105) $n = 0$, получим:

$$w_k = hL_2 g w_{k-1} + g\left(\varepsilon + M_{p+1}C'ah^p\right) = \\ = g\left(\varepsilon + M_{p+1}C'ah^p\right)(1 + hL_2 g) > w_{k-1} \tag{108}$$

и кроме того, $w_k \geqslant \max\limits_{\mu \leqslant k} |\tilde{u}_\mu - u_\mu|$. Нетрудно доказать по индукции, что при любых $n \geqslant 0$ имеет место

$$w_{n+k} \geqslant w_{n+k-1}, \quad w_{n+k} \geqslant \max_{\mu \leqslant n+k} |\tilde{u}_\mu - u_\mu|. \tag{109}$$

Заменим в (105) $n + k$ на $n + k - 1$. Получим:

$$w_{n+k-1} = hL_2 g \sum_{\nu=k-1}^{n+k-2} w_\nu + g\left(\varepsilon + M_{p+1}C'ah^p\right). \tag{110}$$

Вычитая из (105) выражение (110), найдем:

$$w_{n+k} = w_{n+k-1}(1 + hL_2 g). \tag{111}$$

Решение разностного уравнения (111) с начальным условием (106) имеет вид

$$w_{n+k} = (1 + hL_2 g)^{n+1} w_{k-1} < e^{gL_2 a} w_{k-1}. \tag{112}$$

Итак,

$$|\tilde{u}_{n+k} - u_{n+k}| \leqslant g\left(\varepsilon + M_{p+1}C'ah^p\right) e^{gL_2 a}. \tag{113}$$

Искомая оценка получена. Необходимо только обеспечить, чтобы мы не вышли за пределы области G. Обеспечим для этого выполнение условия (91). Это потребует еще одного ограничения. Будем предполагать, что

$$ge^{gL_2 a}\left(\varepsilon + M_{p+1}C'ah^p\right) < r(1 - |\lambda| L). \tag{114}$$

При этом

$$|\tilde{u}_\nu - u_\nu| \leqslant g\theta < g\varepsilon < r(1 - |\lambda| L) \qquad (\nu \leqslant k - 1). \tag{115}$$

Предполагая, что неравенство

$$|\tilde{u}_\nu - u_\nu| < r(1 - |\lambda| L) \tag{116}$$

выполнено при $\nu = k - 1, k, \ldots, k + n - 1$, получаем по (113) и (114):

$$|\tilde{u}_{n+k} - u_{n+k}| < r(1 - |\lambda| L). \tag{117}$$

Таким образом, пока $x_\nu \leqslant a$, неравенство (116) будет иметь место. Неравенство (113) полностью обосновано.

Теперь, использовав неравенство (94), будем иметь:

$$|\tilde{y}_{n+k} - y_{n+k}| < g\left(\varepsilon + M_{p+1}C'ah^p\right) e^{gL_2 a}(1 - |\lambda| L)^{-1}. \tag{118}$$

Это и есть окончательная оценка.

Полученная нами оценка является очень грубой. Но и более точные оценки будут иметь очень ограниченное применение, так как обычно бывает трудно ограничить входящие в оценку величины. До сих пор хороших эффективных оценок не имеется.

Отметим одно следствие из полученной оценки. Если в (118) устремить h к нулю, зафиксировав x и считая $n = \frac{x - x_0}{h}$, то в пределе получим:

$$|\tilde{y}_n - y(x)| < C_2\varepsilon. \tag{119}$$

Это неравенство характерно для устойчивых формул.

Неравенство (118) можно было бы обобщить на случай систем уравнений, если ввести в рассмотрение вместо абсолютных величин соответствующие нормы.

§ 8. Решение краевых задач для обыкновенных дифференциальных уравнений методом конечных разностей

Замена дифференциального оператора разностным может быть использована не только при решении задачи Коши для обыкновенных дифференциальных уравнений, но и при решении краевых задач. Пусть дано дифференциальное уравнение

$$y'' = f(x, y, y'), \tag{1}$$

и нам требуется найти его решение, удовлетворяющее одному из условий следующего вида:

$$y(a) = \alpha, \quad y(b) = \beta \tag{2}$$

или

$$y'(a) = \alpha, \quad y'(b) = \beta, \tag{3}$$

или

$$y'(a) - ky(a) = \alpha; \quad y'(b) + k_1 y(b) = \beta. \tag{4}$$

Будем предполагать, что выполнены все условия, обеспечивающие существование такого решения.

Разобьем отрезок $[a, b]$ на n равных частей точками

$$x_i = a + ih \qquad \left(i = 0, 1, 2, \ldots, n; \; h = \frac{b - a}{n}\right). \tag{5}$$

Эти точки x_i будем называть *узлами*. В каждом из узлов заменим производные через комбинацию значений функции в некоторых узлах по формулам численного дифференцирования. Получим систему $n - 1$ уравнений относительно $y(x_i) = y_i$. Присоединяем к ним уравнения, получающиеся из краевых условий. Будем иметь $n + 1$ уравнений относительно $y_0, y_1, \ldots, y_n$. Решаем эту систему и находим приближенные значения искомого решения в узлах x_i.

Применение такого метода требует решения следующих вопросов:

1. Нужно выбрать формулы численного дифференцирования, достаточно хорошо аппроксимирующие производные и не выводящие нас за пределы промежутка.

2. Проверить разрешимость системы и указать метод ее решения.

3. Дать оценку точности полученных результатов.

Первый вопрос несложен. В главе 3 мы дали много формул численного дифференцирования. Нужно только заметить, что применение более точных формул может усложнить систему, а применение менее точных формул потребует введения большого числа узлов и тем самым увеличения числа уравнений системы.

1. Метод конечных разностей решения краевых задач для линейных дифференциальных уравнений второго порядка. Рассмотрим, как решаются второй и третий вопросы для уравнения [1])

$$y'' - q(x)\,y = r(x) \qquad (q(x) \geqslant 0). \tag{6}$$

Заменим вторую производную в узле x_i выражением

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2}. \tag{7}$$

При этом получим систему линейных алгебраических уравнений

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} - q_i y_i = r_i \qquad (i = 1,\ 2,\ \ldots,\ n-1). \tag{8}$$

Если наши краевые условия имеют вид (2), то к уравнениям (8) еще добавятся

$$y_0 = \alpha; \quad y_n = \beta. \tag{9}$$

Если наши краевые условия имеют вид (4), то добавляем к (8) еще два уравнения:

$$\frac{-y_2 + 4y_1 - 3y_0}{2h} - ky_0 = \alpha; \quad \frac{3y_n - 4y_{n-1} + y_{n-2}}{2h} + k_1 y_n = \beta. \tag{10}$$

Полученная система линейных алгебраических уравнений будет разрешима при любых α, β и r_i, если соответствующая однородная система имеет только тривиальное решение. Обозначим

$$l(y_i) = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} - q_i y_i. \tag{11}$$

Пусть дана произвольная система $n+1$ чисел: $y_0, y_1, \ldots, y_n$; докажем, что если при любых i $l(y_i) \geqslant 0$, то наибольшим положительным числом среди y_i может быть только y_0 или y_n.

---

[1]) Здесь $q(x)$ и $r(x)$ предполагаются дважды непрерывно дифференцируемыми. О разностных схемах решения краевых задач для уравнения $(k(x)\,y')' - q(x)\,y = r(x)$ с разрывными $k(x)$, $q(x)$ и $r(x)$ см. А. Н. Тихонов и А. А. Самарский, ДАН СССР, т. 122, 562—566, 1958.

Действительно, пусть $y_k = M$ — наибольшее положительное число из $y_0, y_1, \ldots, y_n$, такое, что по крайней мере одно из чисел y_{k+1} или y_{k-1} меньше M. Если бы наше утверждение было неверно, то такое y_k обязательно нашлось. По предположению

$$\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} - q_k y_k \geqslant 0. \tag{12}$$

Если заменить здесь y_{k+1} и y_{k-1} на M, то мы увеличим левую часть. Таким образом,

$$\frac{M - 2M + M}{h^2} - q_k M = -q_k M > 0. \tag{13}$$

Но это невозможно, ибо $q_k \geqslant 0$ и $M > 0$.

Точно так же доказывается, что *если мы имеем систему чисел* $y_0, y_1, \ldots, y_n$, *для которой* $l(y_i) \leqslant 0$, *то наименьшим отрицательным числом среди них может быть только* y_0 *или* y_n.

Теперь мы в состоянии доказать, что система (8) при краевых условиях (9) имеет только тривиальное решение, если $\alpha = \beta = r_i = 0$. Действительно, если бы она имела нетривиальное решение, то среди чисел $y_1, \ldots, y_n$ нашлось бы или наименьшее отрицательное или наибольшее положительное, а это противоречит только что доказанным утверждениям. Докажем теперь то же самое для системы (8) с граничными условиями (10) при $\alpha = \beta = r_i = 0$.

В граничных условиях (10) будем предполагать, что $k \geqslant 0$ и $k_1 \geqslant 0$ и по крайней мере одно из этих чисел отлично от нуля. Исключая y_2 из уравнений

$$\frac{y_2 - 2y_1 + y_0}{h^2} - q_1 y_1 = 0, \tag{14}$$

$$\frac{-y_2 + 4y_1 - 3y_0}{2h} - k y_0 = 0 \tag{15}$$

находим

$$y_1 = \frac{1 + kh}{1 - \frac{1}{2} q_1 h^2} y_0. \tag{16}$$

Совершенно так же найдем:

$$y_{n-1} = \frac{1 + k_1 h}{1 - \frac{1}{2} q_{n-1} h^2} y_n. \tag{17}$$

Пусть теперь имеется какая-то система чисел $y_0, y_1, \ldots, y_n$, удовлетворяющая нашей однородной системе, такая, что не все y_i равны нулю. Опять приходим к выводу, что наибольшее положительное из них или наименьшее отрицательное могут быть только на концах. Пусть y_0 будет наибольшим положительным значением. Будем предполагать, что h настолько мало, что $\frac{1}{2} q_1 h^2 < 1$. Из

равенства (16) следует, что $y_1 \geqslant y_0$. Так как в силу нашего предположения не может быть $y_1 > y_0$, то $y_1 = y_0$. Но тогда из доказанных нами утверждений следует, что все y_i равны друг другу. При этом из (8) и (10) при $\alpha = \beta = r_i = 0$ следует, что $y_i = 0$. Таким же образом доказывается, что среди чисел y_i нет наименьшего отрицательного.

Решение системы (8) не встречает затруднений, поэтому мы этого касаться здесь не будем

Перейдем к оценке погрешности. Будем рассматривать только граничные условия вида (9).

Докажем сначала еще одно утверждение. *Если в узловых точках даны две системы значений* $y_0, y_1, \ldots, y_n$ *и* $Y_0, Y_1, \ldots, Y_n$ *такие, что*

$$l(Y_i) \leqslant -|l(y_i)| \qquad (i = 1, 2, \ldots, n-1), \tag{18}$$

а на границе

$$Y_0 \geqslant |y_0|, \quad Y_n \geqslant |y_n|, \tag{19}$$

то при всех i *будет*

$$Y_i \geqslant |y_i|. \tag{20}$$

Действительно, из (18) следует:

$$l(Y_i - y_i) \leqslant 0; \quad l(Y_i + y_i) \leqslant 0, \tag{21}$$

а из (19)

$$\left.\begin{aligned} Y_0 - y_0 \geqslant 0, \quad Y_n - y_n \geqslant 0, \\ Y_0 + y_0 \geqslant 0, \quad Y_n + y_n \geqslant 0. \end{aligned}\right\} \tag{22}$$

Поэтому, в силу доказанных на стр. 373 — 374 утверждений, имеем (20).

Будем, как обычно, обозначать через y_i точное значение решения дифференциального уравнения (6) и через $\tilde{y}_i$ — приближенное решение, полученное путем решения системы (8). Через ε_i обозначаем разность

$$\varepsilon_i = y_i - \tilde{y}_i. \tag{23}$$

Величины ε_i удовлетворяют системе

$$\frac{\varepsilon_{i+1} - 2\varepsilon_i + \varepsilon_{i-1}}{h^2} - q_i\varepsilon_i = R_i \quad (i = 1, 2, \ldots, n-1); \quad \varepsilon_0 = \varepsilon_n = 0, \tag{24}$$

где R_i есть погрешность, вызванная заменой второй производной формулой численного дифференцирования (7), и может быть оценена как

$$|R_i| \leqslant \frac{h^2}{12} M_4, \quad M_4 = \max_{x \in [a, b]} |y^{(4)}(x)|. \tag{25}$$

Рассмотрим систему чисел η_i, удовлетворяющую условиям:

$$\left.\begin{aligned} \frac{\eta_{i+1} - 2\eta_i + \eta_{i-1}}{h^2} - q_i\eta_i = -\frac{h^2}{12} M_4 \quad (i = 1, 2, \ldots, n-1), \\ \eta_0 = \eta_n = 0. \end{aligned}\right\} \tag{26}$$

В силу только что доказанного утверждения будем иметь $\eta_i \geqslant |\varepsilon_i|$. Мы еще увеличим наши величины, если будем рассматривать значения, удовлетворяющие системе

$$\frac{\rho_{i+1} - 2\rho_i + \rho_{i-1}}{h^2} = -\frac{h^2}{12} M_4 \quad (i = 1, 2, \ldots, n-1); \quad \rho_0 = \rho_n = 0. \tag{27}$$

Итак, $\rho_i \geqslant |\varepsilon_i|$. Решение системы (27) находится без труда. Действительно, уравнение (27) говорит о том, что вторая разность величин ρ_i постоянна. Таким образом, величины ρ_i можно рассматривать как значения некоторого многочлена второй степени в точках x_i. Этот многочлен имеет вид

$$\frac{h^2}{12} M_4 \frac{(x-a)(b-x)}{2}. \tag{28}$$

Максимальное значение многочлена (28) на отрезке $[a, b]$ достигается в точке $x = \frac{a+b}{2}$, и оно равно

$$\frac{h^2 M_4 (b-a)^2}{96}. \tag{29}$$

Таким образом,

$$|\varepsilon_i| \leqslant \frac{h^2 M_4}{96} (b-a)^2. \tag{30}$$

Если решение $y(x)$ имеет ограниченную четвертую производную, то из (30) следует, что $\varepsilon_i \to 0$ при $h \to 0$.

Недостатком этой оценки, как и всех аналогичных оценок, является то, что в нее входит четвертая производная от искомого решения, которую обычно бывает трудно оценить.

2. Метод конечных разностей решения краевых задач для нелинейных дифференциальных уравнений второго порядка. Перейдем теперь к нелинейным уравнениям второго порядка. Этот случай потребует более громоздких рассуждений.

Будем рассматривать дифференциальное уравнение

$$y'' = f(x, y, y') \tag{31}$$

и граничные условия

$$\alpha_0 y(a) - \alpha_1 y'(a) = \alpha; \quad \beta_0 y(b) + \beta_1 y'(b) = \beta. \tag{32}$$

При этом будем предполагать, что $f(x, y, z)$ — непрерывная функция в некоторой области G пространства x, y, z, выпуклой относительно y и z, и что $\alpha_0, \alpha_1, \beta_0, \beta_1$ — неотрицательные числа. В дальнейшем мы наложим на функцию $f(x, y, z)$, область G и постоянные $\alpha_0, \alpha_1, \beta_0, \beta_1$ некоторые дополнительные ограничения.

Как и ранее, разбиваем отрезок $[a, b]$ на n равных частей точками:

$$a = x_0 < x_1 < x_2 < \ldots < x_{n-1} < x_n = b; \quad x_i - x_{i-1} = h = \frac{b-a}{n}. \tag{33}$$

Снова аппроксимируем дифференциальное уравнение в точке x_k соотношением

$$\frac{y_{k+1}-2y_k+y_{k-1}}{h^2}=f\left(x_k,\ y_k,\ \frac{y_{k+1}-y_{k-1}}{2h}\right) \qquad (k=1,\ 2,\ \ldots,\ n-1), \tag{34}$$

а граничные условия — соотношениями:

$$\left.\begin{aligned} R_0(y)&=\alpha_0 y_0-\alpha_1\frac{y_1-y_0}{h}=\alpha,\\ R_n(y)&=\beta_0 y_n+\beta_1\frac{y_n-y_{n-1}}{h}=\beta. \end{aligned}\right\} \tag{35}$$

Мы получили систему $n+1$ уравнений относительно $(n+1)$-го неизвестного y_k. Отличие от предыдущего состоит в том, что теперь система, вообще говоря, нелинейна.

Дадим прежде всего итерационный способ решения такой системы. Итерации будем проводить по схеме

$$\left.\begin{aligned} &\frac{y_{k+1}^{(r+1)}-2y_k^{(r+1)}+y_{k-1}^{(r+1)}}{h^2}=f\left(x_k,\ y_k^{(r)},\ \frac{y_{k+1}^{(r)}-y_{k-1}^{(r)}}{2h}\right),\\ &R_0(y^{(r+1)})=\alpha;\quad R_n(y^{(r+1)})=\beta. \end{aligned}\right\} \tag{36}$$

Здесь, как и обычно, индекс вверху означает номер приближения. На каждом шагу нам придется решать несложную систему линейных алгебраических уравнений. На основании предыдущего следует, что эта система всегда имеет единственное решение. Получим это решение в явном виде. Обозначим для краткости

$$f\left(x_k,\ y_k^{(r)},\ \frac{y_{k+1}^{(r)}-y_{k-1}^{(r)}}{2h}\right)=f_k^{(r)}. \tag{37}$$

На каждом шаге — это известная величина.

Будем разыскивать $y_k^{(r+1)}$ в виде суммы

$$y_k^{(r+1)}=z_k+t_k, \tag{38}$$

где z_k удовлетворяет системе

$$\left.\begin{aligned} &z_{k+1}-2z_k+z_{k-1}=h^2 f_k^{(r)} \qquad (k=1,\ 2,\ \ldots,\ n-1),\\ &R_0(z)=R_n(z)=0, \end{aligned}\right\} \tag{39}$$

а t_k удовлетворяет системе

$$\left.\begin{aligned} &t_{k+1}-2t_k+t_{k-1}=0,\\ &R_0(t)=\alpha;\quad R_n(t)=\beta. \end{aligned}\right\} \tag{40}$$

Ясно, что если нам удастся найти такие z_k и t_k, то y_k, полученное по формуле (38), будет удовлетворять (36).

Отыскание t_k не встречает затруднений. Действительно, первое из равенств (40) означает, что вторая разность функции t_k равна нулю. Таким образом, t_k должна иметь вид

$$t_k = C_1 + C_2 k, \tag{41}$$

где C_1 и C_2 — постоянные, которые следует подобрать так, чтобы были выполнены граничные условия. Это накладывает следующие ограничения на C_1 и C_2:

$$\left.\begin{aligned} \alpha_0 t_0 - \alpha_1 \frac{t_1 - t_0}{h} &= \alpha_0 C_1 - \alpha_1 \frac{C_2}{h} = \alpha, \\ \beta_0 t_n + \beta_1 \frac{t_n - t_{n-1}}{h} &= \beta_0 (C_1 + C_2 n) + \frac{\beta_1 C_2}{h} = \beta. \end{aligned}\right\} \tag{42}$$

Получили систему двух линейных уравнений для определения C_1 и C_2. Чтобы эта система имела определенное решение, необходимо и достаточно требовать отличия от нуля определителя

$$\Delta = \begin{vmatrix} \alpha_0 & -\frac{\alpha_1}{h} \\ \beta_0 & \beta_0 n + \frac{\beta_1}{h} \end{vmatrix} = \frac{1}{h}[\alpha_0\beta_0(b-a) + \alpha_0\beta_1 + \alpha_1\beta_0] = \frac{\Delta_1}{h}. \tag{43}$$

Это требование мы будем считать выполненным. Тогда, решая систему (42), найдем:

$$C_1 = \frac{1}{\Delta}\begin{vmatrix} \alpha & -\frac{\alpha_1}{h} \\ \beta & \beta_0 n + \frac{\beta_1}{h} \end{vmatrix} = \frac{1}{\Delta}\left(\alpha\beta_0 n + \frac{\alpha\beta_1}{h} + \frac{\alpha_1\beta}{h}\right), \tag{44}$$

$$C_2 = \frac{1}{\Delta}\begin{vmatrix} \alpha_0 & \alpha \\ \beta_0 & \beta \end{vmatrix} = \frac{1}{\Delta}(\alpha_0\beta - \alpha\beta_0). \tag{45}$$

Для отыскания z_k, удовлетворяющих (39), подберем сначала величины $g_{ik}(i,\ k=0,\ 1,\ 2,\ \ldots,\ n)$, для которых выполнены следующие соотношения:

$$g_{i,\,k+1} - 2g_{ik} + g_{i,\,k-1} = \begin{cases} 0 & (i \neq k) \\ 1 & (i = k) \end{cases} \qquad \begin{aligned} &(i = 0,\ 1,\ 2,\ \ldots,\ n; \\ &k = 1,\ 2,\ \ldots,\ n-1), \end{aligned} \tag{46}$$

$$\alpha_0 g_{i0} - \alpha_1 \frac{g_{i1} - g_{i0}}{h} = 0, \tag{47}$$

$$\beta_0 g_{in} + \beta_1 \frac{g_{in} - g_{i,\,n-1}}{h} = 0. \tag{48}$$

Величины g_{ik} будем искать в виде

$$g_{ik} = \begin{cases} (ci+d)(ek+f) & (i \leqslant k), \\ (ck+d)(ei+f) & (i \geqslant k), \end{cases} \tag{49}$$

где c, d, e, f — постоянные, подлежащие определению. Ясно, что при $i \neq k$ для построенных таким образом величин g_{ik} будут выполнены верхние из условий (46), ибо g_{ik} являются тогда многочленами первой степени по k. При $i = k$ будем иметь:

$$g_{k,\,k+1} - 2g_{kk} + g_{kk-1} = (ck+d)(ek+f+e) - 2(ck+d)(ek+f) + \\ + (ck+d-c)(ek+f) = e(ck+d) - c(ek+f) = ed - cf. \quad (50)$$

Итак, нижнее условие (46) даст

$$ed - cf = 1. \quad (51)$$

Условие (47) повлечет за собой

$$\alpha_0 d(ei+f) - \alpha_1 \frac{c(ei+f)}{h} = \left(\alpha_0 d - \alpha \frac{c}{h}\right)(ei+f) = 0. \quad (52)$$

Вторая скобка $(ei+f)$ не может быть тождественно равна нулю, так как тогда было бы $g_{ik} \equiv 0$. Поэтому из (52) следует:

$$\alpha_0 d - \alpha_1 \frac{c}{h} = 0. \quad (53)$$

Совершенно аналогично из (48) получим:

$$\beta_0(ci+d)(en+f) + \beta_1 \frac{(ci+d)e}{h} = \left[\beta_0(en+f) + \beta_1 \frac{e}{h}\right](ci+d) = 0, \quad (54)$$

или

$$\beta_0(en+f) + \beta_1 \frac{e}{h} = 0. \quad (55)$$

Из (53) и (55) следует:

$$\alpha_0\left(\beta_0 n + \frac{\beta_1}{h}\right)ed + \frac{\alpha_1\beta_0}{h} cf = 0. \quad (56)$$

Будем рассматривать (51) и (56) как систему линейных алгебраических уравнений относительно ed и cf. Определитель этой системы равен

$$\begin{vmatrix} 1 & -1 \\ \alpha_0\left(\beta_0 n + \frac{\beta_1}{h}\right) & \frac{\alpha_1\beta_0}{h} \end{vmatrix} = \frac{\alpha_1\beta_0}{h} + \frac{\alpha_0\beta_1}{h} + \alpha_0\beta_0 n = \Delta \quad (57)$$

и по нашему предположению отличен от нуля. Таким образом,

$$ed = \frac{1}{\Delta}\begin{vmatrix} 1 & -1 \\ 0 & \frac{\alpha_1\beta_0}{h} \end{vmatrix} = \frac{\alpha_1\beta_0}{h\Delta} \quad (58)$$

и

$$cf = \frac{1}{\Delta}\begin{vmatrix} 1 & 1 \\ \alpha_0\left(\beta_0 n + \frac{\beta_1}{h}\right) & 0 \end{vmatrix} = -\frac{\alpha_0\beta_0(b-a) + \alpha_0\beta_0}{h\Delta}. \quad (59)$$

Уравнениям (53), (55), (58), (59) удовлетворяет бесчисленное множество решений. Чтобы закрепить какое-то из них, положим $c=\alpha_0$. Тогда из (53) следует $d=\frac{\alpha_1}{h}$, а (58) и (59) дают $e=\frac{\beta_0}{\Delta}$, $f=-\frac{\beta_0(b-a)+\beta_1}{h\Delta}$. Таким образом,

$$g_{ik}=\begin{cases}\frac{1}{\Delta}\left(\alpha_0 i+\frac{\alpha_1}{h}\right)\left(\beta_0 k-\beta_0 n-\frac{\beta_1}{h}\right) & (i\leqslant k),\\ \frac{1}{\Delta}\left(\alpha_0 k+\frac{\alpha_1}{h}\right)\left(\beta_0 i-\beta_0 n-\frac{\beta_1}{h}\right) & (i\geqslant k).\end{cases} \tag{60}$$

Нетрудно видеть, что g_{ik} является симметрической функцией своих индексов, т. е. что $g_{ik}=g_{ki}$. Таким образом, соотношения (46), (47) и (48) будут справедливы при замене i на k, и наоборот.

Рассмотрим

$$z_k=h^2\sum_{i=1}^{n-1}g_{ik}f_i^{(r)} \tag{61}$$

Очевидно,

$$z_{k+1}-2z_k+z_{k-1}=h^2\sum_{i=1}^{n-1}(g_{i,\,k+1}-2g_{ik}+g_{i,\,k-1})f_i^{(r)}=h^2f_k^{(r)}, \tag{62}$$

т. е. z_k, определенные (61), удовлетворяют первому из равенств (39). Далее,

$$\left.\begin{aligned}R_0(z)&=h^2\sum_{i=1}^{n-1}\left[\alpha_0 g_{i0}-\alpha_1\frac{g_{i1}-g_{i0}}{h}\right]f_i^{(r)}=0,\\ R_n(z)&=h^2\sum_{i=1}^{n-1}\left[\beta_0 g_{in}+\beta_1\frac{g_{in}-g_{i,\,n-1}}{h}\right]f_i^{(r)}=0,\end{aligned}\right\} \tag{63}$$

т. е. и второе условие (39) выполнено.

Окончательно находим:

$$y_k^{(r+1)}=\frac{h}{\Delta}\left[\alpha\beta_0(b-a)+\alpha\beta_1+\alpha_1\beta\right]+\frac{k}{\Delta}(\alpha_0\beta-\alpha\beta_0)+h^2\sum_{i=1}^{n-1}g_{ik}f_i^{(r)}. \tag{64}$$

Мы можем раз и навсегда вычислить g_{ik} и единообразным процессом находить последовательные приближения.

Перейдем теперь к исследованию сходимости этого процесса. Прежде всего произведем некоторые оценки. Оценим $h^2\sum_{i=1}^{n-1}g_{ik}f_i^{(r)}$. Имеем:

$$\left|h^2\sum_{i=1}^{n-1}g_{ik}f_i^{(r)}\right|\leqslant h^2\sum_{i=1}^{n-1}|g_{ik}||f_i^{(r)}|\leqslant h^2\max_i|f_i^{(r)}|\sum_{i=1}^{n-1}|g_{ik}|. \tag{65}$$

Чтобы закончить оценку, нужно вычислить $\sum_{i=1}^{n-1} |g_{ik}|$. Для этого заметим, что все g_{ik} имеют отрицательный знак. Поэтому

$$\sum_{i=1}^{n-1} |g_{ik}| = -\sum_{i=1}^{n-1} g_{ik}. \tag{66}$$

Таким образом, $\sum_{i=1}^{n-1} |g_{ik}|$ есть решение уравнения

$$\bar{z}_{k+1} - 2\bar{z}_k + \bar{z}_{k-1} = -1, \tag{67}$$

удовлетворяющее граничным условиям

$$R_0(\bar{z}) = R_n(\bar{z}) = 0. \tag{68}$$

Но из (67) следует, что

$$\bar{z}_k = -\frac{1}{2}(k^2 + uk + v), \tag{69}$$

где u и v — некоторые постоянные. Потребуем выполнения условий (68). Условие $R_0(\bar{z}) = 0$ даст

$$-\frac{1}{2}\alpha_0 v + \frac{1}{2}\alpha_1 \frac{1+u}{h} = 0, \tag{70}$$

а условие $R_n(\bar{z}) = 0$:

$$-\frac{1}{2}\beta_0(n^2 + un + v) - \frac{1}{2}\beta_1 \frac{n^2 + un - (n-1)^2 - u(n-1)}{h} = 0. \tag{71}$$

Итак, u и v удовлетворяют системе уравнений

$$\left.\begin{aligned} &\alpha_0 hv - \alpha_1 u = \alpha_1, \\ &\beta_0 hv + (\beta_0 hn + \beta_1)u = -\beta_1(2n-1) - \beta_0 hn^2. \end{aligned}\right\} \tag{72}$$

Определитель этой системы равен

$$\begin{vmatrix} \alpha_0 h & -\alpha_1 \\ \beta_0 h & \beta_0 hn + \beta_1 \end{vmatrix} = h^2 \begin{vmatrix} \alpha_0 & -\frac{\alpha_1}{h} \\ \beta_0 & \beta_0 n + \frac{\beta_1}{h} \end{vmatrix} = h^2\Delta \neq 0. \tag{73}$$

Решая систему (72), найдем:

$$v = \frac{\alpha_1}{h^2\Delta}\left[\beta_0 hn(n+1) + 2\beta_1 n\right], \tag{74}$$

$$u = -\frac{1}{h\Delta}\left[nh\Delta + (n-1)(\alpha_0\beta_1 - \alpha_1\beta_0)\right]. \tag{75}$$

Квадратный трехчлен $-\frac{1}{2}(k^2+uk+v)$ примет наибольшее значение при $k=-\frac{u}{2}$, и это значение равно

$$\frac{1}{8}(u^2-4v). \tag{76}$$

Подставляя сюда вместо u и v их значения, найдем:

$$\frac{1}{8}(u^2-4v)=\frac{1}{8h^2\Delta^2}[h^2n^2\Delta^2+2(\alpha_0\beta_1-\alpha_1\beta_0)hn(n-1)\Delta+$$

$$+(n-1)^2(\alpha_0\beta_1-\alpha_1\beta_0)^2-4\alpha_1\beta_0n(n+1)h\Delta-8\alpha_1\beta_1n\Delta]\leqslant$$

$$\leqslant\frac{(b-a)^2}{8h^2}+\frac{n^2(\alpha_0\beta_1-\alpha_1\beta_0)^2}{8h^2\Delta^2}+\frac{2n^2h(\alpha_0\beta_1+\alpha_1\beta_0)}{8h^2\Delta}+\frac{\alpha_1\beta_1n}{h^2\Delta}=\frac{(b-a)^2}{8h^2}+$$

$$+\frac{(b-a)^2(\alpha_0\beta_1-\alpha_1\beta_0)^2}{8h^2\Delta_1^2}+\frac{(b-a)^2(\alpha_0\beta_1+\alpha_1\beta_0)}{4h^2\Delta_1}+\frac{(b-a)\alpha_1\beta_1}{8h^2\Delta_1}, \tag{77}$$

где Δ_1 определено в (43). Таким образом,

$$\left|h\sum_{i=1}^{n-1}g_{ik}f_i^{(r)}\right|\leqslant(b-a)^2\lambda_1^2\max_i|f_i^{(r)}|, \tag{78}$$

где

$$\lambda_1=\frac{1}{8}+\frac{1}{4\Delta_1}\left[\alpha_0\beta_1+\alpha_1\beta_0+\frac{\alpha_1\beta_1}{2(b-a)}+\frac{(\alpha_0\beta_1-\alpha_1\beta_0)^2}{2\Delta_1}\right]. \tag{79}$$

Нам потребуется еще оценка $h\sum_{i=1}^{n-1}\frac{g_{i,\,k+1}-g_{i,\,k-1}}{2}f_i^{(r)}$. Имеем:

$$\left|h\sum_{i=1}^{n-1}\frac{g_{i,\,k+1}-g_{i,\,k-1}}{2}f_i^{(r)}\right|\leqslant h\max_i|f_i^{(r)}|\sum_{i=1}^{n-1}\left|\frac{g_{i,\,k+1}-g_{i,\,k-1}}{2}\right|. \tag{80}$$

Далее,

$$\sum_{i=1}^{n-1}\left|\frac{g_{i,\,k+1}-g_{i,\,k-1}}{2}\right|=\frac{1}{2}\left|\sum_{i=1}^{n-1}|g_{i,\,k+1}|-\sum_{i=1}^{n-1}|g_{i,\,k-1}|\right|. \tag{81}$$

Но

$$\sum_{i=1}^{n-1}|g_{i,\,k+1}|=-\frac{1}{2}[(k+1)^2+u(k+1)+v] \tag{82}$$

и

$$\sum_{i=1}^{n-1}|g_{i,\,k-1}|=-\frac{1}{2}[(k-1)^2+u(k-1)+v]. \tag{83}$$

Таким образом,

$$\frac{1}{2}\left\{\sum_{i=1}^{n-1}|g_{i,\,k+1}|-\sum_{i=1}^{n-1}|g_{i,\,k-1}|\right\}=-\frac{1}{2}(2k+u). \tag{84}$$

Отсюда

$$\max_k\sum_{i=1}^{n-1}\left|\frac{g_{i,\,k+1}-g_{i,\,k-1}}{2}\right|=\frac{1}{2}\left|n-2-\frac{n-1}{h\Delta}(\alpha_0\beta_1-\alpha_1\beta_0)\right|=$$

$$=\frac{1}{2\Delta_1}|(n-2)\Delta_1-(n-1)(\alpha_0\beta_1-\alpha_1\beta_0)|\leqslant$$

$$\leqslant\frac{n}{2\Delta_1}|\alpha_0\beta_0(b-a)+\alpha_0\beta_1+\alpha_1\beta_0+|\alpha_0\beta_1-\alpha_1\beta_0||=$$

$$=\frac{n}{2\Delta_1}[\alpha_0\beta_0(b-a)+2\max(\alpha_0\beta_1,\ \alpha_1\beta_0)]. \tag{85}$$

Окончательно получаем:

$$\left|h\sum_{i=1}^{n-1}\frac{g_{i,\,k+1}-g_{i,\,k-1}}{2}f_i^{(r)}\right|\leqslant(b-a)\lambda_2\max|f_i^{(r)}|, \tag{86}$$

где

$$\lambda_2=\frac{1}{2\Delta_1}[\alpha_0\beta_0(b-a)+2\max(\alpha_0\beta_1,\ \alpha_1\beta_0)]. \tag{87}$$

Заметим, что в оценках (78) и (86) величины $f_i^{(r)}$ не обязательно должны быть определены функцией $f(x,\ y,\ z)$, а могут быть произвольными заданными числами.

Пусть $f(x,\ y,\ z)$ в рассматриваемой области G удовлетворяет условию Липшица по y и z:

$$|f(x,\ y,\ z)-f(x,\ \bar{y},\ \bar{z})|\leqslant L_1|y-\bar{y}|+L_2|z-\bar{z}|. \tag{88}$$

Рассмотрим множество R всевозможных совокупностей $n+1$ чисел $y_0,\ y_1,\ \ldots,\ y_n$. Будем обозначать такие совокупности $\{y_k\}$. Множество R можно сделать метрическим пространством, если определить расстояние между двумя совокупностями $\{y_k\}$ и $\{z_k\}$ как

$$\rho(\{y_k\},\ \{z_k\})=L_1\max_{0\leqslant k\leqslant n}|y_k-z_k|+$$

$$+L_2\max_{1\leqslant k\leqslant n-1}\left|\frac{y_{k+1}-y_{k-1}}{2h}-\frac{z_{k+1}-z_{k-1}}{2h}\right|. \tag{89}$$

Нетрудно проверить, что все аксиомы метрического пространства при этом будут выполнены. В дальнейшем будем рассматривать только такие совокупности $\{y_k\}$, для которых

$$\left(x_k,\ y_k,\ \frac{y_{k+1}-y_{k-1}}{2h}\right)\in G\qquad(k=1,\ 2,\ \ldots,\ n-1). \tag{90}$$

Для каждой такой совокупности формула

$$z_k = \frac{h}{\Delta}[\alpha\beta_0(b-a)+\alpha\beta_1+\alpha_1\beta] + \frac{k}{\Delta}[\alpha_0\beta-\alpha\beta_0]+$$
$$+h^2\sum_{i=1}^{n-1} g_{ik} f\left(x_i,\ y_i,\ \frac{y_{i+1}-y_{i-1}}{2h}\right) \qquad (k=0,\ 1,\ 2,\ \ldots,\ n) \qquad (91)$$

определит отображение

$$\{z_k\} = A(\{y_k\}) \qquad (92)$$

элемента $\{y_k\}$, принадлежащего R, в элемент $\{z_k\}$, принадлежащий R. Если даны две такие совокупности $\{y_k\}$ и $\{\bar{y}_k\}$, то в силу (78) получим:

$$|z_k-\bar{z}_k| = h^2\left|\sum_{i=1}^{n-1} g_{ik}\left[f\left(x_i,\ y_i,\ \frac{y_{i+1}-y_{i-1}}{2h}\right)-f\left(x_i,\ \bar{y}_i,\ \frac{\bar{y}_{i+1}-\bar{y}_{i-1}}{2h}\right)\right]\right| \leqslant$$
$$\leqslant h^2\sum_{i=1}^{n-1}|g_{ik}|\left\{L_1|y_i-\bar{y}_i|+L_2\left|\frac{y_{i+1}-y_{i-1}}{2h}-\frac{\bar{y}_{i+1}-\bar{y}_{i-1}}{2h}\right|\right\}\leqslant$$
$$\leqslant (b-a)^2\lambda_1\left\{L_1\max_{0\leqslant i\leqslant n}|y_i-\bar{y}_i|+\right.$$
$$\left.+L_2\max_{0<i<n}\left|\frac{y_{i+1}-y_{i-1}}{2h}-\frac{\bar{y}_{i+1}-\bar{y}_{i-1}}{2h}\right|\right\}\leqslant (b-a)^2\lambda_1\rho(\{y_k\},\ \{\bar{y}_k\}). \qquad (93)$$

Аналогично, используя (86), найдем:

$$\left|\frac{z_{k+1}-z_{k-1}}{2h}-\frac{\bar{z}_{k+1}-\bar{z}_{k-1}}{2h}\right|\leqslant (b-a)\lambda_2\rho(\{y_k\},\ \{\bar{y}_k\}). \qquad (94)$$

Таким образом,

$$\rho(\{z_k\},\ \{\bar{z}_k\})\leqslant[(b-a)^2L_1\lambda_1+(b-a)L_2\lambda_2]\rho(\{y_k\},\ \{\bar{y}_k\}). \qquad (95)$$

Обозначим

$$(b-a)^2L_1\lambda_1+(b-a)L_2\lambda_2=\gamma. \qquad (96)$$

Если $\gamma<1$, то (95) показывает, что отображение A сжатое. В дальнейшем мы будем предполагать, что это условие выполнено.

Для применения принципа сжатых отображений требуется еще выбрать начальное приближение $\{y_k^{(0)}\}$ и область G такими, чтобы все последующие приближения не выходили из области G. Для этого достаточно потребовать, чтобы для $\{y_k^{(0)}\}$ было выполнено условие (90) и чтобы к области G принадлежали все точки $(x_k,\ y_k,\ z_k)$, для которых

$$|y_k-y_k^{(1)}|\leqslant(b-a)^2\lambda_1\frac{\rho(\{y_k^{(0)}\},\ \{y_k^{(1)}\})}{1-\gamma}, \qquad (97)$$

$$\left|z_k-\frac{y_{k+1}^{(1)}-y_{k-1}^{(1)}}{2h}\right|\leqslant(b-a)\lambda_2\frac{\rho(\{y_k^{(0)}\},\ \{y_k^{(1)}\})}{1-\gamma}. \qquad (98)$$

Действительно, в этом случае неравенства (93) и (94) обеспечивают принадлежность всех последующих приближений к G.

Итак, при всех вышеуказанных предположениях будет применим принцип сжатых отображений, и следовательно, уравнения (34) и (35) имеют решения, удовлетворяющие неравенствам (97) и (98), и эти решения могут быть получены методом последовательных приближений (36) при подходящем выборе начальных приближений $\{y_k^{(0)}\}$.

Нам остается исследовать вопрос о сходимости конечноразностного решения к точному решению краевой задачи и об оценке отклонения этих решений. Сделаем еще два предположения относительно функции $f(x, y, z)$:

1. Функция $f(x, y, z)$ непрерывна в области G вместе со всеми своими производными до второго порядка.

2. Краевая задача имеет решение $\varphi(x)$, лежащее в G. Используемое разностное уравнение также имеет решение, принадлежащее G.

Условимся о некоторых обозначениях. Точное решение краевой задачи (31) и (32) будем обозначать через $\varphi(x)$ и его значение в узле x_k через φ_k. Пусть $f[x, \varphi(x), \varphi'(x)] = F(x)$ и $F(x_k) = F_k$. Через M_k обозначим $\max|\varphi^{(k)}(x)|$ при $x \in [a, b]$. Приближенное решение краевой задачи, полученное приведенным выше способом, будем обозначать через y_k ($k = 0, 1, 2, \ldots, n$). Наконец, обозначим $f\left(x_k, y_k, \frac{y_{k+1} - y_{k-1}}{2h}\right) = f_k$; $\varepsilon_k = \varphi_k - y_k$.

Рассмотрим выражение

$$\varphi_{k+1} - 2\varphi_k + \varphi_{k-1} - h^2\varphi_k'' \qquad (k = 1, 2, \ldots, n-1). \tag{99}$$

Разлагая φ_{k+1} и φ_{k-1} по формуле Тейлора относительно точки x_k до членов, содержащих производные четвертого порядка, получим:

$$\varphi_{k+1} - 2\varphi_k + \varphi_{k-1} = h^2F_k + h^4\rho_k \qquad (k = 1, 2, \ldots, n-1), \tag{100}$$

где

$$|\rho_k| \leqslant \frac{1}{12} M_4. \tag{101}$$

Аналогично найдем:

$$\left.\begin{aligned} \alpha_0\varphi_0 - \alpha_1 \frac{\varphi_1 - \varphi_0}{h} &= \alpha - \alpha_1 h\rho_0, \\ \beta_0\varphi_n + \beta_1 \frac{\varphi_n - \varphi_{n-1}}{h} &= \beta + \beta_1 h\rho_n, \end{aligned}\right\} \tag{102}$$

где

$$|\rho_0| \leqslant \frac{1}{2} M_2; \quad |\rho_n| \leqslant \frac{1}{2} M_2. \tag{103}$$

Так как y_k удовлетворяет уравнению

$$y_{k+1} - 2y_k + y_{k-1} = h^2 f_k \qquad (k = 1, 2, \ldots, n-1) \tag{104}$$

и граничным условиям (35), то, вычитая, получим.

$$\left.\begin{aligned} &\varepsilon_{k+1}-2\varepsilon_k+\varepsilon_{k-1}=h^2(F_k-f_k)+h^4\rho_k \quad (k=1,\ 2,\ \ldots,\ n-1),\\ &\alpha_0\varepsilon_0-\alpha_1\frac{\varepsilon_1-\varepsilon_0}{h}=-\alpha_1 h\rho_0,\\ &\beta_0\varepsilon_n+\beta_1\frac{\varepsilon_n-\varepsilon_{n-1}}{h}=\beta_1 h\rho_n. \end{aligned}\right\} \tag{105}$$

Повторяя рассуждения, при помощи которых мы нашли $y_k^{(r+1)}$ (формула (64)), получим:

$$\varepsilon_k=\frac{h^3}{\Delta_1}[\alpha_1\beta_1\rho_n-\alpha_1\beta_0(b-a)\rho_0-\alpha_1\beta_1\rho_0]+$$

$$+\frac{kh^2}{\Delta_1}(\alpha_0\beta_1\rho_n+\alpha_1\beta_0\rho_0)+h^4\sum_{i=1}^{n-1}g_{ik}[(F_i-f_i)+h^2\rho_i]. \tag{106}$$

Отсюда сразу же следует сходимость. Стремление к нулю при $h\to 0$ первых двух членов правой части видно непосредственно. Стремление к нулю последнего члена следует из неравенства (78).

Рассмотрим теперь

$$\varphi_k'-\frac{y_{k+1}-y_{k-1}}{2h} \qquad (i=1,\ 2,\ \ldots,\ n-1). \tag{107}$$

Это выражение можно записать в виде

$$\varphi_k'-\frac{\varphi_{k+1}-\varphi_{k-1}}{2h}+\frac{\varepsilon_{k+1}-\varepsilon_{k-1}}{2h}. \tag{108}$$

Снова применяя разложение по формуле Тейлора, получим:

$$\varphi_k'-\frac{y_{k+1}-y_{k-1}}{2h}=h^2\rho_k+\frac{\varepsilon_{k+1}-\varepsilon_{k-1}}{2h} \quad (k=1,\ 2,\ \ldots,\ n-1), \tag{109}$$

где

$$|\rho_k|\leqslant\frac{1}{6}M_3. \tag{110}$$

Используя (106), найдем:

$$\frac{\varepsilon_{k+1}-\varepsilon_{k-1}}{2h}=\frac{h}{\Delta_1}(\alpha_0\beta_1\rho_n+\alpha_1\beta_0\rho_0)+$$

$$+h^3\sum_{i=1}^{n-1}[g_{i,\,k+1}-g_{i,\,k-1}][(F_i-f_i)+h^2\rho_i]. \tag{111}$$

Если вспомнить неравенство (85), то из (109) и (111) видно, что $\frac{y_{k+1}-y_{k-1}}{2h}$ стремится к φ_k' при $h\to 0$ $(k=1,\ 2,\ \ldots,\ n-1)$.

Равенство (106) можно использовать и для оценки $|\varepsilon_k|$. Обозначая через N верхнюю границу функции $|f(x, y, z)|$ в G из (106), получим:

$$|\varepsilon_k| \leqslant \frac{h^3}{2\Delta_1} M_2 [2\alpha_1\beta_1 + \alpha\beta(b-a)] + \frac{(b-a)hM_3}{2\Delta_1}(\alpha_1\beta_0 + \alpha_0\beta_1) + \\ + h^2(b-a)\lambda_1^2 2N + \frac{h^2}{12}(b-a)^2\lambda_1^2 M_4. \quad (112)$$

Мы не стремились дать здесь наилучшую оценку. Как и все оценки такого типа, она мало эффективна, а ее получение требовало бы громоздких рассуждений.

Конечноразностные методы могут быть использованы и для решения краевых задач для обыкновенных дифференциальных уравнений более высокого порядка. Они могут быть использованы и для решения краевых задач с другими не рассмотренными здесь типами граничных условий. При этом обычно не возникает никаких принципиальных затруднений при составлении конечноразностного аналога заданной задачи. Но, естественно, возникают трудности при доказательстве сходимости и оценке погрешности.

§ 9. Метод прогонки [1])

Этот метод разработан в Математическом институте им. В. А. Стеклова АН СССР. Проиллюстрируем его на примере линейного дифференциального уравнения второго порядка:

$$y'' = p(x)y + q(x), \quad (1)$$

где $p(x)$ и $q(x)$ — непрерывные функции, $p(x) > 0$. Пусть граничные условия имеют вид

$$y'(a) = \alpha_0 y(a) + \alpha_1, \quad (2)$$

$$y'(b) = \beta_0 y(b) + \beta_1. \quad (3)$$

Задачу (1), (2), (3) можно, как мы видели, решать конечноразностными методами. Можно было бы применить и следующий прием. Задаемся произвольными начальными данными

$$y(a) = \alpha; \quad y'(a) = \alpha_0\alpha + \alpha_1, \quad (4)$$

лишь бы было выполнено граничное условие (2). Находим решение задачи (1), (4). При этом, возможно, придется применить один из численных методов, приспособленных для решения задачи Коши.

---

[1]) При написании данного параграфа использована рукопись неопубликованной статьи И. М. Гельфанда и Локуциевского, любезно предоставленная авторами в наше распоряжение.

Обозначим полученное решение через $y_1(x)$. Тогда общее решение уравнения (1) можно записать в виде

$$y(x) = C_1 z_1(x) + C_2 z_2(x) + y_1(x), \tag{5}$$

где $z_1(x)$ и $z_2(x)$ — линейно независимые решения уравнения

$$z''(x) = p(x)\, z(x). \tag{6}$$

Так как для искомого решения $y(x)$ должно быть выполнено граничное условие (2), то должно иметь место

$$C_1 z_1'(a) + C_2 z_2'(a) + y_1'(a) = C_1 \alpha_0 z_1(a) + C_2 \alpha_0 z_2(a) + \alpha_0 y_1(a) + \alpha_1, \tag{7}$$

или

$$C_1 z_1'(a) + C_2 z_2'(a) = \alpha_0 \left[C_1 z_1(a) + C_2 z_2(a)\right]. \tag{8}$$

Таким образом, нам надо разыскивать все решения (6), удовлетворяющее условию

$$z'(a) = \alpha_0 z(a). \tag{9}$$

Совокупность таких решений образует однопараметрическое семейство. Чтобы получить его, достаточно задать каким-либо образом $z(a) = \gamma \neq 0$ и найти решение $\bar{z}(x)$ уравнения (6), удовлетворяющее условиям:

$$\bar{z}(a) = \gamma; \quad \bar{z}'(a) = \alpha_0 \gamma. \tag{10}$$

При этом

$$y(x) = y_1(x) + C\bar{z}(x), \tag{11}$$

где C подбирается так, что для $y(x)$ выполнено (3). Это дает

$$C = \frac{\beta_1 - y_1'(b) + \beta_0 y_1(b)}{\bar{z}'(b) - \beta_0 \bar{z}(b)}. \tag{12}$$

Если, как мы предполагаем, задача (1), (2), (3) имеет единственное решение, то C должно определиться однозначно.

Теоретически изложенный метод идеален по своей простоте. Однако он может привести к большим вычислительным погрешностям. Для того чтобы показать это, исследуем поведение решения $\bar{z}(x)$ уравнения (6). Дифференцируя $\bar{z}\bar{z}'$, получим:

$$\frac{d(\bar{z}\bar{z}')}{dx} = \bar{z}'^2 + \bar{z}\bar{z}'' = p\bar{z}^2 + \bar{z}'^2 > 0. \tag{13}$$

Отсюда

$$\bar{z}(x)\bar{z}'(x) = \int_a^x \left\{p(\xi)\bar{z}^2(\xi) + \bar{z}'^2(\xi)\right\} d\xi + \bar{z}(a)\bar{z}'(a), \tag{14}$$

$$\bar{z}^2(x) = 2\int_a^x d\xi \int_a^\xi \left\{p(\eta)\bar{z}^2(\eta) + \bar{z}'^2(\eta)\right\} d\eta + 2\bar{z}(a)\bar{z}'(a)(x-a) + \bar{z}^2(a). \tag{15}$$

В нашем случае $\bar{z}(a)\bar{z}'(a)=\alpha_0\bar{z}^2(a)$. Как и ранее, будем предполагать, что $\alpha_0>0$. При этом формула (15) показывает, что $\bar{z}(x)$ будет расти по абсолютной величине вместе с возрастанием x. Следовательно, $\bar{z}(x)$ может быть очень велика при $x=b$, особенно если нижняя граница $p(x)$ большая. Вследствие этого для получения $y(x)$ по формуле (11) с достаточной точностью нужно будет сохранять в $y_1(x)$ и $\bar{z}(x)$ большое количество разрядов.

Метод прогонки и придуман для того, чтобы избежать этих трудностей. Суть метода состоит в следующем. Формула (11) показывает, что совокупность решений дифференциального уравнения (1), удовлетворяющих граничному условию (2), есть семейство, зависящее от одного параметра. Будем разыскивать линейное дифференциальное уравнение первого порядка вида

$$y'(x)=\alpha_0(x)y(x)+\alpha_1(x) \tag{16}$$

такое, что каждое решение (11) принадлежит к числу решений (16). Естественно, при $x=a$ мы должны получить условие (2), и поэтому

$$\alpha_0(a)=\alpha_0,\quad \alpha_1(a)=\alpha_1. \tag{17}$$

Так как (11) должна удовлетворять (16), то

$$y_1'+C\bar{z}'\equiv\alpha_0(x)y_1+C\alpha_0(x)\bar{z}+\alpha_1(x). \tag{18}$$

Это тождество, выполняющееся при любом значении C. Поэтому

$$\bar{z}'=\alpha_0(x)\bar{z};\quad y_1'=\alpha_0(x)y_1+\alpha_1(x). \tag{19}$$

Дифференцируя первое из этих равенств и используя (6), получим:

$$\bar{z}''=p(x)\bar{z}=\alpha_0'(x)\bar{z}+\alpha_0(x)\bar{z}'=\left[\alpha_0'(x)+\alpha_0^2(x)\right]\bar{z}, \tag{20}$$

или

$$\alpha_0'(x)+\alpha_0^2(x)=p(x). \tag{21}$$

Точно так же, дифференцируя второе из равенств (19) и используя (1), получим:

$$y_1''=p(x)y_1+q(x)=\alpha_0'(x)y_1+\alpha_0(x)y_1'+\alpha_1'(x)=$$
$$=\left[\alpha_0'(x)+\alpha_0^2(x)\right]y_1+\alpha_0(x)\alpha_1(x)+\alpha_1'(x), \tag{22}$$

или

$$\alpha_1'(x)+\alpha_0(x)\alpha_1(x)=q(x). \tag{23}$$

Таким образом, наша задача свелась к отысканию функций $\alpha_0(x)$ и $\alpha_1(x)$, удовлетворяющих системе дифференциальных уравнений первого порядка

$$\left.\begin{aligned}\alpha_0'(x)+\alpha_0^2(x)&=p(x),\\ \alpha_1'(x)+\alpha_0(x)\alpha_1(x)&=q(x)\end{aligned}\right\} \tag{24}$$

и начальным данным (17). Сначала мы интегрируем первое из уравнений (24), а затем второе. Найдя $\alpha_0(x)$ и $\alpha_1(x)$, мы можем получить

$$y'(b) = \alpha_0(b)\, y(b) + \alpha_1(b). \tag{25}$$

Этим самым мы совершили *прямую прогонку*, перегнав граничное условие (2) с левого конца на правый. Равенства (25) и (3) рассматриваем как систему уравнений для определения $y(b)$ и $y'(b)$. Уравнения (25) и (3) могут совпадать, и тогда краевая задача имеет бесчисленное множество решений, представляемых формулой (11). Они могут быть несовместны и тогда краевая задача не имеет решения. Если же система (25), (3) имеет единственное решение, то на правом конце мы получим данные Коши для уравнения (1). Но лучше использовать не уравнение (1), а уравнение (16), находя его решение на отрезке $[a, b]$, принимающее при $x = b$ полученное нами значение $y(b)$. Этот процесс называют *обратной прогонкой*.

Исследуем теперь поведение решений уравнений (21) и (23) при прямой прогонке и уравнения (16) при обратной прогонке. Заметим, что уравнение (21) получается из уравнения (6), если произвести замену переменных

$$z = e^{\int_a^x \alpha\,(x)\,dx}. \tag{26}$$

Действительно,

$$z' = \alpha_0(x)\, z; \quad z'' = \alpha_0'(x)\, z + \alpha_0^2(x)\, z \tag{27}$$

и подстановка этих выражений в (6) даст (21). Начальные условия для z, дающие нужное нам решение $\alpha_0(x)$, будут: $z(a) = 1$, $z'(a) = \alpha_0$. При этом, $\alpha_0(x) = \frac{z'}{z}$ будет представляться в виде отношения выражений (14) и (15), которые положительны. Если z быстро растет, то $\alpha_0(x)$ будет быстро убывать. Уравнение (23) для определения $\alpha_1(x)$ линейно и коэффициент при $\alpha_1(x)$ положителен. Следовательно, его решение, имеющее вид

$$\alpha_1(x) = \alpha_1 e^{-\int_a^x \alpha_1(x)\,dx} + e^{-\int_a^x \alpha_0(x)\,dx} \int_a^x e^{\int_a^\xi \alpha_0(\eta)\,d\eta} q(\xi)\,d\xi, \tag{28}$$

благодаря наличию множителя $e^{-\int_a^x \alpha_1 d(x)\,dx}$ не будет быстро расти. При обратной прогонке мы решаем уравнение (16) при заданном значении y на правом конце. Это также не даст быстрого роста.

§ 10. Решение краевых задач для обыкновенных дифференциальных уравнений вариационными методами

В настоящем параграфе мы рассмотрим вариационные методы решения краевых задач для обыкновенных дифференциальных уравнений, позволяющие получить приближенное решение краевой задачи в аналитической форме. Свое название эти методы получили потому, что их первое применение было связано с заменой краевой задачи для дифференциального уравнения некоторой вариационной задачей. Так, решение краевой задачи

$$\frac{d}{dx}[p(x)y'] - q(x)y = f(x), \tag{1}$$

$$y(a) = \alpha, \quad y(b) = \beta \tag{2}$$

заменяется задачей об отыскании функции $y(x)$, удовлетворяющей условиям (2) и обращающей в минимум функционал

$$J = \int_a^b [p(x)\,y'^2 + q(x)y^2 + 2f(x)y]\,dx. \tag{3}$$

Если $p(x) > 0$ и непрерывно дифференцируема, а $q(x)$ и $f(x)$ непрерывны и $q(x) \geqslant 0$, то решение краевой задачи (1) и (2) существует и единственно в классе всех непрерывно дифференцируемых функций, удовлетворяющих (2), и обращает J в минимум.

Более общая краевая задача (1) и

$$\left.\begin{aligned} \alpha_0 y(a) + \alpha_1 y'(a) &= \alpha, \\ \beta_0 y(b) + \beta_1 y'(b) &= \beta \end{aligned}\right\} \tag{4}$$

при $\alpha_1 \neq 0$, $\beta_1 \neq 0$ может быть заменена задачей об отыскании минимума функционала

$$J = \int_a^b [py'^2 + qy^2 + 2fy]\,dx + \frac{p(a)}{\alpha_1}[-\alpha_0 y^2(a) + 2\alpha y(a)] + \\ + \frac{p(b)}{\beta_1}[\beta_0 y^2(b) - 2\beta y(b)] \tag{5}$$

в классе всех непрерывно дифференцируемых функций (выполнения граничных условий можно не требовать).

И в том и в другом случае дифференциальное уравнение является уравнением Эйлера для вариационной задачи. Идя по этому пути, можно найти и другие функционалы, для которых минимум достигается после подстановки в них решения краевой задачи. При этом, если функционал имеет вид

$$J = \int_a^b F(x, y, y', \ldots, y^{(n)})\,dx, \tag{6}$$

то дифференциальное уравнение соответствующей краевой задачи будет представляться в виде

$$F'_y - \frac{d}{dx}F'_{y'} + \frac{d^2}{dx^2}F'_{y''} - \ldots + (-1)^n \frac{d^n}{dx^n}F'_{y^{(n)}} = 0. \qquad (7)$$

Однако вариационное исчисление не является единственным путем для получения функционалов, принимающих минимальное значение при подстановке в них решения краевых задач. Можно, например, решая краевую задачу для дифференциального уравнения

$$L(y) = f, \qquad (8)$$

рассматривать функционал

$$J = \int_a^b [L(y) - f]^2 dx \qquad (9)$$

в классе всех функций, удовлетворяющих граничным условиям и обладающих достаточным количеством непрерывных производных. Можно заменить функционал (9) более общим функционалом

$$J = \int_a^b P(x)[L(y) - f]^2 dx, \qquad (10)$$

где $P(x)$ — некоторая положительная весовая функция. Ясно, что функционалы (9) и (10) принимают наименьшее значение, равное нулю, при подстановке в них решения краевой задачи. Такой способ получения функционалов, минимизирующихся решением краевой задачи, иногда называют методом наименьших квадратов.

1. Вариационные методы решения операторных уравнений в гильбертовом пространстве. Рассмотрим теперь указанные способы с более общей точки зрения. Пусть H — некоторое гильбертово пространство и A — аддитивный оператор, определенный на каком-то всюду плотном в H множестве H_A. Назовем оператор A *положительным*, если для любого элемента $y \in H_A$ имеет место

$$(Ay, y) \geqslant 0, \qquad (11)$$

причем равенство нулю возможно только при $y = 0$. Оператор называется *симметричным*, если для любых двух элементов y, z, принадлежащих H_A, имеет место

$$(Ay, z) = (y, Az). \qquad (12)$$

Заметим, что *для симметричности оператора A в комплексном гильбертовом пространстве необходимо и достаточно, чтобы*

скалярное произведение (Ay, y) *было действительным числом для любого* $y \in H_A$.

Действительно, если оператор A симметричен, то

$$(Ay, y) = (y, Ay) = \overline{(Ay, y)}, \tag{13}$$

т. е. (Ay, y) — действительное число. Для доказательства достаточности рассмотрим тождество

$$4(Ay, z) = (A(y+z), y+z) - (A(y-z), y-z) + \\ + i[(A(y+iz), y+iz) - (A(y-iz), y-iz)]. \tag{14}$$

Поменяв здесь местами y и z, получим:

$$4(Az, y) = (A(y+z), y+z) - (A(z-y), z-y) + \\ + i[(A(z+iy), z+iy) - (A(z-iy), z-iy)]. \tag{15}$$

Если заменить все входящие в (15) величины на комплексно-сопряженные и воспользоваться свойствами скалярного произведения и предположенной действительностью (Ay, y), то найдем:

$$4(y, Az) = (A(y+z), y+z) - (A(y-z), y-z) + \\ + i[(A(y+iz)\ y+iz) - (A(y-iz), y-iz)]. \tag{16}$$

Из (14) и (16) получаем:

$$(Ay, z) = (y, Az), \tag{17}$$

т. е. оператор A симметричен.

В силу только что доказанного *положительный оператор в комплексном гильбертовом пространстве симметричен.*

В дальнейшем мы будем предполагать, что наш оператор A положителен, а если гильбертово пространство H действительно, то дополнительно предположим, что он симметричен.

Рассмотрим уравнение

$$Ay = f, \tag{18}$$

где $f \in H$ — заданный элемент и $y \in H_A$ — искомый элемент. Это уравнение не может иметь более одного решения. Если бы имелось два решения y_1 и y_2 $(y_1 \neq y_2)$, то мы имели бы

$$A(y_1 - y_2) = 0 \tag{19}$$

и

$$(A(y_1 - y_2), \quad y_1 - y_2) = 0, \tag{20}$$

что невозможно, так как A положительный оператор и $y_1 - y_2 \neq 0$.

Далее, *если уравнение* (18) *имеет некоторое решение* y_1, *то оно дает функционалу*

$$J(y) = (Ay, y) - (y, f) - (f, y) \tag{21}$$

минимальное значение. Действительно $J(y)$ принимает только действительные значения. Возьмем произвольный элемент $z \in H_A$. Положим $z = y_1 + t$. Тогда

$$J(z) = (Az, z) - (z, f) - (f, z) = (A(y_1+t), y_1+t) - (y_1+t, f) - (f, y_1+t) =$$
$$= J(y_1) + (At, y_1) + (Ay_1, t) + (At, t) - (t, f) - (f, t) =$$
$$= J(y_1) + (t, Ay_1 - f) + (Ay_1 - f, t) + (At, t), \tag{22}$$

и так как $Ay_1 - f = 0$ и $(At, t) > 0$, то

$$J(z) = J(y_1) + (At, t) > J(y_1). \tag{23}$$

Утверждение доказано.

Обратно, *если найдется такой элемент* $y_1 \in H_A$, *который дает функционалу* $J(y)$ *наименьшее значение, то* y_1 *будет решением уравнения* (18).

Для доказательства возьмем произвольный элемент $z \in H_A$ и произвольное действительное число λ. Тогда $y_1 + \lambda z \in H_A$ и

$$J(y_1 + \lambda z) \geqslant J(y_1). \tag{24}$$

Но

$$J(y_1 + \lambda z) = (A(y_1 + \lambda z), y_1 + \lambda z) - (f, y_1 + \lambda z) - (y_1 + \lambda z, f) =$$
$$= J(y_1) + \lambda(Az, y_1) + \lambda(Ay_1, z) + \lambda^2(Az, z) - \lambda(f, z) - \lambda(z, f) =$$
$$= J(y_1) + \lambda(z, Ay_1 - f) + \lambda(Ay_1 - f, z) + \lambda^2(Az, z). \tag{25}$$

Отсюда

$$2\lambda \operatorname{Re}[(Ay_1 - f, z)] + \lambda^2(Az, z) \geqslant 0, \tag{26}$$

или

$$\left.\begin{aligned} &2\operatorname{Re}[(Ay_1 - f, z)] + \lambda(Az, z) \geqslant 0 \quad \text{при} \quad \lambda > 0, \\ &2\operatorname{Re}[(Ay_1 - f, z)] + \lambda(Az, z) \leqslant 0 \quad \text{при} \quad \lambda < 0. \end{aligned}\right\} \tag{27}$$

Соотношения (27) будут справедливы для любых действительных λ только в том случае, если

$$\operatorname{Re}[(Ay_1 - f, z)] = 0. \tag{28}$$

Заменив z на iz и проведя те же рассуждения, получим:

$$\operatorname{Im}[(Ay_1 - f, z)] = 0. \tag{29}$$

Таким образом,

$$(Ay_1 - f, z) = 0. \tag{30}$$

В случае действительного пространства мы вместо (28) сразу получим (30). Так как H_A всюду плотно в H, то из (30) следует:

$$Ay_1 - f = 0, \tag{31}$$

что и требовалось доказать.

Функционалы (3) и (5) можно было бы получить как раз таким образом. Так, если

$$Ay = -\frac{d}{dx}(py') + qy,$$

то уравнение (1) можно записать в виде

$$Ay = -f.$$

Будем сначала рассматривать однородные краевые условия (4), т. е. положим там $\alpha = \beta = 0$. Пусть y и z — две какие-то функции, удовлетворяющие однородным условиям (4). Тогда

$$(Ay,\ z) = \int_a^b \left[-\frac{d}{dx}(py') + qy\right] z\,dx = -\int_a^b \frac{d}{dx}(py')\,z\,dx + \\ + \int_a^b qyz\,dx = -py'z\Big|_a^b + \int_a^b py'z'\,dx + \int_a^b qyz\,dx = \\ = (-py'z + pz'y)\Big|_a^b + \int_a^b \left[-\frac{d}{dx}(pz') + qz\right] y\,dx.$$

Но, в силу (4) (при $\alpha = \beta = 0$), имеем:

$$(-py'z + pz'y)\Big|_a^b = 0.$$

Таким образом,

$$(Ay,\ z) = \int_a^b \left[-\frac{d}{dx}(pz') + qz\right] y\,dx = (y,\ Az)$$

и оператор A симметричен.

В данном случае функционал (21) примет вид

$$J(y) = (Ay,\ y) + 2(y,\ f) = \int_a^b \left[-\frac{d}{dx}(py') + qy\right] y\,dx + 2\int_a^b yf\,dx = \\ = -py'y\Big|_a^b + \int_a^b [py'^2 + qy^2 + 2fy]\,dx = \\ = \int_a^b [py'^2 + qy^2 + 2fy]\,dx + \frac{\beta_0 p(b)\,y^2(b)}{\beta_1} - \frac{\alpha_0 p(a)\,y^2(a)}{\alpha_1}.$$

Это совпадает с (5) при $\alpha = \beta = 0$.

Отбросим теперь предположение, что $\alpha = \beta = 0$. Обозначим через z произвольную, но фиксированную, достаточное число раз дифференцируемую функцию, удовлетворяющую неоднородным краевым условиям (4), и будем разыскивать y в виде

$$y = y_1 + z.$$

Тогда y_1 будет удовлетворять однородным краевым условиям (4) и дифференциальному уравнению

$$\frac{d}{dx}(py_1')-qy_1=\frac{d}{dx}[p(y'-z')]-q(y-z)=$$
$$=\frac{d}{dx}(py')-qy-\left[\frac{d}{dx}(pz')-qz\right]=f-\left[\frac{d}{dx}(pz')-qz\right].$$

Таким образом, y_1 должна обращать в минимум функционал

$$J(\bar{y})=\int_a^b\left\{p\bar{y}'^2+q\bar{y}^2+2\bar{y}\left[f-\frac{d}{dx}(pz')+qz\right]\right\}dx+$$
$$+\frac{\beta_0 p(b)\,\bar{y}^2(b)}{\beta_1}-\frac{\alpha_0 p(a)\,\bar{y}^2(a)}{\alpha_1}.$$

Но

$$2\int_a^b\left[-\frac{d}{dx}(pz')+qz\right]\bar{y}\,dx=-2pz'\bar{y}\Big|_a^b+2\int_a^b[pz'\bar{y}'+qz\bar{y}]\,dx=$$
$$=\frac{2\beta_0}{\beta_1}p(b)z(b)\bar{y}(b)-\frac{2\alpha_0}{\alpha_1}p(a)z(a)\bar{y}(a)-\frac{2\beta}{\beta_1}p(b)\bar{y}(b)+$$
$$+\frac{2\alpha}{\alpha_1}p(a)\bar{y}(a)+2\int_a^b[pz'\bar{y}'+qz\bar{y}]\,dx.$$

Поэтому $J(\bar{y})$ можно записать в виде

$$J(\bar{y})=\int_a^b[p(\bar{y}'+z')^2+q(\bar{y}+z)^2+2f(\bar{y}+z)]\,dx-$$
$$-\int_a^b\left[pz'^2+qz^2+2fz\right]dx+\frac{\beta_0}{\beta_1}p(b)[\bar{y}(b)+z(b)]^2-$$
$$-\frac{\alpha_0}{\alpha_1}p(a)[\bar{y}(a)+z(a)]^2+\frac{2\alpha}{\alpha_1}p(a)[\bar{y}(a)+z(a)]-$$
$$-\frac{2\beta}{\beta_1}p(b)[\bar{y}(b)+z(b)]-\frac{\beta_0}{\beta_1}p(b)z^2(b)+\frac{\alpha_0}{\alpha_1}p(a)z^2(a)-$$
$$-\frac{2\alpha}{\alpha_1}p(a)z(a)+\frac{2\beta}{\beta_1}p(b)z(b).$$

Так как $z(x)$ — фиксированная функция, то отсюда следует, что искомое решение задачи (1), (4) будет минимизировать функционал

$$J(y)=\int_a^b\left[py'^2+qy^2+2fy\right]dx+$$
$$+\frac{p(b)}{\beta_1}[\beta_0 y^2(b)-2\beta y(b)]-\frac{p(a)}{\alpha_1}[\alpha_0 y^2(a)-2\alpha y(a)].$$

Аналогично можно было бы получить и функционал (3). Функционалы (9) и (10) можно рассматривать как частные случаи

$$J(y) = \|Ay - f\|^2. \tag{32}$$

Перейдем теперь к вопросам, связанным с приближенным отысканием функций, реализующих минимум функционала. Теоретико-функциональная сущность этих методов состоит в следующем. Находят какую-то последовательность подпространств H_A:

$$H_A^{(1)} \subset H_A^{(2)} \subset \ldots \subset H_A^{(n)} \subset \ldots \subset H_A, \tag{33}$$

такую, что задача о минимуме функционала (21) при $y \in H_A^{(n)}$ решается элементарными средствами (хотя бы принципиально). Такое решение y_n и принимается за приближенное решение задачи. Чаще всего каждое из подпространств $H_A^{(n)}$ бывает конечномерным. Если, например, в H_A имеется счетный базис $g_1, g_2, \ldots, g_n, \ldots$, то в качестве $H_A^{(n)}$ можно взять подпространство, порожденное $g_1, g_2, \ldots, g_n$.

При этом приходится сталкиваться со следующими вопросами. Последовательность приближенных минимизирующих функций $y_1, y_2, \ldots, y_n, \ldots$ такова, что

$$J(y_1) \geqslant J(y_2) \geqslant \ldots \geqslant J(y_n) \geqslant \ldots \geqslant m, \tag{34}$$

где через m обозначено минимальное значение функционала (21) при $y \in H_A$. Следовательно, существует

$$\lim_{n \to \infty} J(y_n) = M \geqslant m. \tag{35}$$

Нужно подбирать $H_A^{(n)}$ так, чтобы было $M = m$. Но даже и при этом условии нельзя делать вывода, что $\lim\limits_{n \to \infty} y_n$ существует и

$$\lim_{n \to \infty} y_n = \bar{y}, \tag{36}$$

где через $\bar{y}$ обозначена функция, реализующая минимум (21) при $y \in H_A$. Поэтому $H_A^{(n)}$ нужно выбирать так, чтобы и условие (36) было выполнено. Наконец, на практике мы всегда ограничиваемся каким-то решением y_n, и следовательно, нам необходимо уметь оценивать разность между точным решением задачи $\bar{y}$ и приближенным y_n.

2. Метод Ритца решения вариационных задач. Посмотрим, как решаются эти задачи в случае функционала

$$J(y) = \int_a^b f(x, y, y')\, dx \tag{37}$$

при некоторых краевых условиях. Функцию f будем предполагать непрерывной. Возьмем какую-нибудь систему функций

$$y(x;\ a_1,\ a_2,\ \ldots,\ a_n), \tag{38}$$

зависящих от n параметров $a_1,\ a_2,\ \ldots,\ a_n$, число которых можно неограниченно увеличивать и таких, что при всех значениях параметров в некоторой области они удовлетворяют краевым условиям (последнее требование не обязательно, если краевые условия естественны). Зафиксируем n и будем использовать в качестве допустимых функций в (37) только функции (38). Подставляя (38) в (37), получим некоторую функцию n переменных

$$J = \Phi(a_1,\ a_2,\ \ldots,\ a_n), \tag{39}$$

и наша задача свелась к отысканию минимума такой функции. Этот минимум находится обычным образом. Приравнивая нулю частные производные

$$\frac{\partial \Phi}{\partial a_i} = 0 \qquad (i = 1,\ 2,\ \ldots,\ n), \tag{40}$$

получаем систему уравнений для определения $a_1,\ a_2,\ \ldots,\ a_n$. Пусть $\bar{a}_1,\ \bar{a}_2,\ \ldots,\ \bar{a}_n$ будет решением этой системы (условия существования и единственности решения предполагаются выполненными). Тогда

$$y_n = y(x;\ \bar{a}_1,\ \bar{a}_2,\ \ldots,\ \bar{a}_n). \tag{41}$$

Как мы видели, $\lim\limits_{n\to\infty} J(y_n)$ существует. Наложим некоторые ограничения на семейства $y(x;\ a_1,\ a_2,\ \ldots,\ a_n)$, обеспечивающие выполнение условия $\lim\limits_{n\to\infty} J(y_n) = m$. Для этого достаточно потребовать следующее: для каждого $\varepsilon > 0$ и для каждой допустимой функции краевой задачи можно подобрать такое n и такие значения параметров $a_1^*,\ a_2^*,\ \ldots,\ a_n^*$, что если

$$y_n^* = y(x;\ a_1^*,\ a_2^*,\ \ldots,\ a_n^*), \tag{42}$$

то на отрезке $[a,\ b]$

$$|y - y_n^*| < \varepsilon; \quad |y' - y_n^{*'}| < \varepsilon. \tag{43}$$

Действительно, тогда и для функции $\bar{y}(x)$, дающей точное решение минимальной задачи (37), можно подобрать такую функцию y_n^*, что будут выполнены условия (43). При этом

$$J(y_n^*) - J(\bar{y}) = \int_a^b [f(x,\ y_n^*,\ y_n^{*'}) - f(x,\ \bar{y},\ \bar{y}')]\,dx \tag{44}$$

может быть сделано как угодно малым в силу непрерывности f. Таким образом,

$$0 \leqslant J(y_n^*) - J(\bar{y}) < \eta, \tag{45}$$

где η как угодно мало. Тем более

$$0 \leqslant J(y_n) - J(\bar{y}) < \eta. \tag{46}$$

(Здесь y_n — минимизирующая функция в семействе (38).) Неравенство (46) и показывает, что $J(y_n) \to J(\bar{y})$.

Сейчас мы перейдем к одному эффективному способу построения функций $y(x; a_1, a_2, \ldots, a_n)$. Нам придется наложить некоторые дополнительные ограничения на функцию $f(x, y, y')$. Ограничимся краевыми условиями вида

$$y(a) = \alpha; \quad y(b) = \beta. \tag{47}$$

Как и всегда, будем предполагать, что вариационная задача имеет единственное решение $\bar{y}(x)$ в классе гладких функций, удовлетворяющих краевым условиям (47). Относительно функции $f(x, y, y')$ будем предполагать следующее:

1. Для каждого числа M должно существовать такое число R, что любая функция, удовлетворяющая краевым условиям (47) и неравенству

$$\int_a^b f(x, y, y')\, dx \leqslant M, \tag{48}$$

должна также удовлетворять неравенству

$$|y| \leqslant R; \quad x \in [a, b]. \tag{49}$$

2. Для каждого числа R можно подобрать такие постоянные $c > 0$, d, $p > 1$, что при всех x и y, принадлежащих области G:

$$a \leqslant x \leqslant b, \quad -R \leqslant y \leqslant R, \tag{50}$$

и произвольном z, $-\infty < z < \infty$, имеет место неравенство

$$f(x, y, z) \geqslant c|z|^p + d. \tag{51}$$

Возьмем произвольную систему непрерывно дифференцируемых на $[a, b]$ функций $\varphi_0(x)$, $\varphi_1(x)$, ..., $\varphi_n(x)$, ..., для которых выполнены следующие условия:

1. $\varphi_0(a) = \alpha$; $\varphi_0(b) = \beta$; $\varphi_k(a) = \varphi_k(b) = 0$ $(k = 1, 2, \ldots)$. (52)

2. При любом n функции $\varphi_1'(x)$, $\varphi_2'(x)$, ..., $\varphi_n'(x)$ линейно независимы.

3. Для любой непрерывной на $[a, b]$ функции $F(x)$ и для любого $\varepsilon > 0$ можно найти обобщенный многочлен

$$F_n(x) = a_1\varphi_1'(x) + a_2\varphi_2'(x) + \ldots + a_n\varphi_n'(x) \tag{53}$$

такой, что

$$|F(x) - F_n(x)| < \varepsilon \quad x \in [a, b]. \tag{54}$$

В качестве функции $\varphi_0(x)$ можно, например, взять

$$\varphi_0(x) = \alpha + \frac{\beta - \alpha}{b - a}(x - a), \tag{55}$$

а в качестве функций $\varphi_k(x)$ $(k = 1, 2, \ldots)$

$$\varphi_k(x) = (x - a)^k(x - b), \tag{56}$$

или

$$\varphi_k(x) = \sin k\pi \frac{x - a}{b - a}. \tag{57}$$

За фуикции $y(x; a_1, a_2, \ldots, a_n)$, о которых говорилось выше, будем брать многочлены

$$y_n(x) = \varphi_0(x) + a_1\varphi_1(x) + \ldots + a_n\varphi_n(x). \tag{58}$$

Очевидно, при любых n и a_i многочлены (58) удовлетворяют краевым условиям (47). Подставляя (58) в (37), получим:

$$J(y_n) = \Phi(a_1, a_2, \ldots, a_n). \tag{59}$$

Докажем, что эта функция достигает своего наименьшего значения при некоторых конечных значениях a_i. Для этого возьмем какую-нибудь систему значений a_i: $a_1^*, a_2^*, \ldots, a_n^*$ и обозначим

$$\Phi(a_1^*, a_2^*, \ldots, a_n^*) = M. \tag{60}$$

Ясно, что нам достаточно рассматривать только такие значения a_i, для которых функция Φ имеет значения меньшие или равные M. Тогда, в силу первого условия, наложенного на $f(x, y, z)$, найдется такое R, что будет выполнено неравенство (49). Но при этом мы можем ограничиться областью G определенной (50), где выполнено неравенство (51). В силу этого неравенства будем иметь:

$$M \geqslant \int_a^b f(x, y_n, y_n')\,dx \geqslant \int_a^b \left\{ c \left| \varphi_0' + \sum_{k=1}^{n} a_k\varphi_k' \right|^p + d \right\} dx. \tag{61}$$

Отсюда

$$\int_a^b \left| \varphi_0' + \sum_{k=1}^{n} a_k\varphi_k' \right|^p dx \leqslant \frac{M - d\,(b - a)}{c} = M_1, \tag{62}$$

или

$$\left\{\int_a^b \left|\varphi_0' + \sum_{k=1}^n a_k \varphi_k'\right|^p dx\right\}^{\frac{1}{p}} \leqslant M_1^{\frac{1}{p}}. \tag{63}$$

Применяя неравенство Минковского [1]), найдем:

$$\left\{\int_a^b \left|\sum_{k=1}^n a_k \varphi_k'\right|^p dx\right\}^{\frac{1}{p}} = \left\{\int_a^b \left|\varphi_0' + \sum_{k=1}^n a_k \varphi_k' - \varphi_0'\right|^p dx\right\}^{\frac{1}{p}} \leqslant$$

$$\leqslant M_1^{\frac{1}{p}} + \left\{\int_a^b |\varphi_0'|^p dx\right\}^{\frac{1}{p}} = M_2. \tag{64}$$

Обозначим

$$\lambda_k = \frac{a_k}{\sqrt{a_1^2 + a_2^2 + \ldots + a_n^2}}. \tag{65}$$

Тогда неравенство (64) можно записать в виде

$$\sqrt{a_1^2 + a_2^2 + \ldots + a_n^2}\left\{\int_a^b \left|\sum_{k=1}^n \lambda_k \varphi_k'\right|^p dx\right\}^{\frac{1}{p}} \leqslant M_2. \tag{66}$$

Второй множитель левой части последнего неравенства как непрерывная функция λ_i на единичной сфере $\left(\text{ибо } \sum_{k=1}^n \lambda_k^2 = 1\right)$ достигает там своего наименьшего значения δ. Это δ не может быть нулем, так как функции φ_k' линейно независимы. Таким образом, из неравенства (66) следует:

$$\sqrt{a_1^2 + a_2^2 + \ldots + a_n^2} \leqslant \frac{M_2}{\delta}. \tag{67}$$

Итак, мы получили, что множество тех значений a_i, для которых функция Φ не превышает M, образует замкнутое ограниченное множество пространства R_n. Следовательно, найдется такая система значений a_i, для которых эта функция принимает свое наименьшее значение.

Заметим, что для многочленов (58) выполнены условия полноты (43). Действительно, в силу третьего предположения о функциях $\varphi_k(x)$ для любого $\varepsilon > 0$ и любой непрерывно дифференцируе-

---

[1]) См., например, Л. А. Люстерник, В. И. Соболев, Элементы функционального анализа, ГТТИ, 1951, стр. 347.

мой допустимой функции $y(x)$ найдутся такое n и такие a_i, что

$$\left| y'(x) - \varphi_0'(x) - \sum_{k=1}^{n} a_k \varphi_k'(x) \right| < \min\left(\varepsilon, \frac{\varepsilon}{b-a}\right). \tag{68}$$

Отсюда

$$\left| y(x) - \varphi_0(x) - \sum_{k=1}^{n} a_k \varphi_k(x) \right| =$$

$$= \left| \int_a^x \left[y'(x) - \varphi_0'(x) - \sum_{k=1}^{n} a_k \varphi_k'(x) \right] dx \right| < \varepsilon. \tag{69}$$

Поэтому мы можем утверждать, что $\lim\limits_{n \to \infty} J(y_n) = m$.

При наших предположениях о функции $f(x, y, z)$ мы можем доказать и большее. Покажем, что из последовательности функций $y_n(x)$, минимизирующих функционал (37), можно выделить сходящуюся подпоследовательность. Прежде всего заметим, что мы можем предполагать неравенство (62) выполненным для всех n. Таким образом,

$$\int_a^b |y_n'(x)|^p dx \leqslant M_1. \tag{70}$$

Поэтому неравенство Гельдера даст нам

$$|y_n(x_1) - y_n(x_2)| = \left| \int_{x_1}^{x_2} y_n'(x)\, dx \right| \leqslant \int_{x_1}^{x_2} |y_n'(x)|\, dx \leqslant$$

$$\leqslant |x_1 - x_2|^{\frac{1}{p'}} \left\{ \int_{x_1}^{x_2} |y_n'(x)|^p dx \right\}^{\frac{1}{p}} \leqslant |x_1 - x_2|^{\frac{1}{p'}} M_1^{\frac{1}{p}}, \tag{71}$$

где

$$x_1 \leqslant x_2; \quad x_1, x_2 \in [a, b], \quad \frac{1}{p} + \frac{1}{p'} = 1. \tag{72}$$

Это означает, что последовательность функций $y_n(x)$ равностепенно непрерывна. Она является и равномерно ограниченной, так как R, определяющее область G, можно считать одним и тем же для всех функций последовательности. Тогда на основании теоремы Арцеля мы можем утверждать, что найдется подпоследовательность из функций $y_n(x)$, сходящаяся к некоторой предельной функции $\tilde{y}(x)$: $\lim\limits_{n_i \to \infty} y_{n_i}(x) = \tilde{y}(x)$.

Но $m = \lim\limits_{n_i \to \infty} J(y_{n_i}) = J\left\{\lim\limits_{n_i \to \infty} y_{n_i}\right\} = J(\tilde{y})$. Таким образом, $\tilde{y}(x) = \bar{y}(x)$.

Дальнейшие рассмотрения будем проводить для краевой задачи (1) и (2) при тех предположениях о $p(x)$, $q(x)$ и $f(x)$, которые были сделаны ранее. Прежде всего покажем, что в этом случае выполнены требования 1 и 2 на функцию $f(x, y, z)$. В этом случае

$$f(x, y, y') = p(x)y'^2 + q(x)y^2 + 2f(x)y. \tag{73}$$

Пусть $J(y) < M$. Тогда

$$\int_a^b p(x)y'^2\,dx \leqslant M + 2\int_a^b |f(x)||y|\,dx. \tag{74}$$

Обозначим

$$\min_{x\in[a,\,b]} p(x) = r; \qquad \max_{x\in[a,\,b]} |y(x)| = Y. \tag{75}$$

Из (74) следует:

$$\int_a^b y'^2\,dx \leqslant \frac{M}{r} + \frac{2Y}{r}\int_a^b |f(x)|\,dx. \tag{76}$$

Кроме того, очевидно,

$$y(x) = \alpha + \int_a^x y'(x)\,dx. \tag{77}$$

Поэтому

$$Y \leqslant |\alpha| + \int_a^b |y'(x)|\,dx. \tag{78}$$

Применяя неравенство Буняковского, получим:

$$Y \leqslant |\alpha| + \sqrt{b-a}\sqrt{\int_a^b y'^2\,dx}. \tag{79}$$

Из (79) и (76) следует:

$$Y \leqslant |\alpha| + \sqrt{b-a}\sqrt{\frac{M}{r} + \frac{2Y}{r}\int_a^b |f(x)|\,dx}. \tag{80}$$

Преобразовывая это неравенство, найдем, что Y должна удовлетворять условию

$$Y^2 - 2\left(|\alpha| + \frac{b-a}{r}\int_a^b |f(x)|\,dx\right)Y + |\alpha|^2 - \frac{M(b-a)}{r} \leqslant 0. \tag{81}$$

Последнее неравенство будет выполнено только при Y, заключенных между следующими двумя числами:

$$|\alpha| + \frac{b-a}{r}\int_a^b |f(x)|\,dx \pm$$

$$\pm \frac{\sqrt{b-a}}{r}\sqrt{rM + (b-a)\left[\int_a^b |f(x)|\,dx\right]^2 + 2r|\alpha|\int_a^b |f(x)|\,dx}. \tag{82}$$

Таким образом если обозначить через R значение величины (82), где выбран знак «$+$» перед корнем, то $0 < Y \leqslant R$.

Проверка того, что выполнено второе условие, тривиальна. Достаточно взять следующие значения:

$$p = 2; \quad c = r; \quad d = -2R \max_{x \in [a,\, b]} |f(x)|. \tag{83}$$

Таким образом, все предыдущие рассуждения применимы к нашему случаю.

Но здесь мы можем пойти и дальше. Рассмотрим $J(y_n) - J(\bar{y})$. Обозначим

$$y_n - \bar{y} = \eta(x). \tag{84}$$

Тогда

$$J(y_n) - J(\bar{y}) = \int_a^b [p(x)\,y_n'^2 + q(x)\,y_n^2 + 2f(x)\,y_n - p(x)\,\bar{y}'^2 -$$

$$- q(x)\,\bar{y}^2 - 2f(x)\,\bar{y}]\,dx = \int_a^b [p(x)\,\eta'^2 + q(x)\,\eta^2]\,dx +$$

$$+ 2\int_a^b [p(x)\,\bar{y}'\,\eta' + q(x)\,\bar{y}\eta + f(x)\,\eta]\,dx. \tag{85}$$

Произведем интегрирование по частям в первом члене второго интеграла, учтя, что $\eta(a) = \eta(b) = 0$. Получим:

$$\int_a^b p(x)\,\bar{y}'\eta'\,dx = [p(x)\,\bar{y}'\eta]_a^b - \int_a^b [p(x)\,\bar{y}']'\,\eta\,dx =$$

$$= -\int_a^b [p(x)\,\bar{y}']'\,\eta\,dx. \tag{86}$$

Таким образом,

$$J(y_n) - J(\bar{y}) = \int_a^b \left[p(x)\,\eta'^2 + q(x)\,\eta^2\right] dx - \\ - 2\int_a^b \left\{\left[p(x)\,\bar{y}'\right]' - q(x)\,\bar{y} - f(x)\right\}\eta\,dx, \qquad (87)$$

и так как $\bar{y}$ удовлетворяет дифференциальному уравнению (1), то

$$J(y_n) - J(\bar{y}) = \int_a^b \left[p(x)\left(y_n' - \bar{y}'\right)^2 + q(x)\left(y_n - \bar{y}\right)^2\right] dx. \qquad (88)$$

Следовательно,

$$\int_a^b \left(y_n' - \bar{y}'\right)^2 dx \leqslant \frac{J(y_n) - J(\bar{y})}{r} \qquad (89)$$

и

$$|y_n - \bar{y}| \leqslant \int_a^b |y_n' - \bar{y}'|\,dx \leqslant \sqrt{b-a}\left\{\int_a^b |y_n' - \bar{y}'|^2\,dx\right\}^{\frac{1}{2}} \leqslant \\ \leqslant \sqrt{\frac{b-a}{r}}\,\sqrt{J(y_n) - J(\bar{y})}. \qquad (90)$$

Так как $J(y_n) - J(\bar{y}) \to 0$ при $n \to \infty$, то мы приходим к заключению, что вся последовательность $\{y_n\}$ стремится к $\bar{y}$. При этом неравенство (90) даст оценку погрешности.

В нашем случае функция Ф, минимум которой приходится отыскивать, имеет вид

$$J(y_n) = A_0 + 2\sum_{k=1}^{n} A_k a_k + \sum_{i=1}^{n}\sum_{k=1}^{n} A_{ik} a_i a_k, \qquad (91)$$

где

$$A_0 = \int_a^b \left[p(x)\,\varphi_0'^2 + q(x)\,\varphi_0^2 + 2f(x)\,\varphi_0\right] dx, \qquad (92)$$

$$A_k = \int_a^b \left[p(x)\,\varphi_0'\varphi_k' + q(x)\,\varphi_0\varphi_k + f(x)\,\varphi_k\right] dx \qquad (k = 1, 2, \ldots, n), \qquad (93)$$

$$A_{ki} = A_{ik} = \int_a^b \left[p(x)\,\varphi_i'\varphi_k' + q(x)\,\varphi_i\varphi_k\right] dx \quad (i,\ k = 1, 2, \ldots, n) \qquad (94)$$

— известные постоянные числа.

Таким образом, для отыскания коэффициентов a_k получим систему линейных алгебраических уравнений

$$\frac{1}{2}\frac{\partial J(y_n)}{\partial a_k} = A_k + \sum_{i=1}^{n} A_{ik}a_i = 0 \qquad (k = 1, 2, \ldots, n). \tag{95}$$

Чтобы эта система имела решение и притом единственное, необходимо и достаточно, чтобы соответствующая однородная система имела только тривиальное решение. Но это так и будет в нашем случае. Действительно, если бы имелась нетривиальная система значений $\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_n$, удовлетворяющая системе

$$\sum_{i=1}^{n} A_{ik}\bar{a}_i = 0 \qquad (k = 1, 2, \ldots, n), \tag{96}$$

то, обозначая

$$\psi_n = \bar{a}_1\varphi_1 + \bar{a}_2\varphi_2 + \ldots + \bar{a}_n\varphi_n \tag{97}$$

и используя значения A_{ik}, мы получили бы

$$\int_a^b [p(x)\psi_n'\varphi_k' + q(x)\psi_n\varphi_k]\,dx = 0 \qquad (k = 1, 2, \ldots, n). \tag{98}$$

Умножая каждое из этих равенств на соответствующее $\bar{a}_k$ и складывая полученные равенства, мы нашли бы

$$\int_a^b [p(x)\psi_n'^2 + q(x)\psi_n^2]\,dx = 0. \tag{99}$$

Равенство (99) противоречит нашему предположению о том, что функции $\varphi_1', \varphi_2', \ldots, \varphi_n'$ линейно независимы.

Изложенный нами метод решения вариационных задач был впервые предложен Ритцем и поэтому носит название *метода Ритца.*

Методом Ритца можно решать и краевую задачу (4) с нулевыми α и β. При этом принимаем

$$\varphi_k(x) = (x-a)^k(x-b)^2 \qquad (k = 2, 3, \ldots), \tag{100}$$

а функции φ_0 и φ_1 подбираем так, чтобы при любых c_0 и c_1 были выполнены условия

$$\left.\begin{aligned} \alpha_0[c_0\varphi_0(a) + c_1\varphi_1(a)] + \alpha_1[c_0\varphi_0'(a) + c_1\varphi_1'(a)] = 0, \\ \beta_0[c_0\varphi_0(b) + c_1\varphi_1(b)] + \beta_1[c_0\varphi_0'(b) + c_1\varphi_1'(b)] = 0. \end{aligned}\right\} \tag{101}$$

Это возможно осуществить, если взять

$$\varphi_0(x) = (x-a)^2\left[x - b - \frac{\beta_1(b-a)}{2\beta_1 + \beta_0(b-a)}\right], \tag{102}$$

$$\varphi_1(x) = (x-b)^2\left[x - a - \frac{\alpha_1(b-a)}{\alpha_0(b-a) - 2\alpha_1}\right]. \tag{103}$$

Случай ненулевых α и β можно свести к случаю нулевых заменой искомой функции

$$y(x) = z(x) + \theta(x), \tag{104}$$

где $\theta(x)$ — некоторая функция, удовлетворяющая краевым условиям (4).

3. Понятие о методе Галеркина. Рассмотрим теперь кратко метод академика Б. Г. Галеркина. Хотя он и не связан по своей идее с предыдущим, но часто приводит к тем же вычислениям. Пусть нам требуется решить дифференциальное уравнение

$$L(y) = f \tag{105}$$

при некоторых однородных краевых условиях. Опять выбираем полную систему независимых функций $\{\varphi_k(x)\}$, удовлетворяющих краевым условиям. За $y_n(x)$ принимаем

$$y_n(x) = \sum_{k=1}^{n} a_k \varphi_k(x) \tag{106}$$

и требуем выполнения следующих условий:

$$\int_a^b [L(y_n) - f]\,\varphi_k(x)\,dx = 0 \qquad (k = 1, 2, \ldots, n). \tag{107}$$

Если бы нам удалось так подобрать $y(x)$, удовлетворяющую краевым условиям, что было бы выполнено

$$\int_a^b [L(y) - f]\,\varphi_k(x)\,dx = 0 \qquad (k = 1, 2, \ldots), \tag{108}$$

то в силу полноты системы функций $\{\varphi_k(x)\}$ отсюда следовало бы, что $y(x)$ удовлетворяет уравнению (105). В нашем же случае можно ожидать, что $y_n(x)$, удовлетворяющее (107), будет близко к точному решению $y(x)$ при достаточно больших n. Теория метода Галеркина более сложна, и мы ее здесь излагать не будем. Заметим, что если $L(y)$ — линейный дифференциальный оператор, то для определения коэффициентов получается система линейных алгебраических уравнений. Эта система совпадает с системой (95) для случая задачи (1), (2), если функции $\varphi_k(x)$ выбирать как и в предыдущем случае. Преимущество метода Галеркина состоит в том, что не приходится разыскивать вариационную задачу, эквивалентную краевой задаче.

УПРАЖНЕНИЯ

1. Найти решение уравнения

$$y'' + \frac{1}{x} y' + \left(1 - \frac{n^2}{x^2}\right) y = 0$$

в виде ряда

$$y = a_0 x^{\sigma} + a_1 x^{\sigma+1} + \ldots + a_n x^{\sigma+k} + \ldots$$

2. Методом последовательных приближений найти решение уравнения

$$y' = y,$$

удовлетворяющее начальному условию $y(0) = 1$.

3. Пусть для дифференциального уравнения $y' = f(x, y)$ найдены функции $U_0(x)$ и $u_0(x)$ такие, что:

а) $U_0(x)$ и $u_0(x)$ определены на отрезке $x_0 \leqslant x \leqslant x_0 + a$;

б) $U_0(x_0) = u_0(x_0) = y_0$;

в) $U_0'(x) - f(x_1, U_0(x)) > 0$, $u_0'(x) - f(x, u_0(x)) < 0$ при $x_0 < x \leqslant x_0 + a$. Тогда, если $f(x, y)$ — непрерывная функция в области, определенной неравенствами

$$x_0 \leqslant x \leqslant x_0 + a; \quad u_0(x) \leqslant y \leqslant U_0(x),$$

и является там монотонно возрастающей функцией y при всяком фиксированном x, то последовательность, образованная из $U_0(x)$ при помощи рекуррентной формулы

$$U_n = y_0 + \int_{x_0}^{x} f(x, U_{n-1})\, dx,$$

монотонно убывает и сходится к решению уравнения $y' = f(x, y)$, удовлетворяющему начальному условию $y(x_0) = y_0$. Аналогично последовательность, построенная из u_0 при помощи рекуррентного соотношения

$$u_n = y_0 + \int_{x_0}^{x} f(x, u_{n-1})\, dx,$$

монотонно возрастает и сходится к такому же решению. Доказать.

4. Заменить условие монотонного возрастания функции $f(x, y)$ по y в предыдущем упражнении на монотонное убывание и исследовать поведение введенных там последовательностей функций $U_n(x)$ и $u_n(x)$.

5. Считая μ малой величиной, найти периодическое решение уравнения

$$\frac{d^2y}{dx^2} + \omega^2 y = \mu(\alpha - \gamma^2 y^2)\frac{dy}{dx} + \lambda \sin x$$

(ω, $\alpha > 0$, γ, $\mu > 0$ — постоянные).

6. Исследовать остаточный член формулы Рунге—Кутта при $r=2$ в предположении, что выполнены неравенства (138) § 4.

7. Различными численными методами найти решение уравнения

$$\frac{dy}{dx} = x^2 + y^2,$$

удовлетворяющее начальному условию $y(0) = 1$. Решение разыскивать на отрезке [0, 1] с четырьмя верными значащими цифрами.

8. Различными численными методами найти решение уравнения

$$\frac{d^2y}{dx^2} = -\frac{0{,}0003}{y^2} + 0{,}01\left(\frac{dy}{dx}\right)^2$$

на отрезке [0; 0,75], если начальные условия имеют вид $y(0) = 1$, $y'(0) = 0$.

9. Различными численными и вариационными методами найти решения краевой задачи:

$$y'' - 4x^2y = -2e^{-x^2},$$

$$y(0) = 1; \quad y(1) = e^{-1}.$$

ЛИТЕРАТУРА

1. Л. В. Канторович, В. И. Крылов, Приближенные методы высшего анализа, Гостехиздат, 1952.
2. Л. В. Канторович, Функциональный анализ и прикладная математика, УМН, т. 3, вып. 6, 1948.
3. Л. Коллатц, Численные методы решения дифференциальных уравнений, ИЛ, 1953.
4. А. Н. Крылов, Лекции о приближенных вычислениях, Гостехиздат, 1951.
5. Ш. Е. Микеладзе, Новые методы интегрирования дифференциальных уравнений, Гостехиздат, 1951.
6. В. Э. Милн, Численное решение дифференциальных уравнений, ИЛ, 1955.
7. С. Г. Михлин, Прямые методы в математической физике, Гостехиздат, 1950.
8. М. Р. Шура-Бура, Оценки ошибок численного интегрирования обыкновенных дифференциальных уравнений, ПММ, т. 16, вып. 5, 1952.
9. Н. С. Бахвалов, К оценке ошибки при численном интегрировании дифференциальных уравнений экстраполяционным методом Адамса, ДАН СССР, 1955, т. 104, № 5, 683—686.
10. А. Д. Горбунов и Б. М. Будак, О сходимости некоторых конечно-разностных процессов для уравнений $y' = f(x, y)$ и $y'(x) = f(x, y(x), y(x - r(x)))$, ДАН СССР, 119, № 4, стр. 644—647, 1958, или
Б. М. Будак и А. Д. Горбунов, О сходимости некоторых конечно-разностных процессов для уравнений $y' = f(x, y)$ и $y'(x) = f(x, y(x), y(x - r(x)))$, Вестник МГУ, № 5, 1958, стр. 23—32.

ГЛАВА 10

ПРИБЛИЖЕННЫЕ МЕТОДЫ РЕШЕНИЯ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ В ЧАСТНЫХ ПРОИЗВОДНЫХ И ИНТЕГРАЛЬНЫХ УРАВНЕНИЙ

§ 1. Введение

С дифференциальными уравнениями в частных производных и интегральными уравнениями приходится встречаться в самых разнообразных областях естествознания, причем получить их решение в явном виде, в виде конечной формулы, удается только в самых простейших случаях.

В связи с этим особое значение приобретают приближенные методы решения различных задач для дифференциальных уравнений в частных производных, систем дифференциальных уравнений в частных производных и интегральных уравнений или, как часто говорят, задач математической физики.

В настоящей главе мы и рассмотрим некоторые, наиболее распространенные методы решения задач математической физики. При этом мы ограничимся в основном методами решения задач для линейных дифференциальных уравнений в частных производных второго порядка с двумя независимыми переменными и линейными интегральными уравнениями, в которых искомая функция зависит только от одного независимого переменного. Изложение методов для случая произвольного числа переменных было бы связано с очень громоздкими записями, в то время как основные идеи методов, а также возникающие при их реализации трудности хорошо усматриваются в простейших случаях.

Что касается нелинейных уравнений, то хотя отдельные задачи для нелинейных уравнений и были разрешены, однако общая теория приближенных методов для нелинейных уравнений все еще отсутствует. В последнее время численным методам решения задач для нелинейных уравнений уделяется много внимания, но их разработка еще не достигла такого состояния, при котором их можно было бы включить в учебное пособие.

Как и в случае обыкновенных дифференциальных уравнений, приближенные методы решения различных задач для

дифференциальных уравнений в частных производных можно разбить на две группы:

1) методы, в которых приближенное решение получается в аналитической форме, например в виде отрезка некоторого ряда, и

2) методы, с помощью которых можно получить таблицу приближенных значений искомого решения в некоторых точках рассматриваемой области, — численные методы.

К первой группе относится прежде всего метод Фурье решения краевых задач для дифференциальных уравнений в частных производных, при применении которого точное решение получается в виде некоторого ряда, а за приближенное решение может быть принята сумма некоторого числа первых его членов. Метод Фурье решения классических задач математической физики подробно излагается в курсе математической физики, и мы на нем совсем не будем останавливаться. Из методов первой группы мы рассмотрим лишь вариационные методы решения краевых задач для уравнений в частных производных и близкий к ним метод Галеркина.

Наиболее широко распространенным методом численного решения задач для дифференциальных уравнений в частных производных является метод сеток, или метод конечных разностей, а также метод характеристик решения уравнений и систем уравнений гиперболического типа, который в сущности также является конечноразностным методом, только в этом методе дифференциальное уравнение в частных производных или система таких уравнений предварительно сводится к эквивалентной ей системе обыкновенных дифференциальных уравнений, которая и решается разностным методом. Описанию метода сеток для решения некоторых задач математической физики в основном и посвящена эта глава.

Особое место занимает метод прямых, который в зависимости от способа его реализации может быть отнесен как к той, так и к другой группе методов. В этом методе ищется приближенно решение дифференциального уравнения в частных производных вдоль некоторого семейства прямых. При этом вместо дифференциального уравнения в частных производных получается система обыкновенных дифференциальных уравнений. Если эта система решается в конечном виде, то мы получаем приближенное решение дифференциального уравнения в частных производных в виде системы функций, приближенно представляющих искомое решение вдоль рассматриваемых прямых. Если же система обыкновенных дифференциальных уравнений решается численными методами, то и приближенное решение уравнения в частных производных получается в виде таблицы, и в этом случае этот метод можно отнести к группе численных методов.

В последнем параграфе главы изложены методы приближенного решения линейных интегральных уравнений типа Фредгольма и Вольтерра.

В силу значительных трудностей, возникающих при приближенном решении дифференциальных уравнений в частных производных, мы ограничимся при изложении из педагогических соображений только простейшими уравнениями и простейшими задачами для них, причем во многих случаях не приводятся доказательства сходимости, а также оценки погрешностей, если даже они существуют. Это отнюдь не означает, что описанные методы неприменимы для решения других более сложных задач.

§ 2. Метод сеток решения краевых задач для дифференциальных уравнений эллиптического типа

Метод сеток является одним из самых распространенных методов численного решения краевых задач для дифференциальных уравнений эллиптического типа[1]). При изложении этого метода мы ограничимся краевыми задачами для линейных дифференциальных уравнений с двумя независимыми переменными.

1. Идея метода сеток. Идею метода сеток изложим на примере решения задачи Дирихле для уравнения

$$a\frac{\partial^2 u}{\partial x^2}+b\frac{\partial^2 u}{\partial y^2}+c\frac{\partial u}{\partial x}+d\frac{\partial u}{\partial y}+gu=f, \tag{1}$$

где a, b, c, d, g, f — функции независимых переменных x и y, определенные в конечной области G с границей Γ. Относительно этих функций предположим, что они непрерывны в $G+\Gamma$, a и b положительны в $G+\Gamma$, а g неположительна в ней.

Пусть необходимо найти решение уравнения (1), непрерывное вплоть до границы Γ, принимающее в точках границы заданные значения φ, т. е.

$$u|_\Gamma=\varphi. \tag{2}$$

где φ — непрерывная функция на Γ.

Для отыскания приближенного численного решения этой задачи проведем два семейства параллельных прямых:

$$x=x_0+ih \qquad (i=0,\ \pm 1,\ \pm 2,\ \ldots),$$
$$y=y_0+kl \qquad (k=0,\ \pm 1,\ \pm 2,\ \ldots).$$

Точки пересечения этих прямых назовем *узлами*. Два узла назовем соседними, если они удалены друг от друга в направлении оси x или y на расстояние шага сетки в направлении этой оси. Будем рассматривать только те узлы, которые принадлежат $G+\Gamma$. Те

[1]) По поводу применений метода конечных разностей в теории уравнений эллиптического типа см. обзорную статью О. А. Ладыженской, УМН, т. XII, вып. 5 (77), 1957, стр. 123—148.

из них, у которых все четыре соседних узла принадлежат этому множеству, назовем *внутренними*. Множество внутренних узлов назовем *сеточной областью* и обозначим через G^*. Те узлы, у которых хотя бы один соседний узел не принадлежит к рассматриваемому множеству, назовем *граничными*, а совокупность их назовем *границей сеточной области* и обозначим через Γ^*. Для каждого внутреннего узла (i, k) составим разностное уравнение, заменив в точке (x_0+ih, y_0+kl) производные, входящие в уравнение (1), разностными отношениями, положив, например,

$$\left.\begin{aligned}
&\left(\frac{\partial u}{\partial x}\right)_{(i,k)} \approx \frac{u_{i+1,k}-u_{i-1,k}}{2h}; \quad \left(\frac{\partial u}{\partial y}\right)_{(i,k)} \approx \frac{u_{i,k+1}-u_{i,k-1}}{2l},\\
&\left(\frac{\partial^2 u}{\partial x^2}\right)_{(i,k)} \approx \frac{u_{i+1,k}-2u_{ik}+u_{i-1,k}}{h^2};\\
&\left(\frac{\partial^2 u}{\partial y^2}\right)_{(i,k)} \approx \frac{u_{i,k+1}-2u_{ik}+u_{i,k-1}}{l^2},
\end{aligned}\right\} \qquad (3)$$

где принято обозначение $u_{ik}=u(x_0+ih, y_0+kl)$. Обозначая значения коэффициентов уравнения (1) в узле (i, k) через a_{ik}, b_{ik}, c_{ik}, d_{ik}, g_{ik}, f_{ik}, получим для узла (i, k) разностное уравнение

$$\begin{aligned}
lu_{ik} = a_{ik}&\frac{u_{i+1,k}-2u_{ik}+u_{i-1,k}}{h^2}+\\
+b_{ik}&\frac{u_{i,k+1}-2u_{ik}+u_{i,k-1}}{l^2}+\\
+c_{ik}&\frac{u_{i+1,k}-u_{i-1,k}}{2h}+\\
+d_{ik}&\frac{u_{i,k+1}-u_{i,k-1}}{2l}+g_{ik}u_{ik}=f_{ik}.
\end{aligned} \qquad (4)$$

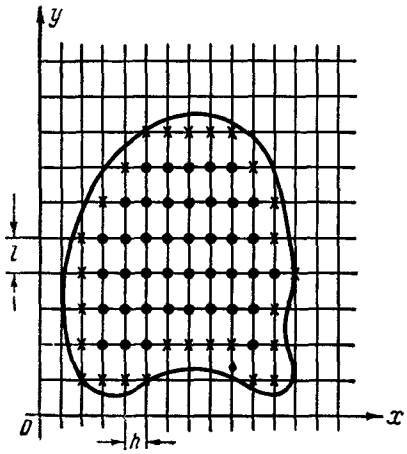

Рис. 26.

Такие уравнения можно записать для каждого внутреннего узла. Если узел (i, k) является граничным узлом, то u_{ik} в этом узле положим равным значению функции φ в точке Γ, ближайшей к этому узлу, т. е. просто снесем в граничные узлы значения функции φ из ближайших к ним точек границы Γ. Таким образом, для отыскания значений u_{ik} решения во внутренних узлах мы получим систему линейных алгебраических уравнений, в которой число уравнений равно числу неизвестных. Если эта система разрешима, то, решив ее, получим приближенные значения искомого решения на конечном множестве точек, являющихся внутренними узлами.

Сразу же возникает ряд вопросов:

1. Разрешима ли полученная система разностных уравнений и если разрешима, то какими способами она может быть решена?

2. Насколько будут близки полученные при этом значения к значениям точного решения задачи Дирихле в соответствующих точках?

Можно сразу сказать, что погрешность, получаемая при этом методе, складывается из трех погрешностей, имеющих разную природу:

а) погрешность, возникающая в результате замены дифференциального уравнения разностным уравнением, зависящая от точности аппроксимации дифференциального уравнения разностным;

б) погрешность, возникающая за счет сноса граничных условий с Γ на границу сеточной области Γ^*;

в) погрешность, возникающая в результате того, что решение разностной системы уравнений, вообще говоря, может быть найдено только приближенно.

Совокупность погрешностей а) и б) дает нам погрешность метода, а погрешность в) есть вычислительная погрешность.

3. Можно ли, неограниченно сгущая сетку, получить решение, сколь угодно близкое к точному решению краевой задачи для уравнения (1), т. е. возникает вопрос о сходимости метода сеток.

В этом параграфе мы постараемся дать ответы на поставленные вопросы для некоторых конкретных краевых задач.

2. Аппроксимация дифференциальных уравнений разностными. Применяя метод сеток для решения краевых задач, мы прежде всего сталкиваемся с задачей замены дифференциальных уравнений разностными уравнениями. Эта замена может быть выполнена разными способами.

Один из способов аппроксимации дифференциального уравнения разностным заключается в том, что производные, входящие в дифференциальное уравнение, заменяются линейными комбинациями значений функции u в узлах сетки по тем или иным формулам численного дифференцирования. Такой прием был применен при построении разностного уравнения (4) в п. 1. В зависимости от того, какими формулами численного дифференцирования будем пользоваться, получим различную точность аппроксимации дифференциального уравнения разностным. Рассмотрим, например, погрешность, получаемую в результате замены дифференциального уравнения (1) разностным уравнением (4). Предполагая, что решение краевой задачи для уравнения (1) имеет непрерывные производные до четвертого

порядка включительно, имеем следующие равенства:

$$\left.\begin{aligned}
&\frac{u(x_i+h, y_k)-u(x_i-h, y_k)}{2h} = \\
&\quad = u'_x(x_i, y_k)+\frac{h^2}{6}u'''_{x^3}(\tilde{x}, y_k) \quad (x_i-h\leqslant\tilde{x}\leqslant x_i+h), \\
&\frac{u(x_i, y_k+l)-u(x_i, y_k-l)}{2l} = \\
&\quad = u'_y(x_i, y_k)+\frac{l^2}{6}u'''_{y^3}(x_i, \tilde{y}) \quad (y_k-l\leqslant\tilde{y}\leqslant y_k+l), \\
&\frac{u(x_i+h, y_k)-2u(x_i, y_k)+u(x_i-h, y_k)}{h^2} = \\
&\quad = u''_{x^2}(x_i, y_k)+\frac{h^2}{12}u^{(\mathrm{IV})}_{x^4}(\tilde{\tilde{x}}, y_k) \quad (x_i-h\leqslant\tilde{\tilde{x}}\leqslant x_i+h), \\
&\frac{u(x_i, y_k+l)-2u(x_i, y_k)+u(x_i, y_k-l)}{l^2} = \\
&\quad = u''_{y^2}(x_i, y_k)+\frac{l^2}{12}u^{(\mathrm{IV})}_{y^4}(x_i, \tilde{\tilde{y}}) \quad (y_k-l\leqslant\tilde{\tilde{y}}\leqslant y_k+l),
\end{aligned}\right\} \tag{5}$$

которые легко получить, разлагая левые части по формуле Тейлора в окрестности точки (x_i, y_k). Используя эти соотношения, имеем:

$$\begin{aligned}
lu_{ik} = &\{a_{ik}u''_{x^2}+b_{ik}u''_{y^2}+c_{ik}u'_x+d_{ik}u'_y+g_{ik}u\}_{(x_i, y_k)}+ \\
&+\frac{h^2}{12}\{a_{ik}u^{(\mathrm{IV})}_{x^4}(\tilde{\tilde{x}}, y_k)+b_{ik}\alpha^2u^{(\mathrm{IV})}_{y^4}(x_i, \tilde{\tilde{y}})+2c_{ik}u'''_{x^3}(\tilde{x}, y_k)+ \\
&+2d_{ik}\alpha^2u'''_{y^3}(x_i, \tilde{y})\} = [L(u)]_{(i, k)}+R_{ik},
\end{aligned} \tag{6}$$

где

$$\alpha = \frac{l}{h},$$

$$\begin{aligned}
R_{ik} = &\frac{h^2}{12}\{a_{ik}u^{(\mathrm{IV})}_{x^4}(\tilde{\tilde{x}}, y_k)+b_{ik}\alpha^2u^{(\mathrm{IV})}_{y^4}(x_i, \tilde{\tilde{y}})+ \\
&+2c_{ik}u'''_{x^3}(\tilde{x}, y_k)+2d_{ik}\alpha^2u'''_{y^3}(x_i, \tilde{y})\}.
\end{aligned} \tag{7}$$

Если ввести обозначения

$$M_3 = \max_G\left\{\left|\frac{\partial^3u}{\partial x^3}\right|, \left|\frac{\partial^3u}{\partial y^3}\right|\right\}; \quad M_4 = \max_G\left\{\left|\frac{\partial^4x}{\partial x^4}\right|, \left|\frac{\partial^4u}{\partial y^4}\right|\right\},$$

то

$$|R_{ik}|\leqslant\frac{h^2}{12}\{(|a_{ik}|+\alpha^2|b_{ik}|)M_4+2(|c_{ik}|+\alpha^2|d_{ik}|)M_3\}. \tag{8}$$

Таким образом,

$$lu_{ik}-f_{ik} = \{L(u)-f\}_{(i, k)}+R_{ik} = R_{ik},$$

т. е., заменяя дифференциальное уравнение (1) разностным уравнением (4), мы совершаем ошибку, равную R_{ik}, которая имеет порядок h^2 относительно шага h $\left(\text{полагая } \alpha = \frac{l}{h} = \text{const}\right)$.

Если мы возьмем уравнение Пуассона

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y) \tag{9}$$

и квадратную сетку $h = l$, то при описанном способе замены производных получим следующее разностное уравнение:

$$\frac{u_{i+1,\,k} + u_{i-1,\,k} + u_{i,\,k+1} + u_{i,\,k-1} - 4u_{ik}}{h^2} = f_{ik}. \tag{10}$$

При этом для погрешности аппроксимации $R_{i,\,k}$ будем иметь оценку

$$|R_{ik}| \leqslant \frac{h^2}{6} M_4. \tag{11}$$

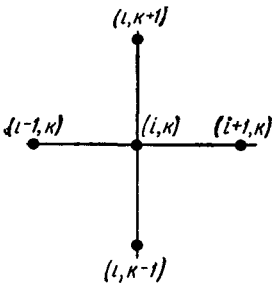

Рис. 27.

В разностной схеме (10) для узла (i, k) участвуют четыре соседних узла, расположенных по «кресту».

Другой способ получения разностных уравнений состоит в следующем. Для получения разностного уравнения, аппроксимирующего дифференциальное уравнение в узле (i, k), рассмотрим N узлов, расположенных определенным образом около точки (i, k). Для простоты записи узел (i, k) будем обозначать 0, а остальные рассматриваемые узлы перенумеруем числами 1, 2, ..., N. Составим линейную комбинацию

$$\sum_{j=0}^{N} c_j u_j$$

с неопределенными коэффициентами c_j, где u_j — значение u в узле j. Предполагая у функции u наличие $n+1$ производных, разложим u_j по формуле Тейлора в окрестности узла 0. Подставим эти разложения в линейную комбинацию и сгруппируем члены с одинаковыми производными от функции u. Получим

$$\sum_{j=0}^{N} c_j u_j = \sum_{i+k \leqslant n} \gamma_{ik} \left(\frac{\partial^{i+k} u}{\partial x^i \partial y^k}\right)_0 + \text{остаточный член}. \tag{12}$$

Заметим, что γ_{ik} линейно выражаются через c_j. Остаточный член будет иметь вид $\theta h^{n+1} K M_{n+1}$, где $|\theta| \leqslant 1$, K — некоторое число, не зависящее от h, $M_{n+1} = \max\limits_{G} \left\{ \left|\frac{\partial^{n+1} u}{\partial x^{n+1}}\right|, \left|\frac{\partial^{n+1} u}{\partial x^n \partial y}\right|, \ldots, \left|\frac{\partial^{n+1} u}{\partial y^{n+1}}\right| \right\}$, h — наименьшая по абсолютной величине разность координат узла 0

и узлов j ($j=0, 1, 2, \ldots, N$). Далее, подбираем коэффициенты c_j таким образом, чтобы правая часть в равенстве отличалась бы возможно меньше от дифференциального выражения $L(u)$ в точке 0. Для этого потребуем, чтобы коэффициенты при производных в уравнении совпадали бы с коэффициентами при соответствующих производных в правой части (12):

$$a_0=\gamma_{20};\quad b_0=\gamma_{02};\quad 0=\gamma_{11};\quad c_0=\gamma_{10};\quad d_0=\gamma_{01};\quad g_0=\gamma_{00},$$

и, кроме того, коэффициенты при старших производных в (12) до порядка r ($2\leqslant r\leqslant n$) обращались бы в нуль, т. е.

$$\gamma_{ik}=0 \quad \text{при} \quad 2<i+k\leqslant r.$$

Если при этом система уравнений относительно c_j имеет решение, то мы найдем такие c_j, при которых

$$\sum_{j=0}^{N} c_j u_j = [L(u)]_0 + \theta K h^{r+1} M_{r+1}. \tag{13}$$

Используя достаточно большое число узлов N, можно получить достаточно хорошую аппроксимацию дифференциального уравнения в узле 0, заменяя дифференциальное уравнение разностным уравнением

$$\sum_{j=0}^{N} c_j u_j = f_0. \tag{14}$$

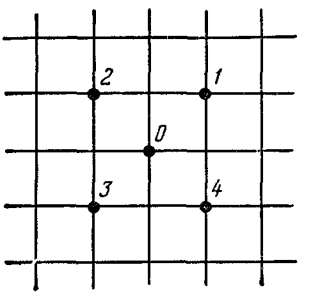

Рис. 28.

Этот способ имеет то преимущество, что можно рассматривать не только прямоугольную сетку, но и другие сетки (треугольную сетку, сетку параллелограммов и др.).

Для внутренних узлов, достаточно удаленных от границы, расположение узлов, участвующих в линейной комбинации для составления разностного уравнения в них, можно и целесообразно сохранять. Для узлов, близких к границе, это не всегда удается. Но этот способ для этих узлов при другой конфигурации узлов, участвующих в линейной комбинации, часто позволяет получать разностные уравнения той же точности. В частности, этот способ позволяет и граничные условия аппроксимировать достаточно точно.

Если имеется произвол в выборе решения системы для c_j, то выбирают наиболее простое решение с тем, чтобы получить наиболее простые разностные уравнения.

Рассмотрим этот метод на примерах аппроксимации уравнений Пуассона для разных сеток и некоторых других примерах.

Сначала рассмотрим разностную аппроксимацию уравнения Пуассона, если сетка квадратная с шагом h и в разностном уравнении для узла 0 участвуют узлы, помеченные номерами 1, 2, 3, 4 на

рис. 28. Учитывая равноправие в уравнении x и y и симметричное расположение узлов, очевидно, имеет смысл искать разностную аппроксимацию вида

$$lu_0 = c_0 u_0 + c_1 (u_1 + u_2 + u_3 + u_4).$$

Предполагая у функции u наличие достаточного числа производных и разлагая u_i по формуле Тейлора в окрестности узла 0, будем иметь:

$$\left.\begin{aligned}
u_1 &= u(x_0 + h,\ y_0 + h) = u_0 + \left\{ h\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right) u + \right. \\
&\quad \left. + \frac{h^2}{2!}\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right)^2 u + \frac{h^3}{3!}\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right)^3 u + \frac{h^4}{4!}\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right)^4 u + \ldots \right\}_0, \\
u_2 &= u(x_0 + h,\ y_0 - h) = u_0 + \left\{ h\left(\frac{\partial}{\partial x} - \frac{\partial}{\partial y}\right) u + \right. \\
&\quad \left. + \frac{h^2}{2!}\left(\frac{\partial}{\partial x} - \frac{\partial}{\partial y}\right)^2 u + \frac{h^3}{3!}\left(\frac{\partial}{\partial x} - \frac{\partial}{\partial y}\right)^3 u + \frac{h^4}{4!}\left(\frac{\partial}{\partial x} - \frac{\partial}{\partial y}\right)^4 u + \ldots \right\}_0, \\
u_3 &= u(x_0 - h,\ y_0 - h) = u_0 + \left\{ -h\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right) u + \right. \\
&\quad \left. + \frac{h^2}{2!}\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right)^2 u - \frac{h^3}{3!}\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right)^3 u + \frac{h^4}{4!}\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right)^4 u - \ldots \right\}_0, \\
u_4 &= u(x_0 - h,\ y_0 + h) = u_0 + \left\{ h\left(-\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right) u + \frac{h^2}{2!}\left(-\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right)^2 u + \right. \\
&\quad \left. + \frac{h^3}{3!}\left(-\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right)^3 u + \frac{h^4}{4!}\left(-\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right)^4 u + \ldots \right\}_0,
\end{aligned}\right\} \tag{13'}$$

а

$$lu_0 = (c_0 + 4c_1) u_0 + \\ + 4c_1 \left\{ \frac{h^2}{2} (u''_{x^2} + u''_{y^2})_0 + \frac{h^4}{4!} (u^{(\mathrm{IV})}_{x^4} + 6u^{(\mathrm{IV})}_{x^2y^2} + u^{(\mathrm{IV})}_{y^4})_0 + \ldots \right\}.$$

Для того чтобы lu_0 аппроксимировало оператор Лапласа

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2},$$

необходимо потребовать выполнения условий

$$c_0 + 4c_1 = 0; \quad 2c_1 h^2 = 1,$$

т. е.

$$c_0 = -\frac{2}{h^2}; \quad c_1 = \frac{1}{2h^2}.$$

Окончательно получаем:

$$lu_0 = \frac{1}{2h^2} [u_1 + u_2 + u_3 + u_4 - 4u_0] = \\ = (\Delta u)_0 + \frac{h^2}{12} (u^{(\mathrm{IV})}_{x^4} + 6u^{(\mathrm{IV})}_{x^2y^2} + u^{(\mathrm{IV})}_{y^4})_0 + \ldots = (\Delta u)_0 + R.$$

Если в разложении (13) ограничиться разложением по формуле Тейлора с остаточным членом в производных четвертого порядка и положить

$$M_4 = \max_G \left\{ \left| \frac{\partial^4 u}{\partial x^4} \right|, \left| \frac{\partial^4 u}{\partial x^3 \partial y} \right|, \ldots, \left| \frac{\partial^4 u}{\partial y^4} \right| \right\},$$

то для R будем иметь следующую оценку:

$$|R| \leqslant 4c_1 \frac{h^4}{24} 2^4 M_4 = \frac{4h^2}{3} M_4.$$

Заменяя $(\Delta u)_0$ через f_0 и отбрасывая R, получим разностное уравнение

$$u_1 + u_2 + u_3 + u_4 - 4u_0 = 2h^2 f_0, \quad (14')$$

аппроксимирующее уравнение Пуассона $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f$, причем погрешность аппроксимации не превосходит $\frac{4h^2}{3} M_4$.

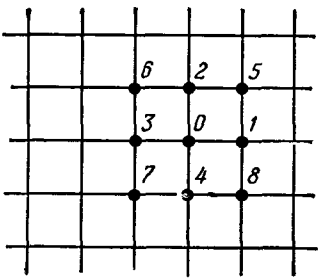

Рис. 29.

Рассмотрим теперь разностную аппроксимацию уравнения Пуассона, в которой используются узлы, изображенные на рис. 29. Из тех же соображений, что и в предыдущем примере, будем искать разностную аппроксимацию вида

$$lu_0 = c_0 u_0 + c_1 (u_1 + u_2 + u_3 + u_4) + c_2 (u_5 + u_6 + u_7 + u_8).$$

Разлагая u_i по формуле Тейлора и приводя подобные члены, будем иметь:

$$u_1 + u_2 + u_3 + u_4 = 4u_0 + \frac{2h^2}{2!} \left(u''_{x^2} + u''_{y^2} \right)_0 +$$

$$+ \frac{2h^4}{4!} \left(u^{(4)}_{x^4} + u^{(4)}_{y^4} \right)_0 + \frac{2h^6}{6!} \left(u^{(6)}_{x^6} + u^{(6)}_{y^6} \right)_0 + \frac{2h^8}{8!} \left(u^{(8)}_{x^8} + u^{(8)}_{y^8} \right)_0 + \ldots,$$

$$u_5 + u_6 + u_7 + u_8 = 4u_0 + 2h^2 \left(u''_{x^2} + u''_{y^2} \right)_0 +$$

$$+ \frac{4h^4}{4!} \left(u^{(4)}_{x^4} + 6u^{(4)}_{x^2y^2} + u^{(4)}_{y^4} \right)_0 + \frac{4h^6}{6!} \left(u^{(6)}_{x^6} + 15u^{(6)}_{x^4x^2} + 15u^{(6)}_{x^2y^4} + u^{(6)}_{y^6} \right)_0 +$$

$$+ \frac{4h^8}{8!} \left(u^{(8)}_{x^8} + 28u^{(8)}_{x^6y^2} + 70u^{(8)}_{x^4y^4} + 28u^{(8)}_{x^2y^6} + u^{(8)}_{y^8} \right)_0 + \ldots.$$

т. е.

$$lu_0 = (c_0 + 4c_1 + 4c_2)\,u_0 + h^2(c_1 + 2c_2)\left(u^{(2)}_{x^2} + u^{(2)}_{y^2}\right)_0 +$$
$$+ \frac{h^4}{12}(c_1 + 2c_2)\left(u^{(4)}_{x^4} + u^{(4)}_{y^4}\right)_0 + c_2 h^4\left(u^{(4)}_{x^2y^2}\right)_0 +$$
$$+ \frac{2h^6}{6!}(c_1 + 2c_2)\left(u^{(6)}_{x^6} + u^{(6)}_{y^6}\right)_0 + \frac{h^6}{12}c_2\left(u^{(6)}_{x^4y^2} + u^{(6)}_{x^2y^4}\right)_0 +$$
$$+ \frac{2h^8}{8!}(c_1 + 2c_2)\left(u^{(8)}_{x^8} + u^{(8)}_{y^8}\right)_0 + \frac{h^8}{720}c_2\left(2u^{(8)}_{x^6y^2} + 5u^{(8)}_{x^4y^4} + 2u^{(8)}_{x^2y^6}\right)_0 + \ldots$$

Для того чтобы lu_0 аппроксимировало оператор Лапласа, положим

$$c_0 + 4c_1 + 4c_2 = 0,$$
$$h^2(c_1 + 2c_2) = 1.$$

Подберем c_2 из того условия, чтобы члены с производными четвертого порядка могли быть получены путем дифференцирования оператора Лапласа. Для этого нужно положить $c_2 = \frac{1}{6h^2}$. Тогда для c_0, c_1 получим следующие значения:

$$c_0 = -\frac{10}{3h^2}; \quad c_1 = \frac{2}{3h^2}$$

и

$$lu_0 = \left(u''_{x^2} + u''_{y^2}\right)_0 + \frac{h^2}{12}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\left(u''_{x^2} + u''_{y^2}\right)_0 +$$
$$+ \frac{2h^4}{6!}\left[\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)^2\left(u''_{x^2} + u''_{y^2}\right) + 2\frac{\partial^4}{\partial x^2\,\partial y^2}\left(u''_{x^2} + u''_{y^2}\right)\right]_0 + R,$$

где R зависит от производных восьмого порядка и имеет порядок h^6.

Так как

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y),$$

то

$$lu_0 = f_0 + \frac{h^2}{12}\left(\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}\right)_0 + \frac{2h^4}{6!}\left(\frac{\partial^4 f}{\partial x^4} + 4\frac{\partial^4 f}{\partial x^2\,\partial y^2} + \frac{\partial^4 f}{\partial y^4}\right)_0 + R$$

и разностное уравнение

$$\frac{4(u_1 + u_2 + u_3 + u_4) + (u_5 + u_6 + u_7 + u_8) - 20u_0}{6} =$$
$$= h^2 f_0 + \frac{h^4}{12}(\Delta f)_0 + \frac{2h^6}{6!}\left(\frac{\partial^4 f}{\partial x^4} + 4\frac{\partial^4 f}{\partial x^2\,\partial y^2} + \frac{\partial^4 f}{\partial y^4}\right)_0 \qquad (15)$$

дает аппроксимацию уравнения Пуассона с точностью до h^6.

Если при выводе разностной аппроксимации воспользоваться разложением по формуле Тейлора с остаточным членом, содержащим

производные восьмого порядка, то, введя обозначение

$$M_8 = \max_G \left\{ \left| \frac{\partial^8 u}{\partial x^8} \right|, \quad \left| \frac{\partial^8 u}{\partial x^7 \partial y} \right|, \quad \ldots, \quad \left| \frac{\partial^8 u}{\partial y^8} \right| \right\}$$

для остаточного члена, дающего погрешность аппроксимации, будем иметь оценку

$$|R| \leqslant \frac{520h^6}{3 \cdot 8!} M_8.$$

Пользоваться полученной разностной аппроксимацией уравнения Пуассона можно только в случае, если функция f задана аналитически. Если же она известна только в узлах сетки или имеет сложное аналитическое выражение, то дифференцирование ее будет затруднительно. Поэтому в этом случае аппроксимацию упрощают, отбрасывая член с h^6 и заменяя $(\Delta f)_0$ через

$$\frac{1}{h^2}(f_1 + f_2 + f_3 + f_4 - 4f_0).$$

В самом деле,

$$\frac{1}{h^2}(f_1 + f_2 + f_3 + f_4 - 4f_0) = (\Delta f)_0 + \bar{R},$$

где

$$|\bar{R}| \leqslant \frac{h^2}{6} \bar{M}_4;$$

$$\bar{M}_4 = \max_G \left\{ \left| \frac{\partial^4 f}{\partial x^4} \right|, \left| \frac{\partial^4 f}{\partial x^3 \partial y} \right|, \quad \ldots, \quad \left| \frac{\partial^4 f}{\partial y^4} \right| \right\}.$$

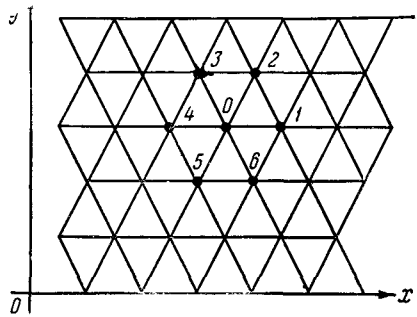

Рис. 30.

Подставляя вместо $(\Delta f)_0$ комбинацию $\frac{1}{h^2}(f_1 + f_2 + f_3 + f_4 - 4f_0)$, мы отбрасываем члены шестого порядка относительно h и получаем разностную аппроксимацию

$$4(u_1 + u_2 + u_3 + u_4) + (u_5 + u_6 + u_7 + u_8) - 20u_0 = \frac{h^2}{2}(8f_0 + f_1 + f_2 + f_3 + f_4), \quad (16)$$

аппроксимирующую уравнение Пуассона с точностью до h^4.

Рассмотрим теперь разностную аппроксимацию уравнения Пуассона для сетки из правильных треугольников со стороной h (рис. 30). При составлении разностного уравнения для узла 0 возьмем шесть ближайших окружающих его узлов 1, 2, 3, 4, 5, 6 и составим линейную комбинацию

$$lu_0 = \sum_{i=0}^{6} c_i u_i.$$

Разлагая u_i по формуле Тейлора в окрестности узла 0, будем иметь, учитывая, что координаты точек 1, 2, ..., 6 соответственно будут (см. стр. 422).

$$(x+h,\ y),\ \left(x+\frac{h}{2},\ y+\frac{h\sqrt{3}}{2}\right),\ \left(x-\frac{h}{2},\ y+\frac{h\sqrt{3}}{2}\right),\ (x-h,\ y),$$

$$\left(x-\frac{h}{2},\ y-\frac{h\sqrt{3}}{2}\right),\ \left(x+\frac{h}{2},\ y-\frac{h\sqrt{3}}{2}\right),$$

$$u_1 = u_0+\left\{hu'_x + \frac{h^2}{2!}u''_{x^2} + \frac{h^3}{3!}u'''_{x^3} + \frac{h^4}{4!}u^{(4)}_{x^4}+ \cdots\right\}_0,$$

$$u_2 = u_0+\left\{h\left(\frac{1}{2}\frac{\partial}{\partial x}+\frac{\sqrt{3}}{2}\frac{\partial}{\partial y}\right)u + \frac{h^2}{2!}\left(\frac{1}{2}\frac{\partial}{\partial x}+\frac{\sqrt{3}}{2}\frac{\partial}{\partial y}\right)^2 u + \frac{h^3}{3!}\left(\frac{1}{2}\frac{\partial}{\partial x}+\frac{\sqrt{3}}{2}\frac{\partial}{\partial y}\right)^3 u + \frac{h^4}{4!}\left(\frac{1}{2}\frac{\partial}{\partial x}+\frac{\sqrt{3}}{2}\frac{\partial}{\partial y}\right)^4 u+ \cdots\right\}_0,$$

$$u = u_0+\left\{h\left(-\frac{1}{2}\frac{\partial}{\partial x}+\frac{\sqrt{3}}{2}\frac{\partial}{\partial y}\right)u + \frac{h^2}{2!}\left(-\frac{1}{2}\frac{\partial}{\partial x}+\frac{\sqrt{3}}{2}\frac{\partial}{\partial y}\right)^2 u + \frac{h^3}{3!}\left(-\frac{1}{2}\frac{\partial}{\partial x}+\frac{\sqrt{3}}{2}\frac{\partial}{\partial y}\right)^3 u + \frac{h^4}{4!}\left(-\frac{1}{2}\frac{\partial}{\partial x}+\frac{\sqrt{3}}{2}\frac{\partial}{\partial y}\right)^4 u+ \cdots\right\}_0,$$

$$u_4 = u_0+\left\{-hu'_x + \frac{h^2}{2!}u''_{x^2} - \frac{h^3}{3!}u'''_{x^3} + \frac{h^4}{4!}u^{(4)}_{x^4}- \cdots\right\}_0,$$

$$u_5 = u_0+\left\{h\left(-\frac{1}{2}\frac{\partial}{\partial x}-\frac{\sqrt{3}}{2}\frac{\partial}{\partial y}\right)u + \frac{h^2}{2!}\left(-\frac{1}{2}\frac{\partial}{\partial x}-\frac{\sqrt{3}}{2}\frac{\partial}{\partial y}\right)^2 u + \frac{h^3}{3!}\left(-\frac{1}{2}\frac{\partial}{\partial x}-\frac{\sqrt{3}}{2}\frac{\partial}{\partial y}\right)^3 u + \frac{h^4}{4!}\left(-\frac{1}{2}\frac{\partial}{\partial x}-\frac{\sqrt{3}}{2}\frac{\partial}{\partial y}\right)^4 u+ \cdots\right\}_0,$$

$$u_6 = u_0+\left\{h\left(\frac{1}{2}\frac{\partial}{\partial x}-\frac{\sqrt{3}}{2}\frac{\partial}{\partial y}\right)u + \frac{h^2}{2!}\left(\frac{1}{2}\frac{\partial}{\partial x}-\frac{\sqrt{3}}{2}\frac{\partial}{\partial y}\right)^2 u + \frac{h^3}{3!}\left(\frac{1}{2}\frac{\partial}{\partial x}-\frac{\sqrt{3}}{2}\frac{\partial}{\partial y}\right)^3 u + \frac{h^4}{4!}\left(\frac{1}{2}\frac{\partial}{\partial x}-\frac{\sqrt{3}}{2}\frac{\partial}{\partial y}\right)^4 u+ \cdots\right\}_0.$$

Учитывая симметрию оператора Лапласа и симметричное расположение рассматриваемых узлов, можно положить $c_1 = c_2 = \ldots = c_6$. Тогда

$$lu_0 = c_0 u_0 + c_1 \sum_{i=1}^{6} u_i = (c_0 + 6c_1) u_0 +$$

$$+ c_1 \left\{ \frac{3h^2}{2} (u''_{x^2} + u''_{y^2})_0 + \frac{9h^4}{4 \cdot 4!} (u^{(4)}_{x^4} + 2u^{4}_{x^2y^2} + u^{(4)}_{y^4})_0 + \right.$$

$$\left. + \frac{3h^6}{16 \cdot 6!} (11u^{(6)}_{x^6} + 15u^{(6)}_{x^4y^2} + 45u^{(6)}_{x^2y^4} + 9u^{(6)}_{y^{(6)}})_0 + \ldots \right\}.$$

Для того чтобы lu_0 аппроксимировало оператор Лапласа, нужно положить

$$c_0 + 6c_1 = 0,$$

$$\frac{3h^2}{2} c_1 = 1,$$

откуда

$$c_0 = -\frac{4}{h^2}; \quad c_1 = \frac{2}{3h^2}.$$

Итак,

$$lu_0 = \frac{2}{3h^2} [u_1 + u_2 + u_3 + u_4 + u_5 + u_6 - 6u_0] =$$

$$= (u''_{x^2} + u''_{y^2})_0 + \frac{h^2}{16} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right) (u''_{x^2} + u''_{y^2})_0 + R.$$

Если разлагать u_i по формуле Тейлора с остаточным членом в производных шестого порядка, то

$$|R| \leqslant \frac{2}{3h^2} \frac{h^6}{6!} \left[2 + 4 \left(\frac{1+\sqrt{3}}{2} \right)^6 \right] M_6 = \frac{10 + 5\sqrt{3}}{6!} M_6 h^4 < \frac{h^4}{36} M_6.$$

Уравнение Лапласа $\Delta u = 0$ аппроксимируется разностным уравнением

$$u_1 + u_2 + u_3 + u_4 + u_5 + u_6 - 6u_0 = 0 \tag{17}$$

с точностью до h^4, а уравнение Пуассона $\Delta u = f$ аппроксимируется разностным уравнением

$$u_1 + u_2 + u_3 + u_4 + u_5 + u_6 - 6u_0 = \frac{3h^2}{2} f_0 + \frac{3h^4}{32} (\Delta f)_0 \tag{18}$$

с точностью h^4, а разностным уравнением

$$u_1 + u_2 + u_3 + u_4 + u_5 + u_6 - 6u_0 = \frac{3h^2}{2} f_0 \tag{19}$$

с точностью до h^2.

В заключение приведем пример разностной аппроксимации уравнения

$$\frac{\partial^4 u}{\partial x^4} + 2\frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^4 u}{\partial y^4} = f(x, y) \tag{20}$$

для случая квадратной сетки с шагом h при выборе узлов, отмеченных на рис. 31. В силу симметрии расположения узлов и симметрии уравнения относительно x и y можно искать разностную аппроксимацию в виде

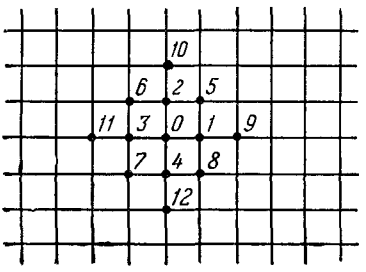

Рис. 31.

$$lu_0 = c_0 u_0 + c_1 \sum_{i=1}^{4} u_i + c_2 \sum_{i=5}^{8} u_i + c_3 \sum_{i=9}^{12} u_i.$$

При наличии производных шестого порядка у функции u разложение по формуле Тейлора дает

$$u_1 + u_2 + u_3 + u_4 = 4u_0 + h^2\left(u''_{x^2} + u''_{y^2}\right)_0 + \frac{h^4}{12}\left(u^{(\mathrm{IV})}_{x^4} + u^{(\mathrm{IV})}_{y^4}\right)_0 + R^{(1)},$$

$$|R^{(1)}| \leqslant \frac{4M_6 h^6}{6!},$$

$$u_5 + u_6 + u_7 + u_8 = 4u_0 + 2h^2\left(u''_{x^2} + u''_{y^2}\right)_0 + \frac{h^4}{6}\left(u^{(4)}_{x^4} + 6u^{(4)}_{x^2y^2} + u^{(4)}_{y^4}\right)_0 + R^{(2)},$$

$$|R^{(2)}| \leqslant \frac{4 \cdot 2^6 M_6 h^6}{6!},$$

$$u_9 + u_{10} + u_{11} + u_{12} = 4u_0 + 4h^2\left(u''_{x^2} + u''_{y^2}\right)_0 + \frac{2 \cdot 2^4 h^4}{4!}\left(u^{(4)}_{x^4} + u^{(4)}_{y^4}\right)_0 + R^{(3)},$$

$$|R^{(3)}| \leqslant \frac{4 \cdot 2^6 \cdot M_6 h^6}{6!},$$

откуда

$$lu_0 = (c_0 + 4c_1 + 4c_2 + 4c_3)u_0 + h^2(c_1 + 2c_2 + 4c_3)\left(u''_{x^2} + u''_{y^2}\right)_0 +$$

$$+ \frac{2h^4}{4!}(c_1 + 2c_2 + 16c_3)\left(u^{(4)}_{x^4} + u^{(4)}_{y^4}\right)_0 + \frac{h^4}{6} 6c_2 \left(u^{(4)}_{x^2y^2}\right)_0 + R,$$

где

$$R = c_1 R^{(1)} + c_2 R^{(2)} + c_3 R^{(3)}.$$

Чтобы разностный оператор lu_0 аппроксимировал дифференциальный оператор $\frac{\partial^4 u}{\partial x^4} + 2\frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^4 u}{\partial y^4}$, нужно потребовать выполне-

ния условий

$$c_0 + 4(c_1 + c_2 + c_3) = 0,$$
$$c_1 + 2c_2 + 4c_3 = 0,$$
$$\frac{h^4}{12}(c_1 + 2c_2 + 16c_3) = 1,$$
$$h^4 c_2 = 2,$$

откуда

$$c_0 = \frac{20}{h^4}; \quad c_1 = -\frac{8}{h^4}; \quad c_2 = \frac{2}{h^4}; \quad c_3 = \frac{1}{h^4}.$$

Таким образом,

$$lu_0 = \frac{20u_0}{h^4} - \frac{8}{h^4}(u_1 + u_2 + u_4) + \frac{2}{h^4}(u_5 + u_6 + u_7 + u_8) +$$
$$+ \frac{1}{h^4}(u_9 + u_{10} + u_{11} + u_{12}) = \left(u^{(4)}_{x^4} + 2u^{(4)}_{x^2y^2} + u^{(4)}_{y^4}\right)_0 + R$$

где

$$|R| = \frac{1}{h^4}|-8R^{(1)} + 2R^{(2)} + R^{(3)}| \leqslant \frac{10}{9} M_6 h^2.$$

Следовательно, разностное уравнение

$$20u_0 - 8(u_1 + u_2 + u_3 + u_4) + 2(u_5 + u_6 + u_7 + u_8) +$$
$$+ (u_9 + u_{10} + u_{11} + u_{12}) = h^4 f_0 \quad (21)$$

аппроксимирует уравнение (20) с точностью до h^2.

Этих примеров достаточно для уяснения способов построения разностных аппроксимаций дифференциальных уравнений, и, следуя им, читатель может построить разностные аппроксимации для других конкретных уравнений, имеющие необходимую точность. Необходимо только помнить, что, строя разностную аппроксимацию той или другой точности, нужно предполагать, что решение уравнения имеет необходимое число производных, а это накладывает определенные требования на коэффициенты уравнения, на область и на функции, входящие в краевые условия. Если они таковы, что решение может иметь производные только до какого-то определенного порядка, то не имеет никакого смысла при решении задачи методом сеток использовать аппроксимации более высокого порядка, так как их использование усложнит работу, но отнюдь не улучшит результата.

3. Аппроксимация граничных условий. Решая методом сеток краевую задачу для дифференциального уравнения в частных производных, мы заменяем заданную область G с границей Γ, на которой заданы граничные условия, сеточной областью G^* с границей Γ^*,

состоящей из граничных узлов. Как правило, граничные узлы не будут находиться на Γ. Поэтому возникает задача замены граничных условий для дифференциального уравнения граничными условиями для разностного уравнения, составляемого только для внутренних узлов, даже в случае задачи Дирихле. При других типах краевых задач условия на границе будут содержать производные искомого решения, которые при переходе к разностным уравнениям должны быть заменены разностными отношениями. Таким образом, почти всегда приходится решать задачу аппроксимации граничных условий.

Остановимся сначала на задаче Дирихле, которую мы уже рассматривали в п. 1. Там был рассмотрен способ аппроксимации граничных условий, заключающийся в том, что значения искомого решения в граничных узлах мы задавали равными значениям заданной функции φ в точках Γ, ближайших к этим узлам. Иногда этот снос выполняют из точки Γ, ближайшей к данному граничному узлу, в направлении одной из осей координат. И в том и в другом случае мы вносим погрешность, величина которой зависит от близости границ Γ и Γ^*. Очевидно, что снесенное в граничный узел значение решения будет отличаться от значения истинного решения краевой задачи в этом узле на величину порядка расстояния этого узла от точки Γ, из которой происходит снос.

Только в том случае, когда все граничные узлы попадут на Γ, перенос граничных условий делать не нужно, и мы не вносим никакой дополнительной погрешности. Поэтому сетку целесообразно выбирать так, чтобы граница Γ^* сеточной области G^* была бы возможно более близка к границе Γ. Для этого иногда целесообразно отказаться от квадратной сетки, а рассматривать прямоугольную сетку или треугольную сетку, или какую-либо другую. Если сетка уже выбрана, выполнен снос граничных условий и методом сеток найдено приближенное решение во всех внутренних узлах, то погрешность, полученная за счет сноса граничного условия, может быть уменьшена. Простейший способ уменьшения этой погрешности заключается в следующем.

По найденным значениям решения во внутренних узлах экстраполируем решение в точки границы Γ, из которых сносились граничные значения в граничные узлы. Находятся разности экстраполированных и заданных значений и в соответствии с ними вносятся поправки значений в граничных узлах. По этим поправкам находятся поправки решения во внутренних узлах путем решения соответствующих однородных разностных уравнений. Это требует повторных пересчетов.

На рис. 32 изображена геометрическая картина описанного процесса уточнения решения. На этом рисунке изображено сечение плоскостью параллельной оси u (в пространстве x, y, u), пересекающейся с плоскостью xOy по узловой линии, на которой

находится граничный узел $\overline{M}$. Предположим, что в узел $\overline{M}$ сносится значение граничной функции φ из точки M, где M — точка пересечения узловой линии с границей Г. Пусть во внутренних узлах M_1, M_2, ..., расположенных на узловой линии, найденные значения решения изображаются ординатами $M_iN_i^0$. Экстраполируя по этим значениям в точку M, получим новое значение u, изображенное отрезком MN_0. График интерполяционного многочлена изображен пунктирной линией. Разность экстраполированного значения и значения $\varphi(M)$ изображается отрезком NN^0. Так как полученное экстраполяцией значение в M больше $\varphi(M)$, то значение в граничном узле $\overline{M}$ изменяем в сторону уменьшения, т. е. вносим поправку $\overline{N}'\overline{N}_0$. Находим поправки в каждом внутреннем узле путем решения однородных разностных уравнений с граничными условиями, равными соответствующим поправкам, внесенным в значения в граничных узлах, и, прибавляя эти поправки к ранее найденным значениям в соответствующих узлах, найдем новое приближение к искомому решению (отрезки $N_i'N_i^0$). Снова выполняем экстраполяцию и т. д.

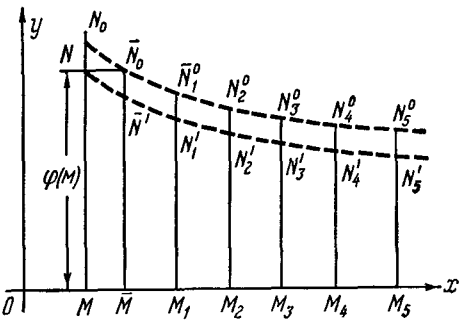

Рис. 32.

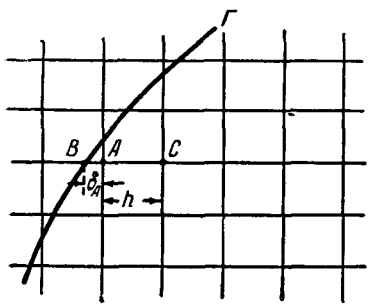

Рис. 33.

Этого пересчета можно избежать, заранее составляя для значений функции в граничных узлах особые уравнения, отличные от уравнений во внутренних узлах. Достаточно удобный для практики способ составления этих уравнений предложил Коллатц. Каждому граничному узлу A (рис. 33) соответствует точка B, лежащая на пересечении границы Г с прямой, принадлежащей сети (узловая прямая), удаленная от A на расстояние $\delta_A < h$, а также ближайший к A внутренний узел сетки C, лежащий на продолжении отрезка BA (предполагается, что сетка достаточно мелкая). Тогда для узла A можно написать уравнение

$$u_A = \frac{\delta_A u_C + h\varphi_B}{\delta_A + h}. \qquad (22)$$

Это уравнение означает, что значение u_A получается линейной интерполяцией значений u_C и φ_B. Получаемая при этом погрешность будет порядка h^2.

Используя при составлении уравнения для значения u в граничном узле большее количество узлов, можно получить более точную аппроксимацию. Например, Микеладзе при решении задачи Дирихле

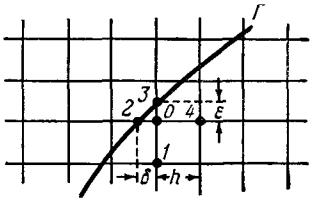

Рис. 34.

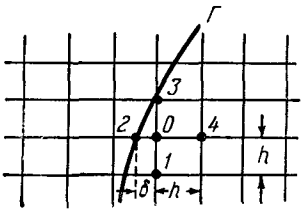

Рис. 35.

для уравнения Пуассона предлагает следующие уравнения для определения значения u_0 в граничном узле 0:

$$u_0 = \frac{h}{h+\delta}\,\frac{h\varphi_2 + \delta u_4}{h+\delta} + \frac{\delta}{h+\delta}\,\frac{u_1+u_3}{2} - \frac{\delta\cdot h^2 f_0}{2(h+\delta)}, \tag{23}$$

(см. рис. 34)

$$u_0 = \frac{\varepsilon}{\varepsilon+\delta}\,\frac{h\varphi_0 + \delta u_4}{h+\delta} + \frac{\delta}{\varepsilon+\delta}\,\frac{h\varphi_3 + \varepsilon u_1}{h+\varepsilon} - \frac{\delta\varepsilon h f_0}{2(\varepsilon+\delta)}. \tag{24}$$

(см. рис. 35)

Порядок погрешности будет в этом случае h^3.

Такого типа формулы могут быть получены с помощью второго способа аппроксимации дифференциального уравнения разностным, описанного в п. 2.

Рассмотрим теперь аппроксимацию других граничных условий, встречающихся при решении краевых задач для эллиптического дифференциального уравнения второго порядка. Наиболее общий вид линейных граничных условий следующий:

$$\left[\alpha\frac{\partial u}{\partial n} + \beta u\right]_\Gamma = \varphi, \tag{25}$$

где α, β и φ — заданные функции точки границы Γ, $\frac{\partial}{\partial n}$ — производная по нормали определенного направления (например, по внутренней нормали), причем предполагается, что $\alpha^2+\beta^2 > 0$. При $\alpha \equiv 0$, $\beta \equiv 1$ получим задачу Дирихле, при $\alpha \equiv 1$, $\beta \equiv 0$ получим задачу Неймана. В зависимости от α и β на различных кусках границы Γ условия могут быть разные (на одних кусках заданы значения решения, на других кусках — значения нормальной производной, на третьих кусках — их линейная комбинация с переменными

коэффициентами). В этом случае переход от заданных граничных условий на Γ к условиям на границе Γ^* сеточной области сильно усложняется из-за наличия в граничных условиях нормальной производной. При этом переходе нормальная производная должна быть заменена через разности значений в узлах сетки. Ограничимся случаем квадратной сетки и укажем простейший способ построения уравнений для граничных узлов. Пусть 0 — один из граничных узлов. Перенесем в этот узел нормаль из точки границы Γ ближайший к 0. Всегда можно найти два таких внутренних или граничных узла, что направления проведенные из узла 0 в них будут образовывать прямой угол. Тогда, обозначая угол направления l_1 с направлением нормали через φ_0 (рис. 36), будем иметь:

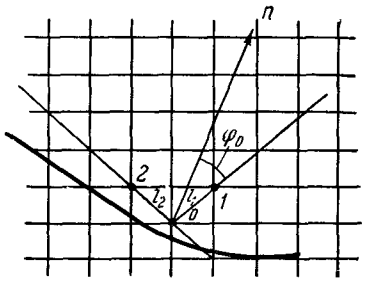

Рис. 36.

$$\frac{\partial u}{\partial n} = \frac{\partial u}{\partial l_1}\cos\varphi_0 + \frac{\partial u}{\partial l_2}\sin\varphi_0.$$

Производную по направлению l_1 приближенно заменим отношением $\frac{u_1 - u_0}{l_1}$, а производную по направлению l_2 — отношением $\frac{u_2 - u_0}{l_2}$, где l_i — расстояние i-го узла от узла 0 ($i = 1, 2$). Тогда будем иметь приближенное равенство

$$\frac{\partial u}{\partial n} \approx \frac{u_1 - u_0}{l_1}\cos\varphi_0 + \frac{u_2 - u_0}{l_2}\sin\varphi_0 \qquad (26)$$

и для граничного узла 0 вместо заданного граничного условия на Γ будем иметь уравнение

$$\alpha_0\left(\frac{u_1 - u_0}{l_1}\cos\varphi_0 + \frac{u_2 - u_0}{l_2}\sin\varphi_0\right) + \beta_0 u_0 = \varphi_0, \qquad (27)$$

где α_0, β_0, φ_0 — значения соответствующих функций в точке границы Γ, ближайшей к узлу 0. При этом мы совершаем погрешность, заменяя нормальную производную линейной комбинацией значений функции в узлах, а также перенося нормаль и функции α, β, φ с Γ в узел 0.

Такие уравнения записываем для каждого граничного узла. Присоединяя их к уравнениям для внутренних узлов, получим систему линейных алгебраических уравнений, в которой число уравнений равно числу неизвестных и равно числу внутренних и граничных узлов.

4. Разрешимость разностных уравнений и способы их решения. Применяя метод сеток для решения краевой задачи для линейного дифференциального уравнения с линейными граничными

условиями, мы получаем систему линейных алгебраических уравнений, число которых равно числу неизвестных. Естественно возникает вопрос о разрешимости этой системы и способах ее решения. Рассмотрим эти вопросы для системы (4), которая получается при решении задачи Дирихле для уравнения (1) методом сеток при разностной аппроксимации, изложенной в п. 1 настоящего параграфа. Эта система имеет вид

$$lu_{ik} = a_{ik}\frac{u_{i+1,\,k} - 2u_{ik} + u_{i-1,\,k}}{h^2} + b_{ik}\frac{u_{i,\,k+1} - 2u_{ik} + u_{i,\,k-1}}{l^2} +$$

$$+ c_{ik}\frac{u_{i+1,\,k} - u_{i-1,\,k}}{2h} + d_{ik}\frac{u_{i,\,k+1} - u_{i,\,k-1}}{2l} + g_{ik}u_{ik} = f_{ik},$$

где число неизвестных u_{ij} равно числу уравнений, т. е. числу внутренних узлов. Перепишем эту систему в другом виде, сгруппировав члены, содержащие одни и те же неизвестные. Получим систему

$$lu_{ik} = A_{ik}u_{i+1,\,k} + B_{ik}u_{i-1,\,k} + C_{ik}u_{i,\,k+1} + D_{ik}u_{i,\,k-1} - E_{ik}u_{ik} = f_{ik},$$

где

$$A_{ik} = \frac{a_{ik}}{h^2} + \frac{c_{ik}}{2h}, \quad B_{ik} = \frac{a_{ik}}{h^2} - \frac{c_{ik}}{2h}, \quad C_{ik} = \frac{b_{ik}}{l^2} + \frac{d_{ik}}{2l},$$

$$D_{ik} = \frac{b_{ik}}{l^2} - \frac{d_{ik}}{2l}, \quad E_{ik} = -\frac{2a_{ik}}{h^2} + \frac{2b_{ik}}{l^2} - g_{ik}.$$

Так как мы предполагаем, что a, b, c, d, g непрерывны в области $G+\Gamma$, причем $a>0$, $b>0$, $g\leqslant 0$ в $G+\Gamma$, то при достаточно малых h, l коэффициенты A_{ik}, B_{ik}, C_{ik}, D_{ik}, E_{ik} будут положительны во всех узлах сеточной области. Будем предполагать, что это условие выполнено. В этом случае имеет место теорема (*принцип максимума*):

Если v_{ik} — какая-либо система значений в узлах сетки и для каждого внутреннего узла $lv_{ik}\geqslant 0$, то v_{ik} во внутренних узлах G^* не могут иметь положительного максимума, а если во всех внутренних узлах $lv_{ik}\leqslant 0$, то v_{ik} во внутренних узлах не могут иметь отрицательного минимума. Исключением является случай $v_{ik}\equiv$ const.

В самом деле, пусть $v_{ik}\not\equiv$ const и во всех внутренних узлах имеет место неравенство $lv_{ik}\geqslant 0$. Предположим, что v_{ik} достигает положительного максимума M в некотором внутреннем узле. Тогда можно найти такой внутренний узел $(i_0,\ k_0)$, в котором $v_{i_0,\,k_0}=M$ и хотя бы в одном соседнем с ним узле значение $v_{i,\,k}$ меньше M. В выражении $lv_{i_0k_0}$ заменим v_{ik} на M, тогда будем иметь строгое неравенство

$$M[A_{i\,k_0} + B_{i_0k_0} + C_{i\,k} + D_{i_0k_0} - E_{i_0k_0}] > 0.$$

Но

$$A_{i_0k_0} + B_{i_0k_0} + C_{i_0k_0} + D_{i_0k_0} - E_{i_0k_0} = g_{i_0k_0}.$$

Таким образом, $g_{i_0 k_0} > 0$, что невозможно. Следовательно, наше допущение было неверно и первая часть утверждения доказана. Вторая часть доказывается совершенно аналогично.

Теперь легко показать, что система (4) имеет решение и притом единственное. Для этого достаточно доказать, что соответствующая однородная система (f_{ik} и все значения в граничных узлах равны нулю) имеет только тривиальное решение, а это сразу следует из принципа максимума. Так как если бы решение однородной системы было отлично от нуля хотя бы в одной точке, то оно должно достигать на Γ^* либо наибольшего положительного значения, либо наименьшего отрицательного значения, что невозможно, так как на Γ^* по условию $u_{ik} \equiv 0$. Следовательно, однородная система имеет лишь тривиальное решение, а неоднородная система (4) имеет одно и только одно решение.

Совершенно аналогично можно доказать разрешимость систем уравнений, которые получаются при решении задачи Дирихле для уравнения Пуассона методом сеток при всех разностных аппроксимациях, которые мы рассмотрели в п. 2, а также системы уравнений, в которой для граничных узлов записываются уравнения по способу Коллатца, так как во всех этих случаях для разностных операторов имеет место принцип максимума [1]).

Для решения получаемых систем могут быть использованы методы, изложенные в главе 6. Часто применяют метод простой итерации или метод Зейделя. Докажем сходимость этих методов применительно к системе (4), дополненной уравнениями Коллатца для определения значений в граничных узлах. Для применения метода простой итерации удобней систему переписать в таком виде:

$$u_{ik} = \frac{A_{ik}}{E_{ik}} u_{i+1,\,k} + \frac{B_{ik}}{E_{ik}} u_{i-1,\,k} + \frac{C_{ik}}{E_{ik}} u_{i,\,k+1} + \frac{D_{ik}}{E_{ik}} u_{i,\,k-1} - \frac{f_{ik}}{E_{ik}}, \quad (28)$$

$$u_A = \frac{\delta_A u_C + h\varphi_B}{h + \delta_A}.$$

В этом случае, когда коэффициент g в уравнении (1) строго отрицателен, а сетка выбрана настолько мелкой, что A_{ik}, B_{ik}, C_{ik}, D_{ik}, E_{ik} положительны, сходимость процесса простой итерации и процесса Зейделя доказывается очень просто, так как тогда в системе (28) все коэффициенты положительны и сумма коэффициентов в правой части каждого уравнения строго меньше единицы, в силу неравенства $\delta_A < h$, а

$$A_{ik} + B_{ik} + C_{ik} + D_{ik} - g_{ik} = E_{ik},$$

т. е.

$$A_{ik} + B_{ik} + C_{ik} + D_{ik} < E_{ik}.$$

[1]) Соответствующие рассуждения аналогичны приведенным на стр. 373 и 374.

Этого достаточно для сходимости обоих процессов итерации, причем скорость сходимости будет определяться величиной

$$q = \max_{A, i, k} \left\{ \frac{\delta_A}{\delta_A + h}, \frac{A_{ik} + B_{ik} + C_{ik} + D_{ik}}{E_{ik}} \right\}$$

(см гл. 6).

Это доказательство не пройдет, если g может обращаться в нуль, так как в этом случае в некоторых уравнениях сумма коэффициентов будет равна единице. Поэтому мы приведем отдельно доказательство сходимости итерационных методов решения системы (28) для случая $g \leqslant 0$.

Начнем с простой итерации. В этом случае последовательные приближения находятся из соотношений

$$u_{ik}^{(n+1)} = \frac{A_{ik}}{E_{ik}} u_{i+1, k}^{(n)} + \frac{B_{ik}}{E_{ik}} u_{i-1, k}^{(n)} + \frac{C_{ik}}{E_{ik}} u_{i, k+1}^{(n)} + \frac{D_{ik}}{E_{ik}} u_{i, k-1}^{(n)} - \frac{f_{ik}}{E_{ik}},$$

$$u_A^{(n+1)} = \frac{\delta_A u_c^{(n)} + h\varphi_B}{\delta_A + h}. \tag{29}$$

Введем обозначения:

$$u_{ij} - u_{ij}^{(n)} = \alpha_{ij}^{(n)}; \quad u_A - u_A^{(n)} = \alpha_A^{(n)}; \quad \max \left\{ \left| \alpha_{ij}^{(n)} \right|, \left| \alpha_A^{(n)} \right| \right\} = M^{(n)},$$

$$\gamma = \min_{i, k} \left\{ \frac{A_{ik}}{E_{ik}}, \frac{B_{ik}}{E_{ik}}, \frac{C_{ik}}{E_{ik}}, \frac{D_{ik}}{E_{ik}} \right\}; \quad \delta = \max \frac{\delta_A}{\delta_A + h} \; (0 < \delta, \; \gamma < 1).$$

Назовем граничные узлы узлами первого разряда. Внутренние узлы, у которых среди соседних имеется хотя бы один граничный узел, назовем узлами второго разряда. Внутренние узлы (не принадлежащие предыдущим разрядам), у которых среди соседних имеется хотя бы один узел второго разряда, назовем узлами третьего разряда и т. д. Таким образом, все граничные и внутренние узлы мы разобьем на конечное число разрядов, причем каждый узел будет принадлежать к одному и только одному разряду. Пусть число разрядов равно m. Тогда из неравенств

$$\left| \alpha_{ik}^{(n+1)} \right| \leqslant \frac{A_{ik}}{E_{ik}} \left| \alpha_{i+1, k}^{(n)} \right| + \frac{B_{ik}}{E_{ik}} \left| \alpha_{i-1, k}^{(n)} \right| + \frac{C_{ik}}{E_{ik}} \left| \alpha_{i, k+1}^{(n)} \right| + \frac{D_{ik}}{E_{ik}} \left| \alpha_{i, k-1}^{(n)} \right|;$$

$$\left| \alpha_A^{(n+1)} \right| \leqslant \frac{\delta_A}{h + \delta_A} \left| \alpha_A^{(n)} \right|; \quad M^{(n+1)} \leqslant M^{(n)}$$

для узлов первого разряда будем иметь неравенство

$$\left| \alpha_A^{(n+1)} \right| \leqslant \delta M^{(n)},$$

для узлов второго разряда — неравенство

$$|\alpha_{ik}^{(n+1)}| \leqslant$$

$$\leqslant \left\{ \frac{A_{ik}+B_{ik}+C_{ik}+D_{ik}}{E_{ik}} - \min\left\{ \frac{A_{ik}}{E_{ik}}, \frac{B_{ik}}{E_{ik}}, \frac{C_{ik}}{E_{ik}}, \frac{D_{ik}}{E_{ik}} \right\} (1-\delta) \right\} M^{(n-1)} \leqslant$$

$$\leqslant [1-\gamma(1-\delta)]\, M^{(n-1)},$$

для узлов третьего разряда — неравенство

$$\alpha_{ik}^{(n+1)}| \leqslant$$

$$\leqslant \left\{ \frac{A_{ik}+B_{ik}+C_{ik}+D_{ik}}{E_{ik}} - \min\left\{ \frac{A_{ik}}{E_{ik}}, \frac{B_{ik}}{E_{ik}}, \frac{C_{ik}}{E_{ik}}, \frac{D_{ik}}{E_{ik}} \right\} \gamma(1-\delta) \right\} M^{(n-2)} \leqslant$$

$$\leqslant [1-\gamma^2(1-\delta)]\, M^{(n-2)},$$

наконец, для узлов m-го разряда — неравенство

$$|\alpha_{ik}^{(n+1)}| \leqslant [1-\gamma^{m-1}(1-\delta)]\, M^{(n-m+1)}$$

откуда будем иметь неравенство

$$M^{(n+1)} \leqslant \beta M^{(n+1-m)},$$

где

$$0 < \beta = \max\{\delta,\ 1-\gamma(1-\delta),\ \ldots,\ 1-\gamma^{m-1}(1-\delta)\} < 1$$

или

$$M^{(km)} \leqslant \beta^k M^{(0)},$$

и так как $\beta^k \to 0$ при $k \to +\infty$, то $M^{(km)} \to 0$ при $k \to +\infty$ и $M^{(n)} \to 0$ при $n \to \infty$, а это означает, что процесс простой итерации сходится.

Рассмотрим теперь сходимость процесса Зейделя в случае, если $g \equiv 0$. Для исследования сходимости этого метода перенумеруем все внутренние и все граничные узлы, которые не находятся на Γ и для которых записываются особые соотношения по Коллатцу следующим образом. За первый узел примем один из граничных узлов, не принадлежащих Γ, или если таковых нет, то один из внутренних узлов, у которого среди соседних есть граничный узел; за второй узел примем один из узлов, среди соседних у которого будет первый узел; за третий примем узел, у которого соседним будет второй узел, и т. д. Перепишем систему уравнений (28) в порядке новой нумерации узлов (т. е. искомых неизвестных) и для ее решения будем применять метод Зейделя. Если мы возьмем фиксированный узел k и в правой части уравнения для этого узла исключим последовательными подстановками значения $(n+1)$-го приближения в узлах предыдущих по номеру, то значение $(n+1)$-го приближения решения в этом узле можно представить в следующем виде:

$$u_k^{(n+1)} = \sum_{i=1}^{N} c_k^{(i)} u_i^{(n)} + \sum_{j=1}^{M} b_k^{(j)} \varphi_j - \sum_{i=1}^{N} d_k^{(i)} f_i \qquad (k = 1,\ 2,\ \ldots,\ N) \quad (30)$$

(N — число узлов, не лежащих на Γ, M — число узлов на Γ). Коэффициенты $c_k^{(i)}$, $b_k^{(i)}$, $d_k^{(i)}$ не зависят от номера n. Все коэффициенты $c_k^{(i)}$ и $b_k^{(j)}$ неотрицательны, так как они получаются сложением и умножением неотрицательных чисел. Далее, при любом k имеет место равенство

$$\sum_{i=1}^{N} c_k^{(i)} + \sum_{j=1}^{M} b_k^{(j)} = 1. \tag{31}$$

Это равенство следует из таких соображений. Коэффициенты $c_k^{(i)}$ и $b_k^{(j)}$ не зависят от граничных значений и правых частей. Поэтому если положить $\varphi \equiv 1$, $f \equiv 0$ и за нулевое приближение $u_k^{(0)}$ принять значения, равные единице, во всех узлах, то при всех k и n будем иметь $u_k^{(n)} \equiv 1$, а отсюда и будет следовать равенство (31). Так как при любом k хотя бы один из коэффициентов $b_k^{(j)} \neq 0$, то $\sum_{i=1}^{N} c_k^{(i)} < 1$.

Пусть $\mu = \max_k \sum_{i=1}^{N} c_k^{(i)} < 1$. Если u_k решение системы (28), то

$$u_k = \sum_{i=1}^{N} c_k^{(i)} u_i + \sum_{j=1}^{M} b_k^{(j)} \varphi_j - \sum_{i=1}^{N} d_k^{(i)} f_i.$$

Вычитая его из равенства (30), получим:

$$u_k^{(n+1)} - u_k = \sum_{i=1}^{N} c_k^{(i)} \left(u_i^{(n)} - u_i\right),$$

откуда

$$\left| u_k^{(n+1)} - u_k \right| \leqslant \sum_{i=1}^{N} c_k^{(i)} \left| u_i^{(n)} - u_i \right|.$$

Отсюда, если ввести обозначение $M^{(n)} = \sup_k \left| u_k^{(n)} - u_k \right|$, будем иметь неравенство

$$\left| u_k^{(n+1)} - u_k \right| \leqslant \sum_{i=1}^{N} c_k^{(i)} M^{(n)} \leqslant \mu M^{(n)},$$

а следовательно,

$$M^{(n+1)} \leqslant \mu M^{(n)}, \quad \text{или} \quad M^{(n+1)} \leqslant \mu^{n+1} M^{(0)}.$$

Так как $\mu < 1$, то $\mu^{n+1} \to 0$ при $n \to \infty$ и $M^{(n+1)} \to 0$ при $n \to \infty$. Таким образом, сходимость метода Зейделя для нашей системы доказана.

5. Оценка погрешности и сходимость метода сеток. Оценим погрешность, которая получается при решении задачи Дирихле для

уравнения (1) п. 1 методом сеток. При этом будем предполагать, что решение $u(x, y)$ уравнения

$$Lu = a\frac{\partial^2 u}{\partial x^2} + b\frac{\partial^2 u}{\partial y^2} + c\frac{\partial u}{\partial x} + d\frac{\partial u}{\partial y} + gu = f, \tag{1}$$

удовлетворяющее на границе Γ области G условию

$$u|_\Gamma = \varphi, \tag{2}$$

имеет в области G ограниченные и непрерывные вплоть до границы Γ производные до четвертого порядка включительно. Для этого коэффициенты уравнения a, b, c, d, g, функции f и φ должны иметь производные до определенного порядка, а граница Γ должна быть достаточно гладкой. Кроме того, будем предполагать, что a, b, c, d, g удовлетворяют требованиям п. 1 и в каждой точке области G выполнено неравенство

$$\frac{a}{p^2} + \frac{b}{q^2} - \frac{|c|}{p} - \frac{|d|}{q} > 0, \tag{3}$$

где p и q — полуоси эллипса с осями, параллельными координатным осям с центром в некоторой точке (x_0, y_0), целиком содержащего внутри себя область G.

Пусть эта задача решается методом сеток с прямоугольной сеткой с шагами h и l по оси x и y соответственно, при этом во внутренних узлах используется разностное уравнение (4), а значения решения в граничных узлах полагаются равными значениям граничной функции φ в соответствующих ближайших к ним точках. Таким образом, обозначая значения приближенного решения задачи через U_{ij}, будем иметь для них систему уравнений

$$lU_{ik} = A_{ik}U_{i+1,k} + B_{ik}U_{i-1,k} + C_{ik}U_{i,k+1} + D_{ik}U_{i,k-1} - E_{ik}U_{ik} = f_{ik}$$

для внутренних узлов,

$$U_{ik} = \varphi_{ik} \tag{4}$$

для граничных узлов, где

$$\left.\begin{aligned} A_{ik} &= \frac{a_{ik}}{h^2} + \frac{c_{ik}}{2h}; \quad B_{ik} = \frac{a_{ik}}{h^2} - \frac{c_{ik}}{2h}; \quad C_{ik} = \frac{b_{ik}}{l^2} + \frac{d_{ik}}{2l}; \\ D_{ik} &= \frac{b_{ik}}{l^2} - \frac{d_{ik}}{2l}; \quad E_{ik} = \frac{2a_{ik}}{h^2} + \frac{2b_{ik}}{l^2} - g_{ik}. \end{aligned}\right\} \tag{5}$$

Будем предполагать h и l настолько малыми, что A_{ik}, B_{ik}, C_{ik}, D_{ik}, E_{ik} положительны во всех узлах (i, k). Мы видели (п. 2), что если функция $u(x, y)$ имеет в G непрерывные и ограниченные производные до четвертого порядка включительно, то

$$lu_{ik} = (Lu)_{(i,k)} + R_{ik}(u), \tag{6}$$

где

$$\left.\begin{aligned} &|R_{ik}(u)| \leqslant \frac{h^2}{12}\{(a_{ik}+\alpha^2 b_{ik})M_4 + 2(|c_{ik}|+\alpha^2|d_{ik}|)M_3\},\\ &M_3 = \max_G\left\{\left|\frac{\partial^3 u}{\partial x^3}\right|, \left|\frac{\partial^3 u}{\partial y^3}\right|\right\}; \quad M_4 = \max_G\left\{\left|\frac{\partial^4 u}{\partial x^4}\right|, \left|\frac{\partial^4 u}{\partial y^4}\right|\right\}, \quad \alpha = \frac{l}{h}. \end{aligned}\right\} \quad (7)$$

Если (i_0, k_0) — граничный узел, а B — ближайшая к нему точка границы Γ, из которой сносится значение φ, то, используя формулу конечных приращений, получим:

$$u_{i_0, k_0} - u_B = \frac{\partial \tilde{u}}{\partial x}\delta_1 + \frac{\partial \tilde{u}}{\partial y}\delta_2$$

или

$$|u_{i,k_0} - \varphi(B)| \leqslant 2\delta M_1 \quad \left(\delta = \max(h, l); \quad M_1 = \max_G\left\{\left|\frac{\partial u}{\partial x}\right|, \left|\frac{\partial u}{\partial y}\right|\right\}\right).$$

Если ввести обозначение ε_{ik} для погрешности, т. е.

$$u_{ik} - U_{ik} = \varepsilon_{ik},$$

то для граничных узлов будет иметь место неравенство

$$|\varepsilon_{ik}| \leqslant 2M_1\delta.$$

Вычитая из уравнения (6) соответствующее уравнение (4), для погрешности ε_{ik} во внутренних узлах получим разностное уравнение

$$l\varepsilon_{ik} = R_{ik}(u). \quad (32)$$

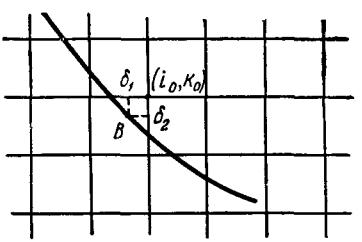

Рис. 37.

Решение системы (32) будем искать в виде

$$\varepsilon_{ik} = \xi_{ik} + \eta_{ik}, \quad (33)$$

где

$l\eta_{ik} = 0$ для внутренних узлов,

$\eta_{ik} = \varepsilon_{ik}$ для граничных узлов

и

$l\xi_{ik} = R_{ik}(u)$ для внутренних узлов,

$\xi_{ik} = 0$ для граничных узлов.

На основании свойства максимума $|\eta_{ik}|$ достигает наибольшего значения на границе сеточной области; таким образом,

$$|\eta_{ik}| \leqslant 2M_1\delta.$$

Для оценки ξ_{ik} докажем следующее утверждение:

Если на сетке заданы две системы значений v_{ij} и V_{ij} такие, что во всех внутренних узлах имеет место неравенство

$$lV_{ij} \leqslant -|lv_{ij}|,$$

а в граничных узлах $V_{ij} \geqslant |v_{ij}|$, *то всюду в сеточной области* G^* *имеет место неравенство*

$$V_{ij} \geqslant |v_{ij}|.$$

В самом деле, неравенство $lV_{ij} \leqslant -|lv_{ij}|$ эквивалентно двум неравенствам:

$$l(V_{ij} - v_{ij}) \leqslant 0; \quad l(V_{ij} + v_{ij}) \leqslant 0,$$

а неравенство $V_{ij} \geqslant |v_{ij}|$ на Γ^* эквивалентно неравенствам

$$V_{ij} - v_{ij} \geqslant 0; \quad V_{ij} + v_{ij} \geqslant 0 \quad \text{на } \Gamma^*.$$

Но в силу принципа максимума и минимума (см. п. 4) $V_{ij} - v_{ij}$ и $V_{ij} + v_{ij}$ в этом случае не могут достигать во внутренних узлах отрицательного минимума, а на Γ^* они неотрицательны. Следовательно, $V_{ij} - v_{ij} \geqslant 0$ и $V_{ij} + v_{ij} \geqslant 0$ во всех внутренних узлах, т. е.

$$V_{ij} \geqslant |v_{ij}|$$

всюду в $G^* + \Gamma^*$.

Рассмотрим теперь вспомогательную функцую

$$W = \lambda\left(1 - \frac{(x-x_0)^2}{p^2} - \frac{(y-y_0)^2}{q^2}\right)$$

с неопределенным пока коэффициентом $\lambda > 0$. На основании равенства (6)

$$lW_{ik} = [L(W)]_{(i,k)} + R_{ik}(W) = (L(W))_{(i,k)},$$

так как $R_{ik}(W) \equiv 0$.

Отсюда

$$lW_{ik} = -2\lambda\left[\frac{a_{ik}}{p^2} + \frac{b_{ik}}{q^2} + \frac{x_i - x_0}{p^2} c_{ik} + \right.$$

$$\left. + \frac{y_k - y_0}{q^2} d_{ik} - \frac{g_{ik}}{2}\left(1 - \frac{(x_i - x_0)^2}{p^2} - \frac{(y_k - y_0)^2}{q^2}\right)\right] \leqslant$$

$$\leqslant -2\lambda\left[\frac{a_{ik}}{p^2} + \frac{b_{ik}}{q^2} - \frac{|c_{ik}|}{p} - \frac{|d_{ik}|}{p}\right].$$

Выберем λ так, чтобы во всех внутренних узлах имело место неравенство

$$|R_{ik}(u)| \leqslant 2\lambda\left[\frac{a_{ik}}{p^2} + \frac{b_{ik}}{q^2} - \frac{|c_{ik}|}{p} - \frac{|d_{ik}|}{q}\right].$$

Для этого достаточно положить

$$\lambda = \frac{h^2}{24} \frac{\max\limits_G \{(a + \alpha^2 b) M_4 + 2(|c| + \alpha^2 |d|) M_3\}}{\min\limits_G \left(\frac{a}{p^2} + \frac{b}{q^2} - \frac{|c|}{p} - \frac{|d|}{q}\right)}.$$

Тогда на основании доказанного утверждения имеет место неравенство

$$|\xi_{ij}| \leqslant W_{ij}$$

в каждом внутреннем узле, так как $\xi_{ij} = 0$ в граничных узлах, а $W_{ij} > 0$. Таким образом,

$$|\varepsilon_{ik}| \leqslant |\eta_{ik}| + |\xi_{ik}| \leqslant 2M_1\delta + \frac{h^2}{24} \frac{\max\limits_G \{(a + \alpha^2 b) M_4 + 2(|c| + \alpha^2 |d|) M_3\}}{\min\limits_G \left[\frac{a}{p^2} + \frac{b}{q^2} - \frac{|c|}{p} - \frac{|d|}{q}\right]} \times$$

$$\times \left(1 - \frac{(x_i - x_0)^2}{p^2} - \frac{(y_k - y_0)^2}{q^2}\right) \leqslant$$

$$\leqslant 2M_1\delta + \frac{h^2}{24} \frac{\max\limits_G \{(a + \alpha^2 b) M_4 + 2(|c| + \alpha^2 |d|) M_3\}}{\min\limits_G \left[\frac{a}{p^2} + \frac{b}{q^2} - \frac{|c|}{p} - \frac{|d|}{q}\right]}. \tag{34}$$

В случае уравнения Пуассона и квадратной сетки оценка записывается значительно проще, так как $a = b = 1$, $c = d = 0$

$$|\varepsilon_{ik}| \leqslant 2M_1 h + \frac{h^2}{12} \frac{p^2 q^2 M_4}{p^2 + q^2}, \tag{35}$$

и если вместо эллипса взять круг радиуса R, то

$$|\varepsilon_{ik}| \leqslant 2M_1 h + \frac{h^2 R^2}{24} M_4. \tag{36}$$

Если для граничных узлов составляются уравнения по Коллатцу и сетка квадратная, то имеет место оценка

$$\left.\begin{aligned} |\varepsilon_{ik}| &\leqslant 3h^2 M_2 + \frac{h^2}{24} \frac{\max\limits_G \{(a + b) M_4 + 2(|c| + |d|) M_3\}}{\min\limits_G \left(\frac{a}{p^2} + \frac{b}{q^2} - \frac{|c|}{p} - \frac{|d|}{q}\right)} \times \\ &\times \left(1 - \frac{(x_i - x_0)^2}{p^2} - \frac{(y_k - y_0)^2}{q^2}\right), \\ M_2 &= \max_G \left\{\left|\frac{\partial^2 u}{\partial x^2}\right|, \quad \left|\frac{\partial^2 u}{\partial x\, \partial y}\right|, \quad \left|\frac{\partial^2 u}{\partial y^2}\right|\right\} \end{aligned}\right\} \tag{37}$$

(см. Коллатц, гл. IV, стр. 281).

Из этих оценок следует, что если мы будем неограниченно измельчать сетку, то последовательность решений, получаемых методом сеток, будет сходиться равномерно к точному решению задачи Дирихле.

Приведенные оценки имеют тот недостаток, что они содержат максимумы модулей производных от искомого решения. Они должны быть определены из дополнительных соображений или получены приближенно по найденным значениям решения в узлах заменой производных разностными отношениями.

На практике для оценки погрешности решения, полученного методом сеток, часто используют *принцип Рунге*, заключающийся в следующем.

Пусть известно, что порядок погрешности решения при использовании квадратной сетки с шагом h есть n, т. е. в точке (x, y) погрешность $\varepsilon_h(x, y)$ может быть приближенно представлена в виде

$$\varepsilon_h(x, y) \approx k(x, y)\, h^n,$$

где $k(x, y)$ от h не зависит. Пусть $U_h(x, y)$ и $U_{2h}(x, y)$ — решения краевой задачи, полученные методом сеток при шаге h и $2h$ соответственно. Тогда, если $u(x, y)$ есть точное решение, то

$$u(x, y) = U_h(x, y) + \varepsilon_h(x, y); \quad u(x, y) = U_{2h}(x, y) + \varepsilon_{2h}(x, y)$$

или

$$U_h - U_{2h} = \varepsilon_{2h} - \varepsilon_h.$$

Далее, по условию

$$\varepsilon_h \approx k(x, y)\, h^n; \quad \varepsilon_{2h} \approx k(x, y)\, 2^n h^n \approx 2^n \varepsilon_h(x, y).$$

Следовательно, $U_h - U_{2h} \approx [2^n - 1]\, \varepsilon_h(x, y)$, откуда

$$\varepsilon_h(x, y) \approx \frac{U_h - U_{2h}}{2^n - 1}. \tag{38}$$

При решении задачи Дирихле для уравнения (1) методом сеток, когда в граничных узлах решение определяется из уравнений Коллатца, погрешность имеет относительно h порядок $n = 2$. Следовательно,

$$\varepsilon_h \approx \frac{U_h - U_{2h}}{3}.$$

Пример. Найдем методом сеток решение задачи Дирихле для уравнения Лапласа $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$ в квадрате со стороной, равной единице, при следующих граничных условиях:

$$u(0, y) = 0; \quad u(1, y) = \sin \pi y; \quad u(x, 0) = 0; \quad u(x, 1) = 0.$$

Возьмем квадратную сетку с шагом $h = 0{,}125$.

Используя простейшую разностную схему, получим следующую систему уравнений для отыскания значений u в узлах сетки:

$$u_{i+1,\,j} + u_{i-1,\,j} + u_{i,\,j+1} + u_{i,\,j-1} - 4u_{ij} = 0 \qquad (i, j = 1, 2, \ldots, 7),$$

$$u_{0j} = u_{i0} = u_{i,\,8} = 0; \quad u_{8j} = \sin \pi j h \qquad (i, j = 0, 1, 2, \ldots, 8).$$

Решение этой системы в общем случае $h = \frac{1}{n}$, где n — целое > 0, найдем с помощью аналога метода Фурье. Будем искать частные решения этой системы вида

$$u_{ij}^{(r)} = \varphi_i^{(r)} \sin r \pi j h \qquad \left(i, j = 0, 1, 2, \ldots, n;\ r = 1, 2, \ldots, n-1;\ h = \frac{1}{n}\right).$$

Это решение удовлетворяет граничным условиям

$$u^{(r)}_{i,0} = u^{(r)}_{i,n} = 0,$$

но не удовлетворяет граничным условиям при $i=0$ и $i=n$. Функцию $\varphi^{(r)}_i$ целочисленного аргумента i найдем из условия, чтобы были удовлетворены уравнения для внутренних узлов. Подстановка дает

$$\left(\varphi^{(r)}_{i+1}+\varphi^{(r)}_{i-1}\right)\sin r\pi jh+\varphi^{(r)}_i\left[\sin r\pi(j+1)h+\sin r\pi(j-1)h-4\sin r\pi jh\right]=$$
$$=\left[\varphi^{(r)}_{i+1}+\varphi^{(r)}_{i-1}-2(2-\cos r\pi h)\,\varphi^{(r)}_i\right]\sin r\pi jh=0$$

или

$$\varphi^{(r)}_{i+1}+\varphi^{(r)}_{i-1}-2(2-\cos r\pi h)\,\varphi^{(r)}_i=0 \qquad (i=0,\ 1,\ 2,\ \ldots,\ n).$$

Общее решение этого однородного разностного уравнения имеет вид

$$\varphi^{(r)}_i = A_r\lambda^i_{r,1}+B_r\lambda^i_{r,2},$$

где $\lambda_{r,1}$ и $\lambda_{r,2}$ — корни характеристического уравнения

$$\lambda^2-2(2-\cos r\pi h)\lambda+1=0,$$

а A_r и B_r — произвольные постоянные.

Таким образом,

$$u^{(r)}_{ij}=\left(A_r\lambda^i_{r,1}+B_r\lambda^i_{r,2}\right)\sin r\pi jh.$$

Решение системы разностных уравнений, удовлетворяющее всем граничным условиям, будем искать в виде

$$u_{ij}=\sum_{r=1}^{n-1}\left(A_r\lambda^i_{r,1}+B_r\lambda^i_{r,2}\right)\sin r\pi jh,$$

где постоянные A_r, B_r $(r=1,\ 2,\ \ldots,\ n-1)$ подберем так, чтобы были удовлетворены граничные условия при $i=0$, $i=n$. В нашем случае из граничных условий имеем:

$$\sum_{r=1}^{n}(A_r+B_r)\sin r\pi jh=0;\quad \sum_{r=1}^{n-1}\left(A_r\lambda^n_{r,1}+B_r\lambda^n_{r,2}\right)\sin r\pi jh=\sin\pi jh.$$

Функции $\sin r\pi jh$ образуют на множестве $\{0,\ 1,\ 2,\ \ldots,\ n\}$ полную ортогональную систему функций, т. е. $\sum\limits_{j=0}^{n}\sin r\pi jh\sin l\pi jh=0$ при $r\neq l$. В нашем случае это дает

$$A_r+B_r=0 \qquad (r=1,\ 2,\ \ldots,\ n-1),$$
$$A_1\lambda^n_{1,1}+B_1\lambda^n_{1,2}=1,$$
$$A_r\lambda^n_{r,1}+B_r\lambda^n_{r,2}=0 \qquad (r=2,\ 3,\ \ldots,\ n-1),$$

откуда

$$A_1 = -B_1 = \frac{1}{\lambda_{11}^n - \lambda_{12}^n}, \qquad A_r = B_r = 0 \qquad (r = 2, 3, \ldots, n-1)$$

и

$$u_{ij} = \frac{\lambda_{11}^i - \lambda_{12}^i}{\lambda_{11}^n - \lambda_{12}^n} \sin \pi jh \qquad (i, j = 0, 1, 2, \ldots, n).$$

В нашем случае $n = 8$. Ниже приведена таблица значений u_{ij} для $i = 0, 1, 2, \ldots, 8$ и $j = 0, 1, 2, 3, 4$. Значения u_{ij} при $j = 5, 6, 7, 8$ не приводятся, так как u_{ij} и точное решение $u(x, y)$ симметричны относительно прямой $y = 0{,}5 = 4h$. (Значения u_{ij} даны в единицах четвертого десятичного знака.)

j \ i	0	1	2	3	4	5	6	7	8
0	0	0	0	0	0	0	0	0	0
1	0	137	295	498	777	1174	1749	2591	3827
2	0	253	545	920	1435	2169	3232	4787	7071
3	0	331	713	1203	1876	2834	4223	6255	9239
4	0	358	771	1302	2030	3067	4571	6771	10000

Если при той же сетке применить более точную разностную аппроксимацию оператора Лапласа

$$(\Delta u)_{ij} \approx \frac{4(u_{i+1,\,j} + u_{i-1,\,j} + u_{i,\,j+1} + u_{i,\,j-1}) +}{6h^2} \rightarrow$$
$$\rightarrow \frac{+ (u_{i+1,\,j+1} + u_{i-1,\,j+1} + u_{i-1,\,j-1} + u_{i+1,\,j-1}) - 20u_{ij}}{6h^2},$$

то получим следующую систему уравнений для u_{ij}:

$$4(u_{i+1,\,j} + u_{i-1,\,j} + u_{i,\,j+1} + u_{i,\,j-1}) +$$
$$+ (u_{i+1,\,j+1} + u_{i-1,\,j+1} + u_{i-1,\,j-1} + u_{i+1,\,j-1}) - 20u_{ij} = 0$$
$$(i, j = 1, 2, \ldots, 7),$$
$$u_{0j} = u_{i0} = u_{i8} = 0, \quad u_{8j} = \sin \pi jh \qquad (i, j = 0, 1, 2, \ldots, 8).$$

Точно так же, как и раньше, можно показать, что решение этой системы имеет вид

$$u_{ij} = \frac{\mu_{11}^i - \mu_{12}^i}{\mu_{11}^8 - \mu_{12}^8} \sin \pi jh \qquad (i, j = 0, 1, 2, \ldots, 8),$$

где μ_{11} и μ_{12} — корни уравнения

$$\mu^2 - 2\,\frac{5 - 2\cos \pi h}{2 + \cos \pi h}\,\mu + 1 = 0.$$

Таблица значений этого решения приведена ниже:

j \ i	0	1	2	3	4	5	6	7	8
0	0	0	0	0	0	0	0	0	0
1	0	134	288	487	763	1157	1732	2578	3827
2	0	247	532	900	1409	2138	3201	4764	7071
3	0	322	695	1176	1841	2793	4.86	6224	9239
4	0	349	752	1273	1993	3024	4527	6737	10000

Точное решение поставленной задачи Дирихле имеет вид

$$u(x, y) = \frac{\operatorname{sh} \pi y}{\operatorname{sh} \pi} \sin \pi x.$$

Таблица его значений в узлах сетки приведена ниже:

j \ i	0	1	2	3	4	5	6	7	8
0	0	0	0	0	0	0	0	0	0
1	0	134	288	488	763	1158	1734	2581	3827
2	0	247	532	901	1410	2140	3204	4769	7071
3	0	323	696	1177	1843	2796	4186	6231	9239
4	0	349	753	1274	1995	3027	4531	6744	10000

Приведем еще таблицу отклонений полученных приближенных решений задачи от значений точного решения во внутренних узлах сетки (в единицах четвертого десятичного знака):

	Простейшая схема							Уточненная схема						
j \ i	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1	3	8	10	14	16	15	10	0	0	−1	0	−1	−2	−3
2	6	13	19	25	29	28	18	0	0	−1	−1	−2	−3	−5
3	8	17	26	33	38	37	24	−1	−1	−1	−2	−3	0	−7
4	9	18	28	35	40	40	27	0	−1	−1	−2	−3	−4	−7

Из этой таблицы видно, что уточненная схема дала значительно лучший результат. Если относительная погрешность результата, полученного в простейшей схеме, достигает 2,5%, то при уточненной схеме она не превосходит 0,3%.

§ 3. Метод сеток решения линейных дифференциальных уравнений гиперболического типа

В этом параграфе мы изложим метод сеток решения основных задач для дифференциальных уравнений вида [1])

$$Lu = a\frac{\partial^2 u}{\partial x^2} - b\frac{\partial^2 u}{\partial y^2} + c\frac{\partial u}{\partial y} + d\frac{\partial u}{\partial y} + gu = f, \tag{1}$$

где a, b, c, d, g, f — заданные функции переменных x и y и $ab > 0$ в рассматриваемой области плоскости x, y. Для определенности будем считать a и b положительными.

Мы рассмотрим решение двух основных задач:

1. Решение *задачи Коши*, заключающейся в отыскании решения $u(x, y)$ уравнения (1) в области $G\{y \geqslant 0\}$, удовлетворяющего начальным условиям:

$$u|_{y=0} = \varphi(x); \quad \left.\frac{\partial u}{\partial y}\right|_{y=0} = \psi(x) \qquad (-\infty < x < +\infty), \tag{2}$$

где φ и ψ — заданные функции переменного x.

2. Решение *смешанной задачи*, заключающейся в отыскании решения $u(x, y)$ уравнения (1) в области $G\{\beta \leqslant x \leqslant \gamma,\ y \geqslant 0\}$, удовлетворяющего начальным условиям:

$$u|_{y=0} = \varphi(x); \quad \left.\frac{\partial u}{\partial y}\right|_{y=0} = \psi(x) \qquad (\beta \leqslant x \leqslant \gamma) \tag{2'}$$

и некоторым условиям на прямых $x = \beta$ и $x = \gamma$. Мы ограничимся тремя типами условий на этих прямых:

краевое условие первого рода

$$u|_{x=\beta} = \Phi(y) \qquad (y \geqslant 0), \tag{3}$$

краевое условие второго рода

$$\left.\frac{\partial u}{\partial x}\right|_{x=\beta} = \Psi(y) \qquad (y \geqslant 0), \tag{4}$$

краевое условие третьего рода

$$\left[\frac{\partial u}{\partial x} + \delta u\right]_{x=\beta} = F(y) \qquad (y \geqslant 0), \tag{5}$$

где Φ, Ψ, F, δ — заданные функции переменного y.

[1]) О применении метода сеток к более общим уравнениям гиперболического типа см. цитированную на стр. 411 обзорную статью О. А. Ладыженской, а также статью С. К. Годунова «Разностный метод численного расчета разрывных решений уравнений гидродинамики», Матем. сб., т. 47 (89): 3, 1959, 271—306.

Если на обеих прямых $x=\beta$ и $x=\gamma$ заданы краевые условия первого рода, то мы будем говорить, что имеем *первую краевую задачу*; если на обеих прямых заданы краевые условия второго рода, то будем говорить, что имеем *вторую краевую задачу*, и, наконец, если на обеих прямых мы имеем краевые условия третьего рода, то будем говорить о *третьей краевой задаче*. Часто встречаются и такие задачи, когда на прямых $x=\beta$ и $x=\gamma$ задаются условия разных типов.

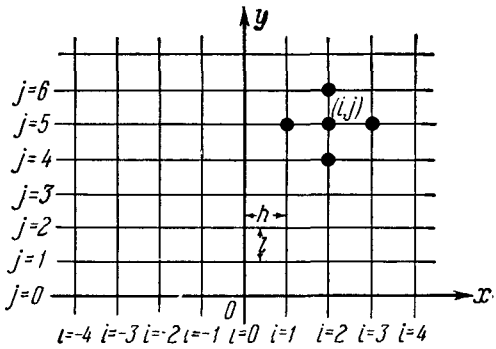

Рис. 38.

1. Метод сеток для решения задачи Коши. Проведем два семейства параллельных прямых:

$$x=ih\ (i=0,\pm 1,\pm 2,\ldots);\qquad y=jl\ (j=0,1,2),$$

т. е. покроем полуплоскость $y\geqslant 0$ сеткой прямоугольников со сторонами h и l по осям x и y соответственно. Вершины прямоугольников назовем *узлами* сетки. Узлы, расположенные на прямой $y=0$, несущей начальные данные, назовем *граничными* узлами. Для каждого внутреннего узла (i, j) составим разностное уравнение, аппроксимирующее дифференциальное уравнение (1) в этом узле, заменив входящие в него производные разностными отношениями:

$$\frac{\partial u}{\partial x}\approx\frac{u_{i+1,\,j}-u_{i-1,\,j}}{2h};\quad \frac{\partial u}{\partial y}\approx\frac{u_{i,\,j+1}-u_{i,\,j-1}}{2l};$$

$$\frac{\partial^2 u}{\partial x^2}\approx\frac{u_{i+1,\,j}-2u_{ij}+u_{i-1,\,j}}{h^2};\quad \frac{\partial^2 u}{\partial y^2}\approx\frac{u_{i,\,j+1}-2u_{ij}+u_{i,\,j-1}}{l^2},\qquad (6)$$

где $u_{ij}=u(ih, jl)$. Получим разностное уравнение

$$lu_{ij}=a_{ij}\frac{u_{i+1,\,j}-2u_{ij}+u_{i-1,\,j}}{h^2}-b_{ij}\frac{u_{i,\,j+1}-2u_{ij}+u_{i,\,j-1}}{l^2}+$$

$$+c_{ij}\frac{u_{i+1,\,j}-u_{i-1,\,j}}{2h}+d_{ij}\frac{u_{i,\,j+1}-u_{i,\,j-1}}{2l}+g_{ij}u_{ij}=f_{ij}\qquad (7)$$

или

$$lu_{ij}=A_{ij}u_{i,\,j+1}+B_{ij}u_{i,\,j-1}+C_{ij}u_{i+1,\,j}+D_{ij}u_{i-1,\,j}+E_{ij}u_{ij}=f_{ij},\qquad (7')$$

где

$$A_{ij}=-\frac{b_{ij}}{l^2}+\frac{d_{ij}}{2l};\quad B_{ij}=-\frac{b_{ij}}{l^2}-\frac{d_{ij}}{2l};\quad C_{ij}=\frac{a_{ij}}{h^2}+\frac{c_{ij}}{2h};$$

$$D_{ij}=\frac{a_{ij}}{h^2}-\frac{c_{ij}}{2h};\quad E_{ij}=-\frac{2a_{ij}}{h^2}+\frac{2b_{ij}}{l^2}+g_{ij}.\qquad (8)$$

Из этого уравнения видно, что если сетка настолько мала, что во всех узлах $A_{ij} \neq 0$ ($A_{ij} < 0$), то, зная значения решения в узлах $(j-1)$-го и j-го горизонтальных рядов, можно найти решение во всех узлах $(j+1)$-го горизонтального ряда.

Для того чтобы найти приближенное решение задачи Коши, необходимо знать значения решения на двух начальных рядах $j=0$ и $j=1$. Их можно найти из начальных условий одним из двух следующих способов.

Первый способ. В начальных условиях (2) заменим производную $\left.\frac{\partial u}{\partial y}\right|_{y=0}$ разностным отношением $\frac{u_{i1}-u_{i0}}{l}$. Тогда для определения значений в узлах первых двух горизонтальных рядов будем иметь:

$$u_{i0} = \varphi(ih) = \varphi_i; \quad \frac{u_{i1}-u_{i0}}{l} = \psi(ih) = \psi_i \tag{9}$$

или

$$u_{i0} = \varphi_i; \quad u_{i1} = \varphi_i + l\psi_i. \tag{9'}$$

Второй способ. Привлечем еще один горизонтальный ряд $j=-1$ и производную $\left.\frac{\partial u}{\partial y}\right|_{y=0}$ заменим разностным отношением $\frac{u_{i,1}-u_{i,-1}}{2l}$. Тогда из начальных условий (2) будем иметь:

$$u_{i0} = \varphi_i; \quad \frac{u_{i1}-u_{i,-1}}{2l} = \psi_i.$$

Значения $u_{i,-1}$ нам не нужны. Исключим их, используя разностное уравнение для узла $(i, 0)$, считая, что уравнение (1) удовлетворяется и на начальной прямой:

$$A_{i0}u_{i1} + B_{i0}u_{i,-1} + C_{i0}u_{i+1,0} + D_{i0}u_{i-1,0} + E_{i0}u_{i0} = f_{i0}.$$

Получим:

$$(A_{i0}+B_{i0})\,u_{i,1} + C_{i0}u_{i+1,0} + D_{i0}u_{i-1,0} + E_{i0}u_{i0} = f_{i0} + 2lB_{i0}\psi_i,$$

или

$$(A_{i0}+B_{i0})\,u_{i,1} = f_{i0} + 2lB_{i0}\psi_i - C_{i0}\varphi_{i+1} - D_{i0}\varphi_{i-1} - E_{i0}\varphi_i.$$

Таким образом, значения решения на первых двух рядах будут определяться следующим образом:

$$u_{i0} = \varphi_i; \quad u_{i1} = \frac{1}{A_{i0}+B_{i0}}[f_{i0} + 2lB_{i0}\psi_i - C_{i0}\varphi_{i+1} - D_{i0}\varphi_{i-1} - E_{i0}\varphi_i]. \tag{10}$$

Второй способ в некоторых случаях предпочтительнее, так как в этом случае мы имеем лучшую аппроксимацию начальных условий.

Особенно простые соотношения получаются для дифференциального уравнения

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f \tag{1'}$$

в случае квадратной сетки $h = l$. В этом случае разностное уравнение имеет вид

$$u_{i,j+1} - u_{i,j-1} - u_{i+1,j} + u_{i-1,j} = -h^2 f_{ij}, \tag{11}$$

а для вычисления значений на первых двух рядах будем иметь:
при первом способе

$$u_{i0} = \varphi_i; \quad u_{i1} = \varphi_i + h\psi_i, \tag{12}$$

при втором способе

$$u_{i0} = \varphi_i; \quad u_{i1} = \frac{1}{2}[-h^2 f_{i0} + 2h\psi_i + \varphi_{i+1} + \varphi_{i-1}]. \tag{13}$$

Погрешность аппроксимации. В результате замены дифференциального уравнения (1) разностным уравнением (7) и начальных условий (2) условиями (9) или (10) мы вносим погрешность, которую назовем *погрешностью аппроксимации* (или *погрешностью метода*). Предполагая у решения $u(x, y)$ задачи Коши существование ограниченных производных до четвертого порядка включительно и используя разложение решения по формуле Тейлора в окрестности узла (i, j), точно так же, как и в случае эллиптических уравнений (см. § 2), получим:

$$[lu_{ij} - f_{ij}] = \{Lu - f\}_{(i,j)} + R_{ij} = R_{ij}, \tag{14}$$

где

$$R_{ij} = \frac{h^2}{12}\left\{a_{ij}\tilde{u}^{(\mathrm{IV})}_{x^4} - b_{ij}\alpha^2\tilde{u}^{(\mathrm{IV})}_{y^4} + 2c_{ij}\tilde{u}'''_{x^3} + 2\alpha^2 d_{ij}\tilde{u}'''_{y^3}\right\}, \tag{15}$$

а $\tilde{u}^{(k)}_{x^k}$, $\tilde{u}^{(k)}_{y^k}$ обозначают k-е производные от $u(x, y)$ в некоторой точке, отличной от узла (i, j), а $\alpha = \frac{l}{h}$. Если ввести обозначение $M_k = \max\limits_{G}\left\{\left|u^{(k)}_{x^k}\right|, \left|u^{(k)}_{y^k}\right|\right\}$, то

$$|R_{ij}| \leqslant \frac{h^2}{12}\{(|a_{ij}| + \alpha^2|b_{ij}|)M_4 + 2(|c_{ij}| + \alpha^2|d_{ij}|)M_3\}. \tag{16}$$

Что касается погрешности аппроксимации начальных условий, то в первом способе аппроксимации начальных условий по формулам (9) первое начальное условие не вносит никакой погрешности, а второе начальное условие вносит следующую погрешность:

$$\frac{u_{i,1} - u_{i0}}{l} = \left(\frac{\partial u}{\partial y}\right)_{(i,0)} + \frac{l}{2}\tilde{u}''_{y^2} = \psi_i + \frac{l}{2}\tilde{u}''_{y^2} = \psi_i + r_i$$

($\tilde{u}''_{y^2}$ — значение второй производной в некоторой точке). Таким образом,

$$u_{i0}=\varphi_i;\quad u_{i1}=\varphi_i+l(\psi_i+r_i), \tag{9}$$

где

$$r_i=\frac{l}{2}\tilde{u}''_{y^2}\quad\text{и}\quad |r_i|\leqslant\frac{\alpha h}{2}M_2. \tag{17}$$

При втором способе аппроксимации начальных условий имеем:

$$\frac{u_{i,1}-u_{i,-1}}{2l}=\left(\frac{\partial u}{\partial y}\right)_{(i,0)}+\frac{l^2}{6}\tilde{u}'''_{y^3}=\psi_i+\eta_i,$$

где

$$\eta_i=\frac{l^2}{6}\tilde{u}'''_{y^3};\quad |\eta_i|\leqslant\frac{l^2}{6}M_3.$$

Далее,

$$lu_{i0}=f_{i0}+R_{i0},$$

где R_{i0} имеет вид (15) с $j=1$. Исключая $u_{i,-1}$, как прежде, получим:

$$u_{i0}=\varphi_i;\quad u_{i1}=\frac{1}{A_{i0}+B_{i0}}[f_i+2lB_{i0}\psi_i-C_{i0}\varphi_{i+1}-D_{i0}\varphi_{i-1}-E_{i0}\varphi_i]+r_i^*, \tag{10}$$

где

$$r_i^*=\frac{1}{A_{i0}+B_{i0}}[R_{i0}+2lB_{i0}\eta_i]=-\frac{1}{2b_{i0}}[l^2R_{i0}-l(2b_{i0}+ld_{i0})\eta_i] \tag{18}$$

или

$$|r_i^*|\leqslant\frac{1}{2|b_{i0}|}\left\{\frac{h^4\alpha^2}{12}[(|a_{i0}|+\alpha^2|b_{i0}|)M_4+2(|c_{i0}|+\alpha^2|d_{i0}|)M_3]+\right.$$
$$\left.+\frac{\alpha^2h^3}{6}(2|b_{i0}|+\alpha h|d_{i0}|)M_3\right. \tag{19}$$

Для уравнения (1′) при $\alpha=1$ эта оценка будет иметь вид

$$|r_i^*|\leqslant\frac{h^4}{12}M_4+\frac{h^3}{6}M_3. \tag{19′}$$

Выбор сетки. Если мы хотим получить методом сеток решение, сколь угодно близкое к точному решению задачи Коши для гиперболического уравнения, то нельзя произвольно выбирать соотношения шагов сетки по осям x и y. Для иллюстрации рассмотрим решение задачи Коши для простейшего дифференциального уравнения гиперболического типа

$$\frac{\partial^2u}{\partial x^2}-\frac{\partial^2u}{\partial y^2}=0$$

с начальными условиями

$$u\Big|_{y=0}=\varphi(x);\quad \frac{\partial u}{\partial y}\Big|_{y=0}=\psi(x)\qquad(-\infty<x<\infty)$$

методом сеток. Будем использовать прямоугольную сетку с шагом h по x и l по y. Пусть $l = \alpha h$. В этом случае будем иметь разностное уравнение

$$lu_{ij} = \frac{u_{i,j+1} - 2u_{ij} + u_{i,j-1}}{\alpha^2 h^2} - \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{h^2} = 0$$

или

$$u_{i,j+1} = 2u_{ij} - u_{i,j-1} + \alpha^2 (u_{i+1,j} - 2u_{ij} + u_{i-1,j}).$$

Зная значения решения в узлах первых двух горизонтальных рядов, можно последовательно вычислить значения решения на втором, третьем и т. д. рядах. При этом значения решения в узле $S(i, j)$ будут определяться начальными данными на отрезке оси x, высекаемом прямыми, выходящими из этого узла и образующими с осью x углы, тангенсы которых равны $\pm\alpha$. Треугольник, образованный этими прямыми, назовем *треугольником определенности разностной схемы*. Если через тот же узел провести две выходящие из него характеристики до пересечения с прямой $y = 0$, то получим еще один треугольник — *треугольник определенности дифференциального уравнения*. Известно, что решение $u(x, y)$ задачи Коши для дифференциального уравнения в узле $S(l, j)$ будет полностью определяться начальными данными на основании последнего треугольника. Пусть шаг l по оси y больше h, т. е. $\alpha > 1$. Тогда треугольник определенности разностной схемы целиком содержится внутри треугольника определенности дифференциального уравнения (рис. 39). Покажем, что в этом случае при постоянном α и при h, стремящемся к нулю таким образом, чтобы точка S была все время узлом сетки, значения решения в узле S, получаемые методом сеток, могут не сходиться к значению истинного решения задачи Коши в этой точке. Это просто показать следующим образом.

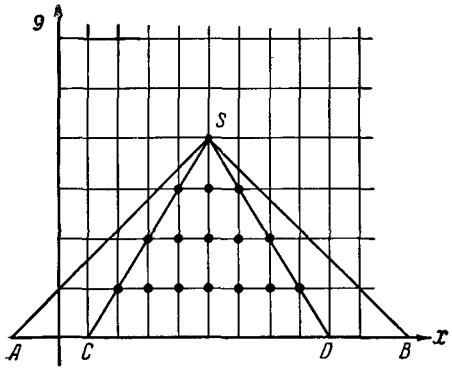

Рис. 39.

При указанном способе стремления к нулю h треугольник определенности разностной схемы все время остается неизменным и приближенное решение в этой точке при любом h будет целиком определяться начальными данными на отрезке CD. Если мы будем изменять начальные данные на отрезках AC и DB, то на решения разностных уравнений в точке S это изменение совершенно влиять не будет, а значения решения задачи Коши для дифференциального уравнения будет существенно зависеть от этих изменений. Следова-

тельно, в этом случае мы не можем говорить о сходимости решений, получаемых методом сеток, к решению задачи Коши для дифференциального уравнения. Отсюда вывод, что *для сходимости последовательности приближенных решений, получаемых методом сеток при постоянстве отношения* $\alpha = \frac{l}{h}$, *при* $h \to 0$ *необходимо выполнение условия* $\alpha \leqslant 1$, т. е. треугольник определенности дифференциального уравнения должен совпадать или содержаться внутри треугольника определенности разностной схемы, имеющего ту же вершину.

В общем случае треугольник определенности дифференциального уравнения становится криволинейным, но и в этом случае для сходимости необходимо выполнение условия, чтобы треугольник определенности дифференциального уравнения содержался бы внутри треугольника определенности разностной схемы. Это требование налагает определенные требования на соотношение шагов, т. е. на выбор сетки. При некоторых требованиях гладкости коэффициентов и начальных функций указанное выше условие является и достаточным для сходимости последовательности решений, получаемых методом сеток, к точному решению задачи Коши для дифференциального уравнения (1).

К вопросу выбора сетки мы еще вернемся в дальнейшем уже в связи с исследованием устойчивости разностных схем (см. § 6).

2. Оценка погрешности и сходимость метода сеток для неоднородного волнового уравнения. Оставляя в стороне общий случай, рассмотрим оценку погрешности решения задачи Коши для неоднородного волнового дифференциального уравнения, получаемого методом сеток:

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f(x, y), \qquad (1')$$

с начальными условиями

$$u\big|_{y=0} = \varphi(x); \quad \frac{\partial u}{\partial y}\Big|_{y=0} = \psi(x) \qquad (2)$$

$$(-\infty < x < \infty)$$

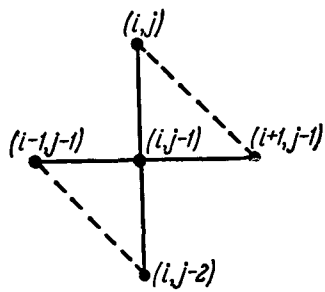

Рис. 40.

для случая квадратной сетки $l = h$. Разностное уравнение в этом случае имеет вид (11). Мы его перепишем в более удобной форме:

$$u_{ij} - u_{i+1,\,j-1} = u_{i-1,\,j-1} - u_{i,\,j-2} - f_{i,\,j-1}h^2. \qquad (11')$$

В левой и правой частях уравнения (11′) входят разности значений u в узлах, соединенных на рис. 40 пунктирной линией. Значения решения на первых двух горизонтальных рядах определим

с помощью равенств (13):

$$u_{i0} = \varphi_i; \quad u_{i1} = \frac{1}{2}[-h^2 f_{i0} + 2h\psi_i + \varphi_{i+1} + \varphi_{i-1}].$$

Используя (11′), будем иметь следующие равенства:

$$u_{i,j} - u_{i+1,j-1} = u_{i-1,j-1} - u_{i,j-2} - h^2 f_{i,j-1} = \ldots = u_{i-j+1,1} - u_{i-j+2,0} - h^2 \sum_{\nu=0}^{j-2} f_{i-\nu,j-\nu-1},$$

$$u_{i+1,j-1} - u_{i+2,j-2} = u_{i-j+3,1} - u_{i-j+4,0} - h^2 \sum_{\nu=0}^{j-3} f_{i-\nu+1,j-\nu-2},$$

. . . . . .

$$u_{i+j-2,2} - u_{i+j-1,1} = u_{i+j-3,2} - u_{i+j-2,0} - h^2 f_{i+j-2,1}.$$

Суммируя их почленно, получим:

$$u_{i,j} = \sum_{\nu=-\frac{j}{2}}^{\frac{j}{2}-1} u_{i+2\nu+1,1} - \sum_{\nu=-\frac{j}{2}}^{\frac{j}{2}-2} u_{i+2\nu+2,0} - h^2 \sum_{\alpha=0}^{j-2} \sum_{\nu=0}^{j-2-\alpha} f_{i-\nu+\alpha,j-\alpha-\nu-1} = S_1 - S_2 - S_3, \tag{20}$$

где

$$S_1 = \sum_{\nu=-\frac{j}{2}}^{\frac{j}{2}-1} u_{i+2\nu+1,1}; \quad S_2 = \sum_{\nu=-\frac{j}{2}}^{\frac{j}{2}-2} u_{i+2\nu+2,0};$$

$$S_3 = h^2 \sum_{\alpha=0}^{j-2} \sum_{\nu=0}^{j-2-\alpha} f_{i-\nu+\alpha,j-\alpha-\nu-1}.$$

Сумма S_1 содержит лишь значения решения u в узлах первого горизонтального ряда, отмеченных на рис. 41 жирными точками, а S_2 — значения решения u в узлах нулевого горизонтального ряда, тоже отмеченных жирными точками. Сумма S_3 содержит значения правой части f в узлах сетки, отмеченных крестиками. Обозначим через v_{ij} значения точного решения задачи Коши в узлах сетки, т. е. $v_{ij} = u(ih, jh)$, а под u_{ij} будем понимать значения решения, полученные методом сеток. Через ε_{ij} обозначим разность $u_{ij} - v_{ij}$, т. е. погрешность приближенного решения. Докажем, что ε_{ij} по абсолютной величине не превосходит некоторой постоянной

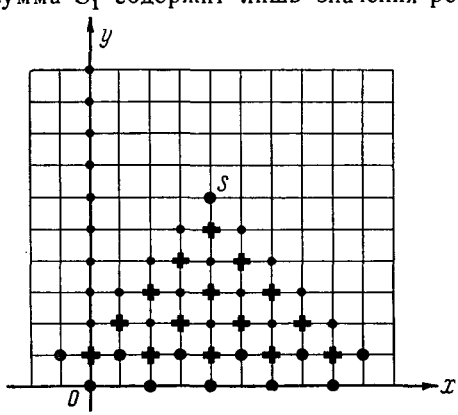

Рис. 41.

умноженной на h^2. Будем предполагать, что задача Коши имеет решение с непрерывными и ограниченными производными до четвертого порядка включительно. Мы видели, что для точного решения v_{ij} имеют место равенства

$$lv_{ij} = f_{ij} + R_{ij}, \tag{14'}$$

где $|R_{ij}| \leqslant \frac{h^2}{6} M_4$ или

$$R_{ij} = \frac{h^2}{6} \theta_{ij} M_4 \quad (0 \leqslant |\theta_{ij}| \leqslant 1), \tag{15'}$$

$$v_{i0} = \varphi_i; \quad v_{i1} = \frac{1}{2}[-h^2 f_{i0} + 2h\psi_i + \varphi_{i+1} + \varphi_{i-1}] + r_i^*, \tag{$\overline{10'}$}$$

где

$$|r_i^*| \leqslant \frac{h^4}{12} M_4 + \frac{h^3}{6} M_3. \tag{19'}$$

Вычитая из (11) и (13) почленно соответствующие равенства (14') и ($\overline{10'}$), получим для погрешности ε_{ij} систему уравнений

$$\left.\begin{aligned} l\varepsilon_{ij} &= -R_{ij} = -\frac{h^2}{6}\theta_{ij} M_4, \\ \varepsilon_{i0} &= 0; \quad \varepsilon_{i1} = -r_i^*. \end{aligned}\right\} \tag{21}$$

Воспользовавшись явным представлением решения разностной системы (20), будем иметь:

$$\varepsilon_{ij} = \sum_{\nu=-\frac{j}{2}}^{\frac{j}{2}-1} \varepsilon_{i+2\nu+1,\,1} - \sum_{\nu=-\frac{j}{2}}^{\frac{j}{2}-2} \varepsilon_{i+2\nu+2,\,0} + \frac{h^4}{6} M_4 \sum_{\alpha=0}^{j-2} \sum_{\nu=0}^{j-2-\alpha} \theta_{i-\nu+\alpha,\,j-\alpha-\nu-1}. \tag{22}$$

Все слагаемые второй суммы нули; первая сумма содержит j слагаемых, каждое из которых не превосходит $\frac{h^4}{12} M_4 + \frac{h^3}{6} M_3$; третья сумма состоит из $\frac{j(j-1)}{2}$ слагаемых, каждое из которых по абсолютной величине не превосходит единицу. Таким образом,

$$|\varepsilon_{ij}| \leqslant j\left(\frac{h^4}{12} M_4 + \frac{h^3}{6} M_3\right) + \frac{h^4}{12} M_4 (j^2 - j).$$

Если узел (i, j) имеет координаты (x_0, y_0), то $jh = y_0$. Отсюда

$$|\varepsilon_{ij}| \leqslant \frac{h^2}{12}(M_4 h + 2M_3) y_0 + \frac{h^2}{12} y_0^2 M_4.$$

При фиксированных x_0, y_0 и $h \to 0$ погрешность ε_{ij} в этой точке стремится к нулю как h^2. Это означает, что последовательность решений задачи Коши, получаемая методом сеток, сходится к точному решению задачи.

3. Метод сеток решения смешанной задачи. Метод сеток решения задачи Коши для уравнения (1) с небольшими изменениями может быть применен и для решения смешанной задачи. Рассмотрим сначала первую краевую задачу:

Найти решение уравнения (1) в области $G\{\beta \leqslant x \leqslant \gamma;\ 0 \leqslant y \leqslant Y\}$, удовлетворяющее начальным условиям

$$u|_{y=0} = \varphi(x); \quad \left.\frac{\partial u}{\partial y}\right|_{y=0} = \psi(x) \qquad (\beta \leqslant x \leqslant \gamma) \tag{2'}$$

и граничным условиям

$$u|_{x=\beta} = \Phi_1(y); \quad u|_{x=\gamma} = \Phi_2(y) \qquad (0 \leqslant y \leqslant Y). \tag{3}$$

Отрезок $[\beta, \gamma]$ можно с помощью линейного преобразования переменного x свести к отрезку $[0, 1]$, поэтому в дальнейшем мы будем полагать $\beta = 0$, $\gamma = 1$.

Проведем две системы параллельных прямых:

$$x = ih \quad \left(i = 0, 1, 2, \ldots, n; \quad h = \frac{1}{n}\right);$$

$$y = jl \ (j = 0, 1, 2, \ldots, m; \quad ml \leqslant Y < (m+1)l),$$

и точки пересечения их назовем узлами сетки. Узлы, лежащие на прямых $x = 0$, $x = 1$, $y = 0$, назовем граничными, а лежащие внутри области G — внутренними узлами. Для каждого внутреннего узла, так же как и в случае задачи Коши, напишем разностное уравнение (7), аппроксимирующее дифференциальное уравнение (1) в этом узле. Значения искомого решения в узлах нулевого горизонтального ряда и во внутренних узлах первого ряда найдем с помощью начальных условий по формулам (9) или (10), значения же решения в граничных узлах, лежащих на прямых $x = 0$ и $x = 1$, определим из граничных условий, положив

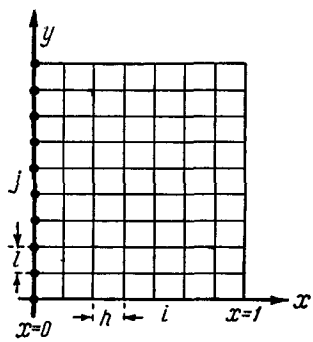

Рис. 42.

$$u_{0j} = \Phi_1(jl) = \Phi_{1j}; \quad u_{nj} = \Phi_2(jl) = \Phi_{2j}.$$

Тогда, используя уравнение (7), мы можем найти последовательно приближенные значения решения во внутренних узлах второго ряда, затем в узлах третьего горизонтального ряда и т. д., т. е. сможем вычислить значения решения во всех узлах сетки.

Если граничные условия заданы на прямых $x = 0$ и $x = 1$ на отрезках разной длины, то решение может быть найдено в узлах сетки, отмеченных на рис. 43 жирными точками. Заметим, что в первой краевой задаче, если предполагать, что значения граничных функций Φ_1 и Φ_2 в граничных узлах вычисляются точно, то граничные

условия не внесут дополнительной погрешности по сравнению со случаем задачи Коши.

Рассмотрим теперь вторую и третью краевые задачи. Для того чтобы не повторять рассуждений, будем рассматривать случай, когда при $x=0$ и $x=1$ заданы условия

$$\left[\frac{\partial u}{\partial x}+\delta_1 u\right]_{x=0}=F_1(y); \quad \left[\frac{\partial u}{\partial x}+\delta_2 u\right]_{x=1}=F_2(y) \quad [0\leqslant y\leqslant Y], \quad (4')$$

т. е. третью краевую задачу, так как, полагая $\delta_1(y)\equiv\delta_2(y)\equiv 0$, получим как частный случай и вторую краевую задачу. Если для решения задачи используется такая же сетка, как и в случае первой краевой задачи, то в этом случае значения в граничных узлах, расположенных на прямых $x=0$ и $x=1$, не могут быть непосредственно определены из граничных условий. Для этих граничных узлов нужно записать разностные уравнения, аппроксимирующие дифференциальные соотношения на границе.

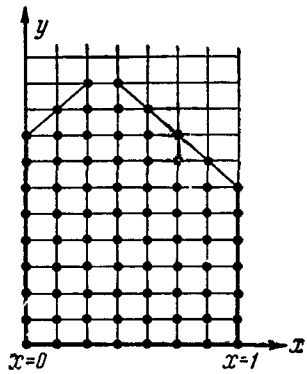

Рис. 43.

Аппроксимация граничных условий может быть выполнена следующим образом. Производные $\left.\frac{\partial u}{\partial x}\right|_{x=0}$, $\left.\frac{\partial u}{\partial x}\right|_{x=1}$ в граничных узлах $(0, j)$, (n, j) заменим соответственно конечными разностями

$$\frac{u_{1j}-u_{0j}}{h}; \quad \frac{u_{nj}-u_{n-1,j}}{h}.$$

Тогда для граничных узлов $(0, j)$, (n, j), используя условия (4'), можно записать следующие разностные уравнения:

$$\frac{u_{1j}-u_{0i}}{h}+\delta_{1j}u_{0j}=F_{1j}; \quad \frac{u_{nj}-u_{n-1,j}}{h}+\delta_{2j}u_{nj}=F_{2j}. \quad (23)$$

Легко видеть, что погрешность аппроксимации граничных условий в этом случае будет первого порядка относительно h. Поэтому начальные условия также достаточно аппроксимировать с этой точностью, т. е. для отыскания значений решения в узлах первых двух горизонтальных рядов можно воспользоваться равенствами (9'). Итак, для определения приближенных значений решения в граничных и внутренних узлах сетки имеем систему уравнений

$$A_{ij}u_{i,j+1}+B_{ij}u_{i,j-1}+C_{ij}u_{i+1,j}+D_{ij}u_{i-1,j}+E_{ij}u_{ij}=f_{ij} \quad (1\leqslant i\leqslant n-1; \quad 1\leqslant j<m), \quad (7')$$

$$u_{1j}+u_{0j}(\delta_{1j}h-1)=hF_{1j}; \quad u_{nj}[1+\delta_{2j}h]-u_{n-1,j}=hF_{2j} \quad (1\leqslant j<m), \quad (23')$$

$$u_{i0}=\varphi_i; \quad u_{i1}=\varphi_i+l\psi_i \quad (i=0, 1, 2, \ldots, n). \quad (9')$$

С помощью уравнений (9′) мы найдем значения решения во всех узлах первого и нулевого горизонтальных рядов, включая и узлы (0, 1), (n, 1); из уравнений (7′) найдутся значения решения во всех внутренних узлах второго горизонтального ряда, а с помощью уравнений (23′) найдем значения решения в граничных узлах второго горизонтального ряда. Далее переходим к третьему горизонтальному ряду и т. д.

Более точную аппроксимацию граничных условий можно получить, заменив входящие в них производные центральными разностями. Это можно сделать двумя способами [1]).

Первый способ. Кроме рассматриваемых узлов, привлечем еще узлы -1 и $n+1$ вертикальных рядов (см. рис. 44) и будем аппроксимировать граничные условия в узлах (0, j) и (n, j) соответственно разностными уравнениями

$$\frac{u_{1j}-u_{-1,j}}{2h}+\delta_{1j}u_{0j}=F_{1j};\quad \frac{u_{n+1,j}-u_{n-1,j}}{2h}+\delta_{2j}u_{nj}=F_{2j}. \tag{24}$$

Для исключения значений $u_{-1,j}$ и $u_{n+1,j}$ используем разностные уравнения, аппроксимирующие дифференциальное уравнение в узлах (0, j) и (n, j), т. е.

$$A_{0j}u_{0,j+1}+B_{0j}u_{0,j-1}+C_{0j}u_{1j}+D_{0j}u_{-1,j}+E_{0j}u_{0j}=f_{0j},$$

$$A_{nj}u_{n,j+1}+B_{nj}u_{n,j-1}+C_{nj}u_{n+1,j}+D_{nj}u_{n-1,j}+E_{nj}u_{nj}=f_{nj}.$$

Подставляя в них $u_{-1,j}$ и $u_{n+1,j}$ из уравнений (24), получим:

$$\left.\begin{aligned}
&A_{0j}u_{0,j+1}+B_{0j}u_{0,j-1}+(C_{0j}+D_{0j})\,u_{1j}+(E_{0j}+2h\delta_{1j}D_{0j})\,u_{0j}=\\
&\qquad\qquad =f_{0j}+2hD_{0j}F_{1j},\\
&A_{nj}u_{n,j+1}+B_{nj}u_{n,j-1}+(C_{nj}+D_{nj})\,u_{n-1,j}+(E_{nj}-2h\delta_{2j}C_{nj})\times\\
&\qquad\qquad \times u_{nj}=f_{nj}-2hC_{nj}F_{2j}.
\end{aligned}\right\} \tag{25}$$

Так как аппроксимация граничных условий в этом случае имеет порядок h^2 (в предположении, что уравнение (1) имеет место и на границах $x=0$ и $x=1$, а также предполагая, что решение можно гладко продолжить за эти прямые), то естественно и начальные условия аппроксимировать с точностью до h^2, т. е. использовать формулы (10) для u_{i0}, u_{i1} ($1\leqslant i\leqslant n$). Для определения значений решения в узлах (0, 1) и (n, 1) поступим следующим образом.

---

[1]) Кроме этих способов, можно для улучшения аппроксимации граничных условий воспользоваться идеей, описаной в разностном методе решения краевой задачи для обыкновенного дифференциального уравнения на стр. 373.

Запишем уравнения (25) для узлов (0, 0) и (n, 0), т. е. просто положим в них $j=0$:

$$A_{00}u_{01}+B_{00}u_{0,-1}+(C_{00}+D_{00})u_{10}+$$
$$+(E_{00}+2h\delta_{10}D_{00})u_{00}=f_{00}+2hD_{00}F_{10},$$
$$A_{n0}u_{n1}+B_{n0}u_{n,-1}+(C_{n0}+D_{n0})u_{n-1,0}+$$
$$+(E_{n0}-2h\delta_{20}C_{00})u_{n0}=f_{n0}-2hC_{n0}F_{20},$$

и, используя аппроксимацию второго начального условия, с помощью центральных разностей

$$\frac{u_{01}-u_{0,-1}}{2l}=\psi_0;\quad \frac{u_{n1}-u_{n,-1}}{2l}=\psi_n$$

исключим $u_{0,-1}$ и $u_{n,-1}$. Получим:

$$(A_{00}+B_{00})u_{01}=f_{00}+2lB_{00}\psi_0+$$
$$+2hD_{00}F_{10}-(C_{00}+D_{00})\varphi_1-(E_{00}+2h\delta_{10}D_{00})\varphi_0,$$
$$(A_{n0}+B_{n0})u_{n1}=f_{n0}+2lB_{n0}\psi_n-2hC_{n0}F_{20}-$$
$$-(C_{n0}+D_{n0})\varphi_{n-1}-(E_{n0}-2h\delta_{20}C_{n0})\varphi_n.$$

Окончательно получим следующую систему уравнений для отыскания значений решения во всех внутренних и граничных узлах:

$$\left.\begin{aligned}
&A_{ij}u_{i,j+1}+B_{ij}u_{i,j-1}+C_{ij}u_{i+1,j}+D_{ij}u_{i-1,j}+E_{ij}u_{ij}=f_{ij}\\
&\qquad(1\leqslant i\leqslant n-1,\ 1\leqslant j<m),\\
&A_{0j}u_{0,j+1}+B_{0j}u_{0,j-1}+(C_{0j}+D_{0j})u_{1j}+\\
&\qquad+(E_{0j}+2h\delta_{1j}D_{0j})u_{0j}=f_{0j}+2hD_{0j}F_{1j},\\
&A_{nj}u_{n,j+1}+B_{nj}u_{n,j-1}+(C_{nj}+D_{nj})u_{n-1,j}+\\
&\qquad+(E_{nj}-2h\delta_{2j}C_{nj})u_{nj}=f_{nj}-2hC_{nj}F_{2j},\\
&(A_{00}+B_{00})u_{01}=f_{00}+2lB_{00}\psi_0+\\
&\qquad+2hD_{00}F_{10}-(C_{00}+D_{00})\varphi_1-(E_{00}+2h\delta_{10}D_{00})\varphi_0,\\
&(A_{n0}+B_{n0})u_{n1}=f_{n0}+2lB_{n0}\psi_n-2hC_{n0}F_{20}-\\
&\qquad-(C_{n0}+D_{n0})\varphi_{n-1}-(E_{n0}-2h\delta_{20}C_{n0})\varphi_n,\\
&u_{i0}=\varphi_i,\\
&u_{i1}=\frac{1}{A_{i0}+B_{i0}}[f_{i0}+2lB_{i0}\psi_i-C_{i0}\varphi_{i+1}-D_{i0}\varphi_{i-1}-E_{i0}\varphi_i].
\end{aligned}\right\}\qquad(26)$$

Используя ее, можно последовательно находить значения решения в узлах первого ряда, затем второго ряда и т. д.

Второй способ. Будем рассматривать сетку, сдвинутую на $\frac{h}{2}$ в направлении оси x (см. рис. 45). При таком выборе сетки граничные узлы (0, j) и (n, j) уже не будут лежать на прямых

$x=0$ и $x=1$. Граничные условия (4′) будем аппроксимировать разностными уравнениями

$$\left.\begin{aligned} \frac{u_{1j}-u_{0j}}{l}+\delta_{1j}\frac{u_{1j}+u_{0j}}{2}&=F_{1j};\\ \frac{u_{nj}-u_{n-1,\,j}}{l}+\delta_{2j}\frac{u_{nj}+u_{n-1,\,j}}{2}&=F_{2j}. \end{aligned}\right\} \tag{27}$$

Очевидно, мы снова имеем аппроксимацию второго порядка относительно h, поэтому и начальные условия нужно аппроксимировать с такой же точностью, т. е. для вычисления значений u_{i0}, u_{i1} использовать равенства (10). Таким образом, окончательно получим следующую систему разностных уравнений:

$$\left.\begin{aligned} &A_{ij}u_{i,\,j+1}+B_{ij}u_{i,\,j-1}+C_{ij}u_{i+1,\,j}+D_{ij}u_{i-1,\,j}+E_{ij}u_{ij}=f_{ij}\\ &\qquad\qquad (1\leqslant i\leqslant n-1;\quad 1\leqslant j<m),\\ &(2+\delta_{1j}l)\,u_{1j}-(2-l\delta_{1j})\,u_{0j}=2lF_{1j};\\ &\qquad (2+l\delta_{2j})\,u_{nj}-(2-l\delta_{2j})\,u_{n-1,\,j}=2lF_{2j}\qquad (1\leqslant j<m).\\ &u_{i0}=\varphi_i,\\ &(A_{i0}+B_{i0})\,u_{i1}=f_{i0}+2lB_{i0}\psi_i-C_{i0}\varphi_{i+1}-D_{i0}\varphi_{i-1}-E_{i0}\varphi_i. \end{aligned}\right\} \tag{28}$$

Нужные нам для счета значения φ_0, φ_n, ψ_0, ψ_n можно получить с помощью экстраполяции или вообще с помощью продолжения

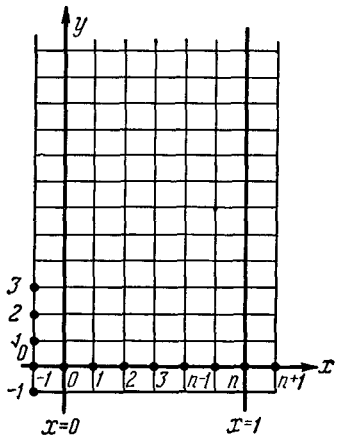

Рис. 44.

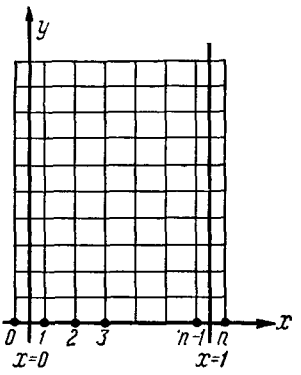

Рис. 45.

функций φ и ψ за пределы их области определения, сохраняя нужную гладкость. Значения приближенного решения на истинных границах $x=0$ и $x=1$ могут быть после решения системы (28) найдены с помощью интерполяции. Система (28), как и во всех

предыдущих случаях, решается последовательно, т. е. находятся значения сначала в узлах первого горизонтального ряда, затем в узлах второго горизонтального ряда и т. д.

4. Другие разностные схемы. Мы рассматривали простейшие разностные схемы. Методами, изложенными в § 2, можно построить множество других разностных схем, которые будут с различной точностью аппроксимировать дифференциальное уравнение (1), начальные условия (2) и в случае смешанной задачи граничные условия. Все схемы можно разбить на два типа: *явные схемы* и *неявные схемы*. Явными схемами называют такие схемы, что при любом j в каждое из уравнений, связывающих значения искомого решения на горизонтальных рядах j, $j-1, \ldots, j-m$, входит лишь одна точка ряда j, так что значение решения в каждом узле j-го горизонтального ряда можно вычислить независимо от его значений в других узлах этого ряда (исключая граничные узлы). Неявными схемами называют такие схемы, когда для определения значений решения в узлах j-го ряда при известных значениях решения во всех предыдущих рядах нужно решать систему уравнений, связывающих значения решения в узлах j-го ряда. Все рассмотренные выше схемы являются явными схемами. Приведем пример простой неявной разностной схемы для простейшего гиперболического уравнения

$$Lu = \frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = 0, \tag{29}$$

причем будем для простоты рассматривать квадратную сетку с шагом h.

Будем строить разностную аппроксимацию дифференциального оператора Lu в узле (i, j), используя значения функции u в семи

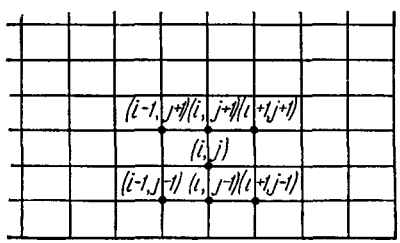

Рис. 46.

узлах, изображенных на рис. 46. Предполагая у функции $u(x, y)$ существование непрерывных производных до четвертого порядка включительно, разложим функцию по формуле Тейлора с остаточ-

ным членом в производных четвертого порядка. Будем иметь:

$$u_{i+1,\,j+1}=\left\{u+h\left(\frac{\partial}{\partial x}+\frac{\partial}{\partial y}\right)u+\frac{h^2}{2!}\left(\frac{\partial}{\partial x}+\frac{\partial}{\partial y}\right)^2 u+\right.$$
$$\left.+\frac{h^3}{3!}\left(\frac{\partial}{\partial x}+\frac{\partial}{\partial y}\right)^3 u\right\}_{(i,\,j)}+h^4R_1,$$

$$u_{i+1,\,j-1}=\left\{u+h\left(\frac{\partial}{\partial x}-\frac{\partial}{\partial y}\right)u+\frac{h^2}{2!}\left(\frac{\partial}{\partial x}-\frac{\partial}{\partial y}\right)^2 u+\right.$$
$$\left.+\frac{h^3}{3!}\left(\frac{\partial}{\partial x}-\frac{\partial}{\partial y}\right)^3 u\right\}_{(i,\,j)}+h^4R_2,$$

$$u_{i-1,\,j+1}=\left\{u+h\left(-\frac{\partial}{\partial x}+\frac{\partial}{\partial y}\right)u+\frac{h^2}{2!}\left(-\frac{\partial}{\partial x}+\frac{\partial}{\partial y}\right)^2 u+\right.$$
$$\left.+\frac{h^3}{3!}\left(-\frac{\partial}{\partial x}+\frac{\partial}{\partial y}\right)^3 u\right\}_{(i,\,j)}+h^4R_3,$$

$$u_{i-1,\,j-1}=\left\{u-h\left(\frac{\partial}{\partial x}+\frac{\partial}{\partial y}\right)u+\frac{h^2}{2!}\left(\frac{\partial}{\partial x}+\frac{\partial}{\partial y}\right)^2 u-\right.$$
$$\left.-\frac{h^3}{3!}\left(\frac{\partial}{\partial x}+\frac{\partial}{\partial y}\right)^3 u\right\}_{(i,\,j)}+h^4R_4,$$

$$u_{i,\,j+1}=\left\{u+h\frac{\partial u}{\partial y}+\frac{h^2}{2!}\frac{\partial^2 u}{\partial y^2}+\frac{h^3}{3!}\frac{\partial^3 u}{\partial y^3}\right\}_{(i,\,j)}+h^4R_5,$$

$$u_{i,\,j-1}=\left\{u-h\frac{\partial u}{\partial y}+\frac{h^2}{2!}\frac{\partial^2 u}{\partial y^2}-\frac{h^3}{3!}\frac{\partial^3 u}{\partial y^3}\right\}_{(i,\,j)}+h^4R_6.$$

Составим линейную комбинацию

$$u_{ij}=c_0u_{ij}+c_1u_{i+1,\,j+1}+c_2u_{i+1,\,j-1}+c_3u_{i-1,\,j+1}+$$
$$+c_4u_{i-1,\,j-1}+c_5u_{i,\,j+1}+c_6u_{i,\,j-1}$$

с неопределенными коэффициентами c_0, c_1, ..., c_6 и подберем их так, чтобы после подстановки разложений в правой части исчезли функция u и производные $\frac{\partial u}{\partial x}$, $\frac{\partial u}{\partial y}$, $\frac{\partial^2 u}{\partial x\,\partial y}$, а члены со вторыми производными давали бы оператор Lu. Для этого нужно потребовать, чтобы c_0, c_1, ..., c_6 удовлетворяли системе уравнений

$$\begin{aligned}
c_0+c_1+c_2+c_3+c_4+c_5+c_6&=0,\\
c_1+c_2-c_3-c_4&=0,\\
c_1-c_2+c_3-c_4+c_5-c_6&=0,\\
c_1-c_2-c_3+c_4&=0,\\
c_1+c_2+c_3+c_4&=\frac{2}{h^2},\\
c_1+c_2+c_3+c_4+c_5+c_6&=-\frac{2}{h^2}.
\end{aligned}$$

Добавим еще одно уравнение

$$c_5=c_6.$$

Тогда, решая систему, получим для $c_0, c_1, c_2, \ldots, c_6$ следующие значения:

$$c_0 = \frac{2}{h^2}; \quad c_1 = c_2 = c_3 = c_4 = \frac{1}{2h^2}; \quad c_5 = c_6 = -\frac{2}{h^2}.$$

При этих значениях коэффициентов после подстановки в lu_{ij} разложений по формуле Тейлора и приведения подобных членов получим:

$$lu_{ij} = \frac{1}{2h^2}[4u_{ij} + u_{i+1,j+1} + u_{i+1,j-1} + u_{i-1,j+1} + u_{i-1,j-1} - 4(u_{i,j+1} + u_{i,j-1})] = \left\{\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2}\right\}_{(i,j)} + \frac{h^2}{2}[R_1 + R_2 + R_3 + R_4 - 4(R_5 + R_6)].$$

Отбрасывая члены с R_i для уравнения $\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = 0$, получим разностную аппроксимацию

$$4u_{ij} + u_{i+1,j+1} + u_{i+1,j-1} + u_{i-1,j+1} + u_{i-1,j-1} - 4(u_{i,j+1} + u_{i,j-1}) = 0, \quad (30)$$

погрешность которой будет порядка h^2.

Используя указанный способ и привлекая большее количество узлов, можно построить разностные аппроксимации как для дифференциального уравнения (29), так и для общего уравнения (2), имеющие больший порядок точности, причем в зависимости от выбора системы узлов получим явные или неявные схемы. Например, набор узлов, изображенный на рис. 47, даст явную схему, а набор узлов, изображенный на рис. 48, даст неявную схему.

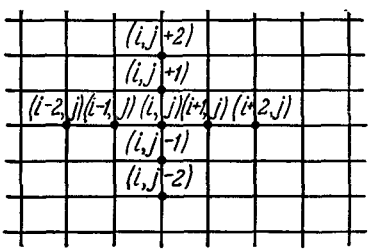

Рис. 47.

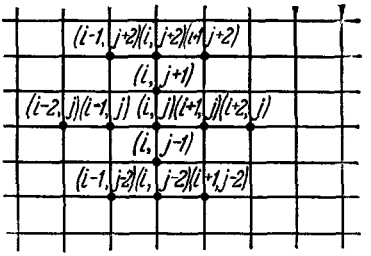

Рис. 48.

Нужно заметить, что разностные аппроксимации высокой точности для дифференциального уравнения можно применять только в том случае, если заранее известно, что решение имеет производные нужных порядков. Кроме того, чтобы не потерять выигрыш в точности решения, который мы хотим получить, применяя разностную аппроксимацию высокого порядка точности для дифферен-

циального уравнения, нужно использовать аппроксимации начальных и граничных условий такой же точности.

Помимо этого нужно всегда помнить, что не всякая разностная схема может быть использована для практического счета, если даже при выборе этой схемы имеет место сходимость последовательности решений, полученных с помощью ее, при неограниченном измельчении сетки к точному решению задачи, так как схема может оказаться непригодной из-за резкого накопления погрешности при счете. Здесь мы приходим к понятию устойчивости разностных схем, которое будет подробно изложено в § 6.

Пример. Найдем методом сеток решение задачи

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = 0;$$

$$u(x, 0) = \varphi(x) = 0{,}2x(1 - x)\sin \pi x; \qquad \left.\frac{\partial u}{\partial y}\right|_{y=0} = 0,$$

$$u|_{x=0} = u|_{x=1} = 0.$$

Для решения применим квадратную сетку с шагом $h = 0{,}05$. Значения искомого решения u_{i0} и u_{i1} $(i = 0, 1, 2, \ldots, 20)$ находим по формулам:

$$u_{i0} = \varphi_i = \varphi(ih),$$

$$u_{i1} = \frac{1}{2}[\varphi_{i+1} + \varphi_{i-1}].$$

Для отыскания приближенных значений на последующих горизонтальных рядах пользуемся уравнением

$$u_{i, j+1} = u_{i+1, j} + u_{i-1, j} - u_{i, j-1}.$$

Результаты приведены в таблице (в единицах четвертого десятичного разряда):

y_i \ x_i	0	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
0	0	15	56	116	185	265	340	405	457	489	500
0,05	0	28	65	122	190	264	335	398	447	478	489
0,10	0	50	94	139	198	260	322	377	419	447	456
0,15	0	66	124	170	209	256	302	343	377	397	405
0,20	0	74	142	194	228	251	277	302	321	335	338
0,25	0	76	144	200	236	249	251	255	260	262	265
0,30	0	70	134	186	221	236	227	209	196	190	186
0,35	0	58	112	155	186	199	194	168	139	120	115
0,40	0	42	79	112	133	144	140	124	92	64	54
0,45	0	21	42	57	70	74	74	64	49	26	13
0,50	0	−1	−1	0	−2	0	−2	−1	−2	−2	−2
$u(x_i, 0{,}5)$	0	0	0	0	0	0	0	0	0	0	0

Значение решения при $x > 0{,}5$ не приведены, так как решение симметрично относительно прямой $x = 0{,}5$. В последней строке приведены значения точного решения задачи при $y = 0{,}5$. Сравнение последней и предпоследней строк показывает, что метод сеток дает очень хорошее совпадение с точным решением.

§ 4. Метод характеристик численного решения гиперболических систем квазилинейных дифференциальных уравнений в частных производных

При изложении метода характеристик мы ограничимся случаем гиперболических систем двух и трех квазилинейных уравнений первого порядка и одним квазилинейным гиперболическим дифференциальным уравнением второго порядка с двумя независимыми переменными [1]).

1. Уравнения характеристик системы квазилинейных дифференциальных уравнений первого порядка. Рассмотрим систему n дифференциальных уравнений в частных производных первого порядка

$$\sum_{j=1}^{n}\left(a_{ij}\frac{\partial u_j}{\partial x}+b_{ij}\frac{\partial u_j}{\partial y}\right)=c_i \qquad (i=1,\ 2,\ \ldots,\ n), \tag{1}$$

где a_{ij}, b_{ij}, c_i — заданные функции переменных x, y, u_1, u_2, $\ldots$, u_n — непрерывные и непрерывно дифференцируемые в некоторой области изменения своих аргументов. Такие системы называют *квазилинейными.*

Пусть $u_1(x,\ y)$, $u_2(x,\ y)$, $\ldots$, $u_n(x,\ y)$ — некоторое непрерывно дифференцируемое в области G плоскости x, y решение системы (1), а C — некоторая гладкая кривая без кратных точек, расположенная в области G. Поставим такой вопрос: можно ли по значениям решения $u_1(x,\ y)$, $u_2(x,\ y)$, $\ldots$, $u_n(x,\ y)$ на кривой C, используя систему (1), найти на кривой C значения частных производных $p_i=\dfrac{\partial u_i}{\partial x}$, $q_i=\dfrac{\partial u_i}{\partial y}$?

Значения частных производных p_i, q_i $(i=1,\ 2,\ \ldots,\ n)$ на кривой C связаны n соотношениями

$$\sum_{j=1}^{n}(a_{ij}p_j+b_{ij}q_j)=c_i \qquad (i=1,\ 2,\ \ldots,\ n), \tag{2}$$

---

[1]) По поводу применения метода характеристик к решению уравнений гиперболического типа см. также статью И. М. Гельфанда, Некоторые задачи теории квазилинейных уравнений, УМН, т. XIV, вып. 2 (86), 1959, 87—158.

получающимися из системы (1), и n дифференциальными соотношениями

$$p_i\,dx + q_i\,dy = du_i \qquad (i = 1,\ 2,\ \ldots,\ n), \tag{3}$$

где дифференциалы берутся вдоль кривой C. Таким образом, для определения p_i, q_i $(i = 1,\ 2,\ \ldots,\ n)$ получаем систему $2n$ линейных уравнений первого порядка.

Предполагая, что в рассматриваемой точке кривой C $dx \neq 0$, систему (3) можно переписать в таком виде:

$$p_i = -q_i \frac{dy}{dx} + \frac{du_i}{dx} \qquad (i = 1,\ 2,\ \ldots,\ n), \tag{4}$$

а затем из системы (2) исключить неизвестные p_i. Для отыскания q_i получим следующую систему уравнений:

$$\sum_{j=1}^{n} (b_{ij}\,dx - a_{ij}\,dy)\,q_j = c_i\,dx - \sum_{j=1}^{n} a_{ij}\,du_j \qquad (i = 1,\ 2,\ \ldots,\ n). \tag{5}$$

Если из этой системы можно найти q_1, q_2, $\ldots$, q_n, то с помощью системы (4) найдутся и p_1, p_2, $\ldots$, p_n (мы предполагали, что в рассматриваемой точке кривой C $dx \neq 0$. Если это не так, то $dy \neq 0$ и систему (3) можно переписать в виде

$$q_i = -p_i \frac{dx}{dy} + \frac{du_i}{dy} \qquad (i = 1,\ 2,\ \ldots,\ n) \tag{4'}$$

и из системы (2) исключить q_i. После исключения получим систему

$$\sum_{j=1}^{n} (a_{ij}\,dy - b_{ij}\,dx)\,p_j = c_i\,dy - \sum_{j=1}^{n} b_{ij}\,du_j \qquad (i = 1,\ 2,\ \ldots,\ n), \tag{5'}$$

определитель которой отличается, может быть, только знаком от определителя системы (5). Обозначим определитель матрицы коэффициентов системы (5) через Δ, т. е.

$$\Delta = \begin{vmatrix} b_{11}\,dx - a_{11}\,dy & b_{12}\,dx - a_{12}\,dy & \ldots & b_{1n}\,dx - a_{1n}\,dy \\ b_{21}\,dx - a_{21}\,dy & b_{22}\,dx - a_{22}\,dy & \ldots & b_{2n}\,dx - a_{2n}\,dy \\ \cdots & \cdots & \cdots & \cdots \\ b_{n1}\,dx - a_{n1}\,dy & b_{n2}\,dx - a_{n2}\,dy & \ldots & b_{nn}\,dx - a_{nn}\,dy. \end{vmatrix} \tag{6}$$

Рассмотрим два случая:

1) определитель Δ не обращается в нуль ни в одной точке кривой C;

2) определитель Δ на кривой C тождественно равен нулю.

В первом случае система (5) имеет относительно q_i единственное решение, а следовательно, в каждой точке кривой C по значениям $u_1(x, y)$, $u_2(x, y)$, $\ldots$, $u_n(x, y)$ на C и системе (1) можно однозначно найти частные производные этих функций.

Во втором случае, так как мы исходим из существующего решения системы (1), система (5) должна быть совместной, но так

как определитель ее равен нулю, то система (5) имеет бесконечно много решений. Таким образом, во втором случае по значениям $u_1(x, y)$, $u_2(x, y)$, ..., $u_n(x, y)$ на C и системе (1) нельзя однозначно определить частные производные решения на кривой C. В этом случае кривую C называют *характеристикой* системы (1), соответствующей данному решению системы. Кривую C вместе со значениями решения вдоль нее, т. е. кривую в $n+2$-мерном пространстве x, y, u_1, u_2, ..., u_n, будем называть *характеристической* кривой. Характеристика C будет ее проекцией на плоскость x, y.

Тангенс угла наклона касательной к характеристике C с осью x $\lambda = \frac{dy}{dx}$ удовлетворяет уравнению

$$\begin{vmatrix} b_{11}-\lambda a_{11} & b_{12}-\lambda a_{12} & \dots & b_{1n}-\lambda a_{1n} \\ b_{21}-\lambda a_{21} & b_{22}-\lambda a_{22} & \dots & b_{2n}-\lambda a_{2n} \\ \dots & \dots & \dots & \dots \\ b_{n1}-\lambda a_{n1} & b_{n2}-\lambda a_{n2} & \dots & b_{nn}-\lambda a_{nn} \end{vmatrix} = 0. \tag{7}$$

В фиксированной точке $(x, y, u_1, u_2, \dots, u_n)$ это будет уравнение степени n относительно λ. Если оно имеет n действительных различных корней, то говорят, что в этой точке система (1) является *гиперболической* системой. Если это свойство имеет место в каждой точке некоторой области пространства x, y, u_1, u_2, ..., u_n, то говорят, что система (1) суть гиперболическая система в этой области. Только такие системы мы и будем рассматривать.

При заданном решении $u_1(x, y)$, $u_2(x, y)$, ..., $u_n(x, y)$ гиперболической системы (1) в каждой точке области G, где это решение определено, уравнение (7) имеет n действительных различных корней, определяющих n направлений касательных к характеристикам, соответствующих данному решению. Обозначая через $\lambda_1, \lambda_2, \dots, \lambda_n$ корни уравнения (7), являющиеся (при заданном решении $u_1(x, y)$, $u_2(x, y)$, ..., $u_n(x, y)$) функциями x и y, мы получим n дифференциальных уравнений

$$dy = \lambda_i(x, y)\, dx \qquad (i = 1, 2, \dots, n). \tag{8}$$

Каждое уравнение определяет однопараметрическое семейство кривых — интегральных кривых этого уравнения, покрывающее область G. Через каждую точку пройдет одна и только одна кривая семейства. Рассматривая все уравнения или, что то же самое, рассматривая уравнение (7) как дифференциальное уравнение первого порядка n-й степени, мы получим при заданном решении $u_1(x, y)$, $u_2(x, y)$, ..., $u_n(x, y)$ системы (1) n налагающихся однопараметрических семейств кривых, или n *семейств характеристик*. Через каждую точку области G будет проходить одна и только одна характеристика каждого семейства.

Если система (1) существенно квазилинейна, то ее характеристики существенно зависят от выбора решения системы и могут быть определены только, если известно решение. В случае линейной системы коэффициенты a_{ij}, b_{ij} в (1) не зависят от u_1, u_2, ..., u_n и из уравнения (7) характеристики могут быть найдены независимо от выбранного решения.

Предположим, что кривая C плоскости x, y есть характеристика системы (1), соответствующая заданному решению $u_1(x, y)$, $u_2(x, y)$, ..., $u_n(x, y)$. На кривой C определитель Δ обращается в нуль, но так как система (5) совместна, то и все определители, получающиеся заменой в Δ k-го столбца столбцом правых частей системы (5), должны также обращаться в нуль. Обозначим определитель, получающийся заменой в Δ i-го столбца столбцом свободных членов системы (5), через Δ_i. Тогда на кривой C функции $u_1(x, y)$, $u_2(x, y)$, ..., $u_n(x, y)$ будут связаны $n+1$ соотношениями

$$\Delta = \Delta_1 = \Delta_2 = \ldots = \Delta_n = 0. \tag{9}$$

Не все эти условия являются независимыми между собой. Так как система (1) по условию гиперболическая, то матрица

$$\begin{pmatrix} b_{11} - \lambda a_{11} & b_{12} - \lambda a_{12} & \ldots & b_{1n} - \lambda a_{1n} \\ b_{21} - \lambda a_{21} & b_{22} - \lambda a_{22} & \ldots & b_{2n} - \lambda a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ b_{n1} - \lambda a_{n1} & b_{n2} - \lambda a_{n2} & \ldots & b_{nn} - \lambda a_{nn} \end{pmatrix}$$

имеет в точках кривой C ранг, равный $n-1$. В этом случае хотя бы одна из матриц, получающихся из последней заменой одного из столбцов столбцом свободных членов системы (5), будет иметь также ранг $n-1$. Пусть это будет матрица, полученная заменой k столбца. Тогда из условий

$$\Delta = 0 \quad \text{и} \quad \Delta_k = 0 \tag{10}$$

будут следовать все остальные условия в (9). Таким образом, на характеристике C решение $u_1(x, y)$, $u_2(x, y)$, ..., $u_n(x, y)$ связано двумя условиями (10), называемыми *уравнениями характеристик*. Первое из них называют *уравнением направления характеристики*, а второе — *дифференциальным соотношением на характеристике*. При данном в области G решении системы (1) мы имеем n семейств характеристик и на каждом из этих семейств имеем свое дифференциальное соотношение.

Заметим еще, что если мы будем рассматривать не характеристику, а характеристическую кривую, то она может принадлежать нескольким решениям системы (1), и если снять требование непрерывной дифференцируемости решения, то при непрерывности решения разрыва первых производных могут быть только на характеристике. Такие решения можно получить следующим образом. Возьмем два решения $u_i^{(1)}(x, y)$ и $u_i^{(2)}(x, y)$ $(i = 1, 2, \ldots, n)$, имею-

щие непрерывные производные в области G, содержащей общую характеристику C, являющуюся проекцией на плоскость (x, y) характеристической кривой, принадлежащей обоим решениям, и рассмотрим решение

$$u_i(x, y) = \begin{cases} u_i^{(1)}(x, y) & \text{с одной стороны } C, \\ u_i^{(2)}(x, y) & \text{с другой стороны } C \end{cases} \qquad (i = 1, 2, \ldots, n).$$

Это решение будет непрерывно в области G, но на C будет иметь разрыв первых производных.

В заключение этого пункта выпишем уравнения характеристик для случаев $n = 1$; $n = 2$; $n = 3$.

$n = 1$. Если уравнение имеет вид

$$a \frac{\partial u}{\partial x} + b \frac{\partial u}{\partial y} = c,$$

то уравнение направления характеристик будет

$$b\,dx - a\,dy = 0, \tag{11}$$

а дифференциальное соотношение на них

$$c\,dx - a\,du = 0. \tag{12}$$

$n = 2$. Уравнения направлений характеристик

$$dy - \lambda_i\,dx = 0 \qquad (i = 1, 2), \tag{13}$$

где λ_i — корни уравнения

$$\begin{vmatrix} b_{11} - \lambda a_{11} & b_{12} - \lambda a_{12} \\ b_{21} - \lambda a_{21} & b_{22} - \lambda a_{22} \end{vmatrix} = 0. \tag{14}$$

Дифференциальные соотношения на характеристиках

$$\begin{vmatrix} c_1\,dx - a_{11}\,du_1 - a_{12}\,du_2 & b_{12} - \lambda_i a_{12} \\ c_2\,dx - a_{21}\,du_1 - a_{22}\,du_2 & b_{22} - \lambda_i a_{22} \end{vmatrix} = 0 \tag{15}$$

или

$$(\lambda_i A + B)\,du_1 + C\,du_2 + M\,dx + N\,dy = 0, \tag{16}$$

где

$$A = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}; \quad B = \begin{vmatrix} b_{12} & a_{11} \\ b_{22} & a_{21} \end{vmatrix}; \quad C = \begin{vmatrix} b_{12} & a_{12} \\ b_{22} & a_{22} \end{vmatrix}; \tag{17}$$

$$M = \begin{vmatrix} c_1 & b_{12} \\ c_2 & b_{22} \end{vmatrix}; \quad N = \begin{vmatrix} a_{12} & c_1 \\ a_{22} & c_2 \end{vmatrix}.$$

$n = 3$. Уравнения направлений характеристик

$$dy - \lambda_i\,dx = 0 \qquad (i = 1, 2, 3), \tag{18}$$

где λ_i — корни уравнения

$$\begin{vmatrix} b_{11}-\lambda a_{11} & b_{12}-\lambda a_{12} & b_{13}-\lambda a_{13} \\ b_{21}-\lambda a_{21} & b_{22}-\lambda a_{22} & b_{23}-\lambda a_{23} \\ b_{31}-\lambda a_{31} & b_{32}-\lambda a_{32} & b_{33}-\lambda a_{33} \end{vmatrix} = 0. \tag{19}$$

Дифференциальные соотношения на характеристиках

$$\begin{vmatrix} c_1\,dx - a_{11}\,du_1 - a_{12}\,du_2 - a_{13}\,du_3 & b_{12}-\lambda_i a_{12} & b_{13}-\lambda_i a_{13} \\ c_2\,dx - a_{21}\,du_1 - a_{22}\,du_2 - a_{23}\,du_3 & b_{22}-\lambda_i a_{22} & b_{23}-\lambda_i a_{23} \\ c_3\,dx - a_{31}\,du_1 - a_{32}\,du_2 - a_{33}\,du_3 & b_{32}-\lambda_i a_{32} & b_{33}-\lambda_i a_{33} \end{vmatrix} = 0 \tag{20}$$

или

$$M_i\,du_1 + N_i\,du_2 + P_i\,du_3 - C_i\,dx = 0, \tag{21}$$

где

$$\left.\begin{aligned} M_i &= \begin{vmatrix} a_{11} & b_{12}-\lambda_i a_{12} & b_{13}-\lambda_i a_{13} \\ a_{21} & b_{22}-\lambda_i a_{22} & b_{23}-\lambda_i a_{23} \\ a_{31} & b_{32}-\lambda_i a_{32} & b_{33}-\lambda_i a_{33} \end{vmatrix}; \quad N_i = \begin{vmatrix} a_{12} & b_{12} & b_{13}-\lambda_i a_{13} \\ a_{22} & b_{22} & b_{23}-\lambda_i a_{23} \\ a_{32} & b_{32} & b_{33}-\lambda_i a_{33} \end{vmatrix}; \\ P_i &= \begin{vmatrix} a_{13} & b_{12}-\lambda_i a_{12} & b_{13} \\ a_{23} & b_{22}-\lambda_i a_{22} & b_{23} \\ a_{33} & b_{32}-\lambda_i a_{32} & b_{33} \end{vmatrix}; \quad C_i = \begin{vmatrix} c_1 & b_{12}-\lambda_i a_{12} & b_{13}-\lambda_i a_{13} \\ c_2 & b_{22}-\lambda_i a_{22} & b_{23}-\lambda_i a_{23} \\ c_3 & b_{32}-\lambda_i a_{32} & b_{33}-\lambda_i a_{33} \end{vmatrix}. \end{aligned}\right\} \tag{22}$$

Может случиться, что условие (20) будет удовлетворяться на какой-либо характеристике тождественно. В этом случае вместо условия (20) нужно на этой характеристике взять другое условие, которое может быть получено заменой в определителе (19) другого столбца столбцом свободных членов системы (5).

2. Примеры: уравнения характеристик для некоторых систем дифференциальных уравнений газовой динамики. В качестве примеров приведем дифференциальные уравнения характеристик для некоторых систем дифференциальных уравнений газовой динамики, где метод характеристик находит широкое применение.

1. *Плоское и осесимметричное сверхзвуковое установившееся течение идеального газа. Случай безвихревого движения.* Система дифференциальных уравнений в этом случае имеет вид

$$H\frac{\partial U}{\partial x} + K\left(\frac{\partial U}{\partial y} + \frac{\partial V}{\partial x}\right) + L\,\frac{\partial V}{\partial y} + \sigma R = 0$$

$$\frac{\partial U}{\partial y} - \frac{\partial V}{\partial x} = 0,$$

где

$$H = 1 - \frac{k+1}{k-1}\,U^2 - V^2; \quad L = 1 - U^2 - \frac{k+1}{k-1}\,V^2; \quad K = -\frac{2}{k-1}\,UV;$$

$$R = \frac{V}{y}\,(1 - U^2 - V^2);$$

U и V — составляющие безразмерной скорости по направлениям x и y, выражающиеся через составляющие действительной скорости u, v по формулам $U = \frac{u}{W_m}$, $V = \frac{v}{W_m}$, $W_m = \sqrt{\frac{2}{k-1}}\,a_0$, где k — постоянная

газового потока, a_0 — скорость звука в покоящемся газе. Для плоского движения $\sigma = 0$, для осесимметричного $\sigma = 1$ (в этом случае x — ось симметрии).

Уравнения характеристик для этой системы будут следующие: уравнения направления характеристик

$$dy - \lambda_i\, dx = 0 \qquad (i = 1,\ 2),$$

где λ_i — корни уравнения

$$\begin{vmatrix} K - \lambda H & L - \lambda K \\ 1 & \lambda \end{vmatrix} = 0,$$

или

$$\lambda_1 = \frac{K - \sqrt{K^2 - LH}}{H}; \quad \lambda_2 = \frac{K + \sqrt{K^2 - LH}}{H}.$$

При сверхзвуковых скоростях λ_1, λ_2 действительны и различны. Вычисляя определители (17), получим:

$$A = -H; \quad B = 0; \quad C = -L; \quad M = 0; \quad N = -\sigma R.$$

Отсюда, используя (16), получим следующие дифференциальные соотношения на характеристиках:

$$\lambda_i H\, dU + L\, dV + \sigma R\, dy = 0.$$

Учитывая, что $\lambda_1\lambda_2 = \frac{L}{H}$, и заменяя dy из уравнений направлений характеристик, получим следующие дифференциальные уравнения характеристик рассматриваемой системы:

1-е семейство	2-е семейство
$dy = \lambda_1\, dx,$	$dy = \lambda_2\, dx,$
$dU + \lambda_2\, dV + \frac{\sigma R}{\lambda_1 H} dx = 0;$	$dU + \lambda_1\, dV + \frac{\sigma R}{\lambda_2 H} dx = 0.$

Иногда уравнения характеристик записывают в другой форме, принимая за искомые функции W и θ, где

$$W = \sqrt{U^2 + V^2}; \quad U = W\cos\theta; \quad V = W\sin\theta$$

(W имеет смысл абсолютной величины безразмерной скорости в данной точке, а θ — угол наклона направления скорости с осью x). Если, кроме того, ввести угол μ (угол Маха) с помощью соотношений

$$\sin\mu = \sqrt{\frac{k-1}{2}}\,\frac{\sqrt{1 - W^2}}{W}; \quad \cos\mu = \sqrt{\frac{k-1}{2}}\,\frac{\sqrt{\frac{k+1}{k-1}W^2 - 1}}{W},$$

то, выполнив замену искомых функций в уравнениях характеристик, придем к следующему результату:

1-е семейство	2-е семейство
$\frac{dy}{dx} = \operatorname{tg}(\theta + \mu),$	$\frac{dy}{dx} = \operatorname{tg}(\theta - \mu),$
$\frac{dW}{W} - \operatorname{tg}\mu\, d\theta - \sigma l \frac{dx}{y} = 0,$	$\frac{dW}{W} + \operatorname{tg}\mu\, d\theta - \sigma m \frac{dx}{y} = 0,$
где	где
$l = \frac{\sin\mu \sin\theta \operatorname{tg}\mu}{\cos(\mu + \theta)}.$	$m = \frac{\sin\mu \sin\theta \operatorname{tg}\mu}{\cos(\theta - \mu)}.$

Из этих уравнений следует, что если в точке (x, y) известно направление скорости θ и мы отложим по ту и другую сторону от него углы, равные μ, то получим направления характеристик, проходящих через эту точку. Эти направления называют направлениями Маха, а линии, имеющие в каждой точке направление Маха, называют линиями Маха. Таким образом, характеристики совпадают с линиями Маха. Физический смысл их состоит в том, что если в некоторой точке сверхзвукового потока поместить источник малых возмущений, то они будут сказываться только в области, ограниченной линиями Маха, выходящими из этой точки (в осесимметричном пространственном случае это будет конус Маха).

2. *Плоское и осесимметричное сверхзвуковое установившееся движение идеального газа. Случай вихревого движения.* Система уравнений в этом случае имеет вид

$$H\frac{\partial U}{\partial x}+K\left(\frac{\partial U}{\partial y}+\frac{\partial V}{\partial x}\right)+L\frac{\partial V}{\partial y}+\sigma R=0,$$

$$V\left(\frac{\partial U}{\partial y}-\frac{\partial V}{\partial x}\right)-Q\frac{\partial S}{\partial x}=0,$$

$$U\left(\frac{\partial U}{\partial y}-\frac{\partial V}{\partial x}\right)+Q\frac{\partial S}{\partial y}=0,$$

где U, V, H, K, L, R имеют прежний смысл, $Q=\frac{1}{2c_p}(1-U^2-V^2)$, S — энтропия, c_p — постоянная величина (удельная теплоемкость газа при постоянном давлении).

Для того чтобы выписать дифференциальные уравнения характеристик, найдем корни уравнения (19) и определители M_i, N_i, P_i, C_i (12). Будем иметь:

$$\begin{vmatrix} K-\lambda H & L-\lambda K & 0 \\ V & \lambda V & \lambda Q \\ U & \lambda U & Q \end{vmatrix} = -Q(V-\lambda U)(\lambda^2 H-2K\lambda+L)=0,$$

или

$$\lambda_1=\frac{V}{U};\quad \lambda_2=\frac{K-\sqrt{K^2-HL}}{H};\quad \lambda_3=\frac{K+\sqrt{K^2-HL}}{H}.$$

Далее,

$$M_i=\lambda_i HQ(V-\lambda_i U);\quad N_i=LQ(V-\lambda_i U);$$

$$P_i=Q^2(L-\lambda_i K);\qquad C_i=-\sigma RQ\lambda_i(V-\lambda_i U).$$

Подставляя их в соотношение (20), получим следующие дифференциальные соотношения на характеристиках:

$$Q[(V-\lambda_i U)(\lambda_i H\,dU+L\,dV-\sigma R\lambda_i\,dx)+Q(L-\lambda_i K)\,dS]=0.$$

При $i=1$ имеем:

$$Q^2(L-\lambda_1 K)\,dS=0,\quad \text{или}\quad dS=0.$$

Окончательно имеем следующие дифференциальные уравнения для характеристик:

1-е семейство

$$U\,dy - V\,dx = 0,$$

$$dS = 0.$$

2-семейство

$$dy - \lambda_2\,dx = 0,$$

$$(V - \lambda_2 U)(H\,dU + \lambda_3 H\,dV - \sigma R\,dx) + Q(H - K)\,dS = 0.$$

3-е семейство

$$dy - \lambda_3\,dx = 0,$$

$$(V - \lambda_3 U)(H\,dU + \lambda_2 H\,dV - \sigma R\,dx) + Q(H - K)\,dS = 0.$$

Первое семейство характерно тем, что направление характеристики в каждой точке совпадает с направлением скорости потока в этой точке, т. е. характеристики первого семейства являются линиями тока; энтропия вдоль них сохраняет постоянное значение. Две другие характеристики, выходящие из данной точки, будут совпадать с линиями Маха.

3. *Одномерное неустановившееся течение в трубах. Безвихревое течение.* Система дифференциальных уравнений движения идеального газа в этом случае имеет вид

$$u\frac{\partial u}{\partial x} + \frac{\partial u}{\partial t} + \frac{2}{k-1}a\frac{\partial a}{\partial x} = 0,$$

$$a\frac{\partial u}{\partial x} + \frac{2}{k-1}\left(u\frac{\partial a}{\partial x} + \frac{\partial a}{\partial t}\right) = -au\frac{a\ln f}{dx},$$

где x — координата вдоль оси трубы, u — скорость в сечении x трубы в момент времени t, a — местная скорость звука, $f = f(x)$ — площадь поперечного сечения трубы.

Уравнения направлений характеристик имеют вид

$$dx - \lambda_i\,dt = 0 \qquad (i = 1, 2),$$

где λ_i — корни уравнения

$$\begin{vmatrix} u - \lambda & \dfrac{2a}{k-1} \\ a & \dfrac{2}{k-1}(u - \lambda) \end{vmatrix} = 0,$$

или

$$\lambda_{1,2} = u \pm a.$$

Здесь мы видим, что система будет гиперболической всегда, в то время как в случае установившихся течений она будет гиперболической лишь в сверхзвуковой области.

Определители A, B, C, M, N будут иметь следующие значения:

$$A = \frac{2}{k-1};\quad B = -\frac{2u}{k-1};\quad C = \frac{4a}{(k-1)^2};\quad M = \frac{2}{k-1}a^2u\frac{d\ln f}{dx};\quad N = 0.$$

Следовательно, дифференциальные соотношения на характеристиках будут:

$$(\lambda_i - u)\,du + \frac{2}{k-1}a\,da + a^2u\frac{d\ln f}{dx}\,dt = 0$$

или, подставляя λ_i и сокращая на a, будем иметь:

$$\frac{2}{k-1}\,a\,da \pm du + au\frac{d\ln f}{dx}\,dt = 0.$$

Итак, дифференциальные уравнения характеристик таковы:

1-е семейство

$$dx - (u+a)\,dt = 0;\quad \frac{2}{k-1}\,a\,da + du + au\frac{d\ln f}{dx}\,dt = 0.$$

2-е семейство

$$dx - (u-a)\,dt = 0;\quad \frac{2}{k-1}\,a\,da - du + au\frac{d\ln f}{dx}\,dt = 0.$$

4. *Одномерное неустановившееся течение в трубах. Вихревое течение.* В этом случае уравнения движения имеют вид

$$u\frac{\partial u}{\partial x} + \frac{\partial u}{\partial t} + \frac{2}{k-1}\,a\frac{\partial a}{\partial x} - \frac{a^2}{k(k-1)\,c_v}\frac{\partial S}{\partial x} = 0,$$

$$a\frac{\partial u}{\partial x} + \frac{2}{k-1}\left(u\frac{\partial a}{\partial x} + \frac{\partial a}{\partial t}\right) = -au\,\frac{d\ln f}{dx},$$

$$u\frac{\partial S}{\partial x} + \frac{\partial S}{\partial t} = 0,$$

где u, a имеют прежний смысл, S — энтропия, c_v — постоянная (удельная теплоемкость газа при постоянном объеме).

Уравнения направлений характеристик

$$dx - \lambda_i\,dt = 0 \qquad (i = 1,\ 2,\ 3),$$

где λ_i — корни уравнения

$$\begin{vmatrix} u-\lambda & \dfrac{2a}{k-1} & -\dfrac{a^2}{k(k-1)\,c_v} \\ a & \dfrac{2}{k-1}(u-\lambda) & 0 \\ 0 & 0 & u-\lambda \end{vmatrix} = 0,$$

или

$$\lambda_1 = u;\quad \lambda_2 = u + a;\quad \lambda_3 = u - a.$$

Дифференциальные соотношения на характеристиках получим из уравнения (20):

$$\begin{vmatrix} du & a & -\dfrac{a^2}{k(k-1)\,c_v} \\ au\,dt\,\dfrac{d\ln f}{dx} + \dfrac{2}{k-1}\,da & u-\lambda_i & 0 \\ ds & 0 & u-\lambda_i \end{vmatrix} =$$

$$= (u-\lambda_i)\left[(u-\lambda_i)\,du - a^2u\,dt\frac{d\ln f}{dx} - \frac{2a}{k-1}\,da + \frac{a^2}{k(k-1)\,c_v}\,dS\right] = 0.$$

При $l=1$ это условие превращается в тождество. Поэтому условие на характеристиках первого семейства получим, используя другое соотношение:

$$\begin{vmatrix} u-\lambda_i & a & du \\ a & u-\lambda_i & au\,dt\,\dfrac{d\ln f}{dx}+\dfrac{2}{k-1}da \\ 0 & 0 & dS \end{vmatrix} = dS\left[(u-\lambda_i)^2-a^2\right]=0,$$

откуда следует, что на характеристиках первого семейства

$$dS=0.$$

Окончательно будем иметь следующие уравнения характеристик:

1-е семейство

$$dx-u\,dt=0,$$

$$dS=0.$$

2-е семейство

$$dx-(u+a)\,dt=0,$$

$$du-\frac{2}{k-1}da+\frac{a}{k(k-1)c_v}dS-au\frac{d\ln f}{dx}dt=0.$$

3-е семейство

$$dx-(u-a)\,dt=0,$$

$$du+\frac{2}{k-1}da-\frac{a}{k(k-1)c_v}dS+au\frac{d\ln f}{dx}dt=0.$$

3. Уравнения характеристик квазилинейного гиперболического дифференциального уравнения второго порядка. Рассмотрим теперь квазилинейное дифференциальное уравнение второго порядка

$$a\frac{\partial^2 u}{\partial x^2}+2b\frac{\partial^2 u}{\partial x\,\partial y}+c\frac{\partial^2 u}{\partial y^2}=f, \tag{23}$$

где a, b, c, f — заданные функции x, y, u, $\dfrac{\partial u}{\partial x}$, $\dfrac{\partial u}{\partial y}$, непрерывные и непрерывно дифференцируемые в некоторой области изменения своих аргументов.

Предположим, что в плоскости x, y задана некоторая гладкая кривая C: $\{x=x(\eta);\ y=y(\eta);\ x'^2(\eta)+y'^2(\eta)>0\}$. Пусть вдоль кривой C задана функция $u(x, y)$, являющаяся дважды непрерывно дифференцируемым в области G, содержащей кривую C, решением уравнения (23), а также заданы и ее производные первого порядка вдоль C: $p=\dfrac{\partial u}{\partial x}$, $q=\dfrac{\partial u}{\partial y}$. Снова поставим вопрос: можно ли на C найти частные производные второго порядка $r=\dfrac{\partial^2 u}{\partial x^2}$, $s=\dfrac{\partial^2 u}{\partial x\,\partial y}$, $t=\dfrac{\partial^2 u}{\partial y^2}$, используя уравнение (23)?

Для отыскания r, s и t вдоль кривой C имеем три соотношения:

$$\left.\begin{aligned} ar+2bs+ct&=f,\\ r\,dx+s\,dy&=dp,\\ s\,dx+t\,dy&=dq,\end{aligned}\right\} \tag{24}$$

где все дифференциалы берутся вдоль кривой C.

Рассмотрим (24) как систему трех линейных алгебраических уравнений для неизвестных r, s и t с определителем

$$\Delta=\begin{vmatrix} a & 2b & c\\ dx & dy & 0\\ 0 & dx & dy\end{vmatrix}=a\,dy^2-2b\,dx\,dy+c\,dx^2.$$

Нас будут интересовать два случая:

1) определитель Δ отличен от нуля на кривой C;

2) определитель Δ тождественно равен нулю на кривой C.

В первом случае вторые производные r, s и t функции $u(x, y)$ определяются вдоль кривой C единственным образом.

Во втором случае, так как мы исходим из существующего решения $u(x, y)$, система (24) будет совместна, и мы получим бесконечное множество значений r, s и t в каждой точке кривой C. В этом случае кривую C называют *характеристикой* уравнения (23), соответствующей заданному решению $u(x, y)$, а кривую C вместе с заданными на ней значениями u, p, q — *характеристической кривой*.

Если кривая C является характеристикой при заданном решении $u(x, y)$, то вдоль нее имеет место соотношение

$$a\,dy^2-2b\,dx\,dy+c\,dx^2=0, \tag{25}$$

или

$$a\left(\frac{dy}{dx}\right)^2-2b\frac{dy}{dx}+c=0. \tag{26}$$

Разрешая это уравнение относительно $\frac{dy}{dx}$, получим:

$$\frac{dy}{dx}=\frac{b\pm\sqrt{b^2-ac}}{a}. \tag{27}$$

Если $b^2-ac>0$, то получим два обыкновенных дифференциальных уравнения первого порядка, которые определят два однопараметрических семейства интегральных кривых, покрывающих область G, где определено решение $u(x, y)$. Эти два семейства называют первым и вторым семействами характеристик, соответствующими данному решению $u(x, y)$ уравнения (23). Через каждую точку области G проходит одна и только одна характеристика каждого семейства.

Если в некоторой области изменения x, y, u, p и q уравнение (26) имеет два действительных различных корня для $\frac{dy}{dx}$, то говорят,

что в этой области уравнение принадлежит к гиперболическому типу. Только такие уравнения мы и будем рассматривать.

Если уравнение (23) линейно, т. е. a, b и c не зависят от u, $\frac{\partial u}{\partial x}$, $\frac{\partial u}{\partial y}$, то характеристики не зависят от выбора решения и оба семейства характеристик можно найти, интегрируя уравнения (27).

Если кривая C для данного решения $u(x, y)$ является характеристикой, то из совместности системы (24) следует, что все определители третьего порядка матрицы

$$\begin{pmatrix} a & 2b & c & f \\ dx & dy & 0 & dp \\ 0 & dx & dy & dq \end{pmatrix}$$

на кривой C должны обращаться в нуль, т. е.

$$\left.\begin{aligned} \Delta_1 &= \begin{vmatrix} 2b & c & f \\ dy & 0 & dp \\ dx & dy & dq \end{vmatrix} = c\,(dp\,dx - dq\,dy) - 2b\,dp\,dy + f\,dy^2 = 0, \\ \Delta_2 &= \begin{vmatrix} a & c & f \\ dx & 0 & dp \\ 0 & dy & dq \end{vmatrix} = -\,c\,dq\,dx - a\,dp\,dy + f\,dx\,dy = 0, \\ \Delta_3 &= \begin{vmatrix} a & 2b & f \\ dx & dy & dp \\ 0 & dx & dq \end{vmatrix} = a\,(dq\,dy - dp\,dx) - 2b\,dq\,dx + f\,dx^2 = 0. \end{aligned}\right\} \quad (28)$$

Используя условие, что C есть характеристика и исключая из соотношений (28) dy с помощью соотношения (27), во всех трех случаях получим:

$$(a\,dp - f\,dx)\left(b \pm \sqrt{b^2 - ac}\right) + ac\,dq = 0. \quad (29)$$

Знаки в (29) соответствуют знакам в (27). Таким образом, имеют место соотношения:

$$\left.\begin{gathered} a\,dy - \left(b \pm \sqrt{b^2 - ac}\right) dx = 0, \\ (a\,dp - f\,dx)\left(b \pm \sqrt{b^2 - ac}\right) + ac\,dq = 0, \\ du = p\,dx + q\,dy, \end{gathered}\right\} \quad (30)$$

называемые *уравнениями характеристик*. Первое из них называют *уравнением направления характеристик*, а последние два — *дифференциальными соотношениями вдоль характеристик*.

Если ввести обозначения

$$\lambda_1 = \frac{b - \sqrt{b^2 - ac}}{a}; \quad \lambda_2 = \frac{b + \sqrt{b^2 - ac}}{a}, \quad (31)$$

то уравнения характеристик можно переписать следующим образом: для первого семейства

$$dy - \lambda_1 dx = 0; \quad a(dp + \lambda_2 dq) - f dx = 0; \quad du = p dx + q dy; \tag{32}$$

для второго семейства

$$dy - \lambda_2 dx = 0; \quad a(dp + \lambda_1 dq) - f dx = 0; \quad du = p dx + q dy. \tag{33}$$

Относительно характеристических кривых уравнения (23) можно сделать такое же замечание, как и о характеристических кривых системы уравнений первого порядка, т. е. она может принадлежать нескольким поверхностям $u = u(x, y)$, которые будут касаться вдоль нее. Можно построить решения уравнения (23), которые на характеристике C будут непрерывны вместе с первыми производными, а вторые производные будут терпеть разрыв. Разрывы такого рода называют слабыми разрывами.

4. Численное решение квазилинейной гиперболической системы двух дифференциальных уравнений первого порядка методом Массо[1]. Для численного решения различных задач для гиперболических систем квазилинейных дифференциальных уравнений первого порядка может быть применен метод Массо, в основе которого лежит замена дифференциальных уравнений характеристик, выведенных в п. 1, соответствующими конечноразностными уравнениями. Мы подробно изложим этот метод применительно к системе двух квазилинейных уравнений.

Идея метода следующая. В плоскости x, y рассмотрим две близкие точки *1* и *2* (рис. 49). Обозначим координаты этих точек через (x_1, y_1) и (x_2, y_2). Пусть в этих точках известны значения искомых функций u и v, удовлетворяющих квазилинейной гиперболической системе уравнений

$$\left.\begin{aligned} a_{11}\frac{\partial u}{\partial x} + b_{11}\frac{\partial u}{\partial y} + a_{12}\frac{\partial v}{\partial x} + b_{12}\frac{\partial v}{\partial y} = c_1, \\ a_{21}\frac{\partial u}{\partial x} + b_{21}\frac{\partial u}{\partial y} + a_{22}\frac{\partial v}{\partial x} + b_{22}\frac{\partial v}{\partial y} = c_2. \end{aligned}\right\} \tag{34}$$

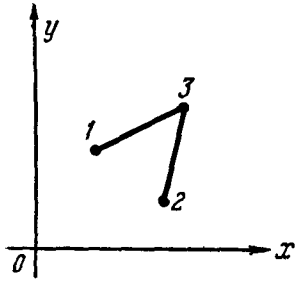

Рис. 49.

Их значения в точках *1* и *2* обозначим соответственно u_1, v_1 и u_2, v_2. Через точку *1* проведем прямую в направлении характеристики первого семейства характеристик, выходящей из точки *1*, а через точку *2* — прямую в направлении характеристики второго семейства, выходящей из точки *2*. Эти пря-

[1]) Метод Массо улучшался и применялся для решения ряда задач различными авторами, в частности Христиановичем.

мые пересекутся в некоторой точке *3*. Координаты $x_3^{(1)}$, $y_3^{(1)}$ этой точки являются решением системы

$$\left.\begin{aligned} y_3^{(1)} - y_1 &= \lambda_{1,1}^{(1)}\left(x_3^{(1)} - x_1\right), \\ y_3^{(1)} - y_2 &= \lambda_{2,2}^{(1)}\left(x_3^{(1)} - x_2\right). \end{aligned}\right\} \tag{35}$$

где $\lambda_{11}^{(1)}$ — угловой коэффициент касательной к характеристике первого семейства в точке *1*, $\lambda_{22}^{(1)}$ — угловой коэффициент касательной к характеристике второго семейства в точке *2*, являющиеся соответствующими корнями уравнения (14) в точках *1* и *2*.

Уравнения (35) получаются из уравнения направления характеристики первого семейства в точке *1* и уравнения направления характеристики второго семейства в точке *2* заменой входящих в них дифференциалов конечными разностями.

Далее, заменяя дифференциалы, входящие в дифференциальные соотношения на соответствующих характеристиках, конечными разностями, получим систему уравнений для определения значений u и v в точке *3*, которые мы обозначим через $u_3^{(1)}$ и $v_3^{(1)}$. Эта система имеет вид

$$\left.\begin{aligned} &\left(\lambda_{11}^{(1)}A_1^{(1)} + B_1^{(1)}\right)\left(u_3^{(1)} - u_1\right) + C_1^{(1)}\left(v_3^{(1)} - v_1\right) + M_1^{(1)}\left(x_3^{(1)} - x_1\right) + \\ &\qquad + N_1^{(1)}\left(y_3^{(1)} - y_1\right) = 0, \\ &\left(\lambda_{22}^{(1)}A_2^{(1)} + B_2^{(1)}\right)\left(u_3^{(1)} - u_2\right) + C_2^{(1)}\left(v_3^{(1)} - v_2\right) + M_2^{(1)}\left(x_3^{(1)} - x_2\right) + \\ &\qquad + N_2^{(1)}\left(y_3^{(1)} - y_2\right) = 0, \end{aligned}\right\} \tag{36}$$

где $A_i^{(1)}$, $B_i^{(1)}$, $C_i^{(1)}$, $M_i^{(1)}$, $N_i^{(1)}$ $(i = 1, 2)$ суть значения определителей (17) в точке i. Решая эту систему относительно $u_3^{(1)}$, $v_3^{(1)}$, найдем первое приближение функций u и v в точке *3*. Это приближение может оказаться недостаточно точным, так как мы заменили характеристики, выходящие из точек *1* и *2*, отрезками прямых, в то время как точка *3* на самом деле должна быть точкой пересечения, вообще говоря, криволинейных характеристик, и, кроме того, дифференциалы всюду мы заменяем конечными приращениями. Поэтому может возникнуть необходимость в уточнении координат точки *3* и значений u и v в этой точке. Это уточнение можно выполнять двумя способами.

Первый способ. Вычисляют угловые коэффициенты $\lambda_{13}^{(1)}$ и $\lambda_{23}^{(1)}$ характеристик первого и второго семейства в точке *3*, найденной в первом приближении, и вводят средние арифметические

$$\lambda_{11}^{(2)} = \frac{1}{2}\left(\lambda_{11}^{(1)} + \lambda_{13}^{(1)}\right); \quad \lambda_{22}^{(2)} = \frac{1}{2}\left(\lambda_{22}^{(1)} + \lambda_{23}^{(1)}\right). \tag{37}$$

Точно так же находят средние арифметические

$$\left.\begin{aligned} &A_i^{(2)}=\frac{1}{2}\left(A_i^{(1)}+A_3^{(1)}\right);\quad B_i^{(2)}=\frac{1}{2}\left(B_i^{(1)}+B_3^{(1)}\right);\quad C_i^{(2)}=\frac{1}{2}\left(C_i^{(1)}+C_3^{(1)}\right);\\ &M_i^{(2)}=\frac{1}{2}\left(M_i^{(1)}+M_3^{(1)}\right);\quad N_i^{(2)}=\frac{1}{2}\left(N_i^{(1)}+N_3^{(1)}\right)\qquad (i=1,\ 2),\end{aligned}\right\} \tag{38}$$

где $A_3^{(1)}$, $B_3^{(1)}$, $C_3^{(1)}$, $M_3^{(1)}$, $N_3^{(1)}$ — значения A, B, C, M, N в найденном первом приближении точки $\left(x_3^{(1)},\ y_3^{(1)},\ u_3^{(1)},\ v_3^{(1)}\right)$. Искомые величины второго приближения точки 3: $x_3^{(2)}$, $y_3^{(2)}$, $u_3^{(2)}$, $v_3^{(2)}$, находят, решая последовательно следующие системы линейных алгебраических уравнений:

$$\left.\begin{aligned} y_3^{(2)}-y_1&=\lambda_{11}^{(2)}\left(x_3^{(2)}-x_1\right),\\ y_3^{(2)}-y_2&=\lambda_{22}^{(2)}\left(x_3^{(2)}-x_2\right),\end{aligned}\right\} \tag{35'}$$

$$\left.\begin{aligned} &\left(\lambda_{11}^{(2)}A_1^{(2)}+B_1^{(2)}\right)\left(u_3^{(2)}-u_1\right)+C_1^{(2)}\left(v_3^{(2)}-v_1\right)+M_1^{(2)}\left(x_3^{(2)}-x_1\right)+\\ &\qquad+N_1^{(2)}\left(y_3^{(2)}-y_1\right)=0,\\ &\left(\lambda_{22}^{(2)}A_2^{(2)}+B_2^{(2)}\right)\left(u_3^{(2)}-u_2\right)+C_2^{(2)}\left(v_3^{(2)}-v_2\right)+M_2^{(2)}\left(x_3^{(2)}-x_2\right)+\\ &\qquad+N_2^{(2)}\left(y_3^{(2)}-y_2\right)=0.\end{aligned}\right\} \tag{36'}$$

Получим уточненные значения координат точки 3: $x_3^{(2)}$, $y_3^{(2)}$, и уточненные значения искомых функций в этой точке $u_3^{(2)}$, $v_3^{(2)}$. Если их еще раз нужно уточнять, поступаем аналогично, принимая в качестве точки 3 вновь полученное приближение. Процесс продолжают до тех пор, пока значения величин x_3, y_3, u_3, v_3 для точки 3, полученные при двух последовательных приближениях, будут совпадать с заданной точностью [1]).

Второй способ. Используя найденные значения $x_3^{(1)}$, $y_3^{(1)}$, $u_3^{(1)}$, $v_3^{(1)}$, находим:

$$\left.\begin{aligned} &x_{i3}^{(1)}=\frac{1}{2}\left(x_i+x_3^{(1)}\right);\quad y_{i3}^{(1)}=\frac{1}{2}\left(y_i+y_3^{(1)}\right);\quad u_{i3}^{(1)}=\frac{1}{2}\left(u_i+u_3^{(1)}\right);\\ &v_{i3}^{(1)}=\frac{1}{2}\left(v_i+v_3^{(1)}\right)\qquad (i=1,\ 2)\end{aligned}\right\} \tag{39}$$

и за $\lambda_{13}^{(2)}$ принимаем первый корень уравнения (14), в котором a_{ij}, b_{ij} взяты для точки $\left(x_{13}^{(1)},\ y_{13}^{(1)},\ u_{13}^{(1)},\ v_{13}^{(1)}\right)$, а за $\lambda_{23}^{(2)}$ принимаем второй корень уравнения (14), в котором a_{ij}, b_{ij} взяты в точке $\left(x_{23}^{(1)},\ y_{23}^{(1)},\ u_{23}^{(1)},\ v_{23}^{(1)}\right)$; далее вычисляем значения определителей A, B,

---

[1]) Если расстояние между точками 1 и 2 невелико, то достаточно сделать два уточнения, так как в дальнейшем точность возрастать не будет.

C, M, N в точке $\left(x_{13}^{(1)}, y_{13}^{(1)}, u_{13}^{(1)}, v_{13}^{(1)}\right)$: $A_{13}^{(2)}$, $B_{13}^{(2)}$, $C_{13}^{(2)}$, $M_{13}^{(2)}$, $N_{13}^{(2)}$ и в точке $\left(x_{23}^{(1)}, y_{23}^{(1)}, u_{23}^{(1)}, v_{23}^{(1)}\right)$: $A_{23}^{(2)}$, $B_{23}^{(2)}$, $C_{23}^{(2)}$, $M_{23}^{(2)}$, $N_{23}^{(2)}$ и находим $x_3^{(2)}$, $y_3^{(2)}$, $u_3^{(2)}$, $v_3^{(2)}$, последовательно решая системы

$$\left.\begin{aligned} y_3^{(2)} - y_1 &= \lambda_{13}^{(2)}\left(x_{32}^{(2)} - x_1\right), \\ y_3^{(2)} - y_2 &= \lambda_{23}^{(2)}\left(x_{32}^{(2)} - x_2\right); \end{aligned}\right\} \tag{35''}$$

$$\left.\begin{aligned} \left(\lambda_{13}^{(2)} A_{13}^{(2)} + B_{13}^{(2)}\right)\left(u_3^{(2)} - u_1\right) + C_{13}^{(2)}\left(v_3^{(2)} - v_1\right) + M_{13}^{(2)}\left(x_3^{(2)} - x_1\right) + \\ + N_{13}^{(2)}\left(y_3^{(2)} - y_1\right) = 0, \\ \left(\lambda_{23}^{(2)} A_{23}^{(2)} + B_{23}^{(2)}\right)\left(u_3^{(2)} - u_2\right) + C_{23}^{(2)}\left(v_3^{(2)} - v_2\right) + M_{23}^{(2)}\left(x_3^{(2)} - x_2\right) + \\ + N_{23}^{(2)}\left(y_3^{(2)} - y_2\right) = 0. \end{aligned}\right\} \tag{36''}$$

Для дальнейшего уточнения процесс продолжаем аналогично, используя вновь найденные значения величин в точке 3 [1]).

Точность, с которой можно получить значения x_3, y_3, u_3, v_3 в точке 3, естественно зависит также и от близости точек 1 и 2.

Умея решать описанную выше элементарную задачу отыскания точки (x_3, y_3, u_3, v_3) по двум известным точкам (x_1, y_1, u_1, v_1) и (x_2, y_2, u_2, v_2), можно численно решать различные задачи для системы (34). Рассмотрим некоторые из них.

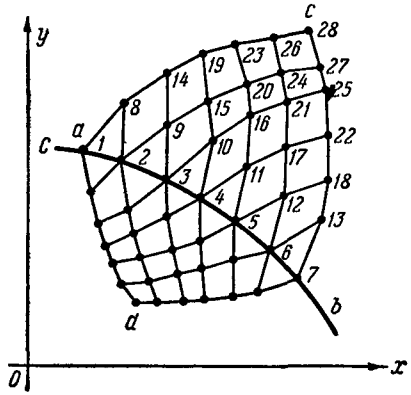

Рис. 50.

А. *Задача Коши.* Задача Коши заключается в отыскании решения системы (34), если функции u и v заданы на некоторой дуге гладкой кривой C, не имеющей характеристических направлений ни в одной точке. Численное решение этой задачи по методу Массо заключается в следующем. На дуге кривой выбираем ряд достаточно близких точек (рис. 50). На этом рисунке выбранные точки занумерованы числами 1, 2, 3, ..., 7. По точкам 1 и 2, выше указанным методом, находим точку 8 (т. е. ее координаты и значения u и v в ней). Это сделать можно, так как для точек 1 и 2 все нужные величины известны из начальных условий. Затем по точкам 2 и 3 находим точку 9, ..., по точкам 6 и 7 — точку 13. Теперь ряд точек 8, 9, ..., 13 рассматриваем как исходный и продолжаем

[1]) См. сноску на стр. 476.

построение. Процесс можно продолжать до тех пор, пока не будет заполнен «треугольник» acb, в котором сторона ac есть ломаная линия, являющаяся некоторым приближением к характеристике первого семейства, выходящей из точки a, а ломаная bc есть приближение к характеристике второго семейства, выходящей из точки b. Указанное построение можно выполнить и с другой стороны кривой C. При этом получим «треугольник» adb, стороны ad и bd которого являются соответственно приближениями к характеристике 2-го семейства, проходящей через точку a, и к характеристике 1-го семейства, проходящей через точку b. При этом в явном виде находится область, в которой можно найти решения системы при начальных условиях, заданных на участке ab кривой C. Для точного решения эта область образуется четырьмя характеристиками, выходящими из концевых точек и соответствующими решению, определяемому начальными условиями.

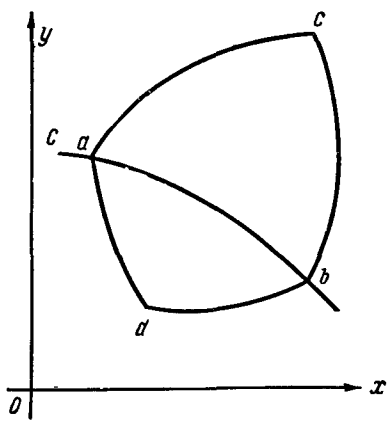

Рис. 51.

Если система нелинейна, то эти характеристики заранее неизвестны, и мы попутно получаем их приближения с помощью ломаных линий. В случае линейной системы сеть характеристик может быть заранее построена, и нужно только в точках их пересечения находить значения u и v, используя дифференциальные соотношения на характеристиках.

Б. *Задача Гурса.* В задаче Гурса требуется найти решение u, v системы (34), если на двух характеристиках ab и ac, выходящих из одной точки a, заданы значения u и v, причем значения соответствующих функций, заданных на характеристиках, совпадают в общей точке a. (Само собой разумеется, что заданные функции u и v на каждой характеристике таковы, что дифференциальные уравнения характеристик удовлетворяются.)

Численное решение этой задачи методом Массо состоит в следующем. На дугах характеристик ab и ac берется ряд близких точек (рис. 52) $1, 2, \ldots, 9$ в нашем случае. В этих точках значения u и v известны. С помощью нашего элементарного построения по точкам 4 и 5 находим точку 10, по точкам 3 и 10 — точку 11; по 2 и 11 — точку 12; по 1 и 12 — точку 13. Далее, принимая ряд точек 5, 10, 11, 12, 13 за новый ряд, продолжаем то же построение. При этом мы заполним элементарными четырехугольниками «четырехугольник», аппроксимирующий криволинейный четырехугольник, две стороны которого суть заданные дуги характеристик ab и ac, а две другие есть дуги характеристик вторых

семейств, выходящие из концов b и c. Таким образом, снова определяется область, в которой можно построить решение по заданным значениям.

В. *Первая смешанная задача.* Эта задача заключается в построении решения u и v системы (34), если на дуге ab, являющейся характеристикой, и на дуге ac, которая ни в одной точке не имеет характеристического направления, заданы значения u и v. При этом предполагается, что в общей точке a значения соответствующих функций согласованы и характеристика второго семейства, выходящая из точки a, лежит внутри угла bac (рис. 53).

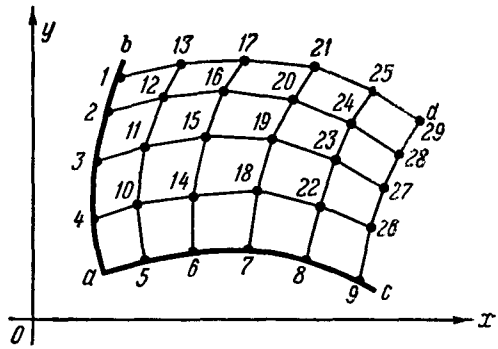

Рис. 52.

Решение первой смешанной задачи сводится к последовательному решению задачи Коши и задачи Гурса изложенным выше методом, нужно только начинать с решения задачи Коши с начальными данными на дуге ac. При этом мы сможем построить решение в «треугольнике», аппроксимирующем треугольник acd, ограниченный дугой ac и дугами двух характеристик разных семейств, выходящих из концов a и c. При этом приближенно определится вторая характеристика ad, выходящая из точки a, которая вначале была неизвестна, а также определятся значения u и v в узлах этой характеристики. Далее, решение задачи в области dab сводится к решению задачи Гурса, так как u и v будут известны на обеих характеристиках, выходящих из точки a

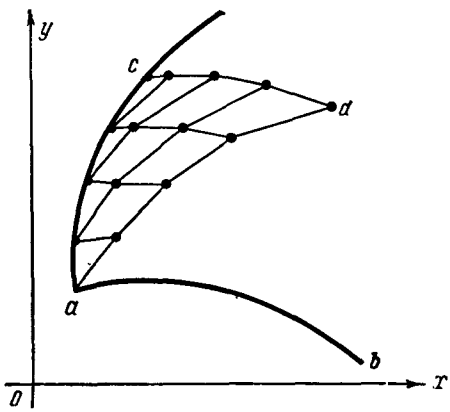

Рис. 53.

Г. *Вторая смешанная задача.* Эта задача заключается в отыскании решения системы (34), если известны значения u и v на характеристике ab и известна линейная комбинация $\alpha u + \beta v = f$ на кривой ac, не имеющей характеристических направлений, где α, β и f — заданные функции точки дуги ac. При этом предполагается, что вторая характеристика, выходящая из точки a, лежит вне угла bac, и,

кроме того, значения u и v в точке a кривой ab удовлетворяют соотношению $\alpha u+\beta v=f$ в этой точке.

Для решения этой задачи поступают следующим образом. На дуге характеристики ab берем ряд точек *1*, *2*, *3*, ... (рис. 54). Из точки *1* проводим в направлении характеристики второго семейства прямую до пересечения с кривой ac. Пусть это будет точка *5*. Из дифференциального условия на характеристике второго семейства и граничного условия находим u и v в этой точке. По найденной точке *5* и точке *2* обычным путем найдем точку *6*, по точкам *6* и *3* — точку *7* и т. д. Ряд точек *5*, *6*, *7*, ... принимаем за исходный ряд и процесс повторяем. Таким образом, мы можем заполнить сетку в области, ограниченной кривыми ab, ac и характеристикой второго семейства, проведенной из точки b до пересечения с кривой ac.

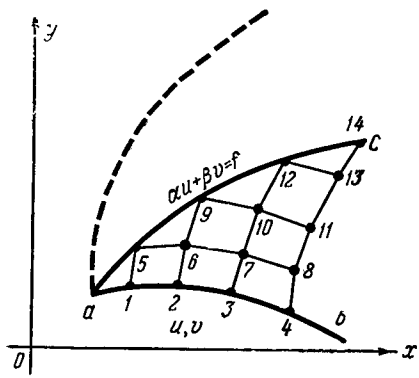

Рис. 54.

5. Численное решение гиперболической системы трех квазилинейных дифференциальных уравнений первого порядка методом Массо. Метод Массо, изложенный в предыдущем пункте, может быть применен и для численного решения гиперболической системы трех квазилинейных уравнений первого порядка:

$$\left.\begin{aligned} a_{11}\frac{\partial u}{\partial x}+b_{11}\frac{\partial u}{\partial y}+a_{12}\frac{\partial v}{\partial x}+b_{12}\frac{\partial v}{\partial y}+a_{13}\frac{\partial w}{\partial x}+b_{13}\frac{\partial w}{\partial y}=c_1,\\ a_{21}\frac{\partial u}{\partial x}+b_{21}\frac{\partial u}{\partial y}+a_{22}\frac{\partial v}{\partial x}+b_{22}\frac{\partial v}{\partial y}+a_{23}\frac{\partial w}{\partial x}+b_{23}\frac{\partial w}{\partial y}=c_2,\\ a_{31}\frac{\partial u}{\partial x}+b_{31}\frac{\partial u}{\partial y}+a_{32}\frac{\partial v}{\partial x}+b_{32}\frac{\partial v}{\partial y}+a_{33}\frac{\partial w}{\partial x}+b_{33}\frac{\partial w}{\partial y}=c_3. \end{aligned}\right\} \quad (39')$$

Элементарная задача в этом случае решается следующим образом.

Пусть 1 и 2 — две близкие точки, в которых известны все величины x_i, y_i, u_i, v_i, w_i $(i=1, 2)$. Будем обозначать через λ_{1i}, λ_{2i}, λ_{3i} корни уравнения (19) в порядке их возрастания, вычисленные для точки i. Обозначим через O середину отрезка, соединяющего точки 1 и 2. Координаты этой точки суть

$$x_0=\frac{1}{2}(x_1+x_2);\quad y_0=\frac{1}{2}(y_1+y_2).$$

Положим, что в этой точке

$$u_0=\frac{1}{2}(u_1+u_2);\quad v_0=\frac{1}{2}(v_1+v_2);\quad w_0=\frac{1}{2}(w_1+w_2).$$

Из точки 1 проведем прямую в направлении характеристики, соответствующей λ_{11}, т. е.

$$y - y_1 = \lambda_{11}(x - x_1),$$

а из точки 2 проведем прямую в направлении характеристики, соответствующей λ_{32}, т. е.

$$y - y_2 = \lambda_{32}(x - x_2).$$

Точку их пересечения обозначим номером 3. Координаты этой точки найдутся из решения системы

$$\left.\begin{aligned} y_3^{(1)} - y_1 &= \lambda_{11}\left(x_3^{(1)} - x_1\right), \\ y_3^{(2)} - y_2 &= \lambda_{32}\left(x_3^{(1)} - x_2\right). \end{aligned}\right\} \tag{40}$$

Из уравнений

$$\left.\begin{aligned} y_3^{(1)} - y_0^{(1)} &= \lambda_{20}\left(x_3^{(1)} - x_0^{(1)}\right), \\ y_1 - y_0^{(1)} &= \frac{y_2 - y_1}{x_2 - x_1}\left(x_1 - x_0^{(1)}\right) \end{aligned}\right\} \tag{41}$$

найдем новые координаты точки $O_1\left(x_0^{(1)}, y_0^{(1)}\right)$. Вводя обозначение

$$\nu^{(1)} = \frac{x_0^{(1)} - x_1}{x_2 - x_0^{(1)}} = \frac{y_1 - y_0^{(1)}}{y_0^{(1)} - y_2}, \tag{42}$$

линейной интерполяцией определяем значения u, v, w в точке O_1:

$$\left.\begin{aligned} u_0^{(1)} &= \frac{u_1 + \nu^{(1)} u_2}{1 + \nu^{(1)}}; \\ v_0^{(1)} &= \frac{v_1 + \nu^{(1)} v_2}{1 + \nu^{(1)}}; \\ w_0^{(1)} &= \frac{w_1 + \nu^{(1)} w_2}{1 + \nu^{(1)}}. \end{aligned}\right\} \tag{43}$$

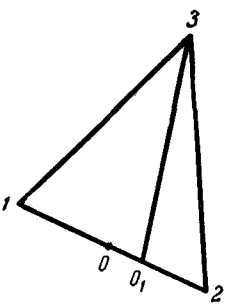

Рис. 55.

Далее, принимая отрезки *13*, *23* и $0_1 3$ за характеристики и используя дифференциальные соотношения (21) вдоль характеристик, пишем систему уравнений для определения первых приближений u, v, w в точке *3*:

$$\left.\begin{aligned} M_1^{(1)}\left(u_3^{(1)} - u_1\right) + N_1^{(1)}\left(v_3^{(1)} - v_1\right) + P_1^{(1)}\left(w_3^{(1)} - w_1\right) + C_1^{(1)}\left(x_3^{(1)} - x_1\right) &= 0, \\ M_2^{(1)}\left(u_3^{(1)} - u_2\right) + N_2^{(1)}\left(v_3^{(1)} - v_2\right) + P_2^{(1)}\left(w_3^{(1)} - w_2\right) + C_2^{(1)}\left(x_3^{(1)} - x_2\right) &= 0, \\ M_0^{(1)}\left(u_3^{(1)} - u_0^{(1)}\right) + N_0^{(1)}\left(v_3^{(1)} - v_0^{(1)}\right) + P_0^{(1)}\left(w_3^{(1)} - w_0^{(1)}\right) + C_0^{(1)}\left(x_3^{(1)} - x_0^{(1)}\right) &= 0, \end{aligned}\right\} \tag{44}$$

где $M_i^{(1)}$, $N_i^{(1)}$, $P_i^{(1)}$, $C_i^{(1)}$ — значения M, N, P, C в точке i ($i = 0_1, 1, 2$). Найдя точку *3*: $\left(x_3^{(1)}, y_3^{(1)}, u_3^{(1)}, v_3^{(1)}, w_3^{(1)}\right)$, производим ее уточнение одним из следующих способов.

Первый способ. Вычисляем значения $\lambda_{13}^{(1)}$, $\lambda_{23}^{(1)}$, $\lambda_{33}^{(1)}$, используя первое приближение точки 3, а также $M_3^{(1)}$, $N_3^{(1)}$, $P_3^{(1)}$, $C_3^{(1)}$. Находим величины

$$\left.\begin{aligned} &\lambda_{11}^{(2)}=\frac{1}{2}\left(\lambda_{11}+\lambda_{13}^{(1)}\right); \qquad \lambda_{32}^{(2)}=\frac{1}{2}\left(\lambda_{32}+\lambda_{33}^{(1)}\right); \quad \lambda_{20}^{(2)}=\frac{1}{2}\left(\lambda_{20}^{(1)}+\lambda_{23}^{(1)}\right),\\ &M_i^{(2)}=\frac{1}{2}\left(M_i^{(1)}+M_3^{(1)}\right); \quad N_i^{(2)}=\frac{1}{2}\left(N_i^{(1)}+N_3^{(1)}\right);\\ &P_i^{(2)}=\frac{1}{2}\left(P_i^{(1)}+P_3^{(1)}\right); \qquad C_i^{(2)}=\frac{1}{2}\left(C_i^{(1)}+C_3^{(1)}\right) \qquad (i=0_1,\ 1,\ 2). \end{aligned}\right\} \tag{45}$$

Находим координаты уточненных точек 0 и 3 по формулам:

$$\left.\begin{aligned} y_3^{(2)}-y_1&=\lambda_{11}^{(2)}\left(x_3^{(2)}-x_1\right),\\ y_3^{(2)}-y_2&=\lambda_{32}^{(2)}\left(x_3^{(2)}-x_2\right), \end{aligned}\right\} \tag{40'}$$

$$\left.\begin{aligned} y_3^{(2)}-y_0^{(2)}&=\lambda_{20}^{(2)}\left(x_3^{(2)}-x_0^{(2)}\right),\\ y_1-y_0^{(2)}&=\frac{y_2-y_1}{x_2-x_1}\left(x_1-x_0^{(2)}\right) \end{aligned}\right\} \tag{41'}$$

и значения $\nu^{(2)}$, $u_0^{(2)}$, $v_0^{(2)}$, $w_0^{(2)}$ — по формулам:

$$\nu^{(2)}=\frac{x_0^{(2)}-x_1}{x_2-x_0^{(2)}}, \qquad u_0^{(2)}=\frac{u_1+\nu^{(2)}u_2}{1+\nu^{(2)}}; \tag{42'}$$

$$v_0^{(2)}=\frac{v_1+\nu^{(2)}v_2}{1+\nu^{(2)}}, \qquad w_0^{(2)}=\frac{w_1+\nu^{(2)}w_2}{1+\nu^{(2)}}. \tag{43'}$$

Используя эти формулы, вычисляем $\lambda_{20}^{(1)}$, $M_0^{(1)}$, $N_0^{(1)}$, $P_0^{(1)}$, $C_0^{(1)}$, а затем $M_0^{(2)}$, $N_0^{(2)}$, $P_0^{(2)}$, $C_0^{(2)}$ и находим значения $u_3^{(2)}$, $v_3^{(2)}$, $w_3^{(2)}$, решая систему уравнений

$$\left.\begin{aligned} &M_1^{(2)}\left(u_3^{(2)}-u_1\right)+N_1^{(2)}\left(v_3^{(2)}-v_1\right)+P_1^{(2)}\left(w_3^{(2)}-w_1\right)+C_1^{(2)}\left(x_3^{(2)}-x_1\right)=0,\\ &M_2^{(2)}\left(u_3^{(2)}-u_2\right)+N_2^{(2)}\left(v_3^{(2)}-v_2\right)+P_2^{(2)}\left(w_3^{(2)}-w_2\right)+C_2^{(2)}\left(x_3^{(2)}-x_2\right)=0,\\ &M_0^{(2)}\left(u_3^{(2)}-u_0^{(2)}\right)+N_0^{(2)}\left(v_3^{(2)}-v_0^{(2)}\right)+P_0^{(2)}\left(w_3^{(2)}-w_0^{(2)}\right)+C_0^{(2)}\left(x_3^{(2)}-x_0^{(2)}\right)=0. \end{aligned}\right\} \tag{44'}$$

Таким образом, мы находим второе приближение точки 3 и продолжаем уточнение до совпадения с заданной точностью двух последовательных приближений [1]).

Второй способ. Этот способ аналогичен первому, но только берутся не средние арифметические величины по формулам (45) в точках i и 3, а значения этих величин вычисляются в средних точках отрезков 13, 23, 03. В остальном процесс уточнения остается такой же.

---

[1]) См. сноску на стр. 476.

С помощью этого элементарного построения задача Коши и задача Гурса решаются точно так же, как было описано в предыдущем пункте. При решении задачи Коши с начальными данными на дуге ab (рис. 56), не имеющей характеристических направлений, решение

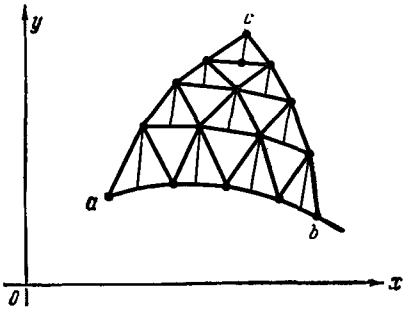

Рис. 56. Задача Коши.

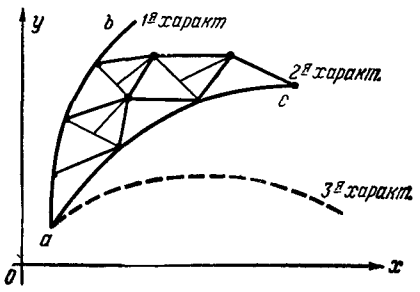

Рис. 57. Задача Гурса.

можно найти в криволинейном четырехугольнике, ограниченном крайними характеристиками, выходящими из точек a и b, а в задаче Гурса в криволинейном четырехугольнике, ограниченном двумя заданными характеристиками и двумя другими характеристиками экстремальных направлений, выходящими из концов заданных характеристик (рис. 57—59). Первая смешанная задача решается путем

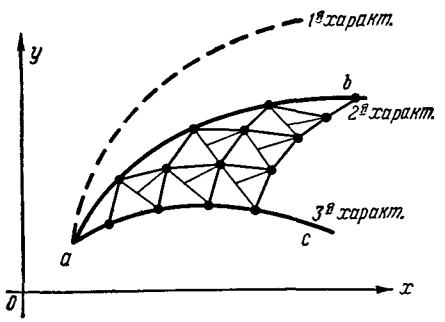

Рис. 58. Задача Гурса.

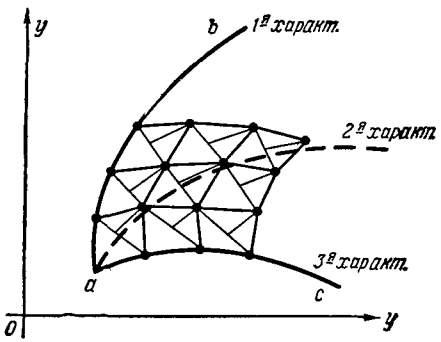

Рис. 59. Задача Гурса.

последовательного решения задачи Коши, а затем задачи Гурса; только при постановке задачи нужно требовать, чтобы кривая ab, не являющаяся характеристикой, лежала бы вне угла, образованного крайними характеристиками, выходящими из точки a, а заданная характеристика ab не должна быть ближайшей к кривой ab (рис. 60 и 61). На остальных задачах в этом случае мы не будем останавливаться ввиду многообразия их постановок.

6. Метод Массо численного решения квазилинейного гиперболического уравнения второго порядка. Рассмотрим теперь квазилинейное дифференциальное уравнение второго порядка гиперболического типа [1])

$$a\frac{\partial^2 u}{\partial x^2}+2b\frac{\partial^2 u}{\partial x\,\partial y}+c\frac{\partial^2 u}{\partial y^2}=f. \tag{46}$$

Для этого уравнения мы получили следующие уравнения характеристик:

1-е семейство	2-е семейство
$dy-\lambda_1 dx=0,$	$dy-\lambda_2 dx=0,$
$a(dp+\lambda_2 dq)-f\,dx=0.$	$a(dp+\lambda_1 dq)-f\,dx=0,$
$du=p\,dx+q\,dy.$	$du=p\,dx+q\,dy.$

Снова предположим, что в двух близких точках 1 и 2 плоскости x, y известны значения u, p, q, т. е. известны точки $(x_1, y_1, u_1, p_1, q_1)$

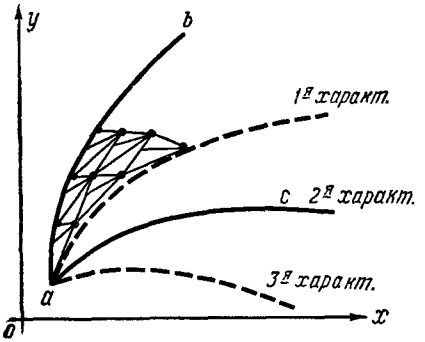

Рис. 60. Первая смешанная задача.

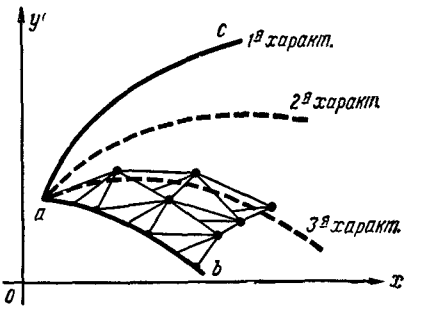

Рис. 61. Первая смешанная задача.

и $(x_2, y_2, u_2, p_2, q_2)$. Проведем через точку 1 прямую в направлении характеристики 1-го семейства, выходящей из этой точки, а через точку 2 — прямую в направлении характеристики 2-го семейства, выходящей из точки 2. Координаты $x_3^{(1)}$, $y_3^{(1)}$ точки их пересечения удовлетворяют уравнениям:

$$y_3^{(1)}-y_1=\lambda_{11}\left(x_3^{(1)}-x_1\right),$$

$$y_3^{(1)}-y_2=\lambda_{22}\left(x_3^{(1)}-x_2\right),$$

где λ_{ij} — значения λ_i, вычисленные в точке j.

---

[1]) См. Ф. И. Франкль, С. А. Христианович, Р. Н. Алексеева, Основы газовой динамики, ЦАГИ, 1938.

Далее, заменяя в дифференциальных соотношениях входящие в них дифференциалы конечными разностями, будем иметь систему уравнений для отыскания $p_3^{(1)}$, $q_3^{(1)}$, $u_3^{(1)}$:

$$\left.\begin{aligned} &a_1\left[\left(p_3^{(1)}-p_1\right)+\lambda_{21}\left(q_3^{(1)}-q_1\right)\right]-f_1\left(x_3^{(1)}-x_1\right)=0,\\ &a_2\left[\left(p_3^{(1)}-p_2\right)+\lambda_{12}\left(q_3^{(1)}-q_2\right)\right]-f_2\left(x_3^{(1)}-x_2\right)=0,\\ &u_3^{(1)}-\frac{u_1+u_2}{2}=\frac{1}{2}\left[p_1\left(x_3^{(1)}-x_1\right)+q_1\left(y_3^{(1)}-y_1\right)\right]+\\ &\qquad+\frac{1}{2}\left[p_2\left(x_3^{(1)}-x_2\right)+q_2\left(y_3^{(1)}-y_2\right)\right]. \end{aligned}\right\} \quad (47)$$

a_i, f_i — значения a и f в точке i ($i=1, 2$). (Последнее соотношение мы получили, заменив в дифференциальных соотношениях для обеих характеристик дифференциалы конечными разностями, а затем взяв их полусумму.)

Таким образом, решая последовательно системы (46) и (47), мы найдем первое приближение точки 3: $\left(x_3^{(1)}, y_3^{(1)}, p_3^{(1)}, q_3^{(1)}, u_3^{(1)}\right)$. Уточнение полученных значений может быть выполнено способами, совершенно аналогичными тем, которые были описаны в п. 3, где было рассмотрено решение системы двух квазилинейных уравнений первого порядка. Решение задач Коши, Гурса и смешанных задач также не будет по существу отличаться от решения соответствующих задач, описанных там, поэтому мы не будем на них останавливаться. Заметим лишь, что при постановке задачи Коши и первой смешанной задачи на кривой, не являющейся характеристикой, мы должны задать функцию u и производную от нее по направлению, не касательному к кривой, несущей начальные значения, так как по этим данным в точках этой кривой могут быть вычислены обе частные производные. Во второй смешанной задаче не на характеристике можно задать функцию u или линейную комбинацию ее частных производных.

Пример. Найти методом характеристик несколько значений решения системы уравнений

$$2u\frac{\partial u}{\partial x}+v\left(\frac{\partial u}{\partial y}+\frac{\partial v}{\partial x}\right)=-2e^{-2x},$$

$$\frac{\partial u}{\partial y}-\frac{\partial v}{\partial x}=0,$$

удовлетворяющего начальным условиям

$$u(0, y)=\cos y, \quad v(0, y)=\sin y \qquad (1\leqslant y\leqslant 1{,}5).$$

Дифференциальные уравнения характеристик этой системы имеют вид:

1-е семейство

$$dy = \frac{v}{u}\,dx;$$

$$v\,du + e^{-2x}\,dy = 0.$$

2-е семейство

$$dy = 0;$$

$$v\,dv + u\,du + e^{-2x}\,dx = 0.$$

Для численного решения задачи возьмем на отрезке, несущем начальные данные, шесть равноотстоящих точек. Координаты этих точек и значения u и v в них приведены в таблице:

i	1	2	3	4	5	6
x_i	0	0	0	0	0	0
y_i	1,0	1,1	1,2	1,3	1,4	1,5
u_i	0,5403	0,4536	0,3624	0,2675	0,1700	0,0707
v_i	0,8415	0,8912	0,9320	0,9636	0,9854	0,9975

Для отыскания координат точки j, лежащий на пересечении характеристик двух разных семейств, выходящих из точек i и $i+1$, и значений u и v в этой точке имеем систему уравнений

$$y_j^{(n)} - y_i = \lambda_{1,ij}^{(n)}\left(x_j^{(n)} - x_i\right); \quad y_j^{(n)} - y_{i+1} = \lambda_{2,i+1,j}^{(n)}\left(x_j^{(n)} - x_{i+1}\right);$$

$$v_{ij}^{(n)}\left(u_j^{(n)} - u_i\right) + e^{-2x_{ij}^{(n)}}\left(y_j^{(n)} - y_i\right) = 0;$$

$$v_{i+1,j}^{(n)}\left(v_j^{(n)} - v_{i+1}\right) + u_{i+1,j}^{(n)}\left(u_j^{(n)} - u_{i+1}\right) + e^{-2x_{i+1,j}^{(n)}}\left(x_j^{(n)} - x_{i+1}\right) = 0,$$

где

$$\lambda_{1,ij}^{(1)} = \frac{v_i}{u_i}; \quad \lambda_{2,i+1}^{(1)} = 0; \quad u_{ij}^{(1)} = u_i; \quad v_{ij}^{(1)} = v_i; \quad x_{i,j}^{(1)} = x_i;$$

$$\lambda_{1,ij}^{(n)} = \frac{\lambda_{1i}^{(1)} + \lambda_{1j}^{(n-1)}}{2}; \quad \lambda_{2,i+1,j}^{(n)} = \frac{\lambda_{2,i+1}^{(1)} + \lambda_{2j}^{(n-1)}}{2} = 0;$$

$$u_{ij}^{(n)} = \frac{u_i + u_j^{(n-1)}}{2}; \quad v_{ij}^{(n)} = \frac{v_i + v_j^{(n-1)}}{2}; \quad x_{ij}^{(n)} = \frac{x_i + x_j^{(n-1)}}{2};$$

$$\lambda_{1j}^{(n-1)} = \frac{v_j^{(n-1)}}{u_j^{(n-1)}}; \quad \lambda_{2j}^{(n-1)} = 0 \qquad (n = 2,\ 3,\ \ldots).$$

Отсюда получаем следующие расчетные формулы:

$$y_j^{(n)} = y_{i+1}; \quad x_j^{(n)} = \frac{1}{\lambda_{1ij}^{(n)}}\left(y_j^{(n)} - y_i\right) + x_i; \quad u_j^{(n)} = u_i - \frac{e^{-2x_{ij}^{(n)}}\left(y_j^{(n)} - y_i\right)}{v_i^{(n)}};$$

$$v_j^{(n)} = v_{i+1} - \frac{u_{i+1,j}^{(n)}\left(u_j^{(n)} - u_{i+1}\right) + e^{-2x_{i+1,j}^{(n)}}\left(x_j^{(n)} - x_{i+1}\right)}{v_{i+1,j}^{(n)}}$$

$$(n = 1,\ 2,\ \ldots).$$

Итерации проводим до тех пор, пока $x_j^{(n)}$, $y_j^{(n)}$, $u_j^{(n)}$, $v_j^{(n)}$ будут с заданной точностью равны соответственно $x_j^{(n+1)}$, $y_j^{(n+1)}$, $u_j^{(n+1)}$, $v_j^{(n+1)}$.

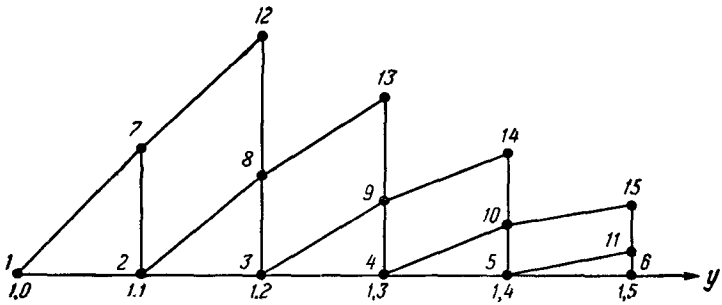

Рис. 62.

Для точки 7 (рис. 62) итерации величин $x_7^{(n)}$, $y_7^{(n)}$, $u_7^{(n)}$, $v_7^{(n)}$ ведут себя следующим образом:

n	$x_7^{(n)}$	$y_7^{(n)}$	$u_7^{(n)}$	$v_7^{(n)}$
1	0,0642	1,1	0,4215	0,8356
2	0,0565	1,1	0,4354	0,8429
3	0,0573	1,1	0,4342	0,8421
4	0,0572	1,1	0,4344	0,8423
5	0,0572	1,1	0,4344	0,8423

Ниже приведены окончательные результаты для двух слоев точек, округленные до третьего десятичного знака. В скобках указаны погрешности приближенных значений u_i и v_i, т. е. разности значений u_i и v_i и значений точного решения u, v в точках x_i, y_i, в единицах третьего десятичного разряда.

i	7	8	9	10	11
x_i	0,057	0,044	0,033	0,021	0,010
y_i	1,1	1,2	1,3	1,4	1,5
u_i	0 434 (6)	0,351 (4)	0,262 (3)	0,168 (2)	0,071 (1)
v_i	0,842 (0)	0,892 (0)	0,933 (0)	0,965 (0)	0,988 (0)

Продолжение

i	12	13	14	15
x_i	0,102	0,077	0,054	0,032
y_i	1,2	1,3	1,4	1,5
u_i	0,333 (5)	0,252 (4)	0,164 (3)	0,070 (2)
v_i	0,842 (0)	0,892 (0)	0,933 (0)	0,966 (0)

На рис. 62 изображено примерное расположение точек (x_i, y_i) (масштаб по оси x в два раза больше, чем оси y).

7. Основные задачи, встречающиеся при исследовании плоского безвихревого сверхзвукового установившегося течения идеального газа. Любые случаи плоского установившегося движения идеального газа при сверхзвуковых скоростях при отсутствии сильных разрывов (разрывов U и V) можно получить, если известны методы решения следующих задач.

Задача 1. Поле скоростей (т. е. U и V) задано в плоскости x, y на дуге ab некоторой линии C, не являющейся характеристикой. Требуется найти поле скоростей в области, ограниченной дугой ab и двумя характеристиками разных семейств, выходящими из точек a и b.

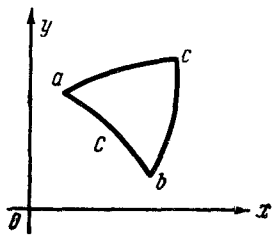

Рис. 63.

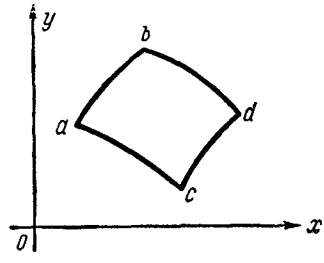

Рис. 64.

Задача 2. Поле скоростей известно на дугах ab и ac двух характеристик разных семейств, выходящих из точки a. Требуется найти поле скоростей в области, ограниченной этими дугами и дугами характеристик, выходящими из точек b и c.

Задача 3. Поле скоростей задано на дуге ab характеристики того или другого семейства, выходящей из точки a, лежащей на твердой стенке ac, заданной уравнением. Предполагается, что

граница стенки лежит между характеристиками, выходящими из точки a. Требуется определить поле скоростей в области, ограниченной дугой ab, стенкой ac и характеристикой второго семейства, выходящей из точки b.

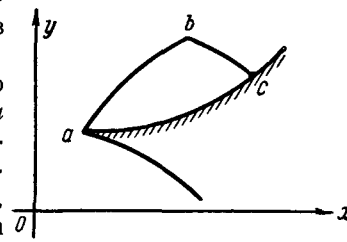

Рис. 65.

Задача 4. Поле скоростей задано на дуге ab характеристики, где a лежит на свободной границе ac, уравнение которой неизвестно (под свободной границей мы понимаем кривую, вдоль которой абсолютная величина скорости постоянна, а направление ее совпадает с касательным направлением к этой кривой в данной точке). Требуется найти уравнение свободной границы и поле скоростей в области, ограниченной дугой ab, свободной границей ac и характеристикой второго семейства, выходящей из точки b. Очевидно, что свободная граница расположена между характеристиками, выходящими из точки a.

Для численного решения этих задач можно применить метод Массо. Подробно на этом методе останавливаться нет необходимости, так как задача 1 есть задача Коши, задача 2 — задача Гурса, задача 3 —задача, которую мы раньше назвали второй смешанной задачей. В задаче 3 нужно только иметь в виду, что на твердой стенке условием на U и V будет требование, что направление скорости совпадает с направлением, касательным к стенке. Решение этих задач

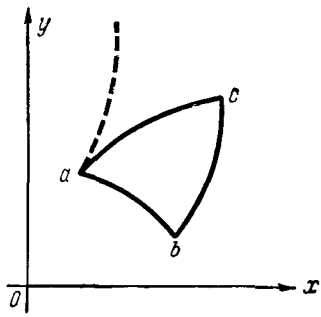

Рис. 66.

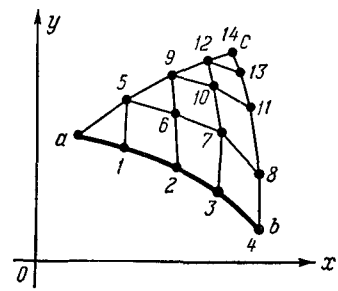

Рис. 67.

было подробно рассмотрено раньше. Остановимся на решении задачи 4, так как раньше мы не рассматривали аналогичную задачу.

Для численного решения задачи 4 возьмем на дуге достаточно густую сетку точек (точки *1*, *2*, *3*, *4* на рис. 67). Так как в точке a известны U и V, то в этой точке можно вычислить абсолютную величину скорости и найти ее направление, т. е. направление свободной границы в этой точке. В направлении ее проводим луч до

пересечения с лучом, выходящим из точки *1* в направлении характеристики второго семейства, проходящей через эту точку (точка *5* на рис. 67). Значения U и V в точке *5* можно найти, используя дифференциальное соотношение вдоль характеристики второго семейства, выходящей из точки 1, если в нем дифференциалы заменим конечными разностями и постоянство абсолютной величины скорости вдоль свободной границы. Таким образом, мы найдем U и V в этой точке, а следовательно и направление свободной границы в этой точке. По точкам *5* и *2* обычным приемом найдем точку *6*, по *6* и *3* — точку *7*, по *7* и *4* — точку *8*. Таким образом, мы найдем новый ряд точек, расположенных на новой характеристике первого семейства. Далее, повторяем изложенный процесс, считая этот ряд исходным. Уточнение можно выполнить каждый раз с помощью приемов, описанных ранее. Таким образом, после конечного числа шагов мы найдем приближенно свободную границу и поле скоростей в рассматриваемой в задаче области.

Умея находить поле скоростей в каждой из четырех задач, можно решать более сложные задачи, комбинируя эти четыре, а также находить другие величины, характеризующие движение газовой среды (например, давление p и плотность ρ), используя соотношения, связывающие их со скоростями.

§ 5. Метод сеток решения линейных дифференциальных уравнений параболического типа

Рассмотрим решение задачи Коши и смешанных задач для линейного дифференциального уравнения параболического типа[1]) вида

$$Lu = \frac{\partial u}{\partial t} - a\frac{\partial^2 u}{\partial x^2} - b\frac{\partial u}{\partial x} - cu = f, \tag{1}$$

где a, b, c, d — заданные функции переменных x и t и $a > 0$.

1. Метод сеток для решения задачи Коши. Пусть необходимо найти решение $u(x, t)$ уравнения (1) в полуплоскости $t > 0$, удовлетворяющее начальному условию

$$u(x, 0) = \varphi(x) \qquad (-\infty < x < \infty), \tag{2}$$

где $\varphi(x)$ — заданная функция.

Для отыскания приближенного решения этой задачи методом сеток рассмотрим прямоугольную сетку узлов, образуемую точками

---

[1]) По поводу применения метода сеток в теории уравнений параболического типа см.: И. Г. Петровский, Лекции об уравнениях с частными производными, ГТТИ, 1953; О. А. Олейник, Разрывные решения нелинейных дифференциальных уравнений, УМН, т. XII, вып. 3 (75) 1957 (§ 7 и библиография в конце статьи); О. А. Ладыженская, обзорная статья, цит. на стр. 412 (и библиография в конце статьи).

пересечения двух семейств параллельных прямых:

$$x = ih\,(i = 0,\ \pm 1,\ \pm 2,\ \ldots);\quad t = jl \qquad (j = 0,\ 1,\ 2,\ \ldots)$$

Для каждого узла (i, j) $(j \geqslant 1)$ запишем разностное уравнение, аппроксимирующее с некоторой точностью уравнение (1). Для этого заменим производные $\frac{\partial u}{\partial x}$, $\frac{\partial^2 u}{\partial x^2}$ в узле (i, j) соответственно разностными отношениями

$$\frac{u_{i+1,\,j} - u_{i-1,\,j}}{2h};$$

$$\frac{u_{i+1,\,j} - 2u_{ij} + u_{i-1,\,j}}{h^2}.$$

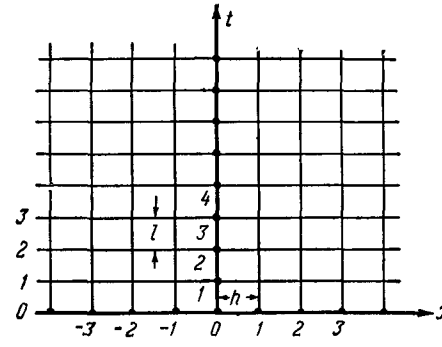

Рис. 68.

Производную $\frac{\partial u}{\partial t}$ в узле (i, j) будем заменять одним из трех разностных отношений:

$$\frac{u_{i,\,j+1} - u_{ij}}{l};\qquad \frac{u_{i,\,j} - u_{i,\,j-1}}{l};$$

$$\frac{u_{i,\,j+1} - u_{i,\,j-1}}{2l}.$$

В соответствии с этими способами аппроксимации производных мы получим три типа разностной аппроксимации дифференциального уравнения (1):

$$l^{(1)}u_{ij} = \frac{u_{i,\,j+1} - u_{i,\,j}}{l} - a_{ij}\frac{u_{i+1,\,j} - 2u_{i,\,j} + u_{i-1,\,j}}{h^2} -$$

$$- b_{ij}\frac{u_{i+1,\,j} - u_{i-1,\,j}}{2h} - c_{ij}u_{ij} = f_{ij}, \qquad (3)$$

$$l^{(2)}u_{ij} = \frac{u_{i,\,j} - u_{i,\,j-1}}{l} - a_{ij}\frac{u_{i+1,\,j} - 2u_{ij} + u_{i-1,\,j}}{h^2} -$$

$$- b_{ij}\frac{u_{i+1,\,j} - u_{i-1,\,j}}{2h} - c_{ij}u_{ij} = f_{ij}, \qquad (4)$$

$$l^{(3)}u_{ij} = \frac{u_{i,\,j+1} - u_{i,\,j-1}}{2l} - a_{ij}\frac{u_{i+1,\,j} - 2u_{ij} + u_{i-1,\,j}}{h^2} -$$

$$- b_{ij}\frac{u_{i+1,\,j} - u_{i-1,\,j}}{2h} - c_{ij}u_{ij} = f_{ij}. \qquad (5)$$

Разностное уравнение (3) содержит значения решения в четырех узлах, изображенных на рис. 69, и аппроксимирует уравнение (1) с точностью до $O(l + h^2)$; разностное уравнение (4) содержит

значения решения в четырех узлах, изображенных на рис. 70, и аппроксимирует уравнение (1) также с точностью до $O(l+h^2)$; в разностное уравнение (5) входят значения решения в пяти узлах (рис. 71), и аппроксимация уравнения (1) в этом случае будет

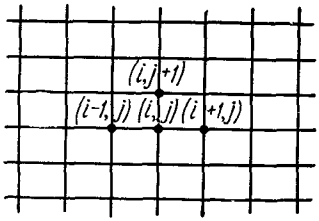

Рис. 69.

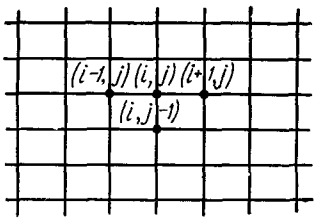

Рис. 70.

$O(l^2+h^2)$. Для узлов нулевого горизонтального ряда $j=0$ из начального условия (2) имеем:

$$u_{i0}=\varphi(ih)=\varphi_i \qquad (i=0,\ \pm 1,\ \pm 2,\ \ldots). \tag{6}$$

Первая и третья разностные схемы являются явными схемами, а вторая — неявная.

Особенно простой вид разностные уравнения (3) — (5) приобретают для уравнения

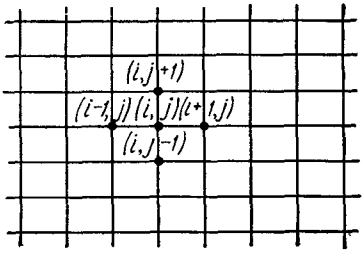

Рис. 71.

$$Lu=\frac{\partial u}{\partial t}-\frac{\partial^2 u}{\partial x^2}=0. \tag{1'}$$

Если ввести обозначение $\alpha=\frac{l}{h^2}$, то будем иметь:

$$u_{i,j+1}=(1-2\alpha)\,u_{ij}+\alpha\,(u_{i+1,j}+u_{i-1,j}), \tag{3'}$$

$$(1+2\alpha)\,u_{ij}-\alpha\,(u_{i+1,j}+u_{i-1,j})=u_{i,j-1}, \tag{4'}$$

$$u_{i,j+1}=2\alpha\,(u_{i+1,j}-2u_{ij}+u_{i-1,j})+u_{i,j-1}. \tag{5'}$$

Естественно возникает вопрос: какую из трех схем целесообразнее использовать и какое соотношение между h и l брать?

С точки зрения простоты расчетных формул целесообразно выбирать α так, чтобы разностное уравнение было наиболее просто. Из этих соображений в уравнениях (3′) — (5′) целесообразно положить $\alpha=\frac{1}{2}$. Тогда будем иметь очень простые разностные уравнения:

$$u_{i,j+1}=\frac{u_{i+1,j}+u_{i-1,j}}{2}, \tag{3''}$$

$$4u_{ij}-(u_{i+1,j}+u_{i-1,j})=2u_{i,j-1}, \tag{4''}$$

$$u_{i,j+1}=u_{i+1,j}-2u_{ij}+u_{i-1,j}+u_{i,j-1}. \tag{5''}$$

Для вычислений проще всего первая схема, так как из начальных условий известны значения решения в узлах начального ряда, по ним легко находятся значения решения в узлах первого ряда, затем второго ряда и т. д. При использовании второй схемы приходится решать систему уравнений. При решении задачи с помощью третьей разностной схемы нужно каким-то образом вычислить значения решения в узлах первого ряда, после чего счет идет так же легко, как и по первой схеме. С точки зрения точности аппроксимации дифференциального уравнения (1) третья разностная схема лучше других. Но эту схему при практических расчетах использовать нельзя по другой причине. При счете по этой схеме вычислительная погрешность, избежать которой невозможно ввиду округлений результатов, возникнув на каком-либо шаге, быстро растет в дальнейшем и через небольшое количество шагов полностью исказит решение. Это просто проследить с помощью так называемой *ε-схемы*.

Пусть счет ведется по схеме (5″) и до k-го горизонтального ряда вычисления велись точно, а при вычислении u_{ik} была допущена погрешность величины ε. Предполагая, что все дальнейшие вычисления ведутся снова точно, эта погрешность будет распространяться при дальнейших вычислениях следующим образом:

j \ l	$l-6$	$l-5$	$l-4$	$l-3$	$l-2$	$l-1$	l
$k-1$	0	0	0	0	0	0	0
k	0	0	0	0	0	0	ε
$k+1$	0	0	0	0	0	ε	-2ε
$k+2$	0	0	0	0	ε	-4ε	7ε
$k+3$	0	0	0	ε	-6ε	17ε	-24ε
$k+4$	0	0	ε	-8ε	31ε	-68ε	89ε
$k+5$	0	ε	-10ε	49ε	-144ε	273ε	-338ε
$k+6$	ε	-12ε	71ε	-260ε	641ε	-1096ε	1311ε

j \ l	$l+1$	$l+2$	$l+3$	$l+4$	$l+5$	$l+6$
$k-1$	0	0	0	0	0	0
k	0	0	0	0	0	0
$k+1$	ε	0	0	0	0	0
$k+2$	-4ε	ε	0	0	0	0
$k+3$	17ε	-6ε	ε	0	0	0
$k+4$	-68ε	31ε	-8ε	ε	0	0
$k+5$	273ε	-144ε	49ε	-10ε	ε	0
$k+6$	-1096ε	641ε	-260ε	71ε	-12ε	ε

Из этой таблицы видно, что малая погрешность, допущенная при вычислении u_{ik}, быстро растет при переходе к следующим слоям. Это, конечно, очень упрощенная схема, так как на самом деле при

счете погрешности возникают на каждом шаге и будут каким-то образом взаимодействовать. Во всяком случае, этот пример показывает, что пользоваться третьей разностной схемой по меньшей мере опасно.

Если мы будем использовать для решения уравнения (1′) разностное уравнение (3″), то распространение ε-погрешности будет иметь вид:

j \ i	$i-5$	$i-4$	$i-3$	$i-2$	$i-1$	i	$i+1$	$i+2$	$i+3$	$i+4$	$i+5$
$k-1$	0	0	0	0	0	0	0	0	0	0	0
k	0	0	0	0	0	ε	0	0	0	0	0
$k+1$	0	0	0	0	$0{,}5\varepsilon$	0	$0{,}5\varepsilon$	0	0	0	0
$k+2$	0	0	0	$0{,}25\varepsilon$	0	$0{,}5\varepsilon$	0	$0{,}25\varepsilon$	0	0	0
$k+3$	0	0	$0{,}125\varepsilon$	0	$0{,}375\varepsilon$	0	$0{,}375\varepsilon$	0	$0{,}125\varepsilon$	0	0
$k+4$	0	$0{,}0625\varepsilon$	0	$0{,}25\varepsilon$	0	$0{,}375\varepsilon$	0	$0{,}25\varepsilon$	0	$0{,}0625\varepsilon$	0

Из таблицы видно, что в этом случае погрешность не только не возрастает, а даже уменьшается.

Мы пришли к понятию *устойчивости* разностной схемы. Разностную схему называют *устойчивой*, если вычислительная погрешность при переходе от одного слоя к другому не возрастает; если же вычислительная погрешность быстро растет, то схема называется *неустойчивой*. Рассуждения, приведенные выше, хотя они и не являются достаточно строгими, показывают, что разностная схема (3″) для решения задачи Коши для уравнения теплопроводности устойчива, а схема (5″) неустойчива. К этому вопросу мы еще вернемся в этом параграфе и подробно его рассмотрим в § 7.

В заключение этого пункта приведем доказательство сходимости последовательности решений задачи Коши для уравнения (1′) с начальными условиями (2), получаемых методом сеток с использованием разностной схемы (3″) при стремлении h к нулю.

Пусть (x_0, t_0) — некоторая фиксированная точка верхней полуплоскости. Построим сетку, удовлетворяющую условиям

$$l = \frac{1}{2}h^2; \quad t_0 = 2nl = nh^2,$$

где n — целое число, причем так, чтобы (x_0, t_0) была узлом сетки. Для удобства применим следующую нумерацию узлов. Через (i, j) будем обозначать узел, находящийся на пересечении прямых $x = x_0 + ih$; $t = jl$. При этой нумерации точка (x_0, t_0) будет узлом $(0, 2n)$. Решение задачи, получаемое методом сеток при выбранном шаге h (а следовательно, и l), будем обозначать через $u_{ij}^{(h)}$.

Таким образом, $u_{ij}^{(h)}$ есть решение системы

$$u_{i,\,j+1}^{(h)} = \frac{u_{i+1,\,j}^{(h)} + u_{i-1,\,j}^{(h)}}{2} \qquad (i = 0,\ \pm 1,\ \pm 2,\ \ldots;\ j = 0,\ 1,\ 2,\ \ldots)$$

$$u_{i0}^{(h)} = \varphi_i \qquad (i = 0,\ \pm 1,\ \pm 2,\ \ldots).$$

Легко видеть, что

$$u_{0\,2}^{(h)} = \frac{1}{2^2}\left[\varphi_{-2} + 2\varphi_0 + \varphi_2\right],$$

$$u_{0\,4}^{(h)} = \frac{1}{2^4}\left[\varphi_{-4} + 4\varphi_{-2} + 6\varphi_0 + 4\varphi_2 + \varphi_4\right],$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$u_{0,\,2n}^{(h)} = \frac{1}{2^{2n}} \sum_{i=-n}^{n} C_{2n}^{n+i} \varphi_{2i}.$$

Так как $t_0 = nh^2$, то $u_{0,\,2n}^{(h)}$ можно еще записать и таким образом:

$$u_{0,\,2n}^{(h)} = \sum_{i=-n}^{n} \frac{1}{2^{2n}} \frac{1}{2h} C_{2n}^{n+i} \varphi_{2i} 2h = \frac{1}{2\sqrt{t_0}} \sum_{i=-n}^{n} \frac{\sqrt{n}}{2^{2n}} \frac{(2n)!}{(n+i)!\,(n-i)!} \varphi_{2i} 2h,$$

где $\varphi_{2i} = \varphi(x_0 + 2ih)$. Рассмотрим выражение

$$g_{ni} = \frac{\sqrt{n}}{2^{2n}} \frac{(2n)!}{(n+i)!\,(n-i)!}.$$

Все входящие в него факториалы заменим по формуле Стирлинга

$$m! = \sqrt{2\pi m} \left(\frac{m}{e}\right)^m e^{\frac{\theta_m}{12m}} \qquad (0 < \theta_m < 1),$$

где e — основание натуральных логарифмов. Тогда $g_{n,\,i}$ будет выглядеть так:

$$g_{ni} = \frac{\sqrt{n}}{2^{2n}} \frac{\sqrt{4\pi n}\left(\frac{2n}{e}\right)^{2n} e^{\frac{\theta_{2n}}{24n}}}{\sqrt{2\pi\,(n+i)}\left(\frac{n+i}{e}\right)^{n+i} e^{\frac{\theta_{n+i}}{12(n+i)}}\ \sqrt{2\pi\,(n-i)}\left(\frac{n-i}{e}\right)^{n-i} e^{\frac{\theta_{n-i}}{12(n-i)}}} =$$

$$= \frac{1}{\sqrt{\pi}} \frac{n}{\sqrt{n^2 - i^2}} \frac{e^{\frac{\theta_{2n}}{24n} - \frac{\theta_{n+i}}{12(n+i)} - \frac{\theta_{n-i}}{12(n-i)}}}{\left(1 + \frac{i}{n}\right)^{n+i}\left(1 - \frac{i}{n}\right)^{n-i}}.$$

Обозначим $x_0 + 2ih$ через η. Будем уменьшать h так, чтобы прямая $x = \eta = \text{const}$ была все время узловой линией. Тогда

$$i = \frac{\eta - x_0}{2h}; \quad n = \frac{t_0}{h^2}; \quad \frac{i}{n} = \frac{\eta - x_0}{2t_0} h.$$

Отсюда g_{ni} есть функция η, h. Обозначим ее через $g_h(\eta)$. Рассмотрим предел $g_h(\eta)$ при $h \to 0$ указанным выше способом. Очевидно, при $h \to 0$ n, i будут стремиться к бесконечности, а $\frac{i}{n}$ к нулю. Будем иметь:

$$\lim_{h\to 0} g_h(\eta) = \frac{1}{\sqrt{\pi}} \lim_{h\to 0} \frac{n}{\sqrt{n^2 - i^2}} \frac{e^{\frac{\theta_{2n}}{24n} - \frac{\theta_{n+i}}{12(n+i)} - \frac{\theta_{n-i}}{12(n-i)}}}{\left(1+\frac{i}{n}\right)^{n+i}\left(1-\frac{i}{n}\right)^{n-i}} =$$

$$= \frac{1}{\sqrt{\pi}} \lim_{h\to 0} \frac{1}{\sqrt{1-\left(\frac{i}{n}\right)^2}} \lim_{n\to 0} \frac{\left(1-\frac{i}{n}\right)^i}{\left(1-\frac{i^2}{n^2}\right)^n \left(1+\frac{i}{n}\right)^i} =$$

$$= \frac{1}{\sqrt{\pi}} \lim_{n\to 0} \frac{\left[1-\frac{\eta_i - x_0}{2t_0}h\right]^{\frac{\eta_i - x_0}{2h}}}{\left[1-\left(\frac{\eta_i - x_0}{2t_0}\right)^2 h^2\right]^{\frac{1}{h^2}} \left[1+\frac{\eta_i - x_0}{2t_0}h\right]^{\frac{\eta_i - x_0}{2h}}} = \frac{1}{\sqrt{\pi}} e^{-\frac{(\eta_i - x_0)^2}{4t_0}}.$$

Далее,

$$u^{(h)}_{0,2n} = \frac{1}{2\sqrt{t_0}} \sum_{i=-n}^{n} g_{ni}\varphi_{2i} 2h \approx \frac{1}{2\sqrt{t_0}} \int_{x_0-2nh}^{x_0+2nh} g_h(\eta)\varphi(\eta)\,d\eta =$$

$$= \frac{1}{2\sqrt{t_0}} \int_{x_0-\frac{2t_0}{h}}^{x_0+\frac{2t_0}{h}} g_h(\eta)\varphi(\eta)\,d\eta \to \frac{1}{2\sqrt{t_0}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{\pi}} e^{-\frac{(\eta-x_0)^2}{4t_0}} \varphi(\eta)\,d\eta.$$

Таким образом,

$$\lim_{h\to 0} u^{(h)}_{0,2n} = \frac{1}{2\sqrt{\pi t_0}} \int_{-\infty}^{+\infty} \varphi(\eta) e^{-\frac{(\eta-x_0)^2}{4t_0}} d\eta.$$

Но известно, что

$$u(x_0, t_0) = \frac{1}{2\sqrt{\pi t_0}} \int_{-\infty}^{+\infty} \varphi(\eta) e^{-\frac{(\eta-x_0)^2}{4t_0}} d\eta$$

есть точное решение уравнения (1) с начальными условиями (2). Следовательно,

$$\lim_{h\to 0} u^{(h)}_{0,2i} = u(x_0, t_0).$$

Хотя здесь и не все рассуждения проведены с достаточной строгостью (не обоснованы предельные переходы), но мы не только показали сходимость последовательности $u^{(h)}_{ij}$ к точному решению, но еще получили интегральную форму точного решения.

2. Метод сеток для решения смешанных задач. Понятие устойчивости разностных схем. Рассмотрим смешанные задачи для уравнения теплопроводности

$$Lu = \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0. \qquad (1')$$

Эти задачи ставятся следующим образом. Требуется найти решение уравнения (1′) в прямоугольнике $R = \{a \leqslant x \leqslant b;\ 0 \leqslant t \leqslant T\}$, удовлетворяющее начальному условию

$$u(x,\ 0) = \varphi(x) \qquad (a \leqslant x \leqslant b) \qquad (2')$$

и граничным условиям

$$\left[\beta_1 \frac{\partial u}{\partial x} + \gamma_1 u\right]_{x=a} = \psi_1(t);\ \left[\beta_2 \frac{\partial u}{\partial x} + \gamma_2 u\right]_{x=b} = \psi_2(t) \quad (0 \leqslant t \leqslant T), \qquad (2'')$$

где β_1, β_2, γ_1, γ_2, ψ_1, ψ_2 — заданные функции переменного t. Выбор функций β_1, β_2, γ_1, γ_2 позволяет получать различные задачи. Например, если $\beta_1 \equiv \beta_2 \equiv 0$; $\gamma_1 \equiv \gamma_2 \equiv 1$, то будем иметь первую краевую задачу; при $\beta_1 \equiv \beta_2 \equiv 1$; $\gamma_1 \equiv \gamma_2 \equiv 0$ — вторую краевую задачу.

При решении смешанных задач методом сеток, кроме аппроксимации дифференциального уравнения и начальных условий, необходимо аппроксимировать также и граничные условия. Простейшие разностные схемы для решения смешанных задач следующие. Рассматриваем сетку точек $(a + ih,\ jl)$, где $i = 0,\ 1,\ 2,\ \ldots,\ n$; $j = 0,\ 1,\ 2,\ \ldots,\ m$; $h = \frac{b-a}{n}$. Будем считать $a = 0$; $b = 1$ и $h = \frac{1}{n}$. Узлы, лежащие на прямых $x = 0$; $x = 1$; $t = 0$, будем считать граничными узлами, все другие — внутренними. Для внутренних узлов выписываем разностные уравнения того или другого типа, аппроксимирующие дифференциальное уравнение (1′), например уравнения (3′) или (4′), или (5′). Для узлов, лежащих на начальной прямой $t = 0$, из начальных условий имеем:

$$u_{i0} = \varphi_i \qquad (i = 0,\ 1,\ 2,\ \ldots,\ n).$$

Для граничных узлов, лежащих на прямых $x = 0$, $x = 1$, запишем соотношения

$$\beta_{1j} \frac{u_{1,j} - u_{0j}}{h} + \gamma_{1j} u_{0j} = \psi_{1j};\quad \beta_{2j} \frac{u_{nj} - u_{n-1,j}}{h} + \gamma_{2j} u_{nj} = \psi_{2j}. \qquad (6')$$

аппроксимирующие с точностью до h граничные условия (2″). Таким образом, мы можем получить три следующие разностные схемы:

$$\left.\begin{gathered} u_{i\,j+1} = (1-2\alpha)\,u_{ij} + \alpha\,(u_{i+1,\,j} + u_{i-1,\,j}) \\ (i = 1,\ 2,\ \ldots,\ n-1;\quad j = 0,\ 1,\ 2,\ \ldots,\ m-1), \\ \beta_{1j}u_{1j} + (h\gamma_{1j} - \beta_{1j})\,u_{0j} = h\psi_{1j};\ (\beta_{2j} + h\gamma_{2j})\,u_{nj} - \beta_{2j}u_{n-1,\,j} = h\psi_{2j} \\ (j = 1,\ 2,\ \ldots,\ m), \\ u_{i0} = \varphi_i \qquad (i = 0,\ 1,\ 2,\ \ldots,\ n); \end{gathered}\right\} \tag{7}$$

$$\left.\begin{gathered} (1+2\alpha)\,u_{ij} - \alpha\,(u_{i+1,\,j} + u_{i-1,\,j}) = u_{i,\,j-1} \\ (i = 1,\ 2,\ \ldots,\ (n-1);\quad j = 1,\ 2,\ \ldots,\ m), \\ \beta_{1j}u_{1j} + (h\gamma_{1j} - \beta_{1j})\,u_{0j} = h\psi_{1j}\quad (\beta_{2j} + h\gamma_{2j})\,u_{nj} - \beta_{2j}u_{n-1,\,j} = h\psi_{2j} \\ (j = 1,\ 2,\ \ldots,\ m), \\ u_{i0} = \varphi_i \qquad (i = 0,\ 1,\ 2,\ \ldots,\ n); \end{gathered}\right\} \tag{8}$$

$$\left.\begin{gathered} u_{i,\,j+1} = 2\alpha\,(u_{i+1,\,j} - 2u_{ij} + u_{i-1\,j}) + u_{i,\,j-1} \\ (i = 1,\ 2,\ \ldots,\ n-1;\quad j = 0,\ 1,\ 2,\ \ldots,\ m-1), \\ \beta_{1j}u_{1j} + (h\gamma_{1j} - \beta_{1j})\,u_{0j} = h\psi_{1j};\ (\beta_{2j} + h\gamma_{2j})\,u_{nj} - \beta_{2j}u_{n-1,\,j} = h\psi_{2j} \\ (j = 1,\ 2,\ \ldots,\ m), \\ u_{i0} = \varphi_i \qquad (i = 0,\ 1,\ 2,\ \ldots,\ n). \end{gathered}\right\} \tag{9}$$

Иногда для лучшей аппроксимации граничных условий привлекают еще два вертикальных ряда узлов $(-h,\ jl)$, $((n+1)\,h,\ jl)$ или рассматривают сетку, сдвинутую на $\frac{h}{2}$ в направлении оси x. В этом случае можно получить для граничных условий аппроксимацию второго порядка относительно h. Построение этой аппроксимации ничем не отличается от аппроксимации граничных условий для смешанных задач уравнений гиперболического типа, которые мы рассматривали в предыдущем параграфе, поэтому здесь мы на этом останавливаться не будем. Так или иначе при любом способе аппроксимации мы получаем для отыскания значений решения смешанной задачи во внутренних и граничных узлах столько уравнений, сколько имеется неизвестных. Решая эту систему линейных алгебраических уравнений, мы найдем приближенные значения решения поставленной задачи во всех узлах сетки. Для явных схем разрешимость полученной системы не вызывает сомнений, для неявных схем ее нужно исследовать в каждом отдельном случае. Для схемы (8) это нетрудно сделать.

Значительно сложней решается вопрос о том, насколько близки полученные методом сеток значения решения в узлах к значениям точного решения смешанной задачи для дифференциального уравнения и можно ли вообще путем измельчения сетки получить методом

сеток приближенное решение, сколь угодно близкое к точному решению. Естественно, что интерес могут представлять только такие разностные схемы, с помощью которых можно получить приближенное решение, достаточно близкое к точному, так называемые *сходящиеся* разностные схемы. Разностная схема называется *сходящейся* при заданном способе стремления h и l к нулю, если решения системы разностных уравнений стремятся при этом к точному решению задачи для дифференциального уравнения. В этом определении предполагается, что мы умеем точно решать системы разностных уравнений, но практически мы можем найти лишь приближенное решение этой системы. Поэтому из сходящихся разностных схем практический интерес могут представлять только те разностные схемы, для которых малые погрешности, допущенные в процессе решения разностных уравнений, не могут привести к большим отклонениям от точного решения системы. Такие схемы мы назвали *устойчивыми*. Пока мы оставим в стороне вопрос об исследовании сходимости разностных схем, а остановимся на исследовании устойчивости разностных схем (7)—(9) для случая первой краевой задачи, т. е. в предположении, что $\beta_1 = \beta_2 = 0$, $\gamma_1 = \gamma_2 = 1$. В дальнейшем мы докажем некоторые общие теоремы о сходимости и устойчивости разностных схем, из которых можно будет сделать заключения о сходимости рассматриваемых нами разностных схем для первой краевой задачи для уравнения теплопроводности.

Сначала уточним понятие устойчивости разностной схемы, о котором пойдет речь. Мы будем предполагать, что значения граничных функций в граничных узлах вычислены точно. Далее, будем предполагать, что при отыскании решения разностных уравнений погрешность допущена на p-м слое, а дальше счет ведется точно. За счет погрешности на p-м слое мы получим добавок v_{ij} к точному решению разностной схемы. Без ограничения общности можно считать, что погрешность допущена на начальном слое. Тогда добавки v_{ij} будут являться решением той же самой системы уравнений, но только значения их в граничных узлах, лежащих на прямых $x = 0$, $x = 1$, равны нулю, а значения в граничных узлах начального слоя равны допущенным погрешностям. Разностную схему будем называть *устойчивой*, если для всякого $\varepsilon > 0$ найдется такое $\delta > 0$, что как только

$$\sum_{i=1}^{n-1} v_{i0}^2 \leqslant \delta$$

будет иметь место неравенство

$$\sum_{i=1}^{n-1} v_{ij}^2 \leqslant \varepsilon$$

для любого j, лишь бы $jl \leqslant T$, причем δ не зависит от h и l. Фактически это понятие непрерывной зависимости решения разностной схемы от начальных значений. Поэтому этот тип устойчивости называют еще *устойчивостью по начальным значениям.*

Перейдем теперь к исследованию на устойчивость схем (7) — (9). Мы докажем, что *схема* (7) *устойчива при* $\alpha \leqslant \frac{1}{2}$ *и неустойчива при* $\alpha > \frac{1}{2}$; *схема* (8) *устойчива при всех* α; *схема* (9) *неустойчива при всех* α. Здесь везде $\alpha = \frac{l}{h^2}$.

Для доказательства этого утверждения рассмотрим функцию

$$w_i^{(k)} = \sin \frac{k\pi l}{n} \qquad (i = 0,\ 1,\ 2,\ \ldots,\ n;\quad k = 1,\ 2,\ \ldots,\ n-1).$$

Легко проверить, что

$$\sum_{i=1}^{n-1} w_i^{(k)} w_i^{(m)} = \sum_{i=1}^{n-1} \sin \frac{k\pi l}{n} \sin \frac{m\pi l}{n} = \begin{cases} 0 & (k \neq m), \\ \frac{n}{2} & (k = m). \end{cases}$$

Будем искать частные решения разностных уравнений (7) — (9) вида

$$v_{ij}^{(k)} = \lambda_k^j \sin \frac{k\pi l}{n},$$

где λ_k — некоторое число, которое нужно определить. Рассмотрим сначала разностную схему (7). Подстановка в (7) дает

$$\lambda_k^{j+1} \sin \frac{k\pi l}{n} = (1 - 2\alpha)\lambda_k^j \sin \frac{k\pi l}{n} + \alpha\lambda_k^j \left(\sin \frac{k\pi (l+1)}{n} + \sin \frac{k\pi (l-1)}{n} \right)$$

или

$$\lambda_k = 1 - 4\alpha \sin^2 \frac{k\pi}{2n}.$$

При каждом фиксированном k $v_{ij}^{(k)}$ удовлетворяет граничным условиям $v_{0j}^{(k)} = v_{nj}^{(k)} = 0$. В силу линейности и однородности разностного уравнения линейная комбинация частных решений будет также решением разностного уравнения, удовлетворяющим граничным условиям:

$$v_{ij} = \sum_{k=1}^{n-1} a_k \lambda_k^j \sin \frac{k\pi l}{n}; \qquad v_{0j} = v_{nj} = 0.$$

Постоянные a_k подберем так, чтобы были удовлетворены и начальные условия в узлах $(l, 0)$ $(l = 1,\ 2,\ \ldots,\ n-1)$, т. е.

$$v_{i0} = \sum_{k=1}^{n-1} a_k \sin \frac{k\pi l}{n}.$$

Для того чтобы найти a_k, умножим обе части равенства на $\sin\frac{k\pi l}{n}$ и просуммируем по l от 1 до $n-1$. Получим:

$$\sum_{i=1}^{n-1} v_{i0}\sin\frac{k\pi l}{n} = a_k\frac{n}{2}.$$

Далее, возводя то же самое равенство в квадрат и суммируя по l от 1 до $n-1$, получим:

$$\sum_{i=1}^{n-1} v_{i0}^2 = \frac{n}{2}\sum_{k=1}^{n-1} a_k^2.$$

В точности таким же приемом для j, отличного от нуля, получим:

$$\sum_{i=1}^{n-1} v_{ij}^2 = \sum_{k=1}^{n-1} a_k^2\lambda_k^{2j}\cdot\frac{n}{2}.$$

Отсюда очевидно, что если при всех k имеет место неравенство $|\lambda_k|\leqslant 1$, то

$$\sum_{i=1}^{n-1} v_{ij}^2 \leqslant \sum_{i=1}^{n-1} v_{i0}^2 \leqslant \delta,$$

и, полагая $\delta=\varepsilon$, докажем устойчивость разностной схемы. Так как в нашем случае $\lambda_k = 1-4\alpha\sin^2\frac{k\pi}{2n}$, то $|\lambda_k|\leqslant 1$ при $\alpha\leqslant\frac{1}{2}$. Таким образом, разностная схема (7) устойчива при $\alpha\leqslant\frac{1}{2}$.

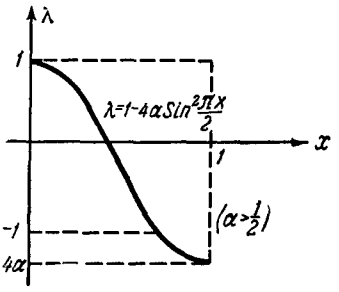

Рис. 72.

Покажем теперь неустойчивость этой схемы при $\alpha>\frac{1}{2}$. В этом случае для каждого достаточно большого n можно найти такое целое число $k_0<n$, что $|\lambda_{k_0}|>1+\mu$, где $\mu>0$ и не зависит от n. Это можно видеть из построенного на рис. 72 графика. Рассмотрим теперь следующее частное решение разностного уравнения:

$$v_{ij}^{(k_0)} = \beta\lambda_{k_0}^j\sin\frac{k_0\pi l}{n}.$$

Для этого решения

$$\sum_{i=1}^{n-1} v_{i0}^{(k_0)\,2} = \beta^2\frac{n}{2},$$

а

$$\sum_{i=1}^{n-1} v_{ij}^{(k_0)\,2} = \beta^2\lambda_{k_0}^{2j}\frac{n}{2} > \frac{n}{2}\beta^2(1+\mu)^{2j} = (1+\mu)^{2j}\sum_{i=1}^{n-1} v_{i0}^{(k_0)\,2}.$$

Следовательно, если сумма квадратов погрешностей в узлах начального ряда $\sum_{i=1}^{n-1} v_{i0}^{(k_0)2}$ не равна нулю, то при очень малом шаге h, когда для отыскания решения в прямоугольнике R нужно сделать большое число шагов в направлении оси t, сумма $\sum_{i=1}^{n-1} v_{ij}^{(k_0)2}$ для больших j будет весьма велика, а это и означает, что разностная схема неустойчива.

Рассмотрим теперь разностную схему (8). Следуя нашему методу, ищем частные решения вида

$$v_{ij}^{(k)} = \lambda_k^j \sin \frac{k\pi l}{n}.$$

Подстановка в разностное уравнение (8) дает

$$(1+2\alpha)\lambda_k^j \sin\frac{k\pi l}{n} - \alpha\lambda_k^j\left[\sin\frac{k\pi(l+1)}{n} + \sin\frac{k\pi(l-1)}{n}\right] = \lambda_k^{j-1}\sin\frac{k\pi l}{n}$$

или

$$\lambda_k\left[1 - 2\alpha\left(\cos\frac{k\pi}{n} - 1\right)\right] = 1.$$

Так как

$$2\alpha\left(\cos\frac{k\pi}{n} - 1\right) < 0 \qquad (k = 1,\ 2,\ \ldots,\ n-1),$$

то при всех α имеет место неравенство

$$0 \leqslant \lambda_k = \frac{1}{1 - 2\alpha\left(\cos\frac{k\pi}{n} - 1\right)} < 1,$$

из которого следует, что при всех α разностная схема (8) устойчива.

Для разностной схемы (9) подстановка $v_{ij}^{(k)} = \lambda_k^j \sin\frac{k\pi l}{n}$ в разностное уравнение дает

$$\lambda_k^{j+1}\sin\frac{k\pi l}{n} = 2\alpha\lambda_k^j\left(\sin\frac{k\pi(l+1)}{n} - 2\sin\frac{k\pi l}{n} + \right.$$
$$\left. + \sin\frac{k\pi(l-1)}{n}\right) + \lambda_k^{j-1}\sin\frac{k\pi l}{n}$$

или

$$\lambda_k^2 + 8\alpha\lambda_k \sin^2\frac{k\pi}{2n} - 1 = 0.$$

Обозначая $4\alpha\sin^2\frac{k\pi}{2n}$ через γ_k, получим:

$$\lambda_k^2 + 2\gamma_k\lambda_k - 1 = 0$$

или

$$\lambda_{k,1} = -\gamma_k + \sqrt{1+\gamma_k^2}; \quad \lambda_{k,2} = -\gamma_k - \sqrt{1+\gamma_k^2}.$$

Таким образом, при всех α имеет место неравенство

$$|\lambda_{k,2}| > 1.$$

Используя частное решение вида

$$\overline{v}_{ij}^{(k)} = \mu\lambda_{k,2}^{j} \sin\frac{k\pi l}{n}$$

и рассуждая точно так же, как и при доказательстве неустойчивости схемы (7) при $\alpha > \frac{1}{2}$, мы убеждаемся, что схема (9) неустойчива при всех α.

Мы рассмотрели простейшие разностные схемы для уравнения теплопроводности. Можно построить другие разностные схемы, например, используя способы построения разностных схем, описанные в § 2, причем можно получить схемы, дающие значительно лучшую аппроксимацию, чем мы имели для рассматриваемых схем, но каждый раз необходимо исследовать их на устойчивость, так как только устойчивые схемы представляют практический интерес.

Все разностные схемы разбиваются на два класса: *явные* схемы и *неявные* схемы. Явные схемы позволяют очень просто вычислить значения искомого решения в узлах m-го горизонтального ряда, если известны значения решения на предыдущих рядах. Но они имеют существенный недостаток: для того чтобы они были устойчивы, необходимо налагать сильные ограничения на сетку. Так, в схеме (7) для устойчивости должно быть выполнено ограничение $\alpha = \frac{l}{h^2} \leqslant \frac{1}{2}$, что требует очень мелкого шага по t, т. е. если нужно найти решение на конечном отрезке изменения t, то количество горизонтальных рядов узлов должно быть очень большим. Кроме того, если в ходе решения нужно уменьшить шаг по x, то нельзя этого сделать, не уменьшая шага по t.

Неявные схемы, например схема (8), свободны от этого недостатка, но использование их связано с другой трудностью: для отыскания значений решения в узлах m-го горизонтального ряда при известных значениях в узлах предыдущих рядов приходится решать систему алгебраических уравнений с большим числом неизвестных.

Если для решения этих систем применять метод итераций, то увеличение шага по времени, допустимое в этом случае, приводит к увеличению числа итераций, необходимых для отыскания решения системы с заданной точностью. Если за начальные приближения принимать соответствующие значения в узлах предыдущего ряда, что вполне естественно, то с увеличением шага по t число необходимых итераций растет, хотя и не пропорционально увеличению шага, а медленней, все же эффект выигрыша времени за счет увеличения шага по t в значительной мере пропадает. В связи с этим возникает необходимость в более эффективных методах решения систем алгебраических уравнений, получающихся при использовании

неявных разностных схем. В следующем параграфе мы изложим *метод прогонки*, разработанный в Математическом институте им. Стеклова АН СССР.

Пример. Методом сеток найти решение уравнения

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2},$$

удовлетворяющее условиям

$$u(x,\ 0) = \sin \pi x \quad (0 \leqslant x \leqslant 1); \quad u(0,\ t) = u(1,\ t) = 0.$$

Воспользуемся простейшей устойчивой явной разностной схемой

$$u_{i,\,j+1} = \frac{u_{i+1,\,j} + u_{i-1,\,j}}{2},$$

выбрав шаг по оси x $h = 0{,}1$ и шаг l по оси t: $l = \frac{h^2}{2} = 0{,}005$.

Ниже приведена таблица значений решения в узлах сетки (в единицах четвертого десятичного знака). Так как решение симметрично

n	x / t	0	0,1	0,2	0,3	0,4	0,5
0	0	0	3090	5878	8090	9511	10000
1	0,005	0	2939	5590	7694	9045	9511
2	0,010	0	2795	5316	7318	8602	9045
3	0,015	0	2658	5056	6959	8182	8602
4	0,020	0	2528	4808	6619	7780	8182
5	0,025	0	2404	4574	6294	7400	7780
6	0,030	0	2287	4349	5987	7037	7400
7	0,035	0	2174	4137	5693	6694	7037
8	0,040	0	2068	3934	5416	6365	6694
9	0,045	0	1967	3742	5150	6055	6365
10	0,050	0	1871	3558	4898	5758	6055
11	0,055	0	1779	3384	4658	5476	5758
12	0,060	0	1692	3218	4430	5208	5476
13	0,065	0	1609	3061	4213	4953	5208
14	0,070	0	1530	2911	4007	4710	4953
15	0,075	0	1456	2768	3810	4480	4710
16	0,080	0	1384	2633	3624	4260	4480
17	0,085	0	1316	2504	3446	4072	4260
18	0,090	0	1252	2381	3288	3853	4072
19	0,095	0	1190	2270	3117	3680	3853
20	0,100	0	1135	2154	2975	3485	3680
21	0,025	0	2414	4593	6321	7431	7813
22	0,050	0	1886	3588	4939	5806	6105
23	0,075	0	1474	2804	3859	4537	4770
24	0,100	0	1152	2191	3015	3545	3727
25	0,025	0	10	19	27	31	33
26	0,050	0	15	30	41	48	50
27	0,075	0	18	36	49	57	60
28	0,100	0	17	37	40	60	47

относительно прямой $x = 0{,}5$, то в таблице даны лишь значения решения для

$$x = 0;\ 0{,}1;\ 0{,}2;\ 0{,}3;\ 0{,}4;\ 0{,}5.$$

В строках 21—24 таблицы приведены значения точного решения задачи в указанных там узлах таблицы, а в строках 25—28 — погрешности приближенного решения в соответствующих узлах. Относительная погрешность не превосходит 2%. При сравнительно большом шаге $h = 0{,}1$ результат совсем неплохой. Используем теперь неустойчивую явную разностную схему

$$u_{i,\,j+1} = 2(u_{i+1,\,j} + u_{i-1,\,j}) - 3u_{ij}.$$

Для того чтобы не ухудшать аппроксимации дифференциального уравнения разностным уравнением, шаг h по оси x возьмем равным 0,05, а шаг по оси t : $l = 2h^2 = 0{,}005$, т. е. оставим его прежним. Ниже приведена таблица полученных значений решения (строки 0—7) снова в единицах четвертого десятичного разряда. В строке 7 приведены значения точного решения задачи для $t = 0{,}030$, а в строке 8 указаны погрешности приближенных значений, полученных по указанной схеме, при $t = 0{,}030$. Как видно из таблицы, после шести шагов по оси t погрешности настолько велики, что по абсолютной величине в некоторых узлах даже превосходят значения точного решения. Это вполне естественно, так как схема неустойчива, поэтому небольшие погрешности в начальных значениях решения очень быстро возросли.

В строке 9 приведены значения решения при $t = 0{,}030$, полученные по неявной схеме

$$u_{i,\,j} - u_{i,\,j-1} = 2(u_{i+1,\,j} - 2u_{ij} + u_{i-1,\,j})$$

при той же сетке ($h = 0{,}05$; $l = 0{,}005$), а в строке 10 — погрешности этих значений. Как видим, неявная схема дает хороший результат, так как после шести шагов относительные погрешности не превосходят 1%.

n	t_j \ x_i	0	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
0	0	0	1564	3090	4540	5878	7071	8090	8910	9511	9877	10000
1	0,005	0	1488	2938	4316	5588	6723	7692	8472	9041	9391	9508
2	0,010	0	1412	2794	4104	5314	6391	7314	8050	8603	8925	9040
3	0,015	0	1352	2650	3904	5048	6083	6940	7684	8141	8511	8580
4	0,020	0	1244	2562	3684	4830	5727	6714	7110	7967	7909	8304
5	0,025	0	1392	2170	3732	4332	5907	5532	8032	6137	8815	6724
6	0,030	0	164	3738	1808	6282	2007	11284	—758	15283	—723	15088
7	0,030	0	1163	2298	3376	4372	5259	6017	6626	7074	7346	7437
8		0	999	—1440	1568	—1910	3252	—5267	7384	—8209	8069	—7651
9	0,030	0	1172	2316	3402	4405	5299	6063	6678	7128	7402	7495
10		0	—9	—18	—26	—33	—40	—46	—52	—54	—56	—58

§ 6. Метод прогонки решения краевых задач для уравнений в частных производных[1])

Метод прогонки мы изложим на примере решения краевых задач для уравнения теплопроводности и для уравнения Пуассона.

1. Уравнение теплопроводности. Пусть требуется найти решение уравнения

$$\frac{\partial u}{\partial t} = \mu^2(x,\ t)\frac{\partial^2 u}{\partial x^2}, \tag{1}$$

удовлетворяющее условиям:

$$u(x,\ 0) = \varphi(x) \qquad (a \leqslant x \leqslant b), \tag{2}$$

$$\left.\begin{aligned} \left[\frac{\partial u}{\partial x} - \alpha_0(t)\,u\right]_{x=a} &= \alpha_1(t), \\ \left[\frac{\partial u}{\partial x} - \beta_0(t)\,u\right]_{x=b} &= \beta_1(t). \end{aligned}\right\} \tag{3}$$

Для решения этой задачи применим следующую разностную схему. Возьмем сетку узлов: $x_i = a + \left(i + \frac{1}{2}\right)h$; $t_j = jl$ $\left(i = -1,\ 0,\ 1,\ \ldots,\ N;\ j = 0,\ 1,\ 2,\ \ldots;\ h = \frac{b-a}{N}\right)$ и для внутреннего узла $(i,\ j)$ запишем разностное уравнение

$$\frac{u_{i,\,j} - u_{i,\,j-1}}{l} = \mu_{ij}^2\,\frac{u_{i+1,\,j} - 2u_{i\,j} + u_{i-1,\,j}}{h^2} \qquad (\mu_{ij} = \mu(x_i,\ t_j)) \tag{4}$$

или

$$u_{ij} = u_{i,\,j-1} + \frac{l\mu_{ij}^2}{h^2}(u_{i+1,\,j} - 2u_{ij} + u_{i-1,\,j})$$

$$(i = 0,\ 1,\ 2,\ \ldots,\ N-1;\ j = 1,\ 2,\ 3,\ \ldots), \tag{5}$$

аппроксимирующее уравнение (1) в узле $(i,\ j)$ с точностью до $0(l + h^2)$. В граничных узлах запишем следующие соотношения:

$$u_{i,0} = \varphi_i = \varphi(x_i) \qquad (i = -1,\ 0,\ 1,\ 2,\ \ldots,\ N), \tag{6}$$

$$\frac{u_{0j} - u_{-1,\,j}}{h} - \alpha_{0j}\frac{u_{-1,\,j} + u_{0j}}{2} = \alpha_{1j}$$

$$(j = 1,\ 2,\ \ldots;\ \alpha_{0j} = \alpha_0(t_j);\ \alpha_{1j} = \alpha_1(t_j)), \tag{7_1}$$

$$\frac{u_{Nj} - u_{N-1,\,j}}{h} - \beta_{0j}\frac{u_{N-1\,j} + u_{Nj}}{2} = \beta_{1j}$$

$$(j = 1,\ 2,\ \ldots;\ \beta_{0j} = \beta_0(t_j);\ \beta_{1j} = \beta_1(t_j)). \tag{7_2}$$

[1]) При написании данного параграфа использована рукопись неопубликованной статьи И. М. Гельфанда и Локуциевского, любезно предоставленная нам авторами.

Из методических соображений рассмотрим сначала предельный случай $h=0$ (так называемый *метод прямых*, о котором подробнее см. § 8). Полагая $u(x, t_j)=v_j(x)$; $\mu(x, t_j)=\mu_j(x)$, в этом случае будем иметь:

$$l\mu_j^2(x)\frac{d^2v_j(x)}{dx^2}=v_j(x)-v_{j-1}(x) \quad (j=1, 2, \ldots), \tag{8}$$

$$v_0(x)=\mu_0(x), \tag{9}$$

$$\left[\frac{dv_j}{dx}-\alpha_{0j}v_j\right]_{x=a}=\alpha_{1j}, \tag{10_1}$$

$$(j=1, 2, \ldots).$$

$$\left[\frac{dv_j}{dx}-\beta_{0j}v_j\right]_{x=b}=\beta_{1j} \tag{10_2}$$

Таким образом, для отыскания $v_j(x)$ (при известной $v_{j-1}(x)$) имеем краевую задачу (8), (10). Для ее решения применим метод прогонки, описанный в § 9 главы 9. В соответствии с этим методом для отыскания $v_j(x)$ находим функции $\alpha_{0j}(x)$ и $\alpha_{1j}(x)$ $(a\leqslant x\leqslant b)$, удовлетворяющие уравнениям:

$$\alpha'_{0j}(x)+\alpha_{0j}^2(x)=\frac{1}{\mu_i^2(x)\,l}, \tag{11}$$

$$\alpha'_{1j}(x)+\alpha_{0j}(x)\,\alpha_{1j}(x)=-\frac{1}{\mu_j^2(x)\,l}\,v_{j-1}(x) \tag{12}$$

и начальным условиям:

$$\alpha_{0j}(a)=\alpha_{0j}; \quad \alpha_{1j}(a)=\alpha_{1j}, \tag{12'}$$

т. е. совершаем *прямую прогонку*. Далее, из системы

$$\left.\begin{aligned} v'_j(b)-\beta_{0j}v_j(b)&=\beta_{1j},\\ v'_j(b)-\alpha_{0j}(b)\,v_j(b)&=\alpha_{1j}(b)\end{aligned}\right\} \tag{13}$$

определяем $v_j(b)$ и, интегрируя уравнение $v'_j(x)=\alpha_{0j}(x)\,v_j(x)+\alpha_{1j}(x)$ с начальным условием $v_j(b)$ на правом конце, находим функцию $v_j(x)$, т. е. выполняем *обратную прогонку*.

Так как $v_0(x)=\varphi(x)$, то, используя этот метод, найдем последовательно $v_1(x)$, $v_2(x)$, ..., т. е. приближенные выражения для решения $u(x, t)$ на прямых $t_j=jl$. Если $\alpha_0(t)>0$, то при численном решении системы (11) и уравнения (8) мы не будем иметь резкой потери точности при отыскании значений $v_j(x)$.

Теперь видоизменим этот метод применительно к методу сеток, т. е. рассмотрим случай $h\neq 0$. В этом случае метод прогонки мы будем применять не к дифференцируемому уравнению, а к граничной задаче для разностного уравнения второго порядка (6) с граничными условиями (7_1), (7_2) (j считаем фиксированным, а $u_{i, j-1}$ — известными).

Уравнения (5) и (7) при $j=m$ можно записать в таком виде:

$$A_{im}u_{i+1,\,m}-2B_{im}u_{im}+C_{im}u_{i-1,\,m}=D_{im}, \tag{14}$$

$$u_{-1,\,m}=P_{0m}u_{0m}+Q_{0m}, \tag{15$_1$}$$

$$u_{N-1,\,m}=R_m u_{Nm}+S_m, \tag{15$_2$}$$

где

$$A_{im}=C_{im}=1;\quad B_{im}=1+\frac{h^2}{2l\mu_{im}^2};\quad D_{im}=-\frac{h^2}{l\mu_{im}^2}u_{i,\,m-1}; \tag{16}$$

$$\left.\begin{aligned} P_{0m}&=\frac{2-h\alpha_{0m}}{2+h\alpha_{0m}};\quad Q_{0m}=-\frac{2h\alpha_{1m}}{2+h\alpha_{0m}};\\ R_m&=\frac{2-h\beta_{0m}}{2+h\beta_{0m}};\quad S_m=-\frac{2h\beta_{1m}}{2+h\beta_{0m}}. \end{aligned}\right\} \tag{17}$$

В соответствии с идеей метода прогонок будем перегонять левое граничное условие (15$_1$) в правый граничный узел, т. е. будем находить такие P_{im} и Q_{im}, чтобы при всех $i=0,\ 1,\ 2,\ \ldots,\ N$

$$u_{i-1,\,m}=P_{im}u_{im}+Q_{im}. \tag{18}$$

Подставляя $u_{i-1,\,m}$ из (18) в (14), будем иметь:

$$A_{im}u_{i+1,\,m}-2B_{im}u_{im}+C_{im}P_{im}u_{im}=D_{im}-C_{im}Q_{im},$$

или, разрешая относительно u_{im},

$$u_{im}=P_{i+1,\,m}u_{i+1,\,m}+Q_{i+1,\,m},$$

где

$$P_{i+1,\,m}=\frac{A_{im}}{2B_{im}-C_{im}P_{im}}, \tag{19}$$

$$Q_{i+1,\,m}=\frac{C_{im}Q_{im}-D_{im}}{2B_{im}-C_{im}P_{im}}=\frac{C_{im}Q_{im}-D_{im}}{A_{im}}P_{i+1,\,m}. \tag{20}$$

Теперь решение находится просто. Зная A_{im}, B_{im}, C_{im}, D_{im}, P_{0m} и Q_{0m}, находим с помощью рекуррентных соотношений (19), (20) P_{im}, Q_{im} и далее с помощью (18) находим последовательно $u_{N-2,\,m}$, $u_{N-3,\,m}$, $\ldots$, $u_{0,\,m}$, $u_{-1,\,m}$.

Погрешности, допущенные при вычислении α_{0j}, α_{1j}, β_{0j}, β_{1j}, при этом методе не могут сильно сказаться на результате, если $\alpha_{0j}>0$. В самом деле, если $|P_{im}|<1$, то из (19) следует, что и $|P_{i+1,\,m}|<1$, так как $A_{im}=C_{im}=1$, а $B_{im}>1$. Но при $\alpha_{0,\,m}>0$ $|P_{0m}|<1$, поэтому $|P_{im}|<1$ при всех i. Погрешности в α_{0j} и α_{1j} вызовут погрешности в P_{0m} и Q_{0m} и будут сказываться на значениях P_{im}, Q_{im}. Покажем, что они не возрастают. Если δP_{im} погрешность в P_{im}, то с точностью до членов первого порядка относительно δP_{im}

$$\delta P_{i+1,\,m}=\frac{A_{im}C_{im}}{(2B_{im}-C_{im}P_{im})^2}\delta P_{im}.$$

В нашем случае $A_{im}=C_{im}=1$ и множитель при δP_{im} в правой части равен $P^2_{i+1,\,m}$, т. е.

$$\delta P_{i+1,\,m}=P^2_{i+1,\,m}\delta P_{im}.$$

Но $|P_{i+1,\,m}|<1$. Следовательно, с возрастанием i погрешность будет убывать. Из равенства (20) видно, что при вычислении $Q_{i+1,\,m}$ значение Q_{im} умножается на $P_{i+1,\,m}$, т. е. на величину, по модулю меньшую единицы, а это означает, что погрешность значения Q_{im} при переходе к $Q_{i+1,\,m}$ тоже не будет возрастать. При обратной прогонке погрешность при вычислении u_{im} не может возрастать, так как каждое предыдущее значение умножается на $P_{i+1,\,m}$, а $|P_{i,\,m}|<1$.

Если $\mu(x,\,t)=\text{const}$; $\alpha_j(t)=\text{const}$; $\beta_j(t)=\text{const}$ $(j=0,\,1)$, то P_{im} не зависят от m и их следует вычислить только один раз. Это уменьшает объем вычислительной работы.

Отметим, что мы специально записали уравнение (14) в общем виде, хотя в нашем случае $A_{im}=C_{im}=1$ и все соотношения имели бы более простой вид. Мы это сделали, желая показать, как можно применять метод прогонок для решения граничных задач для линейных разностных уравнений второго порядка, с которыми приходится встречаться во многих вопросах.

2. Уравнение Пуассона. Пусть в прямоугольнике $R\{a\leqslant x\leqslant b;\ c\leqslant x\leqslant d\}$ требуется найти решение уравнения Пуассона

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=f(x,\,y), \quad (21)$$

удовлетворяющее граничному условию

$$\frac{\partial u}{\partial n}+\alpha u=\beta \text{ на границе прямоугольника } (n \text{ — внешняя нормаль}). \quad (22)$$

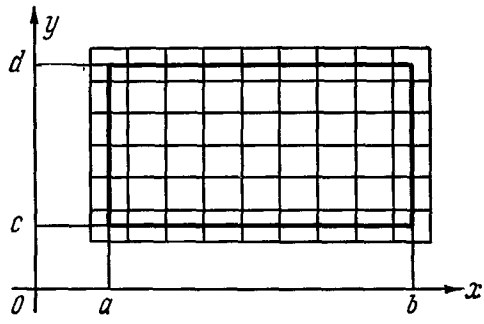

Рис. 73.

Для решения задачи применим метод сеток, выбрав в качестве узлов точки с координатами

$$x_i=a+\left(i+\frac{1}{2}\right)h;\quad y_j=c+\left(j+\frac{1}{2}\right)l;\quad \left(i=-1,\,0,\,1,\,2,\,\ldots,\,n;\right.$$

$$\left.j=-1,\,0,\,1,\,2,\,\ldots,\,m;\quad h=\frac{b-a}{n};\quad l=\frac{d-c}{m}\right).$$

Для внутренних узлов (i, j) запишем разностные уравнения

$$\frac{u_{i+1,j}-2u_{ij}+u_{i-1,j}}{h^2}+\frac{u_{i,j+1}-2u_{ij}+u_{i,j-1}}{l^2}=f_{ij}$$

$$(i=0,\ 1,\ 2,\ \ldots,\ n-1;\quad j=0,\ 1,\ 2,\ \ldots,\ m-1), \quad (23)$$

аппроксимирующие уравнение (21) с точностью $O(h^2+l^2)$, а граничные условия (22) аппроксимируем с той же точностью соотношениями:

$$\frac{u_{-1,j}-u_{0j}}{h}+\alpha_{0j}\frac{u_{0j}+u_{-1,j}}{2}=\beta_{0j}, \quad (24_1)$$

$$(j=0,1,2,\ \ldots,m-1),$$

$$\frac{u_{nj}-u_{n-1,j}}{h}+\alpha_{nj}\frac{u_{n-1,j}+u_{nj}}{2}=\beta_{nj} \quad (24_2)$$

$$\frac{u_{i,-1}-u_{i,0}}{l}+\alpha_{i0}\frac{u_{i,-1}+u_{i0}}{2}=\beta_{i0}, \quad (24_3)$$

$$(i=0,1,2,\ \ldots,n-1).$$

$$\frac{u_{i,m}-u_{i,m-1}}{l}+\alpha_{im}\frac{u_{i,m}+u_{i,m-1}}{2}=\beta_{im} \quad (24_4)$$

Таким образом, мы получим систему $mn+2(m+n)$ уравнений с таким же количеством неизвестных u_{ij}.

Используя граничные условия $(24_3)-(24_4)$, выразим $u_{i,-1}$, u_{im} через u_{i0}, $u_{i,m-1}$. Будем иметь:

$$u_{i,-1}=\frac{2-l\alpha_{i0}}{2+l\alpha_{i0}}u_{i0}+\frac{2l\beta_{i0}}{2+l\alpha_{i0}}=k_{i0}u_{i0}+z_{i0}, \quad (25_1)$$

$$u_{im}=\frac{2-l\alpha_{im}}{2+l\alpha_{im}}u_{i,m-1}+\frac{2l\beta_{im}}{2+l\alpha_{im}}=k_{im}u_{i,m-1}+z_{im}, \quad (25_2)$$

где

$$k_{i0}=\frac{2-l\alpha_{i0}}{2+l\alpha_{i0}};\quad k_{im}=\frac{2-l\alpha_{im}}{2+l\alpha_{im}};\quad z_{i0}=\frac{2l\beta_{i0}}{2+l\alpha_{i0}};\quad z_{im}=\frac{2l\beta_{im}}{2+l\alpha_{im}}. \quad (26)$$

Используя эти соотношения, исключим в системе (23) неизвестные $u_{i,-1}$, u_{im}. Если ввести обозначение $\gamma=\frac{h^2}{l^2}$, то получим систему

$$\left.\begin{aligned}
&u_{i+1,0}-(2+2\gamma-k_{i0}\gamma)u_{i0}+\gamma u_{i1}+u_{i-1,0}=F_{i0},\\
&u_{i+1,j}+\gamma u_{i,j-1}-2(1+\gamma)u_{ij}+\gamma u_{i,j+1}+u_{i-1,j}=F_{ij}\\
&\qquad\qquad (j=1,\ 2,\ \ldots,\ m-2),\\
&u_{i+1,m-1}+\gamma u_{i,m-2}-(2+2\gamma-k_{i,m-1}\gamma)u_{i,m-1}+u_{i-1,m-1}=\\
&\qquad =F_{i,m-1}\qquad (i=0,\ 1,\ 2,\ \ldots,\ n-1),
\end{aligned}\right\} \quad (27)$$

где

$$F_{i0}=h^2f_{i0}-\gamma z_{i0};\quad F_{ij}=h^2f_{ij}\quad (j=1,\ 2,\ \ldots,\ m-2);$$

$$F_{i,m-1}=h^2f_{i,m-1}-\gamma z_{im}. \quad (28)$$

Эту систему коротко можно записать в виде

$$\bar{v}_{i+1} + A_i\bar{v}_i + \bar{v}_{i-1} = \bar{F}_i \qquad (i = 0,\ 1,\ 2,\ \ldots,\ n-1), \tag{29}$$

где

$$\bar{v}_i = (u_{i0},\ u_{i1},\ \ldots,\ u_{i,\,m-1}); \quad \bar{F}_i = (F_{i0},\ F_{i1},\ \ldots,\ F_{i,\,m-1}),$$

$$A_i = \begin{pmatrix} -2(1+\gamma)+k_{i0}\gamma & \gamma & 0 & \ldots 0 & 0 \\ \gamma & -2(1+\gamma) & \gamma & \ldots 0 & 0 \\ 0 & \gamma & -2(1+\gamma) & \ldots 0 & 0 \\ \ldots\ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots\ldots \\ 0 & 0 & 0 & \ldots\gamma & -2(1+\gamma)+k_{im}\gamma \end{pmatrix}. \tag{30}$$

Граничные условия (24_1) и (24_2) можно переписать в виде

$$u_{-1,\,j} = \frac{2-h\alpha_{0j}}{2+h\alpha_{0j}}\,u_{0j} + \frac{2h\beta_{0j}}{2+h\alpha_{0j}} = k_{0j}u_{0j} + z_{0j} \quad (j = 0,\ 1,\ 2,\ \ldots,\ m-1), \tag{31_1}$$

$$u_{n-1,\,j} = \frac{2+h\alpha_{nj}}{2-h\alpha_{nj}}\,u_{nj} - \frac{2h\beta_{nj}}{2-h\alpha_{nj}} = k_{nj}u_{nj} + z_{nj} \quad (j = 0,\ 1,\ 2,\ \ldots,\ m-1), \tag{31_2}$$

где

$$k_{0j} = \frac{2-h\alpha_{0j}}{2+h\alpha_{0j}}; \quad z_{0j} = \frac{2h\beta_{0j}}{2+h\alpha_{0j}}; \quad k_{nj} = \frac{2+h\alpha_{nj}}{2-h\alpha_{nj}}; \quad z_{nj} = -\frac{2h\beta_{nj}}{2-h\alpha_{nj}}. \tag{32}$$

Положив

$$\bar{y}_0 = (z_{00},\ z_{01},\ \ldots,\ z_{0,\,m-1}); \quad \bar{r} = (z_{n0},\ z_{n1},\ \ldots,\ z_{n,\,m-1}),$$

$$X_0 = \begin{pmatrix} k_{00} & 0 & 0 \ldots 0 & 0 \\ 0 & k_{01} & 0 \ldots 0 & 0 \\ \ldots\ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 \ldots 0 & k_{0,\,m-1} \end{pmatrix}; \quad R = \begin{pmatrix} k_{n0} & 0 & 0 \ldots 0 & 0 \\ 0 & k_{n1} & 0 \ldots 0 & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 \ldots 0 & k_{n,\,m-1} \end{pmatrix}, \tag{33}$$

можно записать системы (31_1) и (31_2) в таком виде:

$$\left.\begin{aligned} \bar{v}_{-1} &= X_0\bar{v}_0 + \bar{y}_0, \\ \bar{v}_{n-1} &= R\bar{v}_n + \bar{r}. \end{aligned}\right\} \tag{34}$$

Окончательно имеем следующую систему уравнений:

$$\bar{v}_{-1} = X_0\bar{v}_0 + \bar{y}_0, \tag{35_1}$$

$$\bar{v}_{i+1} = A_i\bar{v}_i + \bar{v}_{i-1} + \bar{F}_i \qquad (i = 0,\ 1,\ 2,\ \ldots,\ n-1), \tag{35_2}$$

$$\bar{v}_{n-1} = R\bar{v}_n + \bar{r}. \tag{35_3}$$

Эту систему будем решать методом прогонки. Прямую прогонку мы совершим, если найдем такие матрицы X_i и векторы $\bar{y}_i$, чтобы при всех i имело место равенство

$$\bar{v}_{i-1} = X_i \bar{v}_i + \bar{y}_i \qquad (i = 0,\ 1,\ 2,\ \ldots,\ n). \tag{36}$$

Для отыскания X_i и $\bar{y}_i$ подставим $\bar{v}_{i-1}$ из (36) в (35_2). Получим:

$$\bar{v}_{i+1} + (A_i + X_i)\bar{v}_i + \bar{y}_i = \bar{F}_i \qquad (i = 0,\ 1,\ 2,\ \ldots,\ n-1)$$

или

$$\bar{v}_i = -(A_i + X_i)^{-1}\bar{v}_{i+1} + (A_i + X_i)^{-1}(\bar{F}_i - \bar{y}_i).$$

Таким образом,

$$X_{i+1} = -(A_i + X_i)^{-1}, \tag{37}$$

$$\bar{y}_{i+1} = (A_i + X_i)^{-1}(\bar{F}_i - \bar{y}_i) = X_{i+1}(\bar{y}_i - \bar{F}_i). \tag{38}$$

Так как X_0 и $\bar{y}_0$ известны, то с помощью (37) и (38) мы сможем найти X_i и $\bar{y}_i$ при всех $i = 1,\ 2,\ 3,\ \ldots,\ n$. Из (35_3) и (36) при $i = n$ получим:

$$(R - X_n)\bar{v}_n = \bar{y}_n - \bar{r} \tag{39}$$

и, следовательно, сможем найти $\bar{v}_n$. Далее, используя (36), последовательно находим $\bar{v}_{n-1}$, $\bar{v}_{n-2}$, $\ldots$, $\bar{v}_0$, $\bar{v}_{-1}$, т. е., выполнив обратную прогонку с помощью (36), найдем все нужные значения u_{ij} $(i = 0,\ 1,\ 2,\ \ldots,\ n;\ \ j = 0,\ 1,\ 2,\ \ldots,\ m)$.

При этом способе вместо решения системы из $mn + 2(m+n)$ уравнений необходимо обратить $n+1$ матриц порядка m, что значительно экономичнее с точки зрения объема вычислительной работы. Естественно оси целесообразно ориентировать так, чтобы было $m < n$.

Покажем, что погрешности в граничных значениях α_{0j}, α_{1j}, β_{0j}, β_{1j}, не приведут к значительным погрешностям в значениях u_{ij}, получаемым этим методом, если только $\alpha_{0j} > 0$. Ограничимся случаем $k_{i0} = k_{im} = 0$ $(i = 0,\ 1,\ 2,\ \ldots,\ m-1)$, что равносильно тому, что на отрезках $y = c$, $y = d$ $(a \leqslant x \leqslant b)$ заданы граничные условия первого рода: $u|_{y=c} = u|_{y=d} = 0$.

Для того чтобы доказать это утверждение, покажем сначала, что $\|X_i\|_3 \leqslant 1$. Если $k_{i0} = k_{im} = 0$ $(i = 0,\ 1,\ \ldots,\ m-1)$, то

$$A_i = A = \begin{pmatrix} -2(1+\gamma) & \gamma & 0 & 0\ldots0\ 0 & 0 \\ \gamma & -2(1+\gamma) & \gamma & 0\ldots0\ 0 & 0 \\ 0 & \gamma & -2(1+\gamma) & \gamma\ldots0\ 0 & 0 \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ 0 & 0 & 0 & 0\ldots0\ \gamma & -2(1+\gamma) \end{pmatrix}. \tag{40}$$

Матрица A симметрична, поэтому она имеет полную ортонормированную систему собственных векторов $\bar{z}_1, \bar{z}_2, \ldots, \bar{z}_m$, соответствующих собственным значениям $\lambda_1, \lambda_2, \ldots, \lambda_m$. Любой m-мерный вектор $\bar{z}$ можно представить в виде

$$\bar{z} = \alpha_1 \bar{z}_1 + \alpha_2 \bar{z}_2 + \ldots + \alpha_m \bar{z}_m,$$

а

$$A\bar{z} = \sum_{i=1}^{m} \alpha_i \lambda_i \bar{z}_i$$

и

$$\|A\bar{z}\|_3^2 = (A\bar{z}, A\bar{z}) = \sum_{i=1}^{m} \lambda_i^2 \alpha_i^2 \geqslant \min_{i=1, 2, \ldots, m} |\lambda_i|^2 \sum_{i=1}^{m} \alpha_i^2 =$$

$$= \min_{i=1, 2, \ldots, m} |\lambda_i|^2 \|\bar{z}\|_3^2.$$

Таким образом,

$$\|A\bar{z}\|_3 \geqslant \min_i |\lambda_i| \, \|\bar{z}\|_3. \tag{41}$$

Покажем, что $\min_i |\lambda_i| > 2$, а для этого просто вычислим характеристический определитель матрицы A, т. е.

$$D_m(\lambda) = \begin{vmatrix} -2(1+\gamma)-\lambda & \gamma & 0 & 0 \ldots 0 & 0 & 0 \\ \gamma & -2(1+\gamma)-\lambda & \gamma & 0 \ldots 0 & 0 & 0 \\ 0 & \gamma & -2(1+\gamma)-\lambda & \gamma \ldots 0 & 0 & 0 \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots\ldots & \ldots\ldots & \ldots\ldots \\ 0 & 0 & 0 & 0 \ldots 0 \; \gamma & -2(1+\gamma)-\lambda & \end{vmatrix}. \tag{42}$$

Раскрывая его по элементам первой строки, имеем:

$$D_m(\lambda) = -[2(1+\gamma)+\lambda] D_{m-1}(\lambda) - \gamma^2 D_{m-2}(\lambda). \tag{43}$$

Это соотношение можно рассматривать как линейное разностное уравнение второго порядка относительно $D_m(\lambda)$. Его общее решение имеет вид

$$D_m(\lambda) = C_1 \mu_1^m + C_2 \mu_2^m, \tag{44}$$

где μ_1 и μ_2 — корни уравнения

$$\mu^2 + 2\left[1 + \gamma + \frac{\lambda}{2}\right]\mu + \gamma^2 = 0,$$

т. е.

$$\left.\begin{aligned} \mu_1 &= -\left(1+\gamma+\frac{\lambda}{2}\right) + \sqrt{\left(1+\gamma+\frac{\lambda}{2}\right)^2 - \gamma^2}; \\ \mu_2 &= -\left(1+\gamma+\frac{\lambda}{2}\right) - \sqrt{\left(1+\gamma+\frac{\lambda}{2}\right)^2 - \gamma^2}. \end{aligned}\right\} \tag{45}$$

Для сокращения записи положим $1+\gamma+\frac{\lambda}{2}=\rho$. Тогда

$$\mu_1=-\rho+\sqrt{\rho^2-\gamma^2};\quad \mu_2=-\rho-\sqrt{\rho^2-\gamma^2}. \tag{46}$$

Для отыскания C_1, C_2 заметим, что

$$D_1(\lambda)=-2(1+\gamma)-\lambda=-2\rho;$$
$$D_2(\lambda)=\begin{vmatrix} -2(1+\gamma)-\lambda & \gamma \\ \gamma & -2(1+\gamma)-\lambda \end{vmatrix}=4\rho^2-\gamma^2,$$

поэтому

$$-2\rho=C_1\left(-\rho+\sqrt{\rho^2-\gamma^2}\right)+C_2\left(-\rho-\sqrt{\rho^2-\gamma^2}\right),$$
$$4\rho^2-\gamma^2=C_1\left(-\rho+\sqrt{\rho^2-\gamma^2}\right)^2+C_2\left(-\rho-\sqrt{\rho^2-\gamma^2}\right)^2,$$

откуда

$$C_1=\frac{-\rho+\sqrt{\rho^2-\gamma^2}}{2\sqrt{\rho^2-\gamma^2}};\quad C_2=\frac{\rho+\sqrt{\rho^2-\gamma^2}}{2\sqrt{\rho^2-\gamma^2}}$$

и

$$D_m(\lambda)=\frac{1}{2\sqrt{\rho^2-\gamma^2}}\left\{\left[-\rho+\sqrt{\rho^2-\gamma^2}\right]^{m+1}-\left[-\rho-\sqrt{\rho^2-\gamma^2}\right]^{m+1}\right\}. \tag{47}$$

Легко проверить, что $D_m(\lambda)=0$ при $\rho_k=\gamma\cos\frac{k\pi}{m+1}$, $1\leqslant k\leqslant m$, или

$$\lambda_k=-2-2\gamma\left(1-\cos\frac{k\pi}{m+1}\right)\qquad (k=1,\ 2,\ \ldots,\ m). \tag{48}$$

Отсюда следует, что $\min\limits_i|\lambda_i|>2$ и из (41)

$$\|A\bar{z}\|_3>2\|\bar{z}\|_3. \tag{49}$$

Докажем теперь, что если $\|X_i\|_3\leqslant 1$, то $\|X_{i+1}\|_3\leqslant 1$.

В самом деле, пусть $\bar{w}$ — произвольный вектор, а

$$\bar{z}=-(A+X_i)^{-1}\bar{w}=X_{i+1}\bar{w}.$$

Так как $\bar{w}=-(A+X_i)\bar{z}$, то

$$\|\bar{w}\|_3=\|(A+X_i)\bar{z}\|_3\geqslant\|A\bar{z}\|_3-\|X_i\bar{z}\|_3.$$

Но $\|A\bar{z}\|_3>2\|\bar{z}\|_3$, а $\|X_i\bar{z}\|_3\leqslant\|X_i\|_3\|\bar{z}\|_3\leqslant\|\bar{z}\|_3$. Отсюда

$$\|\bar{w}\|_3\geqslant 2\|\bar{z}\|_3-\|\bar{z}\|_3=\|\bar{z}\|_3=\|X_{i+1}\bar{w}\|_3,$$

а это означает, что $\|X_{i+1}\|_3\leqslant 1$.

Так как $\|X_0\|_3=\max\limits_{j=0,\,1,\,2,\,\ldots,\,m-1}|k_{0j}|<1$ (при $\alpha_{0j}>0$), то при всех i

$$\|X_i\|_3\leqslant 1.$$

Погрешности в значениях граничных функций вызовут погрешности в X_0, $\bar{F}_i$, R, $\bar{r}$, которые будут распространяться при прямой и обратной прогонках и скажутся на значениях u_{ij}. Вместо точных значений всех величин мы получим приближенные значения

$$X_i+\delta X_i;\quad \bar{y}_i+\delta\bar{y}_i;\quad \bar{F}_i+\delta\bar{F}_i;\quad \bar{v}_i+\delta\bar{v}_i$$

с погрешностями δX_i, $\delta\bar{y}_i$, $\delta\bar{F}_i$, $\delta\bar{v}_i$. Предполагая, что вычислительных погрешностей мы не допускаем, будем иметь:

$$X_{i+1}=-(A+X_i)^{-1};\quad X_{i+1}+\delta X_{i+1}=-(A_i+X_i+\delta X_i)^{-1},$$

откуда

$$X_{i+1}(A+X_i)=-I;\quad (X_{i+1}+\delta X_{i-1})(A_i+X_i+\delta X_i)=-I;$$
$$\delta X_{i+1}(A_i+X_i+\delta X_i)+X_{i+1}\delta X_i=0$$

или

$$\delta X_{i+1}=-X_{i+1}\delta X_i(A_i+X_i+\delta X_i)^{-1}=X_{i+1}\delta X_i(X_{i+1}+\delta X_{i+1}).$$

Таким образом, для линейной части погрешности имеем равенство

$$\delta X_{i+1}=X_{i+1}\delta X_i X_{i+1}.$$

Но так как $\|X_{i+1}\|_3\leqslant 1$, то

$$\|\delta X_{i+1}\|_3\leqslant\|\delta X_i\|_3,$$

т. е. погрешность δX_i не возрастает с возрастанием i.

Далее,

$$\bar{y}_{i+1}+\delta\bar{y}_{i+1}=(X_{i+1}+\delta X_{i+1})(\bar{y}_i-\bar{F}_i+\delta\bar{y}_i-\delta\bar{F}_i);$$
$$\bar{y}_{i+1}=X_{i+1}(\bar{y}_i-\bar{F}_i),$$

т. е. для линейной части погрешности имеет место равенство

$$\delta\bar{y}_{i+1}=X_{i+1}(\delta\bar{y}_i-\delta\bar{F}_i)+\delta X_{i+1}(\bar{y}_i-\bar{F}_i),$$

из которого видно, что в силу $\|X_{i+1}\|_3\leqslant 1$ быстрого роста погрешности $\delta\bar{y}_{i+1}$ не будет.

Так как обратную прогонку мы выполняем с помощью рекуррентного соотношения (36), то

$$\bar{v}_{i-1}=X_i\bar{v}_i+\bar{y}_i;\quad \bar{v}_{i-1}+\delta\bar{v}_{i-1}=(X_i+\delta X_i)(\bar{v}_i+\delta\bar{v}_i)+\bar{y}_i+\delta\bar{y}_i$$

и для линейной части погрешности $\delta\bar{v}_{i-1}$ имеем:

$$\delta\bar{v}_{i-1}=X_i\delta\bar{v}_i+\delta X_i\bar{v}_i+\delta\bar{y}_i,$$

т. е. и при обратной прогонке нет резкого возрастания погрешности.

§ 7. Сходимость и устойчивость разностных схем

В §§ 2, 3, 5 мы изложили метод сеток решения некоторых задач для простейших дифференциальных уравнений в частных производных. Сходимость метода доказывалась в каждом отдельном случае своим приемом. В § 5 было введено понятие устойчивости разностной схемы, играющее важную роль в решении задач методом сеток. В настоящем параграфе мы изложим метод сеток в более общем виде, рассмотрим связь сходимости и устойчивости и изложим некоторые методы исследования устойчивости. Эти вопросы обстоятельно освещены в монографии В. С. Рябенького и А. Ф. Филиппова «Об устойчивости разностных уравнений», которая и была использована при написании данного параграфа.

1. Разностная аппроксимация дифференциального уравнения и граничных условий. Рассмотрим в n-мерном пространстве область G с границей Γ, состоящей из нескольких кусков Γ_i ($i=1, 2, \ldots, m$), которые могут иметь общие части или даже совпадать между собой.

Пусть в области G нужно найти решение дифференциального уравнения

$$L(u)-f=0 \tag{1}$$

с граничными условиями

$$l_i(u)=\varphi_i \qquad (i=1, 2, \ldots, m), \tag{2}$$

где f — заданная в G функция, φ_i — функции заданные на Γ_i, а L и l_i — некоторые дифференциальные операторы.

В замкнутой области $G+\Gamma$ для каждого $h\,(0<h<h_0)$ определим некоторое множество точек, которое назовем *сеткой* и обозначим через G_h. Дифференциальному оператору L поставим в соответствие некоторый разностный оператор R_h, преобразующий функцию u_h, определенную на G_h, в функцию R_hu_h, определенную на некотором множестве $G_h^0 \subset G_h$. При этом будем предполагать, что какова бы ни была точка P области G, при достаточно малом h в любой ее окрестности найдутся точки, принадлежащие G_h и G_h^0. Дифференциальному уравнению (1) поставим в соответствие разностное уравнение

$$R_hu_h=f_h, \tag{3}$$

где f_h определена на G_h^0 и в точках G_h^0 совпадает с f. Каждому граничному условию $l_i(u)=\varphi_i$ на Γ_i поставим в соответствие некоторое разностное граничное условие

$$r_{ih}(u_h)=\varphi_{ih} \qquad (i=1, 2, \ldots, m), \tag{4}$$

где оператор r_{ih} определен на некотором множестве Γ_{ih} из G_h и переводит функцию u_h, определенную на Γ_{ih}, в функцию $r_{ih}(u_h)$.

определенную на множестве $\Gamma_{ih}^0 \subset \Gamma_{ih}$, а φ_{ih} — функции, определенные на Γ_{ih}^0, некоторым образом соответствующие функциям φ_i. Способ этого соответствия зависит от способа переноса граничных условий с Γ_i на Γ_{ih}.

Будем предполагать, что между функциями φ_{ih} выполнены *условия согласования*, под которыми мы будем понимать такие условия, связывающие φ_{ih} в отдельных точках, которые являются необходимыми и достаточными для существования хотя бы одной функции u_h, удовлетворяющей условиям (4).

Пусть U и F — классы функций, определенных на G, а Φ_i $(i = 1, 2, \ldots, m)$ — классы функций, определенных на Γ_i, такие, что при $u \in U$ определены $L(u)$ и $l_i(u)$, при этом $L(u) \in F$, $l_i(u) \in \Phi_i$. Будем предполагать, что в каждом из этих классов определена норма, вообще говоря, своя в каждом классе, обладающая обычными свойствами нормы. Эти нормы обозначим соответственно через $\|u\|_U$, $\|f\|_F$, $\|\varphi_i\|_{\Phi_i}$.

Пусть для функций u_h, определенных на G_h, определена норма $\|u_h\|_{U_h}$; для функций f_h, определенных на G_h^0, — норма $\|f_h\|_{F_h}$ и для функций φ_{ih}, определенных на Γ_{ih}^0, — норма $\|\varphi_{ih}\|_{\Phi_{ih}}$. Функции $u \in U$, $f \subset F$, определенные на G, имеют смысл и на G_h, следовательно, для них имеют смысл нормы $\|u\|_{U_h}$, $\|f\|_{F_h}$, оператор $R_h u$ и т. д. Будем предполагать, что нормы $\|\ \|_{U_h}$, $\|\ \|_{F_h}$, $\|\ \|_{\Phi_{ih}}$ определены так, чтобы для любых функций $u \in U$, $f \in F$ и $\varphi_i \subset \Phi_i$ имеют место предельные соотношения:

$$\|u\|_{U_h} \to \|u\|_U; \quad \|f\|_{F_h} \to \|f\|_F; \quad \|\varphi_{ih}\|_{\Phi_{ih}} \to \|\varphi_i\|_{\Phi_i} \quad (i = 1, 2, \ldots, m) \tag{5}$$

при $h \to 0$. В этом случае мы будем говорить, что соответствующие нормы согласованы.

Говорят, что *разностное уравнение* (3) *и граничные условия* (4) *аппроксимируют дифференциальное уравнение* (1) *и граничные условия* (2) *на классе функций* U, *если для любой функции* $u \in U$ *при* $h \to 0$ *имеют место соотношения:*

$$\|L(u) - R_h u\|_{F_h} \to 0, \tag{6}$$

$$\|[l_i(u)]_{i,h} - r_{ih}(u)\|_{\Phi_{ih}} \to 0 \qquad (i = 1, 2, \ldots, m), \tag{7}$$

где через $[l_i(u)]_{ih}$ обозначен оператор переноса граничных условий с Γ_i на Γ_{ih}.

Далее, говорят, что *порядок разностной аппроксимации равен* k, *если для любой функции* $u \in U$ *и* $0 < h < h_0$ *имеют место*

неравенства

$$\|L(u) - R_h u\|_{F_h} \leqslant h^k M, \tag{8}$$

$$\left\|[l_i(u)]_{ih} - r_{ih}(u)\right\|_{\Phi_{ih}} \leqslant h^k M_i \qquad (i = 1, 2, \ldots, m), \tag{9}$$

где M и M_i не зависят от h.

Пример. Рассмотрим уравнение

$$L(u) = \frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = f(x, t) \qquad (t > 0).$$

в области $G = \{0 < x < 1; \quad 0 < t < T\}$ с граничными условиями:

$$l_0(u) \equiv u(x, 0) = \varphi_0(x) \quad \text{на} \quad \Gamma_0 = \{0 \leqslant x \leqslant 1; \quad t = 0\},$$
$$l_1(u) \equiv u'_t(x, 0) = \varphi_1(x) \quad \text{на} \quad \Gamma_1 = \{0 \leqslant x \leqslant 1; \quad t = 0\},$$
$$l_2(u) \equiv u(0, t) = \varphi_2(t) \quad \text{на} \quad \Gamma_2 = \{x = 0; \quad 0 \leqslant t \leqslant T\},$$
$$l_3(u) \equiv u(1, t) = \varphi_3(t) \quad \text{на} \quad \Gamma_3 = \{x = 1; \quad 0 \leqslant t \leqslant T\}.$$

Под сеткой G_h будем понимать совокупность точек $(x_i = ih, \ t_j = jl)$ $(i = 0, 1, 2, \ldots, M; j = 0, 1, 2, \ldots, N)$, где $h = \frac{1}{M}$; $l = \alpha h$ $(\alpha = \text{const})$ $lN \leqslant T < l(N+1)$. Определим операторы:

$$R_h u_h = \frac{u_h(x, t+l) - 2u_h(x, t) + u_h(x, t-l)}{l^2} -$$
$$- \frac{u_h(x+h, t) - 2u_h(x, t) + u_h(x-h, t)}{h^2} \quad \text{на } G_h^0,$$

$$r_{0h}(u_h) \equiv u_h(x, 0) \quad \text{на } \Gamma_{0h},$$

$$r_{1h}(u_h) \equiv \frac{u_h(x, l) - u_h(x, 0)}{l} \quad \text{на } \Gamma_{1h},$$

$$r_{2h}(u_h) \equiv u_h(0, t) \quad \text{на } \Gamma_{2h},$$

$$r_{3h}(u_h) \equiv u_h(1, t) \quad \text{на } \Gamma_{3h}.$$

Здесь G_h^0 — совокупность точек (ih, jl) $(i = 1, 2, \ldots, M-1;$ $j = 1, 2, \ldots, N-1)$,

Γ_{0h} — совокупность точек $(ih, 0)$ $(i = 0, 1, 2, \ldots, M)$,

Γ_{1h} — совокупность точек $(ih, 0)$, (ih, l) $(i = 0, 1, 2, \ldots, M)$,

Γ_{2h} — совокупность точек $(0, jl)$ $(j = 0, 1, 2, \ldots, N)$,

Γ_{3h} — совокупность точек $(1, jl)$ $(j = 0, 1, 2, \ldots, N)$,

а $\Gamma_{0h}^0 \equiv \Gamma_{0h}$, $\Gamma_{1h}^0 \equiv \Gamma_{0h}$; $\Gamma_{2h}^0 \equiv \Gamma_{2h}$; $\Gamma_{3h}^0 \equiv \Gamma_{3h}$.

Положим, далее,

$$f_h(x_i, t_j) = f(ih, jl); \quad \varphi_{0h}(x_i) = \varphi_0(ih); \quad \varphi_{1h}(x_i) = \varphi_1(ih)$$
$$(i = 0, 1, 2, \ldots, M);$$
$$\varphi_{2h}(t_j) = \varphi_2(jl); \quad \varphi_{3h}(t_j) = \varphi_3(jl) \qquad (j = 0, 1, 2, \ldots, N).$$

Условиями согласования здесь будут условия:

$$\varphi_{0h}(0) = \varphi_{2h}(0); \quad \varphi_{0h}(0) + l\varphi_{1n}(0) = \varphi_{2h}(l);$$
$$\varphi_{0h}(1) = \varphi_{3h}(0); \quad \varphi_{0h}(1) + l\varphi_{1h}(1) = \varphi_{3h}(l),$$

которые получаются из следующих соображений. Точка $(0, 0)$ принадлежит Γ_{0h}, Γ_{1h} и Γ_{2h}, а точка $(0, l)$ принадлежит Γ_{1h} и Γ_{2h}. Следовательно, значения u_h в этих точках можно вычислять различными способами:

$$u_h(0, 0) = \varphi_{0h}(0); \qquad u_h(0, 0) = \varphi_{2n}(0);$$
$$u_h(0, l) = \varphi_{0h}(0) + l\varphi_{1h}(0); \qquad u_h(0, l) = \varphi_{2h}(l),$$

откуда

$$\varphi_{0h}(0) = \varphi_{2h}(0); \quad \varphi_{0h}(0) + l\varphi_{1h}(0) = \varphi_{2h}(l).$$

Аналогично получим и два других условия согласования.

За класс U примем совокупность функций с непрерывными производными второго порядка в замкнутой области $\bar{G}$, за F — совокупность всех непрерывных в $\bar{G}$ функций, а за Φ_i — совокупность всех непрерывных функций на Γ_i. Нормы в этих классах функций введем следующим образом:

$$\|u\|_U = \max_{\bar{G}} |u|; \quad \|f\|_F = \max_{\bar{G}} |f|; \quad \|\varphi_i\|_{\Phi_i} = \max_{\Gamma_i} |\varphi_i|.$$

Для сеточных функций нормы введем так:

$$\|u_h\|_{U_h} = \max_{G_h} |u_h|; \quad \|f_h\|_{F_h} = \max_{G_h^0} |f_h|; \quad \|\varphi_{ih}\|_{\Phi_{ih}} = \max_{\Gamma_{ih}^0} |\varphi_{ih}|.$$

Так определенные нормы будут согласованы, так как при $h \to 0$

$$\|u\|_{U_h} \to \|u\|_U; \quad \|f\|_{F_h} \to \|f\|_F; \quad \|[\varphi_i]_{ih}\|_{\Phi_{ih}} \to \|\varphi_i\|_{\Phi_i}.$$

Далее, при $h \to 0$

$$\|L(u) - R_h u\|_{F_h} \to 0; \quad \|[l_i(u)]_{ih} - r_{ih}(u)\|_{\Phi_{ih}} \to 0 \qquad (i = 0, 1, 2, 3).$$

Если вместо класса U дважды непрерывно дифференцируемых в $\bar{G}$ функций взять класс U_1 всех функций, имеющих непрерывные

производные четвертого порядка, то

$$\|L(u) - R_h u\|_{F_h} \leqslant \frac{h^2}{12}\left(\alpha^2 \max_G \left|\frac{\partial^4 u}{\partial x^4}\right| + \max_G \left|\frac{\partial^4 u}{\partial t^4}\right|\right),$$

$$\|[l_i(u)]_{ih} - r_{ih}(u)\|_{\Phi_{ih}} = 0 \qquad (i = 0,\ 2,\ 3)$$

$$\|[l_1(u)]_{ih} - r_{1h}(u)\|_{\Phi_{1h}} \leqslant \frac{\alpha h}{2} \max_G \left|\frac{\partial^2 u}{\partial t^2}\right|.$$

Таким образом, разностная схема

$$R_h u_h = f_h; \quad r_{ih}(u) = \varphi_{ih} \qquad (i = 0,\ 1,\ 2,\ 3)$$

аппроксимирует в классе U дифференциальное уравнение с граничными условиями, а в классе U_1 будем иметь аппроксимацию первого порядка.

2. Понятие корректности и устойчивости разностной схемы. Совокупность разностного уравнения (3) и разностных граничных условий (4) назовем *разностной схемой* решения задачи (1) — (2).

Введем следующие определения.

Будем говорить, что разностная схема (3) — (4) *корректна*, если при достаточно малом шаге h ее решение существует при любых f_h и φ_{ih} $(i = 1,\ 2,\ \ldots,\ m)$, для которых выполнены условия согласования, и для любого $\varepsilon > 0$ существует такое $\delta = \delta(\varepsilon) > 0$, что для данного решения u_h разностной схемы (3) — (4) и любого $\tilde{u}_h$, где

$$R_h \tilde{u}_h = \tilde{f}_h; \quad r_{ih}(\tilde{u}_h) = \tilde{\varphi}_{ih} \qquad (i = 1,\ 2,\ \ldots,\ m), \tag{10}$$

будет иметь место неравенство

$$\|\tilde{u}_h - u_h\|_{U_h} < \varepsilon \tag{11}$$

сразу для всех h $(0 < h < h_0)$, как только

$$\|f - \tilde{f}\|_{F_h} < \delta; \quad \|\varphi_{ih} - \tilde{\varphi}_{ih}\|_{\Phi_{ih}} < \delta. \tag{12}$$

Это означает, что решение u_h непрерывно зависит от правой части уравнения и правых частей граничных условий, причем зависимость равномерная по h.

Если уравнение (3) и граничные условия (4) линейны, то данное выше определение корректности равносильно следующему: разностная схема (3) — (4) корректна, если решение u_h существует при любых f_h и φ_{ih}, причем

$$\|u_h\|_{U_h} \leqslant N\|f_h\|_{F_h} + \sum_{i=1}^{m} N_i \|\varphi_{i,h}\|_{\Phi_{ih}}. \tag{11'}$$

Разностную схему (3) — (4) называют *устойчивой по правой части*, если ее решение существует и $\|u_h - \tilde{u}_h\|_{U_h} < \varepsilon$ для любых $\tilde{u}_h$, удовлетворяющих (10), при $\|\tilde{f} - f\|_{F_h} < \delta$ и $\tilde{\varphi}_{ih} = \varphi_{ih}$ ($i = 1, 2, 3, \ldots, m$).

Разностную схему называют *устойчивой по р граничным условиям*

$$r_{ih}(u_h) = \varphi_{ih} \qquad (i = 1, 2, \ldots, p \leqslant m),$$

если неравенство (11) имеет место при $\tilde{f} = f$ и $\varphi_{ih} = \tilde{\varphi}_{ih}$ ($i = p+1, \ldots, m$) и

$$\|\tilde{\varphi}_{ih} - \varphi_{ih}\|_{\Phi_{ih}} < \delta \qquad (i = 1, 2, \ldots, p).$$

Если разностная схема такова, что условия, входящие в последнее определение, являются аналогичными начальным условиям для дифференциальных уравнений гиперболического или параболического типа, то их называют *начальными условиями* для разностного уравнения и говорят об *устойчивости* разностной схемы *по начальным условиям*. Так как это понятие в дальнейшем встречается достаточно часто, то уточним его.

Пусть $t, x_1, x_2, \ldots, x_n$ — координаты точек рассматриваемого пространства и сетка G_h лежит в полупространстве $t \geqslant t_0$ и состоит из слоев $S_0, S_1, S_2, \ldots$ Под *слоем* S_j мы понимаем множество точек сетки, лежащих в плоскости $t = t_0(l) + jl$, где $t_0(l) \to t_0$ при $l \to 0$. Пусть первые p граничных условий (4) ($p \leqslant m$) однозначно определяют значения u_h в узлах слоев $S_0, S_1, \ldots, S_{q-1}$. Эти p условий мы и назовем начальными условиями разностного уравнения (3), если для любого j ($j = q, q+1, \ldots$) по значениям u_h в узлах слоев $S_{j-1}, S_{j-2}, \ldots, S_{j-q}$, используя только те уравнения из (3) — (4), которые связывают значения u_h в узлах слоев $S_{j-q}, S_{j-q+1}, \ldots, S_j$, можно однозначно определить значения u_h во всех узлах слоя S_j. При этих условиях решение u_h уравнения (3), удовлетворяющее граничным условиям (4), существует и единственно.

В примере, приведенном в п. 1, условия $r_{0h}(u_h) = \varphi_0$ и $r_{1h}(u_h) = \varphi_{1h}$ являются начальными условиями в этом смысле и $p = q = 2$, а $t_0(l) = t_0 = 0$.

Если разностная схема удовлетворяет требованиям, входящим в определение начальных условий, то начальные условия можно задавать на любых q последовательных слоях $S_{j-q+1}, S_{j-q+2}, \ldots, S_j$, при этом решение будет однозначно определено на всех последующих слоях $S_{j+1}, S_{j+2}, \ldots$ Для таких начальных условий введем обозначение

$$r_{ih,\,j}(u_h) = \varphi_{ih,\,j} \qquad (i = 1, 2, \ldots, p). \tag{13}$$

Пусть $\|r_{ih,\,j}(u_h)\|_{\Phi_{ih}}$ — норма функции $r_{ih,\,j}(u_h)$, аналогичная норме $\|r_{ih}(u_h)\|_{\Phi_{ih}}$ ($i = 1, 2, \ldots, p$). Помимо нормы $\|u_h\|_{U_h}$,

характеризующей поведение u_h на всей сетке G_h, введем нормы, характеризующие поведение u_h на каждом слое S_j, т. е. введем норму $\|u_h\|_{S_j}$, зависящую только от значений u_h в узлах слоя S_j. При всех j эти нормы должны определяться совершенно одинаково. Потребуем, чтобы норма $\|u_h\|_{S_j}$ была *согласована* с нормой $\|u_h\|_{U_h}$ в том смысле, что существует такая константа C, не зависящая от h и u_h, что для любой функции $u_h \subset U_h$ имеет место неравенство

$$\|u_h\|_{U_h} \leqslant C \max_j \|u_h\|_{S_j}. \tag{14}$$

Назовем разностную схему (3) — (4) *равномерно устойчивой по начальным условиям* в области G_h, если найдутся такие постоянные K, δ_0 и h_0, что при любых $j \geqslant q-1$, $\delta < \delta_0$ и $h < h_0$ для любых u_h и $\tilde{u}_h$, заданных на слоях S_{j-q+1}, S_{j-q+2}, ... из соотношений

$$\begin{aligned} &\|r_{ih,\,j}(u_h) - r_{ih,\,j}(\tilde{u}_h)\|_{\Phi_{ih}} < \delta \qquad (i = 1, 2, \ldots, p), \\ &r_{ih}(u_h) \equiv r_{ih}(\tilde{u}_h) \qquad\qquad (i = p+1, \ldots, m), \\ &R_h u_h \equiv R_h(\tilde{u}_h), \end{aligned}$$

при любом $N \geqslant j - q + 1$ и $S_N \subset G_h$ следует неравенство

$$\|u_h - \tilde{u}_h\|_{S_N} < K\delta.$$

Это означает, что при изменении меньше чем на δ начальных условий, задаваемых на любых q последовательных слоях, и сохранении без изменений разностного уравнения и остальных граничных условий решение разностной схемы на любом из последующих слоев изменится не больше чем на $K\delta$.

Устойчивость по начальным условиям всегда следует из равномерной устойчивости по начальным условиям.

В случае линейной разностной схемы равномерная устойчивость по начальным условиям означает, что из соотношений

$$\begin{aligned} &R_h u_h = 0; \quad \|r_{ih,\,j}(u_h)\|_{\Phi_{ih}} < \delta \qquad (i = 1, 2, \ldots, p); \\ &r_{ih}(u_h) = 0 \qquad (i = p+1, \ldots, m) \end{aligned}$$

следует неравенство

$$\|u_h\|_{S_N} \leqslant K\delta$$

при всех $N \geqslant j - q + 1$ и $S_N \subset G_h$.

Пусть область G лежит в полосе $t_0 \leqslant t \leqslant T$. Пусть при любом $j \geqslant q-1$ для любой функции, заданной на q последовательных слоях сетки S_{j-q+1}, ..., S_j, определена норма $\| \ \|_q^{(j)}$, где нижний индекс q указывает количество последовательных слоев, от значений функции u_h на которых зависит норма, а верхний индекс j указывает номер самого верхнего слоя из них.

Теорема. *Если при любом* $j \geqslant q-1$ *для любых функций* u_h *и* $\tilde{u}_h$, *заданных на слоях* S_{j-q+1}, S_{j-q+2}, ..., S_{j+1} *и удовлетворяющих уравнениям*

$$R_h u_h = f;\ R_h \tilde{u}_h = f;\ r_{ih}(u_h) = \varphi_{ih};\ r_{ih}(\tilde{u}_h) = \varphi_{ih}\ (i = p+1, \ldots, m), \tag{15}$$

имеет место неравенство

$$\|u_h - \tilde{u}_h\|_q^{(j+1)} \leqslant (1 + Kl)\|u_h - \tilde{u}_h\|_q^{(j)}, \tag{16}$$

где постоянная K *не зависит от* h, j, u_h, $\tilde{u}_h$ (l — шаг сетки по переменной t), *то разностная схема* (3) — (4) *равномерно устойчива по начальным условиям при любых нормах* $\|\ \|_{\Phi_{ih}}$ *и* $\|\ \|_{S_j}$, *удовлетворяющих неравенствам*

$$\left.\begin{aligned} &\|u_h - \tilde{u}_h\|_{S_j} \leqslant L\|u_h - \tilde{u}_h\|_q^{(j)}; \\ &\|u_h - \tilde{u}_h\|_q^{(j)} \leqslant N_0 \sum_{i=1}^{p} \|r_{ih,\,j}(u_h) - r_{ih,\,j}(\tilde{u}_h)\|_{\Phi_{ih}}, \end{aligned}\right\} \tag{17}$$

в которых L *и* N_0 *не зависят от* h, u_h, $\tilde{u}_h$.

Доказательство. Пусть u_h и $\tilde{u}_h$ определены в узлах слоев S_{j-q+1}, S_{j-q+2}, ... и удовлетворяют уравнениям (15). Тогда по неравенству (16) при всех $N \geqslant j$ имеем:

$$\|u_h - \tilde{u}_h\|_q^{(N)} \leqslant (1 + Kl)^{N-j}\|u_h - \tilde{u}_h\|_q^{(j)}.$$

Так как $(N-j)l \leqslant T - t_0$, то $(1 + Kl)^{N-j} \leqslant e^{K(T-t_0)}$. Отсюда и из неравенств (17) следует, что

$$\|u_h - \tilde{u}_h\|_{S_N} \leqslant LN_0 e^{K(T-t_0)} \sum_{i=1}^{p} \|r_{ih,\,j}(u_h) - r_{ih,\,j}(\tilde{u}_h)\|_{\Phi_{ih}},$$

а это и означает, что разностная схема равномерно устойчива по начальным значениям.

Эту теорему, дающую признак равномерной устойчивости разностной схемы по начальным условиям, грубо можно сформулировать следующим образом:

Для равномерной устойчивости по начальным условиям, достаточно, чтобы ошибка допущенная при вычислении решения при переходе от одного слоя к другому возрастала бы не более чем в $(1 + Kl)$ *раз, где* l *— шаг сетки по* t.

Рассмотрим теперь связь равномерной устойчивости по начальным условиям с устойчивостью по правой части.

Будем рассматривать разностную схему (3) — (4), предполагая, что первые p граничных условий в ней являются начальными условиями в ранее определенном смысле. Разностное уравнение (3)

и граничные условия (4) есть система уравнений, в которой неизвестными являются значения u_h в узлах сетки G_h. Все уравнения этой системы разобьем на группы, включив в k-ю группу все те уравнения, в которые входят значения u_k на k-м слое, но не входят значения в узлах вышележащих слоев S_{k+1}, S_{k+2}, ... Совокупность этих уравнений обозначим так:

$$R_h^{(k)} u_h = f; \quad r_{ih}^{(k)}(u_h) = \varphi_{ih} \qquad (i = p+1, p+2, \ldots, m) \quad (18)$$

Если u_h известна на слоях S_0, S_1, ..., S_{k-1}, то система (18) однозначно разрешима относительно значений u_h на S_k. Обозначим через $\|f\|^{(k)}$ норму, зависящую только от значений f, входящих в систему (18), согласованную с $\|f\|_{F_h}$ в том смысле, что для произвольной функции f, определенной на G_h^0, выполняется неравенство

$$\max_k \|f\|^{(k)} \leqslant \gamma \|f\|_{F_h},$$

где γ — постоянная, не зависящая от h и f.

Теорема. *Если разностная схема* (3) — (4) *равномерно устойчива по начальным значениям для всех таких* $\tilde{f}$, *что* $\|\tilde{f} - f\|_{F_h} \leqslant \delta_0$, *область* G_h *лежит в полосе* $t_0 \leqslant t \leqslant T$, *существуют такие постоянные* C, δ_0, h_0, *что при* $k \geqslant q$, $\delta < \delta_0$, $h < h_0$ *и любых* u_h *и* $\tilde{u}_h$, *совпадающих на слоях* S_{k-1}, S_{k-2}, ... *и удовлетворяющих соотношениям*

$$\left\|R_h^{(k)}(u_h) - R_h^{(k)}(\tilde{u}_h)\right\|^{(k)} \leqslant \delta; \quad r_{ih}^{(k)}(u_h) = r_{ih}^{(k)}(\tilde{u}_h), \qquad (19)$$

функции u_h *и* $\tilde{u}_h$ *на* S_k *удовлетворяют неравенству*

$$\|u_h - \tilde{u}_h\|_{S_k} \leqslant Cl^p\delta \qquad (20)$$

и для любых u_h *и* $\tilde{u}_h$, *совпадающих на слоях* S_{k-1}, S_{k-2}, ..., *выполнено неравенство*

$$\left\|r_{ih,k}(u_h) - r_{ih,k}(\tilde{u}_h)\right\|_{\Phi_{ih}} \leqslant \frac{D}{l^{p-1}} \|u_h - \tilde{u}_h\|_{S_k} \qquad (i = 1, 2, \ldots, p), \quad (21)$$

где D — *постоянная, а* p — *порядок дифференциального уравнения по* t, *то разностная схема* (3) — (4) *устойчива по правой части.*

Доказательство. Пусть u_h и $\tilde{u}_h$ удовлетворяют уравнениям

$$R_h u_h = f_h; \quad R_h \tilde{u}_h = \tilde{f}_h,$$

где $\|f-\tilde{f}\|_{F_h}\leqslant\delta$, и одним и тем же граничным условиям

$$r_{ih}(u_h)=r_{ih}(\tilde{u}_h)=\varphi_{ih}\qquad(i=1,\ 2,\ \ldots,\ m).$$

Обозначим через u_{kh} функцию, удовлетворяющую граничным условиям $r_{ih}(u_{kh})=\varphi_{ih}$, совпадающую с u_h на слоях $S_0,\ S_1,\ \ldots,\ S_k$ (т. е. удовлетворяющую уравнению $R_hu_{kh}=f$), а на слоях S_{k+1}, $S_{k+2},\ \ldots$ удовлетворяющую уравнению $R_hu_{kh}=\tilde{f}$. Здесь $q-1\leqslant k\leqslant N$. Функции $u_{k+1,h}$ и u_{kh} совпадают на слоях $S_0,\ S_1,\ \ldots,\ S_k$, а на S_{k+1} удовлетворяют уравнениям

$$R_h^{(k+1)}u_{kh}=\tilde{f};\qquad r_{ih}^{(k+1)}(u_{kh})=\varphi_{ih};$$
$$R_h^{(k+1)}u_{k+1,h}=f;\quad r_{ih}^{(k+1)}(u_{k+1,h})=\varphi_{ih}.$$

Из неравенства $\|\tilde{f}-f\|_{F_h}\leqslant\delta$ и условия согласования норм $\|f\|^{(k)}$ и $\|f\|_{F_h}$ следует неравенство

$$\|\tilde{f}-f\|^{(k+1)}\leqslant\gamma\delta,$$

т. е. для u_{kh} и $u_{k+1,h}$ выполнены условия, аналогичные условиям (19) теоремы. Поэтому

$$\|u_{kh}-u_{k+1,h}\|_{S_{k+1}}\leqslant C\gamma l^p\delta,$$

а в силу неравенства (20)

$$\|r_{ih,\,k+1}(u_{kh})-r_{ih,\,k+1}(u_{k+1,h})\|_{\Phi_{ih}}\leqslant CD\gamma l\delta. \tag{22}$$

На слоях S_{k+2}, $S_{k+3},\ \ldots$ функции u_{kh} и $u_{k+1,h}$ удовлетворяют одним и тем же уравнениям, а их начальные условия по неравенству (22) мало отличаются друг от друга. Следовательно, в силу равномерной устойчивости по начальным значениям из (22) следует, что при $k\leqslant N$

$$\|u_{kh}-u_{k+1,h}\|_{S_N}\leqslant LCD\gamma l\delta. \tag{23}$$

Так как на $S_0,\ S_1,\ \ldots,\ S_{q-1}$ имеем $u_{q-1,h}=u_h=\tilde{u}_h$, а на S_q, $S_{q+1},\ \ldots$ функции $u_{q-1,h}$ и $\tilde{u}_h$ определяются из одинаковых уравнений, то $u_{q-1,h}\equiv\tilde{u}_h$. С другой стороны, на S_N имеем $u_N=u_h$. Поэтому, написав неравенства (23) для $k=q,\ q+1,\ \ldots,\ N$ и сложив их почленно, получим:

$$\|\tilde{u}_h-u_h\|_{S_N}\leqslant LCD\gamma l\delta\,(N-q+1).$$

Так как область лежит в полосе $t_0\leqslant t\leqslant T$, то $Nl\leqslant T-t_0$. Поэтому в последнем неравенстве правая часть меньше некоторой постоянной, умноженной на δ. В силу условия согласования

норм $\|u_h\|_{U_h}$ и $\|u_h\|_{S_k}$ следует, что при достаточно малом δ для всех $h < h_0$ имеет место неравенство

$$\|\tilde{u}_h - u_h\|_{U_h} < \varepsilon,$$

а это и означает устойчивость по правой части.

Корректность разностной схемы равносильна устойчивости разностной схемы по правой части и всем граничным условиям. Из последней теоремы следует, что для некоторых разностных схем нет необходимости проверять устойчивость схемы по правой части, а достаточно проверить устойчивость по начальным значениям, что сильно упрощает проверку корректности разностной схемы.

3. Связь сходимости с корректностью разностной схемы. Из корректности разностной схемы (3) — (4), аппроксимирующей дифференциальное уравнение (1) с граничными условиями (2), следует сходимость последовательности решений разностной схемы к точному решению граничной задачи для дифференциального уравнения. Поэтому вместо доказательства сходимости достаточно установить корректность разностной схемы. Это следует из теоремы:

Если решение дифференциального уравнения (1) *с граничными условиями* (2) *существует и принадлежит* U, *а разностная схема* (3) — (4) *аппроксимирует уравнение* (1) *с граничными условиями* (2) *на классе* U *и корректна, то при* $h \to 0$ *решения* u_h *разностной схемы сходятся по норме к решению и граничной задачи для дифференциального уравнения, т. е.*

$$\|u - u_h\|_{U_h} \to 0.$$

Если уравнения (1) *и* (3) *и граничные условия* (2) *и* (4) *линейны и порядок аппроксимации равен* k, *то имеет место следующая оценка скорости сходимости*:

$$\|u - u_h\|_{U_h} \leqslant h^k \left(MN + \sum_{i=1}^{m} M_i N_i \right), \tag{24}$$

причем

$$\|u\|_U \leqslant N \|f\|_F + \sum_{i=1}^{m} N_i \|\varphi_i\|_{\Phi_i}. \tag{25}$$

Доказательство. Если $u \in U$: $L(u) = f$; $l_i(u) = \varphi_i$, то, обозначая $R_h u$ через $\tilde{f}$, $r_{ih}(u)$ — через $\tilde{\varphi}_{ih}$, из условия аппроксимации при достаточно малых h имеем:

$$\|f - \tilde{f}\|_{F_h} < \delta; \quad \|\varphi_{ih} - \tilde{\varphi}_{ih}\|_{\Phi_{ih}} < \delta,$$

а тогда из условия корректности следует, что при достаточно малом δ имеет место неравенство

$$\|u_h - u\|_{U_h} < \varepsilon.$$

В линейном случае, так как

$$R_h u_h = f = L(u); \quad r_{ih}(u_h) = \varphi_{ih} = [\varphi_i]_{ih} = [l_i(u)]_{ih},$$

то

$$L(u) - R_h u = R_h u_h - R_h u = R_h(u_h - u) = R_h v_h,$$
$$[l_i(u)]_{ih} - r_{ih}(u) = r_{ih}(u_h) - r_{ih}(u) = r_{ih}(v_h),$$

где $v_h = u_h - u$. Тогда из условия, что мы имеем аппроксимацию порядка k, следует

$$\|R_h v_h\|_{F_h} \leqslant h^k M; \quad \|r_{ik}(v_h)\|_{\Phi_{ih}} \leqslant h^k M_i.$$

В силу корректности разностной схемы для любой функции u_h имеет место неравенство

$$\|u_h\|_{U_h} \leqslant N \|R_h u_h\|_{F_h} + \sum_{i=1}^{m} N_i \|r_{ih}(u_h)\|_{\Phi_{ih}}, \tag{26}$$

откуда и получаем требуемую оценку скорости сходимости, полагая $u_h = v_h$ и используя оценки для v_h.

Для доказательства неравенства (25) по свойству норм имеем:

$$\|u\|_{U_h} \leqslant \|u - u_h\|_{U_h} + \|u_h\|_{U_h}.$$

Но $\|u - u_h\|_{U_h} \to 0$ при $h \to 0$, а для $\|u_h\|_{U_h}$ имеет место неравенство (26). Отсюда, переходя к пределу, и получим неравенство (25).

Заметим следующее:

Если некоторые из граничных условий (2) *аппроксимируются точно*, т. е. при некоторых i $\Gamma_{ih}^0 \subset \Gamma_i$ и для $u \in U$ $r_{ih}(u) = l_i(u)$, $\varphi_{ih} = [\varphi_i]_{ih} = \varphi_i$ на Γ_{ih}^0, *то требование устойчивости по соответствующим граничным условиям в доказанной теореме можно отбросить и требовать лишь устойчивость по правым частям и всем остальным граничным условиям.*

Представляет интерес следующая теорема, используя которую можно обосновать метод Рунге приближенной оценки погрешности метода сеток:

Теорема. *Если уравнение* (3) *и граничные условия* (4) *линейны и выполнены условия предыдущей теоремы, а аппроксимация такова, что существуют пределы*

$$\lim_{h \to 0} h^{-k}(Lu - R_h u) = \psi; \quad \lim_{h \to 0} h^{-k}\left([l_i(u)]_{ih} - r_{ih}(u)\right) = \psi_i, \tag{27}$$

где u — решение задачи (1) — (2), *т. е. существуют такие функции* ψ *и* ψ_i, *что*

$$\left.\begin{aligned} \lim_{h\to 0} \|h^{-k}(L(u)-R_h u)-\psi\|_{F_h} &= 0,\\ \lim_{h\to 0} \|h^{-k}([l_i(u)]_{ih}-r_{ih}(u))-[\psi_i]_{ih}\|_{\Phi_{ih}} &= 0, \end{aligned}\right\} \tag{28}$$

а w есть решение граничной задачи

$$L(w)=\psi;\quad l_i(w)=\psi_i \qquad (i=1,\ 2,\ \ldots,\ m), \tag{29}$$

принадлежащее к некоторому классу W, на котором R_h и r_{ih} аппроксимируют L и l_i, то

$$\lim_{h\to 0} \|h^{-k}(u_h-u)-w\|_{U_h}=0. \tag{30}$$

Доказательство. Пусть u — решение граничной задачи (1) — (2), а u_h — решение разностной схемы (3) — (4). Пусть $R_h u=\tilde{f}$, $r_{ih}(u)=\tilde{\varphi}_{ih}$. По условию теоремы

$$h^{-k}(f-\tilde{f})=\psi+\alpha_h;\quad h^{-k}(\varphi_{ih}-\tilde{\varphi}_{ih})=[\psi_i]_{ih}+\alpha_{ih},$$

где $\|\alpha_h\|_{F_h}\to 0$ и $\|\alpha_{ih}\|_{\Phi_{ih}}\to 0$ при $h\to 0$, т. е.

$$\left.\begin{aligned} h^{-k}(R_h u_h-R_h u)&=\psi+\alpha_h,\\ h^{-k}(r_{ih}(u_h)-r_{ih}(u))&=[\psi_i]_{ih}+\alpha_{ih}. \end{aligned}\right\} \tag{31}$$

Если w есть решение задачи (29), то по определению аппроксимации

$$R_h w=\psi+\beta_h;\quad r_{ih}(w)=[\psi_i]_{ih}+\beta_{ih},$$

где $\|\beta_h\|_{F_h}$ и $\|\beta_{ih}\|_{\Phi_{ih}}\to 0$ при $h\to 0$. Так как R_h и r_{ih} линейны, то из последнего равенства и равенств (31) имеем:

$$R_h\left(h^{-k}(u_h-u)-w\right)=\alpha_h-\beta_h,$$
$$r_{ih}\left(h^{-k}(u_h-u)-w\right)=\alpha_{ih}-\beta_{ih}.$$

Так как при $h\to 0$ правые части по соответствующим нормам стремятся к нулю, то в силу корректности разностной схемы (3) — (4)

$$\lim_{h\to 0} \|h^{-k}(u_h-u)-w\|_{U_h}=0.$$

Эта теорема позволяет оценить погрешность в решении, которую мы получаем, заменяя дифференциальное уравнение разностным.

Пусть u_{h_1} и u_{h_2} — решения разностной схемы при $h=h_1$ и $h=h_2$, где $h_1=ch_2 (c>1)$ и сетка G_{h_1} есть часть сетки G_{h_2}. Если выполнены

условия последней теоремы, то

$$u_{h_1} = u + h_1^k w + o(h_1^k); \quad u_{h_2} = u + h_2^k w + o(h_2^k).$$

Исключая из этих равенств w, получим:

$$u = u_{h_2} + \frac{1}{c^k - 1}(u_{h_2} - u_{h_1}) + o(h_2^k),$$

откуда

$$u - u_{h_2} \approx \frac{1}{c^k - 1}(u_{h_2} - u_{h_1}).$$

Этой формулой иногда пользуются для получения более точного решения, чем u_{h_2}.

Пример. Пусть в области G с границей Γ требуется найти решение уравнения

$$a(x, y)\, u''_{xx} + b(x, y)\, u''_{yy} + c(x, y)\, u'_x + d(x, y)\, u'_y + e(x, y)\, u = f(x, y); \quad u|_\Gamma = \varphi,$$

где a, b, c, d, e, f — заданные функции, непрерывные в $G + \Gamma$, удовлетворяющие следующим условиям:

$$a \geqslant 0; \quad b \geqslant 0; \quad a + b \geqslant q > 0; \quad |c| \leqslant Ma; \quad |d| \leqslant Mb; \quad e \leqslant 0$$

(M и q — положительные постоянные), а φ — заданная функция, непрерывная на Γ. Будем предполагать, что область G лежит в круге $x^2 + y^2 < R^2$. Рассмотрим сетку G_h, состоящую из точек $x = ih$, $y = jh$, лежащих в $G + \Gamma$. Назовем граничными узлами сетки G_h те ее точки, для которых хотя бы одна из четырех ее соседних точек $(x \pm h, y)$, $(x, y \pm h)$ лежит вне $G + \Gamma$.

Рассмотрим следующую разностную схему:

$$R_h u_h = \frac{a_{ij}}{h^2}(u_{i+1,j} - 2u_{ij} + u_{i-1,j}) + \frac{b_{ij}}{h^2}(u_{i,j+1} - 2u_{ij} + u_{i,j-1}) + \frac{c_{ij}}{2h}(u_{i+1,j} - u_{i-1,j}) + \frac{d_{ij}}{2h}(u_{i,j+1} - u_{i,j-1}) + e_{ij}u_{ij} = f_{ij};$$

$$u_h|_{\Gamma_h} = \varphi_1|_{\Gamma_h},$$

где $u_{ij} = u_h(ih, jh)$, а a_{ij}, b_{ij}, c_{ij}, d_{ij}, e_{ij}, f_{ij} — значения соответствующих функций в точке $x = ih$, $y = jh$, а φ_1 — функция, полученная из φ непрерывным продолжением на всю область G.

Докажем корректность этой схемы. Пусть

$$v(x, y) = e^{A(R^2+1)} - e^{A(x^2+y^2)}.$$

Так как

$$e^{A[(x+h)^2+y^2]} - e^{A[(x-h)^2+y^2]} = 2Ah(x+\theta_1 h)\, e^{A[(x+\theta_1 h)^2+y^2]}$$

$$(-1 < \theta_1 < 1),$$

$$e^{A[x^2+(y+h)^2]} - e^{A[x^2+(y-h)^2]} = 2Ah(y+\theta_2 h)\, e^{A[x^2+(y+\theta_2 h)^2]}$$

$$(-1 < \theta_2 < 1),$$

$$e^{A[(x+h)^2+y^2]} - 2e^{A[x^2+y^2]} + e^{A[(x-h)^2+y^2]} = Ah^2\{[1+2A(x+\theta_3 h)^2] \times$$

$$\times e^{A[(x+\theta_3 h)^2+y^2]} + [1+2A(x+\theta_4 h)^2]\, e^{A[(x+\theta_4 h)^2+y^2]}\}$$

$$(0 < \theta_3 < 1;\ -1 < \theta_4 < 0),$$

$$e^{A[x^2+(y+h)^2]} - 2e^{A[x^2+y^2]} + e^{A[x^2+(y-h)^2]} = Ah^2\{[1+2A(y+\theta_5 h)^2] \times$$

$$\times e^{A[x^2+(y+\theta_5 h)^2]} + [1+2A(y+\theta_6 h)^2]\, e^{A[x+(y+\theta_6 h)^2]}\}$$

$$(0 < \theta_5 < 1;\ -1 < \theta_6 < 0),$$

то

$$R_h e^{A(x^2+y^2)} \leqslant A(1+2AR^2)(a_{ij}+b_{ij}) + |c_{ij}|\, AR + |d_{ij}|\, AR \leqslant$$

$$\leqslant A(1+2AR^2)(a_{ij}+b_{ij}) + MAR(a_{ij}+b_{ij}) = C(a_{ij}+b_{ij})$$

$$(C = A(1+2AR^2+MR)).$$

Отсюда

$$R_h v = R_h e^{A(R^2+1)} - R_h e^{A(x^2+y^2)} \leqslant -C(a_{ij}+b_{ij}) < -Cq,$$

а

$$v = e^{AR^2}(e^A - e^{A(x^2+y^2-R^2)}) > e^A - e^{A(x^2+y^2-R^2)} > e^A - 1.$$

Если u_h есть решение разностной схемы при $|f| < Cq\delta$ и $|\varphi| < (e^A - 1)\delta$, то

$$R_h(u_h - v\delta) > 0 \text{ в } G_h \text{ и } u_h - v\delta < 0 \text{ на } \Gamma_h.$$

Следовательно, $u_h - v\delta$ не может иметь положительного максимума в $G_h + \Gamma_h$ (см. § 2) и $u_h - v\delta < 0$. Аналогично получим неравенство $u_h + v\delta > 0$, т. е.

$$|u_h| < v\delta$$

и

$$\|u_h\|_{U_h} = \max_{G_h} |u_h| \leqslant \delta e^{A(R^2+1)},$$

а это означает, что однородная задача $R_h u_h = 0$; $u_h|_{\Gamma_h} = 0$ имеет только тривиальное решение. Следовательно, рассматривая разностную схему как систему линейных алгебраических уравнений с неизвестными значениями u_h в узлах сетки, можно заключить, что ее определитель отличен от нуля, а поэтому разностная схема имеет решение при всех f и φ. Так как оценка u_h не зависит от h, то

разностная схема корректна, если за $\| \ \|_{U_h}$, $\| \ \|_{F_h}$ взять максимумы абсолютных величин соответствующих сеточных функций на соответствующих множествах.

Из теоремы сходимости корректной разностной схемы будет следовать, что если граничная задача для нашего дифференциального уравнения имеет дважды непрерывно дифференцируемое решение, то u_h равномерно сходится к этому решению при $h \to 0$.

4. Некоторые приемы исследования устойчивости разностных схем. *Исследование устойчивости с помощью принципа максимума.* Если для разностной схемы (3) — (4) имеет место в какой-либо форме принцип максимума, то часто удается, используя его, доказать устойчивость по правой части и устойчивость по граничным условиям этой разностной схемы. Такой прием был использован в примере п. 3, где использовался принцип максимума в форме: если $R_h u_h \geqslant 0$ и $r_{ih}(u_h) = 0$, то u_h не может иметь положительного максимума ни внутри, ни на границе области.

Исследование устойчивости с помощью индекса разностной схемы. Если разностная схема линейна и p граничных условий являются начальными условиями в том смысле, в котором они были определены в п. 2, то при любом $k \geqslant q - 1$ значения u_h на любом слое S_{k+1} можно выразить в виде линейной комбинации значений u_h в точках слоев S_k, $S_{k-1}, \ldots$ и члена, зависящего только от правых частей f_h и φ_{ih} разностной схемы (3) — (4). Максимум суммы абсолютных величин коэффициентов этой линейной комбинации по всем узлам сетки называют *индексом* I разностной схемы (3) — (4). *Если область G конечна, а h и l — соответственно шаги сетки по пространственным координатам x_1, $x_2, \ldots, x_n$ и времени t, то если существует постоянная $C > 0$, не зависящая от h и l, такая, что $I < 1 + Cl$, а нормы определены следующими равенствами*:

$$\|u_h\|_{S_k} = \max_{S_k} |u_h|; \quad \|r_{ih}(u_h)\|_{\Phi_{ih}} = \max_{S_0, S_1, \ldots, S_{q-1}} |u_h|,$$

то разностная аппроксимация (3) — (4) *равномерно устойчива по начальным условиям.*

Это утверждение непосредственно следует из теоремы о равномерной устойчивости по начальным условиям, так как неравенства (16) и (17), фигурирующие в условии этой теоремы, будут иметь место, если положить

$$\|u_h\|_q^{(k)} = \max_{S_{k-q+1}, \ldots, S_k} |u_h|.$$

Приведем пример применения этого способа исследования устойчивости.

Рассмотрим дифференциальное уравнение

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0$$

в области $G = \{0 \leqslant t \leqslant T; \quad 0 \leqslant x \leqslant 1\}$ с граничными условиями

$$u(x, 0) = \varphi(x); \quad u(0, t) = u(1, t) = 0,$$

где $\varphi(0) = \varphi(1) = 0$.

Эта задача аппроксимируется с помощью разностной схемы

$$\frac{1}{l}(u_{i, j+1} - u_{ij}) = \frac{1}{h^2}(u_{i+1, j} - 2u_{ij} + u_{i-1, j}); \quad u_{i0} = \varphi(ih);$$

$$u_{0j} = u_{Nj} = 0,$$

где $u_{ij} = u_h(ih, jl)$, а сетка состоит из точек $x = ih$, $t = jl$

$$\left(l = rh^2 \left(r = \text{const} \leqslant \frac{1}{2}\right); \quad i = 1, 2, \ldots, N; \quad Nh = 1;\right.$$

$$\left. j = 0, 1, 2, \ldots, M; \quad Ml \leqslant T < (M+1)l\right).$$

Так как

$$u_{i, j+1} = ru_{i-1, j} + (1 - 2r)\, u_{ij} + ru_{i+1, j},$$

то $I = 1$, и поэтому разностная схема устойчива.

Для того же уравнения и при той же области G рассмотрим граничные условия

$$u(x, 0) = \varphi(x); \quad \left(\frac{\partial u}{\partial x} + au\right)_{x=0} = 0; \quad \left(\frac{\partial u}{\partial x} + bu\right)_{x=1} = 0.$$

Разностное уравнение оставим прежним, а граничные условия для разностного уравнения запишем в виде

$$u_{i0} = \varphi(ih); \quad \frac{u_{1j} - u_{0j}}{h} + a_j u_{0j} = 0; \quad \frac{u_{Nj} - u_{N-1, j}}{h} + b_j u_{Nj} = 0.$$

Тогда

$$u_{i, j+1} = ru_{i-1, j} + (1 - 2r)\, u_{ij} + ru_{i+1, j},$$

$$u_{0, j+1} = \frac{1}{1 - a_j h}\, u_{1, j+1},$$

$$u_{N, j+1} = \frac{1}{1 + b_j h}\, u_{N-1, j+1}.$$

Таким образом,

$$I = \max_j \left\{1, \frac{1}{1 - a_j h}, \frac{1}{1 + b_j h}\right\}$$

и разностная схема будет устойчива, если только имеют место неравенства $a \leqslant 0$; $b \geqslant 0$.

Исследование устойчивости путем изучения роста единичной ошибки. Пусть снова разностная схема (3) — (4) линейна

и первые p граничных условий являются начальными. Если при вычислении решения разностного уравнения допущена ошибка, равная ε, только в одном узле слоя S_k, а во всех других узлах мы не делаем новых ошибок, то она вызовет ошибки в некоторых узлах слоев S_{k+1}, S_{k+2}. Если эти ошибки быстро растут с возрастанием номера слоя, то можно ожидать, что схема не будет устойчивой, а если же они не растут, то можно надеяться на устойчивость. Это — метод исследования устойчивости с помощью ε-*схемы*, о котором мы говорили в § 5. Эти соображения лежат в основе принципа устойчивости, применимого к исследованию многих явных разностных схем.

Пусть область G конечна, а l—шаг сетки по t, $h_1, h_2, \ldots, h_n$ — шаги сетки по $x_1, x_2, \ldots, x_n$. Если разностная схема линейна и первые p граничных условий являются начальными условиями для разностного уравнения, имеющего вид

$$\frac{a}{l^p} u_h(kl, x_1, x_2, \ldots, x_n) + \sum = f,$$

где $a = a(t, x_1, x_2, \ldots, x_n) \geqslant a_0 > 0$, а через $\sum$ обозначена сумма членов, содержащих значения u_h в узлах слоев S_{k-1}, $S_{k-2}, \ldots, S_{k-q}$, и v_h есть решение однородного уравнения $R_h u_h = 0$, $r_{ih}(u_h) = 0$ $(i = p+1, p+2, \ldots, m)$, равное нулю во всех узлах слоев S_{k-q+1}, $S_{k-q+2}, \ldots, S_k$, кроме какого-либо одного узла слоя S_k, где $v_h = 1$ и любое такое решение на любом слое S_K $(K \geqslant k)$ удовлетворяет неравенству

$$\sum_{S_K} |v_h| \leqslant C (K-k)^{p-1}, \tag{32}$$

где C не зависит от $k, K, l, h_1, h_2, \ldots, h_n$ и от выбора точки, в которой $v_h = 1$, то разностная схема (3) — (4) устойчива по правой части в норме

$$\|u_h\|_{U_h} = l h_1 h_2 \ldots h_n \sum_{G_h} |u_h|; \quad \|f\|_{F_h} = l h_1 h_2 \ldots h_n \sum_{G_h^0} |f|.$$

Для доказательства представляем f в виде суммы функций, каждая из которых отлична от нуля лишь в одном узле сетки. Тогда u_h будет суммой решений, аналогичных v_h и оцениваемых по неравенству (32).

Исследование устойчивости методом разделения переменных. Этот метод исследования устойчивости мы уже применяли в § 5. Здесь мы проиллюстрируем его еще на одном примере.

Рассмотрим разностный метод решения задачи о колебании гибкой струны, закрепленной на концах $x = 0$; $x = 1$, рассмотренный в примере п. 1 данного параграфа. Разностная схема и сетка те же, что и в указанном примере. Для исследования устойчивости рассмотрим однородное разностное уравнение с однородными граничными

условиями, кроме начальных, т. е будем рассматривать схему

$$R_h u_h = \frac{u_h(x, t+l) - 2u_h(x, t) + u_h(x, t-l)}{l^2} - \frac{u_h(x+h, t) - 2u_h(x, t) + u_h(x-h, t)}{h^2} = 0,$$

$$r_{0h}(u_h) = u_h(ih, 0) = \varphi_0(ih), \quad r_{1h}(u_h) = \frac{u_h(ih, l) - u_h(ih, 0)}{l} = \varphi_1(ih),$$

$$r_{2h}(u_h) = u_h(0, jl) = 0, \quad r_{3h}(u_h) = u_h(Nh, jl) = 0$$

и применим к ее решению метод разделения переменных, положив

$$u_h(kh, jl) = v(k)\, w(j).$$

Так как коэффициенты уравнений не зависят от t, можно взять $w(j) = \lambda^j$, где λ — пока неизвестная величина. Подставляя $u_h = \lambda^j v(k)$ в разностное уравнение, получим:

$$\frac{(\lambda^{j+1} - 2\lambda^j + \lambda^{j-1})\, v(k)}{l^2} - \frac{\lambda^j [v(k+1) - 2v(k) + v(k-1)]}{h^2} = 0$$

или

$$v(k+1) - 2v(k) + v(k-1) = \frac{h^2}{l^2}\left(\lambda - 2 + \frac{1}{\lambda}\right) v(k).$$

Мы имеем линейное разностное уравнение с постоянными коэффициентами. Для $v(k)$ в силу граничных условий имеем:

$$v(0) = v(N) = 0.$$

Будем искать решение этого уравнения в виде $v(k) = \alpha^k$, где α — искомая величина. Подстановка в уравнение после сокращения на α^{k-2} дает

$$\alpha^2 - 2\alpha + 1 = \frac{h^2}{l^2}\left(\lambda - 2 + \frac{1}{\lambda}\right)\alpha = \rho\alpha \qquad \left(\rho = \frac{h^2}{l^2}\left(\lambda - 2 + \frac{1}{\lambda}\right)\right),$$

или

$$\alpha^2 - (2+\rho)\alpha + 1 = 0,$$

откуда

$$\alpha_1 = \frac{1}{\alpha_2} = \frac{2 + \rho + \sqrt{\rho^2 + 4\rho}}{2}.$$

Общее решение уравнения имеет вид

$$v(k) = C_1 \alpha_1^k + C_2 \alpha_2^k,$$

где C_1, C_2 — произвольные постоянные. Из граничных условий имеем:

$$v(0) = C_1 + C_2 = 0, \qquad \text{т. е.} \quad C_2 = -C_1,$$

$$v(N) = C_1(\alpha_1^N - \alpha_2^N) = 0, \quad \text{т. е.} \quad \frac{\alpha_1}{\alpha_2} = \sqrt[N]{1} = e^{i\frac{2\pi m}{N}},$$

или, так как $\alpha_1 = \frac{1}{\alpha_2}$, $\alpha_1 = \frac{1}{\alpha_2} = e^{\frac{i\pi m}{N}}$. Отсюда

$$v_m(k) = C \sin \frac{k\pi m}{N}.$$

Система функций $v_m(k)$ $(m = 1, 2, \ldots, N-1)$ ортогональна и полна, т. е.

$$\sum_{k=1}^{N-1} \sin \frac{\pi k m}{N} \sin \frac{\pi k m_1}{N} = 0 \qquad (m \neq m_1),$$

и любая функция $\varphi(k)$, определенная на множестве точек $0, 1, 2, \ldots, N$, для которой $\varphi(0) = \varphi(N) = 0$ представляется в виде линейной комбинации этих функций.

Для ρ получим следующее равенство:

$$\rho = \alpha_1 + \alpha_2 - 2 = 2 \cos \frac{m\pi}{A} - 2 = -4 \sin^2 \frac{m\pi}{2N},$$

откуда

$$\lambda^2 - \left(2 - \frac{4l^2}{h^2} \sin^2 \frac{m\pi}{2N}\right)\lambda + 1 = 0.$$

Следовательно,

$$\lambda_1 = \frac{1}{\lambda_2} = 1 - 2r \sin^2 \frac{m\pi}{2N} + \sqrt{4r^2 \sin^4 \frac{m\pi}{2N} - 4r \sin^2 \frac{m\pi}{2N}} \quad \left(r = \frac{l^2}{h^2}\right).$$

Для устойчивости разностной схемы нужно потребовать, чтобы все λ_i были по модулю меньше или равны единице и среди них не должно быть кратных (рассуждения те же, что и в § 5, так как любое решение u_h представляется в виде линейной комбинации решений вида $\left(C_0 + C_1 j + \ldots + C_{p-1} j^{p-1}\right)\lambda^j v(k)$, которое соответствует корню характеристического уравнения λ кратности p). Так как $\lambda_1 \lambda_2 = 1$ и при всех $m = 1, 2, \ldots, N-1$ $\lambda_1 \neq \lambda_2$, то это возможно лишь в том случае, если

$$4r^2 \sin^4 \frac{m\pi}{2N} - 4r \sin^2 \frac{m\pi}{2N} < 0,$$

т. е. при $r \sin^2 \frac{m\pi}{2N} < 1$ или $r < \frac{1}{\sin^2 \frac{m\pi}{2N}}$ $(m = 1, 2, \ldots, N-1)$.

Итак, устойчивость будет иметь место, если при всех $m = 1, 2, \ldots, N-1$ будет выполнено неравенство

$$0 < \frac{l}{h} < \frac{1}{\sin \frac{m\pi}{2N}}$$

или при $\frac{l}{h} \leqslant 1$.

Из теоремы сходимости устойчивой разностной схемы будет следовать, что в нашем случае при $\frac{l}{h} \leqslant 1$ и $h \to 0$ последовательность u_h будет сходиться к точному решению граничной задачи для дифференциального уравнения.

Изложенные выше приемы исследования устойчивости разностных схем применимы в основном для разностных схем, аппроксимирующих дифференциальные уравнения с постоянными коэффициентами. Исследование устойчивости разностных схем для уравнений с переменными коэффициентами, как правило, — очень сложная задача. Если уравнение с непрерывными коэффициентами[1]) рассматривается в конечной области, то на практике применяют следующий принцип:

Заменяем в уравнении все переменные коэффициенты постоянными, полагая их равными значениям переменных коэффициентов в какой-либо точке P области. Если при любом выборе точки P полученная разностная схема с постоянными коэффициентами будет устойчива, то устойчива и разностная схема с переменными коэффициентами.

В случае нелинейного дифференциального уравнения задача исследования устойчивости разностной схемы еще больше усложняется. В этом случае исследуют устойчивость в окрестности искомого решения для линеаризованного уравнения.

5. Некоторые общие замечания. При численном решении краевых задач для дифференциальных уравнений в частных производных методом сеток могут быть использованы только сходящиеся разностные схемы, так как только в этом случае можно рассчитывать на получение приближенного решения, достаточно близкого к точному решению задачи. Но и сходящиеся разностные схемы не всегда могут быть использованы при практическом решении задачи, так как при применении метода сеток при вычислении значений граничных функций и правой части неизбежно возникают погрешности, и чтобы эти погрешности не исказили истинного решения разностной схемы, последняя должна быть устойчивой по граничным условиям и по правой части. При использовании неустойчивой разностной схемы искажение истинного решения тем сильней, чем мельче сетка, при использовании же крупной сетки мы не можем рассчитывать, что решение разностной схемы будет близко к точному решению краевой задачи для дифференциального уравнения в силу плохой разностной аппроксимации уравнения.

Далее, при решении разностной задачи в процессе счета нам неизбежно придется округлять значения решения в узлах сетки.

---

[1]) О разностных схемах для уравнений с коэффициентами, допускающими разрывы см. сноску на стр. 373, а также доклад А. А. Самарского в Трудах конференции по дифференциальным уравнениям в Ереване, ноябрь, 1958 г.

Наличие этих ошибок может также сильно исказить картину решения, поэтому необходимо требование устойчивости разностной схемы относительно ошибок, возникающих в результате округлений значений решения в узлах сетки. Так как ошибки округления значений решения в узлах сетки, по крайней мере, в простейших случаях можно компенсировать изменением правой части разностного уравнения, то особенно существенно требование устойчивости по правой части. Наконец, нужно иметь в виду, что мы всюду рассматривали устойчивость как свойство, связанное лишь с разностным уравнением и граничными условиями для него, совершенно не принимая во внимание алгоритм, используемый для решения разностной схемы. Однако даже в том случае, когда разностная схема устойчива по граничным условиям и по правой части, при неудачном выборе алгоритма для счета решения этой разностной схемы может произойти сильное накопление вычислительной погрешности, в этом случае уже неустойчивым будет сам процесс счета. Неустойчивые алгоритмы счета практически непригодны в случае мелкой сетки. На это явление мы уже указывали при изложении метода прогонки решения краевых задач. В книге В. С. Рябенького и А. Ф. Филиппова «Об устойчивости разностных уравнений» приведены пример разностной схемы и алгоритмы получения решения схемы, из которых одни являются устойчивыми, а другие неустойчивыми.

Вопросы устойчивости разностных схем и вычислительных алгоритмов приобретают особое значение при использовании современных быстродействующих вычислительных машин, и исследованию этих вопросов уже сейчас посвящено большое количество работ.

§ 8. Метод прямых решения граничных задач для дифференциальных уравнений в частных производных

1. Сущность метода прямых. Пусть в прямоугольной области $G\{\alpha < x < \beta;\ y_0 < y < y_0 + l\}$ (рис. 74) необходимо найти решение эллиптического дифференциального уравнения

$$a(x, y)\frac{\partial^2 u}{\partial x^2} + b(x, y)\frac{\partial^2 u}{\partial y^2} + c(x, y)\frac{\partial u}{\partial x} + d(x, y)\frac{\partial u}{\partial y} + e(x, y)u = f(x, y), \tag{1}$$

$$(a,\ b > 0 \text{ в } G + \Gamma),$$

удовлетворяющее граничным условиям:

$$\left.\begin{aligned} &u(x, y_0) = \varphi_0(x); \quad u(x, y_0 + l) = \varphi_1(x) \quad (\alpha \leqslant x \leqslant \beta), \\ &u(\alpha, y) = \psi_0(y); \quad\quad u(\beta, y) = \psi_1(y) \quad (y_0 < y < y_0 + l), \end{aligned}\right\} \tag{2}$$

где $\varphi_i(x)$, $\psi_i(y)$ $(i = 0,\ 1)$ — заданные функции.

Метод прямых приближенного решения этой задачи, предложенный М. Г. Слободянским, заключается в следующем. На отрезке

$[y_0, y_0+l]$ возьмем точки $y_k = y_0 + kh$ $(k=0, 1, 2, \ldots, n)$, $h = \frac{l}{n+1}$ и проведем прямые $y = y_k$. Предполагая существование достаточно гладкого решения $u(x, y)$ задачи (1) — (2), положим в уравнении (1)

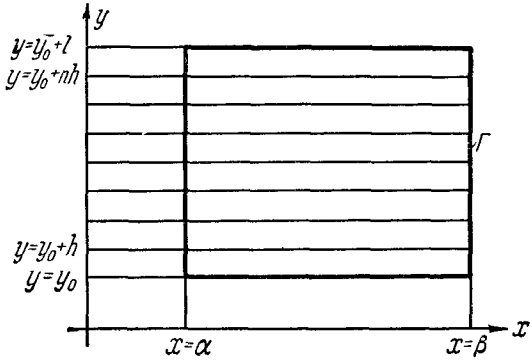

Рис. 74.

$y = y_k$ $(k=1, 2, \ldots, n)$ и заменим производные по y разностными отношениями, например, воспользовавшись равенствами

$$\left.\begin{aligned} \left.\frac{\partial u}{\partial y}\right|_{y=y_k} &= \frac{1}{2h}[u(x, y_{k+1}) - u(x, y_{k-1})] + O(h^2) = \\ &= \frac{1}{2h}[u_{k+1}(x) - u_{k-1}(x)] + O(h^2), \\ \left.\frac{\partial^2 u}{\partial y^2}\right|_{y=y_k} &= \frac{1}{h^2}[u(x, y_{k+1}) - 2u(x, y_k) + u(x, y_{k-1})] + O(h^2) = \\ &= \frac{1}{h^2}[u_{k+1}(x) - 2u_k(x) + u_{k-1}(x)] + O(h^2), \end{aligned}\right\} \quad (3)$$

где $u_k(x) = u(x, y_k)$. Получим следующую систему n обыкновенных линейных дифференциальных уравнений второго порядка:

$$a_k(x)u_k''(x) + \frac{b_k(x)}{h^2}[u_{k+1}(x) - 2u_k(x) + u_{k-1}(x)] + c_k(x)u_k'(x) + \\ + \frac{d_k(x)}{2h}[u_{k+1}(x) - u_{k-1}(x)] + e_k(x)u_k(x) = f_k(x) + O(h^2) \\ (k=1, 2, \ldots, n).$$

Пренебрегая в них членами $O(h^2)$ и обозначая через $U_k(x)$ приближенные значения решения $u(x, y)$ на прямой $y = y_k$ для их определения, получим систему уравнений

$$a_k(x)U_k''(x) + \frac{b_k(x)}{h^2}[U_{k+1}(x) - 2U_k(x) + U_{k-1}(x)] + c_k(x)U_k'(x) + \\ + \frac{d_k(x)}{2h}[U_{k+1}(x) - U_{k-1}(x)] + e_k(x)U_k(x) = f_k(x) \quad (4) \\ (k=1, 2, \ldots, n).$$

Используя граничные условия на Г, имеем:

$$\left.\begin{aligned} U_0(x) &= \varphi_0(x) \qquad (\alpha \leqslant x \leqslant \beta), \\ U_{n+1}(x) &= \varphi_1(k) \qquad (\alpha \leqslant x \leqslant \beta), \\ U_k(\alpha) &= \psi_0(y_k); \quad U_k(\beta) = \psi_1(y_k) \quad (k = 1, 2, \ldots, n). \end{aligned}\right\} \tag{5}$$

Система (4) обыкновенных дифференциальных уравнений с граничными условиями (5) аппроксимирует с точностью до h^2 дифференциальное уравнение (1) с граничным условием (2) и называется системой уравнений метода прямых.

Общее решение системы (4) будет линейно зависеть от $2n$ произвольных постоянных. Используя граничные условия (5) для отыскания этих постоянных, получим систему $2n$ линейных алгебраических уравнений, решив которую мы и найдем функции $U_k(x)$ $(k = 1, 2, \ldots, n)$, аппроксимирующие решение задачи (1) — (2) на прямых $y = y_k$ $(k = 1, 2, \ldots, n)$.

В зависимости от способа замены производных по y разностными отношениями мы будем иметь разные системы метода прямых, с различной точностью аппроксимирующие дифференциальное уравнение (1).

Этот метод удобнее всего применять в том случае, когда коэффициенты в уравнении (1) не зависят от x. В этом случае система (4) будет системой обыкновенных линейных дифференциальных уравнений с постоянными коэффициентами [1]).

Метод прямых можно рассматривать как предельный случай метода сеток, если, используя прямоугольную сетку, шаг сетки по оси x устремить к нулю.

2. Метод прямых решения задачи Дирихле для уравнения Пуассона. Пусть в области G, указанной в п. 1, требуется найти решение уравнения

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y) \tag{6}$$

с граничными условиями

$$\left.\begin{aligned} u(x, y_0) &= \varphi_0(x); \quad u(x, y_0 + l) = \varphi_1(x) \qquad (\alpha \leqslant x \leqslant \beta), \\ u(\alpha, y) &= \psi_0(x); \qquad u(\beta, y) = \psi_1(y) \qquad (y_0 < y < y_0 + l). \end{aligned}\right\} \tag{7}$$

---

[1]) По поводу областей применения метода прямых и построения различных схем метода прямых см. доклад акад. А. А. Дородницына в книге: Конференция «Пути развития советского математического машиностроения и приборостроения», пленарные заседания, Москва, 12—17 марта 1956 г., особенно стр. 47—50.

Применяя для решения задачи (6) — (7) метод прямых и заменяя производную $\left.\frac{\partial^2 u}{\partial y^2}\right|_{y=y_k}$ разностным отношением

$$\frac{1}{h^2}[u(x, y_{k+1}) - 2u(x, y_k) + u(x, y_{k-1})],$$

получим следующую систему уравнений метода прямых:

$$\left.\begin{aligned} U_k''(x) + \frac{1}{h^2}[U_{k+1}(x) - 2U_k(x) + U_{k-1}(x)] = f_k(x) \\ (k = 1, 2, \ldots, n), \\ U_0(x) = \varphi_0(x), \\ U_{n+1}(x) = \varphi_1(x) \end{aligned}\right\} \quad (8)$$

с граничными условиями

$$U_k(\alpha) = \psi_0(y_k); \quad U_k(\beta) = \psi_1(y_k) \qquad (k = 1, 2, \ldots, n), \quad (9)$$

аппроксимирующую уравнение (6) с точностью до h^2.

Рассмотрим более точную аппроксимацию уравнения (6), предполагая большую гладкость решения задачи (6) — (7). Для этого заметим, что из разложения функции $u(x, y)$ как функции переменного y в окрестности точки y_k по формуле Тейлора следует:

$$\begin{aligned} u(x, y_{k+1}) - 2u(x, y_k) + u(x, y_{k-1}) = \\ = u(x, y_k + h) - 2u(x, y_k) + u(x, y_k - h) = \\ = h^2 \frac{\partial^2 u(x, y_k)}{\partial y^2} + \frac{h^4}{12} \frac{\partial^4 u(x, y_k)}{\partial y^4} + O(h^6) \end{aligned}$$

Совершенно аналогично

$$\frac{\partial^2 u(x, y_{k+1})}{\partial y^2} - 2\frac{\partial^2 u(x, y_k)}{\partial y^2} + \frac{\partial^2 u(x, y_{k-1})}{\partial y^2} = h^2 \frac{\partial^4 u(x, y_k)}{\partial y^4} + O(h^4).$$

Исключая из этих двух равенств $\frac{\partial^4 u(x, y_k)}{\partial y^4}$, будем иметь:

$$\begin{aligned} u(x, y_{k+1}) - 2u(x, y_k) + u(x, y_{k-1}) = \\ = \frac{h^2}{12}\left[\frac{\partial^2 u(x, y_{k+1})}{\partial y^2} + \frac{\partial^2 u(x, y_{k-1})}{\partial y^2}\right] + \frac{5h^2}{6}\frac{\partial u(x, y_k)}{\partial y^2} + O(h^6). \end{aligned} \quad (10)$$

Учитывая, что из дифференциального уравнения (6)

$$\frac{\partial^2 u(x, y_k)}{\partial y^2} = f(x, y_k) - \frac{\partial^2 u(x, y_k)}{\partial x^2} = f_k(x) - u_k''(x),$$

и заменяя в (10) все производные $\frac{\partial^2 u}{\partial y^2}$, получим:

$$\begin{aligned} \frac{5}{6}u_k''(x) + \frac{1}{12}[u_{k+1}''(x) + u_{k-1}''(x)] + \frac{1}{h^2}[u_{k+1}(x) - 2u_k(x) + u_{k-1}(x)] = \\ = \frac{5}{6}f_k(x) + \frac{1}{12}[f_{k+1} + f_{k-1}] + O(h^4). \end{aligned}$$

Отбрасывая член с $O(h^4)$, получим следующую систему уравнений метода прямых:

$$\frac{5}{6}U_k''(x)+\frac{1}{12}[U_{k+1}''(x)+U_{k-1}''(x)]+\frac{1}{h^2}[U_{k+1}(x)-2U_k(x)+U_{k-1}(x)]=$$
$$=\frac{5}{6}f_k(x)+\frac{1}{12}[f_{k+1}(x)+f_{k-1}(x)] \qquad (k=1,\ 2,\ \ldots,\ n), \quad (11)$$
$$U_0(x)=\varphi_0(x); \quad U_{n+1}(x)=\varphi_1(x)$$

с граничными условиями

$$U_k(\alpha)=\psi_0(y_k), \quad U_k(\beta)=\psi_1(y_k) \qquad (k=1,\ 2,\ \ldots,\ n), \quad (12)$$

аппроксимирующую задачу (6) — (7) с точностью h^4.

Так как системы уравнений (8) и (11) линейны, то общее решение каждой из них равно сумме некоторого частного решения и общего решения соответствующей однородной системы, последнее не зависит от f, от граничных функций, а также и от размеров области G, если задано n, поэтому его можно найти раз и навсегда, что мы и сделаем сейчас.

Рассмотрим однородную систему уравнений, соответствующую системе (8):

$$U_k''(x)+\frac{1}{h^2}[U_{k+1}(x)-2U_k(x)+U_{k-1}(x)]=0 \quad (k=1,\ 2,\ \ldots,\ n),$$
$$U_0(x)=U_{n+1}(x)\equiv 0. \quad (8')$$

Будем искать частные решения этой системы вида

$$U_k(x)=\gamma(k)\,v(x). \quad (13)$$

Подстановка в (8′) дает

$$\gamma(k)\,v''(x)+\frac{1}{h^2}v(x)[\gamma(k+1)-2\gamma(k)+\gamma(k-1)]=0$$
$$(k=1,\ 2,\ \ldots,\ n),$$
$$\gamma(0)=\gamma(n+1)=0 \quad (14)$$

или

$$\frac{v''(x)}{v(x)}=\frac{\gamma(k+1)-2\gamma(k)+\gamma(k-1)}{-h^2\gamma(k)}=\delta^2=\text{const} \quad (k=1,\ 2,\ \ldots,\ n). \quad (15)$$

Для отыскания $\gamma(k)$ получим однородное разностное уравнение

$$\gamma(k+1)-[2-h^2\delta^2]\gamma(k)+\gamma(k-1)=0 \quad (16)$$

с граничными условиями

$$\gamma(0)=\gamma(n+1)=0. \quad (17)$$

Общее решение разностного уравнения (16) имеет вид

$$\gamma(k)=C_1\lambda_1^k+C_2\lambda_2^k, \quad (18)$$

где C_1 и C_2 — произвольные постоянные, а λ_1 и λ_2 — корни характеристического уравнения

$$\lambda^2 - [2 - h^2\delta^2]\lambda + 1 = 0.$$

Из граничных условий (17) имеем:

$$\gamma(0) = C_1 + C_2 = 0, \quad C_2 = -C_1,$$
$$\gamma(n+1) = C_1\lambda_1^{n+1} + C_2\lambda_2^{n+1} = C_1\left(\lambda_1^{n+1} - \lambda_2^{n+1}\right) = 0.$$

Отсюда

$$\left(\frac{\lambda_1}{\lambda_2}\right)^{n+1} = 1 \quad \text{или} \quad \frac{\lambda_1}{\lambda_2} = \sqrt[n+1]{1} = e^{\frac{2\pi is}{n+1}} \qquad (s = 0,\ 1,\ 2,\ \ldots,\ n).$$

Но так как $\lambda_1\lambda_2 = 1$, то $\lambda_1^2 = e^{\frac{2\pi is}{n+1}}$ и

$$\lambda_1 = \frac{1}{\lambda_2} = e^{\frac{\pi is}{n+1}}.$$

Зная λ_1 и λ_2, можно найти и неизвестную постоянную δ, ибо по свойству корней квадратного уравнения

$$2 - h^2\delta^2 = \lambda_1 + \lambda_2 = e^{\frac{\pi is}{n+1}} + e^{-\frac{\pi is}{n+1}} = 2\cos\frac{\pi s}{n+1},$$

откуда

$$\delta_s^2 = \frac{4}{h^2}\sin^2\frac{\pi s}{2(n+1)} = \frac{4}{h^2}\sin^2\frac{\pi(y_s - y_0)}{2l} \qquad (s = 0,\ 1,\ \ldots,\ n), \quad (18')$$

а

$$\gamma_s(k) = C_1\left(e^{\frac{\pi isk}{n+1}} - e^{-\frac{\pi isk}{n+1}}\right) = C\sin\frac{\pi sk}{n+1} = C\sin\frac{\pi s(y_k - y_0)}{l}. \qquad (19)$$

Нетривиальные решения будут только при $s = 1,\ 2,\ \ldots,\ n$. Из уравнения (15) имеем:

$$v''(x) - \delta_s^2 v(x) = 0$$

или

$$v_s(x) = A_s e^{\delta_s x} + B_s e^{-\delta_s x}. \qquad (20)$$

Итак, мы имеем n частных решений линейной однородной системы (8′):

$$U_{k,\,s}(x) = \left[A_s e^{\delta_s x} + B_s e^{-\delta_s x}\right]\sin\frac{\pi s(y_k - y_0)}{l} \qquad (s = 1,\ 2,\ \ldots,\ n), \quad (21)$$

которые между собой линейно независимы, а следовательно, общее решение этой системы имеет вид

$$U_k(x) = \sum_{s=1}^{n}\sin\frac{\pi s}{l}(y_k - y_0)\left(A_s e^{\delta_s x} + B_s e^{-\delta_s x}\right), \qquad (22)$$

где A_s и B_s — произвольные постоянные.

Совершенно аналогичными рассуждениями можно показать, что общее решение однородной системы, соответствующей системе (11), имеет вид

$$U_k(x) = \sum_{s=1}^{n} \sin\frac{\pi s}{l}(y_k - y_0)\left(A'_s e^{\delta'_s x} - B'_s e^{-\delta'_s x}\right), \tag{23}$$

где

$$\delta'^2_s = \frac{24\sin^2\dfrac{\pi(y_s - y_0)}{2l}}{h^2\left(5 + \cos\dfrac{\pi(y_s - y_0)}{l}\right)} \qquad (s = 1, 2, \ldots, n), \tag{24}$$

а A'_s и B'_s — произвольные постоянные.

Имея общее решение однородных систем, соответствующих системам (8) и (11), в каждом конкретном случае можно найти частное решение этих систем, например, методом вариации постоянных, а следовательно найти общее решение неоднородных систем, а затем, используя граничные условия для функций $U_k(x)$, получить систему линейных алгебраических уравнений для отыскания $2n$ произвольных постоянных A_s и B_s (или A'_s и B'_s), решив которую мы и найдем функции $U_k(x)$, являющиеся приближенными значениями решения $u(x, y)$ задачи (6) — (7) на прямых $y = y_k$.

В случае прямоугольной области сходимость решения задачи (8), (9) к решению задачи (6), (7) в предположении достаточной гладкости последнего и оценка погрешности метода могут быть получены с помощью принципа максимума для решения системы (8), (9); для доказательства сходимости решения задачи (11), (12) к решению задачи (6), (7) и вывода оценки погрешности метода этот подход уже неприменим [1]).

---

[1]) Укажем один подход, не опирающийся на принцип максимума, пригодный как в случае системы (8), (9), так и в случае системы (11), (12). Пусть требуется решить задачу Дирихле:

$$\Delta u = 0, \quad 0 < x < a, \quad 0 < y < b; \quad u|_{x=0} = \varphi(y), \quad u|_{x=a} = u|_{y=0} = u|_{y=b} = 0,$$

где $\varphi(y)$ — нечетная периодическая с периодом $2l$ функция, имеющая при $-\infty < y < +\infty$ непрерывные производные до $(p-1)$-го порядка и кусочно-непрерывную производную p-го порядка; $p \geqslant 1$. Вычитая из точного решения этой задачи

$$u(k, y) = \sum_{m=1}^{+\infty} C_m \frac{\operatorname{Sh}\lambda_m(a-x)}{\operatorname{Sh}\lambda_m a}\sin\lambda_m y, \quad \lambda_m = \frac{\pi m}{b},$$

полученного методом разделения переменных, точное решение задачи (8), (9), представленное в виде

$$u_k(x) = \sum_{m=0}^{+\infty} C_m \frac{\operatorname{Sh}\omega_m(a-x)}{\operatorname{Sh}\omega_m a} Sm\lambda_m y, \quad \omega_m = \frac{2}{h}\sin\frac{\lambda_m h}{2},$$

Если область G имеет вид криволинейной трапеции (рис. 75), ограниченной прямыми $y=y_0$, $y=y_0+l$ и кривыми $x=\alpha(y)$, $x=\beta(y)$ $(y_0\leqslant y\leqslant y_0+l)$, то описанная схема применения метода прямых не является вполне корректной. Оказывается, что существуют очень простые области G, для которых краевая задача для системы обыкновенных дифференциальных уравнений метода прямых при некоторых n будет неразрешима. В связи с этим может быть предложена[1]) другая схема метода прямых, свободная от указанного недостатка. Опишем эту схему на примере решения задачи Дирихле $u|_\Gamma=\varphi(x,\ y)|_\Gamma$ для уравнения Лапласа, если область имеет вид, изображенный на рис. 75.

Строится контур Γ_δ, составленный из линий $y=y_0+\delta$; $y=y_0+l-\delta$; $x=\alpha(y)+\delta$; $x=\beta(y)-\delta$, где $\delta>0$ — достаточно

получим:

$$\delta_k(x)=|u(x,\ y_k)-u_k(x)|=$$

$$=\left|\left(\sum_{m=1}^{m_0-1}+\sum_{m=m_0}^{+\infty}\right)C_m\left(\frac{\operatorname{Sh}\lambda_m(a-x)}{\operatorname{Sh}\lambda_m a}-\frac{\operatorname{Sh}\omega_m(a-x)}{\operatorname{Sh}\omega_m a}\right)\sin\lambda_m y_k\right|.$$

При любом m_0 в силу свойств $\varphi(y)$ будет

$$\left|\sum_{m=m_0}^{+\infty}\right|=O\left(\frac{1}{m_0^p}\right),$$

причем константа, входящая в $O\left(\frac{1}{m_0^p}\right)$, без труда выписывается явно.

Если m_0 уже выбрано, то, исследуя разности $\lambda_m-\omega_m$ и $\omega_m-\frac{1}{2}\lambda_m$ при $1\leqslant m\leqslant m_0$, нетрудно получить:

$$\left|\sum_{m=1}^{m_0-1}\right|=O\left(h^2m_0^3\operatorname{Ch}^2\frac{m_0\pi a}{b}\right).$$

Таким образом, для погрешности метода получаем оценку

$$\delta_k(x)=|u(x,\ y_k)-u_k(x)|=O\left(h^2m_0^3\operatorname{Ch}^2\frac{m_0\pi a}{b}\right)+O\left(\frac{1}{m_0^p}\right).$$

Аналогично в случае применения системы (11), (12) получаем:

$$\delta_k(x)=|u(x,\ y_k)-u_k(x)|=O\left(h^4m_0^3\operatorname{Ch}^2\frac{m_0\pi a}{b}\right)+O\left(\frac{1}{m_0^p}\right).$$

Таким же способом получается оценка погрешности метода при применении систем (8), (9) и (11), (12) к решению задачи Дирихле для уравнения Пуассона. (*Прим. ред.*)

[1]) Е. Х. Костюкович, О сходимости метода прямых ..., ДАН СССР, 118, № 3, 1958.

малое число. Рассмотрим систему прямых $y = y_k = y_0 + \delta + kh$ $\left(h = \frac{l - 2\delta}{n+1}\right)$. Абсциссы точек, а также и сами точки пересечения прямой $y = y_k$ с контуром Γ_δ обозначим через $x_{k,1}$, $x_{k,2}$. Решаем k-е уравнение системы (8), в которой положено $f_k(x) = 0$, только на общей части $[\alpha_k, \beta_k]$ отрезков $[x_{k-1,1}; x_{k-1,2}]$, $[x_{k1}; x_{k2}]$, $[x_{k+1,1}, x_{k+1,2}]$, считая, что h настолько мало, что общая часть

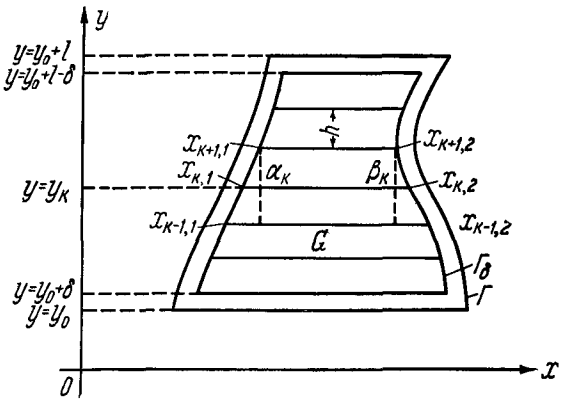

Рис. 75.

этих отрезков есть отрезок. Краевые условия для $U_k(x)$ зададим следующим образом:

$$\left.\begin{aligned} U_k(x) &\equiv \varphi(\alpha(y_k),\ y_k) \quad \text{при} \quad x_{k,1} \leqslant x \leqslant \alpha_k, \\ U_k(x) &\equiv \varphi(\beta(y_k),\ y_k) \quad \text{при} \quad \beta_k \leqslant x \leqslant x_{k,2}, \end{aligned}\right\} \tag{25}$$

где $\varphi(x, y)$ — заданная в задаче Дирихле граничная функция. Решение рассматриваемой таким образом системы (8) с граничными условиями (25) ищем методом последовательных приближений, принимая за первое приближение для $U_k(x)$ на отрезке $\alpha_k \leqslant x \leqslant \beta_k$:

$$U_k^{(1)}(x) = \frac{U_k(\alpha_k) - U_k(\beta_k)}{\alpha_k - \beta_k}(x - \alpha_k) + U_k(\alpha_k) \qquad (k = 1,\ 2,\ \ldots,\ n),$$

а следующие приближения при $m = 2, 3, 4, \ldots$ находим из системы уравнений

$$\left.\begin{aligned} U_k^{(m)''}(x) + \frac{1}{h^2}\left[U_{k-1}^{(m)}(x) - 2U_k^{(m)}(x) + U_{k+1}^{(m-1)}(x)\right] &= 0 \\ (k = 1,\ \ldots,\ n), \qquad & \\ U_0^{(m)} = \varphi(x,\ y_0) = \varphi_0(x); \quad U_{n+1}^{(m)} &= \varphi(x,\ y_0 + l) = \varphi_1(x). \end{aligned}\right\} \tag{26}$$

Обозначая через $U_k^{(h)}(x)$ значения $U_k(x)$ при данном h, можно показать, что при наличии у решения $u(x, y)$ непрерывных производных

по y до третьего порядка включительно в области G и при соответствующем выборе $h = h(\delta)$ ($\lim\limits_{0<\delta\to 0} h(\delta) = 0$) имеет место сходимость приближенного решения к точному решению задачи $u(x, y)$:

$$\lim_{\delta\to 0} \max_{0\leqslant k\leqslant \frac{l-2\delta}{h}} \max_{x_{k,1}\leqslant x\leqslant x_{k,2}} \left| u(x, y_k) - U_k^{(h)}(x) \right| = 0.$$

Если предположить, что $\frac{\partial u^3(x, y)}{\partial y^3}$ равномерно непрерывна в G, то этот процесс можно проводить не строя вспомогательного контура Γ_δ.

Эта схема обобщается на общие линейные эллиптические уравнения с переменными коэффициентами и на области более общего вида, чем изображенные на рис. 75.

Пример. Найти решение уравнения

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = -1$$

в квадрате $-0{,}5 \leqslant x,\ y \leqslant 0{,}5$, если граничные условия нулевые: $u|_{x=\pm 0{,}5} = u|_{y=\pm 0{,}5} = 0$.

Для решения задачи применим метод прямых с тремя промежуточными прямыми $y = 0$, $y = \pm 0{,}25$. Значения $U_k(x)$ решения на этих прямых будем находить используя систему уравнений метода прямых вида (8) с $n = 3$. Будем иметь систему

$$\begin{aligned} U_1''(x) + 16[U_2(x) - 2U_1(x)] &= -1, \\ U_2''(x) + 16[U_3(x) - 2U_2(x) + U_1(x)] &= -1; \qquad U_0(x) = U_4(x) = 0, \\ U_3''(x) + 16[U_2(x) - 2U_3(x)] &= -1 \end{aligned}$$

с краевыми условиями

$$U_i(-0{,}5) = U_i(+0{,}5) = 0 \qquad (i = 1,\ 2,\ 3).$$

Частное решение неоднородной системы ищем в виде $U_i = A_i = \text{const}$. Подстановка в систему дает

$$\begin{aligned} A_2 - 2A_1 &= -\frac{1}{16}, \\ A_1 - 2A_2 + A_3 &= -\frac{1}{16}, \\ A_2 - 2A_3 &= -\frac{1}{16}, \end{aligned}$$

откуда

$$A_1 = A_3 = \frac{3}{32}; \quad A_2 = \frac{1}{8}.$$

Используя выражение (22) для общего решения соответствующей однородной системы, общее решение системы можно записать в виде

$$U_1(x) = \sin 0{,}25\pi\,(C_1 e^{\delta_1 x} + D_1 e^{-\delta_1 x}) + \sin 0{,}5\pi\,(C_2 e^{\delta_2 x} + D_2 e^{-\delta_2 x}) +$$

$$+ \sin 0{,}75\pi\,(C_3 e^{\delta_3 x} + D_3 e^{-\delta_3 x}) + \frac{3}{32} = \frac{\sqrt{2}}{2}(C_1 e^{\delta_1 x} + D_1 e^{-\delta_1 x}) +$$

$$+ (C_2 e^{\delta_2 x} + D_2 e^{-\delta_2 x}) + \frac{\sqrt{2}}{2}(C_3 e^{\delta_3 x} + D_3 e^{-\delta_3 x}) + \frac{3}{32},$$

$$U_2(x) = \sin 0{,}5\pi\,(C_1 e^{\delta_1 x} + D_1 e^{-\delta_1 x}) +$$

$$+ \sin \pi\,(C_2 e^{\delta_2 x} + D_2 e^{-\delta_2 x}) + \sin 1{,}5\pi\,(C_3 e^{\delta_3 x} + D_3 e^{-\delta_3 x}) + \frac{1}{8} =$$

$$= (C_1 e^{\delta_1 x} + D_1 e^{-\delta_1 x}) - (C_3 e^{\delta_3 x} + D_3 e^{-\delta_3 x}) + \frac{1}{8},$$

$$U_3(x) = \sin 0{,}75\pi\,(C_1 e^{+\delta_1 x} + D_1 e^{-\delta_1 x}) + \sin 1{,}5\pi\,(C_2 e^{\delta_2 x} + D_2{}^{-\delta_2 x}) +$$

$$+ \sin 2{,}25\pi\,(C_3 e^{\delta_3 x} + D_3 e^{-\delta_3 x}) + \frac{3}{32} = \frac{\sqrt{2}}{2}(C_1 e^{\delta_1 x} + D_1 e^{-\delta_1 x}) -$$

$$- (C_2 e^{\delta_2 x} + D_2 e^{-\delta_2 x}) + \frac{\sqrt{2}}{2}(C_3 e^{\delta_3 x} + D_3 e^{-\delta_3 x}) + \frac{3}{32},$$

где по (18)

$$\delta_1^2 = 4 \cdot 16 \sin^2 \frac{\pi}{8} = 9{,}3726, \qquad \delta_1 = 3{,}0611,$$

$$\delta_2^2 = 4 \cdot 16 \sin^2 \frac{\pi}{4} = 32, \qquad \delta_2 = 5{,}6568,$$

$$\delta_3^2 = 4 \cdot 16 \sin^2 \frac{3\pi}{8} = 54{,}6274, \quad \delta_3 = 7{,}3910.$$

Удовлетворяя граничным условиям и учитывая симметрию, т. е. считая $C_i = D_i$, получим:

$$\sqrt{2}\,C_1 \operatorname{Ch} 0{,}5\delta_1 + 2C_2 \operatorname{Ch} 0{,}5\delta_2 + \sqrt{2}\,C_3 \operatorname{Ch} 0{,}5\delta_3 = -\frac{3}{32},$$

$$2C_1 \operatorname{Ch} 0{,}5\delta_1 - 2C_3 \operatorname{Ch} 0{,}5\delta_3 = -\frac{1}{8},$$

$$\sqrt{2}\,C_1 \operatorname{Ch} 0{,}5\delta_1 - 2C_2 \operatorname{Ch} 0{,}5\delta_2 + \sqrt{2}\,C_3 \operatorname{Ch} 0{,}5\delta_3 = -\frac{3}{32}.$$

Отсюда

$$C_1 = -0{,}0266, \quad C_2 = 0, \quad C_3 = -0{,}00009.$$

Таким образом,

$$U_1(x) = U_3(x) = 0{,}0938 - 0{,}0376 \operatorname{Ch} 3{,}0611x - 0{,}00013 \operatorname{Ch} 7{,}3910x,$$

$$U_2(x) = 0{,}1250 - 0{,}0532 \operatorname{Ch} 3{,}0611x + 0{,}00018 \operatorname{Ch} 7{,}3910x.$$

Приближенное значение решения в центре квадрата будет $U_2(0) = 0{,}0720$, а точное же значение решения в этой точке $u(0, 0) = 0{,}0736$.

3. Метод прямых решения смешанной задачи для уравнения колебаний струны. Рассмотрим сначала метод прямых приближенного решения простейшего уравнения колебаний струны

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = f(x, t) \tag{27}$$

в области $0 \leqslant x \leqslant l$; $0 \leqslant t < \infty$, при следующих начальных и граничных условиях

$$u|_{t=0} = \varphi_1(x); \quad \left.\frac{\partial u}{\partial t}\right|_{t=0} = \varphi_2(x) \qquad (0 \leqslant x \leqslant l);$$

$$u(0, t) = \psi_1(t); \quad u(l, t) = \psi_2(t) \qquad (0 \leqslant t < \infty). \tag{28}$$

Проведем систему параллельных прямых

$$x = x_k = kh \qquad \left(k = 0, 1, 2, \ldots, n+1; \quad h = \frac{l}{n+1}\right)$$

и обозначим через $u_k(x)$ значения точного решения $u(x, y)$ задачи (27) — (28) на прямой $x = x_k$, т. е. $u_k(x) = u(x_k, t)$. Если $\left.\frac{\partial^2 u}{\partial x^2}\right|_{x=x_k}$ заменить разностным отношением

$$\frac{1}{h^2}[u(x_{k+1}, t) - 2u(x_k, t) + u(x_{k-1}, t)],$$

то получим следующую систему уравнений метода прямых:

$$\left.\begin{aligned} U_k''(t) - \frac{1}{h^2}[U_{k+1}(t) - 2U_k(t) + U_{k-1}(t)] = f_k(t) \\ (k = 1, 2, \ldots, n), \\ U_0(t) = \psi_1(t); \quad U_{n+1}(t) = \psi_2(t) \end{aligned}\right\} \tag{29}$$

с начальными условиями

$$\left.\begin{aligned} U_k(0) = \varphi_1(x_k) = \varphi_{1,k}, \\ U_k'(0) = \varphi_2(x_k) = \varphi_{2,k} \end{aligned} \qquad (k = 1, 2, \ldots, n),\right\} \tag{30}$$

аппроксимирующую уравнение (27) с точностью до h^2.

Чтобы получить систему уравнений метода прямых, более точно аппроксимирующую уравнение (27), воспользуемся равенством, аналогичным равенству (10):

$$u(x_{k+1}, t) - 2u(x_k, t) + u(x_{k-1}, t) =$$

$$= \frac{5h^2}{6}\frac{\partial^2 u(x_k, t)}{\partial x^2} + \frac{h^2}{12}\left[\frac{\partial^2 u(x_{k+1}, t)}{\partial x^2} + \frac{\partial^2 u(x_{k-1}, t)}{\partial x^2}\right] + O(h^6). \tag{10'}$$

Из дифференциального уравнения (27)

$$\frac{\partial^2 u(x_k, t)}{\partial x^2} = \frac{\partial^2 u(x_k, t)}{\partial t^2} - f(x_k, t) = u_k''(t) - f_k(t).$$

Тогда соотношение (10′) после подстановки $\frac{\partial^2 u(x_k, t)}{\partial x^2}$, $\frac{\partial^2 u(x_{k+1}, t)}{\partial x^2}$, $\frac{\partial^2 u(x_{k-1}, t)}{\partial x^2}$ дает

$$\frac{5}{6} u_k''(t) + \frac{1}{12}[u_{k+1}''(t) + u_{k-1}''(t)] - \frac{1}{h^2}[u_{k+1}(t) - 2u_k(t) + u_{k-1}(t)] = \\ = \frac{5}{6} f_k(t) + \frac{1}{12}[f_{k+1}(t) + f_{k-1}(t)] + O(h^4).$$

Отбрасывая член $O(h^4)$ и заменяя при этом $u_k(t)$ на $U_k(t)$, получим следующую систему уравнений метода прямых:

$$\left.\begin{array}{l} \frac{5}{6} U_k''(t) + \frac{1}{12}[U_{k+1}''(t) + U_{k-1}''(t)] - \\ \quad - \frac{1}{h^2}[U_{k+1}(t) - 2U_k(t) + U_{k-1}(t)] = \\ = \frac{5}{6} f_k(t) + \frac{1}{12}[f_{k+1}(t) + f_{k-1}(t)] \qquad (k = 1, 2, \ldots, n), \\ U_0(t) = \psi_1(t); \quad U_{n+1}(t) = \psi_2(t) \end{array}\right\} \quad (31)$$

с начальными условиями

$$U_k(0) = \varphi_1(x_k) = \varphi_{1k}; \quad U_k'(0) = \varphi_2(x_k) = \varphi_{2,k} \quad (k = 1, 2, \ldots, n). \quad (32)$$

Эта система уже дает аппроксимацию порядка h^4.

Здесь, как и в п. 2, легко построить общие решения однородных систем, соответствующих системам дифференциальных уравнений метода прямых (29) и (31).

Построим для примера общее решение системы

$$\left.\begin{array}{l} \frac{5}{6} U_k''(t) + \frac{1}{12}[U_{k+1}''(t) + U_{k-1}''(t)] - \\ - \frac{1}{h^2}[U_{k+1}(t) - 2U_k(t) + U_{k-1}(t)] = 0 \qquad (k = 1, 2, \ldots, n) \\ U_0(t) = U_{n+1}(t) \equiv 0, \end{array}\right\} \quad (33)$$

соответствующей системе (31). Частные решения этой системы будем искать в виде

$$U_k(t) = \gamma(k)\, v(t).$$

Подставляя в систему (33), получим:

$$v''(t)\left[\frac{5}{6}\gamma(k) + \frac{1}{12}\gamma(k+1) + \frac{1}{12}\gamma(k-1)\right] - \\ - \frac{v(t)}{h^2}[\gamma(k+1) - 2\gamma(k) + \gamma(k-1)] = 0 \qquad (k = 1, 2, \ldots, n), \\ \gamma(0) = \gamma(n+1) = 0$$

или

$$\frac{v''(t)}{v(t)} = \frac{\gamma(k+1) - 2\gamma(k) + \gamma(k-1)}{h^2\left[\frac{5}{6}\gamma(k) + \frac{1}{12}\gamma(k+1) + \frac{1}{12}\gamma(k-1)\right]} = -\delta^2 = \text{const}. \quad (34)$$

Для отыскания $\gamma(k)$ получаем разностное уравнение

$$\left[1 + \frac{1}{12}\delta^2 h^2\right]\gamma(k+1) - \left[2 - \frac{5}{6}\delta^2 h^2\right]\gamma(k) + \left[1 + \frac{1}{12}\delta^2 h^2\right]\gamma(k-1) = 0 \quad (35)$$

с граничными условиями

$$\gamma(0) = \gamma(n+1) = 0. \quad (36)$$

Его общее решение имеет вид

$$\gamma(k) = C_1\lambda_1^k + C_2\lambda_2^k,$$

где λ_1 и λ_2 — корни уравнения

$$\lambda^2 - \frac{2[12 - 5\delta^2 h^2]}{12 + \delta^2 h^2}\lambda + 1 = 0.$$

Из граничных условий имеем:

$$\gamma(0) = C_1 + C_2 = 0; \quad C_2 = -C_1;$$

$$\gamma(n+1) = C_1\lambda_1^{n+1} + C_2\lambda_2^{n+1} = C_1\left(\lambda_1^{n+1} - \lambda_2^{n+1}\right) = 0.$$

Таким образом,

$$\frac{\lambda_1}{\lambda_2} = \sqrt[n+1]{1} = e^{\frac{2\pi i s}{n+1}} \qquad (s = 0,\ 1,\ 2,\ \ldots,\ n),$$

или так как $\lambda_1\lambda_2 = 1$, то

$$\lambda_1 = \frac{1}{\lambda_2} = e^{\frac{\pi i s}{n+1}}.$$

Далее,

$$\frac{2[12 - 5\delta^2 h^2]}{12 + \delta^2 h^2} = \lambda_1 + \lambda_2 = 2\cos\frac{\pi s}{n+1},$$

откуда

$$\delta_s^2 = \frac{24\sin^2\frac{\pi s}{2(n+1)}}{h^2\left[5 + \cos\frac{\pi s}{n+1}\right]} = \frac{24\sin^2\frac{\pi(y_s - y_0)}{2l}}{h^2\left(5 + \cos\frac{\pi(y_s - y_0)}{l}\right)}$$

$$(s = 1,\ 2,\ \ldots,\ n), \quad (37)$$

а

$$\gamma_s(k) = C_1\left(e^{\frac{\pi i s k}{n+1}} - e^{-\frac{\pi i s k}{n+1}}\right) = C\sin\frac{\pi s k}{n+1} = C\sin\frac{\pi s x_k}{l}$$

$$(s = 1,\ 2,\ \ldots,\ n) \quad (38)$$

(при $s = 0$ получаем тривиальное решение $\gamma_0(k) \equiv 0$).

Из уравнения (34)

$$v''(t) + \delta_s^2 v(t) = 0,$$

или

$$v_s(t) = A_s \cos \delta_s t + B_s \sin \delta_s t.$$

Таким образом,

$$U_{k,\,s}(t) = \sin \frac{\pi s x_k}{l} (A_s \cos \delta_s t + B_s \sin \delta_s t).$$

Общее решение однородной системы (33), следовательно, имеет вид

$$U_k(t) = \sum_{s=1}^{n} \sin \frac{\pi s x_k}{l} (A_s \cos \delta_s t + B_s \sin \delta_s t), \tag{39}$$

где A_s, B_s — произвольные постоянные. Найдя методом вариации постоянных частное решение неоднородной системы (31), получим общее решение ее как сумму частного решения и построенного общего решения (39) однородной системы. Постоянные A_s и B_s $(s = 1, 2, \ldots, n)$ найдутся из условий (32).

Однородная система, соответствующая системе (29), имеет общее решение

$$U_k(t) = \sum_{s=1}^{n} \sin \frac{\pi s x_k}{l} \left(C_s \cos \delta_s' t + D_s \sin \delta_s' t\right) \quad (k = 1, 2, \ldots, n), \tag{40}$$

где

$$\delta_s'^2 = \frac{4 \sin^2 \frac{\pi}{2l} x_s}{h^2}. \tag{41}$$

Сходимость решений, полученных методом прямых, к обобщенному решению задачи (27) — (28) имеет место [1]) в любом прямоугольнике $0 \leqslant x \leqslant l$; $0 \leqslant t \leqslant T$, если начальные и граничные условия нулевые, а для функции $f(x, t)$ имеют место неравенства

$$\int_0^T \sqrt{\int_0^l f^2\,dx}\,dt < C; \qquad \left[\int_0^l f^2\,dx\right]_{t=0} < C;$$

$$\int_0^T \sqrt{\int_0^l \left(\frac{\partial f}{\partial x}\right)^2 dx}\,dt < C$$

с некоторой положительной константой C. Общий случай начальных и граничных условий сводится к этому случаю при выполнении

[1]) В. И. Лебедев, Уравнения и сходимость дифференциально-разностного метода, Вестник МГУ, № 10, 1955, стр. 45—47.

некоторых требований на гладкость функций φ_1, φ_2, ψ_1, ψ_2 и условий сопряжения.

Остановимся теперь на решении методом прямых задачи о колебаниях неоднородной струны

$$\rho(x)\frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x}\left(\sigma(x)\frac{\partial u}{\partial x}\right) - q(x)u + f(x, t);$$

$$\rho(x) \geqslant \rho_{\min} > 0, \quad \sigma(x) \geqslant \sigma_{\min} > 0, \quad 0 \leqslant x \leqslant l, \quad 0 < t < T, \tag{42}$$

$$u(0, t) = \psi_1(t), \qquad u(l, t) = \psi_2(t), \qquad 0 < t < T, \tag{43}$$

$$u(x, 0) = \varphi_1(x), \quad u_t(x, 0) = \varphi_2(x), \qquad 0 < x < l. \tag{44}$$

Приближенные значения решения этой задачи на прямых $x = x_k = kh$, $k = 0, 1, \ldots, n+1$; $(n+1)h = l$, можно получить, как решение системы обыкновенных дифференциальных уравнений с постоянными коэффициентами

$$\rho_k u_k''(t) = \frac{\sigma_k[u_{k-1}(t) - u_k(t)] - \sigma_{k-1}[u_k(t) - u_{k-1}(t)]}{h^2} - q_k u_k(t) + f_k(t), \tag{45}$$

$$\left.\begin{gathered} \rho_k = \rho(x_k), \quad \sigma_k = \sigma(x_k), \quad q_k = q(x_k), \quad f_k(t) = f(x_k, t), \\ u_k(0) = \varphi_1(x_k), \qquad u_k'(0) = \varphi_2(x_k), \\ u_0(t) = \psi_1(t), \qquad u_{n+1}(t) = \psi_2(t). \end{gathered}\right\} \tag{45'}$$

Может быть предложен следующий способ доказательства сходимости и оценки погрешности метода [1]).

Положим $\gamma_k(t) = u(x_k, t) - u_k(t)$. Для $\gamma_k(t)$ получаем, применяя разложения по формуле Тейлора, систему

$$\rho_k \gamma_k'' = \frac{\sigma_k[\gamma_{k+1} - \gamma_k] - \sigma_{k-1}[\gamma_k - \gamma_{k-1}]}{h^2} - q_k\gamma_k + h\frac{1}{2}R_k(t), \tag{46}$$

$$\gamma_0(t) = \gamma_{n+1}(t) = 0, \quad \gamma_k(0) = \gamma_k'(0) = 0 \quad (k=0, 1, 2, \ldots, n+1), \tag{47}$$

где

$$|R_k(t)| \leqslant M = M_{\sigma''}M_{u_x} + M_{\sigma'}M_{u_{xx}} + \frac{2}{3}M_\sigma M_{u_{xxx}}, \tag{48}$$

причем M_σ, $M_{\sigma'}$, $M_{\sigma''}$, M_{u_x}, $M_{u_{xx}}$, $M_{u_{xxx}}$ — максимумы модулей величин, обозначенных в индексе, при $0 \leqslant x \leqslant l$, $0 \leqslant t \leqslant T$. Положим

$$I(t) = \sum_{k=1}^{n} \rho_k \gamma_k'^2 + \sum_{k=0}^{n} \sigma_k\left(\frac{\gamma_{k+1} - \gamma_k}{h}\right)^2 + \sum_{k=1}^{n} q_k\gamma_k^2. \tag{49}$$

---

[1]) Б. М. Будак, О методе прямых для некоторых краевых задач, Вестник МГУ, № 1, 1956, стр. 3—11.

В силу (46), (47), (48) и неравенства Коши — Буняковского, дифференцируя (49), получаем:

$$\frac{dI}{dt} = 2\sum_{k=1}^{n} \gamma_k' (\rho_k \gamma_k'' + q_k \gamma_k) + 2\sum_{k=0}^{n} \sigma_k \frac{(\gamma_{k+1} - \gamma_k)(\gamma_{k+1}' - \gamma_k')}{h^2} =$$

$$= 2\sum_{k=1}^{n} \frac{\sigma_k(\gamma_{k+1} - \gamma_k) - \sigma_{k-1}(\gamma_k - \gamma_{k-1})}{h^2} \gamma_k' +$$

$$+ 2\sum_{k=0}^{h} \sigma_k \frac{(\gamma_{k+1} - \gamma_k)(\gamma_{k+1}' - \gamma_k')}{h^2} + \sum_{k=1}^{n} \gamma_k' h R_k =$$

$$= 2\sum_{k=1}^{n} \frac{\sigma_k \gamma_k' (\gamma_{k+1} - \gamma_k) - \sigma_{k-1}\gamma_k' (\gamma_k - \gamma_{k-1})}{h^2} +$$

$$+ 2\sum_{k=0}^{n} \frac{\sigma_k \gamma_{k+1}' (\gamma_{k+1} - \gamma_k) - \sigma_k \gamma_k' (\gamma_{k+1} - \gamma_k)}{h^2} +$$

$$+ \sum_{k=1}^{n} \gamma_k' R_k h = \frac{2}{h^2} \{\sigma_n \gamma_{n+1}' [\gamma_{n+1} - \gamma_n] - \sigma_0 \gamma_0' [\gamma_1 - \gamma_0]\} +$$

$$+ \sum_{k=1}^{n} \gamma_k' R_k h = \sum_{k=1}^{n} \gamma_k' R_k h \leqslant \sqrt{\sum_{k=1}^{n} \gamma_k'^2} \sqrt{\sum_{k=1}^{n} R_k^2 h^2} \leqslant$$

$$\leqslant \frac{1}{\sqrt{\rho_{\min}}} \sqrt{I} \sqrt{\sum_{k=1}^{n} R_k^2 h^2} \leqslant 2\sqrt{I} \frac{M\sqrt{lh}}{2\sqrt{\rho_{\min}}}. \qquad (50)$$

Так как $I(0) = 0$, то, следовательно,

$$I(t) \leqslant \frac{M^2 l t^2}{4\rho_{\min}} h.$$

Далее, вследствие $\gamma_0(t) = \gamma_{n+1}(t) = 0$ будет

$$\gamma_k = \sum_{m=1}^{k} h \frac{\gamma_m - \gamma_{m-1}}{h}, \quad \gamma_k = -\sum_{m=k+1}^{n+1} h \frac{\gamma_m - \gamma_{m-1}}{h}.$$

Следовательно, в силу неравенства Коши — Буняковского

$$\gamma_k^2 \leqslant kh^2 \sum_{m=1}^{k} \left(\frac{\gamma_m - \gamma_{m-1}}{h}\right)^2; \quad \gamma_k^2 \leqslant (n+1-k)h^2 \sum_{m=k+1}^{n+1} \left(\frac{\gamma_m - \gamma_{m-1}}{h}\right)^2,$$

откуда

$$\gamma_k^2 \leqslant \frac{k(n+1-k)}{n+1} h^2 \sum_{m=1}^{n+1} \left(\frac{\gamma_m - \gamma_{m-1}}{h}\right)^2 \leqslant h \frac{x_k(l-x_k)}{l} \frac{I(t)}{\sigma_{\min}}. \qquad (51)$$

Сопоставляя (50) и (51), получим оценку

$$|\gamma_k(t)| \leqslant \frac{Mt\sqrt{x_k(l-x_k)}}{2\sqrt{\rho_{\min}\sigma_{\min}}} h. \qquad (52)$$

Здесь константа M выражается через константы $M_0, \ldots, M_{u_{xxx}}$. В цитированной на стр. 551 заметке дается также выражение M через коэффициенты уравнения, свободный член $f(x, t)$ и начальные и граничные функции φ_1, φ_2, ψ_1, ψ_2.

4. Метод прямых решения смешанной задачи для уравнения теплопроводности. Рассмотрим сначала метод прямых решений смешанной задачи для простейшего уравнения теплопроводности

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = f(x, t), \qquad (53)$$

$$\left.\begin{aligned} u|_{t=0} &= \varphi(x) \qquad (0 \leqslant x \leqslant l); \quad u(0, t) = \psi_1(t); \\ u(l, t) &= \psi_2(t) \qquad (0 \leqslant t < \infty), \end{aligned}\right\} \qquad (54)$$

заменяя разностным отношением производную по x. При наборе прямых

$$x = x_k = kh \qquad \left(k = 0, 1, 2, \ldots, n+1; \quad h = \frac{l}{n+1}\right),$$

используя приближенное равенство

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_{x=x_k} \approx \frac{u(x_{k+1}, t) - 2u(x_k, t) + u(x_{k-1}, t)}{h^2}$$

или равенство (10′) с заменой в нем производных $\frac{\partial^2 u}{\partial x^2}$, из дифференциального уравнения получим две системы уравнений метода прямых:

$$\left.\begin{aligned} U_k'(t) - \frac{1}{h^2}[U_{k+1}(t) - 2U_k(t) + U_{k-1}(t)] &= f_k(t) \\ (k = 1, 2, \ldots, n), \\ U_0(t) = \psi_1(t), \quad U_{n+1}(t) &= \psi_2(t) \end{aligned}\right\} \qquad (55)$$

и

$$\left.\begin{aligned}
&\frac{5}{6}U_k'(t)+\frac{1}{12}[U_{k+1}'(t)+U_{k-1}'(t)]- \\
&\qquad -\frac{1}{h^2}[U_{k+1}(t)-2U_k(t)+U_{k-1}(t)]= \\
&=\frac{5}{6}f_k+\frac{1}{12}[f_{k-1}(t)+f_{k+1}(t)] \qquad (k=1,\ 2,\ \ldots,\ n), \\
&U_0(t)=\psi_1(t), \quad U_{n+1}(t)=\psi_2(t),
\end{aligned}\right\} \tag{56}$$

из которых первая дает аппроксимацию уравнения с точностью h^2, а вторая с точностью h^4.

Из начального условия для $u(x, t)$ получаем начальные условия для $U_k(t)$, одинаковые в обоих случаях:

$$U_k(0)=\varphi(x_k)=\varphi_k \qquad (k=1,\ 2,\ \ldots,\ n). \tag{57}$$

Построение общих решений однородных систем уравнений, соответствующих системам (55) и (56), проводится точно так же, как и в случае уравнения колебания струны, так как, отыскивая частные решения однородных систем вида

$$U_k(t)=\gamma(k)\,v(t),$$

мы получим для $\gamma(k)$ в точности те же самые разностные уравнения, что и в п. 3, с граничными условиями $\gamma(0)=\gamma(n+1)=0$, а для отыскания $v(t)$ получим уравнение

$$\frac{v'(t)}{v(t)}=-\sigma_s^2, \tag{58}$$

т. е.

$$v_s(t)=C_s e^{-\sigma_s^2 t}. \tag{59}$$

Следовательно, общее решение однородной системы, соответствующей системе (55), будет иметь вид

$$U_k(t)=\sum_{s=1}^{n} C_s \sin\frac{\pi s x_k}{l}\, e^{-\sigma_s^2 t} \qquad (k=1,\ 2,\ \ldots,\ n), \tag{60}$$

где

$$\sigma_s^2=\frac{4\sin^2\dfrac{\pi x_s}{2l}}{h^2} \qquad (s=1,\ 2,\ \ldots,\ n), \tag{61}$$

а общее решение однородной системы, соответствующей системе (56), имеет вид

$$U_k(t)=\sum_{s=1}^{n} D_s \sin\frac{\pi s x_k}{l}\, e^{-\sigma_s'^2 t} \qquad (k=1,\ 2,\ \ldots,\ n), \tag{62}$$

где

$$\sigma_s'^2 = \frac{24 \sin^2 \frac{\pi x_s}{2l}}{h^2\left[5 + \cos \frac{\pi x_s}{l}\right]} \qquad (s = 1, 2, \ldots, n). \tag{63}$$

Далее, при заданных f, ψ_1, ψ_2 находятся, например, методом вариации постоянных общие решения систем (55) и (56), а неизвестные постоянные $C_s(D_s)$ находятся из начальных условий (57).

В книге А. Н. Тихонова и А. А. Самарского «Уравнения математической физики» 1953 г. сходимость решения задачи (55), (57) к решению задачи (53), (54) и оценка погрешности при соответствующей гладкости свободного члена $f(x, t)$, начальной и граничных функций доказывается с помощью принципа максимума для решения системы (55), (57).

В цитированной на стр. 550 заметке сходимость решений задач (55), (57) и (56), (57) к решению задачи (53), (54) в любом прямоугольнике $0 \leqslant x \leqslant l$; $0 \leqslant t \leqslant T$ доказана на основе применения теорем вложения при условии, что начальные и граничные условия нулевые, а правая часть $f(x, t)$ удовлетворяет условиям

$$\int_0^T \int_0^l f^2\, dx\, dt < C, \quad \left[\int_0^l \left(\frac{\partial f}{\partial x}\right)^2 dx\right]_{t=0} < C, \quad \int_0^T \int_0^l \left(\frac{\partial f}{\partial t}\right)^2 dx\, dt < C,$$

где C — некоторая положительная константа.

Рассмотрим теперь применение метода прямых с заменой производных по x разностными отношениями к уравнению теплопроводности с переменными коэффициентами.

Приближенные значения решения краевой задачи

$$\left.\begin{gathered} \rho(x)\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(\sigma(x)\frac{\partial u}{\partial x}\right) - q(x)u + f(x, t); \\ \rho \geqslant \rho_{\min} > 0, \quad \sigma \geqslant \sigma_{\min} > 0 \qquad (0 \leqslant x \leqslant l,\ 0 \leqslant t \leqslant T), \end{gathered}\right\} \tag{64}$$

$$u(0, t) = \psi_1(t), \quad u(l, t) = \psi_2(t) \qquad (0 \leqslant t \leqslant T), \tag{65}$$

$$u(x, 0) = \varphi(x) \qquad (0 \leqslant x \leqslant l) \tag{66}$$

на прямых $x = x_k = kh$, $(n+1)h = l$ $(k = 0, 1, \ldots, n+1)$ можно получить, решая систему обыкновенных дифференциальных уравнений с постоянными коэффициентами:

$$\rho_k u_k'(t) = \frac{\sigma_k [u_{k+1}(t) - u_k(t)] - \sigma_{k-1}[u_k(t) - u_{k-1}(t)]}{h^2} - q_k u_k(t) + f_k(t), \tag{67}$$

$$\rho_k = \rho(x_k), \quad \sigma_k = \sigma(x_k), \quad q_k = q(x_k), \quad f_k(t) = f(x_k, t) \tag{68}$$

$$u_k(0) = \varphi(x_k), \quad u_0(t) = \psi_1(t), \quad u_{n+1}(t) = \psi_2(t). \tag{69}$$

В цитированной на стр. 551 статье предложен следующий способ доказательства сходимости и оценки погрешности метода. Положим $\gamma_k(t) = u(x_k, t) - u_k(t)$, где $u(x_k, t)$ — значение точного решения задачи (64) — (66) на прямой $x = x_k = kh$. Для величин $\gamma_k(t)$ получаем систему

$$\rho_k \gamma_k' = \frac{\sigma_k[\gamma_{k+1} - \gamma_k] - \sigma_{k-1}[\gamma_k - \gamma_{k-1}]}{h^2} - q_k u_k(t) + R_k(t), \quad (70)$$

$$\gamma_0(t) \equiv \gamma_{n+1}(t) \equiv 0, \quad \gamma_k(0) = 0 \qquad (k = 0, 1, 2, \ldots, n+1), \quad (71)$$

$$|R_k| \leqslant M = M_{\sigma''} M_{u_x} + M_{\sigma'} M_{u_{xx}} + \frac{2}{3} M_\sigma M_{u_{xxx}}, \quad (72)$$

где константы $M_\sigma, \ldots, M_{u_{xxx}}$ определяются как в предыдущем параграфе. Рассмотрим величину

$$I_*(t) = 2 \sum_{k=1}^{n} \int_0^t \rho_k \gamma_k'^2 \, dt + \sum_{k=0}^{n} \sigma_k \left(\frac{\gamma_{k+1} - \gamma_k}{h} \right)^2 + \sum_{k=1}^{n} q_k \gamma_k^2. \quad (73)$$

Дифференцируя (73), мы, в силу (70), (71), (72) и неравенства Коши — Буняковского, аналогично тому, как это сделано в соотношений (50), получим:

$$\frac{dI_*}{dt} = h \sum_{k=1}^{n} \gamma_k' R_k. \quad (73')$$

В силу неравенства Коши — Буняковского и (73′) имеем:

$$I_*^2 \leqslant \left(\int_0^t \frac{dI_*}{dt} \, dt \right)^2 \leqslant t \int_0^t \left(\frac{dI_*}{dt} \right)^2 dt = th^2 \int_0^t \left(\sum_{k=1}^{n} \gamma_k' R_k \right)^2 dt \leqslant$$

$$\leqslant th \int_0^t \left(\sum_{k=1}^{n} \gamma_k'^2 \right) \left(\sum_{k=1}^{n} R_k^2 h \right) dt \leqslant \frac{htM^2 l}{2\rho_{\min}} 2 \int_0^t \sum_{k=1}^{n} \rho_k \gamma_k'^2 \, dt \leqslant \frac{htM^2 l}{2\rho_{\min}} I_*. \quad (73'')$$

Следовательно,

$$I_*(t) \leqslant \frac{M^2 lt}{2\rho_{\min}} h. \quad (74)$$

Для величины γ_k имеет место неравенство

$$\gamma_k^2 \leqslant h \frac{x_k(l - x_k)}{\sigma_{\min} l} I_*(t), \quad (51^*)$$

получающееся аналогично неравенству (51). Из (51*) и (74) вытекает оценка

$$|\gamma_k(t)| \leqslant h \frac{M \sqrt{x_k(l - x_k) t}}{2 \sqrt{\rho_{\min} \sigma_{\min}}} \quad (k = 0, 1, \ldots, n+1; \ 0 \leqslant t \leqslant T). \quad (75)$$

Все это делается совершенно аналогично тому, как в предыдущем параграфе при оценке $I(t)$ и $\gamma_k(t)$. Заметим, что константа M, как и ранее, может быть выражена через известные функции.

В случае замены производной по t разностным отношением, т. е. при выборе семейства прямых $t=t_k=kh$ $(k=0,\ 1,\ 2,\ \ldots)$, можно также построить системы уравнений метода прямых, аппроксимирующие с разной точностью дифференциальное уравнение (53) [1]). Приведем две системы [2]).

Заменим $\left.\frac{\partial u}{\partial t}\right|_{t=t_k}$ разностным отношением $\frac{u(x,\ t_{k+1})-u(x,\ t_{k-1})}{2h}$. Тогда получим систему уравнений для отыскания приближенных значений $U_k(x)$ решения $u(x,\ t)$ на прямых $t=t_k$ вида

$$\left.\begin{aligned} U_k''(x)-\frac{1}{2h}\left[U_{k+1}(x)-U_{k-1}(x)\right]&=f_k(x),\quad k\geqslant 1,\\ U_0(x)&=\varphi(x)\end{aligned}\right\} \tag{76}$$

с граничными условиями

$$U_k(0)=\psi_1(t_k)=\psi_{1,\,k};\quad U_k(l)=\psi_2(t_k)=\psi_{2,\,k}\quad (k=1,\ 2,\ 3,\ \ldots). \tag{77}$$

Эта система аппроксимирует уравнение (53) с точностью h^2.

Систему уравнений, аппроксимирующую уравнение (53) более точно, можно получить следующим образом. Предполагая, что решение $u(x,\ t)$ задачи (53), (54) достаточно гладко, запишем следующие разложения по формуле Тейлора:

$$u(x,\ t_k+h)=u(x,\ t_k)+hu_t'\Big|_{t=t_k}+\frac{h^2}{2}u_{t^2}''\Big|_{t=t_k}+\frac{h^3}{6}u_{t^3}'''\Big|_{t=t_k}+O(h^4),$$

$$u(x,\ t_k-h)=u(x,\ t_k)-hu_t'\Big|_{t=t_k}+\frac{h^2}{2}u_{t^2}''\Big|_{t=t_k}-\frac{h^3}{6}u_{t^3}'''\Big|_{t=t_k}+O(h^4),$$

$$u_t'(x,\ t_k+h)=u_t'(x,\ t_k)+hu_{t^2}''\Big|_{t=t_k}+\frac{h^2}{2}u_{t^3}'''\Big|_{t=t_k}+O(h^3),$$

$$u_t'(x,\ t_k-h)=u_t'(x,\ t_k)-u_{t^2}''\Big|_{t=t_k}+\frac{h^2}{2}u_{t^3}'''\Big|_{t=t_k}+O(h^3).$$

---

[1]) Метод прямых с сохранением производных по x широко применялся также к квазилинейным уравнениям параболического типа; см., например, Rothe, Zweidimensionale parabolische Randwertaufgaben als Grenzfall eindimensionaler Rondwerfaufgaben, Math. Annalen, т. 102, Heft 4/5, 1929; А. Н. Колмогоров, И. Г. Петровский, Н. С. Пискунов, Исследование уравнения диффузии..., Бюлл. МГУ, вып. 6, 1937; О. А. Олейник, А. С. Калашников, Чжоу Юй-линь, Задача Коши и краевые задачи для уравнений типа нестационарной фильтрации, ИАН СССР, т. 22, № 5, 1958; Чжоу Юй-линь, Краевые задачи для нелинейных параболических уравнений, Матем сб., т. 47 (89), № 4, 1959.

[2]) Простейшую схему такого варианта метода прямых, ее сходимость и оценку погрешности см. в книге В. И. Смирнова, Курс высшей математики, т. 4, Гостехиздат, 1951, стр. 737—739.

Умножая третье равенство на $-\frac{h}{2}$, четвертое на $\frac{h}{2}$ и складывая их с первым и вторым равенствами, получим:

$$u(x, t_k+h)+u(x, t_k-h)- \\ -\frac{h}{2}\left[u'_t(x, t_k+h)-u'_t(x, t_k-h)\right]=2u(x, t_k)+O(h^4)$$

или

$$u(x, t_k+h)-2u(x, t_k)+u(x, t_k-h)= \\ =\frac{h}{2}\left[u'_t(x, t_k+h)-u'_t(x, t_k-h)\right]+O(h^4). \qquad (78)$$

Заменяя в (78) производные по t из уравнения (53) и отбрасывая $O(h^4)$, получим для определения приближенных значений $U_k(x)$ решения $u(x, t)$ на прямых $t=t_k$ систему

$$\left.\begin{aligned} &U''_{k+1}(x)-U''_{k-1}(x)-\frac{2}{h}\left[U_{k+1}(x)-2U_k(x)+U_{k-1}(x)\right]= \\ &\qquad =f_{k-1}(x)-f_{k+1}(x), \quad (k=1, 2, \ldots), \\ &U_0(x)=\varphi(x) \end{aligned}\right\} \qquad (79)$$

с граничными условиями

$$U_k(0)=\psi_{1,k}; \quad U_k(l)=\psi_{2,k} \qquad (k=1, 2, \ldots). \qquad (80)$$

Системы (76) и (79) можно решать как рекуррентные системы, если каким-либо способом найти $U_1(x)$.

Пример. Построить методом прямых приближенное решение задачи

$$\frac{\partial u}{\partial t}=\frac{\partial^2 u}{\partial x^2}+1; \quad u|_{t=0}=u|_{x=0}=u|_{x=\pi}=0.$$

Для этого отрезок $[0, \pi]$ разделим на четыре части и проведем через точки деления прямые. Если $U_k(t)$ $(k=0, 1, 2, 3, 4)$ приближенные значения решения на прямых $x=\frac{\pi k}{4}$ $(k=0, 1, 2, 3, 4)$, то для отыскания $U_k(t)$ имеем систему линейных дифференциальных уравнений

$$\frac{5}{6}U'_1+\frac{1}{12}U'_2-\frac{16}{\pi^2}(U_2-2U_1)=1,$$

$$\frac{5}{6}U'_2+\frac{1}{12}(U'_1+U'_3)-\frac{16}{\pi^2}(U_3-2U_2+U_1)=1 \quad U_0(t)=U_4(t)=0,$$

$$\frac{5}{6}U'_3+\frac{1}{12}U'_2-\frac{16}{\pi^2}(U_2-2U_3)=1$$

с начальными условиями

$$U_1(0)=U_2(0)=U_3(0)=0.$$

Частное решение неоднородной системы можно искать в виде

$$U_i = A_i = \text{const}.$$

Подстановка в систему дает

$$A_1 = A_3 = \frac{3\pi^2}{32}; \quad A_2 = \frac{\pi^2}{8}.$$

Общее решение однородной системы получаем из (62). Таким образом, общее решение неоднородной системы имеет вид

$$U_1(t) = C_1 \sin\frac{\pi}{4} e^{-\sigma_1^2 t} + C_2 \sin\frac{\pi}{2} e^{-\sigma_2^2 t} + C_3 \sin\frac{3\pi}{4} e^{-\sigma_3^2 t} + \frac{3\pi^2}{32} =$$

$$= \frac{\sqrt{2}}{2}\left(C_1 e^{-\sigma_1^2 t} + \sqrt{2}\, C_2 e^{-\sigma_2^2 t} + C_3 e^{-\sigma_3^2 t}\right) + \frac{3\pi^2}{32},$$

$$U_2(t) = C_1 \sin\frac{\pi}{2} e^{-\sigma_1^2 t} + C_2 \sin\pi e^{-\sigma_2^2 t} + C_3 \sin\frac{3\pi}{2} e^{-\sigma_3^2 t} + \frac{\pi^2}{8} =$$

$$= C_1 e^{-\sigma_1^2 t} - C_3 e^{-\sigma_3^2 t} + \frac{\pi^2}{8},$$

$$U_3(t) = C_1 \sin\frac{3\pi}{4} e^{-\sigma_1^2 t} + C_2 \sin\frac{3\pi}{2} e^{-\sigma_2^2 t} + C_3 \sin\frac{9\pi}{4} e^{-\sigma_3^2 t} + \frac{3\pi^2}{32} =$$

$$= \frac{\sqrt{2}}{2}\left(C_1 e^{-\sigma_1^2 t} - \sqrt{2}\, C_2 e^{-\sigma_2^2 t} + C_3 e^{-\sigma_3^2 t}\right) + \frac{3\pi^2}{32},$$

где

$$\sigma_1^2 = \frac{16 \cdot 24 \sin^2\frac{\pi}{8}}{\pi^2\left(5 + \cos\frac{\pi}{4}\right)} = 0{,}9984, \quad \sigma_2^2 = \frac{16 \cdot 24 \sin^2\frac{\pi}{4}}{\pi^2\left(5 + \cos\frac{\pi}{2}\right)} = 38{,}9073,$$

$$\sigma_3^2 = \frac{16 \cdot 24 \sin^2\frac{3\pi}{8}}{\pi^2\left(5 + \cos\frac{3\pi}{4}\right)} = 2{,}2458.$$

Для определения C_1, C_2, C_3 из начальных условий получаем систему

$$C_1 + \sqrt{2}\, C_2 + C_3 = -\frac{3\pi^2\sqrt{2}}{32},$$

$$C_1 - C_3 = -\frac{\pi^2}{8},$$

$$C_1 - \sqrt{2}\, C_2 + C_3 = -\frac{3\pi^2\sqrt{2}}{32},$$

откуда

$$C_1 = -1{,}2700; \quad C_2 = 0; \quad C_3 = -0{,}0374.$$

Итак, окончательно

$$U_1(t) = U_3(t) = 0{,}9253 - 0{,}8980e^{-0{,}9984t} - 0{,}0264e^{-2{,}2458t}$$
$$U_2(t) = 1{,}2337 - 1{,}2700e^{-0{,}9984t} + 0{,}0374e^{-2{,}2458t}.$$

Для точного решения на прямой $x = \frac{\pi}{4}$ имеем:

$$u\left(\frac{\pi}{2}, t\right) = \frac{\pi^2}{8} - \frac{4}{\pi}\sum_{k=0}^{\infty}\frac{(-1)^k}{(2k+1)^3}e^{-(2k+1)^2 t}.$$

Для наглядности приведем следующие данные:

t	0,5	1,0
Точное решение	0,4620	0,7663
Приближенное решение	0,4750	0,7698

По поводу других применений метода прямых к решению краевых задач для уравнений в частных производных см. цитированные выше статьи Б. М. Будака и В. И. Лебедева, а также Я. И. Алихашкин, Решение задачи о несовершенной скважине методом прямых, Вычислит. математ., № 1, 1957; Б. М. Будак, А. Д. Горбунов, Метод прямых для решения одной нелинейной краевой задачи в области с криволинейной границей, ДАН, т. 118, № 5, 1958, стр. 858—862 или А. Д. Горбунов, Б. М. Будак, Метод прямых для решения одной нелинейной краевой задачи в области с криволинейной границей, Вестник МГУ, № 3, 1958, стр. 3—11; О. М. Белоцерковский, Расчет обтекания кругового цилиндра с отошедшей ударной волной, Вычислительная математика, сб. 3, 1958; Е. А. Григорьева, Метод прямых в смешанных задачах для параболических систем, ДАН, т. 119, № 4, стр. 649—651, 1958; Л. И. Камынин, О применении метода конечных разностей к решению уравнения теплопроводности, ИАН СССР, Серия математическая, 17 (1953), стр. 163—180 и стр. 249—268; П. И. Чушкин, Обтекание эллипсов и эллипсоидов дозвуковым потоком газа, Вычислит. математ., № 2, 1957.

§ 9. Вариационные методы решения краевых задач для дифференциальных уравнений математической физики

Среди приближенных методов решения уравнений в частных производных значительное место занимают вариационные методы. В некоторых областях механики эти методы являются самыми распространенными. В § 10 главы 9 мы уже рассматривали вариационные методы решения краевых задач для обыкновенных дифференциальных уравнений. Здесь мы рассмотрим применение этих методов к решению краевых задач для линейных дифференциальных уравнений в частных произ-

водных второго порядка эллиптического типа. Как уже говорилось, в основе вариационных методов лежит замена краевой задачи для дифференциального уравнения эквивалентной ей вариационной задачей. Приближенное решение краевой задачи сводится к построению приближенного решения соответствующей ей вариационной задачи. Более подробно мы остановимся на методе Ритца приближенного решения вариационных задач, соответствующих тем или иным краевым задачам, поэтому, чтобы не обосновывать сходимость этого метода в каждом конкретном случае, мы изложим этот метод в общем виде, а в конкретных случаях будем лишь проверять выполнение условий, при которых этот метод применим.

1. Метод Ритца решения операторных уравнений и отыскания собственных значений операторов в гильбертовом пространстве. Пусть на линейном множестве H_A, всюду плотном в гильбертовом пространстве H, определен аддитивный оператор A и f — некоторый элемент из H. В H_A требуется найти элемент, являющийся решением уравнения

$$Ay = f. \tag{1}$$

В § 10 главы 9 было показано, что если оператор A положителен, то уравнение (1) имеет не более одного решения, и если решение уравнения (1) существует, то функционал

$$J(y) = (Ay, y) - (f, y) - (y, f), \tag{2}$$

определенный на H_A, достигает на этом элементе наименьшего значения, т. е. если обозначить через z решение уравнения (1), то

$$J(z) = \inf_{y \in H_A} J(y) = \mu, \tag{3}$$

и наоборот, элемент, реализующий минимум функционала $J(y)$ на H_A, является решением уравнения (1).

В дальнейшем мы всегда будем предполагать существование решения уравнения (1) и будем лишь рассматривать способы приближенного построения этого решения.

Для построения приближенного решения уравнения (1) в предположении, что A — положительный оператор, строят последовательность $\{z_n\}$ $(z_n \in H_A)$, обладающую тем свойством, что

$$\lim_{n \to \infty} J(z_n) = \inf_{y \in H_A} J(y) = \mu. \tag{4}$$

Последовательности, для которых имеет место условие (4), называют *минимизирующими*. Если минимизирующая последовательность окажется сходящейся к элементу $z \in H_A$, то этот элемент будет являться решением задачи о минимуме функционала $J(y)$ в H_A, а следовательно и решением уравнения (1). За приближенное решение уравнения (1) принимают некоторый член z_n этой последовательности.

Как уже отмечалось ранее, не каждая минимизирующая последовательность является сходящейся. Для того чтобы каждая минимизирующая последовательность сходилась к решению z уравнения (1), нужно наложить на оператор A дополнительные ограничения. Таким требованием будет положительная определенность оператора A.

Оператор A называют *положительно определенным*, если существует такая положительная постоянная K^2, что для любого элемента $y \in H_A$ имеет место неравенство

$$(Ay, y) \geqslant K^2 \|y\|^2. \tag{5}$$

Заметим, что положительно определенный оператор является и положительным оператором.

Теорема. *Если A положительно определенный оператор, то любая минимизирующая последовательность $\{z_n\}$ $(z_n \in H_A)$ функционала* (2) *сходится к решению z вариационной задачи, т. е.*

$$\lim_{n \to \infty} \|z_n - z\| = 0. \tag{6}$$

В самом деле, если z — решение вариационной задачи, то $Az = f$ и

$$J(z) = (Az, z) - (z, f) - (f, z) =$$
$$= (Az, z) - (z, Az) - (Az, z) = -(z, Az) = \mu.$$

Далее,

$$J(z_n) - \mu = (Az_n, z_n) - (z_n, f) - (f, z_n) + (z, Az) =$$
$$= (Az_n, z_n) - (z_n, Az) - (Az, z_n) + (z, Az) =$$
$$= (A(z_n - z), z_n) - (z_n - z, Az) = (A(z_n - z), z_n) - (A(z_n - z), z) =$$
$$= (A(z_n - z), z_n - z) \geqslant K^2 \|z_n - z\|^2.$$

Так как $J(z_n) \to \mu$, то $\|z_n - z\|^2 \to 0$, что и доказывает наше утверждение.

Таким образом, если A — положительно определенный оператор, то за приближенное решение уравнения (1) можно принять элемент z_n любой минимизирующей последовательности при достаточно большом n.

Один из способов построения минимизирующей последовательности предложил Ритц. Метод Ритца заключается в следующем. В H_A выбирается последовательность элементов $x_1, x_2, \ldots, x_n, \ldots$, обладающая следующими свойствами:

1) любое конечное число членов этой последовательности линейно независимо;

2) для любого $\varepsilon > 0$ и любого элемента $y \in H_A$ найдется такое m и такие числа $a_1, a_2, \ldots, a_m$, что имеет место неравенство

$$\left(A\left(y - \sum_{k=1}^{m} a_k x_k\right), \quad y - \sum_{k=1}^{m} a_k x_k\right) < \varepsilon. \tag{7}$$

При фиксированном целом n строится линейная комбинация

$$u_n = \sum_{k=1}^{n} \alpha_k x_k \tag{8}$$

с произвольными численными коэффициентами α_k (будем предполагать, что H — действительное гильбертово пространство и α_k — действительные числа). Функционал $J(u_n)$ будет функцией $\alpha_1, \alpha_2, \ldots, \alpha_n$:

$$J(u_n) = \sum_{j,\,k=1}^{n} \alpha_j \alpha_k (Ax_j,\ x_k) - 2\sum_{j=1}^{n} \alpha_j (f,\ x_j). \tag{9}$$

Постоянные $\alpha_1, \alpha_2, \ldots, \alpha_n$ выбираются так, чтобы $J(u_n)$ принимал наименьшее значение на совокупности всевозможных линейных комбинаций (8). Для этих значений

$$\frac{\partial J(u_n)}{\partial \alpha_j} = 0 \qquad (j = 1,\ 2,\ \ldots,\ n), \tag{10}$$

т. е.

$$\sum_{k=1}^{n} \alpha_k (Ax_k,\ x_j) = (f,\ x_j) \qquad (j = 1,\ 2,\ \ldots,\ n). \tag{11}$$

Таким образом, для отыскания α_k получается симметричная система линейных алгебраических уравнений. Определитель этой системы отличен от нуля, так как если на H_A ввести скалярное произведение

$$[x,\ y] = (Ax,\ y), \tag{12}$$

что возможно, так как A — симметричный и положительно определенный оператор, то определитель системы есть определитель Грама системы линейно независимых элементов $x_1, x_2, \ldots, x_n$. Поэтому система (11) имеет единственное решение. Обозначим через z_n линейную комбинацию вида (8), где в качестве коэффициентов взято решение системы (11). Последовательность $\{z_n\}$ будет являться минимизирующей последовательностью. В самом деле, пусть задано $\varepsilon > 0$. Так как $\mu = \inf\limits_{y \in H_A} J(y)$, то найдется элемент $v \in H_A$, для которого

$$\mu \leqslant J(v) \leqslant \mu + \frac{\varepsilon}{2}. \tag{13}$$

В силу свойства 2) последовательности $x_1, x_2, \ldots, x_n, \ldots$ найдутся такое целое число m и такие числа $a_1, a_2, \ldots, a_m$, что при заданном $\eta > 0$ и $v_m = \sum\limits_{i=1}^{m} a_i x_i$ имеет место неравенство $(A(v - v_m),\ v - v_m) < \eta$. Но

$$\begin{aligned} J(v) - J(v_m) &= (Av,\ v) - (Av_m,\ v_m) - 2(f,\ v) + 2(f,\ v_m) = \\ &= (Av,\ v) - (Av_m,\ v_m) + 2(Az,\ v_m - v) = \\ &= -(A(v_m - v),\ v_m - v) + 2(A(v - v_m),\ v) + 2(Az,\ v_m - v) = \\ &= -[v_m - v,\ v_m - v] + 2[v - v_m,\ v] + 2[z,\ v_m - v]. \end{aligned}$$

(Здесь [ ] означает скалярное произведение, определенное равенством (12).) Применяя неравенство Буняковского, будем иметь:

$$|[v-v_m, v]| \leqslant \sqrt{[v-v_m, v-v_m][v, v]} =$$

$$= \sqrt{(A(v-v_m), v-v_m)}\sqrt{(Av, v)} < \sqrt{\eta}\sqrt{(Av, v)},$$

$$|[z, v_m-v]| \leqslant \sqrt{[z, z][v_m-v, v_m-v]} < \sqrt{\eta}\sqrt{(Az, z)}.$$

Отсюда

$$|J(v)-J(v_m)| \leqslant |(A(v_m-v), v_m-v)| + 2|[v-v_m, v]| +$$

$$+ 2|[z, v_m-v]| < \eta + 2\sqrt{\eta}\left[\sqrt{(Av, v)}+\sqrt{(Az, z)}\right].$$

Так как (Av, v), (Az, z) — фиксированные числа, то η можно выбрать так, что будет иметь место неравенство

$$|J(v)-J(v_m)| < \frac{\varepsilon}{2}. \tag{14}$$

Из (13) и (14) следует, что $\mu \leqslant J(v_m) < \varepsilon + \mu$, но тогда заведомо $\mu \leqslant J(z_m) < \mu + \varepsilon$ и при всех $n \geqslant m$ имеем $\mu \leqslant J(z_n) < \varepsilon + \mu$, а это и означает, что

$$\lim_{n\to\infty} J(z_n) = \mu,$$

т. е. $\{z_n\}$ — минимизирующая последовательность.

Заметим, что если вместо системы элементов $x_1, x_2, \ldots, x_n, \ldots$, которые мы будем называть *координатными*, взять новую последовательность координатных элементов $y_1, y_2, \ldots, y_n, \ldots$, связанную с $x_1, x_2, \ldots, x_n, \ldots$ соотношениями

$$y_k = \sum_{j=1}^{k} \beta_{kj} x_j \qquad (\beta_{kl} \neq 0), \tag{15}$$

то минимизирующие последовательности, построенные по методу Ритца, использующие $\{x_n\}$ и $\{y_n\}$, дадут один и тот же результат. Процессом ортогонализации можно построить такую последовательность $\{y_n\}$, что y_k будет выражаться через $x_1, x_2, \ldots, x_k$ с помощью равенств (15), а

$$(Ay_k, y_l) = [y_k, y_l] = \begin{cases} 0 & (k \neq l), \\ 1 & (k = l). \end{cases} \tag{16}$$

В этом случае система, аналогичная системе (11), примет вид

$$\alpha_j = (f, y_j) \qquad (j = 1, 2, \ldots, n) \tag{17}$$

и

$$z_n = \sum_{j=1}^{n} \alpha_j y_j = \sum_{j=1}^{n} (f, y_j) y_j. \tag{18}$$

Заметим также, что вместо свойства 2) системы координатных функций $\{x_i\}$ достаточно требовать полноту системы $\{Ax_i\}$ в H, т. е. потребовать от $\{x_i\}$, чтобы при любом $y \in H$ и $\varepsilon > 0$ существовали такие n и $\alpha_1, \alpha_2, \ldots, \alpha_n$, что

$$\left\| y - \sum_{j=1}^{n} \alpha_j Ax_j \right\| < \varepsilon, \tag{19}$$

так как из этого свойства следует свойство 2). В самом деле, пусть $y \in H_A$ и $\varepsilon > 0$ — заданное число, а $v = Ay$. Тогда по свойству $\{Ax_j\}$ можно найти такое n и такие α_j, что

$$\left\| v - \sum_{j=1}^{n} \alpha_j Ax_j \right\| < \sqrt{\varepsilon}\, K.$$

Далее, используя неравенство (5), имеем:

$$(A(y - v_n),\ y - v_n) \leqslant \|A(y - v_n)\|\ \|y - v_n\| \leqslant$$

$$\leqslant \left\| v - \sum_{j=1}^{n} \alpha_j Ax_j \right\| \frac{1}{K} \sqrt{(A(y - v_n),\ y - v_n)} <$$

$$< \sqrt{\varepsilon}\, \sqrt{(A(y - v_n),\ y - v_n)}.$$

Отсюда, сокращая на $\sqrt{(A(y - v_n),\ y - v_n)}$ и возводя обе части в квадрат, получим:

$$(A(y - v_n),\ y - v_n) < \varepsilon,$$

т. е. получим неравенство (7).

Если имеется способ отыскания чисел δ, меньших μ, но сколь угодно близких к μ, то можно получить оценку точности приближения z_n к решению z уравнения (1). Для этого воспользуемся уже ранее использованным неравенством

$$J(z_n) - \mu \geqslant K^2 \|z_n - z\|^2,$$

из которого следует, что

$$\|z_n - z\| \leqslant \frac{1}{K} \sqrt{J(z_n) - \mu} \leqslant \frac{1}{K} \sqrt{J(z_n) - \delta}, \tag{20}$$

что и позволяет оценить точность приближения z_n через $J(z_n)$.

Рассмотрим теперь задачу отыскания *собственных значений* оператора A, т. е. таких значений λ, для которых уравнение

$$Ay - \lambda y = 0 \tag{21}$$

имеет нетривиальные решения. Последние называются *собственными элементами* оператора A, соответствующими собственному значению λ.

На оператор A наложим следующие ограничения. Будем предполагать, что оператор A симметричен и ограничен снизу.

Оператор A называют *ограниченным снизу*, если существует такое число N, что для любого $y \in H_A$ имеет место неравенство

$$(Ay,\ y) \geqslant N\|y\|^2. \tag{22}$$

Теорема. *Для симметричного оператора A все собственные значения действительны, а собственные элементы, соответствующие различным собственным значениям, ортогональны.*

Действительно, если λ_0 — собственное значение оператора A, а y_0 — соответствующий ему собственный элемент, то

$$Ay_0 - \lambda_0 y_0 = 0. \tag{23}$$

Умножая обе части скалярно справа на y_0, получим:

$$(Ay_0,\ y_0) - \lambda_0 (y_0,\ y_0) = 0,$$

откуда

$$\lambda_0 = \frac{(Ay_0,\ y_0)}{(y_0,\ y_0)}, \tag{24}$$

а так как числитель и знаменатель — действительные числа, то λ_0 — действительное число. Пусть теперь λ_1 и λ_2 — различные собственные значения оператора A, а z_1 и z_2 — соответствующие им собственные элементы. Тогда

$$Az_1 - \lambda_1 z_1 = 0; \quad Az_2 - \lambda_2 z_2 = 0.$$

Умножим первое из них скалярно справа на z_2, а второе скалярно слева на z_1 и вычтем почленно. Получим:

$$(Az_1,\ z_2) - (z_1,\ Az_2) - (\lambda_1 - \lambda_2)(z_1,\ z_2) = 0.$$

В силу симметричности оператора A имеем $(Az_1,\ z_2) = (z_1,\ Az_2)$, т. е.

$$(\lambda_1 - \lambda_2)(z_1,\ z_2) = 0,$$

и так как $\lambda_1 \neq \lambda_2$, то $(z_1,\ z_2) = 0$.

Задача отыскания собственных значений оператора может быть сведена к вариационной задаче.

Теорема. *Если A — ограниченный снизу симметричный оператор, а λ_1 — нижняя грань значений функционала*

$$J(y) = \frac{(Ay,\ y)}{(y,\ y)}, \tag{25}$$

а z_1 — элемент, для которого $J(z_1) = \lambda_1$, то λ_1 есть наименьшее собственное значение оператора A, а z_1 — соответствующий ему собственный элемент.

В самом деле, пусть η — произвольный элемент из H_A, а t — произвольное действительное число. Обозначим через $\varphi(t)$ функцию

$$\varphi(t) = \frac{(A(z_1 + t\eta),\ z_1 + t\eta)}{(z_1 + t\eta,\ z_1 + t\eta)} = \frac{t^2(A\eta,\ \eta) + 2t\,\mathrm{Re}\,\{(Az_1,\ \eta)\} + (Az_1,\ z_1)}{t^2(\eta,\ \eta) + 2t\,\mathrm{Re}\,\{(z_1,\ \eta)\} + (z_1,\ z_1)}.$$

По условию она достигает минимума при $t=0$, но

$$\varphi'(0)=\frac{2\operatorname{Re}\{(Az_1,\ \eta)\}(z_1,\ z_1)-2\operatorname{Re}\{(z_1,\ \eta)\}(Az_1,\ z_1)}{(z_1,\ z_1)^2}.$$

Отсюда

$$(z_1,\ z_1)\operatorname{Re}\{(Az_1,\ \eta)\}-\operatorname{Re}\{(z_1,\ \eta)\}(Az_1,\ z_1)=0,$$

а так как $J(z_1)=\lambda_1$, то $(Az_1,\ z_1)=\lambda_1(z_1,\ z_1)$ и

$$\operatorname{Re}\{(Az_1,\ \eta)\}-\lambda_1\operatorname{Re}\{(z_1,\ \eta)\}=\operatorname{Re}\{(Az_1-\lambda_1z_1,\ \eta)\}=0.$$

Заменяя η на $i\eta$, получим:

$$\operatorname{Im}\{(Az_1-\lambda_1z_1,\ \eta)\}=0,$$

т. е.

$$(Az_1-\lambda_1z_1,\ \eta)=0.$$

Но η — произвольный элемент из H_A, а H_A всюду плотно в H, т. е.

$$Az_1-\lambda_1z_1=0,$$

и утверждение будет доказано, если мы покажем, что λ_1 — наименьшее собственное значение. Пусть λ' — любое другое собственное значение оператора A, а z' — его собственный элемент. Тогда

$$Az'-\lambda'z'=0 \quad \text{и} \quad (Az',\ z')-\lambda'(z',\ z')=0.$$

Отсюда

$$\lambda'=\frac{(Az',\ z')}{(z',\ z')}\geqslant\inf_{y\in H_A}\frac{(Ay,\ y)}{(y,\ y)}=\lambda_1.$$

Это заканчивает доказательство утверждения.

Отыскание следующих по величине собственных значений оператора A тоже может быть сведено к вариационной задаче. Это следует из теоремы:

Если $\lambda_1\leqslant\lambda_2\leqslant\ldots\leqslant\lambda_n$ *— первые* n *собственных значений симметричного ограниченного снизу оператора* A, *а* $z_1,\ z_2,\ \ldots,\ z_n$ *— соответствующие им ортонормированные собственные элементы, и* z_{n+1} *— элемент, реализующий минимум функционала* (25) *на множестве элементов* $y\in H_A$, *удовлетворяющих дополнительным условиям*

$$(y,\ z_i)=0 \qquad (i=1,\ 2,\ \ldots,\ n), \tag{26}$$

то z_{n+1} *— собственный элемент оператора* A, *соответствующий собственному значению* λ_{n+1}, *где*

$$\lambda_{n+1}=\frac{(Az_{n+1},\ z_{n+1})}{(z_{n+1},\ z_{n+1})}. \tag{27}$$

Для доказательства возьмем произвольный элемент $\xi \in H_A$ и положим $\eta = \xi - \sum_{k=1}^{n} (\xi, z_k) z_k$. Тогда $(\eta, z_j) = 0$ $(j = 1, 2, \ldots, n)$, так как

$$(\eta, z_j) = (\xi, z_j) - \sum_{k=1}^{n} (\xi, z_k)(z_k, z_j) = (\xi, z_j) - (\xi, z_j) = 0,$$

ибо $(z_k, z_j) = \begin{cases} 0 & (k \neq j) \\ 1 & (k = j) \end{cases}$. Этим свойством обладает и элемент $t\eta$, где t — любое действительное число, а также и $z_{n+1} + t\eta$. Функция

$$\psi(t) = \frac{(A(z_{n+1} + t\eta),\ z_{n+1} + t\eta)}{(z_{n+1} + t\eta,\ z_{n+1} + t\eta)}$$

достигает минимума при $t = 0$. Но это означает, что $\psi'(0) = 0$. Из этого условия, так же как и раньше, доказывается, что

$$(Az_{n+1} - \lambda_{n+1} z_{n+1},\ \eta) = 0.$$

Докажем, что и $(Az_{n+1} - \lambda_{n+1} z_{n+1},\ \xi) = 0$. Действительно,

$$(Az_{n+1} - \lambda_{n+1}\ z_{n+1},\ \xi) = (Az_{n+1} - \lambda_{n+1} z_{n+1},\ \eta) +$$

$$+ \sum_{k=1}^{n} \overline{(z_k,\ \xi)}\,(Az_{n+1} - \lambda_{n+1} z_{n+1},\ z_k) = \sum_{k=1}^{n} \overline{(z_k,\ \xi)}\,(Az_{n+1} - \lambda_{n+1} z_{n+1},\ z_k).$$

Но

$$(Az_{n+1} - \lambda_{n+1} z_{n+1},\ z_k) = (Az_{n+1},\ z_k) - \lambda_{n+1}(z_{n+1},\ z_k) =$$

$$= (z_{n+1},\ Az_k) - \lambda_{n+1}(z_{n+1},\ z_k) = (\lambda_k - \lambda_{n+1})(z_{n+1},\ z_k) = 0$$

в силу симметричности оператора A и ортогональности z_{n+1} ко всем z_k $(k = 1, 2, \ldots, n)$. Таким образом, для любого $\xi \in H_A$ имеем $(Az_{n+1} - \lambda_{n+1} z_{n+1},\ \xi) = 0$, а поэтому

$$Az_{n+1} - \lambda_{n+1} z_{n+1} = 0.$$

Если λ' — любое собственное значение, следующее по величине за λ_n, а z' — его собственный элемент, то z' ортогонален всем z_k $(k = 1, 2, \ldots, n)$ и

$$\lambda' = \frac{(Az', z')}{(z', z')} \geqslant \lambda_{n+1},$$

что полностью доказывает утверждение.

Вариационная задача, соответствующая задаче отыскания n-го собственного значения оператора A, может быть сформулирована и следующим образом:

Среди всех элементов $y \in H_A$, *удовлетворяющих условиям*

$$(y, y) = \|y\|^2 = 1, \quad (y, z_k) = 0 \qquad (k = 1, 2, \ldots, n - 1), \quad (28)$$

где z_k — собственные элементы, соответствующие собственным значениям $\lambda_1, \lambda_2, \ldots, \lambda_{n-1}$, найти минимум функционала $J_1(y) = (Ay, y)$. Минимум этого функционала и будет λ_n, а элемент, его реализующий, будет собственным элементом, соответствующим λ_n.

Метод Ритца приближенного решения этих задач (предполагая существование их решений) заключается в следующем.

Рассматривается последовательность координатных элементов $\{x_n\}$, обладающая такими свойствами:

1) любое конечное число их: $x_1, x_2, \ldots, x_n$, линейно независимо;

2) для каждого элемента $y \in H_A$ и любого $\varepsilon > 0$ существуют такие целые числа m и n и совокупности действительных чисел $a_1, a_2, \ldots, a_m$ и $b_1, b_2, \ldots, b_n$, что

$$\left\| y - \sum_{i=1}^{m} a_i x_i \right\| < \varepsilon, \tag{29}$$

$$\left(A\left(y - \sum_{i=1}^{n} b_i x_i \right), \quad y - \sum_{i=1}^{n} b_i x_i \right) < \varepsilon. \tag{30}$$

(Предполагается, что H — действительное гильбертово пространство и скалярное произведение элементов — действительное число.) Положим

$$u_n = \sum_{k=1}^{n} \alpha_k x_k \tag{31}$$

и выберем α_k так, чтобы u_n обеспечивало минимум функционала

$$J_1(u_n) = \sum_{k,l=1}^{n} \alpha_k \alpha_l (Ax_k, x_l) \tag{32}$$

при условии, что

$$(u_n, u_n) = \sum_{k,l=1}^{n} \alpha_k \alpha_l (x_k, x_l) = 1. \tag{33}$$

Решение этой задачи выполняем по правилу неопределенных множителей Лагранжа, т. е. составляем вспомогательную функцию

$$\Phi(\alpha_1, \alpha_2, \ldots, \alpha_n) = (Au_n, u_n) - \lambda (u_n, u_n) =$$

$$= \sum_{k,l=1}^{n} \alpha_k \alpha_l (Ax_k, x_l) - \lambda \sum_{k,l=1}^{n} \alpha_k \alpha_l (x_k, x_l)$$

и ищем безусловный минимум этой функции. Отсюда

$$\frac{\partial \Phi}{\partial \alpha_k} = 2 \left[\sum_{l=1}^{n} \alpha_l (Ax_k, x_l) - \lambda \sum_{l=1}^{n} \alpha_l (x_k, x_l) \right] = 0 \qquad (k = 1, 2, \ldots, n)$$

или

$$\sum_{l=1}^{n} \alpha_l [(Ax_k, x_l) - \lambda (x_k, x_l] = 0 \qquad (k = 1, 2, \ldots, n). \tag{34}$$

Определитель этой системы $D_n(\lambda)$ обязан быть равным нулю, так как все α_i не могут быть равны нулю одновременно. Таким образом, для отыскания λ имеем уравнение степени n:

$$D_n(\lambda) = \begin{vmatrix} (Ax_1, x_1)-\lambda(x_1, x_1) & (Ax_2, x_1)-\lambda(x_2, x_1) & \ldots & (Ax_n, x_1)-\lambda(x_n, x_1) \\ (Ax_1, x_2)-\lambda(x_1, x_2) & (Ax_2, x_2)-\lambda(x_2, x_2) & \ldots & (Ax_n, x_2)-\lambda(x_n, x_2) \\ \ldots & \ldots & \ldots & \ldots \\ (Ax_1, x_n)-\lambda(x_1, x_n) & (Ax_2, x_n)-\lambda(x_2, x_n) & \ldots & (Ax_n, x_n)-\lambda(x_n, x_n) \end{vmatrix} = 0. \tag{35}$$

Если λ_{0n} — корень этого уравнения, то $D_n(\lambda_{0n}) = 0$ и система (34) имеет нетривиальное решение при $\lambda = \lambda_{0n}$. Пусть оно $\alpha_1^0, \alpha_2^0, \ldots, \alpha_n^0$. При любом действительном $\mu \neq 0$ система чисел $\mu\alpha_1^0, \mu\alpha_2^0, \ldots, \mu\alpha_n^0$ будет также решением системы (34). Выберем μ так, чтобы выполнялось условие $(u_n, u_n) = 1$ и через $\alpha_1, \alpha_2, \ldots, \alpha_n$ обозначим $\mu\alpha_1^0, \mu\alpha_2^0, \ldots, \mu\alpha_n^0$ при данном значении μ. Тогда будем иметь тождества

$$\sum_{k=1}^{n} \alpha_k (Ax_k, x_l) = \lambda_{0n} \sum_{k=1}^{n} \alpha_k (x_k, x_l) \qquad (l = 1, 2, \ldots, n), \tag{36}$$

$$\sum_{k, l=1}^{n} \alpha_k \alpha_l (x_k, x_l) = 1. \tag{37}$$

Умножим (36) на α_l и просуммируем по l от 1 до n. Получим:

$$\sum_{k, l=1}^{n} \alpha_k \alpha_l (Ax_k, x_l) = \lambda_{0n} \sum_{k, l=1}^{n} \alpha_k \alpha_l (x_k, x_l) = \lambda_{0n}.$$

Но

$$\sum_{k, l=1}^{n} \alpha_k \alpha_l (Ax_k, x_l) = \left(A\left(\sum_{k=1}^{n} \alpha_k x_k \right), \sum_{l=1}^{n} \alpha_l x_l \right) = (Au_n, u_n).$$

Таким образом,

$$\lambda_{0n} = (Au_n, u_n). \tag{38}$$

Это показывает, что все корни уравнения (35) действительны и один из них дает минимум функционалу $J_1(y)$ на множестве $u_n = \sum_{i=1}^{n} \alpha_i x_i$; $(u_n, u_n) = 1$. Этот минимум, очевидно, равен наименьшему по величине корню уравнения (35), который мы обозначим через λ_{1n}.

С возрастанием n λ_{1n} не возрастает и в то же время остается не меньше λ_1. Таким образом,

$$\lim_{n\to\infty}\lambda_{1n}\geqslant\lambda_1.$$

Докажем, что

$$\lim_{n\to\infty}\lambda_{1n}=\lambda_1.$$

Рассмотрим сначала случай $\lambda_1>0$. В этом случае оператор A положительно определенный, так как для любого $y\in H_A$ имеем $(Ay,\ y)\geqslant\lambda_1(y,\ y)$. По определению нижней грани λ_1 для любого $\varepsilon>0$ найдется такой элемент $u\in H_A$, что $\lambda_1\leqslant(Au,\ u)<\lambda_1+\varepsilon$. Из свойства 2) последовательности $\{x_n\}$ следует, что найдется такой элемент $v_n=\sum\limits_{k=1}^{n}b_kx_k$, что

$$[u-v_n,\ u-v_n]=(A(u-v_n),\ u-v_n)<\varepsilon.$$

Отсюда, обозначая

$$\|y\|_0=\sqrt{[y,\ y]},\tag{39}$$

имеем:

$$\|v_n\|_0\leqslant\|u\|_0+\|v_n-u\|_0\leqslant\|u\|_0+\sqrt{\varepsilon}<\sqrt{\lambda_1+\varepsilon}+\sqrt{\varepsilon},$$

или

$$(Av_n,\ v_n)=\|v_n\|_0^2\leqslant\left(\sqrt{\lambda_1+\varepsilon}+\sqrt{\varepsilon}\right)^2.$$

Но $(A(u-v_n),\ u-v_n)\geqslant\lambda_1(u-v_n,\ u-v_n)$. Отсюда $(u-v_n,\ u-v_n)\leqslant\frac{\varepsilon}{\lambda_1}$ или $\|u-v_n\|\leqslant\sqrt{\frac{\varepsilon}{\lambda_1}}$, а $\|v_n\|\geqslant\|u\|-\|v_n-u\|\geqslant1-\sqrt{\frac{\varepsilon}{\lambda_1}}$. Таким образом,

$$\lambda_1\leqslant\frac{(Av_n,\ v_n)}{(v_n,\ v_n)}<\frac{\left(\sqrt{\lambda_1+\varepsilon}+\sqrt{\varepsilon}\right)^2}{1-\sqrt{\frac{\varepsilon}{\lambda_1}}}=\lambda_1+\delta,$$

где $\delta\to0$ вместе с ε. Так как $\lambda_{1n}=\min\frac{(Au_n,\ u_n)}{(u_n,\ u_n)}$, то $\lambda_1\leqslant\lambda_{1n}\leqslant\frac{(Av_n,\ v_n)}{(v_n,\ v_n)}<\lambda_1+\delta$. При $m>n$ имеем $\lambda_{1m}<\lambda_{1n}$ и, следовательно, $\lim\limits_{n\to\infty}\lambda_{1n}=\lambda_1$.

Если $\lambda_1<0$, то введем вспомогательный оператор

$$A_1y=Ay+(1-\lambda_1)y.\tag{40}$$

Это положительно определенный оператор, так как

$$(A_1y,\ y)=(Ay,\ y)+(1-\lambda_1)(y,\ y)\geqslant(y,\ y).$$

Далее,

$$\inf_{y \in H_A} \frac{(A_1 y, y)}{(y, y)} = 1. \tag{41}$$

Если $u_n = \sum_{k=1}^{n} \alpha_k x_k$ и $(u_n, u_n) = 1$, то

$$\inf (A_1 u_n, u_n) = \inf [(Au_n, u_n) + (1 - \lambda_1)(u_n, u_n)] = \lambda_{1n} + (1 - \lambda_1).$$

Отсюда по ранее доказанному

$$\lim_{n \to \infty} \inf (A_1 u_n, u_n) = \lim_{n \to \infty} (\lambda_{1n} + (1 - \lambda_1)) = 1,$$

т. е.

$$\lim_{n \to \infty} \lambda_{1n} = \lambda_1.$$

Для отыскания следующего по величине собственного значения ищем минимум функционала $J_1(u_n) = (Au_n, u_n)$ при условиях

$$(u_n, u_n) = 1; \quad (u_n, z_n) = \sum_{k, l=1}^{n} \alpha_k \beta_l (x_k, x_l) = 0, \tag{42}$$

где $z_n = \sum_{k=1}^{n} \beta_k x_k$ — приближенное значение первой нормированной собственной функции. Для этого составляем вспомогательную функцию

$$\Phi(\alpha_1, \alpha_2, \ldots, \alpha_n) = (Au_n, u_n) - \lambda (u_n, u_n) - \mu (u_n, z_n)$$

и приравниваем нулю ее частные производные по α_k. Это приводит к системе

$$\sum_{k=1}^{n} \{\alpha_k [(Ax_k, x_l) - \lambda (x_k, x_l)] - \mu \beta_k (x_k, x_l)\} = 0. \tag{43}$$

Если обе части равенства (43) умножить на β_l и просуммировать по l от 1 до n, то получим:

$$\sum_{k, l=1}^{n} \alpha_k \beta_l [(Ax_k, x_l) - \lambda (x_k, x_l)] - \mu \sum_{k, l=1}^{n} \beta_k \beta_l (x_k, x_l) = 0,$$

или

$$\sum_{k=1}^{n} \alpha_k \left\{ \sum_{l=1}^{n} \beta_l [(Ax_k, x_l) - \lambda (x_k, x_l)] \right\} - \mu (z_n, z_n) = 0.$$

Но из системы (36)

$$\sum_{l=1}^{n} \beta_l (Ax_k, x_l) = \lambda_{1n} \sum_{l=1}^{n} \beta_l (x_k, x_l),$$

откуда

$$\sum_{k=1}^{n} \alpha_k \left\{ \sum_{l=1}^{n} \beta_l \left[(Ax_k,\ x_l) - \lambda (x_k,\ x_l) \right] \right\} - \mu (z_n,\ z_n) =$$

$$= \sum_{k=1}^{n} \alpha_k \sum_{l=1}^{n} (\lambda_{1n} - \lambda) \beta_l (x_k,\ x_l) - \mu (z_n,\ z_n) =$$

$$= (\lambda_{1n} - \lambda_1)(u_n,\ z_n) - \mu (z_n,\ z_n) = 0.$$

Так как $(u_n,\ z_n) = 0$ и $(z_n,\ z_n) = 1$, то $\mu = 0$ и система (43) совпадает с системой (36). Отсюда, как и прежде, заключаем, что искомый минимум равен второму по величине корню уравнения $D_n(\lambda) = 0$.

Аналогично ищутся и следующие собственные значения оператора A. Они приближенно равны следующим по величине корням уравнения (35). Нужно только отметить, что точность приближения следующих собственных значений меньше.

2. Метод Ритца приближенного решения краевых задач для линейных дифференциальных уравнений в частных производных второго порядка эллиптического типа. В конечной области G, ограниченной кусочно-гладким контуром Γ, задано уравнение

$$Lu = -\frac{\partial}{\partial x}\left(p \frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\left(p \frac{\partial u}{\partial y}\right) + qu = f, \tag{44}$$

где p — положительная непрерывно дифференцируемая в области $G + \Gamma$ функция, а q и f — непрерывные в $G + \Gamma$ функции, при этом $q \geqslant 0$.

Рассмотрим следующие краевые задачи:

Найти решение уравнения (44), удовлетворяющее на границе Γ одному из следующих трех краевых условий:

$$u|_{\Gamma} = 0, \tag{45}$$

$$\left[\frac{du}{dn} + \sigma u\right]_{\Gamma} = 0 \tag{46}$$

(σ — непрерывная неотрицательная функция, не равная тождественно нулю, а n — внешняя нормаль);

$$\frac{du}{dn}\Big|_{\Gamma} = 0. \tag{47}$$

Мы рассматриваем нулевые краевые условия, так как общий случай может быть сведен к рассматриваемому, если в $G + \Gamma$ можно найти какую-нибудь достаточно гладкую функцию u_1, удовлетворяющую ненулевым заданным условиям и вместо функции u искать функцию $v = u - u_1$. Такая замена приведет нас к краевой задаче для уравнения, отличающегося от уравнения (44) только правой частью, но уже с нулевыми начальными условиями.

Рассмотрим гильбертово пространство $L_2(G)$ действительных функций, интегрируемых с квадратом в области G, в котором скалярное произведение определено равенством

$$(\varphi, \psi) = \int_G \int \varphi\psi \, dx\, dy. \tag{48}$$

В этом пространстве выделим три множества функций M, M_σ и M_0, элементами которых являются дважды непрерывно дифференцируемые функции в $G+\Gamma$, удовлетворяющие соответственно краевым условиям (45), (46) или (47).

Покажем, что оператор Lu положителен на множествах M и M_σ, а также и на множестве M_0, если в последнем случае $q(x, y) \not\equiv 0$.

В самом деле,

$$(Lu, u) = \int_G \int uLu \, dx\, dy =$$

$$= -\int_G \int u\left[\frac{\partial}{\partial x}\left(p\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(p\frac{\partial u}{\partial y}\right)\right] dx\, dy + \int_G \int q(x, y)\, u^2 \, dx\, dy =$$

$$= \int_G \int \left\{p\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right] - \frac{\partial}{\partial x}\left(pu\frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\left(pu\frac{\partial u}{\partial y}\right)\right\} dx\, dy +$$

$$+ \int_G \int qu^2 \, dx\, dy =$$

$$= \int_G \int \left\{p\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right] + qu^2\right\} dx\, dy - \int_\Gamma pu\frac{du}{dn}\, ds. \tag{49}$$

Если $u \in M$, то $u|_\Gamma = 0$ и

$$(Lu, u) = \int_G \int \left\{p\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right] + qu^2\right\} dx\, dy \geqslant 0. \tag{50}$$

Если $(Lu, u) = 0$, то $\frac{\partial u}{\partial x} \equiv \frac{\partial u}{\partial y} \equiv 0$ и $u = C = \text{const}$, но $u|_\Gamma = 0$, т. е. $C = 0$ и $u \equiv 0$, а это и показывает, что $(Lu, u) \geqslant 0$ при всех $u \in M$ и $(Lu, u) = 0$ только при $u \equiv 0$, что и означает положительность оператора Lu на M.

Если $u \in M_\sigma$, то из (49) и (46) имеем:

$$(Lu, u) = \int_G \int \left\{p\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right] + qu^2\right\} dx\, dy + \int_\Gamma p\sigma u^2 \, ds. \tag{51}$$

Так как $p > 0$, $q \geqslant 0$, $\sigma \geqslant 0$ и $\sigma \not\equiv 0$, то $(Lu, u) \geqslant 0$ и равенство нулю возможно лишь при $\frac{\partial u}{\partial x} = \frac{\partial u}{\partial y} \equiv 0$, т. е. при $u = C = \text{const}$. Но если $(Lu, u) = 0$, то и $\int_\Gamma p\sigma u^2 \, ds = C^2 \int_\Gamma p\sigma \, ds = 0$, а это озна-

чает, что $C=0$ и $u\equiv 0$, а следовательно, Lu — положительный оператор на M_σ.

Если $u\in M_0$ и $q\not\equiv 0$, то

$$(Lu, u)=\int_G\int\left\{p\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]+qu^2\right\}dx\,dy\geqslant 0. \tag{52}$$

Равенство нулю возможно лишь при $\int_G\int p\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]dx\,dy$ и $\int_G\int qu^2\,dx\,dy=0$. Из первого следует, что $u=C=\text{const}$, а из второго, так как $q\not\equiv 0$, следует, что $u\equiv 0$, т. е. и на M_0 оператор Lu положителен, если $q\not\equiv 0$.

Если $q(x, y)\equiv 0$, то Lu не будет положительным оператором на M_0, так как для любой функции u, тождественно равной в $G+\Gamma$ любой постоянной C, имеем $(Lu, u)=0$, а $u\not\equiv 0$. Но в этом случае решение краевой задачи (44), (47) существует не при всех f. Рассмотрим условия, при которых задача имеет решение. Пусть функция u удовлетворяет уравнению (44) с $q\equiv 0$ и краевому условию (47). Проинтегрируем по области G тождество

$$\frac{\partial}{\partial x}\left(p\frac{\partial u}{\partial x}\right)+\frac{\partial}{\partial y}\left(p\frac{\partial u}{\partial y}\right)=-f. \tag{53}$$

Будем иметь:

$$\int_G\int\left[\frac{\partial}{\partial x}\left(p\frac{\partial u}{\partial x}\right)+\frac{\partial}{\partial y}\left(p\frac{\partial u}{\partial y}\right)\right]dx\,dy=-\int_G\int f\,dx\,dy.$$

Но

$$\int_G\int\left[\frac{\partial}{\partial x}\left(p\frac{\partial u}{\partial x}\right)+\frac{\partial}{\partial y}\left(p\frac{\partial u}{\partial y}\right)\right]dx\,dy=-\int_\Gamma p\frac{\partial u}{\partial n}\,ds=0,$$

так как $\left.\frac{\partial u}{\partial n}\right|_\Gamma=0$. Таким образом, функция f должна удовлетворять условию

$$\int_G\int f\,dx\,dy=0. \tag{54}$$

Выделим из $L_2(G)$ функции, удовлетворяющие этому условию. Их совокупность образует в $L_2(G)$ линейное множество $\tilde{L}_2(G)$. Для $\tilde{L}_2(G)$ сохраним то же скалярное произведение, что и в $L_2(G)$. Полученное гильбертово пространство примем за основное. В нем выделим множество $\tilde{M}_0$ дважды непрерывно дифференцируемых в $G+\Gamma$ функций, удовлетворяющих на Γ условию (47). На этом множестве Lu будет уже положительным оператором. Действительно, при $u\in\tilde{M}_0$

из (52) имеем:

$$(Lu,\ u)=\int_G\int p\left\{\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right\}dx\,dy\geqslant 0.$$

Если $(Lu,\ u)=0$, то $\frac{\partial u}{\partial x}=\frac{\partial u}{\partial y}=0$ и $u=C=\text{const}$. Но так как $u\in\widetilde{M}_0$, то

$$\int\int_G u\,dx\,dy=C\int\int_G dx\,dy=0,$$

откуда $C=0$ и $u(x,\ y)\equiv 0$, что и доказывает положительность оператора Lu на $\widetilde{M}_0$.

Заметим, что в $\widetilde{M}_0$ уравнение (44) имеет единственное решение, чего нет в M_0, так как в M_0 при $q\equiv 0$ решение определяется с точностью до постоянного слагаемого.

На основании общей теории п. 1, если краевые задачи (44) — (45), (44) — (46), (44) — (47) имеют решения, что мы всегда будем предполагать, то они будут также и решениями следующих вариационных задач:

Решение задачи (44) — (45) является решением задачи о минимуме функционала

$$J(u)=\int_G\int\left\{p\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]+qu^2-2fu\right\}dx\,dy \qquad (55)$$

на множестве M,

решение задачи (44) — (46) является решением задачи о минимуме функционала

$$J(u)=\int_G\int\left\{p\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]+qu^2-2fu\right\}dx\,dy+\int_\Gamma p\sigma u^2\,ds \qquad (56)$$

на линейном множестве M_σ;

решение задачи (44) — (47) является решением задачи о минимуме функционала (55) на линейном множестве M_0, если $q(x,\ y)\not\equiv 0$, и решением задачи о минимуме функционала

$$J(u)=\int_G\int\left\{p\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]-2fu\right\}dx\,dy \qquad (55')$$

на линейном множестве $\widetilde{M}_0$ функций, удовлетворяющих условию (47) и условию

$$\int\int_G u\,dx\,dy=0, \qquad (57)$$

при этом функция f должна удовлетворять условию (54).

Для того чтобы к решению вариационных задач можно было применить метод Ритца, нужно показать, что оператор Lu на соответствующих линейных множествах является также и положительно определенным оператором. Тогда на основании общей теории будет иметь место сходимость в среднем минимизирующих последовательностей, полученных по методу Ритца, к точным решениям соответствующих краевых задач.

Положительную определенность оператора Lu нетрудно доказать при некоторых дополнительных ограничениях, воспользовавшись следующими утверждениями, которые мы приведем без доказательства. (Доказательства можно найти в книге Михлина С. Г. «Прямые методы в математической физике» или в книге Куранта и Гильберта «Методы математической физики», т. 2, гл. 7.)

Если функция $u(x, y)$ дважды непрерывно дифференцируема в области $G+\Gamma$ и на кусочно-гладкой границе Γ области G обращается в нуль, то существует такая положительная постоянная A, не зависящая от u, что имеет место неравенство

$$A\int_G\int u^2\,dx\,dy \leqslant \int_G\int \left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]dx\,dy, \tag{58}$$

называемое *неравенством Фридрихса.*

Если функция u дважды непрерывно дифференцируема в области $G+\Gamma$, то существует такая постоянная $B>0$, не зависящая от u, что справедливо неравенство

$$B\int_G\int u^2\,dx\,dy \leqslant \left\{\int_G\int \left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]dx\,dy+\int_\Gamma u^2\,ds\right\}, \tag{59}$$

которое также называют *неравенством Фридрихса*, а также неравенство

$$\int_G\int u^2\,dx\,dy \leqslant D\int_G\int \left\{\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right\}dx\,dy+E\left(\int_G\int u\,dx\,dy\right)^2, \tag{60}$$

где D и E — положительные постоянные, не зависящие от u, которое называется *неравенством Пуанкаре.*

Пусть теперь $u\in M$. Тогда по (50) имеем:

$$(Lu, u)=\int_G\int \left\{p\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]+qu^2\right\}dx\,dy \geqslant$$
$$\geqslant \int_G\int p\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]dx\,dy.$$

Положим $p_0 = \inf\limits_{(x, y)\in G+\Gamma} p(x, y)$. По условию $p_0 > 0$. Используя первое неравенство Фридрихса (58), имеем:

$$(Lu, u) \geqslant p_0 \int_G \int \left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right] dx\, dy \geqslant Ap_0 \int_G \int u^2\, dx\, dy = p_0 A \|u\|^2.$$

Так как $p_0 A = K^2 > 0$, то это и означает положительную определенность оператора Lu на M.

Пусть теперь $u \in M_\sigma$. В этом случае оператор Lu будет положительно определенным, если

$$\inf_{(x, y)\in G+\Gamma} q(x, y) = q_0 > 0, \tag{61}$$

или

$$\inf_{(x, y)\in\Gamma} \sigma(x, y) = \sigma_0 > 0. \tag{62}$$

Пусть $q_0 > 0$. Если $u \in M_\sigma$, то по (51)

$$(Lu, u) = \int_G \int \left\{p\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right] + qu^2\right\} dx\, dy + \int_\Gamma p\sigma u^2\, ds \geqslant$$

$$\geqslant q_0 \int_G \int u^2\, dx\, dy = q_0 \|u\|^2,$$

что и доказывает положительную определенность Lu на M_σ. Если $\sigma_0 > 0$, то

$$(Lu, u) = \int_G \int \left\{p\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right] + qu^2\right\} dx\, dy + \int_\Gamma p\sigma u^2\, ds \geqslant$$

$$\geqslant p_0 \int_G \int \left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right] dx\, dy + p_0\sigma_0 \int_\Gamma u^2\, ds.$$

Обозначим через δ наименьшее из чисел p_0 и $p_0\sigma_0$, и положим $K^2 = \delta B$, где B — постоянная в неравенстве Фридрихса (59) Имеем:

$$(Lu, u) \geqslant \delta\left\{\int_G \int \left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right] dx\, dy + \int_\Gamma u^2\, ds\right\} \geqslant$$

$$\geqslant K^2 \int_G \int u^2\, dx\, dy = K^2 \|u\|^2,$$

т. е. Lu — положительно определенный оператор на M_σ.

Для вариационной задачи, соответствующей краевой задаче (44) — (47), также рассмотрим два случая.

Если $q(x, y) \geqslant q_0 > 0$, то Lu — положительно определенный оператор на M_0, так как по (52) в этом случае

$$(Lu, u) = \int_G \int \left\{ p \left[\left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 \right] + qu^2 \right\} dx\, dy \geqslant$$

$$\geqslant q_0 \int_G \int u^2\, dx\, dy = q_0 \|u\|^2,$$

откуда и следует утверждение.

Если $q(x, y) \equiv 0$, то Lu — положительно определенный оператора на $\widetilde{M}_0$. В этом случае

$$(Lu, u) = \int_G \int p \left[\left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 \right] dx\, dy \geqslant p_0 \int_G \int \left[\left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 \right] dx\, dy,$$

и так как $\int_G \int u\, dx\, dy = 0$, то из неравенства Пуанкаре следует:

$$(Lu, u) \geqslant \frac{p_0}{D} \int_G \int u^2\, dx\, dy = \frac{p_0}{D} \|u\|^2 = K^2 \|u\|^2,$$

где $K^2 = \frac{p_0}{D} > 0$, что и показывает положительную определенность Lu на $\widetilde{M}_0$.

Итак, сходимость метода Ритца будет иметь место при тех ограничениях, при которых мы показали положительную определенность оператора Lu.

Для построения минимизирующей последовательности по методу Ритца выбираем последовательность координатных функций

$$\varphi_1(x, y), \quad \varphi_2(x, y), \quad \ldots, \quad \varphi_n(x, y), \quad \ldots, \tag{63}$$

удовлетворяющих следующим условиям:

1) $\varphi_i(x, y)$ дважды непрерывно дифференцируемы в $G + \Gamma$;

2) $\varphi_i(x, y)$ удовлетворяют заданным краевым условиям;

3) любое конечное число этих функций линейно независимо;

4) для любого $\varepsilon > 0$ и любой функции u, принадлежащей к множеству допустимых функций рассматриваемой вариационной задачи, найдутся такое целое число n и такие числа $a_1, a_2, \ldots, a_n$, что

$$\left(L\left(u - \sum_{i=1}^{n} a_i \varphi_i \right), \quad u - \sum_{i=1}^{n} a_i \varphi_i \right) < \varepsilon. \tag{64}$$

Для выполнения последнего условия достаточно потребовать, чтобы для $u \in L_2(G)$ и любого $\varepsilon > 0$ нашлись такие $b_1, b_2, \ldots, b_m$, что

$$\int_G \int \left[u - \sum_{i=1}^{m} b_i L\varphi_i \right]^2 dx\, dy < \varepsilon. \tag{64'}$$

В случае краевой задачи (44) — (47) от координатных функций φ_i можно не требовать выполнения краевых условий.

Члены минимизирующей последовательности имеют вид

$$u_n(x, y) = \sum_{i=1}^{n} a_i \varphi_i(x, y), \tag{65}$$

где числовые коэффициенты a_i суть решение системы

$$\sum_{k=1}^{n} A_{ik} a_k = B_i \qquad (i = 1, 2, \ldots, n), \tag{66}$$

а A_{ik} и B_i выражаются через коэффициенты уравнения (44) и координатные функции следующим образом:

$$\left.\begin{aligned} A_{ik} &= \int_G \int \left\{ q\varphi_k - \frac{\partial}{\partial x}\left(p \frac{\partial \varphi_k}{\partial x}\right) - \frac{\partial}{\partial y}\left(p \frac{\partial \varphi_k}{\partial y}\right)\right\} \varphi_i \, dx\, dy, \\ B_i &= \int_G \int f \varphi_i \, dx\, dy. \end{aligned}\right\} \tag{67}$$

Все утверждения этого пункта распространяются и на случай любого числа независимых переменных, т. е. на краевые задачи для уравнения

$$Lu = -\sum_{j=1}^{m} \frac{\partial}{\partial x_j} \sum_{k=1}^{m} p_{jk} \frac{\partial u}{\partial x_k} + qu = f, \tag{44'}$$

где p_{ij} — непрерывно дифференцируемые функции переменных $x_1, x_2, \ldots, x_m$ в конечной области G с гладкой границей Γ, удовлетворяющие условию, что в любой точке $G + \Gamma$ квадратичная форма

$$\sum_{j,k=1}^{m} p_{jk} \xi_i \xi_j$$

положительно определенна, q и f — непрерывные функции в $G + \Gamma$, причем $q \geqslant 0$, с граничными условиями одного из следующих видов:

$$u|_\Gamma = 0, \tag{45'}$$

$$\left[\sum_{j,k=1}^{m} p_{jk} \frac{\partial u}{\partial x_k} \cos(n x_j) + \sigma u\right]_\Gamma = 0, \tag{46'}$$

где σ — неотрицательная непрерывная на Γ функция и $\sigma \not\equiv 0$;

$$\sum_{i,j=1}^{m} p_{ij} \frac{\partial u}{\partial x_i} \cos(n x_j)\Big|_\Gamma = 0. \tag{47'}$$

3. Некоторые другие вариационные методы. Кроме метода Ритца существует ряд других приближенных методов решения вариационных задач, соответствующих краевым задачам. Не останавливаясь на них подробно, кратко изложим сущность некоторых из них на примере задачи Дирихле для уравнения (44).

Метод Л. В. Канторовича. Для простоты предположим, что область G, в которой ищется решение уравнения (44) с краевыми условиями (45), ограничена прямыми $x=a$, $x=b$ и двумя кривыми $y=y_1(x)$, $y=y_2(x)$ $(a\leqslant x\leqslant b;\quad y_2(x)>y_1(x))$. Как мы видели, решение задачи сводится к решению задачи о минимуме функционала

$$J(u)=(Lu,\ u)-2(u,\ f)=\int_G\int\left\{p\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]+qu^2-2uf\right\}dx\,dy$$

на множестве M дважды непрерывно дифференцируемых функций, обращающихся в нуль на границе. В методе Канторовича приближенное решение $u_n(x,\ y)$ ищется в виде

$$u_n(x,\ y)=\sum_{k=1}^{n}f_k(x)\,\varphi_k(x,\ y), \tag{68}$$

где $\varphi_k(x,\ y)$ — заданные дважды непрерывно дифференцируемые функции, обращающиеся в нуль на границе Γ, за исключением, быть может, прямых $x=a$; $x=b$, а $f_k(x)$ — неизвестные функции.

Подставляя $u_n(x,\ y)$ в $J(u)$ и выполняя интегрирование по переменному y, получим:

$$J(u_n)=\int_a^b\Phi\left[x,\ f_1(x),\ f_2(x),\ \ldots,\ f_n(x);\ f_1'(x),\ f_2'(x),\ \ldots,\ f_n'(x)\right]dx, \tag{69}$$

где Φ — известная функция своих аргументов. Для отыскания $f_k(x)$ имеем вариационную задачу о минимуме однократного интеграла. Выписывая систему уравнений Эйлера

$$\frac{d}{dx}\frac{\partial\Phi}{\partial f_k'}-\frac{\partial\Phi}{\partial f_k}=0\qquad(k=1,\ 2,\ \ldots,\ n) \tag{70}$$

и присоединяя краевые условия

$$f_k(a)=f_k(b)=0\qquad(k=1,\ 2,\ \ldots,\ n), \tag{71}$$

получим краевую задачу для системы линейных дифференциальных уравнений второго порядка, решая которую, найдем $f_k(x)$ $(k=1,\ 2,\ \ldots,\ n)$, а следовательно и $u_n(x,\ y)$.

Можно показать, что в нашем случае система имеет вид

$$\int_{y_1(x)}^{y_2(x)}L(u_n)\,\varphi_k\,dy=0\qquad(k=1,\ 2,\ \ldots,\ n). \tag{70'}$$

В качестве функций φ_k можно брать, например, функции

$$\varphi_k(x, y) = (y - y_1(x))(y_2(x) - y)\, y^{k-1}, \tag{72}$$

или

$$\varphi_k(x, y) = \sin \frac{\pi k (y - y_1(x))}{y_2(x) - y_1(x)}. \tag{73}$$

При некоторых ограничениях на гладкость решения можно показать, что последовательность $\{u_n\}$, в которой φ_k имеют вид (72) или (73), равномерно сходится к решению краевой задачи (см. Л. В. Канторович и В. И. Крылов, Приближенные методы высшего анализа, гл. 4, ГИТТЛ, 1952).

Метод Куранта. Если в уравнении (44) правая часть имеет непрерывные производные до некоторого порядка m, Курант предложил вместо функционала $J(u)$ рассматривать функционал

$$J_1(u) = J(u) + \sum_{\alpha_1+\alpha_2=0}^{m} \int_G \int \left[\frac{\partial^{\alpha_1+\alpha_2}(Lu - f)}{\partial x^{\alpha_1} \partial y^{\alpha_2}} \right]^2 dx\, dy. \tag{74}$$

Очевидно $J_1(u) \geqslant J(u)$, а для функции u, реализующей минимум функционала $J(u)$ на множестве M, имеет место равенство $J_1(u) = J(u)$, так как $Lu = f$. Таким образом, решение краевой задачи (44) — (45) реализует минимум и функционала (74). Если мы построим минимизирующую последовательность $\{u_n(x, y)\}$ функционала $J_1(u)$, то, очевидно, при $n \to \infty$

$$\int_G \int \left[\frac{\partial^{\alpha_1+\alpha_2}(Lu_n - f)}{\partial x^{\alpha_1} \partial y^{\alpha_2}} \right]^2 dx\, dy \to 0 \qquad (\alpha_1 + \alpha_2 = 0, 1, 2, \ldots, m). \tag{75}$$

Это позволяет получить дополнительные заключения о характере сходимости u_n к u. Например, если решается задача Дирихле для уравнения Пуассона ($p \equiv 1$; $q \equiv 0$) и $m = 0$, то

$$J_1(u) = \int_G \int \left\{ \left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 - 2uf + (\Delta u + f)^2 \right\} dx\, dy. \tag{74'}$$

Построим минимизирующую последовательность $\{u_n\}$; будем иметь:

$$\int_G \int (\Delta u_n + f)^2\, dx\, dy \to 0. \tag{75'}$$

По формуле Грина

$$u_n(x, y) - u(x, y) = \int_G \int \Gamma(x, y; \xi, \eta)\, (\Delta u_n + f)_{(\xi, \eta)}\, d\xi\, d\eta.$$

Отсюда по неравенству Буняковского

$$|u_n(x, y) - u(x, y)| \leqslant$$

$$\leqslant \sqrt{\int_G \int \Gamma^2(x, y; \xi, \eta)\, d\xi\, d\eta} \sqrt{\int_G \int (\Delta u_n + f)^2\, dx\, dy}. \quad (76)$$

Первый множитель в правой части ограничен некоторой постоянной C, а это означает при учете (75′), что $\{u_n(x, y)\}$ равномерно сходится к $u(x, y)$.

Метод Треффтца. В методе Ритца приближенное решение ищется в классе функций, удовлетворяющих краевым условиям, но не удовлетворяющих дифференциальному уравнению. В противоположность этому в методе Треффтца приближенное решение ищется в классе функций, удовлетворяющих уравнению, но не удовлетворяющих краевому условию.

Пусть снова рассматривается краевая задача (44) — (45). Обозначим через $v_0(x, y)$ решение уравнения (44) и пусть $v_1(x, y)$, $v_2(x, y)$, ..., $v_n(x, y)$ — линейно независимые решения соответствующего однородного уравнения, т. е.

$$Lv_0 = f; \quad Lv_k = 0 \qquad (k = 1, 2, \ldots, n). \quad (77)$$

Тогда линейная комбинация

$$u_n(x, y) = v_0 + \sum_{k=1}^{n} \alpha_k v_k \quad (78)$$

будет снова решением уравнения (44): $Lu_n = f$. Требуется так подобрать коэффициенты α_k, чтобы функция $u_n(x, y)$ в каком-то смысле наиболее точно удовлетворяла граничным условиям (45). Например, можно подобрать $\alpha_1, \alpha_2, \ldots, \alpha_n$ так, чтобы интеграл

$$J_2(u_n) = \int_\Gamma u_n^2(x, y)\, ds \quad (79)$$

принимал бы наименьшее значение. В этом случае для отыскания $\alpha_1, \alpha_2, \ldots, \alpha_n$ мы получили бы систему линейных алгебраических уравнений

$$\frac{\partial J_2}{\partial \alpha_k} = 2 \int_\Gamma u_n v_k\, ds = 0 \qquad (k = 1, 2, \ldots, n). \quad (80)$$

В методе Треффтца от u_n требуется, чтобы разность u_n и точного решения задачи u обращала в минимум функционал

$$J_3(u) = \int_G \int \left\{ p \left[\left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 \right] + qu^2 \right\} dx\, dy, \quad (81)$$

т. е. $\alpha_1, \alpha_2, \ldots, \alpha_n$ подбирают так, чтобы обращалась в минимум функция

$$\Phi(\alpha_1, \alpha_2, \ldots, \alpha_n) =$$

$$= \int_G \int \left\{ p \left[\left(\frac{\partial (u_n - u)}{\partial x} \right)^2 + \left(\frac{\partial (u_n - u)}{\partial y} \right)^2 \right] + q (u_n - u)^2 \right\} dx\, dy. \quad (82)$$

Следовательно, $\alpha_1, \alpha_2, \ldots, \alpha_n$ должны являться решением системы

$$\frac{\partial \Phi}{\partial \alpha_k} = 2 \int_G \int \left\{ p \left[\frac{\partial (u_n - u)}{\partial x} \frac{\partial v_k}{\partial x} + \frac{\partial (u_n - u)}{\partial y} \frac{\partial v_k}{\partial y} \right] + \right.$$

$$\left. + q (u_n - u) v_k \right\} dx\, dy = 0 \quad (83)$$

$$(k = 1, 2, \ldots, n).$$

Интеграл в (83) можно преобразовать так, чтобы неизвестное нам решение u не входило. В самом деле, используя формулу Остроградского, интеграл в левой части (83) можно преобразовать так:

$$\int_G \int \left\{ -\frac{\partial}{\partial x} \left(p \frac{\partial v_k}{\partial x} \right) - \frac{\partial}{\partial y} \left(p \frac{\partial v_k}{\partial y} \right) + q v_k \right\} (u_n - u)\, dx\, dy +$$

$$+ \int_G \int \left\{ \frac{\partial}{\partial x} \left[p \frac{\partial v_k}{\partial x} (u_n - u) \right] + \frac{\partial}{\partial y} \left[p \frac{\partial v_k}{\partial y} (u_n - u) \right] \right\} dx\, dy =$$

$$= \int_G \int L v_k (u_n - u)\, dx\, dy + \int_\Gamma \left[p \frac{\partial v_k}{\partial x} \cos nx + p \frac{\partial v_k}{\partial y} \cos ny \right] (u_n - u) ds.$$

Так как $Lv_k = 0$, а $u|_\Gamma = 0$, то систему (83) можно переписать в виде

$$\int_\Gamma p \frac{\partial v_k}{\partial n} u_n\, ds = 0 \qquad (k = 1, 2, \ldots, n). \quad (83')$$

В эту систему $u(x, y)$ уже не входит. Решая ее, находим $\alpha_1, \alpha_2, \ldots, \alpha_n$, а следовательно и $u_n(x, y)$.

Отметим, что если $u_n(x, y)$ — приближенное решение краевой задачи, полученное по методу Треффтца, а $u(x, y)$ — точное решение, то имеет место неравенство

$$J(u_n) \leqslant J(u) = \mu, \quad (84)$$

т. е. метод Треффтца дает приближение к μ снизу.

4. Метод Ритца решения задачи о собственных значениях. При решении ряда задач математической физики, в частности при решении уравнений в частных производных методом Фурье, приходится решать задачу о собственных значениях дифференциальных операторов.

Рассмотрим несколько простейших задач о собственных значениях. Пусть требуется найти значения λ, для которых уравнение

$$Lu - \lambda u = -\frac{d}{dx}\left[p(x)\frac{du}{dx}\right] + qu - \lambda u = 0, \tag{85}$$

где $p(x)$ — положительная непрерывно дифференцируемая функция, а $q(x)$ — непрерывная на отрезке $[a, b]$ функция, имеет нетривиальное решение, удовлетворяющее краевым условиям

$$\alpha_1 u'(a) + \beta_1 u(a) = 0; \quad \alpha_2 u'(b) + \beta_2 u(b) = 0 \quad (\alpha_i^2 + \beta_i^2 > 0; \quad i = 1, 2). \tag{86}$$

Рассмотрим гильбертово пространство $L_2(a, b)$ и линейное множество M функций дважды непрерывно дифференцируемых на $[a, b]$, удовлетворяющих краевым условиям (86). На этом множестве операторов Lu симметричен и ограничен снизу. Действительно, если $u, v \in M$, то

$$(Lu, v) = -\int_a^b v\frac{d}{dx}\left(p\frac{du}{dx}\right)dx + \int_a^b quv\,dx =$$

$$= -\left[pv\frac{du}{dx}\right]_a^b + \int_a^b p\frac{du}{dx}\frac{dv}{dx}dx + \int_a^b quv\,dx =$$

$$= \left[p\left(u\frac{dv}{dx} - v\frac{du}{dx}\right)\right]_a^b - \int_a^b u\frac{d}{dx}\left(p\frac{dv}{dx}\right)dx + \int_a^b quv\,dx = (u, Lv), \tag{87}$$

ибо в силу краевых условий (86) внеинтегральный член обращается в нуль, и

$$(Lu, u) = \int_a^b \left[p\left(\frac{du}{dx}\right)^2 + qu^2\right]dx \geqslant \min_{[a, b]} q(x) \int_a^b u^2\,dx = N\|u\|^2, \tag{88}$$

где $N = \min\limits_{[a, b]} q(x)$.

На основании общей теории п. 1 для отыскания собственных значений λ оператора Lu с краевыми условиями (86) можно применить вариационные методы, в частности метод Ритца.

Выбрав в M систему координатных функций $\{y_k(x)\}$, обладающих свойствами, указанными в п. 1, ищем приближенное выражение для собственных функций в виде

$$u_n(x) = \sum_{k=1}^{n} \alpha_k y_k(x). \tag{89}$$

Для отыскания значений α_k имеем систему уравнений

$$\sum_{j=1}^{n} \alpha_j (A_{ij} - \lambda B_{ij}) = 0 \qquad (i = 1, 2, \ldots, n), \tag{90}$$

где

$$A_{ij} = \int_a^b q y_i y_j dx - \int_a^b \frac{d}{dx}\left(p \frac{dy_i}{dx}\right) y_j dx; \quad B_{ij} = \int_a^b y_i(x)\, y_j(x)\, dx, \tag{91}$$

в которой λ должно быть корнем уравнения

$$D(\lambda) = \begin{vmatrix} A_{11} - \lambda B_{11} & A_{12} - \lambda B_{12} & \ldots & A_{1n} - \lambda B_{1n} \\ A_{21} - \lambda B_{21} & A_{22} - \lambda B_{22} & \ldots & A_{2n} - \lambda B_{2n} \\ \ldots\ldots & \ldots\ldots & \ldots & \ldots\ldots \\ A_{n1} - \lambda B_{n1} & A_{n2} - \lambda B_{n2} & \ldots & A_{nn} - \lambda B_{nn} \end{vmatrix} = 0. \tag{92}$$

По доказанному ранее корни этого уравнения дают приближенные значения первых n собственных значений, а функции (89), в которых α_k есть решения системы (90) при λ, равном соответствующему корню уравнения (92), будут приближенными выражениями для соответствующих собственных функций.

Рассмотрим теперь задачу о собственных значениях для оператора

$$Lu = -\frac{\partial}{\partial x}\left(p \frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\left(p \frac{\partial u}{\partial y}\right) + qu, \tag{93}$$

где $p(x, y)$ — положительная непрерывно дифференцируемая в $G + \Gamma$ функция, а q — непрерывная в $G + \Gamma$ функция. Для примера рассмотрим краевые условия

$$u|_\Gamma = 0. \tag{94}$$

Мы уже видели, что этот оператор симметричен на множестве M дважды непрерывно дифференцируемых функций, обращающихся в нуль на Γ, принадлежащих к гильбертову пространству $L_2(G)$. Так как

$$(Lu, u) = \int_G \int \left\{p\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right] + qu^2\right\} dx\, dy \geqslant$$

$$\geqslant \min_G q(x, y) \int_G \int u^2\, dx\, dy, \tag{95}$$

то он ограничен снизу, т. е. и в этом случае применима общая теория п. 1. В соответствии с этой теорией приближенные выражения для собственных функций ищем в виде

$$u_n(x, y) = \sum_{i=1}^{n} \alpha_i v_i(x, y), \tag{96}$$

где $\{v_i(x, y)\}$ — последовательность координатных функций, обладающая свойствами, указанными в п. 1. Для отыскания α_i имеем систему уравнений вида (90), где теперь

$$\left.\begin{aligned} A_{ij} &= \int_G\int q v_i v_j\,dx\,dy - \int_G\int \left[\frac{\partial}{\partial x}\left(p\,\frac{\partial v_i}{\partial x}\right)+\frac{\partial}{\partial y}\left(p\,\frac{\partial v_i}{\partial y}\right)\right]\times \\ &\times v_j\,dx\,dy = \int_G\int \left[p\left(\frac{\partial v_i}{\partial x}\frac{\partial v_j}{\partial x}+\frac{\partial v_i}{\partial y}\frac{\partial v_j}{\partial y}\right)+q v_i v_j\right]dx\,dy, \\ B_{ij} &= \int_G\int v_i(x, y)\,v_j(x, y)\,dx\,dy, \end{aligned}\right\} \qquad (97)$$

а приближенные значения собственных значений находятся как корни уравнения (92), где A_{ij}, B_{ij} определяются равенствами (97). Для отыскания коэффициентов α_i, входящих в приближенное выражение собственных функций (86), в системе (90) нужно положить λ равным одному из этих корней.

Совершенно аналогично можно было бы рассмотреть и другие виды граничных условий, но на этом мы останавливаться не будем.

5. Метод Галеркина решения краевых задач. Метод Галеркина не является вариационным методом. Он не требует предварительного сведения краевой задачи для дифференциального уравнения в частных производных к вариационной задаче и поэтому он в некотором смысле более универсален, чем метод Ритца. Основная идея этого метода была изложена в § 9 главы 9. Мы рассмотрим кратко его применение к решению краевых задач для уравнений в частных производных, ограничиваясь уравнениями вида

$$Lu = -\frac{\partial}{\partial x}\left(p\,\frac{\partial u}{\partial x}\right)-\frac{\partial}{\partial y}\left(p\,\frac{\partial u}{\partial y}\right)+r\,\frac{\partial u}{\partial x}+s\,\frac{\partial u}{\partial y}+qu = f \qquad (98)$$

и граничными условиями вида (45), (46) или (47). При этом будем предполагать, что p, r, s, q, f — непрерывные функции в рассматриваемой области G, включая границу Γ, p непрерывно дифференцируема и $p(x, y) > 0$ в $G+\Gamma$.

В методе Галеркина приближенное решение краевой задачи для уравнения (98) с граничными условиями вида (45), (46) или (47) ищется в виде

$$u_n(x, y) = \sum_{i=1}^{n} \alpha_i v_i(x, y), \qquad (99)$$

где $v_1, v_2, \ldots, v_n$ — первые n функций последовательности $\{v_k(x, y)\}$, обладающие следующими свойствами:

1) функции $v_i(x, y)$ дважды непрерывно дифференцируемы в $G+\Gamma$;

2) любое конечное число их линейно независимо;

3) для любой дважды непрерывно дифференцируемой функции v, удовлетворяющей граничным условиям, и любого $\varepsilon > 0$ найдется такая линейная комбинация этих функций $\sum_{i=1}^{m} a_i v_i(x, y)$, что

$$-\int_G\int \left\{\frac{\partial}{\partial x}\left[p\frac{\partial}{\partial x}\left(v-\sum_{i=1}^{m}a_iv_i\right)\right]+\right.$$

$$\left.+\frac{\partial}{\partial y}\left[p\frac{\partial}{\partial y}\left(v-\sum_{i=1}^{m}a_iv_i\right)\right]\right\}\left(v-\sum_{i=1}^{m}a_iv_i\right)dx\,dy=$$

$$=\int_G\int p\left\{\left[\frac{\partial}{\partial x}\left(v-\sum_{i=1}^{m}a_iv_i\right)\right]^2+\left[\frac{\partial}{\partial y}\left(v-\sum_{i=1}^{m}a_iv_i\right)\right]^2\right\}dx\,dy<\varepsilon. \tag{100}$$

Коэффициенты $\alpha_i\,(i=1, 2, \ldots, n)$ определяются как решение системы

$$\int_G\int L(u_n)\,v_k(x, y)\,dx\,dy=$$

$$=\int_G\int fv_k(x, y)\,dx\,dy \qquad (k=1, 2, \ldots, n) \tag{101}$$

или

$$\sum_{j=1}^{n} A_{kj}\alpha_j = B_k \qquad (k=1, 2, \ldots, n), \tag{102}$$

где

$$A_{kj}=\int_G\int L(v_j)\,v_k\,dx\,dy; \quad B_k=\int_G\int fv_k(x, y)\,dx\,dy. \tag{103}$$

В случае, если $r\equiv s\equiv 0$ и $q\geqslant 0$, то при одной и той же системе координатных функций $\{v_k(x, y)\}$ метод Ритца и метод Галеркина дают одну и ту же последовательность $\{u_n(x, y)\}$, а это доказывает, что она, как и в методе Ритца, в среднем сходится к точному решению краевой задачи.

Для уравнения (98) с краевыми условиями одного из видов (45), (46) или (47) при условиях на коэффициенты, которые сформулированы выше, и при выборе последовательности координатных функций, удовлетворяющей условиям 1) — 3), последовательности $\{u_n\}$, полученные по методу Галеркина, сходятся в среднем вместе с производными первого порядка к точному решению краевой задачи и соответствующим производным этого решения, если только граница Γ кусочно-гладкая и в случае краевых условий (47) $q(x, y)\not\equiv 0$,

а в случае краевых условий (46) σ — достаточно гладкая на Γ функция.

Метод Галеркина применим и к задаче о собственных значениях для дифференциального оператора Lu при тех же типах краевых условий, но мало надежен для отыскания приближенных выражений для собственных функций.

Обоснование сходимости метода Галеркина мы не приводим, так как это заняло бы много места. Интересующихся отсылаем к книге Михлина С. Г. «Прямые методы в математической физике», где эти вопросы рассмотрены достаточно подробно.

§ 10. Приближенные методы решения интегральных уравнений

В этом параграфе мы рассмотрим некоторые методы приближенного решения интегральных уравнений, ограничиваясь в основном линейными интегральными уравнениями Фредгольма первого рода

$$\lambda \int_a^b K(x, s)\, y(s)\, ds = f(x), \tag{1}$$

интегральными уравнениями Фредгольма второго рода

$$y(x) - \lambda \int_a^b K(x, s)\, y(s)\, ds = f(x) \tag{2}$$

и интегральными уравнениями Вольтерра первого и второго рода, имеющими соответственно вид

$$\int_a^x K(x, s)\, y(s)\, ds = f(x), \tag{3}$$

$$y(x) - \lambda \int_a^x K(x, s)\, y(s)\, ds = f(x), \tag{4}$$

где $f(x)$ и $K(x, s)$ — заданные функции, а $y(x)$ — искомая функция.

1. Решение уравнений Фредгольма методом замены интеграла конечной суммой. При решении интегральных уравнений Фредгольма приходится встречаться с решением двух задач: 1) отыскание решения неоднородного интегрального уравнения при заданном значении параметра λ и заданной правой части $f(x)$; 2) отыскание собственных значений и собственных функций ядра $K(x, s)$, т. е. отыскание таких значений параметра λ, при которых однородное уравнение

$$y(x) - \lambda \int_a^b K(x, s)\, y(s)\, ds = 0 \tag{2'}$$

имеет нетривиальное решение $y(x)$. Эти значения λ и соответствующие им нетривиальные решения и называются, соответственно, собственными значениями и собственными функциями ядра $K(x, s)$.

Будем сначала предполагать, что ядро $K(x, s)$ и правая часть $f(x)$ непрерывны и, даже больше, имеют непрерывные производные до некоторого порядка. Тогда и решение уравнения имеет производные до того же порядка.

Для решения интегральных уравнений можно применять метод замены интеграла, входящего в уравнение, конечной суммой, используя для этого те или иные квадратурные формулы.

Пусть за основу принята некоторая квадратурная формула

$$\int_a^b F(x)\,dx = \sum_{j=1}^{n} A_j F(x_j) + R(F), \tag{5}$$

где абсциссы $x_1, x_2, \ldots, x_n$, принадлежащие отрезку $[a, b]$, и коэффициенты $A_1, A_2, \ldots, A_n$ не зависят от выбора функции $F(x)$, а $R(F)$ — остаточный член квадратурной формулы. Положим в интегральном уравнении (2) $x = x_i$ $(i = 1, 2, \ldots, n)$. Тогда

$$y(x_i) - \lambda \int_a^b K(x_i, s)\, y(s)\, ds = f(x_i) \qquad (i = 1, 2, \ldots, n). \tag{6}$$

Заменим в (6) интеграл с помощью квадратурной формулы (5). Будем иметь:

$$y(x_i) - \lambda \sum_{j=1}^{n} A_j K(x_i, x_j)\, y(x_j) = f(x_i) + \lambda R_i; \quad R_i = R[K(x_i, s)\, y(s)]. \tag{7}$$

Отбрасывая в системе (7) $\lambda(R_i)$, получим для отыскания приближенных значений Y_i решения $y(x)$ в узлах $x_1, x_2, \ldots, x_n$ линейную систему алгебраических уравнений

$$Y_i - \lambda \sum_{j=1}^{n} A_j K_{ij} Y_j = f_i \qquad (i = 1, 2, \ldots, n), \tag{8}$$

где введены обозначения $K_{ij} = K(x_i, x_j)$, $f(x_i) = f_i$. Решив эту систему, мы найдем значения $Y_1, Y_2, \ldots, Y_n$, по которым процессом интерполяции можно получить и приближенное решение интегрального уравнения (2) на всем отрезке $[a, b]$. За аналитическое выражение приближенного решения уравнения (2) можно принять функцию

$$Y(x) = f(x) + \lambda \sum_{j=1}^{n} A_j K(x, x_j)\, Y_j, \tag{9}$$

принимающую в узлах $x_1, x_2, \ldots, x_n$ значения $Y_1, Y_2, \ldots, Y_n$.

В случае уравнений Фредгольма первого рода (1) вместо системы (8) будем иметь систему

$$\lambda \sum_{j=1}^{n} A_j K_{ij} Y_j = f_i \qquad (i = 1, 2, \ldots, n). \tag{8'}$$

Если в качестве квадратурной формулы берется обобщенная формула прямоугольников, то

$$x_1 = a; \quad x_2 = a + h; \quad \ldots; \quad x_n = a + (n-1)h;$$
$$A_1 = A_2 = \ldots = A_n = \frac{b-a}{n};$$

если берется обобщенная формула трапеций, то

$$x_1 = a; \quad x_2 = a + h; \ \ldots; \ x_n = a + (n-1)h = b \qquad \left(h = \frac{b-a}{n-1}\right);$$
$$A_1 = A_n = \frac{h}{2}; \quad A_2 = A_3 = \ldots = A_{n-1} = h;$$

если же берется обобщенная формула Симпсона, то $n = 2m + 1$:

$$x_1 = a; \quad x_2 = a + h; \quad \ldots; \quad x_{2m+1} = a + 2mh = b; \quad h = \frac{b-a}{2m};$$
$$A_1 = A_{2m+1} = \frac{h}{3}; \quad A_2 = A_4 = \ldots = A_{2m} = \frac{4h}{3};$$
$$A_3 = A_5 = \ldots = A_{2m-1} = \frac{2h}{3}$$

и т. д.

Этот метод может быть применен и для решения нелинейных интегральных уравнений вида

$$\int_a^b K(x, s, y(x), y(s))\, ds = f(x), \tag{10}$$

но в этом случае вместо системы (8) получим систему

$$\sum_{j=1}^{n} A_j K(x_i, x_j; Y_i, Y_j) = f_i \qquad (i = 1, 2, \ldots, n), \tag{11}$$

которая уже будет нелинейной.

Вернемся к интегральному уравнению (2). Если это уравнение однородно, то и система (8) будет однородной системой. Она будет иметь нетривиальное решение в том и только в том случае, когда определитель системы (8) равен нулю. Приравнивая нулю этот определитель, получим алгебраическое уравнение, вообще говоря, степени n относительно λ. Решая это уравнение, найдем, вообще говоря, n корней $\bar{\lambda}_1, \bar{\lambda}_2, \ldots, \bar{\lambda}_n$, которые будут приближенными значениями первых n собственных значений ядра $K(x, s)$. Подставляя

в однородную систему, соответствующую системе (8), одно из найденных значений $\overline{\lambda}_i$ и находя линейно независимые решения этой системы, получим приближения к линейно независимым собственным функциям ядра $K(x, s)$, соответствующим данному собственному значению.

Если λ не равно ни одному из этих корней, то однородная система имеет только тривиальное решение, а система (8) — единственное решение.

При выборе квадратурной формулы в этом методе нужно иметь в виду, что чем более точную формулу мы применяем, тем большую гладкость ядра и решения, а следовательно и $f(x)$, нужно требовать. Попытка применения более точных квадратурных формул для получения более точного приближения при несоблюдении этого условия может привести совсем к обратному результату.

В случае, если правая часть или ядро $K(x, s)$ (или их производные) имеет особенности, целесообразно предварительно выполнить некоторые преобразования с тем, чтобы получить более хорошее интегральное уравнение, с помощью которого можно будет получить более точное приближенное решение исходного уравнения. Для этого могут быть полезны следующие приемы.

Если ядро гладкое, а правая часть $f(x)$ имеет особенности, то можно вместо $y(x)$ ввести новую неизвестную функцию:

$$z(x) = y(x) - f(x).$$

Подстановка ее в уравнение дает

$$z(x) - \lambda \int_a^b K(x, s)\, z(s)\, ds = \lambda \int_a^b K(x, s) f(s)\, ds,$$

т. е. мы получим уравнение в точности того же вида, но в котором правая часть будет уже более гладкой, а следовательно, и решение $z(x)$ будет более гладким. Найдя $z(x)$, затем найдем и искомое решение $y(x)$.

Очень часто встречаются уравнения, в которых ядро $K(x, s)$ или его производная по s имеет разрывы на диагонали $x = s$. В этом случае уравнение предварительно выгодно переписать в виде

$$y(x)\left[1 - \lambda \int_a^b K(x, s)\, ds\right] - \lambda \int_a^b K(x, s)\,[y(s) - y(x)]\, ds = f(x).$$

Подынтегральная функция во втором интеграле будет правильной, так как на диагонали $x = s$ разность $y(s) - y(x)$ обращается в нуль, а $\int_a^b K(x, s)\, ds$ уже не будет содержать неизвестной функции и его

часто можно вычислить в явном виде. Применение метода к этому последнему интегральному уравнению даст лучший результат.

Часто встречаются интегральные уравнения с ядрами вида

$$K(x, s)=\frac{H(x, s)}{|x-s|^{\alpha}} \qquad (0<\alpha<1),$$

где $H(x, s)$ — гладкая функция. От уравнений с такими ядрами целесообразно перейти к уравнениям с итерированными ядрами, которые уже не будут иметь особенности при $x=s$. (Об итерации ядер см., например, И. Г. Петровский, Лекции по теории интегральных уравнений.)

Рассмотрим теперь вопрос об оценке погрешности решения, получаемого по этому методу, предполагая в уравнении (2) наличие у ядра $K(x, s)$ и правой части $f(x)$ непрерывных производных до порядка q. Тогда и решение будет иметь непрерывные производные до порядка q.

Если обозначить определитель системы (8) через $D(\lambda)$, а алгебраические дополнения его элементов через $D_{ij}(\lambda)$, то решение системы (8) можно будет записать в виде

$$Y_i=\frac{1}{D(\lambda)}\sum_{j=1}^{n} D_{ij}f_j. \tag{12}$$

Из системы (7) будем иметь:

$$y(x_i)=\frac{1}{D(\lambda)}\sum_{j=1}^{n} D_{ij}(f_j+\lambda R_j). \tag{12'}$$

Обозначим через η_i погрешность приближенного решения в точке x_i, т. е.

$$\eta_i=y(x_i)-Y_i,$$

а через $\eta(x)$ обозначим разность $y(x)-Y(x)$, где $Y(x)$ определяется формулой (9). Тогда из (12) и (12′) имеем:

$$\eta_i=y(x_i)-Y_i=\frac{\lambda}{D(\lambda)}\sum_{j=1}^{n} D_{ij}R_j.$$

Если ввести обозначения

$$B=\max_i \frac{\sum_{j=1}^{n}|D_{ij}|}{|D(\lambda)|}; \quad \rho=\max_{a\leqslant x\leqslant b}|R|; \quad R=R[K(x, s)y(s)],$$

то

$$|\eta_i|\leqslant|\lambda|B\rho. \tag{13}$$

Для $\eta(x)$ получим следующее равенство:

$$\eta(x) = y(x) - Y(x) =$$

$$= \lambda \sum_{j=1}^{n} A_j K(x, x_j)\, y(x_j) + \lambda R - \lambda \sum_{j=1}^{n} A_j K(x, x_j)\, Y_j.$$

Отсюда получается следующая оценка:

$$|\eta(x)| \leqslant |\lambda| \sum_{j=1}^{n} |A_j|\,|K(x, x_j)|\,|y(x_j) - Y_j| + |\lambda|\,|R|$$

или

$$|\eta(x)| \leqslant |\lambda|\rho + |\lambda|^2 M_0 B \rho \sum_{j=1}^{n} |A_j|; \quad M_0 = \max_{a \leqslant x,\ s \leqslant b} |K(x, s)|. \qquad (13')$$

В оценках (13) и (13′) все константы могут быть вычислены, кроме константы ρ. Константа ρ есть максимум абсолютной величины остаточного члена квадратурной формулы для $F = K(x, s)\, y(s)$ при всех $x \in [a, b]$. Для формул трапеций, Симпсона, Гаусса и многих других остаточный член имеет вид

$$R(F) = k_n F^{(m)}(\xi),$$

где k_n — некоторая постоянная, зависящая от n, а ξ — некоторая точка отрезка $[a, b]$. Таким образом,

$$|R(F)| \leqslant k_n \max_{[a,\ b]} |F^{(m)}(s)|.$$

В нашем случае $F(s) = K(x, s)\, y(s)$ (x — параметр). Таким образом,

$$F^{(m)}(s) = \sum_{k=0}^{m} C_m^k \frac{\partial^k K(x, s)}{\partial s^k}\, y^{(m-k)}(s),$$

а

$$\max_{[a,\ b]} |F^{(m)}(s)| \leqslant \sum_{k=0}^{m} C_m^k M_k N_{m-k},$$

где введены обозначения:

$$M_k = \max_{a \leqslant x,\ s \leqslant b} \left\{ \left| \frac{\partial^k K(x, s)}{\partial x^k} \right|, \ \left| \frac{\partial^k K(x, s)}{\partial s^k} \right| \right\}; \quad N_k = \max_{a \leqslant s \leqslant b} |y^{(k)}(s)|. \qquad (14)$$

Постоянные N_k нам неизвестны, так как неизвестно решение. Но их можно оценить. Для этого продифференцируем интегральное уравнение (2) k раз. Будем иметь:

$$y^{(k)}(x) = \lambda \int_a^b \frac{\partial^k}{\partial x^k} K(x, s)\, y(s)\, ds + f^{(k)}(x),$$

откуда

$$|y^{(k)}(x)| \leqslant |\lambda| M_k (b-a) N_0 + P_k; \quad P_k = \max_{a \leqslant k \leqslant b} |f^{(k)}(x)|$$

или

$$N_k \leqslant |\lambda| M_k (b-a) N_0 + P_k.$$

Таким образом,

$$\max_{a \leqslant s \leqslant b} |F^{(m)}(s)| \leqslant |\lambda| N_0 (b-a) \sum_{k=0}^{m} C_m^k M_k M_{m-k} + \sum_{k=0}^{m} C_m^k M_k P_{m-k} = C_1 N_0 + C_2, \quad (15)$$

где

$$C_1 = |\lambda| (b-a) \sum_{k=0}^{m} C_m^k M_k M_{m-k}; \quad C_2 = \sum_{k=0}^{m} C_m^k M_k P_{m-k} \quad (16)$$

— постоянные величины, которые можно найти, так как ядро $K(x, s)$ и $f(x)$ известны. Обозначим через Q_0 максимум абсолютной величины приближенного решения $Y(x)$. Из оценки (13′) будем иметь:

$$|y(x)| \leqslant |\eta(x)| + |Y(x)| \leqslant |\lambda| \rho + |\lambda|^2 M_0 B \rho \sum_{j=1}^{n} |A_j| + Q_0 \leqslant \left\{ |\lambda| + |\lambda|^2 M_0 B \sum_{j=1}^{n} |A_j| \right\} k_n (C_1 N_0 + C_2) + Q_0$$

или

$$N_0 \leqslant \left\{ |\lambda| + |\lambda|^2 M_0 B \sum_{j=1}^{n} |A_j| \right\} k_n (C_1 N_0 + C_2) + Q_0.$$

Если

$$1 - k_n \left\{ |\lambda| + |\lambda|^2 M_0 B \sum_{j=1}^{n} |A_j| \right\} C_1 > 0,$$

то

$$N_0 \leqslant \frac{Q_0 + k_n C_2 \left\{ |\lambda| + |\lambda|^2 M_0 B \sum_{j=1}^{n} |A_j| \right\}}{1 - k_n C_1 \left\{ |\lambda| + |\lambda|^2 M_0 B \sum_{j=1}^{n} |A_j| \right\}}. \quad (17)$$

Следовательно, погрешности η_i и $\eta(x)$ можно оценить через известные величины.

Напомним, что: для обобщенной формулы трапеции с n ординатами $k_n = \frac{(b-a)^3}{12(n-1)^2}$, $m = 2$; для обобщенной формулы Симпсона с $n = 2p + 1$ ординатами $k_n = \frac{(b-a)^5}{90(2p)^4}$, $m = 4$; для формулы Гаусса с n ординатами $k_n = \frac{(b-a)^{2n+1}(n!)^4}{[(2n)!]^3 (2n+1)}$, $m = 2n$.

Пример. Найти приближенное решение уравнения

$$y(x)+\int_0^1 x\,(e^{xs}-1)\,y(s)\,ds=e^x-x.$$

Воспользуемся квадратурной формулой Симпсона:

$$\int_0^1 f(x)\,dx\approx\frac{1}{6}\,[f(0)+4f(0{,}5)+f(1)].$$

Тогда для отыскания приближенного решения в точках $x=0;\ 0{,}5;\ 1$ получим систему

$$\begin{aligned}
Y_1 &= 1,\\
\tfrac{1}{3}(e^{0,25}-1)\,Y_2+\tfrac{1}{12}(e^{0,5}-1)\,Y_3 &= e^{0,5}-0{,}5,\\
\tfrac{2}{3}(e^{0,5}-1)\,Y_2+\tfrac{1}{6}(e+5)\,Y_3 &= e-1,
\end{aligned}$$

или

$$\begin{aligned}
Y_1 &= 1,\\
1{,}0947\,Y_2+0{,}0541\,Y_3 &= 1{,}1487,\\
0{,}4325\,Y_2+1{,}2864\,Y_3 &= 1{,}7183.
\end{aligned}$$

Решая ее, получим:

$$Y_1=1,\quad Y_2=0{,}9999,\quad Y_3=0{,}9996.$$

Точное решение интегрального уравнения $y(x)\equiv 1$. Как видим, результат достаточно хороший.

2. Решение интегральных уравнений Фредгольма второго рода методом замены ядра на вырожденное. Если в уравнении (2) ядро $K(x,\ s)$ вырожденное, то решение этого уравнения может быть найдено в конечном виде. В самом деле, пусть

$$K(x,\ s)=\sum_{i=1}^{n} A_i(x)\,B_i(s).$$

Можно считать, что $A_1(x)$, $A_2(x)$, ..., $A_n(x)$ и $B_1(x)$, $B_2(x)$, ..., $B_n(x)$ суть системы линейно независимых на отрезке $[a,\ b]$ функций. Так как интегральное уравнение будет иметь вид

$$y(x)-\lambda\sum_{i=1}^{n} A_i(x)\int_a^b B_i(s)\,y(s)\,ds=f(x), \tag{18}$$

то решение его можно искать в виде

$$y(x)=f(x)+\lambda\sum_{i=1}^{n} C_iA_i(x),$$

где C_i — некоторые постоянные. Подставляя $y(x)$ в уравнение (18) и сокращая на λ, получим:

$$\sum_{i=1}^{n} C_i A_i(x) - \lambda \sum_{i=1}^{n} A_i(x) \sum_{j=1}^{n} C_j \int_a^b A_j(s) B_i(s)\, ds - \\ - \sum_{i=1}^{n} A_i(x) \int_a^b f(s) B_i(s)\, ds = 0.$$

Вводя обозначения

$$f_i = \int_a^b f(s) B_i(s)\, ds; \quad \alpha_{ij} = \int_a^b A_j(s) B_i(s)\, ds$$

и принимая во внимание линейную независимость функций $A_1(x)$, $A_2(x)$, ..., $A_n(x)$ для отыскания C_i ($i = 1, 2, \ldots, n$), получим систему линейных алгебраических уравнений

$$C_i - \lambda \sum_{j=1}^{n} \alpha_{ij} C_j = f_i \qquad (i = 1, 2, \ldots, n). \tag{19}$$

Если определитель системы (19) $D(\lambda)$ отличен от нуля, то система имеет единственное решение для $C_1, C_2, \ldots, C_n$ и решение интегрального уравнения $y(x)$ будет найдено в явном виде. Если же при данном значении λ определитель $D(\lambda)$ равен нулю, то λ будет собственным значением ядра $K(x, s)$. В этом случае, находя все линейно независимые решения соответствующей однородной системы, мы в явном виде найдем все линейно независимые между собой собственные функции ядра $K(x, s)$, соответствующие данному собственному значению λ.

Метод приближенного решения интегральных уравнений Фредгольма с помощью замены ядра близким к нему вырожденным ядром основан на следующей теореме:

Теорема. *Если*

$$y(x) - \lambda \int_a^b K(x, s)\, y(s)\, ds = f(x), \tag{20}$$

$$z(x) - \lambda \int_a^b H(x, s)\, z(s)\, ds = f_1(x) \tag{21}$$

— два интегральных уравнения, $R(x, s, \lambda)$ — резольвента второго из этих уравнений и существуют такие константы δ, ε, M,

что имеют место неравенства

$$\int_a^b |K(x, s) - H(x, s)|\, ds < \delta, \tag{22}$$

$$|f(x) - f_1(x)| < \varepsilon, \tag{23}$$

$$\int_a^b |R(x, s, \lambda)|\, ds < M \tag{24}$$

и выполнено условие

$$|\lambda|\,\delta(1 + |\lambda|\,M) < 1, \tag{25}$$

то уравнение (20) *имеет единственное решение* $y(x)$ *и*

$$|y(x) - z(x)| < \frac{N|\lambda|(1 + |\lambda|\,M)^2\delta}{1 - |\lambda|\,\delta(1 + |\lambda|\,M)} + \varepsilon(1 + |\lambda|\,M), \tag{26}$$

где $N = \max\limits_{a \leqslant x \leqslant b} |f(x)|$.

Доказательство. Предполагая существование ограниченного решения (20), обозначим через C_0 верхнюю границу его абсолютной величины на отрезке $[a, b]$ и рассмотрим интегральное уравнение

$$y(x) - \lambda \int_a^b H(x, s)\, y(s)\, ds = F(x),$$

где

$$F(x) = f(x) - \lambda \int_a^b (H(x, s) - K(x, s))\, y(s)\, ds.$$

Тогда

$$y(x) = F(x) + \lambda \int_a^b R(x, s, \lambda)\, F(s)\, ds.$$

Далее, так как

$$|F(x)| \leqslant |f(x)| + |\lambda| \int_a^b |K(x, s) - H(x, s)|\,|y(s)|\, ds \leqslant N + |\lambda|\,\delta C_0,$$

то

$$|y(x)| \leqslant |F(x)| + |\lambda| \int_a^b |R(x, s, \lambda)|\,|F(s)|\, ds \leqslant$$

$$\leqslant N + |\lambda|\,\delta C_0 + |\lambda|\,(N + |\lambda|\,\delta C_0)\, M,$$

или

$$C_0 \leqslant |\lambda|\,(N + |\lambda|\,C_0\delta)\, M + N + |\lambda|\,C_0\delta,$$

откуда

$$C_0 \leqslant \frac{N(1 + |\lambda|\,M)}{1 - |\lambda|\,\delta(1 + |\lambda|\,M)}.$$

Таким образом, при выполнении условия (25) все решения уравнения (20) ограничены одной и той же постоянной, как бы мы ни выбирали $f(x)$, а это означает, что λ не является собственным значением и интегральное уравнение (20) имеет только единственное решение, ибо если бы λ было собственным значением ядра, то, прибавляя к какому-либо решению неоднородного уравнения (20) собственную функцию ядра $K(x, s)$, мы снова получили бы решение уравнения (20). Но собственная функция может быть взята такой, что максимум ее абсолютной величины будет больше любого наперед заданного числа, а это означает, что и для уравнения (20) можно было бы найти решение сколь угодно большое по абсолютной величине.

Докажем теперь оценку (26). Мы имеем:

$$y(x) - z(x) - \lambda \int_a^b H(x, s)(y(s) - z(s))\, ds = F(x) - f_1(x) = \Phi(x)$$

или

$$y(x) - z(x) = \Phi(x) + \lambda \int_a^b R(x, s, \lambda)\Phi(s)\, ds.$$

Отсюда

$$|y(x) - z(x)| \leqslant |\Phi(x)| + |\lambda| \int_a^b |R(x, s, \lambda)|\,|\Phi(s)|\, ds.$$

Но

$$|\Phi(x)| = \left|\lambda \int_a^b (H(x, s) - K(x, s))\, y(s)\, ds + f(x) - f_1(x)\right| \leqslant |\lambda| C_0 \delta + \varepsilon.$$

Следовательно,

$$|y(x) - z(x)| \leqslant \varepsilon + |\lambda| M\varepsilon + C_0\delta |\lambda| (1 + |\lambda| M) \leqslant$$
$$\leqslant |\lambda| M\varepsilon + \varepsilon + \frac{N\delta |\lambda| (1 + |\lambda| M)^2}{1 - |\lambda| \delta (1 + |\lambda| M)}.$$

Из доказанной теоремы следует, что если можно построить достаточно близкое к ядру $K(x, s)$ вырожденное ядро $H(x, s)$, то, решив уравнение с вырожденным ядром $H(x, s)$, мы получим решение, близкое к решению уравнения с ядром $K(x, s)$ при той же правой части. Более того, если мы построим последовательность вырожденных ядер $H_n(x, s)$, равномерно сходящуюся к ядру $K(x, s)$, то последовательность решений $z_n(x)$ уравнений с ядрами $H_n(x, s)$ будет равномерно сходиться к решению $y(x)$ уравнения с ядром $K(x, s)$.

Способы построения вырожденных ядер, близких к данному ядру $K(x, s)$, могут быть самыми различными. Например, ядро $K(x, s)$ можно приближать частичными суммами степенного или двойного

тригонометрического ряда, если ядро $K(x, s)$ разлагается в равномерно сходящийся в прямоугольнике $a \leqslant x,\ s \leqslant b$ степенной или тригонометрический ряд, или приближать его алгебраическими или тригонометрическими интерполяционными многочленами.

Бэтмен предложил определять вырожденное ядро $K_n(x, s)$, аппроксимирующее ядро $K(x, s)$, с помощью равенства

$$\begin{vmatrix} K_n(x, s) & K(x, s_1) & K(x, s_2) & \dots & K(x, s_n) \\ K(x_1, s) & K(x_1, s_1) & K(x_1, s_2) & \dots & K(x_1, s_n) \\ \dots & \dots & \dots & \dots & \dots \\ K(x_n, s) & K(x_n, s_1) & K(x_n, s_2) & \dots & K(x_n, s_n) \end{vmatrix} = 0, \qquad (27)$$

где $x_1, x_2, \dots, x_n$; $s_1, s_2, \dots, s_n$ — некоторые точки отрезка $[a, b]$. Представляя элементы первого столбца в виде

$$K_n(x, s) + 0;\quad 0 + K(x_1, s);\quad 0 + K(x_2, s);\ \dots;\ 0 + K(x_n, s)$$

и разлагая определитель (27) на сумму двух определителей, после несложных преобразований получим явный вид ядра:

$$K_n(x, s) = -\frac{\begin{vmatrix} 0 & K(x, s_1) & K(x, s_2) & \dots & K(x, s_n) \\ K(x_1, s) & K(x_1, s_1) & K(x_1, s_2) & \dots & K(x_1, s_n) \\ \dots & \dots & \dots & \dots & \dots \\ K(x_n, s) & K(x_n, s_1) & K(x_n, s_2) & \dots & K(x_n, s_n) \end{vmatrix}}{\begin{vmatrix} K(x_1, s_1) & K(x_1, s_2) & \dots & K(x_1, s_n) \\ K(x_2, s_1) & K(x_2, s_2) & \dots & K(x_2, s_n) \\ \dots & \dots & \dots & \dots \\ K(x_n, s_1) & K(x_n, s_2) & \dots & K(x_n, s_n) \end{vmatrix}}, \qquad (28)$$

откуда видно сразу, что это ядро вырожденное. Это ядро можно также переписать в виде

$$K_n(x, s) = K(x, s) - \frac{\begin{vmatrix} 0 & K(x, s_1) & K(x, s_2) & \dots & K(x, s_n) \\ K(x_1, s) & K(x_1, s_1) & K(x_1, s_2) & \dots & K(x_1, s_n) \\ \dots & \dots & \dots & \dots & \dots \\ K(x_n, s) & K(x_n, s_1) & K(x_n, s_2) & \dots & K(x_n, s_n) \end{vmatrix}}{\begin{vmatrix} K(x_1, s_1) & K(x_1, s_2) & \dots & K(x_1, s_n) \\ K(x_2, s_1) & K(x_2, s_2) & \dots & K(x_2, s_n) \\ \dots & \dots & \dots & \dots \\ K(x_n, s_1) & K(x_n, s_2) & \dots & K(x_n, s_n) \end{vmatrix}} -$$

$$- K(x, s)\frac{\begin{vmatrix} K(x_1, s_1) & K(x_1, s_2) & \dots & K(x_1, s_n) \\ K(x_2, s_1) & K(x_2, s_2) & \dots & K(x_2, s_n) \\ \dots & \dots & \dots & \dots \\ K(x_n, s_1) & K(x_n, s_2) & \dots & K(x_n, s_n) \end{vmatrix}}{\begin{vmatrix} K(x_1, s_1) & K(x_1, s_2) & \dots & K(x_1, s_n) \\ K(x_2, s_1) & K(x_2, s_2) & \dots & K(x_2, s_n) \\ \dots & \dots & \dots & \dots \\ K(x_n, s_1) & K(x_n, s_2) & \dots & K(x_n, s_n) \end{vmatrix}} =$$

$$= K(x, s) - \frac{\begin{vmatrix} K(x, s) & K(x, s_1) & \dots & K(x, s_n) \\ K(x_1, s) & K(x_1, s_1) & \dots & K(x_1, s_n) \\ \dots & \dots & \dots & \dots \\ K(x_n, s) & K(x_n, s_1) & \dots & K(x_n, s_n) \end{vmatrix}}{\begin{vmatrix} K(x_1, s_1) & K(x_1, s_2) & \dots & K(x_1, s_n) \\ K(x_2, s_1) & K(x_2, s_2) & \dots & K(x_2, s_n) \\ \dots & \dots & \dots & \dots \\ K(x_n, s_1) & K(x_n, s_2) & \dots & K(x_n, s_n) \end{vmatrix}}.$$

Из этого представления видно, что $K_n(x, s)$ совпадает с $K(x, s)$ на $2n$ прямых $x = x_i$; $s = s_i$ $(i = 1, 2, \dots, n)$.

Бэтмен предложил также способ построения резольвенты, а следовательно и явного решения интегрального уравнения, если ядро $K(x, s)$ интегрального уравнения представимо в виде суммы ядра $H(x, s)$, для которого известна резольвента, и вырожденного ядра $H_1(x, s)$. Если ядро

$$K(x, s) = H(x, s) - \sum_{i=1}^{n} f_i(x)\, g_i(s), \tag{29}$$

а $r(x, s, \lambda)$ есть резольвента ядра $H(x, s)$, то резольвента $R(x, s, \lambda)$ ядра $K(x, s)$ имеет вид

$$R(x, s, \lambda) = \frac{\begin{vmatrix} r(x, s, \lambda) & \varphi_1(x) & \varphi_2(x) & \dots & \varphi_n(x) \\ \psi_1(s) & 1+\lambda\tau_{11} & \lambda\tau_{12} & \dots & \lambda\tau_{1n} \\ \dots & \dots & \dots & \dots & \dots \\ \psi_n(s) & \lambda\tau_{n1} & \lambda\tau_{n2} & \dots & 1+\lambda\tau_{nn} \end{vmatrix}}{\begin{vmatrix} 1+\lambda\tau_{11} & \lambda\tau_{12} & \lambda\tau_{13} & \dots & \lambda\tau_{1n} \\ \lambda\tau_{21} & 1+\lambda\tau_{22} & \lambda\tau_{23} & \dots & \lambda\tau_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ \lambda\tau_{n1} & \lambda\tau_{n2} & \lambda\tau_{n3} & \dots & 1+\lambda\tau_{nn} \end{vmatrix}}, \tag{30}$$

где

$$\left.\begin{aligned} \varphi_i(x) &= f_i(x) + \lambda \int_a^b r(x, s, \lambda) f_i(s)\, ds, \\ \psi_i(s) &= g_i(s) + \lambda \int_a^b r(x, s, \lambda) g_i(x)\, dx, \\ \tau_{ij} &= \int_a^b \varphi_j(s)\, g_i(s)\, ds. \end{aligned}\right\} \tag{31}$$

Используя этот результат, иногда целесообразно приближать ядро не вырожденным ядром, а суммой вырожденного ядра и ядра с известной резольвентой.

Для ядра $K_n(x, s)$, построенного по способу Бэтмена, резольвента определяется из уравнения

$$\begin{vmatrix} R_n(x, s, \lambda) & K(x, s_1) & K(x, s_2) & \dots & K(x, s_n) \\ K(x_1, s) & A_{11} & A_{12} & \dots & A_{1n} \\ \dots & \dots & \dots & \dots & \dots \\ K(x_n, s) & A_{n1} & A_{n2} & \dots & A_{nn} \end{vmatrix} = 0, \tag{32}$$

где

$$A_{jk} = K(x_j, s_k) - \lambda \int_a^b K(x_j, t) K(t, s_k)\, dt. \tag{33}$$

Тогда приближенное решение уравнения (2) может быть записано в виде

$$y(x) \approx f(x) + \lambda \int_a^b R_n(x, s, \lambda) f(s)\, ds.$$

Результаты Бэтмена мы приводим без доказательства, которое можно найти в книге Л. В. Канторовича и В. И. Крылова «Приближенные методы высшего анализа».

Пример. Найти решение интегрального уравнения

$$y(x) + \int_0^1 x(e^{xs} - 1) y(s)\, ds = e^x - x.$$

Ядро уравнения $K(x, s) = x(e^{xs} - 1)$ аппроксимируем суммой первых трех членов разложения $K(x, s)$ в ряд Тейлора, т. е. положим

$$H(x, s) = x^2 s + \frac{x^3 s^2}{2} + \frac{x^4 s^3}{6},$$

и вместо исходного уравнения рассмотрим интегральное уравнение

$$z(x) + \int_0^1 H(x, s) z(s)\, ds = e^x - x.$$

Решение последнего ищем в виде

$$z(x) = e^x - x + C_1 x^2 + C_2 x^3 + C_3 x^4.$$

Для определения постоянных C_1, C_2, C_3 получим систему

$$\frac{5}{4} C_1 + \frac{1}{5} C_2 + \frac{1}{6} C_3 = -\frac{2}{3},$$

$$\frac{1}{3} C_1 + \frac{13}{6} C_2 + \frac{1}{7} C_3 = \frac{9}{4} - e,$$

$$\frac{1}{6} C_1 + \frac{1}{7} C_2 + \frac{49}{8} C_3 = 2e - \frac{29}{5}.$$

Решая ее, получим следующий результат:

$$C_1 = -0{,}5010, \quad C_2 = -0{,}1671, \quad C_3 = -0{,}0422,$$

т. е.

$$z(x) = e^x - x - 0{,}5010x^2 - 0{,}1671x^3 - 0{,}0422x^4.$$

Точное решение интегрального уравнения: $y(x) \equiv 1$. Из найденного приближенного решения при $x = 0;\ 0{,}5;\ 1{,}0$ имеем:

$$z(0) = 1{,}0000, \quad z(0{,}5) = 1{,}0000, \quad z(1) = 1{,}0080,$$

т. е. расхождение с точным решением всего 0,008.

3. Метод моментов. В методе моментов приближенное решение интегрального уравнения ищется в виде суммы $f(x)$ и линейной комбинации заранее выбранных линейно независимых между собой функций $\varphi_1(x), \varphi_2(x), \ldots, \varphi_n(x)$, т. е.

$$Y_n(x) = f(x) + \sum_{i=1}^{n} C_i \varphi_i(x) \tag{34}$$

с неопределенными коэффициентами $C_1, C_2, \ldots, C_n$. Коэффициенты $C_1, C_2, \ldots, C_n$ отыскиваются следующим образом. Рассмотрим оператор

$$Lu = u(x) - \lambda \int_a^b K(x, s)\, u(s) ds - f(x). \tag{35}$$

Подставив вместо $u(x)$ функцию $Y_n(x)$, получим:

$$LY_n = \sum_{i=1}^{n} C_i \left\{ \varphi_i(x) - \lambda \int_a^b K(x, s)\, \varphi_i(s)\, ds \right\} - \lambda \int_a^b K(x, s) f(s)\, ds =$$
$$= \Phi(x;\ C_1, C_2, \ldots, C_n).$$

Потребуем ортогональность функции $\Phi(x; C_1, \ldots, C_n)$ ко всем функциям $\varphi_1(x), \varphi_2(x), \ldots, \varphi_n(x)$ на отрезке $[a, b]$, т. е. потребуем выполнения условий

$$\int_a^b LY_n \cdot \varphi_i(x)\, dx = 0 \qquad (i = 1, 2, \ldots, n). \tag{36}$$

Получим систему n линейных алгебраических уравнений для отыскания $C_1, C_2, \ldots, C_n$. Система будет иметь вид

$$\sum_{j=1}^{n} C_j \{\alpha_{ij} - \lambda \beta_{ij}\} = \lambda \gamma_i \qquad (i = 1, 2, \ldots, n), \tag{37}$$

где

$$\alpha_{ij}=\int_a^b \varphi_i(x)\,\varphi_j(x)\,dx;\quad \beta_{ij}=\int_a^b dx\int_a^b K(x,\ s)\,\varphi_i(x)\,\varphi_j(s)\,ds;$$

$$\gamma_i=\int_a^b dx\int_a^b K(x,\ s)\,\varphi_i(x)\,f(s)\,ds. \tag{38}$$

Решая эту систему, мы и найдем $C_1,\ C_2,\ \ldots,\ C_n$, а следовательно и приближенное решение $Y_n(x)$ интегрального уравнения (2).

В основе этого метода лежит следующая идея. Пусть $\varphi_1(x)$, $\varphi_2(x),\ \ldots,\ \varphi_n(x)$ — первые n функций полной ортонормированной системы $\{\varphi_k(x)\}$. Для того чтобы функция $Y(x)$ была точным решением интегрального уравнения (2), необходимо и достаточно ортогональности $LY(x)$ ко всем функциям системы $\{\varphi_k(x)\}$, так как в этом случае будем иметь $LY(x)\equiv 0$. При отыскании приближенного решения функция $Y_n(x)$ содержит лишь n параметров $C_1,\ C_2,\ \ldots,\ C_n$, с помощью которых, вообще говоря, можно удовлетворить лишь n условиям ортогональности, что мы и сделали, потребовав ортогональность LY_n к $\varphi_1(x),\ \varphi_2(x),\ \ldots,\ \varphi_n(x)$. Требование ортонормированности системы $\{\varphi_k(x)\}$ в проведенных рассуждениях излишне. Достаточно требовать лишь полноту системы $\{\varphi_k(x)\}$ и линейную независимость любого конечного числа функций этой системы, так как процессом ортогонализации можно из нее получить полную ортонормированную систему $\{\psi_k(x)\}$ такую, что любая функция $\varphi_k(x)$ будет линейной комбинацией конечного числа функций системы $\{\psi_k(x)\}$. Этот метод есть не что иное, как метод Галеркина решения интегральных уравнений.

Применение метода моментов равносильно замене ядра $K(x,\ s)$ вырожденным ядром $K_n(x,\ s)$, строящимся следующим образом. Предполагая ортонормированность системы $\{\varphi_k(x)\}$, разложим ядро $K(x,\ s)$ как функцию x в ряд Фурье по этой ортонормированной системе функций и за $K_n(x,\ s)$ примем n-ю частичную сумму этого ряда. Получим:

$$K_n(x,\ s)=\sum_{i=1}^{n} u_i(s)\,\varphi_i(x),$$

где

$$u_i(s)=\int_a^b K(x,\ s)\,\varphi_i(x)\,dx.$$

Если теперь к уравнению

$$y(x)-\lambda\int_a^b K_n(x,\ s)\,y(s)\,ds-f(x)=0$$

применить метод моментов, то получим точно такое же решение, как и для уравнения (2), ибо система, аналогичная системе (37), может отличаться от нее только коэффициентами β_{ij}. Обозначим их через $\beta_{ij}^{(n)}$. Будем иметь, учитывая ортонормированность системы функций $\{\varphi_k(x)\}$:

$$\beta_{ij}^{(n)} = \int_a^b dx \int_a^b K_n(x, s)\varphi_i(x)\varphi_j(s)\,ds =$$

$$= \int_a^b dx \int_a^b \sum_{k=1}^n u_k(s)\varphi_k(x)\varphi_i(x)\varphi_j(s)\,ds =$$

$$= \sum_{k=1}^n \int_a^b u_k(s)\varphi_j(s)\,ds \int_a^b \varphi_k(x)\varphi_i(x)\,dx = \int_a^b u_i(s)\varphi_j(s)\,ds.$$

С другой стороны,

$$\beta_{ij} = \int_a^b dx \int_a^b K(x, s)\varphi_i(x)\varphi_j(s)\,ds =$$

$$= \int_a^b \left\{\varphi_j(s) \int_a^b K(x, s)\varphi_i(x)\,dx\right\} ds = \int_a^b \varphi_j(s)\,u_i(s)\,ds.$$

Таким образом,

$$\beta_{ij} = \beta_{ij}^{(n)}.$$

Следовательно, приближенные решения обоих интегральных уравнений совпадают. Но решение $Y_n(x)$ уравнения с вырожденным ядром $K_n(x, s)$, полученное методом моментов, будет его точным решением. Это и показывает равносильность метода моментов методу замены ядра вырожденным ядром, строящимся специальным способом.

Это замечание позволяет использовать оценку, полученную в теореме п. 2, для оценки точности решения, полученного по методу моментов.

Метод моментов можно применять и для решения нелинейных интегральных уравнений, но в этом случае вместо системы (37) получим нелинейную систему.

Пример. Найти два первых собственных значения и соответствующие им собственные функции однородного интегрального уравнения

$$Lu = u(x) - \lambda \int_0^1 K(x, s)\,u(s)\,ds = 0,$$

где

$$K(x, s) = \begin{cases} x(1-s) & (0 \leqslant x \leqslant s \leqslant 1), \\ s(1-x) & (0 \leqslant s \leqslant x \leqslant 1). \end{cases}$$

Для решения задачи применим метод моментов. Приближенное решение будем искать в виде

$$u_3(x) = A + Bx(1-x) + Cx(1-x)(1-2x).$$

Для отыскания коэффициентов A, B, C в соответствии с методом моментов имеем три уравнения:

$$\int_0^1 \varphi_i(x)\, Lu_3(x)\, dx = 0 \qquad (i = 1,\ 2,\ 3),$$

где

$$\varphi_1(x) \equiv 1; \quad \varphi_2(x) \equiv x(1-x); \quad \varphi_3(x) \equiv x(1-x)(1-2x).$$

Подстановка $u_3(x)$ в Lu дает следующий результат:

$$Lu_3(x) = A + Bx(1-x) + Cx(1-x)(1-2x) -$$

$$-\lambda\left\{\frac{A}{2}x(1-x) + \frac{B}{12}[4x^2(1-x)^2 + x^4(1-x) + x(1-x)^4] + \right.$$

$$\left. + \frac{C}{60}[5x^2(1-x)^4 - 5x^4(1-x)^2 + x(1-x)^5 - x^5(1-x)]\right\}.$$

Далее,

$$\int_0^1 \varphi_1 Lu_3(x)\, dx = A\left(1 - \frac{\lambda}{12}\right) + \frac{B}{6}\left(1 - \frac{\lambda}{10}\right) = 0,$$

$$\int_0^1 \varphi_2 Lu_3(x)\, dx = \frac{A}{6}\left(1 - \frac{\lambda}{10}\right) + \frac{B}{30}\left(1 - \frac{17\lambda}{168}\right) = 0,$$

$$\int_0^1 \varphi_3 Lu_3(x)\, dx = \frac{C}{210}\left(1 - \frac{\lambda}{40}\right) = 0.$$

Приравнивая нулю определитель полученной системы, после несложных вычислений получим:

$$(\lambda^2 - 180\lambda + 1680)(\lambda - 40) = 0.$$

Корнями этого уравнения будут:

$$\overline{\lambda}_1 = 9{,}8751, \quad \overline{\lambda}_2 = 40, \quad \overline{\lambda}_3 = 170{,}1249.$$

Подставляя в систему найденные значения $\overline{\lambda}_1$ и $\overline{\lambda}_2$ и решая ее относительно A, B, C, получим:

для $\lambda = \overline{\lambda}_1$

$$A = -0{,}01176B, \quad C = 0,$$

или, определяя B из условия нормировки $\int_0^1 y^2(x)\,dx = 1$ для собственной функции, соответствующей значению $\bar{\lambda}_1$, получим выражение

$$\bar{y}_1(x) = -0{,}0684 + 5{,}817x(1-x);$$

для $\lambda = \bar{\lambda}_2$

$$A = B = 0; \quad C \text{ — произвольное число.}$$

Нормируя, получим собственную функцию, соответствующую второму собственному значению:

$$\bar{y}_2(x) = 14{,}49x(1-x)(1-2x).$$

Точные величины собственных значений этого уравнения:

$$\lambda_1 = \pi^2 = 9{,}8696\ldots, \quad \lambda_2 = 4\pi^2 = 39{,}4784\ldots,$$

а соответствующие им собственные функции:

$$y_1(x) = \sqrt{2}\sin\pi x, \quad y_2(x) = \sqrt{2}\sin 2\pi x.$$

Погрешность первого собственного значения около 0,06%, а второго собственного значения примерно 1,3%. Что касается собственных функций, то приближенное значение первой собственной функции близко к точному, в то время как приближение ко второй функции значительно хуже.

4. Метод наименьших квадратов. Для уравнения (2) будем искать приближенное решение $Y_n(x)$ вида

$$Y_n(x) = \sum_{i=1}^{n} C_i\varphi_i(x), \tag{39}$$

где снова $\varphi_1(x)$, $\varphi_2(x)$, $\ldots$, $\varphi_n(x)$ — некоторые заданные линейно независимые функции. Подстановка $Y_n(x)$ в оператор Lu, где

$$Lu = u(x) - \lambda\int_a^b K(x, s)\,u(s)\,ds - f(x), \tag{40}$$

дает

$$LY_n = \sum_{i=1}^{n} C_i\varphi_i(x) - \lambda\int_a^b K(x, s)\sum_{i=1}^{n} C_i\varphi_i(s)\,ds - f(x) =$$
$$= \Phi(x; C_1, C_2, \ldots, C_n). \tag{41}$$

Постоянные $C_1, C_2, \ldots, C_n$ будем находить из условия минимума интеграла

$$J=\int_a^b \Phi^2(x;\ C_1,\ C_2,\ \ldots,\ C_n)\,dx, \tag{42}$$

т. е. из условий, что

$$\frac{\partial J}{\partial C_i}=0 \qquad (i=1,\ 2,\ \ldots,\ n). \tag{43}$$

Используя явное выражение для функции Φ:

$$\Phi=\sum_{i=1}^{n} C_i\left\{\varphi_i(x)-\lambda\int_a^b K(x,\ s)\,\varphi_i(s)\,ds\right\}-f(x), \tag{44}$$

для отыскания $C_1, C_2, \ldots, C_n$ получим систему линейных алгебраических уравнений

$$\sum_{j=1}^{n} a_{ij}C_j=b_i \qquad (i=1,\ 2,\ \ldots,\ n), \tag{45}$$

где

$$a_{ij}=\int_a^b\left\{\varphi_j(x)-\lambda\int_a^b K(x,\ s)\,\varphi_j(s)\,ds\right\}\times$$

$$\times\left\{\varphi_i(x)-\lambda\int_a^b K(x,\ s)\,\varphi_i(s)\,ds\right\}dx,$$

$$b_i=\int_a^b f(x)\left[\varphi_i(x)-\lambda\int_a^b K(x,\ s)\,\varphi_i(s)\,ds\right]dx. \tag{46}$$

Изложенный метод пригоден и для отыскания приближенных значений первых собственных значений ядра $K(x,\ s)$. Для этого полагаем $f(x)\equiv 0$ и приравниваем нулю определитель системы (45). Получим уравнение n-й степени относительно λ, решая которое и найдем приближенные величины первых собственных значений ядра $K(x,\ s)$.

Пример. Найти решение интегрального уравнения первого рода

$$\int_0^1 K(x,\ s)\,u(s)\,ds=x-2x^3+x^4,$$

где

$$K(x,\ s)=\begin{cases} x(1-s) & (0\leqslant x\leqslant s\leqslant 1),\\ s(1-x) & (0\leqslant s\leqslant x\leqslant 1).\end{cases}$$

К аналогичному уравнению приводит задача об отыскании статической нагрузки, под действием которой струна единичной длины, закрепленная на концах $x=0$ и $x=1$, примет форму, описываемую правой частью уравнения.

Первое приближение к решению будем искать в виде

$$u_1(x)=C_1+C_2x.$$

Тогда

$$Lu_1=x-2x^3+x^4-\int_0^1 K(x,\ s)\,(C_1+C_2s)\,ds=$$

$$=x-2x^3+x^4+\frac{C_1}{2}(x^2-x)+\frac{C_2}{6}(x^3-x)$$

и метод наименьших квадратов дает для отыскания C_1 и C_2 следующую систему уравнений:

$$\int_0^1 Lu_1\frac{\partial Lu_1}{\partial C_1}\,dx=\frac{1}{2}\left[-\frac{17}{3\cdot4\cdot5\cdot7}+\frac{C_1}{3\cdot4\cdot5}+\frac{C_2}{4\cdot5\cdot6}\right]=0,$$

$$\int_0^1 Lu_1\frac{\partial Lu_1}{\partial C_2}\,dx=\frac{1}{6}\left[\frac{17}{5\cdot7\cdot8}+\frac{C_1}{5\cdot8}+\frac{4C_2}{5\cdot7\cdot9}\right]=0,$$

или

$$14C_1+7C_2=34,$$
$$63C_1+32C_2=153.$$

Решение этой системы

$$C_1=2{,}4286,\quad C_2=0,$$

т. е.

$$u_1(x)=2{,}4286.$$

Второе приближение будем искать в виде

$$u_2(x)=C_1+C_2x+C_3x^2.$$

Тогда

$$Lu_2(x)=x-2x^3+x^4-\int_0^1 K(x,\ s)\,(C_1+C_2s+C_3s^2)\,ds=$$

$$=x-2x^3+x^4+\frac{C_1}{2}(x^2-x)+\frac{C_2}{6}(x^3-x)+\frac{C_3}{12}(x^4-x)$$

и метод наименьших квадратов дает для отыскания C_1, C_2, C_3 систему

$$-\frac{17}{3\cdot4\cdot5\cdot7}+\frac{C_1}{3\cdot4\cdot5}+\frac{C_2}{4\cdot5\cdot6}+\frac{5C_3}{4\cdot6\cdot6\cdot7}=0,$$

$$-\frac{17}{5\cdot7\cdot8}+\frac{C_1}{5\cdot8}+\frac{4C_2}{5\cdot7\cdot9}+\frac{11C_3}{3\cdot5\cdot8\cdot12}=0$$

$$-\frac{13}{4\cdot5\cdot9}+\frac{5C_1}{4\cdot6\cdot7}+\frac{11C_2}{3\cdot5\cdot6\cdot8}+\frac{9C_3}{9\cdot12}=0$$

или

$$84C_1+42C_2+25C_3=204,$$
$$252C_1+128C_2+77C_3=612,$$
$$450C_1+231C_2+140C_3=1092.$$

Решение этой системы будет:

$$C_1=0,\quad C_2=12,\quad C_3=-12.$$

Таким образом,

$$u_2(x)=12x(1-x).$$

Это — точное решение уравнения, что легко проверить прямой подстановкой в уравнение.

5. Метод последовательных приближений. Для приближенного решения уравнения (2) может быть применен метод последовательных приближений, заключающийся в следующем.

Решение $y(x)$ уравнения (2) ищем в виде ряда по степеням λ:

$$y(x)=\varphi_0(x)+\lambda\varphi_1(x)+\ldots+\lambda^n\varphi_n(x)+\ldots \tag{47}$$

Подставляя его в уравнение и приравнивая коэффициенты при одинаковых степенях λ, получим:

$$\left.\begin{aligned}
&\varphi_0(x)=f(x),\\
&\varphi_1(x)=\int_a^b K(x,\ s)\,\varphi_0(s)\,ds,\\
&\varphi_2(x)=\int_a^b K(x,\ s)\,\varphi_1(s)\,ds,\\
&\cdots\cdots\cdots\cdots\\
&\varphi_n(x)=\int_a^b K(x,\ s\ \varphi_{n-1}(s)\,ds.\\
&\cdots\cdots\cdots\cdots
\end{aligned}\right\} \tag{48}$$

Если ядро $K(x, s)$ ограничено по модулю постоянной M, то ряд сходится равномерно при $|\lambda| < \frac{1}{M(b-a)}$ и $y(x)$ есть решение интегрального уравнения (2). За приближенное решение можно принять n-ю частичную сумму ряда.

Если $|f(x)| \leqslant N$, то

$$|\varphi_1(x)| \leqslant NM(b-a); \quad |\varphi_2(x)| \leqslant NM^2(b-a)^2; \ \ldots;$$

$$|\varphi_n(x)| \leqslant NM^n(b-a)^n, \ \ldots$$

Отсюда

$$|y(x) - Y_n(x)| = \left| \sum_{k=n+1}^{\infty} \lambda^k \varphi_k(x) \right| \leqslant N \sum_{k=n+1}^{\infty} (|\lambda| M(b-a))^n = \frac{N(|\lambda| M(b-a))^{n+1}}{1 - |\lambda| M(b-a)}. \qquad (49)$$

Это и дает оценку погрешности приближенного решения

$$Y_n(x) = \sum_{i=0}^{n} \lambda^i \varphi_i(x),$$

если все квадратуры вычисляются точно. Если же квадратуры нельзя выполнить, то для вычисления интегралов можно применить те или иные квадратурные формулы. В этом случае удобно пользоваться следующей вычислительной схемой.

Пусть $K_{ij} = K(x_i, x_j)$, где $x_1, x_2, \ldots, x_n$ — абсциссы квадратурной формулы; $\varphi_n(x_i) = \varphi_{ni}$; $y(x_i) = y_i$; $f(x_i) = f_i$. Приближенное значение для $\varphi_n(x_i)$ будем обозначать через $\bar{\varphi}_{ni}$, а для $y(x_i)$ — через Y_i. Если A_i — коэффициенты квадратурной формулы, то

$$\bar{\varphi}_{m+1,\,i} = \sum_{k=1}^{n} A_k K_{ik} \bar{\varphi}_{mk}.$$

Исходя из этого соотношения, вычисления можно вести по следующей схеме:

	1	2	...	...	...	n	φ_{0i}	$\bar{\varphi}_{1i}$	$\bar{\varphi}_{2i}$	...	...	...	Y_m
x_1	$\lambda A_1 K_{11}$	$\lambda A_1 K_{21}$	...	...	...	$\lambda A_1 K_{n1}$	φ_{01}	$\lambda\bar{\varphi}_{11}$	$\lambda^2\bar{\varphi}_{21}$	...	...	...	Y_{m1}
x_2	$\lambda A_2 K_{12}$	$\lambda A_2 K_{22}$	...	...	...	$\lambda A_2 K_{n2}$	φ_{02}	$\lambda\bar{\varphi}_{12}$	$\lambda^2\bar{\varphi}_{22}$	...	...	...	Y_{m2}
...	...	...	...	...	...	...	...	...	...	...	..	.	...
...	...	...	...	...	...	...	...	...	...	...	...	. .	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...
x_n	$\lambda A_n K_{1n}$	$\lambda A_n K_{2n}$	...	...	...	$\lambda A_n K_{nn}$	φ_{0n}	$\lambda\bar{\varphi}_{1n}$	$\lambda^2\bar{\varphi}_{2n}$	...	...	...	Y_{mn}

Сначала в столбцах 1, 2, ..., n выписывается указанная в схеме матрица, а в следующем столбце выписываются значения $\varphi_0(x)=f(x)$ в узлах $x_1, x_2, \ldots, x_n$. Для заполнения следующего столбца умножаем элементы первого столбца на соответствующие элементы столбца свободных членов. Сумма попарных произведений их даст первый элемент нового столбца. Для получения второго элемента этого столбца берем сумму попарных произведений элементов второго столбца на соответствующие элементы столбца свободных членов и т. д. После заполнения столбца $\overline{\varphi_{1i}}$ для заполнения следующего столбца поступаем совершенно аналогично, только вместо столбца свободных членов берем последний полученный столбец. Процесс продолжают до тех пор, пока с точностью, с которой ведутся вычисления, следующий столбец становится нулевым. Значения приближенного решения в узлах получим, суммируя соответствующие элементы вычисленных столбцов по строкам.

6. Приближенное решение уравнений Вольтерра. Из теории интегральных уравнений известно, что если ядро $K(x, s)$ есть непрерывная функция в области $R\{a\leqslant s\leqslant x\leqslant b\}$, а $f(x)$ — непрерывная функция на отрезке $[a, b]$, то интегральное уравнение Вольтерра второго рода

$$y(x)-\lambda\int_a^x K(x, s)\,y(s)\,ds=f(x) \tag{4}$$

имеет единственное непрерывное решение $y(x)$ при любом значении λ. Это решение можно искать в виде

$$y(x)=\sum_{k=0}^{\infty}\lambda^k\varphi_k(x). \tag{50}$$

Подставляя этот ряд в уравнение (4) и сравнивая коэффициенты при одинаковых степенях λ, получим:

$$\varphi_0(x)=f(x);\quad \varphi_{k+1}(x)=\int_a^x K(x, s)\,\varphi_k(s)\,ds. \tag{51}$$

Если

$$N=\max_{a\leqslant x\leqslant b}|f(x)|,\ \text{а}\ M=\max_R|K(x, s)|,$$

то

$$|\varphi_k(x)|\leqslant\frac{M^k(b-a)^kN}{k!}. \tag{52}$$

Отсюда, если мы примем за приближенное решение уравнения (4) n-ю частичную сумму ряда (50):

$$Y_n(x)=\sum_{i=0}^{n}\lambda^k\varphi_k(x), \tag{53}$$

то погрешность его может быть оценена следующим образом:

$$|y(x)-Y_n(x)|=\left|\sum_{k=n+1}^{\infty}\lambda^k\varphi_k(x)\right|\leqslant\sum_{k=n+1}^{\infty}\frac{|\lambda|^kM^k(b-a)^kN}{k!}. \qquad (54)$$

Более грубая, но вместе с тем более простая оценка погрешности следующая: обозначим через L произведение $|\lambda|M(b-a)$ и в оценке (54) вынесем общий множитель $\frac{L^{n+1}N}{(n+1)!}$; получим:

$$|y(x)-Y_n(x)|\leqslant\frac{L^{n+1}N}{(n+1)!}\left\{1+\frac{L}{n+2}+\frac{L^2}{(n+2)(n+3)}+\ldots\right\}.$$

Ряд, стоящий в фигурной скобке, мажорируем рядом

$$1+\frac{L}{n+2}+\left(\frac{L}{n+2}\right)^2+\left(\frac{L}{n+2}\right)^3+\ldots$$

Тогда получим следующую оценку:

$$|y(x)-Y_n(x)|\leqslant\frac{L^{n+1}N}{(n+1)!}\cdot\frac{1}{1-\frac{L}{n+2}}. \qquad (55)$$

При этом предполагается, что n настолько велико, что $L<n+2$.

Если в (51) квадратуры не берутся, то для их вычисления можно использовать квадратурные формулы, лучше всего с равноотстоящими абсциссами. Будем, например, использовать обобщенную формулу трапеций. Если отрезок $[a, b]$ разбить на S равных частей и ввести обозначения $h=\frac{b-a}{S}$; $x_k=a+kh$; $K(x_i, x_j)=K_{ij}$; $\varphi_n(x_k)=\varphi_{nk}$, а приближенные значения для φ_{nk} обозначить через $\bar{\varphi}_{nk}$, то будем иметь:

$$\varphi_{n+1,k}=\int_a^{x_k}K(x_k, s)\varphi_n(s)\,ds\approx$$

$$\approx\frac{h}{2}[K_{k0}\varphi_{n0}+2(K_{k1}\varphi_{n1}+K_{k2}\varphi_{n2}+\ldots+K_{k,k-1}\varphi_{n,k-1})+K_{kk}\varphi_{nk}],$$

или

$$\bar{\varphi}_{n+1,k}=\frac{h}{2}[K_{k0}\bar{\varphi}_{n0}+2(K_{k1}\bar{\varphi}_{n1}+K_{k2}\bar{\varphi}_{n2}+\ldots+K_{k,k-1}\bar{\varphi}_{n,k-1})+$$

$$+K_{kk}\bar{\varphi}_{nk}] \qquad (k=0, 1, 2, \ldots, S). \qquad (56)$$

Вычислив $\bar{\varphi}_{n+1,k}$, мы получим приближенные значения решения интегрального уравнения (4) в узлах x_k по формулам

$$Y_{n,k}=\sum_{i=0}^{n}\lambda^i\bar{\varphi}_{i,k} \qquad (k=0, 1, 2, \ldots, S). \qquad (57)$$

При использовании обобщенной формулы Симпсона разбиваем отрезок $[a, b]$ на $2S$ равных частей точками $x_k = a + kh$; $h = \frac{b-a}{2S}$. Тогда, применяя формулу Симпсона для вычисления интеграла

$$\varphi_{n+1,\,2k} = \int_a^{x_{2k}} K(x_{2k},\ s)\,\varphi_n(s)\,ds,$$

будем иметь:

$$\bar{\varphi}_{n+1,\,2k} = \frac{h}{3}\{K_{2k,\,0}\bar{\varphi}_{n0} + 4(K_{2k,\,1}\bar{\varphi}_{n1} + K_{2k,\,3}\bar{\varphi}_{n3} + \ldots + K_{2k,\,2k-1}\bar{\varphi}_{n,\,2k-1}) + 2(K_{2k,\,2}\bar{\varphi}_{n,\,2} + K_{2k,\,4}\bar{\varphi}_{n,\,4} + \ldots + K_{2k,\,2k-2}\bar{\varphi}_{n,\,2k-2}) + K_{2k,\,2k}\bar{\varphi}_{n,\,2k}\} \quad (58)$$

$$(n = 0,\ 1,\ 2,\ \ldots;\quad k = 1,\ 2,\ 3,\ \ldots,\ S).$$

Значения $\bar{\varphi}_{n+1,\,k}$ для нечетных k придется находить интерполяцией.

Для приближенного решения уравнения (4) можно применять также метод прямой замены интеграла, входящего в уравнение, конечной суммой по какой-либо квадратурной формуле. Например, при использовании обобщенной формулы трапеции, разбивая отрезок $[a, b]$ на n частей точками $x_0 = a$; $x_1 = a + h$; $\ldots$; $x_n = a + nh = b$, будем иметь:

$$y_k - \lambda\int_a^{x_k} K(x_k,\ s)\,y(s)\,ds \approx y_k - \frac{h\lambda}{2}[K_{k0}y_0 + 2(K_{k1}y_1 + K_{k2}y_2 + \ldots + K_{k,\,k-1}y_{k-1}) + K_{k,\,k}y_k] \approx f(x_k),$$

или

$$Y_k - \frac{h\lambda}{2}[K_{k0}Y_0 + 2(K_{k1}Y_1 + \ldots + K_{k,\,k-1}Y_{k-1}) + K_{kk}Y_k] - f(x_k) = 0,$$

откуда

$$Y_k = \frac{1}{1 - \frac{\lambda h}{2}K_{kk}}\left\{f_k + \frac{\lambda h}{2}K_{k0}Y_0 + h\lambda\sum_{i=1}^{k-1}K_{ki}Y_i\right\}. \quad (59)$$

Таким образом, шаг за шагом найдем все значения Y_k.

Что касается интегральных уравнений Вольтерра первого рода

$$\lambda\int_a^x K(x,\ s)\,y(s)\,ds = f(x), \quad (3')$$

то при дополнительном предположении, что ядро $K(x, s)$ и $f(x)$ — непрерывно дифференцируемые функции, $K(x, x) > \alpha > 0$, его можно

свести к интегральному уравнению Вольтерра второго рода. В самом деле, дифференцируя уравнение (3), будем иметь:

$$\lambda K(x,\ x)\,y(x)+\lambda\int_a^x K'_x(x,\ s)\,y(s)\,ds=f'(x),$$

и $y(x)$ будет решением интегрального уравнения Вольтерра второго рода

$$y(x)+\int_a^x \frac{K'_x(x,\ s)}{K(x,\ x)}\,y(s)\,ds=\frac{1}{\lambda}\,\frac{f'(x)}{K(x,\ x)}. \tag{60}$$

Пример. Найти решение интегрального уравнения

$$y(x)-\int_0^x e^{-x-s}y(s)\,ds=\frac{e^{-x}+e^{-3x}}{2}.$$

Первый способ. Ищем решение в виде

$$y(x)\approx Y_4(x)=\varphi_0(x)+\varphi_1(x)+\varphi_2(x)+\varphi_3(x)+\varphi_4(x),$$

где

$$\varphi_0(x)=f(x);\quad \varphi_k(x)=\int_0^x K(x,\ s)\,\varphi_{k-1}(s)\,ds.$$

В результате интегрирования получим:

$$\varphi_0(x)=\frac{1}{2}(e^{-x}+e^{-3x}),$$

$$\varphi_1(x)=\int_0^x e^{-x-s}\varphi_0(s)\,ds=\frac{1}{8}\,[3e^{-x}-2e^{-3x}-e^{-5x}],$$

$$\varphi_2(x)=\int_0^x e^{-x-s}\varphi_1(s)\,ds=\frac{1}{48}\,[5e^{-x}-9e^{-3x}+3e^{-5x}+e^{-7x}],$$

$$\varphi_3(x)=\int_0^x e^{-x-s}\varphi_2(s)\,ds=$$

$$=\frac{1}{384}\,[7e^{-x}-20e^{-3x}+18e^{-5x}-4e^{-7x}-e^{-9x}],$$

$$\varphi_4(x)=\int_0^x e^{-x-s}\varphi_3(s)\,ds=$$

$$=\frac{1}{3840}\,[9e^{-x}-35e^{-3x}+50e^{-5x}-30e^{-7x}+5e^{-9x}+e^{-11x}],$$

откуда

$$Y_4(x)=\frac{1}{3840}\left[3839e^{-x}+5e^{-3x}-10e^{-5x}+10e^{-7x}-5e^{-9x}+e^{-11x}\right].$$

Точное решение этого уравнения

$$y(x)=e^{-x}.$$

Для сравнения приведем значения точного решения и приближенного решения при $x=0$ и $x=1$. Имеем:

$$y(0)=1{,}00000,\quad y(1)=0{,}36788,$$
$$Y_4(0)=1{,}00000,\quad Y_4(1)=0{,}36783.$$

Второй способ. Будем вычислять значения решения в точках

$$x=0;\ 0{,}2;\ 0{,}4;\ 0{,}6;\ 0{,}8;\ 1,$$

используя для замены интеграла в уравнении обобщенную формулу трапеций с шагом $h=0{,}2$. Таблица значений K_{ij} и f_k имеет вид:

x_i	$K_{0,i}$	K_{1i}	K_{2i}	K_{3i}	K_{4i}	K_{5i}	f_i
0	1,00000	0,81873	0,67032	0,54881	0,44933	0,36788	1,00000
1	0,81873	0,67032	0,54881	0,44933	0,36788	0,30119	0,68377
2	0,67032	0,54881	0,44933	0,36788	0,30119	0,24660	0,48576
3	0,54881	0,44933	0,36788	0,30119	0,24660	0,20190	0,35706
4	0,44933	0,36788	0,30119	0,24660	0,20190	0,16530	0,27002
5	0,36788	0,30119	0,24660	0,20190	0,16530	0,13534	0,20883

Вычисления дают следующий результат:

$$Y_0=f_0=1{,}0000,$$

$$Y_1=\frac{1}{1-\frac{h}{2}K_{11}}\left[f_1+\frac{h}{2}K_{10}Y_0\right]=0{,}8206,$$

$$Y_2=\frac{1}{1-\frac{h}{2}K_{22}}\left[f_2+\frac{h}{2}K_{20}Y_0+hK_{21}Y_1\right]=0{,}6731,$$

$$Y_3=\frac{1}{1-\frac{h}{2}K_{33}}\left[f_3+\frac{h}{2}K_{30}Y_0+h\left(K_{31}Y_1+K_{32}Y_2\right)\right]=0{,}5518,$$

$$Y_4=\frac{1}{1-\frac{h}{2}K_{44}}\left[f_4+\frac{h}{2}K_{40}Y_0+h\left(K_{41}Y_1+K_{42}Y_2+K_{43}Y_3\right)\right]=0{,}4522,$$

$$Y_5=\frac{1}{1-\frac{h}{2}K_{55}}\times$$
$$\times\left[f_5+\frac{h}{2}K_{50}Y_0+h\left(K_{51}Y_1+K_{52}Y_2+K_{53}Y_3+K_{54}Y_4\right)\right]=0{,}3705.$$

Ниже приведена таблица значений точного решения и погрешность полученного приближенного решения:

x_k	0	0,2	0,4	0,6	0,8	1
Y_k	1,0000	0,8206	0,6731	0,5518	0,4522	0,3705
$y(x_k)$	1,0000	0,8187	0,6703	0,5488	0,4493	0,3679
$Y_k - y(x_k)$	0,0000	0,0019	0,0028	0,0030	0,0029	0,0026

УПРАЖНЕНИЯ

1. Построить разностную аппроксимацию оператора Лапласа, если сетка состоит из вершин правильных шестиугольников со стороной h.

2. Построить разностную аппроксимацию бигармонического оператора

$$Lu = \frac{\partial^4 u}{\partial x^4} + 2\frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^4 u}{\partial y^4},$$

в которой участвуют узлы: (l, j), $(l, j \pm 2)$, $(l, j \pm 1)$, $(l \pm 1, j)$, $(l \pm 2, j)$, $(l \pm 1, j \pm 1)$, где через (l, j) обозначен узел с координатами $x_l = lh$, $y_j = jl$, а h и l — соответственно шаг сетки по осям x и y. Рассмотреть случай $h = l$.

3. Показать, что для уравнения

$$\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = f(x, t) \qquad (0 \leqslant t \leqslant T)$$

с начальными условиями

$$u(x, 0) = \varphi(x)$$

разностная схема

$$\frac{1}{l}\left[u_{i,\, j+1} - \frac{1}{2}(u_{i+1,\, j} + u_{i-1,\, j})\right] - \frac{1}{2h}(u_{i+1,\, j} - u_{i-1,\, j}) = f(lh,\ jl),$$

где сетка состоит из точек с координатами $x_i = lh$, $t_j = jl$, аппроксимирует уравнение только при $l \geqslant Ch$ ($C = \text{const}$) и корректна при $l \leqslant h$.

4. В области $0 \leqslant x,\ t \leqslant \pi$ дано дифференциальное уравнение

$$\frac{\partial^3 u}{\partial t\, \partial x^2} + \frac{\partial u}{\partial t} = f(x, t)$$

с начальными условиями $u(x, 0) = \varphi(x)$ и граничными условиями $u(0, t) = u(\pi, t) = 0$. Пусть сетка состоит из точек с координатами $x_i = lh$; $t_j = jh$ $\left(l, j = 0, 1, 2, \ldots, n;\ h = \frac{\pi}{n}\right)$. Показать, что разностная схема

$$\frac{1}{h^3}(u_{i+1,\, j} - 2u_{ij} + u_{i-1,\, j} - u_{i+1,\, j-1} + 2u_{i,\, j-1} - u_{i-1,\, j-1}) +$$

$$+ \frac{1}{h}(u_{i,\, j} - u_{i,\, j-1}) = f_{ij};$$

$$u_{i0} = \varphi(lh); \quad u_{0j} = 0; \quad u_{nj} = 0$$

аппроксимирует краевую задачу, равномерно устойчива по начальным значениям и неустойчива по правой части.

5. Пусть в области $0 \leqslant t \leqslant T$; $0 \leqslant x \leqslant a$ задано дифференциальное уравнение

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

с начальными условиями $u(x, 0) = \varphi(x)$ и граничными условиями

$$\left.\frac{\partial u}{\partial x}\right|_{x=0} = \left.\frac{\partial u}{\partial x}\right|_{x=a} = 0.$$

Пусть выбрана сетка точек $x_i = ih$; $t_j = jl \left(i = 0, 1, 2, \ldots, n; \quad h = \frac{a}{n}; \right.$ $\left. j = 0, 1, 2, \ldots, m; \; ml \leqslant T\right)$. Показать, что разностное уравнение

$$\frac{1}{l}(u_{i, j+1} - u_{ij}) = \frac{1}{h^2}(u_{i+1, j} - 2u_{ij} + u_{i-1, j})$$

с граничными условиями

$$u_{i0} = \varphi(ih); \quad u_{1j} - u_{0j} = 0; \quad u_{nj} - u_{n-1, j} = 0 \qquad (*)$$

устойчиво при $\alpha = \frac{l}{h^2} \leqslant \frac{1}{2}$, а если заменить условия (*) на следующие:

$$u_{i0} = \varphi(ih); \quad 3u_{2j} - 4u_{1j} + u_{0j} = 0; \quad 3u_{nj} - 4u_{n-1, j} + u_{n-2, j} = 0,$$

то становится неустойчивым.

Указание. Для доказательства последнего утверждения за начальные данные принять $u_{i0} = \frac{\varepsilon}{3^i}$, где ε — сколь угодно малое число, и найти общее решение разностной схемы.

6. Используя метод прогонки, найти решение уравнения Пуассона

$$\frac{\partial_1 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = -1$$

в квадрате $0 \leqslant x, y \leqslant 1$ с граничными условиями

$$u|_{y=0} = u|_{y=1} = 1; \quad \left[\frac{\partial u}{\partial x} + u\right]_{x=0} = 0; \quad u|_{x=1} = 1,$$

взяв квадратную сетку с шагом $h = 0{,}1$ и простейшую разностную аппроксимацию оператора Лапласа.

7. Методом Ритца найти первые два собственных значения оператора Лапласа, если область G квадрат со стороной 1, а функция на границе обращается в нуль.

8. Различными методами найти решение интегрального уравнения

$$y(x) + \frac{2}{\pi}\int_0^{2\pi} \frac{y(s)}{5 - 3\cos(x+s)}\, ds = \begin{cases} \frac{1}{\pi}\sin x & (0 \leqslant x \leqslant \pi), \\ 0 & (-\pi \leqslant x \leqslant 0). \end{cases}$$

ЛИТЕРАТУРА

1. Л. В. Канторович, В. И. Крылов, Приближенные методы высшего анализа, Гостехиздат, 1952.
2. Л. В. Канторович, Функциональный анализ и прикладная математика, УМН, т. 3, вып. 6, 1948.
3. Л. В. Канторович, Приближенное решение функциональных уравнений, УМН, т. 11, вып. 6, 1956.
4. Л. Коллатц, Численные методы решения дифференциальных уравнений, ИЛ, 1953.
5. О. А. Ладыженская, Метод конечных разностей в теории уравнений в частных производных, УМН, т. 12, вып. 5, 1957.
6. Л. А. Люстерник, О разностных аппроксимациях оператора Лапласа, УМН, т. 9, вып. 2, 1954.
7. В. Э. Милн, Численное решение дифференциальных уравнений, ИЛ, 1955.
8. С. Г. Михлин, Прямые методы в математической физике, Гостехиздат, 1950.
9. Д. Ю. Панов, Справочник по численному решению уравнений в частных производных, Гостехиздат, 1951.
10. Д. Ю. Панов, Численное решение квазилинейных гиперболических уравнений в частных производных, Гостехиздат, 1957.
11. В. С. Рябенький, А. Ф. Филиппов, Об устойчивости разностных уравнений, Гостехиздат, 1956.
12. Richtmayer, Difference methods for initial value problems, NY, 1957.